Modern Statistical Methods
Example sheet 2

1. $x, x' \in \mathbb{R}^p \quad \psi \in \{-1, 1\}^p$

$$\mathbb{E}(\psi^T x \, \psi^T x') = \mathbb{E}((\psi^T x)^T \psi^T x') = \mathbb{E}(x^T \psi \psi^T x')$$

$$= x^T \mathbb{E}(\psi \psi^T) x' = x^T x'$$

$\because (\mathbb{E}(\psi\psi^T))_{ij} = \mathbb{E}(\psi_i \psi_j) = \delta_{ij}$ as ψ_i, ψ_j independent when $i \neq j$. and $\mathbb{E}(\psi_i) = 0$.

Note that $\left[\mathbb{E}(\psi^T x \, \psi^T x')\right]^2 = (x^T x')^2$

$(x^T x')^2 = \mathbb{E}(\psi^T x \, \psi^T x') \, \mathbb{E}(\tilde{\psi}^T x \tilde{\psi}^T x')$ where $\psi, \tilde{\psi} \in \{-1, 1\}^p$ as before

$$= \mathbb{E}(\psi^T x \tilde{\psi}^T x \; \psi^T x' \tilde{\psi}^T x')$$

$$= \mathbb{E}(x^T \psi \tilde{\psi}^T x \; x'^T \psi \tilde{\psi}^T x')$$

Hence $\hat{\phi} : \mathbb{R}^p \to \mathbb{R}$

$$x \longmapsto x^T \psi \tilde{\psi}^T x$$

2. ~~Then~~ $p = 1$ ✓

Suppose $\mathbb{P}_p(M = M') = \frac{|z \cap z'|}{|z \cup z'|}$ if $z, z' \neq \phi$, $z, z' \in \mathcal{X}_p$

Take $z, z' \in \mathcal{X}_p$. Examine $k \in \{1, \dots p\} : \pi(k) = 1$,

if $k \in z, z'$ then $M = M' = 1$

~~otherwise~~ if $k \notin z, z'$ then $z, z' \in \mathcal{X}_{p \setminus \{k\}} \sim \mathcal{X}_{p-1}$ and can completely ignore element k

otherwise $M \neq M'$.

Hence,

~~$\mathbb{P}_p \mathbb{P}_p(M = M') = \mathbb{P}(k \in z \cap z') + \mathbb{P}(k \notin z \cup z')$~~

$$\mathbb{P}_p(M = M') = \mathbb{P}(k \in z \cap z') + \mathbb{P}(k \notin z \cup z') \, \mathbb{P}_{p-1}(M = M')$$

$$= \frac{|z \cap z'|}{p} + \frac{p - |z \cup z'|}{p} \frac{|z \cap z'|}{|z \cup z'|} = \frac{|z \cap z'|}{|z \cup z'|} \qquad \square$$

$\mathbb{E}(\psi_M \psi_{M'}) = \mathbb{P}(M = M') = k(z, z')$ $\because$ ~~$\mathbb{E}[\psi_M \psi_{M'}] = 0$~~ if $M \neq M'$

ψ_M and $\psi_{M'}$ independent

$\Rightarrow$ Jaccard similarity kernel corresponds to random feature map

~~$\hat{\phi}$~~ $\hat{\phi} : z \longmapsto \psi_M$ where $\psi \in \{-1, 1\}^P$

and $M = \min\{\pi(k) : k \in z\}$ for random π over $\mathcal{X}$ the input space.

For large n, beneficial to approx using

$$\Phi_{ij} = \frac{1}{\sqrt{L}} \hat{\phi}_j(x_i) \quad \text{where } i = 1, \dots, n, \; j = 1, \dots, L$$

and L is the number of random feature maps ϕ_j used.

then $\hat{\beta} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T Y$.

3. $\log\left(\frac{\mathbb{P}(Y_i = 1)}{\mathbb{P}(Y_i = -1)}\right) = x_i^T \beta^\circ$ $\qquad \log \mathbb{P}(Y_i = 1) - \log \mathbb{P}(Y_i = -1) = x_i^T \beta^\circ$

Log-likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \log \mathbb{P}(Y_i \mid x_i; \beta) = \sum_{i=1}^{n} \frac{Y_i + 1}{2} \log \mathbb{P}(Y_i = 1) - \frac{Y_i - 1}{2} \log \mathbb{P}(Y_i = -1)$$

$$= \frac{1}{2} \sum_{i=1}^{n} Y_i \left(\log \mathbb{P}(Y_i = 1) - \log \mathbb{P}(Y_i = -1) \right) + \frac{1}{2} \log \mathbb{P}(Y_i = 1) + \frac{1}{2} \log \mathbb{P}(Y_i = -1)$$

$$= \sum_{i=1}^{n} \frac{1}{2} Y_i x_i^T \beta - \frac{1}{2} x_i^T \beta + \log \mathbb{P}(Y_i = 1)$$

$$= \sum_{i=1}^{n} \frac{1}{2} Y_i x_i^T \beta - \frac{1}{2} x_i^T \beta + \log \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}$$

$$= \sum_{i=1}^{n} \frac{1}{2} Y_i x_i^T \beta + \frac{1}{2} x_i^T \beta - \log\left(1 + e^{x_i^T \beta}\right)$$

Maximize wrt β

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{2} Y_i x_i^T + \frac{1}{2} x_i^T - \frac{x_i^T e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} = 0$$

$$\Rightarrow \sum_{i=1}^{n} \log \frac{\exp\left(\frac{1}{2} Y_i x_i^T \beta\right) \exp\left(\frac{1}{2} x_i^T \beta\right)}{\left(1 + \exp(x_i^T \beta)\right)} = \sum_{i=1}^{n} \log \frac{\exp\left(\frac{1}{2} Y_i x_i^T \beta\right)}{\left(\exp\left(-\frac{1}{2} x_i^T \beta\right) + \exp\left(\frac{1}{2} x_i^T \beta\right)\right)}$$

Equivalently minimize

$$\sum_{i=1}^{n} \log\left(\exp\left(-\tfrac{1}{2}(Y_i+1) x_i^T \beta\right) + \exp\left(-\tfrac{1}{2}(Y_i-1) x_i^T \beta\right)\right)$$

$$= \sum_{i=1}^{n} \log\left(1 + \exp\left(-Y_i x_i^T \beta\right)\right) \qquad \text{for } Y_i \in \{1, -1\}$$

$$\therefore \hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \log\left\{1 + \exp\left(-Y_i x_i^T \beta\right)\right\}$$

4. Suppose this algorithm at iteration m is equivalent to m steps of forward selection such that

$$R = Y - X\hat{\beta}_{S_m}$$

$$\hat{\beta}_{S_m} = \underset{\beta_{S_m}}{\arg\min} \|Y - X\beta_{S_m}\|_2^2 \qquad \text{with } S_m = \{k \in \{1, \dots, p\} : \beta_{S_m, k} \neq 0\}$$

The algorithm involves choosing $k^* = \underset{k^* \in M^c}{\arg\max}\ \text{Corr}(R, X_{k^*})$, then

$$R \to R - X_{k^*}\hat{\beta}_{k^*} \quad \text{where} \quad \hat{\beta}_{k^*} = \underset{\beta_{k^*}}{\arg\min} \|R - X_{k^*}\beta_{k^*}\|_2^2$$

$$\to \hat{\beta}_{k^*} = (X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T R$$

$$R^* = R - X_{k^*}\hat{\beta}_{k^*} = Y - X\hat{\beta}_{S_m} - X_{k^*}(X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T (Y - X\hat{\beta}_{S_m})$$

$$= Y - X\hat{\beta}_{S_m} - X_{k^*}(X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T Y$$

because $X_{k^*}^T X \hat{\beta}_{S_m} = 0$:

$$X_{k^*}^T X_k = 0 \quad \forall\ k \in M$$

Suppose $X_k^T X_\ell = 0$ at iteration m $\forall\ k \in M$, $\ell \in M^c$

Then for X_{k^*}, $k^* \in M^c$

$$X_\ell \longrightarrow X_\ell - X_{k^*}\tilde{\beta}_{k^*} \quad \text{with } \tilde{\beta}_k = (X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T X_\ell$$

$$= X_\ell - X_{k^*}(X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T X_\ell$$

Now $X_k^T X_\ell^* = X_k^T X_\ell - X_k^T X_{k^*}(X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T X_\ell = 0$

as $X_k^T X_\ell = 0$ and $X_{k^*}^T X_{k^*} = 0$ as $k^* \in M^c$

and $X_{k^*}^T X_l^* = X_{k^*}^T X_l - X_{k^*}^T X_{k^*} (X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T X_l$

$= X_{k^*}^T X_l - X_{k^*}^T X_l = 0$

Hence, $X_k^T X_l = 0$ at iteration $m+1$ $\forall k \in M^* \cup \{k^*\}$, $l \in M^{*c}$ □

{ Suffice to show that

$R^* \stackrel{?}{=} Y - X \hat{\beta}_{S_m^*}$ where $S_m^* = S_m \cup \{k^F\}$

k^F chosen by forward selection }

{ Know that

$\hat{\beta}_{S_m} = (X_{S_m}^T X_{S_m})^{-1} X_{S_m}^T Y$

$\|R^*\|_2^2 = (R - X_{k^*} \hat{\beta}_{k^*})^T (R - X_{k^*} \hat{\beta}_{k^*}) = \|R\|_2^2 - 2\beta_{k^*} X_{k^*}^T R + \|X_{k^*} \beta_{k^*}\|_2^2$

$= \|R^*\|^2 - \|X_{k^*} \beta_{k^*}\|_2^2$ because $X_{k^*}^T R = X_{k^*}^T Y = (X_{k^*}^T X_{k^*})(X_{k^*}^T X_{k^*})^{-1} X_{k^*}^T Y$

$= \|X_{k^*}\|_2^2 \hat{\beta}_{k^*}$

However,

$k^* = \underset{k^* \in M^c}{\operatorname{argmax}} \operatorname{Corr}(R, X_{k^*}) \Rightarrow \dfrac{\operatorname{Cov}(R, X_{k^*})}{\|R\|_2 \|X_{k^*}\|_2} \geq \dfrac{\operatorname{Cov}(R, X_k)}{\|R\|_2 \|X_k\|_2} \quad \forall k \in M^c$

$\Rightarrow \dfrac{X_{k^*}^T R}{\|R\|_2 \|X_{k^*}\|_2} \geq \dfrac{X_k^T R}{\|R\|_2 \|X_k\|_2}$

$\dfrac{\|X_{k^*}\|_2^2 \hat{\beta}_{k^*}}{\|X_{k^*}\|_2} \geq \dfrac{\|X_k\|_2^2 \hat{\beta}_k}{\|X_k\|_2}$

$\Rightarrow \|X_{k^*} \hat{\beta}_{k^*}\|_2 \geq \|X_k \hat{\beta}_k\|_2$

$\therefore \|R^*\|_2^2 = \|R\|^2 - \|X_{k^*} \hat{\beta}_{k^*}\|_2^2 \leq \|R\|_2^2 - \|X_k \hat{\beta}_k\|_2^2 \quad \forall k \in M^c$

Hence, k^* minimizes $\|R - X_k \beta_k\|_2^2$, $k \in M^c$.

On the other hand,

$\|Y - X \hat{\beta}_{S_m} - X_k \hat{\beta}_k\|$

$= \|Y - X \hat{\beta}_{S_m^*}\|$ with cols of X modified according to this algorithm. Hence equivalent to forward selection on transf'm cols

5. $\mathbb{E}W = 0 \qquad \mathbb{E}e^{\alpha W} \le e^{\alpha^2\sigma^2/2} \quad \forall \alpha \in \mathbb{R}$

$\mathrm{Var}(W) = \mathbb{E}\left((W - \mathbb{E}W)^2\right) = \mathbb{E}(W^2)$

$$\frac{d}{d\alpha}\mathbb{E}e^{\alpha W} = \mathbb{E}\frac{d}{d\alpha}e^{\alpha W} = \mathbb{E}We^{\alpha W}$$

$$\frac{d}{d\alpha}\mathbb{E}e^{\alpha W} \le \frac{d}{d\alpha}\left(e^{\alpha^2\sigma^2/2}\right) = \alpha\sigma^2 e^{\alpha^2\sigma^2/2}$$

$$\frac{d^2}{d\alpha^2}\mathbb{E}e^{\alpha W} = \mathbb{E}W^2e^{\alpha W} \le (\sigma^2 + \alpha^2\sigma^4)e^{\alpha^2\sigma^2/2}$$

At $\alpha = 0$,

$\mathbb{E}W^2 \le \sigma^2$ □

6. Hoeffding's lemma: $\mathbb{E}W = 0$, $W \in [a,b] \Rightarrow$ sub-G, $\sigma = (b-a)/2$

If $W \in \{-1, 1\}$ with equal probability,

$\mathbb{E}W = 0$

$$\mathbb{E}e^{\alpha W} = \frac{1}{2}e^{\alpha(-1)} + \frac{1}{2}e^{\alpha} = \frac{1}{2}\left(e^{\alpha} + e^{-\alpha}\right) \le e^{\alpha^2/2}$$

$\Rightarrow \sigma = 1$ as required

7(a) $W \sim \chi^2_d \sim \sum^{d} \chi^2_1$, $W = \sum_{i=1}^{d} W_i$, $W_i \sim \chi^2_1$

$$\mathbb{P}\left(\left|W/d - 1\right| \ge t\right) \le \inf e^{-\alpha dt}\,\mathbb{E}e^{\alpha(W-d)}$$

$$\mathbb{E}e^{\alpha(W-d)} = \mathbb{E}e^{\sum_{i=1}^{d}\alpha(W_i-1)} = \prod_{i=1}^{d}\mathbb{E}e^{\alpha(W_i-1)} = \prod_{i=1}^{d}\frac{\alpha}{e^{\alpha}\sqrt{1-2\alpha}} = (1-2\alpha)^{-d/2}e^{-\alpha d} \text{ for } \alpha < \tfrac{1}{2}$$

$$\Rightarrow \mathbb{P}\left(\left|W/d - 1\right| \ge t\right) \le \inf_{\alpha \ge 0} e^{-\alpha d(t+1)}(1-2\alpha)^{-d/2} \le \inf_{|\alpha| < 1/4} e^{(2\alpha^2 - \frac{t}{2}\alpha)d}$$

$$\le e^{-dt^2/8}$$

(b) $A \in \mathbb{R}^{d\times p} \sim N(0,1)$, $u \in \mathbb{R}^p$

$$\frac{\|Au\|_2^2}{\|u\|_2^2} = \frac{\sum_{i=1}^{d}\left(\sum_{j=1}^{p}A_{ij}u_j\right)^2}{\sum_{j=1}^{p}u_j^2} = \frac{\sum_{i=1}^{d}\|A_i u\|^2}{\sum_{i=1}^{p}u_i^2} = \frac{\sum_{i=1}^{d}\|A_i u\|_2^2}{\|u\|_2^2} \le \sum_{i=1}^{d}\|A_i\|_2^2\frac{\|u\|_2^2}{\|u\|_2^2}$$

$$\sum_{i=1}^{d}\|A_i\|_2^2\,\frac{\|A_i u\|_2^2}{\|u\|_2^2} = \left\|A\frac{u}{\|u\|_2}\right\|_2^2 \quad \text{but} \quad \frac{A_i u}{\|u\|_2} \sim N(0,1)$$

Thus $\|Au\|_2^2/\|u\|_2^2$ is the sum of d $N(0,1)$ vars $\sim \chi^2_d$ and the result follows from previous part.

(c) $u_i \longmapsto Au_i/\sqrt{d} = w_i$

$$\mathbb{P}\left(1-t \le \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \le 1+t \quad \forall i,j \in \{1,\dots,n\}, i \neq j\right)$$

$$= \mathbb{P}\left(1-t \le \frac{\|A(u_i-u_j)\|_2^2}{d\|u_i-u_j\|_2^2} \le 1+t \quad \forall i,j, i \neq j\right)$$

$$= \mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|_2^2}{d\|u_i-u_j\|_2^2} - 1\right| \le t \quad \forall i,j \in \{1,\dots,n\}, \; i<j\right)$$

(can constrain $i<j$ or equal under $u_i \sim u_j$)

$$= 1 - \mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|_2^2}{d\|u_i-u_j\|_2^2} - 1\right| > t \quad \exists i,j \in \{1,\dots,n\}, \; i<j\right)$$

$$\ge 1 - \sum_{\substack{i,j \in \{1,\dots,n\} \\ i<j}} \mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|_2^2}{d\|u_i-u_j\|_2^2} - 1\right| > t\right) \qquad \left(\text{as } \mathbb{P}\left(\bigcup_i \Omega_i\right) \le \sum_i \mathbb{P}(\Omega_i)\right)$$

$$\ge 1 - \sum_{\substack{i,j \in \{1,\dots,n\} \\ i<j}} \mathbb{P}\left(\left|\frac{\|A(u_i-u_j)\|_2^2}{d\|u_i-u_j\|_2^2} - 1\right| \ge t\right) \qquad \text{treating } u_i - u_j \text{ as fixed}$$

$$\ge 1 - \sum_{\substack{i,j \in \{1,\dots,n\} \\ i<j}} 2e^{-dt^2/8} = 1 - \frac{n(n-1)}{2} 2e^{-dt^2/8}$$

$$= 1 - n(n-1)e^{-dt^2/8} \qquad \text{use } d > 16\log(n/\sqrt{\epsilon})/t^2$$

$$\Rightarrow dt^2/8 > 2\log(n/\sqrt{\epsilon})$$

$$> 1 - n(n-1)e^{-2\log(n/\sqrt{\epsilon})} = 1 - \frac{n(n-1)}{n^2}\epsilon = 1 - \epsilon + \frac{\epsilon}{n} > 1 - \epsilon$$

$$\therefore \mathbb{P}\left(1-t \le \frac{\|w_i - w_j\|_2^2}{\|u_i - u_j\|_2^2} \le 1+t \quad \forall i,j \in \{1,\dots,n\}, i \neq j\right) > 1 - \epsilon \quad \square$$

8. $Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$

Any minimizer of Q_λ, $\hat{\beta}_\lambda^L$ must also minimize

$\|Y - X\beta\|_2^2$ subject to $\|\beta\|_1 \le \|\hat{\beta}_\lambda^L\|_1$

$\because$ If $\|\beta\|_1 \le \|\hat{\beta}_\lambda^L\|_1$ and $\|Y - X\beta\|_2^2 < \|Y - X\hat{\beta}_\lambda^L\|_2^2$, then $Q_\lambda(\beta) < Q_\lambda(\hat{\beta}_\lambda^L)$ ##.

Suppose two Lasso solutions $\hat{\beta}_\lambda^1$, $\hat{\beta}_\lambda^2$ have ℓ_1-norms

$\|\hat{\beta}_\lambda^1\|_1 \le \|\hat{\beta}_\lambda^2\|$.

However, from above, $\hat{\beta}_\lambda^2$ must also minimize

$\|Y - X\beta\|_2^2$ subject to $\|\beta\|_1 \le \|\hat{\beta}_\lambda^1\|_1$

$\therefore \|\hat{\beta}_\lambda^1\|_1 = \|\hat{\beta}_\lambda^2\|_1$ □

9. Consider $S = \{-\mathrm{sgn}(\beta_1)X_1, \ldots, -\mathrm{sgn}(\beta_p)X_p, Y\} \subseteq \mathbb{R}^n$

and construct the convex combination

$\alpha_0 Y - \alpha_1 \mathrm{sgn}(\beta_1)X_1 + \ldots - \mathrm{sgn}(\beta_p)\alpha_p X_p = v$, $\sum_{i=0}^{p} \alpha_i = 1$

then by Caratheodory's Lemma

$v = \alpha_0 Y - \alpha_1 \mathrm{sgn}(\beta_{j_1})X_{j_1} - \ldots - \alpha_n \mathrm{sgn}(\beta_{j_n})X_{j_n}$ $\sum_{i=1}^{n} \alpha_i = 1$

$\Rightarrow$ Any $Y - X\beta$ can be expressed with no more than n non-zero coeffs.

Above $\tilde{\beta}_i = \alpha_i \mathrm{sgn}(\beta_{j_i})$, $\|\tilde{\beta}\|_1 = \sum_i \alpha_i = 1$

$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$

$= \frac{1}{2n}\|Y - X\frac{\tilde{\beta}}{\lambda}\|_2^2 + \|\tilde{\beta}\|_1 = \frac{1}{2n\lambda}\|\lambda Y - X\tilde{\beta}\|_2^2 + \|\tilde{\beta}\|_1$

Identify $\lambda \equiv \alpha_0$ to show equivalence.

10. $\lambda \geq \lambda_{max} := \|X^T Y\|_\infty / n$

Then

$$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2n}\|Y - X\beta\|_2^2 + \frac{1}{n}\|X^T Y\|_\infty \|\beta\|_1$$

$$\geq \frac{1}{2n}\|Y - X\beta\|_2^2 + \frac{1}{n}\|(X\beta)^T Y\|_1 \quad \text{by Hölder's inequality}$$

$$= \frac{1}{2n}\left(\|Y\|_2^2 + \|X\beta\|_2^2\right) - 2\frac{1}{n}(X\beta)^T Y + \frac{1}{n}|(X\beta)^T Y|$$

$$\geq \frac{1}{2n}\left(\|Y\|_2^2 + \|X\beta\|_2^2\right) \geq \frac{1}{2n}\|Y\|_2^2 = Q_\lambda(0)$$

Hence, $\hat{\beta}_\lambda^L = 0$ always minimises the objective and is a solution (unique by Qu 2).

11. $$\frac{\partial Q_\lambda(\beta)}{\partial \beta_k} = -\frac{1}{n} X_k^T (Y - X\beta) + \lambda \operatorname{sgn}(\beta_k)$$

$$= -\frac{1}{n} X_k^T Y + \frac{1}{n} X_k^T X\beta + \lambda \operatorname{sgn}(\beta_k)$$

$$= -\frac{1}{n} X_k^T Y + \frac{1}{n}\|X_k\|_2^2 \beta_k + \lambda \operatorname{sgn}(\beta_k) \quad \text{as } X_k \text{ orthogonal}$$

$$= -\frac{1}{n} X_k^T Y + \beta_k + \lambda \operatorname{sgn}(\beta_k)$$

$$\left.\frac{\partial Q_\lambda(\beta)}{\partial \beta_k}\right|_{\hat{\beta}_{\lambda,k}^L} = 0 \quad \Rightarrow -\frac{1}{n} X_k^T Y + \hat{\beta}_{\lambda,k}^L + \lambda \operatorname{sgn}\left(\hat{\beta}_{\lambda,k}^L\right) = 0$$

$\hat{\beta}^{OLS} = (X^T X)^{-1} X^T Y = \frac{1}{n} X^T Y$ when X_k orthogonal, l_2 norm $\sqrt{n}$

$$\hat{\beta}_k^{OLS} = \frac{1}{n} X_k^T Y$$

$$\Rightarrow -\hat{\beta}_k^{OLS} + \hat{\beta}_{\lambda,k}^L + \lambda \operatorname{sgn}\left(\hat{\beta}_{\lambda,k}^L\right) = 0$$

$$\therefore \hat{\beta}_{\lambda,k}^L = \begin{cases} \hat{\beta}_k^{OLS} - \lambda & : \hat{\beta}_{\lambda,k}^L > 0 \\ \hat{\beta}_k^{OLS} + \lambda & : \hat{\beta}_{\lambda,k}^L < 0 \end{cases} \quad \text{no solution otherwise}$$

$$\hat{\beta}_{\lambda,k}^L = \left(|\hat{\beta}_k^{OLS}| - \lambda\right)_+ \operatorname{sgn}\left(\hat{\beta}_k^{OLS}\right) \quad \square$$

For l_0 penalty

$-\hat{\beta}_k^{OLS} + \hat{\beta}_{\lambda,k} = 0$ if $\hat{\beta}_{\lambda,k} \neq 0$

$\lim_{\hat{\beta}_{\lambda,k} \to 0_+} Q_\lambda(\beta^*, \beta_{\lambda,k} \neq 0) - Q_\lambda(\beta, \beta_{\lambda,k} = 0) = \lambda$

Examine when the objective changes as $\beta_k = 0 \to \neq 0$,

$\Delta Q = \frac{1}{2n} \|Y - X^*\beta - X_k\beta_k\|_2^2 + \lambda - \frac{1}{2n}\|Y - X^*\beta\|_2^2$

$= \frac{1}{n}\beta_k X_k^T (Y - X^*\beta) + \frac{1}{2n}\|X_k\beta_k\|_2^2 + \lambda$ where X^* is X with X_k removed

$= -\frac{1}{n}\beta_k X_k^T Y + \frac{1}{2n}\beta_k^2 n + \lambda$ by orthogonality

$= -\beta_k \hat{\beta}_k^{OLS} + \frac{1}{2}\beta_k^2 + \lambda$

$\hat{\beta}_{\lambda,k} \neq 0$ when $-\hat{\beta}_k^{OLS\,2} + \frac{1}{2}\hat{\beta}_k^{OLS\,2} + \lambda \leq 0$

$\Rightarrow \hat{\beta}_{k,\lambda} = \begin{cases} 0 & : \hat{\beta}_k^{OLS} < \sqrt{2\lambda} \\ \hat{\beta}_k^{OLS} & : \hat{\beta}_k^{OLS} \geq \sqrt{2\lambda} \end{cases}$