1 P2

- Let's assume $X\beta_1 \neq X\beta_2$.

  Let f^* be the optimal value, $\alpha = \frac{1}{2}$, $\beta_\alpha = \alpha\beta_1 + (1-\alpha)\beta_2$. Then, by the convexity of $\|.\|_2^2, \|.\|_1$:

  $$\begin{aligned} f^* &\leq ||Y - X\beta_\alpha||_2^2 + \lambda||\beta_\alpha||_1 \\ &= \|\alpha(Y - X\beta_1) + (1-\alpha)(Y - X\beta_2)\|_2^2 + \lambda\|\alpha\beta_1 + (1-\alpha)\beta_2\|_1 \\ &< \alpha\left(||Y - X\beta_1||_2^2 + \lambda||\beta_1||_1\right) + (1-\alpha)\left(||Y - X\beta_2||_2^2 + \lambda||\beta_2||_1\right) p \quad \text{(By strict convexity o} \\ &\leq f^* \end{aligned}$$

  Contradicition.

- $\mathcal{L}(\beta^*, \lambda) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$

  $\partial\|\beta\|_1 = \{\alpha \in [-1,1]^n, \alpha_j = sign(\hat{\beta}_j)$ when $\hat{\beta}_j \neq 0\}$

  Let (β^*, λ^*) be an optimal solution, then $0 \in \partial_\lambda L(\beta^*, \lambda^*)$

  $\partial_{\lambda^*} L(\beta, \lambda^*) = -X^T(Y - X\beta) + \lambda^*\partial\|\beta\|_1$

  Coordinate wise, this gives for all j:

  $X_j^T(Y - X\beta) = \lambda sign(\beta_j)$ if $\beta_j \neq 0$

  $-X(Y - X\beta) = \lambda\alpha_i$ if $\beta_j = 0$

  e.g

  $\lambda^* = -sign(\beta_j^*)X_j^T(Y - X\beta^*)$ if $\beta_j^* \neq 0$

  $\lambda^* \geq |2X_j^T(Y - X\beta^*)|$ if $\beta_j^* = 0$

- Let $\hat{\beta}$ be an optimal solution. Let $\chi = \{j, \hat{\beta}_j \neq 0\}$, and let's suppose it is non empty. Let j such that $\hat{\beta}_j > 0$ (If such j exists)

  By 2.2, $\lambda = X_j^T(Y - X\hat{\beta})$, but since $\lambda > \|X^TY\|_\infty \geq X_j^TY$, then $X_j^TX\hat{\beta} > 0$.

  Similarly, if there for j such that $\hat{\beta} < 0$, $X_j^TX\hat{\beta} < 0$.

  c/c $\beta_j \neq 0 \implies \beta_jX_j^TX\hat{\beta} > 0$

$$
\begin{aligned}
\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 &= \frac{1}{2}\|Y\|_2^2 - \hat{\beta}^T X^T Y + \frac{1}{2}\beta^T X^T X\hat{\beta} + \lambda \sum_{i\in\chi} |\hat{\beta}_i| \\
&\geq \frac{1}{2}\|Y\|_2^2 + \sum_{i\in\chi} |\hat{\beta}_i|(\lambda - |X_i^T Y|) + \frac{1}{2} \underbrace{\sum_{i\in\chi} \hat{\beta}_i X_i^T X\hat{\beta}}_{>0} \\
&> \frac{1}{2}\|Y\|_2^2 \\
&= \frac{1}{2}\|Y - X0\|_2^2 + \lambda\|0\|_1
\end{aligned}
$$

Contradiction, so $\hat{\beta} = 0$

- $$\lambda \in [\lambda_0, \lambda_1]$$

  Let $\chi(\lambda) = \{j, \hat{\beta}_j(\lambda) \neq 0\} := \chi$, $r = |\chi|$ (doesn't depend on λ by assumption)

  We have proved in 2.2 that there exist $\alpha(\lambda)$

  $$X^T(Y - X\hat{\beta}(\lambda)) = \lambda\alpha(\lambda)$$

  where $\alpha(\lambda) \in \partial\|\hat{\beta}(\lambda)\|_1$.

  It is easy to see that this KKT conditions is actualy necessary and sufficient (because we are minimizing a convexe function), since we are assuming uniqueness, $\hat{\beta}(\lambda)$ is the unique solution to :

  $$(\exists\alpha(\lambda) \in \partial\|\hat{\beta}(\lambda)\|_1)\ X^T(Y - X\hat{\beta}(\lambda)) = \lambda\alpha(\lambda)$$

  Note that by uniqueness of $X\beta$ and $\hat{\beta}(\lambda)$, $\alpha(\lambda)$ is unique when $\lambda > 0$.

  Note also, that since we assumed that the signs and support are unchanged, $\partial\|\hat{\beta}(\lambda)\|_1 = \partial\|\hat{\beta}(\lambda_0)\|_1$.

  The last condition becomes:

  $$X^T(Y - X\hat{\beta}(\lambda)) \in \lambda\partial\|\hat{\beta}(\lambda_0)\|_1$$

Notation: $\alpha(\lambda_0) = X^T \underbrace{\frac{(Y - X\hat{\beta}(\lambda_0))}{\lambda_0}}_{v} = X^T v,\ \gamma_0 = X^\dagger v,\ \delta = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\gamma_0$.

Note that:

$X^T X \gamma_0 = X^T X X^\dagger v = (V\Lambda U^T)(U\Lambda V^T)(V\Lambda^{-1}U^T)v = V\Lambda U^T v = X^T v = \alpha(\lambda_0)$

$$\begin{aligned} X^T(Y - X\delta) &= \underbrace{X^T(Y - X\hat{\beta}(\lambda_0))}_{\lambda_0\alpha(\lambda_0)} + (\lambda - \lambda_0)\underbrace{X^T X\alpha_0}_{\alpha(\lambda_0)} \\ &= \lambda\alpha(\lambda_0) \in \lambda\partial\|\hat{\beta}(\lambda_0)\|_1 \end{aligned}$$

Which proves that $\hat{\beta}(\lambda) = \delta = \hat{\beta}(\lambda_0) - (\lambda - \lambda_0)\alpha(\lambda_0)$

- **Notation**: For a vector v, let $v^+ = \max(v, 0), v^- = -\min(-v, 0)$, $sign(v), supp(v)$ the sign and support of v, $\phi(v) = (supp(v^+), supp(v^-))$

  The number of values $\phi(v)$ can take is finite and at most n^2 because $\phi(v) \in \mathcal{P}(\{1 \dots n\})^2$.

  Notice that in the last part, we have proven a stronger result: if for λ_1, λ_2, $\phi(\beta(\lambda_1)) = \phi(\beta(\lambda_2))$, then $\beta(\lambda_2) = \beta(\lambda_1) - (\lambda_2 - \lambda_1)\gamma_0$, where γ_0 depend only on λ_1. This proves a segment of the path C is fully caracterized by the $\phi(v)$ where $v(C)$ is one of the element of C chosen arbitrarly.

  Let $\mathcal{A}$ denote the set of segments that form the lasso path, and consider the following application:

  $\mathcal{A} \to \mathcal{B}; C \to \phi(v(C))$ Where v is an arbitrary element in C.

  We have proven that this application is injective, so $|\mathcal{A}| \leq n^2 < \infty$.

2 P3

- Let's consider the unconstrained optimization problem:

$$\min ||Y - X\beta||^2$$

β is optimal iff $X^TY = X^TX\beta$.

We check easily that $(X^TX)^\dagger X^TY$ is a solution to the last equation, therefore it minimizes the L_2 risk.

If $t > \|(X^TX)^\dagger X^TY\|_{L_1}$, then it is also solution to the following problem: $\min_{\|\beta\|_{L_1}\leq t} ||Y - X\beta||^2$.

- 1.) $X_i, Y_i, i \in V_k$ and $\hat{\beta}_{\hat{t}}^{V_k}$ are independent. $(Y - X^T\hat{\beta}_{\hat{t}}^{V_k})^2 \leq |Y|^2 + \|X\|_\infty^2 \|\hat{\beta}_{\hat{t}}^{V_k}\|_1^2 \leq b^2(1+\hat{t}^2) \leq b^2(1+t_n^2)$

$$\underset{X_i,Y_i,i\in V_k}{\mathbb{P}}\left(\left|\frac{1}{|V_k|}\sum_{i\in V_k}(Y_i - X_i^T\hat{\beta}_{\hat{t}}^{V_k})^2 - \mathbb{E}_{X,Y}[(Y - X^T\hat{\beta}_{\hat{t}}^{V_k})^2]\right| > \varepsilon\right) \leq 2\exp(-\frac{|V_k|\varepsilon^2}{2b^4(1+t_n^2)})$$

$$\mathbb{P}\left(\frac{1}{K}\sum_k R(\hat{\beta}_{\hat{t}}^{V_k}) - \hat{R}_{CV}(\hat{t}) > \varepsilon\right) \leq 2K\exp(-\frac{n\varepsilon^2}{2Kb^4(1+t_n^2)})$$

Since R is convexe, $\frac{1}{K}\sum_k R(\hat{\beta}_{\hat{t}}^{V_k}) \geq R(\frac{1}{K}\sum_k \hat{\beta}_{\hat{t}}^{V_k}) \geq R(\hat{\beta}_{\hat{t}})$, so:

$$\mathbb{P}\left(R(\hat{\beta}_{\hat{t}}) - \frac{1}{K}\sum_k R(\hat{\beta}_{\hat{t}}^{V_k}) - \hat{R}_{CV}(\hat{t}) > \varepsilon\right) \leq 2K\exp(-\frac{n\varepsilon^2}{2Kb^4(1+t_n^2)})$$

2.)

$$\hat{R}_{CV}(\hat{t}) - \hat{R}_{CV}(t_{\max}) \leq 0$$

4.)

$$\hat{R}(\hat{\beta}_{t_{\max}}) = \hat{R}(\hat{\beta}_{t_n})$$

5.) Hoefdding:

$$\mathbb{P}(\hat{R}(\hat{\beta}_{t_n}) - R(\hat{\beta}_{t_n}) > \varepsilon) \leq 2\exp(-\frac{n\varepsilon^2}{2b^4(1+t_n^2)})$$

6.) We did this one in class:

$$\mathbb{P}\left(R(\hat{\beta}) - R(\beta^*) > 2(1+t_n^2)\sqrt{\frac{2b^4\log(\frac{2d^2}{\delta})}{n}}\right) \leq \delta$$

3 Proof of Hoeffding

- $$\begin{aligned}
\mathbb{P}(X \geq t) &= \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) && \text{(because exp is increasing)} \\
&= \mathbb{P}(e^{\lambda(X-t)} \geq 1) \\
&\leq \mathbb{E}[e^{\lambda(X-t)}] && \text{(Markov inequality)} \\
&= e^{-\lambda t} \prod \mathbb{E} e^{\lambda x_i} && \text{(By independence)} \\
&= e^{-\lambda t} (\mathbb{E} e^{\lambda x_i})^k && x_i \sim x_1 \\
&= e^{-\lambda t} (\frac{1}{2}(e^{\lambda} + e^{-\lambda}))^k
\end{aligned}$$

- $e^{\lambda} + e^{-\lambda} = \sum^{\infty} \frac{\lambda^k}{k!} + (-1)^k \frac{\lambda^k}{k!} = 2 \sum^{\infty} \frac{\lambda^{2k}}{(2k!)} \; e^{\frac{1}{2}\lambda^2} = \sum \frac{\lambda^{2k}}{2^k k!}$ Since $\frac{(2k)!}{2^k k!} = \prod_{j=1\ldots 2k} \frac{k+j}{2} \geq 1$, $\frac{1}{2} e^{\lambda} + \frac{1}{2} e^{-\lambda} \leq e^{\frac{1}{2}\lambda^2}$

- $\mathbb{P}(X \geq t) \leq e^{-\lambda t} (\frac{1}{2} e^{\lambda} + \frac{1}{2} e^{-\lambda})^k \leq e^{-\lambda t} e^{\frac{k}{2}\lambda^2}$ $\lambda \to \frac{k}{2}\lambda^2 - \lambda t$ is quadratic function that attain its minimum for $\lambda = \frac{t}{k}$, and the minimum is $-\frac{t^2}{k}$, therefore $\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{k}}$.