2.2 The Lasso estimator

The Lasso (Tibshirani 1996) performs

$$(\hat{\mu}^L_\lambda, \hat{\beta}^L_\lambda) = \underset{(\mu,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Y - \mu\mathbf{1} - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

Here $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$

After centering and scaling the cols of X (to have l_2-norm $\sqrt{n}$) and centering Y, we can remove μ from the objective

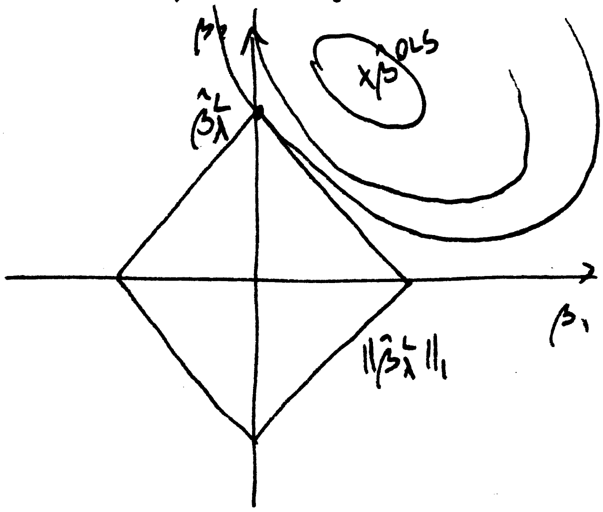
$$Q_\lambda(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

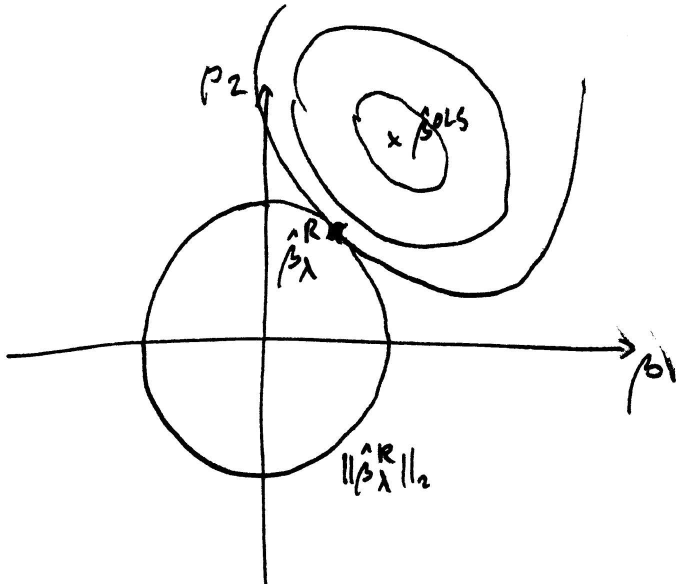
Any minimiser of Q_λ, $\hat{\beta}^L_\lambda$ must also minimise

$$\|Y - X\beta\|_2^2 \text{ subj to } \|\beta\|_1 \le \|\hat{\beta}^L_\lambda\|_1$$

Indeed if β was $\|\beta\|_1 \le \|\hat{\beta}^L_\lambda\|_1$ and $\|Y - X\beta\|_2^2 < \|Y - X\hat{\beta}^L_\lambda\|_2^2$ then $Q_\lambda(\beta) < Q_\lambda(\hat{\beta}^L_\lambda)$, contradicting minimality of $\hat{\beta}^L_\lambda$. Similarly, $\hat{\beta}^R_\lambda$ must minimise

$$\|Y - X\beta\|_2^2 \text{ subj to } \|\beta\|_2 \le \|\hat{\beta}^R_\lambda\|_2$$





The contours of $\beta \longmapsto \|Y - X\beta\|_2^2$ are ellipsoids centred at $\hat{\beta}^{OLS}$ and the Lasso and ridge solutions occur when these intersect the l_1-ball $\{\beta : \|\beta\|_1 \le \|\hat{\beta}^L_\lambda\|_1\}$ and the l_2-ball $\{\beta : \|\beta\|_2^2 \le \|\hat{\beta}^R_\lambda\|_2^2\}$ respectively. The fact that the l_1-ball has 'corners' means some components of $\hat{\beta}^L_\lambda$ are exactly zero.

2.2.1 Prediction error with no design assumptions

Suppose cols of X are centred and scaled to have l_2-norm $\sqrt{n}$. Assume normal linear model (after centring)

$$Y = X\beta^0 + \varepsilon - \bar{\varepsilon}\mathbf{1} \qquad \varepsilon \sim N_n(0, \sigma^2 I)$$

Theorem 9 Let $\hat{\beta}$ be the Lasso estimate when $\lambda = A\sigma\sqrt{\log p/n}$

With probability of at least $1 - 2p^{-(A^2/2-1)}$

$$\frac{1}{n}\|X\beta^0 - X\hat{\beta}\|_2^2 \le 4A\sigma\sqrt{\frac{\log p}{n}}\,\|\beta^0\|_1$$

Proof From def^n of $\hat{\beta}$

$$\frac{1}{2n}\|\underbrace{Y - X\hat{\beta}}_{X(\beta^0-\hat{\beta}) + (\varepsilon - \bar{\varepsilon}\mathbf{1})}\|_2^2 + \lambda\|\hat{\beta}\|_1 \le \frac{1}{2n}\|\underbrace{Y - X\beta^0}_{\varepsilon - \bar{\varepsilon}\mathbf{1}}\|_2^2 + \lambda\|\beta^0\|_1$$

Rearranging

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \le \frac{1}{n}(\varepsilon - \underbrace{\bar{\varepsilon}\mathbf{1})^T X}_{\to 0 \text{ as cols } X \text{ centred}}(\hat{\beta} - \beta^0) + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1$$

$$|\varepsilon^T X(\hat{\beta} - \beta^0)| = |(X^T\varepsilon)^T(\hat{\beta} - \beta^0)| \overset{\text{Holder}}{\le} \|X^T\varepsilon\|_\infty \|\hat{\beta} - \beta^0\|_1$$

[If $v \in \mathbb{R}^p$, $\|v\|_\infty = \max_{j=1\dots p}|v_j|$, Hölder's inequality: if $u, v \in \mathbb{R}^p$ $|u^T v| \le \|u\|_1\|v\|_\infty$]

Let $\Omega = \{\|X^T\varepsilon\|_\infty/n \le \lambda\}$. We will show later (Lemma 13) that $\mathbb{P}(\Omega) \ge 1 - 2p^{-(A^2/2-1)}$. Working on Ω

$$\frac{1}{2n}\|X(\beta^0 - \hat{\beta})\|_2^2 \le \lambda\|\hat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1 - \lambda\|\hat{\beta}\|_1 \le 2\lambda\|\beta^0\|_1$$

by Δ inequality. □

2.2.2 Basic concentration inequalities

Consider the event Ω

$$\mathbb{P}(\|X^T\varepsilon\|_\infty/n > \lambda) = \mathbb{P}\left(\bigcup_{j=1}^p \{|X_j^T\varepsilon|/n > \lambda\}\right)$$

$$\overset{\text{union bd}}{\le} \sum_{j=1}^p \mathbb{P}(|X_j^T\varepsilon|/n > \lambda)$$

Now $X_j^T\varepsilon/n \sim N(0, \sigma^2/n)$.

Markov's inequality: Given a non-negative r.v. W, $t\mathbb{1}_{\{W \ge t\}} \le W$

$$\mathbb{P}(W \geq t) \leq \frac{1}{t}\mathbb{E}W \quad (t>0)$$

Given any strictly increasing function $\varphi: \mathbb{R} \to [0,\infty)$ and any r.v. W

$$\mathbb{P}(W \geq t) = \mathbb{P}(\varphi(W) \geq \varphi(t)) \leq \frac{\mathbb{E}\varphi(W)}{\varphi(t)} \quad (\varphi(t) > 0)$$

Apply this with $\varphi(t) = e^{\alpha t}$ yields the Chernoff bd

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha} e^{-\alpha t} \underbrace{\mathbb{E}e^{\alpha W}}_{\text{mgf } W \text{ (moment generating f}^n\text{)}}.$$

When $W \sim N(0,\sigma^2)$, $\mathbb{E}e^{\alpha W} = e^{\alpha^2\sigma^2/2}$, so

$$\mathbb{P}(W \geq t) \leq \inf_{\alpha} \exp\underbrace{\left(\frac{\alpha^2\sigma^2}{2} - \alpha t\right)}_{\frac{\sigma^2}{2}\left\{\left(\alpha - \frac{t}{\sigma^2}\right) + \text{const}\right\}}$$

$$= e^{-t^2/(2\sigma^2)}.$$

<u>Def^n 3</u> A r.v. W with $\mu = \mathbb{E}W$ is <u>sub-Gaussian</u> if $\exists\, \sigma > 0$ s.t.

$$\mathbb{E}e^{\alpha(W-\mu)} \leq e^{\alpha^2\sigma^2/2}.$$

We say that W is sub-G with parameter σ.

<u>Prop 10</u> If W is sub-G with parameter σ, then $\mathbb{P}(W - \mathbb{E}W \geq t) \leq e^{-t^2/(2\sigma^2)}$

Bounded r.v.-s are sub-G.

<u>Lemma 11</u> (Hoeffding's lemma) If W has mean zero and takes values in $[a,b]$ then W is a sub-G with param $\frac{b-a}{2}$.