## 1. Lasso with Gaussian design

Consider a regression problem where one observes $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$ in the model $y = X\beta^\star + \varepsilon$ where $X$ has iid $N(0,1)$ entries and $\varepsilon$ is a deterministic error vector with

$$\|\varepsilon\|^2 = n.$$

The vector $\beta^\star \in \mathbb{R}^p$ is unknown and $k$-sparse, in the sense that it has $k$ nonzero entries. We consider the Lasso estimator

$$\hat\beta = \mathrm{argmin}_{b\in\mathbb{R}^p} \|Xb - y\|^2 + \lambda\sqrt{n}\|b\|_1$$

where $\|\cdot\|_1$ is the $\ell_1$-norm and $\|\cdot\|$ the Euclidean norm.

You may use without proof the following properties of standard normal vectors $z \sim N(0, I_p)$: the $\ell_\infty$ norm defined as $\|z\|_\infty = \max_{j=1,\dots,p} |z_j|$ satisfies

$$\mathbb{E}[\|z\|_\infty] \le \sqrt{2\log(2p)},$$
$$\mathbb{P}(\|z\|_\infty > \sqrt{2\log p}) \le 1/\sqrt{\pi \log p}.$$

We assume that the sparsity $k$ of $\beta^\star$ times $\log p$ is small compared to $n$ in the sense that

$$\text{(A)} \qquad \sqrt{n-1} - \sqrt{n}/2 \ge \sqrt{2\log p} + 4\sqrt{2k\log(2p)}.$$

**Q1**. Given that $\varepsilon$ is deterministic with $\|\varepsilon\|^2 = n$ and $X$ has iid $N(0,1)$ entries, prove $z = n^{-1/2}X^T\varepsilon$ has distribution $N(0, I_p)$.

**Q2**. Noting that the objective function at $\hat\beta$ is smaller than the objective function at $\beta^\star$, show that

$$\|X(\hat\beta - \beta^\star)\|^2 \le 2\varepsilon^T X(\hat\beta - \beta^\star) + \sqrt{n}\lambda(\|\beta^\star\|_1 - \|\hat\beta\|_1)$$

where $\|\cdot\|$ is the Euclidean norm.

**Q3**. Prove that in an event of probability at least $1 - 1/\sqrt{\pi\log p}$, for $\lambda = 4\sqrt{2\log p}$ we have

$$n^{-1/2}\|X(\hat\beta - \beta^\star)\|^2 \le (\lambda/2)\|\hat\beta - \beta^\star\|_1 + \lambda(\|\beta^\star\|_1 - \|\hat\beta\|_1).$$

From now on, we use the tuning parameter value $\lambda = 4\sqrt{2\log p}$.

**Q4**. In the event of the previous question, show that the vector $\hat\beta - \beta^\star$ belongs to the set $K$ defined as $K = \{u \in \mathbb{R}^p : \sum_{j\in S^c} |u_j| \le 3\sum_{j\in S} |u_j|\}$ where $S = \{j = 1, \dots, p : \beta^\star_j \ne 0\}$ is the support of $\beta^\star$.

**Q5**. Prove that $\|u\|_1 \le 4\sqrt{k}\|u\|$ for any $u \in K$.

**Q6**. In the event of the three previous questions, show that

$$n^{-1/2}\|X(\hat\beta - \beta^\star)\|^2 \le (3\lambda/2)4\sqrt{k}\|\hat\beta - \beta^\star\|.$$

**Q7**. We would like to bound from above $\|\hat\beta - \beta^\star\|_2$ by $\|X(\hat\beta - \beta^\star)\|/\sqrt{n}$ up to a multiplicative constant. To this end, we appeal to the following version of Gordon's theorem that you can use without proof: With probability at least $1 - 1/p$,

$$\inf_{u\in T} \|Xu\| \ge \sqrt{n-1} - w(T) - \sqrt{2\log p}$$

where $T = K \cap \{u \in \mathbb{R}^p : \|u\| = 1\}$ and $w(T) = \mathbb{E}[\sup_{u\in T} z^T u]$ where $z \sim N(0, I_p)$. Using the bound provided on $\mathbb{E}[\|z\|_\infty]$, prove that $w(T) \le 4\sqrt{k}\sqrt{2\log(2p)}$.

**Q8**. In the event of Gordon's theorem from the previous question and using assumption (A), show that $\forall v \in K : \|Xv\| \ge \sqrt{n}\|v\|/2$.

**Q9**. Given two events $E_1$ and $E_2$ with $\mathbb{P}(E_1) \ge 1 - 1/p$ and $\mathbb{P}(E_2) \ge 1 - 1/\sqrt{\pi\log p}$, provide a lower bound on the probability of the intersection $\mathbb{P}(E_1 \cap E_2)$.

**Q10**. Explain how to combine the previous questions to prove that under condition (A), the Lasso satisfies the error bounds

$$\|X(\hat\beta - \beta^\star)\| \le 2(3\lambda/2)4\sqrt{k} \quad \text{and} \quad \|\hat\beta - \beta^\star\| \le 4(3\lambda/2)4\sqrt{k/n}$$

in an event of probability close to 1 when $p$ is large.

**Q11**. How does this bound on $\|X(\hat\beta - \beta^\star)\|$ compare with $\|X(\hat\alpha - \beta^\star)\|$ for $\hat\alpha = (X^TX)^\dagger X^T y$ if $p > n$ and $k\log p \lll n$?

**Q12**. (Bonus). Prove $\mathbb{P}(Z > t) \le e^{-t^2/2}/(t\sqrt{2\pi})$ for $Z \sim N(0,1)$ and $\mathbb{P}(\|z\|_\infty > \sqrt{2\log p}) \le 1/\sqrt{\pi\log p}$ for $z \sim N(0, I_p)$.

**Q13**. (Bonus 2). Prove $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ for $Z \sim N(0,1)$. For $z \sim N(0, I_p)$, deduce $\mathbb{E}[e^{t\|z\|_\infty}] \le 2pe^{t^2/2}$ and taking $t = \sqrt{2\log(2p)}$ prove that $\mathbb{E}[\|z\|_\infty] \le \sqrt{2\log(2p)}$.