MATH 547: HOMEWORK 5 (BONUS)
DUE ON: MONDAY, DECEMBER 9, 9AM.

Reminder: I will drop lowest homework score, so completing this homework is optional.

Problem 1: sparse vectors, 30 points: Given $x \in K \subset \mathbb{R}^p$, let

$$D(K,x) := \{t(z-x) : \; z \in K, \; t \geq 0\}$$

be the descent cone of K at the point x. Moreover, let $S(K,x) = D(K,x) \cap S^{p-1}$, where S^{p-1} is the unit sphere in $\mathbb{R}^p$. The goal of this exercise is to obtain a sharper, compared to the one we proved in class, upper bound on the Gaussian mean width of $S(K,x)$ where K is the unit ball in $\|\cdot\|_1$ norm and **$\mathbf{x}$ is s-sparse, meaning that it has only s non-zero coordinates, and** $\|\mathbf{x}\|_1 = 1$ (the latter condition means that x is on the boundary of K).

(a) Let $S_{p,s} = \{x \in \mathbb{R}^p : \; \|x\|_0 \leq s, \; \|x\|_2 \leq 1\}$, where $\|x\|_0$ denotes the number of non-zero coordinates of x. Show that its Gaussian mean width satisfies $w^2(S_{p,s}) \leq Cs\log(2p/s)$ for some absolute constant $C > 0$.
One way to do it is to use the union bound combined with following *Gaussian concentration inequality*: for any $k \geq 1$ and $g \sim N(0, I_k)$,

$$\Pr(\|g\| \geq \mathbb{E}\|g\|_2 + t) \leq \exp(-t^2/2).$$

You will need the fact that (show it!) $\mathbb{E}\|g\|_2 \leq \sqrt{k}$ and the union bound over $k = 1, \ldots, s$. Finally, use the fact that for any nonnegative random variable Z, $\mathbb{E}Z = \int_0^\infty \Pr(Z \geq t)dt$.

(b) The next step is to relate $w(S(K,x))$ to $w(S_{p,s})$. Show that $S(K,x) \subset 3\,\mathrm{co}(S_{p,s})$, where $\mathrm{co}(\cdot)$ stands for the convex hull of the set. The proof of this fact proceeds in several steps (skip the ones you get stuck on):

(1) Recall that x is s-sparse and let J be the set of non-zero coordinates of x. In class, we showed that for any z in K, $\sum_{j \notin J} |z_j - \underbrace{x_j}_{=0}| \leq \sum_{j \in J} |z_j - x_j|$, so that the vector $z - x$ has its "dominant" coordinates in set J (you can use this fact without the proof).

(2) For the vector $u = \frac{z-x}{\|z-x\|_2}$ for some arbitrary $z \in K, \; z \neq x$, let $|u| = (|u_1|, \ldots, |u_p|)$. For any set of indices I, define u_I to be the vector $(u_i, \; i \in I)$. Next, let J_1 be the set of s largest coordinates of the vector $|u|_{J^c}$, where J^c is the complement of J, J_2 - set of s largest coordinates of $|u|_{(J \cup J_1)^c}$, etc. (first s largest, next s largest, until nothing is left). Finally, show that

$$\sum_{k \geq 2} \|u_{J_k}\|_2 \leq \|u_J\|_2 \leq 1.$$

(3) Deduce that $\sum_{k \geq 2} u_{J_k} \in \mathrm{co}(S_{p,s})$ and conclude that $u \in 3\,\mathrm{co}(S_{p,s})$.

(c) Finally, combine the previous bounds to get an estimate for $w^2(S(K,x))$.

Problem 2: the LASSO, 30 points:

Assume that we observe n noisy linear measurements of an s-sparse vector $\lambda_0 \in \mathbb{R}^d$,

$$Y = \mathbf{X}\lambda_0 + \varepsilon$$

where $Y \in \mathbb{R}^d$, $\mathbf{X}$ is a $n \times d$ matrix and $\varepsilon \in \mathbb{R}^n$ is a noise vector (for example, it is often modeled by a sequence of i.i.d. $N(0,1)$ random variables). In this case, a popular approach to estimating λ_0 is via solving the problem

$$\widehat{\lambda} \in \operatorname*{argmin}_{\lambda \in \mathbb{R}^d} \left[\frac{1}{n} \|X\lambda - Y\|_2^2 + \tau \|\theta\|_1 \right]$$

where $\tau > 0$ is a "penalty coefficient" and $\|\cdot\|_1$ is the ℓ_1 norm. This problem is known as "LASSO" (Least Absolute Shrinkage and Selection Operator).

(1) Assume that $n = d$ and $\mathbf{X}$ is the identity matrix. Find the explicit form of the solution $\widehat{\lambda}$ (this special case explains the term "shrinkage").

(2) Suggest a simple example of the design matrix $\mathbf{X}$ when the Lasso estimator $\widehat{\lambda}$ is not unique.

(3) Prove that any two solutions $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ satisfy

$$\mathbf{X}\widehat{\lambda}_1 = \mathbf{X}\widehat{\lambda}_2 \text{ and } \left\|\widehat{\lambda}_1\right\|_1 = \left\|\widehat{\lambda}_2\right\|_1$$

(Recall that the first of these properties is also valid for any solution to the usual least-squares problem, as we discussed in class)

Thank you for taking the course! I hope that you enjoyed it.