Homework 2
Statistical Machine Learning
10/36-702
Due Friday Feb 10

1. Verify the subdifferentials given in the Convexity chapter, equations (1.128), (1.129) and (1.131).

   Hint for 1.128: By Hölder's inequality, for any $1 < q < \infty$, any $u \in B^{q'}(1)$, and any $z \in \mathbb{R}^p$,
   $$\|z\|_q \geq u^T z$$
   where $1/q + 1/q' = 1$ (why?). But, for any $u \notin B^{q'}(1)$, there exists $z \in \mathbb{R}^p$ such that $\|z\|_q < u^T z$ (consider, for example, $z_i = \text{sign}(u_i)|u_i|^{q'/q}$).

   Hint for 1.131: $\|z\|_\infty \|u\|_1 \geq u^T z$ for all u, z (why?).

2. Consider Example 1.132 in the Optimization chapter. The KKT conditions are given on page 20. Derive these KKT conditions. (Note: there may be a factor of $1/n$ missing in 1.139).

3. Let $Z_i \sim N(\mu_i, 1)$ for $i = 1, \ldots, N$. Suppose we divide the N observations into G groups:
   $$\begin{aligned} G_1 &= \{\mu_1, \ldots, \mu_{n_1}\} \\ G_2 &= \{\mu_{n_1+1}, \ldots, \mu_{n_2}\} \\ \vdots &= \vdots \end{aligned}$$
   Thus there are n_j observations in the j^{th} group and $\sum_{j=1}^G n_j = N$. We want to estimate $\mu = (\mu_1, \ldots, \mu_N)$. Let $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_N)$ minimize
   $$\sum_{j=1}^N (Z_j - \mu_j)^2 + \lambda ||\mu||_1 + \gamma \sum_{j=1}^G ||\mu_{\mathcal{G}_j}||_2$$
   where
   $$||\mu_{\mathcal{G}_j}||_2^2 = \sum_{i \in \mathcal{G}_j} \mu_i^2$$
   and $||\mu||_1 = \sum_{j=1}^N |\mu_j|$.

4. Let $X_1, \ldots, X_n \sim P$ where P has density p on $[0,1]$. Let $\psi_1, \psi_2, \ldots,$ be an orthonormal basis for $L_2[0,1]$. Thus, $\int_0^1 \psi_j^2(x)dx = 1$ and $\int_0^1 \psi_j(x)\psi_k(x)dx = 0$ for $j \neq k$. You may assume that $\sup_j \sup_x |\psi_j(x)| \leq C < \infty$. Assume that $\int p^2(x)dx < \infty$. Thus we can write
   $$p(x) = \sum_{j=1}^\infty \theta_j \, \psi_j(x)$$

where $\theta_j = \int_0^1 \psi_j(x)p(x)dx$. Assume that

$$\sum_{j=1}^{\infty} \theta_j^2 j^{2p} \leq C$$

for some $p > 1/2$. Define an estimator

$$\widehat{p}(x) = \sum_{j=1}^{J_n} \widehat{\theta}_j \, \psi_j(x)$$

where

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \psi_j(X_i).$$

(a) Show that the variance of $\widehat{p}(x)$ is bounded by $C_1 J_n/n$ for some constant C_1.

(b) Show that

$$\int (p(x) - \overline{p}(x))^2 dx \leq \frac{C_2}{J_n^{2p}}$$

for some C_2, where $\overline{p}(x) = \mathbb{E}(\widehat{p}(x))$.

(c) Show that

$$\mathbb{E} \int_0^1 (\widehat{p}(x) - p(x))^2 dx \leq \frac{C_2}{J_n^{2p}} + \frac{C_1 J_n}{n}.$$

(d) Confirm that the upper bound on the risk is minimized by choosing $J_n \asymp n^{\frac{1}{2p+1}}$. With this choice of J_n we see that

$$\mathbb{E} \int_0^1 (\widehat{p}(x) - p(x))^2 dx = O\left(n^{-\frac{2p}{2p+1}}\right).$$