CMSE 820 Homework 3. Due 29 September, 2020.

I. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function and $x_0 \in \mathbb{R}^n$ be a fixed point.

(1) Let $\alpha > 0$ be a positive real number. Prove that

$$\partial(\alpha f)(x_0) = \alpha \partial f(x_0)$$

where $\alpha\partial f(x_0)$ is defined as $\alpha\partial f(x_0) := \{\alpha s : s \in \partial f(x_0)\}$.

Solution. Let $w \in \alpha\partial f(x_0)$. Therefore, $w \in \{\alpha s : s \in \partial f(x_0)\}$. Without loss of generality, let s_0 be an element of $\partial f(x_0)$ such that $w = \alpha s_0$. From the definition of a subgradient, for any $y \in \mathcal{D}(f)$,

$$f(y) \geq f(x_0) + s_0^T(y - x_0).$$

Multiplying the expression by α, we get

$$\alpha f(y) \geq \alpha f(x_0) + (\alpha s_0)^T(y - x_0).$$

Let $\tilde{f}(x) = \alpha f(x)$. Thus,

$$\tilde{f}(y) \geq \tilde{f}(x) + (\alpha s_0)^T(y - x_0).$$

Therefore $\alpha s_0 \in \partial\tilde{f}(x_0)$, which means $\alpha s_0 = w \in \partial\tilde{f}(x_0) = \partial(\alpha f)(x_0)$ and $\alpha\partial f(x_0) \subseteq \partial(\alpha f)(x_0)$.

Now let $w \in \partial(\alpha f)(x_0)$. From the definition of a subgradient, $\forall y \in \mathcal{D}(f)$,

$$\alpha f(y) \geq \alpha f(x_0) + w^T(y - x_0).$$

Since $\alpha > 0$,

$$f(y) \geq f(x) + \left(\frac{1}{\alpha}w\right)^T (y - x_0).$$

Letting $\tilde{w} = \frac{1}{\alpha}w$,

$$f(y) \geq f(x_0) + \tilde{w}^T(y - x_0).$$

Therefore, $\tilde{w} \in \partial f(x_0)$. It follows then that $\alpha\tilde{w} = w \in \{\alpha s : s \in \partial f(x_0)\} = \alpha\partial f(x_0)$. Thus, $\partial(\alpha f)(x_0) \subseteq \alpha\partial f(x_0)$. ■

(2) Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix and $b \in \mathbb{R}^n$ be a vector. Define a function $g(x)$ by $g(x) := f(Ax + b)$. Prove that

$$\partial g(x_0) = A^T \partial f(Ax_0 + b)$$

where the right hand side is defined as $A^T\partial f(Ax_0 + b) := \{A^T s : s \in \partial f(Ax_0 + b)\}$.

Solution. Let $w \in \partial g(x_0)$. It follows that $\forall y \in \mathcal{D}(g)$,

$$\begin{aligned}
g(y) &\geq g(x_0) + w^T(y - x_0)\\
f(Ay + b) &\geq f(Ax_0 + b) + w^T((Ay + b) - (Ax_0 + b))\\
f(Ay + b) &\geq f(Ax_0 + b) + w^T A(y - x_0)\\
f(Ay + b) &\geq f(Ax_0 + b) + (A^T w)^T(y - x_0).
\end{aligned}$$

Therefore, $A^T w \in \partial f(Ax_0 + b) \Rightarrow w \in \{A^T s : s \in \partial f(Ax_0 + b)\}$. Thus, $\partial g(x_0) \subseteq \{A^T s : s \in \partial f(Ax_0 + b)\} = A^T \partial f(Ax_0 + b)$.

Suppose now $w \in A^T \partial f(Ax_0 + b)$. This implies that $A^T w \in \partial f(Ax_0 + b)$, and thus $\forall y \in \mathcal{D}(f)$,

$$
\begin{aligned}
f(Ay + b) &\geq f(Ax_0 + b) + (A^T w)^T (y - x_0) \\
f(Ay + b) &\geq f(Ax_0 + b) + w^T (Ay - Ax_0) \\
f(Ay + b) &\geq f(Ax_0 + b) + w^T ((Ay + b) - (Ax_0 + b)).
\end{aligned}
$$

Since A is invertible, it follows that the column space of A is equal to $\mathbb{R}^n$. Therefore, $\mathcal{D}(f) = \mathcal{D}(g)$. From this and the definition of $g(x)$,

$$g(y) \geq g(x_0) + w^T (y - x_0).$$

$w \in \partial g(x_0) \Rightarrow A^T \partial f(Ax_0 + b) \subseteq \partial g(x_0)$. ■

II. Let $f_1(x) = (x+1)^2$ and $f_2(x) = (x-1)^2, x \in \mathbb{R}$. Define $f(x) := \max\{f_1(x), f_2(x)\}$. Compute $\partial f(x)$.

Solution. Both $f_1(x)$ and $f_2(x)$ are smooth differentiable functions. We can therefore define $\bigtriangledown f_1(x) = 2(x+1)$ and $\bigtriangledown f_2(x) = 2(x-1)$. Additionally, since $\bigtriangledown(\bigtriangledown f_1(x)) = \bigtriangledown(\bigtriangledown f_2(x)) = 2 > 0$, we know that $f_1(x)$ and $f_2(x)$ are convex functions.

Suppose for a given x_0 that $f_1(x_0) < f_2(x_0)$. Therefore,

$$(x_0 + 1)^2 < (x_0 - 1)^2 \Rightarrow x_0 < 0.$$

When $x_0 < 0$, $f(x) = f_2(x)$. Since $f_2(x)$ is differentiable, it follows that $\partial f(x_0) = \bigtriangledown f_2(x_0) = 2(x_0 - 1)$ for $x_0 < 0$.

Suppose now for a given x_0 that $f_1(x_0) > f_2(x_0) \Rightarrow (x_0+1)^2 > (x_0-1)^2 \Rightarrow x_0 > 0$. Therefore, for $x_0 > 0$, $f(x) = f_1(x)$, and by the same argument as before, $\partial f(x_0) = \bigtriangledown f_1(x_0) = 2(x_0 + 1)$ for $x_0 > 0$.

Finally, suppose $x_0 = 0$. It follows $\forall y \in \mathcal{D}(f)$ that

$$
\begin{aligned}
f(y) &\geq f(x_0) + s(y - x_0) \\
\max\{f_1(y), f_2(y)\} &\geq \max\{f_1(0), f_2(0)\} + sy \\
\max\{f_1(y), f_2(y)\} &\geq 1 + sy.
\end{aligned}
$$

The subdifferential will be the collection of slopes s such that the line $1+sy$ lies below $\max\{f_1(x), f_2(x)\}$. For this to hold true, the slope s must be smaller than the slope of either $f_1(x)$ or $f_2(x)$. This is because $f(0) = 1$, therefore the term sy dictates the allowable slopes. At $x_0 = 0$, $\bigtriangledown f_1(0) = 2$ and $\bigtriangledown f_2(0) = -2$. Therefore, $\partial f(x_0) = [-2, 2]$ for $x_0 = 0$ in order for $\max\{f_1(y), f_2(y)\} \geq 1 + sy$.

We can thus write

$$\partial f(x_0) = \begin{cases} 2(x_0+1) & , x_0 > 0 \\ [-2, 2] & , x_0 = 0 \\ 2(x_0 - 1) & , x_0 < 0 \end{cases}. \blacksquare$$

III. For a vector $x \in \mathbb{R}^n$, define the notation

$$\|x\|_0 := \#\{j : x_j \neq 0\},$$

that is, $\|x\|_0$ is the number of nonzero components of x. Let $p \in \mathbb{R}$ be a real number and $p > 0$. Recall the following notation:

$$\|x\|_p^p := \sum_{j=1}^{n} |x_j|^p.$$

(1) Prove that $\forall x \in \mathbb{R}^n$,

$$\lim_{p \to 0^+} \|x\|_p^p = \|x\|_0.$$

Solution. Let $\|x\|_p^p = |x_1|^p + \cdots + |x_n|^p$. As $p \to 0^+$, this expression approaches $|x_1|^0 + \cdots + |x_n|^0$. For $x_j = 0$, we can define $0^0 = 0$ since $0^p = 0$ for any $p > 0$, so in the limit $p \to 0^+$ our definition holds. For all $x_j \neq 0$, $|x_j|^0 = 1$. Suppose there are k nonzero values of x. Then in the limit, $\|x\|_p^p = \sum_{j=1}^{k} 1 + \sum_{k+1}^{n} 0 = k$. Therefore, as $p \to 0^+$, $\|x\|_p^p$ approaches k, the number of nonzero elements of x, which we have defined as $\|x\|_0$.

Alternatively, we can approximate $|x|^p = \exp(p \log |x|) \approx 1 + p \log |x|$ for small p. Therefore,

$$\|x\|_p^p \approx \sum_{j=1}^{n} 1 + p \log |x_j| = n + np \sum_{j=1}^{n} \log |x_j| = n + np \log \left(\prod_{j=1}^{n} |x_j| \right).$$

In this expression, if any $x_j = 0$, then $\prod_{j=1}^{n} |x_j| = 0$ and $\log(0)$ is undefined. Therefore, this expression requires $x_j \neq 0$ for all j. But, as $p \to 0^+$, then $\|x\|_p^p \to n$. Since no $x_j = 0$, then $\|x\|_0 = n$. $\blacksquare$

(2) Prove that $\|x\|_p$ is a convex function of x if and only if $p \geq 1$.

Solution. We write $\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$. For a function f to be convex, we need $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$ for $\lambda \in [0,1]$ and $x, y \in \mathcal{D}(f)$. For $p \geq 1$, we know from the definition of a norm that $\| \bullet \|_p$ satisfies the triangle inequality. Namely, for two vectors $x, y \in \mathbb{R}^n$,

$$\|x+y\|_p \leq \|x\|_p + \|y\|_p \Rightarrow \|\lambda x + (1-\lambda)y\|_p \leq \|\lambda x\|_p + \|(1-\lambda)y\|_p = \lambda \|x\|_p + (1-\lambda)\|y\|_p.$$

Therefore, $f_p(x) = \|x\|_p$ is convex for $p \geq 1$.

Suppose $p < 1$. Let $x = (1, 0, 0, \ldots, 0)^T$ and $y = (0, 1, 0, \ldots, 0)^T$ be vectors in $\mathbb{R}^n$ each with a single non-zero element (which we have set to 1). We can explicitly calculate $\|x+y\|_p$,

$$\|x+y\|_p = (1^p + 1^p)^{1/p} = (2)^{1/p},$$

since $1^p = 1$ for any p. Additionally, we know that $\|x\|_p = \|y\|_p = (1^p)^{1/p} = 1$. Therefore, $\|x\|_p + \|y\|_p = 2$. Since $p < 1$, it follows that $1/p > 1$. Thus, $2^{1/p} > 2$. Since $\|x+y\|_p \not\leq \|x\|_p + \|y\|_p$, $\| \bullet \|_p$ does not satisfy the triangle inequality for $p < 1$. It follows then that f_p is not convex for $p < 1$. $\blacksquare$

IV. Suppose $X \in \mathbb{R}^{p\times n}$ and $XX^T = I_p$, where I_p is the $p \times p$ identity matrix. Define β_0 by

$$\beta_0 := \arg\min_\beta \|y - X^T\beta\|^2 + \lambda\|\beta\|_2^2 + \tau\|\beta\|_1,$$

where $\lambda > 0$ and $\tau > 0$ are two positive constants. Find an expression of β_0 in terms of the soft thresholding function.

Solution. This solution follows the derivation we did in class for β^{lasso}.

$$\begin{aligned}
\beta_0 &= \arg\min_\beta \|y - X^T\beta\|^2 + \lambda\|\beta\|_2^2 + \tau\|\beta\|_1 \\
&= \arg\min_\beta (y - X^T\beta)^T(y - X^T\beta) + \lambda\beta^T\beta + \tau\|\beta\|_1 \\
&= \arg\min_\beta y^Ty - 2(Xy)^T\beta + \beta^T\beta + \lambda\beta^T\beta + \tau\|\beta\|_1 \\
&= \arg\min_\beta y^Ty - 2(\beta^{ls})^T\beta + (1+\lambda)\beta^T\beta + \tau\|\beta\|_1 \\
&= \arg\min_\beta \left(\sum_{j=1}^p (1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|\right) + y^Ty.
\end{aligned}$$

In the above, we substituted $\beta^{ls} = Xy$ for the least squares solution and converted all of the terms pertaining to β into a summation. Since y^Ty effects the minimum value, but not the minimization, it follows that we need only minimize the sum term. We can do this by minimizing $(1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|$ for $j = 1, \ldots, p$. For a fixed j,

$$\begin{aligned}
\partial((1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|)(\beta_y) &= \partial((1+\lambda)\beta_j^2)(\beta_y) - \partial(2\beta_j^{ls}\beta_j)(\beta_y) + \partial(\tau|\beta_j|)(\beta_y) \\
&= 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \begin{cases} \tau & ,\beta_y > 0 \\ \tau[-1,1] & ,\beta_y = 0 \\ -\tau & ,\beta_y < 0 \end{cases} \\
&= \begin{cases} 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \tau & ,\beta_y > 0 \\ 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \tau[-1,1] & ,\beta_y = 0 \\ 2(1+\lambda)\beta_y - 2\beta_j^{ls} - \tau & ,\beta_y < 0 \end{cases}.
\end{aligned}$$

We can minimize the expression for β_0 by satisfying the minimization condition, $0 \in \partial((1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|)(\beta_y)$.

For $\beta_y > 0$,

$$\begin{aligned}
&0 \in \partial((1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|)(\beta_y) \\
\Leftrightarrow\ &0 = 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \tau \\
\Leftrightarrow\ &\beta_y = \frac{1}{1+\lambda}(\beta_j^{ls} - \tau/2) \\
\Leftrightarrow\ &\beta_j^{ls} > \tau/2.
\end{aligned}$$

For $\beta_y < 0$,

$$\begin{aligned}
&0 \in \partial((1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|)(\beta_y) \\
\Leftrightarrow\ &0 = 2(1+\lambda)\beta_y - 2\beta_j^{ls} - \tau \\
\Leftrightarrow\ &\beta_y = \frac{1}{1+\lambda}(\beta_j^{ls} + \tau/2) \\
\Leftrightarrow\ &\beta_j^{ls} < -\tau/2.
\end{aligned}$$

For $\beta_y = 0$,

$$\begin{aligned}
&0 \in \partial((1+\lambda)\beta_j^2 - 2\beta_j^{ls}\beta_j + \tau|\beta_j|)(\beta_y) \\
\Leftrightarrow\ &0 \in [2(1+\lambda)\beta_y - 2\beta_j^{ls} - \tau, 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \tau] \\
\Leftrightarrow\ &[2(1+\lambda)\beta_y - 2\beta_j^{ls} - \tau \leq 0,\ \ 2(1+\lambda)\beta_y - 2\beta_j^{ls} + \tau \geq 0.
\end{aligned}$$

The final condition yields 0 for $-\tau/2 \leq \beta_j^{ls} \leq \tau/2$. Thus, the jth component of β_0 can be written as

$$\begin{aligned}
\beta_{j0} &= \begin{cases} \frac{\beta_j^{ls}}{1+\lambda} + \frac{\tau}{2(1+\lambda)} & , \beta_j^{ls} < -\tau/2 \\ \frac{\beta_j^{ls}}{1+\lambda} - \frac{\tau}{2(1+\lambda)} & , \beta_j^{ls} > \tau/2 \\ 0 & , -\tau/2 \leq \beta_j^{ls} \leq \tau/2 \end{cases} \\
&= \frac{(|\beta_j^{ls}| - \tau/2)_+}{1+\lambda}\operatorname{sgn}(\beta_j^{ls}) \\
&= \frac{1}{1+\lambda} S_{\tau/2}(\beta_j^{ls}). \blacksquare
\end{aligned}$$