CMSE 8: HW 3

Matt Piekenbrock

Problem 1

(1) Let $\alpha > 0$ be a positive real number. The goal is to show:

$$\partial(\alpha f)(x_0) = \{\alpha s : s \in \partial f(x_0)\}$$

Consider the left hand side. It may be rewritten as follows:

$$\begin{aligned}
\partial(\alpha f)(x_0) &= \left[\frac{\partial}{\partial x_0^{(1)}}(\alpha f), \frac{\partial}{\partial x_0^{(2)}}(\alpha f), \dots, \frac{\partial}{\partial x_0^{(n)}}(\alpha f)\right]^T \\
&= \left[\alpha\frac{\partial}{\partial x_0^{(1)}}(f), \alpha\frac{\partial}{\partial x_0^{(2)}}(f), \dots, \alpha\frac{\partial}{\partial x_0^{(n)}}(f)\right]^T \\
&= \alpha\left[\frac{\partial}{\partial x_0^{(1)}}(f), \frac{\partial}{\partial x_0^{(2)}}(f), \dots, \frac{\partial}{\partial x_0^{(n)}}(f)\right]^T \\
&= \alpha\frac{\partial}{\partial x_0}(f) = \alpha\partial f(x_0)
\end{aligned}$$

which yields the desired result, where $x_0^{(i)}$ represents the i^{th} component of $x_0 \in \mathbb{R}^n$.

(2) If a function $g(x)$ is differentiable at a point x_0, then the derivative of g at x_0 is determined uniquely and is represented by the $n \times n$ derivative matrix $D_g(x_0)$, whose columns give the vector partial derivatives with respect to each component. The gradient of g is characterized as follows:

$$\partial g(x) = \begin{bmatrix} \frac{\partial g}{\partial x_0}(x) \\ \frac{\partial g}{\partial x_1}(x) \\ \vdots \\ \frac{\partial g}{\partial x_n}(x) \end{bmatrix} = D_g(x)^T$$

Now if one has two differentiable functions f and g whose composite is $h(x) = g(f(x))$, by the definition of the chain rule, we have:

$$\partial h(x) = D_g(f(x))^T \begin{bmatrix} \frac{\partial f}{x_1}(x) \\ \frac{\partial f}{x_2}(x) \\ \vdots \\ \frac{\partial f}{x_n}(x) \end{bmatrix}$$

Replacing $A = D_g(f(x_0))$ above and noting that $s \in \partial f(Ax_0 + b)$ yields the desired result.

Problem 2

First, consider the component derivatives:

$$\partial f_1(x) = 2(x+1)$$

$$\partial f_2(x) = 2(x-1)$$

Observe that $f_1(x) > f_2(x)$ for all $x > 0$, and $f_2(x) > f_1(x)$ for all $x < 0$, yielding the characterization of $\partial f(x)$ for those subsets of the domain. The only point in $\mathbb{R}$ not described is the derivative at the point $x = 0$. Since f is not differentiable at $x = 0$, $\partial f(x)$ at some fixed point x is defined as the set of vectors s satisfying:

$$f(\hat{x}) \geq f(x) + s^T(\hat{x} - x) \quad \forall\, \hat{x} \in D(f)$$

Note that in this case, $f : \mathbb{R} \to \mathbb{R}$ (presumeably). As a result, if one fixes $x = 0$ then for this inequality to be true we have:

$$f(\hat{x}) \geq \max(1, 1) + s^T(\hat{x} - 0) \quad \forall\, \hat{x} \in D(f)$$

$$\max\{(\hat{x}+1)^2, (\hat{x}-1)^2\} \geq 1 + s\hat{x} \quad \forall\, \hat{x} \in D(f) \qquad \text{since } s, x \in \mathbb{R}$$

$$\implies \begin{cases} (\hat{x}+1)^2 - 1 \geq s\hat{x} & \text{if } \hat{x} > 0 \\ (\hat{x}-1)^2 - 1 \geq s\hat{x} & \text{if } \hat{x} < 0 \end{cases}$$

$$\implies \begin{cases} \hat{x}^2 + 2\hat{x} \geq s\hat{x} & \text{if } \hat{x} > 0 \\ \hat{x}^2 - 2\hat{x} \geq s\hat{x} & \text{if } \hat{x} < 0 \end{cases}$$

$$\implies s \leq \begin{cases} \hat{x} - 2 & \text{if } \hat{x} < 0 \\ \hat{x} + 2 & \text{if } \hat{x} > 0 \end{cases}$$

$$\implies s \in \begin{cases} (-\infty, 2) & \text{if } \hat{x} < 0 \\ (2, +\infty) & \text{if } \hat{x} > 0 \end{cases}$$

I conclude that the characterization of the subdifferential of f is given by:

$$\partial f(x) = \begin{cases} 1 + s\hat{x} & \text{if } x = 0 \\ 2(x-1) & \text{if } x < 0 \\ 2(x+1) & \text{if } x > 0 \end{cases}$$

where s is parameterized by some choice of $\hat{x}$, as described above.

Problem 3

(1) Showing the (one-sided) limit given below

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

holds is equivalent to proving that for any number $\epsilon > 0$, there is a corresponding number $\delta > 0$ such that for all x we have:

$$0 < x < 0 + \delta \implies \left| \|x\|_p^p - \|x\|_0 \right| < \epsilon$$

Rewriting this statement, this is equivalent to showing that the one sided-limit holds:

$$\lim_{p \to 0+} \sum_{j=1}^{n} |x_j|^p = \|x\|_0$$

Consider the case where $x_j \in (0, 1]$ for each $j \in [1, n]$. Observe that the power function $|\cdot|^p : (0, 1] \to \mathbb{R}$ is monotonically non-decreasing for every choice of p, and that for any choice of powers $0 < p < \hat{p} < 1$, the graph of the function of p in the interval $(0, 1]$ is completely above the graph of $\hat{p}$. As $p \to 0^+$, graph becomes a constant function, where for every input $x \in (0, 1]$ the function returns 1. The case is similar when each $x_j > 1$, however instead as $p \to 0^+$, each entry necessary converges to 1, as anything raised to the 0^{th} power must be 1.

(2) Recall that for a given function f to be considered a convex function, its domain $D(f)$ must be a convex set, and it must obey the inequality:

$$f(tx + (1-t)x') \leq tf(x) + (1-t)f(x')$$

To show that the function $\|x\|_p = \sum_{j=1}^{n} |x_j|^p$ obeys this only for $p \geq 1$, I first recall the properties of vector norms. By definition, if $\|\cdot\|$ is a *vector norm*, it satisfies the following three conditions [1] for some vectors $x, x' \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$:

$$\|x\| \geq 0, \text{ and } \|x\| = 0 \text{ only if } x = 0, \tag{1}$$

$$\|x + x'\| \leq \|x\| + \|x'\|, \tag{2}$$

$$\|\alpha x\| = |\alpha|\|x\| \tag{3}$$

For any two vectors $x, \hat{x} \in \mathbb{R}^n$, by the triangle inequality (second above), we have:

$$\|tx + (1-t)\hat{x}\|_p \leq \|tx\|_p + \|(1-t)\hat{x}\|_p \quad \forall\, t \in [0, 1]$$

and by the scaling property (three above), we have:

$$\|tx\|_p + \|(1-t)\hat{x}\|_p = t\|x\|_p + (1-t)\|\hat{x}\|_p \quad \forall\, t \in [0,1]$$

Giving the desired result that any vector norm is convex. To show that $\|x\|_p$ if and only if $p \geq 1$, one needs to show that the above three listed properties hold for $p \geq 1$, and that they don't hold for $p < 1$: but this is given by Minkowski's inequality, which states that for any two measurable functions $f, g : \mathbb{R}^n \to \mathbb{R}$, we have:

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p \quad 1 \leq p \leq \infty$$

The fact that this inequality is only obeyed when $p \geq 1$ indicates that only p-norms with $p \geq 1$ obey the vector norm conditions above, and thus only those such p-norms are convex.

(4) Define β_0 by:

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

Using the notational techniques used in the lecture, and treating the inner terms to optimize as $f(\beta)$ we have that the cost function f is to minimize:

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

At this point, notice that y^Ty is independent of β, and that it suffices to minimize the above for each j independently. Therefore, from now on, assume a fixed choice of j (such as 0). Taking the derivative of $f(\beta)$ with respect to some fixed β_0, we have:

$$\begin{aligned}
\partial f(\beta_0) &= \partial\left(\beta_j^2 - 2\beta_j^{ls}\beta_j + \lambda\beta_j^2 + \tau|\beta_j|\right)(\beta_0) \\
&= \partial(\beta_j^2)(\beta_0) - \partial(2\beta_j^{ls}\beta_j)(\beta_0) + \partial(\lambda\beta_j^2)(\beta_0) + \partial(\tau|\beta_j|)(\beta_0) \\
&= 4\lambda\beta_0 - 2\beta_j^{ls} + \begin{cases} \tau & \beta_0 > 0 \\ \tau[-1,1] & \beta_0 = 0 \\ -\tau & \beta_0 < 0 \end{cases}
\end{aligned}$$

Now, for $0 \in \partial f(\beta_0)$, we have the following two cases when $\beta_0 \neq 0$:

$$\beta_0 = \begin{cases} \frac{1}{2\lambda}\beta_j^{ls} + \frac{\tau}{4\lambda} & \text{if } \beta_0 > 0 \\ \frac{1}{2\lambda}\beta_j^{ls} - \frac{\tau}{4\lambda} & \text{if } \beta_0 < 0 \end{cases}$$

If $\beta_0 = 0$, then for $0 \in \partial f(\beta) \implies 0 \in [-2\beta_j^{ls} - \tau, -2\beta_j^{ls} + \tau]$, which implies that $\tau \geq -2\beta_j^{ls}$ and $\tau \geq 2\beta_j^{ls}$, leading to the final expression:

$$\beta_0 = \begin{cases} \frac{1}{2\lambda}\beta_j^{ls} + \frac{\tau}{4\lambda} & \text{if } \frac{1}{2\lambda}\beta_j^{ls} + \frac{\tau}{4\lambda} > 0 \\ 0 & -\frac{\tau}{2} \leq \beta_j^{ls} \leq \frac{\tau}{2} \\ \frac{1}{2\lambda}\beta_j^{ls} - \frac{\tau}{4\lambda} & \text{if } \frac{1}{2\lambda}\beta_j^{ls} - \frac{\tau}{4\lambda} < 0 \end{cases}$$

This pattern clearly expression β_0 in a form identical to soft thresholding function $S_{\frac{\tau}{4\lambda}}(\beta_j^{ls})$.

References

[1] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.