AM 221: Advanced Optimization Spring 2018

Kojin Oshiba *Problem Set 1 — Due Wed, Jan 31st at 23:59*

Instructions: All your solutions should be prepared in LaTeX and the PDF and .tex should be submitted to Canvas. Please submit all your files as ONE archive of filetype zip, tgz, or tar.gz. For each question, a well-written and correct answer will be selected a sample solution for the entire class to enjoy. If you prefer that we do not use your solutions, please indicate this clearly on the first page of your assignment.

The programming parts can be written in Python, Matlab, or Julia. If you strongly wish to use another language, please contact the instructor to ask for permission.

1. Sequences, Limits, Functions

a. Remember that the field $\mathbb{R}$ is characterized (among ordered fields containing $\mathbb{Q}$) by the least upper bound property: every non-empty bounded set has a least upper bound. Use this property to show that any non-decreasing upper-bounded sequence of real numbers is convergent.

Let $\{a_n\}$ be a sequence of non-decreasing upper-bounded sequence of real numbers. Since $\{a_n\}$ is a non-empty bounded set of real numbers, from the least upper bound property, $\{a_n\}$ has a leat upperbound. Let $b = sup_n\{a_n\}$ be this least upper bound.

Then, for any ϵ, there exists $a_N \in \{a_n\}$ such that $a_N > b - \epsilon$. This is because if not, $b - \epsilon = sup_n\{a_n\}$ and hence there's a contradiction. Since $\{a_n\}$ is non-decreasing, $|b - a_n| \leq |b - a_N| < c, \forall n \geq N, \forall \epsilon$. Hence, by the definition of limit, $a_n \to b$ and so $\{a_n\}$ is convergent.

b. Let $u = (u_n)_{n\geq 0}$ and $v = (v_n)_{n\geq 0}$ be two sequences of real numbers such that:

- u is non-increasing and v is non-decreasing
- $\lim_{n\to\infty} u_n - v_n = 0$

Show that both u and v are convergent and that they have the same limit.

Suppose u is unbounded and v is bounded. Then v_n has a finite limit, whereas u_n doesn't. Hence u_n dominates v_n in $lim_{n\to\infty} u_n - v_n$ and hence it doesn't converge. Similarly suppose u is bounded and v is unbounded. Then v_n dominates u_n in $lim_{n\to\infty} u_n - v_n$ and hence it doesn't converge. Finally, when u and v are both unbounded, because u is non-increasing and v is non-decreasing, $lim_{n\to\infty} u_n - v_n = -\infty - \infty = -\infty$.

Hence, u and v must be both bounded. From the least upper bound property, u has a lower bound and v has an upper bound. Hence, they are both convergent. Moreover because $lim_{n\to\infty} u_n - v_n = lim_{n\to\infty} u_n - lim_{n\to\infty} v_n = 0$, they must have the same limit.

c. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function such that $f(0) > 0$. Show that there exists $\varepsilon > 0$ such that:

$$|x| < \varepsilon \Rightarrow f(x) > 0.$$

Suppose u is unbounded and v is bounded. Then v_n has a finite limit, whereas u_n doesn't. Hence u_n dominates v_n in $lim_{n\to\infty} u_n - v_n$ and hence it doesn't converge. Similarly suppose u is bounded and v is unbounded. Then v_n dominates u_n in $lim_{n\to\infty} u_n - v_n$ and hence it doesn't converge. Finally, when u and v are both unbounded, because u is non-increasing and v is non-decreasing, $lim_{n\to\infty} u_n - v_n = -\infty - \infty = -\infty$.

Hence, u and v must be both bounded. From the least upper bound property, u has a lower bound and v has an upper bound. Hence, they are both convergent. Moreover because $lim_{n\to\infty} u_n - v_n = lim_{n\to\infty} u_n - lim_{n\to\infty} v_n = 0$, they must have the same limit.

2. Linear Algebra A is a matrix in $\mathbb{R}^{m\times n}$ with $m \leq n$.

a. Give the definition of the rank of A. What is the largest possible rank of A?

The rank of A is the dimension of the vector space spanned by its columns. The largest possible rank of A is m because $m \leq n$ and thus there can be at most m linearly independent columns.

b. Let us denote by $\mathbf{a}_1, \ldots, \mathbf{a}_m$ the rows of A, *i.e* $\mathbf{a}_i \in \mathbb{R}^n$ and $A = [\mathbf{a}_1 \ \ldots \ \mathbf{a}_m]^\intercal$. Show that:

$$\text{rank}(A) < m \Leftrightarrow \exists \mathbf{x} \in \mathbb{R}^m \setminus \{0\},\ \sum_{i=1}^{m} \mathbf{a}_i x_i = 0$$

From the fundamental theorem of linear algebra, the dimension of the vector space spanned by its rows is equivalent to the dimension of the vector space spanned by its columns. Hence, the equivalent definition of the rank of A is the dimension of the vector space spanned by its rows. Therefore

$$\begin{aligned} rank(A) < m &\Leftrightarrow \text{the dimension of the vector space spanned by its rows is less than } m \\ &\Leftrightarrow \mathbf{a}_1, ..., \mathbf{a}_m \text{ are linearly dependent} \\ &\Leftrightarrow \exists \mathbf{x} \in \mathbb{R}^m \setminus \{0\}, \sum_{i=1}^{m} \mathbf{a}_i x_i = 0 \quad (\because \text{the definition of linear dependence}) \end{aligned}$$

3. Inner product, norm For $\mathbf{x}$ and $\mathbf{y}$ two vectors of $\mathbb{R}^d$, we write $\mathbf{x}^\intercal \mathbf{y} = \sum_{i=1}^{d} x_i y_i$ their inner product. $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\intercal \mathbf{x}}$ denotes the Euclidean norm of $\mathbf{x}$.

a. Let $\mathbf{a} \in \mathbb{R}^d$. Show that:

$$\mathbf{a} = 0 \Longleftrightarrow \forall \mathbf{x} \in \mathbb{R}^d,\ \mathbf{a}^\intercal \mathbf{x} = 0$$
$$\mathbf{a} \geq 0 \Longleftrightarrow \forall \mathbf{x} \geq 0,\ \mathbf{a}^\intercal \mathbf{x} \geq 0$$

Proof of $\mathbf{a} = 0 \Rightarrow \forall \mathbf{x} \in \mathbb{R}^d,\ \mathbf{a}^\intercal \mathbf{x} = 0$

$$\mathbf{a} = 0 \Rightarrow \forall \mathbf{x} \in \mathbb{R}^d, \sum_{i=1}^{d} a_i x_i = 0 \Rightarrow \forall \mathbf{x} \in \mathbb{R}^d, \mathbf{a}^\intercal \mathbf{x} = 0$$

Proof of $\mathbf{a} = 0 \Leftarrow \forall \mathbf{x} \in \mathbb{R}^d,\ \mathbf{a}^\intercal \mathbf{x} = 0$

$$\forall \mathbf{x} \in \mathbb{R}^d, \mathbf{a}^\intercal \mathbf{x} = 0 \Rightarrow \mathbf{a}^\intercal \mathbf{a} = 0 \Rightarrow \mathbf{a} = 0$$

Proof of $\mathbf{a} \geq 0 \Rightarrow \forall \mathbf{x} \geq 0,\ \mathbf{a}^\intercal \mathbf{x} \geq 0$

$$\mathbf{a} \geq 0 \Rightarrow \forall \mathbf{x} \geq 0, \sum_{i=1}^{d} a_i x_i \geq 0 \Rightarrow \forall \mathbf{x} \geq 0, \mathbf{a}^\intercal \mathbf{x} \geq 0$$

Proof of $\mathbf{a} \geq 0 \Leftarrow \forall \mathbf{x} \geq 0,\ \mathbf{a}^\intercal \mathbf{x} \geq 0$

$$\forall \mathbf{x} \geq 0, \mathbf{a}^\intercal \mathbf{x} \geq 0 \Rightarrow \forall \mathbf{x} \geq 0, \sum_{i=1}^{d} a_i x_i \geq 0$$

Suppose for contradiction that $a_i < 0$ for some i. Then letting $x_i \geq \sum_{j \neq i} x_j$ will result in $\sum_{i=1}^{d} a_i x_i \leq 0$. This is true if at least one element of $\mathbf{a}$ is negative. Hence, contradiction. Therefore, all elements of $\mathbf{a}$ must be positive, meaning $\mathbf{a} \geq 0$.

b. Let $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d$. Show that:

$$\mathbf{x}^\intercal \mathbf{y} = \frac{\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2}{2} = \frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2}{2}$$

$$\mathbf{x}^\intercal \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

$$\begin{aligned}
\frac{||\mathbf{x} + \mathbf{y}||^2 - ||\mathbf{x}||^2 - ||\mathbf{y}||^2}{2} &= \frac{(\mathbf{x} + \mathbf{y})^\intercal (\mathbf{x} + \mathbf{y}) - \mathbf{x}^\intercal \mathbf{x} - \mathbf{y}^\intercal \mathbf{y}}{2} \\
&= \frac{\sum_{i=1}^{d} (x_i + y_i)^2 - \sum_{i=1}^{d} x_i^2 - \sum_{i=1}^{d} y_i^2}{2} \\
&= \sum_{i=1}^{d} x_i y_i
\end{aligned}$$

$$\begin{aligned}\frac{||\mathbf{x}||^2+||\mathbf{y}||^2-||\mathbf{x}+\mathbf{y}||^2}{2} &= \frac{\mathbf{x}^\intercal\mathbf{x}+\mathbf{y}^\intercal\mathbf{y}-(\mathbf{x}-\mathbf{y})^\intercal(\mathbf{x}-\mathbf{y})}{2} \\ &= \frac{\sum_{i=1}^d x_i^2+\sum_{i=1}^d y_i^2-\sum_{i=1}^d(x_i-y_i)^2}{2} \\ &= \sum_{i=1}^d x_i y_i\end{aligned}$$

Hence,

$$x^\intercal y = \frac{||x+y||^2-||x||^2-||y||^2}{2} = \frac{||x||^2+||y||^2-||x+y||^2}{2}$$

c. Deduce the parallelogram law:

$$\|\mathbf{x}+\mathbf{y}\|^2+\|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2+2\|\mathbf{y}\|^2$$

From b,

$$\frac{||\mathbf{x}+\mathbf{y}||^2-||\mathbf{x}||^2-||\mathbf{y}||^2}{2} = \frac{||\mathbf{x}||^2+||\mathbf{y}||^2-||\mathbf{x}+\mathbf{y}||^2}{2} \Leftrightarrow \|\mathbf{x}+\mathbf{y}\|^2+\|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2+2\|\mathbf{y}\|^2$$

d. Let us denote by $B_2(0,1)$ the unit ball of $\mathbb{R}^d$, $B_2(0,1) \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \leq 1\}$ and let us consider $\mathbf{v} \in \mathbb{R}^d$. Show that:

$$\max_{\mathbf{x}\in B_2(0,1)} \mathbf{v}^\intercal\mathbf{x} = \|\mathbf{v}\|$$

$$(\mathbf{v}^\intercal\mathbf{x})^2 = \mathbf{v}^\intercal\mathbf{v}\mathbf{x}^\intercal\mathbf{x} = ||\mathbf{x}||^2||\mathbf{v}||^2$$

Since $||\mathbf{x}|| \leq 1$,

$$\max_{\mathbf{x}\in B_2(0,1)} (\mathbf{v}^\intercal\mathbf{x})^2 = ||\mathbf{v}||^2$$

Hence,

$$\max_{\mathbf{x}\in B_2(0,1)} (\mathbf{v}^\intercal\mathbf{x}) = ||\mathbf{v}||$$

4. Multivariate Calculus

a. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function twice. For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{d} \in \mathbb{R}^n$, we define the function $f_{\mathbf{x},\mathbf{d}} : \mathbb{R} \to \mathbb{R}$ by:

$$f_{\mathbf{x},\mathbf{d}}(\lambda) = f(\mathbf{x}+\lambda\mathbf{d})$$

Express the first and second derivative of $f_{\mathbf{x},\mathbf{d}}$ in terms of the gradient and Hessian of f.

Let $g_{\mathbf{x},\mathbf{d}}(\lambda) = \mathbf{x} + \lambda \mathbf{d}$.

$$\begin{aligned}
\frac{d}{d\lambda} f_{\mathbf{x},\mathbf{d}}(\lambda) &= \frac{d}{d\lambda} f(g_{\mathbf{x},\mathbf{d}}(\lambda)) \\
&= \frac{d}{dg_{\mathbf{x},\mathbf{d}}} f(g_{\mathbf{x},\mathbf{d}}(\lambda)) \frac{d}{d\lambda} g_{\mathbf{x},\mathbf{d}}(\lambda) \\
&= \nabla f(g_{\mathbf{x},\mathbf{d}}(\lambda)) \mathbf{d} \\
&= \mathbf{d}^\intercal \nabla f(g_{\mathbf{x},\mathbf{d}}(\lambda))
\end{aligned}$$

$$\begin{aligned}
\frac{d^2}{d\lambda^2} f_{\mathbf{x},\mathbf{d}}(\lambda) &= \mathbf{d}^\intercal \frac{d}{d\lambda} \nabla f(g_{\mathbf{x},\mathbf{d}}(\lambda)) \\
&= \mathbf{d}^\intercal \frac{d}{dg_{\mathbf{x},\mathbf{d}}} \nabla f(g_{\mathbf{x},\mathbf{d}}(\lambda)) \frac{d}{d\lambda} g_{\mathbf{x},\mathbf{d}}(\lambda) \\
&= \mathbf{d}^\intercal H(\mathbf{x} + \lambda) \mathbf{d}
\end{aligned}$$

b. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\mathbf{x}$ be a local minimum of f, *i.e* there exists $\varepsilon > 0$ such that:

$$\|\mathbf{y} - \mathbf{x}\| \leq \varepsilon \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x})$$

show that $\nabla f(\mathbf{x}) = 0$. **Hint:** remember the Taylor expansion of f at $\mathbf{x}$:

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^\intercal \nabla f(\mathbf{x}) + o(\|\mathbf{h}\|)$$

From Taylor expansion of f at $\mathbf{x}$, we have:

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \mathbf{h}^\intercal \nabla f(\mathbf{x}) + o(\|\mathbf{h}\|)$$

Hence,

$$\mathbf{h}^\intercal \nabla f(\mathbf{x}) = f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - o(\|\mathbf{h}\|)$$

Let $y = h + x$. Then,

$$\mathbf{h}^\intercal \nabla f(\mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}) - o(\|\mathbf{y} - \mathbf{x}\|)$$

Since there exists $\varepsilon > 0$ such that:

$$\|\mathbf{y} - \mathbf{x}\| \leq \varepsilon \Rightarrow f(\mathbf{y}) \geq f(\mathbf{x})$$

We have that

$$\mathbf{h}^\intercal \nabla f(\mathbf{x}) \geq 0$$

From Taylor expansion, we also have:

$$f(\mathbf{x} - \mathbf{h}) = f(\mathbf{x}) - \mathbf{h}^\intercal \nabla f(\mathbf{x}) + o(\|\mathbf{h}\|)$$

In a similar procedure, we have that

$$\mathbf{h}^\intercal \nabla f(\mathbf{x}) \leq 0$$

Therefore, it must be the case that

$$\mathbf{h}^\intercal \nabla f(\mathbf{x}) = 0$$

Because this is true for any $\mathbf{h}$,

$$\nabla f(\mathbf{x}) = 0$$

c. Let $\mathbf{a} \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$. What are the gradients of $f(\mathbf{x}) = \mathbf{a}^\intercal \mathbf{x}$, $g(\mathbf{x}) = \|\mathbf{x}\|^2$ and $h(\mathbf{x}) = \mathbf{x}^\intercal M \mathbf{x}$?

$$\frac{\partial}{\partial x_i} f(\mathbf{x}) = \frac{\partial}{\partial x_i} \sum_i a_i x_i = a_i$$

Hence, $\nabla f(\mathbf{x}) = \mathbf{a}$.

$$\frac{\partial}{\partial x_i} g(\mathbf{x}) = \frac{\partial}{\partial x_i} \sum_i x_i^2 = 2x_i$$

Hence, $\nabla g(\mathbf{x}) = 2\mathbf{x}$.

$$\begin{aligned}
\frac{\partial}{\partial x_i} h(\mathbf{x}) &= \frac{\partial}{\partial x_i} \sum_{j,k} M_{jk} x_j x_k \\
&= \frac{\partial}{\partial x_i} (\sum_{j,k\neq i} M_{jk} x_j x_k + \sum_{k\neq i} M_{ik} x_i x_k + \sum_{j\neq i} M_{ji} x_j x_i \\
&= \sum_{k\neq i} M_{ik} x_k + \sum_{j\neq i} M_{ji} x_j + 2A_{ii} x_i \\
&= \sum_{k} M_{ik} x_k + \sum_{j} M_{ji} x_j \\
&= (M\mathbf{x})_i + (\mathbf{x}^\intercal M)_i \\
&= (M\mathbf{x})_i + (M^\intercal \mathbf{x})_i \\
&= (M + M^\intercal) x_i
\end{aligned}$$

Hence, $\nabla h(\mathbf{x}) = (M + M^\intercal)\mathbf{x}$.

5. Programming Download the file at `http://rasmuskyng.com/am221_spring2018/psets/hw1/access.log`. This file is a server log, each line has the following format:

```
<time>\t<ip-adress>
```

i.e it contains a time and the IP address which accessed the server at that time; the time and the IP address are separated by a tab character. Using the programming language of your choice, write a program to find the list of the ten IP addresses who accessed the server the most (in decreasing order). Report the list you obtained, as a text file with one IP address per line, and report the code you used.

The output text file is `top_ten_ips.txt`. The code used is `hw1.py`.