DATA130026 Optimization Assignment 11

Due Time: Jun. 15, 2024
姓名: 雍崔扬
学号: 21307140051

Problem 1

For each of the following functions on $\mathbb{R}^n$, explain how to calculate a subgradient at a given x.

- **Lemma:**
  假设 f 是具有开凸定义域的凸函数，以保证次微分 $\partial f(x)$ 存在.
  我们有以下微积分规则：
  - **非负缩放 (Nonnegative scaling):** 任意 $\alpha > 0$ 都有 $\partial(\alpha f) = \alpha \partial f$
  - **加法 (Addition):** $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
  - **仿射变换 (Affine transformation):** $\partial(f(Ax + b)) = A^T \partial f(Ax + b)$
  - **有限驻点最大值 (Finite pointwise maximum):**
    若 $f(x) = \max\limits_{i=1,\dots,m} f_i(x)$，则 $\partial f(x) = \mathrm{conv}\{\bigcup\limits_{i=1}^{m} \{\partial f_i(x) : f_i(x) = f(x)\}\}$
    即在 x 处积极函数次微分的并集的凸包 (所谓积极就是 $f_i(x) = f(x)$)
    - 特别地，若 $f_1, \dots, f_m$ 均在 x 处可微，则我们有：
      $$\begin{aligned}\partial f(x) &= \mathrm{conv}\{\bigcup_{i=1}^{m} \{\nabla f_i(x) : f_i(x) = f(x)\}\} \\ &= \{\sum_{i \in I_{\text{active}}(x)} \alpha_i \nabla f_i(x) \Big| \sum_{i \in I_{\text{active}}(x)} \alpha_i = 1,\ \alpha_i \geq 0\}\end{aligned}$$
      其中 x 处的积极函数指标集 $I_{\text{active}}(x) = \{i : f_i(x) = f(x) = \max\limits_{i=1,\dots,m} f_i(x)\}$

(a) $f(x) = \sup\limits_{0 \leq t \leq 1} p(t)$, where $p(t) = x_1 + x_2 t + \ldots + x_n t^{n-1}$.

- **Solution:**
  这是**分段线性函数**的推广.
  记向量 $(1, t, \dots, t^{n-1})^T$ 为 $\nabla_x p(t)$
  定义**积极指标集**:
  $$\begin{aligned}I_{\text{active}}(x) &= \arg\sup_{t \in [0,1]} \{p(t)\} \\ &= \{t \in [0,1] : p(t) = f(x) = \sup_{0 \leq t \leq 1} p(t)\}\end{aligned}$$
  f 在 x 处的**次微分** $\partial f(x) = \mathrm{conv}\{\nabla_x p(t) : t \in I_{\text{active}}(x)\}$
  因此我们只需确定 $I_{\text{active}}(x)$，即可得到 $\partial f(x)$，进而从中选取 f 在 x 处的**次梯度**.

  下面我们描述确定积极指标集 $I_{\text{active}}(x)$ 的方法：
  考虑优化问题 $\sup\limits_{0 \leq t \leq 1} p(t)$，它是凸优化问题，因而 KKT 条件是全局最优解的**充要条件**.
  定义 Lagrange 函数为：
  $$\begin{aligned}L(t, \lambda_1, \lambda_2) &= p(t) + \lambda_1(t-1) + \lambda_2(-t) \\ &= (x_1 - \lambda_1) + (x_2 + \lambda_1 - \lambda_2)t + x_3 t^2 + \cdots + x_n t^{n-1}\end{aligned}$$
  KKT 条件为：
  - **一阶最优条件：**
    $\frac{\partial}{\partial t} L(t, \lambda_1, \lambda_2) = (x_2 + \lambda_1 - \lambda_2) + 2x_3 t + \cdots + (n-1)x_n t^{n-2} = 0$
  - **可行性条件：** $0 \leq t \leq 1$

- **对偶可行性条件：** $\lambda_1, \lambda_2 \geq 0$
- **互补松弛条件：** $\begin{cases} \lambda_1(t-1) = 0 \\ \lambda_2 t = 0 \end{cases}$

分成平凡情况 $(t = 0, 1)$ 和非平凡情况 $(0 < t < 1)$ 讨论后，
我们知道全局最优解的**候选解**有：

- $t = 0, 1$ (边界点)
- 关于 t 的方程 $x_2 + 2x_3 t + \cdots + (n-1)x_n t^{n-2} = 0$ 在 $(0, 1)$ 区间上的解

我们在这些候选解中比较选取 $\sup\limits_{0 \leq t \leq 1} p(t)$ 的**全局最优解**
最后用它们构造**积极指标集** $I_{\text{active}}(x)$

(b) $f(x) = x_{[1]} + x_{[2]} + \ldots + x_{[k]}$, where $x_{[i]}$ denotes the i-th largest elements of x.

- **Solution:**
  我们依然可以将 $f(x)$ 表示为**分段线性函数**：
  定义权重向量集合 $A_k = \{a \in [0,1]^n : 1_n^T a = k\}$
  则我们可以将 $f(x)$ 表示为分段线性函数 $f(x) = \sup\limits_{a \in A_k} a^T x$ 的形式.

  定义**积极权重集**：
  $$\begin{aligned} A_{\text{active}}(x) &= \arg\max_{a \in A_k} \{a^T x\} \\ &= \{a \in A_k : a^T x = f(x) = \sup_{a \in A_k} a^T x\} \\ &= \{a : a_i = \begin{cases} 1 & \text{if } x_i \in \{x_{[1]}, \ldots, x_{[k]}\} \\ 0 & \text{if } x_i \notin \{x_{[1]}, \ldots, x_{[k]}\} \end{cases}\} \end{aligned}$$
  (它不一定是单元素集，存在前 k 大分量的组合不唯一的情形)

  于是 f 在 x 处的次微分 $\partial f(x) = \text{conv}\{A_{\text{active}}(x)\}$
  我们只需在其中选取次梯度即可.

(c) $f(x) = \|Ax - b\|_2 + \|x\|_2$ where $A \in \mathbb{R}^{m \times n}$.

- **Solution:**
  考虑 Euclid 范数 $h(x) = \|x\|_2$ 的次微分，它仅在 $x = 0_n$ 处不可微.
  根据次梯度的定义，我们有 $\|y\|_2 \geq \|0_n\|_2 + g^T(y - 0_n) \ \ (\forall\, y \in \mathbb{R}^n)$
  即 $\|y\|_2 \geq g^T y \ \ (\forall\, y \in \mathbb{R}^n)$
  - 当 $\|g\|_2 \leq 1$ 时，
    根据 **Cauchy-Schwarz 不等式**有 $g^T y \leq \|g\|_2 \|y\|_2 \leq \|y\|_2 \ \ (\forall\, y \in \mathbb{R}^n)$ 成立.
  - 当 $\|g\|_2 > 1$ 时，
    取 $y = g$，我们有 $\|g\|_2 \geq g^T g = \|g\|_2$ 成立，与 $\|g\|_2 > 1$ 相矛盾.

  因此我们有 $\partial h(x) = \begin{cases} \{\frac{x}{\|x\|_2}\} & \text{if } x \neq 0_n \\ \{g : \|g\|_2 \leq 1\} & \text{if } x = 0_n \end{cases}$
  整理可得 $\partial h(x) = \{g \in \mathbb{R}^n : g^T x = \|x\|_2,\ \|g\|_2 \leq 1\}$

因此 $f(x)=\|Ax-b\|_2+\|x\|_2=h(Ax-b)+h(x)$ 的次微分为：(假设 $A\in\mathbb{R}^{m\times n}$)

$$\begin{aligned}\partial f(x)&=\partial(h(Ax-b))+h(x)\\&=A^T\partial h(Ax-b)+h(x)\\&=\{A^Tg_1+g_2:\begin{cases}g_1\in\mathbb{R}^m\\g_2\in\mathbb{R}^n\\g_1^T(Ax-b)=\|Ax-b\|_2\\\|g_1\|_2\le 1\\g_2^Tx=\|x\|_2\\\|g_2\|_2\le 1\end{cases}\}\end{aligned}$$

我们可由这样选取次梯度：

$$g=\begin{cases}A^T\frac{Ax-b}{\|Ax-b\|_2}+\frac{x}{\|x\|_2} & \text{if } Ax-b\neq 0_m \text{ and } x\neq 0_n\\A^T\frac{Ax-b}{\|Ax-b\|_2} & \text{if } Ax-b\neq 0_m \text{ and } x=0_n\\\frac{x}{\|x\|_2} & \text{if } Ax-b=0_m \text{ and } x\neq 0_n\\0_n & \text{if } Ax-b=0_m \text{ and } x=0_n\end{cases}$$

Problem 2

(Subgradient of the maximum eigenvalue function)
Consider the function $f:\mathbb{S}^n\to\mathbb{R}^n$
given by $f(X)=\lambda_{\max}(X)$ (recall that $\mathbb{S}^n$ is the set of all $n\times n$ symmetric matrices).
Let $X\in\mathbb{S}^n$ and let v be a normalized eigenvector of X $(\|v\|_2=1)$
associated with the maximum eigenvalue of X.
Show that $vv^T\in\partial f(X)$.

Solution:
要证明 $vv^T\in\partial f(X)$,
等价于证明 $(vv^T)\bullet(Y-X)\le f(Y)-f(X)=\lambda_{\max}(Y)-\lambda_{\max}(X)\quad(\forall\,Y\in\mathbb{S}^n)$
(其中 $\bullet$ 代表矩阵内积，有 $A\bullet Z=\mathrm{tr}(A^TZ)=\mathrm{tr}(AZ)$ 成立)

对于任意 $Y\in\mathbb{S}^n$ 我们都有：

$$\begin{aligned}(vv^T)\bullet(Y-X)&=\mathrm{tr}(vv^T(Y-X))\\&=\backslash\mathrm{tr}(v^TYv)-\backslash\mathrm{tr}(v^TXv)\\&=v^TYv-v^TXv\\&\le\lambda_{\max}(Y)-\lambda_{\max}(X)\quad(\text{Rayleigh-Ritz})\end{aligned}$$

Hermite 矩阵特征值的变分性质 **(Rayleigh-Ritz 定理)** 参见：
[FDU 高等线性代数 7. 内积空间 & Hermite矩阵特征值 & 奇异值 - 知乎 (zhihu.com)]

Problem 3

Consider the nonsmooth function

$$f(x)=\max_{1\le i\le K}x_i+\frac{1}{2}\|x\|_2^2$$

where $x\in\mathbb{R}^n$, $K\in[2,n]$ is a given positive integer.

(a) Calculate the minimizer $x_\star$ and the associated objective value $f_\star$.

- **Lemma: (非光滑凸函数的最优性条件, 江如俊教授 Note 9 Theorem 1.6)**
  若 $f(x)$ 是具有开凸定义域的凸函数，
  则 $x_\star$ 是 $f(x)$ 的全局最优解，当且仅当 $0_n \in \partial f(x_\star)$ 成立 (即 0_n 是 f 在 $x_\star$ 处的次梯度)

- **Solution:**
  显然目标函数 f 是一个**非光滑凸函数**.
  我们首先计算 f 在 x 处的**次微分**:

$$\begin{aligned}\partial f(x) &= \partial\{\max_{1\le i\le K} x_i\} + \nabla\{\frac{1}{2}\|x\|_2^2\} \\ &= \partial\{\max_{1\le i\le K} e_i^T x\} + x \\ &= \mathrm{conv}\{e_i : i \in I_{\text{active}}(x)\} + x\end{aligned}$$

  其中 e_i 为 $\mathbb{R}^n$ 空间的第 i 个标准正交基，
  $\mathrm{conv}\{\cdot\}$ 代表集合的凸包.
  积极指标集 $I_{\text{active}}(x) = \{i \in \{1,\dots,K\} \text{ such that } x_i = \max_{1\le i\le K} x_i\}$ (它显然不是空集)

  因此 $0_n \in \partial f(x)$ 就等价于:

  存在权重向量 $c \in \mathbb{R}^n$ 满足 $\begin{cases} c \succeq 0_n \\ 1_n^T c = 1 \\ c_i = 0 \text{ for all } i \notin I_{\text{active}}(x) \\ c + x = 0_n \end{cases}$

  - 通过**反证法**可以说明满足上述条件的 x 的前 K 个分量一定是相同的:
    因为**如若不然**，则存在 $\{1,\dots,K\}$ 中的一个 $i_0 \notin I_{\text{active}}(x)$
    (即前 K 个分量中至少有一个小于前 K 个分量的最大值)
    于是我们有 $x_{i_0} = -c_{i_0} = 0$
    由于 $x = -c \preceq 0_n$，故 $x_{i_0} = 0$ 一定是 x 前 K 个分量的最大值，
    这与我们 $i_0 \notin I_{\text{active}}(x)$ 的假设相矛盾.
    因此 x 的前 K 个分量一定是相同的.

  因此满足上述条件的 c 和 x 具有以下结构:

  - x 的前 K 个分量均等于 $-\frac{1}{K}$;
  - $c = -x$

  因此 $x_\star = (\underbrace{-\frac{1}{K},\dots,-\frac{1}{K}}_{K}, 0,\dots,0)$ 是 $\min f(x)$ 的全局最优解，
  全局最优值 $f_\star = f(x_\star) = -\frac{1}{K} + \frac{1}{2}\cdot K \cdot (-\frac{1}{K})^2 = -\frac{1}{2K}$

(b) Show that $f(x)$ is G-Lipschitz continuous for all $\|x\|_2 \le \frac{1}{\sqrt{K}}$, where $G = 1 + \frac{1}{\sqrt{K}}$.

- **Solution:**
  对于任意 x, y 满足 $\begin{cases} \|x\| \le \frac{1}{\sqrt{K}} \\ \|y\| \le \frac{1}{\sqrt{K}} \end{cases}$ 我们都有:

  - 一方面:

$$\begin{aligned}\frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 &= \frac{1}{2}(\|x\|_2 + \|y\|_2)(\|x\|_2 - \|y\|_2) \\ &\le \frac{1}{2}(\|x\|_2 + \|y\|_2)\cdot\|x-y\|_2 \\ &\le \frac{1}{2}(\frac{1}{\sqrt{K}} + \frac{1}{\sqrt{K}})\cdot\|x-y\|_2 \\ &= \frac{1}{\sqrt{K}}\|x-y\|_2\end{aligned}$$

- 另一方面：
  $$\begin{aligned}\max_{1\le i\le K} x_i - \max_{1\le i\le K} y_i &\le \max_{1\le i\le K} |x_i - y_i| \\ &\le \max_{1\le i\le n} |x_i - y_i| \\ &= \|x-y\|_\infty \\ &\le \|x-y\|_2\end{aligned}$$
  最后一步我们使用了 ℓ_p- 范数关于 p 单调递减的结论.

综上所述，我们有：

$$\begin{aligned}f(x) - f(y) &= \{\max_{1\le i\le K} x_i + \frac{1}{2}\|x\|_2^2\} - \{\max_{1\le i\le K} y_i + \frac{1}{2}\|y\|_2^2\} \\ &\le \|x-y\|_2 + \frac{1}{\sqrt{K}}\|x-y\|_2 \\ &= G\|x-y\|_2\end{aligned}$$

其中 $G = 1 + \frac{1}{\sqrt{K}}$

根据 x, y 的任意性可知 $f(x)$ 在 $\|x\|_2 \le \frac{1}{\sqrt{K}}$ 时 G-Lipschitz 连续.

(c) Suppose the initial point $x^{(1)} = 0_n$ and you use subgradient method to solve $\min f(x)$, where the stepsize can be arbitrary chosen.
Suppose the subgradient you use is $g = x + e_j$,
where j is the smallest integer satisfying $x_j = \min_{1\le i\le K} x_i$

Show that after k $(k < K)$ iterations, we have

$$f_{\text{best}}^{(k)} - f_\star \ge G\frac{\|x^{(1)} - x_\star\|}{2(1+\sqrt{K})},$$

where $f_{\text{best}}^{(k)} = \min_{1\le i\le k+1} f(x^{(i)})$.

Solution:
由于 $\min f(x) := \max_{1\le i\le K} x_i + \frac{1}{2}\|x\|_2^2$ 是无约束优化问题，
故**投影次梯度法**退化为**次梯度法**，
迭代公式为 $x^{(k+1)} = x^{(k)} - t_k g^{(k)}$
其中步长 t_k 的选取是任意的，
记 $x^{(k)}$ 前 K 个分量中**最小的最大指标**为 $j = \min\{I_{\text{active}}(x^{(k)})\}$
选取次梯度 $g^{(k)} = x^{(k)} + e_j$

我们简单推理 $x^{(k)}$ $(k \le K)$ 的迭代过程：

- $x^{(1)} = 0_n$ 前 K 个分量中最小的最大指标 $j = 1$
  因此 $g^{(1)} = x^{(1)} + e_1 = e_1$
  于是 $x^{(2)} = x^{(1)} - t_1 g^{(1)} = 0_n - t_1 e_1 = -t_1 e_1$
- $x^{(2)} = -t_1 e_1$ 前 K 个分量中最小的最大指标 $j = 2$
  因此 $g^{(2)} = x^{(2)} + e_2$
  于是 $x^{(3)} = x^{(2)} - t_2 g^{(2)} = (1-t_2)x^{(2)} - t_2 e_2$

  考虑 $x^{(3)}$ 前 K 个分量中最小的最大指标 j：
  - 当 $0 < t_2 < 1$ 时，$j = 3$ (开始优化第三个分量)
  - 当 $t_2 \ge 1$ 时，$j = 1$ (要重新将第一个分量降到 0 以下才可以开始优化第三个分量)

  这暗示 $t_k \ge 1$ 对于迭代来说是不利的.

- 更严格地，假设 $x^{(m)}=(-c_1,\dots,-c_{k-1},0,\dots,0)\ \ (k\le K)$
  其中 $c_1,\dots,c_{k-1}>0$
  其前 K 个分量中最小的最大指标 $j=k$
  因此 $g^{(m)}=x^{(m)}+e_j=x^{(m)}+e_k$
  于是 $x^{(m+1)}=x^{(m)}-t_mg^{(m)}=(1-t_m)x^{(m)}-t_me_k$

  考虑 $x^{(m+1)}$ 前 K 个分量中最小的最大指标 j:

  - 当 $0<t_m<1$ 时，$j=k+1$ (开始优化第 $k+1$ 个分量)
  - 当 $t_m\ge 1$ 时，$j\le k-1$ (要重新将前 $k-1$ 个分量降到 0 以下才可以开始优化第 $k+1$ 个分量)

根据上述推理我们可以预见，理想的迭代过程要求每步步长 $t_k\in(0,1)$
这保证了参与优化的分量数量随着迭代是增加的 (第 $k\le K$ 步中前 k 个分量参与优化)
容易验证，上述理想的迭代过程产生的点列结构如下:

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$\cdots$	$x^{(k)}$
1st	0	$-t_1$	$-t_1(1-t_2)$	$-t_1(1-t_2)(1-t_3)$	$\cdots$	$-t_1\prod_{i=2}^{k-1}(1-t_i)$
2nd	0	0	$-t_2$	$-t_2(1-t_3)$	$\cdots$	$-t_2\prod_{i=3}^{k-1}(1-t_i)$
3rd	0	0	0	$-t_3$	$\cdots$	$-t_3\prod_{i=4}^{k-1}(1-t_i)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$(k-2)$th	0	0	0	0	$\cdots$	$-t_{k-2}(1-t_{k-1})$
$(k-1)$th	0	0	0	0	$\cdots$	$-t_{k-1}$
k-th	0	0	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
n-th	0	0	0	0	$\cdots$	0

我们看到 $0<t_k<1\ (\forall\ k\le K)$ 保证了某一维分量一旦开始优化就会保持负值，
并且随着迭代进行向 0 靠拢.
考虑到我们的优化目标 (即全局最优解) $x_\star=(\underbrace{-\frac{1}{K},\dots,-\frac{1}{K}}_{K},0,\dots,0)$
我们应将 t_k 取值范围缩小到 $(\frac{1}{K},1)$
But does it help us anywhere?

根据上面的分析可知:
经过 $k<K$ 步迭代后，
即使是最理想的情况也只能保证 $x^{(k+1)}$ 的前 $k<K$ 个分量都 <0
因此对于任意 $j=1,\dots,k+1$，$x^{(j)}$ 的前 K 个分量中的最大值都是 0
故 $f(x^{(j)})=\max\limits_{1\le i\le K}x_i^{(j)}+\frac{1}{2}\|x^{(j)}\|^2=0+\frac{1}{2}\|x^{(j)}\|^2\ge 0$
注意到 $f(x^{(1)})=f(0_n)=0$
因此 $f_{\text{best}}^{(k)}=\min\limits_{1\le i\le k+1}f(x^{(i)})=f(x^{(1)})=0$

要证明 $f_{\text{best}}^{(k)}-f_\star\ge G\frac{\|x^{(1)}-x_\star\|}{2(1+\sqrt{K})}\ \ (\forall\ k<K)$
只要证明 $f_\star+G\frac{\|x^{(1)}-x_\star\|}{2(1+\sqrt{K})}\le 0$ 即可.

$$f_\star + G\frac{\|x^{(1)} - x_\star\|}{2(1+\sqrt{K})} = -\frac{1}{2K} + (1+\frac{1}{\sqrt{K}})\frac{\sqrt{K\cdot(-\frac{1}{K})^2}}{2(1+\sqrt{K})} = 0$$

命题得证.