Name: ____________ Student ID: ____________ Scores: ____________

AIE6003 Homework 2
Instructor: Xiaopeng Li **Deadline**: Mar 16

Notice: If you are trying to invoke a result outside of this course, then you should put a reference to it. The questions with 1 star are easy, with 2 stars are medium, and with 3 stars are hard. Easy and medium level of questions are likely to occur in the exams, while hard questions are not likely to occur in the exams.

1**. (10 points) Let $f : \mathbb{R}^n \to \mathbb{R}$ be ℓ-smooth and convex. Let $X^* = \arg\min_{x\in\mathbb{R}^n} f(x)$ be nonempty. Consider the following algorithm

$$x_{k+1} = x_k - \gamma g_k \text{ where } \mathbb{E}[g_k \mid x_k] = \nabla f(x_k) \text{ and } \mathbb{E}[\|g_k - \nabla f(x_k)\|^2] \le \sigma^2, \quad \forall\, k \ge 1,$$

where $\gamma \in (0, 2/\ell)$ and $x_1 \in \mathbb{R}^n$ is an initial point. Denote $f^* := f(x^*)$ for all $x^* \in X^*$.

(a) (4 points) Prove that for all $k \ge 1$,

$$\begin{aligned}\|x_{k+1} - x^*\|^2 \le \|x_k - x^*\|^2 &- (2\gamma - \ell\gamma^2)(f(x_k) - f^*)\\ &- 2\gamma\langle x_k - \gamma\nabla f(x_k) - x^*, g_k - \nabla f(x_k)\rangle + \gamma^2\|g_k - \nabla f(x_k)\|^2.\end{aligned}$$

(b) (3 points) Prove that for all fixed $T \ge 1$,

$$\sum_{k=1}^{T}(2\gamma - \ell\gamma^2)\mathbb{E}[f(x_k) - f^*] \le \|x_1 - x^*\|^2 + \gamma^2\sigma^2 T.$$

(c) (3 points) Fix a $T \ge 1$. Let $\bar{x}_T = \frac{1}{T}\sum_{k=1}^{T} x_k$. Show that

$$\mathbb{E}[f(\bar{x}_T)] - f^* = O\left(\frac{1}{T} + \frac{\sigma}{\sqrt{T}}\right),$$

when appropriate γ is chosen. Notice that γ can depend on T, but it is still a constant step size throughout iterations $k = 1, \ldots, T$.

2**. (15 points) Suppose C_1 and C_2 are two closed convex set in $\mathbb{R}^n$ and $C_1 \cap C_2$. The goal of this exercise is to derive a simple algorithm to find a point $x \in C_1 \cap C_2$. Consider the subgradient method with Polyak step size to minimize a convex function $f : \mathbb{R}^n \to \mathbb{R}$ with $x^* \in \operatorname{argmin} f \neq \emptyset$,

$$\forall k \in \mathbb{N},\ x_{k+1} = x_k - \alpha_k s_k,\ \text{ where } s_k \in \partial f(x_k),\ s_k \neq 0,\ \text{ and } \alpha_k = \frac{f(x_k) - f(x^*)}{\|s_k\|^2}.$$

(a) (2 points) Prove that the function

$$f(x) := \max\{d(x, C_1), d(x, C_2)\}$$

is convex, and any minimizer x^* of f satisfies $x^* \in C_1 \cap C_2$. Here $d(x, \bullet)$ is the distance from a point x to a set $\bullet$.

(b) (2 points) Let $I(x) := \{i \in \{1, 2\} : f(x) = f_i(x)\}$ where $f_1(x) := d(x, C_1)$ and $f_2(x) := d(x, C_2)$. Prove that

$$\bigcup_{i\in I(x)} \partial f_i(x) \subset \partial f(x).$$

(c) (5 points) Verify that if $f(x_k) > 0$, it is valid to choose the s_k in the subgradient method as

$$s_k = \frac{x_k - \Pi_{C_i}(x_k)}{\|x_k - \Pi_{C_i}(x_k)\|_2}$$

for any $i \in I(x_k)$, where $\Pi_{C_i}(x_k)$ denotes the projection of x_k onto C_i.

(d) (3 points) Given any initial point $x_0 \in \mathbb{R}^n$, show that the subgradient method with Polyak step size applied to the function f defined in part (a) is equivalent to

$$\begin{cases} x_{2k+1} = \Pi_{C_1}(x_{2k}) \\ x_{2k+2} = \Pi_{C_2}(x_{2k+1}) \end{cases}, \quad \forall\, k \in \mathbb{N}.$$

(e) (3 points) Show that $x_k \to x^*$ for some $x^* \in C_1 \cap C_2$.

3***. (25 points) In this question, we will look at the *community detection problem* in a social network. Suppose we have n users and build a symmetric nonnegative similarity matrix $M \in \mathbb{R}_+^{n\times n}$, where M_{ij} measures how strongly users i and j are connected (e.g., number of interactions or a weighted adjacency score). If users belong to a small number r of latent communities (e.g., in a university social network, latent communities can be departments, clubs, dorms, or research groups), we can model user i's community memberships by a nonnegative vector $x_i \in \mathbb{R}_+^r$. Collecting these membership vectors as rows of $X \in \mathbb{R}_+^{n\times r}$, the similarity between users i and j is then explained by shared memberships $x_i^\top x_j$, i.e., $(XX^\top)_{ij}$. This leads to the symmetric nonnegative matrix factorization problem

$$\min_{X \in \mathbb{R}_+^{n\times r}} \ f(X) := \frac{1}{2}\|M - XX^\top\|_F^2.$$

(a) (5 points) Is the objective function f in symmetric nonnegative matrix factorization ℓ-smooth? Why?

(b) (5 points) For a C^2 strictly convex function $h : \mathbb{R}^{n\times r} \mapsto \mathbb{R}$, we say f (also C^2) is ℓ-smooth relative to the distance kernel h if and only if for all $X, U \in \mathbb{R}^{n\times r}$,

$$\nabla^2 f(X)[U,U] \le \ell \nabla^2 h(X)[U,U], \text{ where } \nabla^2 f(X)[U,U] := \frac{d^2}{dt^2} f(X+tU)\Big|_{t=0}.$$

Therefore, the classical ℓ-smooth can be recovered by taking $h = \frac{1}{2}\|X\|_F^2$. Now we consider another h defined by

$$h(X) := \tfrac{3}{2}\|X\|_F^4 + \|M\|_F\|X\|_F^2.$$

Prove that the objective function f in symmetric nonnegative matrix factorization is 1-smooth relative to this h.

Hint: here is a simple example for computing the Hessian of $f(X) = \frac{1}{2}\|X\|_F^2 = \frac{1}{2}\langle X, X\rangle$.

$$f(X+tU) = \frac{1}{2}\|X+tU\|_F^2 = \frac{1}{2}\langle X+tU, X+tU\rangle = \frac{1}{2}\|X\|_F^2 + t\langle X, U\rangle + \frac{t^2}{2}\|U\|_F^2.$$

Hence, we can obtain

$$\frac{d^2}{dt^2} f(X+tU) = \|U\|_F^2 \implies \nabla^2 f(X)[U,U] = \frac{d^2}{dt^2} f(X+tU)\Big|_{t=0} = \|U\|_F^2.$$

(c) (5 points) Consider the Bregman proximal gradient algorithm for minimizing function $\Psi := f + g$:

$$X_{k+1} = T_\lambda(X_k), \quad T_\lambda(X) := \arg\min_{U \in \mathbb{R}^{n\times r}} \left\{ g(U) + f(X) + \langle \nabla f(X), U - X\rangle + \tfrac{1}{\lambda} D_h(U, X) \right\}$$

where D_h is the Bregman divergence induced by h and $g(X) := \iota_{\mathbb{R}_+^{n\times r}}(X)$ is the indicator function of nonnegative matrices of size $n \times r$. Prove that $T_\lambda(X) = \frac{1}{\tau(6\|[H]_+\|_F^2)}[H]_+$, where $H := (6\|X\|_F^2 + 2\|M\|_F)X - 2\lambda(XX^\top - M)X$, $[H]_+ = \max\{H, 0\}$ (entry-wise), and $\tau(c)$ denotes the unique real solution to the cubic equation $z^2(z - 2\|M\|_F) = c$.

(d) (5 points) Take $\lambda = 1/2$ in the above algorithm. Denote $\Psi(X) := f(X) + \iota_{\mathbb{R}_+^{n\times r}}(X)$ with f in the symmetric nonnegative matrix factorization problem. Prove that $(X_k)_{k\in\mathbb{N}}$ is bounded and $\min_{1\le k\le K} \operatorname{dist}(0, \partial\Psi(x_k)) = O(1/\sqrt{K})$.

(e) (5 points) In this part you will implement and compare two first-order methods for solving the symmetric nonnegative matrix factorization problem on a real community-detection dataset.

- **Dataset.** Use the SNAP *ego-Facebook* dataset.[1] Fix the ego user `1684` and load the file `1684.edges`. This file contains edges among the ego's friends (the ego node itself does not appear in the file, but is assumed to be connected to all listed friends).
- **Constructing** M**.** Let V be the set of node IDs appearing in `1684.edges` together with the ego node `1684`. Construct a symmetric sparse adjacency matrix $M \in \{0,1\}^{|V|\times|V|}$ by setting $M_{ij} = 1$ if there is an undirected edge listed in `1684.edges` and additionally setting $M_{\text{ego},v} = M_{v,\text{ego}} = 1$ for every friend $v \in V \setminus \{\text{ego}\}$ to include the implicit ego–friend edges.
- **Algorithms to compare.** Use the same initialization $X_0 \ge 0$ and the same rank $r = 10$. Compare the classical projected gradient descent (PGD) with a constant step size $\eta > 0$ and the Bregman proximal gradient algorithm introduced here, also with constant step size $\lambda > 0$.
- **Step-size tuning.** Tune the constant step sizes η and λ to obtain the best practical performance for each method on this dataset. Note that the empirically best step size may be much larger than a conservative bound such as one.
- **Performance comparison.** For each algorithm, plot: objective value $f(X^k)$ versus iteration k or wall-clock time, within a fixed iteration budget.

[1] Download from the SNAP website: `https://snap.stanford.edu/data/egonets-Facebook.html`.

Page 3