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

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

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 $k \geq 1$,

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

Solution: For any $x^* \in X^*$, and all $k \geq 1$, we have

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

Moreover, by refined gradient inequality, we have

$$\frac{1}{\ell}\|\nabla f(x_k)\|^2 \leq \langle \nabla f(x_k), x_k - x^*\rangle.$$

Combining the above two relations, we obtain, for any $k \geq 1$,

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

where the last inequality follows from the convexity of f and the fact that $\gamma \leq 2/\ell$.

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

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

Solution: Taking expectation conditioned on x_k on both sides of the result in part (a),

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \mid x_k] \le \|x_k - x^*\|^2 - (2\gamma - \ell\gamma^2)(f(x_k) - f^*) + \gamma^2 \mathbb{E}[\|g_k - \nabla f(x_k)\|^2 \mid x_k].$$

Taking expectation on both sides yields

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

Rearranging the terms and summing up from $k = 1$ to T, we have

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

(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 $k = 1, \ldots, T$.

Solution: From part (b), we have

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

If $\gamma \le 1/\ell$, then $2 - \ell\gamma \ge 1$ and the above inequality reduces to

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

The RHS is minimized at

$$\gamma^* := \min\left\{\frac{1}{\ell}, \frac{\|x_1 - x^*\|}{\sigma\sqrt{T}}\right\}.$$

By convexity of f, we obtain

$$\begin{aligned}
\mathbb{E}[f(\bar{x}_T)] - f^* &\le \frac{1}{T}\sum_{k=1}^{T} \mathbb{E}[f(x_k) - f^*] \le \frac{\|x_1 - x^*\|^2}{T\gamma^*} + \gamma^*\sigma^2 \\
&\le \frac{\|x_1 - x^*\|^2}{T}\max\left\{\ell, \frac{\sigma\sqrt{T}}{\|x_1 - x^*\|}\right\} + \frac{\|x_1 - x^*\|\sigma}{\sqrt{T}} \\
&\le \frac{\ell\|x_1 - x^*\|^2}{T} + \frac{2\|x_1 - x^*\|\sigma}{\sqrt{T}}.
\end{aligned}$$

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 $\bullet$.

Solution: f is convex because the distance function to a convex set is convex and point-wise maximum of two convex function is convex. The optimal value of f is $f(x^*) = 0$ and $f(x) = 0$ if and only if $d(x, C_1) = d(x, C_2) = 0$. Since C_1 and C_2 are closed, $d(x, C_1) = 0$ implies $x \in C_1$ and $d(x, C_2) = 0$ implies $x \in C_2$. Thus, we conclude that f is minimized at x if and only if $x \in C_1 \cap C_2$.

(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

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

Solution: Let $s(x) \in \cup_{i \in I(x)} \partial f_i(x)$, then there is $i^* \in I(x)$ such that $s(x) \in \partial f_{i^*}(x)$. Thus, for all $y \in \mathbb{R}^n$,

$$f(y) \geqslant f_{i^*}(y) \geqslant f_{i^*}(x) + \langle s(x), y - x\rangle = f(x) + \langle s(x), y - x\rangle.$$

This shows that $s(x) \in \partial f(x)$ and hence the result follows.

(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.

Solution: If $f(x_k) > 0$, then for $i \in I(x_k)$, $x_k \notin C_i$ and hence $\|x_k - \Pi_{C_i}(x_k)\|_2 > 0$. By characterization of projection, for all $y \in \mathbb{R}^n$,

$$\langle x_k - \Pi_{C_i}(x_k), \Pi_{C_i}(y) - \Pi_{C_i}(x_k)\rangle \leqslant 0.$$

Combined with Cauchy-Schwarz, we have

$$\begin{aligned}
\|x_k - \Pi_{C_i}(x_k)\| \cdot \|y - \Pi_{C_i}(y)\| &\geqslant \langle x_k - \Pi_{C_i}(x_k), y - \Pi_{C_i}(y)\rangle \\
&\qquad + \langle x_k - \Pi_{C_i}(x_k), \Pi_{C_i}(y) - \Pi_{C_i}(x_k)\rangle \\
&= \langle x_k - \Pi_{C_i}(x_k), y - \Pi_{C_i}(x_k)\rangle \\
&= \langle x_k - \Pi_{C_i}(x_k), x_k - \Pi_{C_i}(x_k) + y - x_k\rangle \\
&= \|x_k - \Pi_{C_i}(x_k)\|^2 + \langle x_k - \Pi_{C_i}(x_k), y - x_k\rangle.
\end{aligned}$$

Thus, we conclude that for all $y \in \mathbb{R}^n$,

$$\|y - \Pi_{C_i}(y)\| \geqslant \|x_k - \Pi_{C_i}(x_k)\| + \left\langle \frac{x_k - \Pi_{C_i}(x_k)}{\|x_k - \Pi_{C_i}(x_k)\|_2}, y - x_k \right\rangle,$$

which shows $s_k \in \partial f_i(x_k)$ because $f_i(x) = d(x, C_i) = \|x - \Pi_{C_i}(x)\|$. Applying the result in part (b), we have $s_k \in \partial f(x_k)$.

Page 3

(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

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

Solution: Substitute $f(x^*) = 0$ and $\|s_k\| = 1$ into the update rule of subgradient method with Polyak's step size, and we obtain

$$x_{k+1} = x_k - f(x_k)\frac{x_k - \Pi_{C_i}(x_k)}{\|x_k - \Pi_{C_i}(x_k)\|_2} = \Pi_{C_i}(x_k)$$

where $i \in I(x_k)$. Given any x_0, let $x_1 = \Pi_{C_1}(x_0)$ or $x_1 = \Pi_{C_2}(x_0)$ for $k \in \mathbb{N}$. WLOG, we let $x_1 = \Pi_{C_1}(x_0)$. Then $x_1 \in C_1$ and $f_1(x_1) = 0$, so $0 \in I(x_1)$. This means we shall let $x_2 = \Pi_{C_2}(x_1)$. Similarly, $x_3 = \Pi_{C_1}(x_2)$, $x_4 = \Pi_{C_2}(x_3)$ and etc. Thus, starting with the initial point, we alternatively project x_k to C_1 or C_2 to obtain x_{k+1}.

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

Solution: Take any $\bar{x} \in C_1 \cap C_2$. Since $x_{2k+1} = \Pi_{C_1}(x_{2k})$, by characterization of projection,

$$\|x_{2k+1} - \bar{x}\|_2^2 \le \|x_{2k} - \bar{x}\|_2^2 - \|x_{2k+1} - x_{2k}\|_2^2$$

Similarly, by considering $x_{2k+2} = \Pi_{C_2}(x_{2k+1})$, we obtain

$$\|x_{2k+2} - \bar{x}\|_2^2 \le \|x_{2k+1} - \bar{x}\|_2^2 - \|x_{2k+2} - x_{2k+1}\|_2^2$$

From the above two inequalities we can see the sequence $(\|x_k - \bar{x}\|)_{k\in\mathbb{N}}$ is nonnegative and decreasing, hence converges. Thus, $(x_k)_{k\in\mathbb{N}}$ is bounded and there is a subsequence x_{k_j} of x_k converges to x^*. By the convergence result of subgradient method in the lecture, we know $f(x_k) \to 0$. By continuity, $f(x^*) = 0$ and from part (a), we know $x^* \in C_1 \cap C_2$. If we choose $\bar{x} := x^*$, then it is easy to see $(\|x_k - x^*\|)_{k\in\mathbb{N}}$ converges to 0.

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?

Solution: No. ∇f is not globally Lipschitz on $\mathbb{R}^{n\times r}$. A simple counterexample is obtained by taking $M = 0$ and $n = r = 1$, $X = x \in \mathbb{R}$. Then $f(x) = \frac{1}{2}x^4$, where $f'(x) = 2x^3$ and $f''(x) = 6x^2$, which is unbounded as $|x| \to \infty$. Hence f' is not Lipschitz on $\mathbb{R}$, and therefore f is not ℓ-smooth globally.

Page 4

(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 $X, U \in \mathbb{R}^{n\times r}$,

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

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

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

Solution: For h, first compute $\nabla h(X) = (6\|X\|_F^2 + 2\|M\|_F)X$. Then compute

$$\begin{aligned}\nabla^2 h(X)[U,U] &= 6\left(\|X\|_F^2\|U\|_F^2 + 2\langle X, U\rangle^2\right) + 2\|M\|_F\|U\|_F^2 \\ &\ge 6\|X\|_F^2\|U\|_F^2 + 2\|M\|_F\|U\|_F^2.\end{aligned}$$

Similarly, for f, first compute $\nabla f(X) = 2(XX^\top - M)X$. Then compute

$$\nabla^2 f(X)[U,U] = \|UX^\top + XU^\top\|_F^2 + 2\langle XX^\top - M, UU^\top\rangle.$$

Using Cauchy-Schwarz, we have

$$\nabla^2 f(X)[U,U] \;\le\; 6\|X\|_F^2\|U\|_F^2 + 2\|M\|_F\|U\|_F^2.$$

Thus, we conclude

$$\nabla^2 h(X)[U,U] \ge 6\|X\|_F^2\|U\|_F^2 + 2\|M\|_F\|U\|_F^2 \;\ge\; \nabla^2 f(X)[U,U],$$

so f is 1-smooth relative to h.

(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 $z^2(z - 2\|M\|_F) = c$.

Solution: Using the definition of the Bregman divergence and removing constants independent of U, we have

$$T_\lambda(X) = \arg\min_{U\geq 0} \Big\{ h(U) - \langle H, U\rangle \Big\}, \quad H := \nabla h(X) - \lambda \nabla f(X). \tag{1}$$

Using the formula of ∇f and ∇h in part (b),

$$H = (6\|X\|_F^2 + 2\|M\|_F)X - 2\lambda(XX^\top - M)X.$$

Let $Z := [H]_+$ (entrywise), and note that for any $U \geq 0$,

$$\langle H, U\rangle = \langle Z, U\rangle + \langle H - Z, U\rangle \leq \langle Z, U\rangle,$$

because $H - Z \leq 0$ entrywise and $U \geq 0$. Hence

$$h(U) - \langle H, U\rangle \ \geq \ h(U) - \langle Z, U\rangle,$$

and equality holds whenever $U_{ij} = 0$ on the set where $H_{ij} < 0$. Define the index set $S := \{(i,j) : H_{ij} < 0\}$. It is easy to see that if U^* is a minimizer of (1) or (2), then $U^*_{ij} = 0$ for $(i,j) \in S$ and $\langle H, U^*\rangle = \langle Z, U^*\rangle$. Thus, a minimizer of (1) is also a minimizer of

$$\min_{U\geq 0} \{h(U) - \langle Z, U\rangle\}. \tag{2}$$

For any $U \geq 0$ with fixed norm $\|U\|_F = \rho$, Cauchy–Schwarz gives $\langle Z, U\rangle \leq \|Z\|_F\|U\|_F = \rho\|Z\|_F$, with equality if and only if U is a nonnegative scalar multiple of Z. Thus an optimizer of (2) has the form $U^\star = aZ$ for some scalar $a \geq 0$. Plugging $U = aZ$ into the objective and writing $z := \|Z\|_F$ gives the 1-D problem

$$\min_{a\geq 0} \ \Big\{ \tfrac{3}{2}\|aZ\|_F^4 + \|M\|_F\|aZ\|_F^2 - \langle Z, aZ\rangle \Big\} = \min_{a\geq 0} \ \Big\{ \tfrac{3}{2}a^4z^4 + \|M\|_F a^2z^2 - az^2 \Big\}.$$

Differentiate with respect to a and set to zero:

$$0 = 6a^3z^4 + 2\|M\|_F az^2 - z^2.$$

If $z = \|Z\|_F = 0$ then $Z = 0$ and the minimizer is $U^\star = 0$, consistent with the formula. Otherwise, divide by $z^2 > 0$ to obtain

$$6z^2a^3 + 2\|M\|_F a - 1 = 0. \tag{3}$$

Now define $\tau := 1/a$. Multiplying (3) by τ^3 yields

$$6z^2 + 2\|M\|_F\tau^2 - \tau^3 = 0 \iff \tau^3 - 2\|M\|_F\tau^2 = 6z^2 \iff \tau^2(\tau - 2\|M\|_F) = 6\|Z\|_F^2.$$

Thus τ is the unique real solution of $\tau^2(\tau - 2\|M\|_F) = c$ with $c = 6\|Z\|_F^2$. Finally, $U^\star = aZ = \frac{1}{\tau}Z$, hence

$$T_\lambda(X) = U^\star = \frac{1}{\tau(6\|[H]_+\|_F^2)}[H]_+,$$

as claimed. Uniqueness follows since the 1D objective in a is strictly convex and $\tau \mapsto \tau^2(\tau - 2\|M\|_F)$ is strictly increasing on $(2\|M\|_F, \infty)$.

(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

$\min_{1\le k\le K} \operatorname{dist}(0, \partial\Psi(x_k)) = O(1/\sqrt{K})$.

Solution: Fix X, Y and set $\Delta := Y - X$ and $\gamma(t) := X + t\Delta$ for $t \in [0,1]$. Define $\phi(t) := f(\gamma(t))$ and $\psi(t) := h(\gamma(t))$. Then $\phi''(t) = \nabla^2 f(\gamma(t))[\Delta,\Delta]$ and $\psi''(t) = \nabla^2 h(\gamma(t))[\Delta,\Delta]$. Since f is 1-smooth relative to h, we have $\phi''(t) \le \psi''(t)$ for all $t \in [0,1]$. Using the integral remainder formula for C^2 functions,

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 (1-t)\phi''(t)\,dt,\ D_h(Y,X) = \psi(1) - \psi(0) - \psi'(0) = \int_0^1 (1-t)\psi''(t)\,dt,$$

we get

$$f(Y) - f(X) - \langle \nabla f(X), Y - X\rangle = \int_0^1 (1-t)\phi''(t)\,dt \le \int_0^1 (1-t)\psi''(t)\,dt = D_h(Y,X). \quad (4)$$

Optimality condition of the update rule with $U = X_k$ and $\lambda = 1/2$ yields

$$g(X_{k+1}) + \langle \nabla f(X_k), X_{k+1} - X_k\rangle + 2D_h(X_{k+1}, X_k) \le g(X_k).$$

Apply (4) with $(X,Y) = (X_k, X_{k+1})$ and add $g(X_{k+1})$:

$$\Psi(X_{k+1}) \le \Psi(X_k) + \langle \nabla f(X_k), X_{k+1} - X_k\rangle + D_h(X_{k+1}, X_k) + g(X_{k+1}).$$

Substitute the previous bound to eliminate the inner product and $g(X_{k+1})$,

$$\Psi(X_{k+1}) \le \Psi(X_k) - D_h(X_{k+1}, X_k).$$

Thus, $f(X_k) \le f(X_0)$ for all k. Using $\|M - XX^\top\|_F \ge |\|XX^\top\|_F - \|M\|_F|$, we get

$$f(X) \ge \frac{1}{2}(\|XX^\top\|_F - \|M\|_F)^2 \implies \|X_k X_k^\top\|_F \le \|M\|_F + \sqrt{2f(X_0)}.$$

Finally $\|X\|_F^2 \le \sqrt{r}\,\|X^\top X\|_F = \sqrt{r}\,\|XX^\top\|_F$, hence the boundedness follows:

$$\|X_k\|_F \le r^{1/4}\left(\|M\|_F + \sqrt{2f(X_0)}\right)^{1/2} =: R.$$

Define $G_{1/2}(X_k) := 2(\nabla h(X_k) - \nabla h(X_{k+1}))$. On the ball $\|X\| \le R$, ∇h is Lipschitz with some constant M_R, so

$$D_h(X_{k+1}, X_k) \ge \frac{1}{2M_R}\|\nabla h(X_{k+1}) - \nabla h(X_k)\|_F^2 = \frac{\lambda^2}{2M_R}\|G_\lambda(X_k)\|_F^2. \quad (5)$$

With $\lambda = \frac{1}{2}$, combine (5) with the descent inequality and telescope:

$$\sum_{k=0}^{K-1} \|G_{1/2}(X_k)\|_F^2 \le 8M_R(\Psi(X_0) - \Psi_{\inf}) \implies \min_{0\le k\le K-1} \|G_{1/2}(X_k)\|_F = O(K^{-1/2}).$$

Finally, the optimality condition of the Bregman prox step gives

$$0 \in \nabla f(X_k) - G_{1/2}(X_k) + \partial\iota_{\mathbb{R}_+^{n\times r}}(X_{k+1}),$$

so there exists $s_{k+1} \in \partial\iota_{\mathbb{R}_+^{n\times r}}(X_{k+1})$ with $\nabla f(X_k) - G_{1/2}(X_k) + s_{k+1} = 0$. Thus $v_{k+1} := \nabla f(X_{k+1}) + s_{k+1} \in \partial\Psi(X_{k+1})$ and

$$\operatorname{dist}(0, \partial\Psi(X_{k+1})) \le \|v_{k+1}\|_F \le \|\nabla f(X_{k+1}) - \nabla f(X_k)\|_F + \|G_{1/2}(X_k)\|_F.$$

On $\|X\| \le R$, ∇f is Lipschitz and h is strongly convex, so $\|\nabla f(X_{k+1}) - \nabla f(X_k)\|_F \le C_R\|G_\lambda(X_k)\|_F$ for some constant C_R. Therefore

$$\operatorname{dist}(0, \partial\Psi(X_{k+1})) \le (1 + C_R)\|G_\lambda(X_k)\|_F.$$

Thus, we have

$$\min_{1\le k\le K} \operatorname{dist}(0, \partial\Psi(X_k)) = O(K^{-1/2}).$$

Page 7

(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 $X_0 \geq 0$ 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 8