Name: __________ Student ID: __________ Scores: __________

AIE6003 Homework 1

Instructor: Xiaopeng Li **Deadline**: Feb 6

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*. (15 points) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex ℓ-smooth function. Assume an optimal solution $x^* \in \arg\min f$ exists and denote $f^* := f(x^*)$. Consider the following algorithm with initial points $x_{-1} = x_0$:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k(x_k - x_{k-1}), \quad k \in \mathbb{N},$$

Take $\beta_k = \frac{k}{k+2}$ and $\alpha_k = \frac{\alpha_0}{k+2}$ for $k \in \mathbb{N}$, where $\alpha_0 \in (0, 1/\ell]$.

(a) (3 points) Let $p_k = k(x_k - x_{k-1})$ for all $k \in \mathbb{N}$. Prove that

$$x_{k+1} + p_{k+1} = x_k + p_k - \alpha_0 \nabla f(x_k).$$

Solution: Using $p_{k+1} = (k+1)(x_{k+1} - x_k)$, formula of β_k, α_k, and algorithm update, we have

$$\begin{aligned} x_{k+1} + p_{k+1} &= (k+2)x_{k+1} - (k+1)x_k \\ &= x_k - \alpha_0 \nabla f(x_k) + k(x_k - x_{k-1}) \\ &= x_k + p_k - \alpha_0 \nabla f(x_k). \end{aligned}$$

(b) (10 points) For all $k \in \mathbb{N}$, show that the sequence

$$\left(2\alpha_0 k(f(x_{k-1}) - f^*) + \|x_k + p_k - x^*\|^2\right)_{k \in \mathbb{N}}$$

is decreasing.

Solution: From (a), we obtain

$$\begin{aligned} \|x_{k+1} + p_{k+1} - x^*\|^2 &= \|x_k + p_k - x^*\|^2 + \alpha_0^2 \|\nabla f(x_k)\|^2 - 2\alpha_0 \langle x_k + p_k - x^*, \nabla f(x_k)\rangle \\ &= \|x_k + p_k - x^*\|^2 + \alpha_0^2 \|\nabla f(x_k)\|^2 \\ &\quad - 2\alpha_0 \langle x_k - x^*, \nabla f(x_k)\rangle - 2k\alpha_0 \langle x_k - x_{k-1}, \nabla f(x_k)\rangle \\ &\le \|x_k + p_k - x^*\|^2 + \alpha_0^2 \|\nabla f(x_k)\|^2 \\ &\quad - 2\alpha_0 \langle x_k - x^*, \nabla f(x_k)\rangle + 2k\alpha_0 (f(x_{k-1}) - f(x_k)) \\ &\le \|x_k + p_k - x^*\|^2 + \alpha_0^2 \|\nabla f(x_k)\|^2 \\ &\quad - 2\alpha_0 \left(f(x_k) - f(x^*) + \frac{1}{2\ell}\|\nabla f(x_k)\|^2 \right) + 2k\alpha_0 (f(x_{k-1}) - f(x_k)). \end{aligned}$$

Thus, we can conlude

$$2\alpha_0(k+1)(f(x_k) - f^*) + \|x_{k+1} + p_{k+1} - x^*\|^2 \le 2\alpha_0 k(f(x_{k-1}) - f^*) + \|x_k + p_k - x^*\|^2.$$

(c) (2 points) Conclude that for all $T \in \mathbb{N}$,

$$f(x_T) - f^* \le \frac{\|x_0 - x^*\|^2}{2\alpha_0(T+1)}.$$

Solution: From (b), we have

$$2\alpha_0(k+1)(f(x_k) - f^*) + \|x_{k+1} + p_{k+1} - x^*\|^2 \le \|x_0 - x^*\|^2.$$

This implies the desired result.

2**. (15 points) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and ℓ-smooth. Assume an optimal solution $x^\star \in \arg\min f$ exists and denote $f^\star := f(x^\star)$. Consider gradient descent with backtrack line search, with parameters $s > 0$, $\beta \in (0,1)$, $\sigma \in (0,1)$: given x^0, for $k = 0, 1, 2, \dots$ set $g_k := \nabla f(x^k)$ and $d_k := -g_k$. Choose the smallest integer $m_k \ge 0$ such that, with $\alpha_k := s\beta^{m_k}$,

$$f(x^k + \alpha_k d_k) \le f(x^k) + \sigma\alpha_k g_k^\top d_k, \tag{1}$$

and then update $x^{k+1} := x^k + \alpha_k d_k$.

(a) (6 points) Prove that the backtracking procedure terminates for every k, and that the accepted step sizes satisfy the uniform lower bound

$$\alpha_k \ge \underline{\alpha} := \min\left\{s, \frac{2\beta(1-\sigma)}{\ell}\right\}, \quad \forall k \ge 0.$$

Solution: Fix k and write $x := x_k$, $g := g_k = \nabla f(x)$, $d := d_k = -g$. By the descent lemma,

$$f(x + \alpha d) \le f(x) + \alpha\nabla f(x)^\top d + \frac{\ell}{2}\alpha^2\|d\|^2 = f(x) - \alpha\|g\|^2 + \frac{\ell}{2}\alpha^2\|g\|^2.$$

We want (1), i.e.

$$f(x + \alpha d) \le f(x) + \sigma\alpha g^\top d = f(x) - \sigma\alpha\|g\|^2.$$

It suffices that

$$-\alpha\|g\|^2 + \frac{\ell}{2}\alpha^2\|g\|^2 \le -\sigma\alpha\|g\|^2,$$

equivalently,

$$\frac{\ell}{2}\alpha^2 \le (1-\sigma)\alpha \quad \iff \quad \alpha \le \bar{\alpha} := \frac{2(1-\sigma)}{\ell}.$$

Hence, for any $\alpha \in (0, \bar{\alpha}]$, the backtrack condition holds. Since the backtracking candidates are $s, s\beta, s\beta^2, \dots$ and $\beta \in (0,1)$, there exists m large enough so that $s\beta^m \le \bar{\alpha}$, so the procedure terminates.

For the lower bound: if $s \le \bar{\alpha}$, then $m_k = 0$ and $\alpha_k = s \ge \min\{s, \beta\bar{\alpha}\}$. If $s > \bar{\alpha}$, then minimality of m_k implies $\alpha_k = s\beta^{m_k} \le \bar{\alpha}$ but $s\beta^{m_k-1} > \bar{\alpha}$, so

$$\alpha_k = \beta(s\beta^{m_k-1}) > \beta\bar{\alpha} = \frac{2\beta(1-\sigma)}{\ell}.$$

Combining both cases gives

$$\alpha_k \ge \min\left\{s, \frac{2\beta(1-\sigma)}{\ell}\right\} =: \underline{\alpha}, \quad \forall k \ge 0.$$

Page 2

Finally, since $g_k^\top d_k = -\|g_k\|^2$, (1) yields

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) \le f(x_k) + \sigma\alpha_k g_k^\top d_k = f(x_k) - \sigma\alpha_k\|g_k\|^2 \le f(x_k),$$

so $\{f(x_k)\}$ is nonincreasing.

(b) (5 points) Prove that for every $k \ge 0$,

$$f(x^k) - f^\star \le \frac{1}{2\alpha_k}\left(\|x^k - x^\star\|^2 - \|x^{k+1} - x^\star\|^2\right) + \frac{1}{2\sigma}(f(x^k) - f(x^{k+1})).$$

Solution: By convexity, for any k,

$$f(x_k) - f^\star \le \nabla f(x_k)^\top(x_k - x^\star) = g_k^\top(x_k - x^\star). \tag{2}$$

Next, use the update $x_{k+1} = x_k - \alpha_k g_k$ to expand the squared distance:

$$\|x_{k+1} - x^\star\|^2 = \|x_k - \alpha_k g_k - x^\star\|^2 = \|x_k - x^\star\|^2 - 2\alpha_k g_k^\top(x_k - x^\star) + \alpha_k^2\|g_k\|^2.$$

Rearranging gives

$$g_k^\top(x_k - x^\star) = \frac{1}{2\alpha_k}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{\alpha_k}{2}\|g_k\|^2. \tag{3}$$

Moreover, (1) with $d_k = -g_k$ implies the sufficient decrease

$$f(x_{k+1}) \le f(x_k) - \sigma\alpha_k\|g_k\|^2 \quad \Longrightarrow \quad \alpha_k\|g_k\|^2 \le \frac{1}{\sigma}(f(x_k) - f(x_{k+1})).$$

Substitute this bound into (3) and then into (2) to obtain

$$f(x_k) - f^\star \le \frac{1}{2\alpha_k}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{1}{2\sigma}(f(x_k) - f(x_{k+1})),$$

as required.

(c) (4 points) Using parts (a)-(b), prove that there exists an explicit constant $c > 0$ such that

$$f(x^k) - f^\star \le \frac{c}{k} \quad \forall k \ge 1,$$

with

$$c = \frac{\|x^0 - x^\star\|^2}{2\underline{\alpha}} + \frac{f(x^0) - f^\star}{2\sigma}.$$

Solution: Let $e_k := f(x_k) - f^\star \ge 0$. From part (b) and $\alpha_k \ge \underline{\alpha}$ (part (a)),

$$e_k \le \frac{1}{2\underline{\alpha}}\left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{1}{2\sigma}(f(x_k) - f(x_{k+1})).$$

Page 3

Sum this inequality from $k = 0$ to $k = K - 1$:

$$\begin{aligned}
\sum_{k=0}^{K-1} e_k &\le \frac{1}{2\underline{\alpha}} \sum_{k=0}^{K-1} \left(\|x_k - x^\star\|^2 - \|x_{k+1} - x^\star\|^2\right) + \frac{1}{2\sigma} \sum_{k=0}^{K-1} (f(x_k) - f(x_{k+1})) \\
&= \frac{1}{2\underline{\alpha}} \left(\|x_0 - x^\star\|^2 - \|x_K - x^\star\|^2\right) + \frac{1}{2\sigma}(f(x_0) - f(x_K)) \\
&\le \frac{\|x_0 - x^\star\|^2}{2\underline{\alpha}} + \frac{f(x_0) - f^\star}{2\sigma} =: c.
\end{aligned}$$

Since part (a) showed $f(x_{k+1}) \le f(x_k)$, the sequence $\{e_k\}$ is nonincreasing, hence

$$e_K \le \frac{1}{K} \sum_{k=0}^{K-1} e_k \le \frac{c}{K}, \quad \forall K \ge 1.$$

Therefore,

$$f(x_K) - f^\star = e_K \le \frac{c}{K} \quad \forall K \ge 1,$$

with $c = \dfrac{\|x_0 - x^\star\|^2}{2\underline{\alpha}} + \dfrac{f(x_0) - f^\star}{2\sigma}$.

3***. (20 points) Let $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite (SPD) and consider

$$\min_{x\in\mathbb{R}^n} \quad f(x) := \frac{1}{2} x^\top A x - b^\top x,$$

whose unique minimizer is $x^\star = A^{-1}b$. Given $x_0 \in \mathbb{R}^n$, define the residual $r_k := b - Ax_k$, the error $e_k := x_k - x^\star$, and the Krylov subspace

$$\mathcal{K}_k(A, r_0) := \operatorname{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\}. \tag{4}$$

The Conjugate Gradient (CG) iterations (assume $r_k \ne 0$) are:

$$\begin{aligned}
&r_0 = b - Ax_0, \quad p_0 = r_0, \quad \alpha_k = \frac{r_k^\top r_k}{p_k^\top A p_k}, \quad x_{k+1} = x_k + \alpha_k p_k, \\
&r_{k+1} = r_k - \alpha_k A p_k, \quad \beta_{k+1} = \frac{r_{k+1}^\top r_{k+1}}{r_k^\top r_k}, \quad p_{k+1} = r_{k+1} + \beta_{k+1} p_k.
\end{aligned} \tag{5}$$

(a) (5 points) Let $\phi_k(\alpha) := f(x_k + \alpha p_k)$. Show that the minimizer of ϕ_k satisfies

$$\alpha_k = \frac{p_k^\top r_k}{p_k^\top A p_k}, \quad \text{and hence} \quad p_k^\top r_{k+1} = 0. \tag{6}$$

Solution: Fix k and define $\phi_k(\alpha) = f(x_k + \alpha p_k)$. Using the formula of f,

$$\phi_k(\alpha) = \tfrac{1}{2}(x_k + \alpha p_k)^\top A (x_k + \alpha p_k) - b^\top (x_k + \alpha p_k).$$

Differentiate:

$$\begin{aligned}
\phi_k'(\alpha) &= p_k^\top A(x_k + \alpha p_k) - b^\top p_k \\
&= p_k^\top (Ax_k - b) + \alpha p_k^\top A p_k \\
&= -p_k^\top r_k + \alpha p_k^\top A p_k.
\end{aligned}$$

Since $A \succ 0$ and $p_k \neq 0$, we have $p_k^\top A p_k > 0$, hence the unique minimizer is given by $\phi_k'(\alpha) = 0$:

$$\alpha_k = \frac{p_k^\top r_k}{p_k^\top A p_k},$$

which proves the first identity in (6). Next, using $r_{k+1} = b - Ax_{k+1}$ and $x_{k+1} = x_k + \alpha_k p_k$,

$$r_{k+1} = b - A(x_k + \alpha_k p_k) = r_k - \alpha_k A p_k.$$

Therefore

$$p_k^\top r_{k+1} = p_k^\top r_k - \alpha_k p_k^\top A p_k = p_k^\top r_k - \frac{p_k^\top r_k}{p_k^\top A p_k} p_k^\top A p_k = 0.$$

(b) (5 points) Prove that $x_k \in x_0 + \mathcal{K}_k(A, r_0)$ and $r_k \in \mathcal{K}_{k+1}(A, r_0)$ for all $k \geq 0$.

Solution: We first record two closure properties of Krylov subspaces:

$$\mathcal{K}_k(A, r_0) \subseteq \mathcal{K}_{k+1}(A, r_0), \quad A\mathcal{K}_k(A, r_0) \subseteq \mathcal{K}_{k+1}(A, r_0). \tag{7}$$

The first is immediate from (4). For the second, if $v = \sum_{j=0}^{k-1} c_j A^j r_0 \in \mathcal{K}_k$, then $Av = \sum_{j=0}^{k-1} c_j A^{j+1} r_0 \in \mathcal{K}_{k+1}$. We prove by induction that $r_k \in \mathcal{K}_{k+1}$ and $x_k \in x_0 + \mathcal{K}_k$.

- Base: $r_0 \in \mathcal{K}_1$ by definition; and $x_0 \in x_0 + \mathcal{K}_0$.
- Inductive step: assume $r_k \in \mathcal{K}_{k+1}$ and $x_k \in x_0 + \mathcal{K}_k$. We also show $p_k \in \mathcal{K}_{k+1}$ by induction: $p_0 = r_0 \in \mathcal{K}_1$; and if $r_k \in \mathcal{K}_{k+1}$ and $p_{k-1} \in \mathcal{K}_k$, then $p_k = r_k + \beta_k p_{k-1} \in \mathcal{K}_{k+1}$. Thus $x_{k+1} = x_k + \alpha_k p_k \in x_0 + \mathcal{K}_{k+1}$.

For the residual, using $r_{k+1} = r_k - \alpha_k A p_k$ and $p_k \in \mathcal{K}_{k+1}$, we have $Ap_k \in \mathcal{K}_{k+2}$ by (7). Also $r_k \in \mathcal{K}_{k+1} \subseteq \mathcal{K}_{k+2}$, hence $r_{k+1} \in \mathcal{K}_{k+2}$. This completes the induction.

(c) (5 points) Show that for all $k \geq 1$, $p_{k-1}^\top r_k = 0$. Conclude that for all $k \geq 0$, $p_k^\top r_k = r_k^\top r_k$, and therefore the line-search formula in (6) is consistent with the CG step size in (5).

Solution: For $k \geq 1$, apply part (a) with index $k-1$ to obtain $p_{k-1}^\top r_k = 0$. Next, using the recursion $p_k = r_k + \beta_k p_{k-1}$ (for $k \geq 1$), we have

$$p_k^\top r_k = r_k^\top r_k + \beta_k \, p_{k-1}^\top r_k = r_k^\top r_k.$$

(For $k = 0$, $p_0 = r_0$, so the identity holds trivially.)
Substituting $p_k^\top r_k = r_k^\top r_k$ into (6) yields

$$\alpha_k = \frac{p_k^\top r_k}{p_k^\top A p_k} = \frac{r_k^\top r_k}{p_k^\top A p_k},$$

which matches the formula in (5).

(d) (5 points) You may use the following standard CG property without proof: the search directions are A-conjugate, i.e., $p_i^\top A p_j = 0$, for all $i \neq j$. Using this fact together with parts (a)-(b), prove that for all $k \geq 1$, $r_k \perp \mathcal{K}_k(A, r_0)$, and conclude that $r_i^\top r_j = 0$ for all $i \neq j$.

Solution: We prove by induction that $p_i^\top r_k = 0$ for all $0 \le i \le k-1$. Base $k = 1$: by part (a), $p_0^\top r_1 = 0$. Inductive step: assume $p_i^\top r_k = 0$ for all $0 \le i \le k-1$. Using $r_{k+1} = r_k - \alpha_k A p_k$, for any $i \le k-1$,

$$p_i^\top r_{k+1} = p_i^\top r_k - \alpha_k p_i^\top A p_k = 0 - 0 = 0,$$

where we used the induction hypothesis and the given A-conjugacy. Also, part (a) gives $p_k^\top r_{k+1} = 0$. Hence $r_{k+1} \perp \text{span}\{p_0, \dots, p_k\}$.
By part (b), $p_i \in \mathcal{K}_{i+1}(A, r_0) \subseteq \mathcal{K}_k(A, r_0)$ for $i = 0, \dots, k-1$, so $\text{span}\{p_0, \dots, p_{k-1}\} \subseteq \mathcal{K}_k(A, r_0)$. The A-conjugacy implies $p_0, \dots, p_{k-1}$ are linearly independent, hence

$$\dim(\text{span}\{p_0, \dots, p_{k-1}\}) = k.$$

Since $\mathcal{K}_k(A, r_0)$ is spanned by k vectors, $\dim(\mathcal{K}_k(A, r_0)) \le k$, so

$$\mathcal{K}_k(A, r_0) = \text{span}\{p_0, \dots, p_{k-1}\}.$$

Thus, we obtain $r_k \perp \mathcal{K}_k(A, r_0)$.
Fix $i < j$. By part (b), $r_i \in \mathcal{K}_{i+1}(A, r_0) \subseteq \mathcal{K}_j(A, r_0)$. Since $r_j \perp \mathcal{K}_j(A, r_0)$, we get $r_j^\top r_i = 0$. By symmetry, $r_i^\top r_j = 0$ for all $i \ne j$.

Page 6