Accelerated Gradient Methods (Y. Nesterov)

Renato D.C. Monteiro

February 2, 2024

1 Convex Case

1.1 The problem and assumptions

Consider the composite problem

$$\begin{aligned} \phi_* = \min\ & \phi(x) = f(x) + h(x) \\ \text{s.t. } & x \in \mathbb{R}^n \end{aligned}$$

where:

1) $h \in \overline{\text{Conv}}\,(\mathbb{R}^n)$
2) f is convex on $\text{dom}\, h$ (and hence $\text{dom}\, f \supset \text{dom}\, h$);
3) there exists $L > 0$ such that f is L-smooth on $\text{dom}\, h$
4) the set X_* of optimal solutions is nonempty (and hence $\phi_* \in \mathbb{R}$).

Motivation Let us now motivate the class of methods we are interested in studying in this section. Given $A \geq 0$ and $x, y \in \mathbb{R}^n$, we want to find $A^+ > A$ and $x^+, y^+ \in \mathbb{R}^n$ such that

$$\begin{aligned}\eta^+(u) &:= A^+ \left[\phi(y^+) - \phi(u)\right] + \frac{1}{2}\|u - x^+\|^2 \\ &\leq A\left[\phi(y) - \phi(u)\right] + \frac{1}{2}\|u - x\|^2 =: \eta(u) \quad \forall u \in \mathbb{R}^n \quad (1)\end{aligned}$$

Using the above construction, we can then generate a sequence $\{(x_k, y_k, A_k)\} \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+$ such that the quantity

$$\eta_k(u) := A_k\left[\phi(y_k) - \phi(u)\right] + \frac{1}{2}\|u - x_k\|^2$$

satisfies

$$\eta_{k+1}(u) \leq \eta_k(u) \quad \forall k \geq 0,\ \forall u \in \mathbb{R}^n.$$

As a consequence, we have

$$\eta_k(u) \leq \eta_0(u) \quad \forall k \geq 0,\ \forall u \in \mathbb{R}^n,$$

and thus

$$\phi(y_k) - \phi(u) \leq \frac{A_0}{A_k}\left[\phi(y_0) - \phi(u)\right] + \frac{1}{2A_k}\|u - x_0\|^2 \quad \forall k \geq 0,\ \forall u \in \mathbb{R}^n.$$

If A_0 is chosen to be zero, then

$$\phi(y_k) - \phi(u) \leq \frac{1}{2A_k}\|u - x_0\|^2 \quad \forall k \geq 0,\ u \in \mathbb{R}^n.$$

In particular, taking $u = \mathrm{Pr}_{X_*}(x_0)$, we conclude that

$$\phi(y_k) - \phi_* \leq \frac{d_0^2}{2A_k} \quad \forall k \geq 0,$$

where $d_0 := \min\{\|x_* - x_0\| : x_* \in X_*\}$.

Hence, the larger A_k grows, the faster the primal gap $\phi(y_k) - \phi_*$ approaches zero.

Let us now give a sufficient condition for (1) to hold.

Lemma 1.1. *Assume that* $y^+ \in \mathbb{R}^n$, $A^+ > A$, *and* $\gamma \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$, *satisfy*

$$\gamma \leq \phi \tag{2}$$

$$A^+\phi(y^+) \leq A\phi(y) + \min\left\{a\gamma(u) + \frac{1}{2}\|u - x\|^2\right\} \tag{3}$$

where $a := A^+ - A$, *and define*

$$x^+ = \mathrm{argmin}\left\{a\gamma(u) + \frac{1}{2}\|u - x\|^2\right\}. \tag{4}$$

Then, (x^+, y^+, A^+) *satisfies* (1).

Proof: Let

$$\theta := \min\left\{a\gamma(u) + \frac{1}{2}\|u - x\|^2\right\}$$

Since x^+ is the optimal solution of the above problem, we have that for every $u \in \mathbb{R}^n$,

$$\begin{aligned} A\phi(y) + a\gamma(u) + \frac{1}{2}\|u - x\|^2 &\geq A\phi(y) + \theta + \frac{1}{2}\|u - x^+\|^2 \\ &\geq A^+\phi(y^+) + \frac{1}{2}\|u - x^+\|^2. \end{aligned}$$

The above inequality then implies that

$$\begin{aligned} & A^+[\phi(y^+) - \phi(u)] + \frac{1}{2}\|u - x^+\|^2 \\ (\text{since } a = A^+ - A) \quad & \leq A[\phi(y) - \phi(u)] + a[\gamma(u) - \phi(u)] + \frac{1}{2}\|u - x\|^2 \\ (\text{since } a > 0 \text{ and } \gamma \leq \phi) \quad & = A[\phi(y) - \phi(u)] + \frac{1}{2}\|u - x\|^2 \end{aligned}$$

and hence that the conclusion of the lemma holds. ∎

From now on, we will focus on:

Goal: Given (x, y, A), find y^+, $A^+ > A$ and $\gamma \leq \phi$ satisfying condition (3).

Lemma 1.2. *Let $a > 0$ and $\gamma \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ such that $\gamma \leq \phi$ be given and define*

$$\tilde{y} = \frac{Ay + ax^+}{A + a}, \quad \tilde{x} = \frac{Ay + ax}{A + a}$$

where x^+ is as in (4). Then

$$A\phi(y) + \min\left\{a\gamma(u) + \frac{1}{2}\|u - x\|^2\right\} \geq (A + a)\left[\gamma(\tilde{y}) + \frac{A + a}{2a^2}\|\tilde{y} - \tilde{x}\|^2\right].$$

Proof: First observe that the definitions of $\tilde{x}$ and $\tilde{y}$ imply that

$$\|\tilde{y} - \tilde{x}\| = \frac{a}{A + a}\|x^+ - x\|$$

For every $a > 0$, have

$$\begin{aligned}
A\phi(y) + \min\left\{a\gamma(u) + \frac{1}{2}\|u - x\|^2\right\} & \\
\text{(definition of } x^+ \text{ in (4))} \quad &= A\phi(y) + a\gamma(x^+) + \frac{1}{2}\|x^+ - x\|^2 \\
\text{(since } \gamma \leq \phi) \quad &\geq A\gamma(y) + a\gamma(x^+) + \frac{1}{2}\|x^+ - x\|^2 \\
\text{(convexity of } \gamma) \quad &\geq (A + a)\,\gamma\left(\frac{Ay + ax^+}{A + a}\right) + \frac{1}{2}\|x^+ - x\|^2 \\
\text{(definition of } \tilde{y}) \quad &= (A + a)\,\gamma(\tilde{y}) + \frac{1}{2}\|x^+ - x\|^2
\end{aligned}$$

The conclusion of the lemma now follows by combining the above two relations.

Corollary 1.3. *In addition to the assumptions of Lemma 1.2, assume also that $y^+ \in \mathbb{R}^n$ is a point such that*

$$\phi(y^+) \leq \gamma(\tilde{y}) + \frac{A + a}{2a^2}\|\tilde{y} - \tilde{x}\|^2 \tag{5}$$

and set $A^+ = A + a$. Then, (y^+, A^+, γ) satisfy (3), and hence (x^+, y^+, A^+) satisfies (1).

Proof: The first conclusion follows immediately from Lemma 1.2 while the second one follows from Lemma 1.1.

1.2 First ACG variant (Atouch and Teboule)

This variant chooses $a > 0$ such that

$$\frac{A+a}{a^2} = L$$

and sets

$$y^+ = \tilde{y}, \quad \gamma(\cdot) = \ell_f(\cdot; \tilde{x}) + h(\cdot)$$

Then,

$$\begin{aligned}
\gamma(\tilde{y}) + \frac{A+a}{2a^2}\|\tilde{y} - \tilde{x}\|^2 &= \gamma(\tilde{y}) + \frac{L}{2}\|\tilde{y} - \tilde{x}\|^2 \\
&= \gamma(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2 \\
&= \ell_f(y^+; \tilde{x}) + h(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2 \\
&\geq (f+h)(y^+) = \phi(y^+)
\end{aligned}$$

Remark: Instead of setting $y^+ = \tilde{y}$, we can instead choose y^+ such that

$$\gamma(\tilde{y}) + \frac{L}{2}\|\tilde{y} - \tilde{x}\|^2 \geq \gamma(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2$$

e.g., y^+ given by

$$y^+ = \operatorname{argmin} \left\{ \ell_f(u; \tilde{x}) + h(u) + \frac{L}{2}\|u - \tilde{x}\|^2 \right\} \tag{6}$$

The latter variant is more expensive since it solves two subproblems per iteration, namely, subproblems (4) and (6).

Algorithm 1 (AT-ACG variant)

0. Let $x_0 \in \operatorname{dom} h$ be given, and set $y_0 = x_0$, $A_0 = 0$, and $k = 0$;

1. Compute

$$a_k = \frac{1 + \sqrt{1 + 4LA_k}}{2L}, \quad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_k + a_k}; \tag{7}$$

2. Compute

$$x_{k+1} := \operatorname*{argmin}_{u \in \operatorname{dom} h} \left\{ a_k[\ell_f(u; \tilde{x}_k) + h(u)] + \frac{1}{2}\|u - x_k\|^2 \right\}, \tag{8}$$

$$y_{k+1} := \frac{A_k y_k + a_k x_{k+1}}{A_k + a_k}, \tag{9}$$

$$A_{k+1} = A_k + a_k; \tag{10}$$

3. Set $k \leftarrow k + 1$ and go to step 1.

Obs: Formula for a_k and A_{k+1} in (7) imply that

$$\frac{A_{k+1}}{a_k^2} = \frac{A_k + a_k}{a_k^2} = L. \tag{11}$$

Question: How fast does A_k grow?

Have

$$a_k \geq \frac{1 + \sqrt{4LA_k}}{2L} \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}}$$

and so

$$A_{k+1} = A_k + a_k \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}} + A_k \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\frac{1}{L}} \right)^2$$

Hence

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}\sqrt{\frac{1}{L}}$$

This implies that

$$\sqrt{A_k} \geq \sqrt{A_0} + \frac{k}{2}\sqrt{\frac{1}{L}} = \frac{k}{2}\sqrt{\frac{1}{L}}$$

and hence that

$$A_k \geq \frac{k^2}{4L}$$

Hence, the convergence rate of the AT-variant is

$$\phi(y_k) - \phi_* \leq \frac{d_0^2}{2A_k} \leq \frac{2Ld_0^2}{k^2}$$

1.3 Second ACG variant (FISTA)

This variant was originally proposed by Nesterov for set optimization problems and later extended by Teboule et al to composite optimization problems.

While the AT variant solves a composite subproblem for x_{k+1} and obtains y_{k+1} in a straightforward manner, the FISTA variant below solves a composite subproblem for y_{k+1} and easily obtains x_{k+1}.

Algorithm 2 (FISTA)

0. Let $x_0 \in \operatorname{dom} h$ be given, and set $y_0 = x_0$, $A_0 = 0$, and $k = 0$;

1. Compute

$$a_k = \frac{1 + \sqrt{1 + 4LA_k}}{2L}, \quad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_k + a_k}; \tag{12}$$

2. Compute

$$y_{k+1} := \operatorname*{argmin}_{x \in \operatorname{dom} h} \left\{ \ell_f(x; \tilde{x}_k) + h(x) + \frac{L}{2}\|x - \tilde{x}_k\|^2 \right\}, \tag{13}$$

$$x_{k+1} = x_k + La_k(y_{k+1} - \tilde{x}_k), \tag{14}$$

$$A_{k+1} = A_k + a_k; \tag{15}$$

3. Set $k \leftarrow k + 1$ and go to step 1.

The justification of this variant relies on choosing

$$\gamma(\cdot) = \tilde{\gamma}(y^+) + L\langle \tilde{x} - y^+, \cdot - y^+ \rangle \tag{16}$$

where for this section we let

$$\tilde{\gamma}(\cdot) := \ell_f(\cdot; \tilde{x}_k) + h(\cdot)$$

Clearly,

$$\tilde{\gamma}(\cdot) + \frac{L}{2}\| \cdot - \tilde{x}\|^2 \geq \phi(\cdot)$$

and, by (13), have

$$y^+ = \operatorname{argmin} \left\{ \tilde{\gamma}(u) + \frac{L}{2}\|u - \tilde{x}\|^2 \right\}$$

Lemma 1.4. *Affine function $\gamma(\cdot)$ defined in* (16) *satisfies*

$$\gamma(y^+) = \tilde{\gamma}(y^+), \quad \gamma(\cdot) \leq \tilde{\gamma}(\cdot), \quad y^+ = \operatorname{argmin} \left\{ \gamma(u) + \frac{L}{2}\|u - \tilde{x}\|^2 \right\}$$

Proof: Exercise

Using the above lemma, let us verify that (5) holds. Indeed,

$$\begin{aligned}\gamma(\tilde{y})+\frac{A+a}{2a^2}\|\tilde{y}-\tilde{x}\|^2 &= \gamma(\tilde{y})+\frac{L}{2}\|\tilde{y}-\tilde{x}\|^2 \\ &\geq \min\left\{\gamma(u)+\frac{L}{2}\|u-\tilde{x}\|^2\right\} = \gamma(y^+)+\frac{L}{2}\|y^+-\tilde{x}\|^2 \\ &= \tilde{\gamma}(y^+)+\frac{L}{2}\|y^+-\tilde{x}\|^2 \geq \phi(y^+)\end{aligned}$$

showing that (5) holds.

Moreover, by (14), we have

$$x^+ = x + aL(y^+ - \tilde{x})$$

Hence, the definition of the affine function γ implies that

$$x^+ = x - a\nabla\gamma$$

which is the optimality condition for x^+ to be optimal for the subproblem in (4). We have thus shown that x^+ in (14) satisfies (4).

Alternative description of FISTA

Lemma 1.5. *There holds:*

$$x^+ = \frac{A^+ y^+ - Ay}{a}.$$

Proof: Have

$$\begin{aligned} x^+ &= x + aL(y^+ - \tilde{x}) = x + aL\left(y^+ - \frac{Ay + ax}{A+a}\right) \\ &= x + \frac{A+a}{a}\left(y^+ - \frac{Ay+ax}{A+a}\right) = \frac{A^+ y^+ - Ay}{a}. \end{aligned}$$

■

Lemma 1.6. *There holds:*

$$\tilde{x}^+ = y^+ + \frac{1}{t^+}(t-1)(y^+ - y)$$

where

$$t = \frac{A^+}{a} = aL \geq 1.$$

Proof: Have

$$\begin{aligned} \tilde{x}^+ &= \frac{A^+}{A^{++}} y^+ + \frac{a^+}{A^{++}} x^+ \\ &= \left(y^+ - \frac{a^+}{A^{++}} y^+\right) + \frac{a^+}{A^{++}}\left(\frac{A^+ y^+ - Ay}{a}\right) \\ &= y^+ + \frac{a^+}{A^{++}}\left(\frac{A^+}{a} - 1\right)(y^+ - y) \\ &= y^+ + \frac{1}{t^+}(t-1)(y^+ - y). \end{aligned}$$

■

Lemma 1.7. *There holds:*

$$(t^+)^2 - t^+ - t^2 = 0,$$

and hence

$$t^+ = \frac{1 + \sqrt{1 + 4t^2}}{2}.$$

Proof:

$$\begin{aligned} (t^+)^2 - t^+ = t^+(t^+ - 1) &= \frac{A^{++}}{a^+}\left(\frac{A^{++}}{a^+} - 1\right) = \frac{A^{++}}{a^+}\frac{A^+}{a^+} \\ &= LA^+ = (La)\frac{A^+}{a} = t^2. \end{aligned}$$

Algorithm 3 (FISTA-revisted)

0. Let $y_0 \in \operatorname{dom} h$ be given, and set $\tilde{x}_0 = y_0$, $t_0 = 1$, and $k = 0$;

1. Compute

$$y_{k+1} := \operatorname*{argmin}_{x \in \operatorname{dom} h} \left\{ \ell_f(x; \tilde{x}_k) + h(x) + \frac{L}{2}\|x - \tilde{x}_k\|^2 \right\}, \quad (17)$$

and then set

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$\tilde{x}_{k+1} = y_{k+1} + \frac{t_k - 1}{t_{k+1}}(y_{k+1} - y_k)$$

3. Set $k \leftarrow k + 1$ and go to step 1.

Remarks:

- Extrapolated composite gradient method
- Somewhat related to the heavy ball method (due to Polyak) where y_{k+1} is computed as in (17) and

$$\tilde{x}_{k+1} = y_{k+1} + \beta_k(\tilde{x}_k - \tilde{x}_{k-1})$$

for some scalar $\beta_k > 0$.

2 Strongly convex case

2.1 The problem and assumptions

Consider the composite problem

$$(P) \qquad \begin{array}{c} \phi_* = \min \ \phi(x) = f(x) + h(x) \\ \text{s.t. } x \in \mathbb{R}^n \end{array}$$

where:

1) $h \in \overline{\text{Conv}}_\mu(\mathbb{R}^n)$ for some $\mu \geq 0$;

2) $f : \mathbb{R}^n \to \mathbb{R}$ is convex everywhere;

3) there exists $L > 0$ such that f is L-smooth everywhere;

4) the set X_* of optimal solutions of (P) is nonempty (and hence $\phi_* \in \mathbb{R}$).

Motivation: Given $A \geq 0$ and $\tau > 0$, and $x, y \in \mathbb{R}^n$, suppose we can always find $A^+ > A$, $\tau^+ > \tau$, and $x^+, y^+ \in \mathbb{R}^n$ such that

$$\begin{aligned}\eta^+(u) &:= A^+ \left[\phi(y^+) - \phi(u)\right] + \frac{\tau^+}{2}\|u - x^+\|^2 \\ &\leq A\left[\phi(y) - \phi(u)\right] + \frac{\tau}{2}\|u - x\|^2 =: \eta(u) \quad \forall u \in \mathbb{R}^n \quad (18)\end{aligned}$$

Using the above construction, we can then generate an infinite sequence $\{(x_k, y_k, A_k, \tau_k)\} \subset \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_+$ such that the quantity

$$\eta_k(u) := A_k\left[\phi(y_k) - \phi(u)\right] + \frac{\tau_k}{2}\|u - x_k\|^2$$

satisfies

$$\eta_0(u) \geq \eta_k(u) \quad \forall k \geq 0,\ u \in \mathbb{R}^n.$$

Dividing this inequality by A_k yields

$$\begin{aligned}&\frac{A_0}{A_k}\left[\phi(y_0) - \phi(u)\right] + \frac{\tau_0}{2A_k}\|u - x_0\|^2 \\ &\quad \geq \phi(y_k) - \phi(u) + \frac{\tau_k}{2A_k}\|u - x_k\|^2 \geq \phi(y_k) - \phi(u)\end{aligned}$$

If A_0 and τ_0 are chosen to be zero and one, respectively, then

$$\phi(y_k) - \phi(u) \leq \frac{1}{2A_k}\|u - x_0\|^2 \quad \forall k \geq 0,\ u \in \mathbb{R}^n,$$

In particular, taking $u = \mathrm{Proj}_{X_*}(x_0)$, we conclude that

$$\phi(y_k) - \phi_* \leq \frac{d_0^2}{2A_k} \quad \forall k \geq 0$$

where $d_0 := \min\{\|x_* - x_0\| : x_* \in X_*\}$.

Hence, the larger A_k grows, the faster the primal gap $\phi(y_k) - \phi_*$ approaches zero.

Proposition 2.1. *Let* $(x, y, A, \tau) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_+ \times \mathbb{R}_{++}$ *be given and assume that* $(y^+, \gamma(\cdot), a) \in \mathbb{R}^n \times \overline{\mathrm{Conv}}_\mu(\mathbb{R}^n) \times \mathbb{R}_{++}$ *satisfies the KEY condition that*

$$\gamma \leq \phi, \qquad \phi(y^+) \leq \gamma(\tilde{y}) + \frac{\tau(A+a)}{2a^2}\|\tilde{y} - \tilde{x}\|^2 \tag{19}$$

where

$$\tilde{y} := \frac{Ay + ax^+}{A^+}, \quad \tilde{x} := \frac{Ay + ax}{A^+}$$

and

$$A^+ := A + a, \quad x^+ := \operatorname{argmin}\left\{a\gamma(u) + \frac{\tau}{2}\|u - x\|^2\right\}. \tag{20}$$

Then, the quadruple (x^+, y^+, A^+, τ^+) *where*

$$\tau_+ := \tau + a\mu \tag{21}$$

satisfies

$$\begin{aligned}\eta^+(u) &:= A^+\left[\phi(y^+) - \phi(u)\right] + \frac{\tau^+}{2}\|u - x^+\|^2 \\ &\leq A\left[\phi(y) - \phi(u)\right] + \frac{\tau}{2}\|u - x\|^2 =: \eta(u) \quad \forall u \in \mathbb{R}^n\end{aligned} \tag{22}$$

Proof: First observe that the definitions of $\tilde{x}$ and $\tilde{y}$ imply that

$$\|\tilde{y} - \tilde{x}\| = \frac{a}{A+a}\|x^+ - x\| \tag{23}$$

For every $a > 0$, have

$$\begin{aligned}& A\phi(y) + \min\left\{a\gamma(u) + \frac{\tau}{2}\|u - x\|^2\right\} \\ (\text{definition of } x^+) \quad &= A\phi(y) + a\gamma(x^+) + \frac{\tau}{2}\|x^+ - x\|^2 \\ (\text{since } \gamma \leq \phi) \quad &\geq A\gamma(y) + a\gamma(x^+) + \frac{\tau}{2}\|x^+ - x\|^2 \\ (\text{convexity of } \gamma) \quad &\geq (A+a)\,\gamma\left(\frac{Ay + ax^+}{A+a}\right) + \frac{\tau}{2}\|x^+ - x\|^2 \\ (\text{definition of } \tilde{y}) \quad &\geq (A+a)\,\gamma(\tilde{y}) + \frac{\tau}{2}\|x^+ - x\|^2 \\ (\text{equation (23)}) \quad &\geq (A+a)\left[\gamma(\tilde{y}) + \frac{\tau(A+a)}{2a^2}\|\tilde{y} - \tilde{x}\|^2\right] \\ (\text{equation (19)}) \quad &\geq (A+a)\phi(y^+) \\ (\text{def of } A^+) \quad &= A^+\phi(y^+)\end{aligned}$$

This inequality together with the assumption that $\gamma(\cdot) \in \overline{\mathrm{Conv}}_\mu(\mathbb{R}^n)$ yield

$$\begin{aligned}A\phi(y) + a\gamma(u) + \frac{\tau}{2}\|u - x\|^2 &\geq A^+\phi(y^+) + \frac{\tau + a\mu}{2}\|x^+ - u\|^2 \\ &= A^+\phi(y^+) + \frac{\tau^+}{2}\|x^+ - u\|^2\end{aligned}$$

where the last equality is due the the definition of τ^+ in (21). Now, using the assumption that $\gamma \leq \phi$ and the definition of A^+ in (20), we easily see that the above relation implies (22). ■

2.2 First ACG variant (Atouch and Teboule)

This variant chooses $a > 0$ such that

$$\frac{\tau(A+a)}{a^2} = L$$

and sets

$$y^+ = \tilde{y}, \qquad \gamma(\cdot) = \ell_f(\cdot;\tilde{x}) + h(\cdot) \in \overline{\mathrm{Conv}}_\mu(\mathbb{R}^n)$$

Then, $\gamma \leq \phi$ and

$$\begin{aligned}\gamma(\tilde{y}) + \frac{\tau(A+a)}{2a^2}\|\tilde{y}-\tilde{x}\|^2 &= \gamma(\tilde{y}) + \frac{L}{2}\|\tilde{y}-\tilde{x}\|^2 \\ &= \gamma(y^+) + \frac{L}{2}\|y^+-\tilde{x}\|^2 \\ &= \ell_f(y^+;\tilde{x}) + h(y^+) + \frac{L}{2}\|y^+-\tilde{x}\|^2 \\ &\geq (f+h)(y^+) = \phi(y^+)\end{aligned}$$

Instead of $y^+ = \tilde{y}$, we can instead choose y^+ such that

$$\gamma(\tilde{y}) + \frac{L}{2}\|\tilde{y}-\tilde{x}\|^2 \geq \gamma(y^+) + \frac{L}{2}\|y^+-\tilde{x}\|^2$$

e.g., y^+ given by

$$y^+ = \operatorname{argmin}\left\{\ell_f(u;\tilde{x}) + h(u) + \frac{L}{2}\|u-\tilde{x}\|^2\right\} \tag{24}$$

The latter variant is more expensive since it solves two subproblems per iteration, namely, subproblems (20) and (24).

Algorithm 4 (AT-ACG variant)

0. Let $x_0 \in \operatorname{dom} h$ be given, and set $y_0 = x_0$, $A_0 = 0$, $\tau_0 = 1$, and $k = 0$;

1. Compute

$$a_k = \frac{\tau_k + \sqrt{\tau_k^2 + 4L\tau_k A_k}}{2L}, \qquad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_k + a_k}; \tag{25}$$

2. Compute

$$x_{k+1} := \operatorname*{argmin}_{u \in \operatorname{dom} h} \left\{ a_k[\ell_f(u; \tilde{x}_k) + h(u)] + \frac{\tau_k}{2}\|u - x_k\|^2 \right\}, \tag{26}$$

$$y_{k+1} := \frac{A_k y_k + a_k x_{k+1}}{A_k + a_k} \tag{27}$$

$$A_{k+1} = A_k + a_k, \qquad \tau_{k+1} = \tau_k + a_k \mu; \tag{28}$$

3. Set $k \leftarrow k + 1$ and go to step 1.

Remark: Formula for a_k and A_{k+1} in (25) and (28), respectively, imply that

$$\frac{\tau_k A_{k+1}}{a_k^2} = \frac{\tau_k (A_k + a_k)}{a_k^2} = L. \tag{29}$$

Question: How fast does A_k grow?

1) By (25) and the fact that $\tau_k \geq \tau_0 = 1$, have

$$a_k \geq \frac{\tau_k + \sqrt{4L\tau_k A_k}}{2L} \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}}$$

and so

$$A_{k+1} = A_k + a_k \geq \frac{1}{2L} + \sqrt{\frac{A_k}{L}} + A_k \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\frac{1}{L}}\right)^2$$

Hence

$$\sqrt{A_{k+1}} \geq \sqrt{A_k} + \frac{1}{2}\sqrt{\frac{1}{L}}$$

This implies that

$$\sqrt{A_k} \geq \sqrt{A_0} + \frac{k}{2}\sqrt{\frac{1}{L}} = \frac{k}{2}\sqrt{\frac{1}{L}}$$

and hence that

$$A_k \geq \frac{k^2}{4L}$$

2) By (25) again, have

$$a_k \geq \frac{\tau_k + \sqrt{4L\tau_k A_k}}{2L} \geq \frac{\tau_k}{2L} + \sqrt{\frac{\tau_k A_k}{L}}$$

and so

$$A_{k+1} = A_k + a_k \geq \frac{\tau_k}{2L} + \sqrt{\frac{\tau_k A_k}{L}} + A_k \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\frac{\tau_k}{L}}\right)^2$$

Now, since $A_0 = 0$ and $\tau_0 = 1$, we have

$$A_k = A_0 + \sum_{i=0}^{k-1} a_i = \sum_{i=0}^{k-1} a_i, \quad \tau_k = \tau_0 + \mu \sum_{i=0}^{k-1} a_i = 1 + \mu A_k \geq \mu A_k$$

Combining the above two relations, we get

$$A_{k+1} \geq \left(\sqrt{A_k} + \frac{1}{2}\sqrt{\frac{\mu A_k}{L}}\right)^2 = A_k \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^2$$

Since $A_1 = 1/L$, have

$$A_k \geq \frac{1}{L}\left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2(k-1)}$$

Hence, by 1) and 2) above

$$A_k \geq \frac{1}{L} \max\left\{\frac{k^2}{4}, \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{2(k-1)}\right\}$$

Hence, the convergence rate of the AT-variant is

$$\phi(y_k) - \phi_* \leq \frac{d_0^2}{2A_k} \leq \frac{Ld_0^2}{2} \min\left\{ \frac{4}{k^2}\,,\ \left(1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{-2(k-1)} \right\}$$

Remark (iteration-complexity): For any tolerance $\varepsilon > 0$, if

$$k \geq \min\left\{ 2\sqrt{\frac{Ld_0^2}{2\varepsilon}}\,,\ \left[\frac{1}{2} + \sqrt{\frac{L}{\mu}}\right] \log_1^+\left(\frac{Ld_0^2}{2\varepsilon}\right) + 1 \right\}.$$

then have

$$\phi(y_k) - \phi_* \leq \varepsilon$$

Proof: Exercise

2.3 Second ACG variant (S-FISTA)

This variant was originally proposed by Nesterov for set optimization problems and later extended by Teboule et al to composite optimization problems.

While the AT variant solves a composite subproblem for x_{k+1} and obtains y_{k+1} in a straightforward manner, the FISTA variant below solves a composite subproblem for y_{k+1} and easily obtains x_{k+1}.

Algorithm 5 (S-FISTA)

0. Given $x_0 \in \operatorname{dom} h$, set $y_0 = x_0$, $A_0 = 0$, and $k = 0$;
1. Compute

$$a_k = \frac{\tau_k + \sqrt{\tau_k^2 + 4L\tau_k A_k}}{2L}, \qquad \tilde{x}_k = \frac{A_k y_k + a_k x_k}{A_k + a_k}; \tag{30}$$

2. Compute

$$y_{k+1} := \operatorname*{argmin}_{x \in \operatorname{dom} h} \left\{ \ell_f(x; \tilde{x}_k) + h(x) + \frac{L}{2}\|x - \tilde{x}_k\|^2 \right\}, \tag{31}$$

$$x_{k+1} = \frac{1}{\tau_{k+1}} \left[\tau_k x_k + L a_k (y_{k+1} - \tilde{x}_k) + \mu a_k y_{k+1} \right]; \tag{32}$$

$$A_{k+1} = A_k + a_k, \qquad \tau_{k+1} = \tau_k + a\mu \tag{33}$$

3. Set $k \leftarrow k + 1$ and go to step 1.

The justification of this variant relies on choosing

$$\gamma(\cdot) = \tilde{\gamma}(y^+) + L\langle \tilde{x} - y^+, \cdot - y^+ \rangle + \frac{\mu}{2}\| \cdot - y^+\|^2 \tag{34}$$

where for this analysis we let

$$\tilde{\gamma}(\cdot) := \ell_f(\cdot; \tilde{x}) + h(\cdot)$$

Clearly,

$$\tilde{\gamma}(\cdot) + \frac{L}{2}\| \cdot - \tilde{x}\|^2 \geq \phi(\cdot)$$

and, by (13), have

$$y^+ = \operatorname{argmin} \left\{ \tilde{\gamma}(u) + \frac{L}{2}\|u - \tilde{x}\|^2 \right\} \tag{35}$$

Lemma 2.2. *Function $\gamma(\cdot)$ defined in* (16) *satisfies*

$$\gamma \in \overline{\mathrm{Conv}}_{\mu}(\mathbb{R}^n), \quad \gamma(\cdot) \le \tilde{\gamma}(\cdot) \le \phi(\cdot)$$

and

$$\gamma(y^+) = \tilde{\gamma}(y^+), \quad y^+ = \mathrm{argmin}\left\{\gamma(u) + \frac{L}{2}\|u - \tilde{x}\|^2\right\}$$

Proof: The first, third and fourth relations above are straightforward. It remains to show that $\gamma \le \tilde{\gamma}$. By (35) and fundamental result of convex analysis, have

$$\begin{aligned}
&\tilde{\gamma}(u) + \frac{L}{2}\|u - \tilde{x}\|^2 \\
&\quad \ge \tilde{\gamma}(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2 + \frac{L+\mu}{2}\|u - y^+\|^2 \\
\text{(by (34))} \quad &= \gamma(u) + \frac{L}{2}\left(\|y^+ - \tilde{x}\|^2 + 2\langle y^+ - \tilde{x}, u - y^+\rangle + \|u - y^+\|^2\right) \\
&\quad = \gamma(u) + \frac{L}{2}\|u - \tilde{x}\|^2
\end{aligned}$$

and hence $\gamma \le \tilde{\gamma}$ holds. ∎

Using the above lemma, let us verify that (19) holds. Indeed,

$$\begin{aligned}
\gamma(\tilde{y}) + \frac{\tau(A+a)}{2a^2}\|\tilde{y} - \tilde{x}\|^2 &= \gamma(\tilde{y}) + \frac{L}{2}\|\tilde{y} - \tilde{x}\|^2 \\
&\ge \min\left\{\gamma(u) + \frac{L}{2}\|u - \tilde{x}\|^2\right\} = \gamma(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2 \\
&= \tilde{\gamma}(y^+) + \frac{L}{2}\|y^+ - \tilde{x}\|^2 \ge \phi(y^+)
\end{aligned}$$

showing that (19) holds.

Moreover, by (20), x^+ is always given by

$$x^+ = \operatorname{argmin} \left\{ a\gamma(u) + \frac{\tau}{2}\|u - x\|^2 \right\}$$

and hence satisfies

$$a\nabla\gamma(x^+) + \tau(x^+ - x) = 0$$

Since

$$\nabla\gamma(u) = L(\tilde{x} - y^+) + \mu(u - y^+),$$

we get

$$a[L(\tilde{x} - y^+) + \mu(x^+ - y^+)] + \tau(x^+ - x) = 0$$

and hence

$$\tau^+ x^+ = (\tau + a\mu)x^+ = \tau x + aL(y^+ - \tilde{x}) + \mu a y^+$$

which explains the update formula (32) for x_{k+1}.

2.4 Stationarity Complexity Bounds

Let L_f be the smallest L such that f is L-smooth. Let us consider the FISTA version with $L > L_f$.

Definition 2.3. *A pair (y, u) satisfying*

$$u \in \nabla f(x) + \partial h(x), \quad \|u\| \leq \rho$$

is called a a ρ-approximate stationary solution pair of $\phi = f + h$.

The following result describes the iteration-complexity of S-FISTA to find a ρ-approximate stationary solution pair of ϕ.

Lemma 2.4. *Define*

$$u_k = \nabla f(y_k) - \nabla f(\tilde{x}_{k-1}) + L(\tilde{x}_{k-1} - y_k).$$

Then, the following statements hold:

a) for every $k \geq 1$,

$$u_k \in \nabla f(y_k) + \partial h(y_k), \quad \min_{1 \leq i \leq k} \|u_i\|^2 \leq \frac{8L^2 d_0^2}{(L - L_f) \sum_{i=1}^{k} A_i};$$

b) for any $\rho > 0$, S-FISTA generates a ρ-approximate stationary solution pair $(y, u) := (y_k, u_k)$ in at most

$$\left\lceil \min \left\{ \left(\frac{12\zeta d_0^2}{\rho^2} \right)^{1/3}, \left(1 + \frac{2\sqrt{L}}{\sqrt{\mu}} \right) \log \left(1 + \frac{\zeta(c^2 - 1) d_0^2}{\rho^2} \right) \right\} \right\rceil$$

iterations, where

$$\zeta = \zeta(L, L_f) := \frac{8L^3}{L - L_f}, \qquad c = c(\mu, L) = 1 + \frac{1}{2}\sqrt{\frac{\mu}{L}}.$$

3 Nesterov's approximation scheme

Consider the min-max SP problem

$$\phi_* = \min\{\phi(x) := (f+h)(x) + \Phi(x) : x \in \mathbb{R}^n\} \tag{36}$$

where

$$\Phi(x) = \max\{\langle Ax, y\rangle - g(y) : y \in \mathbb{R}^m\} \tag{37}$$

and the following conditions are assumed:

1) $h \in \overline{\mathrm{Conv}}\,(\mathbb{R}^n)$ for some $\mu \geq 0$;

2) for some $L > 0$, $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth;

3) $g \in \overline{\mathrm{Conv}}\,(\mathbb{R}^m)$ and $\mathrm{dom}\, g$ is bounded;

4) $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear map;

5) the set X_* of optimal solutions of (36) is nonempty (and hence $\phi_* \in \mathbb{R}$).

Remarks:

- The boundedness condition on $\mathrm{dom}\, g$ is not needed but guarantees that $\Phi(x)$ is finite everywhere (exercise).
- the saddle function for the above min-max SP problem is

  $$\Psi(x, y) := (f+h)(x) + \langle Ax, y\rangle - g(y) \tag{38}$$

  which is a closed convex-concave one.

Proposition 3.1. *Consider the problem*

$$\tilde{\Phi}(z) = \max\{\langle z, y\rangle - \tilde{g}(y) : y \in \mathbb{R}^m\} \tag{39}$$

where $\tilde{g}(\cdot) \in \overline{\mathrm{Conv}}_\mu(\mathbb{R}^m)$ *for some* $\mu > 0$. *Then:*

a) for every $z \in \mathbb{R}^n$, *problem* (39) *has a unique optimal solution* $\tilde{y}(z)$;

b) $\tilde{\Phi}$ *is a finite everywhere convex function which is* $(1/\mu)$*-smooth and whose gradient is given by*

$$\nabla\tilde{\Phi}(z) = \tilde{y}(z) \quad \forall z \in \mathbb{R}^n.$$

Remark:

- Can not apply this result to (37) since the function $g(\cdot)$ in (37) is not assumed to be in $\overline{\mathrm{Conv}}_\mu(\mathbb{R}^m)$

So let us perturb function g by a positive multiple of a convex function $d \in \overline{\mathrm{Conv}}\,(\mathbb{R}^m)$ such that:

- $d(\cdot)$ is 1-strongly convex on $Y := \mathrm{dom}\, g$;
- there exists $y_0 \in Y$ such that $d(y_0) = 0$ and $d(y) \geq 0$ for every $y \in Y$,

i.e., set $g_\mu(\cdot) := g(\cdot) + \mu d(\cdot)$ and consider the perturbed problem

$$\begin{aligned} \Phi_\mu(x) &= \max\{\langle Ax, y\rangle - g_\mu(y) : y \in \mathbb{R}^m\} \\ &= \max\{\langle Ax, y\rangle - (g + \mu d)(y) : y \in \mathbb{R}^m\}. \end{aligned} \tag{40}$$

Example: $d(y) = \|y - y_0\|^2/2$ for every $y \in \mathbb{R}^m$

Idea of the scheme: If μ is small, then Φ_μ is a smooth finite everywhere convex function which closely approximates Φ. We can then solve the perturbed problem

$$(\phi_\mu)_* = \min\{\phi_\mu(x) := (f + h)(x) + \Phi_\mu(x) : x \in \mathbb{R}^n\} \tag{41}$$

using one of the ACG variants (e.g., FISTA).

Proposition 3.2. *For every $x \in \mathbb{R}^n$ and $\mu > 0$, we have*

a) (40) *has a unique optimal solution $y_\mu(x)$;*

b) Φ_μ is a finite everywhere convex function; moreover, Φ_μ is $(\|A\|^2/\mu)$-smooth,

$$\nabla\Phi_\mu(x) = A^* y_\mu(x) \quad \forall x \in \mathbb{R}^n$$

and $y_\mu(\cdot)$ is $(\|A\|/\mu)$-Lipschitz continuous;

c) there holds

$$0 \le \phi(x) - \phi_\mu(x) = \Phi(x) - \Phi_\mu(x) \le \mu D_Y^2$$

where

$$D_Y := \sup\{[d(y)]^{1/2} : y \in Y\};$$

d) there holds

$$(\phi_\mu)_* \le \phi_* \le (\phi_\mu)_* + \mu D_Y^2$$

Proof: a) Consider problem (39) with $\tilde{g} = g_\mu$ and note that

$$\Phi_\mu(x) = \tilde{\Phi}(Ax) \tag{42}$$

where $\tilde{\Phi}$ is as in (39). Clearly, in view of Prop 3.1(a) with $z = Ax$, (40) has a unique optimal solution $y_\mu(x) = \tilde{y}(Ax)$.

b) It follows from (42) and Prop 3.1(b) that Φ_μ is is a finite everywhere convex function such that

$$\nabla\Phi_\mu(x) = A^*\nabla\tilde{\Phi}(Ax) = A^*\tilde{y}(Ax) = A^* y_\mu(x) \quad \forall x \in \mathbb{R}^n.$$

Hence, for every $x, x' \in \mathbb{R}^n$, have

$$\begin{aligned}
\|\nabla\Phi_\mu(x') - \nabla\Phi_\mu(x)\| &= \|A^*\tilde{y}(Ax') - A^* y(Ax)\| \\
&\le \|A^*\|\|\tilde{y}(Ax') - \tilde{y}(Ax)\| \\
\text{Prop 3.1(b)} \quad &\le \frac{\|A^*\|}{\mu}\|Ax' - Ax\| \\
&\le \frac{\|A^*\|\|A\|}{\mu}\|x' - x\| \\
&= \frac{\|A\|^2}{\mu}\|x' - x\|
\end{aligned}$$

c) The definition of D_Y and the assumption that $d(y) \geq 0$ for every $y \in Y$ imply that for every $x \in \mathbb{R}^n$ and $y \in Y$,

$$\begin{aligned}
&\langle Ax, y\rangle - (g + \mu d)(y) \\
&\qquad \leq \langle Ax, y\rangle - g(y) \\
&\qquad \leq \langle Ax, y\rangle - g(y) + \mu[D_Y^2 - d(y)] \\
&\qquad \leq \langle Ax, y\rangle - (g + \mu d)(y) + \mu D_Y^2
\end{aligned}$$

Taking the supremum of both sides with respect to y, we then conclude that

$$\Phi_\mu(x) \leq \Phi(x) \leq \Phi_\mu(x) + \mu D_Y^2 \qquad \forall x \in \mathbb{R}^n.$$

d) Follows from c) and the definitions of ϕ_* and $(\phi_\mu)_*$.

Algorithm 6 (Nesterov approximation scheme)

0) Let $\varepsilon > 0$ and $x_0 \in \operatorname{dom} h$ be given and set

$$\mu = \frac{\varepsilon}{2D_Y^2};$$

1) Apply an ACG variant to the perturbed problem (41) started from x_0 and stop with an iterate x_k satisfying

$$\phi_\mu(x_k) - (\phi_\mu)_* \leq \frac{\varepsilon}{2} \qquad (*)$$

Obs: $(*)$ can also be replaced by $\phi(x_k) - \phi_* \leq \varepsilon$

Note that

$$\begin{aligned}\phi(x_k) - \phi_* &= [\phi(x_k) - \phi_\mu(x_k)] + [\phi_\mu(x_k) - (\phi_\mu)_*] + [(\phi_\mu)_* - \phi_*] \\ &\leq \mu D_Y^2 + [\phi_\mu(x_k) - (\phi_\mu)_*] + 0 \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + 0 = \varepsilon\end{aligned}$$

Analysis of the scheme:

By the ACG analysis, we have that

$$\begin{aligned}\phi(x_k) - \phi(x) \leq \phi_\mu(x_k) + \mu D_Y^2 - \phi_\mu(x) &\leq \frac{\|x_0 - x\|^2}{2A_k} + \mu D_Y^2 \\ &\leq \left(L + \frac{\|A\|^2}{\mu}\right) \frac{2\|x_0 - x\|^2}{k^2} + \frac{\varepsilon}{2}\end{aligned}$$

Taking $x = \operatorname{Proj}_{X_*}(x_0)$, we have

$$\phi(x_k) - \phi_* \leq \left(L + \frac{\|A\|^2}{\mu}\right) \frac{2d_0^2}{k^2} + \frac{\varepsilon}{2}$$

So, if

$$k^2 \geq \left(L + \frac{\|A\|^2}{\mu}\right) \frac{4d_0^2}{\varepsilon}$$

then $\phi(x_k) - \phi_* \leq \varepsilon$.

So, the iteration complexity of Nesterov's approximation scheme is

$$\mathcal{O}_1\left(\left(\sqrt{L} + \frac{\|A\|}{\sqrt{\mu}}\right) \frac{d_0}{\sqrt{\varepsilon}}\right) = \mathcal{O}_1\left(\left(\sqrt{\frac{L}{\varepsilon}} + \frac{\|A\| D_Y}{\varepsilon}\right) d_0\right).$$

Another Termination:

Let

$$f_\mu = f + \Psi_\mu, \quad L_\mu = L + \frac{\|A\|^2}{\mu}$$

Then f_μ is L_μ-smooth.

Setting

$$u_k = \nabla f_\mu(x_k) - \nabla \tilde{f}_\mu(\tilde{x}_{k-1}) + L_\mu(\tilde{x}_{k-1} - x_k)$$

have

$$u_k \in \nabla f_\mu(x_k) + \partial h(x_k) \quad \forall k \geq 0 \tag{43}$$

Moreover, we have

$$\min_{i \leq k} \|u_i\|^2 = \mathcal{O}\left(\frac{L_\mu^2 d_0^2}{k^3}\right) \tag{44}$$

Hence, the complexity of finding k such that $\|u_k\| \leq \rho$ is

$$\mathcal{O}\left(\left(\frac{L_\mu d_0}{\rho}\right)^{2/3}\right)$$

Now let us interpret (43), i.e., the inclusion

$$u_k \in \nabla f_\mu(x_k) + \partial h(x_k).$$

Have

$$\nabla f_\mu(x) = \nabla f(x) + \nabla \Psi_\mu(x) = \nabla f(x) + A^* y_\mu(x)$$

Now, let $y_k := y_\mu(x_k)$. Then,

$$\nabla f_\mu(x_k) = \nabla f(x_k) + A^* y_k$$

So, (43) reduces to

$$u_k \in \nabla f(x_k) + A^* y_k + \partial h(x_k)$$

Also, by the definition of $y_\mu(\cdot)$ as being the optimal solution of (40), we have

$$0 \in -Ax + \partial g_\mu(y_\mu(x)) = 0 \quad \forall x \in \mathbb{R}^n.$$

Taking $x = x_k$ yields

$$0 \in -Ax_k + \partial g_\mu(y_k)$$

or equivalently

$$0 \in -Ax_k + \partial g(y_k) + \mu \nabla d(y_k)$$

So let

$$v_k := -\mu \nabla d(y_k)$$

Then

$$u_k \in \nabla f(x_k) + A^* y_k + \partial h(x_k), \qquad v_k \in -Ax_k + \partial g(y_k)$$

Now, the optimality condition for (x, y) to be a saddle point for Ψ in (38) is

$$0 \in \nabla f(x) + A^* y + \partial h(x), \qquad 0 \in -Ax + \partial g(y)$$

Hence, if the residual pair (u_k, v_k) is small then (x_k, y_k) is a near saddle point for Ψ.

Now u_k can be made small by (44).

How about v_k? If the quantity

$$\tilde{D}_Y := \sup\{\|\nabla d(y_k)\| : k \geq 1\}$$

is finite and μ is chosen so as to satisfy

$$\mu \leq \frac{\rho}{\tilde{D}_Y}$$

then

$$\|v_k\| = \mu\|\nabla d(y_k)\| \leq \mu\tilde{D}_Y \leq \rho$$

Special case: If $d(y) = \|y - y_0\|^2/2$ and Y is bounded, then

$$\tilde{D}_Y = \sup\{\|y_k - y_0\| : k \geq 1\} \leq \sqrt{2}D_Y$$

Hence, it suffices to choose

$$\mu = \frac{\rho}{\sqrt{2}D_Y}$$

Exercise: Consider the case where Y is unbounded and $d(y) = \|y - y_0\|^2/2$. Show that $\tilde{D}_Y$ is also finite.

Hint:

- show that x_k is bounded;
- show that

$$\|y_k - y_0\| \leq \frac{\|A\|}{\mu}\|x_k - x_0\|$$