Convergence of ISTA

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be   Homepage: angms.science

First draft : December 20, 2019
Last update : January 21, 2020

Problem set up

$$(\mathcal{P}) \quad : \quad \min_{\mathbf{x}} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$$

where

- $f : \mathbb{R}^n \to \mathbb{R}$ is smooth convex function (it is continuous differentiable with Lipschitz gradient : $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|$, where $L_f > 0$)
- $g : \mathbb{R}^n \to \mathbb{R}$ is convex but not necessary smooth
- $(\mathcal{P})$ is assumed solvable : the solution set $\mathcal{X}^* = \arg\min_{\mathbf{x}} F \neq \varnothing$
- Notation : $\mathbf{x}^* \in \mathcal{X}^*$ and $F^* = F(\mathbf{x}^*)$

Comments

- $(\mathcal{P})$ is a convex minimization problem
- If $g \equiv 0$ (corresponds to "no regularization" in applications), $(\mathcal{P})$ reduces to general smooth convex problem
- ISTA is an algorithm for solving the L1-regularized Least Squares
$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$
which is a special case of $(\mathcal{P})$ with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.
See here for details on the problem.

Idea on solving $(\mathcal{P})$ - The quadratic over-approximation Q

The key observation on solving $(\mathcal{P})$ is that f is L_f-smooth :

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_f \|\mathbf{x} - \mathbf{y}\|,$$

which implies for all $\mathbf{x}, \mathbf{y}$ and $L \geq L_f$,

$$f(\mathbf{x}) \leq \underbrace{f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2}_{q(\mathbf{x})},$$

where $q(\mathbf{x})$ is a local quadratic over-approximation of f at the point $\mathbf{y}$. By $f(\mathbf{x}) \leq q(\mathbf{x})$, the objective function $F(\mathbf{x})$ is upper bounded by

$$\begin{aligned} F(\mathbf{x}) \leq Q(\mathbf{x};\mathbf{y}) \quad &:= \quad q(\mathbf{x}) + g(\mathbf{x}) \\ &= \quad f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 + g(\mathbf{x}) \end{aligned}$$

Therefore, if we minimize $Q(\mathbf{x};\mathbf{y})$, we indirectly minimize $F(\mathbf{x})$ as Q is upper bound of F for all $\mathbf{x}$.

Here the philosophy of Majorization-Minimization is used : we indirectly minimize the original F by minimizing an upper bound Q.

Idea on solving $(\mathcal{P})$ - Minimizing Q ... 1/2

$$Q(\mathbf{x};\mathbf{y}) := f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + g(\mathbf{x})$$

is a convex function (but not necessary smooth due to g), so minimizers of Q are global minimizer.

Let $p_L(\mathbf{x}) = \arg\min_{\mathbf{x}} Q(\mathbf{x};\mathbf{y})$, try to see is there anything we can get by studying $p_L(\mathbf{x})$

$$\begin{aligned} p_L(\mathbf{x}) &= \arg\min_{\mathbf{x}} Q(\mathbf{x};\mathbf{y}) \\ &= \arg\min_{\mathbf{x}} \left\{ f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2 + g(\mathbf{x}) \right\} \end{aligned}$$

We continue based on two ideas :

- Idea 1 - The minimization is on $\mathbf{x}$, it is ok to remove or add constant terms that are independent of $\mathbf{x}$
- Item 2 - Q contains quadratic terms, we can try completing squares

Idea on solving $(\mathcal{P})$ - Minimizing Q ... 2/2

By idea 1, remove constant terms that are independent of $\mathbf{x}$ gives

$$\arg\min_{\mathbf{x}} \left\{ \langle \nabla f(\mathbf{y}), \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x}\|_2^2 - \langle \mathbf{x}, \mathbf{y} \rangle + g(\mathbf{x}) \right\}$$

By idea 2, combine the terms gives

$$\arg\min_{\mathbf{x}} \left\{ \langle -L\mathbf{y} + \nabla f(\mathbf{y}), \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x}\|_2^2 + g(\mathbf{x}) \right\}$$

Factorize $\dfrac{L}{2}$ out gives

$$\arg\min_{\mathbf{x}} \left\{ \frac{L}{2}\Big(-2\langle \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}), \mathbf{x} \rangle + \|\mathbf{x}\|_2^2 \Big) + g(\mathbf{x}) \right\}$$

Completing the square by adding constant terms, which is ok (idea 1)

$$\arg\min_{\mathbf{x}} \left\{ \frac{L}{2}\Big(\|\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\|_2^2 - 2\langle \mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y}), \mathbf{x} \rangle + \|\mathbf{x}\|_2^2 \Big) + g(\mathbf{x}) \right\}$$

Hence we have

$$p_L(\mathbf{y}) = \arg\min_{\mathbf{x}} \left\{ \frac{L}{2}\Big\|\mathbf{x} - \Big(\mathbf{y} - \frac{1}{L}\nabla f(\mathbf{y})\Big)\Big\|_2^2 + g(\mathbf{x}) \right\}$$

The update step

$$p_L(\mathbf{y}) = \arg\min_{\mathbf{x}} \left\{ \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{y} - \frac{1}{L} \nabla f(\mathbf{y}) \right) \right\|_2^2 + g(\mathbf{x}) \right\}$$

Let $\mathbf{y} = \mathbf{x}_{k-1}$, we have

$$\mathbf{x}_k = p_L(\mathbf{x}_{k-1}) = \arg\min_{\mathbf{x}} \left\{ \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{x}_{k-1} - \frac{1}{L} \nabla f(\mathbf{x}_{k-1}) \right) \right\|_2^2 + g(\mathbf{x}) \right\}.$$

Hence $p_L(\mathbf{y})$ can be seen as the update operator of a point $\mathbf{y}$, and such operator forms the basis of the iterative algorithm.

- $p_L(\mathbf{y})$ itself is an optimization problem.
  In general $p_L(\mathbf{y})$ may not have close form solution.
- When $g(\mathbf{x}) = \|\mathbf{x}\|_1$ and $f(\mathbf{x}) = \dfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, $p_L(\mathbf{y})$ is the update step of ISTA (Iterative Soft Thresholding Algorithm), which in this case here $p_L(\mathbf{y})$ has close form solution. See here for details.

On solving $p_L(\mathbf{y})$

$$p_L(\mathbf{y}) = \arg\min_{\mathbf{x}} \left\{ \frac{L}{2} \left\| \mathbf{x} - \left(\mathbf{y} - \frac{1}{L} \nabla f(\mathbf{y}) \right) \right\|_2^2 + g(\mathbf{x}) \right\}$$

By sub-gradient version of the first order optimality condition/Fermat Rule, when $\mathbf{x}$ is the minimizer of $p_L(\mathbf{y})$, we have

$$0 \in \partial p_L(\mathbf{y}) \quad \Longleftrightarrow \quad 0 \in \partial g(\mathbf{x}) + L(\mathbf{x} - \mathbf{y}) + \nabla f(\mathbf{y})$$

where $\partial g(\mathbf{x})$ denotes the sub-differential of g at $\mathbf{x}$.

Let $\gamma(\mathbf{y}) \in \partial g(\mathbf{y})$, then one has $\mathbf{x} = p_L(\mathbf{y})$ iff there exists $\gamma(\mathbf{y}) \in \partial g(\mathbf{y})$ is the sub-differential of g s.t. $\nabla f(\mathbf{y}) + L(\mathbf{x} - \mathbf{y}) + \gamma(\mathbf{y}) = 0$.

Lemma

For all $\mathbf{y}$, $L \geq L_f > 0$ s.t. $F(p_L(\mathbf{y})) \leq Q(p_L(\mathbf{y}); \mathbf{y})$, then for all $\mathbf{x}$ we have

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|_2^2 + L\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle.$$

Proof. Consider for all $\mathbf{y}$, $L > L_f > 0$ s.t. $F(p_L(\mathbf{y})) \leq Q(p_L(\mathbf{y}); \mathbf{y})$,

$$\begin{aligned} & F(p_L(\mathbf{y})) \leq Q(p_L(\mathbf{y}); \mathbf{y}) \\ \Longleftrightarrow \quad & -F(p_L(\mathbf{y})) \geq -Q(p_L(\mathbf{y}); \mathbf{y}) \\ \Longleftrightarrow \quad & F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq F(\mathbf{x}) - Q(p_L(\mathbf{y}); \mathbf{y}). \qquad (1) \end{aligned}$$

Key lemma ... 2/4

As f is convex and smooth, and g is convex (possibly not smooth), we have the following inequality

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \\ g(\mathbf{x}) &\geq g(p_L(\mathbf{y})) + \langle \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle \end{aligned}$$

Sum the two inequalities we get

$$\underbrace{f(\mathbf{x}) + g(\mathbf{x})}_{F(\mathbf{x})} \geq f(\mathbf{y}) + g(p_L(\mathbf{y})) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle \quad (2)$$

Note : it is wrong to say $g(\mathbf{x}) \geq g(\mathbf{y}) + \langle \nabla \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$ as g is not smooth (not differentiable).

Finally, based on the definition of Q, put $\mathbf{x} = p_L(\mathbf{y})$, we have

$$\begin{aligned} & Q(p_L(\mathbf{y}); \mathbf{y}) \\ = \; & f(\mathbf{y}) + \langle p_L(\mathbf{y}) - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{L}{2} \| p_L(\mathbf{y}) - \mathbf{y} \|_2^2 + g(p_L(\mathbf{y})) \quad (3) \end{aligned}$$

Put (3) and (2) into right hand side of (1) gives

$$\begin{array}{r} F(\mathbf{x}) \\ -F(p_L(\mathbf{y})) \end{array} \geq \begin{array}{r} f(\mathbf{y}) + g(p_L(\mathbf{y})) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \langle \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle \\ -\Big(+ f(\mathbf{y}) + \langle p_L(\mathbf{y}) - \mathbf{y}, \nabla f(\mathbf{y}) \rangle + \frac{L}{2} \|p_L(\mathbf{y}) - \mathbf{y}\|_2^2 + g(p_L(\mathbf{y})) \Big) \end{array}$$

Cancel terms, combine terms

$$\begin{array}{r} F(\mathbf{x}) \\ -F(p_L(\mathbf{y})) \end{array} \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle + \langle \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle - \frac{L}{2} \|p_L(\mathbf{y}) - \mathbf{y}\|_2^2$$

Further combine the dot product terms gives

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \langle \nabla f(\mathbf{y}) + \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle - \frac{L}{2} \|p_L(\mathbf{y}) - \mathbf{y}\|_2^2$$

Key lemma ... 4/4

By first order optimality condition / Fermat's rule :
$\nabla f(\mathbf{y}) + L(\mathbf{x}-\mathbf{y}) + \gamma(\mathbf{y}) = 0$, we have $\nabla f(\mathbf{y}) + \gamma(\mathbf{y}) = L(\mathbf{y}-\mathbf{x})$ and we can put $\mathbf{x} = p_L(\mathbf{y})$. Hence

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \langle \nabla f(\mathbf{y}) + \gamma(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle - \frac{L}{2} \| p_L(\mathbf{y}) - \mathbf{y} \|_2^2$$

becomes

$$\begin{aligned} F(\mathbf{x}) - F(p_L(\mathbf{y})) &\geq \langle L(\mathbf{y} - p_L(\mathbf{y})), \mathbf{x} - p_L(\mathbf{y}) \rangle - \frac{L}{2} \| p_L(\mathbf{y}) - \mathbf{y} \|_2^2 \\ &= L\langle \mathbf{y} - p_L(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle - \frac{L}{2} \| p_L(\mathbf{y}) - \mathbf{y} \|_2^2 \end{aligned}$$

A tricky step : we have $({\color{red}a} - {\color{blue}c})({\color{green}b} - {\color{blue}c}) = ({\color{green}b} - {\color{red}a})({\color{blue}c} - {\color{green}b}) + ({\color{blue}c} - {\color{green}b}^2)$, then

$$L\langle \mathbf{y} - p_L(\mathbf{y}), \mathbf{x} - p_L(\mathbf{y}) \rangle = L\langle \mathbf{x} - \mathbf{y}, p_L(\mathbf{y}) - \mathbf{y} \rangle + L \| p_L(\mathbf{y}) - \mathbf{y} \|_2^2$$

Put this equality in last inequality finishes the proof. $\square$

ISTA and FISTA

The algorithm ISTA iterate the following step

$$\mathbf{x}_k = p_L(\mathbf{x}_{k-1}).$$

with $L = L_f$ known. If L_f is unknown, backtracking can be used.

The algorithm FISTA iterate the following step

- $\mathbf{x}_k = p_L(\mathbf{y}_{k-1})$
- $t_{k+1} = \dfrac{1 + \sqrt{1 + 4t_k^2}}{2}$
- $\mathbf{y}_{k+1} = \mathbf{x}_k + \dfrac{t_k - 1}{t_{k+1}}(\mathbf{x}_k - \mathbf{x}_{k-1})$

Remarks

- ISTA belongs to the proximal gradient algorithm. See here for details
- FISTA is a kind of Nesterov-type acceleration applied on ISTA
- FISTA belongs to the accelerated proximal gradient algorithm

Convergence of ISTA

The iterates produced by ISTA is monotone :

1. $Q_L(\mathbf{x};\mathbf{y}) = F(\mathbf{x})$ if $\mathbf{x} = \mathbf{y}$ as
   $Q(\mathbf{x};\mathbf{y}) = f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$
2. $p_L(\mathbf{y}) = \arg\min_{\mathbf{x}} Q(\mathbf{x};\mathbf{y})$, based on definition of $p_L(\mathbf{y})$
3. Q is local quadratic over-estimator of F, so $F(\mathbf{x}) \leq Q(\mathbf{x})$

Hence we have

$$F(\mathbf{x}_k) \overset{3}{\leq} Q(\mathbf{x}_k, \mathbf{x}_{k-1}) \overset{2}{\leq} Q(\mathbf{x}_{k-1}, \mathbf{x}_{k-1}) \overset{1}{\leq} F(\mathbf{x}_{k-1})$$

As $F(\mathbf{x}) \geq 0$ for all $\mathbf{x}$, together with the monotonicity, the sequence of the objective function value converge.
As $(\mathcal{P})$ is a convex problem, it converges to global minima with optimal objective value F^*.

Furthermore, it can be shown that the convergence rate of ISTA is

Theorem (Convergence rate of ISTA using constant step size)

$$F(\mathbf{x}_k) - F^* \leq \frac{L_f R_0^2}{2k} \; \forall k, R_0 = \|\mathbf{x}_0 - \mathbf{x}^*\|_2$$

We give the proof starting in the next slide.

Remarks

- As ISTA is a special case of proximal gradient algorithm, the convergence properties of proximal gradient algorithm applies to ISTA. See here for another approach to show the $1/k$ convergence rate.
- The proof starting in the next slide follows this paper.

Proof. By the key lemma

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|_2^2 + L\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle,$$

put $\mathbf{y} = \mathbf{x}_k$, then based on ISTA, $p_L(\mathbf{y}) = \mathbf{x}_{k+1}$. Also put $\mathbf{x} = \mathbf{x}^*$:

$$F(\mathbf{x}^*) - F(\mathbf{x}_{k+1}) \geq \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + L\langle \mathbf{x}_k - \mathbf{x}^*, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle,$$

Based on observation, consider completing squares

$$\begin{aligned}
\frac{2}{L}\Big(F^* - F(\mathbf{x}_{k+1})\Big) &\geq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + 2\langle \mathbf{x}_k - \mathbf{x}^*, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle, \\
&= \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 + 2\langle \mathbf{x}_k - \mathbf{x}^*, \mathbf{x}_{k+1} - \mathbf{x}_k\rangle \\
&\qquad + \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_{k+1} - \mathbf{x}_k + \mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2
\end{aligned}$$

Take summation from $k = 0$ to $k - 1$

$$\frac{2}{L}\Big(kF^* - \sum_{i=0}^{k-1} F(\mathbf{x}_{i+1})\Big) \geq \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 - \underbrace{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}_{R_0^2} \tag{4}$$

Proof of convergence of ISTA ... 2/4

By the key lemma

$$F(\mathbf{x}) - F(p_L(\mathbf{y})) \geq \frac{L}{2}\|p_L(\mathbf{y}) - \mathbf{y}\|_2^2 + L\langle \mathbf{y} - \mathbf{x}, p_L(\mathbf{y}) - \mathbf{y}\rangle,$$

put $\mathbf{x} = \mathbf{y} = \mathbf{x}_k$, $p_L(\mathbf{y}) = \mathbf{x}_{k+1}$

$$F(\mathbf{x}_k) - F(\mathbf{x}_{k+1}) \geq \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$$

$$\iff \qquad \frac{2}{L}\Big(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})\Big) \geq \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$$

Tricky step : multiply the whole inequality by k

$$\frac{2}{L}k\Big(F(\mathbf{x}_k) - F(\mathbf{x}_{k+1})\Big) \geq k\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2$$

Take summation from $k = 0$ to $k - 1$

$$\frac{2}{L}\sum_{i=0}^{k-1} i\Big(F(\mathbf{x}_i) - F(\mathbf{x}_{i+1})\Big) \geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2$$

Continue from last page, a tricky step

$$
\begin{aligned}
\frac{2}{L}\sum_{i=0}^{k-1} i\Big(F(\mathbf{x}_i) - F(\mathbf{x}_{i+1})\Big) {\color{red}+F(\mathbf{x}_{i+1}) - F(\mathbf{x}_{i-1})} &\geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2 \\
\frac{2}{L}\sum_{i=0}^{k-1} iF(\mathbf{x}_i) - (i+1)F(\mathbf{x}_{i+1}) + F(\mathbf{x}_{i+1}) &\geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2 \\
\frac{2}{L}\Bigg(\underbrace{\sum_{i=0}^{k-1}\Big(iF(\mathbf{x}_i) - (i+1)F(\mathbf{x}_{i+1})\Big)}_{-kF(\mathbf{x}_k)} + \sum_{i=0}^{k-1} F(\mathbf{x}_{i+1})\Bigg) &\geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2
\end{aligned}
$$

Finally

$$
\frac{2}{L}\Bigg(-kF(\mathbf{x}_k) + \sum_{i=0}^{k-1} F(\mathbf{x}_{i+1})\Bigg) \geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2 \tag{5}
$$

$$\text{(4)} \qquad \frac{2}{L}\Big(kF^* - \sum_{i=0}^{k-1} F(\mathbf{x}_{i+1})\Big) \geq \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 - R_0^2$$

$$\text{(5)} \quad \frac{2}{L}\Bigg(-kF(\mathbf{x}_k) + \sum_{i=0}^{k-1} F(\mathbf{x}_{i+1})\Bigg) \geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2$$

(4) + (5) gives

$$\frac{2k}{L}\big(F^* - F(\mathbf{x}_k)\big) \geq \sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2 + \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 - R_0^2$$

Multiply -1 to the whole inequality

$$\frac{2k}{L}\big(F(\mathbf{x}_k) - F^*\big) \quad \leq \quad R_0^2 - \underbrace{\sum_{i=0}^{k-1} i\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_2^2}_{\geq 0} - \underbrace{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2}_{\geq 0} \leq R_0^2 \quad \square$$