Convergence of ISTA (Beck & Teboulle 2009)

Andersen Ang

ECS, Uni. Southampton, UK
andersen.ang@soton.ac.uk

Homepage angms.science

Version: July 17, 2023
First draft: Dec 20, 2019

Content

Reference: Amir Beck and Marc Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." SIAM journal on imaging sciences, 2009.

Problem setup: smooth convex optimisation with convex possibly-nonsmooth regularzier

$$(\mathcal{P}) \ : \ \operatorname*{argmin}_{\boldsymbol{x}} \Big\{ F(\boldsymbol{x}) \coloneqq f(\boldsymbol{x}) + g(\boldsymbol{x}) \Big\}. \qquad \text{(a convex optimization problem)}$$

- We consider Euclidean space
- $f : \mathbb{R}^n \to \mathbb{R}$
- f is $L-$smooth
  - f is continuously differentiable — $f \in \mathcal{C}^1$, i.e., $\nabla f(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \operatorname{dom} f$
  - ∇f is L-Lipschitz — $L > 0$ is the least upper bound in $\dfrac{\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|}{\|\boldsymbol{x} - \boldsymbol{y}\|} \leq L$
  - $(\forall \boldsymbol{a} \in \operatorname{dom} f)(\forall \boldsymbol{b} \in \operatorname{dom} f) \Big\{ f(\boldsymbol{a}) - f(\boldsymbol{b}) \leq \langle \nabla f(\boldsymbol{b}), \boldsymbol{a} - \boldsymbol{b} \rangle + \dfrac{L}{2} \|\boldsymbol{a} - \boldsymbol{b}\|_2^2 \Big\}$.
- f is convex — all local minima of $\mathcal{P}$ are global minima
  - $(\forall \boldsymbol{x} \in \operatorname{dom} f)(\forall \boldsymbol{y} \in \operatorname{dom} f) \Big\{ f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \Big\}$
- $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ — $\overline{\mathbb{R}} \coloneqq [-\infty, +\infty]$
  - g is convex, closed, proper and possibly nonsmooth
  - closed & proper is for proximal operator to work
  - possibly nonsmooth means we need to replace gradient with subgradient
- Details of convexity, L-smoothness, closeness, proper, see here
- $(\mathcal{P})$ is assumed solvable: the solution set $\mathcal{X}^* = \operatorname*{argmin}_{\boldsymbol{x}} F \neq \varnothing$
- Notation : $\boldsymbol{x}^* \in \mathcal{X}^*$ and $F^* = F(\boldsymbol{x}^*) \notin \{\pm\infty\}$

Idea on solving $(\mathcal{P})$

- **Key observation** f is $L-$smooth : $(\forall \boldsymbol{x} \in \mathrm{dom} f)(\forall \boldsymbol{y} \in \mathrm{dom} f)$ we have

$$f(\boldsymbol{x}) \quad \leq \quad q(\boldsymbol{x}; \boldsymbol{y}) \coloneqq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

  where $q(\boldsymbol{x})$ is a local quadratic over-approximation of f at a point $\boldsymbol{y}$. Here $\boldsymbol{y}$ is treated as a "parameter".

- **Quadratic over-estimator Q**
  By $f(\boldsymbol{x}) \leq q(\boldsymbol{x})$, for $F(\boldsymbol{x})$ we have

$$\begin{aligned}
& f(\boldsymbol{x}) && \leq && q(\boldsymbol{x}; \boldsymbol{y}) && && \text{by } \blacksquare \\
\iff \quad & f(\boldsymbol{x}) + g(\boldsymbol{x}) && \leq && q(\boldsymbol{x}; \boldsymbol{y}) + g(\boldsymbol{x}) && && \text{add } g \\
\iff \quad & F(\boldsymbol{x}) && \leq && q(\boldsymbol{x}; \boldsymbol{y}) + g(\boldsymbol{x}) && && \text{by definition } F \coloneqq f + g \\
\iff \quad & F(\boldsymbol{x}) && \leq && Q(\boldsymbol{x}; \boldsymbol{y}) && && \text{let } Q \coloneqq q + g \\
\iff \quad & F(\boldsymbol{x}) && \leq && f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + g(\boldsymbol{x}) && && \text{expand } q
\end{aligned}$$

- **Majorization-minimization**:

$$\text{minimizing } Q(\boldsymbol{x}; \boldsymbol{y}) \xrightarrow{Q \text{ is global upper bound of } F} \text{indirectly minimize } F(\boldsymbol{x})$$

Discussion on minimizing Q

$$Q(\boldsymbol{x};\boldsymbol{y}) \coloneqq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + g(\boldsymbol{x}) \quad \text{is a strictly convex function}$$

- $f(\boldsymbol{y})$ is a constant — it is a convex function
- $\langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle$ is a linear function on $\boldsymbol{x}$ — it is a convex function
- $\frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2$ is a quadratic function on $\boldsymbol{x}$ — it is a strictly convex function by $L > 0$
- $g(\boldsymbol{x})$ — it is a convex function by assumption in problem setup

$$\operatorname*{argmin}_{\boldsymbol{x}} \left\{ Q(\boldsymbol{x};\boldsymbol{y}) \coloneqq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + g(\boldsymbol{x}) \right\}$$

- This is a convex optimization problem: all local minimizers are global
- This problem is a strictly convex: the global minimizer is unique
- This problem has nonempty solution set: because we assumed g is closed and proper

Conclusion : $\operatorname*{argmin}_{\boldsymbol{x}} Q(\boldsymbol{x})$ is a $\mathbb{R}^n \to \mathbb{R}^n$ map.

Minimizing $Q \equiv$ computing a proximal operator

$$
\begin{aligned}
& \operatorname*{argmin}_{\boldsymbol{x}} Q(\boldsymbol{x};\boldsymbol{y}) && \\
:= \quad & \operatorname*{argmin}_{\boldsymbol{x}} f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + g(\boldsymbol{x}) && \text{by definition of } Q \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}\rangle - \langle \nabla f(\boldsymbol{y}), \boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{x}\|_2^2 - L\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \frac{L}{2}\|\boldsymbol{y}\|_2^2 + g(\boldsymbol{x}) && \text{expand} \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} \langle \nabla f(\boldsymbol{y}), \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{x}\|_2^2 - L\langle \boldsymbol{x}, \boldsymbol{y}\rangle + g(\boldsymbol{x}) && \text{argmin ignore constant} \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} \langle \nabla f(\boldsymbol{y}) - L\boldsymbol{y},\ \boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{x}\|_2^2 + g(\boldsymbol{x}) && \text{combine inner product} \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} \frac{L}{2}\left(- 2\langle \boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y}),\ \boldsymbol{x}\rangle + \|\boldsymbol{x}\|_2^2\right) + g(\boldsymbol{x}) && \text{factor out } \frac{L}{2} \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} \frac{L}{2}\left(\left\|\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})\right\|_2^2 - 2\langle \boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y}),\ \boldsymbol{x}\rangle + \|\boldsymbol{x}\|_2^2\right) + g(\boldsymbol{x}) && \text{argmin ignore constant} \\
\iff \quad & \operatorname*{argmin}_{\boldsymbol{x}} \frac{L}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})\right)\right\|_2^2 + g(\boldsymbol{x}) && \text{completing squares} \\
\iff \quad & \operatorname{prox}_g^L\left(\boldsymbol{y} - \frac{1}{L}\nabla f(\boldsymbol{y})\right) && \text{proximal operator of } g
\end{aligned}
$$

The update operator

$$
\begin{aligned}
p_L(\boldsymbol{y}) \;=\; \operatorname{prox}_g^L\left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right) \;&:=\; \operatorname*{argmin}_{\boldsymbol{x}} \frac{L}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right)\right\|_2^2 + g(\boldsymbol{x}) \\
&=\; \operatorname*{argmin}_{\boldsymbol{x}} \frac{L}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{x}_{k-1} - \frac{\nabla f(\boldsymbol{x}_{k-1})}{L}\right)\right\|_2^2 + g(\boldsymbol{x}) \quad \text{put } \boldsymbol{y} = \boldsymbol{x}_{k-1}
\end{aligned}
$$

- $p_L(\boldsymbol{y})$ can be seen as an update operator of $\mathcal{P}$ at a point $\boldsymbol{y}$
- $p_L(\boldsymbol{y})$ forms the basis of the iterative algorithm of proximal gradient method.
- $p_L(\boldsymbol{y})$ itself is an optimization problem that is generally not proximable:
  - $p_L(\boldsymbol{y})$ may have no closed-form solution.
  - $p_L(\boldsymbol{y})$ may be expensive to compute
- When $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$, p_L is the update step of ISTA (Iterative Soft Thresholding Algorithm). In this case p_L has a closed-form solution. Details here.

On solving $p_L(\boldsymbol{y}) \;=\; \mathrm{prox}_g^L\left(\boldsymbol{y} - \dfrac{\nabla f(\boldsymbol{y})}{L}\right) \;\coloneqq\; \underset{\boldsymbol{x}}{\mathrm{argmin}}\; \dfrac{L}{2}\left\|\boldsymbol{x} - \left(\boldsymbol{y} - \dfrac{\nabla f(\boldsymbol{y})}{L}\right)\right\|_2^2 + g(\boldsymbol{x}).$

- How to solve p_L: by 1st-order optimality condition
  - p_L contains non-differentiable g, so we have to use subgradient instead of gradient
- **Subgradient 1st-order optimality condition**
  A point $\boldsymbol{x}^*$ is the minimizer of $p_L(\boldsymbol{y})$ if and only if (necessary & sufficient)

$$
\begin{aligned}
& \boldsymbol{x}^* \;\in\; \partial p_L(\boldsymbol{y}) && \text{subgradient 1st-order optimality condition}\\
\iff\; & \boldsymbol{x}^* \;=\; \partial p_L(\boldsymbol{y}) && \text{strictly convex cost so solution is unique}\\
\iff\; & \mathbf{0} \;\in\; \partial\left\{\frac{L}{2}\left\|\boldsymbol{x}^* - \left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right)\right\|_2^2 + g(\boldsymbol{x}^*)\right\} && \text{write down the subdifferential}\\
\iff\; & \mathbf{0} \;\in\; \partial\left\{\frac{L}{2}\left\|\boldsymbol{x}^* - \left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right)\right\|_2^2\right\} + \partial g(\boldsymbol{x}^*) && \mathrm{dom} g \subset \mathrm{int dom} f \text{ so } \partial \text{ is a linear operator}\\
\iff\; & \mathbf{0} \;\in\; \nabla\left\{\frac{L}{2}\left\|\boldsymbol{x}^* - \left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right)\right\|_2^2\right\} + \partial g(\boldsymbol{x}^*) && \partial \text{ on a differentiable function reduces to } \nabla\\
\iff\; & \mathbf{0} \;\in\; L\left\{\boldsymbol{x}^* - \left(\boldsymbol{y} - \frac{\nabla f(\boldsymbol{y})}{L}\right)\right\} + \partial g(\boldsymbol{x}^*) && \text{compute } \nabla\\
\iff\; & \mathbf{0} \;\in\; L(\boldsymbol{x}^* - \boldsymbol{y}) + \nabla f(\boldsymbol{y}) + \partial g(\boldsymbol{x}^*)
\end{aligned}
$$

Let $\gamma(\boldsymbol{y}) \in \partial g(\boldsymbol{y})$, then one has $\boldsymbol{x} = p_L(\boldsymbol{y})$ if and only if

- there exists $\gamma(\boldsymbol{y}) \in \partial g(\boldsymbol{y})$ is the sub-differential of g s.t. $\nabla f(\boldsymbol{y}) + L(\boldsymbol{x} - \boldsymbol{y}) + \gamma(\boldsymbol{y}) = \mathbf{0}$

A key lemma

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

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

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

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

Now we bound F using convexity, first

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

$(*)$ f is convex and differentiable

$(**)$ g is convex and not differentiable

Hence (1) becomes

$$F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \geq f(\boldsymbol{y}) + g(p_L(\boldsymbol{y})) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle + \langle \gamma(\boldsymbol{y}), \boldsymbol{x} - p_L(\boldsymbol{y})\rangle - Q(p_L(\boldsymbol{y}); \boldsymbol{y}) \tag{1'}$$

Next we write down Q: by the definition of Q, put $\boldsymbol{x} = p_L(\boldsymbol{y})$, we have

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

Now $(1')$ becomes

$$\begin{aligned} F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \;\geq\; & f(\boldsymbol{y}) + g(p_L(\boldsymbol{y})) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle + \langle \gamma(\boldsymbol{y}), \boldsymbol{x} - p_L(\boldsymbol{y})\rangle \\ & - f(\boldsymbol{y}) - \langle p_L(\boldsymbol{y}) - \boldsymbol{y}, \nabla f(\boldsymbol{y})\rangle - \tfrac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 - g(p_L(\boldsymbol{y})) \end{aligned} \tag{1''}$$

$$F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \quad \geq \quad \begin{array}{l} f(\boldsymbol{y}) \quad +g(p_L(\boldsymbol{y})) \quad + \underline{\langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle} + \langle \gamma(\boldsymbol{y}), \boldsymbol{x} - p_L(\boldsymbol{y}) \rangle \\ -f(\boldsymbol{y}) \quad \underline{-\langle p_L(\boldsymbol{y}) - \boldsymbol{y}, \nabla f(\boldsymbol{y}) \rangle} - \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 \quad -g(p_L(\boldsymbol{y})) \end{array} \qquad (1'')$$

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

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

$$\iff \quad F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \quad \geq \quad \langle \underbrace{L(\boldsymbol{y} - \boldsymbol{x})}, \boldsymbol{x} - p_L(\boldsymbol{y}) \rangle - \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 \qquad \text{1st-order optimality}$$

$$\iff \quad F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \quad \geq \quad L\langle \boldsymbol{y} - p_L(\boldsymbol{y}), \boldsymbol{x} - p_L(\boldsymbol{y}) \rangle - \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 \qquad \text{put } \boldsymbol{x} = p_L(\boldsymbol{y})$$

Tricky step: $(a - c)(b - c) = (b - a)(c - b) + (c - b^2)$, then

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

Therefore

$$F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \quad \leq \quad L\langle \boldsymbol{x} - \boldsymbol{y}, p_L(\boldsymbol{y}) - \boldsymbol{y} \rangle + \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2. \quad \square$$

ISTA and FISTA

ISTA / Proximal gradient descent

- Iterate the following step

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

  with $L = L_f$ known.

- If L_f is unknown, backtracking can be used.
- ISTA $\in$ proximal gradient algorithm. Details.

FISTA / accelerated proximal gradient method

- Iterate the following step
  - $\boldsymbol{x}_k = p_L(\boldsymbol{y}_{k-1})$
  - $t_{k+1} = \dfrac{1+\sqrt{1+4t_k^2}}{2}$
  - $\boldsymbol{y}_{k+1} = \boldsymbol{x}_k + \dfrac{t_k - 1}{t_{k+1}}(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})$
- FISTA = Nesterov's acceleration applied on ISTA
- FISTA $\in$ accelerated proximal gradient algorithm
- See here for the convergence of FISTA

Convergence of proximal gradient descent / ISTA

- Proximal gradient descent is *monotone*:
  1. $Q(\boldsymbol{x}; \boldsymbol{y}) = f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle + \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ so if $\boldsymbol{x} = \boldsymbol{y}$ then $Q_L(\boldsymbol{x}; \boldsymbol{y}) = F(\boldsymbol{x})$
  2. $p_L(\boldsymbol{y}) \coloneqq \operatorname*{argmin}_{\boldsymbol{x}} Q(\boldsymbol{x}; \boldsymbol{y})$ by definition
  3. Q is local quadratic over-estimator of F, so $F(\boldsymbol{x}) \leq Q(\boldsymbol{x})$

  Hence we have

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

- If $F(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$, 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^*.

Theorem (Convergence rate of ISTA using constant step size)

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

- Here the gradient stepsize is $\frac{1}{L}$
- As ISTA $\in$ 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 Beck & Teboulle 2009.

Proof ... 1/3

$$F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \;\geq\; \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 + L\langle \boldsymbol{y} - \boldsymbol{x}, p_L(\boldsymbol{y}) - \boldsymbol{y}\rangle \text{ (key lemma)}$$

Proof In the key lemma, put $\boldsymbol{x} = \boldsymbol{x}^*$ and $\boldsymbol{y} = \boldsymbol{x}_k$, then based on ISTA, $p_L(\boldsymbol{y}) = \boldsymbol{x}_{k+1}$:

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

Based on observation, consider completing squares

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

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

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

Proof ... 2/3

$$F(\boldsymbol{x}) - F(p_L(\boldsymbol{y})) \;\geq\; \frac{L}{2}\|p_L(\boldsymbol{y}) - \boldsymbol{y}\|_2^2 + L\langle \boldsymbol{y} - \boldsymbol{x}, p_L(\boldsymbol{y}) - \boldsymbol{y}\rangle \text{ (key lemma)}$$

In the key lemma, put $\boldsymbol{x} = \boldsymbol{y} = \boldsymbol{x}_k$, $p_L(\boldsymbol{y}) = \boldsymbol{x}_{k+1}$

$$\begin{aligned}
& F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1}) && \geq && \frac{L}{2}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 && \\
\iff\quad & \frac{2}{L}\Big(F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1})\Big) && \geq && \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 && \\
\iff\quad & \frac{2}{L}k\Big(F(\boldsymbol{x}_k) - F(\boldsymbol{x}_{k+1})\Big) && \geq && k\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|_2^2 && \text{Tricky: multiply } k \\
\iff\quad & \frac{2}{L}\sum_{i=0}^{k-1} i\Big(F(\boldsymbol{x}_i) - F(\boldsymbol{x}_{i+1})\Big) && \geq && \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 && \text{summation from } k=0 \text{ to } k-1 \\
\iff\quad & \frac{2}{L}\sum_{i=0}^{k-1}\Big\{i\Big(F(\boldsymbol{x}_i) - F(\boldsymbol{x}_{i+1})\Big) + F(\boldsymbol{x}_{i+1}) - F(\boldsymbol{x}_{i-1})\Big\} && \geq && \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 && \text{tricky step} \\
\iff\quad & \frac{2}{L}\sum_{i=0}^{k-1}\Big\{iF(\boldsymbol{x}_i) - (i+1)F(\boldsymbol{x}_{i+1}) + F(\boldsymbol{x}_{i+1})\Big\} && \geq && \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 && \\
\iff\quad & \frac{2}{L}\Bigg(\underbrace{\sum_{i=0}^{k-1}\Big(iF(\boldsymbol{x}_i) - (i+1)F(\boldsymbol{x}_{i+1})\Big)}_{=-kF(\boldsymbol{x}_k) \text{ by telescoping sum}} + \sum_{i=0}^{k-1} F(\boldsymbol{x}_{i+1})\Bigg) && \geq && \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 &&
\end{aligned}$$

Therefore

$$\frac{2}{L}\Big(-kF(\boldsymbol{x}_k) + \sum_{i=0}^{k-1} F(\boldsymbol{x}_{i+1})\Big) \;\geq\; \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 \qquad (3)$$

Proof ... 3/3

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

$$\frac{2}{L}\Big(-kF(\boldsymbol{x}_k) + \sum_{i=0}^{k-1} F(\boldsymbol{x}_{i+1})\Big) \geq \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 \quad (3)$$

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

$$\frac{2}{L}\Big(-kF(\boldsymbol{x}_k) + \sum_{i=0}^{k-1} F(\boldsymbol{x}_{i+1})\Big) \geq \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 \quad (3)$$

$$\frac{2}{L}\big(kF^* - kF(\boldsymbol{x}_k)\big) \geq \sum_{i=0}^{k-1} i\|\boldsymbol{x}_{i+1} - \boldsymbol{x}_i\|_2^2 + \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2 - R_0^2 \quad (2)+(3)$$

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

$$\leq R_0^2$$

$$F(\boldsymbol{x}_k) - F^* \leq \frac{LR_0^2}{2k} \quad \square$$

End of document