The Proximal Gradient Method

Instructor: Jin Zhang

Department of Mathematics
Southern University of Science and Technology

Fall 2023

Contents

1. The Composite Model
2. The Proximal Gradient Method
3. Analysis of the PGM—The Nonconvex Case
4. Analysis of the Proximal Gradient Method—The Convex Case
5. Analysis of the Proximal Gradient Method—The Strongly Convex Case
6. FISTA

1. The Composite Model

The Composite Model

$$\min_{x \in \mathbb{R}^n} \{F(x) \equiv f(x) + g(x)\} \tag{1}$$

Standing Assumption (SA):

(A). $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper closed and convex.

(B). $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and closed, $\mathrm{dom}(f)$ is convex, $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$, and f is L_f-smooth over $\mathrm{int}(\mathrm{dom}(f))$.

(C). The optimal set of problem (1) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

Three special cases of the general model (1)

- **Smooth unconstrained minimization.** If $g \equiv 0$ and dom$(f) = \mathbb{R}^n$, then (1) reduces to

$$\min_{x \in \mathbb{R}^n} f(x),$$

  where $f : \mathbb{R}^n \to \mathbb{R}$ is L_f-smooth over $\mathbb{R}^n$.

- **Convex constrained smooth minimization.** If $g = \delta_C$, where $C \subset \mathbb{R}^n$ is a nonempty closed and convex set, then (1) amounts to

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{s.t. } x \in C \qquad \text{or } \min_{x \in C} f(x),$$

  where f is L_f-smooth over int(dom(f)) and $C \subset$ int(dom(f)).

- **l_1-regularized minimization.** If $g(x) = \lambda \|x\|_1$ for some $\lambda > 0$ and dom$(f) = \mathbb{R}^n$, then (1) amounts to

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) + \lambda \|x\|_1 \right\},$$

  where $f : \mathbb{R}^n \to \mathbb{R}$ is L_f-smooth over $\mathbb{R}^n$.

Motivation:

- Solve the Smooth unconstrained model by the gradient method:

$$
\begin{aligned}
x^{k+1} =& x^k - t_k \nabla f\left(x^k\right) \\
=& \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ f\left(x^k\right) + \langle \nabla f\left(x^k\right), x - x^k \rangle + 0 + \frac{1}{2t_k} \left\| x - x^k \right\|^2 \right\}
\end{aligned}
$$

- Solve the Convex constrained smooth model by the projected gradient method:

$$
\begin{aligned}
x^{k+1} =& P_C\left(x^k - t_k \nabla f\left(x^k\right)\right) \\
=& \underset{x \in C}{\arg\min} \left\{ f\left(x^k\right) + \langle \nabla f\left(x^k\right), x - x^k \rangle + \frac{1}{2t_k} \left\| x - x^k \right\|^2 \right\} \\
=& \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ f\left(x^k\right) + \langle \nabla f\left(x^k\right), x - x^k \rangle + \delta_C(x) + \frac{1}{2t_k} \left\| x - x^k \right\|^2 \right\}
\end{aligned}
$$

It's natural to generalize the above idea to the more general model (1):

$$
\begin{aligned}
x^{k+1} &= \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ f\left(x^k\right) + \langle \nabla f\left(x^k\right), x - x^k \rangle + g(x) + \frac{1}{2t_k} \left\| x - x^k \right\|^2 \right\} \\
&= \underset{x \in \mathbb{R}^n}{\arg\min} \left\{ t_k g(x) + \frac{1}{2} \left\| x - \left(x^k - t_k \nabla f\left(x^k\right) \right) \right\|^2 \right\} \\
&= \mathsf{prox}_{t_k g}\left(x^k - t_k \nabla f\left(x^k\right) \right).
\end{aligned}
$$

From now on, we will take the stepsize as $t_k = \frac{1}{L_k}$, leading to the following description.

The Proximal Gradient Method

- **Initialization:** pick $x^0 \in \mathrm{int}(\mathrm{dom} f)$.
- **General Step:** for any $k = 0, 1, 2, \cdots$ execute the following steps:
  (a). pick $L_k > 0$;
  (b). set $x^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(x^k - \frac{1}{L_k}\nabla f(x^k)\right)$.

Definitions:

Prox-grad Operator: Suppose that f and g satisfy (A) and (B) of SA and let $L > 0$. Then $T_L^{f,g}$: int(dom f) $\to \mathbb{R}^n$ is the prox-grad operator associated with f, g, L defined by

$$T_L^{f,g}(x) = \text{prox}_{\frac{1}{L}g}\left(x - \frac{1}{L}\nabla f(x)\right) \text{ for any } x \in \text{int}(\text{dom} f).$$

Gradient Mapping: Suppose that f and g satisfy (A) and (B) of SA and let $L > 0$. Then $G_L^{f,g}$: int(dom f) $\to \mathbb{R}^n$ is the gradient mapping associated with f, g, L defined by

$$G_L^{f,g}(x) = L\left(x - T_L^{f,g}(x)\right) \text{ for any } x \in \text{int}(\text{dom} f).$$

3. Analysis of the PGM—The Nonconvex Case
3.1 Sufficient Decrease

Notations

We set $T_L \equiv T_L^{f,g}$ and $G_L \equiv G_L^{f,g}$ when there's no ambiguity.

Lemma: (sufficient decrease lemma).

Suppose that f and g satisfy properties (A) and (B) of SA. Let $F = f + g$. Then for any $x \in \operatorname{int}(\operatorname{dom} f)$ and $L \in (\frac{L_f}{2}, \infty)$ the following inequality holds:

$$F(x) - F\left(T_L(x)\right) \geq \frac{L - \frac{L_f}{2}}{L^2} \left\|G_L(x)\right\|^2 .$$

Especially,

$$F(x) - F\left(T_{L_f}(x)\right) \geq \frac{1}{2L_f} \left\|G_{L_f}(x)\right\|^2 .$$

The gradient mapping G_L “measures” the optimality.

Theorem

Let f and g satisfy properties (A) and (B) of SA and let $L > 0$. Then

(a). $G_L^{f,g_0}(x) = \nabla f(x)$ *for any* $x \in$ *int*$(\mathrm{dom} f)$, *where* $g_0(x) \equiv 0$;

(b). *for* $x^* \in$ *int*$(\mathrm{dom} f)$, *it holds that* $G_L^{f,g}(x^*) = 0$ *if and only if* x^* *is a stationary point of problem (1).*

Corollary (necessary and sufficient optimality condition under convexity).

Let f and g satisfy properties (A) and (B) of SA and let $L > 0$. Suppose that in addition f is convex. Then for $x^* \in \text{dom}(g)$, it holds that $G_L^{f,g}(x^*) = 0$ if and only if x^* is an optimal solution of problem (1).

The next result establishes monotonicity properties of $\|G_L(x)\|$ w.r.t. the parameter L.

Theorem (monotonicity of the gradient mapping)

Suppose that f and g satisfy properties (A) and (B) of SA. Suppose that $L_1 \geq L_2 > 0$. Then for any $x \in$ int(dom$f$),

$$\|G_{L_1}(x)\| \geq \|G_{L_2}(x)\|, \quad \frac{\|G_{L_1}(x)\|}{L_1} \leq \frac{\|G_{L_2}(x)\|}{L_2}.$$

Lemma: Lipschitz continuity of the gradient mapping

Suppose that f and g satisfy properties (A) and (B) of SA. Then

(a). $\|G_L(x) - G_L(y)\| \leq (2L + L_f)\|x - y\|$ for any $x, y \in \text{int}(\text{dom} f)$.

(b). $\left\|G_{L_f}(x) - G_{L_f}(y)\right\| \leq 3L_f\|x - y\|$ for any $x, y \in \text{int}(\text{dom} f)$.

Lemma: firm nonexpansivity of $\frac{3}{4L_f}G_{L_f}$

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L_f-smooth $(L_f > 0)$ over $\mathbb{R}^n$, and let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper closed and convex. Then

(a). the gradient mapping G_{L_f} satisfies the relation

$$\langle G_{L_f}(x) - G_{L_f}(y), x - y\rangle \geq \frac{3}{4L_f}\left\|G_{L_f}(x) - G_{L_f}(y)\right\|^2$$

for any $x, y \in \mathbb{R}^n$.

(b). $\left\|G_{L_f}(x) - G_{L_f}(y)\right\| \leq \frac{4L_f}{3}\|x - y\|$ for any $x, y \in \mathbb{R}^n$.

Lemma: monotonicity of the norm of the gradient mapping with respect to the pro-grad operator

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L_f-smooth $(L_f > 0)$ over $\mathbb{R}^n$, and let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper closed and convex. Then

$$\left\| G_{L_f}\left(T_{L_f}(x)\right) \right\| \leq \left\| G_{L_f}(x) \right\|.$$

Stepsize Strategies

- **Constant.** $L_k = \bar{L} \in \left(\frac{L_f}{2}, \infty\right)$ for all k.
- **Backtracking procedure B1.**
  The procedure requires three parameters (s, γ, η), where $s > 0$, $\gamma \in (0, 1)$ and $\eta > 1$.
  The choice of L_k is done as follows.
  First, L_k is set to be equal to the initial guess s.
  Then, while

$$F(x^k) - F\left(T_{L_k}(x^k)\right) < \frac{\gamma}{L_k} \left\| G_{L_k}(x^k) \right\|^2,$$

  we set $L_k := \eta L_k$.

- Backtracking procedure B1 (to be continued).
  In other words, L_k is chosen as $L_k = s\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$F(x^k) - F\left(T_{s\eta^{i_k}}(x^k)\right) \geq \frac{\gamma}{s\eta^{i_k}} \left\|G_{s\eta^{i_k}}\right\|^2.$$

  is satisfied.

Remark

Under SA,

1. the backtracking procedure is finite when $L_k \geq \frac{L_f}{2(1-\gamma)}$.
2. Compute the finite upper bound on L_k:

$$L_k \leq \max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}.$$

Lemma (sufficient decrease of the PGM).

Suppose that SA holds. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize defined by $L_k = \bar{L} \in (\frac{L_f}{2}, \infty)$ or with a stepsize chosen by the backtracking procedure B1 with parameter (s, γ, η), where $s > 0$, $\gamma \in (0,1)$, $\eta > 1$. Then for any $k \geq 0$,

$$F(x^k) - F(x^{k+1}) \geq M \left\| G_d(x^k) \right\|^2,$$

where

$$M = \begin{cases} \frac{\bar{L} - \frac{L_f}{2}}{\left(\bar{L}\right)^2}, & \text{constant stepsize,} \\ \frac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}, & \text{backtracking,} \end{cases}$$

and

$$d = \begin{cases} \bar{L}, & \text{constant stepsize,} \\ s, & \text{backtracking.} \end{cases}$$

Theorem (convergence of the PGM-nonconvex case.)

Suppose that SA holds and let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) either with a constant stepsize defined by $L_k = \bar{L} \in (\frac{L_f}{2}, \infty)$ or with a stepsize chosen by the backtracking procedure B1 with parameters (s, γ, η), where $s > 0$, $\gamma \in (0, 1)$, $\eta > 1$. Then

(a). the sequence $\{F(x^k)\}_{k\geq 0}$ is nonincreasing. In addition, $F(x^{k+1}) < F(x^k)$ if and only if x^k is not a stationary of problem (1);

(b). $G_d(x^k) \to 0$ as $k \to \infty$, where

$$d = \begin{cases} \bar{L}, & \text{constant stepsize,} \\ s, & \text{backtracking.} \end{cases}$$

(c).

$$\min_{n=0,\cdots,k} \left\| G_d(x^k) \right\| \leq \frac{\sqrt{F(x^0) - F_{\text{opt}}}}{\sqrt{M(k+1)}},$$

where

$$M = \begin{cases} \dfrac{\bar{L} - \frac{L_f}{2}}{\left(\bar{L}\right)^2}, & \text{constant stepsize,} \\ \dfrac{\gamma}{\max\left\{s, \frac{\eta L_f}{2(1-\gamma)}\right\}}, & \text{backtracking,} \end{cases}$$

(d). all limit points of the sequence $\{x^k\}_{k \geq 0}$ are stationary points of problem (1).

Theorem (fundamental prox-grad inequality).

Suppose that f and g satisfy properties (A) and (B) of SA. For any $x \in \mathbb{R}^n$, $y \in \operatorname{int}(\operatorname{dom} f)$ and $L > 0$ satisfying

$$f\left(T_L(y)\right) \leq f(y) + \langle \nabla f(y), T_L(y) - y \rangle + \frac{L}{2}\left\|T_L(y) - y\right\|^2,$$

it holds that

$$F(x) - F\left(T_L(y)\right) \geq \frac{L}{2}\left\|x - T_L(y)\right\|^2 - \frac{L}{2}\left\|x - y\right\|^2 + l_f(x, y),$$

where

$$l_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Remark

Suppose that f and g satisfy properties (A) and (B) of SA. For any $x \in \mathbb{R}^n$, $y \in \mathrm{int}(\mathrm{dom} f)$, the inequality

$$F(x) - F\left(T_{L_f}(y)\right) \geq \frac{L_f}{2}\left\|x - T_{L_f}(y)\right\|^2 - \frac{L_f}{2}\|x - y\|^2 + l_f(x, y)$$

holds.

Corollary (sufficient decrease lemma-second version).

Suppose that f and g satisfy properties (A) and (B) of SA. For any $x \in \mathrm{int}(\mathrm{dom} f)$ for which

$$f\left(T_L(x)\right) \leq f(x) + \langle \nabla f(x), T_L(x) - x \rangle + \frac{L}{2} \left\| T_L(x) - x \right\|^2,$$

it holds that

$$F(x) - F\left(T_L(x)\right) \geq \frac{1}{2L} \left\| G_L(x) \right\|^2.$$

Stepsize Strategies

- **Constant.** $L_k = L_f$ for all k.
- **Backtracking procedure B2.**
  The procedure requires two parameters (s, η), where $s > 0$, and $\eta > 1$.
  Define $L_{-1} = s$.
  At iteration $k (k \geq 0)$ the choice of L_k is done as follows.
  First, L_k is set to be equal to L_{k-1}.
  Then, while

$$f\left(T_{L_k}(x^k)\right) > f(x^k) + \langle \nabla f(x^k), T_{L_k}(x^k) - x^k \rangle + \frac{L_k}{2} \left\| T_{L_k}(x^k) - x^k \right\|^2,$$

  we set $L_k := \eta L_k$.

- **Backtracking procedure B2 (to be continued).**
  In other words, L_k is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f\left(T_{L_{k-1}\eta^{i_k}}(x^k)\right) \leq f(x^k) + \langle \nabla f(x^k), T_{L_{k-1}\eta^{i_k}}(x^k) - x^k \rangle + \frac{L_k}{2}\left\|T_{L_{k-1}\eta^{i_k}}(x^k) - x^k\right\|^2$$

  is satisfied.

Remark (upper and lower bounds on L_k)

Under SA, the constants L_k that the backtracking procedure B2 produces satisfy the following bounds for all $k \geq 0$:

$$s \leq L_k \leq \max\{\eta L_f, s\},$$

which can be rewritten as $\beta L_f \leq L_k \leq \alpha L_f$, where

$$\alpha = \begin{cases} 1, & \text{constant,} \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking,} \end{cases} \qquad \beta = \begin{cases} 1, & \text{constant,} \\ \frac{s}{L_f}, & \text{backtracking.} \end{cases}$$

Remark (monotonicity of the proximal gradient method)

Under SA and either of two stepsize rules, for any $k \geq 0$, we obtain the inequality

$$F(x^k) - F\left(x^{k+1}\right) \geq \frac{L_k}{2}\left\|x^k - x^{k+1}\right\|^2.$$

Theorem ($O(\frac{1}{k})$ rate of convergence of proximal gradient).

Suppose that SA holds and that in addition f is convex. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B2. Then for any $x^* \in X^*$ and $k \geq 0$,

$$F(x^k) - F_{\text{opt}} \leq \frac{\alpha L_f \left\|x^0 - x^*\right\|^2}{2k},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Definition (Fejér monotonicity)

A sequence $\{x^k\}_{k\geq 0} \subseteq \mathbb{R}^n$ is called Fejér monotone with respect to a set $S \subseteq \mathbb{R}^n$ if

$$\left\|x^{k+1} - y\right\| \leq \left\|x^k - y\right\| \text{ for all } k \geq 0 \text{ and } y \in S.$$

Theorem (convergence under Fejér monotonicity)

Let $\{x^k\}_{k\geq 0} \subseteq \mathbb{R}^n$ be a sequence, and let S be a set satisfying $D \subseteq S$, where D is the set comprising all the limit points of $\{x^k\}_{k\geq 0}$. If $\{x^k\}_{k\geq 0}$ is Fejér monotone with respect to S, then it converges to a point in D.

Theorem (Fejér monotonicity of the sequence generated by the proximal gradient method).

Suppose that SA holds and that in addition f is convex. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B2. Then for any $x^* \in X^*$ and $k \geq 0$,

$$\left\| x^{k+1} - x^* \right\| \leq \left\| x^k - x^* \right\|.$$

Theorem (convergence of the sequence generated by the proximal gradient method).

Suppose that SA holds and that in addition f is convex. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B2. Then the sequence $\{x^k\}_{k\geq 0}$ converges to an optimal solution of problem (1).

Theorem ($O(\frac{1}{k})$ rate of convergence of the minimal norm of the gradient mapping).

Suppose that SA holds and that in addition f is convex. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B2. Then for any $x^* \in X^*$ and $k \geq 1$,

$$\min_{n=0,1,\cdots,k} \left\| G_{\alpha L_f}(x^n) \right\| \leq \frac{2\alpha^{1.5} L_f \left\| x^0 - x^* \right\|}{\sqrt{\beta} k},$$

where $\alpha = \beta = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}, \beta = \frac{s}{L_f}$ if the backtracking rule is employed.

Theorem ($O(\frac{1}{k})$ rate of convergence of the norm of the gradient mapping under the constant stepsize rule).

Suppose that SA holds and that in addition f is convex and L_f-smooth over $\mathbb{R}^n$. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$. Then for any $x^* \in X^*$ and $k \geq 0$,

(a).

$$\left\|G_{L_f}(x^{k+1})\right\| \leq \left\|G_{L_f}(x^k)\right\|;$$

(b).

$$\left\|G_{L_f}(x^k)\right\| \leq \frac{2L_f\left\|x^0 - x^*\right\|}{k+1}.$$

5. Analysis of the PGM—The Strongly Convex Case

Theorem (linear rate of convergence of the proximal gradient method—strongly convex case).

Suppose that SA holds and that in addition f is σ-convex. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B2, and let x^* be the unique minimum of problem (1). Then for any $k \geq 0$,

(a). $\left\|x^{k+1} - x^*\right\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right) \left\|x^k - x^*\right\|^2$;

(b). $\left\|x^{k+1} - x^*\right\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)^k \left\|x^0 - x^*\right\|^2$;

(c). $F(x^{k+1}) - F_{\text{opt}} \leq \frac{\alpha L_f}{2} \left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1} \left\|x^0 - x^*\right\|^2$,

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

no strong:just Fejer monotone

The Composite Model

$$\min_{x \in \mathbb{R}^n} \{F(x) \equiv f(x) + g(x)\} \tag{2}$$

Assumption 2:

(A). $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper closed and convex.

(B). $f : \mathbb{R}^n \to \mathbb{R}$ is L_f-smooth and convex.

(C). The optimal set of problem (1) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

Fast proximal gradient method
Fast iterative shrinkage-thresholding algorithm (FISTA)

FISTA

- **Input:** (f, g, x^0), where f and g satisfy properties (A) and (B) in Assumption 2 and $x^0 \in \mathbb{R}^n$.
- **Initialization:** set $y^0 = x^0$ and $t_0 = 1$.
- **General Step:** for any $k = 0, 1, 2, \cdots$ execute the following steps:
  (a). pick $L_k > 0$;
  (b). set $x^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(y^k - \frac{1}{L_k}\nabla f(y^k)\right)$;
  (c). set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d). compute $y^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(x^{k+1} - x^k\right)$.

Stepsize Strategies

- **Constant.** $L_k = L_f$ for all k.
- **Backtracking procedure B3.**
  The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration $k (k \geq 0)$ the choice of L_k is done as follows:
  First, L_k is set to be equal to L_{k-1}.
  Then, while
  $$f\left(T_{L_k}(y^k)\right) > f(y^k) + \langle \nabla f(y^k), T_{L_k}(y^k) - y^k \rangle + \frac{L_k}{2} \left\| T_{L_k}(y^k) - y^k \right\|^2,$$
  we set $L_k := \eta L_k$.

- Backtracking procedure B3 (to be continued).
In other words, L_k is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which the condition

$$f\left(T_{L_{k-1}\eta^{i_k}}(y^k)\right) \leq f(y^k) + \langle \nabla f(y^k), T_{L_{k-1}\eta^{i_k}}(y^k) - y^k \rangle + \frac{L_k}{2} \left\| T_{L_{k-1}\eta^{i_k}}(y^k) - y^k \right\|^2$$

is satisfied.

Remark (upper and lower bounds on L_k)

Under Assumption 2, the constants L_k that the backtracking procedure B3 produces satisfy the following bounds for all $k \geq 0$:

$$s \leq L_k \leq \max\{\eta L_f, s\},$$

which can be rewritten as $\beta L_f \leq L_k \leq \alpha L_f$, where

$$\alpha = \begin{cases} 1, & \text{constant,} \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking,} \end{cases} \qquad \beta = \begin{cases} 1, & \text{constant,} \\ \frac{s}{L_f}, & \text{backtracking.} \end{cases}$$

Lemma

Let $\{t_k\}_{k\geq 0}$ be the sequence defined by

$$t_0 = 1,\, t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}, \quad k \geq 0.$$

Then $t_k \geq \frac{k+2}{2}$ for all $k \geq 0$.

Theorem ($O(\frac{1}{k^2})$ rate of convergence of FISTA).

Suppose that Assumption 2 holds. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by FISTA for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B3. Then for any $x^* \in X^*$ and $k \geq 0$,

$$F(x^k) - F_{\text{opt}} \leq \frac{2\alpha L_f \left\| x^0 - x^* \right\|^2}{(k+1)^2},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Remark (alternative choice for t_k).

A close inspection of the proof of Theorem ($O(\frac{1}{k^2})$ rate of convergence of FISTA) reveals that the result is correct if $\{t_k\}_{k\geq 0}$ is any sequence satisfying the following two properties for any $k \geq 0$:

(a). $t_k \geq \frac{k+2}{2}$;

(b). $t_{k+1}^2 - t_{k+1} \leq t_k^2$.

The choice $t_k = \frac{k+2}{2}$ also satisfies these two properties. The validity of (a) is obvious; to show (b), note that

$$
\begin{aligned}
t_{k+1}^2 - t_{k+1} = t_{k+1}(t_{k+1} - 1) = \frac{k+3}{2} \cdot \frac{k+1}{2} &= \frac{k^2+4k+3}{4} \\
\leq \frac{k^2+4k+4}{4} = \frac{(k+2)^2}{4} &= t_k^2.
\end{aligned}
$$

Remark

Note that FISTA has an $O(\frac{1}{k^2})$ rate of convergence in function values, while the proximal gradient method has an $O(\frac{1}{k})$ rate of convergence. This improvement was achieved despite the fact that the dominant computational steps at each iteration of both methods are essentially the same: one gradient evaluation and one prox computation.

Example (l_1-regularized minimization)

Consider the following model:

$$\min_{x\in\mathbb{R}^n} f(x) + \lambda \|x\|_1 ,$$

where $\lambda > 0$ and $f : \mathbb{R}^n \to \mathbb{R}$ is assumed to be convex and L_f-smooth.

The proximal gradient method (or iterative shrinkage-thresholding algorithm (ISTA)) with constant stepsize $\frac{1}{L_f}$:

$$x^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(x^k - \frac{1}{L_f}\nabla f(x^k)\right).$$

Recall that $\mathcal{T}_\alpha : \mathbb{R}^n \to \mathbb{R}^n_+$ is the soft thresholding operator associated with $\alpha > 0$ defined by

$$\mathcal{T}_\alpha(x) \equiv ([|x_i| - \alpha]_+ \mathrm{sgn}(x))_{i=1}^n .$$

The fast proximal gradient method (or fast iterative shrinkage-thresholding algorithm (FISTA)) with constant stepsize $\frac{1}{L_f}$:

(a). set $x^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(y^k - \frac{1}{L_f}\nabla f(y^k)\right)$;

(b). set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

(c). compute $y^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}}\left(x^{k+1} - x^k\right)$.

Example (l_1-regularized least squares).

Consider

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1 ,$$

where $\lambda > 0$ and $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$. Note that the function $\frac{1}{2} \|Ax - b\|_{2,2}$ is L-smooth with

$$L = \left\|A^T A\right\|_2^2 = \lambda_{\max}(A^T A).$$

The update step of ISTA:

$$x^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}} \left(x^k - \frac{1}{L_k} A^T (Ax^k - b) \right)$$

The update step of FISTA:

(a). set $x^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}} \left(y^k - \frac{1}{L_k} A^T (Ay^k - b) \right)$;

(b). set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;

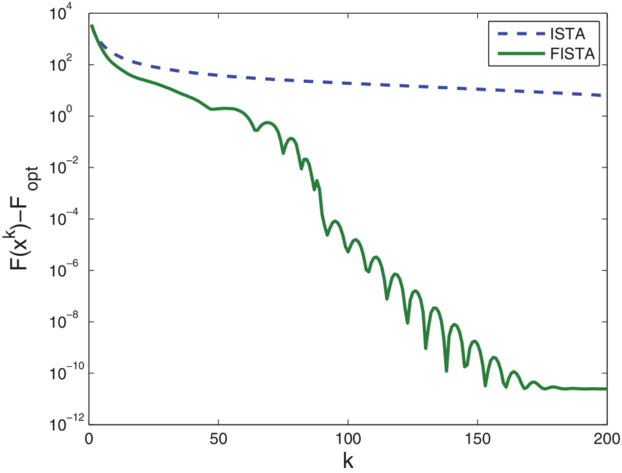
(c). compute $y^{k+1} = x^{k+1} + \frac{t_k - 1}{t_{k+1}} \left(x^{k+1} - x^k \right)$.

Instance performance:

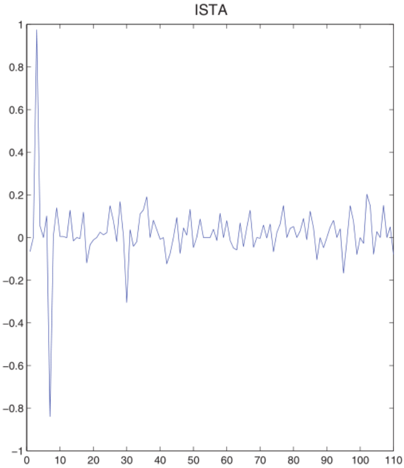
Let

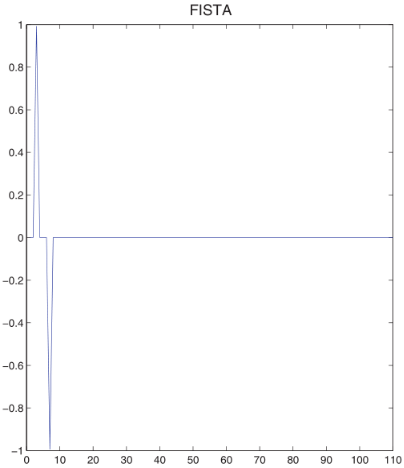
- $\lambda = 1$,
- $A \in \mathbb{R}^{100\times 110}$ where the components of A were independently generated using a standard normal distribution.
- The "true" vector is $x_{\text{true}} = e_3 - e_7$.
- $b = Ax_{\text{true}}$.

Let the initial vector $x = e$. The distances to optimality in terms of function values of the sequences generated by the two methods as a function of the iteration index are plotted:



the vectors that were obtained by 200 iterations of ISTA and FISTA:





- **Input:** (f, g, x^0), where f and g satisfy properties (A) and (B) in Assumption 2 and $x^0 \in \mathbb{R}^n$.
- **Initialization:** set $y^0 = x^0$ and $t_0 = 1$.
- **General Step:** for any $k = 0, 1, 2, \cdots$ execute the following steps:
  (a). pick $L_k > 0$;
  (b). set $z^k = \text{prox}_{\frac{1}{L_k} g}\left(y^k - \frac{1}{L_k}\nabla f(y^k)\right)$;
  (c). choose $x^{k+1} \in \mathbb{R}^n$ such that $F(x^{k+1}) \leq \min\{F(z^k), F(x^k)\}$
  (d). set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (e). compute $y^{k+1} = x^{k+1} + \frac{t_k}{t_{k+1}}\left(z^k - x^{k+1}\right) + \frac{t_k - 1}{t_{k+1}}\left(x^{k+1} - x^k\right)$.

Theorem ($O(\frac{1}{k^2})$ rate of convergence of MFISTA).

Suppose that Assumption 2 holds. Let $\{x^k\}_{k\geq 0}$ be the sequence generated by MFISTA for solving problem (1) with either a constant stepsize rule in which $L_k \equiv L_f$ for all $k \geq 0$ or the backtracking procedure B3. Then for any $x^* \in X^*$ and $k \geq 0$,

$$F(x^k) - F_{\text{opt}} \leq \frac{2\alpha L_f \left\|x^0 - x^*\right\|^2}{(k+1)^2},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.