Accelerated Proximal Gradient Methods

DSAI5104 (2025-26)

Previously

- We considered the problem:

$$\min_{\mathbf{x}} \; f(\mathbf{x}), \qquad f : \mathbb{R}^n \mapsto \mathbb{R} \text{ differentiable.}$$

- GD method:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k \Big(-\nabla f(\mathbf{x}^k) \Big)$$

- Computational Complexity:

$O(1/\sqrt{k})$	for L-Lipschtiz convex functions
$O(1/k)$	for L-smooth convex functions

- LASSO problem:

$$\min_{\boldsymbol{\beta}} \; \frac{1}{2}\|X\boldsymbol{\beta} - \mathbf{b}\|^2 + \lambda \|\boldsymbol{\beta}\|_1 \qquad \text{(NOT differentiable)}$$

Today's Focus

- We consider the composite problem (e.g., LASSO):

$$\min_{\mathbf{x}} \; F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}), \tag{1}$$

  where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is L-smooth convex and $g : \mathbb{R}^n \mapsto \mathbb{R}$ is convex, but usually nonsmooth.
- We extend GD to the composite problem.
- We introduce the celebrated Nesterov acceleration to improve the complexity from $O(1/k)$ to $O(1/k^2)$ (a giant improvement!)

Tools Needed

- Strongly convex functions.
- Proximal operator.
- Acceleration scheme.

Strongly Convex Functions

Definition (μ-strongly convex)

A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be μ-strongly convex if

$$f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$$

is convex with the module $\mu > 0$.

Lemma (Quadratic Growth Lemma)

Suppose $f : \mathbb{R}^n \mapsto \mathbb{R}$ is μ-strongly convex and is differentiable. Then we have

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{2}$$

In particular, we have the quadratic growth away from $\mathbf{x}^*$:

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}^*\|^2 \quad \text{and} \quad \mathbf{x}^* = \arg\min f(\mathbf{x}).$$

Proof

Since the function $h(\mathbf{x}) := f(\mathbf{x}) - (\mu/2)\|\mathbf{x}\|^2$ is convex, we have

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \quad \forall\, \mathbf{x}, \mathbf{y}.$$

Realizing that $\nabla h(\mathbf{x}) = \nabla f(\mathbf{x}) - \mu \mathbf{x}$, the above inequality translate to (2).

Suppose $\mathbf{x}^*$ is an optimal minimizer. Then $\nabla f(\mathbf{x}^*) = 0$. The inequality (2) implies the quadratic growth inequality, which in turn implies the uniqueness of $\mathbf{x}^*$. □

Remark:

If f is both L-smooth and μ-strongly convex, then we have

$$\begin{aligned} & f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2 \\ \leq\ & f(\mathbf{y}) \\ \leq\ & f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \end{aligned}$$

Proximal Operator

Definition

Let $g : \mathbb{R}^n \mapsto \overline{R}$ be proper, closed and convex. We define the proximal operator (or proximal mapping) of g by

$$\mathrm{Prox}_g(\mathbf{x}) := \arg\min_{\mathbf{u}\in\mathbb{R}^n} \left\{ g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2 \right\}.$$

- A function f is proper if $f(\mathbf{x}) > -\infty$ for all $\mathbf{x}$. It is closed if its epigraph $\mathrm{epi} f$ is closed:

$$\mathrm{epi} f = \left\{ (\mathbf{x}, \mu) \in \mathbb{R}^{n+1} \mid \mu \geq f(\mathbf{x}),\ \forall\, \mathbf{x} \in \mathbb{R}^n \right\}.$$

- Since g is proper, closed and convex, the function

$$\mathbf{u} \to g(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2$$

is proper, closed and $1/2$-strongly convex. Thus, it has a unique minimizer. Hence $\mathrm{Prox}_g : \mathbb{R}^n \to \mathbb{R}^n$ is well defined.

Proximal Operator: Projection

When $g = \delta_C$ (indicator function) for a closed convex set C:

$$\delta_C(\mathbf{x}) := \begin{cases} 0 & \text{if } \mathbf{x} \in C \\ +\infty & \text{otherwise,} \end{cases}$$

we have

$$\mathrm{Prox}_g(\mathbf{x}) = \arg\min_{\mathbf{u}\in C} \frac{1}{2}\|\mathbf{u} - \mathbf{x}\|^2.$$

This is the projection from $\mathbf{x}$ to the set C. The projection operator is usually denoted as Π_C. That is

$$\mathrm{Prox}_{\delta_C}(\mathbf{x}) = \Pi_C(\mathbf{x}).$$

Example: Let

$$C := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq 0\} = \mathbb{R}^n_+ \quad \text{(nonnegative orthant)}$$

Then $\Pi_C(\mathbf{x}) = \max\{\mathbf{x}, 0\}$.

Proximal Operator: Soft-Thresholding Operator

Interestingly, there are many functions whose proximal operators can be cheaply calculated. Probably, the best known function is the ℓ_1 norm (e.g., in LASSO):

$$g(\mathbf{x}) = \|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|.$$

Note that

$$\begin{aligned}
\mathsf{Prox}_{\mu\|\cdot\|_1}(\mathbf{x}) &= \arg\min_{\mathbf{u}\in\mathbb{R}^n} \left\{ \sum_{i=1}^{n} \left(\frac{1}{2}(u_i - x_i)^2 + \mu|u_i| \right) \right\} \\
&= \mathtt{sign}(\mathbf{x}) \circ \max\left\{ |\mathbf{x}| - \mu, 0 \right\} \\
&=: \mathcal{T}_\mu(\mathbf{x}),
\end{aligned}$$

where $\circ$ is the componentwise multiplication and $\mathtt{sign}$ is the sign function applied componentwise.

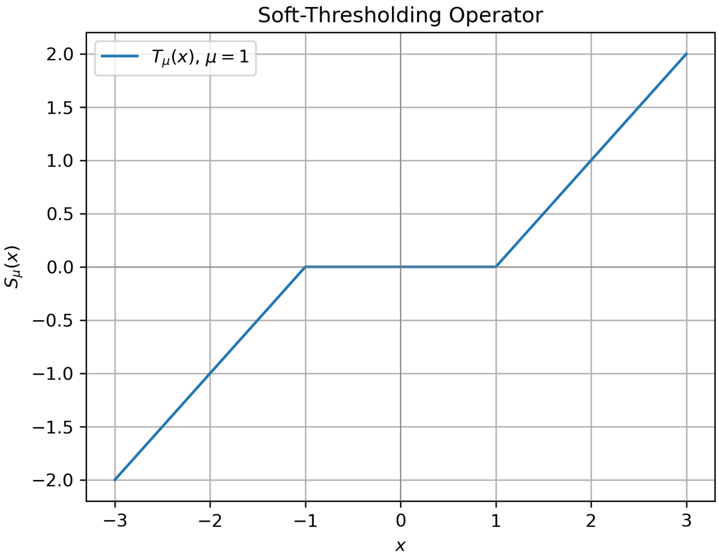


Figure: Soft-Thresholding Operator

Proximal GD for (1)

Proximal gradient descent (PGD): Let $\mathbf{x}^0 \in \mathrm{dom}(g)$. For $k = 0, 1, \ldots,$

$$\mathbf{x}^{k+1} = \mathrm{Prox}_{\frac{1}{L}g}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right) \tag{3}$$

In fact, $\mathbf{x}^{k+1}$ is obtained like this:

$$\begin{aligned}
\mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k),\ \mathbf{x} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}^k\|^2 + g(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{L}{2}\|\mathbf{x} - (\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))\|^2 + g(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x} - (\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k))\|^2 + \frac{1}{L}g(\mathbf{x}) \right\} \\
&= \mathrm{Prox}_{\frac{1}{L}g}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right).
\end{aligned}$$

PGD has $O(1/k)$ Complexity

Consider the composite problem (1) with f being L-smooth and convex, and g being convex. Consider the PGD algorithm (3). We then have

$$F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \frac{L}{2k}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

The proof is similar to GD and is left as an exercise.

APGA: Accelerated Proximal Gradient Algorithm for (1)

Accelerated PGA: Let $\theta_0 = \theta_{-1} = 1$ and $\mathbf{x}^0$ be given. For $k \geq 0$, compute

$$\begin{cases} \mathbf{y}^k & = \mathbf{x}^k + \theta_k(\theta_{k-1}^{-1} - 1)(\mathbf{x}^k - \mathbf{x}^{k-1}), \\ \mathbf{x}^{k+1} & = \mathrm{Prox}_{\frac{1}{L}g}\Big(\mathbf{y}^k - \frac{1}{L}\nabla f(\mathbf{y}^k)\Big), \end{cases} \tag{4}$$

and choose $\theta_{k+1} \in (0, 1]$ so that

$$\frac{1 - \theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}.$$

Choice of θ_k:

$$\theta_k = 2/(k+2) \quad \text{or} \quad \theta_{k+1} = \frac{1}{2}\left(\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2\right).$$

APGA has $O(1/k^2)$ Complexity

Theorem

Consider the composite problem (1) *with f being L-smooth and convex, and g being convex. Consider the APGA algorithm* (3). *We then have*

$$F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \frac{L\theta_{k-1}^2}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

- Compared with PGD, there is a new point $\mathbf{y}^k$ computed each iteration, followed by a proximal step at $\mathbf{y}^k$. Because of this, APGA is not a decreasing algorithm any more. That is, $F(\mathbf{x}^{k+1})$ is not necessarily strictly less than $F(\mathbf{x}^k)$.
- The update condition on θ_k is crucial and the proof of convergence is innovative. When $\theta_k = O(1/k)$, we get $O(1/k^2)$ complexity.

Proof

Step 1: Inequality on the proximal step.

$$
\begin{aligned}
\mathbf{x}^{k+1} &= \mathrm{Prox}_{\frac{1}{L}g}\Big(\mathbf{y}^k - \frac{1}{L}\nabla f(\mathbf{y}^k)\Big) \\
&= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{x} - (\mathbf{y}^k - \frac{1}{L}\nabla f(\mathbf{y}^k))\|^2 + \frac{1}{L}g(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{y}^k\|^2 + \frac{1}{L}\langle \nabla f(\mathbf{y}^k), \mathbf{x} - \mathbf{y}^k \rangle + \frac{1}{L}g(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}} \left\{ \frac{L}{2}\|\mathbf{x} - \mathbf{y}^k\|^2 + \langle \nabla f(\mathbf{y}^k), \mathbf{x} - \mathbf{y}^k \rangle + g(\mathbf{x}) \right\} \\
&= \arg\min_{\mathbf{x}} \underbrace{\left\{ f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \mathbf{x} - \mathbf{y}^k \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}^k\|^2 + g(\mathbf{x}) \right\}}_{=:\Xi(\mathbf{y}^k, \mathbf{x})}.
\end{aligned}
$$

By the quadratic growth lemma, we have

$$
\Xi(\mathbf{y}^k, \mathbf{y}) \geq \Xi(\mathbf{y}^k, \mathbf{x}^{k+1}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}^{k+1}\|^2. \tag{5}
$$

Step 2: Bound on $F(\mathbf{x}^k)$. By the Descent Lemma, we have for any $\mathbf{y} \in \mathbb{R}^n$

$$
\begin{aligned}
F(\mathbf{x}^{k+1}) &= f(\mathbf{x}^{k+1}) + g(\mathbf{x}^{k+1}) \\
&\le \underbrace{f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k),\ \mathbf{x}^{k+1} - \mathbf{y}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2 + g(\mathbf{x}^{k+1})}_{=\Xi(\mathbf{y}^k, \mathbf{x}^{k+1})} \\
&\overset{(5)}{\le} \Xi(\mathbf{y}^k, \mathbf{y}) - \frac{L}{2}\|\mathbf{y} - \mathbf{x}^{k+1}\|^2 \\
&\le \underbrace{f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k),\ \mathbf{y} - \mathbf{y}^k \rangle}_{\le f(\mathbf{y})} + \frac{L}{2}\|\mathbf{y} - \mathbf{y}^k\|^2 \\
&\quad + g(\mathbf{y}) - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{y}\|^2 \\
&\le F(\mathbf{y}) + \frac{L}{2}\|\mathbf{y} - \mathbf{y}^k\|^2 - \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{y}\|^2.
\end{aligned}
$$

Step 3: Convex interpolation step. Set $\mathbf{y} := (1-\theta_k)\mathbf{x}^k + \theta_k \mathbf{x}^*$. This is a convex combination because $\theta_k \in [0,1]$. Hence

$$
\begin{aligned}
F(\mathbf{x}^{k+1}) &\le F((1-\theta_k)\mathbf{x}^k + \theta_k\mathbf{x}^*) + \frac{L}{2}\|(1-\theta_k)\mathbf{x}^k + \theta_k\mathbf{x}^* - \mathbf{y}^k\|^2 \\
&\quad - \frac{L}{2}\|(1-\theta_k)\mathbf{x}^k + \theta_k\mathbf{x}^* - \mathbf{x}^{k+1}\|^2 \\
&= F((1-\theta_k)\mathbf{x}^k + \theta_k\mathbf{x}^*) + \frac{L\theta_k^2}{2}\|\mathbf{x}^* + (\theta_k^{-1}-1)\mathbf{x}^k - \theta_k^{-1}\mathbf{y}^k\|^2 \\
&\quad - \frac{L\theta_k^2}{2}\|\mathbf{x}^* + (\theta_k^{-1}-1)\mathbf{x}^k - \theta_k^{-1}\mathbf{x}^{k+1}\|^2.
\end{aligned}
$$

Step 4: Extrapolation step. This is also regarded as the magic step, which simplifies the inequality in Step 3. Let

$$\begin{aligned}\mathbf{z}^k &:= -(\theta_k^{-1} - 1)\mathbf{x}^k + \theta_k^{-1}\mathbf{y}^k \\ &= -(\theta_k^{-1} - 1)\mathbf{x}^k + \theta_k^{-1}\mathbf{x}^k + (\theta_{k-1}^{-1} - 1)(\mathbf{x}^k - \mathbf{x}^{k-1}) \\ &= -(\theta_{k-1}^{-1} - 1)\mathbf{x}^{k-1} + \theta_{k-1}^{-1}\mathbf{x}^k.\end{aligned}$$

Thus, we further have that

$$\begin{aligned}F(\mathbf{x}^{k+1}) &\le F((1-\theta_k)\mathbf{x}^k + \theta_k\mathbf{x}^*) + \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^k\|^2 - \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^{k+1}\|^2 \\ &\le (1-\theta_k)F(\mathbf{x}^k) + \theta_k F(\mathbf{x}^*) + \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^k\|^2 \\ &\qquad - \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^{k+1}\|^2. \qquad (6)\end{aligned}$$

Rearranging the terms, we have for all $k \geq 0$

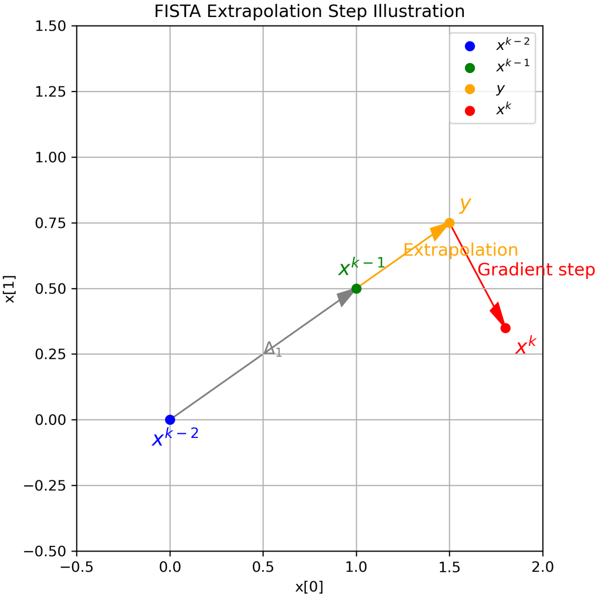
$$F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) \leq (1-\theta_k)[F(\mathbf{x}^k) - F(\mathbf{x}^*)] + \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^k\|^2 - \frac{L\theta_k^2}{2}\|\mathbf{x}^* - \mathbf{z}^{k+1}\|^2.$$

Hence

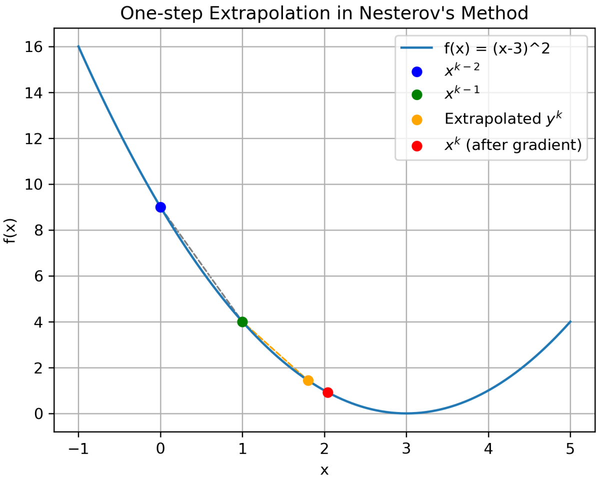
$$\begin{aligned}
&\frac{1-\theta_{k+1}}{\theta_{k+1}^2}[F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*)] + \frac{L}{2}\|\mathbf{x}^* - \mathbf{z}^{k+1}\|^2 \\
&\leq \frac{1}{\theta_k^2}[F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*)] + \frac{L}{2}\|\mathbf{x}^* - \mathbf{z}^{k+1}\|^2 \\
&\leq \frac{1-\theta_k}{\theta_k^2}[F(\mathbf{x}^k) - F(\mathbf{x}^*)] + \frac{L}{2}\|\mathbf{x}^* - \mathbf{z}^k\|^2 \\
&\leq \cdots \leq \frac{1-\theta_0}{\theta_0^2}[F(\mathbf{x}^0) - F(\mathbf{x}^*)] + \frac{L}{2}\|\mathbf{x}^* - \mathbf{z}^0\|^2 = \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}^0\|^2,
\end{aligned}$$

since $\mathbf{z}^0 = \mathbf{x}^0$ and $\theta_0 = 1$. $\square$

Extrapolation Step: Schematic Illustration



Extrapolation Step: Numerical Illustration



ISTA and FISTA: The case $g = \lambda\|\cdot\|_1$

Let $\mathbf{u}^k := \mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)$.
ISTA: Iterative Soft-Thresholding Algorithm:

$$\mathbf{x}^{k+1} = \mathsf{Prox}_{\frac{1}{L}g}\left(\mathbf{u}^k\right) = \mathsf{sign}(\mathbf{u}^k) \circ \max\{|\mathbf{u}^k| - \lambda/L, 0\}.$$

FISTA: Fast ISTA: Take $\mathbf{y}^1 = \mathbf{x}^0 \in \mathbb{R}^n$, and $t_1 = 1$. For $k \geq 1$, compute

$$\begin{cases} \mathbf{x}^k & = \mathsf{Prox}_{\frac{1}{L}g}\left(\mathbf{y}^k - \frac{1}{L}\nabla f(\mathbf{y}^k)\right) \\ t_{k+1} & = \frac{1+\sqrt{1+4t_k^2}}{2} \\ \mathbf{y}^{k+1} & = \mathbf{x}^k + \left(\frac{t_k - 1}{t_{k+1}}\right)\left(\mathbf{x}^k - \mathbf{x}^{k-1}\right) \end{cases}$$

Q: What is the relationship between t_k in FISTA and θ_k in APGA?

Example

$$\min_{x_1,x_2} f(x_1,x_2)+|x_1|+|x_2|, \quad \text{with } f(x_1,x_2) = \frac{1}{2}(x_1+x_2)^2-2x_1-x_2$$

Start with $\mathbf{y}^1 = \mathbf{x}^0 = (0,0)$, use FISTA to compute the first two iterates $\mathbf{x}^1$ and $\mathbf{x}^2$.

Solution: We know $L = 2$ and

$$\nabla f(\mathbf{x}) = \begin{bmatrix} x_1 + x_2 - 2 \\ x_1 + x_2 - 1 \end{bmatrix}$$

$$\mathbf{u}^1 = \mathbf{y}^1 - \frac{1}{L}\nabla f(\mathbf{y}^1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2 \end{bmatrix}$$

$$\mathbf{x}^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \circ \max\left\{\begin{bmatrix} 1 \\ 1/2 \end{bmatrix} - \frac{1}{2}, 0\right\} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$$

$$t_2 = \frac{1+\sqrt{5}}{2}$$

$$\mathbf{y}^2 = \mathbf{x}^1 + \left(\frac{t_1 - 1}{t_2}\right)\left(\mathbf{x}^1 - \mathbf{x}^0\right) = \mathbf{x}^1 = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$$

$$\mathbf{u}^2 = \mathbf{y}^2 - \frac{1}{L}\nabla f(\mathbf{y}^2) = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} -3/2 \\ -1 \end{bmatrix} = \begin{bmatrix} 5/4 \\ 1/2 \end{bmatrix}$$

$$\mathbf{x}^2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \circ \max\left\{\begin{bmatrix} 5/4 \\ 1/2 \end{bmatrix} - \frac{1}{2}, 0\right\} = \begin{bmatrix} 3/4 \\ 0 \end{bmatrix}$$

Numerical Comparison: Lasso

$$\min_{\mathbf{x}} \; F(\mathbf{x}) = \underbrace{\frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2}_{=f(\mathbf{x})} + \underbrace{\lambda\|\mathbf{x}\|_1}_{=g(\mathbf{x})},$$

where $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $\lambda > 0$ is a (penalty) parameter.
Lipschitz constant of $f(\cdot)$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 = \|A^\top A(\mathbf{x} - \mathbf{y})\|_2 \le \rho(A^\top A)\|\mathbf{x} - \mathbf{y}\|_2,$$

where $\rho(A^\top A)$ is the largest eigenvalue of the matrix $(A^\top A)$.

$$L = \rho(A^\top A).$$

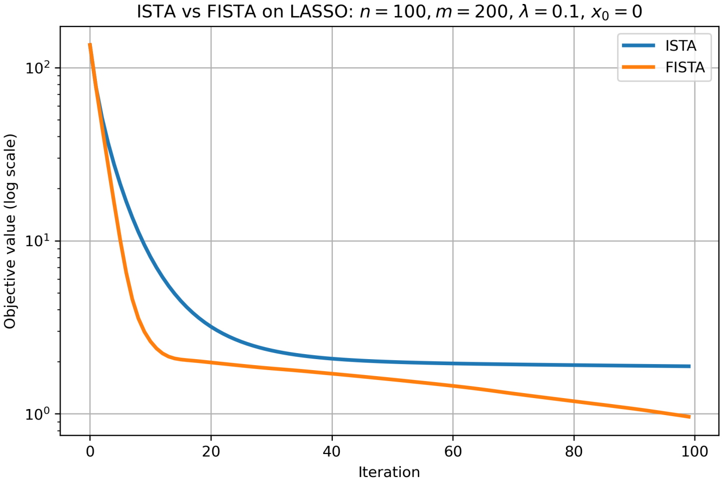


Figure: Comparison of ISTA and FISTA

Comments on Fast Gradient Methods

- Nesterov (1983): a 6-page paper on a gradient method with $O(1/k^2)$ convergence rate.
- Beck and Teboulle (2008): FISTA – Proximal gradient version of Nesterov's 1983 method.
- Tseng (2008): a unified analysis of fast gradient methods (followed by us)

Summary

- We considered the problem:

$$\min_{\mathbf{x}} \; f(\mathbf{x}) + g(\mathbf{x})$$

where f is L-smooth and often convex, g is convex, proximal friendly.

- PGD:

$$\mathbf{x}^{k+1} = \mathrm{Prox}_{\frac{1}{L}g}\left(\mathbf{x}^k - \frac{1}{L}\nabla f(\mathbf{x}^k)\right).$$

- APGA:

$$\begin{cases} \mathbf{y}^k & = \mathbf{x}^k + \theta_k(\theta_{k-1}^{-1} - 1)(\mathbf{x}^k - \mathbf{x}^{k-1}), \\ \mathbf{x}^{k+1} & = \mathrm{Prox}_{\frac{1}{L}g}\left(\mathbf{y}^k - \frac{1}{L}\nabla f(\mathbf{y}^k)\right), \end{cases}$$

Summary

- Function class: L-smooth and convex:

$$\alpha = \frac{1}{L}, \quad \text{complexity} \begin{cases} \text{PGD}: & O(1/k) & F(\mathbf{x}^{k+1}) \leq F(\mathbf{x}^k) \\ \text{APGA}: & O(1/k^2) & F(\mathbf{x}^{k+1}) \not\leq F(\mathbf{x}^k) \end{cases}$$

  If we want a solution that is upto $\epsilon = 10^{-4}$ accuracy, the number of iterations take are of the order of

$$10^4 \quad \text{vs} \quad 10^2.$$

- Question: The constant L is usually hard to estimate. What can be done about its estimation?

Optimality of $O(1/k^2)$ Complexity

GD, PGD, and APGA are of first-order methods because only gradient information was used. APGA belongs to a class of first-order methods commonly known as optimal methods: there exists a convex LC^1 function and $\arg\min f \neq \emptyset$ such that for any first-order method that generates iterates as

$$\mathbf{x}^k \in \mathbf{x}^0 + \mathrm{span}\{\nabla f(\mathbf{x}^0), \nabla f(\mathbf{x}^1), \cdots, \nabla f(\mathbf{x}^{k-1})\}, \quad k \geq 1$$

It holds that for any $\mathbf{x}^* \in \arg\min f$,

$$f(\mathbf{x}^k) - f(\mathbf{x}^*) \geq \frac{3L\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{32(k+1)^2}$$

whenever $1 \leq k \leq (n-1)/2$.