

Lecture: Fast Proximal Gradient Methods

Zaiwen Wen

<http://bicmr.pku.edu.cn/~wenzw/opt-2016-fall.html>

Acknowledgement: this slides is based on Prof. Lieven Vandenberghe's lecture notes

Outline

- 1 fast proximal gradient method (FISTA)
- 2 FISTA with line search
- 3 FISTA as descent method
- 4 Nesterov's second method
- 5 Proof by estimating sequence

Fast (proximal) gradient methods

- Nesterov (1983, 1988, 2005): three projection methods with $1/k^2$ convergence rate
- Beck & Teboulle (2008): FISTA, a proximal gradient version of Nesterov's 1983 method
- Nesterov (2004 book), Tseng (2008): overview and unified analysis of fast gradient methods
- several recent variations and extensions

this lecture

FISTA and Nesterov's 2nd method (1988) as presented by Tseng

FISTA (basic version)

$$\text{minimize } f(x) = g(x) + h(x)$$

- g convex, differentiable, with **dom** $g = \mathbb{R}^n$
- h closed, convex, with inexpensive prox_{th} operator

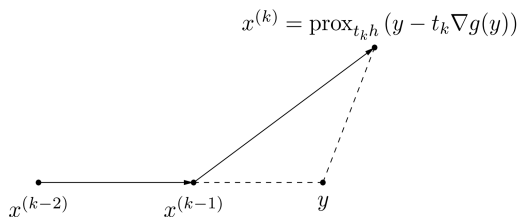
algorithm: choose any $x^{(0)} = x^{(-1)}$; for $k \geq 1$, repeat the steps

$$y = x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$$
$$x^{(k)} = \text{prox}_{t_k h}(y - t_k \nabla g(y))$$

- step size t_k fixed or determined by line search
- acronym stands for ‘Fast Iterative Shrinkage-Thresholding Algorithm’

Interpretation

- first iteration ($k = 1$) is a proximal gradient step at $y = x^{(0)}$
- next iterations are proximal gradient steps at extrapolated points y

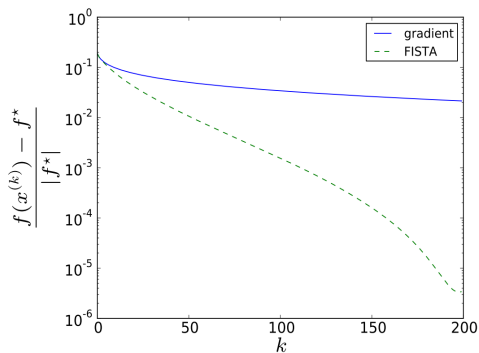


note: $x^{(k)}$ is feasible (in $\mathbf{dom} h$); y may be outside $\mathbf{dom} h$

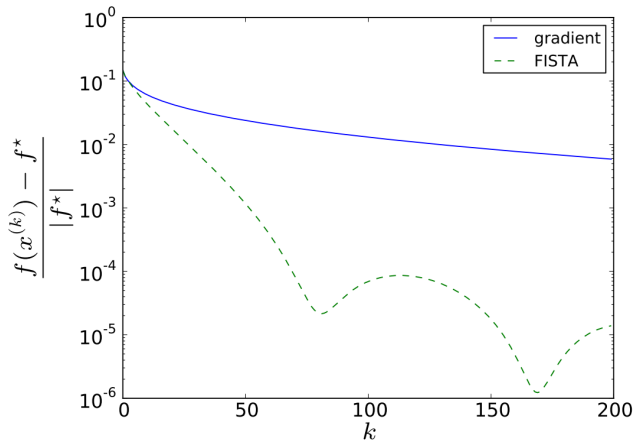
Example

$$\text{minimize} \quad \log \sum_{i=1}^m \exp(a_i^T x + b_i)$$

randomly generated data with $m = 2000$, $n = 1000$, same fixed step size



another instance



FISTA is not a descent method

Convergence of FISTA

assumptions

- g convex with $\text{dom } g = \mathbb{R}^n$; ∇g Lipschitz continuous with constant L :

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$$

- h is closed and convex (so that $\text{prox}_{th}(u)$ is well defined)
- optimal value f^* is finite and attained at x^* (not necessarily unique)

convergence result: $f(x^{(k)}) - f^*$ decreases at least as fast as $1/k^2$

- with fixed step size $t_k = 1/L$
- with suitable line search

Reformulation of FISTA

define $\theta_k = 2/(k + 1)$ and introduce an intermediate variable $v^{(k)}$

algorithm: choose $x^{(0)} = v^{(0)}$; for $k \geq 1$, repeat the steps

$$\begin{aligned}y &= (1 - \theta_k)x^{(k-1)} + \theta_k v^{(k-1)} \\x^{(k)} &= \text{prox}_{t_k h}(y - t_k \nabla g(y)) \\v^{(k)} &= x^{(k-1)} + \frac{1}{\theta_k}(x^{(k)} - x^{(k-1)})\end{aligned}$$

substituting expression for $v^{(k)}$ in formula for y gives FISTA of page 4

Important inequalities

choice of θ_k : the sequence $\theta_k = 2/(k + 1)$ satisfies $\theta_1 = 1$ and

$$\frac{1 - \theta_k}{\theta_k^2} \leq \frac{1}{\theta_{k-1}^2}, \quad k \geq 2$$

upper bound on g from Lipschitz property

$$g(u) \leq g(z) + \nabla g(z)^T (u - z) + \frac{L}{2} \|u - z\|_2^2 \quad \forall u, z$$

upper bound on h from definition of prox-operator

$$h(u) \leq h(z) + \frac{1}{t} (w - u)^T (u - z) \quad \forall w, u = \text{prox}_{th}(w), z$$

Progress in one iteration

define $x = x^{(i-1)}, x^+ = x^{(i)}, v = v^{(i-1)}, v^+ = v^{(i)}, t = t_i, \theta = \theta_i$

- upper bound from Lipschitz property: if $0 < t \leq 1/L$

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t}\|x^+ - y\|_2^2 \quad (1)$$

- upper bound from definition of prox-operator:

$$h(x^+) \leq h(z) + \nabla g(y)^T(z - x^+) + \frac{1}{t}(x^+ - y)^T(z - x^+) \quad \forall z$$

- add the upper bounds and use convexity of g

$$f(x^+) \leq f(z) + \frac{1}{t}(x^+ - y)^T(z - x^+) + \frac{1}{2t}\|x^+ - y\|_2^2 \quad \forall z$$

- make convex combination of upper bounds for $z = x$ and $z = x^*$

$$\begin{aligned}
 & f(x^+) - f^* - (1 - \theta)(f(x) - f^*) \\
 &= f(x^+) - \theta f^* - (1 - \theta)f(x) \\
 &\leq \frac{1}{t}(x^+ - y)^T(\theta x^* + (1 - \theta)x - x^+) + \frac{1}{2t}\|x^+ - y\|_2^2 \\
 &= \frac{1}{2t}(\|y - (1 - \theta)x - \theta x^*\|_2^2 - \|x^+ - (1 - \theta)x - \theta x^*\|_2^2) \\
 &= \frac{\theta^2}{2t}(\|v - x^*\|_2^2 - \|v^+ - x^*\|_2^2)
 \end{aligned}$$

conclusion: if the inequality (1) holds at iteration i , then

$$\begin{aligned}
 & \frac{t_i}{\theta_i^2} \left(f(x^{(i)}) - f^* \right) + \frac{1}{2} \|v^{(i)} - x^*\|_2^2 \\
 & \leq \frac{(1 - \theta_i)t_i}{\theta_i^2} \left(f(x^{(i-1)}) - f^* \right) + \frac{1}{2} \|v^{(i-1)} - x^*\|_2^2
 \end{aligned} \tag{2}$$

Analysis for fixed step size

take $t_i = t = 1/L$ and apply (2) recursively, using $(1 - \theta_i)/\theta_i^2 \leq 1/\theta_{i-1}^2$;

$$\begin{aligned} & \frac{t}{\theta_k^2} \left(f(x^{(k)}) - f^* \right) + \frac{1}{2} \|v^{(k)} - x^*\|_2^2 \\ & \leq \frac{(1 - \theta_1)t}{\theta_1^2} \left(f(x^{(0)}) - f^* \right) + \frac{1}{2} \|v^{(0)} - x^*\|_2^2 \\ & = \frac{1}{2} \|x^{(0)} - x^*\|_2^2 \end{aligned}$$

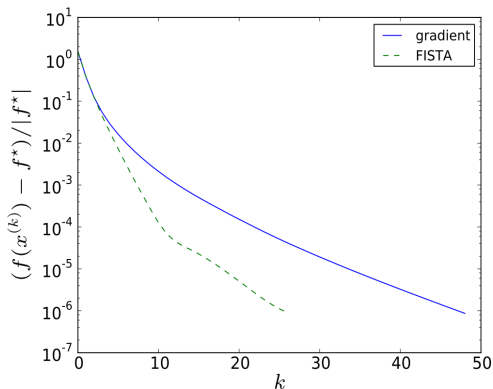
therefore

$$f(x^{(k)}) - f^* \leq \frac{\theta_k^2}{2t} \|x^{(0)} - x^*\|_2^2 = \frac{2L}{(k+1)^2} \|x^{(0)} - x^*\|_2^2$$

conclusion: reaches $f(x^{(k)}) - f^* \leq \epsilon$ after $\mathcal{O}(1/\sqrt{\epsilon})$ iterations

Example: quadratic program with box constraints

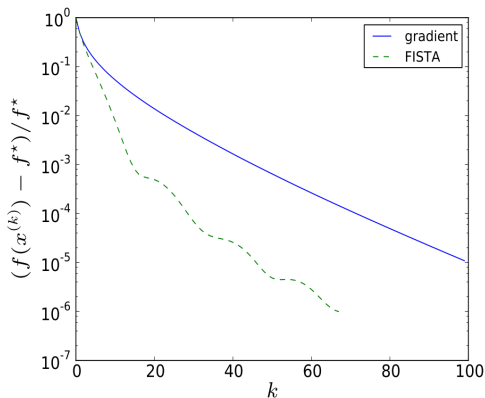
$$\begin{aligned} & \text{minimize} && (1/2)x^T A x + b^T x \\ & \text{subject to} && 0 \leq x \leq \mathbf{1} \end{aligned}$$



$n = 3000$; fixed step size $t = 1/\lambda_{\max}(A)$

1-norm regularized least-squares

$$\text{minimize } \frac{1}{2} \|Ax - b\|_2^2 + \|x\|_1$$



randomly generated $A \in \mathbb{R}^{2000 \times 1000}$; step $t_k = 1/L$ with $L = \lambda_{\max}(A^T A)$

Outline

- 1 fast proximal gradient method (FISTA)
- 2 FISTA with line search**
- 3 FISTA as descent method
- 4 Nesterov's second method
- 5 Proof by estimating sequence

Key steps in the analysis of FISTA

- the starting point (page 11) is the inequality

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t} \|x^+ - y\|_2^2 \quad (1)$$

this inequality is known to hold for $0 < t \leq 1/L$

- if (1) holds, then the progress made in iteration i is bounded by

$$\begin{aligned} & \frac{t_i}{\theta_i^2} \left(f(x^{(i)}) - f^* \right) + \frac{1}{2} \|v^{(i)} - x^*\|_2^2 \\ & \leq \frac{(1 - \theta_i)t_i}{\theta_i^2} \left(f(x^{(i-1)}) - f^* \right) + \frac{1}{2} \|v^{(i-1)} - x^*\|_2^2 \end{aligned} \quad (2)$$

- to combine these inequalities recursively, we need

$$\frac{(1 - \theta_i)t_i}{\theta_i^2} \leq \frac{t_{i-1}}{\theta_{i-1}^2} \quad (i \geq 2) \quad (3)$$

- if $\theta_1 = 1$, combining the inequalities (2) from $i = 1$ to k gives the bound

$$f(x^{(k)}) - f^* \leq \frac{\theta_k^2}{2t_k} \|x^{(0)} - x^*\|_2^2$$

conclusion: rate $1/k^2$ convergence if (1) and (3) hold with

$$\frac{\theta_k^2}{t_k} = \mathcal{O}\left(\frac{1}{k^2}\right)$$

FISTA with fixed step size

$$t_k = \frac{1}{L}, \quad \theta_k = \frac{2}{k+1}$$

these values satisfies (1) and (3) with

$$\frac{\theta_k^2}{t_k} = \frac{4L}{(k+1)^2}$$

FISTA with line search (method 1)

replace update of x in iteration k (page 9) with

$$t := t_{k-1} \quad (\text{define } t_0 = \hat{t} > 0)$$

$$x := \text{prox}_{th}(y - t\nabla g(y))$$

$$\text{while } g(x) > g(y) + \nabla g(y)^T(x - y) + \frac{1}{2t}\|x - y\|_2^2$$

$$t := \beta t$$

$$x := \text{prox}_{th}(y - t\nabla g(y))$$

end

- inequality (1) holds trivially, by the backtracking exit condition
- inequality (3) holds with $\theta_k = 2/(k+1)$ because $t_k \leq t_{k-1}$
- Lipschitz continuity of ∇g guarantees $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
- preserves $1/k^2$ convergence rate because $\theta_k^2/t_k = \mathcal{O}(1/k^2)$:

$$\frac{\theta_k^2}{t_k} \leq \frac{4}{(k+1)^2 t_{\min}}$$

FISTA with line search (method 2)

replace update of y and x in iteration k (page 9) with

$$t := \hat{t} > 0$$

$$\theta := \text{positive root of } t_{k-1}\theta^2 = t\theta_{k-1}^2(1 - \theta)$$

$$y := (1 - \theta)x^{(k-1)} + \theta v^{(k-1)}$$

$$x := \text{prox}_{th}(y - t\nabla g(y))$$

$$\text{while } g(x) > g(y) + \nabla g(y)^T(x - y) + \frac{1}{2t}\|x - y\|_2^2$$

$$t := \beta t$$

$$\theta := \text{positive root of } t_{k-1}\theta^2 = t\theta_{k-1}^2(1 - \theta)$$

$$y := (1 - \theta)x^{(k-1)} + \theta v^{(k-1)}$$

$$x := \text{prox}_{th}(y - t\nabla g(y))$$

end

assume $t_0 = 0$ in the first iteration ($k = 1$), *i.e.*, take $\theta_1 = 1, y = x^{(0)}$

discussion

- inequality (1) holds trivially, by the backtracking exit condition
- inequality (3) holds trivially, by construction of θ_k
- Lipschitz continuity of ∇g guarantees $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
- θ_i is defined as the positive root of $\theta_i^2/t_i = (1 - \theta_i)\theta_{i-1}^2/t_{i-1}$; hence

$$\frac{\sqrt{t_{i-1}}}{\theta_{i-1}} = \frac{\sqrt{(1 - \theta_i)t_i}}{\theta_i} \leq \frac{\sqrt{t_i}}{\theta_i} - \frac{\sqrt{t_i}}{2}$$

combine inequalities from $i = 2$ to k to get $\sqrt{t_i} \leq \frac{\sqrt{t_k}}{\theta_k} - \frac{1}{2} \sum_{i=2}^k \sqrt{t_i}$

- rearranging shows that $\theta_k^2/t_k = \mathcal{O}(1/k^2)$:

$$\frac{\theta_k^2}{t_k} \leq \frac{1}{(\sqrt{t_1} + \frac{1}{2} \sum_{i=2}^k \sqrt{t_i})^2} \leq \frac{4}{(k+1)^2 t_{\min}}$$

Comparison of line search methods

method 1

- uses nonincreasing stepsizes (enforces $t_k \leq t_{k-1}$)
- one evaluation of $g(x)$, one prox_{th} evaluation per line search iteration

method 2

- allows non-monotonic step sizes
- one evaluation of $g(x)$, one evaluation of $g(y)$, $\nabla g(y)$, one evaluation of prox_{th} per line search iteration

the two strategies can be combined and extended in various ways

Outline

- 1 fast proximal gradient method (FISTA)
- 2 FISTA with line search
- 3 FISTA as descent method**
- 4 Nesterov's second method
- 5 Proof by estimating sequence

Descent version of FISTA

choose $x^{(0)} = v^{(0)}$; for $k \geq 1$, repeat the steps

$$y = (1 - \theta_k)x^{(k-1)} + \theta_k v^{(k-1)}$$

$$u = \text{prox}_{t_k h}(y - t_k \nabla g(y))$$

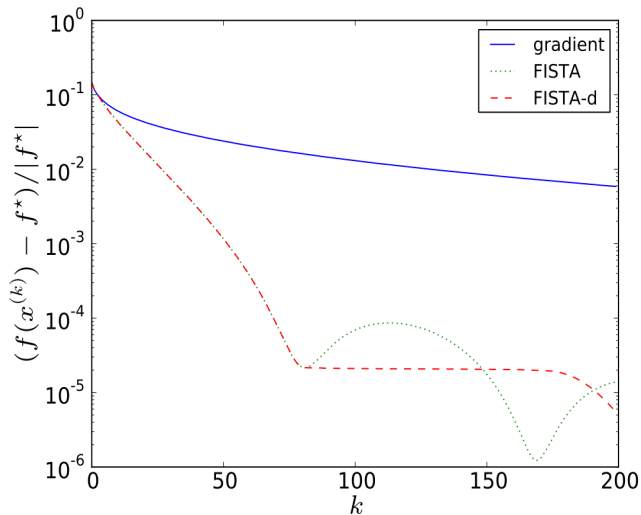
$$x^{(k)} = \begin{cases} u & f(u) \leq f(x^{(k-1)}) \\ x^{(k-1)} & \text{otherwise} \end{cases}$$

$$v^{(k)} = x^{(k-1)} + \frac{1}{\theta_k}(u - x^{(k-1)})$$

- step 3 implies $f(x^{(k)}) \leq f(x^{(k-1)})$
- use $\theta_k = 2/(k+1)$ and $t_k = 1/L$, or one of the line search methods
- same iteration complexity as original FISTA
- changes on page 11: replace x^+ with u and use $f(x^+) \leq f(u)$

Example

(from page 7)



Outline

- 1 fast proximal gradient method (FISTA)
- 2 FISTA with line search
- 3 FISTA as descent method
- 4 Nesterov's second method**
- 5 Proof by estimating sequence

Nesterov's second method

algorithm: choose $x^{(0)} = v^{(0)}$; for $k \geq 1$, repeat the steps

$$y = (1 - \theta_k)x^{(k-1)} + \theta_k v^{(k-1)}$$

$$v^{(k)} = \text{prox}_{(t_k/\theta_k)h} \left(v^{(k-1)} - \frac{t_k}{\theta_k} \nabla g(y) \right)$$

$$x^{(k)} = (1 - \theta_k)x^{(k-1)} + \theta_k v^{(k)}$$

- use $\theta_k = 2/(k+1)$ and $t_k = 1/L$, or one of the line search methods
- identical to FISTA if $h(x) = 0$
- unlike in FISTA, y is feasible (in $\text{dom } h$) if we take $x^{(0)} \in \text{dom } h$

Convergence of Nesterov's second method

assumptions

- g convex; ∇g is Lipschitz continuous on $\mathbf{dom} h \subseteq \mathbf{dom} g$

$$\|\nabla g(x) - \nabla g(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \mathbf{dom} h$$

- h is closed and convex (so that $\text{prox}_{th}(u)$ is well defined)
- optimal value f^* is finite and attained at x^* (not necessarily unique)

convergence result: $f(x^{(k)}) - f^*$ decrease at least as fast as $1/k^2$

- with fixed step size $t_k = 1/L$
- with suitable line search

Analysis of one iteration

define $x = x^{(i-1)}, x^+ = x^{(i)}, v = v^{(i-1)}, v^+ = v^{(i)}, t = t_i, \theta = \theta_i$

- from Lipschitz property if $0 < t \leq 1/L$

$$g(x^+) \leq g(y) + \nabla g(y)^T(x^+ - y) + \frac{1}{2t} \|x^+ - y\|_2^2$$

- plug in $x^+ = (1 - \theta)x + \theta v^+$ and $x^+ - y = \theta(v^+ - v)$

$$g(x^+) \leq g(y) + \nabla g(y)^T((1 - \theta)x + \theta v^+ - y) + \frac{\theta^2}{2t} \|v^+ - v\|_2^2$$

- from convexity of g, h

$$g(x^+) \leq (1 - \theta)g(x) + \theta(g(y) + \nabla g(y)^T(v^+ - y)) + \frac{\theta^2}{2t} \|v^+ - v\|_2^2$$

$$h(x^+) \leq (1 - \theta)h(x) + \theta h(v^+)$$

- upper bound on h from page 10 (with $u = v^+$, $w = v - (t/\theta)\nabla(y)$)

$$h(v^+) \leq h(z) + \nabla g(y)^T(z - v^+) - \frac{\theta}{t}(v^+ - v)^T(v^+ - z) \quad \forall z$$

- combine the upper bounds on $g(x^+)$, $h(x^+)$, $h(v^+)$ with $z = x^*$

$$\begin{aligned} f(x^+) &\leq (1 - \theta)f(x) + \theta f^* - \frac{\theta^2}{t}(v^+ - v)^T(v^+ - x^*) + \frac{\theta^2}{2t}\|v^+ - v\|_2^2 \\ &= (1 - \theta)f(x) + \theta f^* + \frac{\theta^2}{2t}(\|v - x^*\|_2^2 - \|v^+ - x^*\|_2^2) \end{aligned}$$

this is identical to final inequality (2) in the analysis of FISTA on page 12

$$\begin{aligned} &\frac{t_i}{\theta_i^2} \left(f(x^{(i)}) - f^* \right) + \frac{1}{2} \|v^{(i)} - x^*\|_2^2 \\ &\leq \frac{(1 - \theta_i)t_i}{\theta_i^2} \left(f(x^{(i-1)}) - f^* \right) + \frac{1}{2} \|v^{(i-1)} - x^*\|_2^2 \end{aligned}$$

References

surveys of fast gradient methods

- Yu. Nesterov, *Introductory Lectures on Convex Optimization. A Basic Course* (2004)
- P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization* (2008)

FISTA

- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. on Imaging Sciences (2009)
- A. Beck and M. Teboulle, *Gradient-based algorithms with applications to signal recovery*, in: Y. Eldar and D. Palomar (Eds.), *Convex Optimization in Signal Processing and Communications* (2009)

line search strategies

- FISTA papers by Beck and Teboulle
- D. Goldfarb and K. Scheinberg, *Fast first-order methods for composite convex optimization with line search* (2011)
- Yu. Nesterov, *Gradient methods for minimizing composite objective function* (2007)
- O. Güler, *New proximal point algorithms for convex minimization*, SIOPT (1992)

Nesterov's third method (not covered in this lecture)

- Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming (2005)
- S. Becker, J. Bobin, E.J. Candès, *NESTA: a fast and accurate first-order method for sparse recovery*, SIAM J. Imaging Sciences (2011)

Outline

- 1 fast proximal gradient method (FISTA)
- 2 FISTA with line search
- 3 FISTA as descent method
- 4 Nesterov's second method
- 5 Proof by estimating sequence**

FOM Framework: $f^* = \min_x \{f(x), x \in X\}$

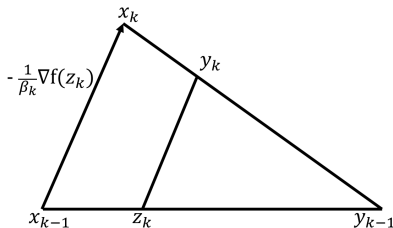
$f(x) \in C_L^{1,1}(X)$ convex. $X \subseteq \mathbb{R}^n$ closed convex. Find $\bar{x} \in X: f(\bar{x}) - f^* \leq \epsilon$

FOM Framework

Input: $x_0 = y_0$, choose $L\gamma_k \leq \beta_k$, $\gamma_1 = 1$. for $k = 1, 2, \dots, N$ do

- 1 $z_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_{k-1}$
- 2 $x_k = \operatorname{argmin}_{x \in X} \left\{ \langle \nabla f(z_k), x \rangle + \frac{\beta_k}{2} \|x - x_{k-1}\|_2^2 \right\}$
- 3 $y_k = (1 - \gamma_k)y_{k-1} + \gamma_k x_k$

- Sequences: $\{x_k\}$, $\{y_k\}$, $\{z_k\}$. Parameters: $\{\gamma_k\}$, $\{\beta_k\}$.



FOM: Techniques for complexity analysis

Lemma 1. (Estimating sequence)

Let $\gamma_t \in (0, 1]$, $t = 1, 2, \dots$, denote $\Gamma_t = \begin{cases} 1 & t = 1 \\ (1 - \gamma_t)\Gamma_{t-1} & t \geq 2 \end{cases}$. If the sequences $\{\Delta_t\}_{t \geq 0}$ satisfies $\Delta_t \leq (1 - \gamma_t)\Delta_{t-1} + B_t$ $t = 1, 2, \dots$, then we have $\Delta_k \leq \Gamma_k(1 - \gamma_1)\Delta_0 + \Gamma_k \sum_{i=1}^k \frac{B_i}{\Gamma_i}$

Remark:

- 1 Let $\Delta_k = f(x_k) - f(x^*)$ or $\Delta_k = \|x_k - x^*\|_2^2$
- 2 Estimate $\{x_k\}$, let $\underbrace{f(x_k) - f(x^*)}_{\Delta_k} \leq (1 - \gamma_k) \underbrace{(f(x_{k-1}) - f(x^*))}_{\Delta_{k-1}} + B_k$
- 3 Note $\Gamma_k = (1 - \gamma_k)(1 - \gamma_{k-1}) \dots (1 - \gamma_2)$; If $\gamma_k = \frac{1}{k} \Rightarrow \Gamma_k = \frac{1}{k}$;
If $\gamma_k = \frac{2}{k+1} \Rightarrow \Gamma_k = \frac{2}{k(k+1)}$; If $\gamma_k = \frac{3}{k+2} \Rightarrow \Gamma_k = \frac{6}{k(k+1)(k+2)}$

FOM Framework: Convergence

Main Goal:
$$\underbrace{f(y_k) - f(x^*)}_{\Delta_k} \leq (1 - \gamma_k) \underbrace{(f(y_{k-1}) - f(x^*))}_{\Delta_{k-1}} + B_k.$$

We have: $f(x) \in C_L^{1,1}(X)$; convexity; optimality condition of subproblem.

$$\begin{aligned} f(y_k) &\leq f(z_k) + \langle \nabla f(z_k), y_k - z_k \rangle + \frac{L}{2} \|y_k - z_k\|^2 \\ &= (1 - \gamma_k)[f(z_k) + \langle \nabla f(z_k), y_{k-1} - z_k \rangle] + \gamma_k[f(z_k) + \langle \nabla f(z_k), x_k - z_k \rangle] + \frac{L\gamma_k^2}{2} \|x_k - x_{k-1}\|^2 \\ &\leq (1 - \gamma_k)f(y_{k-1}) + \gamma_k[f(z_k) + \langle \nabla f(z_k), x_k - z_k \rangle] + \frac{L\gamma_k^2}{2} \|x_k - x_{k-1}\|^2 \end{aligned}$$

Since $x_k = \operatorname{argmin}_{x \in X} \left\{ \langle \nabla f(z_k), x \rangle + \frac{\beta_k}{2} \|x - x_{k-1}\|^2 \right\}$, by the optimal condition

$$\begin{aligned} \Rightarrow \langle \nabla f(z_k) + \beta_k(x_k - x_{k-1}), x_k - x \rangle &\leq 0, \quad \forall x \in X \\ \Rightarrow \langle x_{k-1} - x_k, x_k - x \rangle &\leq \frac{1}{\beta_k} \langle \nabla f(z_k), x - x_k \rangle \end{aligned}$$

$$\begin{aligned} \frac{1}{2} \|x_k - x_{k-1}\|^2 &= \frac{1}{2} \|x_{k-1} - x\|^2 - \langle x_{k-1} - x_k, x_k - x \rangle - \frac{1}{2} \|x_k - x\|^2 \\ &\leq \frac{1}{2} \|x_{k-1} - x\|^2 + \frac{1}{\beta_k} \langle \nabla f(z_k), x - x_k \rangle - \frac{1}{2} \|x_k - x\|^2 \end{aligned}$$

Note $L\gamma_k \leq \beta_k$

FOM Framework: Convergence

Main inequality:

$$f(y_k) - f(x) \leq (1 - \gamma_k)[f(y_{k-1}) - f(x)] + \frac{\beta_k \gamma_k}{2} (\|x_{k-1} - x\|^2 - \|x_k - x\|^2)$$

Main estimation:

$$f(y_k) - f(x) \leq \frac{\Gamma_k(1 - \gamma_1)}{\Gamma_1} (f(y_0) - f(x)) + \underbrace{\frac{\Gamma_k}{2} \sum_{i=1}^k \frac{\beta_i \gamma_i}{\Gamma_i} (\|x_{i-1} - x\|^2 - \|x_i - x\|^2)}_{(*)}$$

$$\begin{aligned} (*) &= \frac{\beta_1 \gamma_1}{\Gamma_1} \|x_0 - x\|^2 + \sum_{i=2}^k \left(\frac{\beta_i \gamma_i}{\Gamma_i} - \frac{\beta_{i-1} \gamma_{i-1}}{\Gamma_{i-1}} \right) \|x_{i-1} - x\|^2 - \beta_k \gamma_k \Gamma_k \|x_k - x\|^2 \\ &\leq \frac{\beta_1 \gamma_1}{\Gamma_1} \|x_0 - x\|^2 + \sum_{i=2}^k \left(\frac{\beta_i \gamma_i}{\Gamma_i} - \frac{\beta_{i-1} \gamma_{i-1}}{\Gamma_{i-1}} \right) \cdot D_X^2 \quad (\text{here } D_X = \sup_{x,y \in X} \|x - y\|) \end{aligned}$$

Observation:

$$\text{If } \frac{\beta_k \gamma_k}{\Gamma_k} \geq \frac{\beta_{k-1} \gamma_{k-1}}{\Gamma_{k-1}} \Rightarrow (*) \leq \frac{\beta_k \gamma_k}{\Gamma_k} D_X^2 \Rightarrow f(y_k) - f(x) \leq \frac{\beta_k \gamma_k}{2} D_X^2$$

$$\text{If } \frac{\beta_k \gamma_k}{\Gamma_k} \leq \frac{\beta_{k-1} \gamma_{k-1}}{\Gamma_{k-1}} \Rightarrow (*) \leq \frac{\beta_1 \gamma_1}{\Gamma_1} \|x_0 - x\|^2 \Rightarrow f(y_k) - f(x) \leq \Gamma_k \frac{\beta_1 \gamma_1}{2} \|x_0 - x\|^2$$

FOM Framework: Convergence

Main results:

1 Let $\beta_k = L$, $\gamma_k = \frac{1}{k} \Rightarrow \Gamma_k = \frac{1}{k}$, $\frac{\beta_k \gamma_k}{\Gamma_k} = L$. We have

$$f(y_k) - f(x^*) \leq \frac{L}{2k} D_X^2, \quad f(y_k) - f(x^*) \leq \frac{L}{2k} \|x_0 - x^*\|^2$$

2 Let $\beta_k = \frac{2L}{k}$, $\gamma_k = \frac{2}{k+1} \Rightarrow \Gamma_k = \frac{2}{k(k+1)}$, $\frac{\beta_k \gamma_k}{\Gamma_k} = 2L$. We have

$$f(y_k) - f(x^*) \leq \frac{2L}{k(k+1)} D_X^2, \quad f(y_k) - f(x^*) \leq \frac{4L}{k(k+1)} \|x_0 - x^*\|^2$$

3 Let $\beta_k = \frac{3L}{k+1}$, $\gamma_k = \frac{3}{k+2} \Rightarrow \Gamma_k = \frac{6}{k(k+1)(k+2)}$, $\frac{\beta_k \gamma_k}{\Gamma_k} = \frac{3Lk}{2} \geq \frac{\beta_{k-1} \gamma_{k-1}}{\Gamma_{k-1}}$.
We have

$$f(y_k) - f(x^*) \leq \frac{9L}{2(k+1)(k+2)} D_X^2$$