Chapter 7

Analysis of (Fast) Proximal Gradient Method
7.1. Convergence of proximal gradient method

to minimize $g + h$, choose $x^0$ and repeat

$$x^k = \text{prox}_{t_k h} \left( x^{k-1} - t \nabla g(x^{k-1}) \right), \quad k \geq 1$$

assumptions

- $g$ convex with dom $g = \mathbb{R}^n$; $\nabla g$ 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 $\text{prox}_{t h}$ is well defined)

- optimal value $f^*$ is finite and attained at $x^*$ (not necessarily unique)

convergence result: $1/k$ rate convergence with fixed step size $t_k = 1/L$
Gradient map

\[ G_t(x) = \frac{1}{t}(x - \text{prox}_{th}(x - t\nabla g(x))) \]

\( G_t(x) \) is the negative “step” in the proximal gradient update

\[ x^+ = \text{prox}_{th}(x - t\nabla g(x)) = x - tG_t(x) \]

- \( G_t(x) \) is not a gradient or subgradient of \( f = g + h \)
- from subgradient definition of prox-operator

\[ G_t(x) \in \nabla g(x) + \partial h(x - tG_t(x)) \]

- \( G_t(x) = 0 \) iff \( x \) minimizes \( f(x) = g(x) + h(x) \)
Consequences of Lipschitz assumption

recall upper bound for convex $g$ with Lipschitz continuous gradient

$$g(y) \leq g(x) + \nabla g(x) \top (y - x) + \frac{L}{2} \|y - x\|_2^2, \quad \forall x, y$$

• substitute $y = x - tG_t(x)$:

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x) \top G_t(x) + \frac{t^2 L}{2} \|G_t(x)\|_2^2$$

• if $0 < t \leq 1/L$, then

$$g(x - tG_t(x)) \leq g(x) - t\nabla g(x) \top G_t(x) + \frac{t}{2} \|G_t(x)\|_2^2 \quad (1)$$
A global inequality

if the inequality (1) holds, then for all $z$,

$$f(x - tG_t(x)) \leq f(z) + G_t(x)^\top (x - z) - \frac{t}{2} \|G_t(x)\|^2_2 \quad (2)$$

proof: (define $v = G_t(x) - \nabla g(x)$)

$$f(x - tG_t(x)) \leq g(x) - t\nabla g(x)^\top G_t(x) + \frac{t}{2} \|G_t(x)\|^2_2 + h(x - tG_t(x))$$
$$\leq g(z) + \nabla g(x)^\top (x - z) - t\nabla g(x)^\top G_t(x) + \frac{t}{2} \|G_t(x)\|^2_2$$
$$+ h(z) + v^\top (x - z - tG_t(x))$$
$$= g(z) + h(z) + G_t(x)^\top (x - z) - \frac{t}{2} \|G_t(x)\|^2_2$$

line 2 follows from convexity of $g$ and $h$, and $v \in \partial h(x - tG_t(x))$
Progress in one iteration

\[ x^+ = x - tG_t(x) \]

- Inequality (2) with \( z = x \) shows the algorithm is a descent method:
  \[
  f(x^+) \leq f(x) - \frac{t}{2} \|G_t(x)\|^2_2
  \]

- Inequality (2) with \( z = x^* \):
  \[
  f(x^+) - f^* \leq G_t(x)^\top (x - x^*) - \frac{t}{2} \|G_t(x)\|^2_2 \\
  = \frac{1}{2t} \left( \|x - x^*\|^2_2 - \|x - x^* - tG_t(x)\|^2_2 \right) \\
  = \frac{1}{2t} \left( \|x - x^*\|^2_2 - \|x^+ - x^*\|^2_2 \right) \\
  \]
  (hence, \( \|x^+ - x^*\|_2 \leq \|x - x^*\|_2 \), i.e., distance to optimal set decreases)
Analysis for fixed step size

add inequalities (3) for $x = x^{i-1}$, $x^+ = x^i$, $t = t_i = 1/L$

$$
\sum_{i=1}^{k}(f(x^i) - f^*) \leq \frac{1}{2t} \sum_{i=1}^{k} \left( \|x^{i-1} - x^*\|_2^2 - \|x^i - x^*\|_2^2 \right)
= \frac{1}{2t} \left( \|x^0 - x^*\|_2^2 - \|x^k - x^*\|_2^2 \right)
\leq \frac{1}{2t} \|x^0 - x^*\|_2^2
$$

since $f(x^i)$ is nonincreasing,

$$
f(x^k) - f^* \leq \frac{1}{k} \sum_{i=1}^{k} (f(x^i) - f^*) \leq \frac{1}{2kt} \|x^0 - x^*\|_2^2
$$

conclusion: reaches $f(x^k) - f^* \leq \epsilon$ after $O(1/\epsilon)$ iterations
Quadratic program with box constraints

\[
\begin{align*}
\text{min} & \quad (1/2)x^\top Ax + b^\top x \\
\text{s.t.} & \quad 0 \leq x \leq 1
\end{align*}
\]

\[n = 3000; \text{ fixed step size } t = 1/\lambda_{\text{max}}(A)\]
1-norm regularized least-squares

\[
\min \frac{1}{2} \|Ax - b\|_2^2 + \|x\|_1
\]

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

- the analysis for fixed step size starts with the inequality

\[ g(x - tG_t(x)) \leq g(x) - t\nabla g(x) \top G_t(x) + \frac{t}{2} \|G_t(x)\|^2 \]  

(1)

This inequality is known to hold for \(0 < t \leq 1/L\)

- if \(L\) is not known, we can satisfy (1) by a backtracking line search: start at some \(t := \hat{t} > 0\) and backtrack \((t := \beta t)\) until (1) holds

- step size \(t\) selected by the line search satisfies \(t \geq t_{\text{min}} = \min\{\hat{t}, \beta/L\}\)

- requires one evaluation of \(g\) and \(\text{prox}_{th}\) per line search iteration
Example

Example: line search for projected gradient method

\[ x^+ = P_C(x - t\nabla g(x)) = x - tG_t(x) \]

backtrack until \( x - tG_t(x) \) satisfies “sufficient decrease” inequality (1)
Analysis with line search

from page 6, if (1) holds in iteration \(i\), then \(f(x^i) < f(x^{i-1})\) and

\[
f(x^i) - f^* \leq \frac{1}{2t_i} (\|x^{i-1} - x^*\|_2^2 - \|x^i - x^*\|_2^2) \\
\leq \frac{1}{2t_{min}} (\|x^{i-1} - x^*\|_2^2 - \|x^i - x^*\|_2^2)
\]

- adding inequalities for \(i = 1\) to \(i = k\) gives

\[
\sum_{i=1}^{k} (f(x^i) - f^*) \leq \frac{1}{2t_{min}} \|x^0 - x^*\|_2^2
\]

- since \(f(x^i)\) is nonincreasing, obtain similar \(1/k\) bound as for fixed \(t_i\):

\[
f(x^k) - f^* \leq \frac{1}{2kt_{min}} \|x^0 - x^*\|_2^2
\]
7.2. Fast proximal gradient methods
Fast (proximal) gradient methods


- several recent variations and extensions

**this lecture:**
FISTA and Nesterov’s 2nd method (1988) as presented by Tseng
FISTA (basic version)

\[ \min f(x) = g(x) + h(x) \]

- \( g \) convex, differentiable, with \( \text{dom } g = \mathbb{R}^n \)
- \( h \) closed, convex, with inexpensive \( \text{prox}_{th} \) operator

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

\[
\begin{align*}
y &= x^{k-1} + \frac{k-2}{k+1}(x^{k-1} - x^{k-2}) \\
x^k &= \text{prox}_{th}(y - t \nabla g(y))
\end{align*}
\]

- step size \( t_k \) fixed or 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
  \(y\)

\[ x^{(k)} = \text{prox}_{t_k h}(y - t_k \nabla g(y)) \]

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

\[
\min \log \sum_{i=1}^{m} \exp(a_i^\top x + b_i)
\]

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 $\text{dom } g = \mathbb{R}^n$; $\nabla g$ 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 $t = 1/L$

• 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

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

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

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

substituting expression for $v^k$ in formula for $y$ gives FISTA
Important inequalities

choice of $\theta_k$: 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

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

upper bound on $h$ from definition of prox-operator

$$h(u) \leq h(z) + \frac{1}{t} (w - u)^\top (u - z) \quad \forall w, u = \text{prox}_{th}(w), z$$
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 $0 < t \leq 1/L$
  \[
g(x^+) \leq g(y) + \nabla g(y)^\top (x^+ - y) + \frac{1}{2t} \|x^+ - y\|_2^2 \tag{1}\]

- upper bound from definition of prox-operator:
  \[
h(x^+) \leq h(z) + \nabla g(y)^\top (z - x^+) + \frac{1}{t}(x^+ - y)^\top (z - x^+) \quad \forall z
  \]

- add the upper bounds and use convexity of $g$
  \[
f(x^+) \leq f(z) + \frac{1}{t}(x^+ - y)^\top (z - x^+) + \frac{1}{2t} \|x^+ - y\|_2^2 \quad \forall z
  \]
• make convex combination of upper bounds for $z = x$ and $z = x^*$

$$f(x^+) - f^* - (1 - \theta)(f(x) - f^*)$$

$$= f(x^+) - \theta f^* - (1 - \theta)f(x)$$

$$\leq \frac{1}{t_i} \left( \frac{1}{2} \right) (x^+ - y)^\top (\theta x^* + (1 - \theta)x - x^+) + \frac{1}{2t} \|x^+ - y\|_2^2$$

$$= \frac{1}{2t_i} \left( \frac{1}{2} \right) \left( \| y - (1 - \theta)x - \theta x^* \|_2^2 - \| x^+ - (1 - \theta)x - \theta x^* \|_2^2 \right)$$

$$= \frac{1}{2t_i} \left( \frac{1}{2} \right) \left( \| v - x^* \|_2^2 - \| v^+ - x^* \|_2^2 \right)$$

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

$$\frac{t_i}{\theta^2_i} \left( f(x^i) - f^* \right) + \frac{1}{2} \| v^i - x^* \|_2^2$$

$$\leq \frac{(1 - \theta_i)t_i}{\theta^2_i} \left( f(x^{i-1}) - f^* \right) + \frac{1}{2} \| v^{i-1} - x^* \|_2^2 \quad (2)$$
**Analysis for fixed step size**

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

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

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 $O(1/\sqrt{\epsilon})$ iterations
Example: quadratic program with box constraints

\[
\begin{align*}
\min & \quad \frac{1}{2} x^\top A x + b^\top x \\
\text{s.t.} & \quad 0 \leq x \leq 1
\end{align*}
\]

\( n = 3000; \) fixed step size \( t = 1/\lambda_{\max}(A) \)
1-norm regularized least-squares

\[
\min \frac{1}{2} \| Ax - b \|_2^2 + \| x \|_1
\]

randomly generated \( A \in \mathbb{R}^{2000 \times 1000} \); step \( t_k = 1/L \) with \( L = \lambda_{\text{max}}(A^T A) \)
7.2.1. FISTA with line search
Key steps in the analysis of FISTA

- the starting point is the inequality

\[ g(x^+) \leq g(y) + \nabla g(y)^\top (x^+ - y) + \frac{1}{2t} \| x^+ - y \|_2^2 \]  

(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

\[
\frac{t_i}{\theta_i^2} \left( f(x^i) - f^* \right) + \frac{1}{2} \| v^i - x^* \|_2^2 \\
\leq \frac{(1-\theta_i) t_i}{\theta_i^2} \left( f(x^{i-1}) - f^* \right) + \frac{1}{2} \| v^{i-1} - x^* \|_2^2
\]

(2)

- to combine these inequalities recursively, we need

\[
\frac{(1 - \theta_i) t_i}{\theta_i^2} \leq \frac{t_i-1}{\theta_i^2} \quad (i \geq 2)
\]

(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

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

**FISTA with fixed step size**

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

these values satisfy (1) and (3) with

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

replace update of $x$ in iteration $k$ with

- $t := t_{k-1}$ (define $t_0 = \hat{t} > 0$)
- $x := \text{prox}_{th}(y - t\nabla g(y))$
- \textbf{while} $g(x) > g(y) + \nabla g(y)^	op (x - y) + \frac{1}{2t}\|x - y\|_2^2$
  - \textit{a)} $t := \beta t$
  - \textit{b)} $x := \text{prox}_{th}(y - t\nabla g(y))$
- \textbf{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 $\nabla g$ guarantees $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
• preserves $1/k^2$ convergence rate because $\theta_k^2/t_k = O(1/k^2)$:
  $$\frac{\theta_k^2}{t_k} \leq \frac{4}{(k + 1)^2 t_{\min}}$$
**Descent version of FISTA**

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

\[
\begin{align*}
  y &= (1 - \theta_k)x^{k-1} + \theta_kv^{k-1} \\
  u &= \text{prox}_{t_k h}(y - t_k \nabla g(y)) \\
  x^k &= \begin{cases} 
    u & \text{if } f(u) \leq f(x^{k-1}) \\
    x^{k-1} & \text{o.w.}
  \end{cases} \\
  v^k &= x^{k-1} + \frac{1}{\theta_k}(x^k - x^{k-1})
\end{align*}
\]

• 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 22: replace $x^+$ with $u$ and use $f(x^+) \leq f(u)$
Example

(from page 18)
Nesterov’s second method

algorithm: choose \( x^0 = v^0 \); for \( k \geq 1 \), repeat

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

- 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; $\nabla g$ is Lipschitz continuous on $\text{dom } h \subset \text{dom } g$

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

• $h$ 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: $f(x^k) - f^*$ decreases 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 $0 < t \leq 1/L$

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

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

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

• from convexity of $g$, $h$

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

$$h(x^+) \leq (1 - \theta)h(x) + \theta h(v^+)$$
• upper bound on $h$ from page 21 (with $u = v^+$, $w = v - (t/\theta)\nabla g(y)$)

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

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

$$f(x^+) \leq (1 - \theta)f(x) + \theta f^* - \frac{\theta^2}{t} (v^+ - v)^\top(v^+ - x^+) + \frac{\theta^2}{2t} \|v^+ - v\|_2^2$$

$$= (1 - \theta)f(x) + \theta f^* + \frac{\theta^2}{2t} (\|v - x^*\|_2^2 - \|v^* - x^*\|_2^2)$$

this is identical to the final inequality (2) in the analysis of FISTA on page 23

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

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