9. Accelerated proximal gradient methods

- Nesterov’s method
- analysis with fixed step size
- line search
Proximal gradient method

Results from lecture 6

- each proximal gradient iteration is a descent step:

\[ f(x^{(k)}) < f(x^{(k-1)}), \quad \|x^{(k)} - x^*\|^2 \leq c \|x^{(k-1)} - x^*\|^2 \]

with \( c = 1 - m/L \)

- suboptimality after \( k \) iterations is \( O(1/k) \):

\[ f(x^{(k)}) - f^* \leq \frac{L}{2k} \|x^{(0)} - x^*\|^2 \]

Accelerated proximal gradient methods

- to improve convergence, we add a momentum term
- we relax the descent property
- originated in work by Nesterov in the 1980s
Assumptions

we consider the same problem and make the same assumptions as in lecture 6:

\[ \text{minimize} \quad f(x) = g(x) + h(x) \]

• \( h \) is closed and convex (so that \( \text{prox}_{th} \) is well defined)

• \( g \) is differentiable with \( \text{dom } g = \mathbb{R}^n \)

• there exist constants \( m \geq 0 \) and \( L > 0 \) such that the functions

\[ g(x) - \frac{m}{2} x^T x, \quad \frac{L}{2} x^T x - g(x) \]

are convex

• the optimal value \( f^* \) is finite and attained at \( x^* \) (not necessarily unique)
Nesterov’s method

Algorithm: choose \( x^{(0)} = v^{(0)} \) and \( \gamma_0 > 0 \); for \( k \geq 1 \), repeat the steps

- define \( \gamma_k = \theta_k^2 / t_k \) where \( \theta_k \) is the positive root of the quadratic equation

\[
\frac{\theta_k^2}{t_k} = (1 - \theta_k) \gamma_{k-1} + m \theta_k
\]

- update \( x^{(k)} \) and \( v^{(k)} \) as follows:

\[
y = x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m \theta_k} (v^{(k-1)} - x^{(k-1)})
\]

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

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

stepsize \( t_k \) is fixed \( (t_k = 1/L) \) or obtained from line search
Momentum interpretation

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

$$
y = x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m\theta_k} (y^{(k-1)} - x^{(k-1)})
= x^{(k-1)} + \frac{\theta_k \gamma_{k-1}}{\gamma_{k-1} + m\theta_k} \left( \frac{1}{\theta_{k-1}} - 1 \right) (x^{(k-1)} - x^{(k-2)})
$$

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

Accelerated proximal gradient methods
Algorithm parameters

\[ \frac{\theta^2_k}{t_k} = (1 - \theta_k)\gamma_{k-1} + m\theta_k, \quad \gamma_k = \frac{\theta^2_k}{t_k} \]

- \( \theta_k \) is positive root of the quadratic equation
- \( \theta_k < 1 \) if \( mt_k < 1 \)
- if \( t_k \) is constant, sequence \( \theta_k \) is completely determined by starting value \( \gamma_0 \)

Example: \( L = 1, m = 0.1, t_k = 1/L \)
FISTA

if we take \( m = 0 \) on page 9-4, the expression for \( y \) simplifies:

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

eliminating the variables \( v^{(k)} \) gives the equivalent iteration (for \( k \geq 2 \))

\[
\begin{align*}
y & = x^{(k-1)} + \theta_k \left( \frac{1}{\theta_{k-1}} - 1 \right) (x^{(k-1)} - x^{(k-2)}) \\
x^{(k)} & = \text{prox}_{t_k h}(y - t_k \nabla g(y))
\end{align*}
\]

this is known as FISTA (‘Fast Iterative Shrinkage-Thresholding Algorithm’)

Accelerated proximal gradient methods
Example

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

- two randomly generated problems with $m = 2000$, $n = 1000$
- same fixed step size used for gradient method and FISTA
- figures show $(f(x^{(k)}) - f^*)/f^*$
Nesterov’s simplest method

• if $m > 0$ and we choose $\gamma_0 = m$, then

$$\gamma_k = m, \quad \theta_k = \sqrt{m t_k} \quad \text{for all } k \geq 1$$

• the algorithm on p. 9-4 and p. 9-5 simplifies:

$$y = x^{(k-1)} + \frac{\sqrt{t_k}}{\sqrt{t_{k-1}}} \frac{1 - \sqrt{m t_{k-1}}}{1 + \sqrt{m t_k}} (x^{(k-1)} - x^{(k-2)})$$

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

• with constant stepsize $t_k = 1/L$, the expression for $y$ reduces to

$$y = x^{(k-1)} + \frac{1 - \sqrt{m/L}}{1 + \sqrt{m/L}} (x^{(k-1)} - x^{(k-2)})$$
Outline

- Nesterov’s method
- analysis with fixed step size
- line search
Overview

• we show that if $t_k = 1/L$, the following inequality holds at each iteration:

\[
\begin{align*}
  f(x^{(k)}) - f^* &+ \gamma_k \frac{\|v^{(k)} - x^*\|^2}{2} \\
  \leq & (1 - \theta_k) \left( f(x^{(k-1)}) - f^* + \frac{\gamma_{k-1}}{2} \|v^{(k-1)} - x^*\|^2 \right)
\end{align*}
\]

• therefore the rate of convergence is determined by $\lambda_k = \prod_{i=1}^{k} (1 - \theta_i)$:

\[
\begin{align*}
  f(x^{(k)}) - f^* &\leq f(x^{(k)}) - f^* + \gamma_k \frac{\|v^{(k)} - x^*\|^2}{2} \\
  &\leq \lambda_k \left( f(x^{(0)}) - f^* + \frac{\gamma_0}{2} \|x^{(0)} - x^*\|^2 \right)
\end{align*}
\]

(here we assume that $x^{(0)} \in \text{dom } h = \text{dom } f$)
Notation for one iteration

quantities in iteration $i$ of the algorithm on page 9-4

- define $t = t_i$, $\theta = \theta_i$, $\gamma = \gamma_i$, and $\gamma^+ = \gamma_i$:

$$
\gamma^+ = (1 - \theta) \gamma + m \theta,
\gamma^+ = \theta^2 / t
$$

- define $x = x^{(i-1)}$, $x^+ = x^{(i)}$, $v = v^{(i-1)}$, and $v^+ = v^{(i)}$:

$$
y = \frac{1}{\gamma + m \theta} (\gamma^+ x + \theta \gamma v)
$$

$$
x^+ = y - t G_t(y)
$$

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

- $v^+$, $v$, and $y$ are related as

$$
\gamma^+ v^+ = (1 - \theta) \gamma v + m \theta y - \theta G_t(y)
$$
\textbf{Proof (last identity):} \\

- combine $v$ and $x$ updates and use $\gamma^+ = \theta^2 / t$: \\

\[
v^+ = x + \frac{1}{\theta}(y - tG_t(y) - x) = \frac{1}{\theta}(y - (1 - \theta)x) - \frac{\theta}{\gamma^+}G_t(y)
\]

- multiply with $\gamma^+ = \gamma + m\theta - \theta \gamma$: \\

\[
\gamma^+ v^+ = \frac{\gamma^+}{\theta}(y - (1 - \theta)x) - \theta G_t(y)
\]

\[
= \frac{(1 - \theta)}{\theta}((\gamma + m\theta)y - \gamma^+x) + \theta my - \theta G_t(y)
\]

\[
= (1 - \theta)\gamma v + \theta my - \theta G_t(y)
\]
Bounds on objective function

recall the results on the proximal gradient update (page 6-13):

• if $0 < t \leq 1/L$ then $g(x^+) = g(y - tG_t(y))$ is bounded by

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

(2)

• if the inequality (2) holds, then $mt \leq 1$ and, for all $z$,

$$f(z) \geq f(x^+) + \frac{t}{2}\|G_t(y)\|^2_2 + G_t(y)^T (z - y) + \frac{m}{2}\|z - y\|^2_2$$

• combine the inequalities for $z = x$ and $z = x^*$:

$$f(x^+) - f^* \leq (1 - \theta)(f(x) - f^*) - G_t(y)^T ((1 - \theta)x + \theta x^* - y)$$
$$- \frac{t}{2}\|G_t(y)\|^2_2 - \frac{m\theta}{2}\|x^* - y\|^2_2$$
Progress in one iteration

• the definition of $\gamma^+$ and (1) imply that

$$\frac{\gamma^+}{2}(\|x^* - v^+\|^2_2 - \|y - v^+\|^2_2)$$

$$= \frac{(1 - \theta)\gamma}{2}(\|x^* - v\|^2_2 - \|y - v\|^2_2) + \frac{m\theta}{2}\|x^* - y\|^2_2 + \theta G_t(y)^T(x^* - y)$$

• combining this with the last inequality on page 9-13 gives

$$f(x^+) - f^* + \frac{\gamma^+}{2}\|x^* - v^+\|^2_2$$

$$\leq (1 - \theta)\left(f(x) - f^* + \frac{\gamma}{2}\|x^* - v\|^2_2 - G_t(y)^T(x - y) - \frac{\gamma}{2}\|y - v\|^2_2\right)$$

$$- \frac{t}{2}\|G_t(y)\|^2_2 + \frac{\gamma^+}{2}\|y - v^+\|^2_2$$
the last term on the right-hand side is

\[
\frac{\gamma^+}{2} \|y - v^+\|_2^2 = \frac{1}{2\gamma^+} \|(1 - \theta)\gamma(y - v) + \theta G_t(y)\|_2^2
\]

\[
= \frac{(1 - \theta)^2 \gamma^2}{2\gamma^+} \|y - v\|_2^2 + \frac{\theta(1 - \theta)\gamma}{\gamma^+} G_t(y)^T(y - v) + \frac{t}{2} \|G_t(y)\|_2^2
\]

\[
= (1 - \theta) \left( \frac{\gamma (\gamma^+ - m\theta)}{2\gamma^+} \|y - v\|_2^2 + G_t(y)^T(x - y) \right) + \frac{t}{2} \|G_t(y)\|_2^2
\]

last step uses definitions of \( \gamma^+ \) and \( y \) (chosen so that \( \theta \gamma(y - v) = \gamma^+(x - y) \))

• substituting this in the last inequality on page 9-14 gives the result on page 9-10

\[
f(x^+) - f^* + \frac{\gamma^+}{2} \|x^* - v^+\|_2^2
\]

\[
\leq (1 - \theta) \left( f(x) - f^* + \frac{\gamma}{2} \|x^* - v\|_2^2 \right) - \frac{(1 - \theta)\gamma m\theta}{2} \gamma^+ \|y - v\|_2^2
\]

\[
\leq (1 - \theta) \left( f(x) - f^* + \frac{\gamma}{2} \|x^* - v\|_2^2 \right)
\]
Analysis for fixed step size

the product $\lambda_k = \prod_{i=1}^{k} (1 - \theta_i)$ determines the rate of convergence (page 9-10)

• the sequence $\lambda_k$ satisfies the following bound (proof on next page)

$$
\lambda_k \leq \frac{4}{(2 + \sqrt{\gamma_0} \sum_{i=1}^{k} \sqrt{t_i})^2}
$$

• for constant step size $t_k = 1/L$, we obtain

$$
\lambda_k \leq \frac{4}{(2 + k\sqrt{\gamma_0/L})^2}
$$

• combined with the inequality on p. 9-10, this shows the $1/k^2$ convergence rate:

$$
f(x^{(k)}) - f^* \leq \frac{4}{(2 + k\sqrt{\gamma_0/L})^2} \left( f(x^{(0)}) - f^* + \frac{\gamma_0}{2} ||x^{(0)} - x^*||_2^2 \right)
$$
Proof.

• recall that $\gamma_k$ and $\theta_k$ are defined by $\gamma_k = (1 - \theta_k)\gamma_{k-1} + \theta_k m$ and $\gamma_k = \theta_k^2 / t_k$

• we first note that $\lambda_k \leq \gamma_k / \gamma_0$; this follows from

$$\lambda_k = (1 - \theta_k)\lambda_{k-1} = \frac{\gamma_k - \theta_k m}{\gamma_{k-1}}\lambda_{k-1} \leq \frac{\gamma_k}{\gamma_{k-1}}\lambda_{k-1}$$

• the inequality follows by combining from $i = 1$ to $i = k$ the inequalities

$$\frac{1}{\sqrt{\lambda_i}} - \frac{1}{\sqrt{\lambda_{i-1}}} \geq \frac{\lambda_{i-1} - \lambda_i}{2\lambda_{i-1}\sqrt{\lambda_i}} \quad \text{(because } \lambda_i \leq \lambda_{i-1})$$

$$= \frac{\theta_i}{2\sqrt{\lambda_i}}$$

$$\geq \frac{\theta_i}{2\sqrt{\gamma_i / \gamma_0}}$$

$$= \frac{1}{2\sqrt{\gamma_0 t_i}}$$
Strongly convex functions

the following bound on $\lambda_k$ is useful for strongly convex functions ($m > 0$)

- if $\gamma_0 \geq m$ then $\gamma_k \geq m$ for all $k$ and

$$\lambda_k \leq \prod_{i=1}^{k} (1 - \sqrt{mt_i})$$

(proof on next page)

- for constant step size $t_k = 1/L$, we obtain

$$\lambda_k \leq \left(1 - \sqrt{m/L}\right)^k$$

- combined with the inequality on p. 9-10, this shows

$$f(x^{(k)}) - f^* \leq \left(1 - \sqrt{\frac{m}{L}}\right)^k \left(f(x^{(0)}) - f^*\right) + \frac{\gamma_0}{2}\|x^{(0)} - x^*\|^2_2$$
Proof.

- if $\gamma_{k-1} \geq m$, then

$$
\gamma_k = (1 - \theta_k)\gamma_{k-1} + \theta_km \\
\geq m
$$

- since $\gamma_0 \geq m$, we have $\gamma_k \geq m$ for all $k$

- it follows that $\theta_i = \sqrt{\gamma_i t_i} \geq \sqrt{mt_i}$ and

$$
\lambda_k = \prod_{i=1}^{k} (1 - \theta_i) \leq \prod_{i=1}^{k} (1 - \sqrt{mt_i})
$$
Outline

- Nesterov’s method
- analysis with fixed step size
- line search
Line search

• the analysis for fixed step size starts with the inequality (2):

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

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

• if \( L \) is not known, we can satisfy (2) by a backtracking line search:
  
  start at some \( t := \hat{t} > 0 \) and backtrack \( t := \beta t \) until (2) holds

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

• for each tentative \( t_k \) we need to recompute \( \theta_k, y, x^{(k)} \) in the algorithm on p. 9-4

• requires evaluations of \( \nabla g, \text{prox}_{\theta_k}, \) and \( g \) (twice) per line search iteration
Analysis with line search

• from page 9-16:

\[ \lambda_k \leq \frac{4}{(2 + \sqrt{\gamma_0} \sum_{i=1}^{k} \sqrt{t_i})^2} \leq \frac{4}{(2 + k \sqrt{\gamma_0 t_{\text{min}}})^2} \]

• from page 9-18, if \( \gamma_0 \geq m \):

\[ \lambda_k \leq \prod_{i=1}^{k} (1 - \sqrt{mt_i}) \leq (1 - \sqrt{mt_{\text{min}}})^k \]

• therefore the results for fixed step size hold with \( 1/t_{\text{min}} \) substituted for \( L \)
Accelerated gradient methods

  The material in the lecture is from §2.2 of this book.

FISTA


Line search strategies

- FISTA papers by Beck and Teboulle.
Interpretation and insight


Implementation