7. Accelerated proximal gradient methods

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

Results from lecture 4

• each proximal gradient iteration is a descent step (page 4.14 and 4.16):

\[ f(x_{k+1}) < f(x_k), \quad \|x_{k+1} - x^*\|^2_2 \leq c \|x_k - x^*\|^2_2 \]

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

• suboptimality after \( k \) iterations is \( O(1/k) \) (page 4.15):

\[ f(x_k) - f^* \leq \frac{L}{2k}\|x_0 - x^*\|^2_2 \]

Accelerated proximal gradient methods

• to improve convergence, we add a momentum term

• we relax the descent properties

• originated in work by Nesterov in the 1980s
Assumptions

we consider the same problem and make the same assumptions as in lecture 4:

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

- \( h \) is closed and convex (so that \( \text{prox}_{\theta h} \) 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

choose $x_0 = v_0$ and $\theta_0 \in (0, 1]$, and repeat the following steps for $k = 0, 1, \ldots$

- if $k \geq 1$, define $\theta_k$ as the positive root of the quadratic equation

$$\frac{\theta_k^2}{t_k} = (1 - \theta_k)\gamma_k + m\theta_k \quad \text{where} \quad \gamma_k = \frac{\theta_{k-1}^2}{t_{k-1}}$$

- update $x_k$ and $v_k$ as follows:

$$y = x_k + \frac{\theta_k \gamma_k}{\gamma_k + m\theta_k} (v_k - x_k) \quad \text{for} \quad k \geq 0 \quad \text{with} \quad y = x_0 \quad \text{if} \quad k = 0$$

$$x_{k+1} = \text{prox}_{t_k h}(y - t_k \nabla g(y))$$

$$v_{k+1} = x_k + \frac{1}{\theta_k} (x_{k+1} - x_k)$$

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

- the first iteration \((k = 0)\) is a proximal gradient step at \(y = x_0\)
- next iterations are proximal gradient steps at extrapolated points \(y\):

\[
y = x_k + \frac{\theta_k \gamma_k}{\gamma_k + m \theta_k} (v_k - x_k) = x_k + \beta_k (x_k - x_{k-1})
\]

where

\[
\beta_k = \frac{\theta_k \gamma_k}{\gamma_k + m \theta_k} \left( \frac{1}{\theta_{k-1}} - 1 \right) = \frac{t_k \theta_{k-1} (1 - \theta_{k-1})}{t_{k-1} \theta_k + t_k \theta^2_{k-1}}
\]

\[x_{k+1} = \text{prox}_{t_k h}(y - t_k \nabla g(y))\]
Parameters $\theta_k$ and $\beta_k$ (for fixed stepsize $t_k = 1/L = 1$)
Parameter $\theta_k$

- for $k \geq 1$, $\theta_k$ is the positive root of the quadratic equation

$$\frac{\theta_k^2}{t_k} = (1 - \theta_k)\frac{\theta_{k-1}^2}{t_{k-1}} + m\theta_k$$

- if $m > 0$ and $\theta_0 = \sqrt{mt_0}$, then $\theta_k = \sqrt{mt_k}$ for all $k$

- $\theta_k < 1$ if $mt_k < 1$

- for constant $t_k$, sequence $\theta_k$ is completely determined by $\theta_0$
FISTA

if we take $m = 0$ on page 7.4, the expression for $y$ simplifies:

$$y = x_k + \theta_k (v_k - x_k)$$
$$x_{k+1} = \text{prox}_{t_k h}(y - t_k \nabla g(y))$$
$$v_{k+1} = x_k + \frac{1}{\theta_k}(x_{k+1} - x_k)$$

eliminating the variables $v^{(k)}$ gives the equivalent iteration

$$y = x_k + \theta_k \left( \frac{1}{\theta_{k-1}} - 1 \right) (x_k - x_{k-1}) \quad (y = x_0 \text{ if } k = 0)$$
$$x_{k+1} = \text{prox}_{t_k h}(y - t_k \nabla g(y))$$

this is known as **FISTA** (Fast Iterative Shrinkage-Thresholding Algorithm)
Example

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

- two randomly generated problems with \( p = 2000, n = 1000 \)
- same fixed step size used for gradient method and FISTA
- figures show \((f(x^{(k)}) - f^*)/f^*)\
A simplification for strongly convex problems

- if $m > 0$ and we choose $\theta_0 = \sqrt{mt_0}$, then
  \[
  \gamma_k = m, \quad \theta_k = \sqrt{mt_k} \quad \text{for all } k \geq 1
  \]

- the algorithm on page 7.4 and page 7.5 simplifies:
  \[
  y = x_k + \frac{\sqrt{t_k}}{\sqrt{t_{k-1}}} \frac{1 - \sqrt{mt_{k-1}}}{1 + \sqrt{mt_k}} (x_k - x_{k-1}) \quad (y = x_0 \text{ if } k = 0)
  \]
  \[
  x_{k+1} = \text{prox}_{t_k h}(y - t_k \nabla g(y))
  \]

- with constant stepsize $t_k = 1/L$, the expression for $y$ reduces to
  \[
  y = x_k + \frac{1 - \sqrt{m/L}}{1 + \sqrt{m/L}} (x_k - x_{k-1}) \quad (y = x_0 \text{ if } k = 0)
  \]
Outline

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

- we show that if $t_i = 1/L$, the following inequality holds at iteration $i$:

$$f(x_{i+1}) - f^* + \frac{\gamma i+1}{2} \|v_{i+1} - x^*\|_2^2$$

$$\leq (1 - \theta_i)(f(x_i) - f^*) + \frac{\gamma i+1 - m\theta i}{2} \|v_i - x^*\|_2^2$$

$$= (1 - \theta_i) \left( f(x_i) - f^* + \frac{\gamma i}{2} \|v_i - x^*\|_2^2 \right) \quad \text{if } i \geq 1$$

- combining the inequalities from $i = 0$ to $i = k - 1$ shows that

$$f(x_k) - f^* \leq \lambda_k \left( (1 - \theta_0)(f(x_0) - f^*) + \frac{\gamma_1 - m\theta_0}{2} \|x_0 - x^*\|_2^2 \right)$$

$$\leq \lambda_k \left( (1 - \theta_0)(f(x_0) - f^*) + \frac{\theta^2_0}{2t_0} \|x_0 - x^*\|_2^2 \right)$$

where $\lambda_1 = 1$ and $\lambda_k = \prod_{i=1}^{k-1} (1 - \theta_i)$ for $k > 1$

(here we assume $x_0 \in \text{dom } f$)
Notation for one iteration

quantities in iteration $i$ of the algorithm on page 7.4

- define $t = t_i$, $\theta = \theta_i$, $\gamma^+ = \gamma_{i+1} = \theta^2/t$
- if $i \geq 1$, define $\gamma = \gamma_i$ and note that $\gamma^+ - m\theta = (1 - \theta)\gamma$
- define $x = x_i$, $x^+ = x_{i+1}$, $v = v_i$, and $v^+ = v_{i+1}$:

\[
\begin{align*}
y & = \frac{1}{\gamma + m\theta} (\gamma^+ x + \theta \gamma v) \quad (y = x = v \text{ if } i = 0) \\
x^+ & = y - tG_t(y) \\
v^+ & = x + \frac{1}{\theta}(x^+ - x)
\end{align*}
\]

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

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

Accelerated proximal gradient methods 7.12
Proof (last identity):

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

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

- for $i = 0$, the equation (1) follows because $y = x = v$
- for $i \geq 1$, multiply with $\gamma^+ = \gamma + m\theta - \theta\gamma$:

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

recall the results on the proximal gradient update (page 4.12):

- 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 $z$,

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

- add $(1 - \theta)$ times the inequality for $z = x$ and $\theta$ times the inequality for $z = x^*$:

$$f(x^+) - f^* \leq (1 - \theta)(f(x) - f^*) - G_t(y)^T ((1 - \theta)x + \theta x^* - y)$$

$$- \frac{t}{2} \|G_t(y)\|_2^2 - \frac{m\theta}{2} \|x^* - y\|_2^2$$
Bound on distance to optimum

- it follows from (1) that

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

\[
\leq \frac{\gamma^+ - m\theta}{2} \|v - x^*\|^2_2 + \theta G_t(y)^T (x^* - v - \frac{m\theta}{\gamma^+} (y - v)) + \frac{t}{2} \|G_t(y)\|^2_2 + \frac{m\theta}{2} \|x^* - y\|^2_2
\]

- \(\gamma^+\) and \(y\) are chosen so that \(\theta (\gamma^+ - m\theta) (y - v) = \gamma^+ (1 - \theta) (x - y)\); hence

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

• combining the bounds on page 7.14 and 7.15 gives

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

this is the first inequality on page 7.11

• if \( i \geq 1 \), we use \( \gamma^+ - m\theta = (1 - \theta)\gamma \) to write this as

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

the product $\lambda_k = \prod_{i=1}^{k-1} (1 - \theta_i)$ determines the rate of convergence (page 7.11)

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

$$\lambda_k \leq \frac{4}{(2 + \sqrt{\sum_{i=1}^{k-1} \sqrt{t_i}})^2} = \frac{4t_0}{(2\sqrt{t_0} + \theta_0 \sum_{i=1}^{k-1} \sqrt{t_i})^2}$$

(3)

• for constant step size and $\theta_0 = 1$, we obtain

$$\lambda_k \leq \frac{4}{(k + 1)^2}$$

• with $t_0 = 1/L$, the inequality on page 7.11 shows a $1/k^2$ convergence rate

$$f(x_k) - f^* \leq \frac{2L}{(k + 1)^2} ||x_0 - x^*||_2^2$$
Proof.

- recall that for $k \geq 1$,
\[ \gamma_{k+1} = (1 - \theta_k) \gamma_k + \theta_k m, \quad \gamma_k = \theta_{k-1}^2 / t_{k-1} \]

- we first note that $\lambda_k \leq \gamma_k / \gamma_1$; this follows from
\[ \lambda_{i+1} = (1 - \theta_i) \lambda_i = \frac{\gamma_{i+1} - \theta_i m}{\gamma_i} \lambda_i \leq \frac{\gamma_{i+1}}{\gamma_i} \lambda_i \]

- the inequality (3) follows by combining from $i = 1$ to $i = k - 1$ the inequalities
\[ \frac{1}{\sqrt{\lambda_{i+1}}} - \frac{1}{\sqrt{\lambda_i}} \geq \frac{\lambda_i - \lambda_{i+1}}{2 \lambda_i \sqrt{\lambda_{i+1}}} \]
\[ = \frac{\theta_i}{2 \sqrt{\lambda_{i+1}}} \]
\[ \geq \frac{\theta_i}{2 \sqrt{\gamma_{i+1} / \gamma_1}} \]
\[ = \frac{1}{2} \sqrt{\gamma_1 t_i} \]
the following bound on $\lambda_k$ is useful for strongly convex functions ($m > 0$)

- if $\theta_0 \geq \sqrt{mt_0}$, then $\theta_k \geq \sqrt{mt_k}$ for all $k$ and

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

(proof on next page)

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

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

- combined with the inequality on page 7.11, this shows linear convergence

$$f(x_k) - f^* \leq \left(1 - \sqrt{\frac{m}{L}}\right)^{k-1} \left((1 - \theta_0)(f(x_0) - f^*) + \frac{\theta_0^2}{2t_0} \|x_0 - x^*\|^2_2\right)$$
Proof.

• if $\theta_{k-1} \geq \sqrt{mt_{k-1}}$, then $\theta_k \geq \sqrt{mt_k}$:

$$\frac{\theta_k^2}{t_k} = (1 - \theta_k)\frac{\theta_{k-1}^2}{t_{k-1}} + m\theta_k$$

$$\geq (1 - \theta_k)m + m\theta_k$$

$$= m$$

• if $\theta_0 \geq \sqrt{mt_0}$, then $\theta_k \geq \sqrt{mt_k}$ for all $k$ and

$$\lambda_k = \prod_{i=1}^{k-1} (1 - \theta_i) \leq \prod_{i=1}^{k-1} (1 - \sqrt{mt_i})$$
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 \(\theta_k, y, x_{k+1}\) in the algorithm on p. 7.4

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

• from page 7.17, if $\theta_0 = 1$:

$$
\lambda_k \leq \frac{4t_0}{(2\sqrt{t_0} + \sum_{i=1}^{k-1} \sqrt{t_i})^2} \leq \frac{4\hat{t}/t_{min}}{(k + 1)^2}
$$

• from page 7.19, if $\theta_0 \geq \sqrt{mt_0}$:

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

• therefore the results for fixed step size hold with $1/t_{min}$ substituted for $L$
References

Acceleration techniques in optimization


Accelerated proximal gradient methods

  Most of the material in this lecture is from §2.2 in Nesterov’s book.

FISTA

References

Line search strategies

- FISTA papers by Beck and Teboulle.

Implementation