Gradient method

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

- gradient method, first-order methods
- quadratic bounds on convex functions
- analysis of gradient method
Algorithms will be covered in this course

**first-order methods**
- gradient method, line search
- subgradient, proximal gradient methods
- accelerated (proximal) gradient methods

**decomposition and splitting**
- first-order methods and dual reformulations
- alternating minimization methods

**interior-point methods**
- conic optimization
- primal-dual methods for symmetric cones

**semi-smooth Newton methods**
Gradient method

To minimize a convex function differentiable function $f$: choose $x^{(0)}$ and repeat

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \ldots$$

Step size rules
- Fixed: $t_k$ constant
- Backtracking line search
- Exact line search: minimize $f(x - t \nabla f(x))$ over $t$

Advantages of gradient method
- Every iteration is inexpensive
- Does not require second derivatives
Quadratic example

\[ f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \quad (\gamma > 1) \]

with exact line search, \( x^{(0)} = (\gamma, 1) \)

\[
\frac{\|x^{(k)} - x^*\|_2}{\|x^{(0)} - x^*\|_2} = \left(\frac{\gamma - 1}{\gamma + 1}\right)^k
\]

Disadvantages of gradient method

- Gradient method is often slow
- Very dependent on scaling
Nondifferentiable example

\[
f(x) = \sqrt{x_1^2 + \gamma x_2^2} (|x_2| \leq x_1), \quad f(x) = \frac{x_1 + \gamma |x_2|}{\sqrt{1 + \gamma}} (|x_2| > x_1)
\]

with exact line search, \( x^{(0)} = (\gamma, 1) \), converges to non-optimal point

gradient method does not handle nondifferential problems
First-order methods

address one or both disadvantages of the gradient method methods with improved convergence
- quasi-Newton methods
- conjugate gradient method
- accelerated gradient method

methods for nondifferentiable or constrained problems
- subgradient methods
- proximal gradient method
- smoothing methods
- cutting-plane methods
Outline

- gradient method, first-order methods
- quadratic bounds on convex functions
- analysis of gradient method
Convex function

A function $f$ is convex if $\text{dom} f$ is a convex set and Jensen’s inequality holds:

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \forall x, y \in \text{dom} f$$

**First-order condition**

For (continuously) differentiable $f$, Jensen’s inequality can be replaced with

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \forall x, y \in \text{dom} f$$

**Second-order condition**

For twice differentiable $f$, Jensen’s inequality can be replaced with

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \text{dom} f$$
Strictly convex function

\( f \) is strictly convex if \( \text{dom } f \) is convex set and

\[
f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y) \quad \forall x, y \in \text{dom } f, x \neq y, \theta \in (0, 1)
\]

hence, if a minimizer of \( f \) exists, it is unique

First-order condition
for differentiable \( f \), Jensen’s inequality can be replaced with

\[
f(y) > f(x) + \nabla f(x)^\top (y - x) \quad \forall x, y \in \text{dom } f, x \neq y
\]

Second-order condition
note that \( \nabla^2 f(x) \succ 0 \) is not necessary for strict convexity (cf., \( f(x) = x^4 \))
Monotonicity of gradient

differentiable $f$ is convex if and only if $\text{dom } f$ is convex and
\[
(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0 \quad \forall x, y \in \text{dom } f
\]
i.e., $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a monotone mapping

differentiable $f$ is strictly convex if and only if $\text{dom } f$ is convex and
\[
(\nabla f(x) - \nabla f(y))^\top (x - y) > 0 \quad \forall x, y \in \text{dom } f, x \neq y
\]
i.e., $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is a strictly monotone mapping
Proof.

if \( f \) is differentiable and convex, then

\[
f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad f(x) \geq f(y) + \nabla f(y)^\top (x - y)
\]

combining the inequalities gives

\[
(\nabla f(x) - \nabla f(y))^\top (x - y) \geq 0
\]

if \( \nabla f \) is monotone, then \( g'(t) \geq g'(0) \) for \( t \geq 0 \) and \( t \in \text{dom } g \), where

\[
g(t) = f(x + t(y - x)), \quad g'(t) = \nabla f(x + t(y - x))^\top (y - x)
\]

hence,

\[
f(y) = g(1) = g(0) + \int_0^1 g'(t)dt \geq g(0) + g'(0)
\]

\[
= f(x) + \nabla f(x)^\top (y - x)
\]
Lipschitz continuous gradient

Gradient of $f$ is Lipschitz continuous with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y \in \text{dom } f$$

- Note that the definition does not assume convexity of $f$
- We will see that for convex $f$ with $\text{dom } f = \mathbb{R}^n$, this is equivalent to

$$\frac{L}{2} x^\top x - f(x) \text{ is convex}$$

(i.e., if $f$ is twice differentiable, $\nabla^2 f(x) \preceq LI$ for all $x$)
Quadratic upper bound

suppose $\nabla f$ is Lipschitz continuous with parameter $L$ and $\text{dom } f$ is convex

- Then $g(x) = (L/2)x^\top x - f(x)$, with $\text{dom } g$, is convex
- convexity of $g$ is equivalent to a quadratic upper bound on $f$:

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

\[ f(y) \quad \quad \quad (x, f(x)) \]
Proof.

- Lipschitz continuity of $\nabla f$ and Cauchy-Schwarz inequality imply
  \[
  (\nabla f(x) - \nabla f(y))^\top (x - y) \leq L \|x - y\|^2_2 \forall x, y \in \text{dom } f
  \]
  this is monotonicity of the gradient $\nabla g(x) = Lx - \nabla f(x)$

- hence, $g$ is a convex function if its domain $\text{dom } g = \text{dom } f$

- the quadratic upper bound is the first-order condition for the convexity of $g$
  \[
  g(y) \geq g(x) + \nabla g(x)^\top (y - x) \quad \forall x, y \in \text{dom } g
  \]
Consequence of quadratic upper bound

if \( \text{dom} \, f = \mathbb{R}^n \) and \( f \) has a minimizer \( x^* \), then

\[
\frac{1}{2L} \| \nabla f(x) \|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2} \| x - x^* \|_2^2 \quad \forall x
\]

- Right-hand inequality follows from quadratic upper bound at \( x = x^* \)
- Left-hand inequality follows by minimizing quadratic upper bound

\[
f(x^*) \leq \inf_{y \in \text{dom} \, f} \left( f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \| y - x \|_2^2 \right) \]

\[
= f(x) - \frac{1}{2L} \| \nabla f(x) \|_2^2
\]

minimizer of upper bound is \( y = x - (1/L) \nabla f(x) \) because \( \text{dom} \, f = \mathbb{R}^n \)
Co-coercivity of gradient

if \( f \) is convex with \( \text{dom} f = \mathbb{R}^n \) and \( (L/2)x^\top x - f(x) \) is convex then

\[
(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2_2 \quad \forall x, y
\]

this property is known as **co-coercivity** of \( \nabla f \)(with parameter \( 1/L \))

- Co-coercivity implies Lipschitz continuity of \( \nabla f \)(by Cauchy-Schwarz)

- Hence, for differentiable convex \( f \) with \( \text{dom} f = \mathbb{R}^n \)

\[
\text{Lipschitz continuity of } \nabla f \Rightarrow \text{convexity of } (L/2)x^\top x - f(x) \\
\Rightarrow \text{co-coercivity of } \nabla f \\
\Rightarrow \text{Lipschitz continuity of } \nabla f
\]

therefore the three properties are equivalent.
proof of co-coercivity: define convex functions \( f_x, f_y \) with domain \( \mathbb{R}^n \):

\[
f_x(z) = f(z) - \nabla f(x)^\top z, \quad f_y(z) = f(z) - \nabla f(y)^\top z
\]

the functions \((L/2)z^\top z - f_x(z)\) and \((L/2)z^\top z - f_y(z)\) are convex

- \( z = x \) minimizes \( f_x(z) \); from the left-hand inequality on page 15,

\[
f(y) - f(x) - \nabla f(x)^\top (y - x) = f_x(y) - f_x(x) \\
\geq \frac{1}{2L} \| \nabla f_x(y) \|_2^2 \\
= \frac{1}{2L} \| \nabla f(y) - \nabla f(x) \|_2^2
\]

- similarly, \( z = y \) minimizes \( f_y(z) \); therefore

\[
f(x) - f(y) - \nabla f(y)^\top (x - y) \geq \frac{1}{2L} \| \nabla f(y) - \nabla f(x) \|_2^2
\]

combing the two inequalities shows co-coercivity
Strongly convex function

\( f \) is strongly convex with parameter \( m > 0 \) if

\[
g(x) = f(x) - \frac{m}{2} x^\top x \text{ is convex}
\]

**Jensen’s inequality:** Jensen’s inequality for \( g \) is

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

**monotonicity:** monotonicity of \( \nabla g \) gives

\[
(\nabla f(x) - \nabla f(y))^\top (x - y) \geq m\|x - y\|^2_2 \quad \forall x, y \in \text{dom}f
\]

this is called **strong monotonicity** (covercivity) of \( \nabla f \)

**second-order condition:** \( \nabla^2 f(x) \succeq mI \) for all \( x \in \text{dom} f \)
Quadratic lower bound

Form 1st order condition of convexity of $g$:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2} \|y - x\|_2^2 \quad \forall x, y \in \text{dom } f$$

- Implies sublevel sets of $f$ are bounded
- If $f$ is closed (has closed sublevel sets), it has a unique minimizer $x^*$ and

$$\frac{m}{2} \|x - x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|_2^2 \quad x \in \text{dom } f$$
Extension of co-coercivity

if \( f \) is strongly convex and \( \nabla f \) is Lipschitz continuous, then

\[
g(x) = f(x) - \frac{m}{2} \|x\|^2
\]

is convex and \( \nabla g \) is Lipschitz continuous with parameter \( L - m \).

co-coercivity of \( g \) gives

\[
(\nabla f(x) - \nabla f(y))^\top (x - y) \geq \frac{mL}{m + L} \|x - y\|^2 + \frac{1}{m + L} \|\nabla f(x) - \nabla f(y)\|^2
\]

for all \( x, y \in \text{dom } f \)
Outline

- gradient method, first-order methods
- quadratic bounds on convex functions
- analysis of gradient method
Analysis of gradient method

\[ x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, \ldots \]

with fixed step size or backtracking line search

assumptions

1. \( f \) is convex and differentiable with \( \text{dom} f = \mathbb{R}^n \)

2. \( \nabla f(x) \) is Lipschitz continuous with parameter \( L > 0 \)

3. Optimal value \( f^* = \inf_x f(x) \) is finite and attained at \( x^* \)
Analysis for constant step size

from quadratic upper bound with \( y = x - t\nabla f(x) \):

\[
f(x - t\nabla f(x)) \leq f(x) - t(1 - \frac{Lt}{2})\|\nabla f(x)\|_2^2
\]

therefore, if \( x^+ = x - t\nabla f(x) \) and \( 0 < t \leq 1/L \),

\[
f(x^+) \leq f(x) - \frac{t}{2}\|\nabla f(x)\|_2^2
\]

\[
\leq f^* + \nabla f(x)^\top(x - x^*) - \frac{t}{2}\|\nabla f(x)\|_2^2
\]

\[
= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x - x^* - t\nabla f(x)\|_2^2)
\]

\[
= f^* + \frac{1}{2t}(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2)
\]
take $x = x^{(i-1)}, x^+ = x^{(i)}, t_i = t$, and add the bounds for $i = 1, \cdots, k$:

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

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

$$\leq \frac{1}{2t} \|x^{(0)} - x^*\|_2^2$$

since $f(x^{(i)})$ is non-increasing,

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

conclusions: iterations to reach $f(x^{(k)}) - f^* \leq \epsilon$ is $O(1/\epsilon)$
Backtracking line search

initialize $t_k$ at $\hat{t} > 0$ (for example, $\hat{t} = 1$); take $t_k := \beta t_k$ until

$$f(x - t_k \nabla f(x)) < f(x) - \alpha t_k \|\nabla f(x)\|_2^2$$

$0 < \beta < 1$; we will take $\alpha = 1/2$ (mostly to simplify proofs)
Analysis for backtracking line search

line search with $\alpha = 1/2$ if $f$ has a Lipschitz continuous gradient

selected step size satisfies $t_k \geq t_{\min} = \min\{\hat{t}, \beta/L\}$
Convergence analysis

from page 23:

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

$$\leq f^* + \frac{1}{2t_{\text{min}}} \left( \|x^{(i-1)} - x^*\|_2^2 - \|x^{(i)} - x^*\|_2^2 \right)$$

add the upper bounds to get

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

conclusion: same $1/k$ bound as with constant step size
better results exist if we add strong convexity to the assumptions

**analysis for constant step size**

if \( x^+ = x - t \nabla f(x) \) and \( 0 < t \leq 2/(m + L) \):

\[
\|x^+ - x^*\|_2^2 = \|x - t \nabla f(x) - x^*\|_2^2 \\
= \|x - x^*\|_2^2 - 2t \nabla f(x)^\top (x - x^*) + t^2 \|\nabla f(x)\|_2^2 \\
\leq (1 - t \frac{2mL}{m + L})\|x - x^*\|_2^2 + t(t - \frac{2}{m + L})\|\nabla f(x)\|_2^2 \\
\leq (1 - t \frac{2mL}{m + L})\|x - x^*\|_2^2
\]

(step 3 follows from result on page 20)
distance to optimum

\[ \|x^{(k)} - x^*\|_2^2 \leq c^k \|x^{(0)} - x^*\|_2^2, \quad c = 1 - t \frac{2mL}{m + L} \]

- implies (linear) convergence
- for \( t = \frac{2}{m+L} \), get \( c = \frac{(\gamma-1)^2}{(\gamma+1)^2} \) with \( \gamma = \frac{L}{m} \)

bound on function value (from page 15),

\[ f(x^{(k)}) - f^* \leq \frac{L}{2} \|x^{(k)} - x^*\|_2^2 \leq \frac{c^k L}{2} \|x^{(0)} - x^*\|_2^2 \]

conclusion: iterations to reach \( f(x^{(k)}) - f^* \leq \epsilon \) is \( O(\log(1/\epsilon)) \)
Limits on convergence rate of first-order methods

**first-order method**: any iterative algorithm that selects $x^{(k)}$ in

$$x^{(0)} + \text{span}\{\nabla f(x^{(0)}), \nabla f(x^{(1)}), \ldots, \nabla f(x^{(k-1)})\}$$

**problem class**: any function that satisfies the assumptions on p. 22

**Theorem (Nesterov)**: for every integer $k \leq (n - 1)/2$ and every $x^{(0)}$, there exist functions in the problem class such that for any first-order method

$$f(x^{(k)}) - f^* \geq \frac{3}{32} \frac{L\|x^{(0)} - x^*\|_2^2}{(k + 1)^2}$$

- suggests $1/k$ rate for gradient method is not optimal
- recent fast gradient methods have $1/k^2$ convergence (see later)
Barzilar-Borwein (BB) gradient method

Consider the problem

$$\min f(x)$$

- Steepest gradient descent method: $x^{k+1} := x^k - \alpha^k g^k$:
  
  $$\alpha^k := \arg\min_\alpha f(x^k - \alpha g^k)$$

- Let $s^{k-1} := x^k - x^{k-1}$ and $y^{k-1} := g^k - g^{k-1}$.
- BB: choose $\alpha$ so that $D = \alpha I$ satisfies $Dy \approx s$:

  $$\alpha = \arg\min_\alpha \|\alpha y - s\|^2 \implies \alpha := \frac{s^\top y}{y^\top y}$$

  $$\alpha = \arg\min_\alpha \|y - s/\alpha\|^2 \implies \alpha := \frac{s^\top s}{s^\top y}$$
Globalization strategy for BB method

Algorithm 1: Raydan’s method

1. Given $x^0$, set $\alpha > 0$, $M \geq 0$, $\sigma$, $\delta$, $\epsilon \in (0, 1)$, $k = 0$.
2. while $\|g^k\| > \epsilon$ do
   3. while $f(x^k - \alpha g^k) \geq \max_{0 \leq j \leq \min(k, M)} f_{k-j} - \sigma \alpha \|g^k\|^2$ do
      4. set $\alpha = \delta \alpha$
   5. Set $x^{k+1} := x^k - \alpha g^k$.
   6. Set $\alpha := \max \left( \min \left( -\frac{\alpha (g^k)^\top y^k}{(g^k)^\top g^k}, \alpha_M \right), \alpha_m \right)$, $k := k + 1$. 
Globalization strategy for BB method

**Algorithm 2: Hongchao and Hagger’s method**

1. Given $x^0$, set $\alpha > 0$, $\sigma, \delta, \eta, \epsilon \in (0, 1)$, $k = 0$.
2. while $\|g^k\| > \epsilon$ do
   3. while $f(x^k - \alpha g^k) \geq C^k - \sigma \alpha \|g^k\|^2$ do
      4. set $\alpha = \delta \alpha$
   5. Set $x^{k+1} := x^k - \alpha g^k$, $Q^{k+1} = \eta Q^k + 1$ and
      $C^{k+1} = (\eta Q^k C^k + f(x^{k+1}))/Q^{k+1}$.
   6. Set $\alpha := \max \left( \min \left( -\frac{\alpha (g^k)^\top g^k}{(g^k)^\top y^k}, \alpha_M \right), \alpha_m \right)$, $k := k + 1$. 
Spectral projected method on convex sets

Consider the problem

\[ \min f(x) \quad \text{s.t.} \quad x \in \Omega \]

**Algorithm 3:** Birgin, Martinez and Raydan’s method

1. Given \( x^0 \in \Omega \), set \( \alpha > 0 \), \( M \geq 0 \), \( \sigma, \delta, \epsilon \in (0, 1) \), \( k = 0 \).
2. while \( \| P(x^k - g^k) - x^k \| \geq \epsilon \) do
3. \hspace{1em} Set \( x^{k+1} := P(x^k - \alpha g^k) \).
4. \hspace{1em} while \( f(x^{k+1}) \geq \max_{0 \leq j \leq \min(k, M)} f_{k-j} + \sigma (x^{k+1} - x^k)^\top g^k \) do
5. \hspace{2em} set \( \alpha = \delta \alpha \) and \( x^{k+1} := P(x^k - \alpha g^k) \).
6. \hspace{1em} if \( (s^k)^\top y^k \leq 0 \) then set \( \alpha = \alpha_M \); \( \alpha_m \).
7. \hspace{1em} else set \( \alpha := \max \left( \min \left( \frac{(s^k)^\top s^k}{(s^k)^\top y^k}, \alpha_M \right), \alpha_m \right) \); \( \alpha_m \).
8. \hspace{1em} Set \( k := k + 1 \).
Spectral projected method on convex sets

Consider the problem

$$\min f(x) \quad \text{s.t. } x \in \Omega$$

Algorithm 4: Birgin, Martinez and Raydan’s method

1. Given $x^0 \in \Omega$, set $\alpha > 0$, $M \geq 0$, $\sigma$, $\delta$, $\epsilon \in (0, 1)$, $k = 0$.
2. \textbf{while} $\|P(x^k - g^k) - x^k\| \geq \epsilon$ \textbf{do}
   3. Compute $d^k := P(x^k - \alpha g^k) - x^k$.
   4. Set $\alpha = 1$ and $x^{k+1} = x^k + d^k$.
   5. \textbf{while} $f(x^{k+1}) \geq \max_{0 \leq j \leq \min(k, M)} f_k - j + \sigma (d^k) \top g^k$ \textbf{do}
      6. \textbf{set} $\alpha = \delta \alpha$ and $x^{k+1} := x^k + \alpha d^k$.
   7. \textbf{if} $(s^k) \top y^k \leq 0$ \textbf{then} \textbf{set} $\alpha = \alpha_M$;
   8. \textbf{else} \textbf{set} $\alpha := \max \left( \min \left( \frac{(s^k) \top s^k}{(s^k) \top y^k}, \alpha_M \right), \alpha_m \right)$.
   9. \textbf{Set} $k := k + 1$.

Question: is $x^k$ feasible?
References


B. T. Polyak, Introduction to Optimization (1987), section 1.4