1. Gradient method

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method
Gradient method

to minimize a convex differentiable function $f$: choose an initial point $x_0$ and repeat

$$x_{k+1} = x_k - t_k \nabla f(x_k), \quad k = 0, 1, \ldots$$

step size $t_k$ is constant or determined by line search

Advantages

- every iteration is inexpensive
- does not require second derivatives

Notation

- $x_k$ can refer to $k$th element of a sequence, or to the $k$th component of vector $x$
- to avoid confusion, we sometimes use $x^{(k)}$ to denote elements of a sequence
Quadratic example

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

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

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

where \( x^* = 0 \)

gradient method is often slow; convergence very dependent on scaling
Nondifferentiable example

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

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

gradient method does not handle nondifferentiable problems
First-order methods

address one or both shortcomings of the gradient method

Methods for nondifferentiable or constrained problems

- subgradient method
- proximal gradient method
- smoothing methods
- cutting-plane methods

Methods with improved convergence

- conjugate gradient method
- accelerated gradient method
- quasi-Newton methods
Outline

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method
Convex function

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

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) \quad \text{for all } x, y \in \text{dom } f, \theta \in [0, 1]$$

**First-order condition**

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

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \text{for all } x, y \in \text{dom } f$$

As in ECE236B, we use $b^T a$ for inner product of $a$ and $b$

**Second-order condition**

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

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

$f$ is strictly convex if $\text{dom } f$ is a convex set and

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta) f(y) \quad \text{for all } x, y \in \text{dom } f, x \neq y, \text{ and } \theta \in (0, 1)$$

Strict convexity implies that if a minimizer of $f$ exists, it is unique.

**First-order condition**

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

$$f(y) > f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y \in \text{dom } f, x \neq y$$

**Second-order condition**

Note that $\nabla^2 f(x) > 0$ is not necessary for strict convexity (cf., $f(x) = x^4$)
Monotonicity of gradient

A differentiable function $f$ is convex if and only if $\text{dom } f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq 0 \quad \text{for all } x, y \in \text{dom } f$$

i.e., the gradient $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a monotone mapping

A differentiable function $f$ is strictly convex if and only if $\text{dom } f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) > 0 \quad \text{for all } x, y \in \text{dom } f, x \neq y$$

i.e., the gradient $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a strictly monotone mapping
Proof

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

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

combining the inequalities gives \((\nabla f(x) - \nabla f(y))^T (x - y) \geq 0\)

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

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

hence

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

\[
= f(x) + \nabla f(x)^T (y - x)
\]

this is the first-order condition for convexity
Outline

• gradient method, first-order methods

• convex functions

• Lipschitz continuity of gradient

• strong convexity

• analysis of gradient method
Lipschitz continuous gradient

The gradient of $f$ is *Lipschitz continuous* with parameter $L > 0$ if

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| \quad \text{for all } x, y \in \text{dom } f$$

- Functions $f$ with this property are also called *$L$-smooth*.
- The definition does not assume convexity of $f$ (and holds for $-f$ if it holds for $f$).
- In the definition, $\| \cdot \|$ and $\| \cdot \|_*$ are a pair of dual norms:

$$\|u\|_* = \sup_{v \neq 0} \frac{u^Tv}{\|v\|} = \sup_{\|v\|=1} u^Tv$$

This implies a generalized Cauchy–Schwarz inequality:

$$|u^Tv| \leq \|u\|_*\|v\| \quad \text{for all } u, v$$
Choice of norm

Equivalence of norms

• for any two norms $\| \cdot \|_a$, $\| \cdot \|_b$, there exist positive constants $c_1$, $c_2$ such that

$$c_1 \| x \|_b \leq \| x \|_a \leq c_2 \| x \|_b \quad \text{for all } x$$

• constants depend on dimension; for example, for $x \in \mathbb{R}^n$,

$$\| x \|_2 \leq \| x \|_1 \leq \sqrt{n} \| x \|_2, \quad \frac{1}{\sqrt{n}} \| x \|_2 \leq \| x \|_\infty \leq \| x \|_2$$

Norm in definition of Lipschitz continuity

• without loss of generality we can use the Euclidean norm $\| \cdot \| = \| \cdot \|_* = \| \cdot \|_2$

• the parameter $L$ depends on choice of norm

• in complexity bounds, choice of norm can simplify dependence on dimensions
Quadratic upper bound

suppose $\nabla f$ is Lipschitz continuous with parameter $L$

- this implies (from the generalized Cauchy–Schwarz inequality) that
  \[
  (\nabla f(x) - \nabla f(y))^T (x - y) \leq L \|x - y\|^2 \quad \text{for all } x, y \in \text{dom } f \quad (1)
  \]

- if $\text{dom } f$ is convex, (1) is equivalent to
  \[
  f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|y - x\|^2 \quad \text{for all } x, y \in \text{dom } f \quad (2)
  \]
Proof (equivalence of (1) and (2) if dom $f$ is convex)

- consider arbitrary $x, y \in \text{dom } f$ and define $g(t) = f(x + t(y - x))$
- $g(t)$ is defined for $t \in [0, 1]$ because dom $f$ is convex
- if (1) holds, then

$$g'(t) - g'(0) = (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x) \leq tL\|x - y\|^2$$

integrating from $t = 0$ to $t = 1$ gives (2):

$$f(y) = g(1) = g(0) + \int_0^1 g'(t) \, dt \leq g(0) + g'(0) + \frac{L}{2}\|x - y\|^2$$

$$= f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|x - y\|^2$$

- conversely, if (2) holds, then (2) and the same inequality with $x, y$ switched, i.e.,

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{L}{2}\|x - y\|^2,$$

can be combined to give

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|^2$$
Consequence of quadratic upper bound

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

\[
\frac{1}{2L} \|
abla f(z) \|^2_2 \leq f(z) - f(x^*) \leq \frac{L}{2} \| z - x^* \|^2 \quad \text{for all } z
\]

- right-hand inequality follows from upper bound property (2) at \( x = x^*, y = z \)
- left-hand inequality follows by minimizing quadratic upper bound for \( x = z \)

\[
\inf_y f(y) \leq \inf_y \left( f(z) + \nabla f(z)^T (y - z) + \frac{L}{2} \| y - z \|^2 \right)
\]

\[
= \inf_{\|v\|=1} \inf_t \left( f(z) + t \nabla f(z)^T v + \frac{Lt^2}{2} \right)
\]

\[
= \inf_{\|v\|=1} \left( f(z) - \frac{1}{2L} (\nabla f(z)^T v)^2 \right)
\]

\[
= f(z) - \frac{1}{2L} \| \nabla f(z) \|^2_2
\]
Co-coercivity of gradient

if \( f \) is convex with \( \text{dom } f = \mathbb{R}^n \) and \( \nabla f \) is \( L \)-Lipschitz continuous, then

\[
(\nabla f(x) - \nabla f(y))^T(x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2 \quad \text{for all } x, y
\]

- this property is known as co-coercivity of \( \nabla f \) (with parameter \( 1/L \))
- co-coercivity in turn implies Lipschitz continuity of \( \nabla f \) (by Cauchy–Schwarz)
- hence, for differentiable convex \( f \) with \( \text{dom } f = \mathbb{R}^n \)

\[
\text{Lipschitz continuity of } \nabla f \quad \Rightarrow \quad \text{upper bound property (2) (equivalently, (1))}
\]

\[
\Rightarrow \quad \text{co-coercivity of } \nabla f
\]

\[
\Rightarrow \quad \text{Lipschitz continuity of } \nabla f
\]

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

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

- the two functions have \( L \)-Lipschitz continuous gradients
- \( z = x \) minimizes \( f_x(z) \); from the left-hand inequality on page 1.14,

\[
f(y) - f(x) - \nabla f(x)^T (y - x) = f_x(y) - f_x(x)
\]

\[
\geq \frac{1}{2L} \| \nabla f_x(y) \|^2_*
\]

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

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

\[
f(x) - f(y) - \nabla f(y)^T (x - y) \geq \frac{1}{2L} \| \nabla f(y) - \nabla f(x) \|^2_*
\]

combining the two inequalities shows co-coercivity
Lipschitz continuity with respect to Euclidean norm

suppose $f$ is convex with $\text{dom } f = \mathbb{R}^n$, and $L$-smooth for the Euclidean norm:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for all } x, y$$

- the equivalent property (1) states that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \leq L(x - y)^T (x - y) \quad \text{for all } x, y$$

- this is monotonicity of $Lx - \nabla f(x)$, i.e., equivalent to the property that

$$\frac{L}{2}\|x\|_2^2 - f(x)$$

is a convex function

- if $f$ is twice differentiable, the Hessian of this function is $LI - \nabla^2 f(x)$:

$$\lambda_{\text{max}}(\nabla^2 f(x)) \leq L \quad \text{for all } x$$

is an equivalent characterization of $L$-smoothness
Outline

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method
Strongly convex function

$f$ is strongly convex with parameter $m > 0$ if $\text{dom } f$ is convex and

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

holds for all $x, y \in \text{dom } f$, $\theta \in [0, 1]$

- this is a stronger version of Jensen’s inequality
- it holds if and only if it holds for $f$ restricted to arbitrary lines:

\[ f(x + t(y - x)) - \frac{m}{2}t^2\|x - y\|^2 \]

is a convex function of $t$, for all $x, y \in \text{dom } f$

- without loss of generality, we can take $\| \cdot \| = \| \cdot \|_2$
- however, the strong convexity parameter $m$ depends on the norm used
Quadratic lower bound

if $f$ is differentiable and $m$-strongly convex, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} \|y - x\|^2 \quad \text{for all } x, y \in \text{dom } f$$

(4)

- follows from the 1st order condition of convexity of (3)
- this implies that the sublevel sets of $f$ are bounded
- if $f$ is closed (has closed sublevel sets), it has a unique minimizer $x^*$ and

$$\frac{m}{2} \|z - x^*\|^2 \leq f(z) - f(x^*) \leq \frac{1}{2m} \|
abla f(z)\|_*^2 \quad \text{for all } z \in \text{dom } f$$

(proof as on page 1.14)
Strong monotonicity

differentiable $f$ is strongly convex if and only if $\text{dom } f$ is convex and

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq m \|x - y\|^2 \quad \text{for all } x, y \in \text{dom } f$$

this is called strong monotonicity (coercivity) of $\nabla f$

Proof

- one direction follows from (4) and the same inequality with $x$ and $y$ switched
- for the other direction, assume $\nabla f$ is strongly monotone and define

$$g(t) = f(x + t(y - x)) - \frac{m}{2} t^2 \|x - y\|^2$$

then $g'(t)$ is nondecreasing, so $g$ is convex
Strong convexity with respect to Euclidean norm

suppose $f$ is $m$-strongly convex for the Euclidean norm:

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

for $x, y \in \text{dom } f$, $\theta \in [0, 1]$

- this is Jensen’s inequality for the function

$$h(x) = f(x) - \frac{m}{2}\|x\|^2_2$$

- therefore $f$ is strongly convex if and only if $h$ is convex

- if $f$ is twice differentiable, $h$ is convex if and only if $\nabla^2f(x) - mI \succeq 0$, or

$$\lambda_{\min}(\nabla^2f(x)) \geq m \quad \text{for all } x \in \text{dom } f$$
Extension of co-coercivity

suppose \( f \) is \( m \)-strongly convex and \( L \)-smooth for \( \| \cdot \|_2 \), and \( \text{dom } f = \mathbb{R}^n \)

- then the function
  \[
  h(x) = f(x) - \frac{m}{2} \|x\|_2^2
  \]
  is convex and \( (L - m) \)-smooth:

  \[
  0 \leq (\nabla h(x) - \nabla h(y))^T (x - y) \\
  = (\nabla f(x) - \nabla f(y))^T (x - y) - m\|x - y\|_2^2 \\
  \leq (L - m)\|x - y\|_2^2
  \]

- co-coercivity of \( \nabla h \) can be written as

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

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

- gradient method, first-order methods
- convex functions
- Lipschitz continuity of gradient
- strong convexity
- analysis of gradient method
Analysis of gradient method

\[ x_{k+1} = x_k - t_k \nabla f(x_k), \quad k = 0, 1, \ldots \]

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 \( L \)-Lipschitz continuous with respect to the Euclidean norm, with \( L > 0 \)

3. optimal value \( f^\star = \inf_x f(x) \) is finite and attained at \( x^\star \)
Basic gradient step

• from quadratic upper bound (page 1.12) with \( y = x - t \nabla f(x) \):

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

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

\[
    f(x^+) \leq f(x) - \frac{t}{2} \| \nabla f(x) \|_2^2 \tag{5}
\]

• from (5) and convexity of \( f \),

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

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

\[
    = \frac{1}{2t} \left( \| x - x^* \|_2^2 - \| x^+ - x^* \|_2^2 \right) \tag{6}
\]
Descent properties

assume $\nabla f(x) \neq 0$ and $0 < t \leq 1/L$

- the inequality (5) shows that
  \[ f(x^+) < f(x) \]

- the inequality (6) shows that
  \[ ||x^+ - x^*||_2 < ||x - x^*||_2 \]

in the gradient method, function value \textit{and} distance to the optimal set decrease
Gradient method with constant step size

\[ x_{k+1} = x_k - t \nabla f(x_k), \quad k = 0, 1, \ldots \]

- take \( x = x_{i-1}, x^+ = x_i \) in (6) and add the bounds for \( i = 1, \ldots, k \):

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

\[
= \frac{1}{2t} \left( \|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2 \right)
\]

\[
\leq \frac{1}{2t} \|x_0 - x^*\|_2^2
\]

- since \( f(x_i) \) is non-increasing (see (5))

\[
f(x_k) - f^* \leq \frac{1}{k} \sum_{i=1}^{k} (f(x_i) - f^*) \leq \frac{1}{2kt} \|x_0 - x^*\|_2^2
\]

**Conclusion:** number of 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, $\hat{t} = 1$) and take $t_k := \beta t_k$ until

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

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 $t_k \geq t_{\text{min}} = \min\{\hat{t}, \beta/L\}$
Gradient method with backtracking line search

- from line search condition and convexity of $f$,

$$f(x_{i+1}) \leq f(x_i) - \frac{t_i}{2} \|\nabla f(x_i)\|_2^2$$

$$\leq f^* + \nabla f(x_i)^T (x_i - x^*) - \frac{t_i}{2} \|\nabla f(x_i)\|_2^2$$

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

- this implies $\|x_{i+1} - x^*\|_2 \leq \|x_i - x^*\|$, so we can replace $t_i$ with $t_{\text{min}} \leq t_i$:

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

- adding the upper bounds gives same $1/k$ bound as with constant step size

$$f(x_k) - f^* \leq \frac{1}{k} \sum_{i=1}^{k} (f(x_i) - f^*) \leq \frac{1}{2kt_{\text{min}}} \|x_0 - x^*\|_2^2$$
Gradient method for strongly convex functions

better results exist if we add strong convexity to the assumptions on p. 1.23

Analysis for constant step size

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

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

(step 3 follows from result on page 1.22)
Distance to optimum

\[ \|x_k - x^*\|^2 \leq c^k \|x_0 - x^*\|^2, \quad c = 1 - t \frac{2mL}{m + L} \]

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

Bound on function value (from page 1.14)

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

Conclusion: number of 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 $x_{k+1}$ in the set

$$x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \ldots, \nabla f(x_k)\}$$

**Problem class:** any function that satisfies the assumptions on page 1.23

**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

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k + 1)^2}$$

- suggests $1/k$ rate for gradient method is not optimal
- more recent accelerated gradient methods have $1/k^2$ convergence (see later)
References


