Unconstrained optimization methods

G. Mastroeni

Department of Computer Science, University of Pisa

Optimization Methods and Game Theory
Master of Science in Artificial Intelligence and Data Engineering
University of Pisa – A.Y. 2024/25

Contents of the lessons

- Gradient method
- Conjugate gradient method
- Newton methods

Gradient method

Consider an unconstrained problem: $\min_{x \in \mathbb{R}^n} f(x)$.

Current point x^k, search direction $d^k = -\nabla f(x^k)$ (steepest descent direction)

Gradient method

1. Choose $x^0 \in \mathbb{R}^n$, set $k = 0$. Go to Step 2.
2. If $\nabla f(x^k) = 0$, STOP. Otherwise go to Step 3.
3. Let $d^k = -\nabla f(x^k)$ [search direction]
   compute an optimal solution t_k of the problem: $\min_{t>0} f(x^k + t\, d^k)$ [step size];
   Set $x^{k+1} = x^k + t_k\, d^k$, $k = k + 1$;
   Go to Step 2.

Example 3.1 $f(x) = x_1^2 + x_2^2 - x_1 x_2$, starting point $x^0 = (1, 1)$.

$$\nabla f(x^0) = (1, 1),\;\; d^0 = (-1, -1),\;\; x^0 + td^0 = (1 - t, 1 - t)$$

,

$$f(x^0 + td^0) = (1 - t)^2 \quad t_0 = 1, \quad x^1 = (0, 0)$$

Gradient method - convergence

Proposition

Let f be continuously differentiable.

- $(d^k)^\mathsf{T} d^{k+1} = 0$ for any iteration k.
- If $\{x^k\}$ converges to x^*, then $\nabla f(x^*) = 0$, i.e. x^* is a stationary point of f.

Theorem

If f is coercive, then for any starting point x^0 the generated sequence $\{x^k\}$ is bounded and any of its cluster points is a stationary point of f.

Corollary

If f is coercive and convex, then for any starting point x^0 the generated sequence $\{x^k\}$ is bounded and any of its cluster points is a global minimum of f.

Corollary

If f is strongly convex, then for any starting point x^0 the generated sequence $\{x^k\}$ converges to the unique global minimum of f.

Gradient method - quadratic case

If $f(x) = \frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x$, with Q positive definite matrix, then, by the Taylor expansion at x^k, we have

$$f(x^k + td^k) = f(x^k) + (td^k)^\mathsf{T} \nabla f(x^k) + \frac{1}{2} (td^k)^\mathsf{T} Q \, td^k =$$

$$= \frac{1}{2} (d^k)^\mathsf{T} Q \, d^k \, t^2 + (d^k)^\mathsf{T} g^k \, t + f(x^k),$$

where $g^k = \nabla f(x^k) = Q \, x^k + c$. Thus the step size is equal to

$$t_k = -\frac{(d^k)^\mathsf{T} g^k}{(d^k)^\mathsf{T} Q \, d^k}.$$

Gradient method - convergence rate

As already observed, two subsequent directions are orthogonal: $(d^k)^\mathsf{T} d^{k+1} = 0$. This implies that the generated sequence has a zig-zag behaviour.

Theorem (Error bound)

If $f(x) = \frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x$, with Q positive definite matrix, and x^* is the global minimum of f, then the sequence $\{x^k\}$ satisfies the following inequality:

$$\|x^{k+1} - x^*\|_Q \leq \left(\frac{\dfrac{\lambda_n}{\lambda_1} - 1}{\dfrac{\lambda_n}{\lambda_1} + 1} \right) \|x^k - x^*\|_Q, \quad \forall\ k \geq 0, \quad \text{(linear convergence)}$$

where $\|x\|_Q = \sqrt{x^\mathsf{T} Q x}$ and $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of Q.

Remark

If λ_n/λ_1 (condition number of Q) is $>> 1$, then the ratio $\left(\frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} \right) \simeq 1$ and the convergence may be slow.

Gradient method - convergence rate

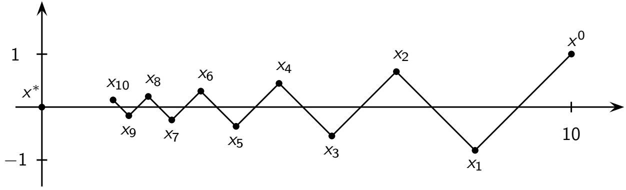
Example 3.2 $f(x) = x_1^2 + 10\,x_2^2$, global minimum is $x^* = (0,0)$.

If the starting point is $x^0 = (10,1)$, then the generated sequence is:

$$x^k = \left(10\left(\frac{9}{11}\right)^k, \left(-\frac{9}{11}\right)^k\right), \qquad \forall\ k \geq 0,$$

hence

$$\|x^{k+1} - x^*\|_Q = \frac{9}{11}\,\|x^k - x^*\|_Q \qquad \forall\ k \geq 0.$$



Gradient method - exercise

Exercise 3.1 Implement in MATLAB the gradient method for solving the problem

$$\begin{cases} \min \ \frac{1}{2}x^\mathsf{T} Q x + c^\mathsf{T} x \\ x \in \mathbb{R}^n \end{cases}$$

where Q is a positive definite matrix. In particular, solve the problem

$$\begin{cases} \min \ 3x_1^2 + 3x_2^2 + 3x_3^2 + 3x_4^2 - 4x_1x_3 - 4x_2x_4 + x_1 - x_2 + 2x_3 - 3x_4 \\ x \in \mathbb{R}^4 \end{cases}$$

starting from the point $(0,0,0,0)$. [Use $\|\nabla f(x)\| < 10^{-6}$ as stopping criterion.]

When f is not a quadratic function, the exact line search may be computationally expensive.

Gradient method with the Armijo inexact line search

1. Set $\alpha, \gamma \in (0, 1)$ and $\bar{t} > 0$. Choose $x^0 \in \mathbb{R}^n$, set $k = 0$. Go to Step 2.
2. If $\nabla f(x^k) = 0$, STOP. Otherwise go to Step 3.
3. Let $d^k = -\nabla f(x^k)$, $t_k = \bar{t}$;
   while $f(x^k + t_k\, d^k) > f(x^k) + \alpha\, t_k\, (d^k)^{\mathsf{T}} \nabla f(x^k)$ **do**
   $t_k = \gamma t_k$
   end
   Set $x^{k+1} = x^k + t_k d^k$, $k = k + 1$
   Go to Step 2.

Theorem

If f is coercive, then for any starting point x^0 the generated sequence $\{x^k\}$ is bounded and any of its cluster points is a stationary point of f.

Example 3.3 Let $f(x_1, x_2) = x_1^4 + x_1^2 + x_2^2$. Set $\alpha = 10^{-4}$, $\gamma = 0.5$, $\bar{t} = 1$, choose $x^0 = (1, 1)$.
$d^0 = -\nabla f(x^0) = (-6, -2)$.

Line search. If $t_0 = 1$ then $x^0 + t_0 d^0 = (-5, -1)$ and

$$f(x^0 + t_0 d^0) = 651 > f(x^0) + \alpha\, t_0 (d^0)^\mathsf{T} \nabla f(x^0) = 2.996,$$

if $t_0 = 0.5$ then

$$f(x^0 + t_0 d^0) = 20 > f(x^0) + \alpha\, t_0 (d^0)^\mathsf{T} \nabla f(x^0) = 2.998,$$

if $t_0 = 0.25$ then

$$f(x^0 + t_0 d^0) = 0.5625 < f(x^0) + \alpha\, t_0 (d^0)^\mathsf{T} \nabla f(x^0) = 2.999$$

hence the step size is $t_0 = 0.25$ and the new iterate is

$$x^1 = x^0 + t_0\, d^0 = (1, 1) + \frac{1}{4}(-6, -2) = \left(-\frac{1}{2}, \frac{1}{2}\right).$$

Gradient method - Armijo inexact line search

Exercise 3.2. Solve the problem

$$\begin{cases} \min \ 2\,x_1^4 + 3\,x_2^4 + 2\,x_1^2 + 4\,x_2^2 + x_1\,x_2 - 3\,x_1 - 2\,x_2 \\ x \in \mathbb{R}^2 \end{cases}$$

by means of the gradient method with the Armijo inexact line search setting $\alpha = 0.1$, $\gamma = 0.9$, $\bar{t} = 1$ and starting from the point $(0,0)$.
[Use $\|\nabla f(x)\| < 10^{-3}$ as stopping criterion.]

Exercise 3.3. Solve the problem

$$\begin{cases} \min \ x_1^4 + x_2^4 - 2\,x_1^2 + 4x_1\,x_2 - 2\,x_2^2 \\ x \in \mathbb{R}^2 \end{cases}$$

by means of the gradient method with the Armijo inexact line search setting $\alpha = 0.1$, $\gamma = 0.9$, $\bar{t} = 1$ and starting from the point $(10, -10)$.
[Use $\|\nabla f(x)\| < 10^{-3}$ as stopping criterion.]

Conjugate gradient method

The conjugate gradient method is a descent method where the search direction involves the gradient computed at the current iteration and the direction computed at the previous iteration.

We first consider the quadratic case:

$$f(x) = \frac{1}{2} x^{\mathsf{T}} Q\, x + c^{\mathsf{T}} x,$$

where Q is positive definite. Set $g^k = \nabla f(x^k) = Qx^k + c$.

At iteration k, the search direction is defined by

$$d^k = \begin{cases} -g^0 & \text{if } k = 0, \\ -g^k + \beta_k\, d^{k-1} & \text{if } k \geq 1, \end{cases}$$

where β_k is such that d^k and d^{k-1} are conjugate with respect to Q, i.e.,

$$(d^k)^{\mathsf{T}} Q\, d^{k-1} = 0.$$

- By the previous relation we can compute β_k:

$$\beta_k = \frac{(g^k)^\mathsf{T} Q\, d^{k-1}}{(d^{k-1})^\mathsf{T} Q\, d^{k-1}}$$

- If we perform an exact line search, then d^k is a descent direction
- The step size given by exact line search is $t_k = -\dfrac{(g^k)^\mathsf{T} d^k}{(d^k)^\mathsf{T} Q\, d^k}$

Conjugate gradient method for quadratic functions

1. Choose $x^0 \in \mathbb{R}^n$, set $g^0 = Q\, x^0 + c$, $k := 0$; go to Step 2.
2. Let $g^k = \nabla f(x^k)$. **If** $g^k = 0$ **then** STOP, **else** go to Step 3.
3. **If** $k = 0$ **then** $d^k = -g^k$

   else $\beta_k = \dfrac{(g^k)^\mathsf{T} Q\, d^{k-1}}{(d^{k-1})^\mathsf{T} Q\, d^{k-1}}$, $\quad d^k = -g^k + \beta_k\, d^{k-1}$

   $$t_k = -\frac{(g^k)^\mathsf{T} d^k}{(d^k)^\mathsf{T} Q\, d^k}$$

   $$x^{k+1} = x^k + t_k\, d^k,\; g^{k+1} = Q\, x^{k+1} + c,\; k = k+1$$

   Go to Step 2.

Conjugate gradient method

Example 3.4 Consider $f(x) = x_1^2 + 10\,x_2^2$, with starting point $x^0 = (10, 1)$.

$$Q = \begin{pmatrix} 2 & 0 \\ 0 & 20 \end{pmatrix} \qquad \nabla f(x) = (2x_1, 20x_2)$$

- $k = 0$: $g^0 = (20, 20)$, $d^0 = -g^0 = (-20, -20)$,
  $t_0 = -((g^0)^\mathsf{T} d^0)/((d^0)^\mathsf{T} Q\, d^0) = 1/11$, and consequently

  $$x^1 = x^0 + t_0\, d^0 = (10 - 20/11, 1 - 20/11) = (90/11, -9/11)$$

- $k = 1$: $g^1 = (180/11, -180/11)$, $\beta_1 = ((g^1)^\mathsf{T} Q\, d^0)/((d^0)^\mathsf{T} Q\, d^0) = 81/121$,
  $d^1 = -g^1 + \beta_1\, d^0 = (-3600/121, 360/121)$,
  $t_1 = -((g^1)^\mathsf{T} d^1)/((d^1)^\mathsf{T} Q\, d^1) = 11/40$,
  and $x^2 = x^1 + t_1\, d^1 = (0, 0)$ which is the global minimum of f.

Conjugate gradient method - convergence

Proposition

- An alternative formula for the step size is $t_k = \dfrac{\|g^k\|^2}{(d^k)^\mathsf{T} Q\, d^k}$
- An alternative formula for β_k is $\beta_k = \dfrac{\|g^k\|^2}{\|g^{k-1}\|^2}$
- If we did not find the global minimum after k iterations, then the gradients $\{g^0, g^1, \dots, g^k\}$ are orthogonal
- If we did not find the global minimum after k iterations, then the directions $\{d^0, d^1, \dots, d^k\}$ are conjugate w.r.t. Q and x^k is the minimum of f on $x^0 + \text{Span}(d^0, d^1, \dots, d^k)$

Theorem (Convergence)

- The CG method finds the global minimum in at most n iterations.
- If Q has r distinct eigenvalues, then CG method finds the global minimum in at most r iterations.

Conjugate gradient method - convergence rate

Theorem (Error bound)

If $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of Q, then the following bounds hold:

- $$\|x^k - x^*\|_Q \leq 2 \left(\frac{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1} \right)^k \|x^0 - x^*\|_Q, \qquad \forall\ k \geq 0,$$
- $$\|x^k - x^*\|_Q \leq \left(\frac{\lambda_{n-k+1} - \lambda_1}{\lambda_{n-k+1} + \lambda_1} \right) \|x^0 - x^*\|_Q, \qquad \forall\ k \geq 0.$$

Conjugate gradient method

Exercise 3.4 Implement in MATLAB the conjugate gradient method for solving the problem

$$\begin{cases} \min \ \frac{1}{2} x^{\mathsf{T}} Q x + c^{\mathsf{T}} x \\ x \in \mathbb{R}^n \end{cases}$$

where Q is a positive definite matrix.
Solve the problem

$$\begin{cases} \min \ 3x_1^2 + 3x_2^2 + 3x_3^2 + 3x_4^2 - 4x_1x_3 - 4x_2x_4 + x_1 - x_2 + 2x_3 - 3x_4 \\ x \in \mathbb{R}^4 \end{cases}$$

starting from the point $(0, 0, 0, 0)$. [Use $\|\nabla f(x)\| < 10^{-6}$ as stopping criterion.]

Newton method – basic version

We want to find a stationary point $\nabla f(x) = 0$.

At iteration k, make a linear approximation of $\nabla f(x)$ at x^k, i.e.

$$\nabla f(x) \simeq \nabla f(x^k) + \nabla^2 f(x^k)(x - x^k),$$

the new iterate x^{k+1} is the solution of the linear system

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0.$$

Note that x^{k+1} is a stationary point of the quadratic approximation of f at x^k:

$$f(x) \simeq f(x^k) + (x - x^k)^\mathsf{T} \nabla f(x^k) + \frac{1}{2}(x - x^k)^\mathsf{T} \nabla^2 f(x^k)(x - x^k).$$

Newton method – basic version

Newton method (basic version)

1. Let $x^0 \in \mathbb{R}^n$, set $k = 0$. Go to Step 2.
2. **If** $\nabla f(x^k) = 0$ then STOP **else** go to Step 3.
3. Let d^k be the solution of the linear system $\nabla^2 f(x^k) d = -\nabla f(x^k)$.
   Set $x^{k+1} = x^k + d^k$, $k = k + 1$ and go to Step 2.

Theorem (Convergence)

If x^* is a local minimum of f and $\nabla^2 f(x^*)$ is positive definite, then there exists $\delta > 0$ such that for any $x^0 \in B(x^*, \delta)$ the sequence $\{x^k\}$ converges to x^* and

$$\|x^{k+1} - x^*\| \leq C \, \|x^k - x^*\|^2 \qquad \forall \; k > \bar{k}, \qquad \text{(quadratic convergence)}$$

for some $C > 0$ and $\bar{k} > 0$.

Newton method – basic version

Example 3.5 $f(x) = 2\,x_1^4 + 3\,x_2^4 + 2\,x_1^2 + 4\,x_2^2 + x_1\,x_2 - 3\,x_1 - 2\,x_2$ is strongly convex because

$$\nabla^2 f(x) = \begin{pmatrix} 24\,x_1^2 + 4 & 1 \\ 1 & 36\,x_2^2 + 8 \end{pmatrix}.$$

k	x^k		$\|\nabla f(x^k)\|$
0	10.000000	5.000000	8189.6317378
1	6.655450	3.298838	2429.6437291
2	4.421132	2.149158	721.6330686
3	2.925965	1.361690	214.6381594
4	1.923841	0.811659	63.7752575
5	1.255001	0.428109	18.6170045
6	0.823359	0.209601	5.0058040
7	0.580141	0.171251	1.0538969
8	0.492175	0.179815	0.1022945
9	0.481639	0.180914	0.0013018
10	0.481502	0.180928	0.0000002

Newton method – basic version

Drawbacks of Newton method:

- at each iteration we need to compute both the gradient $\nabla f(x^k)$ and the hessian matrix $\nabla^2 f(x^k)$
- local convergence: if x^0 is too far from the optimum x^*, then the generated sequence may be not convergent to x^*

Example 3.6 Let $f(x) = -\frac{1}{16}x^4 + \frac{5}{8}x^2$.
Then $f'(x) = -\frac{1}{4}x^3 + \frac{5}{4}x$ and $f''(x) = -\frac{3}{4}x^2 + \frac{5}{4}$.
$x^* = 0$ is a local minimum of f with $f''(x^*) = 5/4 > 0$.
The sequence does not converge to x^* if it starts from $x^0 = 1$:
$x^1 = -1$, $x^2 = 1$, $x^3 = -1$, ...

Newton method with line search

If f is strongly convex, then we have global convergence because d^k is a descent direction, in fact:

$$\nabla f(x^k)^\mathsf{T} d^k = -\nabla f(x^k)^\mathsf{T} [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) < 0.$$

Newton method with (inexact) line search

1. Let $\alpha, \gamma \in (0,1)$, $\bar{t} > 0$, $x^0 \in \mathbb{R}^n$, set $k = 0$. Go to Step 2.
2. **If** $\nabla f(x^k) = 0$ then STOP **else** go to Step 3.
3. Let d^k be the solution of the linear system $\nabla^2 f(x^k) d = -\nabla f(x^k)$.
   Set $t_k = \bar{t}$
   while $f(x^k + t_k\, d^k) > f(x^k) + \alpha\, t_k\, (d^k)^\mathsf{T} \nabla f(x^k)$ **do**
   $t_k = \gamma t_k$
   end
   Set $x^{k+1} = x^k + t_k\, d^k$, $k = k + 1$
   Go to Step 2.

Newton method with line search

Theorem (Convergence)

If f is strongly convex, then for any starting point $x^0 \in \mathbb{R}^n$ the sequence $\{x^k\}$ converges to the global minimum of f. Moreover, if $\alpha \in (0, 1/2)$ and $\bar{t} = 1$ then the convergence is quadratic.

Exercise 3.5. Solve the problem

$$\begin{cases} \min \; 2\,x_1^4 + 3\,x_2^4 + 2\,x_1^2 + 4\,x_2^2 + x_1\,x_2 - 3\,x_1 - 2\,x_2 \\ x \in \mathbb{R}^2 \end{cases}$$

by means of the Newton method with inexact line search setting $\alpha = 0.1$, $\gamma = 0.9$, $\bar{t} = 1$ and starting from the point $(0, 0)$. [Use $\|\nabla f(x)\| < 10^{-3}$ as stopping criterion.]