Unconstrained optimization

Mauro Passacantando

Department of Computer Science, University of Pisa
mauro.passacantando@unipi.it

Numerical Methods and Optimization
Master in Computer Science – University of Pisa

Existence of global minima

An unconstrained optimization problem is defined as

$$\begin{cases} \min \ f(x) \\ x \in \mathbb{R}^n \end{cases}$$

Theorem (Weierstrass)

If f is continuous and $D \subset \mathbb{R}^n$ is closed and bounded, then there exists a global minimum of f on D.

Proof. Let $v^* = \inf_{x \in D} f(x)$. Define a minimizing sequence $\{x^k\} \subseteq D$ s.t. $f(x^k) \to v^*$. The Bolzano-Weierstrass theorem guarantees that there exists a subsequence $\{x^{k_j}\}$ converging to some point $x^* \in D$. Finally, $f(x^{k_j}) \to f(x^*)$ since f is continuous. Therefore, $f(x^*) = v^*$, i.e., x^* is a global minimum. of f on D. □

How can we exploit the Weierstrass Theorem to obtain conditions which guarantee the existence of a global minimum of f on $\mathbb{R}^n$?

Existence of global minima

Corollary 1

If f is continuous and coercive, i.e.,

$$\lim_{\|x\| \to +\infty} f(x) = +\infty,$$

then there exists a global minimum of f on $\mathbb{R}^n$.

Example. $f(x) = x^4 + 3x^3 - 5x^2 + x - 2$.

Corollary 2

If f is strongly convex, then there exists a unique global minimum of f on $\mathbb{R}^n$.

Example. $f(x) = \frac{1}{2} x^\mathsf{T} Q x + c^\mathsf{T} x$, where Q is a symmetric positive definite matrix and $c \in \mathbb{R}^n$.
What if Q is positive semidefinite or indefinite?

First-order optimality conditions

Theorem

If x^* is a local minimum, then $\nabla f(x^*) = 0$.

Proof. By contradiction, assume that $\nabla f(x^*) \neq 0$. Choose direction $d = -\nabla f(x^*)$, define $\varphi(t) = f(x^* + td)$,

$$\varphi'(0) = d^{\mathsf{T}} \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0,$$

thus $f(x^* + td) < f(x^*)$ for all t small enough, which is impossible because x^* is a local optimum. □

Theorem (convex problems)

If f is convex, then x^* is a global minimum if and only if $\nabla f(x^*) = 0$.

Proof. $f(x) \geq f(x^*) + (y - x^*)^{\mathsf{T}} \nabla f(x^*) = f(x^*)$. □

First-order optimality conditions

Exercise. Let $f(x) = \frac{1}{2}x^{\mathsf{T}} Q x + c^{\mathsf{T}} x$, where Q is a symmetric positive definite matrix and $c \in \mathbb{R}^n$. Which is the global minimum of f on $\mathbb{R}^n$?

Exercise. Let $f(x) = -2x_2^3 + x_1\, x_2^2 + x_1^2 - 2\, x_1\, x_2 + 3\, x_2^2$. Find the global and local minima of f on $\mathbb{R}^n$.

Second-order optimality conditions

Theorem (necessary condition)

If x^* is a local minimum, then $\nabla^2 f(x^*)$ is positive semidefinite.

Proof. By contradiction, assume that $\nabla^2 f(x^*)$ is not positive semidefinite, i.e., there is a direction $d \neq 0$ s.t. $d^\mathsf{T} \nabla^2 f(x^*) d = a < 0$. Define $\varphi(t) = f(x^* + td)$, then $\varphi'(0) = d^\mathsf{T} \nabla f(x^*) = 0$ and $\varphi''(0) = a$. The Taylor's formula implies that

$$\begin{aligned} f(x^* + td) &= \varphi(t) \\ &= \varphi(0) + \varphi'(0)t + \frac{1}{2}\varphi''(0)t^2 + o(t^2) \\ &= f(x^*) + \frac{a}{2}t^2 + o(t^2) \\ &= f(x^*) + t^2 \left[\frac{a}{2} + \frac{o(t^2)}{t^2}\right] \\ &< f(x^*) \end{aligned}$$

for all t small enough, which is impossible because x^* is a local minimum. □

Second-order optimality conditions

Theorem (sufficient condition)

If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a local minimum.

Proof. Since $\nabla^2 f(x^*)$ is positive definite, the Weierstrass Theorem guarantees that

$$\min_{d \in \mathbb{R}^n:\ \|d\|=1} d^{\mathsf{T}} \nabla^2 f(x^*) d = a > 0.$$

Then, the Taylor's formula implies that

$$\begin{aligned} f(x^* + h) &= f(x^*) + h^{\mathsf{T}} \nabla f(x^*) + \frac{1}{2} h^{\mathsf{T}} \nabla^2 f(x^*) h + o(\|h\|^2) \\ &= f(x^*) + \|h\|^2 \left[\frac{1}{2} \left(\frac{h}{\|h\|} \right)^{\mathsf{T}} \nabla^2 f(x^*) \left(\frac{h}{\|h\|} \right) + \frac{o(\|h\|^2)}{\|h\|^2} \right] \\ &\geq f(x^*) + \|h\|^2 \left[\frac{a}{2} + \frac{o(\|h\|^2)}{\|h\|^2} \right] \\ &> f(x^*) \end{aligned}$$

for all $h \in \mathbb{R}^n$ such that $\|h\|$ is small enough. Therefore, x^* is a local minimum. □

Gradient method

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

Gradient method

Choose $x^0 \in \mathbb{R}^n$, set $k := 0$

while $\nabla f(x^k) \neq 0$ **do**

compute an optimal solution t_k of the problem

$$\min_{t>0} \ f\left(x^k - t\,\nabla f(x^k)\right),$$

set $x^{k+1} := x^k - t_k\,\nabla f(x^k)$, $k := k+1$.

end

Example. $f(x) = x_1^2 + 2\,x_2^2 - 3\,x_1 - 2\,x_2$, starting point $x^0 = (2,1)$.

$\nabla f(x^0) = (1,2)$, $f(x^0 - t\nabla f(x^0)) = 9\,t^2 - 5\,t - 2$, $t_0 = 5/18$,

$$x^1 = (2,1) - \frac{5}{18}(1,2) = \left(\frac{31}{18}, \frac{4}{9}\right).$$

Gradient method - convergence

Exercise 1. Prove that $\nabla f(x^k)^\mathsf{T} \nabla f(x^{k+1}) = 0$ for any iteration k.

Exercise 2. Prove that if $\{x^k\}$ converges to x^*, then x^* is a stationary point of f, i.e., $\nabla f(x^*) = 0$.

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 strongly 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.

Gradient method - quadratic functions

If $f(x) = \frac{1}{2}x^\mathsf{T} Q x + c^\mathsf{T} x$, with Q positive definite matrix, then

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

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

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

Gradient method - quadratic functions

Exercise 3. Implement in MATLAB the gradient method for solving the problem

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

where Q is a positive definite matrix.

Exercise 4. Run the gradient method for solving 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.]

Gradient method - step size

If f is a general nonlinear function, how to find the step size t_k?

Assume that the restriction $\varphi(t) = f(x^k - t\nabla f(x^k))$ is strongly convex, so we have to find t^* s.t. $\varphi'(t^*) = 0$.

Bisection method

Find $b > 0$ s.t. $\varphi'(b) > 0$ (we know that $\varphi'(0) < 0$).

$a_0 := 0,\ b_0 := b,\ t_0 := \dfrac{a_0 + b_0}{2},\ i := 0.$

while $\varphi'(t_i) \neq 0$ **do**

  if $\varphi'(t_i) > 0$

    then $a_{i+1} := a_i,\ b_{i+1} := t_i$

  else $a_{i+1} := t_i,\ b_{i+1} := b_i$

  end

  $t_{i+1} := \dfrac{a_{i+1} + b_{i+1}}{2},\ i := i + 1$

end

Theorem

The sequence $\{t_i\}$ converges to t^* and $|t_i - t^*| < (b_0 - a_0)/2^{i+1}$.

Gradient method - step size

Example. $\varphi(t) = t^4 + 2\,t^2 - 3\,t.$ $\varphi'(0) = -3,\ \varphi'(1) = 5.$

i	a_i	b_i	t_i	$\varphi'(t_i)$
0	0.000000	1.000000	0.500000	−0.500000
1	0.500000	1.000000	0.750000	1.687500
2	0.500000	0.750000	0.625000	0.476563
3	0.500000	0.625000	0.562500	−0.038086
4	0.562500	0.625000	0.593750	0.212280
5	0.562500	0.593750	0.578125	0.085403
6	0.562500	0.578125	0.570313	0.023241
7	0.562500	0.570313	0.566406	−0.007526
8	0.566406	0.570313	0.568359	0.007831
9	0.566406	0.568359	0.567383	0.000146
10	0.566406	0.567383	0.566895	−0.003692

Gradient method - step size

Newton method (tangent method): write the first order approximation of φ' at t_i:

$$\varphi'(t) \simeq \varphi'(t_i) + \varphi''(t_i)(t - t_i)$$

and solve the linear equation $\varphi'(t_i) + \varphi''(t_i)(t - t_i) = 0$.

Newton method

Choose $t_0 > 0$, set $i := 0$
while $\varphi'(t_i) \neq 0$ **do**

$$t_{i+1} := t_i - \frac{\varphi'(t_i)}{\varphi''(t_i)}, \; i := i + 1$$

end

Theorem

If $\varphi''(t^*) \neq 0$, then $\exists \delta > 0, C > 0$ s.t. for any $t_0 \in (t^* - \delta, t^* + \delta)$ the sequence $\{t_i\}$ converges to t^* and $|t_{i+1} - t^*| \leq C|t_i - t^*|^2$.

Gradient method - step size

Example. $\varphi(t) = t^4 + 2\,t^2 - 3\,t,\ t_0 = 1.$

i	t_i	$\varphi'(t_i)$
0	1.0000000	5.0000000
1	0.6875000	1.0498047
2	0.5789580	0.0920812
3	0.5674799	0.0009093
4	0.5673642	0.0000001

Gradient method - zig-zag behaviour

Two subsequent directions are orthogonal: $\nabla f(x_k)^{\mathsf{T}} \nabla f(x^{k+1}) = 0$

Theorem

If $f(x) = \frac{1}{2}\, x^{\mathsf{T}} Q\, x + c^{\mathsf{T}} x$, with Q positive definite matrix, then

$$\|x^{k+1} - x^*\|_Q \leq \left(\frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1} \right) \|x^k - x^*\|_Q,$$

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.

Gradient method - zig-zag behaviour

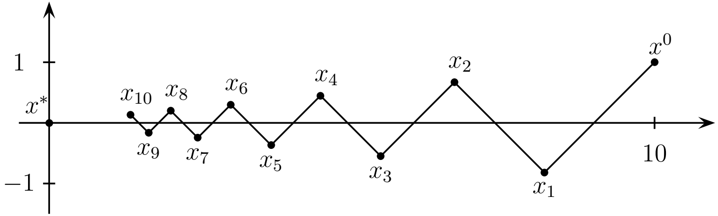
Example. $f(x) = x_1^2 + 10\,x_2^2$, global minimum is $x^* = (0, 0)$.

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

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

hence

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



Gradient method - inexact line search

Gradient method with inexact line search

Set $\alpha, \gamma \in (0,1)$, $\bar{t} > 0$. Choose $x^0 \in \mathbb{R}^n$, set $k := 0$.

while $\nabla f(x^k) \neq 0$ **do**

Compute the smallest nonnegative integer number m such that

$$f(x^k - \gamma^m \, \bar{t} \, \nabla f(x^k)) \leq f(x^k) - \alpha \, \gamma^m \, \bar{t} \, \|\nabla f(x^k)\|^2,$$

$t_k := \gamma^m \, \bar{t}$.
$x^{k+1} := x^k - t_k \nabla f(x^k)$, $k := k+1$

end

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.

Gradient method - inexact line search

Example. $f(x_1, x_2) = x_1^4 + x_1^2 + x_2^2$, starting point $x^0 = (1, 1)$.
$d = -\nabla f(x^0) = (-6, -2)$. Set $\alpha = 10^{-4}$, $\gamma = 0.5$, $\bar{t} = 1$.
Line search. If $m = 0$ then

$$f(x^0 - \nabla f(x^0)) = 651 > f(x^0) - \alpha \|\nabla f(x^0)\|^2 = 2.996,$$

if $m = 1$ then

$$f(x^0 - \gamma \nabla f(x^0)) = 20 > f(x^0) - \alpha \|\nabla f(x^0)\|^2 \gamma = 2.998,$$

if $m = 2$ then

$$f(x^0 - \gamma^2 \nabla f(x^0)) = 0.5625 < f(x^0) - \alpha \|\nabla f(x^0)\|^2 \gamma^2 = 2.999$$

hence the step size is $t_0 = \gamma^2 = 1/4$ and the new iterate is

$$x_1 = x^0 - t_0 \nabla f(x^0) = (1, 1) + \frac{1}{4}(-6, -2) = \left(-\frac{1}{2}, \frac{1}{2}\right).$$

Gradient method with inexact line search

Exercise 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 gradient 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.]

Conjugate gradient method - quadratic functions

Search direction involves directions computed at previous iterations.

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 $g = \nabla f(x) = Qx + c$.

At iteration k, the search direction is

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

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

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

Conjugate gradient method - quadratic functions

- it easy to 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 exact line search, then d^k is a descent direction
- the step size is $t_k = -\dfrac{(g^k)^\mathsf{T} d^k}{(d^k)^\mathsf{T} Q\, d^k}$

Conjugate gradient method (quadratic functions)

Choose $x^0 \in \mathbb{R}^n$, set $g^0 = Q\, x^0 + c$, $k := 0$

while $g^k \neq 0$ **do**

  if $k = 0$ **then** $d^k := -g^k$

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

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

  $x^{k+1} := x^k + t_k\, d^k$, $g_{k+1} = Q\, x^{k+1} + c$, $k := k + 1$

end

Conjugate gradient method - quadratic functions

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

$k = 0$: $g^0 = (20, 20)$, $d^0 = -g^0 = (-20, -20)$,
$t_0 = -((g^0)^\mathsf{T} d^0)/((d^0)^\mathsf{T} Q\, d^0) = 1/11$, hence $x^1 = x^0 + t_0\, d^0 = (90/11, -9/11)$

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

Conjugate gradient method - quadratic functions

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 $x^0 + \mathrm{Span}(d^0, d^1, \dots, d^k)$

Theorem

The conjugate gradient method finds the global minimum after at most n iterations.

Conjugate gradient method - quadratic functions

Theorem

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

- $$\|x^k - x^*\|_Q \leq 2 \left(\frac{\sqrt{\frac{\lambda_n}{\lambda_1}} - 1}{\sqrt{\frac{\lambda_n}{\lambda_1}} + 1} \right)^k \|x^0 - x^*\|_Q$$

- $$\|x^k - x^*\|_Q \leq \left(\frac{\lambda_{n-k+1} - \lambda_1}{\lambda_{n-k+1} + \lambda_1} \right) \|x^0 - x^*\|_Q,$$

  hence the global minimum is found after at most r iterations if Q has r distinct eigenvalues.

Conjugate gradient method - quadratic functions

Exercise 6. Implement in MATLAB the conjugate gradient method for solving the problem

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

where Q is a positive definite matrix.

Exercise 7. Run the conjugate gradient method for solving 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.]

Conjugate gradient method - nonlinear functions

Conjugate gradient method (nonlinear functions)

Choose $x^0 \in \mathbb{R}^n$, set $k := 0$

while $\nabla f(x^k) \neq 0$ **do**

  if $k = 0$ **then** $d^k := -\nabla f(x^k)$

  else $\beta_k := \dfrac{\|\nabla f(x^k)\|^2}{\|\nabla f(x^{k-1})\|^2}$, $d^k := -\nabla f(x^k) + \beta_k \, d^{k-1}$

  Compute the step size t_k

  $x^{k+1} := x^k + t_k \, d^k$, $k := k + 1$

end

Conjugate gradient method - nonlinear functions

If t_k is computed by exact line search, then d^k is a descent direction.

If t_k satisfies the following conditions:

$$\begin{aligned} & f(x^k + t_k\, d^k) \le f(x^k) + \alpha\, t_k\, \nabla f(x^k)^\mathsf{T} d^k, \\ & |\nabla f(x^k + t_k\, d^k)^\mathsf{T} d^k| \le -\beta\, \nabla f(x^k)^\mathsf{T} d^k, \end{aligned} \tag{1}$$

with $0 < \alpha < \beta < 1/2$, then d^k is a descent direction.

Theorem

If f is coercive, then the conjugate gradient method, where (1) holds, generates a sequence $\{x^k\}$ such that

$$\liminf_{k\to\infty} \|\nabla f(x^k)\| = 0.$$

Newton method

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

Newton method

Choose $x^0 \in \mathbb{R}^n$, set $k := 0$

while $\nabla f(x^k) \neq 0$ **do**
  $d^k := -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$
  $x^{k+1} := x^k + d^k$, $k = k + 1$
end

Theorem

Assume that x^* is a local minimum of f and $\nabla^2 f(x^*)$ is positive definite. Then there exists a neighborhood U of x^* such that, for any $x^0 \in U$, the generated sequence $\{x^k\}$ converges to x^* and

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

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

Newton method

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

$$\nabla^2 f(x) = \begin{pmatrix} 24x_1^2 + 4 & 1 \\ 1 & 36x_2^2 + 8 \end{pmatrix} \quad \text{is positive definite for any } x \in \mathbb{R}^2.$$

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

Drawbacks of Newton method:

- at each iteration we have 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 can be not convergent to x^*

Example. 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 the method starts from $x^0 = 1$:
$x^1 = -1$, $x^2 = 1$, $x^3 = -1$, ...

Newton method

If f is strongly convex we get global convergence because d^k is a descent direction: $\nabla f(x^k)^\mathsf{T} d^k = -\nabla f(x^k)^\mathsf{T} [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) < 0$.

Newton method with line search

Set $\alpha, \gamma \in (0,1)$, $\bar{t} > 0$. Choose $x^0 \in \mathbb{R}^n$, $k := 0$

while $\nabla f(x^k) \neq 0$ **do**

$d^k := -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$

Compute the smallest nonnegative integer number m such that

$$f(x^k + \gamma^m \, \bar{t} \, d^k) \leq f(x^k) + \alpha \, \gamma^m \, \bar{t} \, \nabla f(x^k)^\mathsf{T} d^k$$

$t_k := \gamma^m \, \bar{t}$

$x^{k+1} := x^k + t_k \, d^k$, $k := k+1$

end

Theorem

If f is strongly convex then, for any starting point x^0, 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.

Newton method with inexact line search

Exercise 8. 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.]

Quasi-Newton methods

At each iteration $[\nabla^2 f(x^k)]^{-1}$ is approximated by a positive definite matrix H_k

Quasi-Newton method

Choose $x^0 \in \mathbb{R}^n$, a positive definite matrix H_0, $k = 0$

while $\nabla f(x^k) \neq 0$ **do**

$d^k = -H_k \nabla f(x^k)$

Compute step size t_k

$x^{k+1} = x^k + t_k d^k$, update H_{k+1}, $k = k + 1$

end

How to update matrix H_k?

Quasi-Newton methods

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

Set $p^k = x^{k+1} - x^k$ and $g^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, then

$$\nabla^2 f(x^{k+1})\, p^k \simeq g^k, \text{ i.e. } [\nabla^2 f(x^{k+1})]^{-1}\, g^k \simeq p^k.$$

We choose H_{k+1} such that

$$H_{k+1}\, g^k = p^k.$$

Davidon-Fletcher-Powell:

$$H_{k+1} = H_k + \frac{p^k\,(p^k)^{\mathsf{T}}}{(p^k)^{\mathsf{T}} g^k} - \frac{H_k\, g^k (g^k)^{\mathsf{T}}\, H_k}{(g^k)^{\mathsf{T}}\, H_k\, g^k},$$

Quasi-Newton methods

Another approach: find a matrix $B_k = (H_k)^{-1}$ approximating $\nabla^2 f(x^k)$.

Since $\nabla^2 f(x^{k+1})\, p^k \simeq g^k$, we impose that $B_{k+1}\, p^k = g^k$

Update B_k as

$$B_{k+1} = B_k + \frac{g^k (g^k)^\mathsf{T}}{(p^k)^\mathsf{T} g^k} - \frac{B_k\, p^k (p^k)^\mathsf{T} B_k}{(p^k)^\mathsf{T} B_k\, p^k},$$

hence

$$H_{k+1} = H_k + \left(1 + \frac{(g^k)^\mathsf{T} H_k\, g^k}{(p^k)^\mathsf{T} g^k}\right) \frac{p^k (p^k)^\mathsf{T}}{(p^k)^\mathsf{T} g^k} - \frac{p^k (g^k)^\mathsf{T} H_k + H_k\, g^k (p^k)^\mathsf{T}}{(p^k)^\mathsf{T} g^k}.$$

Broyden–Fletcher–Goldfarb–Shanno (BFGS) method

Derivative-free methods

There are situations where derivatives of the objective function do not exist or are computationally expensive.

Derivative-free methods sample the objective function at a finite number of points at each iteration, without any explicit or implicit derivative approximation.

Definition

A set of vectors $\{v_1, \dots, v_p\} \subset \mathbb{R}^n$ is a positive basis if

- any $x \in \mathbb{R}^n$ is a conic combination of $v_1, \dots, v_p$, i.e., there are $\alpha_1, \dots, \alpha_p \geq 0$ such that $x = \sum_{i=1}^{p} \alpha_i v_i$
- no v_i, with $i = 1, \dots, p$, is a conic combination of others $v_1, \dots, v_p$.

Examples: $\{e_1, \dots, e_n, -e_1, \dots, -e_n\}$ is a positive basis of $\mathbb{R}^n$; $\{(1,0), (0,1), (-1,-1)\}$ is a positive basis of $\mathbb{R}^2$.

Proposition. If $\{v_1, \dots, v_p\}$ is a positive basis, then for any $w \in \mathbb{R}^n$, $w \neq 0$ there is $i \in \{1, \dots, p\}$ such that $w^{\mathsf{T}} v_i > 0$.

Directional direct-search method

Directional direct-search method

Choose starting point $x^0 \in \mathbb{R}^n$, step size $t_0 > 0$, $\beta \in (0,1)$, tolerance $\varepsilon > 0$ and a positive basis D. Set $k = 0$.

while $t_k > \varepsilon$ **do**
  Order the poll set $\{x^k + t_k d, \quad d \in D\}$
  Evaluate f at the poll points following the chosen order
  if there is a poll point s.t. $f(x^k + t_k d) < f(x^k)$
    then $x^{k+1} = x^k + t_k d$, $t_{k+1} = t_k$ (successful iteration)
  else $x^{k+1} = x^k$, $t_{k+1} = \beta t_k$ (step size reduction)
  end
  $k = k + 1$
end

The method is called coordinate-search method if $D = \{e_1, \dots, e_n, -e_1, \dots, -e_n\}$.

Directional direct-search method

Theorem

Assume that all the vectors of the positive basis D are in $\mathbb{Z}^n$. If f is coercive and continuously differentiable, then the generated sequence $\{x^k\}$ has a cluster point x^* such that $\nabla f(x^*) = 0$.

Remark 1. The assumption that vectors of D are in $\mathbb{Z}^n$ can be deleted if we accept new iterates which satisfy a "sufficient" decrease condition:

$$f(x^{k+1}) \leq f(x^k) - t_k^2.$$

Remark 2. If a complete poll step is performed, i.e.,

$$f(x^{k+1}) \leq f(x^k + t_k d) \qquad \forall\ d \in D,$$

then *any* cluster point of $\{x^k\}$ is a stationary point of f and $\lim_{k\to\infty} \|\nabla f(x^k)\| = 0$.

Variants of the Directional direct-search method

Directional direct-search method

Choose starting point $x^0 \in \mathbb{R}^n$, step size $t_0 > 0$, $\beta \in (0,1)$, $\gamma \geq 1$, tolerance $\varepsilon > 0$ and a set of positive bases $\mathcal{D}$. Set $k = 0$.

while $t_k > \varepsilon$ **do**

  Choose a positive basis $D \in \mathcal{D}$

  Order the poll set $\{x^k + t_k d, \quad d \in D\}$

  Evaluate f at the poll points following the chosen order

  if there is a poll point s.t. $f(x^k + t_k d) < f(x^k)$

    then $x^{k+1} = x^k + t_k d$, $t_{k+1} = \gamma t_k$ (successful iteration)

  else $x^{k+1} = x^k$, $t_{k+1} = \beta t_k$ (step size reduction)

  end

  $k = k + 1$

end

Directional direct-search method

Exercise 9. 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 directional direct-search method setting $x^0 = (0,0)$, $t_0 = 5$, $\beta = 0.5$, $\varepsilon = 10^{-5}$ and the positive basis $D = \{(1,0),(0,1),(-1,0),(0,-1)\}$.

Exercise 10. Solve the previous problem by means of the directional direct-search method setting the positive basis $D = \{(1,0),(0,1),(-1,-1)\}$.