Nonlinear optimization: theory and algorithms

Giancarlo Bigi

a.y. 2015/16

These lecture notes deal with algorithms for computing the minima (or maxima) of a (generally nonlinear) function $f : \mathbb{R}^n \to \mathbb{R}$ over some set $D \subseteq \mathbb{R}^n$, that is finding $\bar{x} \in \mathbb{R}^n$ such that

(a) $\bar{x} \in D$ (feasibility)

(b) $f(\bar{x}) \leq f(x)$ for any $x \in D$ (optimality)

This problem can be briefly stated in the following way:

$$\min\{f(x) \ : \ x \in D\}$$

while

$$\arg\min\{f(x) \ : \ x \in D\}$$

denotes the set of the optimal solutions, i.e.,

$$\arg\min\{f(x) \ : \ x \in D\} = \{x \in \mathbb{R}^n \ : \ (a) \text{ and } (b) \text{ hold}\}.$$

Chapter 1 provides the basic background material on the topology of the Euclidean space $\mathbb{R}^n$ and multivariate calculus that is needed.

Chapter 1

Topology and calculus background

We consider $\mathbb{R}^n$ endowed with the scalar (or inner) product

$$x^T y = \sum_{i=1}^{n} x_i y_i$$

which induces the Euclidean norm

$$\|x\|_2 = \sqrt{x^T x} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

The following properties hold for any $x, y \in \mathbb{R}^n$ and any $\alpha \in \mathbb{R}$:

$$\|x\|_2 \geq 0$$
$$\|\alpha x\|_2 = |\alpha| \|x\|_2$$
$$\|x\|_2 = 0 \iff x = 0$$
$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$$
$$(\|x - y\|_2 \leq \|x\|_2 + \|y\|_2)$$
$$|x^T y| \leq \|x\|_2 \|y\|_2. \text{ (Schwarz inequality).}$$

In turn, the Euclidean norm induces the well-known Euclidean distance between the points $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$:

$$d(x, y) = \|x - y\|_2$$

and the following properties can be deduced from the above ones:

$$d(x, y) \geq 0$$
$$d(x, y) = 0 \iff x = y$$
$$d(x, y) \leq d(x, z) + d(z, x).$$

1.1 Sequences

A family of points $\{x^k\}_{k\in\mathbb{N}} \subseteq \mathbb{R}^n$ (i.e., $\{x^1, x^2, \dots, x^k, \dots\}$) is called a *sequence*. For instance, the family of points $x^k = (1/k, 1/k^2)$ is a sequence in $\mathbb{R}^2$.

Definition 1.1. $\bar{x} \in \mathbb{R}^n$ is the *limit of a sequence* $\{x^k\}_{k\in\mathbb{N}}$ if for each $\varepsilon > 0$ there exists $\bar{k} \in \mathbb{N}$ such that $d(x^k, \bar{x}) \le \varepsilon$ for all $k \ge \bar{k}$, or equivalently

$$\forall\ \varepsilon > 0 \quad \exists\ \bar{k} \in \mathbb{N} \quad \text{s.t.} \quad \|x^k - \bar{x}\|_2 \le \varepsilon \quad \forall\ k \ge \bar{k}.$$

If it exists, the limit of a sequence is unique. Standard notations to denote a limit are the following: $\lim_{k\to+\infty} x^k = \bar{x}$, $x^k \longrightarrow \bar{x}$ ($k \to +\infty$ below the arrow is often omitted).

Example 1.1. The limit of the sequence $(1/k, 1/k^2)$ is $\bar{x} = (0, 0)$, while the sequence $x^k = (1/k, (-1)^k)$ does not have a limit. Take the sequence obtained just considering odd indices: $x^1, x^3, x^5, \dots$ This sequence converges to $(0, -1)$. Analogously, the sequence obtained considering just even indices converges to $(0, 1)$.

Definition 1.2. $\{x^{k_j}\}_{j\in\mathbb{N}} \subseteq \{x^k\}_{k\in\mathbb{N}}$ is a *subsequence* if $k_j \to +\infty$ as $j \to +\infty$.

Definition 1.3. $\bar{x} \in \mathbb{R}^n$ is a *cluster point* of $\{x^k\}_{k\in\mathbb{N}}$ if there exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $\bar{x}$ is its limit, i.e., $\lim_{j\to+\infty} x^{k_j} = \bar{x}$, or equivalently

$$\forall\ \varepsilon > 0 \quad \forall\ k \in \mathbb{N} \quad \exists\ \bar{k} \ge k \quad \text{s.t.} \quad \|x^{\bar{k}} - \bar{x}\|_2 \le \varepsilon.$$

If a sequence has a limit, then it is the unique cluster point of the sequence.

Example 1.2. The last sequence of Example 1.1 has 2 cluster points: $(0, 1)$ and $(0, -1)$, while the sequence $y^k = (k, 1/k)$ does not have any cluster point.

Theorem 1.1. (Bolzano-Weierstrass) *If the norm of all the points of a sequence $\{x^k\}_{k\in\mathbb{N}}$ do not exceed a threshold value, i.e., there exists $M > 0$ such that $\|x^k\|_2 \le M$ holds for all $k \in \mathbb{N}$, then the sequence has at least one cluster point.*

1.2 Topological properties in the Euclidean space

The open ball of centre $x \in \mathbb{R}^n$ and radius $\varepsilon > 0$ is the set

$$B(x, \varepsilon) = \{y \in \mathbb{R}^n \ :\ \|y - x\|_2 < \varepsilon\}.$$

Definition 1.4.

(*i*) $D \subseteq \mathbb{R}^n$ is called *open* if

$$\forall\ x \in D \quad \exists\ \varepsilon > 0 \quad \text{s.t.} \quad B(x, \varepsilon) \subseteq D.$$

(*ii*) $x \in D$ is called an *interior point* of D if

$$\exists\ \varepsilon > 0 \quad \text{s.t.} \quad B(x, \varepsilon) \subseteq D.$$

The set of the interior points of D is called *the interior of* D and it is generally denoted by $\operatorname{int} D$. Notice that a set D is open if and only if $D = \operatorname{int} D$.

Example 1.3. $B(x,\varepsilon)$, $\mathbb{R}^n$, $\emptyset$ are open sets in $\mathbb{R}^n$ while the interval $]-1,1[$ is an open set in $\mathbb{R}$.

Proposition 1.1.

(*i*) *The union of a family of open sets is an open set.*

(*ii*) *The intersection of a finite family of open sets is an open set.*

The finiteness of the family is crucial for the intersection property:

$$\bigcap_{k=1}^{+\infty} B(0,1/k) = \{0\}.$$

Definition 1.5.

(*i*) $D \subseteq \mathbb{R}^n$ is called *closed* if $\mathbb{R}^n \setminus D = \{x \in \mathbb{R}^n \ : \ x \notin D\}$ is open.

(*ii*) $x \in \mathbb{R}^n$ is called an *closure point* of D if

$$\forall\, \varepsilon > 0 \ : \ B(x,\varepsilon) \cap D \neq \emptyset.$$

The set of the closure points of D is called *the closure of* D and it is generally denoted by $\operatorname{cl} D$ or $\overline{D}$.

Proposition 1.2.

(*i*) *D is closed if and only if $D = \operatorname{cl} D$.*

(*ii*) *D is closed if and only if the limit of any convergent sequence contained in D belongs to D as well, i.e.,*

$$\forall\, \{x^k\}_{k\in\mathbb{N}} \subseteq D \quad s.t. \quad \exists\, \bar{x} \in \mathbb{R}^n \quad s.t. \quad x^k \longrightarrow \bar{x} \ : \ \bar{x} \in D.$$

Example 1.4. $\mathbb{R}^n$, $\emptyset$, $\{y \in \mathbb{R}^n \ : \ \|y - x\|_2 \leq \varepsilon\} = \overline{B(x,\varepsilon)}$ are closed sets in $\mathbb{R}^n$ while the interval $[-1,1]$ is a closed set in $\mathbb{R}$. There exist sets which are neither closed nor open, for instance the interval [-1,1[in $\mathbb{R}$ and

$$D = [-1,0] \times [-1,1] \cup B(0,1) \subseteq \mathbb{R}^2.$$

In fact, $(-1-\varepsilon, 0) \notin D$ but $(-1-\varepsilon, 0) \in B((-1,0),\varepsilon)$ for any $\varepsilon > 0$ so that D is not open, and $x^k = (1 - 1/k, 0) \in D$ for any $k \in \mathbb{N}$ while $x^k \to (1,0) \notin D$ so that D is not closed.

Proposition 1.3.

(*i*) *The union of a finite family of closed sets is an closed set.*

(*ii*) *The intersection of a family of closed sets is a closed set.*

The finiteness of the family is crucial for the union property:

$$\bigcup_{k=2}^{+\infty} \overline{B(0, 1-1/k)} = B(0,1).$$

Definition 1.6. $x \in \mathbb{R}^n$ is called a *boundary point* of D if both

$$B(x,\varepsilon) \cap D \neq \emptyset \quad \text{and} \quad B(x,\varepsilon) \not\subseteq D$$

hold for any $\varepsilon > 0$.

The set of the boundary points of D is called *the boundary (or frontier) of* D and it is generally denoted by ∂D. Notice that $\partial D = \overline{D} \cap \overline{(\mathbb{R}^n \setminus D)}$.

Proposition 1.4. *$D \subseteq \mathbb{R}^n$ is both closed and open if and only if $D = \mathbb{R}^n$ or $D = \emptyset$.*

Definition 1.7.

(*i*) $D \subseteq \mathbb{R}^n$ is called *bounded* if

$$\exists\, M > 0 \quad \text{s.t.} \quad \forall\, x \in D \;:\; \|x\|_2 \leq M.$$

(*ii*) $D \subseteq \mathbb{R}^n$ is called *compact* if it is bounded and closed.

The set D in Example 1.4 is bounded but it is not compact (since it is not closed).

The Bolzano-Weierstrass' theorem can be enhanced in the following way.

Theorem 1.2. (Bolzano-Weierstrass) *A set is compact if and only if any sequence contained in the set has at least one cluster point and all its cluster points belong to the set.*

1.3 Functions of several variables

1.3.1 Continuity

Definition 1.8. $f : \mathbb{R}^n \to \mathbb{R}$ is called *continuous at* $\bar{x} \in \mathbb{R}^n$ if $f(\bar{x})$ is the limit of $f(x)$ as $x \to \bar{x}$, i.e.,

$$\forall\, \varepsilon > 0 \quad \exists\, \delta > 0 \quad \text{s.t.} \quad \|x - \bar{x}\|_2 \leq \delta \implies |f(x) - f(\bar{x})| \leq \varepsilon.$$

f is continuous on a set $D \subseteq \mathbb{R}^n$ if it is continuous at every $x \in D$.

Proposition 1.5. *f is continuous at $\bar{x} \in \mathbb{R}^n$ if and only if any sequence $\{x^k\}_{k\in\mathbb{N}}$ such that $x^k \longrightarrow \bar{x}$ satisfies $f(x^k) \longrightarrow f(\bar{x})$.*

Example 1.5. $f(x) = \|x\|_2$ is a continuous function on $\mathbb{R}^n$, $f(x_1, x_1) = \sin(\pi x_1 x_2)$ is a continuous function on $\mathbb{R}^2$.

Theorem 1.3. (Weierstrass) *Let $D \subseteq \mathbb{R}^n$ be compact and $f : \mathbb{R}^n \to \mathbb{R}$ continuous on D. Then, there exist at least one minimum point $\bar{x} \in D$ and one maximum point $\hat{x} \in D$ for f over D, i.e.,*

$$f(\bar{x}) = \min\{f(x) \ : \ x \in D\} \quad \textit{and} \quad f(\hat{x}) = \max\{f(x) \ : \ x \in D\}.$$

Proof. Let $\ell = \inf\{f(x) \ : \ x \in D\} \in [-\infty + \infty[$ and consider any minimizing sequence, that is any $\{x^k\}_{k\in\mathbb{N}}$ such that $f(x^k) \to \ell$. Since D is compact, there exist a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ and $\bar{x} \in D$ such that $x^{k_j} \to \bar{x}$ (as $j \to +\infty$) by Theorem 1.2. Since f is continuous, $f(x^{k_j}) \to f(\bar{x})$ and therefore $f(\bar{x}) = \ell$ by the uniqueness of the limit. As a consequence, $\ell \neq -\infty$ and $f(\bar{x}) = \min\{f(x) \ : \ x \in D\}$. The existence of $\hat{x}$ can be proved analogously. □

Example 1.6. Take $n = 1$, $f(x) = e^{-x}$ and $D = \mathbb{R}_+$: f is continuous on D, $\inf\{f(x) \ : \ x \in D\} = 0$ but there exists no $x \in D$ such that $f(x) = 0$. Indeed, D is not compact as it is not bounded.

1.3.2 Partial derivatives and differentiability

A point $d \in \mathbb{R}^n$ such that $\|d\|_2 = 1$ is also called a *direction*, and the set

$$\{\bar{x} + td \ : \ t \in \mathbb{R}\}$$

describes the line of direction d passing through $\bar{x} \in \mathbb{R}^n$. If only $t \in \mathbb{R}_+$ are considered, the set describes the corresponding half-line.

Just like the case $n = 1$, the key tool for developing calculus for a function $f : \mathbb{R}^n \to \mathbb{R}$ is the *incremental ratio*

$$icr_{(f,x,d)}(t) = [f(x + td) - f(x)]/t.$$

Definition 1.9. f has a derivative at $\bar{x}$ in the direction d if the derivative of the function of one variable $icr_{(f,\bar{x},d)}$ at $t = 0$ exists, that is $\lim\limits_{t\to 0} [f(\bar{x} + td) - f(\bar{x})]/t$ exists. In that case

$$\frac{\partial f}{\partial d}(\bar{x}) = \lim_{t\to 0} \ \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

is called the *(directional) derivative of f at $\bar{x}$ in the direction d*. For $n = 1$ there exists a unique (up to the sign) direction and the directional derivative coincides with the (usual) derivative and it is also denoted by $f'(\bar{x})$.

If d is one of the vectors of the canonical basis $\{e_1, \ldots, e_n\}$ of $\mathbb{R}^n$, namely $d = e_i$, then the corresponding directional derivative is called *partial derivative* and denoted by $\partial f(x)/\partial x_i$ rather than $\partial f(x)/\partial e_i$. Indeed, the derivative can be computed considering f as a function of x_i while the other variables are kept fixed like parameters:

$$\frac{\partial f}{\partial x_i}(\bar{x}) = \lim_{t\to 0} \ \frac{f(\bar{x}_1, \ldots, \bar{x}_{i-1}, \bar{x}_i + t, \bar{x}_{i+1}, \ldots, \bar{x}_n) - f(\bar{x})}{t}$$

Definition 1.10. If f has all the partial derivatives at $\bar{x} \in \mathbb{R}^n$, the vector

$$\nabla f(\bar{x}) = \Big(\frac{\partial f}{\partial x_1}(\bar{x}), \frac{\partial f}{\partial x_2}(\bar{x}), \ldots, \frac{\partial f}{\partial x_n}(\bar{x})\Big)^T$$

is called the *gradient of f at $\bar{x}$.*

Example 1.7. Take $n = 2$ and $f(x_1, x_2) = \sin(\pi x_1 x_2)$:

$$\frac{\partial f}{\partial x_1}(x) = \pi x_2 \cos(\pi x_1 x_2), \quad \frac{\partial f}{\partial x_2}(\bar{x}) = \pi x_1 \cos(\pi x_1 x_2).$$

Other directional derivatives can be defined just considering the limit of the incremental ratio as $t \to 0^+$, that is $t \to 0$ for only positive t ($t > 0$).

Definition 1.11. The limit

$$f'(\bar{x}; d) = \lim_{t \to 0^+} \frac{f(\bar{x} + td) - f(\bar{x})}{t}$$

is called the *one-sided directional derivative of f at $\bar{x}$ in the direction d.*

Clearly, $f'(\bar{x}; d) = \partial f(\bar{x})/\partial d$ if the latter exists but this is not always the case.

Example 1.8. Consider $f(x) = \|x\|_2$ and take $\bar{x} = 0$:

$$[f(\bar{x} + td) - f(\bar{x})]/t = \|td\|_2/t = |t|\|d\|_2/t = \mathrm{sgn}(t)\|d\|_2$$

where $\mathrm{sgn}(t)$ denotes the sign of t ($\mathrm{sgn}(t) = 1$ if $t \geq 0$ and $\mathrm{sgn}(t) = -1$ if $t < 0$). Therefore, $f'(\bar{x}; d) = \|v\|_2 = 1$ while $\partial f(\bar{x})/\partial d$ does not exist.

Unlike the case $n = 1$, the existence of the directional/partial derivatives does not guarantee the continuity of the function.

Example 1.9. Take $n = 2$ and

$$f(x_1, x_2) = \begin{cases} [x_1^2 x_2/(x_1^4 + x_2^2)]^2 & \text{if } (x_1, x_2) \neq (0, 0) \\ \\ 0 & \text{if } (x_1, x_2) = (0, 0). \end{cases}$$

Consider the parabola $x_2 = \alpha x_1^2$ for $x_1 \neq 0$:

$$f(x_1, \alpha x_1^2) = [\alpha x_1^4/(x_1^4 + \alpha^2 x_1^4)] = \alpha^2/(1 + \alpha^2)^2.$$

Therefore, f is not continuous at $\bar{x} = (0, 0)$: take the sequence $x^k = (1/k, 1/k^2)$ to get $x^k \to \bar{x}$ while $f(x^k) \equiv 1/4$. On the other hand, f has the directional derivative at $\bar{x}$ in each direction d:

$$\frac{\partial f}{\partial d}(\bar{x}) = \lim_{t \to 0} [t^3 d_1^2 d_2/t^2(t^2 d_1^4 + d_2^2)]^2/t = \lim_{t \to 0} t d_1^4 d_2^2/((t^2 d_1^4 + d_2^2)^2 = 0.$$

Definition 1.12. f is called *differentiable at* $\bar{x} \in \mathbb{R}^n$ if there exists a linear function $L : \mathbb{R}^n \to \mathbb{R}$ such that

$$\forall\, v \in \mathbb{R}^n \;:\; f(\bar{x}+v) = f(\bar{x}) + L(v) + r(v)$$

for some residual function r such that $r(v)/\|v\|_2 \to 0$ as $\|v\|_2 \to 0$. If f is differentiable at $\bar{x}$, L is called the *differential of* f *at* $\bar{x}$. Notice that both L and r depend not only on f but also on the considered point $\bar{x}$.
f is differentiable on a set $D \subseteq \mathbb{R}^n$ if it is differentiable at every $x \in D$.

Recall that $L : \mathbb{R}^n \to \mathbb{R}$ is linear if

$$\forall\, x, y \in \mathbb{R}^n\ \forall\, \alpha, \beta \in \mathbb{R} \;:\; L(\alpha x + \beta y) = \alpha L(x) + \beta L(y).$$

L is linear if and only if there exists $\ell \in \mathbb{R}^n$ such that $L(x) = \ell^T x$ for all $x \in \mathbb{R}^n$.

Proposition 1.6. *Suppose f is differentiable at $\bar{x} \in \mathbb{R}^n$. Then,*

(i) f is continuous at $\bar{x}$;

(ii) f has directional derivatives at $\bar{x}$ in each direction d and $\dfrac{\partial f}{\partial d}(\bar{x}) = L(d)$;

(iii) $L(d) = \nabla f(\bar{x})^T d$.

Proof. (i) It is enough to apply Definition 1.12 just taking $h = x - \bar{x}$ as $x \to \bar{x}$.
(ii) Take any direction $d \in \mathbb{R}^n$. Then, Definition 1.12 implies

$$\begin{aligned}
\frac{\partial f}{\partial d}(\bar{x}) &= \lim_{t\to 0} \left(f(\bar{x}+td) - f(\bar{x})\right)/t \\
&= \lim_{t\to 0} \left(L(td) + r(td)\right)/t \\
&= \lim_{t\to 0} \left(tL(d) + r(td)\right)/t \\
&= L(d) + \lim_{t\to 0} r(td)/t \\
&= L(d) + \lim_{t\to 0} \operatorname{sgn}(t)\left(r(td)\right)/\|td\|_2) = L(d).
\end{aligned}$$

(iii) Since $d = \sum_{i=1}^{n} d_i e_i$, (ii) implies

$$\frac{\partial f}{\partial d}(\bar{x}) = L(d) = L\Big(\sum_{i=1}^{n} d_i e_i\Big) = \sum_{i=1}^{n} d_i L(e_i) = \sum_{i=1}^{n} d_i \frac{\partial f}{\partial x_i}(\bar{x}) = \nabla f(\bar{x})^T d. \qquad \square$$

Proposition 1.6 (iii) allows to restate the definition of differentiability through (the first order) Taylor's formula:

Taylor's formula $f(\bar{x}+v) = f(\bar{x}) + \nabla f(\bar{x})^T v + r(v) \quad (r(v)/\|v\|_2 \to 0)$

Considering any $v = x - \bar{x} \approx 0$, Taylor's formula states that $f(x)$ can be approximated by an affine function, namely $f(x) \approx f(\bar{x}) + \nabla f(\bar{x})^T(x-\bar{x})$, and the closer x is to $\bar{x}$ the better the approximation is. Indeed, the set

$$\{\big(x, f(\bar{x}) + \nabla f(\bar{x})^T(x-\bar{x})\big) \;:\; x \in \mathbb{R}^n\}$$

is the tangent hyperplane to the graph $\{(x, f(x)) \;:\; x \in \mathbb{R}^n\}$ of f at $(\bar{x}, f(\bar{x}))$.

Theorem 1.4. *Let $\bar{x} \in \mathbb{R}^n$ and suppose f has all the partial derivatives at each $x \in B(\bar{x}, \varepsilon)$ for some $\varepsilon > 0$. Then, if the functions $x \mapsto \partial f(x)/\partial x_i$ are continuous at $\bar{x}$ for all $i = 1, ..., n$, then f is differentiable at $\bar{x}$.*

Example 1.10. Take $n = 2$ and

$$f(x_1, x_2) = \begin{cases} x_1^2 x_2/(x_1^2 + x_2^2) & \text{if } (x_1, x_2) \neq (0, 0) \\ \\ 0 & \text{if } (x_1, x_2) = (0, 0) \end{cases}$$

and consider $\bar{x} = (0, 0)$: f is continuous but not differentiable at $\bar{x}$. In fact, the derivative of f at $\bar{x}$ in the direction d is

$$\frac{\partial f}{\partial d}(\bar{x}) = \lim_{t \to 0} [t^3 d_1^2 d_2/t^2(d_1^2 + d_2^2)]/t = d_1^2 d_2$$

since $1 = \|d\|_2^2 = d_1^2 + d_2^2$. As a consequence, $\partial f(\bar{x})/\partial x_1 = \partial f(\bar{x})/\partial x_2 = 0$ while $\partial f(\bar{x})/\partial d \neq 0$ for all $d \neq e_1, e_2$ so that $\partial f(\bar{x})/\partial d \neq \nabla f(\bar{x})^T d$ (see Proposition 1.6).

Notice that

$$\frac{\partial f}{\partial x_1}(x) = 2x_1 x_2^3/(x_1^2 + x_2^2)^2 \qquad (x \neq \bar{x})$$

is not continuous at $\bar{x}$ (in accordance with Theorem 1.4): $x^k = (1/k, 1/k) \to \bar{x}$ while $\partial f(x^k)/\partial x_1 \equiv 1/2$ and $\partial f(\bar{x})/\partial x_1 = 0$.

Definition 1.13. f is called *continuously differentiable at $\bar{x} \in \mathbb{R}^n$* if there exists $\varepsilon > 0$ such that f is differentiable at each $x \in B(\bar{x}, \varepsilon)$ and the partial derivatives are continuous at $\bar{x}$. f is continuously differentiable on a set $D \subseteq \mathbb{R}^n$ if it is continuously differentiable at every $x \in D$.

Theorem 1.5. (mean value) *Suppose f is continuously differentiable (on $\mathbb{R}^n$). Given any $\bar{x}, v \in \mathbb{R}^n$, there exists $t \in]0, 1[$ such that*

$$f(\bar{x} + v) = f(\bar{x}) + \nabla f(\bar{x} + tv)^T v.$$

Theorem 1.6. (upper estimate) *Suppose f is continuously differentiable (on $\mathbb{R}^n$) and the gradient mapping ∇f is Lipschitz with modulus $L > 0$, i.e.,*

$$\forall\, x, v \in \mathbb{R}^n \ : \ \|\nabla f(x) - \nabla f(v)\|_2 \leq L\|x - v\|_2.$$

Then, any $x, v \in \mathbb{R}^n$ satisfy $f(x + v) \leq f(x) + \nabla f(\bar{x} + v)^T v + L\|v\|_2^2/2$.

Proposition 1.7. (chain rules)

(i) If $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\bar{x} \in \mathbb{R}^n$ and $h : \mathbb{R} \to \mathbb{R}$ has a derivative at $f(\bar{x})$, then $f = h \circ g$ is differentiable at $\bar{x}$ and $\nabla f(\bar{x}) = h'(g(\bar{x}))\nabla g(\bar{x})$.

(ii) Let $h = (h_1, \ldots, h_n) : \mathbb{R} \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}$. If the functions $h_i : \mathbb{R} \to \mathbb{R}$ have a derivative at $\bar{t} \in \mathbb{R}$ for all $i = 1, ..., n$ and g is differentiable at $h(\bar{t}) \in \mathbb{R}^n$, then $g \circ h$ has a derivative at $\bar{t}$ and $(g \circ h)'(\bar{t}) = \nabla g(h(\bar{t}))^T h'(\bar{t})$ where $h'(\bar{t}) = (h_1'(\bar{t}), \ldots, h_n'(\bar{t}))^T$.

Definition 1.14. Let $F = (f_1, \dots, f_m) : \mathbb{R}^n \to \mathbb{R}^m$. If the functions $f_i : \mathbb{R}^n \to \mathbb{R}$ have all the partial derivatives at $\bar{x} \in \mathbb{R}^n$ for all $i = 1, ..., n$, then

$$JF(\bar{x}) = \begin{bmatrix} \nabla f_1(\bar{x})^T \\ \vdots \\ \nabla f_n(\bar{x})^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\bar{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\bar{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\bar{x}) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

is called the *Jacobian matrix of F at $\bar{x}$*.

1.3.3 Second-order derivatives

If a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on the whole $\mathbb{R}^n$, then each directional derivative exists at each point $x \in \mathbb{R}^n$. In this case, the derivative in the direction d is the function $\partial f/\partial d : \mathbb{R}^n \to \mathbb{R}$ such that $(\partial f/\partial d)(x) = \partial f(x)/\partial d$. If it has a derivative in the direction v, then

$$\frac{\partial}{\partial v}\Big(\frac{\partial f}{\partial d}\Big)(x) = \lim_{t \to 0} \Big[\frac{\partial f}{\partial d}(x + tv) - \frac{\partial f}{\partial d}(x)\Big]/t$$

is generally denoted by $\partial^2 f(x)/\partial v \partial d$.

Definition 1.15. f has *second-order partial derivatives* at $\bar{x} \in \mathbb{R}^n$ if it has the (first-order) partial derivatives at each $x \in B(\bar{x}, \varepsilon)$ for some $\varepsilon > 0$ and they have partial derivatives at $\bar{x}$ as well, namely

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \lim_{t \to 0} \Big[\frac{\partial f}{\partial x_j}(\bar{x} + tv) - \frac{\partial f}{\partial x_j}(\bar{x})\Big]/t$$

for all $i, j = 1, ..., n$. If $i = j$, then the derivative is generally denoted by $\partial^2 f(\bar{x})/\partial x_i^2$. For $n = 1$ there exists a unique second-order directional derivative which coincides with the (usual) second-order derivative and it is also denoted by $f''(\bar{x})$.

Example 1.11. Take the function of Example 1.7:

$$\frac{\partial f}{\partial x_1}(x) = \pi x_2 \cos(\pi x_1 x_2), \quad \frac{\partial f}{\partial x_2}(\bar{x}) = \pi x_1 \cos(\pi x_1 x_2),$$

$$\frac{\partial^2 f}{\partial x_2 \partial x_1}(x) = \pi \cos(\pi x_1 x_2) - \pi^2 x_1 x_2 \sin(\pi x_1 x_2) = \frac{\partial^2 f}{\partial x_1 \partial x_2}(x)$$

$$\frac{\partial^2 f}{\partial x_1^2}(x) = -\pi^2 x_2^2 \sin(\pi x_1 x_2), \quad \frac{\partial^2 f}{\partial x_2^2}(x) = -\pi^2 x_1^2 \sin(\pi x_1 x_2).$$

Theorem 1.7. (Schwarz) *Let $\bar{x} \in \mathbb{R}^n$ and suppose f has the second-order partial derivatives $\partial^2 f/\partial x_i \partial x_j$ and $\partial^2 f/\partial x_j \partial x_i$ at each $x \in B(\bar{x}, \varepsilon)$ for some $\varepsilon > 0$. If both the derivatives are continuous at $\bar{x}$, then*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\bar{x}).$$

Definition 1.16. If f has *second-order partial derivatives* at $\bar{x} \in \mathbb{R}^n$, then

$$\nabla^2 f(\bar{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\bar{x}) & \cdots & \frac{\partial f}{\partial x_1 \partial x_n}(\bar{x}) \\ \vdots & \vdots & \vdots \\ \frac{\partial f}{\partial x_n \partial x_1}(\bar{x}) & \cdots & \frac{\partial f}{\partial x_n^2}(\bar{x}) \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is called the *Hessian matrix of f at $\bar{x}$*.

Definition 1.17. f is called *twice continuously differentiable* at $\bar{x} \in \mathbb{R}^n$ if it has second-order partial derivatives at each $x \in B(\bar{x}, \varepsilon)$ for some $\varepsilon > 0$ and they are continuous at $\bar{x}$. f is twice continuously differentiable on a set $D \subseteq \mathbb{R}^n$ if it is twice continuously differentiable at every $x \in D$.

Notice that the Hessian matrix of a twice continuously differentiable function is symmetric and therefore all its eigenvalues are real numbers.

Theorem 1.8. (Taylor's formulas) *Suppose f is twice continuously differentiable (on $\mathbb{R}^n$). The following statements hold for any $\bar{x} \in \mathbb{R}^n$:*

(i) $\forall\ v \in \mathbb{R}^n\ \exists\ t \in]0,1[$ *such that* $f(\bar{x}+v) = f(\bar{x}) + \nabla f(\bar{x})^T v + \frac{1}{2} v^T \nabla^2 f(\bar{x}+tv) v$;

(ii) $\forall\ v \in \mathbb{R}^n\ :\ f(\bar{x}+v) = f(\bar{x}) + \nabla f(\bar{x})^T v + \frac{1}{2} v^T \nabla^2 f(\bar{x}) v + r(v)$
for some residual function r such that $r(v)/\|v\|_2^2 \to 0$ as $\|v\|_2 \to 0$.

Definition 1.18. f is called *quadratic* if there exist $Q \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$ such that

$$f(x) = \frac{1}{2} x^T Q x + b^T x + c = \frac{1}{2} \sum_{k=1}^{\ell} \sum_{\ell=1}^{n} q_{k\ell} x_k x_\ell + \sum_{k=1}^{n} b_k x_k + c.$$

Without loss of generality, Q can be taken symmetric, eventually replacing it by $(Q + Q^T)/2$ since $q_{k\ell} x_k x_\ell + q_{\ell k} x_\ell x_k = (q_{k\ell} + q_{\ell k}) x_k x_\ell / 2 + (q_{k\ell} + q_{\ell k}) x_\ell x_k / 2$.

The partial derivatives of a quadratic function can be easily computed:

$$\frac{\partial f}{\partial x_i}(x) = \frac{1}{2}\Big(\sum_{\ell=1}^{n} q_{i\ell} x_\ell + \sum_{k=1}^{n} q_{ki} x_k \Big) + b_i = \Big(\sum_{\ell=1}^{n} q_{i\ell} x_\ell \Big) + b_i = (Qx)_i + b_i$$

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(x) = \frac{\partial}{\partial x_j}\Big(\frac{\partial f}{\partial x_i} \Big)(x) = \frac{\partial f}{\partial x_j}\Big(\sum_{\ell=1}^{n} q_{i\ell} x_\ell + b_i \Big) = q_{ij}.$$

Therefore, $\nabla f(x) = Qx + b$ and $\nabla^2 f(x) = Q$.

Considering any $v = x - \bar{x} \approx 0$, the second-order Taylor's formula states that $f(x)$ can be approximated by a quadratic function, namely $f(x) \approx q(x)$ with

$$q(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x}),$$

that is

$$q(x) = \frac{1}{2}x^T\nabla^2 f(\bar{x})x + \big(\nabla f(\bar{x}) - \nabla^2 f(\bar{x})\bar{x}\big)^T x + \big(f(\bar{x}) - \nabla f(\bar{x})^T\bar{x} + \frac{1}{2}\bar{x}^T\nabla^2 f(\bar{x})\bar{x}\big).$$

Example 1.12. Take $n = 2$ and $f(x_1, x_2) = -x_1^4 - x_2^2$:

$$\nabla f(x) = \begin{pmatrix} -4x_1^3 \\ -2x_2 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} -12x_1^2 & 0 \\ 0 & -2 \end{bmatrix}.$$

Considering $\bar{x} = (0, -2/5)$ the quadratic approximation of $f(x)$ near $\bar{x}$ is given by

$$q(x) = -2x_2^2 - 12x_2/5 - 20/25.$$

Chapter 2

Optimization, convex functions and sets

The basic ingredients of an optimization problem are a real-valued function and a subset of its domain over which looking for the minima and/or maxima of the function. Given any $f : \mathbb{R}^n \to \mathbb{R}$ and a set $D \subseteq \mathbb{R}^n$, the minimization problem

$$(P) \qquad \min\{f(x) \ : \ x \in D\}$$

amounts to finding $\bar{x} \in D$ such that $f(\bar{x}) \leq f(x)$ for any $x \in D$. The infimum of the set of real numbers $\{f(x) \ : \ x \in D\}$ is called the *optimal value* of (P) if it is finite. If the infimum is $-\infty$, then the minimization problem (P) is called *unbounded by below*. The corresponding definitions for maximization problems are given just recalling that maximizing f over a set D is equivalent to minimizing $-f$ over the same set. Optimization problems are often called programs, so that optimization and programming are synonyms in this framework.

Optimization problems can be classified according to the kind of function f and set D which are involved. *Unconstrained optimization* deals with the case $D = \mathbb{R}^n$, while *constrained optimization* with the case $D \neq \mathbb{R}^n$. If D is finite or countable, then *discrete optimization* comes into play: *combinatorial optimization* deals specifically with the case $D \subseteq \{0,1\}^n$, while *integer programming* with the generic case $D \subseteq \mathbb{Z}^n$. If D is uncountable (and f is continuous), the most used term is *continuous optimization*. If f is linear and D is a polyhedron, then *linear programming* is used, otherwise *nonlinear optimization or nonlinear programming*.

These lecture notes focus on continuous nonlinear optimization both in the unconstrained and constrained case.

2.1 Optimization and convexity

Definition 2.1. $\bar{x} \in \mathbb{R}^n$ is called a *global minimum point* of (P) if it is feasible and $f(\bar{x})$ is the optimal value of (P), namely if

(i) $\bar{x} \in D$ (feasibility)

(ii) $f(\bar{x}) \leq f(x)$ for any $x \in D$ (optimality).

A global minimum point $\bar{x}$ is called *strict* if the strict inequality $f(\bar{x}) < f(x)$ holds for any $x \in D$ with $x \neq \bar{x}$.

Example 2.1. Take $n = 1$, $f(x) = -x^2$ and $D = \mathbb{R}$: the optimal value does not exist since $f(x) \to -\infty$ as $x \to \pm\infty$, hence no global mimimum point may exist.

Take (P) with f and D as in Example 1.6: the optimal value is 0 but no global minimum point exists all the same.

Definition 2.2. $\bar{x} \in \mathbb{R}^n$ is called a *local minimum point* of (P) if

(i) $\bar{x} \in D$ (feasibility)

(ii) $\exists\ \varepsilon > 0$ such that $f(\bar{x}) \leq f(x)$ for any $x \in D \cap B(\bar{x}, \varepsilon)$ (local optimality).

A local minimum point $\bar{x}$ is called *strict* if there exists $\varepsilon' > 0$ such that $f(\bar{x}) < f(x)$ holds for any $x \in D \cap B(\bar{x}, \varepsilon')$ with $x \neq \bar{x}$.

Clearly, any global minimum point is also a local minimum point but not vice versa. Indeed, finding a global minimum may be much harder than finding a local minimum: the distinction between *global optimization* and *local optimization* is generally very meaningful.

Definition 2.3. $D \subseteq \mathbb{R}^n$ is called *convex* if $\lambda x + (1-\lambda)y \in D$ for any $x, y \in D$ and any $\lambda \in [0, 1]$.

Example 2.2. $\mathbb{R}^n$, $\emptyset$, $B(x, \varepsilon)$, $[\ell, u] = \{x \in \mathbb{R}^n \ : \ \ell_i \leq x_i \leq u_i, \quad i = 1, ..., n\}$ with $\ell, u \in \mathbb{R}^n$ are convex sets in $\mathbb{R}^n$.

Definition 2.4. Let $D \subseteq \mathbb{R}^n$ be convex. Then, $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex on* D if

$$\forall\ x, y \in D,\ \lambda \in [0, 1]\ :\ f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

f is called *strictly convex on* D if the strict inequality

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$$

holds whenever $x, y \in D$ and $\lambda \in [0, 1]$ satisfy $x \neq y$ and $\lambda \neq 0, 1$.

f is called *[strictly] concave on* D if $-f$ is [strictly] convex on D. If $D = \mathbb{R}^n$, then f is simply called [strictly] convex/concave, omitting the (convex) domain of convexity/concavity.

Notice that f is convex on D if and only if given any $k \in \mathbb{N}$ the inequality

$$f(\sum_{i=1}^{k} \lambda_i x_i) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

holds for any $x_1, \ldots, x_k \in D$ and any $\lambda_1, \ldots, \lambda_k \in [0, 1]$ such that $\lambda_1 + \cdots + \lambda_k = 1$.

Theorem 2.1. *Let $D \subseteq \mathbb{R}^n$ be convex. Then,*

(i) If f is convex on D, then any local minimum point of (P) is also a global minimum point.

(ii) If f is strictly convex on D, there exists at most one minimum point of (P).

Proof. (i) Ab absurdo, suppose there exists a local minimum point $\bar{x} \in D$ which is not a global minimum point. Thus, there exists $x \in D$ such that $f(x) < f(\bar{x})$. Consider $x(\lambda) = \lambda x + (1-\lambda)\bar{x}$: it belongs to D for any $\lambda \in [0,1]$ by the convexity of D. Furthermore, the convexity of f implies

$$f(x(\lambda)) \leq \lambda f(x) + (1-\lambda) f(\bar{x}) < \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x})$$

for any $\lambda \neq 0$. Since $x(\lambda) \to \bar{x}$ as $\lambda \to 0$, for any $\varepsilon > 0$ there exists $\bar{\lambda} \in [0,1[$ such that $x(\bar{\lambda}) \in B(\bar{x}, \varepsilon)$. Thus, $\bar{x}$ is not a local minimum point contradicting the assumption.

(ii) Suppose there exist $\bar{x}, \hat{x} \in D$ which are both minimum points of (P). Thus, $f(\hat{x}) = f(\bar{x})$. Take any $\lambda \in]0,1[$ and $x(\lambda) = \lambda\hat{x} + (1-\lambda)\bar{x}$. If $\bar{x} \neq \hat{x}$, then the strict convexity of f implies

$$f(x(\lambda)) < \lambda f(\hat{x}) + (1-\lambda) f(\bar{x}) = \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x}) = f(\hat{x})$$

so that neither $\bar{x}$ nor $\hat{x}$ is a minimum point of (P). □

The theorem shows that local and global optimization coincide if f and D are both convex, therefore the distinction between *convex optimization* and *nonconvex optimization* is very meaningful as well.

Proposition 2.1. *Let $D \subseteq \mathbb{R}^n$ be convex and f be convex on D. Then, the set of all the minimum points of (P) is convex*

Proof. Take any two minimum points $\bar{x}, \hat{x} \in D$ $(\hat{x} \neq \bar{x})$ and any $\lambda \in [0,1]$. Then, $x(\lambda) = \lambda\hat{x} + (1-\lambda)\bar{x} \in D$ by the convexity of D, while the convexity of f implies

$$f(x(\lambda)) \leq \lambda f(\hat{x}) + (1-\lambda) f(\bar{x}) = \lambda f(\bar{x}) + (1-\lambda) f(\bar{x}) = f(\bar{x}) = f(\hat{x}) \leq f(x(\lambda))$$

where the last inequality is due to the optimality of $\bar{x}$. Thus, $f(x(\lambda)) = f(\bar{x}) = f(\hat{x})$ and therefore $x(\lambda)$ is a minimum point as well. □

2.2 Properties of convex functions

Proposition 2.2. *Let $D \subseteq \mathbb{R}^n$ be convex. Then, f is convex on D if and only if the restriction of the epigraph of f to D, namely*

$$\mathrm{epi}_D(f) = \{(x,t) \ : \ x \in D, \ t \geq f(x)\}$$

is a convex set in $\mathbb{R}^{n+1}$.

Proof. *Only if)* Take any $(x,t),(y,\tau) \in \mathrm{epi}_D(f)$ and any $\lambda \in [0,1]$:

$$\lambda t + (1-\lambda)\tau \geq \lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y)$$

where the last inequality is due to the convexity of f. Moreover, $\lambda x + (1-\lambda)y \in D$ since D is convex, and thus $(\lambda x + (1-\lambda)y, \lambda t + (1-\lambda)\tau) \in \mathrm{epi}_D(f)$.

If) Take any $x, y \in D$ and any $\lambda \in [0,1]$. Therefore, $(x, f(x)), (y, f(y)) \in \mathrm{epi}_D(f)$ and the the convexity of $\mathrm{epi}_D(f)$ imply $(\lambda x + (1-\lambda)y, \lambda f(x) + (1-\lambda) f(y)) \in \mathrm{epi}_D(f)$, which reads

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \qquad \square$$

Proposition 2.3. *Let $D \subseteq \mathbb{R}^n$ be convex and f be convex on D. Then, the intersection of the α-sublevel set of f with D, namely*

$$\{x \in \mathbb{R}^n \ : \ f(x) \leq \alpha\} \cap D$$

is a convex set for any $\alpha \in \mathbb{R}$.

Proof. Take any $x, y \in D$ and any $\lambda \in [0,1]$: $\lambda x + (1-\lambda)y \in D$ since D is convex and moreover

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \leq \lambda\alpha + (1-\lambda)\alpha = \alpha \qquad \square$$

Example 2.3. Take $n = 1$, $f(x) = x^3$ and $D = \mathbb{R}$. The α-sublevel set

$$\{x \in \mathbb{R}^n \ : \ x^3 \leq \alpha\} =]-\infty, \sqrt[3]{\alpha}]$$

is convex, while f is not convex. In fact, it is enough to take $x = -1$, $y = 1/2$ and $\lambda = 1/3$ to get $f(\lambda x + (1-\lambda)y) = f(0) = 0 > -1/4 = \lambda f(x) + (1-\lambda) f(y)$.

Proposition 2.4. *Let $D \subseteq \mathbb{R}^n$ be convex and $f_i : \mathbb{R}^n \to \mathbb{R}$ be convex on D for any $i \in I$ for some index set I.*

(i) If I is finite, then $\Big(\sum_{i\in I} f_i\Big)(x) = \sum_{i\in I} f_i(x)$ is convex on D;

(ii) $\Big(\sup_{i\in I} f_i\Big)(x) = \sup_{i\in I} f_i(x)$ is convex on D.

Proof. (i) It is enough to exploit the definition of convexity for each f_i summing all the inequalities.

(ii) Take any $x, y \in D$ and any $\lambda \in [0,1]$. Given any $\varepsilon > 0$, there exists $k = k(\varepsilon) \in I$ such that

$$\begin{aligned}
\Big(\sup_{i\in I} f_i\Big)(\lambda x + (1-\lambda)y) &\leq f_k(\lambda x + (1-\lambda)y) + \varepsilon \\
&\leq \lambda f_k(x) + (1-\lambda) f_k(y) + \varepsilon \\
&\leq \lambda \Big(\sup_{i\in I} f_i\Big)(x) + (1-\lambda)\Big(\sup_{i\in I} f_i\Big)(y) + \varepsilon
\end{aligned}$$

and therefore

$$\Big(\sup_{i\in I} f_i\Big)(\lambda x + (1-\lambda)y) \le \lambda\Big(\sup_{i\in I} f_i\Big)(x) + (1-\lambda)\Big(\sup_{i\in I} f_i\Big)(y)$$

follows since ε is arbitrary. □

Theorem 2.2. *Let $D \subseteq \mathbb{R}^n$ be a convex set with a nonempty interior. If f is convex on D, then f is continuous on $\operatorname{int} D$.*

Theorem 2.3. *Let f be differentiable on D. Then, f is convex on D if and only if*

$$\forall\ x, y \in D\ :\ f(y) \ge f(x) + \nabla f(x)^T (y - x).$$

Proof. *Only if)* Take any $x, y \in D$ and any $\lambda \in [0,1]$: the definition of convexity can be equivalently stated as

$$f(y) - f(x) \ge [f(x + \lambda(y-x)) - f(x)]/\lambda.$$

By Proposition 1.6 $\lim_{\lambda\to 0} [f(x + \lambda(y-x)) - f(x)]/\lambda = \nabla f(x)^T (y-x)$ Therefore, the required inequality follows from the above one just taking the limit as $\lambda \to 0^+$.

If) Take any $x, y \in D$ and any $\lambda \in [0,1]$: the following inequalities follow from the assumption just considering the pairs of points x, $\lambda x + (1-\lambda)y$ and y, $\lambda x + (1-\lambda)y$:

$$\begin{aligned} f(x) &\ge f(\lambda x + (1-\lambda)y) + (1-\lambda)\nabla f(\lambda x + (1-\lambda)y)^T (x - y), \\ f(y) &\ge f(\lambda x + (1-\lambda)y) + \lambda \nabla f(\lambda x + (1-\lambda)y)^T (x - y). \end{aligned}$$

Summing λ times the first inequality with $(1-\lambda)$ times the second inequality gives

$$\lambda f(x) + (1-\lambda) f(y) \ge f(\lambda x + (1-\lambda)y).$$ □

Roughly speaking, the theorem states that the convexity of f means exactly that the graph of f is "above" the tangent hyperplane to the graph at $(x, f(x))$ everywhere on D for any $x \in D$.

An analogous characterization holds for strict convexity.

Theorem 2.4. *Let f be differentiable on D. Then, f is strictly convex on D if and only if*

$$\forall\ x, y \in D \text{ s.t. } x \ne y\ :\ f(y) > f(x) + \nabla f(x)^T (y - x).$$

Theorem 2.5. *Let f be twice continuously differentiable (on $\mathbb{R}^n$). Then, f is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for any $x \in \mathbb{R}^n$, i.e.,*

$$\forall\ x, y \in \mathbb{R}^n\ :\ y^T \nabla^2 f(x) y \ge 0.$$

Proof. *Only if)* Take any $x, y \in \mathbb{R}^n$ and any $t \in \mathbb{R}$: the second-order Taylor's formula and Theorem 2.3 guarantee

$$\frac{1}{2}t^2 y^T \nabla^2 f(x) y + r_{(f,x)}(ty) = f(x+ty) - f(x) - t\nabla f(x)^T y \geq 0$$

and therefore (supposing $y \neq 0$)

$$\frac{1}{2} y^T \nabla^2 f(x) y + \frac{r_{(f,x)}(ty)}{\|ty\|_2^2} \|y\|_2^2 \geq 0.$$

Taking the limit as $t \to 0$, the inequality $y^T \nabla^2 f(x) y \geq 0$ follows.

If) Take any $x, y \in \mathbb{R}^n$. By Theorem 1.8 (i) there exists $t \in [0,1]$ such that

$$f(y) - f(x) - \nabla f(x)^T (y-x) = \frac{1}{2}(y-x)^T \nabla^2 f(x + t(y-x))(y-x) \geq 0$$

where the inequality is due to the positive semidefiniteness of $\nabla^2 f(x + t(y-x))$. Therefore, Theorem 2.3 guarantees that f is convex. □

A similar sufficient condition for strict convexity can be proved in the same way.

Theorem 2.6. *Let f be twice continuously differentiable (on $\mathbb{R}^n$). If $\nabla^2 f(x)$ is positive definite for any $x \in \mathbb{R}^n$, i.e.,*

$$\forall\ x, y \in \mathbb{R}^n\ :\ y^T \nabla^2 f(x) y > 0,$$

then f is strictly convex.

Example 2.4. Take $n = 1$ and $f(x) = x^4$: f is strictly convex but $\nabla^2 f(0) = f''(0) = 0$.

Theorem 2.7. *Let $f(x) = \frac{1}{2} x^T Q x + b^T x + c$ (with $Q = Q^T$). Then,*

(i) f is convex if and only if Q is positive semidefinite;

(ii) f is strictly convex if and only if Q is positive definite.

Proof. (i) and the "if part" of (ii) are just Theorems 2.5 and 2.6 for the quadratic function f since $\nabla^2 f(x) = Q$ for any $x \in \mathbb{R}^n$.

(ii) *Only if)* The residual in the second-order Taylor's formula is zero, that is

$$\frac{1}{2}t^2 y^T Q y = f(x+ty) - f(x) - t\nabla f(x)^T y$$

for any $x, y \in \mathbb{R}^n$. By Theorem 2.4 the right-hand side is always positive, hence Q is positive definite. □

Chapter 3

Optimality conditions for unconstrained optimization

Given any $f : \mathbb{R}^n \to \mathbb{R}$, optimality conditions for the unconstrained minimization problem

$$(P) \qquad \min\{f(x) \ : \ x \in \mathbb{R}^n\}$$

can be achieved exploiting Taylor's formulas whenever f is differentiable or twice continuously differentiable. The corresponding optimality conditions for unconstrained maximization can be obtained replacing f by $-f$.

3.1 Optimality conditions

Theorem 3.1. *Suppose $\bar{x} \in \mathbb{R}^n$ is a local minimum point of (P).*

(i) If f is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) = 0$;

(ii) If f is twice continuously differentiable at $\bar{x}$, then $\nabla^2 f(\bar{x})$ is positive semidefinite.

Proof. Local optimality guarantees the existence of $\varepsilon > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in B(\bar{x}, \varepsilon)$. Let $d \in \mathbb{R}^n$ be any direction and $t \in]0, \varepsilon[$: $\|d\|_2 = 1$ guarantees $\bar{x} + td \in B(\bar{x}, \varepsilon)$ and therefore $f(\bar{x}) \leq f(\bar{x} + td)$.

(i) Taylor's formula implies

$$0 \leq f(\bar{x} + td) - f(\bar{x}) = t\nabla f(\bar{x})^T d + r(td)$$

and therefore

$$\nabla f(\bar{x})^T d + r(td)/t \geq 0.$$

Since $t = \|td\|_2$, the limit of left-hand side as $t \to 0^+$ provides $\nabla f(\bar{x})^T d \geq 0$. Considering $-d$ the same reasoning provides also $\nabla f(\bar{x})^T d \leq 0$. Thus, $\nabla f(\bar{x})^T d = 0$

holds for any $d \in \mathbb{R}^n$. Taking $d = -\nabla f(\bar{x})$, the equality reads $\|\nabla f(\bar{x})\|_2^2 = 0$ and hence $\nabla f(\bar{x}) = 0$ follows.

(ii) The second-order Taylor's formula (see Theorem 1.8) implies

$$0 \leq f(\bar{x} + td) - f(\bar{x}) = t\nabla f(\bar{x})^T d + \frac{1}{2}t^2 d^T \nabla^2 f(\bar{x}) d + r(td).$$

Since (i) guarantees $\nabla f(\bar{x}) = 0$, then

$$d^T \nabla^2 f(\bar{x}) d + r(td)/2t^2 \geq 0$$

holds too. Since $t^2 = \|td\|_2^2$, the limit of the left-hand side as $t \to 0$ provides the inequality $d^T \nabla^2 f(\bar{x}) d \geq 0$. Since d is an arbitrary direction, $\nabla^2 f(\bar{x})$ is positive semidefinite. □

If $\bar{x} \in \operatorname{int} D$ minimizes f over some $D \subseteq \mathbb{R}^n$, then the necessary conditions of Theorem 3.1 hold also in this case: the above proof still works just considering any $\varepsilon > 0$ which in addition satisfies $B(\bar{x}, \varepsilon) \subseteq D$.

Definition 3.1. $\bar{x} \in \mathbb{R}^n$ is called a *stationary point of* f if $\nabla f(\bar{x}) = 0$.

Looking for stationary points of f amounts to solving the system of n equations

$$\begin{cases} \frac{\partial f}{\partial x_1}(x_1, ..., x_n) = 0 \\ \qquad\qquad \vdots \\ \frac{\partial f}{\partial x_n}(x_1, ..., x_n) = 0 \end{cases}$$

in the n unknowns $(x_1, \ldots, x_n)$. This is generally a nonlinear system, but if the quadratic function $f(x) = \frac{1}{2}x^T Q x + b^T x + c$ is considered then it is actually the linear system $Qx = -b$ (since $\nabla f(x) = Qx + b$). If f is strictly convex, then $\nabla^2 f(x) \equiv Q$ is positive definite and therefore invertible: $\bar{x} = -Q^{-1}b$ is the unique stationary point and it is the unique minimum point (see Theorems 3.2 and 3.3 below). On the contrary, if f is not convex, due to Theorem 3.1(ii) no stationary point is a local minimum since Q is not positive semidefinite.

Example 3.1. Take $n = 2$ and $f(x_1, x_2) = (x_2 - x_1^2)(x_2 - 4x_1^2)$:

$$\nabla f(x) = \begin{pmatrix} 16x_1^3 - 10x_1 x_2 \\ 2x_2 - 5x_1^2 \end{pmatrix}, \quad \nabla^2 f(x) = \begin{bmatrix} 48x_1^2 - 10x_2 & -10x_1 \\ -10x_1 & 2 \end{bmatrix}.$$

Then, $\nabla f(x) = 0$ if and only if $x = (0, 0)$ and moreover

$$\nabla^2 f(0, 0) = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

is positive semidefinite (but not definite). Anyway, $(0, 0)$ is not a local minimum point of (P). In fact,

$$f(x_1, 2x_1^2) = -2x_1^2 < 0$$

for any $x_1 \neq 0$. Therefore, f is negative along the parabola $\{x \in \mathbb{R}^2 \ : \ x_2 = 2x_1^2\}$. Notice that f is not even a local maximum point of (P) : $\nabla^2 f(0,0)$ is not negative semidefinite and in fact f is positive along all the parabolas $\{x \in \mathbb{R}^2 \ : \ x_2 = \alpha x_1^2\}$ with $\alpha > 4$.

Theorem 3.2. *Let f be twice continuously differentiable at $\bar{x} \in \mathbb{R}^n$. If $\bar{x}$ is a stationary point of f such that $\nabla^2 f(\bar{x})$ is positive definite, then it is a strict local minimum point of (P) and moreover there exist $\delta, \gamma > 0$ such that*

$$\forall\ x \in B(\bar{x}, \delta) \ : \ f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\|_2^2.$$

Proof. It is enough to prove the above inequality as it guarantees strict local optimality too. Taking any $x \in \mathbb{R}^n$, the second-order Taylor's formula (see Theorem 1.8) implies

$$\begin{aligned} f(x) - f(\bar{x}) &= \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}) + r(x - \bar{x}) \\ &= \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}) + r(x - \bar{x}) \\ &\geq \frac{1}{2}\lambda_{min} \|x - \bar{x}\|_2^2 + r(x - \bar{x}) \end{aligned}$$

and therefore

$$[f(x) - f(\bar{x})]/\|x - \bar{x}\|_2^2 \geq \lambda_{min}/2 + r_{(f,\bar{x})}(x - \bar{x})/\|x - \bar{x}\|_2^2$$

where $\lambda_{min} > 0$ is the minimum eigenvalue of $\nabla^2 f(\bar{x})$.[1] Choose any positive threshold $\varepsilon < \lambda_{min}/2$. Since the limit of the right-hand side as $x \to \bar{x}$ is $\lambda_{min}/2$, there exists $\delta > 0$ such that

$$\forall\ x \in B(\bar{x}, \delta) \ : \ [f(x) - f(\bar{x})]/\|x - \bar{x}\|_2^2 \geq (\lambda_{min}/2 - \varepsilon).$$

Setting $\gamma = \lambda_{min}/2 - \varepsilon$, the thesis follows from the above inequality. □

If f is a strictly convex quadratic function, then the above theorem holds with $\gamma = \lambda_{min}/2$ (where λ_{min} is the minimum eigenvalue of Q) and any $\delta > 0$. In fact, $\bar{x} = -Q^{-1}b$ is the unique stationary point of f and

$$\forall\ x \in \mathbb{R}^n \ : \ f(x) - f(\bar{x}) = \frac{1}{2}(x - \bar{x})^T Q (x - \bar{x}).$$

3.2 Optimality conditions in the convex case

Theorem 3.3. *Let f be convex and differentiable (on $\mathbb{R}^n$). Then, $\bar{x} \in \mathbb{R}^n$ is a minimum point of (P) if and only if $\nabla f(\bar{x}) = 0$.*

[1] Given any symmetric matrix $Q \in \mathbb{R}^{n \times n}$ the inequality $y^T Q y \geq \lambda_{min} \|y\|_2^2$ holds for any $y \in \mathbb{R}^n$ where λ_{min} denotes the minimum eigenvalue of Q

Proof. *Only if)* It is just Theorem 3.1(i).

If) By Theorem 2.3 the convexity of f guarantees

$$f(y) \geq f(\bar{x}) + \nabla f(\bar{x})^T (y - \bar{x})$$

for any $y \in \mathbb{R}^n$. Since $\nabla f(\bar{x}) = 0$, the optimality of $\bar{x}$ follows immediately. □

Notice that any (twice continuously differentiable) convex function f satisfies the second-order optimality condition of Theorem 3.1 at any point (see Theorem 2.5). Moreover, it does not have any global maximum point unless it is a constant function: in fact, a maximum point is a stationary point (just apply Theorem 3.1 to $-f$) and hence it is also a minimum point by Theorem 3.3. The same reasoning applies to local maximum points, which may exist if they are actually also minimum points.

The minimum points of the convex quadratic function $f(x) = \frac{1}{2}x^T Qx + b^T x + c$ are the solutions of the linear system $Qx + b = 0$. If Q is positive definite, then $-Q^{-1}b$ is the unique minimum point. If Q is positive semidefinite but not positive definite, there are infinitely many minimum points if at least one exists but f could be unbounded by below.

Proposition 3.1. *Let $f(x) = \frac{1}{2}x^T Qx + b^T x + c$ be convex. Then, f is unbounded by below if and only if there exists $\hat{x} \in \mathbb{R}^n$ such that $Q\hat{x} = 0$ and $b^T \hat{x} \neq 0$.*

Proof. *If)* Take $x(t) = t\hat{x}$. If $b^T \hat{x} > 0$ (< 0), then

$$f(x(t)) = t(b^T \hat{x}) + c \to -\infty \quad \text{as} \quad t \to -\infty \ (+\infty)$$

Only if) Since Q is symmetric, there exists an orthonormal basis $\{x^1, \ldots, x^n\}$ of $\mathbb{R}^n$ composed by eigenvectors of Q, that is ${x^i}^T x^j = 0$ for all $i \neq j$ and $Qx^i = \lambda_i x^i$ for all $i = 1, \ldots, n$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of Q. Given any $x \in \mathbb{R}^n$, there exist $\gamma_1, \ldots, \gamma_n \in \mathbb{R}$ such that $x = \sum_{i=1}^{n} \gamma_i x^i$. Therefore,

$$f(x) = \frac{1}{2} \sum_{i=1}^{n} \lambda_i \gamma_i^2 + \sum_{i=1}^{n} (b^T x^i)\gamma_i = \sum_{i=1}^{n} [\frac{1}{2}\lambda_i \gamma_i^2 + (b^T x^i)\gamma_i].$$

Ab absurdo, suppose $b^T x = 0$ whenever $Qx = 0$, which implies that $b^T x^i = 0$ if $\lambda_i = 0$. Therefore, each nonzero term in the above sum gets its minimum value for $\gamma_i = \bar{\gamma}_i = -b^T x^i \lambda_i$, and f is bounded by below since

$$f(x) \geq \sum_{i \in I} [\frac{1}{2}\lambda_i \bar{\gamma}_i^2 + (b^T x^i)\bar{\gamma}_i] = -\sum_{i \in I} (b^T x^i)^2 / 2\lambda_i$$

where $I = \{i \ : |\lambda_i \neq 0\}$. □

Chapter 4

Algorithms for unconstrained optimization

This chapter describes some of the most well-known solution methods for the unconstrained minimization problem

$$(P) \qquad \min\{f(x) \ : \ x \in \mathbb{R}^n\}$$

in which $f : \mathbb{R}^n \to \mathbb{R}$ is any (twice) continuously differentiable function.

The main focus will be on *iterative descent methods*, that is iterative algorithms generating a sequence $x^0, x^1, \ldots, x^k, \ldots$ that satisfies the descent property

$$f(x^0) > f(x^1) > \cdots > f(x^k) > f(x^{k+1}) > \ldots$$

or the (weaker) non-monotone descent property

$$\forall\ k \in \mathbb{N} \quad \exists\ m \in \mathbb{N} \quad \text{s.t.} \quad f(x^k) > f(x^{k+m}).$$

The algorithms aim at finding a stationary point, i.e., some $\bar{x} \in \mathbb{R}^n$ such that $\nabla f(\bar{x}) = 0$, which is not necessarily a local minimum point of (P) unless f is convex. Beyond *finite convergence*, that is the existence of some $\bar{k}$ such that $\nabla f(x^{\bar{k}}) = 0$, three different kinds of *asymptotic convergence* may be achieved:

- (i) the sequence $\{x^k\}_{k\in\mathbb{N}}$ has a limit, that is a stationary point of f, i.e., $\lim_{k\to+\infty} x^k = \bar{x}$ for some $\bar{x} \in \mathbb{R}^n$ such that $\nabla f(\bar{x}) = 0$;
- (ii) each cluster point of $\{x^k\}_{k\in\mathbb{N}}$ is a stationary point of f;
- (iii) at least one cluster point of $\{x^k\}_{k\in\mathbb{N}}$ is a a stationary point of f.

The generic iteration can always be described through

$$x^{k+1} = x^k + t_k d^k$$

where $d^k \in \mathbb{R}^n$ identifies the direction along which the algorithm moves away from x^k with stepsize $t_k > 0$. Therefore, a full description of an algorithm can be provided specifying the way d^k and t_k are chosen. Notice that it is not necessary to require $\|d\|_2 = 1$ since the stepsize t_k can be determined accordingly.

4.1 Gradient methods

A *descent direction for* f *at* $x \in \mathbb{R}^n$ is any $d \in \mathbb{R}^n$ such that $f(x+td) < f(x)$ holds whenever $t > 0$ is small enough. Consider any x that is not a stationary point for f, i.e., $\nabla f(x) \neq 0$. Since Proposition 1.6 (ii) guarantees

$$\lim_{t \to 0} \frac{f(x+td) - f(x)}{t} = \nabla f(x)^T d,$$

$\nabla f(x)^T d < 0$ is a sufficient condition for d to be a descent direction. Indeed, the best choice to gain the (asymptotic) maximum decrease is clearly the direction d that provides the minimum value for $\nabla f(x)^T d$.

Proposition 4.1. *Given any* $x \in \mathbb{R}^n$ *which satisfies* $\nabla f(x) \neq 0$*, then* $-\nabla f(x)$ *is a descent direction for* f *at* x *and*

$$\arg\min\{\nabla f(x)^T d \ : \ \|d\|_2 = 1\} = \{-\nabla f(x)/\|\nabla f(x)\|_2\}.$$

Proof. If $d = -\nabla f(x)$, then $\nabla f(x)^T d = -\|\nabla f(x)\|_2^2 < 0$, and the first part of the statement follows immediately. Since $\nabla f(x)^T d = \|\nabla f(x)\|_2 \|d\|_2 \cos\theta$, where θ is the angle formed by the vectors $\nabla f(x)$ and d in the 2-dimensional subspace of $\mathbb{R}^n$ (plane) which contains both, then

$$\min\{\nabla f(x)^T d \ : \ \|d\|_2 = 1\} = \|\nabla f(x)\|_2 \min\{\cos\theta \ : \ \theta \in [0, 2\pi]\}.$$

The minimum value is clearly achieved when $\cos\theta = -1$, that is $\theta = \pi$. Therefore, the direction d, which provides the minimum value, is collinear and opposite to $\nabla f(x)$, that is $d = -\nabla f(x)/\|\nabla f(x)\|_2$. □

The above proposition can be rephrased as "the gradient of a function points in the direction of (asymptotic) maximum increase", or its opposite points in the direction of maximum decrease (steepest descent direction). Notice that the constraint $\|d\|_2 = 1$ is essential in the proposition, otherwise the minimization problem would be unbounded by below as $\nabla f(x)^T d < 0$ implies $\nabla f(x)^T (td) \to -\infty$ as $t \to +\infty$.

Once a descent direction d has been chosen, the ideal choice for the stepsize would be any minimum point of the one dimensional *search function*

$$\varphi(t) = f(x+td),$$

over $\mathbb{R}_+$, i.e., any $t \in \arg\min\{\varphi(t) \ : \ t \geq 0\}$. Such a choice is generally referred to as *exact line search*.

4.1.1 The gradient method with exact line search

Given any x^k, which is not stationary for f, the most straightforward choices are to take the direction $d^k = -\nabla f(x^k)$ and the corresponding stepsize t_k provided by the exact line search. The resulting algorithm is summarized below.

Algorithm 1 – Gradient method with exact line search

0. Choose $x^0 \in \mathbb{R}^n$ and set $k = 0$
1. If $\nabla f(x^k) = 0$, then *STOP*
2. Compute $t_k \in \arg\min\{f(x^k - t\nabla f(x^k)) \ : \ t \geq 0\}$
3. $x^{k+1} = x^k - t_k \nabla f(x^k)$
4. $k = k + 1$ and go to 1

Clearly, Algorithm 1 is a descent method as $-\nabla f(x^k)$ is a descent direction for f at x^k and the exact line search is performed. This can be checked also exploiting the properties of the search function $\varphi_k(t) = f(x^k - t\nabla f(x^k))$.

Proposition 4.2. *Let $\{x^k\}$ be the sequence produced by Algorithm 1. If x^k is not a stationary point of f, then $f(x^{k+1}) < f(x^k)$.*

Proof. The choice of t_k guarantees $\varphi_k(0) = f(x^k) \geq f(x^{k+1}) = \varphi_k(t_k)$. Note that $\varphi_k = f \circ h$ with $h(t) = x^k - t\nabla f(x^k)$. Since f is differentiable at any x and the components of h have a derivative at any t, then φ_k has a derivative at any t and

$$\varphi_k'(t) = -\nabla f(x^k - t\nabla f(x^k))^T \nabla f(x^k)$$

by Proposition 1.7. In particular, $\varphi_k'(0) = -\|\nabla f(x^k)\|_2^2 < 0$ implies $\varphi_k(t) < \varphi_k(0)$ whenever t is small enough. Since t_k minimizes φ_k over $\mathbb{R}_+$, then $\varphi_k(t_k) < \varphi_k(0)$, i.e., $f(x^{k+1}) < f(x^k)$. □

The basic convergence result is a straightforward consequence of the following property stating that any two successive directions in Algorithm 1 are orthogonal.

Proposition 4.3. *Let $\{x^k\}$ be the sequence produced by Algorithm 1. If x^k is not a stationary point of f, then $\nabla f(x^{k+1})^T \nabla f(x^k) = 0$.*

Proof. The proof of Proposition 4.2 shows also that $t_k > 0$. Therefore, since it minimizes φ_k over $\mathbb{R}_+$, then $0 = \varphi_k'(t_k) = -\nabla f(x^{k+1})^T \nabla f(x^k)$. □

Theorem 4.1. *Suppose that Algorithm 1 generates an infinite sequence $\{x^k\}$. If $\lim_{k\to+\infty} x^k = \bar{x}$ for some $\bar{x} \in \mathbb{R}^n$, then $\nabla f(\bar{x}) = 0$.*

Proof. Proposition 4.3 and the continuity of the partial derivatives imply

$$0 = \nabla f(x^{k+1})^T \nabla f(x^k) \to \nabla f(\bar{x})^T \nabla f(\bar{x}) = \|\nabla f(\bar{x})\|_2^2 \quad \text{as} \quad k \to +\infty.$$

Therefore, $\|\nabla f(\bar{x})\|_2 = 0$, or equivalently $\nabla f(\bar{x}) = 0$. □

The above convergence result is not very satisfactory since there is no guarantee that the whole sequence $\{x^k\}$ converges. Actually, it is possible to prove also that each cluster point of the sequence $\{x^k\}$ is a stationary point of f.

The exact line search requires the solution of an additional optimization problem though in a single variable. Actually, if the objective function is the convex quadratic function $f(x) = \frac{1}{2}x^TQx + b^Tx + c$, then the stepsize can be computed explicitly. In fact, the derivative of the search function reads

$$\begin{aligned}\varphi'(t) &= -\nabla f(x - t\nabla f(x))^T\nabla f(x)\\ &= -[Q(x - t\nabla f(x)) + b]^T\nabla f(x)\\ &= -[Qx + b - tQ\nabla f(x)]^T\nabla f(x)\\ &= -[\nabla f(x) - tQ\nabla f(x)]^T\nabla f(x)\\ &= -\nabla f(x)^T\nabla f(x) + t(\nabla f(x)^TQ\nabla f(x)).\end{aligned}$$

If $\nabla f(x)^TQ\nabla f(x) = 0$, then $\varphi'(t) = -\|\nabla f(x)\|_2^2 < 0$ for all $t \in \mathbb{R}$ and therefore $f(x - t\nabla f(x)) = \varphi(t) = -\|\nabla f(x)\|_2^2 t + f(x) \to -\infty$ as $t \to +\infty$. On the other hand, if $\nabla f(x)^TQ\nabla f(x) > 0$, then the exact line search amounts to computing t such that $\varphi'(t) = 0$, that is $t = \nabla f(x)^T\nabla f(x)/(\nabla f(x)^TQ\nabla f(x))$.

If the above quadratic function is strictly convex, stepsizes related to the eigenvalues of Q lead to a finite gradient method.

Theorem 4.2. *Let $f(x) = \frac{1}{2}x^TQx + b^Tx + c$ be strictly convex, and $\lambda_0, \ldots, \lambda_{n-1} > 0$ be the eigenvalues of Q. Given any $x^0 \in \mathbb{R}^n$ and the finite sequence*

$$x^{k+1} = x^k - \lambda_k^{-1}\nabla f(x^k), \qquad k = 0, \ldots, n-1,$$

there exists $j \in \{0, \ldots, n\}$ such that $\nabla f(x^j) = 0$.

Proof. Suppose $\nabla f(x^j) \neq 0$ for all $j < n$. Therefore,

$$\begin{aligned}\nabla f(x^n) &= Qx^n + b\\ &= Qx^{n-1} - \lambda_{n-1}^{-1}Q\nabla f(x^{n-1}) + b\\ &= \nabla f(x^{n-1}) - \lambda_{n-1}^{-1}Q\nabla f(x^{n-1})\\ &= (I - \lambda_{n-1}^{-1}Q)\nabla f(x^{n-1})\\ &= (I - \lambda_{n-2}^{-1}Q)(I - \lambda_{n-1}^{-1}Q)\nabla f(x^{n-2})\\ &\vdots\\ &= \prod_{j=1}^{n}(I - \lambda_{n-j}^{-1}Q)\nabla f(x^0).\end{aligned}$$

Since Q is positive definite, there exists an orthonormal basis $\{u_0, \ldots, u_{n-1}\}$ of $\mathbb{R}^n$ such that $Qu_i = \lambda_i u_i$ for all $i = 0, \ldots, n-1$. Therefore, there exist $\alpha_0, \ldots, \alpha_{n-1} \in \mathbb{R}$

such that $\nabla f(x^0) = \alpha_0 u_0 + \cdots + \alpha_{n-1} u_{n-1}$. As a consequence,

$$\nabla f(x^n) = \Big(\prod_{j=1}^{n} (I - \lambda_{n-j}^{-1} Q) \Big) \sum_{i=0}^{n-1} \alpha_i u_i = \sum_{i=0}^{n-1} \alpha_i \Big(\prod_{j=1}^{n} (1 - \lambda_{n-j}^{-1} \lambda_i) \Big) u_i = 0$$

as the coefficient of each u_i is zero (just consider $j = n - i$). □

4.1.2 Gradient methods with inexact line search

Theorem 4.3. *Suppose f is continuously differentiable (on $\mathbb{R}^n$) and the gradient mapping ∇f is Lipschitz with modulus $L > 0$. Then, any cluster point of the sequence provided by the iterative scheme $x^{k+1} = x^k - \alpha \nabla f(x^k)$ for some given positive $\alpha < 2/L$ is a stationary point of f.*

Proof. Theorem 1.6 guarantess

$$\begin{aligned} f(x^{k+1}) = f(x^k - \alpha \nabla f(x^k)) &\leq f(x^k) - \alpha \nabla f(x^k)^T \nabla f(x^k) + L\alpha^2 \|\nabla f(x^k)\|_2^2/2 \\ &= f(x^k) - \gamma \|\nabla f(x^k)\|_2^2 \end{aligned}$$

where $\gamma = \alpha(2 - L\alpha)/2 > 0$. As a consequence, $f(x^{k+1}) < f(x^k)$. Given any cluster point $\bar{x} \in \mathbb{R}^n$ of $\{x^k\}_{k \in \mathbb{N}}$, there exists a subsequence $\{x^{k_j}\}_{j \in \mathbb{N}}$ such that $x^{k_j} \to \bar{x}$ as $j \to +\infty$. Therefore, the above inequalities imply

$$f(x^{k_{j+1}}) \leq f(x^{k_j+1}) \leq f(x^{k_j}) - \gamma \|\nabla f(x^{k_j})\|_2^2$$

Taking the limit as $j \to +\infty$ yields $\nabla f(x^{k_j}) \leq 0$, that is $\nabla f(\bar{x}) = 0$. □

Given a descent direction d^k for f at x^k, consider the sufficient decrease condition

$$f(x^k + t d^k) \leq f(x^k) + c_1 t \nabla f(x^k)^T d^k \qquad (AJO)$$

where $c_1 \in]0, 1[$. If f is bounded by below, then there exists $\tau > 0$ such that any $t > \tau$ does not satisfy (AJO). In fact, $\nabla f(x^k)^T d^k < 0$ implies $t \nabla f(x^k)^T d^k \to -\infty$ as $t \to +\infty$. In terms of the search function $\varphi_k(t) = f(x^k + t d^k)$, the condition reads

$$\varphi_k(t) \leq \varphi_k(0) + c_1 t \varphi_k'(0). \qquad (AJO)$$

As $\lim_{t \to 0} [\varphi_k(t) - \varphi_k(0)]/t = \varphi_k'(0) < c_1 \varphi_k'(0)$, then (AJO) holds whenever t is small enough. Therefore, a way to compute a stepsize t_k satisfying (AJO) is the so-called *Armijo rule*: given $\bar{t} > 0$ and $\gamma \in]0, 1[$, take $t_k = \bar{t}\gamma^m$ where $m \in \mathbb{N}$ is the smallest natural number such that $\bar{t}\gamma^m$ satisfies (AJO).

Theorem 4.4. *Suppose that Algorithm 2 generates an infinite sequence $\{x^k\}$. If f is bounded by below, then each cluster point of $\{x^k\}$ is a stationary point of f.*

Algorithm 2 – Gradient method with Armijo line search

0. Choose $x^0 \in \mathbb{R}^n$, $\bar{t} > 0$ and $\gamma \in]0,1[$, and set $k = 0$
1. If $\nabla f(x^k) = 0$, then *STOP*
2. Choose $d^k = -\nabla f(x^k)$ and compute $t_k > 0$ through the Armijo rule
3. $x^{k+1} = x^k - t_k \nabla f(x^k)$
4. $k = k + 1$ and go to 1

Proof. $d^k = -\nabla f(x^k)$ implies that (AJO) reads

$$0 \leq c_1 t_k \|\nabla f(x^k)\|_2^2 \leq f(x^k) - f(x^{k+1}),$$

and thus the sequence $\{f(x^k)\}$ is monotone decreasing. Since it is also bounded by below, then it has a limit. As a consequence, $f(x^k) - f(x^{k+1}) \to 0$: either $t_k \to 0$ or $\|\nabla f(x^k)\|_2 \to 0$ holds.

Given any cluster point $\bar{x} \in \mathbb{R}^n$ of $\{x^k\}_{k\in\mathbb{N}}$, there exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $x^{k_j} \to \bar{x}$ as $j \to +\infty$. If $\|\nabla f(x^k)\|_2 \to 0$, then $\|\nabla f(\bar{x})\|_2 = 0$, i.e., $\bar{x}$ is a stationary point for f, since $\|\nabla f(x^{k_j})\|_2 \to \|\nabla f(\bar{x})\|_2$. Therefore, suppose $t_k \to 0$ holds. The Armijo rule guarantess that $t_{k_j}\gamma^1$ does not satisfy (AJO), i.e.,

$$f(x^{k_j} - t_{k_j}\gamma^{-1}\nabla f(x^{k_j})) - f(x^{k_j}) > c_1 t_{k_j}\gamma^{-1}\|\nabla f(x^{k_j})\|_2^2.$$

The mean value Theorem 1.5 guarantees the existence of some $\tau_{k_j} \in [0, t_{k_j}\gamma^{-1}]$ such that $f(x^{k_j} - t_{k_j}\gamma^{-1}\nabla f(x^{k_j})) - f(x^{k_j}) = -t_{k_j}\gamma^{-1}\nabla f(x^{k_j} - \tau_{k_j}\nabla f(x_{k_j}))^T \nabla f(x^{k_j})$ yielding

$$\nabla f(x^{k_j} - \tau_{k_j}\nabla f(x_{k_j}))^T \nabla f(x^{k_j}) > c_1 \|\nabla f(x^{k_j})\|_2^2.$$

Taking the limit as $j \to +\infty$, $(1 - c_1)\|\nabla f(\bar{x})\|_2^2$ follows and hence $\nabla f(\bar{x}) = 0$. □

still an uncomplete draft

$$\nabla f(x^k + td^k)^T d^k \geq c_2 \nabla f(x^k)^T d^k \qquad (CUR)$$

$$\varphi_k'(t) \geq c_2 \varphi_k'(0) \qquad (CUR)$$

Proposition 4.4. *Suppose f is bounded by below. If $x^k \in \mathbb{R}^n$ is not a stationary point of f and $d^k \in \mathbb{R}^n$ is a descent direction for f at x^k, then there exist $\tau_\ell, \tau_u \in \mathbb{R}$ with $\tau_\ell < \tau_u$ such that any $t \in [\tau_\ell, \tau_u]$ satisfies the Wolfe conditions (AJO) and (CUR).*

Proof. The value

$$\tau_u = \sup\{\tau \ : \ (AJO) \text{ is satisfied by any } t \in [0, \tau]\}$$

is positive and finite. Moreover, it satisfies $\varphi_k(\tau_u) = \varphi_k(0) + c_1\tau_u\varphi_k'(0)$: otherwise, by continuity (AJO) would be satisfied by any $t \in [\tau_u, \tau_u + \varepsilon]$ for some $\varepsilon > 0$. Since τ_u is the supremum of a set of real numbers, there exists a sequence $\{t_j\}_{j\in\mathbb{N}}$ such that $t_j > \tau_u$, $t_j \to \tau_u$ as $j \to +\infty$ and (AJO) is not satisfied at t_j, that is

$$\varphi_k(t_j) > \varphi_k(0) + c_1 t_j \varphi_k'(0)$$

or equivalently $\varphi_k(t_j) - \varphi_k(\tau_u) > c_1(t_j - \tau_u)\varphi_k'(0)$. Therefore, dividing both sides by $(t_j - \tau_u)$ and taking the limit as $j \to +\infty$ (which means $t_j \to \tau_u$) leads to $\varphi_k'(\tau_u) \geq c_1\varphi_k'(0)$. Since $c_2 > c_1$ and $\varphi_k'(0) < 0$, $\varphi_k'(\tau_u) > c_2\varphi_k'(0)$ holds and the continuity of φ_k' (f is continuously differentiable) implies that there exists $\delta > 0$ such that $\varphi_k'(t) \geq c_2\varphi_k'(0)$, i.e., (AJO) holds for any $t \in [\tau_u - \delta, \tau_u + \delta]$. Therefore, the thesis follows just taking $\tau_\ell = \tau_u - \delta$. □

Algorithm 3 – Gradient type method with Wolfe line search

0. Choose $x^0 \in \mathbb{R}^n$ and set $k = 0$
1. If $\nabla f(x^k) = 0$, then *STOP*
2. Choose $d^k \in \mathbb{R}^n$ such that $\nabla f(x^k)^T d^k < 0$
3. Compute $t_k > 0$ satisfying the Wolfe conditions (AJO) and (CUR)
4. $x^{k+1} = x^k + t_k d^k$
5. $k = k + 1$ and go to 1

Theorem 4.5. *Suppose that Algorithm 3 generates an infinite sequence $\{x^k\}$. If f is bounded by below and the angle θ_k formed by $\nabla f(x^k)$ and d^k satisfies $\theta_k \geq \pi/2 + \bar{\theta}$ for some fixed $\bar{\theta} \in]0, \pi/2[$ for all iterations $k \in \mathbb{N}$, then each cluster point of $\{x^k\}$ is a stationary point of f.*

Proof. Since d^k is a descent direction for f at x^k and t_k satisfies (AJO), then

$$0 \leq -c_1 t_k \nabla f(x^k)^T d^k = -c_1 t_k \|\nabla f(x^k)\|_2 \|d^k\|_2 \cos\theta_k \leq f(x^k) - f(x^{k+1}).$$

The sequence $\{f(x^k)\}$ is monotone decreasing and it is bounded by below (since f is such), thus it has a limit. As a consequence, $f(x^k) - f(x^{k+1}) \to 0$, which implies $t_k\|\nabla f(x^k)\|_2\|d^k\|_2 \cos\theta_k \to 0$. Since $\cos\theta_k \leq \cos(\pi/2 + \bar{\theta}) = -\sin\bar{\theta} < 0$, then either $t_k\|d^k\|_2 \to 0$ or $\|\nabla f(x^k)\|_2 \to 0$ holds.

Given any cluster point $\bar{x} \in \mathbb{R}^n$ of $\{x^k\}_{k\in\mathbb{N}}$, there exists a subsequence $\{x^{k_j}\}_{j\in\mathbb{N}}$ such that $x^{k_j} \to \bar{x}$ as $j \to +\infty$. If $\|\nabla f(x^k)\|_2 \to 0$, then $\|\nabla f(\bar{x})\|_2 = 0$, i.e., $\bar{x}$ is

a stationary point of f. Therefore, suppose $t_k\|d^k\|_2 \to 0$ holds. Since t_{k_j} satisfies (CUR), then $\hat{d}^{k_j} = d^{k_j}/\|d^{k_j}\|_2$ satisfies

$$\nabla f(x^{k_j} + t_{k_j} d^{k_j})^T \hat{d}^{k_j} \geq c_2 \nabla f(x^{k_j})^T \hat{d}^{k_j}.$$

By construction $\hat{d}^{k_j} \in \overline{B(0,1)}$, and thus $\hat{d}^{k_j} \to \bar{d}$ for some $\bar{d} \in \overline{B(0,1)}$ (eventually taking a further subsequence). Moreover, $x^{k_j} + t_{k_j} d^{k_j} \to \bar{x}$, and thus taking the limit as $j \to +\infty$ in both sides of the above inequality leads to

$$\nabla f(\bar{x})^T \bar{d} \geq c_2 \nabla f(\bar{x})^T \bar{d},$$

which reads also $\nabla f(\bar{x})^T \bar{d} \geq 0$ since $c_2 > 0$. On the other hand, $\nabla f(x^{k_j})^T \hat{d}^{k_j} < 0$ holds for all j, so that it must necessarily be $\nabla f(\bar{x})^T \bar{d} = 0$. Finally,

$$\sin\bar{\theta}\ \|\nabla f(x^{k_j})\|_2 \leq -\cos\theta_{k_j} \|\nabla f(x^{k_j})\|_2 = \nabla f(x^{k_j})^T \hat{d}^{k_j} \to 0$$

guarantees $\|\nabla f(\bar{x})\|_2 = 0$. □

4.2 Conjugate gradient methods

This family of methods provides a concrete alternative to choosing the steepest descent direction by keeping track of the directions that have been exploited in the previous iterations.

4.2.1 The linear case

The linear conjugate gradient method was originally designed to solve the linear system $Ax = b$, where $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is positive definite, through the minimization of the strictly convex quadratic function $f(x) = \frac{1}{2}x^T Ax - b^T x$.

Algorithm 4 – Linear conjugate gradient method

0. Choose $x^0 \in \mathbb{R}^n$ and set $k = 0$
1. If $r^k = b - Ax^k = 0$, then *STOP*
2. $\beta_k = {r^k}^T r^k / {r^{k-1}}^T r^{k-1}$ if $k \geq 1$
3. $d^k = r^k + \beta_k d^{k-1}$ if $k \geq 1$, otherwise $d^0 = r^0$
4. Compute $t_k = {r^k}^T r^k / {d^k}^T A d^k$
5. $x^{k+1} = x^k + t_k d^k$
6. $k = k + 1$ and go to 1

Since $r^k = -\nabla f(x^k)$, the first iteration is the same of the gradient method with exact line search, and afterwards the search direction is modified in such a way that convergence can be achieved in a finite number of iterations.

Proposition 4.5. *Suppose there exists $\bar{k} \in \mathbb{N}$ such that Algorithm 4 generates a sequence $\{r^k\}$ with $r^k \neq 0$ for any $k < \bar{k}$. Then, the relationships*

(i) ${r^k}^T r^j = 0$

(ii) ${d^k}^T A d^j = 0$

(iii) ${r^k}^T d^j = 0$

(iv) ${d^k}^T r^0 = {r^k}^T r^k$

hold for any $k \leq \bar{k}$ and any $j < k$.

Condition (iii) guarantees that Algorithm 4 is a descent method:

$$\nabla f(x^k)^T d^k = -{r^k}^T d^k = -{r^k}^T r^k - \beta_k {r^k}^T d^{k-1} = -{r^k}^T r^k = -\|r^k\|_2^2 < 0.$$

Step 4 of the algorithm identifies the stepsize which minimizes the search function $\varphi_k(t) = f(x^k + t d^k)$ since $t_k > 0$ and

$$\begin{aligned}\varphi_k'(t_k) &= \nabla f(x^k + t_k d^k)^T d^k = (A x^k + t_k A d^k - b)^T d^k = (t_k A d^k - r^k)^T d^k \\ &= t_k {d^k}^T A d^k - {r^k}^T (r^k + \beta_k d^{k-1}) = t_k {d^k}^T A d^k - {r^k}^T r^k = 0.\end{aligned}$$

Condition (i) guarantees that the algorithm stops after at most n iterations: if $r^k \neq 0$ for any $k = 0, \ldots, n-1$, then $r^0, \ldots, r^n$ are linearly independent, which is impossible, unless $r^n = 0$. Furthermore, under the same assumption, condition (ii) implies that also $d^0, \ldots, d^k$ are linearly independent for any $k < n$. In fact, if $d^k = \gamma_0 d^0 + \cdots + \gamma_{k-1} d^{k-1}$ for some $\gamma_0, \ldots, \gamma_{k-1} \in \mathbb{R}$, then $d^k = 0$ since A is positive definite and ${d^k}^T A d^k = \gamma_0 {d^k}^T A d^0 + \cdots + \gamma_{k-1} {d^k}^T A d^{k-1} = 0$, thus $\gamma_0 = \cdots = \gamma_{k-1} = 0$ as $d^0, \ldots, d^{k-1}$ are linearly independent by inductive hypothesis. This further property of linear independence allows proving that the finite sequence $\{x^k\}$ is composed by minimum points of f over nested affine subspaces that invade the whole $\mathbb{R}^n$.

Theorem 4.6. *Let $\{x^k\}$ be the sequence produced by Algorithm 4. Then,*

$$f(x^k) = \min\{f(x) \ : \ (x - x^0) \in S_k\}$$

with S_k denoting the vector subspace of $\mathbb{R}^n$ generated by $d^0, \ldots, d^k$.

Proof. Taking $\psi_k(\alpha_0, \ldots, \alpha_{k-1}) = f(x^0 + \alpha_0 d^0 + \cdots + \alpha_{k-1} d^{k-1})$, the minimization of f over the affine subspace $x^0 + S_k$ can be stated as the unconstrained problem

$$\min\{\psi_k(\alpha_0, \ldots, \alpha_{k-1}) \ : \ \alpha_0, \ldots, \alpha_{k-1} \in \mathbb{R}\}.$$

Moreover, ψ_k is a strictly convex quadratic function since f is quadratic and strictly convex. Therefore, the unique minimum point of the above problem is the unique solution $(\bar{\alpha}_0, \dots \bar{\alpha}_{k-1})$ of the linear system of equations $\nabla\psi_k(\alpha_0, \dots, \alpha_{k-1}) = 0$. Since both

$$0 = \frac{\partial\psi_k}{\partial\alpha_i}(\bar{\alpha}_0, \dots \bar{\alpha}_{k-1}) = \nabla f(x^0 + \bar{\alpha}_0 d^0 + \cdots + \bar{\alpha}_{k-1} d^{k-1})^T d^i$$

and $\nabla f(x^k)^T d^i = -{r^k}^T d^i = 0$ hold for any $i = 0, \dots, k-1$, the uniqueness of the solution implies $x^k = x^0 + \bar{\alpha}_0 d^0 + \cdots + \bar{\alpha}_{k-1} d^{k-1}$. □

Since $S_1 \subset S_2 \subset \cdots \subset S_n = \mathbb{R}^n$, finite convergence follows from Theorem 4.6 as well. An alternative proof of the theorem relies on the explicit expression

$$\psi_k(\alpha_0, \dots, \alpha_{k-1}) = f(x_0) + \sum_{i=0}^{k-1} [\frac{1}{2}({d^i}^T A d^i)\alpha_i^2 - {d^i}^T(b - Ax^0)\alpha_i]$$

since the partial derivative

$$\frac{\partial\psi_k}{\partial\alpha_i}(\alpha_0, \dots, \alpha_{k-1}) = ({d^i}^T A d^i)\alpha_i - {d^i}^T(b - Ax^0)$$

is zero if and only if $\alpha_i = {d^i}^T(b - Ax^0)/{d^i}^T A d^i = {d^i}^T r^0/{d^i}^T A d^i = {r^i}^T r^i/{d^i}^T A d^i = t_i$, and therefore $x^0 + t_0 d^0 + \cdots + t_{k-1} d^{k-1} = x^k$ minimizes f over $x^0 + S_k$.

4.2.2 The nonlinear case

The basic idea to adapt the conjugation approach to the minimization of general nonlinear functions is simply to replace r^k with $-\nabla f(x^k)$. Anyway, some troubles emerge: no formula for the exact line search is available, and in case an inexact search is performed there is no guarantee that $d^k = -\nabla f(x^k) + \beta_k d^{k-1}$ is a descent direction for f at x^k. In fact,

$$\nabla f(x^k)^T d^k = -\|\nabla f(x^k)\|_2^2 + \beta_k \nabla f(x^k)^T d^{k-1}$$

leads to $\nabla f(x^k)^T d^k \leq 0$ if $\nabla f(x^k)^T d^{k-1} \leq 0$, which is true when the exact line search is performed, while the Wolfe conditions are not enough to guarantee it. Actually, it is enough to replace (CUR) by the condition

$$|\nabla f(x^k + t d^k)^T d^k| \leq c_2 |\nabla f(x^k)^T d^k|, \qquad (StrCUR)$$

with $0 < c_1 < c_2 < 1/2$ where c_1 is the parameter chosen for (AJO), for d^k to be a descent direction within an inexact line search framework. Considering the search function $\varphi_k(t) = f(x^k + t d^k)$, $(StrCUR)$ can be equivalently stated as

$$|\varphi_k'(t)| \leq c_2 |\varphi_k'(0)|, \qquad (StrCUR)$$

which clearly implies (CUR) since $\varphi_k'(0) < 0$ and hence

$$\varphi_k'(t) \geq -|\varphi_k'(t)| \geq -c_2|\varphi_k'(0)| = c_2\varphi_k'(0).$$

(AJO) and $(StrCUR)$ are generally referred to as *the strong Wolfe conditions*. The existence of an interval of stepsizes that satisfy both of them can be proved in the same way of Proposition 4.4 if $\varphi_k'(\tau_u) \leq 0$, and exploiting in addition the continuity of φ_k' if $\varphi_k'(\tau_u) > 0$.

Proposition 4.6. *If f is bounded by below, then each direction d^k generated by Algorithm 5 satisfies*

$$-\|\nabla f(x^k)\|_2^2/(1-c_2) \leq \nabla f(x^k)^T d^k \leq [(2c_2-1)/(1-c_2)]\|\nabla f(x^k)\|_2^2.$$

Since any positive c_2 satisfying $c_2 < 1/2$ guarantees $[(2c_2-1)/(1-c_2)] < 0$, the above right inequality guarantees that d^k is a descent direction for f at x^k. Clearly, it is better not to choose c_2 too close to $1/2$[1].

Algorithm 5 – Nonlinear conjugate gradient method

0. Choose $x^0 \in \mathbb{R}^n$ and set $k = 0$
1. If $\nabla f(x^k) = 0$, then *STOP*
2. $\beta_k = \nabla f(x^k)^T \nabla f(x^k) / \nabla f(x^{k-1})^T \nabla f(x^{k-1})$ if $k \geq 1$
3. $d^k = -\nabla f(x^k) + \beta_k d^{k-1}$ if $k \geq 1$, otherwise $d^0 = -\nabla f(x^0)$
4. Compute t_k satisfying the strong Wolfe conditions (AJO) and $(StrCUR)$
5. $x^{k+1} = x^k + t_k d^k$
6. $k = k + 1$ and go to 1

Theorem 4.7. *Suppose that Algorithm 5 generates an infinite sequence $\{x^k\}$. If f is bounded by below and the gradient mapping ∇f is Lipschitz, i.e., there exists $L > 0$ such that*

$$\forall\ x, y \in \mathbb{R}^n\ :\ \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2,$$

then there exists a subsequence $\{x^{k_j}\}$ such that $\lim_{j\to+\infty} \|\nabla f(x^{k_j})\|_2 = 0$.

Corollary 4.1. *Suppose that Algorithm 5 generates an infinite sequence $\{x^k\}$. If f is bounded by below, ∇f is a Lipschitz mapping and the sublevel set*

$$L_f(x^0) = \{x \in \mathbb{R}^n\ :\ f(x) \leq f(x^0)\}$$

is compact, then at least one cluster point of $\{x^k\}$ is a stationary point of f.

[1] $\ell(c) = (2c-1)/(1-c)$ is a monotone increasing function with $\ell(0) = -1$ and $\ell(1/2) = 0$

While in gradient methods with $d^k = -\nabla f(x^k)$ the angle θ_k between d^k and $\nabla f(x^k)$ is always π, in conjugate gradient methods there is no guarantee that it stays bounded away from $\pi/2$. If θ_k gets too close to $\pi/2$, the algorithm may slow down meaningfully. In fact, $\theta_k \approx \pi/2$ implies

$$0 \approx -\cos\theta_k = -\nabla f(x^k)^T d^k / [\|\nabla f(x^k)\|_2 \|d^k\|_2] \geq [(1-2c_2)/(1-c_2)] \|\nabla f(x^k)\|_2 / \|d^k\|_2$$

where the inequality is due to Proposition 4.6. Therefore, it is likely to have $\|\nabla f(x^k)\|_2 << \|d^k\|_2$ and also $t_k \approx 0$ since d^k is almost orthogonal to the steepest descent direction. If $t_k \approx 0$, then $x^{k+1} \approx x^k$ and thus $\nabla f(x^{k+1}) \approx \nabla f(x^k)$ are also probable. In such a case $\beta_{k+1} \approx 1$ and $\|\nabla f(x^{k+1})\|_2 \approx \|\nabla f(x^k)\|_2 << \|d^k\|_2$ lead to

$$d^{k+1} = -\nabla f(x^{k+1}) + \beta_{k+1} d^k \approx -\nabla f(x^{k+1}) + d^k \approx d^k$$

that means $\theta_{k+1} \approx \theta_k$, so that the new iteration will be similar to the previous. Therefore, if $\cos\theta_k \approx 0$, then it is possible that the algorithm will perform a long sequence of almost useless iterations.

The so-called restart technique tries to overcome this issue by performing a steepest descent step after a certain number of iterations, that is setting $\beta_k = 0$ every $\bar{n}$ iterations. The algorithm performs a restart in the sense the effect of the previous directions on the current one is cancelled. It is also possible to prove that the subsequence of the restart iterates x^{k_j} satisfies the convergence property of Theorem 4.7.

Relying on the alternative formula $\beta_k = {r^k}^T (r^k - r^{k-1}) / {r^{k-1}}^T r^{k-1}$ of the linear case, the Polak-Ribiere variant of the method applies the restart technique approximately by choosing $\beta_k = \beta_k^{PR}$ for

$$\beta_k^{PR} = \nabla f(x^k)^T (\nabla f(x^k) - \nabla f(x^{k-1})) / \nabla f(x^{k-1})^T \nabla f(x^{k-1})$$

as $\nabla f(x^k) \approx \nabla f(x^{k-1})$ guarantees $\beta_k^{PR} \approx 0$. Since $\beta_k^{PR} < 0$ may occur, another variant of the method exploits $\beta_k^{PR+} = \max\{\beta_k^{PR}, 0\}$.