Lectures on optimization
Advanced camp

Francisco J. Silva

Université de Limoges

francisco.silva@unilim.fr

http://www.unilim.fr/pages_perso/francisco.silva/

August 2019

Contents of the course

Definition of an optimization problem

◇ An optimization problem has the form

$$\text{Find } \bar{x} \in \mathbb{R}^n \text{ such that} \quad f(\bar{x}) = \min \{ f(x) \mid x \in \mathcal{K} \}, \qquad (P)$$

where $\mathcal{K} \subseteq \mathbb{R}^n$ is a given set. By definition, this mean to find $\bar{x} \in \mathcal{K}$ such that

$$f(\bar{x}) \leq f(x) \ \forall \ x \in \mathcal{K}.$$

◇ In the above, f is called an objective function, $\mathcal{K}$ is called a feasible set (or constraint set) and any $\bar{x}$ solving (P) is called a global solution to problem (P).

◇ Usually one also considers the weaker notion, but easier to characterize, of local solution to problem (P). Namely, $\bar{x} \in \mathcal{K}$ is a local solution to (P) if there exists $\delta > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in \mathcal{K} \cap B(\bar{x}, \delta)$, where

$$B(\bar{x}, \delta) := \{ x \in \mathbb{R}^n \mid \|x - \bar{x}\| \leq \delta \}.$$

◇ In optimization theory one usually studies the following features of problem (P):

1.- Does there exist a solution $\bar{x}$ (global or local)?

2.- Optimality conditions, i.e. properties satisfied by the solutions (or local solutions).

3.- Computation algorithms for finding approximate solutions.

◇ In this course we will mainly focus on points 1 and 2 of the previous program.

◇ We will also consider mainly two cases for the feasible set $\mathcal{K}$:

- ◇ $\mathcal{K} = \mathbb{R}^n$ (unconstrained case).
- ◇ Equality and inequality constraints:

$$\mathcal{K} = \{x \in \mathbb{R}^n \mid g_i(x) = 0, \ i = 1, \dots, m, \ h_j(x) \leq 0, \ j = 1, \dots, \ell\}. \quad (1)$$

◇ In order to tackle point 2 we will assume that f is a smooth function. If the feasible set (1) is considered, we will also assume that g_i and h_j are smooth functions.

Some mathematical tools

◇ In what follows, we will work in the euclidean space $\mathbb{R}^n$. We denote by $\langle \cdot, \cdot \rangle$ the standard scalar product and by $\| \cdot \|$ the corresponding norm. Namely,

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i \quad \forall\, x \in \mathbb{R}^n,\ y \in \mathbb{R}^n \quad \text{and} \quad \|x\| = \sqrt{\langle x, x \rangle}.$$

We will often use the alternative notation $x \cdot y$ for $\langle x, y \rangle$.

◇ [Graph of a function] Let $f : \mathbb{R}^n \to \mathbb{R}$. The graph $\mathsf{Gr}(f) \subseteq \mathbb{R}^{n+1}$ is defined by

$$\mathsf{Gr}(f) := \{(x, f(x)) \mid x \in \mathbb{R}^n\}.$$

◇ [Level sets] Let $c \in \mathbb{R}$. The level set of value c is defined by

$$\mathrm{Lev}_f(c) := \{x \in \mathbb{R}^n \mid f(x) = c\} \, .$$

- When $n = 2$, the sets $\mathrm{Lev}_f(c)$ are useful in order to draw the graph of a function.
- These sets will also be useful in order to solve graphically two dimensional linear programming problems, i.e. $n = 2$, and the function f and the set $\mathcal{K}$ are defined by means of affine functions.

Example 1: We consider the function

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) := x + y + 2 \in \mathbb{R},$$

whose level set is given, for all $c \in \mathbb{R}$, by

$$\mathsf{Lev}_f(c) := \left\{ (x, y) \in \mathbb{R}^2 \mid x + y + 2 = c \right\}.$$

Note that the optimization problem with this f and $\mathcal{K} = \mathbb{R}^2$ does not have a solution.

Example 2: Consider the function

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) := x^2 + y^2 \in \mathbb{R}.$$

Then $\mathsf{Lev}_f(c) = \emptyset$ if $c < 0$ and, if $c \geq 0$,

$$\mathsf{Lev}_f(c) = \{(x, y) \mid x^2 + y^2 = c\},$$

i.e. the circle centered at 0 and of radius $\sqrt{c}$.

Example 3: Consider the function

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) := x^2 - y^2 \in \mathbb{R}.$$

In this case the level sets are given, for all $c \in \mathbb{R}$, by

$$\mathsf{Lev}_f(c) = \{(x, y) \mid y^2 = x^2 - c\}.$$

$\diamond$ [Differentiability] Let $f : \mathbb{R}^n \to \mathbb{R}$. We say that f is differentiable at $\bar{x} \in \mathbb{R}^n$ if for all $i = 1, \ldots, n$ the partial derivatives

$$\frac{\partial f}{\partial x_i}(\bar{x}) := \lim_{\tau \to 0} \frac{f(\bar{x} + \tau \mathbf{e}_i) - f(\bar{x})}{\tau} \quad \text{(where } \mathbf{e}_i := (0, \ldots, \underbrace{1}_{i}, \ldots, 0)\text{)},$$

exist and, defining the gradient of f at $\bar{x}$ by

$$\nabla f(\bar{x}) := \left(\frac{\partial f}{\partial x_1}(\bar{x}), \ldots, \frac{\partial f}{\partial x_n}(\bar{x}) \right) \in \mathbb{R}^n,$$

we have that

$$\lim_{h \to 0} \frac{f(\bar{x} + h) - f(\bar{x}) - \nabla f(\bar{x}) \cdot h}{\|h\|} = 0.$$

If f is differentiable at every x belonging to a set $A \subseteq \mathbb{R}^n$, we say that f is differentiable in A.

Remark 1. Notice that f is differentiable at $\bar{x}$ iff there exists $\varepsilon_{\bar{x}} : \mathbb{R}^n \to \mathbb{R}$, with $\lim_{h\to 0} \varepsilon_{\bar{x}}(h) = 0$ and

$$f(\bar{x}+h) = f(\bar{x}) + \nabla f(\bar{x}) \cdot h + \|h\| \varepsilon_{\bar{x}}(h). \tag{2}$$

In particular, f is continuous at $\bar{x}$.

Lemma 1. [Directional differentiability] *Assume that f is differentiable at $\bar{x}$ and let $h \in \mathbb{R}^n$. Then,*

$$\lim_{\tau\to 0,\, \tau>0} \frac{f(\bar{x}+\tau h) - f(\bar{x})}{\tau} = \nabla f(\bar{x}) \cdot h.$$

Proof. By (2), for every $\tau > 0$, we have

$$f(\bar{x}+\tau h) - f(\bar{x}) = \tau \nabla f(\bar{x}) \cdot h + \tau \|h\| \varepsilon_{\bar{x}}(\tau h).$$

Dividing by τ and letting $\tau \to 0$ gives the result. □

Remark 2. (i) [Simple criterion to check differentiability] Suppose that $A \subseteq \mathbb{R}^n$ is an open set containing $\bar{x}$ and that

$$A \ni x \mapsto \nabla f(x) \in \mathbb{R}^n,$$

is well-defined and continuous at $\bar{x}$. Then, f is differentiable at $\bar{x}$.

As a consequence, if ∇f is continuous in A, then f is differentiable in A. In this case, we say that f is $\mathcal{C}^1$ in A (i.e. differentiability and continuity of ∇f in A). When f is $\mathcal{C}^1$ in $\mathbb{R}^n$ we simply say that f is $\mathcal{C}^1$.

(ii) The notion of differentiability can be extended to a function $f : \mathbb{R}^n \to \mathbb{R}^m$. In this case, f is differentiable at $\bar{x}$ if there exists $L \in \mathcal{M}_{m\times n}(\mathbb{R})$ such that

$$\lim_{\|h\|\to 0} \frac{\|f(\bar{x}+h) - f(\bar{x}) - Lh\|}{\|h\|} \to 0.$$

If f is differentiable at $\bar{x}$, then $L = Df(\bar{x})$, called the Jacobian matrix of f at $\bar{x}$, which is given by

$$Df(\bar{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\bar{x}) \\ \dots & \dots & \dots \\ \frac{\partial f_i}{\partial x_1}(\bar{x}) & \dots & \frac{\partial f_i}{\partial x_n}(\bar{x}) \\ \dots & \dots & \dots \\ \frac{\partial f_m}{\partial x_1}(\bar{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\bar{x}) \end{pmatrix}$$

Note that when $m = 1$ we have that $Df(\bar{x}) = \nabla f(\bar{x})^\top$.

The chain rule says that if $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{x}$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $f(\bar{x})$, then $g \circ f : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\bar{x}$ and the following identity holds

$$D(g \circ f)(\bar{x}) = Dg(f(\bar{x}))Df(\bar{x}).$$

(iii) In the previous definitions the fact that the domain of definition of f is $\mathbb{R}^n$ is not important. The definition can be extended naturally for functions defined on open subsets of $\mathbb{R}^n$.

Basic examples:

(i) Let $c \in \mathbb{R}^n$ and consider the function $f_1 : \mathbb{R}^n \to \mathbb{R}$ defined by $f_1(x) = c \cdot x$. Then, for every $x \in \mathbb{R}^n$, we have $\nabla f_1(x) = c$ and, hence, f is differentiable.

(ii) Let $Q \in M_{n\times n}(\mathbb{R})$ and consider the function $f_2 : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f_2(x) = \tfrac{1}{2}\langle Qx, x\rangle \quad \forall\ x \in \mathbb{R}^n.$$

Then, for all $x \in \mathbb{R}^n$ and $h \in \mathbb{R}^n$, we have

$$\begin{aligned}
f_2(x+h) &= \tfrac{1}{2}\langle Q(x+h), x+h\rangle \\
&= \tfrac{1}{2}\langle Qx, x\rangle + \tfrac{1}{2}\left[\langle Qx, h\rangle + \langle Qh, x\rangle\right] + \tfrac{1}{2}\langle Qh, h\rangle \\
&= \tfrac{1}{2}\langle Qx, x\rangle + \langle \tfrac{1}{2}\left(Q + Q^\top\right)x, h\rangle + \tfrac{1}{2}\langle Qh, h\rangle \\
&= f_2(\bar{x}) + \langle \tfrac{1}{2}\left(Q + Q^\top\right)x, h\rangle + \|h\|\varepsilon_x(h),
\end{aligned}$$

where $\lim_{h\to 0}\varepsilon_x(h)=0$. Therefore, f_2 is differentiable and

$$\nabla f_2(x)=\tfrac{1}{2}\left(Q+Q^\top\right)x \quad \forall\, x\in\mathbb{R}^n.$$

In particular, if Q is symmetric, then $\nabla f_2(x)=Qx$.

(iii) Consider the function $f_3:\mathbb{R}^n\to\mathbb{R}$ defined by $f_3(x)=\|x\|$. Then, since $f_3(x)=\sqrt{\|x\|^2}$, if $x\neq 0$, the chain rule shows that

$$Df(x)=D(\sqrt{\cdot})(\|x\|^2)D(\|\cdot\|^2)(x)=\tfrac{1}{2}\frac{1}{\sqrt{\|x\|^2}}(2x)^\top=\frac{x^\top}{\|x\|},$$

which implies that $\nabla f_3(x)=\frac{x}{\|x\|}$, and, since this function is continuous at every $x\neq 0$, we have that f_3 is $\mathcal{C}^1$ in the set $\mathbb{R}^n\setminus\{0\}$. Let us show that f_3 is not differentiable at $x=0$. Indeed, if this is not the case, then all the partial derivatives $\frac{\partial f_3}{\partial x_i}(0)$ should exists for all $i=1,\dots,n$. Taking, for instance, $i=1$, we have

$$\lim_{\tau\to 0}\frac{\|0+\tau\mathbf{e_1}\|-\|0\|}{\tau}=\lim_{\tau\to 0}\frac{|\tau|}{\tau},$$

which does not exist, because

$$\lim_{\tau\to 0^-}\frac{|\tau|}{\tau}=\lim_{\tau\to 0^-}\frac{-\tau}{\tau}=-1\neq 1=\lim_{\tau\to 0^+}\frac{\tau}{\tau}=\lim_{\tau\to 0^+}\frac{|\tau|}{\tau}.$$

◇ [Second order derivative and Taylor expansion] Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is $\mathcal{C}^1$. In particular, the function $\mathbb{R}^n\ni x\mapsto\nabla f(x)\in\mathbb{R}^n$ is well defined. If this function is differentiable at $\bar{x}$, then we say that f is twice differentiable at $\bar{x}$. If f is twice differentiable at every x belonging to a set $A\subseteq\mathbb{R}^n$, then we say that f is twice differentiable in A.

If this is the case, then, by definition,

$$\frac{\partial^2 f}{\partial x_i\partial x_j}(\bar{x}):=\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right)(\bar{x}),$$

exists for all i, $j=1,\ldots,n$. The following result, due to Clairaut and also known as Schwarz's theorem, says that, under appropriate conditions we can change the derivation order.

Theorem 1. *Suppose that the function f is twice differentiable in an open set $A \subseteq \mathbb{R}^n$ containing $\bar{x}$ and that for all i, $j = 1, \dots, n$ the function $A \ni x \mapsto \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \in \mathbb{R}$ is continuous at $\bar{x}$. Then,*

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

Under the assumptions of the previous theorem, the Jacobian of $\nabla f(\bar{x})$ takes the form

$$D^2 f(\bar{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(\bar{x}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(\bar{x}) \\ \vdots & \dots & \vdots \\ \frac{\partial^2 f}{\partial x_i \partial x_1}(\bar{x}) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_i}(\bar{x}) \\ \vdots & \dots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\bar{x}) & \dots & \frac{\partial^2 f}{\partial x_n^2}(\bar{x}) \end{pmatrix}.$$

This matrix, called the Hessian matrix of f at $\bar{x}$ belongs to $\mathcal{M}_{n \times n}(\mathbb{R})$ and it is a symmetric matrix by the previous result.

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable in an open set $A \subseteq \mathbb{R}^n$ and for all i, $j = 1, \dots, n$ the function

$$A \ni x \mapsto \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \in \mathbb{R}$$

is continuous, we say that f is $\mathcal{C}^2$ in A.

◇ [Taylor's theorem] We admit the following important result:

Theorem 2. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$ in an open set $A \subseteq \mathbb{R}^n$. Then, for all $x \in A$ and h such that $x + h \in A$, we have the following expansion*

$$f(x+h) = f(x) + \nabla f(x) \cdot h + \tfrac{1}{2}\langle D^2 f(x)h, h\rangle + \|h\|^2 R_x(h),$$

where $R_x(h) \to 0$ as $h \to 0$.

Example: Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x,y) = e^x \cos(y) - x - 1$.

Then,

$$\frac{\partial f}{\partial x}(0,0) = e^x \cos(y)\big|_{(x,y)=(0,0)} - 1 = 0,$$
$$\frac{\partial f}{\partial y}(0,0) = -e^x \sin(y)\big|_{(x,y)=(0,0)} = 0,$$
$$\frac{\partial^2 f}{\partial x^2}(0,0) = e^x \cos(y)\big|_{(x,y)=(0,0)} = 1,$$
$$\frac{\partial^2 f}{\partial y^2}(0,0) = -e^x \cos(y)\big|_{(x,y)=(0,0)} = -1,$$
$$\frac{\partial^2 f}{\partial x \partial y}(0,0) = -e^x \sin(y)\big|_{(x,y)=(0,0)} = 0.$$

Note that all the first and second order partial derivatives are continuous in $\mathbb{R}^n$. Therefore, we can apply the previous result and obtain that the Taylor's expansion of f at $(0,0)$ is given by

$$\begin{aligned} f((0,0)+h) &= f(0,0) + \nabla f(0,0)\cdot h + \tfrac{1}{2}\langle D^2 f(0,0)h, h\rangle + \|h\|^2 R_{\bar{x}}(h), \\ &= 0 + 0 + \tfrac{1}{2}h_1^2 - \tfrac{1}{2}h_2^2 + \|h\|^2 R_{(0,0)}(h), \\ &= \tfrac{1}{2}h_1^2 - \tfrac{1}{2}h_2^2 + \|h\|^2 R_{(0,0)}(h). \end{aligned}$$

This expansion shows that locally around $(0,0)$ the function f above is similar to the function in Example 3.

Some good reading for the previous part

◇ Chapters 2 and 3 in “Vector calculus”, sixth edition, by J. E. Marsden and A. Tromba.

◇ Chapter 14 in “Calculus: Early transcendentals”, eight edition, by J. Stewart.

Some basic existence results for problem (P)

◇ [Compactness] Recall that $A \subseteq \mathbb{R}^n$ is called compact if A is closed and bounded (i.e. A is closed and there exists $R > 0$ such that $\|x\| \leq R$ for all $x \in A$).

Let us recall an important result concerning the compactness of a set A.

Theorem 3. [Bolzano-Weierstrass theorem] *A set $A \subseteq \mathbb{R}^n$ is compact if and only if every sequence $(x_k)_{k\in\mathbb{N}}$ of elements of A has a convergence subsequence. This means that there exists $\bar{x} \in A$ and a subsequence $(x_{k_\ell})_{\ell\in\mathbb{N}}$ of $(x_k)_{k\in\mathbb{N}}$ such that*

$$\bar{x} = \lim_{\ell\to\infty} x_{k_\ell}.$$

◇ [The basic existence results] Note that by definition, if $\inf_{x\in\mathcal{K}} f(x) = -\infty$, then f has no lower bounds in $\mathcal{K}$ and, hence, there are no solutions to (P). On the other hand, if $\inf_{x\in\mathcal{L}} f(x)$ is finite, then the existence of a solution can also fail to hold as the following example shows.

Example: Consider the function $\mathbb{R} \ni x \mapsto f(x) := e^{-x}$ and take $\mathcal{K} := [0, +\infty[$. Then, $\inf_{x \in \mathcal{K}} f(x) = 0$ and there is no $x \in \mathcal{K}$ such that $f(x) = 0$.

Definition 1. *We say that $(x_k)_{k \in \mathbb{N}} \subseteq \mathcal{K}$ is a minimizing sequence for (P) if*

$$\inf_{x \in \mathcal{K}} f(x) = \lim_{k \to \infty} f(x_k).$$

By definition, a minimizing sequence always exists if $\mathcal{K}$ is non-empty.

Theorem 4. [Weierstrass theorem, $\mathcal{K}$ compact] *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and that $\mathcal{K}$ is non-empty and compact. Then, problem (P) admits at least one global solution.*

Proof. Let $(x_k)_{k \in \mathbb{N}} \in \mathcal{K}$ be a minimizing sequence. By compactness, there exists $\bar{x} \in \mathcal{K}$ and a subsequence $(x_{k_\ell})_{\ell \in \mathbb{N}}$ of $(x_k)_{k \in \mathbb{N}}$ such that $\bar{x} = \lim_{\ell \to \infty} x_{k_\ell}$. Then, by continuity

$$f(\bar{x}) = \lim_{\ell \to \infty} f(x_{k_\ell}) = \inf_{x \in \mathcal{K}} f(x).$$

□

Example: Suppose that $f : \mathbb{R}^3 \to \mathbb{R}$ is given by $f(x, y, z) = x^2 - y^3 + \sin z$ and $\mathcal{K} = \{(x, y, z) \mid x^4 + y^4 + z^4 \leq 1\}$. Then f is continuous and $\mathcal{K}$ is compact. As a consequence, problem (P) admits at least one solution.

Theorem 5. [$\mathcal{K}$ closed but not bounded] *Suppose that $\mathcal{K}$ is non-empty, closed, and that f is continuous and "coercive" or "infinity at the infinity" in $\mathcal{K}$, i.e.*

$$\lim_{x\in\mathcal{K},\ \|x\|\to\infty} f(x) = +\infty. \tag{3}$$

Then, problem (P) admits at least one global solution.

Proof. Let $(x_k)_{k\in\mathbb{N}} \in \mathcal{K}$ be a minimizing sequence. Since $\inf_{x\in\mathcal{K}} f(x) = -\infty$ or $\inf_{x\in\mathcal{K}} f(x) \in \mathbb{R}$ and $\lim_{k\to\infty} f(x_k) = \inf_{x\in\mathcal{K}} f(x)$, there exists $R > 0$ such that $(x_k)_{k\in\mathbb{N}} \subseteq \mathcal{K}_R := \{x' \in \mathcal{K} \mid f(x') \leq R\} \subseteq \mathcal{K}$. By continuity of f, this set is closed and bounded because f is coercive. Arguing as in the previous proof, the compactness of $\mathcal{K}_R$ implies the existence of $\bar{x} \in \mathcal{K}_R$ such that a subsequence of $(x_k)_{k\in\mathbb{N}}$ converges to $\bar{x}$, which, by continuity of f, implies that $\bar{x}$ solves (P). $\square$

Example: Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is given by

$$f(x) = \langle Qx, x\rangle + c^\top x \quad \forall\, x \in \mathbb{R}^n,$$

where $Q \in \mathcal{M}_{n,n}(\mathbb{R})$ is symmetric and positive definite, and $c \in \mathbb{R}^n$. Since

$$\langle Qx, x\rangle \geq \lambda_{\min}(Q)\|x\|^2 \quad \forall\, x \in \mathbb{R}^n$$

(where $\lambda_{\min}(Q) > 0$ is the smallest eigenvalue of Q), we have that

$$f(x) \geq \lambda_{\min}(Q)\|x\|^2 - \|c\|\|x\| \quad \forall\, x \in \mathbb{R}^n.$$

This implies that $f(x) \to \infty$ as $\|x\| \to \infty$. Therefore,

$$\lim_{x\in\mathcal{K}, \|x\|\to\infty} f(x) = \infty, \tag{4}$$

holds for every closed set $\mathcal{K}$. Since f is also continuous, problem (P) admits at least one global solution for any given non-empty closed set $\mathcal{K}$.

Example: Suppose that $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x, y) = x^2 + y^3 \quad \forall\, (x, y) \in \mathbb{R}^2,$$

and

$$\mathcal{K} = \{(x, y) \in \mathbb{R}^2 \mid y \geq -1\}.$$

Then,

$$\lim_{x \in \mathcal{K}, \|x\| \to \infty} f(x) = +\infty \tag{5}$$

holds (exercise) and, hence, (P) admits at least one global solution.

Optimality conditions for unconstrained problems

◇ Notice that, by the second existence theorem, if f is continuous and satisfies that

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

then, if $\mathcal{K} = \mathbb{R}^n$, problem (P) admits at least one global solution.

◇ [First order optimality conditions for unconstrained problems]

We have the following result

Theorem 6. [Fermat's rule] *Suppose that $\mathcal{K} = \mathbb{R}^n$ and that $\bar{x}$ is a local solution to problem (P). If f is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) = 0$.*

Proof. For every $h \in \mathbb{R}^n$ and $\tau > 0$, the local optimality of $\bar{x}$ yields

$$f(\bar{x}) \leq f(\bar{x} + \tau h) = f(\bar{x}) + \tau \nabla f(\bar{x}) \cdot h + \tau \|h\| \varepsilon_{\bar{x}}(\tau h),$$

where $\lim_{z \to 0} \varepsilon_{\bar{x}}(z) = 0$. Therefore,

$$0 \leq \tau \nabla f(\bar{x}) \cdot h + \tau \|h\| \varepsilon_{\bar{x}}(\tau h).$$

Dividing by τ and letting $\tau \to 0$, we get

$$\nabla f(\bar{x}) \cdot h \geq 0.$$

Since h is arbitrary, we get that $\nabla f(\bar{x}) = 0$ (take for instance $h = -\nabla f(\bar{x})$ in the previous inequality). $\square$

◇ [Second order optimality conditions for unconstrained problems]

We have the following second order necessary condition for local optimality:

Theorem 7. *Suppose that $\mathcal{K} = \mathbb{R}^n$ and that $\bar{x}$ is a local solution to problem (P). If f is $\mathcal{C}^2$ in an open set A containing $\bar{x}$, then $D^2 f(\bar{x})$ is positive semidefinite. In other words,*

$$\langle D^2 f(\bar{x}) h, h \rangle \geq 0 \quad \forall\, h \in \mathbb{R}^n.$$

Proof. Let us fix $h \in \mathbb{R}^n$. By Taylor's theorem, for all $\tau > 0$ small enough, we have

$$f(\bar{x} + \tau h) \quad = \quad f(\bar{x}) + \nabla f(\bar{x}) \cdot (\tau h) + \tfrac{1}{2}\langle D^2 f(\bar{x})\tau h, \tau h\rangle + \|\tau h\|^2 R_{\bar{x}}(\tau h),$$

where $R_{\bar{x}}(\tau h) \to 0$ as $\tau \to 0$. Using the local optimality of $\bar{x}$, the previous result implies that $\nabla f(\bar{x}) = 0$ and, hence,

$$0 \leq f(\bar{x} + \tau h) - f(\bar{x}) = \frac{\tau^2}{2}\langle D^2 f(\bar{x})h, h\rangle + \tau^2 \|h\|^2 R_{\bar{x}}(\tau h).$$

Dividing by τ^2 and passing to the limit with $\tau \to 0$, we get the result. $\square$

We have the following second order sufficient condition for local optimality.

Theorem 8. *Suppose that* $f : \mathbb{R}^n \to \mathbb{R}$ *is* $\mathcal{C}^2$ *in an open set* A *containing* $\bar{x}$ *and that*

(i) $\nabla f(\bar{x}) = 0$.

(ii) *The matrix* $D^2 f(\bar{x})$ *is positive definite. In other words,*

$$\langle D^2 f(\bar{x})h, h\rangle > 0 \quad \forall\, h \in \mathbb{R}^n,\ h \neq 0.$$

Then, $\bar{x}$ *is a local solution to* (P).

Proof. Let $\lambda > 0$ be the smallest eigenvalue of $D^2 f(\bar{x})$, then

$$\forall h \in \mathbb{R}^n, \quad \langle D^2 f(\bar{x})h, h\rangle \geq \lambda \|h\|^2.$$

Using this inequality, the hypothesis $\nabla f(\bar{x}) = 0$, and the Taylor's expansion, for all

$h \in \mathbb{R}^n$ such that $\bar{x} + h \in A$ we have that

$$\begin{aligned} f(\bar{x}+h) - f(\bar{x}) &= \nabla f(\bar{x}) \cdot h + \frac{1}{2}\langle D^2 f(\bar{x})h, h\rangle + \|h\|^2 R_{\bar{x}}(h) \\ &\geq \frac{\lambda}{2}\|h\|^2 + \|h\|^2 R_{\bar{x}}(h) \\ &= \left(\frac{\lambda}{2} + R_{\bar{x}}(h)\right)\|h\|^2. \end{aligned}$$

Since $R_{\bar{x}}(h) \to 0$ as $h \to 0$, we can choose $\delta > 0$ such that $\|h\| \leq \delta$, $\bar{x} + h \in A$ and $|R_{\bar{x}}(h)| \leq \frac{\lambda}{4}$. As a consequence,

$$f(\bar{x}+h) - f(\bar{x}) \geq \frac{\lambda}{4}\|h\|^2 \quad \forall\, h \in \mathbb{R}^n \ \text{ with } \|h\| \leq \delta,$$

which proves the local optimality of $\bar{x}$. $\square$

Example: Let us study problem (P) with $\mathcal{K} = \mathbb{R}^2$ and

$$\mathbb{R}^2 \ni (x, y) \mapsto f(x, y) = 2x^3 + 3y^2 + 3x^2 y - 24y.$$

First, consider the sequence $(x_k, y_k) = (-k, 0)$ for $k \in \mathbb{N}$. Then,

$$f(x_k, y_k) = -2k^3 \to -\infty \text{ as } k \to \infty.$$

Therefore, $\inf_{(x,y)\in\mathbb{R}^2} f(x, y) = -\infty$ and problem (P) does not admit global solutions. Let us look for local solutions. We know that if (x, y) is a local solution, then it should satisfy $\nabla f(x, y) = (0, 0)$. This equation gives

$$\begin{aligned} 6x^2 + 6xy &= 0, \\ 6y + 3x^2 &= 24. \end{aligned}$$

From the first equation, we get that $x = 0$ or $x = -y$. In the first case, the second equation yields $y = 4$, while in the second case we obtain that $x^2 - 2x - 8 = 0$ which yields the two solutions $(4, -4)$ and $(-2, 2)$. Therefore, we have the three candidates $(0, 4)$, $(4, -4)$ and $(-2, 2)$. Let us check what can be obtained from the Hessian at these three points. We have that

$$D^2 f(x,y) = \begin{pmatrix} 12x + 6y & 6x \\ 6x & 6 \end{pmatrix}.$$

For the first candidate, we have

$$D^2 f(0,4) = \begin{pmatrix} 24 & 0 \\ 0 & 6 \end{pmatrix}.$$

which is positive definite. This implies that $(0,4)$ is a local solution of (P). For the second candidate, we have

$$D^2 f(4,-4) = \begin{pmatrix} 24 & 24 \\ 24 & 6 \end{pmatrix} = 6 \begin{pmatrix} 4 & 4 \\ 4 & 1 \end{pmatrix},$$

whose determinant is given by $36(-12) < 0$, which implies that $D^2 f(4,-4)$ is indefinite (the sign of the eigenvalues is not constant). Finally,

$$D^2 f(-2,2) = \begin{pmatrix} -12 & -12 \\ -12 & 6 \end{pmatrix}$$

which is also indefinite because the sign of the diagonal terms are not constant. Therefore, $(0,4)$ is the unique local solution to (P).

◇ [Maximization problems] If instead of problem (P) we consider the problem

$$\text{Find } \bar{x} \in \mathbb{R}^n \text{ such that} \quad f(\bar{x}) = \max\left\{ f(x) \mid x \in \mathcal{K} \right\}, \qquad (P')$$

then $\bar{x}$ is a local (resp. global) solution to (P') iff $\bar{x}$ is a local (resp. global) solution to (P) with f replaced by $-f$. In particular, if $\bar{x}$ is a local solution to (P') and f is regular enough, then we have the following first order necessary condition

$$\nabla f(\bar{x}) = 0,$$

as well as the following second order necessary condition

$$\langle D^2 f(\bar{x})h, h\rangle \leq 0 \quad \forall\, h \in \mathbb{R}^n.$$

In other words, $D^2 f(\bar{x})$ is negative semidefinite.

Conversely, if $\bar{x} \in \mathbb{R}^n$ is such that $\nabla f(\bar{x}) = 0$ and

$$\langle D^2 f(\bar{x})h, h\rangle < 0 \quad \forall\, h \in \mathbb{R}^n,\ h \neq 0.$$

Then, $\bar{x}$ is a local solution to (P').

Convexity

- ◇ [Convexity of a set] A non-empty set $C \subseteq \mathbb{R}^n$ is called convex if for any $\lambda \in [0,1]$ and x, $y \in C$, we have that

$$\lambda x + (1-\lambda) y \in C.$$

Let us fix a convex set $C \subseteq \mathbb{R}^n$.

- ◇ [Convexity of a function] A function $f : C \to \mathbb{R}$ is said to be convex if for any $\lambda \in [0,1]$ and x, $y \in C$, we have that

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

- ◇ [Relation between convex functions and convex sets] Given $f : \mathbb{R}^n \to \mathbb{R}$, let us define its epigraph epi(f) by

$$\text{epi}(f) := \left\{ (x, y) \in \mathbb{R}^{n+1} \mid y \geq f(x) \right\}.$$

Proposition 1. *The function f is convex iff the set epi(f) is convex.*

Proof. Indeed, suppose that f is convex and let (x_1, z_1), $(x_2, z_2) \in$ epi(f). Then, given $\lambda \in [0, 1]$ set

$$P_\lambda := \lambda(x_1, z_1) + (1 - \lambda)(x_2, z_2) = (\lambda x_1 + (1 - \lambda)x_2, \lambda z_1 + (1 - \lambda)z_2)$$

Since, by convexity,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2) \leq \lambda z_1 + (1 - \lambda)z_2,$$

we have that $P_\lambda \in$ epi(f). Conversely, assume that epi(f) is convex and let x_1, $x_2 \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. Since $(x_1, f(x_1))$, $(x_2, f(x_2)) \in$ epi(f), we deduce that

$$(\lambda x_1 + (1 - \lambda)x_2, \lambda f(x_1) + (1 - \lambda)f(x_2)) \in \text{epi}(f),$$

and, hence,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2),$$

which proves the convexity of f. $\square$

◇ [Strict convexity of a function] A function $f : C \to \mathbb{R}$ is said to be strictly convex if for any $\lambda \in (0,1)$ and x, $y \in C$, with $x \neq y$, we have that

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

◇ [Concavity and strict concavity of a function] A function $f : C \to \mathbb{R}$ is said to be concave if $-f$ is convex. Similarly, the function f is strictly concave if $-f$ is strictly convex.

Example: Let us show that the function $\mathbb{R}^n \ni x \mapsto \|x\| \in \mathbb{R}$ is convex but not strictly convex. Indeed, the convexity follows from the triangle inequality

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|.$$

Now, if we have that for some $\lambda \in (0,1)$

$$\|\lambda x + (1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|$$

the equality case in the triangle inequality ($\|a+b\| = \|a\| + \|b\|$ iff $a = 0$ and $b = 0$ or $a = \alpha b$ with $\alpha > 0$) shows that the previous inequality holds iff that $x = y = 0$

or $x = \gamma y$ for some $\gamma > 0$. By taking $x \neq 0$ and $y = \gamma x$ with $\gamma \in (0, \infty) \setminus \{1\}$ we conclude that $\| \cdot \|$ is not strictly convex.

Example: Let $\beta \in (1, +\infty)$. Let us show that the function $\mathbb{R}^n \ni x \mapsto \|x\|^\beta \in \mathbb{R}$ is strictly convex. Indeed, the real function $[0, +\infty) \ni t \mapsto \alpha(t) := t^\beta \in \mathbb{R}$ is increasing and strictly convex because

$$\alpha'(t) = \beta t^{\beta-1} > 0 \text{ and } \alpha''(t) = \beta(\beta-1)t^{\beta-2} > 0 \;\forall\, t \in (0, +\infty).$$

As a consequence, for any $\lambda \in [0, 1]$, using that

$$\|\lambda x + (1-\lambda)y\| \leq \lambda\|x\| + (1-\lambda)\|y\|,$$

we get that

$$\begin{aligned} \|\lambda x + (1-\lambda)y\|^\beta &\leq (\lambda\|x\| + (1-\lambda)\|y\|)^\beta \\ &\leq \lambda\|x\|^\beta + (1-\lambda)\|y\|^\beta, \end{aligned} \tag{6}$$

which implies the convexity of $\| \cdot \|^\beta$. Now, in order to prove the strict convexity, assume that for some $\lambda \in (0,1)$ we have

$$\|\lambda x + (1-\lambda)y\|^{\beta} = \lambda\|x\|^{\beta} + (1-\lambda)\|y\|^{\beta},$$

and let us prove that $x = y$. Then, all the inequalities in (6) are equalities and, hence,

$$\begin{aligned} &\|\lambda x + (1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|, \\ \text{and } &(\lambda\|x\| + (1-\lambda)\|y\|)^{\beta} = \lambda\|x\|^{\beta} + (1-\lambda)\|y\|^{\beta}. \end{aligned}$$

The equality case in the triangle inequality and the first relation above imply that $x = y = 0$ or $x = \gamma y$ for some $\gamma > 0$. The strict convexity of α and the second inequality above imply that $\|x\| = \|y\|$. Therefore, either $x = y = 0$ or both x and y are not zero and $x = \gamma y$ for some $\gamma > 0$ and $\|x\| = \|y\|$. In the latter case, we get that $\alpha = 1$ and, hence, $x = y$ from which the strict convexity follows.

◇ [Convexity and differentiability] We have the following result:

Theorem 9. *Let $f : C \to \mathbb{R}$ be a differentiable function. Then,*

(i) *f is convex in $\mathbb{R}^n$ if and only if for every $x \in C$ we have*

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x), \quad \forall\, y \in C. \tag{7}$$

(ii) *f is strictly convex in $\mathbb{R}^n$ if and only if for every $x \in C$ we have*

$$f(y) > f(x) + \nabla f(x) \cdot (y - x), \quad \forall\, y \in C,\ y \neq x. \tag{8}$$

Proof. (i) By definition of convex function, for any x, $y \in C$ and $\lambda \in (0, 1)$, we have

$$f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x),$$

and, hence,

$$f(\lambda y + (1 - \lambda)x) - f(x) \leq \lambda\,(f(y) - f(x))$$

Since, $\lambda y + (1-\lambda)x = x + \lambda(y-x)$, we get

$$\frac{f(x+\lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x).$$

Letting $\lambda \to 0$, Lemma 1 yields

$$\nabla f(x) \cdot (y-x) \leq f(y) - f(x).$$

Conversely, take x_1 and x_2 in C, $\lambda \in]0,1[$ and define $x_\lambda := \lambda x_1 + (1-\lambda)x_2$. By assumption,

$$\forall i \in \{1,2\}\,, \quad f(x_i) \geq f(x_\lambda) + \nabla f(x_\lambda) \cdot (x_i - x_\lambda),$$

and we obtain, by taking the convex combination, that

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(x_\lambda) + \nabla f(x_\lambda) \cdot (\lambda x_1 + (1-\lambda)x_2 - x_\lambda) = f(x_\lambda),$$

which shows that f is convex.

(ii) Since f is convex, by (i) we have that

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x), \quad \forall\, y \in C. \tag{9}$$

Suppose that there exists $z \in C$ such that $z \neq x$ and

$$f(z) = f(x) + \nabla f(x) \cdot (z - x).$$

Let $y = \frac{1}{2}x + \frac{1}{2}z$. Then, by (9), and strict convexity, we get

$$f(x)+\nabla f(x)\cdot(\tfrac{1}{2}z-\tfrac{1}{2}x) \leq f(y) < \tfrac{1}{2}f(x)+\tfrac{1}{2}f(z) = f(x)+\nabla f(x)\cdot(\tfrac{1}{2}z-\tfrac{1}{2}x),$$

which is impossible. The proof that (8) implies that f is strictly convex is completely analogous to the proof that (7) implies that f is convex. The result follows. $\square$

Theorem 10. *Let $f : C \to \mathbb{R}$ be $\mathcal{C}^2$ in C and suppose that C is open (besides being convex). Then*

(i) *f is convex if and only if $D^2 f(x)$ is positive semidefinite for all $x \in C$.*
(ii) *f is strictly convex if $D^2 f(x)$ is positive definite for all $x \in C$*

Proof. (i) Suppose that f is convex. Then, by Taylor's theorem for every $x \in C$, $h \in \mathbb{R}^n$ and $\tau > 0$ small enough such that $x + \tau h \in C$ we have

$$f(x+\tau h) = f(x) + \tau \nabla f(x) \cdot h + \frac{\tau^2}{2}\langle D^2 f(x)h, h\rangle + \tau^2 \|h\|^2 R_x(\tau h),$$

which implies, by the first order characterization of convexity, that

$$0 \leq \tfrac{1}{2}\langle D^2 f(x)h, h\rangle + \|h\|^2 R_x(\tau h).$$

Using that $\lim_{\tau \to 0} R_x(\tau h) = 0$, and the fact that h is arbitrary, we get that

$$\langle D^2 f(x)h, h\rangle \geq 0 \;\; \forall\, h \in \mathbb{R}^n.$$

Suppose that $D^2 f(x)$ is positive semidefinite for all $x \in C$ and assume, for the time being, that for every x, $y \in C$ there exists $c_{xy} \in \{\lambda x + (1-\lambda)y \mid \lambda \in (0,1)\}$ such that

$$f(y) = f(x) + \nabla f(x) \cdot (y-x) + \tfrac{1}{2}\langle D^2 f(c_{xy})(y-x), y-x\rangle. \qquad (10)$$

Then, have that

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) \quad \forall\, x,\ y \in C,$$

and, hence, f is convex. It remains to prove (10). Defining $g(\tau) := f(x + \tau(y - x))$ for all $\tau \in [0, 1]$, formula (10) follows from the equality

$$g(1) = g(0) + g'(0) + \tfrac{1}{2} g''(\hat{\tau})$$

for some $\hat{\tau} \in (0, 1)$.
(ii) The assertion follows directly from (10), with $y \neq y$, and Theorem 9(ii). $\square$

Remark 3. Note that the positive definiteness of $D^2 f(x)$, for all $x \in C$, is only a sufficient condition for strict convexity but not necessary. Indeed, the function $\mathbb{R} \ni x \mapsto f(x) = x^4 \in \mathbb{R}$ is strictly convex but $f''(0) = 0$.

Example: Let $Q \in \mathcal{M}_{n,n}(\mathbb{R})$ be symmetric and let $f : \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f(x) = \tfrac{1}{2} x^\top Q x + c^\top x.$$

Then, $D^2 f(x) = Q$ and hence f is convex if Q is positive semidefinite and strictly convex if Q is positive definite.
In this case, the fact that Q is positive definite is also a necessary condition for strict convexity. Indeed, for simplicity suppose that $c = 0$ and write $Q = PDP^\top$, where the set of columns of P is an orthonormal basis of eigenvectors of Q (which exists because Q is symmetric), and D is the diagonal matrix containing the corresponding eigenvalues $\{\lambda_i\}_{i=1}^N$ in the diagonal. Set $y(x) = P^\top x$. Then,

$$f(x) = \sum \lambda_{i=1}^n y_i(x)^2.$$

If Q is not definite positive, then there exists $j \in \{1, \ldots, N\}$ such that $\lambda_j \leq 0$. Then, it is easy to see that f is not strictly convex on the set $\{x \in \mathbb{R}^n \mid y_i(x) = 0, \text{ for all } i \in \{1, \ldots, n\} \setminus \{j\}\}$.

Convex optimization problems

◇ [Optimality conditions for convex problems] Let us begin with a definition.

Definition 2. *Problem (P) is called convex if f is convex and $\mathcal{K}$ is a non-empty closed and convex set.*

We have the following result.

Theorem 11. [Characterization of solutions for convex problems] *Suppose that problem (P) is convex and that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable in $\mathcal{K}$. Then, the following statements are equivalent:*

(i) *$\bar{x}$ is a local solution to (P).*

(ii) *The following inequality holds:*

$$\langle \nabla f(\bar{x}), x - \bar{x} \rangle \geq 0 \quad \forall\, x \in \mathcal{K}. \tag{11}$$

(iii) *$\bar{x}$ is a global solution to (P).*

Proof. Let us prove that (i) implies (ii). Indeed, by convexity of $\mathcal{K}$ we have that given $x \in \mathcal{K}$ for any $\tau \in [0,1]$ the point $\tau x + (1-\tau)\bar{x} = \bar{x} + \tau(x - \bar{x}) \in \mathcal{K}$. Therefore, by the differentiability of f, if τ is small enough, we have

$$0 \leq f(\bar{x} + \tau(x - \bar{x})) - f(\bar{x}) = \tau \nabla f(\bar{x}) \cdot (x - \bar{x}) + \tau \|x - \bar{x}\| \varepsilon_{\bar{x}}(\tau \|x - \bar{x}\|),$$

where $\lim_{h \to 0} \varepsilon_{\bar{x}}(h) = 0$. Dividing by τ and letting $\tau \to 0$, we get (ii).

The proof that (ii) implies (iii) follows directly from the inequalities

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) \geq f(\bar{x}) \quad \forall\, x \in \mathcal{K}.$$

Finally, (iii) implies (i) is trivial. The result follows. $\square$

Remark 4. In particular, if $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and $\mathcal{K} = \mathbb{R}^n$, the relation

$$\nabla f(\bar{x}) = 0,$$

is a necessary and sufficient condition for $\bar{x}$ to be a global solution to (P).

Proposition 2. *Suppose that $\mathcal{K}$ is convex and that f is strictly convex in $\mathcal{K}$. Then, there exists at most one solution to problem (P).*

Proof. Assume, by contradiction, that $x_1 \neq x_2$ are both solutions to (P). Then, $\frac{1}{2}x_1 + \frac{1}{2}x_2 \in \mathcal{K}$ and

$$f(\tfrac{1}{2}x_1 + \tfrac{1}{2}x_2) < \tfrac{1}{2}f(x_1) + \tfrac{1}{2}f(x_2) = \tfrac{1}{2}\min_{x\in\mathcal{K}} f(x) + \tfrac{1}{2}\min_{x\in\mathcal{K}} f(x) = \min_{x\in\mathcal{K}} f(x),$$

which is impossible. $\square$

◇ [Least squares] Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ and consider the system $Ax = b$. Suppose that $m > n$. This type of systems appear, for instance, in data fitting problem and it is often ill-posed, in the sense that there is no x satisfying the equation. In this case, one usually considers the optimization problem

$$\min_{x\in\mathcal{K}:=\mathbb{R}^n} f(x) := \|Ax - b\|^2. \tag{12}$$

Note that

$$f(x) = \langle A^\top Ax, x\rangle - 2\langle A^\top b, x\rangle + \|b\|^2.$$

and, hence, $D^2 f(x) = 2A^\top A$, which is symmetric positive semidefinite, and, hence, f is convex. Let us assume that the columns of A are linearly independent. Then, for any $h \in \mathbb{R}^n$,

$$\langle A^\top A h, h\rangle = 0 \Leftrightarrow Ah = 0 \Leftrightarrow h = 0,$$

i.e. for all $x \in \mathbb{R}^n$, the matrix $D^2 f(x)$ is symmetric positive definite and, hence, f is strictly convex. Moreover, denoting by $\lambda_{\min} > 0$ the smallest eigenvalue of $2A^\top A$, we have

$$f(x) \geq \lambda_{\min}\|x\|^2 - 2\langle A^\top b, x\rangle + \|b\|^2.$$

and, hence, f is infinity at the infinity. Therefore, problem (12) admits only one solution $\bar{x}$. By Remark 4, the solution $\bar{x}$ is characterized by

$$A^\top A\bar{x} = A^\top b, \quad \text{i.e.} \quad \bar{x} = (A^\top A)^{-1} A^\top b.$$

◇ [Projection on a closed and convex set] Suppose that $\mathcal{K}$ is a nonempty closed and convex set and let $y \in \mathbb{R}^n$. Consider the problem the projection problem

$$\inf\left\{\|x - y\| \mid x \in \mathcal{K}\right\}. \qquad (Proj_{\mathcal{K}})$$

Note that $\mathcal{K}$ being closed and the cost functional being coercive, we have the existence of at least one solution $\bar{x}$. In order, to characterize $\bar{x}$ notice that the set of solutions to $(Proj_{\mathcal{K}})$ is the same as the set of solutions to the problem

$$\inf\left\{\tfrac{1}{2}\|x-y\|^2 \mid x \in \mathcal{K}\right\}.$$

Then, since the cost functional of the problem above is strictly convex, Proposition 2 implies that $\bar{x}$ is its unique solution and, hence, is also the unique solution to $(Proj_{\mathcal{K}})$. Moreover, by Theorem 11(ii), we have that $\bar{x}$ is characterized by the inequality

$$(y-\bar{x})\cdot(x-\bar{x}) \leq 0 \quad \forall\, x \in \mathcal{K}. \tag{13}$$

Example: Let $b \in \mathbb{R}^m$ and $A \in \mathcal{M}_{m\times n}$ be such that

$$b \in \mathrm{Im}(A) := \{Ax \mid x \in \mathbb{R}^m\}.$$

Suppose that

$$\mathcal{K} = \{x \in \mathbb{R}^n \mid Ax = b\}. \tag{14}$$

Then, $\mathcal{K}$ is closed, convex and nonempty. Moreover, for any $h \in \mathrm{Ker}(A)$ we have that $\bar{x} + h \in \mathcal{K}$. As a consequence, (13) implies that

$$(y - \bar{x}) \cdot h \leq 0 \ \ \forall\, h \in \mathrm{Ker}(A),$$

and, using that $h \in \mathrm{Ker}(A)$ iff $-h \in \mathrm{Ker}(A)$, we get that

$$(y - \bar{x}) \cdot h = 0 \ \ \forall\, h \in \mathrm{Ker}(A). \tag{15}$$

Conversely, since for every $x \in \mathcal{K}$ we have $x - \bar{x} \in \mathrm{Ker}(A)$, relation (15) implies (13), and, hence, (15) characterizes $\bar{x}$. Note that (15) can be written as[1]

$$y - \bar{x} \in \mathrm{Ker}(A)^{\perp} = \left\{ v \in \mathbb{R}^n \mid v^{\top} h = 0 \ \ \forall\, h \in \mathrm{Ker}(A) \right\},$$

[1] Recall that given a subspace V of $\mathbb{R}^n$, the orthogonal space $V^{\perp}$ is defined by

$$V^{\perp} := \{ z \in \mathbb{R}^n \mid z^{\top} v = 0 \ \forall\, v \in V \}.$$

Two important properties of the orthogonal space are $V \oplus V^{\perp} = \mathbb{R}^n$, and $(V^{\perp})^{\perp} = V$.

or, equivalently,

$$y = \bar{x} + z \text{ for some } z \in \mathrm{Ker}(A)^{\perp}. \tag{16}$$

◇ [Convex problems with equality constraints] Now, we consider the same set $\mathcal{K}$ as in (14) but we consider a general differentiable convex objective function $f : \mathbb{R}^n \to \mathbb{R}$. We will need the following result from Linear Algebra.

Lemma 2. *Let* $A \in \mathcal{M}_{m,n}(\mathbb{R})$. *Then,* $\mathrm{Ker}(A)^{\perp} = \mathrm{Im}(A^{\top})$.

Proof. By the previous footnote, the desired relation is equivalent to $\mathrm{Im}(A^{\top})^{\perp} = \mathrm{Ker}(A)$. Now, $x \in \mathrm{Im}(A^{\top})^{\perp}$ iff $\langle x, A^{\top} y\rangle = 0$ for all $y \in \mathbb{R}^m$, and this holds iff $\langle Ax, y\rangle = 0$ for all $y \in \mathbb{R}^m$, i.e. $x \in \mathrm{Ker}(A)$. □

Proposition 3. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable and convex and suppose that the set* $\mathcal{K}$ *in* (14) *is nonempty. Then* $\bar{x}$ *is a global solution to* (P) *iff* $\bar{x} \in \mathcal{K}$ *and there exists* $\lambda \in \mathbb{R}^m$ *such that*

$$\nabla f(\bar{x}) + A^{\top}\lambda = 0. \tag{17}$$

Proof. We are going to show that (17) is equivalent to (11) from which the result follows. Indeed, exactly as in the previous example, we have that (11) is equivalent to

$$\nabla f(\bar{x}) \cdot h = 0 \quad \forall\, h \in \mathrm{Ker}(A),$$

i.e.

$$\nabla f(\bar{x}) \in \mathrm{Ker}(A)^{\perp}.$$

Lemma 2 implies the existence of $\mu \in \mathbb{R}^m$ such that $\nabla f(\bar{x}) = A^{\top}\mu$. Setting $\lambda = -\mu$ we get (17). $\square$

Example: Let $Q \in \mathcal{M}_{n,n}(\mathbb{R})$ be symmetric and positive definite, and $c \in \mathbb{R}^n$. In the framework of the previous proposition, suppose that f is given by

$$f(x) = \frac{1}{2}\langle Qx, x\rangle + c^{\top}x \quad \forall\, x \in \mathbb{R}^n,$$

and that A has m linearly independent columns. A classical linear algebra result states that this is equivalent to the fact that the m lines of A are linearly independent. In this case, we say that A has full rank.

Under the previous assumptions on Q, we have seen that f is strictly convex. Moreover, the condition on the columns of A implies that $\mathrm{Im}(A) = \mathbb{R}^m$ and, hence, $\mathcal{K} \neq \emptyset$. Now, by Proposition 3 the point $\bar{x}$ solves (P) iff $\bar{x} \in \mathcal{K}$ and there exists $\lambda \in \mathbb{R}^m$ such that (17) holds. In other words, there exists $\lambda \in \mathbb{R}^m$ such that

$$A\bar{x} = b, \quad \text{and} \quad Q\bar{x} + c + A^{\top}\lambda = 0.$$

The second equation above yields $\bar{x} = -Q^{-1}(c + A^\top \lambda)$ and, hence, by the first equation, we get

$$AQ^{-1}c + AQ^{-1}A^\top \lambda + b = 0. \tag{18}$$

Let us show that $M := AQ^{-1}A^\top$ is invertible. Indeed, since $M \in \mathcal{M}_{m,m}(\mathbb{R})$ it suffices to show that $My = 0$ implies that $y = 0$. Now, let $y \in \mathbb{R}^m$ such that $My = 0$. Then, $\langle My, y \rangle = 0$ and, hence, $\langle Q^{-1}A^\top y, A^\top y \rangle = 0$, which implies, since Q^{-1} is also positive definite, that $A^\top y = 0$. Now, since the columns of $A^\top$ are also linearly independent, we deduce that $y = 0$, i.e. M is invertible. Using this fact, we can solve for λ in (18), obtaining

$$\lambda = -M^{-1}\left(AQ^{-1}c + b\right).$$

We deduce that

$$\bar{x} = -Q^{-1}\left(c - A^\top M^{-1}\left(AQ^{-1}c + b\right)\right), \tag{19}$$

is the unique solution to this problem.

Example: Let us now consider the projection problem

$$\begin{aligned} \min \ & \tfrac{1}{2}\|x-y\|^2 \\ \text{s.t.} \quad & Ax = b. \end{aligned}$$

Noting that $\frac{1}{2}\|x-y\|^2 = \frac{1}{2}\|x\|^2 - y^\top x + \frac{1}{2}\|y\|^2$, the previous problem has the same solution than

$$\begin{aligned} \min \ & \tfrac{1}{2}\|x\|^2 - y^\top x \\ \text{s.t.} \quad & Ax = b, \end{aligned}$$

which corresponds to $Q = I_{n\times n}$ (the $n \times n$ identity matrix) and $c = -y$. Then, (19) implies that the solution of this problem is given by

$$\bar{x} = (I - A^\top (AA^\top)^{-1} A)y + A^\top (AA^\top)^{-1} b.$$

Note that if $h \in \mathrm{Ker}(A)$

$$
\begin{aligned}
\langle y - \bar{x}, h \rangle &= \langle A^\top (AA^\top)^{-1} Ay - A^\top (AA^\top)^{-1} b, h \rangle \\
&= \langle (AA^\top)^{-1} Ay - (AA^\top)^{-1} b, Ah \rangle \\
&= 0,
\end{aligned}
$$

confirming (16).

◇ [Separation of a point and a closed convex set] We have the following result:

Proposition 4. *Let $\mathcal{K}$ be a nonempty closed and convex set and let $y \notin \mathcal{K}$. Then, there exists $p \in \mathbb{R}^n$ such that*

$$
\langle p, x \rangle < \langle p, y \rangle \quad \forall\, x \in \mathcal{K}.
$$

Proof. Let $\bar{x} \in \mathcal{K}$ be the projection of y onto $\mathcal{K}$. Let us define the affine function

$$
\ell(x) := \langle y - \bar{x}, x \rangle - \langle y - \bar{x}, \bar{x} \rangle.
$$

Then, the tangent plane to $\mathcal{K}$ at $\bar{x}$ is given by $\Pi_{\bar{x}} := \{x \in \mathbb{R}^d \mid \ell(x) = 0\}$. Indeed,

$\bar{x} \in \Pi_{\bar{x}}$ and, by (13), $\ell(x) \leq 0$ for all $x \in \mathcal{K}$. Now, since $\ell(y) = \|y - x\|^2 > 0$ (because $y \notin \mathcal{K}$), we deduce that $\ell(x) < \ell(y)$ for all $x \in \mathcal{K}$, which yields $\langle y - \bar{x}, x \rangle < \langle y - \bar{x}, y \rangle$ for all $x \in \mathcal{K}$. Setting $p := y - \bar{x}$, we have proven the result. $\square$

◇ [Cones and polar cones]

Definition 3. (i) *A set $C \subseteq \mathbb{R}^n$ is a cone if*

$$\forall\ h \in C, \quad \forall\ \tau \geq 0 \quad \textit{we have } \tau h \in C.$$

(ii) *The set*

$$C^\circ := \{u \in \mathbb{R}^n \mid \langle u, h \rangle \leq 0\ \forall\ h \in C\},$$

is called the polar cone of C.

The simplest example of a cone is any subspace V of $\mathbb{R}^n$. In this case, V is a convex cone and we have $V^\circ = V^\perp$, the orthogonal space to V. In particular, if $a_i \in \mathbb{R}^n$ $(i = 1, \ldots, m)$, then, the set

$$C = \{h \in \mathbb{R}^n \mid \langle a_i, h \rangle = 0 \quad \forall\ i = 1, \ldots, m\},$$

which, denoting by $A \in \mathcal{M}_{m,n}(\mathbb{R})$ the matrix whose ith row is a_i, can be written as

$$C = \{h \in \mathbb{R}^n \mid Ax = 0\} = \mathrm{Ker}(A),$$

is a convex cone and, as a consequence of Lemma 2, we have

$$C^\circ = \mathrm{Ker}(A)^\perp = \mathsf{Im}(A^\top) = \left\{ A^\top \lambda \mid \lambda \in \mathbb{R}^m \right\}.$$

Now, suppose that C is given by

$$C = \{h \in \mathbb{R}^n \mid \langle a_i, h \rangle \leq 0 \quad \forall\, i = 1, \ldots, m\}.$$

Our purpose now is to compute C°.

Lemma 3. *Denote, as before, by $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ the matrix whose ith row is given by a_i. Then*

$$\begin{aligned} C^\circ &= \left\{ A^\top \lambda \mid \lambda \in \mathbb{R}^m, \quad \lambda_i \geq 0 \ \ \forall\, i = 1, \ldots, m \right\}, \\ &= \left\{ \sum_{i=1}^m \lambda_i a_i \mid \lambda_i \geq 0 \ \ \forall\, i = 1, \ldots, m \right\}. \end{aligned} \tag{20}$$

Proof. If $\lambda \in \mathbb{R}^m$ with $\lambda_i \geq 0 \ \ \forall\, i = 1, \ldots, m$, then for every $h \in C$ we have

$$\langle A^\top \lambda, h\rangle = \langle \lambda, Ah\rangle = \sum_{i=1}^{m} \lambda_i \langle a_i, h\rangle \leq 0,$$

and, hence, $A^\top \lambda \in C^\circ$. Now, denote by B the set on the right hand side of (20). Clearly, B is convex and nonempty. Moreover, B can also be shown to be closed. Suppose that $u \in C^\circ$ and $u \notin B$. Then, by Proposition 4, there exists $p \in \mathbb{R}^n$ such that

$$\langle p, x\rangle < \langle p, u\rangle \ \ \forall\, x \in B,$$

i.e.

$$\langle p, A^\top \lambda\rangle < \langle p, u\rangle \quad \forall\, \lambda \in \mathbb{R}^m, \quad \lambda_i \geq 0 \quad \forall\, i = 1, \ldots, m. \tag{21}$$

Now, the previous inequality has the following two consequences:

(i) $\langle p, u\rangle > 0$. Indeed, it suffices to take $\lambda = 0$ in (21).

(ii) $\langle p, a_i\rangle \leq 0$ for all $i = 1, \ldots, m$. Indeed, fix $i \in \{1, \ldots, m\}$ and $\gamma > 0$. By taking

$$\lambda = (0, \ldots, \underbrace{\gamma}_{i}, \ldots, 0)$$

in (21) we get $\gamma\langle p, a_i\rangle < \langle p, u\rangle$ which implies that $\langle p, a_i\rangle \leq 0$ (if this is not the case, by taking γ large enough we get a contradiction).

From (i)-(ii), we conclude that $p \in C$ and $\langle p, u\rangle > 0$ which contradicts the fact that $u \in C^\circ$. $\square$

Now, let us consider the case where C is defined by both linear equalities and inequalities. Let $a_i \in \mathbb{R}^n$ $(i = 1, \dots, m)$ and $a'_j \in \mathbb{R}^n$ $(j = 1, \dots, p)$. Suppose that C is given by

$$C = \left\{ h \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle a_i, h\rangle = 0 \quad \forall\, i = 1, \dots, m \\ \langle a'_j, h\rangle \leq 0 \quad \forall\, j = 1, \dots, p \end{array} \right\}.$$

Lemma 4. *We have*

$$C^\circ = \left\{ \sum_{i=1}^{m} \lambda_i a_i + \sum_{j=1}^{p} \mu_j a'_j \;\middle|\; \mu_j \geq 0 \;\; \forall\, j = 1, \dots, p \right\}.$$

Proof. The set C can be written as

$$C = \left\{ h \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle a_i, h \rangle \le 0 \quad \forall\, i = 1, \ldots, m, \\ \langle -a_i, h \rangle \le 0 \quad \forall\, i = 1, \ldots, m, \\ \langle a_j', h \rangle \le 0 \quad \forall\, j = 1, \ldots, p \end{array} \right\}.$$

Lemma 3 implies that $u \in C^\circ$ iff there exist $\alpha_1 \ge 0$, $\beta_1 \ge 0, \ldots, \alpha_m \ge 0,\ \beta_m \ge 0$ and $\mu_1 \ge 0, \ldots, \mu_p \ge 0$ such that

$$\begin{aligned} u &= \textstyle\sum_{i=1}^m \alpha_i a_i + \sum_{i=1}^m \beta_i(-a_i) + \sum_{j=1}^p \mu_j a_j' \\ &= \textstyle\sum_{i=1}^m (\alpha_i - \beta_i) a_i + \sum_{j=1}^p \mu_j a_j'. \end{aligned}$$

The result follows by setting $\lambda_i = \alpha_i - \beta_i$ for all $i = 1, \ldots, m$. □

◇ [Application to convex problems with affine equality and inequality constraints] Let $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ $(i = 1, \ldots, m)$, $a_j' \in \mathbb{R}^n$ and $b_j' \in \mathbb{R}$ $(j = 1, \ldots, p)$. Suppose that the constraint set $\mathcal{K}$ is given by

$$\mathcal{K} = \left\{ x \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle a_i, x\rangle + b_i = 0 \quad \forall\, i = 1, \dots, m, \\ \langle a_j', x\rangle + b_j' \le 0 \quad \forall\, j = 1, \dots, p \end{array} \right\}. \tag{22}$$

The main result in this section is the following.

Proposition 5. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\mathcal{C}^1$. Moreover, assume that the constraint set $\mathcal{K}$ is nonempty and given by* (22)*. Then, $\bar{x}$ is a global solution to (P) iff $\bar{x} \in \mathcal{K}$ and there exist $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$ such that*

$$\begin{array}{l} \nabla f(\bar{x}) + \sum_{i=1}^m \lambda_i a_i + \sum_{j=1}^p \mu_j a_j' = 0, \\ \text{and} \quad \forall\, j \in \{1, \dots, p\} \;\; \text{we have } \mu_j \ge 0, \quad \text{and} \quad \mu_j(\langle a_j', \bar{x}\rangle + b_j') = 0 \end{array} \tag{23}$$

Proof. Suppose that $\bar{x}$ solves (P). Then, by Theorem 11(ii), we have

$$\langle \nabla f(\bar{x}), x - \bar{x}\rangle \ge 0 \;\; \forall\, x \in \mathcal{K}. \tag{24}$$

Now, let us define the set of active inequality constraints

$$I(\bar{x}) = \left\{ j \in \{1, \ldots, p\} \mid \langle a'_j, \bar{x} \rangle + b'_j = 0 \right\},$$

and consider the set

$$C := \left\{ h \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle a_i, h \rangle = 0 \quad \forall\, i = 1, \ldots, m \\ \langle a'_j, h \rangle \leq 0 \quad \forall\, j \in I(\bar{x}) \end{array} \right\}.$$

Let $\tau > 0$ and $h \in C$. Then, we claim that $\bar{x} + \tau h \in \mathcal{K}$ if τ is small enough. Indeed,

$$\langle a_i, \bar{x} + \tau h \rangle + b_i = \langle a_i, \bar{x} \rangle + b_i = 0 \quad \forall\, i = 1, \ldots, m,$$

$$\langle a'_j, \bar{x} + \tau h \rangle + b'_j \leq \langle a'_j, \bar{x} \rangle + b'_j \leq 0 \quad \forall\, j \in I(\bar{x}),$$

$$\langle a'_j, \bar{x} + \tau h \rangle + b'_j = \langle a'_j, \bar{x} \rangle + b'_j + \tau \langle a'_j, h \rangle < 0 \quad \forall\, j \in \{1, \ldots, p\} \setminus I(\bar{x}),$$

where the last inequality holds because $\langle a'_j, \bar{x} \rangle + b'_j < 0$ for all $j \in \{1, \ldots, p\} \setminus I(\bar{x})$ and we can pick τ small enough in order to ensure that $\langle a'_j, \bar{x} \rangle + b'_j + \tau \langle a'_j, h \rangle < 0$

for all $j \in \{1, \ldots, p\} \setminus I(\bar{x})$. Indeed, it suffices to take $\tau > 0$ such that

$$\tau \max_{j \in \{1,\ldots,m\} \setminus I(\bar{x})} |\langle a_j', h \rangle| < \min_{j \in \{1,\ldots,m\} \setminus I(\bar{x})} -\langle a_j', \bar{x} \rangle - b_j'.$$

Thus, by (24) we deduce that

$$\langle \nabla f(\bar{x}), h \rangle \geq 0 \;\; \forall \; h \in C,$$

which means that

$$-\nabla f(\bar{x}) \in C^{\circ}.$$

Thus, from Lemma 4 we get the existence of $\lambda \in \mathbb{R}^m$, $\mu \in \mathbb{R}^p$, such that $\mu_j \geq 0$ for all $j \in \{1, \ldots, p\}$, $\mu_j = 0$ for all $j \in \{1, \ldots, p\} \setminus I(\bar{x})$, and

$$-\nabla f(\bar{x}) = \sum_{i=1}^{m} \lambda_i a_i + \sum_{j=1}^{p} \mu_j a_j'.$$

Condition (23) follows directly from the previous relation. Conversely, assume that

$(\bar{x}, \lambda, \mu)$ satisfies (23) and define the function $L(\cdot, \lambda, \mu)$ as

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i(\langle a_i, x\rangle + b_i) + \sum_{j=1}^{p} \mu_j(\langle a_j', x\rangle + b_j').$$

Then, $L(\cdot, \lambda, \mu)$ is convex, $\mathcal{C}^1$, and satisfies

$$L(x, \lambda, \mu) = f(\bar{x}), \quad \text{and} \quad \nabla_x L(\bar{x}, \lambda, \mu) = 0.$$

Thus, using that for all $x \in \mathcal{K}$ we have $L(x, \lambda, \mu) \leq f(x)$, we get

$$f(\bar{x}) = L(\bar{x}, \lambda, \mu) + \langle \nabla_x L(\bar{x}, \lambda, \mu), x - \bar{x}\rangle \leq L(x, \lambda, \mu) \leq f(x),$$

for all $x \in \mathcal{K}$, which implies that $\bar{x}$ is a global solution to (P). $\square$

Example: Let us consider the projection problem

$$\inf_{x \in \mathcal{K}} \tfrac{1}{2}\|x - y\|^2,$$

where

$$\mathcal{K} = \{x \in \mathbb{R}^n \mid x_1 \leq x_2 \leq \ldots \leq x_n\}.$$

Clearly $\mathcal{K}$ is nonempty, closed and convex. Therefore, the projection $\bar{x}$ of y onto $\mathcal{K}$ exists and it is unique. Let us define

$$a_j' := (0, \ldots, \underbrace{1}_{j}, \underbrace{-1}_{j+1}, \ldots, 0), \text{ and } \; b_j' = 0.$$

Then,

$$\mathcal{K} = \left\{ x \in \mathbb{R}^n \mid \langle a_j', x\rangle + b_j' \leq 0 \;\; \forall\, j = 1, \ldots, n-1 \right\}.$$

By the previous proposition, $\bar{x}$ solves (P) iff $\bar{x} \in \mathcal{K}$ and there exists $\mu \in \mathbb{R}^{n-1}$ such that (23) holds. Namely,

$$
\begin{gathered}
\bar{x}_1 - y_1 + \mu_1 = 0 \\
\bar{x}_2 - y_2 + \mu_2 - \mu_1 = 0 \\
\vdots \\
\bar{x}_{n-1} - y_{n-1} + \mu_{n-1} - \mu_{n-2} = 0 \\
\bar{x}_n - y_n - \mu_{n-1} = 0 \\
\mu_i \geq 0 \quad \mu_i h_i(\bar{x}) = 0 \quad \forall\ i = 1, \ldots, n-1,
\end{gathered}
$$

which is equivalent to

$$
\begin{gathered}
\bar{x}_1 - y_1 \leq 0 \\
\bar{x}_2 + \bar{x}_1 - (y_2 + y_1) \leq 0 \\
\vdots \\
\textstyle\sum_{k=1}^{n-1} \bar{x}_k - \sum_{k=1}^{n-1} y_k \leq 0 \\
\textstyle\sum_{k=1}^{n} \bar{x}_k - \sum_{k=1}^{n} y_k = 0, \\
\left(\textstyle\sum_{k=1}^{i} \bar{x}_k - \sum_{k=1}^{i} y_k\right) h_i(\bar{x}) = 0 \quad \forall\ i = 1, \ldots, n-1.
\end{gathered}
$$

Let us compute $\bar{x}$ when $n = 4$ and $y = (2, 1, 5, 4)$. In this case, we have

$$\bar{x}_1 - 2 \leq 0$$
$$\bar{x}_2 + \bar{x}_1 - 3 \leq 0$$
$$\bar{x}_3 + \bar{x}_2 + \bar{x}_1 - 8 \leq 0$$
$$\bar{x}_4 + \bar{x}_3 + \bar{x}_2 + \bar{x}_1 - 12 = 0$$
$$(\bar{x}_1 - 2)(\bar{x}_1 - \bar{x}_2) = 0$$
$$(\bar{x}_2 + \bar{x}_1 - 3)(\bar{x}_2 - \bar{x}_3) = 0$$
$$(\bar{x}_3 + \bar{x}_3 + \bar{x}_1 - 8)(\bar{x}_3 - \bar{x}_4) = 0$$

The first two equations and the constraint $\bar{x}_1 \leq \bar{x}_2$ imply that $\bar{x}_1 = \bar{x}_2 < 2$. Then, taking $\bar{x}_1 = \bar{x}_2 = 3/2$, the second equation is satisfied, but the third inequality cannot be satisfied with an equality and, hence, we take $\bar{x}_3 = \bar{x}_4$ and the fourth relation yields $\bar{x}_3 = \bar{x}_4 = 9/2$. Thus, the point $(\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4) = (3/2, 3/2, 9/2, 9/2)$ satisfies the previous system and, hence, solves the projection problem.

Optimization problems with equality and inequality constraints

◇ [Abstract optimality condition] In this section we establish an abstract optimality condition for the general problem (P) with $\mathcal{K}$ being a nonempty closed set and $f : \mathbb{R}^n \to \mathbb{R}$ being differentiable. We need first the following definition.

Definition 4. *Let $\bar{x} \in \mathcal{K}$. We say that $h \in \mathbb{R}^n$ is a tangent vector to $\mathcal{K}$ at $\bar{x}$ if there exist $(\tau_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ and $(\epsilon_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^n$ such that*

$$\begin{gathered} \tau_n > 0 \;\; \forall\, n \in \mathbb{N}, \quad \tau_n \overset{n\to+\infty}{\longrightarrow} 0, \quad \epsilon_n \overset{n\to+\infty}{\longrightarrow} 0, \\ \bar{x} + \tau_n h + \tau_n \epsilon_n \in \mathcal{K} \;\; \forall\, n \in \mathbb{N}. \end{gathered} \tag{25}$$

The set of tangent vectors to $\mathcal{K}$ at $\bar{x}$ is called the tangent cone to $\mathcal{K}$ at $\bar{x}$ and it is denoted by $T_{\mathcal{K}}(\bar{x})$.

Remark 5. (i) By definition $h \in T_{\mathcal{K}}(\bar{x})$ iff there exists $(\tau_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}$ and $(h_n)_{n\in\mathbb{N}} \subseteq \mathbb{R}^n$ such that $\tau_n \geq 0$, $\tau_n \overset{n\to+\infty}{\longrightarrow} 0$, $h_n \overset{n\to+\infty}{\longrightarrow} h$, and $\bar{x} + \tau_n h_n \in \mathcal{K}$ for all $n \in \mathbb{N}$.

(ii) It is easy to see that $T_{\mathcal{K}}(\bar{x})$ is indeed a closed cone.

Using this notion, we can prove the following result.

Theorem 12. [Abstract optimality condition] *Suppose that $\bar{x} \in \mathcal{K}$ a local solution to (P) and that f is differentiable at $\bar{x}$. Then,*

$$\langle \nabla f(\bar{x}), h \rangle \geq 0 \quad \forall\, h \in T_{\mathcal{K}}(\bar{x}).$$

Proof. Let $h \in T_{\mathcal{K}}(\bar{x})$ and let $(\tau_n)_{n\in\mathbb{N}}$ and $(h_n)_{n\in\mathbb{N}}$ be as in Remark 5(i). Then, for n large enough we have

$$f(\bar{x}) \leq f(\bar{x} + \tau_n h_n) = f(\bar{x}) + \tau_n \langle \nabla f(\bar{x}), h_n \rangle + \tau_n \|h_n\| \varepsilon_{\bar{x}}(\tau_n h_n),$$

which yields

$$\langle \nabla f(\bar{x}), h_n \rangle + \|h_n\| \varepsilon_{\bar{x}}(\tau_n h_n) \geq 0.$$

Therefore, letting $n \to \infty$, we get

$$\langle \nabla f(\bar{x}), h \rangle \geq 0,$$

from which the result follows. $\square$

Now, we need the following definition

Definition 5. *The normal cone* $N_{\mathcal{K}}(\bar{x})$ *to* $\mathcal{K}$ *at* $\bar{x} \in \mathcal{K}$ *is defined by*

$$N_{\mathcal{K}}(\bar{x}) := (T_{\mathcal{K}})^{\circ}.$$

Corollary 1. *Suppose that* $\bar{x} \in \mathcal{K}$ *a local solution to* (P) *and that* f *is differentiable at* $\bar{x}$. *Then,*

$$-\nabla f(\bar{x}) \in N_{\mathcal{K}}(\bar{x}).$$

◇ [Optimization problems with equality and inequality constraints] We suppose now that the constraint system is given by

$$\mathcal{K} := \{x \in \mathbb{R}^n \mid g_i(x) = 0, \ \forall\, i = 1, \ldots, m, \ \ h_j(x) \leq 0 \ \forall\, j = 1, \ldots, p\},$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \ldots, m)$, and $h_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 1, \ldots, p)$ are differentiable functions. In this case, Problem (P) is usually written as

$$\begin{aligned} &\min \ f(x) \\ &\text{s.t.} \ \ g_i(x) = 0, \ \forall \, i = 1, \ldots, m, \\ &\qquad\ h_j(x) \leq 0, \ \forall \, j = 1, \ldots, p. \end{aligned}$$

Let $\bar{x} \in \mathcal{K}$ and set

$$I(\bar{x}) := \{ j \in \{1, \ldots, p\} \mid h_j(\bar{x}) = 0 \},$$

for the set of indexes of active inequality constraints at $\bar{x}$.

Let us study the tangent cone $T_{\mathcal{K}}(\bar{x})$.

Lemma 5. *The following inclusion holds*

$$T_{\mathcal{K}}(\bar{x}) \subseteq \left\{ h \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle \nabla g_i(\bar{x}), h \rangle = 0 \;\; \forall\, i = 1, \ldots, m, \\ \langle \nabla h_j(\bar{x}), h \rangle \leq 0 \;\; \forall\, j \in I(\bar{x}) \end{array} \right\}. \tag{26}$$

Proof. Let $h \in T_K(\bar{x})$ and let $(\tau_n)_{n\in\mathbb{N}}$ and $(h_n)_{n\in\mathbb{N}}$ be as in Remark 5(i). Then, for every $i = 1, \ldots, m$, we have

$$\begin{aligned} 0 = g_i(\bar{x} + \tau_n h_n) &= g_i(\bar{x}) + \tau_n \langle \nabla g_i(\bar{x}), h_n \rangle + \tau_n \|h_n\| \varepsilon_{g_i,\bar{x}}(\tau_n h_n) \\ &= \tau_n \langle \nabla g_i(\bar{x}), h_n \rangle + \tau_n \|h_n\| \varepsilon_{g_i,\bar{x}}(\tau_n h_n). \end{aligned}$$

Then, dividing by τ_n and letting $n \to \infty$, we get

$$\langle \nabla g_i(\bar{x}), h \rangle = 0.$$

Similarly, for every $j \in I(\bar{x})$,

$$\begin{aligned} 0 \geq h_j(\bar{x} + \tau_n h_n) &= h_j(\bar{x}) + \tau_n \langle \nabla h_j(\bar{x}), h_n \rangle + \tau_n \|h_n\| \varepsilon_{h_j,\bar{x}}(\tau_n h_n) \\ &= \tau_n \langle \nabla h_j(\bar{x}), h_n \rangle + \tau_n \|h_n\| \varepsilon_{h_j,\bar{x}}(\tau_n h_n). \end{aligned}$$

Then, dividing by τ_n and letting $n \to \infty$, we get

$$\langle \nabla h_j(\bar{x}), h \rangle \leq 0.$$

The result follows. $\square$

Unfortunately, the converse inclusion does not always hold.

Example: Consider the set

$$\mathcal{K} := \left\{ (x, y) \in \mathbb{R}^2 \mid x^2 - y^3 = 0, \;\; -x \leq 0 \right\}.$$

Then, it is easy to see that $T_K((0,0)) = \{(0, \gamma) \mid \gamma \geq 0\}$ and the right hand side of (26) is given by

$$\left\{ (h_1, h_2) \in \mathbb{R}^2 \mid h_1 \geq 0 \right\}.$$

Definition 6. (i) *We say that the constraint functions* g_i $(i = 1, \ldots, m)$ *and* h_j $(j = 1, \ldots, p)$ *are qualified at* $\bar{x}$ *if*

$$T_{\mathcal{K}}(\bar{x}) = \left\{ h \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \langle \nabla g_i(\bar{x}), h \rangle = 0 \;\; \forall \, i = 1, \ldots, m, \\ \langle \nabla h_j(\bar{x}), h \rangle \leq 0 \;\; \forall \, j \in I(\bar{x}) \end{array} \right\}. \tag{27}$$

(ii) *Any condition ensuring that the constraint functions are qualfied is called a constraint qualification condition.*

Remark 6. In general, the qualified character of the constraints is not a geometrical property of the set $\mathcal{K}$. Indeed, consider the set

$$\mathcal{K} = \left\{(x, y) \in \mathbb{R}^2 \mid y - x^2 = 0\right\},$$

which can also be written as

$$\mathcal{K} = \left\{(x, y) \in \mathbb{R}^2 \mid y^2 - x^4 = 0, \quad -y \leq 0\right\}.$$

Then, it is easy to check that the constraint functions are qualified at $(0, 0)$ in the first formulation but they are not qualified at $(0, 0)$ in the second one.

$\diamond$ [The Karush-Kuhn-Tucker theorem] The main result here is the following first order optimality condition.

Theorem 13. [Karush-Kuhn-Tucker] *Let $\bar{x} \in \mathcal{K}$ be a local solution to (P). Assume that f, g_i $(i = 1, \ldots, m)$, h_j $(j = 1, \ldots, p)$ are $\mathcal{C}^1$ and that the constraint functions are qualified at $\bar{x}$. Then, there exist $(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ and $(\mu_1, \ldots, \mu_p) \in \mathbb{R}^p$ such that*

$$\begin{aligned} \nabla f(\bar{x}) + \textstyle\sum_{i=1}^m \lambda_i \nabla g_i(\bar{x}) + \sum_{j=1}^p \mu_j \nabla h_j(\bar{x}) = 0, \\ \text{and} \quad \forall\, j \in \{1, \ldots, p\} \ \text{ we have } \mu_j \geq 0, \quad \text{and} \quad \mu_j h_j(\bar{x}) = 0. \end{aligned} \tag{28}$$

Equivalently, setting $g(x) = (g_1(x), \ldots, g_m(x))$ and $h = (h_1(x), \ldots, h_p(x))$ for all $x \in \mathbb{R}^n$, there exists $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ such that

$$\begin{aligned} \nabla f(\bar{x}) + Dg(\bar{x})^\top \lambda + Dh(\bar{x})^\top \mu = 0, \\ \text{and} \quad \forall\, j \in \{1, \ldots, p\} \ \text{ we have } \mu_j \geq 0, \quad \text{and} \quad \mu_j h_j(\bar{x}) = 0. \end{aligned} \tag{29}$$

Proof. Since the constraint functions are qualified at $\bar{x}$, by Lemma 4 we have that

$$N_{\mathcal{K}}(\bar{x}) = \left\{ \sum_{i=1}^m \lambda_i \nabla g_i(\bar{x}) + \sum_{j \in I(\bar{x})} \mu_j \nabla h_j(\bar{x}) \;\middle|\; \begin{array}{l} \lambda_i \in \mathbb{R} \ \ \forall\, i = 1, \ldots, m, \\ \mu_j \geq 0 \ \ \forall\, j \in I(\bar{x}) \end{array} \right\}.$$

Then, by Corollary 1, there exist $\lambda_i \in \mathbb{R}$ $(i = 1, \ldots, m)$ and $\mu_j \geq 0$ $(j \in I(\bar{x}))$ such that

$$-\nabla f(\bar{x}) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) + \sum_{j \in I(\bar{x})} \mu_j \nabla h_j(\bar{x}).$$

Relation (28) follows by setting $\mu_j = 0$ for all $j \in \{1, \ldots, p\} \setminus I(\bar{x})$. $\square$

Let $g(x) = (g_1(x), \ldots, g_m(x))$ and $h(x) = (h_1(x), \ldots, h_p(x))$. The Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined by

$$L(x, \lambda, \mu) = f(x) + \langle \lambda, g(x) \rangle + \langle \mu, h(x) \rangle,$$

and, at a local solution $\bar{x}$, the optimality system (29) reads: there exists $(\lambda, \mu) \in \mathbb{R}^{m+p}$ such that

$$\begin{aligned} & \nabla_x L(\bar{x}, \lambda, \mu) = 0, \\ \text{and} \quad & \forall\ j \in \{1, \ldots, p\} \ \text{ we have } \mu_j \geq 0, \quad \text{and} \quad \mu_j h_j(\bar{x}) = 0. \end{aligned} \tag{30}$$

System (30) is usually called KKT system and (λ, μ) are called Lagrange multipliers.

[The KKT system as a sufficient condition for convex problems] The KKT condition is also sufficient for convex problems.

Proposition 6. *Suppose that f is convex and $\mathcal{C}^1$, $g_i(x) = \langle a_i, x\rangle + b_i$, for $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ $(i = 1, \dots, m)$, and h_j $(j = 1, \dots, p)$ is convex and $\mathcal{C}^1$. Moreover, assume that there exists $(\bar{x}, \lambda, \mu) \in \mathbb{R}^{n+m+p}$ such that $\bar{x} \in \mathcal{K}$ and the KKT system* (30) *holds at $(\bar{x}, \lambda, \mu)$. Then $\bar{x}$ is a global solution to (P).*

Proof. Note that $L(\bar{x}, \lambda, \mu) = f(\bar{x})$ and that $L(\cdot, \lambda, \mu)$ is convex. Then, for any $x \in \mathcal{K}$,

$$\begin{aligned} f(\bar{x}) = L(\bar{x}, \lambda, \mu) + \langle \nabla_x L(\bar{x}, \lambda, \mu), x - \mu\rangle &\leq L(\bar{x}, \lambda, \mu) \\ &= f(x) + \textstyle\sum_{i=1}^m \lambda_i g_i(x) \\ &\quad + \textstyle\sum_{j=1}^p \mu_j h_j(x) \\ &\leq f(x), \end{aligned}$$

which implies that $\bar{x}$ is a global solution to (P). $\square$

Note that no qualification condition is needed in Proposition 6.

Remark 7. (i) [Equality constraints only] Suppose that we only have equality constraints. Then, if $\bar{x} \in \mathcal{K}$ solves (P) and the constraint functions are qualified at $\bar{x}$, then there exists $\lambda \in \mathbb{R}^m$ such that

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\bar{x}) = 0.$$

In this case the Lagrangian is given by $L(x, \lambda) = f(x) + \langle \lambda, g(x) \rangle$, and the previous relation can be written as

$$\nabla_x L(\bar{x}, \lambda) = 0. \tag{31}$$

(ii) [Inequality constraints only] Suppose that we only have inequality constraints. Then, if $\bar{x} \in \mathcal{K}$ solves (P) and the constraint functions are qualified at $\bar{x}$, then there exists $\mu \in \mathbb{R}^p$ such that

$$\begin{aligned} &\nabla f(\bar{x}) + \textstyle\sum_{j=1}^{p} \mu_j \nabla h_j(\bar{x}) = 0, \\ &\mu_j \geq 0 \ \text{ and } \ \mu_j h_j(\bar{x}) = 0 \ \ \forall\, j = 1, \ldots, p. \end{aligned} \tag{32}$$

In this case the Lagrangian is given by $L(x, \mu) = f(x) + \langle \mu, h(x) \rangle$, and the previous relation can be written as

$$\nabla_x L(\bar{x}, \mu) = 0,$$
$$\mu_j \geq 0 \text{ and } \mu_j h_j(\bar{x}) = 0 \ \forall\, j = 1, \ldots, p.$$

(iii) [Maximization problems] Consider the maximization problem

$$\begin{aligned} \max \ & f(x) \\ \text{s.t.} \ & g_i(x) = 0, \ \forall\, i = 1, \ldots, m, \\ & h_j(x) \leq 0, \ \forall\, j = 1, \ldots, p. \end{aligned}$$

In this case, if $\bar{x}$ is a local solution and the constraint functions are qualified at $\bar{x}$, then there exists $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ such that

$$\begin{aligned} & -\nabla f(\bar{x}) + Dg(\bar{x})^\top \lambda + Dh(\bar{x})^\top \mu = 0, \\ \text{and} \quad & \forall\, j \in \{1, \ldots, p\} \text{ we have } \mu_j \geq 0, \ \text{ and } \ \mu_j h_j(\bar{x}) = 0. \end{aligned} \tag{33}$$

It is important to notice that, differently from the case where only equality constraints were present, when inequality constraints are present the optimality system for local solutions of the minimization and maximization problems differ. The coincidence of both optimality systems is generally false and it is a specific feature of the problem with equality constraints and of the unconstrained problem.

(iv) The following example shows that the assumption on the qualification of the constraints plays an important role in the necessary condition. Consider the problem

$$\begin{aligned} &\min\ y \\ &\text{s.t.}\ \ x^2 - y^3 = 0, \\ &\qquad -x \leq 0. \end{aligned}$$

In this case, $(\bar{x}, \bar{y}) = (0,0)$ is a global solution and (29) reads

$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \lambda \begin{pmatrix} 2x \\ -3y^2 \end{pmatrix}\Bigg|_{(x,y)=(0,0)} + \mu \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which is impossible. Note that $\nabla g_1(0,0) = (0,0)$ and $\nabla h_1(0,0) = (-1,0)$. Therefore,

$$\{h \in \mathbb{R}^2 \mid \langle \nabla g_1(0,0), h\rangle = 0,\ \ \langle \nabla h_1(0,0), h\rangle \leq 0\} = \{h \in \mathbb{R}^2 \mid h_1 \geq 0\},$$

and $T_{\mathcal{K}}(0,0) = \{h \in \mathbb{R}^2 \; | h_1 = 0, \;\; h_2 \geq 0\}$. Thus, g_1 and h_1 are not qualified at $(0,0)$.

◇ [On constraint qualifications] Let us now comment on some well-known constraint qualifications. The first two conditions are easy to check but they can be applied only when $\mathcal{K}$ is convex.

- [Affine constraints] If $g_i(x) = \langle a_i, x\rangle + b_i$ for all $i = 1,\dots,m$, and $h_j(x) = \langle a'_j, x\rangle + b'_j$ for all $j = 1,\dots,m$, where $a_i \in \mathbb{R}^n$, $a'_j \in \mathbb{R}^n$, $b_i \in \mathbb{R}$ and $b'_j \in \mathbb{R}$, then the constraint functions are qualified at every $x \in \mathcal{K}$.

- [Slater condition] If $g_i(x) = \langle a_i, x\rangle + b_i$ for all $i = 1,\dots,m$, with $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$, and, for all $j = 1,\dots,p$, the function h_j is convex, then the constraint functions are qualified at every $x \in \mathcal{K}$ if the following condition holds:

$$\begin{aligned} &\text{There exists } x_0 \in \mathbb{R}^n \text{ such that } \; g_i(x_0) = 0 \;\; \forall i = 1,\dots,m, \\ &\text{and } \; h_j(x_0) < 0 \;\; \forall j = 1,\dots,p. \end{aligned} \qquad (SLC)$$

The following two conditions are more general but, at the same time, they are more difficult to check.

- [Mangasarian-Fromovitz] The constraint functions are qualified at $\bar{x} \in \mathcal{K}$ if the following conditions hold:

  (i) the set of vectors $\{\nabla g_1(\bar{x}), \ldots, \nabla g_m(\bar{x})\}$ are linearly independent.

  (ii) there exists $\bar{d} \in \mathrm{Ker}(Dg(\bar{x}))$ such that $\qquad (MF)$

$$\nabla h_j(\bar{x}) \cdot \bar{d} < 0 \quad \forall \; j \in I(\bar{x}).$$

- [Linear independence constraint qualification] The constraint functions are qualified at $\bar{x} \in \mathcal{K}$ if the following condition holds: the vectors

$$\left\{ (\nabla g_i(\bar{x}))_{i=1}^{m}, (\nabla h_j(\bar{x}))_{j \in I(\bar{x})} \right\} \text{ are linearly independent.} \qquad (LICQ)$$

Remark 8. [On $(LICQ)$ and the uniqueness of Lagrange multipliers] It is easy to check that $(LICQ)$ implies (MF), but the converse is false. Moreover, it is easy to check that $(LICQ)$ implies that there exist at most one $(\lambda, \mu) \in \mathbb{R}^{m+p}$ such that

(28) holds. In general, (MF) only implies that the set of $(\lambda, \mu) \in \mathbb{R}^{m+p}$ such that (28) is a compact set.

◇ [Some examples] In the first example, we consider a problem where $\mathcal{K}$ is defined by equality constraints only.

Example: Let us consider the problem

$$\begin{aligned} &\min \; xy \\ &\text{s.t.} \;\; x^2 + (y+1)^2 = 1. \end{aligned}$$

In this case $f : \mathbb{R}^2 \to \mathbb{R}$, is given by $f(x, y) = xy$, and $\mathcal{K} = \{(x, y) \in \mathbb{R}^2 \mid g(x, y) = 0\}$, with $g : \mathbb{R}^2 \to \mathbb{R}$ being given by $g(x, y) = x^2 + (y+1)^2 - 1$. Note that $\mathcal{K}$ is given by the cercle centered at $(0, -1)$ with radius 1. Hence, $\mathcal{K}$ is a compact subset of $\mathbb{R}^2$. The function f being continuous, the Weierstrass theorem implies that the optimization problem has at least one solution $(\bar{x}, \bar{y}) \in \mathcal{K}$.
Let us study the qualification condition (MF) (when only equality constraints are present). We have $\nabla g(x, y) = (2x, 2(y+1))$ and, hence, $\nabla g(x, y) = 0$ iff

$x = 0, y = -1$. Thus, every $(x, y) \in \mathbb{R}^2 \setminus \{(0, -1)\}$ satisfies (MF). Since $(0, -1) \notin \mathcal{K}$ we deduce that (MF) holds for every $(x, y) \in \mathcal{K}$.
The Lagrangian $L : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$ of this problem is given by

$$L(x, y, \lambda) = xy + \lambda(x^2 + (y + 1)^2 - 1).$$

By Theorem 13, we have the existence of $\lambda \in \mathbb{R}$ such that (31) holds at $(\bar{x}, \bar{y}, \lambda)$. Now,

$$\begin{aligned} \nabla_{(x,y)} L(\bar{x}, \bar{y}, \lambda) = 0 \quad &\Leftrightarrow \quad \begin{array}{rcl} \bar{y} + 2\lambda\bar{x} &=& 0, \\ \bar{x} + 2\lambda(\bar{y} + 1) &=& 0, \end{array} \\ &\Leftrightarrow \quad \begin{array}{rcl} \bar{y} &=& -2\lambda\bar{x}, \\ (1 - 4\lambda^2)\bar{x} &=& -2\lambda. \end{array} \end{aligned} \tag{34}$$

Now, $1 - 4\lambda^2 = 0$ iff $\lambda = 1/2$ or $\lambda = -1/2$, and both cases contradict the last equality above. Therefore, $1 - 4\lambda^2 \neq 0$ and, hence,

$$\bar{x} = \frac{2\lambda}{4\lambda^2 - 1} \quad \text{and} \quad \bar{y} = \frac{-4\lambda^2}{4\lambda^2 - 1}.$$

Since $\nabla_\lambda L(\bar{x}, \bar{y}, \lambda) = g(\bar{x}, \bar{y}) = 0$, we get

$$\begin{aligned}
&\left(\tfrac{2\lambda}{4\lambda^2-1}\right)^2 + \left(1 - \tfrac{4\lambda^2}{4\lambda^2-1}\right)^2 = 1, \\
&\Leftrightarrow 4\lambda^2 + 1 = (4\lambda^2 - 1)^2 \\
&\Leftrightarrow (4\lambda^2 - 1)^2 - (4\lambda^2 - 1) - 2 = 0,
\end{aligned}$$

which yields

$$4\lambda^2 - 1 = \tfrac{1+\sqrt{9}}{2} \quad \text{or} \quad 4\lambda^2 - 1 = \tfrac{1-\sqrt{9}}{2}$$
$$\text{i.e. } \lambda^2 = 3/4 \quad \text{or} \quad \lambda^2 = 0.$$

If $\lambda = 0$, then (34) yields $\bar{x} = \bar{y} = 0$. If $\lambda = \sqrt{3}/2$ we get $\bar{x} = \sqrt{3}/2$ and $\bar{y} = -3/2$. If $\lambda = -\sqrt{3}/2$ we get $\bar{x} = -\sqrt{3}/2$ and $\bar{y} = -3/2$. Thus, the candidates to solve the problem are

$$(\bar{x}_1, \bar{y}_1) = (0, 0), \quad (\bar{x}_2, \bar{y}_2) = (\sqrt{3}/2, -3/2) \text{ and } (\bar{x}_3, \bar{y}_3) = (-\sqrt{3}/2, -3/2).$$

We have $f(\bar{x}_1, \bar{y}_1) = 0$, $f(\bar{x}_2, \bar{y}_2) = -3\sqrt{3}/4$ and $f(\bar{x}_3, \bar{y}_3) = 3\sqrt{3}/4$. Therefore, the global solution is $(\bar{x}_2, \bar{y}_2)$. $\square$

In the second example, we consider a problem where $\mathcal{K}$ is defined by inequality constraints only.

Example: Consider the problem

$$\begin{aligned} \min\ & 4x^2 + y^2 - x - 2y \\ \text{s.t.}\ & 2x + y \leq 1, \\ & x^2 \leq 1. \end{aligned}$$

Note that, setting $Q = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}$ and $c = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$, in the notation for (P) we have

$$\begin{aligned} f(x) &= \tfrac{1}{2}\langle Q(x,y),(x,y)\rangle + c^\top(x,y) \\ h_1(x,y) &= 2x + y - 1 \\ h_2(x,y) &= x^2 - 1. \end{aligned}$$

Note that the feasible set is nonempty, convex, closed and the Slater condition is satisfied (for instance $h_1(0,0) < 0$, $h_2(0,0) < 0$). Moreover, f is continuous, strictly convex, differentiable and infinity at the infinity (Q is positive definite). We deduce that there exists a unique solution $(\bar{x}, \bar{y}) \in \mathbb{R}^2$ to problem (P) and $(\bar{x}, \bar{y})$ is characterized by the KKT system (32). A point $(\bar{x}, \bar{y}, \mu_1, \mu_2)$ satisfies (32) iff

$$Q \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} + c + \mu_1 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + \mu_2 \begin{pmatrix} 2\bar{x} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\mu_1 \geq 0, \quad \mu_2 \geq 0, \quad \mu_1 h_1(\bar{x}, \bar{y}) = 0, \quad \mu_2 h_2(\bar{x}, \bar{y}) = 0,$$

iff

$$8\bar{x} + 2\mu_1 + 2\mu_2\bar{x} = 1$$

$$2\bar{y} + \mu_1 = 2,$$

$$\mu_1 \geq 0, \quad \mu_2 \geq 0, \quad \mu_1 h_1(\bar{x}, \bar{y}) = 0, \quad \mu_2 h_2(\bar{x}, \bar{y}) = 0,$$

Case 1: $\mu_1 = \mu_2 = 0$. We obtain $(\bar{x}, \bar{y}) = (1/8, 1) \notin \mathcal{K}$.
Case 2: $\mu_1 > 0$, $\mu_2 > 0$. In this case, we obtain $2\bar{x} + \bar{y} = 1$, $\bar{x}^2 = 1$, which gives $(\bar{x}, \bar{y}) = (1, -1)$ or $(\bar{x}, \bar{y}) = (-1, 3)$. In the first case, we should have

$$2\mu_1 + 2\mu_2 = -7,$$

which is impossible, because $\mu_1 > 0$ and $\mu_2 > 0$. In the second case, we should have $6 + \mu_1 = 2$, which is also impossible.

Case 3: $\mu_1 = 0$, $\mu_2 > 0$. We obtain $\bar{x}^2 = 1$ and $\bar{y} = 1$, which gives $(\bar{x}, \bar{y}) = (1, 1)$ or $(\bar{x}, \bar{y}) = (-1, 1)$. If $(\bar{x}, \bar{y}) = (1, 1)$ we should have $8 + 2\mu_2 = 1$, which is impossible. If $(\bar{x}, \bar{y}) = (-1, 1)$, we should have $-8 - 2\mu_2 = 1$, which is also impossible.

Case 4: $\mu_1 > 0$, $\mu_2 = 0$. We obtain

$$\begin{aligned} 2\bar{x} + \bar{y} &= 1, \\ 8\bar{x} + 2\mu_1 &= 1 \\ 2\bar{y} + \mu_1 &= 2 \end{aligned}$$

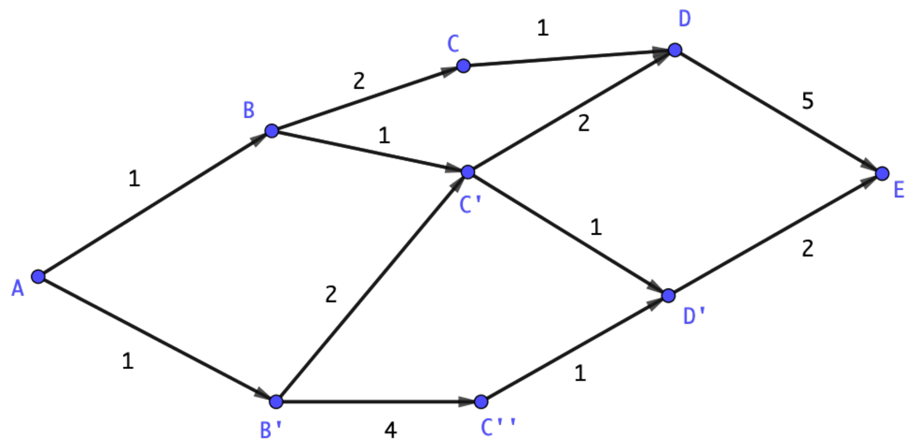
which implies

$$\begin{aligned} 2\bar{x} + \bar{y} &= 1, \\ 8\bar{x} - 4\bar{y} &= -3, \end{aligned}$$

which gives $(\bar{x}, \bar{y}) = (1/16, 7/8)$ and $\mu_1 = 1/4$. This point $(\bar{x}, \bar{y})$ belongs to $\mathcal{K}$. Therefore, we conclude that $(\bar{x}, \bar{y}, \mu_1, \mu_2) = (1/16, 7/8, 1/4, 0)$ is the unique solution to the KKT system, and, hence, $(\bar{x}, \bar{y}) = (1/16, 7/8)$ is the unique global solution to (P).

Dynamic programming in discrete time: the finite horizon case

◇ [Introduction: Shortest path between two vertices A and E on a graph] Consider a salesman who has to go from city A to city E according to the following graph



The data in the graph G are

- For every vertex x we denote by $\Gamma(x)$ the set of successors. For instance, $\Gamma(B) = \{C, C'\}$.

- The "travel time" (in hours) $F(x, x')$ of each $V' \in \Gamma(x)$. For instance, $F(C', D) = 2$.

- In order to compute the shortest path one could enumerate all the paths and choose the one with the smaller travel time.

- However, it is more convenient to notice that if a path is optimal between A and E, and this path passes through a vertex x, then the "sub-path" between x and E will be optimal for the shortest path problem between x and E.

- This suggest to parametrize the optimal travel time by the departure point. Let us set $V(x)$ as the smallest time needed to go from x to E. Then,

$$\begin{gathered}
V(E) = 0, \\
V(D) = 5, \quad V(D') = 2, \\
V(C) = 6, \quad V(C') = \min\{2 + V(D), 1 + V(D')\} = 3, \quad V(C'') = 3, \\
V(B) = \min\{2 + V(C), 1 + V(C')\} = 4, \quad V(B') = \min\{2 + V(C'), 4 + V(C'')\} = 5, \\
V(A) = \min\{1 + V(B), 1 + V(B')\} = 5.
\end{gathered}$$

- We deduce that the shortest travel time is $V(A) = 5$ and the shortest path is $ABC'D'E$.

◇ [The general framework] We are interested in the problem

$$\sup_{(x_t)} \left\{ \sum_{t=0}^{T-1} F_t(x_t, x_{t+1}) + F_T(x_T) \right\}, \qquad (P_{fh})$$

where

- $x_t \in X$ for all $t = 0, \ldots, T$. The set X is called the state space.
- $x_{t+1} \in \Gamma_t(x_t)$ for all $t = 0, \ldots, T-1$. For all $x \in X$, and $t = 0, \ldots, T-1$, $\Gamma_t(x)$ is a nonempty subset of X.
- $F_t(x_t, x_{t+1})$ denotes the profit at time t for the pair (x_t, x_{t+1}), and $F_T(x_T)$ denotes the final profit for the final state x_T. Notice that redefining $F_{T-1}(x_{T-1}, x_T)$ as

$$\tilde{F}_{T-1}(x_{T-1}, x_T) = F_{T-1}(x_{T-1}, x_T) + F_T(x_T),$$

  we can assume, without loss of generality, that $F_T \equiv 0$.

◇ [Dynamic Programming relation] As in the shortest path problem, it is a good idea to parametrize problem (P_{fh}). Given, $(k, x) \in \{1, \ldots, T-1\} \times X$, the value function at (k, x) is defined as

$$V(k, x) = \sup \left\{ \sum_{t=k}^{T-1} F_t(x_t, x_{t+1}) \mid x_k = x, \quad x_{t+1} \in \Gamma_t(x_t) \;\; \forall\, t = k, \dots, T-1 \right\}. \tag{35}$$

For $k = T$ and $x \in X$, we set $V(T, x) = 0$ (recall that we have a zero final cost).

The main result here is the following

Theorem 14. [Dynamic Programming] *The following relations holds:*

$$\begin{aligned} V(k, x) &= \sup \left\{ F_k(x, x_{k+1}) + V(k+1, x_{k+1}) \mid x_{k+1} \in \Gamma_k(x) \right\}, \\ &\quad \forall\, k = 0, \dots, T-1, \quad x \in X. \\ V(T, x) &= 0 \quad \forall\, x \in X. \end{aligned} \tag{36}$$

Proof. Let $(x_k, x_{k+1}, \dots, x_T)$ be a feasible sequence for the problem defining

$V(k,x)$, i.e. $x_k = x$ and $x_{t+1} \in \Gamma_t(x_t)$, for all $t = k, \ldots, T-1$. Then,

$$V(k,x) \geq \sum_{t=k}^{T-1} F_t(x_t, x_{t+1}) = F_t(x, x_{k+1}) + \sum_{t=k+1}^{T-1} F_t(x_t, x_{t+1}).$$

Using that the previous inequality holds for any $(x_{k+1}, \ldots, x_T)$ such that $x_{t+1} \in \Gamma_t(x_t)$, for all $t = k+1, \ldots, T-1$, by taking the supremum of the right hand side with respect to those $(x_{k+1}, x_{k+2}, \ldots, x_T)$, we get

$$V(k,x) \geq F_t(x, x_{k+1}) + V(k+1, x_{k+1})$$

Therefore, by taking the supremum with respect to $x_{k+1} \in F(x_k)$, we get

$$V(k,x) \geq \sup \left\{ F_k(x, x_{k+1}) + V(k+1, x_{k+1}) \mid x_{k+1} \in \Gamma_k(x) \right\}.$$

Conversely, for any $(x'_k, x'_{k+1}, \ldots, x'_T)$ such that $x'_k = x$ and $x'_{t+1} \in \Gamma_t(x'_t)$, for all

$t = k, \ldots, T-1$, we have

$$\begin{aligned} \textstyle\sum_{t=k}^{T-1} F_t(x'_t, x'_{t+1}) &= F_t(x', x'_{k+1}) + \textstyle\sum_{t=k+1}^{T-1} F_t(x'_t, x'_{t+1}) \\ &\leq F_t(x, x'_{k+1}) + V(k+1, x'_{k+1}) \\ &\leq \sup \left\{ F_k(x, x_{k+1}) + V(k+1, x_{k+1}) \;\middle|\; x_{k+1} \in \Gamma_k(x) \right\}. \end{aligned}$$

Using that $(x'_k, x'_{k+1}, \ldots, x'_T)$ is an arbitrary admissible sequence, by taking the supremum on the left hand side, we get

$$V(k, x) \leq \sup \{F_k(x, x_{k+1}) + V(k+1, x_{k+1}) \mid x_{k+1} \in \Gamma_k(x)\}.$$

The result follows. $\square$

◇ [Backward solution] By Theorem 14, we can solve backward for V using relations (36). In particular, (36) characterizes the value function V. Now, et us assume that for all $(k, x) \in \{0, \ldots, T-1\} \times X$ there exists $s(k, x) \in \Gamma_k(x)$ such that

$$V(k, x) = F_k(x, s(k, x)) + V(k+1, s(k, x)).$$

Then, by the very definition we have that $(x_k, x_{k+1}, \ldots, x_T)$, with

$$x_k = x \text{ and } x_{t+1} := s(t, x_t) \quad \forall\, t = k, \ldots, T-1,$$

solves the problem in the right hand side of (35). In particular, this problem admits a solution if

- $X \subseteq \mathbb{R}^n$.
- For all $(k, x) \in \{0, \ldots, T-1\} \times X$ the set $\Gamma_t(x)$ is nonempty and compact,
- and F_t and $V(t, \cdot)$ are continuous for all $t = k, \ldots, T-1$.

Remark 9. Under the previous assumptions, $V(t, \cdot)$ $(t = 0, \ldots, T-1)$ is continuous if F_k and Γ_k are continuous for all $k = t \ldots, T-1$. Concerning the latter continuity, we recall that the correspondence Γ_k is called continuous at $x \in X$ if the following conditions hold:

(i) [Lower semicontinuity] For every $y \in \Gamma(x)$ and every sequence $(x_n)_{n\in\mathbb{N}}$ such that $x_n \to x$, as $n \to \infty$, there exists a sequence $(y_n)_{n\in\mathbb{N}}$ such that $y_n \in \Gamma_k(x_n)$, for all

$n \in \mathbb{N}$, and $y_n \to y$ as $n \to \infty$.

(ii) [Upper semicontinuity] If $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are two sequences such that $x_n \to x$, as $n \to \infty$, and $y_n \in \Gamma_k(x_n)$, for all $n \in \mathbb{N}$, then $(y_n)_{n\in\mathbb{N}}$ has a subsequence which converges to a point $y \in \Gamma_k(x)$.

Example: Let $y \geq 0$ and consider the problem

$$\inf\left\{\sum_{i=1}^{N} x_i^2 \mid \sum_{i=1}^{N} x_i = y, \quad x_i \geq 0 \quad \forall\, i = 1, \ldots, N\right\}. \tag{37}$$

- Let us solve (37) by using nonlinear programming tools. Note first that the cost function is continuous, strictly convex and the feasible set is nonempty, convex and compact. Therefore, there exists a unique solution $\bar{x}$ to (37). Since the set $\mathcal{K}$ is defined by affine-constraints, $\bar{x}$ is characterized by the KKT system. Consider the Lagrangian $L : \mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}$ defined by

$$L(x, \lambda, \mu) = \sum_{i=1}^{N} x_i^2 + \lambda\left(\sum_{i=1}^{N} x_i - y\right) - \sum_{i=1}^{N} \mu_i x_i.$$

Then, $(\bar{x}_1, \ldots, \bar{x}_N)$ is characterized by the existence of $\lambda \in \mathbb{R}$ and $\mu \in \mathbb{R}^N$ such that

$$\begin{aligned} &\partial_{x_i} L(x, \lambda, \mu) = 2\bar{x}_i + \lambda - \mu_i = 0 \quad \forall\, i = 1, \ldots, N. \\ &\textstyle\sum_{i=1}^{N} \bar{x}_i = y \\ &\mu_i \geq 0 \text{ and } \quad \mu_i x_i = 0 \quad \forall\, i = 1, \ldots, N. \end{aligned} \tag{38}$$

If $y = 0$, the only feasible point is $x_i = 0$ for all $i = 0, \ldots, N$ and, hence, $\bar{x} = (0, \ldots, 0)$ is the solution. If $y > 0$ and there exists $\hat{i} \in \{1, \ldots, N\}$ such that $\bar{x}_{\hat{i}} = 0$, then, the first equation in (38) yields $\lambda = \mu_{\hat{i}} \geq 0$. On the other hand, there must exist $i' \in \{1, \ldots, N\}$ such that $\bar{x}_{i'} > 0$ (otherwise $\sum_i \bar{x}_i = y$, $\bar{x}_i \geq 0$ for all $i \in \{1, \ldots, N\}$ would not hold). The first and third equation in (38) imply that $\lambda = -2\bar{x}_{i'} < 0$, which is a contradiction. As a consequence, $\bar{x}_i > 0$ for all $i \in \{1, \ldots, N\}$. Then, the first and the second conditions in (38) yield $\bar{x}_i = y/N$ for all $i \in \{1, \ldots, N\}$. Thus,

$\bar{x} = (y/N, \ldots, y/N)$ is the solution to the problem and the optimal cost is y^2/N.

- Let us find the same conclusion by using dynamic programming techniques. Let us define

$$V(k, y) := \inf \left\{ \sum_{i=k}^{N} x_i^2 \;\middle|\; \sum_{i=k}^{N} x_i = y, \quad x_i \geq 0 \quad \forall\, i = k, \ldots, N \right\}.$$

Note that we are interested in $V(1, y)$. Notice that the problem defining $V(1, y)$ has not the form discussed in the previous subsection. However, arguing as in the proof of Theorem 14, we can prove that (exercise)

$$\begin{aligned} V(k, y) &= \inf \left\{ x_k^2 + V(k+1, y - x_k) \mid 0 \leq x_k \leq y \right\} \;\; \forall\, y \geq 0, \\ V(N, y) &= y^2 \;\; \forall\, y \geq 0. \end{aligned} \tag{39}$$

Solving backwards, we get

$$\begin{aligned} V(N-1, y) &= \inf \left\{ x_{N-1}^2 + V(N, y - x_{N-1}) \mid 0 \leq x_{N-1} \leq y \right\} \\ &= \inf \left\{ x_{N-1}^2 + (y - x_{N-1})^2 \mid 0 \leq x_{N-1} \leq y \right\}. \end{aligned}$$

From the last expression we get $s(N-1, y) = y/2$ and $V(N-1, y) = y^2/2$.

Similarly,

$$
\begin{aligned}
V(N-2, y) &= \inf \left\{x_{N-2}^2 + V(N-1, y - x_{N-2}) \mid 0 \leq x_{N-2} \leq y\right\} \\
&= \inf \left\{x_{N-2}^2 + \tfrac{(y - x_{N-2})^2}{2} \mid 0 \leq x_{N-2} \leq y\right\},
\end{aligned}
$$

from which we get $s(N-2, y) = y/3$, and $V(N-2, y) = y^2/3$. Recursively, for all $k = 1, \ldots, N$ we get

$$
s(k, y) = y/(N-k+1) \text{ and } V(k, y) = y^2/(N-k+1).
$$

Thus, we recover $V(1, y) = y^2/N$ for the optimal cost and, adapting the definition of successor according to the dynamic programming principle (39), for the solution we get

$$
\begin{aligned}
&x_1 = s(1, y) = y/N, \\
&x_2 = s(2, y - x_1) = (y - y/N)/(N - 1) = y/N, \\
&x_3 = s(3, y - x_1 - x_2) = y(1 - 2/N)/(N - 2) = y/N, \\
&\vdots \\
&x_N = s\left(N, y - \textstyle\sum_{i=1}^{N-1} x_i\right) = y/N.
\end{aligned}
$$

recovering our previous result.
Another way to tackle the problem is to perform a change of variable in order to be in the framework of the previous subsection.
Let us define

$$z_i := \sum_{j=1}^{N-i} x_j \ \forall \ i = 0, \ldots, N - 1.$$

Then, $z_0 = y$, $z_{N-1} = x_1$ and $z_i - z_{i+1} = x_{N-i}$ for all $i = 0, \ldots, N - 2$. Then, problem (37) can be written as

$$\inf \left\{ \sum_{i=0}^{N-2} (z_{i+1} - z_i)^2 + z_{N-1}^2 \mid z_0 = y, \quad 0 \leq z_{i+1} \leq z_i \;\; \forall\, i = 0, \ldots, N-2 \right\}.$$

By applying Theorem 14 to the problem above, we can also recover the desired solution (exercise).

Dynamic programming in discrete time: the infinite horizon case

◇ [Introduction: A model of optimal growth] We consider an economy where at each period t a single good is produced. This good can be used for consumption or for investment. At period t we denote by c_t the consumption, by i_t the investment, by k_t the capital and by y_t the production of the good. By assuming that $y_t = F(k_t)$, for some production function F, and that the capital depreciates at constant rate $\delta \in (0,1)$, we get the following relations

$$c_t + i_t = y_t = F(k_t) \quad \text{and} \quad k_{t+1} = (1-\delta)k_t + i_t.$$

Thus, setting $f(k_t) = F(k_t) + (1-\delta)k_t$, we obtain

$$c_t = F(k_t) + (1-\delta)k_t - k_{t+1} = f(k_t) - k_{t+1}.$$

Naturally, we impose $c_t \geq 0$ and $k_t \geq 0$, conditions which imply $0 \leq k_{t+1} \leq f(k_t)$.

Finally, the preference over consumption is supposed to have the form

$$\sum_{t=0}^{\infty} \beta^t U(c_i) \quad \text{for some discount factor } \beta \in (0,1),$$

and some utility function U. For a given initial capital $k_0 > 0$, the utility maximization problem is

$$\begin{aligned} &\sup \ \textstyle\sum_{t=0}^{\infty} \beta^t U(f(k_t) - k_{t+1}) \\ &\text{s.t.} \quad 0 \le k_{t+1} \le f(k_t). \end{aligned}$$

◇ [The mathematical framework] For $\beta \in (0,1)$, we consider the problem

$$\sup_{(x_t)} \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}), \tag{P_{ih}}$$

where

- $x_t \in X$ for all $t \ge 0$, with X being a given set and $x_0 \in X$ being prescribed.

- $x_{t+1} \in \Gamma(x_t)$, where, for all $x \in X$, $\Gamma(x)$ is a nonempty subset of X.
- $F(x_t, x_{t+1})$ denotes the profit for the pair (x_t, x_{t+1}). In this section F is assumed to be bounded.

◇ [Dynamic Programming equation] Given $x \in X$, we define the value function

$$V(x) := \sup \left\{ \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \,\middle|\, x_0 = x, \quad x_{t+1} \in \Gamma(x_t) \;\; \forall\, t \geq 0 \right\}.$$

A first important consequence of our assumptions is that $V : X \to \mathbb{R}$ is well-defined. The main result here is the following

Theorem 15. [Dynamic Programming] *For all $x \in X$ we have*

$$V(x) = \sup \left\{ F(x, x_1) + \beta V(x_1) \mid x_1 \in \Gamma(x) \right\}. \tag{40}$$

Proof. The proof is similar to the proof of Theorem 15. Indeed, for any admissible sequence $(x_t)_{t\geq 0}$ we have

$$\begin{aligned} V(x) &\geq \sum_{t=0}^{\infty} \beta^t F(x_t, x_{t+1}) \\ &= F(x, x_1) + \beta \sum_{t=1}^{\infty} \beta^{t-1} F(x_t, x_{t+1}), \\ &= F(x, x_1) + \beta \sum_{t=0}^{\infty} \beta^t F(x_{t+1}, x_{t+2}). \end{aligned} \tag{41}$$

Now, let us fix $x_1 \in \Gamma(x)$. Note that if $(x'_t)_{t\geq 1}$ is an admissible sequence for the problem defining $V(x_1)$, then $x, x'_1, x'_2, x'_3, \ldots$ is an admissible sequence for the problem defining $V(x)$. This remark and (41) imply

$$V(x) \geq F(x, x_1) + \beta V(x_1),$$

and, hence, since $x_1 \in \Gamma(x)$ is arbitrary, we get

$$V(x) \geq \sup \left\{ F(x, x_1) + \beta V(x_1) \;\middle|\; x_1 \in \Gamma(x) \right\}.$$

Conversely, for any admissible sequence $(x'_t)_{t\geq 0}$ for the problem defining $V(x)$, we have

$$
\begin{aligned}
\sum_{t=0}^{\infty} \beta^t F(x'_t, x'_{t+1}) &= F(x, x'_1) + \beta \sum_{t=0}^{\infty} \beta^t F(x'_{t+1}, x'_{t+2}), \\
&\leq F(x, x'_1) + \beta V(x'_1) \\
&\leq \sup \left\{ F(x, x_1) + \beta V(x_{t+1}) \;\middle|\; x_1 \in \Gamma(x) \right\}.
\end{aligned}
$$

We conclude that

$$
V(x) \leq \sup \left\{ F(x, x_1) + \beta V(x_1) \;\middle|\; x_1 \in \Gamma(x) \right\}.
$$

Relation (40) follows. □

Remark 10. (i) Differently from the finite horizon case, in which the dynamic programming relations characterize the value function, in general equation (40) can have more than one solution. However, under our boundedness assumption on F, which in practice are rather restrictive, it is possible to show that the functional equation (40) admits a unique solution. As a consequence (40) characterizes the value function V.

Moreover, this solution can be computed as the limit of the following sequence of functions

$$V^{\ell+1}(x) = \sup\left\{F(x, x_{t+1}) + \beta V^{\ell}(x_{t+1}) \;\middle|\; x_{t+1} \in \Gamma(x)\right\}, \quad \forall\, x \in X,$$

with $V^0 : X \to \mathbb{R}$ being arbitrary.

(ii) If in addition we assume that X is a compact subset of $\mathbb{R}^n$, $\Gamma(x)$ closed for all $x \in X$, and that F and the correspondence Γ are continuous, then the value function V can also be shown to be continuous. In particular, for every $x \in X$, there exists $s(x) \in \Gamma(x)$ such that

$$V(x) = F(x, s(x)) + \beta V(s(x)).$$

As a consequence, the sequence defined recursively by $\bar{x}_0 = x$, $\bar{x}_{t+1} = s(\bar{x}_t)$ for all $t \geq 0$ solves the problem defining $V(x)$.