3 The Study of Smooth Optimization Problems

This chapter plays a central role in this monograph. In its first section, we present smooth optimization problems and deduce existence conditions for minimality. We take this opportunity to prove and discuss the Ekeland Variational Principle and its consequences. We then obtain necessary conditions for optimality as well as sufficient optimality conditions of the first and second-order for smooth objective functions under geometric restrictions (i.e., restrictions of the type $x \in M$, where M is an arbitrary set). The second section is dedicated to the investigation of optimality conditions under functional restrictions (with equalities and inequalities). The main aim is to deduce Karush-Kuhn-Tucker conditions and to introduce and compare several qualification conditions. Special attention is paid to the case of convex and affine data. Subsequently, we derive second-order optimality conditions for the case of functional restrictions. The last section of this chapter includes two examples which show that, for practical problems, the computational challenges posed by the optimality conditions are sometimes not easy to solve.

3.1 General Optimality Conditions

Let $U \subset \mathbb{R}^p$ be a nonempty open set, $f : U \to \mathbb{R}$ be a function and $M \subset U$ a nonempty set. We are interested in studying the minimization problem for the function f when its argument belongs to M. Formally, we write this problem in the following form:

$$(P)\ \min f(x),\ \text{subject to } x \in M.$$

The function f is called the objective function, or cost function, and the set M is called the set of feasible points of the problem (P), or the set of constraints, or the set of restrictions.

We should say from the very beginning that we shall study the minimization of f, but, by virtue of the relation $\max f = -\min(-f)$, similar results for its maximization could be obtained. Let us start by defining the notion of a solution associated to the problem (P).

Definition 3.1.1. *One says that $\overline{x} \in M$ is a local solution (or, simply, solution) of the problem (P) or minimum point of the function f on the set M if there exists a neighborhood V of $\overline{x}$ such that $f(\overline{x}) \leq f(x)$ for every $x \in M \cap V$. If $V = \mathbb{R}^p$, one says that $\overline{x}$ is a global solution of (P) or minimal global point of f on M.*

Of course, for the maximization problem, the corresponding solution is clear. Let us mention that in this chapter we will only deal with the case of smooth functions (up to the order two, i.e., $f \in C^2$).

Remark 3.1.2. *We shall distinguish two main situations for the study of problem* (P) *: (i) the case where* $M = U$ *and (ii) the case where* M *appears as an intersection between a closed set of* $\mathbb{R}^p$ *and* U*. In the former case, we say that the optimization problem* (P) *has no constraints (or restrictions), while in the latter, we call this a problem with constraints. In the case of a problem without constraints, we use as well the term "local minimum point of* f*" instead of local solution. Let us observe that in Definition 3.1.1, if* $\overline{x} \in \operatorname{int} M$*, then* $\overline{x}$ *is a local solution of the unconstrained problem (it is enough to take a smaller neighborhood* V *such that* $V \subset M$*). So, in the case of problems with restrictions, the interesting situation is when* $\overline{x} \in \operatorname{bd} M$*. If* $\overline{x} \in \operatorname{int} M$ *we say that the restriction is inactive.*

In the next sections, the two cases mentioned above will be treated together, but afterwards the discussion will split.

The basis of Optimization Theory consists of two fundamental results: the Weierstrass Theorem which ensures the existence of extrema and the Fermat Theorem (on stationary, or critical points) which gives a necessary condition for a point to be an extremum (without constraints) of a function. The theory follows the main trajectory of these fundamental results: on one hand, the study of existence conditions, and, on the other hand, the study of the (necessary and sufficient) optimality conditions.

We now recall these basic results. The classical Weierstrass Theorem, also given in the first chapter, states that a continuous function on a compact interval has a global minimum point on that interval. We have shown already that some conditions can be relaxed. Here is the theorem again.

Theorem 3.1.3 (Weierstrass Theorem)**.** *If* $f : K \subset \mathbb{R}^p \to \mathbb{R}$ *is a lower (upper) semicontinuous function and* K *is a compact set, then* f *is lower (upper) bounded on* K *(i.e.,* $f(K)$ *is a bounded below (above) set) and* f *attains its minimum (maximum) on* K*, i.e., there exists* $\overline{x} \in K$ *such that* $\inf\limits_{x \in K} f(x) = f(\overline{x})$ *(*$\sup\limits_{x \in K} f(x) = f(\overline{x})$*, respectively).*

Remark 3.1.4. *The compactness of* K *is essential in the Weierstrass Theorem because otherwise the conclusion does not hold. As an example, let us consider the continuous function* $f : (0, 1] \to \mathbb{R}$*,* $f(x) = x^{-1} \sin(x^{-1})$*. Clearly,* $\inf\limits_{(0,1]} f(x) = -\infty$*, and* $\sup\limits_{(0,1]} f(x) = +\infty$*.*

We now present Fermat Theorem. The particular case of a function of one real variable was presented in Section 1.3. The proof of the theorem will be given later in this chapter.

Theorem 3.1.5 (Fermat Theorem). *Let $S \subset \mathbb{R}^p$ be a set and $a \in \operatorname{int} S$. If $f : S \to \mathbb{R}$ is of class C^1 in a neighborhood of a, and a is a local minimum or maximum point of f, then a is also a stationary (or critical) point of f, i.e., $\nabla f(a) = 0$.*

Some remarks are in order to understand the applications of Fermat Theorem.

Remark 3.1.6. *1. The converse of Fermat Theorem is not true: for instance, the derivative of $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ vanishes at 0, but this point is neither minimum nor maximum of f.*

2. The interiority condition for a is essential since, without this assumption, the conclusion does not hold: take $f : [0, 1] \to [0, 1]$, $f(x) = x$ which has at $\overline{x} = 0$ a minimum point where the derivative is not 0. Therefore, in view of Remark 3.1.2, one can say that Fermat Theorem applies only to unconstrained problems.

3. If S is compact and f is continuous on S and differentiable on $\operatorname{int} S$, it is possible to have $\nabla f(x) \neq 0$ for every $x \in \operatorname{int} S$, and in such a case the extreme points of f on S, which surely exist from Weierstrass Theorem, lie on the boundary of S.

We now start our discussion on the existence conditions for the minimum points.

Theorem 3.1.7. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a lower semicontinuous function and $M \subset \mathbb{R}^p$ be a nonempty, closed set. If there exists $\nu > \inf_{x \in M} f(x)$ such that the level set of f relative to M, i.e., $M \cap N_\nu f = \{x \in M \mid f(x) \leq \nu\}$, is bounded, then f attains its global minimum on M.*

Proof It is obvious that if f has a global minimum on M. Similarly, it also has a global minimum on $M \cap N_\nu f$. Since this set is compact and f is lower semicontinuous, from Weierstrass Theorem 3.1.3, we infer that f is lower bounded and attains its global minimum on $M \cap N_\nu f$, whence on M, too. □

Obviously, in the above result, if M is bounded, then the hypothesis is automatically fulfilled. The interesting case is where M is unbounded and in this situation we ensure the boundedness assumption by imposing a certain condition on f.

Proposition 3.1.8. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function and $M \subset \mathbb{R}^p$ be a closed, unbounded set. If $\lim_{x \in M, \|x\| \to \infty} f(x) = \infty$ (i.e., for every sequence $(x_k) \subset M$ with $\|x_k\| \to \infty$, one has $f(x_k) \to \infty$), then the set $N_\nu f \cap M$ is bounded for every $\nu > \inf_{x \in M} f(x)$.*

Proof Let $\nu > \inf_{x \in M} f$. If the set $N_\nu f \cap M$ would be unbounded, then there exists $(x_k) \subset N_\nu f \cap M$ with $\|x_k\| \to \infty$. On one hand, by our assumption, $\lim f(x_k) = \infty$ and, on the other hand, $f(x_k) \leq \nu$ for every $k \in \mathbb{N}$, which is absurd. So $N_\nu f \cap M$ is bounded. □

If the condition $\lim_{x\in M, \|x\|\to\infty} f(x) = \infty$ holds, we say that f is coercive relative to M. If $\lim_{\|x\|\to\infty} f(x) = \infty$, the we say that f is coercive.

Some special conditions to ensure the existence and the attainment of the minimum on a bounded, not closed set can be given as well. We present here such a result which we are going to use in Chapter 6.

Proposition 3.1.9. *Let D, $C \subset \mathbb{R}^p$ be two sets such that C is compact and $D \cap C \neq \emptyset$. Let $\varphi : D \cap C \to \mathbb{R}$ be a continuous function. Suppose that the following condition holds: for every sequence $(x_k) \subset D \cap C$, $x_k \to \overline{x} \in \operatorname{bd} D \cap C$, the sequence $(\varphi(x_n))$ is unbounded above. Then φ is lower bounded and attains its minimum on $D \cap C$.*

Proof Firstly, let us observe that φ cannot be constant. Let $x_0 \in D \cap C$ such that $\varphi(x_0) > \inf_{x\in D\cap C} \varphi(x)$. It is enough to show that the level set $A := \{x \in D \cap C \mid \varphi(x) \leq \varphi(x_0)\}$ is compact (see the proof of Theorem 3.1.7). Obviously, A is bounded (as a subset of C). It remains to show that A is closed. Let $(x_k) \subset A$, $x_k \to \overline{x}$. Suppose, by contradiction, that $\overline{x} \notin A$. Then, from the closedness of C,

$$\overline{x} \in (\operatorname{cl} D \setminus D) \cap C \subset \operatorname{bd} D \cap C.$$

By assumption, $(\varphi(x_k))$ is unbounded above, which contradicts the definition of the set A. Hence, A is compact and the conclusion follows. □

In general, a global minimum is local minimum, but, of course, the converse is false. We shall derive a condition for the fulfillment of this converse. We now give a necessary and sufficient condition for a point to be a global minimum.

Theorem 3.1.10. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a continuous function and $\overline{x} \in \mathbb{R}^p$. Then the following assertions are equivalent:*

(i) $\overline{x}$ is a global minimum point of f;

(ii) every $x \in \mathbb{R}^p$ with $f(x) = f(\overline{x})$ is a local minimum point of f.

Proof The implication $(i) \Rightarrow (ii)$ is obvious. Suppose that (ii) holds, but $\overline{x}$ is not a global minimum point. Then, there exists $u \in \mathbb{R}^p$ with $f(u) < f(\overline{x})$. We define $\varphi : [0, 1] \to \mathbb{R}$ by $\varphi(t) := f(t\overline{x} + (1 - t)u)$. The set $S := \{t \in [0, 1] \mid \varphi(t) = f(\overline{x})\}$ is nonempty ($1 \in S$), closed (from the continuity of f) and bounded. Then there is $t_0 = \min S$. Clearly, $t_0 \in (0, 1]$. From the hypothesis, $f(t_0\overline{x}+(1-t_0)u) = f(\overline{x})$ tells us that the point $t_0\overline{x}+(1-t_0)u$ is a local minimum of f. Consequently, there is $\varepsilon > 0$ such that for every $t \in [0, 1]\cap(t_0 - \varepsilon, t_0 + \varepsilon)$, $\varphi(t) \geq \varphi(t_0)$. Since $t_0 = \min S$, if one takes $t_1 \in [0, 1] \cap (t_0 - \varepsilon, t_0)$, the strict inequality $\varphi(t_1) > \varphi(t_0) > \varphi(0)$ holds. The function f being continuous, it has the Darboux property (or intermediate value property), whence there exists $t_2 \in (0, t_1)$ with $\varphi(t_2) = \varphi(t_0) = f(\overline{x})$, and this contradicts the minimality of t_0. Therefore, the assumption made was false, hence the conclusion holds. □

Notice that, if f is not continuous, the result does not hold. For this, it is sufficient to analyze the following function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = \begin{cases} -x - 1, x \in (-\infty, -1] \\ x + 1, x \in (-1, 0) \\ -1, x = 0 \\ -x + 1, x \in (0, 1) \\ x - 1, x \in [1, \infty). \end{cases}$$

Let us observe that, by using the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ and $\overline{x} = 0$, that one cannot replace in the item (ii) above the local minimality by the stationarity.

The above results ensure sufficient conditions for the existence of minimum points under compactness assumptions for the level sets. Conversely, it is clear that the lower boundedness of the function is a necessary condition for the existence of minimum points, but the boundedness of the level sets is not. For instance, the function $f : (0, \infty) \to \mathbb{R}$, $f(x) = (x - 1)^2 e^{-x}$ attains its minimum at $\overline{x} = 1$ and the minimal value is 0, but $N_v f$ is not bounded for every value of $v > 0 = \inf\{f(x) \mid x \in (0, \infty)\}$.

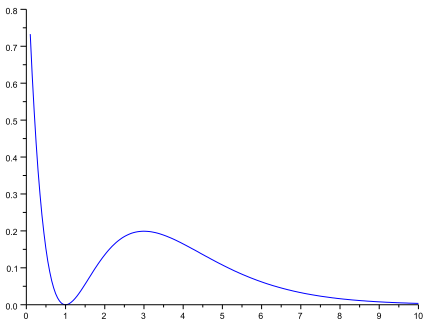

Figure 3.1: The graph of $(x - 1)^2 e^{-x}$.

One may like to have a notion of approximate minimum which is advantageous to always exist if f is lower bounded. We work here with unconstrained problems in order to better illustrate the main ideas.

Definition 3.1.11. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a lower bounded function and take $\varepsilon > 0$. A point x_ε is called ε–minimum of f if*

$$f(x_\varepsilon) \leq \inf_{x \in \mathbb{R}^p} f(x) + \varepsilon.$$

Since $\inf_{x\in\mathbb{R}^p} f(x) \in \mathbb{R}$, the existence of $\varepsilon-$minima for every positive ε is ensured. We use the generic term of approximate minima for $\varepsilon-$minima.

We now present a very important result, the Ekeland Variational Principle, which states that close to an approximate minimum point one can find a genuine minimum point for some perturbation of the initial function. This results was proved by the French mathematician Ivar Ekeland in 1974.

Theorem 3.1.12 (Ekeland Variational Principle)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a lower semicontinuous and lower bounded function. Let $\varepsilon > 0$ and let x_ε be an $\varepsilon-$minimum of f. Then, for every $\delta > 0$ there exists $\overline{x}_\varepsilon \in \mathbb{R}^p$ that have the following properties:*

$$\begin{aligned}
&f(\overline{x}_\varepsilon) \le f(x_\varepsilon),\\
&\|\overline{x}_\varepsilon - x_\varepsilon\| \le \delta,\\
&f(\overline{x}_\varepsilon) \le f(x) + \varepsilon\delta^{-1}\|x - \overline{x}_\varepsilon\|\,,\ \forall x \in \mathbb{R}^p.
\end{aligned}$$

Proof Let us consider the function $g : \mathbb{R}^p \to \mathbb{R}$

$$g(x) = f(x) + \varepsilon\delta^{-1}\|x - x_\varepsilon\|\,.$$

Using the assumptions on f, we infer that g is lower semicontinuous and lower bounded. Moreover, g is coercive, i.e.,

$$\lim_{\|x\|\to\infty} g(x) = +\infty.$$

From Theorem 3.1.7 and Proposition 3.1.8, the function g has a global minimum point, which we denote by $\overline{x}_\varepsilon$. Consequently,

$$f(\overline{x}_\varepsilon) + \varepsilon\delta^{-1}\|\overline{x}_\varepsilon - x_\varepsilon\| \le f(x) + \varepsilon\delta^{-1}\|x - x_\varepsilon\|\,,\ \forall x \in \mathbb{R}^p. \tag{3.1.1}$$

For $x = x_\varepsilon$, we get

$$f(\overline{x}_\varepsilon) \le f(x_\varepsilon).$$

That is the first relation in the conclusion. On the other hand, using this inequality

$$f(x_\varepsilon) \le \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon,$$

with the relation (3.1.1), for $x = x_\varepsilon$, implies that:

$$\delta^{-1}\|\overline{x}_\varepsilon - x_\varepsilon\| \le 1.$$

This is the second part of the conclusion. Relation (3.1.1) allows us to write, successively, for any $x \in \mathbb{R}^p$,

$$\begin{aligned}
f(\overline{x}_\varepsilon) &\le f(x) + \varepsilon\delta^{-1}(\|x - x_\varepsilon\| - \|\overline{x}_\varepsilon - x_\varepsilon\|)\\
&\le f(x) + \varepsilon\delta^{-1}\|x - \overline{x}_\varepsilon\|\,.
\end{aligned}$$

That is the last part of the conclusion. The proof is complete. □

Remark 3.1.13. *Notice that the Ekeland Variational Principle holds (with minor changes in the proof) if instead of the whole space $\mathbb{R}^p$ one takes a closed subset of it.*

Clearly, the point $\overline{x}_\varepsilon$ is a global minimum point for the function

$$x \mapsto f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\| .$$

On the other hand, if we want $\overline{x}_\varepsilon$ to be close to x_ε (i.e., δ to be small), then the perturbation term, $\varepsilon\delta^{-1} \|\cdot - \overline{x}_\varepsilon\|$ is big. A compromise would be to choose $\delta := \sqrt{\varepsilon}$, and in this case one gets the next consequence.

Corollary 3.1.14. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a lower semicontinuous and lower bounded function. Take $\varepsilon > 0$ and let x_ε be an ε–minimum of f. Then there exists $\overline{x}_\varepsilon \in \mathbb{R}^p$ having the properties:*

$$\begin{aligned} &f(\overline{x}_\varepsilon) \leq f(x_\varepsilon), \\ &\|\overline{x}_\varepsilon - x_\varepsilon\| \leq \sqrt{\varepsilon}, \\ &f(\overline{x}_\varepsilon) \leq f(x) + \sqrt{\varepsilon} \|x - \overline{x}_\varepsilon\| , \ \forall x \in \mathbb{R}^p. \end{aligned}$$

The Ekeland Variational Principle has many applications. Some of these refer to the same issues of extreme points. The next example of such an application asserts that every differentiable function has approximate critical points (for which the norm of the differential is arbitrarily small).

Theorem 3.1.15. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable, lower bounded function. Then for every $\varepsilon, \delta > 0$, there exists $\overline{x}_\varepsilon \in \mathbb{R}^p$ with $f(\overline{x}_\varepsilon) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon$ and $\left\|\nabla f(\overline{x}_\varepsilon)\right\| \leq \varepsilon\delta^{-1}$. In particular, there exists $(x_n) \subset \mathbb{R}^p$ with*

$$f(x_n) \to \inf_{x\in\mathbb{R}^p} f(x), \ \nabla f(x_n) \to 0.$$

Proof Let x_ε be an ε–minimum of f. According to Theorem 3.1.12, there exists $\overline{x}_\varepsilon \in \mathbb{R}^p$ with the three mentioned properties. Since $f(\overline{x}_\varepsilon) \leq f(x_\varepsilon)$, we infer that $f(\overline{x}_\varepsilon) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon$. Let $x := \overline{x}_\varepsilon + tu$ with $u \in \mathbb{R}^p$ and $t > 0$. The relation $f(\overline{x}_\varepsilon) \leq f(x) + \varepsilon\delta^{-1} \|x - \overline{x}_\varepsilon\|$ holds for any $x \in \mathbb{R}^p$, so we have

$$\frac{f(\overline{x}_\varepsilon + tu) - f(\overline{x}_\varepsilon)}{t} \geq -\varepsilon\delta^{-1} \|u\| .$$

Passing to the limit with $t \to 0$, we deduce

$$\nabla f(\overline{x}_\varepsilon)(u) \geq -\varepsilon\delta^{-1} \|u\| , \ \forall u \in \mathbb{R}^p,$$

that is,

$$-\nabla f(\overline{x}_\varepsilon)(u) \leq \varepsilon\delta^{-1} \|u\| , \ \forall u \in \mathbb{R}^p.$$

Changing u into $-u$, we get

$$\nabla f(\overline{x}_\varepsilon)(u) \leq \varepsilon\delta^{-1} \|u\| , \ \forall u \in \mathbb{R}^p,$$

whence

$$\left|\nabla f(\overline{x}_\varepsilon)(u)\right| \leq \varepsilon\delta^{-1}\|u\|, \ \forall u \in \mathbb{R}^p,$$

and this implies $\left\|\nabla f(\overline{x}_\varepsilon)\right\| \leq \varepsilon\delta^{-1}$. For $\varepsilon := n^{-1}$, $\delta = \sqrt{n^{-1}}$, $n \in \mathbb{N}^*$ we obtain the second part of the conclusion. □

The Ekeland Variational Principle also allows us to prove the equivalence of several existence conditions for minimum points.

Theorem 3.1.16. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a differentiable, lower bounded function. The following assertions are equivalent:*

(i) $\lim_{\|x\|\to\infty} f(x) = \infty$;

(ii) $N_v f$ *is bounded for any* $v > \inf_{x\in\mathbb{R}^p} f(x)$;

(iii) every sequence (x_n) *for which* $(f(x_n))$ *is convergent and* $\nabla f(x_n) \to 0$ *has a convergent subsequence.*

Proof The implication $(i) \Rightarrow (ii)$ was already proved in Proposition 3.1.8, while $(ii) \Rightarrow (iii)$ is a consequence of the fact that every bounded sequence has a convergent subsequence. Let us show that $(iii) \Rightarrow (i)$. Suppose, by way of contradiction, that there exist $c \in \mathbb{R}$ and a sequence $(x_n) \subset \mathbb{R}^n$ with $\|x_n\| \to \infty$ and $f(x_n) \leq c$, for every $n \in \mathbb{N}$. Clearly, $c \geq \inf_{x\in\mathbb{R}^p} f(x)$. For every $n \in \mathbb{N}^*$ we choose

$$\varepsilon_n := c + n^{-1} - \inf_{x\in\mathbb{R}^p} f(x) > 0,$$

so,

$$f(x_n) < \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon_n.$$

Let $\delta_n := 2^{-1}\|x_n\| > 0$. Like in Theorem 3.1.15 (and its proof) there exists $\overline{x}_n$ with

$$\begin{aligned} &f(\overline{x}_n) \leq f(x_n) \leq \inf_{x\in\mathbb{R}^p} f(x) + \varepsilon_n, \\ &\|\overline{x}_n - x_n\| \leq \delta_n, \\ &\left\|\nabla f(\overline{x}_n)\right\| \leq \varepsilon_n\delta_n^{-1}. \end{aligned}$$

But,

$$\|\overline{x}_n\| \geq \|x_n\| - \|\overline{x}_n - x_n\| \geq \|x_n\| - 2^{-1}\|x_n\| = 2^{-1}\|x_n\|,$$

whence $\|\overline{x}_n\| \to \infty$. On the other hand,

$$\left\|\nabla f(\overline{x}_n)\right\| \leq \frac{2}{\|x_n\|}(c + n^{-1} - \inf_{x\in\mathbb{R}^p} f(x)) \to 0.$$

Since $(f(\overline{x}_n))$ is bounded, it has a convergent subsequence. From (iii), one deduces that $(\overline{x}_n)$ should also have such a subsequence, but this is not possible. Consequently, (i) holds. □

The condition (iii) in the above result is called Palais-Smale condition.

On the basis of Ekeland Variational Principle one obtains a new condition for the existence of the minimum points.

Theorem 3.1.17. *Let $\alpha > 0$ and $f : \mathbb{R}^p \to \mathbb{R}$ be a lower semicontinuous and lower bounded function. Suppose that for every $x \in \mathbb{R}^p$ with $\inf_{x\in\mathbb{R}^p} f(x) < f(x)$ there exists $z \in \mathbb{R}^p \setminus \{x\}$ such that*

$$f(z) < f(x) - \alpha \|z - x\| .$$

Then f has a minimum global point.

Proof Suppose, to obtain a contradiction, that the conclusion does not hold. Then, for every $x \in \mathbb{R}^p$, $\inf_{x\in\mathbb{R}^p} f(x) < f(x)$, so, by the assumptions made, there exists $z_x \in \mathbb{R}^p \setminus \{x\}$ such that

$$f(z_x) < f(x) - \alpha \|z_x - x\| .$$

By the Ekeland Variational Principle for $\varepsilon > 0$, $\delta > 0$ with $\varepsilon\delta^{-1} = \alpha$, there is an element $u \in \mathbb{R}^p$ with

$$f(u) \le f(v) + \alpha \|v - u\| , \ \forall v \in \mathbb{R}^p.$$

Then

$$f(u) \le f(z_u) + \alpha \|z_u - u\| < f(u),$$

which is absurd. Hence the conclusion hold. □

A straightforward example of a function which satisfies the conditions of the above result is $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \beta |x|$ for $\beta > \alpha$.

In the second part of this section, we present necessary optimality conditions and sufficient optimality conditions. At first, we deduce necessary optimality conditions that use the ideas developed around the construction and the study of the Bouligand tangent cone.

Theorem 3.1.18 (First-order necessary optimality condition)**.** *If $\overline{x}$ is a local solution of (P) and f is differentiable at $\overline{x}$, then $\nabla f(\overline{x})(u) \ge 0$ for every $u \in T_B(M, \overline{x})$.*

Proof Let V be a neighborhood of $\overline{x}$ where $f(\overline{x}) \le f(x)$ for every $x \in V \cap M$. Let $u \in T_B(M, \overline{x})$. Then there exists $(t_n) \subset (0, \infty)$ with $t_n \to 0$ and $(u_n) \to u$ such that for every n,

$$\overline{x} + t_n u_n \in M.$$

Obviously, the sequence $(t_n u_n)$ converges towards $0 \in \mathbb{R}^p$ and, for n large enough, $\overline{x} + t_n u_n$ belongs to V. Taking into account the differentiability of f at $\overline{x}$, there exists $(\alpha_n) \subset \mathbb{R}$, $\alpha_n \to 0$ such that for every $n \in \mathbb{N}$,

$$f(\overline{x} + t_n u_n) = f(\overline{x}) + t_n \nabla f(\overline{x})(u_n) + t_n \|u_n\| \alpha_n,$$

whence,

$$\nabla f(\overline{x})(u_n) + \|u_n\| \alpha_n \ge 0,$$

for all n large enough. Passing to the limit for $n \to \infty$, we get the conclusion. □

Remark 3.1.19. *The conclusion of Theorem 3.1.18 could equivalently be written as*

$$-\nabla f(\overline{x}) \in N_B(M, \overline{x}).$$

Remark 3.1.20. *Taking into account Proposition 2.1.12, if $\overline{x} \in \operatorname{int} M$ (inactive restriction), Theorem 3.1.18 gives $\nabla f(\overline{x})(u) \geq 0$ for every $u \in \mathbb{R}^p$. The linearity of $\nabla f(\overline{x})$ implies $\nabla f(\overline{x}) = 0$, i.e., the Fermat Theorem on stationary points.*

We present now a second-order necessary optimality condition for the problem without restrictions.

Theorem 3.1.21 (Second-order necessary optimality condition)**.** *Let $U \subset \mathbb{R}^p$ be an open set and $\overline{x} \in U$. If $f : U \to \mathbb{R}$ is of class C^2 on a neighborhood of $\overline{x}$, and $\overline{x}$ is a local minimum point of f, then $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is positive semidefinite (that is, $\nabla^2 f(\overline{x})(u, u) \geq 0$ for every $u \in \mathbb{R}^p$).*

Proof Let $V \subset U$ be a neighborhood of $\overline{x}$ such that $f(\overline{x}) \leq f(x)$ for every $x \in V$ and f is of class C^2 on V. The fact that $\nabla f(\overline{x}) = 0$ follows from Fermat Theorem. As before, take $u \in \mathbb{R}^p$ and $(t_n) \subset (0, \infty)$ with $t_n \to 0$. Taylor Theorem 1.3.4 says that for every $n \in \mathbb{N}$ there exists $c_n \in (\overline{x}, \overline{x} + t_n u)$ such that

$$f(\overline{x} + t_n u) - f(\overline{x}) = t_n \nabla f(\overline{x})(u) + \frac{1}{2} t_n^2 \nabla^2 f(c_n)(u, u) = \frac{1}{2} t_n^2 \nabla^2 f(c_n)(u, u).$$

For n sufficiently large, $f(\overline{x} + t_n u) - f(\overline{x}) \geq 0$, whence

$$\nabla^2 f(c_n)(u, u) \geq 0,$$

and passing to the limit as $n \to \infty$ we get $c_n \to \overline{x}$. Since f is of class C^2, we infer

$$\nabla^2 f(\overline{x})(u, u) \geq 0,$$

whence $\nabla^2 f(\overline{x})$ is positive semidefinite. □

Obviously, if $\overline{x} \in \operatorname{int} U$ is a local maximum of f, then $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(\overline{x})$ is negative semidefinite (i.e., $\nabla^2 f(\overline{x})(u, u) \leq 0$ for every $u \in \mathbb{R}^p$). Further, if $\nabla^2 f(\overline{x})$ is neither positive semidefinite, or negative semidefinite, then $\overline{x}$ is not an extreme point of f.

In fact, many results in this book are generalizations, refinements of these results, or answer to different issues which naturally arise from their analysis.

One may consider whether the converses of these results are true. The answer is negative in both cases: it is sufficient to consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ and $\overline{x} = 0$.

One can however impose supplementary conditions in order to get some equivalences, and the most important of these relates to convex functions.

Theorem 3.1.22. *Let $U \subset \mathbb{R}^p$ be an open convex set and let $f : U \to \mathbb{R}$ be a convex differentiable function. The next assertions are equivalent:*

(i) $\overline{x}$ is a global minimum point of f (on U);

(ii) $\overline{x}$ is a local minimum point of f;

(iii) $\overline{x}$ is a critical point of f (i.e., $\nabla f(\overline{x}) = 0$).

Proof The implication $(i) \Rightarrow (ii)$ is obvious for every function, and $(ii) \Rightarrow (iii)$ follows from Fermat Theorem. Finally, the implication $(iii) \Rightarrow (i)$ relies on the convexity of f and follows from Theorem 2.2.10. □

Therefore, for convex functions, the first-order necessary optimality condition (in the unconstrained case) is also sufficient. In this situation, the second order condition is automatically satisfied (according to Theorem 2.2.10).

Concerning the nature of the extreme points for convex functions, we record here some important aspects.

Proposition 3.1.23. *Let $M \subset \mathbb{R}^p$ be a convex set and let $f : M \to \mathbb{R}$ be a convex function. If $\overline{x} \in M$ is a local minimum point of f on M, then $\overline{x}$ is in fact a global minimum point of f on M. If $u \in \operatorname{int} M$ is a local maximum point of f, then u is a global minimum point of f.*

Proof Let $\overline{x}$ be a local minimum point of f on M. Then there exists a convex neighborhood V of $\overline{x}$ such that for every $x \in V \cap M$, $f(\overline{x}) \leq f(x)$. Let $x \in M$. There exists $\lambda \in (0, 1)$ such that $y := (1 - \lambda)\overline{x} + \lambda x \in M \cap V$. Then,

$$f(\overline{x}) \leq f(y) = f((1 - \lambda)\overline{x} + \lambda x) \leq (1 - \lambda)f(\overline{x}) + \lambda f(x),$$

that is,

$$\lambda f(\overline{x}) \leq \lambda f(x),$$

and the conclusion of the first part follows.

For the second part, there is a convex symmetric neighborhood V of 0 (a ball with the center 0, for instance) such that for every $v \in V$, $f(u + v) \leq f(u)$ and $f(u - v) \leq f(u)$. Then

$$f(u) = f\left(\frac{1}{2}(u + v) + \frac{1}{2}(u - v)\right) \leq \frac{1}{2}f(u + v) + \frac{1}{2}f(u - v) \leq f(u),$$

for all $v \in V$. Consequently, $f(u + v) = f(u)$ for every $v \in V$. Therefore, u is a local (hence global) minimum point of f. □

Proposition 3.1.24. *Let $M \subset \mathbb{R}^p$ be a convex set and let $f : M \to \mathbb{R}$ be a convex function. If nonempty, the set of minimum points of f on M is convex. If, moreover, f is strictly convex, then this set has at most one element.*

Proof From the preceding result, if $x_1, x_2 \in M$ are (global) minima of f on M, then $f(x_1) = f(x_2)$. The convexity implies $f(x) = f(x_1)$ for every $x \in [x_1, x_2]$. Therefore, the first part is proved. Suppose now that f is strictly convex. If this is so, then we would have two different global minima, then $f(x) < f(x_1)$ for every $x \in (x_1, x_2)$, which is not possible. □

For constrained problems involving convex functions, the first-order necessary optimality condition is, again, a sufficient optimality condition.

Proposition 3.1.25. *Let $U \subset \mathbb{R}^p$ be a convex open set and let $f : U \to \mathbb{R}$ be a convex, differentiable function. Let $M \subset U$ be convex. The element $\overline{x} \in M$ is a minimum point of f on M if and only if*

$$-\nabla f(\overline{x}) \in N(M, \overline{x}).$$

Proof Let $\overline{x} \in M$ be a minimum point of f on M. Then, according to Theorem 3.1.18, $\nabla f(\overline{x})(u) \geq 0$ for every $u \in T(M, \overline{x})$, that is

$$-\nabla f(\overline{x}) \in T(M, \overline{x})^- = N(M, \overline{x}).$$

Conversely, we know from the convexity of f (Theorem 2.2.10), that

$$f(x) \geq f(\overline{x}) + \nabla f(\overline{x})(x - \overline{x}), \ \forall x \in U.$$

But, using the hypothesis and the convexity of M (Proposition 2.1.15),

$$-\nabla f(\overline{x}) \in N(M, \overline{x}) = \{u \in \mathbb{R}^p \mid \langle u, x - \overline{x} \rangle \leq 0, \ \forall x \in M\},$$

whence $\nabla f(\overline{x})(x - \overline{x}) \geq 0$ for every $x \in M$. From these relations we know $f(x) \geq f(\overline{x})$ for every $x \in M$. □

Coming back to Theorems 3.1.18 and 3.1.21, in order to formulate sufficient optimality conditions, we strengthen the conclusion of these results. The good point is that we get stronger minimality concepts.

Definition 3.1.26. *Let $\alpha > 0$. One says that $\overline{x} \in M$ is a strict local solution of order α for (P), or a strict local minimum point of order α for f on M if there exist two constants $r, l > 0$ such that for every $x \in M \cap B(\overline{x}, r)$,*

$$f(x) \geq f(\overline{x}) + l \|x - \overline{x}\|^{\alpha}.$$

The announced results are as follows.

Theorem 3.1.27. *Suppose that f is differentiable at $\overline{x} \in M$ and*

$$\nabla f(\overline{x})(u) > 0, \ \forall u \in T_B(M, \overline{x}) \setminus \{0\}.$$

Then $\overline{x}$ is a strict local solution of order $\alpha = 1$ for (P).

Proof Suppose, by way of contradiction, that $\overline{x}$ is not a strictly local solution of order 1. Then, there exists a sequence $(x_n) \to \overline{x}$, $(x_n) \subset M$ such that for every $n \in \mathbb{N}^*$,

$$f(x_n) < f(\overline{x}) + n^{-1} \|x_n - \overline{x}\| .$$

By virtue of this inequality,

$$x_n \neq \overline{x}, \ \forall n \in \mathbb{N}^*.$$

Since f is differentiable, there exists a sequence of real numbers $(y_n) \to 0$ such that for every $n \in \mathbb{N}$,

$$f(x_n) = f(\overline{x}) + \nabla f(\overline{x})(x_n - \overline{x}) + y_n \|x_n - \overline{x}\| .$$

The combination of these two relation yields

$$n^{-1} \|x_n - \overline{x}\| > \nabla f(\overline{x})(x_n - \overline{x}) + y_n \|x_n - \overline{x}\| ,$$

whence, by division with $\|x_n - \overline{x}\|$, we deduce

$$n^{-1} > \nabla f(\overline{x}) \left(\frac{x_n - \overline{x}}{\|x_n - \overline{x}\|} \right) + y_n, \ \forall n \in \mathbb{N}^*. \tag{3.1.2}$$

Since the sequence $\left(\frac{x_n - \overline{x}}{\|x_n - \overline{x}\|} \right)$ is bounded, there exists a convergent subsequence of it. The limit, denoted by u, of this subsequence is not zero (being of norm 1) and, furthermore, from $\|x_n - \overline{x}\| \to 0$, we infer that $u \in T_B(M, \overline{x})$. Consequently, $u \in T_B(M, \overline{x}) \setminus \{0\}$, and passing to the limit in the relation (3.1.2) we have

$$0 \geq \nabla f(\overline{x})(u),$$

which is in contradiction with the hypothesis. □

Notice that for differentiable functions, the concept of a local strict solution of order 1 is specific to the case of active restrictions (that is, $\overline{x} \in M \setminus \operatorname{int} M$): if f is differentiable at $\overline{x} \in \operatorname{int} M$, then $\overline{x}$ cannot be a local strict solution of order 1. Indeed, if $\overline{x} \in \operatorname{int} M$ would be local strict solution of order 1, then, on one hand, $\nabla f(\overline{x}) = 0$ (Fermat Theorem), and, on the other hand, $\nabla f(\overline{x}) \neq 0$ from the definition of strict solutions.

Concerning second-order optimality conditions, one has the following results.

Theorem 3.1.28. *Suppose that f is of class C^2, $\nabla f(\overline{x}) = 0$ and*

$$\nabla^2 f(\overline{x})(u, u) > 0, \ \forall u \in T_B(M, \overline{x}) \setminus \{0\}.$$

Then $\overline{x}$ is a local strict solution of order $\alpha = 2$ for problem (P).

Proof As before, one supposes, by contradiction, that the conclusion does not hold. Then there exists a sequence $(x_n) \to \overline{x}$, $(x_n) \subset M \setminus \{\overline{x}\}$ such that for every $n \in \mathbb{N}^*$,

$$f(x_n) < f(\overline{x}) + n^{-1} \|x_n - \overline{x}\|^2 .$$

From Taylor Theorem 1.3.4, for every $n \in \mathbb{N}$ there exists c_n on the segment joining $\overline{x}$ and x_n such that

$$f(x_n) - f(\overline{x}) = \nabla f(\overline{x})(x_n - \overline{x}) + \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x})$$
$$= \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x}).$$

We get

$$n^{-1}\|x_n - \overline{x}\|^2 > \frac{1}{2}\nabla^2 f(c_n)(x_n - \overline{x}, x_n - \overline{x}),$$

whence, in order to finish the proof, we divide by $\|x_n - \overline{x}\|^2$ and we repeat the above arguments. □

In the unconstrained case, this result gives the following consequence.

Corollary 3.1.29. *Let $U \subset \mathbb{R}^p$ be a nonempty, open set and $f : U \to \mathbb{R}$ be a C^2 function. If $\overline{x} \in U$ is a critical point of f and $\nabla^2 f(\overline{x})$ is positive definite (i.e., $\nabla^2 f(\overline{x})(u, u) > 0$ for every $u \in \mathbb{R}^p \setminus \{0\}$), then $\overline{x}$ is a local strict solution of order $\alpha = 2$ for f.*

One can identify $\nabla^2 f(\overline{x})$ with the Hessian matrix of f at $\overline{x}$, $\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(\overline{x})\right)_{i,j\in\overline{1,p}}$, and a sufficient condition for the positive definiteness of it is given by the next criterion, known from linear algebra (Sylvester criterion): all the determinants of the matrices $\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(\overline{x})\right)_{i,j\in\overline{1,k}}$, $k \in \overline{1,p}$ are strictly positive. Analogously (for $-f$), if the determinants of the matrices $\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(\overline{x})\right)_{i,j\in\overline{1,k}}$, $k \in \overline{1,p}$ are not zero and change the signs starting with minus, then $\nabla^2 f(\overline{x})$ is negative definite and $\overline{x}$ is a maximum point. Furthermore, if all these determinants are not zero, then any other distribution of their signs leads to the conclusion that the reference point is not an extreme point.

3.2 Functional Restrictions

The restriction of the problem (P) introduced in the previous section is $x \in M$. Many times, in practice this set M of feasible points is defined by means of functions. Let us consider $g : \mathbb{R}^p \to \mathbb{R}^n$ and $h : \mathbb{R}^p \to \mathbb{R}^m$ as C^1 functions. As usual, g and h can be thought of as $g = (g_1, g_2, \ldots, g_n)$, and $h = (h_1, h_2, \ldots, h_m)$, respectively, where $g_i : \mathbb{R}^p \to \mathbb{R}$ ($i \in \overline{1,n}$) and $h_j : \mathbb{R}^p \to \mathbb{R}$ ($j \in \overline{1,m}$) are C^1 real valued functions.

Let the set of feasible points be defined as:

$$M := \{x \in U \mid g(x) \leq 0,\ h(x) = 0\} \subset \mathbb{R}^p.$$

Let us observe that we have two types of constraints: equalities and inequalities. Let $x \in M$. If for an $i \in \overline{1,n}$, one has that $g_i(x) < 0$, then the continuity of g ensures the existence of a neighborhood V of x such that $g_i(y) < 0$ for all $y \in V$. Therefore,

when one looks for a certificate that x is a local solution of (P), the restriction $g_i \leq 0$ does not effectively influence the set of points u where one should compare $f(x)$ and $f(u)$. For this reason, one says that the restriction $g_i \leq 0$ is inactive at x and these kind of restrictions should be eliminated from the discussion. In the opposite case, when $g_i(x) = 0$, we call this active (inequality) restriction. For $\overline{x} \in M$, we denote the set of indexes corresponding to active inequality type restrictions by

$$A(\overline{x}) = \{i \in \overline{1,n} \mid g_i(\overline{x}) = 0\}.$$

We are now going to present two types of optimality conditions for problem (P) with functional constraints as described above. These two types of conditions are formally very close, but their differences are important for the detection of extreme points. We start with the Fritz John necessary optimality conditions where the objective function does not play any special role with respect to the functions which define the restrictions. We shall consider the drawbacks of these conditions, and next we shall impose supplementary conditions in order to eliminate then. By this procedure, we get the famous Karush-Kuhn-Tucker necessary optimality conditions which will be extensively used for solving nonlinear optimization problems.

3.2.1 Fritz John Optimality Conditions

The result of this subsection refers to necessary optimality conditions for problem (P) with functional restrictions without any additional assumption to the general framework already described. These conditions were obtained in 1948 by the German mathematician Fritz John.

Theorem 3.2.1 (Fritz John). *Let $\overline{x} \in M$ be a solution of (P). Then there exist $\lambda_0 \in \mathbb{R}$, $\lambda_0 \geq 0$, $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{R}^n$, $\mu = (\mu_1, \mu_2, \ldots, \mu_m) \in \mathbb{R}^m$, with $\lambda_0 + \|\lambda\| + \|\mu\| \neq 0$ such that*

$$\lambda_0 \nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0$$

and

$$\lambda_i \geq 0, \ \lambda_i g_i(\overline{x}) = 0, \ \textit{for every } i \in \overline{1,n}.$$

Proof Let us take $\delta > 0$ such that $D(\overline{x}, \delta) \subset U$ and for every $x \in M \cap D(\overline{x}, \delta)$, $f(\overline{x}) \leq f(x)$. For all $k \in \mathbb{N}^*$ we consider the function $\varphi_k : D(\overline{x}, \delta) \to \mathbb{R}$ given by

$$\varphi_k(x) = f(x) + \frac{k}{2} \sum_{i=1}^{n} \left(g_i^+(x)\right)^2 + \frac{k}{2} \sum_{j=1}^{m} \left(h_j(x)\right)^2 + \frac{1}{2} \|x - \overline{x}\|^2,$$

where $g_i^+(x) = \max\{g_i(x), 0\}$. Clearly, φ_k attains its minimum on $D(\overline{x}, \delta)$ and we denote by x_k such a minimum point. We also observe that

$$0 \le \varphi_k(x_k) = f(x_k) + \frac{k}{2}\sum_{i=1}^{n}\left(g_i^+(x_k)\right)^2 + \frac{k}{2}\sum_{j=1}^{m}\left(h_j(x_k)\right)^2 + \frac{1}{2}\|x_k - \overline{x}\|^2$$
$$\le \varphi_k(\overline{x}) = f(\overline{x}).$$

Since the sequence (x_k) is bounded and f in continuous on $D(\overline{x}, \delta)$, we infer that $(f(x_k))$ is also a bounded sequence. Letting $k \to \infty$ in the above relation, we get

$$\lim_{k\to\infty}\sum_{i=1}^{n}\left(g_i^+(x_k)\right)^2 = 0$$
$$\lim_{k\to\infty}\sum_{j=1}^{m}\left(h_j(x_k)\right)^2 = 0.$$

The boundedness of (x_k) ensures that one can extract a convergent subsequence of it. Without relabeling, we can write $x_k \to x^* \in D(\overline{x}, \delta)$, and the previous relations yield $x^* \in M$. Consequently, passing to the limit in the inequality above, we have

$$f(x^*) + \frac{1}{2}\left\|x^* - \overline{x}\right\|^2 \le f(\overline{x}).$$

On the other hand, $f(\overline{x}) \le f(x^*)$, so $\left\|x^* - \overline{x}\right\| = 0$, that is $x^* = \overline{x}$. Therefore $x_k \to \overline{x}$.

An essential remark here is that φ_k is differentiable since the (nondifferentiable) scalar functions $g_i^+(x)$ are squared, whence $\nabla\left(g_i^+(x)\right)^2 = g_i^+(x)\nabla g(x)$. Since x_k is a minimum for φ_k on $D(\overline{x}, \delta)$, we deduce that

$$-\nabla\varphi_k(x_k) \in N(D(\overline{x}, \delta), x_k).$$

For k sufficiently large, x_k belongs to the interior of the ball $D(\overline{x}, \delta)$ and we conclude that for these numbers k, one has $N(D(\overline{x}, \delta), x_k) = \{0\}$. The combination of these facts allow us to write

$$\nabla f(x_k) + k\sum_{i=1}^{n} g_i^+(x_k)\nabla g(x_k) + k\sum_{j=1}^{m} h_j(x_k)\nabla h_j(x_k) + x_k - \overline{x} = 0, \tag{3.2.1}$$

for every k large enough. For $i \in \overline{1, n}$, $j \in \overline{1, m}$, we denote $\alpha_i^k := kg_i^+(x_k)$, $\beta_j^k := kh_j(x_k)$ and $y^k = \sqrt{1 + \sum_{i=1}^{n}\left(\alpha_i^k\right)^2 + \sum_{j=1}^{m}\left(\beta_i^k\right)^2}$. It is clear that $y^k > 1$ and we take $\lambda_0^k := \frac{1}{y^k}$, $\lambda_i^k := \frac{\alpha_i^k}{y^k}$, $\mu_j^k := \frac{\beta_j^k}{y^k}$. We observe that

$$\left(\lambda_0^k\right)^2 + \sum_{i=1}^{n}\left(\lambda_i^k\right)^2 + \sum_{j=1}^{m}\left(\mu_j^k\right)^2 = 1,$$

whence the sequences (λ_0^k), (λ_i^k), (μ_j^k) $(i \in \overline{1,n}, j \in \overline{1,m})$ are bounded. Then there exist subsequences (we keep the indexes) convergent to some real numbers, respectively denoted by

$$\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_n, \ \mu_1, \mu_2, \ldots, \mu_m.$$

These numbers cannot be zero simultaneously. The positivity of the terms of the sequences (λ_0^k), (λ_i^k) $(i \in \overline{1,n})$ implies the positivity of their limits $\lambda_0, \lambda_1, \lambda_2, \ldots, \lambda_n$. Now, we divide relation (3.2.1) by y^k, and we get

$$\lambda_0^k \nabla f(x_k) + \sum_{i=1}^{n} \lambda_i^k \nabla g(x_k) + \sum_{j=1}^{m} \mu_j^k \nabla h_j(x_k) + \frac{1}{y^k}(x_k - \overline{x}) = 0.$$

Letting $k \to \infty$ we have the first relation in the conclusion. Now we show the second one. Let $i \in \overline{1,n}$. If $\lambda_i = 0$, there is nothing to prove. Otherwise, if $\lambda_i > 0$, from the definition of λ_i we infer that for k sufficiently large, $g_i^+(x_k) > 0$, whence $g_i^+(x_k) = g_i(x_k)$. The relation

$$0 < g_i(x_k) \to g_i(\overline{x}) \leq 0$$

leads us to the conclusion $g_i(\overline{x}) = 0$. So, the second part of the conclusion holds and the theorem is completely proved. □

The relations in the conclusion of Theorem 3.2.1 are called Fritz John necessary optimality conditions. The major drawback of this result is that it does not eliminate the possibility that the real number associated to the objective function (i.e., λ_0) can be zero. This means that it would be possible to have too many points where the conditions in the conclusion are satisfied and therefore, in such a case, the result would not give important practical hints on the solutions. For instance, if a feasible point x satisfies $\nabla g_i(x) = 0$ for a certain $i \in A(x)$ or $\nabla h_j(x) = 0$ for an $j \in \overline{1,m}$, then it satisfies Fritz John conditions (with $\lambda_0 = 0$), the objective function being then completely eliminated. In the next subsection we shall impose a condition in order to avoid $\lambda_0 = 0$.

Let us first illustrate the possibilities created by Theorem 3.2.1 through two concrete examples.

Example 3.2.2. *Let us consider the problem of minimization of* $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 2)^2$$

under the restriction $g(x) \leq 0$, *where* $g : \mathbb{R}^2 \to \mathbb{R}^4$,

$$g(x_1, x_2) = (x_1^2 + x_2^2 - 5, x_1 + 2x_2 - 4, -x_1, -x_2).$$

One can observe graphically that $\overline{x} = (2, 1)$ *is solution of the problem and* $A(\overline{x}) = \{1, 2\}$. *We want to verify Fritz John condition at this point. From the second condition, since* $3, 4 \notin A(\overline{x})$, *we get* $\lambda_3 = \lambda_4 = 0$. *Since* $\nabla f(\overline{x}) = (-2, -2)$, $\nabla g_1(\overline{x}) = (4, 2)$, $\nabla g_2(\overline{x}) =$

$(1,2)$, *we have to find positive real numbers* $\lambda_0, \lambda_1, \lambda_2 \geq 0$, *not simultaneously zero, such that*

$$\lambda_0(-2,-2) + \lambda_1(4,2) + \lambda_2(1,2) = (0,0).$$

We get $\lambda_1 = \frac{1}{3}\lambda_0$ *and* $\lambda_2 = \frac{2}{3}\lambda_0$, *whence, by taking* $\lambda_0 > 0$, *the first Fritz John condition is fulfilled.*

Let us now have a look to the point $x = (0,0)$. *This time* $A(x) = \{3,4\}$, *whence* $\lambda_1 = \lambda_2 = 0$. *We have that* $\nabla f(\overline{x}) = (-6,-4)$, $\nabla g_3(\overline{x}) = (-1,0)$, $\nabla g_4(\overline{x}) = (0,-1)$. *A computation shows that the equation*

$$\lambda_0(-6,-4) + \lambda_3(-1,0) + \lambda_4(0,-1) = (0,0)$$

has no solution $(\lambda_0, \lambda_3, \lambda_4)$ *different to zero with positive components. Then* x *does not fulfill the Fritz John conditions, hence it is not a minimum point for the given problem.*

Example 3.2.3. *Let us consider the problem of minimization of* $f : (0,\infty)\times(0,\infty) \to \mathbb{R}$, $f(x_1,x_2) = -2x_2$ *under the restriction* $g(x) \leq 0$, *where* $g : \mathbb{R}^2 \to \mathbb{R}^3$, $g(x_1,x_2) = (x_1 - x_2 - 2, -x_1 + x_2 + 2, x_1 + x_2 - 6)$. *The set* M *of feasible points is* $[(2,0),(4,2)] \setminus \{(2,0)\}$, *and the minimum point is* $\overline{x} = (4,2)$. *It is easy to observe that any feasible point satisfies the Fritz John conditions, but the solution is the only point where one can choose* $\lambda_0 \neq 0$. *Indeed, if* x *is a feasible point different to* $\overline{x}$, *then* $\lambda_3 = 0$, $\lambda_1 = \lambda_2$ *and* $\lambda_0 = 0$.

3.2.2 Karush-Kuhn-Tucker Conditions

As seen before, it is desirable to have a Fritz John type result, but with $\lambda_0 \neq 0$. We could directly impose an extra condition in Theorem 3.2.1 in order to ensure this, but we prefer a direct approach because we aim at working with weak assumptions.

Let consider the sets

$$G(\overline{x}) = \left\{ \sum_{i\in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) \mid \lambda_i \geq 0,\ \forall i \in A(\overline{x}), \mu_j \in \mathbb{R}, \forall j \in \overline{1,m} \right\} \subset \mathbb{R}^p.$$

(where, as usual, we used the identification between $L(\mathbb{R}^p,\mathbb{R})$ and $\mathbb{R}^p$) and

$$D(\overline{x}) = \left\{ u \in \mathbb{R}^p \mid \nabla g_i(\overline{x})(u) \leq 0,\ \forall i \in A(\overline{x}) \text{ and } \nabla h_j(\overline{x})(u) = 0,\ \forall j \in \overline{1,m} \right\}.$$

Before the main result, we need to shed some light on some important relations for these sets.

Proposition 3.2.4. *For every* $\overline{x} \in M$ *we have:*

(i) $G(\overline{x}) = D(\overline{x})^-$;

(ii) $T_B(M,\overline{x}) \subset D(\overline{x})$.

Proof (i) The inclusion $G(\overline{x}) \subset D(\overline{x})^-$ is obvious, while the reverse one is a direct consequence of Farkas Lemma (Theorem 2.1.8).

(ii) Clearly, $0 \in D(\overline{x})$. Let $u \in T_B(M, \overline{x}) \setminus \{0\}$. By the definition of tangent vectors, there exist $(t_n) \subset (0, \infty)$, $t_n \to 0$ and $(u_n) \to u$ such that for every n,

$$\overline{x} + t_n u_n \in M.$$

The sequence $(t_n u_n)$ converges towards 0 in $\mathbb{R}^p$. Taking into account the differentiability of h at $\overline{x}$, there exists $(\alpha_n) \subset \mathbb{R}^p$, $\alpha_n \to 0$ such that for every $n \in \mathbb{N}$,

$$h(\overline{x} + t_n u_n) = h(\overline{x}) + t_n \nabla h(\overline{x})(u_n) + t_n \|u_n\| \alpha_n.$$

Since $h(\overline{x} + t_n u_n) = h(\overline{x}) = 0$, dividing by t_n and passing to the limit as $n \to \infty$, we get $\nabla h(\overline{x})(u) = 0$. Now, for every $i \in A(\overline{x})$ there exist $(\alpha_n^i) \subset \mathbb{R}$, $\alpha_n^i \to 0$ such that for every $n \in \mathbb{N}$,

$$g_i(\overline{x} + t_n u_n) = g_i(\overline{x}) + t_n \nabla g_i(\overline{x})(u_n) + t_n \|u_n\| \alpha_n^i.$$

As before, since $g_i(\overline{x} + t_n u_n) \leq 0$ and $g_i(\overline{x}) = 0$, we have $\nabla g_i(\overline{x})(u) \leq 0$, and the proposition is proved. □

The next example shows that the reverse inclusion in the item (*ii*) above is false.

Example 3.2.5. *Let $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x_1, x_2) = -x_1 - x_2$, $h : \mathbb{R}^2 \to \mathbb{R}$, $h(x_1, x_2) = x_1 x_2$ and the feasible point $\overline{x} = (0, 0)$. Then:*

$$D(\overline{x}) = \{(u_1, u_2) \mid -u_1 - u_2 \leq 0\},$$
$$T_B(M, \overline{x}) = \{(u_1, u_2) \mid u_1 \geq 0, u_2 \geq 0, u_1 u_2 = 0\}.$$

We establish now a generalized form of a classical result known under the name of Karush-Kuhn-Tucker Theorem, since it was obtained (with stronger assumptions) by the American mathematicians William Karush, Harold William Kuhn and Albert William Tucker. It is interesting to note that William Karush obtained the result in 1939, but the mathematical community become aware of its importance when Harold William Kuhn and Albert William Tucker got the result, in a different way, in 1950.

Theorem 3.2.6 (Karush-Kuhn-Tucker)**.** *Let $\overline{x} \in M$ be a solution of the problem (P). Suppose that $T_B(M, \overline{x})^- = D(\overline{x})^-$. Then there exist $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \mathbb{R}^n$, $\mu = (\mu_1, \mu_2, \ldots, \mu_m) \in \mathbb{R}^m$, such that*

$$\nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0 \tag{3.2.2}$$

and

$$\lambda_i \geq 0, \ \lambda_i g_i(\overline{x}) = 0, \ \textit{for every } i \in \overline{1, n}. \tag{3.2.3}$$

Proof From Theorem 3.1.18, $\nabla f(\overline{x})(u) \geq 0$ for every $u \in T_B(M, \overline{x})$, whence $-\nabla f(\overline{x}) \in T_B(M, \overline{x})^-$. We use now the assumption $T_B(M, \overline{x})^- = D(\overline{x})^-$ to infer that $-\nabla f(\overline{x}) \in D(\overline{x})^-$. From Proposition 3.2.4 (i), we get $-\nabla f(\overline{x}) \in G(\overline{x})$. Consequently, there exist $\lambda_i \geq 0$, $i \in A(\overline{x})$, $\mu_j \in \mathbb{R}$, $j \in \overline{1, m}$ such that $-\nabla f(\overline{x}) = \sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x})$. Now, for indexes $i \in \overline{1, n} \setminus A(\overline{x})$ we take $\lambda_i = 0$, and we obtain the conclusion. □

If one compares Theorem 3.2.6 and Theorem 3.2.1, one notices the announced difference concerning the real number associated to the objective function.

The function $L : U \times \mathbb{R}^{n+m} \to \mathbb{R}$,

$$L(x, (\lambda, \mu)) := f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x)$$

is called the Lagrangian of (P). Therefore, the conclusion given by relation $(3.2.2)$ can be written as

$$\nabla_x L(\overline{x}, (\lambda, \mu)) = 0,$$

and the elements $(\lambda, \mu) \in \mathbb{R}^n_+ \times \mathbb{R}^m$ are called Lagrange multipliers. This name is due to the fact that the first time this method was used to investigate constrained optimization problems was given in some of Lagrange's works on calculus of variations problems.

The preceding theorem does not ensure the uniqueness of these multipliers. We denote by $M(\overline{x})$ the set of Lagrange multipliers at $\overline{x}$, i.e.,

$$M(\overline{x}) := \{(\lambda, \mu) \in \mathbb{R}^n_+ \times \mathbb{R}^m \mid \nabla_x L(\overline{x}, (\lambda, \mu)) = 0\},$$

where $\mathbb{R}^n_+ := [0, \infty)^n$.

On the other hand, $L(x, (\lambda, \mu))$ is an affine function with respect to the variables (λ, μ). We can observe the following fact which will appear later in the discussion: if $\overline{x} \in M$ and $(\overline{\lambda}, \overline{\mu})$ is a maximum on $\mathbb{R}^n_+ \times \mathbb{R}^m$ for $(\lambda, \mu) \mapsto L(\overline{x}, (\lambda, \mu))$, then $\overline{\lambda}_i g_i(\overline{x}) = 0$ for every $i \in \overline{1, n}$.

Theorem 3.2.6 gives necessary optimality conditions for (P). If, instead of minimization, we are looking for maximization of the objective function f under the same constraints, then, from $\max f = -\min(-f)$, the necessary condition (3.2.2) can be written as

$$-\nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

Furthermore, let us notice that if one has only equalities as constraints, taking into account that $h(x) = 0$ is equivalent to $-h(x) = 0$, the necessary optimality condition can by written, for both maxima and minima, as

$$\nabla f(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

Coming back to the main results, let us observe two more things. Firstly, if the problem has no restrictions (for instance, $U = M = \mathbb{R}^p$), then relation (3.2.2) reduces to the first-order necessary optimality condition (Fermat Theorem): $\nabla f(\overline{x}) = 0$. Secondly, the key relation (3.2.2) does not hold without supplementary conditions (here, $T_B(M, \overline{x})^- = D(\overline{x})^-$). To illustrate this consider the following example.

Example 3.2.7. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}^2$ given by $f(x_1, x_2) = x_1$ and $g(x_1, x_2) = (-x_2 + (1 - x_1)^3, x_2)$. It is easy to see that $\overline{x} = (1, 0)$ is a minimum point of the associated problem, but (3.2.2) does not hold. Clearly, Fritz John conditions are fulfilled for $\lambda_0 = 0$.*

So, in the next section, every condition which ensures the validity of the Karush-Kuhn-Tucker Theorem is called a qualification condition, and in view of the decisive importance of such requirements, we shall discuss it into detail in the next section.

Before that, let us observe that under certain assumptions, Karush-Kuhn-Tucker conditions (3.2.2) and (3.2.3) are also sufficient for minimality.

Theorem 3.2.8. *Suppose that U is convex, f is convex on U, h is affine and g_i, $i \in \overline{1, n}$ are convex. Let $\overline{x} \in M$. If there exists $(\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^m$ such that* (3.2.2) *and* (3.2.3) *hold, then $\overline{x}$ is a minimum point for (P) (or minimum of f on M).*

Proof The condition (3.2.2) expresses the fact that

$$\nabla_x L(\overline{x}, (\lambda, \mu)) = 0.$$

Under our assumptions, L is a convex function in x, so according to Theorem 3.1.22, $\overline{x}$ is a minimum (without constraints) of the map $x \mapsto L(x, (\lambda, \mu))$. Therefore, for every $x \in U$,

$$L(x, (\lambda, \mu)) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x) \geq L(\overline{x}, (\lambda, \mu)) = f(\overline{x}).$$

But, for any $x \in M$,

$$\sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x) \leq 0,$$

whence $f(x) \geq f(\overline{x})$. The proof is complete. □

Concerning the structure of the set of Lagrange multipliers, we have the following result.

Proposition 3.2.9. *For data with the structure mentioned in the above theorem, the set $M(\overline{x})$ of the Lagrange multipliers is the same for all minimum points of f an M.*

Proof Clearly, M is a convex set. Let $x_1, x_2 \in M$ be two minimum points of (P). According to Proposition 3.1.23, one has $f(x_1) = f(x_2)$. Let $(\lambda, \mu) \in M(x_1)$. Then

$$\nabla f(x_1) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x_1) + \sum_{j=1}^{m} \mu_j \nabla h_j(x_1) = 0$$

and

$$\lambda_i \geq 0, \ \lambda_i g_i(x_1) = 0, \ \text{for every } i \in \overline{1, n}.$$

As before,

$$f(x_2) + \sum_{i=1}^{n} \lambda_i g_i(x_2) \geq f(x_1) = f(x_2).$$

Taking into account the information on the numbers λ_i and $g_i(x_2)$, we infer that $\lambda_i g_i(x_2) = 0$ for every $i \in \overline{1, n}$. From

$$L(x_2, (\lambda, \mu)) = f(x_2) = f(x_1) = L(x_1, (\lambda, \mu)),$$

we get that x_2 is a minimum point for the convex function $L(\cdot, (\lambda, \mu))$ on U. Hence

$$\nabla f(x_2) + \sum_{i=1}^{n} \lambda_i \nabla g_i(x_2) + \sum_{j=1}^{m} \mu_j \nabla h_j(x_2) = 0.$$

We have that $(\lambda, \mu) \in M(x_2)$. The other inclusion follows by exchanging x_1 and x_2 in the above proof. □

We now interpret Theorem 3.2.6 by using the concept of saddle point applied to the Lagrangian function. Firstly, we define the concept.

Definition 3.2.10. *Let X, Y be two sets and $F : X \times Y \to \mathbb{R}$. A saddle point of F is a pair $(\overline{x}, \overline{y}) \in X \times Y$ with the property that*

$$\max_{y \in Y} F(\overline{x}, y) = F(\overline{x}, \overline{y}) = \min_{x \in X} F(x, \overline{y}). \tag{3.2.4}$$

It is clear that the relation (3.2.4) is equivalent to

$$F(\overline{x}, y) \leq F(\overline{x}, \overline{y}) \leq F(x, \overline{y}), \ \forall (x, y) \in X \times Y$$

and to

$$F(\overline{x}, y) \leq F(x, \overline{y}), \ \forall (x, y) \in X \times Y.$$

For instance, the point $(0, 0)$ is a saddle point of $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $F(x, y) = x^2 - y^2$ (the figure below).

The following general result is in order.

Proposition 3.2.11. *For all saddle points $(\overline{x}, \overline{y})$ of F, the value $F(\overline{x}, \overline{y})$ is constant. If (x_1, y_1) and (x_2, y_2) are saddle points, then (x_1, y_2) and (x_2, y_1) are saddle points as well.*

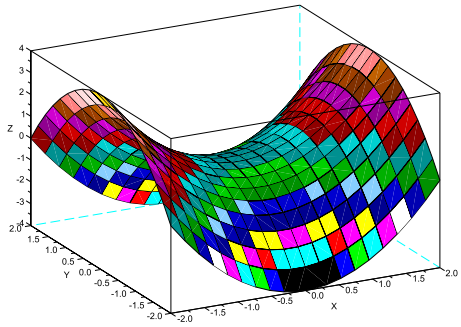

Figure 3.2: A saddle point.

Proof The following relations hold:

$$F(x_1, y) \le F(x_1, y_1) \le F(x, y_1),\ \forall (x, y) \in X \times Y$$
$$F(x_2, y) \le F(x_2, y_2) \le F(x, y_2),\ \forall (x, y) \in X \times Y.$$

If, in the first one, we take $x = x_2$ and $y = y_2$, and in the second one we put $x = x_1$ and $y = y_1$, we get $F(x_1, y_1) = F(x_2, y_2) = F(x_2, y_1) = F(x_1, y_2)$. Moreover, we can write for every $(x, y) \in X \times Y$,

$$F(x_1, y) \le F(x_1, y_2) \le F(x, y_2),$$

whence (x_1, y_2) is a saddle point. For (x_2, y_1), the proof is similar. □

For the general form of problem (P), we consider again the Lagrangian function $L : U \times (\mathbb{R}^n_+ \times \mathbb{R}^m) \to \mathbb{R}$,

$$L(x, (\lambda, \mu)) = f(x) + \sum_{i=1}^{n} \lambda_i g_i(x) + \sum_{j=1}^{m} \mu_j h_j(x).$$

Theorem 3.2.12. *An element $(\overline{x}, (\overline{\lambda}, \overline{\mu})) \in U \times (\mathbb{R}^n_+ \times \mathbb{R}^m)$ is a saddle point for the Lagrangian function L if and only if the following relations hold:*

(i) $\overline{x}$ is a minimum point for $L(\cdot, (\overline{\lambda}, \overline{\mu}))$ on the open set U;

(ii) $\overline{x} \in M$;

(iii) $\lambda_i g_i(\overline{x}) = 0$, for every $i \in \overline{1, n}$.

Proof Let $(\overline{x}, (\overline{\lambda}, \overline{\mu})) \in U \times (\mathbb{R}^n_+ \times \mathbb{R}^m)$ be a saddle point for L. Then, according to the definition,

$$\max_{(\lambda,\mu)\in\mathbb{R}^n_+\times\mathbb{R}^m} L(\overline{x}, (\lambda, \mu)) = L(\overline{x}, (\overline{\lambda}, \overline{\mu})) = \min_{x\in U} L(x, (\overline{\lambda}, \overline{\mu})).$$

The second part of this relation is equivalent to (i). It remains to be shown that the first equality is equivalent to the combination of (ii) and (iii), and this is based on the fact that L is affine with respect to (λ, μ) and, moreover, the particular form of $\mathbb{R}^n_+ \times \mathbb{R}^m$ allows us to easily compute the polar of its Bouligand tangent cone. According to Proposition 3.1.25, $(\overline{\lambda}, \overline{\mu})$ with the property

$$\max_{(\lambda,\mu)\in\mathbb{R}^n_+\times\mathbb{R}^m} L(\overline{x}, (\lambda, \mu)) = L(\overline{x}, (\overline{\lambda}, \overline{\mu}))$$

is characterized by the relation

$$-\nabla_{(\lambda,\mu)} L(\overline{x}, (\overline{\lambda}, \overline{\mu})) \in N(\mathbb{R}^n_+ \times \mathbb{R}^m, (\overline{\lambda}, \overline{\mu})).$$

It is not difficult to see that

$$N(\mathbb{R}^n_+ \times \mathbb{R}^m, (\overline{\lambda}, \overline{\mu})) = \{u \in \mathbb{R}^n \mid u_i = 0 \text{ if } \overline{\lambda}_i > 0,\, u_i \leq 0 \text{ if } \overline{\lambda}_i = 0\} \times \{0\}_{\mathbb{R}^m}.$$

Then

$$\frac{\partial L}{\partial \lambda_i}(\overline{x}, (\overline{\lambda}, \overline{\mu})) = g_i(\overline{x}) : \begin{cases} = 0, \text{ if } \overline{\lambda}_i > 0 \\ \leq 0, \text{ if } \overline{\lambda}_i = 0 \end{cases}, \ \forall i = \overline{1, n},$$

and

$$\frac{\partial L}{\partial \mu_j}(\overline{x}, (\overline{\lambda}, \overline{\mu})) = h_j(\overline{x}) = 0, \ \forall j = \overline{1, m}.$$

The proof is complete. □

Corollary 3.2.13. *If $(\overline{x}, (\overline{\lambda}, \overline{\mu})) \in U \times (\mathbb{R}^n_+ \times \mathbb{R}^m)$ is a saddle point for the Lagrangian function L, then $\overline{x}$ is a solution of (P).*

Proof The preceding result shows that $\overline{x} \in M$ and

$$f(\overline{x}) = L(\overline{x}, (\overline{\lambda}, \overline{\mu})) \leq L(x, (\overline{\lambda}, \overline{\mu})), \ \forall x \in U.$$

Since for $x \in M$,

$$L(x, (\overline{\lambda}, \overline{\mu})) \leq f(x),$$

we get $f(\overline{x}) \leq f(x)$ for every $x \in M$. □

As usual, for convex data the converse holds as well.

Theorem 3.2.14. *Suppose that U is convex, f is convex on U, h is affine and g_i, $i \in \overline{1, n}$ are convex. The next relations are equivalent:*

(i) $(\overline{x}, (\overline{\lambda}, \overline{\mu})) \in U \times (\mathbb{R}^n_+ \times \mathbb{R}^m)$ is a saddle point for the Lagrangian function L;

(ii) $\overline{x}$ is a minimum point for (P) and $(\overline{\lambda}, \overline{\mu})$ is a Lagrange multiplier.

Proof According to Theorem 3.2.12, relation (i) above is equivalent to all three relations in that result. One applies now Theorem 3.1.22 and the conclusion follows. □

3.2.3 Qualification Conditions

The qualification condition $T_B(M, \overline{x})^- = D(\overline{x})^-$ imposed in Theorem 3.2.6 is called the Guignard condition at $\overline{x}$ (after the name of the French mathematician Monique Guignard who proposed it back in 1969) and it is one of the weakest qualification conditions. The difficulty with this condition is that the effective calculations of the involved objects can be tricky in certain situation, and for this reason we want to investigate and to compare it with other qualification conditions as well. Clearly, relation $T_B(M, \overline{x}) = D(\overline{x})$ is in turn a qualification condition (called the quasiregularity condition), since implies Guignard condition. As expected, the two conditions are not equivalent, as one can see from the next example (see also Example 2.1.7).

Example 3.2.15. *Let $g : \mathbb{R}^2 \to \mathbb{R}^2$, $g(x_1, x_2) = (-x_1, x_2)$ and $h : \mathbb{R}^2 \to \mathbb{R}$, $h(x_1, x_2) = x_1 x_2$. Let us consider the feasible point $\overline{x} = (0, 0)$. Then:*

$$D(\overline{x}) = \{(u_1, u_2) \mid u_1 \geq 0, u_2 \leq 0\},$$
$$T_B(M, \overline{x}) = \{(u_1, u_2) \mid u_1 \geq 0, u_2 \leq 0, u_1 u_2 = 0\}$$

and

$$T_B(M, \overline{x})^- = D(\overline{x})^- = \{(u_1, u_2) \mid u_1 \leq 0, u_2 \geq 0\}.$$

The qualification conditions are linked to the reference point ($\overline{x}$ in our notation). Every time when no confusion concerning the reference point could appear, we avoid, for simplicity, writing it explicitly.

Two of the most important (from a practical point of view) qualification conditions are listed below. The first one is called the linear independence qualification condition (at $\overline{x}$) and is as follows:

$$\text{the set } \{\nabla g_i(\overline{x}) \mid i \in A(\overline{x})\} \cup \{\nabla h_j(\overline{x}) \mid j \in \overline{1, m}\} \text{ is linearly independent.}$$

The second one is called Mangasarian-Fromovitz qualification condition (at $\overline{x}$):

$$\text{the set } \{\nabla h_j(\overline{x}) \mid j \in \overline{1, m}\} \text{ is linearly independent and}$$
$$\exists u \in \mathbb{R}^p : \nabla h(\overline{x})(u) = 0 \text{ and } \nabla g_i(\overline{x})(u) < 0, \ \forall i \in A(\overline{x}).$$

(The American mathematicians Olvi Leon Mangasarian and Stanley Fromovitz published this condition in 1967.)

We will now establish the relations between these conditions and then show that they are indeed qualification conditions.

Theorem 3.2.16. *If the linear independence qualification condition at $\overline{x} \in M$ holds, then the Mangasarian-Fromovitz qualification condition at $\overline{x}$ is satisfied.*

Proof Without loss of generality, we suppose that $A(\overline{x}) = \{1, \ldots, q\}$. Let T be the matrix of dimensions $(q + m) \times p$ with the lines $\nabla g_i(\overline{x})$, $i \in \overline{1, q}$, $\nabla h_j(\overline{x})$, $j \in \overline{1, m}$ and let b be the column vector with $b_i = -1$, $i \in \overline{1, q}$, $b_j = 0$, $j \in \overline{q + 1, q + m}$. Since the lines of T are linearly independent, the system $Td = b$ has a solution. If one denotes by u such a solution, then

$$\nabla g_i(\overline{x})(u^t) = -1, \ \forall i \in \overline{1, q} \text{ and } \nabla h_j(\overline{x})(u^t) = 0, \ \forall j \in \overline{1, m},$$

hence the Mangasarian-Fromovitz condition at $\overline{x}$ is satisfied. □

The two conditions are not, however, equivalent.

Example 3.2.17. *Let $g_i : \mathbb{R}^2 \to \mathbb{R}$, $i \in \overline{1, 3}$ defined by:*

$$g_1(x) = (x_1 - 1)^2 + (x_2 - 1)^2 - 2$$
$$g_2(x) = (x_1 - 1)^2 + (x_2 + 1)^2 - 2$$
$$g_3(x) = -x_1$$

and the feasible point $\overline{x} = (0, 0)$. Surely, the set

$$\{\nabla g_1(\overline{x}), \nabla g_2(\overline{x}), \nabla g_3(\overline{x}), \ i \in \overline{1, 3}\}$$

is not linearly independent since it consists of three elements in the two dimensional space $\mathbb{R}^2$. On the other hand, for $u = (1, 0)$, $\nabla g_i(\overline{x})(u) < 0$ for every $i \in \overline{1, 3}$.

Theorem 3.2.16 tells us that in order to show that the two conditions above are qualifications conditions, it is enough to show this only for Mangasarian-Fromovitz condition. This becomes obvious if one applies Theorem 3.2.1 and argues by contradiction. Suppose that $\lambda_0 = 0$. Then

$$\sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0.$$

We multiply by the vector u from Mangasarian-Fromovitz condition and deduce that

$$\sum_{i \in A(\overline{x})} \lambda_i \left\langle \nabla g_i(\overline{x}), u \right\rangle = 0,$$

whence $\lambda_i = 0$ for every $i \in A(\overline{x})$. Therefore,

$$\sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0,$$

and the linear independence of the gradients $\{\nabla h_j(\overline{x}) \mid j \in \overline{1, m}\}$ implies that $\mu_j = 0$ for every $j \in \overline{1, m}$. Putting together these remarks, we get the contradiction to $|\lambda_0| + \|\lambda\| + \|\mu\| \neq 0$. Consequently, $\lambda_0 \neq 0$.

In order to more precisely classify the qualification conditions introduced so far, we show that the Mangasarian-Fromovitz condition implies the quasiregularity condition.

We need an auxiliary result.

Lemma 3.2.18. *Let $\varepsilon > 0$ and $y : (-\varepsilon, \varepsilon) \to \mathbb{R}^p$ be a differentiable function such that $y(0) = \overline{x}$, $y'(0) = u \neq 0$. Then there exists a sequence $(x_k) \subset \operatorname{Im} y \setminus \{\overline{x}\}$, $(x_k) \to \overline{x}$ such that*

$$\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} \to \frac{u}{\|u\|}.$$

Proof We have

$$\lim_{t \to 0} \frac{y(t) - \overline{x}}{t} = \lim_{t \to 0} \frac{y(t) - y(0)}{t} = y'(0) = u \neq 0.$$

In particular, for $t \neq 0$ sufficiently small one has $y(t) \neq \overline{x}$. We consider a sequence $(t_k) \to 0$ of positive numbers and we define $x_k = y(t_k)$. Then,

$$\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} = \frac{x_k - \overline{x}}{t_k} \frac{t_k}{\|x_k - \overline{x}\|} \to \frac{u}{\|u\|}.$$

This ends the proof. □

Theorem 3.2.19. *If the Mangasarian-Fromovitz condition is satisfied at $\overline{x} \in M$, then $T_B(M, \overline{x}) = D(\overline{x})$.*

Proof As already observed, one inclusion is always true. We show only the opposite one, so we start with an element $u \in D(\overline{x})$. Denote by $\overline{u} \in \mathbb{R}^p$ the vector given by the Mangasarian-Fromovitz condition. Let $\lambda \in (0, 1)$ and $d_\lambda := (1 - \lambda)u + \lambda\overline{u}$. We show that $d_\lambda \in T_B(M, \overline{x})$ for every $\lambda \in (0, 1)$, and then, taking $\lambda \to 0$ and using the closedness of $T_B(M, \overline{x})$ the conclusion will follow. Suppose that $d_\lambda \neq 0$, since otherwise, there is nothing to prove.

Let P be the operator defined by the matrix (of dimensions $m \times p$) which has on the lines the vectors $\nabla h_j(\overline{x})$, $j \in \overline{1, m}$ of $\mathbb{R}^p$. These vectors are linearly independent and form a basis in the linear space $\operatorname{Im}(P)$. Clearly, from the linear independence of $\nabla h_j(\overline{x})$, $j \in \overline{1, m}$ one deduces that $m \leq p$. But $p = \dim(\operatorname{Im}(P)) + \dim(\operatorname{Ker}(P))$, and we complete the above linear independent set up to a base of $\mathbb{R}^p$ with a set of vectors $\{v_1, v_2, \ldots, v_{p-m}\}$, and we denote by Z the matrix (of dimensions $(p - m) \times p$) which has on the lines these vectors (which give a base in $\operatorname{Ker}(P)$). Then the square matrix $\begin{pmatrix} P \\ Z \end{pmatrix}$ is nonsingular. We define $\varphi : \mathbb{R}^{p+1} \to \mathbb{R}^p$ by

$$\varphi(x, \tau) = (h(x), (Z(x - \overline{x} - \tau d_\lambda)^t)^t).$$

Then $\nabla_x \varphi(\overline{x}, 0) = \begin{pmatrix} P \\ Z \end{pmatrix}$ is a nonsingular matrix, whence, from Implicit Functions Theorem (Theorem 1.3.5), there exist $\varepsilon > 0$ and a differentiable function $y : (-\varepsilon, \varepsilon) \to$

$\mathbb{R}^p$ such that

$$\varphi\left(y(\tau), \tau\right) = 0,$$

for every $\tau \in (-\varepsilon, \varepsilon)$. Then

$$h(y(\tau)) = 0 \text{ and } Z(y(\tau) - \overline{x} - \tau d_\lambda)^t = 0. \tag{3.2.5}$$

At the same time, for every $\tau \in (-\varepsilon, \varepsilon)$ and every x close enough to $\overline{x}$ we have

$$\varphi(x, \tau) = 0 \Rightarrow x = y(\tau).$$

Since $\varphi(\overline{x}, 0) = 0$, we infer that $y(0) = \overline{x}$. According to the relations (3.2.5), we get, on one hand (by differentiation),

$$Py'(0) = 0,$$

and, on the other hand (by dividing with $\tau \neq 0$ and passing to the limit),

$$Zy'(0)^t = Zd_\lambda^t.$$

Since $u, \overline{u} \in D(\overline{x})$, we get $P(d_\lambda) = 0$. We obtain

$$\begin{pmatrix} P \\ Z \end{pmatrix} (y'(0)) = \begin{pmatrix} P \\ Z \end{pmatrix} (d_\lambda),$$

that is $d_\lambda = y'(0)$. Using the above lemma, there exists a sequence $(x_k) \subset \operatorname{Im} y \setminus \{\overline{x}\}$, $(x_k) \to \overline{x}$ with

$$\frac{x_k - \overline{x}}{\|x_k - \overline{x}\|} \to \frac{d_\lambda}{\|d_\lambda\|}.$$

Then $h(x_k) = 0$. In order to deduce that $d_\lambda \in T_B(M, \overline{x})$, it is sufficient to prove that, for k large enough, $g(x_k) \leq 0$. If $i \notin A(\overline{x})$, then $g_i(\overline{x}) < 0$, and the continuity of g_i implies that $g_i(x_k) < 0$ for large k. If $i \in A(\overline{x})$, $\left\langle \nabla g_i(\overline{x}), u \right\rangle \leq 0$ and $\left\langle \nabla g_i(\overline{x}), \overline{u} \right\rangle < 0$, hence $\left\langle \nabla g_i(\overline{x}), d_\lambda \right\rangle < 0$. Since g_i is smooth (of class C^1), there exists a sequence $(\alpha_k) \to 0$ such that for every $k \in \mathbb{N}$,

$$g_i(x_k) = g_i(\overline{x}) + \nabla g_i(\overline{x})(x_k - \overline{x}) + \alpha_k \|x_k - \overline{x}\|.$$

Therefore,

$$\frac{g_i(x_k)}{\|x_k - \overline{x}\|} = \frac{\nabla g_i(\overline{x})(x_k - \overline{x})}{\|x_k - \overline{x}\|} + \alpha_k \overset{k \to \infty}{\to} \nabla g_i(\overline{x}) \left(\frac{d_\lambda}{\|d_\lambda\|} \right) < 0.$$

Then $g_i(x_k) < 0$ for sufficiently large k . Since there are a finite number of indexes i, we obtain the conclusion. □

In order to show that all four qualification conditions introduced are different, it remains to prove that the quasiregularity condition does not imply the Mangasarian-Fromovitz condition.

Example 3.2.20. *Let $g : \mathbb{R}^2 \to \mathbb{R}^2$, $g(x_1, x_2) = (-x_1^2 + x_2, -x_1^2 - x_2)$ and the feasible point $\overline{x} = (0, 0)$. Then, $D(\overline{x}) = \{(u_1, 0) \mid u_1 \in \mathbb{R}\}$. On the other hand, is it easy to check that $T_B(M, \overline{x}) \supset D(\overline{x})$ (whence the equality holds), but there is no $\overline{u} \in \mathbb{R}^2$ with $\nabla g(\overline{x})(\overline{u}) < 0$.*

We have shown the following implications :

Linear independence condition
$\Downarrow$
Mangasarian-Fromovitz condition
$\Downarrow$
Quasiregularity condition
$\Downarrow$
Guignard condition

and none of the converses hold.

Remark 3.2.21. *Let us notice that, in particular, Theorem 3.2.19 shows as well that if $h : \mathbb{R}^p \to \mathbb{R}$ is a C^1 function, and $\overline{x} \in \mathbb{R}^p$ has the property that $\nabla h(\overline{x}) \neq 0$, then the Bouligand tangent cone to the level curve $\{x \in \mathbb{R}^p \mid h(x) = h(\overline{x})\}$ at $\overline{x}$ is the hyperplane $\{u \in \mathbb{R}^p \mid \nabla h(\overline{x})(u) = 0\}$ (or $\operatorname{Ker} \nabla h(\overline{x})$). Therefore, $\nabla h(\overline{x})$ is a normal vector to this hyperplane. We recall here that the affine subspace (of $\mathbb{R}^{p+1}$) tangent to the graph of h at $(\overline{x}, h(\overline{x}))$ has the equation*

$$y = h(\overline{x}) + \nabla h(\overline{x})(x - \overline{x}),$$

and a normal vector to it is $(\nabla h(\overline{x}), -1)$.

Therefore, the Mangasarian-Fromovitz condition ensures that the set $M(\overline{x})$ is nonempty at $\overline{x}$, which is local minimum of the problem (P). Moreover, we will now show that this condition implies special properties of the set of Lagrange multipliers.

Proposition 3.2.22. *If the Mangasarian-Fromovitz condition at $\overline{x}$ holds, then $M(\overline{x})$ is convex and compact (in $\mathbb{R}^{n+m}$).*

Proof According to the definition of $M(\overline{x})$, an element $(\lambda, \mu) \in M(\overline{x})$ satisfies

$$\nabla f(\overline{x}) + \sum_{i=1}^{n} \lambda_i \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x}) = 0$$

and

$$\lambda_i \geq 0, \ \lambda_i g_i(\overline{x}) = 0, \ \text{for every } i \in \overline{1, n}.$$

Therefore, checking the convexity and the closedness of $M(\overline{x})$ is straightforward. We will now show that $M(\overline{x})$ is bounded. Let, from the Mangasarian-Fromovitz condition,

$u \in \mathbb{R}^p$ such that

$$\nabla h(\overline{x})(u) = 0 \text{ and } \nabla g_i(\overline{x})(u) < 0, \forall i \in A(\overline{x}).$$

Then, for every $(\lambda, \mu) \in M(\overline{x})$,

$$\nabla f(\overline{x})(u) + \sum_{i \in A(\overline{x})} \lambda_i \nabla g_i(\overline{x})(u) + \sum_{j=1}^{m} \mu_j \nabla h_j(\overline{x})(u) = 0,$$

whence

$$\sum_{i \in A(\overline{x})} \lambda_i (-\nabla g_i(\overline{x})(u)) = \nabla f(\overline{x})(u),$$

from where we deduce

$$\sum_{i \in A(\overline{x})} \lambda_i \min_{i \in A(\overline{x})} \left(-\nabla g_i(\overline{x})(u)\right) \leq \nabla f(\overline{x})(u),$$

so

$$\sum_{i \in A(\overline{x})} \lambda_i \leq \frac{\nabla f(\overline{x})(u)}{\min_{i \in A(\overline{x})} \left(-\nabla g_i(\overline{x})(u)\right)}.$$

Since the right-hand side is constant, we deduce that the set of multipliers associated to inequalities constraints is bounded. Suppose, by contradiction, that there exists a sequence $(\mu_k)_{k \in \mathbb{N}} \subset \mathbb{R}^m$ unbounded (without loss of generality, we can suppose that $\|\mu_k\| \to \infty$) and a sequence $(\lambda_k)_{k \in \mathbb{N}} \subset \mathbb{R}^n_+$ such that $(\lambda_k, \mu_k) \in M(\overline{x})$. Then, for every $k \in \mathbb{N}$,

$$\nabla f(\overline{x}) + \sum_{i \in A(\overline{x})} (\lambda_i)_k \nabla g_i(\overline{x}) + \sum_{j=1}^{m} (\mu_j)_k \nabla h_j(\overline{x}) = 0.$$

We divide by $\|\mu_k\|$ and we infer that

$$\|\mu_k\|^{-1} \nabla f(\overline{x}) + \sum_{i \in A(\overline{x})} \|\mu_k\|^{-1} (\lambda_i)_k \nabla g_i(\overline{x}) + \sum_{j=1}^{m} \|\mu_k\|^{-1} (\mu_j)_k \nabla h_j(\overline{x}) = 0. \tag{3.2.6}$$

From the previous step of the proof we have

$$\sum_{i \in A(\overline{x})} \|\mu_k\|^{-1} (\lambda_i)_k \nabla g_i(\overline{x}) \overset{k \to \infty}{\to} 0$$

and,

$$\|\mu_k\|^{-1} \nabla f(\overline{x}) \overset{k \to \infty}{\to} 0.$$

On the other hand, the sequence $(\|\mu_k\|^{-1} \mu_k)$ is bounded (in $\mathbb{R}^m$), whence, without relabeling, we can suppose that $(\|\mu_k\|^{-1} \mu_k)$ is convergent towards a limit denoted by $\overline{\mu} \in \mathbb{R}^m \setminus \{0\}$. Passing to the limit in (3.2.6), we get

$$\sum_{j=1}^{m} \overline{\mu}_j \nabla h_j(\overline{x}) = 0.$$

Since $\overline{\mu} \neq 0$, this is in contradiction to the linear independence assumed in the Mangasarian-Fromovitz condition. Then $M(\overline{x})$ is bounded. □

Let us discuss now two special cases of the problem data.

Firstly, we consider the situation where the inequality restrictions are convex functions, while the equality restriction is affine. The Slater condition takes place if h is affine, g_i, $i \in \overline{1,n}$ are convex and there exists $u \in \mathbb{R}^p$ such that $h(u) = 0$ and $g(u) < 0$. This condition was introduced in 1950 by the American mathematician Morton Slater.

Theorem 3.2.23. *The Slater condition implies $T(M, x) = D(x)$ for every $x \in M$ whence, in particular, is a qualification condition.*

Proof Let $\overline{x} \in M$. The inclusion $T(M, \overline{x}) \subset D(\overline{x})$ is always true. Let $v \in D(\overline{x})$. By the Slater condition (using the convexity of g_i) we deduce (by virtue of Theorem 2.2.10) that

$$0 > g_i(u) \geq g_i(\overline{x}) + \nabla g_i(\overline{x})(u - \overline{x}),$$

whence, for $i \in A(\overline{x})$, $\nabla g_i(\overline{x})(u - \overline{x}) < 0$. We denote $w := u - \overline{x}$, and for $\lambda \in (0, 1)$, we define

$$w_\lambda := (1 - \lambda)v + \lambda w.$$

We show that $w_\lambda \in T(M, \overline{x})$ for every $\lambda \in (0, 1)$. For $i \in A(\overline{x})$,

$$\nabla g_i(\overline{x})(v) \leq 0, \ \nabla g_i(\overline{x})(w) < 0,$$

hence $\nabla g_i(\overline{x})(w_\lambda) < 0$. By Taylor's Formula, there exists $t > 0$ such that $g_i(\overline{x} + tw_\lambda) < g_i(\overline{x}) = 0$ for every $i \in A(\overline{x})$. Let $(t_k) \subset (0, \infty)$, $t_k \to 0$. Then

$$x_k := (1 - t_k)\overline{x} + t_k(\overline{x} + tw_\lambda) = \overline{x} + t_k t w_\lambda \overset{k\to\infty}{\to} \overline{x}.$$

In order for the conclusion to follow, we need to show that for k sufficiently large all (x_k) are in M. As usual, for $i \notin A(\overline{x})$, the continuity of g ensures this, while for $i \in A(\overline{x})$, we have

$$g_i(x_k) \leq (1 - t_k)g_i(\overline{x}) + t_k g_i(\overline{x} + tw_\lambda) < 0.$$

Since h is affine and $h(\overline{x}) = 0$, we get

$$h(x_k) = h(\overline{x} + t_k t w_\lambda) = t_k t \nabla h(\overline{x})(w_\lambda).$$

But $v \in D(\overline{x})$, $\nabla h(\overline{x})(v) = 0$, so

$$\nabla h(\overline{x})(w_\lambda) = \lambda \nabla h(\overline{x})(w) = \lambda \nabla h(\overline{x})(u - \overline{x}) = \lambda h(u) = 0.$$

Therefore, $h(x_k) = 0$ for any k, so, finally, $(x_k)_{k \geq k_0} \subset M$, and this means that $w_\lambda \in T(M, \overline{x})$. We let now $\lambda \to 0$; since $T(M, \overline{x})$ is closed, we get $v \in T(M, \overline{x})$, and the proof is complete. □

We consider now the case of affine restrictions. Take a matrix A of dimensions $n\times p$, a matrix B of dimensions $m\times p$ and $b\in\mathbb{R}^n$, $c\in\mathbb{R}^m$. Therefore the set M become $M=\{x\in\mathbb{R}^p \mid Ax^t\leq b^t,\ Bx^t=c^t\}$, where the relationship "$\leq$" is understood in the componentwise sense. Hence $g(x)=(Ax^t-b^t)^t$, $h(x)=(Bx^t-c^t)^t$.

Theorem 3.2.24. *In the above conditions and notation, the quasiregularity condition is automatically fulfilled.*

Proof As before, it is enough to prove that $D(\overline{x})\subset T(M,\overline{x})$. Without loss of generality, one can suppose that $A(\overline{x})=\overline{1,n}$. Let $v\in D(\overline{x})$. Then $Av^t\leq 0$, $Bv^t=0$. If $v=0$, there is nothing to prove. Otherwise, we define

$$x_k:=\overline{x}+\frac{1}{k}v,\ \forall k\in\mathbb{N}^*.$$

The relations

$$Ax_k^t\leq b^t,\ Bx_k^t=c^t,\ x_k\to\overline{x}$$

show that $v\in T(M,\overline{x})$. □

For affine restrictions, every minimum point of (P) satisfies the conclusions of Theorem 3.2.6.

3.3 Second-order Conditions

In this section we obtain second-order optimality conditions for the optimization problem with functional constraints and to this end, we assume that the data are C^2 functions. Let $\overline{x}\in M$ and $(\overline{\lambda},\overline{\mu})\in\mathbb{R}^{n+m}$ be a vector which satisfies Karush-Kuhn-Tucker conditions (i.e., the conclusion of Theorem 3.2.6). We define the set of critical directions $C(\overline{x},(\overline{\lambda},\overline{\mu}))$ as the set of vectors $u\in\mathbb{R}^p$ for which

$$\begin{cases}\nabla g_i(\overline{x})(u)=0, \text{ if } i\in A(\overline{x}) \text{ and } \overline{\lambda}_i>0,\\ \nabla g_i(\overline{x})(u)\leq 0, \text{ if } i\in A(\overline{x}) \text{ and } \overline{\lambda}_i=0,\\ \nabla h_j(\overline{x})(u)=0, \text{ for every } j\in\overline{1,m}.\end{cases}$$

Clearly, $C(\overline{x},(\overline{\lambda},\overline{\mu}))$ is a cone.

Remark 3.3.1. *Obviously, $C(\overline{x},(\overline{\lambda},\overline{\mu}))\subset D(\overline{x})$. In particular, under quasiregularity qualification condition, i.e., $T_B(M,\overline{x})=D(\overline{x})$, one has the inclusion $C(\overline{x},(\overline{\lambda},\overline{\mu}))\subset T_B(M,\overline{x})$. Moreover, if one has only equalities constraints, then one has the equality, since $\overline{\lambda}$ does not intervene in such a case.*

Theorem 3.3.2. *Let $\overline{x}\in M$ be a solution of the problem (P) and $(\overline{\lambda},\overline{\mu})\in\mathbb{R}^{n+m}$ a vector which satisfies Karush-Kuhn-Tucker conditions. If the linear independence condition*

holds at $\overline{x}$, then

$$\nabla^2_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(u, u) \geq 0$$

for every $u \in C(\overline{x}, (\overline{\lambda}, \overline{\mu}))$.

Proof Without loss of generality, we suppose that all the inequality constraints are active. We split the proof into several steps.

At the first step, we repeat, with some modifications, several arguments from the proof of Theorem 3.2.19 in order to get a sequence of feasible points with special properties. Let $d \in D(\overline{x})$, and let P be the operator defined by the matrix (of dimensions $(n+m)\times p$) with the lines consisting of vectors $\nabla g_i(\overline{x})$, $i \in \overline{1,n}$, $\nabla h_j(\overline{x})$, $j \in \overline{1,m}$ in $\mathbb{R}^p$. These vectors are linearly independent and form a basis in the linear subspace $\mathrm{Im}(P)$. Let us denote by Z the matrix (of dimensions $(p-(n+m))\times p$) whose lines are some vectors that form a basis in $\mathrm{Ker}(P)$. The the square matrix $\begin{pmatrix} P \\ Z \end{pmatrix}$ is nonsingular. We define $\varphi : \mathbb{R}^{p+1} \to \mathbb{R}^p$ by

$$\varphi(x, \tau) = ((g(x), h(x)) - \tau(Pd^t)^t, (Z(x - \overline{x} - \tau d)^t)^t).$$

Then $\nabla_x \varphi(\overline{x}, 0) = \begin{pmatrix} P \\ Z \end{pmatrix}$ is nonsingular and, from Implicit Function Theorem (i.e., Theorem 1.3.5), there exists $\varepsilon > 0$ and a differentiable function $y : (-\varepsilon, \varepsilon) \to \mathbb{R}^p$ such that

$$\varphi\left(y(\tau), \tau\right) = 0,$$

for every $\tau \in (-\varepsilon, \varepsilon)$. Moreover, for every $\tau \in (-\varepsilon, \varepsilon)$ and every x close enough to $\overline{x}$,

$$\varphi(x, \tau) = 0 \Rightarrow x = y(\tau).$$

Let $(t_k) \subset (0, \infty)$, $(t_k) \to 0$. Then, using the fact that $\varphi(y(t_k), t_k) = 0$, there exists, for every k large enough, $z_k = y(t_k)$ such that

$$\begin{aligned} g_i(z_k) &= t_k \nabla g_i(\overline{x})(d) \leq 0, \ \forall i \in \overline{1,n} \\ h_j(z_k) &= t_k \nabla h_j(\overline{x})(d) = 0, \ \forall j \in \overline{1,m}. \end{aligned} \tag{3.3.1}$$

Therefore, $(z_k) \subset M$ and the sequence $\left(t_k^{-1}(z_k - \overline{x})\right)$ is convergent. We show that $t_k^{-1}(z_k - \overline{x}) \to d$, which would imply $d \in T_B(M, \overline{x})$. According to the Taylor Theorem, from $\varphi(z_k, t_k) = 0$, there exists $(\mu_k) \subset \mathbb{R}^{n+m}$, $\mu_k \to 0$ such that

$$0 = \left((P(z_k - \overline{x} - t_k d)^t)^t, (Z(z_k - \overline{x} - t_k d)^t)^t\right) + \|z_k - \overline{x}\| \mu_k,$$

whence

$$\left(\frac{z_k - \overline{x}}{t_k} - d\right)^t = \begin{pmatrix} P \\ Z \end{pmatrix}^{-1} (-t_k^{-1} \|z_k - \overline{x}\| \mu_k),$$

from where, after passing to the limit, one gets the announced relation.

Let now $u \in C(\overline{x}, (\overline{\lambda}, \overline{\mu})) \subset D(\overline{x})$. We use now the above construction of the sequence $(z_k) \to \overline{x}$ corresponding to u. We have

$$L(z_k, (\overline{\lambda}, \overline{\mu})) = f(z_k) + \sum_{i=1}^{n} \overline{\lambda}_i g_i(z_k) + \sum_{j=1}^{m} \overline{\mu}_j h_j(z_k) = f(z_k) - t_k \sum_{i \in A(\overline{x})} \overline{\lambda}_i \nabla g_i(\overline{x})(u) = f(z_k).$$

From Taylor second-order condition, there exists $(y_k) \to 0$ such that for every k,

$$\begin{aligned} L(z_k, (\overline{\lambda}, \overline{\mu})) = L(\overline{x}, (\overline{\lambda}, \overline{\mu})) + \nabla_x L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}) \\ + \frac{1}{2} \nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) + y_k \|z_k - \overline{x}\|^2 . \end{aligned}$$

But, from the Karush-Kuhn-Tucker conditions, $L(\overline{x}, (\overline{\lambda}, \overline{\mu})) = f(\overline{x})$ and $\nabla_x L(\overline{x}, (\overline{\lambda}, \overline{\mu})) = 0$, whence

$$f(z_k) = f(\overline{x}) + \frac{1}{2} \nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) + y_k \|z_k - \overline{x}\|^2 .$$

Since $z_k \to \overline{x}$ and $\overline{x}$ is a solution for the problem (P), we obtain $f(z_k) - f(\overline{x}) \geq 0$ for every sufficiently large k. Then

$$\frac{1}{2} \nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) + y_k \|z_k - \overline{x}\|^2 \geq 0.$$

We divide by t_k^2 and we pass to the limit in order to get

$$\nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(u, u) \geq 0.$$

The proof is complete. □

We formulate now a converse of the previous result. As shown before, the sufficient optimality condition returns a stronger type of solution (i.e., strict solution).

Theorem 3.3.3. *Let $\overline{x} \in M$ and $(\overline{\lambda}, \overline{\mu}) \in \mathbb{R}^{n+m}$ a vector which satisfies Karush-Kuhn-Tucker conditions. Suppose that*

$$\nabla^2_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(u, u) > 0$$

for every $u \in C(\overline{x}, (\overline{\lambda}, \overline{\mu})) \setminus \{0\}$. Then $\overline{x}$ is a local strict solution of second order of (P).

Proof Since the set $C(\overline{x}, (\overline{\lambda}, \overline{\mu})) \cap \{u \in \mathbb{R}^p \mid \|u\| = 1\}$ is compact, and $C(\overline{x}, (\overline{\lambda}, \overline{\mu}))$ is a cone, the relation $\nabla^2_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(u, u) > 0$ for every $u \in C(\overline{x}, (\overline{\lambda}, \overline{\mu})) \setminus \{0\}$ is equivalent to the existence of a strictly positive number ρ with the property

$$\nabla^2_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(u, u) \geq \rho \|u\|^2 , \forall u \in C(\overline{x}, (\overline{\lambda}, \overline{\mu})).$$

Suppose that there exists $(z_k) \to \overline{x}$, $(z_k) \subset M$ such that

$$f(z_k) < f(\overline{x}) + k^{-1} \|z_k - \overline{x}\|^2 ,$$

for every k sufficiently large. Then, without loss of generality, we suppose that $\|z_k - \overline{x}\|^{-1}(z_k - \overline{x}) \to d \in T_B(M, \overline{x}) \setminus \{0\} \subset D(\overline{x})$. On the other hand,

$$L(z_k, (\overline{\lambda}, \overline{\mu})) = f(z_k) + \sum_{i \in A(\overline{x})} \overline{\lambda}_i g_i(z_k) \leq f(z_k),$$

and, as before, there exists $(y_k) \to 0$ such that for every k,

$$L(z_k, (\overline{\lambda}, \overline{\mu})) = f(\overline{x}) + \frac{1}{2}\nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) + y_k \|z_k - \overline{x}\|^2 . \tag{3.3.2}$$

Suppose that $d \notin C(\overline{x}, (\overline{\lambda}, \overline{\mu}))$. Then there exists $i_0 \in A(\overline{x})$ with $\overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x})(d) < 0$. For the other indices $i \in A(\overline{x})$ we have $\overline{\lambda}_i \nabla g_i(\overline{x})(d) \leq 0$. Then, it exists $(\tau_k) \to 0$ such that for every k,

$$\begin{aligned}\overline{\lambda}_{i_0} g_{i_0}(z_k) &= \overline{\lambda}_{i_0} g_{i_0}(\overline{x}) + \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x})(z_k - \overline{x}) + \tau_k \overline{\lambda}_{i_0} \|z_k - \overline{x}\| \\ &= \|z_k - \overline{x}\| \, \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x}) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) + \tau_k \overline{\lambda}_{i_0} \|z_k - \overline{x}\| .\end{aligned}$$

Hence

$$\begin{aligned}L(z_k, (\overline{\lambda}, \overline{\mu})) &= f(z_k) + \sum_{i \in A(\overline{x})} \overline{\lambda}_i g_i(z_k) \leq f(z_k) + \overline{\lambda}_{i_0} g_{i_0}(z_k) \\ &= f(z_k) + \|z_k - \overline{x}\| \, \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x}) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) + \tau_k \overline{\lambda}_{i_0} \|z_k - \overline{x}\| .\end{aligned}$$

From (3.3.2), we get

$$\begin{aligned}&f(\overline{x}) + \frac{1}{2}\nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) + y_k \|z_k - \overline{x}\|^2 \\ &\leq f(z_k) + \|z_k - \overline{x}\| \, \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x}) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) + \tau_k \overline{\lambda}_{i_0} \|z_k - \overline{x}\| .\end{aligned}$$

Furthermore,

$$\begin{aligned}&\lim_k \|z_k - \overline{x}\|^{-1} \nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu}))(z_k - \overline{x}, z_k - \overline{x}) \\ &= \lim_k \|z_k - \overline{x}\| \, \nabla_{xx} L(\overline{x}, (\overline{\lambda}, \overline{\mu})) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|}, \frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) = 0.\end{aligned}$$

After relabeling, one can see that there exists $(\nu_k) \to 0$ such that

$$f(z_k) \geq f(\overline{x}) - \|z_k - \overline{x}\| \, \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x}) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) + \nu_k \|z_k - \overline{x}\| .$$

From the assumption made, $f(z_k) < f(\overline{x}) + k^{-1} \|z_k - \overline{x}\|^2$, whence

$$f(\overline{x}) + k^{-1} \|z_k - \overline{x}\|^2 \geq f(\overline{x}) - \|z_k - \overline{x}\| \, \overline{\lambda}_{i_0} \nabla g_{i_0}(\overline{x}) \left(\frac{z_k - \overline{x}}{\|z_k - \overline{x}\|} \right) + \nu_k \|z_k - \overline{x}\| ,$$

that is

$$k^{-1}\|z_k-\overline{x}\| \geq -\overline{\lambda}_{i_0}\nabla g_{i_0}(\overline{x})\left(\frac{z_k-\overline{x}}{\|z_k-\overline{x}\|}\right)+v_k.$$

Passing to the limit, we arrive at a contradiction to the relation $\overline{\lambda}_{i_0}\nabla g_{i_0}(\overline{x})(d)<0$. Consequently, $d\in C(\overline{x},(\overline{\lambda},\overline{\mu}))\setminus\{0\}$, whence $\nabla^2_{xx}L(\overline{x},(\overline{\lambda},\overline{\mu}))(d,d)\geq\rho$. Since $L(z_k,(\overline{\lambda},\overline{\mu}))\leq f(z_k)$, coming back to (3.3.2), we can write

$$f(z_k)\geq f(\overline{x})+\frac{1}{2}\nabla_{xx}L(\overline{x},(\overline{\lambda},\overline{\mu}))(z_k-\overline{x},z_k-\overline{x})+y_k\|z_k-\overline{x}\|^2.$$

But $\nabla^2_{xx}L(\overline{x},(\overline{\lambda},\overline{\mu}))$ is continuous, whence for k large enough,

$$\nabla_{xx}L(\overline{x},(\overline{\lambda},\overline{\mu}))(z_k-\overline{x},z_k-\overline{x})>2^{-1}\rho\|z_k-\overline{x}\|^2.$$

Finally,

$$f(\overline{x})+k^{-1}\|z_k-\overline{x}\|^2\geq f(\overline{x})+\frac{\rho}{4}\|z_k-\overline{x}\|^2+y_k\|z_k-\overline{x}\|^2,$$

and a new contradiction occurs. □

3.4 Motivations for Scientific Computations

In this section, we examine the computational limits of the theoretical results from the previous sections. In some cases, it is not possible to get the exact solution of the optimization problems. The theory leads us to solve some nonlinear (systems of) equations, which do not admit analytical expressions for the solutions. This motivates us to subsequently study numerical algorithms for solving such equations, and this will be done in Chapter 6.

1. (Least squares method) We discuss now a special case of an optimization problem without restrictions. This problem belongs to the general approach called the method of least squares which is designed for the interpretation of numerical data issued from experiments in physics, biology, astronomy, chemistry. From historical perspective, this kind of problem arose from the study of the movements of planets and in questions linked to navigation techniques. The mathematician who founded this method is considered to be Gauss, but the method was published for the first time by Legendre.

In few words, the method of least squares refers to the following situation: we dispose of a data set $v_1, v_2, \ldots, v_N$ obtained after some measurements made at the (different) moments $t_1, t_2, \ldots, t_N$. The objective is to determine the best model function of the form $t\mapsto\varphi(t,x)$ (where $x=(x_1,x_2,\ldots,x_k)$ are parameters to optimize) which fits with the measurements. Therefore, for every $i\in\overline{1,N}$, one defines the residual at the moment t_i as the absolute difference between the measurement v_i and the value of the model at same time:

$$r_i:=\left|v_i-\varphi(t_i,x)\right|,$$

and now the problem is to minimize the function

$$f(x) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} \left[v_i - \varphi(t_i, x)\right]^2 .$$

It should be said that another possible objective function (even more natural to be considered) would be

$$\sum_{i=1}^{N} \left|v_i - \varphi(t_i, x)\right|$$

but this construction does not preserve the differentiability. For this reason, one prefers the sum of the squares of the residuals, whence the name of the method.

Let us now consider the simplest case of a linear dependence. Let us suppose that one has made N measurements at the different moments of time $t_1, t_2, \ldots, t_N > 0$ and, correspondingly, one has the values $v_1, v_2, \ldots, v_N$. We know that the dependence between these two sets of data is linear, and we are interested in obtaining a line which better fits the collection of observations. Let be a line $v = at + b$. As above, the residual at the moment t_i is $\left|v_i - (at_i + b)\right|$ and in order to "measure" the sum of these residuals, we consider the function $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(a, b) = \sum_{i=1}^{N} \left[v_i - (at_i + b)\right]^2 .$$

The line with respect to which this sum of residuals will be the smallest, will be that which we seek. Then, we arrive at the problem of minimization (without restrictions) of the function f. We compute the partial derivatives of f:

$$\frac{\partial f}{\partial a}(a, b) = \sum_{i=1}^{N} 2(-t_i)\left[v_i - (at_i + b)\right]$$
$$\frac{\partial f}{\partial b}(a, b) = \sum_{i=1}^{N} -2\left[v_i - (at_i + b)\right],$$

and the calculus of critical points is reduced to computation of the solutions of the system:

$$\begin{cases} \left(\sum_{i=1}^{N} t_i^2\right) a + \left(\sum_{i=1}^{N} t_i\right) b = \sum_{i=1}^{N} t_i v_i \\ \left(\sum_{i=1}^{N} t_i\right) a + Nb = \sum_{i=1}^{N} v_i. \end{cases}$$

The determinant of this system is

$$\Delta := N\left(\sum_{i=1}^{N} t_i^2\right) - \left(\sum_{i=1}^{N} t_i\right)^2 = \left(\sum_{i=1}^{N} 1^2\right)\left(\sum_{i=1}^{N} t_i^2\right) - \left(\sum_{i=1}^{N} t_i\right)^2 .$$

From the Hölder inequality, this number is positive (equality would be possible only if all the values t_i are equal, but this is not possible). Then the system admits a unique solution:

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N t_i^2 & \sum_{i=1}^N t_i \\ \sum_{i=1}^N t_i & N \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^N t_i v_i \\ \sum_{i=1}^N v_i \end{pmatrix}.$$

Let us observe that, furthermore, the function is coercive since $\lim_{\|(a,b)\|\to\infty} f(a,b) = \infty$, hence, according to Theorem 3.1.7, it admits a minimum point which is necessarily the critical point determined above. Therefore this pair (a, b) is the solution of the problem.

Another remark is that an important part of the above calculations can be repeated with obvious changes if one supposes a dependence of the type $v = a \cdot p(t) + b \cdot q(t)$, where $p, q : \mathbb{R} \to \mathbb{R}$. One obtains the system

$$\begin{pmatrix} \sum_{i=1}^N p^2(t_i) & \sum_{i=1}^N p(t_i)q(t_i) \\ \sum_{i=1}^N p(t_i)q(t_i) & \sum_{i=1}^N q^2(t_i) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N p(t_i)v_i \\ \sum_{i=1}^N q(t_i)v_i \end{pmatrix}.$$

Again, the Hölder inequality ensures that the associated matrix is invertible if and only if $(p^2(t_i))_{i=\overline{1,N}}$ and $(q^2(t_i))_{i=\overline{1,N}}$ are not proportional.

In general, for more complicated models (nonlinear in x) the method of least squares does not have an easily computable solution and this will be an impetus for us to study several algorithm in order to get good approximation of solution in fast computational time.

2. (The projection on a closed convex set) Let $a_1, a_2, \ldots, a_p \in (0, \infty)$ and the set (generalized ellipsoid)

$$M := \left\{ x \in \mathbb{R}^p \mid \sum_{i=1}^p \left(\frac{x_i}{a_i}\right)^2 \le 1 \right\}.$$

Obviously, this is a convex and compact set. Let $v \notin M$. From Theorem 2.1.5, there exists $\overline{v} \in M$, the projection of v on M. Again, we want to find an expression of this element.

As before, $\overline{v}$ is the solution of the minimization problem of $f(x) = \|x - v\|^2$ under the restriction $x \in M$. If it would exist a solution $\overline{x} \in \operatorname{int} M$, then $\nabla f(\overline{x}) = 0$, whence $\overline{x} - v = 0$, that is $v \in M$, which is false. Therefore, the restriction is active in $\overline{v}$, that is $\sum_{i=1}^p \left(\frac{\overline{v}_i}{a_i}\right)^2 = 1$. Moreover, the function which defines (with inequality) the constraint $x \in M$, i.e., $g : \mathbb{R}^p \to \mathbb{R}$,

$$g(x) := \sum_{i=1}^p \left(\frac{x_i}{a_i}\right)^2 - 1,$$

is convex, and the Slater condition holds. Moreover, f is convex as well, so we can conclude that $\overline{v}$ is a solution of the problem if and only of there exists $\lambda \ge 0$ such that

for all indexes i,

$$\overline{v}_i - v_i + \lambda \frac{\overline{v}_i}{a_i^2} = 0.$$

Since $v \neq \overline{v}$, $\lambda > 0$ and

$$\overline{v}_i = \frac{a_i^2 v_i}{a_i^2 + \lambda}.$$

On the other hand, by $\sum_{i=1}^{p} \left(\frac{\overline{v}_i}{a_i}\right)^2 = 1$,

$$\sum_{i=1}^{p} \frac{a_i^2 v_i^2}{\left(a_i^2 + \lambda\right)^2} = 1,$$

so finding λ (and then $\overline{v}$) requires solving the above equation. Let us remark that the equation has a unique solution, since the mapping

$$0 \leq \lambda \mapsto \sum_{i=1}^{p} \frac{a_i^2 v_i^2}{\left(a_i^2 + \lambda\right)^2}$$

is strictly decreasing, its value at 0 is strictly greater than 1 (notice that $v \notin M$), while its limit at $+\infty$ is 0. So, to get $\overline{v}$ one must to solve an algebraic equation of degree $2p$, and, in general, this is impossible. We will be interested in approximation methods of the solutions of nonlinear equations, and this will be one of the subjects of Chapter 6.