Chapter 9

Optimization and multipliers

> *I think the restriction to smooth manifolds is dictated only by technical reasons and is unsatisfactory...non-smooth manifolds exist naturally.*
> Shiing-Shen Chern
>
> *Consider God's handiwork*: *who can straighten what He hath made crooked?*
> Ecclesiastes 7:13

The abstract optimization problem $\min_A f$ consists of minimizing a *cost function* $f(x)$ over the points x belonging to the *admissible set A*. The set A incorporates the constraints imposed upon the points that are allowed to compete in the minimization. The nature of A, and also of the function f, determine whether our problem is classical or modern, discrete or continuous, finite or infinite dimensional, smooth or convex.

Optimization is a rich and varied subject with numerous applications. The core mathematical issues, however, are always the same:

- Existence: Is there, in fact, a solution of the problem? (This means a point $x_* \in A$ at which $\min_A f$ is attained.)

- Necessary conditions: What special properties must a solution have, properties that will help us to identify it?

- Sufficient conditions: Having identified a point that is suspected of being a solution, what tools can we apply to confirm the suspicion?

Many other issues than these can, and do arise, depending on the nature of the problem. Consider for example the calculus of variations, which we take up later on, in which the variable x refers to a function. The regularity of the minimizing function x_* reveals itself to be a central question, one that we shall study rather thoroughly. In contrast, issues such as modeling, computation, and implementation, which are crucial to applied optimization of all sorts, are not on our agenda.

Deductive and inductive methods. A familiar optimization problem that the reader has encountered is that of minimizing a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ over all points x belonging to $A = \mathbb{R}^n$. This is a "free" optimization problem, since the admissible points x are subject to no explicit constraint (except, of course, that they reside in the underlying space $\mathbb{R}^n$).

The existence question might be treated by imposing a supplementary hypothesis on f, for example, a growth condition: $f(x) \to +\infty$ as $|x| \to +\infty$. Then, bearing in mind that f is continuous (since it is differentiable), it follows that a solution x_* does exist. We would like to identify it, however.

For that purpose, we turn to necessary conditions, proceeding to invoke *Fermat's rule*: a solution x_* must satisfy $\nabla f(x_*) = 0$. (It is in order to write this equation that f was taken to be differentiable, rather than merely continuous.) Thus, we search for a solution among the critical points of f.

Then, we could conclude in one of two ways. If we know that a solution exists, and if we have examined the critical points in order to find the best critical point x_* (that is, the one assigning the lowest value to the cost f), it follows logically that x_* is the solution to the minimization problem. This approach is known as the **deductive method**.[1]

There is a potential fallacy lurking here, one that is rather common in certain areas of application. It consists of applying the deductive reasoning above without knowing with certainty that a solution exists. In the absence of guaranteed existence, it is quite possible for the necessary conditions to identify a unique admissible point x, which then fails to be a solution (because there isn't one).

An alternative approach, one that does not necessitate finding all the critical points, or knowing ahead of time that a solution exists, is to find an argument tailored precisely to a given suspect x_*. Let us give three examples of how this might work. First, suppose we find it possible to rewrite $f(x)$ as follows:

$$f(x) = \big[\varphi(x) - \varphi(x_*) \big]^2 + c,$$

for some function φ and constant c. Then, evidently, x_* minimizes f.

Another strategy would be to postulate the convexity of the function f; then, the stationarity condition $\nabla f(x_*) = 0$ implies, without further argument, that x_* is a global minimum for f.

Finally, let us mention a third tactic: if f is twice continuously differentiable, the condition $\nabla^2 f(x_*) > 0$ (positive definite) together with $\nabla f(x_*) = 0$, is enough to imply that f is at least a local minimum.

Note that all three of these alternate arguments do *not* require an existence theorem. They are examples of the **inductive method**.[2] These two approaches to solving optimization problems, the deductive and the inductive, will play a role in shaping the things to come.

We turn now to the issue of necessary conditions in the presence of constraints.

---

[1] Deductive: reasoning from the general to the particular.

[2] Inductive reasoning: wherein one argues from the particular to the general.

9.1 The multiplier rule

Consider the following special case of our problem $\min_A f$:

$$\text{Minimize } f(x) \text{ subject to } h(x) = 0. \tag{P_0}$$

Here, the admissible set is given by $A = \{x \in \mathbb{R}^n : h(x) = 0\}$, and we are dealing with a *constrained optimization* problem, one in which admissibility is defined via an equality constraint. To keep things simple, we suppose for now that f and h are continuously differentiable, and that h is real-valued.

There is a famous technique for obtaining necessary conditions in this case, known as *Lagrange multipliers*. It should be part of any mathematical education, for it is a serious nominee for the most useful theorem in applied mathematics.

The method consists of seeking the solutions of the constrained problem (P_0) above among the critical points, not of f (for this would ignore the constraint), but of $f + \lambda h$, where the *multiplier* λ is a parameter whose value is not known for the moment. The resulting equation $\nabla(f + \lambda h)(x) = 0$ may appear to be a step in the wrong direction, since it involves an additional unknown λ, but this is compensated for by the constraint equation $h(x) = 0$. The idea is to solve the two equations for x and λ simultaneously, and thus identify x (and λ, for whatever that's worth).

The theorem we have alluded to is known as the *multiplier rule*. We now discuss in some detail (but in general terms) how to prove such necessary conditions for optimality, as they are known.

Terminology: Various branches of optimization employ different synonyms for a "solution" of the underlying problem. A point x_* that solves the minimization problem can be called *optimal*, or it can be referred to as a *minimizer*, or it can be said that it provides a minimum. The word "local" is used in addition, when the minimum in question is a local one in some prescribed sense.

The first approach to proving the multiplier rule is geometric. Let x_* solve (P_0), and consider, for $\varepsilon > 0$, the relation of the set

$$f_\varepsilon^{-1} := \{x \in \mathbb{R}^n : f(x) = f(x_*) - \varepsilon\}$$

to the admissible set $A = \{x \in \mathbb{R}^n : h(x) = 0\}$. Clearly, these two sets (which we imagine as surfaces in $\mathbb{R}^n$) do not intersect, for otherwise x_* cannot be a solution of (P_0). As ε decreases to 0, the surfaces f_ε^{-1} "converge" to the level set

$$\{x \in \mathbb{R}^n : f(x) = f(x_*)\},$$

which *does* have a point in common with A, namely x_*. (Have we mentioned that we are arguing in general terms?) Thus, the value $\varepsilon = 0$ corresponds to a point of first contact (or "osculation") between these surfaces. We conclude that the normal

vectors at x_* of these two surfaces are parallel, that is, multiples of one another. Since normals to level sets are generated by gradients, we deduce the existence of a scalar λ such that $\nabla f(x_*) + \lambda \nabla h(x_*) = 0$. This is precisely the multiplier rule we seek to establish.

The argument given above has the merit of explaining the geometric meaning behind the multiplier rule. It is difficult to make it rigorous, however. A more manageable classical approach is to consider the nature of the solutions $(x,r) \in \mathbb{R}^n \times \mathbb{R}$ of the equation

$$F(x,r) := \big(f(x) - f(x_*) + r, \, h(x) \big) = (0,0).$$

The reader will observe that the point $(x_*, 0)$ satisfies the equation.

If the Jacobian matrix $D_x F(x_*, 0)$ has (maximal) rank 2, then, by the implicit function theorem, the equation $F(x,r) = (0,0)$ admits a solution $x(r)$ for every r near 0, where $\lim_{r \to 0} x(r) = x_*$. But then, for $r > 0$ sufficiently small, we obtain a point $x(r)$ arbitrarily near x_* which is admissible, and for which $f(x(r)) < f(x_*)$. This contradicts even the local optimality of x_*. It follows that the rows of $D_x F(x_*, 0)$, namely the vectors $\nabla f(x_*)$ and $\nabla h(x_*)$ (modulo transpose), must be linearly dependent. If we assume that $\nabla h(x_*) \neq 0$ (as is usually done), this implies that, for some λ, we have $\nabla f(x_*) + \lambda \nabla h(x_*) = 0$. Ergo, the multiplier rule.

This classical argument is satisfyingly rigorous, but it is difficult to adapt it to different types of constraints, notably inequality constraints $g(x) \leqslant 0$, and unilateral constraints $x \in S$. Other considerations, such as replacing $\mathbb{R}^n$ by an infinite dimensional space, or allowing the underlying functions to be nondifferentiable, further complicate matters.

Let us turn, then, to an entirely different argument for proving the multiplier rule, one that we invite the reader to criticize. It is based upon considering the following *perturbed* problem (P_α):

$$\text{Minimize } f(x) \text{ subject to } h(x) = \alpha. \qquad (P_\alpha)$$

Note that the original problem (P_0) has been imbedded in a family of problems depending on the parameter α. We define $V(\alpha)$ to be the value of the minimum in the problem (P_α). Thus, by definition of V, and since x_* solves (P_0) by assumption, we have $V(0) = f(x_*)$. On the other hand, for any x, the very definition of V implies that $V(h(x)) \leqslant f(x)$. (There is a pause here while the reader checks this.) We may summarize these two observations as follows:

$$f(x) - V\big(h(x) \big) \geqslant 0 \;\; \forall x, \text{ with equality for } x = x_*.$$

By Fermat's rule, the gradient of the function in question must vanish at x_*. By the chain rule, we obtain:

$$\nabla f(x_*) - V'\big(h(x_*) \big) \nabla h(x_*) \; = \; \nabla f(x_*) - V'(0) \nabla h(x_*) = 0.$$

Behold, once again, the multiplier rule, with $\lambda = -V'(0)$. We have also gained new insight into the meaning of the multiplier λ: it measures the sensitivity of the problem with respect to perturbing the equality constraint $h = 0$ to $h = \alpha$. (This interpretation is well known in such fields as operations research, mechanics, or economics.)

Nonsmoothness. The "value function proof" that we have just presented is completely rigorous, if it so happens that V is differentiable at 0. It must be said at once, however, that value functions are notoriously nonsmooth. Note that V above is not even finite-valued, necessarily: $V(\alpha) = +\infty$ when the set $\{x : h(x) = \alpha\}$ is empty. And simple examples show that V is not necessarily differentiable, even when it is finite everywhere. This raises the issue of rescuing the proof through the use of generalized derivatives and nonsmooth calculus, subjects that we develop in subsequent chapters.

Let us mention one more approach to proving the multiplier rule, one that uses an important technique in optimization: *exact penalization*. Our interest remains focused on the problem $\min_A f$, but we consider the (free!) minimization of the function $f(x) + k d_A(x)$, where k is a positive number and, as usual, d_A denotes the distance function associated with A. Under mild hypotheses, it turns out that for k sufficiently large, the solution x_* of the constrained problem will be a local solution of this unconstrained problem. We might say that the constraint has been absorbed into the cost by penalization.

At this point, we are tempted to write Fermat's rule: $\nabla(f + k d_A)(x_*) = 0$. There is a difficulty, once more having to do with regularity, in doing so: distance functions like d_A are not differentiable. Once again, then, we require some generalized calculus in order to proceed. A further issue also arises: given that A is the set $\{h = 0\}$, how may we interpret the generalized derivative of d_A? Is it characterized by ∇h somehow, and would this lead (yet again) to the multiplier rule? We shall develop later the "nonsmooth geometry" required to answer such questions (positively).

We have explained how considerations of theory lead to nonsmoothness. In fact, there are many important problems that feature data that are nondifferentiable from the start. They arise in such areas as elasticity and mechanics, shape optimization and optimal design, operations research, and principal-agent analysis in economics. However, we begin our study with the smooth case, in a more general setting as regards the constraints that define admissibility.

The basic problem. The focus of this chapter is the following basic problem of constrained optimization:

$$\text{Minimize } f(x) \text{ subject to } g(x) \leqslant 0, \ h(x) = 0, \ x \in S \qquad \text{(P)}$$

where the functions

$$f : X \to \mathbb{R}, \quad g : X \to \mathbb{R}^m, \quad h : X \to \mathbb{R}^n,$$

together with the set S in the Banach space X, constitute the given data. The vector inequality $g(x) \leqslant 0$ means, of course, that each component $g_i(x)$ of $g(x)$ satisfies $g_i(x) \leqslant 0$ $(i = 1, 2, \ldots, m)$. An optimization problem of this type is sometimes referred to as a *program*, which is why the term *mathematical programming* is a synonym for certain kinds of optimization.

Terminology. We say that $x \in X$ is *admissible* for the problem (P) if it lies in S and satisfies both the inequality constraint $g(x) \leqslant 0$ and the equality constraint $h(x) = 0$. The requirement $x \in S$ is also referred to as the *unilateral constraint*. A *solution* x_* of (P) is an admissible point which satisfies $f(x_*) \leqslant f(x)$ for all other admissible points x, where f is the *cost function*. We also say that x_* is *optimal* for the problem, or is a *minimizer*.

9.1 Theorem. (Multiplier rule) *Let x_* be a solution of* (P) *that lies in the interior of S. Suppose that all the functions involved are continuously differentiable in a neighborhood of x_*. Then there exists $(\eta, \gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$ satisfying the **nontriviality** condition*

$$(\eta, \gamma, \lambda) \neq 0,$$

*together with the **positivity** and **complementary slackness** conditions*

$$\eta = 0 \text{ or } 1, \ \gamma \geqslant 0, \ \langle \gamma, g(x_*) \rangle = 0,$$

*and the **stationarity** condition*

$$\big\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}'(x_*) = 0.$$

Remarks on the multiplier rule. The triple (η, γ, λ) is called a **multiplier**. The theorem asserts that the existence of such a multiplier is a necessary condition for x_* to be a solution. The term *Lagrange multiplier* is often used, in honor of the person who used the concept to great effect in classical mechanics; in fact, the idea goes back to Euler (1744). The hypothesis that the functions involved are smooth, and that x_* lies in the interior of S, makes the setting of the theorem a rather classical one, though the combination of an infinite dimensional underlying space with the presence of mixed equality/inequality constraints is modern.

Because we have assumed $x_* \in \text{int } S$, the set S plays no role in the necessary conditions. In fact, S merely serves (for the present) to localize the optimization problem. Suppose, for example, that x_* is merely a *local* minimum for (P), when the unilateral constraint $x \in S$ is absent. Then, by adding the constraint $x \in S$, where S is a sufficiently small neighborhood of x_*, we transform x_* into a *global* minimum. Another possible role of S is to define a neighborhood of x_* in which certain hypotheses hold (in this case, continuous differentiability).

The reader will observe that the nontriviality condition is an essential component of the multiplier rule, since the theorem is vacuous in its absence: the triple $(0,0,0)$ satisfies all the other conclusions, for any x_*.

The complementary slackness condition $\langle \gamma, g(x_*) \rangle = 0$ is equivalent to

$$i \in \{1, 2, \ldots, m\}, \ \ g_i(x_*) < 0 \implies \gamma_i = 0.$$

This equivalence follows from the observation that, since $\gamma \geqslant 0$ and $g(x_*) \leqslant 0$, the inner product $\langle \gamma, g(x_*) \rangle$ is necessarily nonpositive, and equals zero if and only if each term $\gamma_i g_i(x_*)$ is zero. Thus, we may rephrase the complementary slackness condition as follows: if the constraint $g_i \leqslant 0$ is not *saturated* at x_* (that is, if $g_i(x_*) < 0$), then the function g_i does not appear in the necessary conditions (the corresponding γ_i is equal to 0). This makes perfect sense, for if $g_i(x_*) < 0$, then (by the continuity of g) we have $g_i(x) < 0$ for all nearby points x, so that (locally) the constraint is redundant, and can be ignored.

The case $\eta = 0$ of the multiplier rule yields necessary conditions that do not involve the cost function f. Typically, this rather pathological situation arises when the equality and inequality constraints are so "tight" that they are satisfied by just one point x_*, which is then *de facto* optimal, independently of f. The case $\eta = 0$ is referred to as the **abnormal** case. In contrast, when $\eta = 1$, we say that we are in the **normal** case.

The proof of Theorem 9.1 is postponed to § 10.4, where, using techniques of nonsmooth analysis, a more general result can be proved.

Absence of certain constraints. Either equality or inequality constraints can be absent in the problem treated by Theorem 9.1, which then holds without reference to the missing data. Consider first the case of the problem in which there are no inequality constraints. We can simply introduce a function g that is identically -1, and then apply the theorem. When we examine the resulting necessary conditions, we see that the multiplier γ corresponding to g must be 0, and therefore the conclusions can be couched entirely in terms of a nontrivial multiplier (η, λ), with no reference to g. Note that the resulting multiplier must be normal if the vectors $h_j'(x_*)$ ($j = 1, 2, \ldots, n$) are independent. This assumption, a common one, is said to correspond to the nondegeneracy of the equality constraints.

Consider next the problem having no equality constraints. Let us introduce another variable $y \in \mathbb{R}$ on which f and g have no dependence. In $X \times \mathbb{R}$, we redefine S to be $S \times \mathbb{R}$, and we impose the equality constraint $h(y) := y = 0$. We then proceed to apply Theorem 9.1 to this augmented problem. There results a multiplier (η, γ, λ); the stationarity with respect to y yields $\lambda = 0$. Then all the conclusions of the theorem hold for a nontrivial multiplier (η, γ) (with no reference to h and λ).

The meaning of the multiplier rule. Consider the problem (P) in the case when only inequality constraints are present. For any admissible x, we denote by $I(x)$ the set of indices for which the corresponding inequality constraint is active at x:

$$I(x) = \{i \in \{1, 2, \ldots, m\} : g_i(x) = 0\}.$$

Now let x_* be optimal. As a consequence of this optimality, we claim that there cannot exist $v \in X$ such that, simultaneously,

$$\langle f'(x_*), v \rangle < 0 \text{ and } \langle g_i'(x_*), v \rangle < 0 \ \forall i \in I(x_*).$$

Such a v would provide a direction in which, for a small variation, the function f, as well as each g_i for $i \in I(x_*)$, would simultaneously decrease. Thus, for all $t > 0$ sufficiently small, we would have

$$f(x_* + tv) < f(x_*), \ g_i(x_* + tv) < g_i(x_*) = 0 \ \forall i \in I(x_*).$$

But then, by further reducing t if necessary, we could arrange to have

$$g_i(x_* + tv) \leqslant 0 \text{ for } \textit{all} \text{ indices } i \in \{1, 2, \ldots, m\},$$

as well as $f(x_* + tv) < f(x_*)$. This would contradict the optimality of x_*.

The nonexistence of such a direction v is equivalent to the positive linear dependence of the set

$$\{ f'(x_*), \ g_i'(x_*) : i \in I(x_*) \},$$

as Exer. 2.40 points out. We conclude, therefore, that the necessary conditions of the multiplier rule correspond to the nonexistence of a decrease direction (in the above sense). (In fact, this is a common feature of first-order necessary conditions in various contexts.) We remark that in the presence of equality constraints, it is much harder to argue along these lines, especially in infinite dimensions.

9.2 Exercise. We wish to minimize $f(x)$ subject to the constraints $g_1(x) \leqslant 0$ and $g_2(x) \leqslant 0$, where f, g_1 and g_2 are continuously differentiable functions defined on $\mathbb{R}^3$. At four given points x_i in $\mathbb{R}^3$ ($i = 1, 2, 3, 4$), we have the following data:

	g_1	g_2	∇f	∇g_1	∇g_2
x_1	0	0	$(2, -2, 4)$	$(-2, 0, 0)$	$(0, 1, -2)$
x_2	0	-1	$(0, 1, 1)$	$(0, -1, 0)$	$(0, 0, -1)$
x_3	0	1	$(0, 0, 1)$	$(0, 0, -1)$	$(0, 0, -1)$
x_4	0	0	$(1, 1, 1)$	$(0, -1, 0)$	$(1, 0, 1)$

(a) Only one of these four points could solve the problem. Which one is it?

(b) For each point x_i that is admissible but definitely not optimal, find a direction in which a small displacement can be made so as to attain a "better" admissible point.   □

9.3 Example. We allow ourselves to hope that the reader has seen the multiplier rule applied before. However, just in case the reader's education has not included this topic, we consider now a simple 'toy problem' (to borrow a phrase from the

physicists) of the type that the author seems to recall having seen in high school (but that was long ago). Despite its simplicity, some useful insights emerge.

The problem is that of designing a soup can of maximal volume V, given the area q of tin that is available for its manufacture. It is required, for reasons of solidity, that the thickness of the base and of the top must be double that of the sides. More specifically then, we wish to find the radius x and the height y of a cylinder such that the volume $V = \pi x^2 y$ is maximal, under the constraint $2\pi xy + 4\pi x^2 = q$. We do not doubt the reader's ability to solve this problem without recourse to multipliers, but let us do so by applying Theorem 9.1.

We could view the constraint $2\pi xy + 4\pi x^2 = q$ as an equality constraint (which it is), but we can also choose to replace it by the inequality constraint

$$g(x,y) := 2\pi xy + 4\pi x^2 - q \leqslant 0,$$

since it is clear that the solution will use all the available tin. Doing so offers the advantage of knowing beforehand the sign of the multiplier that will appear in the necessary conditions.

The problem has a natural (implicit) constraint that x and y must be nonnegative, a feature of many optimization problems that motivates the following definition.

Notation.   We denote by $\mathbb{R}_+^n$ the set $\{x \in \mathbb{R}^n : x \geqslant 0\}$, also referred to as the positive orthant.

To summarize, then, we have the case of problem (P) in which $(x,y) \in \mathbb{R}^2$ and

$$f(x,y) := -\pi x^2 y, \quad g(x,y) = 2\pi xy + 4\pi x^2 - q, \quad S = \mathbb{R}_+^2$$

with the equality constraint $h = 0$ being absent. (Note the minus sign in f, reflecting the fact that our theory was developed for minimization rather than maximization.) If we take q strictly positive, it is easy to prove that a solution (x_*, y_*) of the problem exists, and that we have $x_* > 0$, $y_* > 0$ (thus, the solution lies in int S).

The usual first step in applying Theorem 9.1 is to rule out the abnormal case $\eta = 0$; we proceed to do so. If $\eta = 0$, then the necessary conditions imply that $\nabla g(x_*, y_*)$ equals $(0,0)$, which leads to $x_* = y_* = 0$, which is absurd. Thus, we may take $\eta = 1$. (Note that the abnormal case corresponds to an exceedingly tight constraint, the case $q = 0$.) With $\eta = 1$, the resulting stationarity condition becomes

$$-2\pi x_* y_* + \gamma (2\pi y_* + 8\pi x_*) = 0, \quad -\pi x_*^2 + \gamma (2\pi x_*) = 0.$$

The second equation gives $x_* = 2\gamma$, whence $\gamma > 0$; substituting in the first equation then produces $y_* = 8\gamma$. Since $\gamma > 0$, the inequality constraint is saturated (as expected). The equality $g(x_*, y_*) = 0$ then leads to

$$\gamma = \sqrt{q}/(4\sqrt{3\pi}), \quad y_* = 4x_*, \quad f(x_*, y_*) = -q^{3/2}/(6\sqrt{3\pi}).$$

Thus the height of the optimal soup can is four times its radius. As regards our own (non soup oriented) intentions, it is more to the point to note that the derivative of the optimal volume $q^{3/2}/(6\sqrt{3\pi})$ with respect to q is precisely the multiplier γ, thus confirming the interpretation of γ (suggested in the introduction) as a sensitivity with respect to changing the constraint (that is, the amount of available tin). Economists would refer to γ as a *shadow price*.[3]   □

9.2 The convex case

The next item on the agenda is to impart to the reader an appreciation of the "convex case" of the problem (P). We shall see in this section that the multiplier rule holds in a stronger form in this setting, and that it is normally sufficient as well as necessary. In the following section, another characteristic feature of convex optimization is examined: the possibility of defining a useful *dual problem*. Together, these elements explain why, other things being equal, the convex case of (P) is preferred, if we can so arrange things.

The problem (P) is unaltered: it remains that of minimizing $f(x)$ subject to the constraints

$$g(x) \leqslant 0, \ \ h(x) = 0, \ \ x \in S,$$

but in the following framework, referred to as the *convex case*:

- S is a convex subset of a real vector space X;

- The following functions are convex:

$$f : S \to \mathbb{R} \ \text{ and } \ g_i : S \to \mathbb{R} \ (i = 1, 2, \ldots, m);$$

- Each function $h_j : S \to \mathbb{R}$ $(j = 1, 2, \ldots, n)$ is affine; that is, h_j is of the form $\langle \zeta_j, x \rangle + c_j$, where ζ_j is a linear functional on S and $c_j \in \mathbb{R}$.

Note that these functions need only be defined on S. In the following counterpart to Theorem 9.1, it is *not* required that x_* lie in the interior of S; indeed, no topology is imposed on X.

9.4 Theorem. (Kuhn-Tucker) *Let x_* be a solution of* (P) *in the convex case. Then there exists* $(\eta, \gamma, \lambda) \in \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$ *satisfying the* **nontriviality** *condition*

$$(\eta, \gamma, \lambda) \neq 0,$$

---

[3] The shadow price would be used, for example, to decide whether the soup can (which is optimal for the specified volume) should be made larger (in order to increase profit). To decide, one compares $p\gamma$ (the marginal effect on revenue of using more tin, where p is the unit price of soup) to the marginal cost of tin; at optimality, the two marginal effects are equal.

the **positivity** *and* **complementary slackness** *conditions*

$$\eta = 0 \ \text{or} \ 1, \ \gamma \geqslant 0, \ \langle \gamma, g(x_*) \rangle = 0,$$

and the **minimization** *condition*

$$\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \}(x) \ \geqslant \ \{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \}(x_*) = \eta f(x_*) \ \forall x \in S.$$

Proof. We consider the following subset of $\mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^n$:

$$C = \{ (f(x) + \delta, g(x) + \Delta, h(x)) : \delta \geqslant 0, \ \Delta \geqslant 0, \ x \in S \}.$$

It is easy to see that C is convex (that's what the convexity hypotheses on the data are for). We claim that the point $(f(x_*), 0, 0)$ lies in the boundary of C.

If this were false, C would contain, for some $\varepsilon > 0$, a point of the form

$$(f(x) + \delta, g(x) + \Delta, h(x)) = (f(x_*) - \varepsilon, 0, 0), \text{ where } x \in S, \ \delta \geqslant 0, \ \Delta \geqslant 0.$$

But then x is admissible for (P) and assigns to f a strictly lower value than does x_*, contradicting the optimality of x_*.

Since C is finite dimensional, the normal cone (in the sense of convex analysis) to C at this boundary point is nontrivial (Cor. 2.48). This amounts to saying that there exists $(\eta, \gamma, \lambda) \neq 0$ such that

$$\eta(f(x) + \delta) + \langle \gamma, g(x) + \Delta \rangle + \langle \lambda, h(x) \rangle \ \geqslant \ \eta f(x_*) \ \forall x \in S, \ \delta \geqslant 0, \ \Delta \geqslant 0.$$

Note that this yields the minimization condition of the theorem. It also follows readily that $\eta \geqslant 0$ and $\gamma \geqslant 0$. Taking

$$x = x_*, \ \delta = 0, \ \Delta = 0$$

in the inequality gives $\langle \gamma, g(x_*) \rangle \geqslant 0$, which is equivalent to the complementary slackness condition $\langle \gamma, g(x_*) \rangle = 0$, since $g(x_*) \leqslant 0$ and $\gamma \geqslant 0$. Finally, if $\eta > 0$, note that we can normalize the multiplier (η, γ, λ); that is, replace it by $(1, \gamma/\eta, \lambda/\eta)$. Thus, in all cases, we can assert that η equals 0 or 1. □

Remark. We refer to the vector (η, γ, λ) as a *multiplier in the convex sense*. The difference between such a multiplier and a classical one (as given in Theorem 9.1) is that the stationarity is replaced by an actual minimization. Furthermore, no differentiability of the data is assumed here, and, as we have said, there is no requirement that x_* lie in the interior of S.

In the same vein as our discussion following Theorem 9.1, it is easy to see that the theorem above adapts to the cases in which either the equality or inequality constraint is absent, by simply deleting all reference to the missing constraint.

9.5 Example. (du Bois-Raymond lemma) The following situation arises in the calculus of variations, where the conclusion below will be of use later on. We are given an element $\theta \in L^1(a,b)$ such that

$$\int_a^b \theta(t)\,\varphi'(t)\,dt \,=\, 0 \;\; \forall \varphi \in \mathrm{Lip}_0[a,b],$$

where $\mathrm{Lip}_0[a,b]$ is the set of Lipschitz functions on $[a,b]$ that vanish at a and b. Evidently, the stated condition holds if θ is constant; our goal is to prove that this is the only case in which it holds.

Let X be the vector space of all $\varphi \in \mathrm{Lip}[a,b]$ satisfying $\varphi(a) = 0$, and define

$$f(\varphi) \,=\, \int_a^b \theta(t)\,\varphi'(t)\,dt\,, \quad h(\varphi) \,=\, \varphi(b)\,.$$

Then, by hypothesis, we have $f(\varphi) \geqslant 0$ for all $\varphi \in X$ satisfying $h(\varphi) = 0$. Thus the function $\varphi_* \equiv 0$ solves the corresponding version of the optimization problem (P). We proceed to apply Theorem 9.4. Accordingly, there exists a multiplier $(\eta,\lambda) \neq 0$ with η equal to 0 or 1, such that

$$\eta \int_a^b \theta(t)\,\varphi'(t)\,dt + \lambda\,\varphi(b) \,=\, \int_a^b \big\{\,\eta\,\theta(t) + \lambda\,\big\}\,\varphi'(t)\,dt \,\geqslant\, 0 \;\; \forall \varphi \in X\,.$$

It follows that $\eta = 0$ cannot occur, for then we would have $\lambda = 0$ too, violating nontriviality. Thus we may set $\eta = 1$, and we obtain

$$\int_a^b \big\{\,\theta(t) + \lambda\,\big\}\,\varphi'(t)\,dt \,\geqslant\, 0 \;\; \forall \varphi \in X\,.$$

For a positive integer k, let A_k be the set $\{t \in (a,b) : |\theta(t)| \leqslant k\}$, and let χ_k be its characteristic function. Taking

$$\varphi(t) \,=\, -\int_a^t \big\{\,\theta(s) + \lambda\,\big\}\,\chi_k(s)\,ds$$

in the inequality above (note that $\varphi \in X$) yields

$$-\int_{A_k} \big\{\,\theta(t) + \lambda\,\big\}^2\,dt \,\geqslant\, 0\,.$$

Thus the integral is 0 for every k, and we discover $\theta(t) + \lambda = 0$ a.e. $\square$

9.6 Exercise. Let x_* be a solution of the problem encountered in Exer. 5.53. Show that the problem fits into the framework of Theorem 9.4 if one takes

$$g(x) = \sum_{i=1}^{\infty} x_i - 1, \;\; S = \big\{ x \in \ell^r : 0 \leqslant x_i\,\forall i,\, \sum_{i=1}^{\infty} x_i < \infty,\, f(x) < \infty \big\}$$

and if no equality constraint is imposed. Deduce the existence of a nonnegative constant γ such that, for each i, the value x_{*i} minimizes the function $t \mapsto f_i(t) + \gamma t$ over $[0, \infty)$. In that sense, all the x_{*i} are determined by a single scalar γ.   □

Remark. In many cases, the minimization condition in the theorem can be expressed in equivalent terms as a stationarity condition, via the subdifferential and the normal cone of convex analysis. This brings out more clearly the common aspects of Theorem 9.1 and Theorem 9.4, as we now see.

9.7 Exercise.

(a) Let x_* be admissible for (P) in the convex case, and suppose there exists a normal multiplier $(1, \gamma, \lambda)$ associated with x_*. Prove that x_* is optimal.

(b) In addition to the hypotheses of Theorem 9.4, suppose that X is a normed space, and that f, g, and h are convex and continuous on X. Prove that the minimization condition in the conclusion of the theorem is equivalent to

$$0 \in \partial \big\{ \eta f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}(x_*) + N_S(x_*).$$

Under what additional hypotheses would this be equivalent to the stationarity conclusion in Theorem 9.1?   □

The exercise above expresses the fact that in the convex case of (P), the necessary conditions, when they hold normally, are also sufficient for optimality. Another positive feature of the convex case is the possibility of identifying reasonable conditions in the presence of which, *a priori*, the necessary conditions must hold in normal form. We are referring to the **Slater condition**, which is said to hold when:

- X is a normed space;

- There exists a *strictly admissible* point x_0 for (P):

$$x_0 \in \operatorname{int} S, \quad g(x_0) < 0, \quad h(x_0) = 0;$$

- The affine functions of the equality constraint are independent, meaning that the set $\{ h'_j : j = 1, 2, \ldots, n \}$ is independent.

9.8 Theorem. *In the convex case of problem* (P), *when the Slater condition holds, the multiplier whose existence is asserted by* Theorem 9.4 *is necessarily normal*: $\eta = 1$.

Proof. We reason *ad absurdum*, by supposing that $(0, \gamma, \lambda)$ is an abnormal (nontrivial) multiplier. The minimization condition, when expressed at the point x_0 provided by the Slater condition, gives $\langle \gamma, g(x_0) \rangle \geqslant 0$. Since every component of $g(x_0)$ is strictly negative, and since $\gamma \geqslant 0$, we deduce $\gamma = 0$. Then the minimization condition becomes: $\langle \lambda, h(x) \rangle \geqslant 0 \ \forall x \in S$. Since equality holds at $x_0 \in \operatorname{int} S$, we have

$\sum_i \lambda_i h'_i = 0$ by Fermat's rule. Then the linear independence implies $\lambda = 0$, contradicting the nontriviality of the multiplier.   □

We now illustrate the use of the Slater condition by means of a simple problem arising in statistics.

9.9 Exercise. Let $z_1, z_2, \ldots, z_n$ be the n distinct values of a random variable Z, and let p_i be the probability that $Z = z_i$. Let us suppose that we know from observation that Z has mean value m, so that $\sum_i z_i p_i = m$. However, the probabilities p_i are not known. A common way to estimate the probability distribution $p = (p_1, p_2, \ldots, p_n)$ in this case is to postulate that it maximizes the *entropy*

$$E = -\sum_{i=1}^{n} p_i \ln p_i.$$

The optimization problem, then, is to maximize E subject to the constraints

$$p \in \mathbb{R}^n_+, \ \ \sum_{i=1}^{n} p_i = 1, \ \ \sum_{i=1}^{n} z_i p_i = m.$$

We place this in the context of Theorem 9.4 by taking $X = \mathbb{R}^n$ and $S = \mathbb{R}^n_+$, and by defining

$$f(p) = \sum_i p_i \ln p_i, \ \ h_1(p) = \left(\sum_i p_i\right) - 1, \ \ h_2(p) = \left(\sum_i z_i p_i\right) - m.$$

Thus, the equality constraint has two components, and the inequality constraint is absent. Note that the function $t \mapsto t \ln t$ has a natural value of 0 at $t = 0$.

(a) Prove that f is convex on S.

(b) Prove that a solution to the problem exists, and that it is unique.

We suppose henceforth that $\min_i z_i < m < \max_i z_i$. If this were not the case, m would equal either $\min_i z_i$ or $\max_i z_i$, which means that the distribution has all its mass on a single value: a case of overly tight constraints which, of themselves, identify the solution.

(c) Prove that the Slater condition is satisfied. Deduce that the solution admits a normal multiplier in the convex sense.

(d) Deduce from the minimization condition of the multiplier that the solution p satisfies $p_i > 0 \ \forall i \in \{1, 2, \ldots, n\}$.

(e) Prove that the solution p corresponds to an *exponential* distribution: for certain constants c and k, we have

$$p_i = \exp\left(c + k z_i\right), \ \ i = 1, 2, \ldots, n.$$ □

The next result gives a precise meaning (in the current convex setting) to the interpretation of multipliers in terms of sensitivity.

9.10 Theorem. *Let there exist a solution x_* of the problem* (P) *in the convex case, where the Slater condition is satisfied. We define the value function V on $\mathbb{R}^m \times \mathbb{R}^n$ as follows*:

$$V(\alpha, \beta) = \inf \left\{ f(x) : x \in S,\ g(x) \leqslant \alpha,\ h(x) = \beta \right\}.$$

Then V is a convex function with values in $(-\infty, +\infty]$. The vector $(1, \gamma, \lambda)$ is a multiplier associated with x_* if and only if $(\gamma, \lambda) \in -\partial V(0,0)$.

Proof. According to Theorems 9.4 and 9.8, there exists a normal multiplier $(1, \gamma, \lambda)$ associated to x_*. Let $P(\alpha, \beta)$ denote the optimization problem that defines the value of $V(\alpha, \beta)$, and let x be any point admissible for $P(\alpha, \beta)$:

$$x \in S,\ g(x) \leqslant \alpha,\ h(x) = \beta.$$

Then, using $\gamma \geqslant 0$ and $-g(x) \geqslant -\alpha$, we have

$$\begin{aligned}
f(x) &= f(x) + \langle \gamma, g(x) \rangle + \langle \gamma, -g(x) \rangle + \langle \lambda, h(x) - \beta \rangle \\
&\geqslant f(x) + \langle \gamma, g(x) \rangle + \langle \gamma, -\alpha \rangle + \langle \lambda, h(x) - \beta \rangle \\
&\geqslant f(x_*) - \langle \gamma, \alpha \rangle - \langle \lambda, \beta \rangle = V(0,0) - \langle (\gamma, \lambda), (\alpha, \beta) \rangle,
\end{aligned}$$

by the minimization condition of the multiplier $(1, \gamma, \lambda)$. Taking the infimum over x, we deduce

$$V(\alpha, \beta) \geqslant V(0,0) - \langle (\gamma, \lambda), (\alpha, \beta) \rangle,$$

which confirms $V > -\infty$, and that $-(\gamma, \lambda)$ belongs to $\partial V(0,0)$, the subdifferential of V at $(0,0)$. As for the convexity of V, it follows easily from its definition (or it can be deduced from Exer. 8.10).

As the reader well knows, it is not our custom to abandon a proof in midstream. On this occasion, however, we would ask the reader to kindly supply the converse; it happens to be the subject of the exercise that follows. □

9.11 Exercise. Under the hypotheses of Theorem 9.10, prove that an element (γ, λ) belonging to $-\partial V(0,0)$ determines a normal multiplier $(1, \gamma, \lambda)$. □

9.3 Convex duality

An important feature of convex optimization is the possibility of developing a theory in which one associates to the original, or *primal*, problem another optimization problem, the *dual*, which is linked to the primal through multipliers (of some type or other). This idea has important theoretical and even numerical consequences, in such areas as game theory, optimal transport, operations research, mechanics, and

economics. We take a brief look at the topic in this section, in order to establish its connection to the multiplier rule.

We continue to be interested in the problem (P) of the preceding section, which we think of as the primal. The **dual problem** (D) associated to (P) is defined as follows:

$$\text{Maximize } \varphi(\gamma, \lambda) \text{ subject to } (\gamma, \lambda) \in \mathbb{R}_+^m \times \mathbb{R}^n \tag{D}$$

where the (concave) function $\varphi : \mathbb{R}^m \times \mathbb{R}^n \to [-\infty, +\infty)$ is defined by

$$\varphi(\gamma, \lambda) = \inf_{x \in S} \big\{ f + \langle \gamma, g \rangle + \langle \lambda, h \rangle \big\}(x).$$

The dual problem is of greatest interest when it can be rendered more explicit. Let's illustrate this now.

9.12 Example. We are given $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and a matrix M which is $m \times n$, and we consider the following instance of the problem (P):

$$\text{Minimize } \langle c, x \rangle \text{ subject to } x \in \mathbb{R}_+^n, \quad Mx \leqslant b.$$

(As usual, points in Euclidean space, in their dealings with matrices, are viewed as columns.) This is a problem in what is called *linear programming*. We are dealing, then, with the convex case of (P), in the absence of equality constraints. Let us make explicit the dual problem (D). We have

$$\varphi(\gamma) = \inf \big\{ \langle c, x \rangle + \langle \gamma, Mx - b \rangle : x \in \mathbb{R}_+^n \big\}$$

$$= \inf \big\{ \langle c + M^* \gamma, x \rangle - \langle \gamma, b \rangle : x \in \mathbb{R}_+^n \big\} = \begin{cases} -\langle \gamma, b \rangle & \text{if } c + M^* \gamma \geqslant 0 \\ -\infty & \text{otherwise.} \end{cases}$$

It turns out then, that (D) can be expressed as follows:

$$\text{Maximize } \langle -b, \gamma \rangle \text{ subject to } \gamma \in \mathbb{R}_+^m, \quad -M^* \gamma \leqslant c.$$

Thus, the dual problem has essentially the same form as the primal; this fact is exploited to great effect in the subject.   □

We now describe the link between the primal and the dual problem.

9.13 Theorem. (Lagrangian duality) *We consider the basic problem* (P) *in the convex case. We suppose that there is a solution of* (P) *which admits a normal multiplier. Then*

$$\min (P) = \max (D).$$

Furthermore, any solution x_* of (P) *admits a normal multiplier, and a vector $(1, \gamma_*, \lambda_*) \in \mathbb{R} \times \mathbb{R}_+^m \times \mathbb{R}^n$ is a multiplier for x_* if and only if (γ_*, λ_*) solves the dual problem* (D).

Proof.

A. Let $(\gamma, \lambda) \in \mathbb{R}^m_+ \times \mathbb{R}^n$. Observe that

$$\varphi(\gamma, \lambda) \leqslant \{f + \langle \gamma, g \rangle + \langle \lambda, h \rangle\}(x_*) \leqslant f(x_*) = \min(\text{P}).$$

It follows that sup (D) $\leqslant$ min (P). Now let $(1, \gamma_*, \lambda_*)$ be a multiplier for a solution x_* of (P). (By hypothesis, at least one such normal multiplier and solution exist.) The minimization condition asserts

$$\{f + \langle \gamma_*, g \rangle + \langle \lambda_*, h \rangle\}(x) \geqslant \{f + \langle \gamma_*, g \rangle + \langle \lambda_*, h \rangle\}(x_*) = f(x_*) \ \forall x \in S,$$

whence

$$\sup(\text{D}) \geqslant \varphi(\gamma_*, \lambda_*) \geqslant f(x_*) = \min(\text{P}) \geqslant \sup(\text{D}).$$

We deduce that (γ_*, λ_*) solves the dual problem, and that min (P) = max (D).

B. Now let (γ_*, λ_*) be any solution of the dual problem, and let x_* be any solution of the primal problem. Then $\gamma_* \in \mathbb{R}^m_+$, and we have

$$\sup(\text{D}) = \varphi(\gamma_*, \lambda_*) \leqslant \{f + \langle \gamma_*, g \rangle + \langle \lambda_*, h \rangle\}(x_*) \leqslant f(x_*) = \min(\text{P}) = \sup(\text{D}),$$

which implies $\langle \gamma_*, g(x_*) \rangle = 0$, the complementary slackness condition. We also have, for any $x \in S$,

$$\{f + \langle \gamma_*, g \rangle + \langle \lambda_*, h \rangle\}(x) \geqslant \varphi(\gamma_*, \lambda_*) = \sup(\text{D}) = \min(\text{P}) = f(x_*),$$

which yields the minimization condition for $(1, \gamma_*, \lambda_*)$, and confirms that this vector has all the properties of a multiplier for x_*. $\qquad\square$

9.14 Exercise. Under the hypotheses of Theorem 9.13, let x_* be a solution of (P), and let $(1, \gamma_*, \lambda_*)$ be a (normal) multiplier associated to x_*. The *Lagrangian L* of the problem is defined to be the function

$$L(x, \gamma, \lambda) = \{f + \langle \gamma, g \rangle + \langle \lambda, h \rangle\}(x).$$

Prove that $(x_*, \gamma_*, \lambda_*)$ is a *saddle point* of L, meaning that

$$L(x_*, \gamma, \lambda) \leqslant L(x_*, \gamma_*, \lambda_*) \leqslant L(x, \gamma_*, \lambda_*) \ \forall x \in S, \ \gamma \in \mathbb{R}^m_+, \ \lambda \in \mathbb{R}^n. \qquad\square$$

Remark. If x is admissible for (P), we know, of course, that min (P) $\leqslant f(x)$. Similarly, if (γ, λ) is admissible for (D), we obtain max (D) $\geqslant \varphi(\gamma, \lambda)$. But now, suppose that duality holds: min (P) = max (D). Then we deduce

$$\varphi(\gamma, \lambda) \leqslant \min(\text{P}) \leqslant f(x).$$

The generating of *bilateral* bounds of this type is of evident interest in developing numerical methods, a task to which duality has been effectively applied. Under more subtle hypotheses than those of Theorem 9.13 (in linear programming, or in

the presence of infinite-dimensional equality constraints, for example), it can be a delicate matter to establish duality.

Another salient point, as evidenced by Theorem 9.10, is the possibility of finding the multipliers for the primal problem by solving the dual problem. We illustrate this procedure now.

9.15 Exercise. We are given n continuous, convex functions $f_i : [0, \infty) \to \mathbb{R}$, and a positive parameter q. We study the following simple *allocation* problem, of a type that frequently arises in economics and operations research:

$$\text{Minimize } f(x) = \sum_{i=1}^{n} f_i(x_i) \text{ subject to } x \in S := \mathbb{R}^n_+, \quad \sum_{i=1}^{n} x_i \leqslant q. \quad \text{(P)}$$

Note that this is a convex case of (P), with no equality constraints.

The i-th cost component f_i depends only on x_i; the difficulty (especially when n is large) lies in determining what optimal $x_i \geqslant 0$ to allocate to f_i, while respecting the upper bound on the sum of the x_i.

(a) Prove that a solution x_* exists, verify the Slater condition, and deduce that x_* admits a normal multiplier $(1, \gamma_*)$.

(b) Prove that, for each index i, the value x_{*i} is a solution of the problem

$$\min \left\{ f_i(u) + \gamma_* u : u \in \mathbb{R}_+ \right\}.$$

It turns out then, that if we know γ_*, we may use it to calculate x_* one coordinate at a time, while completely ignoring the constraint $\sum_i x_i \leqslant q$. The problem is said to have been *decomposed*.[4]

How might we effectively calculate γ_*, however? For this purpose, we define the following function closely related to the conjugate of f_i:

$$\theta_i(\gamma) = \inf \left\{ \gamma u + f_i(u) : u \geqslant 0 \right\}, \quad \gamma \in \mathbb{R}.$$

(c) Show that the dual problem (D) consists of maximizing over $\mathbb{R}_+$ the following function of a single variable:

$$\varphi(\gamma) = \sum_{i=1}^{n} \theta_i(\gamma) - \gamma q.$$

Why is this problem relatively simple? How do we know a maximum exists?

We suppose henceforth, in order to permit explicit calculation, that each f_i has the form

$$f_i(u) = p_i u + u \ln u \text{ for } u > 0, \text{ with } f_i(0) = 0.$$

---

[4] Thus, the computation could be envisaged on a *decentralized* basis, where each component, having been informed of the internal unit cost γ_*, can calculate its own allocation x_{*i} by maximizing its own profit. These individually motivated calculations would lead to *global* optimality: Adam Smith's invisible hand at work.

(d) Let $\psi : \mathbb{R}_+ \to \mathbb{R}$ be defined by

$$\psi(u) = u \ln u \text{ for } u > 0, \ \psi(0) = 0,$$

and let $c \in \mathbb{R}$. Prove that the function $u \mapsto \psi(u) - cu$ attains a minimum over $\mathbb{R}_+$ at $u = e^{c-1}$, the corresponding value of ψ being $-e^{c-1}$.

e) Deduce from this the evident solution to problem (P) when q is no less than

$$\sigma := \sum_{i=1}^n e^{-p_i - 1}.$$

Prove that when $q < \sigma$, the solution γ_* of the dual problem is $\ln(\sigma/q)$. Use this to show that the optimal allocation is given by $x_{*i} = e^{-p_i - 1} q / \sigma$.

(f) Prove that the value $V(q)$ of the problem (P) is given by

$$V(q) = \begin{cases} -\sigma & \text{if } q \geqslant \sigma \\ q\left(\ln q - 1 - \ln \sigma\right) & \text{if } q \leqslant \sigma. \end{cases}$$

Show that $V'(q) = -\gamma_*$ (the expected sensitivity relation).   □