Lecture 4

Convex Programming and Lagrange Duality

(Convex Programming program, Convex Theorem on Alternative, convex duality Optimality Conditions in Convex Programming)

4.1 Convex Programming Program

In Lecture 2 we have discussed Linear Programming model which cover numerous applications. Whenever applicable, LP allows to obtain useful quantitative and qualitative information on the problem at hand. The analytic structure of LP programs gives rise to a number of general results (e.g., those of the LP Duality Theory) which provide us in many cases with valuable insight. Nevertheless, many "real situations" cannot be covered by LP models. To handle these "essentially nonlinear" cases, one needs to extend the basic theoretical results and computational techniques known for LP beyond the realm of Linear Programming. There are several equivalent ways to define a general convex optimization problem. Here we follow the "traditional" and the most direct way: when passing from a generic LP problem

$$\{\min c^T x |\ Ax \geq b\} \quad [A \in \mathbf{R}^{m\times n}] \tag{LP}$$

to its nonlinear extensions, we simply allow for some of linear functions involved – the *linear* objective function $f(x) = c^T x$ and *linear* inequality constraints $g_i(x) = a_i^T x \leq b_i,\ i = 1, ..., m$, to become *nonlinear*.

A (constrained) Mathematical Programming program is a problem as follows:

$$\text{(P)} \quad \min \left\{ f(x) \mid x \in X, \quad \begin{array}{rcl} g(x) & \equiv & (g_1(x), ..., g_m(x)) \leq 0, \\ h(x) & \equiv & (h_1(x), ..., h_k(x)) = 0 \end{array} \right\}. \tag{4.1.1}$$

(P) is called *convex* (or *Convex Programming* program), if

- X is a *convex* subset of $\mathbf{R}^n$

- $f, g_1, ..., g_m$ are *real-valued convex* functions on X,
  and
- there are no equality constraints at all.

Note that instead of saying that there are no equality constraints, we could say that there are constraints of this type, but only *linear* ones; this latter case can be immediately reduced to the one without equality constraints by replacing $\mathbf{R}^n$ with the affine set given by the (linear) equality constraints.

4.1.1 Convex Theorem on Alternative

The question we are interested for now is the same we asked (and answered with success) in Section 2.3:

Given an Convex Programming program (P) find a systematic way to bound from below its optimal value.

Let us start with the same simple observation that the fact that a point x^* is an optimal solution can be expressed in terms of solvability/unsolvability of certain systems of inequalities: these systems are

$$x \in X,\ f(x) \leq c,\ g_j(x) \leq 0,\ j = 1, ..., m \tag{4.1.2}$$

and

$$x \in X,\ f(x) < c,\ g_j(x) \leq 0,\ j = 1, ..., m; \tag{4.1.3}$$

here c is a parameter. Optimality of x^* for the problem means exactly that for appropriately chosen c (this choice, of course, is $c = f(x^*)$) the first of these systems is solvable and x^* is its solution, while the second system is unsolvable. Given this trivial observation, what we need is to convert the "negative" part of it – the claim that (4.1.3) is unsolvable – into a positive statement. This will be done by means of what is called *Convex Theorem on Alternative*.

Frow what we already know, it is easy to *guess* the result we need. The question we are interested in is, basically, how to express in an affirmative manner the fact that a system of inequalities has no solutions; to this end we observe that if we can combine, in a linear fashion, the inequalities of the system and get an obviously false inequality like $0 \leq -1$, then the system is unsolvable; this condition is certain affirmative statement with respect to the weights with which we are combining the original inequalities.

Indeed, consider *an arbitrary* inequality system of the type (4.1.3):

$$\begin{array}{rcl} f(x) & < & c \\ g_j(x) & \leq & 0,\ j = 1, ..., m \\ x & \in & X; \end{array} \tag{I}$$

all we assume is that X is a nonempty subset in $\mathbf{R}^n$ and $f, g_1, ..., g_m$ are real-valued functions on X. It is absolutely evident that

if there exist nonnegative $\lambda_1, ..., \lambda_m$ *such that the inequality*

$$f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) < c \tag{4.1.4}$$

has no solutions in X*, then* (I) *also has no solutions.*

Indeed, a solution to (I) clearly is a solution to (4.1.4) – the latter inequality is nothing but a combination of the inequalities from (I) with the weights 1 (for the first inequality) and λ_j (for the remaining ones).

Now, what does it mean that (4.1.4) has no solutions? A necessary and sufficient condition for this is that the infinum of the left hand side of (4.1.4) in $x \in X$ is $\geq c$. Thus, we come to the following evident

Proposition 4.1.1 [Sufficient condition for unsolvability of (I)] *Consider a system* (I) *with arbitrary data and assume that the system*

$$\begin{array}{rcl} \inf_{x\in X}\left[f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)\right] & \geq & c \\ \lambda_j & \geq & 0,\ j = 1, ..., m \end{array} \tag{II}$$

with unknowns $\lambda_1, ..., \lambda_m$ *has a solution. Then* (I) *is infeasible.*

Let me stress that this result is completely general; it does not require any assumptions on the entities involved.

The result we have obtained, unfortunately, does not help us very much: what we expect is not only the *sufficiency* of the condition of Proposition 4.1.1 for infeasibility of (I), but the *necessity* of this condition. Same as in the case of the Linear Programming program, justifying the necessity of the condition in question has nothing in common with the evident reasoning which gives the sufficiency. We have established the necessity for the linear case in Lecture 2 via the Farkas Lemma. We will now prove the necessity of the condition for the general *convex* case, and, unlike the situation in Lecture 2, we need some additional assumptions; and in the general nonconvex case the condition in question simply is *not* necessary for infeasibility of (I) [and this is very bad – this is the reason why there exist difficult optimization problems which we do not know how to solve efficiently].

The just presented "preface" explains what we should do; now let us carry out our plan. We start with the aforementioned "minor regularity assumptions".

Definition 4.1.1 [Slater Condition] *Let* $X \subset \mathbf{R}^n$ *and* $g_1, ..., g_m$ *be real-valued functions on* X*. We say that these functions satisfy the Slater condition on* X*, if there exists* $x \in X$ *such that* $g_j(x) < 0,\ j = 1, ..., m$*.*

An inequality constrained program

$$f(x) \to \min \mid\ g_j(x) \leq 0,\ j = 1, ..., m,\ x \in X \tag{IC}$$

($f, g_1, ..., g_m$ *are real-valued functions on* X*) is called to satisfy the Slater condition, if* $g_1, ..., g_m$ *satisfy this condition on* X*.*

We are about to establish the following fundamental fact:

Theorem 4.1.1 [Convex Theorem on Alternative]
Let $X \subset \mathbf{R}^n$ be convex, let $f, g_1, ..., g_m$ be real-valued convex functions on X, and let $g_1, ..., g_m$ satisfy the Slater condition on X. Then system (I) *is solvable if and only if system* (II) *is unsolvable.*

The "only if" part of the statement – "if (II) has a solution, then (I) has no solutions" – is given by Proposition 4.1.1. What we need is to prove the inverse statement. Thus, let us assume that (I) has no solutions, and let us prove that then (II) has a solution.

1^0. Let us set

$$F(x) = \begin{pmatrix} f(x) \\ g_1(x) \\ ... \\ g_m(x) \end{pmatrix}$$

and consider two sets in $\mathbf{R}^{m+1}$:

$$S = \{u = (u_0, ..., u_m) \mid \exists x \in X : F(x) \leq u\}$$

and

$$T = \{(u_0, ..., u_m) \mid u_0 < c, u_1 \leq 0, u_2 \leq 0, ..., u_m \leq 0\}.$$

I claim that

- (i) S and T are nonempty convex sets;
- (ii) S and T do not intersect.

Indeed, convexity and nonemptiness of T is evident, same as nonemptiness of S. Convexity of S is an immediate consequence of the fact that X and $f, g_1, ..., g_m$ are convex. Indeed, assuming that $u', u'' \in S$, we conclude that there exist $x', x'' \in X$ such that $F(x') \leq u'$ and $F(x'') \leq u''$, whence, for every $\lambda \in [0, 1]$.

$$\lambda F(x') + (1 - \lambda)F(x'') \leq \lambda u' + (1 - \lambda)u''.$$

The left hand side in this inequality, due to convexity of X and $f, g_1, ..., g_m$, is $\geq F(y)$, $y = \lambda x' + (1 - \lambda)x''$. Thus, for the point $v = \lambda u' + (1 - \lambda)u''$ there exists $y \in X$ with $F(y) \leq v$, whence $v \in S$. Thus, S is convex.

The fact that $S \cap T = \emptyset$ is an evident equivalent reformulation of the fact that (I) has no solutions.

2^0. Since S and T are nonempty convex sets with empty intersection, according to the Separation Theorem (Lecture 1) they can be separated by a linear form: there exist $a = (a_0, ..., a_m) \neq 0$ such that

$$\inf_{u \in S} \sum_{j=0}^{m} a_j u_j \geq \sup_{u \in T} \sum_{j=0}^{m} a_j u_j. \tag{4.1.5}$$

3^0. Let us look what can be said about the vector a. I claim that, first,

$$a \geq 0 \tag{4.1.6}$$

and, second,

$$a_0 > 0. \tag{4.1.7}$$

Indeed, to prove (4.1.6) note that if some a_i were negative, then the right hand side in (4.1.5) would be $+\infty$ [1)], which is forbidden by (4.1.5).

Thus, $a \geq 0$; with this in mind, we can immediately compute the right hand side of (4.1.5):

$$\sup_{u\in T} \sum_{j=0}^{m} a_j u_j = \sup_{u_0<c, u_1,\ldots,u_m\leq 0} \sum_{j=0}^{m} a_j u_j = a_0 c.$$

Since for every $x \in X$ the point $F(x)$ belongs to S, the left hand side in (4.1.5) is not less that

$$\inf_{x\in X} \left[a_0 f(x) + \sum_{j=1}^{m} a_j g_j(x) \right];$$

combining our observations, we conclude that (4.1.5) implies the inequality

$$\inf_{x\in X} \left[a_0 f(x) + \sum_{j=1}^{m} a_j g_j(x) \right] \geq a_0 c. \tag{4.1.8}$$

Now let us prove that $a_0 > 0$. This crucial fact is an immediate consequence of the Slater condition. Indeed, let $\bar{x} \in X$ be the point given by this condition, so that $g_j(\bar{x}) < 0$. From (4.1.8) we conclude that

$$a_0 f(\bar{x}) + \sum_{j=0}^{m} a_j g_j(\bar{x}) \geq a_0 c.$$

If a_0 were 0, then the right hand side of this inequality would be 0, while the left one would be the combination $\sum_{j=0}^{m} a_j g_j(\bar{x})$ of *negative* reals $g_j(\bar{x})$ with *nonnegative* coefficients a_j *not all equal to 0* [2)], so that the left hand side is strictly negative, which is the desired contradiction.

4^0. Now we are done: since $a_0 > 0$, we are in our right to divide both sides of (4.1.8) by a_0 and thus get

$$\inf_{x\in X} \left[f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \right] \geq c, \tag{4.1.9}$$

where $\lambda_j = a_j/a_0 \geq 0$. Thus, (II) has a solution. ∎

4.1.2 Lagrange Function and Lagrange Duality

The result of Convex Theorem on Alternative brings to our attention the function

$$\underline{L}(\lambda) = \inf_{x\in X} \left[f_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \right], \tag{4.1.10}$$

[1)] look what happens when all coordinates in u, except the ith one, are fixed at values allowed by the description of T and u_i is a large in absolute value negative real

[2)] indeed, from the very beginning we know that $a \neq 0$, so that if $a_0 = 0$, then not all a_j, $j \geq 1$, are zeros

same as the aggregate

$$L(x,\lambda) = f_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) \tag{4.1.11}$$

from which this function comes. Aggregate (4.1.11) has a special name – it is called the *Lagrange function* of the inequality constrained optimization program

$$\text{(IC)} \qquad f(x) \to \min \; g_j(x) \le 0,\; j = 1, ..., m,\; x \in X.$$

The Lagrange function of an optimization program is a very important entity: most of optimality conditions are expressed in terms of this function. Let us start with translating of what we already know to the language of the Lagrange function.

Convex Programming Duality Theorem

Theorem 4.1.2 *Consider an arbitrary inequality constrained optimization program* (IC). *Then*

(i) *The infinum*

$$\underline{L}(\lambda) = \inf_{x \in X} L(x,\lambda)$$

of the Lagrange function in $x \in X$ is, for every $\lambda \ge 0$, a lower bound for the optimal value in (IC), *so that the optimal value in the optimization program*

$$\text{(IC}^*\text{)} \qquad \sup_{\lambda \ge 0} \underline{L}(\lambda)$$

also is a lower bound for the optimal value in (IC)*;*

(ii) [Convex Duality Theorem] *If* (IC)

- *is convex,*
- *is below bounded*

and

- *satisfies the Slater condition,*

then the optimal value in (IC*) *is attained and is equal to the optimal value in* (IC).

Proof. (i) is nothing but Proposition 4.1.1 (please understand why); it makes sense, however, to repeat here the corresponding one-line reasoning:

Let $\lambda \ge 0$; in order to prove that

$$\underline{L}(\lambda) \equiv \inf_{x \in X} L(x,\lambda) \le c^* \quad [L(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x)],$$

c^* being the optimal value in (IC), note that if x is feasible for (IC), then evidently $L(x,\lambda) \le f(x)$, so that the infinum of L in $x \in X$ is $\le$ the infinum c^* of f over the feasible set of (IC). ∎

(ii) is an immediate consequence of the Convex Theorem on Alternative. Indeed, let c^* be the optimal value in (IC). Then the system

$$f(x) < c^*,\ g_j(x) \leq 0,\ j = 1, ..., m$$

has no solutions in X, and by the above Theorem the system (II) associated with $c = c^*$ has a solution, i.e., there exists $\lambda^* \geq 0$ such that $\underline{L}(\lambda^*) \geq c^*$. But we know from (i) that the strict inequality here is impossible and, besides this, that $\underline{L}(\lambda) \leq c^*$ for every $\lambda \geq 0$. Thus, $\underline{L}(\lambda^*) = c^*$ and λ^* is a maximizer of $\underline{L}$ over $\lambda \geq 0$. ∎

The Dual Program

Theorem 4.1.2 establishes certain connection between two optimization programs – the "primal" program

$$\text{(IC)} \qquad f(x) \to \min \mid\ g_j(x) \leq 0,\ j = 1, ..., m,\ x \in X.$$

and its *Lagrange Dual*

$$(\text{IC}^*) \qquad \sup_{\lambda \geq 0} \underline{L}(\lambda), \quad [\underline{L}(\lambda) = \inf_{x \in X} L(x, \lambda)]$$

(the variables λ of the dual problem are called the *Lagrange multipliers* of the primal problem). The Theorem says that the optimal value in the dual problem is $\leq$ the one in the primal, and under some favorable circumstances (the primal problem is convex below bounded and satisfies the Slater condition) the optimal values in the programs are equal to each other.

In our formulation there is some asymmetry between the primal and the dual programs. In fact both of the programs are related to the Lagrange function in a quite symmetric way. Indeed, consider the program

$$\min_{x \in X} \overline{L}(x), \quad \overline{L}(x) = \sup_{\lambda \geq 0} L(\lambda, x).$$

The objective in this program clearly is $+\infty$ at every point $x \in X$ which is not feasible for (IC) and is $f(x)$ at the feasible set of (IC), so that the program is equivalent to (IC). We see that both the primal and the dual programs come from the Lagrange function: in the primal problem, we minimize over X the result of maximization of $L(x, \lambda)$ in $\lambda \geq 0$, and in the dual program we maximize over $\lambda \geq 0$ the result of minimization of $L(x, \lambda)$ in $x \in X$. This is a particular (and the most important) example of a *zero sum two person game* (cf. the non-obligatory Section Section 4.3).

We have said that the optimal values in (IC) and (IC*) are equal to each other under some convexity and regularity assumptions. There is also another way to say when these optimal values are equal – this is always the case when the Lagrange function possesses a saddle point, i.e., there exists a pair $x^* \in X, \lambda^* \geq 0$ such that at the pair $L(x, \lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X, \lambda \geq 0.$$

It can be easily demonstrated (do it by yourself or look at Theorem 4.3.1 in Section 4.3) that

Proposition 4.1.2 *(x^*, λ^*) is a saddle point of the Lagrange function L of* (IC) *if and only if x^* is an optimal solution to* (IC)*, λ^* is an optimal solution to* (IC*) *and the optimal values in the indicated problems are equal to each other.*

Our current goal is to extract from what we already know optimality conditions for convex programs.

4.1.3 Optimality Conditions in Convex Programming

We start from the *saddle point* formulation of the Optimality Conditions.

Theorem 4.1.3 [Saddle Point formulation of Optimality Conditions in Convex Programming]
Let (IC) *be an optimization program, $L(x, \lambda)$ be its Lagrange function, and let $x^* \in X$. Then*

(i) *A sufficient condition for x^* to be an optimal solution to* (IC) *is the existence of the vector of Lagrange multipliers $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrange function $L(x, \lambda)$, i.e., a point where $L(x, \lambda)$ attains its minimum as a function of $x \in X$ and attains its maximum as a function of $\lambda \geq 0$:*

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X, \lambda \geq 0. \tag{4.1.12}$$

(ii) *if the problem* (IC) *is convex and satisfies the Slater condition, then the above condition is necessary for optimality of x^*: if x^* is optimal for* (IC)*, then there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrange function.*

Proof. (i): assume that for a given $x^* \in X$ there exists $\lambda^* \geq 0$ such that (4.1.12) is satisfied, and let us prove that then x^* is optimal for (IC). First of all, x^* is feasible: indeed, if $g_j(x^*) > 0$ for some j, then, of course, $\sup_{\lambda \geq 0} L(x^*, \lambda) = +\infty$ (look what happens when all λ's, except λ_j, are fixed, and $\lambda_j \to +\infty$); but $\sup_{\lambda \geq 0} L(x^*, \lambda) = +\infty$ is forbidden by the second inequality in (4.1.12).

Since x^* is feasible, $\sup_{\lambda \geq 0} L(x^*, \lambda) = f(x^*)$, and we conclude from the second inequality in (4.1.12) that $L(x^*, \lambda^*) = f(x^*)$. Now the first inequality in (4.1.12) says that

$$f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x) \geq f(x^*) \quad \forall x \in X.$$

This inequality immediately implies that x^* is optimal: indeed, if x is feasible for (IC), then the left hand side in the latter inequality is $\leq f(x)$ (recall that $\lambda^* \geq 0$), and the inequality implies that $f(x) \geq f(x^*)$. ∎

(ii): Assume that (IC) is a convex program, x^* is its optimal solution and the problem satisfies the Slater condition; we should prove that then there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrange function, i.e., that (4.1.12) is satisfied. As we know from the Convex Programming Duality Theorem (Theorem 4.1.2.(ii)), the dual problem (IC*) has

a solution $\lambda^* \geq 0$ and the optimal value of the dual problem is equal to the optimal value in the primal one, i.e., to $f(x^*)$:

$$f(x^*) = \underline{L}(\lambda^*) \equiv \inf_{x \in X} L(x, \lambda^*). \tag{4.1.13}$$

We immediately conclude that

$$\lambda_j^* > 0 \Rightarrow g_j(x^*) = 0$$

(this is called *complementary slackness*: positive Lagrange multipliers can be associated only with active (satisfied at x^* as equalities) constraints. Indeed, from (4.1.13) it for sure follows that

$$f(x^*) \leq L(x^*, \lambda^*) = f(x^*) + \sum_{j=1}^{m} \lambda_j^* g_j(x^*);$$

the terms in the $\sum_j$ in the right hand side are nonpositive (since x^* is feasible for (IC)), and the sum itself is nonnegative due to our inequality; it is possible if and only if all the terms in the sum are zero, and this is exactly the complementary slackness.

From the complementary slackness we immediately conclude that $f(x^*) = L(x^*, \lambda^*)$, so that (4.1.13) results in

$$L(x^*, \lambda^*) = f(x^*) = \inf_{x \in X} L(x, \lambda^*).$$

On the other hand, since x^* is feasible for (IC), we have $L(x^*, \lambda) \leq f(x^*)$ whenever $\lambda \geq 0$. Combining our observations, we conclude that

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$$

for all $x \in X$ and all $\lambda \geq 0$. ∎

Note that (i) is valid for an arbitrary inequality constrained optimization program, not necessarily convex. This is another story that in the nonconvex case the *sufficient* condition for optimality given by (i) is extremely far from being necessary and is satisfied "almost never". In contrast to this, in the convex case the condition in question is not only sufficient, but also "nearly necessary" – it for sure is necessary when (IC) is a convex program satisfying the Slater condition.

Theorem 4.1.3 is basically the strongest optimality condition for a Convex Programming program, but it is, in a sense, "implicit" – it is expressed in terms of saddle point of the Lagrange function, and it is unclear how to verify that something is the saddle point of the Lagrange function. Let us try to understand what does it mean that (x^*, λ^*) is a saddle point of the Lagrange function. By definition, it means that

- (A) $L(x^*, \lambda)$ attains its maximum in $\lambda \geq 0$ at the point $\lambda = \lambda^*$
- (B) $L(x, \lambda^*)$ attains its minimum in $x \in X$ at the point $x = x^*$.

It is immediate to understand what (A) means: it means exactly that

x^* is feasible for (IC) *and the complementary slackness condition*

$$\lambda_j^* g_j(x^*) = 0$$

holds (positive λ_j^* can be associated only with the constraints $g_j(x) \leq 0$ active at x^*, i.e., with those satisfying at the point as equalities).

Indeed, the function

$$L(x^*, \lambda) = f(x^*) + \sum_{j=1}^{m} \lambda_j g_j(x^*)$$

is affine in λ, and we of course understand when and where such a function attains its maximum on the nonnegative orthant: it is above bounded on the orthant if and only if all the coefficients at λ_j are nonpositive (i.e., if and only if x^* is feasible for (IC)), and if it is the case, then the set of maximizers is exactly the set

$$\{\lambda \geq 0 \mid \lambda_j g_j(x^*) = 0,\ j = 1, ..., m\}.$$

Now, what does it mean that the function $L(x, \lambda^*)$ attains its minimum over $x \in X$ at x^*?

Karush-Kuhn-Tucker Optimality Conditions in Convex case The answer depends on how "good" is the Lagrange function as a function of x. E.g., when (IC) is a convex program, then

$$L(x, \lambda^*) = f(x) + \sum_{j=1}^{m} \lambda_j^* g_j(x)$$

is convex in $x \in X$ (recall that $\lambda^* \geq 0$); when $f, g_1, ..., g_m$ are differentiable at x^*, so is $L(x, \lambda^*)$, etc. Recall that a general answer to this question is provided by the optimality condition of Theorem 3.5.2 – we should have $0 \in \partial_x L(x^*, \lambda^*)$. The latter rule is easy to interpret when x^* is an interior point of X and $L(x, \lambda^*)$ is differentiable at x^*: in this case we have

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) = 0$$

(cf. Theorem 3.5.4).

A natural question is:

what does it mean that $L(x, \lambda^*)$ attains it minimum in $x \in X$ at x^* when x^* is not interior to X?

Let us assume, for the sake of simplicity, that $L(x, \lambda^*)$ is differentiable at x^* From the discussion of Section 3.5 we know that in this case the necessary and sufficient condition for this is that the gradient of $L(x, \lambda^*)$ at x^* should belong to the normal cone of the set X at x^*. Moreover, we know at least two cases when this "belongs to the normal cone" can be said in quite explicit words – these are the cases when

- (a) X is an arbitrary convex set and $x^* \in \operatorname{int} X$ – here "to belong to the normal cone" simply means to be zero;

- (b) X is a polyhedral convex set:

$$X = \{x \in \mathbf{R}^n \mid a_i^T x - b_i \leq 0,\ i = 1, ..., M\}$$

and x^* is an arbitrary point from X; here "to belong to the normal cone of X at x^*" means "to be a combination, with nonpositive coefficients, of a_i corresponding to "active" i – those with $a_i^T x^* = b_i$.

Now consider a "mixture" of these two cases; namely, assume that X in (IC) is the intersection of arbitrary convex set X' and a *polyhedral* convex set X'':

$$X = X' \cap X'',$$

$$X'' = \{x \mid g_{i+m}(x) \equiv a_i^T x - b_i \leq 0,\ i = 1, ..., M\}.$$

Let x^* be a feasible solution to (IC) which is *interior* for X', and let $f, g_1, ..., g_m$ be convex functions which are differentiable at x^*. When x^* is optimal for (IC)?

As we already know, the *sufficient* condition (which is also *necessary*, if $g_1, ..., g_m$ satisfy the Slater condition on X) is that there exist nonnegative Lagrange multipliers $\lambda_1^*, ..., \lambda_m^*$ such that

$$\lambda_j^* g_j(x^*) = 0,\ \ j = 1, ..., m \tag{4.1.14}$$

and

$$x^* \in \operatorname*{Argmin}_X [f(x) + \sum_{j=1}^m \lambda_j^* g_j(x)] \tag{4.1.15}$$

Now let us look what the latter condition actually means. By assumption, x^* is an interior point of X'. It follows that if x^* is a minimizer of the function $\phi(x) = f(x) + \sum_{j=1}^m \lambda_j^* g_j(x)$ on X, it is also a local minimizer of the function on X''; since ϕ is convex, x^* is also a global minimizer of ϕ on X''. Vice versa, if x^* is a minimizer of ϕ on X'', it is, of course, a minimizer of the function on the smaller set X. Thus, (4.1.15) says exactly that ϕ attains at x^* its minimum on the polyhedral set X''. But we know when a convex differentiable at x^* function ϕ attains at x^* its minimum on polyhedral set: this is the case if and only if

$$\nabla\phi(x^*) + \sum_{i \in I} \mu_i^* a_i = 0 \tag{4.1.16}$$

where $\mu_i^* \geq 0$ and I is the set of indices of those linear constraints $g_{m+i}(x) \equiv a_i^T x - b \geq 0$ in the description of X'' which are active (are satisfied as equalities) at x^*.

Now let us set $\lambda_{m+i}^* = \mu_i^*$ for $i \in I$ and $\lambda_{m+i}^* = 0$ for $i \notin I$, $i \leq M$. With this notation, we clearly have

$$\lambda_j^* \geq 0,\ \lambda_j^* g_j(x^*) = 0,\ j = 1, ..., m + M \tag{4.1.17}$$

while (4.1.16) says that

$$\nabla f(x^*) + \sum_{i=1}^{m+M} \lambda_j^* \nabla g_j(x^*) = 0. \tag{4.1.18}$$

Summarizing our considerations, we conclude that under our assumptions (the problem is convex, the data are differentiable at x^*, x^* is a feasible solution which is an interior point of X') *sufficient (and necessary and sufficient, if $g_1, ..., g_m$ satisfy the Slater condition on X) condition for optimality of x^* is existence of Largange multipliers λ_j^*, $j = 1, ..., m + M$, satisfying (4.1.17) and (4.1.18).*

Note that this optimality condition looks exactly as if we treat both the constraints $g_1(x) \leq 0, ..., g_m(x) \leq 0$ and the linear constraints defining X'' as functional constraints, and treat X', and not $X = X' \cap X''$, as the domain of the problem. There is, anyhow, great difference: with this new interpretation of the data, in order to get necessity of our optimality condition, we were supposed to assume that all $m+M$ our new functional constraints satisfy the Slater condition: there exists $\bar{x} \in X'$ such that $g_j(\bar{x}) < 0$, $j = 1, ..., m + M$. With our approach we got necessity under weaker assumption: there should exist $\bar{x} \in X'$ where the "complicated" constraints $g_1(x) \leq 0, ..., g_m(x) \leq 0$ are satisfied as strict inequalities, while the linear constraints $g_{m+1}(x) \leq 0$,..., $g_{m+M}(x) \leq 0$ simply are satisfied (so that $\bar{x}$ also belongs to X'' and thus to X).

The results of our considerations definitely deserve to be formulated as a theorem (where we slightly change the notation: what will be m and X, in the above considerations were $m + M$ and X'):

Theorem 4.1.4 [Karush-Kuhn-Tucker Optimality Conditions in Convex case]
Let (IC) *be a convex program, let $x^* \in X$ be an interior feasible solution to* (IC) *($x^* \in \operatorname{int} X$), and let $f, g_1, ..., g_m$ be differentiable at x^*.*

(i) [Sufficiency] *The Karush-Kuhn-Tucker condition:*

There exist nonnegative Lagrange multipliers λ_j^*, $j = 1, ..., m$, such that

$$\lambda_j^* g_j(x^*) = 0,\ j = 1, ..., m \quad \text{[complementary slackness]} \tag{4.1.19}$$

and

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda_j^* \nabla g_j(x^*) = 0, \tag{4.1.20}$$

is sufficient for x^* to be optimal solution to (IC).

(ii) [Necessity and sufficiency] *Under the "restricted Slater assumption":*

there exists $\bar{x} \in X$ such that the nonlinear g_j are strictly negative, and linear g_j are nonpositive at $\bar{x}$

the Karush-Kuhn-Tucker condition from (i) *is necessary and sufficient for x^* to be optimal solution to* (IC).

Note that the optimality conditions from Theorem 3.5.4 is a particular case of the above Theorem.

4.2 Duality in Linear and Convex Quadratic Programming

The fundamental role of the Lagrange function and Lagrange Duality in Optimization is clear already from the Optimality Conditions given by Theorem 4.1.3, but this role is not restricted by this theorem only. There are some cases when we can explicitly write down the Lagrange dual, and whenever it is the case, we get a pair of explicitly formulated and closely related to each other optimization programs – the *primal-dual pair*; analyzing the problems simultaneously, we get more information about their properties (and get a possibility to solve the problems numerically in a more efficient way) than it is possible when we we restrict ourselves with only one problem of the pair. Let us look at two of such cases.

4.2.1 Linear Programming Duality

Let us start with some general observation. Note that the Karush-Kuhn-Tucker condition under the assumption of the Theorem ((IC) is convex, x^* is an interior point of X, $f, g_1, ..., g_m$ are differentiable at x^*) is exactly the condition that $(x^*, \lambda^* = (\lambda_1^*, ..., \lambda_m^*))$ is a saddle point of the Lagrange function

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) : \tag{4.2.21}$$

(4.1.19) says that $L(x^*, \lambda)$ attains its maximum in $\lambda \geq 0$, and (4.1.20) says that $L(x, \lambda^*)$ attains its minimum in x at $x = x^*$.

Now consider the particular case of (IC) where $X = \mathbf{R}^n$ is the entire space, the objective f is convex and everywhere differentiable and the constraints $g_1, ..., g_m$ are linear. For this case, Theorem 4.1.4 says to us that the KKT (Karush-Kuhn-Tucker) condition is necessary and sufficient for optimality of x^*; as we just have explained, this is the same as to say that the necessary and sufficient condition of optimality for x^* is that x^* along with certain $\lambda^* \geq 0$ form a saddle point of the Lagrange function. Combining these observations with Proposition 4.1.2, we get the following simple result:

Proposition 4.2.1 *Let* (IC) *be a convex program with $X = \mathbf{R}^n$, everywhere differentiable objective f and linear constraints $g_1, ..., g_m$. Then x^* is optimal solution to* (IC) *if and only if there exists $\lambda^* \geq 0$ such that (x^*, λ^*) is a saddle point of the Lagrange function (4.2.21) (regarded as a function of $x \in \mathbf{R}^n$ and $\lambda \geq 0$). In particular,* (IC) *is solvable if and only if L has saddle points, and if it is the case, then both* (IC) *and its Lagrange dual*

$$(\mathrm{IC}^*): \quad \underline{L}(\lambda) \to \max \mid \lambda \geq 0$$

are solvable with equal optimal values.

Let us look what this proposition says in the Linear Programming case, i.e., when (IC) is the program

$$(P) \quad f(x) = c^T x \to \min \mid g_j(x) \equiv b_j - a_j^T x \leq 0,\ j = 1, ..., m.$$

In order to get the Lagrange dual, we should form the Lagrange function

$$L(x,\lambda) = f(x) + \sum_{j=1}^{m} \lambda_j g_j(x) = [c - \sum_{j=1}^{m} \lambda_j a_j]^T x + \sum_{j=1}^{m} \lambda_j b_j$$

of (IC) and to minimize it in $x \in \mathbf{R}^n$; this will give us the dual objective. In our case the minimization in x is immediate: the minimal value is $-\infty$, if $c - \sum_{j=1}^m \lambda_j a_j \neq 0$, and is $\sum_{j=1}^m \lambda_j b_j$ otherwise. We see that the Lagrange dual is

$$(D) \quad b^T\lambda \to \max \mid \sum_{j=1}^{m} \lambda_j a_j = c,\ \lambda \geq 0.$$

The problem we get is the usual LP dual to (P), and Proposition 4.2.1 is one of the equivalent forms of the Linear Programming Duality Theorem which we already know from Lecture 2.

4.2.2 Quadratic Programming Duality

Now consider the case when the original problem is linearly constrained convex quadratic program

$$(P) \quad f(x) = \frac{1}{2}x^T D x + c^T x \mid\ g_j(x) \equiv b_j - a_j^T x \leq 0,\ j = 1, ..., m,$$

where the objective is a strictly convex quadratic form, so that $D = D^T$ is positive definite matrix: $x^T D x > 0$ whenever $x \neq 0$. It is convenient to rewrite the constraints in the vector-matrix form

$$g(x) = b - Ax \leq 0,\ b = \begin{pmatrix} b_1 \\ ... \\ b_m \end{pmatrix},\ A = \begin{pmatrix} a_1^T \\ ... \\ a_m^T \end{pmatrix}.$$

In order to form the Lagrange dual to (P) program, we write down the Lagrange function

$$\begin{array}{rcl} L(x,\lambda) & = & f(x) + \sum_{j=1}^m \lambda_j g_j(x) \\ & = & c^T x + \lambda^T(b - Ax) + \frac{1}{2}x^T D x \\ & = & \frac{1}{2}x^T D x - [A^T\lambda - c]^T x + b^T\lambda \end{array}$$

and minimize it in x. Since the function is convex and differentiable in x, the minimum, if exists, is given by the Fermat rule

$$\nabla_x L(x,\lambda) = 0,$$

which in our situation becomes

$$Dx = [A^T\lambda - c].$$

Since D is positive definite, it is nonsingular, so that the Fermat equation has a unique solution which is the minimizer of $L(\cdot,\lambda)$; this solution is

$$x = D^{-1}[A^T\lambda - c].$$

Substituting the value of x into the expression for the Lagrange function, we get the dual objective:

$$\underline{L}(\lambda) = -\frac{1}{2}[A^T\lambda - c]^T D^{-1}[A^T\lambda - c] + b^T\lambda,$$

and the dual problem is to maximize this objective over the nonnegative orthant. Usually people rewrite this dual problem equivalently by introducing additional variables

$$t = -D^{-1}[A^T\lambda - c] \quad [[A^T\lambda - c]^T D^{-1}[A^T\lambda - c] = t^T Dt];$$

with this substitution, the dual problem becomes

$$(D) \quad -\frac{1}{2}t^T Dt + b^T\lambda \to \max \mid A^T\lambda + Dt = c,\ \lambda \geq 0.$$

We see that the dual problem also turns out to be linearly constrained convex quadratic program.

Note also that in the case in question feasible problem (P) automatically is solvable[3)]

With this observation, we get from Proposition 4.2.1 the following

Theorem 4.2.1 [Duality Theorem in Quadratic Programming]
Let (P) be feasible quadratic program with positive definite symmetric matrix D in the objective. Then both (P) and (D) are solvable, and the optimal values in the problems are equal to each other.

The pair $(x; (\lambda, t))$ of feasible solutions to the problems is comprised of the optimal solutions to them

(i) *if and only if the primal objective at x is equal to the dual objective at (λ, t)* ["zero duality gap" optimality condition]
same as

(ii) *if and only if*

$$\lambda_i(Ax - b)_i = 0,\ i = 1, ..., m, \quad \text{and} \quad t = -x. \tag{4.2.22}$$

Proof. (i): we know from Proposition 4.2.1 that the optimal value in minimization problem (P) is equal to the optimal value in the maximization problem (D). It follows that the the value of the primal objective at any primal feasible solution is $\geq$ the value of the dual objective at any dual feasible solution, and equality is possible if and only if these values coincide with the optimal values in the problems, as claimed in (i).

(ii): Let us compute the difference Δ between the values of the primal objective at primal feasible solution x and the dual objective at dual feasible solution (λ, t):

$$\begin{array}{rcl} \Delta & = & c^T x + \frac{1}{2}x^T Dx - [b^T\lambda - \frac{1}{2}t^T Dt] \\ & = & [A^T\lambda + Dt]^T x + \frac{1}{2}x^T Dx + \frac{1}{2}t^T Dt - b^T\lambda \\ & & [\text{since } A^T\lambda + Dt = c] \\ & = & \lambda^T[Ax - b] + \frac{1}{2}[x + t]^T D[x + t] \end{array}$$

[3)] since its objective, due to positive definiteness of D, goes to infinity as $|x| \to \infty$, and due to the following general fact:
Let (IC) *be a feasible program with closed domain X, continuous on X objective and constraints and such that $f(x) \to \infty$ as $x \in X$ "goes to infinity" (i.e., $|x| \to \infty$). Then* (IC) *is solvable.*
You are welcome to prove this simple statement (it is among the exercises accompanying the Lecture)

Since $Ax - b \geq 0$ and $\lambda \geq 0$ due to primal feasibility of x and dual feasibility of (λ, t), both terms in the resulting expression for Δ are nonnegative. Thus, $\Delta = 0$ (which, by (i). is equivalent to optimality of x for (P) and optimality of (λ, t) for (D)) if and only if $\sum_{j=1}^m \lambda_j (Ax - b)_j = 0$ and $(x + t)^T D(x + t) = 0$. The first of these equalities, due to $\lambda \geq 0$ and $Ax \geq b$, is equivalent to $\lambda_j (Ax - b)_j = 0$, $j = 1, ..., m$; the second, due to positive definiteness of D, is equivalent to $x + t = 0$. ∎

4.3 Saddle Points

This section is not obligatory

4.3.1 Definition and Game Theory interpretation

When speaking about the "saddle point" formulation of optimality conditions in Convex Programming, we touched a very interesting in its own right topic of Saddle Points. This notion is related to the situation as follows. Let $X \subset \mathbf{R}^n$ and $\Lambda \subset \mathbf{R}^m$ be two nonempty sets, and let

$$L(x, \lambda) : X \times \Lambda \to \mathbf{R}$$

be a real-valued function of $x \in X$ and $\lambda \in \Lambda$. We say that a point $(x^*, \lambda^*) \in X \times \Lambda$ is a *saddle point* of L on $X \times \Lambda$, if L attains in this point its maximum in $\lambda \in \Lambda$ and attains at the point its minimum in $x \in X$:

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall (x, \lambda) \in X \times \Lambda. \tag{4.3.23}$$

The notion of a saddle point admits natural interpretation in *game terms*. Consider what is called a *two person zero sum game* where player I chooses $x \in X$ and player II chooses $\lambda \in \Lambda$; after the players have chosen their decisions, player I pays to player II the sum $L(x, \lambda)$. Of course, I is interested to minimize his payment, while II is interested to maximize his income. What is the natural notion of the equilibrium in such a game – what are the choices (x, λ) of the players I and II such that every one of the players is not interested to vary his choice independently on whether he knows the choice of his opponent? It is immediately seen that the equilibria are exactly the saddle points of the cost function L. Indeed, if (x^*, λ^*) is such a point, than the player I is not interested to pass from x to another choice, given that II keeps his choice λ fixed: the first inequality in (4.3.23) shows that such a choice cannot decrease the payment of I. Similarly, player II is not interested to choose something different from λ^*, given that I keeps his choice x^* – such an action cannot increase the income of II. On the other hand, if (x^*, λ^*) is not a saddle point, then either the player I can decrease his payment passing from x^* to another choice, given that II keeps his choice at λ^* – this is the case when the first inequality in (4.3.23) is violated, or similarly for the player II; thus, equilibria are exactly the saddle points.

The game interpretation of the notion of a saddle point motivates deep insight into the structure of the set of saddle points. Consider the following two situations:

(A) player I makes his choice first, and player II makes his choice already knowing the choice of I;

(B) vice versa, player II chooses first, and I makes his choice already knowing the choice of II.

In the case (A) the reasoning of I is: If I choose some x, then II of course will choose λ which maximizes, for my x, my payment $L(x,\lambda)$, so that I will pay the sum

$$\overline{L}(x) = \sup_{\lambda\in\Lambda} L(x,\lambda);$$

Consequently, my policy should be to choose x which minimizes my *loss function* $\underline{L}$, i.e., the one which solves the optimization problem

$$\text{(I)} \qquad \min_{x\in X} \overline{L}(x);$$

with this policy my anticipated payment will be

$$\inf_{x\in X} \overline{L}(x) = \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda).$$

In the case (B), similar reasoning of II enforces him to choose λ maximizing his *profit function*

$$\underline{L}(\lambda) = \inf_{x\in X} L(x,\lambda),$$

i.e., the one which solves the optimization problem

$$\text{(II)} \qquad \max_{\lambda\in\Lambda} \underline{L}(\lambda);$$

with this policy, the anticipated profit of II is

$$\sup_{\lambda\in\Lambda} \underline{L}(\lambda) = \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda).$$

Note that these two reasonings relate to two *different* games: the one with priority of II (when making his decision, II already knows the choice of I), and the one with similar priority of I. Therefore we should not, generally speaking, expect that the anticipated loss of I in (A) is equal to the anticipated profit of II in (B). What can be guessed is that the anticipated loss of I in (B) is *less than or equal to* the anticipated profit of II in (A), since the conditions of the game (B) are better for I than those of (A). Thus, we may guess that independently of the structure of the function $L(x,\lambda)$, there is the inequality

$$\sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) \le \inf_{x\in X} \sup_{\lambda\in\Lambda} L(x,\lambda). \tag{4.3.24}$$

This inequality indeed is true; which is seen from the following reasoning:

$$\begin{array}{l} \forall y \in X: \quad \inf_{x\in X} L(x,\lambda) \le L(y,\lambda) \Rightarrow \\ \forall y \in X: \sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda) \le \sup_{\lambda\in\Lambda} L(y,\lambda) \equiv \overline{L}(y); \end{array}$$

consequently, the quantity $\sup_{\lambda\in\Lambda} \inf_{x\in X} L(x,\lambda)$ is a lower bound for the function $\overline{L}(y)$, $y \in X$, and is therefore a lower bound for the infinum of the latter function over $y \in X$, i.e., is a lower bound for $\inf_{y\in X} \sup_{\lambda\in\Lambda} L(y,\lambda)$.

Now let us look what happens when the game in question has a saddle point (x^*, λ^*), so that

$$L(x, \lambda^*) \geq L(x^*, \lambda^*) \geq L(x^*, \lambda) \quad \forall (x, \lambda) \in X \times \Lambda. \tag{4.3.25}$$

I claim that if it is the case, then

(*) *x^* is an optimal solution to* (I), *λ^* is an optimal solution to* (II) *and the optimal values in these two optimization problems are equal to each other* (and are equal to the quantity $L(x^*, \lambda^*)$).

Indeed, from (4.3.25) it follows that

$$\underline{L}(\lambda^*) \geq L(x^*, \lambda^*) \geq \overline{L}(x^*),$$

whence, of course,

$$\sup_{\lambda \in \Lambda} \underline{L}(\lambda) \geq \underline{L}(\lambda^*) \geq L(x^*, \lambda^*) \geq \overline{L}(x^*) \geq \inf_{x \in X} \overline{L}(x).$$

the very first quantity in the latter chain is $\leq$ the very last quantity by (4.3.24), which is possible if and only if all the inequalities in the chain are equalities, which is exactly what is said by (A) and (B).

Thus, if (x^*, λ^*) is a saddle point of L, then (*) takes place. We are about to demonstrate that the inverse also is true:

Theorem 4.3.1 [Structure of the saddle point set] *Let $L : X \times \Lambda \to \mathbf{R}$ be a function. The set of saddle points of the function is nonempty if and only if the related optimization problems* (I) *and* (II) *are solvable and the optimal values in the problems are equal to each other. If it is the case, then the saddle points of L are exactly all pairs (x^*, λ^*) with x^* being an optimal solution to* (I) *and λ^* being an optimal solution to* (II), *and the value of the cost function $L(\cdot, \cdot)$ at every one of these points is equal to the common optimal value in* (I) *and* (II).

Proof. We already have established "half" of the theorem: if there are saddle points of L, then their components are optimal solutions to (I), respectively, (II), and the optimal values in these two problems are equal to each other and to the value of L at the saddle point in question. To complete the proof, we should demonstrate that if x^* is an optimal solution to (I), λ^* is an optimal solution to (II) and the optimal values in the problems are equal to each other, then (x^*, λ^*) is a saddle point of L. This is immediate: we have

$$\begin{array}{rcll} L(x, \lambda^*) & \geq & \underline{L}(\lambda^*) & [\text{ definition of } \underline{L}] \\ & = & \overline{L}(x^*) & [\text{by assumption}] \\ & \geq & L(x^*, \lambda) & [\text{definition of } \overline{L}] \end{array}$$

whence

$$L(x, \lambda^*) \geq L(x^*, \lambda) \quad \forall x \in X, \lambda \in \Lambda;$$

substituting $\lambda = \lambda^*$ in the right hand side of this inequality, we get $L(x, \lambda^*) \geq L(x^*, \lambda^*)$, and substituting $x = x^*$ in the right hand side of our inequality, we get $L(x^*, \lambda^*) \geq L(x^*, \lambda)$; thus, (x^*, λ^*) indeed is a saddle point of L. ∎

4.3.2 Existence of saddle points

It is easily seen that a "quite respectable" cost function, say, $L(x,\lambda) = (x-\lambda)^2$ on the unit square $[0,1]\times[0,1]$ has no saddle points. Indeed, here

$$\underline{L}(x) = \sup_{\lambda\in[0,1]} (x-\lambda)^2 = \max\{x^2, (1-x)^2\},$$

$$\overline{L}(\lambda) = \inf_{x\in[0,1]} (x-\lambda)^2 = 0,\ \ \lambda\in[0,1],$$

so that the optimal value in (I) is $\frac{1}{4}$, and the optimal value in (II) is 0; according to Theorem 4.3.1 it means that L has no saddle points.

On the other hand, there are generic cases when L has a saddle point, e.g., when

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) : X\times \mathbf{R}^m_+ \to \mathbf{R}$$

is the Lagrange function of a solvable convex program satisfying the Slater condition. Note that in this case L is convex in x for every $\lambda\in\Lambda\equiv\mathbf{R}^m_+$ and is linear (and therefore concave) in λ for every fixed X. As we will see in a while, these are the structural properties of L which take upon themselves the "main responsibility" for the fact that in the case in question the saddle points exist. Namely, there exists the following

Theorem 4.3.2 [Existence of saddle points of a convex-concave function (Sion-Kakutani)] *Let X and Λ be convex compact sets in $\mathbf{R}^n$ and $\mathbf{R}^m$, respectively, and let*

$$L(x,\lambda) : X\times\Lambda \to \mathbf{R}$$

be a continuous function which is convex in $x\in X$ for every fixed $\lambda\in\Lambda$ and is concave in $\lambda\in\Lambda$ for every fixed $x\in X$. Then L has saddle points on $X\times\Lambda$.

Proof. According to Theorem 4.3.1, we should prove that

- (i) Optimization problems (I) and (II) are solvable
- (ii) the optimal values in (I) and (II) are equal to each other.

(i) is valid independently of convexity-concavity of L and is given by the following routine reasoning from the Analysis:

Since X and Λ are compact sets and L is continuous on $X\times\Lambda$, due to the well-known Analysis theorem L is uniformly continuous on $X\times\Lambda$: for every $\epsilon>0$ there exists $\delta(\epsilon)>0$ such that

$$|x-x'|+|\lambda-\lambda'|\le\delta(\epsilon) \Rightarrow |L(x,\lambda)-L(x',\lambda')|\le\epsilon \quad [4)] \qquad (4.3.26)$$

[4)] for those not too familiar with Analysis, I wish to stress the difference between the usual continuity and the uniform continuity: continuity of L means that given $\epsilon>0$ *and a point* (x,λ), it is possible to choose $\delta>0$ such that (4.3.26) is valid; the corresponding δ may depend on (x,λ), not only on ϵ. Uniform continuity means that this positive δ may be chosen as a function of ϵ only. The fact that a continuous on a compact set function automatically is uniformly continuous on the set is one of the most useful features of compact sets

In particular,

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |L(x, \lambda) - L(x'\lambda)| \leq \epsilon,$$

whence, of course, also

$$|x - x'| \leq \delta(\epsilon) \Rightarrow |\overline{L}(x) - \overline{L}(x')| \leq \epsilon,$$

so that the function $\overline{L}$ is continuous on X. Similarly, $\underline{L}$ is continuous on Λ. Taking in account that X and Λ are compact sets, we conclude that the problems (I) and (II) are solvable.

(ii) is the essence of the matter; here, of course, the entire construction heavily exploits convexity-concavity of L.

0^0. To prove (ii), we first establish the following statement, which is important by its own right:

Lemma 4.3.1 [Minmax Lemma] *Let X be a convex compact set and $f_0, ..., f_N$ be a collection of $N + 1$ convex and continuous functions on X. Then the minmax*

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) \tag{4.3.27}$$

of the collection is equal to the minimum in $x \in X$ of certain convex combination of the functions: there exist nonnegative μ_i, $i = 0, ..., N$, with unit sum such that

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) = \min_{x \in X} \sum_{i=0}^{N} \mu_i f_i(x)$$

Remark 4.3.1 Minimum of *any* convex combination of a collection of *arbitrary* functions is $\leq$ the minmax of the collection; this evident fact can be also obtained from (4.3.24) as applied to the function

$$M(x, \mu) = \sum_{i=0}^{N} \mu_i f_i(x)$$

on the direct product of X and the standard simplex

$$\Delta = \{\mu \in \mathbf{R}^{N+1} \mid \mu \geq 0, \sum_i \mu_i = 1\}.$$

The Minmax Lemma says that if f_i are convex and continuous on a convex compact set X, then the indicated inequality is in fact equality; you can easily verify that this is nothing but the claim that the function M possesses a saddle point. Thus, the Minmax Lemma is in fact a particular case of the Sion-Kakutani Theorem; we are about to give a direct proof of this particular case of the Theorem and then to derive the general case from this particular one.

Proof of the Minmax Lemma. Consider the optimization program

$$(S) \quad t \to \min \mid f_0(x) - t \le 0,\ f_1(x) - t \le 0, ..., f_N(x) - t \le 0,\ x \in X.$$

This clearly is a convex program with the optimal value

$$t^* = \min_{x \in X} \max_{i=0,...,N} f_i(x)$$

(note that (t, x) is feasible solution for (S) if and only if $x \in X$ and $t \ge \max_{i=0,...,N} f_i(x)$). The problem clearly satisfies the Slater condition and is solvable (since X is compact set and f_i, $i = 0, ..., N$, are continuous on X; therefore their maximum also is continuous on X and thus attains its minimum on the compact set X); let (t^*, x^*) be an optimal solution to the problem. According to Theorem 4.1.3, there exists $\lambda^* \ge 0$ such that $((t^*, x^*), \lambda^*)$ is a saddle point of the corresponding Lagrange function

$$L(t, x; \lambda) = t + \sum_{i=0}^{N} \lambda_i (f_i(x) - t) = t(1 - \sum_{i=0}^{N} \lambda_i) + \sum_{i=0}^{N} \lambda_i f_i(x),$$

and the value of this function at $((t^*, x^*), \lambda^*)$ is equal to the optimal value in (S), i.e., to t^*.

Now, since $L(t, x; \lambda^*)$ attains its minimum in (t, x) over the set $\{t \in \mathbf{R}, x \in X\}$ at (t^*, x^*), we should have

$$\sum_{i=0}^{N} \lambda_i^* = 1$$

(otherwise the minimum of L in (t, x) would be $-\infty$). Thus,

$$[\min_{x \in X} \max_{i=0,...,N} f_i(x) =] \quad t^* = \min_{t \in \mathbf{R}, x \in X} \left[t \times 0 + \sum_{i=0}^{N} \lambda_i^* f_i(x) \right],$$

so that

$$\min_{x \in X} \max_{i=0,...,N} f_i(x) = \min_{x \in X} \sum_{i=0}^{N} \lambda_i^* f_i(x)$$

with some $\lambda_i^* \ge 0, \sum_{i=0}^{N} \lambda_i^* = 1$, as claimed. ∎

From the Minmax Lemma to the Sion-Kakutani Theorem. We should prove that the optimal values in (I) and (II) (which, by (i), are well defined reals) are equal to each other, i.e., that

$$\inf_{x \in X} \sup_{\lambda \in \Lambda} L(x, \lambda) = \sup_{\lambda \in \Lambda} \inf_{x \in X} L(x, \lambda).$$

we know from (4.3.26) that the first of these two quantities is greater than or equal to the second, so that all we need is to prove the inverse inequality. For me it is convenient to assume that the right quantity (the optimal value in (II)) is 0, which, of course, does not restrict generality; and all we need to prove is that the left quantity – the optimal value in (I) – cannot be positive.

1^0. What does it mean that the optimal value in (II) is zero? When it is zero, then the function $\underline{L}(\lambda)$ is nonpositive for every λ, or, which is the same, the convex continuous function of $x \in X$ – the function $L(x,\lambda)$ – has nonpositive minimal value over $x \in X$. Since X is compact, this minimal value is achieved, so that the set

$$X(\lambda) = \{x \in X \mid L(x,\lambda) \leq 0\}$$

is nonempty; and since X is convex and L is convex in $x \in X$, the set $X(\lambda)$ is convex (as a level set of a convex function, Lecture 4). Note also that the set is closed (since X is closed and $L(x,\lambda)$ is continuous in $x \in X$).

2^0. Thus, if the optimal value in (II) is zero, then the set $X(\lambda)$ is a nonempty convex compact set for every $\lambda \in \Lambda$. And what does it mean that the optimal value in (I) is nonpositive? It means exactly that there is a point $x \in X$ where the function $\overline{L}$ is nonpositive, i.e., the point $x \in X$ where $L(x,\lambda) \leq 0$ for all $\lambda \in \Lambda$. In other words, to prove that the optimal value in (I) is nonpositive is the same as to prove that *the sets $X(\lambda)$, $\lambda \in \Lambda$, have a point in common.*

3^0. With the above observations we see that the situation is as follows: we are given a family of closed nonempty convex subsets $X(\lambda)$, $\lambda \in \Lambda$, of a compact set X, and we should prove that these sets have a point in common. To this end, in turn, it suffices to prove that every *finite* number of sets from our family have a point in common.[5] Let $X(\lambda_0), ..., X(\lambda_N)$ be $N+1$ sets from our family; we should prove that the sets have a point in common. In other words, let

$$f_i(x) = L(x,\lambda_i),\ i = 0, ..., N;$$

all we should prove is that there exists a point x where all our functions are nonpositive, or, which is the same, that the minmax of our collection of functions – the quantity

$$\alpha \equiv \min_{x \in X} \max_{i=1,...,N} f_i(x)$$

is nonpositive.

The proof of the inequality $\alpha \leq 0$ is as follows. According to the Minmax Lemma (which can be applied in our situation – since L is convex and continuous in x, all f_i are convex and continuous, and X is compact), α is the minimum in $x \in X$ of certain convex combination $\phi(x) = \sum_{i=0}^N \nu_i f_i(x)$ of the functions $f_i(x)$. We have

$$\phi(x) = \sum_{i=0}^N \nu_i f_i(x) \equiv \sum_{i=0}^N \nu_i L(x,\lambda_i) \leq L(x, \sum_{i=0}^N \nu_i \lambda_i)$$

(the last inequality follows from concavity of L in λ; this is the only – and crucial – point where we use this assumption). We see that $\phi(\cdot)$ is majorated by $L(\cdot,\lambda)$ for a properly chosen λ; it follows that the minimum of ϕ in $x \in X$ – and we already know that this minimum is exactly α – is nonpositive (recall that the minimum of L in x is nonpositive for every λ). ∎

[5] To justify this claim, I refer to the *Helly's Theorem* as follows: *Let $X_1, ..., X_N$ is a finite collection of convex subsets of $\mathbf{R}^n$ with $N > n$. If the intersection of every $(n+1)$ sets of the family is not empty that all sets have a point in common.* For the proof of this classical result see, e.g., section 2.2 of this course or any other source.

4.4 Exercises

Exercise 4.4.1 *Prove the following statement:*
Assume that the optimization program

$$f(x) \to \min \mid g_j(x) \le 0,\ j = 1, ..., m,\ h_l(x) = 0,\ l = 1, ..., k,\ x \in X \subset \mathbf{R}^n$$

is feasible, the domain X of the problem is closed, and the functions $f, g_1, ..., g_m, h_1, ..., h_k$ are continuous on X. Assume, besides this, that the problem is "coercive", i.e., there exists a function $s(t) \to \infty$, $t \to \infty$, on the nonnegative ray such that

$$\max\{f(x), g_1(x), ..., g_m(x), |h_1(x)|, ..., |h_k(x)|\} \ge s(|x|) \quad \forall x \in X.$$

Prove that under this assumption the problem is solvable.
Hint: consider what is called *minimizing sequence* $\{x_i\}$, i.e., a sequence of feasible solutions to the problem with the values of the objective converging, as $i \to \infty$, to the optimal value of the problem. Prove that the sequence is bounded and therefore possesses limiting points; verify that every such point is an optimal solution to the problem.

Exercise 4.4.2 *Find the minimizer of a linear function*

$$f(x) = c^T x$$

on the set

$$V_p = \{x \in \mathbf{R}^n \mid \sum_{i=1}^n |x_i|^p \le 1\};$$

here p, $1 < p < \infty$, is a parameter.

Exercise 4.4.3 *Solve the problems*

- $\sum_{i=1}^n c_i x_i \to \min \mid \sum_{i=1}^n x_i^4 \le 1$
- $\sum_{i=1}^n c_i x_i^{-2} \to \min \mid \sum_{i=1}^n x_i^4 \le 1,\ x_i > 0,\ i = 1, ..., n,$
  where c_i are positive reals.

Exercise 4.4.4 *Consider the function*

$$I(u, v) = \sum_{i=1}^k u_i \ln(u_i/v_i)$$

regarded as a function of nonnegative $u \in \mathbf{R}^k$ and positive $v \in \mathbf{R}^k$; here $0 \ln 0 = 0$.
1) Prove that the function is convex in (u, v) on the indicated set
2) Prove that if $u, v \in \Delta = \{z \in \mathbf{R}^k_+ : \sum_i z_i = 1\}$ and $u \ge 0$, then

$$I(u, v) \ge 0,$$

with the inequality being strict provided that $u \neq v$.
Hint: apply Jensen's inequality to the strictly convex on $(0,\infty)$ function $-\ln t$
Comment: vector $z \in \Delta$ can be regarded as probability distribution on k-point set: z_i is the probability assigned to ith element of the set. With this interpretation, $I(u,v)$ is a kind of "directed distance" between probability distributions: it sets into correspondence to an *ordered* pair of distributions certain nonnegative real which is positive whenever the distributions are distinct, and is zero otherwise. This quantity is called the *Kullback distance* (this is not a distance in the standard definition, since it is not symmetric: $I(u,v)$ is not the same as $I(v,u)$). The Kullback distance between distributions plays important role in the Theory of Statistical Decisions (see example in Exercise 4.4.8).

Exercise 4.4.5 *Prove the following statement:*
if $r > 0$ and $\mu \in \mathbf{R}^k$ are given real and vector, then

$$\inf_{v\in\mathbf{R}^k}[r\ln\left(\sum_{i=1}^k \exp\{v_i\}\right) - \mu^T v]$$

differs from $-\infty$ if and only if

$$\mu \geq 0, \sum_i \mu_i = r,$$

and if this is the case, then the indicated inf *is either 0 (when $r = 0$), or is*

$$-\sum_{i=1}^k \mu_i \ln(\mu_i/r) \quad [0\ln 0 = 0].$$

Hint: *it is immediately seen that $\mu \geq 0$ is necessary condition for the infinum in question to be finite. To complete the proof of necessity, you should verify that the indicated* inf *is $-\infty$ also in the case of $\mu \geq 0$ and $\sum_{i=1}^k \mu_i \neq r$; to see this, look what happens if $v_i = t$, $i = 1,...,k$, and t runs over $\mathbf{R}$.*

To prove sufficiency and to get the required representation of the optimal value, assume first that all μ_i are positive and use the Fermat rule to find the minimizer exactly, and then think how to eliminate the zero components of μ, if they are present.

Exercise 4.4.6 *Let*

$$f(x) = \min_y\{|x-y|^2, \;\; Ay \leq b\}, \qquad (P)$$

where $x \in \mathbf{R}^n$, $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$. In other words, $f(x)$ is a squared distance from x to the polyhedron $Y = \{y \in \mathbf{R}^n |\; Ay \leq b\}$, which is assumed to be non-empty.

1. *Write down the Lagrange dual to the optimization problem (P).*
2. *Verify that f is convex.*
3. *Compute the subdifferential of f at x.*

Exercise 4.4.7 *Prove the following Theorem of Karhu-Bonnenblast (which is very close to the Minmax Lemma):*

Let $X \subset \mathbf{R}^k$ be a convex set and $f_1, ..., f_m$ be real-valued convex functions on X. Prove that
- either the system of strict inequalities

$$(*) \quad f_i(u) < 0,\ i = 1, ..., m,$$

has a solution in X,
- or there exist nonzero μ_i with the unit sum such that the function

$$\sum_{i=1}^{m} \mu_i f_i(u)$$

is nonnegative for all $u \in X$.

Hint: follow the proof of the Minmax Lemma

Optional problems

The below exercise presents statistical application of the Kullback distance (Exercise 4.4.4)

Exercise 4.4.8 Consider the situation as follows. You observe a "signal" s – a single random output of certain statistical experiment, and you know in advance that the output belongs to a given finite set S of all possible outputs. E.g. you look at the radar screen and see certain mark – here or there, this or that bright. You know in advance that the there are two possibile distributions, u_s^1 and u_s^2, of the output, given, respectively, by two statistical hypotheses H_1 ("the mark comes from the target") and H_2 ("false mark due to noises"). Your goal is to decide, given the observation, which of these hypotheses is valid. The only deterministic policy you could use is as follows: partition in advance the set of all tentative outputs S into two non-overlapping subsets S_1 and S_2, and check whether the observed s belongs to S_1 or to S_2; in the first case you claim that the signal comes from the hypothesis H_1, in the second – that it comes from H_2. If every output can be obtained, with positive probability, from both the hypotheses, you clearly cannot find a 100% reliable decision rule ($\equiv$ partitioning $S = S_1 \cup S_2$); what we can speak about are the probabilities of errors

$$\pi_1 = \sum_{s \in S_2} u_s^1$$

("missing" probability – the mark came from the target and you decided that it came from noise), and

$$\pi_2 = \sum_{s \in S_1} u_s^2$$

("false alarm" probability – you decided that the mark came from the target when it came from noise).

We would be interested to make both p_1 and p_2 small. A very natural (and an extremely important) question here is what are the lower bounds on the errors. A simple answer is given by the following *necessary* condition (which in many cases turns out to be "sharp"):

if α, β are two given positive reals less than 1/2 each, then a decision rule which ensures $\pi_1 \leq \alpha$ and $\pi_2 \leq \beta$ exists only if the Kullback distance from the distribution u^1 to the distribution u^2 is not too small, namely,

$$I(u^1, u^2) \geq (1-\alpha)\ln\left(\frac{1-\alpha}{\beta}\right) + \alpha\ln\left(\frac{\alpha}{1-\beta}\right).$$

The exercise is to prove the necessity of the indicated condition.

Remark 4.4.1 Both the Kullback distance and the indicated necessary condition have continuous analogies. Namely, the Kullback distance between two distributions on $\mathbf{R}^l$ with densities, respectively, $v^1(x)$, $v^2(x)$, is defined as

$$I(v_1, v_2) = \int v_1(x)\ln(v_1(x)/v_2(x))dx$$

("directed distance" from v_1 to v_2), and with this definition of the distance the indicated necessary condition remains valid for continuous random signals.

Note that for Gaussian distributions v_1 and v_2 with the unit covariance matrix and the mean values m_1, m_2 the Kullback distance is simply

$$|m_1 - m_2|_2^2.$$