Optimization and Computational Linear Algebra for Data Science
Lecture 11: Optimality conditions

Léo Miolane · leo.miolane@gmail.com

November 12, 2020

Warning: *This material is not meant to be lecture notes. It only gathers the main concepts and results from the lecture, without any additional explanation, motivation, examples, figures...*

1 Local and global minimizers

We aim at minimizing a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We say that $x \in \mathbb{R}^n$ is

- a critical point of f if $\nabla f(x) = 0$,
- a *global* minimizer of f if for all $x' \in \mathbb{R}^n$, $f(x) \leq f(x')$,
- a *local* minimizer of f if there exists $\delta > 0$ such that for all $x' \in B(x, \delta)$, $f(x) \leq f(x')$,

where $B(x, \delta) = \{x' \,|\, \|x' - x\| \leq \delta\}$ denotes the ball a radius δ centered at x. Of course, a global minimizer if also a local minimizer but the converse is not true.

Proposition 1.1

Let $x \in \mathbb{R}^n$ be a point at which f is differentiable. Then

$$x \text{ is a local minimizer of } f \implies \nabla f(x) = 0.$$

As we saw in Lecture 9, if f is convex then the converse is true:

Proposition 1.2

Assume that f is convex. Let $x \in \mathbb{R}^n$ be a point at which f is differentiable. Then

$$\nabla f(x) = 0 \iff x \text{ is a global minimizer of } f.$$

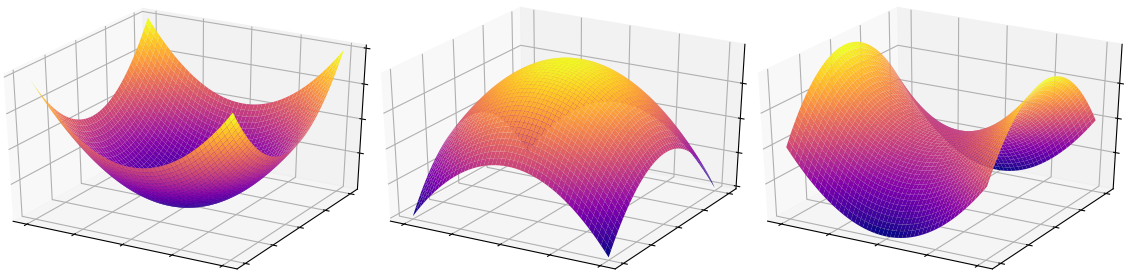


Figure 1: Left: minimum. Middle: maximum. Right: saddle point.

Proposition 1.3

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Let $x \in \mathbb{R}^n$ be a critical point of f, i.e. $\nabla f(x) = 0$. Then

- *If $H_f(x)$ is positive definite (that is, if all the eigenvalues of $H_f(x)$ are strictly positive), then x is a local minimizer of f.*
- *If $H_f(x)$ is negative definite (that is, if all the eigenvalues of $H_f(x)$ are strictly negative), then x is a local maximizer of f.*
- *If $H_f(x)$ admits strictly positive eigenvalues and strictly negative eigenvalues, then x is neither a local maximum nor a local minimum. We call x a saddle point.*

2 Constrained optimization

We would now like to investigate constrained optimization problems:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p, \end{array} \tag{1}$$

with variable $x \in \mathbb{R}^n$. Here we have m inequality constraints $g_1(x) \leq 0, \dots, g_m(x) \leq 0$ and p equality constraints $h_1(x) = 0, \dots, h_p(x) = 0$ to satisfy. We denote by $p^* \in \mathbb{R} \cup \{\pm\infty\}$ the optimal value of (1).

Definition 2.1 (*Feasible point*)

A point $x \in \mathbb{R}^n$ is feasible *if it satisfies all the constraints: $g_1(x) \leq 0, \dots, g_m(x) \leq 0$ and $h_1(x) = 0, \dots, h_p(x) = 0$. We will denote by F the set of feasible points. F is called the feasible set.*

Definition 2.2 (*Active constraints*)

We say that the inequality constraint $g_i(x) \leq 0$ is active at x if $g_i(x) = 0$. We let $\mathcal{A}(x)$ be the set of active constraints at x: $\mathcal{A}(x) = \{i \,|\, g_i(x) = 0\}$.

We would now get for the problem (1) the analog of Proposition 1.1. Since an equality constraint $h_i(x) = 0$ can be equivalently written in two inequality constraints $h_i(x) \leq 0$ and $-h_i(x) \leq 0$, we can assume to have only inequality constraints. For simplicity, we first assume to have only one inequality constraint $g(x) \leq 0$ so that (1) reduces to

$$\text{minimize } f(x) \text{ subject to } g(x) \leq 0. \tag{2}$$

Let x be a solution of (2), i.e. $g(x) \leq 0$ and $f(x) \leq f(x')$ for all x' such that $g(x') \leq 0$ We distinguish two cases:

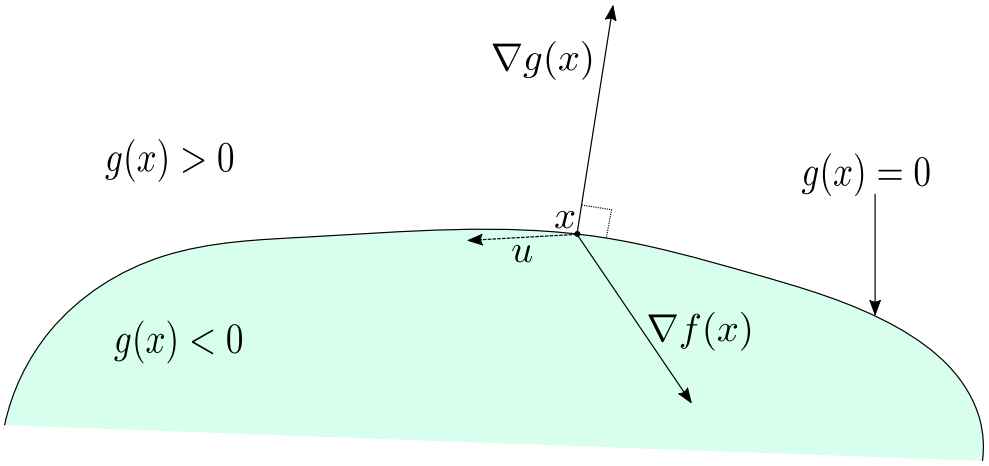
Case 1: the constraint is not active at x; $g(x) < 0$. In that case x is in the interior of F: one can find $\delta > 0$ such that $B(x, \delta) \subset F$. Since x is a solution of (2) we have for all $x' \in B(x, \delta)$, $f(x) \leq f(x')$. One can therefore apply Proposition 1.1 to get that $\nabla f(x) = 0$.

We conclude that in the case where the constraint is not active, the constraint does not play any role and one gets the same optimality condition as in the unconstrained setting.

Case 2: the constraint is active at x; $g(x) = 0$. In that case, there exists $\lambda \geq 0$ such that

$$\nabla f(x) = -\lambda \nabla g(x). \tag{3}$$

To see that, assume that (3) does not hold. Then we are in the following situation:



As we can see on the figure, we can find a vector u such that

$$\langle u, \nabla g(x) \rangle < 0 \quad \text{and} \quad \langle u, \nabla f(x) \rangle < 0.$$

Starting from x and following the direction u one remains in the feasible set because for small $\delta > 0$

$$g(x + \delta u) \simeq g(x) + \delta \langle u, \nabla g(x) \rangle \leq 0.$$

Moreover, f decreases locally on the direction u:

$$f(x + \delta u) \simeq f(x) + \delta \langle u, \nabla f(x) \rangle < f(x).$$

This means that one can find $\delta > 0$ such that $x + \delta u$ is feasible and such that $f(x + \delta u) < f(x)$. This contradicts the assumption that x is solution of (2). We conclude that (3) holds, i.e. that there exists $\lambda \geq 0$ such that

$$\nabla f(x) + \lambda \nabla g(x) = 0.$$

This can be generalized to the case (1) where we have multiple constraints.

Definition 2.3 (*LICQ*)

Let x be a feasible point. We say that the linear independence constraint qualification *(LICQ) holds at x if the set of gradients*

$$\{\nabla g_i(x) \,|\, i \in \mathcal{A}(x)\} \cup \{\nabla h_i(x) \,|\, i \in \{1, \dots, p\}\}$$

is linearly independent.

Theorem 2.1 (*First-order optimality conditions*)

Assume that the functions f, $g_1, \dots, g_m, h_1, \dots, h_p$ are continuously differentiable. If x is solution of (1) and if LICQ holds at x then there exists $\lambda_1, \dots, \lambda_m \geq 0$ and $\nu_1, \dots, \nu_p \in \mathbb{R}$ such that:

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0. \tag{4}$$

Moreover, for all $i \in \{1, \dots, m\}$, $\lambda_i = 0$ if $g_i(x) < 0$.

The scalars λ_i, ν_i are called *Lagrange multipliers*. The first-order conditions of Theorem 2.1 are often called the Karush-Kuhn-Tucker (KKT) conditions and Theorem 2.1 is presented the following way. If x is a solution of (1) then there exists numbers λ_i, ν_i such that:

(i) *Primal feasibility*: $g_i(x) \leq 0$ for $i = 1, \ldots, m$ and $h_i(x) = 0$ for $i = 1, \ldots, p$.

(ii) *Dual feasibility*: $\lambda_i \geq 0$ for $i = 1, \ldots, m$.

(iii) *Stationarity*: $\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$.

(iv) *Complementary slackness*: $\lambda_i g_i(x) = 0$ for $i = 1, \ldots, m$.

The terms "primal" and "dual" will be explained in the next section.

3 The Lagrangian and the dual problem

We define the Lagrangian L associated with the problem (1) by

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{p} \nu_i h_i(x), \tag{5}$$

where $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m_{\geq 0}$ and $\nu \in \mathbb{R}^p$. We define the Lagrange dual function by

$$\ell(\lambda, \nu) = \inf_{x \in \mathbb{R}^n} L(x, \lambda, \nu).$$

Notice that for all feasible point x,

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \leq f(x)$$

because $h_i(x) = 0$ and $\lambda_i g_i(x) \leq 0$. By taking the infimum in x on both sides of the inequality we get a lower bound on the value of the optimization problem (1):

Proposition 3.1

For all $\lambda_1, \ldots, \lambda_m \geq 0$ and all $\nu_1, \ldots, \nu_p \in \mathbb{R}$ we have:

$$\ell(\lambda, \nu) \leq p^*. \tag{6}$$

3.1 Dual problem

We would like to make the lower bound (6) as tight as possible: one would like therefore to solve the so-called *dual problem*:

$$\begin{array}{lll} \text{maximize} & \ell(\lambda, \nu) & \\ \text{subject to} & \lambda_i \geq 0, & i = 1, \ldots, m \\ & \nu_i \in \mathbb{R}, & i = 1, \ldots, p. \end{array} \tag{7}$$

Notice that the Lagrange dual function is always concave, as an infimum of affine functions. Hence, the dual problem might be easier to solve than the original problem.

We denote by d^* the optimal value of the dual problem (7). From (6) we deduce that the optimal value of the primal problem is greater or equal than the one of the dual problem:

$$d^* \leq p^*. \tag{8}$$

This is known as *weak duality*. Notice that $p^* = \inf_{x\in\mathbb{R}^n} F(x)$ where

$$F(x) \stackrel{\text{def}}{=} \sup_{\lambda\geq 0,\nu} L(x,\lambda,\nu) = \begin{cases} f(x) & \text{if } x \text{ is feasible,} \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, the weak duality inequality can be rewritten as:

$$\sup_{\lambda\geq 0,\nu} \inf_{x\in\mathbb{R}^n} L(x,\lambda,\nu) \leq \inf_{x\in\mathbb{R}^n} \sup_{\lambda\geq 0,\nu} L(x,\lambda,\nu). \tag{9}$$

When there is equality in (8) (or equivalently in (9)) we say that there is *strong duality*. We will see in Section 4 that strong duality holds for convex problems under mild assumptions.

3.2 Saddle-points

Definition 3.1 (*Saddle-point*)

We say that $(x;\lambda,\nu) \in \mathbb{R}^n \times \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$ *is a* saddle-point[1] *of* L *if*

$$\forall (\lambda',\nu') \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p \quad L(x,\lambda',\nu') \leq L(x,\lambda,\nu) \leq L(x',\lambda,\nu) \quad \forall x' \in \mathbb{R}^n. \tag{10}$$

Notice that if $(x;\lambda,\nu)$ is a saddle-point of L, then x is solution of (1). Indeed, by taking the supremum over (λ',ν') in (10) we get:

$$F(x) \leq L(x,\lambda,\nu) \leq L(x',\lambda,\nu) \leq F(x')$$

for all $x' \in \mathbb{R}^n$. This gives that $F(x) = \min F = L(x,\lambda,\nu)$. By a "symmetric" argument one also gets that (λ,ν) is a solution of the dual problem (7) and $\ell(\lambda,\nu) = \max_{\lambda\geq 0,\nu} \ell(\lambda',\nu') = L(x,\lambda,\nu)$.

We conclude that if (x,λ,ν) is a saddle-point of L, then x is primal optimal, that (λ,ν) is dual optimal and that strong duality holds. The next Theorem shows that the converse is true.

Theorem 3.1

Let $(x;\lambda,\nu) \in \mathbb{R}^n \times \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. *Then*

$$(x;\lambda,\nu) \text{ is a saddle-point of } L \iff \begin{cases} x \text{ is a solution of the primal problem (1),} \\ (\lambda,\nu) \text{ is a solution of the dual problem (7),} \\ \text{strong duality holds.} \end{cases}$$

Proof. Assume that x is a solution of the primal problem (1), that (λ,ν) is a solution of the dual problem (7) and strong duality holds. We have

$$\ell(\lambda,\nu) = \inf_{x'} L(x',\lambda,\nu) \leq L(x,\lambda,\nu) \leq \sup_{\lambda'\geq 0,\nu'} L(x,\lambda',\nu') = F(x).$$

Since there is strong duality, we have $\ell(\lambda,\nu) = F(x)$: hence the inequalities above are in fact equalities. This implies that $(x;\lambda,\nu)$ is a saddle-point of L. □

3.3 Solving the primal problem via the dual

Assume that strong duality holds and that we have found a solution (λ^*,ν^*) of the dual problem. By Theorem 3.1 we get that a point $x^* \in \mathbb{R}^n$ is a solution of the primal problem if and only if $(x^*;\lambda^*,\nu^*)$ is a saddle-point of L:

$$x^* \text{ is solution of the primal problem} \iff \begin{cases} x^* & \text{minimizes} & L(\cdot,\lambda^*,\nu^*), \\ (\lambda^*,\nu^*) & \text{maximizes} & L(x^*,\cdot,\cdot). \end{cases}$$

[1] Unfortunately, "saddle-point" has multiple meanings in maths: the one of Section 1 and the one of this definition.

Notice that (λ^*, ν^*) maximizes the affine function $L(x^*, \cdot, \cdot)$ if and only if x^* is feasible. We obtain:

$$x^* \quad \text{is solution of the primal problem} \quad \Longleftrightarrow \quad \begin{cases} x^* & \text{minimizes} \quad L(\cdot, \lambda^*, \nu^*), \\ x^* & \text{is feasible.} \end{cases}$$

This is particularly useful, because the dual problem might be easier to solve than the primal one. In the case when there is strong duality, the equivalence above tells us that it suffices then to solve the **unconstrained** optimization problem $\min_{x \in \mathbb{R}^n} L(x, \lambda^*, \nu^*)$. Assume for simplicity that $x \mapsto L(x, \lambda^*, \nu^*)$ admits a unique minimizer x^*. Then if x^* is feasible then it is a solution of the problem (1). If x^* is not feasible then the minimum of (1) can not be attained.

4 Kuhn Tucker Theorem

In this section, we assume that the functions f, $g_1, \dots, g_m$ are **convex** and that $h_1, \dots, h_p$ are **affine**, that is $h_i(x) = \langle a_i, x \rangle + b_i$ for some $a_i \in \mathbb{R}^n$, $b_i \in \mathbb{R}$. We say then that the optimization problem (1) is convex. Notice in that case that the set of feasible points is convex, as the intersection of the convex sets $\{x | g_i(x) \leq 0\}$ and $\{x | h_i(x) = 0\}$.

We will see in this section that **strong duality holds for convex problems under mild assumptions**, known as "Slater's condition".

Definition 4.1 (*Slater's condition*)

We say that the problem (1) verifies Slater's condition if there exists a feasible point x such that $g_i(x) < 0$ for all $i \in \{1, \dots, m\}$.

Proposition 4.1

If the problem (1) is convex and verifies Slater's condition, then strong duality holds. Moreover if $p^* = d^*$ is finite then the optimal value of the dual problem is attained at some $(\lambda, \nu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$.

We refer to Section 5.2.3 of [1] for a proof of Proposition 4.1.

Theorem 4.1 (*Kuhn Tucker*)

*Assume that the functions f, $g_1, \dots, g_m$ are **convex**, differentiable and that $h_1, \dots, h_p$ are affine. Assume that strong duality holds, that $p^* = d^*$ is finite, and that the optimal value of the dual problem is attained at some $(\lambda, \nu) \in \mathbb{R}^m_{\geq 0} \times \mathbb{R}^p$. (This is for instance the case under Slater's condition).*

Then $x \in \mathbb{R}^n$ is solution of (1) if and only if x is feasible and

$$\begin{cases} \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla g_i(x) + \sum_{i=1}^p \nu_i \nabla h_i(x) = 0 \\ \lambda_i g_i(x) = 0 \quad \text{for all} \quad i = 1, \dots, m. \end{cases} \tag{11}$$

In other words, a feasible point x is a solution of (1) if and only if $(x; \lambda, \nu)$ is a saddle-point of the Lagrangian L.

Such couple (λ, ν) is sometimes called a "dual certificate", since it certifies the optimality of x.

Proof of Theorem 4.1. Let x be a feasible point. By Theorem 3.1 we have

$$x \text{ solution of (1)} \Longleftrightarrow (x; \lambda, \nu) \text{ is a saddle-point of the Lagrangian } L.$$

We assumed that the problem (1) is convex. The function $L(\cdot, \lambda, \nu)$ is thus convex and the function $L(x, \cdot, \cdot)$ is concave (in fact it is affine). Hence

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0 \quad \iff \quad x \text{ minimizes } L(\cdot, \lambda, \nu)$$
$$\lambda_i g_i(x) = 0 \quad \text{for all} \quad i = 1, \ldots, m \quad \iff \quad (\lambda, \nu) \text{ maximizes } L(x, \cdot, \cdot).$$

We get that (11) $\iff$ $(x; \lambda, \nu)$ is a saddle-point of L, which concludes the proof. □

Further reading

See Chapter 12 from [2] for a proof of Theorem 2.1. See in particular section 12.6 for constraint qualifications that are more general than LICQ. See Chapters 4 and 5 of [1] a more detailed introduction to convex optimization problems and duality.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, https://web.stanford.edu/~boyd/cvxbook/, 2004.

[2] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.