Optimization Models

EECS 127 / EECS 227AT

Laurent El Ghaoui

EECS department
UC Berkeley

Fall 2018

LECTURE 20

Optimality Conditions

Duality, in mathematics, principle whereby one true statement can be obtained from another by merely interchanging two words.

Britannica

Outline

Overview

In this lecture, we describe the so-called "optimality conditions" that characterize optimality for convex programs, and generalize the "zero-gradient" condition that arises in convex unconstrained problems.

Theese conditions have many uses, in particular in

- the theoretical analysis of solutions to convex problems;
- the design of convex optimization algorithms.

We will first look at an "abstract" form of optimality conditions that offer geometric insight and work well for equality constraints only; then develop optimality conditions for the general case.

Primal problem

In this lecture, we consider the following "primal" problem

$$p^* = \min_{x \in \mathbb{R}^n} f_0(x) \text{ subject to: } \quad f_i(x) \le 0, \quad i = 1, \ldots, m,$$
$$Ax = b,$$

where

- $f_0, \ldots, f_m$ are convex differentiable functions, which we assume to be defined everywhere (hence the domain of the problem is $\mathcal{D} = \mathbb{R}^n$);
- matrix $A \in \mathbb{R}^{q \times n}$ and vector $b \in \mathbb{R}^q$ are given.

We denote by $\mathcal{D}$ the domain of the problem: $\mathcal{D} \doteq \bigcap_{i=0}^m \mathrm{dom}\, f_i$.

We make a few assumptions on the above problem:

- it is strictly feasible (so that Slater's condition holds);
- it is attained: there exist $x^* \in \mathcal{D}$ such that $p^* = f_0(x^*)$.

Abstract form of optimality conditions

The primal problem can be written in abstract form

$$\min_{x \in \mathcal{X}} \ f_0(x),$$

where $\mathcal{X} \subseteq \mathcal{D}$ denotes the feasible set.

Proposition 1

Consider the optimization problem $\min_{x \in \mathcal{X}} f_0(x)$*, where* f_0 *is convex and differentiable, and* $\mathcal{X}$ *is convex. Then,*

$$x \in \mathcal{X} \text{ is optimal} \quad \Leftrightarrow \quad \nabla f_0(x)^\top (y - x) \geq 0, \quad \forall y \in \mathcal{X}. \tag{1}$$

Note: the above conditions are often hard to work with, due to the presence of the "$\forall\, y \ldots$" statement, which requires checking a condition over the entire feasible set.

Proof

First let us show the implication from right to left in (1). Since f_0 is convex, for every $x, y \in \mathrm{dom}\, f_0$, we have

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^\top (y - x). \tag{2}$$

The implication from right to left in (1) is immediate, since

$$\nabla f_0(x)^\top (y - x) \geq 0 \text{ for every } y \in \mathcal{X}$$

implies, from (2), that $f_0(y) \geq f_0(x)$ for all $y \in \mathcal{X}$, i.e., that x is optimal.

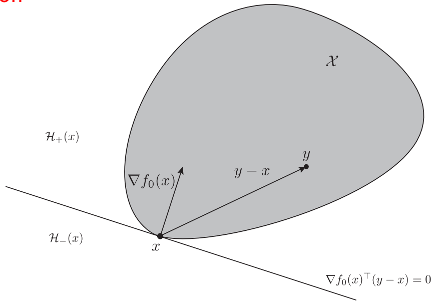
Conversely, assume that x is optimal. We show that then $\nabla f_0(x)^\top (y - x) \geq 0$ for all $y \in \mathcal{X}$. If $\nabla f_0(x) = 0$, then the claim holds trivially. Assume now that $\nabla f_0(x) \neq 0$, and that there exist $y \in \mathcal{X}$ such that $\nabla f_0(x)^\top (y - x) < 0$. Consider the function

$$g \;:\; t \in [0, 1] \to f_0(x(t)),$$

where $x(t) = ty + (1 - t)x$; note that $x(t) \in \mathcal{X}$ for every $t \in [0, 1]$, since $\mathcal{X}$ is convex. Further, $g'(0) = \nabla f_0(x)^\top (y - x)$. Hence, for sufficiently small $t > 0$, $g(t) < g(0)$, which translates as $f(x(t)) < f(x)$; with $x(t) \in \mathcal{X}$, this contradicts the optimality of x. $\square$

Optimality conditions

Geometric interpretation



If $\nabla f_0(x) \neq 0$, then $\nabla f_0(x)$ is a normal direction defining an hyperplane $\{y : \nabla f_0(x)^\top (y - x) = 0\}$ such that:

- x is on the boundary of the feasible set $\mathcal{X}$, and
- the whole feasible set lies on one side of this hyperplane, that is in the halfspace defined by

$$\mathcal{H}_+(x) = \{y : \nabla f_0(x)^\top (y - x) \geq 0\}.$$

Optimality conditions

Geometric interpretation

Notice that the gradient vector $\nabla f_0(x)$ defines two set of directions:

- for directions v_+ such that $\nabla f_0(x)^\top v_+ > 0$ (i.e., directions that have positive inner product with the gradient), if we make a move away from x in direction v_+, then the objective f_0 *increases*.
- for directions v_- such that $\nabla f_0(x)^\top v_+ < 0$ (i.e., *descent* directions, that have negative inner product with the gradient), if we make a sufficiently small move away from x in direction v_-, then the objective f_0 locally *decreases*.

Condition (1) then says that x is an optimal point if and only if there is no feasible direction along which we may improve (decrease) the objective.

Optimality conditions for unconstrained problems

Proposition 2

In a convex unconstrained problem with differentiable objective, x is optimal if and only if

$$\nabla f_0(x) = 0. \tag{3}$$

Proof: When the problem is unconstrained, i.e., $\mathcal{X} = \mathbb{R}^n$, then the optimality condition (1) requires that

$$\begin{aligned} \forall\, y \in \mathbb{R}^n \ : \ \nabla f_0(x)^\top (y - x) \geq 0 \quad &\iff \quad \forall\, z \in \mathbb{R}^n \ : \ \nabla f_0(x)^\top z \geq 0 \\ &\iff \quad \forall\, z \in \mathbb{R}^n \ : \ \nabla f_0(x)^\top z = 0 \\ &\iff \quad \nabla f_0(x) = 0. \end{aligned}$$

Optimality conditions for equality-constrained problems

Consider the problem

$$\min_x f_0(x) \ : \ Ax = b, \tag{4}$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ are given. We assume that $b \in \mathcal{R}(A)$, so the problem is feasible. Here the feasible set is

$$\mathcal{X} = \{y \ : \ Ay = b\}.$$

Proposition 3

A point x is optimal for problem (4) if and only if

$$Ax = b \text{ and } \exists\, \nu \in \mathbb{R}^m \ : \ \nabla f_0(x) + A^\top \nu = 0.$$

Proof

The point $x \in \mathcal{X}$ is optimal iff

$$\nabla f_0(x)^\top (y - x) \geq 0, \quad \forall y \in \mathcal{X}.$$

Since $Ax = b$, the feasible set can be written as

$$\mathcal{X} = \{x + z \ : \ z \in \mathcal{N}(A)\}.$$

The optimality condition becomes

$$\forall\, z \in \mathcal{N}(A) \ : \ \nabla f_0(x)^\top z \geq 0.$$

Since $z \in \mathcal{N}(A)$ if and only if $-z \in \mathcal{N}(A)$, we see that the condition is equivalent to

$$\forall\, z \in \mathcal{N}(A) \ : \ \nabla f_0(x)^\top z = 0.$$

That is, $\nabla f_0(x) \in \mathcal{N}(A)^\perp$. Recall the fundamental theorem of linear algebra, which states that $\mathcal{N}(A)^\perp = \mathcal{R}(A^\top)$; we obtain that there exist $\nu \in \mathbb{R}^m$ such that $\nabla f_0(x) + A^\top \nu = 0$.

Example

Minimum-norm solutions to linear equations

Consider the Euclidean projection problem seen in lecture 8:

$$\min_x \ \frac{1}{2}x^\top x \ : \ Ax = b.$$

(The solution is the projection of 0 on the affine subspace $\mathcal{X}$.)

We obtain that x is optimal if and only if there exist $\nu \in \mathbb{R}^m$ such that

$$Ax = b, \ \ x + A^\top \nu = 0. \tag{5}$$

Assuming that A is full row rank (hence, $AA^\top \succ 0$), we get the unique solution:

$$\nu^* = -(AA^\top)^{-1}b, \ \ x^* = -A^\top \nu^* = A^\top (AA^\top)^{-1}b.$$

General case

Dual problem

Turning to the general problem (1), recall the expression of the problem dual to (1), as seen in lecture 18:

$$d^* = \max_{\lambda \geq 0} g(\lambda), \tag{6}$$

where g is the dual function

$$g(\lambda) = \min_x \mathcal{L}(x, \lambda, \nu),$$

with $\mathcal{L}$ the Lagrangian

$$\mathcal{L}(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Since Slater's condition hold, we have strong duality: $p^* = d^*$.
- We make the further assumption that d^* is attained by some $\lambda^* \geq 0$.

Karush-Kuhn-Tucker (KKT) conditions

For the convex problem (1), we say that a pair $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfies the Karush-Kuhn-Tucker (KKT) conditions if

1. Primal feasibility: x is feasible for the primal problem:

$$x \in \mathcal{D}, \quad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

2. Dual feasibility: $\lambda \geq 0$.
3. Complementary slackness: $\lambda_i f_i(x) = 0$, $i = 1, \ldots, m$.
4. Lagrangian stationarity: $x \in \arg\min \mathcal{L}(\cdot, \lambda)$, which, in the case when the functions f_i, $i = 0, \ldots, m$ are differentiable, writes

$$\nabla_x f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla_x f_i(x) = 0.$$

Proposition 4

Assume that the primal problem (1) is convex, and attained; that its dual is also attained; and that strong duality holds. Then, a primal-dual pair (x, λ) is optimal if and only if it satisfies the KKT conditions.

Proof: sufficiency

Assume that the KKT conditions are satisfied for some pair (x^*, λ^*). The first two conditions imply that x^* is primal feasible, and λ^* is dual feasible. Further, since $\mathcal{L}(x, \lambda^*)$ is convex in x, the fourth condition states that x^* is a global minimizer of $\mathcal{L}(x, \lambda^*)$, hence

$$\begin{aligned} g(\lambda^*, \nu^*) &= \min_{x \in \mathcal{D}} \mathcal{L}(x, \lambda^*) = \mathcal{L}(x^*, \lambda^*) \\ &= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \\ &= f_0(x^*), \end{aligned}$$

where the last equality follows from complementary slackness.

The above proves that the primal-dual feasible pair (x^*, λ^*) is optimal: the corresponding duality gap $p^* - d^*$ is zero, since x^* (resp. λ^*) attains the lower bound d^* (resp. upper bound p^*).

Proof: necessity

Assume that (x^*, λ^*) is an optimal primal-dual pair.

- Since $p^* = f_0(x^*)$, $d^* = g(\lambda^*)$, and $p^* = d^*$, we have

$$f_0(x^*) = g(\lambda^*) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x, \lambda^*), \quad \forall\, x \in \mathcal{D}.$$

- Since the last inequality holds for all $x \in \mathcal{D}$, it must hold also for x^*, hence

$$f_0(x^*) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \leq f_0(x^*),$$

  where the last inequality follows from the fact that x^* is optimal, hence feasible, for the primal problem, therefore $f_i(x^*) \leq 0$, and λ^* is optimal, hence feasible, for the dual, therefore $\lambda_i^* \geq 0$, whereby each term $\lambda_i^* f_i(x^*)$ is ≤ 0.

- Observing the last chain of inequalities, since the first and the last terms are equal, we must conclude that all inequalities must actually hold with equality, that is

$$f_0(x^*) = \inf_{x \in \mathcal{D}} \mathcal{L}(x, \lambda^*) = \mathcal{L}(x^*, \lambda^*).$$

Complementary slackness and Lagrangian stationarity

These two conditions are at the heart of the KKT conditions.

The complementary slackness property prescribes that a primal and the corresponding dual inequality cannot be slack simultaneously, that is, if $f_i(x^*) < 0$, then it must be $\lambda_i^* = 0$, and if $\lambda_i^* > 0$, then it must be $f_i(x^*) = 0$.

The second property (i.e., the fact that x^* is a minimizer of $\mathcal{L}(x, \lambda^*)$) can, in some cases, be used to recover a primal-optimal variable from the dual-optimal variables (see later).

Recovering primal solutions from the dual

- First observe that if the primal problem is convex, then $\mathcal{L}(x, \lambda^*)$ is also convex in x. Global minimizers of this function can then be determined by unconstrained minimization techniques. For instance, if $\mathcal{L}(x, \lambda^*)$ is differentiable, a necessary condition for x to be a global minimizer is determined by the zero-gradient condition $\nabla_x \mathcal{L}(x, \lambda^*) = 0$, that is,

$$\nabla_x f_0(x) + \sum_{i=1}^{m} \lambda_i^* \nabla_x f_i(x) = 0.$$

- However, $\mathcal{L}(x, \lambda^*)$ may have multiple global minimizers, and it is *not* guaranteed that *every* global minimizer of $\mathcal{L}$ is a primal-optimal solution—what is guaranteed is that the primal-optimal solution x^* is among the global minimizers of $\mathcal{L}(\cdot, \lambda^*)$.
- A particular case arises when $\mathcal{L}(\cdot, \lambda^*)$ has an *unique* minimizer. In this case the unique minimizer x^* of $\mathcal{L}$ is either primal feasible, and hence it is the primal-optimal solution, or it is not primal feasible, and then we can conclude that the no primal-optimal solution exists.

Example

Power allocation in a communication channel[1]

We seek to best allocate a power level to n communication channels. The problem can be formulated as

$$p^* = \min_x -\sum_{i=1}^{n} \log(\alpha_i + x_i) \ : \ x \geq 0, \ \sum_{i=1}^{m} x_i = 1.$$

where $\alpha_i > 0$ is a measure of the noise over the channel. Here the objective function is related to the communication rate. We use the Lagrangian

$$\mathcal{L}(x, \lambda, \nu) = -\sum_{i=1}^{n} \log(\alpha_i + x_i) - \lambda^\top x + \nu(\sum_{i=1}^{m} x_i - 1),$$

with $\lambda \in \mathbb{R}_+^n$, $\nu \in \mathbb{R}$.

[1] From Boyd & Vandenberghe's book, *Convex Optimization*.

KKT conditions

Slater's conditions are satisfied. The KKT conditions are:

- Primal feasibility: $x \geq 0$ and $\mathbf{1}^\top x = 1$;
- Dual feasibility: $\lambda \geq 0$;
- Stationarity: $\lambda_i + 1/(x_i + \alpha_i) = \nu$, $i = 1, \ldots, n$.
- Complementarity: $\lambda_i x_i = 0$, $i = 1, \ldots, n$.

For an optimal pair (x^*, λ^*, ν^*):

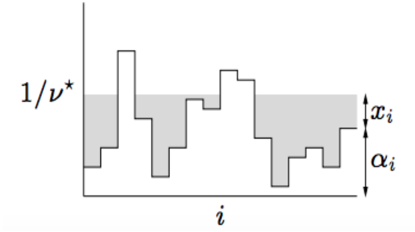
- if $\nu^* \leq 1/\alpha_i$, then $0 \leq \lambda_i^* \leq 1/\alpha_i - 1/(x_i^* + \alpha_i) = x_i/(\alpha_i(\alpha_i + x_i))$. If $\lambda_i^* > 0$, then $x_i^* = 0$ from the complementarity conditions; this yields a contradiction. Hence $\lambda_i = 0$ and $x_i^* = 1/\nu^* - \alpha_i (\geq 0)$ in that case.
- otherwise, $\nu^* > 1/\alpha_i$; this leads to $\lambda_i^* + 1/(x_i^* + \alpha_i) > 1/\alpha_i$. Again, assuming $x_i^* > 0$ leads to $\lambda_i = 0$ and a contradiction; hence $x_i^* = 0$ in that case.

We have obtained $x_i^* = \max(0, 1/\nu^* - \alpha_i)$ for every i. Summing, we obtain a condition that characterizes ν^*:

$$1 = \sum_{i=1}^{n} x_i^* = \sum_{i=1}^{n} \max(0, 1/\nu^* - \alpha_i).$$

Waterfilling algorithm

We can solve this 1D equation using a simple method called the waterfilling algorithm. Once ν^* is found, we then recover a primal optimal point via $x_i^* = \max(0, 1/\nu^* - \alpha_i)$, $i = 1, \ldots, n$.



The height of patch i is given by α_i. The region is flooded to a level $1/\nu$, using a total quantity of water equal to one. The height of the water (shown shaded) above each patch is the optimal value of x_i.

Example

Maximum entropy distribution

Consider the problem

$$\min_x \; f_0(x) \doteq \sum_{i=1}^{n} x_i \log x_i \;\; : \;\; x \geq 0, \;\; \mathbf{1}^\top x = 1.$$

The feasible set is the set of discrete distributions in $\mathbb{R}^n$; The objective function is called the negative entropy of the distribution x.

- Lagrangian: $\mathcal{L}(x, \lambda, \nu) = f_0(x) - \lambda^\top x + \nu(1 - \mathbf{1}^\top x)$.
- KKT conditions: $x \geq 0$, $\mathbf{1}^\top x = 1$, $\lambda \geq 0$, and

$$\lambda_i x_i = 0, \;\; \log x_i = \lambda_i + \nu - 1, \;\; i = 1, \ldots, n.$$

The stationarity conditions imply that $x^* > 0$, hence $\lambda^* = 0$, and thus x_i does not depend on i. Since $\mathbf{1}^\top x = 1$, we obtain that $x^* = (1/n)\mathbf{1}$, which is the uniform distribution.

This fact illustrates why the (negative) entropy function is used as a measure of "distance" between a distribution, to the uniform one.

Example

Risk parity portfolio

Consider a portfolio optimization problem: to find a portfolio weight vector $x \in \mathbb{R}^n_{++}$, containing positive dollar amounts to invest in various assets, such that the risk parity condition holds:

$$\forall\, i \;:\; x_i(Cx)_i = \frac{1}{n} x^\top Cx,$$

where $C = C^\top \succ 0$ is the (positive-definite) covariance of the assets. The interpretation of a risk-parity portfolio is that, since

$$\sum_{i=1}^{n} x_i(Cx)_i = x^\top Cx,$$

all the partial contributions $x_i(Cx)_i (> 0)$ of each asset i to the total risk in the portfolio, as measured by its variance $x^\top Cx$, are equal ("at parity").

Risk parity portfolio

Consider the optimization problem

$$\min_x \; f_0(x) + x^\top C x, \tag{7}$$

where

$$f_0(x) \doteq \begin{cases} -\sum_{i=1}^n \log x_i & \text{if } x > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Lagrangian:

$$\mathcal{L}(x, \lambda) = -\sum_{i=1}^n \log x_i + x^\top C x - \lambda^\top x.$$

KKT conditions: $x > 0$ (since $\mathcal{D} = \mathbb{R}^n_{++}$), $\lambda \geq 0$,

$$\lambda_i x_i = 0, \quad -\frac{1}{x_i} + (Cx)_i = \lambda_i, \quad i = 1, \ldots, n.$$

Since $x > 0$, we have $\lambda = 0$, and we obtain $x_i(Cx)_i = 1$, $i = 1, \ldots, n$; summing, we get $x^\top C x = n$, which implies that the risk parity conditions hold.

This means that by solving the convex problem (7), we obtain a risk parity portfolio.