Optimization - Optimality Conditions

Qamrul Hasan Ansari
Aligarh Muslim University, India
KFUPM, Dhahran, Saudi Arabia
E-mail: qhansari@gmail.com

October 15, 2020

This talk is based on Chapter 2 of the following books:

- Q. H. Ansari, C. S. Lalitha and M. Mehta, *Generalized Convexity, Nonsmooth Variational Inequalities and Nonsmooth Optimization*, Taylor & Francis Group (2014)
- W. Forst and D. Hoffmann, *Optimization – Theory and Practice*, Springer Science & Business Media (2010)

Helped by:
Pradeep, Jauny, Ramsurat, Gourav, Debdas

Table of contents

PART - II
OPTIMALITY CONDITIONS

Constrained Optimization

We study first order (that uses gradient) optimality conditions for the following constrained optimization problem:

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, 2, \ldots, m \\ & h_j(x) = 0,\ j = 1, 2, \ldots, \ell, \\ & x \in X\,(\neq \emptyset) \subseteq \mathbb{R}^n, \end{cases} \tag{1}$$

where f and all of g_i's and h_j's are continuously differentiable on X.

$$\mathscr{F} = \{x \in X : g_i(x) \leq 0 \text{ for } i = 1, 2, \ldots, m \text{ and } h_i(x) = 0 \text{ for } j = 1, 2, \ldots, \ell\}$$

Given a feasible point x_0, a constraint $g_i(x) \leq 0$ is called

- *active* if $g_i(x_0) = 0$
- *inactive* if $g_i(x_0) < 0$.

(An equality constraint $h_j(x) = 0$ is active at any feasible point.)

For a feasible point x_0, we set

$$\mathscr{A}(x_0) := \{i \in \{1, 2, \ldots, m\} : g_i(x_0) = 0\},$$

which describes the inequality restrictions that are active at x_0.

The active constraints have a special significance: if a constraint is inactive ($g_i(x_0) < 0$) at the feasible point x_0, it is possible to move from x_0 a bit in any direction without violating this constraint.

Cone of feasible direction

Definition

Let $d \in \mathbb{R}^n$ and $x_0 \in \mathscr{F}$. A vector d is called the *feasible direction* of $\mathscr{F}$ at x_0 if

$$\exists\ \delta > 0 \text{ such that } x_0 + \tau d \in \mathscr{F}, \quad \forall \tau \in [0, \delta].$$

(That is, a 'small' movement from x_0 along such a direction gives feasible points.)

Set of all feasible directions of $\mathscr{F}$ at x_0 constitutes a cone, which is denoted by $\mathcal{C}_{fd}(x_0)$ and it is called *cone of feasible directions*. That is,

$$\mathcal{C}_{fd}(x_0) = \{d \in \mathbb{R}^n : \exists\ \delta > 0 \text{ such that } x_0 + \tau d \in \mathscr{F}, \forall \tau \in [0, \delta]\}. \quad (2)$$

The cone $\mathcal{C}_{fd}(x_0)$ is also referred as *radial cone* at x_0.

Let d be a feasible direction of $\mathscr{F}$ at x_0. Then, for an $i \in \mathscr{A}(x_0)$, there exists $\delta > 0$ such that for $0 < \tau \leq \delta$, we have

$$\underbrace{g_i(x_0 + \tau d)}_{\leq 0} = \underbrace{g_i(x_0)}_{=0} + \tau \langle \nabla g_i(x_0), d \rangle + \mathrm{o}(\tau).$$

Dividing by τ and passing the limit as $\tau \to 0^+$ gives $\langle \nabla g_i(x_0), d \rangle \leq 0$. In the same way, we get $\langle \nabla h_j(x_0), d \rangle = 0$ for all $j = 1, 2, \ldots, \ell$.

Linearizing cone

Definition

For any $x_0 \in \mathscr{F}$, the set (cone)

$$\mathcal{C}_l(x_0) := \Big\{ d \in \mathbb{R}^n : \langle \nabla g_i(x_0), d \rangle \leq 0 \, \forall i \in \mathscr{A}(x_0), \\ \langle \nabla h_j(x_0), d \rangle = 0 \, \forall j = 1, 2, \ldots, \ell \Big\}$$

is called *linearizing cone*, of the constrained optimization problem (1), at x_0.

Evidently, $\mathcal{C}_l(x_0)$ contains at least all feasible directions of $\mathscr{F}$ at x_0, i.e., $\mathcal{C}_{fd}(x_0) \subset \mathcal{C}_l(x_0)$.

Definition (Cone of descent directions)

For any $x_0 \in X$, the set (cone)

$$\mathcal{C}_{dd}(x_0) := \{d \in \mathbb{R}^n : \langle \nabla f(x_0), d \rangle < 0\}$$

is called *cone of descent directions* of f at x_0.

Note that for any $d \in \mathcal{C}_{dd}(x_0)$,

$$\begin{aligned}
& f(x_0 + \tau d) = f(x_0) + \tau \underbrace{\langle \nabla f(x_0), d \rangle}_{<\ 0} + o(\tau) \\
\Longrightarrow\ & \lim_{\tau \to 0+} \frac{1}{\tau}(f(x_0 + \tau d) - f(x_0)) = \langle \nabla f(x_0), d \rangle < 0 \\
\Longrightarrow\ & \exists \delta > 0 \text{ such that } \frac{1}{\tau}(f(x_0 + \tau d) - f(x_0)) < 0\, \forall \tau \in (0, \delta] \\
& (\text{since } \frac{1}{\tau}(f(x_0 + \tau d) - f(x_0) \text{ is continuous for } \tau > 0) \\
\Longrightarrow\ & \exists \delta > 0 \text{ such that } f(x_0 + \tau d) < f(x_0),\ \forall \tau \in (0, \delta].
\end{aligned} \tag{3}$$

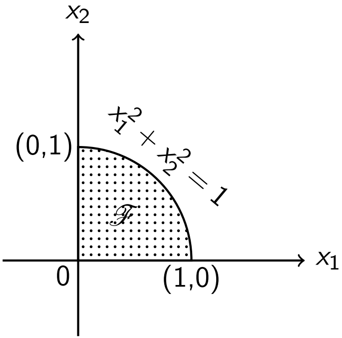
Thus, $d \in \mathcal{C}_{dd}(x_0)$ guarantees that the objective function f can be reduced along this direction.

Examples of $\mathcal{C}_{fd}(x_0)$, $\mathcal{C}_l(x_0)$ and $\mathcal{C}_{dd}(x_0)$

Let $f(x) := x_1 + x_2$ and

$$\mathscr{F} = \left\{ x = (x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 - 1 \leq 0,\ -x_1 \leq 0,\ -x_2 \leq 0 \right\},$$

where $g_1(x) := x_1^2 + x_2^2 - 1$, $g_2(x) := -x_1$ and $g_3(x) := -x_2$.
Evidently, f attains a (strict, global) minimum at $(0, 0)$.

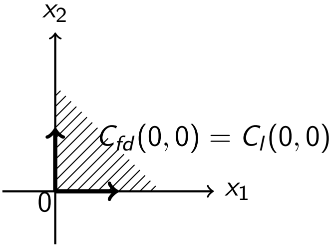


Examples of $\mathcal{C}_{fd}(x_0)$ and $\mathcal{C}_l(x_0)$

Take $x_0 = (0,0)$. Then, $\mathscr{A}(x_0) = \{2,3\}$,

$$\mathcal{C}_{fd}(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_1 \geq 0,\ d_2 \geq 0\}$$

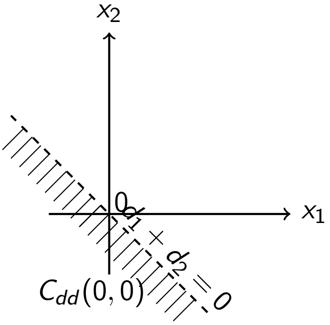
$$\begin{aligned} \mathcal{C}_l(x_0) &= \{d \in \mathbb{R}^2 : \langle \nabla g_2(x_0), d \rangle \leq 0,\ \langle \nabla g_3(x_0), d \rangle \leq 0\} \\ &= \left\{(d_1, d_2) \in \mathbb{R}^2 : d_1 \geq 0, d_2 \geq 0\right\} \end{aligned}$$



Examples of $\mathcal{C}_{dd}(x_0)$

At $x_0 = (0, 0)$,

$$\begin{aligned}\mathcal{C}_{dd}(x_0) &= \{d \in \mathbb{R}^2 : \langle \nabla f(x_0), d \rangle < 0\} \\ &= \{(d_1, d_2) \in \mathbb{R}^2 : d_1 + d_2 < 0\}.\end{aligned}$$

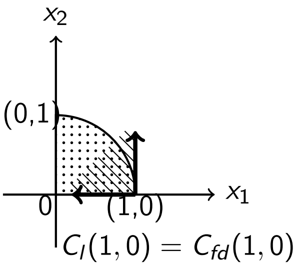


Examples of $\mathcal{C}_{fd}(x_0)$ and $\mathcal{C}_l(x_0)$

Take $x_0 = (1,0)$. Then, $\mathscr{A}(x_0) = \{1,3\}$,

$$\mathcal{C}_{fd}(x_0) = \{(d_1,d_2) \in \mathbb{R}^2 : d_1 \leq 0, d_2 \geq 0\},$$

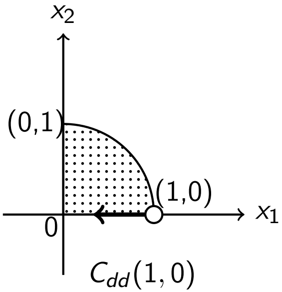
$$\begin{aligned} \mathcal{C}_l(x_0) &= \{d \in \mathbb{R}^2 : \langle \nabla g_1(x_0), d\rangle \leq 0,\ \langle \nabla g_3(x_0) d\rangle \leq 0\} \\ &= \left\{(d_1,d_2) \in \mathbb{R}^2 : d_1 \leq 0, d_2 \geq 0\right\}. \end{aligned}$$



Examples of $\mathcal{C}_{dd}(x_0)$

At $x_0 = (1, 0)$,

$$\mathcal{C}_{dd}(x_0) = \{d \in \mathbb{R}^2 : \langle \nabla f(x_0), d \rangle < 0\} = \{(d_1, d_2) \in \mathbb{R}^2 : d_1 < 0\}.$$



Geometric necessary optimality condition 1

Theorem

If x_0 is a local minimizer of the constrained optimization problem (1), then

$$\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0) = \emptyset.$$

Proof

Assume contrary that there exists $d \in \mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0)$. Then, from (2) and (3), there exists $\delta > 0$ such that

$$x_0 + \tau d \in \mathscr{F} \text{ and } f(x_0 + \tau d) < f(x_0) \text{ for all } \tau \in (0, \delta],$$

which is contradictory to x_0 being a local minimizer of (1). Hence, $\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0) = \emptyset$.

Remark

The condition $\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0) = \emptyset$ is only a necessary optimality condition not sufficient. For example, consider

$$\begin{cases} \text{minimize} & f(x_1, x_2) := x_2 \\ \text{subject to} & h_1(x) := x_1^2 + x_2^2 - 1 = 0. \end{cases}$$

Take $x_0 = (0, 1)$. In this problem, $\mathcal{C}_{fd}(x_0) = \{\mathbf{0}\}$ and
$\mathcal{C}_{dd}(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_2 < 0\}$.
Hence, $\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0) = \emptyset$ but x_0 is not a minimizer.

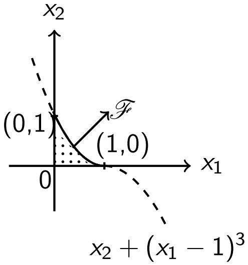
Question arises whether not just $\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{fd}(x_0) = \emptyset$ but even $\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_{l}(x_0) = \emptyset$ is true for any local minimizer $x_0 \in \mathscr{F}$.

The next example gives a negative answer to this question.

$\mathcal{C}_l(x_0) \cap \mathcal{C}_{dd}(x_0) = \emptyset$ is not necessary for optimality

Consider $X = \mathbb{R}^2$ and the following optimization problem:

$$\begin{cases} \text{minimize} & f(x) := -x_1 \\ \text{subject to} & g_1(x) := x_2 + (x_1 - 1)^3 \leq 0 \\ & g_2(x) := -x_1 \leq 0 \\ & g_3(x) := -x_2 \leq 0 \\ & x \in X. \end{cases}$$

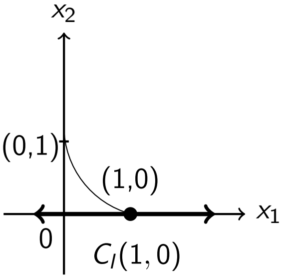


$\mathcal{C}_l(x_0) \cap \mathcal{C}_{dd}(x_0) = \emptyset$ is not necessary for optimality

Here $x_0 = (1, 0)$ is a minimizer. At x_0, we have $\mathscr{A}(x_0) = \{1, 3\}$,

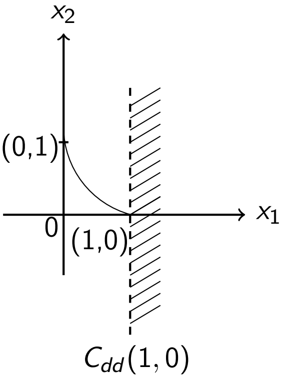
$$\nabla f(x_0) = (-1, 0),\ \nabla g_1(x_0) = (0, 1),\ \nabla g_2(x_0) = (-1, 0)$$

and $\nabla g_3(x_0) = (0, -1)$. Thus, $\mathcal{C}_l(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_2 = 0\}$



$\mathcal{C}_l(x_0) \cap \mathcal{C}_{dd}(x_0) = \emptyset$ is not necessary for optimality

At $x_0 = (1, 0)$, $\mathcal{C}_{dd}(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_1 > 0\}$.



Evidently, $\mathcal{C}_l(x_0) \cap \mathcal{C}_{dd}(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_1 > 0,\ d_2 = 0\} \neq \emptyset$.

As feasible directions allows movements along straight line segments those are embedded in $\mathscr{F}$, the cone of feasible direction turns out to be the singleton set containing the null vector if the feasible set $\mathscr{F}$ is a curved surface.

For example, consider $\mathscr{F} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ and any $x_0 \in \mathscr{F}$.
Note that $\mathcal{C}_{fd}(x_0) = \{\mathbf{0}\} = \{(0,0)\}$.
Thus, often there would be no point of $\mathscr{F}$ from which progress could be made by a descent method for minimization.

As feasible directions allows movements along straight line segments those are embedded in $\mathscr{F}$, the cone of feasible direction turns out to be the singleton set containing the null vector if the feasible set $\mathscr{F}$ is a curved surface.

For example, consider $\mathscr{F} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ and any $x_0 \in \mathscr{F}$.
Note that $\mathcal{C}_{fd}(x_0) = \{\mathbf{0}\} = \{(0,0)\}$.
Thus, often there would be no point of $\mathscr{F}$ from which progress could be made by a descent method for minimization.

As $\mathcal{C}_{fd}(x_0)$ is too limited to include quite a few descent direction, to development of a descent algorithms and to establish optimality conditions that are general enough, the notion of tangent cone is introduced. Although the notion of tangent cone is less intuitive but contains $\mathcal{C}_{fd}(x_0)$.

As feasible directions allows movements along straight line segments those are embedded in $\mathscr{F}$, the cone of feasible direction turns out to be the singleton set containing the null vector if the feasible set $\mathscr{F}$ is a curved surface.

For example, consider $\mathscr{F} = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ and any $x_0 \in \mathscr{F}$.
Note that $\mathcal{C}_{fd}(x_0) = \{\mathbf{0}\} = \{(0,0)\}$.
Thus, often there would be no point of $\mathscr{F}$ from which progress could be made by a descent method for minimization.

As $\mathcal{C}_{fd}(x_0)$ is too limited to include quite a few descent direction, to development of a descent algorithms and to establish optimality conditions that are general enough, the notion of tangent cone is introduced.
Although the notion of tangent cone is less intuitive but contains $\mathcal{C}_{fd}(x_0)$.

We say that a sequence $\{x_k\}$ in $\mathbb{R}^n$ converges in direction $d \in \mathbb{R}^n$ to x_0 if there exist sequences $\{r_k\}$ in $\mathbb{R}^n$ and $\{\alpha_k\}$ in $\mathbb{R}_+$ with $r_k \to 0$ and $\alpha_k \downarrow 0$ such that

$$x_k = x_0 + \alpha_k(d + r_k).$$

Then, we use the *notation*: $x_k \xrightarrow{d} x_0$.

Note that $x_k \xrightarrow{d} x_0$ simply means there exists a sequence of positive numbers $\{\alpha_k\}$ such that $\alpha_k \downarrow 0$ and

$$\frac{1}{\alpha_k}(x_k - x_0) \to d.$$

Tangent Cone

Definition

Let M be a nonempty subset of $\mathbb{R}^n$ and $x_0 \in M$.
A vector $d \in \mathbb{R}^n$ is said to be a *tangent* or *tangent directions* to M at x_0 if there exist a sequence $\{x_k\}$ in M and a sequence $\{\alpha_k\}$ of positive real numbers such that $x_k \to x_0$ and $\alpha_k(x_k - x_0) \to d$.

The set $\mathcal{C}_t(x_0)$ of all tangents to M at x_0 is called the *tangent cone* to M at x_0.

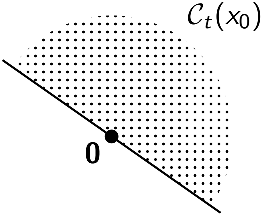
The set

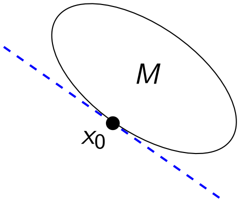
$$\mathcal{C}_t(x_0) := \left\{ d \in \mathbb{R}^n : \exists \{x_k\} \text{ in } M \text{ such that } x_k \xrightarrow{d} x_0 \right\}$$

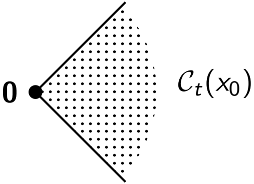
is called the *tangent cone* of M at x_0.

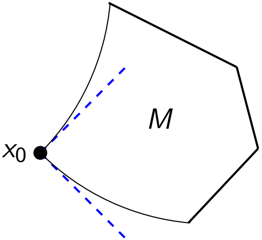
The cone $\mathcal{C}_t(x_0)$ is also referred as *Bouligand's contingent / tangent cone*.

Tangent Cone









Some Properties of Tangent Cone

It is evident that the tangent cone is really a cone.

If M is a subset of $\mathbb{R}^n$ with nonempty interior, then for every $x_0 \in \text{int}(M)$, we have $\mathcal{C}_t(x_0) = \mathbb{R}^n$.

Theorem

Let M be a nonempty subset of $\mathbb{R}^n$.

(a) *$\mathcal{C}_t(x_0)$ is convex for every $x_0 \in M$.*

(b) *$\mathcal{C}_t(x_0)$ is closed for every $x_0 \in \overline{M}$ (closure of M).*

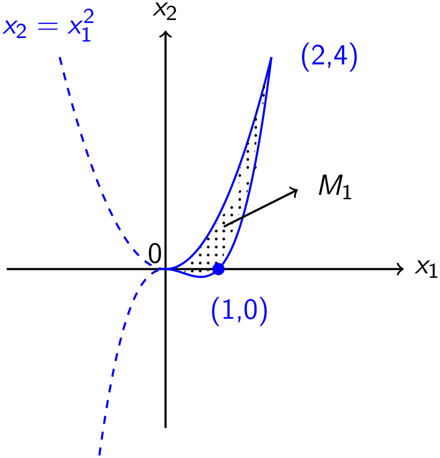
For further details on tangent cone, we refer to the following book:

J. Jahn, Introduction to the Theory of Nonlinear Optimization, Fourth Edition, Springer (2020)

Examples of tangent cone $\mathcal{C}_t(x_0)$

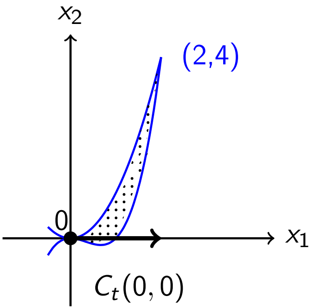
Consider the set

$$M_1 := \left\{ x = (x_1, x_2) \in \mathbb{R}^2 : x_1 \geq 0,\ x_1^2 \geq x_2 \geq x_1^2(x_1 - 1) \right\}.$$



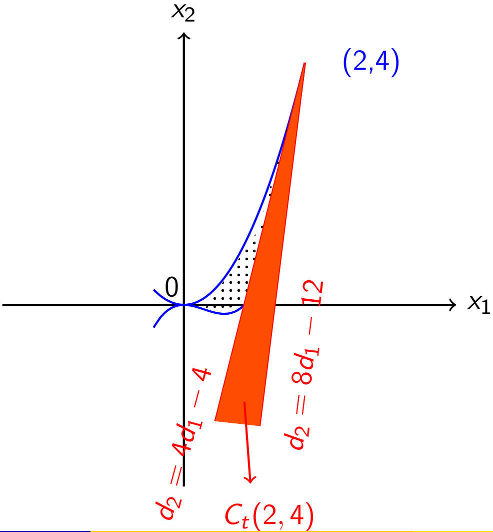
Examples of tangent cone $\mathcal{C}_t(x_0)$

At $x_0 = (0,0)$, $\mathcal{C}_t(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_1 \geq 0, d_2 = 0\}$.



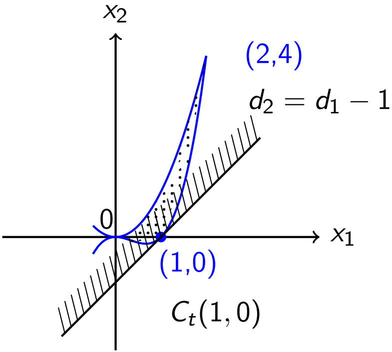
Examples of tangent cone $\mathcal{C}_t(x_0)$

At $x_0 = (2,4)$, $\mathcal{C}_t(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : 8d_1 - 12 \leq d_2 \leq 4d_1 - 4\}$.



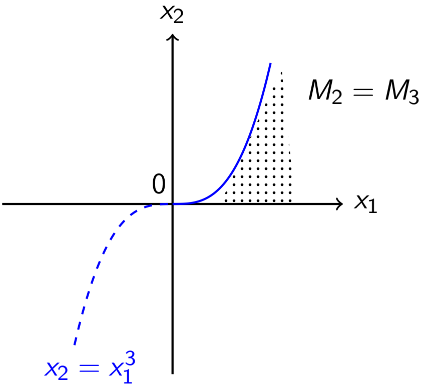
Examples of tangent cone $\mathcal{C}_t(x_0)$

At $x_0 = (1, 0)$, $\mathcal{C}_t(x_0) = \{(d_1, d_2) \in \mathbb{R}^2 : d_2 \geq d_1 - 1\}$.



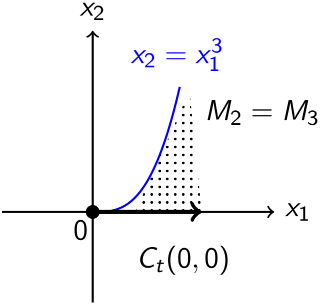
Examples of tangent cone $\mathcal{C}_t(x_0)$

Consider $M_2 := \left\{x \in \mathbb{R}^2 : -x_1^3 + x_2 \leq 0, -x_2 \leq 0\right\}$.



Examples of tangent cone $\mathcal{C}_t(x_0)$

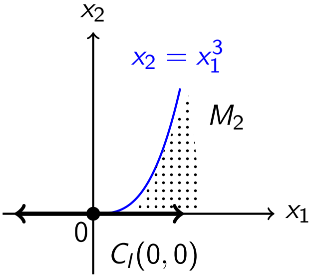
At $x_0 = (0,0)$, $\mathcal{C}_t(x_0) = \left\{d \in \mathbb{R}^2 : d_1 \geq 0,\ d_2 = 0\right\}$.



Examples of tangent cone $\mathcal{C}_t(x_0)$

Consider $M_3 := \left\{x \in \mathbb{R}^2 : -x_1^3 + x_2 \leq 0, -x_1 \leq 0, -x_2 \leq 0\right\}$. Actually, it is equal to M_2.

At $x_0 = (0,0)$, $\mathcal{C}_t(x_0) = \left\{d \in \mathbb{R}^2 : d_1 \geq 0,\ d_2 = 0\right\}$.



Examples of tangent cone $\mathcal{C}_t(x_0)$

Theorem (Relationship among $\mathcal{C}_{fd}(x_0)$, $\mathcal{C}_l(x_0)$ and $\mathcal{C}_t(x_0)$)

$$\overline{\mathcal{C}_{fd}(x_0)} \subseteq \mathcal{C}_t(x_0) \subseteq \mathcal{C}_l(x_0).$$

$\mathcal{C}_t(x_0)$ is strictly a subset of $\mathcal{C}_l(x_0)$

Example

Consider $M_2 := \left\{x \in \mathbb{R}^2 : -x_1^3 + x_2 \leq 0, -x_2 \leq 0\right\}$.

At $x_0 = (0,0)$, $\mathcal{C}_t(x_0) = \left\{d \in \mathbb{R}^2 : d_1 \geq 0,\ d_2 = 0\right\}$.

At $x_0 = (0,0)$, $\mathscr{A}(x_0) = \{1,2\}$.

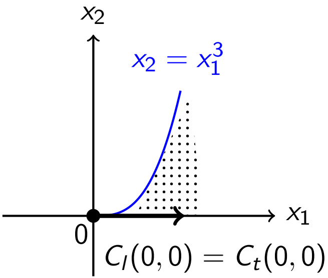
which gives $\mathcal{C}_l(x_0) = \left\{d \in \mathbb{R}^2 : d_2 = 0\right\}$.
Hence, in this example, $\mathcal{C}_t(x_0) \subset \mathcal{C}_l(x_0)$.

$\mathcal{C}_t(x_0)$ is identical to $\mathcal{C}_l(x_0)$

Example

Consider $M_3 := \{x \in \mathbb{R}^2 : -x_1^3 + x_2 \leq 0,\ -x_1 \leq 0,\ -x_2 \leq 0\}$ and $x_0 = (0,0)$. Then, $\mathscr{A}(x_0) = \{1,2,3\}$ and therefore

$$\mathcal{C}_l(x_0) = \{d \in \mathbb{R}^2 : d_1 \geq 0,\ d_2 = 0\} = \mathcal{C}_t(x_0).$$



Remark

From the above two examples, we notice that although $M_2 = M_3$, the linearizing cones are different. Thus, linearizing cone is dependent on the algebraic representation of the set of feasible points! However, it is true in general that tangent cone depends only on the geometric representation and not on algebraic representation of the set.

Theorem (Geometric necessary optimality condition 2)

If x_0 is a local minimizer of the optimization problem (1)*, then* $\nabla f(x_0) \in \mathcal{C}_t(x_0)^*$*, and hence*

$$\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_t(x_0) = \emptyset.$$

Theorem (First order Karush-Kuhn-Tucker necessary optimality condition)

Suppose that x_0 is a local minimizer of (1)*, and (the constraint qualification) $\mathcal{C}_\ell(x_0)^* = \mathcal{C}_t(x_0)^*$ is fulfilled. Then, there exist vectors $\lambda \in \mathbb{R}^m_+$ and $\mu \in \mathbb{R}^\ell$ such that*

$$\nabla f(x_0) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_0) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(x_0) = \mathbf{0} \tag{4}$$

$$\text{and } \lambda_i g_i(x_0) = 0 \text{ for } i = 1, 2, \ldots, m. \tag{5}$$

Theorem (Geometric necessary optimality condition 2)

If x_0 is a local minimizer of the optimization problem (1), then $\nabla f(x_0) \in \mathcal{C}_t(x_0)^*$, and hence

$$\mathcal{C}_{dd}(x_0) \cap \mathcal{C}_t(x_0) = \emptyset.$$

Theorem (First order Karush-Kuhn-Tucker necessary optimality condition)

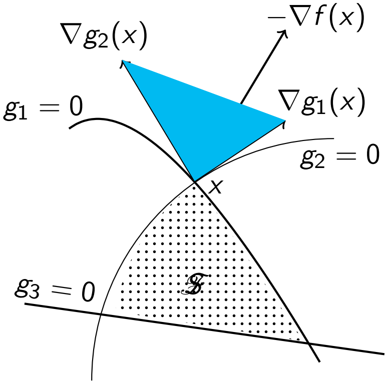
Suppose that x_0 is a local minimizer of (1), and (the constraint qualification) $\mathcal{C}_\ell(x_0)^* = \mathcal{C}_t(x_0)^*$ is fulfilled. Then, there exist vectors $\lambda \in \mathbb{R}^m_+$ and $\mu \in \mathbb{R}^\ell$ such that

$$\nabla f(x_0) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_0) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(x_0) = \mathbf{0} \tag{4}$$

$$\text{and } \lambda_i g_i(x_0) = 0 \text{ for } i = 1, 2, \ldots, m. \tag{5}$$

KKT Conditions

(4) and (5) are referred as KKT conditions.



KKT point

Definition

A point $x_0 \in X$ is called a *KKT point* for the problem (1) if it satisfies the following four (KKT) conditions:

1. (Primal Feasibility). x_0 is a feasible point of (1).
2. (Stationarity). $\nabla f(x_0) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_0) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(x_0) = \mathbf{0}$ for some $\lambda = (\lambda_1 \ldots, \lambda_m) \in \mathbb{R}^m$ and $\mu = (\mu_1, \ldots, \mu_\ell) \in \mathbb{R}^\ell$.
3. (Dual Feasibility). The stationarity condition is satisfied for a $\lambda \in \mathbb{R}^m_+$ whose all the components are nonnegative.
4. (Complementary Slackness). With λ in dual feasibility, $\lambda_i g_i(x_0) = 0$.

Examples of a KKT point

Consider the problem

$$\begin{cases} \text{minimize} & f(x_1, x_2) := (x_1 - 3)^2 + (x_2 - 2)^2 \\ \text{subject to} & g_1(x_1, x_2) := x_1^2 + x_2^2 - 5 \leq 0 \\ & g_2(x_1, x_2) := -x_1 \leq 0 \\ & g_3(x_1, x_2) := -x_2 \leq 0 \\ & h_1(x_1, x_2) := x_1 + 2x_2 = 4. \end{cases}$$

Let us check if the points $x_0 = (2, 1)$ is a KKT points for this problem.
The point x_0 is a (primal) feasible point.

Examples of a KKT point continue ....

We see that $\mathscr{A}(x_0) = \{1\}$ and

$$\nabla f(x_0) = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \nabla g_1(x_0) = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \text{ and } \nabla h_1(x_0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Let us check if the stationarity holds at x_0. For this, consider the system

$$\begin{aligned} \nabla f(x_0) + \lambda_1 \nabla g_1(x_0) + \mu_1 \nabla h_1(x_0) = 0 \\ \text{i.e., } \begin{bmatrix} -2 \\ -2 \end{bmatrix} + \lambda_1 \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \mu_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{6}$$

Notice that $\lambda_1 = \frac{1}{3}$ and $\mu_1 = \frac{2}{3}$ is a solution to (6). Thus, x_0 satisfies stationarity.

Examples of a KKT point continue ....

Taking $\lambda_2 = \lambda_3 = 0$, we see that

$$\lambda_1 g_1(x_0) = 0, \lambda_2 g_2(x_0) = 0 \text{ and } \lambda_3 g_3(x_0) = 0,$$

which shows that complementary slackness holds at x_0.
Further notice that $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_3 \geq 0$. Thus, dual feasibility holds.
Therefore, x_0 is a KKT point for the considered problem.

First order KKT sufficient optimality condition

Theorem

Consider the problem (1) where f and g_i, $i = 1, 2, \ldots, m$ are convex and $h_j(x) = \langle a_j, x \rangle - b_j$ for some $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$, $j = 1, 2, \ldots, \ell$. Suppose at an $x_0 \in \mathscr{F}$ the KKT conditions are satisfied. Then, x_0 is a global minimum point.

Remark

There is one important point to note in regard to KKT conditions that is often a source of error. Namely, despite the usually the well behaved nature of convex programming problems and the sufficiency of KKT conditions under convexity assumptions, the KKT conditions are not necessary for optimality for convex programming problems. For example, consider

$$\begin{cases} \text{minimize} & x_1 \\ \text{subject to} & (x_1-1)^2+(x_2-1)^2 \leq 1 \\ & (x_1-1)^2+(x_2+1)^2 \leq 1. \end{cases}$$

The only feasible point $(1,0)$ is naturally optimal. However, this is not a KKT point.

For the problem in Remark 44, a careful reader must have observed that the point $(1,0)$ does not satisfy the hypothesis of Theorem 11 since $\mathcal{C}_l(x_0)^* = \{(0,0)\} \neq \mathbb{R}^2 = \mathcal{C}_t(x_0)^*$. Hence, although $(1,0)$ is optimizer, it is not a KKT point.

In order to ensure that the KKT conditions are necessary for optimality, a constraint qualification (CQ) is needed. A CQ is an assumption made about the constraint functions that, when satisfied by a local minimizer, ensures stationarity.

However, the CQ $\mathcal{C}_l(x_0)^* = \mathcal{C}_t(x_0)^*$ is very abstract, extremely general, but not easily verifiable. Therefore, for practical problems, we will try to find regularity assumptions called *constraint qualification* (CQ) which are more specific, easily verifiable, but also somewhat restrictive.
For the moment, we consider the case that we only have inequality constraints.
As affine constraints poses fewer problems than nonlinear constraints, we the partition the set inequality indices $\mathscr{I} = \{1, 2, \ldots, m\}$ by

$$\mathscr{I} = \mathscr{I}_1 \cup \mathscr{I}_2,$$

where $\mathscr{I}_2$ consists the indices corresponding to the affine inequalities and $\mathscr{I}_1$ are for nonaffine constraints.
Corresponding to this partition, we will also split up the set of active constraints $\mathscr{A}(x_0)$ for $x_0 \in \mathscr{F}$ into

$$\mathscr{A}_j(x_0) := \mathscr{I}_j \cap \mathscr{A}(x_0), \quad \text{for } j = 1, 2.$$

We now will focus on following Constraint Qualifications:

(GCQ) Guignard Constraint Qualification: $\mathcal{C}_l(x_0)^* = \mathcal{C}_t(x_0)^*$

(ACQ) Abadie Constraint Qualification: $\mathcal{C}_l(x_0) = \mathcal{C}_t(x_0)$

(MFCQ) Mangasarian-Fromovitz Constraint Qualification:

$$\exists\ d \in \mathbb{R}^n \text{ such that } \begin{cases} g_i'(x_0)d < 0 & \text{for } i \in \mathscr{A}_1(x_0) \\ g_i'(x_0)d \leq 0 & \text{for } i \in \mathscr{A}_2(x_0) \end{cases}$$

(SCQ) Slater Constraint Qualification: The functions g_i are convex for all $i \in \mathscr{I}$ and there exists $\tilde{x} \in \mathscr{F}$ such that $g_i(\tilde{x}) < 0$ for $i \in \mathscr{I}_1$.

The conditions $g_i'(x_0)d < 0$ and $g_i'(x_0)d \leq 0$ each define half space. (MFCQ) means nothing else but that the intersection of all of these half spaces is nonempty.

Theorem

(SCQ) $\implies$ *(MFCQ)* $\implies$ *(ACQ)* $\implies$ *(GCQ).*

(LICQ) The vectors $(\nabla g_i(x_0) : i \in \mathscr{A}(x_0))$ and $(\nabla h_j(x_0) : j \in 1, 2, \ldots, \ell)$ are linearly independent.

(LICQ) greatly reduces the number of active inequality constraints. Instead of (LICQ) we will now consider the following weaker constraint qualification which is a variant of (MFCQ), and is often cited as the Arrow-Hurwitz-Uzawa Constraint Qualification:

(AHUCQ) There exists a vector $d \in \mathbb{R}^n$ such that

$$\begin{cases} g_i'(x_0)d < 0 & \text{for } i \in \mathscr{A}_1(x_0) \\ h_j'(x_0)d = 0 & \text{for } j \in \mathscr{J}, \end{cases}$$

and the vectors $(\nabla h_j(x_0) : j \in \mathscr{J})$ are linearly independent.

Theorem

(LICQ) $\Longrightarrow$ *(AHUCQ)* $\Longrightarrow$ *(ACQ)* $\Longrightarrow$ *(GCQ).*

Consider the following constrained optimization problem with equality constraints:

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & h_j(x) = 0, \quad j = 1, 2, \ldots, \ell, \end{cases} \tag{P}$$

where f and h_j $(j = 1, 2, \ldots, \ell)$ are continuously differentiable.

In order to solve (P), in Lagrange multiplier method, we (loosely) follow the following steps:

- Construct the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^\ell \to \mathbb{R}$ given by

$$L(x, \lambda) = f(x) + \sum_{j=1}^{\ell} \lambda_j h_j(x).$$

- Solve the following system for (x, λ):

$$\left.\begin{aligned} \nabla_x L(x, \lambda) &\equiv \nabla f(x) + \sum_{j=1}^{\ell} \lambda_j \nabla h_j(x) = 0, \\ \nabla_\lambda L(x, \lambda) &\equiv h_j(x) = 0, \ j = 1, 2, \ldots, \ell. \end{aligned}\right\} \tag{7}$$

- Let S be the set of x's which satisfy (7). Find the points in S for which f gets minimum most value and declare that those points are solutions to (7).

One should be mathematically careful when applying the above three steps to solve (P):

Caution

- Without a *regularity* assumption, the system (7) may be inconsistent at a local minimum of (P).
- *Step* 3 can be misleading without an extra check—a point of S that gives minimum most value of f in S may not be a local minimum for the problem (P).

Example for Careful 1

Consider the following minimization problem

$$\begin{cases} \text{minimize} & f(x_1, x_2) := x_1 + x_2 \\ \text{subject to} & h_1(x_1, x_2) := (x_1 - 1)^2 + x_2^2 - 1 = 0 \\ & h_2(x_1, x_2) := (x_1 - 2)^2 + x_2^2 - 4 = 0. \end{cases}$$

This problem has only one feasible point $(x_1, x_2) = (0, 0)$, and hence the objective function is minimized at this point. At the point $(0, 0)$, we have

$$\nabla f(0,0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \nabla h_1(0,0) = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad \text{and} \quad \nabla h_2(0,0) = \begin{bmatrix} -4 \\ 0 \end{bmatrix}.$$

Clearly, there does not exist (λ_1, λ_2) such that

$$\nabla f(0,0) + \lambda_1 \nabla h_1(0,0) + \lambda_2 \nabla h_1(0,0) = 0.$$

The issue here is $\nabla h_1(0,0)$ and $\nabla h_2(0,0)$ are linearly dependent.

Example for Careful 2

Consider the following minimization problem

$$\begin{cases} \text{minimize} & f(x_1, x_2) := x_1 x_2 \\ \text{subject to} & h_1(x_1, x_2) := x_1 + x_2 - 4 = 0. \end{cases}$$

The system (7) for this example is

$$\begin{cases} \nabla f(x_1, x_2) + \lambda_1 \nabla h_1(x_1, x_2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ h_1(x_1, x_2) = 0 \end{cases}$$

$$\text{i.e.,} \quad \begin{cases} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \lambda_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ x_1 + x_2 = 4 \end{cases}$$

$$\text{i.e.,} \quad x_1 = x_2 = -\lambda_1 = 2.$$

Therefore, $S = \{(2,2)\}$. However, notice that $(2,2)$ is not a minimizer of x_1x_2 over $x_1 + x_2 = 4$. In fact, x_1x_2 reduces indefinitely on $x_1 + x_2 = 4$. The basic issue with the Step 3 is that without checking the definiteness Hessian of the Lagrangian (on certain set) we are declaring that a point is minimum.

Let X be a nonempty subset of $\mathbb{R}^n$, $f : X \to \mathbb{R}$ and $g_i : X \to \mathbb{R}$ be functions for $\mathscr{I} := \{1, 2 \ldots, m\}$. Consider the following optimization problem with equality constraints:

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) = 0,\ i = 1, 2, \ldots, m, \end{cases} \tag{P}$$

For the problem (P), define a Lagrangian function $L : \mathbb{R}^n \times \mathbb{R}^m_+$ by

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = f(x) + \langle \lambda, g(x) \rangle, \quad \text{for } x \in X \text{ and } \lambda \in \mathbb{R}^m_+, \tag{8}$$

where $g(x) = (g_1(x), \ldots, g_m(x))$ and $\lambda = (\lambda_1, \ldots, \lambda_m)$ with $\lambda_i \geq 0$.

Definition

A pair $(\bar{x}, \bar{\lambda}) \in X \times \mathbb{R}^m_+$ is said to be a *saddle point* of L if and only if

$$L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}), \quad \forall\, x \in X \text{ and } \lambda \in \mathbb{R}^m_+,$$

that is, $\bar{x}$ minimizes $L(\cdot, \bar{\lambda})$ and $\bar{\lambda}$ maximizes $L(\bar{x}, \cdot)$.

Lemma

If $(\bar{x}, \bar{\lambda})$ is a saddle point of L, then

(a) *$\bar{x}$ is a global minimizer of* (P)*;*

(b) *$L(\bar{x}, \bar{\lambda}) = f(\bar{x})$;*

(c) *$\bar{\lambda}_i g_i(\bar{x}) = 0$ for all $i =, 1, 2, \ldots, m$.*

Definition

A pair $(\bar{x}, \bar{\lambda}) \in X \times \mathbb{R}^m_+$ is said to be a *saddle point* of L if and only if

$$L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}), \quad \forall\, x \in X \text{ and } \lambda \in \mathbb{R}^m_+,$$

that is, $\bar{x}$ minimizes $L(\cdot, \bar{\lambda})$ and $\bar{\lambda}$ maximizes $L(\bar{x}, \cdot)$.

Lemma

If $(\bar{x}, \bar{\lambda})$ is a saddle point of L, then

(a) *$\bar{x}$ is a global minimizer of* (P)*;*

(b) *$L(\bar{x}, \bar{\lambda}) = f(\bar{x})$;*

(c) *$\bar{\lambda}_i g_i(\bar{x}) = 0$ for all $i =, 1, 2, \ldots, m$.*

If X is an open and convex subset of $\mathbb{R}^n$ and $f : X \to \mathbb{R}$ and $g_i : C \to \mathbb{R}$ are continuously differentiable and convex for $i = 1, 2, \ldots, m$, then we write (CP) instead of (P)

Theorem

If the SLATER *constraint qualification*[a] *holds and $\bar{x}$ is a minimizer of (CP), then there exists a vector $\bar{\lambda} \in \mathbb{R}^m_+$ such that $L(\bar{x}, \bar{\lambda})$ is a saddle point of L.*

[a] SLATER Constraint Qualification: The functions g_i are convex for all $i \in \mathscr{I}$ and there exists $\tilde{x} \in \mathscr{F}$ such that $g_i(\tilde{x}) < 0$ for $i \in \mathscr{I}_1$.

If X is an open and convex subset of $\mathbb{R}^n$ and $f : X \to \mathbb{R}$ and $g_i : C \to \mathbb{R}$ are continuously differentiable and convex for $i = 1, 2, \ldots, m$, then we write (CP) instead of (P)

Theorem

If the SLATER *constraint qualification[a] holds and $\bar{x}$ is a minimizer of (CP), then there exists a vector $\bar{\lambda} \in \mathbb{R}^m_+$ such that $L(\bar{x}, \bar{\lambda})$ is a saddle point of L.*

[a] SLATER Constraint Qualification: The functions g_i are convex for all $i \in \mathscr{I}$ and there exists $\tilde{x} \in \mathscr{F}$ such that $g_i(\tilde{x}) < 0$ for $i \in \mathscr{I}_1$.

In the following example we see that the SLATER constraint qualification is essential in the above theorem.

If X is an open and convex subset of $\mathbb{R}^n$ and $f : X \to \mathbb{R}$ and $g_i : C \to \mathbb{R}$ are continuously differentiable and convex for $i = 1, 2, \ldots, m$, then we write (CP) instead of (P)

Theorem

If the SLATER *constraint qualification*[a] *holds and* $\bar{x}$ *is a minimizer of (CP), then there exists a vector* $\bar{\lambda} \in \mathbb{R}^m_+$ *such that* $L(\bar{x}, \bar{\lambda})$ *is a saddle point of* L.

[a] SLATER Constraint Qualification: The functions g_i are convex for all $i \in \mathscr{I}$ and there exists $\tilde{x} \in \mathscr{F}$ such that $g_i(\tilde{x}) < 0$ for $i \in \mathscr{I}_1$.

In the following example we see that the SLATER constraint qualification is essential in the above theorem.

Consider the following convex problem

$$\begin{cases} \text{minimize} & f(x) := -x \\ \text{subject to} & g(x) := x^2 \leq 0. \end{cases} \tag{9}$$

The only feasible point is $\bar{x} = 0$ with value $f(0) = 0$. So 0 minimizes $f(x)$ subject to $g(x) \leq 0$.

$L(x, \lambda) := -x + \lambda x^2$ for $\lambda \geq 0$, $x \in \mathbb{R}$. There is no $\bar{\lambda} \in [0, \infty)$ such that $(\bar{x}, \bar{\lambda})$ is a saddle point of L.

Can you find some other example?

Consider the following convex problem

$$\begin{cases} \text{minimize} & f(x) := -x \\ \text{subject to} & g(x) := x^2 \leq 0. \end{cases} \tag{9}$$

The only feasible point is $\bar{x} = 0$ with value $f(0) = 0$. So 0 minimizes $f(x)$ subject to $g(x) \leq 0$.

$L(x, \lambda) := -x + \lambda x^2$ for $\lambda \geq 0$, $x \in \mathbb{R}$. There is no $\bar{\lambda} \in [0, \infty)$ such that $(\bar{x}, \bar{\lambda})$ is a saddle point of L.

Can you find some other example?

Let $f : X \to \mathbb{R}$ and $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable functions with f and g_i ($\mathscr{I} := \{1, 2 \dots, m\}$) are convex,
and h_j ($\mathscr{J} := \{1, 2, \dots, \ell\}$) is (affine) linear.
Consider the following convex optimization problem:

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) = 0,\ i = 1, 2, \dots, m, \\ & h_j(x) = 0,\ j = 1, 2, \dots, \ell. \end{cases} \tag{CP}$$

Theorem

Suppose $\bar{x} \in \mathscr{F}$ *and there exist vector* $\lambda \in \mathbb{R}^m_+$ *and* $\mu \in \mathbb{R}^\ell$ *such that*

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

and

$$\lambda_i \nabla g_i(x) = 0, \quad \textit{for } i = 1, 2, \ldots, m,$$

then the problem (CP) *attains its global minimum at* $\bar{x}$*.*

The following example shows that even if we have convex problems the KKT conditions are not necessary for minimal points.

Example

Let $x = (x_1, x_2) \in \mathbb{R}^2$.

$$\begin{cases} \text{minimize} & f(x) := x_1 \\ \text{subject to} & g_1(x) := x_1^2 + (x_2 - 1)^2 - 1 \leq 0, \\ & g_2(x) := x_1^2 + (x_2 + 1)^2 - 1 \leq 0. \end{cases} \tag{CPE}$$

Obviously, $\bar{x} = (0, 0)$ is a feasible point. Hence, $\bar{x}$ is the (global) minimal point. Since $\nabla f(\bar{x}) = (1, 0)$, $\nabla g_1(\bar{x}) = (0, -2)$ and $\nabla g_2(\bar{x}) = (0, 2)$, the gradient condition of the KKT conditions is not met. f is linear, g_i's are convex. However, the SLATER condition is not fulfilled.

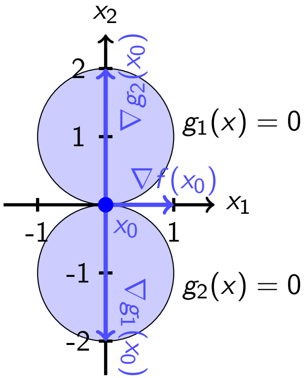
One can also argue in the following way that if $\mathcal{C}_l(x_0) \cap \mathcal{C}_{dd}(x_0) \neq \emptyset$ at $x_0 \in \mathscr{F}$, then KKT conditions do not hold. Clearly,

$$\mathcal{C}_{dd}(x_0) = \left\{ d = (d_1, d_2) \in \mathbb{R}^2 : \langle \nabla f(x_0), d \rangle < 0 \right\} = \{ d \in \mathbb{R}^2 : d_1 < 0 \}$$

and

$$\mathcal{C}_l(x_0) = \left\{ d \in \mathbb{R}^2 : \forall i \in \mathscr{A}(x_0),\ \langle \nabla g_i(x_0), d \rangle \leq 0 \right\} = \{ d \in \mathbb{R}^2 : d_2 = 0 \}$$

are not disjoint.



W. Forst and D. Hoffmann (2010). Optimization—theory and practice. Springer Science & Business Media. (Chapter 2)

M. S. Bazaraa, H. D. Sherali and C. M. Shetty (2013). *Nonlinear programming: theory and algorithms*. Third Edition. John Wiley & Sons. (Chapter 6)

N. Andréasson, A. Evgrafov, and M. Patriksson (2020). *An introduction to continuous optimization: foundations and fundamental algorithms*, Courier Dover Publications. (Chapter 6)

S. Boyd and L. Vandenberghe (2004). *Convex optimization*. Cambridge University Press. (Chapter 5)

W. S. Dorn (1960). Duality in quadratic programming. *Quarterly of Applied Mathematics*, 18(2), 155–162.

P. Wolfe (1961). A duality theorem for nonlinear programming, *Quarterly of Applied Mathematics*, 19, 239–244.

B. Mond (2009). Mond–Weir duality. Chapter 8, In: Optimization, C.

Questions?

&

Questions?

&

Thanks!

Questions?

&

Thanks!