

MATH3027: Optimization (UK 22/23)

Week 10: The KKT Optimality Conditions

Prof. Richard Wilkinson
School of Mathematical Sciences
University of Nottingham, United Kingdom
Please send any comments or mistakes to
r.d.wilkinson@nottingham.ac.uk

Last week we introduced the concept of stationarity, which we showed is a necessary optimality condition for problems with closed convex feasible sets. However, it can be difficult to prove whether stationarity is satisfied or not, and in practice it is not a useful way of solving optimization problems. This week we study an equivalent set of optimality conditions that are much easier to work with, known as the Karush-Kuhn-Tucker (KKT) conditions. These give us necessary and sufficient optimality conditions for linearly constrained problems. Before introducing the KKT conditions, we first need to introduce an *alternative* and a *separation* theorem, which will lead us to the KKT conditions.

| | |
|--|----|
| Reformulating optimization problems | 1 |
| Separation Theorem | 2 |
| KKT Conditions for Linearly Constrained Problems | 5 |
| Orthogonal Regression | 11 |
| Bonus questions | 13 |
| Summary | 13 |
| Checklist | 15 |

Reformulating optimization problems

The methods we study in this module rely upon functions being differentiable¹, either to characterize solutions, or to algorithms to make good choices for search directions. However, sometimes we can transform non-differentiable objectives to a set of smooth constraints.

¹ There are generalizations of some of the algorithms that use the sub-gradient instead for non-differentiable convex functions, but this beyond the scope of the module.



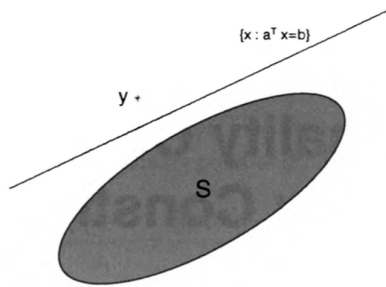


Figure 1: Illustration of a hyperplane separating point y from convex set S . If S is a closed convex set, then for any point $y \notin S$ we can always find a hyperplane separating y from S .



Consider the constraint $\|x\|_1 \leq 1$ where $x \in \mathbb{R}^2$. Show that this can be described by a set of four smooth constraints. Hint: one of these constraints is $x_1 + x_2 \leq 1$.



Consider the non-differentiable function $f(x) = \max(x^2, x)$, and the unconstrained optimization problem

$$\min_{x \in \mathbb{R}} f(x).$$

Show that this can be written as a smooth constrained optimization problem.

Separation Theorem

To begin our study of optimality conditions for linearly constrained problems we need first some technical results, known as *Alternative and Separation Theorems*.

A hyperplane is of the form

$$H = \{x \in \mathbb{R}^n : a^T x = b\} \quad (a \in \mathbb{R}^n \setminus \{0\}, b \in \mathbb{R}).$$

H is said to strictly separate a point $y \notin S$ from S if

$$a^T y > b,$$

and

$$a^T x \leq b \text{ for all } x \in S.$$

Theorem (Separation of a Point from a Closed and Convex Set). *Let $C \subseteq \mathbb{R}^n$ be a nonempty closed and convex set, and let $y \notin C$. Then there exists $p \in \mathbb{R}^n \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that $p^T y > \alpha$ and $p^T x \leq \alpha$ for all $x \in C$.*



Proof. By the second orthogonal projection theorem, the vector $\bar{\mathbf{x}} = P_C(\mathbf{y}) \in C$ satisfies

$$(\mathbf{y} - \bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \leq 0 \quad \text{for all } \mathbf{x} \in C,$$

which is the same as

$$(\mathbf{y} - \bar{\mathbf{x}})^\top \mathbf{x} \leq (\mathbf{y} - \bar{\mathbf{x}})^\top \bar{\mathbf{x}} \quad \text{for all } \mathbf{x} \in C.$$

Denote $\mathbf{p} = \mathbf{y} - \bar{\mathbf{x}} \neq \mathbf{0}$ and $\alpha = (\mathbf{y} - \bar{\mathbf{x}})^\top \bar{\mathbf{x}}$. Then,

$$\mathbf{p}^\top \mathbf{x} \leq \alpha \quad \text{for all } \mathbf{x} \in C.$$

On the other hand, we have

$$\mathbf{p}^\top \mathbf{y} = (\mathbf{y} - \bar{\mathbf{x}})^\top \mathbf{y} = (\mathbf{y} - \bar{\mathbf{x}})^\top (\mathbf{y} - \bar{\mathbf{x}}) + (\mathbf{y} - \bar{\mathbf{x}})^\top \bar{\mathbf{x}} = \|\mathbf{y} - \bar{\mathbf{x}}\|^2 + \alpha > \alpha$$

as $\mathbf{y} \neq \bar{\mathbf{x}}$. □

This theorem allows us to prove Farkas' Lemma, which says that exactly one of two alternative systems has a solution. We will rely upon Farkas' Lemma to derive the KKT conditions.

Lemma (Farkas' Lemma - an Alternative Theorem). *Let $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then exactly one of the following systems has a solution:*

I. $\mathbf{Ax} \leq \mathbf{0}, \mathbf{c}^\top \mathbf{x} > 0.$

II. $\mathbf{A}^\top \mathbf{y} = \mathbf{c}, \mathbf{y} \geq \mathbf{0}.$

An alternative way of writing this lemma is as follows:

Lemma (Farkas' Lemma - Second Formulation). *Let $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then the following two claims are equivalent:*

(A) $\mathbf{Ax} \leq \mathbf{0} \Rightarrow \mathbf{c}^\top \mathbf{x} \leq 0.$

(B) *There exists $\mathbf{y} \in \mathbb{R}_+^m$ such that $\mathbf{A}^\top \mathbf{y} = \mathbf{c}$.*

It is easy to see the equivalence of these two ways of writing the lemma. Condition (B) is the same as (II), and (A) is the converse of (I). The first formulation says not (I) if and only if (II), or equivalently, (A) if and only if (B), which is the second formulation.

Proof. Suppose that (B) is true, i.e., there exists $\mathbf{y} \in \mathbb{R}_+^m$ such that $\mathbf{A}^\top \mathbf{y} = \mathbf{c}$. We will show that this implies (A). Suppose that $\mathbf{Ax} \leq \mathbf{0}$ for some $\mathbf{x} \in \mathbb{R}^n$. Multiplying this inequality from the left by \mathbf{y}^\top we obtain:

$$\mathbf{y}^\top \mathbf{Ax} \leq 0 \quad (\text{as } \mathbf{y} \geq \mathbf{0}),$$

and hence,

$$\mathbf{c}^\top \mathbf{x} \leq 0.$$



So we have shown that (B) implies (A).

Conversely, now suppose that (A) is true. We will show that (A) implies (B) using proof by contradiction. Start by assuming that (B) is false, i.e., there does not exist $\mathbf{y} \in \mathbb{R}_+^m$ with $\mathbf{A}^\top \mathbf{y} = \mathbf{c}$. Consider the following closed and convex set

$$S = \{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{A}^\top \mathbf{y} \text{ for some } \mathbf{y} \in \mathbb{R}_+^m \},$$

where by assumption $\mathbf{c} \notin S$. By the separation theorem, there exists a hyperplane $\mathbf{p}^\top \mathbf{x} = \alpha$ that separates \mathbf{c} from S , i.e. there exists $\mathbf{p} \in \mathbb{R}^n \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that $\mathbf{p}^\top \mathbf{c} > \alpha$ and

$$\mathbf{p}^\top \mathbf{x} \leq \alpha \text{ for all } \mathbf{x} \in S.$$


Because $\mathbf{0} \in S$ we have that $\alpha \geq 0$. This implies that $\mathbf{p}^\top \mathbf{c} > 0$. Moreover, the inequality above is equivalent to

$$\mathbf{p}^\top \mathbf{A}^\top \mathbf{y} \leq \alpha \text{ for all } \mathbf{y} \geq \mathbf{0}$$

or to

$$(\mathbf{A}\mathbf{p})^\top \mathbf{y} \leq \alpha \text{ for all } \mathbf{y} \geq \mathbf{0}.$$

Therefore, $\mathbf{A}\mathbf{p} \leq \mathbf{0}$ as if any element of the vector $\mathbf{A}\mathbf{p}$ was greater than 0 we could set \mathbf{y} to be a vector of zeros with a positive non-zero value in the position corresponding to that element. Thus we have a vector \mathbf{p} such that $\mathbf{A}\mathbf{p} \leq \mathbf{0}$ and $\mathbf{c}^\top \mathbf{p} > 0$, and thus we have contradicted the assertion that (A) is true. Thus we see that (A) implies (B). \square

Example.  What does this mean for $\mathbf{A} = \begin{pmatrix} 1 & 5 \\ -1 & 2 \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} -1 \\ 9 \end{pmatrix}$?

Finally, we state another *Alternative* theorem.

Theorem (Gordan's Alternative Theorem). *Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then exactly one of the following two systems has a solution:*

- I. $\mathbf{A}\mathbf{x} < \mathbf{0}$.
- II. $\mathbf{p} \neq \mathbf{0}, \mathbf{A}^\top \mathbf{p} = \mathbf{0}, \mathbf{p} \geq \mathbf{0}$.

These results will be important in the following section where we derive the KKT conditions that can be used to solve constrained optimization problems.



KKT Conditions for Linearly Constrained Problems

Recall that if \mathbf{x}^* is a local minimum of a constrained optimization problem, then \mathbf{x}^* is a stationary point of the problem. However, proving stationarity is difficult in general. Instead we use the Karush-Kuhn-Tucker (or KKT) conditions, which are a set of fundamental characterizations of the solution of convex optimization problems. In this module we focus on the linearly constrained case, but the same conditions apply in the case where we have non-linear constraints (but we need an additional condition in this case as well).

Theorem (KKT conditions for Linearly Constrained Problems - Necessary Optimality Conditions). *Consider the minimization problem*

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to } \mathbf{a}_i^\top \mathbf{x} \leq b_i, \quad i = 1, 2, \dots, m \end{cases} \quad (\text{LCP})$$

where f is continuously differentiable over \mathbb{R}^n , $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m \in \mathbb{R}^n$, $b_1, b_2, \dots, b_m \in \mathbb{R}$. Let \mathbf{x}^* be a local minimum point of (LCP). Then, there exist $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \mathbf{a}_i = \mathbf{0}, \quad (1)$$

and

$$\lambda_i (\mathbf{a}_i^\top \mathbf{x}^* - b_i) = 0, \quad i = 1, 2, \dots, m. \quad (2)$$

Equations (1) and (2) are known as the **KKT conditions**. These can be seen as a generalization of the method of Lagrange multipliers that you have studied previously (it generalizes from only allowing equality constraints, to also allowing inequality constraints). Consequently the scalars $\lambda_1, \dots, \lambda_m$ that appear in the KKT conditions are sometimes called **Lagrange multipliers**. The equations (2) are often referred to as the **complimentary slackness conditions**.

Proof. If \mathbf{x}^* is a local minimum, this implies \mathbf{x}^* is a stationary point (as stationarity is a necessary condition for optimality), meaning

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0$$

for every $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{a}_i^\top \mathbf{x} \leq b_i$ for $i = 1, 2, \dots, m$. Now, denote the set of active constraints by

$$I(\mathbf{x}^*) = \{i : \mathbf{a}_i^\top \mathbf{x}^* = b_i\}.$$

Making the change of variables $\mathbf{y} = \mathbf{x} - \mathbf{x}^*$, we have $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$ for any $\mathbf{y} \in \mathbb{R}^n$ satisfying

$$\mathbf{a}_i^\top (\mathbf{y} + \mathbf{x}^*) \leq b_i, \quad i = 1, 2, \dots, m.$$

Or equivalently $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$ for any \mathbf{y} satisfying

$$\begin{aligned} \mathbf{a}_i^\top \mathbf{y} &\leq 0 & i \in I(\mathbf{x}^*) \\ \mathbf{a}_i^\top \mathbf{y} &\leq b_i - \mathbf{a}_i^\top \mathbf{x}^* & i \notin I(\mathbf{x}^*) \end{aligned} \quad (3)$$



Lemma. $\mathbf{a}_i^\top \mathbf{y} \leq 0$ for all $i \in I(\mathbf{x}^*) \Rightarrow \nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$.

We'll prove the lemma in a moment. Let's first use it to complete the proof of the theorem. If we enumerate the elements in $I(\mathbf{x}^*)$, i.e., suppose $I(\mathbf{x}^*) = \{i_1, \dots, i_k\}$, then if we write

$$\mathbf{A} = \begin{pmatrix} - & \mathbf{a}_{i_1}^\top & - \\ \vdots & \vdots & \vdots \\ - & \mathbf{a}_{i_k}^\top & - \end{pmatrix}$$

we can see that the lemma says

$$\mathbf{A}\mathbf{y} \leq 0 \Rightarrow \nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$$

or if we let $\mathbf{c} = -\nabla f(\mathbf{x}^*)$ it says

$$\mathbf{A}\mathbf{y} \leq 0 \Rightarrow \mathbf{c}^\top \mathbf{y} \leq 0.$$

Then by Farkas' Lemma we have that this is equivalent to

$$\exists \lambda \in \mathbb{R}_+^k \text{ such that } \mathbf{A}^\top \lambda = \mathbf{c}$$

or equivalently that

$$\sum_{i \in I(\mathbf{x}^*)} \lambda_i \mathbf{a}_i = -\nabla f(\mathbf{x}^*).$$

If we define $\lambda_i = 0$ for all $i \notin I(\mathbf{x}^*)$ we get

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \mathbf{a}_i = 0$$

and

$$\lambda_i (\mathbf{a}_i^\top \mathbf{x}^* - b_i) = 0 \text{ for all } i.$$

I.e., we have shown that the KKT conditions must hold (they are necessary conditions).

Finally, it just remains to prove the lemma. Suppose $\mathbf{a}_i^\top \mathbf{y} \leq 0$ for all $i \in I(\mathbf{x}^*)$. Since by definition $b_i - \mathbf{a}_i^\top \mathbf{x}^* > 0$ for $i \notin I(\mathbf{x}^*)$, there must exist $\alpha > 0$ such that

$$b_i - \mathbf{a}_i^\top \mathbf{x}^* > \alpha (\mathbf{a}_i^\top \mathbf{y}) \text{ for all } i \notin I(\mathbf{x}^*).$$

Thus, $\alpha \mathbf{y}$ meets the stationary conditions (3), and hence

$$\nabla f(\mathbf{x}^*)^\top (\alpha \mathbf{y}) \geq 0 \Rightarrow \nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$$

as required.

□

The previous theorem can be improved to necessary and sufficient conditions when f is convex.



Theorem (KKT Conditions for Convex Linearly Constrained Problems - Necessary and Sufficient Optimality Conditions). Consider the minimization problem (LCP) where in addition f is a convex continuously differentiable function over \mathbb{R}^n , and let \mathbf{x}^* be a feasible solution. Then \mathbf{x}^* is an optimal solution if and only if there exist $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \mathbf{a}_i = 0 \quad (4)$$

and

$$\lambda_i (\mathbf{a}_i^\top \mathbf{x}^* - b_i) = 0, \quad i = 1, 2, \dots, m \quad (5)$$

Proof. Necessity was already proven. For sufficiency, suppose that \mathbf{x}^* is a feasible solution of (LCP) satisfying the optimality conditions (4) and (5). Let \mathbf{x} be an arbitrary feasible solution of (LCP). Define the function

$$h(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i (\mathbf{a}_i^\top \mathbf{x} - b_i) .$$

Then $\nabla h(\mathbf{x}^*) = \mathbf{0}$. As h is convex, by the theorem that says convexity + stationarity implies global optimality for unconstrained problems, we can conclude that \mathbf{x}^* minimizes h over \mathbb{R}^n .

From here, it follows that

$$f(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i (\mathbf{a}_i^\top \mathbf{x}^* - b_i) = h(\mathbf{x}^*) \leq h(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i (\mathbf{a}_i^\top \mathbf{x} - b_i) \leq f(\mathbf{x}) .$$

The last inequality follows from the fact $\lambda \geq 0$ and $\mathbf{a}_i^\top \mathbf{x} \leq b_i$. Thus we have shown that \mathbf{x}^* is a minimizer of f . \square

We conclude by stating without proof the KKT conditions associated to linear problems with equality and inequality constraints.

Theorem (KKT conditions for Linearly Constrained Problems). Consider the minimization problem

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to} & \mathbf{a}_i^\top \mathbf{x} \leq b_i, \quad i = 1, 2, \dots, m \\ & \mathbf{c}_j^\top \mathbf{x} = d_j, \quad j = 1, 2, \dots, p \end{cases} \quad (\text{LCPI})$$

where f is continuously differentiable, $\mathbf{a}_i, \mathbf{c}_j \in \mathbb{R}^n, b_i, d_j \in \mathbb{R}$.

(i) (necessity of the KKT conditions) If \mathbf{x}^* is a local minimum of (LCPI) then there exist $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ and $\mu_1, \mu_2, \dots, \mu_p \in \mathbb{R}$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \mathbf{a}_i + \sum_{j=1}^p \mu_j \mathbf{c}_j = 0 \quad (6)$$

$$\lambda_i (\mathbf{a}_i^\top \mathbf{x}^* - b_i) = 0, \quad i = 1, 2, \dots, m . \quad (7)$$



(ii) (sufficiency in the convex case) If f is convex over \mathbb{R}^n and \mathbf{x}^* is a feasible solution of (LCPI) for which there exist $\lambda_1, \dots, \lambda_m \geq 0$ and $\mu_1, \dots, \mu_p \in \mathbb{R}$ such that the conditions are satisfied, then \mathbf{x}^* is an optimal solution of (LCPI).

A feasible point \mathbf{x}^* is called a KKT point if there exists Lagrange multipliers $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ and $\mu_1, \mu_2, \dots, \mu_p \in \mathbb{R}$ satisfying the KKT conditions (6) and (7). Note that if the feasibility constraints are all equality constraints (i.e. no constraints of the form $\mathbf{a}_i^\top \mathbf{x} \leq b_i$) then this result just restates the method of Lagrange multipliers that you saw in the first year.

Examples:  Solve the problem

$$\begin{aligned} \min \quad & \frac{1}{2} (x_1^2 + x_2^2 + x_3^2) \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 3. \end{aligned}$$



Let C be the affine space $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Use the KKT conditions to show that

$$P_C(\mathbf{y}) = \mathbf{y} - \mathbf{A}^\top (\mathbf{A}\mathbf{A}^\top)^{-1} (\mathbf{A}\mathbf{y} - \mathbf{b}).$$

Non-linear constraints - WILL NOT BE EXAMINED

The information in this section will not be tested in the exam, but may be used in the coursework.

In general (i.e., not just for linear constraints), given the problem

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m \\ h_j(\mathbf{x}) = 0, \quad j = 1, 2, \dots, p, \end{cases} \quad (\text{NLP})$$

we define the Lagrangian as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^p \mu_j h_j(\mathbf{x}).$$



The KKT equations are then

$$\begin{aligned}\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^p \mu_j \nabla h_j(\mathbf{x}^*) &= 0, \\ \lambda_i g_i(\mathbf{x}^*) &= 0, \quad i = 1, 2, \dots, m \\ g_i(\mathbf{x}) &\leq 0, \quad i = 1, \dots, m \\ h_j(\mathbf{x}) &= 0, \quad j = 1, \dots, p \\ \lambda_i &\geq 0.\end{aligned}$$

We say that a feasible point \mathbf{x}^* is a **KKT point** if there exists $\lambda_1, \dots, \lambda_m \geq 0$ and $\mu_1, \dots, \mu_p \in \mathbb{R}$ satisfying the KKT equations.

In the case where the constraints g_i and h_j are affine functions, we have shown that any minimum \mathbf{x}^* must satisfy the KKT equations, and that when f is convex, the KKT conditions are sufficient conditions for \mathbf{x}^* being a minimum. In general (i.e. for non-linear constraints), we additionally need a regularity condition to apply.

To define this condition we first need some terminology. Consider the set of inequalities

$$g_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m.$$

We say the active constraints at \mathbf{x} are the constraints satisfied as equalities at \mathbf{x} , denoted as

$$I(\mathbf{x}) = \{i : g_i(\mathbf{x}) = 0\}.$$

Definition. Consider optimization problem (NLP). A feasible point \mathbf{x}^* is called **regular** if the gradients of the active inequality constraints and the gradients of the equality constraints

$$\{\nabla g_i(\mathbf{x}^*) : i \in I(\mathbf{x}^*)\} \cup \{\nabla h_j(\mathbf{x}^*) : j = 1, 2, \dots, p\}$$

are linearly independent.

Theorem (Necessity of the KKT conditions). Consider the minimization problem (NLP) where $f, g_1, \dots, g_m, h_1, \dots, h_p$ are continuously differentiable functions. Suppose \mathbf{x}^* is a regular point of (NLP). If \mathbf{x}^* is a minimum point, then \mathbf{x}^* is a KKT point.



Solve

$$\begin{aligned}\min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 1.\end{aligned}$$

In the case where the objective f and the inequality constraints are convex functions, and the equality constraints are affine functions, then the KKT conditions are also sufficient conditions and we do not require regularity.



Theorem (Sufficiency of the KKT conditions in convex problems). Consider the minimization problem (NLP) where f, g_1, \dots, g_m are continuously differentiable convex functions, and where h_1, \dots, h_p are affine functions. If \mathbf{x}^* is a KKT point then \mathbf{x}^* is a minimum point of problem (NLP).

For convex optimization problems, there is a different condition than regularity that guarantees the necessity of the KKT conditions called **Slater's condition**.

Definition. Consider the set of convex inequalities

$$g_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m.$$

We say **Slater's condition** is satisfied if there exists $\hat{\mathbf{x}} \in \mathbb{R}^n$ such that

$$g_i(\mathbf{x}) < 0, \quad i = 1, \dots, m.$$

This essentially says that \mathbf{x} is an interior point of the feasible region.

Note that Slater's condition is much easier to check than regularity, as it only requires that there exists a point that strictly satisfies the constraints (whereas the regularity conditions require *a priori* knowledge of the candidate optimal solution).

Theorem (Necessity of the KKT conditions in convex problems under Slater's condition). Consider the minimization problem

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to } g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m \end{cases} \quad (\text{NLP2})$$

where f, g_1, \dots, g_m are continuously differentiable convex functions. If Slater's condition is satisfied, then if \mathbf{x}^* is a minimum point of problem (NLP2) it is also a KKT point.



Solve

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 \leq 1. \end{aligned}$$

Note that this is the problem we solved in week 1 using an ad hoc method.

More generally, for optimization problems with a mixture of constrain types such as

$$\begin{aligned} g_i(\mathbf{x}) &\leq 0, & i = 1, 2, \dots, m \\ h_j(\mathbf{x}) &\leq 0, & j = 1, 2, \dots, p \\ s_k(\mathbf{x}) &= 0, & k = 1, \dots, q \end{aligned}$$

where g_i are convex functions, and h_j and s_k are affine functions, then we say that the generalized Slater's condition is satisfied if there exists $\hat{\mathbf{x}} \in \mathbb{R}^n$ with

$$\begin{aligned} g_i(\mathbf{x}) &< 0, & i = 1, 2, \dots, m \\ h_j(\mathbf{x}) &\leq 0, & j = 1, 2, \dots, p \\ s_k(\mathbf{x}) &= 0, & k = 1, \dots, q \end{aligned}$$



Theorem (Necessity of the KKT conditions under the generalized Slater's condition).
Consider the minimization problem

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to} & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, m \\ & h_j(\mathbf{x}) \leq 0, \quad j = 1, 2, \dots, p \\ & s_k(\mathbf{x}) = 0, \quad k = 1, \dots, q \end{cases} \quad (\text{NLP3})$$

where f, g_1, \dots, g_m are continuously differentiable convex functions, and h_j, s_k are affine. If the generalized Slater's condition is satisfied, then if \mathbf{x}^* is a minimum point of problem (NLP3) it is also a KKT point, i.e., there exists $\lambda_i \geq 0, \eta_j \geq 0$ and $\mu_k \in \mathbb{R}$ such that

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^p \eta_j \nabla h_j(\mathbf{x}^*) + \sum_{k=1}^q \mu_k \nabla s_k(\mathbf{x}^*) &= 0, \\ \lambda_i g_i(\mathbf{x}^*) &= 0, \quad i = 1, 2, \dots, m \\ \eta_j h_j(\mathbf{x}^*) &= 0, \quad j = 1, 2, \dots, p. \end{aligned}$$

We will see an application of these results in the computer lab.

Orthogonal Regression

We conclude by studying a statistical problem known as orthogonal regression. We begin by developing some results about projections onto hyperplanes.

Orthogonal Projection onto Hyperplanes

Consider the hyperplane

$$H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\} \quad \text{where } \mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R}.$$

Then by the previous result about the projection onto an affine space we have

$$P_H(\mathbf{y}) = \mathbf{y} - \mathbf{a} (\mathbf{a}^\top \mathbf{a})^{-1} (\mathbf{a}^\top \mathbf{y} - b) = \mathbf{y} - \frac{\mathbf{a}^\top \mathbf{y} - b}{\|\mathbf{a}\|^2} \mathbf{a}.$$

Lemma (distance of a point from a hyperplane). Let $H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\}$, where $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then

$$d(\mathbf{y}, H) = \frac{|\mathbf{a}^\top \mathbf{y} - b|}{\|\mathbf{a}\|}$$



Proof.

$$d(\mathbf{y}, H) = \|\mathbf{y} - P_H(\mathbf{y})\| = \left\| \mathbf{y} - \left(\mathbf{y} - \frac{\mathbf{a}^\top \mathbf{y} - b}{\|\mathbf{a}\|^2} \mathbf{a} \right) \right\| = \frac{|\mathbf{a}^\top \mathbf{y} - b|}{\|\mathbf{a}\|}$$

□

Similarly, it follows the computation of the orthogonal projection onto half-spaces.

Lemma. Let $H^- = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} \leq b\}$ where $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then

$$P_{H^-}(\mathbf{x}) = \mathbf{x} - \frac{[\mathbf{a}^\top \mathbf{x} - b]_+}{\|\mathbf{a}\|^2} \mathbf{a}.$$

The Orthogonal Regression Problem

For a given $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$, we define the hyperplane:

$$H_{\mathbf{x}, y} := \{\mathbf{a} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{a} = y\}.$$

Suppose we are given data points $\mathbf{a}_1, \dots, \mathbf{a}_m$. In the orthogonal regression problem, we seek to find a nonzero vector $\mathbf{x} \in \mathbb{R}^n$ and $y \in \mathbb{R}$ such that the sum of squared Euclidean distances between the points $\mathbf{a}_1, \dots, \mathbf{a}_m$ to $H_{\mathbf{x}, y}$ is minimal:

$$\min_{\mathbf{x}, y} \left\{ \sum_{i=1}^m d(\mathbf{a}_i, H_{\mathbf{x}, y})^2 : \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R} \right\}.$$

We begin by noting that $d(\mathbf{a}_i, H_{\mathbf{x}, y})^2 = \frac{(\mathbf{a}_i^\top \mathbf{x} - y)^2}{\|\mathbf{x}\|^2}$, for $i = 1, \dots, m$. Therefore, the Orthogonal Regression problem is the same as

$$\min \left\{ \sum_{i=1}^m \frac{(\mathbf{a}_i^\top \mathbf{x} - y)^2}{\|\mathbf{x}\|^2} : \mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R} \right\}.$$

Fixing \mathbf{x} and minimizing first with respect to y we obtain that the optimal y is given by $y = \frac{1}{m} \sum_{i=1}^m \mathbf{a}_i^\top \mathbf{x} = \frac{1}{m} \mathbf{e}^\top \mathbf{A} \mathbf{x}$. Using the above expression for y we obtain that

$$\begin{aligned} \sum_{i=1}^m (\mathbf{a}_i^\top \mathbf{x} - y)^2 &= \sum_{i=1}^m \left(\mathbf{a}_i^\top \mathbf{x} - \frac{1}{m} \mathbf{e}^\top \mathbf{A} \mathbf{x} \right)^2 \\ &= \sum_{i=1}^m (\mathbf{a}_i^\top \mathbf{x})^2 - \frac{2}{m} \sum_{i=1}^m (\mathbf{e}^\top \mathbf{A} \mathbf{x}) (\mathbf{a}_i^\top \mathbf{x}) + \frac{1}{m} (\mathbf{e}^\top \mathbf{A} \mathbf{x})^2 \\ &= \sum_{i=1}^m (\mathbf{a}_i^\top \mathbf{x})^2 - \frac{1}{m} (\mathbf{e}^\top \mathbf{A} \mathbf{x})^2 = \|\mathbf{A} \mathbf{x}\|^2 - \frac{1}{m} (\mathbf{e}^\top \mathbf{A} \mathbf{x})^2 \\ &= \mathbf{x}^\top \mathbf{A}^\top \left(\mathbf{I}_m - \frac{1}{m} \mathbf{e} \mathbf{e}^\top \right) \mathbf{A} \mathbf{x}. \end{aligned}$$



Therefore, a reformulation of the problem is

$$\min_{\mathbf{x}} \left\{ \frac{\mathbf{x}^\top \left[\mathbf{A}^\top \left(\mathbf{I}_m - \frac{1}{m} \mathbf{e} \mathbf{e}^\top \right) \mathbf{A} \right] \mathbf{x}}{\|\mathbf{x}\|^2} : \mathbf{x} \neq \mathbf{0} \right\}.$$



From this, show that an optimal solution of the orthogonal regression problem (\mathbf{x}, y) is to take \mathbf{x} to be an eigenvector of associated with the minimum eigenvalue of $\mathbf{A}^\top \left(\mathbf{I}_m - \frac{1}{m} \mathbf{e} \mathbf{e}^\top \right) \mathbf{A}$, and $y = \frac{1}{m} \sum_{i=1}^m \mathbf{a}_i^\top \mathbf{x}$. The optimal function value of the problem is $\lambda_{\min} \left[\mathbf{A}^\top \left(\mathbf{I}_m - \frac{1}{m} \mathbf{e} \mathbf{e}^\top \right) \mathbf{A} \right]$.

Orthogonal regression is a special case of total least squares regression, which is used when there are errors on the covariates and response variables (least squares only considers errors on the response).

Bonus questions



Find the optimal solution of the problem

$$\min \{ \mathbf{x}^\top \mathbf{Q} \mathbf{x} + 2 \mathbf{c}^\top \mathbf{x} : \mathbf{A} \mathbf{x} = \mathbf{b} \}$$

where \mathbf{Q} is a symmetric positive definite matrix.



Consider the problem

$$\begin{aligned} & \min -x_1 x_2 x_3 \\ & \text{subject to } x_1 + 3x_2 + 6x_3 \leq 48 \\ & x_1, x_2, x_3 \geq 0. \end{aligned}$$

Summary

For linearly constrained optimization problems

- the KKT conditions are always necessary,
- for convex optimization problems with linear constraints, the KKT conditions are also sufficient.

For optimization problems with non-linear constraints,



- the KKT conditions are necessary for regular points, and for convex problems, the KKT conditions are necessary if Slater's condition is satisfied.
- for convex problems, the KKT conditions are sufficient conditions.



Checklist

The idea of this checklist is to help you to self-evaluate your progress and understanding of the subject, and to give you some guidance on where to focus. If you can tick all the boxes it means you're doing alright, otherwise you need to study a bit more, grab a book, watch the videos, or seek help from classmates, the lecturers, or the demonstrators. Try to fill as many gaps as quickly as possible.



And remember to do the 's!

| Learning Outcome | Check |
|--|-------|
| I can formulate the KKT system for equality and inequality linear constraints. | |
| I can solve the KKT system. | |
| I understand how to use convexity for sufficiency. | |
| I can compute orthogonal projections using the KKT conditions. | |
| I understand the statement of the orthogonal regression problem. | |