14

Karush-Kuhn-Tucker Conditions

For convex problems of the form

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{w}) \\ \text{subject to} \quad & \boldsymbol{f}(\boldsymbol{w}) \leq \boldsymbol{0} \\ & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}, \end{aligned} \tag{P}$$

we introduced the Lagrangian $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ and defined the Lagrange dual as

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{w} \in \mathcal{D}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

We saw that $g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a lower bound on the optimal value of (P). Note that we wrote the linear equality conditions $h_j(\boldsymbol{w}) = 0$ as system of linear equations $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$. We will derive conditions under which the lower bound provided by the Lagrange dual matches the upper bound, and derive a system of equations and inequalities that certify optimality, the Karush-Kuhn-Tucker (KKT) conditions. These conditions can be seen as generalizations of the first-order optimality conditions to the setting when equality and inequality constraints are present.

Constraint qualification

Let p^* and d^* denote the primal and dual optimal values, so that

$$d^* = \sup_{\boldsymbol{\lambda} \geq 0, \boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \inf_{\boldsymbol{w} \in \mathcal{D}} \{f(\boldsymbol{w}) \mid f_i(\boldsymbol{w}) \leq 0,\ \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}\} = p^*.$$

Once certain conditions, called **constraint qualifications**, hold, we can ensure that **strong duality** holds, which means $d^* = p^*$. One particular such constraint qualification is Slater's Theorem.

Theorem 14.1. *(Slater conditions) Assume that the interior of the domain $\mathcal{D}$ of* (P) *is non-empty, that the problem* (P) *is convex, and that there exists a point $\boldsymbol{w} \in \mathcal{D}$ such that*

$$f_i(\boldsymbol{w}) < 0,\ 1 \leq i \leq m, \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b},\ 1 \leq j \leq p.$$

Assume that $\boldsymbol{A}$ has maximal rank. Then $d^* = p^*$.

The proof makes use of a fundamental result on convex sets, the **separating hyperplane theorem**. For an affine hyperplane $H = \{\boldsymbol{w} \colon \boldsymbol{a}^T\boldsymbol{w} + b = 0\}$, we denote by $H_+ = \{\boldsymbol{w} \colon \boldsymbol{a}^\top\boldsymbol{w} + b \geq 0\}$ one of the two closed halfspaces defined by the hyperplane, and by H_- the other.

Theorem 14.2 (Separating Hyperplane Theorem)**.** *Let C and D be disjoint, nonempty convex subsets of $\mathbb{R}^d$. Then there exists an affine hyperplane H such that $C \subset H_+$ and $D \subset H_-$.*

A hyperplane with the properties of Theorem 14.2 is called a **separating hyperplane**. The hyperplane separates the two sets strictly, if each of the sets is contained in the interior of the respective halfspaces. It can be shown that for *compact* convex sets, there exists a hyperplane separating them strictly.

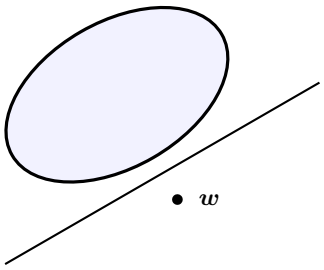

Figure 14.1: A hyperplane separating a convex sets and a point

Proof of Theorem 14.1. Assume $\boldsymbol{A}$ has rank p (the number of rows). Assume moreover that p^* is finite, since if $p^* = -\infty$, then by weak duality we already have $d^* = p^*$. Define the convex set

$$\mathcal{A} = \{(\boldsymbol{u}, \boldsymbol{v}, t) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid \forall \boldsymbol{w} \in \mathcal{D}, f_i(\boldsymbol{w}) \leq u_i, \boldsymbol{A}\boldsymbol{w} - \boldsymbol{b} = \boldsymbol{v}, f(\boldsymbol{w}) \leq t\}.$$

Then the optimal value of (P) is

$$p^* = \inf\{t \mid (\boldsymbol{0}, \boldsymbol{0}, t) \in \mathcal{A}\}.$$

Define the convex set $\mathcal{B}$ as

$$\mathcal{B} = \{(\boldsymbol{0}, \boldsymbol{0}, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} \mid s < p^*\}.$$

The sets $\mathcal{A}$ and $\mathcal{B}$ are disjoint. To see this, assume to the contrary that there is a point $\boldsymbol{z} \in \mathcal{A} \cap \mathcal{B}$. Then since $\boldsymbol{z} \in \mathcal{B}$, $\boldsymbol{z} = (\boldsymbol{0}, \boldsymbol{0}, s)$ with $s < p^*$, but also since $\boldsymbol{z} \in \mathcal{A}$, ther exists an $\boldsymbol{w} \in \mathcal{D}$ with $f_i(\boldsymbol{w}) \leq 0$, $\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b} = \boldsymbol{0}$, and $f(\boldsymbol{w}) \leq s < p^*$, in contradiction to the optimality of p^*.

By the separating hyperplane theorem, there exists a hyperplane separating $\mathcal{A}$ and $\mathcal{B}$ (but not necessarily strictly!), defined by a vector $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}}, \nu) \neq \mathbf{0}$ and $\alpha \neq 0$ with the property that

$$(\boldsymbol{u}, \boldsymbol{v}, t) \in \mathcal{A} \Longrightarrow \tilde{\boldsymbol{\lambda}}^\top \boldsymbol{u} + \tilde{\boldsymbol{\mu}}^\top \boldsymbol{v} + \nu t \geq \alpha \tag{14.1}$$

and

$$(\boldsymbol{u}, \boldsymbol{v}, t) \in \mathcal{B} \Longrightarrow \tilde{\boldsymbol{\lambda}}^\top \boldsymbol{u} + \tilde{\boldsymbol{\mu}}^\top \boldsymbol{v} + \nu t \leq \alpha. \tag{14.2}$$

If $\tilde{\boldsymbol{\lambda}} < \mathbf{0}$ or $\nu < 0$, we could find values of $\boldsymbol{u}$ and t for which $(\boldsymbol{u}, \boldsymbol{v}, t) \in \mathcal{A}$, but that make the right-hand side of (14.1) arbitrary small, contradicting the bound by α. It follows that $\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}$ and $\nu \geq 0$. Condition (14.2) simply means that $\nu t \leq \alpha$ for all $t < p^*$, so that $\nu p^* \leq \alpha$. Combining this bound with (14.1), we get the two inequalities, valid for any $\boldsymbol{w} \in \mathcal{D}$,

$$\sum_{i=1}^{m} \tilde{\lambda}_i f_i(\boldsymbol{w}) + \tilde{\boldsymbol{\mu}}^\top (\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}) + \nu f(\boldsymbol{w}) \geq \alpha \geq \nu p^*. \tag{14.3}$$

Note that the left-hand side has the form of a Lagrangian function scaled by ν. If $\nu > 0$ we can divide by ν and set $\lambda_i = \tilde{\lambda}_i / \nu$, $\mu_i = \tilde{\mu}_i / \nu$, to obtain

$$\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \geq p^*.$$

By weak duality we have $p^* \geq g(\boldsymbol{\lambda}, \boldsymbol{\mu})$, so that we get strong duality if $\nu > 0$. If $\nu = 0$, then (14.3) implies

$$\sum_{i=1}^{m} \tilde{\lambda}_i f_i(\boldsymbol{w}) + \tilde{\boldsymbol{\mu}}^\top (\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}) \geq 0, \tag{14.4}$$

and for a point $\tilde{\boldsymbol{w}}$ satisfying the conditions $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$ and $f_i(\boldsymbol{w}) < 0$ for $1 \leq i \leq m$, this means that $\tilde{\boldsymbol{\lambda}} = \mathbf{0}$. As $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\mu}}, \nu) \neq \mathbf{0}$, we have $\boldsymbol{\mu} \neq \mathbf{0}$. Since $\tilde{\boldsymbol{\mu}}^\top (\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}) \geq 0$ and $= 0$ for some $\tilde{\boldsymbol{w}}$ in the interior of $\mathcal{D}$, there must be a $\boldsymbol{w}$ in the interior such that $\tilde{\boldsymbol{\mu}}^\top (\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}) < 0$, in contradiction to (14.4) (unless $\tilde{\boldsymbol{\mu}}^\top \boldsymbol{A} = \mathbf{0}$, which would contradict the condition that $\boldsymbol{A}$ has maximal rank p). □

Example 14.3. Consider a linear programming problem of the form

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^\top \boldsymbol{w} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \\ & \boldsymbol{w} \geq \mathbf{0}. \end{aligned}$$

The inequality constraints are $-w_i \leq 0$, while the equality constraints are $\boldsymbol{a}_i^\top \boldsymbol{w} - b_i$. The Lagrangian has the form

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \boldsymbol{c}^\top \boldsymbol{w} - \sum_{i=1}^{m} \lambda_i w_i + \sum_{j=1}^{p} \mu_j (\boldsymbol{a}_j^\top \boldsymbol{w} - b_j) \\ &= (\boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu})^\top \boldsymbol{w} - \boldsymbol{b}^\top \boldsymbol{\mu}. \end{aligned}$$

The infimum over $\boldsymbol{w}$ of this function is $-\infty$ unless $\boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0}$. The Lagrange dual is therefore

$$g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{\mu}^\top \boldsymbol{b} & \text{if } \boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0} \\ -\infty & \text{else.} \end{cases}$$

From the previous chapter we conclude that

$$\max\{-\boldsymbol{b}^\top \boldsymbol{\mu} \mid \boldsymbol{c} - \boldsymbol{\lambda} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0},\ \boldsymbol{\lambda} \geq \boldsymbol{0}\} \leq \min\{\boldsymbol{c}^\top \boldsymbol{w} \mid \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b},\ \boldsymbol{w} \geq \boldsymbol{0}\}.$$

Note that if we write $\boldsymbol{v} = -\boldsymbol{\mu}$ and $\boldsymbol{s} = \boldsymbol{\lambda}$, then we get the dual version of the linear programming problem we started out with, and in this case it is known that

$$\max_{\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{\mu}} g(\boldsymbol{\lambda}, \boldsymbol{\mu}) = p^*.$$

A more common way to express **linear programming duality** is as follows: the optimal value of

$$\text{maximize} \quad \boldsymbol{b}^\top \boldsymbol{v} \quad \text{subject to} \quad \boldsymbol{A}^\top \boldsymbol{v} \leq \boldsymbol{c} \tag{D}$$

coincides with the optimal value of

$$\text{minimize} \quad \boldsymbol{c}^\top \boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b},\ \boldsymbol{w} \geq \boldsymbol{0}, \tag{P}$$

provided both (D) and (P) have a finite solution (here, we set $\boldsymbol{v} = -\boldsymbol{\mu}$).

Example 14.4. Consider the problem

$$\text{minimize } \|\boldsymbol{w}\|^2 \quad \text{subject to} \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}.$$

Note that $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^\top \boldsymbol{w}$. The Lagrangian is $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\mu}) = \|\boldsymbol{w}\|^2 + \boldsymbol{\mu}^\top (\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})$. For any $\boldsymbol{\mu}$, we can find the infimum

$$g(\boldsymbol{\mu}) = \inf_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\mu})$$

by setting the derivative of the Lagrangian to $\boldsymbol{w}$ to zero:

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, \boldsymbol{\mu}) = 2\boldsymbol{w} + \boldsymbol{A}^\top \boldsymbol{\mu} = \boldsymbol{0},$$

which gives the solution

$$\boldsymbol{w} = -\frac{1}{2} \boldsymbol{A}^\top \boldsymbol{\mu}.$$

The dual function is therefore

$$g(\boldsymbol{\mu}) = -\frac{1}{4} \boldsymbol{\mu}^\top \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{\mu} - \boldsymbol{b}^\top \boldsymbol{\mu}.$$

As the negative of a positive semidefinite quadratic function, it is concave. Moreover, we get the lower bound

$$-\frac{1}{4} \boldsymbol{\mu}^\top \boldsymbol{A}^\top \boldsymbol{A} \boldsymbol{\mu} - \boldsymbol{b}^\top \boldsymbol{\mu} \leq \inf\{\|\boldsymbol{w}\|^2 \mid \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}\}.$$

The problem we started out with is convex, and if we assume that there exists a feasible primal point, then the above inequality is in fact an equality by Slater's conditions.

Karush-Kuhn-Tucker optimality conditions

Consider now a not necessarily convex problem of the form

$$\begin{array}{ll} \text{minimize} & f(\boldsymbol{w}) \\ \text{subject to} & \boldsymbol{f}(\boldsymbol{w}) \leq \boldsymbol{0} \\ & \boldsymbol{h}(\boldsymbol{w}) = \boldsymbol{0}, \end{array} \tag{14.5}$$

If p^* is the optimal solution of (14.5) and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ dual variables, then we have seen that (this holds even in the non-convex case)

$$p^* \geq g(\boldsymbol{\lambda}, \boldsymbol{\mu}).$$

From this is follows that for any primal feasible point $\boldsymbol{w}$,

$$f(\boldsymbol{w}) - p^* \leq f(\boldsymbol{w}) - g(\boldsymbol{\lambda}, \boldsymbol{\mu}).$$

The difference $f(\boldsymbol{w}) - g(\boldsymbol{\lambda}, \boldsymbol{\mu})$ between the primal objective function at a primal feasible point and the dual objective function at a dual feasible point is called the **duality gap** at $\boldsymbol{w}$ and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$. For any such points we know that

$$p^*, q^* \in [g(\boldsymbol{\lambda}, \boldsymbol{\mu}), f(\boldsymbol{w})],$$

and if the gap is small we have a good approximation of the primal and dual optimal values. The duality gap can be used in iterative algorithms to define stopping criteria: if the algorithm generates a sequence of primal-dual variables $(\boldsymbol{w}^k, \boldsymbol{\lambda}^k, \boldsymbol{\mu}^k)$, then we can stop if the duality gap is less than, say, a predefined tolerance ε.

Now suppose that $(\boldsymbol{w}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is such that the duality gap is zero. Then

$$\begin{aligned} f(\boldsymbol{w}^*) = g(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) &= \inf_{\boldsymbol{w}} \left(f(\boldsymbol{w}) + \sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{w}) + \sum_{j=1}^{p} \mu_j^* h_j(\boldsymbol{w}) \right) \\ &\leq f(\boldsymbol{w}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{w}^*) + \sum_{j=1}^{p} \mu_j^* h_j(\boldsymbol{w}^*) \\ &\leq f(\boldsymbol{w}^*), \end{aligned}$$

where the last inequality follows from the fact that $h_j(\boldsymbol{w}^*) = 0$ and $\lambda_i^* f_i(\boldsymbol{w}^*) \leq 0$ for $1 \leq j \leq p$ and $1 \leq i \leq m$. It follows that the inequalities are in fact equalities. From the identity

$$f(\boldsymbol{w}^*) = f(\boldsymbol{w}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\boldsymbol{w}^*)$$

and $\lambda_i^* \geq 0$ and $f_i(\boldsymbol{w}^*) \leq 0$ we also conclude that at such optimal points,

$$\lambda_i^* f_i(\boldsymbol{w}^*) = 0, \quad 1 \leq i \leq m.$$

This condition is known as **complementary slackness**. From the above we also see that $\boldsymbol{w}^*$ minimizes the Lagrangian $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, so that

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0}.$$

Collecting these conditions (primal and dual feasibility, complementary slackness, vanishing gradient), we arrive at a set of optimality conditions known as the Karush-Kuhn-Tucker (KKT) conditions.

Theorem 14.5. *(KKT conditions) Let $\boldsymbol{w}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be primal and dual optimal solutions of* (14.5) *with zero duality gap. The the following conditions are satisfied:*

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{w}^*) &\leq \mathbf{0} \\ \boldsymbol{h}(\boldsymbol{w}^*) &= \mathbf{0} \\ \boldsymbol{\lambda}^* &\geq \mathbf{0} \\ \lambda_i^* f_i(\boldsymbol{w}^*) &= 0, \ 1 \leq i \leq m \\ \nabla_{\boldsymbol{w}} f(\boldsymbol{w}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_{\boldsymbol{w}} f_i(\boldsymbol{w}^*) + \sum_{j=1}^{p} \mu_j^* \nabla_{\boldsymbol{w}} h_j(\boldsymbol{w}^*) &= \mathbf{0}. \end{aligned}$$

Moreover, if the problem is convex and the Slater Conditions (Theorem 14.1) are satisfied, then any points satisfying the KKT conditions have zero duality gap.

Notes

The Karush-Kuhn-Tucker conditions were introduced by Kuhn and Tucker [1], and the necessity was shown by William Karush in his 1939 MSc thesis at the University of Chicago. These conditions generalize and unify different previously known results: the optimality conditions for linear programming based on linear programming duality, and Lagrange duality in the presence of inequalities. The KKT conditions also form the basis of algorithms for non-linear optimization, such as interior-point methods. A detailed treatment can be found in [2]. The separating hyperplane theorem is a central result in convexity, and lies at the heart of many duality results. In particular, it is the basis of a classic "theorem of the alternative" known as Farkas' Lemma, which states that given a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{b} \in \mathbb{R}^m$, there exists a vector $\boldsymbol{w}$ such that

$$\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}, \quad \boldsymbol{w} \geq 0$$

if and only if there is no $\boldsymbol{v} \in \mathbb{R}^m$ such that

$$\boldsymbol{A}^\top \boldsymbol{v} \geq \mathbf{0}, \quad \boldsymbol{v}^\top \boldsymbol{b} < 0.$$

This result, in turn, is an ingredient for deriving linear programming duality.

[1] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2014.

[2] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006.