Kuhn-Tucker-Lagrange conditions: basics

This handout contains all you really need to know about KT in order to be able to solve problems.

General non-linear optimisation problem: let $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$. Namely, $f(\boldsymbol{x})$ is an objective function, and the notation $\boldsymbol{g}(\boldsymbol{x}) = (g_1(\boldsymbol{x}), \ldots, g_m(\boldsymbol{x}))$ embraces the constraint functions, with $\boldsymbol{x} = (x_1, \ldots, x_n)$. All the functions are smooth. Consider the problem

$$\text{Min } f(\boldsymbol{x}) \text{ such that } \boldsymbol{g}(\boldsymbol{x}) \geq 0. \tag{1}$$

Let

$$F = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{g}(\boldsymbol{x}) \geq 0\}$$

be the feasible set for (1).

The problem is handled via the Lagrange multipliers method. The key difference will be now that due to the fact that the constraints are formulated as inequalities, Lagrange multipliers will be non-negative. Kuhn-Tucker conditions, henceforth KT, are the necessary conditions for some feasible $\boldsymbol{x}$ to be a local minimum for the optimisation problem (1).

In general, one can proclaim the following alternative: either $\boldsymbol{x}$ is a local minimum or it is not. Let's call the former side of the alternative ($\boldsymbol{x}$ is a local minimum) positive, and the latter side (it is not local minimum) negative. If the positive side of the alternative is true, then the following scenario cannot happen.

There cannot exist a curve γ, emanating from $\boldsymbol{x}$ and contained in the feasible set F – let us refer to γ as a *feasible curve* beginning at $\boldsymbol{x}$ – such that $f(\boldsymbol{x})$ decreases along this curve. In particular, if $\boldsymbol{v}$ is the tangent vector to the curve γ at its initial point $\boldsymbol{x}$, then the directional derivative of f in the direction $\boldsymbol{v}$ cannot be negative. Indeed, otherwise, arbitrarily closely to $\boldsymbol{x}$ in F there will be points $\boldsymbol{x}'$, where $f(\boldsymbol{x}')$ is smaller than $f(\boldsymbol{x})$.

Given $\boldsymbol{x}$, let us introduce the set of *True Feasible Directions* at $\boldsymbol{x}$ as the set of all vectors $\boldsymbol{v}$, such that there exists a feasible curve γ, beginning at $\boldsymbol{x}$, and such that $\boldsymbol{v}$ is the tangent to γ at $\boldsymbol{x}$. Denote this set $TFD(\boldsymbol{x})$. So the set $TFD(\boldsymbol{x})$ is just the set of tangent vectors at $\boldsymbol{x}$ to all feasible curves beginning at $\boldsymbol{x}$.

Also, given $\boldsymbol{x}$, let us say that the ith constraint *matters* at $\boldsymbol{x}$ if it is tight at $\boldsymbol{x}$, i.e $g_i(\boldsymbol{x}) = 0$.

Then if $\boldsymbol{v} \in TFD(\boldsymbol{x})$ and the ith constraint matters at $\boldsymbol{x}$, one must have

$$\boldsymbol{v} \cdot \nabla g_i(\boldsymbol{x}) \geq 0. \tag{2}$$

Otherwise, if it were < 0, there will be points in F – on any feasible curve γ, to which $\boldsymbol{v}$ is tangent at $\boldsymbol{x}$ – where $g_i < 0$, which contradicts the notion of feasibility.

The definition of the set of *True Feasible Directions* is geometrically clear, but it is not at all clear how it can be put into formulae. One would like to use (2) instead. So let us call the set of all $\boldsymbol{v}$, such that for all constraints that matter in $\boldsymbol{x}$, they satisfy (2) the set of *Feasible Directions* at $\boldsymbol{x}$. Denote this set $FD(\boldsymbol{x})$. What we've shown so far is that $TFD(\boldsymbol{x}) \subseteq FD(\boldsymbol{x})$: a *true* feasible direction is always a feasible direction.

Just like the Lagrange multipliers' under equality constraints theorem, KT conditions will work only under the non-degeneracy assumption. This assumption is $TFD(\boldsymbol{x}) = FD(\boldsymbol{x})$, rather than $\subset$. This assumption is called *Constraint Qualification*, in short CQ. So, if CQ is satisfied, the method below will work. If CQ is not satisfied – then it may fail.

Let us now formulate the theorem and elaborate on it.

`Theorem (Kuhn-Tucker)` If $\boldsymbol{x}$ is a local minimum for the optimisation problem (1) and CQ is satisfied at $\boldsymbol{x}$, then the gradient $\nabla f(\boldsymbol{x})$ must be represented as a linear combination of the gradients of the constraints $g_i(\boldsymbol{x})$ *that matter* (are tight) at $\boldsymbol{x}$, *with non-negative coefficients.*

These coefficients are called, once again, Lagrange multipliers. To eliminate "constraints that matter" notion from the formulation, observe that if we can just set $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ (assign a Lagrange multiplier to each constraint) and then require

$$\boldsymbol{\lambda} \cdot \boldsymbol{g}(\boldsymbol{x}) = 0, \text{ i.e. } \lambda_1 g_1(\boldsymbol{x}) + \ldots + \lambda_m g_m(\boldsymbol{x}) = 0.$$

Then we can only have $\lambda_i \neq 0$ when $g_i(\boldsymbol{x}) = 0$ (tight), while as soon as $g_i(x) > 0$ we may not have $\lambda_i > 0$, because this will never give us zero in the right-hand side above. Therefore, we can reformulate the theorem as follows.

`Theorem (Kuhn-Tucker, reformulated)` If $\boldsymbol{x}$ is a local minimum for the optimisation problem (1) and CQ is satisfied at $\boldsymbol{x}$, then $\boldsymbol{x}$ must satisfy the following system of equations-inequalities:

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= \lambda_1 \nabla g_1(\boldsymbol{x}) + \ldots + \lambda_m \nabla g_m(\boldsymbol{x}), \\ 0 &= \lambda_1 g_1(\boldsymbol{x}) + \ldots + \lambda_m g_m(\boldsymbol{x}), \\ 0 &\leq \boldsymbol{g}(\boldsymbol{x}), \end{aligned} \tag{3}$$

with $\boldsymbol{\lambda} \in \mathbb{R}^m_+$.

This is a practical formulation – the system (3) is referred to as Kuhn-Tucker (Lagrange) conditions. Practically, one can solve it, find all $\boldsymbol{x}$ that satisfy it – and these will be suitable candidates for local minima, provided that CQ is satisfied.

Note that the first equation in (3) is, in fact, n equations, and is equivalent to obtaining critical points with respect to $\boldsymbol{x}$ of the Lagrangian

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \boldsymbol{\lambda} \cdot \boldsymbol{g}(\boldsymbol{x}),$$

with $\boldsymbol{\lambda} \geq 0$ and the minus sign being therefore important! Observe that for the *MAXIMUM* problem, all one needs to do is to change the *minus* sign in the Lagrangian to *plus*, because finding a maximum for f is the same as finding a minimum for $-f$.

`Proof of KT theorem:` Follows immediately from the Farkas alternative. Given $\boldsymbol{x}$, let A be a matrix, whose columns are the vectors $\nabla g_i(\boldsymbol{x})$ for the constraints that matter at $\boldsymbol{x}$. Let $\boldsymbol{b} = \nabla f(\boldsymbol{x})$. By the Farkas alternative, one of the two occurs: either $A\boldsymbol{\lambda} = \boldsymbol{b}$ for some $\boldsymbol{\lambda} \geq 0$, or there exists some $\boldsymbol{v}$, such that $\boldsymbol{v}^T A \geq 0$ and $\boldsymbol{v} \cdot \boldsymbol{b} < 0$. I.e., there exists a feasible direction $\boldsymbol{v}$, such that the directional derivative of f in the direction $\boldsymbol{v}$ is negative. Under the non-degeneracy assumption, $\boldsymbol{v}$ is a *true* feasible direction. So, if $\boldsymbol{x}$ is a local minimum, the latter side of the Farkas alternative cannot occur. Then the former must occur. But the former side of Farkas *is* (3).□

This is really it. Let us make some final remarks addressing the longer handout.

1. Often it happens that among the constraints one has $x_1, \ldots, x_n \geq 0$. These have a particularly simple form, because their gradients are just the coordinate unit vectors $(1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$ respectively, often denoted as $\boldsymbol{e}^j$. These constraints can be singled out from the rest, $\boldsymbol{g}(\boldsymbol{x}) = 0$ then describing the rest of "more difficult" constraints. In literature the Lagrange multipliers, corresponding to the "easy" constraints $\boldsymbol{x} \geq 0$ are often denoted as $\boldsymbol{\mu}$, while $\boldsymbol{\lambda}$ stand for the Lagrange multipliers corresponding to the rest of the constraints. The Lagrangian is then $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, with $\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0$, the extra term $-\boldsymbol{\mu} \cdot \boldsymbol{x}$, and the second relation in (3) then adds to itself $\boldsymbol{x} \cdot \boldsymbol{\mu} = 0$. (Besides, many books use the letter Ψ for the Lagrangian, rather than L.)

2. To list the constraints tight at $\boldsymbol{x}$, one can run in literature into the notation $J(\boldsymbol{x})$ for the list of all the components of $\boldsymbol{x}$ which are zero (provided that $\boldsymbol{x} \geq 0$ is given as a constraint) and $I(\boldsymbol{x})$ for the list of the remaining constraints that are tight. With this notation, the main formula of the KT theorem becomes

$$\nabla f(\boldsymbol{x}) = \sum_{j \in J(\boldsymbol{x})} \mu_j \boldsymbol{e}^j + \sum_{i \in I(\boldsymbol{x})} \lambda_i \nabla g_i(\boldsymbol{x}).$$

   Following Franklin, I used the above notations in the long handouts, rather than introducing the notion of *constraints that matter*. Both ways express equivalently the same concept. And one gets rid of those by rather writing the second expression in (3) and dealing with it (adding $\mu \cdot \boldsymbol{x} = 0$ if the constraints $\boldsymbol{x} \geq 0$ are given separate treatment.)

3. The second line in (3) is often referred to as complementary slackness. Indeed, if λ_i is the i the constraint's shadow price, then it can only be nonzero when the constraint is tight. In exactly the same way as with the equality constraints, the Lagrange multipliers $\boldsymbol{\lambda}$ are the constraints' shadow prices.

4. If there is an equality constraint $h(\boldsymbol{x}) = 0$ involved, by rewriting it as $h(\boldsymbol{x}) \geq 0$ and $-h(\boldsymbol{x}) \geq 0$, assigning the Lagrange multiplier λ_1 to the first one and λ_2 to the second one, one gets the term $(\lambda_1 - \lambda_2)h(\boldsymbol{x})$ in the lagrangian, and then lets $\lambda = \lambda_1 - \lambda_2$. I.e., the Largange multiplier for an equality constraint – as we know – is unsigned.

5. Finally I copy from the long handout how KT implies the duality theory for LP. Consider the manufacturing problem Max $\boldsymbol{c}^T\boldsymbol{x}$, such that $\boldsymbol{x} \geq 0$ and $A\boldsymbol{x} \leq \boldsymbol{b}$, $\boldsymbol{x} \in \mathbb{R}^n_+$, $\boldsymbol{b} \in \mathbb{R}^m$.

   Denote $f(\boldsymbol{x}) = \boldsymbol{c}^T\boldsymbol{x}$ and $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{b} - A\boldsymbol{x}$. From linearity of the constraints, CQ are always satisfied: the gradients of the constraints are the rows of A, which are linearly independent. Also, in fact, all the functions involved are both convex and concave, and so KT are necessary and sufficient, because when one has convexity, as we know, a local extremum is the global one. As LP singles out the constraints $\boldsymbol{x} \geq 0$ from the rest, let us introduce Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m_+$ for the constraints $A\boldsymbol{x} \leq \boldsymbol{b}$ and $\boldsymbol{\mu} \in \mathbb{R}^n_+$ for the constraints $\boldsymbol{x} \geq 0$.

   The Lagrangian (note: there are *plus* signs, due to *Max*) is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = (\boldsymbol{c}^T + \boldsymbol{\mu}^T)\boldsymbol{x} + \boldsymbol{\lambda}^T(\boldsymbol{b} - A\boldsymbol{x}),$$

   and by KT, $\hat{\boldsymbol{x}}$ is the maximum production strategy if and only if together with some $\hat{\boldsymbol{\lambda}} \geq 0$, it satisfies the inequalities/equations:

$$\begin{array}{ll} \boldsymbol{\lambda}^T A \geq \boldsymbol{c}^T + \boldsymbol{\mu}^T, & \boldsymbol{c}^T\boldsymbol{x} = \boldsymbol{\lambda}^T A\boldsymbol{x}, \\ A\boldsymbol{x} \leq \boldsymbol{b}, & \boldsymbol{\lambda}^T\boldsymbol{b} = \boldsymbol{\lambda}^T A\boldsymbol{x}, \end{array} \quad \boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0.$$

   In other words, $\hat{\boldsymbol{\lambda}}$ is an optimal solution for the dual problem `min` $\boldsymbol{\lambda}^T\boldsymbol{b}$ for $\boldsymbol{\lambda} \geq 0$, such that $\boldsymbol{\lambda}^T A \geq \boldsymbol{c}^T$, reached when $\hat{\boldsymbol{\lambda}}^T\boldsymbol{b} = \boldsymbol{c}^T\hat{\boldsymbol{x}}$. Recall that for a pair $(\boldsymbol{x}, \boldsymbol{\lambda})$ of feasible solutions of the primal $A\boldsymbol{x} \leq \boldsymbol{b}$ and the dual $\boldsymbol{\lambda}^T A \geq \boldsymbol{c}^T$ problems, one always has $\boldsymbol{\lambda}^T\boldsymbol{b} \geq \boldsymbol{c}^T\boldsymbol{x}$ by the so-called weak duality theorem: to get it just multiply the primal from the left by $\boldsymbol{\lambda}^T$, the dual from the right by $\boldsymbol{x}$ and compare, using that both $\boldsymbol{x}, \boldsymbol{\lambda} \geq 0$).

   Complementary slackness theorem is also there: by definition of $\hat{\boldsymbol{\lambda}}$, a component $\hat{\lambda}_i$ may be positive only if the ith constraint for the primal is satisfied as an equality. In the same fashion, the jth feasibility inequality for the dual optimal solution (shadow price) $\hat{\boldsymbol{\lambda}}$ may not be an equality only if the corresponding component of $\hat{\boldsymbol{x}}$ is zero, that is the decision variable x_j is free (the dual inequalities for the basic components of $\hat{\boldsymbol{x}}$ are satisfied as the equalities). The vector $\hat{\boldsymbol{\mu}}^T = \hat{\boldsymbol{\lambda}}^T A - \boldsymbol{c}^T$, whose components $\hat{\mu}_j$ may be strictly positive only for non-basic j, shows the amount by which the market price c_j should increase, so that j becomes basic, that is the optimal pair $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\lambda}})$ should change, as $\mu_j < 0$ is not allowed. So it gives a reduced cost of the non-basic decision variable x_j.

   Of course, the same can be done when the primal problem is not the MP, but Canonical form, which involves equalities. Then the Lagrange multipliers $\boldsymbol{\lambda}$, or the shadow prices, will be unsigned and solve the dual problem.