TMA 4180 Optimeringsteori
KARUSH-KUHN-TUCKER THEOREM

H. E. Krogstad, IMF, Spring 2012

Karush-Kuhn-Tucker (KKT) Theorem is the most central theorem in constrained optimization, and since the proof is scattered around in Chapter 12 of N&W (more in the first edition than in the second), it may be good to give a summary of what is going on. The complete proof of the theorem is difficult, and we shall actually skip the proof of a central lemma.

This note is not at all simple, and certainly the hardest so far in the course. The goal for the reader should be to get a reasonable understanding for what is going on, that is, try to understand what the KKT theorem says and how it is applied, without trying to get all details in the proofs.

The note is somewhat extended compared to N&W and discusses the KKT theorem in connection with convexity, as well as the Sensitivity Theorem for the Lagrange multipliers.

1 Preliminaries

We consider problems of the form

$$\min_{x\in\Omega} f(x), \tag{1}$$

where the *feasibility domain,* Ω, is defined in terms of a set of *constraints*. In general, f itself is defined on a larger set than Ω. The constraints are *equality constraints*, which we write in the short form as

$$c_i(x) = 0,\ i \in \mathcal{E}, \tag{2}$$

and *inequality constraints*,

$$c_i(x) \geq 0,\ i \in \mathcal{I}. \tag{3}$$

In order to keep the exposition simple, we shall always assume that f and $\{c_i\}_{i\in\mathcal{I}\cup\mathcal{E}}$ are sufficiently smooth functions.

We could, in principle, only deal with inequality constraints, since

$$\{c_i(x) = 0\} \Longleftrightarrow \{c_i(x) \geq 0\ \wedge -c_i(x) \geq 0\}. \tag{4}$$

However, apart from in some theoretical arguments, it turns out to be convenient to keep the distinction and we therefore define

$$\Omega = \{x\ ;\ c_i(x) = 0,\ i \in \mathcal{E}\ ,\ c_i(x) \geq 0,\ i \in \mathcal{I}\}. \tag{5}$$

For a given $x_0 \in \Omega$, inequality constraints may be *active*, $c_i(x_0) = 0$, or *inactive*, $c_i(x_0) > 0$, as illustrated in Fig. 1 (there is also a concept *weakly active*, which is not needed right now). Thus, equality constraints are always active, and it is convenient to write

$$c_i(x) = 0,\ i \in \mathcal{A}, \tag{6}$$

in order to specify the active constraints. To be even more specific, $\mathcal{A} = \mathcal{E} \cup (\mathcal{I} \cap \mathcal{A})$.

We have previously defined a *feasible direction* as a non-zero vector d from a feasible point $x_0 \in \Omega$ in terms of a continuous *curve* in Ω, $x(t)$, $t \geq 0$, $x(t) \underset{t\to 0}{\longrightarrow} x_0$, such that

$$\frac{x(t) - x_0}{\|x(t) - x_0\|} \underset{t\to 0}{\longrightarrow} \frac{d}{\|d\|}. \tag{7}$$

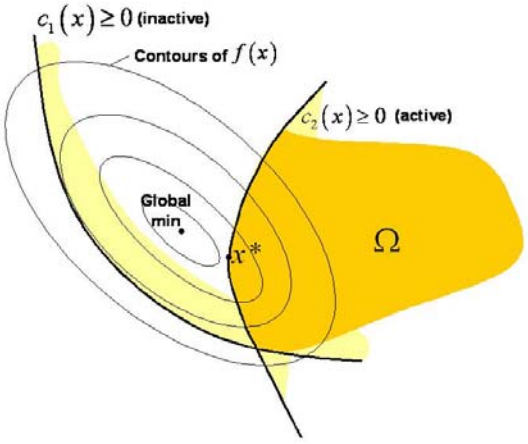

Figure 1: Active and inactive constraints: At the solution x^*, $c_2(x^*) = 0$, and this constraint is active (also called *binding*). The constraint $c_1(x) \geq 0$ is not active since $c_1(x^*) > 0$.

An equality constraint $c_i(x) = 0,\ i \in \mathcal{E}$, defines a *contour* for the function $z = c_i(x)$. Since moving around in the feasible domain means to move on this contour, i.e. "horizontally", it is obvious and easy to prove that all feasible directions from a point $x \in \Omega$ have to fulfil

$$\nabla c_i(x)\, d = 0,\ i \in \mathcal{E}. \tag{8}$$

(Otherwise, the value of $c_i(x)$ along the d-direction will immediately start to change). Similarly, check that for *active* inequality constraints, we should have

$$\nabla c_i(x)\, d \geq 0,\ i \in \mathcal{I} \cap \mathcal{A}. \tag{9}$$

When an inequality constraint is *inactive*, it puts no immediate restriction on d.

For a point $x \in \Omega$, we denote all feasible directions by $\mathcal{T}(x)$,

$$\mathcal{T}(x) = \{d;\ d \text{ feasible direction out from } x\}. \tag{10}$$

This is also called the *tangent cone* of Ω at x, since the d-s are tangents to the curves in the definition above (Note that N&W 1st Ed. used a different notation). The set is called a *cone* in $\mathbb{R}^n$, since $d \in \mathcal{T}(x) \Longrightarrow \alpha d \in \mathcal{T}(x)$ for $\alpha > 0$ (This is just the mathematical definition of a *cone*).

Exactly as before (Thm. 12.3 in N&W 2nd ed.), it is easy to prove that *if x^* is a local minimum, then*

$$\nabla f(x^*)\, d \geq 0, \text{ for all } d \in \mathcal{T}(x^*). \tag{11}$$

On the contrary, if (11) holds at a point x^*, the point will be a *candidate* for a local minimum. The KKT theorem is based on condition 11 and some rather deep additional results.

Following the notation in N&W 2nd ed., we define the following set (which is also a cone):

$$\mathcal{F}(x) = \{d;\ \nabla c_i(x)\, d = 0,\ i \in \mathcal{E},\ \nabla c_i(x)\, d \geq 0,\ i \in \mathcal{I} \cap \mathcal{A}\ \}. \tag{12}$$

The set $\mathcal{F}(x)$ is called the set of *linearized feasible directions* in N&W. Since the feasible directions in x already satisfy Eq. 8 and 9, we must have

$$\mathcal{T}(x) \subset \mathcal{F}(x), \tag{13}$$

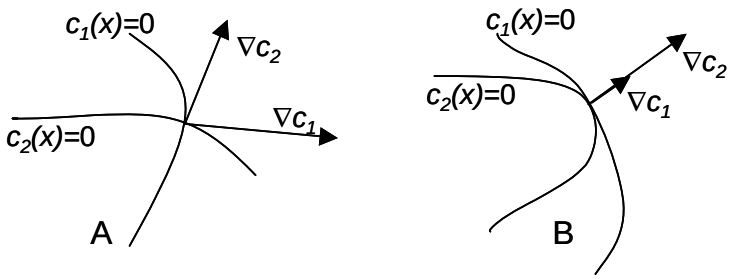

Figure 2: The LICQ holds in case A, but fails in case B, since the gradients in the point are parallel.

but it *not* obvious, and not true in general, that all directions in $\mathcal{F}(x)$ are feasible (Counter-example: N&W Example 12.4, p. 317 – 18).

This is all basic material that we need in order to be able to state the Karush-Kuhn-Tucker Theorem, but there is one extra technical condition that is used in the proof, and which may be a bit difficult to grasp. A point $x_0 \in \Omega$ is *regular* if the gradients of the active constraints at x_0,

$$\{\nabla c_i(x_0)\}, \ i \in \mathcal{A}, \tag{14}$$

are linearly independent. N&W call this the *Linearly Independence Constraint Qualification* (LICQ), see Fig. 2. If we have m active constraints, and the LICQ holds, then m must be less or equal to n, and the matrix

$$A(x_0) = \begin{bmatrix} - & \nabla c_1(x_0) & - \\ - & \nabla c_2(x_0) & - \\ & \vdots & \\ - & \nabla c_m(x_0) & - \end{bmatrix} \tag{15}$$

will have rank m. This matrix will be central in the proofs below.

2 The Main Theorem

The *Karush-Kuhn-Tucker Theorem* is Theorem 12.1 in N&W. The problem is defined in Eq. 1 and 5:

$$\min_{x \in \Omega} f(x),$$
$$\Omega = \{x \ ; \ c_i(x) = 0, \ i \in \mathcal{E} \ , \ c_i(x) \geq 0, \ i \in \mathcal{I}\}. \tag{16}$$

The formulation here is a bit more compact than the one in N&W (Thm. 12.1 p. 321).

Assume that $x^* \in \Omega$ is a local minimum and that the LICQ holds at x^*. Then it is possible to write

$$\nabla f(x^*) = \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i^* \nabla c_i(x^*), \tag{17}$$

where

$$\begin{array}{ll} (i) & \lambda_i^* \cdot c_i(x^*) = 0, \ i \in \mathcal{E} \cup \mathcal{I}, \\ (ii) & \lambda_i^* \geq 0 \text{ for } i \in \mathcal{I}. \end{array} \tag{18}$$

Note the following:

- A point x^* where Eq. 17 and 18 hold is called a *KKT-point*.
- The parameters $\{\lambda_i\}$ are called the *Lagrange multipliers*.
- It follows from 18 (i) that if an inequality constraint is inactive ($c_i(x^*) > 0$) then the corresponding λ_i^* will be equal to 0. The sum in Eq. 17 is therefore, in effect, only over $i \in \mathcal{A}$,
$$\nabla f(x^*) = \sum_{i \in \mathcal{A}} \lambda_i^* \nabla c_i(x^*). \tag{19}$$
- The Lagrange multipliers for inequality constraints must be *non-negative*, whereas equality constraints have no such restriction (since $c_i(x) = 0 \Longleftrightarrow -c_i(x) = 0$)
- Equation 17 may be written as
$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in \mathcal{A}} \lambda_i^* \nabla c_i(x^*) = 0, \tag{20}$$
where the so-called *Lagrange function*, $\mathcal{L}$, is defined as
$$\mathcal{L}(x, \lambda) = f(x) - \lambda' c(x), \tag{21}$$
and where $\lambda = (\lambda_1, \cdots, \lambda_m)'$ and $c(x) = (c_1(x), \cdots, c_m(x))'$. The notation ∇_x means the gradient taken with respect to x.
- The theorem states only *necessary* conditions. Even if Eq. 17 and 18 hold at a point $x^* \in \Omega$, it is not always this is a minimum. Unless, e.g.
  (i) if we know that a minimum exists and the problem turns out to have only one KKT-point.
  (ii) if we have a *convex problem* (see below).

In order to apply the theorem, let us first assume that we only have equality constraints. We then form the Lagrange function and simply solve (!) the $n + m$ *nonlinear equations for* (x^*, λ^*):

$$\begin{aligned} \nabla_x \mathcal{L}(x, \lambda) &= \nabla f(x) - \sum_{i \in \mathcal{A}} \lambda_i \nabla c_i(x) = 0, \\ -\nabla_\lambda \mathcal{L}(x, \lambda) &= c(x)' = (c_1(x), \cdots, c_m(x))' = 0. \end{aligned} \tag{22}$$

If we also have inequality constraints, it is in general impossible to say which constraints are active or inactive at the solution. In principle, we must then include all possible combinations of the inequality constraints as equalities and check which combination makes the smallest value of $f(x^*)$. We also have to check the sign of the Lagrange multipliers for the active constraints, and, of course, that inactive constraints are not violated. After the proof of the theorem, we shall demonstrate this for a simple example.

3 Proof of the KKT Theorem

The proof of the KKT-theorem is not simple, but can be streamlined by first establishing a couple of lemmas. Start by reading Lemma A and what is called *Farkas Lemma* below. Then read the

actual proof of the KKT theorem (follows after Farkas Lemma), and finally the proof of Lemma A, which is rather tricky.

Lemma A (part of Lemma 12.2 in N&W 2nd ed.): *If the LICQ holds at a point $x \in \mathbb{R}^n$, then $\mathcal{T}(x) = \mathcal{F}(x)$.*

Since we always have $\mathcal{T}(x) \subset \mathcal{F}(x)$, Lemma A establishes the opposite inclusion, $\mathcal{T}(x) \supset \mathcal{F}(x)$ when LICQ holds.

Proof: Form the matrix A of all gradients of the active constraints ($i \in \mathcal{A}$) at x (also stated in Eq. 15):

$$A = \begin{bmatrix} \nabla c_1(x) \\ \vdots \\ \nabla c_m(x) \end{bmatrix}. \tag{23}$$

Because LICQ holds, A has full rank, and hence, $m \leq n$. We know that $\dim \mathcal{N}(A) + \operatorname{rank}(A) = n$, so let $Z = [z_1, \cdots, z_{n-m}]$ be a basis for $\mathcal{N}(A)$. The $n \times n$ matrix

$$\begin{bmatrix} A \\ Z' \end{bmatrix} \tag{24}$$

will then be non-singular (Think about it!).

Assume that $d \in \mathcal{F}(x)$. We need to show that $d \in \mathcal{T}(x)$. The idea of the proof is to establish the existence of a curve $y(t) \in \Omega$ such that $y(t) \to x$ when $t \to 0$, and Eq. 7 holds. The construction uses the *Implicit Function Theorem* (see Basic Tools Note).

We start by forming a (smart!) non-linear system of equations for $y(t)$:

$$R(y,t) = \begin{bmatrix} c(y) - tAd \\ Z'(y - x - td) \end{bmatrix} = 0, \tag{25}$$

which obviously has the solution $y = x$ for $t = 0$. We now claim that it also has a solution $y(t) \in \Omega$ for all $t \geq 0$ in a neighborhood of $t = 0$. The existence of $y(t)$ follows from the *Implicit Function Theorem*, since

$$\left.\frac{\partial R}{\partial y}\right|_{t=0} = \left\{\frac{\partial R_i}{\partial y_j}\right\}_{i,j=1}^{n} = \begin{bmatrix} A \\ Z' \end{bmatrix} \tag{26}$$

is non-singular (fill in the details!). If we assume that all constraints are differentiable,

$$c(y) = c(x) + \nabla c(x)(y - x) + o(\|y - x\|) = A(y - x) + o(\|y - x\|). \tag{27}$$

We obtain

$$0 = R(y(t), t) = \begin{bmatrix} A \\ Z' \end{bmatrix}(y(t) - x - td) + o(\|y(t) - x\|), \tag{28}$$

which implies, again since the matrix in front is non-singular, that

$$y(t) - x = td + o(\|y(t) - x\|). \tag{29}$$

Since $y(t)$ already satisfies

$$c(y(t)) - tAd = 0, \tag{30}$$

we see that

$$c_i(y(t)) = t\nabla c_i(x)\, d = \begin{cases} 0, & i \in \mathcal{E}, \\ \geq 0, & i \in \mathcal{I} \cap \mathcal{A}. \end{cases} \tag{31}$$

In addition, since $y(t) \to x$ when $t \to 0$, and we have assumed that all elements in the c-vector are differentiable and hence continuous functions, $y(t)$ will also satisfy all *inactive* constraints for small enough t-s. All this taken together, $y(t)$ is in Ω for sufficiently small non-negative t-s. Finally, from Eq. 29,

$$\frac{y(t) - x}{\|y(t) - x\|} \xrightarrow[t \to 0]{} \frac{d}{\|d\|}, \tag{32}$$

and d is really a feasible direction, that is, $d \in \mathcal{T}(x)$.

The next lemma was not explicitly mentioned in N&W, 1st Ed., but is quite famous and was first proved in 1902.

Farkas Lemma: *Let g and $\{a_i\}_{i=1}^m$ be n-dimensional row vectors and*

$$\mathcal{S} = \{d \in \mathbb{R}^n \ ; \ gd < 0 \text{ and } a_i d \geq 0, \ i = 1, \cdots, m\}. \tag{33}$$

Then, $\mathcal{S} = \varnothing$ if and only if there is a non-negative vector $\lambda \in \mathbb{R}^m$ such that

$$g = \sum_{i=1}^{m} \lambda_i a_i. \tag{34}$$

The complete proof is surprisingly difficult, and all "simple proofs" of Farkas Lemma, or similar statements, as the one in the 1st edition of N&W, are fakes,– they use results already proved by Farkas Lemma, or other not-so-obvious results (It looks however that the proof in N&W 2nd Ed. is complete).

The general idea of the proof is simple. First of all, if Eq. 34 holds, then it is easy to see that $\mathcal{S} = \varnothing$. For the converse, let

$$\mathcal{C} = \left\{ z; \ z = \sum_{i=1}^{m} \lambda_i a, \ \lambda_i \geq 0 \right\}.$$

The set $\mathcal{C}$ is also a *cone*. It will be convex and closed (tricky to prove!). If $g \notin \mathcal{C}$, it is possible to put a plane between g and $\mathcal{C}$ (*Separating hyperplane theorem*). One of the normal vectors to this plane will be in $\mathcal{S}$, which therefore in this case is not empty.

The lemma is often formulated in terms of the matrix

$$A = \begin{Bmatrix} - & a_1 & - \\ - & a_2 & - \\ & \vdots & \\ - & a_m & - \end{Bmatrix}. \tag{35}$$

Given g and A.

Problem $\mathcal{P}$: *Find solutions d to the inequalities*

$$\begin{aligned} gd &< 0, \\ Ad &\geq 0. \end{aligned} \tag{36}$$

Problem $\mathcal{D}$: *Find a $\lambda \geq 0$ such that*

$$g = \lambda' A. \tag{37}$$

The Alternative Farkas Lemma: *Either $\mathcal{P}$ or $\mathcal{D}$ have solutions, but never both simultaneously.*

Prove the "*never both simultaneously*" yourself! (No, this does *not* prove Farkas Lemma. There could be cases where neither $\mathcal{P}$ nor $\mathcal{D}$ hold!).

Equipped with Lemma A and Farkas Lemma, the proof of the KKT-theorem is a *piece of cake*:

We know by assumption that x^* is a local minimum. Therefore,

$$\nabla f(x^*)\, d \geq 0 \text{ for all } d \in \mathcal{T}(x^*). \tag{38}$$

Since the LICQ holds, Lemma A tells us that $\mathcal{T}(x^*) = \mathcal{F}(x^*)$. Let $\nabla f(x^*)$ be g in Farkas Lemma, and the a_i-s equal to the gradients $\nabla c_i(x^*)$, $i \in \mathcal{A}$, where $c_i(x^*) \geq 0$ and $-c_i(x^*) \geq 0$ have been used for *all* equality constraints. Since $\mathcal{F}(x^*)$ contains exactly all feasible directions (and not more!) and 38 holds, problem $\mathcal{P}$ has no solution. Hence Problem $\mathcal{D}$ has a solution, and

$$\nabla f(x^*) = \sum \lambda_i^* \nabla c_i(x^*), \ \lambda_i^* \geq 0. \tag{39}$$

The remaining details (i.e., putting the sum back to the form in Eq. 17 or 19, and the conditions in 18) are left to the reader.

4 A worked example for the KKT theorem

Consider the objective function

$$f(x) = 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2, \tag{40}$$

and the constraints

$$c_1(x) = 5 - x_1^2 - x_2^2 \geq 0, \tag{41}$$
$$c_2(x) = 6 - 3x_1 - x_2 \geq 0. \tag{42}$$

Since the objective function is continuous and Ω is bounded (why?), we are sure to have minima.

As mentioned above, we first form the Lagrange function

$$\mathcal{L}(x, \lambda) = f(x) - \lambda_1 c_1(x) - \lambda_2 c_2(x). \tag{43}$$

The KKT-points (candidates for minima!) have to satisfy the following set of equations

$$\frac{\partial \mathcal{L}}{\partial x_1}(x, \lambda) = 4x_1 + 2x_2 - 10 + 2\lambda_1 x_1 + 3\lambda_2 = 0, \tag{44}$$
$$\frac{\partial \mathcal{L}}{\partial x_2}(x, \lambda) = 2x_1 + 2x_2 - 10 + 2\lambda_1 x_2 + \lambda_2 = 0, \tag{45}$$
$$\lambda_1 \left(5 - x_1^2 - x_2^2\right) = 0, \tag{46}$$
$$\lambda_2 (6 - 3x_1 - x_2) = 0, \tag{47}$$
$$\lambda_1, \lambda_2 \geq 0. \tag{48}$$

There are 4 possible combinations of active constraints at the solution:

1. No active constraints
2. c_1 active and c_2 inactive
3. c_1 inactive and c_2 active
4. Both c_1 and c_2 active

4.1 Case 1: No active constraints

SInce there are no active constraints, the theorem says that $\lambda_1 = \lambda_2 = 0$, and the minimum will occur for a point where

$$\nabla \mathcal{L}(x, 0) = \nabla f(x) = 0. \tag{49}$$

This leads to

$$4x_1 + 2x_2 - 10 = 0, \tag{50}$$
$$2x_1 + 2x_2 - 10 = 0, \tag{51}$$

with the solution

$$x_1^* = 0, \tag{52}$$
$$x_2^* = 5. \tag{53}$$

However, x^* needs to be in Ω, so we must check the constraints:

$$c_1(x^*) = 5 - 0 - 5^2 = -20 \text{ (Violation!)} \tag{54}$$
$$c_2(x^*) = 6 - 0 - 5 = 1 \text{ (OK!)} \tag{55}$$

This eliminates Case 1.

4.2 Case 4: Both constraints active

Now we have

$$c_1(x) = 5 - x_1^2 - x_2^2 = 0, \tag{56}$$
$$c_2(x) = 6 - 3x_1 - x_2 = 0, \tag{57}$$

which leads (by squaring the second condition) to a quadratic equation for x_1,

$$10x_1^2 - 36x_1 + 31 = 0. \tag{58}$$

There are two solutions and two possible points:

$$x_a = (2.17... \ , \ -0.52...), \tag{59}$$
$$x_b = (1.43... \ , \ 1.72...). \tag{60}$$

We need to check the Lagrange multipliers ($\nabla_x \mathcal{L} = 0$):

$$4x_1 + 2x_2 - 10 + 2\lambda_1 x_1 + 3\lambda_2 = 0, \tag{61}$$
$$2x_1 + 2x_2 - 10 + 2\lambda_1 x_2 + \lambda_2 = 0. \tag{62}$$

Hence,

$$\lambda_1 = \frac{10 - 2x_2 - x_1}{3x_2 - x_1}, \lambda_2 = -\left(2x_1 + 2x_2 - 10 + 2\lambda_1 x_2\right). \tag{63}$$

The point x_a gives

$$\lambda_1 = -2.37..., \ \lambda_2 = 4.22... \tag{64}$$

Since λ_1 and λ_2 should be positive, *x_a is unacceptable.*

Similarly, the point x_b gives

$$\lambda_1 = 1.37..., \ \lambda_2 = -1,02... \tag{65}$$

Also x_b is unacceptable.

4.3 Case 3: c_1 inactive, c_2 active

Since c_2 is active,

$$6 - 3x_1 - x_2 = 0.$$

Thus,

$$x_2 = 6 - 3x_1, \tag{66}$$

and

$$h\left(x_1\right) = f\left(x_1, 6 - 3x_1\right) = 5x_1^2 - 4x_1 - 24. \tag{67}$$

The (global) minimum occurs for $dh/dx = 0$, or

$$x_1 = \frac{2}{5}, \ x_2 = \frac{24}{5}. \tag{68}$$

However,

$$c_1\left(x_1, x_2\right) = 5 - \left(\frac{2}{5}\right)^2 - \left(\frac{24}{5}\right)^2 = -\frac{91}{5} < 0! \tag{69}$$

We assumed that c_1 was inactive, but this is not a guarantee for not violating it!

Only one case it left, and so far we have not found a single KKT-point.

4.4 Case 2: Only c_1 is active

Now $\lambda_2 = 0$,

$$\left(\frac{\partial \mathcal{L}}{\partial x_1} =\right) \ 4x_1 + 2x_2 - 10 + 2\lambda_1 x_1 = 0, \tag{70}$$

$$\left(\frac{\partial \mathcal{L}}{\partial x_2} =\right) \ 2x_1 + 2x_2 - 10 + 2\lambda_1 x_2 = 0, \tag{71}$$

$$x_1^2 + x_2^2 = 5. \tag{72}$$

One solution of these 3 equations is easily seen to be

$$\begin{aligned} x_1^* &= 1, \\ x_2^* &= 2, \\ \lambda_1^* &= 1. \end{aligned} \tag{73}$$

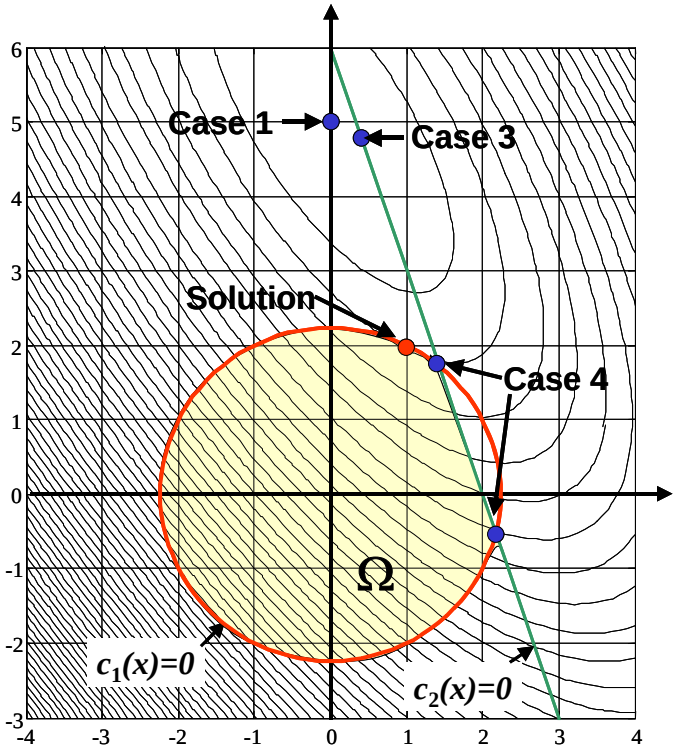

Figure 3: Contours of the objective function and the various constraints. Can you imagine where the other stationary point for Case 4 is?

This looks promising, but we must also check c_2:

$$c_2\,(1,2) = 6 - 3 - 2 = 5 > 0 \quad \text{(OK!)}$$

Finally we found a KKT-point.

There *is* another solution of Eq. 70 – 72, which is *not* a KKT-point (Left for you to find numerically!).

A summary of the problem is given in Fig. 3.

5 Convexity and the KKT-conditions

We have several times pointed out the superiority of convex optimization problems, and these occur when a convex objective function f is minimized over a convex feasibility domain Ω. Theorem 12.7 in N&W (p. 350–351) gives us simple and usable sufficient conditions for Ω to be convex. We repeat the result here for completeness:

Lemma: *Let (as above)*

$$\Omega = \{x;\ c_i(x) = 0,\ i \in \mathcal{E},\ c_i(x) \geq 0,\ i \in \mathcal{I}\}, \tag{74}$$

and assume that $c_i(x)$ for $i \in \mathcal{E}$ are linear functions, whereas $c_i(x)$ for $i \in \mathcal{I}$ are concave (that is, $-c_i$ is convex). Then Ω is convex.

Proof: Look back into the Basic Tools Note about convex functions (Proposition 2) and try to prove this lemma yourself before you look in N&W (*Hint:* Write all equality constraints in inequality form by requiring that $c_i(x) \geq 0$ and $-c_i(x) \geq 0$. Note that when c_i is linear, both c_i and $-c_i$ are linear functions and hence convex!).

An optimization problem where both Ω and f are convex is called a *Convex Problem* or a *Convex Programme*, and the theory is called *Convex Programming*.

Let us consider the convex problem

$$\min f(x) \tag{75}$$

$$c_i(x) = 0,\ i \in \mathcal{E},\ c_i(x) \geq 0,\ i \in \mathcal{I}, \tag{76}$$

where the constraints fulfill the conditions in the lemma.

We recall from the unconstrained theory that for a differentiable convex function, the condition

$$\nabla f\left(x^*\right) = 0 \tag{77}$$

is both necessary and sufficient for $x^* \in \Omega$ to be a global minimum. It turns out that we have a similar situation for constrained problems. However, it does not seem that the following theorem is stated explicitly in N&W:

The Convex KKT Theorem: *Consider a convex problem where Ω fulfills the conditions in the lemma above, and where (for the argument in the proof) all linear equality constraints have been expressed as inequality constraints. Assume that x^* is a KKT-point, that is,*

$$\begin{array}{ll} \text{(i)} & \nabla f\left(x^*\right) = \sum_{i \in \mathcal{I}} \lambda_i^* \nabla c_i\left(x^*\right), \\ \text{(ii)} & \lambda_i^* c_i\left(x^*\right) = 0, \\ \text{(iii)} & c_i\left(x^*\right) \geq 0, \\ \text{(iv)} & \lambda_i^* \geq 0. \end{array} \tag{78}$$

Then x^* is a global minimum.

Proof: The feasible set Ω is convex, and the Lagrangian evaluated for $x \in \Omega$ and $\lambda = \lambda^*$,

$$\mathcal{L}\left(x, \lambda^*\right) = f\left(x\right) - \sum_{i \in \mathcal{I}} \lambda_i^* c_i\left(x\right) = f\left(x\right) + \sum_{i \in \mathcal{I}} \lambda_i^* \left(-c_i\left(x\right)\right), \tag{79}$$

is also convex in x, because $\lambda_i^* \geq 0$, and f, as well as $-c_i$, are convex.

A differentiable convex function g lies above its tangent planes, that is,

$$g(x) \geq g(x_0) + \nabla g(x_0)'(x - x_0) \tag{80}$$

(The proof is in the Basic Tools Note). By first observing that

$$-\sum_{i \in \mathcal{I}} \lambda_i^* c_i\left(x^*\right) \leq 0, \tag{81}$$

we have

$$\begin{aligned} f\left(x\right) &\geq \mathcal{L}\left(x, \lambda^*\right) \\ &\geq \mathcal{L}\left(x^*, \lambda^*\right) + \nabla_x \mathcal{L}\left(x^*, \lambda^*\right)\left(x - x^*\right) \\ &= f\left(x^*\right) - 0 + 0 \times \left(x - x^*\right) \\ &= f\left(x^*\right). \end{aligned} \tag{82}$$

Note that in this theorem, the KKT-point is *assumed* to exist, and no LICQ-condition is necessary for that.

However, if we have a convex problem and a global minimum x^* (since all minima are global), and the LICQ condition holds in x^*, *then* the KKT-conditions (78) also hold because of the KKT Theorem.

For convex problems, the KKT-conditions are therefore *sufficient* for having a global minimum!

Since the solution we found for Case 2 in the example above was an isolated KKT-point, checking the other solution was not really necessary.

6 Second Order Conditions

Similar to the unconstrained case and the general (non-convex) situation, the first order conditions in the KKT theorem can not ensure that the KKT point you find is a minimum. From the non-constrained case, we recall that if $\nabla f(x^*) = 0$, it was possible to go on and look at the second derivative, the Hessian, $\nabla^2 f(x^*)$. If x^* was a local minimum, then $\nabla^2 f(x^*)$ was necessarily positive semi-definite, and if $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$, then we could conclude that x^* *was* a local minimum.

The constrained case is more complicated for two reasons:

1. The feasible directions from a point $x \in \Omega$ may be just be a small subset of all directions from x.
2. The gradient of f at x^* is in general not 0.

We know that at a local minimum, $\nabla f(x^*)\, d \geq 0$ for all feasible directions. If $\nabla f(x^*)\, d > 0$ it is clear that $f(x)$ increases (locally) along curves out from x^* with limiting direction d, but this is impossible to say if $\nabla f(x^*)\, d = 0$.

Let us now consider a point x^* where the KKT-conditions and the LICQ hold. Since LICQ holds, Lemma A says that the feasible directions is also equal to the set $\mathcal{F}(x^*)$, as defined in Eq. 12.

Below, we simplify the analysis somewhat compared to N&W by assuming *strict complementarity* (N&W, Definition 12.5, p. 321). This implies that the Lagrange multipliers to all active inequality constraints are strictly positive.

Following in essence N&W Section 12.5, we introduce the subset $\mathcal{C}(x^*) \subset \mathcal{F}(x^*)$ consisting of the problematic directions for which $\nabla f(x^*)\, d = 0$. Since x^* is a KKT-point, we have for $d \in \mathcal{C}(x^*)$ that

$$0 = \nabla f(x^*)\, d = \sum_{i \in \mathcal{E}} \lambda_i^* \nabla c_i(x^*)\, d + \sum_{i \in \mathcal{I} \cap \mathcal{A}} \lambda_i^* \nabla c_i(x^*)\, d. \tag{83}$$

The first sum is always 0, and in the second we know that $\nabla c_i(x^*)\, d \geq 0$. However, we also know that $\lambda_i^* \geq 0$ according to the KKT theorem, and in fact strictly positive, according to strict complementarity. Hence

$$\nabla c_i(x^*)\, d = 0,\ i \in \mathcal{I} \cap \mathcal{A}, \tag{84}$$

and we may characterize $\mathcal{C}(x^*)$ simply as

$$\mathcal{C}(x^*) = \{d;\ \nabla c_i(x^*)\, d = 0,\ i \in \mathcal{A}\}. \tag{85}$$

Theorem N&W 12.5 and 12.6 *Let x^* be a KKT-point where LICQ and strict complementarity apply.*

(a) If x^* is a local minimum, then

$$d'\nabla^2\mathcal{L}(x^*,\lambda^*)\,d \geq 0 \text{ for all } d \in \mathcal{C}(x^*). \tag{86}$$

(b) If

$$d'\nabla^2\mathcal{L}(x^*,\lambda^*)\,d > 0 \text{ for all } d \in \mathcal{C}(x^*),\ d \neq 0, \tag{87}$$

then x^* is a strict local minimum.

(The theorem is also valid without strict complementarity, see N&W, Section 12.5)

Proof: Recall the proof of Lemma A and the solution $y(t) \to x^*$ when $t \to 0$. Since

$$c(y(t)) - tAd = 0, \tag{88}$$

we have for all $i \in \mathcal{A}$,

$$c_i(y(t)) = t\nabla c_i(x^*)\,d = 0 \text{ for all } d \in \mathcal{C}(x^*). \tag{89}$$

Thus,

$$\mathcal{L}(y(t),\lambda^*) = f(y(t)) - \lambda^{*\prime}c(y(t)) = f(y(t)). \tag{90}$$

Along $y(t)$ we may therefore check the Lagrangian instead of f.

$$\begin{aligned} f(y(t)) &= \mathcal{L}(y(t),\lambda^*) \\ &= \mathcal{L}(x^*,\lambda^*) + \nabla_x\mathcal{L}(x^*,\lambda^*)(y(t)-x^*) + \qquad (91)\\ &+ \frac{1}{2}(y(t)-x^*)'\nabla_x^2\mathcal{L}(x^*,\lambda^*)(y(t)-x^*) + o\left(\|y(t)-x^*\|^2\right) \qquad (92)\\ &= f(x^*) + \frac{1}{2}(y(t)-x^*)'\nabla_x^2\mathcal{L}(x^*,\lambda^*)(y(t)-x^*) + o\left(\|y(t)-x^*\|^2\right). \end{aligned}$$

Note that it follows from the KKT theorem that $\mathcal{L}(x^*,\lambda^*) = f(x^*)$, and $\nabla_x\mathcal{L}(x^*,\lambda^*) = 0$.

As in the proof of Lemma A, we can for any $d \in \mathcal{C}(x^*)$ find an $y(t)$ such that

$$y(t) - x^* = td + o(\|y(t)-x^*\|), \tag{93}$$

and the rest of the proof of **(a)** is left to the reader.

For Part **(b)**, the proof is by contradiction: Assume that Eq. 87 holds, but x^* is not a strict minimum. From the definition of a strict minimum we then have that there is a sequence $\{x_i\} \subset \Omega$, converging to x^*, and such that $f(x_i) \leq f(x^*)$. By a compactness argument, there also is even a subsequence $\{x_{i_n}\}$ such that

$$\frac{x_{i_n}-x^*}{\|x_{i_n}-x^*\|} \underset{n\to\infty}{\longrightarrow} d. \tag{94}$$

(**Digression:** *The compactness argument.* The left hand sides of Eq. 94 will for all i's be vectors of length 1. Their end-points are lying on the sphere $\|x\| = 1$ in $\mathbb{R}^n$. This sphere is a closed and bounded set in $\mathbb{R}^n$ and therefore compact. *Then* we apply the definition of compactness).

Thus, d is a feasible direction in x^*(N&W Def. 12.2). If d is not in $\mathcal{C}(x^*)$, we have $\nabla c_i(x^*)\,d > 0$ for at least one of the active constraints, and then (since in that case $\lambda_i^* > 0$ by strict complementarity), also $\nabla f(x^*)\,d > 0$. But that is simply impossible if $f(x_{i_n}) \leq f(x^*)$ (Remember Taylor's

formula, $f(x_{i_n}) = f(x^*) + \nabla f(x_\theta)(x_{i_n} - x^*)$). The only remaining case is that $d \in \mathcal{C}(x^*)$. From Eq. 90 we have in general that

$$f(x_{i_n}) \geq \mathcal{L}(x_{i_n}, \lambda^*) = f(x^*) + \frac{1}{2}(x_{i_n} - x^*)' \nabla_x^2 \mathcal{L}(x^*, \lambda^*)(x_{i_n} - x^*) + o\left(\|y(t) - x^*\|^2\right). \quad (95)$$

This leads, by a similar argument to the proof of **(a)**, to

$$0 \geq d' \nabla^2 \mathcal{L}(x^*, \lambda^*) d > 0, \quad (96)$$

which is impossible.

The above result may look a little obscure, but it has a simple geometric content. Assume for simplicity that we only have one (differentiable) equality constraint. $c_0(x^*) = 0$. The constraint limits our motion around x^*, and if we magnify the neighborhood, it looks for us that the constraint forces us to move in the *tangent (hyper) plane* through x^*, which for $\nabla c_0(x^*) \neq 0$, is defined as

$$\{x \ ; \ \nabla c_0(x^*)(x - x^*) = 0\}. \quad (97)$$

Adding another equality constraint limits our motion to the *intersection* of the two planes. E.g., in $\mathbb{R}^3$, the intersection of two different planes will be a *line*. If the LICQ holds, the intersection of all tangent planes defines what is called the *tangent space* in x^*. It is clear from Eq. 85 that $\mathcal{C}(x^*)$ *is* the tangent space in x^* if we only have equality constraints (and will continue to be so if the LICQ holds and we have strict complementarity). If we introduce, as in the proof of Lemma A, the matrix of all gradients of the active constraints,

$$A = \begin{bmatrix} \nabla c_1(x^*) \\ \vdots \\ \nabla c_m(x^*) \end{bmatrix}, \quad (98)$$

then

$$\mathcal{C}(x^*) = \mathcal{N}(A). \quad (99)$$

This is illustrated in Fig. 4. If $F = [z_1 \ z_2 \cdots z_{n-m}]$ is a basis for $\mathcal{N}(A)$, we can write all vectors $d \in \mathcal{C}(x^*)$ as

$$d = Zu, \ u \in \mathbb{R}^{n-m}. \quad (100)$$

The expression in Eq. 86 then reads

$$d' \nabla^2 \mathcal{L}(x^*, \lambda^*) d = u' \left(Z' \nabla^2 \mathcal{L}(x^*, \lambda^*) Z\right) u. \quad (101)$$

It is common to call $Z' \nabla^2 L(x^*, \lambda^*) Z$ the *projected Lagrangian*, where the *projection* is onto the linear operators on $\mathcal{N}(A)$. The conditions in the theorem may therefore be tested on the projected Lagrangian on the $n - m$ dimensional space $\mathcal{N}(A)$, where *the tests are identical to the non-constrained case.*

6.1 Example

We illustrate the above theory by the following very simple problem (are you able to see the practical origin of this problem?):

$$\min f(x) = \min \{-(x_1 x_2 + x_2 x_3 + x_1 x_3)\}$$

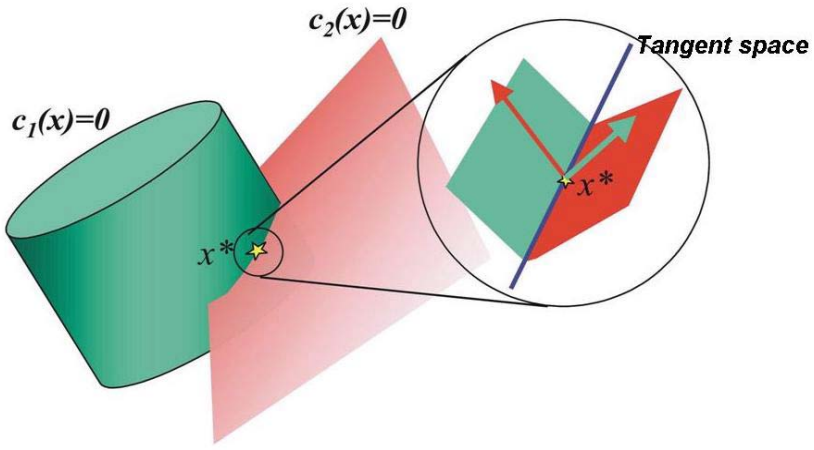

Figure 4: At a *regular point* (i.e., where the LICQ holds), the tangent space is defined in terms of the gradients of the active (equality) constraints, and is equal to $\mathcal{N}(A)$.

when

$$c(x) = x_1 + x_2 + x_3 - 3 = 0.$$

The Lagrangian is

$$\mathcal{L}(x, \lambda) = -(x_1x_2 + x_2x_3 + x_1x_3) - \lambda(x_1 + x_2 + x_3 - 3),$$

leading to the equations

$$\begin{aligned} -x_2 - x_3 - \lambda &= 0, \\ -x_1 - x_3 - \lambda &= 0 \\ -x_2 - x_1 - \lambda &= 0 \\ x_1 + x_2 + x_3 &= 3 \end{aligned}$$

The KKT-point is $x_1^* = x_2^* = x_3^* = 1$ and $\lambda^* = -2$ (which is acceptable for an *equality* constraint!). Is this point a minimum? Let us compute $\nabla^2\mathcal{L}$:

$$\nabla^2\mathcal{L} = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}$$

The matrix has eigenvalues -2 and 1, and is therefore by itself *not* positive semidefinite. However, close to the solution, we only move (approximately) around in a set which is $\mathcal{N}(A)$ (with origin shifted to x^*, and where $A = [1\ 1\ 1]$. The columns of

$$Z = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

span $\mathcal{N}(A)$, and we check the projected Lagrangian,

$$Z'\nabla^2\mathcal{L}Z = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} > 0. \qquad (102)$$

Thus, the KKT point is a strict, and in fact, global minimum (Which, of course, is seen easier by inserting the constraint directly into the objective and thus eliminating x_3, for example).

7 Sensitivity and the Meaning of the Lagrange Multipliers

The size of the Lagrange multipliers can tell us something about the importance of the various constraints at the solution, but the treatment in N&W (p. 341–343) is rather sketchy.

The discussion below is adapted from D. G. Luenberger: *Linear and Nonlinear Programming*, 2nd Ed., Addison Westley, 1984, p. 313 – 318, and states the so-called *Sensitivity Theorem*.

Consider a KKT point x^* where the LICQ holds and where also

$$d' \nabla^2 \mathcal{L}(x^*, \lambda^*) d > 0$$

for all $d \in \mathcal{C}(x^*)$, such that x^* is a strict local minimum. Let us again introduce the matrix A as in Eq. 98 such that the KKT equations may be expressed as

$$\nabla f(x) - \lambda' A(x) = 0, \tag{103}$$

$$c(x) = 0. \tag{104}$$

The vector c consists of the constraints that are active at x^*.

Let us change the active constraints a tiny amount to

$$c(x) = \delta, \tag{105}$$

where $\delta \in \mathbb{R}^m$, but such that none of the inactive constraints become active. The KKT equations then change to

$$\nabla f(x) - \lambda' A(x) = 0, \tag{106}$$

$$c(x) = \delta. \tag{107}$$

We now claim that these equations have, for sufficiently small δ, a unique continuous solution $x^*(\delta)$ (and $\lambda^*(\delta)$) such that

$$x^*(\delta) \underset{\|\delta\| \to 0}{\longrightarrow} x^*. \tag{108}$$

This follows again from the *Implicit Function Theorem*: The equations 106 and 107 have the solutions x^* and λ^* for $\delta = 0$. Furthermore, the Jacobian of the left hand side at x^* is (derivation left to the reader!)

$$\begin{bmatrix} \nabla^2 \mathcal{L}(x^*, \lambda^*) & A'(x^*) \\ A(x^*) & 0 \end{bmatrix}, \tag{109}$$

and *this matrix is non-singular*:

Lemma: *Let $L \in \mathbb{R}^{n \times n}$ and $A \in \mathbb{R}^{m \times n}$, where A has full rank m, and $d'Ld > 0$ for all $d \in \mathcal{N}(A)$, $d \neq 0$. Then*

$$\begin{bmatrix} L & A' \\ A & 0 \end{bmatrix} \tag{110}$$

is non-singular.

Proof: Also left for the reader (Hint: Assume that

$$\begin{bmatrix} L & A' \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{111}$$

and show that this implies that both x and λ are equal to 0! In case of problems, look up N&W, Lemma 16.1 in Chapter 16).

After this, we can then state the

The Sensitivity Theorem:

Assume that the conditions above hold. Then

$$f\left(x^*\left(\delta\right)\right) = f\left(x^*\right) + \lambda^{*\prime}\delta + o\left(\|\delta\|\right)$$

Proof: Let us introduce the notation

$$\nabla_\delta f\left(x^*\left(\delta\right)\right)|_{\delta=0} = \left\{ \frac{\partial}{\partial \delta_i} f\left(x^*\left(\delta\right)\right)\Big|_{\delta=0} \right\} \tag{112}$$

The result will follow if we are able to to prove that

$$\frac{\partial}{\partial \delta_i} f\left(x^*\left(\delta\right)\right)\Big|_{\delta=0} = \lambda_i^* \tag{113}$$

and this follows by applying the *Chain Rule* and Eq. 107: First of all,

$$\nabla_\delta f\left(x^*\left(\delta\right)\right)|_{\delta=0} = \nabla f\left(x^*\right) J, \tag{114}$$

where

$$J = \left\{ \frac{\partial x_i^*}{\partial \delta_j}\left(0\right) \right\}. \tag{115}$$

Moreover,

$$\nabla_\delta c\left(x^*\left(\delta\right)\right)|_{\delta=0} = AJ. \tag{116}$$

But also, since $c\left(x^*\left(\delta\right)\right) = \delta$,

$$\nabla_\delta c\left(x^*\left(\delta\right)\right)|_{\delta=0} = I_{m\times m}. \tag{117}$$

Hence, by using the first KKT-equation,

$$\nabla_\delta f\left(x^*\left(\delta\right)\right)|_{\delta=0} = \nabla f\left(x^*\right) J = \left(\lambda^{*\prime} A\right) J = \lambda^{*\prime}. \tag{118}$$

Note that the sign of δ here is the opposite of what is used in N&W.

A change in the active constraint $c_i\left(x\right) = 0$ to $c_i\left(x\right) = \delta_i$ thus leads to a first order change in the optimal value of the objective $f\left(x^*\right)$ by $\lambda_i^*\delta_i$. Note the definition *strongly active,* or *binding*, for an active inequality constraint where $\lambda_i^* > 0$, and *weakly active* if $\lambda_i^* = 0$ (N&W, Definition 12.3). This is illustrated in Fig. 5. For an inactive constraint, a small change does not influence optimal solution at all ($\lambda_i^* = 0$), whereas for a weakly active constraint, the value of the objective function does not change to the first order if the condition is perturbed.

The larger the Lagrange multipliers, the more dramatic the change in the optimal value!

The changes of the optimal values of x^* and λ^* due to changes in δ have to be found by solving the equations 106 and 107. To first order, this amounts to solve the (regular) linear system

$$\begin{bmatrix} \nabla^2\mathcal{L}\left(x^*,\lambda^*\right) & A'\left(x^*\right) \\ A\left(x^*\right) & 0 \end{bmatrix} \begin{bmatrix} x^*\left(\delta\right) - x^* \\ \lambda^*\left(\delta\right) - \lambda^* \end{bmatrix} = \begin{bmatrix} 0 \\ \delta \end{bmatrix}. \tag{119}$$

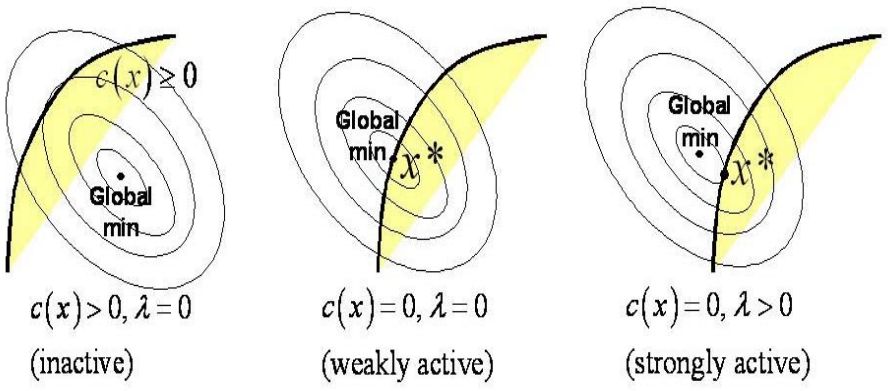

Figure 5: Constraints and Lagrange coefficients for the three different cases *inactive*, *weakly active*, and *active* constraints.