# First Order Optimality Conditions for Constrained Nonlinear Programming

Lecture 9, Continuous Optimisation

Oxford University Computing Laboratory, HT 2006
Notes by Dr Raphael Hauser (hauser@comlab.ox.ac.uk)

Additions by Dr Nick Gould (n.gould@rl.ac.uk)

**Optimality Conditions: What We Know So Far**

- Necessary optimality conditions for unconstrained optmization: $\nabla f(x) = 0$ and $D^2 f(x) \succeq 0$.
- Sufficient optimality conditions: $\nabla f(x) = 0$, $D^2 f(x) \succ 0$.
- Sufficiency occurs because $D^2 f(x) \succ 0$ guarantees that $f$ is locally strictly convex.
- Indeed, if convexity of $f$ is a given, $\nabla f(x^*) = 0$ is a necessary and sufficient condition.

- In the exercises, we used the fundamental theorem of linear inequalities to derive the LP duality theorem. This yielded the necessary and sufficient optimality conditions

$$\begin{aligned} A^\top y = c, \quad & y \geq 0 \\ Ax \leq b & \\ c^\top x - b^\top y = 0 & \end{aligned}$$

for the LP problem

$$\text{(P)} \qquad \max_{x \in \mathbb{R}^n} c^\top x \quad \text{s.t.} \quad Ax \leq b.$$

- Writing (P) in the form

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } g_i(x) \geq 0 \quad (i = 1, \ldots, m), \end{aligned}$$

the optimality conditions can be rewritten as

$$\begin{aligned} &\nabla f(x) - \sum_{i=1}^m y_i \nabla g_i(x) = 0 \\ &g_i(x) \geq 0 \quad (i = 1, \ldots, m) \\ &y^\top (Ax - b) = 0, \text{ that is,} [\, g_1(x) \; \ldots \; g_m(x) \,] y = 0. \end{aligned}$$

- We will see that the last condition could have been strengthened to $y_i g_i(x) = 0$ for all $i$.
- LP is the simplest example of a constrained convex optimisation problem: minimise a convex function over a convex domain. Again convexity implies that first order conditions are enough.

More generally, let

$$\text{(NLP)} \qquad \min_{x\in\mathbb{R}^n} f(x)$$
$$\text{s.t.} \quad g_i(x) = 0, \quad (i \in \mathcal{E}),$$
$$g_j(x) \geq 0 \quad (j \in \mathcal{I}).$$

The following will emerge under appropriate regularity assumptions:

i) Convex problems have first order necessary and sufficient optimality conditions.

ii) In general problems, second order conditions introduce local convexity.

If $\mathcal{J} \subset \mathcal{E} \cup \mathcal{I}$ is a subset of indices, we will write

- $g_{\mathcal{J}}$ for the vector-valued map that has $g_i$ $(i \in \mathcal{J})$ as components in some specific order,

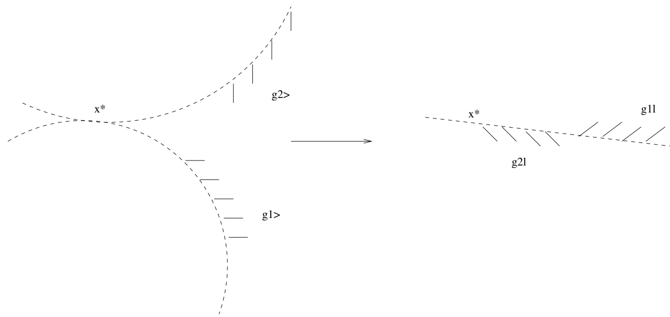
- $g$ for $g_{\mathcal{E}\cup\mathcal{I}}$.

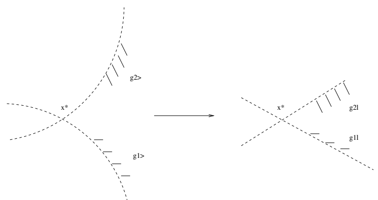
**Definition 2:** If $\{\nabla g_i : i \in \mathcal{E} \cup \mathcal{A}(x^*)\}$ is a linearly independent set of vectors, we say that the *linear independence constraint qualification* (LICQ) holds at $x^*$.

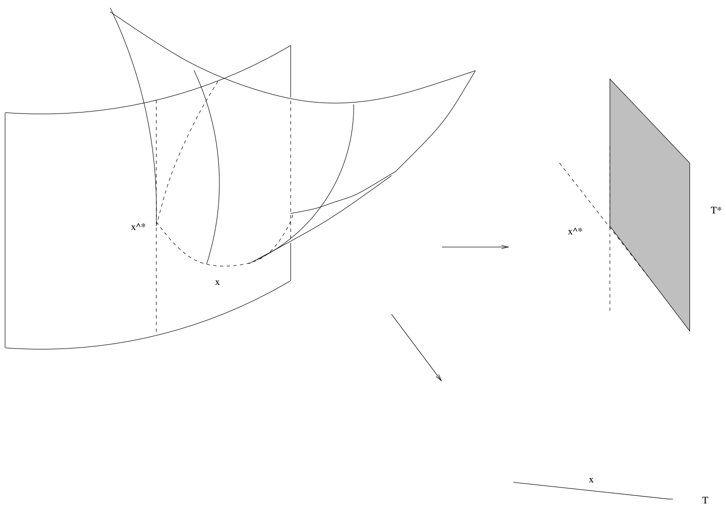
## I. First Order Necessary Optimality Conditions

**Definition 1** Let $x^* \in \mathbb{R}^n$ be feasible for the problem (NLP). We say that the inequality constraint $g_j(x) \geq 0$ is *active* at $x^*$ if $g(x^*) = 0$. We write $\mathcal{A}(x^*) := \{j \in \mathcal{I} : g_j(x^*) = 0\}$ for the set of indices corresponding to active inequality constraints.

Of course, equality constraints are always active, but we will account for their indices separately.







**Lemma 1:** Let $x^*$ be a feasible point of (NLP) where the LICQ holds and let $d \in \mathbb{R}^n$ be a vector such that

$$
\begin{aligned}
d &\neq 0, \\
d^{\mathsf{T}} \nabla g_i(x^*) &= 0, \qquad (i \in \mathcal{E}), \\
d^{\mathsf{T}} \nabla g_j(x^*) &\geq 0, \qquad (j \in \mathcal{A}(x^*)).
\end{aligned} \tag{1}
$$

Then for $\epsilon > 0$ small enough there exists a path $x \in C^k\Big((-\epsilon, +\epsilon), \mathbb{R}^n\Big)$ such that

$$
\begin{aligned}
x(0) &= x^*, \\
\frac{d}{dt}x(0) &= d, \\
g_i(x(t)) &= t d^{\mathsf{T}} \nabla g_i(x^*) \quad (i \in \mathcal{E} \cup \mathcal{A}(x^*), t \in (-\epsilon, \epsilon)),
\end{aligned} \tag{2}
$$

so that

$$
\begin{aligned}
g_i(x(t)) &= 0 \quad (i \in \mathcal{E}, t \in (-\epsilon, \epsilon)), \\
g_j(x(t)) &\geq 0 \quad (j \in \mathcal{I}, t \geq 0).
\end{aligned}
$$

*Proof:*

- Let $l = |\mathcal{A}(x^*) \cup \mathcal{E}|$. Since the LICQ holds, it is possible to choose $Z \in \mathbb{R}^{(n-l)\times n}$ such that $\left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) \\ Z \end{smallmatrix}\right]$ is a nonsingular $n \times n$ matrix.

- Let $h : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ be defined by
$$(x,t) \mapsto \left[\begin{smallmatrix} g_{\mathcal{A}(x^*)\cup\mathcal{E}}(x) - tDg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*)d \\ Z(x - x^* - td) \end{smallmatrix}\right]$$

- Then $Dh(x^*, 0) = [\, D_x h(x^*,0) \;\; D_t h(x^*,0) \,]$, where
$$
\begin{aligned}
D_x h(x^*, 0) &= \left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*) \\ Z \end{smallmatrix}\right] \qquad \text{and} \\
D_t h(x^*, 0) &= -\left[\begin{smallmatrix} Dg_{\mathcal{A}(x^*)\cup\mathcal{E}}(x^*)d \\ Zd \end{smallmatrix}\right] = -D_x h(x^*, 0)d
\end{aligned}
$$

- Since $D_x h(x^*, 0)$ is nonsingular, the Implicit Function Theorem implies that for $\tilde{\epsilon} > 0$ small enough there exists a unique $C^k$ function $x : (-\tilde{\epsilon}, \tilde{\epsilon}) \to \mathbb{R}^n$ and a neighbourhood $\mathfrak{V}(x^*)$ such that for $x \in \mathfrak{V}(x^*)$, $t \in (-\tilde{\epsilon}, \tilde{\epsilon})$,
$$h(x,t) = 0 \Leftrightarrow x = x(t).$$

- In particular, we have $x(0) = x^*$ and $g_i(x(t)) = td^{\mathsf{T}} \nabla g(x^*)$ for all $i \in \mathcal{A}(x^*) \cup \mathcal{E}$ and $t \in (-\tilde{\epsilon}, \tilde{\epsilon})$. (1) therefore implies that $g_i(x(t)) = 0$ $(i \in \mathcal{E})$ and $g_i(x(t)) \geq 0$ $(i \in \mathcal{A}(x^*), t \in [0, \tilde{\epsilon}))$.

- On the other hand, since $g_i(x^*) > 0$ ($i \notin \mathcal{A}(x^*)$), the continuity of $x(t)$ implies that there exists $\epsilon \in (0, \tilde{\epsilon})$ such that $g_j(x(t)) > 0$ ($j \in \mathcal{I} \setminus \mathcal{A}(x^*), t \in (-\epsilon, \epsilon)$).

- Finally,
$$\frac{d}{dt}x(0) = -\Big(D_x h(x^*, 0)\Big)^{-1} D_t h(x^*, 0) = d$$
follows from the second part of the Implicit Function Theorem. $\square$

- Since $d$ satisfies (1), Lemma 1 implies that there exists a path $x : (-\epsilon, \epsilon) \to \mathbb{R}^n$ that satisfies (2).

- Taylor's theorem then implies that
$$f(x(t)) = f(x^*) + td\nabla f(x^*) + O(t^2) < f(x^*)$$
for $0 < t \ll 1$.

- Since (2) shows that $x(t)$ is feasible for $t \in [0, \epsilon)$, this contradicts the assumption that $x^*$ is a local minimiser. $\square$

**Theorem 1:** If $x^*$ is a local minimiser of (NLP) where the LICQ holds then
$$\nabla f(x^*) \in \text{cone}\Big(\{\pm\nabla g_i(x^*) : i \in \mathcal{E}\} \cup \{\nabla g_j(x^*) : j \in \mathcal{A}(x^*)\}\Big).$$

*Proof:*

- Suppose our claim is wrong. Then the fundamental theorem of linear inequalities implies that there exists a vector $d \in \mathbb{R}^n$ such that
$$\begin{aligned} d^\mathsf{T}\nabla g_j(x^*) &\geq 0, \qquad (j \in \mathcal{A}(x^*)),\\ \pm d^\mathsf{T}\nabla g_i(x^*) &\geq 0, \quad (\text{i.e., } d^\mathsf{T}\nabla g_i(x^*) = 0) \qquad (i \in \mathcal{E}),\\ d^\mathsf{T}\nabla f(x^*) &< 0. \end{aligned}$$

*Comments:*

- The condition
$$\nabla f(x^*) \in \text{cone}\Big(\{\pm\nabla g_i(x^*) : i \in \mathcal{E}\} \cup \{\nabla g_j(x^*) : j \in \mathcal{A}(x^*)\}\Big)$$
is equivalent to the existence of $\lambda \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ such that
$$\nabla f(x^*) = \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla g_i(x^*), \tag{3}$$
where $\lambda_j \geq 0$ ($j \in \mathcal{A}(x^*)$) and $\lambda_j = 0$ for ($j \in \mathcal{I} \setminus \mathcal{A}(x^*)$).

- $x^*$ was assumed feasible, that is, $g_i(x^*) = 0$ for all $i \in \mathcal{E}$ and $g_j(x^*) \geq 0$ for all $j \in \mathcal{I}$.

Thus, Theorem 1 shows that when $x^*$ is a local minimiser where the LICQ holds, then the following so-called Karush-Kuhn-Tucker (KKT) conditions must hold:

**Corollary 1:** There exist *Lagrange multipliers* $\lambda \in \mathbb{R}^{|\mathcal{I}\cup\mathcal{E}|}$ such that

$$\begin{aligned} \nabla f(x) - \sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i \nabla g_i(x) &= 0 \\ g_i(x) &= 0 \qquad (i \in \mathcal{E}) \\ g_j(x) &\geq 0 \qquad (j \in \mathcal{I}) \\ \lambda_j g_j(x) &= 0 \qquad (j \in \mathcal{I}) \\ \lambda_j &\geq 0 \qquad (j \in \mathcal{I}). \end{aligned}$$

**Corollary 2: First Order Necessary Optimality Conditions.** If $x^*$ is a local minimiser of (NLP) where the LICQ holds then there exists $\lambda^* \in \mathbb{R}^m$ such that $(x^*, \lambda^*)$ solves the following system of inequalities,

$$\begin{aligned} D_x\mathcal{L}(x^*, \lambda^*) &= 0, \\ \lambda_j^* &\geq 0 \quad (j \in \mathcal{I}), \\ \lambda_i^* g_i(x^*) &= 0 \quad (i \in \mathcal{E}\cup\mathcal{I}), \\ g_j(x^*) &\geq 0 \quad (j \in \mathcal{I}), \\ g_i(x^*) &= 0 \quad (i \in \mathcal{E}). \end{aligned}$$

We can formulate this result in slightly more abstract form in terms of the Lagrangian associated with (NLP):

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m &\to \mathbb{R} \\ (x, \lambda) &\mapsto f(x) - \sum_{i=1}^m \lambda_i g_i(x). \end{aligned}$$

The balance equation

$$\nabla f(x) - \sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i \nabla g_i(x) = 0$$

says that the derivative of the Lagrangian with respect to the $x$ coordinates is zero.

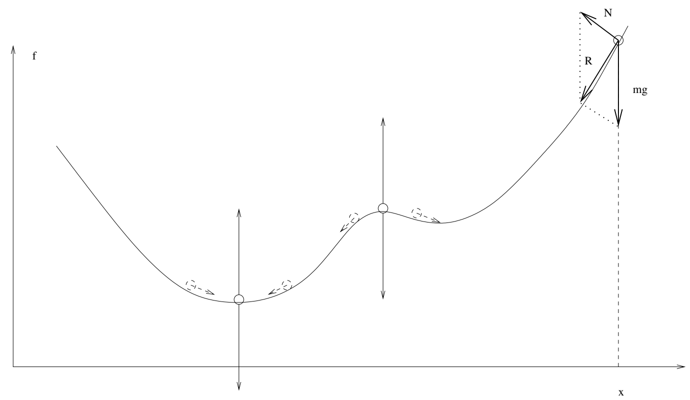
Putting all the pieces together, we obtain the following result:

**Mechanistic Motivation of KKT Conditions:**

A useful picture in unconstrained optimisation is to imagine a point mass $m$ or an infinitesimally small ball that moves on a hard surface

$$F := \left\{(x, f(x)) : x \in \mathbb{R}^n\right\}$$

without friction.




- The external forces acting on the point mass are the gravity force $m\vec{g} = \binom{0}{-mg}$ and the reaction force

$$\vec{N}_f = \frac{mg}{1 + \|\nabla f(x)\|^2}\binom{-\nabla f(x)}{1}.$$

- The total external force

$$\vec{R} = m\vec{g} + \vec{N}_f = \frac{mg}{1 + \|\nabla f(x)\|^2}\begin{bmatrix} -\nabla f(x) \\ -\|\nabla f(x)\|^2 \end{bmatrix} \perp \vec{N}_f$$

equals zero if and only if $\nabla f(x) = 0$ (i.e., a stationary point).

- When the test mass is slightly moved from a local maximiser, then the external forces will pull it further away.

- In a neighbourhood of a local minimiser they will restore the point mass to its former position.

- This is expressed by the second order optimality conditions: an equilibrium position is *stable* if $D^2 f(x) \succ 0$ and instable if $D^2 f(x) \prec 0$.

**Extension to constrained optimisation:**

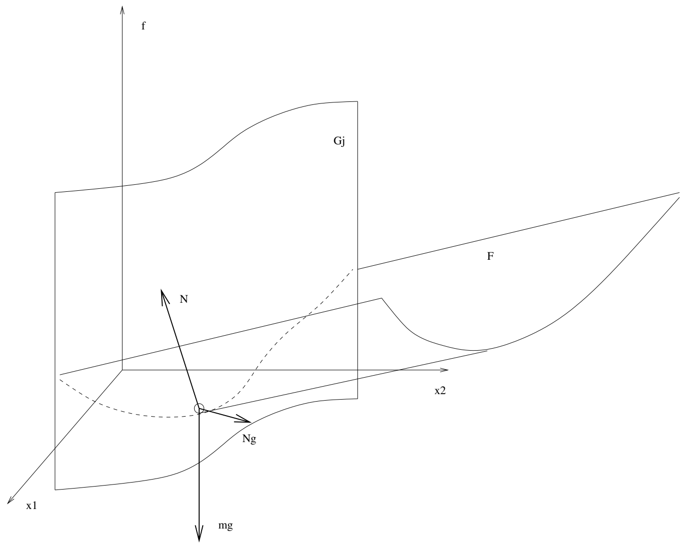
We can interpret an inequality constraint $g(x) \geq 0$ as a hard smooth surface

$$G := \left\{(x, z) \in \mathbb{R}^n \times \mathbb{R} : \, g(x) = 0\right\}$$

that is parallel to the $z$-axis everywhere and keeps the point mass from rolling into the domain where $g(x) < 0$.

Such a surface can exert only a normal force that points towards the domain $\{x : \, g_j(x) > 0\}$.

Therefore, the reaction force must be of the form $\vec{N}_g = \mu_g\binom{\nabla g(x)}{0}$, where $\mu_g \geq 0$.




- In the picture the point mass is at rest and does not roll to lower terrain if the sum of external forces is zero, that is, $\vec{N}_f + \vec{N}_g + m\vec{g} = 0$.

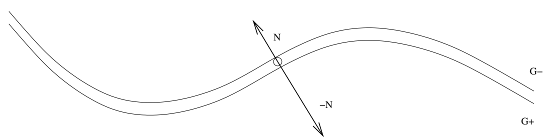
- Since $\vec{N}_f = \mu_f \binom{-\nabla f(x)}{1}$ for some $\mu_f \geq 0$, we find
$$\mu_f \begin{bmatrix} -\nabla f(x) \\ 1 \end{bmatrix} + \mu_g \begin{bmatrix} \nabla g(x) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -mg \end{bmatrix} = 0,$$
from where it follows that $\mu_f = mg$ and
$$\nabla f(x) = \lambda \nabla g(x) \tag{4}$$
with $\lambda = \mu/mg \geq 0$.

- When multiple inequality constraints are present, the the balance equation (4) must thus be replaced with
$$\nabla f(x) = \sum_{j \in \mathcal{I}} \lambda_j \nabla g_j(x)$$
for some $\lambda_j \geq 0$.

- Since constraints for which $g_j(x) > 0$ cannot excert a force on the test mass, we must set $\lambda_j > 0$ for these indices, or equivalently, the equation $\lambda_j g_j(x) = 0$ must hold for all $j \in \mathcal{I}$.

**What about equality constraints?**

Replacing $g_i(x) = 0$ by the two inequality constraints $g_i(x) \geq 0$ and $-g_i(x) \geq 0$, our mechanistic interpretation yields two parallel surfaces $G_i^+$ and $G_i^-$, leaving an infinitesimally thin space between them within which our point mass is constrained to move.



The net reaction force of the two surfaces is of the form

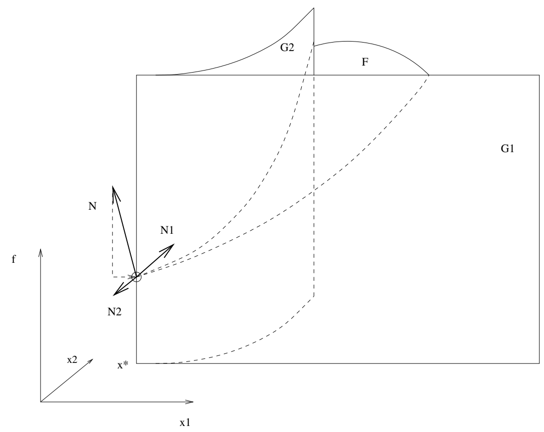
$$\lambda_i^{+}\nabla g_i(x) + \lambda_i^{-}\nabla(-g_i)(x) = \lambda_i \nabla g_i(x),$$

where we replaced the difference $\lambda_i^{+} - \lambda_i^{-}$ of the bound-constrained variables $\lambda_i^{+}, \lambda_i^{-} \geq 0$ by a single unconstrained variable $\lambda_i = \lambda_i^{+} - \lambda_i^{-}$.

Note that in this case the conditions $\lambda_i^{+} g_i(x) = 0$, $\lambda_i^{-}(-g_i(x)) = 0$ are satisfied automatically, since $g_i(x) = 0$ if $x$ is feasible.

There are situations in which our mechanical picture is flawed: if two inequality constraints have first order contact at a local minimiser then they cannot annul the horizontal part of $\vec{N}_f$.

When there are more constraints constraints, then generalisations of this situation can occur. In order to prove the KKT conditions, we must therefore make a regularity assumption like the LICQ.




**Reading Assignment:** Lecture-Note 9.