MATHEMATICAL OPTIMIZATION

A. M. ATTO, E. TROUVÉ

Polytech Annecy Chambéry - University of Savoie
E-mails: {abatt, etrou}@univ-savoie.fr

IAI - MATH 642

Version provisoire - March 24, 2015

Contents

1 Introduction to mathematical optimization

1.1 Definitions - Notations

$$\min_{x \in \mathcal{S}} f(x) \tag{1}$$

where

$$\begin{cases} \mathcal{S} = \{x \in X \mid g_i(x) \leqslant 0 : i = 1, \ldots, m;\; h_k(x) = 0 : k = 1, \ldots, m'\} \\ X \subset \mathbb{R}^n \\ f, g_i, h_k : X \longrightarrow \mathbb{R}^q \end{cases}$$

⇢ *Objective* or *Cost* function : f.

- ⇝ *Scalar objective* / *Scalar optimization*, function $f : D \subset X \rightarrow \mathbb{R}$.
- ⇝ *Vector objective* / *Vector optimization*, function $f : D \subset X \rightarrow \mathbb{R}^m$, $m > 1$. Imposes redefining the concept of optimization since several definitions of optimal points are possible (several possibilities for $\mathbb{R}^m$ ordering).

⇢ *Minimization* / *Maximization* problems are 'equivalent' in the sense that function f satisfies

$$\min_{x \in \mathcal{S}} f(x) = -\max_{x \in \mathcal{S}} \{-f(x)\}.$$

⇢ *Set of feasible solutions* / *Feasible set* / *Constraint set*, $\mathcal{S}$.

⇢ *Global minimum* / *Minimum* at x^*: $f(x^*) \leqslant f(x)$ for every $x \in \mathcal{S}$.

⇢ *Local minimum* at x°: $f(x^\circ) \leqslant f(x)$ for every $x \in \mathcal{S} \cap V(x^\circ)$, where $V(x^\circ)$ is a neighborhood of x°.

⇢ *Unconstrained* / *free problem*, when $\mathcal{S}$ coincides with the domain (open set) D of f, or when $\mathcal{S}$ is an open subset of D; *constrained* otherwise.

⇢ *Inconsistent problem*: no feasible solution exists due to inconsistent constraints (example of $5 - x \leqslant 0$ and $x - 1 \leqslant 0$).

⇢ *Level set*: $\{x \in X : f(x) = C\}$, where C is a constant.

1.2 Set and function properties

⇢ Convex sets / non-convex sets.

Let $a, b \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

Define the closed line segment $[a, b]$ as:

$$[a, b] = \{x \in \mathbb{R}^n \mid x = \lambda a + (1 - \lambda) b, \;\; 0 \leqslant \lambda \leqslant 1\}.$$

The open line segment $]a, b[$ is defined similarly by:

$$]a, b[= \{x \in \mathbb{R}^n \mid x = \lambda a + (1 - \lambda) b, \;\; 0 < \lambda < 1\}.$$

The set X is convex if it contains the closed line segment joining every two points $a, b \in X$.

⇢ Convex objective functions / non-convex objective.

A function $f : X \subset \mathbb{R}^n \longrightarrow \mathbb{R}$, where X is a convex set, is said to be convex on X if:

$$f(\lambda a + (1 - \lambda) b) \leqslant \lambda f(a) + (1 - \lambda) f(b)$$

for every $a, b \in X$ and every $\lambda \in [0, 1]$. A particular case of non-convex function: if $-f$ is convex on X, then f is said to be concave on X.

⇢ Convexity preservation / adding to f, a convex function, multiplying f by a scalar, composition of f with a linear function.

⇢ Convexity and minimisation / if f and $\mathcal{S}$ are convex, then the set of solutions of Eq. (1) is convex. This implies no local minima in $\mathcal{S}$.

⇢ Strict convexity and minimisation / at most one minimiser of f in $\mathcal{S}$ when $\mathcal{S}$ is convex and f is strctly convex.

non-convex objective.

⇢ Smooth objective / ~~non-smooth optimization problem~~.

Smoothness of the optimization problem relates to the differentiability of f. In particular, for a function $f : X \longrightarrow \mathbb{R}$ being $\mathcal{C}^2(X)$ on the open convex set $X \subset \mathbb{R}^n$, the gradient and the Hessian of f are useful for quantities for solving optimization problems.

The Gradient vector of f at $a \in X$ is given by

$$\nabla f(a) = \left(\frac{\partial f(a)}{\partial a_1}, \frac{\partial f(a)}{\partial a_2}, \ldots, \frac{\partial f(a)}{\partial a_n}\right)^t, \qquad \text{with}(x \in X) \quad f(x) - f(a) = (x-a)^t \nabla f(a) + o(||x-a||).$$

The Hessian matrix $Hf(a)$ of f at $a \in \mathbb{R}^n$ is the $n \times n$ matrix:

$$\nabla^2 f(a) = \begin{bmatrix} \frac{\partial^2 f(a)}{\partial a_1^2} & \frac{\partial^2 f(a)}{\partial a_1 \partial a_2} & \cdots & \frac{\partial^2 f(a)}{\partial a_1 \partial a_n} \\ | & | & & | \\ \frac{\partial^2 f(a)}{\partial a_n \partial a_1} & \frac{\partial^2 f(a)}{\partial a_n \partial a_2} & \cdots & \frac{\partial^2 f(a)}{\partial a_n^2} \end{bmatrix}, \quad \text{with} \quad f(x)-f(a) = (x-a)^t \nabla f(a) + \frac{1}{2}(x-a)^t \nabla^2 f(a)(x-a) + o(||x-a||^2).$$

⇢ Smooth and convex function over a Convex set

When f is differentiable on the open convex set $X \subset \mathbb{R}^n$, then f is convex on X if and only if its gradient is a monotone map on X, *i.e.*

$$(a-b)^t [\nabla f(a) - \nabla f(b)] \geqslant 0.$$

Let $f : X \longrightarrow \mathbb{R}$ be $\mathcal{C}^{\in}(\mathcal{X})$ on the open convex set $X \subset \mathbb{R}^n$. Function f is convex on X if and only if its Hessian matrix $Hf(x)$ is semidefinite positive on X, *i.e.*

$$x^t [Hf(x)] x \geqslant 0 \quad \forall x \in X$$

(in practice, this is equivalent to the non-negativity for the determinants of submatrices of $Hf(x)$, or equivalently, the non-negativity for the eigenvalues of $Hf(x)$).

Specific points in optimization problems

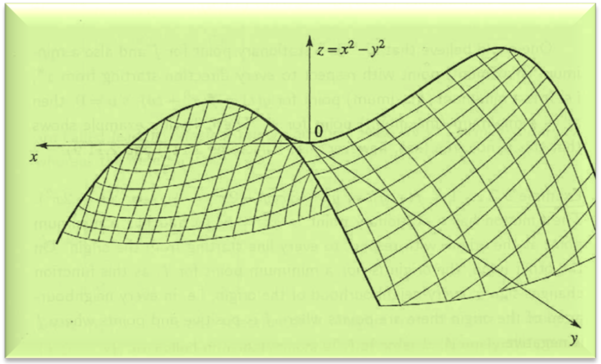
⇢ Critical point, x such that $\nabla f(x)$ is not well-defined.

⇢ Stationary point, x such that $\nabla f(x) = 0$.

⇢ Saddle point, stationary point but not a local extremum.

Saddle point of function $z = x^2 - y^2$.

At $x^\circ = 0$, we have:
f has a minimum with respect to x-axis,
f has a maximum with respect to y-axis.



Problem: $\min_{\mathcal{S}} f(x)$,
$f(x) = |x_1 - 2| + |x_2 - 2|$,
$\mathcal{S} = \{x, | \, x \in \mathbb{R}^2, g_1(x) \leqslant 0 \,;\, h_1(x) = 0\}$
$g_1(x) = (x_2)^2 - x_1$
$h_1(x) = (x_1)^2 + (x_2)^2 - 1$

Feasible set: within parabola,
Level sets: dashed lines,

Solution: $x^* = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$.

If we remove h_1, solution is $x^* = \left(2, \sqrt{2}\right)$.

If we remove g_1 and h_1, solution is $x^* = (2, 2)$.

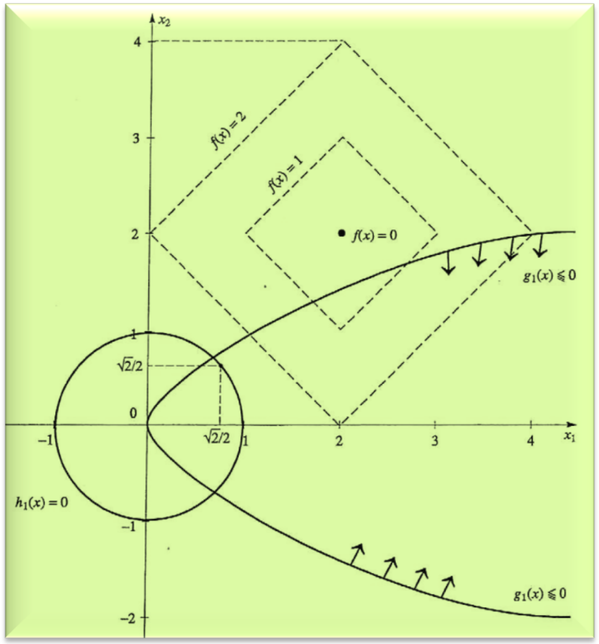


Figure 1: Example of objective functions and some stationary points.

1.3 Mathematical optimization problems

- *Linear programming problem* : all the functions involved in Eq. (1) are linear.
- *Non-Linear programming problem* : Eq. (1) involves at least one non-linear function.
  - Standard optimization problem, $\mathbf{X}$ is an open subset of $\mathbb{R}^n$ and there are no inequality.
  - Convex programming problem f and g_i are convex functions, all h_k are linear functions.
  - Quadratic programming problem, f is a quadratic form and the constraints are linear
  - Separable programming problem, f and g_i and h_k are separable functions

$$f(x) = \sum_{j=1}^{n} f_j(x_j), \qquad g_i(x) = \sum_{j=1}^{n} g_{i,j}(x_j), \qquad h_k(x) = \sum_{j=1}^{n} h_{k,j}(x_j).$$

  - Fractional or hyperbolic programming problem,

$$f(x) = \frac{u(x)}{v(x)}.$$

- **Issues** raised by *Mathematical optimization problems* :
  - Smoothness (differentiability) property of f;
  - Optimality conditions / existennce of solutions for Eq. (1);
  - Effective methods for finding solutions of Eq. (1).
- **Specific formulations** of optimization problems:
  - *Stochastic programming problem* (involves random variables) ;
  - *Integer programming problem* (some components $x \in \mathcal{S}$ must be integers, for instance in image processing, management science or operational research) ;
  - *Dynamic optimization problem* (time variation, for instance in control theory and calculus of variations).

2 Linear programming problem [all functions involved in Eq. (1) are linear]

$$f(x) = c^t x$$

for some $c \in \mathbb{R}^n$ and, for some $A \in \mathbb{R}^m \times \mathbb{R}^n$ and $b \in \mathbb{R}^m$, the inequality constraints have the form:

$$Ax \leqslant b$$

- f is a linear function, thus f is a convex function over $\mathbb{R}^n$.
- $X = \{x \in \mathbb{R}^n, Ax \leqslant b\}$ is a convex set (polyhedron, polytope).
- If the linear optimization problem is consistent and the optimal value is finite (bounded problem), then the problem admits a solution (there may exist several solutions) on a corner (vertex) [sommet] of the polyhedron.
- For $\mathbb{R}^n$ polyhedrons, vertices occur at points where n constraints intersect.
- Finding where n constraints intersect relates solving a system of n equations (Gauss-Seidel, Jacobi, QR decomposition, Cholesky factorization, *etc.*). Finding vertices then involves solving a series of systems of n equations.

$$\begin{array}{ll} \text{Maximize:} & x - y; \\ \text{Subject to:} & \frac{1}{3}x + y \leqslant 4, \\ & -x + y \leqslant 2, \\ & 0 \leqslant x \leqslant 3, \\ & y \geqslant 0. \end{array}$$

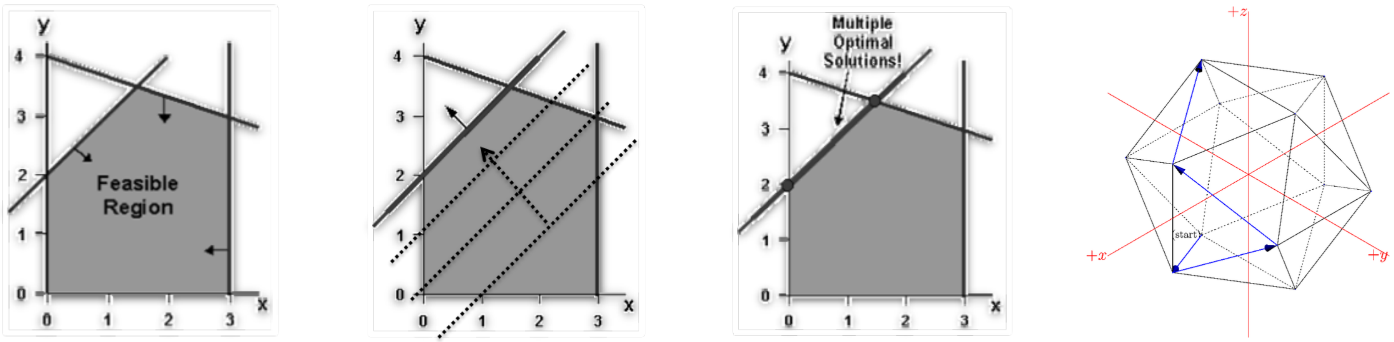


Figure 2: Simplex algorithm moves adaptively from vertex to vertex until an optimal solution has been found. Each vertex that it visits is an improvement over the previous one. Once it can't find a better vertex, it decides that an optimal solution has been reached.

3 Non-Linear programming problem

3.1 Non-linear optimization: Standard optimization problems

[$\mathbf{X}$ is an open subset of $\mathbb{R}^n$ and there are no inequality constraints]

3.1.1 Optimization with no constraints

We assume that $f : \mathbf{X} \longrightarrow \mathbb{R}$ is $\mathcal{C}^2(\mathbf{X})$ on the open convex set $\mathbf{X} \subset \mathbb{R}^n$. Let x^* be a solution of Eq. (1), $f(x^*) = \min_{x \in \mathbb{R}^n} f(x)$.

Optimality conditions

Theorem 1 (Necessary condition of 1$^{\text{rst}}$ order)
If x^* is a local minimum of f on $\mathbb{R}^n$, then $\nabla f(x^*) = 0$ [x^* is thus a critical point].

(Nota: null gradient does not imply a local minimum, examples of saddle points).

Theorem 2 (Necessary condition of 2$^{\text{nd}}$ order)
If x^* is a local minimum of f on $\mathbb{R}^n$, then $\nabla f(x^*) = 0$ and

$$y^t \nabla^2 f(x^*) y \geq 0 \quad \forall y \in \mathbb{R}^n$$

[x^* is thus a critical point having positive semi-definite Hessian matrix $\nabla^2 f(x^*)$].

Theorem 3 (Sufficient condition of 2$^{\text{nd}}$ order)
Let $x^* \in \mathbb{R}^n$. If

$$\nabla f(x^*) = 0 \quad \text{and} \quad y^t \nabla^2 f(x^*) y > 0 \quad \forall y \in \mathbb{R}^n \quad \text{[positive definite Hessian]}$$

then x^* is the minimum of f on $\mathbb{R}^n$.

Gradient descent

Principle: generating a series $(x^k)_k$ such that $\left(f(x^k)\right)_k$ is decreasing, with

$$x^{k+1} = x^k + \lambda^k \mathbf{d}^k$$

[$\mathbf{d}^k$ is the descent direction : $\mathbf{d}^k = -\nabla f(x^k)$ and the descent step λ^k is positive].

--→ $\mathbf{d} \in \mathbb{R}^n$ is a descent direction of f in $x \in \mathbb{R}^n$ iff $f(x + \lambda\mathbf{d}) < f(x)$ for every small value $\lambda > 0$, thus iff $\mathbf{d}^t \nabla f(x) < 0$. The descent direction reduces to $\mathbf{d}^k = -\nabla f(x^k)$;

$$x^{k+1} = x^k - \lambda^k \nabla f(x^k).$$

Algorithm of the gradient descent

1. Initializing : $x^k = x^0 \in \mathbb{R}^n$ ($k = 0$).
2. Computing the gradient descent direction $\mathbf{d}^k = -\nabla f(x^k)$.
3. Convergence test: If $||\nabla f(x^k)|| \cong 0$, stop (critical point x^k). Else, continue:
4. Selecting / Computing step λ^k
5. $x^k \leftarrow x^k + \lambda^k \mathbf{d}^k$; $k \leftarrow k + 1$; return to item 2.

Remark [Selection of the steps]

--→ Constant step: $\lambda^k = \lambda$: guaranty neither descent, nor convergence.

--→ Linear search of the step: $\lambda^k = \arg\min_\lambda f(x^k - \lambda^k \nabla f(x^k))$, convergent, but non-realistic in many cases (non-linearity).

--→ Decreasing sequence of steps:

$$\begin{cases} \lim_{k\to\infty} \sum_k \lambda^k = \infty \\ \lim_{k\to\infty} \sum_k (\lambda^k)^2 < \infty \end{cases}$$

--→ Adaptive steps ...

3.1.2 Optimization with equality constraints

Function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $\mathcal{S} \subset \mathbb{R}^n$ involves m equality constraints:

$$\mathcal{S} = \{x \in \mathbb{R}^n \,/ h_j(x) = k_j,\ 1 \leq j \leq m\},$$

where $h_j : \mathbb{R}^n \longrightarrow \mathbb{R}$, with h_j differentiable and k_j constant. Let x^* be a solution:

$$f(x^*) = \min_{x \in \mathcal{S}} f(x), \qquad h_j(x^*) = k_j,\ 1 \leq j \leq m$$

Lagrange multiplier for one equality constraint

We consider hereafter a single constraint having the form $\mathcal{S} = \{x \in \mathbb{R}^n \,/ h(x) = k\}$, with $f, h \in \mathcal{C}^1(\mathcal{S})$.

Theorem 4 (First order Lagrange necessary condition)
If x^* is a stationary point of f on $\mathcal{S}$ and if $\nabla h(x^*) \neq 0$, then the gradient of f and the gradient of h are colinear:

$$\exists \lambda_0 \in \mathbb{R} \,/\, \nabla f(x^*) = \lambda_0 \nabla h(x^*) \tag{2}$$

The Lagrange method consists in introducing a new variable λ called a Lagrange multiplier and study the Lagrange function (or Lagrangian) defined by

$$\mathcal{L}(x, \lambda) = f(x) + \lambda(h(x) - k),$$

given that: ***if $f(x^*)$ is a minimum of the original constrained problem, then there exists λ_0 such that (x^*, λ_0) is a stationary point for the Lagrange function $\mathcal{L}$.***

Remarks :

1. A stationary point is not necessarily an extremum. A study of the solutions of Eq. (2) in order to check for minimality of f on $\mathcal{S}$.

2. The Lagrange multiplier λ represents the rate of variation of the critical value $f(x^*(k))$ over the constraint $h(x) = k$ when k varies.

Lagrange multipliers for m equality constraints

Consider m equality constraints expressed as $\mathcal{S} = \{x \in \mathbb{R}^n / h_j(x) = k_j\}$.

Theorem 5 (First order Lagrange necessary condition)
If x^* is a local minimum of f on $\mathcal{S}$ and if $\{\nabla h_j(x^*) : 1 \leq j \leq m\}$ is a linearly independent set, then there exist $(\lambda_j)_j$ such that:

$$\nabla f(x^*) = \sum_{j=1}^{m} \lambda_j \nabla h_j(x^*) \tag{3}$$

Coefficients $(\lambda_j)_j$ are called Lagrange multipliers associated with extremum x^*.

Procedure 1 (Solution search by using Lagrange Multipliers)

1. *Provide the following system associated with $(n+m)$ variables $(x_i)_{1\leq i\leq n}$, $(\lambda_j)_{1\leq j\leq m}$ and $(n+m)$ equations:*

$$\left\{\begin{array}{rcl}
\frac{\partial f(x)}{\partial x_1} & = & \lambda_1 \frac{\partial h_1(x)}{\partial x_1} + \ldots + \lambda_m \frac{\partial h_m(x)}{\partial x_1} \\
\cdots & & \\
\frac{\partial f(x)}{\partial x_n} & = & \lambda_1 \frac{\partial h_1(x)}{\partial x_n} + \ldots + \lambda_m \frac{\partial h_m(x)}{\partial x_n} \\
h_1(x) & = & k_1 \\
\cdots & & \\
h_m(x) & = & k_m
\end{array}\right.$$

2. *Search stationary points of this system,*

3. *Look for minimum on $\mathcal{S}$ among these stationary points.*

3.2 Non-linear optimization: Convex programming problem

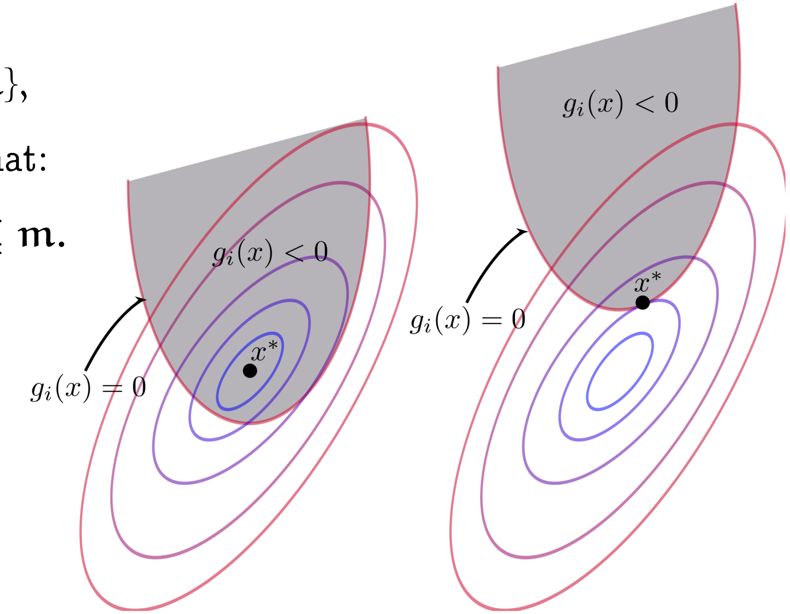
[f and g_i are convex functions, all h_k are linear functions] (we assume no equality constraint first)

Given a set $\mathcal{S} \subset \mathbb{R}^n$ associated with m equality constraints:

$$\mathcal{S} = \{x \in \mathbb{R}^n / g_j(x) \leq 0,\ 1 \leq j \leq m\},$$

and $f, g \in \mathcal{C}^1(\mathcal{S})$, a solution x^* of the optimization problem of Eq.(1) is such that:

$$f(x^*) = \min_{x \in \mathcal{S}} f(x),\ g_j(x^*) \leq 0,\ 1 \leq j \leq m.$$



Definition 1 (Regularity conditions (Constraints qualifications))
Let $x_0 \in \mathcal{S}$ and $\mathbf{I}(x_0)$ the set of indices of constraints such that x_0 satisfy:

$$\mathbf{I}(x_0) = \{1 \leq j \leq m / g_j(x_0) = 0\}$$

Constraints are said to be qualified at point x_0 if:

- *either all functions g_j, $j \in \mathbf{I}(x_0)$, are affines,*
- *or there exist $w \in \mathbb{R}^n / \forall j \in \mathbf{I}(x_0)$:*

$$\begin{aligned} \nabla g_j(x_0)^t w &\leq 0, \\ \nabla g_j(x_0)^t w &< 0 \text{ if } g_j \text{ is not affine.} \end{aligned}$$

Theorem 6 (Necessary Kuhn-Tucker conditions)
If x^* is a local minimum of f on $\mathcal{S}$ and if the constraints are qualified at x^*, then there exist coefficients λ_j, $j \in I(x^*)$ such that:

$$\nabla f(x^*) + \sum_{j \in I(x^*)} \lambda_j \nabla g_j(x^*) = 0 \text{ and } \lambda_j \geq 0,\ j \in I(x^*) \tag{4}$$

Conditions expressed in Eq (4) are called Kuhn and Tucker conditions.

Remarks:

1. Given a local minimum x^*, coefficients λ_j, $j \in I(x^*)$ are not necessarily unique, provided if gradients $\nabla g_j(x^*)$, $j \in I(x^*)$ are linearly independants.

2. If $I(x^*) = \emptyset$, Kuhn-Tucker conditions reduce to $\nabla f(x^*) = 0$, (standard necessary condition for extremum over an open set and x^* given in the interior of $\mathcal{S}$).

3. By letting $\lambda_j = 0$, $j \notin I(x^*)$, we obtain:

$$\begin{cases} \nabla f(x^*) + \sum_{j=1}^{m} \lambda_j \nabla g_j(x^*) = 0, \\ \lambda_j \geq 0,\ 1 \leq j \leq m,\ \sum_{j=1}^{m} \lambda_j g_j(x^*) = 0, \\ g_j(x^*) \geq 0,\ 1 \leq j \leq m. \end{cases}$$

Procedure 2 (Optimisation under inequality constraint)

1. *Identify the stationary points strictly pertaining to the $\mathcal{S}$ (pertaining to the interior of $\mathcal{S}$):*

$$x^* \,/\, g(x^*) < 0 \text{ et } \nabla f(x^*) = 0;$$

2. *Identify the stationary points on the frontier of $\mathcal{S}$ ($g(x) = 0$) by using Lagrange multiplier method (Procedure 1),*

3. *Evaluate f at stationary points and deduce the global minimum.*

3.3 Non-linear optimization: Quadratic programming problem

f is a quadratic form and the constraints are linear

3.4 Non-linear optimization: Separable programming problem

f and g_i and h_k are separable functions

$$f(x) = \sum_{j=1}^{n} f_j(x_j)$$

$$g_i(x) = \sum_{j=1}^{n} g_{i,j}(x_j)$$

$$h_k(x) = \sum_{j=1}^{n} h_{k,j}(x_j)$$

3.5 Non-linear optimization: fractional or hyperbolic programming problem

$$f(x) = \frac{u(x)}{v(x)}.$$