Introduction to optimization

Jean-François Aujol

CMLA, ENS Cachan, CNRS, UniverSud,
61 Avenue du Président Wilson, F-94230 Cachan, FRANCE

Email : Jean-Francois.Aujol@cmla.ens-cachan.fr
http://www.cmla.ens-cachan.fr/membres/aujol.html

27 March 2008

Note: This document is a working and uncomplete version, subject to errors and changes. Readers are invited to point out mistakes by email to the author.

This document is intended as course notes for master students. The author encourage the reader to check the references given in this manuscript. In particular, it is recommended to look at [5] which is closely related to the subject of this course.

Contents

1. Generalities

For this introductory section, we refer the reader to [5, 4, 6, 3].

1.1 Introduction

We consider a function $J : U \to \mathbb{R}$, with $U \subset \mathbb{R}^N$. We aim at minimizing it, i.e. at finding some $v \in U$ such that $J(v) = \text{argmin}_{u \in U} J(u)$.

Principle As we will see, all the methods are iterative ones. Starting from u_0, a sequence u_n is built. If the method works, $u_n \to u$ minimizer of J.

Some basic questions:

1. J is differentiable on U ?
2. $N = 1$ or $N > 1$?
3. Is U an open set in $\mathbb{R}^N$?

Differentiability: If J is not differentiable, then the problem is much harder to handle.

Number of variables: The case $N = 1$ is much simpler. In particular, we can make use of the fact that $\mathbb{R}$ is ordered.

In theory, there is no difference between $N = 2$, $N = 3$, ... Notice however that the larger N, the most time consuming the minimization algorithm is.

The general idea in the case when $N > 1$ is to try to boil down to the case $N = 1$. Starting from u_0, one tries to find a good descent direction v, and then minimize $t \mapsto J(u + tv)$.

Topology of U: This is also a major issue. The easiest case is the one when U is an open set of $\mathbb{R}^N$.

Indeed, we have the following standard result:

Theorem 1.1. *Let J a differentiable function on an open set U. If $u \in U$ is a minimizer of J, then $\nabla J(u) = 0$.*

1.2 Existence of a minimizer in finite dimension

J is defined on $U \subset \mathbb{R}^N$. We denote by $E = \mathbb{R}^N$.

In this lecture, we will always assume that $E = \mathbb{R}^N$. Notice however that all the results hold if E is a separable Hilbert space.

We remind the reader that $\mathbb{R}^N$ is embeded with the standard euclidean inner product.

$$\langle u, v \rangle = \sum_{i=1}^{N} u_i v_i \tag{1.1}$$

and $\|u\|^2 = \langle u, u \rangle$.

We have the Cauchy Scwhartz inequality:

$$\langle u, v\rangle \leq \|u\|\|v\| \tag{1.2}$$

We remind the reader that in $\mathbb{R}^N$, a set is compact iff it is closed and bounded.

Theorem 1.2. *Assume U a closed non empty set in E. If U is not bounded, then we also assume that J is coercive on U. Then there exists $\hat{u} \in U$ such that*

$$J(\hat{u}) = \operatorname*{argmin}_{u\in U} J(u) \tag{1.3}$$

The proof is straightforward using compacity and a minimizing sequence.

1.3 Differentiability

First derivative: We consider here $J : U \subset X \to \mathbb{R}$ where $E = \mathbb{R}^N$. J is differentiable at some point $u \in U$ iff there exists $v \in E'$ which we denote by $\nabla J(u)$ such that:

$$J(u + h) = J(u) + \langle \nabla J(u), h\rangle + \|h\|\epsilon(h) \tag{1.4}$$

with $\lim_{h\to 0} \epsilon(h) = 0$. If v exists, it is then easy to show that it is unique (Riesz representation theorem). $\langle ., .\rangle$ stands for the duality product. In the case when $E = \mathbb{R}^N$, then $\langle ., .\rangle$ is simply the usual euclidian inner product.

Notice that if J is differentiable, then J admits partial derivative with respect to each of its variables. But the controverse does not hold. Cosider for instance $f(x_1x_1) = 0$ if $x_1x_2 = 0$, and 1 otherwise.

We have the following basic result:

Proposition 1.1. *J differentiable on U. If the segment $[u, v] \subset U$:*

$$\|J(u) - J(v)\| \leq \sup_{[u,v]} \|\nabla J\|\|u - v\| \tag{1.5}$$

Proposition 1.2. *Taylor formula (order 1):*

$$J(u + v) = J(u) + \int_0^1 \langle \nabla J(u + tv), v\rangle \, dt \tag{1.6}$$

Second derivative If $\nabla J : X \to X'$ is differentiable in u, then we denote by $\nabla^2 J(u)$ its derivative, which belongs to $\mathcal{L}(X; X')$. Since this last space is isomorphic to $\mathcal{L}_2(X; \mathbb{R})$ of bi-linear continuous applications from X to $\mathbb{R}$, the second derivative of J is identified with a continuous bilinear application. Moreover, it is easy to see that $\nabla^2 J(u)$ is a *symetric* bilinear application (using theoreme des accroissements finis).

The Taylor expansion to the second order is

$$J(u + h) = J(u) + \langle \nabla J(u), h\rangle + \frac{1}{2}\langle h, \nabla^2 J(u)h\rangle + o(\|h\|^2)\epsilon(h) \tag{1.7}$$

with $\lim_{h\to 0} \epsilon(h) = 0$.

Proposition 1.3. *Taylor formula (order 2):*

$$J(u+v) = J(u) + \langle \nabla J(u), v\rangle + \int_0^1 (1-t)\langle \nabla^2 J(u+tv)v, v\rangle \, dt \tag{1.8}$$

Stokes fomula We define div $= \nabla^T$. We thus have:

$$\langle \nabla u, \nabla v\rangle = \langle \Delta u, v\rangle \tag{1.9}$$

Typical funtional

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle \tag{1.10}$$

with A symmetric.

We have $\nabla J(u) = Au - b$, and $\nabla^2 J(u) = A$.

Remember Tychonov regularization:

$$J(u) = \frac{1}{2}\|Au - f\|^2 + \|\nabla u\|^2 \tag{1.11}$$

We have: $\|\nabla u\|^2 = -\langle \Delta u, \rangle$ and $\|Au - f\|^2 = \|Au\|^2 - 2\langle Au, f\rangle + \|f\|^2$. Hence:

$$J(u) = \langle \frac{1}{2}(A^T A - \Delta)u, u\rangle - \langle u, A^T f\rangle + Cste \tag{1.12}$$

1.4 Conditions

Necessary condition:

Theorem 1.3. *Let J a differentiable function on an open set U. If $u \in U$ is a minimizer of J, then $\nabla J(u) = 0$.*

The proof is straightforward with Taylor expansion (order 1).

$$0 \leq J(u+h) - J(u) \leq \langle \nabla J(u), h\rangle + o(\|h\|) \tag{1.13}$$

and then with $-h$.

Necessary condition:

Theorem 1.4. *Let $J : U \to \mathbb{R}$ a twice differentiable function on an open set U. If $u \in U$ is a minimizer of J, then $\nabla^2 J(u)(w, w) \geq 0$ for all $w \in U$.*

The proof is straightforward with Taylor-Young expansion (order 2) and the previous result.

$$0 \leq J(u+h) - J(u) \leq \langle \underbrace{\nabla J(u)}_{=0}, h\rangle + \frac{1}{2}\langle h, \nabla^2 J(u)h\rangle + o(\|h\|^2) \tag{1.14}$$

Sufficient condition:

Theorem 1.5. *Let $J : U \to \mathbb{R}$ a twice differentiable function on an open set U. We assume that there exists $u \in U$ such that $\nabla J(u) = 0$. Let us assume that there exists $\alpha > 0$ such that $\nabla^2 J(u)(w, w) \geq \alpha\|w\|^2$ for all $w \in U$. Then J admits a strict minimum in u.*

The proof is straightforward with Taylor expansion (order 2).

1.5 Convexity

1.5.1 Characterization

Definition 1.1. $U \subset E$ is convex if

$$\lambda x + (1-\lambda)y \in U \tag{1.15}$$

for all x, y in U and $\lambda \in [0,1]$.

Necessary condition:

Theorem 1.6. *Let $J : U \to \mathbb{R}$ a differentiable function on a convex set U. If $u \in U$ is a minimizer of J, then $\langle \nabla J(u), (v-u)\rangle \geq 0$ for all $v \in U$.*

Proof: Remark that if u and $v = u + w$ in U convex, then $u + tw$ is in U for all $t \in [0,1]$. We then apply Taylor formula:

$$0 \leq J(u+tw) - J(u) = t\langle \nabla J(u), w\rangle + o(t\|w\|) \tag{1.16}$$

and $w = v - u$.

■

Notice that if U is a subspace of E, then the necessary condition becomes: $\langle \nabla J(u), v\rangle = 0$ for ll $v \in U$.

Notice also that if $U = E$, then the condition is the classical Euler equation $\nabla J(u) = 0$.

Definition 1.2. $J : U \to \mathbb{R}$ is convex if

$$J(\lambda x + (1-\lambda)y) \leq \lambda J(x) + (1-\lambda)J(y) \tag{1.17}$$

for all x, y in E and $\lambda \in [0,1]$.

Recall also the notion of strict convexity, and of concavity.

Theorem 1.7. *Let $J : U \to \mathbb{R}$ a differentiable function on a convex set U.*

- *J is convex iff*

$$J(v) \geq J(u) + \langle \nabla J(u), v-u\rangle \tag{1.18}$$

  for all u, v in U.

- *J is strictly convex iff*

$$J(v) > J(u) + \langle \nabla J(u), v-u\rangle \tag{1.19}$$

  for all $u \neq v$ in U.

Theorem 1.8. *Let $J : U \to \mathbb{R}$ a twice differentiable function on a convex set U. J is convex iff*

$$\langle \nabla^2 J(u)(v-u), v-u\rangle \geq 0 \tag{1.20}$$

for all u, v in U.

Example:

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle \tag{1.21}$$

with $A = A^T$ (i.e. A symmetric).

J is convex iff A is positive.

J is strictly convex iff A is strictly positive.

1.5.2 Global minimizer

Convex function: local minimizer is a global minimizer!

Theorem 1.9. *Let $J : U \to \mathbb{R}$ a convex function defined on a convex set U.*

1. *If J admits a local minimum in $u \in U$, then this is in fact a global minimum in U.*
2. *If J is strictly convex, then it has at most one minimum and it is strict.*
3. *Assume J is differentiable in $u \in U$. Then J admits a minimum in u iff*

$$\langle \nabla J(u), (v - u)\rangle \geq 0 \tag{1.22}$$

   for all v in U.

4. *If U is open, then the previous condition is equivalent to the Euler equation $\nabla J(u) = 0$.*

1.6 Ellipticity

Definition 1.3. $J : E \to \mathbb{R}$ is said to be *elliptic* if it is continuously differentiable on U, and if there exists a constant $\alpha > 0$ such that:

$$\langle \nabla J(v) - \nabla J(u), v - u\rangle \geq \alpha \|v - u\|^2 \tag{1.23}$$

for all u, v in U.

Theorem 1.10.

1. *If $J : E \to \mathbb{R}$ is elliptic, then J is strictly convex and coercive, and satisfies:*

$$J(v) - J(u) \geq \langle \nabla J(u), v - u\rangle + \frac{\alpha}{2}\|v - u\|^2 \tag{1.24}$$

2. *If U is a non empty convex closed set in E, and if J is elliptic, then J admits one and only one minimizer on U.*
3. *If J is elliptic, and U convex, then $u \in U$ is a minimizer of J iff for all $v \in U$:*

$$\langle \nabla J(u), v - u\rangle \geq 0 \tag{1.25}$$

   or if $E = U$:

$$\nabla J(u) = 0 \tag{1.26}$$

4. If J is twice differentiable on E, then J is elliptic iff

$$\langle \nabla^2 J(u)w, w\rangle \geq \alpha \|w\|^2 \tag{1.27}$$

for all $w \in E$.

Indeed:

$$J(v)-J(u) = \int_0^1 \langle \nabla J(u+t(v-u)), v-u\rangle \, dt = \langle \nabla J(u), v-u\rangle + \int_0^1 \langle \nabla J(u+t(v-u))-\nabla J(u), v-u\rangle \, dt \tag{1.28}$$

Hence:

$$J(v) - J(u) \geq \langle \nabla J(u), v-u\rangle + \int_0^1 \alpha t \|v-u\|^2 \, dt = \langle \nabla J(u), v-u\rangle + \frac{\alpha}{2}\|v-u\|^2 \tag{1.29}$$

In particular, we have if $u \neq v$:

$$J(v) - J(u) > \langle \nabla J(u), v-u\rangle \tag{1.30}$$

And

$$J(v) \geq J(0) + \langle \nabla J(0), v\rangle + \frac{\alpha}{2}\|v\|^2 \geq J(0) - \|J(0)\|\|v\| + \frac{\alpha}{2}\|v\|^2 \tag{1.31}$$

1.7 Algorithm

Definition 1.4. $x_0 \in E = \mathbb{R}^N$.

$$x_{k+1} = \mathcal{A}(x_k) \tag{1.32}$$

Definition 1.5. The algorithm $\mathcal{A}$ is said to be convergent if the sequence x_k converges towards some $x \in E$.

Definition 1.6. Convergence rate:

Let x_k a sequence defined by an algorithm $\mathcal{A}$ and convergent towards some $x \in E$. The convergence of $\mathcal{A}$ is said to be:

- *Linear* if the error $e_k = \|x_k - x\|$ is linearly decreasing, i.e. $e_{k+1} \leq Ce_k$
- *Supra-linear* if the error $e_k = \|x_k - x\|$ decreases as $e_k \leq \alpha_k e_k$ with α_k a non-negative sequence decreasing to 0. If α_k is a geometric sequence, then the convergence is said to be geometric.
- *Of order p* if the error $e_k = \|x_k - x\|$ decreases as $e_k \leq C(e_k)^p$. If $p = 2$, the convergence is said to be quadratic.
- Local convergence if x_0 needs to be close to x; otherwise global convergence.

Fixed point theorem $F : X \to X$ is said to be a contraction if there exists $\gamma \in (0, 1)$ such that: $\|F(u) - F(v)| \leq \gamma \|u - v\|$.

Theorem 1.11. *Let X a Banach space. If $F : X \to X$ is a contraction, then F admits a unique fixed point $\hat{u}$ such that $F(\hat{u}) = \hat{u}$.*

2. Unconstrained optimization

2.1 The 1 dimensional case

For this subsection, we refer the reader to [6].

Here, we will denote the function J by f.

2.1.1 Basic algorithms

These algorithms do not require to estimate the derivative of f.

Assumption: f unimodal on $[a, b]$, i.e. f strictly decreasing on $[a, x^*[$ and strictly increasing on $]x^*, b]$.

Dichotomie algorithm Two points a and b such that $f(a)f(b) < 0$. The aim is the to find a 0 of f.

Computation of 5 values of f in $a = x_1 < x_2 < x_3 < x_4 < x_5 = b$.

Depending on f at least two points among the x_i can be removed. We are then in the situation: $a \leq y_1 < y_3 < y_5 \leq b$ and $f(y_1) > f(y_3) < f(y_5)$. Thus x^* lies in $]y_1, y_5[$.

The Dichotomie method consists in dividinc by 2 the intervall at each iteration. To this end, one just needs to divide the original intervall in 4 equal segments.

Golden section search It is the same principle as before, excpet that the intervall is divided in 3 at each iterations (and not in 4) (and thus only oneevaluation of f is needed at each iteration).

To find the minimizer of a function, 3 points a required: a, b, c, with $f(b) < \min(f(a), f(c))$. Look at x in (a, b) or (b, c). It is possible to choose the new point x such that the length of the new interval is $\gamma|a - c|$ where $\gamma = (\sqrt{5} - 1)/2$ (inverse of the golden number). This is a *linear convergent* algorithm.

More concretly: Computation of 5 values of f in $a = x_1 < x_2 < x_3 < x_4 = b$.

Depending on f, x^* will lie either in $]x_1, x_3[$ or in $]x_2, x_4[$. Assume for instance that x^* is in $]x_2, x_4[$. If x_3 is the middle of $[x_2, x_4]$, then the next intervalls cannot all have the same size.

To fix this problem, the idea is to make the whole length of the intervall L_k decreases at each iteration k. One wants to have: $\frac{L_{k+1}}{L_k} = \gamma < 1$. This implies that $L_k = L_{k+1} + L_{k+2}$. Divided this last equation by L_k, one gets the value of γ.

2.1.2 Steepest gradient descent method

Let us consider $f : \mathbb{R} \rightarrow$ Assume $a < b < c$, and $f(b) < \min(f(a), f(c))$. The next point to test is of the type: $b - \alpha f'(b)$ whith $\alpha > 0$.

Algorithm: Given $x_0 \in \Omega$, define the sequence:

$$x_{k+1} = x_k - \alpha f'(x_k) \tag{2.1}$$

2.1.3 Newton method

Basic idea: use of the condition $f'(u) = 0$. To find a minimum of J, one looks for a zero of f'. Notice that then it is needed to check that indeed the zero of f' is a minimizer of f.

Finding a zero of g Let us consider $g : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}$.

The principle of Newton method is to linearise the equation $g(x) = 0$ around x_k the current iteration:

$$g(x_k) + g'(x_k)(x - x_k) = 0 \tag{2.2}$$

If $g'(x_k) = 0$ then one needs to use a higher order Taylor expansion. Otherwise, the solution is

$$x = x_k - \frac{g(x_k)}{g'(x_k)} \tag{2.3}$$

And we thus choose $x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)}$.

Given $x_0 \in \Omega$, define the sequence:

$$x_{k+1} = x_k - \frac{g(x_k)}{g'(x_k)} \tag{2.4}$$

It has an immediate geometric interpetation: x_{k+1} is the intersection of the horizontal axis and the tangent to g in x_k.

To be well defined, x_k is to remain in Ω for all k, f is to be derivable in Ω, and $g' \neq 0$.

When it converges, Newton method has a quadratic convergence speed.

It is interesting to notice that such a method is immediate to generealize to the N dimensional case.

Finding a minimum of f One just applies the above algorithm to f'.

Given $x_0 \in \Omega$, define the sequence:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_{k+1})} \tag{2.5}$$

To be well defined, x_k is to remain in Ω for all k, f is to be twice derivable in Ω, and $f'' \neq 0$.

2.1.4 Secant method

Finding a zero of g Let us consider $g : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}$. We look for the zeros of a linear approximation of g.

Given $x_0, x_1 \in \Omega$, define the sequence:

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{g(x_k) - g(x_{k-1})} g(x_k) \tag{2.6}$$

2.2 Unconstrained problems

For this subsection, we refer the reader to [5, 6, 1].

2.2.1 Introduction

J functional defined on $E = \mathbb{R}^N$. Problem: find $u \in E$ such that:

$$J(u) = \inf_{v \in E} J(v) \tag{2.7}$$

$\Longrightarrow$ Iterative methods:

From an arbitrary $u_0 \in E$, a sequence u_k is built. u_k is expected to converge to a solution of problem (2.7).

To build u_{k+1} from u_k, we get back to a problem which is easy to solve numerically, i.e. to a problem with only one single real variable. To do so:

- A descent direction d_k is chosen from u_k.
- The minimum of J on the line passing in u_k parallel to d_k is computed. This defines u_{k+1} if there exists a unique solution $\rho(u_k, d_k)$ minimizing: $\rho \mapsto J(u_k + \rho d_k)$.

  Notice that in the case of a quadratic functional, finding ρ amounts to solving a second order polynomial.

2.2.2 Relaxation method

The simplest choice to define the successive descent directions consist in choosing them in advance. A canonical choice is the directions of the coordinates axis, taken in a cyclic way.

Algorithm: u_0 fixed in $\mathbb{R}^N$. $u_{k=1}$ is computedfrom u_k by sucessively solving the following minimizationproblems:

$$\begin{cases} J([u_1^{k+1}], u_2^k, \dots, u_N^k) = \inf_{\xi \in \mathbb{R}} J(\xi, u_2^k, \dots, u_N^k) \\ \dots \\ J(u_1^{k+1}, \dots, u_{N-1}^{k+1}, [u_N^{k+1}]) = \inf_{\xi \in \mathbb{R}} J(u_1^{k+1}, \dots, u_{N-1}^{k+1}, \xi) \end{cases} \tag{2.8}$$

We set:

$$u_k = u_{k,0} = \left(u_1^k, \dots, u_N^k\right) \tag{2.9}$$

and

$$\begin{cases} u_{k,1} = \left(u_1^{k+1}, u_2^k, \dots, u_N^k\right) \\ \dots \\ u_{k,N-1} = \left(u_1^{k+1}, \dots, u_{N-1}^{k+1}, u_N^k\right) \end{cases} \tag{2.10}$$

and

$$u_{k+1} = u_{k,N} = \left(u_1^{k+1}, \dots, u_N^{k+1}\right) \tag{2.11}$$

The minimization problems are thus (denoting by e_i the canonical basis of $\mathbb{R}^N$):

$$\begin{cases} J(u_{k,1}) = \inf_{\rho \in \mathbb{R}} J(u_{k,0} + \rho e_1) \\ \dots \\ J(u_{k,N}) = \inf_{\rho \in \mathbb{R}} J(u_{k,N-1} + \rho e_N) \end{cases} \tag{2.12}$$

Theorem 2.1. *If J is elliptic on $E = \mathbb{R}^N$, then the relaxation method converge.*

Remark: The differentiability of the functional is essential. Consider for instance:

$$J(v_1, v_2) = v_1^2 + v_2^2 - 2(v_1 + v_2) + 2|v_1 - v_2| \tag{2.13}$$

Exercice:

J is coercive, strictly convex, almost quadatric, but non differentiable.

If one chooses $u_0 = (0, 0)$, then the relaxation method leads to a stationary sequence $u_k = u_0$ for all k, although $\inf_{v \in \mathbb{R}^2} J(v) = J(1, 1)$.

Nevertheless, the relaxation method works for function of the type:

$$J(v) = J_0(v) + \sum_{i=1}^{N} \alpha_i |v_i| \tag{2.14}$$

with $\alpha_i \geq 0$ and J_0 elliptic.

Computation time: In general, the relaxation method is N times slower then the optimal step gradient method.

2.2.3 Optimal step gradient method

Intuitively, the metod will perform better if the differences $J(u_k) - J(u_{k+1})$ are large. The choice of the coordinates axis is thus not optimal. The idea to choose the descent direction is to use the opposite direction to the gradient (since it is locally the steepest descent) (clear using Young formula at first order: $J(u_k + w) = J(u_k) + \langle \nabla J(u_k), w \rangle + o(\|w\|)$).

Algorithm u_0 in E.

$$\begin{cases} J(u_k - \rho(u_k)\nabla J(u_k) = \inf_{\rho \in \mathbb{R}} J(u_k - \rho \nabla J(u_k)) \\ u_{k+1} = u_k - \rho(u_k)\nabla J(u_k) \end{cases} \tag{2.15}$$

Theorem 2.2. *Assume J to be elliptic. Then the optimal step gradient method converges.*

Sketch of the proof:

(i) We can assume that $\nabla J(u_k) \neq 0$ for all k (otherwise, convergence in a finite number of iterations.

$$\phi_k(\rho) = J(u_k - \rho \nabla J(u_k)) \tag{2.16}$$

$\phi_k : \mathbb{R} \to \mathbb{R}$. It is easy to see that ϕ_k is strictly convex, coercive, and therefore admits a unique minimizer characterized by: $\phi_k^{'}(\rho(u_k)) = 0$.

$$\phi_k^{'}(\rho) = -\langle \nabla J(u_k - \rho \nabla J(u_k)), \nabla J(u_k) \rangle \tag{2.17}$$

Hence

$$\langle \nabla J(u_k), \nabla J(u_{k+1}) \rangle = 0 \tag{2.18}$$

Since $u_{k+1} = u_k - \rho \nabla J(u_k)$, we also have:

$$\langle u_{k+1} - u_k, \nabla J(u_{k+1}) \rangle = 0 \tag{2.19}$$

Since J elliptic, we get

$$J(u_k) - J(u_{k+1}) \geq \frac{\alpha}{2} \|u_k - u_{k+1}\|^2 \tag{2.20}$$

(ii) Since by construction, $J(u_k)$ is non-oncreasing and larger than $J(u)$, it converges and in particular we have $J(u_{k+1}) - J(u_k) \to 0$ as $k \to +\infty$. Hence from (i), $\|u_k - u_{k+1}\| \to 0$ as $k \to +\infty$.

(ii) Thanks to the orthogonality of successive gradients, we have:

$$\|\nabla J(u_k)\|^2 = \langle \nabla J(u_k), \nabla J(u_k) - \nabla J(u_{k+1}\rangle \leq \|\nabla J(u_k)\| \|\nabla J(u_k) - \nabla J(u_{k+1}\| \quad (2.21)$$

and thus:

$$\|\nabla J(u_k)\| \leq \|\nabla J(u_k) - \nabla J(u_{k+1})\| \quad (2.22)$$

(iii) Since $J(u_k)$ non-increasing, and since J coercive, we have u_k bounded. ∇J being continuous, it is uniformly continuous on any compact set of E (Heine theorem). Hence from (ii),

$$\|u_k - u_{k+1}\| \to 0 \quad (2.23)$$

as $k \to +\infty$. And from (iii), we thus get $\nabla J(u_k) \to 0$ as $k \to +\infty$.

(iv) We have (using ellipticity and $\nabla J(u) = 0$):

$$\alpha \|u_k - u\|^2 \leq \langle \nabla J(u_k) - \nabla J(u), u_k - u\rangle = \langle \nabla J(u_k), u_k - u\rangle \leq \|\nabla J(u_k)\| \|u - u_k\| \quad (2.24)$$

Hence:

$$\|u_k - u\| \leq \frac{1}{\alpha} \|\nabla J(u_k)\| \quad (2.25)$$

Notice in particular that it gives an estimation of the error.
Notice also that during the proof, we have shown that:

$$\langle \nabla J(u_k), \nabla J(u_{k+1})\rangle = 0 \quad (2.26)$$

Remark: In general, one does not try to compute the optimal ρ. There exists a large number of methods in the litterature for a linear search of an optimal ρ. In particular, let us mention Wolfe rule.

Case of a quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \quad (2.27)$$

Notice that $\nabla J(v) = Av - b$.
Since $\langle \nabla J(u_k), \nabla J(u_{k+1})\rangle = 0$, the computation of $\rho(u_k)$ is straightforward:

$$0 = \langle \nabla J(u_k), \nabla J(u_{k+1})\rangle = \langle A(u_k - \rho(u_k)(Au_k - b)) - b, Au_k - b\rangle \quad (2.28)$$

We deduce:

$$\rho(u_k) = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle} \quad (2.29)$$

with

$$w_k = Au_k - b = \nabla J(u_k) \quad (2.30)$$

The algorithm in this case reduces at each iteration to:

- Compute $w_k = Au_k - b$.
- Compute $\rho(u_k) = \frac{\|w_k\|^2}{\langle Aw_k, w_k\rangle}$
- Compute $u_{k+1} = u_k - \rho(u_k) w_k$.

Remark: sufficient condition of convergence. Notice that we only give sufficient condition of convergence. In practice, these conditions will not always be fullfilled, but the algorithm may nevertheless converges. In practice, if the algorithm does not converge, most often the solution explodes (although sometimes it may oscillate).

2.2.4 Variable step gradient method

In the previous algorithm, finding the optimal step ρ can be high time consuming. It is therefore sometimes simpler to use a constant step ρ for all iterations. Although there will be more iterations needed, since the iterations will be faster it may be a good strategy. It can also be ρ_k depending on iteration k, but not "optimal".

The variable step algorithm is therefore:
u_0 in $\mathbb{R}^N$, and:

$$u_{k+1} = u_k - \rho_k \nabla J(u_k) \tag{2.31}$$

The Fixed step algorithm is therefore:
u_0 in $\mathbb{R}^N$, and:

$$u_{k+1} = u_k - \rho \nabla J(u_k) \tag{2.32}$$

Theorem 2.3. *Let us consider an elliptic functional J (with ellipticity constant α). Let us furthermore assume that ∇J is M Lipshitz. If for all k:*

$$0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2} \tag{2.33}$$

then the variable step gradient method converges, and the speed of convergence is geometric. There exists $\beta \in (0,1)$ such that

$$\|u_k - u\| \leq \beta^k \|u_0 - u\| \tag{2.34}$$

Proof: We use the characterization $\nabla J(u) = 0$. Hence we can write:

$$u_{k+1} - u = (u_k - u) - \rho_k (\nabla J(u_k) - \nabla J(u)) \tag{2.35}$$

Thus:

$$\|u_{k+1} - u\|^2 = \|u_k - u\|^2 - 2\rho_k \langle u_k - u, \nabla J(u_k) - \nabla J(u)\rangle + \rho_k^2 \|\nabla J(u_k) - \nabla J(u)\|^2 \tag{2.36}$$

And using the ellipticity of J and the Lipshitz constant of ∇J (and assuming $\rho_k > 0$):

$$\|u_{k+1} - u_k\|^2 \leq (1 - 2\alpha\rho_k + M^2\rho_k^2)\|u_k - u\|^2 \tag{2.37}$$

It is easy to see that if $0 < \rho_k < \frac{2\alpha}{M^2}$ then $0 < \sqrt{1 - 2\alpha\rho_k + M^2\rho_k^2} = \beta < 1$.

■

Case of a quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \tag{2.38}$$

with A symmetric definite positive matrix.

In this case the method is $u_{k=1} = u_k - \rho_k(Au_k - b)$. And the algorithm converges if

$$0 < a \leq \rho_k \leq b < \frac{2\lambda_1}{\lambda_N^2} \tag{2.39}$$

with λ_1 smallest eigenvalue of A, and λ_N largest eigenvalue of A.

But here we have (using the fact that $Au = b$):

$$u_{k+1} - u = (u_k - u) - \rho_k A(u_k - u) = (Id - \rho_k A)(u_k - u) \tag{2.40}$$

Hence:

$$\|u_{k+1} - u\| \leq \|Id - \rho_k A\|_2 \|u_k - u\| \tag{2.41}$$

But since $(Id - \rho_k A)$ is a symmetric matrix, its norm $\|.\|_2$ is $\|Id - \rho_k A\|_2 = \max\{|1 - \rho_k\lambda_1|, |1 - \rho_k\lambda_N|\}$. Hence $\|Id - \rho_k A\|_2 < 1$ if $0 < \rho_k < \frac{2}{\lambda_N}$. Therefore the algorithm converges if:

$$0 < a \leq \rho_k \leq b < \frac{2}{\lambda_N} \tag{2.42}$$

And if $\lambda_1 << \lambda_N$, this is a much sharper result.

Armijo condition: J convex funtional. We have

$$J(u_k + \rho d_k) \geq J(u_k) + \rho\langle \nabla J(u_k), d_k\rangle \tag{2.43}$$

If $\epsilon \in (0, 1)$ is fixed, there exists ρ sufficiently small such that:

$$J(u_k + \rho d_k) \leq J(u_k) + \epsilon\rho\langle \nabla J(u_k), d_k\rangle \tag{2.44}$$

This last inequality is *Armijo* condition. This a sufficient decrease condition.

Typical value for ϵ in practice is 10^{-4}.

Armijo test: Fix $\epsilon \in (0, 1)$ and $\rho_0 > 0$. We set $\rho_i = \rho_0 2^{-i}$. The value of ρ chosen is the largest ρ_i which satisfies the *Armijo* condition.

With such a selection rule, it can be shown that the variable gradient descent converges (under reasonable hypotheses).

Wolfe conditions This is Armijo condition plus a condition on the curvature:

$$\langle \nabla J(u_k + \rho_k d_k), d_k\rangle \geq \tilde{\epsilon}\langle \nabla J(u_k), d_k\rangle \tag{2.45}$$

with $\tilde{\epsilon} \in (\epsilon, 1)$.

2.2.5 Conjugate gradient method

Introduction: J functional defined on $E = \mathbb{R}^N$. Problem: find $u \in E$ such that:

$$J(u) = \inf_{v \in E} J(v) \tag{2.46}$$

To improve the convergence with respect to the optimal gradient descent (best local choice), one needs to use more information about the funtional.

One solution is to use second order derivative $\Longrightarrow$ Newton like method.

But it is possible to improve the direction choice without resorting to second order derivative.

Example: Let us consider

$$J(v_1, v_2) = \frac{1}{2}(\alpha_1 v_1^2 + \alpha_2 v_2^2) \tag{2.47}$$

with $0 < \alpha_1 < \alpha_2$. We have $J(0) = \inf_{v \in \mathbb{R}^2} J(v)$.

Assume that we use the optimal gradient descent to solve this problem. Assume $u_0 = (u_1^0, u_2^0)$ has its two components non zero (otherwise the method converges in 1 iteration).

Indeed, a necessary and sufficient condition for u_{k+1} to be equal to 0 is that 0 belongs to the line $\{u_k - \rho \nabla J(u_k); \rho \in \mathbb{R}\}$., i.e. there exists $\rho \in \mathbb{R}$ such that: $u_1^k = \rho \alpha_1 u_1^k$ and $u_2^k = \rho \alpha_2 u_2^k$. But this is possible only if one of the u_i^k is zero (since $\alpha_1 < \alpha 2$). But it is easy to prove by induction that, if $u_1^0 \neq 0$ and $u_2^0 \neq 0$, then for all k we have: $u_1^k \neq 0$ and $u_2^k \neq 0$.

In the optimal gradient method, we have the relation

$$\langle \nabla J(u_k), \nabla J(u_{k+1}) \rangle = 0 \tag{2.48}$$

The basic idea of the conjugate gradient method for a quadratic functional is that the direction descent d_k are orthogonal with respect to $\langle ., . \rangle_A$ inner product, to take into account the geometry of J.

We recall that a quadratic functional is given by:

$$J(v) = \frac{1}{2}\langle Av, v \rangle - \langle b, v \rangle \tag{2.49}$$

with A symmetric definite positive matrix.

Principle: $u_0 \in \mathbb{R}^N$. We define:

$$G_k = Vect_{0 \leq i \leq p}\{\nabla J(u_k)\} \tag{2.50}$$

u_{k+1} is defined as the minimizer of the restiction of J to $u_k + G_k = \{u_k + v_k, v_k \in G_k\}$, i.e.:

$$\begin{cases} u_{k+1} \in (u_k + G_k) \\ J(u_{k+1}) = \inf_{v \in (u_k + G_k)} J(v) \end{cases} \tag{2.51}$$

Notice that since $u_k + G_k$ is closed and convex, J being coercive and strictly convex, the above minimization problem admits a solution and only one.

Notice also that comparing with the optimal gradient descent, one optimizes on a larger set, $u_k + G_k$, instead of $\{u_k + \rho \nabla J(u_k)\}$: the result is therefore better.

Proposition 2.1. *For all $p \neq q$, we have:*

$$\langle \nabla J(u_p), \nabla J(u_q) \rangle = 0 \tag{2.52}$$

Notice that in the optimal gradient descent, only 2 consecutive gradients are orthogonal.

Proof: This is an immediate consequence of the fact that $J(u_{k+1}) = \inf_{v \in (u_k + G_k)} J(v)$, i.e. $J(u_{k+1}) = \inf_{v \in G_k} J(u_k + v)$: this implies that $\langle \nabla J(u_{k+1}), w \rangle = 0$ for all w in G_k. In particular, $\langle \nabla J(u_{k+1}), \nabla J(u_i) \rangle = 0$ if $0 \leq i \leq k$.

■

In particular, this implies that the method converges in at most N iterations.

Proposition 2.2. *For all $p \neq q$, we have:*

$$\langle d_p, d_q \rangle_A = \langle Ad_p, d_q \rangle = 0 \tag{2.53}$$

where $d_k = u_{k+1} - u_k$ is the descent direction.

Hence the name of the method.

Proof: Assume the first $p+1$ vectors constructed.

$$J(v) = \frac{1}{2}\langle Av, v \rangle - \langle v, b \rangle \tag{2.54}$$

with A symmetric. Hence $\nabla J(v) = Av - b$.

$$\nabla J(v+w) = A(v+w) - b = \nabla J(v) + Aw \tag{2.55}$$

In particular:

$$\nabla J(u_{k+1}) = \nabla J(u_k + d_k) = \nabla J(u_k) + Ad_k \tag{2.56}$$

Hence $d_k \neq 0$ for all k.

From the orthogonality of the gradients, we get:

$$0 = \|\nabla J(u_k)\|^2 + \langle Ad_k, \nabla J(u_k) \rangle \tag{2.57}$$

Moreover:

$$0 = \langle \nabla J(u_{k+1}), \nabla J(u_i) \rangle = \langle \nabla J(u_k, \nabla J(u_i) \rangle + \langle Ad_k, \nabla J(u_i) \rangle \tag{2.58}$$

hence:

$$0 = \langle Ad_k, \nabla J(u_i) \rangle \tag{2.59}$$

Since all d_m are linear combinations of the $\nabla J(u_i)$, this implies that: $0 = \langle Ad_k, d_m \rangle$.

■

Algorithm (conjugate gradient method):

$$J(v) = \frac{1}{2}\langle Av, v \rangle - \langle b, v \rangle \tag{2.60}$$

with A symmetric definite positive matrix.

$u_0 \in \mathbb{R}^N$. We set $d_0 = \nabla J(u_0)$. If $\nabla J(u_0) = 0$, then the algorithm is finished. Otherwise, we set:

$$r_0 = \frac{\langle \nabla J(u_0), d_0 \rangle}{\langle Ad_0, d_0 \rangle} \tag{2.61}$$

and then $u_1 = u_0 - r_0 d_0$.

To build u_{k+1} from u_k: if $\nabla J(u_k) = 0$, then the algorithm is finished. Otherwise, we set:

$$d_k = \nabla J(u_k) + \frac{\|\nabla J(u_k)\|^2}{\|\nabla J(u_{k-1})\|^2} d_{k-1} \tag{2.62}$$

and

$$r_k = \frac{\langle \nabla J(u_k), d_k \rangle}{\langle A d_k, d_k \rangle} \tag{2.63}$$

and then $u_{k+1} = u_k - r_k d_k$.

Theorem 2.4. *The conjugate gradient method applied to a quadratic elliptic functional converges in at most N iterations.*

Notice that this method is particularly usefull when A is sparse (since A plays a role in the computation only in the product Ad_k).

Practical method: In practice, due to numerical imprecision, the convergence is not reached in a finite number of iterations. One needs to use a stopping criterion.

Matrix inversion Find u such that

$$J(u) = \inf_v J(v) \tag{2.64}$$

with

$$J(v) = \frac{1}{2}\langle Av, v \rangle - \langle v, b \rangle \tag{2.65}$$

with A symmetric. We have $\nabla J(v) = Av - b$, and the optimality condition $\nabla J(u) = 0$, i.e. $Au = b$, i.e. $u = A^{-1}b$.

Hence the conjugate gradient method can be used to inverse positive definite matrices. It is in particular used if A has some sparse structure (then the method can become faster then the Cholesky factorization).

Polak-Ribière conjugate gradient method: The previous method can be extended to general convex functional (although with no guarantie of finite number iterations to converge). First notice that the orthogonality of the gradient in the conjugate gradient method enables to write:

$$d_k = \nabla J(u_k) + \frac{\|\nabla J(u_k)\|^2}{\|\nabla J(u_{k-1})\|^2} d_{k-1} = \nabla J(u_k) + \frac{\langle \nabla J(u_k), \nabla J(u_k) - \nabla J(u_{k-1}) \rangle}{\|\nabla J(u_{k-1})\|^2} d_{k-1} \tag{2.66}$$

Using the first part of the equation leads to Fletcher-Reeves conjugate gradient method. Using the last part leads to Ploak Ribière conjugate gradiente method (which in practice is more efficient):

$u_0 \in \mathbb{R}^N$. $d_0 = \nabla J(u_0)$.

To build u_{k+1} from u_k: if $\nabla J(u_k) = 0$, then the algorithm is finished. Otherwise, we set:

$$d_k = \nabla J(u_k) + \frac{\langle \nabla J(u_k), \nabla J(u_k) - \nabla J(u_{k-1}) \rangle}{\|\nabla J(u_{k-1})\|^2} d_{k-1} \tag{2.67}$$

and

$$u_{k+1} = u_k - r_k d_k \tag{2.68}$$

with $J(u_{k+1}) = \inf_{r \in \mathbb{R}} J(u_k - r d_k)$.

2.2.6 Quasi-Newton method

Basic idea: use of the condition $\nabla J(u) = 0$. To find a minimum of J, one looks for a zero of ∇J. Notice that then it is needed to check that indeed the zero of ∇J is a minimizer of J.

Finding a zero of F Let us consider $F : \Omega \to \mathbb{R}$ where $\Omega \subset \mathbb{R}$.

The principle of Newton method is to linearise the equation $F(x) = 0$ around x_k the current iteration:

$$0 = F(x) = F(u_k) + F'(u_k).(x - u_k) \tag{2.69}$$

$F'(u_k)$ is the differential of F.

By analogy with the one dimensional case, we set:

$$u_{k+1} = u_k - (\nabla F(u_k))^{-1} F(u_k) \tag{2.70}$$

Given $u_0 \in \Omega$, define the sequence:

$$u_{k+1} = u_k - (\nabla F(u_k))^{-1} F(u_k) \tag{2.71}$$

To be well defined, u_k is to remain in Ω for all k, F is to be differentiable in Ω, and ∇F needs to be invertible.

The main difficulty in Newton method relies in good guess of u_0.

Notice that at each iteration, the main computation time is required by $\nabla F(u_k))^{-1}$. A natural idea is to keep the same matrix during a bunch of iterations, or even to replace it by a matrix easy to inverse $\Longrightarrow$ quasi-newton method.

We get an algorithm of the type:

Given $u_0 \in \Omega$, define the sequence:

$$u_{k+1} = u_k - (A_k(u_{k'}))^{-1} F(u_k) \tag{2.72}$$

with $0 \leq k' \leq k$, and $A_k(u_{k'})$ invertible.

For such an algorithm to converge, we need the following intuitive hyptheses: $F(u_0)$ sufficiently small, $\nabla F(u)$ does not vary too much around u_0, $A_k(u)$ and $A_k^{-1}(u)$ do not vary too much with respect to k and for u close to u_0.

Theorem 2.5. *$F : \Omega \subset X \to Y$ where X is a Banach space. F differentiable. Let us assume that there exist r, M, β such that: $r > 0$ and $B = \{x \in X, \|x - x_0\| \leq r\} \subset \Omega$. $A_k \in Isom(X, Y)$ (i.e. set of continuous linear applications bijectives from X to Y with continuous inverse).*

1.

$$\sup_{k \geq 0} \sup_{x \in B} \|A_k^{-1}(x)\| \leq M \tag{2.73}$$

2.

$$\sup_{k \geq 0} \sup_{x, x' \in B} \|\nabla F(x) - A_k(x')\| \leq \frac{\beta}{M} \tag{2.74}$$

and $\beta < 1$.

3.

$$\|F(x_0)\| \leq \frac{r}{M}(1 - \beta) \tag{2.75}$$

Then the sequence x_k defined by:

$$x_{k+1} = x_k - A_k(x_{k'}^{-1})F(x_k) \tag{2.76}$$

with $k \geq k' \geq 0$ is entirely contained inside B, and converges to a zero of F, which is the only zero of F inside B. Moreover, the convergence is geometric:

$$\|x_k - a\| \leq \frac{\|x_1 - x_0\|}{1-\beta}\beta^k \tag{2.77}$$

It relies on the fixed point theorem in complete spaces (for contractant applications).

Finding a minimum of J :

Newton method:

Given $u_0 \in \Omega$, define the sequence:

$$u_{k+1} = u_k - (\nabla^2 J(u_k))^{-1}\nabla J(u_k) \tag{2.78}$$

Quasi-Newton method:

Given $u_0 \in \Omega$, define the sequence:

$$u_{k+1} = u_k - A_k^{-1}(u_k')\nabla J(u_k) \tag{2.79}$$

- If $A_k(u_k') = \rho^{-1}Id$, then this the fixed step gradient method.
- If $A_k(u_k') = \rho_k^{-1}Id$, then this the variable step gradient method.
- If $A_k(u_k') = (\rho(u_k))^{-1}Id$, then this is the optimal step gradient method, with $\rho(u_k)$ determined by:

$$J(u_k - \rho(u_k)\nabla J(u_k)) = \inf_{\rho\in\mathbb{R}} J(u_k - \rho\nabla J(u_k)) \tag{2.80}$$

3. Constrained optimization

3.1 Problems with general constraints

For this subsection, we refer the reader to [5].

3.1.1 Projection on a convex closed set

Necessary condition:

Theorem 3.1. *Let $J : U \to \mathbb{R}$ a differentiable function on a convex set U. If $u \in U$ is a minimizer of J, then $\langle \nabla J(u), (v-u)\rangle \geq 0$ for all $v \in U$.*

Proof: Remark that if u and $v = u + w$ in U convex, then $u + tw$ is in U for all $t \in [0, 1]$. We then apply Taylor formula:

$$0 \leq J(u + tw) - J(u) = t\langle \nabla J(u), w\rangle + o(t\|w\|) \tag{3.1}$$

and $w = v - u$.

■

Notice that if U is a subspace of E, then the necessary condition becomes: $\langle \nabla J(u), v\rangle = 0$ for ll $v \in U$.

Notice also that if $U = E$, then the condition is the classical Euler equation $\nabla J(u) = 0$.

Projection theorem :

Theorem 3.2. *Let U a closed non empty convex set in $X = \mathbb{R}^N$. Let $w \in X$. Then there exists a unique element denoted by Pw such that:*

$$Pw \in U \text{ and } \|w - Pw\| = \inf_{v \in U} \|w - v\| \tag{3.2}$$

Theorem 3.3. *Let U a closed non empty convex set in $X = \mathbb{R}^N$. $u = Pw$ iff*

$$\langle u - w, v - u\rangle \geq 0 \tag{3.3}$$

for all $v \in U$.

Proposition 3.1. *The projection $P : X \to U$ is 1 contractant, i.e.:*

$$\|Pw_1 - Pw_2\| \leq \|w_1 - w_2\| \tag{3.4}$$

for all w_1, w_2 in X.

Proposition 3.2. *The projection $P : X \to U$ is linear iff U is a subspace of X. In this case, the characterization inequality becomes:*

$$\langle Pw - w, v\rangle = 0 \tag{3.5}$$

for all v in U.

Remaks:

- The characterization of the projection has an immediate geometric interpretation on the angle between the vectors $(Pw - w)$ and $(v - Pw)$.
- Notice that $Pw = w$ iff $w \in U$.
- The characterization inequality of the projection is nothing but the Euler equation associated to the following problem: w fixed in X, consider

$$J(v) = \frac{1}{2}\|w - v\|^2 \tag{3.6}$$

  J is differentiable and strictly convex, and has a minimum in U in $v = Pw$.
- In the case when U is a subspace, then $Pw - w \perp u$ for all u in U.

If U is of the type:

$$U = \Pi_{i=1}^N [a_i, b_i] \tag{3.7}$$

then $(Pw)_i = \min(\max(w_i, b_i), b_i)$.

3.1.2 Relaxation method for a rectangle

U being contained in X, find u such that:

$$J(u) = \inf_{v \in U} J(u) \tag{3.8}$$

Here we only consider the case when

$$U = \Pi_{i=1}^N [a_i, b_i] \tag{3.9}$$

with possibly $a_i = -\infty$ and $b_i = +\infty$.

In this case the relaxation method is:

$$\begin{cases} J([u_1^{k+1}], u_2^k, \ldots, u_N^k) = \inf_{\xi \in [a_1, b_1]} J(\xi, u_2^k, \ldots, u_N^k) \\ \qquad \ldots \\ J(u_1^{k+1}, \ldots, u_{N-1}^{k+1}, [u_N^{k+1}]) = \inf_{\xi \in [a_N, b_N} J(u_1^{k+1}, \ldots, u_{N-1}^{k+1}, \xi) \end{cases} \tag{3.10}$$

Theorem 3.4. *If J is elliptic, and if $U = \Pi_{i=1}^N [a_i, b_i]$, then the relaxation method converges.*

Remark: It is not possible to extend the relaxation algorithm to more general sets. Indeed, consider the case when $J(v) = v_1^2 + v_2^2$ and $U = \{v = (v_1, v_2) \in \mathbb{R}^2; v_1 + v_2 \geq 2\}$. Assume $u^0 = (u_1^0, u_2^0)$ with $u_1^0 \neq 1$ and $u_2^0 \neq 1$.

3.1.3 Projected gradient method

U convex closed non empty set. J convex.

Problem: find $u \in U$ such that $J(u) = \inf_{v \in U} J(v)$.
$\Leftrightarrow u \in U$ and $\langle \nabla J(u), v - u \rangle \geq 0$ for all $v \in U$.
$\Leftrightarrow u \in U$ and $\langle u - (u - \rho \nabla J(u)), v - u \rangle \geq 0$ for all $v \in U$ and $\rho > 0$.

$\Leftrightarrow u = P(u - \rho\nabla J(u))$ for all $\rho > 0$.

Hence the solution appears as a fixed point of the application $g : v \to g(v) = P(v - \rho\nabla J(v))$

It thus natural to consider the sequence:

$u_0 \in E$, and

$$u_{k+1} = g(u_k) = P(u_k - \rho\nabla J(u_k)) \tag{3.11}$$

Notice that in the case when $U = E = \mathbb{R}^N$, then $P = Id$ and thus $u_{k+1} = u_k - \rho\nabla J(u_k)$, i.e. this is the fixed step gradient method for unconstrained problems.

The method we have just described is called *fixed step projected gradient method*.

To show the convergence of the fixed step projected gradient method, it suffices to show that $g : E \to E$ is a contraction, i.e. there exists $\gamma \in (0, 1)$ such that $\|g(u) - g(v)\| \leq \gamma\|u - v\|$ for all u, v in E.

Thanks to the fixed point theorem in Banach spaces, it shows that g has a fixed point, and thus the convergence of the algorithm.

More generally, the convergence (under reasonable hypothese) of the variable step projected gradient method can be shown.

$u_0 \in E$, and

$$u_{k+1} = g(u_k) = P(u_k - \rho_k\nabla J(u_k)) \tag{3.12}$$

with $\rho_k > 0$ for all $k \geq 0$.

Theorem 3.5. *Let us consider an elliptic functional J (with ellipticity constant α). Let us furthermore assume that ∇J is M Lipshitz. If for all k:*

$$0 < a \leq \rho_k \leq b < \frac{2\alpha}{M^2} \tag{3.13}$$

then the variable step projected gradient method converges, and the speed of convergence is geometric. There exists $\beta \in (0, 1)$ such that

$$\|u_k - u\| \leq \beta^k\|u_0 - u\| \tag{3.14}$$

Proof:

$$g_k(v) = P(v - \rho_k\nabla J(v)) \tag{3.15}$$

We have:

$$\begin{aligned} \|g_k(v_1) - g_k(v_2)\|^2 &= \|P(v_1 - \rho_k\nabla J(v_1)) - P(v_2 - \rho_k\nabla J(v_2))\|^2 \\ &\leq \|(v_1 - v_2) - \rho_k(\nabla J(v_1)) - \nabla J(v_2))\|^2 \\ &\leq (1 - 2\alpha\rho_k + M^2\rho_k^2)\|v_1 - v_2\|^2 \end{aligned}$$

It is easy to see that if $0 < \rho_k < \frac{2\alpha}{M^2}$ then $0 < \sqrt{1 - 2\alpha\rho_k + M^2\rho_k^2} = \beta < 1$.

Since the solution u is a fixed point of each application g_k, we can write:

$$\|u_{k+1} - u\| = \|g_k(u_k) - g_k(u)\| \leq \beta\|u - u_k\| \tag{3.16}$$

■

Case of a quadratic functional

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \tag{3.17}$$

with A symmetric definite positive matrix.

In this case the method is $u_{k+1} = P(u_k - \rho_k(Au_k - b))$. And the algorithm converges if

$$0 < a \leq \rho_k \leq b < \frac{2\lambda_1}{\lambda_N^2} \tag{3.18}$$

with λ_1 smallest eigenvalue of A, and λ_N largest eigenvalue of A.

But here we have:

$$u_{k+1} - u = P(u_k - \rho_k Au_k) - P(u) \tag{3.19}$$

Using the fact that $Au = b$,

$$\|u_{k+1} - u\| \leq \|u_k - \rho_k A(u_k - u)) - u\| = (Id - \rho_k A)(u_k - u) \tag{3.20}$$

Hence:

$$\|u_{k+1} - u\| \leq \|Id - \rho_k A\|_2 \|u_k - u\| \tag{3.21}$$

But since $(Id - \rho_k A)$ is a symmetric matrix, its norm $\|.\|_2$ is $\|Id - \rho_k A\|_2 = \max\{|1 - \rho_k\lambda_1|, |1 - \rho_k\lambda_N|\}$. Hence $\|Id - \rho_k A\|_2 < 1$ if $0 < \rho_k < \frac{2}{\lambda_N}$. Therefore the algorithm converges if:

$$0 < a \leq \rho_k \leq b < \frac{2}{\lambda_N} \tag{3.22}$$

And if $\lambda_1 << \lambda_N$, this is a much sharper result.

Practical remark: From a practical point of view, the projected gradient method can be used only if the projection operator is explicitely known (which is not the case in general). A notable excpetion is the case when:

$$U = \Pi_{i=1}^N [a_i, b_i] \tag{3.23}$$

We have already written the projection operator in this case.

Consider for instance the following problem: Minimize $J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$ on $U = \mathbb{R}_+^N$. The projected gradient algorithm in this case is:

$u_0 \in \mathbb{R}^N$, and, for $1 \leq i \leq N$:

$$u_i^{k+1} = \max\left(u_i^k - \rho_k(Au_k - b)_i, 0\right) \tag{3.24}$$

But except in such particular cases, constrained minimization problems are to be handled with different methods, such as penalization methods.

3.1.4 Penalization method

Theorem 3.6. *Let $J : \mathbb{R}^N \to \mathbb{R}$ a coercive strictly convex (and thus continuous) function, U a non empty convex set in $\mathbb{R}^N$, and $\psi : \mathbb{R}^N \to \mathbb{R}$ a convex (and thus continuous) function such that*

$$\psi(v) \geq 0 \text{ for all } v \in \mathbb{R}^N \text{ and } \psi(v) = 0 \Leftrightarrow v \in U \tag{3.25}$$

Then for all $\epsilon > 0$, there exists a unique u_ϵ such that:

$$u_\epsilon \in \mathbb{R}^N \text{ and } J_\epsilon(u_\epsilon) = \inf_{v\in\mathbb{R}^N} J_\epsilon(v) \text{ where } J_\epsilon(v) = J(v) + \frac{1}{\epsilon}\psi(v) \tag{3.26}$$

and $\lim_{\epsilon\to 0} u_\epsilon = u$, where u is the unique solution of the problem:

$$\text{Find } u \in U \text{ and } J(u) = \inf_{v\in U} J(v) \tag{3.27}$$

Proof: It is clear that both (3.26) and (3.27) have a unique solution. We have:

$$J(u_\epsilon) \le J(u_\epsilon) + \frac{1}{\epsilon}\psi(u_\epsilon) = J_\epsilon(u_\epsilon) \le J_\epsilon(u) = J(u) \tag{3.28}$$

Since J is coercive, u_ϵ is bounded. Up to an extraction, there exists $\tilde{u}$ such that $u_\epsilon \to \tilde{u}$. Since J continuous, and since $J(u_\epsilon) \le J(u)$, we get:

$$J(\tilde{u}) \le J(u) \tag{3.29}$$

Remark that:

$$0 \le \psi(u_\epsilon) \le \epsilon(J(u) - J(u_\epsilon)) \tag{3.30}$$

But since u_ϵ bounded, $J(u) - J(u_\epsilon)$ is also bounded (independently from ϵ). Hence: $\psi(\tilde{u}) = \lim \psi(u_\epsilon) = 0$. We thus see that $\tilde{u}$ is in U. Moreover, since $J(\tilde{u}) \le J(u)$ we get that $\tilde{u} = u$. By uniqueness of the solution, we deduce that the result is true for any cluster point of u_ϵ.

■

Example: $J : \mathbb{R}^N \to \mathbb{R}$ strictly convex, and $\phi_i : \mathbb{R}^N \to \mathbb{R}$ convex. Consider the problem: find $u \in U$ such that:

$$J(u) = \inf_{v\in U} J(v) \tag{3.31}$$

with

$$U = \{v \in \mathbb{R}^N;\ \phi_i(v) \le 0,\ 1 \le i \le m\} \tag{3.32}$$

We can choose for instance:

$$\psi(v) = \sum_{i=1}^{m} \max(\phi_i(v), 0) \tag{3.33}$$

or (differentiable constraint):

$$\psi(v) = \sum_{i=1}^{m} (\max(\phi_i(v), 0))^2 \tag{3.34}$$

Remarks: The main point of a penalization method is to replaced a constrained minimization problem by an unconstrained one.

In practice, the problem is to find good penalization functions ψ (for instance differentiable).

3.2 Problems with equality constraints

We refer the reader to [5].

Definition 3.1. X and Y two normed vectorial spaces.
Isom(X, Y) is the set of linear continuous bijective applications with continuous inverse.

Theorem 3.7. Implicit functions theorem
Let $\phi : \Omega \subset X_1 \times X_2 \to Y$ an application continuously differentiable on Ω, and (a_1, a_2) in Ω, b in Y, points such that:

$$\phi(a_1, b_1) = b,\ \partial_2\phi(a_1, a_2) \in Isom(X_2, Y) \tag{3.35}$$

X_2 is assumed to be a Banach space.
Then there exists an open set $O_1 \subset X_1$, an open set $O_2 \subset X_2$, and a continuous application called implicit function

$$f : 0_1 \subset X_1 \to X_2 \tag{3.36}$$

such that $(a_1, a_2) \in 0_1 \times 0_2 \subset \Omega$ and

$$\{(x_1, x_2) \in 0_1 \times 0_2,\ \phi(x_1, x_2) = b\} = \{(x_1, x_2) \in 0_1 \times X_2,\ x_2 = f(x_1)\} \tag{3.37}$$

Moreover, f is differentiable in a_1 and

$$f'(a_1) = -\left(\partial_2\phi(a_1, a_2)\right)^{-1} \partial_1\phi(a_1, a_2) \tag{3.38}$$

The following theorem is a consequence of the classical implicit funtion theorem (in Banach spaces). It gives a necessary condition of linked extrema.

Theorem 3.8. *Let $\Omega \subset \mathbb{R}^N$ open set. Let $\phi_i : \Omega \to \mathbb{R}$, $1 \le i \le m$ C^1 functions on Ω. Let u in*

$$U = \{v \in \Omega,\ \phi_i(v) = 0,\ 1 \le i \le m\} \tag{3.39}$$

such that the differential $\nabla\phi_i(u)$ in $\mathcal{L}(\mathbb{R}^N, \mathbb{R})$, $1 \le i \le m$ are linearly independants.
Let $J : \Omega \to \mathbb{R}$ a function differentiable in u. If J has some local extremum with respect to U, then there exist m numbers $\lambda_i(u)$, $1 \le i \le m$, uniquely defined, such that:

$$\nabla J(u) + \sum_{i=1}^{m} \lambda_i(u)\nabla\phi_i(u) = 0 \tag{3.40}$$

Basic idea: $J(u_1, u_2)$ with u_2 in the constraints set $\{\psi(v) = 0\}$. Implicit function theorem: $\Longrightarrow u_2 = f(u_1)$ and we apply the classical conditon (derivative 0) to $J(u, f(u))$.

Remarks:

- $\nabla\phi_i(u)$ in $\mathcal{L}(\mathbb{R}^N, \mathbb{R})$, $1 \le i \le m$ are linearly independant means that the matrix $\partial_j\phi_i(u)$, $1 \le i \le m$, $1 \le j \le N$, is of rank m.
- The numbers $\lambda_i(u)$, $1 \le i \le m$, are called *Lagrange multipliers* associated to the linked extremum u.
- Notice that the result of the theorem is a necessary condition, but not a sufficient one. It is therefore needed to carry out a local analysis to check whether it is indeed an extremum.

Quadratic functional

$$J(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle \tag{3.41}$$

with A symmetric.

Assume we are interested in the problem: find u in U such that

$$J(u) = \inf_{v\in U} J(v) \tag{3.42}$$

with

$$U = \{v \in \mathbb{R}^N,\ Cv = d\} \tag{3.43}$$

where C is a matrix of size $m \times N$, and d a vector of size m. We assume that $m < N$.

Let us set $\phi(v) = Cv - d$. We have $\nabla\phi(v) = C$.

Hence, if C is of rank m, we can apply the above theorem to get a necessary condition for J to have an extremum in $u \in U$ with respect to U, is that the following linear system admits a solution (u, λ) in $\mathbb{R}^{N+m}$:

$$\left\{ \begin{array}{lcl} Au + C^T\lambda & = & b \\ Cu & = & d \end{array} \right. \tag{3.44}$$

i.e.:

$$\left(\begin{array}{cc} A & C^T \\ C & 0 \end{array} \right) \left(\begin{array}{c} u \\ \lambda \end{array} \right) = \left(\begin{array}{c} b \\ d \end{array} \right) \tag{3.45}$$

which can be solved with any classical method for linear systems.

Lagrange-Newton algorithm Consider the minimization problem

$$\inf J(u)$$

under the constraint $\phi_i(u) = 0$. Formally, the first order conditions are:

$$\left\{ \begin{array}{lcl} \nabla_u\mathcal{L}(u, \lambda) & = & 0 \\ \phi(u) & = & 0 \end{array} \right. \tag{3.46}$$

where $\mathcal{L}$ is the Lagrangian of the problem:

$$\mathcal{L}(u, \lambda) = J(u) + \sum_{i=1}^{m} \lambda_i\phi_i(u) = J(u) + \langle\lambda, \phi(u)\rangle \tag{3.47}$$

with $\phi : \Omega \subset \mathbb{R}^N \to \mathbb{R}^m$, $\phi(u) = (\phi_1(u), \ldots, \phi_m(u))$.

The Lagrange-Newton algorithm is Newton method applied to this non-linear system: choose an initial guess $(u_0, \lambda_0) \in \mathbb{R}^N \times \mathbb{R}^m$. If (u_k, λ_k) known, then:

$$\left(\begin{array}{cc} \nabla^2_{uu}\mathcal{L}(u_k, \lambda_k)) & \nabla\phi(u_k)^T \\ \nabla\phi(u_k) & 0 \end{array} \right) \left(\begin{array}{c} d_k \\ y_k \end{array} \right) = - \left(\begin{array}{c} \nabla\mathcal{L}(u_k, \lambda_k)^T \\ \phi(x_k) \end{array} \right) \tag{3.48}$$

Then: $u_{k+1} = u_k + d_k$, $\lambda_{k+1} = \lambda_k + y_k$.

3.3 Problems with inequality constraints

We refer the reader to [5, 7].

3.3.1 Characterization of a minimum on a general set

Theorem 3.9. Farkas-Minkowski lemma*: Let a_i, $i \in I$, where I is a finite set of indices, and b elements in $\mathbb{R}^N$. Then the following inclusion*

$$\{w \in \mathbb{R}^N; \langle a_i, w\rangle \geq 0, i \in I\} \subset \{w \in \mathbb{R}^N, \langle b, w\rangle \geq 0\} \tag{3.49}$$

is satisfied iff:

$$\text{There exists } \lambda_i \geq 0,\ i \in I, \text{ such that } b = \sum_{i\in I} \lambda_i a_i \tag{3.50}$$

Definition 3.2. U a non empty set in $E = \mathbb{R}^N$. For u in U, we define $C(u)$ the cone of admissible directions. It is the union of $\{0\}$ and the set of $w \in \mathbb{R}^N$ such that there exists at least one sequence of points u_k such that:

$$\begin{cases} u_k \in U,\ u_k \neq u \text{ for all } k,\ \lim_k u_k = u \\ \lim_k \frac{u_k - u}{\|u_k - u\|} = \frac{w}{\|w\|},\ w \neq 0 \end{cases} \tag{3.51}$$

This last condition can be equivalently written:

$$u_k = u + \|u - u_k\| \frac{w}{\|w\|} + \|u_k - u\|\delta_k \tag{3.52}$$

with $\lim \delta_k = 0$ and $w \neq 0$.

Notice that $C(u)$ is a cone (with sommet 0), not necessarily convex.

Proposition 3.3. *Let U a non empty set in $E = \mathbb{R}^N$.*

- *For all $u \in U$, $C(u)$ is a closed set.*
- *If U is convex, then $U \subset \{u + C(u)\}$.*

Theorem 3.10. *Let $J : \Omega \subset E \to \mathbb{R}$ a function defined on an open set Ω which contains U. If J admits in $u \in U$ a local minimum with respect to U, and if J is differentiable in u, then*

$$\langle \nabla J(u), v - u\rangle \geq 0 \tag{3.53}$$

for all v in $\{u + C(u)\}$, i.e for all w in $C(u)$ we have:

$$\langle \nabla J(u), w\rangle \geq 0 \tag{3.54}$$

We remind the reader that we proved before that if U is convex, then a necessary and sufficient condition for J to have a local minimum in u with respect to U is:

$$\langle \nabla J(u), v - u\rangle \geq 0 \tag{3.55}$$

for all v in U.

Proof of the Theorem 3.10: Let $w = (v - u)$ a non zero vector in $C(u)$, and let u_k a sequence of points in $U \backslash \{u\}$ such that: $\lim u_k = u$,

$$u_k - u = \|u - u_k\| \frac{w}{\|w\|} + \|u_k - u\| \delta_k \tag{3.56}$$

with $\lim \delta_k = 0$, and $J(u) \leq J(u_k)$.

Using the derivability of J in u, we get:

$$0 \leq J(u_k) - J(u) = \langle \nabla J(u), u_k - u \rangle + \|u_k - u\| \epsilon_k \tag{3.57}$$

with $\lim \epsilon_k = 0$. Hence:

$$0 \leq \frac{\|u_k - u\|}{\|w\|} \langle \nabla J(u), w + \|w\| \delta_k \rangle + \epsilon_k \|u_k - u\| \tag{3.58}$$

which implies:

$$0 \leq \langle \nabla J(u), w + \|w\| \delta_k \rangle + \epsilon_k \|w\| \tag{3.59}$$

Hence, since $\lim \delta_k = 0$, we get $\langle \nabla J(u), w \rangle \geq 0$

■

3.3.2 Kuhn and Tucker relations

We now consider a particular case of set U:

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.60}$$

with Ω open set of $E = \mathbb{R}^N$.

Definition 3.3. *Active indices:* if u is in U, then:

$$I(u) = \{1 \leq i \leq m;\ \phi_i(u) = 0\} \tag{3.61}$$

Definition 3.4.

$$C^*(u) = \{w \in E, \langle \nabla \phi_i(u), w \rangle \leq 0, i \in I(u)\} \tag{3.62}$$

Notice that if $I(u)$ is empty, then $C^*(u) = E$. Notice also that $C^*(u)$ is a convex set.

Definition 3.5. *Qualified constraints:* The constraints in $u \in U$ are said to be qualified if one of the two following alternative holds:

- ϕ_i is an affine function ($\phi_i(x) = ax + b$) for all $i \in I(u)$.
- there exists w in E such that for all $i \in I(u)$:

$$\langle \nabla \phi_i(u), w \rangle \leq 0 \tag{3.63}$$

with strict inequality if ϕ_i is not affine.

Proposition 3.4. *Let u in*

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.64}$$

on which the fuctions $\phi_i : \Omega \subset E \to \mathbb{R}$, $i \in I(u)$ are differentiable. Then we have the inclusion:

$$C(u) \subset C^*(u) \tag{3.65}$$

Proof: Each function ϕ_i has a local maximum in u with respect to U by definition of the set $I(u)$. Using Theorem 3.10, we get:

$$\langle \nabla \phi_i(u), w \rangle \leq 0 \text{ for all } w \text{ in } C(u) \text{ and } i \text{ in } I(u). \tag{3.66}$$

which proves the proposition.

■

Theorem 3.11. *Let u in*

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.67}$$

on which the functions $\phi_i : \Omega \subset E \to \mathbb{R}$, $i \in I(u)$ are differentiable.

If the constraints are qualified in u, and if the functions ϕ_i, $i \neq I(u)$, are continuous in u, we have the equality:

$$C(u) = C^*(u) \tag{3.68}$$

Example: affine constraints in $\mathbb{R}^N$.

$$U = \{v \in \mathbb{R}^N; \sum_{j=1}^{N} c_{i,j} v_j \leq d_i, 1 \leq i \leq m\} = \{v \in \mathbb{R}^N;\ Cv \leq d\} \tag{3.69}$$

with $C = (c_{i,j})$ is an $m \times N$ matrix with real coefficients and $d \in \mathbb{R}^m$. Then the constraints are qualified in $u \in U$, and from the above result we have:

$$C(u) = C^*(u) = \{w \in \mathbb{R}^N, \sum_{j=1}^{N} c_{i,j} w_j \leq 0, i \in I(u)\} \tag{3.70}$$

We are now in position to state one of the most important results in optimization.
ă

Theorem 3.12. Kuhn and Tucker
Let $\phi_i : \Omega \subset E \to \mathbb{R}$ defined on a open set Ω, and

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.71}$$

Let u in U, and $I(u) = \{1 \leq i \leq m; \phi_i(u) = 0\}$.

We assume that ϕ_i is differentiable in u if $i \in I(u)$, and continuous in u if $i \neq I(u)$.

Let $J : \Omega \subset E \to \mathbb{R}$ differentiable in u.

If J admits a local minimum in u with respect to U, and if the constraints are qualified in u, then there exists $\lambda_i(u), i \in I(u)$ such that:

$$\nabla J(u) + \sum_{i \in I(u)} \lambda_i(u) \nabla \phi_i(u) = 0 \tag{3.72}$$

and $\lambda_i(u) \geq 0$ for all $i \in I(u)$.

Notice that it is only a necessary condition.

Proof: From Theorem 3.10, we have $\langle \nabla J(u), w \rangle \geq 0$ for all $w \in C(u)$. But here

$$C(u) = C^*(u) = \{w \in E, \langle \nabla \phi_i(u), w \rangle \leq 0, i \in I(u)\} \tag{3.73}$$

Hence:

$$\{w \in E, \langle -\nabla \phi_i(u), w \rangle \geq 0, i \in I(u)\} \subset \{w \in E, \langle \nabla J(u), w \rangle \geq 0\} \tag{3.74}$$

And we can conclude with the Farkas-Minkowski lemma.

■

Remarks:

- The relations, $\nabla J(u) + \sum_{i \in I(u)} \lambda_i \nabla \phi_i(u) = 0$ and $\lambda_i(u) \geq 0$ for all $i \in I(u)$, are called the Kuhn and Tucker relations. They can be written in a closer form to the Lagrange multipliers (equality constraints):

$$\begin{cases} \nabla J(u) + \sum_{i=1}^{m} \lambda_i(u) \nabla \phi_i(u) = 0 \\ \lambda_i(u) \geq 0, \ 1 \leq i \leq m, \ \sum_{i=1}^{m} \lambda_i \phi_i(u) = 0 \end{cases} \tag{3.75}$$

  $(\lambda_i(u))_i$ are often called general Lagrange multipliers.

- In practice, these relations remain difficult to handle. It leads to a very large system of equations and inequations to solve.

- A sufficient condition for the $(\lambda_i(u))_i$ to be unique is that the differential $\nabla \phi_i$ are linearly independent.

- If $I(u)$ is empty, then we find the classical relation $\nabla J(u) = 0$.

3.3.3 Convex case

The above definition of qualified constraints is not easy to handle, since it depends on u. But it can be simplified in the case of convex constraints.

Definition 3.6. The constraints $\phi_i : \Omega \subset E \to \mathbb{R}$, $1 \leq i \leq m$ are qualified if one of the two following alternative holds:

- all the functions ϕ_i are affines, and the set (convex if Ω is convex)

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.76}$$

  is not empty.

- there exists w in Ω such that for all i:

$$\phi_i(w) \leq 0 \tag{3.77}$$

  with strict inequality if ϕ_i is not affine.

We get the following necessary and sufficient condition: ă

Theorem 3.13. Kuhn and Tucker (convex case)
Let $J : \Omega \subset E \to \mathbb{R}$ defined on a convex set Ω, and

$$U = \{v \in \Omega; \phi_i(v) \leq 0, 1 \leq i \leq m\} \tag{3.78}$$

a subset of Ω. The constraints $\phi_i : \Omega \subset E \to \mathbb{R}$ are assumed to be convex.
Let u in U in which ϕ_i and J are differentiable.

1. *If J admits a local minimum in u with respect to U, and if the constraints are qualified, then there exists $\lambda_i(u),\ 1 \leq i \leq m$ such that:*

$$\begin{cases} \nabla J(u) + \sum_{i=1}^m \lambda_i(u) \nabla \phi_i(u) = 0 \\ \lambda_i(u) \geq 0,\ 1 \leq i \leq m,\ \sum_{i=1}^m \lambda_i \phi_i(u) = 0 \end{cases} \tag{3.79}$$

2. *If $J : U \to \mathbb{R}$ convex, and if there exists λ_i which satisfy the above equation, then J admits in u a local minimum with respect to U.*

Proof:

1. It suffices to show that if the convex constraints ϕ_i are qualified in the above sens, then they are qualified in the previous sens (and we then can conclude with Kuhn and Tucker's theorem).

   We denote by $\tilde{v}$ the element of U of the above definition. For all i: $\phi_i(\tilde{v}) \leq 0$ with strict inequality if ϕ_i is not affine.

   Now, if $u \neq \tilde{v}$, take $\tilde{w} = \tilde{v} - u$. $\tilde{w}$ will satisfies the previous definition. Indeed, we have if $i \in I(u)$ (thus $\phi_i(u) = 0$):

$$\langle \nabla \phi_i(u), \tilde{w} \rangle = \phi_i(u) + \langle \nabla \phi_i(u), \tilde{w} \rangle \leq \phi_i(\tilde{v}) \tag{3.80}$$

   since ϕ_i convex. This conclude the proof if $u \neq \tilde{v}$.

   Otherwise, if $u = \tilde{v}$ then necessarily all the constraints ϕ_i are affine (since $\phi_i(u) = \phi_i(\tilde{v}) \leq 0$ with strict inequality if ϕ_i not affine).

2. Let v in U. Then, since $\lambda_i(u) = \lambda_i \geq 0$ and $\phi_i(v) \leq 0$:

$$J(u) \leq J(u) - \sum_{i=1}^m \lambda_i \phi_i(v) \tag{3.81}$$

   Since $\lambda_i = 0$ if i not in $I(u)$ and $\phi_i(u) = 0$ if $i \in I(u)$, we get:

$$J(u) \leq J(u) - \sum_{i=1}^m \lambda_i (\phi_i(v) - \phi_i(u)) \tag{3.82}$$

   Now, using the convexity of ϕ_i:

$$J(u) \leq J(u) - \sum_{i=1}^m \lambda_i \langle \nabla \phi_i(u), v - u \rangle \tag{3.83}$$

   Using Kuhn and Tucker relation, we have:

$$J(u) \leq J(u) + \langle \nabla J(u), v - u \rangle \tag{3.84}$$

   And since J convex, we eventually get $J(u) \leq J(v)$.

■

Interpretation: Let us consider:

$$F_u(v) = J(v) + \sum_{i=1}^{N} \lambda_i(u)\phi_i(v) \tag{3.85}$$

If the condition of the above theorem are satisfied, it is such that:

$$\begin{cases} J(u) = \inf_{v\in U} J(v) \implies \nabla F_u(u) = 0 \\ J(u) = F_u(u) \end{cases} \tag{3.86}$$

It means that if the λ_i are known, than we are back to an unconstrained minimization problem.

Example:

$$\begin{aligned} U &= \left\{ v \in \mathbb{R}^N, \sum_{j=1}^{N} c_{i,j} v_j \leq d_i, 1 \leq i \leq m \right\} \\ &= \left\{ v \in \mathbb{R}^N, \langle C_i, v\rangle \leq d_i, 1 \leq i \leq m \right\} \\ &= \left\{ v \in \mathbb{R}^N, Cv \leq d \right\} \end{aligned}$$

C_i column vectors of $C = (c_{i,j})$.
The constraints are: $\phi_i(v) = \langle C_i, v\rangle - d_i$. Thus $\nabla\phi_i = C_i$. Hence

$$\sum_i \lambda_i \nabla\phi_i = \sum_i \lambda_i C_i = C^T\lambda \tag{3.87}$$

$J : U \to \mathbb{R}$ convex. The constraints are qualified iff U is not empty.
J admits a local minimum with respect to U if there exists $\lambda \in \mathbb{R}^m$ such that:

$$\begin{cases} \nabla J(u) + C^T\lambda = 0 \\ \lambda \in \mathbb{R}^m_+, \text{ and } \lambda_i = 0 \text{ if } \langle C_i, u\rangle < d_i \end{cases} \tag{3.88}$$

3.3.4 Ideas from duality

Let V and M two sets, and a L a function:

$$L : V \times M \to \mathbb{R} \tag{3.89}$$

Definition 3.7. A point (u, λ) is said to be a saddle point (*point selle*) of L if u is a minimizer of $L(., \lambda) : v \in V \to L(v, \lambda) \in \mathbb{R}$ and if λ is a maximizer of $L(u, .) : \mu \in M \to L(u, \mu) \in \mathbb{R}$, i.e. if:

$$\sup_{\mu\in M} L(u, \mu) = L(u, \lambda) = \inf_{v\in V} L(v, \lambda) \tag{3.90}$$

Notice that in the definition, the two variables play different roles (cannot be changed).
In practice, $\mu \in M$ will denote a generalized lagrange multiplier.

Theorem 3.14. *If (u, λ) is a saddle point of L, then:*

$$\sup_{\mu\in M} \inf_{v\in V} L(v, \mu) = L(u, \lambda) = \inf_{v\in V} \sup_{\mu\in M} L(v, \mu) \tag{3.91}$$

Proof: Notice that we always have:

$$\sup_{\mu\in M}\inf_{v\in V} L(v,\mu) \leq \inf_{v\in V}\sup_{\mu\in M} L(v,\mu) \tag{3.92}$$

Indeed, if $\tilde{v}\in V$ and if $\tilde{\mu}\in M$, then:

$$\inf_{\tilde{v}\in V} L(\tilde{v},\mu) \leq L(\tilde{v},\tilde{\mu}) \leq \sup_{\mu\in M} L(\tilde{v},\mu) \tag{3.93}$$

To get the opposite inequality, we use the fact that (u,λ) is saddle point of L.

$$\inf_{v\in V}\sup_{\mu\in M} L(v,\mu) \leq \sup_{\mu\in M} L(u,\mu) = L(u,\lambda) \tag{3.94}$$

And

$$L(u,\lambda) = \inf_{v\in V} L(v,\lambda) \leq \sup_{\mu\in M}\inf_{v\in V} L(v,\mu) \tag{3.95}$$

■

From now on, we assume all the functions to be defined on the whole space $V=\mathbb{R}^N$ (for the sake of clarity).

We consider $J: V\to\mathbb{R}$ and $\phi_i: V\to\mathbb{R}$, $1\leq i\leq m$, and

$$U=\{v\in V=\mathbb{R}^N, \phi_i(v)\leq 0, 1\leq i\leq m\} \tag{3.96}$$

We consider the problem (primal problem) (P):

$$\text{Find } u\in U \text{ such that } J(u)=\inf_{v\in U} J(v) \tag{3.97}$$

Under some specific conditions, any solution u of (P) is the first argument of a saddle point (u,λ) of a certain function L called *Lagrangian* associated to problem (P). The second argument λ is called *generalized Lagrange multiplier* associated to u (since as we will see it is the vector given by the Kuhn and Tucker relations).

Let us define the *Lagrangien* associated to problem (P) as:

$$L:(v,\mu)\in V\times\mathbb{R}^m_+ \to L(v,\mu)=J(v)+\sum_{i=1}^m \mu_i\phi_i(v) \tag{3.98}$$

Theorem 3.15.

1. *If $(u,\lambda)\in V\times\mathbb{R}^m_+$ is a saddle point of the Lagrangien L, then u, which belongs to U, is solution of problem (P).*

2. *We assume that J and ϕ_i, $1\leq i\leq m$ are convex and differentiable in $u\in U$, and the constraints are qualified. Then, if u is solution of problem (P), there exists at least one vector $\lambda\in\mathbb{R}^m_+$ such that $(u,\lambda)\in V\times\mathbb{R}^m_+$ is a saddle point of L.*

Proof:

1. From the inequalities $L(u,\mu) \leq L(u,\lambda)$ for all $\mu \in \mathbb{R}_+^m$, we get:

$$\sum_{i=1}^{m} (\mu_i - \lambda_i)\phi_i(u) \leq 0 \tag{3.99}$$

Hence $\phi_i(u) \leq 0$ (by letting $\mu \to +\infty$). Moreover, with $\mu = 0$, we get $\sum_{i=1}^m \lambda_i \phi_i(u) \geq 0$. But since $\lambda_i \geq 0$ and $\phi_i(u) \leq 0$, we know that $\sum_{i=1}^m \lambda_i \phi_i(u) \leq 0$. As a consequence, we have:

$$u \in U \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i \phi_i(u) = 0 \tag{3.100}$$

Now using the fact that $L(u,\lambda) \leq L(v,\lambda)$ for all $v \in V$ we get:

$$\underbrace{\sum_{i=1}^{m} \lambda_i \phi_i(u)}_{=0} + J(u) \leq J(v) + \sum_{i=1}^{m} \lambda_i \phi_i(v) \tag{3.101}$$

And since $\phi_i(v) \leq 0$, we get $J(u) \leq J(v)$.

2. We can apply Kuhn and Tucker's theorem: if u is a solution of problem (P), then there exists $\lambda \in \mathbb{R}_+^m$ such that:

$$\sum_{i=1}^{m} \lambda_i \phi_i(u) = 0 \quad \text{and} \quad \nabla J(u) + \sum_{i=1}^{m} \lambda_i \nabla \phi_i(u) = 0 \tag{3.102}$$

From the first equality, we get for all $\mu \in \mathbb{R}_+^m$:

$$L(u,\mu) = J(u) + \sum_{i=1}^{m} \mu_i \phi_i(u) \leq J(u) = L(u,\lambda) \tag{3.103}$$

And the second equality is a sufficient condition for the follwing convex function (as sum of convex functions) to have a minimum:

$$L(.,\lambda) : v \to J(v) + \sum_{i=1}^{m} \lambda_i \phi_i(v) \tag{3.104}$$

Hence $L(u,\lambda) \leq L(v,\lambda)$ for all $v \in V$. And thus (u,λ) is saddle point of the Lagragien L.

■

Up to now, we have established that, under some hypotheses, a solution u of the primal problem (P) is the first argument of a saddle point of its associated Lagrangien.

Assume that we know one of the second argument λ of the saddle points of L. Then the constrained problem (P) would be replaced by an unconstrained problem (P_λ): find u_λ such that:

$$u_\lambda \in V \text{ and } L(u_\lambda,\lambda) = \inf_{v \in V} L(v,\lambda) \tag{3.105}$$

Now the question is how to find such a $\lambda \in \mathbb{R}^m_+$. Recall the equality verified by saddle points:

$$L(u_\lambda, \lambda) = \inf_{v \in V} L(v, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{v \in V} L(v, \mu) \tag{3.106}$$

We are therefore led to find λ as a solution of the dual problem (Q):

$$\text{Find } \lambda \in \mathbb{R}^m_+ \text{ such that } G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu) \tag{3.107}$$

where $G : \mathbb{R}^m_+ \to \mathbb{R}$ is defined by:

$$G : \mu \in \mathbb{R}^m_+ \to G(\mu) = \inf_{v \in V} L(v, \mu) \tag{3.108}$$

$\mu \in \mathbb{R}^m_+$ is called the dual variable of the primal variable $v \in V = \mathbb{R}^N$.

Notice that the dual problem (Q) is also a constrained problem, but the constraints $\mu_i \geq 0$ are very easy to handle (since we know explicitely the projection operator). On the contrary, the constraints $\phi_i(u) \leq 0$ are in general impossible to handle numerically. This is the basic idea of Uzawa algorithm.

Theorem 3.16.

1. *We assume that $\phi_i : V \to \mathbb{R}$ is continuous, and that for all $\mu \in \mathbb{R}^m_+$, problem (P_μ):*

$$\textit{find } u_\mu \in V \textit{ such that } L(u_\mu, \mu) = \inf_{v \in V} L(v, \mu) \tag{3.109}$$

   has a unique solution u_μ which depends continuously of $\mu \in \mathbb{R}^m_+$.

   Then, if λ is a solution of the dual problem (Q), the solution u_λ of the corresponding problem (P_λ) is a solution of the primal problem (P).

2. *We assume that the primal problem (P) has at least one solution u, that J and ϕ_i are convex functions, differentiable in u, and that the constraints are qualified. Then the dual problem (Q) has at least one solution.*

Sketch of the proof:

1. Let λ a solution of problem (Q). we already have:

$$\lambda \in \mathbb{R}^m_+ \quad \text{and} \quad G(\lambda) = L(u_\lambda, \lambda) = \inf_{v \in V} L(v, \lambda) \tag{3.110}$$

   We want to show that:

$$\sup_{\mu \in \mathbb{R}^m_+} L(u_\lambda, \mu) = L(u_\lambda, \lambda) \tag{3.111}$$

   These two relations are exactly the defintion of a saddle point (u_λ, λ) of the Lagrangien L, and we will thus deduce that u_λ is a solution of problem (P).

   The first point is to show that the function G is differentiable. It can be shown that if $\xi \in \mathbb{R}^m$:

$$\langle \nabla G(u), \xi \rangle = \sum_{i=1}^{m} \xi_i \phi_i(u_\mu) \tag{3.112}$$

Since G has a maximum in λ on the convex $\mathbb{R}^m_+$, we have for all $\mu \in \mathbb{R}^m_+$:

$$\langle \nabla G(\lambda), \mu - \lambda \rangle \leq 0 \tag{3.113}$$

i.e. for all $\mu \in \mathbb{R}^m_+$:

$$\sum_{i=1}^{m} \mu_i \phi_i(u_\lambda) \leq \sum_{i=1}^{m} \lambda_i \phi_i(u_\lambda) \tag{3.114}$$

We thus have:

$$\begin{aligned} L(u_\lambda, \mu) &= J(u_\lambda) + \sum_{i=1}^{m} \mu_i \phi_i(u_\lambda) \\ &\leq J(u_\lambda) + \sum_{i=1}^{m} \lambda_i \phi_i(u_\lambda) \\ &= L(u_\lambda, \lambda) \end{aligned}$$

which is the second inequality characterizing a saddle point.

2. From the previous theorem, there exists at least one $\lambda \in \mathbb{R}^m_+$ such that (u, λ) is a saddle point of the Lagrangien L. The theorem on saddle point then imply that:

$$L(u, \lambda) = \inf_{v \in V} L(v, \lambda) = \sup_{\mu \in \mathbb{R}^m_+} \inf_{v \in V} L(v, \mu) \tag{3.115}$$

i.e: $G(\lambda) = \sup_{\mu \in \mathbb{R}^m_+} G(\mu)$.

■

Example: Consider the example of a quadratic functional:

$$J(v) = \frac{1}{2}\langle Av, v \rangle - \langle b, v \rangle \tag{3.116}$$

with A symmetric definite positive matrix of size N^2, and $b \in \mathbb{R}^N$.

$$U = \left\{ v \in \mathbb{R}^N, \sum_{j=1}^{N} c_{i,j} v_j \leq d_i, 1 \leq i \leq m \right\} = \left\{ v \in \mathbb{R}^N, Cv \leq d \right\} \tag{3.117}$$

where C is a real $m \times N$ matrix, and $d \in \mathbb{R}^m$. We assume U non empty (and thus the constraints are qualified).

We want to find u such that $J(u) = \inf_{v \in U} J(v)$. It is easy to see that there exists a unique solution for the primal problem (P).

The lagrangien L is:

$$L(v, \mu) = \frac{1}{2}\langle Av, v \rangle_{\mathbb{R}^N} - \langle b - C^T \mu, v \rangle_{\mathbb{R}^N} - \langle \mu, d \rangle_{\mathbb{R}^n} \tag{3.118}$$

Moreover, the gradient of $L(., \mu)$ is zero in u_μ., i.e.

$$Au_\mu = b - C^T \mu \iff u_\mu = A^{-1}(b - C^T \mu) \tag{3.119}$$

Then, after some computations:

$$G(\mu) = -\frac{1}{2}\langle CA^{-1}C^T\mu, \mu\rangle_m + \langle CA^{-1}b - d, \mu\rangle_m - \frac{1}{2}\langle A^{-1}b, b\rangle_N \tag{3.120}$$

We can then show that the dual problem (Q) has always a solution. This solution is unique if rank(C)=m.

Remark that:

$$\nabla G(\mu) = Cu_\mu - d = -CA^{-1}C^T\mu + CA^{-1}b - d \tag{3.121}$$

i.e. $(\nabla G(\mu))_i = \phi_i(u_\mu)\ ,\ 1 \leq i \leq m$.

3.3.5 Uzawa algorithm

The idea of the method is that the projection operator $P_+ : \mathbb{R}^m \to \mathbb{R}^m_+$ in the dual problem (Q) is very simple:

$$(P_+\lambda)_i = \max(\lambda_i, 0) \tag{3.122}$$

Uzawa algorithm is in fact the projected gradient method applied to the dual problem (Q): $\lambda_0 \in \mathbb{R}^m_+$ arbitrary, and the sequence λ_k in $\mathbb{R}^m_+$ is defined by:

$$\lambda_{k+1} = P_+(\lambda_k - \rho\nabla G(\lambda_k)) \tag{3.123}$$

Since in the dual problem (Q), one is interested in a maximum (and not a minimum), it is therefore natural to change the sign of the parameter ρ with respect to the classical method.

Under some hypotheses, it is possible to compute the gradient of G:

$$(\nabla G(\mu))_i = \phi_i(u_\mu)\ ,\ 1 \leq i \leq m \tag{3.124}$$

the vector u_μ being the solution of the *unconstrained* minimization problem:

$$u_\mu \in V,\ J(u_\mu) + \sum_{i=1}^m \mu_i\phi_i(u_\mu) = \inf_{v\in V}\left\{J(v) + \sum_{i=1}^m \mu_i\phi_i(v)\right\} \tag{3.125}$$

Uzawa algorithm: $\lambda_0 \in \mathbb{R}^m_+$ arbitrary. We define by induction $(\lambda^k, u^k) \in \mathbb{R}^m_+ \times V$ by (for the sake of clarity, we write $u^k = u_{\lambda_k}$):

$$\left\{\begin{array}{ll} \text{Computation of } u^k: & J(u^k) + \sum_{i=1}^m \lambda_i^k\phi_i(u^k) = \inf_{v\in V}\left\{J(v) + \sum_{i=1}^m \lambda_i^k\phi_i(v)\right\} \\ \text{Computation of } \lambda_i^{k+1}: & \lambda_i^{k+1} = \max\left\{\lambda_i^k + \rho\phi_i(u^k), 0\right\} \end{array}\right. \tag{3.126}$$

Uzawa method is a way to replace a constrained minimization problem by a sequence of unconstrained minimization problem.

Notice that u^k can converge while λ^k does not.

Theorem 3.17. Convergence of Uzawa method

$V = \mathbb{R}^N$. J elliptic (with contant α), and:

$$U = \{v \in \mathbb{R}^N, Cv \leq d\},\ C \in \mathcal{A}_{m,N}(\mathbb{R}),\ d \in \mathbb{R}^m \tag{3.127}$$

is non empty. Then, if

$$0 < \rho < \frac{2\alpha}{\|C\|^2} \tag{3.128}$$

the sequence u_k converges to the unique solution of the primal problem (P).

If the rank of C is m, then the sequence λ^k also converges towards the unique solution of the dual problem (Q).

Notice that $\|C\| = \sup_{v\in\mathbb{R}^N} \frac{\|Cv\|_m}{\|v\|_n}$.

Case of a quadratic functional:

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle \tag{3.129}$$

with A symmetric definite positive matrix of size N^2, and $b \in \mathbb{R}^N$.

$$U = \left\{ v \in \mathbb{R}^N, \sum_{j=1}^{N} c_{i,j} v_j \leq d_i, 1 \leq i \leq m \right\} = \left\{ v \in \mathbb{R}^N, Cv \leq d \right\} \tag{3.130}$$

where C is a real $m \times N$ matrix, and $d \in \mathbb{R}^m$. We assume U non empty (and thus the constraints are qualified).

An iteration of Uzawa algorithm is:

$$\left\{ \begin{array}{ll} \text{Computation of } u^k\text{:} & Au^k - b + C^T\lambda^k = 0 \\ \text{Computation of } \lambda_i^{k+1}\text{:} & \lambda_i^{k+1} = \max\{(\lambda^k + \rho(Cu^k - d))_i, 0\} \end{array} \right. \tag{3.131}$$

And the method converges if

$$0 < \rho < \frac{2\lambda_1(A)}{\|C\|^2} \tag{3.132}$$

where $\lambda_1(A)$ is the smallest eigenvalue of A (it is the ellipticity constant).

References

[1] G. Allaire. *Analyse numérique et optimisation*. Ecole Polytechnique, 2005.

[2] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing*, volume 147 of *Applied Mathematical Sciences*. Springer-Verlag, 2002.

[3] M. Bergounioux. *Optimisation et controle des systèmes linéaires*. Dunod, 2001.

[4] J-F. Bonnans, J-C. Gilbert, C. Lemaréchal, and C. Sagastizabal. *Optimisation numérique.* Springer, 1997.

[5] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et l'optimisation*. Mathématiques appliquées pour la maitrise. Masson, 1984.

[6] J-C. Culioli. *Introduction à l'optimisation*. Ellipses, 1994.

[7] J.B. Hiriart-Urruty and C. Lemarechal. *Convex Analysis ans Minimisation Algorithms I*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, 1993.

[8] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[9] C.R. Vogel. *Computational Methods for Inverse Problems*, volume 23 of *Frontiers in Applied Mathematics*. SIAM, 2002.

[10] E. Weiszfeld. Sur le point pour lequel la somme des distances de points donnés est minimum. *Tôhoko Mathematics Journal*, 43:355–386, 1937.