SEDOCONF SEDOCONF2024

Numerical methods for optimisation

CIMPA School in Cape Coast, Ghana,June 2024

Marie Postel

Laboratoire Jacques-Louis Lions

Sorbonne Université

marie.postel@sorbonne-universite.fr

https://github.com/letsop/optim

Introduction to optimization

- Reminders : Differential calculus and convexity Memory help
- Isovalues, contouring, level curves
- Numerical approximation of derivatives and gradient

Unconstrained optimisation

- Optimality conditions in the unconstrained case
- Solving systems of non linear equations
- Descent methods
- Linear and non linear regression

Optimisation with constraints

- Duality
- Optimality conditions for equality constraints
- Optimality conditions for inequality constraints
- Algorithms for constrained optimization

Outline

Course objectives

- Introduction to numerical methods of Optimization
- Test of algorithms in Python:
  - Install `jupyter notebook` for `python 3`
  - Download material at https://github.com/letsop/optim
  - Run notebooks locally or via googlecolab (needs a google ID)

Bibliography

- Stephen Boyd, Lieven Vandenberghe, Convex Optimization. Cam-bridge University Press (2004) https://stanford.edu/ boyd/cvxbook/
- Jorge Nocedal , Stephen J. Wright, Numerical Optimization, Spinger (2006)

  https://www.math.uci.edu/ qnie/Publications/NumericalOptimization.pdf

Different categories of optimization

- Discrete optimization : variables in a discrete set
  - Combinatorial <-> linear programming
  - "NP-complete" (nondeterministic polynomial-time complete)
  - Logistics, Economy (Traveling salesman, Knapsack, etc.)
  - Heuristic methods : Hill climbing, Simulated annealing, Ant colony, etc.
- Continuous optimization : variables within a range of values
  - Infinite dimensions : calculus of variations, shape optimization, control theory
  - Finite dimension : includes the discretization of above problems

Definition of a minimum

a neigborhood of x^*

Def : Let $f : V \to \mathbb{R}$ with V normed vector space.
$x^\star \in D_a \subset V$ **achieves**

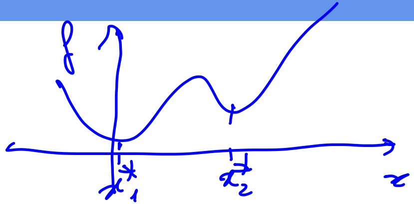
- a local minimum on D_a if there exists $\varepsilon > 0$ such that

$$f(x^\star) \leq f(x) \quad \text{for all} \quad x \in D_a \quad \text{t.q. } \|x - x^\star\| \leq \varepsilon.$$

- a strict local minimum if there exists $\varepsilon > 0$ such that

$$f(x^\star) < f(x) \quad \text{for all} \quad x \in D_a \quad \text{s. t. } x \neq x^\star \text{ and } \|x - x^\star\| \leq \varepsilon.$$

Definition of a minimum



Def : $x^\star \in D_a$ **achieves**

- a global minimum on D_a if

$$f(x^\star) \leq f(x) \quad \text{for all} \quad x \in D_a.$$

$$f(x^*) = \min_{x \in D_a} f(x)$$

- a strict global minimum if

$$f(x^\star) < f(x) \quad \text{for all} \quad x \in D_a \quad \text{s. t. } x \neq x^\star.$$

It is sometimes said that $x^\star$ **is a minimum** of $f(x)$, but this is a misnomer. The exact term, if $x^\star$ **realizes** a minimum of f, is that it is a **minimizer** of f, denoted

$$x^\star = \operatorname*{argmin}_{x \in D_a} f(x)$$

Definition of a maximum

To find the maximum of f we search the minimum of $-f$.

Optimization applied to differential problems : Calculus of Variations

$< f, g >_{L^2([0,1])} = \int_0^1 f(x) g(x)\, dx$

Let $V_0 = \{u \in C^2([0,1]),\ u(0) = u(1) = 0\}$ and $g : \mathbb{R}^3 \to \mathbb{R}$, $g \in \mathcal{C}^1$.

$\mathcal{J} : V_0 \longrightarrow \mathbb{R}$

$$\mathcal{J}(u) = \int_0^1 g(x, u(x), u'(x))dx, \quad u \in V_0.$$

$$D\mathcal{J}(u)(v) = \langle -\frac{d}{dx}\frac{\partial g}{\partial u'}(x, u, u') + \frac{\partial g}{\partial u}(x, u, u'), v\rangle_{L^2([0,1])}.$$

Euler-Lagrange Theorem: An extremum of $\mathcal{J}$ satisfies

$$-\frac{d}{dx}\frac{\partial g}{\partial u'}(x, u, u') + \frac{\partial g}{\partial u}(x, u, u') = 0$$

Infinite dimension example

Let $V_0 = \{u \in C^2([0,1]),\ u(0) = u(1) = 0\}$, $f \in C^1([0,1])$ and $g : \mathbb{R}^3 \to \mathbb{R}$, $g \in \mathcal{C}^1$.

$$\mathcal{J}(u) = \int_0^1 \frac{1}{2}u'(x)^2 + \frac{1}{2}u(x)^2 - f(x)u(x)dx, \quad u \in V_0.$$

$$g(x,y,z) = \frac{1}{2}z^2 + \frac{1}{2}y^2 - f(x)y.$$

$$\begin{aligned} D\mathcal{J}(u)(v) &= \langle -\frac{d}{dx}\frac{\partial g}{\partial z}(x,u(x),u'(x)) + \frac{\partial g}{\partial y}(x,u(x),u'(x)), v\rangle \\ &= \int_0^1 (-u'' + u - f)vdx. \end{aligned}$$

$$\boxed{\mathcal{J}(\bar{u}) = \min \mathcal{J}(u) \quad \overset{\Longleftrightarrow}{\text{iff}} \quad \begin{cases} -u'' + u = f \\ u(0) = u(1) = 0 \end{cases}}$$

Canonical Continuous Optimization problem on $\mathbb{R}^n$

Find the extrema of a function $f(x)$ defined on $\mathbb{R}^n$ (or part of $\mathbb{R}^n$ in the case of a optimization with constraints).

Find

$$\inf_{x\in\mathbb{R}^n} f(x),$$

infimum
→ min

under constraints

$$\begin{array}{rcl} C^E(x) & = & 0, \\ C^I(x) & \preceq & 0 \quad (\Leftrightarrow C^I_i(x) \leq 0,\ i = 1, \ldots, p) \end{array}$$

equality constraints $C_j^E(x) = 0$ $j=1,\dots m$

inequality constraints

with

$$\begin{array}{rcl} f & : & \mathbb{R}^n \longrightarrow \mathbb{R}, \\ C^I & : & \mathbb{R}^n \longrightarrow \mathbb{R}^p, \\ C^E & : & \mathbb{R}^n \longrightarrow \mathbb{R}^m, \quad f, C^I, C^E, \text{smooth} \end{array}$$

Admissible domain

$$D_a = \{x \in \mathbb{R}^n,\ C^E(x) = 0,\ C^I(x) \preceq 0\}$$

Example 1: linear programming

$c^T x = \langle x, c \rangle = c \cdot x = \sum_{i=1}^{n} c_i x_i$

Generic problem

$$(P)\left\{\begin{array}{cc} \text{Minimize} & \boxed{c^T x} \\ \text{under constraints} & Ax \preceq b \\ & x \succeq 0 \end{array}\right. \quad \text{with} \quad \begin{array}{l} c \text{ and } x \in \mathbb{R}^n, \\ b \in \mathbb{R}^m, \\ A \in \mathcal{M}_{m\times n}(\mathbb{R}) \end{array}$$

$(Ax)_i \leq b_i$

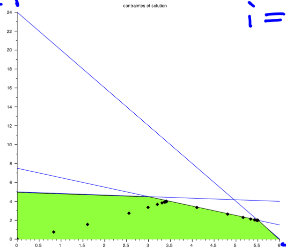
$i = 1, \dots, m$

Example and admissible domain

$$A = \begin{pmatrix} 1 & 6 \\ 2 & 2 \\ 4 & 1 \end{pmatrix}, \; b = \begin{pmatrix} 30 \\ 15 \\ 24 \end{pmatrix}$$

$$c = \begin{pmatrix} -2 \\ -1 \end{pmatrix}$$

x_2



x_1

Rewrite example 1 in canonical form

$$\inf_{x\in\mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

$\inf c^T x$

$Ax \leq b$

$x \geq 0$

- $f(x) = c^T x = c \cdot x$
- $C^E(x) = \emptyset$
- $C^I(x)$ $\quad C^I : \mathbb{R}^n \to \mathbb{R}^5 \quad x \mapsto C^I(x) = \begin{cases} (Ax)_1 - b_1 \\ (Ax)_2 - b_2 \\ (Ax)_3 - b_3 \\ -x_1 \\ -x_2 \end{cases}$
- D_a, def, nature

$D_a = \{ x \in \mathbb{R}^2 , \; C^I(x) \leq 0_{\mathbb{R}^5} \}$

Solve with the Python toolbox linprog

```
scipy.optimize.linprog (c, A_ub=None,
B_ub=None, A_eq=None, B_eq=None, bounds=None,
method='interior-point', callback=None,
options=None, x0=None)
```

The problem must be written in the form expected by the program

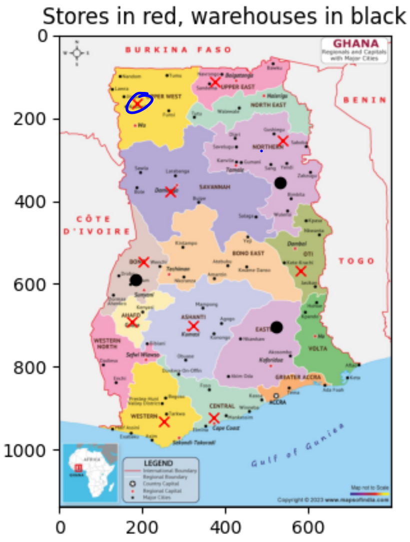
$$\begin{aligned} \min_x \quad & c^T x \\ \text{such that} \quad & A_{ub} x \preceq B_{ub} \\ & A_{eq} x = B_{eq} \\ & \ell \preceq x \preceq u \end{aligned}$$

Solve with the Python toolbox linprog

notebook ddlivraries.ipynb

```
c = [-2, -1]
Aub = [[1, 6], [2, 2], [4,1]]
Bub = [30,15,24]
lu = (0., None)
bounds=2*[lu]
res = scipy.optimize.linprog(c, A_ub=Aub,
b_ub=Bub, bounds=bounds)
```

Practical example



- A company stores a commodity in M warehouses.
- Each warehouse i ($i = 1, \ldots, M$) has a quantity q_i of goods in stock.
- The company has a network of N stores.
- Each store j ($j = 1, \ldots, N$) ordered a quantity r_j of goods.
- The problem is to minimize the cost of delivering goods to stores.

Mathematical modelling

Let us denote

- $v_{i,j}$ the quantity of merchandise shipped from warehouse i to store j
- $Q = \sum_{i=1}^{M} q_i$ the total quantity of goods available in the warehouses
- $R = \sum_{j=1}^{N} r_j$ the total quantity of goods ordered by the stores, assuming $Q \geq R$
- $D_{i,j}$ the cost of unit transport from the warehouse i to the store j, directly proportional to the distance between the store and the warehouse.

Rewriting as a linear programming problem

v_{ij} amount of good delivered from Wh_i to S_j

The problem (whose unknowns are the $v_{i,j}$) is therefore to minimize

$$\sum_{i=0}^{M-1}\sum_{j=0}^{N-1} D_{i,j} v_{i,j}$$

with respect to v, under the constraints

(i) $v_{i,j} \geq 0$ we do not return goods from a store to a warehouse

inequality

(ii) $\sum_{j=0}^{N-1} v_{i,j} \leq q_i$ a warehouse cannot supply more than its stock

(iii) $\sum_{i=0}^{M-1} v_{i,j} = r_j$ each store must receive the requested quantity

equality constraints

Rewriting under canonical form

$$\inf_{x\in\mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

$n = N \times M$ # wh & # stores

$v = (v_{ij})_{i=1,N,\ j=1,\dots M} \rightsquigarrow v_k,\ k=1,\dots N\times M$

- f : $\mathbb{R}^{N\times M} \longrightarrow \mathbb{R}$
- C^E : $\mathbb{R}^{N\times M} \longrightarrow \mathbb{R}^M$
- C^I : $\mathbb{R}^{N\times M} \longrightarrow \mathbb{R}^N$
- D_a, def, nature

$f(v) = \sum_i \sum_j v_{ij} d_{ij}$

$C^E(v) = A_{eq} V - b_{eq}$ with $A_{eq} \in \mathcal{M}_{M\times NM}(\mathbb{R})$

$C^I(v) = A_{in} V - b_{in}$ with $A_{in} \in \mathcal{M}_{N\times NM}(\mathbb{R})$

Solve with the Python toolbox linprog

```
scipy.optimize.linprog (c, A_ub=None,
B_ub=None, A_eq=None, B_eq=None, bounds=None,
method='interior-point', callback=None,
options=None, x0=None)
```

The problem must be written in the form expected by the program

$$\begin{array}{rl} \min_x & c^T x \\ \text{such that} & A_{ub} x \preceq B_{ub} \\ & A_{eq} x = B_{eq} \\ & \ell \preceq x \preceq u \end{array}$$

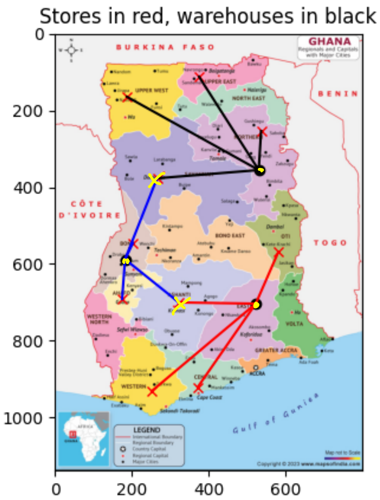
`c,x, B_ub, b_eq` are 1-D arrays or lists, `A_ub, A_eq` are 2-D arrays or lists. `bounds` is a list of 2 1-D lists

Solve with the Python toolbox linprog

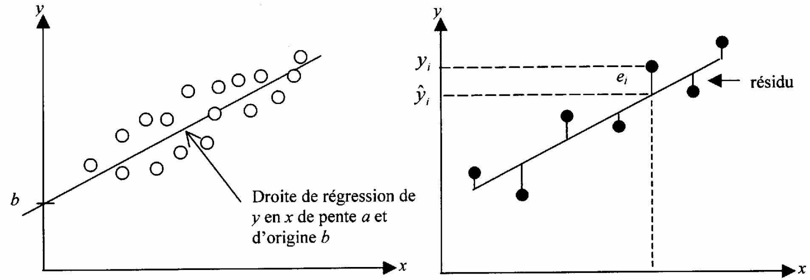
We must therefore define Python structures

- $x \in \mathbb{R}^{NM \times 1}$ contains the solution matrix v unrolled in columns
- $c \in \mathbb{R}^{NM \times 1}$ contains the matrix D unrolled in columns
- $Aub \in \mathcal{M}_{M,MN}(\mathbb{R})$ contains 1s in the right places so that $(Aub\ x)_i = \sum_{j=0}^{N-1} v_{i,j}$ for $i = 0, \dots, M-1$
- $Bub \in \mathbb{R}^M$ contains q (the warehouse stocks)
- $Aeq \in \mathcal{M}_{N,MN}(\mathbb{R})$ contains 1s in the right places so that $(Aeq\ x)_j = \sum_{i=0}^{M-1} v_{i,j}$ for $j = 0, \dots, N-1$
- $Beq \in \mathbb{R}^N$ contains r (the store orders)
- $\ell \in \mathbb{R}^{NM \times 1} = 0$

Python solution of this Linear Programming Problem

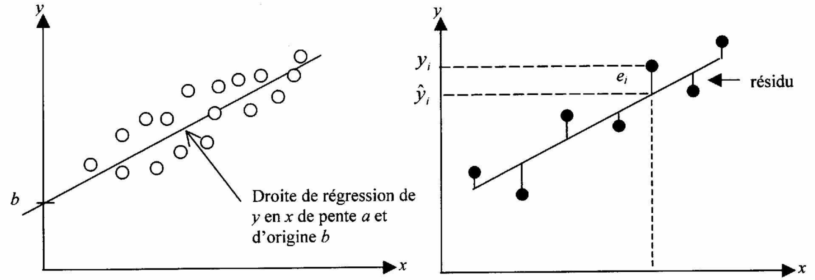


Example 2 : Least Squares



Data $x_i, y_i, i = 1, \ldots, n$
Linear model $y = ax + b$
Minimize $\sum_{i=1}^{n}(ax_i + b - y_i)^2$ with respect to $(a, b) \in \mathbb{R}^2$

Example 2 : Least Squares - vector formulation



Data $x_i, y_i, i = 1, \ldots, n$
Linear model $y = ax + b$

$$X = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad P = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathbb{R}^2$$

Minimize $\|Y - XP\|_2^2$ with respect to $P \in \mathbb{R}^2$

Rewrite Least Square problem in canonical form

$X = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}$ $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$\inf_{x \in \mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

$\min \| XP - Y \|^2$

$f : \mathbb{R}^2 \to \mathbb{R}$

$P \mapsto f(P) = \| XP - Y \|^2$

- f
- C^E $\emptyset$
- C^I $\emptyset$
- D_a, def, nature $D_a = \mathbb{R}^2$

Example 3 : $\boxed{\text{Non differentiable}}$ convex Optimization

Parsimonious Least Squares Lasso (least absolute shrinkage and selection operator)

- Sociological models (e.g. explanation of academic success as a function of social, family, medical factors, etc.)
- Data $Y = (y_i)_{i=1,\dots,n}$, $X = (x_{i,j})_{i=1,\dots,n,j=1,\dots,p}$
- Linear model $\tilde{Y} = XP$, with $P \in \mathbb{R}^p$, using as few factors as possible
- Minimize $\underbrace{\|Y - XP\|_2^2}_{\text{LSQ}} + \alpha\|P\|_1$

$\|P\|_1 = \sum_{i=1}^{p} |P_i|$

weight.

adjusted

Example 4 : Classification

How to define a linear model $f(x) = w^t x + b$ with $w \in \mathbb{R}^N$ allowing to divide a family of points $(x_i, y_i)_{i=1,\dots,M}$ with coordinates $(x_i^j)_{j=1,\dots,N}$ so that the safety band on either side of the hyperplane delimiting the two classes is as wide as possible?

Simple case $y_i = \pm 1$ (yes or no), model answer $f(x) \geq 0$ for yes and $f(x) < 0$ no

The product $yf(x)$ should be "as positive as possible"

Keep $\|w\|$ as small as possible

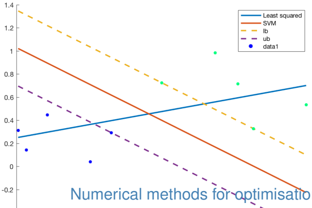
- Hard Margin classifier: we do not accept any intruders in the band
- Soft Margin classifier: we accept intruders in the band

Exemple 4 : Classification avec Linear Support Vector Machine

$$\min_{w,b} \frac{1}{M} \sum_{i=1}^{M} L(y_i(w^t x_i + b)) + \gamma \|w\|^2$$

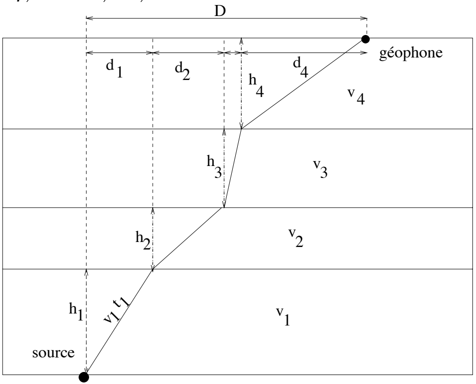
with L decreasing function

- Hard Margin classifier : $L(u) = \begin{cases} +\infty & \text{if } u < 1 \\ 0 & \text{otherwise} \end{cases}$
- Soft Margin classifier : $L(u) = \begin{cases} 1 - u & \text{if } u < 1 \\ 0 & \text{otherwise} \end{cases}$



Example 5: Wave propagation in a stratified medium by "ray tracing"

n parallel layers of thickness $h_i, i = 1, ..., n$.
In each layer the speed of propagation is constant and equals $v_i, i = 1, ..., n$.



Optimization problem

Path followed by the seismic wave from the source to a geophone on the surface, at distance D from the vertical of the epicenter.

Descartes' law

$$\frac{\sin(\theta_i)}{v_i} = constant$$

Find the minimum travel time $\sum_i t_i$ under constraints

$D = \sum_{i=1}^{n} d_i$.
$d_i^2 + h_i^2 = v_i^2 t_i^2$

Canonical form of the optimization problem

$$\inf_{x\in\mathbb{R}^n} f(x) \text{ under constraints } \left\{ \begin{array}{l} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{array} \right.$$

- f
- C^E
- C^I
- D_a, def, nature

Canonical form of the optimization problem

Choice of unknowns: $(t_i)_{i=1,\dots,n}$ or $(d_i)_{i=1,\dots,n}$

- If $X = (t_i)_{i=1,\dots,n}$

$$f(X) = \sum_{i=1}^{n} x_i \text{ and } C^E(x) = \sum_{i=1}^{n} \sqrt{v_i^2 {x_i}^2 - h_i^2} - D$$

- If $X = (d_i)_{i=1,\dots,n}$

$$f(X) = \sum_{i=1}^{n} \frac{\sqrt{{x_i}^2 + h_i^2}}{v_i} \text{ and } C^E(x) = \sum_{i=1}^{n} x_i - D$$

Equality or inequality constraints

$$C^E(x) = \sum_{i=1}^{n} x_i - D = 0 \quad \Leftrightarrow \quad \left\{ \begin{array}{l} C_1^I(x) = \sum_{i=1}^{n} x_i - D \leq 0 \\ C_2^I(x) = D - \sum_{i=1}^{n} x_i \leq 0 \end{array} \right.$$

Special case of absolute values

$$\inf_{x \in \mathbb{R}^n} f(x) \text{ under constraints } |g(x)| \preceq b$$
$$\text{with } g : \mathbb{R}^n \to \mathbb{R}^d \text{ and } b \in \mathbb{R}^d_+$$

Define C^I

Example 6: Epidemy model

$f: \mathbb{R}^6 \to \mathbb{R}$
$P \mapsto f(P)$

$f(P) = \sum_{i=1}^{N_d} |I(t_i) - I_i|^2$

SIRC model

- $S(t)$, proportion of *susceptibles* persons
- $I(t)$, proportion of *infected* persons $\longrightarrow$ Data $(I_i)_{i=1}^{N_d}$
- $R(t)$, proportion of *recovered* persons
- $C(t)$, proportion of *cross immuned* persons

I_i = # infected at time t_i, $i = 1, .. N_d$

$$
\left\{
\begin{array}{lcl}
\dot{S}(t) & = & \mu(1-S) - \beta SI + \gamma C, \\
\dot{I}(t) & = & \beta SI + \sigma\beta CI - (\mu+\alpha)I, \\
\dot{R}(t) & = & (1-\sigma)\beta CI + \alpha I - (\mu+\delta)R, \\
\dot{C}(t) & = & \delta R - \beta CI - (\mu+\gamma)C,
\end{array}
\right.
\tag{1}
$$

$I(t)$ is the solution of the ODE system at time t

Parameters $P = (\mu, \alpha, \beta, \gamma, \delta, \sigma)$ $(M = 6)$

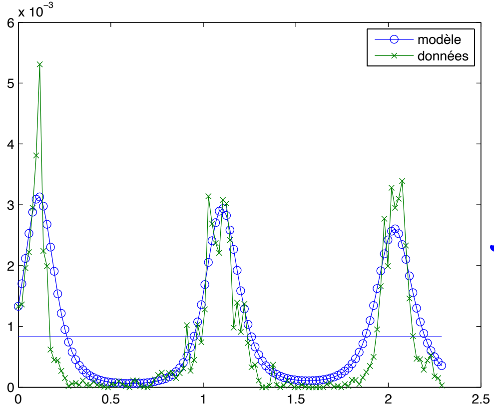
$I(t, P)$

Adequation of model with data

$f(P) = \sum_j \left(I(t_j, P) - \tilde{I}_j \right)^2$

Data $(\tilde{I}_j)_{j=1,\dots,d}$ to be compared with values predicted by the model $(I(t_j))_{j=1,\dots,d}$

Proportion of flu in Paris region between Jan 2007 and April 2009 (source : "Reseau Sentinelle")



$\min_{\substack{P \in \mathbb{R}^6 \\ P \geq 0}} f(P)$

$I(t, P)$ solution of ODE

$\nabla_P I(t, P)$

Rewrite Epidemic problem in canonical form

$$\inf_{x\in\mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

- f
- C^E
- C^I
- D_a, def, nature

Shape optimization

Optimize the shape of a plane wing to minimize the CO_2 impact factor of the plane.
https://interstices.info/la-forme-ideale-dune-aile/

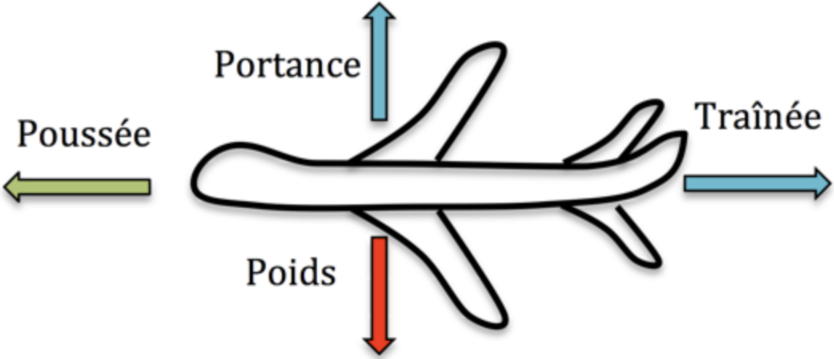


Figure 1 : forces exercées sur un avion.
Forces aérodynamiques en bleu, poussée motrice en vert, poids en rouge.

Shape optimization - scientific computing

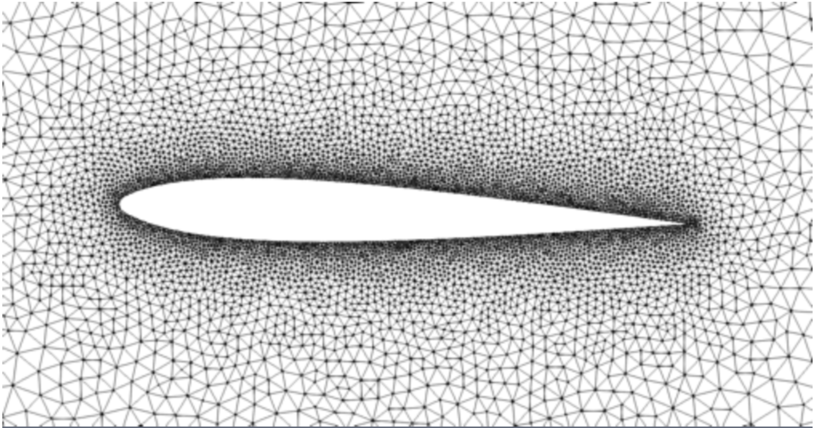


Figure 2 : Maillage autour d'un profil d'aile.

Shape optimization - model parameters

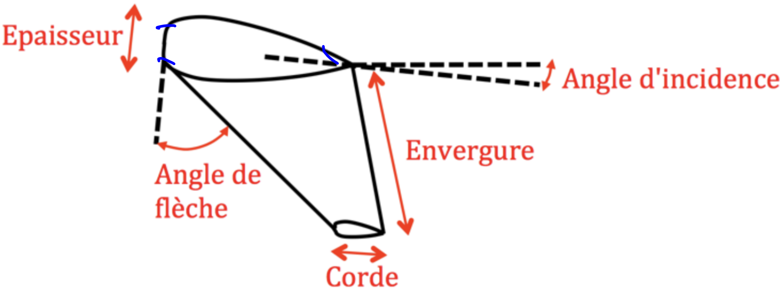


Figure 3 : paramètres décrivant globalement une forme d'aile.

Canonical form of the optimization problem

$$\inf_{x\in\mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

$f : \mathbb{R}^5 \longrightarrow \mathbb{R}$

$f(x) \longrightarrow$ cost of making this wing fly

- f
- C^E
- C^I

solution of PDE on a domain including wing(x)

- D_a, def, nature

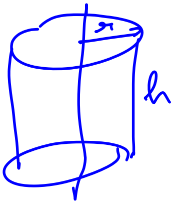
Exercise

A cylindrical container should hold $20\pi m^3$. The price of the material constituting the bottom and the cover is 10 euros $/m^2$, that of the material constituting the sides is 8 euros $/m^2$. Write the optimisation problem to find the dimensions (radius r and height h) of the most economical container:

$$\inf_{x \in \mathbb{R}^n} f(x) \text{ under constraints } \begin{cases} C^E(x) = 0, \\ C^I(x) \preceq 0 \end{cases}$$

$f : \mathbb{R}^2 \to \mathbb{R}$

$x = (r, h) \mapsto f(x) = 20\pi r^2 + 16\pi r h$



- f
- $C^E(x) = \text{Vol} - 20\pi = \pi r^2 h - 20\pi$
- $C^I(x) = -x$
- D_a, definition, nature $= \{x \in \mathbb{R}^2 \quad C^E(x) = 0, \ C^I(x) \leq 0_{\mathbb{R}^2}\}$

Two classes of methods

- Deterministic methods: use of the regularity properties of the objective function, descent methods
  - advantages: speed
  - disadvantage: possibility of being trapped near a local minimum
- Stochastic methods: random exploration of the search domain, no need for regularity
  - advantages: robustness, global minimum
  - disadvantage: slowness

Continuous Optimization Methods

- Least Squares
- Linear Programming
- Convex minimization
- Role of regularity

Least squares

$$\text{Minimize } \|Ax - b\|^2, \quad A \in \mathbb{R}^{m \times n} \quad x \in \mathbb{R}^n, \quad b \in \mathbb{R}^m$$

- Explicit solution: $f(x) = \|Ax - b\|^2$, $\nabla f(x) = 2A^T(Ax - b)$, $x^\star = (A^TA)^{-1}A^Tb$
- Cost at most $O(n^2m)$
- Robust and generalized method

Linear Programming

$$\text{Minimize } a^T x, \quad a, x \in \mathbb{R}^n,$$
$$\text{under constraints } b_i^T x \leq c_i,$$
$$b_i \in \mathbb{R}^n, \quad c_i \in \mathbb{R}, \text{ for } i = 1, \dots, m$$

- No explicit solution
- Robust and efficient algorithms (simplex, Karmarkar, ...)
- Cost at most in $O(n^2 m)$ if $m \geq n$

Convex optimization

$$\text{Minimize } f(x), \quad x \in \mathbb{R}^n,$$
$$\text{under constraints } g_i(x) \leq 0, \text{ for } i = 1, \ldots, m$$

- f and g_i convex:

  $$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), \text{ for } 0 \leq \alpha \leq 1$$

- Includes the two previous problems (least squares and linear)
- Cost: $O(\max(n^3, mn^2, c))$ c number of operations to calculate the gradients and hessians of f and g_i
- Robust and efficient algorithms
- Difficulty: identifying the convexity of the problem

Optimizing a differentiable function

$$\text{Minimize } f(x), \quad x \in \mathbb{R}^n$$

- f differentiable: $\nabla f(x^\star) = 0$
- f twice differentiable: we can solve $\nabla f(x^\star) = 0$ with Newton
- Two levels of difficulty: differentiability and calculation of the differential

Important notions on the "memory list"

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

2D

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that
$$\frac{\partial f}{\partial x}(x,y) = \frac{\partial f}{\partial y}(x,y) = 0$$

Why are the critical points important?

if $f(x^*, y^*) = \min\limits_{V(x^*,y^*)} f(x,y)$ ← open neighborhood of x^*, y^*

then (x^*, y^*) is a critical point

1D $f : \mathbb{R} \to \mathbb{R}$ $f(x^*) = \min\limits_{]x^*-\varepsilon,\, x^*+\varepsilon[} f(x) \Rightarrow f'(x^*) = 0$

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that $\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 0$
- Monge notation: let (x^*, y^*) be a critical point of f

$$r = \frac{\partial^2 f}{\partial x^2}(x^*, y^*), \quad s = \frac{\partial^2 f}{\partial x \partial y}(x^*, y^*), \quad t = \frac{\partial^2 f}{\partial y^2}(x^*, y^*)$$

$H_f(x^*, y^*) = \begin{pmatrix} r & s \\ s & t \end{pmatrix}$ hessian matrix of f

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that $\dfrac{\partial f}{\partial x}(x,y) = \dfrac{\partial f}{\partial y}(x,y) = 0$
- Monge notation: let (x^*, y^*) be a critical point of f

$$r = \frac{\partial^2 f}{\partial x^2}(x^*, y^*), \quad s = \frac{\partial^2 f}{\partial x \partial y}(x^*, y^*), \quad t = \frac{\partial^2 f}{\partial y^2}(x^*, y^*)$$

- Extrema: i.e. $\Delta = rt - s^2$ $= \det H_f(x^*, y^*)$

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that $\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 0$
- Monge notation: let (x^*, y^*) be a critical point of f

$$r = \frac{\partial^2 f}{\partial x^2}(x^*, y^*), \quad s = \frac{\partial^2 f}{\partial x \partial y}(x^*, y^*), \quad t = \frac{\partial^2 f}{\partial y^2}(x^*, y^*)$$

- Extrema: i.e. $\Delta = rt - s^2$
  - If $\Delta > 0$ then (x^*, y^*) is $\begin{cases} \text{a maximum if } r < 0 \\ \text{a minimum if } r > 0 \end{cases}$

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that $\dfrac{\partial f}{\partial x}(x, y) = \dfrac{\partial f}{\partial y}(x, y) = 0$
- Monge notation: let (x^*, y^*) be a critical point of f

$$r = \frac{\partial^2 f}{\partial x^2}(x^*, y^*), \quad s = \frac{\partial^2 f}{\partial x \partial y}(x^*, y^*), \quad t = \frac{\partial^2 f}{\partial y^2}(x^*, y^*)$$

- Extrema: i.e. $\Delta = rt - s^2$
  - If $\Delta > 0$ then (x^*, y^*) is $\begin{cases} \text{a maximum if } r < 0 \\ \text{a minimum if } r > 0 \end{cases}$
  - If $\Delta < 0$ then (x^*, y^*) is not an extremum

Reminders for the functions of $\mathbb{R}^2$ in $\mathbb{R}$

Let $f : \mathbb{R}^2 \to \mathbb{R}$ in $\mathcal{C}^2$

- The critical points of f are the points such that $\dfrac{\partial f}{\partial x}(x, y) = \dfrac{\partial f}{\partial y}(x, y) = 0$
- Monge notation: let (x^*, y^*) be a critical point of f

$$r = \frac{\partial^2 f}{\partial x^2}(x^*, y^*), \quad s = \frac{\partial^2 f}{\partial x \partial y}(x^*, y^*), \quad t = \frac{\partial^2 f}{\partial y^2}(x^*, y^*)$$

- Extrema: i.e. $\Delta = rt - s^2$
  - If $\Delta > 0$ then (x^*, y^*) is $\begin{cases} \text{a maximum if } r < 0 \\ \text{a minimum if } r > 0 \end{cases}$
  - If $\Delta < 0$ then (x^*, y^*) is not an extremum
  - If $\Delta = 0$ nothing can be said

Definition of "isovalues" or "contour lines" or "level curves" = set of points

1 $f^{-1}(-1) = \emptyset$ $f^{-1}(c) = \emptyset$ for $c < 0$
$f^{-1}(0) = \{(0, 0)\}$ $f^{-1}(c)$ for $c > 0$
circle of center s and radius $\sqrt{c}$

For $f : \mathbb{R}^n \to \mathbb{R}$, $c \in \mathbb{R}$

$$f^{-1}(c) = \{x \in \mathbb{R}^n,\ f(x) = c\}$$

Exemples

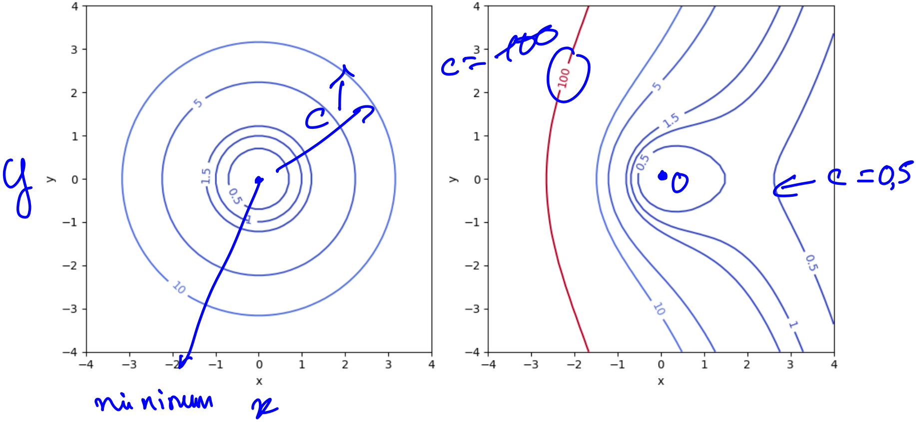
1 ▶ $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) = \|x - s\|^2 = (x_1 - s_1)^2 + (x_2 - s_2)^2$
▶ $f : \mathbb{R}^2 \to \mathbb{R}$, $a \in \mathbb{R}^{\cancel{n}\,2}$, $b \in \mathbb{R}$, $f(x) = a \cdot x + b$

2 $f^{-1}(c) = \{ a_1 x_1 + b_1 x_2 + b = c \}$ line
parallel lines

Example: $f(x, y) = (x^2 + y^2)e^{-x}$

$f(x,y) = x^2 + y^2$

$f(x,y) = (x^2 + y^2)e^{-x}$



Example: $f(x, y) = (x^2 + y^2)e^{-x}$

Experiment numerically with notebook 0

Numerical approximation of derivatives and gradient

Finite differences

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is not known explicitly with a formula. Computing its gradient can be achieved numerically using Taylor formula.

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

For $f : \mathbb{R} \to \mathbb{R}$, C^2

$$f(x+h) = f(x) + f'(x)h + o(h)$$ Taylor

finite difference ratio

$$\boxed{\frac{f(x+h) - f(x)}{h}} = f'(x) + o(1).$$

true value

error

1st order approximation

Or, for h small enough

$$\boxed{\frac{f(x+h) - f(x)}{h} \approx f'(x).}$$

Numerical approximation of derivatives

Better approximation using

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + o(h^2),$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) + o(h^2).$$

Substracting (2) from (1) provides

$$\frac{f(x+h) - f(x-h)}{2h} = f'(x) + o(h)$$

Or, for h small enough

$$\boxed{\frac{f(x+h) - f(x-h)}{2h} \approx f'(x).}$$

Why is it a better approximation than $\boxed{\dfrac{f(x+h) - f(x)}{h}}$?

Numerical approximation of derivatives

$$f(x+h) = f(x) + h f'(x) + \frac{h^2}{2} f''(x) + o(h^2)$$

$$f(x-h) = f(x) - h f'(x) + \frac{h^2}{2} f''(x) + o(h^2)$$

$$f(x+h) - f(x-h) = 2h f'(x) + o(h^2)$$

$$\frac{f(x+h)-f(x-h)}{2h} = f'(x) + o(h)$$

2nd order approximation

notebook 1

Can we do better using higher order Taylor expansions ?

Find a 4^th order approximation of $f'(x)$ using expansions like

$$f(x+h)=f(x)+hf'(x)+\frac{h^2}{2}f''(x)+\frac{h^3}{6}f^{(3)}(x)+\frac{h^4}{24}f^{(4)}(x)+o(h^4)$$

Expand...

$$f'(x) + o(h^3) = \frac{a f(x+h) + b f(x-h) + c f(x+2h) + d f(x-2h)}{h}$$

$$a \quad f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f^{(3)}(x) + \frac{h^4}{24} f^{(4)}(x) + o(h^4)$$

$$b \quad f(x-h) =$$

$$c \quad f(x+2h) =$$

$$d \quad f(x-2h) =$$

Choose correct coefficients before summing

$$\frac{8f(x+h) - 8f(x-h) - f(x+2h) + f(x-2h)}{12h} =$$

$$f'(x) + o(h^3)$$

Numerical approximation of gradient

to approximate $\nabla f(x)$ $f: \mathbb{R}^n \to \mathbb{R}$ you need to evaluate f at $n+1$ points in $\mathbb{R}^n$

And if $f : \mathbb{R}^n \to \mathbb{R}$?

$$\frac{\partial f}{\partial x_i}(x) = \lim_{h \to 0} \frac{f(x + he_i) - f(x)}{h}$$

with e_i the canonical basis vector $(e_i)_j = \delta_{i,j} = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$

Therefore $\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + he_i) - f(x)}{h}$ or

$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + he_i) - f(x - he_i)}{2h}$ for h small enough

Experiment numerically with notebook 1

Outline

Optimality conditions in the unconstrained case

Find extrema of a function defined on E.

Find

$$\inf_{x \in E} f(x),$$

with

$$f : E \longrightarrow \mathbb{R}$$

E normed finite dimension vector space

$E = \mathbb{R}^n$

Necessary optimality conditions

if $f(x^*) = \min\limits_{\mathbb{R} \atop I} f(x)$ then $f'(x) = 0$

I open interval of $\mathbb{R}$

Let $x^\star$ local minimum of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$.

1. First order optimality condition : if f is differentiable on an open neighborhood V of $x^\star$, then $\nabla f(x^\star) = 0$
2. Second order optimality condition : moreover if f is twice differentiable on V , then H$f(x^\star)$ is positive semi definite and f is locally convex in $x^\star$.

scalar case: if $f \in C^2$ $f'(x^*) = 0$ if x^* is a minimum then $f''(x^*) \geq 0$

$$f(x+th) = f(x) + Df(x)th + o(\|th\|)$$
$$= f(x) + \nabla f(x).th + o(\|th\|)$$

$$f(x^*) = \min_{x \in V(x^*)} f(x)$$

$t \in \mathbb{R}$

$$f(x+th) = f(x) + t\nabla f(x).h + \|h\| o(t)$$

$x + th \in V(x^*)$ for t small enough

$\boxed{t>0}$

$$f(x+th) = f(x) + t\nabla f(x).h + \|h\| o(t)$$
$$f(x+th) - f(x) = t\nabla f(x).h + \|h\| o(t)$$
$$0 \le \frac{f(x+th) - f(x)}{t} = \nabla f(x).h + \|h\| \frac{o(t)}{t} \quad t \to 0$$

$\frac{o(t)}{t} \to 0$

$$\nabla f(x).h \ge 0$$

$t<0 \Rightarrow \nabla f(x).h \le 0$) $\forall h \Rightarrow \nabla f(x) = 0$

Examples

$f(x) = x^4$ $\qquad 0 = \min f(x)$

$f'(0) = 0$

$f''(0) = 12x^2 = 0$

Counter example : $f(x) = x^4$.
Counter example : $f(x) = x^3$.

$f(x) = x^3$

$f'(0) = 0$

$f''(0) = 6x = 0$

but 0 is not a minimum

Sufficient optimality conditions

If f is twice differentiable in $x^\star$, and if $\nabla f(x^\star) = 0$ and moreover

- either $Hf(x^\star)$ is positive definite
- either f is twice differentiable in a neighborhood V of $x^\star$ and $Hf(x)$ is positive semi definite on V $\longrightarrow$ $f(x) = x^4$

then $x^\star$ is a strict (isolated) minimizer of f on V.

$f'(0) = 0$

$f''(x) = 12x^2 \geq 0$

Uniqueness condition in the convex case

Proof by contraposée

(i) If f is convex on a convex subset $C \in \mathbb{R}^n$, any local minimum of f on C is global.
(ii) If f is strictly convex it has at most one global minimum.

Necessary and sufficient optimality condition in the convex case

If f is convex on $\mathbb{R}^n$ and C^1, $x^\star \in \mathbb{R}^n$ realizes a global minimum of f if and only if $\nabla f(x^\star) = 0$.

$$f(x) = \frac{1}{2} Ax \cdot x + b \cdot x + c$$

$x \in \mathbb{R}^n$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$, $A \in S^n(\mathbb{R})$

$\nabla f(x) = Ax + b$

$\neq$ $\nabla f(x) = \frac{1}{2}(A + A^T)x + b$

$H f(x) = A$

$Ax + b = 0_{\mathbb{R}^n}$

1) if $A \in S^n_{++}(\mathbb{R})$ then

$\exists !\, x^*$ $Ax^* = b$

$x^* = -A^{-1} b$

2) if $A \notin S^n_+(\mathbb{R})$

$\hookrightarrow f(x) \to -\infty$

for at least one V_A

$f''(x) = a$

$f(x) = \frac{1}{2} a x^2 + bx + c$ $\qquad$ $f'(x) = ax + b$

Optimality condition for quadratic problems

for $A \in S^n(\mathbb{R})$, $\lambda_1 \leq \dots \leq \lambda_n$ eigenvalues in $\mathbb{R}$

$\lambda_1 \|x\|^2 \leq Ax.x \leq \lambda_n \|x\|^2$

$$f(x) = \frac{1}{2}x^t Ax + b^t x + c = \frac{1}{2}(Ax) \cdot x + b \cdot x + c$$

with $A \in S_n^{++}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$.

$$\inf_{x \in \mathbb{R}^n} f(x) \tag{1}$$

- If A is not positive semi definite then the problem (1) has no solution : no $x \in \mathbb{R}^n$ realizes a local minimum.
- If A is positive definite then $x^\star = -A^{-1}b$ is the only global minimum.

Solving systems of non linear equations

if $f(x^*) = \min_{x \in V(x^*)} f(x) \Rightarrow \nabla f(x^*) = 0$

$$\boxed{\nabla f(x) = 0_{\mathbb{R}^n}}$$

$\nabla f(x) \in \mathbb{R}^n$

$x \in \mathbb{R}^n$

$g(x) = 0$

$g : \mathbb{R}^n \to \mathbb{R}^n$

set of n equations with n unknowns

Fixed point method

$\boxed{g(x) = 0}$

$Id - \varphi = g$

$\boxed{x = \varphi(x)}$; $\boxed{\varphi = Id - g}$

Definition

A fixed point $x \in \mathbb{R}^N$ of a function $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is a point such that $\boxed{x = \varphi(x)}$

Definition

A fixed point $x \in \mathbb{R}^N$ of a function $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is said to be attractive if there exists a neighborhood V of x such that for all x_0 in V, the sequence defined by $x_{n+1} = \varphi(x_n)$ converges to x. Otherwise, the point is said to be repulsive.

$x_0, \quad x_{n+1} = \varphi(x_n)$

$\longrightarrow \quad x_n \xrightarrow[n \to +\infty]{} x$

Picard's Theorem

Theorem (Picard's Theorem)

Let F be a closed ~~state~~ set of $\mathbb{R}^N$ and let $\varphi : F \subset \mathbb{R}^N \to \mathbb{R}^N$ be a map such that $\varphi(F) \subset F$. We assume that φ is contracting, i.e. there exists $k \in]0, 1[$ such that:

$$\forall x, y \in F, \qquad \|\varphi(x) - \varphi(y)\| \leq k\|x - y\|. \tag{2}$$

Then there exists a unique $x^* \in F \subset \mathbb{R}^N$ such that $\boxed{\varphi(x^*) = x^*}$ and, for all $x_0 \in F$, the sequence defined by $x_{n+1} = \varphi(x_n)$ converges to x^* (i.e. $x^\star$ is an attractive fixed point). Furthermore, there exists a constant C (depending on the choice of x_0 and the function φ) such that

$$e_n := \|x_n - x^*\| \leq Ck^n. \tag{3}$$

Fixed point algorithm

$g(x) = 0$

Simplest method

```
def g(x):
    y = g(x)
    return y

def phi(x):
    return x - g(x)
```

choose x_0, $x_1 = \mathrm{phi}(x_0)$, ε, n_{max}

while $\|x_1 - x_0\| \geq \varepsilon$ and $\underline{n \leq n_{max}}$

$x_{n+1} = \varphi(x_n)$

$n = n+1$

$g(x) = 0$, scalar case

Newton method

Zeros of function $g : \mathbb{R} \longrightarrow \mathbb{R}$

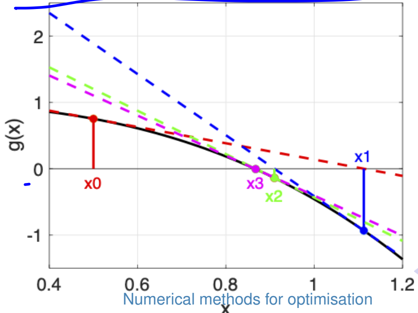
$$g(x^\star) = g(x) + g'(x)(x^\star - x) + o(\|x^\star - x\|).$$

Fixed point algorithm to solve a nonlinear equation

$$\psi(x) = x, \text{ with } \psi(x) = x - g(x)/g'(x).$$

Approximation by a sequence x_n

$$\boxed{x_{n+1} = x_n - g(x_n)/g'(x_n)}$$



$$\psi(x) = x - \frac{g(x)}{g'(x)}$$

$$\psi'(x) = 1 - \frac{g'(x)}{g'(x)} + \frac{g''(x)\,g(x)}{g'(x)^2}$$

$$= \frac{g''(x)\,g(x)}{g'(x)^2}$$

at the solution of $g(x^*) = 0$

$$\psi'(x^*) = 0$$

by continuity ψ is contracting in a neighborhood of x^*.

Newton algorithm (scalar case)

Data: Function $g(x)$ derivative $g'(x)$, tolerance ε, max number of iterations $k_{\max}$
Result: $x^\star$ such that $g(x^\star) = 0$
Initialisation : $k = 0$, x_0 initial guess for $g(x_0) = 0$.
while $|g(x_k)| > \varepsilon$ *and* $\boxed{k \leq k_{\max}}$ **do**

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

$$k \leftarrow k + 1$$

end
$x^\star \leftarrow x_k$

if x_0 is close enough to x^* it converges very fast
if not it may diverge

Convergence of Newton algorithm

Let g in C^2 on $I = [x^\star - r, x^\star + r]$ with $g(x^\star) = 0$ and $g' \neq 0$ on I. Let

$$M = \max_{x \in I} \left| \frac{g''(x)}{g'(x)} \right|, \quad \text{and } h = \min\left(r, \frac{1}{M}\right).$$

Then for any $x_0 \in]x^\star - h, x^\star + h[$ we have

$$e_k := |x_k - x^\star| \leq \frac{1}{M} \left(M|x_0 - x^\star|\right)^{2k},$$

from which we deduce $\lim_{k\to+\infty} |x_k - x^\star| = 0$.

Convergence speed

for any iterative algorithm $x_0, (x_k)_k$

Denote by $e_k = x^k - x^\star$ the error at iteration k. We say that

- the algorithm converges if $\lim_{k\to\infty} \|e_k\| = 0$
- the algorithm converges linearly if $c \in]0,1[$ tel que $\|e_k\| \leq c\|e_{k-1}\|$ for $k > K(c)$
- the algorithm converges supra-linearly if $(c_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} c_k = 0$ such that $\|e_k\| \leq c_k\|e_{k-1}\|$
- the algorithm converges geometrically if the sequence c_k is geometric
- the algorithm~~e~~ is of order p if there exists $c \in]0,1[$ such that $\|e_k\| \leq c\|e_{k-1}\|^p$ for $k > K(c)$

  $p = 2$

The co~~n~~vergence can be global or local

Newton method in dimension n

$g'(x) \Longleftrightarrow Jg(x)$

$G(x) = 0$ with $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$.

$$\boxed{x_{n+1} = x_n - \frac{g(x_n)}{g'(x_n)}}$$

Scalar case

Let $JG(x) \in \mathbb{R}^{n \times n}$ the jacobian matrix of G in x,

$$JG_{i,j}(x) = \frac{\partial G_i(x)}{\partial x_j}.$$

$$G(x^\star) = G(x) + JG(x)(x - x^\star) + o(||x - x^\star)||.$$

Fixed point algorithm to solve a nonlinear equation

$$\Psi(X) = X, \text{ with } \Psi(\cancel{X}) = X - JG(X)^{-1}G(X).$$

Approximation by a fixed point sequence $X_{n+1} = \Psi(X_n)$

X_0 ;

$$X_{n+1} = X_n - JG(X_n)^{-1}G(X_n)$$

Except that in practice for n large, one never computes the inverse of a matrix

Newton-Ralphson algorithm

extension to $\mathbb{R}^n$

$x_{n+1} = x_n - JG(x_n)^{-1} G(x_n)$

Data: Function $G(x)$, jacobian matrix $JG(x)$, tolerance ε,
max number of iterations $k_{\max}$

Result: $x^\star$ such that $G(x^\star) = 0$

Initialisation : $k = 0$, x_0

while $\|G(x_k)\| > \varepsilon$ *and* $k \leq k_{\max}$ **do**

  Solve $JG(x_k)d_k = -G(x_k)$ do not compute $JG(x_k)^{-1}$

  $x_{k+1} = x_k + d_k$

  $k \leftarrow k + 1$

end

$x^\star \leftarrow x_k$

Convergence of Newton-Ralphson algorithm

Suppose :

- G of class C^2
- $G(x^\star) \cancel{\neq} 0$
- the tangent linear map $JG(x^\star) \in \mathcal{L}(\mathbb{R}^m, \mathbb{R}^m)$ is inversible.

Jacobian matrix

Then $x^\star$ is a superattractive fixed point of

$$\Psi(x) = x - (JG(x))^{-1} G(x).$$

Local result.

Scalar case : the secant method

in the case where $g'(x)$ is not accessible

Data: Function $g(x)$, tolerance ε, max number of iterations $k_{\max}$

Result: $x^\star$ such that $g(x^\star) = 0$

Initialisation : $k = 0$, x_0 initial guess for $g(x) = 0$.

a_0 initial guess for $g'(x_0)$ (default =1)

while $|g(x_k)| > \varepsilon$ *and* $k < k_{\max}$ **do**

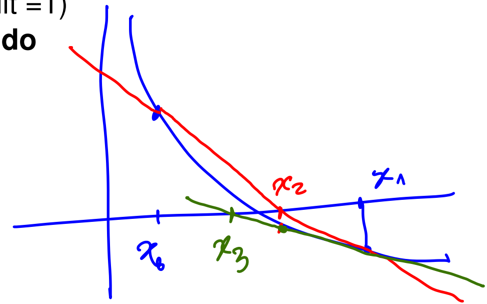
$$x_{k+1} = x_k - \frac{g(x_k)}{a_k}$$

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

$$k \leftarrow k + 1$$

end

$x^\star \leftarrow x_k$



Vector case : the quasi-Newton method

Data: $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$
$\varepsilon > 0$.
Result: $x^\star$ such that $g(x^\star) = 0$
Initialisation : a first approximation of $x_0 \in \mathbb{R}^n$
$A_0 \approx J(x_0)$ or $W_0 \approx J(x_0)^{-1}$
$x_1 = x_0 - W_0 G(x_0)$
$d_0 = x_1 - x_0$,
$y_0 = G(x_1) - G(x_0)$,
$k = 1$
while $\|G(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**
  Update : $W_k = W_{k-1} + B_{k-1}$
  Compute d_k solution of $d_k = -W_k G(x_k)$
  $x_{k+1} = x_k + d_k$
  $y_k = G(x_{k+1}) - G(x_k)$
  $k \leftarrow k + 1$
end
$x^\star \leftarrow x_k$

$JG(x_k)\, d_k = -G(x_k)$
$x_{k+1} = x_k + d_k$

$A_k d_k = -G(x_k)$
$x_{k+1} = x_k + d_k$
$A_{k+1} = \mathcal{F}(A_k)$

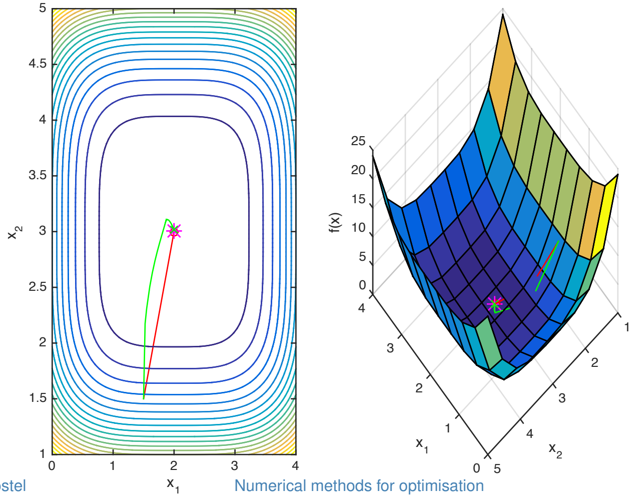
$x_{k+1} = x_k - W_k G(x_k)$
$W_{k+1} = \mathcal{F}(W_k)$
$\vdots$

Comparison of Newton and quasi Newton methods

Minimum of the quadratic function $f(x) = ((x_1 - 2)^4 + (x_2 - 3)^4)/2$

Newton method : 12 iterations

quasi Newton (BFGS) method : 21 iterations



Update in the quasi Newton method

$W_k \simeq JG(x_k)^{-1}$

Update :

$W_k = W_{k-1} + B_{k-1}$

$W_k = \mathcal{F}(W_{k-1})$

Compute d_k solution of $d_k = -W_k G(x_k)$

$x_{k+1} = x_k + d_k$

Conditions on the W_k matrix

1. W_k should remain symetric positive definite for all k.
2. The quasi-Newton equation $W_k y_k = d_k$ is satisfied for each k
3. The difference between two consecutive approximations $W_{k+1} - W_k$ is minimum in some sense (for some norm), for example for the Frobenius norm $\|A\| = \sum_i \sum_j |A_{ij}|$

Examples of update rules satisfying the conditions

$$W_k = W_{k-1} + B_{k-1}$$
$$d_k = -W_k G(x_k)$$
$$x_{k+1} = x_k + d_k$$
$$y_k = G(x_{k+1}) - G(x_k)$$

- The Davidon-Fletcher-Powell method

$$(DFP) \quad W_{k+1} = W_k + \frac{d_k d_k^T}{y_k \cdot d_k} - \frac{W_k y_k y_k^T W_k}{(y_k \cdot W_k y_k)}.$$

- The Broyden-Fletcher-Goldfarb-Shanno method

$$\boxed{(BFGS)} \quad W_{k+1} = W_k - \frac{d_k y_k^T W_k + W_k y_k d_k^T}{y_k \cdot d_k}$$
$$+ \left(1 + \frac{y_k \cdot W_k y_k}{y_k \cdot d_k}\right) \frac{d_k d_k^T}{y_k \cdot d_k}.$$

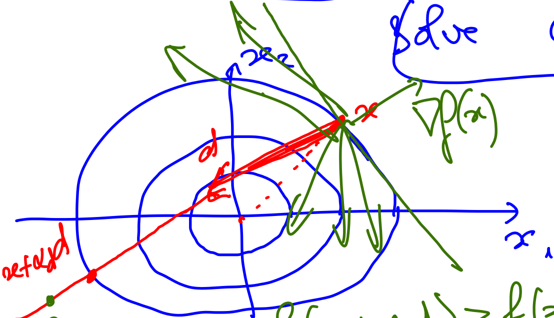
Experiment with notebook 2

$f(x^*) = \min f(x) \Rightarrow \nabla f(x) = 0$

1st set of methods

solve $g(x) = 0_{\mathbb{R}^n}$

$f(x) = \|x\|^2$



for $\alpha > \alpha_d$ $\quad f(x+\alpha d) > f(x)$

Descent method

build sequence $(x_n)_n$
such that $f(x_{n+1}) \leq f(x_n)$

Definition : We say that $d \in \mathbb{R}^n$ is a descent direction at x for the function f if there exists $\alpha_d > 0$ such that

$$f(x + \alpha d) < f(x) \quad \boxed{\forall 0 < \alpha \leq} \alpha_d.$$

when you cross the level curve $f(x)$ again

Property : If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable , $d \in \mathbb{R}^n$ is a descent direction at x if and only if

$$\nabla f(x) \cdot d < 0.$$

$$f(x + \alpha d) = f(x) + \alpha \underbrace{\nabla f(x) . d}_{<0} + \|d\| o(|\alpha|)$$

$$\underbrace{\frac{f(x+\alpha d) - f(x)}{\alpha}}_{<0} = \underbrace{\nabla f(x) . d + o(|\alpha|)/\alpha}_{<0 \text{ for } \alpha \text{ small enough}}$$

Example $f(x) = \frac{1}{2} Ax.x + b.x + c$

necessary condition to have a solution $\min f(x)$

$A \in S_n^+$; $\|\nabla f(x) - \nabla f(y)\| = \|A(x-y)\|$

$\|A\|.\|x-y\|$

sufficient: $A \in S_n^{++}$ $Ax^* = -b$ $L = \lambda_n$

eigenvalues: $0 < \lambda_1 \leq \dots \leq \lambda_n$

$$\underline{\lambda_1 \|x\|^2 \leq Ax.x}$$

f is strongly convex or λ_1-elliptic

$$Hf(x)d.d \geq \lambda_1 \|d\|^2$$

the gradient method will converge if we choose

$$0 < \alpha < \frac{2\lambda_1}{\lambda_n^2}$$

General descent algorithm

(x_k) $f(x_{k+1}) < f(x_k)$

Data: function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
required precision $\varepsilon > 0$.
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for the solution $x_0 \in \mathbb{R}^n$
while $\|\nabla f(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**
  Choose descent direction d_k (such that $\nabla f(x_k) \cdot d_k < 0$)
  Choose step α_k in direction d_k, such that
   $f(x_k + \alpha_k d_k) \le f(x_k)$
  $x_{k+1} = x_k + \alpha_k d_k$
  $k \leftarrow k + 1$
end
$x^\star \leftarrow x_k$

Convergence of a descent algorithm

Sufficient conditions

Let f a function C^1 from $\mathbb{R}^n$ into $\mathbb{R}$ and $x^\star$ minimizer of f. If the following conditions are satisfied

1. f α-elliptic

$Hf(x)d \cdot d \geq \alpha \|d\|^2$ 2nd ord

2. ∇f L-lipschitz

$(\nabla f(x) - \nabla f(y)) \cdot (x-y) \geq \alpha \|x-y\|^2$ 1st order

Then for all $(\alpha_k)_{k\in\mathbb{N}}$ sequence such that there exist $a, b \in \mathbb{R}$, s.t.

$\|\nabla f(x) - \nabla f(y)\|^2 \leq L\|x-y\|^2$

$$0 < a \leq \alpha_k \leq b < \frac{2\alpha}{L^2}, \quad \forall k \in \mathbb{N},$$

$d_k = -\nabla f(x_k)$

$d_k \cdot \nabla f(x_k) = -\|\nabla f(x_k)\|^2 < 0$ for $x_k \neq x^*$

The gradient method

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

converges geometrically for all initial guess, i.e.

$$\exists \beta \in]0,1[, \|x_k - x^\star\| \leq \beta^k \|x_0 - x^\star\|$$

$$e_{k+1} = x_{k+1} - x^* \quad \text{with} \quad \nabla f(x^*) = 0$$

$$\|x_{k+1} - x^*\|^2 = \| x_k - \alpha_k \nabla f(x_k) - x^* \|^2$$

$$= \| x_k - x^* - \alpha_k (\nabla f(x_k) - \nabla f(x^*)) \|^2$$

$$= \|x_k - x^*\|^2 - 2\alpha_k \underbrace{(x_k - x^*)\cdot(\nabla f(x_k) - \nabla f(x^*))}_{\geq \alpha \|x_k - x^*\|^2} + \alpha_k^2 \underbrace{\|\nabla f(x_k) - \nabla f(x^*)\|^2}_{\leq L^2 \|x_k - x^*\|^2}$$

$$\|e_{k+1}\|^2 \leq \|x_k - x^*\|^2 \underbrace{(1 - 2\alpha_k \alpha + L^2 \alpha_k^2)}_{< 1}$$

choose $L^2 \alpha_k^2 - 2\alpha_k \alpha < 0$

$$\alpha_k (L^2 \alpha_k - 2\alpha) < 0$$

if $\alpha_k < \dfrac{2\alpha}{L^2}$

Examples of possible choices for the descent direction

- Gradient Algorithm (*steepest descent*)

$$d_k = -\nabla f(x_k).$$

- Newton algorithm based on direction

$$d_k = \boxed{-Hf(x_k)^{-1}\nabla f(x_k).}$$

$\alpha_k = 1$

- Quasi-Newton with

$$d_k = -W_k \nabla f(x_k),$$

where $W_k \approx Hf(x_k)$

only for $f(x) = \frac{1}{2}Ax.x + b.x + c$

- Conjugate gradient method (in the quadratic case)

$x_{k+1} = x_k + \alpha_k d_k$

$$d_k = \begin{cases} -\nabla f(x_1) & \text{for } k = 1 \\ -\nabla f(x_k) + \beta_k d_{k-1} & \text{for } k > 1. \end{cases}$$

Choice of the step in a given direction

at iteration k

$$h_k : \alpha \mapsto h_k(\alpha) = f_{\cancel{k}}(x_k + \alpha d_k)$$

Cauchy's rule

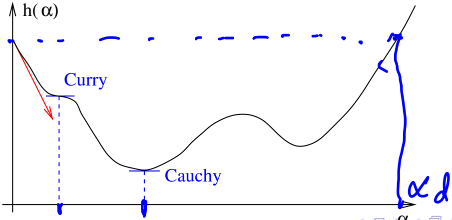
$h_k'(\alpha) = d_k \cdot \nabla f(x_k + \alpha d_k) = 0$!!)

$$\alpha_k = \operatorname*{argmin}_{\alpha>0} h_k(\alpha)$$

difficult.

Curry's rule

$$\alpha_k = \inf \{\alpha > 0; h_k'(\alpha) = 0, h_k(\alpha) < h_k(0)\}$$



Convergence of optimal step gradient

Suppose $\nabla f(x)$ L-Lipschitz on $\{x, f(x) \leq f(x^0)\}$.
Then the gradient algorithm with:

- $d_k = -\nabla f(x^k)$
- α_k fixed by Curry's rule

satisfies

- either $f(x^k)$ non-bounded below
- either $\nabla f(x^k) \to 0$ when $k \to \infty$.

$$f(x) = \frac{1}{2} Ax \cdot x + b \cdot x + c \qquad A \in S_n^{++}$$

$$h(\alpha) = f(x + \alpha d)$$

$$h'(\alpha) = d \cdot \nabla f(x + \alpha d)$$

$$= d \cdot \left(A(x + \alpha d) + b \right)$$

$$= d \cdot (Ax + b) + \alpha\, d \cdot Ad$$

$$h'(\alpha) = 0 \iff \alpha = \frac{d \cdot \nabla f(x)}{d \cdot Ad}$$ better optimal

optimal step in the quadratic case

$$0 < \alpha < \frac{2 \lambda_1}{\lambda_n^2}$$ sufficient

Optimal step in the quadratic case

$$f(x) = \frac{1}{2}Ax \cdot x + b \cdot x$$

with $A \in S^n$ and $b \in \mathbb{R}^n$,

$$h(\alpha) = f(x + \alpha d) = f(x) + \frac{\alpha^2}{2}Ad \cdot d + \alpha(Ax + b) \cdot d$$

$\alpha^\star = \operatorname{argmin} h(\alpha)$

$$\alpha^\star = -\frac{g \cdot d}{(Ad) \cdot d}$$

with $g = \nabla f(x) = Ax + b$.

Optimal step gradient method in the quadratic case

Data: A, b, ε
Result: $x^\star$ s.t. $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$, $x_0 \in \mathbb{R}^n$
$g_0 = Ax_0 + b$
while $\|g_k\| > \varepsilon$ *and* $k < k_{\max}$ **do**

$$d_k = -g_k$$
$$v_k = Ad_k$$
$$\alpha_k = \frac{d_k \cdot d_k}{v_k \cdot d_k}$$
$$x_{k+1} = x_k + \alpha_k d_k$$
$$g_{k+1} = g_k + \alpha_k v_k$$
$$k \leftarrow k + 1$$

end
$x^\star \leftarrow x_k$

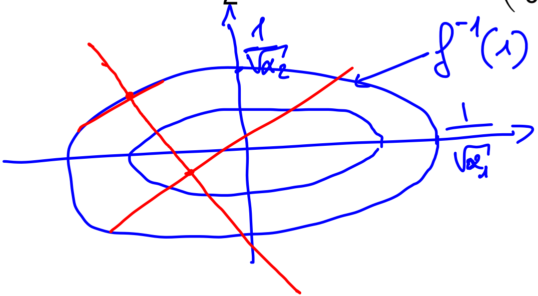
Conjugate gradient method : motivation

$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

with α_k optimal

will converge only asymptotically

$$
\begin{aligned}
f(x^*) &= \inf_{x \in \mathbb{R}^n} f(x) \\
f(x) &= \frac{1}{2}(\alpha_1 x_1^2 + \alpha_2 x^2), \quad \text{with } 0 < \alpha_1 < \alpha_2 \\
&= \frac{1}{2}(Ax) \cdot x \quad \text{with } A = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix}
\end{aligned}
$$



A-conjugate directions

if $f(x) = \frac{1}{2} Ax.x + b.x + c$

$Ax.x$ is a scalar product in the eigen vector basis

Definition : Let $A \in S^n_{++}$.

- 2 non zero vectors v, w are called $A-$conjugate iff $Av \cdot w = 0$.
- A family of non zero vectors $(v_i)_{i=1,\dots m}$, is called $A-$conjugate iff $Av_i \cdot v_j = 0$ for all $i = 1, \dots, m$, $j = 1, \dots, m$, $i \neq j$.

Property : $A-$conjugate vectors are linearly independent. If $m = n$ a $A-$conjugate family is a basis of $\mathbb{R}^n$.
Definition : a conjugate descent method is a method where the successive descent directions form a $A-$conjugate family

Expression of the minimum of f in a $A-$conjugate basis

$$f(x) = \frac{1}{2}Ax \cdot x + b \cdot x$$

Suppose we have a basis $(d_i)_{i=1,\dots n}$, such that $Ad_i \cdot d_j = 0$ for $j \neq i$

$$x^\star = \sum_{i=1}^{n} \alpha_i d_i, \text{ and } Ax^\star + b = 0,$$

$$\text{therefore } Ax^\star = -b = \sum_{i=1}^{n} \alpha_i A d_i, \text{ then for any } j = 1, \dots, n$$

$$-b \cdot d_j = \sum_{i=1}^{n} \alpha_i A d_i \cdot d_j = \alpha_j A d_j \cdot d_j$$

$$\alpha_j = \frac{-b \cdot d_j}{Ad_j \cdot d_j}$$

Construction of the $A-$conjugate basis

Let $g_k = \nabla f(x_k) = Ax_k + b$ be the gradient at step k
Choose $d_0 = -g_0$ (The first step is a standard gradient descent step)
Then $d_k = -g_{k-1} + \beta_{k-1} d_{k-1}$ satisfying:
(CG1) $Ad_k \cdot d_j = 0$ for $j = 0, \dots, k-1$ and
(CG2) $g_k \cdot d_j = 0$ for $j = 0, \dots, k-1$
Update at step k : $x_{k+1} = x_k + \alpha_k d_k$
Next gradient $g_{k+1} = Ax_{k+1} + b = g_k + \alpha_k Ad_k$
Property : For all initial guess x_0 there exists $(\alpha_k)_k$ and $(\beta_k)_k$ such that (CG1) and (CG2) are satisfied.
Property : (CG1) and (CG2) $\Rightarrow g_k \cdot g_j = 0$ for $j \neq k$

Convergence of a conjugate method

Property : A conjugate descent method using directions satisfying conditions (CG1) and (CG2) converges in at most n steps.

Property : $\beta_k = -\dfrac{Ad_{k-1} \cdot g_k}{Ad_{k-1} \cdot d_{k-1}} = \dfrac{\|g_k\|^2}{\|g_{k-1}\|^2}$. ← to update the direction

Property : $\alpha_k = -\dfrac{g^k \cdot d^k}{Ad^k \cdot d^k}$ ← same optimal step as before

$$d_{k+1} = -\nabla f(x_k) + \beta_k d_k$$

Conjugate gradient algorithm

optimal step
Gradient Method

Data: Matrix A, vector b, tolerance ε
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for solution $x^0 \in \mathbb{R}^n$
$g^0 = Ax^0 + b$
$d^0 = -g^0$
while $\|g^k\| > \varepsilon$ **do**

- Compute directionnal minimum :
  $v^k = Ad^k$
  $\alpha_k = -\dfrac{g^k \cdot d^k}{v^k \cdot d^k}$
  $x^{k+1} = x^k + \alpha_k d^k$
- Update gradient :
  $g^{k+1} = g^k + \alpha_k v^k$
- Compute new direction :
  $\beta_{k+1} = \dfrac{g^{k+1} \cdot g^{k+1}}{g^k \cdot g^k}$
  $d^{k+1} = -g^{k+1} + \beta_{k+1} d^k$

$k \leftarrow k + 1$
end
$x^\star \leftarrow x^k$

$v^k = Ad_k$
$\alpha_k = -\dfrac{g^k \cdot d^k}{v^k \cdot d^k}$
$x^{k+1} = x^k + \alpha_k d^k$
$g^{k+1} = g^k + \alpha_k v^k$

Monotonicity of the conjugate gradient algorithm

Property : If $d_k \neq 0$ and $\alpha_{k+1} \neq 0$ then $f(x_{k+1}) < f(x_k)$.
If $\alpha_{k+1} = 0$, x_k is the minimizer of f and $Ax_k + b = 0$

Polak-Ribière method

for f non quadratic

Data: Function f, gradient ∇f, tolerance ε
Result: $x^\star$ such that $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Initial guess for $x^0 \in \mathbb{R}^n$
$g^0 = \nabla f(x^0)$
$d^0 = -g^0$
while $\|g^k\| > \varepsilon$ *and* $k < k_{\max}$ **do**

- Compute the step in direction d_k:
  $f(x^k + \alpha_k d^k) \leq f(x^k + \alpha d^k) < f(x^k)$ for all $0 < \alpha \leq \alpha_k$
- Compute new position :
  $x^{k+1} = x^k + \alpha_k d^k$
- Compute new direction :
  $g^{k+1} = \nabla f(x^{k+1})$
  $$c_{k+1} = \frac{(g^{k+1} - g^k) \cdot g^{k+1}}{g^k \cdot g^k}$$
  $d^{k+1} = -g^{k+1} + c_{k+1} d^k$

$k \leftarrow k + 1$

end

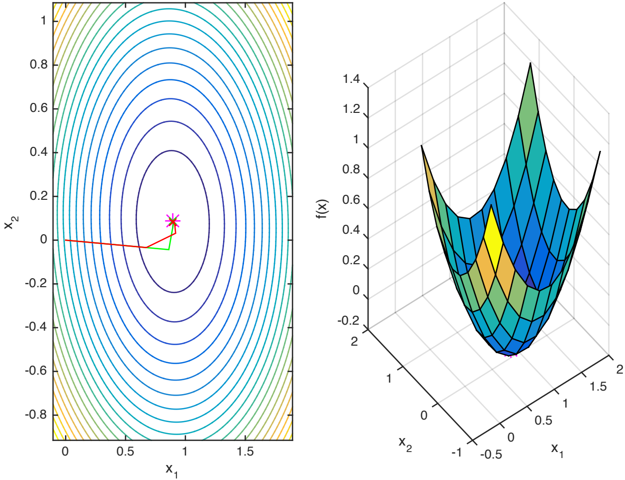
$x^\star \leftarrow x^k$

1) will not converge in a finite number of iteration

2) no explicit formula for α optimal

Comparaison of conjugate Gradient (green, 4 steps)and Polak-Ribière (red, 8 steps) methods.

f quadratic function in $\mathbb{R}^5$. Projection on $(0, x_1, x_2)$.



Choice of the step in the general case

Data: Function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
Required precision $\varepsilon > 0$.
Result: $x^\star$ s.t. $f(x^\star) = \min_x f(x)$
Initialisation : $k = 0$,
Intial guess $x_0 \in \mathbb{R}^n$
while $\|\nabla f(x_k)\| > \varepsilon$ *and* $k < k_{\max}$ **do**
    Choose d_k, s.t. $\nabla f(x_k) \cdot d_k < 0$)
    Choose step α_k in direction d_k, s.t. $f(x_k + \alpha_k d_k) \leq f(x_k)$
    $x_{k+1} = x_k + \alpha_k d_k$
    $k \leftarrow k + 1$
end
$x^\star \leftarrow x_k$

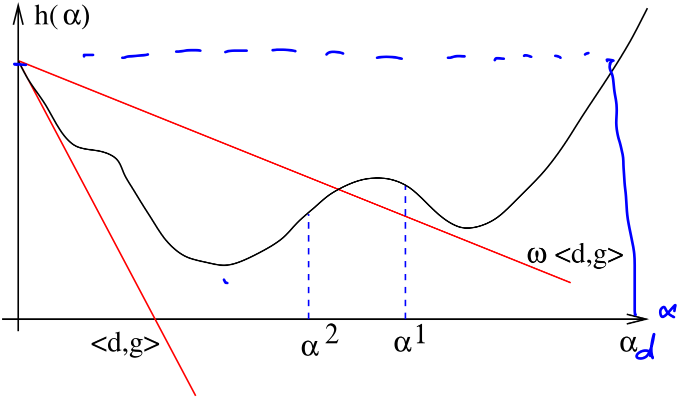
Directional minimisation -Line search

BLS
Backtracking
line search

Armijo's rule linearization of the constraint on α_k

$$f(x^k + \alpha_k d^k) < f(x^k) + \omega \alpha_k g^k \cdot d^k$$

$g = \nabla f(x)$



Armijo's rule

BLS

this converges unconditionally for any $d.\nabla f(x) < 0$

Data: Function f, current position x, descent direction d, coefficients $\tau \in]0,1[$ and $\omega \in]0,1[$
Result: α s.t. $f(x+\alpha d) < f(x)$
Initialisation : $k = 0$, initial guess α_0
while $f(x+\alpha_k d) > f(x) + \omega\alpha_k d\cdot\nabla f(x)$ **do**
  Choose $\alpha_{k+1} = \tau\alpha_k$
  $k \leftarrow k+1$
end
$\alpha = \alpha_k$

red line in the previous slide

Choice of the first value α_k^1

Assume a quadratic model for $\varphi(\alpha) = f(x^k + \alpha d^k)$

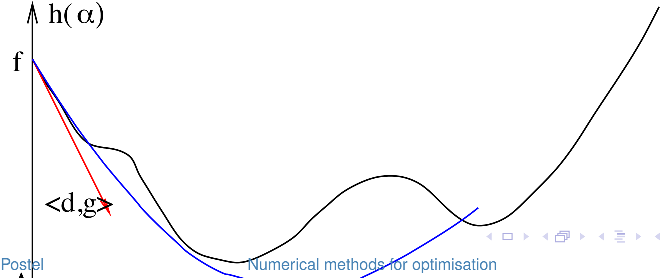
$$h(\alpha) = a_0 + a_1\alpha + a_2\alpha^2/2$$

with

$$\begin{cases} a_0 = f(x^k) \\ a_1 = d^k \cdot \nabla f(x^k) \end{cases}$$

a_2 is fixed by setting Δ, the maximal decrease of φ

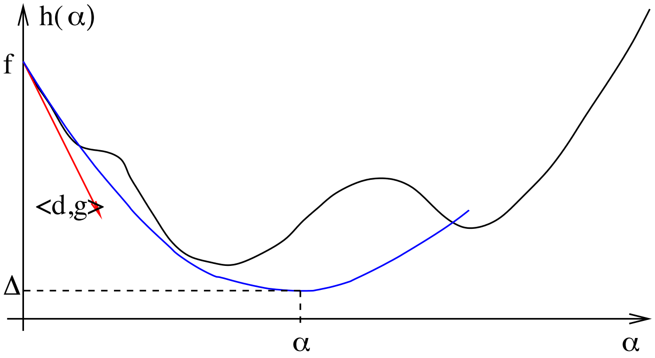
$$\Delta = \varphi(0) - \varphi_{\min} = a_1^2/(2a_2).$$



Fletcher's rule

α_k^1 is chosen to minimize the quadratic model

$$\alpha_k^1 = \frac{2\Delta}{d^k \cdot \nabla f(x^k)}$$



Convergence of gradient + Armijo methods

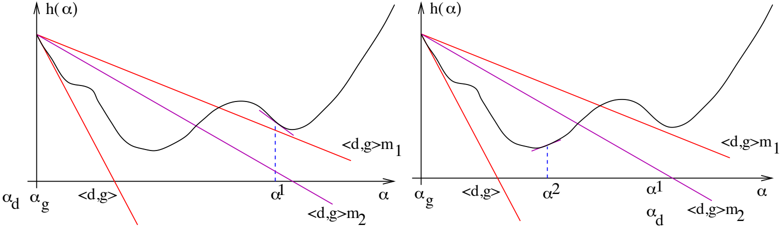
If $f : \mathbb{R}^n \to \mathbb{R}$ is C^1 and $\nabla f(x)$ is γ-lipschitz then Armijo's rule is satisfied for all

$$\alpha \in [0, \omega], \quad \text{with } \omega = \frac{(\omega_1 - 1)\nabla f(x) \cdot d}{\gamma \|d\|^2}.$$

Drawback of Armijo strategy: $\alpha_k^{i+1} < \alpha_k^i$, slow convergence.

Wolfe's method

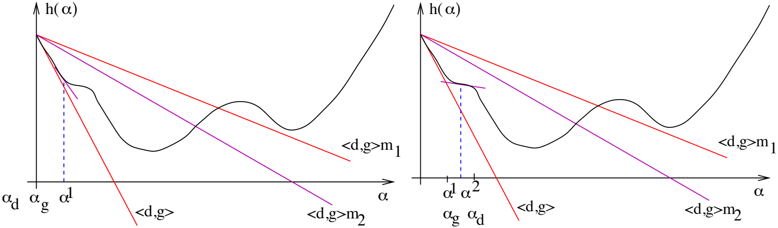
$$
\begin{array}{lrcl}
(A) & f(x^k + \alpha_k d^k) & < & f(x^k) + \omega_1 \alpha_k g^k \cdot d^k \\
(W) & \nabla f((x^k + \alpha_k d^k) \cdot d^k & > & \omega_2 g^k \cdot d^k, \text{ with } 0 < \omega_1 < \omega_2 < 1.
\end{array}
$$



Wolfe algorithm

Data: Function f, gradient ∇f, current point x^k, descente direction d^k, coefficients $0 < \omega_1 < \omega_2 < 1$
Result: α_k s.t. (A) and (W) are satisfied
Initialisation : Fix $\alpha_D = -1$ and $\alpha_G = 0$.
$p = 0$, Fix initial guess α_k^p (Fletcher's rule)
while $f(x^k + \alpha_k^p d^k) > f(x^k) + \omega_1 \alpha_k^p d^k \cdot \nabla f(x^k)$ *or* $\nabla f((x^k + \alpha_k^p d^k) \cdot d^k < \omega_2 g^k \cdot d^k$ **do**
    if $f(x^k + \alpha_k^p d^k) > f(x^k) + \omega_1 g^k \cdot d^k \alpha_k^p$ **then**
        $\alpha_D = \alpha_k^p$
    end
    else
        $\alpha_G = \alpha_k^p$
    end
    if $\alpha_D < 0$ *(not yet updated)* **then**
        $\alpha_k^{p+1} = 2\alpha_G$
    end
    else
        $\alpha_k^{p+1} = (\alpha_G + \alpha_D)/2$
    end
    $p \leftarrow p + 1$
end
$\alpha_k = \alpha_k^p$

Wolfe example



Convergence of Wolfe method

Let $f : \mathbb{R}^n \to \mathbb{R}$, bounded below, C^1 différentiable on

$$\mathcal{N} = \{f(x) \leq f(x_0)\}$$

and gradient $\nabla f(x)$ L-lipschitz. Then, if coefficients $(\alpha_k)_k$ satisfy conditions (A) and (W)

$$\sum_k \cos\theta_k^2 \|\nabla f(x^k)\|^2 < \infty, \quad \text{with } \cos\theta_k = \frac{-d^k \cdot \nabla f(x^k)}{\|d^k\|, \|\nabla f(x^k)\|}.$$

Experiment with notebook 5

X is rectangular

$$E(\theta) = \frac{1}{m} \| X\theta - y \|^2$$

$$= \frac{1}{m} (X\theta - y) \cdot (X\theta - y)$$

$$\nabla E(\theta) = \frac{2}{m} X^T (X\theta - y)$$

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix}$$

m lines × n columns

$$\nabla E(\theta^*) = 0_{\mathbb{R}^n}$$

$$X^T (X\theta^* - y) = 0$$

$$X^T X \theta^* = X^T y$$

$$\theta^* = (X^T X)^{-1} X^T y$$

No simplification

$X^T X$ invertible if n points at least are different among the m $\iff$ rank $X = n$

Linear regression

$$\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} \qquad \theta . x = \sum \theta_i x_i$$

Find θ defining a linear model

$$\hat{y} = h_\theta(x) = \theta^{\cancel{T}}.x$$

Let m measurements (x_i, y_i), $i = 1, \dots, m$, where explaining variables are in $\mathbb{R}^n$ ($x_i = (x_i^j)_{j=1,\dots,n}$. θ is found by minimizing the least squared error

$$E(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta^{\cancel{T}}.x_i - y_i \right)^2 = \frac{1}{m} \| X\theta - y \|^2$$

$$\nabla E(\theta) =$$

The "normal equation" gives the best solution

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_m^T \end{pmatrix} \qquad \hat{\theta} = (X^T.X)^{-1}.X^T.y \in$$

complexity in $O(n^3)$ and $O(m)$.

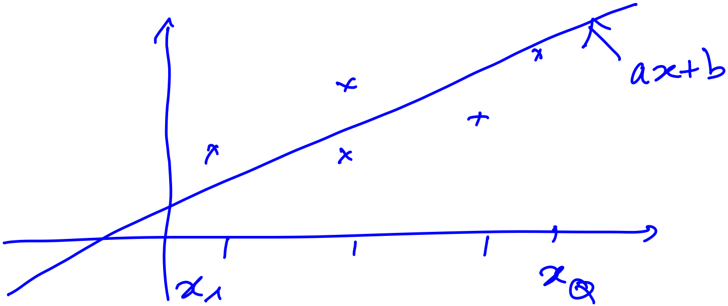
Nonlinear least squares

$$f : \begin{cases} \mathbb{R}^P & \to & \mathbb{R}^Q \\ x = (x_1, \dots, x_P)^t & \mapsto & (f_1(x), \dots, f_Q(x))^t \end{cases}$$

for $Q > P$ we seek a solution to the problem $f(x) = 0$.

Examples

- Find a line that passes through Q points with $Q > 2$



$$f : \mathbb{R}^{2=P} \longrightarrow \mathbb{R}^Q$$

$$z = \begin{pmatrix} a \\ b \end{pmatrix} \longmapsto \begin{pmatrix} a x_1 + b \\ \vdots \\ a x_Q + b \end{pmatrix} = \begin{pmatrix} f_1(z) \\ \vdots \\ f_Q(z) \end{pmatrix}$$

Examples

- Find the parameters N_0 and λ of a radioactive material whose emissions are monitored over time $N(t) = N_0 e^{-\lambda t}$

$\tilde{N}_1 = N(t_1)$ measurements $\tilde{N}_i$ at t_i

$\vdots \quad \vdots$

$\tilde{N}_q - t_q$

Toy example

(N_0, λ)

$f : \mathbb{R}^2 \to R^Q$ with Q large
$(N_i)_{i=1,\dots,Q}$ radioactivity measurements at times $(t_i)_{i=1,\dots,Q}$

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix} = \begin{pmatrix} x_1 e^{-x_2 t_1} - N_1 \\ x_1 e^{-x_2 t_2} - N_2 \\ \vdots \\ x_1 e^{-x_2 t_Q} - N_Q \end{pmatrix}. = 0_{\mathbb{R}^q}$$

Calculate the Jacobian matrix $Jf(x)$

Toy example

$f : \mathbb{R}^2 \to R^Q$ with Q large

$(N_i)_{i=1,\dots,Q}$ radioactivity measurements at times $(t_i)_{i=1,\dots,Q}$

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_Q(x) \end{pmatrix} = \begin{pmatrix} x_1 e^{-x_2 t_1} - N_1 \\ x_1 e^{-x_2 t_2} - N_2 \\ \vdots \\ x_1 e^{-x_2 t_Q} - N_Q \end{pmatrix}.$$

Calculate the Jacobian matrix $Jf(x)$

$$Jf(x) = \begin{pmatrix} e^{-x_2 t_1} & -x_1 t_1 e^{-x_2 t_1} \\ e^{-x_2 t_2} & -x_1 t_2 e^{-x_2 t_2} \\ \vdots & \\ e^{-x_2 t_Q} & -x_1 t_Q e^{-x_2 t_Q} \end{pmatrix}.$$

Reminders: linear least squares

$Ax = b$ for $b \in \mathbb{R}^Q$ and $A \in \mathcal{M}_{Q,P}(\mathbb{R})$ with $Q > P$ and $\mathrm{rg}(A) = P$.
The problem: find $x \in \mathbb{R}^P$ such that

$$\|Ax - b\|^2 = \min_{y \in \mathbb{R}^P} \|Ay - b\|^2$$

admits a unique solution given by the normal equation

$$A^t Ax = A^t b.$$

Nonlinear case

$\nabla \|f(x)\|^2 =$

$$\begin{cases} \text{Find } x^* \in \mathbb{R}^P \text{ such that} \\ \|f(x^*)\|^2 = \min_{x \in \mathbb{R}^P} \|f(x)\|^2 \quad \left(\|f(x)\|^2 = \sum_{k=1}^{Q} (f_k(x))^2 \right), \end{cases}$$

We suppose that :

$$\forall x \in \mathbb{R}^P, \qquad J_f(x) \in \mathcal{M}_{Q,P}(\mathbb{R}) \text{ has rank } P.$$

In particular, we will have $\underbrace{(J_f(x))^t J_f(x)}_{\text{can be inverted}}$ symmetric defined positive.

Nonlinear case (continued)

We note

$$g : \begin{cases} \mathbb{R}^P & \to & \mathbb{R} \\ x & \mapsto & \|f(x)\|^2 \end{cases}$$

If g is strictly convex and coercive then the problem $g(x^*) = \min_x g(x)$ admits a unique solution x^*

$$\nabla g(x^*) = 0.$$

Calculating the gradient of g

$Jf(x) \underset{q}{=} \Big(\qquad \Big)$

2 columns

$g = \|f(x)\|^2 = N o f$, composition of
$N : \mathbb{R}^Q \to \mathbb{R}$, $N(y) = ||y||^2$ and $f : \mathbb{R}^P \to \mathbb{R}^Q$.
The rule for differentiating a composite function gives
$Dg(x) = DN(f(x))Df(x)$
For $y, \delta \in \mathbb{R}^Q$, $DN(y)\delta = 2y \cdot \delta$
For $x, h \in \mathbb{R}^P$, $Df(x)h = Jf(x)h \in \mathbb{R}^Q$

$$h, x \in \mathbb{R}^P, \quad Dg(x)h = 2f(x) \cdot Jf(x)h = 2Jf(x)^T f(x) \cdot h$$

$$\nabla g(x) = 2Jf(x)^T f(x).$$

Pay attention to the dimensions of the different terms. The order of the operations is important

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = \underbrace{2Jf(x)^T f(x)}_{}$ with Newton method requires $Hf(x)$

$\nabla g(x) \in \mathbb{R}^P$ (toy problem $P=2$)

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = 2Jf(x)^T f(x)$ with Newton method requires $Hf(x)$
- If $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^P$ we find the zeros by Newton's algorithm

$$x_{k+1} = x_k + d_k$$

$$\text{with } Jf(x_k)d_k = -f(x_k).$$

Here it is not possible because

$$Jf(x) \in \mathcal{M}_{Q \times P}$$

Find the zeros of ∇g or the zeros of $f(x)$

- Zeroing $\nabla g(x) = 2Jf(x)^T f(x)$ with Newton method requires $Hf(x)$
- If $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^P$ we find the zeros by Newton's algorithm

$$x_{k+1} = x_k + d_k$$

$$\text{with } Jf(x_k)d_k = -f(x_k).$$

- Here $f(x)$ is a function of $\mathbb{R}^P$ in $\mathbb{R}^Q$ so the system $Jf(x_k)d_k = -f(x_k)$ of size $Q \times P$ is solved in the least squares sense

$$Jf(x_k)^T Jf(x_k)d_k = -Jf(x_k)^T f(x_k)$$

$$\Leftrightarrow \quad d_k = -(Jf(x_k)^T Jf(x_k))^{-1} Jf(x_k)^T f(x_k).$$

Gauss Newton method

to minimize $\|f(x_k)\|^2$

$d_k = \begin{pmatrix}1\\1\\1\end{pmatrix} \in \mathbb{R}^P$

- Initialize $x_0 \in \mathbb{R}^P$
- While $\cancel{\|f(x_k)\|}$ $\|d_k\| > \varepsilon$ $> \varepsilon$ and $k < k_{\max}$
  - Solve $(Jf(x_k)^T Jf(x_k))d_k = -Jf(x_k)^T f(x_k)$ ← normal equation in the linear case
  - Update $x_{k+1} = x_k + d_k$
  - Update $k \to k + 1$

Same pros and cons as Newton method

Convergence of the Gauss Newton method

We recall that $Jf(x)$ of rank P and $g(x)$ is strictly convex coercive

- Let $x_k \in \mathbb{R}^P$, then the direction $d_k = -(Jf(x_k)^T Jf(x_k))^{-1} Jf(x_k)^T f(x_k)$ satisfies

$$\nabla g(x_k) \cdot d_k \leq 0.$$

  If $x_k \neq x^*$ then

$$\nabla g(x_k) \cdot d_k < 0.$$

  So d_k is a descent direction for g at x_k.
- If the sequence $(x_k)_k$ converges, then its limit is x^*.

Experiment with notebook 4

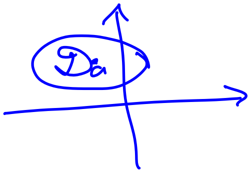
Toy problem
with radioactive
materials

Outline

Canonical problem

$\mathcal{D}_a = \{x, \|x-c\|^2 = r^2\}$

$\mathcal{D}_a$ is bounded



$f(x) = \|x\|^2$ on $\mathbb{R}^2$

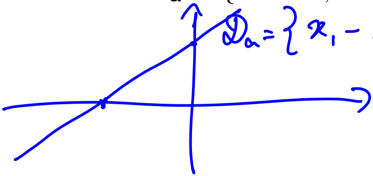
$c^E(x) = \|x-c\|^2 - r^2$

$$\inf f(x) \quad \begin{cases} c^E(x) = 0 \\ c^I(x) \le 0 \\ x \in \mathbb{R}^n \end{cases} \qquad \text{with} \qquad \begin{cases} f : \mathbb{R}^n \longrightarrow \mathbb{R}, \\ c^E : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \\ c^I : \mathbb{R}^n \longrightarrow \mathbb{R}^p, \\ f, c, \text{ smooth.} \end{cases}$$

$$D_a = \{x \in \mathbb{R}^n,\ c^E(x) = 0,\ c^I(x) \le 0\}$$

$\mathcal{D}_a = \{x_1 - x_2 = -1\}$ line

$\mathcal{D}_a$ is unbounded



General existence theorem

We consider f continuous from $C \subset \mathbb{R}^n$ into $\mathbb{R}$ with C closed.
If one of the following hypotheses is satisfied

- C bounded
- C not bounded and f coercive

$f(x) \longrightarrow +\infty$
$\|x\| \longrightarrow +\infty$

then f has a minimum on C

$c^E : \mathbb{R}^n \to \mathbb{R}^m$ $\quad \nabla c_i^E = a_i^E \in \mathbb{R}^n$

$c^I : \mathbb{R}^n \to \mathbb{R}^p$

Notations for the gradient and the hessian of the i^{th} constraint

$$a_i^E(x) = \nabla c_i^E(x), \quad H_i^E(x) = Hess\ c_i^E(x),$$
$$a_i^I(x) = \nabla c_i^I(x), \quad H_i^I(x) = Hess\ c_i^I(x).$$

Jacobian matrices of the constraints :

$$A^E(x) = \nabla c^E(x) = \begin{pmatrix} a_1^E(x)^T \\ \vdots \\ a_m^E(x)^T \end{pmatrix}, \quad A^I(x) = \nabla c^I(x) = \begin{pmatrix} a_1^I(x)^T \\ \vdots \\ a_p^I(x)^T \end{pmatrix}.$$

$$JC^E = A^E = \underbrace{\begin{pmatrix} \nabla c_1^{E\,T} \\ \\ \nabla_m^{E\,T} \end{pmatrix}}_{n \text{ columns}} m \text{ rows}$$

$\ell : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$

Lagrangian and Lagrange multipliers (P)

$\inf f(x)$
$c^E(x) = 0$
$c^I(x) \leq 0$

Let y a vector of $\mathbb{R}^m$, z a vector of $\mathbb{R}^p$, Lagrange multipliers.
The Lagrangien is defined by → associated to the problem (P)

$$\ell(x,y,z) = f(x) + y \cdot c^E(x) + z \cdot c^I(x)$$

$$= f(x) + \sum_{i=1}^{m} y_i \, c_i^E(x) + \sum_{j=1}^{p} z_j \, c_j^I(x)$$

The gradient and the hessian of the Lagrangien with respect to x are

$$g(x,y,z) = \nabla_x \ell(x,y,z) = \nabla f(x) + \sum_{i=1}^{m} y_i a_i^E(x) + \sum_{i=1}^{p} z_i a_i^I(x),$$

$$H(x,y,z) = Hess_x \ell(x,y,z) = Hf(x) + \sum_{i=1}^{m} y_i H_i^E(x) + \sum_{i=1}^{p} z_i H_i^I(x).$$

Example 1: $f : \mathbb{R}^2 \to \mathbb{R} \quad f(x) = x_1 + x_2, \quad \inf_{x_1^2+x^2=2} f(x)$

$$\ell : \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}$$

$$(x, y) \longrightarrow \ell(x,y) = x_1 + x_2 + y(x_1^2 + x_2^2 - 2)$$

Example 2: $f : \mathbb{R}^n \to \mathbb{R} \quad f(x) = ||x||^2, \quad \inf_{\substack{x_{i+1} - x_i \leq 2 \\ i=1,\dots,n-1}} f(x)$

$$\mathbb{R}^n \times \mathbb{R}^{n-1} \longrightarrow \mathbb{R}$$

$$(x, z) \longrightarrow \ell(x,z) = ||x||^2 + z \cdot C^I(x) = ||x||^2 + \sum_{j=1}^{n-1} z_j (x_{j+1} - x_j - 2)$$

primal

$$\nabla_x \ell(x,z) = 2x + \sum_{j=1}^{n-1} z_j \nabla(x_{j+1} - x_j - 2)$$

$$= 2x + z_1 \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} + \dots + z_j \begin{pmatrix} 0 \\ -1 \\ 1 \\ 0 \end{pmatrix} \dots + z_{n-1} \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$$

$$= 2x + \begin{pmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{pmatrix} z$$

n rows

n-1 columns

Examples

Actives constraints

$(P)\ \inf f(x)$

$\mathcal{D}_a = \{c^E(x)=0,\ c^I(x)\le 0\}$

Let $x^\star$ a minimizer of f. on $\mathcal{D}_a$

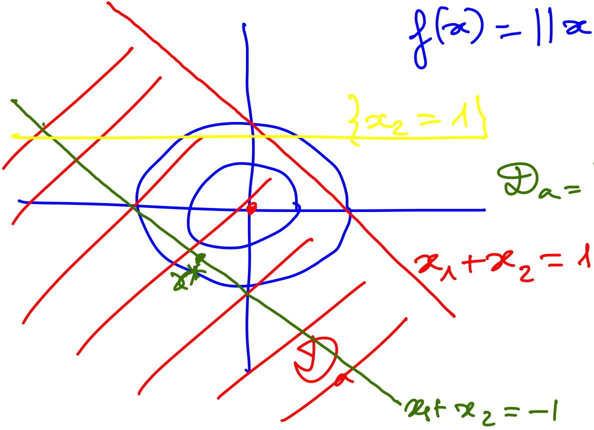
The i^{th} inequality constraint is active if $c_i^I(x^\star) = 0$.

$f(x) = \|x\|^2$

$\mathcal{D}_a = \{x_1+x_2 \le 1\}$

$x^* = (0,0)$

c^I is not active



$\mathcal{D}_a = \{x_1+x_2 \le -1\}$

$x^* = \left(\frac{-1}{2}, \frac{-1}{2}\right)$

c^I is active

$c^F(x) = \begin{pmatrix} x_1+x+1 \\ x_2-1 \end{pmatrix}$

Actives constraints

Let $x^\star$ a minimizer of f.

The i^{th} inequality constraint is active if $c_i^I(x^\star) = 0$.

1. Exemple $f : \mathbb{R}^2 \to \mathbb{R} \quad f(x) = ||x||^2, \inf_{x_1+x_2 \leq 1} f(x)$

Actives constraints

Let $x^\star$ a minimizer of f.
The i^{th} inequality constraint is active if $c_i^I(x^\star) = 0$.

2. Exemple $f : \mathbb{R}^2 \to \mathbb{R} \quad f(x) = ||x||^2, \inf_{x_1+x_2 \leq -1} f(x)$

proof of next slide

$$g(y, z) = \inf_{\mathbb{R}^n} f(x) + y \cdot c^E(x) + z \cdot c^I(x)$$

$$\leq \ell(x^*, y, z) = f(x^*) + \underbrace{z}_{\geq 0} \cdot \underbrace{c^I(x^*)}_{\leq 0}$$

$$\leq f(x^*) = p^*$$

Lagrange dual function

$\ell(x,y,z) = f(x) + y\cdot c^E(x) + z\cdot c^I(x)$

for each x

$g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ $\quad (y,z) \to \ell(x,y,z)$ is affine

$(y,z) \to g(y,z)$

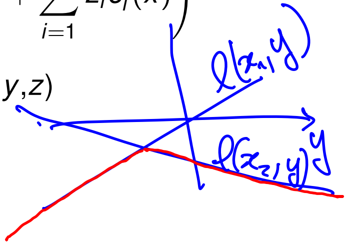
$$g(y,z) = \inf_{x\in D_a} \ell(x,y,z)$$

$$= \inf_{x\in D_a} \left(f(x) + \sum_{i=1}^{m} y_i c_i^E(x) + \sum_{i=1}^{p} z_i c_i^I(x) \right)$$

g is concave (can be unbounded for some y,z)

Property : inferior bound:

If $z \geq 0$ then $g(y,z) \leq p^\star = \inf_{x\in D_a} f(x)$



Example : solution of a linear system with minimal norm

Solve

$$p^\star = \inf_{Ax=b} x^T x$$

- Lagrangian : $\ell(x,y) = x^T x + y^T(Ax-b)$

$= f(x) + y.c^E(x)$

$f(x) = x^T x = \|x\|^2$

$c^E(x) = Ax - b$

$A \in \mathcal{M}_{m\times n}(\mathbb{R})$

$b \in \mathbb{R}^m$

$g(y,z) = \inf_{x\in\mathcal{D}} \ell(x,y)$

$\nabla_x \ell(x,y) = 2x + A^T y$

$\nabla_x \ell(x,y) = 0$ when

$x = -\frac{1}{2}A^T y$

$g(y,z) \leq \frac{1}{4}\|A^T y\|^2 + y.\left(A\left(-\frac{1}{2}A^T y\right) - b\right)$

$= -\frac{1}{4}\|A^T y\|^2 - \underline{y.b} \quad \leq p^\star$ for all y, z

Example : solution of a linear system with minimal norm

Solve

$$p^\star = \inf_{Ax=b} x^T x$$

- Lagrangian : $\ell(x,y) = x^T x + y^T(Ax-b)$
- In order to minimize $\ell(x,y)$ with respect to x we seek gradient zeros

$$\nabla_x \ell(x,y) = 2x + A^T y = 0 \;\Leftarrow\; x = -A^T y/2$$

Example : solution of a linear system with minimal norm

Solve

$$p^\star = \inf_{Ax=b} x^T x$$

- Lagrangian : $\ell(x,y) = x^T x + y^T(Ax-b)$
- In order to minimize $\ell(x,y)$ with respect to x we seek gradient zeros

$$\nabla_x \ell(x,y) = 2x + A^T y = 0 \Leftarrow x = -A^T y/2$$

- Inject in the definition of the dual function

$$g(y) = \ell(-A^T y/2, y) = -\frac{1}{4} y^T A A^T y - b^T y$$

concave in y

Example : solution of a linear system with minimal norm

Solve

$$p^\star = \inf_{Ax=b} x^T x$$

the primal is on $\mathbb{R}^n$

the dual is on $\mathbb{R}^m$

- Lagrangian : $\ell(x,y) = x^T x + y^T(Ax-b)$
- In order to minimize $\ell(x,y)$ with respect to x we seek gradient zeros

$$\nabla_x \ell(x,y) = 2x + A^T y = 0 \Leftarrow x = -A^T y/2$$

- Inject in the definition of the dual function

$$g(y) = \ell(-A^T y/2, y) = -\frac{1}{4} y^T A A^T y - b^T y$$

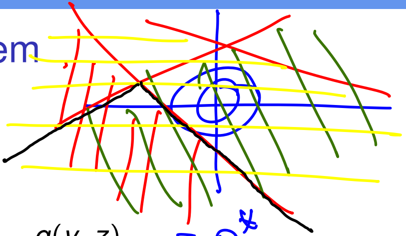
concave in y

- Inferior bound property

$$\boxed{p^\star \geq -\frac{1}{4} y^T A A^T y - b^T y \ \forall y}$$

easier to solve this than p

Resolution of the dual problem



$$JC(x^*) = \begin{cases} \nabla c_1^T(x^*) \\ \vdots \\ \nabla c_p^T(x^*) \end{cases} \quad n=2$$

Solve

$$d^\star = \sup_{y \in \mathbb{R}^m, z \in \mathbb{R}^p, z \geq 0} g(y,z) = p^*$$

- Best inferior bound for $p^\star \geq d^\star$
- The dual problem is concave : existence of an optimal problem $d^\star$

Weak duality $d^\star \leq p^\star$ ← always true I.B.P.

$\boxed{\text{Strong duality } d^\star = p^\star}$

Qualified constraints : condition for $d^\star = p^\star$

$JC(x^*) = \begin{matrix} JC^E(x^*) \; \} m \text{ rows} \\ JC^I(x^*) \; \} q \text{ rows} \end{matrix}$

$\max_{y, z \geq 0} g(y, z)$ $\inf_{x \in D_a} f(x)$

- Slater condition for a convex problem f , C^I_i convex and C^E : $Ax - b$ affine : there exists an interior point in D_a. *i.e.* There exists x such that $C^E(x) = 0$ and $C^I_k < 0$ for all $k = 1, \cdots, p$.
- Linear independence constraint qualification : the rank of the matrix formed by the union of the Jacobian matrix of equality constraints and the Jacobian matrix of q constraints of active inequality in $x^\star$ is equal to $m + q$ $\leq n$, then called maximal rank.

$C^I : \mathbb{R}^n \to \mathbb{R}^p$ at $C^I(x^*) = \begin{pmatrix} 0 \\ <0 \\ \vdots \\ \leq 0 \\ 0 \end{pmatrix}$ p constraints and $q \leq p$ are 0 at x^*

if $m > n$ $JC^E(x^*)$ cannot be max rank.

Case of equality constraints

$$\left\{\begin{array}{ll} \inf & f(x) \\ s.c. & C(x) = 0 \\ & x \in \mathbb{R}^n \end{array}\right.$$

with

$$\begin{array}{ccl} f & : & \mathbb{R}^n \longrightarrow \mathbb{R}, \\ C & : & \mathbb{R}^n \longrightarrow \mathbb{R}^m, \\ f & C & \text{smooth.} \end{array}$$

Lagrange multipliers Theorem

equality constraint

Let f and C in C^1, and $x^\star$ a local minimizer of f satisfying

$$C(x^\star) = 0 \quad \text{primal feasability}$$

rank $\nabla C(x^*) = m$

If the constraints are qualified, there exists a vector of Lagrange multipliers $y^\star \in \mathbb{R}^m$ s. t. $\nabla \ell(x^*, y^*) = 0$

$$\nabla f(x^\star) + \sum_{i=1}^{m} y_i^\star \nabla C_i(x^\star) = 0 \quad \text{dual feasability}$$

- Linear constraints special case
- $n = 2$, $m = 1$ special case

Linear constraints special case

$c : \mathbb{R}^n \to \mathbb{R}^m$
$x \mapsto Cx$ with $C \in \mathcal{M}_{m\times n}(\mathbb{R})$

$$c(x) = (c_1 \cdot x, \dots, c_m \cdot x)^\perp$$

$(c_i)_{i=1,\dots,m}$independendant vectors family in R^n $m \le n$

$\mathcal{D}_a$ $\cancel{K} = \{x, c_i \cdot x = 0, i = 1, \dots, m\}$ linear subspace

$\mathcal{D}_a = \mathrm{span}((c_i)_{i=1,m})^\perp$

$$\inf_{x \in \cancel{K}\,\mathcal{D}_a} f(x) \Leftrightarrow \inf_{\alpha \in \mathbb{R}^p} g(\alpha)$$

with $g(\alpha) = f\left(\sum_{i=1}^{p} \alpha_i k_i\right)$, $(k_i)_{i=1,\dots,p}$, basis of $\cancel{K}\,\mathcal{D}_a$

$p = n - m$

$\inf_{\alpha \in \mathbb{R}^p} g(\alpha)$ is unconstrained minimization in $\mathbb{R}^p$ $\leftarrow \nabla g(\alpha) = 0$ $\qquad g(\alpha) = f\left(\sum_{i=1}^{p} \alpha_i k_i\right)$

$$\nabla g(\alpha) = \left(\nabla f\left(\sum \alpha_i k_i\right) \cdot k_i\right)_{i=1,\dots p}$$

$$\Rightarrow \nabla f(x^*) \cdot k_i = 0 \quad \text{for } i=1,\dots p$$

$$\text{span } k_i = \mathcal{D}_a \Rightarrow \nabla f(x^*) \in \mathcal{D}_a^{\perp}$$

$$\mathcal{D}_a = \{(c_i)_i\}^{\perp} \Longrightarrow \nabla f(x^*) \in \text{span}\{(c_i)_{i=1,\dots m}\}$$

$$\Rightarrow \exists \lambda_i \quad \nabla f(x^*) = \sum \lambda_i c_i$$

$$y_i^* = -\lambda_i \qquad \nabla f(x^*) + \sum y_i^* c_i = 0$$

$$\nabla \ell(x^*, y^*) = 0$$

Special case $n = 2$, $m = 1$

Qualification condition for one single constraint $m = 1$:
$\nabla_x c_1(x^\star) \neq 0$, we can suppose $\partial_{x_2} c_1(x^\star) \neq 0$.
Implicit function theorem : $\exists V_1 \times V_2$ containing $x^\star$ and φ unique and differentiable in $x^\star$ s. t. $\forall x_1 \in V_1$ $c_1([x_1, \varphi(x_1)]) = 0$ and $x_2^\star = \varphi(x_1^\star)$ with

$$\varphi'(x_1) = \frac{-1}{\partial_{x_2} c_1(x)} \partial_{x_1} c_1(x).$$

Proof

$$\inf_{c_1(x)=0} f(x) \quad \Leftrightarrow \quad \inf_{x_1 \in V_1} \tilde{f}(x_1), \quad \text{with } \tilde{f}(x_1) = f([x_1, \varphi(x_1)])$$

First order optimality conditions for $\tilde{f}$ (without constraints since V_1 is an open set)

$$\tilde{f}'(x_1^\star) = 0 \Leftrightarrow \frac{\partial f}{\partial x_1}([x_1^\star, \varphi(x_1^\star)]) + \varphi'(x_1^\star)\frac{\partial f}{\partial x_2}([x_1^\star, \varphi(x_1^\star)]) = 0.$$

$$y = -\frac{\partial_{x_2} f(x^\star)}{\partial_{x_2} c_1(x^\star)}$$

Example 1

$\mathcal{D}_a =$ circle of radius 1, center 0,0

$$\inf_{x_1^2+x_2^2=1} x_1^4 + x_2^4.$$

Resolution by changing variables in polar coordinates

Set $x_1 = \cos(\theta), x_2 = \sin(\theta)$, problem (4) becomes $\inf_{\theta\in[0,2\pi]}(\cos\theta^4 + \sin\theta^4)$ whose solution is obtained by finding the zero of the derivative:

$$4\cos\theta\sin\theta(-\cos\theta^2 + \sin\theta^2) = -2\sin(2\theta)\cos(2\theta) = 0,$$

4 local minima $(\pm\sqrt{2}/2, \pm\sqrt{2}/2)$, where $f(x) = 1/2$,
4 local maxima $\{(1,0),(0,1),(-1,0),(0,-1))\}$, where $f(x) = 1$.

Resolution using Lagrange multipliers

$$\ell : \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}$$
$$(x, y) \longmapsto x_1^4 + x_2^4 + y(x_1^2 + x_2^2 - 1)$$

We seek $x^\star \in \mathbb{R}^2$ and $y^\star \in \mathbb{R}$ s. t.

$C(x) = x_1^2 + x_2^2 - 1$

$JC(x^*) = 2x^*$

rank $JC(x^*) = 1$

because $(0,0) \notin \mathcal{D}_a$

$$C(x^*) = 0 \Leftarrow \quad (x_1^\star)^2 + (x_2^\star)^2 = 1$$

$$\left\{ \begin{aligned} 4(x_1^\star)^3 + y^\star 2x_1^\star &= 0 \\ 4(x_2^\star)^3 + y^\star 2x_2^\star &= 0 \end{aligned} \right.$$

$$\nabla_x \ell(x, y) = \begin{pmatrix} 4x_1^3 \\ 4x_2^3 \end{pmatrix} + y \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Resolution using Lagrange multipliers

We seek $x^\star \in \mathbb{R}^2$ and $y^\star \in \mathbb{R}$ s. t.

$$\begin{aligned} (x_1^\star)^2 + (x_2^\star)^2 &= 1 \\ 4(x_1^\star)^3 + y^\star 2x_1^\star &= 0 \\ 4(x_2^\star)^3 + y^\star 2x_2^\star &= 0 \end{aligned}$$

Resolution using Lagrange multipliers

We seek $x^\star \in \mathbb{R}^2$ and $y^\star \in \mathbb{R}$ s. t.

$$
\begin{array}{rcl}
(x_1^\star)^2 + (x_2^\star)^2 & = & 1 \\
2x_1(2x_1^2+y) = \quad 4(x_1^\star)^3 + y^\star 2x_1^\star & = & 0 \\
2x_2(2x_2^2+y) = \quad 4(x_2^\star)^3 + y^\star 2x_2^\star & = & 0
\end{array}
$$

	$x_1^\star = 0$	$y^\star = -2(x_1^\star)^2$
$x_2^\star = 0$	$(x_1^\star)^2 + (x_2^\star)^2 \neq 1$	$(x_1^\star)^2 = 1$ et $y^\star = -2$ $f(x^\star) = 1$
$y^\star = -2(x_2^\star)^2$	$(x_2^\star)^2 = 1$ et $y^\star = -2$ $f(x^\star) = 1$	$(x_1^\star)^2 = (x_2^\star)^2 = 1/2$ et $y^\star = -1$, $f(x^\star) = 1/2$

Second order optimality conditions

$\nabla_x \ell(x^*, y^*) = 0_{\mathbb{R}^n}$

Let f and c in C^2, and $x^\star$ be a local minimizer of f verifying the constraints of equality $c(x^\star) = 0$. If the constraints are qualified, there exists a vector of Lagrange multipliers $y^\star \in \mathbb{R}^m$ such that

$$s \cdot H(x^\star, y^\star)s \geq 0 \quad \text{for all } s \in \mathcal{N}$$

where

$$\mathcal{N} = \{s \in \mathbb{R}^n, A(x^\star)s = 0\}.$$

$A = JC(x^*)$

$H(x^*, y^*) = Hf(x^*) + \sum_{i=1}^{m} y_i^* HC_i^E(x^*)$

Interpretation of Lagrange multipliers

The Lagrange multiplier y_i measures the sensitivity of the minimum $x^\star$ with respect to the corresponding constraint.

Initial primal and dual problems

$\inf_{c(x)=0} f(x)$	$\sup_y g(y)$ avec $g(y) = \inf_x f(x) + y^t c(x)$

Perturbated primal and dual problems

$\inf_{c(x)=\varepsilon} f(x)$	$\sup_y g(y) - \varepsilon^T y$

- x is the primal variable, ε a parameter
- $p^\star(\varepsilon)$ the optimal value when ε varies

Global interpretation of Lagrange multipliers

Hyp: strong duality for the undisturbed problem, that is $y^\star$ t.q. $g(y^\star) = d^\star = p^\star(0)$
For the perturbated problem we have

$$\begin{aligned} p^\star(\varepsilon) &\geq \max_y g(y) - \varepsilon^T y \\ &\geq g(y^\star) - \varepsilon^T y^\star \\ &\geq p^\star(0) - \varepsilon^T y^\star \end{aligned}$$

hence

- if $y_i^\star > 0$ and large, $p^\star$ increases a lot if $\varepsilon_i < 0$
- if $y_i^\star < 0$ and large, $p^\star$ disminishes a lot if $\varepsilon_i > 0$

Local interpretation of Lagrange multipliers

$\mathbb{R}^n = \text{span}(e_i)$

$e_i = (\delta_{ij})_{j=1,\dots n}$

$p^*(\varepsilon) = \inf_{C(x)=\varepsilon e_i} f(x)$

$$y_i^\star = -\frac{\partial p^\star(0)}{\partial \varepsilon_i}$$

Proof : $\varepsilon = te_i$ in the global sensitivity
$p^\star(te_i) \geq p^\star(0) - ty_i^\star$

$$\lim_{t\searrow 0} \frac{p^\star(te_i) - p^\star(0)}{t} \geq -y_i^\star$$
$$\lim_{t\nearrow 0} \frac{p^\star(te_i) - p^\star(0)}{t} \leq -y_i^\star$$

Example 2 : Diagonalization of a symetric matrix

$\ell : \mathbb{R}^2 \times \mathbb{R} \longrightarrow \mathbb{R}$ $\qquad \ell(x, y) = Ax.x + y\,(\|x\|^2 - 1)$

$$\inf_{\|x\|=1} Ax \cdot x$$

$A \in S^n$

with A a symetric matrix in $\cancel{\mathbb{R}^{n \times n}}$.

ensure differentiability

$$\inf_{c(x)=0} f(x) \quad \text{with } f(x) = Ax \cdot x \text{ and } c(x) = \|x\|^2 - 1$$

Existence of a minimum since f is continuous and $\{x, \|x\| = 1\}$ bounded closed set.
f differentiable and $\{c(x) = 0\}$ Lagrange multipliers $\Rightarrow \exists y^\star \in \mathbb{R}$ s.t. $2Ax^\star + 2y^\star x^\star = 0$ $\qquad \lambda = -y^\star$ is a real eigenvalue
$\Rightarrow \exists (\lambda, v) \in \mathbb{R} \times \mathbb{R}^n,\ Av = \lambda v$ and $f(v) = \inf_{\|x\|=1} f(x)$.
Induction hypothesis H_n : existence of an orthonormal eigenvector basis of A with n related eigenvalues
$n = 1$ easy

Suppose H_n true

We know that there is one eigenvalue

$\nabla \ell(x^*, y^*) = 0 \qquad A x^* = -y^* x^*$

For $A \in \mathbb{R}^{n+1 \times n+1}$ we consider the subspace $H = \{\text{vect}(x^\star)\}^\perp$.
dim $H = n$
H is stable by A. Indeed

$$\text{if } x^\star \cdot x = 0 \text{ then } x^\star \cdot Ax = Ax^\star \cdot x = -y^\star x^\star \cdot x = 0$$

The restriction of A to H is a matrix $n \times n$ therefore using H_n existence of a orthonormal eigenvector basis of the restriction of A to H.
We divide $x^\star$ by $\|x^\star\|$ in order to complete this basis on $\mathbb{R}^{n+1}$.

Example 3 : Minimization of a quadratic function under linear constraints of equality

$$\begin{aligned} f(x) &= \frac{1}{2}Ax\cdot x + b\cdot x \\ c(x) &= Bx - C \end{aligned}$$

symetric definite positive.

with A ~~defined symetric positive~~ matrix in $\mathbb{R}^{n\times n}$, b vector in $\mathbb{R}^n$, B matrix in $\mathbb{R}^{m\times n}$ and C vector in $\mathbb{R}^m$.

Qualified constraints $\Leftrightarrow$ rank$(B) = m$.

Lagrangian : $\ell : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$

$$\ell(x,y) = \underbrace{\frac{1}{2}Ax\cdot x + b\cdot x}_{f(x)} + \underbrace{y\cdot(Bx - C)}_{y\cdot c(x)}$$

Theorem of Lagrange multipliers

$y \in \mathbb{R}^m$ $\quad \ell(x,y) = \frac{1}{2} Ax \cdot x + (Bx - C) \cdot y$

$A \in S_n^{++}$

$$\nabla_x \ell(x,y) = \qquad Ax + b + B^t y = 0$$
$$Bx = C$$

A defined symetric positive matrix $\Rightarrow x = -A^{-1}(b + B^t y)$.
$B(-A^{-1}(b + B^t y)) = C$ $\qquad -BA^{-1}b - BA^{-1}B^t y = C$
$\text{rank}(B) = m \Rightarrow BA^{-1}B^t$ is invertible
$BA^{-1}B^t y = -(BA^{-1}b + C)$ from which we get y then x.

$$y = -(BA^{-1}B^t)^{-1}(BA^{-1}b + C)$$

Application

0) write f as a quadratic function

$f(x) = \frac{1}{2} A x . x + b . x$

$\nabla f(x) : \begin{cases} 6x_1 - 3x_2 \\ 10x_2 - 3x_1 \end{cases}$ $Hf(x) = \begin{pmatrix} 6 & -3 \\ -3 & 10 \end{pmatrix}$

Find

$$\inf_{C(x)=0} f(x) \quad \text{with} \begin{cases} \boxed{f(x) = 3x_1^2 + 5x_2^2 - 3x_1 x_2,} \\ C(x) = x_1 + x_2 - 1 \end{cases}$$

$f(x) = \frac{1}{2} Hf(0) x . x + \nabla f(0) . x + f(0)$

- Lagrangian $\ell(x, y) = 3x_1^2 + 5x_2^2 - 3x_1 x_2 + y(x_1 + x_2 - 1)$

to check your computations:

$x_2 = 1 - x_1$

$g(x) = f(x, 1-x)$

$f(x) = \frac{1}{2} \begin{pmatrix} 6 & -3 \\ -3 & 10 \end{pmatrix} x . x$

Application

Find

$$\inf_{C(x)=0} f(x) \quad \text{with} \begin{cases} f(x) = 3x_1^2 + 5x_2^2 - 3x_1x_2, \\ C(x) = x_1 + x_2 - 1 \end{cases}$$

$= f(x) + y\,C(x)$

- Lagrangian $\ell(x,y) = 3x_1^2 + 5x_2^2 - 3x_1x_2 + y(x_1 + x_2 - 1)$
- Gradient $\nabla\ell(x,y) = \begin{pmatrix} 6x_1 - 3x_2 + y \\ -3x_1 + 10x_2 + y \end{pmatrix}$

Can we apply Lag. Mult. th ?

$f \in C^1$, $JC(x) = \begin{pmatrix} 1 & 1 \end{pmatrix}$ Rank $JC(x) = 1$

Application

Find

$$\inf_{C(x)=0} f(x) \quad \text{with} \begin{cases} f(x) = 3x_1^2 + 5x_2^2 - 3x_1x_2, \\ C(x) = x_1 + x_2 - 1 \end{cases}$$

- Lagrangian $\ell(x, y) = 3x_1^2 + 5x_2^2 - 3x_1x_2 + y(x_1 + x_2 - 1)$
- Gradient $\nabla\ell(x, y) = \begin{pmatrix} 6x_1 - 3x_2 + y \\ -3x_1 + 10x_2 + y \end{pmatrix}$
- If $f(x^\star) = \inf_{C(x)=0} f(x)$ then $\exists y^\star \in \mathbb{R}$ s.t. $\nabla\ell(x^\star, y^\star) = 0$

Application

Find

$$\inf_{C(x)=0} f(x) \quad \text{with} \left\{ \begin{array}{l} f(x) = 3x_1^2 + 5x_2^2 - 3x_1 x_2, \\ C(x) = x_1 + x_2 - 1 \end{array} \right.$$

- Lagrangian $\ell(x,y) = 3x_1^2 + 5x_2^2 - 3x_1x_2 + y(x_1 + x_2 - 1)$
- Gradient $\nabla\ell(x,y) = \begin{pmatrix} 6x_1 - 3x_2 + y \\ -3x_1 + 10x_2 + y \end{pmatrix}$
- If $f(x^\star) = \inf\limits_{C(x)=0} f(x)$ then $\exists y^\star \in \mathbb{R}$ s.t. $\nabla\ell(x^\star, y^\star) = 0$
- Plus the primal condition $C(x^\star) = 0$

Application - see notebook 6

$$\begin{cases} 6x_1 - 3x_2 + y = 0 \\ -3x_1 + 10x_2 + y = 10 \\ x_1 + x_2 - 1 = 0 \end{cases}$$

- Solve the system of 3 equations to find $x^\star, y^\star$

Application - see notebook 6

- Solve the system of 3 equations to find $x^\star, y^\star$
- Other method ? change of variable

Utilisation : SQP algorithm -linear equality constraints

Let the minimization problem with linear equality constraints

$$\left\{\begin{array}{ll} \inf & f(x) \\ s.c. & Bx - c = 0 \\ & x \in \mathbb{R}^n \end{array}\right. \quad \text{with} \quad \left\{\begin{array}{l} f : \mathbb{R}^n \longrightarrow \mathbb{R}, \text{ twice differentiable} \\ B \in \mathcal{M}_{m\times n}(\mathbb{R}), \\ c \in \mathbb{R}^m. \end{array}\right.$$

The Lagrangian is

$$\ell(x, y) = f(x) + ay \cdot (Bx - c)$$

The 1st order optimality constrainst are

$$\begin{array}{rcl} \nabla_x \ell(x, y) & = & \nabla f(x) + B^T y = 0_{\mathbb{R}^n} \\ \nabla_y \ell(x, y) & = & Bx - c = 0_{\mathbb{R}^m} \end{array}$$

Let $G = \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$, $G(x, y) = \begin{pmatrix} \nabla f(x) + B^T y \\ Bx - c \end{pmatrix}$ and use the Newton method to find its zero in $\mathbb{R}^{n+m}$

SQP algorithm : Newton method in $\mathbb{R}^{n+m}$

$$G(x,y) = \begin{pmatrix} \nabla f(x) + B^T y \\ Bx - c \end{pmatrix} \quad JG(x,y) = \begin{pmatrix} Hf(x) & B^T \\ B & 0_{m\times m} \end{pmatrix}$$

Newton method : $\begin{cases} x_{k+1} = x_k + d_k \\ y_{k+1} = y_k + \delta_k \end{cases}$

$$JG(x_k, y_k) \begin{pmatrix} x_k \\ \delta_k \end{pmatrix} = -G(x_k, y_k)$$

Leads to

$$\begin{cases} Hf(x_k)d_k + \nabla f(x_k) + B^T y_k + B^T \delta_k = 0 \\ Bd_k = 0 \end{cases}$$

Which is equivalent to solve

$$\inf_{Bd=0} \frac{1}{2} Hf(x_k)d \cdot d + (\nabla f(x_k) + B^T y_k) \cdot d$$

Algorithme SQP - linear equality constraints

At each iteration k

- let $J_k(d) = \frac{1}{2}Hf(x_k)d \cdot d + (\nabla f(x_k) + B^T y_k) \cdot d$
- minimize $J_k(d)$ under constraint $C(d) = Bd = 0$

$$\ell(d, \delta) = J_k(d) + \delta \cdot Bd$$

$\nabla_d \ell(d, \delta) = 0$ and $Cd = 0$ leads to the linear system

$$\begin{pmatrix} Hf(x^k) & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} d \\ \delta \end{pmatrix} = \begin{pmatrix} -\nabla f(x^k) - B^T y_k \\ 0_{\mathbb{R}^m} \end{pmatrix}$$

Algorithme SQP - linear equality constraints

3 pitfalls

1. $Hf(x^k)$ may be hard to compute : quasi-Newton approximation $\hat{H}$
2. $\hat{H}$ may be not invertible : penalize with $\max(0, -\min(\lambda_{\hat{H}}) + \varepsilon)$
3. $\nabla_\delta \ell(\delta, y)$ might not decrease -> line search for a better step

Algorithme SQP - linear equality constraints

Data: Function f, gradient ∇f, hessien Hf, tolerance τ, max number of iterations $k_{\max}$
Result: $f(x^\star) = \min_{x \in \mathbb{R}^n} f(x)$
Initialisation : choose $x_0 \in \mathbb{R}^n$, $d_0 \in \mathbb{R}^n$ t.q. $\|d_0\| > \tau$
while $\|d^k\| \geq \tau$ *and* $k < k_{\max}$ **do**
  Compute $f(x^k)$, $\nabla f(x^k)$ et $Hf(x^k)$
  Minimize $J_k(d) = \frac{1}{2} Hf(x_k) d \cdot d + (\nabla f(x_k) + B^T y_k) \cdot d$
  under constraints $Bd = 0 \quad \rightarrow$ find $d^\star$ and $\delta^\star$
  Update $d_k = d^\star$,
  Update $x^{k+1} = x^k + d^\star, \quad y^{k+1} = y^k + \delta^\star$
  Update $k \leftarrow k + 1$
end
$x^\star = x_k$

Back to the general situation

$$\left\{\begin{array}{l} \inf \quad f(x) \\ c^E(x) = 0 \\ c^I(x) \leq 0 \\ x \in \mathbb{R}^n \end{array}\right. \quad \text{with} \quad \left\{\begin{array}{l} f : \mathbb{R}^n \longrightarrow \mathbb{R}, \\ c^E : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \\ c^I : \mathbb{R}^n \longrightarrow \mathbb{R}^p, \\ f, c, \text{ smooth.} \end{array}\right.$$

$$D_a = \{x \in \mathbb{R}^n,\ c^E(x) = 0,\ c^I(x) \leq 0\}$$

$z_i \geq 0$ for $i = 1, \dots, p$

$$\begin{array}{rcl} \ell : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{+p} & \to & \mathbb{R} \\ \ell(x, y, z) & = & f(x) + y \cdot C^E(x) + z \cdot C^I(x) \end{array}$$

Optimality conditions for the optimization with constraints of inequality

Theorem

$x^\star$ is a local minimizer of f verifying the constraints of inequality $c_I(x^\star) \leq 0$ and the constraints of equality $c_E(x^\star) = 0$. If the constraints are qualified, there exists a vector $y^\star \in \mathbb{R}^m$ and a vector $z^\star \in {\mathbb{R}^+}^p$ of Lagrange multipliers such as

$$
\begin{array}{rcll}
c^E(x^\star) = 0, c^I(x^\star) & \leq & 0 & \textit{primal feasibility} \\
\boxed{\forall x \in \mathbb{R}^n} \quad \ell(x^\star, y^\star, z^\star) & \leq & \ell(x, y^\star, z^\star) & \textit{dual feasibility} \\
z^\star & \geq & 0 & \textit{dual feasibility} \\
c_i^I(x^\star) z_i^\star & = & 0 & \textit{complementary relaxation}
\end{array}
$$

Conditions of complementary relaxation

$g(y,z) = \inf_{x \in D_a} \ell(x, y, z)$

$$\begin{aligned} d^\star &:= g(y^\star, z^\star) = \sup_{y,z \in \mathbb{R}^{+p}} g(y,z) \\ &= \inf_{x \in D_a} f(x) + y^\star \cdot C^E(x) + z^\star \cdot C^I(x) = \inf_{x \in D_a} \ell(x, y^\star, z^\star) \end{aligned}$$

Suppose qualified constraints $\Leftrightarrow \quad p^\star = d^\star$

$$\begin{aligned} \boxed{p^\star := \inf_{x \in D_a} f(x)} &= \inf_x f(x) + y^\star \cdot C^E(x) + z^\star \cdot C^I(x) \\ &\leq f(x^\star) + y^\star \cdot \cancel{C^E(x^\star)} + \underbrace{z^\star \cdot C^I(x^\star)}_{= \sum_{j=1}^{p} z_j^\star C_j^I(x^\star)} \leq f(x^\star) \end{aligned}$$

$= f(x^\star)$

≤ 0

then

$$\boxed{z^\star \cdot C^I(x^\star) = 0} \Rightarrow z_j^\star C^I(x^\star)_j = 0 \ \forall\, j = 1, \ldots, p$$

Conditions of complementary relaxation

First order optimality conditions for the optimization with constraints of inequality

for $x \in \mathcal{D}_a$ $\quad \ell(x^*, y^*, z^*) \leq \ell(x, y^*, z^*)$

Theorem

Karush-Kuhn-Tucker (KKT) conditions
Let f, c^I and c^E in C^1, and $x^\star$ a local minimizer of f satisfying the inequality constraints $c^I(x) \leq 0$ and equality constraints $c^E(x^\star) = 0$. If the constraints are qualified, there exists $y^\star \in \mathbb{R}^m$ and $z^\star \in \mathbb{R}^{+p}$ Lagrange multipliers such that

$$\begin{array}{rcll}
c^E(x^\star) = 0, c^I(x^\star) & \leq & 0 & \textit{primal feasibility} \\
g(x^\star) + {A^E}^T(x^\star)y^\star + {A^I}^T(x^\star)z^\star & = & 0 & \textit{dual feasibility} = \nabla_x \ell(x^*, y^*, z^*) \\
z^\star & \geq & 0 & \textit{dual feasibility} \\
\forall i = 1, \ldots, m \quad c_i^I(x^\star)z_i^\star & = & 0 & \textit{complementary relaxation}
\end{array}$$

KKT Conditions deduced from Lagrange multipliers theorem

Replace the original inegality constrained problem by
$\inf x \in \mathbb{R}^n, t \in \mathbb{R}^p F(x,t)$ with
$F(x,t) = f(x)$
and equality constraints
$c_i^E(x) = 0$ for $i = 1, \dots, m$
$c_j^I(x) + t_j^2 = 0$ for $j = 1, \dots, p$. $\Longleftrightarrow$ $c_j^I(x) \leq 0$
The lagrangian of the modified problem is

$$L(x,t,y,z) = F(x,t) + \sum_{i=1}^{m} y_i c_i^E(x) + \sum_{j=1}^{p} z_j (c_j^I(x) + t_j^2)$$

Lagrange multipliers theorem provides

$$\nabla_{x,t} F(x,t) + \sum_{i=1}^{m} y_i \nabla_{x,t} c_i^E(x) + \sum_{j=1}^{p} z_j \nabla_{x,t} (c_j^I(x) + t_j^2) \quad = \quad 0$$

KKT proof...

$\ell(x, y, z) = f(x) + \sum y_i c_i^E + \sum_{j=1}^{p} z_j (c_j^I(x) + t_j^2)$

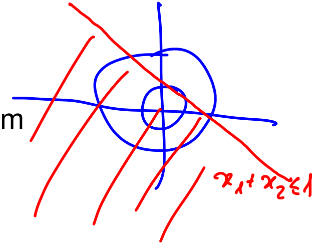
$$
\begin{aligned}
\nabla_x f(x) + \sum_{i=1}^{m} y_i \nabla_x c_i^E(x) + \sum_{j=1}^{p} z_j \nabla_x (c_j^I(x) &= 0 \qquad n \text{ equations} \\
2 z_j t_j &= 0, \quad j = 1, \ldots, p
\end{aligned}
$$

either $z_j = 0$ either $t_j = 0 = c_j^I(x) = 0 \Leftarrow$

Condition $z_j \geq 0$ To find this condition, we apply the 2nd order optimality condition on the Lagrangian of $F(x, t)$:

$$
H_{x,t} L(x, t, y, z) = \begin{pmatrix} H_x \ell(x, y, z)_{n \times n} & 0_{n \times p} \\ 0_{p \times n} & \begin{pmatrix} 2z_1 & 0 & \ddots & 0 \\ 0 & 2z_2 & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & 0 & 2z_p \end{pmatrix} \end{pmatrix} \Rightarrow z_j \geq 0
$$

Example of KKT application



We look at the quadratic minimization problem

$$\inf_{x_1+x_2-1\leq 0} x_1^2 + x_2^2.$$

Trivial solution: $(0,0)$ checks the inequality constraint therefore the constraint is inactive in $x^\star$, the solution of the problem is the solution of the unconstrained problem, i.e. $(0,0)$.
KKT check : We seek $(x^\star, z^\star)$ with $z^\star \geq 0$ s. t.

$$\begin{aligned} \begin{pmatrix} 2x_1^\star \\ 2x_2^\star \end{pmatrix} + z^\star \begin{pmatrix} 1 \\ 1 \end{pmatrix} &= 0 \\ z^\star(x_1^\star + x_2^\star - 1) &= 0. \end{aligned}$$

Example of KKT application

From the two first equalities $x_1^\star = x_2^\star = -z^\star/2$
replace in the third one leads to
either $z^\star = 0$ then $x_1^\star = x_2^\star = 0$
either $x_1^\star = x_2^\star = 1/2$ then $z^\star = -1 < 0$ impossible.

Modified example

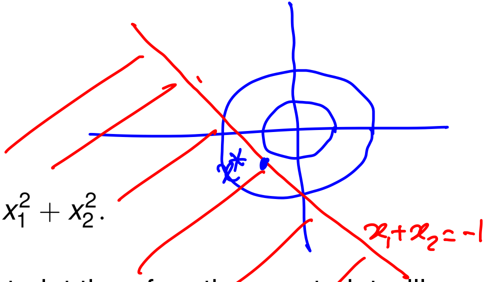
$x_1 = x_2 = -3/2$

$x_1 + x_2 = -1$

$-\frac{3}{2} - \frac{3}{2} = -1 \quad 3 = 1$

$x_1 = x_2 = -1/2$

$$\inf_{x_1+x_2+1\leq 0} x_1^2 + x_2^2.$$



$(0,0)$ does not satisfy the constraint therefore the constraint will be active in $x^\star$.

The third KKT condition is now $z^\star(x_1^\star + x_2^\star + 1) = 0 \Rightarrow$
either $z^\star = 0$ then $x_1^\star = x_2^\star = 0$, does not satisfy the constraint
either $x_1^\star = x_2^\star = -1/2$ then $z^\star = 1$, correct solution.
Verification : change of variable $x_2 = -1 - x_1$ in the function :
$\inf_{x_1} x_1^2 + (1 + x_1)^2$ is attained at $x_1 = -1/2$.

Exemple Kepler's problem

Find the parallelepiped of maximum volume inscribed in the ellipsoid

$$\mathcal{E} = \{x \in \mathbb{R}^3, {x_1}^2/a_1^2 + x_2^2/a_2^2 + x_3^2/a_3^2 = 1\}$$

Write the problem as a canonical optimisation problem

Exemple Kepler's problem

Find the parallelepiped of maximum volume inscribed in the ellipsoid

$$\mathcal{E} = \{x \in \mathbb{R}^3, {x_1}^2/a_1^2 + x_2^2/a_2^2 + x_3^2/a_3^2 = 1\}$$

Write the problem as a canonical optimisation problem

$$\inf_{\substack{{x_1}^2/a_1^2 + x_2^2/a_2^2 + x_3^2/a_3^2 = 1 \\ x_1 \geq 0 \\ x_2 \geq 0 \\ x_3 \geq 0}} f(x), \quad f(x) = -\prod_{i=1}^{3} x_i$$

- Can we apply KKT theorem?
- Which constraints are active ?
- Lagrangian ?

Exemple Kepler's problem

$x_1 = 0$ or $x_2 = 0$ or $x_3 = 0 \Rightarrow f(x) = 0$! inequality constraints are inactive

$$\ell(x, y) = -\prod_{i=1}^{3} x_i + y({x_1}^2/a_1^2 + x_2^2/a_2^2 + x_3^2/a_3^2 - 1)$$

Gradient

Exemple Kepler's problem

$$\begin{aligned}
-x_2x_3 + 2yx_1/a_1^2 &= 0\\
-x_1x_3 + 2yx_2/a_2^2 &= 0\\
-x_2x_1 + 2yx_3/a_3^2 &= 0
\end{aligned}$$

Exemple Kepler's problem

$$
\begin{aligned}
-x_2 x_3 + 2y x_1 / a_1^2 &= 0 \\
-x_1 x_3 + 2y x_2 / a_2^2 &= 0 \\
-x_2 x_1 + 2y x_3 / a_3^2 &= 0
\end{aligned}
$$

We multiply each equation by the corresponding x_i component

$$
\begin{aligned}
-x_2 x_3 x_1 + 2y x_1^2 / a_1^2 &= 0 \\
-x_1 x_3 x_2 + 2y x_2^2 / a_2^2 &= 0 \\
-x_2 x_1 x_3 + 2y x_3^2 / a_3^2 &= 0
\end{aligned}
$$

Then sum $-3 \prod x_i + 2y = 0$ from which $x_2 x_3 = 2y/(3x_1)$ Then inserting in the first equation leads to

$$
-2y/(3x_1) + 2y x_1 / a_1^2 = 2y(x_1/a_1^2 - 1/(3x_1))
$$

Exemple Kepler's problem

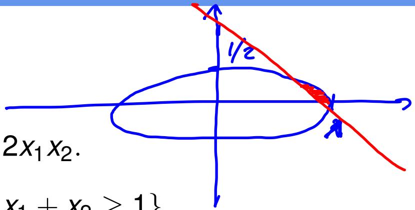
$x_1 = 0$ is impossible (or $\prod x_i = 0$ therefore $3x_1^2 = a_1^2$. We obtain similarly

$$x_i = \frac{a_i}{\sqrt{3}}, \quad i = 1, \ldots 3$$

Finally we get

$$Vol = 8 \prod x_i = 8 \frac{\prod x_i}{3\sqrt{3}}, \quad y = \frac{3 \prod x_i}{2} = \frac{\prod x_i}{2\sqrt{3}}$$

Exemple



Let the function

$$f(x) = 2x_1^2 + 3x_2^2 + 2x_1x_2.$$

and domain $\mathcal{D}_a$ $K = \{x_1^2 + 4x_2^2 \leq 1 \text{ and } x_1 + x_2 \geq 1\}$

Seek $\inf\limits_{x \in K} f(x)$

1. Show that f has a global minimum over $\mathbb{R}^2$ and calculate it
2. Show that K is nonempty convex.
3. Draw K in solid lines and isovalues of $f(x)$ in dotted lines
4. Write the Lagrangian for the problem (P).
5. Write the 1st order optimality conditions (or KKT conditions).
6. Use the drawing and the result of question 1 to decide which of the two constraints will be active at least.
7. Calculate $x^\star$ and the associated Lagrange multipliers.

Example

1. Compute $\nabla f(x) = (4x_1 + 2x_2, 2x_1 + 6x_2)^T$ and $Hf(x) = \begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix}$. The Hessian eigenvalues are $(5 \pm \sqrt{5})/2$. Hence

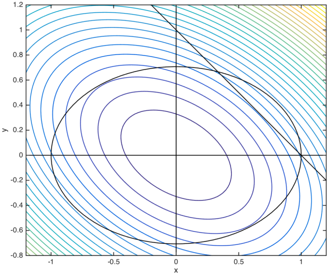
$$f(x) = f(0) + \nabla f(0) \cdot x + \frac{1}{2} Hf(0) x \cdot x$$

is a quadratic form with a symmetric positive definite matrix, so the minimum 0 is reached in $x = (0, 0)$ and unique.

Example

1. Consider $K = \{x_1^2 + 4x_2^2 \leq 1 \text{ and } x_1 + x_2 \geq 1\}$. The point $(1, 0) \in K \neq \emptyset$.
   The set $\{x_1^2 + 4x_2^2 \leq 1\}$ is the ellipse with semi-major axis $x_1 \in [-1, 1]$, $x_2 = 0$ and semi-minor axis $x_1 = 0$, $x_2 \in [-1/2, 1/2]$. Thus, it is convex. The half-plane $\{x_1 + x_2 \geq 1\}$ is also convex. The non-empty intersection of two convex sets is convex.

Example



Example

1. We apply the method of Lagrange multipliers:

$$\ell(x,y) = f(x) + y_1(x_1^2 + 4x_2^2 - 1) + y_2(1 - x_1 - x_2)$$

Example

1. If $(x^\star, y^\star)$ minimizes f on K, there exist $y_1 \geq 0$ and $y_2 \geq 0$ such that

$$
\begin{aligned}
\nabla_x \ell(x^\star, y^\star) &= \begin{pmatrix} 4x_1^\star + 2x_2^\star + 2y_1^\star x_1^\star - y_2^\star \\ 2x_1^\star + 6x_2^\star + 8y_1^\star x_2^\star - y_2^\star \end{pmatrix} = 0_{\mathbb{R}^2}, \\
y_1^\star((x_1^\star)^2 + 4(x_2^\star)^2 - 1) &= 0, \\
y_2^\star(1 - x_1^\star - x_2^\star) &= 0
\end{aligned}
$$

2 complementary relaxation conditions

4 cases

	$y_1=0$	$x_1, x_2 \in$ ellipse
$y_2=0$		
$x_1,x_2 \in$ line		

Example

Since the global minimum of $f(x)$ does not belong to K, at least one of the two constraints must be active. Given the convexity of f, we anticipate that the minimum will be reached on the segment $\{x_1 + x_2 = 1\} \cap K$ and the maximum on the arc $\{x_1^2 + 4x_2^2 = 1\} \cap K$.
For the minimum, we seek x, $y_2 > 0$ such that

$$\begin{aligned} 4x + 2y &= y_1^\star \\ 2x + 6y &= y_2^\star \\ x^2 + 4y^2 &< 1 \\ x + y &= 1 \end{aligned}$$

which leads to $x = \frac{2}{3}$, $y = \frac{1}{3}$, $y_2^\star = \frac{10}{3}$, $y_1^\star = 0$ and $f(x^\star) = \frac{5}{3}$.

Algorithms for constrained optimization

$$\inf_{\begin{cases} c^E(x) = 0 \\ c^I(x) \leq 0 \\ x \in \mathbb{R}^n \end{cases}} f(x) \qquad \text{with} \qquad \begin{cases} f : \mathbb{R}^n \longrightarrow \mathbb{R}, \\ c^E : \mathbb{R}^n \longrightarrow \mathbb{R}^m, \\ c^I : \mathbb{R}^n \longrightarrow \mathbb{R}^p, \\ f, c, \ \text{smooth.} \end{cases}$$

- Change of unknowns
- Projection
- Penalisation
- Methods using Lagrange multipliers

Change of unknowns

$\min_{(\mathbb{R}^+)^n} f(x) = \min_{y \in \mathbb{R}^n} F(y)$

with $F(y) = f(y^2)$

diffeomorphism $\varphi : \mathbb{R}^n \to K \subset \mathbb{R}^n$

Examples:

abuse of notation

- $K = (\mathbb{R}+)^n \longrightarrow$ set $x = y^2$ and optimize without constraints with respect to y.

  $\varphi : (\mathbb{R}^*)^n \to (\mathbb{R}^{+*})^n$, $x_i = \varphi_i(y) = y_i^2$

$F(y^*) = \min F(y)$

y^2 is impossible

$x_i^* = y_i^{*2} \iff x^* = y^{*2}$

not the matrix multiplication

Change of unknowns

example for $f : \mathbb{R} \to \mathbb{R}$
$D_a = [a, b]$ $a \leq x \leq b$

$$x = \frac{b-a}{2} \cos\theta + \frac{a+b}{2} \qquad \theta : \mathbb{R}_{\setminus 2\pi}$$

diffeomorphism $\varphi : \mathbb{R}^n \to K \subset \mathbb{R}^n$

Examples:

- $K = (\mathbb{R}+)^n \longrightarrow$ set $x = y^2$ and optimize without constraints with respect to y.
  $\varphi : (\mathbb{R}^*)^n \to (\mathbb{R}^{+*})^n$, $x_i = \varphi_i(y) = y_i^2$
- $K = \prod_{i=1,\dots,n}[a_i, b_i] \longrightarrow$ set $x_i = \frac{a_i+b_i}{2} + \frac{b_i-a_i}{2} cos\theta_i$ and optimize without constraints with respect to θ_i
  $\varphi : \mathbb{R}^n \to K$, $x_i = \varphi_i(\theta_i)$

Consequence of the change of unknowns on the calculation of the gradient

φ diffeomorphism $\Rightarrow J\varphi$ defined on $\varphi^{-1}(\mathring{K})$

Consequence of the change of unknowns on the calculation of the gradient

φ diffeomorphism $\Rightarrow J\varphi$ defined on $\varphi^{-1}(\overset{\circ}{K})$
$\tilde{f}(y) = fo\varphi(y) = f(x)$

Consequence of the change of unknowns on the calculation of the gradient

φ diffeomorphism $\Rightarrow J\varphi$ defined on $\varphi^{-1}(\overset{\circ}{K})$
$\tilde{f}(y) = fo\varphi(y) = f(x)$ $\nabla_y \tilde{f}(y) = J(y)\nabla_x f(x)$ with $J(x)$ jacobian matrix of φ

Example of change of unknowns

Example $\varphi : (\mathbb{R}^*)^n \to (\mathbb{R}^{+*})^n$, $x_i = \varphi_i(y) = y_i^2$

Example of change of unknowns

Example $\varphi : (\mathbb{R}^*)^n \to (\mathbb{R}^{+*})^n$, $x_i = \varphi_i(y) = y_i^2$

$$J(y) \quad = \quad \begin{pmatrix} 2y_1 & 0 & \cdots & 0 \\ 0 & 2y_2 & \cdots & 0 \\ \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2y_n \end{pmatrix} = \begin{pmatrix} 2\sqrt{x_1} & 0 & \cdots & 0 \\ 0 & 2\sqrt{x_2} & \cdots & 0 \\ \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2\sqrt{x_n} \end{pmatrix},$$

Example of change of unknowns

Example $\varphi : (\mathbb{R}^*)^n \to (\mathbb{R}^{+*})^n$, $x_i = \varphi_i(y) = y_i^2$

$$J(y) = \begin{pmatrix} 2y_1 & 0 & \cdots & 0 \\ 0 & 2y_2 & \cdots & 0 \\ \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2y_n \end{pmatrix} = \begin{pmatrix} 2\sqrt{x_1} & 0 & \cdots & 0 \\ 0 & 2\sqrt{x_2} & \cdots & 0 \\ \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 2\sqrt{x_n} \end{pmatrix},$$

$$\begin{aligned} \tilde{f}(y) &= f(y_1^2, \dots, y_n^2), \quad \nabla_y \tilde{f}(y) = J(y) \nabla f(x) \\ \frac{\partial \tilde{f}(y)}{\partial y_i} &= 2y_i \frac{\partial f}{\partial x_i}(y_1^2, \dots, y_n^2) \end{aligned}$$

Projection and projected gradient methods

Inequality constraints $\Leftrightarrow$ Belong to a non empty closed convex set $K \subset \mathbb{R}^n$.

$$\inf_{x \in K} f(x)$$

Projection and projected gradient methods

Inequality constraints $\Leftrightarrow$ Belong to a non empty closed convex set $K \subset \mathbb{R}^n$.

$$\inf_{x \in K} f(x)$$

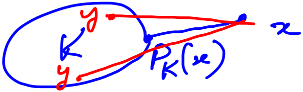
Theorem

Necessary local optimality condition
Let $f : \mathbb{R}^n \to \mathbb{R}$ a differentiable function and K a convex non empty subset of $\mathbb{R}^n$. Let $x^\star$ a local minimizer of f in K then

$$\boxed{\nabla f(x^\star) \cdot (x - x^\star) \geq 0, \forall x \in K}$$

If moreover f is convex the condition becomes necessary, ie any point $x^\star$ satisfying it is a global minimum of f on K.

Definition/proposition of the projection on a convex set



Let K a non empty closed convex set in $\mathbb{R}^n$. The projection of a point $x \in \mathbb{R}^n$ on K, denoted $P_K(x)$, is defined as the unique solution

$$\|x - P_K(x)\| = \inf_{y\in K} \|x - y\|_2^2.$$

$$f(y) = \|y - x\|^2$$

$$Hf(y) = 2\,Id$$

$$\left(P_K(x) - x, \left(y - P_K(x)\right)\right) \geq 0$$

Definition/proposition of the projection on a convex set

$f_x(y) = \|y - x\|^2$ $\quad \nabla f_x(y) = 2(y - x)$

Let K a non empty closed convex set in $\mathbb{R}^n$. The projection of a point $x \in \mathbb{R}^n$ on K, denoted $P_K(x)$, is defined as the unique solution

$$\inf_{y \in K} \|x - y\|_2^2.$$

Furthermore $P_K(x)$ is the only point in K satisfying

$$\boxed{(P_K(x) - x) \cdot (y - P_K(x)) \geq 0,} \quad \forall y \in K$$

Proof

- Existence of $P_K(x)$ as minimum on a closed set of $f(y) = \|x - y\|^2$ coercive (Theorem 1.1)
-
-

Proof

- Existence of $P_K(x)$ as minimum on a closed set of $f(y) = \|x - y\|^2$ coercive (Theorem 1.1)
- Unicity : $f(y)$ is convex (Hessien $=2I_{\mathbb{R}^n}$) therefore unique minimum
-

Proof

- Existence of $P_K(x)$ as minimum on a closed set of $f(y) = \|x - y\|^2$ coercive (Theorem 1.1)
- Unicity : $f(y)$ is convex (Hessien $=2I_{\mathbb{R}^n}$) therefore unique minimum
- Equivalence between the two properties:
  $\Rightarrow$ for $y \in K$ and $\theta \in [0, 1]$, $\theta y + (1 - \theta)P_K(x) \in K$
  - Develop $\|x - P_K(x)\|^2 \leq \|\theta y + (1 - \theta)P_K(x) - x\|^2$
  - Simplify θ
  - Do $\theta = 0$

Proof

- Existence of $P_K(x)$ as minimum on a closed set of $f(y) = \|x - y\|^2$ coercive (Theorem 1.1)
- Unicity : $f(y)$ is convex (Hessien $=2I_{\mathbb{R}^n}$) therefore unique minimum
- Equivalence between the two properties:
  $\Rightarrow$ for $y \in K$ and $\theta \in [0,1]$, $\theta y + (1-\theta)P_K(x) \in K$
  - Develop $\|x - P_K(x)\|^2 \leq \|\theta y + (1-\theta)P_K(x) - x\|^2$
  - Simplify θ
  - Do $\theta = 0$

  $\Leftarrow$ for $y \in K$ develop

$$
\begin{aligned}
\|x - y\|^2 &= \|x - P_K(x) + P_K(x) - y\|^2 \\
&= \|x - P_K(x)\|^2 + \|P_K(x) - y\|^2 + 2(x - P_K(x)) \cdot (P_K(x) - y) \\
&\geq \|x - P_K(x)\|^2
\end{aligned}
$$

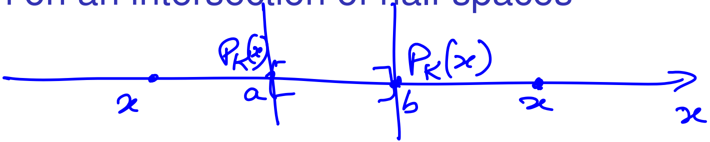
Particular cases of projection on a convex.

Orthogonal projection of x on a subset K

$$P_K(x) := \underset{y \in K}{\operatorname{argmin}} \|x - y\|.$$

Specific cases that can be solved explicitly

Projection on an intersection of half spaces



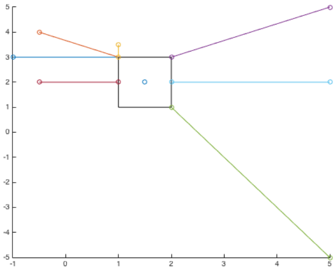
Assume $K = \{x \in \mathbb{R}^n, x_i \geq a_i, i \in I, x_j \leq b_j, j \in J\}$ with $I, J \subset \{1, \dots, n\}$. Then

$$P_K(x)_i = \begin{cases} \max(a_i, x_i), \text{ for } i \in I \backslash J \\ \min(b_i, x_i), \text{ for } i \in J \backslash I \\ \min(b_i, \max(a_i, x_i)), \text{ for } i \in I \cap J \end{cases}$$

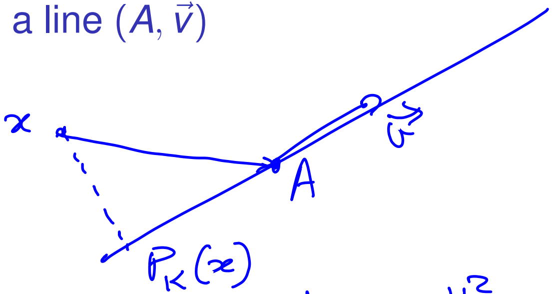
Projection on an intersection of half spaces

Assume $K = \{x \in \mathbb{R}^n, x_i \geq a_i, i \in I, x_j \leq b_j, j \in J\}$ with $I, J \subset \{1, \dots, n\}$. Then

$$P_K(x)_i = \begin{cases} \max(a_i, x_i), \text{ for } i \in I \backslash J \\ \min(b_i, x_i), \text{ for } i \in J \backslash I \\ \min(b_i, \max(a_i, x_i)), \text{ for } i \in I \cap J \end{cases}$$



Projection on a line $(A, \vec{v})$



$$\|P_K(x) - x\|^2 = \min_{y \in \text{line}} \|y - x\|^2 = \min_{\lambda \in \mathbb{R}} \|A - x + \lambda \vec{v}\|^2$$

$$y = A + \lambda \vec{v}$$

$$= \min_{\lambda} \|\vec{xA} + \lambda \vec{v}\|^2$$

$$= \|\vec{xA}\| + \lambda^2 \|\vec{v}\| + 2\lambda\, \vec{xA} \cdot \vec{v}$$

Projection on a line $(A, \vec{v})$

$$f(\lambda) = \|\overrightarrow{xA}\|^2 + \lambda^2 \|\vec{v}\|^2 + 2\lambda\, \vec{v} \cdot \overrightarrow{xA}$$

$$f'(\lambda) = 2\lambda \|\vec{v}\|^2 + 2\, \vec{v} \cdot \overrightarrow{xA} = 0$$

when $\lambda = -\dfrac{\vec{v} \cdot \overrightarrow{xA}}{\|v\|^2}$

$$P_K(x) = A + \lambda \vec{v} = A - \frac{\vec{v} \cdot \overrightarrow{xA}}{\|\vec{v}\|^2} \vec{v}$$

$$\overrightarrow{AP_K(x)} = -\frac{\vec{v} \cdot \overrightarrow{xA}}{\|v\|^2} \vec{v}$$

Euler inequation

the orthogonal projection of x on K exists and is unique.

Let $f : K \subset E \to \mathbb{R}$, where K is a convex included in E, a Hilbert space. We suppose that f is differentiable in $x^\star \in K$. If $x^\star$ is a local minimum of f over K, then $x^\star$ satisfies the Euler inequality:

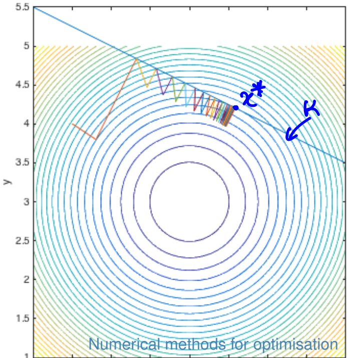
$$Df(x^\star)(y - x^\star) \geq 0, \forall y \in K.$$

Algorithm of the projected gradient

Data: Function f, convex K, step $(\alpha_k)_{k\geq 0}$, tolerance τ, max number of iterations $k_{\max}$
Result: $\min_{x\in K} f(x)$
Initialisation : choice of $x_0 \in \mathbb{R}^n$
while $\|x^{k+1} - x^k\| \geq \tau$ *and* $k < k_{\max}$ **do**
  Solve $x^{k+1} = P_K(x^k - \alpha_k \nabla f(x^k))$
  $k \leftarrow k+1$
end
$x^\star = x_k$

standard gradient

$x^{k+1} = x^k - \alpha_k \nabla f(x_k)$



Convergence of projected gradient algorithm

$\nabla f(x^*).(y-x^*) \geq 0$ for $y \in K$

Theorem

Let f differentiable sur $\mathbb{R}^n$ and $K \in \mathbb{R}^n$ closed non empty convex subset. Denote by x_k the current solution of the projected gradient algorithm and

$x_{k+1} = x_k + d(\alpha)$ with

$$d(\alpha) = P_K(x_k - \alpha \nabla f(x_k)) - x_k$$

If $d(\alpha) \neq 0$ then $d(\alpha)$ is a descent direction $\forall \alpha > 0$

$d(\alpha).\nabla f(x_k) < 0$

$\tilde{x}_{k+1} = x_k - \alpha \nabla f(x_k)$

$x_{k+1} = P_K(\tilde{x}_{k+1})$

Proof

Let $\alpha > 0$ fixed. Suppose : $d(\alpha) = p_K(x_k - \alpha\nabla f(x_k)) - x_k \neq 0$.
$d(\alpha)$ descent direction f in $x_k \Leftrightarrow \nabla f(x_k) \cdot d(\alpha) < 0$

Proof

Let $\alpha > 0$ fixed. Suppose : $d(\alpha) = p_K(x_k - \alpha\nabla f(x_k)) - x_k \neq 0$.
$d(\alpha)$ descent direction f in $x_k \Leftrightarrow \nabla f(x_k) \cdot d(\alpha) < 0$
Indeed $\forall y \in K$, $\Leftrightarrow$
$(P_k(x_k - \alpha\nabla f(x_k)) - (x_k - \alpha\nabla f(x_k))) \cdot (y - P_K(x_k - \alpha\nabla f(x_k))) \geq 0$
Therefore for all $y \in K$ $(d(\alpha) + \alpha\nabla f(x_k))) \cdot (y - x_k - d(\alpha)) \geq 0$

Proof

Let $\alpha > 0$ fixed. Suppose : $d(\alpha) = p_K(x_k - \alpha\nabla f(x_k)) - x_k \neq 0$.
$d(\alpha)$ descent direction f in $x_k \Leftrightarrow \nabla f(x_k) \cdot d(\alpha) < 0$
Indeed $\forall y \in K$, $\Leftrightarrow$
$(P_k(x_k - \alpha\nabla f(x_k)) - (x_k - \alpha\nabla f(x_k))) \cdot (y - P_K(x_k - \alpha\nabla f(x_k))) \geq 0$
Therefore for all $y \in K$ $(d(\alpha) + \alpha\nabla f(x_k))) \cdot (y - x_k - d(\alpha)) \geq 0$
Since $x_k \in K$ choose $y = x_k$ on a
$(d(\alpha) + \alpha\nabla f(x_k)) \cdot d(\alpha) \leq 0$ or $\alpha\nabla f(x_k) \cdot d(\alpha) \leq -\|d(\alpha)\|^2 < 0$

Convergence of projected gradient algorithm

Theorem

If f is differentiable, α-elliptic and its gradient is C-Lipschitzien, the projected gradient algorithm converges towards $x^\star$ when $k \to \infty$ for $(\alpha_k)_{k\geq 0}$ sufficiently small: $\alpha_k \leq \dfrac{2\alpha}{C^2}$ where α and C are two constants such that

$$
\begin{aligned}
(\nabla f(x) - \nabla f(y), x - y) &\geq \alpha \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n, \\
\|\nabla f(x) - \nabla f(y)\| &\leq C \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.
\end{aligned}
$$

Proof

Compose two contracting applications

- $x \to x - \alpha \nabla f(x)$ with conditions on α_k (See unconstrained gradient convergence theorem)

Proof

Compose two contracting applications

- $x \to x - \alpha\nabla f(x)$ with conditions on α_k (See unconstrained gradient convergence theorem)
- $x \to P_K(x)$. indeed

$$
\begin{aligned}
\|x-y\|^2 &= \|x-y-(P_K(x)-P_K(y))+(P_K(x)-P_K(y))\|^2 \\
&= \|x-y-(P_K(x)-P_K(y))\|^2 + \\
&\quad (x-y-(P_K(x)-P_K(y)))\cdot(P_K(x)-P_K(y))+\|P_K(x)-P_K(y)\|^2 \\
&= \|x-y-(P_K(x)-P_K(y))\|^2 \\
&\quad +(x-P_K(x))\cdot(P_K(x)-P_K(y)) \\
&\quad +(-y+P_K(y)))\cdot(P_K(x)-P_K(y))+\|P_K(x)-P_K(y)\|^2
\end{aligned}
$$

but $(x-P_K(x))\cdot(vP_K(x)-P_K(y))\geq 0$ car $P_K(y)\in K$
Therefore

$$
\begin{aligned}
\|x-y\|^2 &\geq \|x-y-(P_K(x)-P_K(y))\|^2+\|P_K(x)-P_K(y)\|^2 \\
&\geq \|P_K(x)-P_K(y)\|^2
\end{aligned}
$$

Hands on. - notebook 6

Real Problem
minimize $f(x)$ where
x depends on the solution of
a PDE, ODE

Projection on a triangle

Penalization method

Idea:

- Replace the constrained problem by an unconstrained one

Penalization method

Idea:

- Replace the constrained problem by an unconstrained one
- The new function to be minimized

Penalization method

Idea:

- Replace the constrained problem by an unconstrained one
- The new function to be minimized
  - takes large values when the constraints are not satisfied,

Penalization method

Idea:

- Replace the constrained problem by an unconstrained one
- The new function to be minimized
  - takes large values when the constraints are not satisfied,
  - is similar to the original function when the constraints are satisfied.

Penalization method

Idea:

- Replace the constrained problem by an unconstrained one
- The new function to be minimized
  - takes large values when the constraints are not satisfied,
  - is "similar" to the original function when the constraints are satisfied.
- Find a penalization function $p(x)$ such that

$$(P_K) \qquad \min_{x \in K} f(x), \quad = f(x^*)$$

with $K \subset \mathbb{R}^n$ convex and

$$(P_{K_\varepsilon}) \qquad \min_{x \in \mathbb{R}^n} \Theta_\varepsilon(x), \quad \text{with} \quad \Theta_\varepsilon(x) = f(x) + \frac{1}{\varepsilon} p(x) \quad = \Theta_\varepsilon(x^*)$$

are equivalent (exact penalization)

Exact penalization method

Definition

A penalisation function $p(x)$ associated to the problem (P_K) is exact if any solution of (P_K) minimizes Θ_ε.

Exact penalization method

Definition

A penalisation function $p(x)$ associated to the problem (P_K) is exact if any solution of (P_K) minimizes Θ_ε.

For instance an exact penalization function could be

$$p(x) = \begin{cases} 0 \text{ if } x \in K \\ +\infty \text{ if } x \notin K \end{cases}$$

exact OK but not smooth...

Inexact penalization method

Interior Penalization : $x_\varepsilon^\star \in K$

Example logarithmic penalization $p(x) = -\log x$, pour $x \in \mathbb{R}^+$.

$$\Theta_\varepsilon(x) = f(x) - \varepsilon \log(x).$$

example for
$K = \mathbb{R}^+$

Inexact penalization method

Interior Penalization : $x_\varepsilon^\star \in K$

Example logarithmic penalization $p(x) = -\log x$, pour $x \in \mathbb{R}^+$.

$$\Theta_\varepsilon(x) = f(x) - \varepsilon \log(x).$$

General case with a constraint $C^I(x) \leq 0$ in $\mathbb{R}^p$

$$\Theta_\varepsilon(x) = f(x) - \varepsilon \sum_{i=1}^{p} \log(-c_i^I(x))$$

Inexact penalization method

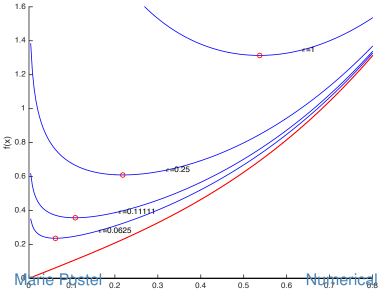
Interior Penalization : $x_\varepsilon^\star \in K$

Example logarithmic penalization $p(x) = -\log x$, pour $x \in \mathbb{R}^+$.

$$\Theta_\varepsilon(x) = f(x) - \varepsilon \log(x).$$

General case with a constraint $C^I(x) \leq 0$ in $\mathbb{R}^p$

$$\Theta_\varepsilon(x) = f(x) - \varepsilon \sum_{i=1}^{p} \log(-c_i^I(x))$$



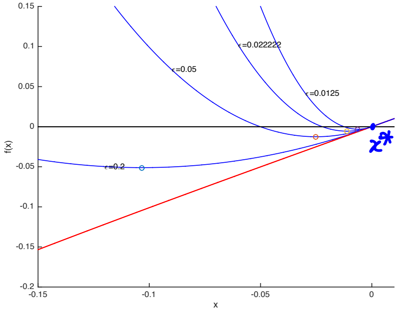
$\Theta_\varepsilon(x) = x + x^3 - \varepsilon \log(x)$

Minimisation of
$f(x) = x + x^3$ on $\mathbb{R}^+$
$f(x) + \varepsilon p_{int}(x)$
with $p_{int} = -\log x$

Inexact penalization method

Exterior Penalization : $x^\star_\varepsilon \to x^\star$ with $x^\star_\varepsilon \notin K$ for $\varepsilon \neq 0$.

- if $K = {\mathbb{R}^+}^n$ $p(x) = \|x^-\|^2$, where x^- is the vector of negative parts of components of x

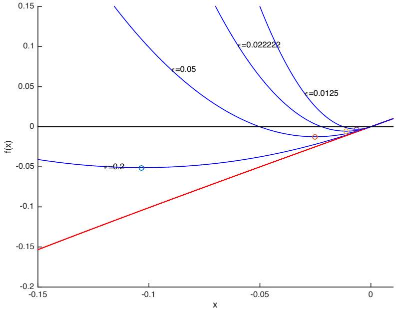


Minimization of
$f(x) = x + x^3$ on $\mathbb{R}^+$
$f(x) + p_{ext}(x)/\varepsilon$
with $p_{ext} = (x^-)^2$

Inexact penalization method

Exterior Penalization : $x^\star_\varepsilon \to x^\star$ with $x^\star_\varepsilon \notin K$ for $\varepsilon \neq 0$.

- if $K = \mathbb{R}^{+n}$ $p(x) = \|x^-\|^2$, where x^- is the vector of negative parts of components of x
- if $K = \{x, c(x) = 0\}$, $p(x) = \|c(x)\|^2$

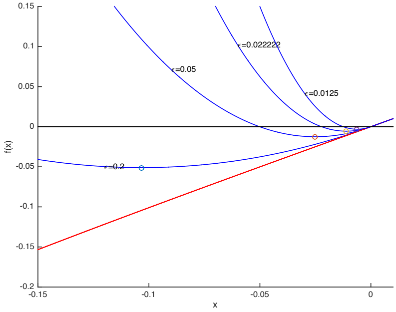


Minimization of
$f(x) = x + x^3$ on $\mathbb{R}^+$
$f(x) + p_{ext}(x)/\varepsilon$
with $p_{ext} = (x^-)^2$

Inexact penalization method

Exterior Penalization : $x^\star_\varepsilon \to x^\star$ with $x^\star_\varepsilon \notin K$ for $\varepsilon \neq 0$.

- if $K = {\mathbb{R}^+}^n$ $p(x) = \|x^-\|^2$, where x^- is the vector of negative parts of components of x
- if $K = \{x, c(x) = 0\}$, $p(x) = \|c(x)\|^2$
- if $K = \{x, c(x) \leq 0\}$, $p(x) = \|c(x)^+\|^2$



Minimization of
$f(x) = x + x^3$ on $\mathbb{R}^+$
$f(x) + p_{ext}(x)/\varepsilon$
with $p_{ext} = (x^-)^2$

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$.
Suppose $p(x)$ satisfies

1. *$p(x)$ continuous on $\mathbb{R}^n$,*

Then

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$.
Suppose $p(x)$ satisfies

1. $p(x)$ *continuous on* $\mathbb{R}^n$,
2. $\forall x \in \mathbb{R}^n, p(x) \geq 0$,

Then

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$.
Suppose $p(x)$ satisfies

1. $p(x)$ *continuous on* $\mathbb{R}^n$,
2. $\forall x \in \mathbb{R}^n, p(x) \geq 0$,
3. $p(x) = 0 \Leftrightarrow x \in K$.

Then

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$. Suppose $p(x)$ satisfies

1. $p(x)$ *continuous on* $\mathbb{R}^n$,
2. $\forall x \in \mathbb{R}^n, p(x) \geq 0$,
3. $p(x) = 0 \Leftrightarrow x \in K$.

Then

1. $\forall \varepsilon > 0$ *problem* (P_{K_ε}) *has at least one solution* $x_\varepsilon^\star$,

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$. Suppose $p(x)$ satisfies

1. *$p(x)$ continuous on $\mathbb{R}^n$,*
2. *$\forall x \in \mathbb{R}^n$, $p(x) \geq 0$,*
3. *$p(x) = 0 \Leftrightarrow x \in K$.*

Then

1. *$\forall \varepsilon > 0$ problem (P_{K_ε}) has at least one solution $x^\star_\varepsilon$,*
2. *Sequence $(x^\star_\varepsilon)_{\varepsilon>0}$ is bounded*

Convergence of inexact penalization methods

Theorem

Let f continuous and coercive on a closed set $K \in \mathbb{R}^n$.
Suppose $p(x)$ satisfies

1. $p(x)$ *continuous on* $\mathbb{R}^n$,
2. $\forall x \in \mathbb{R}^n$, $p(x) \geq 0$,
3. $p(x) = 0 \Leftrightarrow x \in K$.

Then

1. $\forall \varepsilon > 0$ *problem* (P_{K_ε}) *has at least one solution* $x^\star_\varepsilon$,
2. *Sequence* $(x^\star_\varepsilon)_{\varepsilon>0}$ *is bounded*
3. *All converging subsequence* $(x^\star_\varepsilon)_{\varepsilon>0}$ *converges towards a solution of* (P_K) *when* $\varepsilon \searrow 0$.

$x^*_\varepsilon \xrightarrow[\varepsilon \to 0]{} x^*$

Inexact penalization algorithm

Data: Function Θ_ε, tolerance τ, max number of iterations $k_{\max}$
Result: $\min_{x\in K} f(x)$
Initialisation : Choice of $x_0 \in \mathbb{R}^n$, $\varepsilon_0 > 0$
while $\|x_{k+1} - x_k\| \geq \tau$ *and* $k < k_{\max}$ **do**
  Solve $x_{k+1} = \min_{x\in\mathbb{R}^n} \Theta_\varepsilon(x)$ with x_k as initial value
  Choose $\varepsilon_{k+1} < \varepsilon_k$
  $k \leftarrow k + 1$
end
$x^\star = x_k$

tricky step how do you make ε diminish.

$\varepsilon_{k+1} = \tau\,\varepsilon_k$ with $\tau \in]0,1[$

Saddle point definition

$\max_{z \geq 0} g(z) = g(z^*) = d^* = p^*$

A "saddle point" of $\ell(x, z)$ is a point $(x^\star, z^\star) \in \mathbb{R}^n \times \mathbb{R}^{+p}$ s. t.

$$\forall x \in \mathbb{R}^n, \quad \forall z \in \mathbb{R}^{+p},$$
$$\ell(x^\star, z) \leq \ell(x^\star, z^\star) \leq \ell(x, z^\star).$$

For $f(x)$ and $c^I(x)$ convex, a saddle point $(x^\star, z^\star)$ of the Lagrangian

$$\ell(x, z) = f(x) + z \cdot c^I(x)$$

with $x, z \in \mathbb{R}^n \times \mathbb{R}^p_+$ satisfies KKT conditions.

$$\begin{cases} \nabla_x \ell(x^*, z^*) = 0 \\ z^*_j \cdot c^I_j(x^*) = 0 \\ c^I_j(x^*) \leq 0 \end{cases}$$

Principle of Uzawa's method $(z^\star - z)\cdot P($

By definition we have

$$\forall z \in \mathbb{R}^p_+ \quad \ell(x^\star, z) \leq \ell(x^\star, z^\star) \leq \ell(x, z^\star) \quad \forall x \in \mathbb{R}^n$$

from which $(P_K(x) - x)\cdot(y - P_K(x)) \geq 0$

$$\begin{aligned}
& f(x^\star) + z\cdot c^I(x^\star) \leq f(x^\star) + z^\star\cdot c^I(x^\star) \\
\Leftrightarrow \quad & (z^\star - z)\cdot c^I(x^\star) \geq 0 \\
\Leftrightarrow \quad & (z^\star - z)\cdot \lambda c^I(x^\star) \geq 0 \quad \forall \lambda > 0 \\
\Leftrightarrow \quad & (z^\star - z)\cdot (z^\star + \lambda c^I(x^\star) - z^\star) \geq 0 \quad \forall \lambda > 0
\end{aligned}$$

with $x = z + \lambda c^I(x^\star)$

which implies $(P_K(x) - z)\cdot(x - P_K(x)) \geq 0$

$$z^\star = \Pi_{\mathbb{R}^p_+}(z + \lambda c^I(x^\star)) \quad \forall \lambda > 0$$

Uzawa's method

$\ell(x,z) = f(x) + z \cdot c^T(x)$

a step of gradient "maximization" of $\ell(x,z)$ with respect to z

- choose $z_0 \in \mathbb{R}_+^p$ and $\lambda > 0$
- for all $k \geq 0$ set

$$\begin{aligned} \ell(x_k, z_k) &= \inf_{x \in \mathbb{R}^n} \ell(x, z_k) \\ z_{k+1} &= \Pi_{\mathbb{R}_+^p} \boxed{(z_k + \lambda c^I(x_k))} \end{aligned}$$

- $\Rightarrow$ 2 imbricated unconstrained minimization problems, plus a projection on $\mathbb{R}_+^p$

$z_{k+1} = z_k + \lambda \nabla_z \ell(x, z_k)$

Convergence of Uzawa's algorithm

Theorem : Let f α-elliptic and differentiable, $c^I : \mathbb{R}^n \to \mathbb{R}^p$ convex and C-Lipschitz. If the Lagrangian has a saddle point $(x^\star, z^\star)$ then

$$0 < \lambda < \frac{2\alpha}{C^2} \to x_n \to x^\star \text{ when } n \to \infty$$

Uzawa's algoritm

Scipy. optimize.

Data: Lagrangian $\ell(x, y, z)$ gradients with respect to x, y and z, dual step $\rho > 0$, tolerance ε

Result: Solution $x^\star, y^\star, z^\star$ of problem (P)

Initialisation : For $k = 0$, choose estimate x^0, y^0, z^0 and $\rho > 0$

repeat

(i) Seek x^{k+1} solution of primal unconstrained problem

$$\forall x \in \mathbb{R}^n, \quad \ell(x^{k+1}, y^k, z^k) \leq \ell(x, y^k, z^k)$$

(ii) Compute $y^{k+1} = y^k + \rho c^E(x^{k+1})$

(ii) Compute $z^{k+1} = \max(0, z^k + \rho c^I(x^{k+1}))$

$k \leftarrow k + 1$

until $\|x^k - x^{k-1}\| > \varepsilon$;

$x^\star = x_k$

$y^\star = y_k$

$z^\star = z_k$

Exercise 1

We consider the function of $\mathbb{R}^3$ in $\mathbb{R}$

$$f(x) = e^{x_2} + x_1 x_2 + x_1^2 - 2x_1 x_3 + x_3^2$$

and the constrained minimization problem

$$(P_{IE}) \quad x^* = \underset{\substack{C_E(x)=0 \\ C_I(x)\leq 0}}{\operatorname{argmin}} \quad f(x),$$

with

$$\begin{cases} C_I(x) & = & x_1^2 + x_2^2 + x_3^2 - 10, \\ C_E(x) & = & a, x - 2, \end{cases}$$

with $a \in \mathbb{R}^3$.
Determine a so that $x^* = (1, 0, 2)^T$ is a solution of (P_{IE}).

Exercise 2

A company manufactures two types of bicycles. Model X is sold for 500 euros the unit and model Y is sold for 1000 euros. Production costs monthly s élèvent à

$$c(x, y) = 5x^2 + 5y^2 - 2.5xy + 1000$$

where x (respectively y) denotes the number of bicycles of model X (resp. Y) manufactured. It is assumed that all the bicycles manufactured are sold immediately.

1. Give the monthly profit function $p(x, y)$
2. We assume that the production capacity is 150 bicycles per month, find the distribution between the two models allowing to maximize the profit.
3. Is profit maximum when maximum production capacity is achievement ?

Exercise 3

Let $A \in \mathbb{R}^{m\times n}$ be symmetric positive definite. We define

$$K = \{x \in \mathbb{R}^n, x, Ax \leq 1\}.$$

1. Show that K is convex. If A is an homothety (i.e. $A = \lambda Id$), what does K represent?
2. Show that for a suitable orthonormal basis $(e_i)_{1\leq i\leq n}$, there exist strictly positive coefficients $(\alpha_i)_{1\leq i\leq n}$ such that

$$K = \{\sum_{i=1}^{n} x_i e_i, \sum_{i=1}^{n} \alpha_i x_i^2 \leq 1\}.$$

   What does the set K represent geometrically?
3. Show that for any vector $x = \sum_{i=1}^{n} x_i e_i \in \mathbb{R}^n \backslash K$, there exists a unique $\lambda \in \mathbb{R}^+$ which we will note $lambda_x$ such that

$$\sum_{i=1}^{n} \alpha_i \frac{x_i^2}{(1+\lambda_x \alpha_i)^2} = 1$$

4. Show that the projection of x on K is defined by

$$P_K(x) = \sum_{i=1}^{n} \frac{x_i^2}{1 + \lambda_x \alpha_i} e_i$$

5. We now consider $f : \mathbb{R}^n \to \mathbb{R}$ δ-elliptic and M-Lipschitzian and the minimization problem $f(x^\star) = \min_{x \in K} f(x)$. Explain the projected gradient algorithm. For which choice of descent parameter α does it converge?
6. We choose $f(x) = \|x - y\|^2$ with $y \in \mathbb{R}^n$ fixed. What does $x^\star$ correspond to in this case?

Exercise 4 - Tartaglia's problem

Decompose the number 8 into the sum of two positive numbers p_1 and p_2 such that the product of their product by their difference is maximum

Example Tartaglia's problem

Decompose the number 8 into the sum of two positive numbers p_1 and p_2 such that the product of their product by their difference is maximum

Write the problem as a canonical optimisation problem

$$\inf_{\substack{p_1+p_2=8\\ p_1\geq 0\\ p_2\geq 0}} f(p_1,p_2) = -p_1p_2(p_1-p_2)$$

- Can we apply KKT theorem?

Example Tartaglia's problem

Decompose the number 8 into the sum of two positive numbers p_1 and p_2 such that the product of their product by their difference is maximum

Write the problem as a canonical optimisation problem

$$\inf_{\substack{p_1+p_2=8\\ p_1\geq 0\\ p_2\geq 0}} f(p_1,p_2) = -p_1 p_2 (p_1 - p_2)$$

- Can we apply KKT theorem?
- Which constraints are active ?

Example Tartaglia's problem

Decompose the number 8 into the sum of two positive numbers p_1 and p_2 such that the product of their product by their difference is maximum

Write the problem as a canonical optimisation problem

$$\inf_{\substack{p_1+p_2=8 \\ p_1 \geq 0 \\ p_2 \geq 0}} f(p_1, p_2) = -p_1 p_2 (p_1 - p_2)$$

- Can we apply KKT theorem?
- Which constraints are active ?
- Lagrangian ?

Example Tartaglia's problem

$$\ell(p_1, p_2, y, z_1, z_2) = -p_1 p_2 (p_1 - p_2) + y(p_1 + p_2 - 8) - z_1 p_1 - z_2 p_2$$

Inactive constraints $\Rightarrow z_1 = z_2 = 0$

$$\ell(p_1, p_2, y, z_1, z_2) = -p_1^2 p_2 + p_1 p_2^2 + y(p_1 + p_2 - 8)$$

Gradient

$$\nabla_p \ell(p_1, p_2, y, z_1, z_2) = \begin{pmatrix} -2p_2p_1 + p_2^2 + y \\ 2p_1p_2 - p_1^2 + y \end{pmatrix} \Rightarrow \begin{cases} -2p_2p_1 + p_2^2 + y = 0 \\ 2p_1p_2 - p_1^2 + y = 0 \end{cases}$$

Substract and get

$$p_2^2 + p_1^2 - 4p_1p_2 = (p_1 + p_2)^2 - 6p_1p_2 = 0$$

then, since $p_1 + p_2 = 8$, $p_1p_2 = 64/6 = 32/3$.
p_1, p_2 the roots of the polynomial

$$x^2 - 8x + \frac{32}{3}$$

from which

$$p_1 = 4 - \frac{4}{\sqrt{3}} \quad p_2 = 4 + \frac{4}{\sqrt{3}}$$