Chapter 1

Prerequisites

1.1 Norms

1.1.1 Norms of vectors

We will start this chapter with some definitions and results that will be used in the resolution of optimisation problems.

Definition 1. *Let X be a vector space. The function $||.|| : X \to \mathbb{R}$ is called norm if and only if*

- *For all $x \in X$ $||x|| = 0 \iff x = 0$,*
- *For all $k \in \mathbb{R}, x \in X$ $||kx|| = |k|||x||$,*
- *For all $x, y \in X$ $||x + y|| \leq ||x|| + ||y||$.*

The $(X, ||.||)$ is called a norm space.

The following properties can be proven as an exercice.

Proposition 1. *Let $(X, ||.||)$ be a norm space, then the following hold.*

- *For all $x \in X$, $||x|| \geq 0$,*
- *For all $x \in X$, $||x|| = |||-x||$,*
- *For all $x, y \in X$, $||x - y|| = |||y - x||$,*
- *For all $x, y \in X$ $\big|||x|| - ||y|||\big| \leq ||y \pm x||$*
- *For all $x_1, x_2, ..., x_n \in X$ $||x_1 + x_2 + ... + x_n|| \leq ||x_1|| + ||x_2|| + ... + ||x_n||$.*

Remark: $||.||$ is a continuous function.
Examples of norms.

1. In $X = \mathbb{R}^n$ we can define the following norms for $x = (x_1, ..., x_n)$
   (a) $||x||_1 = \sum_{i=1}^n |x_i|$,
   (b) $||x||_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$,
   Hint to prove that this is a norm use the Minkowski inequality in $\mathbb{R}^n$, $(\sum_{k=1}^n |x_k + y_k|^p)^{\frac{1}{p}} \leq (\sum_{k=1}^n |x_k|^p)^{\frac{1}{p}} + (\sum_{k=1}^n |y_k|^p)^{\frac{1}{p}}$, $p \geq 1$.
   (c) $||x||_\infty = max_{i=1,2,..,n}\{|x_i|\}$.

2. In $X = B(A)$ the space of bounded functions defined on A we define

$$\begin{aligned} ||.||_\infty : B(A) &\to \mathbb{R} \\ f &\mapsto sup\{|f(x)|, x \in A\} \end{aligned}$$

3. In $X = C([a, b])$ the space of continous functions on $[a, b]$ we can define

$$\begin{aligned} ||.||_\infty : C[a, b] &\to \mathbb{R} \\ f &\mapsto max\{|f(x)|, x \in [a, b]\} \end{aligned}$$

   or

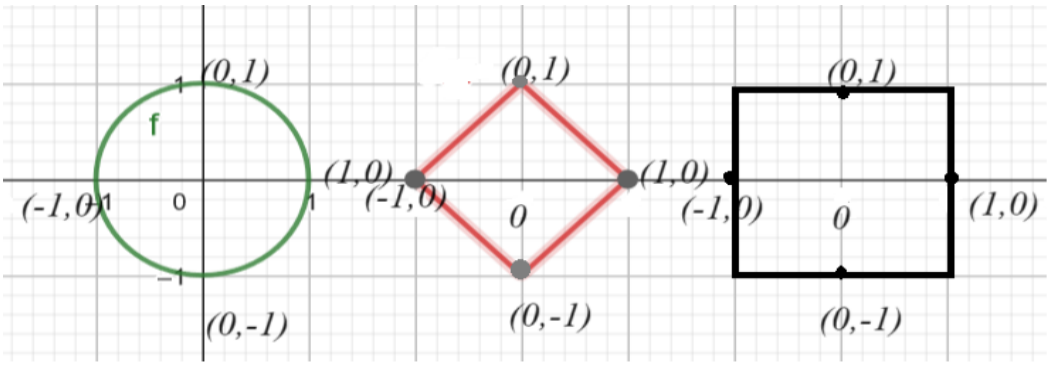
$$\begin{aligned} ||.||_1 : C[a, b] &\to \mathbb{R} \\ f &\mapsto \int_a^b |f(x)|dx \end{aligned}$$

Exercice In $X = \mathbb{R}^2$ plot $||x||_p = 1$ for $p = 1, 2, \infty$.

Proof. For $p = 2$ we have the usual euclidian norm so $||x||_2 = 1$ represents the unit circle. For $p = 1$ observe that $(1, 0), (-1, 0), (0, 1), (0, -1)$ always have norm 1. Then the equation

$$|x| + |y| = 1$$

represents a line. In the first quadrant it is $x + y = 1$, in the second $y - x = 1$, in the third $-x - y = 1$ and in the fourth $x - y = 1$. So $||x||_1 = 1$ is a diamond. Similarly for $p = \infty$ the points $(1, 0), (-1, 0), (0, 1), (0, -1)$ always have norm 1. The equation $||x||_\infty = 1$ represents the points for which at least one of the two coordinates is equal to 1. These points lie on the unit square. The curves are shown in the following figure,



Starting from $p = 1$ and changing its value progressively up to $p = \infty$ we observe that the diamond swells up to be a circle at $p = 2$ and it keeps swelling to a square at $p = \infty$. The advantage of using the $||.||_1$ norm in some optimization problems will be explained later. □

Example Let S be a symmetric positive definite matrix of $M_n(\mathbb{R})$. For $x \in \mathbb{R}^n$ a column vector

$$||x||_S = \sqrt{x^T S x}$$

is a norm in $\mathbb{R}^n$.

Exercice Prove that $||x||_S$ is a norm. Plot $||x||_S = 1$ for $x \in \mathbb{R}^2$ and $S = I_2$,$S = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$.

Commented solution. To verify that it is a norm we will need some properties of symmetric positive definite matrices which we state here:

- Since S is symmetric all its eigenvalues are real and its eigenvectors form an orthonormal basis of the ambient space.
- A square symmetric matrix is diagonalizable, it can then be written as $S = Q^T D Q$ where Q is an orthogonal matrix whose columns are the eigenvectors of S and D a diagonal matrix of the eigenvalues of S.
- S is positive definite so all its eigenvalues are positive.

- For every $x \neq 0$ $x^T S x > 0$.
- S is symmetric positive definite so it can be written in the form $B^T B$ where B has independent columns.

We are now ready to prove that $||.||_S$ is a norm.

- If $x = 0$ then trivially $||0||_S = 0$. If $||x||_S = 0$ then

$$x^T S x = 0 \implies x^T Q^T D Q x = 0 \implies (Qx)^T D Q x = 0.$$

  Since all the elements of D are positive and D is diagonal we deduce that $Qx = 0$ and since Q is invertible $x = 0$.

- With simple calculations we deduce that $||kx||_S = k^2 ||x||_S$ for any $k \in \mathbb{R}$ and with the squre root we have the result.

- The triangular inequality results from the observation that there is a matrix B with independent columns such that

$$||x||_S^2 = x^T S x = x^T B^T B x = (Bx)^T (Bx) = ||Bx||_2^2$$

  which is the euclidian norm and hence satisfies the triangular inequality. Consequently

$$||x + y||_S = ||B(x + y)||_2 \leq ||Bx||_2 + ||By||_2 = ||x||_S + ||y||_S.$$

  Note that using B the other properties can also be proven.

When $S = I_2$, $||x||_S = ||x||_2$ and when $S = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ it is a weighted norm whose unit ball has the shape of an ellipse given by the equation

$$2x^2 + 3y^3 = 1,$$

so $a^2 = \frac{1}{2}$ and $b^2 = \frac{1}{3}$.

1.1.2 Norm of a matrix

In this paragraph we will define norms on matrices, which will then be used in the definition of the condition number of a matrix.

Definition 2. *Consider $M_{n\times m}(\mathbb{R})$ be the vector space of matrices of size $n \times m$. A function $||.|| : M_{n\times m}(\mathbb{R}) \to \mathbb{R}$ is a norm if it satisfies the norm properties. Defined on the space of square matrices $M_n(\mathbb{R})$ a matrix norm must also satisfie*

$$||AB|| \leq ||A||||B||.$$

We say that the norm is sub-multiplicative.

Example The norm defined for $A \in M_{m\times n}(\mathbb{R})$ by

$$||A||_F = \sqrt{\sum_{1\leq i\leq m, 1\leq j\leq n} |A_{i,j}|^2}$$

is called the Frobenius norm. It measures the 'size' of a matrix since a matrix with small entries will have a small norm. Show that when A is square the Frobenius norm is sub-multiplicative. *Hint* Use the Cauchy-Schwartz inequality

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2$$

.

Subordinate norm

Consider $(X = \mathbb{R}^n, ||.||_X)$ and $(Y = \mathbb{R}^m, ||.||_Y)$ endowed with vector norms. Let A be an $n \times m$ matrix and x a column vector of $\mathbb{R}^n$. Then Ax is a column vector of $\mathbb{R}^m$. We want to measure the maximum blow up when A maps $x \in \mathbb{R}^n$ to $y = Ax \in \mathbb{R}^m$. We define thus

$$||A|| = inf\{c : ||Ax||_Y \leq c||x||_X, \forall x\} \tag{1.1}$$

Remarks

1. Observe that

$$||Ax||_Y \leq ||A||||x||_X$$

so $||A|| \in inf\{c : ||Ax||_Y \leq c||x||_X\}$.

2. If for all $x \neq 0$, $||Ax||_Y \leq c||x||_X \implies \frac{1}{||x||_X}||Ax||_Y \leq c$. Hence

$$||A(\frac{x}{||x||_X})||_Y \leq c,$$

which holds if and only if for all $w \in X$ with $||x||_X = 1$ we have that

$$||Aw||_Y \leq c.$$

If $||w||_X \leq 1$ then $||Aw||_Y \leq \frac{||Aw||_Y}{||w||_X} = ||A(\frac{w}{||w||_X})||_Y \leq c$. This results says that the image of the unit ball under A is included in a bounded ball of a potentially different radius. If $c > 1$ the ball may be blown up but the increase cannot be more than c.

Definition 3. $||A||$ *defined by* (1.1) *is called subordinate norm or induced norm on the space of* $M_{n\times m}(\mathbb{R})$ *matrices.*

Proposition 2.
- *Let $A \in M_{n\times m}(\mathbb{R})$. Then*

$$||A|| = sup_{x\in X, x\neq 0}\frac{||Ax||_Y}{||x||_X} = sup_{x\in X,||x||\leq 1}||Ax||_Y = sup_{x\in X,||x||=1}||Ax||.$$

- *The supremum in the definition of the subordinate norm is a maximum.*
- *The subordinate norm is sub-multiplicative. Let $A, B \in M_{n\times n}(\mathbb{R})$ be square matrices, then*

$$||AB|| \leq ||A||||B||.$$

Proof.
- The proof is left as an exercice.
- The function $g : \mathbb{R}^n \to \mathbb{R}$ with $g(x) = ||Ax||_Y$ is continuous. The unit shpere $\{x \in \mathbb{R}^n, ||x|| = 1\}$ is a compact set in $\mathbb{R}^n$ so g atteins its boundaries. There is x_0 such that

$$||Ax_0|| = ||A||.$$

- $$||ABx||_X \leq ||A||||Bx||_X \leq ||A||||B||||x||_X.$$

  This implies

$$||AB|| \leq ||A||||B||.$$

□

Proposition 3. $||A||$ *defined by* (1.1) *is a norm in* $M_{n\times m}(\mathbb{R})$.

Proof. Let A, B be two matrices in $M_{n\times m}(\mathbb{R})$, then $A + B \in M_{n\times m}(\mathbb{R})$. Let $x \in \mathbb{R}^n$ be a column vector.

$$||(A + B)x||_Y \leq ||Ax||_Y + ||Bx||_Y \leq ||A||||x||_X + ||B||||x||_Y \leq (||A|| + ||B||)||x||_X$$

so

$$||A + B|| \leq ||A|| + ||B||$$

$$||\lambda T|| = sup_{x\neq 0, x\in X}\frac{||\lambda Ax||}{||x||_X} = sup_{x\neq 0, x\in X}\frac{|\lambda|||Ax||}{||x||_X} = ||\lambda\frac{||Ax||_Y}{||x||_X} = |\lambda|||Ax||_Y.$$

If A is the zero matrix then $||A|| = 0$ since

$$||Ax||_Y = 0 = 0||x||_X$$

hence

$$||A|| \leq 0$$

If $||A|| = 0$ then for all $x \in \mathbb{R}^n$

$$||Ax||_X \leq 0||x||_X \implies ||Ax||_Y = 0,$$

so $Ax = 0$ for all x which implies that A is the zero matrix. □

Proposition 4. *Let A be a square matrix of size n.*

- *If $\mathbb{R}^n$ is endowed with the $||.||_\infty$ then*

$$||A|| = max_{i=1,..,n} \sum_{j=1}^{n} |a_{i,j}|,$$

in that case $||.||$ is also denoted as $||.||_\infty$.

- *If $\mathbb{R}^n$ is endowed with the $||.||_1$ then*

$$||A|| = max_{j=1,..,n} \sum_{i=1}^{n} |a_{i,j}|,$$

in that case $||.||$ is also denoted as $||.||_1$.

- *If $\mathbb{R}^n$ is endowed with the $||.||_2$ then*

$$||A|| = \sqrt{\rho(A^T A)}$$

where ρ stands for the spectral radius. In that case $||.||$ is also denoted as $||.||_2$.

1.1.3 Condition number

We are now ready to define the condition number of a matrix. Let's say that we want to solve

$$Ax = b$$

with $A \in M_n(\mathbb{R})$ invertible representing a data matrix. Errors can occur either due to inaccurate measurements of the elements of A and b or due to rounding these measurements while storage. We are interested in how the errors in A and b impact the accuracy of the solution x. The condition number will give us an estimate of whether the approximate solution is reasonable or not. Let $\delta_b \in \mathbb{R}^n$ and $\delta_A \in M(\mathbb{R})$ be the errors committed on b and A. We want to evaluate δ_x where $x + \delta_x$ is a solution of

$$(A + \delta_A)(x + \delta_x) = b + \delta_b.$$

If δ_A is not "very big" then $A + \delta_A$ is invertible and we can estimate δ_x in terms of δ_A and δ_B.

Definition 4. *Let $\mathbb{R}^n$ be endowed with a norm $||.||$ and $M_n(\mathbb{R})$ endowed with the subordinate norm $||.||_{sub}$ induced from $\mathbb{R}^n$. We call condition number of an invertible matrix A with respect to $||.||$ the positive real number*

$$cond(A) = ||A||_{sub}||A^{-1}||_{sub}.$$

Theorem 1. *(Caracterization of the condition number for the induced euclidean norm) Let $\mathbb{R}^n$ be endowed with the $||.||_2$ norm and the matrices $M_n(\mathbb{R})$ with the subordinate norm. $A \in M_n(\mathbb{R})$ invertible.*

- *If A is symmetric definite positive then*

$$cond(A) = \frac{\lambda_1}{\lambda_n}$$

where λ_1 is the largest and λ_n the smallest eigenvalue of A.

- *If A is invertible then A^TA is symmetric positive definite with largest and smallest eigenvalue σ_1 and σ_n. Then the condition number is*

$$cond(A) = \sqrt{\frac{\sigma_1}{\sigma_n}}.$$

Proof. It is left as an exercice. □

Let us denote the relative errors in A, b and x as

$$\epsilon(A) = \frac{||\delta A||_{sub}}{||A||_{sub}}, \quad \epsilon(b) = \frac{||\delta b||}{||b||}, \quad \epsilon(x) = \frac{||\delta x||}{||x||}$$

respectively.

Theorem 2. *Let $A \in M_n(\mathbb{R})$ be an invertible matrix, $b \in \mathbb{R}^n, b \neq 0$. $\mathbb{R}^n$ is endowed with a norm $||.||$ and $M_n(\mathbb{R})$ with the subordinate norm. Assume that $||\delta_A||||A^{-1}||_{sub} < 1$. Then $A + \delta_A$ is invertible and*

$$\epsilon(x) \leq \frac{cond(A)}{1 - cond(A)\epsilon(A)}(\epsilon(A) + \epsilon(b)).$$

If the condition number is not significantly larger than one, the matrix is well-conditioned, which means that its inverse can be computed with good accuracy. If the condition number is very large, then the matrix is said to be ill-conditioned.

Exercice

1. If A is invertible then $cond(A) \geq 1$ for any subordinate norm.
2. If A is invertible and $a \in \mathbb{R}^*$ then $cond(aA) = cond(A)$.
3. If Q is orthogonal $cond(Q) = 1$ in $||.||_2$ norm.
4. Prove theorem (2).

1.2 Rates of convergence

Most of the methods of resolution of numerical problems of optimization consist of trying to make a sequence converge towards a local extremum. Typically, we have an interative algorithm that is trying to find the extremum of a function f and we want an estimate of how long it will take to reach it. One way to compare algorithms is by using their rates of convergence. In practice the rates of convergence depend on how much information about the function we use. Algorithms that use little about f converge slowly.

Definition 5. *Let x_n be a sequence that converges to x_0 in norm $||.||$ of $\mathbb{R}^n$, that is $lim_{n\to+\infty}||x_n - x_0|| = 0$. We say that the convergence is linear, or of order, 1 if there is $a \in (0,1)$ and $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$*

$$||x_{k+1} - x_0|| \leq a||x_k - x_0||$$

The rate of linear convergence is r

$$r = limsup_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||},$$

which is the inifimum of all the values of a that satisfy the above inequality.
We say that the convergence is superlinear if

$$\lim_{k\to+\infty} \frac{||x_{k+1} - x_0||}{||x_k - x_0||} = 0$$

We say that the convergence is at least of oder β if there is an $a > 0$ and a $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$

$$||x_{k+1} - x_0|| \leq a||x_k - x_0||^\beta.$$

The order of convergence is defined as the upper bound of the values β that satisfy this inequality. For $\beta = 2, 3$ we refer to quadratic and cubic convergence.

Remark Note the following, which is kind of the inverse of the previous proposition. Assume $||x_n - x_0||$ converges to 0 and set

$$r = lim_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||}.$$

- If the limit exists then the convergence is linear for $r \in [0, 1)$. In particular $r = 0$ stands for superlinear convergence.
- If the limit does not exist, but

  $$a = limsup_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||} < 1,$$

  then the convergence is linear with a rate not exceeding a.

- If
$$m = liminf_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||} = 1,$$
the convergence is sublinear and in all other cases, i.e
$$liminf_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||} < 1 \leq limsup_k \frac{||x_{k+1} - x_0||}{||x_k - x_0||}$$
we cannot claim anything about the convergence rate.

Examples

- The sequence $x_n = 1 + (\frac{1}{2})^n$ converges linearly to $x_0 = 1$.
$$\frac{||x_{k+1} - 1||}{||x_k - 1||} = \frac{1}{2}$$

- The sequence $x_n = 1 + (\frac{1}{n})^n$ converges superlinearly to $x_0 = 1$.
$$\frac{||x_{k+1} - 1||}{||x_k - 1||} = \frac{1}{n+1}\left(\frac{1}{(1+\frac{1}{n})^n}\right) \to 0$$

- The sequence $x_n = 1 + (\frac{1}{n})^{2^n}$ converges quadratically to 1.
$$\frac{||x_{k+1} - 1||}{||x_k - 1||^2} = (\frac{n}{n+1})^{2^{n+1}} \leq 1.$$

Note that if there is $a \in (0,1)$ and $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$
$$||x_{k+1} - x_0|| \leq a||x_k - x_0||$$
then the sequence is convergent to x_0 since this implies that
$$||x_k - x_0|| \leq a^{k-k_0}||x_{k_0} - x_0||$$
which goes to 0. This is not the case for a β order of convergence which does not guarantee convergence.

Proposition 5. *Practical criterion of convergence and estimation of the order*

- *If there is an $a \in (0,1)$ and a $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$*
$$||x_{k+1} - x_k|| \leq a||x_k - x_{k-1}||,$$
then the sequence converges towards an x_0, the convergence is linear and the rate of convergence is $r \leq a$.

- *Assume $\beta > 1$. If there is $a > 0$, $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$*

$$||x_{k+1} - x_k|| \leq a||x_k - x_{k-1}||^{\beta},$$

and there is $k_1 \geq k_0$ such that $a||x_{k_1} - x_{k_1-1}||^{\beta-1} < 1$ then the sequence converges towards an x_0 and the convergence is at least of order β.

Exercice

1. Show with the definition that the rate of convergence is not linear,

$$x_n = \frac{1}{n}.$$

2. Show with the definition that the rate of convergence is not quadratic,

$$x_n = \frac{1}{n^n}.$$

   Prove that actually x_n converges superlinearly to 0 but does not converge to 0 with any other order $a > 1$.

3. Show that a sequence that converges with order $a > 1$ must converge superlinearly.

4. Determine the convergence or divergence of $x_n = 0.7^n$ and $u_n = 0.7^{2^n}$.

5. Check whether the following sequences converge. In the case of convergence, find the rate of convergence or the convergence order.

   - $x_0 = 1$ and $x_{n+1} = \frac{1}{2}ln(\sqrt{x_n + 1}), n > 0$.
   - $x_n = \sum_{k=1}^{n} \frac{1}{k}$.
   - $x_1 = 2$, $x_{n+1} = \frac{1}{2}(x_n + \frac{1}{x_n})$.
   - $x_0 > 0$ and $x_{n+1} = (2^{-n} + 3^{-n})x_n$.

1.3 Analytical methods of optimization

Optimization consists in seeking the minimum or the maximum of a function of one or several variables. This function is called objective function. In optimization, we are concerned about continuous functions at least piecewise, as well as discrete functions (for example, the price of goods, number of given equipment). A function is called unimodal when it presents only one maximum or one minimum. If it presents several maxima or minima, it is called multimodal.

Definition 6. *Let X be a convex subset of a vector space and $f : X \to \mathbb{R}$ a function. f is convex if for all $t \in [0, 1]$ and $x, y \in X$*

$$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y).$$

f is called concave of $-f$ is convex.

Example

1. If $S \in M_n(\mathbb{R})$ and $x \in \mathbb{R}^n$ is a symmetric positive definite matrix then

$$f(x) = x^T Sx - a^T x + b$$

is (strictly) convex.

1.3.1 Functions of one variable

Let the objective function be

$$f : I \to \mathbb{R}$$

and search the solution to the problem

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

that is find the x such that for all $y \in I$, $f(x) \leq f(y)$. We can also look for local solutions where $f(x) \leq f(y)$ for all y in a neighborhood of x in I.
We recall the following results.

- If I is compact there is at least one solution to the problem.
- If x is a minimiser and f is differentiable at x then $f'(x) = 0$. If f is twice differentiable then $f''(x) > 0$.
- If f is strictly convex any local minimum is a global minimum.
- Possibile extrema are the solutions of $f'(x) = 0$. Such points are called stationary points. Boundary points are also possible extrema. To find global extrema we compare the values of f at all stationary and boundary points.

The previous results do not hold for discontinuous functions or functions with discontinuous derivatives but still hold in each of the subintervals of continuity of f and f'.

Exercice

- Find the global extrema of $f(x) = x^2 + \frac{4}{x}$ in $[-1, 2]$.
- True-False
  1. Let C be the set below the graph of $f(x) = |x|$, $x \in [-1, 1]$. C is convex.
  2. $[-1, 1] \times [-1, 1] \setminus C$ is convex.
  3. The function $(-1, 0) \cup (1, 2) \to \mathbb{R}$ given by $f(x) = x^2$ is convex.
  4. Any convex function is continuous in the interior of its domain.

5. If f is not contiuously differentiable, it cannot be convex.
6. If f and g are convex then $f + g$ is convex.

- Find any stationary point and sketch the function

$$f(x) = \frac{x-5}{x}, x \in \mathbb{R}^*.$$

1.3.2 Functions of several variables

Let f be a real function defined on $\mathbb{R}^n$. The stationary points of f are the solutions of

$$\nabla f(x) = 0.$$

This condition is necessary for an extremum but not sufficient. For a stationary point x^* to be a minimum the Hessian matrix at x^* must be definite positive. Similarly for a maximum it must definite negative. A matrix is definite positive if all its eigenvalues have strictly positive real part. If f is defined on a bounded interval possible extrema may lie on the limits of the domain.
Exercice Find the extremum points of the following functions in $\mathbb{R}^2$.

- $f(x, y) = log(x^2 + y^2 + 1)$
- $f(x, y) = x^2 - y^2$
- $f(x, y) = (y - 3x^2)(y - x^2)$.
- $f(x, y) = x^5y + y^5x + xy$
- $f(x, y) = x^2 - 2xy + y^2$
- $f(x, y) = x^2 + y^2 - x - y - 1$ in $D = \{(x, y)/x^2 + y^2 \leq 1\}$

1.3.3 Function subject to constraints

Lagrange multipliers
A common case of optimization in physical and economical problems is the search of the minimum of the objective function subject to m constraints:

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

under m equality constraints:

$$g_i(x) = 0, \quad i = 1, .., m.$$

The problem of maximization is equivalent to minimizing $-f$. One way to solve this is by using the Lagrange multipliers method. Define the following function

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

which is called the Lagrangian. L is a function of x and m scalars $\lambda_i, i = 1, ..., m$. Finding the stationary points of L in the unconstrained domain, is equivalent to solving the constrained problem. Remark that $\frac{\partial L}{\partial \lambda_i} = 0$ gives back the constraints.

Theorem 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be the objective function and $g : \mathbb{R}^n \to \mathbb{R}^m$ the constraints function, both belonging to $C^1(\mathbb{R}^n)$. Let x^* be the optimal solution to the following optimization problem*

$$\begin{cases} min_x f(x) \\ subject\ to \quad g(x) = 0 \end{cases}.$$

If $D(g), D(f)$ stand for the matrices of partial derivatives of f and g then there is a unique Lagrange multiplier $\lambda \in \mathbb{R}^m$ such that

$$D(f)|_{x^*} = \lambda^T D(g)|_{x^*},$$

where λ is a column vector.

Corollary 1. *Let $f, g : A \subset \mathbb{R}^n \to \mathbb{R}$ be two functions, A an open subset of $\mathbb{R}^n$ and $f, g \in C^1(A)$. If P_0 is a local extremum of f under the constraint $g(P) = k$, $k \in \mathbb{R}$ a constant, and $\nabla g(P_0) \neq 0$ then, there is $\lambda \neq 0$ such that*

$$\nabla f(P_0) = \lambda \nabla g(P_0).$$

Proof. Here is a description of the proof. Let $g(x_1, .., x_n)$ be $C^1(A)$ with level curve for $k \in \mathbb{R}$

$$g(x_1, ..., x_n) = k.$$

Since $\nabla g(P_0) \neq 0$, the gradient at P_0 will be perpendicular to the level curve. Consequently, it is perpendicular to the tangent vectors of any curve passing from the point P_0. Let σ denote a curve

$$\begin{cases} \sigma : [a, b] \to \mathbb{R}^n \\ g(t) = (g_1(t), ..., g_n(t)) \end{cases}$$

for which $\sigma(t_0) = P_0$ for some $t_0 \in [a, b]$. Define $\phi : [a, b] \to \mathbb{R}$,

$$\phi = f \circ \sigma.$$

By the chain rule $\phi'(t) = \nabla f(P)\sigma'(t)$ and since we have an extremum at P_0, $\phi'(t_0) = 0$. This implies in particular that

$$\nabla f(P_0)\sigma'(t_0) = 0$$

hence, $\nabla f(P_0)$ and $\nabla g(P_0)$ are collinear.

□

Remarks

- Lagrange theorem describes possible extremum points.
- Possible extrema are also the points where $\nabla g(P) = 0$.
- If the constraints describe a closed, bounded set of $\mathbb{R}^n$ there is always a minimum and a maximum. In that case to determine them, compare the values of f at the possible extremum points.
- If the constraint set is unbounded there may not be any extremum on the set. One way to decide whether the stationary point is a local extremum is to study f locally using its Taylor approximation and estimate the sign of $df = f(P) - f(P_0)$.

Exercices

1. Find the extremum points of the following function f under constraints.
   - $f(x, y) = x - y + z$ under $x^2 + y^2 + z^2 = 2$.
   - $f(x, y) = x^2 + y^2$ under $y - x = 1$
   - $f(x, y) = x - y$ under $x^2 - y^2 = 2$.
   - $f(x, y, z) = x + y + z$ under $x^2 + y^2 = 2, \quad x + z = 1$.
   - $f(x, y, z) = x + y + z$ under $x^2 - y^2 = 1, \quad 2x + z = 1$.
   - $f(x, y) = x^2 + y^2 - x - y + 1$ on $S = \{(x, y) \in \mathbb{R}^2 / x^2 + y^2 \leq 1\}$.
2. Find the maximum of the surface area of a rectangle while imposing that the perimeter is constant.
3. Find the $min||x||_p$ for $p = 1, 2, +\infty$ under $Ax = b$ where $A = \begin{bmatrix} 3 & 4 \end{bmatrix}$ and $b = [1]$.
   Remark: The optimum in $p = 1$ norm gives a sparse solution.
4. (PENALTY) Consider the matrix $A = \begin{bmatrix} \sigma \end{bmatrix}$, with $0 \leq \sigma << 1$ very small, $b \in \mathbb{R}$. Solve the problem
   $$min|||Ax - b||_2^2.$$
   Hint: When $\sigma << 1$ the eigenvalue of A is very small which entails that the eigenvalue of A^{-1} is very big. If $\sigma = 0$, A is not invertible. In optimization, this situation is in general very difficult to handle especially for large matrices. Try to solve the problem and see what happens with σ. Add now a penalty term and solve the problem
   $$min|||Ax - b||_2^2 + \delta^2 ||x||_2^2.$$
   Let $\delta > 0$ go to 0. How is the problem simplified?
   Comment: (LASSO) In statistics, we usually add a penalty term in the $||.||_1$ norm and solve the problem
   $$min|||Ax - b||_2^2 + \delta^2 ||x||_1^2$$
   because it gives more genuine, sparse solutions.

Chapter 2

Numerical methods

Analytical methods represent an important theoretical base for optimization but in general, optimization is performed by means of iterative numerical methods, especially when it deals with large scale problems such as those encountered in engineering.

Definition 7. *Let*

$$\begin{cases} minf(x) \\ g_i(x) = 0, i = 1, .., m \end{cases}$$

be the optimization problem. We call feasible set the set of values in the domain of f satisfying the constraints g_i.

For functions of several variables there are two main classes of methods, the methods of direct search and the gradient methods. They both lie on the same principles.

1. Choose a point on the feasible set, called basis point, and evaluate the value of the function at this point.

2. Choose a second point based on the method that you use and evaluate the value of the objective function at this point.

3. Compare the values. If the second point is better then it is the new basis point. If the initial point is better modify the search direction, or the strategy or stop.

4. Continue until a stopping criterion is satisfied (or a desired accuracy is obtained).

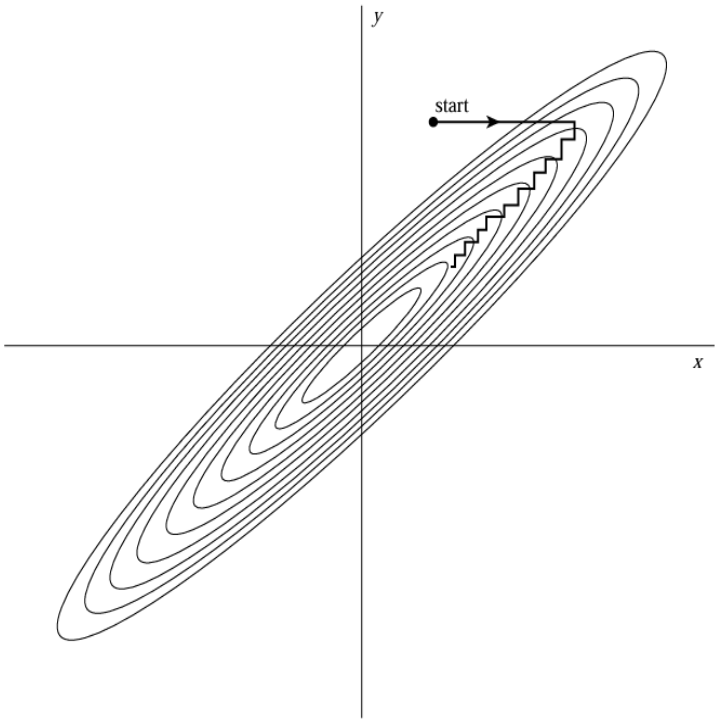
2.1 Methods of direct search

Direct search is a method of solving optimization problems that does not require information about the gradient and as a result it can be used for functions that are not differentiable or even continuous. A direct search algorithm searches a set of points around the current point, looking for one point where the value of the objective function is lower than the value at the current point.

1. **One variable search** Let's say that we want to minimize f in $\mathbb{R}^n$ without constraints. Consider $\{e_1, e_2, ..., e_n\}$ the unit basis of $\mathbb{R}^n$ and start at a random point $x_0 = x_0^1 e_1 + ... + x_0^n e_n$. The simple one variable search will modify the value of only one variable at each step. If you move along direction v, the new point will be

$$x_{n+1} = x_n + \alpha_n v_n,$$

with α_n to be appropriately determined. Start with $v = e_1$ for the first step, continue with $v = e_2$ at the second step etc. The difficulty stands in choosing α for which no information exists. One way to proceed is to move along the first direction to its minimum, then from there along the second direction to its minimum, and so on, cycling through the whole set of directions as many times as necessary, until the function stops decreasing.
This simple method can work well for many functions but there are examples where it is extremely inefficient. For instance if you consider a function whose level sets form narrow and long valleys at some angle to the coordinate basis system like in the picture below



Unless the valley is optimally oriented the only way to move down to the valley along the coordinate vectors, is to make tiny steps. For that reason we need a better set of directions. There are many methods that define the set of directions, we will explain the Hooke Jeaves algorithm which is simple to understand and implement.

2. **Hooke and Jeeves acceleration method**

   The Hooke Jeaves acceleration method is a method that converges faster than the simple direction search since it keeps track of the direction of the travel as the process moves from point to point.

   The method will be given in 2 dimensions but it can be generalized to n dimensions in the same way.

   1. Step 1: Start from an initial point $x_0^1 = (x_0, y_0)$ and a fixed variation $\pm\delta$ (for instance $\delta = 0.1$), and move parallel to Ox_1. Find the values of f at $(x_0 + \delta, y_0)$ and $(x_0 - \delta, y_0)$ and compare. Save the best value and call the new point x_1^1.
   2. Step 2: Start from the new point and move parallel to Ox_2 with a variation $\pm\delta$. If there is some improvement the new point is x_2^1. It is supposed that at least one of the two values gives a better result than the initial point x_0^1.
   3. Step 3: The accelaration is introduced here. Define the vector formed by the two points x_0^1 and x_2^1, called v and define a new point x_0^2

      $$x_0^2 = x_0^1 + \alpha v = x_0^1 + \alpha(x_2^1 - x_0^1)$$

      with recommended factor for $\alpha = 2$. Search in that direction until you stop seeing improvement.
   4. Step 4: When you stop seeing improvement in the direction of the vector v repeat steps 1 and 2. When step 1 gives back the original point, you move to step 5:
   5. Step 5: Decrease the incremental interval. If you started by adding and subtracting 0.1, now you might add and subtract 0.01. Then repeat everything until you have the desired level of accuracy. The stopping criterion can be determined by the user and could be $||\delta|| < \epsilon$, $\epsilon > 0$ a desired accuracy level.

   Since the search largely depends on the moves along the coordinate directions the algorithm may converge to a wrong solution, especially in the case of functions with highly nonlinear interactions among variables. Note also that the algorithm terminates only by exhaustively searching the neighborhood of the converged point. This requires a large number of function evaluations for convergence to a solution with a reasonable degree of accuracy. The convergence to the optimum point depends on the parameter α.

3. **Simplex Method** In geometry a simplex is a generalization of the notion of the triangle or tetrahedron to arbitrary dimensions.

   Definition 8. *The convex hull of a given set E is the unique minimal convex set that contains E, or equivalently, the intersection of all the convex sets containing E.*

Definition 9. *A k-simplex is a k dimensionnal polytope, the convex hull of its $k+1$ vertices. The points $u_0, u_1, ..., u_{k+1}$ define k linearly independent vectors $u_1 - u_0, u_2 - u_0, ..., u_n - u_0$ and the simplex is determined by them as*

$$\{\lambda_0 u_0 + \lambda_1 u_1 + ... + \lambda_n u_n / \sum_{i=0}^{n} \lambda_i = 1 \quad and\ \lambda_i \geq 0\}.$$

The simplex is the simplest polytope in any given dimension. Hence

- In 0 dimension, a simplex is a point.
- In 1 dimension with 2 points, a simplex is a line segment.
- In 2 dimensions with 3 points, a simplex is a triangle.
- In 3 dimensions with 4 points, a simplex is a tetrahedron (triangular pyramid).

For the simplex method we will consider regular geometric figures, for instance in 2 dimensions an equilateral triangle. Then we proceed as follows

a) Choose an initial base figure with $n+1$ points $x_j, j = 1, .., n+1$ with respective coordinates $x_{i,j}$ for $i = 1, .., n$.

b) Calculate the objective function at the vertices and determine the point with the worst score (in terms of minimum or maximum based on the optimization problem). Note this points x_R with coordinates $x_{i,R}$.

c) Find the centroid of the other n points x_C

$$x_{i,C} = \frac{1}{n}\left(\sum_{j=1}^{n+1} x_{i,j} - x_{i,R}\right)$$

This point is an arithmetic mean position.

d) x_R is rejected and it is replaced by its symmetrical with respect to the centroid noted x_N. Since $\frac{x_N + x_R}{2} = x_C$

$$x_{i,N} = \frac{2}{n}\left(\sum_{j=1}^{n+1} x_{i,j} - x_{i,R}\right) - x_{i,R}.$$

e) Evaluate the value of the function at this new point.

- If this new point gives a better result than the rejected point, this new point is adopted and we are sure to move away from the bad direction. Restart at b. If this new point gives a worse result than the rejected point, don't reject it. Choose the second worse point as the rejected point, and then act like in b/ by replacing the rejected point by its symmetrical with respect to the centroid of the other points.

– It may also happen that the new point is not feasible. In this case, it cannot be accepted. The second worse point will be chosen as the worst point.

- **Nedler-Mead Simplex** This method of direct search has many features in common with the simplex method but it has the advantage of using a deformable figure, a polytope that can be adapted to the shape of the objective function. It is one of the best methods used in unconstrained problems with a convex objective functions. Note that depending on the dimension the problem either converges or may not converge or even converge to a non stationary point on problems that can be solved alternatively. Moreover the optimization problem with lower dimension will converge faster as compared to higher dimensions. It corresponds to the method "fminsearch" of Matlab, optim() function in R and minimize(method='Nelder-Mead') in Python.

  Starting from an initial point in n dimensions, Nelder–Mead algorithm generates a series of vertices to come closer to the sought optimum. At each iteration, the vertices of the simplex are ordered with respect to the value of the objective function as

  $$f(x_1) \leq f(x_2) \leq ... \leq f(x_{n+1})$$

  with x_1 being the best vertex and x_{n+1} the worst. Note $f_i = f(x_i)$.

  1 Step 1. Reflect the worst point with respect to the centroid of the other points x_C. The new point is

  $$x_r = x_C + \rho(x_C - x_{n+1}),$$

  where ρ is a parameter.

  2 Step 2. Evaluate $f_r = f(x_r)$ and proceed as follows. If $f_1 \leq f_r < f_n$ replace x_{n+1} by x_r and terminate the iteration.

  3 Step 3. If the above ordering is not satisfied then if the point x_r is better than the current best point we stretch exponentially along this line. If it is not much better than the previous value we shrink the simplex towards a better point.

  * If $f_r < f_1$ calculate the expansion point

  $$x_e = x_C + \chi(x_r - x_C),$$

  where χ is a parameter. If $f_e = f(x_e) < f_r$ replace x_{n+1} by x_e and terminate the iteration otherwise replace x_{n+1} by x_r and terminate the iteration.

  * If $f_r \geq f_n$ do contraction with parameter γ.

  a) If $f_n \leq f_r < f_{n+1}$ do external contraction

  $$x_{ec} = x_C + \gamma(x_r - x_C)$$

with $f_{ec} = f(x_{ec})$. If $f_{ec} \leq f_r$, replace x_{n+1} by x_{ec} and terminate the iteration. Otherwise, proceed to the shrink stage.

b) If $f_r \geq f_{n+1}$ do internal contraction

$$x_{ic} = x_C - \gamma(x_C - x_{n+1})$$

Evaluate $f_{ic} = f(x_{ic})$. If $f_{ic} < f_{n+1}$, replace x_{n+1} by x_{ic} and terminate the iteration. Otherwise, proceed to the shrink stage.

$*$ Shrink with parameter $0 < \sigma < 1$.
For $2 \leq i \leq n+1$,calculate the new points v_i $v_i = x_1 + \sigma(x_i - x_1)$ The simplex is then formed by the unordered points $x_1, v_2, ..., v_{n+1}$

The recommended values for the parameters are, $\chi > \rho$, $\rho = 1, \chi = 2$ and $0 < \gamma < 1$.

Nelder and Mead used the sample standard deviation of the function values of the current simplex. If these fall below some tolerance, then the cycle is stopped and the lowest point in the simplex returned as a proposed optimum.

2.2 Gradient methods

In this section we will be solving an optimization problem using the gradient of the objective function. In the case of minimization the algorithms will be generating a minimizing sequence of points $x_0, x_1, ...$ where

$$x_{k+1} = x_k + t_k(\Delta x)_k.$$

The $(\Delta x)_k$ is called search direction and the t_k is called the step size. The search direction is a vector in the direction of which the objective functions decreases its value, it is thus called a descent direction and implies

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

except when x_k is optimal.

Definition 10. *Let $f : A \subset \mathbb{R}^n \to \mathbb{R}$ be a $C^1(A)$ function with A open. We say that v is a descent direction in $\mathbb{R}^n$ at x if the derivative of $h(t) = f(x + tv)$ at 0 is*

$$h'(0) < 0.$$

The general algorithm has three steps at each iteration. Fix an initial point $x_0 \in A$.

- Choose a descent direction $(\Delta x)_k$ at the current point.
- Choose a step size t_k.

- Update $x_{k+1} = x_k + t_k(\Delta x)_k$.

until stopping criterion is satisfied. Since looking for a minimim of a C^1 function starts from searching the stationary points, hence the solutions of $\nabla f(x) = 0$ a possible stopping criterion would be

$$||\nabla f(x)|| < \epsilon, for \quad \epsilon > 0.$$

The second step is called the line search since selection of the step size t_k determines where along the line $\{x_k + t(\Delta x)_k / t \in \mathbb{R}^+\}$ the next iterate will be.

The different algorithms consist in different choices of the descent direction and the step size.

Since the gradient of a a function is the direction of largest increase (or largest decrease for $-\nabla f$) in a descent direction v the angle formed between $-\nabla f$ and v must be acute or

$$\nabla f^T.v < 0$$

Remark If f is convex and twice differentiable then the Hessian is positive definite so from its Taylor development we deduce that if $\nabla f(x_k)^T.(\Delta x)_k > 0$ then $f(x_{k+1}) > f(x_k)$ which implies that if we want f to decrease we must move in a direction of descent.

For these algorithms the choice of the initial point may be crucial. If the function is convex there is no problem, the algorithm will converge to the unique minimum since all the decreasing paths lead to this point. But if the function is not convex there may be many local minima. The initialization impacts the convergence leading to one local minimum which may not be the global one since we only follow a descent direction. One should use all information given about f to initialize as close to the real minimum as possible.

2.2.1 Gradient descent with constant step size

All the algorithms of gradient descent choose the $-\nabla f(x_k)$ as search direction at x_k. The first algorithm that we will explain uses a fixed step size a and is as follows

- Fix a step a and an initial point $x_0 \in \mathbb{R}^n$.
- Set $x_{k+1} = x_k - a\nabla f(x_k)$
- Stop iterating when $||\nabla f(x_k)|| < \epsilon$.

Theorem 4. *Let $f : A \to \mathbb{R}$ be a C^2 function and lower bounded. Fix a point $x_0 \in A$ and assume*

- *The set $S = \{x / f(x) \leq f(x_0)\}$ is closed in $\mathbb{R}^n$,*
- *For all $x \in S$ the Hessian satisfies $lI_n \leq H(x) \leq LI_n$, with $0 < l \leq L$.*

If we chose $a < \frac{2}{L}$ the sequence x_k converges linearly towards a local minimum and the rate of convergence is less than or equal to $r = r_a = max\,(|1 - la|, |1 - La|)$.

This results says that if the Hessian is positive definite in a neighborhood of a local minimum then the algorithm will converge to the local minimum and the rate of convergence is linear. In the general case, the convergence of the algorithm (towards a point) depends on the assumptions that we make on f and similarly the convergence towards the local minimum.

To test if the conditions of convergence are satisfied we need to verify that the set S is closed. For that, we give a result that helps us conclude.

Definition 11. *(Coercive) Let $f : X \subset \mathbb{R}^n \to \mathbb{R}$ is called coercive on X if $lim_{||x||\to+\infty, x\in X} f(x) = +\infty$ if and only if for any sequence in X with $lim_{k\to+\infty}||x_k|| = +\infty$ we have $lim_{k\to+\infty} f(x_k) = +\infty$.*

Proposition 6. *(Characterization of coerciveness) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. Then f is coercive if and only if for every $a \in \mathbb{R}$ the set $S = \{x/f(x) \leq a\}$ is compact.*

We will prove the result for a specific function.

$$f(x) = \frac{1}{2}x^T Sx - b^T x$$

with $x, b \in \mathbb{R}^n$ column vectors, b constant, S symmetric positive definite.

- The gradient is $\nabla f(x) = Sx - b$ and $H = S$.
- The iteration step $x_{k+1} = x_k - a(Sx_k - b)$.

This function is convex (already proven) so we can analytically find its minimum which happens at $S^{-1}b$ with minimum value

$$f(S^{-1}b) = \frac{1}{2}(S^{-1}b)^T SS^{-1}b - b^T S^{-1}b$$

S is symmetric so S^{-1} is also symmetric so

$$f(S^{-1}b) = \frac{1}{2}b^T S^{-1}SS^{-1}b - b^T S^{-1}b = -\frac{1}{2}b^T S^{-1}b.$$

We will now determine how quickly the algorithm converges to this minimum.

Proposition 7. *If S is a symmetric matrix of size n and $\lambda_1, ..., \lambda_n$ its eigenvalues. Then for all $x \in \mathbb{R}^n$ we have*

$$||Sx||_2 \leq max\{\lambda_i\}||x||_2$$

Proof. Choose an orthnormal basis of eigenvector $e_1, .., e_n$ and express x in this basis $x = \sum_{i=0}^{n} x_i e_i$. Then

$$||Sx||_2^2 =< Sx, Sx >=< \sum_{i=1}^{n} x_i Se_i, \sum_{i=1}^{n} x_i Se_i >=< \sum_{i=1}^{n} \lambda_i x_i e_i, \sum_{i=1}^{n} \lambda_i x_i e_i >$$

the basis is orthonormal hence it gives

$$||Sx||_2^2 = \sum_{i=1}^{n} \lambda_i^2 x_i^2 \leq max_i \lambda_i^2 ||x||_2.$$

□

Proposition 8. *If l, L are the smallest and the largest eigenvalues of S and $a \in (0, \frac{2}{L})$ the convergence is linear and the rate of convergence is at most $max\,|1 - al|, |1 - aL|)$. The rate takes its minimum value if $a = \frac{2}{l+L}$ with a rate $r = \frac{L-l}{L+l}$.*

Proof. In the following expression we substitute the values of the iterates with respect to their previous point.

$$x_{k+1} - x_k = x_k - a(Sx_k - b) - x_{k-1} + a(Sx_{k-1} - b) = (I_n - aS)(x_k - x_{k-1})$$

The eigenvalues of $I_n - aS$ are $1 - a\lambda$ for λ eigenvalue of S so

$$||x_{k+1} - x_k||_2 \leq max\{|1 - al|, |1 - aL|\}||x_k - x_{k-1}||_2$$

hence we have linear convergnece once the rate is less than 1, which is guaranteed if $a \in (0, \frac{2}{l})$ and $a \in (0, \frac{2}{L})$. The minimum is taken when $|1-al| = |1-aL|$ since otherwise we could decrease the value of the max by changing the value of a. If $L > l$ the only possibility for these values to be equal is $1 - al = aL - 1$ which gives $a = \frac{2}{l+L}$. Then the rate is

$$1 - \frac{2l}{l + L} = \frac{L - l}{l + L}.$$

If $l << L$ the rate is very close to 1 and in that case the matrix is ill-conditioned with a very big condition number $\frac{L}{l}$. □

2.2.2 Gradient descent with optimal step

The next algorithm chooses the step a_k in an optimal way, in the sense that the objective function decreases as much as possible in a given direction.

Definition 12. *We say that the gradient descent given by*

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

is a descent direction with optimal step if a_k minimizes the real function $h(t) = f(x_k - t\nabla f(x_k))$ in $\mathbb{R}^+$.

Proposition 9. *If a_k is an optimal step, then $-\nabla f(x_k)$ and $\nabla f(x_{k+1})$ are orthogonal.*

Proof. Differentiate the function h to get $h'(t) = -\nabla f(x_k - t\nabla f(x_k)).\nabla f(x_k)$. Since we follow a descent direction $a_k > 0$. Also $h'(0) < 0$ so the minimum happens at $(0, +\infty)$ which is a stationary point so $h'(a_k) = 0$. Consequently

$$< \nabla f(x_{k+1}), -\nabla f(x_k) >= 0.$$

□

This results holds for any descent direction.

Advantages and disadvantages of the method Firstly, a variant step permits a certain flexibility when applying the method in constrat to a fixed step. Moreover when the step is constant, we need to choose it very small and the algorithm may take time to start especially if we initialize away from the minimum. An inconvience is that we need to search for the optimal step. The gradient descent with a constant step converges linearly whereas the gradient descent with optimal step does not perform much better, in the sense that the convergence is linear for the $||.||_S$ norm ($||x||_S = \sqrt{x^T S x}$, with S symmetric definite positive).

Proposition 10. *For the function $f(x) = \frac{1}{2}x^T S x - b^T x$ in $\mathbb{R}^n$ with b a constant vector and S a symmetric definite positive matrix with l, L its minimum and maximum eigenvalues, the unique solution of $Sx - b = 0$ we have*

$$||x_{k+1} - x^*||_S \leq \frac{L - l}{L + l}||x_k - x^*||_S$$

with x_k the sequence of iterates of the gradient descent with optimal step. Moreover there is an initial point x_0 for which we get the equality in the previous estimation.

If the condition number $\frac{L}{l}$ is close to 1 the rate of convergence is close to 0 and we have no problem of convergence, whereas when the condition number is very big $l << L$ the matrix is ill-conditioned □.

We will now determine the value of optimal a_k and the iterates in the case of a well-conditioned matrix S and the quadratic function

$$f(x) = \frac{1}{2}x^T S x - b^T x + c.$$

Note that $\nabla f(x) = Sx - b$. Set $h(t) = f(x + tv)$ with v the stepest descent direction at x. We have $h'(t) = v^T.\nabla f(x + tv)$. At the optimum $h'(t) = 0$ hence

$$v^T S(x + tv) - v^T b = 0 \implies v^T(Sx - b) + v^T S v t = 0 \implies t = \frac{v^T(b - Sx)}{v^T S v}.$$

Since $-\nabla f(x) = b - Sx$ is the steepest descent direction v

$$t = \frac{v^T v}{v^T S v} = \frac{||v||^2}{v^T S v}$$

Note as well that $v_{k+1} = b - Sx_{k+1} = b - S(x_k + a_k S v_k)$ giving

$$v_{k+1} = v_k - a_k S v_k.$$

In total, the iterates go as follows

$$\begin{cases} x_{k+1} = x_k + a_k v_k \\ v_k = -\nabla f(x_k) \\ a_k = \frac{||v_k||^2}{v_k^T S v_k} \\ v_1 = b - Sx_1 \end{cases}$$

Remark When S is ill-conditioned we can use a relaxed gradient descent algorithm with respect to the optimal gradient descent where the iterates are

$$x_{k+1} = x_k + a_k r_k,$$

with the only change $a_k = \beta_k a_{k_0}$ where a_{k_0} is

$$a_{k_0} = \frac{||v_k||^2}{v_k^T S v_k},$$

and $0 < \beta_k < 2$ with $\beta_k = 1$ corresponding to the optimal gradient descent. It is proven that for $0 < \delta \leq 1$, $\delta \leq \beta_k \leq 2 - \delta$ and

$$L = 1 - \delta(2 - \delta)\frac{4lL}{(l + L)^2}$$

the convergence of f to the minimum x_0 is ruled by

$$f(x_{k+1}) - f(x_0) \leq L^k \left(f(x_1) - f(x_0)\right)$$

so that $f(x_k)$ tends linearly to $f(x_0)$ with a constant L and x_k tends linearly to x_0.

2.2.3 Gradient descent with momentum

The gradient descent method is used to minimize convex and non convex functions but does not always perform well. It is sure to find the minimum under reasonnable hypothesis when the function is convex but for other functions it may get stuck to stationary points that are not extrema (such as saddle points) or to local mimima. Sometimes, even in the case of convergence of a convex function the convergence may be extremely slow and hence we wish to accelerate it. To do so, we use momentum, and the motivation behind it is given with a simple example.

- Consider the quadratic function

$$f(x) = \frac{1}{2}x^T Sx$$

with $S = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ and $x \in \mathbb{R}^2$, $\lambda_i > 0, \lambda_1 < \lambda_2$. Note that we have already seen a specific case of this function for $\lambda_1 = 1$ and $\lambda_2 = b$. f has a global minimum at $(0,0)$, $\nabla f(x) = Sx$ and λ_i are the eigenvalues of S.

If we run a gradient descent with a constant step size a we get

$$x_{k+1} = x_k - aSx_k$$

which gives

$$(x_{1_{k+1}}, x_{2_{k+1}}) = (x_{1_k}, x_{2_k}) - a(\lambda_1 x_{1_k}, \lambda_2 x_{2_k}) = ((1 - a\lambda_1)x_{1_k}, (1 - a\lambda_2)x_{2_k})$$

Since we want iterates to go to 0 as fast as possible we set

$$|1 - a\lambda_1| << 1, |1 - a\lambda_2| << 1$$

If λ_1, λ_2 are of the same order of magnitude there is no problem of convergence, we can set $a \approx \frac{1}{\lambda_1}$ but if λ_2 is much larger than λ_1 (in which case the condition number $\frac{\lambda_2}{\lambda_1}$ is very big and S is ill-conditioned) there is no good choice of a. If $a \approx \frac{1}{\lambda_1}$ then $1 - a\lambda_2 < -1$ and the second coordinate of the iterates diverges as k goes to infinity. If $a \approx \frac{1}{\lambda_2}$ the decay is very slow $1 - a\lambda_1 \approx 1$.

One way to face this problem is to use a memory term (momentum) when defining the search direction since using more information from previous steps can help us improve the approximation. Instead of simply taking $-\nabla f(x_k)$ we take another vector z_k whose leading term is still the gradient but it is also corrected by a memory term, that is

$$\begin{cases} x_{k+1} = x_k - sz_k \\ z_k = \nabla f(x_k) + \beta z_{k-1} \end{cases},$$

with s, β constants. With s, β constants we call this algortithm "heavy ball" and we can imagine a ball going down the valley instead of a point since with this algorithm we have important speed ups. Using momentum instead of gradient descent permits us use a largest step size. With a large step size the gradient descent may diverge (as was the case with the second component in the previous example) but sz_k may remain bounded in which case the momentum algorithm does not diverge.

Theorem 5. *Let $0 < l < L$ be fixed. Let f be a quadratic function which is differentiable and satisfies for all $x, y \in \mathbb{R}^n$*

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

and

$$f(y) \geq f(x) + \nabla f(x).(y-x) + \frac{l}{2}||y-x||^2.$$

For the gradient descent with momentum and fixed step size $a = \frac{1}{lL}$ *there exists a constant* $C = C_{l,L}$, x^* *such that for any* k

$$f(x_k) - f(x^*) \leq Ck^2(\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}})^{2k}||x_0 - x^*||^2.$$

a) **Example :** Consider the function

$$f(x) = \frac{1}{2}x^T Sx,$$

with $S \in M_2(\mathbb{R})$, symmetric positive definite matrix and $x \in \mathbb{R}^2$. We will apply gradient descent with momentum.

Note that when we introduce the previous step we have a three level method which we will write as two equations of the same order in a slightly different way that above,

$$\begin{cases} x_{k+1} = x_k - sz_k \\ z_{k+1} - \nabla f(x_{k+1}) = \beta z_k \end{cases},$$

or using the gradient of f

$$\begin{cases} x_{k+1} = x_k - sz_k \\ z_{k+1} - Sx_{k+1} = \beta z_k \end{cases},$$

we know that if one of the eigenvalues of S is very big the convergence of the gradient descent is not guaranteed and it is this eigenvalue that causes divergence. S is a symmetric positive definite matrix so its eigenvectors form a basis and any x can be written in this basis. We will go in the direction of an eigenvector q such that $Sq = \lambda q$. Let's say that c_k, d_k are scalars such that $x_k = c_k q$ and $z_k = d_k q$. The system is rewritten

$$\begin{cases} c_{k+1} = c_k - sd_k \\ d_{k+1} - c_{k+1}\lambda = \beta d_k \end{cases},$$

or in a matrix form

$$\begin{bmatrix} c_{k+1} \\ d_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda & 1 \end{bmatrix} \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} \begin{bmatrix} c_k \\ d_k \end{bmatrix}$$

by matrix multiplication

$$R = \begin{bmatrix} 1 & 0 \\ \lambda & 1 \end{bmatrix} \begin{bmatrix} 1 & -s \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} 1 & -s \\ \lambda & \beta - \lambda s \end{bmatrix}$$

Knowing that $l \leq \lambda \leq L$ and that R depends on s, β the idea is to minimize the eigenvalues of R. R mulltiplies every step so if its eigenvalues are small the convergence is fast. It is proven that the values of β, s giving the optimum are

$$s_{opt} = \left(\frac{2}{\sqrt{l}+\sqrt{L}}\right)^2, \beta_{opt} = \left(\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}}\right)^2,$$

with the largest eigenvalue of R being negative. This result can be generalized to matrices S of size n with l, L being the min and max of its eigenvalues. We state here without proving it that the decay rate with momentum is

$$\left(\frac{\sqrt{L}-\sqrt{l}}{\sqrt{L}+\sqrt{l}}\right)^2$$

where l, L are the smallest and the largest eigenvalues of S. For ill-conditioned problems it is much faster than the gradient descent. Indeed gradient descent converges linearly with rate $\frac{L-l}{L+l}$ and if $\frac{l}{L}$ is small (giving $1-\frac{l}{L}$ close to 1 and so slow convergence) then $\frac{\sqrt{l}}{\sqrt{L}}$ is bigger than $\frac{l}{L}$ and speeds up convergence.

Remark It should be made clear that in the general case the convergence of $f(x_k)$ does not guarantee that x_k is convergent or even if it does, that it converges to the minimum. What is guaranteed is that $||f(x_{k+1}) - f(x_k)|| \leq \epsilon$ or $||\nabla f(x_k)|| \leq \epsilon$ hence the algorithm will terminate. Everything depends on the assumptions that we make and the sets that we work on. If x_k is convergent, then $f(x_k)$ is also convergent under the simple assumption of continuity of f.

2.2.4 Projected gradient

The gradient descent already discussed is an algorithm that worked for unconstrained problems with the update rule

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

For constrained optimization, however, the update rule does not necessarily generate a feasible solution. A natural fix for this is that we take the projection of the point $x_{k+1} = x_k - a_k \nabla f(x_k)$ onto the feasible set C which is taken closed and convex. Let us denote $proj_C()$ the Euclidean projection operator which is defined as

$$proj_C(x_0) = argmin_{x \in C} ||x - x_0||_2$$

which is equivalent to $\frac{1}{2} argmin_{x \in C} ||x - x_0||_2^2$ where we are squaring the function to make it differentiable. The gradient descent with projection has the following steps

- Pick an initial point $x_0 \in \mathbb{R}^n$.
- Choose a step size a_k,

- Find the iterate $y_{k+1} = x_k - a_k \nabla f(x_k)$,
- $x_{k+1} = proj_C(y_{k+1}) = \frac{1}{2} argmin_{x \in C} ||x - y_{k+1}||_2^2$.
- Continue until a condition is met.

Finding the *argmin* is an opitimization problem itself which is easy to solve and fast if it has a closed expression. If C is not convex or if the problem has no closed form, it may be expensive to compute. Concerning projection, if $x_0 \in C$ the projected point is x_0 itself whereas if x_0 is not in C the projection is the point where the euclidean distance from x_0 to C becomes minimal. The point where this happens is orthogonal to C. To determine it in 2 dimensions, consider circles centered at x_0 with increasing radius until the circle meets C in a tangent way. There

$$argmin_{x \in C} \frac{1}{2} ||x - x_0||_2^2 \in \partial C.$$

Proposition 11. *(Bourbaki-Cheney-Goldstein inequality- Obtuse angle criterion) If $x \in C$ and $z \in \mathbb{R}^n$ then*

$$(proj_C(z) - z)^T (proj_C(z) - x) \leq 0$$

Proof. By definition $proj_C(z) = \frac{1}{2} argmin_{x \in C} ||x - z||_2^2$ for $z \in \mathbb{R}^n$. The gradient of $f(x) = \frac{1}{2}||x - z||_2^2$ at $proj_C(z)$ is $proj_C(z) - z$. The statement is the optimality condition for the projection. Indeed

$$0 \geq f(proj_C(z)) - f(x) = \nabla f(proj_C z))^T (proj_C(z) - x).$$

□

Proposition 12. *A function f is called non-expansive if for any x, z in the domain of f*

$$||f(x) - f(z)|| \leq L||x - z||,$$

with $L \leq 1$. The projection operator is non-expansive. If Pz stands for the projection of z in C for the shake of simplicity

$$||Pz - Px||_2 \leq ||z - x||_2.$$

Proof. From the obtuse angle inequality we have $(Pz - z)^T (Pz - x) \leq 0$ for x in C. Replace x by Px to get $(Pz - z)^T (Pz - Px) \leq 0$ and z by x and x by Pz to get $(Px - x)^T (Px - Pz) \leq 0$. Equivalently we have

$$< Pz - z, Pz - Px >\leq 0, \quad < x - Px, Pz - Px >\leq 0$$

Add them together

$$< Pz - z + -Px + x, Pz - Px >\leq 0$$

which can be rewritten as

$$< x - z, Pz - Px > + < Pz - Px, Pz - Px > \leq 0.$$

Hence

$$< z - x, Pz - Px > \geq ||Pz - Px||_2^2$$

from Cauchy-Schwarz $< z - x, Pz - Px > \leq ||z - x||_2 ||Pz - Px||_2$, all together imply

$$||Pz - Px||_2^2 \leq ||z - x||_2 ||Pz - Px||_2,$$

which gives the desired result

$$||Pz - Px||_2 \leq ||z - x||_2.$$

□

We claim without proving it, that the non-expansiveness of the projection and the obtuse angle inequality guarantee that the iterates of the algorithm are monotonic

$$f(x_{k+1}) \leq f(x_k)$$

and these are used to find the convergence rate of the projected gradient descent. One result concerning convergence is given below.

Theorem 6. *(On ergodic convergence rate) If f is convex the projected gradient descent with constant step satisfies*

$$f(\frac{1}{K+1}\sum_{k=0}^{K} x_k) - f(x^*) \leq \frac{||x_0 - x^*||_2^2}{2a(K+1)} + \frac{a}{2(K+1)}\sum_{k=0}^{K} ||\nabla f(x_k)||_2^2,$$

where x^* is the global minimizer, a is the constant step size, K is the total number of iterations performed.

Interpretation

$\bar{x} = \frac{1}{K+1}\sum_0^K x_k$ is the average of the sequence x_k after K iterations. The theorem reads

$$f(\bar{x}) - f(x^*) \leq \frac{||x_0 - x^*||_2^2}{2a(K+1)} + a \text{ positive term}$$

and the rate of convergence of $f(\frac{1}{K+1}\sum_{k=0}^K x_k)$ is like $O(\frac{1}{K})$.

The positive term goes to 0 if the series $\sum_{k=0}^K ||\nabla f(x_k)||_2^2$ is not divergent or if it is summable with rate less than K. This convergence theorem is different than the theorems already seen, since what it permits us conclude is that the centroid of the points x_k goes to x^* but not each point x_i itself (as it would be the case in sequence convergence). Ergodic convergence means that the average of $x_0, x_1, ..., x_K$ is going closer to x^* but some of them may move away from x^* as long as the centroid gets closer.

2.2.5 Penalty

Consider the constrained minimization problem

$$min_C f(x),$$

C a closed convex subset of $\mathbb{R}^n$. We wish to transform this problem into an unconstrained optimization problem by penalising the values x that are not in C.

Definition 13. *Let $C \subset \mathbb{R}^n$ a closed convex subset of $\mathbb{R}^n$. We call $P : \mathbb{R}^n \to \mathbb{R}$ a penalty function any continuous function that satisfies*

- $P(x) \geq 0$,
- $P(x) = 0 \iff x \in C$.

Example (Beltrami function) A penalty function for $C = \{x \in \mathbb{R}^n / g_i(x) \leq 0, h_j(x) = 0, i \in I, j \in J\}$ is

$$P(x) = \sum_{i \in I} max(0, g_i(x))^2 + \sum_{j \in J} (h_j(x))^2.$$

To apply the penalty method, we remove the constraint and put it in the objective function via a scalar μ

$$min_{\mathbb{R}^n} f(x) + \mu P(x)$$

we have a new problem whose solution gives, under suitable assumptions, a solution to the constrained problem. Intuitively, we can see that if μ is sufficiently large then the new objective function $f(x) + \mu P(x)$ increases so the minimum will be found in C (since outside of C the penalty P is positive). The best value of μ is of course unknown hence to find a possible solution we consider a sequence $\mu_k > 0$ strictly increasing that tends to infinity. We solve each system

$$min_{\mathbb{R}^n} f(x) + \mu_k P(x)$$

and note x_k its solution. The following properties hold

- At each iteration the objective function increases.
- At each iteration the penalty decreases.
- At each iteration f iteself increases.

If x^* stands for the optimal solution of the initial problem then

$$f(x_k) \leq f(x_k) + \mu_k P(x_k) \leq f(x^*), for\ all\ k = 1, 2, ..$$

and under reasonnable hypothesis a limit exists and solves the constraint problem.

Theorem 7. *Let f be a C^1 function strictly convex with $lim_{||x|| \to +\infty} f(x) = +\infty$, C a convex set, closed and non empty. P is a penalty function. Then for all k the penalized problem has a solution x_k, the sequence of solutions has a limit $lim_k x_k = x^*$ and the limit is the unique solution of the constrained problem on C.*

2.3 Linear programming

In the case where the function to minimize is linear and the constraints are linear the appropriate method for solving the optimization problem is linear programming (LP). In the case of maximization of f we can minimize $-f$, hence we always refer to minimization.

The problem is posed under the following form. Let $f(x) = c^T x$, $x \in \mathbb{R}^n$

$$minf(x) = min \sum_{i=1}^{n} c_i x_i$$

subject to linear positive or negative inequality constraints

$$g_i(x) \leq 0, i = 1, .., n, \quad h_j(x) \geq 0, j = 1, .., m$$

and possibly some linear equality constraints

$$u_k(x) = 0, k = 1, .., r$$

with the supplementary assumption that $x_i \geq 0$ for all $i = 1, .., n$.

The above form of the LP contains both equalities and inequalities. The first step that must be done is to transform the problem into an equivalent one which contains only equality constraints and is then said to be in *canonical form.*

Definition 14. *The LP problem is in canonical form if it written with $n+m$ variables and m equations*

$$minc^T x$$

and

$$\begin{cases} a_{11}x_1 + ... + a_{1n}x_n \pm x_{n+1} - b_1 = 0 \\ a_{21}x_1 + ... + a_{2n}x_n \pm x_{n+2} - b_2 = 0 \\ ... \\ a_{m1}x_1 + ... + a_{mn}x_n \pm x_{n+m} - b_m = 0 \\ x_i \geq 0, i = 1, 2, ..., n + m. \end{cases}$$

with $b_i \geq 0$. The varibles added (or substracted) to transform the inequalities into equalities are called slack variables. The variables $x_1, .., x_n$ are the main or structural variables. Any set of variables $x_j, j = 1, .., n+m$ satisfing the equalities of the canonical form is called a solution and when in addition they are positive the solution is called feasible or realizable.

Example Transform the following LP into canonical form.

$$min \quad 2x_1 - x_2,$$

subject to

$$\begin{cases} 3x_1 - 2x_2 \geq -2 \\ 2x_1 - 4x_2 \leq 3 \\ x_1 + x_2 \leq 6 \\ x_1 \geq 0 \\ x_2 \geq 0. \end{cases}$$

Note that $b_1 = -2 < 0$ but the constants must also be positive so the first step consistis in changing the first inequality

$$\begin{cases} -3x_1 + 2x_2 \leq 2 \\ 2x_1 - 4x_2 \leq 3 \\ x_1 + x_2 \leq 6 \\ x_1 \geq 0 \\ x_2 \geq 0. \end{cases}$$

now add x_3, x_4, x_5 to get equalities

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2 \\ 2x_1 - 4x_2 + x_4 = 3 \\ x_1 + x_2 + x_5 = 6 \\ x_i \geq 0. \end{cases} .$$

With m slack variables, a basis is constituted of a set of m variables whose coefficients in the m equations form a square nonsingular (of nonzero determinant) matrix. These m variables are called basic variables. The other variables are nonbasic variables. If the basic variables are all positive or zero, they are called feasible basic solution. When the basic variables are strictly positive, they are nondegenerate.

A feasible solution can be found by setting the non basic variables equal to zero, hence we are left with m equations and m unknowns corresponding to a square matrix nonsignular so their values are determined uniquely.

For instance in the previous example, we immediately see that x_3, x_4, x_5 are basic and if we set $x_1 = x_2 = 0$ we obtain $f(0, 0) = 0$ and $x_3 = 2, x_4 = 3, x_5 = 6$. The costraints are satisfied so $(0, 0, 2, 3, 6)$ is a feasible solution (the value of f is not necessarily the minimum).

If it happens that by adding slack variables and setting the non-basic variables to 0 the solution has negative components we add as many new variables as the ones found negative (called artificial variables) in their corresponding compartments. In that, the number of equations does not change and we can transfer the variables that were negative to the non basic (and set them to 0) while we will be left with the same number of basic variables as previously (basis comprised of the new ones).

Consider for instance

$$min - 4x_1 - 5x_2$$

subject to

$$\begin{cases} 2x_1 + x_2 \leq 6 \\ x_1 + 2x_2 \leq 5 \\ x_1 + x_2 \geq 6 \\ x_1 + 4x_2 \geq 2 \\ x_1 \geq 0 \\ x_2 \geq 0. \end{cases}$$

All $b_i > 0$ so there is no need to change the inequalities, add simply the slack variables

$$\begin{cases} 2x_1 + x_2 + x_3 - 6 = 0 \\ x_1 + 2x_2 + x_4 - 5 = 0 \\ x_1 + x_2 - x_5 - 6 = 0 \\ x_1 + 4x_2 - x_6 - 2 = 0 \\ x_i \geq 0 \end{cases} .$$

First, a feasible solution must be found. The first trial consists in taking the nonbasic variables $x_1 = 0$ and $x_2 = 0$ and calculating the basic variables from x_3 to x_6. We notice that the obtained solution cannot be accepted as two variables, x_5 and x_6, would take negative values. This first trial having failed, we use the method of artificial variables. On both equations that gave negative basic variables, we add an artificial variable which must be positive or zero.

$$\begin{cases} x_1 + x_2 - x_5 + x_7 - 6 = 0 \\ x_1 + 4x_2 - x_6 + x_8 - 2 = 0 \\ x_i \geq 0 \end{cases}$$

We still take x_1 and x_2 as nonbasic variables, to which we add x_5 and x_6, which posed a difficulty. All the nonbasic variables are set equal to zero. $x_1 = x_2 = x_5 = x_6 = 0$. The value of the basic variables results $x_3 = 6, x_4 = 5, x_7 = 6, x_8 = 2$.

2.3.1 The simplex method

On way to solve the LP problem, once it is written in the form

$$minc^T x$$

under the constraint $Ax = b$ and $x_i \geq 0$ is with Lagrange multipliers. Note that in that case we also have the inequality constraints $x_i \geq 0$ that is why the Lagrangian that we consider is the function

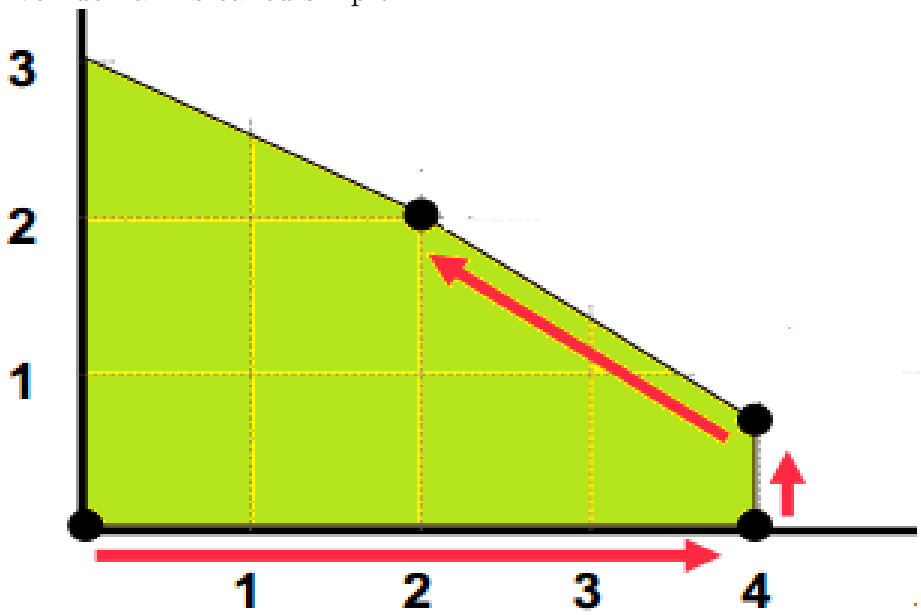
$$L(x, \lambda, \mu) = c^T x + \lambda^T (Ax - b) - \mu^T x,$$

μ is called the Karush–Kuhn–Tucker (KKT) parameter. Finding the stationary points comes to

$$\begin{cases} \frac{\partial L}{\partial x_i} = 0, for\ all\ i \\ \frac{\partial L}{\partial \lambda_i} = 0, for\ all\ i \\ \mu_i \geq 0 \\ x_i \geq 0 \\ \mu_i x_i = 0. \end{cases}$$

Another way to solve it is by using the simplex method that is explained in this section.

According to a basic theorem of linear programming, a feasible basic solution of the linear programming problem is a vertex of the convex polytope of the feasible solutions. This convex domain is called simplex.



If this is the feasible set then the simplex method works as follows: start at any corner of the simplex and evaluate the value of the objective function there. Follow an edge where the value of f is continually decreasing until you arrive at the next corner. If such corner does not exist stop, f is unbounded. Otherwise continue until no adjacent corner has a better objective value.

Description

After the addition of the slack variables, we identify the basic and the non-basic variables and we express everything (f and basic variables) in terms of the non-basic variables. Then, based on the expression of f, we see which variable can give a better value of the objective function when increasing it. The variables that don't, remain in non basic (since we wish to make them equal to zero). Among those that improve f, choose the one with the best improvement and atribute to it a new value based on the constraints (the minimum possible is taken so that no constraint is violated).

Example Consider an example already treated and in canonical form

$$min 2x_1 - x_2$$

under the constraints

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2 \\ 2x_1 - 4x_2 + x_4 = 3 \\ x_1 + x_2 + x_5 = 6 \\ x_i \geq 0. \end{cases} .$$

Start by identifing one feasible solution (in that example it is $(0, 0)$ which is always one corner of the feasible set and any other vertex can only have larger values of x_1, x_2). As previously explained, the basic variables are x_3, x_4, x_5 so express everything in terms of x_1, x_2. Since f is already written in terms of x_1, x_2 we only need to rewrite the constraints

$$\begin{cases} x_3 = 2 + 3x_1 - 2x_2 \\ x_4 = 3 - 2x_1 + 4x_2 \\ x_5 = 6 - x_1 - x_2 \\ x_i \geq 0. \end{cases} .$$

An increase of x_1 can only increase f, so keep it in the non basic variables and set it to 0. On the contrary, we can increase x_2 but not in any proportion. All variables must be positive so

$$\begin{cases} 2 - 2x_2 \geq 0 \\ 3 + 4x_2 \geq 0 \\ 6 - x_2 \geq 0 \end{cases} .$$

The boundary values are $x_2 = 1, x_2 = -\frac{3}{4}, x_2 = 6$. We keep $x_2 = 1$ hence no constraint is violated. A new estimate of the solution is found

$$x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 7, x_5 = 5.$$

The new nonbasic variables are thus x_1 and x_3, and the basic variables are x_2, x_4 and x_5. The function is now $f = -1$. Now, express the basic variables with respect to the nonbasic variables.

$$\begin{cases} x_2 = 1 + 1.5x_1 - 0.5x_3 \\ x_4 = 7 + 4x_1 - 2x_3 \\ x_5 = 5 - 2.5x_1 + 0.5x_3 \end{cases}$$

and the function f

$$f = -1 + 0.5x_1 + 0.5x_3$$

If we increase x_1 or x_3 f increases so there is no variable that can improve f and we are done.

Exercice Solve the following problem with the simplex method

$$max f = 2x_1 - x_2$$

subject to

$$\begin{cases} -3x_1 + 2x_2 \leq 2 \\ 2x_1 - 4x_2 \leq 3 \\ x_1 + x_2 \leq 6 \\ x_i \geq 0 \end{cases}$$

The operations that we do correspond to elementary operations between the rows, that is the well known Gauss elimination method. To automatize the algorithm we will work with matrices, as in Gauss, that is why we will arrange our equation in a matrix called the simplex table.

Simplex table

Once the system is put into canonical form it is possible to represent the stages in terms of martices. Create a matrix wich includes the system, an extra column where we note the basis and an extra line where we put the coefficients of f.

$$max 2x_1 - x_2.$$

Basis	x_1	x_2	x_3	x_4	x_5	b
x_3	-3	2	1	0	0	2
x_4	2	-4	0	1	0	3
x_5	1	1	0	0	1	6
f	2	-1	0	0	0	0

Table 2.1: Simplex table first iteration.

1. In the first step check the final line of the table and find the largest positive (in the case of maximization) or the smallest negative (in the case of minimization) coefficient of f. It corresponds to the variable with the best contribution that will move to the basis. (We put one new variable per step to the basis).

   If no such coefficient exists the solution is optimal.

2. We must find which variable will move out of the current basis. To do so, divide each element of the vector b by the corresponding elements of the chosen vector (to be moved into the basis). Among all the strictly positive ratios, choose the smallest one. (At this stage we take into account the constraints and the limitations they pose). Indeed, when we decide which variable will become basic, the others previously non basic will be set to 0. The constraints give the acceptable

value of the new basic variable (so that they are not violated) by taking the smallest (positive) that results from them. The corresponding constraint line makes one (previously basic) variable equal to 0 and consequently this is the variable that will become non basic and will be replaced. In the end of this step change the basic variables of the first column.

If all ratios are strictly negative, then no constraint is limiting. Consquently, there is nothing that prevents us from increasing as much as we want the variable which is thus nonlimited. The domain is unbounded, stop the problem.

3. The element in the intersection of the row of the variable leaving the basis and the column of the variable entering the basis is the pivot element.

4. For the next iteration apply Gauss's algorithm to make zeros above and below the pivot element. Once this is done, the last line gives the new coeffients of f in terms of the new non basic variables. Repeat from step 1 until the optimum is found (all coefficients negative in case of max and all positive in case on min).

The next simplex table is

Basis	x_1	x_2	x_3	x_4	x_5	b
x_3	0	-4	1	1.5	0	6.5
x_1	1	-2	0	0.5	0	1.5
x_5	0	3	0	-0.5	1	4.5
f	0	3	0	-1	0	-3

Table 2.2: Simplex table second iteration.

The highest contribution corresponds now to x_2 and the smallest quotient to x_5 which will leave from the basis and will be replaced by x_2. The pivot element is 3 and we apply the Gauss reduction method to make zeros above and below it.

Basis	x_1	x_2	x_3	x_4	x_5	b
x_3	0	0	1	0.833	1.333	12.5
x_1	1	0	0	0.167	0.667	4.5
x_2	0	1	0	-0.167	0.333	1.5
f	0	0	0	-0.5	-1	-7.5

Table 2.3: Simplex table third iteration.

All coefficients are negative so no contribution can improve f. We are done. The solution reads as follows:

$$f = -0.5x_4 - 1.5x_5 + 7.5$$

and $x_1 = 4.5$, $x_2 = 1.5$, $x_3 = 12.5$, $x_4 = x_5 = 0$ with $maxf = 7.5$.

Remarks

- If several variables are likely to be selected for the pivot rule, choose the entering variable which has the smallest index (similarly for the exiting variable).
- A basic solution is called degenerate if one or several basic variables are equal to 0. The degeneracy is not an exception, but it is the rule.
- It may occur that the optimal solution is not restricted to a single point but is constituted by a whole hyperplane when the function to minimize is parallel to that hyperplane. The degeneracy is called of first kind.

Exercice

	P_1	P_2	P_3
parts C_1	1	2	4
parts C_2	2	1	2
parts C_3	3	2	2

A company constructs three products P_1, P_2, P_3 using three materials C_1, C_2 and C_3. The company has no more than 70 of C_1, 80 of C_2 and 60 of C_3 per week. The marginal costs are 3 euros for P_1, 5 euros for P_2 and 6 euros for P_3. Present the canonical form of the LP. Find the optimal solution with the simplex method.

Proof. We put the data in equations :

$$\begin{cases} P_i \geq 0 \\ P_1 + 2P_2 + 4P_3 \leq 70 \\ 2P_1 + P_2 + 2P_3 \leq 80 \\ 3P_1 + 2P_2 + 2P_3 \leq 60 \end{cases}$$

with total gain $f = 3P_1 + 5P_2 + 6P_3$. The canonical form is

$$\begin{cases} P_i \geq 0 \\ P_1 + 2P_2 + 4P_3 + e_1 = 70 \\ 2P_1 + P_2 + 2P_3 + e_2 = 80 \\ 3P_1 + 2P_2 + 2P_3 + e_3 = 60 \\ maxf = 3P_1 + 5P_2 + 6P_3 \end{cases}$$

The simplex method produces the following tables

Basis	P_1	P_2	P_3	e_1	e_2	e_3	b
e_1	1	2	4	1	0	0	70
e_2	2	1	2	0	1	0	80
e_3	3	2	2	0	0	1	60
f	3	5	6	0	0	0	0

Table 2.4: Simplex table first iteration.

Basis	P_1	P_2	P_3	e_1	e_2	e_3	b
P_3	0.25	0.5	1	0.25	0	0	17.5
e_2	1.5	0	0	-0.5	1	0	45
e_3	2.5	1	0	-0.5	0	1	25
f	1.5	2	0	-1.5	0	0	-105

Table 2.5: Simplex table second iteration.

Basis	P_1	P_2	P_3	e_1	e_2	e_3	b
P_3	-1	0	1	0.5	0	-0.5	5
e_2	1.5	0	0	-0.5	1	0	45
P_2	2.5	1	0	-0.5	0	1	25
f	-3.5	0	0	-0.5	0	-2	-155

Table 2.6: Simplex table third iteration.

Solution

$$f = -3.5P_1 - 0.5e_1 - 2e_3 + 155$$

with $P_1 = 0 = e_1 = e_3$, $P_2 = 25, P_3 = 5, e_2 = 45$ and $maxf = 155$. □

Slack and artificial variables

Consider the problem

$$max 4x_1 + 5x_2$$

subject to

$$\begin{cases} 2x_1 + x_2 \leq 6 \\ x_1 + 2x_2 \leq 5 \\ x_1 + x_2 \geq 1 \\ x_1 + 4x_2 \geq 2 \\ x_i \geq 0 \end{cases}$$

Introduce the slack variables to put it into canonical form

$$\begin{cases} 2x_1 + x_2 + e_1 = 6 \\ x_1 + 2x_2 + e_2 = 5 \\ x_1 + x_2 - e_3 = 1 \\ x_1 + 4x_2 - e_4 = 2 \\ x_i \geq 0, e_i \geq 0 \end{cases}$$

A first feasible solution must be found. Trying to put $x_1 = x_2 = 0$ for the non basic variables yields $e_1 = 6, e_2 = 5, e_3 = -1, e_4 = -2$ where two solutions are negative so the

solution is not in the feasible set. We have to find one feasible solution to initialize the algorithm. For the two compartments that gave negative solutions we will introduce two more artificial variables, e_5, e_6 asking them to be positive. These two variables can replace e_3, e_4 in the basis.

$$\begin{cases} 2x_1 + x_2 + e_1 = 6 \\ x_1 + 2x_2 + e_2 = 5 \\ x_1 + x_2 - e_3 + e_5 = 1 \\ x_1 + 4x_2 - e_4 + e_6 = 2 \\ x_i \geq 0, e_i \geq 0 \end{cases}$$

with basic e_5, e_6 and the variables e_3, e_4 (that posed problem) become non basic so they can be set equal to zero giving

$$e_3 = e_4 = x_1 = x_2 = 0, e_1 = 6, e_2 = 5, e_5 = 1, e_6 = 2.$$

The obtained solution is not feasible for the main problem since the constraints are not satisfied.

Note that it is sufficient to solve the problem

$$mine_5 + e_6$$

under the same constraints. If the minimum is zero (which implies from positivity that $e_5 = e_6 = 0$) the solution of this system gives a feasible solution of the initial problem. Otherwise, if the minimum is positive the artificial variables cannot be zero giving a non feasible solution of the initial problem which consequently has no solution. We will solve the new problem using the simplex table to see if we can find a feasible solution.

$$minf_2 = e_5 + e_6$$

under

$$\begin{cases} 2x_1 + x_2 + e_1 = 6 \\ x_1 + 2x_2 + e_2 = 5 \\ x_1 + x_2 - e_3 + e_5 = 1 \\ x_1 + 4x_2 - e_4 + e_6 = 2 \\ x_i \geq 0, e_i \geq 0 \end{cases}$$

since the non basic are x_1, x_2, e_3, e_4 the objective function must be written in terms of these variables

$$f_2 = (1 - x_1 - x_2 + e_3) + (2 - x_1 - 4x_2 + e_4) = 3 - 2x_1 - 5x_2 + e_3 + e_4.$$

Basis	x_1	x_2	e_1	e_2	e_3	e_4	e_5	e_6	b
e_1	2	1	1	0	0	0	0	0	6
e_2	1	2	0	1	0	0	0	0	5
e_5	1	1	0	0	-1	0	1	0	1
e_6	1	4	0	0	0	-1	0	1	2
f_2	-2	-5	0	0	1	1	0	0	-3
e_1	1.75	0	1	0	0	0.25	0	-0.25	5.5
e_2	0.5	0	0	1	0	0.5	0	-0.5	4
e_5	0.75	0	0	0	-1	0.25	1	-0.25	0.5
x_2	0.25	1	0	0	0	-0.25	0	0.25	0.5
f_2	-0.75	0	0	0	1	-0.25	0	1.25	-0.5
e_1	0	0	1	0	2.33	-0.33	-2.33	0.33	4.33
e_2	0	0	0	1	0.67	0.33	-0.67	-0.33	3.67
x_1	1	0	0	0	-1.33	0.33	1.33	-0.33	0.67
x_2	0	1	0	0	0.33	-0.33	-0.33	0.33	0.33
f_2	0	0	0	0	0	0	1	1	0

Table 2.7: Simplex table

Now that a solution is available for the main problem, the second simplex tableau is built by taking the part of the first tableau corresponding to the last iteration and by suppressing the columns of the artificial variables. Now, the function f is used in the way it was originally defined

$$max 4x_1 + 5x_2$$

under the constraints

$$\begin{cases} 2x_1 + x_2 + e_1 = 6 \\ x_1 + 2x_2 + e_2 = 5 \\ x_1 + x_2 - e_3 = 1 \\ x_1 + 4x_2 - e_4 = 2 \\ x_i \geq 0, e_i \geq 0 \end{cases} .$$

As the last simplex table indicates the basic variables are x_1, x_2, e_1, e_2 hence f must be expressed in terms of the non basic variables

$$f = 4(\frac{2}{3} + \frac{4}{3}e_3 - \frac{1}{3}e_4) + 5(\frac{1}{3} - \frac{1}{3}e_3 + \frac{1}{3}e_4) = \frac{13}{3} + \frac{11}{3}e_3 + \frac{1}{3}e_4$$

Basis	x_1	x_2	e_1	e_2	e_3	e_4	b
e_1	0	0	1	0	2.33	−0.33	4.33
e_2	0	0	0	1	0.67	0.33	3.67
x_1	1	0	0	0	−1.33	0.33	0.67
x_2	0	1	0	0	0.33	−0.33	0.33
f	0	0	0	0	3.67	0.33	−4.33
e_1	0	−7	1	0	0	2	2
e_2	0	−2	0	1	0	1	3
x_1	1	4	0	0	0	−1	2
e_3	0	3	0	0	1	−1	1
f	0	−11	0	0	0	4	−8
e_4	−3.5	0.5	1	0	0	1	1
e_2	0	1.5	−0.5	1	0	0	2
x_1	1	0.5	0.5	0	0	0	3
e_3	0	−0.5	0.5	0	1	0	2
f	0	3	−2	0	0	0	−12
e_4	0	0	−0.67	2.33	0	1	5.67
x_2	0	1	−0.33	0.67	0	0	1.33
x_1	1	0	0.67	−0.33	0	0	2.33
e_3	0	0	0.33	0.33	1	0	2.67
f	0	0	−1	−2	0	0	−16

All the coefficients of the row of f are negative. Thus, the maximum of f has been reached and it is equal to 16. We have

$$f = -e_1 - 2e_2 + 16$$

with values $e_1 = e_2 = 0, x_1 = 2.33, x_2 = 1.33, e_3 = 2.67, e_4 = 5.67$.

2.3.2 Duality

Let us pose the optimization problem (called primal)

$$max c^T x$$

subject to

$$Ax \leq b, x \geq 0$$

where the inequality is takes row per row. The dual problem associated to the primal is

$$min b^T w$$

subject to

$$A^T w \geq c, w \geq 0.$$

An easy way to remember the transformation it to take

$$\begin{bmatrix} A & b \\ c^T & 0 \end{bmatrix}$$

for the primal and transpose it

$$\begin{bmatrix} A^T & c \\ b^T & 0 \end{bmatrix}$$

for the dual.

Theorem 8. *The dual of the dual is the primal.*

Proof. Take the dual problem $minb^T w$ and write it as $max - b^T w$, subject to

$$-A^T w \leq -c, w \geq 0.$$

Take now its dual

$$min - c^T x$$

subject to

$$-Ax \geq -b, x \geq 0,$$

which is equivalent to

$$maxc^T x$$

under $Ax \leq b, x \geq 0$. □

Theorem 9. *If both primal and dual problems admit a feasible solution, each of them has a finite optimum and the optimal values of the objective functions are equal*

Remark If one of the two problems has a non finite solution then the other has no realizable solution.

To understand a bit better what happens, to maximize $c^T x$ we need to find the best upper bound for it, that is the smallest of all the upper bounds. We allege that the best upper bound is the minimum of $b^T w$. Indeed

$$Ax \leq b \implies w^T Ax \leq w^T b, since\ w \geq 0.$$

When $A^T w \geq c$ then $(A^T w)^T \geq c^T$ or $w^T A^T \geq c^T$. Consequently

$$c^T x \leq w^T Ax \leq w^T b = (b^T w)^T$$

which proves that $b^T w$ is indeed an upper bound and to choose the best such upper bound we take its minimum.

- To the constraint i of the primal problem, corresponds the variable w_i of the dual and vice versa.

- When one variable of the dual is 0, an associated constraint is saturated and the marginal cost (coefficient of the objective function of the variable in the column corresponding to this variable) is equal in absolute value to the corresponding dual variable.
- If a dual variable is 0, the corresponding primal variable is non zero (a constraint in not saturated).

Example A horticulturist wants to plant at least 504 apple trees, 256 cherry trees and 420 pear trees. A supplier suggests two possible packs

A. With 4 apple trees, 5 cherry trees, 2 pear trees with unit price 35 euros.

B. With 3 apple trees, 1 cherry tree, 5 pear trees with unit price 34 euros.

Determine the number of A and B to minimize the cost of the operation.

Proof. Let us denote by x_1 the number of packs A and x_2 the number of packs B he buys. Then he will have $4x_1 + 3x_2$ aplle trees, $5x_1 + x_2$ cherry trees and $2x_1 + 5x_2$ pear trees. The quantities he needs satisfy

$$\begin{cases} 4x_1 + 3x_2 \geq 504 \\ 5x_1 + x_2 \geq 256 \\ 2x_1 + 5x_2 \geq 420 \\ minf = 35x_1 + 34x_2 \end{cases}$$

If we put the problem into canonical form by introducing $e_1, e_2, e_3 \geq 0$

$$\begin{cases} 4x_1 + 3x_2 - e_1 = 504 \\ 5x_1 + x_2 - e_2 = 256 \\ 2x_1 + 5x_2 - e_3 = 420 \\ minf = 35x_1 + 34x_2 \end{cases}$$

we end up with negative $e_i, i = 1, 2, 3$ and artificial variables are needed. In such a case we can also pass from the dual program. The primal will be transformed by considering y_1 the price of one apple tree, y_2 the price of one cherry tree and y_3 the price of one pear tree. Then

$$\begin{cases} 4y_1 + 5y_2 + 2y_3 \leq 35 \\ 3y_1 + y_2 + 5y_3 \leq 34 \\ maxf = 504y_1 + 256y_2 + 420y_3 \end{cases}$$

by assuming that the supplier fixes the maximum price possible (that is exactly 35 and 34 in the primal case). We add u_1, u_2 two slack variables to put this problem in canonical form.

$$\begin{cases} 4y_1 + 5y_2 + 2y_3 + u_1 = 35 \\ 3y_1 + y_2 + 5y_3 + u_2 = 34 \\ maxf = 504y_1 + 256y_2 + 420y_3 \end{cases}.$$

- Each equation i of the primal corresponds to y_i of the dual. Indeed, the first equation of the primal for instance, regards the number of apple trees and y_1 is the price of each apple tree.
- The coefficient 504 of the dual's objective function is the quantity of apples needed and it corresponds in absolute value to e_1 of the first step. Similarly for the other coefficients. We interpret in the same way the coefficients of the objective function in the last step (and all steps).

The simplex table is

Basis	y_1	y_2	y_3	u_1	u_2	b
u_1	4	5	2	1	0	35
u_2	3	1	5	0	1	35
f	504	256	420	0	0	0
y_1	1	5/4	1/2	1/4	0	31/4
u_2	0	$-11/4$	7/2	$-3/4$	1	31/4
f	0	-374	168	-126	0	-4410
y_1	1	23/14	0	5/14	$-1/7$	107/14
y_3	0	$-11/14$	1	$-3/4$	2/7	31/14
f	0	-242	0	-90	-48	-4782

Table 2.8: Simplex table.

Hence

$$f = -242y_2 - 90u_1 - 48u_3 + 4782$$

and $y_2 = u_1 = u_3 = 0$, $maxf = 4782$, $y_1 = 107/14, y_3 = 31/14$. But the objective was to solve the primal. The association of variables is

$$y_i \leftrightarrow e_i, i = 1, 2, 3, x_i \leftrightarrow u_i, i = 1, 2.,$$

so $x_1 = 90, x_2 = 48, e_1 = 0 = e_3, e_2 = 242$.

Note that in the dual case the problem reads as : A supplier wants to sell separately his trees with prices y_1, y_2, y_3 for each apple, cherry and pear tree respectively. He offers two packs A, B with maximal price 35 and 34 euros per pack. Pack A is comprised of 4 apple trees, 2 cherry trees and 2 pear trees and pack B of 3 apple trees, 1 cherry tree and 5 pear trees. He wishes to sell 504 apple trees, 256 cherry trees and 420 pear trees. What is the price $y_i, i = 1, 2, 3$ should he fix to maximize his profit?

□