Handout: A first approach of optimization

Haberkorn Thomas

July 2005

Contents

1 Differentiation: or how to compute speed 3
  1.1 Instantaneous vs. Average speed 3
  1.2 Velocity of a particle which evolution is linear 4
  1.3 Differentiation and speed (scalar case) 4
  1.4 Vector case 5
  1.5 Acceleration 6
  1.6 Main properties of differentiation 6
  1.7 An example: simplified satellite 7

2 Integration 8
  2.1 Antiderivative 8
  2.2 The sigma notation 9
  2.3 Area 10
  2.4 How to determine position with velocity 12

3 Extrema of a function 12
  3.1 Continuity and differentiability 12
  3.2 What does $f'$ and $f''$ says about $f$ 13
  3.3 Maximum and minimum values 15
  3.4 Finding extrema 16
  3.5 Some Maximum and minimum problems 17

4 Numerical aspect 21
  4.1 Representation of the numbers 21
  4.2 Numerical differentiation 22
  4.3 Numerical integration 23
  4.4 Finding zeros: The dichotomy 23
  4.5 Finding zeros: Newton’s method 24

5 Modeling and the different kinds of optimization problems 24
  5.1 Modeling 24
  5.2 Parameter optimization problem 27
  5.3 Mixed Integer Optimization Problem 28
  5.4 Optimal Control Problem 28
  5.5 An example of two close problem with different modeling 29

6 Some Optimal Control Problem 32
  6.1 The Brachystochrone 32
  6.2 The AUV 33
  6.3 Low-thrust orbital transfer 38
Introduction

Optimal control, or more generally optimization is a very useful mathematical tool. It consists of minimizing a function (the cost or criterion) subject to some constraints. A very common optimization problem is the one of minimizing the time needed for a controlled mechanical system to go from one point to another, or to minimize the energy needed for this same motion. For example, the controlled mechanical system could be an Autonomous Underwater Vehicle (AUV) that we want to steer from one position to another by applying a sequence of controls (the thrusts of the propellers) that will do the job in the minimum time. Figure 1 shows two way of steering a spherical AUV from the origin to the position $(10, 8, 5)$ (in meters).

Figure 1: Two trajectories for ODIN to go from the origin to the position $(10, 8, 5)$ (m).

In this figure, it is quite obvious that one trajectory is faster than the other, but in general it is not always the shortest path (in terms of distance) that is the shortest in time since all directions of motion are not equivalent.

Another example could be the minimization of the transfer time of an orbitary satellite to go from one orbit to another, or the minimization of the fuel required to do the same change of orbit (cf. figure 2).

The minimum time problems are very common optimization problems. You can also encounter shape optimization problems (e.g. shape of the nose of an airplane in order to minimize the drag) and many others. Historically, the first interesting optimization problem is the search of the brachystochrone which corresponds to the search of the faster path connecting two position when considering a particle moving only thanks to gravity. This problem will be presented in the last section of this document.

The first section of this document is dedicated to the notion of differentiation as the way to compute speed of a moving object. The second section introduces the concept of integration with the help of the area of a surface. In the third section, we use the differentiation to find extrema of functions. We then explain, in the fourth section, how to numerically compute derivatives, integrals, and zeros of a function. The fifth section outlines the importance of the modeling step in optimization problem. Then, we present different kinds of optimization problems, their formulations, and the principles of their numerical solving. The final section gives some examples of optimal control problem and their solutions.
1 Differentiation: or how to compute speed

1.1 Instantaneous vs. Average speed

Imagine a car that is going from town A to town B by following different kinds of roads (highway, trail ...etc). Say that the distance (with the chosen itinerary) between the two towns is $d$ km, and the car spends $T$ hours on the road. Figure 3 illustrates the trip of the car.

It is quite obvious that the average speed, $\bar{v}$ of the car is equal to the total distance divided by the total time, that is:

$$\bar{v} = \frac{d}{T},$$

where $\bar{v}$ is expressed in km/h. Since the state of the road is not uniform over the whole trip, it is doubtful that the speed $v_1$ at time $t_1$ is the same as the speed $v_2$ at time $t_2$. At each time $t$ of the trip, we can associate a speed $v(t)$ at this specific time. We say here that $v(t)$ is the instantaneous speed of the car at time $t$. Actually, we usually say that $v(t)$ is the speed of the car at time $t$ since we are generally considering instantaneous (or local) speed. Henceforth we will specify when we are dealing with average speed.

Note that knowing the average speed $\bar{v}$ does not give us the speed at a specific time $t$. And knowing the speed at time $t$ does not tell us what was or what will be the speed at another time $\tilde{t}$. However,
the knowledge of the speed and position of the car at time \( t \) can be used to approximate (more or less accurately) the position of the car at another time \( \hat{t} \) which is close to \( t \) (this is one of the ideas of the integration operation).

In physics and mathematics, we reserve the word speed for the positive value of \( v \). And we call \( v \) (which can be positive or negative, for example if the car is on a rear motion) the velocity.

We will now try to be more general in our definition of velocity.

### 1.2 Velocity of a particle which evolution is linear

Consider a particle moving on a straight line according to the equation \( r(t) = at + b \) where \( r \) is the position in meters of the particle, \( t \) is the time in second and \( a \) and \( b \) are some constants. Figure 4 shows the evolution of the position of this particle with respect to time.

![Figure 4: \( r(t) = at + b \).](image)

The velocity of the particle is the rate of change of its position with respect to the time. Between time \( t_1 \) and \( t_2 \) \( (t_2 > t_1) \), the particle goes from position \( r(t_1) \) to position \( r(t_2) \). The slope of the line connecting \( (t_1, r(t_1)) \) to \( (t_2, r(t_2)) \) is:

\[
\delta r = \frac{r(t_2) - r(t_1)}{t_2 - t_1} = \frac{at_2 + b - (at_1 + b)}{t_2 - t_1} = a
\]

So, the slope \( \delta r \) does not depend on times \( t_1 \) and \( t_2 \). And the rate of change of the position of the particle is always the same, that means that the velocity of the particle is constant and is equal to \( a \) \( (m.s^{-1}) \).

### 1.3 Differentiation and speed (scalar case)

Actually, we have a constant velocity because the evolution of the particle with respect to time is linear. Usually, when computing speed of an object, one has a less trivial evolution. Figure 5 shows an example of nonlinear evolution of a particle with respect to time.

Here the rate of change of the position of the particle at time \( t_0 \) is given by the slope of the tangent to the curve at the point \( (t_0, r(t_0)) \). An approximation of this slope is given by the line connecting \( (t_0, r(t_0)) \) to another position of the particle at a time \( t_1 \) close to \( t_0 \), and is equal to \( (r(t_1) - r(t_0))/(t_1 - t_0) \). The closer the time \( t_1 \) is to time \( t_0 \), the more accurate the approximation will be. Ultimately, the idea is that the velocity of the particle at time \( t_0 \) is its rate of change when considering the time \( t_1 \) as close as possible to \( t_0 \). The operation of taking the closest time \( t_1 \) to \( t_0 \) is the one of passage to the limit. If we note \( v(t_0) \) the velocity of the particle at time \( t_0 \), we note:

\[
v(t_0) = \lim_{t_1 \to t_0} \frac{r(t_1) - r(t_0)}{t_1 - t_0}
\]

we say here that \( v(t_0) \) is equal to the limit, when \( t_1 \) tends toward \( t_0 \), of \( (r(t_1) - r(t_0))/(t_1 - t_0) \).

Note that if we let \( h = t_1 - t_0 \), we have that:
Figure 5: Example of evolution of a particle with respect to time.

\[ v(t_0) = \lim_{h \to 0} \frac{r(t_0 + h) - r(t_0)}{h} \]  
Equation (3)

The velocity \( v(t_0) \) is more generally called the derivative of the function \( r(.) \) at time \( t_0 \), and is usually denoted \( r'(t_0) \). In the specific case of (3), the operation of passage to the limit is called differentiation.

In this example, \( v(t_0) \) depends on the chosen time \( t_0 \), and it is possible (under some regularity assumptions) to define a function \( v(.) \) that associates its velocity to each time \( t \). And we have:

\[ v(t) = r'(t) = \lim_{h \to 0} \frac{r(t + h) - r(t)}{h} \]  
Equation (4)

Other notations for the derivative are:

\[ r'(t) = \frac{dr}{dt} = Dr \]

The first notation is from Leibniz, and means that the derivative represents the change in \( r \) for an infinitesimal time step \( dt \). The second notation, put simply that \( D \) is the differentiation operator.

Now, we extend our definition of velocity to motions that are not necessarily in a straight line.

1.4 Vector case

To define velocity, we do not have to restrict ourselves to the case of an object moving in a straight line. Indeed, we can consider objects moving in the \( \mathbb{R}^2 \) plane, or points moving in a 3 dimensional space (\( \mathbb{R}^3 \)). For such objects, it is of course still possible to define and compute the velocity. And this computation still involves the differentiation operator.

Consider a particle moving in the \( \mathbb{R}^2 \) plane and whose position is given by \( r(t) = (x(t), y(t)) \) where \( x(t) \) and \( y(t) \) are real functions (that is \( x(t) \in \mathbb{R} \) and \( y(t) \in \mathbb{R} \) for all \( t \in \mathbb{R} \)). Similarly to the scalar case, we define the velocity \( v(t) \) of this particle to be the derivative \( r'(t) \):

\[ r'(t) = (x'(t), y'(t)) \]  
Equation (5)

In other words, the velocity of a particle moving in \( \mathbb{R}^2 \) is the pair of the velocities of each component of the motion of this particle.

We also define the speed of the particle as the length (or norm) of \( r'(t) \).

\[ |v(t)| = \sqrt{x'^2(t) + y'^2(t)} \]  
Equation (6)

We can similarly define the velocity and speed of a particle moving in the space \( \mathbb{R}^3 \) \( (r(t) = (x(t), y(t), z(t)), v(t) = r'(t) = (x'(t), y'(t), z'(t)) \) and \( |v(t)| = \sqrt{x'^2(t) + y'^2(t) + z'^2(t)} \).

Now, we define the notion of acceleration.
1.5 Acceleration

As speed is the rate of change of the position with respect to time, acceleration is the rate of change of the speed with respect to time. A first implication is that an acceleration is expressed in $\text{distance/time}^2$ units.

An acceleration that is experienced by everyone is the Earth gravity. This is approximately:

\[ g_0 \approx 9.8 \text{m/s}^2 \]  

(7)

As for the mathematical definition of the acceleration, it is quite straightforward using the same idea as velocity. Given a particle moving in a straight line according to the equation $r(t)$, its velocity $v(t) = r'(t)$ and we have $a(t) = v'(t) = r''(t)$.

Here $r''(t)$ means the derivative of the derivative of $r(.)$ and is called the second derivative of $r(.)$. Other notations for the second derivative are

\[ r''(t) = \frac{d^2r}{dt^2} = D^2r = r^{(2)}(t) \]

Similarly to the velocity, we can define acceleration for an object moving in $\mathbb{R}^2$ or $\mathbb{R}^3$ as the pair or triple of the acceleration of each of its components.

\[
\begin{align*}
\mathbb{R}^2 & \quad \mathbb{R}^3 \\
r(t) & = (x(t), y(t)) & r(t) & = (x(t), y(t), z(t)) \\
v(t) & = (x'(t), y'(t)) & v(t) & = (x'(t), y'(t), z'(t)) \\
a(t) & = (x''(t), y''(t)) & a(t) & = (x''(t), y''(t), z''(t))
\end{align*}
\]

1.6 Main properties of differentiation

We have defined the velocity and acceleration of an object in motion using the derivatives (first or second order) of the equations of motion of the considered object. To compute velocity and acceleration, we should be able to compute these derivatives.

We begin with the derivative of an elementary function. Let $f(t) = at^2 + bt + c$ where $a$, $b$ and $c$ are real constants. We previously defined $f'(t)$ as:

\[ f'(t) = \lim_{h \to 0} \frac{f(t + h) - f(t)}{h} \]

For the moment, it might not be very clear to you how the limit operation works. In this specific case, it will become easier once we developed and simplified the fraction:

\[
\begin{align*}
\frac{f(t+h) - f(t)}{h} & = \frac{a(t+h)^2 + b(t+h) + c - at^2 - bt - c}{h} \\
& = \frac{a(t^2 + 2ah + h^2) + b(t + h) + c - at^2 - bt - c}{h} \\
& = 2at + ah \\
\end{align*}
\]

Now, if we let $h$ tend toward 0, it simply corresponds in our case to replace $h$ by 0, which gives us:

\[ f'(t) = 2at + b \]

More generally, we have that for all $n \in \mathbb{Z} - \{0\}$

\[ (x^n)' = nx^{n-1} \]

The trigonometric functions $\cos$ and $\sin$ have the following derivatives:

\[
\begin{align*}
\cos' t & = -\sin t \\
\sin' t & = \cos t
\end{align*}
\]

We now give some rules of differentiation.
Proposition 1.1 (Differentiation rules). Given $u$ and $v$ are two real functions and $a$ is a constant. The following differentiation rules apply:

(i) $(au)' = au'$.

(ii) $(u + v)' = u' + v'$ (derivative of a sum).

(iii) $(uv)' = u'v + uv'$ (derivative of a product).

(iv) $(u(v))' = v'u'(v)$ (derivative of a composition).

With these rules, one can differentiate a lot of functions, including the polynomials and the composition of them with trigonometric functions.

We now apply the above notions to an example.

1.7 An example: simplified satellite

Consider a satellite in orbit around the Earth at an height $h$ 600 km. Assume that the orbit is circular and periodic of period $T = 96$ min and that the Earth’s radius is $R_{Earth} = 6400$ km. The question is: what is the velocity, speed, and acceleration of this satellite?

If we set $d = R_{Earth} + h$ the distance between the satellite and the center of the Earth, and $\theta(t)$ the angle (oriented) between the $Ox$ axis and the segment connecting $O$ to the satellite at time $t$, we can use the cosine and sine functions to define the position of the satellite as follows:

$$r(t) = (d \cos \theta(t), d \sin \theta(t))$$  \hfill (8)

This coordinate system is called the polar coordinate system.

Since the orbit is circular, it is obvious that the rate of change of $\theta$ is constant. We call $\omega$ the angular speed of the satellite and we have:

$$\theta(t) = \omega t$$  \hfill (9)

So,

$$r(t) = (d \cos \omega t, d \sin \omega t)$$  \hfill (10)

Using the given information, we can now compute the angular speed $\omega$. Indeed, we know that the orbit is periodic of period $T$, but we also know that cosine and sine are periodic of period $2\pi$. From this, we can deduce the following equation:
We now compute the velocity of the satellite using the results of the previous subsection.

\[ v(t) = r'(t) = ([d \cos \omega t]' , [d \sin \omega t]') = (-d\omega \sin \omega t, d\omega \cos \omega t) \]  \hspace{1cm} (12)

Note that the velocity is tangent to the orbit, which here (because the orbit is circular) means that the vector defined by \( v(t) \) is orthogonal to the radius (the vector joining the origin \( O \) and the satellite).

The speed is:

\[ |v(t)| = \sqrt{d^2 \omega^2 \sin^2 \omega t + d^2 \omega^2 \sin \omega t} = \frac{d\omega}{d\omega} \approx 27.49 \times 10^3 \text{ km.h}^{-1} \]  \hspace{1cm} (13)

The acceleration is the second derivative of the position and is equal to:

\[ a(t) = v'(t) = ([-d\omega \sin \omega t]' , [d\omega \cos \omega t]') = (-d\omega^2 \cos \omega t, -d\omega^2 \sin \omega t) \]  \hspace{1cm} (14)

The acceleration is orthogonal to the velocity and is in the opposite direction to the radius. The length (or norm) of this acceleration is:

\[ |a(t)| = \sqrt{(d\omega^2 \cos \omega t)^2 + (d\omega^2 \sin \omega t)^2} = \frac{d\omega^2}{d\omega^2} \approx 10.79 \times 10^4 \text{ km.h}^{-2} \approx 8.33 \text{ m.s}^{-2} \]  \hspace{1cm} (15)

Note that the norm of the acceleration is less than that at the surface of the Earth.

## 2 Integration

Integration can be viewed as the inverse of differentiation and as with differentiation, its use is not restricted to the domain of optimization.

In the case of controlled mechanical systems, integration is used to determine the position of the mechanical system. In physics, the given information is usually not the position but the acceleration (think of what you control when driving a car). Another application of integration is the computation of areas or volumes. It is with this approach that we will define the integral.

But first, let’s introduce the definition of antiderivative.

### 2.1 Antiderivative

When considering the position and velocity of an object, we defined the derivative by saying that the velocity is the derivative of the position. If we view the relation between position and velocity from the other side, we say that the position is one of the antiderivatives of the velocity.

To put it more rigorously, we have the following definition.

**Definition 2.1 (Antiderivative).** A function \( F \) is called an antiderivative of \( f \) if \( F'(t) = f(t) \) for all \( t \).

Note that the antiderivative of a function is not unique since \( F(t) + C \) (\( C \) constant) has the same derivative than \( F(t) \).

**Example 2.1 (Some usual antiderivatives).** (i) Since \( \cos t = -\sin t \), we can conclude that \( \cos t \) is an antiderivative of \( -\sin t \) (and \( \sin t \) is an antiderivative of \( \cos t \)).
(ii) Since \((t^n)' = nt^{n-1}\) for any \(n \in \mathbb{Z} - 0\). From this, we have that for any \(m \in \mathbb{Z} - \{-1\}\), \(t^{m+1}/(m+1)\) is an antiderivative of \(t^m\).

We also have antiderivative rules that are similar to the differentiation ones.

**Proposition 2.1 (Antiderivative rules).** Let \(F\) be an antiderivative of \(f\), \(G\) an antiderivative of \(g\) and \(c\) a real constant. Then:

(i) \(cF(t)\) is an antiderivative of \(cf(t)\).

(ii) \(F(t) + G(t)\) is an antiderivative of \(f(t) + g(t)\).

**Example 2.2.** Find an antiderivative \(F\) of the function:

\[ f(t) = 2\sin t - 4t^3 + 2/t^2 \]

Using the previous antiderivative rules we have:

\[ F(t) = 2(-\cos t) - 4(t^4/4) + 2(t^{-3}/(-3)) \]

\[ = -2\cos t - t^4 - 2/(3t^3) \]

We now introduce very useful notation that allows to write a sum into a contracted form.

### 2.2 The sigma notation

When evaluating integrals, we often use sums with many terms. A convenient way a writing such sums is to use the sigma notation (from the capital Greek letter \(\Sigma\)).

**Definition 2.2 (sigma notation).** If \(x_m, x_{m+1}, ..., x_n\) are real numbers and \(m\) and \(n\) are integers such that \(m \leq n\) then

\[ \sum_{i=m}^{n} x_i = x_m + x_{m+1} + ... + x_n \]

this operation is called a summation and \(i\) is called the index of summation.

**Example 2.3.** (a) \( \sum_{i=1}^{4} \cos i\pi = \sum_{j=1}^{4} \cos j\pi = \cos \pi + \cos 2\pi + \cos 3\pi + \cos 4\pi = 0 \)

(b) \( \sum_{i=0}^{5} i = 0^0 + 1^1 + 2^2 + 3^3 + 4^4 + 5^5 = 63 \)

(c) \( \sum_{i=1}^{n} i = 1 + 2 + ... + n = n(n + 1)/2 \)

(d) \( \sum_{i=1}^{n} 2 = 2 \sum_{i=1}^{n} 1 = 2n \)

Here are three rules for the sigma notation

**Proposition 2.2.** If \(c\) is a constant (does not depend on \(i\)), then

(i) \( \sum_{i=m}^{n} cx_i = c \sum_{i=m}^{n} x_i \)

(ii) \( \sum_{i=m}^{n} (x_i + y_i) = \sum_{i=m}^{n} x_i + \sum_{i=m}^{n} y_i \)

(iii) \( \sum_{i=m}^{n} (x_i - y_i) = \sum_{i=m}^{n} x_i - \sum_{i=m}^{n} y_i \)

We will now use the sigma notation to evaluate area.
2.3 Area

Considering a function \( f(x) > 0 \) on an interval \([a, b]\), we would like to evaluate the area between the curve \( y = f(x) \) and the segment \([a, b]\) of the \( x\)-axis. Figure 7 gives an example of such a function.

As shown in the figure, an approximation of the area could be the sum of the areas \( A_i \) of all the small rectangles. This approximated area \( \tilde{A}_n \) is equal to

\[
\tilde{A}_n = (x_2 - x_1)f(x_1) + (x_3 - x_2)f(x_2) + \cdots + (x_{n+1} - x_n)f(x_n)
\]

\[
= \sum_{i=1}^{n}(x_{i+1} - x_i)f(x_i) \tag{16}
\]

One can easily see that the greater \( n \) is, the closer \( \tilde{A}_n \) will be to the actual area, \( A \).

Example 2.4. Consider the function \( f(x) = 0.5x + 3 \). On the interval \([0, 10]\), the area defined by this function is the one of the rectangle which vertices are \((0, 0), (10, 0), (10, f(0)), (0, f(0))\) plus the one of the triangle which vertices are \((0, f(0)), (10, f(0)), (10, f(10))\) (cf figure 8). So this area \( A \) is

\[
A = 10f(0) + 10(f(10) - f(0))/2 = 30 + 25 = 55
\]

If we apply the formula (16) to the function \( f \), we obtain:

\[
\tilde{A}_2 = 5(f(0) + f(5)) = 42.5 \\
\tilde{A}_5 = 50 \\
\tilde{A}_{10} = 52.5 \\
\tilde{A}_{100} = 54.75
\]

We can see that we can accurately estimate to the real value of the area \( A \). Actually, we can be as close as we want providing that we take an integer \( n \) large enough.

This example leads us to the idea of a limit as \( n \) tends toward \( \infty \), and thus, to the following definition.
Definition 2.3 (Integral). Let $f$ be a positive function defined on an interval $[a, b]$. The area defined by the curve $y = f(x)$ on $[a, b]$ is

$$A = \lim_{n \to \infty} \tilde{A}_n = \int_a^b f(x)dx$$

The term on the right hand side is called the integral of $f$ over the interval $[a, b]$.

We will now connect the idea of antiderivatives with that of integrals. To do so, we begin with an example.

Example 2.5 (Integral of $f(x) = x$). The integral of the function $f$ from 0 to a real positive number $x$ is a function depending on $x$:

$$g(x) = \int_0^x t \, dt$$

Using approximate area $\tilde{A}_n$ of (16), we have:

$$\tilde{A}_n = \sum_{i=1}^n \frac{x(i-1)x}{n}$$

$$= \frac{x^2}{n^2} \sum_{i=1}^n (i - 1)$$

$$= \frac{x^2}{n^2} \frac{n - 1}{2}$$

$$= \frac{x^2}{2} - \frac{x}{n}$$

Even if the notion of limit had not been defined precisely, one can intuitively understand that when $n$ tends toward $\infty$, the fraction $\frac{n - 1}{n}$ tends toward 1 since the $-1$ in the numerator is neglectable when dealing with very large values of $n$ (actually, any constant is neglectable in this case).

So, we have:

$$g(x) = \lim_{n \to \infty} \tilde{A}_n = \frac{x^2}{2}$$

And $g$ is an antiderivative of $f$ since $g'(x) = f(x)$. Actually, $g$ is the unique antiderivative of $f$ that is equal to 0 at $x = 0$.

This result can be extended to all functions (under some regularity assumptions that we do not consider here).

Theorem 2.1. If $f$ is a (continuous) function on the interval $[a, b]$, then the function $g$ defined by:

$$g(x) = \int_a^x f(t) \, dt \quad a \leq x \leq b$$

is the unique antiderivative of $f(x)$ such that $g(a) = 0$.

This is a very important result that can be used in both directions. On one hand, it greatly eases the calculation of an integral which is tedious if only considering the limit of the approximated area. On the other hand, it gives us a way of evaluating (approximatively) antiderivatives of a function, even if this is not analytically easy.

Remark 2.1 (extension). Integrals are not restricted to positive functions, we first use this restriction to be able to approach the notion by the means of the area. Furthermore, not all functions are integrable, which means that integrals of some functions are not defined (typically a function with a non finite value at a given point, e.g. $x^{-1}$ at 0). Moreover, it is also possible to extend the notion of integral to non scalar functions by simply applying the integration componentwise.

We now use integrals to define position of an object for which we know the velocity.
2.4 How to determine position with velocity

Considering an object in motion with a known velocity \( v(t) = 2at \) and a known position at initial time \( r(0) = b \). The problem is to determine the position \( r \) of this object at any time \( t \).

Since we know that \( r(t) \) is such that \( r'(t) = v(t) \), we know that \( v(t) \) is an antiderivative of \( r(t) \). And, since the antiderivative of a function is defined up to a constant, we can write

\[
r(t) = \int_0^t v(s) \, ds + c \quad (c \text{ a constant})
\]

And the constant \( c \) is such that \( r(0) = b \), so \( c = b \), because an integral over an interval of length zero is equal to zero. So we have

\[
r(t) = at^2 + b
\]

More generally, we have the following result.

**Proposition 2.3 (position from velocity).** Let \( v(t) \) be the velocity of a moving object whose position at initial time \( 0 \) is \( r(0) \). Then, the position \( r(t) \) of this object at each time is:

\[
r(t) = r(0) + \int_0^t v(s) \, ds
\]

Similarly, we can determine the velocity if we know the acceleration.

**Proposition 2.4 (velocity from acceleration).** Let \( a(t) \) be the acceleration of a moving object which velocity at initial time \( 0 \) is \( v(0) \). Then, the velocity \( v(t) \) of this object at each time is:

\[
v(t) = v(0) + \int_0^t a(s) \, ds
\]

We now have defined the necessary mathematical tools to deal with controlled mechanical systems. The next section is a first step toward optimization, the search for extrema of scalar functions.

3 Extrema of a function

When dealing with optimization problems, one tries to find the minimum or maximum of a function with respect to some cost or criteria. In this section we first expose properties of functions. We then see what information can be extracted from the derivatives of a function. After that, we will give some leads on how to evaluate maximum or minimum values of a function. Finally we illustrate the search for extrema with some examples.

3.1 Continuity and differentiability

In the first two sections, we didn’t really care about the regularity of the functions we want to differentiate or integrate. Unfortunately, we need to be more rigorous about those properties, that’s why we begin by giving an idea (followed by one of the proper definitions) of what a continuous function is.

But first a definition.

**Definition 3.1 (Domain).** The domain of a function \( f \) is the set of numbers \( x \) such that this function is properly defined. In particular, it is the set of number \( x \) such that \( f(x) \) exists and is finite.

**Example 3.1.**

- The domain of definition of the function \( f(x) = ax^2 + bx + c \) \((a, b, c \in \mathbb{R})\) is \( \mathbb{R} \) since this polynomial is defined for all \( x \in \mathbb{R} \).

- The function \( f(x) = 1/x \) is not defined for \( x = 0 \) but is defined for every real number not equal to 0. So the domain of definition of this function is \( \mathbb{R} - \{0\} \) (also denoted \( \mathbb{R}^* \)).
• In general, a function $f(x) = 1/g(x)$ is defined everywhere except at numbers $\bar{x}$, such that $g(\bar{x}) = 0$.

**Remark 3.1.** The domain of definition of a vector function is the Cartesian product of the domain of definition of each of its components.

The continuity of a function $f(x)$ means that if two numbers $x_1$ and $x_2$ are close, then the two numbers $f(x_1)$ and $f(x_2)$ are also not too far from each other. Graphically, a function $f(x)$ is said to be continuous if you can draw its representative curve $y = f(x)$ with one single move (the pen do not leave the paper). More rigorously, we give the following definition.

**Definition 3.2 (Continuity at a point).** Let $f$ be a function (real valued), defined at the point $\bar{x}$, and on a neighborhood of $\bar{x}$ (open interval containing $\bar{x}$). Then, $f$ is said to be continuous at the point $\bar{x}$ if:

$$\forall \varepsilon > 0, \exists \eta > 0, |x - \bar{x}| \leq \eta \Rightarrow |f(x) - f(\bar{x})| \leq \varepsilon$$

The meaning of this definition might not seem very clear, so let’s look at it graphically (Figure 9).

![Figure 9: Illustration of continuity and discontinuity](image-url)

**Definition 3.3 (Continuity).** A function is continuous on an interval if it is continuous at each point of this interval.

Another regularity property of a function is its differentiability.

**Definition 3.4 (Differentiability).** A function $f$ is differentiable at a point $\bar{x}$ of its domain if the limit (3) exists (and is finite). A function is differentiable on an interval if it is differentiable at each point of this interval.

**Proposition 3.1.** If a function is differentiable, then it is continuous. And, a discontinuous function is not differentiable.

From now on, except when indicated, the considered function are infinitely differentiable (which means that the derivative of the function is differentiable and so on).

## 3.2 What does $f'$ and $f''$ says about $f$

The derivative of a function represents the slope of its tangent. So, from the value of the derivative we can gain insight about the monotonicity of the function.

**Proposition 3.2.** Let $f$ be a function defined on an interval $[a, b]$.

(i) If $f'(x) \geq 0$ for all $x \in [a, b]$ then $f$ is increasing on $[a, b]$.

(ii) If $f'(x) \leq 0$ for all $x \in [a, b]$ then $f$ is decreasing on $[a, b]$. 

13
Figure 10 illustrates this property. Here the function is increasing, then decreasing then once increasing again.

Since the second derivative of a function is the derivative of its derivative, it says something about the monotonicity of the first derivative. This says something about the curvature or concavity of the original function.

**Definition 3.5 (Concavity).** Let $f$ be a function defined on an interval $[a, b]$. Then,

(1) If $\forall (x, y) \in [a, b]^2, \forall t \in [0, 1], (1 - t)f(x) + tf(y) > f((1 - t)x + ty)$ then $f$ is concave upward on $[a, b]$.

(2) If $\forall (x, y) \in [a, b]^2, \forall t \in [0, 1], (1 - t)f(x) + tf(y) < f((1 - t)x + ty)$ then $f$ is concave downward on $[a, b]$.

This definition simply means that a concave upward (respectively downward) function is a function whose representing curve has all lines which connect two of its points (in $[a, b]$) above (respectively below) it. Figure 11 illustrate this definition.

We have the following property connecting the second derivative and the concavity.

**Proposition 3.3.** Let $f$ be a function defined on the interval $[a, b]$. Then,

(i) if $\forall x \in [a, b], f''(x) > 0$ then $f$ is concave upward on $[a, b]$.

(ii) if $\forall x \in [a, b], f''(x) < 0$ then $f$ is concave downward on $[a, b]$.
That means that if the second derivative of a function is positive (resp. negative), then its derivative is increasing (resp. decreasing) and then this function is concave upward (resp. downward). Observing a concave upward function, it is quite obvious that its derivative is increasing (the slope of the tangent is increasing).

**Example 3.2.** Take the function \( f(x) = x^3 \), it is an infinitely differentiable function (every polynomial is infinitely differentiable). Its derivative is \( f'(x) = 3x^2 \) and its second derivative is \( f''(x) = 6x \). So \( f'(x) \) is non negative on \( \mathbb{R} \), and then \( f \) is increasing on \( \mathbb{R} \). \( f''(x) \) is negative on \( \mathbb{R}^+ \) and positive on \( \mathbb{R}^- \) (\( = \) \( ]-\infty, 0[ \)) and positive on \( \mathbb{R}^+ \) (\( = (0, \infty) \)). So \( f \) is concave downward on \( \mathbb{R}^- \) and concave upward on \( \mathbb{R}^+ \).

We can use this information to find maximum and minimum values of a function.

### 3.3 Maximum and minimum values

**Definition 3.6 (Extrema).** A function \( f \) has a global (or absolute) maximum at \( x_M \) if \( f(x_M) \geq f(x) \) for all \( x \) in the domain \( D \) of \( f \). The number \( f(x_M) \) is called the maximum value of \( f \) on \( D \). Similarly, \( f \) has a global (or absolute) minimum at \( x_m \) if \( f(x_m) \leq f(x) \) for all \( x \) in \( D \). The number \( f(x_m) \) is called the minimum value of \( f \) on \( D \). The maximum and minimum values of \( f \) are called extreme values (or extrema) of \( f \).

**Definition 3.7 (Local maximum and minimum).** A function \( f \) has a local (or relative) maximum at \( \tilde{x}_M \) if \( f(\tilde{x}_M) \geq f(x) \) for all \( x \) near \( \tilde{x}_M \) (on an open interval containing \( \tilde{x}_M \)). Similarly, \( f \) has a local minimum at \( \tilde{x}_m \) if \( f(\tilde{x}_m) \leq f(x) \) for all \( x \) near \( \tilde{x}_m \).

Figure 12 shows a function \( f \) defined on \([a, b] \) with a global maximum at \( d \) and a global minimum at \( c \), and with local maximum and minimum (\( e, g, h \)).

![Figure 12: Illustration of global extrema and local one. On \([a, b] \), \( f(c) \) is the global minimum, \( f(d) \) is the global maximum, \( f(e) \) and \( f(h) \) are local minima and \( f(g) \) is a local maximum.](image)

Since optimization consists in finding the extrema of a function, it is interesting to know whether an extrema exists or not. Figure 13 shows some examples for which extrema don’t exist.

The function \( f(x) = x \) (called the identity) has no maximum or minimum values on \( \mathbb{R} \) since whatever number you choose, a larger number will have a higher value and a smaller number will have a lower value. However, if you restrict your search of extrema to a closed interval, this function will have extrema at the boundaries of the closed interval considered. The function \( f(x) = 1/x \) is not defined at zero and has no extreme values on every set which contains \( 0^- \) and \( 0^+ \) (sets like \( [-10, 0) \cup (0, 2] \)). The function \( f(x) = 1/(1 + |x|) \) is defined on \( \mathbb{R} \) but has no minimum value on \( \mathbb{R} \) since any number that is larger (in absolute value) has a smaller function value.

**Proposition 3.4.** If \( f \) is a continuous function on a closed interval \([a, b] \), then \( f \) possess a minimum and maximum value on \([a, b] \).
These minimum and maximum value could either be on the boundary of the interval \((a, b)\), or strictly inside the interval. Figure 14 shows examples of functions that possess or not extreme values either considered on a closed or open interval.

### 3.4 Finding extrema

**Definition 3.8 (Critical number).** A function \(f\) has a critical number, \(\bar{x}\), if \(f'(\bar{x}) = 0\).

Note that \(f'(x) = 0\) implies that the slope of the tangent is zero. Figure 15 shows some examples of critical numbers.

**Theorem 3.1 (Fermat’s Theorem).** If \(f\) has a local maximum or minimum at \(\bar{x}\) (on an open interval), and if \(f'(\bar{x})\) exists, then \(\bar{x}\) is a critical number of \(f\).

**Remark 3.2.** The fact that a number is critical does not always imply that it is a maximum or minimum value.
Proposition 3.5 (The Closed interval method). To find the global maximum and minimum values of a function (continuous and differentiable) \( f \) on a closed interval \([a,b]\):

1. Find the critical numbers of \( f \) in \((a,b)\).
2. Find the values of \( f \) at the endpoints \( a \) and \( b \).
3. The largest value of steps 1 and 2 is the global maximum, the smallest of these values is the global minimum.

We also have some tests on the derivative of a function that tells us whether a critical number is an extrema or not.

Proposition 3.6 (First derivative test). Assume that \( \bar{x} \) is a critical number of a (continuous and differentiable) function \( f \).

(i) If \( f' \) changes from positive to negative at \( \bar{x} \), then \( f \) has a local maximum at \( \bar{x} \).

(ii) If \( f' \) changes from negative to positive at \( \bar{x} \), then \( f \) has a local minimum at \( \bar{x} \).

(iii) If \( f' \) does not change sign at \( \bar{x} \), then \( f \) has no local maximum or minimum at \( \bar{x} \).

The following second derivative test implies the first derivative test.

Proposition 3.7 (Second derivative test). Assume \( f'' \) is continuous near \( \bar{x} \)

(i) If \( f'(\bar{x}) = 0 \) and \( f''(\bar{x}) > 0 \), then \( f \) has a local minimum at \( \bar{x} \).

(ii) If \( f'(\bar{x}) = 0 \) and \( f''(\bar{x}) < 0 \), then \( f \) has a local maximum at \( \bar{x} \).

Figure 16 illustrates the cases of the two derivative test.

![Figure 16: Illustration of first and second derivative test.](image)

We now give examples of maximum and minimum problems.

3.5 Some Maximum and minimum problems

Example 3.3. A farmer has 1.6 km of fencing and wants to fence off a rectangular field that borders a straight river. He needs no fence along the river. What are the dimensions of the field that has the largest area.

Solution. The first step to solve this problem (after reading and understanding it of course) is to introduce some notations. Figure 17 represents the chosen notation and their meaning.

Here, we choose to call \( A \) the area of the rectangle, \( x \) the length of the sides of the rectangle that are perpendicular to the river and \( y \) the length of the other side. We need now to express the function we want to optimize (maximize here). This function is the area:

\[
A = xy
\]
A depends on the two variables \( x \) and \( y \), but we have a relation between those two variables that is given by the value of the perimeter of the field:

\[ 2x + y = 1600 \text{ m} \iff y = 1600 - 2x \]

We can plug the expression of \( y \) with respect to \( x \) into \( A \) so that \( A \) only depends on the variable \( x \).

\[ A(x) = (1600 - 2x)x = 1600x - 2x^2, \quad 0 \leq x \leq 800 \]

We want to maximize \( A \) on the closed interval \([0, 800]\). We can use the closed interval method, and to do so, we need to find the critical numbers of \( A \). We have:

\[ A'(x) = 1600 - 4x \]

And the only critical number of \( A \) is

\[ \bar{x} = 400 \]

Now, the maximum area is the largest value of \( A \) among \( A(0) \) (= \( 0 \) m\(^2\)), \( A(\bar{x}) = 320000 \) m\(^2\), and \( A(800) \) (= \( 0 \) m\(^2\)). The maximum value of the field is then:

\[ A(800) = 0.32 \text{ km}^2 \]

**Example 3.4.** A cylindrical can is to be made to hold 1 L of liquid. Find the dimensions that will minimize the cost of the metal to manufacture the can.

**Solution.** We first draw a diagram where we put some notations for the radius \( r \) of the cylinder, and for its height \( h \) (Figure 18).

Let \( A \) be the total surface of the can, this is equal to the areas of the top and bottom circles plus the surface of a rectangle with \( 2\pi r \) and \( h \) as the length of its sides. We have:

\[ A = 2\pi r^2 + 2\pi rh \]

We need to take into account the fact that the can must have a volume of 1 L (= 1000 cm\(^3\)):

\[ \pi r^2 h = 1000 \iff h = \frac{1000}{\pi r^2} \]

We now plug the expression of \( h \) into \( A \) so that \( A \) only depends on the radius \( r \):

\[ A(r) = 2\pi r^2 + 2\pi r \left( \frac{1000}{\pi r^2} \right) = 2\pi r^2 + \frac{2000}{r} \]
We need to minimize $A$ for $r > 0$. We can not apply the closed interval method since the considered interval is open, but we can still find the critical numbers of $A$:

$$A'(r) = 4\pi r - \frac{2000}{r^2} = \frac{4(\pi r^3 - 500)}{r^2}$$

Then, $A'(r) = 0$ when $\pi r^3 = 500$, so the only critical number is $\bar{r} = \sqrt[3]{\frac{500}{\pi}}$. With the help of the first derivative test, we observe that $A'(r) < 0$ for $r < \bar{r}$ and $A'(r) > 0$ for $r > \bar{r}$. Thus, $A(\bar{r})$ is a global minimum of $A$. The corresponding $h$ value is:

$$\bar{h} = \frac{1000}{\pi \bar{r}^2} = \frac{1000}{\pi (\sqrt[3]{\frac{500}{\pi}})^2} = 2\sqrt[3]{\frac{500}{\pi}} = 2\bar{r}$$

\[ \square \]

**Example 3.5.** A conical drinking cup is made from a circular piece of paper of radius $R$ by cutting out a sector and joining the edges $CA$ and $CB$. Find the maximum capacity of such a cup.

**Solution.** Figure 19 shows the chosen notations.

![Figure 19](image)

**Figure 19:** On the left: the circular piece of paper, on the middle and right: the cup and its view from the top.

The volume of the cup is a third of the volume of the cylinder with the same radius and height

$$V = \frac{1}{3} \pi b^2 h$$

The variable we can play with to design the cup is the angle $\theta$. Therefore, we need to express $V$ with respect to $\theta$. To this end, we first observe that the perimeter of the circle on the top of the cup is the
perimeter of the remaining circular arc $AB$

$$2\pi b = \theta R \Rightarrow b = \frac{\theta R}{2\pi}$$

In addition, we have

$$R^2 = b^2 + h^2 \Rightarrow h = \sqrt{R^2 - b^2} = R\sqrt{1 - \frac{\theta^2}{4\pi^2}}$$

We can now express $V$ with respect to $\theta$

$$V(\theta) = \frac{\theta^2 R^3}{24\pi^2} \sqrt{4\pi^2 - \theta^2}, 0 < \theta < 2\pi$$

We can exclude $\theta = 0$ and $2\pi$ since the corresponding volumes are zero (but essentially because $\sqrt{}$ is not differentiable at $0$). We search for the critical numbers of $V$:

$$V'(\theta) = \frac{2\theta R^3 \sqrt{4\pi^2 - \theta^2}}{24\pi^2} - \frac{\theta^2 R^3}{24\pi^2} \frac{\theta}{\sqrt{4\pi^2 - \theta^2}} = \frac{\theta R^3(8\pi^2 - 3\theta^2)}{24\pi^2 \sqrt{4\pi^2 - \theta^2}}$$

Since we excluded $\theta = 0$, the only critical number is $\theta = \sqrt{8\pi^2/3}$. Thanks to the first derivative test (check yourself), we can say that $V(\theta)$ is the maximum volume of the drinking cup, and

$$V(\theta) = \frac{2\pi R^3}{9\sqrt{3}}$$

Example 3.6. Let $v_1$ be the velocity of light in air and $v_2$ the velocity of light in water. According to Fermat’s Principle, a ray of light that travels from point $A$ in the air to point $B$ in the water will follow the path $ACB$ that minimizes the time taken. Show that

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2}$$

where $\theta_1$ is the angle of incidence and $\theta_2$ is the angle of refraction. This equation is known as Snell’s Law.

Solution. Figure 20 shows the different information of the problem and some additional notations.
The given information is the position of \( A \) and \( B \), and the velocity of light \( v_1 \) and \( v_2 \). The unknowns for the problem are the position of the point \( C \) and incidentally the angles \( \theta_1 \) and \( \theta_2 \) and also \( a \) and \( b \).

We can list the relations between each of the variables

\[
\begin{align*}
d &= a + b \\
AC &= \sqrt{h_A^2 + a^2} \\
CB &= \sqrt{h_B^2 + b^2} = \sqrt{h_B^2 + (d-a)^2} \\
\sin \theta_1 &= \frac{a}{AC} = \frac{a}{\sqrt{h_A^2 + a^2}} \\
\sin \theta_2 &= \frac{b}{CB} = \frac{b}{(d-a)/\sqrt{h_B^2 + (d-a)^2}}
\end{align*}
\]

The ray of light minimizes the time taken to follow \( ACB \). This time \( t \) is

\[
t(a) = \frac{AC}{v_1} + \frac{CB}{v_2} = \frac{\sqrt{h_A^2 + a^2}}{v_1} + \frac{\sqrt{h_B^2 + (d-a)^2}}{v_2}
\]

So we need now to find the critical numbers of \( t \) with respect to \( a \)

\[
t'(a) = \frac{a}{v_1\sqrt{h_A^2 + a^2}} - \frac{d-a}{v_2\sqrt{h_B^2 + (d-a)^2}} = \frac{\sin \theta_1}{v_1} - \frac{\sin \theta_2}{v_2}
\]

And \( a \) is a critical number if and only if

\[
\frac{\sin \theta_1}{v_1} = \frac{\sin \theta_2}{v_2} \implies \frac{\sin \theta_1}{v_1} = \frac{\sin \theta_2}{v_2}
\]

Which is Snell’s law.

\[\square\]

### 4 Numerical aspect

This section is dedicated to the computation of derivatives and integrals with a computer. We will first quickly explain how numbers are represented in a computer.

#### 4.1 Representation of the numbers

When we are thinking of numbers, we represent them in the decimal system. That means that, for instance, the number 7301\(_{10}\) (the subscript indicates the base in which we are counting) is for us \( 7 \times 10^3 + 3 \times 10^2 + 0 \times 10^1 + 1 \times 10^0 \). Note that the Babylonians had a system with base 60, so for them 7301 = 2 \times 60^2 + 1 \times 60^1 + 41 \times 60^0 = (02)(01)(41)\(_{60}\), and other cultures had other bases, but usually 5 or 10. If you wonder why 5 or 10, look at your hands!

Unlike us, the computer does not represent numbers in the decimal system, it uses the binary system, as if it only has one “finger”, a bit, that can only represent 0 or 1. So for the computer, the number (in the decimal system) 135\(_{10}\) = 1 \times 2^7 + 0 \times 2^6 + 0 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 10000111\(_2\) in the binary system. With \( n \) bits, the binary system can represent \( 2^n \) different numbers (from 0 to \( 2^n - 1 \)).

But since the memory capability of a computer is limited, it’s set of integer numbers is finite (despite our set \( \mathbb{N} \)). Usually the computer uses integers that are coded with 32 bits, which means that the biggest integer it can represent is \( 2^{32} - 1 = 4.295 \times 10^9 \). Note that the computer actually represents signed integers, and uses one bit to indicate the sign of the integer.

To represent real numbers, the computer transforms them into the form \( \pm 1, XXXXXXXX \times 2^n \). When using a 32 bits coding, the IEEE norm states that:

- The first bit (the more to the left) gives the sign of the number.
The next 8 bits are used to code the power \( n \).

The last 23 bits are used to code the mantissa (the XXXX after the coma).

The important thing to understand is that the computer cannot represent every real number. For example, there exists a non-zero number \( \varepsilon \), called the machine epsilon and such that for a computer:

\[
1 + \varepsilon = 1
\]

So, when dealing with a numerical program on a computer, one must understand that they are only approximating (though, pretty accurately). For instance, when looking for the zero of a function \( f(x) \), one cannot hope to find \( \bar{x} \) such that \( f(\bar{x}) = 0 \) but more a \( \bar{x}_\eta \) such that \( |f(\bar{x}_\eta)| < \eta \).

Keeping this in mind, we now give an introduction to how to numerically differentiate a function.

### 4.2 Numerical differentiation

To compute an approximation of the derivative of a function \( f(t) \) we simply use the definition of the derivative:

\[
f'(t) = \lim_{h \to 0} \frac{f(t + h) - f(t)}{h}
\]

and instead of going to the limit, we choose a small \( h \). We have three main ways of approximating \( f'(t) \):

\[
f'(t) \approx \begin{cases} 
\frac{f(t)-f(t-h)}{h} & \text{(backward differences)} \\
\frac{f(t+h)-f(t)}{h} & \text{(forward differences)} \\
\frac{f(t+h)-f(t-h)}{2h} & \text{(center differences)}
\end{cases}
\]

These three schemes are called finite difference schemes and are represented in Figure 21.

![Finite difference schemes](image)

**Figure 21**: Finite difference schemes (the step size \( h \) is obviously too large).

When using finite difference, one might think that the smaller the stepsize \( h \), the more accurate the approximation will be. That would be true if the computer was able to exactly evaluate the expression we are giving to it. However this is not the case, and by taking a stepsize that is very small, we take the risk to differentiate the function \( f \) but also the evaluation errors (which is actually always the case but a very small \( h \) emphasize the error participation).

There exist other differentiation methods such as automatic differentiation which uses the fact that any function defined with a computer program is only an accumulation of elementary functions that are easily differentiable (no matter how many there are).
4.3 Numerical integration

To numerically integrate a function $f$, we recall that the result of the integration is an antiderivative of $f$. It means that this antiderivative $g$ is such that at each point, the representing curve $y = g(x)$ has a tangent whose slope is $f(x)$.

So, if we know the value of $g$ at a point $\bar{x}$, it is possible to approximate the $g(x)$ for $x$ close to $\bar{x}$. One way of doing so is to use the so called Euler integration scheme:

$$g(x) \approx g(\bar{x}) + (x - \bar{x})f(\bar{x})$$

This approximation is illustrated by Figure 22.

![Euler integration scheme](image)

Figure 22: Euler integration scheme

Note that there exist a lot of different integration schemes that are more or less accurate, depending of the required number of evaluations of $f$ and on the regularity of this function.

4.4 Finding zeros: The dichotomy

In optimization problems we usually need to compute critical number or more generally, the zero of a function. To do so, we use the Newton method which will be presented (in its most simple version) in the next section. Here, we present a basic method that does not even use the idea of the derivative. This method is the dichotomy.

Assume that you are looking for the zero of a continuous function $f : [a, b] \to \mathbb{R}$ and such that $f(a) \times f(b) < 0$. Then, thanks to the continuity of $f$, there exist at least one number $c$ for which $f(c) = 0$. Applying the following algorithm will give an approximation of one of this zero.

1. Choose a precision $\varepsilon > 0$.
2. Let $x = a$, $y = b$ and $z = (a + b)/2$.
3. Do while $|f(z)| > \varepsilon$
   a. If $f(x)f(z) < 0$ then let $y = z$ and go to step c.
   b. Else let $x = z$.
   c. Let $z = (x + y)/2$
4. The approximation of the zero is $z$.

Figure 23 shows an example of the dichotomy method.

The problem with this method is that is is not possible to directly extend it to the search of zeros of a non scalar function. Moreover, it is impossible to find the zeros of a function that has zero as a minimum (see Figure 24) since this function would not have numbers $a$ and $b$ such that $f(a) \times f(b) < 0$.  

23
The Newton method can deal with the previous drawbacks of the dichotomy method and is usually faster.

4.5 Finding zeros: Newton’s method

The Newton method iteratively approximates the function by its tangent line. Actually, for a differentiable function $f : D \subset \mathbb{R} \to \mathbb{R}$, a first order approximation states that

$$f(x + h) \approx f(x) + hf'(x)$$

(17)

So, if we start from a point $x_n$ and want to find a point $x_{n+1}$ such that the first order approximation (which is in fact the tangent of $y = f(x)$ at the point $x_n$), we need

$$f(x_n + h) = 0 = f(x_n) + hf'(x_n) \iff h = -\frac{f(x_n)}{f'(x_n)} \Rightarrow x_{n+1} = x_n + h$$

Of course, for this method to work, we need to have $f'(x_n) \neq 0$ otherwise we cannot compute the step $h$: in this case the method does not know in which direction is the zero.

Figure 25 illustrates the newton method.

5 Modeling and the different kinds of optimization problems

5.1 Modeling

Except for academic exercises, optimization problems do not come to us modeled in a straight mathematical form. Usually, they are related to some real world problem. For instance, controlled mechanical
problems (e.g. finding the optimal strategy to steer a mechanical system from one point to another in the minimum time or with the minimum cost), chemistry problems (e.g. which amount of chemical products do we need to accelerate a prescribed chemical reaction), economics problems, and so on.

The modeling step is a very important one since depending on how you convert your real world problem into a mathematical optimization problem, solving it might be more or less easy. Here is an easy example (still academic) that shows that the solving of a problem can be greatly eased by choosing an appropriate model.

Example 5.1. Find the area of the largest rectangle that can be inscribed in a semicircle of radius $r$.

Solution 1. Let’s take the semicircle to be the upper half of the circle $x^2 + y^2 = r^2$ with center at the origin. The word inscribed means that the rectangle has two vertices on the semicircle and two vertices on the $x$-axis as shown in Figure 26.

Let $(x, y)$ be the vertex that lies in the first quadrant. Then the rectangle has sides of length $2x$ and $y$, so its area is

$$A = 2xy$$

Since $(x, y)$ lies in the semicircle, we have:

$$x^2 + y^2 = r^2 \iff y = \sqrt{r^2 - x^2}$$

Plugging this relation into the area expression, we eliminate $y$ and we have:

$$A(x) = 2x\sqrt{r^2 - x^2}$$
The domain of this function is $0 \leq x \leq r$ (reflects the fact that $x$ lies in the semicircle). Its derivative is:

$$A'(x) = 2\sqrt{r^2 - x^2} - \frac{2x^2}{\sqrt{r^2 - x^2}}$$

since $x \geq 0$, the only critical number of $A$ is $\bar{x} = r/\sqrt{2}$. This value of $x$ gives a maximum value of $A$ since $A(0) = A(r) = 0$. Therefore, the area of the largest inscribed rectangle is

$$A(\bar{x}) = 2\frac{r}{\sqrt{2}}\sqrt{r^2 - \frac{r^2}{2}} = r^2$$

Another solution is:

**Solution 2.** If, instead of the Cartesian coordinate, we use the angle as a variable, the solution comes more easily. Let $\theta$ be the angle shown in Figure 27.

![Figure 27:](image)

The area of the rectangle is

$$A(\theta) = (2r \cos \theta)(r \sin \theta) = r^2(2 \sin \theta \cos \theta) = r^2 \sin 2\theta$$

We know that $\sin 2\theta$ has a maximum value of one when $2\theta = \pi/2$. So $A(\theta)$ has a maximum value of $r^2$ and it occurs when $\theta = \pi/4$.

Notice that the second solution does not involve derivatives, and is much easier than the first one. This illustrates how the choice of the model (here of reference frame: angular instead of Cartesian) can be of critical importance.

The choice of an appropriate reference frame is one of the numerous aspect of modeling, we will see another example in the AUV case, and in the orbital transfer case in the next section. However, the choice of a reference frame is not the only aspect of modeling. Indeed, especially for controlled mechanical systems, one should also decide how robust the model.

For instance, consider that you want to optimize the throwing of a ball; that is, throw it the furthest distance. Assume that the force with which you launch the ball is always the same, and that the variable you can control is the angle at which you launch the ball (you can also consider that the force depends on the angle). Figure 28 illustrates this problem.

To model this problem, we must first observe the different forces acting on the ball. Then using the Newton’s second law ($\sum \text{Force} = \text{mass} \times \text{acceleration}$), we will have a differential equation to determine the velocity of the ball at each time ($v(t) = \text{acceleration}$). The question here is: how accurate do we want our model to be? Indeed, the first force one will think to take into account is obviously the gravitational one. However, one might also think about considering the drag force due to the air resistance (how do we model this, with which precision?). But if one wants to be excessively picky, one may also want to consider the coriolis force (effect of the rotation of the Earth). Or the fact that the gravitational force depends on the altitude (who knows, maybe the ball will enter into orbit around the Earth!). Even the fact that due to the Earth oblateness, its gravitational field is not
exactly centered. Of course, if we consider that the throwing is done with near human strength, we can neglect almost every force except the central gravitational field (assuming it is constant) and maybe the drag force. Even with this simplified model, we will be confident that the result will be close to what happens in reality.

Once the modeling of an optimization problem is done, one can distinguish between different kinds of mathematical formulation. For instance, the previous example is a very simple form of parameter optimization problem, which general form is given in the next section.

5.2 Parameter optimization problem

The main characteristic of a parameter optimization problem is that the number of variables (parameters) you can modify, in order to optimize the cost, is finite. This number can be very large (hundreds, thousands ...) but it is finite. The general form of a parameter optimization problem is

\[
\min f(x) \quad \text{subject to} \quad h_i(x) = 0, \quad i = 1, \ldots, m \\
g_i(x) \leq 0, \quad i = 1, \ldots, r
\]

where \( x \in \mathbb{R}^n \) (\( n \in \mathbb{N}^* \)) are the parameters of the problem. The functions \( h_i \) are the equality constraints and the functions \( g_i \) are the inequality constraints. They define a set over which the function \( f \) (the cost or criterion) has to be minimized. The method to solve such problems consists in raising it to an higher dimension by introducing other variables associated with the constraints, called Lagrangian multipliers. With these additional variables, \( \lambda \), we can introduce a scalar valued function called the Lagrangian, which has to be minimized with respect to the parameters \( x \) and \( \lambda \) by using an extension of the search of extrema presented in the previous sections (the search for critical numbers and the second derivative test).

Remark 5.1. Note that a maximization problem can always be rewritten as a minimization one since

\[
\max f(x) \leftrightarrow \min -f(x)
\]

The solving of (POP) is very effective when all functions \( f, h_i \) and \( g_i \) are linear (or quadratic) with respect to \( x \) because it is possible to transform the problem into a linear system of equations.

For (POP) which are nonlinear (the trigonometric functions or polynomials of degree greater than 2) the solving is more tricky and needs to be done iteratively. The methods used are based on Newton’s method with additional refinements: we try to improve our guess step by step until the desired accuracy is achieved.

A more effective (POP) solver uses Sequential Quadratic Programming (SQP), which basically approximates the Lagrangian as a quadratic function at each step. After an iteration is done, the approximation is updated.

In the class of Parameter Optimization Problems, we can outline one specific and troublesome subclass, the Mixed Integer Optimization Problem (MIOOP).
5.3 Mixed Integer Optimization Problem

A (MIOP) is a (POP) in which some of the optimization parameters are restricted to be integer. So it has the following general form.

\[
\begin{align*}
\text{(MIOP)} \quad & \begin{cases} 
\min f(z, x) \\
\text{subject to} \\
z \in \mathbb{Z}^m \\
x \in \mathbb{R}^n \\
h_i(z, x) = 0, \ i = 1, \ldots, s \\
g_i(z, x) \leq 0, \ i = 1, \ldots, r
\end{cases}
\end{align*}
\]

(19)

If \( m = 0 \) then the (MIOP) is a (POP). Otherwise, the (MIOP) is an NP-complete problem, which means that it is combinatorial (you would have to explore all the combination of integers if you don’t find an appropriate method).

The general method to solve such problems is the Branch & Bound method. Note that there also exists a Branch & Cut, and a Branch & Price method whose principles are based on the same idea. The Branch & Bound method consists of dividing the (MIOP) into subproblems by, for instance, cutting the domain for the integer parameters in half. Then, the idea is to find a way to bound (even roughly) the minimum and maximum value of the criterion of each subproblem. If one of the subproblems has a lower bound that is higher than the upper bound of the other problem, then the search for the solution of the original (MIOP) can be restricted to the search of the solution of one of the two subproblems. The 'set' of integer in which we have to search is then reduced. By applying this idea recursively, one can drastically reduce the set of integer in which to search and ultimately find the solution.

Of course, this method needs a pretty clever way to choose the subproblems and to estimate lower and upper bounds. For the moment, the solving of (MIOP) is restricted either to very specific cases or to a few number of integer parameter (usually no more than a hundred).

Since the solving of (MIOP) problems generally does not depend on the same kind of methods used for optimization with real parameters, we do not go further in their presentation. We now present the optimal control problem (OCP).

5.4 Optimal Control Problem

Usually, when dealing with controlled mechanical systems, the parameter on which one can play, to optimize the cost function, is variable in time and therefore, should be controlled at each time. One can think about the AUV or the satellite where the control is the thrust at each time. The modeling of such a system usually gives birth to Optimal Control Problems. Their general form is

\[
\begin{align*}
\text{(OCP)} \quad & \begin{cases} 
\min f_0(t_f, x(t_f)) + \int_{t_0}^{t_f} l(t, x(t), u(t))dt \\
x(t) = f(t, x(t), u(t)) \text{ (Dynamics)} \\
u(t) \in \mathcal{U} \subset \mathbb{R}^m \\
h_0(t_0, x(t_0)) = 0 \text{ (Initial conditions)} \\
h_f(t_f, x(t_f)) = 0 \text{ (Final conditions)}
\end{cases}
\end{align*}
\]

(20)

Here, the parameter you can play with in order to minimize the cost is the function \( u: t \in [t_0, t_f] \rightarrow u(t) \), called the control. Contrary to the parameter optimization, the dimension of the control is not finite (there is an infinite number of times \( t \in [t_0, t_f] \)), so one cannot use the same solution methods as for the (POP). Furthermore, the control is usually restricted to a domain \( \mathcal{U} \) which reflects the fact that the propellers of the AUV have a finite power, so the thrust is bounded.

The criterion \( f_0(t_f, x(t_f)) + \int_{t_0}^{t_f} l(t, x(t), u(t))dt \) has a part \( f_0 \) that only depends on the final time \( t_f \) and final state \( x(t_f) \), called the Mayer part of the criterion, and an integral part \( \int_{t_0}^{t_f} l(t, x(t), u(t))dt \) called the Lagrange part of the criterion. The Mayer part is most often the final time, \( t_f \) but can also be a distance between the final state and a given point in case of the final state is not fixed. The Lagrange part reflects a criterion that has to be minimized over the whole time. One of these two parts can be zero, but not both since we need a criterion.
The terminal conditions $h_0$ and $h_f$, fix or relate between them the initial and final conditions. For instance, it can be the initial and final positions and velocities of a dynamical system. Note that there is no need to completely fix the initial and final states since one can prescribe a final position but not a final velocity to a dynamical system, or fix the final speed but not the direction of the velocity.

The dynamic, $\dot{x}(t) = f(t, x(t), u(t))$, of the system tells us how the states change with respect to the control and the other states. For a mechanical system, it can reflect the fact that the derivative of the position is the velocity and that the derivative of the velocity is the natural acceleration plus the acceleration due to external forces that can be controlled. In this case the dynamic is of the following form (with $x(t) = (r(t), v(t))$)

$$\begin{align*}
\dot{r}(t) &= v(t) \\
\dot{v}(t) &= a(t, r(t)) + g(u(t))
\end{align*}$$

As for the solving method, it consists once more in raising the problem into a higher dimension by introducing additional states, called costates or adjoint states. These costates verify a specific dynamic based on the derivatives of a function called the Hamiltonian, which depends on the states, the costates, the control, and the time. Thanks to this Hamiltonian, we can express a necessary condition for a control to be optimal. This necessary condition is that the control maximizes the Hamiltonian at each time. From this, we can define another function, called the shooting function, that depends on $n$ unknowns and that is zero if and only if the unknowns define a trajectory and a control that satisfies the necessary condition and the terminal conditions. The solving of the problem then involves the search for the zero of the shooting function, a bit like for the Lagrangian in $(POP)$. Another way to solve $(OCP)$ is to transform it into a $(POP)$. Indeed, using an integration scheme (e.g. the Euler one), one can discretize the dynamic and transform it into equality constraints. The discretization of $(OCP)$ can be as follows

1. Choose time steps $t_i, \ i = 0, \cdots , ns$ such that $t_0 < t_1 < t_2 < \cdots < t_{ns-1} < t_{ns} = t_f$ (e.g $t_i = t_0 + i(t_f - t_0)/ns$ the uniform discretization with time step $\delta t = 1/\text{ns}$).
2. Set $x_i = x(t_i)$ and $u_i = u(t_i)$.
3. Choose an integration scheme for the dynamics. For instance, the Euler scheme would give the equality constraints: $x_{i+1} = x_i + \delta t f(t_i, x_i, u_i), \ i = 0, \cdots , ns - 1$.
4. Transform the domain of the control into equality or inequality constraints (the domain $U$ is often already in the appropriate form).
5. If there is a Lagrange part in the criterion, use an integration scheme (not necessarily the same as for the dynamics) to write it as a function of $t_i, x_i$ and $u_i$.
6. The obtained problem is a $(POP)$ whose solution approximates the solution of the original $(OCP)$.

The transformation into a $(POP)$ usually gives less accurate results, but the actual solving is often much easier than with the Hamiltonian approach. Note that this transformation is only one of the many possibilities.

In the next section, we will presents some $(OCP)$ and their solutions.

The next paragraph gives an interesting example where, depending on the assumptions made on the control, the modeling gives either a $(POP)$ or a $(OCP)$.

5.5 An example of two close problem with different modeling

We considerer the transfer of a satellite from one orbit around the Earth to the geostationary orbit (see Figure 2). The criterion we are interested in is the maximization of the final mass (or equivalently the minimization of the consumption).
Let $r$ be the position ($\in \mathbb{R}^3$) of the satellite in the Earth-fixed cartesian frame (the satellite is viewed as a point with mass $m$) and $T$ the thrust (in $\text{N} = \text{kg.m.s}^{-2}$) delivered by the satellite engine(s). Then, the equation of motion of the satellite is given by:

$$\ddot{r} = -\frac{\mu_0}{|r|^3}r + \frac{T}{m}$$

(22)

where we only consider the central gravitational field. This is actually a very good approximation considering the involved altitudes (the farther we are from Earth, the less the gravitational perturbations will be).

Now, we introduce two different ways of providing the thrust $T$. Actually, the commonly used propulsor are chemical ones whose principle is to eject gas in order to provide thrust. Another way of providing the thrust is by an ion thruster whose principle is to eject ions instead of gas. Table 1 gives the main characteristics of common thrusters.

<table>
<thead>
<tr>
<th>Propulsion</th>
<th>Thrust (N)</th>
<th>Specific impulse (s)</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>propellant rocket</td>
<td>0.1 – 100</td>
<td>100 – 400 minutes</td>
<td></td>
</tr>
<tr>
<td>ion thruster</td>
<td>$10^{-7}$ – 1</td>
<td>1500 – 8000 month</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparison between two thruster technologies.

The main drawback of ion thrusters is that their thrust magnitude is weak compared to traditional propellers. Actually, when traditional propellers take a day to achieve a transfer, the ion thrusters will take a few months to do the same transfer. However, ion thrusters have an interesting property, which is the high value of their specific impulse ($I_{sp}$), which is related to the acceleration provided to the ejected ions (or gas for traditional thrusters). This is interesting because the evolution of the mass of the satellite (which is the consumption) is given by:

$$\dot{m} = \frac{|T|}{I_{sp}g_0}.$$  

(23)

So the consumption is proportional to the magnitude of the thrust, but also to the inverse of $I_{sp}$. This indicates that the greater the specific impulse, the more effective in terms of consumption the thruster is. The effectiveness of ion thrusters is interesting because of the economic cost of the launch of a satellite, which is directly related to its weight.

Furthermore, there is a great difference in the modeling of a transfer with traditional propellers and with ion thrusters. Indeed, for traditional propellers, the magnitude of the thrust is predominant with respect to the Earth gravitational field, and the duration of the thrust is short. So one can consider that a high thrust transfer consists into jumps from one orbit to another as illustrated in Figure 29.

Here, the consumption is a function of the two impulses $\Delta V_1$ and $\Delta V_2$ which change the velocity of the satellite (and then its orbit since knowing the velocity of the satellite at one point gives the orbit it is following). So the implied optimization problem is a (POP) with very few parameters and it is very easy to solve. Note that in this case, even if the impulse $\Delta V_1$ implies a thrust magnitude too high for the propellers, it is always possible to divide it into two or more impulses in the same direction and separated by one revolution.

Now, if we consider a low thrust transfer, we cannot assume that the satellite jumps from one orbit to another. Indeed, the magnitude of the thrust implies that we should consider the transfer as a continuous modification of the orbit of the satellite. Figure 30 shows an example of a low-thrust transfer.

With this low-thrust transfer, we must consider the continuous evolution of the position and velocity of the satellite given by the dynamic equation (22). If we do so, the parameter we can play with to find the optimal (in the sense of minimum consumption) way of steering the satellite from one orbit to another is the sequence of thrusts provided by the thruster. This is then a function of time and the
optimization problem is an (OCP) of the form:

\[
\min_{|T| \leq T_{\text{max}}} \int_{t_0}^{t_f} |T(t)| dt
\]

\[
\begin{aligned}
\dot{r}(t) &= v(t) \\
\dot{v}(t) &= -\frac{\mu}{r(t)^3} r(t) + \frac{T(t)}{m(t)} \\
\dot{m}(t) &= -\frac{\mu v(t)}{|T(t)|} l_{\text{cl}} \\
|T(t)| &\leq T_{\text{max}}, \forall t \in [t_0, t_f] \\
r(t_0) &= r_0 \\
v(t_0) &= v_0 \\
h_f(r(t_f), v(t_f)) &= 0
\end{aligned}
\]

Figure 29: High thrust orbital transfer with two impulses

Figure 30: Low thrust orbital transfer
Note that the final condition \( h_f(r(t_f), v(t_f)) = 0 \) means that the final position and velocity is not completely fixed, since we only specified the final orbit and do not impose the final position of the satellite on this orbit.

This (OCP) is much harder to solve than the high-thrust transfer, but the considered propulsion yields much better final mass. An example of low-thrust minimum fuel orbital transfer will be given in the section 6.

6 Some Optimal Control Problem

This section is dedicated to some examples of optimal control problems and their solutions. We will not detail the optimization procedure, but just give an outline.

6.1 The Brachystochrone

The brachystochrone problem was posed by Johann Bernoulli in Acta Eruditorum in June 1696. He introduced the problem as follows:

**Invitation to all mathematicians to solve a new problem.**

If in a vertical plane two points \( A \) and \( B \) are given, then it is required to specify the orbit \( AMB \) of the movable point \( M \), along which it, starting from \( A \), and under the influence of its own weight, arrives at \( B \) in the shortest possible time. So that those who are keen of such matters will be tempted to solve this problem, it is good to know that it is not, as it may seem, purely speculative and without practical use. Rather it even appears, and this may be hard to believe, that it is very useful also for other branches of science than mechanics. In order to avoid a hasty conclusion, it should be remarked that the straight line is certainly the line of shortest distance between \( A \) and \( B \), but it is not the one which is traveled in the shortest time. However, the curve \( AMB \) - which I shall divulge if by the end of this year nobody else has found it - is very well known among geometers.

This problem is historically very interesting since it is the one that launched the calculus of variations and more specifically the optimal control theory which derived from it. It is called the *Brachystochrone* from the Greek words brachisthos (\( \beta\rho\omicron\chi\sigma\sigma\omicron\omicron\omicron \)) : shortest, and chronos (\( \chi\rho\omicron\omicron\omicron\omicron \)) : time.

At that time, different mathematicians solved the problem (some after an extension of the deadline): Johann Bernoulli, Leibniz, Jakob Bernouilly (Johann’s elder brother), Tschirnhaus, l’Hospital and Newton. They did not all use the same method, but all found the correct answer. We will only present Johann’s method since it is quite elegant. But first, let us see (Figure 31) three different paths connecting the points \( A \) and \( B \), one of those being the solution.

![Figure 31: Three different path to connect points A and B.](image)

The solution of the brachystochrone problem is the third path, which is the cycloid. The cycloid corresponds to the trajectory defined by a point fixed on the circumference of a circle rolling on the horizontal axis.

Let us now write more precisely the brachystochrone problem. Choose a cartesian coordinate system with the \( y \) axis pointing downward. Denote \((0,0)\) and \((a, b)\) the respective coordinates of points \( A \) and \( B \). A path \( r(t) = (x(t), y(t)) \) defined on the time interval \([0,T]\) is a feasible trajectory if:
(i) \( r(0) = (0, 0), r(T) = (a, b) \) and \( r \) is continuous.

(ii) \( \frac{1}{2}(\dot{x}^2(t) + \dot{y}^2(t)) = gy(t) \) for all \( t \in [0, T] \).

Here \( g \) is the gravitational constant. Condition (i) states that the path \( r \) must connect \( A \) and \( B \) continuously. Condition (ii) is the conservation of energy at each time \( t \): the kinetic energy of the body \( (mv^2/2) \) must equal the decrease of potential energy \( (mgy) \) due to its loss of altitude. The brachystochrone problem is the one of finding the feasible path with the smallest final time \( t_f \).

In order to solve this problem, Johann Bernoulli used an analogy with a light ray by the mean of the Fermat’s principle and more precisely, Snell’s law (see the example of section 3). To do so, he divided the half-plane into horizontal strips \( S_k = \{(x, y) : y_k \leq y \leq y_{k+1}\} \) of height \( \delta \), for \( k = 0, 1, \ldots \), where \( y_k = k\delta \) and treating the speed \( c_k \) in the strip \( S_k \) as a constant \( (c_k = \sqrt{2gy_k}) \), then the light rays of this discretized problem should approach those of the original problem as \( \delta \) tends toward 0.

Clearly, the paths will be straight line segments within each strip \( S_k \), the question is then to find how those light rays bend at the frontiers of the strips. Using the Snell’s law, we have that:

\[
\frac{c_k}{\sin \theta_k} = \text{constant},
\]

where \( \theta_k \) is the refraction angle of the light ray in the strip \( S_k \). Following the Leibniz notation, we denote \( dx \) and \( dy \) as infinitesimal changes in \( x \) and \( y \) respectively. We now consider the problem when \( \delta \) tends toward 0, and then \( \theta \) is the refraction angle of the tangent to the path at point \((x, y)\). Then we have:

\[
\sin \theta = \frac{dx}{\sqrt{dx^2 + dy^2}}
\]

Combining equations (24) and (25) and the fact that \( c_k = \sqrt{2gy_k} \), we obtain:

\[
\frac{dx^2 + dy^2}{dx^2} = \frac{1}{Ky},
\]

for some constant \( K \). Then \( \frac{dx^2 + dy^2}{dx^2} = \frac{1}{Ky} \), and considering \( y \) as a function of \( x \) we have \( 1 + y'(x) = \frac{C}{y} \), where \( C = 1/K \). So the brachystochrone curve will satisfy the differential equation:

\[
y'(x) = \sqrt{\frac{C - y(x)}{y(x)}},
\]

with \( C \) a constant. The curves given by the parametric equations:

\[
\begin{align*}
x(\varphi) &= x_0 + \frac{C}{2}(\varphi - \sin \varphi) \\
y(\varphi) &= \frac{C}{2}(1 - \cos \varphi) \\
\varphi &\in [0, 2\pi]
\end{align*}
\]

satisfy (26). These equations describe a unique cycloid generated by a point \( P \) on a circle of diameter \( C \) that rolls on the horizontal axis, in such a way that \( P \) is at \((x_0, 0)\) when \( \varphi = 0 \).

This approach used by Bernoulli has some shortcomings. Indeed, first the equation (26) has spurious solutions because. The curve \( y(x) = \text{constant} \) is a solution, but clearly not a feasible path. Moreover, this solving assumed the possibility to write \( y \) as a function of \( x \) which reduces the domain of path. The use of modern control theory, allows one to overcome those problems, but we will not give them here since they are a bit too complicated.

Instead, let us see another example.

### 6.2 The AUV

In this subsection, we are interested in the minimum time transfer of an AUV from one position at rest to another position at rest. This is an optimal control problem as we will see after the explanation of the modeling.
The AUV is a material object, and contrary to the case of the satellite, we do not reduce it to a material point because some forces that act on the AUV depend on the shape of the AUV and on its angular velocity. These forces are neglectable in the void, not in water. To know the position of the AUV on an Earth fixed frame, we use 3 coordinates \((x, y, z)\) to indicate the position of the center of the AUV and 3 angles \((\phi, \theta, \psi)\) (known as the Euler angles) to give its orientation. We then have 6 coordinates for the position and orientation of the AUV, we let \(\eta = (x, y, z, \phi, \theta, \psi)\), we also need to know its velocity. There will be one velocity associated to each position and orientation coordinates, but instead of expressing them in the Earth-fixed frame, we prefer to express them in the body-fixed reference frame as shown in Figure 32.

The velocities corresponding to the positions are noted \((u, v, w)\) (in m/s), the ones corresponding to the orientations are noted \((p, q, r)\) (in rad/s) and the velocity of the AUV is denoted \(\nu = (u, v, w, p, q, r)\). Now, we must relate the velocity \(\nu\) to the position \(\eta\) knowing that they are not expressed in the same frame. Actually, to express \(\nu\) in the Earth-fixed frame, we just need to know the orientation of the body-fixed frame, and this orientation is given by the Euler angles \((\phi, \theta, \psi)\). The velocity \(\dot{\eta}\), expressed in the Earth-fixed frame, is then:

\[
\begin{align*}
\dot{x} &= u \cos \psi \cos \theta + v(- \sin \psi \cos \phi + \cos \psi \sin \theta \sin \phi) + w(\sin \psi \sin \phi + \cos \psi \cos \phi \sin \theta) \\
\dot{y} &= u \sin \psi \cos \theta + v(\cos \psi \cos \phi + \sin \phi \sin \theta \sin \phi) + w(- \cos \psi \sin \phi + \sin \psi \cos \phi \sin \theta) \\
\dot{z} &= -u \sin \theta + v \cos \theta \sin \phi + w \cos \theta \cos \phi \\
\dot{\phi} &= p + q \sin \phi \tan \theta + r \cos \phi \tan \theta \\
\dot{\theta} &= q \cos \phi - r \sin \phi \\
\dot{\psi} &= \frac{q \sin \phi \cos \theta}{\cos \phi} + \frac{r \sin \phi}{\cos \theta}
\end{align*}
\]

These equations can be summarized using the so called matricial notation:

\[
\dot{\nu} = J(\eta)\nu \tag{27}
\]

Now, we should express the dynamic of the velocity \(\nu\) (\(\dot{\nu} = ?\)). To do so, we need to know what external forces we should consider for this model. The first force we consider is the control we can apply to the AUV by means of its thrusters. We assume that the AUV is fully actuated, which means that every component of the velocity can be controlled and so the control \(\tau\) is in \(\mathbb{R}^6\) (6 components). Of course \(\tau\) is bounded:

\[
\tau \in U = \{\tau \in \mathbb{R}^6 | \alpha_i \leq \tau_i \leq \beta_i, i = 1, \cdots, 6\}
\]
The control $\tau$ is the force the user can play with; other forces have to be taken into account. The first one is the restoring forces which corresponds to the gravity and the buoyancy. If we denote $(x_G, y_G, z_G)$ the coordinate of the center of gravity and $(x_B, y_B, z_B)$ of the center of buoyancy, the restoring force vector $g$ is as follows:

$$g(\eta) = \begin{bmatrix} (W_g - B) \sin(\theta) \\
(-W_g + B) \cos(\theta) \sin(\phi) \\
(-W_g + B) \cos(\theta) \cos(\phi) \\
(-y_G W_g + y_B B) \cos(\theta) \cos(\phi) + (z_G W_g - z_B B) \cos(\theta) \sin(\phi) \\
(z_G W_g - z_B B) \sin(\theta) + (x_G W_g - x_B B) \cos(\theta) \cos(\phi) \\
(-x_G W_g + x_B B) \cos(\theta) \sin(\phi) - (y_G W_g - y_B B) \sin(\theta) \end{bmatrix}$$

where $W_g$ is the gravity coefficient ($\approx mg_0$ in Newtons) and $B$ is a buoyancy coefficient. Note that the restoring force depends on the orientation of the vehicle except for the orientation $\psi$ which does not intervene in the calculus of the righting arm produced by the inclination of the AUV. Moreover, one can see that this righting arm depends on the distance between the center of gravity and the center of buoyancy. If it is large then the AUV will be very stable and we need a great force to compensate any change of inclination. If the two centers are close, a small force will be enough to change the orientation, implying that the AUV is less stable.

Another force that would be wise to take into account is the force that water opposes on a moving object. This damping force is modeled with a force that quadratically depends on the magnitude of velocity in each direction, and is opposed to this velocity. So, we can model it as follows:

$$D(\nu) = \begin{bmatrix} -X_u u - X_{uu} u |u| \\
-Y_v v - Y_{vv} v |v| \\
-Z_w w - Z_{ww} w |w| \\
-K_{pp} p - K_{pp} p |p| \\
-M_{qq} q - M_{qq} q |q| \\
-N_{rr} r - N_{rr} r |r| \end{bmatrix}$$

Here the absolute value is used to take into account the direction of the velocity. The coefficients $X_u, X_{uu}, \cdots, N_{rr}$ depend on the AUV, and can be determined by experimentation or with the help of hydrodynamic mechanics (usually both are used in conjunction).

The last force we consider is the Coriolis force, $C(\nu)$ which comes from the eventual rotation of the AUV. Its expression being rather cumbersome, we do not give it, we simply outline the fact that a pure translation motion, in the sense that $\dot{\nu} = (\dot{u}, 0, 0, \cdots, 0)$ (pure Surge) or $(0, \dot{v}, 0, 0, \cdots, 0)$ (pure Sway) or $(0, 0, \dot{w}, 0, 0, 0)$ (pure Heave), has a zero Coriolis force. Moreover, the Coriolis force depends on the center of gravity, and on the principal and secondary inertia coefficients of the AUV.

Now, all these forces $\tau$, $g(\eta)$, $D(\nu)$ and $C(\nu)$ are not coupled with the acceleration of the AUV. Since we are considering an object that is not a point and which is not homogeneous (in terms of mass repartition), applying a force at a point of the AUV will creates inertia momentum. To represent this inertia effect, we introduce what we call the inertia matrix $M$ that will be multiplied to the previously mentioned forces (actually it is its inverse $M^{-1}$ which we multiply). We will not give the expression of this matrix, one just has to accept that the matrix-vector multiplication $M^{-1}(\tau - C(\nu) - D(\nu) - g(\eta))$ returns a vector which represents the acceleration of the AUV expressed in the body-fixed frame.

The equations of motion of the AUV are as follows:

$$\begin{cases} \ddot{\eta} &= J(\eta)\nu \\ \dot{\nu} &= M^{-1}(\tau - C(\nu) - D(\nu) - g(\eta)) \end{cases} \quad (28)$$

Now, the problem we want to solve is the one of steering the AUV from an initial configuration $(\eta_0, \nu_0)$ to a final configuration $(\eta_f, \nu_f)$ in the minimum time. To simplify, we will only consider terminal configurations at rest, i.e. $\nu_0 = \nu_f = (0, \cdots, 0)$. 

35
To give numerical results of this optimal control problem, we need to give values to the different hydrodynamic parameters. Here are parameters corresponding to the AUV ODIN.

\[
\begin{align*}
  m &= 125 \text{ kg}, & W_g &= 1125, & X_u = Y_v = Z_w = 0, & B = 2.5578\pi + W_g \\
  I_x &= 6.8 \text{ kg.m}, & x_G &= 0, & K_p = M_q = N_r = -130, & \tau_{\text{max}} = 20\text{N} \\
  I_y &= 7.2 \text{ kg.m}, & y_G &= 0, & X_{uu} = Y_{vu} = Z_{wv} = -148 \\
  I_z &= 9.1 \text{ kg.m}, & z_G &= 0.5 \text{ m}, & K_{pp} = M_{qq} = N_{rr} = -180 \\
  I_{xy} &= 0, & x_B &= 0, & X_{ud} = Y_{vd} = Z_{wd} = -m/2 \\
  I_{xz} &= 0, & y_B &= 0, & K_{pd} = M_{pd} = N_{rd} = 0 \\
  I_{yz} &= 0, & z_B &= 0
\end{align*}
\]

Where \((I_x, I_y, I_z)\) are the primary inertia coefficients and \((I_{xy}, I_{xz}, I_{yz})\) are the secondary ones. And \(\beta_i = -\alpha_i = \tau_{\text{max}}, \ i = 1, \cdots, 6.\)

Before exposing the minimum time results, we first want to outline the fact that since we are in the water, the damping force is very restrictive. To understand this, we consider a pure Surge motion, which means that we consider a motion for which \(\dot{v} = (\dot{u}, 0, \cdots, 0).\) To do so, we need to consider the equations of motion and compute the control \(\tau\) that will yield the desired acceleration. Moreover, we fix the control \(\tau_1\) to be maximum if we want to go forward, minimum if we want to go backward. After computation, we have:

\[
\begin{align*}
\dot{u} &= \frac{2}{3m}(\beta_1 - g_1 + (X_u + X_{uu}|u|)u) \\
\tau_2 &= g_2 \\
\tau_3 &= g_3 \\
\tau_4 &= g_4 \\
\tau_5 &= m\dot{u}z_G + g_5 \\
\tau_6 &= -m\dot{u}y_G + g_6
\end{align*}
\]

The equations for \(\tau_i\) first reflect that we want to keep the same inclination along the pure motion and that is why the control should compensate the restoring force \(g_i.\) Secondly, the pure surge acceleration also produce torques \(\dot{q}\) and \(\dot{r}\) and for which the control should also compensate.

At this point we should make a remark concerning the feasibility of such a pure motion. Indeed, for this pure motion to be feasible, we need all the controls \(\tau_i\) to stay in the admissible domain \(U.\) According to the equations of \(\tau_2, \cdots, 6\) it depends on the restoring force and on the created inertia forces (in \(\tau_{5,6}\)).

This leads us to another similar concerning the acceleration \(\dot{u}.\) As we are considering the forward motion, we need the maximum surge thrust \(\beta_1\) to be greater than the restoring force and the damping force. Moreover, the higher the surge velocity, the higher the damping force will be. So the acceleration \(\dot{u}\) will tend to decrease until it reaches zero at the maximum surge velocity. With the chosen numerical values, the \(\dot{u}\) equation depends on the angle \(\theta_i,\) and Figure 33 gives the evolution of the surge when starting at rest.

One can see that after 10 s, the surge velocity reaches its limit, which is not greater than 0.45 m/s (= 1.62 km/h).

In order to do a complete surge motion from one configuration to another, we need a full thrust phase in one direction (acceleration and maximum velocity) and another full thrust phase in the opposite direction (to decelerate and reach a the final configuration at rest). The principle of the deceleration phase is the same as the acceleration phase and to compute the full pure surge motion, one only needs to know the duration of both phases.

Figure 34 shows an example of a concatenation of pure motion transfer (Surge, Sway and then Heave) from the initial configuration \(\eta_0 = (0, \cdots, 0)\) to the final configuration \(\eta_f = (10, 8, 5, 0, 0, 0)\) (in meters).
For this trajectory, the pure surge acceleration and constant velocity phase take place from the initial time \((t_0 = 0)\) to \(t_1 \approx 28.5251\) s, then the surge deceleration is from \(t_1\) to \(t_2 \approx 31.1062\) s which is also the time \(t_{\text{surge}}\) needed for the surge translation. The sway acceleration (and constant velocity) take place from \(t_2\) to \(t_3 \approx 54.1540\) s, the deceleration from \(t_3\) to \(t_4 \approx 56.7351\) s and the time needed for the sway motion motion is \(t_{\text{sway}} = t_4 - t_2 \approx 25.6289\) s. Finally, the heave acceleration (and constant velocity) is from \(t_4\) to \(t_5 \approx 76.6068\) s, the deceleration from \(t_5\) to \(t_f \approx 78.2153\) s and \(t_{\text{heave}} \approx 21.4802\) s. As one could expect we have \(t_{\text{surge}} > t_{\text{sway}} > t_{\text{heave}}\) but of course those times are not proportional to the distances \(|\eta_f - \eta_0|\) (the AUV is not mechanically symmetric).

In the figure we see that the position \(x\) first evolves nearly linearly during the pure Surge phase,
then it is \( y \) which does the same during the pure Sway phase and finally \( z \) mimics its two predecessors during the pure Heave phase. The angles \( (\phi, \theta, \psi) \) and the angles velocities \( (p, q, r) \) are zero along the entire trajectory as expected.

As for the velocities \( (u, v, w) \), they clearly represent the full acceleration phases, then the constant velocity phases and finally the deceleration ones.

This trajectory is rather simplistic and one can easily understand that it is possible to complete it quicker. To compute the time optimal trajectory, we transform the optimal control problem into a parameter optimization problem (see section 5 for explanation). When applying the numerical method to find the time optimal trajectory, we find the trajectory shown in Figure 35. The corresponding minimum time is \( t_{\text{min}}^f \approx 27.6830 \) s.

![Figure 35: Time optimal trajectory for \( \eta_0 = (0, \cdots, 0) \) and \( \eta_f = (10, 8, 5, 0, 0, 0) \).](image)

On this trajectory, we can see some similarities with the pure translation one. Indeed, the evolution of the positions \((x, y, z)\) are also nearly linear except that they occur simultaneously instead of one after the other. If we look at the velocities \((u, v, w)\) we can also see that they have three phases each, an acceleration, a constant velocity and a deceleration.

A great difference with the pure translations trajectory is that the orientation of the AUV is not constant, but varies along the trajectory. This is the case because in order to be time optimal, we always use the full power of the thrusters as shown in Figure 36.

We see that during the transfer, the different components of the control are maximum (or minimum). The first three controls do not evolve a lot since the corresponding velocities \((u, v, w)\) are quite regular. On the other hand, the controls \( \tau_4, \tau_5 \) switch a lot at the beginning and end of the strategy, which corresponds to the variation of the angles \((\phi, \theta)\) in the Figure 35. Finally, the control \( \tau_6 \) has a very important number of switchings which most likely means that the chosen orientation \( \psi \) is very unstable. Indeed, one can see in Figure 35 that the angular velocity \( r \) rapidly switches sign.

### 6.3 Low-thrust orbital transfer

Here we consider the satellite problem introduced in section 5. Since solving this problem involved numerical computation, we first need one remark. Providing that the considered thrust is weak with respect to the Earth gravitational field, we can expect the transfer to have a lot of revolutions around
the Earth. And, since the solving will necessarily involve the integration of the differential equation (22) we should be able to find a more appropriate reference frame than the cartesian one. Indeed, the cartesian coordinates of the satellite will greatly evolve during each revolution, which could impair the precision of the integration.

Usually, one prefers to use reference frames which are numerically more stable. Such a frame could be the Gauss coordinate system. The idea is to describe the geometry of the osculating orbit of the satellite (the orbit the satellite would follow if just moved by gravity).

Figure 37 shows the different components used in Gauss coordinates.

The Gauss coordinates of the osculating orbit (OsOr) are:
- a: semi-major axis of the OsOr.
- e: eccentricity of the OsOr.
- i: inclination of the OsOr with respect to the equatorial plan.
- Ω: ascending node longitude.
- ω: argument of the perigee (gives the position of the perigee).
- ν: true anomaly (position of the satellite on its OsOr with respect to the perigee).

The numerical stability of these coordinates comes from the fact that when only subjected to gravitation, the only coordinate that evolves is ν, the others remaining unchanged. Unfortunately, (Ω, ω, ν) are not uniquely defined for the geostationary orbit. To eliminate this last singularity, we introduce the modified set of Gauss coordinates (P, e_x, e_y, h_x, h_y, L) as follows:

- P = a(1 - e^2): parameter of the OsOr.
- (e_x, e_y) = (e \cos(Ω + ω), e \sin(Ω + ω)): the eccentricity vector.
- (h_x, h_y) = (\tan(i/2) \cos(Ω), \tan(i/2) \sin(Ω)): the slope vector.
- L = Ω + ω + ν: the cumulative longitude.

Note that here the cumulative longitude L stands for the position of the satellite in its OsOr, but also for the number of revolutions the satellite has already done.

As for the control, ie the thrust, we express it in the body-fixed reference frame attached to the satellite. Figure 38 shows the considered ortho-radial frame.

![Ortho-radial frame](image)

Figure 38: Ortho-radial frame.

This (q, s, w) frame is defined by:

\[
\begin{align*}
q &= r/|r| \\
s &= w \times q \\
w &= (r \times v)/|r \times v|
\end{align*}
\]

Here the operator \(\times\) stands for the vector product and its result is a vector that is orthogonal to the one involved in the binary operation. So q is directed towards the position of the satellite (the ’radius’ of the OsOr), w is orthogonal to the orbital plan and s is more or less oriented towards the velocity vector of the satellite.
In these frames, the dynamic of the satellite is of the form:

\[
\begin{align*}
\dot{x} &= f_0(x) + \frac{1}{m} \sum_{i=1}^{3} T_i f_i(x) \\
\dot{m} &= -\frac{\dot{\gamma}}{I_{sp} g_0}
\end{align*}
\]

where \(x = (P, e_x, e_y, h_x, h_y, L)\) and \(f_0(x)\) is zero except for the component corresponding to \(L\).

The initial orbit we consider is defined by the parameters \(P^0 = 11.625\) Mm, \((e^0_x, e^0_y) = (0.75, 0)\), \((h^0_x, h^0_y) = (0.0612, 0)\) \((\Rightarrow i = 7^\circ)\) and \(L_0 = \pi\) \((\Rightarrow\) apogee\)). The final orbit is defined by \(P^f = 42.165\) Mm, \((e^f_x, e^f_y) = (0, 0)\) (circular orbit), \((h^f_x, h^f_y) = (0, 0)\) (equatorial orbit) and \(L^f\) is free or fixed depending on whether or not we want to fix the final position of the satellite on its orbit.

As for the final time \(t^f\), we set it as a multiple of the minimum time needed to achieve the prescribed transfer. This multiple \(c_{tf}\) is of course taken strictly greater than 1 since if it is less than 1, there is no admissible trajectory. If it is equal to 1 then the minimum fuel trajectory is the same as the minimum time one.

To solve this problem, as for the one with the AUV, we have two main options. The first one being to solve this as a real optimal control problem, the second being the transformation of the optimal control problem into a parameter optimization problem. Contrary to the AUV case, we choose the first method for the following reasons:

- This method is more accurate than the transformation into a (POP).
- It is faster in terms of computation time.
- It is more fun (the (POP) method needs less subtlety to work).

The main drawback of this method is that it may be quite hard to make it work, mainly because its initialization is extremely sensitive. This is a common drawback of shooting methods, and it is amplified in our case. Let us try to explain this sensitivity of initialization.

Considering the problem from a physical point of view, one can easily guess the general form of an optimal thrust strategy. The minimization of the transfer time implies that we use the thrust to its maximum possibility along all the trajectory. For the maximization of the final mass, the strategy will not be the same as we have some spare time \((c_{tf} > 1)\) to save fuel. This saving will be realized by arcs of maximum thrust were its use is the more effective and arcs of zero thrust when thrust is less efficient. As we will see in the numerical results, the thrust is the more effective when the satellite is near its apogee because this is the place of least kinetic energy.

Still concerning the final time, it is interesting to note that the minimum transfer time, \(t^f_{\text{min}}\) numerically verified the following law:

\[t^f_{\text{min}} \ast T_{\text{max}} \approx 850\ \text{h.N} \tag{35}\]

This means that the minimum transfer time is inversely proportional to the maximum thrust (of course it is only true if we consider the same terminal configurations). This legitimate our choice of \(c_{tf}\) rather than directly \(t^f\) because it will allow us to effectively compare solutions for different \(T_{\text{max}}\). The longitude and the time being close, we also have the following relation:

\[(L^f_{\text{min}} - L^0) \ast T_{\text{max}} \approx 267\text{rad.N} \tag{36}\]

where \((L^f_{\text{min}} - L^0)\) corresponds to the minimum number \((\in \mathbb{R}^+)\) of revolutions needed to achieve the prescribed transfer.

The main difficulty of the minimum fuel thrust strategy is its discontinuity. Because of this, the function of which we are trying to find a zero (the shooting function) loses smoothness. This loss of smoothness takes the following form: between two thrust strategies with a different number of discontinuities, there is a \textit{frontier} where the shooting function is not differentiable. Since we use Newton like methods to find a zero of the shooting function, we cannot (or at least our success rate
will decrease) find a zero if we start from the wrong thrust strategy. This basically means than for the moment, in order to find the optimal thrust strategy, we need to know quite accurately what it looks like \textit{a priori}.

To overcome this problem, the idea is to consider a similar optimal control problem, namely the minimization of the energy. This criterion is the following:

\[ \min_{|T| \leq T_{\max}} \int_{t_0}^{t_f} |T(t)|^2 \, dt \]

The minimum energy thrust strategies are continuous. A zero of the corresponding shooting function is less sensitive to the initialization. We then introduce the homotopic criteria

\[ J_\lambda = \min_{|T| \leq T_{\max}} \int_{t_0}^{t_f} \left( 1 - \lambda \right) |T(t)|^2 + \lambda |T(t)| \, dt \]

For \( \lambda = 0 \), \( J_0 \) is the minimization of energy, and for \( \lambda = 1 \), \( J_1 \) is the minimization of the consumption. The idea is then to connect the solution of the optimal control with the criterion \( J_0 \), to the one with criterion \( J_1 \). We legitimate this approach with Figure 39, which shows the evolution of the optimal control with respect to time for various values of \( \lambda \).

\[ \text{Figure 39: Optimal control w.r.t time for various } \lambda \text{ and two different homotopic criteria} \]

One can see that the solution of minimization of energy already shows the premises of the structure of the minimum fuel thrust strategy. One can also imagine that the evolution of the thrust strategy with respect to \( \lambda \) is quite smooth which helps if we want to obtain the minimum fuel thrust strategy from the minimum energy strategy.

Now, concerning the trajectory, Figure 40 gives the 3D optimal trajectory for a thrust of 0.1 N. Here, the final time is \( t_f = 1.5 \times t_{\min} \approx 12750 \) h \( \approx 17.7 \) months and there is approximatively 750 revolutions around the Earth. Concerning the repartition of the thrust arcs, one can easily see that the
most numerous ones are centered around apogees and that some (of the last ones) are centered around perigees.

As for the thrust strategy, Figure 41 shows an optimal thrust strategy for $T_{\text{max}} = 10 \text{ N}$ and $c_{\text{tf}} = 1.75$ ($\Rightarrow t^f \approx 127.5 \text{ h}$).

Figure 41: Optimal thrust strategy for $T_{\text{max}} = 10 \text{ N}$ and $c_{\text{tf}} = 1.75$. From the top to the bottom: $q$, $s$, $w$, $|T|$ (all scaled in $[-1, 1]$) w.r.t time.

Note that thrust strategies for very low-thrust have the same structure except that their graphs
are not very exploitable. The dotted line corresponds to the scaled longitude and when it continuously crosses the zero axis, it is approximatively an apogee and when the crossing is discontinuous, it is approximatively a perigee.

First of all, we can see that nearly all the thrust arcs are centered on apogees, as in the 3D trajectory. We also have that the last thrust arcs are centered on perigees. As for the direction of the thrust, the $s$ (transverse or ortho-radial) component is dominant. Moreover, the radial component $q$ of the thrust has a nearly zero average and the $w$ component is weak because the correction of inclination is small. It is interesting to note that the low-thrust strategy mimics the behavior of the high-thrust strategy.

The trajectory corresponding to the preceding thrust strategy is given in Figure 42.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{optimal_trajectory}
\caption{Optimal trajectory for $T_{\text{max}} = 10$ N and $c_{t_f} = 1.75$.}
\end{figure}

This figure shows that the evolution of the parameter of OsOr is increasing except at the end of the trajectory where it decreased a little. This decrease is due to the thrusts around perigees. As for the eccentricity and the inclination, they decrease regularly.

Furthermore, we can see that the evolution of the longitude $L$ is more and more linear. This reflects the fact that the orbit of the satellite is becoming circular as we approach to the final geostationary orbit.

Now, it could be interesting to see the evolution of the final mass with respect to the final time. Figure 43 shows this evolution with respect to $c_{t_f}$ for various $T_{\text{max}}$.

The first interesting thing that is uncovered in this figure is that the evolution of the final mass with respect to $c_{t_f}$ is numerically independent of $T_{\text{max}}$. An application of this could be to fix $c_{t_f}$ according to a final mass we want to obtain, no matter what the maximum thrust is. Note also that this independency is a result of the first numerical independency of the product $T_{\text{max}} \ast t_{\text{min}}$.

In this figure, the horizontal line corresponds to a high-thrust (impulse) transfer with the same specific impulse (which is not possible in practice). This line seems to represent the limit value of the final mass with infinite allowed transfer time ($c_{t_f} \rightarrow \infty$). This make sense because intuitively, a
low-thrust transfer where the final time is not restricted will only put a thrust where its effectiveness is maximum, so the resulting transfer will just have a lot of small impulse thrusts.

We can exhibit another empirical law which states that the evolution of the final longitude with respect to the final time is linear and does not depend on $T_{\text{max}}$.

References


