Mechatronic systems simulation (146029) - MOD1

A Practical Approach

Enrico Bertolazzi

Copyright © 2025 Enrico Bertolazzi

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 4.0 License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at https://creativecommons.org/licenses/by-nc-sa/4.0. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

First printing, May 2025



Contents

. 9

I Laplace Transform

1 Laplace Transform definition 15

1.1 Transform Properties 15

1.1.1 Scale Change and Translation Properties 17

1.1.2 Scaling of the Time Axis 17

1.1.3 Translation in the s and t axes 19

1.2 Existence of the Laplace Transform 20

1.2.1 Exponential Order Functions 21

2 Usage of the Laplace Transform 23

3 Inversion of the Laplace Transform 27

3.1 Partial Fraction Expansion 27

3.2 Inverting the Transform 27

3.2.1 Real roots 28

3.2.2 Complex roots 31

3.3 Table for the Laplace transform 38

4 Boundary value problem 41

II KKT

5 Calculus Review 49

5.1 Higher Regularity Conditions 49

5.2 Local Minimum 50

5.3 Conditions for a Local Minimum . . . 50
5.3.1 First-Order Conditions . . . 50
5.3.2 Second-Order Conditions . . . 50

6 Lagrange Multipliers . . . 53
6.1 Summary of constrained minimization with proofs . . . 56

7 Sylvester theorem . . . 63
7.1 Why use sylvester theorem . . . 64
7.1.1 Trick for semi positive matrices . . . 66

8 Inequality constraints . . . 69
8.0.1 Karush-Kuhn-Tucker conditions . . . 71

9 Summary of KKT with proofs . . . 81
9.1 Constrained minima, NLP problem . . . 85
9.2 Constrained minima and KKT . . . 86
9.3 Other constraints qualifications . . . 89

10 Worked examples . . . 91
10.1 Example 1: Detailed Example of Constrained Minimization . . . 91
10.1.1 Parametric Solution . . . 91
10.1.2 Lagrange Multipliers Method . . . 93
10.1.3 Conclusion . . . 94
10.2 Example 2: Detailed example of KKT . . . 94
10.2.1 Study of solution 1 . . . 97
10.2.2 Study of solution 2 . . . 98
10.2.3 Study of solution 3 . . . 99
10.3 Example 3 . . . 99
10.3.1 Lagrangian Function . . . 100
10.3.2 First-Order Conditions . . . 100
10.3.3 Solutions of First-Order Conditions . . . 101
10.3.4 Analysis of First Candidate . . . 101
10.3.5 Analysis of Second Candidate . . . 102
10.4 Example 4 . . . 102
10.4.1 Lagrangian Function . . . 104
10.4.2 First-Order Conditions . . . 104
10.4.3 Solutions of First-Order Conditions . . . 104
10.4.4 Analysis of First Candidate . . . 104
10.4.5 Analysis of Second Candidate . . . 105
10.5 Example 5 . . . 106
10.5.1 Lagrangian function . . . 106
10.5.2 First-Order Conditions . . . 107
10.5.3 Solution First-Order Conditions . . . 107
10.5.4 Analisys of The Candidate . . . 107

III Minimization of a Functional

11 Introduction . . . 113
11.1 Analogies with linear algebra . . . 116
11.2 First Variation . . . 118
11.2.1 Equivalent Forms . . . 118
11.3 Integrals and Limits . . . 119
11.3.1 Commuting Limit and Integral . . . 119
11.3.2 A Counterexample . . . 119

12 Fundamental lemma of the calculus of variations . . . 123

13 Euler Lagrange equation . . . 129
13.1 Pendulum example . . . 129
13.2 General formulation . . . 131
13.3 Application to the brachistochrone problem . . . 133

14 Boundary condition for a functional . . . 137
14.1 Free boundary conditions . . . 139

15 Functional minimization with generic constraints . . . 143
15.1 Heuristic deduction of general boundary condition . . . 145

16 Minimization with Integral Constraints . . . 151
16.1 Approximate solution using finite differences . . . 153
16.1.1 Discretization Setup . . . 153
16.1.2 Discretized Problem Formulation . . . 154
16.1.3 Lagrange function formulation . . . 154
16.1.4 First-Order conditions . . . 154
16.2 Heuristic deduction of integral constraint for functional . . . 155
16.2.1 A simple example . . . 155
16.2.2 First Variation of $\mathscr{L}$ . . . 156
16.2.3 General Solution . . . 156
16.2.4 Extension to Multiple Constraints . . . 157

17 Minimization with algebraic (path) constraints . . . 159
17.1 Heuristic deduction of path constraints . . . 162
17.1.1 Problem Statement . . . 162
17.1.2 Formulating the augmented functional . . . 162
17.1.3 Deriving the necessary conditions via first variation . . . 163
17.1.4 Applying the fundamental Lemma of calculus of variations . . . 163

18 Summary . . . 165
18.1 Spaces . . . 165
18.1.1 Variation of problem 2. . . . 166
18.1.2 Variation of problem 3. . . . 166
18.1.3 Variation of problem 4. . . . 167

18.1.4 Problem 5 . . . 167
18.2 General Problem . . . 167

19 Worked exercizes . . . 169
19.1 Exercize 1 . . . 169
19.1.1 Applying the First Variation . . . 169
19.1.2 Integration by Parts . . . 169
19.1.3 Applying the Fundamental Lemma of Calculus of Variations . . . 170
19.1.4 Boundary Conditions . . . 170
19.1.5 Final Problem . . . 170
19.1.6 Applying the First Variation . . . 170
19.1.7 Integration by Parts . . . 171
19.1.8 Applying the Fundamental Lemma of Calculus of Variations . . . 171
19.1.9 Final Problem . . . 171
19.2 Exercize 3 . . . 172
19.2.1 Applying the First Variation . . . 172
19.2.2 Integration by Parts . . . 172
19.2.3 Applying the Fundamental Lemma of Calculus of Variations . . . 173
19.2.4 Final Problem . . . 173

IV Optimal Control
General formulation from Euler-Lagrange . . . 179

20 Optimal Control problems . . . 183
20.1 Pontryagin maximum (minimum) principle . . . 186
20.1.1 Examples with Pontryagin maximum (minimum) principle . . . 186
20.2 Problems with linear controls and singular arc . . . 191
20.2.1 Singular controls . . . 192
20.2.2 Test for singular controls . . . 193

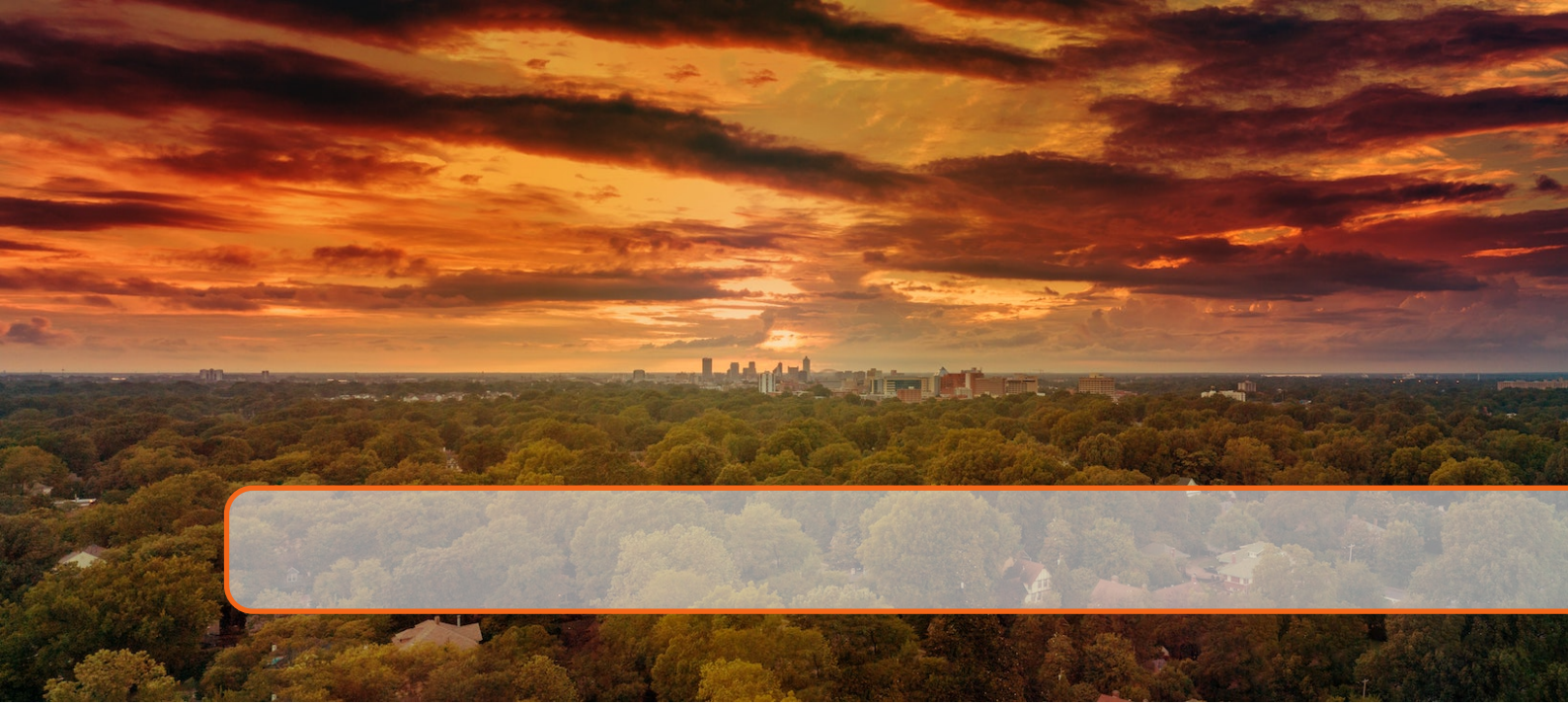
21 OCP Transformation . . . 195
21.1 Lagrange to Mayer Transformation . . . 197
21.2 Mayer to Lagrange Transformation . . . 199
21.3 Final Transformed Problem (Lagrange Form) . . . 200
21.4 Constant Parameters to Constant Functions . . . 201
21.5 Integral Constraints . . . 203
21.6 Free-Time . . . 205
21.7 Free-Time 2 . . . 206

22 Examples of optimal control problems . . . 207
22.1 The Brachistochrone Problem . . . 207
22.1.1 Problem Statement . . . 207
22.1.2 Path Length and Travel Time . . . 207
22.1.3 Constraints . . . 208
22.1.4 Solution Using Calculus of Variations . . . 208
22.1.5 The brachistochrone as an optimal control problem . . . 209

22.1.6 A Numerical Example . . . 209
22.2 The DIDO Problem . . . 212
22.2.1 The Divergence Theorem . . . 212
22.3 An Optimal Control Problem with constraints . . . 218
22.4 Dubins Car . . . 218
22.5 Four Singular Optimal Control Problems . . . 223
22.5.1 Singular 1 . . . 224
22.5.2 Singular 2 . . . 227
22.5.3 Singular 3 . . . 229
22.5.4 Singular 4 . . . 231
22.6 Free time problems . . . 232
22.6.1 Example 1 . . . 232
22.6.2 Example 2 (exam's simulation) . . . 234
22.6.3 Example 3 (free time problem, 2 controls) . . . 235

23 Summary . . . 239
23.1 The BVP . . . 241

References . . . 243



Ordinary differential equations are a powerful and essential tool in both theoretical and applied mathematics. Their ability to model dynamic systems, whether simple or complex, makes them indispensable in understanding the behavior of many real-world systems.

With advanced techniques and numerical methods, ODEs continue to play a crucial role in scientific research, engineering design, and beyond. Understanding ODEs is therefore central to many fields of study, from physics and engineering to biology and economics.

Acknowledgement



Figure 1: License: CC BY 4.0

This project is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0). Some pages may contain materials that are subject to copyright, in which case you will see the copyright notice. make

Laplace Transform

1 Laplace Transform definition 15
1.1 Transform Properties 15
1.2 Existence of the Laplace Transform 20

2 Usage of the Laplace Transform 23

3 Inversion of the Laplace Transform 27
3.1 Partial Fraction Expansion 27
3.2 Inverting the Transform 27
3.3 Table for the Laplace transform 38

4 Boundary value problem 41

The **Laplace Transform**, denoted by $\mathscr{L}$, is a powerful mathematical tool that converts a time-dependent function $f(t)$ into a function $\hat{f}(s)$ of a complex variable s.
This change of domain — from the **time domain** to the **complex domain** — is not just a formal manipulation: it provides a way to simplify the analysis and solution of problems that would otherwise be difficult to handle directly in the time domain.

Formally, the transform is defined as:

$$f(t) \;\mapsto\; \hat{f}(s) = \mathscr{L}\{f(t)\}(s)$$

Not every function can be transformed. The existence of the Laplace Transform depends on the **convergence** of the defining integral:

$$\hat{f}(s) = \int_{0^-}^{\infty} f(t)\, e^{-st}\, \mathrm{d}t = \lim_{\varepsilon \to 0^+} \lim_{M \to \infty} \int_{-\varepsilon}^{M} f(t)\, e^{-st}\, \mathrm{d}t$$

This introduces the notion of the **region of convergence (ROC)**: the set of complex values of s for which the integral is well-defined. The ROC is crucial, since it determines both whether the Laplace Transform exists and where it is valid.

Why it is powerful

The strength of the Laplace Transform lies in its ability to convert **differential operations** (difficult to manage) into **algebraic operations** (much easier to solve).

> **Note**
>
> A helpful analogy is with logarithms:
> - multiplying two numbers directly may be complicated,
> - but taking logarithms turns multiplication into addition:
>
> $$a \cdot b \;\mapsto\; \log a + \log b$$
>
> Similarly, the Laplace Transform translates differentiation and integration into multiplication and division by s, greatly simplifying calculations.

Applications in dynamic system

In engineering and physics, the behavior of mechanical, electrical, and other systems is often described by **linear differential equations**. Traditional analytical methods, such as the *method of variation of parameters*, may be cumbersome.

The Laplace Transform offers a systematic approach:

1. **Transformation** – the differential equation is converted into an algebraic equation in the s-domain.
2. **Analysis** – the algebraic equation is studied to extract system properties such as stability and frequency response.
3. **Inverse transformation** – the solution is brought back to the time domain via the **Inverse Laplace Transform**, yielding the explicit form $f(t)$.

This procedure provides both a theoretical framework and a practical tool, making the Laplace Transform indispensable in the study and design of complex systems.

Learning objectives

After this part, you should be able to state when a Laplace transform exists, use its main operational rules, and solve a linear initial- or boundary-value problem by transforming it to the s-domain and inverting the result.



1. Laplace Transform definition

Visualization of the Laplace Transform

To visualize the effect of the Laplace Transform, consider the following example with an exponential function $f(t) = e^{-at}$. The Laplace Transform of this function is well-known:

$$f(t) = te^{-at} \quad \mapsto \quad \hat{f}(s) = \frac{1}{(s+a)^2}$$

Here's a plot to illustrate the function and its Laplace Transform.

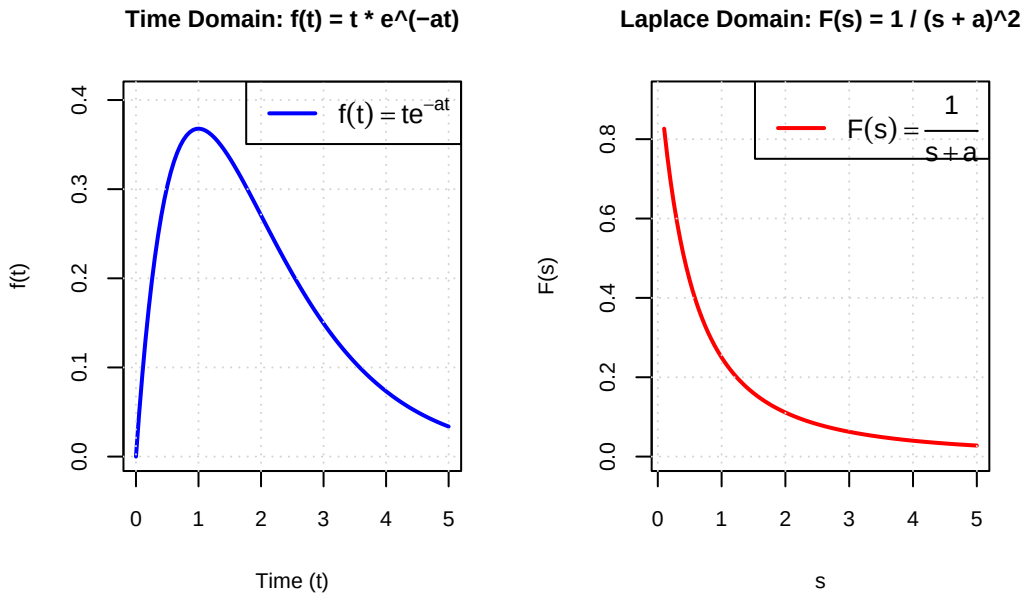


Figure 1.1: The left plot shows the function $f(t) = e^{-at}$ in the time domain, and the right plot shows the corresponding Laplace Transform $\hat{f}(s) = \frac{1}{s+a}$.

1.1 Transform Properties

The Laplace Transform possesses several important properties that facilitate calculations. One of the key properties to remember is that it acts as a **linear operator**:

Theorem 1.1 — linearity. The Laplace Transform exhibits a fundamental property known as linearity. For any functions

$$f(t) \mapsto \hat{f}(s) \quad \text{and} \quad g(t) \mapsto \hat{g}(s),$$

and for any real constants $a, b \in \mathbb{R}$, the linearity of the Laplace Transform can be expressed as follows:

$$h(t) := a f(t) + b g(t) \quad \mapsto \quad \hat{h}(s) := a \hat{f}(s) + b \hat{g}(s).$$

This means that the Laplace Transform of a linear combination of functions is equal to the same linear combination of their respective transforms.

Proof. The linearity property of the Laplace Transform can be demonstrated by applying the transform to a linear combination of two functions:

$$\mathscr{L}\{a f(t) + b g(t)\}(s) = \int_{0^-}^{\infty} (a f(t) + b g(t)) e^{-st} \, \mathrm{d}t.$$

Using the linearity of the integral, we can separate this into two distinct integrals:

$$\begin{aligned} &= a \int_{0^-}^{\infty} f(t) e^{-st} \, \mathrm{d}t + b \int_{0^-}^{\infty} g(t) e^{-st} \, \mathrm{d}t \\ &= a \hat{f}(s) + b \hat{g}(s). \end{aligned}$$

■

This confirms that the Laplace Transform of a linear combination of functions is equal to the same linear combination of their respective transforms.

However, it is important to note that not all functions allow for this property to hold. For specific cases where the integral may not converge, refer to Example Example 1.1 for further clarification.

■ **Example 1.1 — Limits of intergral sequence.** Consider a piecewise function $f_n(t)$, parameterized by n, defined for $t \geq 0$ as follows:

$$f_n(t) = \begin{cases} nt & 0 \leq t \leq \frac{1}{n} \\ 2 - nt & \frac{1}{n} < t \leq \frac{2}{n} \\ 0 & \text{otherwise} \end{cases}$$

As $n \to \infty$, this function converges to $f(t)$, which is zero everywhere:

$$f(t) = \lim_{n \to \infty} f_n(t).$$

To confirm this, we can compute the integral of $f_n(t)$ over the interval $[0, \infty)$:

$$F_n = \int_0^{\infty} f_n(t) \, \mathrm{d}t = \frac{1}{n} \quad \xrightarrow{n \to \infty} \quad 0.$$

Thus, as $n \to \infty$, the area under the curve tends to zero, consistent with the function becoming zero everywhere.

Misleading Intuition

At first glance, one might assume that the limit and the integral can be interchanged:

$$F = \int_0^\infty \lim_{n\to\infty} f_n(t)\,\mathrm{d}t = \lim_{n\to\infty}\int_0^\infty f_n(t)\,\mathrm{d}t.$$

However, this is not always true. To illustrate this, consider another piecewise function $g_n(t)$, defined as:

$$g_n(t) = \begin{cases} n^2 t & 0 \le t < \frac{1}{n} \\ 2n - n^2 t & \frac{1}{n} < t < \frac{2}{n} \\ 0 & \text{otherwise} \end{cases}$$

Taking the limit as $n \to \infty$, we find that the function $g(t) = \lim_{n\to\infty} g_n(t) = 0$ everywhere. However, the computed area differs:

$$\lim_{n\to\infty}\int_0^\infty g_n(t)\,\mathrm{d}t = 1 \quad \neq \quad \int_0^\infty g(t)\,\mathrm{d}t = 0.$$

This demonstrates that, in general, the limit of the integral and the integral of the limit are not equivalent:

$$\lim_{n\to\infty}\int g_n(t)\mathrm{d}t \neq \int \lim_{n\to\infty} g_n(t)\mathrm{d}t.$$

Visualization

To further illustrate these concepts, we can visualize the behavior of $f_n(t)$ and $g_n(t)$ for various values of n. As n increases, observe how each function behaves and how their areas change.

■

1.1.1 Scale Change and Translation Properties

In the study of Laplace transforms, the **scaling** and **translation** properties of the time axis play a significant role. These properties are essential for manipulating functions, allowing us to simplify their transformations and gain deeper insights into their behavior in the frequency domain.

1.1.2 Scaling of the Time Axis

Theorem 1.2 — Stretching/Expanding the Time Axis. When we stretch or expand the time axis of a function $f(t)$ by a factor $a > 0$, the corresponding Laplace Transform $\hat{f}(s)$ is modified as follows:

$$f(at) \mapsto \frac{1}{a}\hat{f}\left(\frac{s}{a}\right).$$

Proof. The proof directly follows from the definition of the Laplace transform by applying a change of variables. Let $at = z$, which implies $t = \frac{z}{a}$:

$$\begin{aligned} \mathscr{L}\{f(at)\}(s) &= \int_{0^-}^\infty f(at)\,\mathrm{d}t \\ &= \int_{0^-}^\infty f(z)\,\mathrm{e}^{-s\frac{z}{a}}\,\frac{\mathrm{d}z}{a} \\ &= \frac{1}{a}\hat{f}\left(\frac{s}{a}\right) \end{aligned}$$

■

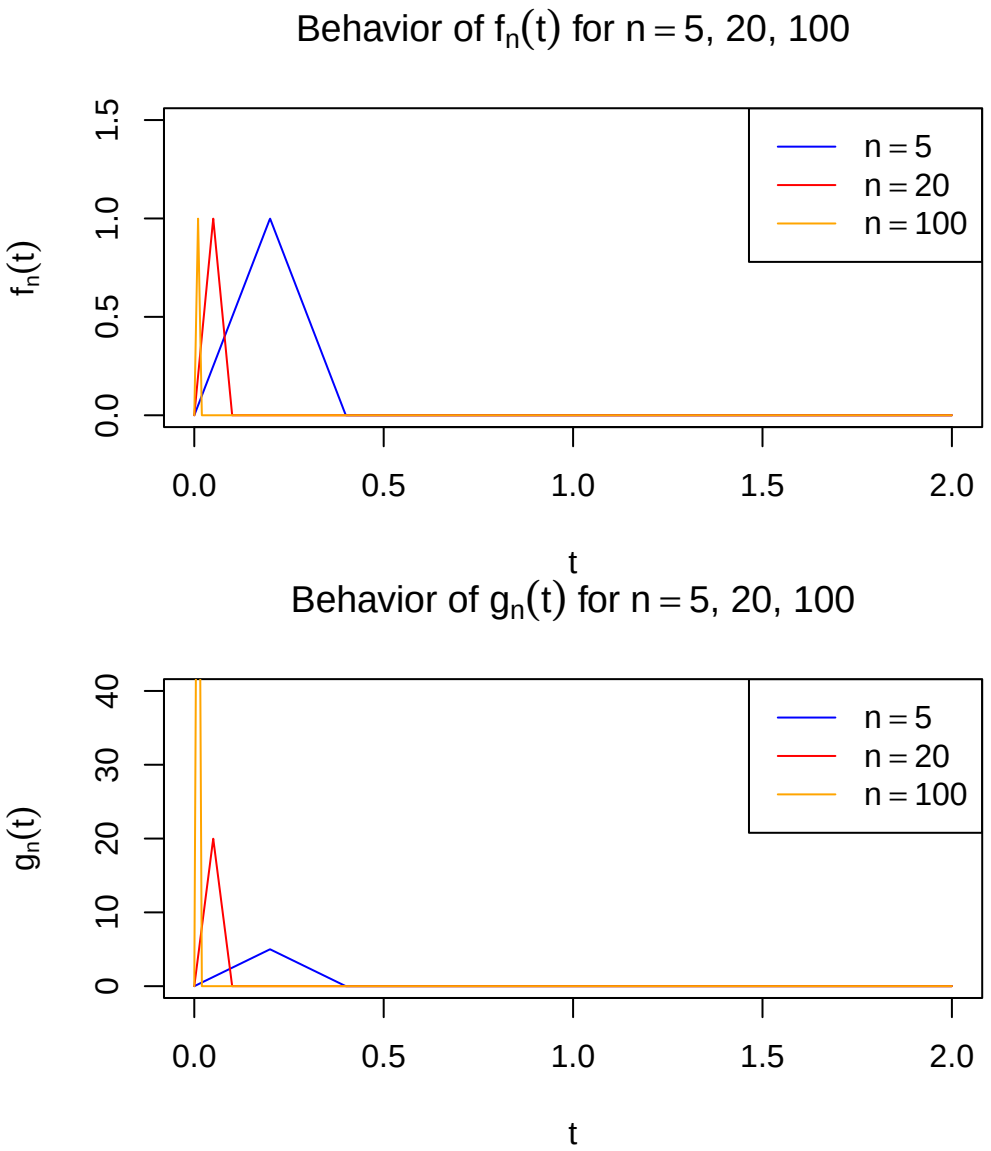


Figure 1.2: In these plots, we observe the behavior of both $f_n(t)$ and $g_n(t)$ as n increases. While both functions appear to converge to zero for large n, their areas behave differently. This distinction highlights the important concept that integrating limits and taking limits of integrals are not always interchangeable.

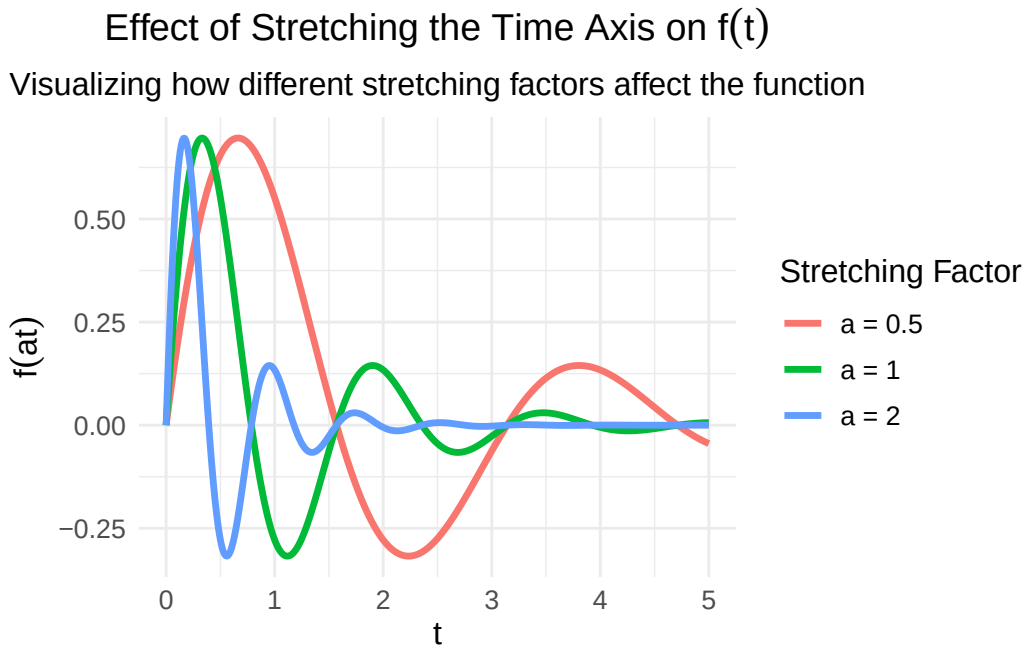


Figure 1.3: Visualization of the effect of stretching the time axis of $f(t)$ by a factor a, leading to the corresponding change in the Laplace transform.

1.1.3 Translation in the s and t axes

The **translation** property is a crucial aspect of Laplace transforms, affecting both the s axis and the t axis by a coefficient $a > 0$. These translations lead to the following relationships:

Theorem 1.3 — Translation in the s axes.

$$e^{at} f(t) \mapsto \hat{f}(s-a)$$

Proof. To prove the translation property in the s complex variable, we start with the Laplace Transform of the function $e^{at} f(t)$:

$$\begin{aligned} \mathscr{L}\{e^{at} f(t)\}(s) &= \int_{0^-}^{\infty} e^{at} f(t) e^{-st}\, dt \\ &= \int_{0^-}^{\infty} f(t) e^{(a-s)t}\, dt \\ &= \hat{f}(s-a). \end{aligned}$$

This shows that multiplying by e^{at} translates the Laplace Transform along the s-axis. ■

Theorem 1.4 — Translation in the t axes.

$$f(t-a) \mapsto e^{-as} \hat{f}(s)$$

Proof. For translation along the t-axis, we apply a change of variables with $z = t - a$:

$$\begin{aligned}\mathscr{L}\{f(t-a)\}(s) &= \int_{0^-}^{\infty} f(t-a)\,e^{-st}\,\mathrm{d}t \\ &= \int_{-a}^{\infty} f(z)e^{-s(z+a)}\,dz \\ &= e^{-as}\int_{0^-}^{\infty} f(z)e^{-sz}\,dz \\ &= e^{-as}\hat{f}(s).\end{aligned}$$

In this derivation, the first integral is effectively canceled because we typically assume that $f(t) = 0$ for $t < 0$. Therefore, this integral evaluates to zero. ■

Summary

These properties illustrate how translations in time and frequency domains impact the behavior of functions under Laplace transformations. The translation in the s-axis shifts the transform horizontally, while translation in the t-axis introduces an exponential decay factor.

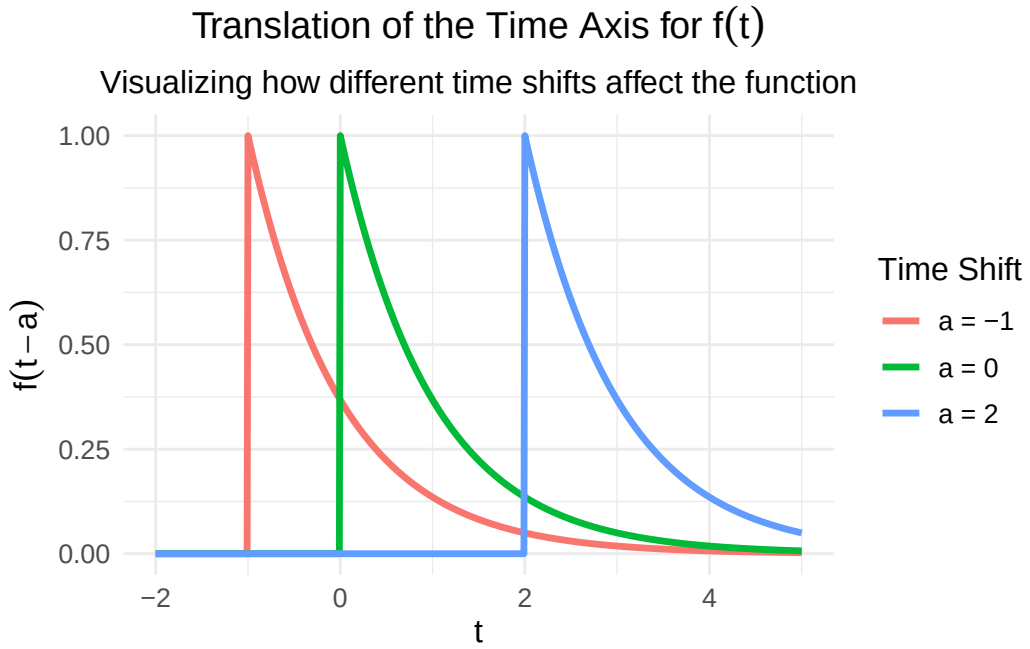


Figure 1.4: Visual representation of the effect of translating the function $f(t)$ in both the s and t axes.

1.2 Existence of the Laplace Transform

Not every function can be transformed using the Laplace operator $\mathscr{L}$. For instance, consider the function $f(t) = \mathrm{e}^{t^2}$ cannot be transformd. In fadct, if we apply the Laplace transform to this function, we obtain the following integral:

$$\mathscr{L}\{\mathrm{e}^{t^2}\}(s) = \int_{0^-}^{\infty} \mathrm{e}^{t^2-st}\,\mathrm{d}t = \int_{0^-}^{T} \mathrm{e}^{(t-s)t}\,\mathrm{d}t + \int_{T}^{\infty} \mathrm{e}^{(t-s)t}\,\mathrm{d}t.$$

If we choose a constant $T > \Re(s)$ (the real part of s), we can observe that the second integral $\int_T^{\infty} \mathrm{e}^{(t-s)t}\,\mathrm{d}t$ **does not converge** for any value $s \in \mathbb{C}$.

Proof. To prove this result, we start with the function $f(t) = \mathrm{e}^{t^2}$. If its Laplace transform exists (denoted by $\hat{f}(s)$), it must equal:

$$\hat{f}(s) = \lim_{M\to\infty} \int_{0^-}^{M} \mathrm{e}^{t(t-s)}\,\mathrm{d}t.$$

Since s is a complex variable, we express it as $s = a + \mathrm{i}b$, where $a, b \in \mathbb{R}$ and $\mathrm{i} = \sqrt{-1}$. Substituting this into the Laplace transform expression gives:

$$\mathrm{e}^{t(t-a-\mathrm{i}b)} = \mathrm{e}^{t(t-a)}\mathrm{e}^{-\mathrm{i}bt}.$$

Using the De-Moivre formula [1] we rewrite the exponential with the imaginary term:

$$\mathrm{e}^{-\mathrm{i}bt} = \cos(bt) - \mathrm{i}\sin(bt).$$

Thus, the integral becomes:

$$\lim_{M\to\infty} \int_{0^-}^{M} \mathrm{e}^{t(t-a)} \left(\cos(bt) - \mathrm{i}\sin(bt)\right) \mathrm{d}t.$$

This expression can be split into real and imaginary parts:

- **Real part**:

$$R = \lim_{M\to\infty} \int_{0^-}^{M} \mathrm{e}^{t(t-a)} \cos(bt)\,\mathrm{d}t.$$

- **Imaginary part**:

$$I = \lim_{M\to\infty} \int_{0^-}^{M} \mathrm{e}^{t(t-a)} \sin(bt)\,\mathrm{d}t.$$

Focusing on the real part R, we note that it involves the product of a cosine function and an exponentially growing function. By analyzing this product at large values of t, we can see that the integral does not converge. To formalize this observation:

$$\int_{0}^{M+\Delta M} \mathrm{e}^{t(t-a)} \cos(bt)\,\mathrm{d}t \geq \int_{0}^{M} \mathrm{e}^{t(t-a)} \cos(bt)\,\mathrm{d}t + \mathrm{e}^{M(M-a)} \int_{M}^{M+\Delta M} \cos(bt)\,\mathrm{d}t.$$

Since the exponential term $\mathrm{e}^{t(t-a)}$ grows without bound as t increases, it follows that the integral does not converge. Consequently, we conclude that the Laplace transform for $f(t) = \mathrm{e}^{t^2}$ **does not exist**. ■

1.2.1 Exponential Order Functions

A function $f(t)$ is said to be of **exponential order** if there exist constants $M, N \in \mathbb{R}$ such that:

$$|f(t)| \leq Me^{Nt}$$

for all $t \geq T$, where T is a finite constant.

This definition implies that the growth of the function $f(t)$ is bounded by an exponential function as t approaches infinity. Specifically, the term Me^{Nt} serves as an upper limit for the absolute value of $f(t)$.

If a function is of exponential order, its Laplace transform exists in at least part of the complex plane. This means that we can find values of s for which the Laplace transform converges, providing valuable insights into the behavior of the function in the frequency domain.

[1] $\mathrm{e}^{A+\mathrm{i}B} = \mathrm{e}^{A}\mathrm{e}^{iB} = \mathrm{e}^{A}(\cos B + \mathrm{i}\sin B)$.

Proof. To demonstrate that functions of exponential order always have a Laplace transform, we focus on the residual part of the integral from T to infinity:

$$\int_T^\infty |f(t)|\mathrm{e}^{-st}\,\mathrm{d}t.$$

Using the property $|f(t)| \leq M\mathrm{e}^{Nt}$, we can write:

$$\int_T^\infty |f(t)|\mathrm{e}^{-st}\,\mathrm{d}t \leq \int_T^\infty M\mathrm{e}^{Nt}\mathrm{e}^{-st}\,\mathrm{d}t = M\int_T^\infty \mathrm{e}^{(N-a)t}\,\mathrm{d}t,$$

where $s = a + ib$ and we focus on the real part $a = \Re(s)$.

For convergence of the integral, we require $a > N$. Thus, if $a > N$, the term $\mathrm{e}^{(N-a)t}$ tends to zero as $t \to \infty$, ensuring that the integral has a finite value.

Consequently, the domain of s for which the Laplace transform of an exponential order function exists is at least:

$$\Re(s) > N.$$

■

This result confirms that functions of exponential order have well-defined Laplace transforms within specific regions of the complex plane.

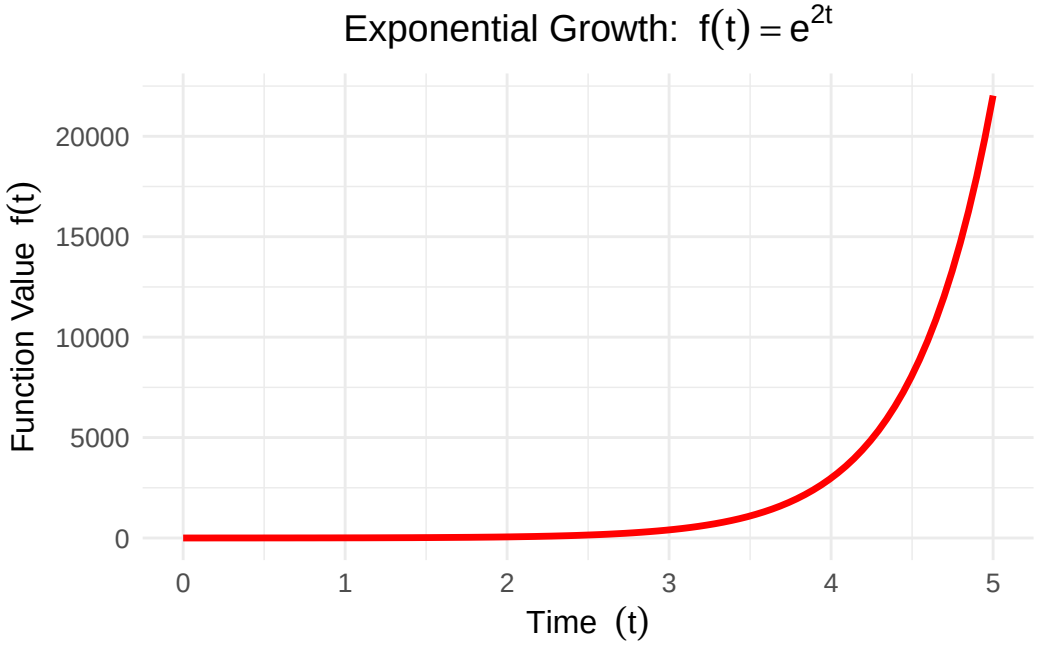


Figure 1.5: Exponential Growth of e^{Nt} with $N = 2$

According to the definition, we can also observe that the **domain** of the variable s for the Laplace transform of a function of exponential order is **at least** given by:

$$\forall s = a + \mathrm{i}b \in \mathbb{C}, \quad \text{such that} \quad a = \mathrm{Re}(s) \geq N.$$

This means that the real part of s must be greater than or equal to N for the Laplace transform to exist. In other words, the Laplace transform converges in the half-plane where the real part of s meets this condition.



2. Usage of the Laplace Transform

Consider the problem of solving the following differential equation:

$$\begin{cases} y'(t) = a\,y(t) + t \\ y(0) = 1 \end{cases}$$

Solving this equation in the time domain can be challenging. This is where the Laplace transform comes into play, allowing us to analyze the problem in the complex variable s. By applying the Laplace transform to the differential equation, we obtain:

$$\begin{aligned} \mathscr{L}\{y'(t)\}(s) &= \mathscr{L}\{a\,y(t) + t\}(s) \\ &= a\mathscr{L}\{y(t)\}(s) + \mathscr{L}\{t\}(s). \end{aligned}$$

At this point, we need to highlight an important **Laplace transform property**. This property states that the transform of the first derivative $f'(t)$ of a function $f(t)$ is given by:

$$f'(t) \mapsto s\hat{f}(s) - f(0^+),$$

where $f(0^+)$ represents the value of the function $f(t)$ at the origin of the time axis and corresponds to the initial (Cauchy) condition of the ordinary differential equation.

This property allows us to convert **differential equations** in the time domain into **algebraic equations** in the complex variable domain. Returning to our initial problem, we can rewrite the differential equation as:

$$s\hat{y}(s) - \underbrace{1}_{y(0^+)} = a\hat{y}(s) + \underbrace{\frac{1}{s^2}}_{\mathscr{L}\{t\}(s)}.$$

This simplifies to:

$$\hat{y}(s) = \frac{1 + \dfrac{1}{s^2}}{s - a}.$$

■ **Example 2.1 — System of Ordinary Differential Equations.** Consider the following system of two ordinary differential equations, subject to their initial conditions:

$$\begin{cases} x'(t) = y(t) + 1 \\ y'(t) = x(t) + \cos t \\ x(0) = 1 \\ y(0) = 1 \end{cases}$$

To solve this system, we can apply the Laplace transform to convert the equations into the complex variable domain. The Laplace transforms yield the following equations:

$$\begin{aligned} (i) \quad & \underbrace{s\tilde{x}(s) - x(0)}_{\mathscr{L}\{x'(t)\}(s)} = \tilde{y}(s) + \frac{1}{s} \\ (ii) \quad & \underbrace{s\tilde{y}(s) - y(0)}_{\mathscr{L}\{y'(t)\}(s)} = \tilde{x}(s) + \frac{1}{1+s^2} \end{aligned}$$

Knowing the initial conditions of the system allows us to express the system of functions $\tilde{x}$ and $\tilde{y}$ using matrix notation:

$$\underbrace{\begin{bmatrix} s & -1 \\ -1 & s \end{bmatrix}}_{A(s)} \begin{pmatrix} \tilde{x}(s) \\ \tilde{y}(s) \end{pmatrix} = \begin{pmatrix} 1 + \frac{1}{s} \\ \frac{1}{1+s^2} \end{pmatrix}$$

To find explicit values for $\tilde{x}(s)$ and $\tilde{y}(s)$, we can invert the matrix $A(s)$:

$$\tilde{x}(s) = \frac{s + 1 - \frac{1}{1+s}}{s^2 - 1}, \qquad \tilde{y}(s) = \frac{\frac{s}{1+s} - 1 - \frac{1}{s}}{s^2 - 1}.$$

■

■ **Example 2.2 — Computation of the Laplace Transform.** Consider the ordinary differential equation:

$$\begin{cases} x'(t) + x(t) = \cos t \\ x(0) = 2 \end{cases}$$

To find the solution in the Laplace domain, we apply the Laplace transform operator $\mathscr{L}$ to the entire equation. We utilize known transforms, specifically:

$$\begin{aligned} \mathscr{L}\{x(t)\}(s) &= \hat{x}(s), \\ \mathscr{L}\{x'(t)\}(s) &= s\hat{x}(s) - x(0), \\ \mathscr{L}\{\cos(t)\}(s) &= \frac{s}{s^2+1}. \end{aligned}$$

Applying the Laplace transform to the differential equation yields:

$$\mathscr{L}\{x'(t) + x(t)\}(s) = \mathscr{L}\{\cos t\}(s),$$

which translates to:

$$s\hat{x}(s) - 2 + \hat{x}(s) = \frac{s}{s^2+1}.$$

Now, we can combine terms to solve for $\hat{x}(s)$:

$$(s+1)\hat{x}(s) - 2 = \frac{s}{s^2+1}.$$

Rearranging gives us:

$$(s+1)\hat{x}(s) = \frac{s}{s^2+1} + 2.$$

Finally, solving for $\hat{x}(s)$ results in:

$$\hat{x}(s) = \frac{s}{(s+1)(s^2+1)} + \frac{2}{s+1} = \frac{2s^2+s+2}{(s+1)(s^2+1)}.$$

This expression represents the Laplace transform of the solution to the original ordinary differential equation.

■



3. Inversion of the Laplace Transform

3.1 Partial Fraction Expansion

The **inversion** of a Laplace transform is a crucial step that allows us to *transport* the algebraic solution found in the s domain back into a time-domain description. This process is represented by the operator $\mathscr{L}^{-1}$.

Rational Polynomial Representation

When dealing with (systems of) ordinary differential equations, the result of the Laplace transform (such as $\hat{x}(s)$) is typically expressed as a rational function in the form:

$$\frac{P(s)}{Q(s)},$$

where $P(s)$ and $Q(s)$ are polynomials with real coefficients in the variable s.

An important observation is that, in many practical applications, the degree of the numerator polynomial ∂P is less than that of the denominator polynomial ∂Q. If this condition does not hold naturally, we can apply polynomial long division. Specifically, if $P(s)$ can be expressed as:

$$P(s) = T(s)Q(s) + R(s),$$

where $T(s)$ is the quotient and $R(s)$ is the remainder, we can rewrite the rational function as:

$$\frac{P(s)}{Q(s)} = T(s) + \frac{R(s)}{Q(s)}.$$

The term represented by the polynomial $T(s)$, which has real coefficients, corresponds to the Laplace transform of Dirac delta functions. For example:

$$\mathscr{L}\{\delta(t)\}(s) = \int_{0^-}^{\infty} \delta(t)e^{-st}\,\mathrm{d}t = e^{-st}\Big|_{t=0} = 1,$$
$$\mathscr{L}\{\delta'(t)\}(s) = -\int_{0^-}^{\infty} \delta'(t)e^{-st}\,\mathrm{d}t = -\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{-st}\right)\Big|_{t=0} = s.$$

3.2 Inverting the Transform

Assuming we have a transform solution expressed as a simplified rational polynomial

$$\frac{P(s)}{Q(s)} \quad \text{such that} \quad \partial P < \partial Q$$

we can compute its inverse to express it in the time domain. This is commonly done using **partial fraction expansion**, which allows us to rewrite the solution as a sum of simpler elements that can be easily inverted using a Laplace transform table.

To properly perform partial fraction expansion, it is essential to rewrite the denominator of the polynomial ratio in factored form:

$$\frac{P(s)}{Q(s)} = \frac{b_0 + b_1 s + b_2 s^2 + \cdots + b_m s^m}{(s-p_1)^{m_1}(s-p_2)^{m_2}\cdots(s-p_n)^{m_n}},$$

where p_i are the roots of the denominator and each root has a multiplicity m_i.

Depending on the multiplicity and nature (real or complex conjugate) of these roots, the procedure for performing partial fraction expansion may vary.

In the following sections, we will present examples illustrating each case of partial fraction expansion based on different root types and multiplicities.

3.2.1 Real roots

Single real root

Let's consider the following rational polynomial with no multiple roots:

$$G(s) = \frac{s}{(s-1)(s+3)(s-4)}$$

it's factorization is based on determining the coefficients α_i such that

$$\frac{\alpha_1}{s-1} + \frac{\alpha_2}{s+3} + \frac{\alpha_3}{s-4} = \frac{s}{(s-1)(s+3)(s-4)} \tag{3.1}$$

This can be accomplished in two ways. The first method involves computing the sum of the three *basic* elements in a parametric form that depends on α_i, and then solving the resulting linear system to determine the parameters that satisfy the equality.

A second, more efficient approach is to recognize that each coefficient can be determined as follows:

$$\alpha_i = \lim_{s \to p_i} \big(s - p_i\big) G(s)$$

In the context of the example, to compute the coefficient α_1 associated with the root $p = 1$, we can multiply Equation 3.1 by the factor $(s-1)$. This yields the following equation:

$$\frac{\cancel{s(s-1)}}{\cancel{(s-1)}(s+3)(s-4)} = \frac{\alpha_1}{\cancel{s-1}}\cancel{(s-1)} + (s-1)\left(\frac{\alpha_2}{s+3} + \frac{\alpha_3}{s-4}\right)$$

Evaluating this expression at the point $s = 1$ will yield an equation in which the only unknown coefficient is α_1. This occurs because the coefficients α_2 and α_3 are effectively *eliminated* due to the multiplication by $(s-1)$, which equals zero at this point. Therefore, we have:

$$\xrightarrow{s=1} \quad \alpha_1 = \frac{1}{(1+3)(1-4)} = -\frac{1}{12}$$

By applying the same method to the other two roots, -3 and 4, we can derive the coefficients:

- For the root -3, we obtain $\alpha_2 = -\frac{3}{28}$.
- For the root 4, we find $\alpha_3 = \frac{4}{21}$.

With these coefficients in hand, we can now verify that:

$$-\frac{1}{12}\frac{1}{s-1} - \frac{3}{28}\frac{1}{s+3} + \frac{4}{21}\frac{1}{s-4} = \frac{s}{(s-1)(s+3)(s-4)}$$

Each basic element can be easily transform using the Laplace table, and in particular

$$\mathscr{L}^{-1}\{G(s)\}(t) \; : \quad -\frac{1}{12}\mathrm{e}^{-t} - \frac{3}{28}\mathrm{e}^{3t} + \frac{4}{21}\mathrm{e}^{-4t}$$

Multiple real roots

Let's now consider a case where the rational polynomial has multiple roots, specifically:

$$G(s) = \frac{1+s}{(s-4)^4}.$$

In this scenario, the partial fraction decomposition should not consist of four identical terms with the denominator $s-4$. Instead, we need to consider four distinct exponential terms that increase in complexity, represented as follows:

$$\frac{\alpha_1}{s-4} + \frac{\alpha_2}{(s-4)^2} + \frac{\alpha_3}{(s-4)^3} + \frac{\alpha_4}{(s-4)^4} = \frac{1+s}{(s-4)^4} \tag{3.2}$$

This problem can be addressed by summing all the elements and solving the resulting linear system for the coefficients α_i. However, a more efficient approach is to recognize that the element a_{n-k}—where n denotes the multiplicity of the root (in this case, $n=4$)—can be expressed as follows:

$$\frac{1}{k!}\lim_{s\to p}\frac{\mathrm{d}^k}{\mathrm{d}s^k}\left[(s-p)^n G(s)\right]$$

This process becomes clearer when we consider the example provided. The first step is to multiply Equation 3.2 by the polynomial $(s-4)^4$, which leads us to the following relationship:

$$1+s = \alpha_1(s-4)^3 + \alpha_2(s-4)^2 + \alpha_3(s-4) + \alpha_4$$

Evaluating this expression at $s=4$ allows us to determine the coefficient $\alpha_4 = 5$. To find the other coefficients, we can employ the proposed *scheme*, which involves differentiating the previous equation with respect to s and then evaluating it at the point $s=4$.

$$\downarrow \mathrm{d}/\mathrm{d}s$$

$$1 = 3\alpha_1(s-4)^2 + 2\alpha_2(s-4) + \alpha_3 \qquad \xrightarrow{s=4} \qquad \alpha_3 = 1$$

$$\downarrow \mathrm{d}/\mathrm{d}s$$

$$0 = 6\alpha_1(s-4) + 2\alpha_2 \qquad \xrightarrow{s=4} \qquad \alpha_2 = 0$$

$$\downarrow \mathrm{d}/\mathrm{d}s$$

$$0 = 6\alpha_1(s-4) \qquad \xrightarrow{s=4} \qquad \alpha_1 = 0$$

Once the coefficients α_i associated with the partial fraction expansion have been determined, we can compute the inverse transform of the original function $G(s)$, which is given by:

$$\mathscr{L}^{-1}\{G(s)\}(t) = \mathscr{L}^{-1}\left\{\frac{1}{(s-4)^3}\right\} + 5\mathscr{L}^{-1}\left\{\frac{1}{(s-4)^4}\right\}$$

$$= \frac{1}{2}\mathrm{e}^{4t}t^2 + 5\frac{1}{6}\mathrm{e}^{4t}t^3$$

■ **Example 3.1 — partial fraction expansion with different roots, one which is multiple.** Let's consider the following rational polynomial $G(s)$ that's already been split in the factor for the partial expansion:

$$\begin{aligned} G(s) &:= \frac{1+s}{(s-1)(s+2)^3} \\ &= \frac{A}{s-1} + \frac{B}{s+2} + \frac{C}{(s+2)^2} + \frac{D}{(s+2)^3} \end{aligned}$$

To calculate the first coefficient A associated with the root $s = 1$, we can follow these straightforward steps:

1. **Multiply by a Factor**: Multiply the last equation by the factor $(s-1)(s+2)^2$. This step isolates the behavior of the function around the specified root.
2. **Evaluate at** $s = 1$: After performing the multiplication, evaluate the resulting expression at $s = 1$. This evaluation will yield the value of the coefficient A.

By following these steps, we can effectively determine the coefficient A related to the root $s = 1$.

$$\begin{aligned} 1+s &= A(s+2)^3 + (s+1)\Big[B(s+2)^2 + C(s+2) + D\Big] \\ &\downarrow \quad s = 1 \\ 2 &= A3^3 \qquad \Longrightarrow \quad A = \frac{2}{27} \end{aligned}$$

The next step is to calculate the remaining coefficients B, C, D associated with the multiple root $s = -2$. However, the current form of the equation does not allow us to apply the previously demonstrated method effectively.

3.2.1.1 Deflating the Equation

To create a suitable equation for our analysis, we need to **deflate** it. This involves expanding the term $(s+2)^3$, which corresponds to the coefficient A. By expanding this term and moving it to the left-hand side of the equation, we can simplify our calculations and isolate the coefficients B, C, D more effectively.

$$\begin{aligned} 1+s-\frac{2}{27}\Big(s^2+6s^2+12s+8\Big) &= (s+1)\Big[B(s+2)^2 + C(s+2) + D\Big] \\ &= \frac{1}{27}\Big(-2s^3-12s^2+35s+11\Big) \end{aligned}$$

Once we have deflated the equation properly, we can proceed with determining the coefficients associated with the multiple root. This will provide clarity on the polynomial's behavior around $s = -2$.

At this stage, we can divide both sides of the equation by the factor $s-1$. If division on the left-hand side is not possible, it indicates that an error has occurred in the calculations leading up to this point. By performing polynomial division, we can simplify the equation and obtain the following result:

$$\frac{1}{27}\left(-2s^2-14s-11\right) = B(s+2)^2 + C(s+2) + D$$

From this point onward it's possible to use the method shown to calculate the coefficient of the multiple roots; starting evaluating the previous expression at point $s = -2$ we get the value

$$D = \frac{-8+28-11}{27} = \frac{1}{3}$$

Continuing with differentiation we get

$$\downarrow \mathrm{d}/\mathrm{d}s$$

$$\frac{-4s-14}{27} = 2B(s+2)+C \qquad \xrightarrow{s=-2} \qquad C = -\frac{2}{9}$$

$$\downarrow \mathrm{d}/\mathrm{d}s$$

$$\frac{-4}{27} = 2B \qquad \xrightarrow{s=-2} \qquad B = -\frac{2}{27}$$

To end the inversion of the starting transform $G(s)$ we can see that

$$\begin{aligned}\mathscr{L}^{-1}\{G(s)\}(t) &= \frac{2}{27}\mathscr{L}^{-1}\left\{\frac{1}{s-1}\right\}(t) - \frac{2}{27}\mathscr{L}^{-1}\left\{\frac{1}{s+2}\right\}(t) \\ &\quad - \frac{2}{9}\mathscr{L}^{-1}\left\{\frac{1}{(s+2)^2}\right\}(t) + \frac{1}{3}\mathscr{L}^{-1}\left\{\frac{1}{(s+2)^3}\right\}(t) \\ &= \frac{2}{27}\mathrm{e}^{t} - \frac{2}{27}\mathrm{e}^{-2t} - \frac{2}{9}\mathrm{e}^{-2t}t + \frac{1}{3}\cdot\frac{1}{2}\mathrm{e}^{-2t}t^2\end{aligned}$$

■

3.2.2 Complex roots

Handling complex roots in the denominator $Q(s)$ of the solution to an ordinary differential equation presents greater computational challenges compared to dealing with real roots. One important observation is that if the complex value

$$\alpha = a + \mathrm{i}b \in \mathbb{C}$$

is a root of the denominator, so that $Q(\alpha) = 0$. It's conjugate

$$\overline{\alpha} = a - \mathrm{i}b$$

is also a root, i.e. $Q(\overline{\alpha}) = 0$. Considering in fact that the polynomial denominator $Q(s)$, evaluated in the point $s = \alpha$, can be expressed as

$$Q(\alpha) = q_0 + \alpha q_1 + \alpha^2 q_2 + \cdots + \alpha^n q_n = 0$$

where q_i are real coefficients. At this point we have to consider the following property of the conjugate that are true for all couple of complex variables α, β:

$$\begin{aligned}&(i) && \overline{(\alpha+\beta)} = \overline{\alpha}+\overline{\beta} \\ &(ii) && \overline{(\alpha\beta)} = \overline{\alpha}\,\overline{\beta} \\ &(iii) && \alpha\overline{\alpha} = |\alpha|^2\end{aligned}$$

Now known that $Q(\alpha) = 0$, than also the conjugate $\overline{Q(\alpha)}$ should be equal to zero (because $\overline{0} = 0$) and so by applying the properties we can see that

$$\begin{aligned}0 = Q(\alpha) = \overline{Q(\alpha)} &= \overline{q_0 + \alpha q_1 + \alpha^2 q_2 + \cdots + \alpha^n q_n} \\ &= \overline{q_0} + \overline{\alpha}\overline{q_1} + \overline{\alpha^2 q_2} + \cdots + \overline{\alpha^n q_n} \\ &= q_0 + \overline{\alpha} q_1 + \overline{\alpha}^2 q_2 + \cdots + \overline{\alpha}^n q_n = Q\big(\overline{\alpha}\big)\end{aligned}$$

This verifies that if α is a root of $Q(s)$, than also it's conjugate $\overline{\alpha}$ is a root of the same polynomial and so we can rewrite the denominator in the form

$$Q(s) = \ldots \underbrace{\big(s-\alpha\big)\big(s-\overline{\alpha}\big)}_{\text{conj. roots}} \ldots$$

Single conjugated complex roots

Let's now assume that we have a rational polynomial that present two conjugated complex roots expressed in the form

$$\frac{P(s)}{Q(s)} = \frac{P(s)}{(s-\alpha)(s-\overline{\alpha})} = \frac{A}{s-\alpha} + \frac{B}{s-\overline{\alpha}}$$

By trying to apply the partial fraction expansion as for the real root previously described, we can demonstrate that the two coefficients A and B are conjugated, so such that $B = \overline{A}$. In fact by summing the two elements we can see that

$$\begin{aligned}\frac{A}{s-\alpha} + \frac{B}{s-\overline{\alpha}} &= \frac{A(s-\overline{\alpha}) + B(s-\alpha)}{(s-\alpha)(s-\overline{\alpha})} \\ &= \frac{(A+B)s - (A\overline{\alpha} + B\alpha)}{(s-\alpha)(s-\overline{\alpha})}\end{aligned}$$

By expanding the definition of the complex variable (so expanding $A = a + \mathrm{i}b$, $B = c + \mathrm{i}d \ldots$) and considering that the polynomial $P(s)$ at the numerator has only real coefficients it's possible to see that, in order to verify the equality, it must be $a = c$ and $b = -d$, and so the two complex values A and B must be conjugated.

Proven that $B = \overline{A}$ now the problem is to find such complex value A in order to satisfy the partial fraction expansion; in order to do so we can try to use the same approach shown for real roots by multiplying both members of the equation by $Q(s)$ and then evaluating the result for $s = \alpha$:

$$\frac{P(s)}{(s-\alpha)(s-\overline{\alpha})} = \frac{A}{s-\alpha} + \frac{B}{s-\overline{\alpha}}$$

$$\xrightarrow[Q(s)]{\times} \qquad P(s) = A(s-\overline{\alpha}) + \overline{A}(s-\alpha)$$

$$\Longrightarrow \qquad P(\alpha) = A(\alpha-\overline{\alpha}) + \overline{A}\cancel{(\alpha-\alpha)}$$

Considering now that $\alpha - \overline{\alpha}$ conresponds to a purely imaginary number equals 2 times the imaginary part of the root α, we can invert that relation in order to explicitly get A:

$$A = \frac{P(\alpha)}{2\mathrm{i}\Im(\alpha)} = -\mathrm{i}\frac{P(\alpha)}{2\Im(\alpha)}$$

Let's now try to consider a real case scenario where we want to deal with the inversion of the following rational polynomial (that's consequently expanded):

$$\frac{s}{\big(s-(1+\mathrm{i})\big)\big(s-(1-\mathrm{i})\big)} = \frac{A}{s-(1+\mathrm{i})} + \frac{\overline{A}}{s-(1-\mathrm{i})}$$

We can now multiply both members of the equation by the denominator $Q(s)$ and then evaluate the result for the point $s = 1+\mathrm{i}$ (root of the denominator):

$$s = A\big(s-(1-\mathrm{i})\big)+\overline{A}\big(s-(1+\mathrm{i})\big)$$

$$\downarrow \quad s = 1+\mathrm{i}$$

$$1+\mathrm{i} = A\big(1+\mathrm{i}-1+\mathrm{i}\big)$$

$$\Longrightarrow \qquad A = \frac{1+\mathrm{i}}{2\mathrm{i}} = \frac{1}{2}\big(1-\mathrm{i}\big)$$

With this value retrieved we can now state that

$$\frac{s}{\big(s-(1+\mathrm{i})\big)\big(s-(1-\mathrm{i})\big)} = \frac{1}{2}\left(\frac{1-\mathrm{i}}{s-(1+\mathrm{i})}+\frac{1+\mathrm{i}}{s-(1-\mathrm{i})}\right)$$

Unfortunately watching the Laplace table we can see there's no possible match between the known transform (where the numerator is expressed as a pure real number) and the complex coefficient retrieved. But an interesting fact that we can note is that by *merging* the two components we can get a rational polynomial with real coefficient: in a general way we can in fact state that

$$\begin{aligned}\frac{A}{s-\alpha}+\frac{\overline{A}}{s-\overline{\alpha}} &= \frac{A(s-\overline{\alpha})+\overline{A}(s-\alpha)}{(s-\alpha)(s-\overline{\alpha})}\\ &= \frac{s\big(A+\overline{A}\big)-A\overline{\alpha}-\overline{A}\alpha}{s^2-s\big(\alpha+\overline{\alpha}\big)+|\alpha|^2}\end{aligned}$$

This rational polynomial can be matched with the following transform:

$$\begin{aligned}\mathscr{L}\{C\mathrm{e}^{at}\cos(\omega t)+D\mathrm{e}^{at}\sin(\Omega t)\} &= C\frac{s-a}{(s-a)^2+\omega^2}+D\frac{\omega}{(s-a)^2+\omega^2}\\ &= \frac{Cs-\big(D\omega-Ca\big)}{(s-a)^2+\omega^2}\end{aligned}$$

Given now the values of α, A, the problem is to find the expression of the coefficients a, ω, C and D; in order to to so we can perform two steps:

1. match the denominator in order to find a and ω; in order to compare the denominator it's useful to expand the denominator $(s-a)^2+\omega^2 = s^2-2sa+a^2+\omega^2$. At this point by relating the function we can see that

$$\cancel{s^2}-s\big(\alpha+\overline{\alpha}\big)+|\alpha|^2 = \cancel{s^2}-2sa+a^2+\omega^2$$

$$\Longrightarrow \quad \begin{cases} -2sa = -s\big(\alpha+\overline{\alpha}\big) \\ \alpha^2+\omega^2 = |\alpha|^2 \end{cases} \quad \Longrightarrow \quad a = \frac{\alpha+\overline{\alpha}}{2} = \Re(\alpha)$$

$$\Longrightarrow \quad \omega = \sqrt{|\alpha|^2-a^2} = \sqrt{|\alpha|^2-\Re(\alpha)^2} = \sqrt{\Im(\alpha)^2} = \pm\Im(\alpha)$$

2. at this point, known a and ω, it's possible to match the numerator in order to get the coefficients C and D:

$$s\big(A+\overline{A}\big)-A\overline{\alpha}-\overline{A}\alpha = Cs-\big(D\omega-Ca\big)$$

$$\implies \quad \begin{cases} A+\overline{A}=C \\ -A\overline{\alpha}-\overline{A}\alpha = D\omega - Ca \end{cases}$$

$$\implies \quad C = A+\overline{A} = 2\Re(\alpha)$$

$$\implies \quad D = \frac{Ca - A\overline{\alpha} - \overline{A}\alpha}{\omega}$$

■ **Example 3.2 — inversion of a transform with two conjugated complex roots.** Let's consider the following rational polynomial that has to be inverted:

$$\frac{s+1}{s^2+s+10}$$

At this point we can verify that the roots of the denominator are complex conjugated by simply applying the quadratic formula to retrieve them

$$s_{1,2} = \frac{-1 \pm \sqrt{-9}}{2}.$$

At this point the problem of the inversion if finding the parameters a, ω, C, D such that

$$\frac{s+1}{s^2+s+10} = C\frac{s-a}{(s-a)^2+\omega^2} + D\frac{\omega}{(s-a)^2+\omega^2}$$

By comparing the denominators we can observe that $-2sa = s$ and so $a = -\frac{1}{2}$ and so

$$a^2+\omega^2 = 10 \qquad \implies \qquad \omega = \sqrt{10-a^2} = \frac{\sqrt{39}}{2}$$

At this point it's possible to compare the denominator (considering that now a, ω are known) that's lead to

$$s+1 = C(s-a)+D\omega \quad \implies \quad C=1, \quad D = \frac{1+Ca}{\omega} = \frac{1}{\sqrt{39}}$$

Retrieved all the coefficient we can explicitly state the inversion of the original rational polynomial:

$$\mathscr{L}^{-1}\left\{\frac{s+1}{s^2+s+10}\right\}(t) = \mathrm{e}^{-\frac{1}{2}t}\left[\cos\left(\frac{\sqrt{39}}{2}t\right) + \frac{1}{\sqrt{39}}\sin\left(\frac{\sqrt{39}}{2}t\right)\right]$$

■

Multiple complex roots

Dealing with multiple complex roots requires even more computational steps; let's consider the generic case of a multiple complex root of multiplicity of 2 that's already expanded:

$$\frac{P(s)}{(s-\alpha)^2(s-\overline{\alpha})^2} = \frac{A}{s-\alpha} + \frac{B}{(s-\alpha)^2} + \frac{\overline{A}}{s-\overline{\alpha}} + \frac{\overline{B}}{(s-\overline{\alpha})^2}$$

As in all the previously cases in order to compute the first parameter B it's possible to multiply both sides of the equation by the denominator $Q(s) = (s-\alpha)^2(s-\overline{\alpha})^2$ and then evaluate the relation at the root $s = \alpha$:

$$P(s) = A(s-\alpha)(s-\overline{\alpha})^2 + B(s-\overline{\alpha})^2 + \overline{A}(s-\overline{\alpha})(s-\alpha)^2 + \overline{B}(s-\alpha)^2$$

$$\Longrightarrow \quad P(\alpha) = B\underbrace{(\alpha-\overline{\alpha})^2}_{=2\Im(\alpha)} \quad \Longrightarrow \quad B = \frac{P(\alpha)}{4\Im(\alpha)^2}$$

Now in order to calculate the second parameter A of the relation we have to deflate te elements associated to B and then, as for the multiple real roots, compute the derivative in respect to the variable s:

$$\downarrow \text{deflation}$$

$$P(s) - B(s-\overline{\alpha})^2 - \overline{B}(s-\alpha)^2 = A(s-\alpha)(s-\overline{\alpha})^2 + \overline{A}(s-\overline{\alpha})(s-\alpha)^2$$

$$\downarrow \ \mathrm{d}/\mathrm{d}s$$

$$P'(s) - 2B(s-\overline{\alpha}) - 2\overline{B}(s-\alpha) = A(s-\overline{\alpha})\left[s-\overline{\alpha}+s(s-\alpha)\right] + \overline{A}(s-\alpha)\left[s-\alpha+2(s-\overline{\alpha})\right]$$

By now evaluating the expression at the point $s=\alpha$ it's possible to have a clear definition of the complex variable A:

$$P'(\alpha) - 2B(s-\overline{\alpha}) = A(\alpha-\overline{\alpha})^2 \quad \Longrightarrow \quad A = \frac{P'(\alpha) - 2B(s-\overline{\alpha})}{4\mathrm{Im}^2(\alpha)}$$

■ **Example 3.3 — inversion of a transform with multiple complex roots.** Let's consider the following rational transform that has to be inverted:

$$\frac{s^2}{(s^2+s+1)^2}$$

The first thing to notice that the multiple roots are complex because calculating their value by using the quadratic formula it happens that

$$s_{1,2} = \frac{-1 \pm \sqrt{1-4}}{2} = -\frac{1}{2} \pm \mathrm{i}\frac{\sqrt{3}}{2}$$

For sake of semplicity of representation of the equation, from now on these roots value will be reported as $z = -\frac{1}{2} + \frac{\sqrt{3}}{2}$ and $\overline{z} = -\frac{1}{2} - \frac{\sqrt{3}}{2}$. At this point it's possible to express the parametric partial fraction expansion of the transform as

$$\frac{s^2}{(s^2+s+1)^2} = \frac{A}{s-z} + \frac{\overline{A}}{s-\overline{z}} + \frac{B}{(s-z)^2} + \frac{\overline{B}}{(s-\overline{z})^2}$$

In order to retrieve the value of the coefficient B we multiply both the sides by the factor

$$Q(s) = (s^2+s+1)^2 = (s-z)^2(s-\overline{z})^2$$

that, evaluated at the point $s=z$, will give the expression

$$z^2 = B\underbrace{\left(z-\overline{z}\right)^2}_{i\sqrt{3}} \quad \Longrightarrow \quad B = -\frac{z^2}{3} = -\frac{\left(-\frac{1}{2}+\mathrm{i}\frac{\sqrt{3}}{2}\right)^2}{3} = -\frac{1}{6}-\mathrm{i}\frac{\sqrt{3}}{6}$$

By deflating the term presenting the parameter B just calculate we can get the relation

$$s^2 - B(s-\overline{z})^2 - \overline{B}(s-z)^2 = A(s-z)(s-\overline{z})^2 + \overline{A}(s-\overline{z})(s-z)^2$$

$$\downarrow \quad \mathrm{d}/\mathrm{d}s$$

$$\begin{aligned} 2s - 2B(s-\overline{z}) - 2\overline{B}(s-z) = {} & A(s-\overline{z})^2 + \overline{A}(s-z)^2 \\ & + 2(A+\overline{A})(s-z)(s-\overline{z}) \end{aligned}$$

Having computed the derivative of the relation in respect to the variable s it's possible to compute the parameter A just by evaluating the result in the point $s = z$:

$$2z - 2B\left(z-\overline{z}\right) = A\left(z-\overline{z}\right)^2 \quad \Longrightarrow \quad A = \frac{4iB\Im(z) - 2z}{4\Im(z)^2} = -\frac{2\mathrm{i}}{3\sqrt{3}}$$

■

Parametric solution for multiple complex roots

To solve the partial fraction expansion of multiple complex roots it's possible to use a parametric formula where the numerator are defined as polynomials A, B in a generic variable t; in particular for the double complex root the expansion can be expressed as

$$\frac{P(s)}{\left(s^2-as+t\right)^2 Q(s)} = \frac{A'(t)s + B'(t)}{s^2-as+t} - \frac{A(t)s+B(t)}{\left(s^2-as+t\right)^2} + \frac{q_t(s,t)}{Q(s)}$$

where the term $q_t(s,t)$ represent the derivative $\partial q/\partial t$ of the remaining term on the numerator after the partial fraction expansion. If the polynomial has 3 complex roots than the parametric decomposition becomes

$$\begin{aligned} 2\frac{P(s)}{\left(s^2-2as+t\right)^3 Q(s)} = {} & 2\frac{A(t)s+B(t)}{\left(s^2-2as+t\right)^3} - 2\frac{A'(t)s+B'(t)}{\left(s^2-2as+t\right)^2} \\ & + \frac{A''(t)s+B''(t)}{s^2-as+t} + \frac{q_{tt}(s,t)}{Q(s)} \end{aligned} \tag{3.3}$$

where $q_{tt} = \partial^2 q/\partial t^2$.

This process can be better seen by considering a numerical example like the one that follows:

$$\begin{aligned} \frac{s^2+1}{(s-1)(s^2-2s+2)^3} = {} & \frac{A}{s-1} + \frac{B_1 s + B_2}{(s^2-2s+2)^3} \\ & + \frac{C_1 s + C_2}{(s^2-2s+2)^2} + \frac{D_1 s + D_2}{s^2-2s+2} \end{aligned}$$

The first coefficient A of the partial fraction expansion (associated to a single real root) can be determined easily with the methods previously described and so

$$A = \left.\frac{s^2+1}{(s^2-2s+2)^3}\right|_{s=1} = 2$$

At this point it's important to acknowledge hot compute the parametric functions A, B, q; in particular this can be done by writing the following relation:

$$\frac{s^2+1}{(s-1)(s^2-2s+2)} = \frac{A(t)s+B(t)}{s^2-2s+t} + \frac{q(t)}{s-1}$$

We can note that the left hand side is equal to the original rational polynomial with the only difference that the term s^2-2s+2 is now considered once; the right hand side is instead the *correct* partial fraction expansion of the other side with the difference that the coefficient A, B, q are now function of the variable t and the last known coefficient of the complex root (in this case the $+$) has been replaced with the variable t.

By multiplying both member of the equation by the denominator of the left hand side we derive the following equation:

$$s^2+1 = \Big(A(t)s+B(t)\Big)(s-1) + q(t)\big(s^2-2s+t\big)$$

This expression allows now us to compute the parametric form of the function A, B, q; in particular considering that $s^2-2s+t=0$ and so $s^2 = 2s-t$ we can rewrite the previous expression as

$$2s-t+1 = A(t)(2s-t) - A(t)s + B(t)s - B(t) + \cancel{q(t)\big(2s-t-2s+t\big)}$$

$$\implies \quad \begin{cases} 2 = A(t)+B(t) \\ -t+1 = -tA(t) - B(t) \end{cases}$$

$$\implies \quad A(t) = 1 - \frac{2}{t-1}\,, \quad B(t) = 1 + \frac{2}{t-1}$$

In order to have the complete the partial fraction expansion we need to compute the derivative (up to the second order) and evaluate them for $t=3$ (noting that \$A(2) = -1 \$ and $B(2)=3$):

$$A' = \frac{2}{(t-1)^2}\bigg|_{t=2} = 2 \qquad B' = -\frac{2}{(t-1)^2}\bigg|_{t=2} = 2$$
$$A'' = -\frac{4}{(t-1)^3}\bigg|_{t=2} = -4 \qquad B'' = \frac{4}{(t-1)^3}\bigg|_{t=2} = 4$$

With all the coefficient defined is possible to express numerically the partial fraction expansion by using Equation 3.3:

$$\frac{2}{s-1} + \frac{-s+3}{(s^2-2s+2)^3} - 2\frac{s-1}{(s^2-2s+2)^2} + \frac{1}{2}\frac{-4s+4}{s^2-2s+2}$$

▪ **Example 3.4 — computation of a partial fraction expansion.** Considering the rational polynomial in s

$$\xi: \quad \frac{P(s)}{Q(s)} = \frac{s(s^2+7)}{(s-1)(s+1)^3} = \frac{A}{s-1} + \frac{B}{s+1} + \frac{C}{(s+1)^2} + \frac{D}{(s+1)^3}$$

finding the partial fraction expansion means determining the coefficients $A, B, C, D \in \mathbb{R}$ that satisfies the equation. In order to solve this problem we can define ζ as the multiplication of ξ by $Q(s)$ and so

$$\zeta: \quad s(s^2+7) = A(s+1)^3 + B(s-1)(s+1)^2 + C(s-1)(s+1) + D(s-1)$$

In order to find some coefficient we can note that such polynomial equation must be verified if we evaluate ζ for some *strategic* point, in particular for $s=1$ (that's a root of Q) we have $8=2^3A$ that determines $A=1$, while if we evaluate for $s=-1$ we retrieve $-8=-2D$ and so $D=4$. We can so inflate ζ by moving on the left hand side the associated terms on A and D determining the rewritten expression:

$$\begin{aligned}\zeta_r: \quad & s^3+7s-A(s^3+3s^2+3s+1)-D(s-1)\\ &= B(s-1)(s+1)^2+C(s-1)(s+1)\end{aligned}$$

$$\begin{aligned}& s^3+7s-s^3-3s^2-3s-1-4s+4\\ &= B(s-1)(s^2+2s+1)+C(s^2-1)\end{aligned}$$

$$-3s^2+3=B(s^3+s^2-s-1)+C(s^2-1)$$

In order to determine the value of the other coefficient B,C we can start deriving ζ in the variable s:

$$\zeta'=\frac{\mathrm{d}\zeta}{\mathrm{d}s}: \qquad -6s=B(3s^2+2s-1)+2Cs$$

Evaluating this expression for $s=-1$ (associated to the root of both B and C) we have that $6=-2Cs$ determining $C=-3$; after inflation we so have

$$\zeta_r': \qquad 0=B(3s^2+2s-1)$$

$$\zeta''=\frac{\mathrm{d}\zeta'}{\mathrm{d}s}: \qquad 0=B(6s+2)$$

With the last evaluation on $s=-1$ we obtain $0=-4B$ requiring $B=0$; with that said the overall partial fraction expansion is

$$\frac{s(s^2+7)}{(s-1)(s+1)^3}=\frac{1}{s-1}-\frac{3}{(s+1)^2}+\frac{4}{(s+1)^3}$$

■

3.3 Table for the Laplace transform

	$f(t)$	$\mathscr{L}\{f\}(s)$
1	$af(t)+bg(t)$	$af(s)+bg(s)$
2	$f(at)$	$\frac{1}{a}f\left(\frac{s}{a}\right) \qquad [a>0]$
3	1	$\frac{1}{s}$
4	t	$\frac{1}{s^2}$
5	t^k	$\frac{k!}{s^{k+1}}$
6	$t^n f(t)$	$(-1)^n\frac{\mathrm{d}^n}{\mathrm{d}s^n}f(s)$
7	$f'(t)$	$sf(s)-f(0^+)$
8	$f''(t)$	$s^2f(s)-f'(0^+)-sf(0^+)$

	$f(t)$	$\mathscr{L}\{f\}(s)$
9	$\frac{\mathrm{d}^n}{\mathrm{d}t^n} f(t)$	$s^n f(s) - \sum_{j=0}^{n-1} s^{n-j-1} f^{(j)}(0^+)$
10	$e^{\alpha t} - e^{\beta t}$	$\frac{\alpha - \beta}{(s-\alpha)(s-\beta)}$
11	$e^{at} f(t)$	$f(s-a)$
12	$f(t-a)$	$e^{-as} f(s)$
13	$(f \star g)(t)$	$f(s)\, g(s)$
14	$e^{at} \cos \omega t$	$\frac{s-a}{(s-a)^2 + \omega^2}$
15	$e^{at} \sin \omega t$	$\frac{\omega}{(s-a)^2 + \omega^2}$
16	$e^{at} \cosh \omega t$	$\frac{s-a}{(s-a)^2 - \omega^2}$
17	a^{bt}	$\frac{1}{s - b \log a}$
18	$e^{at} \sinh \omega t$	$\frac{\omega}{(s-a)^2 - \omega^2}$
19	$\int_0^t f(z)\, \mathrm{d}z$	$\frac{1}{s} f(s)$
20	$e^{at} t^n$	$\frac{n!}{(s-a)^{n+1}}$



4. Boundary value problem

Usually some differential equation problems requires some condition that are not defined at the initial time $t = 0$ of the time axes, but in a more general point $t = t^\star$: in order to solve this problem a parametric approach is recomended. As example let's consider the following problem:

$$\begin{cases} x''(t) + y'(t) = \sin(t) \\ y'(t) + x'(t) = 1 + \sin(t) \\ \qquad x(0) = 0, \quad y(0) = -1, \quad x(1) = 1 \end{cases}$$

As we can see the last condition constrains the solution $x(t), y(t)$ at the time $t = 1$; in order to solve the problem in the Laplace domain we apply the transform on the system of differential equation and so

$$\begin{cases} s^2\hat{x}(s) - x'(0) - sx(0) + s\hat{y}(s) - y(0) = \dfrac{1}{s^2+1} \\ \qquad s\hat{y}(s) - y(0) + s\hat{x}(s) - x(0) = \dfrac{1}{s} + \dfrac{1}{s^2+1} \end{cases}$$

If we now try to substitute the boundary condition on the problem we can see that we have no information for $x'(t)$ at the initial time: the idea now is to solve the problem parametrically (assign a variable, in this case A, to the initial condition $x'(0)$) and solve the problem this way. At the and we have to find the constant value A such that $x(1) = 1$. So by substituting we get

$$\begin{cases} s^2\hat{x}(s) - A + s\hat{y}(s) + 1 = \dfrac{1}{s^2+1} \\ \qquad s\hat{y}(s) + 1 + s\hat{x}(s) = \dfrac{1}{s} + \dfrac{1}{s^2+1} \end{cases}$$

To simplify the calculation of the algebraic solution of the functions $\hat{x}, \hat{y}$ we rewrite the system in matrix form that, by inversion, can determine the solutions (by using the Kramer method):

$$\begin{bmatrix} s^2 & s \\ s & s \end{bmatrix} \begin{pmatrix} \hat{x}(s) \\ \hat{y}(s) \end{pmatrix} = \begin{pmatrix} \dfrac{1}{s^2+1} + A - 1 \\ \dfrac{1}{s} + \dfrac{1}{s^2+1} - 1 \end{pmatrix}$$

$$\implies \qquad \hat{x}(s) = \frac{\det \begin{bmatrix} \dfrac{1}{s^2+1} + A - 1 & s \\ \dfrac{1}{s} + \dfrac{1}{s^2+1} - 1 & s \end{bmatrix}}{\det \begin{bmatrix} s^2 & s \\ s & s \end{bmatrix}} = \frac{sA-1}{s^2(s-1)}$$

Having the boundary condition set on $x(t)$, our focus is to invert only the related transform $\hat{x}(s)$ (because, for the sake of the problem, it's not relevant determining $y(t)$); by doing the parametric partial fraction expansion we have

$$\frac{sA-1}{s^2(s-1)} = \frac{c_1}{s-1} + \frac{c_2}{s} + \frac{c_3}{s^2}$$

where, by solving the linear system, the coefficients are equals to $c_1 = A-1$, $c_2 = 1-A$ and $c_3 = 1$. At this point we have everything in order to compute the inverted function of $\hat{x}(s)$:

$$x(t) = \mathscr{L}^{-1}\{\hat{x}(s)\}(t) = (A-1)\mathrm{e}^t + (1-A) + t$$

The last thing to do now is to evaluate the function $x(t)$ at the time $t = 1$ and equal the result to 1 in order to determine the coefficient A related to the initial first derivative of x:

$$x(1) = (A-1)e + (1-A) + 1 = 1 \qquad \implies \qquad A = 1$$

■ **Example 4.1 — boundary value problem.** Considering the parametric differential equation problem depending on the parameter $A \in \mathbb{R}$

$$\begin{cases} x''(t) - x'(t) = t \\ x(0) = 1 \\ x'(0) = A \end{cases}$$

the solution can be found in the Laplace domain; starting off with the transformation of the problem using the $\mathscr{L}$ operator we have that

$$\begin{aligned} \mathscr{L}\{x''(t)\}(s) &= s^2\hat{x}(s) - x'(0) - sx(0) \\ \mathscr{L}\{x'(t)\}(s) &= s\hat{x}(s) - x(0) \\ \mathscr{L}\{t\}(s) &= \frac{1}{s^2} \end{aligned}$$

$$\mathscr{L}\{x''(t) - x'(t) = t\}(s) \quad \mapsto \quad s^2\hat{x}(s) - A - s - s\hat{x}(s) + 1 = \frac{1}{s^2}$$

Solving this expression for $\hat{x}(s)$ and performing the partial fraction expansion on the result (whose computation is here skipped) we obtain

$$\begin{aligned} \hat{x}(s) &= \frac{\frac{1}{s^2} + s + A - 1}{s^2 - s} = \frac{s^3 + As^2 - s^2 + 1}{s^3(s-1)} \\ &= \frac{A+1}{s-1} - \frac{A}{s} - \frac{1}{s^2} - \frac{1}{s^3} \end{aligned}$$

If we now want to compute the parameter A such that $x(1) = -\frac{1}{2}$ we need to invert $\hat{x}(s)$ in order to have the solution in the time domain. Considering so that

$$\mathscr{L}^{-1}\left\{\frac{1}{s-1}\right\}(t) = \mathrm{e}^t$$
$$\mathscr{L}^{-1}\left\{\frac{1}{s}\right\}(t) = 1$$
$$\mathscr{L}^{-1}\left\{\frac{1}{s^2}\right\}(t) = t$$
$$\mathscr{L}^{-1}\left\{\frac{1}{s^3}\right\}(t) = \frac{t^2}{2!}$$

and so

$$x(t) = \mathscr{L}^{-1}\left\{\hat{x}(s)\right\}(t) = (A+1)\mathrm{e}^t - A - t - \frac{t^2}{2}$$
$$x(1) = (A+1)e - A - 1 - \frac{1}{2} = -\frac{1}{2}$$

Solving the equation we retrieve the value of A that's -1.

■

II KKT

5 Calculus Review 49
5.1 Higher Regularity Conditions 49
5.2 Local Minimum 50
5.3 Conditions for a Local Minimum 50

6 Lagrange Multipliers 53
6.1 Summary of constrained minimization with proofs 56

7 Sylvester theorem 63
7.1 Why use sylvester theorem 64

8 Inequality constraints 69

9 Summary of KKT with proofs 81
9.1 Constrained minima, NLP problem 85
9.2 Constrained minima and KKT 86
9.3 Other constraints qualifications 89

10 Worked examples 91
10.1 Example 1: Detailed Example of Constrained Minimization 91
10.2 Example 2: Detailed example of KKT 94
10.3 Example 3 99
10.4 Example 4 102
10.5 Example 5 106

The **Karush-Kuhn-Tucker (KKT) conditions** are a set of criteria used to determine whether a solution to a constrained optimization problem is optimal. They extend the method of Lagrange multipliers to handle not only equality constraints but also inequality constraints.

Learning objectives

After this part, you should be able to formulate a constrained minimization problem, distinguish necessary from sufficient optimality conditions, and apply the KKT conditions while checking an appropriate constraint qualification.



5. Calculus Review

When minimizing a function, certain regularity conditions must be assumed to simplify the analysis and prove meaningful results. The first condition is a minimal regularity requirement for the function being studied.

Definition 5.1 — first order regularity. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a multivariate function satisfying the following properties:

- $f \in C^1(\mathbb{R}^n)$, meaning $f(\boldsymbol{x})$ is continuously differentiable, and the gradient (row) vector $\nabla f(\boldsymbol{x})$ exists.
- The gradient $\nabla f(\boldsymbol{x})$ is **Lipschitz continuous**, i.e.,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq \gamma \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

With these regularity assumptions, the Taylor series for f can be expanded up to the first order as follows:

$$f(\boldsymbol{x} + t\boldsymbol{d}) = f(\boldsymbol{x}) + t\nabla f(\boldsymbol{x})\boldsymbol{d} + \wr(t)$$

where the remainder term $\wr(t)$ satisfies

$$\lim_{t \to 0} \frac{\wr(t)}{t} = 0$$

This expansion does not require the existence of second-order derivatives, but it does require that the gradient is **Lipschitz continuous**.

5.1 Higher Regularity Conditions

In the classification of minimum or maximum points, additional regularity conditions are needed.

Definition 5.2 — second order regularity. Consider again a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, but now assume the following:

- $f \in C^2(\mathbb{R}^n)$, meaning $f(\boldsymbol{x})$ is twice continuously differentiable. Both the gradient vector $\nabla f(\boldsymbol{x})$ and the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ exist.
- The Hessian $\nabla^2 f(\boldsymbol{x})$ is **Lipschitz continuous**, i.e.,

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{y})\| \leq \gamma \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

With this stronger regularity assumption, the Taylor series can be expanded up to the second order as:

$$f(\boldsymbol{x}+t\boldsymbol{d}) = f(\boldsymbol{x}) + t\nabla f(\boldsymbol{x})\boldsymbol{d} + \frac{t^2}{2}\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{d} + \wr(t^2)$$

Here, the remainder term $\wr(t^2)$ does not require the existence of third-order derivatives but does depend on the **Lipschitz continuity of the Hessian**.

In summary, the first-order expansion requires $f \in C^1$ and the Lipschitz continuity of the gradient, while second-order expansions demand $f \in C^2$ and the Lipschitz continuity of the Hessian matrix. These conditions provide the necessary regularity to perform effective analyses of function behavior.

A point $\boldsymbol{x}^\star \in \mathbb{R}^n$ is considered a **global minimum** if

$$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n.$$

5.2 Local Minimum

A point $\boldsymbol{x}^\star \in \mathbb{R}^n$ is classified as a **local minimum** of a function $f : \mathbb{R}^n \to \mathbb{R}$ if:

$$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in B(\boldsymbol{x}^\star, \delta),$$

where $B(\boldsymbol{x}^\star, \delta)$ represents the open ball centered at $\boldsymbol{x}^\star$ with radius $\delta > 0$. This implies that within a small neighborhood around $\boldsymbol{x}^\star$, the value of the function $f(\boldsymbol{x})$ is always greater than or equal to $f(\boldsymbol{x}^\star)$.

5.2.0.1 Strict Local Minimum

If the inequality is **strict**, meaning:

$$f(\boldsymbol{x}^\star) < f(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in B(\boldsymbol{x}^\star, \delta) \text{ and } \boldsymbol{x} \neq \boldsymbol{x}^\star,$$

then $\boldsymbol{x}^\star$ is classified as a **strict local minimum**. In this case, no other point in the neighborhood has a function value equal to or less than $f(\boldsymbol{x}^\star)$.

5.3 Conditions for a Local Minimum

5.3.1 First-Order Conditions

For first order condition we assume first order regularity (Definition 5.1). A **necessary condition** for $\boldsymbol{x}^\star \in \mathbb{R}^n$ to be a local minimum is that the gradient of f at $\boldsymbol{x}^\star$ must vanish. Mathematically, this is expressed as:

$$\nabla f(\boldsymbol{x}^\star)^T = \boldsymbol{0}$$

This condition is known as the **first-order necessary condition**. However, it is not sufficient to confirm that $\boldsymbol{x}^\star$ is a minimum. This condition alone could indicate a minimum, a maximum, or even a **saddle point** (a point that is neither a minimum nor a maximum).

5.3.2 Second-Order Conditions

To further classify $\boldsymbol{x}^\star$, we need to examine the second-order derivatives of the function and assume seond order regularity (Definition 5.2).

5.3.2.1 Necessary Conditions

Assuming that $f \in C^2(\mathbb{R}^n)$, with Lipshitz continuous second derivatives, then the Hessian matrix at $\boldsymbol{x}^\star$ must be **semi-positive definite**, which means:

$$\boldsymbol{d}^T\nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0 \quad \text{for all } \boldsymbol{d} \in \mathbb{R}^n$$

5.3.2.2 Sufficient Conditions

A **sufficient condition** for $\boldsymbol{x}^\star$ to be a **strict local minimum** is that the Hessian matrix $\nabla^2 f(\boldsymbol{x}^\star)$ is **positive definite**. That is:

$$\boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{d} > 0 \quad \text{for all} \, \boldsymbol{d} \neq \boldsymbol{0} \in \mathbb{R}^n$$

If the Hessian is positive definite, then $\boldsymbol{x}^\star$ is guaranteed to be a strict local minimum. This condition is equivalent to saying that all the eigenvalues of the Hessian matrix $\nabla^2 f(\boldsymbol{x}^\star)$ are positive.

In summary:

- A point where the gradient vanishes may be a local minimum, maximum, or saddle point.
- To classify a point as a **(strict) local minimum**, the Hessian must be **positive definite**.
- If the Hessian matrix is not **semi-positive definite** the point is surely not a minimum point.

There is a grey zone when the point satisfy necessary condition:

- the Hessian matrix is **semi-positive definite**

but not the sufficient condition, i.e:

- the Hessian matrix is not **positive definite**

in this case the point can be a mininimal point or not a require further investigation.



6. Lagrange Multipliers

In constrained optimization, the goal is to minimize a function $f \in C^2(\mathbb{R}^n)$ (twice continuously differentiable) subject to m equality constraints defined by $h_k \in C^2(\mathbb{R}^n)$. This problem can be formulated as:

$$\begin{aligned} &\text{Minimize:} && f(\boldsymbol{x}) \\ &\text{Subject to:} && h_k(\boldsymbol{x}) = 0, \quad k = 1, \ldots, m \end{aligned}$$

This means that we want to find the point $\boldsymbol{x}$ that minimizes the objective function $f(\boldsymbol{x})$ while also satisfying the m constraints given by $h_k(\boldsymbol{x}) = 0$ for each k.

The challenging analytical problem of constrained minimization can be simplified using the **Lagrange multiplier theorem**. The idea is to incorporate the constraints directly into the optimization problem through additional variables called **Lagrange multipliers**.

Let f be the function to be minimized, and let $\boldsymbol{h}$ represent the constraint functions (so that $f, \boldsymbol{h} \in C^2(\mathbb{R}^n)$). If $\boldsymbol{x}^\star$ is a local minimum of f that satisfies all the constraints (i.e., $\boldsymbol{h}(\boldsymbol{x}^\star) = \boldsymbol{0}$), then if $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ has maximum rank, there exist m scalars $\lambda_1, \ldots, \lambda_m$ such that:

$$\nabla f(\boldsymbol{x}^\star) = \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^\star)$$

This equation indicates that at the optimal point $\boldsymbol{x}^\star$, the gradient of the objective function f is a linear combination of the gradients of the constraint functions h_k. The scalars λ_k are called **Lagrange multipliers**.

6.0.0.1 Lagrangian Function

The problem is transformed into a simpler form using the **Lagrangian function** $\mathscr{L}$, which incorporates both the objective function and the constraints:

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x})$$

Here, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)$ is the vector of Lagrange multipliers. The goal now is to find stationary points of the Lagrangian $\mathscr{L}$ by solving a system of equations derived from the first-order conditions.

6.0.0.2 Solving the Constrained Minimization Problem

To solve the problem, we need to find the points $\boldsymbol{x}^\star$ that satisfy the Lagrange multiplier conditions. This involves solving the following system of equations:

$$\begin{cases} \nabla_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}) = \mathbf{0} \\ \nabla_{\boldsymbol{\lambda}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = -\boldsymbol{h}(\boldsymbol{x}) = \mathbf{0} \end{cases}$$

The first equation corresponds to finding points where the gradient of the Lagrangian $\mathscr{L}$ with respect to $\boldsymbol{x}$ is zero. The second equation ensures that the constraints $h_k(\boldsymbol{x}) = 0$ are satisfied.

All points $\boldsymbol{x}^\star$ that solve this system are **candidates** for being local minima, maxima, or saddle points. Since we are searching for stationary points of the Lagrangian, further analysis is required to determine whether these points are indeed minima or maxima.

6.0.0.3 Second-Order Conditions

To determine whether a candidate point $\boldsymbol{x}^\star$ is a local minimum in a constrained optimization problem, we must examine the **second-order conditions**. These conditions involve the Hessian matrix of the **Lagrangian** function. The second-order condition checks whether this Hessian is **positive definite** when restricted to the null space of the Jacobian of the constraints, $\nabla \boldsymbol{h}(\boldsymbol{x})$. This is necessary to ensure that the second derivative of the objective function along feasible directions is positive, which would confirm the existence of a local minimum.

Sufficient Condition Mathematically, the second-order condition is expressed as:

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) \boldsymbol{z} > 0 \quad \text{for all } \boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$$

Here:

- $\boldsymbol{z}$ is any vector in the **null space** (or kernel) of the constraint Jacobian, $\ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$.
- The condition means that in the directions allowed by the constraints, the Hessian of the Lagrangian should be **positive**.

This condition ensures that the Lagrangian's second derivative is positive in the feasible directions, confirming that the point $\boldsymbol{x}^\star$ is a **local minimum** rather than a maximum or saddle point.

Necessary Condition For $\boldsymbol{x}^\star$ to be a local minimum, a necessary second-order condition is that the Hessian $\nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})$ is **semi-positive definite** in the feasible directions, which implies:

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) \boldsymbol{z} \geq 0 \quad \forall \boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$$

In other words, the Hessian of the Lagrangian must not have negative curvature along any direction that lies in the kernel of the constraint gradients. This is a **necessary condition** but not sufficient to confirm a local minimum.

▪ **Example 6.1 — constrained minimization problem.** Let's consider the problem on where we want to minimize the function $f : \mathbb{R}^2 \to R$ using the constraint h defined as

$$f(x,y) = \mathrm{e}^{x^2 - y^2}, \qquad h(x,y) = x - y^2$$

In order to solve this problem we at first need to build the lagrangian (having only one constraint $\boldsymbol{\lambda}$ reduces to a scalar) and so

$$\mathscr{L}(x,y,\lambda) = \mathrm{e}^{x^2-y^2} - \lambda\left(x - y^2\right)$$

We now need to compute the stationary points of the lagrangian and this means solving the following non linear system of equations:

$$\begin{cases} \nabla_x \mathscr{L}(x,y,\lambda) = 2xe^{x^2-y^2} - \lambda = 0 \\ \nabla_y \mathscr{L}(x,y,\lambda) = -2ye^{x^2-y^2} + 2\lambda y = 0 \\ \nabla_\lambda \mathscr{L}(x,y,\lambda) = -x + y^2 = 0 \end{cases}$$

$$\implies \quad (x,y,\lambda) = (0,0,0),$$

$$\implies \quad (x,y,\lambda) = \left(\frac{1}{2}, \frac{1}{\sqrt{2}}, e^{-\frac{1}{4}}\right),$$

$$\implies \quad (x,y,\lambda) = \left(\frac{1}{2}, -\frac{1}{\sqrt{2}}, e^{-\frac{1}{4}}\right)$$

To determine now it this points are local maximum or minimum we have to firstly define the general gradient of the constraint map and then the hessian of the Lagrangian in respect to the variable $\boldsymbol{x} = (x,y)$:

$$\nabla h(x,y) = (1, -2y)$$

$$\nabla^2_{(x,y)} \mathscr{L} = \begin{bmatrix} (4x^2+2)e^{x^2-y^2} & -4xye^{x^2-y^2} \\ -4xye^{x^2-y^2} & (4y^2-2)e^{x^2-y^2} + 2\lambda \end{bmatrix}$$

Now we have to check each stationary point independently:

1. when $x = y = \lambda = 0$ we have that $\nabla h = (1,0)$ while

$$\nabla^2_{(x,y)} \mathscr{L} = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

   By computing the null space of the vector $(1,0)$ we can see that all the vectors in the form $(0,\alpha)$ match the definition; we can now check if the point is of maximum/minimum be determining if the matrix $\nabla^2_{(x,y)} \mathscr{L}$ is positive or negative defined:

$$\begin{pmatrix} 0 & \alpha \end{pmatrix} \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix} \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = -2\alpha^2 \leq 0$$

   for all $\alpha \in \mathbb{R}$. The hessian matrix is negative defined and so the point $(x,y) = (0,0)$ is a local maximum.

2. evaluating for the second point

$$x = \frac{1}{2}, \quad y = \frac{1}{\sqrt{2}}, \quad \text{and} \quad \lambda = e^{-\frac{1}{4}}$$

   we can compute the gradient $\nabla h = (1, -\sqrt{2})$ of the constraint map that determines a null space of the form $(\alpha\sqrt{2}, \alpha)$. Given the hessian matrix of the transform we can see that it's positive defined, in fact

$$e^{-\frac{1}{4}} \begin{pmatrix} \alpha\sqrt{2} & \alpha \end{pmatrix} \begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix} \begin{pmatrix} \alpha\sqrt{2} \\ \alpha \end{pmatrix} = 4e^{-\frac{1}{2}} \alpha^2 > 0$$

   for all $\alpha \in \mathbb{R}$. This means that the point is a local minimum.

3. considering instead the last point

$$x = \frac{1}{2}, \quad y = -\frac{1}{\sqrt{2}} \quad \text{and} \quad \lambda = e^{-\frac{1}{4}}$$

we have a similar gradient $\delta h = (1, \sqrt{2})$ that determines a kernel in the form $(\alpha\sqrt{2}, -\alpha)$. Evaluating the hessian on the null space base we can see that the matrix is positive defined, in fact

$$e^{-\frac{1}{4}} \begin{pmatrix} \alpha\sqrt{2} & -\alpha \end{pmatrix} \begin{bmatrix} 3 & -\sqrt{2} \\ -\sqrt{2} & 2 \end{bmatrix} \begin{pmatrix} \alpha\sqrt{2} \\ -\alpha \end{pmatrix} = 4e^{-\frac{1}{2}}\alpha^2 > 0$$

for all $\alpha \in \mathbb{R}$

We can see that the both points

$$\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) \quad \text{and} \quad \left(\frac{1}{2}, -\frac{1}{\sqrt{2}}\right)$$

are local minimum and they are both also global minimum because we can see that

$$f\left(\frac{1}{2}, \frac{1}{\sqrt{2}}\right) = f\left(\frac{1}{2}, -\frac{1}{\sqrt{2}}\right) = e^{-\frac{1}{4}}.$$

■

■ **Example 6.2 — determination of the non-linear system.** Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h_1(x,y,z) = x^2 + y^2 - 2 = 0 \\ & h_2(x,y,z) = x + z - 1 = 0 \end{aligned}$$

the solution can be computed by firstly determining the lagrangian

$$\mathscr{L} = f - \lambda_1 h_1 - \lambda_2 h_2$$

of the problem that's

$$\mathscr{L}(x,y,z,\lambda_1,\lambda_2) = x - y + z^2 - \lambda_1\left(x^2 + y^2 - 2\right) - \lambda_2\left(x + z - 1\right)$$

The first order necessary condition related to the Lagrange multipliers states that a point, to be a minimum, must be a stationary one and so the candidate minimum points for this problem can be computed by solving the following system of non-linear equation:

$$\begin{cases} 1 - 2\lambda_1 x - \lambda_2 = 0 & : \dfrac{\partial}{\partial x}\mathscr{L} \\ -1 - 2\lambda_1 y = 0 & : \dfrac{\partial}{\partial y}\mathscr{L} \\ 2z - \lambda_2 = 0 & : \dfrac{\partial}{\partial z}\mathscr{L} \\ x^2 + y^2 = 2 & : -\dfrac{\partial}{\partial \lambda_1}\mathscr{L} = h_1 \\ x + z = 1 & : -\dfrac{\partial}{\partial \lambda_2}\mathscr{L} = h_2 \end{cases}$$

■

6.1 Summary of constrained minimization with proofs

Consider the following constrained optimization problem:

$$\begin{aligned} \text{Minimize:} \quad & f(\boldsymbol{x}) \\ \text{Subject to:} \quad & h_i(\boldsymbol{x}) = 0 \quad \text{for} \quad i = 1, 2, \ldots, m \end{aligned} \tag{6.1}$$

Solution Algorithm

1. **Form the Lagrangian**: Define the Lagrangian function by incorporating the constraints into the objective function:

$$\mathscr{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x})$$

   where λ_k are the Lagrange multipliers.
2. **Solve the first order conditions**: Find the critical points by solving the system of equations given by the gradients of the Lagrangian:

$$\nabla_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x},\boldsymbol{\lambda}) = \mathbf{0}^T \quad \text{and} \quad \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}$$

   This nonlinear system must be solved simultaneously for $\boldsymbol{x}$ and $\boldsymbol{\lambda}$.
3. **Check linear independence of constraints**: For each solution $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, compute the Jacobian matrix $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ of the constraint functions. Verify that the rows of this matrix are linearly independent, meaning the matrix has full rank.
4. **Calculate the kernel of the constraint Jacobian**: Compute the matrix $\boldsymbol{K}$, which forms the null space (kernel) of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$. This satisfies:

$$\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{K} = \mathbf{0}$$

5. **Compute the reduced Hessian**: The reduced Hessian $\boldsymbol{H}$ is a key matrix for analyzing the second-order conditions. It is defined as:

$$\boldsymbol{H} = \boldsymbol{K}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)\boldsymbol{K}$$

6. **Verify second-order conditions**:
   - **Necessary condition**: $\boldsymbol{H}$ must be **semi-positive** definite for $\boldsymbol{x}^\star$ to be a candidate for a local minimum.
   - **Sufficient condition**: $\boldsymbol{H}$ is **positive** definite, ensuring that $\boldsymbol{x}^\star$ is a strict local minimum.

The sufficient condition for optimality states that if the Hessian matrix, $\boldsymbol{H}$, is positive definite at the point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$, then $\boldsymbol{x}^\star$ is a strict local minimizer of the original constrained problem.

The following theorem formalizes this sufficient condition for optimality.

Theorem 6.1 — of Lagrange Multipliers. Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$ and $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$ with $\boldsymbol{x}^\star$ be a local minimum of problem Equation 6.1, i.e.,

- $f(\boldsymbol{x}^\star)$ is a minimum
- subject to $\boldsymbol{h}(\boldsymbol{x}^\star) = \mathbf{0}$.
- and $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ is full rank

and define the Lagrangian

$$\mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) = f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}^\star)$$

then there exist scalars λ_k, $k = 1, \ldots, m$, such that:

1. **first order necessary condition**:

$$\nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) = \nabla f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^\star) = 0. \tag{A}$$

2. **second order condition**:
   for all $\boldsymbol{z} \in \mathbb{R}^n$ such that $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{z} = \boldsymbol{0}$ and $\boldsymbol{z} \neq \boldsymbol{0}$, it holds that:
   - **necessary**:
   $$\boldsymbol{z}^T \left(\nabla^2 f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla^2 h_k(\boldsymbol{x}^\star) \right) \boldsymbol{z} \geq 0, \tag{B}$$
   i.e., the matrix $\nabla_x^2 \big(f(\boldsymbol{x}^\star) - \boldsymbol{\lambda} \cdot \boldsymbol{h}(\boldsymbol{x}^\star) \big)$ is **positive semi-definite** in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$.
   - **sufficient**:
   $$\boldsymbol{z}^T \left(\nabla^2 f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla^2 h_k(\boldsymbol{x}^\star) \right) \boldsymbol{z} \geq \varepsilon \|\boldsymbol{z}\|^2 > 0, \tag{C}$$
   i.e., the matrix $\nabla_x^2 \big(f(\boldsymbol{x}^\star) - \boldsymbol{\lambda} \cdot \boldsymbol{h}(\boldsymbol{x}^\star) \big)$ is **positive definite** in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$.

Proof.

Let $\boldsymbol{x}^\star$ be a local minimum. Then, there exists $\varepsilon > 0$ such that

$$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}), \quad \text{for all } \boldsymbol{x} \in \mathscr{B} \text{ with } \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0}, \tag{6.2}$$

where

$$\mathscr{B} = \{ \boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^\star\| \leq \varepsilon \}. \tag{6.3}$$

We define a sequence of functions:

$$f_k(\boldsymbol{x}) = f(\boldsymbol{x}) + k\|\boldsymbol{h}(\boldsymbol{x})\|^2 + \alpha\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2, \quad \alpha > 0, \tag{6.4}$$

with the corresponding sequence of unconstrained local minima $\boldsymbol{x}_k \in \mathscr{B}$, i.e.,

$$\boldsymbol{x}_k = \operatorname*{argmin}_{\boldsymbol{x} \in \mathscr{B}} f_k(\boldsymbol{x}).$$

Since $\boldsymbol{x}_k$ is contained within the compact ball $\mathscr{B}$, a convergent subsequence $\boldsymbol{x}_{k_j} \to \bar{\boldsymbol{x}} \in \mathscr{B}$ exists by compactness.
The rest of the proof to verify that $\bar{\boldsymbol{x}} = \boldsymbol{x}^\star$ and it a minimum.

Step 1:

$\boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}$.
Notice that the sequence x_k satisfy $f_k(\boldsymbol{x}_k) \leq f(\boldsymbol{x}^\star)$, in fact

$$f_k(\boldsymbol{x}_k) \leq f_k(\boldsymbol{x}^\star) = f(\boldsymbol{x}^\star) + k\|\boldsymbol{h}(\boldsymbol{x}^\star)\|^2 + \alpha\|\boldsymbol{x}^\star - \boldsymbol{x}^\star\|^2 = f(\boldsymbol{x}^\star).$$

and by definition in Equation 6.4 we have

$$\begin{aligned} k_j\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\|^2 + \alpha\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\|^2 &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j}) \\ &\leq f(\boldsymbol{x}^\star) - \min_{\boldsymbol{x} \in B} f(\boldsymbol{x}) = C < +\infty \end{aligned} \tag{6.5}$$

so that

$$\lim_{j\to\infty} \|\boldsymbol{h}(\boldsymbol{x}_{k_j})\| = 0$$
$$\Downarrow$$
$$\left\|\boldsymbol{h}\left(\lim_{j\to\infty} \boldsymbol{x}_{k_j}\right)\right\| = \|\boldsymbol{h}(\bar{\boldsymbol{x}})\| = 0$$
$$\Downarrow$$
$$\boldsymbol{h}(\bar{\boldsymbol{x}}) = \boldsymbol{0}.$$

Step 2:

$\bar{\boldsymbol{x}} = \boldsymbol{x}^\star$.
From equation Equation 6.5, we have:

$$\begin{aligned}\alpha\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\|^2 &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j}) - k_j\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\|^2 \\ &\leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{k_j}).\end{aligned}$$

Taking the limit as j approaches infinity gives us:

$$\begin{aligned}\alpha\|\lim_{j\to\infty} \boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\|^2 &\leq \alpha\|\bar{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2 \\ &\leq f(\boldsymbol{x}^\star) - \lim_{j\to\infty} f(\boldsymbol{x}_{k_j}) \\ &\leq f(\boldsymbol{x}^\star) - f(\bar{\boldsymbol{x}}).\end{aligned}$$

Given that $\|\boldsymbol{h}(\bar{\boldsymbol{x}})\| = 0$, we can infer from equation Equation 6.2 that:

$$f(\boldsymbol{x}^\star) \leq f(\bar{\boldsymbol{x}}).$$

Thus, we can combine the inequalities:

$$\alpha\|\bar{\boldsymbol{x}} - \boldsymbol{x}^\star\|^2 \leq f(\boldsymbol{x}^\star) - f(\bar{\boldsymbol{x}}) \leq 0.$$

This implies that:

$$\bar{\boldsymbol{x}} = \boldsymbol{x}^\star.$$

Step 3:

Build multiplier.
Since $\boldsymbol{x}_{k_j}$ are *unconstrained local minima* for $f_{k_j}(\boldsymbol{x})$, it follows that:

$$\nabla f_{k_j}(\boldsymbol{x}_{k_j}) = \nabla f(\boldsymbol{x}_{k_j}) + k_j\nabla\|\boldsymbol{h}(\boldsymbol{x}_{k_j})\|^2 + \alpha\nabla\|\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star\|^2 = \boldsymbol{0}.$$

Recalling the gradient identities:

$$\begin{aligned}\nabla\|\boldsymbol{x}\|^2 &= \nabla(\boldsymbol{x}\cdot\boldsymbol{x}) = 2\boldsymbol{x}^T, \\ \nabla\|\boldsymbol{h}(\boldsymbol{x})\|^2 &= \nabla(\boldsymbol{h}(\boldsymbol{x})\cdot\boldsymbol{h}(\boldsymbol{x})) = 2\boldsymbol{h}(\boldsymbol{x})^T\nabla\boldsymbol{h}(\boldsymbol{x}),\end{aligned}$$

we can express the gradient condition as:

$$\nabla f(\boldsymbol{x}_{k_j})^T + 2k_j\nabla\boldsymbol{h}(\boldsymbol{x}_{k_j})^T\boldsymbol{h}(\boldsymbol{x}_{k_j}) + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star) = \boldsymbol{0}. \tag{6.6}$$

Next, we left-multiply by $\nabla\boldsymbol{h}(\boldsymbol{x}_{k_j})$:

$$\nabla\boldsymbol{h}(\boldsymbol{x}_{k_j})\left[\nabla f(\boldsymbol{x}_{k_j})^T + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star)\right] + 2k_j\nabla\boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla\boldsymbol{h}(\boldsymbol{x}_{k_j})^T\boldsymbol{h}(\boldsymbol{x}_{k_j}) = \boldsymbol{0}.$$

Since $\nabla \boldsymbol{h}(\boldsymbol{x}^\star) \in \mathbb{R}^{m\times n}$ is full rank for sufficiently large j, by continuity, $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})$ is also full rank. Therefore, the term $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T$ is nonsingular in $\mathbb{R}^{m\times m}$.
This leads us to:

$$2k_j \boldsymbol{h}(\boldsymbol{x}_{k_j}) = -\left(\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j}) \left[\nabla f(\boldsymbol{x}_{k_j})^T + 2\alpha(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star)\right].$$

Taking the limit as j approaches infinity, we obtain:

$$\lim_{j\to\infty} 2k_j \boldsymbol{h}(\boldsymbol{x}_{k_j}) = -\left(\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T\right)^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^\star)\nabla f(\boldsymbol{x}^\star)^T = -\boldsymbol{\lambda}. \tag{6.7}$$

Finally, substituting this limit back into equation Equation 6.6 along with equation Equation 6.7 yields:

$$\nabla f(\boldsymbol{x}^\star) - \nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{\lambda} = \boldsymbol{0}. \tag{6.8}$$

Step 4:

Build a special sequence of $\boldsymbol{z}_j$.
We need to construct a sequence $\boldsymbol{z}_j \to \boldsymbol{z}$ such that

$$\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}_j = \boldsymbol{0} \tag{6.9}$$

for all j. To achieve this, we define the sequence $\boldsymbol{z}_j$ as the projection of $\boldsymbol{z}$ onto the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})$. This is given by:

$$\boldsymbol{z}_j = \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}.$$

To verify that this construction satisfies our requirement, we compute:

$$\begin{aligned}\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}_j &= \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} \\ &= \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} = \boldsymbol{0}.\end{aligned}$$

Next, we consider the limit as j approaches infinity:

$$\begin{aligned}\lim_{j\to\infty} \boldsymbol{z}_j &= \boldsymbol{z} - \lim_{j\to\infty} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z} \\ &= \boldsymbol{z} - \nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T \left[\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\nabla \boldsymbol{h}(\boldsymbol{x}^\star)^T\right]^{-1} \nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{z}.\end{aligned}$$

Thus, if $\boldsymbol{z}$ lies in the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$, i.e., $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\boldsymbol{z} = 0$ then the limit

$$\lim_{j\to\infty} \boldsymbol{z}_j = \boldsymbol{z}.$$

exists.

Step 5:

Necessary conditions.
Since $\boldsymbol{x}_{k_j}$ are *unconstrained local minima* for $f_{k_j}(\boldsymbol{x})$, it follows that the matrices $\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})$ are semi-positive definite. This means:

$$\boldsymbol{z}^T \nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})\boldsymbol{z} \geq 0, \quad \forall \boldsymbol{z} \in \mathbb{R}^n.$$

Moreover, we can express the Hessian as:

$$\begin{aligned}\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j}) &= \nabla^2 f(\boldsymbol{x}_{k_j}) + k_j \nabla^2 \|\boldsymbol{h}(\boldsymbol{x}_{k_j})\|^2 + 2\alpha\nabla(\boldsymbol{x}_{k_j} - \boldsymbol{x}^\star) \\ &= \nabla^2 f(\boldsymbol{x}_{k_j}) + k_j \nabla^2 \sum_{i=1}^{m} h_i(\boldsymbol{x}_{k_j})^2 + 2\alpha \boldsymbol{I}.\end{aligned} \tag{6.10}$$

Using the identity for the Hessian of the squared function:

$$\nabla^2 h_i(\boldsymbol{x})^2 = \nabla(2h_i(\boldsymbol{x})\nabla h_i(\boldsymbol{x})^T) = 2\nabla h_i(\boldsymbol{x})^T \nabla h_i(\boldsymbol{x}) + 2h_i(\boldsymbol{x})\nabla^2 h_i(\boldsymbol{x}),$$

we can substitute into equation Equation 6.10 to obtain:

$$\begin{aligned}\nabla^2 f_{k_j}(\boldsymbol{x}_{k_j}) = \nabla^2 f(\boldsymbol{x}_{k_j}) &+ 2k_j \sum_{i=1}^{m} \nabla h_i(\boldsymbol{x}_{k_j})^T \nabla h_i(\boldsymbol{x}_{k_j}) \\ &+ 2k_j \sum_{i=1}^{m} h_i(\boldsymbol{x}_{k_j}) \nabla^2 h_i(\boldsymbol{x}_{k_j}) + 2\alpha \boldsymbol{I}.\end{aligned} \tag{6.11}$$

Let $\boldsymbol{z} \in \mathbb{R}^n$, since $0 \leq \boldsymbol{z}^T \nabla^2 f_{k_j}(\boldsymbol{x}_{k_j})\boldsymbol{z}$, we can write:

$$\begin{aligned}0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}_{k_j})\boldsymbol{z} &+ \sum_{i=1}^{m} (2k_j h_i(\boldsymbol{x}_{k_j})) \boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}_{k_j})\boldsymbol{z} \\ &+ 2k_j \|\nabla \boldsymbol{h}(\boldsymbol{x}_{k_j})\boldsymbol{z}\|^2 + 2\alpha\|\boldsymbol{z}\|^2.\end{aligned}$$

This inequality holds for all $\boldsymbol{z} \in \mathbb{R}^n$, and specifically for any $\boldsymbol{z}_j$ of step 4 that salisfy Equation 6.9:

$$0 \leq \boldsymbol{z}_j^T \nabla^2 f(\boldsymbol{x}_{k_j})\boldsymbol{z}_j + \sum_{i=1}^{m} (2k_j h_i(\boldsymbol{x}_{k_j})) \boldsymbol{z}_j^T \nabla^2 h_i(\boldsymbol{x}_{k_j})\boldsymbol{z}_j + 2\alpha\|\boldsymbol{z}_j\|^2.$$

Taking the limit as j approaches infinity, we apply Equation 6.7:

$$0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{z} + \sum_{i=1}^{m} \lambda_i \boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}^\star)\boldsymbol{z} + 2\alpha\|\boldsymbol{z}\|^2.$$

Since $\alpha > 0$ can be chosen arbitrarily, it follows that:

$$0 \leq \boldsymbol{z}^T \nabla^2 f(\boldsymbol{x}^\star)\boldsymbol{z} - \sum_{i=1}^{m} \lambda_i \left[\boldsymbol{z}^T \nabla^2 h_i(\boldsymbol{x}^\star)\boldsymbol{z}\right].$$

Step 6:

Sufficient conditions.
Using Equation 6.8 and sufficient condition (C) for all $\boldsymbol{z} \in \ker \nabla \boldsymbol{h}(\boldsymbol{x}^\star)$

$$\begin{aligned}\mathscr{L}(\boldsymbol{x}^\star + \boldsymbol{z}, \boldsymbol{\lambda}) &= \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) + \cancel{\nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\boldsymbol{z}} + \frac{1}{2}\boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\boldsymbol{z} + o\left(\|\boldsymbol{z}\|^2\right) \\ &= \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) + \frac{1}{2}\boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\boldsymbol{z} + o\left(\|\boldsymbol{z}\|^2\right) \\ &\geq \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) + \frac{\varepsilon}{2}\|\boldsymbol{z}\|^2 + o\left(\|\boldsymbol{z}\|^2\right) \\ &\geq \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}) + \frac{\varepsilon}{4}\|\boldsymbol{z}\|^2, \qquad \text{for small } \boldsymbol{z} \in \ker \nabla \boldsymbol{h}(\boldsymbol{x}^\star)\end{aligned} \tag{6.12}$$

moreover

$$
\begin{aligned}
f(\delta \boldsymbol{x}+\boldsymbol{x}^\star) &= f(\delta \boldsymbol{x}+\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k h_k(\delta \boldsymbol{x}+\boldsymbol{x}^\star) \\
&= f(\boldsymbol{x}^\star) + \nabla f(\boldsymbol{x}^\star)\delta \boldsymbol{x} + \frac{1}{2}\delta \boldsymbol{x}^T \nabla^2 f(\boldsymbol{x}^\star)\delta \boldsymbol{x} \\
&- \sum_{k=1}^{m} \lambda_k \left(\cancel{h_i(\boldsymbol{x}^\star)} + \nabla h_i(\boldsymbol{x}^\star)\delta \boldsymbol{x} + \frac{1}{2}\delta \boldsymbol{x}^T \nabla^2 h_i(\boldsymbol{x}^\star) \right) + o(\|\delta \boldsymbol{x}\|^2) \\
&= f(\boldsymbol{x}^\star) + \cancel{\nabla \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x}} + \frac{1}{2}\delta \boldsymbol{x}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} + o(\|\delta \boldsymbol{x}\|^2)
\end{aligned}
$$

if $\delta \boldsymbol{x} = \boldsymbol{z} + o(\boldsymbol{z})$ then
Using Equation 6.12 for all $\boldsymbol{z} \in \ker \nabla \boldsymbol{h}(\boldsymbol{x}^\star)$

$$
\begin{aligned}
f(\delta \boldsymbol{x}+\boldsymbol{x}^\star) &= f(\boldsymbol{x}^\star) + \frac{1}{2}\delta \boldsymbol{x}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} + o(\|\delta \boldsymbol{x}\|^2) \\
&\geq f(\boldsymbol{x}^\star) - \frac{1}{2}\left| \delta \boldsymbol{x}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} - \boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{z}^\star, \boldsymbol{\lambda})\boldsymbol{z} \right| \\
&+ \frac{1}{2}\boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{z}^\star, \boldsymbol{\lambda})\boldsymbol{z} + o(\|\delta \boldsymbol{x}\|^2) \\
&\geq f(\boldsymbol{x}^\star) - \frac{1}{2}E(\boldsymbol{z}, \delta \boldsymbol{x}) + \frac{\varepsilon}{4}\|\boldsymbol{z}\|^2 + o(\|\delta \boldsymbol{x}\|^2)
\end{aligned}
\tag{6.13}
$$

where

$$
\begin{aligned}
E(\boldsymbol{z}, \delta \boldsymbol{x}) &= \left| \delta \boldsymbol{x}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} - \boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{z}^\star, \boldsymbol{\lambda})\boldsymbol{z} \right| \\
&\leq \left| \delta \boldsymbol{x}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} - \boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{z}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} \right| \\
&+ \left| \boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda})\delta \boldsymbol{x} - \boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{z}^\star, \boldsymbol{\lambda})\boldsymbol{z} \right| \\
&\leq C\|\delta \boldsymbol{x} - \boldsymbol{z}\| \cdot (\|\boldsymbol{z}\| + \|\delta \boldsymbol{x}\|)
\end{aligned}
$$

if $\delta \boldsymbol{x} = \boldsymbol{z} + o(\|\boldsymbol{z}\|)$ then $o(\|\delta \boldsymbol{x}\|) = o(\|\boldsymbol{z}\|)$ and

$$
E(\boldsymbol{z}, \delta \boldsymbol{x}) \leq C\, o(\|\boldsymbol{z}\|) \cdot (\|\boldsymbol{z}\| + \|\boldsymbol{z} + o(\boldsymbol{z})\|) = o(\|\boldsymbol{z}\|^2)
$$

used in Equation 6.13 for small $\delta \boldsymbol{x}$

$$
\begin{aligned}
f(\delta \boldsymbol{x}+\boldsymbol{x}^\star) &\geq f(\boldsymbol{x}^\star) - o(\|\delta \boldsymbol{x}\|^2) + \frac{\varepsilon}{4}\|\boldsymbol{z}\|^2 + o(\|\delta \boldsymbol{x}\|^2) \\
&= f(\boldsymbol{x}^\star) + \frac{\varepsilon}{4}\|\delta \boldsymbol{x}\|^2 + o(\|\delta \boldsymbol{x}\|^2) \\
&\geq f(\boldsymbol{x}^\star) + \frac{\varepsilon}{8}\|\delta \boldsymbol{x}\|^2 \qquad \text{for small } \delta \boldsymbol{x}
\end{aligned}
$$

and thus $\boldsymbol{x}^\star$ is a strict local minimum.
To prove $\delta \boldsymbol{x} = \boldsymbol{z} + o(\|\boldsymbol{z}\|)$
TO BE FINISHED...
This establishes all the relation we aimed to prove. □



7. Sylvester theorem

Sylvester's theorem provides an operational criterion to determine whether a matrix is symmetric positive definite (SPD). We now state Sylvester's theorem:

Theorem 7.1 — Sylvester's Criterion. A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is **positive** definite if and only if all its leading principal minors are positive. That is, $\boldsymbol{A}$ is positive definite if and only if:

$$\det(\boldsymbol{A}_k) > 0 \quad \text{for} \quad k = 1, 2, \ldots, n$$

where $\boldsymbol{A}_k$ is the upper-left $k \times k$ submatrix (leading principal minor) of $\boldsymbol{A}$.

Proof.

(1) Necessity (if A is positive definite, then all leading principal minors are positive): Assume $\boldsymbol{A}$ is a symmetric positive definite matrix. This means that for any nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$, we have:

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$$

Consider a sequence of nested subspaces corresponding to the leading principal minors of $\boldsymbol{A}$. Let $\boldsymbol{A}_k$ denote the upper-left $k \times k$ submatrix of $\boldsymbol{A}$. For each k, $\boldsymbol{A}_k$ must be symmetric and positive definite because any submatrix corresponding to a smaller subspace inherits positive definiteness.
To verify that each leading principal minor is positive, consider $\boldsymbol{x}_k = (x_1, x_2, \ldots, x_k)^\top \in \mathbb{R}^k$. For $\boldsymbol{A}_k$, we have:

$$\boldsymbol{x}_k^\top \boldsymbol{A}_k \boldsymbol{x}_k > 0 \quad \forall \boldsymbol{x}_k \in \mathbb{R}^k, \boldsymbol{x}_k \neq 0$$

This implies that the determinant of each $\boldsymbol{A}_k$, which represents the product of the eigenvalues of $\boldsymbol{A}_k$, must be positive. Hence, $\det(\boldsymbol{A}_k) > 0$ for all $k = 1, 2, \ldots, n$.
(2) Sufficiency (if all leading principal minors are positive, then A is positive definite):

- **Base case:** For $n = 1$, $\boldsymbol{A}_1 = [a_{11}]$. If $\det(\boldsymbol{A}_1) > 0$, then $a_{11} > 0$, and $\boldsymbol{A}_1$ is positive definite.
- **Induction Step:**
  Now, assume that all leading principal minors of $\boldsymbol{A}$ are positive:

  $$\det(\boldsymbol{A}_k) > 0 \quad \text{for} \quad k = 1, 2, \ldots, n$$

  We aim to show that $\boldsymbol{A}$ is positive definite. Let us use induction on the dimension n.

– **Block Matrix Decomposition:**
We decompose the matrix $\boldsymbol{A}_k$ into block form. The matrix $\boldsymbol{A}_k$ can be written as:

$$\boldsymbol{A}_k = \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{bmatrix}$$

where:
- $\boldsymbol{A}_{k-1}$ is the upper-left $(k-1)\times(k-1)$ submatrix,
- $\boldsymbol{b} \in \mathbb{R}^{k-1}$ is a column vector, and
- $c \in \mathbb{R}$ is a scalar, corresponding to the k-th diagonal element of $\boldsymbol{A}_k$.

– **Factorize $\boldsymbol{A}$**

$$\begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{b}^\top \boldsymbol{A}_{k-1}^{-\top} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & d \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}_{k-1}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^\top & 1 \end{bmatrix} \tag{7.1}$$

where $d = c - \boldsymbol{b}^\top \boldsymbol{A}_{k-1}\boldsymbol{b}$

– **Positivity of d:**
From $\det(\boldsymbol{A}_k) > 0$ and $\det(\boldsymbol{A}_{k-1}) > 0$ using the product of determinant, from Equation 7.1

$$\det(\boldsymbol{A}_{k-1})d = \det(\boldsymbol{A}_k) > 0, \quad \Longrightarrow \quad d > 0$$

– **SPD of $\boldsymbol{A}_k$:**
From Equation 7.1 for $\boldsymbol{x} \neq \boldsymbol{0}$

$$\begin{aligned}
\boldsymbol{x}^\top \boldsymbol{A}_k \boldsymbol{x} &= \begin{bmatrix} \boldsymbol{y} & z \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ z \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{y}^\top & z \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{b}^\top \boldsymbol{A}_{k-1}^{-\top} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & d \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}_{k-1}^{-1}\boldsymbol{b} \\ \boldsymbol{0}^\top & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ z \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{y}^\top + z\boldsymbol{b}^\top \boldsymbol{A}_{k-1}^{-\top} & z \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & d \end{bmatrix} \begin{bmatrix} \boldsymbol{y} + z\boldsymbol{A}_{k-1}^{-1}\boldsymbol{b} \\ z \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{w}^\top & z \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{k-1} & \boldsymbol{0} \\ \boldsymbol{0}^\top & d \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ z \end{bmatrix} \\
&= \boldsymbol{w}^\top \boldsymbol{A}_{k-1}\boldsymbol{w} + dz^2 > 0
\end{aligned}$$

where $\boldsymbol{w} = \boldsymbol{y} + z\boldsymbol{A}_{k-1}^{-1}\boldsymbol{b}$ and $[\boldsymbol{w}^\top, z] \neq \boldsymbol{0}^\top$.
this conlude the proof. □

7.1 Why use sylvester theorem

The tedious and error prone operation of finding the minimum with the Lagrangian multiplier is the one that's performed to determine if the matrix is (semi) positive defined in the kernel

$$\ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$$

of the gradient of the constraints map. In fact for every stationary point $\boldsymbol{x}^\star$ of the lagrangian we have to check that

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)\, \boldsymbol{z} \geq 0 \qquad \forall \boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$$

For sake of simplicity from now we will denote the matrix $\nabla^2_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star)$ as $\boldsymbol{A}$. We can note that the vector $\boldsymbol{z} \in \ker\{\nabla \boldsymbol{h}(\boldsymbol{x}^\star)\}$ (and from now on we refer to the matrix $\nabla \boldsymbol{h}$ as $\boldsymbol{B}$) can be expressed as a linear combination of the vectors $\boldsymbol{k}_i$ (that are representing the base of $\boldsymbol{B}$) in the way

$$\boldsymbol{z} = \boldsymbol{k}_1\alpha_1 + \boldsymbol{k}_2\alpha_2 + \cdots + \boldsymbol{k}_p\alpha_p = \boldsymbol{K}\boldsymbol{\alpha} \qquad \boldsymbol{\alpha} \in \mathbb{R}^p$$

We can see that this expression can be reduced to a multiplication of a matrix $K \in \mathbb{R}^{n\times p}$ (whose columns are the vector $\boldsymbol{k}_i$ of the kernel base) and a p-dimensional vector $\boldsymbol{\alpha}$ (where p is the number of constraints in the map $\boldsymbol{h}(\boldsymbol{x})$).

We can now rewrite the second order necessary condition considering that

$$\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} = \boldsymbol{\alpha}^T\boldsymbol{K}^T\boldsymbol{A}\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\boldsymbol{C}\boldsymbol{\alpha}, \quad \boldsymbol{C} := \boldsymbol{K}^T\boldsymbol{A}\boldsymbol{K} \in \mathbb{R}^{p\times p}$$

■ **Example 7.1** Let's consider the numerical problem when the values of the matrix

$$\boldsymbol{A} = \nabla_{\boldsymbol{x}}\mathscr{L} \quad \text{and} \quad \boldsymbol{B} = \nabla\boldsymbol{h}$$

are given with values

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

The first thing now is to compute the manually compute the kernel of the matrix $\boldsymbol{B}$ in $\mathbb{R}^3$ solving the linear system

$$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \boldsymbol{0} \quad \Longrightarrow \quad \begin{cases} z_1 = 0 \\ z_2 = \alpha \\ z_3 = \beta \end{cases} \qquad \alpha, \beta \in \mathbb{R}$$

At this point we can rewrite the kernel of B using the linear combination of the vector composing the base:

$$\ker\{\boldsymbol{B}\} = \begin{bmatrix} 0 \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \beta = \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{=\boldsymbol{K}} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

The last thing is now to compute the matrix C that should be analyzed to know if it's (semi) positive defined or not:

$$\boldsymbol{K}^T\boldsymbol{A}\boldsymbol{K} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

■

This definition reduces the complexity of the problem at analyzing the matrix $\boldsymbol{C}$ (that's smaller than the original matrix A) and determining if that particular matrix is (semi) positive defined using two methods:

1. Considering the fact that $\boldsymbol{C}$ is a symmetric matrix we know for sure that it can be diagonalised

$$\boldsymbol{C} = \boldsymbol{T}^T\boldsymbol{\Lambda}\boldsymbol{T}$$

(where $\boldsymbol{\Lambda}$ is a diagonal matrix containing all the eigenvalues of $\boldsymbol{C}$); considering the expression

$$\boldsymbol{\alpha}^T\boldsymbol{C}\boldsymbol{\alpha} = \boldsymbol{\alpha}^T\boldsymbol{T}^T\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{\alpha}$$

in order to have a semi positive defined matrix all the eigenvalue λ_i must be positive or at least equals to zero. If it happens that

$$\lambda_i > 0 \qquad \forall i$$

then the matrix is positive defined and the point is a local minimum;

2. One other solution is to use the **Sylvester theorem** that states:
a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is **positive defined** if and only if all the **principal minors** of $\boldsymbol{A}$ are strictly **positive**
(note that if one minor is equal to zero, no information can be retrieved with this method).

■ **Example 7.2 — application of the Sylvester theorem.** Let's consider now the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$$

In order to determine if $\boldsymbol{A}$ is positive defined we have to compute all the principle minors starting from the first $\boldsymbol{A}_1$ that's represented only by the element a_{11} of $\boldsymbol{A}$ and so

$$\boldsymbol{A}_1 = \det[1] = 1 > 0$$

The second (and last) principle minor of $\boldsymbol{A}$ is the determinant of the whole matrix and it happens that

$$\boldsymbol{A}_2 = \det \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix} = 1 > 0$$

Having all minors greater then zero, this means that the matrix $\boldsymbol{A}$ is positive defined.

The same result can be achieved by determining all the eigenvalues of the matrix and showing they are all positive; we can in fact see that the characteristic polynomial of $\boldsymbol{A}$ is equal to

$$\begin{aligned} p(\lambda) &= \det\left[\boldsymbol{A} - \lambda \boldsymbol{I}\right] \\ &= \det \begin{bmatrix} 1-\lambda & 2 \\ 2 & 5-\lambda \end{bmatrix} \\ &= \lambda^2 - 6\lambda + 1 \end{aligned}$$

and thus

$$\lambda_{1,2} = \frac{6 \pm \sqrt{36-4}}{2} = 3 \pm \frac{\sqrt{32}}{2}$$

nothing that

$$\frac{\sqrt{32}}{2} < \frac{\sqrt{36}}{2} = 3$$

then all the eigenvalues λ_1, λ_2 are strictly positive and so $\boldsymbol{A}$ is positive defined.

■

7.1.1 Trick for semi positive matrices

The Sylvester theorem gives no hint to determine if a matrix is semi-positive defined when a minor is equal to zero. However a way to determine if the matrix is semi-positive defined is by considering that the matrix

$$\boldsymbol{A} + \varepsilon \boldsymbol{I}$$

should be positive defined for values of ε approaching zero from the positive direction (for $\varepsilon \to 0^+$).

■ **Example 7.3 — Trick for usage of the Sylvester theorem on semi-positive matrrices.** Let's consider now the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$$

The first minor $\boldsymbol{A}_1 = \det[1] = 1$ is positive, and so we have to compute the second principal minor noting that it's equal to zero:

$$\boldsymbol{A}_2 = \det \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} = 4 - 2 \cdot 2 = 0$$

This relation gives no clue on determining if $\boldsymbol{A}$ is positive defined or not, but considering the trick yet defined we can compute

$$\det \begin{bmatrix} 1+\varepsilon & 2 \\ 2 & 4+\varepsilon \end{bmatrix} = (1+\varepsilon)(4+\varepsilon) - 4 = 5\varepsilon$$

We can see that this expression, for $\varepsilon \to 0^+$, is strictly greater then zero and this might conclude our analyses stating that $\boldsymbol{A}$ is semi positive defined.

The same result can be defined by computing the eigenvalues of $\boldsymbol{A}$ and so calculating the roots of the polynomial

$$\begin{aligned} p(\lambda) &= \det \begin{bmatrix} 1-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} \\ &= (1-\lambda)(4-\lambda) - 4 \\ &= \lambda^2 - 5\lambda \qquad \Longrightarrow \qquad \lambda_1 = 0,\ \lambda_2 = 5 \end{aligned}$$

Having one eigenvalues zero (and the other positive) determines that $\boldsymbol{A}$ is semi positive defined.

■

■ **Example 7.4 — Counter example of the Sylvester theorem.** Let's consider now the matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

The first minor $\boldsymbol{A}_1 = \det[1] = 1$ is positive, while the second one is zero, in fact

$$\boldsymbol{A}_2 = \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = 0$$

This result gives no information but we can use the trick by considering the determinant

$$\det \begin{bmatrix} 1+\varepsilon & 1 \\ 1 & 1+\varepsilon \end{bmatrix} = (1+\varepsilon)^2 - 1 = \varepsilon^2 + 2\varepsilon > 0$$

for $\varepsilon > 0$.

Being $\boldsymbol{A}$ a 3×3 matrix we also need to compute the third minor that leads to another zero:

$$\begin{aligned} \boldsymbol{A}_3 &= \det \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \\ &= 1 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} - 1 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + 0 \det \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= 0 \end{aligned}$$

We can use another time the Sylvester theorem trick considering the determinant

$$\begin{aligned} p(\varepsilon) &= \det \begin{bmatrix} 1+\varepsilon & 1 & 1 \\ 1 & 1+\varepsilon & 1 \\ 1 & 1 & \varepsilon \end{bmatrix} \\ &= (1+\varepsilon)\left(\varepsilon^2+\varepsilon-1\right)-(\varepsilon-1)-\varepsilon \\ &= \varepsilon\left(\varepsilon^2+2\varepsilon-2\right) \end{aligned}$$

Having that the derivative

$$p'(\varepsilon) = 3\varepsilon^2+4\varepsilon-2$$

is negative for ε approaching zero, then we can conclude that $\boldsymbol{A}$ is **not** semi positive defined. The same result can be confirmed calculating the eigenvalues of the matrix and so

$$p(\lambda) = \det \begin{bmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & -\lambda \end{bmatrix} = -\lambda\left(\lambda^2-2\lambda-2\right)$$

$$\implies \qquad \lambda_1 = 0 \qquad \lambda_{2,3} = \frac{2\pm\sqrt{12}}{2} = 1\pm\sqrt{3}$$

Noting that $\lambda_3 = 1-\sqrt{3} < 0$ is negative, than the matrix $\boldsymbol{A}$ is for sure **not** semi positive defined.

■

This trick can be resumend in the following Lemma

Lemma 7.1 — Sylvester's Trick. A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is **semi-positive definite** if and only if all its leading principal minors $\boldsymbol{A}_k$ satisfy the following conditions:

- $\det(\boldsymbol{A}_k) > 0$, or
- $\det(\boldsymbol{A}_k) = 0$ and the derivative of the perturbed determinant at $\varepsilon = 0$ is positive:

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\det(\boldsymbol{A}_k+\varepsilon\boldsymbol{I})\Big|_{\varepsilon=0} > 0$$



8. Inequality constraints

Let's introduce now the problem of minimizing a function $f(\boldsymbol{x}) : \mathbb{R}^n \to R$ considering a set of p inequalities constrains in the form $g_k(\boldsymbol{x}) \geq 0$.

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) \\ &\text{subject to:} && g_k(\boldsymbol{x}) \geq 0 \qquad k = 1, \dots, p \end{aligned}$$

The first approach to this problem is by trying to transform the inequality constraints g_k into equality one (so having $h_k(\boldsymbol{x}) = 0$). This can be done considering that each constraint can be expressed as

$$g_k(\boldsymbol{x}) = \varepsilon_k^2 \geq 0 \quad \Longrightarrow \quad h_k(\boldsymbol{x}, \varepsilon_k) = g_k(\boldsymbol{x}) - \varepsilon_k^2 = 0$$

Doing this process for each constraint we can determine a minimization problem that depends both on the function variable $\boldsymbol{x}$ but also on the so called **slack variables** $\boldsymbol{\varepsilon}$:

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) \\ &\text{subject to:} && h_k(\boldsymbol{x}, \boldsymbol{\varepsilon}) = 0 \qquad k = 1, \dots, p \end{aligned}$$

This kind of problem increases the computational costs because it increases the variable to minimize from n to $n + p$.

> **Note**
>
> We used the expression ε_k^2 and not ε_k because with this expression we don't need to specify one more inequality $\varepsilon_k \geq 0$.

■ **Example 8.1** Let's consider the following problem:

$$\begin{aligned} &\text{minimize:} && f(x,y) = x^2 \\ &\text{subject to:} && x^2 + y^2 \leq 1 \\ & && x + y \geq 0 \end{aligned}$$

In order to solve this problem we have to reduce the inequality constraints to equality ones; considering the first constraint, that has to be expressed in the form $g_1(x,y) \geq 0$ and so

$$g_1(x,y) = 1 - x^2 - y^2 \geq 0$$
$$\Downarrow$$
$$h_1(x,y,\varepsilon_1) = 1 - x^2 - y^2 - \varepsilon_1^2$$

At the same way it's possible to express the second inequality as the a constraint

$$h_2(x,y,\varepsilon_2) = x + y - \varepsilon_2^2$$

With this expression being set the problem reduces to the form

$$\begin{aligned} \text{minimize:} \quad & f(x,y,\varepsilon_1,\varepsilon_2) = x^2 \\ \text{subject to:} \quad & h_1(x,y,\varepsilon_1,\varepsilon_2) = 1 - x^2 - y^2 - \varepsilon_1^2 = 0 \\ & h_2(x,y,\varepsilon_1,\varepsilon_2) = x + y - \varepsilon_2^2 = 0 \end{aligned}$$

With the problem this stated we can build the lagrangian that depend's on both the original variables x, y but also the slack variables $\varepsilon_1, \varepsilon_2$ and so

$$\begin{aligned} \mathscr{L}(\underbrace{x,y,\varepsilon_1,\varepsilon_2}_{\boldsymbol{x}}, \underbrace{\lambda_1,\lambda_2}_{\boldsymbol{\lambda}}) &= f(\boldsymbol{x}) - \lambda_1 h_1(\boldsymbol{x}) - \lambda_2 h_2(\boldsymbol{x}) \\ &= x^2 - \lambda_1\left(1 - x^2 - y^2 - \varepsilon_1^2\right) - \lambda_2\left(x + y - \varepsilon_2^2\right) \end{aligned}$$

To find the local minimum point we can start using the first necessary condition of the lagrangian multiplier, so by determining all the stationary point such that $\nabla\mathscr{L} = \mathbf{0}$; in practise this means solving the following non linear system of 6 equations in 6 variables:

$$\begin{cases} \partial/\partial_x \mathscr{L}: & 2x - 2\lambda_1 x - \lambda_2 = 0 \\ \partial/\partial_y \mathscr{L}: & 2\lambda_1 y - \lambda_2 = 0 \\ \partial/\partial_{\varepsilon_1} \mathscr{L}: & 2\lambda_1 \varepsilon_1 = 0 \\ \partial/\partial_{\varepsilon_2} \mathscr{L}: & 2\lambda_2 \varepsilon_2 = 0 \\ \partial/\partial_{\lambda_1} \mathscr{L}: & x^2 + y^2 + \varepsilon_1^2 - 1 = 0 \\ \partial/\partial_{\lambda_2} \mathscr{L}: & x + y - \varepsilon_2^2 = 0 \end{cases}$$

Solving minimization problems with inequality constraints can be tricky using the lagrangian multiplier method and so other algorithm (like the KKT that's going to be explained) are usually preferred. However this kind of system can be solved by cases; considering in fact the second and third equations we can discriminate the solution considering the various combination when ε_1, ε_2 are equal (or not) to zero.

Considering for example the case $\varepsilon_1 = \varepsilon_2 = 0$ the non linear system reduces to the form

$$\begin{cases} 2x - 2\lambda_1 x - \lambda_2 = 0 \\ 2\lambda_1 y - \lambda_2 = 0 \\ x^2 + y^2 - 1 = 0 \\ x + y = 0 \end{cases}$$

Considering that the first two expressions are linear in respect to λ_1, λ_2 it's possible to determine their value directly in terms of x, y:

$$\lambda_1 = \frac{x}{y - x} \qquad \lambda_2 = 2\frac{xy}{y - x}$$

Considering instead the last 2 equations we have a non linear system that can be also computed by cases; from the last equation we have in fact that $y = -x$ and so

$$1-x^2-x^2=0 \qquad \Longrightarrow \qquad x=\pm\frac{1}{\sqrt{2}}$$

At this end of this process we defined two stationary points for the lagrangian:

$$(x,y,\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2)=\left(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}},0,0,-\frac{1}{2},\frac{\sqrt{2}}{2}\right),$$
$$(x,y,\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2)=\left(-\frac{1}{\sqrt{2}},\frac{1}{\sqrt{2}},0,0,\frac{1}{2},-\frac{\sqrt{2}}{2}\right).$$

Solving symbolically this system of equation with a computer the lonely real solution of the problem (that determines a local minimum) is expressed in the form

$$(x,y,\varepsilon_1,\varepsilon_2,\lambda_1,\lambda_2)=\left(0,\sqrt{-\varepsilon_1^2+1},\varepsilon_1,\sqrt[4]{1-\varepsilon_1^2},1,1\right)$$

In order for the square roots to exists the value ε_1^2 should fit inside the range $[0,1]$ and so the local minimum of the problem can be found on the vertical line

$$(x,y)=(0,k) \qquad k\in[0,1]$$

■

8.0.1 Karush-Kuhn-Tucker conditions

The **Karush-Kuhn-Tucker** (KKT) conditions are a set of necessary and sufficient conditions (first and second order) that allows to describe if a point $\boldsymbol{x}^\star$ is a minimum point of a minimization problem subjected to equality and inequality constraints, so for the problem

$$\begin{array}{lll} \text{minimize:} & f(\boldsymbol{x}) & \\ \text{subject to:} & h_k(\boldsymbol{x})=0 & k=1,\dots,m \\ & g_k(\boldsymbol{x})\geq 0 & k=1,\dots,p \end{array}$$

where $f,h_k,g_k:\mathbb{R}^n\to\mathbb{R}$ are real evaluated function.
In order to describe this conditions it's important to define *set of the active constraints** $\mathscr{A}$ as

$$\mathscr{A}(\boldsymbol{x}^\star)=\{k\mid g_k(\boldsymbol{x}^\star)=0\}$$

This means that given the constraints $g\in C(\mathbb{R}^n)$ it's defined as *active* if it happens that the point $\boldsymbol{x}^\star$ is such that $g(\boldsymbol{x}^\star)=0$, so in practical way it's in the *border* between activating and not activating the constraints.

8.0.1.1 Qualified constraints

Given $\boldsymbol{g}\in C^1(\mathbb{R}^p,\mathbb{R}^n)$ and $\boldsymbol{h}\in C^1(\mathbb{R}^m,\mathbb{R}^n)$ respectively the inequality and equality constraints map, then a point $\boldsymbol{x}^\star$ is defined **qualified** if the gradients of the active constraints (and the equality ones) are linearly independent, and so

$$\{\nabla g_k(\boldsymbol{x}^\star)\ :\ k\in\mathscr{A}(\boldsymbol{x}^\star)\}\bigcup\{\nabla h_1(\boldsymbol{x}^\star),\dots,\nabla h_m(\boldsymbol{x}^\star)\}$$

are linearly independent

■ **Example 8.2 — qualification of a point.** Considering the minimization problem started in example Example 6.2 stated as

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h_1(x,y,z) = x^2 + y^2 - 2 = 0 \\ & h_2(x,y,z) = x + z - 1 = 0 \end{aligned}$$

and considering the candidate point of minimum (resulting from the solution of the non-linear system)

$$\begin{aligned} x^\star &= \frac{\sqrt{3}}{2} - \frac{1}{2} \\ y^\star &= \frac{1}{2} + \frac{\sqrt{3}}{2} \\ z^\star &= \frac{3}{2} - \frac{\sqrt{3}}{2} \\ \lambda_1^\star &= \frac{1}{2} - \frac{\sqrt{3}}{2} \\ \lambda_2^\star &= 3 - \sqrt{3} \end{aligned}$$

In order to check if the chosen point is qualified it's necessary to check if the gradients of the equality constraint and the active ones (that in this case are not present because there aren't inequality constraints) are linearly independent. In order to solve this problem firstly we need to build the matrix H represent the gradients of the constraints:

$$H(x,y,z) = \begin{bmatrix} \nabla h_1(x,y,z) \\ \nabla h_2(x,y,z) \end{bmatrix} = \begin{bmatrix} 2x & 2y & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\implies \qquad H^\star = H(x^\star, y^\star, z^\star) = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

It's straightforward to see that the lines of the matrix $H^\star$ are linearly independent, considering in fact the rank of $H^\star$ by analysing the columns we see that's equal to 2, so equalizing the number of the constraints.

∎

8.0.1.2 KKT first order necessary condition

Let $f \in C^1(\mathbb{R}^n)$ a function to minimize subjected to the inequality

$$\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$$

and equality

$$\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$$

constraint maps. If $\boldsymbol{x}^\star$ satisfy constraint qualification then necessary condition for $\boldsymbol{x}^\star$ to be a local minima is that that exists $m+p$ scalars λ_i, μ_j such that all the following conditions are satisfied:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathscr{L}\left(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star\right) &= \mathbf{0} \\ h_k\left(\boldsymbol{x}^\star\right) &= 0 \qquad k = 0, \dots, m \\ g_k\left(\boldsymbol{x}^\star\right) &\geq 0 \qquad k = 0, \dots, p \\ \mu_k^\star g_k\left(\boldsymbol{x}^\star\right) &= 0 \qquad k = 0, \dots, p \\ \mu_k^\star &\geq 0 \qquad k = 0, \dots, p \end{aligned} \tag{8.1}$$

where the lagrangian is in this case defined as

$$\mathscr{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k g_k(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x})$$

■ **Example 8.3** Solving minimization problem with the KKT conditions is generally more easier; considering the case of the minimization problem in Example 8.1 the lagrangian will depend only on 4 parameter (instead of 6 as in that example):

$$\mathscr{L}(x,y,\mu_1,\mu_2) = x^2 - \mu_1\left(1-x^2-y^2\right) - \mu_2\left(x+y\right)$$

Using the first order KKT condition the main non linear system to solve is in the form

$$\begin{cases} \left.\begin{aligned} 2x + 2\mu_1 x - \mu_2 &= 0 \\ 2\mu_1 y - \mu_2 &= 0 \end{aligned}\right\} & \nabla_{\boldsymbol{x}}\mathscr{L}\left(\boldsymbol{x},\boldsymbol{\lambda}\right) \\ \left.\begin{aligned} \mu_1\left(1-x^2-y^2\right) &= 0 \\ \mu_2\left(x+y\right) &= 0 \end{aligned}\right\} & \mu_k g_k\left(\boldsymbol{x}\right) = 0 \\ \mu_1,\mu_2 \geq 0 \end{cases}$$

This system of four non linear equations (the conditions $\mu_1,\mu_2 \geq 0$ are *easier* to consider) can be solved by cases:

1. If $\mu_1 = \mu_2 = 0$ then the last two equalities are always verified, while the first two becomes

$$\begin{cases} 2x = 0 \\ 0 = 0 \end{cases}$$

this means that $x = 0$, while y is a free parameter that must satisfy both the constraints $x+y \geq 0$ (so that $y \geq 0$) and $1-x^2-y^2 \geq 0$ (and so $1-y^2 \geq 0$ that determines $-1 \leq y \leq 1$): considering both condition at the same time we have that

$$\begin{cases} y \geq 0 \\ -1 \leq y \leq 1 \end{cases} \qquad \Longrightarrow$$

$$\text{solution: } x = \mu_1 = \mu_2 = 0 \quad y \in [0,1]$$

2. When considering $\mu_1 \neq 0$ and $\mu_2 = 0$ the system of equations become

$$\begin{cases} 2x + 2\mu_1 x = 0 \\ 2\mu_1 y = 0 \\ \mu_1\left(1-x^2-y^2\right) = 0 \end{cases}$$

Based on the assumption of the value μ_1, from the second equation we determine that y must be equal to 0 and so the third equation becomes $1-x^2 = 0$ whose solutions are ± 1. Considering that the first equation can be factorised as $2x(1+\mu_1) = 0$ (and $x \neq 0$), then in order to satisfy the relation μ_1 must be -1, but this imply that μ_1 is negative and it's an unacceptable solution, so for $\mu_1 \neq 0, \mu_2 = 0$ no solutions of the system can be found;

3. Considering instead $\mu_1 = 0$ and $\mu_2 \neq 0$ the system becomes

$$\begin{cases} 2x - \mu_2 = 0 \\ -\mu_2 = 0 \\ \mu_2(x+y) = 0 \end{cases}$$

The second equation states that μ_2 must be equal to 0 and so the solution of this problem is the same as the first case considered;

4. Considering both variables $\mu_1, \mu_2 \neq 0$ we have to solve the full systems; considering the last two equations we have that

$$\begin{cases} x+y=0 \\ 1-x^2-y^2=0 \end{cases}$$
$$\implies \quad x=-y$$
$$\implies \quad 1-2x^2=0$$

and so it means that the two possible solution for the systems are

$$(x,y) = \left(\pm\frac{1}{\sqrt{2}}, \mp\frac{1}{\sqrt{2}}\right).$$

Considering the first case where $x = 1/\sqrt{2}$ the first two equations become

$$\begin{cases} \frac{2}{\sqrt{2}}(1+\mu_1)-\mu_2=0 \\ -\frac{2}{\sqrt{2}}\mu_1-\mu_2=0 \end{cases}$$
$$\implies \quad \mu_2 = -\frac{2}{\sqrt{2}}\mu_1$$
$$\implies \quad \frac{4}{\sqrt{2}}\mu_1+\frac{2}{\sqrt{2}}=0$$
$$\implies \quad \mu_1 = -\frac{1}{2} \qquad \mu_2 = \frac{1}{\sqrt{2}}$$

In this case $\mu_1 < 0$ is an unacceptable solution of the problem, and also considering the point $x = -1/\sqrt{2}$ the system becomes with no considerable solutions

$$\begin{cases} -\frac{2}{\sqrt{2}}(1+\mu_1)-\mu_2=0 \\ \frac{2}{\sqrt{2}}\mu_1-\mu_2=0 \end{cases}$$
$$\implies \quad \mu_1 = -\frac{1}{2} \quad \mu_2 = -\frac{1}{\sqrt{2}}$$

With all this considerations done the point of local minimum relies on the points

$$(x,y) = (0,k) \qquad k \in [0,1]$$

■

8.0.1.3 KKT second order necessary conditions

Given a function $f \in C^1(\mathbb{R}^n)$ to minimize subject to the equality $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ and inequality $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ constraint maps, then necessary conditions for a point $\boldsymbol{x}^\star$ (that satisfies the constraints) to be a local minima is that exists $m+p$ scalars λ_i, μ_i such that satisfy the first order conditions (Equation 8.1) and

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}\left(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star\right) \boldsymbol{z} \geq 0 \tag{8.2}$$

for all vector $\boldsymbol{z}$ such that

$$\begin{aligned} &(i) && \nabla h_k(\boldsymbol{x}^\star)\boldsymbol{z} = 0 \quad k = 1, \dots, m \\ &(ii) && \nabla g_k(\boldsymbol{x}^\star)\boldsymbol{z} = 0 \quad \text{for all } k \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_k > 0 \\ &(iii) && \nabla g_k(\boldsymbol{x}^\star)\boldsymbol{z} \geq 0 \quad \text{for all } k \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_k = 0 \end{aligned}$$

In general the conditions (ii) and (iii) are hard to verify and so to have a less accurate necessary condition (by having a smaller set of possible vector $\boldsymbol{z}$), and so improving the chance of considering point that don't belong to the original domain, this two conditions can be substitute with the expression

$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{z} = 0 \qquad \text{for all } k \in \mathscr{A}(\boldsymbol{x}^\star)$$

In other words, the second-order necessary condition (and similarly the sufficient one) states (from Equation 8.2) that the matrix $\nabla^2_{\boldsymbol{x}}\mathscr{L}$ must be positive semidefinite on the kernel of the active and equality constraints.

▪ **Example 8.4 — kernel of the active constraints.** Continuing examples Example 6.2 and Example 8.2, on the minimum candidate point $(x^\star, y^\star, z^\star)$ we have the following matrix for the active constraints:

$$H^\star = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Computing the kernel of the active constraints means so solving the following system of linear equation determined by $H^\star \boldsymbol{x} = \boldsymbol{0}$ and so

$$\begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

resulting in

$$\begin{cases} (\sqrt{3}-1)x + (1+\sqrt{3})y = 0 \\ x + z = 0 \end{cases}$$

Having 3 unknowns but only 2 equation means that we have to solve the problem in a parametric form; chosen $z = t$ we have $x = -t$ and so $y = t\frac{\sqrt{3}-1}{\sqrt{3}+1}$; choosing $t = \sqrt{3}+1$ (to simplify the results) we obtain the generator K of the kernel as

$$\boldsymbol{K} = \begin{pmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{pmatrix}$$

▪

8.0.1.4 KKT second order sufficient conditions

Given a function $f \in C^1(\mathbb{R}^n)$ to minimize subject to the equality $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$ and inequality $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ constraint maps, then sufficient condition for a point $\boldsymbol{x}^\star$ (that satisfies the constraints) to be a local minima is that exists $m+p$ scalars λ_i, μ_i such that satisfy the first order conditions (Equation 8.1) and

$$\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}\left(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star\right) \boldsymbol{z} > 0$$

for all vector $\boldsymbol{z}$ such that

$$\begin{aligned} &(i) && \nabla h_k(\boldsymbol{x}^\star)\boldsymbol{z} = 0 && k = 1, \dots, m \\ &(ii) && \nabla g_k(\boldsymbol{x}^\star)\boldsymbol{z} = 0 && \text{for all } k \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_k > 0 \\ &(iii) && \nabla g_k(\boldsymbol{x}^\star)\boldsymbol{z} \geq 0 && \text{for all } k \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_k = 0 \end{aligned}$$

As in the previous case in order to have simpler calculation the third equation (iii) can be dropped (and so we obtain less accurate solution of the local minima).

■ **Example 8.5 — second order check.** Given the problem started in example Example 6.2 and carried on up to Example 8.4 states as

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h_1(x,y,z) = x^2 + y^2 - 2 = 0 \\ & h_2(x,y,z) = x + z - 1 = 0 \end{aligned}$$

we found a candidate for minimum point in

$$\begin{aligned} x^\star &= \frac{\sqrt{3}}{2} - \frac{1}{2} \\ y^\star &= \frac{1}{2} + \frac{\sqrt{3}}{2} \\ z^\star &= \frac{3}{2} - \frac{\sqrt{3}}{2} \\ \lambda_1^\star &= \frac{1}{2} - \frac{\sqrt{3}}{2} \\ \lambda_2^\star &= 3 - \sqrt{3} \end{aligned}$$

that present the matrix $\boldsymbol{H}^\star$ for the active constraints with related kernel $\boldsymbol{K}$:

$$\boldsymbol{H}^\star = \begin{bmatrix} \sqrt{3}-1 & 1+\sqrt{3} & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad \boldsymbol{K} = \begin{bmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{bmatrix}$$

In order to check the second order necessary/sufficient condition we have to compute the hessian matrix $\nabla^2_{\boldsymbol{x}}\mathscr{L}$ of the lagrangian that's

$$\nabla^2_{\boldsymbol{x}}\mathscr{L} = \boldsymbol{L} = \begin{bmatrix} -2\lambda_1 & 0 & 0 \\ 0 & -2\lambda_1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \Longrightarrow$$

$$\boldsymbol{L}^\star = \begin{bmatrix} \sqrt{3}-1 & 0 & 0 \\ 0 & \sqrt{3}-1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We have now to determine if this resulting matrix is (semi-)positive defined or not in the kernel of active constraints in order to determine if the point is a minimum or not; this mean performing the following matrix multiplication:

$$\begin{aligned} \alpha \boldsymbol{K}^T \boldsymbol{L}^\star \boldsymbol{K} \alpha &= \alpha \begin{bmatrix} -\sqrt{3}-1 & \sqrt{3}-1 & \sqrt{3}+1 \end{bmatrix} \\ &\quad \begin{bmatrix} \sqrt{3}-1 & 0 & 0 \\ 0 & \sqrt{3}-1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -\sqrt{3}-1 \\ \sqrt{3}-1 \\ \sqrt{3}+1 \end{bmatrix} \alpha \\ &= \alpha \begin{bmatrix} -\sqrt{3}-1 & \sqrt{3}-1 & \sqrt{3}+1 \end{bmatrix} \begin{bmatrix} -2 \\ 4-2\sqrt{3} \\ \sqrt{3}+1 \end{bmatrix} \alpha \\ &= \alpha \left(2\sqrt{3}+2+4\sqrt{3}-4-6+2\sqrt{3}+3+2\sqrt{3}+1\right)\alpha \\ &= \left(10\sqrt{3}-4\right)\alpha^2 \end{aligned}$$

Observing that for $\alpha \neq 0$ the product is always greater then zero we have that the hessian $L^\star$ is positive defined in the kernel of the active constraints, satisfying the second order sufficient condition that allow so to state that

$$\boldsymbol{x}^\star = \left(\frac{\sqrt{3}}{2} - \frac{1}{2}, \frac{1}{2} + \frac{\sqrt{3}}{2}, \frac{3}{2} - \frac{\sqrt{3}}{2} \right)$$

is a minimum.

■

■ **Example 8.6** Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y) = x^2 - xy \\ \text{subject to:} \quad & g_1(x,y) = 1 - x^2 - y^2 \geq 0 \\ & g_2(x,y) = 1 - (x-1)^2 - y^2 \geq 0 \end{aligned}$$

To solve this kind of problem using the KKT conditions we firstly need to build the lagrangian of the problem that's equal to

$$\begin{aligned} \mathscr{L}(x,y,\mu_1,\mu_2) = x^2 - xy - \mu_1\left(1 - x^2 - y^2\right) \\ - \mu_2\left[1 - (x-1)^2 - y^2\right] \end{aligned}$$

At this point it's possible to construct the non linear system of equation whose solution are the stationary points candidate to be local minimum:

$$\begin{cases} \left.\begin{aligned} 2x - y + 2\mu_1 x + 2\mu_2(x-1) = 0 \\ -x + 2\mu_1 y + 2\mu_2 y = 0 \end{aligned}\right\} & \nabla_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \\ \left.\begin{aligned} \mu_1\left(1 - x^2 - y^2\right) = 0 \\ \mu_2\left[1 - (x-1)^2 - y^2\right] = 0 \end{aligned}\right\} & \mu_k g_k(\boldsymbol{x}) = 0 \\ \mu_1, \mu_2 \geq 0 \end{cases}$$

This system can be solved by cases of the value of μ_i and in particular

1. when $\mu_1 = \mu_2 = 0$ the system reduces to the form

$$\begin{cases} 2x - y = 0 \\ -x = 0 \end{cases} \qquad \Longrightarrow \qquad x = y = 0$$

   This solution satisfy the constraints, in fact $g_1(0,0) = 1 \geq 0$ and $g_2(0,0) = 0 \geq 0$, so this point can be used to compute the second order conditions;
2. considering all the other cases of μ_j the only other analytical solution to the system, found with `Mathematica`, determines the point

$$(x,y,\mu_1,\mu_2) = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}, \frac{1}{\sqrt{3}} - \frac{1}{2}, \frac{1}{2} - \frac{\sqrt{3}}{6} \right)$$

In order to consider now the second order necessary/sufficient conditions we have firstly to evaluate the gradient $\nabla \boldsymbol{g}$ of the inequality constraint map (in respect to the variables x, y) resulting in the matrix

$$\nabla \boldsymbol{g} = \begin{bmatrix} -2x & -2y \\ -2(x-1) & -2y^2 \end{bmatrix}$$

For the second order conditions is also important to define the hessian matrix of the lagrangian respect to the variables x, y:

$$\nabla^2_{\boldsymbol{x}} \mathscr{L} = \begin{bmatrix} 2 + 2\mu 1 + 2\mu 2 & -1 \\ -1 & 2\mu_1 + 2\mu_2 \end{bmatrix}$$

We have now to check the conditions for the two local minima candidates:

1. in the first case when $x = y = \mu_1 = \mu_2 = 0$ the set of the active constraints is just the second inequality g_2: in fact we have that $g_1(0,0) = 1 \neq 0$ while $g_2(0,0) = 0$. In respect to this vector we have to compute it's kernel (to determine the direction z on which verify the KKT condition) and in particular

$$\nabla g_2(0,0) = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \quad \Longrightarrow \quad \ker\{\nabla g_2(0,0)\} = \alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Evaluating the expression $\boldsymbol{z}^T \nabla^2_{\boldsymbol{x}} \mathscr{L}\, \boldsymbol{z}$ for $\boldsymbol{z} \in \ker\{\nabla g_2\}$ we obtain that

$$\alpha \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \alpha = 2\alpha^2$$

■

■ **Example 8.7 — Sample Exam Question.** Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x - y + z^2 \\ \text{subject to:} \quad & h(x,y,z) = x^2 + y^2 - 2 = 0 \\ & g(x,y,z) = x^2 - 1 \geq 0 \end{aligned}$$

the solution can be obtained by using the KKT condition; in order to use them we firstly need to define the lagrangian

$$\mathscr{L} = f - \lambda h - \mu g$$

of the problem that in this is of the form

$$\mathscr{L}(x,y,z,\lambda,\mu) = x - y + z^2 - \lambda\left(x^2 + y^2 - 2\right) - \mu\left(x^2 - 1\right)$$

The first order necessary condition allows to build the system of non-linear equation whose solution are candidates to be minimum point of f (respect to the constrains set), and so we have

$$\begin{cases} 1 - 2\lambda x - 2\mu x = 0 & : \partial/\partial_x \mathscr{L} \\ -1 - 2\lambda y = 0 & : \partial/\partial_y \mathscr{L} \\ -2z = 0 & : \partial/\partial_z \mathscr{L} \\ x^2 + y^2 - 2 = 0 & : h \\ x^2 - 1 \geq 0 & : g \\ \mu(x^2 - 1) = 0 & : \mu g \\ \mu \geq 0 & : \mu \end{cases}$$

Verifying that the point $\boldsymbol{x}^\star = (1,1,0)$ is a candidate minimum with $\lambda^\star = -\frac{1}{2}$ and $\mu^\star = 1$ (substituting the values in the system, all the relation are satisfied), the associated set of active constraints is represented by both h (in fact $1 + 1 - 2 = 0$) and g (indeed $1 - 1 = 0$) whose gradient can be regarded as

$$\nabla \mathscr{A} = \begin{bmatrix} \nabla h \\ \nabla g \end{bmatrix} = \begin{bmatrix} 2x & 2y & 0 \\ 2x & 0 & 0 \end{bmatrix} \quad \Longrightarrow$$
$$\boldsymbol{H} = \nabla \mathscr{A}(\boldsymbol{x}^\star) = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 0 & 0 \end{bmatrix}$$

The constraints are so qualified because the gradients are linearly independent and related kernel, obtained by solving the equation $H\boldsymbol{x} = \boldsymbol{0}$, gives a generator of the space in the form of $\boldsymbol{K} = (0,0,1)$.

In order to determine if $\boldsymbol{x}^\star$ is a minimum point using the second order necessary/sufficient KKT conditions it's mandatory to compute the hessian $\nabla^2_{\boldsymbol{x}}\mathscr{L}$ of the lagrangian:

$$\nabla^2_{\boldsymbol{x}}\mathscr{L} = \begin{bmatrix} -2(\lambda+\mu) & 0 & 0 \\ 0 & -2\lambda & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad \Longrightarrow$$

$$\boldsymbol{L} = \nabla^2_{\boldsymbol{x}}\mathscr{L}(\boldsymbol{x}^\star) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We now need to check if the matrix $\boldsymbol{L}$ is (semi-)positive defined in the kernel of the active constraints, and so we have to compute

$$\begin{aligned} \alpha\boldsymbol{K}^T\boldsymbol{L}\boldsymbol{K}\alpha &= \alpha \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \alpha \\ &= \alpha \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix} \alpha \\ &= 2\alpha^2 \end{aligned}$$

Observing that this product is always greater then zero for $\alpha \neq 0$, then it means that L is positive defined in the kernel of the active constraints and so

$$\boldsymbol{x}^\star = (1,1,0) \qquad \text{is a minimum point}$$

■

■ **Example 8.8 — Sample Exam Question.** Given the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x^2 - y + z \\ \text{subject to:} \quad & h(x,y,z) = y - x = 1 \\ & g_1(x,y,z) = x \geq 1 \\ & g_2(x,y,z) = z \geq 1 \end{aligned}$$

the associated lagrangian is

$$\begin{aligned} \mathscr{L}(x,y,z,\lambda,\mu_1,\mu_2) = x^2 - y + z - \lambda(y-x-1) \\ - \mu_1(x-1) - \mu_2(z-1) \end{aligned}$$

The nonlinear system associated with the first-order necessary KKT condition is therefore

$$\begin{cases} 2x + \lambda - \mu_1 = 0 & : \partial/\partial_x\mathscr{L} \\ -1 - \lambda = 0 & : \partial/\partial_y\mathscr{L} \\ 1 - \mu_2 = 0 & : \partial/\partial_z\mathscr{L} \\ y - x - 1 = 0 & : h \\ x - 1 \geq 0 & : g_1 \\ z - 1 \geq 0 & : g_2 \\ \mu_1(x-1) = 0 & : \mu_1 g_1 \\ \mu_2(z-1) = 0 & : \mu_2 g_2 \\ \mu_1, \mu_2 \geq 0 \end{cases}$$

The lonely solution of this system is the one

$$\begin{aligned} x^\star &= 1 \\ y^\star &= 2 \\ z^\star &= 1 \\ \lambda^\star &= -1 \\ \mu_1^\star = \mu_2^\star \quad &= 1 \end{aligned}$$

For such point the set of active constraint $\mathscr{A}$ is determined by all the constraints $\{h, g_1, g_2\}$ (in fact $g_1(\boldsymbol{x}^\star) = g_2(\boldsymbol{x}^\star) = 0$) whose gradient is so

$$\boldsymbol{H} = \begin{bmatrix} \nabla h \\ \nabla g_1 \\ \nabla g_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

This is a 3×3 matrix with linearly independent columns/rows and so the kernel obtained by solving $\boldsymbol{Hz} = \boldsymbol{0}$ consist only in the null vector $\boldsymbol{z} = \boldsymbol{0}$. Determined the hessian

$$\boldsymbol{L} = \nabla^2_{\boldsymbol{x}} \mathscr{L} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we have that $\boldsymbol{z}^T \boldsymbol{L} \boldsymbol{z}$ is always zero in the kernel of the active constraints, meaning that the necessary condition is satisfied, but not the sufficient and so we are in the *grey zone* on which the point might (or not) be a local minima.

■



9. Summary of KKT with proofs

The Lagrange-multiplier theorem can be extended to inequality constraints. Consider the nonlinear programming (NLP) problem

$$\begin{aligned} \text{minimize:} \quad & f(\boldsymbol{x}) \\ \text{subject to:} \quad & h_i(\boldsymbol{x}) = 0 \qquad i = 1,2,\ldots,m \\ & g_i(\boldsymbol{x}) \geq 0 \qquad i = 1,2,\ldots,p \end{aligned} \tag{9.1}$$

introducing the **slack** variables e_i for $i = 1,2,\ldots,p$ and the vector $\boldsymbol{y}^T = (\boldsymbol{x}^T, \boldsymbol{e}^T)$ the new problem

$$\begin{aligned} \text{minimize:} \quad & \mathsf{f}(\boldsymbol{y}) = f(\boldsymbol{x}) \\ \text{subject to:} \quad & \mathsf{h}_i(\boldsymbol{y}) = h_i(\boldsymbol{x}) = 0 \qquad i = 1,2,\ldots,m \\ & \mathsf{h}_{i+m}(\boldsymbol{y}) = g_i(\boldsymbol{x}) - e_i^2 = 0 \qquad i = 1,2,\ldots,p \end{aligned} \tag{9.2}$$

Using this formulation we deduce:

Lemma 9.1 — First-order KKT conditions. Defining the Lagrangian:

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{\ell=1}^{p} \mu_\ell g_\ell(\boldsymbol{x}) \tag{9.3}$$

Assume that f, h_i, and g_i are continuously differentiable and that a constraint qualification, such as LICQ, holds at the local minimizer $\boldsymbol{x}^\star$. The first-order KKT conditions for Equation 9.1 are:

$$\begin{aligned} & \nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{0}^T, \\ & h_k(\boldsymbol{x}^\star) = 0, && k = 1,2,\ldots m \\ & \mu_\ell g_\ell(\boldsymbol{x}^\star) = 0, \quad g_\ell(\boldsymbol{x}^\star) \geq 0 && \ell = 1,2,\ldots p, \\ & \mu_\ell \geq 0, && \ell = 1,2,\ldots,p. \end{aligned}$$

The last two lines are, respectively, complementary slackness and dual feasibility. Thus

$g_\ell(\boldsymbol{x}^\star) > 0$ implies $\mu_\ell = 0$. LICQ means that the gradients of the equality constraints and active inequality constraints are linearly independent.

Proof.

with the **extended** Lagrangian function:

$$\mathscr{L}^e(\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{\ell=1}^{p} \mu_\ell \left(g_\ell(\boldsymbol{x}) - e_\ell^2\right) \tag{9.4}$$

The first order condition becomes

$$\nabla_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla h_k(\boldsymbol{x}^\star) - \sum_{k=1}^{p} \mu_k \nabla g_k(\boldsymbol{x}^\star) = \mathbf{0}^T,$$

$$\nabla_e \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 2(\mu_1 e_1, \ldots, \mu_p e_p) = \mathbf{0}^T,$$

$$\nabla_{\lambda_k} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = h_k(\boldsymbol{x}^\star) = 0,$$

$$\nabla_{\mu_\ell} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = g_\ell(\boldsymbol{x}^\star) = e_\ell^2 \geq 0,$$

Notice that $\mu_\ell e_\ell = 0$ is equivalent to $\mu_\ell e_\ell^2 = 0$ and thus $\mu_\ell g_\ell(\boldsymbol{x}^\star) = 0$. Moreover $\nabla_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu})$. The sign condition $\mu_\ell \geq 0$ does not follow from stationarity with respect to e_ℓ alone; it follows from the second-order necessary condition in a pure slack direction for an active inequality. □

second order condition become:

Lemma 9.2 — second order necessary condition. Using the Lagrangian Equation 9.3, the second order condition for problem Equation 9.2 become:

$$\boldsymbol{z}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq 0,$$
$$\mu_\ell \geq 0, \qquad \ell = 1, 2, \ldots, p$$

for all $\boldsymbol{z}$ in the kernel of the gradient of the **active** constraints.

Proof.

Using Lagrangian Equation 9.4 with the slack functions the necessary condition become

$$\boldsymbol{z}^T \nabla^2_{(x,e)} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq 0$$

for all $\boldsymbol{z}$ in the kernel of matrix

$$\begin{pmatrix} \nabla_x \boldsymbol{h}(\boldsymbol{x}^\star) & \mathbf{0} \\ \nabla_x \boldsymbol{g}(\boldsymbol{x}^\star) & -2\,\mathrm{diag}(e_1, \ldots, e_p) \end{pmatrix} \tag{9.5}$$

where

$$\nabla^2_{(x,e)} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{pmatrix} \nabla_x^2 \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) & \mathbf{0} \\ \mathbf{0} & \nabla_e^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \end{pmatrix} \tag{9.6}$$

and $\nabla_x \nabla_e^T \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$. Moreover

$$\nabla_x^2 \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla^2 f(\boldsymbol{x}^\star) - \sum_{k=1}^{m} \lambda_k \nabla^2 h_k(\boldsymbol{x}^\star) - \sum_{\ell=1}^{p} \mu_\ell \nabla^2 g_\ell(\boldsymbol{x}^\star),$$
$$\nabla_e^2 \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = 2\,\mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_p).$$

After reordering, we can separate the constraints as follows:

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{g}^{(A)}(\boldsymbol{x}) \\ \boldsymbol{g}^{(I)}(\boldsymbol{x}) \end{pmatrix} \text{ and } \begin{cases} g_\ell(\boldsymbol{x}^\star) = e_\ell^2 = 0, & \ell = 1, 2, \ldots, r \\ g_\ell(\boldsymbol{x}^\star) = e_\ell^2 > 0, & \ell = r+1, r+2, \ldots, p \end{cases}$$

The group of constraints $\boldsymbol{g}^{(A)}(\boldsymbol{x}^\star)$ that are zeros are the active constraints, the remaining $\boldsymbol{g}^{(I)}(\boldsymbol{x}^\star)$ are the inactive ones. Thus, Equation 9.5 in more details becomes

$$\begin{pmatrix} \nabla \boldsymbol{h}(\boldsymbol{x}^\star) & \boldsymbol{0} & \boldsymbol{0} \\ \nabla \boldsymbol{g}^{(A)}(\boldsymbol{x}^\star) & \boldsymbol{0} & \boldsymbol{0} \\ \nabla \boldsymbol{g}^{(I)}(\boldsymbol{x}^\star) & \boldsymbol{0} & -\boldsymbol{E} \end{pmatrix}, \quad \boldsymbol{E} = 2\,\mathrm{diag}(e_{k+1}, \ldots, e_p). \tag{9.7}$$

and

$$\nabla_e^2 \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{pmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{M} = 2\,\mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_r) \tag{9.8}$$

The kernel of Equation 9.7 can be written as

$$\underline{\mathbf{K}} = \begin{pmatrix} \boldsymbol{K}^a & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{E}^{-1} \nabla_x \boldsymbol{g}^{(I)}(\boldsymbol{x}^\star) \boldsymbol{K}^a & \boldsymbol{0} \end{pmatrix}, \tag{9.9}$$

where $\boldsymbol{K}^a$ is the kernel of active constraints given by the matrix

$$\begin{pmatrix} \nabla \boldsymbol{h}(\boldsymbol{x}^\star) \\ \nabla \boldsymbol{g}^{(A)}(\boldsymbol{x}^\star) \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} \nabla \boldsymbol{h}(\boldsymbol{x}^\star) \\ \nabla \boldsymbol{g}^{(A)}(\boldsymbol{x}^\star) \end{pmatrix} \boldsymbol{K}^a = \boldsymbol{0} \tag{9.10}$$

thus $\boldsymbol{z}$ can be written as $\boldsymbol{z} = \underline{\mathbf{K}}\boldsymbol{d}$ and thus second order condition $\boldsymbol{z}^T \nabla^2_{(x,e)} \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq 0$ become

$$0 \leq \boldsymbol{d}^T \left[\underline{\mathbf{K}}^T \nabla^2_{(x,e)} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \underline{\mathbf{K}} \right] \boldsymbol{d}, \qquad \boldsymbol{d} \in \mathbb{R}^s$$

and using Equation 9.9 with Equation 9.6 and Equation 9.8

$$
\begin{aligned}
&\left[\underline{\mathbf{K}}^T \nabla^2_{(x,e)} \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \underline{\mathbf{K}}\right] \\
&\quad = \underline{\mathbf{K}}^T \begin{pmatrix} \nabla^2_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \underline{\mathbf{K}}, \\
&\quad = \underline{\mathbf{K}}^T \begin{pmatrix} \nabla^2_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{K}^a & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \\
&\quad = \begin{pmatrix} \boldsymbol{K}^{a,T} \nabla^2_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{K}^a & \mathbf{0} \\ \mathbf{0} & \boldsymbol{M} \end{pmatrix}
\end{aligned} \tag{9.11}
$$

Using the equality

$$
\nabla^2_x \mathscr{L}^e(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla^2_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{9.12}
$$

the second order condition becomes: $\nabla^2_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu})$ must be positive semi-definite in the kernel of the gradient of the **active** constraints. The matrix $\boldsymbol{M}$ must be semi-positive definite or equivalenty $\mu_\ell \geq 0$ for all the multipliers of the *not-active* constraints (remember that for the active constraints must be $\mu_\ell = 0$). □

Lemma 9.3 — second order sufficient condition. Using the Lagrangian Equation 9.3, the second order sufficient condition for problem Equation 9.2 become:

$$
\begin{aligned}
\boldsymbol{z}^T \nabla^2_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq \varepsilon \|\boldsymbol{z}\|^2, & \\
\mu_\ell = 0, \qquad & \text{for } g_\ell(\boldsymbol{x}^\star) > 0 \\
\mu_\ell > 0, \qquad & \text{for } g_\ell(\boldsymbol{x}^\star) = 0
\end{aligned}
$$

for all $\boldsymbol{z}$ in the kernel of the gradient of the **active** constraints.

Proof.

Using Equation 9.11 with Equation 9.12 the sufficient condition become that the hessian satisfy

$$
\boldsymbol{z}^T \nabla^2_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{z} \geq \varepsilon \|\boldsymbol{z}\|^2
$$

for all $\boldsymbol{z}$ in the kernel of the gradients of **active** constraints.
And matrix $\boldsymbol{M}$ must be positive definite, or $\mu_\ell > 0$ for all **inactive** constraints. □

The last two lemma can be resumed in:

First and second order KKT condition

- **First order condition:** for the NLP Equation 9.1

$$
\begin{aligned}
(a) \qquad & \nabla_x \mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0} \\
(b) \qquad & h_k(\boldsymbol{x}) = 0 \\
(c) \qquad & \mu_\ell g_\ell(\boldsymbol{x}) = 0 \\
(d) \qquad & g_\ell(\boldsymbol{x}) \geq 0 \\
(e) \qquad & \mu_\ell \geq 0
\end{aligned}
$$

- **Second order condition:**
  The matrix $\boldsymbol{K}$ is the kernel of the gradient of the active constraints.
  - **Necessary:** the matrix

    $$\boldsymbol{K}^{a,T} \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{K}^a,$$

    must be **semi-positive** defined.
  - **Sufficient:** the matrices

    $$\boldsymbol{K}^{a,T} \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \boldsymbol{K}^a,$$

    must be **positive defined** and $\mu_\ell > 0$ for the **active** constraints, i.e., when $g_\ell(\boldsymbol{x}^\star) = 0$.

9.1 Constrained minima, NLP problem

Consider the constrained minimization problem

$$
\begin{aligned}
\text{minimize:} \quad & f(\boldsymbol{x}) \\
\text{subject to:} \quad & h_i(\boldsymbol{x}) = 0 \qquad i = 1, 2, \ldots, m \\
& g_\ell(\boldsymbol{x}) \geq 0 \qquad \ell = 1, 2, \ldots, p
\end{aligned}
\tag{9.13}
$$

Solution Algorithm

1. **Compute the Lagrangian Function**: The Lagrangian function is defined as:

   $$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{\ell=1}^{p} \mu_\ell g_\ell(\boldsymbol{x}).$$

2. **Solve the Nonlinear System**: We need to solve the following system of equations:

   $$
   \begin{aligned}
   \nabla_x \mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \mathbf{0}, \\
   h_k(\boldsymbol{x}) &= 0, \qquad k = 1, 2, \ldots, m, \\
   \mu_\ell g_\ell(\boldsymbol{x}) &= 0, \qquad \ell = 1, 2, \ldots, p.
   \end{aligned}
   $$

   Ensure to keep only the solutions where $\mu_k^\star \geq 0$ and $g_k(\boldsymbol{x}^\star) \geq 0$.
3. **Check Active Constraints**: For each solution point $(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star)$:
   - Compute $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ and $\nabla g_\ell(\boldsymbol{x}^\star)$ for the active constraints ($g_\ell(\boldsymbol{x}^\star) = 0$).
   - Verify that these gradients are linearly independent.
4. **Compute the Kernel Matrix**: Construct the matrix $\boldsymbol{K}^a$ representing the kernel of $\nabla \boldsymbol{h}(\boldsymbol{x}^\star)$ along with the gradients $\nabla g_k(\boldsymbol{x}^\star)$ for the active constraints where $\mu_k > 0$.
5. **Compute the Reduced Hessian**: The reduced Hessian is calculated as:

   $$\boldsymbol{H} = \boldsymbol{K}^{a,T} \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^\star) \boldsymbol{K}^a$$

6. **Evaluate Conditions**:
   - **Necessary Condition**: Ensure that $\boldsymbol{H}$ is semi-positive definite.
   - **Sufficient Condition**: Confirm that $\boldsymbol{H}$ is positive definite and that $\mu_\ell > 0$ for all active constraints.

Definition 9.1 — ammissibile point. A point $\boldsymbol{x}^\star$ is admissible if

$$\begin{aligned} h_k(\boldsymbol{x}^\star) &= 0 \qquad k = 1,2,\ldots,m \\ g_\ell(\boldsymbol{x}^\star) &\geq 0 \qquad \ell = 1,2,\ldots,p \end{aligned}$$

Definition 9.2 — Feasible region.

$$\mathscr{F} = \Big\{ \boldsymbol{x} \in \mathbb{R}^n \mid h_k(\boldsymbol{x}) = 0, \qquad k = 1,2,\ldots,m, \\ g_\ell(\boldsymbol{x}) \geq 0, \qquad \ell = 1,2,\ldots,p, \Big\}$$

is called the *feasible region* or set of *feasible points*.

Definition 9.3 — active constraints. The following set

$$\mathscr{A}(\boldsymbol{x}^\star) = \{k \mid g_k(\boldsymbol{x}^\star) = 0\}$$

is named *active constraints set*. This set can be split in two subsets

$$\begin{aligned} \mathscr{A}^+(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) &= \{k \mid g_k(\boldsymbol{x}^\star) = 0, \quad \mu_k^* > 0\} \\ \mathscr{A}^0(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) &= \{k \mid g_k(\boldsymbol{x}^\star) = 0, \quad \mu_k^* = 0\} \end{aligned}$$

$\mathscr{A}^+(\boldsymbol{x}^\star, \boldsymbol{\mu}^*)$ are the **strongly active** constraints and $\mathscr{A}^0(\boldsymbol{x}^\star, \boldsymbol{\mu}^*)$ are the **weakly active** constraints.

Obviously

$$\mathscr{A}^0(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) \bigcap \mathscr{A}^+(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) = \emptyset \quad \text{and} \quad \mathscr{A}^0(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) \bigcup \mathscr{A}^+(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) = \mathscr{A}(\boldsymbol{x}^\star)$$

9.2 Constrained minima and KKT

Here, we present classical theorems of constrained minimization, including those by Karush, Kuhn, Tucker, and others, without providing detailed proofs.

The following theorem (see [4]) establishes the necessary conditions for constrained minima. Note that no specific conditions on the constraints are required.

Theorem 9.1 — Fritz John. If the functions $f(\boldsymbol{x})$, $g_1(\boldsymbol{x})$, $\ldots$, $g_p(\boldsymbol{x})$, are differentiable, then a necessary condition that $\boldsymbol{x}^\star$ be a local minimum to problem:

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{x}) \\ &\text{subject to:} && g_i(\boldsymbol{x}) \geq 0 \qquad i = 1,2,\ldots,p \end{aligned}$$

is that there exist scalars $\mu_0^\star, \mu_1^\star, \ldots, \mu_p^\star$, (not all zero) such that the following inequalities and equalities are satisfied:

$$\begin{aligned}\nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\mu}^*) &= \mathbf{0}^T \\ \mu_k^* g_k(\boldsymbol{x}^\star) &= 0, \qquad k = 1, 2, \ldots, p; \\ \mu_k^* &\geq 0, \qquad k = 0, 1, 2, \ldots, p;\end{aligned}$$

where

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\mu}) = \mu_0 f(\boldsymbol{x}) - \sum_{k=1}^{p} \mu_k g_k(\boldsymbol{x})$$

In [6] the authors showed that if a condition, called the *first order constraint qualification*, holds at $\boldsymbol{x}^\star$, $\boldsymbol{\lambda}^\star$ then λ_0 can be taken equal to 1.

Definition 9.4 — Constraints qualification LI. Let be the unilateral and bilateral constraints $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$, the point $\boldsymbol{x}^\star$ is admissible if

$$g_k(\boldsymbol{x}^\star) \geq 0, \qquad h_k(\boldsymbol{x}^\star) = 0.$$

The constraints $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x})$ are **qualified** at $\boldsymbol{x}^\star$ if the point $\boldsymbol{x}^\star$ is admissible and the vectors

$$\{\nabla g_k(\boldsymbol{x}^\star) \ : \ k \in \mathscr{A}(\boldsymbol{x}^\star)\} \bigcup \{\nabla h_1(\boldsymbol{x}^\star), \nabla h_2(\boldsymbol{x}^\star), \ldots, \nabla h_m(\boldsymbol{x}^\star)\}$$

are *linearly independent*.

The next theorems are taken from [10] and [9].

Theorem 9.2 — First order necessary conditions. Let $f \in C^1(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^1(\mathbb{R}^n, \mathbb{R}^m)$. Suppose that $\boldsymbol{x}^\star$ is a local minima of Equation 9.13 and that the constraints qualification LI holds at $\boldsymbol{x}^\star$. Then there are Lagrange multiplier vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ such that the following conditions are satisfied at $(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu})$:

$$\begin{aligned}\nabla_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \mathbf{0}^T \\ h_k(\boldsymbol{x}^\star) &= 0, \qquad k = 1, 2, \ldots, m; \\ \mu_\ell^* g_k(\boldsymbol{x}^\star) &= 0, \qquad \ell = 1, 2, \ldots, p; \\ \mu_\ell^* &\geq 0, \qquad \ell = 1, 2, \ldots, p;\end{aligned}$$

where

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) - \sum_{k=1}^{m} \lambda_k h_k(\boldsymbol{x}) - \sum_{\ell=1}^{p} \mu_\ell g_\ell(\boldsymbol{x})$$

Theorem 9.3 — Second order necessary conditions. Let $f \in C^2(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^2(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$.

Let $\boldsymbol{x}^\star$ satisfying the first order necessary conditions, a **necessary** condition for $\boldsymbol{x}^\star$ be a *local minima* is that the $m + p$ scalars (Lagrange Multiplier) of the first order necessary condition satisfy:

$$\boldsymbol{d}^T \nabla_x^2 \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \boldsymbol{d} \geq 0$$

for all $\boldsymbol{d}$ such that:

$$\nabla h_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \text{for} \quad k = 1,2,\dots,m$$

and for all active inequalities $\ell \in \mathscr{A}(\boldsymbol{x}^\star)$:

easy condition

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0$$

more accurate condition

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0 \quad \text{and} \quad \mu_\ell > 0$$

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0 \quad \text{and} \quad \mu_\ell = 0$$

Remark 9.1 The conditions

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \text{if } \ell \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_\ell > 0$$

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \qquad \text{if } \ell \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_\ell = 0$$

restrict the space of direction to be considered. If changed with

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \text{if } \ell \in \mathscr{A}(\boldsymbol{x}^\star)$$

Theorem 9.3 is still valid cause necessary condition is tested in a **smaller** set.

Theorem 9.4 — Second order sufficient conditions. Let $f \in C^2(\mathbb{R}^n)$ and the constraints $\boldsymbol{g} \in C^2(\mathbb{R}^n, \mathbb{R}^p)$ and $\boldsymbol{h} \in C^2(\mathbb{R}^n, \mathbb{R}^m)$.

Let $\boldsymbol{x}^\star$ satisfying the First order necessary conditions, a **sufficient** condition for $\boldsymbol{x}^\star$ be a *local minima* is that the $m+p$ scalars (Lagrange Multiplier) of the first order necessary condition satisfy:

$$\boldsymbol{d}^T \nabla^2_x \mathscr{L}(\boldsymbol{x}^\star, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)\boldsymbol{d} \;>\; 0$$

for all $\boldsymbol{d} \neq \boldsymbol{0}$ such that

$$\nabla h_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \qquad \text{for} \quad k = 1,2,\dots,m$$

and for all active inequalities $\ell \in \mathscr{A}(\boldsymbol{x}^\star)$:

easy condition

$$\nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \quad \text{and} \quad \mu_\ell > 0$$

more accurate condition

$$\begin{aligned} \nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \quad &\text{and} \quad \mu_k > 0 \\ \nabla g_k \ell(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \quad &\text{and} \quad \mu_\ell = 0 \end{aligned}$$

Remark 9.2 The condition

$$\nabla g_k(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \qquad \text{if } k \in \mathscr{A}(\boldsymbol{x}^\star) \text{ and } \mu_k = 0$$

restrict the space of direction to be considered. If omitted the Theorem 9.4 is still valid cause sufficient condition is tested in a **larger** set.

9.3 Other constraints qualifications

In the study of optimality condition the constraints and its gradients cannot be arbitrary. They must satisfy additional analytic/geometric properties. This properties are named **constraints qualification**. The easiest qualification (but also compelling) is linear independence (LI)

Definition 9.5 — Constraints qualification LI. Given the inequality constraints $\boldsymbol{g}(\boldsymbol{x})$ and equality constraints $\boldsymbol{h}(\boldsymbol{x})$, we will say than an admissible point $\boldsymbol{x}^\star$ is **qualified** if the vectors

$$\{\nabla g_k(\boldsymbol{x}^\star) \,:\, k \in \mathscr{A}(\boldsymbol{x}^\star)\} \bigcup \{\nabla h_1(\boldsymbol{x}^\star), \nabla h_2(\boldsymbol{x}^\star), \ldots, \nabla h_m(\boldsymbol{x}^\star)\}$$

are linearly independent.

9.3.0.1 Mangasarian-Fromovitz qualification

This qualification is less stringent of the previous

Definition 9.6 — Mangasarian-Fromovitz qualification. Given the inequality constraints $\boldsymbol{g}(\boldsymbol{x})$ and equality constraints $\boldsymbol{h}(\boldsymbol{x})$, we will say than an admissible point $\boldsymbol{x}^\star$ is **qualified** if *does not exists* a linear combination

$$\sum_{\ell \in \mathscr{A}(\boldsymbol{x}^\star)}^{m} \alpha_\ell \nabla g_\ell(\boldsymbol{x}^\star) + \sum_{k=1}^{m} \beta_k \nabla h_k(\boldsymbol{x}^\star) = \boldsymbol{0}$$

with $\alpha_\ell \geq 0$ for $\ell \in \mathscr{A}(\boldsymbol{x}^\star)$ and α_ℓ and β_k not all zero. That is, there is no non trivial linear combination for the null vector with $\alpha_\ell \geq 0$ for $\ell \in \mathscr{A}(\boldsymbol{x}^\star)$.

9.3.0.2 Garth P. McCormick qualification

Definition 9.7 — Garth P. McCormick constraints qualification (1 order). Given an admissible point $\boldsymbol{x}^\star$ the constraints are **first order qualified** if for all direction $\boldsymbol{d}$ that satisfy

$$\begin{aligned} \nabla h_k(\boldsymbol{x}^\star)\boldsymbol{d} = 0, \quad & k \in \{1,2,\ldots,m\}, \\ \nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} \geq 0, \quad & \ell \in \mathscr{A}(\boldsymbol{x}^\star), \end{aligned}$$

exists a curve $\boldsymbol{x} \in \mathrm{C}^1(\mathbb{R}, \mathbb{R}^n)$ and an $\varepsilon > 0$ such that for $0 < t < \varepsilon$.

$$\begin{aligned} \boldsymbol{x}(0) = \boldsymbol{x}^\star, \qquad & h_k(\boldsymbol{x}(t)) = 0, \quad k \in \{1,2,\ldots,m\}, \\ \boldsymbol{x}'(0) = \boldsymbol{d}, \qquad & g_\ell(\boldsymbol{x}(t)) \geq 0, \quad \ell \in \{1,2,\ldots,p\}. \end{aligned}$$

Definition 9.8 — Garth P. McCormick constraints qualification (2 order). Given an admissible point $\boldsymbol{x}^\star$ the constraints are **second order qualified** if for all direction $\boldsymbol{d}$ that satisfy

$$\begin{aligned} \nabla h_k(\boldsymbol{x}^\star)\boldsymbol{d} &= 0, \qquad k \in \{1,2,\ldots,m\}, \\ \nabla g_\ell(\boldsymbol{x}^\star)\boldsymbol{d} &= 0, \qquad \ell \in \mathscr{A}(\boldsymbol{x}^\star), \end{aligned}$$

exists a curve $\boldsymbol{x} \in \mathtt{C}^2(\mathbb{R},\mathbb{R}^n)$ and an $\varepsilon > 0$ such that for $0 < t < \varepsilon$.

$$\begin{aligned} \boldsymbol{x}(0) &= \boldsymbol{x}^\star, \qquad h_k(\boldsymbol{x}(t)) = 0, \qquad k \in \{1,2,\ldots,m\}, \\ \boldsymbol{x}'(0) &= \boldsymbol{d}, \qquad g_\ell(\boldsymbol{x}(t)) = 0, \qquad \ell \in \mathscr{A}(\boldsymbol{x}^\star). \end{aligned}$$



10. Worked examples

10.1 Example 1: Detailed Example of Constrained Minimization

In this example, we will minimize the function

$$\begin{aligned} \text{minimize:} \quad & f(x,y) = x+y, \\ \text{subject to the constraint:} \quad & h(x,y) = x^2+y^2-1 = 0, \end{aligned}$$

where $h(x,y) = 0$ describes a unit circle. We will approach this problem using two methods:

1. **Parametrization with $x = \cos(t)$ and $y = \sin(t)$.**
2. **Lagrange Multipliers.**

10.1.1 Parametric Solution

The constraint $h(x,y) = x^2+y^2-1$ represents a unit circle. We can parametrize the constraint as:

$$x(t) = \cos(t), \quad y(t) = \sin(t), \quad t \in [0, 2\pi].$$

Substituting the parametrization into the objective function yields:

$$f(\cos(t), \sin(t)) = \cos(t) + \sin(t).$$

To find the critical points, we can differentiate this parametric function and analyze it:

10.1.1.1 Critical Points

The maximum and minimum values of f can be found by solving for t:

- The function $f(\cos(t), \sin(t))$ achieves its extrema when:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\cos(t) + \sin(t)) = -\sin(t) + \cos(t) = 0.$$

This gives us:

$$\sin(t) = \cos(t) \quad \Rightarrow \quad t = \frac{\pi}{4} + k\pi, \, k \in \mathbb{Z}.$$

10.1.1.2 Function Values

Evaluating the function at these points:

- At $t = \dfrac{5\pi}{4}$ (minimum):

$$\left.\begin{aligned} x &= \cos\left(\frac{5\pi}{4}\right) = -\frac{\sqrt{2}}{2}, \\ y &= \sin\left(\frac{5\pi}{4}\right) = -\frac{\sqrt{2}}{2}, \end{aligned}\right\} \quad \Longrightarrow \quad f(x,y) = -\sqrt{2}.$$

- At $t = \dfrac{\pi}{4}$ (maximum):

$$\left.\begin{aligned} x &= \cos\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}, \\ y &= \sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}, \end{aligned}\right\} \quad \Longrightarrow \quad f(x,y) = \sqrt{2}.$$

10.1.1.3 Visualization

Let's visualize the unit circle and the level sets of the objective function.

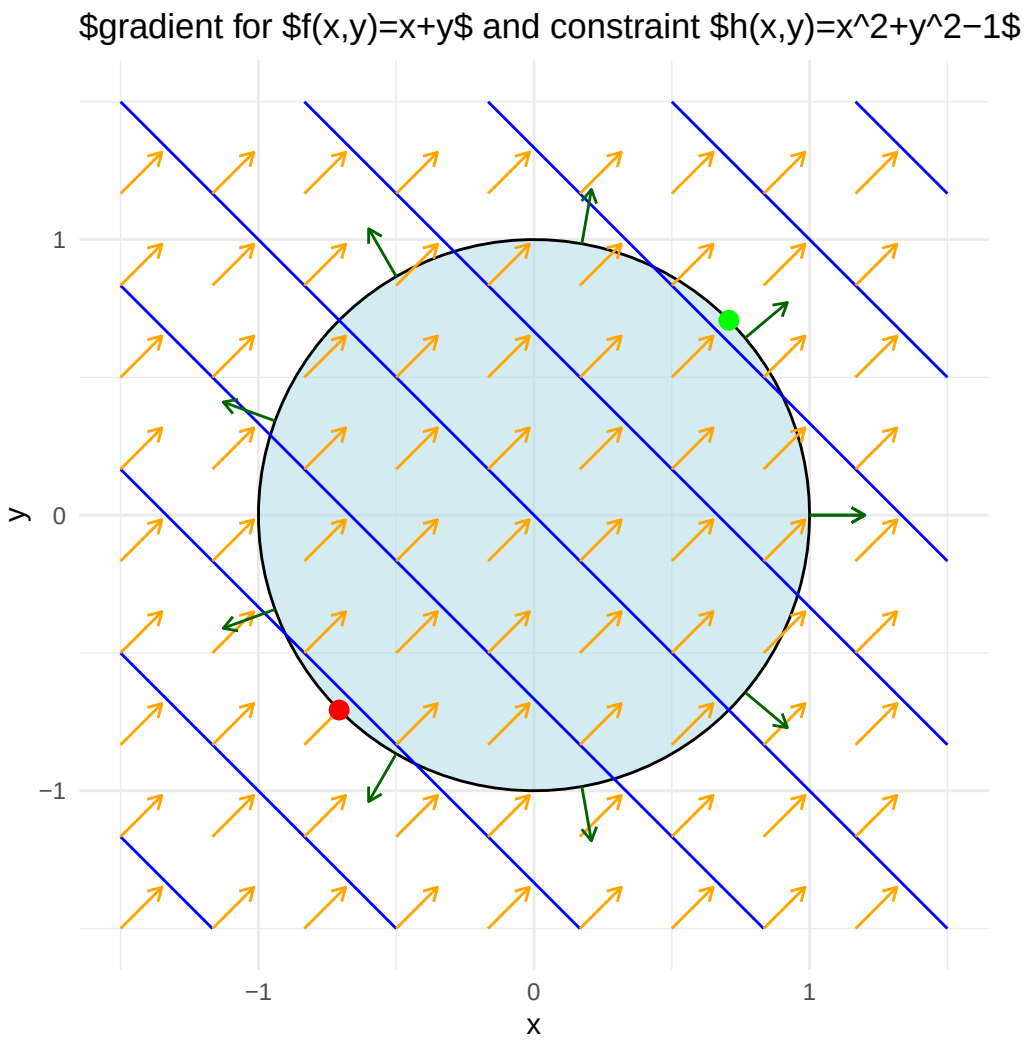


Figure 10.1: The plot illustrates that the minimum and maximum points occur where the gradient vector field of the function $f(x,y) = x+y$ is aligned with the normal vectors of the constraint curve $h(x,y) = x^2+y^2-1$. This alignment indicates that the gradient is parallel to the constraint's normal, satisfying the first order necessary condition.

10.1.2 Lagrange Multipliers Method

Now, we will apply the method of Lagrange multipliers to our constrained optimization problem. We define the Lagrangian as follows:

$$\mathscr{L}(x,y,\lambda) = f(x,y) - \lambda h(x,y) = x + y - \lambda(x^2 + y^2 - 1).$$

10.1.2.1 First-Order Conditions

To find the extrema, we take the gradients of the Lagrangian with respect to x, y, and λ. This leads to the first-order conditions:

1. For x:
$$\nabla_x \mathscr{L} = 1 - 2\lambda x = 0 \quad \Rightarrow \quad \lambda = \frac{1}{2x}.$$
2. For y:
$$\nabla_y \mathscr{L} = 1 - 2\lambda y = 0 \quad \Rightarrow \quad \lambda = \frac{1}{2y}.$$
3. The constraint condition:
$$h(x,y) = x^2 + y^2 - 1 = 0.$$

10.1.2.2 Solving the nonlinear system

Setting the two expressions for λ equal gives us:

$$\frac{1}{2x} = \frac{1}{2y} \quad \Rightarrow \quad x = y.$$

Next, substituting $x = y$ into the constraint:

$$2x^2 = 1 \quad \Rightarrow \quad x = \pm\frac{1}{\sqrt{2}} \quad \Rightarrow \quad y = \pm\frac{1}{\sqrt{2}}.$$

10.1.2.3 Second-Order Conditions

To determine whether the critical points are minima or maxima, we need to evaluate the second-order conditions.

1. **Second-Order Necessary Condition**: The Hessian of the Lagrangian, $\nabla^2_{(x,y)}\mathscr{L}(x,y,\lambda)$, must be **semi-positive definite** in the null space of the constraint gradients.
2. **Second-Order Sufficient Condition**: The Hessian of the Lagrangian, $\nabla^2_{(x,y)}\mathscr{L}(x,y,\lambda)$, must be **positive definite** in the null space of the constraint gradients.

10.1.2.4 Computing the Hessian

The Hessian matrix of the Lagrangian with respect to x and y is given by:

$$\nabla^2 \mathscr{L}(x,y,\lambda) = \begin{bmatrix} \frac{\partial^2 \mathscr{L}}{\partial x^2} & \frac{\partial^2 \mathscr{L}}{\partial x \partial y} \\ \frac{\partial^2 \mathscr{L}}{\partial y \partial x} & \frac{\partial^2 \mathscr{L}}{\partial y^2} \end{bmatrix} = \begin{bmatrix} -2\lambda & 0 \\ 0 & -2\lambda \end{bmatrix}.$$

10.1.2.5 Kernel of the Constraint

The kernel of the constraint

$$h(x,y) = 0$$

consists of the vectors $\boldsymbol{z}$ that satisfy

$$\nabla h(x,y) \cdot \boldsymbol{z} = 0.$$

The gradient of the constraint is:

$$\nabla h(x,y) = [2x, 2y]$$

and it is easy to verify that the kernel of the constraint gradient can be spanned by vector

$$z = \begin{pmatrix} -y \\ x \end{pmatrix}$$

10.1.2.6 Check qualification

Gradient of active constraint $\nabla h(x,y)$ consist of a single nonzero row. Thus, is full rank and satify LI (linear indipendence) qualification.

10.1.2.7 Evaluating the Hessian at Critical Points

1. At point $(x^\star, y^\star) = \left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ the multiplier $\lambda = -\frac{1}{\sqrt{2}}$. The kernel of $h(x,y)$ is spanned by
$$z = t \begin{pmatrix} y^\star \\ -x^\star \end{pmatrix} = \frac{t}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix} \in \ker(\nabla(x^\star, y^\star))$$
Thus the Hessian and second order condition becomes:
$$\nabla^2 \mathscr{L} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad z^2 \nabla^2 \mathscr{L} z \quad \Longrightarrow$$
$$\frac{t^2}{2} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = t^2 > 0,$$
for all $t \neq 0$. Therefore, this point is a **local minimum**.
2. At point $(x^\star, y^\star) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ the multiplier $\lambda = \frac{1}{\sqrt{2}}$.
Thus the Hessian and second order condition becomes:
$$\nabla^2 \mathscr{L} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad z^2 \nabla^2 \mathscr{L} z \quad \Longrightarrow$$
$$\frac{t^2}{2} \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -t^2 < 0, \qquad \forall t \neq 0$$
Therefore, this point is a **local maximum**.

10.1.3 Conclusion

Both methods yield the same results:

- The **minimum** occurs at $\left(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)$ with $f(x,y) = -\sqrt{2}$.
- The **maximum** occurs at $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ with $f(x,y) = \sqrt{2}$.

10.2 Example 2: Detailed example of KKT

In this example, we will minimize the function

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x^2 + y + z, \\ \text{equality constraint:} \quad & x + y^2 + z^2 = 1, \\ \text{and inequalities:} \quad & z \leq 1, \quad x \geq 0, \end{aligned}$$

10.2.0.1 Rewriting in Standard Form

We can rewrite the problem in a more standard form:

$$\begin{aligned} \text{minimize:} \quad & f(x,y,z) = x^2 + y + z, \\ \text{equality constraint:} \quad & h(x,y,z) = x + y^2 + z^2 - 1 = 0, \\ \text{and inequalities:} \quad & g_1(x,y,z) = 1 - z \geq 0, \\ & g_2(x,y,z) = x \geq 0, \end{aligned}$$

Here, $h(x,y,z) = 0$ describes a unit circle. We will now solve this problem using the KKT (Karush-Kuhn-Tucker) conditions.

10.2.0.2 Lagrangian Function

We define the Lagrangian function as:

$$\mathscr{L}(x,y,z,\lambda,\mu_1,\mu_2) = f(x,y,z) - \lambda h(x,y,z) - \mu_1 g_1(x,y,z) - \mu_2 g_2(x,y,z)$$

Substituting the expressions for f, h, g_1, and g_2:

$$\mathscr{L}(x,y,z,\lambda,\mu_1,\mu_2) = x^2 + y + z - \lambda(x + y^2 + z^2 - 1) - \mu_1(1 - z) - \mu_2 x$$

10.2.0.3 First-Order Conditions

To satisfy the first-order conditions, we compute the partial derivatives of $\mathscr{L}$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 2x - \lambda - \mu_2 = 0 \\ \dfrac{\partial \mathscr{L}}{\partial y} = 1 - 2\lambda y = 0 \\ \dfrac{\partial \mathscr{L}}{\partial x} = 1 - 2\lambda z + \mu_1 = 0 \end{cases}$$

The KKT conditions also require the original constraints to hold:

$$\begin{cases} x + y^2 + z^2 - 1 = 0 \\ \mu_1(1 - z) = 0 \\ \mu_2 x = 0 \end{cases}$$

We now proceed with different cases based on the values of μ_1 and μ_2

- **Case 1:** $\mu_1 = \mu_2 = 0$
  In this case, the first-order conditions simplify to:

$$\begin{cases} 2x - \lambda = 0 \\ 1 - 2\lambda y = 0 \\ 1 - 2\lambda z = 0 \\ x + y^2 + z^2 - 1 = 0 \end{cases} \implies \begin{cases} x = \lambda/2 \\ y = 1/(2\lambda) \\ z = 1/(2\lambda) \\ \lambda/2 + 2/(2\lambda)^2 - 1 = 0 \end{cases} \implies \lambda^3 - 2\lambda^2 + 1 = 0$$

  solving the cubic (with Maple) the real roots are

$$\lambda = 1, \ \frac{1+\sqrt{5}}{2}, \ \frac{1-\sqrt{5}}{2}$$

  From these values, we compute the corresponding x, y, and z:

$$\mu_1=\mu_0=0,\quad \begin{cases} x=y=z=1/2, & \lambda=1\\ x=(1+\sqrt{5})/4, & y=z=1/(\sqrt{5}+1), \quad \lambda=(\sqrt{5}+1)/2\\ x=(1-\sqrt{5})/4<0, & y=z=1/(1-\sqrt{5}), \quad \lambda=(1-\sqrt{5})/2 \end{cases}$$

the last solution must be discarded

- **Case 2:** $\mu_1=0$**, and** $\mu_2>0$
  In this case, $x=0$, and the system becomes:

$$\begin{cases} 2x-\lambda-\mu_2=0\\ 1-2\lambda y=0\\ 1-2\lambda z=0\\ x+y^2+z^2-1=0\\ x=0 \end{cases} \Longrightarrow \begin{cases} \lambda=-\mu_2\\ y=1/(2\lambda)=-1/(2\mu_2)\\ z=1/(2\lambda)=-1/(2\mu_2)\\ 2/(2\mu_2)^2-1=0\\ x=0 \end{cases} \Longrightarrow \mu_2^2=\frac{1}{2}$$

the positive root is $\mu_2=1/\sqrt{2}$ so that

$$x=0,\quad y=z=-\frac{1}{2\mu_2}=-\frac{\sqrt{2}}{2},\quad \lambda=-\frac{1}{\sqrt{2}}=-\frac{\sqrt{2}}{2},\quad \mu_1=0,\quad \mu_2=\frac{1}{\sqrt{2}}$$

- **Case 3:** $\mu_2=0$**, and** $\mu_1>0$

$$\begin{cases} 2x-\lambda=0\\ 1-2\lambda y=0\\ 1-2\lambda z+\mu_1=0\\ x+y^2+z^2-1=0\\ z=1 \end{cases} \Longrightarrow \begin{cases} 2x-\lambda=0\\ 1-2\lambda y=0\\ 1-2\lambda+\mu_1=0\\ x+y^2=0\\ z=1 \end{cases} \Longrightarrow \begin{cases} x=\lambda/2\\ y=1/(2\lambda)\\ \mu_1=2\lambda-1\\ x+y^2=0\\ z=1 \end{cases}$$

so that the real root of λ is

$$\frac{\lambda}{2}+\frac{1}{4\lambda^2}=0 \Longrightarrow 2\lambda^3+1=0 \Longrightarrow \lambda=-2^{-1/3} \Longrightarrow$$

and

$$x=-2^{-4/3}<0,\quad y=2^{-2/3},\quad z=1,\quad \lambda=-2^{-1/3},\quad \mu_1=2^{2/3}-1>0,\quad \mu_2=0$$

the solutoion must be discarded.

- **Case 4:** $\mu_2>0$ **and** $\mu_1>0$

$$\begin{cases} 2x-\lambda-\mu_2=0\\ 1-2\lambda y=0\\ 1-2\lambda z+\mu_1=0\\ x+y^2+z^2-1=0\\ z=1\\ x=0 \end{cases} \Longrightarrow \begin{cases} -\lambda-\mu_2=0\\ 1-2\lambda y=0\\ 1-2\lambda+\mu_1=0\\ y^2=0\\ z=1\\ x=0 \end{cases} \Longrightarrow 1-2\lambda\cdot 0=0, \quad \text{no solution}$$

10.2.0.4 All the solutions

x	y	z	λ	μ_1	μ_2
$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	0	0
$\frac{1+\sqrt{5}}{4}$	$\frac{1}{\sqrt{5}+1}$	$\frac{1}{\sqrt{5}+1}$	$\frac{\sqrt{5}+1}{2}$	0	0
0	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$	$-\frac{\sqrt{2}}{2}$	0	$\frac{\sqrt{2}}{2}$

10.2.1 Study of solution 1

$x = y = z = \frac{1}{2}$, $\lambda = 1$, $\mu_1 = \mu_2 = 0$

10.2.1.1 Compute active constraints

$$g_1(x,y,z) = 1 - z = 1 - \frac{1}{2} = \frac{1}{2} > 0, \qquad \text{(not-active)}$$

$$g_2(x,y,z) = x = \frac{1}{2} > 0, \qquad \text{(not-active)}$$

Gradient of active constraints

$$\nabla h(x,y,z) = \nabla(x + y^2 + z^2 - 1) = [1, 2y, 2z] = [1,1,1]$$

find the kernel of the active constraint

$$[1,1,1]\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = z_1 + z_2 + z_3 \implies \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} -\alpha - \beta \\ \alpha \\ \beta \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\boldsymbol{K}^a} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

10.2.1.2 Check qualification

Gradient of active constraint $\nabla h(x,y,z)$ consist of a single nonzero row. Thus, is full rank and satify LI (linear indipendence) qualification.

10.2.1.3 Compute hessian and reduced hessian

$$\nabla^2_{(x,y,z)} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2\lambda & 0 \\ 0 & 0 & -2\lambda \end{pmatrix} \implies \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}$$

reduced hessian

$$\begin{aligned} \boldsymbol{K}^{\top,a} \nabla^2_{(x,y,z)} \boldsymbol{K}^a &= \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -2 & -2 \\ -2 & 0 \\ 0 & -2 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} \end{aligned}$$

Use Sylverster to check positivity

- $\boldsymbol{H}_1 = (0)$, determinant 0, check perturbed $\det(0+\varepsilon) = \varepsilon$ and $\det(0+\varepsilon)' = 1 > 0$ [OK]
- $\det \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} = -4 < 0$ [NO]

the point is **not a minimum**.

10.2.2 Study of solution 2

$x = \frac{\sqrt{5}+1}{4}, y = z = \frac{\sqrt{5}-1}{4}, \lambda = \frac{\sqrt{5}+1}{2}, \mu_1 = \mu_2 = 0$

10.2.2.1 Compute active constraints

$$g_1(x,y,z) = 1 - z = 1 - \frac{\sqrt{5}-1}{4} = \frac{5-\sqrt{5}}{4} > 0, \qquad \text{(not-active)}$$

$$g_2(x,y,z) = x = \frac{\sqrt{5}+1}{4} > 0, \qquad \text{(not-active)}$$

Gradient of active constraints

$$\nabla h(x,y,z) = \begin{pmatrix} 1 & 2y & 2z \end{pmatrix} = \begin{pmatrix} 1 & \frac{\sqrt{5}-1}{2} & \frac{\sqrt{5}-1}{2} \end{pmatrix}$$

find the kernel of the active constraint

$$\begin{pmatrix} 1 & \frac{\sqrt{5}-1}{2} & \frac{\sqrt{5}-1}{2} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = z_1 + \frac{\sqrt{5}-1}{2}(z_2+z_3)$$

$$\Downarrow$$

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \begin{pmatrix} \frac{1-\sqrt{5}}{2}(\alpha+\beta) \\ \alpha \\ \beta \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1-\sqrt{5}}{2} & \frac{1-\sqrt{5}}{2} \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\boldsymbol{K}^a} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

10.2.2.2 Check qualification

Gradient of active constraint $\nabla h(x,y,z)$ consist of a single nonzero row. Thus, is full rank and satify LI (linear indipendence) qualification.

10.2.2.3 Compute hessian and reduced hessian

$$\nabla^2_{(x,y,z)} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -2\lambda & 0 \\ 0 & 0 & -2\lambda \end{pmatrix} \implies \begin{pmatrix} 2 & 0 & 0 \\ 0 & -(\sqrt{5}+1) & 0 \\ 0 & 0 & -(\sqrt{5}+1) \end{pmatrix}$$

reduced hessian

$$\begin{aligned} \boldsymbol{K}^{\top,a} \nabla^2_{(x,y,z)} \boldsymbol{K}^a &= \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & -(\sqrt{5}+1) & 0 \\ 0 & 0 & -(\sqrt{5}+1) \end{pmatrix} \begin{pmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} -2 & -2 \\ -(\sqrt{5}+1) & 0 \\ 0 & -(\sqrt{5}+1) \end{pmatrix} \\ &= \begin{pmatrix} 1-\sqrt{5} & 2 \\ 2 & 1-\sqrt{5} \end{pmatrix} \end{aligned}$$

Use Sylverster to check positivity

- $\boldsymbol{H}_1 = (01-\sqrt{5})$, determinant $1-\sqrt{5} < 0$ [NO]

the point is **not a minimum**.

10.2.3 Study of solution 3

$x=0, y=z=\lambda=-\frac{\sqrt{2}}{2}, \mu_1=0, \mu_2=\frac{\sqrt{2}}{2}$

10.2.3.1 Compute active constraints

$$g_1(x,y,z)=1-z=1+\frac{\sqrt{2}}{2}=\frac{2+\sqrt{2}}{2}>0, \qquad \text{(not-active)}$$
$$g_2(x,y,z)=x=0, \qquad \text{(active)}$$

Gradient of active constraints

$$\begin{pmatrix}\nabla h(x,y,z)\\ \nabla g_2(x,y,z)\end{pmatrix}=\begin{pmatrix}1 & 2y & 2z\\ 1 & 0 & 0\end{pmatrix}=\begin{pmatrix}1 & -\sqrt{2} & -\sqrt{2}\\ 1 & 0 & 0\end{pmatrix}$$

find the kernel of the active constraint

$$\begin{pmatrix}1 & -\sqrt{2} & -\sqrt{2}\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=\begin{pmatrix}z_1-\sqrt{2}(z_2+z_3)\\ z_1\end{pmatrix}$$
$$\Downarrow$$
$$\begin{pmatrix}z_1\\ z_2\\ z_3\end{pmatrix}=\begin{pmatrix}0\\ \alpha\\ -\alpha\end{pmatrix}=\underbrace{\begin{pmatrix}0\\ 1\\ -1\end{pmatrix}}_{\boldsymbol{K}^a}\alpha \tag{10.1}$$

10.2.3.2 Check qualification

Gradient of active constraint Equation 10.1 is full rank and satify LI (linear indipendence) qualification.

10.2.3.3 Compute hessian and reduced hessian

$$\nabla^2_{(x,y,z)}=\begin{pmatrix}2 & 0 & 0\\ 0 & -2\lambda & 0\\ 0 & 0 & -2\lambda\end{pmatrix}\Longrightarrow\begin{pmatrix}2 & 0 & 0\\ 0 & \sqrt{2} & 0\\ 0 & 0 & \sqrt{2}\end{pmatrix}$$

reduced hessian

$$\begin{aligned}\boldsymbol{K}^{\top,a}\nabla^2_{(x,y,z)}\boldsymbol{K}^a&=\begin{pmatrix}0 & 1 & -1\end{pmatrix}\begin{pmatrix}2 & 0 & 0\\ 0 & \sqrt{2} & 0\\ 0 & 0 & \sqrt{2}\end{pmatrix}\begin{pmatrix}0\\ 1\\ -1\end{pmatrix}\\ &=\begin{pmatrix}0 & 1 & -1\end{pmatrix}\begin{pmatrix}0\\ \sqrt{2}\\ -\sqrt{2}\end{pmatrix}\\ &=2\sqrt{2}>0\end{aligned}$$

Use Sylverster to check positivity
- $\boldsymbol{H}_1=2\sqrt{2}>0$, [OK]

the point **is a minimum**.

10.3 Example 3

Consider the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y) = y \\ \text{inequalities:} \quad & g_1(x,y,z) = x^2 + y^2 - 1 \geq 0, \\ & g_2(x,y,z) = 1 - x^2 - \frac{1}{4}y^2 \geq 0, \end{aligned}$$

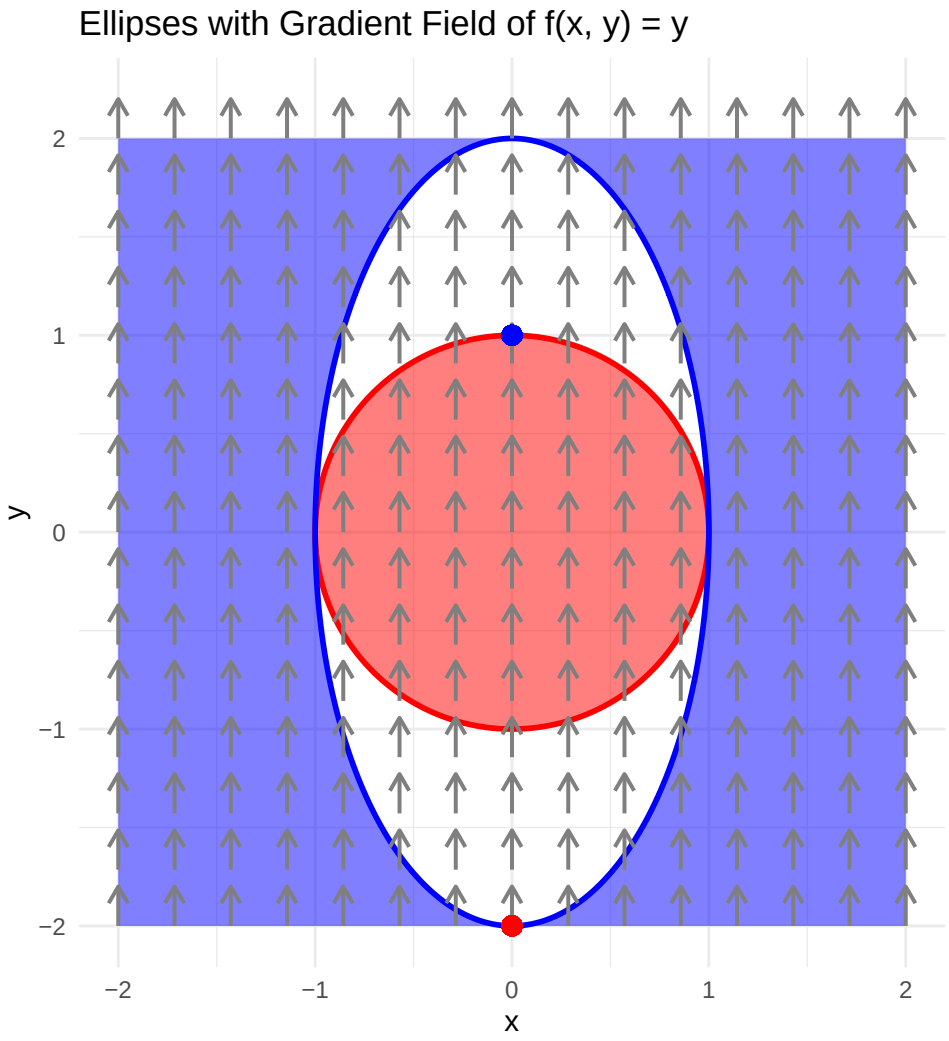


Figure 10.2: The white area is the admissible region of the problem. The red point is the constrained minima.

10.3.1 Lagrangian Function

We define the Lagrangian function as follows:

$$\mathscr{L}(x,y,\mu_1,\mu_2) = y - \mu_1(x^2 + y^2 - 1) - \mu_2\left(1 - x^2 - \frac{1}{4}y^2\right)$$

10.3.2 First-Order Conditions

The first-order conditions derived from the Lagrangian are:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial x} &= 2x(\mu_2 - \mu_1), \\
\frac{\partial \mathscr{L}}{\partial y} &= 1 + y\left(\frac{1}{2}\mu_2 - 2\mu_1\right), \\
0 &= \mu_1(x^2 + y^2 - 1), \\
0 &= \mu_2\left(1 - x^2 - \frac{1}{4}y^2\right).
\end{aligned}$$

10.3.3 Solutions of First-Order Conditions

The solutions to the first-order conditions yield the following critical points:

$$\begin{aligned}
&x = 0, \quad y = 1, \quad \mu_1 = \frac{1}{2}, \quad \mu_2 = 0, \\
&x = 0, \quad y = -2, \quad \mu_1 = 0, \quad \mu_2 = 1.
\end{aligned}$$

10.3.4 Analysis of First Candidate

10.3.4.1 Compute Active Constraints

We evaluate the constraints at the first candidate $(x, y) = (0, 1)$:

$$\begin{aligned}
g_1(x,y) &= x^2 + y^2 - 1 = 0^2 + 1^2 - 1 = 0, && \text{(active)} \\
g_2(x,y) &= 1 - x^2 - \frac{1}{4}y^2 = 1 - 0^2 - \frac{1}{4}1^2 = \frac{3}{4} > 0, && \text{(not-active)}
\end{aligned}$$

10.3.4.2 Gradient of Active Constraints

The gradient of the active constraint g_1 is computed as follows:

$$\nabla g_1(x,y) = [2x, 2y] = [0, 2].$$

10.3.4.3 Check qualification

Gradient of active constraint $\nabla g_1(x,y)$ is full rank and satify LI (linear indipendence) qualification.

10.3.4.4 Kernel of the Gradient of Active Constraints

The kernel of the gradient can be expressed as:

$$[0,2]\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 2z_2 = 0 \implies \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\boldsymbol{K}^a} \alpha.$$

10.3.4.5 Reduced Hessian

The reduced Hessian is given by:

$$\begin{aligned}
\nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) &= \begin{pmatrix} 2(\mu_2 - \mu_1) & 0 \\ 0 & \frac{1}{2}(\mu_2 - 4\mu_1) \end{pmatrix} \\
&\implies \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.
\end{aligned}$$

Evaluating the quadratic form yields:

$$\boldsymbol{K}^{a,\top} \nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) \boldsymbol{K}^a = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = -1.$$

Since the Hessian is not **semi-definite positive**, this point is not a minimum.

10.3.5 Analysis of Second Candidate

10.3.5.1 Compute Active Constraints

Next, we evaluate the constraints at the second candidate $(x,y) = (0,-2)$:

$$g_1(x,y) = x^2 + y^2 - 1 = 0^2 + (-2)^2 - 1 = 3 > 0, \quad \text{(not-active)}$$
$$g_2(x,y) = 1 - x^2 - \frac{1}{4}y^2 = 1 - 0^2 - \frac{1}{4}(-2)^2 = 0, \quad \text{(active)}$$

10.3.5.2 Gradient of Active Constraints

The gradient of the active constraint g_2 is computed as:

$$\nabla g_2(x,y) = [-2x, -\frac{1}{2}y] = [0,-1].$$

10.3.5.3 Check qualification

Gradient of active constraint $\nabla g_1(x,y)$ is full rank and satify LI (linear indipendence) qualification.

10.3.5.4 Kernel of the Gradient of Active Constraints

The kernel of the gradient can be expressed as:

$$[0,-1]\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = 2z_2 = 0 \implies \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\boldsymbol{K}^a} \alpha.$$

10.3.5.5 Reduced Hessian

The reduced Hessian is given by:

$$\nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) = \begin{pmatrix} 2(\mu_2 - \mu_1) & 0 \\ 0 & \frac{1}{2}(\mu_2 - 4\mu_1) \end{pmatrix}$$
$$\implies \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Evaluating the quadratic form yields:

$$\boldsymbol{K}^{a,\top} \nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) \boldsymbol{K}^a = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 1.$$

Since the Hessian is **definite positive**, this point is a minimum.

10.4 Example 4

Consider the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y) = x \\ \text{inequalities:} \quad & g_1(x,y,z) = x^2 + y^2 - 1 \geq 0, \\ & g_2(x,y,z) = 1 - x^2 - \frac{1}{4}y^2 \geq 0, \end{aligned}$$

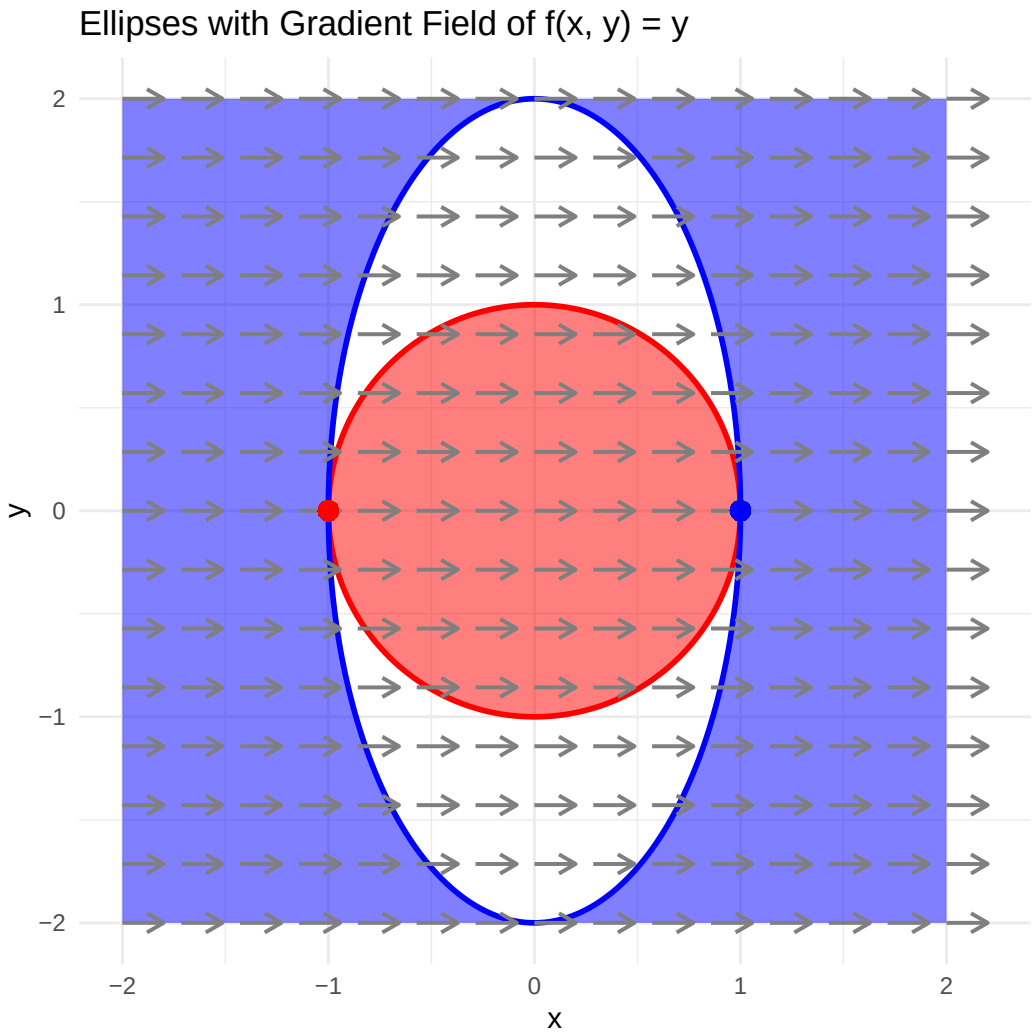


Figure 10.3: The white area is the admissible region of the problem. The red point is the constrained minima.

10.4.1 Lagrangian Function

We define the Lagrangian function as follows:

$$\mathscr{L}(x,y,\mu_1,\mu_2) = x - \mu_1(x^2+y^2-1) - \mu_2\left(1-x^2-\frac{1}{4}y^2\right)$$

10.4.2 First-Order Conditions

The first-order conditions derived from the Lagrangian are:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial x} &= 1+2x(\mu_2-\mu_1),\\
\frac{\partial \mathscr{L}}{\partial y} &= y\left(\frac{1}{2}\mu_2-2\mu_1\right),\\
0 &= \mu_1(x^2+y^2-1),\\
0 &= \mu_2\left(1-x^2-\frac{1}{4}y^2\right).
\end{aligned}$$

10.4.3 Solutions of First-Order Conditions

- **Case** $y=0$

$$\begin{aligned}
0 &= 1+2x(\mu_2-\mu_1),\\
0 &= \mu_1(x^2-1),\\
0 &= \mu_2(1-x^2).
\end{aligned}$$

  in this case if $\mu_1=\mu_2=0$ the first equation fails. Thus $x=1$ or $x=-1$
  - **Case** $x=1$ solving $1+2(\mu_2-\mu_1)=0$ we have $\mu_1 = 1/2+\mu_2$
  - **Case** $x=-1$ solving $1-2(\mu_2-\mu_1)=0$ we have $\mu_2 = 1/2+\mu_1$

  thus we have two family of solution
  - $x=1, y=0, \mu_1 = 1/2+\mu_2$ with $\mu_1 \geq 0$
  - $x=-1, y=0, \mu_2 = 1/2+\mu_1$ with $\mu_2 \geq 0$
- **Case** $y\neq 0$ in this case $\mu_2 = 4\mu_1$ and

$$\begin{cases}
0 = 1+6x\mu_1,\\
0 = \mu_1(x^2+y^2-1),\\
0 = 4\mu_1\left(1-x^2-\frac{1}{4}y^2\right).
\end{cases} \implies \mu_1 \neq 0 \implies \begin{cases}
0 = x^2+y^2-1,\\
0 = 1-x^2-\frac{1}{4}y^2.
\end{cases}$$

  but the last system has solutioon $y=0$ and $x=\pm 1$. Thus the solution must be discarded.

The solutions to the first-order conditions yield the following critical points:

$$\begin{aligned}
x=1, \quad y=0, \quad \mu_1 &= \frac{1}{2}+\mu_2,\\
x=-1, \quad y=0, \quad \mu_2 &= \frac{1}{2}+\mu_1
\end{aligned}$$

10.4.4 Analysis of First Candidate

10.4.4.1 Compute Active Constraints

We evaluate the constraints at the first candidate $(x,y)=(1,0)$:

$$\begin{aligned}
g_1(x,y) &= x^2+y^2-1 = 1^2+0^2-1 = 0, && \text{(active)}\\
g_2(x,y) &= 1-x^2-\frac{1}{4}y^2 = 1-1^2-\frac{1}{4}0^2 = 0, && \text{(active)}
\end{aligned}$$

10.4.4.2 Gradient of Active Constraints

The gradient of the active constraint g_1 and g_2 is computed as follows:

$$\begin{pmatrix} \nabla g_1(x,y) \\ \nabla g_2(x,y) \end{pmatrix} = \begin{pmatrix} 2x & 2y \\ -2x & -y/2 \end{pmatrix} \implies \begin{pmatrix} 2 & 0 \\ -2 & 0 \end{pmatrix} \tag{10.2}$$

10.4.4.3 Check qualification

Gradient of active constraint Equation 10.2 **is not** full rank and **do not** satify LI (linear indipendence) qualification.

Remark 10.1 However, by choosing the solution with $\mu_2 = 0$, we can eliminate the second constraint and still obtain the same candidate point. Thus, by removing the first inequality, the result for this point remains unchanged, and the constraint qualification holds.

10.4.4.4 Kernel of the Gradient of Active Constraints

The kernel of the gradient can be expressed as:

$$\begin{pmatrix} 2 & 0 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 2z_1 \\ -2z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\Downarrow$$
$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{K}^a} \alpha.$$

10.4.4.5 Reduced Hessian

The reduced Hessian (with $\mu_2 = 0$ and $\mu_1 = 1/2$) is given by:

$$\nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) = \begin{pmatrix} 2(\mu_2 - \mu_1) & 0 \\ 0 & \frac{1}{2}(\mu_2 - 4\mu_1) \end{pmatrix}$$
$$\implies \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Evaluating the quadratic form yields:

$$\boldsymbol{K}^{a,\top} \nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) \boldsymbol{K}^a = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{4}.$$

Since the Hessian is not **semi-definite positive**, this point is not a minimum.

10.4.5 Analysis of Second Candidate

10.4.5.1 Compute Active Constraints

Next, we evaluate the constraints at the second candidate $(x,y) = (-1,0)$:

$$g_1(x,y) = x^2 + y^2 - 1 = (-1)^2 + 0^2 - 1 = 0, \quad \text{(active)}$$
$$g_2(x,y) = 1 - x^2 - \frac{1}{4}y^2 = 1 - (-1)^2 - \frac{1}{4}(0)^2 = 0, \quad \text{(active)}$$

10.4.5.2 Gradient of Active Constraints

The gradient of the active constraint g_1 and g_2 is computed as:

$$\begin{pmatrix} \nabla g_1(x,y) \\ \nabla g_2(x,y) \end{pmatrix} = \begin{pmatrix} 2x & 2y \\ -2x & -y/2 \end{pmatrix} \implies \begin{pmatrix} 2 & 0 \\ -2 & 0 \end{pmatrix}$$

10.4.5.3 Kernel of the Gradient of Active Constraints

The kernel of the gradient can be expressed as:

$$\begin{gathered} \begin{pmatrix} 2 & 0 \\ -2 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 2z_1 \\ -2z_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ \Downarrow \\ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \alpha \end{pmatrix} = \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\boldsymbol{K}^a} \alpha. \end{gathered} \tag{10.3}$$

10.4.5.4 Check qualification

Gradient of active constraint Equation 10.3 **is not** full rank and **do not** satify LI (linear indipendence) qualification.

Remark 10.2 However, by choosing the solution with $\mu_1 = 0$, we can eliminate the first constraint and still obtain the same candidate point. Thus, by removing the first inequality, the result for this point remains unchanged, and the constraint qualification holds.

10.4.5.5 Reduced Hessian

The reduced Hessian (with $\mu_1 = 0$ and $\mu_2 = 1/2$) is given by:

$$\begin{aligned} \nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) &= \begin{pmatrix} 2(\mu_2 - \mu_1) & 0 \\ 0 & \frac{1}{2}(\mu_2 - 4\mu_1) \end{pmatrix} \\ &\implies \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}. \end{aligned}$$

Evaluating the quadratic form yields:

$$\boldsymbol{K}^{a,\top} \nabla^2_{(x,y)} \mathscr{L}(x,y,\mu_1,\mu_2) \boldsymbol{K}^a = \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{4}.$$

Since the Hessian is **definite positive**, this point is a **minimum**.

10.5 Example 5

Consider the problem

$$\begin{aligned} \text{minimize:} \quad & f(x,y) = 6 + 2y - x^2 \\ \text{equality:} \quad & h(x,y) = y - x^2 \end{aligned}$$

10.5.1 Lagrangian function

$$\mathscr{L}(x,y,\lambda) = 6 + 2y - x^2 - \lambda(y - x^2)$$

10.5.2 First-Order Conditions

$$\frac{\partial \mathscr{L}}{\partial x} = 2x(\lambda - 1) = 0$$
$$\frac{\partial \mathscr{L}}{\partial y} = 2 - \lambda = 0$$
$$0 = y - x^2$$

10.5.3 Solution First-Order Conditions

- $\lambda = 2$
- $x = 0$
- $y = 0$

10.5.4 Analisys of The Candidate

Gradient of the constraints

$$\nabla(y - x^2) = \begin{pmatrix} -2x & 1 \end{pmatrix}, \quad \Longrightarrow \quad \begin{pmatrix} 0 & 1 \end{pmatrix}$$

10.5.4.1 Check Qualification

Only one vector non zero is linearly independent. The constraint is qualified.

10.5.4.2 Kernel of the Gradient of Active Constraints

$$\begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = z_2 = 0, \quad \Longrightarrow \quad \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}_{\boldsymbol{K}^a} \alpha$$

10.5.4.3 Reduced Hessian

$$\nabla_{(x,y)} \mathscr{L}(x, y, \lambda) = \begin{pmatrix} \lambda - 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \Longrightarrow \quad \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\boldsymbol{K}^{a,\top} \nabla_{(x,y)} \mathscr{L}(0, 0, 2) \boldsymbol{K}^a = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = (1)$$

the reduced Hessian is definite positive, the point is a minimum.

III Minimization of a Functional

11 Introduction 113
11.1 Analogies with linear algebra 116
11.2 First Variation 118
11.3 Integrals and Limits 119

12 Fundamental lemma of the calculus of variations 123

13 Euler Lagrange equation 129
13.1 Pendulum example 129
13.2 General formulation 131
13.3 Application to the brachistochrone problem . 133

14 Boundary condition for a functional . 137
14.1 Free boundary conditions 139

15 Functional minimization with generic constraints 143
15.1 Heuristic deduction of general boundary condition 145

16 Minimization with Integral Constraints 151
16.1 Approximate solution using finite differences . 153
16.2 Heuristic deduction of integral constraint for functional 155

17 Minimization with algebraic (path) constraints 159
17.1 Heuristic deduction of path constraints 162

18 Summary 165
18.1 Spaces 165
18.2 General Problem 167

19 Worked exercizes 169
19.1 Exercize 1 169
19.2 Exercize 3 172

This part develops the passage from finite-dimensional constrained minimization to variational problems. The key output is an Euler–Lagrange equation together with the boundary and integral conditions that complete it.

Learning objectives

After this part, you should be able to compute a first variation, derive the Euler–Lagrange equation, and identify the natural boundary, integral, and path constraints associated with a variational problem.



11. Introduction

A classical problem in minimization is the brachistrochrone, i.e., find the curve (meaning determining the fastest trajectory) to move from a point A to a (lower) point B considering that the only action is the one of the gravity.

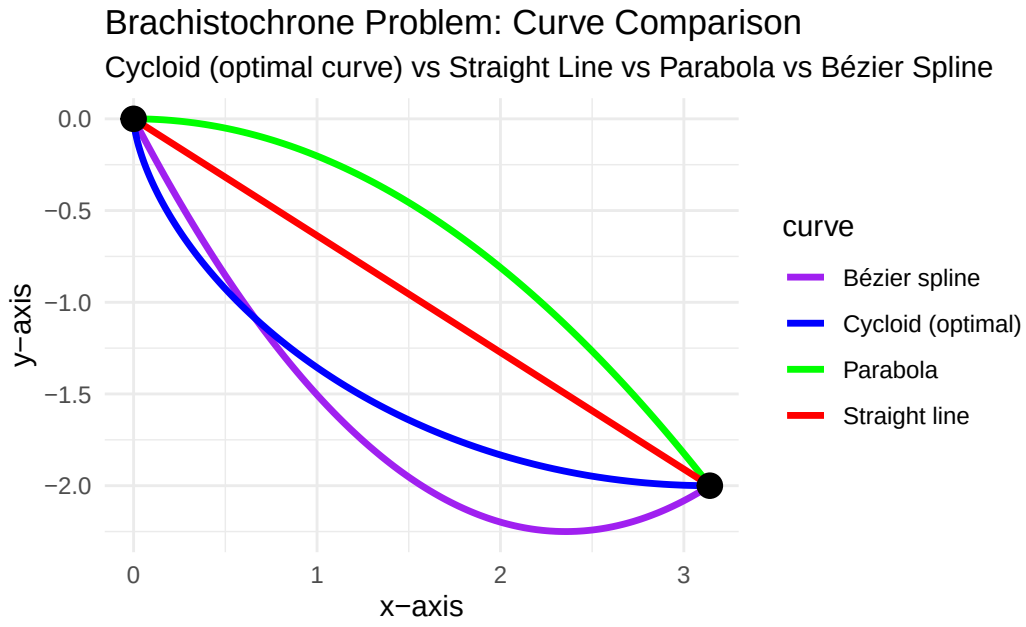


Figure 11.1: example of trajectories that an object can follow for moving from point A to B.

Mathematically the problem is:

Definition 11.1 — Brachistrochrone problem. Determine the curve

$$\mathscr{C} : [a,b] \to \mathbb{R}$$

such that

$$\mathscr{C}(a) = y_a \quad \text{and} \quad \mathscr{C}(b) = y_b$$

that minimize the function $T(\mathscr{C})$ which represent the time to travel from $A = (a, y_a)$ to $B = (b, y_b)$. Thus, given $T(\mathscr{C})$ the problem become:

$$\text{minimize } T(\mathscr{C}) \qquad \text{for all possible curves } \mathscr{C}$$

A **function** takes a number as input and returns a number, like $f(x) = xe^x$. A **functional** is instead something that takes as input a function and returns a number and an example is

$$\mathscr{F}(x) = \int_a^b x(t)\,\mathrm{d}t$$

where x is a function. Considering for example $x(t) = t^2$ the previous functional becomes

$$\mathscr{F}(x) = \int_a^b t^2\,\mathrm{d}t = \frac{t^3}{3}\Big|_a^b = \frac{b^3 - a^3}{3}$$

Another example of functional can be

$$\mathscr{G}(z) = z'(0)\int_0^1 z^2(t)\,\mathrm{d}t$$

and this expression can be evaluated for every generic function $z(t)$.

The question connected to Definition 11.1 is how to define a *minimum* for a functional; in order to do so we have to firstly understand how to minimize a simple function and in particular given a function

$$f : A \subseteq \mathbb{R} \mapsto \mathbb{R}$$

has a minimum in the point $x^\star$ if

$$f(x^\star) \leq f(x) \qquad \forall x \in A$$

Similarly given a function $\mathscr{F}(x)$ a *point* $x^\star$ (that in reality is a function) is a minimum if

$$\mathscr{F}(x^\star) \leq \mathscr{F}(x) \qquad \forall x \in ?$$

In this case we have to specify the *class* (functional space) of functions we are considering, as example x can be a function is continuous in the domain $[a,b]$ and so we can define

$$x \in C([a,b]) = \left\{g : [a,b] \mapsto R \mid g \text{ is continuous}\right\}$$

In general changing the *domain* of the functional $\mathscr{F}$ may change the problem.

▪ **Example 11.1 — domain change.** Considering the function $f(x) = x^2 + 2$, it's roots can be computed if we allow complex solutions $z \in \mathbb{C}$ (and in fact $z = \pm \mathrm{i}\sqrt{2}$), while if the domain of the solution is the real set $z \in \mathbb{R}$ no solutions exists.

Considering now the functional $\mathscr{F}$ defined as

$$\mathscr{F}(x) = \int_{-1}^1 \left(x(t) - |t|\right)^2 \mathrm{d}t$$

The $|t|$ introduce a cuspid in $t = 0$ that cannot be derived. If we minimize the functional in the continuous interval $x \in C([-1,1])$, the solution is the function $x^\star(t) = |t|$, in fact

$$\mathscr{F}(x^\star) = \int_{-1}^1 \left(|t| - |t|\right)^2 \mathrm{d}t = \int_{-1}^1 0\,\mathrm{d}t = 0$$

Choosing any other continuous function will result in a functional with a positive value.

Considering now to minimize the function in the domain of functions with continuous derivatives, and so $x \in C^1([-1,1])$. In this case $|t| \notin C^1([-1,1])$ (due to the cuspid). An approximation of the function $|t|$ that is continuous with also continuous first derivative is the function

$$x_\varepsilon(t) = \begin{cases} t & t \geq \varepsilon \\ \frac{1}{2}\left(\frac{t^2}{\varepsilon} + \varepsilon\right) & -\varepsilon < t < \varepsilon \\ -t & t \leq -\varepsilon \end{cases}$$

By pushing the limit $\varepsilon \to 0$ we can have the function that minimize the functional $\mathscr{F}$ in the domain $C^1([-1,1])$. In fact we can see that the functional of x_ε becomes

$$\mathscr{F}(x_\varepsilon) = \int_{-\varepsilon}^{\varepsilon} \left(\frac{t^2}{2\varepsilon} + \frac{\varepsilon}{2} - |t|\right)^2 \mathrm{d}t = \frac{\varepsilon^3}{10}$$

■

11.0.0.1 Some examples of space of functions

- Functions $f : [a,b] \mapsto \mathbb{R}$ continuous
- Functions $f : [a,b] \mapsto \mathbb{R}$ which are C^k (continuous with derivative up to order k).
- Functions $f : [a,b] \mapsto \mathbb{R}$ piecewise continuos

11.0.0.2 Less trivial space of functions examples

- Space $L_1([a,b])$ of functions $f : [a,b] \mapsto \mathbb{R}$ which are module integrable

$$\int_a^b |f(x)|\,\mathrm{d}t < \infty$$

  [the integral must exists and is finite]
- Space $L_2([a,b])$ of functions $f : [a,b] \mapsto \mathbb{R}$ which are square integrable

$$\int_a^b |f(x)|^2\,\mathrm{d}t < \infty$$

  [the integral must exists and is finite]

■ **Example 11.2** Let $f : [0,1] \mapsto \mathbb{R}$, $f(x) = 1/x$ is not $L_1([0,1])$ in fact

$$\int_0^1 \frac{1}{x}\mathrm{d}x \to \infty$$

The function $g : [0,1] \mapsto \mathbb{R}$, $g(x) = 1/\sqrt{x}$, is in $L_1([0,1])$ in fact

$$\int_0^1 \frac{1}{\sqrt{x}}\mathrm{d}x = \left[2\sqrt{|x|}\right]_{x=0}^{x=1} = 2 < \infty$$

however this function $g(x)$ is not in $L_2([0,1])$, in fact

$$\int_0^1 g(x)^2\mathrm{d}x \int_0^1 \frac{1}{(\sqrt{x})^2}\mathrm{d}x = \int_0^1 \frac{1}{|x|}\mathrm{d}x \to \infty$$

the function $h(x) = x^{-1/4}$ is in $L_2([0,1])$, in fact

$$\begin{aligned} \int_0^1 h(x)^2\mathrm{d}x &= \int_0^1 (x^{-1/4})2\mathrm{d}x \\ &= \int_0^1 \frac{1}{\sqrt{x}}\mathrm{d}x \\ &= \left[2\sqrt{|x|}\right]_{x=0}^{x=1} \\ &= 2 < \infty \end{aligned}$$

■

11.1 Analogies with linear algebra

To solve the problem of minimizing a functional we can see some relations with the linear algebra. The domain of the functional can be in fact see as a **functional space** $\mathbb{V}$ analogous to the vectorial one which the vectors are represented by the functions and the scalars are represented by real values. With this definition example of functional spaces might be

$$\mathbb{V}_1 = \{f : [0,1] \mapsto \mathbb{R} \text{ continuous}\}$$
$$\mathbb{V}_2 = \{f : [a,b] \mapsto \mathbb{R} \text{ such that } f \in C^k([a,b])\}$$

with $a, b \in \mathbb{R}$ and $k \in \mathbb{N}$.

Given in fact two function $f, g \in \mathbb{V}$ that are member of the same functional space, we can see that each linear combination of the functions determine a function h that's still in the vectorial space:

$$h(x) = \alpha f(x) + \beta g(x) \in \mathbb{V} \qquad \forall \alpha, \beta \in \mathbb{R},\ f, g \in \mathbb{V}$$

As example let's consider the two continuous function

$$f(x) = x^2 \quad \text{and} \quad g(x) = \sin x$$

then we can clearly see that the function

$$h(x) = 2f(x) + \frac{1}{3}g(x) = 2x^2 + \frac{1}{3}\sin x$$

is still continuous. This in general means that the functional space is *closed* respect to the operations of function summation and multiplication by a scalar.

11.1.0.1 Scalar product

In linear algebra given two vector $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ it exists the bilinear operator scalar product $\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle = \boldsymbol{v}_1 \cdot \boldsymbol{v}_2$ that satisfy the following rules:

$$\langle \alpha \boldsymbol{v} + \beta \boldsymbol{w}, \boldsymbol{z} \rangle = \langle \boldsymbol{z}, \alpha \boldsymbol{v} + \beta \boldsymbol{w} \rangle = \alpha \langle \boldsymbol{v}, \boldsymbol{z} \rangle + \beta \langle \boldsymbol{w}, \boldsymbol{z} \rangle$$

for all $\alpha, \beta \in \mathbb{R}$, $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z} \in \mathbb{V}$ and

$$\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0 \quad \text{and} \quad \langle \boldsymbol{v}, \boldsymbol{v} \rangle = 0 \quad \Leftrightarrow \quad \boldsymbol{v} = \boldsymbol{0}$$

Also in the functional space $\mathbb{V}$ can exists definitions of product scalar such the one here presented:

$$\langle f, g \rangle = \int_0^1 f(x) g(x)\, \mathrm{d}x \tag{11.1}$$

> **Note**
>
> This is not the lonely function that can serve as product scalar of function and in this case the relation holds for the functional space
>
> $$\mathbb{V} = \{f : [0,1] \mapsto \mathbb{R} \text{ integrable}\}$$

In this case we can prove that this definitions meets the requirements stated for the scalar product considering the linear properties of the integrals as follows:

$$\begin{aligned}\langle \alpha f(\cdot) + \beta g(\cdot), h\rangle &= \int_0^1 \big(\alpha f(x) + \beta g(x)\big) h(x)\,\mathrm{d}x \\ &= \alpha \int_0^1 f(x)h(g)\,\mathrm{d}x + \beta \int_0^2 g(x)h(x)\,\mathrm{d}x \\ &= \alpha\langle f,h\rangle + \beta\langle g,h\rangle\end{aligned}$$

and also

$$\langle f,f\rangle = \int_0^1 f(x)f(x)\,\mathrm{d}x = \int_0^1 f^2(x) \geq 0$$

and in particular we can see that the scalar product $\langle f,f\rangle$ will give as result 0 if and only if the function f is identically null, so such that $f(x) = 0$ for all x in it's domain (in this case $[0,1]$).

11.1.0.2 Norm

In linear algebra it's also defined the norm of a vector as

$$\|\boldsymbol{v}\| := \sqrt{\boldsymbol{v}\cdot\boldsymbol{v}}$$

and this expression is used to compute, from a vector, a single positive value (and in particular $\|\boldsymbol{v}\| = 0$ if and only if $\boldsymbol{v} = \mathbf{0}$). This definition also holds for the functional space and considering the scalar product defined (as example) in equation Equation 11.1 we can see that one definition of norm for function can be the one

$$\|f(x)\| = \sqrt{\langle f(x), f(x)\rangle} = \sqrt{\int_0^1 f^2(x)\,\mathrm{d}x}$$

In this case we can clearly see that the norm $\|f\| \geq 0$ is always positive defined (and is zero only in the case on which the function f is identically null).

■ **Example 11.3 — Spaces of Functions.** A *space of functions* can be thought of as a set of functions satisfying certain properties.

For example, one such space is the set of continuous functions

$$f : [a,b] \to \mathbb{R}, \quad \text{with } f \in C([a,b]).$$

More generally, we may consider the space of k-times continuously differentiable functions

$$f \in C^k([a,b]), \quad k \in \mathbb{N}.$$

Similarly, we can work with spaces of piecewise continuous functions.

A less trivial example is the space of *integrable* functions, namely

$$f : [a,b] \to \mathbb{R}, \quad \text{such that } \int_a^b |f(x)|\,dx < \infty.$$

This is precisely the space of functions that are absolutely integrable on $[a,b]$.

In general we define as $L^p([a,b])$ the space of p-integrable functions, so such that

$$f \in L^p([a,b]) \qquad \Longleftrightarrow \qquad \int_a^b |f(x)|^p\,\mathrm{d}x < \infty$$

■

11.2 First Variation

Let $\mathscr{J} : \mathbb{V} \to \mathbb{R}$ be a functional defined on a function space $\mathbb{V}$.
To study conditions for $\mathscr{J}$ to attain a minimum, we introduce the notion of the **derivative of a functional**, called the **first variation** (or *directional derivative*).

Given a function $x \in \mathbb{V}$ and a perturbation $\eta \in \mathbb{V}$, we consider the **variation**

$$x_\alpha(\cdot) = x(\cdot) + \alpha\eta(\cdot), \qquad \alpha \in \mathbb{R}.$$

Defining the auxiliary function

$$g(\alpha) = \mathscr{J}\big(x(\cdot) + \alpha\eta(\cdot)\big).$$

then the **first variation** of $\mathscr{J}$ at x in the direction η is

$$\delta\mathscr{J}(x;\eta) = g'(0). \tag{11.2}$$

11.2.1 Equivalent Forms

From the definition of the derivative, we can write

$$\delta\mathscr{J}(x;\eta) = \lim_{\alpha\to 0} \frac{\mathscr{J}(x+\alpha\eta) - \mathscr{J}(x)}{\alpha}.$$

Equivalently, using the small-o notation,

$$\mathscr{J}(x+\alpha\eta) = \mathscr{J}(x) + \alpha\,\delta\mathscr{J}(x;\eta) + o(\alpha).$$

▪ **Example 11.4 — directional derivative of a functional — first variation.** Given the functional

$$\mathscr{F}(x) = \int_0^1 \big(x^2(t) + 1\big)\,\mathrm{d}t \; + x(1)$$

it's first variation respect to a generic direction $d(t)$ can be calculated by firstly determining the associated function $g : \mathbb{R} \mapsto \mathbb{R}$ defined as

$$\begin{aligned} g(\alpha) &= \mathscr{F}\big(x(\cdot) + \alpha\, d(\cdot)\big) \\ &= \int_0^1 \big(x(t) + \alpha\, d(t)\big)^2 + 1\,\mathrm{d}t + x(1) + \alpha\, d(1) \end{aligned}$$

As reported in equation Equation 11.2, the first variation of the functional $\mathscr{F}(x)$ can be regarded as the derivative of $g(\alpha)$ respect to α evaluated for $\alpha = 0$; the first step is so determining

$$\begin{aligned} g'(\alpha) &= \frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) \\ &= \frac{\mathrm{d}}{\mathrm{d}\alpha} \int_0^1 \big(x(t) + \alpha\, d(t)\big)^2 + 1\,\mathrm{d}t \\ &+ \frac{\mathrm{d}}{\mathrm{d}\alpha}\big(x(1) + \alpha\, d(1)\big) \end{aligned}$$

Assuming that $\frac{\mathrm{d}}{\mathrm{d}\alpha}\int = \int \frac{\mathrm{d}}{\mathrm{d}\alpha}$ (operation that cannot always be performed) then the derivative of $g(\alpha)$ can be regarded as

$$\begin{aligned} g'(\alpha) &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big(x(t) + \alpha\, d(t)\Big)^2 \mathrm{d}t + \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big(x(1) + \alpha\, d(1)\Big) \\ &= \int_0^1 2\big(x(t) + \alpha\, d(t)\big) d(t)\,\mathrm{d}t + d(1) \end{aligned}$$

Evaluating this expression for $\alpha = 0$ then the first variation of $\mathscr{F}(x)$ with direction $d(t)$ becomes

$$\delta\mathscr{F}(x;d) = g'(0) = \int_0^1 2xd\,\mathrm{d}t + d(1)$$

■

11.3 Integrals and Limits

In the computation of variations it is often assumed that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \circ \int_a^b \quad = \quad \int_a^b \circ \frac{\mathrm{d}}{\mathrm{d}\alpha}.$$

However, this property does **not** hold in general: some regularity conditions on the functions involved are required.
To better understand the issue, let us first examine a simpler problem.

11.3.1 Commuting Limit and Integral

Since a derivative is defined as a limit of a quotient, the question reduces to whether the operations of **limit** and **integration** can always be interchanged:

$$\lim \circ \int \quad \stackrel{?}{=} \quad \int \circ \lim.$$

Concretely, let $\{f_n\}_{n=1}^{\infty}$ be a sequence of functions such that

$$f(x) = \lim_{n\to\infty} f_n(x), \qquad \forall x \in \mathbb{R}.$$

We ask whether

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f_n(t)\,dt \stackrel{?}{=} \int_{-\infty}^{\infty} \lim_{n\to\infty} f_n(t)\,dt = \int_{-\infty}^{\infty} f(t)\,dt.$$

11.3.2 A Counterexample

Consider the sequence

$$f_n(x) = \begin{cases} 0, & x \le 0, \\ nx, & 0 < x \le \frac{1}{n}, \\ 2 - nx, & \frac{1}{n} < x \le \frac{2}{n}, \\ 0, & x > \frac{2}{n}. \end{cases}$$

Each f_n is supported in the interval $(0, 2/n)$ and takes the shape of a triangular "bump" of base $2/n$ and height 1.

For every fixed $x \in \mathbb{R}$ we have

$$\lim_{n\to\infty} f_n(x) = 0,$$

so the pointwise limit is the zero function:

$$f(x) \equiv 0.$$

However,

$$\int_{-\infty}^{\infty} f_n(t)\,dt = \text{area of the triangle } = \frac{1}{2} \cdot \frac{2}{n} \cdot 1 = \frac{1}{n}.$$

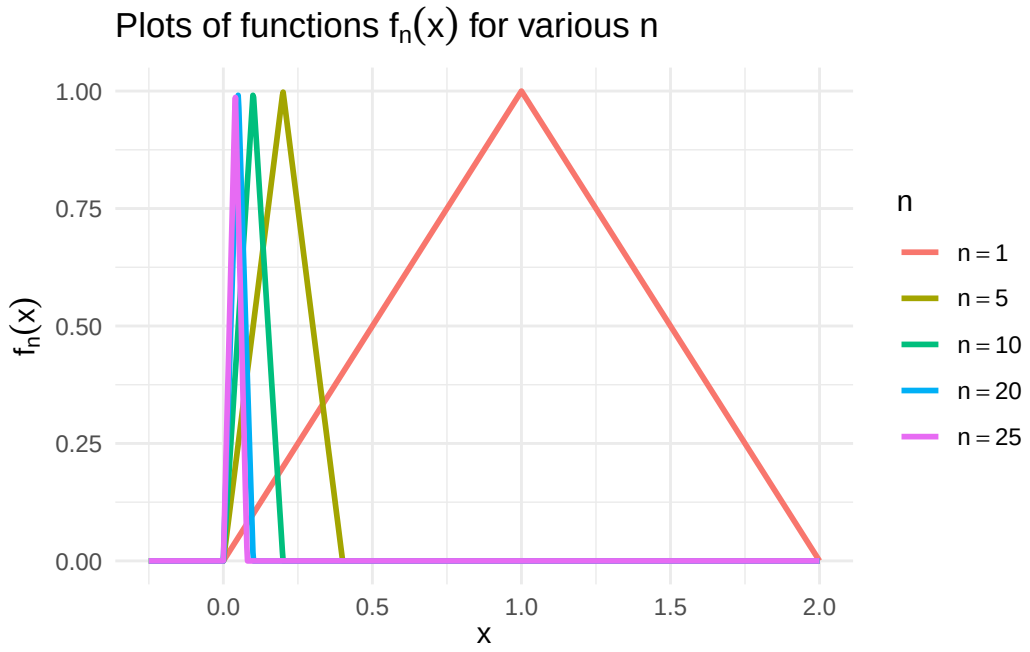


Figure 11.2

or in more detail

$$\begin{aligned}\int_{-\infty}^{\infty} f_n(x)\,\mathrm{d}x &= \int_0^{1/n} nx\,\mathrm{d}x + \int_{1/n}^{2/n} 2-nx\,\mathrm{d}x \\ &= \left[n\frac{x^2}{2}\right]_{x=0}^{x=1/n} + \left[2x - n\frac{x^2}{2}\right]_{x=1/n}^{x=2/n} \\ &= \left[n\frac{(1/n)^2}{2} - n\frac{(0)^2}{2}\right] + \left[2\frac{2}{n} - n\frac{(2/n)^2}{2} - 2\frac{1}{n} + n\frac{(1/n)^2}{2}\right] \\ &= \frac{1}{2n} + \frac{4}{n} - \frac{2}{n} - \frac{2}{n} + \frac{1}{2n} = \frac{1}{n}\end{aligned}$$

Thus

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} f_n(t)\,dt = 0, \qquad \int_{-\infty}^{\infty} f(t)\,dt = 0,$$

so in this case the interchange works.
Consider now the functions

$$g_n(x) = nf_n(t) = \begin{cases} 0 & x \le 0 \\ n^2x & 0 < x \le \frac{1}{n} \\ 2n - n^2x & \frac{1}{n} < x \le \frac{2}{n} \\ 0 & x > \frac{2}{n} \end{cases}$$

it is clear that

$$\lim_{n\to\infty} g_n(x) = 0, \qquad \forall x \in \mathbb{R}$$

and

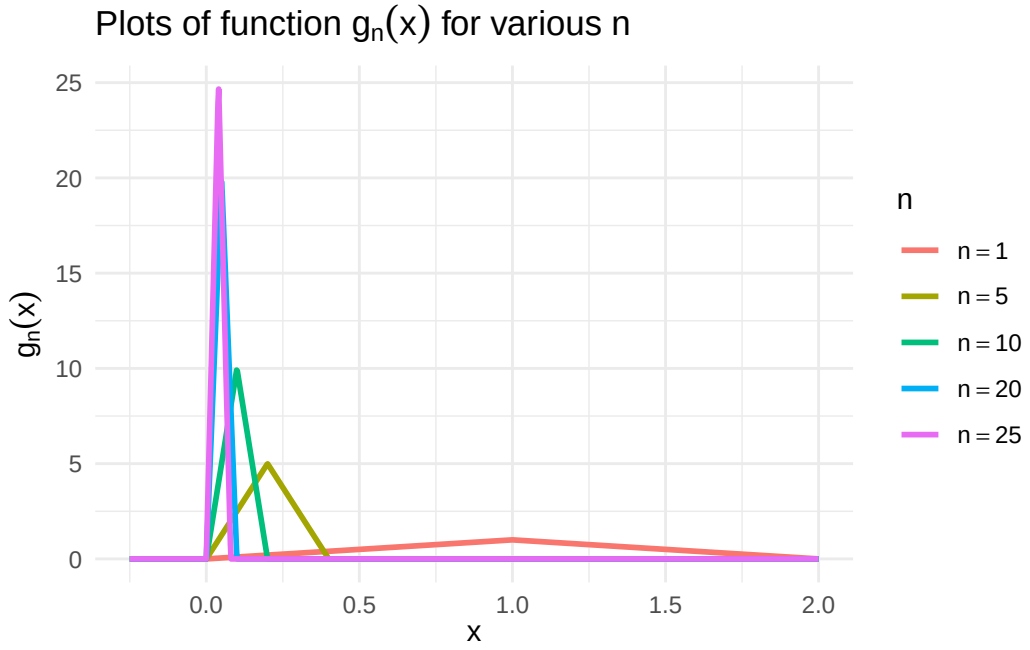


Figure 11.3

$$\int_{-\infty}^{\infty} g_n(x)\,\mathrm{d}x = \int_0^{1/n} n^2 x\,\mathrm{d}x + \int_{1/n}^{2/n} 2n - n^2 x\,\mathrm{d}x$$
$$= \frac{1}{2} + \frac{1}{2} = 1$$

and thus

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g_n(x)\mathrm{d}x = \lim_{n\to\infty} 1 = 1$$
$$\int_{-\infty}^{\infty} \lim_{n\to\infty} g_n(x)\mathrm{d}x = \int_{-\infty}^{\infty} 0\mathrm{d}x = 0$$

and in this case limit and integral **cannot** be exchanged.

Thus in general, when limits, derivative and integral are exhanged we **shuold** check if the opearation is possibile.



12. Fundamental lemma of the calculus of variations

Considering the minimization of a a function

$$f : A \subseteq \mathbb{R}^n \mapsto \mathbb{R}$$

it has been shown that the first order necessary condition for a point $\boldsymbol{x}^\star \in A$ to be a minimum point is that it's gradient $\nabla f(\boldsymbol{x}^\star)$ must be null. Similarly in calculus of variation it's proven that the **first order necessary condition** for the **optimality** of the solution is that the first variation of the functional $\mathscr{J}$ must be

$$\delta \mathscr{J}(x;\eta) = 0$$

for every *admissible perturbation* $\eta(t)$.

In particular we define a **perturbation** $\eta \in \mathbb{V}$ **admissible** for the functional $\mathscr{J} : A \times A \mapsto \mathbb{R}$ respect to the function $x(t)$ if it happens that $x(\cdot) + \alpha\eta(\cdot) \in A$ for all value α *sufficiently close to* 0.

Lemma 12.1 — Fundamental lemma of the calculus of variations. Given a function

$$f : [a,b] \mapsto \mathbb{R}, \qquad \text{(piecewise continuous)}$$

such that the integral satisfy

$$\int_a^b f(x)\, g(x)\, \mathrm{d}x = 0 \tag{12.1}$$

for all $g : [a,b] \mapsto \mathbb{R}$ smooth, i.e. $g \in C^\infty([a,b])$, with

$$g^{(k)}(a) = g^{(k)}(b) = 0, \quad \forall k = 0,1,2,3,\ldots$$

then $f(x)$ is **identically null**.

12.0.0.1 Sign permanence

In order to later proof the fundamental lemma yet described, we have to remark the *sign permanence* that states:

Lemma 12.2 Given a function $f : [a,b] \mapsto \mathbb{R}$ continuous such that $f(c) > 0$, then there exists a $\delta > 0$ such that

$$f(x) \geq \frac{f(c)}{2} \qquad \forall x \in [c-\delta, c+\delta]$$

Proof. The proof of this lemma can be done defining the parameter $\varepsilon = f(c)/2$; based on the assumption that $f(x)$ is continuous, then exists a $\delta > 0$ such that:

$$|f(x) - f(c)| \leq \varepsilon = \frac{f(c)}{2} \qquad \forall |c-x| \leq \delta$$

Expanding the module operation the inequality becomes:

$$\begin{aligned} -\frac{f(c)}{2} = -\varepsilon \leq f(x) - f(c) \leq \varepsilon = \frac{f(c)}{2} \\ \implies \underbrace{-\frac{f(c)}{2} + f(c) \leq f(x)}_{\frac{f(c)}{2} \leq f(x)} \leq f(c) + \frac{f(c)}{2} \end{aligned}$$

and so we prove the lemma of sign permanence. □ ■

12.0.0.2 Preparation of the proof of the fundamental lemma

The proof of the fundamental lemma of the calculus of variation can be done by contradiction.

Let consider the integral Equation 12.1 and for the given function f, exists a point c such that $f(c) > 0$ (in general the proof can be done for $f(c) \neq 0$). From the sign permanence lemma we can state that exists an interval $[c-\delta, c+\delta]$ such that $f(x) \geq \frac{f(c)}{2} \geq 0$. Determining the function g as

$$g(x) = \begin{cases} \dfrac{x-(c-\delta)}{\delta} & x \in [c-\delta, c] \\ \dfrac{c-x+\delta}{\delta} & x \in [c, c+\delta] \\ 0 & \text{otherwise} \end{cases}$$

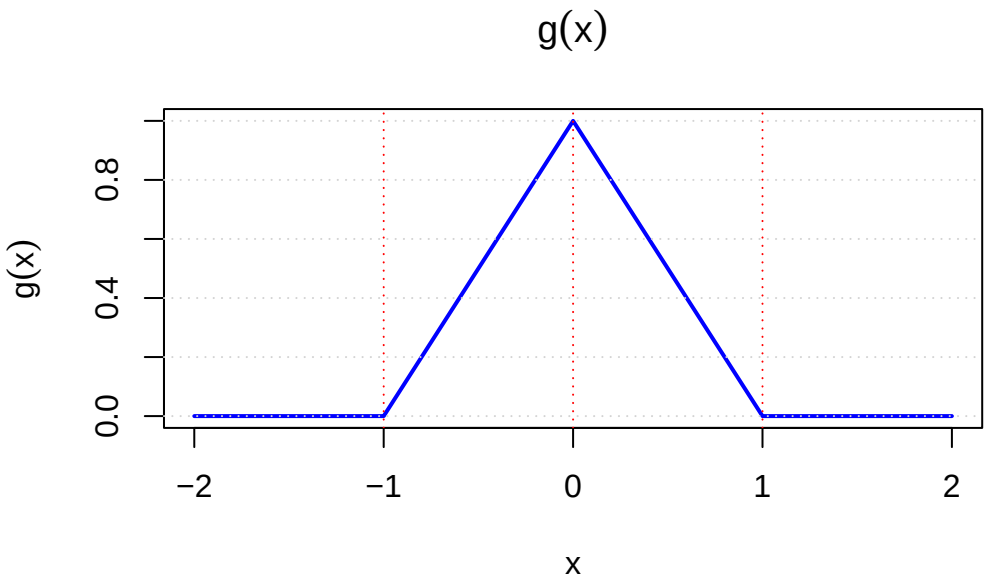


Figure 12.1: Example of function $g(x)$ with $c = 0$ and $\delta = 1$.

Note that this function is continuous but $g \notin C^\infty$ (later will be described a function that presents this feature). With this definition the integral $\int_a^b fg\,dx$ becomes

$$\int_{c-\delta}^{c+\delta} f(x)g(x)\,\mathrm{d}x \geq \frac{f(c)}{2}\int_{c-\delta}^{c+\delta} g(x)\,\mathrm{d}x \geq \frac{f(c)}{2}\frac{\delta}{2} > 0$$

We can clearly see that the integral $\int_a^b fg\,\mathrm{d}x$ has a positive real value grater than zero that's (considering that f has at least on point c such that $f(c) > 0$) in contraposition with the initial request that the integral should have been zero evaluated.

To create a direction function g that's in the set C^∞ we can use the function h(for whose those condition is demonstrated) as in Figure 12.2 defined as:

$$h(t) = \begin{cases} \mathrm{e}^{-1/t} & t > 0 \\ 0 & t \leq 0 \end{cases} \tag{12.2}$$

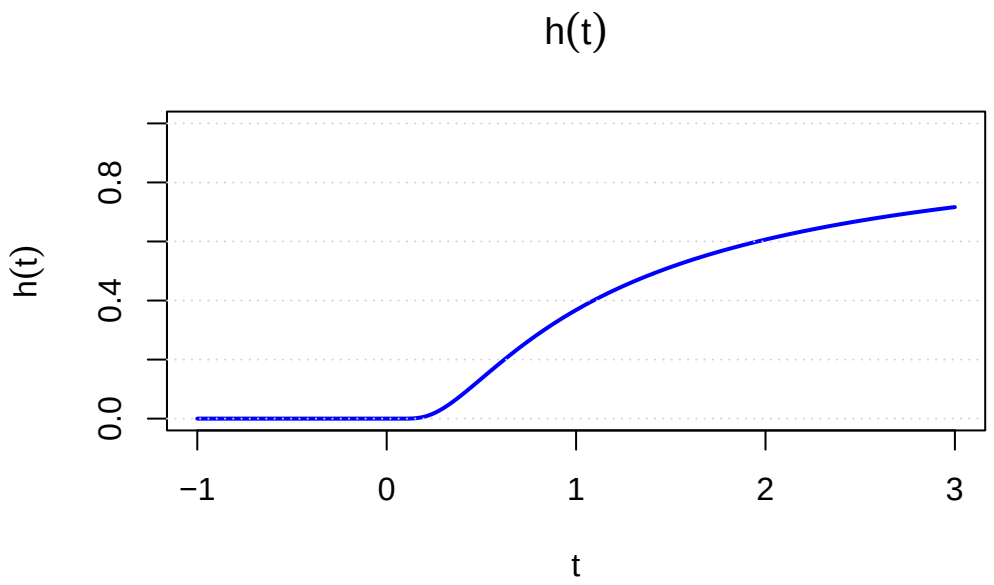


Figure 12.2: Representation of $h(t)$ defined in Equation 12.2

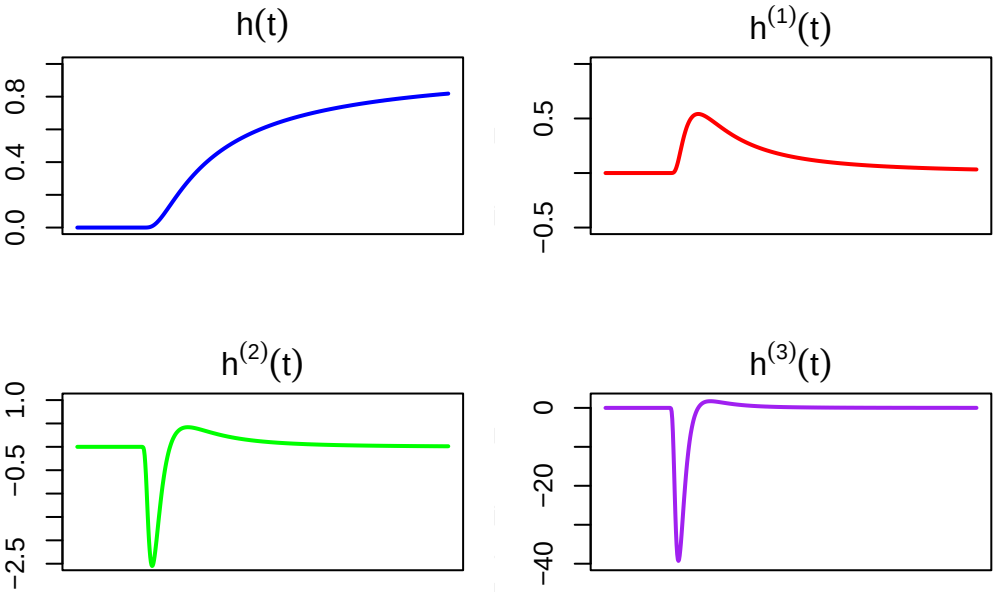


Figure 12.3: Function $h(t)$ defined in Equation 12.2 with its first three derivatives.

A way to create a function that presents a *bell shape* in the range $[0,1]$ is by computing

$$g(t) := h(t)h(1-t)$$

how's graph in the range $[0,1]$ is similar to the one shown in figure Figure 12.4.

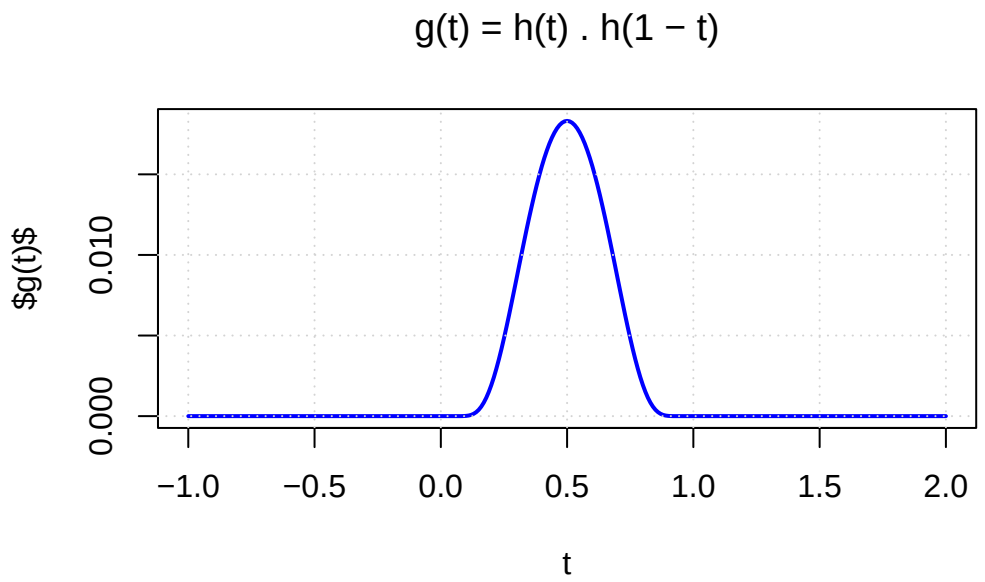


Figure 12.4: Eepresentation of the function $g(t) := h(t)h(1-t)$, where $h(t)$ is defined in equation Equation 12.2

In particular, to demonstrate the fundamental lemma of the calculus of variations, we need to rescale the function g so that it has a bell shape centered at the point c with a *bell width* δ. We consider the following formulation:

$$g(t) = K\,h\left(\frac{t-c+\delta}{2\delta}\right) h\left(\frac{\delta+c-t}{2\delta}\right)$$

where the constant $K \in \mathbb{R}$ is chosen such that $g(c) = 1$ or $\int_a^b g(x)\,\mathrm{d}x = 1$.
In this case, we observe that:

- $g \in C^\infty(\mathbb{R})$,
- $g(x) \geq 0$ for all values of x, and specifically,
- $g(x) = 0$ for all $x \notin [c-\delta, c+\delta]$.

Lemma 12.1 can now be proven by contradiction.

Proof. As in the previous case, let us consider a function $f(x)$ that is not identically zero. For this argument, we assume there exists at least one point c where $f(c) > 0$. By the sign permanence theorem, we can assert that:

$$f(x) \geq \frac{f(c)}{2} \qquad \forall x \in [c-\delta, c+\delta].$$

This inequality indicates that the function $f(x)$ remains bounded away from zero within the interval $[c-\delta, c+\delta]$.

Next, we examine the original integral $\int_a^b fg\,\mathrm{d}x$. Given our previous findings, we can conclude that this integral is not equal to zero. Specifically, we have:

$$\int_a^b f(x)g(x)\,\mathrm{d}x \geq \frac{f(c)}{2}\int_a^b g(x)\,\mathrm{d}x > 0.$$

Here, the term $\frac{f(c)}{2}$ is positive, and since $g(x)$ is non-negative for all x, the product $\int_a^b f(x)g(x)\,\mathrm{d}x$ yields a positive value. This result constradict the assumption that $f(x)$ in not identically zero. □ ■



13. Euler Lagrange equation

The **Euler Lagrange equation** is the generalization of the minimum action principle that's used in physics.

13.1 Pendulum example

Considering as a practical example the motion of a pendulum of a mass m that's free to oscillate in respect to a pivot point using a rope of length l, the kinetic energy T and the potential term V of the system can be expressed as

$$T = \frac{1}{2}mv^2, \qquad V = mgy.$$

Considering θ as the angle that the rope determines with the vertical axis, then the we can rewrite the energies as functions of the angular position θ and velocity $\dot{\theta}$ as

$$T(\theta, \dot{\theta}) = \frac{1}{2}ml^2\dot{\theta}^2, \qquad V(\theta) = -mgl\sin\theta.$$

To solve the dynamic equation $\theta(t)$ of the mechanism we can compute the lagrangian L of the system defined as

$$L(\theta, \dot{\theta}) = T(\theta, \dot{\theta}) - V(\theta) = \frac{m}{2}l^2\dot{\theta}^2 + lmg\cos\theta$$

As law that's analyzed in mechanics physics we can state that the solution of the dynamics of the system is the one the function that minimize the **action** $\mathscr{A}$ of the system defined as

$$A(\theta) = \int_{t_0}^{t_1} L(\theta, \dot{\theta})\,\mathrm{d}t \tag{13.1}$$

where the values $\theta(t_0) = \theta_0$ and $\theta(t_1) = \theta_1$ are known parameters.

In practice to determine the required solution we can use analytical tool to find the trajectory that can then be demonstrated to be the minimum of the action $\mathscr{A}$ (that's indeed a functional).

However we can also try to analytically determine the function $\theta^\star(t)$ that minimize the functional $\mathscr{A}(\cdot)$ by computing the directional derivatives.

Let now consider the function $\theta^\star(t)$ and a direction $\delta\theta(t)$ that satisfy

$$\delta\theta(t_0) = \delta\theta(t_1) = 0$$

we can then use the fundamental lemma of the calculus of variation to determine the minimal function (considering that the expression Equation 13.1 of the action is comparable to equation Equation 12.1 of the lemma).

To determine the minimum point we have in fact to determine the function $\theta^\star$ whose first variation of the action $\mathscr{A}$ is zero for each direction of approach $\delta\theta$ to the point, and so in this example we need to compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{A}\left(\theta^\star+\alpha\,\delta\theta\right) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\int_{t_0}^{t_1} L\left(\theta^\star+\alpha\,\delta\theta,\dot{\theta}^\star+\alpha\,\delta\dot{\theta}\right)\mathrm{d}t\\ &= \int_{t_0}^{t_1}\left[\frac{\partial L}{\partial\theta}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\theta^\star+\alpha\,\delta\theta\right)+\frac{\partial L}{\partial\dot{\theta}}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\dot{\theta}^\star+\alpha\,\dot{\delta}\theta\right)\right]\mathrm{d}t\\ &= \int_{t_1}^{t_2}\left(-lmg\sin(\theta^\star+\alpha\,\delta\theta)\right)\delta\theta+ml^2\left(\dot{\theta}^\star+\alpha\,\delta\dot{\theta}\right)\delta\dot{\theta}\,\mathrm{d}t\end{aligned}$$

Note that from in the first step we made the implicit assumption that

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\int=\int\frac{\mathrm{d}}{\mathrm{d}\alpha}$$

while however this is not always possible. Evaluating the previous expression for $\alpha=0$ gives the first variation of the action $\mathscr{A}$ that's

$$\begin{aligned}\delta\mathscr{A}(\theta^\star;\delta\theta) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{A}\left(\theta^\star+\alpha\,\delta\theta\right)\Big|_{\alpha=0}\\ &= \int_{t_0}^{t_1}\left(-lmg\sin\theta^\star\delta\theta+ml^2\dot{\theta}^\star\delta\dot{\theta}\right)\mathrm{d}t=0\end{aligned}$$

for all $\delta\theta$. Performing an integration by part allows to remove the term $\delta\dot{\theta}$ that's hard to determine, and in fact

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\,\delta\theta\right) &= \frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\right)\delta\theta+ml^2\dot{\theta}^\star\delta\dot{\theta}\\ &\Downarrow\\ ml^2\dot{\theta}^\star\delta\dot{\theta} &= \frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\,\delta\theta\right)-\frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\right)\delta\theta\end{aligned}$$

Performing the substitution of the integration by parts determines the result

$$\begin{aligned}\mathscr{A}\left(\theta^\star;\delta\theta\right) &= \int_{t_0}^{t_1}\left[-lmg\sin\theta^\star-\frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\theta^\star\right)^2\right]\delta\theta\,\mathrm{d}t+\cancel{\left[ml^2\dot{\theta}^\star\delta\theta\right]\Big|_{t_0}^{t_1}}\\ &= \int_{t_0}^{t_1}\underbrace{\left[-lmg\sin\theta^\star-\frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\right)\right]}_{f(t)}\delta\theta\,\mathrm{d}t=0\end{aligned}$$

We can now see in this formulation that the marked expression represent the function $f(x)$ of the fundamental lemma considering that the relation must be true for all approaching direction $\delta\theta$, and so it must be

$$-lmg\sin\theta^\star-\frac{\mathrm{d}}{\mathrm{d}t}\left(ml^2\dot{\theta}^\star\right)=0$$

The problem now to complete the analyses of the pendulum motion is determining the function $\theta^\star(t)$ that satisfy this expression and also match the boundary conditions $\theta(t_0)=\theta_0$ and $\theta(t_1)=\theta_1$.

13.2 General formulation

Given a generic functional of the form

$$\mathscr{A}(x) = \int_a^b L\big(x(t), x'(t), t\big)\,\mathrm{d}t$$

the problem is to minimize the functional $\mathscr{A}(x)$ for a function $x \in \mathbb{V}$ subject to the boundary conditions $x(a) = x_a$ and $x(b) = x_b$ in the function space $\mathbb{V}$ defined as

$$\mathbb{V} = \big\{x \mid x \in C^2([a,b]) \text{ with } x(a) = x_a, x(b) = x_b\big\}$$

Let $\mathbb{D}$ the function space of all the feasible directions of derivatives defined as

$$\mathbb{D} = \big\{\delta x \mid \delta x \in C^\infty([a,b]) \text{ with } \delta x(a) = \delta x(b) = 0\big\}$$

the function $x(t)$ that minimize the functional is the one whose first variation is zero for any admissible perturbation, and so such that

$$\delta\mathscr{A}(x;\delta x) = \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{A}(x+\alpha\,\delta x)\Big|_{\alpha=0} = 0 \qquad \text{for all } \delta x \in \mathbb{D}$$

Assuming the possibility to correctly apply the rule

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\int = \int\frac{\mathrm{d}}{\mathrm{d}\alpha}$$

we can express the directional derivative of the functional $\mathscr{A}(x)$ evaluated in $x + \alpha\,\delta x$ as

$$\begin{aligned}
\frac{\mathrm{d}\mathscr{A}(x+\alpha\delta x)}{\mathrm{d}\alpha} &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\int_a^b L\big(x+\alpha\,\delta x, x'+\alpha\,\delta x', t\big)\,\mathrm{d}t \\
&= \int_a^b \left(\frac{\partial L}{\partial x}\delta x + \frac{\partial L}{\partial x'}\delta x'\right)\mathrm{d}t \\
\frac{\mathrm{d}\mathscr{A}}{\mathrm{d}\alpha}\Big|_{\alpha=0} &= \int_a^b \left(\frac{\partial L}{\partial x}(x,x',t)\delta x + \frac{\partial L}{\partial x'}(x,x',t)\delta x'\right)\mathrm{d}t
\end{aligned}$$

By performing the integration by part it's possible do reconvert the term associated do $\delta x'$ into pieces depending on δx one of which, when evaluated, becomes zero due to the fact that $\delta x(a) = \delta x(b) = 0$:

$$\begin{aligned}
\frac{\mathrm{d}\mathscr{A}}{\mathrm{d}\alpha}\Big|_{\alpha=0} &= \int_a^b \left[\frac{\partial L}{\partial x}(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\right)\right]\delta x\,\mathrm{d}t \\
&\quad + \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\delta x\right)\mathrm{d}t \\
&= \int_a^b \left[\frac{\partial L}{\partial x}(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\right)\right]\delta x\,\mathrm{d}t \\
&\quad + \cancel{\left[\frac{\partial L}{\partial x'}(x,x',t)\delta x\right]_{t=a}^{t=b}}
\end{aligned}$$

> **Note**
>
> In this case the substitution by part has been obtained by performing the operation
>
> $$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\delta x\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\right)\delta x + \frac{\partial L}{\partial x'}(x,x',t)\delta x'$$
>
> If we in fact explicit the term containing $\delta x'$ we have

$$\frac{\partial L}{\partial x'}(x,x',t)\delta x' = \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\delta x\right)}_{\text{term \#1}} - \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t)\right)\delta x}_{\text{term \#2}}$$

The second term still remains in the variation of the functional $\mathscr{A}$, while the first term is eliminated because it's evaluation becomes

$$\begin{aligned}\int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}\delta x\right)\mathrm{d}t &= \left[\frac{\partial L}{\partial x'}(x(t),x'(t),t)\delta x(t)\right]_a^b \\ &= \frac{\partial L}{\partial x'}(x(b),x'(b),b)\delta x(b) - \frac{\partial L}{\partial x'}(x(a),x'(a),a)\delta x(t)\end{aligned}$$

Looking at the function space $\mathbb{D}$, we see that $\delta x(a) = \delta x(b) = 0$; hence the boundary term vanishes.

We can now see that the condition to have a minimum for the functional $\mathscr{A}$, by using the fundamental lemma of calculus of variation, is requiring that the function $f(t)$ defined as follows is identically null:

$$\frac{\mathrm{d}\mathscr{A}(x+\alpha\,\delta x)}{\mathrm{d}\alpha}\Big|_{\alpha=0} = \int_a^b \underbrace{\left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}\right)}_{=f(t)}\delta x\,\mathrm{d}t = 0 \tag{13.2}$$

This in general means that the function that minimise the functional $\mathscr{A}$ must solve the following second order ordinary differential equation:

$$\begin{cases}\dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial x'} - \dfrac{\partial L}{\partial x} = 0 \\ x(a) = x_a \\ x(b) = x_b\end{cases}$$

■ **Example 13.1 — computation of the first variation.** Given the generic functional defined as

$$\mathscr{F}(x) = \int_a^b G(x,x',t)\,\mathrm{d}t + x(a)x(b)$$

in order to compute it's first variation we can use the *standard* approach by evaluating the perturbation of the functional respect to a function x:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{F}\big(x+\alpha\,\delta x\big) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\int_a^b G\big(x+\alpha\,\delta x, x'+\alpha\,\delta x', t\big)\,\mathrm{d}t \\ &+ \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big(\big(x(a)+\alpha\delta x(a)\big)\big(x(b)+\alpha\,\delta x(b)\big)\Big) \\ &= \int_a^b \left(\frac{\partial G}{\partial x}x\,\delta x + \frac{\partial G}{\partial x'}\delta x'\right)\mathrm{d}t \\ &+ \delta x(a)\big(x(b)+\alpha\,\delta x(b)\big) + \big(x(a)+\alpha\,\delta x(a)\big)\delta x(b) \\ &\downarrow \quad \alpha = 0 \\ &= \int_a^b \left(\frac{\partial G}{\partial x}\delta x + \frac{\partial G}{\partial x'}\delta x'\right)\mathrm{d}t + \delta x(a)x(b) + x(a)\delta x(b)\end{aligned}$$

This redundant formulation can be simplified by using the Gateaux derivative notation δ that allows to express the first variation as

$$\begin{aligned}\delta\mathscr{F}(x) &= \delta\left(\int_a^b G(x,x',t)\,\mathrm{d}t + x(a)x(b)\right)\\ &= \int_a^b \delta G(x,x',t)\,\mathrm{d}t + \delta\big(x(a)x(b)\big)\\ &= \int_a^b \left(\frac{\partial G}{\partial x}(x,x',t)\delta x + \frac{\partial G}{\partial x'}(x,x',t)\delta x'\right)\mathrm{d}t\\ &\quad + \delta x(a)x(b) + x(a)\delta x(b)\end{aligned}$$

using integration by part

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial G}{\partial x'}(x,x',t)\delta x\right) = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial G}{\partial x'}(x,x',t)\delta x + \frac{\partial G}{\partial x'}(x,x',t)\delta x'$$

and substituting

$$\begin{aligned}\delta\mathscr{F}(x) = \int_a^b &\underbrace{\left(\frac{\partial G}{\partial x}(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial G}{\partial x'}(x,x',t)\right)}_{(A)}\delta x\,\mathrm{d}t\\ &+ \underbrace{\left[x(a) + \frac{\partial G}{\partial x'}(x(b),x'(b),b)\right]}_{(B)}\delta x(b)\\ &+ \underbrace{\left[x(b) - \frac{\partial G}{\partial x'}(x(a),x'(a),a)\right]}_{(C)}\delta x(a)\end{aligned}$$

using foundamental lemma of calculus of variations $(A) = 0$ and then using variation differents from zero on th eborder it follows that $(B) = (C) = 0$ and the stationary points of the functional $F(x)$ satrisfy the BVP

$$\begin{cases}\dfrac{\partial G}{\partial x}(x,x',t) - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial G}{\partial x'}(x,x',t) = 0\\ x(a) + \dfrac{\partial G}{\partial x'}(x(b),x'(b),b) = 0\\ x(b) - \dfrac{\partial G}{\partial x'}(x(a),x'(a),a) = 0\end{cases}$$

■

13.3 Application to the brachistochrone problem

As saw at the introduction, functional minimization can be used to solve real problem. As example the brachistochrone problem wants to determine the path that minimize the time T to reach point B from a point A for a mass m subjected only to the force of gravity: this means minimizing the function

$$T = \int_A^B \frac{1}{v}\,\mathrm{d}s$$

where v is the velocity of the point over the arc length $\mathrm{d}s$. In order to describe the problem we consider x as the horizontal axis and y the vertical one (with positive direction the one *facing upwards*);

- the point A is described by the coordinates $(x_a, x_b) = (0,0)$ (so it's in the center of the reference system)
- the point $B = (x_b, y_b)$ is placed anywhere in the lower semi-plane (in fact we must require that $y_b < 0$ in order to have a free fall of the object).

Considering that at the initial position A both the kinetic and potential energy are zero, due to the conservation of energy we can derive that

$$0 = mgy + \frac{1}{2}mv^2 \quad \Longrightarrow \quad v = \sqrt{-2gy}$$

We so have the velocity v as depending from the vertical coordinate y of the point, and so the next step is to relate the arc length $\mathrm{d}s$ with the horizontal variation $\mathrm{d}x$: we can see that

$$\mathrm{d}s = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2} = \sqrt{\mathrm{d}x^2 + y'^2(x)\mathrm{d}x^2} = \sqrt{1+y'^2}\mathrm{d}x$$

The initial brachistochrone problem can so be states as the following functional minimization problem:

$$\begin{aligned} &\text{minimize:} && \mathscr{F}(y) = \int_{x_a}^{x_b} L(y, y')\,\mathrm{d}x \\ &\text{subject to:} && y(x_a) = y_a = 0 \qquad y(x_b) = y_b < 0 \end{aligned}$$

where the independent variable is the x coordinate, and

$$L(y, y') = \sqrt{\frac{1+y'^2}{-2gy}}$$

the integral part of the target function to be minimized can be computed starting from

$$\begin{aligned} \frac{\partial L}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial L}{\partial y'} &= -\frac{1 + 3y'^2 + 2y'^4 - 2yy''}{2\sqrt{2}g^2y^3\sqrt{\left(-\dfrac{1+y'^2}{gy}\right)^3}} \\ &= \frac{1 + 3y'^2 + 2y'^4 - 2yy''}{2\sqrt{2}gy^2(1+(y')^2)}\sqrt{-\frac{gy}{1+y'^2}} \end{aligned}$$

and the functional minimization is equivalent to the following boundary value problem:

$$\begin{cases} 1 + 3\,y'(x)^2 + 2\,y'(x)^4 - 2\,y(x)y''(x) = 0 \\ y(0) = 0 \\ y(x_b) = y_b \end{cases}$$

Even if it's not clear at first sight, it can be proven that the solution of the differential equation is a cycloid described by the equations

$$x(\theta) = \frac{1}{2}k^2\big(\theta - \sin\theta\big), \qquad y(\theta) = \frac{1}{2}k^2\big(1 - \cos\theta\big)$$

where the constant k and the range of the new angle θ of the parametrization can be found by using the boundary conditions.

In particular, the initial condition on points A allows to state that $\theta_a = 0$ (in fact $x(0) = \frac{1}{2}k^2 0 = 0$ and also $y(0) = 0$), while considering the final conditions we can determine both θ_b and k solving this system of non-linear equation in such variables:

$$\begin{cases} x_b = \dfrac{1}{2}k^2\left(\theta_b - \sin\theta_b\right) \\ y_b = \dfrac{1}{2}k^2\left(1 - \cos\theta_b\right) \end{cases}$$



14. Boundary condition for a functional

Considering now the functional $\mathscr{B}(x)$ defined as

$$\mathscr{B}(x) = \int_a^b L(x(t), x'(t), x''(t), t)\,\mathrm{d}t \tag{14.1}$$

the problem to solve now is the minimization of $\mathscr{B}(x)$ for all the function $x \in \mathbb{V}$ where the functional space is defined as

$$\mathbb{V} = \left\{ x \text{ such that } \begin{array}{l} x \in C^4([a,b]) \\ x(a) = x_a, x'(a) = x'_a, x(b) = x_b, x'(b) = x'_b \end{array} \right\}$$

and directional derivatives $\delta x \in \mathbb{D}$ in the functional domain

$$\mathbb{D} = \left\{ \delta x \text{ such that } \begin{array}{l} x \in C^\infty([a,b]) \\ x(a) = x'(a) = x(b) = x'(b) = 0 \end{array} \right\}$$

When trying to calculate the directional derivative of the functional $\mathscr{B}$ (skipping all the unnecessary computation that's similar to the cases yet studies) we end up to the following results:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big|_{\alpha=0} \mathscr{B}(x+\alpha, \delta x) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big|_{\alpha=0} \int_a^b L\big(x+\alpha\,\delta x, x'+\alpha\,\delta x', x''+\alpha\,\delta x'', t\big)\,\mathrm{d}t \\ &= \int_a^b \left(\frac{\partial L}{\partial x}(x, x', x'', t)\delta x + \frac{\partial L}{\partial x'}\delta x' + \frac{\partial L}{\partial x''}\delta x'' \right) \mathrm{d}t \end{aligned}$$

To *cancel out* the terms involving the terms $\delta x'$, $\delta x''$ it's necessary to use integration by parts considering the following derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}\delta x \right) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x}\delta x \right) + \frac{\partial L}{\partial x'}\delta x'$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x''}\delta x' \right) = \left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\delta x' \right) + \frac{\partial L}{\partial x''}\delta x''$$

and so with that said the derivative becomes

$$\begin{aligned}\delta\mathscr{B} &= \int_a^b \frac{\partial L}{\partial x}\delta x + \frac{\mathrm{d}}{\mathrm{d}t}\cancel{\left(\frac{\partial L}{\partial x'}\delta x\right)} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}\delta x \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}t}\cancel{\left(\frac{\partial L}{\partial x''}\delta x'\right)} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\delta x'\,\mathrm{d}t \\ &= \int_a^b \left[\left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}\right)\delta x - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\delta x'\right]\mathrm{d}t\end{aligned}$$

In the first line the terms are cancelled because if evaluated singularly we can see that they become in the form

$$\int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}\delta x\right)\mathrm{d}t = \left[\frac{\partial L}{\partial x'}\delta x\right]_a^b \xrightarrow{\delta x(a)=\delta x(b)=0} 0$$

To finish the analysis of the derivation we have to consider one last integration by part for the element

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\delta x\right) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial x''}\delta x + \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\delta x'$$

$$\implies \qquad \delta\mathscr{B}(x) = \int_a^b \underbrace{\left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'} + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial x''}\right)}_{=f}\delta x\,\mathrm{d}t \tag{14.2}$$

Using so the fundamental lemma of calculus of variation in order to determine the function x that minimize the functional $\mathscr{B}$ we need to solve the system of differential equation

$$\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'} + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial x''} = 0$$

subject to the boundary conditions initially defined.

▪ **Example 14.1 — minimization of a functional with the Euler-Lagrange method.** Let's consider the problem

$$\begin{aligned}\text{minimize:} \quad & \int_a^b \left(\left(x''\right)^2 - x^2 + t\right)\mathrm{d}t \\ \text{with:} \quad & x(a) = 0, \quad x'(a) = 1 \\ & x(b) = 1, \quad x'(b) = 2\end{aligned}$$

In this case the function in minimization functional is

$$L(x, x', x'', t) = (x'')^2 - x^2 + t$$

and all the boundaries conditions are well defined. In this case to calculate the variation of the functional

$$\mathscr{A}(x) = \int_a^b L(x, x', x'', t)\,\mathrm{d}t$$

by using equation Equation 14.2:

$$\delta\mathscr{A} = \int_a^b \left(-2x - 0 + \frac{\mathrm{d}^2 x''}{\mathrm{d}t^2}\right)\delta x\,\mathrm{d}t$$

Using the fundamental lemma of calculus of variation the terms in the bracket must be always equal to zero: this so represent, in conjunction with the boundary conditions, the ordinary differential equation that has to be solved to determine the solution of the problem:

$$\begin{cases} x'''' - x = 0 \\ x(a) = 0, x'(a) = 1 \\ x(b) = 1, x'(b) = 2 \end{cases}$$

As we can see the differential equation is of order 4 and having 4 boundary condition and so it's possible to compute the solution.

■

14.1 Free boundary conditions

Until now we considered minimization problems with all boundary conditions values set, while however this might not always be the case: Equation 14.2 in fact is derived in the assumption of having all the variation constants $\delta x = \delta x^{(k)} = 0$ (for all k), but when this won't happens and so the general expression of the first variation is

$$\begin{aligned} \delta\mathscr{B}(x;\delta x) &= \int_a^b \left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'} + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial x''} \right) \delta x \, \mathrm{d}t \\ &+ \left[\left(\frac{\partial L}{\partial x'} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''} \right) \delta x \right]_a^b + \left[\left(-\frac{\partial L}{\partial x''} \right) \delta x' \right]_a^b \\ &= \int_a^b (A)\delta x \, \mathrm{d}t + (B)\delta x(b) - (C)\delta x(a) + (D)\delta x'(b) - (E)\delta x'(a) \end{aligned}$$

where

$$(A) = \frac{\partial L}{\partial x}(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}(x,x',t) + \frac{\mathrm{d}^2}{\mathrm{d}t^2}\frac{\partial L}{\partial x''}(x,x',t)$$

$$(B) = \frac{\partial L}{\partial x'}(x(b),x'(b),b) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}(x(b),x'(b),b)$$

$$(C) = \frac{\partial L}{\partial x'}(x(a),x'(a),a) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}(x(a),x'(a),a)$$

$$(D) = -\frac{\partial L}{\partial x''}(x(b),x'(b),b)$$

$$(E) = -\frac{\partial L}{\partial x''}(x(a),x'(a),a)$$

In this case if we consider that the boundary conditions fixed are only the one

$$x(a) = x_a \qquad x'(b) = x'_b$$

this also reflect on the functional domain of the variations $\delta x \in \mathbb{D}$ that becomes

$$\widetilde{\mathbb{D}} = \left\{ \delta x \in C^\infty([a,b]) \text{ such that } \delta x(a) = \delta x'(b) = 0 \right\}$$

We can now see that in general the domain $\widetilde{\mathbb{D}}$ of the variation set condition for value of the function (and it's derivative) on points only where the boundary conditions are defined. In the case

described considering the values fixed it means that the solution for the minimum is the one that satisfies:

$$\delta\mathscr{B}(x;\delta x) = \int_a^b (A)\delta x\,\mathrm{d}t \quad + (B)\delta x(b) - (E)\delta x'(a) \tag{14.3}$$

for all $\delta x \in \widetilde{\mathbb{D}}$. In this case we have 2 boundary degrees of freedom for the variation that we can define $\delta x(b) = \delta_{xb}$ and $\delta x'(a) = \delta'_{xa}$ (that are real evaluated variable); the simplest function that we might determine in order to create a variation $\delta x(t)$ is a polynomial function depending on this parameters, and so

$$\begin{aligned}\delta x(t) = \delta'_{xa}(t-a) + \frac{2\delta'_{xa}(a-b)+3\delta_{xb}}{(a-b)^2}(t-a)^2 \\ + \frac{\delta'_{xa}(a-b)+2\delta_{xb}}{(a-b)^3}(t-a)^3\end{aligned}$$

Considering that the Gateaux derivative $\delta\mathscr{B}(x;\delta x)$ should always be zero for each function $\delta x(t)$ in it's domain, we can consider the special function with parameters $\delta_{xb} = 1$ and $\delta'_{xa} = 0$ becoming

$$\delta x(t) = \frac{3}{(a-b)^2}(t-a)^2 + \frac{2}{(a-b)^3}(t-a)^3$$

Considering now that $\delta x'(a) = 0$ we can see that in expression Equation 14.3 the terms (E) is free and in order to have a null derivative it must be that $(B) = 0$. Similarly we can create a particular polynomial (in particular considering

$$\delta_{xb} = \delta x(b) = 0 \quad \text{and} \quad \delta'_{xa} = 1$$

that set to zero the coefficients associated to (B) (and so leaving it *free* to change) and so, to have a zero evaluated derivative, it must be that $(E) = 0$.

In this case the differential equation associated to the terms B, E are called **trasversality conditions** and are necessary when the boundary conditions of the problem are not enough (in fact that expressions are evaluated at precise point on the *boarder* of the domain). At this point the solution of the original problem becomes

$$\begin{cases} \dfrac{\partial L}{\partial x} - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial x'} + \dfrac{\mathrm{d}^2}{\mathrm{d}t^2}\dfrac{L}{\partial x''} = 0 \\ \dfrac{\partial L}{\partial x'}\Big|_b - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial x''}\Big|_b = 0 \\ -\dfrac{\partial L}{\partial x''}\Big|_a = 0 \\ x(a) = x_a,\ x'(b) = x'_b \end{cases}$$

■ **Example 14.2 — minimization with less boundary conditions.** Let's consider the problem of minimizing the functional $\mathscr{A}$ as in Example 14.1 here reported

$$\text{minimize:} \qquad \int_a^b \left(\left(x''\right)^2 - x^2 + t \right) \mathrm{d}t$$

where the boundary conditions in this case are only

$$x(a) = 0 \qquad x'(b) = 2$$

In this case the boundary conditions are not enough and the case is similar to the theory yet described: in this case the domain of the variations δx is described as

$$\mathbb{D} = \left\{ \delta x \in C^\infty([a,b]) \text{ such that } \delta x(a) = \delta x'(b) = 0 \right\}$$

No information are set for the values $\delta x(b)$, $\delta x'(a)$ that are so *free* to have different values in $\mathbb{R}$ so determining the following two trasverality conditions:

$$\begin{aligned} \frac{\partial L}{\partial x'}\Big|_b - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x''}\Big|_b = 0 - 2x'''\Big|_b = -2x'''(b) = 0 \\ -\frac{\partial L}{\partial x''}\Big|_a = -\big(-2x''\big)\Big|_a = 2x''(a) = 0 \end{aligned}$$

This, in conjunction with the differential equation determined by the function in the integral and the known boundary conditions, determines the following ordinary system of equation that's the solution that minimize the functional:

$$\begin{cases} x'''' - x = 0 \\ x'''(b) = 0 \\ x''(a) = 0 \\ x(a) = 0, x'(b) = 2 \end{cases}$$

■

■ **Example 14.3 — minimization of a functional.** Let's consider the problem

$$\begin{aligned} \text{minimize:} \quad & \mathscr{A}(y) = \int_0^1 \left(\frac{\big(y'(x)\big)^2}{2} + y(x)y'(x) + y(x) \right) \mathrm{d}x \\ \text{with:} \quad & y(1) = 1 \end{aligned}$$

In this case the lagrangian of the problem is defined as

$$L(y, y', x) = \frac{(y')^2}{2} + yy' + y$$

the first variation of this functional so becomes

$$\delta\mathscr{A}(y;\delta y) = \int_0^1 \left(\frac{\partial L}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial L}{\partial y'} \right) \delta y \,\mathrm{d}x + \left[\frac{\partial L}{\partial y'} \delta y \right] \Big|_{x=0}^{x=1} = 0$$

To formally compute the derivative we firstly need to define the domain of the variation δy that, having only one boundary condition, is

$$\mathbb{D} = \left\{ \delta y \in C^\infty([0,1]) \text{ with } \delta y(1) = 0 \right\}$$

At this point the Gateaux derivative can be computed as

$$\begin{aligned} \delta\mathscr{A}(y;\delta y) &= \int_0^1 \underbrace{\big(y' - (y'' - y') + 1\big)}_{=f} \delta y \,\mathrm{d}x + \Big[\big(y' + y\big)\delta y \Big]_{x=0}^{x=1} \\ &= \int_0^1 1 - y''\delta y \,\mathrm{d}x - \underbrace{\big(y' + y\big)\delta y(0)}_{\text{trasv. cond.}} \end{aligned}$$

As we can see we have that the term into the integral (associated to y'') related to the fundamental lemma must be zero and so represent the differential equation associated to the solution of the

problem while the second term $y' + y$ evaluated for $x = 0$ represent the transversality condition that allow to have a unique solution of the system of ordinary differential equation that is:

$$\begin{cases} y''(x) = 1 \\ y'(0) + y(0) = 0 \\ y(1) = 1 \end{cases}$$

By integration of the first differential equation we can determine the parametric solution of the system whose coefficients can be matched considering the boundary conditions:

$$y(x) = \frac{1}{2}x^2 + c_1 x + c_2$$

Substituting the parametric solution on the boundary conditions we can solve for the parameters:

$$\begin{cases} c_1 + c_2 = 0 \\ \frac{1}{2} + c_1 + c_2 = 0 \end{cases}$$

In this case the system of linear equation has no solution (in fact we have that $c_1 + c_2 = 0 \neq -1/2$) and so this means that the functional $\mathscr{A}$ cannot be minimized.

■



15. Functional minimization with generic constraints

Let's consider the problem of the minimization of a functional $\mathscr{F}$ subjected to an inequality constraints *at the boarder* as follows:

$$\begin{aligned} &\text{minimize:} && \mathscr{F}(z) = \int_a^b L(z,z',t)\,\mathrm{d}t \\ &\text{subject to:} && w\big(z(a),z(b)\big) = 0 \end{aligned} \tag{15.1}$$

In this case we want to find the solutions for the function $z(t)$ in the functional spaced defined as

$$\mathbb{V} = \Big\{ z \in C^2([a,b]) \text{ such that } w\big(z(a),z(b)\big) = 0 \Big\}$$

In this case we cannot consider the linearity of the functional space (in fact two functions z_1, z_2 can satisfy the constraint b, but their sum $z_1 + z_2$ doesn't belong to the domain), and more difficult is the definition of the directional domain $\mathbb{D}$ on which we can compute all the possible derivatives.

15.0.0.1 Discretization approach

A way to solve this problem is by discretizing the problem: given the domain $[a,b]$ we can subdivide him in n subintervals (in this case equally spaced) having length $h = \frac{b-a}{2}$; the axis t is so discretized in values

$$t_k = t_0 + kh = a + k\frac{b-a}{n}$$

With this definition we can compute the function z in the various point considering that $z(t_k) = z_k$ (and indeed we can also note that $t_0 = a$ and $t_n = b$). Using the mid-point squaring numerical method to integrate the original function, we can approximate the functional as

$$\mathscr{F}(z) = \int_a^b L(z,z',t)\,\mathrm{d}t \approx h \sum_{k=1}^{n} L\left(z_{k-\frac{1}{2}}, z'_{k-\frac{1}{2}}, t_{k-\frac{1}{2}}\right)$$

where

$$z_{k-\frac{1}{2}} = \frac{z_k + z_{k-1}}{2} \qquad z'_{k-\frac{1}{2}} = \frac{z_k - z_{k-1}}{h} \qquad t_{k-\frac{1}{2}} = t_k - \frac{h}{2}$$

With this being said the initial constrained minimization problem (Equation 15.1) can be discretized so obtaining the form

$$\begin{aligned} &\text{minimize:} && f(\boldsymbol{z}) = h\sum_{k=1}^{n} L\left(z_{k-\frac{1}{2}}, z'_{k-\frac{1}{2}}, t_{k-\frac{1}{2}}\right) \\ &\text{subject to:} && w(z_0, z_n) = 0 \end{aligned} \tag{15.2}$$

where $\boldsymbol{z} = (z_0, z_1, \ldots, z_n)^t$ is the vector off all the discretized values of the initial function $z(t)$. This formulation recall the constrained minimization problem with equality constraints and so we can use the Lagrange multiplier method by defining the expression

$$\mathscr{L}(\boldsymbol{z}, \lambda) = f(\boldsymbol{z}) - \lambda w(z_0, z_n)$$

To solve this kind of problem we need to find the stationary point of the yet built lagrangian $\mathscr{L}$,and so this means solving the following non-linear system determined by the equations

$$\frac{\partial \mathscr{L}}{\partial z_i} = 0 \qquad \forall i = 0, 1, \ldots, n \qquad \frac{\partial \mathscr{L}}{\partial \lambda} = 0$$

Starting with $i = 0$ we can compute that derivative of the lagrangian as

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial z_0} &= \frac{\partial}{\partial z_0}\left(h\sum_{k=1}^{n} L\left(\frac{z_k + z_{k-1}}{2}, \frac{z_k - z_{k-1}}{h}, t_{k-\frac{1}{2}}\right) - \lambda w(z_0, z_n)\right) \\ &= \frac{\partial}{\partial z_0}\left(hL(\underbrace{\frac{z_1 + z_0}{2}, \frac{z_1 - z_0}{h}, t_{\frac{1}{2}}}_{[1]}) - \lambda w(z_0, z_n)\right) \\ &= h\frac{\partial L([1])}{\partial z}\frac{\partial}{\partial z_0}\left(\frac{z_1 + z_0}{2}\right) + h\frac{\partial L([1])}{\partial z'}\frac{\partial}{\partial z_0}\left(\frac{z_1 - z_0}{h}\right) - \lambda\frac{\partial w(z_0, z_n)}{\partial z_0} \\ &= \frac{h}{2}\frac{\partial L([1])}{\partial z} - \frac{\partial L([1])}{\partial z'} - \lambda\frac{\partial w(z_0, z_n)}{\partial z_0} \end{aligned}$$

Note that passing from the first to the second line only the terms associated to $k = 1$ present terms depending on z_0, and so only that part of the summation has been considered.

For all the other values $i \neq 0, n$, the mathematical expression of the derivative $\partial\mathscr{L}/\partial z_i$ becomes more *complex* (due to the fact that we have to consider two elements of the summation) and so we can use the simplified notation to express the partial terms such

$$\frac{\partial L}{\partial z}\Big|_{k+\frac{1}{2}} := \frac{\partial}{\partial z} L\left(\frac{z_{k+1} + z_k}{2}, \frac{z_{k+1} - z_k}{h}, t_{k+\frac{1}{2}}\right)$$

With this, doing the steps as previously shown, we can compute the partial derivatives as

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial z_k} &= \frac{\partial}{\partial z_k}\left(h\sum_{j=1}^{n} L\left(\frac{z_j + z_{j-1}}{2}, \frac{z_j - z_{j-1}}{h}, t_{j-\frac{1}{2}}\right) - \lambda w(z_0, z_n)\right) \\ &= \frac{h}{2}\left(\frac{\partial L}{\partial z}\Big|_{k-\frac{1}{2}} + \frac{\partial L}{\partial z}\Big|_{k+\frac{1}{2}}\right) + \frac{\partial L}{\partial z'}\Big|_{k-\frac{1}{2}} - \frac{\partial L}{\partial z'}\Big|_{k+\frac{1}{2}} \\ &= h\left(\frac{\frac{\partial L}{\partial z}\Big|_{k-\frac{1}{2}} + \frac{\partial L}{\partial z}\Big|_{k+\frac{1}{2}}}{2} - \frac{\frac{\partial L}{\partial z'}\Big|_{k+\frac{1}{2}} - \frac{\partial L}{\partial z'}\Big|_{k-\frac{1}{2}}}{h}\right) \end{aligned}$$

The last partial derivative (computed for $k = n$) is instead

$$\frac{\partial \mathscr{L}}{\partial z_n} = \frac{h}{2}\frac{\partial L}{\partial z}\Big|_{k-\frac{1}{2}} + \frac{\partial L}{\partial z'}\Big|_{k-\frac{1}{2}} - \lambda\frac{\partial w(z_0, z_n)}{\partial z_n}$$

With all this calculation being done we determine that the non-linear system of equations representing the first order necessary condition for the minimum point of the lagrangian $\mathscr{L}$ is

$$\begin{cases} \dfrac{\frac{\partial L}{\partial z}\Big|_{k-\frac{1}{2}} + \frac{\partial L}{\partial z}\Big|_{k+\frac{1}{2}}}{2} - \dfrac{\frac{\partial L}{\partial z'}\Big|_{k+\frac{1}{2}} - \frac{\partial L}{\partial z'}\Big|_{k-\frac{1}{2}}}{h} = 0 & : A \\ \dfrac{h}{2}\dfrac{\partial L}{\partial z}\Big|_{\frac{1}{2}} - \dfrac{\partial L}{\partial z'}\Big|_{\frac{1}{2}} - \lambda \dfrac{\partial w(z_0, z_n)}{\partial z_0} = 0 & : B \\ \dfrac{h}{2}\dfrac{\partial L}{\partial z}\Big|_{n-\frac{1}{2}} + \dfrac{\partial L}{\partial z'}\Big|_{n-\frac{1}{2}} - \lambda \dfrac{\partial w(z_0, z_n)}{\partial z_n} = 0 & : C \end{cases}$$

By pushing the limit for $h \to 0$ (to have a *continuous discretization*), we can see a correlation of the two terms composing equation A, in fact

$$\begin{aligned} \frac{\frac{\partial L}{\partial z}\Big|_{k-\frac{1}{2}} + \frac{\partial L}{\partial z}\Big|_{k+\frac{1}{2}}}{2} &\approx \frac{\partial L}{\partial z}\left(z_k, z'_k, t_k\right) \\ \frac{\frac{\partial L}{\partial z'}\Big|_{k+\frac{1}{2}} - \frac{\partial L}{\partial z'}\Big|_{k-\frac{1}{2}}}{h} &\approx \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial z'}\left(z_k, z'_k, t_k\right)\right) \end{aligned}$$

Similarly both the terms A and B presents the terms similar to the previous definition and they also presents another contribute due to the Lagrange multiplier λ. By so considering the limit $h \to 0$ we can see that $k = \frac{1}{2}$ tends to be a while $n - \frac{1}{2}$ tends to b, and so we can rewrite the systems of non-linear equations as

$$\begin{cases} \dfrac{\partial L}{\partial z}(z, z', t) - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial z'}(z, z', t) = 0 \\ -\dfrac{\partial L}{\partial z}(z(a), z'(a), a) - \lambda \dfrac{\partial w(z(a), z(b))}{\partial z(a)} = 0 \\ -\dfrac{\partial L}{\partial z}(z(b), z'(b), b) - \lambda \dfrac{\partial w(z(a), z(b))}{\partial z(b)} = 0 \\ w\big(z(a), z(b)\big) = 0 \end{cases}$$

This is so the general formulation of the initial problem of minimizing a functional

$$\mathscr{F}(z) = \int_a^b L(z, z', t)\,\mathrm{d}t$$

subject to an equality constraint w.

15.1 Heuristic deduction of general boundary condition

The same result can also be achieved with an heuristic formulation. Given so the problem of minimizing a functional $\mathscr{F}$ subjected to an equality constraint w (as in Equation 15.1), the solution can be achieved by determining a new functional $\mathscr{L}$ defined as the lagrangian of the system:

$$\mathscr{L}(z, \lambda) = \int_a^b L(z, z', t)\,\mathrm{d}t - \lambda w\big(z(a), z(b)\big)$$

At this point we can compute the variation of this functional considering it's Gateaux derivative δ that's

$$
\begin{aligned}
\delta\mathscr{L}(z,\lambda;\delta z,\delta\lambda) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\Big|_{\alpha=0}\mathscr{L}\big(z+\alpha\,\delta z,\lambda+\alpha\,\delta\lambda\big)\\
&= \delta\int_a^b L(z,z',t)\,\mathrm{d}t - \delta\lambda\, w\big(z(a),z(b)\big) - \lambda\,\delta w\big(z(a),z(b)\big)\\
&= \int_a^b\left(\frac{\partial\mathscr{L}}{\partial z}\delta z+\frac{\partial\mathscr{L}}{\partial z'}\delta z'\right)\mathrm{d}t - \delta\lambda\, w\big(z(a),z(b)\big)\\
&\quad -\lambda\left(\frac{\partial w(z(a),z(b))}{z(a)}\delta z(a)+\frac{\partial w(z(a),z(b))}{z(b)}\delta z(b)\right)\\
&= \int_a^b\left(\frac{\partial\mathscr{L}}{\partial z}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathscr{L}}{\partial z'}\right)\delta z\,\mathrm{d}t+\left[\frac{\partial\mathscr{L}}{\partial z'}\delta z\right]\Big|_a^b - \delta\lambda\, w\big(z(a),z(b)\big)\\
&\quad -\lambda\left(\frac{\partial w}{\partial z(a)}\delta z(a)+\frac{\partial w}{\partial z(b)}\delta z(b)\right)
\end{aligned}
$$

By evaluating the partial derivative $\partial\mathscr{L}/\partial z'$ and collecting common terms we can reduce the variation of the functional to the form

$$
\delta\mathscr{L}(z,\lambda;\delta z,\delta\lambda) = \int_a^b (A)\,\delta z\,\mathrm{d}t - \delta\lambda\, w\big(z(a),z(b)\big) + (B)\delta z(a) + (C)\delta z(b)
$$

where

$$
\begin{aligned}
(A) &= \frac{\partial\mathscr{L}}{\partial z}(z,z',t) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\mathscr{L}}{\partial z'}(z,z',t)\\
(B) &= -\frac{\partial\mathscr{L}}{\partial z'}(z(a),z'(a),a) - \lambda\frac{\partial w}{\partial z(a)}(z(a),z(b))\\
(C) &= \frac{\partial\mathscr{L}}{\partial z'}(z(b),z'(b),b) - \lambda\frac{\partial w}{\partial z(b)}(z(a),z(b))
\end{aligned}
$$

We start by noting that the variation $\delta\mathscr{L}(z,\lambda;\delta z,\delta\lambda)$ must be equal to zero for all variations $\delta z\in C^\infty([a,b])$ and for all constants $\delta\lambda\in\mathbb{R}$.

First, consider the case where $\delta\lambda=0$, $\delta z(a)=0$, and $\delta z(b)=0$. In this scenario, we can apply the fundamental lemma of the calculus of variations, which implies that the term A must equal zero.

Next, let's examine the case where $\delta\lambda\neq 0$ while maintaining $\delta z(a)=0$ and $\delta z(b)=0$. Under these conditions, we arrive at the initial equality constraint:

$$
w(z(a),z(b)) = 0.
$$

Now, consider the situation where $\delta\lambda=0$, $\delta z(a)\neq 0$, and $\delta z(b)=0$. In this case, we can conclude that the term (B) must also equal zero. Similarly, if we have $\delta z(b)\neq 0$ while keeping $\delta\lambda=0$ and $\delta z(a)=0$, we find that the term (C) must be zero as well.

This means that the resultant system of non-linear equation that solves the problem can be expressed as

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial z}(z,z',t) - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial \mathscr{L}}{\partial z'}(z,z',t) = 0 \\ w\big(z(a),z(b)\big) = 0 \\ \boxed{\dfrac{\partial \mathscr{L}}{\partial z'}(z(a),z'(a),a) + \lambda \dfrac{\partial w\big(z(a),z(b)\big)}{\partial z(a)} = 0} \\ \boxed{\dfrac{\partial \mathscr{L}}{\partial z'}(z(b),z'(b),b) - \lambda \dfrac{\partial w\big(z(a),z(b)\big)}{\partial z(b)} = 0} \end{cases} \tag{15.3}$$

We can observe that the resulting system is equivalent to the one obtained by pushing the limit $h \to 0$ with the discretized version previously performed.

15.1.0.1 Verification

Considering the common case described by the problem

$$\begin{aligned} &\text{minimize:} && \mathscr{F}(x) = \int_a^b L(x,x',t)\,\mathrm{d}t \\ &\text{subject to:} && x(a) = x_a,\ x(b) = x_b \end{aligned}$$

than the solution can be still achieved using the method yet shown. By in fact building the Lagrangian functional

$$\mathscr{L}(x,\lambda_1,\lambda_2) = \int_a^b L(x,x',t)\,\mathrm{d}t - \lambda_1\big(x(a) - x_a\big) - \lambda_2\big(x(b) - x_b\big)$$

substituting this function in the result of Equation 15.3 we get the following differential system of equations:

$$\begin{cases} \dfrac{\partial L}{\partial x}(x,x',t) - \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial L}{\partial x'}(x,x',t) = 0 \\ x(a) = x_a, \quad x(b) = x_b \\ \boxed{-\lambda_1 - \dfrac{\partial L}{\partial x'}(x(a),x'(a),a) = 0} \\ \boxed{-\lambda_2 + \dfrac{\partial L}{\partial x'}(x(b),x'(b),b) = 0} \end{cases}$$

The first two lines represent the terms that were always presents in the previous statement of the problem (without the equality constraints), while the last two depending from λ_i gives no real information because they can always be verified (in fact there will always exists a parameter λ_i that equals the derivative $\partial L/\partial x'$); in general the variable λ_i can appear in the other equations, and so this last can be used to determine the stationary solution of the lagrangian $\mathscr{L}$.

▪ **Example 15.1 — computation of first variation.** Given the functional

$$\mathscr{F}(x) = x(0) + \int_0^1 tx^2 + \big(x'\big)^2\,\mathrm{d}t$$

it's first variation can be computed considering that

$$L(x,x',t) = tx^2 + \big(x'\big)^2$$

so that

$$\frac{\partial L}{\partial x}(x,x',t) = 2tx \qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}(x,x',t) = \frac{\mathrm{d}}{\mathrm{d}t}(2x') = 2x''$$

To compute the complete variation of the system it's mandatory to consider the terms resulting from the integration by part (due to the fact that no boundary conditions are set) related to the term

$$\left[\frac{\partial L}{\partial x'}\delta x\right]_a^b$$

With that said, the overall first variation can be computed as

$$\begin{aligned}\delta\mathscr{F}(x;\delta x) &= \delta x(0) + 2\int_0^1 \left(tx - x''\right)\delta x\,\mathrm{d}t + 2x'\delta x(1) - 2x'\delta x(0)\\ &= 2\int_0^1 \left(tx - x''\right)\delta x\,\mathrm{d}t + 2x'(1)\delta x(1) + \left(1 - 2x'(0)\right)\delta x(0)\end{aligned}$$

and the resulting BVP by the application of foundamental lemma of calculus of variations is

$$\begin{cases} tx - x'' = 0 \\ x'(1) = 0 \\ 2x'(0) = 1 \end{cases}$$

■

■ **Example 15.2 — boundary value problem.** Given the problem

$$\begin{aligned} \text{minimize:} \quad & \mathscr{F}(x) = x(0) + \int_0^1 x^2 + \left(x' - t\right)^2 \mathrm{d}t \\ \text{subject to:} \quad & \int_0^1 x\,\mathrm{d}t = 0 \\ & x(1) = 2 \end{aligned}$$

the resulting boundary value problem can be computed considering the lagrangian $\mathscr{L}$ of the system that's

$$\begin{aligned}\mathscr{L}(x,\lambda,\mu) &= \mathscr{F}(x) - \lambda\int_0^1 x\,\mathrm{d}t - \mu\left(x(1) - 2\right)\\ &= \int_0^1 \underbrace{x^2 + \left(x' - t\right)^2 - \lambda x}_{L}\,\mathrm{d}t + x(0) - \mu\left(x(1) - 2\right)\end{aligned}$$

Starting from this point we can so compute the first variation of $\mathscr{L}$ that's:

$$\begin{aligned}\delta\mathscr{L} &= \int_0^1 \left(\frac{\partial L}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}\right)\delta x\,\mathrm{d}t + \left[\frac{\partial L}{\partial x'}\delta x\right]_0^1 - \int_0^1 x\,\delta\lambda\,\mathrm{d}t\\ &\quad + \delta x(0) - \mu\delta x(1) - \left(x(1) - 2\right)\delta\mu\\ &= \int_0^1 \left(-2x'' + 2x + 2 - \lambda\right)\delta x\,\mathrm{d}t\\ &\quad + 2\left(x'(1) - 1\right)\delta x(1) - 2x'(0)\delta x(0)\\ &\quad - \int_0^1 x\delta\lambda\,\mathrm{d}t + \delta x(0) - \mu\delta x(1) - \left(x(1) - 2\right)\delta\mu\end{aligned}$$

The resulting boundary valued problem is so determined by setting to zero the terms that multiply every variation $\delta\cdot$ and so:

$$\begin{cases} \delta x: & 2x''-2x-2+\lambda=0 \\ \delta x(0): & 2x'(0)=1 \\ \delta x(1): & \cancel{2x'(1)-2=\mu} \quad \text{: trivially satisfied} \\ \delta\lambda: & \int_0^1 x=0 \\ \delta\mu: & x(1)=2 \end{cases}$$

■

■ **Example 15.3 — exam's simulation.** Given the functional

$$\begin{aligned} &\text{minimize:} && \mathscr{F}(x,y,z)=y(1)+\int_0^1 x^2+z^2+x'y'+z'^2\,\mathrm{d}t \\ &\text{subject to:} && x(1)=2 \end{aligned}$$

the resulting boundary value problem can be stated by determining the first variation of the lagrangian

$$\mathscr{L}(x,y,z,\lambda)=y(1)+\int_0^1 \underbrace{x^2+z^2+x'y'+z'^2}_{=L(x,y,z,x',y',z')}\,\mathrm{d}t-\lambda\big(x(1)-2\big)$$

that results in

$$\begin{aligned} \delta\mathscr{L}(x,y,z,\lambda)=\delta y(1)&+\int_0^1\left(\frac{\partial L}{\partial x}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial x'}\right)\delta x\,\mathrm{d}t+\left[\frac{\partial L}{\partial x'}\delta x\right]_0^1 \\ &+\int_0^1\left(\frac{\partial L}{\partial y}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial y'}\right)\delta y\,\mathrm{d}t+\left[\frac{\partial L}{\partial y'}\delta y\right]_0^1 \\ &+\int_0^1\left(\frac{\partial L}{\partial z}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial z'}\right)\delta z\,\mathrm{d}t+\left[\frac{\partial L}{\partial z'}\delta z\right]_0^1 \\ &+\lambda\delta x(1)+\big(x(1)-2\big)\delta\lambda \\ \\ &=\delta y(1)+\int_0^1\big(2x-y''\big)\delta x\,\mathrm{d}t+y'(1)\delta x(1)-y'(0)\delta x(0) \\ &+\int_0^1-x''\delta y\,\mathrm{d}t+x'(1)\delta y(1)-x'(0)\delta y(0) \\ &+\int_0^1\big(2z-2z''\big)\delta z\,\mathrm{d}t \\ &+2z'(1)\delta z(1)-2z'(0)\delta z(0)+\lambda\delta x(1)+\big(x(1)-2\big)\delta\lambda \end{aligned}$$

Setting to zero the terms associated to each variation $\delta\cdot$ determines the following boundary value problem, solution of the functional minimization:

$$\begin{cases} 2x - y'' = 0 & : \delta x \\ x'' = 0 & : \delta y \\ z - z'' = 0 & : \delta z \\ y'(0) = 0 & : \delta x(0) \\ \cancel{y'(1) + \lambda = 0} & : \delta x(1), \text{ trivially solved in } \lambda \\ x'(0) = 0 & : \delta y(0) \\ 1 + x'(1) = 0 & : \delta y(1) \\ z'(0) = 0 & : \delta z(0) \\ z'(1) = 0 & : \delta z(1) \\ x(1) = 2 & : \delta \lambda \end{cases}$$

■



16. Minimization with Integral Constraints

We consider the minimization of a functional $\mathscr{A}(x)$ defined as

$$\mathscr{A}(x) = \int_0^1 (x')^2 \, \mathrm{d}t,$$

subject to the following constraints:

1. **Integral Constraint:** $\int_0^1 x \, \mathrm{d}t = 1$,
2. **Boundary Conditions:** $x(0) = 2$ and $x(1) = 1$.

Thus, the minimization problem can be stated as

$$\begin{aligned} \text{minimize} \quad & \mathscr{A}(x) = \int_0^1 (x')^2 \, \mathrm{d}t, \\ \text{subject to} \quad & \int_0^1 x \, \mathrm{d}t = 1, \\ & x(0) = 2, \quad x(1) = 1. \end{aligned}$$

To address this problem, we define a feasible space $\mathbb{W}$ for x, which includes functions satisfying the boundary conditions:

$$\mathbb{W} = \left\{ x \in C^1([0,1]) \mid x(0) = 2, x(1) = 1 \right\}.$$

The problem can thus be reformulated as finding $x \in \mathbb{W}$ that minimizes $\mathscr{A}(x)$.

Feasible directions and conditions

To analyze feasible directions for variations of x, we introduce a perturbation $x + \alpha \delta x$ with δx satisfying

$$\begin{aligned} x(0) + \alpha \delta x(0) &= 2, \\ x(1) + \alpha \delta x(1) &= 1, \\ \int_0^1 (x(t) + \alpha \delta x(t)) \, \mathrm{d}t &= 1. \end{aligned}$$

Expanding these conditions yields:

1. $\delta x(0) = 0$,
2. $\delta x(1) = 0$,
3. $\int_0^1 \delta x(t) \, \mathrm{d}t = 0$.

This means the set of feasible variations δx lies in the space

$$\mathbb{V} = \left\{ \delta x \in C^\infty([0,1]) \mid \delta x(0) = 0,\, \delta x(1) = 0,\, \int_0^1 \delta x \,\mathrm{d}t = 0 \right\}.$$

Stationarity Condition

A function $x \in \mathbb{W}$ is stationary for this problem if

$$\left.\frac{\mathrm{d}}{\mathrm{d}\alpha}\right|_{\alpha=0} \mathscr{A}(x + \alpha\delta x) = 0, \quad \forall \delta x \in \mathbb{V}.$$

Calculating this derivative, we get

$$\left.\frac{\mathrm{d}}{\mathrm{d}\alpha}\right|_{\alpha=0} \int_0^1 (x' + \alpha\delta x')^2 \,\mathrm{d}t = \int_0^1 2x'x'' \,\mathrm{d}t.$$

Using integration by parts on $\int_0^1 2x'x'' \,\mathrm{d}t$, we find

$$\int_0^1 2x'x'' \,\mathrm{d}t = 2x'\delta x\Big|_0^1 - 2\int_0^1 x''\delta x \,\mathrm{d}t.$$

Since $\delta x(0) = \delta x(1) = 0$, the boundary term vanishes, leaving

$$0 = \int_0^1 x''\delta x \,\mathrm{d}t \quad \forall \delta x \in \mathbb{V}.$$

At this point, we cannot directly conclude that $x'' = 0$ because $\mathbb{V}$ is not rich enough for the fundamental lemma of the calculus of variations to apply.

Constructing an Appropriate Variation

To circumvent this limitation, consider an auxiliary function $\delta y \in C^\infty([0,1])$ with $\delta y(0) = \delta y(1) = 0$. Define δx as

$$\delta x(t) = \delta y(t) - \lambda t(1-t) \int_0^1 \delta y(\tau) \,\mathrm{d}\tau.$$

Choosing $\lambda = 6$ ensures that $\delta x \in \mathbb{V}$.

Concluding the Solution

Using this construction in our stationarity condition, we obtain

$$0 = \int_0^1 \left(x''(t) - 6\left(x(1) + x(0) - 2\int_0^1 x(\tau) \,\mathrm{d}\tau \right) \right) \delta y(t) \,\mathrm{d}t.$$

Now, since δy satisfies the conditions for the fundamental lemma of calculus of variations, we conclude that

$$x''(t) = 6\left(x(1) + x(0) - 2\int_0^1 x(\tau) \,\mathrm{d}\tau \right).$$

Applying the Boundary Conditions

Substituting the boundary conditions and integral constraint, we get

$$x''(t) = 6, \quad x(0) = 2, \quad x(1) = 1.$$

Solving this system, we find that

$$x(t) = 3t^2 - 4t + 2,$$

which satisfies all the boundary conditions and integral constraint.

Final Remark

While this approach correctly finds the solution, it would be advantageous to simplify the derivation by seeking a direct application of the fundamental lemma of calculus of variations. This simplification would streamline the solution process and avoid the need for constructing additional variations.

16.1 Approximate solution using finite differences

To approximate the variational problem

$$\begin{aligned} \text{minimize} \quad & \mathscr{A}(x) = \int_0^1 (x')^2 \, \mathrm{d}t, \\ \text{subject to} \quad & \int_0^1 x \, \mathrm{d}t = 1, \\ & x(0) = 2, \quad x(1) = 1, \end{aligned}$$

we apply the finite difference method. This approach transforms the continuous problem into a discrete one, allowing us to approximate the solution numerically.

16.1.1 Discretization Setup

1. **Divide the Interval:** Let n be the number of intervals, and define $h = \frac{1}{n}$ as the interval width. We set sample points $t_k = kh$ for $k = 0, 1, \ldots, n$, where $t_0 = 0$ and $t_n = 1$.
2. **Approximate the Solution:** Define $y_k \approx x(t_k)$ as the approximate value of the solution at each sample point t_k.
3. **Quadrature Formula for the Integral:** To approximate the integral of $(x')^2$, we use the midpoint rule:

$$\int_0^1 (x')^2 \, \mathrm{d}t \approx h \sum_{k=1}^{n} \left(x'(t_{k-\frac{1}{2}}) \right)^2 + \mathscr{O}(h^2),$$

where $x'(t_{k-\frac{1}{2}})$ represents the approximate derivative at the midpoint between t_{k-1} and t_k.
4. **Finite Difference Approximation for the Derivative:** The first derivative at $t_{k-\frac{1}{2}}$ can be approximated by

$$x'(t_{k-\frac{1}{2}}) = \frac{x(t_k) - x(t_{k-1})}{h} + \mathscr{O}(h^2).$$

5. **Quadrature Formula for the Integral constraint:** To approximate the integral of x, we use the trapezoidal rule:

$$\int_0^1 (x')^2 \, \mathrm{d}t = \frac{h}{2}(x(0) + x(1)) + h \sum_{k=1}^{n-1} x(t_k) + \mathscr{O}(h^2),$$

Using this setup, we can now write the discretized form of the problem.

16.1.2 Discretized Problem Formulation

Substituting the finite difference approximation into the functional, the discretized problem becomes

$$\begin{aligned}
\text{minimize} \quad & \mathscr{A}_h(\boldsymbol{x}) = h \sum_{k=1}^{n} \left(\frac{x_k - x_{k-1}}{h} \right)^2, \\
\text{subject to} \quad & h_1(\boldsymbol{x}) = \frac{h}{2}(x_0 + x_n) + h \sum_{k=0}^{n} x_k - 1 = 0, \\
& h_2(\boldsymbol{x}) = x_0 - 2 = 0, \\
& h_3(\boldsymbol{x}) = x_n - 1 = 0.
\end{aligned}$$

This formulation transforms our continuous problem into a finite-dimensional optimization problem, where we seek values of $\boldsymbol{x}$ that minimize $\mathscr{A}_h(\boldsymbol{x})$ while satisfying the discretized constraints.

16.1.3 Lagrange function formulation

Let λ_1, λ_2, and λ_3 be the Lagrange multipliers corresponding to constraints $h_1(\boldsymbol{x})$, $h_2(\boldsymbol{x})$, and $h_3(\boldsymbol{x})$, respectively. We define the **Lagrange function** as

$$\begin{aligned}
\mathscr{L}(\boldsymbol{x}, \lambda_1, \lambda_2, \lambda_3) = & \frac{1}{h} \sum_{k=1}^{n} (x_k - x_{k-1})^2 \\
& - \lambda_1 \left(\frac{h}{2}(x_0 + x_n) + h \sum_{k=1}^{n-1} x_k - 1 \right) \\
& - \lambda_2 \cdot (x_0 - 2) + \lambda_3 \cdot (x_n - 1).
\end{aligned}$$

16.1.4 First-Order conditions

To find the stationary points of $\mathscr{L}$, we compute the partial derivatives with respect to each x_k and set them equal to zero:

1. **For $k = 1, 2, \ldots, n-1$:**

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial x_k} &= \frac{2}{h}(x_k - x_{k-1}) - \frac{2}{h}(x_{k+1} - x_k) - \lambda_1 \cdot h \\
&= h \left(\frac{2}{h^2}(2x_k - x_{k-1} - x_{k+1}) - \lambda_1 \right) = 0.
\end{aligned}$$

2. **For $k = 0$ (boundary point):**

$$\frac{\partial \mathscr{L}}{\partial x_0} = \frac{2}{h}(x_1 - x_0) - \lambda_1 \cdot \frac{h}{2} - \lambda_2 = 0.$$

3. **For $k = n$ (boundary point):**

$$\frac{\partial \mathscr{L}}{\partial x_n} = -\frac{2}{h}(x_n - x_{n-1}) - \lambda_1 \cdot \frac{h}{2} - \lambda_3 = 0.$$

These equations, combined with the constraints, form a system of equations that can be solved for $\boldsymbol{x}$, λ_1, λ_2, and λ_3.

$$\begin{cases} 2\dfrac{x_{k+1}-2x_k+x_{k-1}}{h^2}+\lambda_1=0. \\ x_0=2 \\ x_n=1 \\ \dfrac{2}{h}(x_1-x_0)-\lambda_1\cdot\dfrac{h}{2}=\lambda_2 \\ -\dfrac{2}{h}(x_n-x_{n-1})-\lambda_1\cdot\dfrac{h}{2}=\lambda_3 \end{cases}$$

notice that λ_2 and λ_3 are trivially solved and can be removed to the nonlinear system to obtain

$$\begin{cases} 2\dfrac{x_{k+1}-2x_k+x_{k-1}}{h^2}+\lambda_1=0. \\ x_0=2 \\ x_n=1 \end{cases}$$

which can be interpreted as the discrete approximation of the Boundary Value Problem (BVP)

$$\begin{cases} 2x''+\lambda_1=0. \\ x(0)=2 \\ x(1)=1 \end{cases}$$

solving this last BVP

$$x(t)=-\frac{\lambda_1}{4}t^2+C_1t+C_2$$

using BC

$$\begin{aligned} x(0)=C_2=2, &\qquad C_2=2, \\ x(1)=-\frac{\lambda_1}{4}+C_1+C_2=1, &\qquad C_1=\frac{\lambda_1}{4}-1 \end{aligned}$$

Using integral constraint

$$\begin{aligned} 0&=\int_0^1 x(t)\,\mathrm{d}t-1 \\ &=\int_0^1 -\frac{\lambda_1}{4}t^2+\left(\frac{\lambda_1}{4}-1\right)t+2\,\mathrm{d}t-1 \\ &=-\frac{\lambda_1}{12}+\frac{\lambda_1-4}{8}+2-1 \\ &=\frac{\lambda_1}{24}+\frac{1}{2} \end{aligned}$$

so that $\lambda_1=-12$ and

$$x(t)=-\frac{\lambda_1}{4}t^2+C_1t+C_2=3t^2-4t+2$$

the same solution.

16.2 Heuristic deduction of integral constraint for functional

16.2.1 A simple example

The finite difference approach motivates solving the problem

$$\begin{aligned} \text{minimize} \quad & \mathscr{A}(x) = \int_0^1 (x')^2 \,\mathrm{d}t, \\ \text{subject to} \quad & \int_0^1 x \,\mathrm{d}t = 1, \\ & x(0) = 0, \quad x(1) = 2, \end{aligned}$$

by introducing a Lagrangian functional that incorporates the integral constraint. We define:

$$\begin{aligned} \mathscr{L}(x,\lambda) &= \mathscr{A}(x) - \lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right) \\ &= \int_0^1 (x')^2 \,\mathrm{d}t - \lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right). \end{aligned}$$

The stationary points of this Lagrangian, $\mathscr{L}(x,\lambda)$, are candidates for the minimizers. To find these stationary points, we perform the **first variation** of $\mathscr{L}$.

16.2.2 First Variation of $\mathscr{L}$

The variation of $\mathscr{L}(x,\lambda)$ with respect to both x and λ is:

$$\begin{aligned} \delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda) &= \delta \int_0^1 (x')^2 \,\mathrm{d}t - \delta\lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right) - \lambda\delta \int_0^1 x \,\mathrm{d}t \\ &= \int_0^1 2x'\delta x' \,\mathrm{d}t - \delta\lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right) - \lambda \int_0^1 \delta x \,\mathrm{d}t. \end{aligned}$$

Using integration by parts on the term involving x' and using $\delta x(0) = \delta x(1) = 0$, we get:

$$\begin{aligned} \delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda) &= \int_0^1 2x'\delta x' - \lambda\delta x \,\mathrm{d}t - \delta\lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right) \\ &= -\int_0^1 (2x'' + \lambda)\delta x \,\mathrm{d}t - \cancel{2x'\delta x\Big|_{t=0}^{t=1}} - \delta\lambda \left(\int_0^1 x \,\mathrm{d}t - 1 \right). \end{aligned}$$

Applying the **fundamental lemma of calculus of variations** yields the following conditions:

$$\begin{cases} 2x'' + \lambda = 0, \\ x(0) = 0, \\ x(1) = 2, \\ \displaystyle\int_0^1 x(t) \,\mathrm{d}t = 1. \end{cases}$$

Solving this boundary value problem gives the same solution as obtained before.

16.2.3 General Solution

Consider a variational problem with an objective functional and one or more integral constraints. The problem can be stated as follows:

16.2.3.1 Problem Formulation

Given:

$$\begin{aligned} \text{minimize} \quad & \mathscr{F}(x) = \int_a^b F(x,x',x'',\ldots,x^{(n)},t) \,\mathrm{d}t, \\ \text{subject to} \quad & \mathscr{G}(x) = \int_a^b G(x,x',x'',\ldots,x^{(n)},t) \,\mathrm{d}t = 0. \end{aligned}$$

we want to find functions $x(t)$ that minimize the objective functional $\mathscr{F}(x)$ while satisfying the constraint $\mathscr{G}(x)$.

16.2.3.2 Solution Using the Lagrangian Functional

To solve this constrained problem, we introduce a **Lagrange multiplier** λ and form a **Lagrangian functional** $\mathscr{L}(x,\lambda)$:

$$\begin{aligned}\mathscr{L}(x,\lambda) &= \mathscr{F}(x) - \lambda \mathscr{G}(x) \\ &= \int_a^b \left(F(x,x',x'',\ldots,x^{(n)},t) - \lambda G(x,x',x'',\ldots,x^{(n)},t) \right) \mathrm{d}t.\end{aligned}$$

16.2.3.3 Stationary Points of the Lagrangian Functional

The functions $x(t)$ that are candidates for minimizing $\mathscr{F}(x)$ must be stationary points of the Lagrangian $\mathscr{L}(x,\lambda)$. To find these stationary points, we compute the **first variation** of $\mathscr{L}(x,\lambda)$ with respect to $x(t)$ and λ, and set it to zero. This will lead to a set of Euler-Lagrange equations that must be satisfied by $x(t)$ and λ.

16.2.4 Extension to Multiple Constraints

If the problem includes multiple integral constraints, we can generalize the formulation. Let the constraints be given by:

$$\begin{aligned}\text{minimize} \quad & \mathscr{F}(x) = \int_a^b F(x,x',x'',\ldots,x^{(n)},t)\,\mathrm{d}t, \\ \text{subject to} \quad & \mathscr{G}_1(x) = \int_a^b G_1(x,x',x'',\ldots,x^{(n)},t)\,\mathrm{d}t = 0, \\ & \mathscr{G}_2(x) = \int_a^b G_2(x,x',x'',\ldots,x^{(n)},t)\,\mathrm{d}t = 0, \\ & \vdots \\ & \mathscr{G}_p(x) = \int_a^b G_p(x,x',x'',\ldots,x^{(n)},t)\,\mathrm{d}t = 0.\end{aligned}$$

16.2.4.1 Lagrangian for Multiple Constraints

To incorporate these constraints, we introduce a Lagrange multiplier λ_k for each constraint $\mathscr{G}_k(x)$, resulting in the Lagrangian functional:

$$\begin{aligned}\mathscr{L}(x,\lambda_1,\lambda_2,\ldots,\lambda_p) &= \mathscr{F}(x) - \sum_{k=1}^{p} \lambda_k \mathscr{G}_k(x) \\ &= \int_a^b \left(F(x,x',x'',\ldots,x^{(n)},t) - \sum_{k=1}^{p} \lambda_k G_k(x,x',x'',\ldots,x^{(n)},t) \right) \mathrm{d}t.\end{aligned}$$

16.2.4.2 Finding Stationary Points

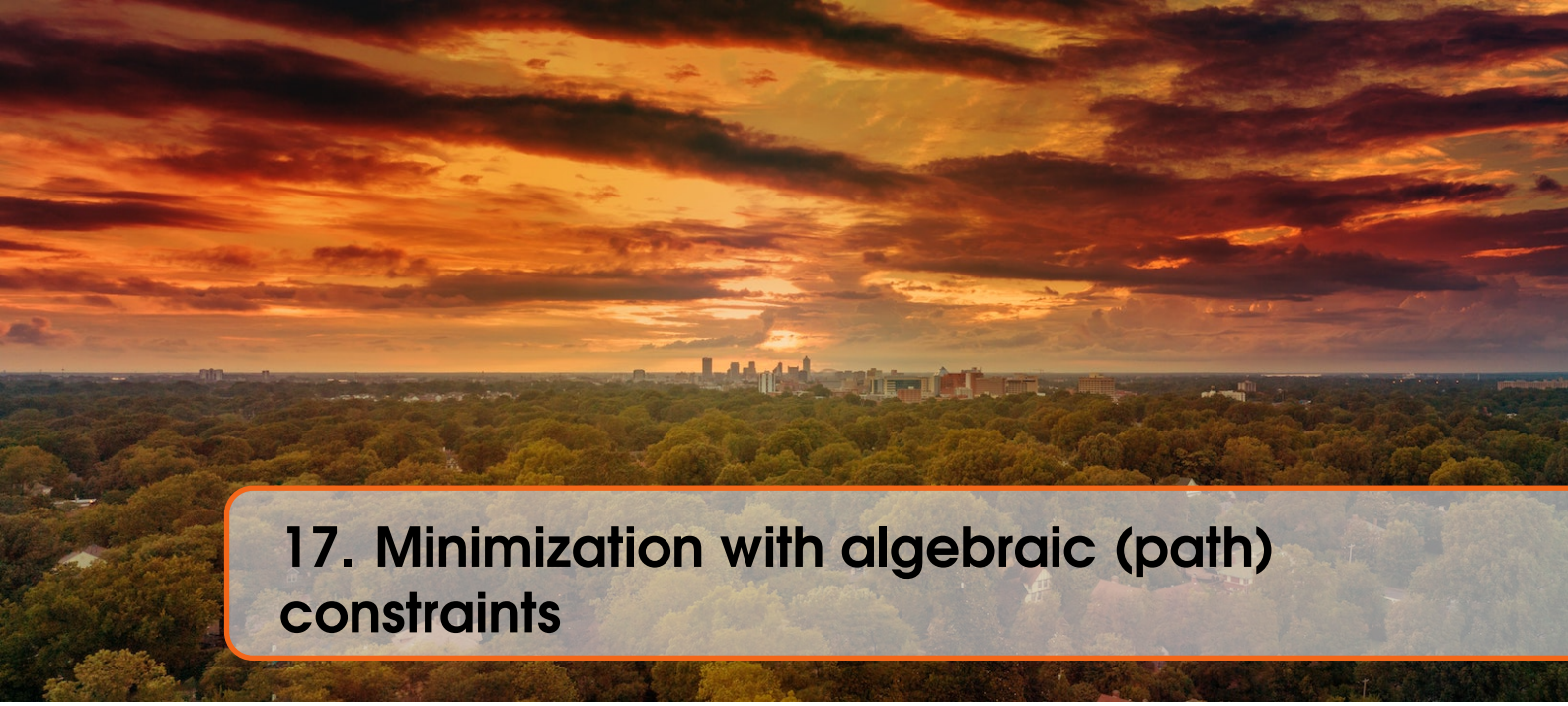
The stationary points of $\mathscr{L}$ with respect to $x(t)$ and λ_k (for $k = 1,2,\ldots,p$) provide candidates for the solution. Setting the first variations of $\mathscr{L}(x,\lambda_1,\ldots,\lambda_p)$ to zero leads to a system of equations analogous to the Euler-Lagrange equations for unconstrained problems, now modified to account for the integral constraints.

16.2.4.3 Conclusion

This approach provides a general framework for handling variational problems with one or more integral constraints. By introducing appropriate Lagrange multipliers and forming the Lagrangian

functional, we can convert a constrained variational problem into an unconstrained problem whose solutions must satisfy a system of differential equations determined by the first variation of the Lagrangian.

These equations, together with the boundary conditions, yield the function $x(t)$ that minimizes $\mathscr{F}(x)$ under the given integral constraints.



17. Minimization with algebraic (path) constraints

We consider a variational problem in which we seek to minimize an objective functional with respect to a **path constraint**—an algebraic constraint that must hold continuously along the entire path of the function $x(t)$.

Problem formulation

The problem can be stated as follows:

$$\begin{aligned} &\text{minimize} && \mathscr{F}(x) = \int_a^b L(x,x',t)\,\mathrm{d}t, \\ &\text{subject to} && g(x,x',t) = 0. \end{aligned}$$

where $L(x,x',t)$ is the Lagrangian, which depends on $x(t)$, its derivative $x'(t)$, and t; and $g(x,x',t) = 0$ represents the algebraic (path) constraint that must be satisfied at each point $t \in [a,b]$.

Defining the Feasible Space

We define a feasible space $\mathbb{W}$ for x, which includes all functions that satisfy both the path constraint $g(x,x',t) = 0$ and the boundary conditions of the problem:

$$\mathbb{W} = \left\{ x \in C^1([a,b]) \mid g(x,x',t) = 0,\ t \in [a,b] \right\}.$$

The goal, then, is to find an $x \in \mathbb{W}$ that minimizes $\mathscr{F}(x)$.

Feasible Directions and Conditions

To analyze feasible directions, we introduce a perturbation of the form $x + \alpha\delta x$, where δx is a variation such that:

$$g(x(t),x'(t),t) = 0, \quad \text{for all } t \in [a,b].$$

To obtain the variation of $\mathscr{F}(x)$, we compute:

$$\begin{aligned} \delta\mathscr{F}(x;\delta x) &= \int_a^b \left(\frac{\partial L}{\partial x}(x,x',t)\delta x + \frac{\partial L}{\partial x'}(x,x',t)\delta x' \right) \mathrm{d}t \\ &= \int_a^b \left(\frac{\partial L}{\partial x}(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial x'}(x,x',t) \right) \right) \delta x\,\mathrm{d}t \\ &\quad + \left[\frac{\partial L}{\partial x'}(x,x',t)\delta x \right]_{t=a}^{t=b}. \end{aligned}$$

However, for δx to be feasible, it must also satisfy the constraint:

$$\delta g(x,x',t) = 0, \quad \Rightarrow \quad \frac{\partial g}{\partial x}(x,x',t)\delta x + \frac{\partial g}{\partial x'}(x,x',t)\delta x' = 0.$$

this approach is impratical, thus, we try to use finite difference.

Discretization with Finite Differences

To discretize the problem, we divide the interval $[a,b]$ into n subintervals of width $h = \frac{b-a}{n}$ and define points $t_k = a + kh$ for $k = 0,1,\ldots,n$. Let x_k approximate the solution at each point t_k, so that $x_k \approx x(t_k)$.

1. **First Derivative Approximation:** We approximate the first derivative $x'(t)$ at each point t_k using a finite difference formula:

   $$x'(t_{k+1/2}) = \frac{x(t_{k+1}) - x(t_k)}{h} + \mathscr{O}(h^2)$$

   (check using Taylor series)
2. **Midpoint Approximation:** We approximate the midipoint $x(t_{k+1/2})$ using left and rigth values:

   $$x(t_{k+1/2}) = \frac{x(t_{k+1}) + x(t_k)}{2} + \mathscr{O}(h^2)$$

   (check using Taylor series)
3. **Applying the Constraint:** We impose the path constraint $g(x,x',t) = 0$ at each (mid-)point $t_{k+1/2}$ by substituting the finite difference approximation:

   $$g\left(\frac{x_{k+1}+x_k}{2}, \frac{x_{k+1}-x_k}{h}, \frac{t_{k+1}+t_k}{2}\right) = 0.$$

   This transforms the continuous path constraint into a set of discrete constraints, one at each point $t_{k+1/2}$.

Discrete Formulation of the Objective Functional

To approximate the integral $\mathscr{F}(x)$, we use the midpoint rule:

$$\mathscr{F}_h(\boldsymbol{x}) = h\sum_{k=0}^{n-1} L\left(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}\right),$$

$$x_{k+1/2} = \frac{x_{k+1}+x_k}{2}, \quad x'_{k+1/2} = \frac{x_{k+1}-x_k}{h}, \quad t_{k+1/2} = \frac{t_{k+1}+t_k}{2},$$

Our goal is now to minimize $\mathscr{F}_h(\boldsymbol{x})$ with respect to $\boldsymbol{x} = (x_0, x_1, \ldots, x_n)$, subject to the discrete constraints on $g(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}) = 0$ for each k.

Lagrange Multiplier Method for Discrete Problem

We introduce Lagrange multipliers λ_k at each point t_k to incorporate the constraints. The discrete Lagrangian becomes:

$$\begin{aligned}\mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda}) &= \mathscr{F}_h(\boldsymbol{x}) - \sum_{k=0}^{n-1} \lambda_k \cdot g\left(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}\right). \\ &= h\sum_{k=0}^{n-1} L\left(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}\right) - \sum_{k=0}^{n-1} \lambda_k \cdot g\left(x_{k+1/2}, x_{k+1/2}, t_{k+1/2}\right) \\ &= h\sum_{k=0}^{n-1} M\left(x_{k+1/2}, x'_{k+1/2}, \lambda_{k+1/2}, t_{k+1/2}\right)\end{aligned}$$

where

$$M(x,x',\lambda,t) = L(x,x',t) - \lambda g(x,x',t)$$

The nonlinear system

To find stationary points of $\mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})$, we differentiate with respect to each x_k and λ_k, leading to a system of equations:

1. **Derivative with respect to x_k:**

$$\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_k} = 0, \qquad k = 0,1,\ldots,n$$

2. **Constraint equations for each k:**

$$g\left(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}\right) = 0, \qquad k = 0,1,\ldots,n-1$$

in details:

$$\begin{aligned}\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_0} &= h\frac{\partial}{\partial x_0} M\left(x_{1/2}, x'_{1/2}, \lambda_{1/2}, t_{1/2}\right) \\ &= \frac{h}{2}\frac{\partial M}{\partial x}\left(x_{1/2}, x'_{1/2}, \lambda_{1/2}, t_{1/2}\right) - \frac{\partial M}{\partial x'}\left(x_{1/2}, x'_{1/2}, \lambda_{1/2}, t_{1/2}\right) \\ \frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_n} &= h\frac{\partial}{\partial x_n} M\left(x_{n-1/2}, x'_{n-1/2}, \lambda_{n-1/2}, t_{n-1/2}\right) \\ &= \frac{h}{2}\frac{\partial M}{\partial x}\left(x_{n-1/2}, x'_{n-1/2}, \lambda_{n-1/2}, t_{n-1/2}\right) + \frac{\partial M}{\partial x'}\left(x_{n-1/2}, x'_{n-1/2}, \lambda_{n-1/2}, t_{n-1/2}\right)\end{aligned}$$

for $k = 1,2,\ldots n-1$

$$\begin{aligned}\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_k} &= h\frac{\partial}{\partial x_k} M\left(x_{k-1/2}, x'_{k-1/2}, \lambda_{k-1/2}, t_{k-1/2}\right) + h\frac{\partial}{\partial x_k} M\left(x_{k+1/2}, x'_{k+1/2}, \lambda_{k+1/2}, t_{k+1/2}\right) \\ &= \frac{h}{2}\left[\frac{\partial M}{\partial x}\left(x_{k-1/2}, x'_{k-1/2}, \lambda_{k-1/2}, t_{k-1/2}\right) + \frac{\partial M}{\partial x}\left(x_{k+1/2}, x'_{k+1/2}, \lambda_{k+1/2}, t_{k+1/2}\right)\right] \\ &+ \left[\frac{\partial M}{\partial x'}\left(x_{k-1/2}, x'_{k-1/2}, \lambda_{k-1/2}, t_{k-1/2}\right) - \frac{\partial M}{\partial x'}\left(x_{k+1/2}, x'_{k+1/2}, \lambda_{k+1/2}, t_{k+1/2}\right)\right]\end{aligned}$$

consider now the discrete nonlinear system

$$\begin{cases} \dfrac{1}{h}\dfrac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_k} = 0, & k = 1,2,\ldots,n-1 \\ \dfrac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_0} = 0 \\ \dfrac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_n} = 0 \\ g\left(x_{k+1/2}, x'_{k+1/2}, t_{k+1/2}\right) = 0, & k = 0,1,\ldots,n-1 \end{cases}$$

As $h \to 0$, this finite difference approximation corresponds to the following approxinations

$$\begin{aligned}
\frac{1}{h}\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_k} &\approx \left[\frac{\partial M(x,x',\lambda,t)}{\partial x} + \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial M(x,x',\lambda,t)}{\partial x'}\right]_{t=t_k} \\
\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_0} &\approx -\frac{\partial M}{\partial x'}\left(x(a),x'(a),\lambda(a),a\right) \\
\frac{\partial \mathscr{L}_h(\boldsymbol{x},\boldsymbol{\lambda})}{\partial x_n} &\approx \frac{\partial M}{\partial x'}\left(x(b),x'(b),\lambda(b),b\right) \\
g\left(x_{k+1/2},x'_{k+1/2},t_{k+1/2}\right) &\approx g\left(x,x',t\right)\Big|_{t=t_k}
\end{aligned}$$

thus the finite differencre approximation can be viewed as the discretization of the following BVP

$$\begin{cases}
\dfrac{\partial M(x,x',\lambda,t)}{\partial x} + \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial M(x,x',\lambda,t)}{\partial x'} = 0 \\
-\dfrac{\partial M}{\partial x'}\left(x(a),x'(a),\lambda(a),a\right) = 0 \\
\dfrac{\partial M}{\partial x'}\left(x(b),x'(b),\lambda(b),b\right) = 0 \\
g\left(x,x',t\right) = 0
\end{cases}$$

17.1 Heuristic deduction of path constraints

In this section, we develop an analogy for handling constrained variational problems, where we aim to minimize an objective functional subject to an algebraic (or path) constraint. This technique introduces an augmented functional by incorporating a Lagrange multiplier to enforce the constraint continuously over the interval.

17.1.1 Problem Statement

We start with the following constrained minimization problem:

$$\begin{aligned}
&\text{minimize} \quad \mathscr{F}(x) = \int_a^b L(x,x',t)\,\mathrm{d}t, \\
&\text{subject to} \quad g(x,x',t) = 0.
\end{aligned}$$

Where:

- $\mathscr{F}(x)$ is the objective functional, with $L(x,x',t)$ as the Lagrangian depending on the function $x(t)$, its derivative $x'(t)$, and the independent variable t.
- The path constraint $g(x,x',t) = 0$ must hold continuously over the interval $[a,b]$.

17.1.2 Formulating the augmented functional

To incorporate the constraint into the minimization process, we construct an augmented functional $L(x,\lambda)$ by introducing a Lagrange multiplier function $\lambda(t)$ that enforces the constraint $g(x,x',t) = 0$ at every point in the interval:

$$\mathscr{L}(x,\lambda) = \int_a^b M(x,x',\lambda,t)\,\mathrm{d}t$$

where we define the augmented Lagrangian (or the modified integrand) as:

$$M(x,x',\lambda,t) = L(x,x',t) - \lambda g(x,x',t).$$

17.1.3 Deriving the necessary conditions via first variation

To find stationary points of $\mathscr{L}(x,\lambda)$, we calculate the first variation $\delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda)$ with respect to x and λ:

$$\delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda)=\int_a^b \frac{\partial M}{\partial x}\delta x+\frac{\partial M}{\partial x'}\delta x'+\frac{\partial M}{\partial \lambda}\delta\lambda\,\mathrm{d}t$$

To further simplify, we integrate by parts on the term involving $\delta x'$:

$$\delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda)=\int_a^b \left[\frac{\partial M}{\partial x}-\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial M}{\partial x'}\right]\delta x+\frac{\partial M}{\partial \lambda}\delta\lambda\,\mathrm{d}t+\frac{\partial M}{\partial x'}\delta x\Big|_{t=a}^{t=b}$$

17.1.4 Applying the fundamental Lemma of calculus of variations

To ensure that $\delta\mathscr{L}(x,\lambda;\delta x,\delta\lambda)=0$ for arbitrary variations δx and $\delta\lambda$, we apply the Fundamental Lemma of Calculus of Variations. This yields the following system of Euler-Lagrange equations:

$$\begin{cases}\dfrac{\partial M}{\partial x}(x,x',\lambda,t)-\dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\partial M}{\partial x'}(x,x',\lambda,t)=0\\ \dfrac{\partial M}{\partial x'}(x(b),x'(tb),\lambda(b),b)=0\\ -\dfrac{\partial M}{\partial x'}(x(a),x'(a),\lambda(a),a)=0\\ \dfrac{\partial M}{\partial \lambda}(x,x',\lambda,t)=g(x,x',\lambda,t)=0\end{cases}$$

that is the previous result obtained with finite difference.



18. Summary

1. **Notation**

$$\left.\frac{\mathrm{d}}{\mathrm{d}\alpha}\right|_{\alpha=0} = \delta$$

2. **Problem: minimize**

$$\int_a^b L(x,x',t)\,\mathrm{d}t \; + \; \Phi(x(a),x(b))$$

3. **Boundary conditions:**

$$x(a) = x_a, \quad x(b) = \text{free}, \quad b_k(x(a),x(b)) = 0, \cdots$$

4. **Integral contraints:**

$$\int_a^b g_k(x,x',t)\,\mathrm{d}t \; + \; \Psi_k(x(a),x(b)) = 0$$

5. **Higher order problem: minimize**

$$\int_a^b L(x,x',x'',\ldots,x^{(n)},t)\,\mathrm{d}t \; + \; \Phi(x(a),x(b))$$

18.1 Spaces

- $\mathbb{V} = \left\{ x \in C^m([a,b]), \ldots \right\}$
- $\mathbb{D} = \left\{ \delta x \in C^\infty([a,b]), \delta x(a) = \delta x(b) = 0, \delta x'(a) = \delta x'(b) = 0, \ldots \right\}$

18.1.1 Variation of problem 2.

$$
\begin{aligned}
\mathscr{F}(x) &= \int_a^b L(x,x',t)\,\mathrm{d}t + \Phi(x(a),x(b),x'(a),x'(b)) \\
\delta\mathscr{F}(x) &= \delta\int_a^b L(x,x',t)\,\mathrm{d}t + \delta\Phi(x(a),x(b),x'(a),x'(b)) \\
&= \int_a^b \delta L(x,x',t)\,\mathrm{d}t + \delta\Phi(x(a),x(b)) \\
&= \int_a^b L_x(x,x',t)\delta x + L_{x'}(x,x',t)\delta x'\,\mathrm{d}t \\
&\quad + \Phi_{x_a}(x(a),x(b))\delta x(a) + \Phi_{x_b}(x(a),x(b))\delta x(b)
\end{aligned}
$$

integration by part

$$
\begin{aligned}
&= \int_a^b \left[L_x(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}L_{x'}(x,x',t)\right]\delta x\,\mathrm{d}t \\
&\quad + \left(\Phi_{x_a}(x(a),x(b)) - L_{x'}(x(a),x'(a),a)\right)\delta x(a) \\
&\quad + \left(\Phi_{x_b}(x(a),x(b)) + L_{x'}(x(b),x'(b),b)\right)\delta x(b)
\end{aligned}
$$

18.1.2 Variation of problem 3.

Define

$$
\begin{aligned}
\mathscr{F}(x) &= \int_a^b L(x,x',t)\,\mathrm{d}t - \sum_{k+1}^{p}\lambda_k b_k(x(a),x(b)) \\
\delta\mathscr{F}(x) &= \delta\int_a^b L(x,x',t)\,\mathrm{d}t + \delta\sum_{k+1}^{p}\lambda_k b_k(x(a),x(b)) \\
&= \int_a^b L_x(x,x',t)\delta x + L_{x'}(x,x',t)\delta x'\,\mathrm{d}t - \sum_{k=1}^{p} b_k(x(a),x(b))\delta\lambda_k \\
&\quad - \sum_{k=1}^{p}\lambda_k\left(\frac{\partial b_k}{\partial x_a}(x(a),x(b))\delta x(a) + \frac{\partial b_k}{\partial x_b}(x(a),x(b))\delta x(b)\right)
\end{aligned}
$$

integration by part

$$
\begin{aligned}
&= \int_a^b \left[L_x(x,x',t) - \frac{\mathrm{d}}{\mathrm{d}t}L_{x'}(x,x',t)\right]\delta x\,\mathrm{d}t - \sum_{k=1}^{p} b_k(x(a),x(b))\delta\lambda_k \\
&\quad - \left(\sum_{k=1}^{p}\lambda_k\frac{\partial b_k}{\partial x_a}(x(a),x(b)) + L_{x'}(x(a),x'(a),a)\right)\delta x(a) \\
&\quad - \left(\sum_{k=1}^{p}\lambda_k\frac{\partial b_k}{\partial x_b}(x(a),x(b)) - L_{x'}(x(b),x'(b),b)\right)\delta x(b)
\end{aligned}
$$

18.1.3 Variation of problem 4.

Define

$$M(x,x',\boldsymbol{\lambda},t) = L(x,x',t) - \sum_{k=1}^{m} \lambda_k g_k(x,x',t)$$
$$N(x_a,x_b,\boldsymbol{\lambda}) = \Phi(x_a,x_b) - \sum_{k=1}^{m} \lambda_k \Psi_k(x_a,x_b)$$

$$\begin{aligned}
\mathscr{F}(x,\boldsymbol{\lambda}) &= \int_a^b M(x,x',\boldsymbol{\lambda},t)\,\mathrm{d}t + N(x(a),x(b),\boldsymbol{\lambda}) \\
\delta\mathscr{F}(x,\boldsymbol{\lambda}) &= \delta \int_a^b M(x,x',\boldsymbol{\lambda},t)\,\mathrm{d}t + \delta N(x(a),x(b),\boldsymbol{\lambda}) \\
&= \int_a^b M_x(x,x',\boldsymbol{\lambda},t)\delta x + M_{x'}(x,x',\boldsymbol{\lambda},t)\delta x' + M_{\boldsymbol{\lambda}}(x,x',\boldsymbol{\lambda},t)\cdot\delta\boldsymbol{\lambda}\,\mathrm{d}t \\
&\quad + N_{x_a}(x(a),x(b),\boldsymbol{\lambda})\delta x(a) + N_{x_b}(x(a),x(b),\boldsymbol{\lambda})\delta x(b) + N_{\boldsymbol{\lambda}}\cdot\delta\boldsymbol{\lambda}
\end{aligned}$$

integration by part

$$\begin{aligned}
&= \int_a^b \left[M_x(x,x',\boldsymbol{\lambda},t) - \frac{\mathrm{d}}{\mathrm{d}t} M_{x'}(x,x',t) \right] \delta x\,\mathrm{d}t \\
&\quad - \sum_{k=1}^{m} \delta\lambda_k \left[\int_a^b g_k(x,x',t)\,\mathrm{d}t + \Psi_k(x(a),x(b)) \right] \\
&\quad + \Big(N_{x_a}(x(a),x(b),\boldsymbol{\lambda}) - M_{x'}(x(a),x'(a),\boldsymbol{\lambda},a) \Big)\delta x(a) \\
&\quad + \Big(N_{x_b}(x(a),x(b),\boldsymbol{\lambda}) + M_{x'}(x(b),x'(b),\boldsymbol{\lambda},b) \Big)\delta x(b)
\end{aligned}$$

18.1.4 Problem 5

Transform to optimal control problem

$$\text{minimize:} \quad \int_a^b L(x_1,x_2,x_3,\ldots,x_n,u,t)\,\mathrm{d}t$$

$$\text{subject to:} \quad \begin{cases} x_1' = x_2 \\ x_2' = x_3 \\ \quad\vdots \\ x_{n-1}' = x_n \\ x_n' = u \end{cases}$$

18.2 General Problem

$$\text{minimize:} \quad \int_a^b L(x,x',t)\,\mathrm{d}t \;+\; \Phi(x(a),x(b))$$

$$\text{subjedt to:} \quad b_k(x(a),x(b)) = 0, \qquad k = 1,2,\ldots,p$$

$$\int_a^b g_k(x,x',t)\,\mathrm{d}t \;+\; \Psi_k(x(a),x(b)) = 0, \qquad k = 1,2,\ldots,m$$

defining

$$M(x,x',\boldsymbol{\lambda},t) = L(x,x',t) - \sum_{k=1}^{m} \lambda_k g_k(x,x',t)$$

$$N(x_a,x_b,\boldsymbol{\lambda},\boldsymbol{\mu}) = \Phi(x_a,x_b) - \sum_{k=1}^{m} \lambda_k \Psi_k(x_a,x_b) - \sum_{k=1}^{p} \mu_k b_k(x_a,x_b)$$

the stationary points satisfy the BVP

$$\begin{cases} M_x(x,x',\boldsymbol{\lambda},t) - \dfrac{\mathrm{d}}{\mathrm{d}t} M_{x'}(x,x',\boldsymbol{\lambda},t) = 0 & \text{Euler Lagrange} \\ N_{x_a}(x(a),x(b),\boldsymbol{\lambda},\boldsymbol{\mu}) - M_{x'}(x(a),x'(a),\boldsymbol{\lambda},a) = 0 & \text{Adjoint BC} \\ N_{x_b}(x(a),x(b),\boldsymbol{\lambda},\boldsymbol{\mu}) + M_{x'}(x(b),x'(b),\boldsymbol{\lambda},b) = 0 & \text{Adjoint BC} \\ b_k(x(a),x(b)) = 0, & k = 1,2,\ldots,p \\ \displaystyle\int_a^b g_k(x,x',t)\,\mathrm{d}t + \Psi_k(x(a),x(b)) = 0, & k = 1,2,\ldots,m \end{cases}$$



19. Worked exercizes

19.1 Exercize 1

Find stationary point of the functional

$$\mathscr{F}(x) = \int_0^2 \left(x(t)^2 + x(t)(x'(t))^2 + tx(t)\right) \mathrm{d}t$$

with constaint

$$x(0) = 0, \qquad x(1) = 1$$

19.1.1 Applying the First Variation

We start by applying the first variation to the functional: $\mathscr{F}(x)$. To compute the variation, we first take the variation of the integrand:

$$\begin{aligned}
\delta\mathscr{F}(x) &= \delta \int_0^2 \left(x(t)^2 + x(t)(x'(t))^2 + tx(t)\right) \mathrm{d}t \\
&= \int_0^2 \delta\left(x(t)^2 + x(t)(x'(t))^2 + tx(t)\right) \mathrm{d}t \\
&= \int_0^2 \left(2x(t)\delta x(t) + \delta x(t)(x'(t))^2 + 2x(t)x'(t)\delta x'(t) + t\delta x(t)\right) \mathrm{d}t \\
&= \int_0^2 \left(\left(2x(t) + (x'(t))^2 + t\right)\delta x(t) + 2x(t)x'(t)\delta x'(t)\right) \mathrm{d}t
\end{aligned}$$

Here, we expanded the variation term by term and grouped the expressions. The blue term $2x(t)x'(t)\delta x'(t)$ is crucial for the integration by parts.

19.1.2 Integration by Parts

To eliminate the $\delta x'(t)$ term, we apply integration by parts to the expression $2x(t)x'(t)\delta x'(t)$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(2x(t)x'(t)\delta x(t)\right) &= 2x(t)x'(t)\delta x'(t) + \frac{\mathrm{d}}{\mathrm{d}t}\left(2x(t)x'(t)\right)\delta x(t) \\
&= 2x(t)x'(t)\delta x'(t) + 2\left((x'(t))^2 + x(t)x''(t)\right)\delta x(t)
\end{aligned}$$

This allows us to rewrite the variation of $\mathscr{F}(x)$ as:

$$\begin{aligned}\delta\mathscr{F}(x) &= \int_0^2 \left(2x(t) + (x'(t))^2 + t\right)\delta x(t) + 2x(t)x'(t)\delta x'(t)\,\mathrm{d}t \\ &= \int_0^2 \left(2x(t) + (x'(t))^2 + t\right)\delta x(t)\,\mathrm{d}t \\ &\quad + \left[\frac{\mathrm{d}}{\mathrm{d}t}\left(2x(t)x'(t)\delta x(t)\right) - 2\left((x'(t))^2 + x(t)x''(t)\right)\delta x(t)\right]\mathrm{d}t\end{aligned}$$

Now, after applying the integration by parts, we can group terms again:

$$\begin{aligned}\delta\mathscr{F}(x) &= \int_0^2 \left(2x(t) - (x'(t))^2 + t - 2x(t)x''(t)\right)\delta x(t)\,\mathrm{d}t \\ &\quad + \left[2x(t)x'(t)\delta x(t)\right]_0^2\end{aligned}$$

Since $\delta x(0) = \delta x(2) = 0$ (from the boundary conditions), the boundary term vanishes. Therefore, we are left with:

$$\delta\mathscr{F}(x) = \int_0^2 \left(2x(t) - (x'(t))^2 + t - 2x(t)x''(t)\right)\delta x(t)\,\mathrm{d}t$$

19.1.3 Applying the Fundamental Lemma of Calculus of Variations

For the variation $\delta\mathscr{F}(x)$ to be zero for all variations $\delta x(t)$, the integrand must vanish. Therefore, we obtain the Euler-Lagrange equation:

$$2x(t) - (x'(t))^2 + t - 2x(t)x''(t) = 0$$

19.1.4 Boundary Conditions

The problem also specifies boundary conditions:

$$x(0) = 0, \quad x(2) = 1$$

19.1.5 Final Problem

Thus, the full problem is to solve the following boundary value problem:

$$\begin{cases} 2x(t) - (x'(t))^2 + t - 2x(t)x''(t) = 0, \\ x(0) = 0, \\ x(2) = 1. \end{cases}$$

Find stationary point of the functional

$$\mathscr{F}(x) = \int_a^b x^2 + 4xx' + 2(x')^2\,\mathrm{d}t$$

with constaint

$$x(a) = x_a$$

19.1.6 Applying the First Variation

$$\mathscr{F}(x + \alpha\delta x) = \int_a^b (x + \alpha\delta x)^2 + 4(x + \alpha\delta x)(x' + \alpha\delta x') + 2(x' + \alpha\delta x')^2\,\mathrm{d}t$$

perform variation

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{F}(x+\alpha\delta x) &= \underbrace{\frac{\mathrm{d}}{\mathrm{d}\alpha}\int_a^b}_{\longleftrightarrow}\cdots\mathrm{d}t \\ &= \underbrace{\int_a^b\frac{\mathrm{d}}{\mathrm{d}\alpha}}_{\longleftrightarrow}\cdots\mathrm{d}t \\ &= \int_a^b (x+\alpha\delta x)(2\delta x+4\delta x')+4(x'+\alpha\delta x')(\delta x+\delta x')\,\mathrm{d}t\end{aligned}$$

evaluating at $\alpha = 0$:

$$\begin{aligned}\delta\mathscr{F}(x) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\mathscr{F}(x+\alpha\delta x)\Big|_{\alpha=0} \\ &= \int_a^b x(2\delta x+4\delta x')+4x'(\delta x+\delta x')\,\mathrm{d}t \\ &= \int_a^b (2x+4x')\delta x+4(x+x')\delta x'\,\mathrm{d}t\end{aligned}$$

19.1.7 Integration by Parts

For the term $4(x+x')\delta x'$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left((x+x')\delta x\right) &= (x+x')\delta x'+\frac{\mathrm{d}}{\mathrm{d}t}\left((x+x')\right)\delta x \\ &= (x+x')\delta x'+(x'+x'')\delta x\end{aligned}$$

and

$$\begin{aligned}\int_a^b (x+x')\delta x'\,\mathrm{d}t &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}t}\left((x+x')\delta x\right)\mathrm{d}t-\int_a^b (x'+x'')\delta x\,\mathrm{d}t \\ &= (x+x')\delta x\Big|_a^b-\int_a^b (x'+x'')\delta x\,\mathrm{d}t\end{aligned}$$

so that

$$\begin{aligned}\delta\mathscr{F}(x) = \int_a^b (2x+\cancel{4x'}-\cancel{4x'}-4x'')\delta x\,\mathrm{d}t+ \\ (x(b)+x'(b))\delta x(b)-\underbrace{\cancel{(x(a)+x'(a))\delta x(a)}}_{\delta x(a)=0}\end{aligned}$$

19.1.8 Applying the Fundamental Lemma of Calculus of Variations

$$\begin{aligned}\delta x(b) &\implies x(b)+x'(b)=0 \\ \delta x &\implies 2x(t)-4x''(t)=0\end{aligned}$$

19.1.9 Final Problem

$$\begin{cases} x-2x''=0 \\ x(a)=x_a \\ x(b)+x'(b)=0\end{cases}$$

The same result can be obtained using Euler-Lagrange equation

$$L(x,x',t) = x^2+4xx'+2(x')^2$$

so that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L(x,x',t)}{\partial x'}-\frac{\partial L(x,x',t)}{\partial x}&=0\\ \frac{\mathrm{d}}{\mathrm{d}t}(4x+4x')-(2x+4x')&=0\\ 4x''-2x&=0\\ -2(x-2x'')&=0\end{aligned}$$

and for the free boundary

$$0=\frac{\partial L(x,x',t)}{\partial x'}\bigg|_{x=b}=(4x+4x')\big|_{x=a}=4(x(b)+x'(b))=0$$

19.2 Exercize 3

Find stationary point of the functional

$$\mathscr{F}(x)=\int_0^1 x'(\log x')^2\mathrm{d}t$$

with no constaint

19.2.1 Applying the First Variation

$$\begin{aligned}\delta\mathscr{F}(x)&=\delta\int_0^1 x'(\log x')^2\mathrm{d}t\\ &=\int_0^1\delta\left(x'(\log x')^2\right)\mathrm{d}t\\ &=\int_0^1\delta x'(\log x')^2+2x'\log x'\frac{\delta x'}{x'}\mathrm{d}t\\ &=\int_0^1\left((\log x')^2+2\log x'\right)\delta x'\mathrm{d}t\end{aligned}$$

Remark 19.1 Do not do this wrong argument: δx is arbitrary and so that using foundamental lemma of calculus o f variation it follows

$$(\log x')^2+2\log x'=0$$

the lemma is true for δx not $\delta x'$.

19.2.2 Integration by Parts

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left[\left((\log x')^2+2\log x'\right)\delta x\right]=\frac{\mathrm{d}}{\mathrm{d}t}\left[(\log x')^2+2\log x'\right]\delta x\\ +\left((\log x')^2+2\log x'\right)\delta x'\end{aligned}$$

so that

$$\delta\mathscr{F}(x)=\left((\log x')^2+2\log x'\right)\delta x\Big|_0^1-\int_0^1\frac{\mathrm{d}}{\mathrm{d}t}\left[(\log x')^2+2\log x'\right]\delta x\,\mathrm{d}t$$

19.2.3 Applying the Fundamental Lemma of Calculus of Variations

$$\begin{aligned}
\delta x(a) &\implies (\log x'(a))^2 + 2\log x'(a) = 0 \\
\delta x(b) &\implies (\log x'(b))^2 + 2\log x'(b) = 0 \\
\delta x(t) &\implies \frac{\mathrm{d}}{\mathrm{d}t}\left[(\log x')^2 + 2\log x'\right] = 0,
\end{aligned}$$

19.2.4 Final Problem

$$\begin{cases}
(\log x')^2 + 2\log x' = C, \\
(\log x'(a))^2 + 2\log x'(a) = 0 \\
(\log x'(b))^2 + 2\log x'(b) = 0
\end{cases}$$

notice that the ODE part is now first order and the solution depend on a constant C (determined by the boundary conditions).

IV Optimal Control

General formulation from Euler-Lagrange 179

20 Optimal Control problems 183
20.1 Pontryagin maximum (minimum) principle . . . 186
20.2 Problems with linear controls and singular arc 191

21 OCP Transformation 195
21.1 Lagrange to Mayer Transformation 197
21.2 Mayer to Lagrange Transformation 199
21.3 Final Transformed Problem (Lagrange Form) . . 200
21.4 Constant Parameters to Constant Functions . . 201
21.5 Integral Constraints 203
21.6 Free-Time 205
21.7 Free-Time 2 206

22 Examples of optimal control problems 207
22.1 The Brachistochrone Problem 207
22.2 The DIDO Problem 212
22.3 An Optimal Control Problem with constraints . 218
22.4 Dubins Car 218
22.5 Four Singular Optimal Control Problems 223
22.6 Free time problems 232

23 Summary 239
23.1 The BVP 241

References 243

The **optimal control problems** can be seen as a functional minimization, called **target**, subjected to ordinary differential constraints.

Learning objectives

After this part, you should be able to formulate a Bolza optimal-control problem, convert integral constraints to augmented states, derive the adjoint system, and interpret the stationarity or minimum condition for the control.

Isoperimetric problem

Let's consider a function

$$y(x) \in [a,b] \mapsto \mathbb{R}$$

where the are known the point $y(a) = y(b) = 0$. If we consider the function as a rope having fixed length l that, by calculus I, can be computed as

$$l = \int_a^b \sqrt{1 + y'(x)^2}\,\mathrm{d}x$$

an example of optimal control problem is the one to determine the function f that minimize it's integral over the it's domain:

$$\begin{aligned} \text{minimize:} \quad & \mathscr{F}(y) = -\int_a^b y(x)\,\mathrm{d}x \\ \text{subject to:} \quad & y(a) = 0, \quad y(b) = 0 \\ & \int_a^b \sqrt{1 + y'(x)^2}\,\mathrm{d}x - l = 0 \end{aligned}$$

Discretized solution

A way to solve the problem is by discretization, dividing the domain $[a,b]$ in n edges having length $h = \frac{b-a}{n}$ and so $x_k = a + kh$. By founding the discrete solution y_k of the system we can approximate also the continuous solution as

$$y_k \approx y(x_k)$$

Considering the integral approximation

$$\int_{x_{k-1}}^{x_k} y(x)\,\mathrm{d}x \approx h y_{k-\frac{1}{2}} = h\frac{y_k + y_{k-1}}{2}$$

and the derivative

$$y'(x_{k-\frac{1}{2}}) \approx \frac{y_k - y_{k-1}}{h}$$

With this being stated the initial isoperimetric problem can be rewritten (with some analytical simplification for the calculus) in the discretized form as

$$\begin{aligned} \text{minimize:} \quad & -\frac{h}{2}(y_0 + y_n) - h\sum_{k=1}^{n-1} y_k \\ \text{subject to:} \quad & y_0 = 0, \quad y_n = 0 \\ & l - h\sum_{k=1}^{n} \sqrt{1 + \left(\frac{y_k - y_{k-1}}{h}\right)^2} = 0 \end{aligned}$$

As here stated, this problems become a constrained minimization one that can so be solved using the Lagrange multiplier method on which the unknowns is the vector of the discretized function $\boldsymbol{y} = (y_1, \ldots, y_n)$ and the 3 Lagrange multipliers λ_0, λ_n, λ associated to the constraints:

$$\mathscr{L}(\boldsymbol{y}, \lambda_0, \lambda_n, \lambda) = -\frac{h}{2}\sum_{k=1}^{n}\left(y_k + y_{k-1}\right) - \lambda_0 y_0 - \lambda_n y_n - \lambda\left(l - h\sum_{k=1}^{n}\sqrt{1+\left(\frac{y_k - y_{k-1}}{h}\right)^2}\right)$$

By calculating the gradient of the lagrangian and setting it to zero, we satisfy the first necessary condition for the solutions that relates to the following non-linear system that can be solved by approximate solutions:

for $k = 0$:

$$-\frac{h}{2} - \lambda_0 - \lambda\frac{\frac{y_1 - y_0}{h}}{\sqrt{1+\left(\frac{y_1 - y_0}{h}\right)^2}} = 0$$

for $k = 1, 2, \ldots, n-1$

$$-h + \frac{\lambda\frac{y_k - y_{k-1}}{h}}{\sqrt{1+\left(\frac{y_k - y_{k-1}}{h}\right)^2}} - \frac{\lambda\frac{y_{k+1} - y_k}{h}}{\sqrt{1+\left(\frac{y_{k+1} - y_k}{h}\right)^2}} = 0$$

for $k = n$:

$$-\frac{h}{2} - \lambda_n + \lambda\frac{\frac{y_n - y_{n-1}}{h}}{\sqrt{1+\left(\frac{y_n - y_{n-1}}{h}\right)^2}} = 0$$

multipliers conditions:

$$\begin{cases} y_0 = 0, \qquad y_n = 0 \\ l - h\sum_{k=1}^{n}\sqrt{1+\left(\frac{y_k - y_{k-1}}{h}\right)^2} = 0 \end{cases}$$

Continuous interpretation

Considering now the limit for $h \to 0$ the discretized systems seems to become continuous; in particular the 4-th conditions become

$$y_0 \to y(a) = 0 \quad \text{and} \quad y_n \to y(b) = 0$$

The first condition (associated to $k = 0$) and the third ($k = n$) instead becomes

$$-\frac{h}{2} - \lambda_0 - \lambda\frac{\frac{y_1 - y_0}{h}}{\sqrt{1+\left(\frac{y_1 - y_0}{h}\right)^2}} \quad \to \quad -\lambda_0 - \lambda\frac{y'(a)}{\sqrt{1+y'(a)^2}} = 0$$

$$-\frac{h}{2}-\lambda_n+\lambda\frac{\frac{y_n-y_{n-1}}{h}}{\sqrt{1+\left(\frac{y_n-y_{n-1}}{h}\right)^2}} \quad \rightarrow \quad -\lambda_n+\lambda\frac{y'(b)}{\sqrt{1+y'(b)^2}}=0$$

With some analytical manipulation the second lagrange condition can be stated as

$$-1-\lambda\frac{\mathrm{d}}{\mathrm{d}x}\left.\frac{y'(x)}{\sqrt{1+y'(x)^2}}\right|_{x=x_k}=0$$

Observation

By pushing the limit $h\rightarrow 0$ the discretized problem becomes continuous and can relate to the general formulation of minimization of a functional $\mathscr{F}$ subjected to an equality constraints (depending on $\mathscr{G}$):

$$\begin{aligned}
&\text{minimize:} \quad \mathscr{F}(y)=-\int_a^b y(x)\,\mathrm{d}x \qquad \Longrightarrow \int_a^b L(y,y',t)\,\mathrm{d}t\\
&\text{subject to:} \quad y(a)=0, \quad y(b)=0\\
&\qquad\qquad \int_a^b \sqrt{1+y'(x)^2}\,\mathrm{d}x-l=0 \quad \Longrightarrow \int_a^b \mathscr{G}(y,y',y)\,\mathrm{d}t=0
\end{aligned}$$

In the isoperimetric problem we can determine the functions $L(y,y',x)=-y$ whose derivative are

$$\frac{\partial L}{\partial y}=-1 \quad \text{and} \quad \frac{\partial L}{\partial y'}=0$$

Using so the Euler-Lagrange approach we can see that it fails because

$$\frac{\partial L}{\partial y}-\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial L}{\partial y'}=-1\neq 0$$

Considering instead now the definition of the constraint

$$\mathscr{G}(y,y',x)=\frac{l}{b-a}-\sqrt{1+y'^2}$$

we can build another lagrangian $\tilde{L}$ that also considers (with a Lagrange multiplier λ) the constraint:

$$\tilde{L}(y,y',x,\lambda)=L(y,y',x)-\lambda\mathscr{G}(y,y',x)$$

Applying the Euler-Lagrange method on this expression it's possible to get the same result achieved with discretization, in fact

$$\frac{\partial\tilde{L}}{\partial y}-\frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial\tilde{L}}{\partial y'}=-1-\lambda\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{y'}{\sqrt{1+y'^2}}\right)$$

General formulation from Euler-Lagrange

Considering the optimal control problem as

$$\begin{aligned}
&\text{minimize:} && \int_a^b L(x,x',t)\,\mathrm{d}t \\
&\text{subject to:} && x(a)=x_a, \quad x(b)=x_b \\
& && \int_a^b \mathscr{G}(x,x',t)\,\mathrm{d}t = 0
\end{aligned}$$

a way to reach a solution is by constructing the lagrangian $\mathscr{L}$ defined as

$$\begin{aligned}
\tilde{L}(x,x',t,\lambda) &= L(x,x',t) - \lambda\mathscr{G}(x,x',t) \\
\mathscr{L}(x,\lambda,\lambda_a,\lambda_b) &= \int_a^b \tilde{L}(x,x',t,\lambda)\,\mathrm{d}t - \lambda_a x(a) - \lambda_b x(b)
\end{aligned}$$

At this point the solution of the problem is the stationary *point* (that in reality is a function) x of the lagrangian, and this means setting to zero the Gateaux derivative of $\mathscr{L}$:

$$\begin{aligned}
\delta\mathscr{L} = \int_a^b \left(\frac{\partial\tilde{L}}{\partial x}\delta x + \frac{\partial\tilde{L}}{\partial x'}\delta x' + \frac{\partial\tilde{L}}{\partial\lambda}\delta\lambda\right)\mathrm{d}t \\
- \delta\lambda_a\big(x(a)-x_a\big) - \lambda_a\delta x(a) - \delta\lambda_b\big(x(b)-x_b\big) - \lambda_b\delta x(b)
\end{aligned}$$

Performing the integration by part in order to remove the term $\delta x'$ the derivative so becomes

$$\begin{aligned}
\delta\mathscr{L} = \int_a^b A\,\delta x\,\mathrm{d}x - \delta\lambda_a\big(x(a)-x_a\big) - \delta\lambda_b\big(x(b)-x_b\big) \\
+ B\delta x(a) + C\delta x(b)
\end{aligned}$$

where

$$A = \frac{\partial\tilde{L}}{\partial x} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\tilde{L}}{\partial x'} \quad B = -\lambda_a - \frac{\partial\tilde{L}}{\partial y'}\Big|_a \quad C = -\lambda_b + \frac{\partial\tilde{L}}{\partial y'}\Big|_b$$

At this point the overall solution of the minimization problem becomes the following ordinary differential equation system:

$$\begin{cases}
\left(\dfrac{\partial L}{\partial x} - \lambda\dfrac{\partial\mathscr{G}}{\partial x}\right) - \dfrac{\mathrm{d}}{\mathrm{d}x}\left(\dfrac{\partial L}{\partial x'} - \lambda\dfrac{\partial\mathscr{G}}{\partial x'}\right) = 0 \\
x(a)=x_a, \quad x(b)=x_b \\
-\lambda_a - \left(\dfrac{\partial L}{\partial x'}\Big|_a - \lambda\dfrac{\partial\mathscr{G}}{\partial x'}\Big|_a\right) = 0 \\
-\lambda_b + \left(\dfrac{\partial L}{\partial x'}\Big|_b - \lambda\dfrac{\partial\mathscr{G}}{\partial x'}\Big|_b\right) = 0
\end{cases}$$

Ordinary differential equation constraint

Let's consider the general formulation of the isoperimetric problem as

$$\begin{aligned}
&\text{minimize:} && \mathscr{F}(x) = \int_a^b L(x,x',t)\,\mathrm{d}t \\
&\text{subject to:} && w\big(x(a),x(b)\big) = 0 \\
& && \int_a^b \mathscr{G}(x,x',t)\,\mathrm{d}t = l
\end{aligned}$$

the problem can be solved introducing the function $z(t)$ such that $z' = \mathscr{G}(x,x',t)$: this means that $z(a)=0$ and $z(b)=l$ and so the problem if transformed, considering also the introduction of the relation $y=x'$, to the form

$$
\begin{aligned}
\text{minimize:} \quad & \mathscr{F}(x,y) = \int_a^b L(x,y,t)\,\mathrm{d}t \\
\text{subject to:} \quad & w\big(x(a),x(b)\big) = 0 \\
& z(a) = 0, \quad z(b) = l \\
& x' = y \\
& z' = \mathscr{G}(x,y,t)
\end{aligned}
$$

With the problem as here stated we refer to x and z as **states**, while y is the **control**. We can so see that an optimal control problem can be easily transform into a boundary value problem that can be solved using the technique shown on functional minimization.



20. Optimal Control problems

Considering the problem

$$\begin{aligned}
\text{minimize:} \quad & \overbrace{\underbrace{\phi\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big)}_{\text{Mayer}} + \int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t}^{\text{Lagrange}} && \text{: target} \\
& \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{Bolza}} \\
\text{subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}\big(\boldsymbol{x},\boldsymbol{u},t\big) && \text{: dynamical system} \\
& \boldsymbol{b}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big) = 0 && \text{: boundary conditions} \\
& \int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t = \boldsymbol{g}_0 && \text{: integral constraints}
\end{aligned}$$

where $\boldsymbol{x}$ is the state vector and $\boldsymbol{u}$ is the control vector. An integral constraint can be converted into a boundary condition by adding the state $\boldsymbol{z}$ defined by $\boldsymbol{z}' = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t)$. With $\boldsymbol{z}(a) = \boldsymbol{0}$, integration gives

$$z(a) = 0 \qquad z(b) = g_0$$

With that said the minimization problem becomes

$$\begin{aligned}
\text{minimize:} \quad & \phi\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big) + \int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t \\
\text{subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}\big(\boldsymbol{x},\boldsymbol{u},t\big) \\
& \boldsymbol{z}' = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t) \\
& \boldsymbol{b}\big(\boldsymbol{x}(a),\boldsymbol{x}(b)\big) = 0 \\
& \boldsymbol{z}(a) = 0 \quad \boldsymbol{z}(b) = \boldsymbol{g}_0
\end{aligned}$$

Compacting the vector of states $\boldsymbol{x}$ and added variable $\boldsymbol{z}$ we can define a new variable vector $\boldsymbol{w}$, and similarly all the dynamical system can be described by a multi-variable function F and boundary conditions B and so the problem can be compacted as

$$\begin{aligned}
\text{minimize:} \quad & \psi\big(\boldsymbol{w}(a),\boldsymbol{w}(b)\big) + \int_a^b \mathscr{M}(\boldsymbol{w},\boldsymbol{u},t)\,\mathrm{d}t \\
\text{subject to:} \quad & \boldsymbol{w}' = \boldsymbol{F}\big(\boldsymbol{w},\boldsymbol{u},t\big) \\
& \boldsymbol{B}\big(\boldsymbol{w}(a),\boldsymbol{w}(b)\big) = 0
\end{aligned}$$

To solve the problem as here state it's necessary to build the function of the boundary conditions $\mathscr{B}$ (depending by the Lagrange multipliers $\boldsymbol{\mu}$), also known as **utility function**, and the **Hamiltonian** $\mathscr{H}$ (depending by the multipliers $\boldsymbol{\lambda}$):

$$\begin{aligned}\mathscr{B}\big(\boldsymbol{w}(a),\boldsymbol{w}(b),\boldsymbol{\mu}\big) &= \psi\big(\boldsymbol{w}(a),\boldsymbol{w}(b)\big) + \boldsymbol{\mu}\cdot\boldsymbol{B}\big(\boldsymbol{w}(a),\boldsymbol{w}(b)\big)\\ \mathscr{H}(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{u},t) &= \mathscr{M}(\boldsymbol{w},\boldsymbol{u},t) + \boldsymbol{\lambda}\cdot\boldsymbol{F}(\boldsymbol{w},\boldsymbol{u},t)\end{aligned}$$

At this point it's possible to compute the lagrangian $\mathscr{L}$ on which the variations can be performed:

$$\begin{aligned}\mathscr{L}\big(\boldsymbol{w},\boldsymbol{u},\boldsymbol{\lambda},\boldsymbol{\mu},t\big) &= \mathscr{B}\big(\boldsymbol{w}(a),\boldsymbol{w}(b),\boldsymbol{\mu}\big) + \mathscr{H}(\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{u},t)\\ \delta\mathscr{L} &= \int_a^b \Big(A\delta w + B\delta u + C\delta\lambda\Big)\,\mathrm{d}t\\ &\quad + D\,\delta w(a) + E\,\delta w(b) - \boldsymbol{B}\big(w(a),w(b)\big)\,\delta\mu\end{aligned}$$

where

$$A = \lambda' + \frac{\partial H}{\partial w} \qquad B = \frac{\partial H}{\partial u} \qquad C = f(w,u,t) - x'$$

$$D = \frac{\partial B}{\partial w(a)}\Big|_a + \lambda(a) \qquad\qquad E = \frac{\partial B}{\partial w(b)}\Big|_b - \lambda(b)$$

With that said the resulting boundary valued problem becomes

$$\begin{cases} w' = f(x,u,t) & : \text{original ODE}\\ \lambda' = -\dfrac{\partial H}{\partial w} & : \text{adjoint ODE}\\ B\big(w(a),w(b)\big) = 0 & : \text{original boundary condition}\\ \left.\begin{aligned}&\dfrac{\partial B}{\partial w(a)}\Big|_a + \lambda(a) = 0\\ &\dfrac{\partial B}{\partial w(b)}\Big|_b - \lambda(b) = 0\end{aligned}\right\} & : \text{adjoint boundary conditions}\\ \dfrac{\partial H}{\partial u} = 0 & : \text{control equation}\end{cases} \tag{20.1}$$

■ **Example 20.1 — optimal control problem.** Let's consider the problem of a mass m sliding on a plane (coordinate x) subjected only to an external applied force F, the optimal control problem is

$$\begin{aligned}\text{minimize:} \quad & \int_0^1 F^2\,\mathrm{d}t\\ \text{subject to:} \quad & x' = v, \qquad v' = \frac{F}{m}\\ & x(0) = 0 \qquad x(1) = 1\\ & v(0) = 0 \qquad v(1) = 0\end{aligned}$$

In this case the description of the dynamical system is known from the physical domain, in fact the velocity is the derivative of the position and the acceleration (derivative of velocity) is equal to the force applied divided by the mass; the boundaries conditions are set by the problem.

At this point to solve the problem we have to compute the two function $\mathscr{B},\mathscr{H}$ defined as

$$\begin{aligned}\mathscr{B} &= \mu_1\big(x(a)-0\big) + \mu_2\big(x(b)-1\big) + \mu_3\big(v(a)-0\big) + \mu_4\big(v(b)-0\big)\\ \mathscr{H} &= F^2 + \lambda_1 v + \lambda_2\frac{F}{m}\end{aligned}$$

Following the results of equation Equation 20.1 the boundary valued problem that has to be solved to find the minimum point is

$$\begin{cases} x' = v \\ v' = F/m \\ \lambda_1' = 0 \\ \lambda_2' = \lambda_1 \\ x(0) = 0 \qquad x(1) = 1 \\ v(0) = 0 \qquad v(1) = 0 \\ \mu_1 + \lambda_1(0) = 0 \\ \mu_3 + \lambda_2(0) = 0 \\ \mu_2 - \lambda_1(1) = 0 \\ \mu_4 - \lambda_2(1) = 0 \\ 2F + \frac{\lambda_2}{m} = 0 \end{cases}$$

In this case the adjoin boundary conditions are trivially solved (in fact will exists μ_i that will satisfy the equation); considering the last equation in the expression of the velocity the system becomes

$$\begin{cases} x' = v \\ v' = -\frac{\lambda_2}{2m^2} \\ \lambda_1' = 0 \\ \lambda_2' = \lambda_1 \\ x(0) = 0 \qquad x(1) = 1 \\ v(0) = 0 \qquad v(1) = 0 \end{cases}$$

Considering that λ_1 is a constant c_1 (in order to have null derivative), then it means that λ_2 is in the form $c_1 t + c_2$ and so the expression of the acceleration and velocity becomes

$$\begin{aligned} v' &= -\frac{c_1 t + c_2}{2m^2} \\ \xrightarrow{\int} \quad v &= -\frac{c_1}{4m^2} t^2 - \frac{c_2}{2m^2} t + c_3 \\ \xrightarrow{\int} \quad x &= -\frac{c_1}{12m^2} t^3 - \frac{c_2}{4m^2} t^2 + c_3 t + c_4 \end{aligned}$$

Considering the boundary conditions than the integration constant are $c_3 = c_4 = 0$ and the solution of the linear system given by $-c_1 - 2c_2 = 0$ and $-c_1 - 3c_2 = 12m^2$ that determines the last full solution

$$\begin{aligned} x &= -6t^2 + 6t \\ v &= -2t^3 + 3t^2 \\ \lambda_1 &= 24m^2 \\ \lambda_2 &= 24m^2 t - 12m^2 \end{aligned}$$

and so the control history to determine this minimal solution is

$$F = -\frac{\lambda_2}{2m} = 6\big(1 - 2t\big)m$$

■

20.1 Pontryagin maximum (minimum) principle

The **Pontryagin maximum** (minimum) **principle** has been developed in order to solve problem where the controls must stay bounded to a certain range due to the physical implementation of the system (if, as example, a motor is powered with $20W$ of power, than it cannot move object that require more than that power).

Mathematically this kind of problem is described by

$$\begin{aligned} \text{minimize:} \quad & \phi\big(x(a),x(B)\big) + \int_a^b L(x,u,t)\,\mathrm{d}t \\ \text{subject to:} \quad & x' = f(x,u,t) \\ & b\big(x(a),x(b)\big) = 0 \\ & u(t) \in \mathscr{U} \end{aligned}$$

where $\mathscr{U}$ is the **domain of the controls** that, in order for the principle to work, must be convex and compact.

To solve this problem we compute the Hamiltonian $\mathscr{H}$ and the utility function $\mathscr{B}$ as in the previous cases:

$$\begin{aligned} \mathscr{H}(x,u,\lambda,t) &= L(x,u,t) + \lambda f(x,u,t) \\ \mathscr{B}(x_a,x_b,\mu) &= \phi(x_a,x_b) + \mu b(x_a,x_b) \end{aligned}$$

With that stated the boundary valued problem becomes similar to the formulation in Equation 20.1 with the addition of the Pontryagin minimum principle:

$$\begin{cases} x' = f(x,u,t) = \dfrac{\partial \mathscr{H}}{\partial \lambda} & : \text{original ODE} \\ \lambda' = -\dfrac{\partial \mathscr{H}}{\partial x} & : \text{adjoint ODE - co-equations} \\ b\big(w(a),w(b)\big) = 0 & : \text{original BC} \\ \left.\begin{array}{l} \dfrac{\partial \mathscr{B}}{x_a} + \lambda(a) = \mathbf{0} \\ \dfrac{\partial \mathscr{B}}{x_b} - \lambda(b) = \mathbf{0} \end{array}\right\} & : \text{adjoint BC} \\ u(t) = \underset{\overline{u}\in\mathscr{U}}{\operatorname{argmin}}\big\{\mathscr{H}\big(x,\overline{u},\lambda,t\big)\big\} & : \text{Pontryagin min principle} \\ \dfrac{\partial \mathscr{H}}{\partial u} = 0 & : \text{control equation} \end{cases}$$

20.1.1 Examples with Pontryagin maximum (minimum) principle

▪ **Example 20.2 — maximum travel optimal control problem.** Let's consider the system in example Example 20.1 where in this case the goal is to maximize the travel of the mass m having a force $F \leq |1|$; this means solving the following optimal control problem

$$\begin{aligned} \text{minimize:} \quad & -x(1) \\ \text{subject to:} \quad & x' = v, \qquad v' = \frac{F}{m} \\ & x(0) = 0 \\ & v(0) = 0 \qquad v(1) = 0 \\ & F \in [-1,1] \end{aligned}$$

In order to solve this problem it's necessary to determine the Hamiltonian

$$\mathscr{H}(x,v,\lambda_1,\lambda_2,F,t) = \lambda_1 v + \lambda_2 \frac{F}{m}$$

and the utility function

$$\mathscr{B}(x_a,v_a,x_b,v_b,\mu_1,\mu_2,\mu_3) = -x_b + \mu_1 x_a + \mu_2 v_a + \mu_3 v_b$$

The adjoint equation so becomes

$$\lambda_1' = -\frac{\partial \mathscr{H}}{\partial x} = 0$$

and

$$\lambda_2' = -\frac{\partial \mathscr{H}}{\partial v} = -\lambda_1$$

Among the adjoint boundary conditions, only $-1-\lambda_1(1)=0$ constrains the state–costate trajectory; the other equations determine the multipliers μ_i. The boundary-value problem therefore becomes

$$\begin{cases} x' = v \\ v' = F/m \\ \lambda_1' = 0 \\ \lambda_2' = -\lambda_1 \\ -1-\lambda_1(1) = 0 \\ x(0) = v(0) = v(1) = 0 \\ F(t) = \underset{\overline{F}\in[-1,1]}{\operatorname{argmin}} \left\{ \lambda_1(t) v + \lambda_2(t) \dfrac{\overline{F}}{m} \right\} \end{cases}$$

In this case in order to compute the argument minimum associated to the Pontryagin principle we have to consider only the term related to $\overline{F}$, and in particular this means

$$\begin{aligned} F(t) &= \underset{\overline{F}\in[-1,1]}{\operatorname{argmin}} \left\{ \lambda_2(t) \frac{\overline{F}}{m} \right\} \\ &= \begin{cases} +1 & \text{if } \lambda_2(t) < 0 \\ -1 & \text{if } \lambda_2(t) > 0 \\ [-1,1] & \text{if } \lambda_2(t) = 0 \end{cases} \\ &= -\operatorname{sign}\lambda_2(t) \end{aligned}$$

and so the simplified system that determines the solution becomes

$$\begin{cases} x' = v \\ v' = -\dfrac{\operatorname{sign}\lambda_2}{m} \\ \lambda_1' = 0 \\ \lambda_2' = -\lambda_1 \\ \lambda_1(1) = -1 \\ x(0) = v(0) = v(1) = 0 \end{cases}$$

By integration it can be seen that λ_1 can be regarded as a constant c_1 that, for the adjoint boundary condition, is equal to -1 and so λ_2 is in the form

$$\lambda_2 = t + c_2$$

From now on the solution becomes analytically complex, however it can be found that $c_2 = -\frac{1}{2}$ determining the solution

$$F(t) = \begin{cases} 1 & t \leq \frac{1}{2} \\ -1 & t > \frac{1}{2} \end{cases}$$

■

■ **Example 20.3 — optimal control problem.** Given the optimal control problem

$$\begin{aligned} \text{minimize:} \quad & \mathscr{F}(x,u) = \int_0^2 2x - u^2 \,\mathrm{d}t \\ \text{subject to:} \quad & x' = 1 - 2u \\ & x(0) = 1 \qquad x(2) = 0 \end{aligned}$$

the solution can be obtained by computing the hamiltonian $\mathscr{H}$ and the utility function $\mathscr{B}$ related to the problem:

$$\begin{aligned} \mathscr{H}(x,u,\lambda) &= 2x - u^2 - \lambda\big(1 - 2u\big) \\ \mathscr{B}(x_0,x_2,\mu_1,\mu_2) &= -\mu_1\big(x_0 - 1\big) - \mu_2 x_2 \end{aligned}$$

With that stated, the lonely resulting co-equation is $\lambda' = \frac{\partial \mathscr{H}}{\partial x} = 2$, determining so that the solution of $\lambda(t)$ is in the form $2t + c_1$ (where c_1 is a constant depending from the boundary conditions). The adjoint boundary conditions aren't useful to determine the solution because they trivially solves the parameters μ_1, μ_2. We still have the control equation that states

$$\frac{\partial \mathscr{H}}{\partial u} = -2u + 2\lambda = 0 \qquad \Longrightarrow \qquad u(t) = \lambda(t)$$

Using the original ordinary differential equation we have that $x' = 1 - 4t - 2c_1$ that by integration becomes

$$x(t) = -2t^2 + \big(1 - 2c_1\big)t + c_2$$

From the boundary condition $x(0) = 1$ we obtain $c_2 = 1$, while considering $x(2) = 0$ we have $c_1 = -5/4$ and so the final control law is

$$u(t) = 2t - \frac{5}{4}$$

■

■ **Example 20.4 — optimal control problem.** To solve the optimal control problem

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 t + x - y^2 \,\mathrm{d}t \\ \text{subject to:} \quad & x' = x + y \qquad y' = x + yu \\ & \int_0^1 \sin y \,\mathrm{d}t = 3 \\ & x(0) = 1 \qquad y(1) = 0 \\ & u \in [-2, 3] \end{aligned}$$

the first thing to do is removing the integral constraint by transforming him in a differential equation with 2 boundary condition as follows:

$$\begin{aligned}
\text{minimize:} \quad & \int_0^1 t + x - y^2 \,\mathrm{d}t \\
\text{subject to:} \quad & x' = x + y \\
& y' = x + yu \\
& z' = \sin y \\
& x(0) = 1, \qquad z(0) = 0 \\
& y(1) = 0, \qquad z(1) = 3 \\
& u \in [-2, 3]
\end{aligned}$$

The associated hamiltonian $\mathscr{H}$ and utility function $\mathscr{B}$ of the problem are

$$\begin{aligned}
\mathscr{H}(x, y, z, \lambda_1, \lambda_2, \lambda_3) &= t + x - y^2 \\
&\quad - \lambda_1(x + y) - \lambda_2(x + yu) - \lambda_3(\sin y) \\
\mathscr{B}(x_0, y_1, z_0, z_1, \mu_{1\ldots4}) &= -\mu_1\big(x_0 - 1\big) - \mu_2 y_1 \\
&\quad - \mu_3 z_0 - \mu_4\big(z_1 - 3\big)
\end{aligned}$$

With that stated the co-equations are computed as

$$\begin{aligned}
\lambda_1' &= -\frac{\partial \mathscr{H}}{\partial x} = -1 + \lambda_1 + \lambda_2 \\
\lambda_2' &= -\frac{\partial \mathscr{H}}{\partial y} = 2y + \lambda_1 + \lambda_2 u + \lambda_3 \cos y \\
\lambda_3' &= -\frac{\partial \mathscr{H}}{\partial z} = 0
\end{aligned}$$

Evaluating the adjoint boundary conditions for the point that present no explicit constraints (so x_1, y_0) we retrieve $\lambda_1(1) = 0$ and $\lambda_2(0) = 0$. With the definition of the control equation $\frac{\partial \mathscr{H}}{\partial u} = -\lambda_2 y$ the resulting boundary value problem of the optimal control one is

$$\begin{cases}
x' = x + y \\
y' = x + yu \\
z' = \sin y \\
\lambda_1' = \lambda_1 + \lambda_2 - 1 \\
\lambda_2' = 2y + \lambda_1 + \lambda_2 y + \lambda_3 \cos y \\
\lambda_3' = 0 \\
x(0) = 1 \qquad \lambda_1(1) = 0 \\
y(1) = 0 \qquad \lambda_2(0) = 0 \\
z(0) = 0 \qquad z(1) = 3 \\
-\lambda_2 = 0 \\
u = \underset{\overline{u} \in [-2,3]}{\operatorname{argmin}} \{-\lambda_2 y \overline{u}\}
\end{cases}$$

where the solution of the Pontryagin minimum principle is

$$u(t) = \begin{cases} -2 & \lambda_2 y < 0 \\ 3 & \lambda_2 y \geq 0 \end{cases}$$

■

■ **Example 20.5 — optimal control problem.** Given the optimal control problem

$$\begin{aligned}
\text{minimize:} \quad & \mathscr{F}(x,y,u) = x(0) + \int_0^1 x^2 + y^2 + u \,\mathrm{d}t \\
\text{subject to:} \quad & x' = y - u, \qquad y' = x + u \\
& x(1) = 2 \\
& u(t) \in [-1,2]
\end{aligned}$$

the solution can be obtained by firstly computing the hamiltonian $\mathscr{H}$ and the utility function $\mathscr{B}$ of the problems resulting in

$$\begin{aligned}
\mathscr{H}\big(x,y,u,\lambda_1,\lambda_2\big) &= x^2 + y^2 + u + \lambda_1\big(y-u\big) + \lambda_2\big(x+u\big) \\
&= x^2 + y^2 + u\big(1 - \lambda_1 + \lambda_2\big) + \lambda_1 y + \lambda_2 x \\
\mathscr{B}\big(x_0,x_1,\mu\big) &= x_0 + \mu\big(x_1 - 2\big)
\end{aligned}$$

We can so explicit: the adjoint ordinary differential equations

$$\begin{aligned}
\lambda_1' &= -\frac{\partial \mathscr{H}}{\partial x} = -2x - \lambda_2 \\
\lambda_2' &= -\frac{\partial \mathscr{H}}{\partial y} = -2y - \lambda_1
\end{aligned}$$

the adjoint boundary conditions

$$\begin{aligned}
\lambda_1(0) &= -\frac{\partial \mathscr{B}}{\partial x_0} = -1 \\
\lambda_1(1) &= \frac{\partial \mathscr{B}}{\partial x_1} = \mu \\
\lambda_2(0) &= -\frac{\partial \mathscr{B}}{\partial y_0} = 0 \\
\lambda_2(1) &= \frac{\partial \mathscr{B}}{\partial y_1} = 0
\end{aligned}$$

and the control equation

$$\frac{\partial \mathscr{H}}{\partial u} = 1 - \lambda_1 + \lambda_2 = 0$$

With the addition of the original differential equation the boundary value problem becomes

$$\begin{cases}
x' = y - u \\
y' = x + u \\
\lambda_1' = -2x - \lambda_2 \\
\lambda_2' = -2y - \lambda_1 \\
\lambda_1(0) = -1 \\
\cancel{\lambda_1(1) = \mu} \qquad\qquad \text{: trivially satisfied} \\
\lambda_2(0) = \lambda_2(1) = 0 \\
1 + \lambda_1 - \lambda_2 = 0 \\
u = \underset{\overline{u} \in [-1,2]}{\operatorname{argmin}} \big\{ \overline{u}\big(1 + \lambda_1 - \lambda_2\big) \big\}
\end{cases}$$

In this case the solution of the Pontryagin minimum principle is reduced to the form

$$u(t) = \begin{cases} -1 & 1+\lambda_1-\lambda_2 > 0 \\ 2 & 1+\lambda_1-\lambda_2 < 0 \end{cases}$$

■

20.2 Problems with linear controls and singular arc

In this section, we examine problems in which the Hamiltonian is a linear function of the control variable u. Although linear problems are generally expected to be easier to handle, this is not always the case. The complexity arises from the fact that the necessary condition for minimizing the Hamiltonian with respect to the control along an optimal trajectory does not yield a well-defined expression for synthesizing the optimal control.

Another challenge associated with this class of problems is the occurrence of discontinuities in the control, which can be difficult to identify. Additionally, there may be an infinite number of switching points that can accumulate at a specific time instant. Given that general results and theorems regarding existence and uniqueness are limited in this context, it becomes necessary to introduce additional relationships by manipulating other necessary conditions.

Consequently, problems with a Hamiltonian that is linear in terms of the control u represent an almost independent area of research within optimal control theory. Most theoretical results in this domain are derived using a geometric approach, involving fiber bundles and symmetry groups rather than traditional variational techniques.

To maintain clarity and simplicity in notation, we will focus on problems with a single control variable and fixed endpoints. The formulation of these problems is as follows:

Definition 20.1 Minimize the cost functional J subject to the dynamic $\boldsymbol{x}'$ with the control u constrained in $|u| \leq 1$ and the initial state $\boldsymbol{x}(0) = \boldsymbol{x}_0$, where

$$\begin{aligned} J(u) &= \int_0^T f_0(\boldsymbol{x}) + b_0(\boldsymbol{x})u \, \mathrm{d}t \\ \boldsymbol{x}'(t) &= \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{b}(\boldsymbol{x})u \qquad \text{i.e.} \end{aligned} \tag{20.2}$$

$f_0(\boldsymbol{x})$ and $b_0(\boldsymbol{x})$ are scalar functions of the state $\boldsymbol{x}$, u is the scalar control.

To further simplify the computations, we consider the scalar equation of the functional as an extra state variable posing

$$x_0' = f_0(\boldsymbol{x}) + b_0(\boldsymbol{x})u, \qquad x_0(0) = 0,$$

(we have already seen that this is not a loss of generality), but the key point here is that the control appears only linearly in both dynamics and cost functional. We can write the Hamiltonian for this system, namely

$$\mathscr{H}(\boldsymbol{x}, \boldsymbol{\lambda}, t) = \sum_{i=0}^{n} f_i(\boldsymbol{x})\lambda_i + u \sum_{i=0}^{n} b_i(\boldsymbol{x})\lambda_i.$$

It is convenient to rename the two addends as

$$\begin{aligned} A(\boldsymbol{x}, \boldsymbol{\lambda}, t) &= \sum_{i=0}^{n} f_i(\boldsymbol{x})\lambda_i, \\ B(\boldsymbol{x}, \boldsymbol{\lambda}, t) &= \sum_{i=0}^{n} b_i(\boldsymbol{x})\lambda_i, \end{aligned}$$

so that the Hamiltonian becomes

$$\mathscr{H}(\boldsymbol{x},\boldsymbol{\lambda},t) = A(\boldsymbol{x},\boldsymbol{\lambda},t) + B(\boldsymbol{x},\boldsymbol{\lambda},t)u. \tag{20.3}$$

The adjoint variables of the costate are given by the usual relation

$$\lambda_i' = -\frac{\partial \mathscr{H}}{\partial x_i} = -\sum_{j=0}^{n} \lambda_j \frac{\partial f_j(\boldsymbol{x})}{\partial x_i} - u\sum_{j=0}^{n} \lambda_j \frac{\partial b_j(\boldsymbol{x})}{\partial x_i}.$$

We state now the necessary conditions for problem Equation 20.2.

Theorem 20.1 If $u^\star(t)$ is an optimal control and if $\boldsymbol{x}^\star(t)$ is the corresponding optimal trajectory, then there are a costate $\lambda_i^\star(t)$ such that

- for $i = 1,\dots,n$

$$x_i'^\star = f_i(\boldsymbol{x}^\star) + b_i(\boldsymbol{x}^\star)u^\star,$$
$$\lambda_i'^\star = -\sum_{j=0}^{n} \lambda_j^\star \frac{\partial f_j(\boldsymbol{x}^\star)}{\partial x_i^\star} - u^\star \sum_{j=0}^{n} \lambda_j^\star(t) \frac{\partial b_j(\boldsymbol{x}^\star)}{\partial x_i^\star};$$

- for all $t \in [0,T]$ and all $u(t)$ s.t. $|u(t)| \le 1$, holds

$$A(\boldsymbol{x}^\star,\lambda^\star,t) + u^\star(t)B(\boldsymbol{x}^\star,\lambda^\star,t) \le A(\boldsymbol{x}^\star,\lambda^\star,t) + u(t)B(\boldsymbol{x}^\star,\lambda^\star,t);$$

- If T is free, then for all $t \in [0,T]$

$$\mathscr{H}^\star(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) := A(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) + u^\star(t)B(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) = 0, \tag{20.4}$$

while, if T is fixed, then

$$\mathscr{H}^\star(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) := A(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) + u^\star(t)B(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t) = c, \tag{20.5}$$

for a real constant c.

It follows easily from equation Equation 20.3 with the PMP, that the optimal control is given by

$$u^\star(t) = -\text{sign}\{B(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t)\} = -\text{sign}\left\{\sum_{i=0}^{n} b_i(\boldsymbol{x}^\star(t))\lambda_i^\star(t)\right\},$$

as long as the function $B(\boldsymbol{x}^\star,\boldsymbol{\lambda}^\star,t)$ is not zero. If B becomes zero, then the sign function is not defined. If B is zero only in a point we speak of *bang-bang* controls, if B is identically zero for an entire interval $t \in (t_1,t_2]$, the controls are called **singular**. In this case, the associated trajectory $\boldsymbol{x}(t)$ is called a **singular arc**.

20.2.1 Singular controls

If problem Equation 20.2 results singular, controls, costate and trajectory have the following property: there is at least one half-open interval $(t_1,t_2] \subset [0,T]$ such that

$$\sum_{i=0}^{n} \lambda_i^\star(t) b_i(\boldsymbol{x}^\star(t)) = 0, \qquad t \in (t_1,t_2]. \tag{20.6}$$

As before, the existence of an extremal singular control, does not imply that the **optimal** control is singular: we need additional information as uniqueness to conclude its optimality. The function Equation 20.6 is often called **switching function**.

20.2.2 Test for singular controls

Suppose we are in the free terminal time case, so that Equation 20.4 holds, we test if it is possible to have a singular control as follows. First we need to assume that the switching function is zero in an unknown interval $(t_1,t_2]$, then, because of the free end time hypothesis, for $t \in (t_1,t_2]$,

$$\mathscr{H}(\boldsymbol{x},\boldsymbol{\lambda},t) = \sum_{i=0}^{n} f_i(\boldsymbol{x})\lambda_i + u\sum_{i=0}^{n} b_i(\boldsymbol{x})\lambda_i = 0$$

$$\Downarrow$$

$$\sum_{i=0}^{n} f_i(\boldsymbol{x})\lambda_i = 0.$$

Moreover, for each $k \in \mathbb{N}$, we have that

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\sum_{i=0}^{n}\lambda_i(t)b_i(\boldsymbol{x}) = 0, \quad \begin{cases} k = 1,2,3,\ldots, \\ t \in (t_1,t_2], \end{cases} \tag{20.7}$$

and similarly,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}\sum_{i=0}^{n}\lambda_i f_i(\boldsymbol{x}) = 0, \quad \begin{cases} k = 1,2,3,\ldots, \\ t \in (t_1,t_2], \end{cases} \tag{20.8}$$

However, the canonical equations (we omit explicit time dependence) are

$$\begin{aligned} x_i' &= f_i + ub_i \\ \lambda_i' &= -\sum_{j=0}^{n}\lambda_j(t)\frac{\partial f_j}{\partial x_i} - u\sum_{j=0}^{n}\lambda_j\frac{\partial b_j}{\partial x_i}, \quad \begin{cases} i,j = 1,2,\ldots,n \\ k \in \mathbb{N}, \end{cases} \end{aligned}$$

Let $k = 1$ in Equation 20.7, then by the chain rule we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=0}^{n}\lambda_i b_i = \sum_{i=0}^{n}\left(\lambda_i' b_i + \lambda_i\sum_{j=0}^{n}\frac{\partial b_j}{\partial x_i}x_j'\right) = 0.$$

Substituting the canonical equations in the previous one, yields

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\left(f_j\lambda_i\frac{\partial b_i}{\partial x_j} - b_i\lambda_j\frac{\partial f_j}{\partial x_i}\right) + u\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i b_j\frac{\partial b_i}{\partial x_j} - \lambda_j b_i\frac{\partial b_j}{\partial x_i}\right) = 0.$$

We notice that the coefficient of u is zero, and we conclude that

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j} - f_j\frac{\partial b_i}{\partial x_j}\right) = 0. \tag{20.9}$$

Applying the same argument to equation Equation 20.8, gives

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\sum_{i=0}^{n}\lambda_i f_i(\boldsymbol{x}) \\ &= \sum_{i=0}^{n}\left(\lambda_i' f_i + \lambda_i\sum_{j=0}^{n}\frac{\partial f_i}{\partial x_j}x_j'\right) \\ &= \sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i f_i\frac{\partial f_i}{\partial x_j} - \lambda_j f_i\frac{\partial f_j}{\partial x_i}\right) + u\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\lambda_i b_j\frac{\partial f_i}{\partial x_j} - \lambda_j f_i\frac{\partial b_j}{\partial x_i}\right) \\ &= u\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j} - f_j\frac{\partial b_i}{\partial x_j}\right) = 0. \end{aligned}$$

This last line implies that either $u = 0$ or

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0.$$

Noticing that the previous line is equal to equation Equation 20.9, we can not conclude that $u = 0$, and this is true for all $k \in \mathbb{N}$.\ In conclusion, a necessary but not sufficient test for an extremal singular control is that in the interval $(t_1, t_2]$ the following relations are satisfied:

$$\left.\begin{aligned}\sum_{i=0}^{n} b_i\lambda_i = 0\\ \sum_{i=0}^{n} f_i\lambda_i = 0\\ \sum_{i=0}^{n}\sum_{j=0}^{n}\lambda_i\left(b_j\frac{\partial f_i}{\partial x_j}-f_j\frac{\partial b_i}{\partial x_j}\right)=0\end{aligned}\right\}$$

We show now how to compute the explicit expression for the singular control, we return back to the equation of the switching function Equation 20.7 and let $k = 2,3,\ldots$. It is shown in [1:pag.487] that those derivatives for $k = 2,3,\ldots$ require extensive manipulations but have all the following same structure:

$$\sum_{i=0}^{n}\lambda_i\psi_{ki}(\boldsymbol{x})+u\sum_{i=0}^{n}\lambda_i\phi_{ki}(\boldsymbol{x})=0.$$

If $\sum\lambda_i\phi_{ki}(\boldsymbol{x}) = 0$ we can perform successive derivatives of Equation 20.7, in general there will be a finite value of $k = m$ for which $\sum\lambda_i\phi_{mi}(\boldsymbol{x}) \neq 0$ and then the singular control u can be solved:

$$u=-\frac{\displaystyle\sum_{i=0}^{n}\lambda_i\psi_{mi}(\boldsymbol{x})}{\displaystyle\sum_{i=0}^{n}\lambda_i\phi_{mi}(\boldsymbol{x})}.$$

Up to now we have dealt with free final time, if the final time is fixed the we have to take in account the presence of the constant c in the Hamiltonian, but Equation 20.7 and Equation 20.8 still hold and hence also the conclusions.\ In practical problems, we must check if

$$\sum\lambda_i\phi_{ki}(\boldsymbol{x}) = 0$$

for all $k = 1,\ldots,m-1$ and

$$\sum\lambda_i\phi_{mi}(\boldsymbol{x}) \neq 0$$

If any of the relations are violated then this represent a violation of the necessary conditions and so singular controls can not occur, if these conditions are all met, then there may be a singular extremal control.



21. OCP Transformation

Let:

$$\begin{aligned} \boldsymbol{x}(t) &\in \mathbb{R}^n && \text{(state variables)} \\ \boldsymbol{\lambda}(t) &\in \mathbb{R}^n && \text{(costates or adjoint variables)} \\ \boldsymbol{u}(t) &\in \mathscr{U} \subseteq \mathbb{R}^m && \text{(control variables)} \end{aligned}$$

We aim to solve the following **Optimal Control Problem**:

Objective: Minimize the performance index

$$\underbrace{\underbrace{\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b))}_{\text{Mayer term}} + \int_a^b \underbrace{L(\boldsymbol{x},\boldsymbol{u},t)}_{\text{Lagrange term}}\, \mathrm{d}t}_{\text{Bolza form}}$$

Subject to:

1. **System dynamics**:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$$

2. **Boundary conditions**:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0}$$

3. **Control constraints**:

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Hamiltonian

The Hamiltonian function is introduced as follows:

$$\begin{aligned} H(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u},t) &= L(\boldsymbol{x},\boldsymbol{u},t) + \boldsymbol{\lambda}\cdot\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) \\ B(\boldsymbol{x}_a,\boldsymbol{x}_b,\boldsymbol{\omega}) &= \Phi(\boldsymbol{x}_a,\boldsymbol{x}_b) + \boldsymbol{\omega}\cdot\boldsymbol{b}(\boldsymbol{x}_a,\boldsymbol{x}_b) \end{aligned}$$

Where:

$$\boldsymbol{\lambda} \cdot \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) = \sum_{k=1}^{n} \lambda_k f_k(\boldsymbol{x},\boldsymbol{u},t)$$
$$\boldsymbol{\omega} \cdot \boldsymbol{b}(\boldsymbol{x}_a,\boldsymbol{x}_b) = \sum_{k=1}^{p} \omega_k b_k(\boldsymbol{x}_a,\boldsymbol{x}_b)$$

Boundary Value Problem (BVP)

The candidate for the constrained optimal solution must satisfy the following Boundary Value Problem (BVP):

$$\begin{cases} \boldsymbol{x}' = \dfrac{\partial H}{\partial \boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u},t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) & \text{(state equation)} \\ \boldsymbol{\lambda}' = -\dfrac{\partial H}{\partial \boldsymbol{x}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{u},t) & \text{(adjoint equation)} \\ \boldsymbol{0} = \dfrac{\partial B}{\partial \boldsymbol{\omega}}(\boldsymbol{x}_a,\boldsymbol{x}_b,\boldsymbol{\omega}) = \boldsymbol{b}(\boldsymbol{x}_a,\boldsymbol{x}_b) & \text{(boundary condition)} \\ \boldsymbol{0} = \dfrac{\partial B}{\partial \boldsymbol{x}_a}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\omega}) - \boldsymbol{\lambda}(a) & \text{(additional boundary condition)} \\ \boldsymbol{0} = \dfrac{\partial B}{\partial \boldsymbol{x}_b}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\omega}) + \boldsymbol{\lambda}(b) & \text{(additional boundary condition)} \\ \boldsymbol{u}(t) = \underset{\boldsymbol{v}\in\mathscr{U}}{\operatorname{argmin}} H(\boldsymbol{x}(t),\boldsymbol{\lambda}(t),\boldsymbol{v},t) & \text{(control equation)} \end{cases}$$

Remark 21.1 — Remark on Boundary Conditions. From:

$$\boldsymbol{0} = \frac{\partial B}{\partial \boldsymbol{x}_a}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\omega}) - \boldsymbol{\lambda}(a),$$
$$\boldsymbol{0} = \frac{\partial B}{\partial \boldsymbol{x}_b}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\omega}) + \boldsymbol{\lambda}(b),$$

It is often possible to eliminate $\boldsymbol{\omega}$ from the boundary conditions.

21.1 Lagrange to Mayer Transformation

Consider the following Optimal Control Problem (OCP):

Objective

Minimize the integral of the running cost $L(\boldsymbol{x},\boldsymbol{u},t)$:

$$\text{minimize} \quad \int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t$$

Subject to:

1. **System dynamics** (state evolution):

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$$

2. **Boundary conditions**:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0}$$

Introducing an Auxiliary Variable

To transform the problem from Lagrange form to Mayer form, introduce a new auxiliary variable z to represent the accumulation of the running cost $L(\boldsymbol{x},\boldsymbol{u},t)$:

$$z' = L(\boldsymbol{x},\boldsymbol{u},t), \qquad z(a) = 0$$

Reformulating the Objective

The integral of $L(\boldsymbol{x},\boldsymbol{u},t)$ can be rewritten using the new variable z:

$$\int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t = \int_a^b z'\,\mathrm{d}t = z(b) - z(a) = z(b)$$

Thus, the original objective of minimizing the integral can now be expressed as minimizing $z(b)$:

Transformed Objective

$$\text{minimize} \quad z(b)$$

Transformed System Dynamics

The system dynamics now include both the original state $\boldsymbol{x}$ and the new variable z:

$$\begin{cases} \boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) \\ z' = L(\boldsymbol{x},\boldsymbol{u},t) \end{cases}$$

Transformed Boundary Conditions

The boundary conditions now include the original condition on $\boldsymbol{x}$ and the initial condition for z:

$$\begin{cases} \boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0} \\ z(a) = 0 \end{cases}$$

Reformulating the Problem Using New Variables

We can define the augmented state $\boldsymbol{w}$ to combine $\boldsymbol{x}$ and z, and correspondingly define the new system dynamics and boundary conditions:

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ z \end{pmatrix}, \qquad \boldsymbol{F}(\boldsymbol{w}, \boldsymbol{u}, t) = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) \\ L(\boldsymbol{x}, \boldsymbol{u}, t) \end{pmatrix}$$

The new objective becomes:

$$\Phi(\boldsymbol{w}(a), \boldsymbol{w}(b)) = z(b)$$

The boundary conditions are expressed as:

$$\boldsymbol{c}(\boldsymbol{w}(a), \boldsymbol{w}(b)) = \begin{pmatrix} \boldsymbol{b}(\boldsymbol{x}(a), \boldsymbol{x}(b)) \\ z(a) \end{pmatrix}$$

Final Transformed Problem

The transformed OCP now takes the Mayer form:

Objective

$$\text{minimize} \quad \Phi(\boldsymbol{w}(a), \boldsymbol{w}(b))$$

Subject to:

1. **Augmented system dynamics**:

$$\boldsymbol{w}' = \boldsymbol{F}(\boldsymbol{w}, \boldsymbol{u}, t)$$

2. **Augmented boundary conditions**:

$$\boldsymbol{c}(\boldsymbol{w}(a), \boldsymbol{w}(b)) = \boldsymbol{0}$$

21.2 Mayer to Lagrange Transformation

Consider the following Optimal Control Problem (OCP) given in Mayer form.

Objective

Minimize the terminal cost $\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b))$:

$$\text{minimize} \quad \Phi(\boldsymbol{x}(a),\boldsymbol{x}(b))$$

Subject to:

1. **System dynamics** (state evolution):

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$$

2. **Boundary conditions**:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0}$$

Introducing an Auxiliary Function

To transform this problem from Mayer form to Lagrange form (which involves a running cost), we introduce a new auxiliary function $g(t)$ that reconstructs the Mayer terminal cost $\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b))$ as a cumulative function of t. This allows us to spread the cost over the time horizon and thus define a running cost function.

Define $g(t)$ as:

$$g(t) = \Phi(\boldsymbol{x}(a),\boldsymbol{x}(t)) + \frac{t-a}{b-a}\Phi(\boldsymbol{x}(a),\boldsymbol{x}(a))$$

Here, $\boldsymbol{x}(t)$ is the state at time t, and $\Phi(\boldsymbol{x}(a),\boldsymbol{x}(t))$ represents the contribution to the cost as the system evolves. The second term in $g(t)$, $\frac{t-a}{b-a}\Phi(\boldsymbol{x}(a),\boldsymbol{x}(a))$, ensures that the cost is scaled proportionally to the time horizon.

Deriving the running cost

Next, we differentiate $g(t)$ with respect to time t to obtain a running cost.

First, compute $g'(t)$:

$$g'(t) = \frac{\partial\Phi(\boldsymbol{x}(a),\boldsymbol{x}(t))}{\partial\boldsymbol{x}_b}\boldsymbol{x}'(t) + \frac{\Phi(\boldsymbol{x}(a),\boldsymbol{x}(a))}{b-a}$$

Since $\boldsymbol{x}'(t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$ (from the system dynamics), we substitute this into the expression:

$$g'(t) = \frac{\partial\Phi(\boldsymbol{x}(a),\boldsymbol{x}(t))}{\partial\boldsymbol{x}_b}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) + \frac{\Phi(\boldsymbol{x}(a),\boldsymbol{x}(a))}{b-a}$$

Integrating the running cost

By integrating $g'(t)$ over the time interval $[a,b]$, we can reconstruct the original Mayer cost $\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b))$:

$$\int_a^b g'(t)\,\mathrm{d}t = g(b) - g(a)$$

Substituting the values of $g(b)$ and $g(a)$, we get:

$$g(b) - g(a) = \Phi(\boldsymbol{x}(a), \boldsymbol{x}(b)) + \cancel{\Phi(\boldsymbol{x}(a), \boldsymbol{x}(a))} - \cancel{\Phi(\boldsymbol{x}(a), \boldsymbol{x}(a))}$$

Thus, the integral of $g'(t)$ gives us the original Mayer cost:

$$\int_a^b g'(t)\,\mathrm{d}t = \Phi(\boldsymbol{x}(a), \boldsymbol{x}(b))$$

Defining the Lagrange running cost

Now, we can define the running cost $L(\boldsymbol{x}, \boldsymbol{u}, t)$ for the Lagrange formulation of the problem. From the expression for $g'(t)$, we set:

$$L(\boldsymbol{x}, \boldsymbol{u}, t) = \frac{\partial \Phi(\boldsymbol{x}(a), \boldsymbol{x}(t))}{\partial \boldsymbol{x}_b} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) + \frac{\Phi(\boldsymbol{x}(a), \boldsymbol{x}(a))}{b - a}$$

This running cost $L(\boldsymbol{x}, \boldsymbol{u}, t)$ captures both the immediate cost rate associated with the evolution of the state and a constant term related to the total cost scaling over time.

21.3 Final Transformed Problem (Lagrange Form)

The problem is now expressed in Lagrange form:

Objective

Minimize the integral of the running cost $L(\boldsymbol{x}, \boldsymbol{u}, t)$:

$$\text{minimize} \quad \int_a^b L(\boldsymbol{x}, \boldsymbol{u}, t)\,\mathrm{d}t$$

Subject to:

1. **System dynamics**:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t)$$

2. **Boundary conditions**:

$$\boldsymbol{b}(\boldsymbol{x}(a), \boldsymbol{x}(b)) = \boldsymbol{0}$$

By introducing the auxiliary function $g(t)$ and deriving the running cost $L(\boldsymbol{x}, \boldsymbol{u}, t)$, we have successfully transformed the Mayer problem into an equivalent Lagrange problem.

Remark 21.2 In practice Mayer form, Lagrange form or Bolza form of the target are equivalent.

21.4 Constant Parameters to Constant Functions

Consider the following Optimal Control Problem (OCP), where the performance index depends also on constant parameters $\boldsymbol{\mu} \in \mathbb{R}^q$.

Objective

Minimize the performance index, which is a combination of the terminal cost Φ and the integral of the running cost L:

$$\Phi(\boldsymbol{x}(a), \boldsymbol{x}(b), \boldsymbol{\mu}) + \int_a^b L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\mu}, t)\, \mathrm{d}t$$

Subject to:

1. **System dynamics**:
   The system evolves according to the following differential equation, which depends on both the state $\boldsymbol{x}$ and the control $\boldsymbol{u}$, as well as the constant parameters $\boldsymbol{\mu}$:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{\mu}, t)$$

2. **Boundary conditions**:
   The system must satisfy specific boundary conditions at the initial and terminal time points:

$$\boldsymbol{b}(\boldsymbol{x}(a), \boldsymbol{x}(b), \boldsymbol{\mu}) = \mathbf{0}$$

3. **Control constraints**:
   The control $\boldsymbol{u}(t)$ must remain within a predefined admissible set $\mathscr{U}$:

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Here, $\boldsymbol{\mu} \in \mathbb{R}^q$ are constant parameters that remain fixed over time. However, we can transform the problem by introducing constant functions that represent these parameters.

Introducing Constant Functions

To rewrite the problem in a form suitable for certain optimization techniques, we introduce constant functions for each component of the parameter vector $\boldsymbol{\mu}$:

$$\mu_k(t) \equiv \mu_k, \qquad k = 1, 2, \ldots, q$$

Each $\mu_k(t)$ is a constant function that satisfies the differential equation:

$$\mu_k'(t) = 0 \quad \text{for all} \quad t \in [a, b]$$

In vector form, this is written as:

$$\boldsymbol{\mu}'(t) = \mathbf{0}$$

Reformulated Problem

By introducing these constant functions, we can now treat the parameters $\boldsymbol{\mu}$ as state variables that evolve according to the trivial dynamic equation $\boldsymbol{\mu}'(t) = \mathbf{0}$. This allows us to apply standard optimal control techniques to problems involving constant parameters. The reformulated problem is as follows:

Objective

Minimize the modified performance index:

$$\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\mu}(a)) + \int_a^b L(\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{\mu}(t),t)\,\mathrm{d}t$$

Subject to:

1. **Extended system dynamics**:

$$\begin{cases} \boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{\mu},t) \\ \boldsymbol{\mu}' = \boldsymbol{0} \end{cases}$$

2. **Boundary conditions**:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{\mu}(a)) = \boldsymbol{0}$$

3. **Control constraints**:

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Remark 21.3 By introducing the constant functions $\mu_k(t)$, we have transformed the original OCP with fixed parameters $\boldsymbol{\mu}$ into an equivalent OCP where these parameters are treated as constant functions governed by the trivial dynamic equation $\boldsymbol{\mu}'(t) = \boldsymbol{0}$. This reformulation allows the use of standard optimal control methods to solve problems with constant parameters.

21.5 Integral Constraints

Consider the following Optimal Control Problem (OCP), where the performance index is influenced by certain integral constraints.

Objective

Minimize the performance index, which consists of a terminal cost and the integral of a running cost:

$$\text{Minimize} \quad \Phi(\boldsymbol{x}(a),\boldsymbol{x}(b)) + \int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t$$

Subject to:

1. **System Dynamics**:
   The evolution of the state variables is governed by the following differential equation:

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$$

2. **Integral Constraints**:
   The problem must satisfy specific integral constraints over the time interval ([a, b]):

$$\int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t = \boldsymbol{0}$$

3. **Boundary Conditions**:
   The system must adhere to specified boundary conditions at both the initial and terminal time points:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0}$$

4. **Control Constraints**:
   The control input $\boldsymbol{u}(t)$ is required to remain within a predefined admissible set $\mathscr{U}$:

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Introducing Auxiliary Functions

To effectively incorporate the integral constraints into our OCP, we introduce auxiliary functions $\boldsymbol{z}(t)$ that capture the cumulative behavior of the integral constraints over time. We define $\boldsymbol{z}(t)$ as follows:

$$\boldsymbol{z}(t) = \int_a^t \boldsymbol{g}(\boldsymbol{x}(s),\boldsymbol{u}(s),s)\,\mathrm{d}s$$

This function represents the accumulated effect of the constraint $\boldsymbol{g}$ from the starting time (a) up to any time (t). The derivative of $\boldsymbol{z}(t)$ is given by:

$$\boldsymbol{z}'(t) = \boldsymbol{g}(\boldsymbol{x}(t),\boldsymbol{u}(t),t)$$

Reformulated Problem

By introducing the auxiliary function $\boldsymbol{z}(t)$, we can reformulate the OCP as follows:

Objective

The objective remains to minimize the modified performance index:

$$\text{Minimize} \quad \Phi(\boldsymbol{x}(a),\boldsymbol{x}(b)) + \int_a^b L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t$$

Subject to:

1. **Extended System Dynamics**:
   The dynamics of the state variables and the auxiliary function can now be described as:

$$\begin{cases} \boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) \\ \boldsymbol{z}' = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t) \end{cases}$$

2. **Integral Constraints**:
   The integral constraints are now effectively incorporated into the system through the auxiliary function $\boldsymbol{z}(t)$, allowing them to be handled dynamically.
3. **Boundary Conditions**:
   The boundary conditions remain unchanged:

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b)) = \boldsymbol{0}$$

4. **Control Constraints**:
   The control constraints continue to apply:

$$\boldsymbol{u}(t) \in \mathscr{U}$$

In summary, the introduction of the auxiliary function $\boldsymbol{z}(t)$ enables the reformulation of the original OCP, allowing for the integral constraints to be treated as part of the system dynamics.

21.6 Free-Time

In some optimal control problems, the time horizon T is not fixed but instead is an unknown variable that must be determined as part of the solution. This setup leads to what is known as the **free-time problem**, which can be formalized as:

$$\begin{aligned} \text{Minimize:} \quad & \phi\big(\boldsymbol{x}(0),\boldsymbol{x}(T)\big) + \int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t, \\ \text{Subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}(x,u,t), \\ & \boldsymbol{b}\big(\boldsymbol{x}(0),\boldsymbol{x}(T)\big) = 0, \end{aligned}$$

where:

- $\boldsymbol{x}(t)$ is the state variable,
- $\boldsymbol{u}(t)$ is the control input,
- T is the free (unknown) time horizon,
- $\phi(\boldsymbol{x}(0),\boldsymbol{x}(T))$ represents the Mayer cost,
- $L(\boldsymbol{x},\boldsymbol{u},t)$ is the (Lagrange) running cost,
- $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$ is the system dynamics,
- $\boldsymbol{b}(\boldsymbol{x}(0),\boldsymbol{x}(T)) = \boldsymbol{0}$ represents boundary conditions.

Normalizing Time

To handle the unknown time horizon T, a common approach is to normalize the time interval to a fixed domain, say $[0,1]$. This is achieved by introducing a change of variables:

$$t = sT, \quad s \in [0,1],$$

where s is a normalized time variable. Under this transformation:

1. The state variable becomes $\tilde{\boldsymbol{x}}(s) = \boldsymbol{x}(sT)$, so its derivative transforms as:

$$\tilde{x}'(s) = \frac{\mathrm{d}\tilde{\boldsymbol{x}}}{\mathrm{d}s} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}s} = \boldsymbol{x}'(sT)T.$$

2. The control variable becomes $\tilde{\boldsymbol{u}}(s) = \boldsymbol{u}(sT)$.

Substituting into the system dynamics, $\boldsymbol{x}'(t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$ transforms to:

$$\tilde{\boldsymbol{x}}'(s) = T\boldsymbol{f}\big(\tilde{\boldsymbol{x}}(s),\tilde{\boldsymbol{u}}(s),sT\big).$$

Similarly, the running cost $\int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t$ becomes:

$$\int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t = \int_0^1 L\big(\tilde{\boldsymbol{x}}(s),\tilde{\boldsymbol{u}}(s),sT\big)T\,\mathrm{d}s.$$

The Mayer cost $\phi(\boldsymbol{x}(0),\boldsymbol{x}(T))$ remains unchanged under the transformation.

Reformulated Problem

Using the normalized variables and the change of variable $t = sT$, the free-time problem can be rewritten as:

$$\begin{aligned} \text{Minimize:} \quad & \phi\big(\tilde{\boldsymbol{x}}(0),\tilde{\boldsymbol{x}}(1)\big) + \int_0^1 L\big(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{u}},sT\big)T\,\mathrm{d}s, \\ \text{Subject to:} \quad & \tilde{\boldsymbol{x}}'(s) = T\boldsymbol{f}\big(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{u}},sT\big), \\ & b\big(\tilde{\boldsymbol{x}}(0),\tilde{\boldsymbol{x}}(1)\big) = 0. \end{aligned}$$

Here, T is now treated as an additional optimization variable alongside the state $\tilde{\boldsymbol{x}}(s)$ and the control $\tilde{\boldsymbol{u}}(s)$. The problem is now defined over the fixed interval $[0,1]$, which simplifies the numerical treatment.

21.7 Free-Time 2

Let:

$$\begin{aligned} \text{Minimize:} \quad & \phi\big(\boldsymbol{x}(0),\boldsymbol{x}(T)\big) + \int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t, \\ \text{Subject to:} \quad & \boldsymbol{x}' = \boldsymbol{f}(x,u,t), \\ & \boldsymbol{b}\big(\boldsymbol{x}(0),\boldsymbol{x}(T)\big) = 0, \end{aligned}$$

Normalizing Time

Introduce a change of variables:

$$t = sT(s), \quad s \in [0,1], \qquad T'(s) = 0$$

where s is a normalized time variable. Under this transformation:

1. The state variable becomes $\tilde{\boldsymbol{x}}(s) = \boldsymbol{x}(sT(s))$, so its derivative transforms as:

$$\begin{aligned} \tilde{x}'(s) &= \frac{\mathrm{d}\tilde{\boldsymbol{x}}}{\mathrm{d}s} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}s} \\ &= \boldsymbol{x}'(sT(s))(sT(s))' \\ &= \boldsymbol{x}'(sT(s))(T(s)+sT'(s)) \\ &= \boldsymbol{x}'(sT(s))T(s) \end{aligned}$$

2. The control variable becomes $\tilde{\boldsymbol{u}}(s) = \boldsymbol{u}(sT(s))$.
   Substituting into the system dynamics, $\boldsymbol{x}'(t) = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t)$ transforms to:

$$\tilde{\boldsymbol{x}}'(s) = T(s)\boldsymbol{f}\big(\tilde{\boldsymbol{x}}(s),\tilde{\boldsymbol{u}}(s),sT(s)\big).$$

Similarly, the running cost $\int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t$ becomes:

$$\int_0^T L(\boldsymbol{x},\boldsymbol{u},t)\,\mathrm{d}t = \int_0^1 L\big(\tilde{\boldsymbol{x}}(s),\tilde{\boldsymbol{u}}(s),sT(s)\big)T(s)\,\mathrm{d}s.$$

The Mayer cost $\phi(\boldsymbol{x}(0),\boldsymbol{x}(T))$ remains unchanged under the transformation.

Reformulated Problem

Using the normalized variables and the change of variable $t = sT(s)$, the free-time problem can be rewritten as:

$$\begin{aligned} \text{Minimize:} \quad & \phi\big(\tilde{\boldsymbol{x}}(0),\tilde{\boldsymbol{x}}(1)\big) + \int_0^1 L\big(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{u}},sT(s)\big)T(s)\,\mathrm{d}s, \\ \text{Subject to:} \quad & \tilde{\boldsymbol{x}}'(s) = T(s)\boldsymbol{f}\big(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{u}},sT(s)\big), \\ & b\big(\tilde{\boldsymbol{x}}(0),\tilde{\boldsymbol{x}}(1)\big) = 0. \end{aligned}$$



22. Examples of optimal control problems

22.1 The Brachistochrone Problem

One of the earliest and most renowned challenges in the calculus of variations is the brachistochrone problem, originally posed by Johann Bernoulli in 1696. This problem can be approached from various perspectives, particularly through the lenses of calculus of variations and optimal control theory, which involve ordinary and differential algebraic equations (ODEs and DAEs), as well as the principles of energy conservation.

22.1.1 Problem Statement

The goal of the brachistochrone problem is to determine the path of minimum time that connects two points under the influence of gravity alone. Specifically, we consider motion in two dimensions, transitioning from a starting point A to a fixed endpoint B. The challenge is to find the shape of the curve along which a particle, subject solely to gravitational force, travels from A to B in the least amount of time.

To facilitate analysis, we can establish a coordinate system with the origin at point A. In this system, the vertical axis y points upward (opposite to the direction of gravitational force), while the horizontal axis x is perpendicular to y, ensuring that the x-coordinate of point B is positive. Given that the optimal path γ must be a continuous curve without loops, we can represent γ as a function $y(x)$.

For simplicity in calculations, we assume that $y > 0$, which eliminates complications arising from the negative direction of gravity and allows for straightforward interpretation of heights. By applying the principle of conservation of energy, we can relate the height y, gravitational acceleration g, and particle velocity v:

$$mgy = \frac{1}{2}mv^2 \quad \Longrightarrow \quad v(y(x)) = \sqrt{2gy(x)}.$$

This relationship indicates that the solution is independent of the mass m.

22.1.2 Path Length and Travel Time

The distance L traveled by the particle as it approaches endpoint B can be expressed using a line integral:

$$L(y,x) = \oint_{\gamma} \mathrm{d}s = \int_0^{x_B} \sqrt{1 + (y'(x))^2}\,\mathrm{d}x.$$

The time taken to traverse this path is given by:

$$T(y,x) = \int_0^{x_B} \frac{\sqrt{1+(y'(x))^2}}{\sqrt{2gy(x)}}\,\mathrm{d}x. \tag{22.1}$$

22.1.3 Constraints

The solution to this problem involves minimizing the functional defined in equation Equation 22.1 under the following constraints:

$$y(0) = 0, \qquad y(x_B) = -y_B. \tag{22.2}$$

22.1.4 Solution Using Calculus of Variations

To solve the brachistochrone problem using calculus of variations, we employ the Euler-Lagrange equation. Since the functional

$$T(y,x) = \int F(y,x)\,\mathrm{d}x$$

does not explicitly depend on x, the Euler-Lagrange equation simplifies to Beltrami's identity:

$$\frac{\partial F}{\partial y} - \frac{\mathrm{d}}{\mathrm{d}x}\frac{\partial F}{\partial y'} = 0 \quad \Longrightarrow \quad F - y'\frac{\partial F}{\partial y'} = c, \tag{22.3}$$

where c is a constant determined by the constraints given in equation Equation 22.2.
For the brachistochrone problem, we can express F as:

$$\sqrt{2g}T = \int_0^{x_B} \sqrt{\frac{1+y'(x)^2}{y(x)}}\,\mathrm{d}x \quad \Longrightarrow \quad F(y,x) = \sqrt{\frac{1+y'(x)^2}{y(x)}}.$$

Thus, the Euler-Lagrange equation Equation 22.3 reduces to:

$$\sqrt{\frac{1+y'^2}{y}} - \frac{y'^2}{\sqrt{y(1+y'^2)}} = c \quad \Longrightarrow \quad y(1+y'^2) = \frac{1}{c^2} = C > 0.$$

This results in an autonomous nonlinear differential equation that can be solved using separation of variables:

$$\int_{y_A}^{y_B} \sqrt{\frac{y}{C-y}}\,\mathrm{d}y = \int_{x_A}^{x_B} 1\,\mathrm{d}x.$$

To compute this integral, we use the substitution $y = C\left(\sin\frac{\phi}{2}\right)^2$, which gives us the differential

$$\mathrm{d}y = \sin\frac{\phi}{2}\cos\frac{\phi}{2}\,\mathrm{d}\phi$$

Thus, we have:

$$\begin{aligned}\int \sqrt{\frac{y}{C-y}}\,\mathrm{d}y &= C\int \sqrt{\frac{\left(\sin\frac{\phi}{2}\right)^2}{1-\left(\sin\frac{\phi}{2}\right)^2}}\sin\frac{\phi}{2}\cos\frac{\phi}{2}\,\mathrm{d}\phi \\ &= C\int \left(\sin\frac{\phi}{2}\right)^2 \mathrm{d}\phi \\ &= \frac{C}{2}(\phi - \sin\phi) + k,\end{aligned}$$

where k is the integration constant.

Evaluating this integral at the extrema defined by the constraints in equation Equation 22.2 allows us to determine the constants C and k. Since the starting point A coincides with the origin, we have $k = 0$. The constant C depends on the coordinates of point $B = (x_B, y_B)$. Specifically, we find:

$$\begin{aligned} x_B &= \frac{C}{2}(\phi_B - \sin\phi_B), \\ y_B &= -\frac{C}{2}(1 - \cos\phi_B). \end{aligned} \tag{22.4}$$

In conclusion, the parametric equations for the brachistochrone are given by:

$$\begin{aligned} x(\phi) &= \frac{C}{2}(\phi - \sin\phi), \\ y(\phi) &= -\frac{C}{2}(1 - \cos\phi). \end{aligned} \tag{22.5}$$

These equations describe a cycloid curve.

The geometric proof provided by Bernoulli contains spurious solutions, as noted in [11]. While the variational approach presents challenges in proving the existence of the brachistochrone, applying optimal control theory simplifies this task. The existence of a solution can be directly established using the Ascoli-Arzelà theorem.

22.1.5 The brachistochrone as an optimal control problem

This example can be treated as an optimal control problem, the control is the angle ϑ of descent of the particle. Splitting the velocity v in its two components along the xy axes, the problem can be states as

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= v\sin\vartheta, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= -v\cos\vartheta, \\ \frac{\mathrm{d}v}{\mathrm{d}t} &= g\cos\vartheta \end{aligned} \tag{22.6}$$

with border conditions, if T is the time used to travel from A to B,

$$\begin{aligned} x(0) = 0, \quad y(0) = 0, \quad v(0) = 0, \\ x(T) = x_B, \quad y(T) = y_B, \quad v(T) = v_T, \end{aligned} \tag{22.7}$$

for a free final velocity v_T. The functional to be minimized is T.

22.1.6 A Numerical Example

Consider the brachistochrone problem for a particle released from the origin and traveling to the point $B = (10, -3)$. In the variational formulation, we need to determine the constant C. This can be achieved using Equation 22.4. Solving the nonlinear system for C and ϕ_B yields:

$$\phi_B \approx 4.17, \qquad C \approx 3.97.$$

The solution is illustrated in Figure Figure 22.1.

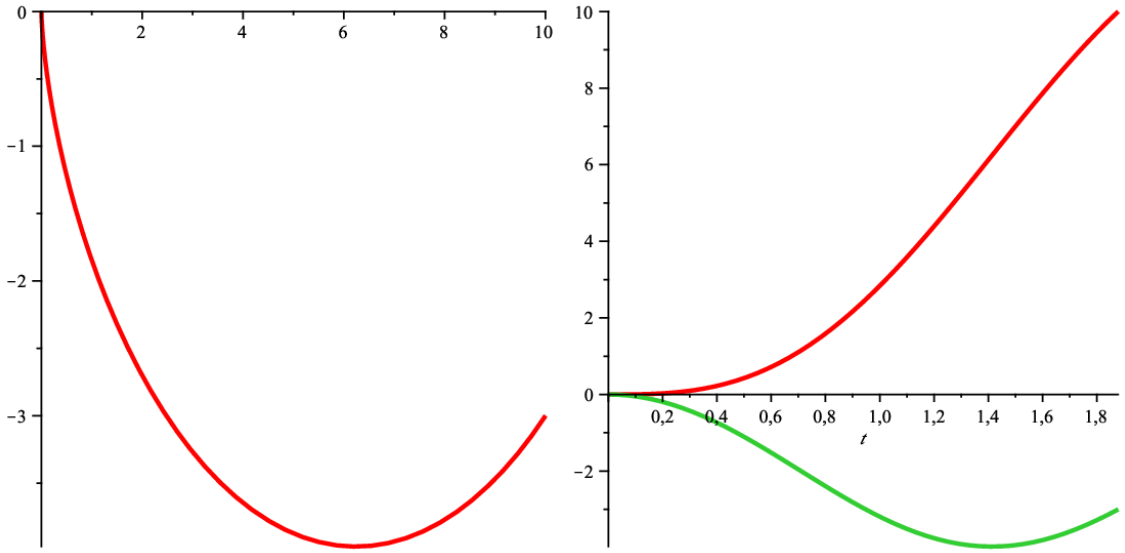


Figure 22.1: **Variational solution to the brachistochrone problem:** The left plot shows the trajectory of the particle from the origin to point B, while the right plot displays the two components of the solution: x in red and y in green.

When treated as an optimal control problem, as stated in equation Equation 22.6 with boundary conditions Equation 22.7, finding an analytical solution becomes challenging. For instance, the Hamiltonian function is given by:

$$\mathscr{H}(x,y,v,\lambda,\mu,\xi,\theta) = 1 + \lambda v \sin\vartheta - \mu v \cos\vartheta + \xi g \cos\vartheta.$$

According to Pontryagin's Maximum Principle (PMP), the associated boundary value problem is:

$$\begin{aligned} \lambda' &= -\frac{\partial \mathscr{H}}{\partial x} = 0, \\ \mu' &= -\frac{\partial \mathscr{H}}{\partial y} = 0, \\ \xi' &= -\frac{\partial \mathscr{H}}{\partial v} = \lambda \sin\vartheta - \mu \cos\vartheta. \end{aligned} \tag{22.8}$$

Additionally, we have:

$$0 = \frac{\partial \mathscr{H}}{\partial \vartheta} = \lambda v \cos\vartheta + \mu v \sin\vartheta - \xi g \sin\vartheta.$$

To avoid solving this analytically, we can use the variational formulation's results to inform our optimal control problem (OCP) solution. However, since equations Equation 22.5 are functions of the spatial angle ϕ rather than time t, we need to establish a relationship between them using the time functional:

$$T(x,y) = \int_0^{\phi_B} \frac{\sqrt{x'(\phi)^2 + y'(\phi)^2}}{\sqrt{2gy(\phi)}}\,\mathrm{d}\phi. \tag{22.9}$$

A brief calculation shows that:

$$T(x,y) = \int_0^{\phi_B} \sqrt{\frac{C}{2g}}\,\mathrm{d}\phi = \phi_B \sqrt{\frac{C}{2g}} = k\phi_B \approx 1.87,$$

demonstrating the tautochrone (isochrone) property of the cycloid. Thus, we can parameterize time as $t = k\phi$, for $t \in [0, k\phi_B]$.

From this perspective, the control $\vartheta(t)$ must satisfy:

$$\begin{aligned}\vartheta(t) &= \arctan\left(\frac{y'(t)}{x'(t)}\right) + \frac{\pi}{2} \\ &= \arctan\left(-\frac{\sin(t/k)}{1-\cos(t/k)}\right) + \frac{\pi}{2}.\end{aligned}$$

This leads to:

$$\begin{aligned}\sin\vartheta &= \frac{1-\cos(t/k)}{\sqrt{2}\sqrt{1-\cos(t/k)}}, \\ \cos\vartheta &= \frac{\sin(t/k)}{\sqrt{2}\sqrt{1-\cos(t/k)}}.\end{aligned}$$

Consequently, the velocity becomes:

$$v(t) = \frac{\sqrt{2}C}{2k}\sqrt{1-\cos(t/k)}.$$

The plots of velocity $v(t)$ and control $\vartheta(t)$ are shown in Figure Figure 22.2.

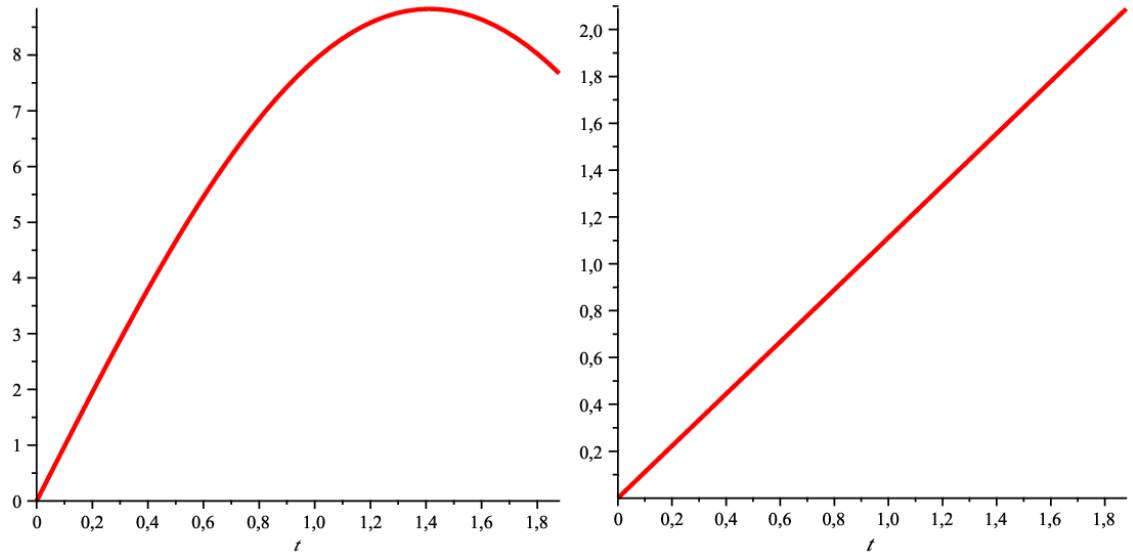


Figure 22.2: On the left is the plot of velocity $v(t)$, and on the right is the control variable $\vartheta(t)$.

The differential equations for the spatial components become:

$$\begin{aligned}x'(t) &= v(t)\sin\vartheta = \frac{C}{2}(1-\cos(t/k)), \\ y'(t) &= -v(t)\cos\vartheta = -\frac{C}{2}(\sin(t/k)).\end{aligned}$$

To compute the multipliers λ, μ, and ξ, we can use the Hamiltonian function. By substituting for ξ, we are left with a single function in one unknown, λ. This gives us $\lambda = -k/C$. Thus, we find that the three multipliers are:

$$\lambda = -\frac{k}{C} \approx -0.11,$$
$$\mu = -\frac{\lambda \sin(T/k)}{1-\cos(T/k)} \approx -0.06,$$
$$\xi(= \frac{\sqrt{2}C}{2gk} \frac{\lambda \sin(t/k) + \mu(1-\cos(t/k))}{\sqrt{1-\cos(t/k)}}.$$

The plot of these costate variables is shown in Figure 22.3.

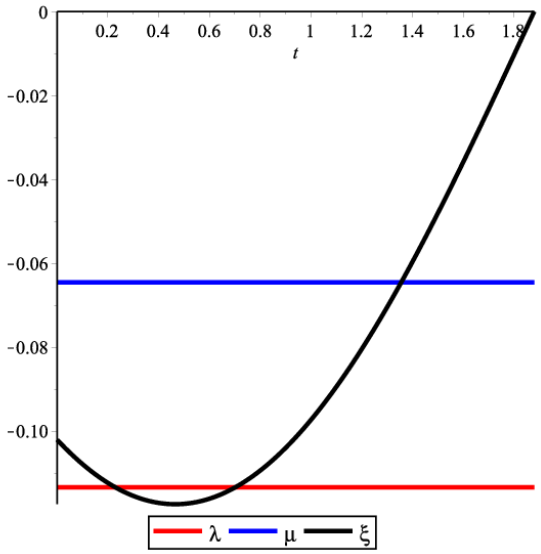


Figure 22.3: The three multipliers are represented as follows: λ in red, μ in blue, and ξ in black.

Finally, using the ACADO toolkit to solve this same problem yields a solution depicted in Figure Figure 22.4.

22.2 The DIDO Problem

Consider a 2D curve defined parametrically, with the following conditions:

- The curve begins and ends on the x-axis.
- The total length of the curve is fixed.

The objective is to **maximize the area** enclosed between the curve and the x-axis, while maintaining the constraint that the curve's length remains constant.

Formally, we seek the shape of the curve that maximizes the enclosed area, given that:

1. The curve starts and ends on the x-axis.
2. The total length of the curve is fixed.

This problem is often referred to as the *DIDO problem*, and it is a well-known problem in **variational calculus** and **optimal control theory**.

To compute the area inside a curve we require the Divergence Theorem:

22.2.1 The Divergence Theorem

The **Divergence Theorem** is a fundamental result in vector calculus, connecting the flow of a vector field through a closed surface to the behavior of the field within the volume enclosed by the surface. It provides a powerful tool for converting complex surface integrals into simpler volume integrals, and vice versa.

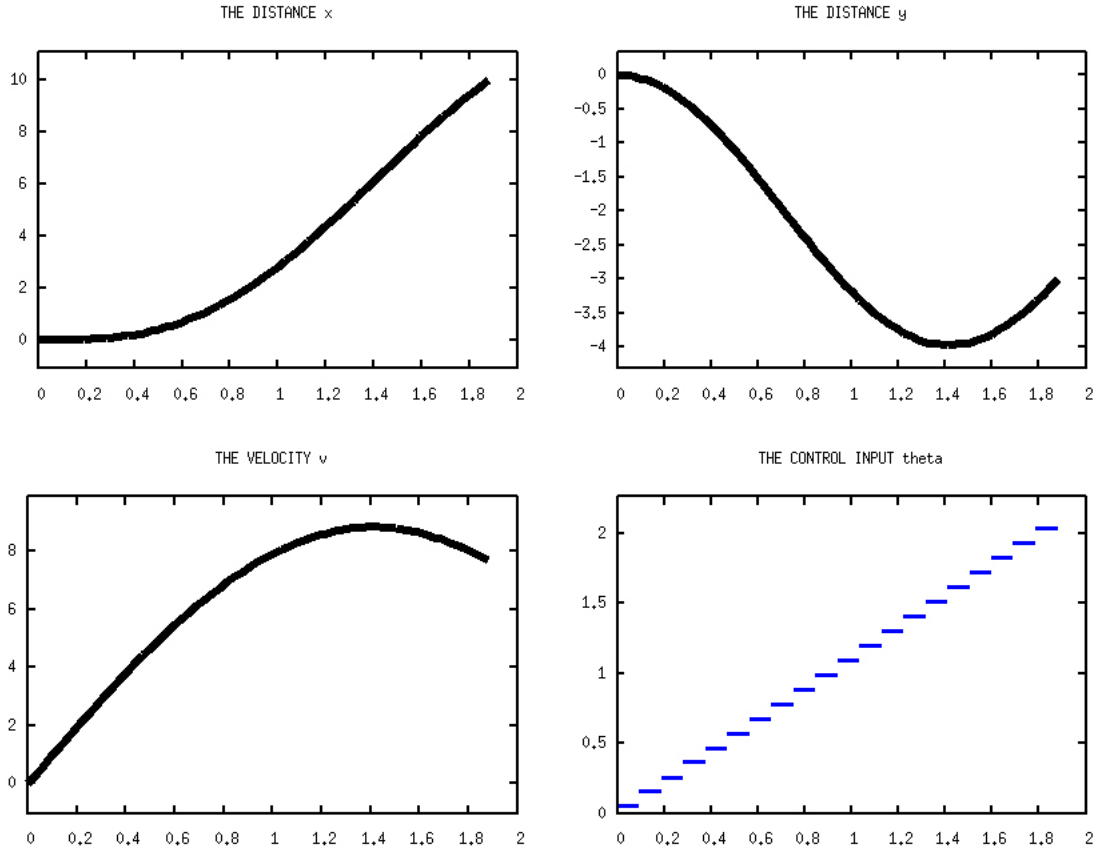


Figure 22.4: **Solution provided by ACADO.**

Statement of the Divergence Theorem (2D version)

Let S be a solid region in 2-dimensional space with a smooth (or piecewise smooth), closed boundary curve $C = \partial S$, oriented outward. If $\boldsymbol{F}$ is a continuously differentiable vector field defined on S, the Divergence Theorem states:

$$\int_C \boldsymbol{F} \cdot \boldsymbol{n} \, \mathrm{d}s = \iint_S (\nabla \cdot \boldsymbol{F}) \, \mathrm{d}S,$$

where:

- $\boldsymbol{F}$: The vector field (e.g., $\boldsymbol{F}(x,y)$).
- $\boldsymbol{n}$: The unit normal vector to the curve C, oriented outward.
- $\nabla \cdot \boldsymbol{F}$: The divergence of $\boldsymbol{F}$, defined as:

$$\nabla \cdot \boldsymbol{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y}.$$

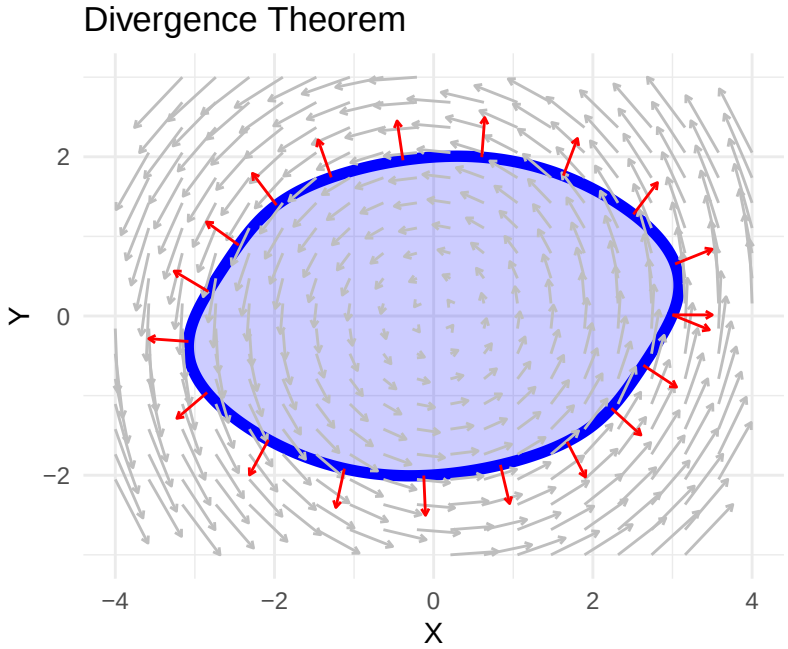


Figure 22.5: 2D schema of divergence theorem. Vector field, curve 2D border and normals.

Application of the Divergence Theorem for area computation

Consider a simple two-dimensional vector field:

$$\boldsymbol{F}(x,y) = \begin{pmatrix} \alpha x \\ (1-\alpha) y \end{pmatrix}$$

In this case, the vector field is a linear combination of x and y coordinates, where α is a constant that determines the relative contribution of the x-component and the y-component of the vector field.

22.2.1.1 Divergence of the Vector Field

The divergence of a vector field is the sum of the partial derivatives of its components. For the vector field $\boldsymbol{F}(x,y)$, we compute:

$$\nabla \cdot \boldsymbol{F}(x,y) = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y}$$

Substituting the components of $\boldsymbol{F}(x,y)$:

$$\nabla \cdot \boldsymbol{F}(x,y) = \frac{\partial(\alpha x)}{\partial x} + \frac{\partial((1-\alpha)y)}{\partial y}$$

Taking the derivatives:

$$\nabla \cdot \boldsymbol{F}(x,y) = \alpha + (1-\alpha) = 1$$

Thus, the divergence of $\boldsymbol{F}(x,y)$ is 1 everywhere in the plane.

Surface Integral and the Divergence Theorem

Now, we apply the **Divergence Theorem** to compute the flux of the vector field through a closed curve (C). The curve is parametrized by:

$$C : (x(t), y(t)), \quad t \in [0, L]$$

The area enclosed by this curve can be computed as a surface integral of the divergence of the vector field over the region enclosed by C. Using the Divergence Theorem, the flux through the closed curve is:

$$|S| = \iint_S 1 \, \mathrm{d}S$$

This is the same as integrating the divergence $\nabla \cdot \boldsymbol{F} = 1$ over the surface, which reduces to the boundary integral:

$$|S| = \int_0^L \boldsymbol{F}(x(t), y(t)) \cdot \frac{\begin{pmatrix} y'(t) \\ -x'(t) \end{pmatrix}}{\sqrt{x'(t)^2 + y'(t)^2}} \sqrt{x'(t)^2 + y'(t)^2} \, \mathrm{d}t$$

In this expression:

- The first term is the **vector field** $\boldsymbol{F}(x(t), y(t))$.
- The second term is the **normal to the curve**, which is given by the vector $\frac{\begin{pmatrix} y'(t) \\ -x'(t) \end{pmatrix}}{\sqrt{x'(t)^2 + y'(t)^2}}$ (the unit normal).
- The third term is the **length element** of the curve, $\sqrt{x'(t)^2 + y'(t)^2} \, \mathrm{d}t$.

Expanding the integral:

$$|S| = \int_0^L \left(\alpha x(t) y'(t) - (1-\alpha) y(t) x'(t) \right) \mathrm{d}t$$

Simplifying for a Specific Curve

Suppose we have a curve that passes through the x-axis (i.e., $y(t) = 0$ for some segment of the curve). On this segment, $y'(t) = 0$. In this case, the expression simplifies, and if we choose $\alpha = 1$, the integral becomes:

$$|S| = \int_0^L x(t) y'(t) \, \mathrm{d}t$$

This formula computes the area enclosed by the curve, with the integral focusing on the (x)-coordinate's contribution when the curve intersects the (x)-axis.

Final Expression for the Area

Finally, if the curve is closed and the initial and final points on the curve satisfy $y(0) = y(L) = 0$, the total enclosed area $|S|$ is given by:

$$|S| = \pm \int_0^L x(t)y'(t)\,\mathrm{d}t$$

The sign of the area depends on the orientation of the curve:

- **Positive** if the curve is counterclockwise (the normal points outward).
- **Negative** if the curve is clockwise (the normal points inward).

This sign convention reflects the direction of the normal vector and is crucial for consistent computation of area in physics and engineering applications.

Formulation of Dido problem

Now we are ready to formulate the Dido problem

Minimize

$$s\int_0^L x\sin\theta\,\mathrm{d}t, \quad s = \operatorname{sign}(b-a)$$

Under the constraints

$$x' = \cos\theta$$
$$y' = \sin\theta$$

And boundary conditions

$$x(0) = b, \quad y(0) = 0, \quad x(L) = a, \quad y(L) = 0$$

Solution of Dido problem

We start by defining two key functions:

$$H(x,y,\lambda_1,\lambda_2,\theta) = \lambda_1\cos\theta + (sx+\lambda_2)\sin\theta,$$
$$B(x_0,y_0,x_1,y_1,\omega_1,\omega_2,\omega_3,\omega_4) = \omega_1(x_0-b) + \omega_2(y_0) + \omega_3(x_1-a) + \omega_4(y_1),$$

where H represents the Hamiltonian function and B encodes boundary conditions. These functions are used to compute the adjoint equations and boundary conditions for the control problem.

Deriving the Adjoint Equations

From the Hamiltonian H, the adjoint equations are given by:

$$\lambda_1' + H_x = \lambda_1' + s\sin\theta = 0,$$
$$\lambda_2' + H_y = \lambda_2' = 0,$$

which simplify to:

1. $\lambda_2(t) = \text{constant} = -sc_y$,
2. $\lambda_1' = -s\sin\theta = -sy'$,

yielding:

$$\lambda_1(t) = s(c_y - y(t)).$$

Boundary Conditions

The adjoint boundary conditions are derived from the B-function:

$$\begin{aligned} B_{x_0} + \lambda_1(0) &= \omega_1 + \lambda_1(0) = 0, \\ B_{y_0} + \lambda_2(0) &= \omega_2 + \lambda_2(0) = 0, \\ B_{x_1} - \lambda_1(L) &= \omega_3 - \lambda_1(L) = 0, \\ B_{y_1} - \lambda_2(L) &= \omega_4 - \lambda_2(L) = 0. \end{aligned}$$

Since these equations can all be satisfied by appropriately choosing the ω parameters, we effectively eliminate the boundary conditions.

Control Equation

The control equation is derived from the partial derivative of H with respect to θ:

$$\begin{aligned} H_\theta &= -\lambda_1 \sin\theta + (sx + \lambda_2)\cos\theta, \\ &= -s(c_y - y)\sin\theta + s(x - c_x)\cos\theta = 0. \end{aligned}$$

This eliminates $\lambda_1(t)$ and $\lambda_2(t)$ entirely. The reduced system is a Boundary Value Problem (BVP) that follows directly.

Reduced BVP

The resulting BVP is:

$$\begin{cases} x' = \cos\theta, \\ y' = \sin\theta, \\ 0 = (x - c_x)\cos\theta + (y - c_y)\sin\theta, \\ x(0) = b, \quad y(0) = 0, \quad x(L) = a, \quad y(L) = 0. \end{cases}$$

Geometric Interpretation

Define $X = x - c_x$ and $Y = y - c_y$. The third equation becomes:

$$\begin{aligned} 0 &= (x - c_x)x' + (y - c_y)y', \\ &= XX' + YY', \\ &= \frac{1}{2}\frac{d}{dt}\left(X^2 + Y^2\right). \end{aligned}$$

Integrating this relation shows that:

$$R^2 = X^2 + Y^2 = (x - c_x)^2 + (y - c_y)^2,$$

which is the equation of a circle with radius R, centered at (c_x, c_y).

The trajectory described by $(x(t), y(t))$ is constrained to lie on a circle of radius R centered at (c_x, c_y). The constansts are otained by matching the boundary conditions.

22.3 An Optimal Control Problem with constraints

The following problem involves a linear control which is bounded, and a constraint on both state and control. The formulation is

$$\begin{aligned} \text{minimize:} \quad & \int_0^3 (t-4)u\,\mathrm{d}t, \\ \text{subject to the dynamic:} \quad & x' = u, \\ \text{constrained by:} \quad & x(0) = 0, \quad x(3) = 3, \quad 0 \le u \le 2, \\ \text{and:} \quad & g(t,x,u) = x - t - u + 1 \le 0. \end{aligned}$$

The Hamiltonian for the problem is given by:

$$\begin{aligned} \mathscr{H} &= (t-4)u + \lambda u + \mu(x - t - u + 1) \\ &= (t - 4 + \lambda - \mu)u + \mu(x - t + 1). \end{aligned}$$

At $t = 0$, we find that the constraint g is satisfied only if $g(0) = 1 - u(0) \le 0$. This implies that the initial control must be $u(0) \ge 1$.

From the Hamiltonian, we derive the equations necessary to form the boundary value problem:

$$\begin{aligned} \frac{\partial \mathscr{H}}{\partial u} &= t - 4 + \lambda - \mu = 0, \\ \lambda' &= -\frac{\partial \mathscr{H}}{\partial x} = -\mu, \quad \mu \ge 0. \end{aligned}$$

The multiplier of the constraint, μ, should be non-negative when the bound is active (i.e., when $g = 0$) and should equal zero when the bound is inactive.

Suppose the constraint is inactive, which implies $\mu = 0$. In this case, the boundary value problem yields $\lambda = c$, where c is a constant derived from the equation $\frac{\partial \mathscr{H}}{\partial x}$. However, from the equation $\frac{\partial \mathscr{H}}{\partial u}$, we get $\lambda = 4 - t$.

This leads to a contradiction: there cannot be an arc where the constraint is inactive. Therefore, on the interval $[0,3]$, the constraint must be sharp. By combining the two equations $\lambda' = -\mu$ and $t - 4 + \lambda - \mu = 0$, we obtain the differential equation:

$$t - 4 + \lambda + \lambda' = 0.$$

Its solution is $\lambda(t) = \alpha \mathrm{e}^{-t} - t + 5$ for an unknown constant α. From the multiplier we can get $\mu = -\lambda' = \alpha \mathrm{e}^{-t} + 1$ which is non negative on $[0,3]$ for $\alpha \ge \mathrm{e}^3 \approx 20.08553692$.

The control is given differentiating $g = 0$ with the combination of $x' = u$, that is $x' - 1 - u' = u - u' - 1 = 0$. The initial value of the control $u(0) = 1$ allows to solve the differential equation yielding $u(t) = 1$. Now from $g = 0$ we obtain $x(t) = t$.

The presence of a minimum is given by the application of the Weierstrass condition $\mathscr{H}(u) - \mathscr{H}(u^\star) > 0$, where $u^\star$ is the optimal control and $u > 1$ is another admissible control.

Observing that $\mu g = 0$ the Weierstrass condition must be checked only for $(t - 4 + \lambda)(u - u^\star) > 0$. Notice that $(t - 4 + \lambda)$ is strictly positive for $\alpha \ge \mathrm{e}^3$ and $u > u^\star = 1$, see Figure 22.6.

22.4 Dubins Car

The Dubins car model represents a simple car with three degrees of freedom, conceptualized as a rigid body moving in a two-dimensional plane. In this model, the back wheels can only slide, which complicates maneuvers like parallel parking.

If all wheels could rotate freely, parking would be a straightforward task.

The position of the car is described by the triplet $(x, y, \theta) \in \mathbb{R}^2 \times \mathbb{S}$, where x and y represent the coordinates in the plane, and θ denotes the angle of the car relative to the x-axis.

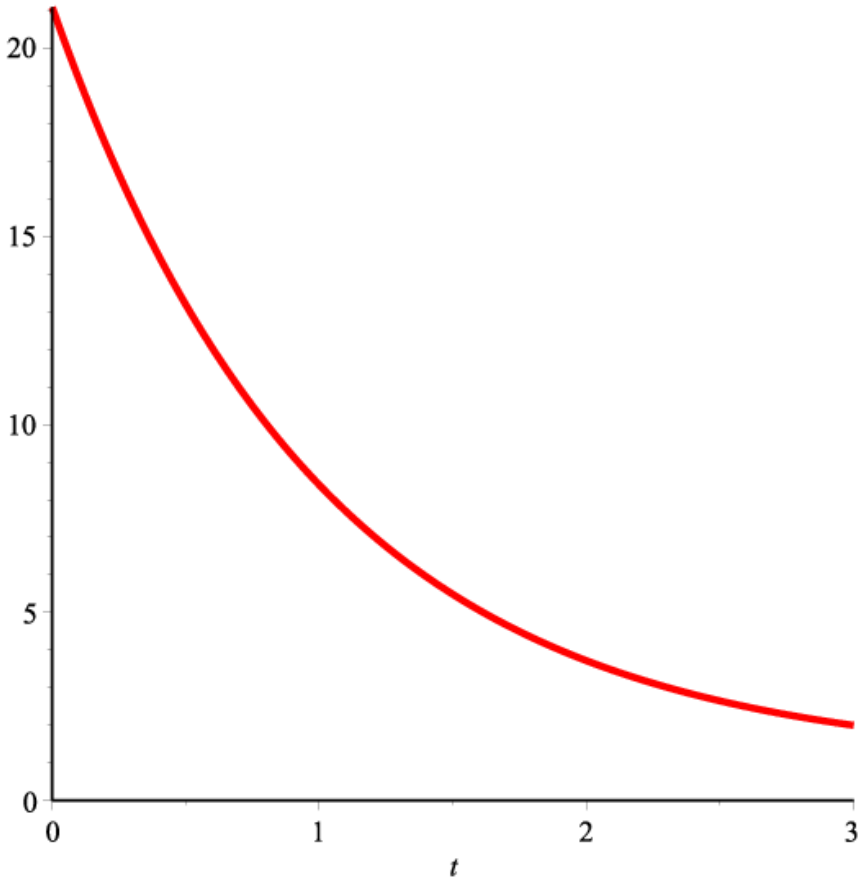


Figure 22.6: The plot of $t-4+\lambda$.

From the geometry of this model (with certain constants simplified to unity), we can derive the following system of differential equations:

$$\begin{aligned} x' &= \cos(\theta), \\ y' &= \sin(\theta), \\ \theta' &= u. \end{aligned}$$

The objective is to drive the car from a specified position to the origin in minimum time. The control variable u, which represents the steering angle rate change, is constrained to lie within the interval $[-2,2]$. Therefore, the optimal control problem can be formulated as follows:

$$\min T = \min \int_0^T 1 \, \mathrm{d}t \qquad s.t. \qquad |u| \leq 2 \qquad \text{with}$$

$$\begin{aligned} x' &= \cos(\theta), & x(0) &= 4, & x(T) &= 0, \\ y' &= \sin(\theta), & y(0) &= 0, & y(T) &= 0, \\ \theta' &= u, & \theta(0) &= \frac{\pi}{2}. \end{aligned}$$

The control variable u appears linearly in the equations, indicating the potential for a singular arc in the solution. The Hamiltonian for this problem is given by:

$$\mathscr{H}(x, y, \theta, \lambda_1, \lambda_2, \lambda_3, u) = 1 + \lambda_1 \cos(\theta) + \lambda_2 \sin(\theta) + \lambda_3 u.$$

From the PMP,

$$u = \underset{u \in [-2,2]}{\operatorname{argmin}} \mathscr{H}(x, y, \theta, \lambda_1, \lambda_2, \lambda_3, u) = u = \underset{u \in [-2,2]}{\operatorname{argmin}} \lambda_3 u$$

therefore we can write

$$u = \begin{cases} 2 & \text{if} \quad \lambda_3 < 0, \\ ? & \text{if} \quad \lambda_3 = 0, \\ -2 & \text{if} \quad \lambda_3 > 0. \end{cases}$$

The equation of the costate are derived from the Hamiltonian,

$$\begin{aligned} -\lambda_1' &= \frac{\partial \mathscr{H}}{\partial x} = 0, \\ -\lambda_2' &= \frac{\partial \mathscr{H}}{\partial y} = 0, \\ -\lambda_3' &= \frac{\partial \mathscr{H}}{\partial \theta} = -\lambda_1 \sin\theta + \lambda_2 \cos\theta. \end{aligned}$$

From the previous equations, the multipliers λ_1 and λ_2 are real constants. Upon further differentiation of λ_3', we find that within the singular arc, $\lambda_3'' = 0$. Therefore, we can conclude that:

$$\lambda_3'' = \lambda_1 \theta' \cos\theta + \lambda_2 \theta' \sin\theta = \lambda_1 u \cos\theta + \lambda_2 u \sin\theta = 0,$$

In the singular arc, we have the condition $u(\lambda_1 \cos\theta + \lambda_2 \sin\theta) = 0$, which implies that $u = 0$. Given this information about u, we can reconstruct the singular arc. Since $\theta' = u = 0$, it follows that $\theta(t) = K$, where K is a constant. Consequently, this leads to:

$$x' = \cos K \quad \text{and} \quad y' = \sin K.$$

Integrating these equations yields the singular arc:

$$x(t) = t\cos K + K_1,$$
$$y(t) = t\sin K + K_2,$$

where K_1 and K_2 are two real constants.

Next, we analyze the non-singular arc, where we have:

$$\theta(t) = -2t\,\text{sign}(\lambda_3) + k_3.$$

Using the initial condition $\theta(0) = \frac{\pi}{2}$, we can solve for k_3:

$$k_3 = \frac{\pi}{2}.$$

Letting $m = -2\,\text{sign}(\lambda_3)$, the associated arc becomes:

$$x'(t) = \cos(mt + k_3) \quad \Longrightarrow \quad x(t) = \frac{1}{m}\sin(mt + \pi/2) + k_1,$$
$$y'(t) = \sin(mt + k_3) \quad \Longrightarrow \quad y(t) = -\frac{1}{m}\cos(mt + \pi/2) + k_2.$$

We now assume that there is only a singular arc in the interval $(t_A, T]$, where t_A represents the unknown switching instant. This assumption is supported by the fact that the final condition for θ is free, leading to the conclusion that the associated multiplier λ_3 is zero at $t = T$.

Consequently, the arc is non-singular in the interval $[0, t_A]$. By examining the initial conditions, we can infer that $\lambda_3 < 0$ in this interval. This results in:

$$k_1 = 4 - \frac{1}{2} = \frac{7}{2} \quad \text{and} \quad k_2 = -\cos\left(\frac{\pi}{2}\right) = 0.$$

Thus, the associated trajectory is given by:

$$x(t) = \frac{1}{2}\cos(2t) + \frac{7}{2},$$
$$y(t) = \frac{1}{2}\sin(2t),$$
$$\theta(t) = 2t + \frac{\pi}{2}.$$

We can now connect the first arc with the second, as the trajectory is a continuous function at $t = t_A$. This ensures a smooth transition between the two arcs.

$$x(t_A) = t_A\cos K + K_1 = \frac{1}{2}\cos(2t_A) + \frac{7}{2},$$
$$y(t_A) = t_A\sin K + K_2 = \frac{1}{2}\sin(2t_A),$$
$$\theta(t_A) = K = 2t_A + \frac{\pi}{2}.$$

We can also impose the final conditions

$$x(T) = T\cos K + K_1 = 0,$$
$$y(T) = T\sin K + K_2 = 0.$$

This forms a nonlinear system of five equations involving the five unknowns K, K_1, K_2, t_A, and T. We will solve this system next.

By performing some algebraic manipulations, we can simplify the relationships among K, K_1, and K_2. This reduction leads us to two equations with two unknowns: T and t_A.

$$-T\sin(2t_A)+t_A\sin(2t_A)+\frac{1}{2}\cos(2t_A)+\frac{7}{2}=0,$$
$$T\cos(2t_A)-t_A\cos(2t_A)+\frac{1}{2}\sin(2t_A)=0.$$

By multiplying the first equation by $\cos(2t_A)$ and adding it to the second equation multiplied by $\sin(2t_A)$, we can combine these terms. Similarly, if we multiply the first equation by $\sin(2t_A)$ and add it to the second equation multiplied by $\cos(2t_A)$, we obtain:

$$\frac{1}{2}+\frac{7}{2}\cos(2t_A)=0,$$
$$-T+t_A+\frac{7}{2}\sin(2t_A)=0.$$

From this pair of expressions, it becomes straightforward to solve for the switching instant t_A and the final (minimum) time T. Consequently, we can also determine the three constants K, K_1, and K_2.

$$\begin{aligned}
t_A &= \frac{\pi}{2}-\frac{1}{2}\arccos\frac{1}{7} && \approx 0.857071948,\\
T &= \frac{\pi}{2}-\frac{1}{2}\arccos\frac{1}{7}+2\sqrt{3} && \approx 4.321173564,\\
K &= \frac{3\pi}{2}-\arccos\frac{1}{7} && \approx 3.284940223,\\
K_1 &= \frac{2\sqrt{3}}{7}\left(\pi-\arccos\frac{1}{7}+4\sqrt{3}\right) && \approx 4.276852666,\\
K_2 &= \frac{\pi}{14}-\frac{1}{14}\arccos\frac{1}{7}+\frac{2\sqrt{3}}{7} && \approx 0.6173105091.
\end{aligned}$$

These constants allow us to fully characterize the state of the system at any time $t \in [0,T]$. Next, we need to specify the costate. We have already established that λ_1 and λ_2 are constant. For λ_3, we have the differential equation given by

$$-\frac{\partial \mathscr{H}}{\partial \theta}=0$$

for the singular arc. To set up a nonlinear system involving the three unknowns λ_1, λ_2, and λ_3, we require two additional equations. One of these can be derived from the Hamiltonian itself, which is autonomous and thus equal to zero. The third equation can be expressed as the multiplier in the interval $[0,t_A]$:

$$\lambda_3(t)=-\frac{\lambda_1}{2}\cos(2t+k_3)-\frac{\lambda_2}{2}\sin(2t+k_3)+\lambda_3(0).$$

Here, we introduce a fourth unknown, $\lambda_3(0)$, but we do not require an additional equation for it. Consequently, the nonlinear system at $t=t_A$ is given by:

$$\begin{aligned}
\mathscr{H} &= 1+\lambda_1\cos(\theta)+\lambda_2\sin(\theta)+\lambda_3 u=0,\\
\lambda_3'(t_A) &= -\frac{\partial \mathscr{H}}{\partial \theta}=\lambda_1\sin\theta-\lambda_2\cos\theta=0,\\
\lambda_3(t_A) &= -\lambda_1/2\cos(2t_A+k_3)-\lambda_2/2\sin(2t_A+k_3)+\lambda_3(0).
\end{aligned}$$

The solution of the system gives the following expressions,

$$\lambda_1 = \frac{\cos(k_3)}{\sin(2t_A)\cos k_3 - \sin K \cos(2t_A)} = \frac{4}{7}\sqrt{3} \quad \approx 0.9897433188,$$
$$\lambda_2 = \frac{\sin(k_3)}{\sin(2t_A)\cos k_3 - \sin K \cos(2t_A)} = \frac{1}{7} \quad \approx 0.1428571429,$$
$$\lambda_3(0) = -\frac{1}{2}.$$

In Figure Figure 22.7 the plots for state, costate, control and trajectory.

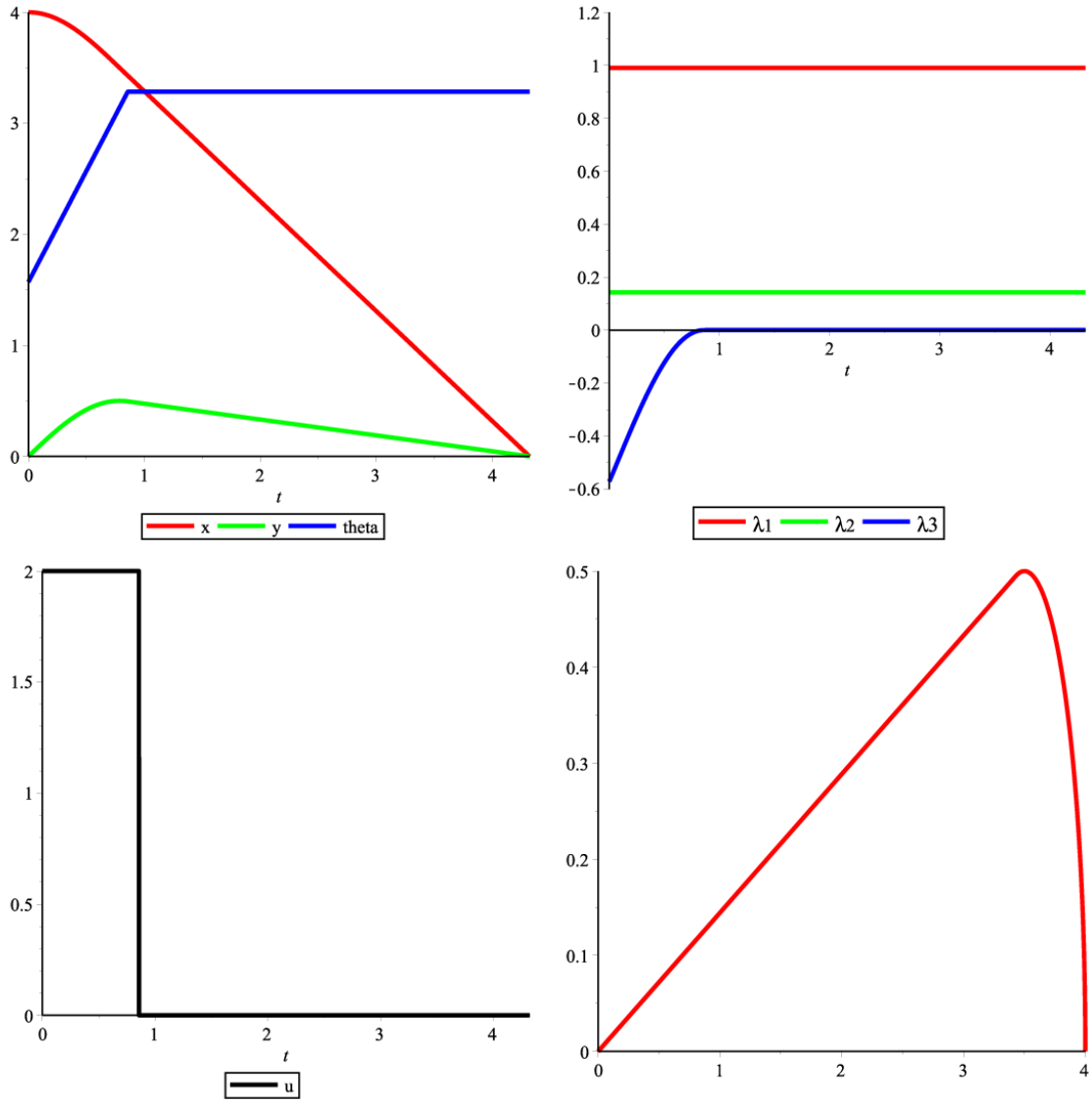


Figure 22.7: Variational solution to the Dubins car problem. On the top left the states , on the right the costates, below the control and the trajectory.

22.5 Four Singular Optimal Control Problems

In this chapter, we present four optimal control problems (OCPs) that feature singular arcs. Three of these examples are drawn from the PROPT handbook, which accompanies a commercial software tool for solving OCPs.

Most of these problems are discussed in [8], though only a few provide exact solutions. Here, we present the analytical solutions to some of these problems, enabling a precise comparison with the numerical solutions generated by the software used for benchmarking.

Optimal Control Problem with Singular Controls

In optimal control theory, when the Hamiltonian $\mathscr{H}$ is linear in the control variable u, the solution may exhibit discontinuities. If the switching function associated with the control is not sustained over an interval—meaning the coefficient of u in $\mathscr{H}$ equals zero only at isolated points—the control is classified as bang-bang. A bang-bang control always takes on extreme values from the control set.

Conversely, if the coefficient of u in $\mathscr{H}$ is zero over an interval, the control is termed singular. In this case, the choice of u must be derived from information beyond Pontryagin's maximum principle. The moments when the optimal control transitions between states or switches to singular controls are referred to as switch times, which can be challenging to identify.

22.5.1 Singular 1

Consider the problem of minimizing the functional:

$$\begin{aligned} \text{minimize:} \quad & \int_{-1}^{1} (x-1+t^2)^2 \, \mathrm{d}t, \\ \text{subject to:} \quad & x' = u, \\ \text{with control bounded by:} \quad & |u| \leq 1. \end{aligned}$$

The Hamiltonian for this problem is given by:

$$\mathscr{H} = (x-1+t^2)^2 + \lambda u + p_\varepsilon(u),$$

where $p_\varepsilon(u)$ is a penalty function designed to manage constrained controls within the interval $[-1,1]$. An example of such a function is

$$p_\varepsilon(u) = -\varepsilon \ln\left(\cos\left(\frac{\pi}{2}u\right)\right).$$

As $\varepsilon \to 0$, this function approaches zero within the interval $[-1,1]$ and increases to infinity outside that range. Consequently, the Pontryagin maximum principle becomes ineffective in this scenario.

The stationarity condition for the optimal control problem is:

$$\frac{\partial \mathscr{H}}{\partial u} = \lambda + \frac{\partial p_\varepsilon(u)}{\partial u} = 0.$$

This leads to:

$$\begin{aligned} &\lambda(t) = -\varepsilon \frac{\pi}{2} \tan\left(\frac{\pi}{2}u(t)\right) \implies \\ &u(t) = \lim_{\varepsilon \to 0} -\frac{2}{\pi} \arctan\left(\frac{2\pi\lambda(t)}{\varepsilon}\right) = -\operatorname{sign}(\lambda(t)). \end{aligned}$$

Thus, when $\lambda \neq 0$, the control takes values $u = \pm 1$. When $\lambda = 0$, the control is singular.

The state is not specified at the boundaries, leading to the transversality conditions $\lambda(-1) = \lambda(1) = 0$. Differentiating the Hamiltonian with respect to x provides information about the multiplier λ:

$$\lambda' = -\frac{\partial \mathscr{H}}{\partial x} = -2(x-1+t^2).$$

Integrating this expression yields:

$$\lambda(t) = -2\int_{-1}^{t}(x(s)-1+s^2)\,\mathrm{d}s + k, \tag{22.10}$$

where $k \in \mathbb{R}$ can be determined using the initial condition $\lambda(-1) = 0$:

$$0 = \lambda(-1) = -2\int_{-1}^{-1}(x(s)-1+s^2)\,\mathrm{d}s + k \implies k = 0.$$

Substituting $k = 0$ into equation Equation 22.10 allows us to write:

$$0 = \lambda(1) = -2\int_{-1}^{-1}(x(s)-1+s^2)\,\mathrm{d}s \implies$$
$$\int_{-1}^{-1}(x(s)-1+s^2)\,\mathrm{d}s = 0.$$

For t in the interval $[t_A, t_B] \subset [-1,1]$, where $\lambda(t) = 0$, we have:

$$\begin{aligned} 0 &= \lambda(t) \\ &= -2\int_{-1}^{t_A}(x(s)-1+s^2)\,\mathrm{d}s - 2\int_{t_A}^{t}(x(s)-1+s^2)\,\mathrm{d}s \\ &= \lambda(t_A) - 2\int_{t_A}^{t}(x(s)-1+s^2)\,\mathrm{d}s. \end{aligned} \tag{22.11}$$

Differentiating the right-hand side gives:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left(-2\int_{t_A}^{t}(x(s)-1+s^2)\,\mathrm{d}s\right)$$
$$\Downarrow$$
$$x(t)-1+t^2 = 0, \quad t \in [t_A, t_B].$$

Thus, when the control is singular, we find that

$$x(t) = 1 - t^2,$$

and from the dynamics of the problem,

$$u(t) = -2t \quad \text{for } t \in [t_A, t_B].$$

It is important to note that for $|t| > 1/2$, the control $u(t) = -2t$ does not satisfy the imposed bounds. Therefore, we need to identify two switching times, t_A and t_B, such that $x' = u$ and $|u| \leq 1$. For $t \in [-1, t_A)$, we have $u(t) = 1$, and for $t \in (t_B, 1]$, we have $u(t) = -1$. The corresponding state equations are then given by:

- For $t < t_A$: $x(t) = t + a$
- For $t > t_B$: $x(t) = -t + b$.

From Equation 22.11 and the fact that for $t \in [t_A, t_B]$, $x(t) = 1 - t^2$, it follows that $\lambda(t_A) = 0$, i.e,

$$\begin{aligned} 0 = \lambda(t_A) &= -2\int_{-1}^{t_A}(x(s)-1+s^2)\,\mathrm{d}s \\ &= -2\int_{-1}^{t_A}(s+a-1+s^2)\,\mathrm{d}s \\ &= -2\left[\frac{1}{3}(t_A+1)\left(t_A^2 + \frac{1}{2}t_A - \frac{7}{2} + 3a\right)\right]. \end{aligned}$$

This expression is equal to zero for

$$a = -\frac{1}{3}t_A^2 - \frac{1}{6}t_A + \frac{7}{6} \implies x(t) = t - \frac{1}{3}t_A^2 - \frac{1}{6}t_A + \frac{7}{6}. \tag{22.12}$$

Now, because x is continuous, in the switching point must hold $t_A + a = 1 - t_A^2$. From that equation it is possible to solve $t_A = -\frac{1}{4}$ and $a = \frac{19}{16}$.

With similar consideration it is possible to solve b and t_B: in fact,

$$\begin{aligned} 0 = \lambda(1) = \lambda(t_B) - 2\int_{t_B}^{1} x(s) - 1 + s^2 \,\mathrm{d}s \implies \\ \int_{t_B}^{1} x(s) - 1 + s^2 \,\mathrm{d}s = 0. \end{aligned}$$

Using the expression $x(t) = -t + b$ for $t \in (t_B, 1]$, the previous integral yields

$$\begin{aligned} 0 &= \int_{t_B}^{1} -s + b - 1 + s^2 \,\mathrm{d}s \\ &= -\frac{1}{3}(t_B - 1)\left(t_B^2 - \frac{1}{2}t_B - \frac{7}{2} + 3b\right), \end{aligned}$$

which gives the same result of equation Equation 22.13,

$$b = -\frac{1}{3}t_B^2 - \frac{1}{6}t_B + \frac{7}{6} \implies x(t) = -t - \frac{1}{3}t_B^2 - \frac{1}{6}t_B + \frac{7}{6}. \tag{22.13}$$

Again, for the continuity of x, $-t_B + b = 1 - t_B^2$, implies that $t_B = \frac{1}{4}$ and $b = \frac{19}{16}$. This also shows the symmetry of the solution. With this optimal control u and state x the minimum of the functional is

$$\begin{aligned} \int_{-1}^{1} (x - 1 + t^2)^2 \,\mathrm{d}t &= \int_{-1}^{-1/4} \left(t + \frac{9}{16} - 1 + t^2\right)^2 \mathrm{d}t \\ &\quad + \int_{1/4}^{1} \left(-t + \frac{9}{16} - 1 + t^2\right)^2 \mathrm{d}t \\ &= \frac{9}{1280} = 0.00703125. \end{aligned}$$

Collecting the three phases the analytical solution of the optimal control problem is,

$$x(t) = \begin{cases} t + a \text{ for} & t \in [-1, t_A) = \left[-1, -\frac{1}{4}\right) \\ 1 - t^2 \text{ for} & t \in [t_A, t_B] = \left[-\frac{1}{4}, \frac{1}{4}\right], \\ -t + b \text{ for} & t \in (t_B, 1] = \left(\frac{1}{4}, 1\right] \end{cases} \quad a = b = \frac{19}{16},$$

$$u(t) = \begin{cases} 1 \text{ for} & t \in [-1, t_A) = \left[-1, -\frac{1}{4}\right) \\ -2t \text{ for} & t \in [t_A, t_B] = \left[-\frac{1}{4}, \frac{1}{4}\right] \\ -1 \text{ for} & t \in (t_B, 1] = \left(\frac{1}{4}, 1\right] \end{cases}$$

$$\lambda(t) = \begin{cases} -\frac{2}{3}t^3 - t^2 - \frac{3}{8}t - \frac{1}{24} \text{ for} & t \in [-1, t_A) = \left[-1, -\frac{1}{4}\right) \\ 0 \text{ for} & t \in [t_A, t_B] = \left[-\frac{1}{4}, \frac{1}{4}\right], \\ -\frac{2}{3}t^3 + t^2 - \frac{3}{8}t + \frac{1}{24} \text{ for} & t \in (t_B, 1] = \left(\frac{1}{4}, 1\right]. \end{cases}$$

A graphical representation of the state x, the control u and the costate λ is shown in Figure 22.8.

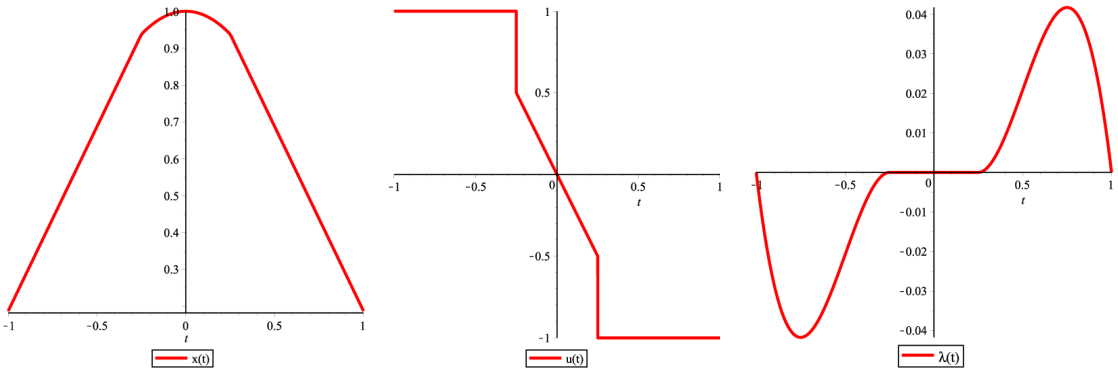


Figure 22.8: Variational solution to the singular arc problem. On the left the trajectory of the state $x(t)$, in the middle the control $u(t)$ and on the right the costate $\lambda(t)$.

In Figure 22.9 the numerical solution obtained with ACADO.

22.5.2 Singular 2

The first example is relatively straightforward from an analytical standpoint and could be approached using intuition alone. However, it has been explored by several researchers—including Jacobson, Chen, Huang, and Luus (refer to Example 10.2.1)—who have shown that achieving a reliable numerical solution is more challenging than it initially appears. The dynamical system is characterized by two equations and a functional:

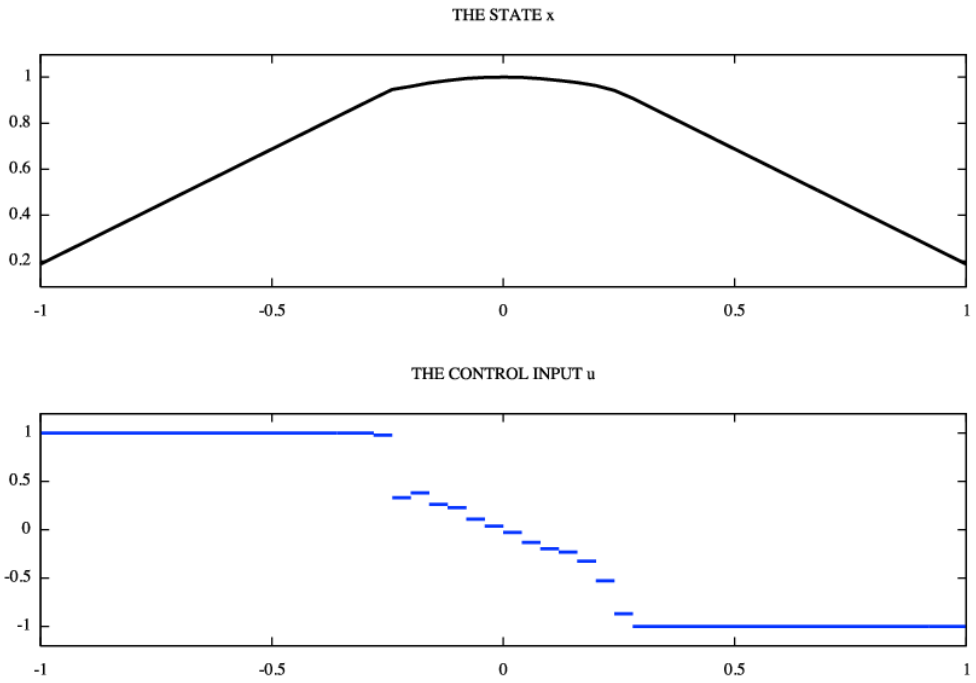


Figure 22.9: Numerical solution given by ACADO.

$$\begin{aligned} \text{minimize:} \quad & x_2(T) \\ \text{subject to the dynamic:} \quad & x_1(t)' = u(t), \\ & x_2(t)' = \frac{1}{2}x_1(t)^2, \\ \text{with control bounded by:} \quad & |u| \leq 1. \end{aligned}$$

The boundary conditions provided are $x_1(0) = 1$ and $x_2(0) = 0$, leading to the conditions $\lambda_1(T) = 0$ and $\lambda_2(T) = 1$, where T is the final time, set to $T = 2$. The Hamiltonian for this problem is given by:

$$\mathscr{H} = \lambda_1 u + \frac{1}{2}\lambda_2 x_1^2.$$

The equation of the costate are derived from the Hamiltonian,

$$\begin{aligned} -\lambda_1' &= \frac{\partial \mathscr{H}}{\partial x_1} = -\lambda_2 x_1, \\ -\lambda_2' &= \frac{\partial \mathscr{H}}{\partial x_2} = 0, \end{aligned}$$

Thus, we have $\lambda_2(t) = 1$, which is constant. Assuming there is only one switching point, we can infer that the singular arc is the second one. By deriving the multiplier λ_1 in the singular part, we obtain the singular control. From the equation $\lambda_1' = -x_1$, we find that $\lambda_1'' = -x_1' = -u = 0$. This indicates that the control is zero within the singular arc.

Next, we can integrate the dynamical system to determine both the state and the costate. The resulting expressions are:

$$
\begin{aligned}
x_1(t) &= K, \\
x_2(t) &= \frac{1}{2}K^2 t + K_2, \\
\lambda_1(t) &= -Kt + 2K, \\
\lambda_2(t) &= 1,
\end{aligned}
$$

for some constants K and K_2 that need to be determined, and for $t \in [t_A, T]$, where t_A is the unknown switching time.

In the first arc, the control is derived using Pontryagin's maximum principle, resulting in a constant control of $u(t) = -1$ for $t \in [0, t_A)$. By applying the initial conditions, we can solve for the state, yielding:

$$
\begin{aligned}
x_1(t) &= -t + 1, \\
x_2(t) &= \frac{1}{6}t^3 - \frac{1}{2}t^2 + \frac{1}{2}t, \\
\lambda_1(t) &= \frac{1}{2}t^2 - t + \lambda_0, \\
\lambda_2(t) &= 1,
\end{aligned}
$$

for $t \in [0, t_A)$. By applying the continuity of the states at the switching point, we can establish a nonlinear system involving the unknowns K, K_2, λ_0, and t_A. This system yields two solutions: one lacks physical significance, while the other provides $t_A = 1$, $K = 0$, $K_2 = \frac{1}{6}$, and $\lambda_0 = \frac{1}{2}$. Consequently, the corresponding value for $x_2(T)$ is $\frac{1}{6}$. The plots of the system are illustrated in Figure 22.10.

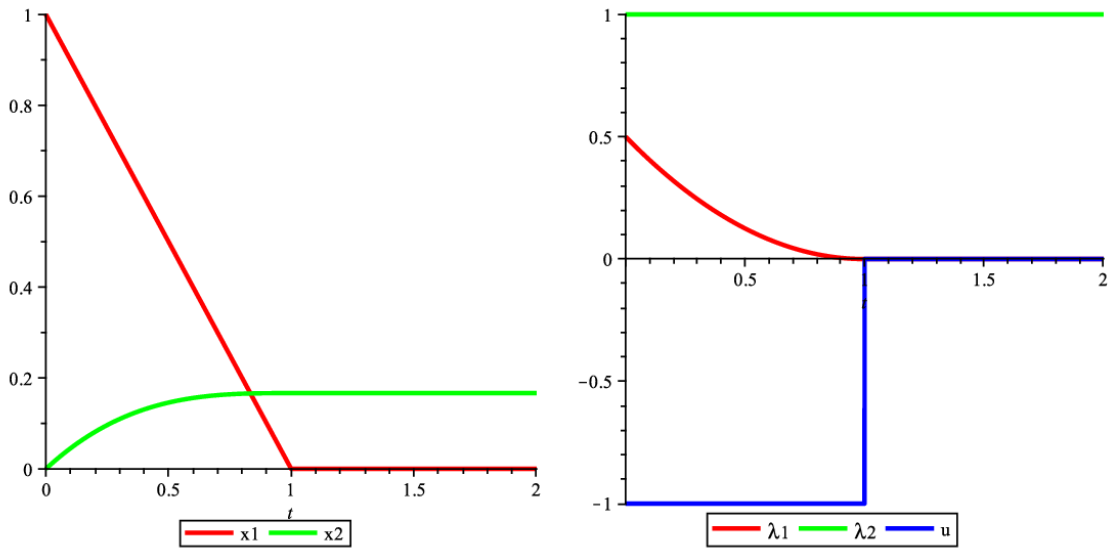


Figure 22.10: Variational solution to Singular Problem 1. On the left the states, on the right the costates and the control.

22.5.3 Singular 3

Constrained 1d vehicle

In this section, we analyze the model of a one-dimensional vehicle. We will explore various types of control strategies: a linear control based on jerk and a quadratic control on jerk. The linear control results in both singular and bang-bang arcs, while the quadratic control provides a smoother control profile.

The optimal control problem can be formulated as follows:

$$\min \int_0^T 1 + wJ^2 \, \mathrm{d}t$$

where w is set to zero in the first case, and $w = 0.1$ in the second case, subject to the following dynamic and path constraints:

$$\begin{aligned} s' &= v \\ v' &= a(p) \\ p' &= J \end{aligned}$$

with boundary conditions:

$$\begin{aligned} s(0) &= 0, & s(T) &= 200 \\ v(0) &= 10, & v(T) &= 10 \\ p(0) &= 0, & p(T) &= \text{free}. \end{aligned}$$

The path constraint is

$$a_{\min} - a(p) \leq 0, \qquad a(p) = p \cdot \begin{cases} a_1 & p \geq 0, \\ a_2 & p < 0. \end{cases}$$

where the constants are $a_{\min} = -5$, $a_1 = 3$ and $a_2 = 10$. The state p is constrained in $[-1, 1]$ and the control $J \in [-1, 1]$.

Jerk as quadratic control

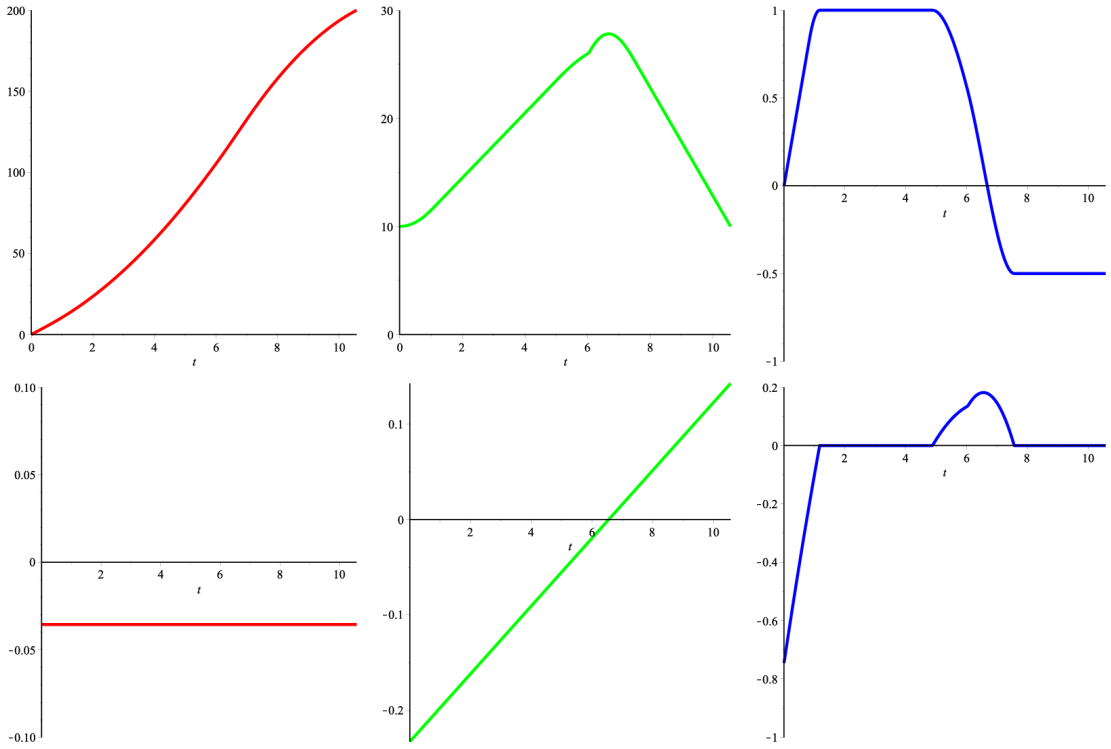


Figure 22.11: In red, green and blue respectively, on top the three states s, v and p, below the corresponding multipliers λ_1, λ_2, λ_3.}

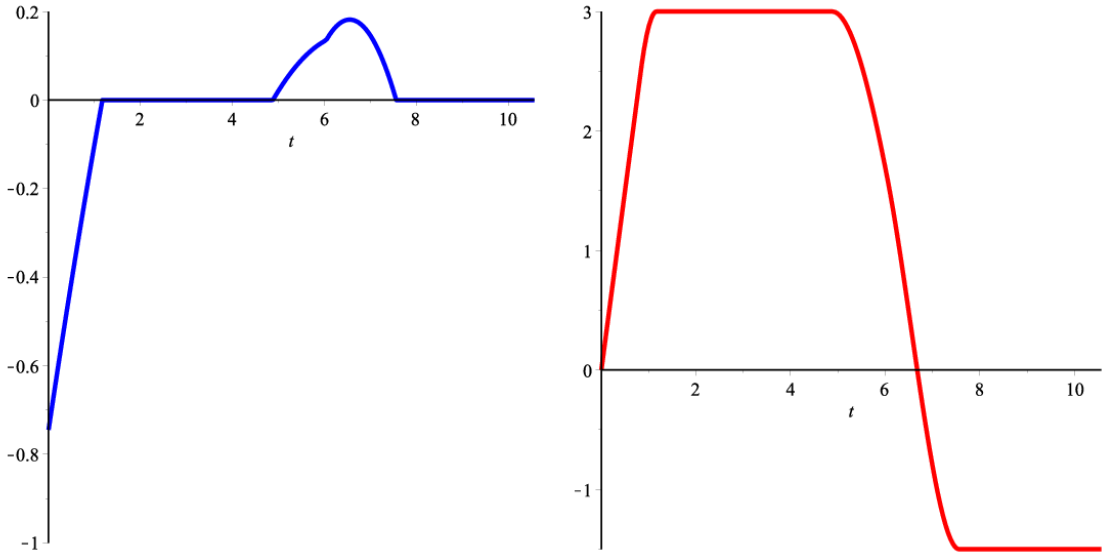


Figure 22.12: On the left the control J, on the right the acceleration $a(t)$.

22.5.4 Singular 4

Dadebo example 1

Here we analyze the problem proposed by many researchers which is stated in the paper by Dadebo et McAuley.

$$\text{minimize:} \quad y(1)$$

subject to the dynamical evolution

$$x' = u$$
$$y' = x^2 + u^2,$$

with boundary conditions $x(0) = x(1) = 1$ and $y(0) = 0$.
The Hamiltonian for this problem is

$$\mathscr{H} = \lambda u + \mu(x^2 + u^2),$$

The adjoint equations are given by $\lambda' = -2\mu x$ and $\mu' = 0$, with the boundary condition $\mu(1) = 1$. The control is defined as $u = -\frac{\lambda}{2\mu}$. Consequently, the multiplier μ is constant, so we have $\mu(t) = 1$.

Differentiating the equation for λ', we obtain $\lambda'' = -2x' = -2u = \lambda$. This leads to the characteristic equation, which has the general solution:

$$\lambda(t) = \alpha e^t + \beta e^{-t},$$

where α and β are unknown constants. From the multiplier λ, we can express the control as:

$$u(t) = -\frac{\alpha}{2}e^t - \frac{\beta}{2}e^{-t}.$$

Finally, the state x is determined by the differential equation $x' = u$, yielding:

$$x(t) = -\frac{\alpha}{2}e^t + \frac{\beta}{2}e^{-t} + c,$$

where c is a constant. The state y can be obtained from the equation $y' = x^2 + u^2$, which results in:

$$\begin{aligned} x^2 &= \frac{\alpha^2}{4}\mathrm{e}^{2t} + \frac{\beta^2}{4}\mathrm{e}^{-2t} + c^2 - \frac{\alpha\beta}{2} - \alpha c\mathrm{e}^{t} + \beta c\mathrm{e}^{-t} \\ u^2 &= \frac{\alpha^2}{4}\mathrm{e}^{2t} + \frac{\beta^2}{4}\mathrm{e}^{-2t} + \frac{\alpha\beta}{2} \\ x^2 + u^2 &= \frac{\alpha^2}{2}\mathrm{e}^{2t} + \frac{\beta^2}{2}\mathrm{e}^{-2t} + c^2 - \alpha c\mathrm{e}^{t} + \beta c\mathrm{e}^{-t}. \end{aligned}$$

Therefore, by integrating the last equation, we can compute the state y, which is given by:

$$y(t) = \frac{\alpha^2}{4}\mathrm{e}^{2t} - \frac{\beta^2}{4}\mathrm{e}^{-2t} + c^2 t - \alpha c\mathrm{e}^{t} - \beta c\mathrm{e}^{-t} + k$$

for an unknown constant k, which can be determined using the boundary conditions. We can find the four constants α, β, c, and k by solving the following nonlinear system:

$$\begin{aligned} x(0) &= -\frac{\alpha}{2} + \frac{\beta}{2} + c = 1 \\ x(1) &= -\frac{\alpha}{2}e + \frac{\beta}{2}\mathrm{e}^{-1} + c = 1 \\ y(0) &= \frac{\alpha^2}{4} - \frac{\beta^2}{4} - \alpha c - \beta c + k = 0 \\ \mathscr{H}(t=0) = \mathscr{H}(t=1) \quad &\Longrightarrow \quad -\alpha c + \beta c = -\alpha c e + \beta c\mathrm{e}^{-1}. \end{aligned}$$

The last equation arises from the constancy of the Hamiltonian along an optimal control and can be simplified to $c(\beta - \alpha) = c(\beta e^{-1} - \alpha e)$. Thus, if $c \neq 0$, we find that $\alpha = -\beta = 0$ and $c = 1$. Substituting these values into the third equation of the nonlinear system yields $k = 0$, resulting in one extremal solution given by $x(t) = 1$, $y(t) = t$, and $u(t) = 0$. Consequently, the minimized functional is $y(1) = 1$.

However, the nonlinear system also admits another solution when $c = 0$. From the first equation, we have $\beta = \alpha + 2$, which, when substituted into the second equation, gives $\alpha = \frac{-2}{1+e}$ and $\beta = \frac{2e}{1+e}$. Plugging these values into the third equation results in $k = \frac{e-1}{e+1}$.

This alternative solution yields a lower minimized functional compared to the previous one, achieving an optimal value of:

$$y(1) = 2\frac{e-1}{e+1} \approx 0.9242343144.$$

22.6 Free time problems

22.6.1 Example 1

Given the free time optimal control problem

$$\begin{aligned} \text{minimize:} \quad & \int_0^T \left(1 + t^2\right)u^2 + y\,\mathrm{d}t \\ \text{subject to:} \quad & x' = y + u \qquad y' = xy \\ & x(0) = 0 \qquad y(T) = 1 \end{aligned}$$

the first thing to do is to rewrite the statement in a form where the integral bound are known by performing a change of variable ζ that we arbitrary set to be defined in the domain $[0, 1]$ resulting in the transformation $\zeta = t/T$. The optimal control problem can so be stated as

$$\begin{aligned}
\text{minimize:} \quad & \int_0^1 \Big(\big(1+\zeta^2 T(\zeta)^2\big) u^2(\zeta) + y(\zeta) \Big) T(\zeta) \mathrm{d}\zeta \\
\text{subject to:} \quad & \begin{cases} x'(\zeta) = T(\zeta)\big(y(\zeta)+u(\zeta)\big) \\ y'(\zeta) = T(\zeta) x(\zeta) y(\zeta) \\ T'(\zeta) = 0 \end{cases} \\
\text{and BC:} \quad & \begin{cases} x(0) = 0 \\ y(1) = 1 \end{cases}
\end{aligned}$$

The hamiltonian $\mathscr{H}$ related to such problem is so

$$\mathscr{H}\left(\begin{pmatrix} x \\ y \\ T \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}, u, \zeta \right) = T\big(1+\zeta^2 T^2\big)u^2 + Ty - \lambda_1 T\big(y+u\big) - \lambda_2 Txy + 0\lambda_3$$

Considering the original ordinary differential equation of the problem and the addition of the co-equations the system that needs to be solved is

$$\begin{cases}
x'(\zeta) = T(\zeta)\big(y(\zeta)+u(\zeta)\big) \\
y'(\zeta) = T(\zeta) x(\zeta) y(\zeta) \\
T'(\zeta) = 0 \\
\lambda_1'(\zeta) = -\dfrac{\partial \mathscr{H}}{\partial x} = \lambda_2(\zeta) T(\zeta) y(\zeta) \\
\lambda_2'(\zeta) = -\dfrac{\partial \mathscr{H}}{\partial y} = -T(\zeta) + \lambda_1(\zeta) + \lambda_2(\zeta) T(\zeta) x(\zeta) \\
\lambda_3'(\zeta) = -\dfrac{\partial \mathscr{H}}{\partial T} = -u^2(1+3\zeta^2 T^2) - \lambda_1(y+u) - (1+\lambda_2 x) y
\end{cases}$$

on which the boundary conditions are

$$\begin{aligned}
x(0) &= 0 \\
y(1) &= 1 \\
\lambda_1(1) &= 0 \\
\lambda_2(0) &= 0 \\
\lambda_3(0) = \lambda_3(1) &= 0
\end{aligned}$$

We also need to consider the control law

$$\frac{\partial \mathscr{H}}{\partial u}$$

that becomes

$$2T\big(1+\zeta^2 T^2\big)u - \lambda_1 = 0$$

and so we can explicit the control u as

$$u(\zeta) = \frac{\lambda_1(\zeta)}{2T(\zeta)\big(1+\zeta^2 T(\zeta)^2\big)}$$

22.6.2 Example 2 (exam's simulation)

Given the optimal control problem in the control u and states x, y

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 \left(1+x^2\right)u^2 \,\mathrm{d}t \\ \text{subject to:} \quad & x' = xy - u \\ & y' = x + u \\ & \int_0^1 \left(x+u\right) \mathrm{d}t = 2 \\ & y(0) = 2 \end{aligned}$$

the first thing to do is remove the integral constraint by adding a new state variable z such that

$$z' = x + u$$

with boundary conditions

$$z(0) = 0 \quad \text{and} \quad z(1) = 2$$

with that the problem becomes

$$\begin{aligned} \text{minimize:} \quad & \int_0^1 \left(1+x^2\right)u^2 \,\mathrm{d}t \\ \text{subject to:} \quad & x' = xy - u \\ & y' = x + u \\ & z' = x + u \\ & y(0) = 2 \quad z(0) = 0 \quad z(1) = 2 \end{aligned}$$

We can so compute the Hamiltonian $\mathscr{H}$ and the utility function $\mathscr{B}$ of the problem as

$$\mathscr{H}\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}, u, t\right) = \left(1+x^2\right)u^2 + \lambda_1\left(xy-u\right) + \lambda_2\left(x+u\right) + \lambda_3\left(x+u\right)$$

$$\mathscr{B}\left(y_0, z_0, z_1, \mu_1, \mu_2, \mu_3\right) = \mu_1\left(y_0 - 2\right) + \mu_2 z_0 + \mu_3\left(z_1 - 2\right)$$

The control obtained by the Pontryagin maximum principle is the solution of u that verifies $\frac{\partial \mathscr{H}}{\partial u} = 0$, and so we have

$$\begin{aligned} \frac{\partial \mathscr{H}}{\partial u} &= 2\left(1+x^2\right)u - \lambda_1 + \lambda_2 + \lambda_3 \\ &\Downarrow \\ u(t) &= \frac{\lambda_1(t) - \lambda_2(t) - \lambda_3(t)}{2\left(1+x(t)^2\right)} \end{aligned}$$

The additional adjoint ODE are described by the following derivatives:

$$\begin{aligned} \lambda_1' &= -\frac{\partial \mathscr{H}}{\partial x} = -2xu^2 - \lambda_1 y - \lambda_2 - \lambda_3 \\ \lambda_2' &= -\frac{\partial \mathscr{H}}{\partial y} = -\lambda_1 x \\ \lambda_3' &= -\frac{\partial \mathscr{H}}{\partial z} = 0 \end{aligned}$$

while the adjoint boundary conditions are

$$\begin{aligned}
&\lambda_1(0) = -\frac{\partial \mathscr{B}}{\partial x_0} = 0 && \lambda_1(1) = \frac{\partial \mathscr{B}}{\partial x_1} = 0\\
&\lambda_2(0) = -\frac{\partial \mathscr{B}}{\partial y_0} = -\mu_1 && \lambda_2(1) = \frac{\partial \mathscr{B}}{\partial y_1} = 0\\
&\lambda_3(0) = -\frac{\partial \mathscr{B}}{\partial z_0} = -\mu_2 && \lambda_3(1) = \frac{\partial \mathscr{B}}{\partial z_1} = \mu_3
\end{aligned}$$

Some expression have been cancelled out because they are trivially solved for μ_i.
Thus the BVP become

$$\begin{cases}
\left.\begin{aligned} x' &= xy - u \\ y' &= x + u \\ z' &= x + u \end{aligned}\right\} & \text{original ODE}\\
\left.\begin{aligned} \lambda_1' &= -2xu^2 - \lambda_1 y - \lambda_2 - \lambda_3 \\ \lambda_2' &= -\lambda_1 x \\ \lambda_3' &= 0 \end{aligned}\right\} & \text{adjoint ODE}\\
\left.\begin{aligned} y(0) &= 2 \\ z(0) &= 0 \\ z(1) &= 2 \end{aligned}\right\} & \text{original BC}\\
\left.\begin{aligned} \lambda_1(0) &= 0 \\ \lambda_1(1) &= 0 \\ \lambda_2(1) &= 0 \end{aligned}\right\} & \text{adjoint BC}\\
u = \dfrac{\lambda_1 - \lambda_2 - \lambda_3}{2\left(1+x^2\right)} & \text{Pontryagin}
\end{cases}$$

22.6.3 Example 3 (free time problem, 2 controls)

Given the free time optimal control problem in the controls u, w and states x, y

$$\begin{aligned}
\text{minimize:} \quad & \int_0^T \left(x^2 + y^2\right) \mathrm{d}t \\
\text{subject to:} \quad & x' = yu \\
& y' = x + w \\
& x(0) = 0 \quad y(0) = 0 \quad y(T) = 1 \\
& u, w \in [-1, 1]
\end{aligned}$$

the first thing to do is to transform the problem in fixed boundary optimal control problem by performing the change of variable $t = \zeta\, T(\zeta)$ (with $\zeta \in [0,1]$) and so

$$\begin{aligned}
\text{minimize:} \quad & \int_0^1 T(\zeta)\left(x^2(\zeta) + y^2(\zeta)\right) \mathrm{d}\zeta \\
\text{subject to:} \quad & x'(\zeta) = T(\zeta) y(\zeta) u(\zeta) \\
& y'(\zeta) = T(\zeta)\Big(x(\zeta) + w(\zeta)\Big) \\
& T'(\zeta) = 0 \\
& x(0) = 0 \quad y(0) = 0 \quad y(1) = 1 \\
& u(\zeta), w(\zeta) \in [-1, 1]
\end{aligned}$$

From now on all the functions will be represented as depending from the independent variable ζ and it won't be written any more. With that said we can compute the hamiltonian $\mathscr{H}$ and the utility function $\mathscr{B}$ of the problem as

$$\mathscr{H}\left(\begin{pmatrix} x \\ y \\ T \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}, \begin{pmatrix} u \\ w \end{pmatrix}\right) = T\left(x^2+y^2\right) + \lambda_1 T y u + \lambda_2 T\left(x+w\right) + 0 \cdot \lambda_3$$
$$\mathscr{B}\left(x_0, y_0, y_1, \mu_1, \mu_2, \mu_3\right) = \mu_1 x_0 + \mu_2 y_0 + \mu_3\left(y_1 - 1\right)$$

The solution of the controls obtained by the Pontryagin principle becomes

$$\begin{aligned}
(u, w) &= \underset{(u,w)\in[-1,1]\times[-1,1]}{\operatorname{argmin}} \mathscr{H}\left(\begin{pmatrix} x \\ y \\ T \end{pmatrix}, \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}, \begin{pmatrix} u \\ w \end{pmatrix}\right) \\
&= \underset{(u,w)\in[-1,1]\times[-1,1]}{\operatorname{argmin}} T\left(x^2+y^2\right) + \lambda_2 T x + T(\lambda_1 y u + \lambda_2 w) \\
&= \underset{(u,w)\in[-1,1]\times[-1,1]}{\operatorname{argmin}} T(\lambda_1 y u + \lambda_2 w) \\
&= \underset{(u,w)\in[-1,1]\times[-1,1]}{\operatorname{argmin}} \lambda_1 y u + \lambda_2 w
\end{aligned}$$

to find the minimum you can use KKT

minimize

$$f(u, w) = \lambda_1 y u + \lambda_2 w$$

subject to

$$-1 \le u \le 1, \qquad -1 \le w \le 1,$$

Lagrangian

$$\mathscr{L}(u, w, \mu_1, \mu_2, \mu_3, \mu_4) = \lambda_1 y u + \lambda_2 w - \mu_1(1-u) - \mu_2(1+u) - \mu_3(1-w) - \mu_4(1+w)$$

Nonlinear sistem for Pontryagin

$$\begin{cases}
\partial_u \mathscr{L} = \lambda_1 y + \mu_1 - \mu_2 = 0 \\
\partial_w \mathscr{L} = \lambda_2 + \mu_3 - \mu_4 = 0 \\
\mu_1(1-u) = 0 \\
\mu_2(1+u) = 0 \\
\mu_3(1-w) = 0 \\
\mu_4(1+w) = 0
\end{cases} \tag{22.14}$$

The solution of Equation 22.14 is the following:

$$u(\zeta) = \begin{cases} -1 & \lambda_1 y > 0 \\ \text{undefined} & \lambda_1 y = 0 \\ +1 & \lambda_1 y < 0 \end{cases}$$

$$w(\zeta) = \begin{cases} -1 & \lambda_2 > 0 \\ \text{undefined} & \lambda_2 = 0 \\ +1 & \lambda_2 < 0 \end{cases}$$

remember that must be $T > 0$ for $T \leq 0$ solution is discarded.

We that stated we can build the boundary value problem solution of the free time optimal control problem by adding at the initial ODEs and boundary conditions the the adjoint ones and the control law (related to Pontryagin minimum principle):

$$\begin{cases} \left.\begin{aligned} x' &= Tyu \\ y' &= T(x+w) \\ T' &= 0 \end{aligned}\right\} \text{ original ODE} \\ \left.\begin{aligned} \lambda_1' &= -\frac{\partial \mathscr{H}}{\partial x} = -2xT - \lambda_2 T \\ \lambda_2' &= -\frac{\partial \mathscr{H}}{\partial y} = -2Ty - \lambda_1 Tu \\ \lambda_3' &= -\frac{\partial \mathscr{H}}{\partial T} = -x^2 - y^2 - \lambda_1 yu - \lambda_2(x+w) \end{aligned}\right\} \text{ adjoint ODE} \\ \left.\begin{aligned} x(0) &= 0 \\ y(0) &= 0 \\ y(1) &= 1 \end{aligned}\right\} \text{ original BC} \\ \left.\begin{aligned} \lambda_1(1) &= 0 \\ \lambda_3(0) &= 0 \\ \lambda_3(1) &= 0 \end{aligned}\right\} \text{ adjoint BC} \\ \left.\begin{aligned} u &= \begin{cases} -1 & \lambda_1 y > 0 \\ \text{undefined} & \lambda_1 y = 0 \\ 1 & \lambda_1 y < 0 \end{cases} \\ w &= \begin{cases} -1 & \lambda_2 > 0 \\ \text{undefined} & \lambda_2 = 0 \\ 1 & \lambda_2 < 0 \end{cases} \end{aligned}\right\} \text{ Pontryagin} \end{cases}$$

The solution of the boundary value problem are in the independent variable ζ, so to have the final solution in the time variable t every control/state law must be evaluated in $t = \zeta\, T(\zeta)$.



23. Summary

The OCP problem

Minimize

$$\underbrace{\underbrace{\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p})}_{\text{Mayer}} + \int_a^b \underbrace{L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)}_{\text{Lagrange}}\,\mathrm{d}t}_{\text{Bolza}}$$

Differential constraints

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

Boundary conditions

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}) = \boldsymbol{0}$$

Integral constraints

$$\boldsymbol{\Psi}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\big) + \int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)\,\mathrm{d}t = 0$$

Controls bounds

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Tranformation

Minimize

$$\Phi(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}) + \int_a^b L(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)\,\mathrm{d}t$$

Differential constraints

$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

$$\boldsymbol{z}' = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

Boundary conditions

$$\boldsymbol{b}(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}) = \boldsymbol{0}$$

$$\boldsymbol{z}(a) = \boldsymbol{0}$$

$$\boldsymbol{\Psi}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\big) + \boldsymbol{z}(b) = \boldsymbol{0}$$

Controls bounds

$$\boldsymbol{u}(t) \in \mathscr{U}$$

Functional

$$B\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\big) + \boldsymbol{\eta}\cdot\boldsymbol{z}(b) + \int_a^b \Big(H\big(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},\boldsymbol{p},t\big) - \boldsymbol{\lambda}\cdot\boldsymbol{x}' - \boldsymbol{\mu}\cdot\boldsymbol{z}'\Big)\mathrm{d}t$$

where

$$H\big(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t\big) = L(\boldsymbol{x},\boldsymbol{u},t) + \boldsymbol{\lambda}\cdot\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},t) + \boldsymbol{\mu}\cdot\boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},t)$$

$$B\big(\boldsymbol{x}_a,\boldsymbol{x}_b,\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\big) = \Phi\big(\boldsymbol{x}_a,\boldsymbol{x}_b,\boldsymbol{p}\big) + \boldsymbol{\omega}\cdot\boldsymbol{b}\big(\boldsymbol{x}_a,\boldsymbol{x}_b,\boldsymbol{p}\big) + \boldsymbol{\eta}\cdot\boldsymbol{\Psi}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\big)$$

variation and integration by part of $-\boldsymbol{\lambda}\cdot\delta\boldsymbol{x}'$ and $-\boldsymbol{\mu}\cdot\delta\boldsymbol{z}'$

$$\begin{aligned}\int_a^b -\delta\boldsymbol{\lambda}\cdot\boldsymbol{x}' - \boldsymbol{\lambda}\cdot\delta\boldsymbol{x}'\,\mathrm{d}t &= \int_a^b -\delta\boldsymbol{\lambda}\cdot\boldsymbol{x}' - \left[\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{\lambda}\cdot\delta\boldsymbol{x}) - \boldsymbol{\lambda}'\cdot\delta\boldsymbol{x}\right]\mathrm{d}t \\ &= \int_a^b \Big(-\delta\boldsymbol{\lambda}\cdot\boldsymbol{x}' + \boldsymbol{\lambda}'\cdot\delta\boldsymbol{x}\Big)\mathrm{d}t + \boldsymbol{\lambda}(a)\delta\boldsymbol{x}(a) - \boldsymbol{\lambda}(b)\delta\boldsymbol{x}(b)\end{aligned}$$

$$\begin{aligned}\int_a^b -\delta\boldsymbol{\mu}\cdot\boldsymbol{z}' - \boldsymbol{\mu}\cdot\delta\boldsymbol{z}'\,\mathrm{d}t &= \int_a^b -\delta\boldsymbol{\mu}\cdot\boldsymbol{z}' - \left[\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{\mu}\cdot\delta\boldsymbol{z}) - \boldsymbol{\mu}'\cdot\delta\boldsymbol{z}\right]\mathrm{d}t \\ &= \int_a^b \Big(-\delta\boldsymbol{\mu}\cdot\boldsymbol{z}' + \boldsymbol{\mu}'\cdot\delta\boldsymbol{z}\Big)\mathrm{d}t + \underbrace{\cancel{\boldsymbol{\mu}(a)\delta\boldsymbol{z}(a)}}_{\delta\boldsymbol{z}(a)=\boldsymbol{0}} - \boldsymbol{\mu}(b)\delta\boldsymbol{z}(b)\end{aligned}$$

after variations:

Variation of

$\delta\boldsymbol{x}(t)$, $\delta\boldsymbol{x}(a)$ e $\delta\boldsymbol{x}(b)$

$$\boldsymbol{\lambda}' + \frac{\partial H}{\partial \boldsymbol{x}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t) = \boldsymbol{0}$$

$$\frac{\partial B}{\partial \boldsymbol{x}_a}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\big) + \boldsymbol{\lambda}(a) = \boldsymbol{0}$$

$$\frac{\partial B}{\partial \boldsymbol{x}_b}\big(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\big) - \boldsymbol{\lambda}(b) = \boldsymbol{0}$$

Variation of

$\delta\boldsymbol{z}(t)$ e $\delta\boldsymbol{z}(b)$

$$\left.\begin{aligned}\boldsymbol{\mu}' &= \boldsymbol{0} \\ \boldsymbol{\eta} - \boldsymbol{\mu}(b) &= \boldsymbol{0}\end{aligned}\right\} \quad\Longrightarrow\quad \boldsymbol{\mu}(t) = \boldsymbol{\eta}$$

Variation of

$\delta \boldsymbol{p}$

$$\frac{\partial B}{\partial \boldsymbol{p}}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\right)+\int_a^b \frac{\partial H}{\partial \boldsymbol{p}}\left(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t\right)\mathrm{d}t=\boldsymbol{0}$$

Variation of

$\delta \boldsymbol{\lambda}(t)$ e $\delta \boldsymbol{\mu}(t)$

$$\boldsymbol{x}'=\frac{\partial H}{\partial \boldsymbol{\lambda}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t)=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

$$\boldsymbol{z}'=\frac{\partial H}{\partial \boldsymbol{\mu}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t)=\boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

Variation of

$\delta \boldsymbol{\eta}$ e $\boldsymbol{\omega}$

$$\boldsymbol{\Psi}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\right)+\boldsymbol{z}(b)=\boldsymbol{0}$$

$$\boldsymbol{b}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\right)=\boldsymbol{0}$$

eliminate $z(t)$

$$\boldsymbol{z}(b)=\underbrace{\boldsymbol{z}(a)}_{=\boldsymbol{0}}+\int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)\,\mathrm{d}t$$

thus, BC become

$$\boldsymbol{\Psi}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\right)+\int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)\,\mathrm{d}t=\boldsymbol{0}$$

23.1 The BVP

BVP after variation

The original ODE

$$\boldsymbol{x}'=\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)$$

Adjoint equations

$$\boldsymbol{\lambda}'=-\frac{\partial H}{\partial \boldsymbol{x}}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{u},t)$$

Original Boundary Conditions

$$\boldsymbol{b}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\right)=\boldsymbol{0}$$

Adjoint Boundary Conditions

$$\frac{\partial B}{\partial \boldsymbol{x}_a}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\right)+\boldsymbol{\lambda}(a)=\boldsymbol{0}$$

$$\frac{\partial B}{\partial \boldsymbol{x}_b}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\right)-\boldsymbol{\lambda}(b)=\boldsymbol{0}$$

Integral Constraints

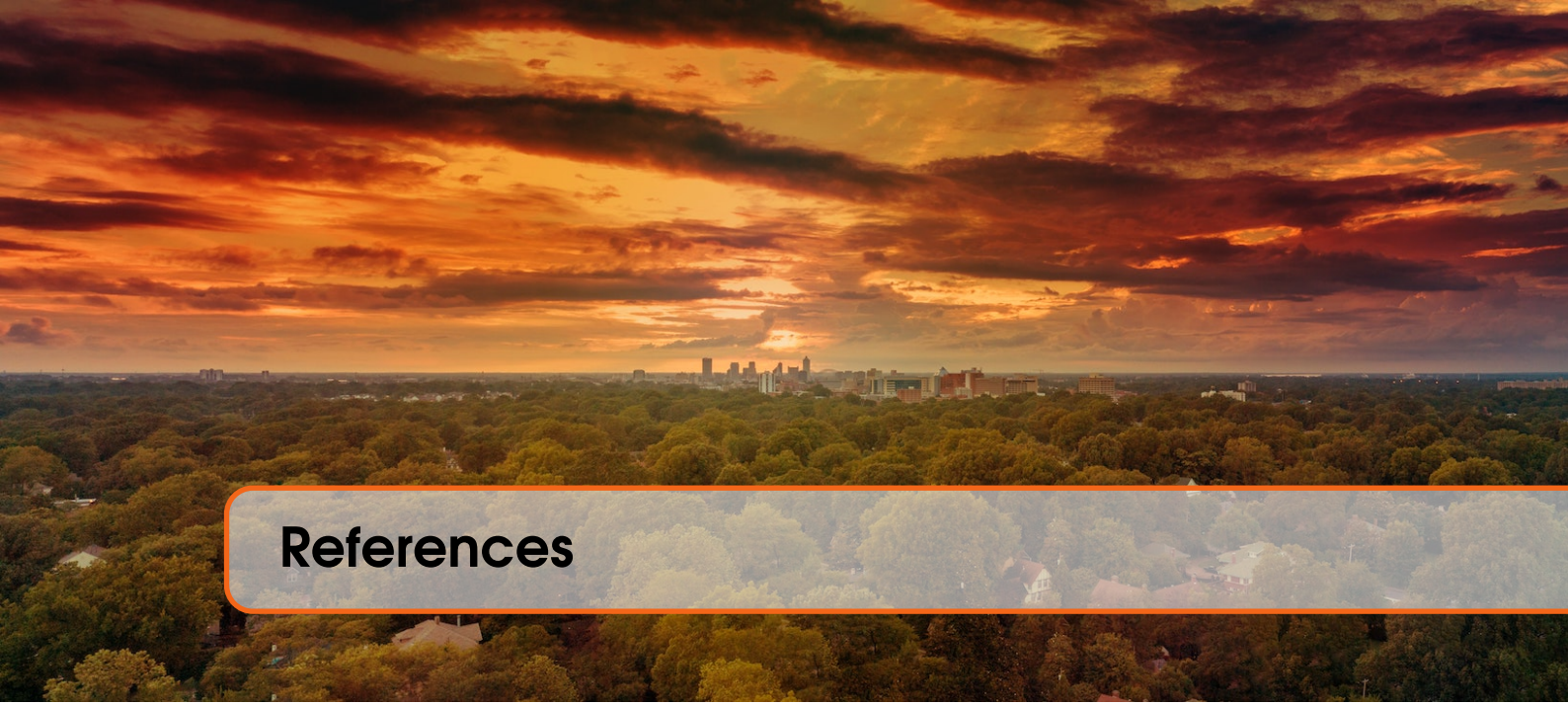
$$\boldsymbol{\Psi}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p}\right)+\int_a^b \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u},\boldsymbol{p},t)\,\mathrm{d}t=\mathbf{0}$$

Optimization Parameters

$$\frac{\partial B}{\partial \boldsymbol{p}}\left(\boldsymbol{x}(a),\boldsymbol{x}(b),\boldsymbol{p},\boldsymbol{\omega},\boldsymbol{\eta}\right)+\int_a^b \frac{\partial H}{\partial \boldsymbol{p}}\left(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{u},t\right)\mathrm{d}t=\mathbf{0}$$

Pontryagin maximum principle (minimum principle)

$$\boldsymbol{u}(t)=\underset{\boldsymbol{v}\in\mathscr{U}}{\arg\min}\,H\left(\boldsymbol{x}(t),\boldsymbol{\lambda}(t),\boldsymbol{\eta},\boldsymbol{v},t\right)$$



References

[1] Michael Athans and Peter L. Falb. 1966. *Optimal control : An introduction to the theory and its applications*. McGraw-Hill, New York, Saint Louis, San Francisco.

[2] John T. Betts. 2010. *Practical methods for optimal control using nonlinear programming* (3rd ed.). Society for Industrial; Applied Mathematics.

[3] Arthur E. Bryson and Yu-Chi Ho. 1975. *Applied optimal control: Optimization, estimation, and control*. Wiley.

[4] F. John. 1948. Extremum Problems with Inequalities as Subsidiary Conditions. In *Studies and essays: Courant anniversary volume*, K. O. Friedrichs, O. E. Neugebauer and J. J. Stoker (eds.). Wiley-Interscience, New York, 187–204.

[5] Donald E. Kirk. 2004. *Optimal control theory: An introduction*. Dover Publications.

[6] H. W. Kuhn and A. W. Tucker. 1951. Nonlinear programming. In *Proceedings of the second berkeley symposium on mathematical statistics and probability*, 1951. University of California Press, Berkeley, California, 481–492.

[7] Daniel Liberzon. 2012. *Calculus of variations and optimal control theory: A concise introduction*. Princeton University Press.

[8] Rein Luus. 2000. *Iterative dynamic programming* (1st ed.). CRC Press, Inc., Boca Raton, FL, USA.

[9] G. McCormick. 1967. Second order conditions for constrained minima. *SIAM Journal on Applied Mathematics* 15, 3 (1967), 641–652. https://doi.org/10.1137/0115056

[10] Jorge Nocedal and Stephen J. Wright. 2006. *Numerical optimization* (2nd ed.). Springer Science & Business Media. https://doi.org/10.1007/978-0-387-40065-5

[11] Hector J. Sussmann and Jan C. Willems. 2001. The brachystrochrone problem and modern control theory. In *Contemporary trends in nonlinear geometric control theory and its, monroy-perez (eds); world scientific publishers*, 2001.