People's Democratic Republic of Algeria
Minister of Higher Education and Scientific Research
Abderrahmane Mira-Bejaia University

Faculty of Exact Sciences
Department of Operations Research

Domain : Mathematics and Informatic
Field : Applied Mathematics
Speciality : Operational Research and Decision Support

THEME

TO OBTAIN A MASTER'S DIPLOMA

Resolution Methods for Norm Minimization in an Optimal Control problem

Realized by:
KHEDDOUCI *Ouarda*

Jury composition :

Chairperson	**KHIMOUM** *Noureddine*	MCB	A.Mira Bejaia University.
Supervisor	**BIBI** *Mohand Ouamer*	Professor	A.Mira Bejaia University.
Examiner	**GHELLAB** *Fouzia*	MCB	A.Mira Bejaia University.
Examiner	**LAOUAR** *Abdelhak*	MAA	A.Mira Bejaia University.

Academic year : 2022/2023

Acknowledgements

I am grateful to Almighty God for granting me the strength and patience needed to successfully complete this humble work.

I would like to express my gratitude to Mr. BIBI Mohand Ouamer for accepting the role of my supervisor and providing valuable guidance throughout the entire process.

I extend my appreciation to the members of the jury: Mr. KHIMOUM Noureddine, whose constant support and recognition of individual effort have uplifted my spirits, Ms. GHELLAB Fouzia and Mr. LAOUAR Abdelhak, they honored me by examining and evaluating my work.

I would like to extend my heartfelt thanks to Mr. ALLICHE Abdennour for his valuable assistance throughout this endeavor.

And may he rest in peace Mr. Allouache Athmane for always being there by my side when needed.

A big thank you to all of my teachers without exception and to everyone who has supported me throughout my journey.

Dedication

I dedicate this work to my parents who have always supported and encouraged me throughout these years of study. I couldn't do anything without your prayers.

To my brothers Smail, Saliha, Wassim , Manel and my other brothers, as well as my beloved uncle Hmimi. Your steadfast support and love have been a source of strength throughout my journey.

To my dear friends Rimouch, Sona, Celina, Nora and Wassila, who have become like a second family to me.

To Kamel, for his support that has been a constant source of strength for me.

Your unwavering friendship, support, and presence in my life have made a profound impact.

Ouarda.

Contents

General Introduction

Optimal control problems play an indispensable role in a multitude of fields, encompassing engineering, economics, and robotics, among others. These problems lie at the heart of designing control strategies that optimize a given objective functional while effectively accounting for the system's dynamic behavior and constraints. They provide a framework for determining the best course of action to achieve desired system performance.

In numerous practical applications, there exists a compelling need to minimize the norm associated with the control input. This norm, typically representing a measure of the control effort, holds great significance as it directly impacts system behavior, performance, and resource utilization. Minimizing the state norm has far reaching benefits, such as enhancing system efficiency, reducing energy consumption, improving stability, and increasing robustness against uncertainties and disturbances. As a result, norm minimization has emerged as a problem of paramount significance in the realm of optimal control.

The central focus of this dissertation lies in the exploration of resolution methods dedicated to norm minimization in optimal control problems. This comprehensive study entails a thorough examination of diverse approaches and algorithms devised to tackle the challenge of norm minimization. The goal is to identify control strategies that effectively minimize the state norm while satisfying the system dynamics and constraints.

The dissertation provides a detailed review of the existing literature on norm minimization techniques, analyzing their applicability within our specific problem domain. It critically evaluates the strengths and limitations of various methods, algorithms, and optimization techniques employed in norm minimization. Special attention is given to the mathematical foundations underlying these techniques, including convex optimization, quadratic programming, and constrained optimization.

To facilitate a deeper understanding of the subject matter, the dissertation is structured as follows:

Chapter 1 serves as an introductory chapter, elucidating fundamental concepts related to the control of dynamic systems. It covers topics such as controllability, stability, and the optimality of control problems.

Chapter 2 offers a comprehensive overview of convex quadratic programming, placing particular emphasis on the Support Method (SM) and its application in the pres-

ence of simple constraints. It provides a solid mathematical foundation for understanding optimization techniques used in norm minimization.

Chapter 3 delves into a comprehensive discussion of a linear quadratic optimal control problem featuring constraints. It explores the formulation of the problem, the design of cost functionals, and the incorporation of constraints. Various methods for solving this class of problems are examined, with a focus on norm minimization.

Finally, the last chapter serves as a culmination of the research, presenting a variety of numerical methods for solving equations and differential equations. It discusses the numerical techniques employed to solve optimal control problems, including direct and indirect methods. Additionally, it engages in an extensive discourse on the methods employed to minimize norms within the context of optimal control. This includes model predictive control, dynamic programming, optimal control parameterization, and iterative algorithms.

The significance of this research lies not only in providing insights into norm minimization in optimal control but also in contributing to the broader body of knowledge in the field. By shedding light on the intricacies of norm minimization and offering a comprehensive exploration of the techniques and algorithms employed, this dissertation serves as a valuable resource for researchers, practitioners, and engineers working in the field of optimal control and related disciplines.

Chapter 1

Introduction to the control of linear dynamic systems

1.1 Introduction

Linear dynamics systems are principle components. It's important to understand how they can be controlled to ensure that they operate as required and to avoid any unintented consequences .

1.2 Controlled system

A controlled system is a differential system which can be influenced by one or more parameters known as control :

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(t_0) = x^0, \tag{1}$$

where $x(t)$ is a state vector in R^n, $u(t)$ is a control which belongs to a set Ω of piecewise continuous functions with values in a compact set U in R^k, x_0 is the inital state at the moment t_0 and $\dot{x}(t) = \frac{dx}{dt}$.

We suppose that the function $f : R^n \times R^k \to R^n$ is enough regular to ensure the existence and uniqueness for the solution of the system (1). To control the system (1), it consists by finding controls $u(t)$ to apply that evolve its state in a desirable way [3].

1.3 Control strategies of a dynamic system

Control strategies are techniques used to manipulate the behavior of a dynamic system in order to achieve a desired outcome. There are two types of control strategies [21] :

Open-loop control: It involves setting predetermined control inputs without regard for the system reply.

Closed-loop control: It involves using feedback from the system to adjust the control inputs in real-time.

There are more advanced control techniques like model predictive control, adaptive control and optimal control. They use mathematical models of the system to forecast its behavior and optimize the control inputs correspondingly.

1.4 Linear approximation of a controlled system

Generally, control systems are nonlinear. Their analysis is not very developed like the linear case which was deeply studied. So we must linearize them around the equilibrium point $(x^e, u^e) \neq 0$, such that $f(x^e, u^e) = 0$. If we set

$$y = x - x^e, \qquad v = u - u^e,$$
$$A = \frac{\partial f}{\partial x}(x, u), \qquad B = \frac{\partial f}{\partial u}(x, u),$$

then we get

$$\dot{y} = Ay + Bv + \circ(\|(y, v)\|),$$

where $\dot{y} = Ay + Bv$ is called the linear approximation of the nonlinear system (1).

1.5 Controllability of dynamic systems

1.5.1 Controllability

The controllability is a leading concept in control theory. In a dynamic system, it determinates whether it's possible to guide in a finite time the system from any initial state to any desired final state using an appropriate control input [21, 29].

1.5.2 Accessible set

We consider the controlled system (1) and we have the following definitions :

Definition 1.5.2

The set of the accessible points from x^0 at a moment $t_1 \geq 0$ is determinated by :

$$Acc(x^0, t_1) = \{x_u(t_1) : u(t) \in U, t \in T = [0, t_1]\},$$

where $x_u(t)$ is the solution of the system (1) with the control $u(t)$. We note :

$$Acc(x^0, 0) = \{x^0\}.$$

Definition 1.5.3

The dynamic system $\dot{x}(t) = f(x(t), u(t))$ is controllable at the moment t_1 if $Acc(x^0, t_1) = R^n$, i.e, for every $x^0, x^1 \in R^n$, there is a control $u \in \Omega$ with values in $U = R^k$, such as the associated trajectory connects x^0 to x^1 at the moment t_1.

1.5.3 Controllability criterion of Kalman

There is an algebraic characterization of a stationary linear system controllability, due to Kalman.

Theorem 1.5.1

The system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad t \geq 0,$$

is controllable if and only if the rank of the controllability matrix of Kalman $K = (B, AB, ..., A^{n-1}B)$ is equal to n, where $x(t) \in R^n$, $u(t) \in R^k$, $A \in R^{n \times n}$ and $B \in R^{n \times k}$.

1.6 Example

If we consider the dynamic system :

$$\begin{cases} \frac{dx_1}{dt} = -x_1 + 2x_2 + u, & x_1(0) = x_1^0 = -3; \\ \frac{dx_2}{dt} = x_1 + x_2, & x_1(0) = x_2^0 = -4, \end{cases}$$

then we have :

$$A = \begin{pmatrix} -1 & 2 \\ 1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad AB = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \Longrightarrow K = (B, AB) = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}.$$

Since $detK = 1 \neq 0$ and $rank(K) = 2$, this system is controllable.

1.7 Linear systems stability

In control theory, stability refers to the behavior of a system over time. A stable system is one that will eventually settle down to a steady state behavior, even after experiencing some initial disturbance. For linear systems, many types of stability are considered [25] :

Bibo stability (Bounded-output): No matter how large the input signal is, the output will never become infinite.

Internal stability: The system will not exhibit runaway behavior or uncontrollable oscillations.

If we consider the system below :

$$\dot{x} = Ax, \tag{2}$$

then it admits $x = 0$ as an equilibrium point, since $f(0, 0) = A \times 0 + B \times 0 = 0$.

1.7.1 Definition

We say that the equilibrium point $x = 0$ of the system (2) is stable, if for every $\varepsilon > 0$, it exists $\eta > 0$, such as for every solution $x(t)$ of (2) , we have

$$\|x(0)\| < \eta \Longrightarrow \|x(t)\| < \varepsilon, \quad \forall t \geq 0. \tag{3}$$

1.7.2 Definition

We say that the equilibrium point $x = 0$ of the system (2) is attractive, if it exists $r > 0$, such as for every solution $x(t)$ of (2) , we have

$$\|x(0)\| < r \Longrightarrow \lim_{t \longrightarrow \infty} x(t) = 0. \tag{4}$$

1.7.3 Remark

If the equilibrium point is not stable, it's called : unstable. Thus, a stable system is more predictable and easier to control, while unstable system can lead to unpredictable behavior.

1.8 Optimal control problem

It's a mathematical optimization problem that involves finding the optimal control strategy that minimizes or maximizes a given performance criterion subject to certain constraints on the control inputs and the state of the system. The general problematic of an optimal control is by considering the dynamic system [23] :

$$\dot{x}(t) = f(x(t), u(t)), \qquad x(0) = x^0, \qquad t \geq 0. \tag{5}$$

We suppose that the admissible controls $u(.)$ belong to Ω which is a set of piecewise continuous functions with the values in a compact set $U \subset R^k$. For every control $u(t) \in U$, the associated trajectory $x(t)$ is defined on the interval $[0, t_1]$, and we define the associated cost :

$$J(u) = \varphi(x(t_1)) + \int_0^{t_1} L(x(t), u(t))dt \tag{6}$$

where $L : R^n \times R^k \to R$ and $\varphi : R^n \to R$.

The problem is to determinate an optimal control $u(t) \in U$, where the corresponding trajectory $x(.)$ is a solution of the system

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x^0,$$

and both minimize the cost $J(u)$.

1.9 The Pontryagin Maximum Principle (PMP)

It's a powerfull tool used in optimal control theory to solve certain types of optimization problems. It was developed by the Russian mathematician Lev Pontryagin in 1956.
If we consider the system (5), then the Hamiltonian function is a function that combines the performance criterion and the constraints into a single expression. So, the Hamiltonian of the system (5) and the cost (6) is [23] :

$$H : R^n \times (R^n \setminus \{0\}) \times R^k \to R$$

$$(x, \quad \psi, \quad u) \longmapsto H(x, \psi, u) =< \psi, f(x, u) > -L(x, u),$$

where $< \psi, f(x, u) >$ is the usual scalar product in R^n.
We can also write

$$H(x(t), \psi(t), u(t)) = \psi^T(t) f(x(t), u(t)) - L(x(t), u(t)), t \in T = [0, t_1]. \tag{7}$$

The vector that corresponds to the Lagrange multipliers vector $\psi : [0, t_1] \to R^n \setminus \{0\}$ is called a conjugate state vector.

1.9.1 General statement of the PMP

We consider the system (5) and we note U the admissible controls set. We define the cost of a control u :

$$J(u) = \varphi(x(t_1)) + \int_0^{t_1} L(x(t), u(t)) dt, \tag{8}$$

where $x(.)$ is the solution trajectory of (5) associated to the control u, $L : R^n \times R^k \to R$ and $\varphi : R^n \to R$.

We consider the problem which is about determining a control $u(t) \in U$ and a trajectory $x(t)$ which minimize the cost (8). If the control $u(t)$ associated to the trajectory $x(t)$ is optimal, then it exists an absolutely continuous application $\psi(.) : [0, t_1] \to R^n$ such as the Hamiltonian achieves its maximum [23]:

$$H(x(t), \psi(t), u(t)) = \max_{v \in U} H(x(t), \psi(t), v), \quad t \in T = [0, t_1],$$

where $H(x(t), \psi(t), u(t)) =< \psi(t), f(x(t), u(t)) > -L(x(t), u(t))$ is the Hamiltonian of the system, with

$$\dot{x} = \frac{\partial H}{\partial \psi}(x(t), \psi(t), u(t), \qquad x(0) = x^0, \tag{9}$$

$$\dot{\psi} = -\frac{\partial H}{\partial x}(x(t), \psi(t), u(t)), \quad \psi(t_1) = -\frac{\partial \varphi(x(t_1))}{\partial x}. \tag{10}$$

1.10 Conclusion

In this chapter, we have introduced the fundamental concepts of the control of linear dynamic systems such as controllability of dynamic systems, their stability and the Pontryagin maximum principle.

Chapter 2

Recalls on convex quadratic programming

2.1 Introduction

Convex quadratic programming plays a main role in solving a wide range of optimization problems, particularly in problems involving quadratic objectives and linear constraints.

2.2 Quadratic forms and their properties

[4]

2.2.1 Quadratic form

A quadratic form in R^n is an homogeneous polynomial of degree two of n variables. It is represented as follows :

$$F(x) = F(x_1, ..., x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j, \tag{1}$$

where $x_1, ..., x_n$ are the n variables, $x = (x_1, ..., x_n)^T$ and $A = (a_{ij}, 1 \leq i, j \leq n)$. So we get

$$F(x) = x^T A x. \tag{2}$$

The matrix A is supposed always symmetric, otherwise we define a new symmetric matrix D such as

$$D = \frac{A + A^T}{2} \Rightarrow D^T = \frac{(A + A^T)^T}{2} = \frac{A^T + A}{2} = D,$$

and we always have

$$F(x) = x^T A x = x^T D x, \quad \forall x \in R^n. \tag{3}$$

In the sequel, we consider that the associated matrix to a quadratic form is always symmetric.

2.2.2 Quadratic form gradient

The gradient of a quadratic form is the vector of its first partial derivatives. If we consider a quadratic form with a symmetric matrix D :

$$F(x) = x^T Dx, \tag{4}$$

and write the matrix D as a column-vectors :

$$D = (d_1, d_2, ..., d_j, \cdots, d_n), \tag{5}$$

then we get

$$F(x) = (x_1, x_2, ..., x_j, ..., x_n) \begin{pmatrix} d_1^T x \\ d_2^T x \\ \vdots \\ d_j^T x \\ \vdots \\ d_n^T x \end{pmatrix} = x_1 d_1^T x + x_2 d_2^T x + ... + x_j d_j^T x + ... + x_n d_n^T x.$$

Thus, the gradient of $F(x)$ is :

$$\nabla F(x) = \begin{pmatrix} \frac{\partial F}{\partial x_1} \\ \frac{\partial F}{\partial x_2} \\ \vdots \\ \frac{\partial F}{\partial x_j} \\ \vdots \\ \frac{\partial F}{\partial x_n} \end{pmatrix}, \tag{6}$$

with :

$$\begin{aligned} \frac{\partial F}{\partial x_1} &= d_1^T x + d_{11} x_1 + d_{12} x_2 + ... + d_{1j} x_j + ... + d_{1n} x_n = 2 d_1^T x, \\ \frac{\partial F}{\partial x_2} &= d_2^T x + d_{21} x_1 + d_{22} x_2 + ... + d_{2j} x_j + ... + d_{2n} x_n = 2 d_2^T x, \\ &\vdots \\ \frac{\partial F}{\partial x_j} &= d_j^T x + d_{j1} x_1 + d_{j2} x_2 + ... + d_{jj} x_j + ... + d_{jn} x_n = 2 d_j^T x, \\ &\vdots \\ \frac{\partial F}{\partial x_n} &= d_n^T x + d_{n1} x_1 + d_{n2} x_2 + ... + d_{nj} x_j + ... + d_{nn} x_n = 2_n^T x, \end{aligned}$$

and hence

$$\nabla F(x) = 2 \begin{pmatrix} d_1^T x \\ d_2^T x \\ \vdots \\ d_j^T x \\ \vdots \\ d_n^T x \end{pmatrix} = 2Dx. \tag{7}$$

2.2.3 Positive definite quadratic form

We say that $F(x)$ is a positive definite quadratic form if

$$x^T Dx > 0, \forall x \in R^n, x \neq 0.$$

Thus D is called a positive definite matrix ($D > 0$).

2.2.4 Positive semi-definite quadratic form

We say that $F(x)$ is positive semi-definite quadratic form if

$$x^T Dx \geq 0, \quad \forall x \in R^n \quad and \; \exists x_0 \neq 0 : x_0^T Dx_0 = 0.$$

Thus, D is called a positive semi-definite matrix ($D \geq 0$).

2.2.5 Matrix minors

We consider the symmetric matrix below :

$$D = \begin{pmatrix} d_{11} & d_{12} & ... & d_{1n} \\ d_{21} & d_{22} & ... & d_{2n} \\ \vdots & \vdots & ... & \vdots \\ d_{n1} & d_{n2} & ... & d_{nn} \end{pmatrix}.$$

The minor of the matrix D, which is formed by rows $i_1, i_2, ..., i_p$ and columns $j_1, j_2, ..., j_p$, is noted as follows :

$$D\begin{pmatrix} i_1, i_2, ..., i_p \\ j_1, j_2, ..., j_p \end{pmatrix} = \begin{vmatrix} d_{i_1 j_1} & ... & d_{i_1 j_p} \\ d_{i_2 j_1} & ... & d_{i_2 j_p} \\ \vdots & ... & \vdots \\ d_{i_p j_1} & ... & d_{i_n j_p} \end{vmatrix}.$$

This minor is called principal if $i_1 = j_1, i_2 = j_2, ..., i_p = j_p$, which means that it's formed by rows and columns wearing the same numbers. The following minors

$$D_1 = d_{11}, \quad D_2 = \begin{vmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{vmatrix}, \quad ... \quad , D_n = \begin{vmatrix} d_{11} & d_{12} & ... & d_{1n} \\ d_{21} & d_{22} & ... & d_{2n} \\ \vdots & \vdots & ... & \vdots \\ d_{n1} & d_{n2} & ... & d_{nn} \end{vmatrix},$$

are called successive principal minors.

2.2.6 Sylvester criterion

Theorem 2.2.1

(i) For a matrix D to be positive definite ($D > 0$), it's necessary and sufficient that the successive principal minors of D are positive :

$$D_1 > 0, \quad D_2 > 0, \quad ..., \quad D_n > 0 \; ;$$

(ii) For a matrix D to be positive semi-definite ($D \geq 0$), it's necessary and sufficient that all the principal minors of D are not negative :

$$D\begin{pmatrix} i_1, i_2, ..., i_p \\ i_1, i_2, ..., i_p \end{pmatrix} \geq 0, \qquad 1 \leq i_1 \leq i_2 \leq ... \leq i_p \leq n, \qquad p = \overline{1, n}.$$

2.3 Convexity

Convexity is a leading concept which is used in optimization theory and its applications.

2.3.1 Convex set

A set S in R^n is called convex if

$$\forall x_1, x_2 \in S, \forall \lambda \in [0,1] \Rightarrow \lambda x_1 + (1-\lambda)x_2 \in S.$$

2.3.2 Convex function

A function f which is called defined on a convex set S in R^n is convex if the following inequality is verified :

$$f(\lambda x+(1-\lambda)y) \leq \lambda f(x)+(1-\lambda)f(y), \quad \forall x, y \in S, \quad \forall \lambda \in [0,1]. \tag{8}$$

2.3.3 Convex quadratic form

Property 2.3.3.1

A quadratic form $F(x) = x^T Dx$ is convex if and only if D is a positive semi-definite matrix ($D \geq 0$).

2.4 Non-linear programming

2.4.1 Minimization without constraints

Let f be a non-linear function that is defined from R^n to R and continuously differentiable. The non-linear programming problem consists of finding $x^* \in R^n$ such as

$$f(x^*) = min f(x), x \in R^n. \tag{9}$$

2.4.2 Local minimum

Let f be a function which is defined on R^n. The function f admits a local minimum x^* if

$$\exists \ B(x^*, \varepsilon) = \{x \in R^n : \|x-x^*\| < \varepsilon\} \ \Rightarrow \ f(x) \geq f(x^*), \quad \forall x \in B(x^*, \varepsilon). \tag{10}$$

2.4.3 Global minimum

The function f admits a global minimum $x^* \in R^n$ if :

$$f(x) \geq f(x^*), \quad \forall x \in R^n. \tag{11}$$

Theorem 2.4.1

If x^* is a local (or a global) minimum of f on R^n and f is differentiable at x^* , then

$$\nabla f(x^*) = 0. \tag{12}$$

A stationary point is a point that verifies the condition (12).

Theorem 2.4.2

If x^* is a local (or global) minimum of f on R^n and f is twice differentiable, then

i) $\nabla f(x^*) = 0$ (stationarity).

ii) $H(x)$ is positive semi-definite, where $H(x)$ is the Hessian matrix of f at the point x^*

$$H(x^*) = \nabla^2 f(x^*). \tag{13}$$

2.4.4 Minimization with constraints

Let f be a non-linear function that is defined from R^n to R and continuously differentiable. The non-linear programming problem consists of finding $x^* \in S \subset R^n, S \neq R^n$, such as

$$f(x^*) = minf(x), x \in S. \tag{14}$$

The constraints' set S is generally given by equations and / or inequations, linear or not :

$$S = \{x \in R^n : g(x) \leq 0\}, \tag{15}$$

where g is a vectorial function which is defined from R^n to R^m and continuously differentiable, with

$$g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{pmatrix},$$

$$\nabla g(x) = (\nabla g_1(x), \nabla g_2(x), ..., \nabla g_m(x)) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1}(x) & \frac{\partial g_2}{\partial x_1}(x) & ... & \frac{\partial g_m}{\partial x_1}(x) \\ \frac{\partial g_1}{\partial x_2}(x) & \frac{\partial g_2}{\partial x_2}(x) & ... & \frac{\partial g_m}{\partial x_2}(x) \\ \vdots & \vdots & ... & \vdots \\ \frac{\partial g_1}{\partial x_n}(x) & \frac{\partial g_2}{\partial x_n}(x) & ... & \frac{\partial g_m}{\partial x_n}(x) \end{pmatrix}.$$

The matrix $\nabla g(x)$ is called the matrix of the gradients and $Jg(x) = [\nabla g(x)]^T$ is the Jacobian of g.

2.4.5 Admissible direction

The vector $d \in R^n, d \neq 0$, is an admissible direction at a point $x \in S$ if it exists a real parameter $\alpha > 0$ such as $x + \theta d \in S, \forall \theta \in [0, \alpha]$.

Lemma 2.4.1

If f is differentiable on S and x^* is its local minimum, then for every admissible direction d, we have

$$d^T \nabla f(x^*) \geq 0. \tag{16}$$

2.5 Support method for minimizing a convex quadratic form with simple constraints

2.5.1 Statement of the problem

We consider the following problem of convex quadratic programming with bounded variables [7, 8] :

$$\min\ F(x) = \frac{1}{2}x^T Dx + c^T x, \tag{1}$$

$$l \le x \le u,$$

where $D = D(J,J)$ is a symmetric matrix ($D^T = D$) and it's positive semi-definite ($D \ge 0$); $c, l, u \in R^n$, $\|l\| < \infty$, $\|u\| < \infty$, $J = \{1, \cdots, j, \cdots, n\}$ and the symbol $()^T$ represents the transposition operation.

Since the admissible set $S = \{x \in R^n : l \le x \le u\}$ is compact, the problem (1) has an optimal solution $x^0 \in S$ according to Weirstrass' theorem :

$$F(x^0) = \min_{x \in S} F(x)\,,$$

where x runs through the set of feasible solutions S. This optimal solution x^0 is unique if D is positive definite ($D > 0$).

2.5.2 Optimality criterion

Let x be a feasible solution of the problem (1), so the vector

$$E(x) = g(x) = \nabla F(x) = Dx + c$$

is called the gradient of F at the point x or the vector of estimations.

Besides x, we consider an other arbitrary feasible solution $\bar{x} = x + \triangle x$. Then the increment of the function F is written as :

$$\begin{aligned}
\triangle F(x) = {} & F(\bar{x}) - F(x) \\
= {} & \tfrac{1}{2}\bar{x}^T D\bar{x} + c^T\bar{x} - \tfrac{1}{2}x^T Dx - c^T x \\
= {} & \tfrac{1}{2}(x + \triangle x)^T D(x + \triangle x) + c^T(x + \triangle x) - \tfrac{1}{2}x^T Dx - c^T x \\
= {} & \tfrac{1}{2}x^T Dx + \tfrac{1}{2}x^T D\triangle x + \tfrac{1}{2}\triangle x^T Dx + \tfrac{1}{2}\triangle x^T D\triangle x \\
& + c^T x + c^T\triangle x - \tfrac{1}{2}x^T Dx - c^T x \\
= {} & x^T D\triangle x + \tfrac{1}{2}\triangle x^T D\triangle x c^T\triangle x.
\end{aligned}$$

If we set

$$\gamma = \frac{1}{2}\triangle x^T D\triangle x \ge 0,$$

then we get

$$\begin{aligned} \triangle F(x) = & F(\bar{x}) - F(x) = (Dx + c)^T \triangle x + \gamma \\ & = E^T(x) \triangle (x) + \gamma, \quad \gamma \geq 0. \end{aligned} \tag{2}$$

So, we have the following optimality criterion :

Theorem (Optimality criterion)

Let x be a feasiable solution of the problem (1). Then the relations

$$\begin{aligned} & E_j(x) \geq 0, \quad if \quad x_j = l_j; \\ & E_j(x) \leq 0, \quad if \quad x_j = u_j; \\ & E_j(x) = 0, \quad if \quad l_j < x_j < u_j, j \in J, \end{aligned} \tag{3}$$

are both necessary and sufficient for the optimality of the vector x.

Proof.Sufficiency

Let x be a feasible solution verifying (3) and we consider an other arbitrary feasible solution $\bar{x}$ of the problem (1). So for $j \in J$, we have :

$$\triangle x_j = \bar{x}_j - x_j \geq l_j - x_j = l_j - l_j = 0, \qquad if \quad x_j = l_j ;$$

$$\triangle x_j = \bar{x}_j - x_j \leq u_j - x_j = u_j - u_j = 0, \quad if \quad x_j = u_j ;$$

According to (3), we get

$$E_j(x) \triangle x_j \geq 0, \ \forall j \in J.$$

Since $\gamma \geq 0$, we deduce from the formula (2) :

$$F(\bar{x}) - F(x) = \sum_{j=1}^{n} E_j(x) \triangle x_j + \gamma \geq 0 ,$$

what provides that x is an optimal solution of the problem (1).

Necessity

We suppose that x is an optimal solution, but the relations (3) are not verified : it exists an index $j_0 \in J$ such as :

$$\begin{aligned} & E_{j_0} < 0 \quad and \quad x_{j_0} = l_{j_0}, \\ & \qquad or \\ & E_{j_0} > 0, \quad and \quad x_{j_0} = u_{j_0}, \\ & \qquad or \\ & E_{j_0} \neq \quad and \quad l_{j_0} < x_{j_0} < u_{j_0}. \end{aligned} \tag{4}$$

In this case, we have always $|E_{j_0}| \neq 0$. So we construct an other feasible solution $\bar{x}$ such as $\bar{x} = x + \theta d$, where $d = (d_1, \cdots, d_{j_0}, \cdots, d_n)$, with $d_{j_0} = -sign E_{j_0}$ and

$d_j = 0, \forall j \in J, j \neq j_0$. So for every case of relations (4), there exists always $\theta > 0$, which is sufficiently small such that the vector $\bar{x}$ is a feasble solution. Then we get from the increment

$$F(\bar{x}) - F(x) = E^T(x) \triangle x + \gamma = \theta \sum_{j=1}^{n} E_j(x) d_j + \frac{1}{2}\theta^2 d^T D d$$

$$= -\theta E_{j_0} sign E_{j_0} + \frac{1}{2}\theta^2 (d_{j_0})^2 d_{j_0 j_0}$$

$$= \theta[-|E_{j_0}| + \frac{1}{2}\theta d_{j_0 j_0}].$$

For $\theta > 0$ sufficiently small, we get $F(\bar{x}) < F(x)$, contradicting thus the optimality of x.

In summary, the relations (3) are necessarily verified if x is an optimal solution.

2.5.3 Suboptimality criterion

For the problem (1), we can define suboptimal solutions or ε-optimal ones, where $\varepsilon \geq 0$ is a certain precision which is already given. Also, x^ε is a suboptimal or an ε-optimal solution if

$$0 \leq F(x^\varepsilon) - F(x^0) \leq \varepsilon,$$

where x^0 is an optimal solution of the problem (1).

We consider the increment (2) after replacing the feasible solution $\bar{x}$ by an optimal solution x^0 :

$$F(x^0) - F(x) = \sum_{j=1}^{n} E_j(x)(x_j^0 - x_j) + \gamma \geq \sum_{j=1}^{n} E_j(x)(x_j^0 - x_j).$$

Then

$$F(x) - F(x^0) \leq \sum_{j=1}^{n} E_j(x)(x_j - x_j^0). \tag{5}$$

Since $l_j \leq x_j^0 \leq u_j, j \in J$, we can write :

$$x_j^0 \geq l_j \;\Rightarrow\; -x_j^0 \leq -l_j \;\Rightarrow\; x_j - x_j^0 \leq x_j - l_j \;;$$

$$x_j^0 \leq u_j \;\Rightarrow\; -x_j^0 \geq -u_j \;\Rightarrow\; x_j - x_j^0 \geq x_j - u_j \;.$$

Then we have

$$E_j(x)(x_j - x_j^0) \leq E_j(x)(x_j - l_j), \;\; if \quad E_j(x) > 0;$$

$$E_j(x)(x_j - x_j^0) \leq E_j(x)(x_j - u_j), \;\; if \quad E_j(x) < 0.$$

So, we deduce from the inequality (5) :

$$0 \leq F(x) - F(x^0) \leq \sum_{E_j > 0, j \in J} E_j(x)(x_j - l_j) \quad + \sum_{E_j < 0, j \in J} E_j(x)(x_j - u_j).$$

The number

$$\beta(x) = \sum_{E_j>0, j\in J} E_j(x_j - l_j) \quad + \sum_{E_j<0, j\in J} E_j(x_j - u_j) \tag{6}$$

is called the suboptimality estimate, because we have always :

$$0 \le F(x) - F(x^0) \le \beta(x).$$

Hence, if $\beta(x) \le \varepsilon$, then x will be a suboptimal or an ε-optimal solution of the problem (1).

2.6 Support of the objective function

The support of the objective function is an element related to the non linearity of F, i.e, to its curvature. For a subset $J_S \subset J$, we say that J_S is a support of the objective function if

$$detD_S = detD(J_S, J_S) \neq 0.$$

The pair $\{x, J_S\}$, formed from the feasible solution x and the support J_S is called a Support Feasible Solution (SFS) of the problem (1). The SFS $\{x, J_S\}$ is called non degenerate if

$$l_j < x_j < u_j, \quad \forall j \in J_S.$$

It's called consistent if

$$E_s = E(J_S) = D(J_S, J)x + c(J_S) = 0\,.$$

For a consistent SFS $\{x, J_S\}$, we can reformulate the optimality criterion of the problem (1). Indeed, for an other arbitrary feasible solution $\bar{x} = x + \Delta x$, the increment holds :

$$F(\bar{x}) - F(x) = \sum_{j\in J} E_j(x)(\bar{x}_j - x_j) + \gamma = \sum_{j\in J_N} E_j(x)(\bar{x}_j - x_j) + \gamma,$$

where $\gamma = \frac{1}{2} \Delta x^T D x \ge 0$ and $J_N = J \setminus J_S$.

So, we have the following theorem :

2.6.1 Theorem (The Support Optimality Criterion)

Let $\{x, J_S\}$ be a consistent SFS of the problem (1). Then these relations

$$\begin{aligned} &E_j(x) \ge 0, \quad if \quad x_j = l_j; \\ &E_j(x) \le 0, \quad if \quad x_j = u_j; \\ &E_j(x) = 0, \quad if \quad l_j < x_j < u_j; j \in J_N, \end{aligned} \tag{7}$$

are sufficient, and they are also necessary for the optimality of x in the case of non degeneracy of the SFS $\{x, J_S\}$.

We note that for a consistent SFS $\{x, J_S\}$, the suboptimality estimate is written as :

$$\beta(x, J_S) = \sum_{E_j>0, j\in J_N} E_j(x_j - l_j) \quad + \sum_{E_j<0, j\in J_N} E_j(x_j - u_j). \tag{8}$$

2.6.2 Remark 1

If we set $D = 0$, then the objective function in the problem (1) becomes linear and it has no curvature. So, we can set $J_S = \emptyset$, with $detD(J_S, J_S) \neq 0$ by convention.

2.6.3 Remark 2

The index $j \in J$ such as $E_j(x) = 0$ is an index of universal optimality, in the sense that it verifies the optimality criterion (7) whatever the value of x_j in the interval $[l_j, u_j]$. We note that such indices are susceptible even to the slightest variation of the feasible solution x. To keep their optimality, J_S recieves them when it is possible. Generally, we chose J_S (eventually empty) in the following set :

$$J_S \subset J_0 = \{j \in J : E_j(x) = 0\},\ with\ detD(J_S, J_S) \neq 0.$$

2.7 Algorithm of resolution

Let $\{x, J_S\}$ be a consistent SFS of the problem (1), such as $\beta(x, J_S) > \varepsilon$, where $\varepsilon \geq 0$ is an already chosen accuracy. The purpose of the algorithm is to construct an ε-optimal solution x^ε or a completely optimal solution x^0. It passes from a consistent SFS $\{x, J_S\}$ to an other consistent SFS $\{\bar{x}, \bar{J}\}$ such as $F(\bar{x}) \leq F(x)$. For this, we form a new feasible solution $\bar{x} = x + \theta d$, where d is a direction of improvement and θ is the maximal step along this direction. Indeed, let J_{NNO} be the subset of indices of J_N, which do not verify the optimality criterion (7) :

$$J_{NNO} = \{j \in J_N : x_j = l_j, E_j < 0;\quad x_j = u_j, E_j > 0;\quad l_j < x_j < u_j, E_j \neq 0\}.$$

After finding the index j_0 such as :

$$|Ej_0| = \max_{j \in J_{NNO}} |Ej|,$$

then we form the direction $d = d(J) = (d(J_S), d(J_N)) = (d_S, d_N)$ as follows :

$$d_{j_0} = -signE_{j_0},\quad d_j = 0,\quad \forall j \in J_N, j \neq j_0.$$

The sub-vector $d_S = d(J_S)$ will be calculated such that

$$E_j(\bar{x}) = E_j(x) = 0,\quad \forall j \in J_S.$$

Then we get

$$E_S(\bar{x}) = D(J_S, J)\bar{x} + c_S = D_S(J_S, J)(x + \theta d) + c_s.$$
$$E_S(\bar{x}) = E_S + \theta D(J_S, J_S)d_S + \theta D(J_S, J_N)d_N = 0$$

So

$$d_S = -D_S^{-1}(J_S, J_S)D(J_S, J_N)d_N$$
$$= D_S^{-1}D(J_S, j_0)signE_{j_0}.$$

Therefore the direction d is written as :

$$\begin{cases} d_{j_0} = -signE_{j_0},\quad d_j = 0,\quad j \neq j_0,\quad j \in J_N, \\ d_s = +D_s^{-1}D(J_s, j_0)signE_{j_0}. \end{cases}$$

As for the step θ, it must firstly verify the relation :

$$l \le x + \theta d \le u \quad \Leftrightarrow \begin{cases} l_{j_0} - x_{j_0} \le \theta d_{j_0} \le u_{j_0} - x_{j_0}, \\ l_j - x_j \le \theta d_j \le u_j - x_j; \quad j \in J_S. \end{cases}$$

For the index j_0, we have

$$l_{j_0} - x_{j_0} \le -\theta_{j_0} sign E_{j_0} \le u_{j_0} - x_{j_0} \quad \Rightarrow \theta_{j_0} = \begin{cases} x_{j_0} - l_{j_0} > 0, \ if \quad E_{j_0} > 0, \\ u_{j_0} - x_{j_0} > 0, \ if \quad E_{j_0} < 0. \end{cases}$$

For the indices $j \in J_S$ of the relations (10), the maximal step will be :

$$\theta_s = \theta_{j_s} = \min_{j \in J_s} \theta_j \ , \ with \ \ \theta_j = \begin{cases} (u_j - x_j)/d_j, \ if \quad d_j > 0 \ ; \\ (l_j - x_j)/d_j, \ if \quad d_j < 0 \ ; \\ +\infty \qquad , \ if \quad d_j = 0 \ . \end{cases}$$

Also, the maximal step for which $\bar{x} = x + \theta d$ is a feasible solution is :

$$\theta = \min\{\theta_{j_0}, \theta_{j_s}\}. \tag{11}$$

For a quadratic function which is not linear, we must calculate the relaxation step θ_F which minimizes the function $F(x + \theta d) = \varphi(\theta)$ along the direction d, it is equivalent to have :

$$\varphi'(\theta_F) = \nabla F^T(x + \theta_F d)d = E^T d + \theta_F d^T D d = -|E_{j_0}| + \theta_F \alpha = 0,$$

where

$$\alpha = d^T D d = (d_S, d_N)^T D \begin{pmatrix} d_S \\ d_N \end{pmatrix},$$

$$\alpha = (d_S, d_N)^T [D(J, J_S)d_S + D(J, J_N)d_N],$$

$$\alpha = d_S^T D(J_S, J_S)d_S + d_S^T D(J_S, J_N)d_N + d_N^T D(J_N, J_S)d_S + d_N^T D(J_N, J_N)d_N.$$

Using the relations (9), we find that :

$$\alpha = D(j_0, J_S)D_S^{-1} D_S D_S^{-1} D(J_S, j_0) + 2(-sign E_{j_0})D(j_0, J_S)D_S^{-1} D(J_S, j_0)(sign E_{j_0}) + d_{j_0 j_0}$$

So

$$\alpha = d_{j_0 j_0} - D(j_0, J_s)D_S^{-1} D(J_S, j_0). \tag{12}$$

We deduce that

$$\theta_F = \begin{cases} \frac{|E_{j_0}|}{\alpha} \ , \quad if \ \alpha \neq 0; \\ \infty \quad , \ if \quad \alpha = 0. \end{cases} \tag{13}$$

In summary, the new feasible solution $\bar{x}$ will be formed with the maximal step θ such as :

$$\theta^0 = \min\{\theta_{j_0}, \theta_{j_s}, \theta_F\}, \ \ \bar{x} = x + \theta^0 d. \tag{14}$$

As for the support $\bar{J}_S$, it will be formed according to the specefic value of θ^0 :

$$\theta^0 = \theta_{j_0} \vee \theta_{j_s} \vee \theta_F.$$

We observe these three possible cases :

(i) $\theta^0 = \theta_{j_0}$: In this case, the element $\bar{x}_{j_0}$ becomes critical ($\bar{x}_{j_0} = l_{j_0} \vee u_{j_0}$). The index j_0 becomes an optimal index. So it's not necessary to change the support, then we set $\bar{J}_S := J_S$;
The SFS $\{\bar{x}, \bar{J}_S\} = \{\bar{x}, J_S\}$ is a consistent SFS.

(ii) $\theta^0 = \theta_{j_s}$: In this case, the element $\bar{x}_{j_s}$ becomes critical ($\bar{x}_{j_s} = l_{j_s} \vee u_{j_s}$). The index j_s will be picked up from the support for not disturbing the algorithm's progression.
So we set $\bar{J}_S := J_S \backslash j_s$, and it's clear that

$$E_j(\bar{x}) = 0, \quad j \in \bar{J}_S, \ and \ detD(\bar{J}_S, \bar{J}_S) \neq 0.$$

The pair $\{\bar{x}, \bar{J}_S\}$ is a consistent SFS.

(iii) $\theta = \theta_F$: In this case, the index j_0 becomes optimal because :

$$\begin{cases} E_{j_0}(\bar{x}) = E_{j_0}(x) + \theta_F D(j_0, J)d = E_{j_0} - \theta_F \alpha \ \ signE_{j_0} \\ \\ E_{j_0}(\bar{x}) = E_{j_0} - \frac{|E_{j_0}|}{\alpha}\alpha \ \ signE_{j_0} = E_{j_0} - E_{j_0} = 0. \end{cases}$$

Since $\alpha > 0$ (θ_F is the minimum), the matrix $D(J_S \cup j_0, J_S \cup j_0)$ is inversible.
So we can include j_0 in the new support. Therefore, we set $\bar{J}_S := J_S \cup j_0$.
Then the new SFS $\{\bar{x}, \bar{J}_S\}$ is consistent.

2.8 The convergence of the algorithm

Generally, the algorithm described above is not finite. But there is a rule that can hold it to be finite. It's the property of non critical index in the non optimal set J_{NNO}. Indeed, if we consider the set of the non critical and non optimal indices :

$$J_{NC} = \{j \in J_{NNO} : l_j < x_j < u_j\},$$

then the algorithm becomes finite if we set $j_0 = \min_{j \in J_{NC}} j$.

According to this rule, the improvement of the support feasible solution $\{x, J_s\}$ is done firstly by the non basic and non critical elements.

2.9 Example

We solve the following problem of convex quadratic programming with bounded variables :

$$F(x) = 4x_1^2 + x_2^2 - 4x_1x_2 - x_3 + 6x_4 \;\rightarrow\; \min$$
$$0 \le x_1 \le 6\ ;$$
$$0 \le x_2 \le 2\ ;$$
$$-1 \le x_3 \le 3\ ;$$
$$-3 \le x_4 \le 1\ .$$

Since the function F is convex, it admits a global minimum on R^4 at x if and only if :

$$\frac{\partial F}{\partial x} = \nabla F(x) = 0 \quad \Leftrightarrow \quad \begin{cases} \frac{\partial F}{\partial x_1} = 8x_1 - 4x_2 = 0, \\ \frac{\partial F}{\partial x_2} = -4x_1 + 2x_2 = 0, \\ \\ \frac{\partial F}{\partial x_3} = -1 \neq 0, \\ \frac{\partial F}{\partial x_4} = 6 \neq 0, \end{cases}$$

Since $\frac{\partial F}{\partial x} \neq 0, \forall x \in R^4$, then F does not admit a finite global minimum over R^4. But, F admits a global minimum on the feasible compact set

$$S = \{x = (x_1, x_2, x_3, x_4) \in R^4 : 0 \le x_1 \le 6;\ \ 0 \le x_2 \le 2;\ \ -1 \le x_3 \le 3;\ \ -3 \le x_4 \le 1\}.$$

The standard form of the problem is :

$$F(x) = \frac{1}{2}(8x_1^2 + 2x_2^2 - 8x_1x_2) - x_3 + 6x_4 \;\rightarrow\; \min,$$
$$0 \le x_1 \le 6\ ;$$
$$0 \le x_2 \le 2\ ;$$
$$-1 \le x_3 \le 3\ ;$$
$$-3 \le x_4 \le 1\ ;$$

with

$$D = \begin{pmatrix} 8 & -4 & 0 & 0 \\ -4 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad c = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 6 \end{pmatrix}, \quad l = \begin{pmatrix} 0 \\ 0 \\ -1 \\ -3 \end{pmatrix} and \quad u = \begin{pmatrix} 6 \\ 2 \\ 3 \\ 1 \end{pmatrix}.$$

We set $\varepsilon = 0$ and we start with the initial SFS $\{x, J_S\}$, where $x = (6, 2, 0, 0)$, $J_S = \phi$, and we compute F and its gradient :

$$F(x) = 100, \qquad g(x) = \frac{\partial F}{\partial x} = \begin{pmatrix} 8x_1 - 4x_2 \\ -4x_1 + 2x_2 \\ -1 \\ 6 \end{pmatrix} = \begin{pmatrix} 40 \\ -20 \\ -1 \\ 6 \end{pmatrix}.$$

We have

$$E_1 = g_1 = 40 > 0; \quad E_2 = g_3 = -20 < 0; \quad E_3 = g_3 = -1 < 0; \quad E_4 = g_4 = 6 > 0;$$

$$\begin{aligned}\beta(x,J_s) &= E_1(x_1-l_1)+E_2(x_2-u_2)+E_3(x_3-u_3)+E_4(x_4-l_4)\\ &= 40(6-0)\ +(-20)(2-2)+(-1)(0-3)+6(0+3)\\ &= 240+0+3+18=261>\varepsilon.\end{aligned}$$

The indices $\{1,3,4\}$ are not optimal, so

$$J_{NNO}=\{1,3,4\}\ \Rightarrow\ |E_{j_0}|=\max\{|E_1|,|E_3|,|E_4|\}=|E_1|=40\ \Rightarrow\ j_0=1.$$

Therefore, we set :

$$d_{j_0}=d_1=-sign\,E_1=-1,\quad d_2=d_3=d_4=0,\ (because\ \ d_{14}=0),$$

$$\theta_{j_0}=\theta_1=x_1-l_1=6,\qquad \theta_{j_s}=+\infty.$$

We calculate the relaxation step that minimizes the function $F(x+\theta d)$:

$$\alpha=d_{j_0j_0}-D(j_0,J_s)D_s^{-1}D(J_s,j_0)$$

$$\alpha=d_{11}-D(1,J_s)D_s^{-1}D(J_s,1)=d_{11}=8\ (because\ J_s=\phi)\ \Rightarrow\ \theta_F=\frac{40}{8}=5.$$

The new feasible solution $\bar{x}$ will be formed with the maximal step θ^0 such as :

$$\theta^0=\min\{\theta_{j_0},\theta_{j_s},\theta_F\}=\min\{6,+\infty,5\}=5=\theta_F.$$

So we get

$$\bar{x}=x+\theta^0 d=\begin{pmatrix}6\\2\\0\\0\end{pmatrix}+5\begin{pmatrix}-1\\0\\0\\0\end{pmatrix}=\begin{pmatrix}1\\2\\0\\0\end{pmatrix},\qquad F(\bar{x})=0.$$

We start the second iteration with the new SFS $\{x,J_S\}$, with

$$x:=\bar{x}=(1,2,0,0),\quad J_S:=J_S\cup j_0=\{1\}\ and\ \ J_N=\{2,3,4\}.$$

So we have :

$$g(x)=\frac{\partial F}{\partial x}=\begin{pmatrix}0\\0\\-1\\6\end{pmatrix},\quad E_1=g_1=0;\quad E_2=g_2=0;\quad E_3=g_3=-1<0;\quad E_4=g_4=6>0.$$

The indices $\{3,4\}$ are not optimal, so :

$$J_{NNO}=\{3,4\}\ \Rightarrow\ |E_{j_0}|=max\{|E_3|,|E_4|\}=|E_4|=6\ \Rightarrow\ j_0=4.$$

Therefore, we set :

$$d_{j_0}=d_4=-sign\,E_4=-1,\quad d_2=d_3=0,$$

$$d_{j_s}=D_S^{-1}D\,(J_S,j_0)sign\,E_{j_0},$$

$$d_1=D_S^{-1}D\,(1,4)sign\,E_4=\tfrac{1}{d_{11}}d_{14}=0.$$

So

$$\theta_{j_0}=\theta_4=x_4-l_4=3,\qquad \theta_{j_s}=+\infty\ (because\ \ \ d_1=0).$$

We calculate the relaxation step that minimizes the function $F(x+\theta^0 d)$:

$$\alpha = d_{j_0 j_0} - D(j_0, J_s) D_s^{-1} D(J_s, j_0)$$

$$\alpha = d_{44} - D(4,1) D_s^{-1} D(1,4) = d_{44} - d_{41} \frac{1}{d_{11}} d_{14} = 0 \quad \Rightarrow \quad \theta_F = +\infty.$$

The new feasible solution $\bar{x}$ will be formed with the maximal step θ^0 such as :

$$\theta^0 = \min\{\theta_{j_0}, \theta_{j_s}, \theta_F\} = min\{3, +\infty, +\infty\} = 3 = \theta_{j_0}.$$

We get

$$\bar{x} = x + \theta^0 d = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 0 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -3 \end{pmatrix}, \quad F(x) = -18.$$

We start the third iteration with the new SFS $\{x, J_S\}$, with

$$x := \bar{x} = (1, 2, 0, -3), \quad J_S = \{1\} \; and \; J_N = \{2, 3, 4\}.$$

So we have :

$$g(x) = \frac{\partial F}{\partial x} = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 6 \end{pmatrix}, \quad E_1 = g_1 = 0; \quad E_2 = g_2 = 0; \quad E_3 = g_3 = -1 < 0; \quad E_4 = g_4 = 6 > 0.$$

The index $\{3\}$ is not optimal, so :

$$J_{NNO} = \{3\} \;\Rightarrow\; |E_{j_0}| = |E_3| = 1 \;\Rightarrow\; j_0 = 3.$$

Therefore, we set :

$$d_{j_0} = d_3 = -sign\, E_3 = +1, \quad d_2 = d_4 = 0,$$
$$d_{j_s} = D_s^{-1} D\,(J_s, j_0) sign\, E_{j_0},$$
$$d_1 = D_s^{-1} D\,(1,3) sign\, E_3 = \tfrac{1}{d_{11}} d_{13}(-1) = 0 \; (because \; d_{13} = 0).$$

$$\theta_{j_0} = \theta_3 = \frac{u_3 - x_3}{d_3} = 3, \qquad \theta_{j_s} = +\infty \; (because \; d_1 = 0).$$

We calculate the relaxation step that minimizes the function $F(x+\theta^0 d)$:

$$\alpha = d_{j_0 j_0} - D(j_0, J_s) D_s^{-1} D(J_s, j_0)$$

$$\alpha = d_{33} - D(3,1) D_s^{-1} D(1,3) = d_{33} - d_{31} \frac{1}{d_{11}} d_{13} = 0 \;\Rightarrow\; \theta_F = +\infty.$$

The new feasible solution $\bar{x}$ will be formed with the maximal step θ^0 such as :

$$\theta^0 = min\{\theta_{j_0}, \theta_{j_s}, \theta_F\} = \min\{3, +\infty, +\infty\} = 3 = \theta_3.$$

We get

$$\bar{x} = x + \theta^0 d = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -3 \end{pmatrix} + 3 \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ -3 \end{pmatrix}, \quad F(x) = -21.$$

We start the fourth iteration with the new SFS $\{x, J_S\}$, with

$$x := \bar{x} = (1, 2, 3, -3), \quad J_S = \{1\} \; and \; J_N = \{2, 3, 4\}.$$

So we have :

$$g(x) = \frac{\partial F}{\partial x} = \begin{pmatrix} 0 \\ 0 \\ -1 \\ 6 \end{pmatrix}, \quad E_1 = g_1 = 0; \quad E_2 = g_2 = 0; \quad E_3 = g_3 = -1 < 0; \quad E_4 = g_4 = 6 > 0;$$

In this case, all the indices are optimal,so the algorithm is finished .

In summary, $x = (1, 2, 3, -3)$ is the optimal solution, with $F(x) = -21$.

2.10 Conclusion

In this chapter, we have provided an introduction to convex optimization and its significance in the context of quadratic programming. We have also explored the fundamental properties of convex sets and functions, which enable efficient and reliable optimization algorithms.

Chapter 3

Linear quadratic optimal control problem with constraints

3.1 Introduction

Understanding the role of constraints is crucial for developing control strategies that satisfy operational limitations and achieve desired objectives.

3.2 Problem Statement

Let be a stationary linear dynamic system defined on the interval $T = [0, t_1]$:

$$\dot{x}(t) = Ax(t) + Bu(t) + r(t), \quad x(0) = x^0, \quad t \in T,$$

where $\dot{x}(t) = \dfrac{dx}{dt}$, $x(t) \in R^n$ is the state vector at the moment t, x^0 is an initial state, $u = u(t) \in R^k$ is a control, $r(t) \in R^n$ is a disturbance function (it can be considered null). The matrices A, B are of n and $(n \times k)$ order respectively.

For the above dynamic system, various controls $u(t)$ generate different trajectories $x(t)$ and processes $(u(t), x(t)), t \in T$. We can affect the choice of $u(t)$ to get a desired trajectory by optimizing a certain predefined quality criterion. So we obtain an optimal process $(u^0(t), x^0(t)), t \in T$, in the sens of this criterion.

We consider the following optimal control problem :

$$J(u) = \frac{\alpha}{2}\|x(t_1)\|^2 + c^T x(t_1) + K \to \min, \tag{1}$$

$$\dot{x} = Ax + Bu + r, \qquad x(0) = x^0, \tag{2}$$

$$u(t) \in U = [-L, L]^k, \qquad t \in T = [0, t_1], \tag{3}$$

where $c = (c_1, ..., c_n)^T \in R^n$, $x = (x_1, ..., x_n)^T \in R^n$, $u = (u_1, ..., u_k)^T \in R^k$, $K, \alpha, L \in R$, $\alpha > 0$ and $L > 0$.

The control $u(t)$ is an admissible control of the problem $(1) - (3)$ if :

- Its values are in U which is a convex and compact set in R^k;

- It's piecewise continuous on T, which means that the function $u(t)$ admits a finite number s of discontinuity points τ_i of the first kind, where the limits on the right and on the left exist :

$$\lim_{t\to\tau_i+0} u(t) = u(\tau_i+0), \qquad \lim_{t\to\tau_i-0} u(t) = u(\tau_i-0), i = \overline{1,s}.$$

Also, we consider that the control $u(t)$ is continuous on the right at the point τ_i :

$$u(\tau_i) = u(\tau_i+0), \quad i = \overline{1,s} \qquad and \qquad u(t_1-0) = u(t_1).$$

We say that an admissible control $u^0(t), t \in T$, is optimal if it minimizes the quality criterion (1) on the set of admissible controls. Thus, the corresponding trajectory $x^0(t)$ is called an optimal trajectory [5, 10].

3.3 Increment of the functional

Let $u(t)$ be an admissible control of the problem (1) – (3) and $x(t)$ its corresponding trajectory. We consider an other arbitrary admissible control

$$\bar{u}(t) = u(t) + \triangle u(t),$$

and its corresponding trajectory $\bar{x}(t) = x(t) + \triangle x(t), t \in T$.
We have

$$\triangle x(t) = \bar{x}(t) - x(t), \quad \triangle x(0) = \bar{x}(0) - x(0) = x^0 - x^0 = 0.$$

So, the functional's increment is :

$$\begin{aligned}
\triangle J(u) &= J(\bar{u}) - J(u) \\
&= \tfrac{\alpha}{2}\bar{x}^T(t_1)\bar{x}(t_1) + c^T\bar{x}(t_1) + K - \tfrac{\alpha}{2}x^T(t_1)x(t_1) - c^Tx(t_1) - K \\
&= \tfrac{\alpha}{2}(x+\triangle x)^T(t_1)(x+\triangle x)(t_1) + c^T(x+\triangle x)(t_1) - \tfrac{\alpha}{2}x^T(t_1)x(t_1) - c^Tx(t_1) \\
&= \tfrac{\alpha}{2}x^T(t_1)x(t_1) + \tfrac{\alpha}{2}x^T(t_1)\triangle x(t_1) + \tfrac{\alpha}{2}\triangle x^T(t_1)x(t_1) + \tfrac{\alpha}{2}\triangle x^T(t_1)\triangle x(t_1) \\
&\quad + c^Tx(t_1) + c^T\triangle x(t_1) - \tfrac{\alpha}{2}x^T(t_1)x(t_1) - c^Tx(t_1) \\
&= \alpha x^T(t_1)\triangle x(t_1) + \tfrac{\alpha}{2}\triangle x^T(t_1)\triangle x(t_1) + c^T\triangle x(t_1).
\end{aligned}$$

We set

$$\gamma = \frac{\alpha}{2}\|\triangle x(t_1)\|^2 \geq 0,$$

and we get

$$\triangle J(u) = J(\bar{u}) - J(u) = (\alpha x(t_1) + c)^T\triangle x(t_1) + \gamma, \qquad \gamma \geq 0. \tag{4}$$

Using the Cauchy's formula, the solution of the system (2) is :

$$x(t) = e^{At}x^0 + \int_0^t e^{A(t-\tau)}[Bu(\tau) + r(\tau)]d\tau, \quad t \in T.$$

So

$$x(t_1) = e^{At_1}x^0 + \int_0^{t_1} e^{A(t_1-t)}[Bu(t) + r(t)]dt,$$

and

$$\bar{x}(t_1) = e^{At_1}x^0 + \int_0^{t_1} e^{A(t_1-t)}[B\bar{u}(t) + r(t)]dt.$$

Then we have :

$$\triangle x(t_1) = \bar{x}(t_1) - x(t_1) = \int_0^{t_1} e^{A(t_1-t)}[B\bar{u}(t) - Bu(t)]dt = \int_0^{t_1} e^{A(t_1-t)}B \triangle u(t)dt,$$

and we deduce from (4) :

$$\triangle J(u) = J(\bar{u}) - J(u) = \int_0^{t_1} (\alpha x(t_1) + c)^T e^{A(t_1-t)}B \triangle u(t)dt + \gamma. \tag{5}$$

We define the following function ψ as :

$$\psi(t) = -e^{A^T(t_1-t)}(\alpha x(t_1) + c), t \in T. \tag{6}$$

The function ψ satisfies the following differential equation :

$$\begin{aligned} \dot{\psi}(t) &= A^T e^{A^T(t_1-t)}(\alpha x(t_1) + c), \\ \dot{\psi}(t) &= -A^T\psi(t), \qquad \psi(t_1) = -(\alpha x(t_1) + c). \end{aligned} \tag{7}$$

The system (7) is called the conjugate system of the problem $(1) - (3)$.
We consider the Hamiltonian of the system $(1) - (3)$:

$$H(x, \psi, u) = \psi^T(Ax + Bu + r), \tag{8}$$

and we set

$$\triangle_{\bar{u}} H(x, \psi, u) = H(x, \psi, \bar{u}) - H(x, \psi, u). \tag{9}$$

Then we have

$$\triangle_{\bar{u}} H(x, \psi, u) = \psi^T B \triangle u. \tag{10}$$

From (5) and (6), we get

$$\triangle J(u) = -\int_0^{t_1} \psi^T(t)B \triangle u(t)dt + \gamma,$$

Hence, from (10) we get

$$\triangle J(u) = J(\bar{u}) - J(u) = -\int_0^{t_1} \triangle_{\bar{u}} H(x(t), \psi(t), u(t))dt. \tag{11}$$

The expression of the Hamiltonian allows to deduce the primal dynamic system (2) and its conjugate one (7) as follows :

$$\dot{x} = \frac{\partial H}{\partial \psi}, \qquad x(0) = x^0, \tag{12}$$

$$\dot{\psi} = -\frac{\partial H}{\partial x}, \qquad \psi(t_1) = -(\alpha x(t_1) + c). \tag{13}$$

3.4 Pontryagin Maximum Principle (PMP)

3.4.1 Theorem 3.1

An admissible control $u^0(t), t \in T$, is optimal in the problem $(1)-(3)$ if and only if along $u^0(t)$ and the corresponding trajectories $x^0(t), \psi^0(t), t \in T$, of the direct system (2) and the conjugate one (7), the Hamiltonian achieves its maximum [23] :

$$H(x^0(t), \psi^0(t), u^0(t)) = \max_{v \in U} H(x^0(t), \psi^0(t), v), \quad t \in T = [0, t_1]. \tag{14}$$

3.4.2 Resolution of the boundary value problem, derived from the (PMP)

Thus, in order to find the optimal control u^0, we will have to solve a differential system of $2n$ equations with $(2n+1)$ unknowns (u^0, x^0, ψ^0) at two boundaries : $t = t_0 = 0$ and $t = t_1$, where we suppose that the disturbance function is null : $r(t) = 0, t \in T$, and $k = 1$, with $B = b \in R^n$ [1].

From the equation (14), if we can express the extremal control u as a function of x and ψ, i.e, $u(t) = u(t, x, \psi)$, then we get a two boundary value problem of $2n$ equations with $2n$ unknowns :

$$\begin{cases} \dot{x}(t) = Ax(t) + bu(t, x, \psi), & x(0) = x^0; \\ \\ \dot{\psi}(t) = -A^T \psi(t), & \psi(t_1) = -(\alpha x(t_1) + c). \end{cases}$$

This problem is difficult to solve, since it is not a Cauchy's problem.To avoid this difficulty, we set the following Cauchy's problem :

$$\begin{cases} \dot{x}(t) = Ax(t) + bu(t, x, \psi), & x(0) = x^0; \\ \\ \dot{\psi}(t) = -A^T \psi(t), & \psi(0) = \psi^0, \end{cases}$$

where ψ^0 is a parameter representing the initial condition of ψ at $t = 0$; it must be found such as the trajectories $\psi(t, \psi^0)$ and $x(t, \psi^0)$ verify the final condition :

$$\psi(t_1, \psi^0) = -(\alpha x(t_1, \psi^0) + c),$$

i.e,

$$\psi(t_1, \psi^0) + \alpha x(t_1, \psi^0) + c = 0.$$

In order to solve this equation, we set $\psi^0 = s$ as a free parameter and we define the shooting function :

$$F(s) = \psi(t_1, s) + \alpha x(t_1, s) + c, \tag{15}$$

where

$$s = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_j \\ \vdots \\ s_n \end{pmatrix}, \qquad F(s) = \begin{pmatrix} f_1(s) \\ f_2(s) \\ \vdots \\ f_i(s) \\ \vdots \\ f_n(s) \end{pmatrix}, \qquad JF(s) = \left(\frac{\partial f_i(s)}{\partial s_j}, \quad 1 \le i, j \le n \right).$$

Consequently, the optimal solution $u^0(t)$ of the problem $(1)-(3)$ is deduced from (14), with $\psi^0 = s^*$ such as :

$$F(s^*) = 0. \tag{16}$$

We apply the Newton's method to solve (16). For this, if s^k is an approximation of order k, then the approximation of order $(k+1)$ is written as follows :

$$s^{k+1} = s^k - JF^{-1}(s^k)F(s^k), \quad k = 0, 1, 2, ... \tag{17}$$

During the application of the shooting method, we will write with more details the shooting function $F(s)$(15) and the formula (17).

3.4.3 Linear case

If we set $K = 0$, $\alpha = 0$ and $r = 0$, then we get the linear case :

$$\begin{cases} J(u) = c^T x(t_1) \to min, \\ \dot{x} = Ax + bu, \qquad x(0) = x^0, \\ -L \le u(t) \le L, \qquad t \in T = [0, t_1], \end{cases}$$

and the following boundary value problem :

$$\begin{cases} \dot{x} = Ax + bu, \qquad x(0) = x^0, \\ \dot{\psi} = -A^T\psi, \qquad \psi(t_1) = -c. \end{cases}$$

Example

We consider the following optimal control problem :

$$\begin{aligned} J(u) &= 2x_1(t_1) - 2x_2(t_1) \to min, \\ \dot{x_1} &= u, \qquad x_1(0) = 3; \\ \dot{x_2} &= x_1, \qquad x_2(0) = -1; \\ -2 \le u(t) &\le 2, \qquad t \in T = [0, t_1], \ t_1 = 5. \end{aligned} \tag{19}$$

This problem is easy to solve directly without using the shooting method. Indeed, we start by finding the Hamiltonian :

$$H(x, \psi, u) = \psi_1\dot{x_1} + \psi_2\dot{x_2} = \psi_1 u + \psi_2 x_1 = \psi_2 x_1 + \psi_1 u.$$

The conjugate system is expressed as follows :

$$\begin{cases} \dot{\psi_1} = -\dfrac{\partial H}{\partial x_1} = -\psi_2, \qquad \psi_1(t_1) = -c_1 = -2, \\ \dot{\psi_2} = -\dfrac{\partial H}{\partial x_2} = 0, \qquad \psi_2(t_1) = -c_2 = 2 \Rightarrow \psi_2(t) = 2, \ t \in T. \end{cases}$$

So

$$\dot{\psi_1}(t) = -\psi_2(t) = -2, \;\; \psi_1(5) = -2 \;\; \Rightarrow \;\; \psi_1(t) = -2t + 8, \;\; t \in T = [0,5].$$

The optimal control is equal to :

$$u^0(t) = 2sign\triangle(t), \; t \in T = [0,5],$$

where $\triangle(t) = \psi_1(t), \; t \in T$, is the switching function. Hence, we have :

$$u^0(t) = 2sign(-2t+8) = \begin{cases} +2, \;\; if \;\; t \in [0,4[; \\ \\ -2, \;\; if \;\; t \in [4,5]. \end{cases}$$

To calculate $J(u^0)$, we must solve the primal system (19) setting $u = u^0$. Indeed on the interval $[0,4]$, we have :

$$\dot{x_1}(t) = u^0 = 2, \quad x_1(0) = 3 \quad \Rightarrow \quad x_1(t) = 2t + 3, \;\; with \;\; x_1(4) = 11,$$

and on the interval $[4,5]$, we have :

$$\dot{x_1}(t) = u^0 = -2, \quad x_1(4) = 11 \quad \Rightarrow \quad x_1(t) = -2t + 19; \;\; with \;\; x_1(5) = 9.$$

On the interval $[0,4]$, we get for the second equation :

$$\dot{x_2}(t) = x_1(t) = 2t + 3; \quad x_2(0) = -1 \quad \Rightarrow \quad x_2(t) = t^2 + 3t - 1, \;\; with \;\; x_2(4) = 27,$$

and on the interval $[4,5]$, we obtain :

$$\dot{x_2}(t) = x_1(t) = -2t + 19; \quad x_2(4) = 27 \quad \Rightarrow \quad x_2(t) = -t^2 + 19t - 33, \;\; with \;\; x_2(5) = 37.$$

So

$$J(u^0) = 2x_1(t_1) - 2x_2(t_1) = 2 \times 9 - 2 \times 37 = -56.$$

3.4.4 Weakly linear case

If we set $K = 0$, $\alpha > 0$ sufficiently small and $c \in R^n$, then we get the weakly linear case :

$$J(u) = \frac{\alpha}{2}\|x(t_1)\|^2 + c^T x(t_1) \rightarrow min,$$

with

$$\psi(t_1) = -\alpha x(t_1) - c, \quad 0 < \alpha << 1.$$

Example

We consider the following optimal control problem :

$$\begin{aligned} J(u) &= \frac{\alpha}{2}\|x(t_1)\|^2 + 2x_1(t_1) - 2x_2(t_1) \rightarrow min, \\ \dot{x_1} &= u, \qquad x_1(0) = 3; \\ \dot{x_2} &= x_1, \qquad x_2(0) = -1; \\ -2 &\leq u(t) \leq 2, \qquad t \in T = [0, t_1], \ \ t_1 = 5, \end{aligned} \tag{20}$$

where α is a small real positif parameter.

Here, to solve this problem, we use the solution of the problem (19). For this, we start by finding the Hamiltonian :

$$H(x, \psi, u) = \psi^T(Ax + bu) = \psi_1 \dot{x_1} + \psi_2 \dot{x_2} = \psi_1 u + \psi_2 x_1 = \psi_2 x_1 + \psi_1 u.$$

The conjugate system is expressed as follows :

$$\begin{cases} \dot{\psi_1} = -\dfrac{\partial H}{\partial x_1} = -\psi_2, & \psi_1(t_1) = -\alpha x_1(t_1) - 2, \ \ t \in T; \\ \dot{\psi_2} = -\dfrac{\partial H}{\partial x_2} = 0, & \psi_2(t_1) = -\alpha x_2(t_1) + 2, \ \ t \in T. \end{cases}$$

The optimal control is equal to :

$$u^0(t) = 2sign\triangle(t, \alpha), \ \ t \in T = [0, 5],$$

where $\triangle(t, \alpha) = \psi_1(t, \alpha)$ is the switching function. From the conjugate system, we deduce :

$$\begin{cases} \psi_2(t) = -\alpha x_2(t_1) + 2; \ \ t \in T; \\ \dot{\psi_1}(t) = -\psi_2(t) = \alpha x_2(t_1) - 2, \ \ t \in T, \qquad \psi_1(t_1) = -\alpha x_1(t_1) - 2. \end{cases}$$

So we get

$$\begin{cases} \psi_1(t) = (\alpha x_2(t_1) - 2)t + k_1, \\ \psi_1(t_1) = 5(\alpha x_2(t_1) - 2) + k_1 = -\alpha x(t_1) - 2, \end{cases}$$

with

$$k_1 = -\alpha x_1(t_1) - 2 - 5\alpha x_2(t_1) + 10 = -\alpha x_1(5) - 5\alpha x_2(5) + 8.$$

We replace t_1 by its value :

$$\psi_1(t) = (\alpha x_2(5) - 2)t - \alpha x_1(5) - 5\alpha x_2(5) + 8.$$

Let θ_α be the zero of the switching function :

$$\psi_1(\theta_\alpha) = 0 \Leftrightarrow (\alpha x_2(5) - 2)\theta_\alpha - \alpha x_1(5) - 5\alpha x_2(5) + 8 = 0,$$

i.e,

$$(\alpha x_2(5)-2)\theta_\alpha = \alpha x_1(5) + 5\alpha x_2(5) - 8,$$
$$\theta_\alpha = \frac{\alpha x_1(5) + 5\alpha x_2(5) - 8}{\alpha x_2(5) - 2}.$$

When α tends to zero, then $\theta_0 = 4$, and for $\alpha > 0$ sufficiently small, then we get :

$$u^0(t,\alpha) = 2sign(-2\theta_\alpha + 8) = \begin{cases} +2, \ if \ t \in [0, \theta_\alpha[; \\ \\ -2, \ if \ t \in [\theta_\alpha, 5], \end{cases}$$

where $\theta_\alpha \in [4-\varepsilon, 4+\varepsilon]$.

To calculate $J(u^0)$, we must solve the primal system (20) setting $u = u^0$.
On the interval $t \in [0, \theta_\alpha[$, for x_1 we have :

$$\dot{x_1}(t,\alpha) = u^0(t,\alpha) = +2; \ \ x_1(0) = 3 \ \ \Rightarrow \ \ x_1(t,\alpha) = 2t + 3; \ \ x_1(\theta_\alpha) = 2\theta_\alpha + 3.$$

On the interval $t \in [\theta_\alpha, 5]$, we write :

$$\dot{x_1}(t,\alpha) = u^0(t,\alpha) = -2; \ \ x_1(\theta_\alpha) = 2\theta_\alpha + 3 \ \ \Rightarrow \ \ x_1(t,\alpha) = -2t + 4\theta_\alpha + 3, \ with \ x_1(5,\alpha) = 4\theta_\alpha - 7.$$

On the interval $t \in [0, \theta_\alpha[$, for x_2 we have :

$$\dot{x_2}(t,\alpha) = x_1(t,\alpha) = 2t + 3; \ \ x_2(0) = -1 \ \ \Rightarrow \ \ x_2(t,\alpha) = t^2 + 3t - 1, \ with \ x_2(\theta_\alpha) = \theta_\alpha^2 + 3\theta_\alpha - 1.$$

On the interval $t \in [\theta_\alpha, 5]$, we write :

$$\dot{x_2}(t,\alpha) = x_1(t,\alpha) = -2t + 4\theta_\alpha + 3; \ \ x_2(\theta_\alpha) = \theta_\alpha^2 + 3\theta_\alpha - 1 \ \ \Rightarrow$$
$$x_2(t,\alpha) = -t^2 + 4\theta_\alpha t + 3t - 1 - 2\theta_\alpha^2, \ with \ x_2(5,\alpha) = -2\theta_\alpha^2 + 20\theta_\alpha - 11.$$

Then :

$$J(u^0,\alpha) = \frac{\alpha}{2}(4\theta_\alpha - 7)^2 + \frac{\alpha}{2}(-2\theta_\alpha^2 + 20\theta_\alpha - 11)^2 + 2(4\theta_\alpha - 7) - 2(-2\theta_\alpha^2 + 20\theta_\alpha - 11)$$

When α tends to zero, then $\theta_0 = 4$ and we find the minimum value of the linear case :

$$\lim_{\alpha \longrightarrow 0} J(u^0,\alpha) = -56.$$

3.5 Conclusion

In this chapter, we have focused on formulating the constrained linear quadratic optimal control problem and then solving it for the linear and weakly linear cases.

Chapter 4

Methods for norm minimization in an optimal control problem

4.1 Introduction

This final chapter focuses on our main objective : norm minimization in the context of optimal control. We will review existing methods and algorithms that aim to minimize norms associated with control inputs. Through a numerical example, we will illustrate the simple shooting method.

4.2 Problem statement

Let (P) be the following optimal control problem :

$$J(u) = \frac{1}{2}\alpha\|x(t_*))\|^2 + a^T x(t_*) \;\; \rightarrow \;\; min, \tag{1}$$

$$\dot{x} = Ax(t) + bu(t), \qquad x(0) = x^0, \tag{2}$$

$$-L \leq u(t) \leq L, \qquad t \in T = [0, t_*], \tag{3}$$

where $\dot{x} = \frac{dx}{dt}$; $x(t)$ is the vector of state at the moment t ; $x(0) = x^0$ is the initial state; $u(t), t \in T$, is a scalar control, taken from the set of piecewise continuous functions, and it has a finite number of discontinuity points of first kind. The matrix A is square of order n; $b \in R^n$; $\alpha, L, t_* \in R$, with $\alpha > 0$, $L > 0$ and $t_* > 0$ (fixed terminal time).

We note that the problem (P) has an optimal solution in the set of piecewise constant admissible controls, and this, without searching the existence of solutions in the space of measurable functions.

The resolution methods of the problem (P) can be classified into two types : Direct and Indirect methods [28].

4.2.1 Direct methods

The direct methods for solving an optimal control problem are based on total or partial discretization, which transforms it into a nonlinear optimization problem of finite dimension. The resolution of this latter provides an approximate solution for the

original problem (P).

Among direct methods of resolution of the problem (P), we can cite :

- Method of total discretization;
- Method of partial discretization, where we treat the problem (P) in the class of impulsive controls, without discretizing the linear differential system (2).

The two methods yield an approximate solution for the problem (P) after solving a quadratic programming problem in finite dimension.

Advantages

- Simple to use and to understand.
- Computational time is fast.
- The user does not have to be concerned with adjoint variables or switching structures.
- Work well for small scale problems.

Disdvantages

- Producing less accurate solutions than indirect methods.
- The discretized optimal control problem has sometimes several minima, if the fonctional $J(u)$ is not convex. Applying the direct methods often ends up in one of these pseudominima. This solution, however, can be quite a step away from the true solution satisfying all the necessary conditions from variational calculus resulting.
- Increasing the dimension of the finite dimensional space does not necessarily yield better values for the extremely complicated problems.

4.2.2 Indirect methods

The indirect methods for solving an optimal control problem are based on Pontryagin Maximum Principle. They consist of reducing the problem to a two-boundary value problem, then solving it by shooting method, we get the controls analytically or numerically using for example Newton's method.

Advantages

- All kinds of constraints are allowed and very accurate results can be obtained.
- Handle a larger class of optimal control problems.
- Derive analytical solutions and provide explicit feedback laws.

Disdvantages

- Computationally expensive is the numerical integration of differential equations.
- May not always converge to the solution because of the sensitivity to the initial conditions.
- Difficult to implement for complex problems with nonlinear dynamics and constraints.

4.2.3 Support method or hybrid method

Support method is a resolution algorithm which combines simultanuously the direct and indirect methods. It solves the problem (P) in the admissible set of piecewise continuous functions, and without discretization of the linear dynamic system (2). First, it solves a quadratic programming problem in finite dimension in order to improve the current admissible control u, getting thus a new admissible control $\bar{u}$ such that $J(\bar{u}) \leq J(u)$. Then it applies a finishing procedure, close to the shooting method, and that, in order to calculate an exact solution with a necessary accuracy. For this, the finishing procedure uses the Newton's method, known for its fast quadratic convergence [27] .

Before presenting the two first methods, we recall a numerical method for solving equations, called Newton's method.

4.3 Newton's method

Newton's method (1686) is a numerical method used to find the roots of a given function or non-linear equation. It's based on the idea of approximating the root of a function by finding the tangent line to the function at a given point and then finding where the tangent line intersects the x-axis. This process is repeated iteratively until a satisfactory approximation of the root is obtained. The general structure of Newton's method is as follows :

For scalar functions

We consider $f \in C^1(I)$, with $I = [a, b] \subset R$ and $f'(\alpha) \neq 0$ (α is a simple root of f).

1. We choose an initial guess for the root, denoted by x_0. For this, we take in $I = [a, b]$ two points x_1 and x_2 such as :

$$f(x_1) \times f(x_2) < 0, \quad x_0 \in [x_1, x_2];$$

2. We evaluate the function and its derivative at the initial guess $f(x_0)$ and $f'(x_0)$;
3. Compute the next approximation of the root :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

4. Repeat steps (2) and (3) with x_1 as the initial guess until convergence is achieved, i.e, $f(x)$ is sufficiently small or the absolute difference between two consecutive approximations is less than a prescribed tolerance level.

Newton's method algorithm :

We have x_0 is an initial guess of the root, x_n is the n^{th} approximation of the root , N is the maximum number of iterations that we can do and ε is a chosen accuracy.

1. Compute :
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \forall n \geq 0;$$
2. If : $|\frac{x_{n+1}-x_n}{x_{n+1}}| < \varepsilon$, then
   - Convergence achieved.
   - Write the root x_{n+1}.
   - Stop.
   - Else ;
3. If the maximum number of iterations N is achieved :
   - Convergence not achieved in N iterations;
   - Stop.

For vectorial functions

Let F be a function defined on R^n with values in R^n :
$$F = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$
If we consider that the equation
$$F(x) = 0$$
has at least a solution x^* and the Jacobian matrix $JF(x^*)$ is an inversible matrix, then the continuity of JF ensures the invertibility of $JF(x_k)$ for all x_k nearby x^* and the exixtence of x_{k+1} at the second step of the algorithm.

Let us solve the following system :
$$[JF(x_k)]\delta_k = F(x_k).$$
Then we set :
$$x_{k+1} = x_k - \delta_k$$
1.Initialization : $k = 0$: choice of x_0 nearby x^*.

2.Iteration k **:** $x_{k+1} = x_k - [JF(x_k)]^{-1}F(x_k)$;

3.Stopping criterion : If $\|x_{k+1} - x_k\| < \varepsilon$, stop.
Else we set $k = k + 1$ and we return to 2.

4.4 Method of total Discretization in optimal control

The discretization method in an optimal control problem refers to the process of approximating a continuous time optimal control problem by discretizing the state and the control, then converting the problem into a finite dimensional optimization problem. Here, the smaller the step of the discretization, the closer we get the original problem (P). In this case, if we want to obtain diserable approximate solutions, it's clear that the problems to solve, are necessarily of large size [11, 12].

Also, we choose a subdivision on the interval of time $T = [0, t_*]$, i.e, a set of isolated moments τ_j such as :

$$0 = \tau_0 < \tau_1 < \ldots < \tau_j < \ldots < \tau_N = t_*, \quad T_N = \{\tau_j, j = \overline{0, N}\},$$
$$\tau_{j+1} - \tau_j = h = const, \qquad j \in J = \{0, 1, ..., N-1\},$$

where $h = \frac{t_*}{N} > 0$ is the step of subdivision.

On the basis of this subdivision, we replace the derivative in the problem (P) by a finite difference and the integral by a sum, so we consider that the control takes a constant value on every interval of the subdivision :

$$u(t) = u_j, \quad t \in [jh, (j+1)h[, \quad j \in J = \{0, 1, ..., N-1\}. \tag{4}$$

Besides, if we replace the derivative in the system (2) by a simple formula of numerical derivation, by occuring Euler's formula, then we get

$$\dot{x}(t) \simeq \frac{x(t+h) - x(t)}{h}.$$

The dynamic system (2) is then approximated as follows :

$$\frac{x(t+h) - x(t)}{h} = \dot{x}(t) = Ax(t) + bu(t),$$

i.e,

$$x(t+h) = x(t) + h[Ax(t) + bu(t)], \quad t \in T = [0, t_*].$$

At the points of the subdivision $t = \tau_j, \ \ j \in J$, we will have :

$$x(\tau_j + h) = x(\tau_j) + h[Ax(\tau_j) + bu(\tau_j)], \quad j \in J = \{0, 1, ..., N-1\}.$$

We set $x(\tau_j) = x^j$, $u(\tau_j) = u_j$, and we get the following reccurent relation :

$$x^{j+1} = x^j + h[Ax^j + bu_j], \quad j \in J = \{0, 1, ..., N-1\}. \tag{5}$$

By using the last formula , we write the final state $x(t_*)$, which depends only on the variables $u_0, u_1, ..., u_{N-1}$. Then, we deduce $J(u)$, and we get the following problem of convex quadratic programming in R^N :

$$J(u^N) = \frac{1}{2}\alpha \|x(t_*)\|^2 + a^T x(t_*) \ \rightarrow \ min,$$
$$-L \leq u_j \leq L, \quad j \in J = \{0, 1, ..., N-1\},$$

whith $u^N = (u_0, u_1, ..., u_{N-1})^T \in R^N$ and $x(t_*) = x(\tau_N) = x^N$.

To solve this problem of finite dimension, we use one of the many algorithms that treat the problem of convex quadratic programming with simple constraints. Here, we can apply the Support Method which is described above.

4.5 Method of Partial Discretization with impulsive controls

For the problem (P), we also consider on the interval $T = [0, t_*]$ the subdivision [12,13] :

$$0 = \tau_0 < \tau_1 < \ldots < \tau_j < \ldots < \tau_N = t_*, \qquad \tau_{j+1} - \tau_j = h = \frac{t_*}{N}, \quad j \in J,$$

where the moments τ_j are called the points of the discretization.

A piecewise constant function $u = (u(t), t \in T)$ is called an impulsive control, if it changes its value only at the points of discretization τ_j, $j \in J$, i.e,

$$u(t) = u_j = const, \quad t \in [\tau_j, \tau_{j+1}[, \quad j \in J.$$

The set of the impulsive controls is a subset of the class of piecewise constant functions. So, the minimization will be done on this reduced subspace. But, the dynamic system (2) will not be discretized in this case. Since it's linear, its solution, under the initial condition $x(s)$, $s \geq 0$, is formulated by using the following formula of Cauchy :

$$x(t) = e^{A(t-s)}x(s) + \int_s^t e^{A(t-\tau)}bu(\tau)d\tau, \qquad t \geq s. \tag{7}$$

If the states of the system are only measured at the moments τ_j, $j \in J$, then under the impulsive control $u(t) = u_j$ on the interval $[\tau_j, \tau_{j+1}[$, the formula (7) yields for the consecutive states $x(\tau_j) = x^j$ and $x(\tau_{j+1}) = x^{j+1}$ the following reccurent relation :

$$x(\tau_{j+1}) = e^{Ah}x(\tau_j) + \int_{\tau_j}^{\tau_{j+1}} e^{A(\tau_{j+1}-\tau)}bu_j d\tau.$$

In an other way, that means that

$$x^{j+1} = A(h)x^j + b(h)u_j,$$

where

$$A(h) = e^{Ah}, \qquad b(h) = \int_{\tau_j}^{\tau_j+h} e^{A(\tau_j+h-\tau)}bd\tau,$$

and with change of the variable $t = \tau - \tau_j$, the term $b(h)$ does not depend ultimately on the index $j \in J$:

$$b(h) = \int_{\tau_j}^{\tau_j+h} e^{A(\tau_j+h-\tau)}bd\tau = \int_0^h e^{A(h-t)}bdt.$$

Thus, for the successive states, we get the following reccurent relation :

$$x^{j+1} = A(h)x^j + b(h)u_j, \quad j \in J = \{0, 1, ..., N-1\}. \tag{8}$$

Besides, the continuous dynamic system (2), moving under an impulsive control, behaves as the discret system (8). The obtained discretized problem (P') is written as follows :

$$F(u^N) = \frac{1}{2}\alpha\|x(\tau_N)\|^2 + a^T x(\tau_N) = \frac{1}{2}\alpha\|x^N\|^2 + a^T x^N \;\rightarrow\; min, \tag{9}$$

$$x^{j+1} = A(h)x^j + b(h)u_j, \quad x(0) = x(\tau_0) = x^0, \tag{10}$$

$$-L \le u_j \le L, \qquad j \in J = \{0, 1, ..., N-1\}. \tag{11}$$

The constraint (10) can be deleted, because, in this problem, it can only be used to express the terminal state $x(t_*) = x(\tau_N) = x^N$ as a function of the variables $u^N = (u_0, u_1, ..., u_{N-1})^T$. Indeed, we have

$x^1 = A(h)x^0 + b(h)u_0;$

$x^2 = A(h)x^1 + b(h)u_1 = A(h)[A(h)x^0 + b(h)u_0] + b(h)u_1$

$x^2 = A^2(h)x^0 + A(h)b(h)u_0 + b(h)u_1;$

$x^3 = A(h)x^2 + b(h)u_2 = A(h)[A^2(h)x^0 + A(h)b(h)u_0 + b(h)u_1] + b(h)u_2$

$x^3 = A^3(h)x^0 + A^2(h)b(h)u_0 + A(h)b(h)u_1 + b(h)u_2;$

then we get

$x^N = A^N(h)x^0 + A^{N-1}(h)b(h)u_0 + A^{N-2}(h)b(h)u_1 + ... + A(h)b(h)u_{N-2} + b(h)u_{N-1};$

i.e,

$x^N = A^N(h)x^0 + \sum_{j\in J} A^{N-j-1}(h)b(h)u_j.$

If we set

$$q_j = A^{N-j-1}(h)b(h) \;\Rightarrow\; x^N = A^N(h)x^0 + \sum_{j\in J} q_j u_j,$$

then we have

$$a^T x^N = a^T A^N(h)x^0 + \sum_{j\in J} a^T q_j u_j. \tag{12}$$

Therefore

$\frac{1}{2}\ \|x^N\|^2 = \frac{1}{2}\ \|A^N(h)x^0 + \sum_{j\in J} q_j u_j\|^2,$

$= \frac{1}{2}\ < A^N(h)x^0 + \sum_{j\in J} q_j u_j,\ A^N(h)x^0 + \sum_{j\in J} q_j u_j >,$

$= \frac{1}{2}\ \|A^N(h)x^0\|^2 + (A^N(h)x^0)^T \sum_{j\in J} q_j u_j + \frac{1}{2}\ \|\sum_{j\in J} q_j u_j\|^2,$

where

$$\frac{1}{2}\,\|\sum_{j\in J} q_j u_j\|^2 = \frac{1}{2} < \sum_{i\in J} q_i u_i,\ \sum_{j\in J} q_j u_j > = \frac{1}{2}\sum_{i\in J}\sum_{j\in J} q_i^T q_j u_i u_j.$$

If we set

$$(Q = q_j, \quad j \in J),$$

So we have

$$\frac{1}{2}\,\|\sum_{j\in J} q_j u_j\|^2 = \frac{1}{2} < Qu^N, Qu^N > = \frac{1}{2}(u^N)^T Q^T Q u^N = \frac{1}{2}\,\|Qu^N\|^2 \ge 0.$$

We also set

$$p_j = q_j^T A^N(h)x^0, \quad j \in J.$$

Then we get

$$\frac{1}{2}\|x^N\|^2 = \frac{1}{2}(u^N)^T Q^T Q u^N + \sum_{j\in J} p_j u_j + \frac{1}{2}\|A^N(h)x^0\|^2. \tag{13}$$

Using the formula (12) and (13), the objective function (9) is written as follows :

$$F(u^N) = \frac{1}{2}\alpha\|x^N\|^2 + a^T x^N = \frac{1}{2}\alpha(u^N)^T Q^T Q u^N + \sum_{j\in J}\alpha p_j u_j$$

$$+\frac{1}{2}\alpha\|A^N(h)x^0\|^2 + \sum_{j\in J} a^T q_j u_j + a^T A^N(h)x^0,$$

$$F(u^N) = \frac{1}{2}\alpha\,(u^N)^T Q^T Q u^N + \sum_{j\in J}(a^T q_j + \alpha p_j)u_j + \frac{1}{2}\alpha\,\|A^N(h)x^0\|^2 + a^T A^N(h)x^0,$$

$$F(u^N) = \tfrac{1}{2}\alpha\,(u^N)^T Q^T Q u^N + \textstyle\sum_{j\in J}(a + \alpha A^N(h)x^0)^T q_j u_j$$

$$+\frac{1}{2}\alpha\,\|A^N(h)x^0\|^2 + a^T A^N(h)x^0.$$

If we set

$$D = Q^T Q \geq 0,\ c^N = (c_j, j \in J),\ \ with\ \ c_j = (a + \alpha A^N(h)x^0)^T q_j,$$

$$K = \frac{1}{2}\alpha\|A^N(h)x^0\|^2 + a^T A^N(h)x^0,$$

then the discretized problem (P') is written as follows :

$$F(u^N) = \frac{1}{2}\alpha(u^N)^T D u^N + (c^N)^T u^N + K \ \to \min,$$

$$-L \leq u_j \leq L, \qquad j \in J = \{0, 1, ..., N-1\}. \tag{14}$$

4.6 Simple shooting method

The simple shooting method is one of the indirect methods used to solve ordinary differential equations (ODEs) by converting them into boundary value problems (BVPs), using initial values that satisfy the boundary conditions. It's based on Pontryagin maximum principle. It makes the problem more tractable and allows for a simpler implementation, but may require a larger number of intervals to achieve a desired level of accuracy[16, 27, 29].

The general procedure of the shooting method :

1. Guess a set of control inputs that are practicable and sufficiently close to the optimal solution.

2. Integrate the differential equations forward in time from the initial state using the guessed control inputs.

3. Check whether the final state achieved by the integration matches the desired final state.

4. Adjust the guessed control inputs and repeat steps 2 – 3 until a satisfactory solution is found.

For illustration, we consider the following optimal control problem :

$$J(u) = \frac{\alpha}{2}\|x(t_1)\|^2 + c^T x(t_1) + K \rightarrow \min, \tag{1}$$

$$\dot{x} = Ax + bu + r, \qquad x(0) = x^0, \tag{2}$$

$$u(t) \in U = [-L, L], \quad t \in T = [0, t_1]. \tag{3}$$

The optimal control $u(t)$ verifies :

$$H(x(t), \psi(t), u(t)) = \max_{v \in U} H(x(t), \psi(t), v),$$

where

$$\dot{\psi}(t) = -A^T \psi(t), \qquad \psi(t_1) = -(\alpha x(t_1) + c), \tag{4}$$

and

$$H(x(t), \psi(t), u(t)) = \psi^T(t)[Ax(t) + Bu(t) + r(t)].$$

Hence, in order to find the optimal control u^0, we will have to solve a differential system of $2n$ equations with $(2n+1)$ unknowns (u^0, x^0, ψ^0) at two boundaries : $t = t_0 = 0$ and $t = t_1$, where we suppose that the disturbance function is null : $r(t) = 0, t \in T$.

From the equation (4), if we can express the extremal control u as a function of x and ψ, i.e, $u(t) = u(t, x, \psi)$, then we get a two boundary value problem of $2n$ equations with $2n$ unknowns. This problem is difficult to solve, since it is not a Cauchy's problem.To avoid this difficulty, we set the following Cauchy's problem :

$$\begin{cases} \dot{x}(t) = Ax(t) + bu(t, x, \psi), & x(0) = x^0; \\ \dot{\psi}(t) = -A^T \psi(t), & \psi(0) = \psi^0. \end{cases}$$

We have

$$\psi(t) = e^{-A^T t}\psi^0,$$

Hence, we have

$$\dot{\psi}(t) = -A^T e^{-A^T t}\psi^0 = -A^T \psi(t), \quad t \in T,$$

where ψ^0 is a parameter representing the initial condition of ψ at $t = 0$; it must be found such as the trajectories $\psi(t, \psi^0)$ and $x(t, \psi^0)$ verify the final condition :

$$\psi(t_1, \psi^0) = -(\alpha x(t_1, \psi^0) + c),$$

i.e,

$$\psi(t_1, \psi^0) + \alpha x(t_1, \psi^0) + c = 0.$$

The Cauchy's formula allows to write :

$$x(t_1) = e^{At_1}x^0 + \int_0^{t_1} e^{A(t_1-t)}bu(t)dt.$$

According to the Maximum Principle, we get :

$$u(t) = L\ sign\Delta(t) = L\ sign\psi^T(t,\psi^0)b = L\delta(t,\psi^0),$$

with $\delta(t,\psi^0) = sign\psi^T(t,\psi^0)b$,
and we deduce that

$$x(t_1) = x(t_1,\psi^0) = e^{At_1}x^0 + L\int_0^{t_1} e^{A(t_1-t)}b\delta(t,\psi^0)dt.$$

Thus, we must have :

$$\psi(t_1,\psi^0) = e^{-A^T t_1}\psi^0 = -(\alpha x(t_1,\psi^0) + c).$$

In order to solve this equation, we set $\psi^0 = s$ as a free parameter, so we write :

$$x(t_1,s) = e^{At_1}x^0 + L\int_0^{t_1} e^{A(t_1-t)}b\delta(t,s)dt,$$

$$\psi(t_1,s) = e^{-A^T t_1}s = -(\alpha x(t_1,s) + c).$$

We define the shooting function :

$$F(s) = \psi(t_1,s) + \alpha x(t_1,s) + c, \tag{5}$$

where

$$s = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_j \\ \vdots \\ s_n \end{pmatrix}, \qquad F(s) = \begin{pmatrix} f_1(s) \\ f_2(s) \\ \vdots \\ f_i(s) \\ \vdots \\ f_n(s) \end{pmatrix}, \qquad JF(s) = \left(\frac{\partial f_i(s)}{\partial s_j}, \quad 1 \le i,j \le n \right).$$

By replacing $\psi(t_1),s)$ and $x(t_1),s)$ by their values, we obtain :

$$F(s) = e^{-A^T t_1}s + \alpha e^{At_1}x^0 + \alpha L\int_0^{t_1} e^{A(t_1-t)}b\delta(t,s)dt.$$

The optimal solution $u^0(t)$ of the problem $(1)-(3)$ is deduced from (4), with $\psi^0 = s^*$ such as :

$$F(s^*) = 0. \tag{6}$$

We apply the Newton's method to solve (6). For this, if s^k is an approximation of order k, then the approximation of order $(k+1)$ is written as follows :

$$s^{k+1} = s^k - JF^{-1}(s^k)F(s^k), \quad k = 0,1,2,... \tag{7}$$

4.7 Example

We apply the simple shooting method to the problem (P_α) :

$$J(u) = \frac{1}{2}\alpha\|x(t_1)\|^2 + c^T x(t_1) \to \min,$$

$$\dot{x}(t) = Ax(t) + bu(t), \quad x(0) = x^0,$$

$$-L \le u(t) \le L, \quad t \in T = [0, t_1]$$

where $\alpha = 1$, $c = (c_1, c_2, c_3)^T = (-18, -4, \frac{-14}{3})$,

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad x^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, L = 1, \quad t_1 = 6.$$

The Hamiltonian is :

$$H(x, \psi, u) = \psi^T \dot{x} = \psi_1 x_2 + \psi_2 x_3 + \psi_3 u.$$

The boundary value problem is written as follows :

$$\begin{cases} \dot{x}(t) = Ax(t) + bu(t), \quad x(0) = x^0, \\ \dot{\psi}(t) = -A^T \psi(t), \quad \psi(t_1) = -(\alpha x(t_1) + c), \end{cases}$$

where

$$H(x(t), \psi(t), u(t)) = \max_{v \in [-L,L]} H(x(t), \psi(t), v) \Rightarrow \psi_3(t)u(t) = \max_{v \in [-L,L]} \psi_3(t)v \Rightarrow u(t) = Lsign\psi_3(t), t \in T.$$

So we obtain the differential system of 6 equations and 6 unknowns at two boundary value $t = t_0 = 0$ and $t = t_1 = 6$:

$$\begin{cases} \dot{x}(t) = Ax(t) + Lbsign\psi_3(t), \quad x(0) = x^0, \\ \dot{\psi}(t) = -A^T \psi(t), \quad \psi(t_1) = -(\alpha x(t_1) + c). \end{cases}$$

We set the Cauchy's problem :

$$\begin{cases} \dot{x}(t) = Ax(t) + Lbsign\psi_3(t), \quad x(0) = x^0 & (1) \\ \dot{\psi}(t) = -A^T \psi(t), \quad \psi(0) = \psi^0 & (2) \end{cases}$$

In this problem of Cauchy, the solutions $x(t)$ and $\psi(t)$ depend on the parameter ψ^0, i.e,

$$x(t) = x(t, \psi^0), \quad \psi(t) = \psi(t, \psi^0), \quad \triangle(t) = \triangle(t, \psi^0) = \psi_3(t, \psi^0).$$

We note that

$$sign \triangle(t, \psi^0) = \delta(t, \psi^0) = \begin{cases} 1, & if\ \delta(t, \psi^0) > 0, \\ -1, & if\ \delta(t, \psi^0) < 0. \end{cases}$$

We get from (2) :

$$\psi(t,\psi^0) = e^{-A^T t}\psi^0.$$

Indeed, we have :

$$\dot{\psi}(t,\psi^0) = -A^T e^{-A^T t}\psi^0 = -A^T\psi(t,\psi^0), \qquad \psi(0,\psi^0) = \psi^0.$$

We deduce :

$$\psi(t_1,\psi^0) = e^{-A^T t_1}\psi^0,$$

$$x(t_1,\psi^0) = e^{At_1}x^0 + L\int_0^{t_1} e^{A(t_1-t)}b\delta(t,\psi^0)dt.$$

If we set $\psi^0 = s$, then we get the shooting function :

$$F(s) = \psi(t_1,s) + \alpha x(t_1,s) + c,$$

$$F(s) = e^{-A^T t_1}s + \alpha e^{At_1}x^0 + c + \alpha L\int_0^{t_1} e^{A(t_1-t)}b\delta(t,s)dt.$$

Since $\delta(t,s) = \pm 1$, then $J\delta(t,s) = 0$ and we get

$$JF(s) = e^{-A^T t_1} + 0 = e^{-6A^T} = (e^{-6A})^T.$$

We have

$$e^{At} = I_3 + tA + \frac{1}{2}t^2A^2 \quad (because\ \ A^3 = 0), \quad [e^{At}]^{-1} = e^{-At}.$$

Thus,

$$e^{At} = \begin{pmatrix} 1 & t & \frac{1}{2}t^2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}, \quad e^{-At} = e^{A(-t)} = \begin{pmatrix} 1 & -t & \frac{1}{2}t^2 \\ 0 & 1 & -t \\ 0 & 0 & 1 \end{pmatrix}.$$

Then we obtain

$$e^{-At_1} = \begin{pmatrix} 1 & -t_1 & \frac{1}{2}t_1^2 \\ 0 & 1 & -t_1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & -6 & 18 \\ 0 & 1 & -6 \\ 0 & 0 & 1 \end{pmatrix},$$

So we get

$$e^{-A^T t_1} = (e^{-6A})^T = \begin{pmatrix} 1 & 0 & 0 \\ -6 & 1 & 0 \\ 18 & -6 & 1 \end{pmatrix},$$

$$JF(s) = const = \begin{pmatrix} 1 & 0 & 0 \\ -6 & 1 & 0 \\ 18 & -6 & 1 \end{pmatrix} = e^{-6A^T},$$

$$JF^{-1}(s) = const = \begin{pmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 18 & 6 & 1 \end{pmatrix}.$$

We have the formula :

$$s^{k+1} = s^k - JF^{-1}(s^k)F(s^k),$$

where

$$JF^{-1}(s^k) = \begin{pmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 18 & 6 & 1 \end{pmatrix} \quad and \quad F(s^k) = e^{-6A^T} s^k + \int_0^6 e^{A(6-t)} b\delta(t, s^k) dt + c,$$

with

$$e^{A(6-t)} b = \begin{pmatrix} 1 & 6-t & \frac{1}{2}(6-t)^2 \\ 0 & 1 & 6-t \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(6-t)^2 \\ -t+6 \\ 1 \end{pmatrix},$$

$$e^{-6A^T} s^k = \begin{pmatrix} 1 & 0 & 0 \\ -6 & 1 & 0 \\ 18 & -6 & 1 \end{pmatrix} \begin{pmatrix} s_1^k \\ s_2^k \\ s_3^k \end{pmatrix} = \begin{pmatrix} s_1^k \\ -6s_1^k + s_2^k \\ 18s_1^k - 6s_2^k + s_3^k \end{pmatrix},$$

$$\delta(t, s^k) = sign\psi_3(t, s^k), \quad F = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}.$$

Then we have

$$F(s^k) = \begin{pmatrix} s_1^k \\ -6s_1^k + s_2^k \\ 18s_1^k - 6s_2^k + s_3^k \end{pmatrix} + \int_0^6 \begin{pmatrix} \frac{1}{2}(6-t)^2 \\ -t+6 \\ 1 \end{pmatrix} sign\psi_3(t, s^k) + \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix},$$

with

$$F_1(s^k) = s_1^k + \frac{1}{2} \int_0^6 (6-t)^2 sign\psi_3(t, s^k) dt + c_1;$$

$$F_2(s^k) = -6s_1^k + s_2^k + \int_0^6 (-t+6) sign\psi_3(t, s^k) dt + c_2;$$

$$F_3(s^k) = 18s_1^k - 6s_2^k + s_3^k + \int_0^6 sign\psi_3(t, s^k) dt + c_3.$$

Remark 1

We obtain $\psi_3(t, s^k)$ by solving the following conjugate system of Cauchy :

$$\dot{\psi} = \frac{-\partial H}{\partial x} = -A^T \psi, \quad \psi(0) = s^k \Rightarrow \psi(t, s^k) = e^{-A^T t} s^k,$$

i.e,

$$\psi(t, s^k) = \begin{pmatrix} \psi_1(t, s^k) \\ \psi_2(t, s^k) \\ \psi_3(t, s^k) \end{pmatrix} = (e^{-At})^T s^k = \begin{pmatrix} 1 & 0 & 0 \\ -t & 1 & 0 \\ \frac{1}{2}t^2 & -t & 1 \end{pmatrix} \begin{pmatrix} s_1^k \\ s_2^k \\ s_3^k \end{pmatrix} = \begin{pmatrix} s_1^k \\ -s_1^k t + s_2^k \\ \frac{1}{2}s_1^k t^2 - s_2^k t + s_3^k \end{pmatrix}$$

Hence, we get

$$\psi_3(t, s^k) = \frac{1}{2} s_1^k t^2 - s_2^k t + s_3^k,$$

$$sign\psi_3(t, s^k) = sign(\frac{1}{2} s_1^k t^2 - s_2^k t + s_3^k)$$

We obtain the following formula :

$$F(s^k) = \begin{cases} F_1(s^k) = s_1^k + \frac{1}{2}\int_0^6 (t-6)^2 sign(\frac{1}{2}s_1^k t^2 - s_2^k t + s_3^k)dt + c_1, \\ F_2(s^k) = -6s_1^k + s_2^k + \int_0^6 (-t+6) sign(\frac{1}{2}s_1^k t^2 - s_2^k t + s_3^k)dt + c_2, \\ F_1(s^k) = 18s_1^k - 6s_2^k + s_3^k + \int_0^6 sign(\frac{1}{2}s_1^k t^2 - s_2^k t + s_3^k)dt + c_3. \end{cases}$$

So we get

$$s^{k+1} = s^k - JF^{-1}(s^k)F(s^k)$$

$$s^{k+1} = s^k - \begin{pmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 18 & 6 & 1 \end{pmatrix} \begin{pmatrix} F_1(s^k) \\ F_2(s^k) \\ F_3(s^k) \end{pmatrix},$$

$$s^{k+1} = s^k - \begin{pmatrix} F_1(s^k) \\ 6F_1(s^k) + F_2(s^k) \\ 18F_1(s^k) + 6F_2(s^k) + F_3(s^k) \end{pmatrix}.$$

Remark2

We use the following algorithm to find the sign of $\psi_3(t, s^k)$ on the interval $[0,6]$: $s0 = [18, 112, 1058/3]$

```
    coefficients = [-18*s0(1), -4*s0(2), -14/3*s0(3)];
x = 0:0.01:6;
y = polyval(coefficients, x);
signes = sign(y);
indices = find(diff(sign(y))  = 0);
pointsdechangement = x(indices);
disp(pointsdechangement);
plot(x, signes);
intervals = [];
signeprecedent = sign(y(1));
intervaldebut = x(1);
for i = 2:length(x)
signeactuel = sign(y(i));

    if signeactuel  = signeprecedent
intervalfin = x(i-1);
intervals = [intervals; intervaldebut, intervalfin, signeprecedent];
intervaldebut = x(i);
signeprecedent = signeactuel;
end end intervalfin = x(end);
intervals = [intervals; intervaldebut, intervalfin, signeprecedent];
disp(intervals);
```

We calculate the Newton's iterations with Matlab in order to calculate $\psi^0 = s^*$ such as $F(s^*) = 0$, which corresponds to the optimal control $u^0(t), t \in [0,6]$:

$$u^0(t) = \begin{cases} 1, & if \quad t \in [0,2[; \\ -1, & if \quad t \in [2,4[; \\ 1, & if \quad t \in [4,6], \end{cases}$$

with

$$J(u^0) = \frac{-1576}{9} < 0 \ \ (\simeq -175,111111).$$

4.8 conclusion

In this chapter,we have presented the Newton's method for solving equations, then we have explored different norm minimization techniques commonly used in optimal control problems. For illustration, we have solved a numerical example via the Simple Shooting method.

General conclusion

In conclusion, norm minimization stands out as a fundamental aspect of optimal control problems, with its wide-ranging applications spanning numerous fields. The endeavor to minimize norms necessitates the utilization of various optimization techniques, including both convex and nonconvex optimization. Each approach exhibits its own strengths and weaknesses, and the selection of the most suitable method depends on the nature and complexity of the specific problem.

Convex optimization, in particular, holds a preference in many cases due to its favorable mathematical properties. The guarantee of optimality in the solution and the efficiency of the algorithm make convex optimization an attractive choice. It provides a solid framework for norm minimization problems and yields efficient and effective control strategies.

The primary objective of this dissertation was to explore and determine efficient ways to minimize the norms of terminal states, thereby facilitating the achievement of desired objectives in control problems. By delving into the existing literature, analyzing various techniques, and investigating optimization methods, this research aimed to contribute to the development of practical and effective approaches for norm minimization in optimal control.

Throughout the dissertation, we have discussed the mathematical foundations of norm minimization, reviewed different optimization techniques, and examined their applicability to optimal control problems. By critically evaluating the strengths and weaknesses of these methods, we aimed to identify efficient and effective ways to minimize norms while considering the complexities and constraints associated with the control problem.

By providing valuable insights into norm minimization in the context of optimal control, this research contributes to the broader field of control theory and its related disciplines. The findings of this dissertation can assist researchers, practitioners, and engineers in selecting appropriate methods and designing control strategies that effectively minimize norms, leading to improved system performance, reduced energy consumption, and enhanced robustness in a variety of applications.

Bibliography

[1] Ascher, U. M., Mattheij, R. M., and Russell, R. D. Numerical solution of boundary value problems for ordinary differential equations. Prentice Hall, (1988).

[2] Betts, J. T. Practical methods for optimal control and estimation using non-linear programming. SIAM, Philadelphia, PA, (2010).

[3] Bibi, M. O. Contrôle optimal. Cours de Master 1 en Mathématiques Appliquées, Université de Béjaia, (2021).

[4] Bibi, M. O. Programmation linéaire et quadratique. Cours de Master 1 en Mathématiques Appliquées, Université de Béjaia, (2021).

[5] Bibi, M. O. Methods for solving linear-quadratic problems of optimal control. Thèse de Doctorat en Mathématiques Appliquées, Université de Minsk, (1985).

[6] Bibi, M. O. Optimization of a linear dynamic system with double terminal constraint on the trajectories. Optimization, vol.30, 4(1994), 359-366.

[7] Bibi, M. O. Support method for solving a linear-quadratic problem with polyhedral constraints on control. Optimization, vol.37, 4(1996), 139-147.

[8] Chernushevich, A. Method of support problems for solving a linear-quadratic problem of terminal control. International Journal of Control, vol.52,6(1990), 1475-1488.

[9] Fedorenko, R. Approximate solution of optimal control problems. Moscow : Nauka, (1978).

[10] Frank, L. L., Draguna, V., and Vassilis, L. S. Optimal control. John Wiley, third Edition, Canada, (2011).

[11] Frederic, B., and Pierre, R. Commande et optimisation de systèmes dynamiques. Editions Ecole Polytechnique Amazon, France, (2005).

[12] Gabasov, R., and Kirillova, F.M. Constructive methods of optimization, Part 2 : control problems. University Press, Minsk, (1984).

[13] Gabasov, R., Kirillova, F.M., Kostyukova, I. O., and Raketsky, V.M. Constructive methods of optimization, Part 4 : convex problems. University Press, Minsk, (1987).

[14] Gabasov, R., Kirillova, F.M., and Pavlenok, N. Constructing open-loop and closed-loop solutions of linear-quadratic optimal control problems. Computational Mathematics and Mathematical Physics, vol. 48, 10(2008), 1715-1745.

[15] Gabasov, R., Kirillova, F.M., and Prischepova, S. V. Optimal feedback control. Springer-Verlag, London, (1995).

[16] Ghellab, F. Principe du maximum de support dans un problème de contrôle optimal quadratique avec une entrée non linéaire. Thèse de Doctorat en Mathématiques Appliquées, Université de Béjaia, (2020).

[17] Gornov, A. Y., Tyatyushkin, A. I., and Finkelstein, E. A. Numerical methods for solving terminal optimal control problems. Computational Mathematics and Mathematical Physics, vol. 56, 2(2016), 221-234.

[18] Khimoum, N. Contrôle optimal multivariable et applications à un Jeu différentiel linéaire-quadratique. Thèse de Doctorat en Mathématiques Appliquées, Université de Béjaia, (2019).

[19] Khimoum, N., and Bibi, M. O. Primal-dual method for solving a linear-quadratic multi-input optimal control problem. Optimization Letters, vol. 14, 3(2020), 653-669.

[20] Lee, E. B., and Markus, L. Foundations of optimal control theory, John Wiley, New York, (1967).

[21] Bergounioux, M. Optimisation et contrôle des systèmes linéaires. Deuxième cycle. Ecole d'Ingénieurs. Dunod, Orléans, (2001).

[22] Moissiev, N. N. Numerical methods in optimal systems theory. Nauka, Moskow, (1971).

[23] Pontryagin, L., Boltyanski, V., Gamkrelidze, R., and Mishchenko, E. The mathematical theory of optimal processes. Interscience, New York, (1962).

[24] Rao, A. V. A survey of numerical methods for optimal control. Advances in the Astronautical Sciences, vol. 135, 1(2009), 497-528.

[25] Sandro, S., and Annamaria S. Dynamical Systems and Optimal Control. Bocconi University Press, first Edition, Italy, (2018).

[26] Sklab M., and Tighremt S. Contrôle Optimal linéaire quadratique et applications. Mémoire de Master, Université de Béjaia, (2017).

[27] Stoer, J., and Bulirsch, R. Introduction to numerical analysis, Springer-Verlag, Berlin, (1980).

[28] Stryk, O., and Bulirsch, R. Direct and Indirect methods for trajectory optimization. Mathematics Institue, University of Munchen 2, Germany, (1992).

[29] Trélat, E. Contrôle optimal : théorie et applications. Vuibert, collection Mathématiques Concrètes, Paris, (2005).

[30] Zaitri, M. A., Bibi, M. O., and Bentobache, M. A hybrid direction algorithm for solving optimal control problems. Cogent Mathematics and Statistics, vol.6, 1(2019), 1-12.

[31] Zhao, S., and Zhou, J. Solutions to constrained optimal control problems with linear systems. Journal of Optimization Theory and Applications, vol.178, 2(2018), 349-362.

Agzul ⵣⴰⵣⵓⵍ

Iswi ugemmir ayi d tazrewt n tarrayin n ferru i-ussaddey n ulugen deg yegna n weswaḍ akkay. Di tazwara, nsaher-d s tewzel tiẓri n weswaḍ akkay aked umsihel asnuzmir afesnu. Sakin, numa igna imzirgen-isnuzmiren s tmariwin u nemmel-d snat n tarrayin tigejdayin : asefraray abruyan s yeswaḍen n udemmer aked tarrayt n walday aḥerfi. Iswi n umahil ayi d tiwsi ɣer tigzi talqayant n titiknitin n ussaddey n ulugen deg yiwen wegnu n weswaḍ akkay.

Awalen n tsura : Aswaḍ akkay, assaddey n ulugen, tarrayt n usefraray abruyan s yeswaḍen n udemmer, tarrayt n walday aḥerfi.

Abstract

This dissertation explores resolution methods for norm minimization in optimal control problems. It begins by introducing general concepts of optimal control, followed by an overview on convex quadratic programming. Subsequently, it focuses on linear quadratic problems with constraints and then delves into two principle methods: partial discretization with impulsive controls and the simple shooting method. The ultimate goal of this research is to contribute to a deeper understanding of norm minimization techniques in an optimal control problem.

Keywords : Optimal control, norm minimization, method of partial discretization with impulsive controls, simple shooting method.

ملخص

تستكشف هذه الأطروحة طرق الحل لتقليل المعايير في مشاكل التحكم المثلى. يبدأ بتقديم مفاهيم عامة للتحكم الأمثل، متبوعا بنظرة عامة على البرمجة التربيعية المحدبة. بعد ذلك، يركز على المشكلات التربيعية الخطية مع القيود ثم يتعمق في طريقتين أساسيتين: التمييز الجزئي مع عناصر التحكم الاندفاعية وطريقة التصوير البسيطة. الهدف النهائي من هذا البحث هو المساهمة في فهم أعمق لتقنيات تقليل المعايير في مشكلة التحكم المثلى

.

الكلمات الرئيسية :التحكم الأمثل، تقليل المعايير، طريقة التمييز الجزئي، طريقة التصوير البسيطة .

Résumé

Ce mémoire examine les méthodes de résolution pour la minimisation de la norme dans les problèmes de contrôle optimal. Il commence par présenter des généralités sur le contrôle optimal, suivi d'un aperçu sur la programmation quadratique convexe. Ensuite, il aborde les problèmes linéaires quadratiques avec contraintes, avant de se pencher sur deux méthodes principales : la discrétisation partielle avec des contrôles impulsionnels et la méthode de tir simple. L'objectif ultime de ce travail est de contribuer à une meilleure compréhension des techniques de minimisation de la norme dans un problème de contrôle optimal.

Mots clés : Contrôle optimal, minimisation de la norme, méthode de discrétisation partielle avec des contrôles impulsionnels, méthode de tir simple.