MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU
FACULTE DES SCIENCES
DEPARTEMENT DE MATHEMATIQUES

THÈSE DE DOCTORAT

SPECIALITE : MATHÉMATIQUES
OPTION : RECHERCHE OPERATIONNELLE ET OPTIMISATION

Présentée par :
Mme Saliha TITOUCHE

Sujet :

Résolution d'un problème de contrôle optimal avec contraintes sur l'état

Devant le jury d'examen composé de :

Mr. OUANES Mohand ;	Professeur ;	U.M.M.T.O ;	Président
Mr. AIDENE Mohamed ;	Professeur ;	U.M.M.T.O ;	Rapporteur
Mr. SPITERI Pierre ;	Professeur ;	U. Toulouse ;	Examinateur
Mr. AIDER Meziane ;	Professeur ;	U.S.T.H.B ;	Examinateur
Mr. BIBI Mohand Ouamer ;	Professeur ;	U. Béjaia ;	Examinateur
Mr. OUKACHA Brahim ;	M. de Conférences A ;	U.M.M.T.O ;	Examinateur

Soutenue le :

Résumé : Dans le présent travail, nous mettons en œuvre la méthode de relaxation couplée à la méthode de tir pour résoudre un problème de contrôle optimal quadratique avec contraintes sur l'état et la valeur de l'état final fixé ; la convergence de la méthode itérative étudiée est analysée. Nous appliquons ensuite la méthode proposée pour la détermination de la commande d'un grand système thermique composé d'un four vertical, dans la cheminée duquel est placé un barreau. L'objectif est d'amener la température relevée en n points du barreau à une température désirée, en un temps fini T.

Mots clés : contrôle optimal, méthode de realaxation, méthode de tir, sous-différentiel, processus thermique.

Abstract : In the present study, we apply the relaxation method coupled with the shooting method to solve a quadratic optimal control problems with constraints on the state and the final state fixed ; the convergence of the iterative method used is analyzed. Then, we apply the proposed method for the determination of the control of a large thermal system composed of a vertical oven, in the chimney of which is placed a bar. The goal is to bring the temperature identified in n points of the bar at a desired temperature, in a finite time T.

Keywords : optimal control, relaxation method, shooting method, sub-differential, thermic process.

Remerciements

Je tiens à exprimer toute ma reconnaissance au professeur Mohamed AIDENE mon directeur de thèse, de l'université Mouloud Mammeri de Tizi-Ouzou, qui a proposé et dirigé ce travail, et qui par son engagement et ses orientations a contribué grandement à son aboutisement. Qu'il soit à cette occasion assuré de ma profonde reconnaissance.

Je remercie également toute l'équipe du laboratoire de l'IRIT-ENSEEIHT, en particulier les pofesseurs Pierre SPITERI et Frédéric MESSINE qui m'ont fait profiter de leur experience dans le domaine de l'optimisation, ainsi que pour leur investissement indéniable dans les travaux présentés ici. Leurs qualités humaines et scientifiques ont contribué à améliorer et parfaire mes traveaux de thèse.

Il m'est très agreable de remercier le professeur Mohand OUANES de l'université Mouloud Mammeri de Tizi Ouzou, qui m'a fait l'honneur de présider le jury de cette thèse.

Mes remerciements vont particulièrement vers le professeur Pierre SPITERI de l'ENSEEIHT de Toulouse pour avoir accepté de faire partie de ce jury.

Je remercie également le Professeur Meziane AIDER de l'université USTHB de Bab Ezzouar de faire partie de ce jury et d'avoir accepté d'examiner ce travail.

J'ai l'honneur de remercier le professeur Mohand Ouamer BIBI de l'université de Béjaia pour avoir accepté d'examiner ce travail et de faire partie du jury.

Je tiens aussi à remercier chaleureusement M^r Brahim OUKACHA, maître de conference ; classe A à l'université Mouloud Mammeri de Tizi Ouzou pour avoir accepté de faire partie de ce jury et d'examiner ce travail.

Enfin je remercie mon mari, mes parents, mes frères et sœurs et ma belle famille, pour leur soutien moral et leurs encouragements.

Table des matières

Introduction

La théorie du contrôle analyse les propriétés des systèmes commandés, c'est-à-dire des systèmes dynamiques sur lesquels on peut agir au moyen d'une commande. L'objectif d'un problème de contrôle est d'amener le système d'un état initial donné à un état final en respectant certains critères. Une voiture sur laquelle on agit avec les pédales d'accélérateur et de frein, et que l'on guide avec le volant est un exemple de système dynamique commandé.

Un système de contrôle est un système dynamique sur lequel on peut agir au moyen d'une commande. Pour définir précisément le concept de système de contrôle, il faut utiliser le langage mathématique. Chaque système a une structure, des propriétés et des finalités spécifiques. Notons que ce concept peut aussi bien décrire des transformations discrètes que continues. Cela permet donc de modéliser le fonctionnement de robots, de systèmes adaptatifs à structure variable, etc. En considérant tous ces objets comme des systèmes de contrôle, on s'intéresse à leur comportement et à leurs caractéristiques fonctionnelles, sans forcément attacher d'importance à leurs propriétés internes ou intrinsèques. Par conséquent, deux systèmes de contrôle ayant, en un certain sens, même comportement et des caractéristiques similaires, sont considérés comme identiques. De nos jours, les systèmes automatisés font complètement partie de notre quotidien ; le but est d'améliorer notre qualité de vie et de faciliter certaines tâches.

L'objectif peut être aussi de stabiliser le système pour le rendre insensible à certaines perturbations, ou encore de déterminer des solutions optimales pour un certain critère d'optimisation (contrôle optimal). Du point de vue mathématique, un système de contrôle

est un système dynamique dépendant d'un paramétre dynamique appelé le contrôle. Pour le modéliser, on peut avoir recours à des équations différentielles, des équations intégrales, des équations fonctionnelles, des équations aux différences finies, des équations aux dérivées partielles, des équations stochastiques, etc. Pour cette raison la théorie du contrôle est à l'interconnexion de nombreux domaines mathématiques. Une fois le problème de contrôlabilité résolu, on peut de plus vouloir passer de l'état initial à l'état final en minimisant un certain critère ; on parle alors d'un problème de contrôle optimal [78].

Historiquement, le problème de contrôle optimal est apparu après la seconde guerre mondiale dans le cadre du calcul des variations, répondant à des besoins pratiques de guidage, notament dans le domaine de l'aéronotique et de la dynamiques de vol. La formalisation de cette théorie a posé des questions nouvelles ; par exemple dans la théorie des équations différentielles ordinaires elle a motivé un concept de solution généralisée et a engendré de nouveaux résultats d'existence de trajectoires optimales. La théorie de contrôle optimal est trés liée à la mécanique classique, en particulier aux principes variationnels de la mécanique [44], [82] (principe de Fermat, équations d'Euler-Lagrange, etc).

Les problèmes de commande optimale sont appliqués à de nombreux domaines, par exemple l'optimisation de trajectoire, la robotique, la chimie, la biologie, l'économie, etc. Pour résoudre ces problèmes, deux grandes théories ont émergé indépendamment depuis une cinquantaine d'années : le principe du maximum de Pontryagin et le principe de la programmation de Bellman. La première théorie, basée sur le principe du maximum du Pontryagin [73], découvert par L. S. Pontryagin en 1956, donne une condition nécessaire d'optimalité [39]. Cette théorie est développée dans différentes branches mathématiques : le problème de contrôle optimal d'équations aux dérivées partielles, la théorie de contrôle stochastique, la théorie des jeux, etc. La deuxième théorie, apparue dans les années 60, est basée sur le principe de la programmation dynamique de Bellman [5], qui fournit une condition suffisante d'optimalité.

Il existe différentes méthodes pour résoudre les problèmes de commande optimale, chacune ayant ses avantages et ses inconvénients. Le choix de la méthode dépend du problème considéré. Généralement, les problèmes de commande optimale sont résolus

de façon numérique ; par conséquent les méthodes de résolution ont nettement évolué ces dernières années. La plupart des anciennes méthodes étaient basées sur l'obtention d'une solution qui satisfait soit les équations d'Euler-Lagrange, qui sont des conditions nécessaires d'optimalité, soit l'équation de Hamilton -Jacobi-Bellman [5], [6], qui est une condition suffissante d'optimalité. Ces méthodes sont appelées les méthodes indirectes. L'inconvénient principal des méthodes indirectes, est la résolution fastidieuse de l'équation d'Hamilton-Jacobi-Bellman. Ce qui a amené plusieurs chercheurs à utiliser des méthodes directes pour résoudre le problème de la commande optimale. Ces méthodes consistent à discrétiser les équations du problème, et ainsi se ramener à un problème de programmation non linéaire, c'est-à-dire un problème d'optimisation non linéaire en dimension finie. Le problème discrétisé peut ensuite être résolu par n'importe quel algorithme d'optimisation en dimension finie, par exemple par programmation quadratique séquentielle (voir par exemple Betts [8], Bonnans et Launay [18]), ou par une méthode de points intérieurs (voir Laurent-Varin et al. [54]).

Dans cette thèse, nous avons mis en œuvre la méthode de relaxation couplée à la méthode de tir pour résoudre un problème de contrôle optimal avec contraintes sur l'état et sur l'état final. Par rapport aux travaux de ([67]) où il n'y a pas de contraintes sur l'état final, la situation est plus complexe. En général, les conditions de la méthode de tir se traduisent par la formulation d'un problème aux deux bouts qui possède une structure particulière, car elles découlent de la dérivation du Hamiltonien. Le tir simple consiste à trouver un zéro de la fonction de tir associée au problème original. Il n'y a pas ici de discrétisation explicite, même si la méthode requiert l'intégration numérique du système différentiel et par conséquent la discrètisation temporelle explicite ou implicite ; généralement on utilise des schémas explicite en temps. Le choix de ces méthodes se justifie par leurs avantages bien connus, à savoir une grande précision. Toutefois, l'inconvénient de cette méthode est la nécessité de disposer d'une donnée initiale de la commande. Une des démarches classiques consiste à appliquer un algorithme de quasi-Newton à la fonction de tir, ceci est particulièrement vrai pour des problèmes à contrôle Bang-Bang.

La contribution de cette thèse est de présenter une nouvelle méthode qui est la méthode

de relaxation couplée à la méthode de tir pour résoudre un problème de contrôle optimal avec contrainte sur l'état et l'état final. Puis nous appliquons cette méthode à la régulation d'un processus thermique de grande dimension ainsi qu'à un système en anneau.

La suite des chapitres est organisée comme suit :

Au premier chapitre, nous rappelons les notions de dérivées clasiques au sens de Fréchet et au sens de Gâteaux et nous introduisons la notion de sous-différentiel dans le cas où la fonction n'est pas dérivable.

Le second chapitre est consacré à la résolution numérique d'équations différentielles qui sont la base fondamentale du contrôle optimal.

Au troisième chapitre, nous rappelons la formulation d'un problème de contrôle optimal, et nous présentons la notion de contrôlabilité pour les systèmes linéaires et les systèmes non linéaires. La dernière section de ce chapitre est consacrée à l'énoncé général du principe du maximum de Pontryagin.

Au quatrième chapitre, nous présentons deux types de méthode numérique ; les méthodes directes et les méthodes indirectes.

Au cinquième chapitre, on développe une nouvelle méthode de résolution d'un problème de contrôle optimal qui consiste à un couplage de deux méthodes : la méthode de relaxation et la méthode de tir.

Au dernier chapitre, nous appliquons la méthode décrite au cinquième chapitre pour la détermination de la commande optimale d'un grand système physique de regulation thermique.

Le manuscrit se termine par une conclusion et des recommandations pour les traveaux futurs.

Liste des travaux scientifiques

Article accepté dans une revue internationale :

1. Saliha Titouche, Pierre Spiteri, Frédéric Messine, Mohamed Aidene, "Optimal control of a large thermic process", Journal of Process and Control, vol 25, 50-58, 2015.

Article soumis pour acceptation :

2. Saliha Titouche, Pierre Spiteri, Frédéric Messine, Mohamed Aidene, "A relaxation based method for solving optimal control problems with a constraint on the state", Revue internationale des technologie avancées. Octobre 2014.

Présentation dans des manifestations internationale :

3. Saliha Titouche, Pierre Spiteri, Frédéric Messine, Mohamed Aidene, "Résolution d'un problème de contrôle optimal avec une contrainte sur l'état final et sur l'état par la méthode de relaxation", Colloque sur l'optimisation et les systèmes d'information COSI'13 organisé par le CDTA à Alger du 09 au 11 juin 2013.

Chapitre 1

Introduction au calcul différentiel

1.1 Introduction

En optimisation, le calcul différentiel joue un rôle important. Dans ce chapitre nous rappelons les notions de dérivées classiques au sens de Fréchet et au sens de Gâteaux en dimension infinie et en dimension finie. Lorsque une fonction n'est pas dérivable, nous introduisons la notion de sous-différentiel qui sera utile dans la suite.

1.2 Notions de différentielle

Nous rappelons dans ce chapitre certaines notions de base de calcul en dimension n, en particulier, la derivée au sens de Gâteaux et la dérivée au sens de Fréchet. On rappelle qu'une fonction réelle f d'une variable réelle est différentiable en un point x s'il existe un nombre réel $a = f'(x)$ tel que

$$\lim_{t\to 0}(\frac{1}{t})[f(x+t) - f(x) - at] = 0.$$

Cette définition se prolonge d'une manière simple en dimension n.

1.2.1 Différentielle au sens de Gâteaux

Définition 1.1. Soit V un espace vectoriel normé. Une fonction $F(x)$ de V dans $\mathbb{R}$ est une fonction différentiable au sens de Gâteaux en $x \in V$ s'il existe une forme linéaire continue

sur V, notée $F'_G(x)$ telle que :

$$\forall h \in V : \lim_{t \to 0} \frac{1}{t}[F(x+th) - F(x)] =< F'_G(x), h >,$$

où $F'_G(x)$ est la dérivée au sens de Gâteaux de F en x.

1.2.2 Différentielle au sens de Fréchet

Définition 1.2. Soit V un espace vectoriel normé. Une application $F(x)$ de V dans $\mathbb{R}$ est différentiable au sens de Fréchet en $x \in V$, s'il existe une application linéaire continue de V dans $\mathbb{R}$, notée $F'_F(x)$ telle que :

$$\forall h \in V : F(x+h) = F(x) + < F'_F(x), h > +\epsilon(h), \quad \text{avec} \lim_{\|h\| \to 0} \frac{|\epsilon(h)|}{\|h\|} = 0,$$

où $F'_F(x)$ est la dérivée au sens de Fréchet de F en x.

Une fonction différentiable au sens de Fréchet est différentiable au sens de Gâteaux, l'inverse n'est pas toujours vrai.

Remarque 1.1. Si V est un espace de dimension finie, toutes les normes sont équivalentes et il résulte de la définition précédente que la différentiabilité ne dépend pas de la norme choisie [70].

Si V est de dimension infinie, une fonction peut être différentiable pour une norme sans l'être pour une autre norme non équivalente [70].

1.2.3 Résultats essentiels

Les résultats, les plus fréquemment utilisés sur les dérivées concernent les théorèmes de la moyenne et des accroissement finis ; dans cette section, nous rappelons divers résultats de ce type, avec quelques applications ; nous emploierons aussi la notation $[x, y], x, y \in \mathbb{R}^n$, pour dénoter l'intervalle fermé $[x, y] = \{z/z = tx + (1-t)y, 0 \leq t \leq 1\}$.

Nous commençons en rappelant le théorème des accroissement finis pour des fonctions d'une variable réelle.

Théorème 1.1. *(accroissements finis). Si $F : [a,b] \subset \mathbb{R} \to \mathbb{R}$ est continue sur $[a,b]$ et différentiable sur $[a,b]$, alors il existe un point $t \in (a,b)$ tel que*

$$F(b) - F(a) = F'(t)(b-a).$$

Comme conséquence immédiate de ce résultat unidimentionnel, nous avons le résultat correspondant pour des fonctionnelles.

Théorème 1.2. *Soit $F : V \subset \mathbb{R}^n \to \mathbb{R}$; supposons que F soit G-différentiable en tout point du convexe $V_0 \subset V$. Alors, pour deux points quelconques $x, y \in V_0$, il existe $t \in]0,1[$ tel que :*

$$F(y) - F(x) = F'_G(x + t(y-x)).(y-x).$$

Preuve. Pour $x, y \in V_0$, il découle immédiatement que la fonction $\phi(s) = F(x + s(y-x))$ est différentiable et continue sur $[0,1]$ et que $\phi'(s) = F'_G(x+s(y-x)).(y-x), \quad \forall s \in [0,1]$. Par consequent, en utilisant le Théoréme **1.1**,

$$F(y) - F(x) = \phi(1) - \phi(0) = F'(x + t(y-x)).(y-x),$$

pour $t \in]0,1[$.

Remarque 1.2. Il est important de noter que le Théorème **1.2** n'est pas verifié en général pour des fonctions $F : \mathbb{R}^n \to \mathbb{R}^m,\ m > 1$. Il n'est donc valable que pour les fonctionnelles.

Définition 1.3. Une application $F : V \subset \mathbb{R}^n \to \mathbb{R}^m$ est hémicontinue au point $x \in V$, si pour tout $h \in \mathbb{R}^n$ et $\varepsilon > 0$, il existe $\delta = \delta(\varepsilon, h)$ tel que $|t| < \delta$ et $x + th \in V$, alors

$$\|F(x+th) - F(x)\| < \varepsilon.$$

Théorème 1.3. *[70]. Si $F : V \subset \mathbb{R}^n \to \mathbb{R}^m$ est G-différentiable en chaque point d'un ensemble convexe $V_0 \subset V$, et F' est hémicontinue sur V_0, alors, pour tout $x,\ y \in V_0$, la relation suivante est vérifiée*

$$F(y) - F(x) = \int_0^1 F'_G(x + t(y-x)).(y-x)dt. \tag{1.1}$$

Théorème 1.4. *[70]. Si $F : V \subset \mathbb{R}^n \to \mathbb{R}^m$ est G− différentiable au point $x \in V$, alors F est hémicontinue au point x.*

1.2.4 Gradient et dérivées partielles

V est à présent un espace de dimension finie. Soit F une fonction G-différentiable en x. Il existe alors un vecteur, notée $\nabla F(x)$, appelé gradient, tel que

$$< F'(x), h >=< \nabla F(x), h > .$$

Si $V = \mathbb{R}^n$ muni du produit scalaire standard, on retrouve la définition usuelle du gradient

$$\nabla F(x) = (\frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_i}, \ldots, \frac{\partial F}{\partial x_n})^T$$

1.2.5 Matrice Jacobienne

Définition 1.4. Soit $F : V \subset \mathbb{R}^n \to \mathbb{R}^m$. Soient $f_1, \ldots, f_m$ les composantes de la fonction F, on définit alors la matrice Jacobienne de F notée par J_F par :

$$J_F = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}.$$

1.2.6 Sous-différentiel

Nous considérons à présent une situation où la fonction n'est plus dérivable.

Définition 1.5. Soit E un espace vectoriel. Soit χ une fonction convexe dans E et μ un point de E. On note par $\partial\chi(\mu)$ l'ensemble des $\mu' \in E'$ tel que

$$\chi(\upsilon) \geq \chi(\mu) + < \upsilon - \mu, \mu' >, \text{ pour tout } \upsilon \in E, \tag{1.2}$$

où $<,>$ est le produit de dualité de E dans E' et E' est l'espace topologique dual de E ; un tel élément μ' est appelé sous-gradiant de χ en μ, et $\partial\chi(\mu)$ est appelé le sous-différenciel de χ en μ.

Remarque 1.3. Le produit de dualité de E et E' est une application bilinéaire de $E \times E'$ dans $\mathbb{R}$. Si E est un espace de Hilbert, alors $<,>$ est le produit scalaire de E.

Exemple 1.1. Soit $\chi(u) = |u|$ non-différentiable à l'origine ; le sous différentel de cette application est $\partial\chi(u) \equiv sign(u)$:

$$\partial\chi(u) \equiv sign(u) = \begin{cases} -1, & \text{si } u < 0\,; \\ [-1,+1], & \text{si } u = 0\,; \\ +1, & \text{si } u > 0. \end{cases}$$

et admet le graphe représenté en Figure 1.1

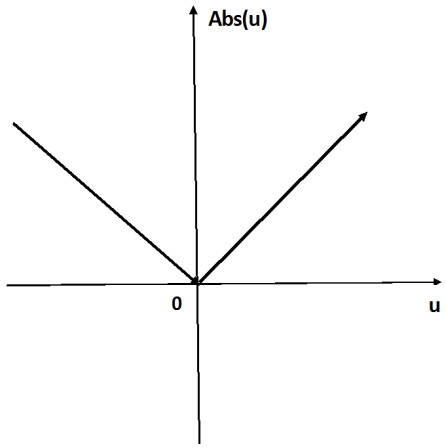

FIG. 1.1 – Sous différentiel de la fonction $\chi(u) = |u|$

Remarque 1.4. Soit χ une fonction différentiable (Fréchet ou Gâteaux différentiable) en μ ; alors $\partial\chi(\mu)$ est un opérateur univoque qui coïncide avec la différentielle au sens de Fréchet ou au sens de Gâteaux de χ en μ. On montre que $\partial\chi(\mu)$ est un ensemble convexe fermé (éventuellement vide voir [3]).

Nous avons également les deux résultats principaux classiques suivants :

Lemme 1.1. *$\mu \in E$ est tel que $\chi(\mu) = \min_{\upsilon\in E}(\chi(\upsilon))$ si et seulement si $0 \in \partial\chi(\mu)$.*

Preuve. Soit $\mu \in E$ tel que $\chi(\mu) = \min_{\upsilon\in E}(\chi(\upsilon))$; nous avons alors trivialement $\chi(\upsilon) \geq \chi(\mu) + < \upsilon - \mu, 0 >$ et donc $0 \in \partial\chi(\mu)$.

Lemme 1.2. *Le sous-différentiel $\partial\chi(\mu)$ est un opérateur monotone (en général multivoque) de E dans E'.*

Preuve. Soit $w' \in \partial\chi(w)$, alors $\chi(v) \geq \chi(w) + < v - w, w' >,\ \forall v \in E$. Soit encore $\mu' \in \partial\chi(\mu)$, alors $\chi(v) \geq \chi(\mu) + < v - \mu, \mu' >,\ \forall v \in E$. On considère la première équation pour $v = \mu$ et la deuxième équation pour $v = w$; on additionne terme à terme et on obtient alors :

$$< w - \mu, w' - \mu' > \geq 0.$$

La fonction indicatrice du sous ensemble convexe $\mathcal{K}$ jouera un rôle important dans la suite ; elle est définie ci-dessous.

Définition 1.6. Soit $\mathcal{K}$ un sous ensemble convexe fermé de E. On appelle fonction indicatrice de $\mathcal{K}$, la fonctionnelle $\Psi_{\mathcal{K}}$ définie par :

$$\Psi_{\mathcal{K}}(\mu) = \begin{cases} 0, & \text{si } \mu \in \mathcal{K}, \\ +\infty, & \text{sinon.} \end{cases}$$

On montre que $\Psi_{\mathcal{K}}(\mu)$ est convexe (voir [53]).

Par conséquent, il résulte du Lemme **1.1** que chercher le minimum de χ sur $\mathcal{K} \subset E$ revient à résoudre une équation multivoque $0 \in \mathcal{A}(v)$, où $\mathcal{A} = \partial(\chi + \Psi_{\mathcal{K}})$, $\Psi_{\mathcal{K}}$ fonction indicatrice du convexe $\mathcal{K}$. En utilisant la définition du sous différentiel, on a (voir [3]) :

$$\partial\Psi_{\mathcal{K}}(v) = \{v' \in E' / < v - w, v' > \geq 0, \textit{ pour tout } w \in \mathcal{K}\}.$$

Ce qui montre que $D(\partial\Psi_{\mathcal{K}}) = D(\Psi_{\mathcal{K}}) = \mathcal{K}$ et $\partial\Psi_{\mathcal{K}}(v) = \{0\}$ pour tout $v \in int(\mathcal{K})$. Par ailleurs, si v se trouve sur la frontière de $\mathcal{K}$, alors $\partial\Psi_{\mathcal{K}}(v)$ est confondu avec le cône normal à $\mathcal{K}$ au point v.

Chapitre 2

Méthodes numériques de résolution des systèmes différentiels ordinaires

2.1 Introduction

Les équations différentielles ordinaires apparaisent dans un nombre important d'applications lieés à des disciplines variées par exemple en physique ou en chimie. Elles forment un cadre naturel au sein duquel un grand nombre de systèmes complexes peuvent être modélisés. Elles interviennent également pour la détermination de la loi de commande optimale de systèmes gouvernés par des équations différentielles ordinaires.

Il n'est pas toujours facile d'obtenir la solution exacte de ces équations différentielles ordinaires, autrement dit, la résolution de ces équations différentielles n'est pas toujours possible analytiquement ; on fait alors appel dans ce cas à des méthodes numériques qui permettent d'obtenir des solutions approchées. On est donc amené à déterminer une majoration d'erreur entre la solution approchée et la solution exacte. Il ne suffit pas de se donner une méthode numérique pour avoir la solution d'une équation différentielle ; en effet la méthode numérique doit aussi vérifier les notions de consistance, de stabilité et de convergence qui garantissent une bonne approximation de la solution. Ces trois notions sont liées par un résultat théorique qui spécifie que la consistance et la stabilité d'un schéma entraîne la convergence de ce schéma.

Les méthodes numériques de résolution d'équations différentielles ordinaires sont nombreuses. Dans ce chapitre nous nous limitons aux méthodes à un pas : la méthode d'Euler,

la méthode de Taylor et la méthode de Runge-Kutta.

2.2 Problème de Cauchy

Le problème de Cauchy (aussi appelé problème aux valeurs initiales) consiste à trouver la solution d'une EDO (équation différentielle ordinaire), scalaire ou vectorielle, satisfaisant des conditions initiales. Soit I_0 un intervalle de $\mathbb{R}$ contenant le point t_0 ; on se donne une fonction f définie et continue sur $I_0 \times \mathbb{R}^m$ à valeurs dans $\mathbb{R}^m$; ainsi qu'un élément y_0 de $\mathbb{R}^m$; le problème de Cauchy asssocié à une équation différentielle ordinaire (EDO) du premier ordre s'écrit :
déterminer une fonction y continue et dérivable sur l'intervalle I_0, à valeurs dans $\mathbb{R}^m$, telle que

$$y'(t) = f(t, y(t)); \quad t \in I_0, \tag{2.1}$$

$$y(t_0) = y_0, \tag{2.2}$$

la condition (2.2) s'appelle la condition initiale. Une fonction y qui vérifie les équations (2.1)-(2.2) est appelée une intégrale du système différentiel (2.1)-(2.2). Nous nous intéresserons plus particulièrement au cas où I_0 est de la forme $[t_0, T]$; les cas où I_0 est de la forme $[t_0, T[$ ou $[t_0, +\infty[$ se traiteraient de même.
Dans de nombreux exemples physiques, la variable t représente le temps ; l'instant t_0 est alors appelé instant initial.

2.3 Théorèmes d'existence et d'unicité

2.3.1 Théorème d'existence

Théorème 2.1 (*Cauchy-Péano*). *[25] On suppose que la fonction f est continue dans un voisinage du point (t_0, y_0) dans $I_0 \times \mathbb{R}^m$; alors il existe un intervalle $J_0 \subset I_0$, au voisinage de t_0 et une fonction $y \in C^1(J_0)$ tels que*

$$\forall t \in J_0, \ y'(t) = f(t, y(t)), y(t_0) = y_0.$$

Définition 2.1. On appelle solution locale du problème (2.1),(2.2) la donnée d'un couple (I, y) où I est un intervalle de $\mathbb{R}$ inclus dans I_0 et où y est une fonction appartenant à $C^1(I)$ telle que

$$y(t_0) = y_0 \quad \text{et} \quad \forall t \in I,\ y'(t) = f(t, y(t))$$

Définition 2.2. On dit que la solution locale (J, z) prolonge la solution locale (I, y) si on a $I \subset J$, et $\forall t \in I;\ y(t) = z(t)$; si de plus $I \neq J$, on dit que (J, z) prolonge strictement (I, y).

Définition 2.3. On dit que la solution locale (I, y) est une solution maximale du problème (2.1),(2.2) s'il n'existe pas de solution locale de ce problème qui la prolonge strictement.

Définition 2.4. On dit que (I_0, y) est une solution globale du problème (2.1), (2.2) dans I, (ou encore que y est solution du problème (2.1), (2.2)), si (I_0, y) est une solution locale de ce problème, et $I_0 = I$.

2.3.2 Théorème d'unicité

Définition 2.5. On dira que le problème (2.1), (2.2) admet une solution et une seule, s'il admet une solution globale et si toute solution locale est la restriction de cette solution globale.

Théorème 2.2. *[25] On suppose que I_0 est de la forme $[t_0,\ T]$ ou $[t_0,\ T[$ ou $[t_0,\ +\infty[$, de plus f est continue sur $I_0 \times \mathbb{R}^m$ et qu'il existe une fonction $l \in \mathfrak{L}(I_0)$ telle que*

$$\forall t \in I_0,\ \forall y,\ z \in \mathbb{R}^m,\ (f(t, y) - f(t, z),\ y - z) \leq l(t)|y - z|^2, \tag{2.3}$$

alors le problème (2.1), (2.2) admet une solution et une seule.

Dans le Théorème **2.2**, $\mathfrak{L}(I_0)$ est un espace vectoriel normé des fonctions réelles mesurables sur I_0 telle que :

$$\|f\| = \int_{I_0} |f(x)|dx < +\infty.$$

Une conséquence immédiate du Théorème **2.2** est le résultat suivant :

Corollaire 2.1 (*Cauchy-Lipschitz*). *[25] On suppose que la fonction f est continue sur $I_0 \times \mathbb{R}^m$ et qu'il existe un réel L tel que*

$$\forall (t, y) \text{ et } (t, z) \in I_0 \times \mathbb{R}^m, \ |f(t, y) - f(t, z)| \leq L|y - z|;$$

alors le problème (2.1),(2.2) admet une solution et une seule.

2.4 Méthode d'Euler

2.4.1 Présentation de la méthode pour résoudre numériquement une EDO

La méthode numérique la plus simple pour résoudre le problème de Cauchy décrit précédement est la méthode d'Euler. Considérons donc le problème différentiel

$$\forall t \in [0, T], y'(t) = f(t, y(t)) \tag{2.4}$$

$$y(0) = \eta \text{ donné dans } \mathbb{R}^m. \tag{2.5}$$

On va se donner des points $t_0 = 0 < t_1 < t_2 < t_3 < \cdots < t_N = T$, et on va essayer de calculer une approximation des valeurs de la solution en tous ces points, c'est-à-dire des valeurs $y_i \simeq y(t_i),\ i = 1, 2, \ldots, N$.

Essentiellement il y a deux approches pour calculer les y_i qui, dans le cas de la méthode d'Euler, vont aboutir au même résultat : soit on approche directement l'EDO par différences finies en utilisant une approximation de la dérivée $y'(t)$, soit on intègre l'EDO, de manière analogue à la preuve d'existence. L'approche par différences finies consiste à trouver une approximation de $y'(t_i)$; par exemple :

$$y'(t_i) \simeq \frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i}. \tag{2.6}$$

L'EDO se réécrit alors sous la forme :

$$\frac{y(t_{i+1}) - y(t_i)}{t_{i+1} - t_i} \simeq f(t_i, y(t_i)), \tag{2.7}$$

et donc :

$$y(t_{i+1}) \simeq y(t_i) + (t_{i+1} - t_i) f(t_i, y(t_i)). \tag{2.8}$$

Soit $y_0 = \eta_h$ une approximation de $y(0) = \eta$; nous construisons par récurrence une approximation y_i de $y(t_i)$ par :

$$y_{i+1} = y_i + (t_{i+1} - t_i) f(t_i, y_i),\ i = 0, \ldots, N-1. \tag{2.9}$$

Le terme y_0 est la condition initiale. La formule (2.9) represente la méthode d'Euler.

La deuxième approche, qui va nous conduire au même résultat mais avec une philosophie très différente, consiste à intégrer l'équation de t_i à t_{i+1} :

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(s, y(s))ds. \tag{2.10}$$

En appliquant la méthode des rectangles, on obtient :

$$\int_{t_i}^{t_{i+1}} f(s, y(s))ds \simeq (t_{i+1} - t_i) f(t_i, y(t_i)), \tag{2.11}$$

et en posant $h = t_{i+1} - t_i = \frac{T}{N}$, la reltion (2.10) devient

$$y_{i+1} = y_i + hf(t_i, y_i),\ i = 0, 1, \ldots, N-1, \tag{2.12}$$

Le shéma défini par (2.12) (ou (2.9)) s'appelle le shéma d'Euler.

2.4.2 Étude de l'erreur

Nous cherchons à obtenir une estimation de l'erreur :

$$e_i = y(t_i) - y_i,$$

entre la solution exacte de (2.4)-(2.5) et la solution approchée donnée par (2.12). Pour cela nous supposons que f est continue sur $[0, T] \times \mathbb{R}^m$ et lipschitzienne.
L'erreur commise au point t_i est :

$$\varepsilon_i = y(t_{i+1}) - y(t_i) - hf(t_i, y(t_i)), \tag{2.13}$$

où $y(.)$ désigne la solution de (2.4)-(2.5) ; cette quantité mesure avec quelle précision la solution exacte vérifie le schéma (2.12) ; elle s'appelle l'erreur de consistance à l'instant t_i de la méthode d'Euler. Pour évaluer l'erreur, on procède comme suit :

$$\begin{aligned} e_{i+1} &= y(t_{i+1}) - y_{i+1} \\ &= [y(t_{i+1}) - y(t_i) - h\ f(t_i, y(t_i))] + [y(t_i) - y_i] + [y_i + h\ f(t_i, y_i)] \\ &+ h[f(t_i, y(t_i)) - f(t_i, y_i)] - y_{i+1} \\ &= \varepsilon_i + e_i + h[f(t_i, y(t_i)) - f(t_i, y_i)] \end{aligned}$$

D'où :

$$\begin{aligned} |e_{i+1}| &\leq |\varepsilon_i| + |e_i| + h|f(t_i, y(t_i)) - f(t_i, y_i)| \\ &\leq |\varepsilon_i| + (1 + L\ h)|e_i|, \end{aligned}$$

en utilisant la condition de lipschitz de f.
Il reste à estimer $|\varepsilon_i|$. En utilisant l'équation (2.13), on a :

$$\begin{aligned} |\varepsilon_i| &= |y(t_{i+1}) - y(t_i) - h\ f(t_i, y(t_i))| \\ &= |\int_{t_i}^{t_{i+1}} y'(s)ds - \int_{ti}^{t_{i+1}} y'(t_i)ds| \\ &= |\int_{t_i}^{t_{i+1}} [y'(s) - y'(t_i)]ds| \\ &\leq h \sup_{s\in[t_i,t_{i+1}]} |y'(s) - y'(t_i)| \\ &\leq h.\omega(h, y'), \end{aligned}$$

où $\omega(., y')$ est le module de continuité de la fonction (continue) y' sur $[0, T]$.
Finalement on a l'estimation suivante de l'erreur e_{i+1} :

$$|e_{i+1}| \leq (1 + Lh)|e_i| + h.\omega(h, y').$$

Pour en déduire une majoration de $|e_i|$, nous allons utiliser le Lemme de Gronwall discret suivant :

Lemme 2.1. *(Lemme de Gronwall discret)[25]. Si $(\theta_i)_i$ est une suite de réels positifs qui satisfait :*

$$\theta_{i+1} \leq (1+A)\theta_i + B,$$

où A, B sont des constantes strictement positives alors :

$$\theta_i \leq \exp(iA)\theta_0 + \frac{exp(iA)-1}{A}B.$$

On utilise d'abord le lemme avec $A = L\ h$ et $B = h.\omega(h, y')$:

$$|e_i| \leq \exp(Lih)|e_0| + \frac{\exp(Lih)-1}{Lh}h.\omega(h, y').$$

Mais $ih = t_i$ et $e_0 = 0$ donc :

$$|e_i| \leq \frac{\exp(Lt_i)-1}{L}\omega(h, y') \leq \frac{\exp(LT)-1}{L}\omega(h, y').$$

Pour conclure on a la majoration de l'erreur e_{i+1} :

$$|e_{i+1}| \leq \omega(h, y')[(1+Lh)\frac{\exp(LT)-1}{L} + h].$$

2.5 Étude générale des méthodes à un pas

On conserve, dans cette section, une grille uniforme de pas $h = \frac{T}{N}$. Les méthodes à un pas sont les méthodes de résolution numérique qui peuvent s'écrire sous la forme :

$$\begin{cases} y_{i+1} = y_i + h\Phi(t_i, y_i, h), i = 0, ..., N-1, \\ y_0 = \eta_h, \text{ donné dans } \mathbb{R}, \end{cases}$$

où Φ est une fonction continue sur $[0, T] \times \mathbb{R}^n \times [0, H]$, H désignant le pas de discrétisation maximal. Notons que la méthode d'Euler est la méthode à un pas qui correspond à

$$\Phi(t, y, h) = f(t, y); \tag{2.14}$$

Dans ce cas Φ est indépendant de h.

Propriétés importantes d'une méthode à un pas

- CONSISTANCE

  Définition 2.6. Soit $y(.)$ une solution exacte de $y'(t) = f(t, y(t))$. On appelle erreur de consistance relative à $y(.)$ de la méthode à un pas, la quantité :

$$\varepsilon_h(y) = \sum_{i=0}^{N-1} |y(t_{i+1}) - y(t_i) - h\ \Phi(t_i, y(t_i), h)|.$$

  On dit que la méthode numérique est consistante si, pour toute solution exacte y de $y'(t) = f(t, y(t))$, $\varepsilon_h(y) \to 0$ quand $h \to 0$.

  La quantité $y(t_{i+1}) - y(t_i) - h\Phi(t_i, y(t_i), h)$ est donnée à la formule (2.13).

- STABILITÉ

  Une autre notion importante est la notion de stabilité. Dans la pratique, le calcul récurrent des points y_i est en effet entaché d'erreurs d'arrondi ε_i. Pour que les calculs soient significatifs, il est indispensable que la propagation de ces erreurs reste contrôlable. Ce qui nous amène à la définition suivante.

  Définition 2.7. La méthode à un pas est dite stable s'il existe une constante $S \geq 0$ telles que, pour toutes suites (y_i), $(\tilde{y}_i)$ définies par :

$$\begin{cases} \tilde{y}_{i+1} = \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i, & 0 \leq i \leq N-1. \\ y_{i+1} = y_i + h\Phi(t_i, y_i, h), & 0 \leq i \leq N-1. \end{cases},$$

  on a :

$$\max_i |y_i - \tilde{y}_i| \leq S|y_0 - \tilde{y}_0| + \sum_{i=0}^{N-1} |\epsilon_i|.$$

- CONVERGENCE

  Une autre notion importante est la suivante.

  Définition 2.8. On dit que la méthode est convergente si

$$\max_i |y_i - y(t_i)| \to 0 \text{ quand } h \to 0. \tag{2.15}$$

  Théorème 2.3. *Toute méthode stable et consistante converge à condition que $y_0 \to y(0)$ quand $h \to 0$.*

Preuve. Si on note $\tilde{y}_i = y(t_i)$, on a par définition de ε_i :

$$\tilde{y}_{i+1} = \tilde{y}_i + h\Phi(t_i, \tilde{y}_i, h) + \varepsilon_i.$$

Avec $\tilde{y}_0 = y(0)$. Puisque la méthode est stable :

$$\max_i |y_i - y(t_i)| \leq S|y_0 - y(0)| + \sum_{i=0}^{N-1} |\varepsilon_i|.$$

Hors, les deux quantités du membre de droite tendent vers *0* quand h tend vers *0* par consistance, donc le résultat est acquis.

La dernière condition étant toujours satisfaite en pratique, l'étude de la convergence des méthodes à un pas se réduit à l'étude de leur consistance et de leur stabilité, ce qui est plus simple comme on va le voir.

2.6 Méthode de Taylor d'ordre p.

Supposons que f soit de classe C^p, alors toute solution exacte $y(.)$ est de classe C^{p+1} ; on définit des fonction $f^{(k)}$, construite par récurrence à partir de f et de ses dérivées partielles telles que $y^{(k)}(t) = f^{(k-1)}(t, y(t))$, *pour* $k = 1, \ldots, p+1$. La formule de Taylor à l'ordre $p+1$ s'écrit alors :

$$y(t_i+h) = y(t_i)+\sum_{k=1}^{p} \frac{1}{k!}h^k f^{(k-1)}(t_i, y(t_i))+\frac{1}{(p+1)!}f^{(p)}(t_i, y(t_i))h^{p+1}+o(h^{p+1}),\ i = 0, \ldots, N-1.$$

ou avec la formule de Taylor Lagrange :

$$y(t_i+h) = y(t_i)+\sum_{k=1}^{p} \frac{1}{k!}h^k f^{(k-1)}(t_i, y(t_i))+\frac{1}{(p+1)!}f^{(p)}(t_i+\theta h, y(t_i+\theta h))h^{p+1},\ i = 0, \ldots, N-1\, et\, \theta \in]0, 1[.$$

Ceci suggère le schéma numérique suivant obtenu en remplaçant les valeurs inconnues $y(t_k)$ par les y_k.

$$\begin{cases} y_{i+1} = y_i + \sum\limits_{k=1}^{p} \frac{1}{k!}h^k f^{(k-1)}(t_i, y_i), \\ t_{i+1} = t_i + h,\ i = 0, ..., N-1, \end{cases}$$

La fonction Φ associée à cette méthode est :

$$\Phi(t,y,h) = \sum_{k=1}^{p} \frac{1}{k!} h^{k-1} f^{(k-1)}(t,y)$$

Calculons l'erreur de consistance ε_i. En supposant $y_i = y(t_i)$, la formule de Taylor d'ordre $p+1$ donne

$$\begin{aligned}
\varepsilon_i &= y(t_{i+1}) - y_{i+1} \\
&= y(t_i + h) - \sum_{k=0}^{p} \frac{1}{k!} h^k y^{(k)}(t_i) \\
&= \frac{1}{(p+1)!} h^{p+1} f^{(p)}(t_i, y_i) + o(h^{p+1}).
\end{aligned}$$

L'erreur est donc de l'ordre de h^{p+1}. On dira d'une manière générale qu'une méthode est d'*ordre p* si l'erreur de consistance est en h^{p+1}, chaque fois que f est de classe C^p au moins. La méthode d'Euler est le cas particulier où $p = 1$ pour la méthode de Taylor.

2.7 Méthode du point milieu.

Notons M_i le point de coordonnées $(t_i, y(t_i))$ pour $i = 0, ..., N-1$ du graphe de y. Le segment $[M_i, M_{i+1}]$ a une pente plus proche en général de $y'(t_i + \frac{h}{2})$ (pente de la tangente au point milieu) que de $y'(t_i)$ (pente de la tangente de M_i).

On peut considérer qu'une approximation de $y(t_{i+1})$ à partir de $y(t_i)$ meilleure que l'expression $y(t_i) + hf(t_i, y(t_i))$ de la méthode d'Euler est :

$$y(t_i) + hy'(t_i + \frac{h}{2}) = y(t_i) + hf(t_i + \frac{h}{2}, y(t_i + \frac{h}{2})).$$

On prend par récurrence y_i approximation de $y(t_i)$. Comme la valeur de $y(t_i + \frac{h}{2})$ n'est pas connue, il convient d'en chercher une approximation notée $y_{i+\frac{1}{2}}$. Le schéma d'Euler suggère de prendre

$$y_{i+\frac{1}{2}} = y_i + \frac{h}{2} f(t_i, y_i).$$

On aboutit ainsi donc au schéma numérique :

$$\begin{cases}
y_{i+\frac{1}{2}} = y_i + \frac{h}{2} f(t_i, y_i), \\
p_i = f(t_i + \frac{h}{2}, y_{i+\frac{1}{2}}), \\
y_{i+1} = y_i + hp_i, \\
t_{i+1} = t_i + h,
\end{cases}$$

Ce schéma est encore une méthode à un pas dans laquelle l'expression de Φ obtenue en développant p_i est :

$$\Phi(t, y, h) = f(t_i + \frac{h}{2}, y_i + \frac{h}{2} f(t_i, y_i)).$$

2.8 Méthodes de Runge- Kutta

Ces méthodes sont les plus utilisées : elles sont implémentées dans la plupart des logiciels. On considère le problème de Cauchy suivant :

$$\begin{cases} y'(t) = f(t, y(t)), \ \forall t \in [0, T], \\ y(0) = y_0, \end{cases}$$

avec une solution exacte $y(t)$ sur $[0, T]$ et une subdivision $t_0 = 0 < t_1 < \cdots < t_N = T$. Introduisons q points intermédiaires dans chaque intervalle $[t_i, t_{i+1}]$ notés $t_{i,1}, t_{i,2}, \ldots, t_{i,q}$. On se donne $c_1, c_2, \ldots, c_q$ réels dans l'intervalle $[0, 1]$, avec $t_{i,j} = t_i + c_j h$, $i = 0, ..., N-1$. et $j = 1, ..., q$

A chacune de ces points, on associe la pente correspondante

$$p_{i,j} = f(t_{i,j}, y_{i,j}).$$

On part de l'expression intégrale de l'accroissement $y(t_{i+1}) - y(t_i)$, dans laquelle on ramène l'intervalle d'intégration de $[t_i, t_{i+1}]$, à $[0, 1]$, par le changement de variables $t = t_i + uh$:

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + \int_{t_i}^{t_{i+1}} f(t, y(t)) dt \\ &= y(t_i) + h \int_0^1 f(t_i + uh, y(t_i + uh)) du \end{aligned}$$

ou encore

$$y(t_{i+1}) = y(t_i) + h \int_0^1 g(u) du \tag{2.16}$$

avec $g(u) = f(t_i + uh, y(t_i + uh))$. On se donne alors une méthode d'intégration approchée sur $[0, 1]$ pour calculer l'intégrale qui apparaît dans l'équation (2.16) :

$$\int_0^1 g(u) du \simeq \sum_{j=1}^{q} b_j g(c_j). \tag{2.17}$$

Comme les valeurs des $g(c_j) = f(t_i + c_j h, y(t_i + c_j h)) = f(t_{i,j}, y(t_{i,j}))$ ne sont pas connues, il faut aussi évaluer la fonction y aux points $t_{i,j} = t_i + c_j h$ par un calcul similaire. On a

$$\begin{aligned} y(t_{i,j}) &= y(t_i) + \int_{t_i}^{t_{i,j}} f(t, y(t))dt \\ &= y(t_i) + h \int_0^{c_j} f(t_i + uh, y(t_i + uh))du \\ &= y(t_i) + h \int_0^{c_j} g(u)du. \end{aligned}$$

On se donne également pour chaque $j = 1, 2, \ldots, q$ une méthode d'intégration approchée

$$\int_0^{c_j} g(u)du \simeq \sum_{1 \leq k \leq j-1} a_{j,k} g(c_k), \tag{2.18}$$

En appliquant ces méthodes à $g(u) = f(t_i + uh, y(t_i + uh))$, il vient

$$\begin{aligned} y(t_{i,j}) &\simeq y(t_i) + h \sum_{1 \leq k \leq j-1} a_{j,k} f(t_{i,j}, y(t_{i,j})), \\ y(t_{i+1}) &\simeq y(t_i) + h \sum_{1 \leq j < q} b_j f(t_{i,j}, y(t_{i,j})). \end{aligned}$$

La méthode de Runge-Kutta (RK_q) correspondante est définie par l'algorithme

- $c_1 = 0$, $t_{i,1} = t_i$, $y_{i,1} = y_i$, $p_{i,1} = f(t_i, y_i)$.
- Pour $j = 2, ..., q$,

$$\begin{cases} \begin{cases} t_{i,j} = t_i + c_j h, \\ y_{i,j} = y_i + h \sum\limits_{1 \leq k < j-1} a_{j,k} p_{i,k}, \\ p_{i,j} = f(t_{i,j}, y_{i,j}) \end{cases} \\ \\ \begin{cases} t_{i+1} = t_i + h, \\ y_{i+1} = y_i + h \sum\limits_{1 \leq k \leq q} b_k p_{i,k}, \end{cases} \end{cases}$$

On la représente traditionnellement par le tableau

$$\begin{array}{cc|ccccc} (M_1) & c_1 & 0 & 0 & \dots & 0 & 0 \\ (M_2) & c_2 & a_{21} & 0 & \dots & 0 & 0 \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & & 0 & 0 \\ (M_q) & c_q & a_{q1} & a_{q2} & \cdots & a_{qq-1} & 0 \\ \hline (M) & & b_1 & b_2 & \cdots & b_{q-1} & b_q \end{array}$$

dans lequel les méthodes d'intégration approchées correspondent aux lignes. On pose par convention $a_{j,k} = 0$ pour $k \geq j$.

Exemples

- Pour $q = 1$: C'est la méthode d'Euler basée sur la méthode des rectangles. Le seul choix possible est

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

On a ici

$c_1 = 0, a_{11} = 0, b_1 = 1$. L'algorithme est donné par

$$\begin{cases} p_{i,1} = f(t_i, y_i), \\ t_{i+1} = t_i + h, \\ y_{i+1} = y_i + hp_{i,1}, \end{cases}$$

- Pour $q = 2$: on considère les tableaux de la forme

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \beta & \beta & 0 \\ \hline & 1 - \frac{1}{2\beta} & \frac{1}{2\beta} \end{array}$$

L'algorithme s'écrit ici

$$\begin{cases} p_{i,1} = f(t_i, y_i), \\ t_{i,2} = t_i + \beta h, \\ y_{i,2} = y_i + \beta h p_{i,1}, \\ p_{i,2} = f(t_{i,2}, y_{i,2}), \\ t_{i+1} = t_i + h, \\ y_{i+1} = y_i + h((1 - \frac{1}{2\beta})p_{i,1} + \frac{1}{2\beta}p_{i,2}), \end{cases}$$

ou encore, sous forme condensée :

$$y_{i+1} = y_i + h\big((1 - \frac{1}{2\beta})f(t_i, y_i) + \frac{1}{2\beta}f(t_i + \beta h, y_i + \beta h f(t, y_i))\big),$$

– Pour $\beta = \frac{1}{2}$, on retrouve la méthode du point milieu

$$y_{i+1} = y_i + hf(t_i + \frac{h}{2}, y_i + \frac{h}{2}f(t_i, y_i)),$$

qui est basée sur la formule d'intégration du point milieu :

$$(M) \qquad \int_0^1 g(t)dt \simeq g(\frac{1}{2}).$$

– Pour $\beta = 1$, on obtient la *méthode de Heun* :

$$y_{i+1} = y_i + h\big((\frac{1}{2}f(t_i, y_i) + \frac{1}{2}f(t_i + h, y_i + hf(t_i, y_i)))\big),$$

qui repose sur la formule d'intégration des trapèzes :

$$(M) \qquad \int_0^1 g(t)dt \simeq \frac{1}{2}(g(0) + g(1)).$$

- $q = 4$: la méthode de Runge-Kutta "classique" (la plus utilisée) basée sur la formule de Simpson :

0	0	0	0	0
$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
$\frac{1}{2}$	0	$\frac{1}{2}$	0	0
1	0	0	1	0
	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

La méthode s'écrit aussi :

$$\begin{cases} p_{i,1} = f(t_i, y_i), \\ t_{i,2} = t_i + \frac{1}{2}h, \\ y_{i,2} = y_i + \frac{1}{2}hp_{i,1}, \\ p_{i,2} = f(t_{i,2}, y_{i,2}), \\ y_{i,3} = y_i + \frac{1}{2}hp_{i,2}, \\ p_{i,3} = f(t_{i,3}, y_{i,3}), \\ t_{i+1} = t_i + h, \\ y_{i,4} = y_i + hp_{i,3}, \\ p_{i,4} = f(t_{i,4}, y_{i,4}), \\ y_{i+1} = y_i + h(\frac{1}{6}p_{i,1} + \frac{2}{6}p_{i,2} + \frac{2}{6}p_{i,3} + \frac{1}{6}p_{i,4}), \end{cases}$$

Cette méthode est d'ordre 4. Dans ce cas les méthodes d'intégration (2.17) et (2.18) utilisées sont respectivement :

$$(M_2) \qquad \int_0^{\frac{1}{2}} g(t)dt \simeq \frac{1}{2}g(0) : \qquad \text{formule des rectangles à gauche,}$$

$$(M_3) \qquad \int_0^{\frac{1}{2}} g(t)dt \simeq \frac{1}{2}g(\frac{1}{2}) : \qquad \text{formule des rectangles à droite,}$$

$$(M_3) \qquad \int_0^{1} g(t)dt \simeq g(\frac{1}{2}) : \qquad \text{formule du point milieu,}$$

$$(M) \qquad \int_0^{1} g(t)dt \simeq \frac{1}{6}g(0) + \frac{2}{6}g(\frac{1}{2}) + \frac{2}{6}g(\frac{1}{2}) + \frac{1}{6}g(1) : \qquad \text{formule de simpson.}$$

Remarque 2.1. Dans MATLAB, *ode*45 résout l'équation différentielle ordinaire par une méthode d'ordre 4 similaire à la méthode de Runge-Kutta classique d'ordre 4 ; en plus,

*ode*45 utilise un pas de temps variable et choisit à chaque instant le pas le plus convenable de façon à satisfaire une tolérance fixée.

Chapitre 3

Introduction à la commande optimale

3.1 Introduction

La théorie du contrôle analyse les propriétés des systèmes commandés, c'est-à-dire des systèmes dynamiques sur lesquels on peut agir au moyen d'une commande (ou contrôle).

Le but est alors d'amener le système d'un état initial donné à un certain état final, en respectant certains critères ; les systèmes abordés sont multiples : systèmes différentiels, systèmes discrets, systèmes avec bruit, systèmes avec retard, etc. Leurs origines sont très diverses : mécanique, électrique, électronique, biologique, chimique, économique, etc. L'objectif peut être de stabiliser le système pour le rendre insensible à certaines perturbations (stabilisation), ou encore de déterminer des solutions optimales pour un certain critère d'optimisation (contrôle optimal, ou commande optimale).

Dans ce chapitre nous rappelons la formulation d'un problème général de commande optimale, et nous présentons la notion de contrôlabilité pour les systèmes linéaires et les systèmes non linéaires. La dernière section du chapitre est consacrée à l'énoncé général du principe du maximum de Pontryagin.

3.2 Formulation générale d'un problème de contrôle optimal

La formulation d'un problème de contrôle optimal exige une description mathématique du processus à contrôler, une proclamation des contraintes physiques et la détermination du

critère de performance. Après modélisation, on obtient un système comportant beaucoup de variables et de paramètres.

Les variables, nommées variables d'état seront notées $x_i,\ i = 1, ..., n$. Le système évolue dans le temps, donc les x_i sont des fonctions de t : $x_i(t),\ i = 1, ..., n$, où t désigne le temps défini dans un intervalle $[0, T]$. Les n variables d'état vont être gouvernées par n équations différentielles du premier ordre ; elles sont sous la forme :

$$\dot{x}(t) = \frac{dx}{dt} = f(t, x, u),$$

où f est un vecteur de n composantes $f_i,\ i = 1, ..., n$.
Les variables de contrôle seront notées $u_j(t),\ j = 1, ..., m$; elles doivent être intégrables par rapport à t. On définit aussi l'ensemble des commandes admissibles U qui peut être non borné, borné ou du type Bang-Bang.

Commande bornée

Dans beaucoup de problèmes de contrôle, on peut minorer et majorer les $u_j(t)$ par des constantes. Dans la suite, nous considérons ce type de problème avec $a_j \leq u_j \leq b_j$. Notons que l'on peut remplacer u_j par v_j en posant $u_j = \frac{1}{2}(a_j + b_j) + \frac{1}{2}(a_j - b_j)v_j$ et ainsi v_j est aussi intégrable et l'on a $-1 \leq v_j \leq 1$. Donc lorsque U est borné, il est toujours pratique de se ramener à des commandes entre -1 et 1.

Commande Bang-Bang

On suppose que U est un polyèdre (cube) $[-1, 1]^m$ dans $\mathbb{R}^m$. Un contrôle $u \in U$ est appelé contrôle Bang-Bang si pour chaque instant t et chaque indice $j = 1, ..., m$, on a $|u_j(t)| = 1$. En d'autres termes, une commande Bang-Bang est une commande qui possède au moins un switch.

3.3 Contrôle optimal

Position du problème

Un problème de contrôle optimal se formule comme suit :

$$J(x,u) = g(T,x(T)) + \int_0^T f_0(t,x(t),u(t))dt \to \min_u, \tag{3.1}$$

$$\dot{x}(t) = f(t,x(t),u(t)), \tag{3.2}$$

$$x(0) = x_0 \in M_O, \tag{3.3}$$

$$x(T) = x_1 \in M_1, \tag{3.4}$$

$$u \in U,\ t \in I = [0,T], \tag{3.5}$$

où M_0 et M_1 sont deux variétés de $\mathbb{R}^n$, I un intervalle de $\mathbb{R}$, $x_0 = x(0)$ est la position initiale du système (3.2), $x(T)$ est sa position terminale. En pratique, l'état du système peut représenter à la fois la vitesse, la position, la température et d'autres paramètres mesurables. U est l'ensemble des applications mesurables, localement bornées sur I à valeurs dans $U \subset \mathbb{R}^m$. Le but de la commande consiste à ramener l'objet de la position initiale $x_0 \in M_0$ à une autre position $x_1 \in M_1$, où M_0 est l'ensemble de départ, et M_1 l'ensemble d'arrivée. Le but est d'optimiser la fonction décrite par la formule suivante :

$$J(x,u) = g(T,x(T)) + \int_0^T f_0(t,x(t),u(t))dt.$$

On appelle $J(x,u)$ le coût du contrôle ou fonction objectif. Cette fonctionnelle comporte deux parties : $g(T,x(T))$ est le coût terminal, c'est une sorte de pénalité liée à la fin de l'évolution du système au temps final T ; il a son importance lorsque T est libre, sinon il est constant. Le second terme intervenant dans la fonction objectif $\int_0^T f_0(t,x(t),u(t))dt$ dépend de l'état du système tout au long de la trajectoire de la solution, définie par les variables d'état. Cette trajectoire dépend aussi du temps t mais surtout des variables de contrôle u. C'est une fonction d'efficacité de chaque commande sur l'intervalle T.

On dinstigue trois problèmes importants :

Problème de Lagrange

C'est le problème dont le critére à minimiser est égal à :

$$J(x,u) = \int_0^T f_0(t,x(t),u(t))dt,$$

c'est à dire $g = 0$.

Problème de Mayer

Dans ce cas le coût s'écrit :

$$J(x,u) = g(T,x(T)),$$

c'est à dire $f_0 = 0$, $J(x,u)$ est le coût terminal.

Problème de Mayer-Lagrange

Le problème de Mayer-Lagrange est donné sous la forme suivante :

$$J(x,u) = g(T,x(T)) + \int_0^T f_0(t,x(t),u(t))dt,$$

3.4 Contrôlabilité

L'objectif d'un problème de contrôle est d'amener le système d'un état initial donné à un état final tout en respectant certaines contraintes. Plus précisément on pose la définition suivante :

Définition 3.1. le système $\dot{x}(t) = f(t,x(t),u(t))$, $x(0) = x_0$ est dit contrôlable si pour tous points $x_0 \in M_0$ et $x_1 \in M_1$, il existe un contrôle $u(.)$ tel que la trajectoire associée à u relie x_0 à x_1 en un temps fini.

La notion de contrôlabilité a été introduite en 1960 par Kalman [48] pour des systèmes linéaires de la forme $\dot{x} = Ax+Bu$. Pour les systèmes non linéaires, le problème mathématique de contrôlabilté est beaucoup plus compliqué. Il constitue un domaine de recherche actif.

3.4.1 Contrôlabilité des systèmes linéaires

Considérons le système de contrôle linéaire suivant :

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0,\ \forall t \in I, \tag{3.6}$$

où I est un intervalle de $\mathbb{R}$, A,B, et r sont trois applications localement intégrables sur I à valeurs respectivement dans $\mathcal{M}_n(\mathbb{R})$, $\mathcal{M}_{n,m}(\mathbb{R})$ et $\mathbb{R}^n$, où $\mathcal{M}_n(\mathbb{R})$ est l'ensemble des matrices réelles de dimension n, et $\mathcal{M}_{n,m}(\mathbb{R})$ est l'ensemble des matrices de n lignes et de m colonnes.
L'ensemble des contrôles u considérés est l'ensemble des applications mesurables localement bornées sur I à valeurs dans un sous ensemble $U \subset \mathbb{R}^m$.

Les théorèmes d'existence de solutions d'équations différentielles nous assurent que, pour tout contrôle u, le système (3.6) admet une unique solution $x(.) : I \to \mathbb{R}^n$, absolument continue. Soit $M(.) : I \to \mathcal{M}_n(\mathbb{R})$ la résolvante du système linéaire homogène $\dot{x}(t) = A(t)x(t)$, définie par $\dot{M}(t) = A(t)M(t)$, $M(0) = Id$. Alors, la solution $x(.)$ du système (3.6) associée au contrôle u est donnée par

$$x(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u(s) + r(s))ds,$$

pour tout $t \in I$.

Contrôlabilité des systèmes linéaires stationnaires

Le système $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$ est dit stationnaire lorsque les matrices A et B ne dépendent pas de t. Dans ce cas, $A(t) = A$, $B(t) = B$ sont des constantes sur I, alors $M(t) = e^{At}$.

Le théorème suivant donne une condition nécessaire et suffisante de contrôlabilité dans le cas sans contrainte sur le contrôle

Théorème 3.1. *[48]. On suppose que $U = \mathbb{R}^m$. Le système stationnaire $\dot{x}(t) = Ax(t) + Bu(t) + r(t)$ est contrôlable en temps T (quelconque) si et seulement si la matrice*

$$C = [B, AB, A^2B, ..., A^{n-1}B]$$

est de rang n.

La matrice C est appelée matrice de Kalman, et la condition $rang\ C = n$ est appelée condition de Kalman.

Remarque 3.1. La condition de Kalman ne dépend ni de T ni de x_0. Autrement dit, si un système linéaire stationnaire est contrôlable en temps T depuis x_0, alors il est contrôlable en tout temps depuis tout point.

Notons que la matrice C est de rang n si et seulement si l'application linéaire

$$\Phi : L^\infty([0,T],\mathbb{R}^m) \to \mathbb{R}^n$$

$$u \mapsto \int_0^T e^{(T-t)A} B u(t) dt$$

est surjective.

Contrôlabilité des systèmes linéaires non stationnaires

Le théorème suivant donne une condition nécessaire et suffisante de contrôlabilité dans le cas non stationnaire.

Théorème 3.2. *Le système* $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$ *est contrôlable en temps* T *si et seulement si la matrice*

$$C(T) = \int_0^T M(t)^{-1} B(t) B(t)^t M(t)^{-1} dt,$$

dite matrice de contrôlabilité, est inversible.

Remarque 3.2. Cette condition dépend de T , mais ne dépend pas du point initial x_0. Autrement dit, si un système linéaire non stationnaire est contrôlable en temps T depuis x_0, alors il est contrôlable en temps T depuis tout point.

3.4.2 Contrôlabilité des systèmes non-linéaires

Pour les systèmes de contrôle non linéaires, il est impossible d'étudier la contrôlabilité globale ; le problème est beaucoup plus compliqué du fait qu'on ne peut pas utiliser la caractérisation de Kalman. Dans ce cas, on s'intéressera à l'étude de la contrôlabilité locale du système $\dot{x} = f(t, x, u)$, $x(0) = x_0$, où la fonction f est C^1 sur $\mathbb{R}^{1+n+m}$.

Proposition 3.1. *Considérons le système $\dot{x}(t) = f(t, x(t), u(t))$, $x(0) = x_0$ avec $f(x_0, u_0) = 0$. On note $A = \frac{\partial f}{\partial x}(x_0, u_0)$ et $B = \frac{\partial f}{\partial u}(x_0, u_0)$. Si*

$$rang(B, AB, A^2B, ..., A^{n-1}B) = n,$$

alors le système est localement contrôlable en x_0.
En général, le problème de contrôlabilité est difficile. Cependant, il existe des techniques qui permettent de déduire la contrôlabilité locale dans le cas des systèmes linéarisés.

3.5 Problème de contrôle optimal

Un problème de contrôle optimal se décompose en deux parties : pour déterminer une trajectoire optimale joignant un ensemble initial à une cible, il faut d'abord savoir si cette cible est atteignable. C'est le problème de contrôlabilité. Ensuite, une fois ce problème résolu, il faut chercher parmi toutes ces trajectoires possibles celles qui minimisent un certain critère ; on parle alors d'un problème de contrôle optimal. Historiquement, la théorie du contrôle optimal est très liée à la mécanique classique, en particulier aux principes variationnels de la mécanique. Le point clé de cette théorie est le principe du maximum de Pontryagin, formulé par L. S. Pontryagin en 1956, qui donne une condition nécessaire d'optimalité et permet ainsi de calculer les trajectoires optimales.

Le théorème suivant est l'énoncé général du principe du maximum de Pontryagin.

Théorème 3.3. *[73]. Considérons le système de contrôle dans $\mathbb{R}^n$*

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{3.7}$$

où $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ est de classe C^1, les contrôles sont des applications mesurables bornées à valeurs dans $U \subset \mathbb{R}^m$. Soient M_O et M_1 deux sous ensembles de $\mathbb{R}^n$. Notons par $U(t)$ l'ensemble des contrôles admissibles u dont les trajectoires associées relient un point initial de M_0 à un point final de M_1 en temps t.

On définit le coût

$$J(x,u) = \int_0^T f_0(t, x(t), u(t))dt, \tag{3.8}$$

où $f_0 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ de classe C^1, $x(.)$ est la solution de (3.7) associée au contrôle u.

S'il n'existe pas de contrôle $u \in U$ satisfaisant le système $\dot{x} = f(t,x,u)$, $x(0) = x_0$ et $x(T) = x_1$, on dit que le système n'est pas contrôlable. Dans ce cas, le problème n'admet pas de solution. Si le système est contrôlable, il existe en général beaucoup de contrôles possibles et pour chacun de ces contrôles correspond une valeur pour J. Le problème est de déterminer un contrôle optimal $u^* \in U$ associé à des trajectoires optimales x^* et qui minimise le coût J. Le temps final peut être fixé ou non.

Si le contrôle $u \in U$ associé à la trajectoire $x(.)$ est optimal sur $[0,T]$, alors il existe une application $p(.) : [0,T] \to \mathbb{R}^n$ absolument continue, appelée vecteur adjoint, et un réel $p^0 \leq 0$ tel que le couple $(p(.), p^0)$ est non trivial et tels que pour presque tout $t \in [0,T]$,

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t, x(t), p(t), p^0, u(t)) \tag{3.9}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t), p^0, u(t)) \tag{3.10}$$

où $H(t,x,p,p^0,u) = p^t(t)f(t,x,u) + p^0 f_0(t,x,u)$ est l'Hamiltonien du système ; on a alors la condition du maximum presque partout sur $[0,T]$

$$H(t, x(t), p(t), p^0, u(t)) = \max_{v \in U} H(t, x(t), p(t), p^0, v) \tag{3.11}$$

Si de plus le temps final pour joindre M_1 n'est pas fixé, on a la condition au temps final T

$$\max_{v \in U} H(T, x(T), p(T), p^0, v) = 0. \tag{3.12}$$

Si de plus M_0 et M_1 (ou juste l'un des deux ensembles) sont des variétés de $\mathbb{R}^n$ ayant des espaces tangents en $x(0) = x_0 \in M_0$ et $x(T) = x_1 \in M_1$, alors le vecteur adjoint peut être construit de manière à vérifier les conditions de transversalités aux deux extrémités (ou juste l'une des deux)

$$p(0) \perp T_{x(0)} M_0, \tag{3.13}$$

$$p(T) \perp T_{x(T)} M_1. \tag{3.14}$$

Remarque 3.3. La convention $p_0 \leq 0$ conduit au principe du maximum, tandis que $p_0 \geq 0$ conduit au principe du minimum.

Remarque 3.4. Dans le cas où il n'y a pas de contrainte sur le contrôle ($U = \mathbb{R}^m$), la condition du maximum (3.11) devient $\frac{\partial H}{\partial u} = 0$.

Chapitre 4

Méthode de Tir et méthode de Newton

4.1 Introduction

Dans ce chapitre, on présente deux types de méthode numérique pour résoudre le problème de contrôle optimal : **les méthodes directes** et **les méthodes indirectes**. Ces méthodes peuvent traiter un problème de contrôle optimal plus général qui cherche par exemple à minimiser une fonction non linéaire. Parmi les méthodes directes, on trouve la méthode de résolution par l'approche de la programmation linéaire, qui est la méthode adaptée appelée aussi méthode du support [1, 58, 71]. Une autre méthode directe est la méthode de discrétisation du problème initial. Elle consiste à discrétiser la solution du système et le contrôle en transformant le problème de contrôle optimal en un problème d'optimisation non linéaire sous contraintes. Les méthodes indirectes sont basées sur une méthode de tir aprés l'application du principe du maximum de Pontryagin [72, 73]. Ces méthodes ont l'extrême précision numérique, mais elles sont très sensibles au choix de la condition initiale. L'algorithme de Tir nécessite, l'utilisation de la méthode de Newton.

On se réfère au livre de Emmanuel Trélat [78] et à l'article de O. Von Stryk et R. Bulirsch [81] pour plus de détails sur les différentes méthodes numériques.

4.2 Méthodes indirectes

Méthode de tir simple

Considérons le problème de contrôle optimal $(3.1) - (3.5)$, et supposons le temps final T fixé. La méthode de tir permet d'obtenir la valeur de $p(0)$ nécessaire à la résolution du problème à résoudre qui est caractérisé par l'application du principe du maximum ou du minimum de Pontryagin. Cette méthode permet de résoudre un système d'équations non linéaires

$$x^{p_0}(T) - x_f = 0,$$

où $x^{p_0}(t)$ est obtenu en résolvant le système d'équations :

$$\begin{cases} \frac{dx(t)}{dt} = \frac{\partial H}{\partial p} \\ \\ -\frac{dp(t)}{dt} = \frac{\partial H}{\partial x} \end{cases} \quad \forall t \in [0, T],$$

avec les conditions initiales $x(0) = x_0$ et $p(0) = p_0$.

Notons que ce système peut être résolu en utilisant un intégrateur tel que les méthodes d'Euler ou Runge Kutta ; ainsi, de la valeur initiale p_0, il est possible de construire la solution du problème de commande optimale correspondant.

Par la notation $G(p_0) = x^{p_0}(T) - x_f$, définissons G la fonction implicite de $\mathbb{R}^n$ dans $\mathbb{R}^n$. Il s'agit de déterminer un zéro de cette équation ; pour la résoudre, on utilisera la méthode de Newton.

Remarque 4.1. Si le temps final T est libre, on peut se ramener à la formulation précédente, en considérant T comme une inconnue auxiliaire. On augmente alors la dimension de l'état en considérant l'équation supplémentaire $\frac{dT}{dt} = 0$.

On peut utiliser le même artifice si le contrôle est bang-bang, pour déterminer les temps de commutation. Il peut cependant s'avérer préférable, lorsque le temps final est libre, d'utiliser la condition de transversalité sur l'Hamiltonien.

Méthode de Tir multiple

Par rapport à la méthode de tir simple, la méthode de tir multiple découpe l'intervalle $[0, T]$ en N intervalles $[t_i, t_{i+1}]$, et se donne comme inconnues les valeurs $z(t_i)$ au début de

chaque sous-intervalle. Il faut prendre en compte des conditions de recollement en chaque temps t_i (conditions de continuité). L'intérêt est d'améliorer la stabilité de la méthode.

De manière plus précise, considérons un problème de contrôle optimal général. L'application du principe du maximum réduit le problème à un problème aux valeurs limites du type

$$\dot{z}(t) = F(t, z(t)) = \begin{cases} F_0(t, z(t)), & \text{si } t_0 \leq t < t_1\,; \\ F_1(t, z(t)), & \text{si } t_1 \leq t < t_2\,; \\ \vdots & \\ F_N(t, z(t)), & \text{si } t_N \leq t < T, \end{cases} \tag{4.1}$$

où $z = (x, p) \in \mathbb{R}^{2n}$, x est l'état, p est l'état adjoint, et $t_1, t_2, \ldots, t_N \in [0, T]$ sont les temps de commutation.

Lorsqu'il est possible de dériver une expression analytique de la commande optimale, la méthode de tir multiple ([22], [32]) se trouve être bien adaptée. Elle consiste à résoudre le problème d'optimisation comme dans la méthode de tir simple, avec des conditions de continuité (ou conditions de jonction) sur les variables d'état et d'état adjoint, lors d'un changement du système différentiel.

Ce type d'algorithme se trouve donc bien adapté à la résolution de problèmes d'optimisation avec contraintes d'état. En effet, dans ce cas les temps $t_1, t_2, \ldots, t_N \in [0, T]$ peuvent être des temps de jonction avec un arc frontière (contrainte sur l'état active pendant un temps donné), ou bien des temps de contact avec la frontière (correspondant à des contraintes saturées sur l'état).

Un très bon exemple se trouve dans [15], exemple qui est repris dans [75] : il s'agit de connaître la trajectoire optimale d'une corde élastique attachée à ses deux extrémités, et qui repose sur un support plan. Lorsque la corde ne touche pas le support, l'équation représentant la trajectoire de la corde est parfaitement connue. En revanche, lorsque la corde touche le support, il devient difficile de connaître sa trajectoire, notamment à quels endroits elle touche le support puis le quitte.

On peut alors découper la trajectoire en 3 tronçons. Chaque partie est décrite par une équation différentielle différente. Les temps de commutation représentent alors les abscisses auxquels la corde touche puis quitte le support. Dans certains problèmes très simples, on

peut résoudre le problème, en écrivant les conditions d'optimalité associées. Différents cas sont énumérés dans [22] : système continu avec état final fixé à un temps donné, état final fixé à un temps final libre (incluant les problèmes à temps minimal), etc. L'auteur résout les problèmes posés en adjoignant l'équation d'état et les contraintes associées à l'état, à la fonction à minimiser, puis définit les conditions d'optimalité associées au problème via une méthode de perturbation. Le lecteur pourra aussi se référer à [76] pour des applications relatives aux problèmes économiques.

Dans le cas de contraintes du premier ordre sur l'état, il a été prouvé dans [15] a prouvé que l'état adjoint était continu sous certaines hypothèses. Pour des contraintes d'ordre supérieur cependant, des conditions de saut sur l'état adjoint doivent être considérées dans le problème d'optimisation, la continuité de celui-ci n'étant pas assurée.

L'inconvénient majeur de ces méthodes provient du fait qu'il soit nécessaire de connaître la forme de la trajectoire optimale pour pouvoir intégrer une équation différentielle valide, ce qui revient à connaître le nombre de contraintes actives.

4.3 Méthodes directes

Discrétisation totale : méthode directe

C'est la méthode la plus simple lorsqu'on aborde un problème de contrôle optimal. En discrétisant l'état et le contrôle, on se ramène à un problème d'optimisation non linéaire en dimension finie de la forme

$$\min_{Z \in C} F(Z), \tag{4.2}$$

où $Z = (x_1, x_2, ..., x_N, u_1, u_2, ..., u_n)$, et

$$C = \{Z / g_i(Z) = 0,\ i \in 1, ..., r, g_j(Z) \leq 0,\ j \in r+1, ..., m\}. \tag{4.3}$$

Plus précisément, la méthode consiste à choisir les contrôles dans un espace de dimension finie, et à utiliser une méthode d'intégration numérique des équations différentielles. Considérons donc une subdivision $0 = t_0 < t_1 < t_2 < .... < t_N = T$ de l'intervalle $[0, T]$. Réduisons l'espace des contrôles en considérant (par exemple) les contrôles constants par

morceaux selon cette subdivision. Par ailleurs, choisissons une discrétisation de l'équation différentielle, par exemple choisissons ici (pour simplifier) la méthode d'Euler. On obtient alors, en posant

$$h_i = t_{i+1} - t_i,$$

$$x_{i+1} = x_i + h_i f(t_i, x_i, u_i).$$

Remarque 4.2. Il existe une infinité de méthodes d'intégration numérique. D'une part, on peut discrétiser l'ensemble des contrôles admissibles par des contrôles constants par morceaux, ou affines par morceaux, ou par des fonctions splines, etc. D'autre part, il existe de nombreuses méthodes pour discrètiser une équation différentielle ordinaire : méthode d'Euler (explicite ou implicite), méthode du point milieu, méthode de Heun, méthode Runge-Kutta, méthode d'Adams Moulton, etc [25]. De plus l'introduction d'éventuelles contraintes sur l'état ne pose aucun problème.

La discrétisation précédente conduit donc au problème de programmation non linéaire

$$x_{i+1} = x_i + h_i f(t_i, x_i, u_i),\ i = 0, ..., N-1,$$

$$\min\ C(x_0, x_1, ..., x_N, u_0, u_1, ..., u_N),$$

$$u_i \in U,\ i = 0, ..., N-1,$$

qui est un problème du type (4.2).

D'un point de vue général, cela revient à choisir une discrètisation des contrôles, ainsi que de l'état, dans des espaces de dimension finie :

$$u \in Vect(U_1, ..., U_N),\ \text{i.e. } u(t) = \sum_{i=1}^{N} u_i U_i(t),\ u_i \in \mathbb{R},$$

$$x \in Vect(X_1, ..., X_N),\ \text{i.e. } x(t) = \sum_{i=1}^{N} x_i X_i(t),\ x_i \in \mathbb{R},$$

où les $U_i(t)$ et $X_i(t)$ représentent une base. Typiquement, on peut choisir des approximations polynomiales par morceaux. L'équation différentielle, ainsi que les éventuelles

contraintes sur l'état ou le contrôle, ne sont vérifiées que sur les points de discrétisation. On se ramène bien à un problème d'optimisation non linéaire en dimension finie de la forme (4.2).

La résolution numérique d'un problème de programmation non linéaire du type (4.2) est standard. Elle peut être effectuée, par exemple, par une méthode de pénalisation, ou par une méthode SQP (séquentiel quadratic programming). Dans ces méthodes, le but est de se ramener à des sous-problèmes plus simples, sans contraintes, en utilisant des fonctions de pénalisation pour les contraintes, ou bien d'appliquer les conditions nécessaires de Kuhn-Tucker pour des problèmes d'optimisation avec contraintes. Pour le problème (4.2), (4.3), les conditions de Kuhn-Tucker s'écrivent

$$\nabla F(Z) + \sum_{i=1}^{m} \lambda_i \nabla g_i(Z) = 0,$$

où les multiplicateurs de Lagrange λ_i vérifient

$$\lambda_i g_i(Z) = 0,\ i \in \{1, ..., r\},\ et\ \lambda_i \geq 0,\ i \in \{r+1, ..., m\}.$$

Les méthodes SQP consistent à calculer de manière itérative ces multiplicateurs de Lagrange, en utilisant des méthodes de Newton ou quasi-Newton. A chaque itération, on utilise une méthode de quasi-Newton pour estimer le hessien du Lagrangien associé au problème de programmation non linéaire, et on résout un sous-problème de programmation quadratique du Lagrangien.

4.4 Méthode de Newton discrète

Soit $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ une application, ou D est un ouvert. On définit un système non linéaire implicite de n équations à n inconnues :

$$F(x_0) = 0. \tag{4.4}$$

Pour le résoudre, on utilisera la méthode de Newton. Le principe de la méthode est le suivant : à une étape k donné, soit x_0^k une approximation d'un zéro x_0 de F ; donc x_0

s'écrit $x_0 = x_0^k + \Delta x_0^k$, et on a alors :

$$0 = F(x_0) = F(x_0^k + \Delta x_0^k) = F(x_0^k) + F'(x_0^k).(x_0 - x_0^k) + o(x_0 - x_0^k),$$

ce qui entraîne la résolution de

$$F'(x_0^k).(x_0 - x_0^k) = -F(x_0^k),$$

où $F'(x_0^k)$ est la matrice Jacobienne de l'application $x_0 \to F(x_0)$ calculée quand $x_0 = x_0^k$; or on ne connait la fonction $x_0 \to F(x_0)$ que numériquement. On va donc utiliser un procédé de dérivation numérique basé sur la méthode des différences finies. Pour éviter le calcul de $F'(x_0^k)$, il suffit de trouver une approximation de $F'(x_0^k)$; conformément à [70], on utilise l'une ou l'autre des approximations par différences finies.

$$\frac{\partial F_i}{\partial x_{0_j}}(x_0^k) \approx \frac{1}{h_{ij}}[F_i(x_0 + \sum_{k=1}^{j} h_{ik}e^k) - F_i(x_0 + \sum_{k=1}^{j-1} h_{ik}e^k)],$$

ou bien

$$\frac{\partial F_i}{\partial x_{0_j}}(x_o^k) \approx \frac{1}{h_{ij}}[F_i(x_0 + h_{ij}e^j) - F_i(x_0)],$$

où les h_{ij} sont des paramètres de discrétisation correspondant à la i^{eme} équation et à la j^{eme} variable, et e^k est le k^{eme} vecteur de la base canonique ; notons que, classiquement, on peut toujours choisir les valeurs de h_{ij} égales entre elles à une valeur h. Soit $\Delta_{ij}(x_0, h)$ une approximation par différences finies consistante ; alors, on a :

$$\lim_{h \to 0} \Delta_{ij}(x_0, h) = \frac{\partial F_i}{\partial x_{0j}}(x_0), i, j = 1, ..., n.$$

On pose,

$$J(x_0, h) = (\Delta_{ij}(x_0, h)),$$

qui est une approximation de la matrice Jacobienne. De manière générale, on a à considerer à chaque itération :

$$x_0^{k+1} = x_0^k - J(x_0^k, h^k)^{-1}.F(x_0^k), \quad k = 0, 1, ..., \tag{4.5}$$

Définition 4.1. Soit $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ une application G-différentiable sur $D_0 \subset D$ et $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n)$. Alors J est une approximation consistante de F' sur $D_0 \subset D_J$, si $0 \in \mathbb{R}^m$ est un point limite de D_h et

$$\lim_{h \to 0;\ h \in D_h} J(x,h) = F'(x), \text{ uniformément pour } \quad x \in D_0. \tag{4.6}$$

Si de plus, il existe c et $r > 0$ tel que

$$\|F'(x) - J(x,h)\| \leq c\|h\|, \ \forall x \in D_0, \ h \in D_h \cap S(0,r), \tag{4.7}$$

Alors J est une approximation fortement consistante de F' sur D_0.

La base des résultats dans cette section est le lemme suivant.

Lemme 4.1. *Supposons que $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ est G-différentiable sur un voisinage ouvert $S_0 \subset D$ contenant $x_0 \in D$ pour lequel $F(x_0) = 0$. De plus on suppose que F' est continue au point x_0 et $F'(x_0)$ non singulière. Soit $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n)$ une approximation consistante de F' sur S_0. Il existe alors $\delta > 0$ et $r > 0$ tel que l'application*

$$G(x,h) = x - (J(x,h))^{-1}F(x). \tag{4.8}$$

soit bien définie pour tout $x \in S = S(x_0, \delta), \ h \in D'_h = D_h \cap S(0,r)$, et satisfait

$$\|x_0 - G(x,h)\| \leq w(x,h)\|x - x_0\|, \ \forall x \in S, \ h \in D'_h, \tag{4.9}$$

où

$$w(x,h) \to 0 \quad x \to x_0 \ et \quad h \to 0, \ h \in D'_h. \tag{4.10}$$

De plus, si J est une approximation fortement consistante de F' sur S_0 et si

$$\|F'(x) - F'(x_0)\| \leq \gamma\|x - x_0\|, \ \forall x \in S_0. \tag{4.11}$$

alors il existe deux constentes $\alpha_1, \ \alpha_2$ tel que :

$$\|x_0 - G(x,h)\| \leq \alpha_1\|x - x_0\|^2 + \alpha_2\|h\|\|x - x_0\|, \ \forall x \in S, \ h \in D'_h. \tag{4.12}$$

Preuve. On pose $\beta = \|F'(x_0)^{-1}\|$ et soit $\varepsilon \in (0, \frac{1}{2}\beta^{-1})$. J étant une approximation consistante sur S_0 , il existe $r > 0$ tel que D'_h soit non vide et

$$\|F'(x) - J(x,h)\| \leq \frac{1}{2}\varepsilon, \ \forall x \in S_0, \ h \in D'_h.$$

De plus, d'après la continuité de F' au point x_0, il existe $\delta > 0$ tel que $S = S(x_0, \delta) \subset S_0$ et

$$\|F'(x) - F'(x_0\| \leq \frac{1}{2}\varepsilon, \ \forall x \in S.$$

D'où

$$\|F'(x_0) - J(x,h)\| \leq \varepsilon, \ \forall x \in S, \ h \in D'_h.$$

En effet

$$\|F'(x_0) - F'(x) + F'(x) - J(x,h)\| \leq \|F'(x) - F'(x_0)\| + \|F'(x) - J(x,h)\| \leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon$$

D'après le lemme de Perturbation **4.8** de l'annexe 1, alors $(J(x,h))^{-1}$ existe et vérifie :

$$\|J(x,h)^{-1}\| \leq \eta = \frac{\beta}{1 - \beta\varepsilon}, \quad \forall x \in S, \ h \in D'_h.$$

G est bien définie sur $S \times D'_h$ et

$$\begin{aligned}
\|G(x,h) - x_0\| &= \|J(x,h)^{-1}[J(x,h)(x - x_0) - F(x)]\| \\
&= \|J(x,h)^{-1}[J(x,h)(x - x_0) - F'(x)(x - x_0) + F'(x)(x - x_0) \\
&- F'(x_0)(x - x_0) + F'(x_0)(x - x_0) - F(x_0) - F(x)]\| \\
&\leq \eta[\|J(x,h) - F'(x)\| + \|F'(x) - F'(x_0)\|]\|x - x_0\| \\
&+ \eta[\|F(x) - F(x_0) - F'(x_0)(x - x_0)\|].
\end{aligned}$$

et la relation (4.9) est vérifiée avec

$$w(x,h) = \eta[\|J(x,h) - F'(x)\| + \|F'(x) - F'(x_0)\| + q(x)], \tag{4.13}$$

où

$$q(x) = \frac{\|F(x) - F(x_0) - F'(x_0)(x - x_0)\|}{\|x - x_0\|}, \text{ pour } x \neq x_0, q(x_0) = 0;$$

donc la relation (4.10) est vérifiée. On a besoin de la continuité de F' au point x_0 pour assurer que l'application $q(x) \to 0$ et que $F'(x) - F'(x_0) \to 0$ quand $x \to x_0$, alors que la convergence uniforme dans la définition **4.1** implique $J(x, h) - F'(x) \to 0$ quand $h \to 0$ et $x \to x_0$.

A présent, supposons que la relation (4.11) soit vérifiée. Alors, d'après le lemme **4.4** de l'annexe 1, on obtient :

$$\|q(x)\| \leq \gamma \|x - x_0\|, \ \forall x \in S \tag{4.14}$$

avec

$$\gamma = \sup_{0 \leq t \leq 1} \frac{\|F'(x_0 + t(x - x_0)) - F'(x)\|}{\|x - x_0\|};$$

alors la relation (4.12) est vérifiée immédiatement grâce aux relations (4.9) et (4.13) avec $\alpha_1 = 2\gamma\eta$ et $\alpha_2 = \eta c$, où c est la constante intervenant dans la relation (4.7).

Comme première application du lemme **4.1** , nous donnons le résultat simple suivant, qui prouve que la vitesse de la convergence de la suite (4.5) est superlinéaire quand $\lim_{k \to \infty} h^k = 0$.

Corollaire 4.1. *Supposons que $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ est $G-$différentiable dans un voisinage ouvert $S_0 \subset D$ de $x_0 \in D$ tel que $F(x_0) = 0$, et que F' soit continue au point x_0 où $F'(x_0)$ est non singulière. Soit $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n)$ une approximation consistante de F' sur S_0. Alors il existe une boule $S_1 = S(x_0, \delta_1) \subset S_0$ et un nombre réel $r_1 > 0$ tel que pour tout $x^0 \in S_1$ et pour toute suite $\{h^k\} \subset D_h \cap S(0, r_1)$, les itérations $\{x^k\}$ donnée par la relation* (4.5) *appartiennent toujours à S_1 et convergent vers x_0. De plus, si*

$$\lim_{h \to \infty} h^k = 0,$$

alors

$$R_1\{x^k\} = Q_1\{x^k\} = 0.$$

Preuve. Soit $\delta > 0$ et $r > 0$ des constantes obtenues au lemme **4.1** ; alors, pour $\alpha \in]0,1[$ donné, la relation (4.10) assure qu'on peut choisir $\delta_1 \leq \delta$ et $r_1 \leq r$ tel que

$$w(x,h) \leq \alpha,\ \forall x \in S(x_0,\delta_1),\ h \in D_h \cap S(0,r_1)$$

D'où, l'existence et la convergence d'une suite $\{x^k\}$ d'après le lemme **4.10** de l'annexe 1.

De plus, si $\lim\limits_{k\to\infty} h^k = 0$, alors grâce au lemme **4.10** , et aux relations (4.9) et (4.10), on déduit que :

$$R_1\{x^k\} \leq Q_1\{x^k\} \leq \lim_{k\to\infty} \sup w(x^k,h^k) = 0$$

Afin d'appliquer correctement le lemme **4.1** et le corollaire **4.1** , il est nécessaire d'assurer que J soit une approximation consistante.

Corollaire 4.2. *Soit F et x_0 satisfaisant les conditions du corollaire **4.1** ; alors il existe deux constantes $1 > c_1 > 0, c_2 > 0$ et une boule $S_1 = S(x_0,\delta_1) \subset S_0$ telles que, pour tout $x_0 \in S_1$ et les suites $\{w_k\}$, $\{\lambda_k\}$ satisfassent :*

$$1 - c_1 \leq w_k \leq 1 + c_1, \qquad -c_2 \leq \lambda_k \leq c_2, \quad k = 0,1,...$$

Et les itérations

$$x^{k+1} = x^k - w_k[F'(x^k) + \lambda_k I]^{-1} F(x^k),\ k = 0,1,...$$

sont contenues dans S_1 et convergent vers x_0. De plus, si

$$\lim_{k\to+\infty} \lambda_k = 0,$$

et

$$\lim_{k\to+\infty} w_k = 0,$$

alors

$$R_1\{x^k\} = Q_1\{x^k\} = 0.$$

Preuve. On définit $J : S_0 \times D_h \subset \mathbb{R}^n \times \mathbb{R}^2 \to L(\mathbb{R}^n)$ par

$$J(x,h) = (1-h_1)^{-1}[F'(x) + h_2 I],$$

où $D_h = \{h \in \mathbb{R}^2 / h_1 \neq 1\}$. Comme F' est continue en x_0, il existe $\eta > 0$ donnée, $\delta > 0$ telle que $\|F'(x)\| \leq \|F'(x_0)\| + \eta = \eta_1$ pour tout $x \in S(x_0, \delta) \subset D_0$. On a

$$\begin{aligned} \|J(x,h) - F'(x)\| &= \|\frac{h_1}{1-h_1}F'(x) + \frac{h_2}{1-h_1}I\| \\ &\leq \frac{|h_1|\eta_1 + |h_2|}{|1-h_1|}; \end{aligned}$$

en effet :

$$\begin{aligned} \|J(x,h) - F'(x)\| &= \|(1-h_1)^{-1}(F'(x) + h_2 I) - F'(x)\| \\ &= \|\frac{1}{1-h_1}F'(x) + \frac{h_2}{1-h_1}I - F'(x)\| \\ &= \|\frac{h_1}{1-h_1}F'(x) + \frac{h_2}{1-h_1}I\| \\ &\leq \frac{|h_1|\eta_1 + |h_2|}{|1-h_1|}. \end{aligned}$$

Ce qui montre que J est une approximation consistante de F' sur $S(x_0, \delta)$. Le résultat suivant s'obtient grâce au corollaire **4.1** . Considérons, par exemple, que les composantes de la matrice $J(x,h) \in L(\mathbb{R}^n)$ soient définies par :

$$[J(x,h)]_{i,j} = \begin{cases} \frac{1}{h_{ij}}[F_i(x + \beta \sum_{k=1}^{j-1} h_{ik}e^k + h_{ij}e^j) - F_i(x + \beta \sum_{k=1}^{j-1} h_{ik}e^k)], \ si \ h_{ij} \neq 0; \\ \\ \frac{\partial F_i}{\partial x_j}(x + \beta \sum_{k=1}^{j-1} h_{ik}e^k + h_{ij}e^k), \ si \ h_{ij} = 0, \end{cases} \tag{4.15}$$

où $\beta \in [0,1]$ et $e^1, ..., e^n$ les vecteurs de la base canonique.

Si $\beta = 1$, alors (4.15) correspond à l'approximation de $\frac{\partial F_i}{\partial x_j}(x)$

$$\frac{\partial F_i}{\partial x_j}(x) = \frac{1}{h_{ij}}[F_i(x + \sum_{i=1}^{j} h_{ik}e^k) - F_i(x + \sum_{i=1}^{j-1} h_{ik}e^k)],$$

Si $\beta = 0$, alors la relation (4.15) correspond à

$$\frac{\partial F_i}{\partial x_j}(x) = \frac{1}{h_{ij}}[F_i(x + h_{ij}e^j) - F_i(x)].$$

Si F est $F-$différentiable au voisinage de x, alors $J(x,h) \to F'(x)$ quand $h \to 0$,

Le prochain lemme prouve que, sous certaines conditions, la limite peut être réalisée uniformément de sorte que J soit une approximation consistante de F'.

Lemme 4.2. *Supposons que $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ soit continûment différentiable sur un ensemble ouvert D. Alors, pour tout ensemble compact $D_0 \subset D$, il existe un $r > 0$ tel que l'application $J : D_0 \times D_h \subset \mathbb{R}^n \times \mathbb{R}^{n^2} \to L(\mathbb{R}^n)$, donnée par la relation (4.15) soit bien définie pour tout $\beta \in [0,1]$, soit une approximation consistante de F' sur D_0, avec $D_h = \{h \in \mathbb{R}^{n^2} / |h_{ij}| \leq r,\ i,j = 1,...,n\}$. Si de plus*

$$\|F'(x) - F'(y)\| \leq \gamma \|x - y\|,\ \forall x,\ y \in D. \tag{4.16}$$

Alors J est une approximation fortement consistante de F' sur D_0.

Preuve. Comme D_0 est un compact et D est un ouvert, il existe $\delta > 0$ tel que l'ensemble compact $D_1 = \{x / \|x - y\|_1 \leq \delta, \text{pour un certain} \quad y \in D_0\}$ soit inclus dans D.
Clairement, F' est uniformément continue sur D_1, et, par conséquence pour $\varepsilon > 0$ donnée, il existe $\delta_1 \in (0, \delta)$ tel que

$$|\frac{\partial F_i}{\partial x_j}(x) - \frac{\partial F_i}{\partial x_j}(y)| \leq \varepsilon,\ \ i,\ j = 1,...,n,\ \ \forall\ x,\ y \in D_1,\ \ \|x - y\|_1 \leq \delta_1.$$

On pose $r = \frac{\delta_1}{n}$ et $\Delta_{ij}(h) = \beta \sum_{k=1}^{j-1} h_{ik} e^k$. Alors, pour tout $h \in D_h$,

$$\|\Delta_{ij}(h) + h_{ij} e^j\|_1 \leq nr \leq \delta_1 < \delta,\ i,j = 1,...,n,$$

ce qui montre que $x + \Delta_{ij}(h) + h_{ij} e^j \in D_1,\ \forall x \in D_0$.

Par conséquent, d'après le Théorème **4.2** de l'annexe 1, on obtient :

$$\begin{aligned} | \quad & \frac{1}{h_{ij}}[F_i(x + \Delta_{ij}(h) + h_{ij} e^j) - F_i(x + \Delta_{ij}(h))] - \frac{\partial F_i}{\partial x_j}(x)| \\ \leq \ & |\frac{1}{h_{ij}}[F_i(x + \Delta_{ij}(h) + h_{ij} e^j) - F_i(x + \Delta_{ij}(h))] - \frac{\partial F_i}{\partial x_j}(x + \Delta_{ij}(h))| \\ + \ & |\frac{\partial F_i}{\partial x_j}(x + \Delta_{ij}(h)) - \frac{\partial F_i}{\partial x_j}(x)| \leq 2\varepsilon, \end{aligned} \tag{4.17}$$

et par conséquent

$$\|F'(x) - J(x,h)\|_1 \leq 2n\varepsilon,\ \forall x \in D_0,\ h \in D_h.$$

Comme ε est arbitraire, cela prouve que J est bien une approximation consistante de F' sur D_0. Si la relation (4.16) est vérifiée, alors,

$$|\frac{\partial F_i}{\partial x_j}(x) - \frac{\partial F_i}{\partial x_j}(y)| \leq \gamma_1 \|x - y\|_1, \ \forall \ x, \ y \in D.$$

D'après le Théorème **4.2** de l'annexe 1, on en déduit que la partie droite de l'inégalité (4.17) peut être remplacé par :

$$\gamma_1[|h_{ij}| + |\Delta_{ij}(h)|] \leq \gamma_1 \sum_{k=1}^{n} |h_{ik}|.$$

D'où

$$\|F'(x) - J(x,h)\|_1 \leq \gamma_1 \sum_{i,j=1}^{h} |h_{ij}| = \gamma_1 \|h\|_1, \ \forall \ x \in D_0,$$

Ce qui prouve que J est une approximation fortement consistante sur D_0.

Remarque 4.3. Pour assurer que J soit bien définie, il faut que h soit petit. Si F est bien définie sur tout $\mathbb{R}^n$, on prend $D_h = \mathbb{R}^{n^2}$.
D'après les corollaire **4.1** et le lemme **4.2** on obtient un résultat de convergence résumé dans le paragraphe suivant :

Convergence de la méthode de Newton discrète

Théorème 4.1. *[70]. Supposons que $F : \mathbb{R}^n \to \mathbb{R}^n$ soit continûment différentiable, et qu'il existe une solution x_0 de $F(x) = 0$ tel que $F'(x_0)$ soit non singulière. Définissons $J : \mathbb{R}^n \times \mathbb{R}^{n^2} \to L(\mathbb{R}^n)$ par (4.15). Alors il existe $r_1 > 0$ et $\delta_1 > 0$ tel que , pour tout $x^0 \in S(x_0, \delta_1)$ et toute suite $\{h^k\} \subset S(0, r_1) \subset \mathbb{R}^{n^2}$, les itérations $\{x^k\}$ donné par (4.5) sont bien définies et convergent vers x_0. En plus, si $\lim_{k\to\infty} h^k = 0$, alors $R_1\{x^k\} = Q_1\{x^k\} = 0$.*

Afin d'obtenir une convergence rapide, il est nécessaire d'introduire les trois conditions suivantes :

1. *J est une approximation fortement consistante.*
2. *La fonction F est suffisamment lisse.*
3. *Le taux de décroissance de h^k est suffisamment rapide.*

Si F' satisfait la condition de Lipschitz (4.11) *et J est une approximation fortement consistante, alors d'après* (4.12) *on a :*

$$\|x^{k+1} - x_0\| \leq \alpha_1 \|x^k - x_0\|^2 + \alpha_2 \|h^k\| \|x^k - x_0\|.$$

- *Si $\alpha_2 = 0$, on a une convergence quadratique de la suite $\{x^k\}$.*
- *Si $\alpha_2 \neq 0$, alors le comportement de h^k quand $k \to \infty$, joue un rôle important pour estimer le taux de convergence ; le résultat suivant indique deux conditions sur h^k pour assurer la convergence rapide.*

Corollaire 4.3. *Supposons que $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$ soit G-différentiable dans un voisinage ouvert $S_0 \subset D$ de $x_0 \in D$, où $F(x_0) = 0$; si de plus la condition de Lipschitz* (4.11) *est vérifiée et si $F'(x_0)$ est non singulière alors :*
soit $J : D_J \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to L(\mathbb{R}^n)$ une approximation fortement consistante de F' sur S_0. Supposons que pour un certain $\{h^k\} \subset D_h$ les itérations $\{x^k\}$ donnée par la relation (4.5) *sont bien définies et convergent vers x_0. Si, de plus, la condition*

$$\|h^k\| \leq \beta_1 \|F(x^k)\|, \ \forall \ k \geq k_0, \tag{4.18}$$

est vérifiée, alors, $O_R\{x^k\} \geq O_Q\{x^k\} \geq 2$; sinon, si

$$\|h^k\| \leq \beta_2 \|x^k - x^{k-1}\|, \ \forall \ k \geq k_0, \tag{4.19}$$

*est vérifiée, alors, $O_R\{x^k\} \geq \frac{1}{2}(1 + \sqrt{5})$, où $O_R\{x^k\}$ est donné dans la définition **4.4** de l'annexe 1, et $O_Q\{x^k\}$ est donné dans la définition **4.3** de l'annexe 1.*

Preuve. D'après le lemme **4.1** , il existe $\delta > 0$, $r > 0$ tel que, pour tout $x \in S = S(x_0, \delta) \subset S_0$ et $h \in D'_h = D_h \cap S(0, r)$, la relation (4.12) est vérifiée.
Supposons que la relation (4.18) soit aussi vérifiée, alors $\lim_{k\to\infty} x^k = x_0$ implique que $\lim_{k\to+\infty} F(x^k) = 0$, et $x^k \in S$, $h^k \in D'_h$, pour tout $k \geq k_1 \geq k_0$. Par conséquent, d'après la relation (4.12) on a

$$\|x^{k+1} - x_0\| \leq \alpha_1 \|x^k - x_0\|^2 + \alpha_2 \beta_1 \|F(x^k)\| \|x^k - x_0\|, \ \forall \ k \geq k_1.$$

Or

$$\begin{aligned} \|F(x^k)\| &\leq \|F(x^k) - F(x_0) - F'(x_0)(x^k - x_0)\| + \|F'(x_0)(x_0 - x^k)\| \\ &\leq [\varepsilon_k + \|F'(x_0)\|]\|x^k - x_0\|, \end{aligned}$$

où $\lim_{k\to\infty} \varepsilon_k = 0$; en effet, ce résultat est une conséquence directe du lemme **4.11** de l'annexe 1.

De manière similaire, si (4.19) est verifiée, alors

$$\begin{aligned} \|x^{k+1} - x_0\| &\leq \alpha_1\|x^k - x_0\|^2 + \alpha_2\beta_2\|x^k - x^{k-1}\|\|x^k - x_0\| \\ &\leq (\alpha_1 + \alpha_2\beta_2)\|x^k - x_0\|^2 + \alpha_2\beta_2\|x^{k-1} - x_0\|\|x^k - x_0\|, \ \forall k \geq k_1. \end{aligned} \tag{4.20}$$

Il s'en suit que $O_R\{x^k\} \geq \tau$ où $\tau = \frac{1}{2}(1+\sqrt{5})$ est la racine positive du polynôme $t^2 - t - 1 = 0$.

Annexe 1 du chapitre 4

Lemme 4.3. *Supposons que $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ soit G-différentiable sur un ensemble convexe $D_0 \subset D$. Alors, pour tout, $x,\ y \in D_0$,*

$$\|F(y) - F(x)\| \leq \sup_{0\leq t\leq 1} \|F'(x + t(y - x)))\|\|x - y\|. \tag{4.21}$$

Preuve. Supposons que $M = \sup_{0\leq t\leq 1} \|F'(x + t(y - x)))\| < \infty$ et, pour $\varepsilon > 0$, soit Γ l'ensemble des $t \in [0, 1]$ pour lequel

$$\|F(x + t(y - x)) - F(x)\| \leq Mt\|y - x\| + \varepsilon t\|y - x\|, \tag{4.22}$$

est vérifiée. Clairement, $0 \in \Gamma$, et $\gamma = \sup_{t\in\Gamma} t$ est bien définie ; comme la définition **1.4** implique que $F(x + t(y - x))$ est continue par rapport à t, on a

$$\|F(x + \gamma(y - x)) - F(x)\| \leq M\gamma\|y - x\| + \varepsilon\gamma\|y - x\|. \tag{4.23}$$

Comme ε est arbitraire, clairement le résultat est prouvé si $\gamma = 1$.

Supposons que $\gamma < 1$. Alors, comme F' existe au point $x + \gamma(y - x)$, il existe $\beta \in (\gamma, 1)$ tel que

$$\begin{aligned} |F(x + \beta(y - x)) - F(x + \gamma(y - x)) - F'(x + \gamma(y - x))(\beta - \gamma)(y - x)\| \\ \leq \varepsilon(\beta - \gamma)\|y - x\|, \end{aligned}$$

d'où

$$\|F(x + \beta(y - x)) - F(x + \gamma(y - x))\| \leq M(\beta - \gamma)\|y - x\| + \varepsilon(\beta - \gamma)\|y - x\|.$$

Or, d'après la relation (4.23), on a

$$\begin{aligned} \|F(x + \beta(y - x) - F(x)\| &\leq (M\gamma + \varepsilon\gamma)\|y - x\| + (M + \varepsilon)(\beta - \gamma)\|y - x\| \\ &= (M + \varepsilon)\beta\|y - x\|; \end{aligned}$$

qui est en contradiction avec la définition de γ. Donc la relation (4.22) est vérifiée pour $1 > \beta > \gamma$.

Lemme 4.4. *Si $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ est $G-$ différentiable sur un ensemble convexe $D_0 \subset D$, alors, pour tout $x,\ y,\ z \in D_0$.*

$$\|F(y) - F(z) - F'(x)(y-z)\| \leq \sup_{0\leq t\leq 1} \|F'(z + t(y-z)) - F'(x)\|\|y-z\|. \tag{4.24}$$

Preuve. Pour $x \in D_0$, on définit l'application

$$G(\omega) = F(\omega) - F'(x)\omega,\ \omega \in D.$$

Les conditions de la définition **1.4** sont satisfaites pour G et comme

$$G'(\omega) = F'(\omega) - F'(x),$$

la relation (4.24) s'écrit

$$\|G(y) - G(z)\| \leq \sup_{0\leq t\leq 1} \|G'(z + t(y-z))\|\|y-z\|.$$

Lemme 4.5. *Si $G : [a,b] \subset \mathbb{R}^1 \to \mathbb{R}^m$ est continue sur $[a,b]$, alors*

$$\|\int_a^b G(t)dt\| \leq \int_a^b \|G(t)\|dt.$$

Preuve. Comme la norme est une fonction continue, $\|G(.)\|$ est intégrable au sens de Riemann, et, pour $\varepsilon > 0$ arbitraire, il existe une partition $a < t_0 < ... < t_p < b$ telle que

$$\|\int_a^b G(t)dt - \sum_{i=1}^{p} G(t_i)(t_i - t_{i-1})\| \leq \varepsilon,$$

et

$$|\int_a^b \|G(t)\|dt - \sum_{i=1}^{p} \|G(t_i)\|(t_i - t_{i-1}| \leq \varepsilon.$$

D'où,

$$\|\int_a^b G(t)dt\| \leq \|\sum_{i=1}^{p} G(t_i)(t_i - t_{i-1})\| + \varepsilon \leq \sum_{i=1}^{p} \|G(t_i)\|(t_i - t_{i-1}) + \varepsilon \leq \int_a^b \|G(t)\|dt + 2\varepsilon,$$

Théorème 4.2. *Soit $F : D \subset \mathbb{R}^n \to \mathbb{R}^m$ continûment différentiable sur un ensemble convexe $D_0 \subset D$ et supposons que, pour des constantes $\alpha \geq 0$ et $p \geq 0$, F' satisfasse*

$$\|F'(u) - F'(v)\| \leq \alpha \|u - v\|^p, \ u, \ v \in D_0.$$

Alors, pour tout $x, \ y \in D_0$

$$\|F(y) - F(x) - F'(x)(y - x)\| \leq [\frac{\alpha}{(p+1)}]\|y - x\|^{p+1}.$$

Preuve. D'après le Théorème **1.3** et le lemme **4.5** on a

$$\begin{aligned} \|F(y) - F(x) - F'(x)(y - x)\| &= \| \int_0^1 [F'(x + t(y - x)) - F'(x)](y - x)dt\| \\ &\leq \int_0^1 \|F'(x + t(y - x)) - F'(x)\|\|y - x\|dt \\ &\leq \alpha\|y - x\|^{p+1} \int_0^1 t^p dt. \end{aligned}$$

Corollaire 4.4. *[70] Soit $A \in L(\mathbb{C}^n)$, alors, pour tout $\epsilon > 0$, il existe une norme dans $\mathbb{C}^n$ telle que*

$$\|A\| \leq \rho(A) + \epsilon. \tag{4.25}$$

Lemme 4.6. *Soit $A \in L(\mathbb{C}^n)$. Alors $\lim_{k\to\infty} A^k = 0$ si et seulement si $\rho(A) < 1$.*

Preuve. Si $\rho(A) < 1$, alors, d'après le corollaire **4.4** , il existe une norme telle que $\|A\| < 1$. Par conséquent, on a $\|A^k\| \leq \|A\|^k$, il s'en suit que $A^k \to 0$ quand $k \to \infty$. Par contre, supposons que A admet des valeurs propres λ tel que $|\lambda| \geq 1$ et des vecteurs propres correspondants $x \neq 0$. Alors $A^k x = \lambda^k x$ pour tout k, donc $A^k x$ ne tend pas vers zéro.

Lemme 4.7. ***(Lemme de Neumann)***
Soit $B \in L(\mathbb{R}^n)$ et supposons que $\rho(B) < 1$ alors $(I - B)^{-1}$ existe et

$$(I - B)^{-1} = \lim_{k\to\infty} \sum_{i=0}^{k} B^i. \tag{4.26}$$

Preuve. Comme $\rho(B) < 1$, clairement $I - B$ n'admet pas des valeurs propres nulles et est donc inversible. Vérifions maintenant le relation (4.26), Notons que

$$(I - B)(I + ... + B^{k-1}) = I - B^k,$$

de plus

$$I + B + ... + B^{k-1} = (I - B)^{-1} - (I - B)^{-1}B^k.$$

D'après le lemme **4.6** , la partie droite tend vers $(I - B)^{-1}$. Comme $(I - B)$ est inversible quand $\|B\| \leq 1$, d'après (4.26) on a

$$\|(I - B)^{-1}\| \leq \sum_{i=0}^{\infty} \|B\|^i = \frac{1}{1 - \|B\|}. \tag{4.27}$$

Lemme 4.8. ***(Lemme de Perturbation)***
Soit $A,\ C \in L(\mathbb{R}^n)$ et supposons que A soit inversible, avec $\|A^{-1}\| \leq \alpha$. Si $\|A - C\| \leq \beta$ et $\beta\alpha < 1$, alors C est aussi inversible, et

$$\|C^{-1}\| \leq \frac{\alpha}{(1 - \alpha\beta)}.$$

Preuve. On a $\|I - A^{-1}C\| = \|A^{-1}(A - C)\| \leq \alpha\beta < 1$ et $A^{-1}C = I - (I - A^{-1}C)$; d'après le lemme **4.7** on déduit que $A^{-1}C$ est inversible. D'où, C est inversible. De plus, d'après la relation (4.27) on déduit que

$$\|C^{-1}\| = \|[I - (I - A^{-1}C)]^{-1}A^{-1}\| \leq \alpha \sum_{i=0}^{\infty} (\alpha\beta)^i = \frac{\alpha}{(1 - \alpha\beta)}.$$

Définition 4.2. Soit $\{x^k\} \subset \mathbb{R}^n$ une suite qui convergent vers x_0. Pour $p \in [1, \infty[$, on définit les quantités suivantes :

$$Q_p\{x^k\} = \begin{cases} 0, & \text{si } x^k = x_0,\ k < \infty\,; \\ \lim\limits_{k \to +\infty} \sup \frac{\|x^{k+1} - x_0\|}{\|x^k - x_0\|^p}, & \text{si } x^k \neq x_0,\ k < \infty\,; \\ +\infty, & \text{sinon}\,; \end{cases}$$

appelée $Q-$facteur, et

$$R_p\{x^k\} = \begin{cases} \lim\limits_{k\to+\infty} \sup \|x^k - x_0\|^{\frac{1}{k}}, & \text{si } p = 1\,; \\ \lim\limits_{k\to+\infty} \sup \|x^k - x_0\|^{\frac{1}{p^k}}, & \text{si } p > 1\,; \end{cases}$$

appelée $R-$facteur.

Définition 4.3. Soit $Q_p(\mathfrak{F}, x_0)$ un $Q-$ facteurs du processus itératif $\mathfrak{F}$ avec un point limite x_0 dans $\mathbb{R}^n$; alors la quantité

$$O_Q(\mathfrak{F}, x_0) = \begin{cases} +\infty, \text{ si } Q_p(\mathfrak{F}, x_0) = 0, \ \forall p \in [1, \infty[, \\ \inf\{p \in [1, \ \infty[, \ Q_p(\mathfrak{F}, x_0) = +\infty\}, \text{ sinon}, \end{cases}$$

est appelée l'ordre de Q du processus $\mathfrak{F}$ au point x_0.

Définition 4.4. Soit un processus itératif $\mathfrak{F}$ avec un point limite x_0 ; alors la quantité :

$$O_R(\mathfrak{F}, x_0) = \begin{cases} +\infty, \text{ si } R_p(\mathfrak{F}, x_0) = 0, \ \forall p \in [1, \infty[, \\ \inf\{p \in [1, \ \infty[, \ R_p(\mathfrak{F}, x_0) = 1\}, \text{ sinon}, \end{cases}$$

est appelée l'ordre de R du processus $\mathfrak{F}$ au point x_0.

Lemme 4.9. *Soit $\{x^k\} \subset \mathbb{R}^n$ une suite convergente vers x_0, alors*

$$R_1\{x^k\} \leq Q_1\{x^k\},$$

pour chaque norme. Par suite, si $\mathfrak{F}$ est un processus itératif avec un point limite x_0, alors

$$R_1(\mathfrak{F}, x_0) \leq Q_1(\mathfrak{F}, x_0),$$

pour chaque norme.

Preuve. Supposons que $Q_1\{x^k\} < \infty$, et posons $\varepsilon_k = \|x^k - x_0\|$, alors, pour tout $\varepsilon > 0$ et $\gamma = Q_1\{x^k\} + \varepsilon$, il existe $k_0 \geq 0$ tel que

$$\varepsilon_k \leq \gamma\varepsilon_{k-1} \leq ... \leq \gamma^{k-k_0}\varepsilon_{k0}, \ \forall \ k \geq k_0.$$

Par conséquent,

$$R_1\{x^k\} \leq \gamma \lim_{k\to\infty} \sup[\frac{\varepsilon_{k0}}{\gamma^{k0}}]^{\frac{1}{k}} = \gamma,$$

et comme ε est arbitraire

$$R_1\{x^k\} \leq Q_1\{x^k\}.$$

Si $\mathbb{C}(\mathfrak{F}, x_0)$ est l'ensemble de toute les suites générées par $\mathfrak{F}$ qui convergent vers x_0, on obtient

$$\begin{aligned} R_1(\mathfrak{F}, x_0) = \sup\{R_1\{x^k\}/\{x^k\} \in \mathbb{C}(\mathfrak{F}, x_0)\} &\leq \sup\{Q_1\{x^k\}/\{x^k\} \in \mathbb{C}(\mathfrak{F}, x_0)\} \\ &= Q_1(\mathfrak{F}, x_0). \end{aligned}$$

Lemme 4.10. *Soit $G : D \times D_h \subset \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$; supposons qu'il existe un ensemble $S = S(x_0, \delta) \subset D$ et $D'_h \subset D_h$, tel que $\alpha < 1$ vérifie*

$$\|G(x, h) - x_0\| \leq \alpha\|x - x_0\|, \ \forall x \in S, \ \forall h \in D'_h. \tag{4.28}$$

Alors, pour tout $x^0 \in S$ et toute suite $\{h^k\} \subset D'_h$, les itérations $\{x^k\}$ générées par :

$$x^{k+1} = G(x^k, h^k), \ k = 0, 1, ... \tag{4.29}$$

sont contenues dans S et convergent vers x_0. De plus,

$$R_1\{x^k\} \leq Q_1\{x^k\} \leq \alpha. \tag{4.30}$$

Preuve. La preuve est immédiate. Une simple induction montre que

$$\|x^{k+1} - x_0\| = \|G(x^k, h^k) - x_0\| \leq \alpha\|x^k - x_0\| \leq .... \leq \alpha^{k+1}\|x^0 - x_0\|;$$

d'où, pour toute suite x^k contenue dans S et convergeant vers x. La première inégalité de la relation (4.30) est donnée par le lemme **4.9** , la seconde découle de la relation (4.28) et de la définition de Q_1.

Lemme 4.11. *[70] Soit $\mathfrak{F}$ un processus itératif de point limite x_0, supposons qu'il existe $p \in [1, \infty[$ et une constante c_2 tel que, pour tout suite $\{x^k\}$ on a :*

$$\|x^{k+1} - x_0\| \leq c_2\|x^k - x_0\|^p, \ \forall \ k \geq k_0, \tag{4.31}$$

alors

$$O_R(\mathfrak{F}, x_0) \geq O_Q(\mathfrak{F}, x_0) \geq p.$$

D'autres part, s'il existe une constante $c_1 > 0$ *et pour une certain suite* $\{x^k\}$ *on a :*

$$\|x^{k+1} - x_0\| \geq c_1 \|x^k - x_0\|^p > 0, \ \forall \ k \geq k_0, \tag{4.32}$$

alors

$$O_Q(\mathfrak{F}, x_0) \leq O_R(\mathfrak{F}, x_0) \leq p.$$

D'ou si (4.31) *et* (4.32) *sont vérifiées, alors*

$$O_R(\mathfrak{F}, x_0) = O_Q(\mathfrak{F}, x_0) = p.$$

Chapitre 5

Résolution d'un problème de contrôle optimal avec contrainte sur l'état par la méthode de relaxation

5.1 Introduction

Dans cette étude, nous présentons une méthode numérique pour résoudre un problème de contrôle optimal avec un temps terminal fixé et une contrainte sur l'état ainsi que sur l'état final. Nous considérons, sous le même formalisme, deux cas distincts de problèmes de contrôle optimal : le cas sans et avec contrainte sur l'état. Dans les deux cas, en vue d'une résolution numérique et en utilisant la notion de sous-différentiel [3] et [53] pour prendre en compte si nécessaire, la projection sur le convexe des contraintes, nous reformulerons les équations d'optimalité issues du principe de minimum de Pontryagin. Ces dernières forment un système algébro-différentiel où l'équation d'état est munie d'une condition initiale et d'une condition finale. Par contre, l'équation d'état adjoint n'est munie d'aucune condition initiale ou terminale utilisable de manière algorithmique. Pour déterminer la condition initiale sur l'état adjoint, nous utiliserons dans cette étude, la méthode de tir [78], couplée à la méthode de relaxation (voir [67],[42] et [51]). Sous des hypothèses convenables, nous analysons la convergence de la méthode itérative considérée et nous terminons en exposant des résultats d'expérimentations numériques.

5.2 Position du problème

5.2.1 Cas sans contrainte sur l'état

Soit le système dynamique suivant :

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t),\ t \in [0,T], \\ x(0) = x_0,\ x(T) = x_f, \\ u(t) \in U, \end{cases} \tag{5.1}$$

où $x(t)$ est un n-vecteur représentant l'état du système à l'instant t, $x(0) = x_0$ est la condition initiale, $x(T) = x_f$ est l'état final. $u(t)$ est un m-vecteur représentant la commande agissant sur le système à l'instant $t \in [0,T]$; U est l'ensemble des commandes admissibles qui est un ensemble ouvert. $A,\ B$ sont des $n \times n$ et $n \times m$ matrices données.

On cherche une commande admissible $\hat{u}$ qui transfère le système d'un état initial x_0 fixé vers un état final x_f fixé et minimisant la fonction coût J définie par :

$$J(u) = \frac{1}{2}\int_0^T [(x - x_d)^t Q(x - x_d) + ku^t u]dt,$$

où x_d représente un état désiré ; la matrice Q est symétrique définie non-négative. L'Hamiltonien du système est donné par :

$$H(x,p,u,t) = \frac{1}{2}[(x - x_d)^t Q(x - x_d) + ku^t u] + p^t[Ax + Bu],$$

où p est le vecteur d'état adjoint. Cherchons maintenant la commande $\hat{u}$ qui minimise l'Hamiltonien, tel que :

$$H(\hat{x},\hat{p},\hat{u}) \leq H(\hat{x},\hat{p},u);\ \forall u \in U,\ \forall t \in [0,T].$$

Les équations d'optimalité s'écrivent donc :

$$\begin{cases} \frac{dx}{dt} = \frac{\partial H}{\partial p} = Ax + Bu;\ x(0) = x_0,\ x(T) = x_f,\ \forall t \in [0,T], \\ -\frac{dp}{dt} = \frac{\partial H}{\partial x} = A^t p + Q(x - x_d),\ p(0) \text{ à déterminer}, \\ \frac{\partial H}{\partial u} = 0 = ku + B^t p. \end{cases} \tag{5.2}$$

Ces équations sont connues sous le nom d'équations d'Hamilton-Pontryagin. On aboutit à la résolution d'un système algébro-différentiel ; l'équation d'état décrivant le système physique est munie d'une condition initiale $x(0) = x_0$ et d'une condition finale $x(T) = x_f$. Par contre, la seconde équation correspondant à l'équation d'état adjoint, n'est munie d'aucune condition initiale ni d'aucune condition terminale utilisable pratiquement. On va donc utiliser la méthode de tir présentée au chapitre 4 pour calculer la valeur de $p(0)$.

Condition de transversalité

De manière générale, lorsque l'on prend en compte un coût terminal, le critère à minimiser s'écrit :

$$J = g(T, x(T)) + \int_0^T f_0(x(t), u(t), t)dt,$$

où g est le coût terminal, l'état final étant fixé. Conformément à [78], soient M_0 et M_1 deux sous ensembles de $\mathbb{R}^n$; on cherche à déterminer une trajectoire reliant M_0 à M_1 tout en minimisant le coût. Si de plus M_0 et M_1 sont des variétés de $\mathbb{R}^n$ ayant des espaces tangents $T_{x(0)}M_0$ et $T_{x(T)}M_1$ respectivement en $x(0) \in M_0$ et en $x(T) \in M_1$, alors le vecteur $p(t)$ peut être construit de manière à vérifier les conditions de transversalité :

$$p(0) \perp T_{x(0)}M_0, \tag{5.3}$$

$$p(T) - p^0 \nabla_x g(T, x(T)) \perp T_{x(T)}M_1, \tag{5.4}$$

où p^0 est un réel tel que $p^0 < 0$ conduit au principe du maximum de Pontryagin et $p^0 > 0$ conduit au principe du minimum de Pontryagin [78]. Le cas $p^0 = 0$ correspond à un cas singulier qui n'est pas étudié dans ce travail. Si $M_0 = \{x_0\}$, la condition (5.3) devient vide et si la variété M_1 s'écrit sous la forme :

$$M_1 = \{x \in \mathbb{R}^n / F_1(x) = ... = F_q(x) = 0\},$$

où les F_i sont des fonctions de classe C^1 sur $\mathbb{R}^n$, alors l'espace tangent à M_1 en un point $x \in M_1$ est donné par :

$$T_x M_1 = \{s \in \mathbb{R}^n / \nabla F_i(x)s = 0,\ i = 1, ..., q\};$$

la condition (5.4) s'écrit alors :

$$\exists s_1, ..., s_q \in \mathbb{R}/p(T) = \sum_{i=1}^{q} s_i \nabla_x F_i(x(T)) + p^0 \nabla_x g(T, x(T)),$$

où s_i sont les multiplicateurs de *Lagrange*. Dans notre problème, $g(T, x(T)) = 0$; donc la condition de transversalité sur le vecteur adjoint s'écrit :

$$p(T) = \sum_{i=1}^{q} s_i \nabla_x F_i(x(T)), s_i \in \mathbb{R}.$$

5.2.2 Cas avec contraintes sur l'état

Dans le cas où l'état est soumis à certaines contraintes notées respectivement par x^{min} et x^{max}, soit X_{ad} un ensemble convexe des trajectoires admissibles. Par la suite, nous reformulerons les conditions nécessaires d'optimalité ; pour cela on va donc utiliser la notion de sous-différentiel pour obtenir des conditions d'optimalité. Considérons le problème avec contraintes suivant :

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t),\ t \in [0, T],\ x(t) \in X_{ad}, \\ \\ x(0) = x_0,\ x(T) = x_f, \end{cases} \tag{5.5}$$

où l'ensemble convexe fermé X_{ad} est défini par $X_{ad} = \{x \in \mathbb{R}^n / x_i^{min} \leq x_i \leq x_i^{max}, i = 1, \ldots, n\}$; notons $\Psi_{X_{ad}}$ la fonction indicatrice de X_{ad} dont la *ieme* composante satisfait :

$$(\Psi_{X_{ad}})_i = \begin{cases} 0, & \text{si } \ x_i^{min} \leq x_i \leq x_i^{max}, \\ \\ +\infty, & \text{sinon.} \end{cases}$$

La *ieme* composante du sous-différentiel $\partial\Psi_{X_{ad}}$ est donnée par :

$$(\partial\Psi_{X_{ad}})_i = \begin{cases}]-\infty, 0], & \text{si } \ x_i = x_i^{min}, \\ \\ 0, & \text{si } \ x_i^{min} < x_i < x_i^{max}, \\ \\ [0, +\infty[, & \text{si } \ x_i = x_i^{max}, \\ \\ \emptyset, & \text{sinon,} \end{cases}$$

et admet le graphe représenté par la Figure 5.1. Notons que le sous-différentiel $\partial\Psi_{X_{ad}}$ est monotone. Appliquons le Lemme **1.1** du chapitre 1 ; on cherche $\hat{u}$ qui minimise l'Hamiltonien H ; ceci peut s'écrire sous la forme :

$$0 \in \partial H(\hat{u}).$$

Comme H est un opérateur continu [53], nous obtenons la nouvelle formulation des conditions nécessaires d'optimalité :

$$\begin{cases} 0 \in \frac{dx}{dt} + \partial\Psi_{X_{ad}} - Ax - B\hat{u};\ x(0) = x_0,\ x(T) = x_f,\ \forall t \in [0,T], \\ -\frac{dp}{dt} = \frac{\partial H}{\partial x} = A^t p + Q(x - x_d),\ p(0) \text{ à déterminer}, \\ \frac{\partial H}{\partial u} = 0 = B^t p + k\hat{u}. \end{cases} \tag{5.6}$$

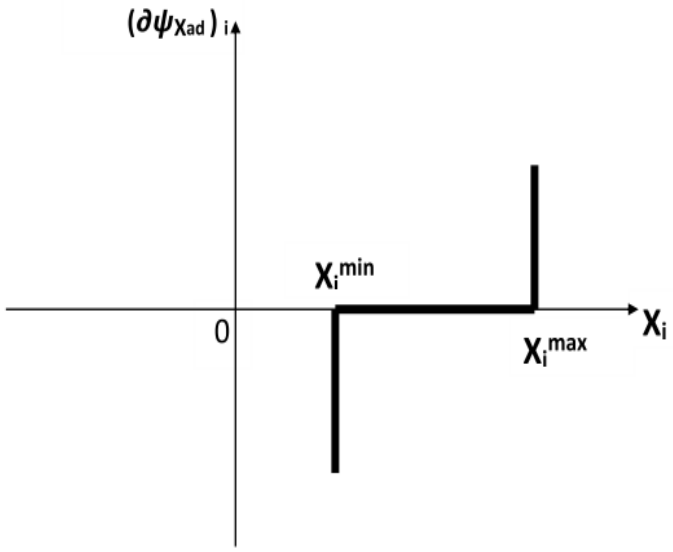

FIG. 5.1 – La i^{eme} composante du sous-différentiel de la fonction $\Psi_{X_{ad}}$

5.3 Méthode de résolution numérique

Pour la résolution du problème, avec et sans contraintes sur l'état, nous effectuons le couplage de la méthode de relaxation (voir [42], [51] et [67]) avec la méthode de tir (voir

[78]), cette dernière étant destinée à calculer $p(0)$ nécessaire à la résolution du système algébro-différentiel obtenus par application du principe de minimum de Pontryagin. Les étapes de la méthode de résolution numérique sont résumées ci-dessous :

5.3.1 Cas avec contrainte

1. Approximation de la commande initiale u^0 pour $t \in [0, T]$, et de l'état adjoint initial $p^0(0)$,
2. $r \leftarrow 0$ (où r permet de compter les itérations),
3. **Tant que** $\mid u^{(r+1)} - u^{(r)} \mid > \epsilon$ (où ϵ définit le seuil de convergence) **faire** :
   - Détermination de l'équation d'état $x^{(r)}$, par intégration numérique de l'équation d'état avec projection sur le convexe X_{ad} :

$$\begin{cases} \frac{d\bar{x}}{dt} = A\bar{x} + Bu^{(r)}, \quad 0 < t \leq T, \\ \\ \bar{x}(0) = x_0, \end{cases} \qquad et \quad x^{(r)} = Proj(\bar{x}), \tag{5.7}$$

     où $Proj(.)$ est l'opérateur de projection sur le convexe fermé X_{ad}, puis détermination de l'état adjoint $p^{(r)}$ en résolvant :

$$\begin{cases} -\frac{dp^{(r)}}{dt} = A^t p^{(r)} + Q(x^{(r)} - x_d), \\ \\ p^{(r)}(0), \end{cases} \tag{5.8}$$

     où $p^{(r)}(0)$ est calculé par la méthode de tir,
   - Détermination de la commande $u^{(r+1)}$:

$$u^{(r+1)} \leftarrow -(\frac{1}{k} B^t p^{(r)}), \tag{5.9}$$

   - $Convergence \leftarrow \mid u^{(r+1)} - u^{(r)} \mid$,
   - Détermination de la fonction de tir :

$$G(p) = x^{(r)}(T) - x_f,$$

   - Résolution de l'équation de tir par la méthode de Newton et détermination de la nouvelle valeur de $p(0)$:

$$p^{(r+1)}(0) \leftarrow p^{(r)}(0) + correction,$$

- $r \leftarrow r + 1$.

Fin de tant que.

5.3.2 Cas sans contrainte

La démarche est analogue sauf que l'étape (5.7) est remplacée par

$$\begin{cases} \frac{dx^{(r)}}{dt} = Ax^{(r)} + Bu^{(r)}, \quad 0 < t \leq T, \\ x(0) = x_0. \end{cases}$$

Remarque 5.1. Les étapes (5.7)-(5.9) de la boucle correspondent à la méthode de relaxation alors que les étapes suivantes correspondent à la mise en œuvre de la méthode de tir.

5.4 Convergence de la méthode

Ecrivons les équation d'optimalité sous forme matricielle :

$$\begin{pmatrix} \frac{dx}{dt} + \partial\Psi_{X_{ad}} \\ -\frac{dp}{dt} \\ 0 \end{pmatrix} + \begin{pmatrix} \bar{A} & 0 & -B \\ -Q & \bar{A}^t & 0 \\ 0 & B^t & kI \end{pmatrix} \begin{pmatrix} x \\ p \\ u \end{pmatrix} \ni \begin{pmatrix} 0 \\ -Qx_d \\ ku_d \end{pmatrix}, \quad x(0) = x_0,$$

où $\bar{A} = -A$ et I est la matrice identité. La valeur du paramètre $k > 0$ permet de réaliser le dosage entre la précision du calcul et la minimisation de l'énergie dépensée, pour réaliser la commande optimale. Le problème s'écrit donc comme la somme d'un système linéaire perturbé par une application diagonale. Notons Θ la matrice suivante :

$$\Theta = \begin{pmatrix} \bar{A} & 0 & -B \\ -Q & \bar{A}^t & 0 \\ 0 & B^t & kI \end{pmatrix}.$$

Définition 5.1. Une matrice inversible $\bar{A}$ est une M-matrice si $\bar{A}^{-1} \geq 0$ et $\bar{a}_{ij} \leq 0$ pour $i \neq j$.

Remarque 5.2. Les M-matrices ont de nombreuses propriétés importantes ; notamment le rayon spectral de la matrice de Jacobi associée $J = I - \bar{D}^{-1}.\bar{A}$, où $\bar{D}$ est la diagonale de $\bar{A}$, est inférieur à un ; propriété que nous utiliserons dans la suite.

Proposition 5.1. *Si les conditions suivantes sont vérifiées :*

- $\bar{A}$ *est une M-matrice*
- $k \geq k_0 > 0$
- $p^2(0) - p^2(T) > 0$,

alors l'algorithme permettant de calculer numériquement la loi de commande optimale, par la méthode de relaxation couplée à la méthode de tir, converge quelque soit la donnée initiale u^0.

Preuve. La preuve est analogue à celle utilisée dans [52, 67] dans le cas plus simple où la valeur de l'état final $x(T)$ est libre et où il n'y a pas de contrainte sur l'état. En effet, on a vu dans le Lemme **1.2** du chapitre 1 que le sous-différentiel est une application continue monotone ; de plus si $x(0)$ est nul ce qui est toujours possible par un changement de variable, alors trivialement on a,

$$< \frac{dx}{dt}, x >= \int_0^T x\frac{dx}{dt}dt = \frac{1}{2}\int_0^T \frac{dx^2}{dt}dt = \frac{1}{2}x^2(T) = \frac{1}{2}x_f^2 > 0,$$

où $<,>$ est le produit scalaire standart dans l'espace des fonctions continues. De plus, $x \rightarrow (dx/dt)$, avec $x(0) = 0$, est un opérateur monotone.

Par ailleurs, puisque l'état adjoint $p(T)$ étant en général différent de zero, nous avons :

$$< -\frac{dp}{dt}, p >= -\frac{1}{2}\int_0^T \frac{dp^2}{dt}dt = \frac{1}{2}p^2(0) - \frac{1}{2}p^2(T) > 0.$$

En utilisant la dernière hypothèse, l'opérateur $p \rightarrow -(dp/dt)$ est également monotone.

Ainsi, nous avons :

$$\begin{cases} \frac{dx_i}{dt} + \bar{a}_{ii}x_i + \sum_{j\neq i} \bar{a}_{ij}x_j - \sum_j b_{ij}u_j + \partial\Psi_i \ni 0, \\ -\frac{dp_i}{dt} + \bar{a}^t_{ii}p_i + \sum_{j\neq i} \bar{a}^t_{ij}p_j - \sum_j q_{ij}x_j = -\sum_j q_{ij}x_{jd}, \\ ku_i + \sum_{j\neq i} b^t_{ij}p_j = 0. \end{cases} \tag{5.10}$$

Soit (w, π, ν) la valeur des itérés obtenus par un algorithme itératif tel que la méthode de Jacobi ou bien celle de Gauss-Seidel :

$$\begin{cases} \frac{dw_i}{dt} + \bar{a}_{ii}w_i + \sum_{j\neq i} \bar{a}_{ij}w_j - \sum_j b_{ij}\nu_j + \partial\bar{\Psi}_i \ni 0, \\ -\frac{d\pi_i}{dt} + \bar{a}^t_{ii}\pi_i + \sum_{j\neq i} \bar{a}^t_{ij}\pi_j - \sum_j q_{ij}w_j = -\sum_j q_{ij}x_{jd}, \\ k\nu_i + \sum_{j\neq i} b^t_{ij}\pi_j = 0, \end{cases} \tag{5.11}$$

où à la r^{eme} itération on a $w_i = x^r_i$, $\pi = p^r_i$ et $\nu_i = u^r_i$. En soustrayant membre à membre les équations des systèmes (5.10) et (5.11) et en multipliant respectivement par $(x_i - w_i)$, $(p_i - \pi_i)$ et $(u_i - \nu_i)$, on aura :

$$\begin{cases} < \frac{d}{dt}(x_i - w_i), x_i - w_i > + \bar{a}_{ii} \mid x_i - w_i \mid^2 + \sum_{j\neq i} a_{ij} < x_j - w_j, x_i - w_i > \\ \qquad\qquad - \sum_j b_{ij} < u_j - \nu_j, x_i - w_i > + < \partial\Psi_i - \partial\bar{\Psi}_i, x_i - w_i > \ni 0, \\ < -\frac{d}{dt}(p_i - \pi_i), p_i - \pi_i > + \bar{a}^t_{ii} \mid p_i - \pi_i \mid^2 = \sum_{j\neq i} a^t_{ij} < p_j - \pi_j, p_i - \pi_i > \\ \qquad\qquad + \sum_j q_{ij} < x_j - w_j, p_i - \pi_i >, \\ k < u_i - \nu_i, u_i - \nu_i > = -\sum_j b^t_{ij} < p_j - \pi_j, u_i - \nu_i > . \end{cases}$$

Compte tenu de la monotonie des trois opérateurs diagonaux précédents, on obtient aisément

les inégalités suivantes :

$$\begin{cases} \mid x_i - w_i \mid \leq \sum\limits_{j \neq i} \frac{|\bar{a}_{ij}|}{\bar{a}_{ii}} \mid x_j - w_j \mid + \sum\limits_{j} \frac{|b_{ij}|}{\bar{a}_{ii}} \mid u_j - \nu_j \mid, \\ \\ \mid p_i - \pi_i \mid \leq \sum\limits_{j \neq i} \frac{|\bar{a}^t_{ij}|}{\bar{a}^t_{ii}} \mid p_j - \pi_j \mid + \sum\limits_{j} \frac{|q_{ij}|}{\bar{a}_{ii}} \mid x_j - w_j \mid, \\ \\ \mid u_i - \nu_i \mid \leq \sum\limits_{j} \frac{|b^t_{ij}|}{k} \mid p_j - \pi_j \mid, \end{cases}$$

qui peuvent aussi s'écrire :

$$\mid s_i - s_i^r \mid \leq \sum_{j \neq i} \frac{\mid \theta_{ij} \mid}{\theta_{ij}} \mid s_j - v_j \mid,$$

où $S = (x, p, u)$ est la solution exacte du problème (5.10). Si k est supérieur à un nombre $k_0 > 0$ donné, alors la matrice Θ est une H-matrice [70] (c'est à dire, la matrice $\bar{\Theta}$ de cœfficients $\mid \theta_{ii} \mid$ et $- \mid \theta_{ij} \mid$ est une M-matrice) ; dans ces conditions, on peut définir la norme uniforme avec poids :

$$\parallel S - S^r \parallel_J = \max_j \frac{\mid s_j - s_j^r \mid}{\Gamma_j},$$

où Γ_j est la composante du vecteur propre associé au rayon spectral $\rho(J)$ de la matrice de Jacobi J associée à la matrice $\bar{\Theta}$ (voir [70]).

D'aprés le Théorème de Perron-Frobenius [70], on a :

$$J\Gamma \leq \rho(J)\Gamma, \quad avec \quad 0 \leq \rho(J) < 1,$$

où Γ est un vecteur strictement positif. Alors, nous obtenons finalement :

$$\parallel S - S^r \parallel_J \leq \rho(J) \parallel S - S^{r-1} \parallel_J,$$

et comme $\rho(J) < 1$, alors la convergence de la méthode est assurée.

Remarque 5.3. La preuve de la convergence est valable dans le cas avec et sans contrainte sur l'état. En effet, dans ce dernier cas, le sous-différentiel de la fonction indicatrice est nul et le raisonement est encore valable.

5.5 Exemple numérique : le problème en anneau

5.5.1 Cas sans contrainte

On considère le problème de contrôle optimal suivant :

$$\begin{cases} \textit{Déterminer} \quad \hat{u} \in U, \quad \textit{tel que}, \\ \\ J(\hat{u}) \leq J(u), \forall u \in U, \end{cases} \tag{5.12}$$

où

$$J(u) = \frac{1}{2} \int_0^T \{\| x - x_d \|_2^2 + k \| u \|_2^2\} dt,$$

sous les contraintes suivantes :

$$\begin{cases} \dot{x}_i = -wx_i + ax_{i+1} + bu_i, \ x_i(0) = 0.5, \ i \in \{1, 2, ..., n-1\} \ et \ n \geq 2, \\ \\ \dot{x}_n = ax_1 - wx_n + bu_n, \ x_n(0) = 0.5, \\ \\ x_i(T) = 0.5, \ i \in \{1, 2, ..., n\}, \end{cases} \tag{5.13}$$

où a, b et w sont des constantes réelles positives. L'Hamiltonien relatif à ce problème est donné par :

$$\begin{aligned} H(x, p, u, t) &= \frac{1}{2}(\| x - x_d \|_2^2 + k \| u \|_2^2) + \sum_{i=1}^{n-1} p_i(-wx_i + ax_{i+1} + bu_i) \\ &+ p_n(ax_1 - wx_n + bu_n). \end{aligned}$$

Les équations d'optimalité s'écrivent :

$$\begin{cases} \dot{x}_i = -wx_i + ax_{i+1} + bu_i, \ x_i(0) = 0.5, \ i \in \{1, 2, ..., n-1\}, \\ \\ \dot{x}_n = ax_1 - wx_n + bu_n, \ x_n(0) = 0.5, \\ \\ \dot{p}_1 = -x_1 + wp_1 - ap_n + x_{1d}, \ p_1(0) \text{ à déterminer}, \\ \\ \dot{p}_i = -x_i - ap_{i-1} + wp_i + x_{id}, \ p_i(0) \text{ à déterminer pour } i \in \{2, ..., n\}, \\ \\ ku_i + bp_i = 0, \ i \in \{1, ..., n\}. \end{cases}$$

Solution numérique

la solution numérique est calculée pour différentes valeurs de n. En posant $z(t) = (x(t), p(t))$ notre système devient :

$$\begin{cases} \dot{z}_i = -wz_i + az_{i+1} + bu_i,\ i \in \{1, ..., n-1\}, \\ \dot{z}_n = az_1 - wz_n + bu_n, \\ \dot{z}_{n+1} = -z_1 + wz_{n+1} - az_{2n} + x_{1d}, \\ \dot{z}_i = -z_{i-n} - az_{i-1} + wz_i + x_{(i-n)d},\ i \in \{n+2, ..., 2n\} \\ ku_{i-n} + bz_i = 0,\ i \in \{n+1, ..., 2n\}, \\ z_i(0) = 0.5,\ i \in \{1, ..., n\}, \\ z_i(0) \in \mathbb{R},\ \ i \in \{n+1, ..., 2n\}. \end{cases}$$

Soit $z(t)$ une solution du système précédent au temps t avec les conditions initiales $z(0) = (z_1(0), \ldots, z_i(0), \ldots, z_{2n}(0))$.

Pour $T = 4$, on doit avoir :

$$z_i(T = 4, z(0)) = \begin{cases} 0.5, & \text{pour } i \in \{1, ..., n\}, \\ z_i(0), & \text{pour } i \in \{n+1, ..., 2n\}, \end{cases}$$

où $z_i(0)$ pour $i \in \{n+1, ..., 2n\}$ sont à déterminer. On construit une fonction de tir qui est une équation algébrique non linéaire de la variable p à l'instant $T = 4$; cette fonction de tir est calculée par une procédure d'intégration numérique d'équations différentielles ordinnaires (en utilisant par exemple la méthode d'Euler, la méthode de Runge-Kutta, etc). La fonction de tir s'écrit :

$$G(z) = \overline{z} - I \times 0.5,$$

où

$$\overline{z} = (z_i \text{ pour } 1 \leq i \leq n).$$

Le problème à résoudre s'écrit alors comme suit : Déterminer $p(0)$ tel que $G(p(0))$ donne la valeur de $x(T) = x_f$ désiré. L'algorithme de résolution numérique de ce problème sera alors complètement défini, si l'on se donne :

1. L'algorithme d'intégration d'un système différentiel à valeur initiale (par exemple une procédure d'Euler ou de Runge-Kutta), pour calculer la fonction de tir G. Dans notre cas, nous utiliserons 'ode45' de Matlab qui est une méthode de Runge-Kutta 4/5 à pas variable.
2. L'algorithme de résolution de l'équation $G(z) = 0$ qui dans notre cas utilise la méthode de quasi-Newton ('fsolve' de Matlab).

Solution exacte dans le cas sans contrainte

Pour calculer de manière analytique la commande optimale $u(t)$, et sa trajectoire correspondante $x(t)$ du problème $(5.12) - (5.13)$, afin de limiter les calculs analytiques, nous limitons la valeur de n à 2 ; nous avons utilisé les équations d'optimalité ainsi que la condition de transversalité sur $p(t)$. Puisque il n'y a pas de coût terminal, la condition de transversalité sur $p(t)$ s'écrit :

$$\exists s_1, s_2 \in \mathbb{R}/p(T) = \sum_{i=1}^{2} s_i \nabla F_i(x(t)).$$

On pose $F_1(x) = x_1(T) - 0.5,\ F_2(x) = x_2(T) - 0.5,\ p(T) = (s_1, s_2)$ où s_1, s_2 sont les multiplicateurs de Lagrange.

Pour trouver la solution exacte du problème de contrôle optimal, on utilise la méthode de dérivation au niveau des équations. On a :

$$\dot{p}_1 = -x_1 + wp_1 - ap_2 + x_{1d};$$

en dérivant par rapport à t, on obtient :

$$\ddot{p}_1 = -\dot{x}_1 + w\dot{p}_1 - a\dot{p}_2,$$

soit,

$$\ddot{p}_1 = -(-wx_1 + ax_2 - \frac{b^2}{k}p_1) + w(-x_1 + wp_1 - ap_2 + x_{1d}) - a(-x_2 - ap_1 + wp_2 + x_{2d}),$$

$$\ddot{p}_1 = wx_1 - ax_2 + \frac{b^2}{k}p_1 - wx_1 + w^2p_1 - awp_2 + wx_{1d} + ax_2 + a^2p_1 - wap_2 - ax_{2d},$$

d'où

$$\ddot{p}_1 = (w^2 + a^2 + \frac{b^2}{k})p_1 - 2awp_2 + wx_{1d} - ax_{2d}. \tag{5.14}$$

De la même manière, on obtient :

$$\dot{p}_2 = -x_2 - ap_1 + wp_2 + x_{2d},$$

soit

$$\ddot{p}_2 = -\dot{x}_2 - a\dot{p}_1 + w\dot{p}_2,$$

$$\ddot{p}_2 = -(ax_1 - wx_2 - \frac{b^2}{k}p_2) - a(-x_1 + wp_1 - ap_2 + x_{1d}) + w(-x_2 - ap_1 + wp_2 + x_{2d}),$$

$$\ddot{p}_2 = -ax_1 + wx_2 + \frac{b^2}{k}p_2 + ax_1 - awp_1 + a^2p_2 - ax_{1d} - wx_2 - wap_1 + w^2p_2 + wx_{2d};$$

par conséquent

$$\ddot{p}_2 = (w^2 + a^2 + \frac{b^2}{k})p_2 - 2awp_1 - ax_{1d} + wx_{2d}. \tag{5.15}$$

Dérivons deux fois l'équation (5.14), on obtient :

$$p_1^{(4)} = (w^2 + a^2 + \frac{b^2}{k})\ddot{p}_1 - 2aw\ddot{p}_2,$$

$$p_1^{(4)} = (w^2 + a^2 + \frac{b^2}{k})\ddot{p}_1 - 2aw[(w^2 + a^2 + \frac{b^2}{k})p_2 - 2awp_1 - ax_{1d} + wx_{2d}],$$

d'où

$$p_1^{(4)} = (w^2 + a^2 + \frac{b^2}{k})\ddot{p}_1 - 2aw(w^2 + a^2 + \frac{b^2}{k})p_2 + 4a^2w^2p_1 + 2wa^2x_{1d} - 2aw^2x_{2d}. \tag{5.16}$$

(5.14) entraîne :

$$-2awp_2 = \ddot{p}_1 - (w^2 + a^2 + \frac{b^2}{k})p_1 - wx_{1d} + ax_{2d}. \tag{5.17}$$

En injectant (5.17) dans (5.16), on obtient :

$$\begin{aligned} p_1^{(4)} &= (w^2 + a^2 + \frac{b^2}{k})\ddot{p}_1 + (w^2 + a^2 + \frac{b^2}{k})[\ddot{p}_1 - (w^2 + a^2 + \frac{b^2}{k})p_1 - wx_{1d} + ax_{2d}] + 4a^2w^2p_1 \\ &+ 2wa^2x_{1d} - 2aw^2x_{2d}, \end{aligned}$$

d'où

$$\begin{aligned} p_1^{(4)} &- 2(w^2+a^2+\frac{b^2}{k})\ddot{p}_1 + [(w^2+a^2+\frac{b^2}{k})^2 - 4a^2w^2]p_1 \\ &= w(a^2-w^2-\frac{b^2}{k})x_{1d} + a(a^2-w^2+\frac{b^2}{k})x_{2d}. \end{aligned} \tag{5.18}$$

L'équation caractéristique correspondant à l'équation (5.18) s'écrit comme suit :

$$C^4 - 2(w^2+a^2+\frac{b^2}{k})C^2 - [4a^2w^2 - (w^2+a^2+\frac{b^2}{k})^2] = 0.$$

Les racines de l'équation caractéristique sont données par :

$$C_1^2 = (a-w)^2 + \frac{b^2}{k},$$

$$C_2^2 = (a+w)^2 + \frac{b^2}{k}.$$

D'où, on obtient

$$p_1(t) = \lambda e^{C_1 t} + \beta e^{-C_1 t} + \gamma e^{C_2 t} + \alpha e^{-C_2 t} + \nu. \tag{5.19}$$

Pour déterminer ν, on a :

$$\dot{p}_1(t) = \lambda C_1 e^{C_1 t} - \beta C_1 e^{-C_1 t} + \gamma C_2 e^{C_2 t} - \alpha C_2 e^{-C_2 t},$$

$$\ddot{p}_1(t) = \lambda C_1^2 e^{C_1 t} + \beta C_1^2 e^{-C_1 t} + \gamma C_2^2 e^{C_2 t} + \alpha C_2^2 e^{-C_2 t},$$

$$p_1^{(3)}(t) = \lambda C_1^3 e^{C_1 t} - \beta C_1^3 e^{-C_1 t} + \gamma C_2^3 e^{C_2 t} - \alpha C_2^3 e^{-C_2 t},$$

$$p_1^{(4)}(t) = \lambda C_1^4 e^{C_1 t} + \beta C_1^4 e^{-C_1 t} + \gamma C_2^4 e^{C_2 t} + \alpha C_2^4 e^{-C_2 t}.$$

On remplacant dans (5.18), on obtient :

$$\begin{aligned} \lambda &\ [C_1^4 - 2C_1^2(w^2+a^2+\frac{b^2}{k}) + ((w^2+a^2+\frac{b^2}{k}) - 4a^2w^2)]e^{C_1 t} \\ + &\ \beta[C_1^4 - 2C_1^2(w^2+a^2+\frac{b^2}{k}) + ((w^2+a^2+\frac{b^2}{k}) - 4a^2w^2)]e^{-C_1 t} \\ + &\ \gamma[C_2^4 - 2C_2^2(w^2+a^2+\frac{b^2}{k}) + ((w^2+a^2+\frac{b^2}{k}) - 4a^2w^2)]e^{C_2 t} \\ + &\ \alpha[C_2^4 - 2C_2^2(w^2+a^2+\frac{b^2}{k}) + ((w^2+a^2+\frac{b^2}{k}) - 4a^2w^2)]e^{-C_2 t} \\ + &\ ((w^2+a^2+\frac{b^2}{k}) - 4a^2w^2)\nu \\ = &\ w(a^2-w^2-\frac{b^2}{k})x_{1d} + a(a^2-w^2+\frac{b^2}{k})x_{2d}. \end{aligned}$$

Par identification, on obtient la valeur exacte de ν donnée par :

$$\nu = \frac{w(a^2 - w^2 - \frac{b^2}{k})x_{1d} + a(a^2 - w^2 + \frac{b^2}{k})x_{2d}}{(a^2 + w^2 + \frac{b^2}{k})^2 - 4a^2w^2}. \tag{5.20}$$

On en déduit aisément $p_2(t)$ de (5.17), et on obtient :

$$p_2(t) = -\frac{1}{2aw}\ddot{p}_1 + \frac{1}{2aw}(w^2 + a^2 + \frac{b^2}{k})p_1 + \frac{1}{2a}x_{1d} - \frac{1}{2w}x_{2d},$$

$$\begin{aligned} p_2(t) &= -\frac{1}{2aw}(\lambda C_1^2 e^{C_1 t} + \beta C_1^2 e^{-C_1 t} + \gamma C_2^2 e^{C_2 t} + \alpha C_2^2 e^{-C_2 t}) \\ &+ \frac{1}{2aw}(w^2 + a^2 + \frac{b^2}{k})(\lambda e^{C_1 t} + \beta e^{-C_1 t} + \gamma e^{C_2 t} + \alpha e^{-C_2 t} + \nu) + \frac{1}{2a}x_{1d} - \frac{1}{2w}x_{2d}, \end{aligned}$$

d'où

$$\begin{aligned} p_2(t) &= \lambda(\frac{w^2 + a^2 + \frac{b^2}{k} - C_1^2}{2aw})e^{C_1 t} + \beta(\frac{w^2 + a^2 + \frac{b^2}{k} - C_1^2}{2aw})e^{-C_1 t} \\ &+ \gamma(\frac{w^2 + a^2 + \frac{b^2}{k} - C_2^2}{2aw})e^{C_2 t} + \alpha(\frac{w^2 + a^2 + \frac{b^2}{k} - C_2^2}{2aw})e^{-C_2 t} \\ &+ \frac{\nu}{2aw}(w^2 + a^2 + \frac{b^2}{k}) + \frac{1}{2a}x_{1d} - \frac{1}{2w}x_{2d}. \end{aligned} \tag{5.21}$$

De même compte tenu des équations suivantes :

$$x_1(t) = -\dot{p}_1 + wp_1 - ap_2 + x_{1d}, \tag{5.22}$$

$$x_2(t) = -\dot{p}_2 - ap_1 + wp_2 + x_{2d}. \tag{5.23}$$

En remplaçant (5.19), (5.21), $\dot{p}_1$ et $\dot{p}_2$ dans (5.22) et (5.23), on obtient les résultats suivants :

$$\begin{aligned} x_1(t) &= -[\lambda C_1 e^{C_1 t} - \beta C_1 e^{-C_1 t} + \gamma C_2 e^{C_2 t} - \alpha C_2 e^{-C_2 t}] + w[\lambda e^{C_1 t} + \beta e^{-C_1 t} + \gamma e^{C_2 t} + \alpha e^{-C_2 t} + \nu] \\ &- a[\lambda(\frac{w^2 + a^2 + \frac{b^2}{k} - C_1^2}{2aw})e^{C_1 t} + \beta(\frac{w^2 + a^2 + \frac{b^2}{k} - C_1^2}{2aw})e^{-C_1 t} + \gamma(\frac{w^2 + a^2 + \frac{b^2}{k} - C_2^2}{2aw})e^{C_2 t} \\ &+ \alpha(\frac{w^2 + a^2 + \frac{b^2}{k} - C_2^2}{2aw})e^{-C_2 t} + \frac{\nu}{2aw}(w^2 + a^2 + \frac{b^2}{k}) + \frac{1}{2a}x_{1d} - \frac{1}{2w}x_{2d}] + x_{1d}. \end{aligned}$$

D'ou

$$\begin{aligned} x_1(t) &= \lambda(\frac{w^2 - a^2 + C_1^2 - \frac{b^2}{k}}{2w} - C_1)e^{C_1 t} + \beta(\frac{w^2 - a^2 + C_1^2 - \frac{b^2}{k}}{2w} + C_1)e^{-C_1 t} \\ &+ \gamma(\frac{w^2 - a^2 + C_2^2 - \frac{b^2}{k}}{2w} - C_2)e^{C_2 t} + \alpha(\frac{w^2 - a^2 + C_2^2 - \frac{b^2}{k}}{2w} + C_2)e^{-C_2 t} \\ &+ \frac{(w^2 - a^2 - \frac{b^2}{k})}{2w}\nu + \frac{1}{2}x_{1d} + \frac{a}{2w}x_{2d}, \end{aligned} \tag{5.24}$$

et

$$\begin{aligned} x_2(t) &= -[\lambda(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})C_1e^{C_1t}+\beta(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})C_1e^{-C_1t}-\gamma(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})C_2e^{C_2t} \\ &+ \alpha(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})C_2e^{-C_2t}]-a[\lambda e^{C_1t}+\beta e^{-C_1t}+\gamma e^{C_2t}+\alpha e^{-C_2t}+\nu] \\ &+ w[\lambda(\frac{w^2+a^2+\frac{b^2}{k}-C_1^2}{2aw})e^{C_1t}+\beta(\frac{w^2+a^2+\frac{b^2}{k}-C_1^2}{2aw})e^{-C_1t}+\gamma(\frac{w^2+a^2+\frac{b^2}{k}-C_2^2}{2aw})e^{C_2t} \\ &+ \alpha(\frac{w^2+a^2+\frac{b^2}{k}-C_2^2}{2aw})e^{-C_2t}+\frac{\nu}{2aw}(w^2+a^2+\frac{b^2}{k})+\frac{1}{2a}x_{1d}-\frac{1}{2w}x_{2d}]+x_{2d}. \end{aligned}$$

Donc

$$\begin{aligned} x_2(t) &= \lambda[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a}-C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{C_1t}+\beta[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} \\ &+ C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{-C_1t}+\gamma[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a}-C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{C_2t} \\ &+ \alpha[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a}+C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{-C_2t}+(\frac{w^2-a^2+\frac{b^2}{k}}{2a})\nu+\frac{w}{2a}x_{1d} \\ &+ \frac{1}{2}x_{2d}. \end{aligned} \tag{5.25}$$

L'expression de $u_1(t)$ et $u_2(t)$ sont données par :

$$u_1(t)=-\frac{b^2}{k}[\lambda e^{c_1t}+\beta e^{-c_1t}+\gamma e^{c_2t}+\alpha e^{-c_2t}+\nu],$$

et

$$\begin{aligned} u_2(t) &= -\frac{b^2}{k}[(\frac{w^2+a^2-c_1^2+\frac{b^2}{k}}{2aw})e^{c_1t}\lambda+(\frac{w^2+a^2-c_1^2+\frac{b^2}{k}}{2aw})e^{-c_1t}\beta \\ &+ (\frac{w^2+a^2-c_2^2+\frac{b^2}{k}}{2aw})e^{c_2t}\gamma+(\frac{w^2+a^2-c_2^2+\frac{b^2}{k}}{2aw})e^{-c_2t}\alpha \\ &+ (\frac{w^2+a^2+\frac{b^2}{k}}{2aw})\nu+\frac{1}{2a}x_{1d}-\frac{1}{2w}x_{2d}]. \end{aligned}$$

Les constantes étant déterminées par les conditions aux limites suivantes :

$x_1(0)=0.5,\ x_2(0)=0.5,\ p_1(T)=s_1,\ p_2(T)=s_2,\ x_1(T)=0.5,\ x_2(T)=0.5.$

On résoud le système linéaire numériquement :

$$
\begin{cases}
0.5 = \lambda(\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} - C_1) + \beta(\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} + C_1) + \gamma(\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} - C_2) \\
+\alpha(\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} + C_2) + \frac{(w^2-a^2-\frac{b^2}{k})}{2w}\nu + \frac{1}{2}x_{1d} + \frac{a}{2w}x_{2d}, \\
\\
0.5 = \lambda[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} - C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{C_1 t} + \beta[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} + C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{-C_1 t} \\
+\gamma[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} - C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{C_2 t} + \alpha[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} + C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{-C_2 t} \\
+(\frac{w^2-a^2+\frac{b^2}{k}}{2a})\nu + \frac{w}{2a}x_{1d} + \frac{1}{2}x_{2d}, \\
\\
0.5 = \lambda(\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} - C_1)e^{2C_1} + \beta(\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} + C_1)e^{-2C_1} + \gamma(\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} - C_2)e^{2C_2} \\
+\alpha(\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} + C_2)e^{-2C_2} + \frac{(w^2-a^2-\frac{b^2}{k})}{2w}\nu + \frac{1}{2}x_{1d} + \frac{a}{2w}x_{2d}, \\
\\
0.5 = \lambda[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} - C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{2C_1} + \beta[\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} + C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{-2C_1} \\
+\gamma[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} - C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{2C_2} + \alpha[\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} + C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})]e^{-2C_2} \\
+(\frac{w^2-a^2+\frac{b^2}{k}}{2a})\nu + \frac{w}{2a}x_{1d} + \frac{1}{2}x_{2d}, \\
\\
s_1 = \lambda e^{2C_1} + \beta e^{-2C_1} + \gamma e^{2C_2} + \alpha e^{-2C_2} + \nu, \\
\\
s_2 = \lambda(\frac{w^2+a^2+\frac{b^2}{k}-C_1^2}{2aw})e^{2C_1} + \beta(\frac{w^2+a^2+\frac{b^2}{k}-C_1^2}{2aw})e^{-2C_1} + \gamma(\frac{w^2+a^2+\frac{b^2}{k}-C_2^2}{2aw})e^{2C_2} \\
+\alpha(\frac{w^2+a^2+\frac{b^2}{k}-C_2^2}{2aw})e^{-2C_2} + \frac{\nu}{2aw}(w^2+a^2+\frac{b^2}{k}) + \frac{1}{2a}x_{1d} - \frac{1}{2w}x_{2d}.
\end{cases}
\tag{5.26}
$$

Soit encore sous la forme matricielle :

$$\begin{pmatrix} a_1 & a_2 & a_3 & a_4 & 0 & 0 \\ b_1 & b_2 & b_3 & b_4 & 0 & 0 \\ d_1 & d_2 & d_3 & d_4 & 0 & 0 \\ v_1 & v_2 & v_3 & v_4 & 0 & 0 \\ e^{2C_1} & e^{-2C_1} & e^{2C_2} & e^{-2C_2} & -1 & 0 \\ h_1 & h_2 & h_3 & h_4 & 0 & -1 \end{pmatrix} \begin{pmatrix} \lambda \\ \beta \\ \gamma \\ \alpha \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ -\nu \\ f_5 \end{pmatrix},$$

avec

$$a_1 = \frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} - C_1,\ a_2 = \frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} + C_1,\ a_3 = \frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} - C_2,$$

$$a_4 = \frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} + C_2,\ f_1 = 0.5 - \frac{(w^2-a^2-\frac{b^2}{k})}{2w}\nu - \frac{1}{2}x_{1d} - \frac{a}{2w}x_{2d},$$

$$f_2 = 0.5 - \frac{(w^2-a^2+\frac{b^2}{k})}{2a}\nu - \frac{1}{2}x_{2d} - \frac{w}{2a}x_{1d},\ f_3 = 0.5 - \frac{(w^2-a^2-\frac{b^2}{k})}{2w}\nu - \frac{1}{2}x_{1d} - \frac{a}{2w}x_{2d},$$

$$f_4 = 0.5 - \frac{(w^2-a^2+\frac{b^2}{k})}{2a}\nu - \frac{1}{2}x_{2d} - \frac{w}{2a}x_{1d},\ f_5 = -\frac{(w^2+a^2+\frac{b^2}{k})}{2aw}\nu - \frac{1}{2a}x_{1d} + \frac{1}{2w}x_{2d},$$

$$b_1 = \frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} - C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw}),\ b_2 = \frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} + C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw}),$$

$$b_3 = \frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} - C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw}),\ b_4 = \frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} + C_2(\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw}),$$

$$d_1 = (\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} - C_1)e^{2C_1},\ d_2 = (\frac{w^2-a^2+C_1^2-\frac{b^2}{k}}{2w} + C_1)e^{-2C_1},\ d_3 = (\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} + C_2)e^{2C_2},$$

$$d_4 = (\frac{w^2-a^2+C_2^2-\frac{b^2}{k}}{2w} + C_2)e^{-2C_2},\ v_1 = [\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} - C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{2C_1},$$

$$v_2 = [\frac{w^2-a^2-C_1^2+\frac{b^2}{k}}{2a} + C_1(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{-2C_1},\ v_3 = [\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} - C_2(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{2C_2},$$

$$v_4 = [\frac{w^2-a^2-C_2^2+\frac{b^2}{k}}{2a} + C_2(\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})]e^{-2C_2},\ h_1 = (\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})e^{2C_1},\ h_2 = (\frac{w^2+a^2-C_1^2+\frac{b^2}{k}}{2aw})e^{2C_1},$$

$h_3 = (\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})e^{2C_2},\ h_4 = (\frac{w^2+a^2-C_2^2+\frac{b^2}{k}}{2aw})e^{-2C_2},$

Comparaison des deux approches

Cette section est consacrée à la comparaison des méthodes analytiques et numériques dans le cas ($n = 2$ et $T = 2$), afin de valider la solution numérique. Les expériences numériques ont été réalisées pour $k = 2.5,\ w = 2; a = 0.5; b = 1,\ x_{1d} = x_{2d} = 0.2$. On déduit que la solution exacte et la solution numérique sont concordantes (voir Figure 5.2 et Figure 5.3).

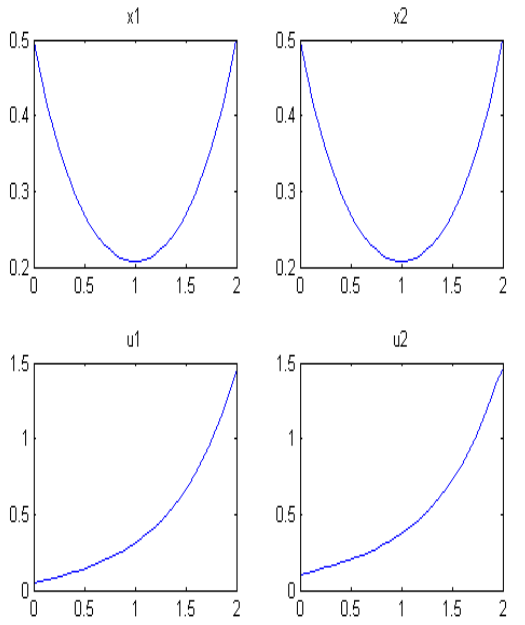

FIG. 5.2 – Solution exacte sans contraintes

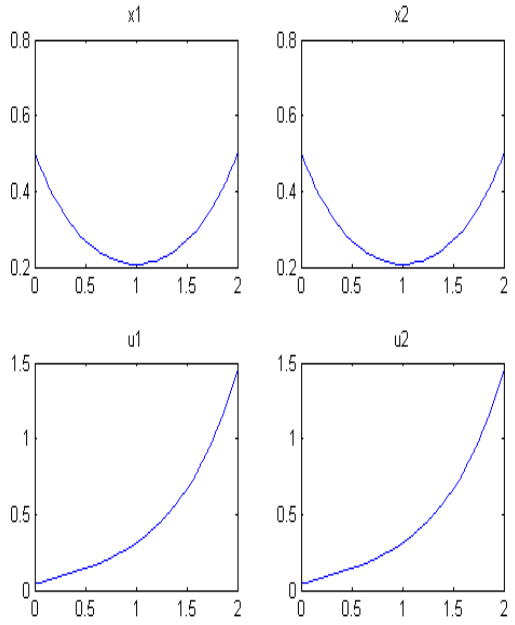

FIG. 5.3 – Solution numérique sans contraintes

Les performances de la procédure numérique sont résumées dans le Tableau 1 ci dessous, pour différentes valeurs de k. Notons que la convergence est rapide et que de plus, le temps de calcul est très faible.

k	C.P.U. temps	Nombre d'itérations
0.5	0.1248	2
1	0.1716	2
1.5	0.0468	1
2	0.1092	1
2.5	0.1092	1

Table 1 : Nombre d'itérations nécessaires à la convergence et temps de calcul en secondes dans le cas sans contraintes (n=2).

5.5.2 Cas avec contrainte

En présence de contrainte sur l'état, le calcul de la solution exacte par une méthode analytique est difficile à effectuer. Nous nous limiterons ici à la recherche d'une solution numérique. Les données sont les mêmes que celles considérées dans le cas sans contrainte sauf qu'en plus du cas où $n = 2$ et $T = 2$, nous considérons également le cas $n = 5$ et $T = 4$. L'algorithme est analogue à celui proposé en section 3 en rajoutant les contraintes sur l'état $x(t)$ définies par :

$$si\ x > x^{max},\ alors\ x = x^{max}\ sinon\ si\ x < x^{min}\ alors\ x = x^{min}.$$

Pour $x_i^{min} = 0.35\ et\ x_i^{max} = 0.5$ pour $i \in \{1, ..., 5\}$, les résultats numériques sont présentés dans le Tableau 2 et les valeurs de l'état et le contrôle sont représentées sur la Figure 5.5 (pour le cas $n = 2, T = 2$) et en Figure 5.7 (pour le cas $n = 5, T = 4$). Comme précédement la convergence, exprimée en nombre d'itérations est rapide et les temps de calculs très faibles. Notons que dans les résultats obtenus, les contraintes sont saturées, compte-tenu des paramètres utilisées.

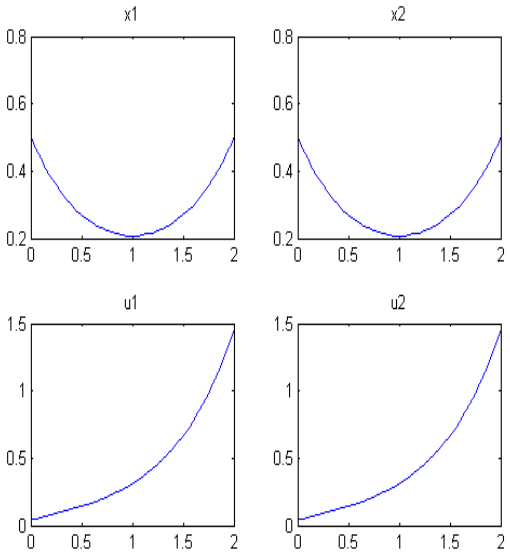

FIG. 5.4 – Solution numérique sans contrainte (n=2)

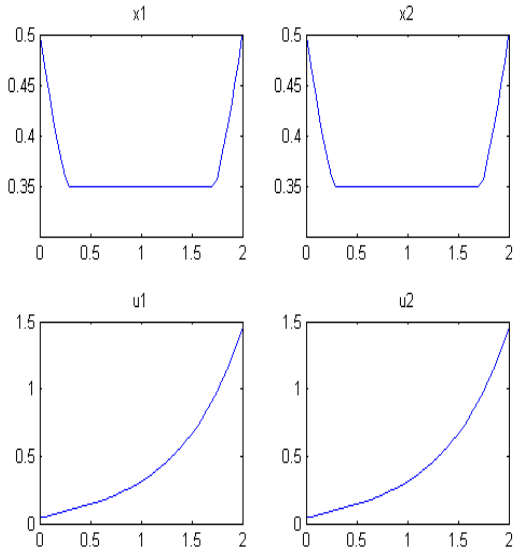

FIG. 5.5 – Solution numérique avec contrainte (n=2)

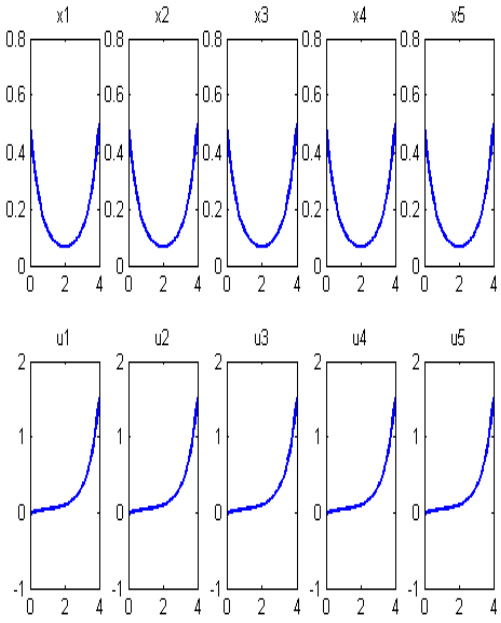

FIG. 5.6 – Solution numérique sans contrainte (n=5)

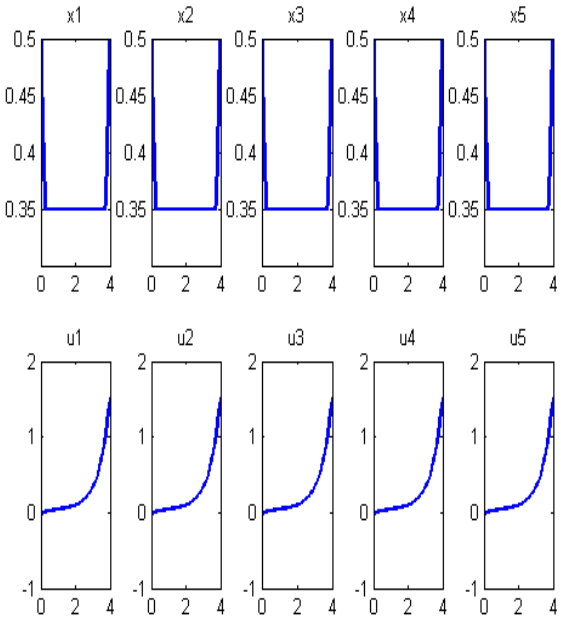

FIG. 5.7 – Solution numérique avec contrainte (n=5)

k	C.P.U. temps	Nombre d'itérations
0.5	0.1872	2
1	0.1716	2
1.5	0.1872	2
2	0.1716	2
2.5	0.2340	2

Table 2 : Nombre d'itérations nécessaires à la convergence et temps de calcul en secondes dans le cas avec contraintes (n=2).

Nous calculons également la solution numérique des systèmes à grande dimension. Le Tableau 3 représente pour différentes valeurs de n le nombre d'itérations et le temps nécessaire pour la convergence.

n	C.P.U. temps	Nombre d'itérations
10	3.8220	3
20	5.0388	3
50	10.9825	3
100	43.6179	4

Table 3 : Nombre d'itérations nécessaires à la convergence et temps de calcul en secondes dans le cas avec contraintes.

5.6 Conclusion

Dans ce chapitre, nous avons proposé, sous certaines hypothèses, l'utilisation d'un algorithme de relaxation pour la résolution d'un problème de contrôle optimal non linéaire dans le cas où il y a une contrainte sur l'état et l'état final. Nous avons testé cette méthode sur un problème modèle et il s'avère que la convergence est rapide et que les temps de calculs sont petits.

Chapitre 6

Application de l'algorithme à l'étude de la régulation d'un processus thermique

6.1 Introduction

Dans ce chapitre, nous appliquons la méthode numérique étudiée au chapitre 5 pour la détermination de la commande d'un grand système thermique lorsque l'état est soumis à certaines contraintes et la valeur de l'état final est fixé. Le système est composé d'un four vertical, dans la cheminée duquel est placé un barreau.

Ce système présente de grands couplages internes en raison de la convection naturelle dans la cheminée et de la conduction thermique ; l'objectif est de maintenir, malgré les perturbations, une distribution prescrite de température sur un objet vertical placé dans la cheminée, les observations étant réalisées en n points équidistants. Le modèle du processus thermique est décrit grâce à une équation d'état linéaire avec un critère quadratique à minimiser. En présence de contraintes sur l'état, nous utiliserons la notion de sous-différentiel permettant de prendre en compte, si nécessaire, la projection sur l'ensemble convexe des contraintes. Par la suite nous reformulerons les conditions d'optimalité issues du principe de Pontryagin, on aboutit à la résolution d'un système algébro-différentiel. En raison des propriétés de la monotonie d'une part de sous-différentiel de la fonction indicatrice du convexe et d'autre part, sous des hypothèses convenables satisfaites par les dérivées de dérivation

de l'état et de l'état adjoint par rapport au temps, nous analysons la convergence de la méthode itérative considérée.

Ce chapitre est organisé comme suit : Dans la section 2, nous définissons le problème dans les deux cas, avec et sans contraintes sur l'état. La section 3 est consacrée à la description de la méthode de relaxation couplée à la méthode de tir. La convergence de la méthode proposée est présentée à la section 4, tandis que la section 5 contient les résultats des expériences numériques.

6.2 Régulation d'un processus thermique

6.2.1 Cas sans contraintes sur l'état

On considère un système physique composé d'un four vertical, dans la cheminée duquel est placé un barreau (voir Figure 6.1) ; le but de l'étude est d'amener la température z relevée en n points du barreau à une température z_d, en un temps fini T, qui représentera l'horizon de commande. Dans la suite, $x(t) \in \mathbb{R}^n$ représente la température de la cheminée en n points. $u(t) \in \mathbb{R}^m$ est le vecteur contrôle qui représente l'intensité des courants envoyés dans chacun des n enroulements de chauffage. On aura donc à déterminer la commande u, de telle sorte que, au bout du temps T, la température z du barreau soit uniformément égal à z_d, compte-tenu d'un critère à minimiser qui sera précisé ultérieurement.

Soit $y = (z_1, x_1, \ldots, z_i, x_i, \ldots, z_n, x_n)$ le vecteur d'état du système ; alors le modèle mathématique est obtenu par linéarisation de l'équation de la chaleur autour d'un point de fonctionnement est représenté sous la forme suivante :

$$\begin{cases} \dot{y}(t) = Ay(t) + Bu(t),\ t \in [0, T], \\ y(0) = y_0,\ y(T) = y_f, \\ u(t) \in U, \end{cases} \tag{6.1}$$

où $y(0) = y_0$ est l'état initial donné, $y(T) = y_f$ est l'état final. U est l'ensemble des commandes admissibles, qui est un ensemble ouvert. $A \in \mathbb{R}^{2n \times 2n}$ et $B \in \mathbb{R}^{2n \times m}$ sont deux matrices constantes données. L'observation du système est constituée par une partie du

vecteur y à savoir le vecteur z ; on pose

$$z = C\ y, \tag{6.2}$$

où $C \in \mathbb{R}^{n\times 2n}$ est la matrice d'observation. Le problème est complété par la minimisation de la fonction objective $J(u)$:

$$J(u) = \frac{1}{2}\int_0^T [\frac{\alpha}{\|z_d\|_2^2}\|z - z_d\|_2^2 + \frac{\beta}{\|u_d\|_2^2}\|u - u_d\|_2^2]dt, \tag{6.3}$$

où u_d est la commande conduisant asymptotiquement à la température prescrite z_d donnée par

$$u_d = -(CA^{-1}B)^{-1}z_d.$$

Les deux coefficients α et β indiquent le dosage entre les deux composantes de la fonction coût, i.e., les valeurs α et β sont choisies afin de fournir respectivement plus de poids à la précision et à l'énergie dépensée.

On cherche une commande admissible qui transfère le système d'un état initial y_0 vers un état final y_f fixé et minimisant la fonction coût donnée en (6.3). L'Hamiltonien du système $(6.1) - (6.3)$ est donnée par :

$$H(y, p, u, t) = \frac{1}{2}[\frac{\alpha}{\|z_d\|_2^2}\|z - z_d\|_2^2 + \frac{\beta}{\|u_d\|_2^2}\|u - u_d\|_2^2] + p^t[A\ y + B\ u],$$

où p est le vecteur d'état adjoint. Cherchons maintenant la commande $\hat{u}$ qui minimise l'Hamiltonien, tel que

$$H(\hat{y}, \hat{p}, \hat{u}) \leq H(\hat{y}, \hat{p}, u);\ \forall u \in U,\ \forall t \in [0, T].$$

Dans le cas sans contraintes sur l'état, les équations d'optimalité s'écrivent comme suite :

$$\begin{cases} \frac{dy}{dt} = \frac{\partial H}{\partial p} = Ay + Bu;\ y(0) = y_0,\ y(T) = y_f,\ \forall t \in [0, T], \\ -\frac{dp}{dt} = \frac{\partial H}{\partial y} = A^t p + C^t C y - C^T z_d,\ p(0) \text{ à déterminer}, \\ \frac{\partial H}{\partial u} = 0 = B^t p + k(u - u_d), \end{cases} \tag{6.4}$$

où

$k = \frac{\|z_d\|^2}{\|u_d\|^2}\frac{\beta}{\alpha}$.

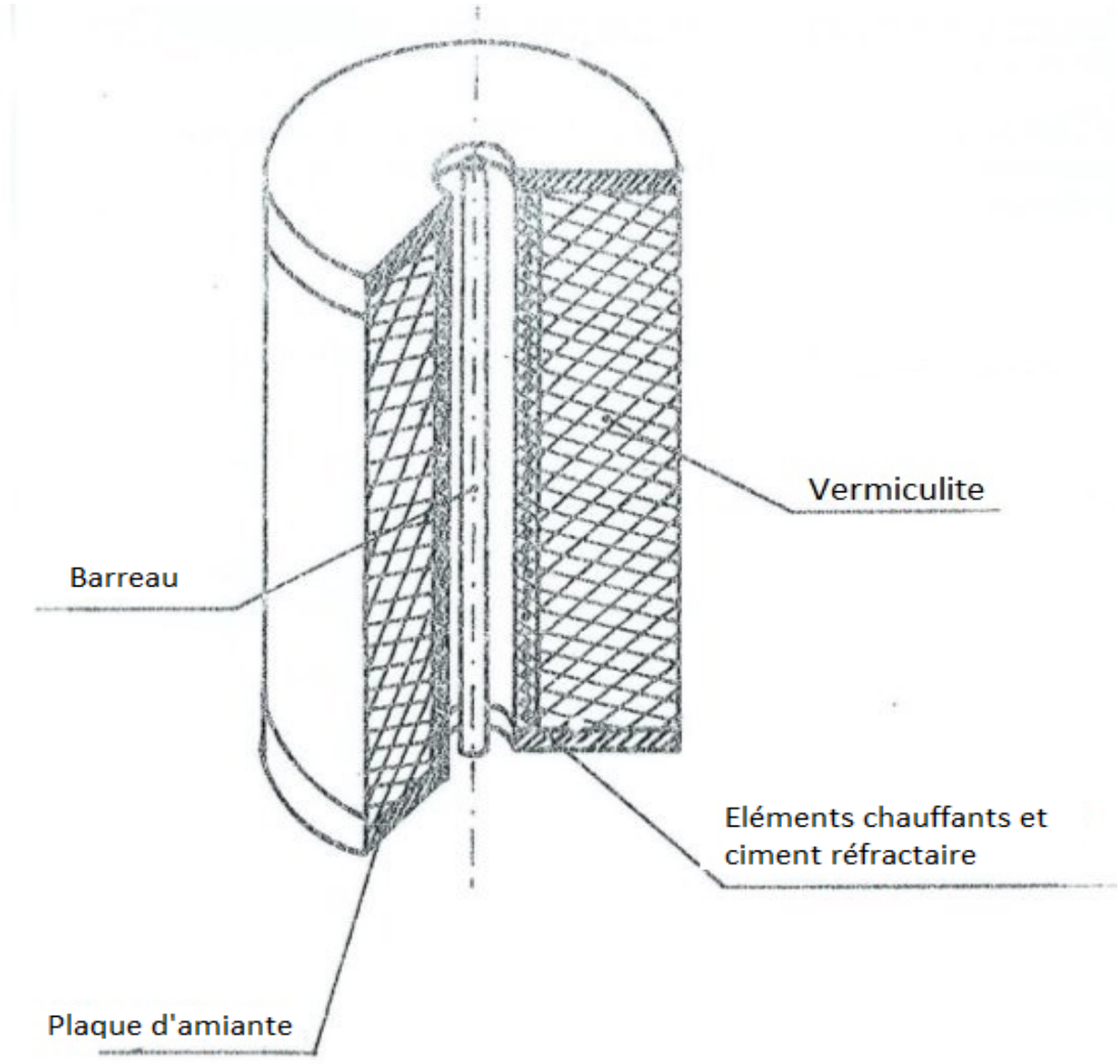

FIG. 6.1 – Représentation du four

6.2.2 Cas avec contraintes sur l'état

Dans le cas où l'état est soumis à certaines contraintes notées respectivement par y^{min} et y^{max}, soit Y_{ad} un ensemble convexe des trajectoires admissibles. Par la suite, nous reformulerons les conditions nécessaires d'optimalité ; pour cela on va donc utiliser la notion de sous-différentiel pour obtenir des conditions d'optimalité. Considérons le problème sous contraintes suivant :

$$\begin{cases} \dot{y}(t) = Ay(t) + Bu(t),\ t \in [0,T],\ y(t) \in Y_{ad}, \\ \\ y(0) = y_0,\ y(T) = y_f, \end{cases} \tag{6.5}$$

où l'ensemble convexe fermé Y_{ad} est définie par $Y_{ad} = \{y \in \mathbb{R}^n / y_i^{min} \leq y_i \leq y_i^{max},\ i = 1, \ldots, 2n\}$; notons $\partial\Psi_{Y_{ad}}$ le sous différentiel de la fonction indicatrice $\Psi_{Y_{ad}}$ définie au chapitre 5 et admet le graphe représenté par la Figure 5.1. Notons que le sous-différentiel $\partial\Psi_{Y_{ad}}$ est monotone. Appliquons le Lemme **1.1** du chapitre 1 : on cherche $\hat{u}$ qui minimise l'Hamiltonien H ; ceci peut s'écrire sous la forme :

$$0 \in \partial H(\hat{u});$$

puisque H est un opérateur continue [53], nous obtenons la nouvelle formulation des conditions nécessaires d'optimalité :

$$\begin{cases} 0 \in \frac{dy}{dt} + \partial\Psi_{Y_{ad}} - Ay - B\hat{u};\ y(0) = y_0,\ y(T) = y_f,\ \forall t \in [0,T], \\ \\ -\frac{dp}{dt} = \frac{\partial H}{\partial y} = A^t p + C^t C y - C^T z_d,\ p(0) \text{ à determiner}, \\ \\ \frac{\partial H}{\partial u} = 0 = B^t p + k(\hat{u} - u_d). \end{cases} \tag{6.6}$$

6.3 Méthode de résolution numérique

Pour la résolution du problème, avec et sans contraintes sur l'état, nous effectuons le couplage de la méthode de relaxation (voir [42], [51] et [67]) avec la méthode de tir [78], cette dernière étant destinée à calculer $p(0)$ nécessaire à la résolution du système algébro-différentiel obtenus par application du principe de minimum de Pontryagin. Les étapes de la méthode de résolution numérique sont résumées ci-dessous :

6.3.1 Cas avec contrainte

1. Approximation de la commande initiale u^0 pour $t \in [0, T]$, et de l'état adjoint initial $p^0(0)$,
2. $r \leftarrow 0$ (où r permet de compter les itérations),
3. **Tant que** $\mid u^{(r+1)} - u^{(r)} \mid > \epsilon$ (où ϵ définit le seuil de convergence) **faire** :
   - Détermination de l'équation d'état $y^{(r)}$, par intégration numérique de l'équation d'état avec projection sur le convexe Y_{ad} :

$$\begin{cases} \frac{d\bar{y}}{dt} = A\bar{y} + Bu^{(r)}, \quad 0 < t \leq T, \\ \\ \bar{y}(0) = y_0, \end{cases} \qquad et \quad y^{(r)} = Proj(\bar{y}), \tag{6.7}$$

   où $Proj(.)$ est l'opérateur de projection sur le convexe fermé Y_{ad}, puis détermination de l'état adjoint $p^{(r)}$ en résolvant :

$$\begin{cases} -\frac{dp^{(r)}}{dt} = A^t p^{(r)} + C^t(Cy^{(r)} - z_d), \\ \\ p^{(r)}(0), \end{cases} \tag{6.8}$$

   où $p^{(r)}(0)$ est calculé par la méthode de tir,
   - Détermination de la commande $u^{(r+1)}$:

$$u^{(r+1)} \leftarrow (u_d - \frac{1}{k} B^t p^{(r)}), \tag{6.9}$$

   - $Convergence \leftarrow \mid u^{(r+1)} - u^{(r)} \mid$,
   - Détermination de la fonction de tir :

$$G(p) = y^{(r)}(T) - y_f,$$

   - Résolution de l'équation de tir par la méthode de Newton et détermination de la nouvelle valeur de $p(0)$:

$$p^{(r+1)}(0) \leftarrow p^{(r)}(0) + correction,$$

   - $r \leftarrow r + 1$.

Fin de tant que.

6.3.2 Cas sans contrainte

La démarche est analogue sauf que l'étape (6.7) est remplacée par

$$\begin{cases} \frac{dy^{(r)}}{dt} = Ay^{(r)} + Bu^{(r)}, \quad 0 < t \leq T, \\ y(0) = y_0. \end{cases}$$

6.4 Convergence de la méthode

Ecrivons les équations d'optimalité sous forme matricielle :

$$\begin{pmatrix} \frac{dy}{dt} + \partial\Psi_{Y_{ad}} \\ -\frac{dp}{dt} \\ 0 \end{pmatrix} + \begin{pmatrix} \bar{A} & 0 & -B \\ -Q & \bar{A}^t & 0 \\ 0 & B^t & kI \end{pmatrix} \begin{pmatrix} y \\ p \\ u \end{pmatrix} \ni \begin{pmatrix} 0 \\ -C^t z_d \\ ku_d \end{pmatrix}, \quad y(0) = y_0,$$

où $\bar{A} = -A$, $Q = C^tC$ *et* I est la matrice identité. La valeur du paramètre $k > 0$ permet de réaliser le dosage entre la précision du calcul et la minimisation de l'énergie dépensée, pour réaliser la commande optimale. Le problème s'écrit donc comme la somme d'un système linéaire perturbé par une application diagonale. Notons Θ la matrice suivante :

$$\Theta = \begin{pmatrix} \bar{A} & 0 & -B \\ -Q & \bar{A}^t & 0 \\ 0 & B^t & kI \end{pmatrix}.$$

Remarque 6.1. Dans l'étude présentée, la principale hypothèse considérée, est que la matrice $\overline{A}$ est l'opposé d'une M-matrice. Ainsi, les parties réelles des valeurs propres de $\overline{A}$ sont négatives et par conséquent toute solution du problème est Lyapunov stable.

Proposition 6.1. *Si les conditions suivantes sont vérifiées :*

- $\bar{A}$ *est une M-matrice*
- $k \geq k_0 > 0$
- $p^2(0) - p^2(T) > 0$,

alors l'algorithme permettant de calculer numériquement la loi de commande optimale, par la méthode de relaxation couplée à la méthode de tir, converge quelque soit la donnée initiale u^0.

Preuve. La preuve est analogue à celle de la proposition 5.1.

6.5 Les expériences numériques

Les expérimentations numériques ont été effectuées pour la régulation de deux grands processus thermiques qui sont les fours à 3 et 12 zones de chauffage. La matrice A est déterminée par linéarisation et identification de l'équation de la chaleur autour d'un point de fonctionnement. La forme de la matrice B tient compte du fait que seul le contrôle est à l'intérieur de la cheminée. Par conséquent, la ligne correspondant à un indice impair est égal à zéro. La matrice C est définie de telle sorte que l'on peut extraire uniquement les composantes de la température du barreau. Les conditions initiales de l'état considéré correspondent aux conditions initiales réels du problème physique. Cependant, afin de satisfaire dans la suite les hypothèses permettant à la convergence de la méthode de relaxation couplée à la méthode de tir, il est nécessaire de considérer des conditions initiales sur l'état égal à zéro. Puisque, l'équation d'état est linéaire, nous pouvons satisfaire facilement cette hypothèse par un simple changement de variables.

La mise en œuvre de l'algorithme est obtenue en utilisant Matlab en conformité avec le code structurée donnée dans la section 3.

Dans la suite, notons que la compatibilité des conditions initiales peut être vérifiée en tenant compte des résultats des calculs et plus particulièrement par la vérification de la valeur de l'état final est conforme à la valeur désirée. Ceci est bien vérifiée pour toutes les expériences présentées dans cette section.

6.5.1 Le four à trois zones de chauffage

L'exemple étudié concerne la loi de commande optimale d'un four vertical avec trois zones de chauffage qui à six variables d'état et trois variables de contrôle ; dans notre cas

$n = 6$ et $m = 3$. Les constantes de temps sont en minutes et les contrôles sont en calories par minute.

T est égal à 10, $z_d = \begin{pmatrix} 30°c & 30°c & 30°c \end{pmatrix}$, et $u_d = \begin{pmatrix} 164.55 & 245.30 & 419.69 \end{pmatrix}$.

Les valeurs numériques des matrices A, B et C, sont données ci-dessous :

$$A = \begin{pmatrix} -0.030 & 0.013 & 0.0077 & 0.0071 & 0.00017 & 0.00065 \\ 0.0017 & -0.012 & 0.00009 & 0.00033 & 0.00008 & 0.00029 \\ 0.0075 & 0 & -0.040 & 0.016 & 0.0077 & 0.00073 \\ 0 & 0.0030 & 0.0019 & -0.014 & 0.00009 & 0.0033 \\ 0 & 0 & 0.0075 & 0 & -0.029 & 0.012 \\ 0 & 0 & 0 & 0.0030 & 0.0014 & -0.013 \end{pmatrix},$$

$$B = \begin{pmatrix} 0 & 0 & 0 \\ 0.00125 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.00125 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.00125 \end{pmatrix}, \; C = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Cas sans contraintes

Les performances de la procédure numérique sont résumées dans le Tableau 1 et représentés sur les Figures 6.2 et 6.3.

β/α	k	C.P.U. temps	Nombre d'itérations
1/10	0.0010	8.2837	17
1/4	0.0026	9.4537	18
1/2	0.0051	9.3601	18
1/1	0.0103	9.4069	19
2/1	0.0205	9.7501	20
4/1	0.0410	10.2025	21
10/1	0.1025	10.3897	22

Tableau 1 : Nombre d'itérations nécessaires à la convergence et temps C.P.U de calcul en secondes dans le cas sans contraintes.

Cas avec contraintes

Les contraintes pour les composantes paires sont définies comme suit :

$$\begin{cases} y_i^{min} = 0, & \text{for } \ i \in \{2,4,6\}, \\ y_2^{max} = 130, \\ y_4^{max} = 250, \\ y_6^{max} = 260, \end{cases}$$

Les résultats des expériences sont résumées dans le Tableau 2 et représentés sur les Figures 6.4 et 6.5. Pour ce qui concerne les résultats obtenus dans le cas du four avec 3 zones du

β/α	k	C.P.U. temps	Nombre d'itérations
1/10	0.0010	9.1729	17
1/4	0.0026	9.2041	18
1/2	0.0051	9.7345	18
1/1	0.0103	9.9685	19
2/1	0.0205	10.077	20
4/1	0.0410	10.8733	21
10/1	0.1025	11.2789	22

Table 2 : Nombre d'itérations nécessaires à la convergence et temps C.P.U de calcul en secondes dans le cas avec contraintes.

chauffage, nous pouvons observer sur les Figure 6.4 et 6.5 une situation où les contraintes sont saturées.

6.5.2 Le four à douze zones de chauffage

L'exemple étudié concerne la loi de commande optimale d'un four vertical avec douze zones de chauffage comportant 24 variables d'état et 12 variables de contrôle ; dans notre cas $n = 24$ et $m = 12$. Les constantes de temps sont en minutes et les contrôles sont en calories par minute.

Les expériences numériques ont été réalisées dans le cas où les composantes du vecteur z_d est égal à $30°C$, $T = 5mn$ et

$u_d = (1075.7 \ \ 826.3 \ \ 840.9 \ \ 842.7 \ \ 845.7 \ \ 850.7 \ \ 858.8 \ \ 872.2 \ \ 894.3 \ \ 930.7 \ \ 959.7 \ \ 1484.1).$

La matrice d'état A a la forme suivante :

$$A = \begin{pmatrix} D_1 & S+\theta & \epsilon S & \epsilon^2 S & . & . & . & \epsilon^k S & . & . & . & \epsilon^{10} S \\ \theta & D & S+\theta & \epsilon S & . & . & . & . & . & . & \epsilon^8 S & \epsilon^9 S \\ 0 & \theta & D & S+\theta & . & . & . & . & . & . & . & . \\ 0 & 0 & \theta & D & S+\theta & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & . & . & . & . & S+\theta & \epsilon S & \epsilon^2 S \\ . & . & . & . & . & . & . & . & . & . & S+\theta & \epsilon S \\ . & . & . & . & . & . & . & . & . & \theta & D & S+\theta \\ 0 & 0 & 0 & . & . & . & . & . & 0 & 0 & \theta & D_2 \end{pmatrix},$$

où $\epsilon = 0.67$ et chacun des blocs D_1, D, D_2, S et θ sont définis comme les matrices 2×2 suivantes :

$$D_1 = \begin{pmatrix} -0.38 & 0.196 \\ 0.0068 & -0.0629 \end{pmatrix}, \; D = \begin{pmatrix} -0.39 & 0.196 \\ 0.0068 & -0.0682 \end{pmatrix},$$

$$D_2 = \begin{pmatrix} -0.44 & 0.196 \\ 0.0068 & -0.0559 \end{pmatrix}, \; S = \begin{pmatrix} 0.0057 & 0.012 \\ 0.0003 & 0.0009 \end{pmatrix},$$

$$\theta = \begin{pmatrix} 0.0651 & 0 \\ 0 & 0.0030 \end{pmatrix}, \; S+\theta = \begin{pmatrix} 0.0708 & 0.012 \\ 0.0003 & 0.0039 \end{pmatrix}.$$

Les éléments des matrices B et C ont les valeurs suivantes :

$$b_{ij} = \begin{cases} 0.00195, & \text{si } i \text{ est pair et } j = i/2, \\ 0, & \text{sinon.} \end{cases}$$

$$c_{ij} = \begin{cases} 1, & \text{pour } 1 \leq i \leq 12 \text{ et } j = 2i - 1, \\ 0, & \text{sinon.} \end{cases}$$

Cas sans contraites

Les résultats des expériences sont résumées dans le Tableau 3 et représentés sur les Figures 6.6 et 6.7.

β/α	k	C.P.U. temps	Nombre d'itérations
1/10	9.8325×10^{-5}	17.1445	10
1/4	2.4581×10^{-4}	17.0509	10
1/2	4.9162×10^{-4}	19.6405	11
1/1	9.8325×10^{-4}	19.4533	12
2/1	0.0020	19.8901	12
4/1	0.0039	21.7933	13
10/1	0.0098	23.3377	14

Tableau 3 : Nombre d'itérations nécessaires à la convergence et temps C.P.U de calcul en secondes dans le cas sans contraintes.

Cas avec contraintes

Les contraintes pour les composantes paires sont définies comme suit :

$$\begin{cases} y_i^{min} = 0, & \text{pour } \ i \in \{2, 4, \ldots, 24\}, \\ y_2^{max} = 40, & \\ y_i^{max} = 35, & \text{pour } \ i = 2l \ \text{ et } \ l \in \{2, \ldots, 9\}, \\ y_i^{max} = 40, & \text{pour } \ i = 20 \ \text{ et } \ i = 22, \\ y_{24}^{max} = 60, & \end{cases}$$

Les résultats des expériences sont résumées dans le Tableau 4 et représentés sur les Figures 6.8 et 6.9.

β/α	k	C.P.U. temps	Nombre itérations
1/10	9.8325×10^{-5}	17.3473	10
1/4	2.4581×10^{-4}	17.0977	10
1/2	4.9162×10^{-4}	18.1117	11
1/1	9.8325×10^{-4}	19.5781	12
2/1	0.0020	20.4205	12
4/1	0.0039	21.1381	13
10/1	0.0098	21.3877	14

Tableau 4 : Nombre d'itérations nécessaires à la convergence et temps C.P.U de calcul en secondes dans le cas avec contraintes.

Commentaires des résultats numériques

D'après les résultats obtenus dans les deux cas, on remarque que, les conditions relatives à la valeur de l'état final $y(T) = y_f$ ainsi que l'état $y^{min} \leq y \leq y^{max}$ sont satisfaites.

La comparaison entre les cas avec et sans contraintes sur l'état, montre que dans le premier cas, les valeurs de la température de la cheminée sont saturées, comme dans le deuxième cas, la température de la cheminée est élevée. Néanmoins, dans les deux cas, l'algorithme fonctionne efficacement puisque la valeur fixée de l'état est atteinte. Pour la régulation du premier (respectivement seconde) four, il faut résoudre 12 (respectivement 48) équations différentielles ordinaires et 3 (respectivement 12) équations algébriques. Dans les Tableaux 1 et 2 (respectivement 3 et 4) on montre que 17 à 22 (respectivement 10 à 14) itérations sont nécessaires pour la convergence. Ainsi, pour le grand processus thermique considéré quelques itérations sont nécessaires pour la convergence et de plus, le temps de calcul est petit dans les deux cas.

Dans le travail précédent [50], il a été considéré les mêmes problèmes avec un modèle similaire, mais lorsque l'état n'est pas soumis à des contraintes et quand la valeur de l'état final est libre. Dans ce cas l'utilisation de la méthode de tir n'est pas nécessaire. Dans ces expériences, l'utilisation de la méthode de relaxation avec la méthode de descente et la méthode de gradient conjugué ont été comparées. Il apparaît que, le temps de calcul est plus important lorsque la méthode de descente et la méthode de gradient conjugué sont utilisées.

Par conséquent, dans notre cas, lorsque l'état est soumis à des contraintes et la valeur de l'état final est fixé, l'algorithme proposé est également bien adapté au contrôle du processus thermique.

Enfin, lorsque on compare les résultats obtenus par les deux simulations concernant les deux modèles mathématiques qui décrivent l'évolution du processus thermique, nous notons que les résultats obtenus avec le four à douze zones de chauffage sont plus réalistes. En effet, la division en douze zones est plus pertinente comparée à la division en trois zones.

6.6 Conclusion

Le but de ce chapitre concerne la régulation optimale d'un grand processus thermique, le four à 3 et 12 zones de chauffage. Nous avons utilisé la méthode de relaxation associée à la méthode de tir pour résoudre ce problème de contrôle optimal lorsque la variable d'état est soumis à certaines contraintes. La convergence de la procédure est assurée. La vitesse de convergence est élevée et le temps de calcul est rapide.

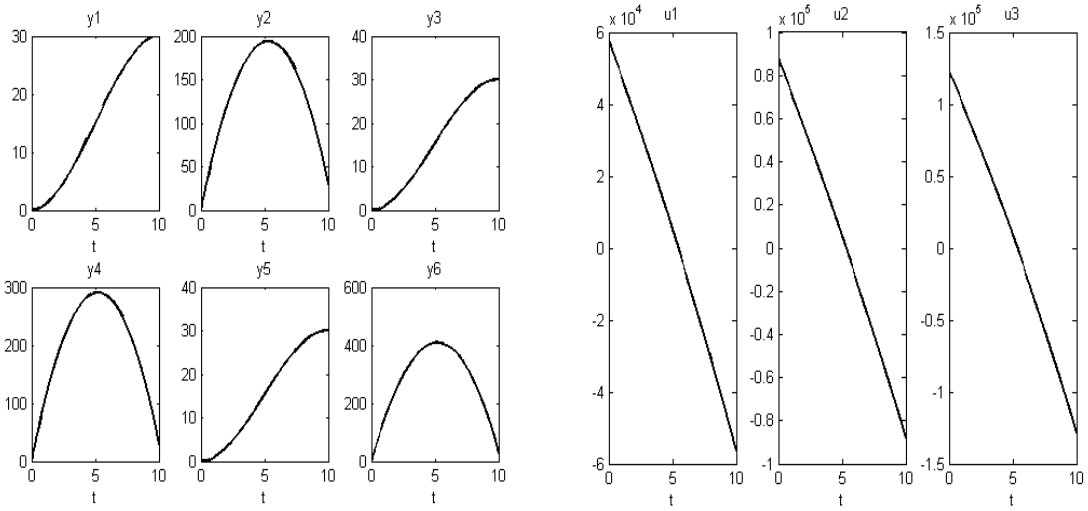

FIG. 6.2 – L'état et le contrôle pour $\beta = 10$ et $\alpha = 1$

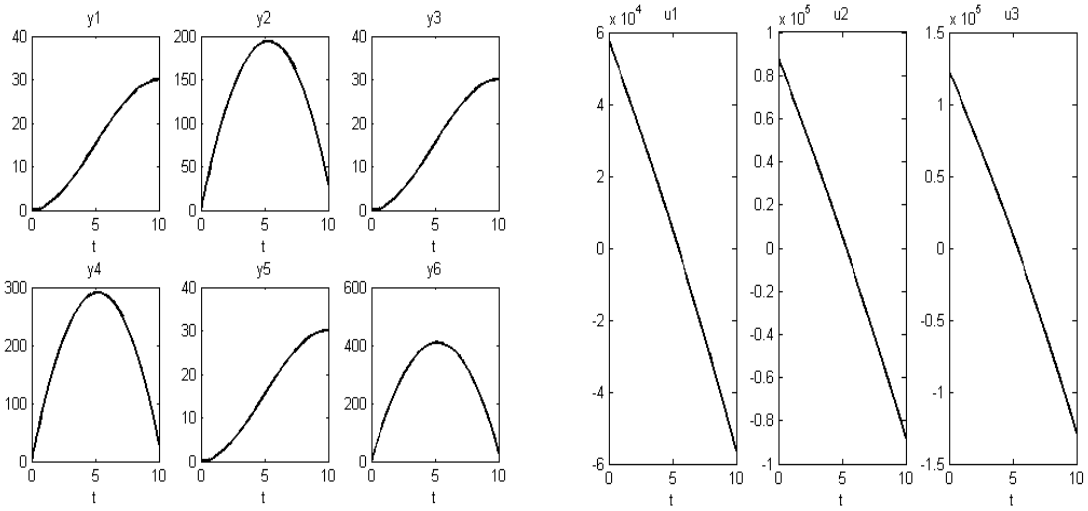

FIG. 6.3 – L'état et le contrôle pour $\beta = 1$ et $\alpha = 10$

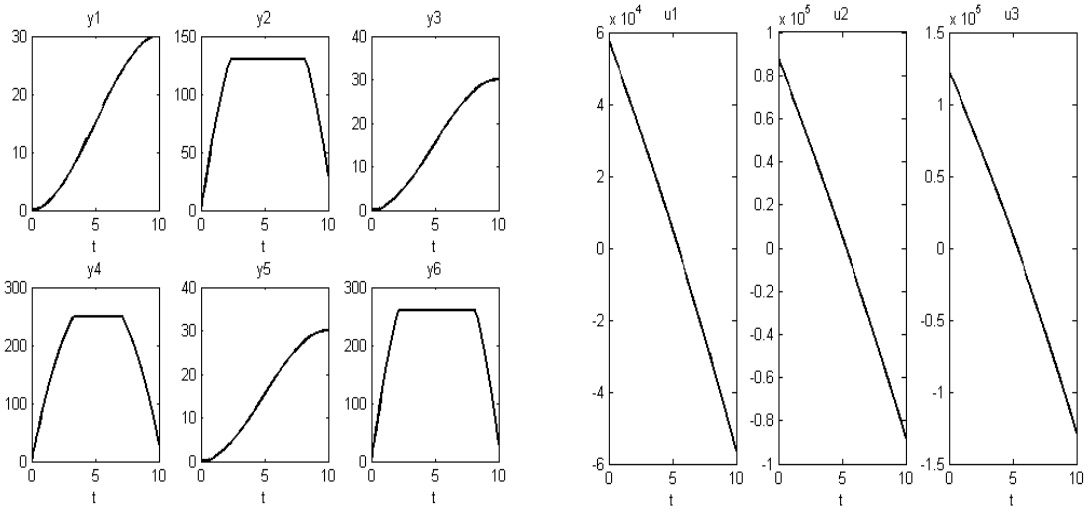

FIG. 6.4 – L'état et le contrôle pour $\beta = 10$ et $\alpha = 1$

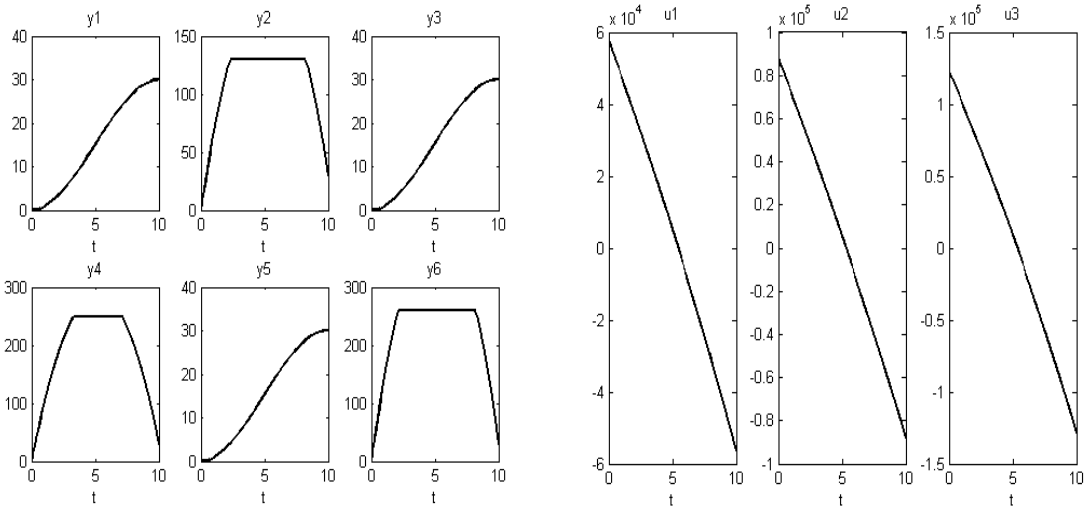

FIG. 6.5 – L'état et le contrôle pour $\beta = 1$ et $\alpha = 10$

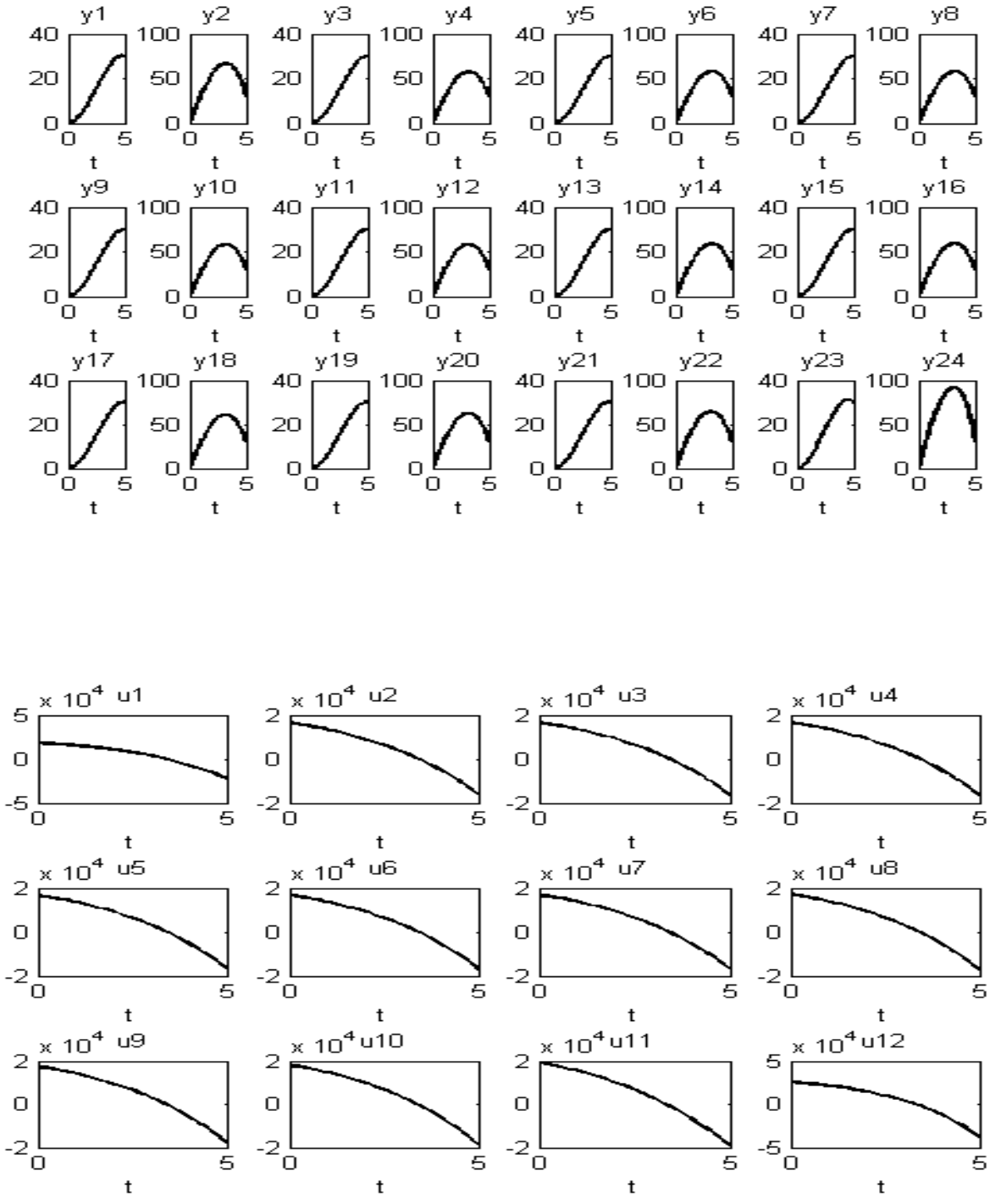

FIG. 6.6 – L'état et le contrôle pour $\beta = 10$ et $\alpha = 1$

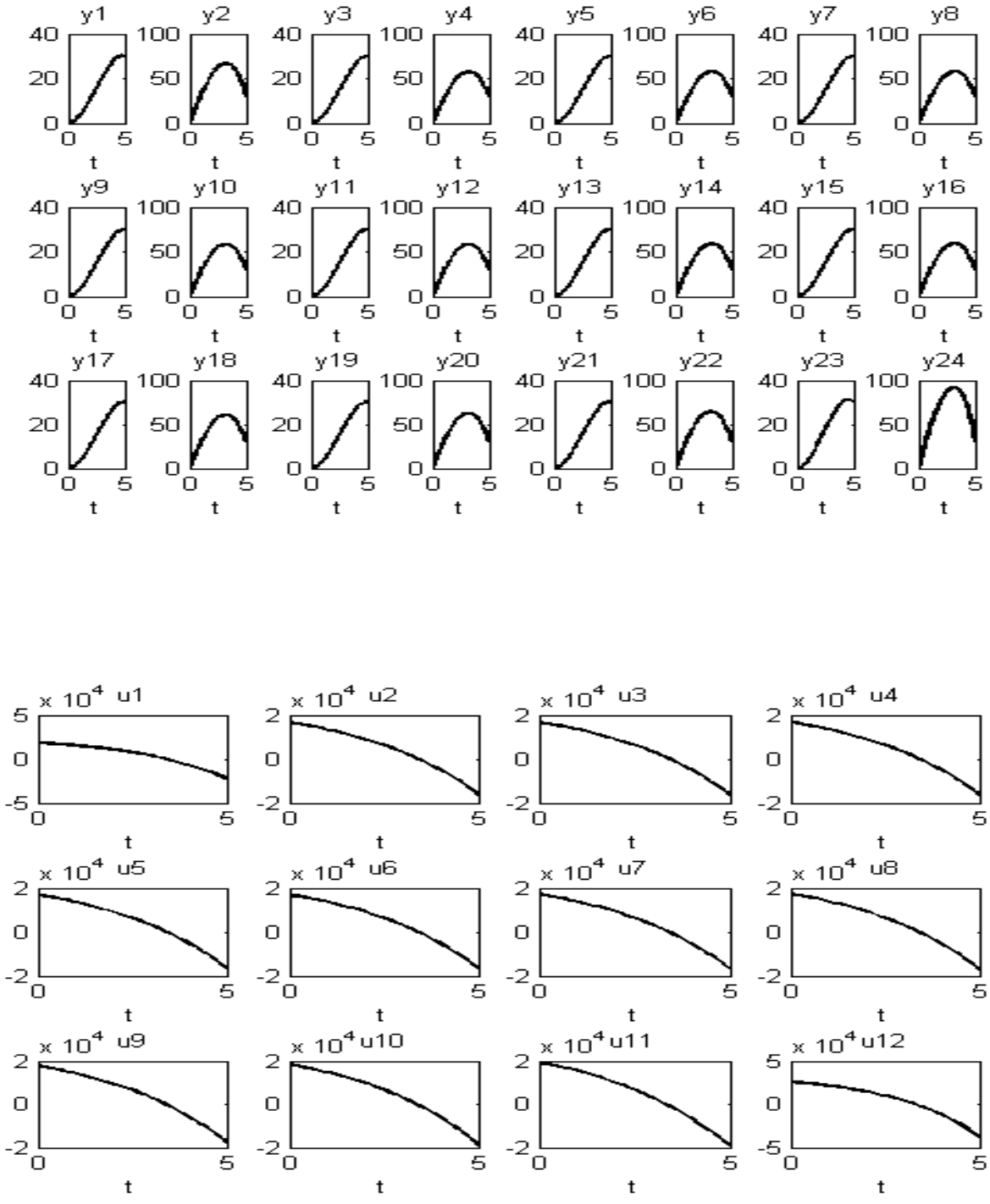

FIG. 6.7 – L'état et le contrôle pour $\beta = 1$ et $\alpha = 10$

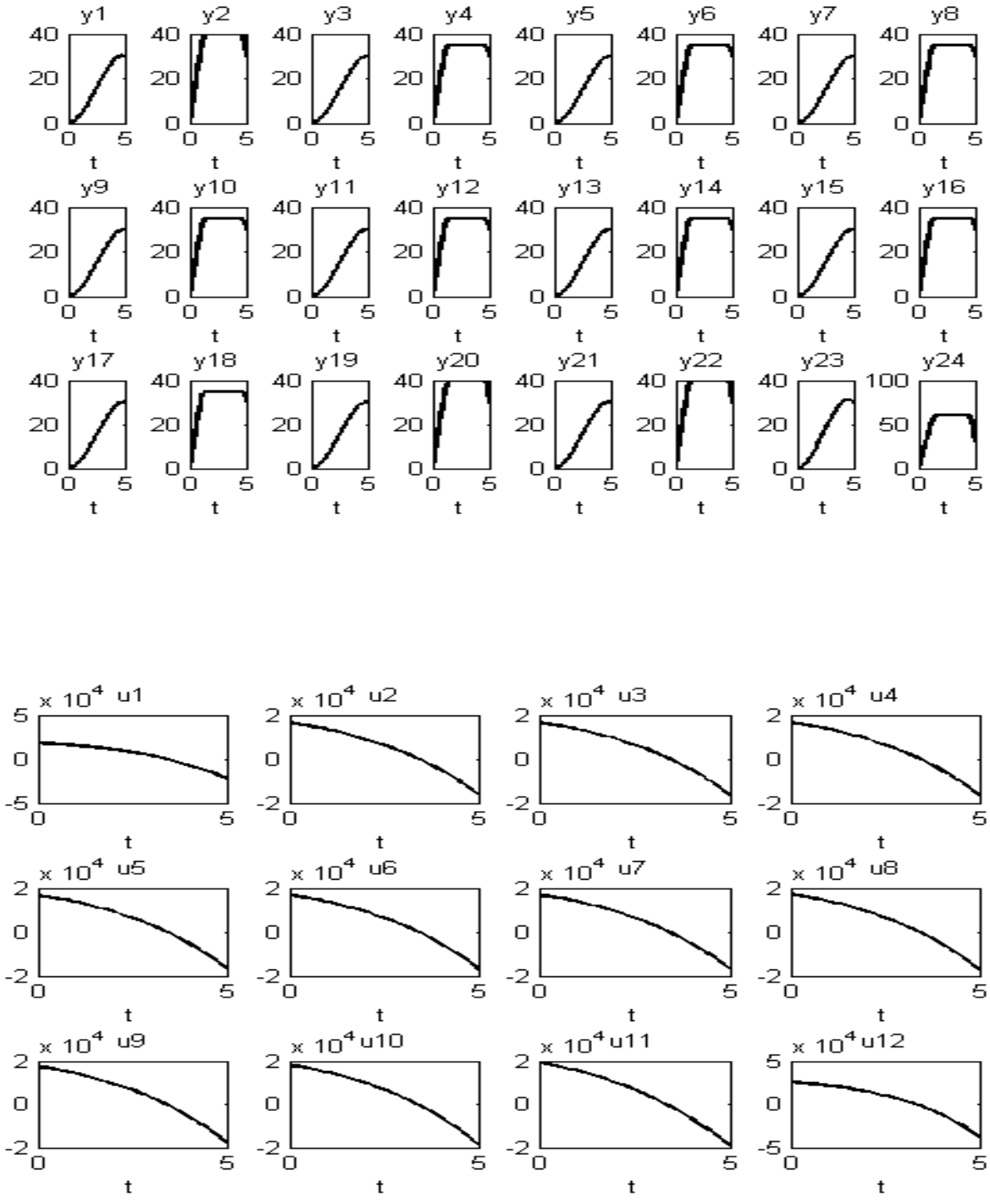

FIG. 6.8 – L'état et le contrôle pour $\beta = 10$ et $\alpha = 1$

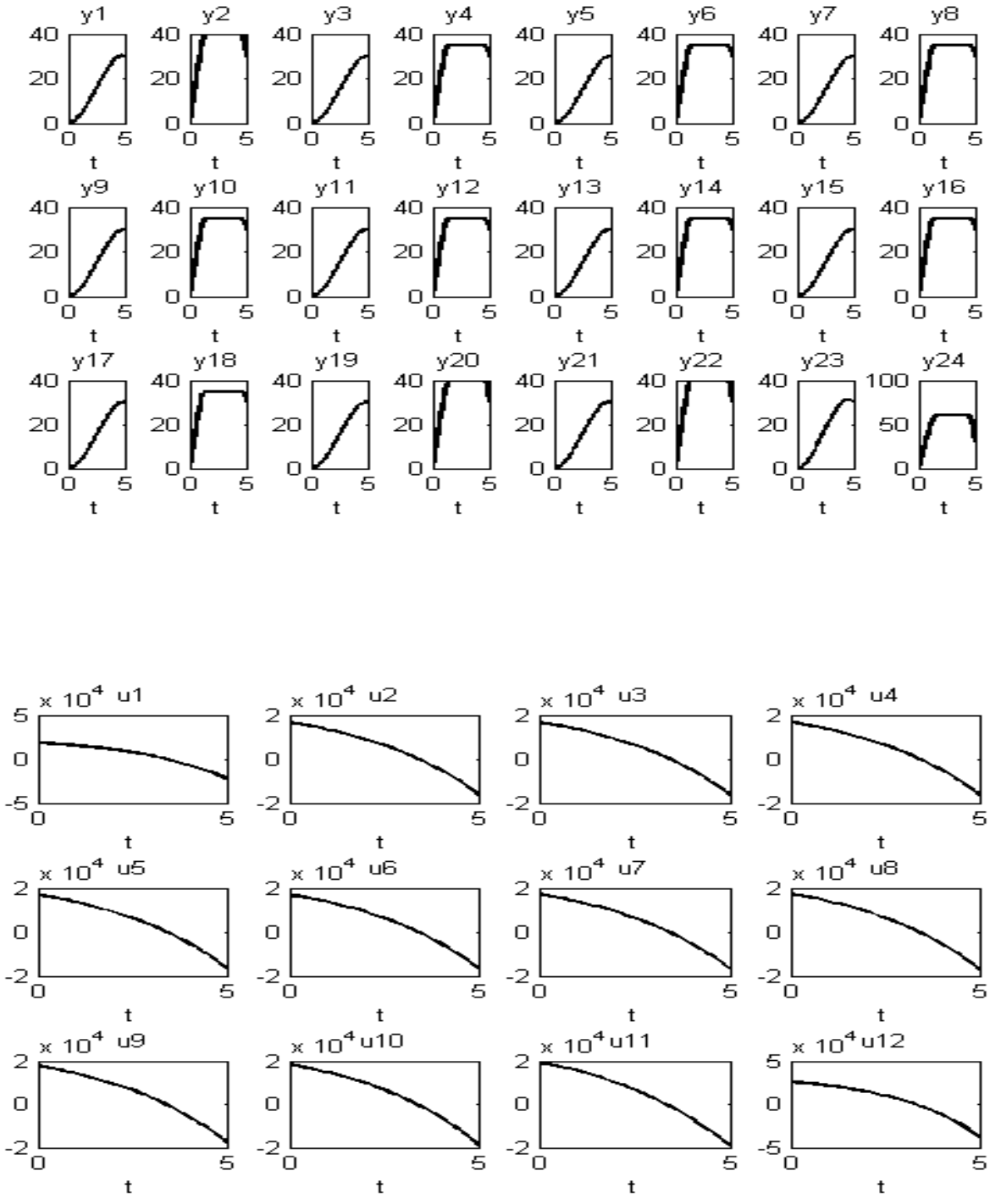

FIG. 6.9 – L'état et le contrôle pour $\beta = 1$ et $\alpha = 10$

Conclusion

Cette étude est consacrée à la résolution d'un problème de contrôle optimal non linéaire lorsque l'état est soumis à certaines contraintes et la valeur de l'état final est fixé. Pour sa résolution, on a utilisé la méthode de relaxation couplée à la méthode de tir ; on a considéré le cas avec et sans contrainte sur l'état. Nous avons testé cette méthode sur des problèmes numériques et il s'avère que la convergence est rapide et que les temps de calculs sont petits.

Ensuite on a appliqué la méthode proposée pour la régulation de deux fours verticaux décomposées respectivement en trois et douze zones de chauffage.

Ce domaine de recherche offre de nombreuses perspectives qu'elles soient théoriques ou pratiques :

- En théorie, il est intéressant d'appliquer cet algorithme à des problèmes de contrôle optimal non linéaire, stochastique, feedback, multicritère, etc.
- En pratique, différents problèmes que ce soit en économie, en agriculture, en automatique, en aéronautique, etc.., peuvent être modélisés par des problèmes de contrôle optimal et résolus par la méthode proposée.

Bibliographie

[1] M. Aidene, I.L. Vorobev, and B. Oukacha. Algorithm for solving a linear optimal control problem with minimax Performance Index. Computational Mathematics and mathematical Physics 45 vol 10, pages 1691–1700, 2005.

[2] N.V. Balashevich, R. Gabasov, and F.M. Kirillova. Numerical methods of program and positional optimization of the linear control systems. Zh. Vychisl. Mat. Mat. Fiz., 40(6) :838-859, 2000.

[3] V. Barbu, Nonlinear Semigroups and Differential Equations in Banach Spaces, Noordhoff International Publishing, Leyden, The Netherlands, 1976.

[4] G.M. Baudet. Asynchronous iterative methods for multiprocesseurs. Journal of ACM, 25 :226-244, 1978.

[5] R.E. Bellman. Dynamic programming. Princeton University Press, Princeton, NJ, 1963.

[6] R. Bellman, I. Glicksberg, and O. Gross. On the bang-bang control problem. Quart. Appl. Math., 14 :11–18, 1956.

[7] R.E. Bellman, I. Glicksberg, and O.A. Gross. Some aspects of the mathematical theory of control processes. Report R-313, Rand Corporation, Santa Monica, CA, 1958.

[8] J.T. Betts. Practical methods for optimal control using nonlinear programming. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[9] M.O. Bibi. Support method for solving a linear-quadratic problem with polyedral constraints on control. Optimization, 37(2) :139-147, 1996.

[10] M.O. Bibi and M. Bentobache. The adaptive method with hybrid direction for solving linear programming problems with bounded variables. Optimization, Prceedings of COSI'2011, University of Guelma, Algeria :80-91, 2011.

[11] M.O. Bibi and M. Bentobache. An hybrid direction algorithm for solving linear programs. Proceedings of DIMACOS'11, University of Mohammadia, Morocco, pages 28-30, 2011.

[12] J.F. Bonnans and A. Hermant. Second-order analysis for optimal control problems with pure state constraints and mixed control-state constraints. INRIA Research Report 6199, to appear in Annales de l'Institut Henri Poincaré (C) Analyse Non Linéaire.

[13] J.F. Bonnans and A. Hermant. Conditions d'optimalité du second ordre nécessaires ou suffisantes pour les problèmes de commande optimale avec une contrainte sur l'état et une commande scalaire. C. R. Math. Acad. Sci. Paris, 343(7) :473-478, 2006.

[14] J.F. Bonnans and A. Hermant. Well-posedness of the shooting algorithm for state constrained optimal control problems with a single constraint and control. SIAM J. on Control and Optimization, 46(4) :1398-1430, 2007.

[15] J.F. Bonnans and A. Hermant. Stability and sensitivity analysis for optimal control problems with a rst-order state constraint. ESAIM : COCV, 2008.

[16] J.F. Bonnans and A. Hermant. No-gap second-order optimality conditions for optimal control problems with a single state constraint and control. Mathematical Programming, 117 :21-50, 2009.

[17] J.F. Bonnans and P. Rouchon. Commande et optimisation de systèmes dynamiques. Editions de l'Ecole Polytechnique, Palaiseau, 2005.

[18] J.F. Bonnans and G. Launay. Large scale direct optimal control applied to a re-entry problem. AIAA J. of Guidance, Control and Dynamics, 21 :996-1000, 1998.

[19] B. Bonnard, L. Faubourg, G. Launay, and E. Trélat. Optimal control with state constraints and the space shuttle re-entry problem. J. Dynam. Control Systems, 9(2) :155-199, 2003.

[20] B. Brahmi and M.O. Bibi. Dual support method for solving convex quadratic programs. Optimisation, 59(6) :851-872, 2010.

[21] H. Brézis. Analyse fonctionnelle. théorie et applications. Masson Paris, 1983.

[22] A.E. Bryson and Yu-Chi Ho. Applied optimal control. Blaisdell, Toronto, Canada, 1969.

[23] A.E. Bryson, W.F. Denham, and S.E. Dreyfus. Optimal programming problems with inequality constraints I : necessary conditions for extremal solutions. AIAA Journal, 1 :2544-2550, 1963.

[24] F.H. Clarke. Optimization and nonsmooth analysis. Wiley, New York, 1983.

[25] M. Crouzeix and A.L. Mignot. Analyse numérique des équations différentielles. 2ème Edition Masson, 1989.

[26] G.B. Dantzig. Maximisation of a linear function of variables subject to linear inequalities. in Koopmans RC (ed.) , Activity Analysis of Production and Allocation, Wiley, New-York, pages 339-347, 1951.

[27] G.B. Dantzig. Linear programming and extensions. Princeton University Press Princeton N.J, 1963.

[28] J.P. Denailly. Analyse numérique et équations différentielles. EDP Sciences, 2006.

[29] A.V. Dmitruk. Quadratic conditions for the Pontryagin minimum in an optimal control problem linear with respect to control. I. Decoding theorem. Izv. Akad. Nauk SSSR Ser. Mat., 50(2) :284-312, 1986.

[30] A.L. Dontchev and W.W. Hager. Lipschitzian stability for state constrained nonlinear optimal control. SIAM J. on Control and Optimization, 36(2) :698-718 (electronic), 1998.

[31] A.L. Dontchev and W.W. Hager. The Euler approximation in state constrained optimal control. Mathematics of Computation, 70 :173-203, 2001.

[32] L.C. Evans. An Introduction To Mathematical Optimal Control Theory. University of California Berkeley, 2000.

[33] H. O. Fattorini. Time-optimal control of solutions of operational differential equations. J. Soc. Indust. Appl. Math. Ser. A Control, 2 :54–59, 1964.

[34] H.O. Fattorini. Infinite dimensional linear control systems, volume 201 of North-Holland Mathematics Studies. Elsevier Science B.V., Amsterdam, 2005. The time optimal and norm optimal problems.

[35] R. Gabasov and F.M. Kirillova. Methods of linear programming. in 3 parts, Edition of University Press, Minsk, 63, (1977, 1978 and 1980).

[36] R. Gabasov and F.M. Kirillova. Constructive methods of optimization part 2. control problems. University Press, Minsk ,(in Russian), 39(4), 1984.

[37] R. Gabasov and F.M. Kirillova. Adaptive method of solving linear programing problems. Preprints Series of University of Karlsruhe, Institute for Satistics and Mathematics, 1994.

[38] R. Gabasov, F.M. Kirillova, and O.I. Kostyukova. Solution of linear quadratic extremal problems. Soviet Mathematics Doklady, 31 : 99-103, 1985.

[39] R. V. Gamkrelidze. Discovery of the maximum principle. J. Dynam. Control Systems, 5(4) :437–451, 1999.

[40] D.M. Gay and B.W. Kernighan. Ampl : A modeling language for mathematical programming. Duxbury Press Second edition, page 540, 2002.

[41] H. P. Geering. Optimal control with engineering applications. Springer-Verlag Berlin Heidelberg 2007.

[42] D. Gien , B. Lang, J.C. Miellou, L. Raffort, P. Spiteri, Commande optimale de systèmes complexes, RAIRO Automatique, Systems Analysis and Control vol. 18, 1984, pp.209-224.

[43] A. Girad. Optimal control of linear system. a multiresolution approch. 43rd IEEE conference on d´ecision and control ; Nassau ; Bahamas, 2004.

[44] H.H. Goldstine. A history of the calculus of variations from the 17th through the 19th century, volume 5 of Studies in the History of Mathematics and Physical Sciences. Springer-Verlag, New York, 1980.

[45] R.F. Hartl, S.P. Sethi, R.G. Vickson, A survey of the maximum principles for optimal control problems with state constraints. Society for industrial and applied mathematics 17 (1995) 181-218.

[46] A. Hermant. Homotopy algorithm for optimal control problems with a second-order state constraint. INRIA Research Report RR-6626, 2008.

[47] D.H. Jacobson, M.M. Lele, and J.L. Speyer. New necessary conditions of optimality for control problems with state-variable inequality contraints. J. of Mathematical Analysis and Applications, 35 :255-284, 1971.

[48] R.E. Kalman. Mathematical description of linear dynamical systems. J. SIAM control, 1 : 152-192, 1963.

[49] E. Kreindler. Additional necessary conditions for optimal control with state-variable inequality constraints. J. Optim. Theory Appl., 38(2) :241-250, 1982.

[50] B. Lang, A.W. El Awtani, P. Spiteri, N.E. Cheik Obeidh. Decentralized calculations in optimal control of a large thermic process : methods and results, Proceedings of the international conferences of Large Scale Systems, Toulouse, Pergamon Press, (1981) 505-516.

[51] B. Lang, J.C. Miellou, P. Spiteri. Asynchronous relaxation algorithms for optimal control problems. Mathematical and Computers in Simulations 28(1986) 227-242.

[52] B. Lang, P. Spiteri. Decomposition and coordination using asynchronous iterations. Encyclopedia of systems and control, M. Singh ed., Pergamon Press, (1987) 3475-3481.

[53] P.J. Laurent. Approximation et Optimisation. Collection Enseignement des Sciences, 1972.

[54] J. Laurent-Varin, F. Bonnans, N. Bérend, M. Haddou, and C. Talbot. Interior-point approach to trajectory optimization. Journal of Guidance, Control, and Dynamics, 30(5) :1228-1238, 2007.

[55] G. Leborgne. Notes du cours d'équations aux dérivées partielles de l'isima première année. page 46, 2008.

[56] A. Ledoux. Sur l'algorithme de tir pour les problèmes de commande optimale avec contraintes sur l'état. Thèse de doctorat, école plytechnique INRIA. 2008

[57] J.L. Lions. Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles. Avant propos de P. Lelong. Dunod, Paris, 1968.

[58] K. Louadj. Résolution des problèmes paramétrés en contrôle optimal. Thèse de doctorat, université Mouloud Mammeri de Tizi-Ouzou, 2012.

[59] K. Malanowski. Sufficient optimality conditions for optimal control subject to state constraints. SIAM J. on Control and Optimization, 35 :205-227, 1997.

[60] K. Malanowski. Stability analysis for nonlinear optimal control problems subject to state constraints. SIAM J. on Optimization, 18(3) :926-945, 2007.

[61] P. Martinon. Résolution numérique de problèmes de contrôle optimal par une méthode homotopique simpliciale. thèse de doctorat, 2005.

[62] H. Mauer, C. Buskens, J.H.R. Kim, and C.Y. Kaya. Optmization methods for the verification of second order sufficient conditions for bang-bang controls. Journal optimal control, Applications and methods, 26 :129-156, 2005.

[63] H. Maurer. On optimal control problems with bounded state variables and control appearing linearly. SIAM J. Control and optimisation 15 (1977).

[64] H. Maurer. On the minimum principle for optimal control problems with state constraints. Schriftenreihe des Rechenzentrum 41, Universität Münster, 1979.

[65] H. Maurer. First and second order sufficient optimality conditions in mathematical programming and optimal control. Math. Programming Stud., (14) :163-177, 1981.

[66] A. Merakeb. Optimisation multicritère en contrôle optimal : Application au véhicule électrique. Thèse de doctorat, l'université Mouloud Mammeri de Tizi-Ouzou, 2011.

[67] J.C. Miellou, P. Spiteri, A parallel asynchronous relaxation algorithm for optimal control problem. Proceeding of the International Conference on Mathematical Analysis and its Appllications, Kowait 1985.

[68] A.A. Milyutin. The maximum principle in the general problem of optimal control. Fizmatlit, Moscow, 2001.

[69] N. Moussouni. Contrôle optimal : optimisation d'une production céréalière. Thèse de doctorat, l'université Mouloud Mammeri de Tizi-Ouzou, 2011. 2012.

[70] J.M. Ortega and W.C. Rheinboldt. Iterative solution of nonlinear equations in several variables. Academic Press, New York, 1970.

[71] B. Oukacha. Résolution de problème de contrôle optimal. Thèse de doctorat, université Mouloud Mammeri de Tizi-Ouzou, 2005.

[72] L. Pontryagin and al. Mathematical theory of optimal processes. Eds Mir Moscou, 1974.

[73] L.S. Pontryagin, V.G. Boltyanski, R.V. Gamkrelidze, and E.F. Mischenko. The mathematical theory of optimal processes. Interscience Publishers New York, 1962.

[74] F.M. Ramos and A. Giovannini. Résolution d'un problème inverse multidimensional de diffusion de la chaleur par la méthode des éléments analytiques et par le principe de l'entropie maximale. Internat.J.Heat Mass Transfer, 38 :101-111, 1995.

[75] G. Rousseau, Q.H. Tran and D. Sinoquet. Scop : a sequential constraint-free optimal control problem algorithm. Chinese Control and Decision Conference (CCDC), Yantai, China, 2008.

[76] S.P. Sethi and G.L. Thompson. Optimal control theory : applications to management science and economics, 2006.

[77] J. Stoer and R. Bulirsch. Introduction to numerical analysis, volume 12 of Texts in Applied Mathematics. Springer-Verlag, New York, third edition, 2002. Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.

[78] E. Trélat. Contrôle optimal. Mathématiques. Concrètes. Vuibert, Paris, 2005. Théorie et applications.

[79] E. Trelat and J.M. Coron. Tout est sous contrôle. Laboratoire de Mathématique, Equipe AN-EDP, Université Paris-Sud, MatAplic 83, 1-15, Juillet 2007.

[80] R. Vinter. Optimal control. Systems and Control : Foundations and Applications. Birkhäuser Boston Inc., Boston, MA, 2000.

[81] O. Von Stryk and R. Bulirsch. Direct and indirect methods for trajectory optimization. Ann. Oper. Res., 37(1-4) :357–373, 1992. Nonlinear methods in economic dynamics and optimal control (Vienna, 1990).

[82] R.Weinstock. Calculus of variations, chapitre 3. Dover publications, Inc., New-York, 1974.

[83] V. Zeidan. The Riccati equation for optimal control problems with mixed state-control constraints : necessity and sufficiency. SIAM J. on Control and Optimization, 32 :1297-1321, 1994.