Contrôle optimal : théorie et applications

Emmanuel Trélat

Université Pierre et Marie Curie (Paris 6)
et Institut Universitaire de France

Laboratoire Jacques-Louis Lions
CNRS, UMR 7598
4 place Jussieu, BC 187
75252 Paris cedex 05, FRANCE

– Première édition: 2005, Vuibert, Collection "Mathématiques Concrètes", 246 pages. ISBN 2 7117 7175 X.
– Seconde édition: 2008, Vuibert, Collection "Mathématiques Concrètes", 250 pages. ISBN-10: 2711722198.
(correction de misprints)
– Présente version électronique: 2013. Ajout de quelques exercices, correction de quelques misprints.

Si vous trouvez des misprints ou des choses incorrectes, merci de m'envoyer un mail: `emmanuel.trelat@upmc.fr`

Table des matières

Notations

$\forall$: pour tout.

$\exists$: il existe.

| ou t.q. : tel que

$A \backslash B$: ensemble A privé de l'ensemble B.

$Conv(A)$: enveloppe convexe de A.

$\bar{A}$: adhérence de A.

$\overset{\circ}{A}$: intérieur de A.

∂A : frontière de A, *i.e.* $\bar{A} \setminus \overset{\circ}{A}$.

max : maximum.

min : minimum.

sup : borne supérieure.

inf : borne inférieure.

lim : limite.

lim sup : limite supérieure.

lim inf : limite inférieure.

$\mathbb{N}$: ensemble des entiers naturels.

$\mathbb{Z}$: ensemble des entiers relatifs.

$\mathbb{Q}$: ensemble des nombres rationnels.

$\mathbb{R}$: ensemble des nombres réels.

$\mathbb{R}^+$: ensemble des nombres réels positifs ou nuls.

$\mathbb{C}$: ensemble des nombres complexes.

Re z : partie réelle du nombre complexe z.

Im z : partie imaginaire du nombre complexe z.

| | : valeur absolue, ou module.

Vect : espace vectoriel engendré par.

$\mathcal{M}_{n,p}(\mathbb{K})$: ensemble des matrices à n lignes et p colonnes, à coefficients dans $\mathbb{K}$.

$\mathcal{M}_n(\mathbb{K})$: ensemble des matrices carrées d'ordre n, à coefficients dans $\mathbb{K}$.

ker l : noyau de l'application linéaire l.

Im l : image de l'application linéaire l.

det : déterminant.

tr : trace.

rg ou rang : rang.

$com(A)$: comatrice de la matrice A.

$\chi_A(X)$: polynôme caractéristique de la matrice A.

$\pi_A(X)$: polynôme minimal de la matrice A.

$\exp(A)$, ou e^A : exponentielle de la matrice A.

A^{T} : transposée de la matrice A.

x^{T} : transposée du vecteur x.

$f^{(n)}$ (où f est une fonction numérique) : n-ème dérivée de la fonction f.

$df(x).h$ (où f est une application d'un Banach E dans un Banach F) : différentielle de Fréchet de f au point x, appliquée au vecteur h.

$\frac{\partial f}{\partial x}(x,y)h$ (où f est une application de $E\times F$ dans G, et E, F, G sont des espaces de Banach) : différentielle de Fréchet de f par rapport à la variable x, au point $(x,y) \in E \times F$, appliquée au vecteur $h \in E$.

∇f (où f est une fonction) : gradient de f.

$C^p(\Omega, \mathbb{K})$: ensemble des applications de Ω dans $\mathbb{K}$, de classe C^p.

$L^p(\Omega, \mathbb{K})$: ensemble des applications mesurables de Ω dans $\mathbb{K}$, de puissance p intégrable.

$L^p_{loc}(\Omega, \mathbb{K})$: ensemble des applications mesurables de Ω dans $\mathbb{K}$, dont la puissance p est intégrable sur tout compact de Ω.

$H^1(\Omega, \mathbb{K})$: ensemble des applications mesurables f de Ω dans $\mathbb{K}$, telles que $f, f' \in L^2(\Omega, \mathbb{K})$.

$\rightharpoonup$: flèche de convergence faible.

$\mathcal{L}$: transformation de Laplace.

$Acc(x_0, T)$: ensemble accessible en temps T depuis le point x_0.

$Acc_\Omega(x_0, T)$: ensemble accessible en temps T depuis le point x_0, pour des contrôles à valeurs dans Ω.

$E_{x_0,T}$, ou E_T (si le point x_0 est sous-entendu) : application entrée-sortie en temps T depuis le point x_0.

$\|x\|_W$ (où $x \in \mathbb{K}^n$ et $W \in \mathcal{M}_n(\mathbb{K})$) : abbréviation pour $x^{\mathrm{T}}Wx$.

T_xM (où M est une variété, et $x \in M$) : espace tangent à M au point x.

T^*_xM : espace cotangent à M au point x.

$[X, Y]$ (où X et Y sont des champs de vecteurs) : crochet de Lie des champs X et Y.

Avant-propos

Qu'est-ce que la théorie du contrôle ? 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 éventuellement certains critères. Les systèmes abordés sont multiples : systèmes différentiels, systèmes discrets, systèmes avec bruit, avec retard... Leurs origines sont très diverses : mécanique, électricité, électronique, biologie, chimie, économie... 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).

Dans les industries modernes où la notion de rendement est prépondérante, le rôle de l'automaticien est de concevoir, de réaliser et d'optimiser, tout au moins d'améliorer les méthodes existantes. Ainsi les domaines d'application sont multiples : aérospatiale, automobile, robotique, aéronautique, internet et les communications en général, mais aussi le secteur médical, chimique, génie des procédés, etc.

Du point de vue mathématique, un système de contrôle est un système dynamique dépendant d'un paramètre dynamique appelé le contrôle. Pour le modéliser, on peut avoir recours à des équations différentielles, intégrales, fonctionnelles, aux différences finies, aux dérivées partielles, stochastiques, etc. Pour cette raison la théorie du contrôle est à l'interconnexion de nombreux domaines mathématiques. Les contrôles sont des fonctions ou des paramètres, habituellement soumis à des contraintes.

Contrôlabilité. Un système de contrôle est dit contrôlable si on peut l'amener (en temps fini) d'un état initial arbitraire vers un état final prescrit. Pour les systèmes de contrôle linéaires en dimension finie, il existe une caractérisation très simple de la contrôlabilité, due à Kalman. Pour les systèmes non linéaires, le problème mathématique de contrôlabilité est beaucoup plus difficile.

Origine du contrôle optimal. 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. En mathématiques, la théorie du contrôle optimal s'inscrit dans la continuité du calcul des variations. Elle est apparue après la seconde guerre mondiale, répondant à des besoins pratiques de guidage, notamment dans le domaine de l'aéronautique et de la dynamique du vol. Historiquement, la théorie du contrôle optimal est très liée à la mécanique classique, en particulier aux principes variationnels de la mécanique (principe de Fermat, de Huygens, équations d'Euler-Lagrange). 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 (voir [31] pour l'histoire de cette découverte). Les points forts de la théorie ont été la découverte de la

méthode de programmation dynamique, l'introduction de l'analyse fonctionnelle dans la théorie des systèmes optimaux, la découverte des liens entre les solutions d'un problème de contrôle optimal et des résultats de la théorie de stabilité de Lyapunov. Plus tard sont apparues les fondations de la théorie du contrôle stochastique et du filtrage de systèmes dynamiques, la théorie des jeux, le contrôle d'équations aux dérivées partielles.

Notons que l'allure des trajectoires optimales dépend fortement du critère d'optimisation. Par exemple pour réaliser un créneau et garer sa voiture, il est bien évident que la trajectoire suivie diffère si on réalise l'opération en temps minimal (ce qui présente un risque) ou bien en minimisant la quantité d'essence dépensée. Le plus court chemin entre deux points n'est donc pas forcément la ligne droite. En 1638, Galilée étudie le problème suivant : déterminer la courbe sur laquelle une bille roule, sans vitesse initiale, d'un point A à un point B, avec un temps de parcours minimal, sous l'action de la pesanteur (toboggan optimal). C'est le fameux problème de la brachistochrone (du grec brakhistos, "le plus court", et chronos, "temps"). Galilée pense (à tort) que la courbe cherchée est l'arc de cercle, mais il a déjà remarqué que la ligne droite n'est pas le plus court chemin en temps. En 1696, Jean Bernoulli pose ce problème comme un défi aux mathématiciens de son époque. Il trouve lui-même la solution, ainsi que son frère Jacques Bernoulli, Newton, Leibniz et le marquis de l'Hospital. La solution est un arc de cycloïde commençant par une tangente verticale. Ce résultat a motivé le développement de la théorie du calcul des variations, devenue, plus tard, la théorie du contrôle optimal (pour plus de détails sur l'histoire du problème de la brachistochrone, voir [68]).

Contrôle optimal moderne et applications. On considère que la théorie moderne du contrôle optimal a commencé dans les années 50, avec la formulation du principe du maximum de Pontryagin, qui généralise les équations d'Euler-Lagrange du calcul des variations. Dès lors, la théorie a connu un essor spectaculaire, ainsi que de nombreuses applications. De nos jours, les systèmes automatisés font complètement partie de notre quotidien (nous en sommes souvent inconscients), ayant pour but d'améliorer notre qualité de vie et de faciliter certaines tâches : système de freinage ABS, assistance à la conduite, servomoteurs, thermostats, régulation hygrométrique, circuits frigorifiques, contrôle des flux routiers, ferroviaires, aériens, boursiers, fluviaux, barrages EDF, photographie numérique, filtrage et reconstruction d'images, lecteurs CD et DVD, réseaux informatiques, moteurs de recherche sur internet, circuits électriques, électroniques, télécommunications en général, contrôle des procédés chimiques, raffinage pétrolier, chaînes industrielles de montage, peacemakers et autres systèmes médicaux automatisés, opérations au laser, robotique, satellites, guidages aérospatiaux, bioréacteurs, distillation, ... La liste est infinie, les applications concernent tout système sur lequel on peut avoir une action, avec une notion de rendement optimal.

Résumé du livre

L'objectif de ce livre est de présenter, du point de vue mathématique, les bases théoriques du contrôle optimal, ainsi que des applications concrètes de cette théorie. Il a été rédigé à partir de notes de cours d'Automatique et de Contrôle Optimal enseignés par l'auteur dans le master d'Ingénierie Mathématique de l'Université d'Orsay, Option Automatique.

Il est accessible à un élève suivant une formation universitaire (licence, master) ou une école d'ingénieurs.

Dans une première partie, on présente la théorie du contrôle optimal pour des systèmes de contrôle linéaires, ainsi que la théorie dite linéaire-quadratique et ses applications : régulation, stabilisation, filtrage de Kalman.

Dans une seconde partie, on présente la théorie du contrôle optimal pour des systèmes de contrôle généraux (non linéaires), avec notamment le principe du maximum de Pontryagin dans toute sa généralité, ainsi que la théorie d'Hamilton-Jacobi. Un chapitre est consacré aux méthodes numériques en contrôle optimal.

Enfin, en appendice on effectue quelques rappels :
- généralisations des théorèmes de Cauchy-Lipschitz pour des équations différentielles ordinaires ;
- bases de l'automatique : fonctions de transfert, stabilisation, observateurs.

Ce livre est résolument orienté vers les applications concrètes de l'automatique et du contrôle optimal, et de nombreux exercices et applications sont présentés. Les applications numériques sont également détaillées ; elles sont effectuées à l'aide de logiciels standards comme *Matlab* et *Maple*, ou bien, si nécessaire, implémentées en *C++*. Parmi les applications détaillées dans cet ouvrage, figurent le contrôle optimal d'un ressort (linéaire ou non linéaire) ; le filtrage de Kalman ; différents problèmes de régulation ; le contrôle optimal et la stabilisation d'une navette spatiale en phase de rentrée atmosphérique ; le transfert orbital d'un satellite à poussée faible ; le contrôle optimal et la stabilisation d'un pendule inversé. Des exercices concernent aussi différents problèmes d'aéronautique, transfert de fichiers informatiques, contrôle d'un réservoir, problème de Bolzano en économie, dynamique des populations (système prédateurs-proies), réactions chimiques, mélangeurs, circuits électriques, contrôle d'épidémies. Ils sont présentés avec des éléments de correction et, si nécessaire, des algorithmes d'implémentation numérique.

Chapitre 1

Introduction : contrôle optimal d'un ressort

Pour expliquer et motiver la théorie nous allons partir d'un problème concret simple : le contrôle optimal d'un ressort. Cet exemple, leitmotiv de cet ouvrage, sera résolu complètement, de manière théorique puis numérique.

Dans une première partie, nous nous placerons dans le cas linéaire : c'est le problème de l'oscillateur harmonique (traité en totalité dans [52]), et nous développerons la théorie du contrôle optimal linéaire.

Dans une deuxième partie nous traiterons le cas de l'oscillateur non linéaire et introduirons des outils généraux de théorie du contrôle optimal. Les applications numériques seront effectuées à l'aide des logiciels *Maple* et *Matlab*.

1.1 Présentation du problème

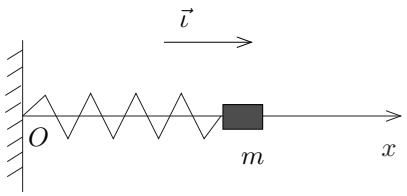


FIGURE 1.1 – Le ressort

Considérons une masse ponctuelle m, astreinte à se déplacer *le long d'un axe* (Ox), attachée à un ressort (voir figure 1.1). La masse ponctuelle est alors attirée vers l'origine par une force que l'on suppose égale à $-k_1(x-l)-k_2(x-l)^3$, où l est la longueur du ressort au repos, et k_1, k_2 sont des coefficients de raideur. On applique à cette masse ponctuelle une force extérieure horizontale $u(t)\vec{\imath}$. Les

lois de la physique nous donnent l'équation du mouvement,

$$m\ddot{x}(t) + k_1(x(t) - l) + k_2(x(t) - l)^3 = u(t). \tag{1.1}$$

De plus on impose une *contrainte* à la force extérieure,

$$|u(t)| \leqslant 1.$$

Cela signifie qu'on ne peut pas appliquer n'importe quelle force extérieure horizontale à la masse ponctuelle : le module de cette force est borné, ce qui traduit le fait que notre puissance d'action est limitée et rend ainsi compte des limitations techniques de l'expérience.
Supposons que la position et la vitesse initiales de l'objet soient $x(0) = x_0$, $\dot{x}(0) = y_0$. Le problème est d'amener la masse ponctuelle à la position d'équilibre $x = l$ en un *temps minimal* en *contrôlant la force externe* $u(t)$ appliquée à cet objet, et en tenant compte de la *contrainte* $|u(t)| \leqslant 1$. La fonction u est appelée le *contrôle*.

Des conditions initiales étant données, le but est donc de trouver une fonction $u(t)$ qui permet d'amener la masse ponctuelle à sa position d'équilibre en un temps minimal.

1.2 Modélisation mathématique

Pour la simplicité de l'exposé, nous supposerons que $m = 1$ kg, $k_1 = 1$ N.m^{-1}, $l = 0$ m (on se ramène à $l = 0$ par translation). Dans la première partie sur le contrôle linéaire, nous supposerons que $k_2 = 0$, et dans la deuxième partie sur le contrôle non linéaire, nous prendrons $k_2 = 2$ (ces valeurs n'étant pas limitatives dans le problème).

Dans *l'espace des phases* $(x, \dot{x})$, le système différentiel correspondant à l'équation du mouvement est

$$\begin{cases} \dot{x}(t) = y(t), \\ \dot{y}(t) = -x(t) - k_2 x(t)^3 + u(t), \end{cases}$$

$$x(0) = x_0,\ \dot{x}(0) = y_0.$$

Posons

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ X = \begin{pmatrix} x \\ y \end{pmatrix},\ X_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix},\ f(X) = \begin{pmatrix} 0 \\ -k_2 x^3 \end{pmatrix}.$$

On obtient

$$\dot{X}(t) = AX(t) + f(X(t)) + Bu(t)\,, \quad X(0) = X_0.$$

On dit qu'il s'agit d'un *système différentiel contrôlé*. C'est un *système linéaire* dans le cas où $k_2 = 0$.

1.3 Quelques remarques sur l'équation

Faisons quelques remarques sur l'équation (1.1) dans le cas non linéaire, avec $k_2 = 2$.

Le ressort libre

Dans ce paragraphe on suppose que $u(t) = 0$, c'est-à-dire qu'aucune force n'est appliquée au ressort. L'équation (1.1) se réduit alors à

$$\ddot{x}(t) + x(t) + 2x(t)^3 = 0,$$

qui s'appelle l'*équation de Duffing*. Il est très facile de vérifier que toute solution $x(\cdot)$ de cette équation est telle que

$$x(t)^2 + x(t)^4 + \dot{x}(t)^2 = \text{Cste}.$$

Autrement dit, dans le plan de phase, toute solution est périodique, et son image est incluse dans une courbe algébrique. Ci-dessous nous utilisons *Maple* pour tracer dans le plan de phase $(x, \dot{x})$ plusieurs trajectoires solutions et le champ de vecteurs associé, ainsi que les courbes $x(t)$ en fonction de t.

Les commandes suivantes donnent la figure 1.2.

```
> with(DEtools):
> eq1 := D(x)(t)=y(t) :
  eq2 := D(y)(t)=-x(t)-2*x(t)^3 :
  sys := eq1,eq2 :
  ic := [x(0)=1,y(0)=0],[x(0)=2,y(0)=0] :
> DEplot([sys],[x(t),y(t)],t=0..6,[ic],stepsize=0.05,
                scene=[x(t),y(t)],linecolor=[blue,red]);
> DEplot([sys],[x(t),y(t)],t=0..6,[ic],stepsize=0.05,
                   scene=[t,x(t)],linecolor=[blue,red]);
```

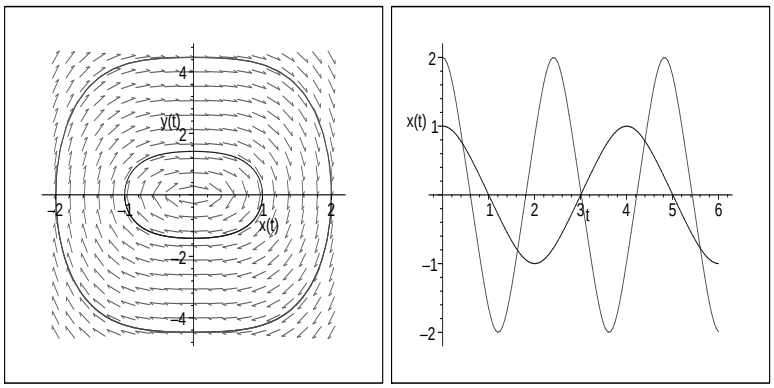


FIGURE 1.2 –

Le ressort amorti

Dans ce paragraphe on suppose que $u(t) = -\dot{x}(t)$. L'équation (1.1) devient

$$\ddot{x}(t) + x(t) + 2x(t)^3 + \dot{x}(t) = 0.$$

A l'aide de Maple, traçons dans le plan de phase $(x, \dot{x})$ plusieurs trajectoires solutions et le champ de vecteurs associé.

```
> eq1 := D(x)(t)=y(t) :
  eq2 := D(y)(t)=-x(t)-2*x(t)^3-y(t) :
  sys := eq1,eq2 :
  ic := [x(0)=1,y(0)=0],[x(0)=2,y(0)=0] :
  DEplot([sys],[x(t),y(t)],t=0..15,[ic],stepsize=0.05,
                scene=[x(t),y(t)],linecolor=[blue,red]);
  DEplot([sys],[x(t),y(t)],t=0..15,[ic],stepsize=0.05,
                  scene=[t,x(t)],linecolor=[blue,red]);
```

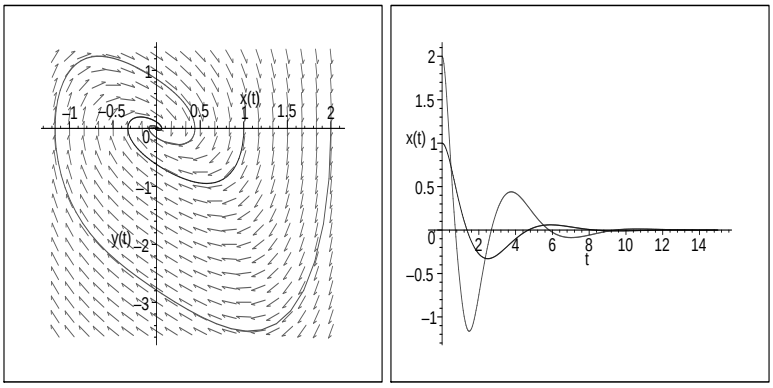


FIGURE 1.3 –

On observe un *amortissement* : les solutions tendent vers l'origine (voir figure 1.3). En fait il est aisé, à l'aide de la théorie de Lyapunov, de montrer que l'origine est globalement asymptotiquement stable. Notons cependant que ce contrôle $u(t)$ ne résout pas notre problème, car le temps pour amener le ressort à sa position d'équilibre est infini !

Le ressort entretenu

Dans ce paragraphe on suppose que

$$u(t) = -(x(t)^2 - 1)\dot{x}(t).$$

L'équation (1.1) devient

$$\ddot{x}(t) + x + 2x(t)^3 + (x(t)^2 - 1)\dot{x}(t) = 0.$$

C'est une équation dite de *Van der Pol*.

A l'aide de Maple, traçons dans le plan de phase $(x, \dot{x})$ plusieurs trajectoires solutions et le champ de vecteurs associé, ainsi que les courbes $x(t)$ en fonction de t.

```
> eq1 := D(x)(t)=y(t) :
  eq2 := D(y)(t)=-x(t)-2*x(t)^3-(x(t)^2-1)*y(t) :
  sys := eq1,eq2 :
  ic := [x(0)=1,y(0)=0], [x(0)=4,y(0)=0] :
  DEplot([sys],[x(t),y(t)],t=0..10,[ic],stepsize=0.05,
              scene=[x(t),y(t)],linecolor=[blue,red]);
  DEplot([sys],[x(t),y(t)],t=0..10,[ic],stepsize=0.05,
                 scene=[t,x(t)],linecolor=[blue,red]);
```

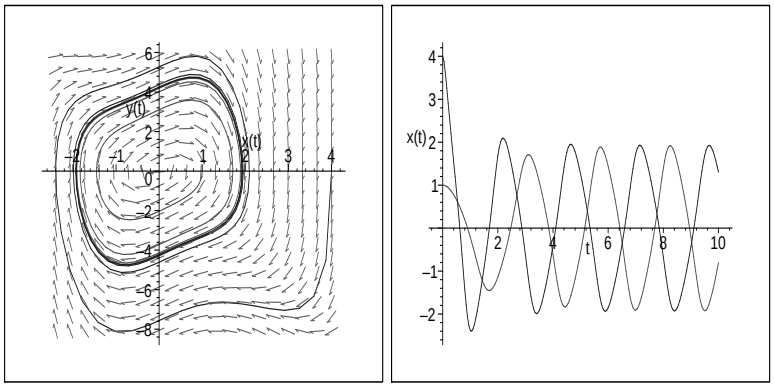


FIGURE 1.4 –

Numériquement (voir figure 1.4) on constate l'existence d'une solution périodique qui semble "attirer" toutes les autres solutions. En fait on peut montrer rigoureusement, toujours à l'aide de la théorie de Lyapunov, que cette solution périodique existe et est *attractive*.

Qualitativement on peut comprendre le comportement d'un tel oscillateur en discutant le signe de $x(t)^2 - 1$. En effet si $x(t)$ est grand il y a un amortissement et le rayon polaire des solutions dans le plan de phase a tendance à décroître. Au contraire si $x(t)$ est petit alors le terme $(x(t)^2 - 1)\dot{x}(t)$ apporte de l'énergie et le rayon a tendance à augmenter. On retrouve bien ce comportement sur la figure.

L'équation de Van der Pol est en fait le modèle d'une *horloge*.

Première partie

Contrôle optimal de systèmes linéaires

Le problème général étudié dans cette partie est le suivant. Soient n et m deux entiers naturels non nuls, I un intervalle de $\mathbb{R}$, et soient A, B et r trois applications L^∞ sur I (en fait, localement intégrables, L^1_{loc}, suffit), à valeurs respectivement dans $\mathcal{M}_n(\mathbb{R})$, $\mathcal{M}_{n,m}(\mathbb{R})$, et $\mathcal{M}_{n,1}(\mathbb{R})$ (identifié à $\mathbb{R}^n$). Soit Ω un sous-ensemble de $\mathbb{R}^m$, et soit $x_0 \in \mathbb{R}^n$. Le système de contrôle linéaire auquel on s'intéresse est

$$\begin{aligned} \forall t \in I \quad & \dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t), \\ & x(0) = x_0, \end{aligned} \tag{1.2}$$

où l'ensemble des contrôles u considérés est l'ensemble des applications mesurables et localement bornées sur I, à valeurs dans le sous-ensemble $\Omega \subset \mathbb{R}^m$.

Les théorèmes d'existence de solutions d'équations différentielles (cf section 11.3) nous assurent que, pour tout contrôle u, le système (1.2) admet une unique solution $x(\cdot) : I \to \mathbb{R}^n$, absolument continue. Soit $M(\cdot) : I \to \mathcal{M}_n(\mathbb{R})$ la résolvante du système linéaire homogène $\dot{x}(t) = A(t)x(t)$, définie par $\dot{M}(t) = A(t)M(t)$, $M(0) = Id$. Notons que si $A(t) = A$ est constante sur I, alors $M(t) = e^{tA}$. Alors, la solution $x(\cdot)$ du système (1.2) 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 $t \in I$.

Cette application dépend de u. Donc si on change la fonction u, on obtient une autre trajectoire $t \mapsto x(t)$ dans $\mathbb{R}^n$ (voir figure 1.5).

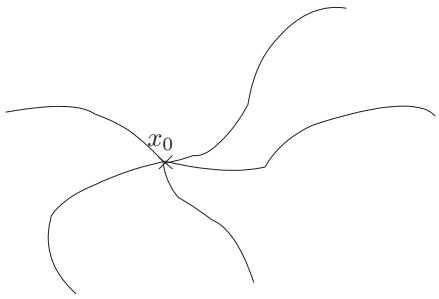


FIGURE 1.5 –

Deux questions se posent alors naturellement :

– Etant donné un point $x_1 \in \mathbb{R}^n$, existe-t-il un contrôle u tel que la trajectoire associée à ce contrôle joigne x_0 à x_1 en un temps fini T ? (voir figure 1.6)
C'est le problème de contrôlabilité.

– Si la condition précédente est remplie, existe-t-il un contrôle joignant x_0 à x_1, et qui de plus minimise une certaine fonctionnelle $C(u)$? (voir figure 1.7)

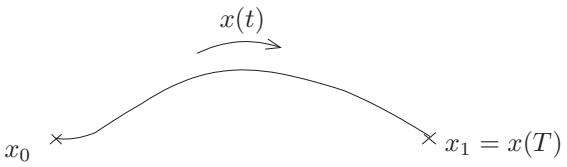


FIGURE 1.6 – Problème de contrôlabilité

C'est le problème de contrôle optimal.
La fonctionnelle $C(u)$ est un critère d'optimisation, on l'appelle le *coût*. Par exemple ce coût peut être égal au temps de parcours ; dans ce cas c'est le problème du *temps minimal*.

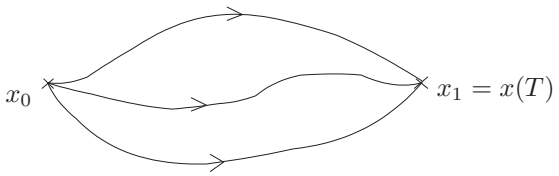


FIGURE 1.7 – Problème de contrôle optimal

Les théorèmes suivants vont répondre à ces questions, et permettre en particulier de résoudre le problème de l'oscillateur harmonique linéaire ($k_2 = 0$) vu en introduction.

Chapitre 2

Contrôlabilité

2.1 Ensemble accessible

2.1.1 Définition

Considérons le système contrôlé (1.2),

$$\begin{aligned} \forall t \in I \quad & \dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t), \\ & x(0) = x_0, \end{aligned}$$

Définition 2.1.1. L'ensemble des points accessibles à partir de x_0 en un temps $T > 0$ est défini par

$$Acc(x_0, T) = \{x_u(T) \mid u \in L^\infty([0, T], \Omega)\},$$

où $x_u(\cdot)$ est la solution du système (1.2) associée au contrôle u.

Autrement dit $Acc(x_0, T)$ est l'ensemble des extrémités des solutions de (1.2) au temps T, lorsqu'on fait varier le contrôle u (voir figure 2.1). Pour la cohérence on pose $Acc(x_0, 0) = \{x_0\}$.

2.1.2 Topologie des ensembles accessibles

Théorème 2.1.1. *Considérons le système de contrôle linéaire dans $\mathbb{R}^n$*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$$

où $\Omega \subset \mathbb{R}^m$ est compact. Soient $T > 0$ et $x_0 \in \mathbb{R}^n$. Alors pour tout $t \in [0, T]$, $Acc(x_0, t)$ est compact, convexe, et varie continûment avec t sur $[0, T]$.

Remarque 2.1.1. La convexité de $Acc(x_0, t)$ est facile à établir si Ω est convexe. En effet, dans ce cas, soient $x_1^1, x_2^1 \in Acc(x_0, t)$, et $\lambda \in [0, 1]$. On veut montrer que $\lambda x_1^1 + (1 - \lambda)x_2^1 \in Acc(x_0, t)$. Par définition, pour $i = 1, 2$, il existe un contrôle $u_i : [0, t] \to \Omega$ tel que la trajectoire $x_i(\cdot)$ associée à u_i vérifie

$$x_i(0) = x_0, \ x_i(t) = x_i^1, \ x_i'(s) = A(s)x_i(s) + B(s)u(s) + r(s).$$

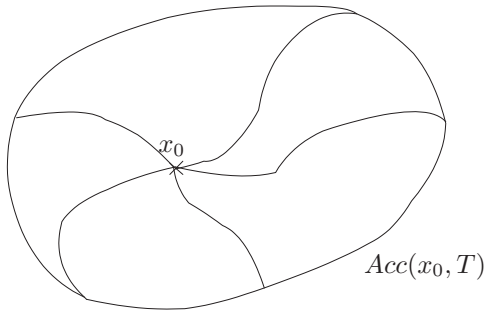


FIGURE 2.1 – Ensemble accessible

D'après la formule de variation de la constante,

$$x_i^1 = x_i(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u_i(s) + r(s))ds.$$

Pour tout $s \in [0,t]$, posons $u(s) = \lambda u_1(s) + (1-\lambda)u_2(s)$. Le contrôle u est dans L^2, à valeurs dans Ω car Ω est convexe. Soit $x(\cdot)$ la trajectoire associée à u. Alors, par définition de $A(x_0,t)$, on a

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

Or,

$$\begin{aligned}\lambda x_1^1 + (1-\lambda)x_2^1 &= M(t)x_0 + (1-\lambda)M(t)x_0 \\ &\quad + \int_0^t M(t)M(s)^{-1}(B(s)(\lambda u_1(s) + (1-\lambda)u_2(s)) + \lambda r(s) + (1-\lambda)r(s))ds \\ &= x(t)\end{aligned}$$

donc $\lambda x_1^1 + (1-\lambda)x_2^1 \in Acc(x_0,t)$, ce qui prouve la convexité de $Acc(x_0,t)$.

Pourtant, et ce résultat est surprenant, la conclusion de ce théorème est encore vraie si Ω n'est pas convexe. Ceci implique en particulier le résultat suivant.

Corollaire 2.1.2. *Supposons que Ω soit compact. Si on note $Acc_\Omega(x_0,t)$ l'ensemble accessible depuis x_0 en temps t pour des contrôles à valeurs dans Ω, alors on a*

$$Acc_\Omega(x_0,t) = Acc_{Conv(\Omega)}(x_0,t),$$

où $Conv(\Omega)$ est l'enveloppe convexe de Ω. En particulier, on a $Acc_{\partial\Omega}(x_0,t) = Acc_\Omega(x_0,t)$, où $\partial\Omega$ est la frontière de Ω.

Ce dernier résultat illustre le *principe bang-bang* (voir théorème 3.2.1).

Démonstration du théorème 2.1.1. Démontrons d'abord ce théorème dans le cas où Ω est compact et convexe. La convexité de $Acc(x_0, t)$ résulte alors de la remarque 2.1.1. Montrons maintenant la compacité de $Acc(x_0, t)$. Cela revient à montrer que toute suite $(x_n^1)_{n\in\mathbb{N}}$ de points de $Acc(x_0, t)$ admet une sous-suite convergente. Pour tout entier n, soit u_n un contrôle reliant x_0 à x_n^1 en temps t, et soit $x_n(\cdot)$ la trajectoire correspondante. On a donc

$$x_n^1 = x_n(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u_n(s) + r(s))ds. \tag{2.1}$$

Par définition, les contrôles u_n sont à valeurs dans le compact Ω, et par conséquent la suite $(u_n)_{n\in\mathbb{N}}$ est bornée dans $L^2([0, t], \mathbb{R}^m)$. Par réflexivité de cet espace (voir [19]), on en déduit que, à sous-suite près, la suite $(u_n)_{n\in\mathbb{N}}$ converge faiblement vers un contrôle $u \in L^2([0, t], \mathbb{R}^m)$. Comme Ω est supposé convexe, on a de plus $u \in L^2([0, t], \Omega)$. Par ailleurs, de la formule de représentation (2.1) on déduit aisément que la suite $(x_n(\cdot))_{n\in\mathbb{N}}$ est bornée dans $L^2([0, t], \mathbb{R}^n)$. De plus, de l'égalité $\dot{x}_n = Ax_n + Bu_n + r$, et utilisant le fait que A, B et r sont dans L^∞ sur $[0, T]$, on conclut que la suite $(\dot{x}_n(\cdot))_{n\in\mathbb{N}}$ est également bornée dans $L^2([0, t], \mathbb{R}^n)$, autrement dit que cette suite est bornée dans $H^1([0, t], \mathbb{R}^n)$. Mais comme cet espace de Sobolev est réflexif et se plonge de manière compacte dans $C^0([0, t], \mathbb{R}^n)$ muni de la topologie uniforme, on conclut que, à sous-suite près, la suite $(x_n(\cdot))_{n\in\mathbb{N}}$ converge uniformément vers une application $x(\cdot)$ sur $[0, t]$. En passant à la limite dans (2.1) il vient alors

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

et en particulier

$$\lim_{i\to+\infty} x_{n_i}^1 = \lim_{i\to+\infty} x_{n_i}(t) = x(t) \ \in Acc(x_0, t),$$

ce qui prouve la compacité.

Montrons enfin la continuité par rapport à t de $Acc(x_0, t)$. Soit $\varepsilon > 0$. On va chercher $\delta > 0$ tel que

$$\forall t_1, t_2 \in [0, T] \quad |t_1 - t_2| \leqslant \delta \Rightarrow d(Acc(t_1), Acc(t_2)) \leqslant \varepsilon,$$

où on note pour simplifier $Acc(t) = Acc(x_0, t)$, et où

$$d(Acc(t_1), Acc(t_2)) = \sup\left(\sup_{y\in Acc(t_2)} d(y, Acc(t_1)), \sup_{y\in Acc(t_1)} d(y, Acc(t_2))\right).$$

Par la suite, on suppose $0 \leqslant t_1 < t_2 \leqslant T$. Il suffit de montrer que

1. $\forall y \in Acc(t_2) \quad d(y, Acc(t_1)) \leqslant \varepsilon$,
2. $\forall y \in Acc(t_1) \quad d(y, Acc(t_2)) \leqslant \varepsilon$.

Montrons juste le premier point (2. étant similaire). Soit $y \in Acc(t_2)$. Il suffit de montrer que

$$\exists z \in Acc(t_1) \mid d(y,z) \leqslant \varepsilon.$$

Par définition de $Acc(t_2)$, il existe un contrôle $u \in L^2([0,T],\Omega)$ tel que la trajectoire associée à u, partant de x_0, vérifie $x(t_2) = y$ (voir figure 2.2). On va voir

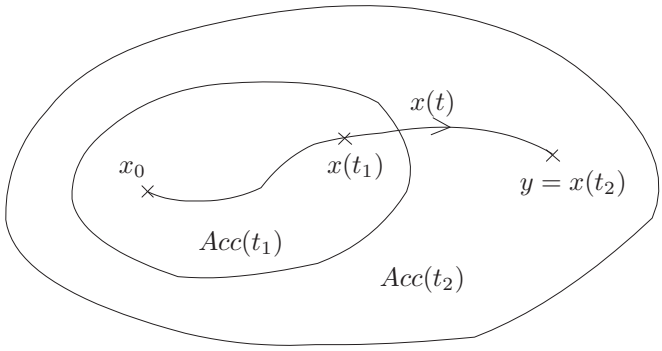


FIGURE 2.2 –

que $z = x(t_1)$ convient. En effet on a

$$\begin{aligned} x(t_2) - x(t_1) &= M(t_2)x_0 + \int_0^{t_2} M(t_2)M(s)^{-1}(B(s)u(s) + r(s))ds \\ &\quad - \left(M(t_1)x_0 + \int_0^{t_1} M(t_1)M(s)^{-1}(B(s)u(s) + r(s))ds \right) \\ &= M(t_2) \int_{t_1}^{t_2} M(s)^{-1}(B(s)u(s) + r(s))ds \\ &\quad + (M(t_2) - M(t_1)) \left(x_0 + \int_0^{t_1} M(s)^{-1}(B(s)u(s) + r(s))ds \right) \end{aligned}$$

Si $|t_1 - t_2|$ est petit, le premier terme de cette somme est petit par continuité de l'intégrale ; le deuxième terme est petit par continuité de $t \longmapsto M(t)$. D'où le résultat.

Dans le cas général où Ω est seulement compact (mais pas forcément convexe), la preuve est plus difficile est fait appel au lemme de Lyapunov en théorie de la mesure (démontré par exemple dans [52, Lemma 4A p. 163]) et plus généralement au théorème d'Aumann (voir par exemple [37]), grâce auquel on a les

égalités

$$\begin{aligned}&\{\int_0^T M(t)^{-1}B(t)u(t)dt \mid u \in L^\infty([0,T],\Omega)\}\\ =&\{\int_0^T M(t)^{-1}B(t)u(t)dt \mid u \in L^\infty([0,T],\partial\Omega)\}\\ =&\{\int_0^T M(t)^{-1}B(t)u(t)dt \mid u \in L^\infty([0,T],\mathrm{Conv}(\Omega))\},\end{aligned}$$

et de plus ces ensembles sont compacts convexes. La preuve du théorème et du corollaire s'ensuivent. Notons que la preuve du lemme de Lyapunov et du théorème d'Aumann évoqués ici reposent sur le théorème de Krein-Milman en dimension infinie (du moins, sur le fait que tout compact convexe d'un espace localement convexe admet au moins un point extrémal, voir [37] pour des précisions). □

Remarque 2.1.2. Si $r = 0$ et $x_0 = 0$, la solution de $\dot{x}(t) = A(t)x(t)+B(t)u(t), x(0) = 0$, s'écrit

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

et est linéaire en u.

Cette remarque nous mène à la proposition suivante.

Proposition 2.1.3. *On suppose que $r = 0$, $x_0 = 0$ et $\Omega = \mathbb{R}^m$. Alors, pour tout $t > 0$, l'ensemble $Acc(0,t)$ est un sous-espace vectoriel de $\mathbb{R}^n$. Si on suppose de plus que $B(t) \equiv B$ est constante, alors, pour tous $0 < t_1 < t_2$, on a $Acc(0,t_1) \subset Acc(0,t_2)$.*

Démonstration. Soient $x_1^1, x_2^1 \in Acc(0,T)$, et $\lambda, \mu \in \mathbb{R}$. Pour $i = 1, 2$, il existe par définition un contrôle u_i et une trajectoire associée $x_i(\cdot)$ vérifiant $x_i(t) = x_i^1$. D'où

$$x_i^1 = M(t)\int_0^t M(s)^{-1}B(s)u_i(s)ds.$$

Pour tout $s \in [0,T]$, posons $u(s) = \lambda u_1(s) + \mu u_2(s)$. Alors

$$\lambda x_1^1 + \mu x_2^1 = M(t)\int_0^t M(s)^{-1}B(s)u(s)ds = x(t) \ \in Acc(0,t).$$

Pour la deuxième partie de la proposition, soit $x_1^1 \in Acc(0,t_1)$. Par définition, il existe un contrôle u_1 sur $[0,t_1]$ tel que la trajectoire associée $x_1(\cdot)$ vérifie $x_1(t_1) = x_1^1$. D'où

$$x_1^1 = M(t_1)\int_0^{t_1} M(s)^{-1}Bu_1(s)ds.$$

Définissons u_2 sur $[0,t_2]$ par

$$\left\{\begin{array}{lll} u_2(t) = 0 & \text{si} & 0 \leqslant \ t \ \leqslant t_2 - t_1 \\ u_2(t) = u_1(t_1 - t_2 + t) & \text{si} & t_2 - t_1 \leqslant \ t \ \leqslant t_2 \end{array}\right. .$$

Soit $x_2(\cdot)$ la trajectoire associée à u_2 sur $[0, t_2]$. Alors

$$\begin{aligned} x_2(t_2) &= M(t_2) \int_0^{t_2} M(t)^{-1} B u_2(t) dt \\ &= M(t_2) \int_{t_2 - t_1}^{t_2} M(t)^{-1} B u_2(t) dt \quad \text{car } u_{2|[0, t_2 - t_1]} = 0 \\ &= M(t_2) \int_0^{t_1} M(t_2)^{-1} M(t_1) M(s)^{-1} B u_2(t_2 - t_1 + s) ds \quad \text{si } s = t_1 - t_2 + t \\ &= M(t_1) \int_0^{t_1} M(s)^{-1} B u_1(s) ds \\ &= x_1^1 \end{aligned}$$

Ainsi, $x_1^1 \in Acc(0, t_2)$. □

Remarque 2.1.3. Dans le cadre de la deuxième partie de la proposition, $Acc(0) = \bigcup_{t \geqslant 0} Acc(0, t)$, l'ensemble des points accessibles (en temps quelconque), est un sous-espace vectoriel de $\mathbb{R}^n$. En effet, une union croissante de sous-espaces vectoriels de $\mathbb{R}^n$ est un sous-espace vectoriel.

2.1.3 Définition de la contrôlabilité

Définition 2.1.2. Le système contrôlé $\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$ est dit *contrôlable en temps* T si $Acc(x_0, T) = \mathbb{R}^n$, *i.e.* , pour tous $x_0, x_1 \in \mathbb{R}^n$, il existe un contrôle u tel que la trajectoire associée relie x_0 à x_1 en temps T (voir figure 2.3).

FIGURE 2.3 – Contrôlabilité

2.2 Contrôlabilité des systèmes linéaires autonomes

2.2.1 Cas sans contrainte sur le contrôle : condition de Kalman

Le théorème suivant nous donne une condition nécessaire et suffisante de contrôlabilité dans le cas où A et B ne dépendent pas de t.

Théorème 2.2.1. *On suppose que $\Omega = \mathbb{R}^m$ (pas de contrainte sur le contrôle). Le système $\dot{x}(t) = Ax(t) + Bu(t) + r(t)$ est contrôlable en temps T (quelconque) si et seulement si la matrice*

$$C = \big(B, AB, \ldots, A^{n-1}B\big)$$

est de rang n.

La matrice C est appelée *matrice de Kalman*, et la condition rg $C = n$ est appelée *condition de Kalman*.

Remarque 2.2.1. La condition de Kalman ne dépend ni de T ni de x_0. Autrement dit, si un système linéaire autonome est contrôlable en temps T depuis x_0, alors il est contrôlable en tout temps depuis tout point.

Démonstration. L'essentiel de la preuve est contenu dans le lemme suivant.

Lemme 2.2.2. *La matrice C est de rang n si et seulement si l'application linéaire*

$$\begin{array}{rccc} \Phi: & L^\infty([0,T],\mathbb{R}^m) & \to & \mathbb{R}^n \\ & u & \mapsto & \int_0^T e^{(T-t)A}Bu(t)dt \end{array}$$

est surjective.

Preuve du lemme. Supposons tout d'abord que rg $C < n$, et montrons qu'alors Φ n'est pas surjective. L'application C étant non surjective, il existe un vecteur $\psi \in \mathbb{R}^n\backslash\{0\}$, que l'on supposera être un vecteur ligne, tel que $\psi C = 0$. Par conséquent,

$$\psi B = \psi AB = \ldots = \psi A^{n-1}B = 0.$$

Or d'après le théorème d'Hamilton-Cayley, il existe des réels $a_0, a_1, \ldots, a_{n-1}$ tels que

$$A^n = a_0 I + \cdots a_{n-1}A^{n-1}.$$

On en déduit par récurrence immédiate que, pour tout entier k,

$$\psi A^k B = 0,$$

et donc, pour tout $t \in [0,T]$,

$$\psi e^{tA}B = 0.$$

Par conséquent, pour tout contrôle u, on a

$$\psi \int_0^T e^{(T-t)A}Bu(t)dt = 0,$$

i.e. $\psi\Phi(u) = 0$, ce qui montre que Φ n'est pas surjective.

Réciproquement, si Φ n'est pas surjective, alors il existe un vecteur ligne $\psi \in \mathbb{R}^n\backslash\{0\}$ tel que pour tout contrôle u on ait

$$\psi \int_0^T e^{(T-t)A}Bu(t)dt = 0.$$

Ceci implique que, pour tout $t \in [0,T]$,

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

En $t = T$ on obtient $\psi B = 0$. Ensuite, en dérivant par rapport à t, puis en prenant $t = T$, on obtient $\psi AB = 0$. Ainsi, par dérivations successives, on obtient finalement

$$\psi B = \psi AB = \cdots = \psi A^{n-1}B = 0,$$

donc $\psi C = 0$, et donc rg $C < n$. □

Ce lemme permet maintenant de montrer facilement le théorème.

Si la matrice C est de rang n, alors d'après le lemme l'application Φ est surjective, *i.e.* $\Phi(L^\infty) = \mathbb{R}^n$. Or, pour tout contrôle u, l'extrémité au temps T de la trajectoire associée à u est donnée par

$$x(T) = e^{TA}x_0 + \int_0^T e^{(T-t)A}(Bu(t) + r(t))dt,$$

de sorte que l'ensemble accessible en temps T depuis un point $x_0 \in \mathbb{R}^n$ est

$$Acc(T, x_0) = e^{TA}x_0 + \int_0^T e^{(T-t)A}r(t)dt + \phi(L^\infty) = \mathbb{R}^n,$$

ce qui montre que le système est contrôlable.

Réciproquement si le système est contrôlable, alors il est en particulier contrôlable depuis x_0 défini par

$$x_0 = -e^{-TA}\int_0^T e^{(T-t)A}r(t)dt.$$

Or en ce point l'ensemble accessible en temps T s'écrit

$$Acc(T, x_0) = \phi(L^\infty),$$

et le système étant contrôlable cet ensemble est égal à $\mathbb{R}^n$. Cela prouve que Φ est surjective, et donc, d'après le lemme, que la matrice C est de rang n. □

Remarque 2.2.2. Si $x_0 = 0$ et si $r = 0$, la démonstration précédente est un peu simplifiée puisque dans ce cas, d'après la remarque 2.1.3, $Acc(0)$ est un sous-espace vectoriel.

2.2.2 Cas avec contrainte sur le contrôle

Dans le théorème 2.2.1, on n'a pas mis de contrainte sur le contrôle. Cependant en adaptant la preuve on obtient aisément le résultat suivant.

Corollaire 2.2.3. *Sous la condition de Kalman précédente, si* $r = 0$ *et si* $0 \in \overset{\circ}{\Omega}$, *alors l'ensemble accessible* $Acc(x_0, t)$ *en temps* t *contient un voisinage du point* $\exp(tA)x_0$.

Remarque 2.2.3. Les propriétés de contrôlabilité globale sont reliées aux propriétés de stabilité de la matrice A. Par exemple il est clair que si

1. la condition de Kalman est remplie,
2. $r = 0$ et $0 \in \overset{\circ}{\Omega}$,
3. toutes les valeurs propres de la matrice A sont de partie réelle strictement négative (*i.e.* la matrice A est *stable*),

alors tout point de $\mathbb{R}^n$ peut être conduit à l'origine en temps fini (éventuellement grand).

Dans le cas mono-entrée $m = 1$, on a un résultat plus précis que nous admettrons (voir [52]).

Théorème 2.2.4. *Soit $b \in \mathbb{R}^n$ et $\Omega \subset \mathbb{R}$ un intervalle contenant 0 dans son intérieur. Considérons le système $\dot{x}(t) = Ax(t) + bu(t)$, avec $u(t) \in \Omega$. Alors tout point de $\mathbb{R}^n$ peut être conduit à l'origine en temps fini si et seulement si la paire (A, b) vérifie la condition de Kalman et la partie réelle de chaque valeur propre de A est inférieure ou égale à 0.*

2.2.3 Similitude de systèmes, forme de Brunovski

Définition 2.2.1. Les systèmes de contrôle linéaires $\dot{x}_1 = A_1x_1 + B_1u_1$ et $\dot{x}_2 = A_2x_2 + B_2u_2$ sont dits *semblables* s'il existe $P \in GL_n(\mathbb{R})$ tel que $A_2 = PA_1P^{-1}$ et $B_2 = PB_1$.

Remarque 2.2.4. On a alors $x_2 = Px_1$.

Proposition 2.2.5. *La propriété de Kalman est intrinsèque, i.e.*

$$(B_2, A_2B_2, \ldots, A_2^{n-1}B_2) = P(B_1, A_1B_1, \ldots, A_1^{n-1}B_1),$$

En particulier, le rang de la matrice de Kalman est invariant par similitude.

Considérons une paire (A, B) où $A \in \mathcal{M}_n(\mathbb{R})$ et $B \in \mathcal{M}_{n,m}(\mathbb{R})$.

Proposition 2.2.6. *La paire (A, B) est semblable à une paire (A', B') de la forme*

$$A' = \begin{pmatrix} A'_1 & A'_3 \\ 0 & A'_2 \end{pmatrix}, \quad B' = \begin{pmatrix} B'_1 \\ 0 \end{pmatrix},$$

où $A'_1 \in \mathcal{M}_r(\mathbb{R})$, $B'_1 \in \mathcal{M}_{r,m}(\mathbb{R})$, r étant le rang de la matrice de Kalman de la paire (A, B). De plus, la paire (A'_1, B'_1) est contrôlable.

Démonstration. Supposons que le rang r de la matrice de Kalman C de la paire (A, B) soit strictement plus petit que n (sinon il n'y a rien à montrer). Le sous-espace

$$F = \text{Im } C = \text{Im } B + \text{Im } AB + \cdots + \text{Im } A^{n-1}B$$

est de dimension r, et d'après le théorème d'Hamilton-Cayley il est clairement invariant par A. Soit G un supplémentaire de F dans $\mathbb{R}^n$, et soient $(f_1, \ldots, f_r)$ une base de F, et $(f_{r+1}, \ldots, f_n)$ une base de G. Notons P la matrice de passage

de la base $(f_1, \dots, f_n)$ à la base canonique de $\mathbb{R}^n$. Alors, puisque F est invariant par A, on a

$$A' = PAP^{-1} = \begin{pmatrix} A'_1 & A'_3 \\ 0 & A'_2 \end{pmatrix},$$

et d'autre part, puisque Im $B \subset F$, on a

$$B' = PB = \begin{pmatrix} B'_1 \\ 0 \end{pmatrix}.$$

Enfin, on voit facilement que le rang de la matrice de Kalman de la paire (A'_1, B'_1) est égal à celui de la paire (A, B). □

Théorème 2.2.7 (Forme de Brunovski)**.** *Si $m = 1$ et si la paire (A, B) est contrôlable, alors elle est semblable à la paire $(\tilde{A}, \tilde{B})$, où*

$$\tilde{A} = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

et où les coefficients a_i sont ceux du polynôme caractéristique de A, i.e.

$$\chi_A(X) = X^n + a_1 X^{n-1} + \cdots + a_{n-1} X + a_n.$$

Remarque 2.2.5. Dans ces nouvelles coordonnées, le système est alors équivalent à l'équation différentielle scalaire d'ordre n

$$x^{(n)}(t) + a_1 x^{(n-1)}(t) + \cdots + a_n x(t) = u(t).$$

Démonstration. Raisonnons par analyse et synthèse. S'il existe une base $(f_1, \dots, f_n)$ dans laquelle la paire (A, B) prend la forme $(\tilde{A}, \tilde{B})$, alors on a nécessairement $f_n = B$ à scalaire près, et

$$Af_n = f_{n-1} - a_1 f_n, \ \dots, \ Af_2 = f_1 - a_{n-1} f_n, \ Af_1 = -a_n f_n.$$

Définissons donc les vecteurs $f_1, \dots, f_n$ par les relations

$$f_n = B, \ f_{n-1} = Af_n + a_1 f_n, \ \dots, \ f_1 = Af_2 + a_{n-1} f_n.$$

La famille $(f_1, \dots, f_n)$ est bien une base de $\mathbb{R}^n$ puisque

$$\begin{gathered} \text{Vect } \{f_n\} = \text{Vect } \{B\}, \\ \text{Vect } \{f_n, f_{n-1}\} = \text{Vect } \{B, AB\}, \\ \vdots \\ \text{Vect } \{f_n, \dots, f_1\} = \text{Vect } \{B, \dots, A^{n-1}B\} = \mathbb{R}^n. \end{gathered}$$

Il reste à vérifier que l'on a bien $Af_1 = -a_n f_n$. On a

$$\begin{aligned} Af_1 &= A^2 f_2 + a_{n-1} A f_n \\ &= A^2(Af_3 + a_{n-2} f_n) + a_{n-1} A f_n \\ &= A^3 f_3 + a_{n-2} A^2 f_n + a_{n-1} A f_n \\ &\vdots \\ &= A^n f_n + a_1 A^{n-1} f_n + \cdots + a_{n-1} A f_n \\ &= -a_n f_n \end{aligned}$$

puisque d'après le théorème d'Hamilton-Cayley, on a $A^n = -a_1 A^{n-1} - \cdots - a_n I$. Dans la base $(f_1, \ldots, f_n)$, la paire (A, B) prend la forme $(\tilde{A}, \tilde{B})$. □

Remarque 2.2.6. Lorsque $m > 1$, ce théorème admet la généralisation suivante. Si la paire (A, B) est contrôlable, alors on peut la conjuguer à une paire $(\tilde{A}, \tilde{B})$ telle que

$$\tilde{A} = \begin{pmatrix} \tilde{A}_1 & * & \cdots & * \\ 0 & \tilde{A}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ 0 & \cdots & 0 & \tilde{A}_s \end{pmatrix},$$

où les matrices $\tilde{A}_i$ sont des matrices compagnons (*i.e.*, ayant la forme de Brunovski du théorème précédent) ; par ailleurs, il existe une matrice $G \in \mathcal{M}_{m,s}(\mathbb{R})$ telle que

$$\tilde{B}G = \begin{pmatrix} \tilde{B}_1 \\ \vdots \\ \tilde{B}_s \end{pmatrix},$$

où tous les coefficients de chaque matrice $\tilde{B}_i$ sont nuls, sauf celui de la dernière ligne, en i-ème colonne, qui est égal à 1.

Exercice 2.2.1. Tester la contrôlabilité des systèmes suivants.

– *Wagon*

$$m\ddot{x}(t) = u(t).$$

– *Oscillateur harmonique linéaire*

$$m\ddot{x}(t) + kx(t) = u(t).$$

– *Systèmes de ressorts amortis*

$$\begin{cases} m_1\ddot{x}_1 = -k_1(x_1 - x_2) - d_1(\dot{x}_1 - \dot{x}_2) + u, \\ m_2\ddot{x}_2 = k_1(x_1 - x_2) - k_2 x_2 + d_1(\dot{x}_1 - \dot{x}_2) - d_2\dot{x}_2. \end{cases}$$

– *Amortisseurs d'une voiture*

$$\begin{cases} \ddot{x}_1 = -k_1 x_1 - d_1 \dot{x}_1 + l_1 u, \\ \ddot{x}_2 = -k_2 x_2 - d_2 \dot{x}_2 + l_2 u. \end{cases}$$

– *Vitesse angulaire d'un rotor*

$$I\dot{\omega}(t) = u(t).$$

– *Circuit RLC*

$$L\frac{di}{dt} + Ri + \frac{q}{C} = u,$$

où $q(t) = \int^t i$ est la charge du condensateur. D'où

$$\begin{cases} \dfrac{dq}{dt} = i, \\ \dfrac{di}{dt} = -\dfrac{R}{L} i - \dfrac{1}{LC} q + \dfrac{1}{L} u. \end{cases}$$

– *Servomoteur à courant continu.* On note R la résistance, L l'inductance, e la force contre-électromotrice, k_1, k_2 des constantes, J le moment d'inertie du moteur, f le coefficient de frottement du moteur, $\Gamma = k_2 i$ le couple moteur, Γ_c le couple antagoniste, θ l'angle moteur. On a

$$\begin{cases} u = Ri + L\dfrac{di}{dt} + e, \\ e = k_1 \dot{\theta}, \\ J\ddot{\theta} = k l_2 i - f\dot{\theta} - \Gamma_c, \end{cases}$$

d'où

$$\frac{d}{dt}\begin{pmatrix} i \\ \theta \\ \dot{\theta} \end{pmatrix} = \begin{pmatrix} -R/L & 0 & k_1/L \\ 0 & 0 & 1 \\ k_2/J & 0 & -f/J \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} u + \begin{pmatrix} 0 \\ 0 \\ -\Gamma_c \end{pmatrix}.$$

– *Système de ressorts couplés (train à deux wagons)*

$$\begin{cases} \ddot{x} = -k_1 x + k_2 (y - x), \\ \ddot{y} = -k_2 (y - x) + u. \end{cases}$$

Exercice 2.2.2. Pour quelles valeurs de α le système

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 2 & \alpha - 3 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ \alpha^2 - \alpha & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}$$

est-il contrôlable ?

2.3 Contrôlabilité des systèmes linéaires instationnaires

Les deux théorèmes suivants donnent une condition nécessaire et suffisante de contrôlabilité dans le cas instationnaire.

Théorème 2.3.1. *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*

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

dite matrice de contrôlabilité, est inversible.

Remarque 2.3.1. Cette condition dépend de T, mais ne dépend pas du point initial x_0. Autrement dit, si un système linéaire instationnaire est contrôlable en temps T depuis x_0, alors il est contrôlable en temps T depuis tout point.

Remarque 2.3.2. On a $C(T) = C(T)^{\mathrm{T}}$, et $x^{\mathrm{T}} C(T) x \geqslant 0$ pour tout $x \in \mathbb{R}^n$, *i.e.* $C(T)$ est une *matrice carrée réelle symétrique positive.*

Démonstration. Pour toute solution $x(t)$, on a, d'après la formule de variation de la constante,

$$x(T) = x^* + M(T) \int_0^T M(t)^{-1} B(t) u(t) dt,$$

où

$$x^* = M(T)x_0 + M(T) \int_0^T M(t)^{-1} r(t) dt.$$

Si $C(T)$ est inversible, posons $u(t) = B(t)^{\mathrm{T}} M(t)^{-1^{\mathrm{T}}} \psi$, avec $\psi \in \mathbb{R}^n$. Alors

$$x(T) = x^* + M(T)C(T)\psi,$$

et il suffit de prendre $\psi = C(T)^{-1} M(T)^{-1} (x_1 - x^*)$.

Réciproquement, si $C(T)$ n'est pas inversible, alors il existe $\psi \in \mathbb{R}^n \setminus \{0\}$ tel que $\psi^{\mathrm{T}} C(T) \psi = 0$. On en déduit que

$$\int_0^T \| B(t)^{\mathrm{T}} M(t)^{-1^{\mathrm{T}}} \psi \|^2 dt = 0,$$

d'où $\psi^{\mathrm{T}} M(t)^{-1} B(t) = 0$ p.p. sur $[0, T]$. Ainsi, pour tout contrôle u, on a

$$\psi^{\mathrm{T}} \int_0^T M(t)^{-1} B(t) u(t) dt = 0.$$

Posons $\psi_1 = M(T)^{-1^{\mathrm{T}}} \psi$; on a, pour tout contrôle u,

$$\psi^{\mathrm{T}} (x_u(T) - x^*) = 0,$$

i.e. $x_u(T) \in x^* + \psi^{\perp}$, et donc le système n'est pas contrôlable. □

Remarque 2.3.3. Ce théorème peut se montrer beaucoup plus facilement en contrôle optimal, le contrôle utilisé dans la preuve étant optimal pour un certain critère.

Remarque 2.3.4. Si le système est autonome, on a $M(t) = e^{tA}$, et donc

$$C(T) = \int_0^T e^{-sA} B B^{\mathrm{T}} e^{-sA^{\mathrm{T}}} ds.$$

Dans ce cas, $C(T_1)$ est inversible si et seulement si $C(T_2)$ est inversible, et en particulier la condition de contrôlabilité ne dépend pas de T (ce qui est faux dans le cas instationnaire).

Théorème 2.3.2. *Considérons le système*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$$

où les applications A, B sont de classe C^∞ sur $[0,T]$. Définissons par récurrence

$$B_0(t) = B(t),\ B_{i+1}(t) = A(t)B_i(t) - \frac{dB_i}{dt}(t).$$

1. *S'il existe $t \in [0,T]$ tel que*

$$Vect\,\{B_i(t)v \mid v \in \mathbb{R}^m,\ i \in \mathbb{N}\} = \mathbb{R}^n,$$

   alors le système est contrôlable en temps T.
2. *Si de plus les applications A, B sont analytiques sur $[0,T]$, alors le système est contrôlable en temps T si et seulement si*

$$\forall t \in [0,T] \quad Vect\,\{B_i(t)v \mid v \in \mathbb{R}^m,\ i \in \mathbb{N}\} = \mathbb{R}^n.$$

Ce théorème se montre aisément en théorie du contrôle optimal, par l'application du principe du maximum (voir plus loin).

Remarque 2.3.5. Dans le cas autonome, on retrouve la condition de Kalman.

Exercice 2.3.1. Montrer que le système $\dot{x}(t) = A(t)x(t) + B(t)u(t)$, avec

$$A(t) = \begin{pmatrix} t & 1 & 0 \\ 0 & t^3 & 0 \\ 0 & 0 & t^2 \end{pmatrix},\ \text{et}\ B(t) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},$$

est contrôlable en temps quelconque.

Exercice 2.3.2. Montrer que le système

$$\begin{cases} \dot{x}(t) = -y(t) + u(t)\cos t, \\ \dot{y}(t) = x(t) + u(t)\sin t, \end{cases}$$

n'est pas contrôlable.

Exercice 2.3.3. Soient m et n des entiers naturels non nuls, et soient $A \in \mathcal{M}_n(\mathbb{R})$ et $B \in \mathcal{M}_{n,m}(\mathbb{R})$. On suppose que le système de contrôle $\dot{x}(t) = Ax(t) + Bu(t)$ est contrôlable. Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction de classe C^∞ ; on pose, pour tout $t \in \mathbb{R}$,

$$A(t) = A + f(t)I,$$

où I est la matrice identité d'ordre n. Montrer que le système de contrôle $\dot{x}(t) = A(t)x(t) + Bu(t)$ est contrôlable en temps quelconque.

Chapitre 3

Temps-optimalité

3.1 Existence de trajectoires temps-optimales

Il faut tout d'abord formaliser, à l'aide de $Acc(x_0, t)$, la notion de temps minimal. Considérons comme précédemment le système de contrôle dans $\mathbb{R}^n$

$$\dot{x}(t) = A(t)x(t) + b(t)u(t) + r(t),$$

où les contrôles u sont à valeurs dans un compact d'intérieur non vide $\Omega \subset \mathbb{R}^m$. Soient x_0 et x_1 deux points de $\mathbb{R}^n$. Supposons que x_1 soit accessible depuis x_0, c'est-à-dire qu'il existe au moins une trajectoire reliant x_0 à x_1. Parmi toutes les trajectoires reliant x_0 à x_1, on aimerait caractériser celles qui le font en temps minimal t^* (voir figure 3.1).

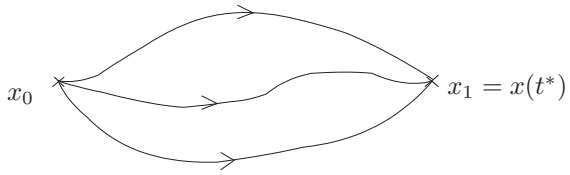


FIGURE 3.1 –

Si t^* est le temps minimal, alors pour tout $t < t^*$, $x_1 \notin Acc(x_0, t)$ (en effet sinon x_1 serait accessible à partir de x_0 en un temps inférieur à t^*). Par conséquent,

$$t^* = \inf\{t > 0 \mid x_1 \in Acc(x_0, t)\}.$$

Ce temps t^* est bien défini car, d'après le théorème 2.1.1, $Acc(x_0, t)$ varie continûment avec t, donc l'ensemble $\{t > 0 \mid x_1 \in Acc(x_0, t)\}$ est fermé dans $\mathbb{R}$. En particulier cette borne inférieure est atteinte.

Le temps $t = t^*$ est le premier temps pour lequel $Acc(x_0, t)$ contient x_1 (voir figure 3.2).

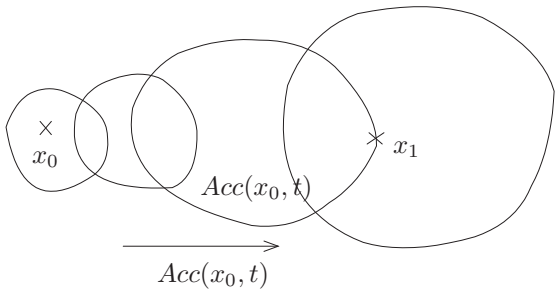


FIGURE 3.2 – Temps minimal

D'autre part, on a nécessairement

$$x_1 \in \partial Acc(x_0, t^*) = A(x_0, t^*) \setminus \overset{o}{Acc}(x_0, t^*).$$

En effet, si x_1 appartenait à l'intérieur de $Acc(x_0, t^*)$, alors pour $t < t^*$ proche de t^*, x_1 appartiendrait encore à $Acc(x_0, t)$ car $Acc(x_0, t)$ varie continûment avec t. Mais ceci contredit le fait que t^* soit le temps minimal.

En particulier on a prouvé le théorème d'existence suivant.

Théorème 3.1.1. *Si le point x_1 est accessible depuis x_0 alors il existe une trajectoire temps-minimale reliant x_0 à x_1.*

Remarque 3.1.1. On peut aussi se poser le problème d'atteindre une cible non réduite à un point. Ainsi, soit $(M_1(t))_{0 \leqslant t \leqslant T}$ une famille de sous-ensembles compacts de $\mathbb{R}^n$ variant continûment en t. Tout comme précédemment, on voit que s'il existe un contrôle u à valeurs dans Ω joignant x_0 à $M_1(T)$, alors il existe un contrôle temps-minimal défini sur $[0, t^*]$ joignant x_0 à $M(t^*)$.

Ces remarques donnent une vision géométrique de la notion de temps minimal, et conduisent à la définition suivante.

Définition 3.1.1. Le contrôle u est dit *extrémal* sur $[0, t]$ si la trajectoire du système (1.2) associée à u vérifie $x(t) \in \partial Acc(x_0, t)$.

En particulier, tout contrôle temps-minimal est extrémal. La réciproque est évidemment fausse car l'extrémalité ne fait pas la différence entre la minimalité et la maximalité.

Dans le paragraphe suivant on donne une caractérisation de cette propriété.

3.2 Condition nécessaire d'optimalité : principe du maximum dans le cas linéaire

Le théorème suivant donne une condition nécessaire et suffisante pour qu'un contrôle soit extrémal.

Théorème 3.2.1. *Considérons le système de contrôle linéaire*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0,$$

où le domaine de contraintes $\Omega \subset \mathbb{R}^m$ sur le contrôle est compact. Soit $T > 0$. Le contrôle u est extrémal sur $[0,T]$ si et seulement s'il existe une solution non triviale $p(t)$ de l'équation $\dot{p}(t) = -p(t)A(t)$ telle que

$$p(t)B(t)u(t) = \max_{v\in\Omega} p(t)B(t)v \tag{3.1}$$

pour presque tout $t \in [0,T]$. Le vecteur ligne $p(t) \in \mathbb{R}^n$ est appelé vecteur adjoint.

Remarque 3.2.1. La condition initiale $p(0)$ dépend en fait du point final x_1, comme on le voit dans la démonstration. Comme elle n'est pas directement connue, l'usage de ce théorème sera plutôt indirect, comme on le verra dans les exemples.

Remarque 3.2.2. Dans le cas mono-entrée (contrôle scalaire), et si de plus $\Omega = [-a,a]$ où $a > 0$, la condition de maximisation implique immédiatement que $u(t) = a\ \text{signe}(p(t)B(t))$. La fonction $\varphi(t) = p(t)B(t)$ est appelée *fonction de commutation*, et un temps t_c auquel le contrôle extrémal $u(t)$ change de signe est appelé un *temps de commutation*. C'est en particulier un zéro de la fonction φ.

Démonstration. On a vu que $Acc_\Omega(x_0,T) = Acc_{\text{Conv}(\Omega)}(x_0,T)$, et par conséquent on peut supposer que Ω est convexe. Si u est extrémal sur $[0,T]$, soit x la trajectoire associée à u. On a $x(T) \in \partial Acc(x_0,T)$. Par convexité de $Acc(x_0,T)$, il existe d'après le théorème du convexe (voir par exemple [19]) un hyperplan séparant au sens large $x(T)$ et $Acc(x_0,T)$. Soit p_T un vecteur normal à cet hyperplan (voir figure 3.3).

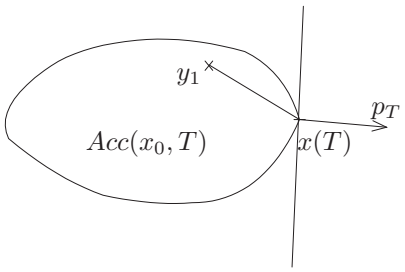


FIGURE 3.3 –

D'après le théorème du convexe,

$$\forall y_1 \in Acc(x_0,T) \quad p_T(y_1 - x(T)) \leqslant 0. \tag{3.2}$$

Par définition de $Acc(x_0, T)$, il existe un contrôle u_1 tel que la trajectoire associée $y(t)$ vérifie $y_1 = y(T)$. L'inégalité (3.2) se réécrit

$$p_T x(T) \geqslant p_T y(T).$$

D'où

$$\int_0^T p_T M(T)M(s)^{-1}(B(s)u(s)+r(s))ds \geqslant \int_0^T p_T M(T)M(s)^{-1}(B(s)u_1(s)+r(s))ds.$$

Appelons $p(t)$ la solution sur $[0,T]$ de $\dot{p} = -pA$, telle que $p(T) = p_T$. Alors il est clair que $p(t) = p(0)M(t)^{-1}$ et $p_T = p(T) = p(0)M(T)^{-1}$. Il s'ensuit que

$$\forall s \in [0,T] \quad p_T M(T)M(s)^{-1} = p(0)M(s)^{-1} = p(s),$$

et donc que

$$\int_0^T p(s)B(s)u_1(s)ds \leqslant \int_0^T p(s)B(s)u(s)ds \tag{3.3}$$

Si (3.1) n'est pas vraie alors

$$p(t)B(t)u(t) < \max_{v\in\Omega} p(t)B(t)v.$$

sur un sous-ensemble de $[0,T]$ de mesure positive. Soit alors $u_1(\cdot)$ sur $[0,T]$ à valeurs dans Ω tel que

$$p(t)B(t)u_1(t) = \max_{v\in\Omega} p(t)B(t)v.$$

En appliquant un lemme de sélection mesurable de théorie de la mesure, on peut montrer que l'application $u_1(\cdot)$ peut être choisie mesurable sur $[0,T]$ (voir [52, Lem. 2A, 3A p. 161]).

Comme u_1 est à valeurs dans Ω, l'inégalité (3.3) est vraie, alors que par ailleurs la définition de u_1 conduit immédiatement à l'inégalité stricte inverse, d'où la contradiction. Par conséquent (3.1) est vraie.

Réciproquement, supposons qu'il existe un vecteur adjoint tel que le contrôle u vérifie (3.1). Notons $x(\cdot)$ la trajectoire associée à u. On voit facilement en remontant le raisonnement précédent que

$$\forall y_1 \in Acc(x_0,T) \quad p(T)(y_1 - x(T)) \leqslant 0. \tag{3.4}$$

Raisonnons alors par l'absurde, et supposons que $x(T) \in \text{Int } Acc(x_0,T)$. Alors il existerait un point y_1 de $Acc(x_0,T)$ qui serait sur la demi-droite d'origine $x(T)$ et de direction $p(T)$ (voir figure 3.4). Mais alors $p(T)(y_1 - x(T)) > 0$, ce qui contredirait (3.4). Donc $x(T) \in \partial Acc(x_0,T)$, et u est extrémal. □

Remarque 3.2.3. Si u est extrémal sur $[0,T]$ alors u est aussi extrémal sur $[0,t]$ pour tout $t \in [0,T]$, et de plus $p(t)$ est un vecteur normal extérieur à $Acc(x_0,t)$. Cela découle facilement de la preuve et de la propriété (3.1).

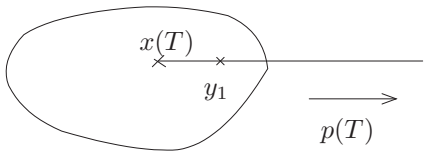


FIGURE 3.4 –

Remarque 3.2.4. Puisque tout contrôle temps-minimal est extrémal, le théorème précédent, qui est le principe du maximum dans le cas linéaire, donne une *condition nécessaire* d'optimalité.

Remarque 3.2.5. Si u est un contrôle temps-minimal joignant en temps T une cible M_1, où $M_1 \subset \mathbb{R}^n$ est convexe, alors on peut de plus choisir le vecteur adjoint pour que le vecteur $p(T)$ soit unitaire et normal à un hyperplan séparant (au sens large) $Acc(x_0, T)$ et M_1. C'est une condition dite de *transversalité*, obtenue facilement dans la preuve précédente.

Comme exemple théorique d'application, montrons le résultat suivant.

Proposition 3.2.2. *Considérons dans $\mathbb{R}^n$ le système linéaire autonome $\dot{x}(t) = Ax(t) + Bu(t)$, avec $B \in \mathbb{R}^n$ et $|u(t)| \leqslant 1$, et où la paire (A, B) vérifie la condition de Kalman.*

1. *Si toute valeur propre de A est réelle, alors tout contrôle extrémal a au plus $n-1$ commutations sur $\mathbb{R}^+$.*
2. *Si toute valeur propre de A a une partie imaginaire non nulle, alors tout contrôle extrémal a un nombre infini de commutations sur $\mathbb{R}^+$.*

Démonstration. D'après le théorème 2.2.7, le système peut s'écrire sous forme de Brunovski, et il est alors équivalent à une équation différentielle scalaire d'ordre n de la forme

$$x^{(n)} + a_1 x^{(n-1)} + \cdots + a_n x = u, \ |u| \leqslant 1.$$

De plus, tout contrôle extrémal est de la forme $u(t) = \text{signe } \lambda(t)$, où $\lambda(t)$ est la dernière coordonnée du vecteur adjoint, qui vérifie l'équation différentielle

$$\lambda^{(n)} - a_1 \lambda^{(n-1)} + \cdots + (-1)^n a_n \lambda = 0.$$

En effet le vecteur adjoint vérifie $p'(t) = -p(t)A(t)$.

1. Si toute valeur propre de A est réelle, alors $\lambda(t)$ s'écrit sous la forme

$$\lambda(t) = \sum_{j=1}^{r} P_j(t) e^{\lambda_j t},$$

où P_j est un polynôme de degré inférieur ou égal à $n_j - 1$, et où $\lambda_1, \ldots, \lambda_r$, sont les r valeurs propres distinctes de $-A$, de multiplicités respectives $n_1, \ldots, n_r$. Notons que $n = n_1 + \cdots + n_r$. On montre alors facilement, par récurrence, que $\lambda(t)$ admet au plus $n-1$ zéros.

2. Si toute valeur propre de A a une partie imaginaire non nulle, alors, comme précédemment, on peut écrire

$$\lambda(t) = \sum_{j=1}^{r} (P_j(t)\cos\beta_j t + Q_j(t)\sin\beta_j t)e^{\alpha_j t},$$

où $\lambda_j = \alpha_j + i\beta_j$, et P_j, Q_j sont des polynômes réels non nuls. En mettant en facteur un terme $t^k e^{\alpha_i t}$ de plus haut degré (*i.e.* dominant), on voit facilement que $\lambda(t)$ a un nombre infini de zéros.

□

3.3 Exemples

3.3.1 Synthèse optimale pour le problème de l'oscillateur harmonique linéaire

Appliquons la théorie précédente à l'exemple de l'oscillateur harmonique présenté en introduction, pour $k_2 = 0$, et répondons aux deux questions suivantes :

1. Pour toute condition initiale $x(0) = x_0, \dot{x}(0) = y_0$, existe-t-il une force extérieure horizontale (un contrôle), vérifiant la contrainte, qui permette d'amener la masse ponctuelle à sa position d'équilibre $x(T) = 0, \dot{x}(T) = 0$ en un temps fini T ?
2. Si la première condition est remplie, peut-on de plus déterminer cette force de manière à minimiser le temps ?

Enfin, ces deux problèmes résolus, nous représenterons dans le plan de phase la trajectoire optimale obtenue.

Contrôlabilité du système

Le système s'écrit

$$\left\{ \begin{array}{rcl} \dot{X} & = & AX + Bu \\ X(0) & = & X_0 \end{array} \right. \quad \text{avec } A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

On a facilement $\mathrm{rg}(B, AB) = 2$; par ailleurs les valeurs propres de A sont de partie réelle nulle. Donc, d'après le théorème 2.2.4, le système est contrôlable à 0, *i.e.* il existe des contrôles u vérifiant la contrainte $|u| \leqslant 1$ tels que les trajectoires associées relient X_0 à 0, ce qui répond à la première question.

Interprétation physique

- Si l'on n'applique aucune force extérieure, *i.e.* $u = 0$, alors l'équation du mouvement est $\ddot{x}+x = 0$. La masse ponctuelle oscille, et ne s'arrête jamais, donc ne parvient pas à sa position d'équilibre en un temps fini.
- Si l'on applique certaines forces extérieures, on a tendance à amortir les oscillations. La théorie prévoit qu'on parvient à stopper l'objet en un temps fini.

Calcul du contrôle optimal

D'après le paragraphe précédent, il existe des contrôles permettant de relier X_0 à 0. On cherche maintenant à le faire en temps minimal. Pour cela, on applique le théorème 3.2.1, selon lequel

$$u(t) = \text{signe}(p(t)B),$$

où $p(t) \in \mathbb{R}^2$ est solution de $\dot{p} = -pA$. Posons $p = (p_1, p_2)$. Alors $u(t) = \text{signe}(p_2(t))$, et $\dot{p}_1 = p_2, \dot{p}_2 = -p_1$, d'où $\ddot{p}_2+p_2 = 0$. Donc $p_2(t) = \lambda\cos t+\mu\sin t$. En particulier, la durée entre deux zéros consécutifs de $p_2(t)$ est exactement π. Par conséquent le contrôle optimal est constant par morceaux sur des intervalles de longueur π, et prend alternativement les valeurs ± 1.

- Si $u = -1$, on obtient le système différentiel

$$\begin{cases} \dot{x} = y, \\ \dot{y} = -x - 1. \end{cases} \tag{3.5}$$

- Si $u = +1$,

$$\begin{cases} \dot{x} = y, \\ \dot{y} = -x + 1. \end{cases} \tag{3.6}$$

La trajectoire optimale finale, reliant X_0 à 0, sera constituée d'arcs successifs, solutions de (3.5) et (3.6).

Solutions de (3.5). On obtient facilement $(x + 1)^2 + y^2 = \text{cste} = R^2$, donc les courbes solutions de (3.5) sont des cercles centrés en $(-1, 0)$, et de période 2π (en fait, $x(t) = -1 + R\cos t, y(t) = R\sin t$).

Solutions de (3.6). On obtient $x(t) = 1 + R\cos t$ et $y(t) = R\sin t$. Les solutions sont des cercles centrés en $(1, 0)$, de période 2π.

La trajectoire optimale de X_0 à 0 doit donc suivre alternativement un arc de cercle centré en $(-1, 0)$, et un arc de cercle centré en $(1, 0)$.

Quitte à changer t en $-t$, nous allons raisonner en temps inverse, et construire la trajectoire optimale menant de 0 à X_0. Pour cela, nous allons considérer toutes les trajectoires optimales partant de 0, et nous sélectionnerons celle qui passe par X_0.

En faisant varier $p(0)$, on fait varier la trajectoire optimale. En effet, d'après le théorème de Cauchy-Lipschitz, $p(0)$ détermine $p(t)$ pour tout t, ce qui définit un contrôle optimal $u(t)$, et donc une trajectoire optimale.

Prenons des exemples pour commencer à représenter l'allure des trajectoires optimales possibles.

– Si $p_1(0) = 1, p_2(0) = 0$, alors $p_2(t) = -\sin t$, donc sur $]0, \pi[$ on a $u(t) = \text{signe}(p_2(t)) = -1$. La trajectoire optimale correspondante, partant de 0, suit donc pendant un temps π l'arc de cercle Γ_- solution de (3.5), passant par 0 (voir figure 3.5).

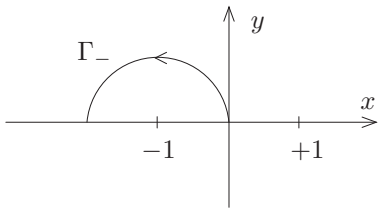


FIGURE 3.5 –

– Si $p_1(0) = -1, p_2(0) = 0$, alors $p_2(t) = \sin t$, donc sur $]0, \pi[$ on a $u(t) = \text{signe}(p_2(t)) = +1$. La trajectoire optimale correspondante, partant de 0, suit donc pendant un temps π l'arc de cercle Γ_+ solution de (3.6), passant par 0 (voir figure 3.6).

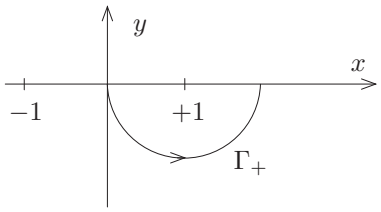


FIGURE 3.6 –

– Pour tout autre choix de $p(0)$ tel que $p_2(0) > 0$, la trajectoire optimale correspondante part de l'origine en suivant Γ_+ *jusqu'à ce que* $p_2(t) = 0$. Au-delà de ce point, $p_2(t)$ change de signe, donc le contrôle *commute* et prend la valeur -1, pendant une durée π (*i.e.* jusqu'à ce que $p_2(t)$ change à nouveau de signe). La trajectoire optimale doit alors être solution de (3.5), en partant de ce point de commutation M, et doit donc suivre un arc de cercle centré en $(-1, 0)$, pendant un temps π. C'est donc un demi-cercle, vu la paramétrisation des courbes de (3.5) (voir figure 3.7).
La trajectoire optimale rencontre un deuxième point de commutation N lorsque à nouveau $p_2(t)$ change de signe. On remarque que M et N sont symétriques par rapport au point $(-1, 0)$ (en effet ce sont les extrémités d'un

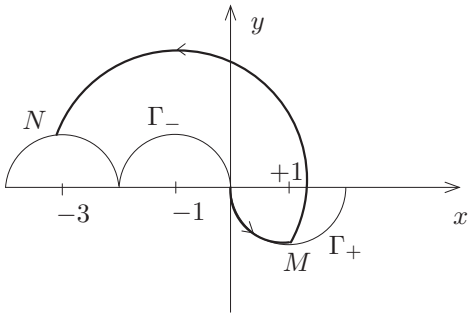


FIGURE 3.7 –

demi-cercle centré en ce point). Le point M appartenant au demi-cercle Γ_+, le point N appartient au demi-cercle image de Γ_+ par la symétrie par rapport au point $(-1, 0)$ qui est aussi, comme on le voit facilement, le translaté à gauche de Γ_- par la translation de vecteur $(-2, 0)$.

Poursuivons alors notre raisonnement. On se rend compte que les points de commutation de cette trajectoire optimale partant de 0 sont situés sur la courbe W construite de la manière suivante : W est l'union de tous les translatés à gauche de Γ_- par la translation précédente, et aussi de tous les translatés à droite de Γ_+ (voir figure 3.8).

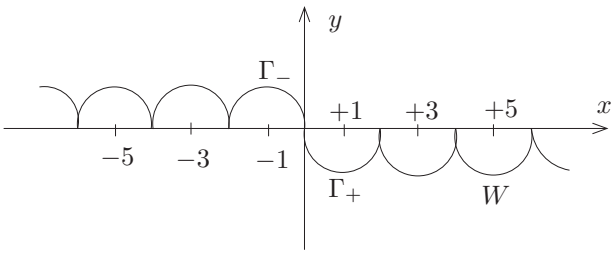


FIGURE 3.8 – Ensemble W

Les trajectoires optimales sont alors construites de la manière suivante : on part de 0 et l'on suit un morceau de Γ_+ ou Γ_-, jusqu'à un premier point de commutation. Si par exemple on était sur Γ_+, alors partant de ce point on suit un arc de cercle centré en $(-1, 0)$, au-dessus de W, jusqu'à ce qu'on rencontre W. De ce deuxième point de commutation, on suit un arc de cercle centré en $(1, 0)$ jusqu'à rencontrer W en un troisième point de commutation, etc (voir figure 3.9).

On est maintenant en mesure de répondre à la deuxième question, du moins graphiquement. Le but est de relier 0 et X_0 par une trajectoire optimale. La théorie prévoit qu'on peut effectivement le faire. Une trajectoire partant de 0

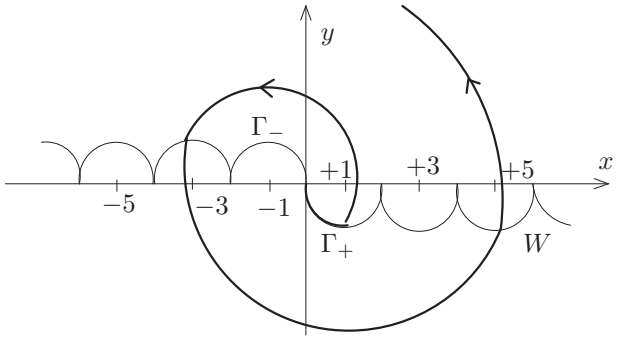


FIGURE 3.9 –

est, comme on vient de le voir ci-dessus, déterminée par deux choix :

1. partant de 0, on peut suivre un morceau de Γ_+ ou de Γ_-.
2. il faut choisir le premier point de commutation.

Si maintenant on se donne un point $X_0 = (x_0, y_0)$ du plan de phase, on peut déterminer graphiquement ces deux choix, et obtenir un tracé de la trajectoire optimale (voir figure 3.10). Dans la pratique il suffit d'inverser le temps, *i.e.* de partir du point final et d'atteindre le point initial.

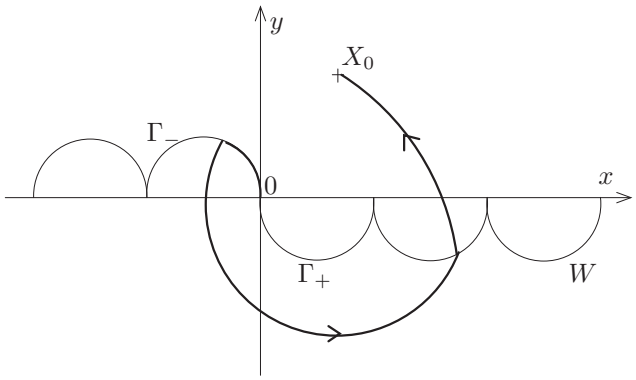


FIGURE 3.10 – Synthèse optimale

Remarque 3.3.1. L'implémentation numérique de cet exemple est très facile à faire. Nous la ferons plutôt dans le cas non linéaire où elle est plus intéressante.

3.3.2 Autres exemples

Exemple 3.3.1. [52] Considérons le système de contrôle

$$\dot{x} = y + u,\ \dot{y} = -y + u,\ |u| \leqslant 1.$$

Le but est de joindre en temps minimal la droite $x = 0$, puis de rester sur cette droite.

Remarquons tout d'abord que si une trajectoire reste dans $x = 0$, cela implique $y(t) = -u(t)$, et donc $|y| \leqslant 1$. Réciproquement de tout point $(0, y)$ avec $|y| \leqslant 1$ part une trajectoire restant dans le lieu $x = 0, |y| \leqslant 1$; il suffit de choisir $u(t) = -ye^{-2t}$. Par conséquent la cible est

$$M_1 = \{(0, y) \mid |y| \leqslant 1\}.$$

C'est un compact convexe.

Le système est du type $\dot{X} = AX + Bu$ avec

$$A = \begin{pmatrix} 0 & 1 \\ 0 & -1 \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

On vérifie facilement la condition de Kalman, et d'autre part les valeurs propres de A sont 0 et -1. D'après le théorème 2.2.4 le système est donc contrôlable à 0, et donc la cible M_1 est atteignable de tout point.

Comme dans le cas précédent, raisonnons en temps inverse en calculant les trajectoires optimales joignant M_1 à tout point final. Le système extrémal s'écrit alors

$$\dot{x} = -y - u,\ \dot{y} = y - u,\ \dot{p}_x = 0,\ \dot{p}_y = p_x - p_y,$$

où $u(t) = -\text{signe}(p_x(t) + p_y(t))$. On intègre aisément $p_x(t) = \text{cste} = p_x$ et $p_y(t) = p_x + (p_y(0) - p_x)e^{-t}$. En particulier $p_x + p_y$ est strictement monotone et donc le contrôle u admet au plus une commutation.

Par ailleurs la condition de transversalité (voir remarque 3.2.5) impose que si $x(0) = 0, |y(0)| < 1$ alors $p_x(0) = \pm 1$ et $p_y(0) = 0$. Mais alors $p_x(t) + p_y(t) = \pm(1 - e^{-t})$, et u ne commute pas sur $\mathbb{R}^+$. Par exemple si $p_x = 1$ on obtient $u(t) = -1$ pour tout $t \geqslant 0$, ce qui donne les courbes en pointillé sur la figure 3.11. La courbe limite est obtenue pour $u = -1$, partant du point $x = 0$, $y = 1$, et s'écrit

$$\Gamma_- = \{(-2e^t + 2t + 2, 2e^t - 1) \mid t \geqslant 0\}.$$

Calculons maintenant les extrémales partant du point $(0, 1)$. La condition de transversalité s'écrit alors $p_x(0) = \cos\alpha$, $p_y(0) = -\sin\alpha$, avec $0 \leqslant \alpha \leqslant \pi$. Par conséquent

$$u(t) = -\text{signe}(2\cos\alpha - (\sin\alpha + \cos\alpha)e^{-t}),$$

et l'on a une commutation si et seulement s'il existe $t \geqslant 0$ tel que

$$e^{-t} = \frac{\cos\alpha}{\sin\alpha + \cos\alpha}.$$

– Si $0 \leqslant \alpha < \frac{\pi}{4}$ alors $\frac{2\cos\alpha}{\sin\alpha+\cos\alpha} > 1$ donc l'équation ci-dessus n'a pas de solution, et donc $u(t) = +1$ sur $\mathbb{R}^+$.
– Si $\frac{\pi}{4} \leqslant \alpha < \frac{\pi}{2}$, l'équation a une solution $t(\alpha) \geqslant 0$, et l'on voit facilement que $t(\alpha)$ est strictement croissante de $[\frac{\pi}{4}, \frac{\pi}{2}]$ dans $[0, +\infty[$. Le contrôle vaut alors -1 sur $[0, t(\alpha)[$ et $+1$ ensuite.
– Si $\frac{\pi}{2} \leqslant \alpha \leqslant \pi$, l'équation n'a pas de solution dans $\mathbb{R}^+$, et on trouve $u(t) = -1$ sur $\mathbb{R}^+$.

Ainsi dans le deuxième cas, l'extrémale partant du point $(0, 1)$ suit pendant un moment la courbe Γ_-, puis commute sur $u = +1$.

Enfin, on définit Γ_+, symétrique de Γ_- par rapport à l'origine (voir figure 3.11). Finalement, le lieu de commutation est $\Gamma_- \cup \Gamma_+$, et l'on peut exprimer en fonction de x, y la loi de commande optimale $u(x, y) = -1$ (resp. $+1$) si (x, y) est au-dessus de W ou sur Γ_- (resp. en dessous de W ou sur Γ_+), où $W = \Gamma_- \cup M_1 \cup \Gamma_+$.

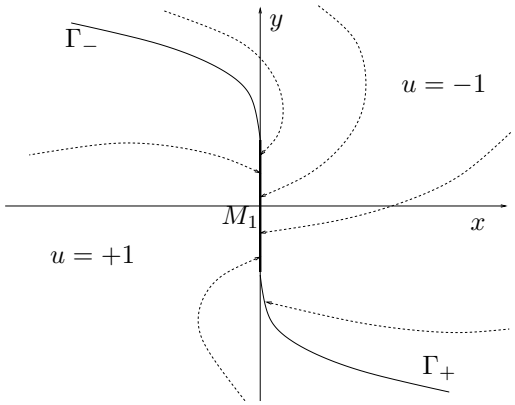


FIGURE 3.11 – Synthèse optimale

Exemple 3.3.2. Considérons le système dans $\mathbb{R}^2$

$$\dot{x}_1 = x_2,\ \dot{x}_2 = 2x_2 + u,\ |u| \leqslant 1.$$

On se pose le problème de relier en temps minimal le point origine $(0, 0)$ à tout point $(a, 0)$, où $a \in \mathbb{R}$. Sans perte de généralité on peut supposer que $a > 0$.

On peut facilement vérifier que le système est contrôlable. Par ailleurs le système adjoint s'écrit

$$\dot{p}_1 = 0,\ \dot{p}_2 = -p_1 - 2p_2,$$

et le contrôle extrémal est $u = \text{signe}(p_2)$. On a facilement $p_1 = \text{cste}$, puis $p_2(t) = -\frac{1}{2}p_1 + \lambda e^{-2t}$. En particulier $p_2(t)$ est strictement monotone donc le

contrôle a au plus une commutation. En posant $u = \varepsilon = \pm 1$ on intègre aisément

$$x_1(t) = -\frac{\varepsilon}{2}(t - t_0) + \frac{1}{2}(x_2(t_0) + \frac{\varepsilon}{2})(e^{2(t-t_0)} - 1) + x_1(t_0),$$
$$x_2(t) = -\frac{\varepsilon}{2} + (x_2(t_0) + \frac{\varepsilon}{2})e^{2(t-t_0)}.$$

On peut alors représenter le flot extrémal (voir figure 3.12). Notons le changement de monotonie en $x_2 = \pm\frac{1}{2}$.

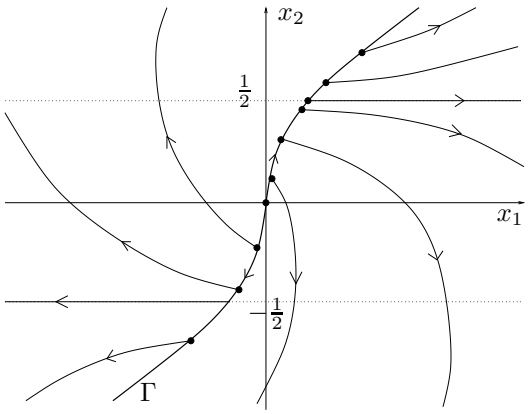


FIGURE 3.12 – Extrémales de l'exemple 3.3.2

On note Γ la courbe, en gras sur la figure, réunion des deux extrémales passant par l'origine et associées respectivement aux contrôles $u = +1$ et $u = -1$. Il est clair que Γ est la courbe de commutation, et que $u = +1$ si on est au-dessus de Γ, ou sur Γ avec $x_2 > 0$, et $u = -1$ si on est en dessous de Γ ou sur Γ avec $x_2 < 0$. Il est alors clair que, pour aller en temps minimal de l'origine à un point $(a, 0)$ où $a > 0$, il faut d'abord prendre $u = +1$, *i.e.* suivre un morceau de la courbe Γ, puis commuter (avant d'arriver à $x_2 = \frac{1}{2}$) et suivre un arc associé à $u = -1$. Par exemple si $a > 0$ est très grand, le point de commutation doit être très proche de la droite $x_2 = \frac{1}{2}$.

Chapitre 4

Théorie linéaire-quadratique

Dans ce chapitre on s'intéresse aux systèmes de contrôle linéaires avec un coût quadratique. Ces systèmes sont d'une grande importance dans la pratique, comme on le verra en section 4.4. En effet un coût quadratique est souvent très naturel dans un problème, par exemple lorsqu'on veut minimiser l'écart au carré par rapport à une trajectoire nominale (problème de poursuite). Par ailleurs même si les systèmes de contrôle sont en général non linéaires, on est très souvent amené à linéariser le système le long d'une trajectoire, par exemple dans des problèmes de stabilisation.

Nous allons donc considérer un système de contrôle linéaire dans $\mathbb{R}^n$

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \tag{4.1}$$

muni d'un coût quadratique du type

$$C(u) = x(T)^{\mathrm{T}} Q x(T) + \int_0^T \Big(x(t)^{\mathrm{T}} W(t) x(t) + u(t)^{\mathrm{T}} U(t) u(t) \Big)\, dt, \tag{4.2}$$

où $T > 0$ est fixé, et où, pour tout $t \in [0, T]$, $U(t) \in \mathcal{M}_m(\mathbb{R})$ est symétrique définie positive, $W(t) \in \mathcal{M}_n(\mathbb{R})$ est symétrique positive, et $Q \in \mathcal{M}_n(\mathbb{R})$ est une matrice symétrique positive. On suppose que les dépendances en t de A, B, W et U sont L^∞ sur $[0, T]$. Par ailleurs le coût étant quadratique, l'espace naturel des contrôles est $L^2([0, T], \mathbb{R}^m)$.

Le problème de contrôle optimal est alors le suivant, que nous appellerons *problème LQ* (linéaire-quadratique).

Problème LQ : Un point initial $x_0 \in \mathbb{R}^n$ étant fixé, l'objectif est de déterminer les trajectoires partant de x_0 qui minimisent le coût $C(u)$.

Notons que l'on n'impose aucune contrainte sur le point final $x(T)$. Pour toute la suite, on pose

$$\|x(t)\|_W^2 = x(t)^{\mathrm{T}} W(t) x(t),\ \|u(t)\|_U^2 = u(t)^{\mathrm{T}} U(t) u(t),\ \text{et } g(x) = x^{\mathrm{T}} Q x,$$

de sorte que

$$C(u) = g(x(T)) + \int_0^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt.$$

Les matrices Q, W, U sont des matrices de *pondération*.

Remarque 4.0.2. Par hypothèse les matrices Q et $W(t)$ sont symétriques positives, mais pas nécessairement définies. Par exemple si $Q = 0$ et $W = 0$ alors le coût est toujours minimal pour le contrôle $u = 0$.

Remarque 4.0.3. Comme dans le chapitre précédent, on suppose pour alléger les notations que le temps initial est égal à 0. Cependant tous les résultats qui suivent sont toujours valables si on considère le problème LQ sur un intervalle $[t_0, T]$, avec des contrôles dans l'espace $L^2([t_0, T], \mathbb{R}^m)$.

Remarque 4.0.4. Les résultats des sections 4.1 et 4.2 seront en fait valables pour des systèmes linéaires perturbés $\dot{x} = Ax + Bu + r$, et avec une fonction g de $\mathbb{R}^n$ dans $\mathbb{R}$ continue ou C^1. Nous préciserons pour chaque résultat les extensions possibles.

De même nous envisagerons le cas où $T = +\infty$.

4.1 Existence de trajectoires optimales

Introduisons l'hypothèse suivante sur U.

$$\exists \alpha > 0 \mid \forall u \in L^2([0, T], \mathbb{R}^m) \quad \int_0^T \|u(t)\|_U^2 dt \geqslant \alpha \int_0^T u(t)^{\mathrm{T}} u(t) dt. \tag{4.3}$$

Par exemple cette hypothèse est vérifiée si l'application $t \mapsto U(t)$ est continue sur $[0, T]$ et $T < +\infty$, ou encore s'il existe une constante $c > 0$ telle que pour tout $t \in [0, T]$ et pour tout vecteur $v \in \mathbb{R}^m$ on ait $v^{\mathrm{T}} U(t) v \geqslant c v^{\mathrm{T}} v$.

On a le théorème d'existence suivant.

Théorème 4.1.1. *Sous l'hypothèse (4.3), il existe une unique trajectoire minimisante pour le problème LQ.*

Démonstration. Montrons tout d'abord l'existence d'une telle trajectoire. Considérons une suite minimisante $(u_n)_{n \in \mathbb{N}}$ de contrôles sur $[0, T]$, *i.e.* la suite $C(u_n)$ converge vers la borne inférieure des coûts. En particulier cette suite est bornée. Par hypothèse, il existe une constante $\alpha > 0$ telle que pour tout $u \in L^2([0, T], \mathbb{R}^m)$ on ait $C(u) \geqslant \alpha \|u\|_{L^2}$. On en déduit que la suite $(u_n)_{n \in \mathbb{N}}$ est bornée dans $L^2([0, T], \mathbb{R}^m)$. Par conséquent à sous-suite près elle converge faiblement vers un contrôle u de L^2. Notons x_n (resp. x) la trajectoire associée au contrôle u_n (resp. u) sur $[0, T]$. D'après la formule de variation de la constante, on a, pour tout $t \in [0, T]$,

$$x_n(t) = M(t) x_0 + M(t) \int_0^t M(s)^{-1} B(s) u_n(s) ds \tag{4.4}$$

(et la formule analogue pour $x(t)$). On montre alors aisément que, à sous-suite près, la suite (x_n) converge simplement vers l'application x sur $[0,T]$ (en fait on peut même montrer que la convergence est uniforme).

Passant maintenant à la limite dans (4.4), on obtient, pour tout $t \in [0,T]$,

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

et donc x est une solution du système associée au contrôle u. Montrons qu'elle est minimisante. Pour cela on utilise le fait que puisque $u_n \rightharpoonup u$ dans L^2, on a l'inégalité

$$\int_0^T \|u(t)\|_U^2 dt \leqslant \liminf \int_0^T \|u_n(t)\|_U^2 dt,$$

et donc $C(u) \leqslant \liminf C(u_n)$. Mais comme (u_n) est une suite minimisante, $C(u)$ est égal à la borne inférieure des coûts, *i.e.* le contrôle u est minimisant, ce qui montre l'existence d'une trajectoire optimale.

Pour l'unicité on a besoin du lemme suivant.

Lemme 4.1.2. *La fonction C est strictement convexe.*

Preuve du lemme. Tout d'abord remarquons que pour tout $t \in [0,T]$, la fonction $f(u) = u^{\mathrm{T}}U(t)u$ définie sur $\mathbb{R}^m$ est strictement convexe puisque par hypothèse la matrice $U(t)$ est symétrique définie positive. Ensuite, notons $x_u(\cdot)$ la trajectoire associée à un contrôle u. On a pour tout $t \in [0,T]$,

$$x_u(t) = M(t)x_0 + M(t)\int_0^t M(s)^{-1}B(s)u(s)ds.$$

Par conséquent, comme dans la preuve du théorème 2.1.1, l'application qui à un contrôle u associe $x_u(t)$ est convexe, ceci pour tout $t \in [0,T]$. La matrice $W(t)$ étant symétrique positive, ceci implique la convexité de l'application qui à un contrôle u associe $x(t)^{\mathrm{T}}W(t)w(t)$. On raisonne de même pour le terme $x(T)^{\mathrm{T}}Qx(T)$. Enfin, l'intégration respectant la convexité, on en déduit que le coût est strictement convexe en u. □

L'unicité de la trajectoire optimale en résulte trivialement.

□

Remarque 4.1.1 (Extension du théorème 4.1.1). Si la fonction g apparaissant dans le coût est une fonction continue quelconque de $\mathbb{R}^n$ dans $\mathbb{R}$, bornée inférieurement ou convexe, et/ou si le système de contrôle est perturbé par une fonction $r(t)$, alors le théorème précédent reste vrai.

Remarque 4.1.2 (Cas d'un intervalle infini). Le théorème est encore valable si $T = +\infty$, avec $g = 0$, pourvu que le système (4.1) soit contrôlable (en temps quelconque).

En effet il suffit juste de montrer qu'il existe des trajectoires solutions du système (4.1) sur $[0,+\infty[$ et de coût fini. Or si le système est contrôlable, alors

il existe un contrôle u et un temps $T > 0$ tel que la trajectoire associée à u relie x_0 à 0 sur $[0, T]$. On étend alors le contrôle u par 0 sur $]T, +\infty[$, de sorte que la trajectoire reste en 0. On a ainsi construit une trajectoire solution du système sur $[0, +\infty[$ et de coût fini. Ceci permet d'affirmer l'existence d'une suite de contrôles minimisants. Les autres arguments de la preuve sont inchangés. On obtient donc le résultat suivant.

Proposition 4.1.3. *Considérons le problème de déterminer une trajectoire solution de*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t)$$

sur $[0, +\infty[$ *et minimisant le coût*

$$C(u) = \int_0^{+\infty} \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt.$$

Si le système est contrôlable en un temps $T > 0$*, et si l'hypothèse (4.3) est vérifiée sur* $[0, +\infty[$*, alors il existe une unique trajectoire minimisante.*

Remarque 4.1.3.
- Si l'on suppose de plus que les applications $A(\cdot)$ et $B(\cdot)$ sont L^2 sur $[0, +\infty[$, et si $W(\cdot)$ vérifie comme U une hypothèse de coercivité (4.3), alors la trajectoire minimisante tend vers 0 lorsque t tend vers l'infini.
  En effet on montre facilement en utilisant l'inégalité de Cauchy-Schwarz que l'application $\dot{x}(\cdot)$ est dans L^1, et par conséquent que $x(t)$ converge. Sa limite est alors forcément nulle.
- Dans le cas autonome (A et B sont constantes), si $W(\cdot)$ vérifie comme U une hypothèse de coercivité (4.3), alors la trajectoire minimisante tend vers 0 lorsque t tend vers l'infini.
  En effet il suffit d'écrire l'inégalité

  $$\|\dot{x}(t)\| \leqslant \|A\|\|x(t)\| + \|B\|\|u(t)\| \leqslant Cste(\|x(t)\|^2 + \|u(t)\|^2),$$

  puis en intégrant on montre de même que l'application $\dot{x}(\cdot)$ est dans L^1.

4.2 Condition nécessaire et suffisante d'optimalité : principe du maximum dans le cas LQ

Théorème 4.2.1. *La trajectoire x, associée au contrôle u, est optimale pour le problème LQ si et seulement s'il existe un vecteur adjoint $p(t)$ vérifiant pour presque tout $t \in [0, T]$*

$$\dot{p}(t) = -p(t)A(t) + x(t)^{\mathrm{T}} W(t) \tag{4.5}$$

et la condition finale

$$p(T) = -x(T)^{\mathrm{T}} Q. \tag{4.6}$$

De plus le contrôle optimal u s'écrit, pour presque tout $t \in [0, T]$,

$$u(t) = U(t)^{-1} B(t)^{\mathrm{T}} p(t)^{\mathrm{T}}. \tag{4.7}$$

Démonstration. Soit u un contrôle optimal et x la trajectoire associée sur $[0,T]$. Le coût est donc minimal parmi toutes les trajectoires solutions du système, partant de x_0, le point final étant non fixé. Considérons alors des perturbations du contrôle u dans $L^2([0,T],\mathbb{R}^m)$ du type

$$u_{pert}(t) = u(t) + \delta u(t),$$

engendrant les trajectoires

$$x_{pert}(t) = x(t) + \delta x(t) + \mathrm{o}(\|\delta u\|_{L^2}),$$

avec $\delta x(0) = 0$. La trajectoire x_{pert} devant être solution du système $\dot{x}_{pert} = Ax_{pert} + Bu_{pert}$, on en déduit que

$$\delta\dot{x} = A\delta x + B\delta u,$$

et par conséquent, pour tout $t \in [0,T]$,

$$\delta x(t) = M(t)\int_0^t M(s)^{-1}B(s)\delta u(s)ds. \tag{4.8}$$

Par ailleurs il est bien clair que le coût $C(\cdot)$ est une fonction lisse sur $L^2([0,T],\mathbb{R}^m)$ (elle est même analytique) au sens de Fréchet. Le contrôle u étant minimisant on doit avoir

$$dC(u) = 0.$$

Or

$$C(u_{pert}) = g(x_{pert}(T)) + \int_0^T (\|x_{pert}(t)\|_W^2 + \|u_{pert}(t)\|_U^2)dt,$$

et comme Q, $W(t)$ et $U(t)$ sont symétriques, on en déduit que

$$\frac{1}{2}dC(u).\delta u = x(T)^{\mathrm{T}}Q\delta x(T) + \int_0^T (x(t)^{\mathrm{T}}W(t)\delta x(t) + u(t)^{\mathrm{T}}U(t)\delta u(t))dt = 0, \tag{4.9}$$

ceci étant valable pour toute perturbation δu. Cette équation va nous conduire à l'expression du contrôle optimal u. Mais introduisons tout d'abord le vecteur adjoint $p(t)$ comme solution du problème de Cauchy

$$\dot{p}(t) = -p(t)A(t) + x(t)^{\mathrm{T}}W(t),\ p(T) = -x(T)^{\mathrm{T}}Q.$$

La formule de variation de la constante nous conduit à

$$p(t) = \Lambda M(t)^{-1} + \int_0^t x(s)^{\mathrm{T}}W(s)M(s)ds\ M(t)^{-1}$$

pour tout $t \in [0,T]$, où

$$\Lambda = -x(T)^{\mathrm{T}}QM(T) - \int_0^T x(s)^{\mathrm{T}}W(s)M(s)ds.$$

Revenons alors à l'équation (4.9). Tout d'abord, en tenant compte de (4.8) puis en intégrant par parties, il vient

$$\begin{aligned}\int_0^T x(t)^{\mathrm{T}} W(t)\delta x(t)dt &= \int_0^T x(t)^{\mathrm{T}} W(t)M(t)\int_0^t M(s)^{-1}B(s)\delta u(s)ds\ dt \\ &= \int_0^T x(s)^{\mathrm{T}} W(s)M(s)ds \int_0^T M(s)^{-1}B(s)\delta u(s)ds \\ &\quad - \int_0^T \int_0^t x(s)^{\mathrm{T}} W(s)M(s)ds\ M(t)^{-1}B(t)\delta u(t)\ dt.\end{aligned}$$

Or

$$p(t) - \Lambda M(t)^{-1} = \int_0^t x(s)^{\mathrm{T}} W(s)M(s)ds\ M(t)^{-1},$$

et d'après l'expression de Λ on arrive à

$$\int_0^T x(t)^{\mathrm{T}} W(t)\delta x(t)dt = -x(T)^{\mathrm{T}} QM(T)\int_0^T M(t)^{-1}B(t)\delta u(t)dt - \int_0^T p(t)B(t)\delta u(t)dt.$$

Injectons cette égalité dans (4.9), en tenant compte du fait que

$$x(T)^{\mathrm{T}} Q\delta x(T) = x(T)^{\mathrm{T}} QM(T)\int_0^T M(t)^{-1}B(t)\delta u(t)dt.$$

On trouve alors que

$$\frac{1}{2}dC(u).\delta u = \int_0^T (u(t)^{\mathrm{T}} U(t) - p(t)B(t))\delta u(t)\ dt = 0,$$

ceci pour toute application $\delta u \in L^2([0,T],\mathbb{R}^m)$. Ceci implique donc l'égalité pour presque tout $t \in [0,T]$

$$u(t)^{\mathrm{T}} U(t) - p(t)B(t) = 0,$$

ce qui est la conclusion souhaitée. Réciproquement s'il existe un vecteur adjoint $p(t)$ vérifiant (4.5) et (4.6) et si le contrôle u est donné par (4.7), alors il est bien clair d'après le raisonnement précédent que

$$dC(u) = 0.$$

Or C étant strictement convexe ceci implique que u est un minimum global de C. □

Remarque 4.2.1. Si le système de contrôle est perturbé par une fonction $r(t)$, alors le théorème précédent reste vrai. Il le reste, de même, si la fonction g apparaissant dans le coût est une fonction convexe C^1 quelconque de $\mathbb{R}^n$ dans $\mathbb{R}$, sauf que la condition finale sur le vecteur adjoint (4.6) devient

$$p(T) = -\frac{1}{2}\nabla g(x(T)), \tag{4.10}$$

comme on le voit facilement dans la démonstration (en l'absence de convexité, la condition nécessaire reste vraie). Cette condition s'appelle *condition de transversalité*.

Remarque 4.2.2. Dans le cas d'un intervalle infini ($T = +\infty$) la condition devient

$$\lim_{t \to +\infty} p(t) = 0. \tag{4.11}$$

Remarque 4.2.3. Définissons la fonction $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ par

$$H(x, p, u) = p(Ax + Bu) - \frac{1}{2}(x^{\mathrm{T}} W x + u^{\mathrm{T}} U u),$$

en utilisant toujours la convention que p est un vecteur ligne de $\mathbb{R}^n$. Alors les équations données par le principe du maximum LQ s'écrivent

$$\dot{x} = \frac{\partial H}{\partial p} = Ax + Bu,$$
$$\dot{p} = -\frac{\partial H}{\partial x} = -pA + x^{\mathrm{T}} W,$$

et

$$\frac{\partial H}{\partial u} = 0,$$

puisque $pB - u^{\mathrm{T}} U = 0$. Ceci annonce le principe du maximum général. Mais en fait ici dans le cas LQ on peut dire mieux : d'une part le principe du maximum LQ est une condition nécessaire et suffisante de minimalité (alors que dans le cas général c'est une condition nécessaire seulement), d'autre part il est possible d'exprimer le contrôle sous forme de boucle fermée, grâce à la théorie de Riccati (voir section suivante).

Exemple 4.2.1. Considérons, avec $n = m = 1$, le système de contrôle $\dot{x} = u$, $x(0) = x_0$, et le coût

$$C(u) = \int_0^T (x(t)^2 + u(t)^2) dt.$$

Si la trajectoire x associée au contrôle u est optimale alors d'après le théorème précédent on doit avoir

$$\dot{x} = u,\ \dot{p} = x,\ p(T) = 0,$$

avec $u = p$. On en déduit que $\ddot{x} = x$, et donc

$$x(t) = x_0 \mathrm{ch\ t} + p(0) \mathrm{sh\ } t,\ p(t) = x_0 \mathrm{sh\ t} + p(0) \mathrm{ch\ } t.$$

Or $p(T) = 0$, d'où finalement

$$x(t) = x_0 \left(\mathrm{ch\ t} - \frac{\mathrm{sh\ } T}{\mathrm{ch\ } T} \mathrm{sh\ } t \right).$$

Exemple 4.2.2. Considérons le problème du véhicule se déplaçant en ligne droite, modélisé par le système de contrôle

$$\ddot{x} = u,\ x(0) = \dot{x}(0) = 0.$$

On souhaite, pendant un temps T fixé, maximiser la distance parcourue tout en minimisant l'énergie fournie. On choisit donc le critère

$$C(u) = -x(T) + \int_0^T u(t)^2 dt.$$

En appliquant le théorème 4.2.1 on obtient les équations

$$\dot{x} = y,\ \dot{y} = u,\ \dot{p}_x = 0,\ \dot{p}_y = -p_x,$$

et la condition (4.10) donne

$$p_x(T) = \frac{1}{2},\ p_y(T) = 0.$$

En intégrant on trouve le contrôle

$$u(t) = \frac{T-t}{2}$$

et la distance parcourue

$$x(T) = \frac{1}{6}T^3.$$

Remarque 4.2.4. Dans l'exemple précédent on aurait pu mettre des poids différents dans le coût, suivant qu'on accorde plus d'importance à maximiser la distance parcourue ou minimiser l'énergie. On peut aussi choisir le coût

$$C(u) = -x(T)^2 + \int_0^T u(t)^2 dt,$$

qui conduit à $u(t) = x(T)(T-t)$ et $x(T) = \frac{T^3}{3T^3-6}$.

Remarque 4.2.5. L'approche développée dans la démonstration du théorème 4.2.1 est variationnelle. On trouvera une autre approche dans [52], qui permet notamment une extension au cas où on impose que le point final appartienne à une cible. Nous avons ici préféré l'approche du calcul des variations classique, car elle permet une preuve plus rapide et élégante. L'autre approche est en fait plus générale et sera privilégiée dans le cas général (non linéaire) où elle conduit au principe du maximum de Pontryagin général.

4.3 Fonction valeur et équation de Riccati

4.3.1 Définition de la fonction valeur

Soit $T > 0$ fixé, et soit $x \in \mathbb{R}^n$. Considérons le problème LQ de trouver une trajectoire solution de

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x, \tag{4.12}$$

minimisant le coût quadratique

$$C_T(u) = x(T)^{\mathrm{T}} Q x(T) + \int_0^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt. \tag{4.13}$$

Définition 4.3.1. La *fonction valeur* S_T au point x est la borne inférieure des coûts pour le problème LQ. Autrement dit

$$S_T(x) = \inf\{C_T(u) \mid x_u(0) = x\}.$$

Remarque 4.3.1. Sous l'hypothèse (4.3) on a existence d'une unique trajectoire optimale d'après le théorème 4.1.1, et dans ce cas cette borne inférieure est un minimum.

4.3.2 Equation de Riccati

Théorème 4.3.1. *Sous l'hypothèse (4.3), pour tout $x \in \mathbb{R}^n$ il existe une unique trajectoire optimale x associée au contrôle u pour le problème (4.12), (4.13). Le contrôle optimal se met sous forme de boucle fermée*

$$u(t) = U(t)^{-1} B(t)^{\mathrm{T}} E(t) x(t), \tag{4.14}$$

où $E(t) \in \mathcal{M}_n(\mathbb{R})$ est solution sur $[0,T]$ de l'équation matricielle de Riccati

$$\dot{E}(t) = W(t) - A(t)^{\mathrm{T}} E(t) - E(t) A(t) - E(t) B(t) U(t)^{-1} B(t)^{\mathrm{T}} E(t), \; E(T) = -Q. \tag{4.15}$$

De plus, pour tout $t \in [0,T]$, la matrice $E(t)$ est symétrique, et

$$S_T(x) = -x^{\mathrm{T}} E(0) x. \tag{4.16}$$

Remarque 4.3.2. En particulier le théorème affirme que le contrôle optimal u se met sous forme de *boucle fermée*

$$u(t) = K(t) x(t),$$

où $K(t) = U(t)^{-1} B(t)^{\mathrm{T}} E(t)$. Cette forme se prête bien aux problèmes de stabilisation, comme nous le verrons plus loin.

Démonstration. D'après le théorème 4.1.1, il existe une unique trajectoire optimale qui, d'après le théorème 4.2.1, est caractérisée par le système d'équations

$$\begin{aligned} \dot{x} &= Ax + BU^{-1} B^{\mathrm{T}} p^{\mathrm{T}}, \\ \dot{p} &= -pA + x^{\mathrm{T}} W, \end{aligned}$$

avec $x(0) = x$ et $p(T) = -x(T)^{\mathrm{T}} Q$. De plus le contrôle s'écrit

$$u = U^{-1} B^{\mathrm{T}} p^{\mathrm{T}}.$$

Il faut donc montrer que l'on peut écrire $p(t) = x(t)^{\mathrm{T}} E(t)$, où $E(t)$ est solution de (4.15). Notons que si p s'écrit ainsi, alors, d'après l'équation vérifiée par

le couple (x, p), on trouve facilement que $E(t)$ doit être solution de l'équation (4.15). En utilisant l'unicité de la trajectoire optimale, on va maintenant montrer que p s'écrit effectivement ainsi. Soit $E(t)$ solution de l'équation

$$\dot{E} = W - A^{\mathrm{T}}E - EA - EBU^{-1}B^{\mathrm{T}}E,\ E(T) = -Q.$$

Tout d'abord $E(t)$ est symétrique car le second membre de l'équation différentielle l'est, et la matrice Q est symétrique. A priori on ne sait pas cependant que la solution est bien définie sur $[0, T]$ tout entier. On montrera cela plus loin (lemme 4.3.2).

Posons maintenant $p_1(t) = x_1(t)^{\mathrm{T}}E(t)$, où x_1 est solution de

$$\dot{x}_1 = Ax_1 + Bu_1,$$

et $u_1 = U^{-1}B^{\mathrm{T}}Ex_1$. On a alors

$$\begin{aligned}\dot{p}_1 &= \dot{x}_1^{\mathrm{T}}E + {x_1}^{\mathrm{T}}\dot{E}\\ &= (Ax_1 + BU^{-1}B^{\mathrm{T}}Ex_1)^{\mathrm{T}}E + {x_1}^{\mathrm{T}}(W - A^{\mathrm{T}}E - EA - EBU^{-1}B^{\mathrm{T}}E)\\ &= -p_1A + {x_1}^{\mathrm{T}}W.\end{aligned}$$

Autrement dit le triplet (x_1, p_1, u_1) vérifie exactement les équations du théorème 4.2.1. Par conséquent la trajectoire x_1 est optimale, et par unicité il vient $x_1 = x$, $u_1 = u$, puis $p_1 = p$. En particulier on a donc $p = x^{\mathrm{T}}E$, et $u = U^{-1}B^{\mathrm{T}}Ex$. Déduisons-en la formule (4.16). Pour cela calculons d'abord, le long de la trajectoire $x(t)$,

$$\begin{aligned}\frac{d}{dt}x(t)^{\mathrm{T}}E(t)x(t) = \frac{d}{dt}p(t)x(t) &= \dot{p}(t)x(t) + p(t)\dot{x}(t)\\ &= (-p(t)A(t) + x(t)^{\mathrm{T}}W(t))x(t) + p(t)(A(t)x(t) + B(t)u(t))\\ &= x(t)^{\mathrm{T}}W(t)x(t) + p(t)B(t)u(t).\end{aligned}$$

Par ailleurs de l'expression de u on déduit

$$u^{\mathrm{T}}Uu = (U^{-1}B^{\mathrm{T}}Ex)^{\mathrm{T}}UU^{-1}B^{\mathrm{T}}Ex = x^{\mathrm{T}}EBU^{-1}B^{\mathrm{T}}Ex = pBu.$$

Finalement on a l'égalité

$$\frac{d}{dt}x(t)^{\mathrm{T}}E(t)x(t) = x(t)^{\mathrm{T}}W(t)x(t) + u(t)^{\mathrm{T}}U(t)u(t),$$

et par conséquent

$$S_T(x) = x(T)^{\mathrm{T}}Qx(T) + \int_0^T \frac{d}{dt}x(t)^{\mathrm{T}}E(t)x(t)\ dt.$$

Or puisque $E(T) = -Q$ et $x(0) = x$, il vient $S_T(x) = -x^{\mathrm{T}}E(0)x$.

Lemme 4.3.2. *L'application $t \mapsto E(t)$ est bien définie sur $[0, T]$ tout entier.*

Preuve du lemme. Si l'application $E(t)$ n'est pas définie sur $[0,T]$ entier, alors il existe $0 < t_* < T$ tel que $\|E(t)\|$ tend vers $+\infty$ lorsque t tend vers t_* par valeurs supérieures. En particulier pour tout $\alpha > 0$ il existe $t_0 \in]t_*, T]$ et $x_0 \in \mathbb{R}^n$, avec $\|x_0\| = 1$, tels que

$$|{x_0}^{\mathrm{T}} E(t_0) x_0| \geqslant \alpha. \tag{4.17}$$

D'après le théorème 4.1.1, il existe une unique trajectoire optimale $x(\cdot)$ pour le problème LQ sur $[t_0, T]$, telle que $x(t_0) = x_0$ (voir remarque 4.0.3). Cette trajectoire est caractérisée par le système d'équations

$$\begin{aligned} \dot{x} &= Ax + BU^{-1}B^{\mathrm{T}}p^{\mathrm{T}},\ x(t_0) = x_0, \\ \dot{p} &= -pA + x^{\mathrm{T}}W,\ p(T) = -x(T)^{\mathrm{T}}Q. \end{aligned}$$

Le raisonnement précédent, en remplaçant l'intervalle $[0,T]$ par l'intervalle $[t_0,T]$, montre que $S_{T-t_0}(x_0) = -{x_0}^{\mathrm{T}} E(t_0) x_0$. Par ailleurs, $S_{T-t_0}(x_0)$ est inférieur au coût de la trajectoire solution du systéme, partant de x_0, associée (par exemple) au contrôle nul sur l'intervalle $[t_0, T]$; or il est facile de voir que ce coût est majoré, à une constante multiplicative $C > 0$ prés, par $\|x_0\|^2$. On en déduit donc que $|{x_0}^{\mathrm{T}} E(t_0) x_0| \leqslant C\|x_0\|^2$, ce qui contredit (**??**). □

Ceci achève la preuve du théorème. □

Remarque 4.3.3. Il est clair d'après l'expression (4.16) du coût minimal que la matrice $E(0)$ est symétrique négative. On peut améliorer ce résultat si la matrice Q est de plus définie (voir lemme suivant).

Lemme 4.3.3. *Si la matrice Q est symétrique définie positive, ou bien si pour tout $t \in [0,T]$ la matrice $W(t)$ est symétrique définie positive, alors la matrice $E(0)$ est symétrique définie négative.*

Preuve du lemme 4.3.3. Soit x_0 tel que ${x_0}^{\mathrm{T}} E(0) x_0 = 0$, et montrons que $x_0 = 0$. Pour cela on considère le problème LQ

$$\begin{aligned} &\dot{x} = Ax + Bu,\ x(0) = x_0, \\ &\min\ x(T)^{\mathrm{T}} Q x(T) + \int_0^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt, \end{aligned}$$

pour lequel, d'après le théorème 4.3.1, le coût minimal vaut $-{x_0}^{\mathrm{T}} E(0) x_0 = 0$. Par conséquent, puisque pour tout t la matrices $U(t)$ est définie positive, on a $u(t) = 0$ sur $[0,T]$. Si par ailleurs Q est définie positive on a aussi $x(T) = 0$. Donc la trajectoire $x(\cdot)$ est solution du problème de Cauchy $\dot{x} = Ax, x(T) = 0$, et par unicité $x(\cdot)$ est identiquement nulle. En particulier $x(0) = x_0 = 0$, ce qui achève la preuve. Dans le deuxième cas où $W(t)$ est définie positive, la conclusion est immédiate. □

Exercice 4.3.1. Considérons le problème LQ pour le système $\dot{x} = \frac{1}{2}x + u$ $(n = m = 1)$ et le coût

$$C(u) = \int_0^T (2e^{-t}u(t)^2 + \frac{1}{2}e^{-t}x(t)^2)dt.$$

Montrer que l'on obtient les résultats suivants :

$$E(t) = -\frac{1 - e^t e^{-T}}{e^t + e^{2t} e^{-T}},\ u(t) = \frac{1}{2}\frac{1 - e^t e^{-T}}{e^t + e^t e^{-T}} x(t),\ S_T(x) = \frac{1 - e^{-T}}{1 + e^{-T}} x^2.$$

Exercice 4.3.2 (Contrainte finale imposée)**.** Montrer que le problème LQ avec le coût modifié

$$C(u) = \int_0^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt + \lim_{n \to +\infty} n \|x(T)\|^2$$

conduit à une trajectoire minimisante telle que $x(T) = 0$. Montrer que $F(t) = E(t)^{-1}$ existe et est solution sur l'intervalle $[0, T]$ entier d'une équation de Riccati, avec $F(T) = 0$.

Variante du problème précédent. Soit $T > 0$ fixé. Pour tout $t \in [0, T]$ et tout $x \in \mathbb{R}^n$, considérons le problème LQ de trouver une trajectoire solution de

$$\dot{x} = Ax + Bu,\ x(t) = x, \tag{4.18}$$

minimisant le coût quadratique

$$C_T(t, u) = g(x(T)) + \int_t^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2 \right) dt. \tag{4.19}$$

Définition 4.3.2. La *fonction valeur* S au point (t, x) est la borne inférieure des coûts pour ce problème LQ. Autrement dit

$$S_T(t, x) = \inf\{C_T(t, u) \mid x_u(t) = x\}.$$

Théorème 4.3.4. *Sous l'hypothèse (4.3), pour tout $x \in \mathbb{R}^n$ et tout $t \in [0, T]$ il existe une unique trajectoire optimale x associée au contrôle u pour le problème (4.18), (4.19). Le contrôle optimal se met sous forme de boucle fermée*

$$u(s) = U(s)^{-1} B(s)^{\mathrm{T}} E(s) x(s), \tag{4.20}$$

pour tout $s \in [t, T]$, et où $E(s) \in \mathcal{M}_n(\mathbb{R})$ est solution sur $[t, T]$ de l'équation matricielle de Riccati

$$\dot{E} = W - A^{\mathrm{T}} E - EA - EBU^{-1} B^{\mathrm{T}} E,\ E(T) = -Q. \tag{4.21}$$

De plus pour tout $s \in [t, T]$ la matrice $E(s)$ est symétrique, et pour tout $t \in [0, T]$ on a

$$S_T(t, x) = -x^{\mathrm{T}} E(t) x. \tag{4.22}$$

Démonstration. La différence par rapport au cas précédent est que l'on paramétrise le temps initial. Le seul changement est donc la formule (4.22). Comme dans la démonstration précédente, on a

$$S_T(t, x) = x(T)^{\mathrm{T}} Q x(T) + \int_t^T \frac{d}{ds} x(s)^{\mathrm{T}} E(s) x(s)\ ds.$$

Or puisque $E(T) = -Q$ et $x(t) = x$, il vient $S_T(t, x) = -x^{\mathrm{T}} E(t) x$. □

Remarque 4.3.4. L'équation de Riccati étant fondamentale, notamment dans les problèmes de régulateur (voir section suivante), la question de son implémentation numérique se pose naturellement. On peut procéder de manière directe : il s'agit alors, en tenant compte du fait que $E(t)$ est symétrique, d'intégrer un système différentiel non linéaire de $n(n+1)/2$ équations. Dans le paragraphe suivant on donne une alternative à cette méthode. Ci-dessous, nous traitons en *Matlab* un exemple implémentant directement l'équation de Riccati.

Exemple 4.3.1. Considérons le problème LQ pour le système dans $\mathbb{R}^3$

$$\dot{x} = y,\ \dot{y} = z,\ \dot{z} = u,$$

et le coût

$$C_T(u) = \int_0^T (x(t)^2 + y(t)^2 + z(t)^2 + u(t)^2)dt.$$

Notons que pour implémenter l'équation de Riccati (4.21), une condition finale étant donnée, on inverse le temps de façon à se ramener à une condition initiale. Pour rétablir le bon sens du temps, on utilise la fonction *flipud*, cf programme ci-dessous.

```
function riccati1

% Systeme   dx/dt=y, dy/dt=z, dz/dt=u
% min int_0^T (x^2+y^2+z^2+u^2)

clc ; clear all ;
range = [0 : 0.01 : 10 ];

global tricca ricca ;
minit = [ 0 ; 0 ; 0 ; 0 ; 0 ; 0 ] ;
[tricca,ricca] = ode113(@matriccati,range,minit);
ricca=flipud(ricca);    % on remet le temps dans le bon sens

xinit = [ 1 ; 2 ; 3 ] ;
[t,X] = ode113(@systriccati,range,xinit);

plot(t,X(:,1))

%-------------------------------------------------------

function dXdt = systriccati(t,X)

global tricca ricca ;
x=X(1) ; y=X(2) ; z=X(3) ;
[bla,k]=min(abs(tricca-t));

e=ricca(k,5) ; f=ricca(k,6) ; c=ricca(k,3) ;
```

```
u=e*x+f*y+c*z ;  % controle feedback u=U^{-1}B'EX

dXdt= [ y
        z
        u ] ;

%-------------------------------------------------------

function dXdt = matriccati(t,X)

% Eq de Riccati  dE/dt=W-A'E-EA-EBU^{-1}B'E, E(T)=-Q, en temps inverse
a=X(1) ; b=X(2) ; c=X(3) ; d=X(4) ; e=X(5) ; f=X(6) ;

dXdt= - [ 1-e^2
         -2*d-f^2+1
         -2*f-c^2+1
         -a-e*f
         -d-e*c
         -e-b-f*c  ] ;
```

4.3.3 Représentation linéaire de l'équation de Riccati

On a la propriété suivante.

Proposition 4.3.5. *Plaçons-nous dans le cadre du théorème 4.3.1. Soit*

$$R(t) = \begin{pmatrix} R_1(t) & R_2(t) \\ R_3(t) & R_4(t) \end{pmatrix}$$

la résolvante du système linéaire

$$\dot{x} = Ax + BU^{-1}B^{\mathrm{T}}p^{\mathrm{T}},$$
$$\dot{p}^{\mathrm{T}} = -A^{\mathrm{T}}p^{\mathrm{T}} + Wx,$$

telle que $R(T) = Id$. Alors pour tout $t \in [0,T]$ on a

$$E(t) = (R_3(t) - R_4(t)Q)\,(R_1(t) - R_2(t)Q)^{-1}.$$

Démonstration. Par définition de la résolvante on a

$$x(t) = R_1(t)x(T) + R_2(t)p(T)^{\mathrm{T}},$$
$$p(t)^{\mathrm{T}} = R_3(t)x(T) + R_4(t)p(T)^{\mathrm{T}}.$$

Or on sait que $p(T)^{\mathrm{T}} = -Qx(T)$, donc

$$x(t) = (R_1(t) - R_2(t)Q)x(T) \quad \text{et} \quad p(t)^{\mathrm{T}} = (R_3(t) - R_4(t)Q)x(T).$$

On conclut en remarquant que $p(t)^{\mathrm{T}} = E(t)x(t)$. Notons que la matrice $R_1(t) - R_2(t)Q$ est inversible sur $[0,T]$ car le problème LQ est bien posé, comme nous l'avons vu précédemment. □

Par conséquent pour résoudre l'équation de Riccati (4.15), il suffit d'intégrer un système linéaire (il faut calculer une résolvante), ce qui est très facile à programmer. Cette méthode (due à Kalman-Englar) est notamment préférable à la méthode directe dans le cas stationnaire (voir [47]).

4.4 Applications de la théorie LQ

4.4.1 Problèmes de régulation

Le problème du régulateur d'état (ou "problème d'asservissement", ou "problème de poursuite", en anglais "tracking problem")

Considérons le système de contrôle linéaire perturbé

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0, \tag{4.23}$$

et soit $\xi(t)$ une certaine trajectoire de $\mathbb{R}^n$ sur $[0,T]$, partant d'un point ξ_0 (et qui n'est pas forcément solution du système (4.23)). Le but est de déterminer un contrôle tel que la trajectoire associée, solution de (4.23), suive le mieux possible la trajectoire de référence $\xi(t)$ (voir figure 4.1).

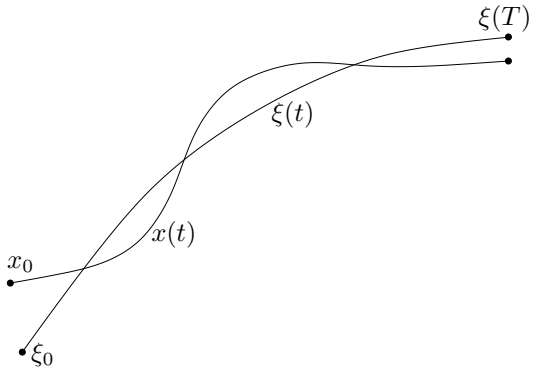


FIGURE 4.1 – Problème du régulateur

On introduit alors l'*erreur* sur $[0,T]$

$$z(t) = x(t) - \xi(t),$$

qui est solution du système de contrôle

$$\dot{z}(t) = A(t)z(t) + B(t)u(t) + r_1(t),\ z(0) = z_0, \tag{4.24}$$

où $z_0 = x_0 - \xi_0$ et $r_1(t) = A(t)\xi(t) - \dot{\xi}(t) + r(t)$. Il est alors raisonnable de vouloir minimiser le coût

$$C(u) = z(T)^{\mathrm{T}} Q z(T) + \int_0^T \left(\|z(t)\|_W^2 + \|u(t)\|_U^2 \right) dt,$$

où Q, W, U sont des matrices de pondération. Pour absorber la perturbation r_1, on augmente le système d'une dimension, en posant

$$z_1 = \begin{pmatrix} z \\ 1 \end{pmatrix},\ A_1 = \begin{pmatrix} A & r_1 \\ 0 & 0 \end{pmatrix},\ B_1 = \begin{pmatrix} B \\ 0 \end{pmatrix},\ Q_1 = \begin{pmatrix} Q & 0 \\ 0 & 0 \end{pmatrix},\ W_1 = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix},$$

de sorte que l'on se ramène à minimiser le coût

$$C(u) = z_1(T)^{\mathrm{T}} Q_1 z_1(T) + \int_0^T \left(\|z_1(t)\|^2_{W_1} + \|u(t)\|^2_U \right) dt,$$

pour le système de contrôle

$$\dot{z}_1 = A_1 z_1 + B_1 u,$$

partant du point $z_1(0)$.

La théorie LQ faite précédemment prévoit alors que le contrôle optimal existe, est unique, et s'écrit

$$u(t) = U(t)^{-1} B_1(t)^{\mathrm{T}} E_1(t) z_1(t),$$

où $E_1(t)$ est solution de l'équation de Riccati

$$\dot{E}_1 = W_1 - A_1{}^{\mathrm{T}} E_1 - E_1 A_1 - E_1 B_1 U^{-1} B_1{}^{\mathrm{T}} E_1,\ E_1(T) = -Q_1.$$

Posons

$$E_1(t) = \begin{pmatrix} E(t) & h(t) \\ h(t)^{\mathrm{T}} & \alpha(t) \end{pmatrix}.$$

En remplaçant dans l'équation précédente, on établit facilement les équations différentielles de E, h, α :

$$\begin{array}{lll} \dot{E} &= W - A^{\mathrm{T}} E - EA - EBU^{-1} B^{\mathrm{T}} E, & E(T) = -Q, \\ \dot{h} &= -A^{\mathrm{T}} h - Er_1 - EBU^{-1} B^{\mathrm{T}} h, & h(T) = 0, \\ \dot{\alpha} &= -2 r_1{}^{\mathrm{T}} h - h^{\mathrm{T}} BU^{-1} B^{\mathrm{T}} h, & \alpha(T) = 0. \end{array} \tag{4.25}$$

Résumons tout ceci dans la proposition suivante.

Proposition 4.4.1. *Soit ξ une trajectoire de $\mathbb{R}^n$ sur $[0, T]$. Considérons le problème de poursuite pour le système de contrôle*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0,$$

où l'on veut minimiser le coût

$$C(u) = (x(T) - \xi(T))^{\mathrm{T}} Q (x(T) - \xi(T)) + \int_0^T \left(\|x(t) - \xi(t)\|^2_W + \|u(t)\|^2_U \right)\ dt.$$

Alors il existe un unique contrôle optimal, qui s'écrit

$$u(t) = U(t)^{-1} B(t)^{\mathrm{T}} E(t)(x(t) - \xi(t)) + U(t)^{-1} B(t)^{\mathrm{T}} h(t),$$

où $E(t) \in \mathcal{M}_n(\mathbb{R})$ *et* $h(t) \in \mathbb{R}^n$ *sont solutions sur* $[0,T]$ *de*

$$\begin{aligned} \dot{E} &= W - A^{\mathrm{T}}E - EA - EBU^{-1}B^{\mathrm{T}}E, & E(T) &= -Q, \\ \dot{h} &= -A^{\mathrm{T}}h - E(A\xi - \dot{\xi} + r) - EBU^{-1}B^{\mathrm{T}}h, & h(T) &= 0, \end{aligned}$$

et de plus $E(t)$ *est symétrique. Par ailleurs le coût minimal est alors égal à*

$$\begin{aligned} &- (x(0) - \xi(0))^{\mathrm{T}} E(0)(x(0) - \xi(0)) - 2h(0)^{\mathrm{T}}(x(0) - \xi(0)) \\ &\quad - \int_0^T \Big(2(A(t)\xi(t) - \dot{\xi}(t) + r(t))^{\mathrm{T}} h(t) + h(t)^{\mathrm{T}} B(t)U(t)^{-1}B(t)^{\mathrm{T}}h(t) \Big)\, dt. \end{aligned}$$

Remarque 4.4.1. Notons que le contrôle optimal s'écrit bien sous forme de boucle fermée

$$u(t) = K(t)(x(t) - \xi(t)) + H(t).$$

Remarque 4.4.2. Si $\dot{\xi} = A\xi + r$, *i.e.* la trajectoire de référence est solution du système sans contrôle, alors dans les notations précédentes on a $r_1 = 0$, et d'après les équations (4.25) on en déduit que $h(t)$ et $\alpha(t)$ sont identiquement nuls. On retrouve alors le cadre LQ de la section précédente. En fait,

- si $\xi = 0$ et $r = 0$, le problème est un problème LQ standard ;
- si $r = 0$, il s'agit d'un problème de poursuite de la trajectoire ξ ;
- si $\xi = 0$, c'est un problème de régulation avec la perturbation r.

Exercice 4.4.1. Résoudre le problème de poursuite sur $[0, \frac{\pi}{2}]$ pour le système $\dot{x} = x + u, x(0) = 0$, la fonction $\xi(t) = t$, et des poids tous égaux à 1.

Exercice 4.4.2. Considérons l'oscillateur harmonique

$$\ddot{x} + x = u,\ x(0) = 0, \dot{x}(0) = 1.$$

On désire asservir le mouvement de cet oscillateur à la courbe $(\cos t, \sin t)$ sur $[0, 2\pi]$, *i.e.* décaler la phase de $\pi/2$. Ecrire les équations permettant de résoudre le problème, puis réaliser l'implémentation numérique.

Variante : le problème de poursuite d'une sortie (ou "output tracking")

On ajoute au problème précédent une variable de sortie :

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0, \\ &y(t) = C(t)x(t), \end{aligned}$$

et étant donné un signal de référence $\xi(t)$ on cherche un contrôle tel que, le long de la trajectoire associée, l'observable $z(\cdot)$ soit proche de $\xi(\cdot)$. Notons qu'on retrouve le cas précédent si $y(t) = x(t)$.

Posant $z(t) = y(t) - \xi(t)$, on cherche à minimiser le coût

$$C(u) = z(T)^{\mathrm{T}} Q z(T) + \int_0^T \left(\|z(t)\|_W^2 + \|u(t)\|_U^2 \right) dt.$$

Posons alors

$$x_1 = \begin{pmatrix} x \\ 1 \end{pmatrix},\ Q_1 = \begin{pmatrix} C(T)^{\mathrm{T}} Q C(T) & -C(T)^{\mathrm{T}} Q \xi(T) \\ -\xi^{\mathrm{T}}(T) Q C(T) & \xi^{\mathrm{T}}(T) Q \xi(T) \end{pmatrix},\ W_1 = \begin{pmatrix} C^{\mathrm{T}} W C & -C^{\mathrm{T}} W \xi \\ -\xi^{\mathrm{T}} W C & \xi^{\mathrm{T}} W \xi \end{pmatrix},$$

et A_1, B_1 comme précédemment (avec $r_1 = r$). Alors on cherche un contrôle u, associé à la trajectoire x_1 solution de $\dot{x}_1 = A_1 x_1 + B_1 u$, minimisant le coût

$$C(u) = x_1(T)^{\mathrm{T}} Q_1 x_1(T) + \int_0^T \left(\|x_1(t)\|^2_{W_1} + \|u(t)\|^2_U \right) dt.$$

En raisonnant comme précédemment, on arrive au résultat suivant.

Proposition 4.4.2. *Soit ξ une trajectoire de $\mathbb{R}^p$ sur $[0,T]$. Considérons le problème de poursuite de la sortie r pour le système de contrôle avec sortie*

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0, \\ &y(t) = C(t)x(t), \end{aligned}$$

où l'on veut minimiser le coût

$$C(u) = (y(T) - \xi(T))^{\mathrm{T}} Q (y(T) - \xi(T)) + \int_0^T \left(\|y(t) - \xi(t)\|^2_W + \|u(t)\|^2_U \right)\ dt.$$

Alors il existe un unique contrôle optimal, qui s'écrit

$$u(t) = U(t)^{-1} B(t)^{\mathrm{T}} E(t) x(t) + U(t)^{-1} B(t)^{\mathrm{T}} h(t),$$

où $E(t) \in \mathcal{M}_n(\mathbb{R})$ et $h(t) \in \mathbb{R}^p$ sont solutions sur $[0,T]$ de

$$\begin{aligned} \dot{E} &= C^{\mathrm{T}} W C - A^{\mathrm{T}} E - E A - E B U^{-1} B^{\mathrm{T}} E, \quad E(T) = -C(T)^{\mathrm{T}} Q C(T), \\ \dot{h} &= -C^{\mathrm{T}} W \xi - A^{\mathrm{T}} h - E r - E B U^{-1} B^{\mathrm{T}} h, \quad h(T) = -C(T)^{\mathrm{T}} Q \xi(T), \end{aligned}$$

et de plus $E(t)$ est symétrique. Par ailleurs le coût minimal est alors égal à

$$-x(0)^{\mathrm{T}} E(0) x(0) - 2h(0)^{\mathrm{T}} x(0) - \alpha(0),$$

où $\alpha(t)$ est solution de

$$\dot{\alpha} = \xi^{\mathrm{T}} W \xi - 2 r^{\mathrm{T}} h - h^{\mathrm{T}} B U^{-1} B^{\mathrm{T}} h,\ \alpha(T) = \xi(T)^{\mathrm{T}} Q \xi(T).$$

Remarque 4.4.3. On trouvera dans [64] d'autres variantes de ce problème, notamment le même problème que ci-dessus, sauf que le coût s'écrit

$$C(u) = x(T)^{\mathrm{T}} Q x(T) + \int_0^T \left(\|y(t) - \xi(t)\|^2_W + \|u(t)\|^2_U \right)\ dt.$$

Le seul changement est dans la matrice augmentée Q_1, et donc dans les conditions aux limites de E et h, qui deviennent dans ce cas $E(T) = -Q$ et $h(T) = 0$.

On trouvera aussi dans [52, ex. 9, p. 203] une autre variante du problème LQ, où la fonction g apparaissant dans le coût est linéaire en x. Nous laissons l'écriture de toutes ces variantes au lecteur, la méthode étant de toute façon la même que précédemment.

Exercice 4.4.3. On considère le système proies-prédateurs contrôlé

$$\dot{x} = x + y + u_1,\ x(0) = 1,$$
$$\dot{y} = x - y + u_2,\ y(0) = 1.$$

Trouver l'expression des contrôles permettant d'asservir la variable $x(t)$ à la valeur 1 sur l'intervalle $[0, 10]$.

4.4.2 Filtre de Kalman déterministe

Ce problème célèbre est le suivant. Connaissant un signal de référence $\xi(t)$ sur $[0, T]$, on cherche une trajectoire solution sur $[0, T]$ de

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),$$

minimisant le coût

$$C(u) = x(0)^{\mathrm{T}} Q x(0) + \int_0^T \left(\|(C(t)x(t) - \xi(t))\|_W^2 + \|u(t)\|_U^2 \right) dt.$$

Il s'agit d'une variante des problèmes de poursuite précédents, sauf que l'on n'impose aucune condition sur $x(0)$ et $x(T)$, et de plus le coût pénalise le point initial $x(0)$. En revanche dans ce problème on suppose que la matrice Q est symétrique *définie* positive.

Pour se ramener aux cas précédents, il convient donc tout d'abord d'inverser le temps, de façon à ce que le coût pénalise, comme avant, le point final. On pose donc, pour tout $t \in [0, T]$,

$$\tilde{x}(t) = x(T-t),\ \tilde{u}(t) = u(T-t),\ \tilde{A}(t) = -A(T-t),\ \tilde{B}(t) = -B(T-t),$$
$$\tilde{\xi}(t) = \xi(T-t),\ \tilde{W}(t) = W(T-t),\ \tilde{U}(t) = U(T-t),\ \tilde{C}(t) = C(T-t),$$

de sorte que l'on se ramène au problème de déterminer une trajectoire solution de $\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}\tilde{u}$, minimisant le coût

$$\tilde{C}(\tilde{u}) = \tilde{x}(T)^{\mathrm{T}} Q \tilde{x}(T) + \int_0^T \left(\|(\tilde{C}(t)\tilde{x}(t) - \tilde{\xi}(t))\|_{\tilde{W}}^2 + \|\tilde{u}(t)\|_{\tilde{U}}^2 \right) dt.$$

Notons que, par construction, on a $\tilde{C}(\tilde{u}) = C(u)$.

Fixons une donnée initiale $\tilde{x}(0)$, et appliquons, pour cette donnée initiale, le même raisonnement que dans les cas précédents. On obtient alors

$$\tilde{u}(t) = \tilde{U}^{-1}\tilde{B}^{\mathrm{T}}\tilde{E}\tilde{x} + \tilde{U}^{-1}\tilde{B}^{\mathrm{T}}\tilde{h},$$

où

$$\begin{array}{rcll} \dot{\tilde{E}} & = & \tilde{C}^{\mathrm{T}}\tilde{W}\tilde{C} - \tilde{A}^{\mathrm{T}}\tilde{E} - \tilde{E}\tilde{A} - \tilde{E}\tilde{B}\tilde{U}^{-1}\tilde{B}^{\mathrm{T}}\tilde{E}, & \tilde{E}(T) = -Q, \\ \dot{\tilde{h}} & = & -\tilde{C}^{\mathrm{T}}\tilde{W}\tilde{\xi} - \tilde{A}^{\mathrm{T}}\tilde{h} - \tilde{E}\tilde{B}\tilde{U}^{-1}\tilde{B}^{\mathrm{T}}\tilde{h}, & \tilde{h}(T) = 0, \\ \dot{\tilde{\alpha}} & = & \tilde{\xi}^{\mathrm{T}}\tilde{W}\tilde{\xi} - \tilde{h}^{\mathrm{T}}\tilde{B}\tilde{U}^{-1}\tilde{B}^{\mathrm{T}}\tilde{h}, & \alpha(T) = 0, \end{array}$$

et le coût minimal pour cette donnée initiale fixée $\tilde{x}(0)$ vaut

$$-\tilde{x}(0)^{\mathrm{T}}\tilde{E}(0)\tilde{x}(0) - 2\tilde{x}(0)^{\mathrm{T}}\tilde{h}(0) - \tilde{\alpha}(0).$$

Il faut maintenant trouver $\tilde{x}(0)$ tel que ce coût soit minimal. Posons donc

$$f(x) = -x^{\mathrm{T}}\tilde{E}(0)x - 2x^{\mathrm{T}}\tilde{h}(0) - \alpha(0).$$

Il faut donc déterminer un minimum de f. Notons tout d'abord que, la matrice Q étant par hypothèse définie positive, la matrice $\tilde{E}(0)$ est d'après le lemme 4.3.3 symétrique définie négative. En particulier la fonction f est strictement convexe et de ce fait admet un unique minimum. En un tel point on doit avoir $f'(x) = 0$, d'où $x = -\tilde{E}(0)^{-1}\tilde{h}(0)$.

Finalement, en reprenant le cours positif du temps, et en posant pour tout $t \in [0, T]$

$$E(t) = -\tilde{E}(T-t),\ h(t) = -\tilde{h}(T-t),$$

on arrive au résultat suivant.

Proposition 4.4.3. *Soit $\xi(\cdot)$ une trajectoire définie sur $[0, T]$ à valeurs dans $\mathbb{R}^p$. On considère le problème de déterminer une trajectoire solution sur $[0, T]$ de*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),$$

minimisant le coût

$$C(u) = x(0)^{\mathrm{T}}Qx(0) + \int_0^T \left(\|(C(t)x(t) - \xi(t))\|_W^2 + \|u(t)\|_U^2\right) dt,$$

où la matrice Q est de plus supposée définie positive. Alors il existe une unique trajectoire minimisante, associée au contrôle

$$u(t) = U(t)^{-1}B(t)^{\mathrm{T}}E(t)x(t) + U(t)^{-1}B(t)^{\mathrm{T}}h(t),$$

et à la condition finale

$$x(T) = -E(T)^{-1}h(T),$$

où

$$\begin{array}{rcll} \dot{E} & = & C^{\mathrm{T}}WC - A^{\mathrm{T}}E - EA - EBU^{-1}B^{\mathrm{T}}E, & E(0) = Q, \\ \dot{h} & = & -C^{\mathrm{T}}W\xi - A^{\mathrm{T}}h - EBU^{-1}B^{\mathrm{T}}h, & h(0) = 0, \end{array}$$

et le coût minimal vaut alors

$$-h(T)^{\mathrm{T}}E(T)^{-1}h(T) + \int_0^T \Big(\xi(t)^{\mathrm{T}}W(t)\xi(t) - h(t)^{\mathrm{T}}B(t)U(t)^{-1}B(t)^{\mathrm{T}}h(t)\Big)\, dt.$$

L'état final $x(T) = -E(T)^{-1}h(T)$ est la donnée qui nous intéresse principalement dans le problème du filtre de Kalman, qui est un problème d'estimation, comme nous le verrons dans les exemples à suivre. L'estimation de cet état final peut être simplifiée de la manière suivante.

Posons $F(t) = E(t)^{-1}$. On trouve facilement, puisque $\dot{F} = -F\dot{E}F$,

$$\dot{F} = BU^{-1}B^{\mathrm{T}} + AF + FA^{\mathrm{T}} - FC^{\mathrm{T}}WCF,\ F(0) = Q^{-1}.$$

Par ailleurs si on pose $z(t) = -F(t)h(t)$, on trouve que

$$\dot{z} = (A - FC^{\mathrm{T}}WC)z + FC^{\mathrm{T}}W\xi,\ z(0) = 0.$$

Finalement on arrive au résultat suivant.

Proposition 4.4.4. *Sous les hypothèses de la proposition 4.4.3, l'état final $x(T)$ de la solution optimale est égal à $z(T)$, où*

$$\begin{array}{rcll} \dot{z} & = & (A - FC^{\mathrm{T}}WC)z + FC^{\mathrm{T}}W\xi, & z(0) = 0, \\ \dot{F} & = & BU^{-1}B^{\mathrm{T}} + AF + FA^{\mathrm{T}} - FC^{\mathrm{T}}WCF, & F(0) = Q^{-1}. \end{array}$$

Application au filtrage. Le problème est d'estimer, d'après une observation, un signal bruité. Le modèle est

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \\ &y(t) = C(t)x(t) + v(t), \end{aligned}$$

où les fonctions u et v sont des *bruits*, *i.e.* des perturbations affectant le système. La donnée initiale x_0 est inconnue. Le signal $\xi(t)$ représente une observation de la variable $y(t)$, et à partir de cette observation on veut construire une estimation de l'état final $x(T)$. On cherche une estimation optimale dans le sens que les perturbations u et v, ainsi que la donnée initiale x_0, doivent être aussi petites que possible. On cherche donc à minimiser un coût de la forme

$$x(0)^{\mathrm{T}}Qx(0) + \int_0^T (\|w(t)\|_W^2 + \|u(t)\|_U^2)dt.$$

Il s'agit donc exactement du problème LQ

$$\begin{aligned} &\dot{x}(t) = A(t)x(t) + B(t)u(t), \\ &y(t) = C(t)x(t), \\ &C(u) = x(0)^{\mathrm{T}}Qx(0) + \int_0^T (\|y(t) - \xi(t)\|_W^2 + \|u(t)\|_U^2)dt, \end{aligned}$$

i.e. le problème que l'on vient d'étudier ($x(0)$ non fixé).

L'estimation optimale de l'état est donc égale à $z(T)$ (voir proposition 4.4.4).

Remarque 4.4.4. La bonne manière d'interpréter le filtre de Kalman est statistique, ce qui dépasse le cadre de cet ouvrage. En fait il faut interpréter les perturbations u et b comme des bruits blancs gaussiens, et x_0 comme une variable aléatoire gaussienne, tous supposés centrés en 0 (pour simplifier). Les matrices $Q, W(t), U(t)$ sont alors les matrices de variance de $x_0, v(t), u(t)$, et le problème de minimisation s'interprète comme le problème d'estimer l'état

final de variance minimale, connaissant l'observation $\xi(t)$ (à ce sujet, voir par exemple [3]).

Par ailleurs les pondérations doivent être choisies en fonction de l'importance des bruits. Par exemple si le bruit v est très important comparé au bruit u et à l'incertitude sur la condition initiale alors on choisit une matrice $W(t)$ petite.

Exemple 4.4.1. On veut estimer $x(T)$ pour le système bruité

$$\dot{x} = u,\ y = x + v,$$

d'après l'observation $\xi(t)$.

Les équations de la proposition 4.4.4 donnent

$$\dot{z} = -FWz + FW\xi,\ z(0) = 0,$$
$$\dot{F} = U^{-1} - FWF,\ F(0) = Q^{-1}.$$

Choisissons les poids $Q = 1, U(t) = 1, W(t) = w^2$. On trouve

$$F(t) = \frac{1}{w} + \frac{e^{-wt}(-1+w)}{w}.$$

En particulier si le bruit v est petit alors on peut choisir le paramètre w très grand, de sorte que pour tout $t > 0$ on a $F(t) \simeq 1$. On montre alors facilement, avec l'équation de z, que $z(t) \simeq \xi(t)$, ce qui est bien cohérent : en effet s'il n'y a pas de bruit alors on observe directement l'état que l'on cherche à estimer !

Dans le cas général, on calcule (numériquement) $z(T)$, ce qui fournit l'estimation de $x(T)$ souhaitée.

4.4.3 Régulation sur un intervalle infini et rapport avec la stabilisation

Considérons le problème LQ sur l'intervalle $[0, +\infty[$. Il s'agit d'un problème de régulation où l'on cherche à rendre l'erreur petite pour tout temps. Nous nous restreignons au cas de systèmes stationnaires. Le cadre est le suivant.

On cherche à déterminer une trajectoire solution de

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = x_0,$$

minimisant le coût

$$C(u) = \int_0^\infty \left(\|x(t)\|_W^2 + |u(t)\|_U^2 \right) dt,$$

où de même les matrices W et U sont constantes.

On a la résultat suivant.

Théorème 4.4.5. *On suppose que les matrices W et U sont symétriques définies positives, et que le système est contrôlable. Alors il existe une unique trajectoire minimisante pour ce problème, associée sur $[0,+\infty[$ au contrôle optimal*

$$u(t) = U^{-1}B^{\mathrm{T}}Ex(t), \tag{4.26}$$

où $E \in \mathcal{M}_n(\mathbb{R})$ est l'unique matrice symétrique définie négative solution de l'équation de Riccati stationnaire

$$A^{\mathrm{T}}E + EA + EBU^{-1}B^{\mathrm{T}}E = W. \tag{4.27}$$

De plus le coût minimal vaut $-{x_0}^{\mathrm{T}}Ex_0$.

Par ailleurs le système bouclé

$$\dot{x} = (A + BU^{-1}B^{\mathrm{T}}E)x$$

est globalement asymptotiquement stable, et la fonction $V(x) = -x^{\mathrm{T}}Ex$ est une fonction de Lyapunov stricte pour ce système.

Remarque 4.4.5. En particulier, la trajectoire minimisante associée à ce problème en horizon infini tend vers 0 lorsque t tend vers l'infini.

Démonstration. On sait déjà (voir proposition 4.1.3 et remarque 4.2.2) qu'il existe une unique trajectoire optimale, vérifiant les équations

$$\dot{x} = Ax + Bu,\ \dot{p} = -pA + {x}^{\mathrm{T}}W,\ \lim_{t\to+\infty} p(t) = 0,$$

avec $u = U^{-1}B^{\mathrm{T}}{p}^{\mathrm{T}}$. De manière tout à fait similaire à la preuve du théorème 4.2.1 on montre, par un argument d'unicité, que $p(t) = {x(t)}^{\mathrm{T}}E$, où E est solution, pourvu qu'elle existe, de l'équation (4.27). Il faut donc montrer l'existence d'une telle solution. C'est l'objet du lemme suivant.

Lemme 4.4.6. *Il existe une unique matrice E symétrique définie négative solution de l'équation (4.27).*

Preuve du lemme. Il est bien clair que si $x(\cdot)$ est minimisante pour le problème LQ sur $[0,+\infty[$, alors elle l'est aussi sur chaque intervalle $[0,T]$, $T>0$. Considérons donc le problème LQ sur $[0,T]$

$$\dot{x} = Ax + Bu,\ x(0) = x_0,$$
$$C(T,u) = \int_0^T \left(\|x(t)\|_W^2 + \|u(t)\|_U^2\right) dt,$$

et appelons $E(T,t)$ la solution de l'équation de Riccati associée

$$\dot{E} = W - A^{\mathrm{T}}E - EA - EBU^{-1}B^{\mathrm{T}}E,\ E(T,T) = 0.$$

On sait que de plus le coût minimal est $C(T,u) = -{x_0}^{\mathrm{T}}E(T,0){x_0}^{\mathrm{T}}$. Posons alors $D(T,t) = -E(T,T-t)$. Il est bien clair que

$$\dot{D} = W + A^{\mathrm{T}}D + DA - DBU^{-1}B^{\mathrm{T}}D,\ D(T,0) = 0.$$

Cette équation étant en fait indépendante de T, on peut poser $D(t) = D(T,t)$, et $D(t)$ est solution de l'équation de Riccati ci-dessus sur $\mathbb{R}^+$. De plus pour tout $T > 0$ on a $D(T) = -E(T,0)$, et comme la matrice W est symétrique définie positive on déduit du lemme 4.3.3 que $D(T)$ est symétrique définie positive.

Par ailleurs on a, pour tout $T > 0$, $C(T,u) = {x_0}^{\mathrm{T}} D(T) x_0$. Il est clair que si $0 < t_1 \leqslant t_2$ alors $C(t_1,u) \leqslant C(t_2,u)$, et donc ${x_0}^{\mathrm{T}} D(t_1) x_0 \leqslant {x_0}^{\mathrm{T}} D(t_2) x_0$. Ceci est en fait indépendant de x_0, car l'équation de Riccati ne dépend nullement de la donnée initiale. Ainsi pour tout $x \in \mathbb{R}^n$ la fonction $t \mapsto x^{\mathrm{T}} D(t) x$ est croissante.

Montrons qu'elle est également majorée. Le système étant contrôlable, l'argument de la remarque 4.1.2 montre qu'il existe au moins un contrôle v sur $[0, +\infty[$ de coût fini. Comme le contrôle u est optimal, on en déduit que la fonction $t \mapsto C(t,u)$ est majorée (par $C(v)$).

Pour tout $x \in \mathbb{R}^n$, la fonction $t \mapsto x^{\mathrm{T}} D(t) x$ étant croissante et majorée, on en déduit qu'elle converge. En appliquant cette conclusion aux éléments d'une base (e_i) de $\mathbb{R}^n$, on en déduit que chaque élément $d_{ij}(t)$ de la matrice $D(t)$ converge, car en effet

$$d_{ij}(t) = {e_i}^{\mathrm{T}} D(t) e_j = \frac{1}{2} e_i + {e_j}^{\mathrm{T}} D(t)(e_i + e_j) - {e_i}^{\mathrm{T}} D(t) e_i - {e_j}^{\mathrm{T}} D(t) e_j.$$

Ainsi la matrice $D(t)$ converge vers une matrice $-E$, qui est nécessairement symétrique définie négative d'après la croissance de la fonction $t \mapsto x^{\mathrm{T}} D(t) x$.

Par ailleurs, de l'équation différentielle vérifiée par D, on déduit que $\dot{D}(t)$ converge, et cette limite est alors nécessairement nulle. En passant à la limite dans cette équation différentielle on obtient finalement l'équation de Riccati stationnaire (4.27).

Enfin, en passant à la limite on a $C(u) = -{x_0}^{\mathrm{T}} E x_0$, d'où on déduit aisément l'unicité de la solution. □

Pour montrer la deuxième partie du théorème, il suffit de voir que la fonction $V(x) = -x^{\mathrm{T}} E x$ est une fonction de Lyapunov pour le système bouclé $\dot{x} = (A + BU^{-1}B^{\mathrm{T}}E)x$. La forme quadratique V est bien définie positive puisque E est symétrique définie négative. Par ailleurs on calcule facilement le long d'une trajectoire $x(t)$ solution du système bouclé

$$\frac{d}{dt} V(x(t)) = -x(t)^{\mathrm{T}} \left(W + EBU^{-1}B^{\mathrm{T}}E \right) x(t).$$

Or la matrice W est par hypothèse définie positive, et la matrice $EBU^{-1}B^{\mathrm{T}}E$ est positive, donc cette quantité est strictement négative si $x(t) \neq 0$. On a donc bien une fonction de Lyapunov stricte, ce qui prouve que le système bouclé est asymptotiquement stable. □

Remarque 4.4.6. Le contrôle optimal s'écrit sous forme de boucle fermée $u = Kx$, avec $K = U^{-1}B^{\mathrm{T}}E$. On retrouve le fait que si le système est contrôlable alors il est stabilisable par feedback linéaire (voir le théorème 13.1.5 de placement de pôles). Cependant, alors que la méthode de stabilisation décrite par le

théorème 13.1.5 consiste à réaliser un placement de pôles, ici la matrice K est choisie de manière à minimiser un certain critère. On parle de stabilisation par retour d'état optimal. C'est donc une méthode (parmi beaucoup d'autres) de stabilisation.

Remarque 4.4.7. En général l'équation (4.27) admet plusieurs solutions, mais elle n'admet qu'une seule solution symétrique définie négative.

Exemple 4.4.2. Considérons le système scalaire $\dot{x} = -x + u, x(0) = x_0$ et le coût $C(u) = \int_0^\infty (x(t)^2 + u(t)^2)dt$. L'équation de Riccati stationnaire est $-2E + E^2 = 1$, et conduit à $E = 1 - \sqrt{2} < 0$, d'où la trajectoire optimale

$$u(t) = (1 - \sqrt{2})x(t),\ x(t) = x_0 e^{-\sqrt{2}t}.$$

Exemple 4.4.3. On considère le système contrôlé

$$\begin{aligned} \dot{x} &= x + y + u_1,\ x(0) = 1, \\ \dot{y} &= x - y + u_2,\ y(0) = 1. \end{aligned}$$

On désire stabiliser la solution de ce système vers l'origine, en minimisant le coût

$$C(u) = \int_0^{+\infty} (x(t)^2 + y(t)^2 + u_1(t)^2 + u_2(t)^2)dt.$$

Pour cela, écrivons l'équation de Riccati stationnaire, avec les matrices

$$A = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},\ B = U = W = Id.$$

En posant

$$E = \begin{pmatrix} a & c \\ c & b \end{pmatrix},$$

on arrive au système d'équations

$$\begin{aligned} 2a + 2c + a^2 + c^2 &= 1, \\ 2c - 2b + c^2 + b^2 &= 1, \\ a + b + ac + cb &= 0. \end{aligned}$$

En particulier la troisième équation conduit à

$$(a + b)(1 + c) = 0,$$

et par conséquent $a = -b$ ou $c = -1$. Si $a = -b$, les valeurs propres de la matrice E sont alors $\pm\sqrt{a^2 + c^2}$, ce qui est exclu puisque la matrice E doit être définie négative. Par conséquent $c = -1$, et on trouve alors

$$a = -1 \pm \sqrt{3},\ b = 1 \pm \sqrt{3}.$$

Parmi ces 4 possibilités, la seule façon d'obtenir une matrice E définie négative est de prendre $a = -1 - \sqrt{3}$ et $b = 1 - \sqrt{3}$. Donc finalement

$$E = \begin{pmatrix} -1-\sqrt{3} & -1 \\ -1 & 1-\sqrt{3} \end{pmatrix},$$

et le système bouclé est alors

$$\dot{x} = -\sqrt{3}x,\ x(0) = 1,$$
$$\dot{y} = -\sqrt{3}y,\ y(0) = 1.$$

Exercice 4.4.4. On considère le système contrôlé :

$$\ddot{x} + x = u,\ x(0) = 0,\ \dot{x}(0) = 1.$$

1. Quel est le comportement de la solution en l'absence de contrôle ?
2. On désire stabiliser la solution de ce système vers l'origine par la méthode de Riccati stationnaire, en minimisant le coût

$$C(u) = \int_0^{+\infty} (x(t)^2 + \dot{x}(t)^2 + u(t)^2)dt.$$

   (a) Montrer que la solution de l'équation de Riccati stationnaire est

$$E = \begin{pmatrix} -\alpha\sqrt{2} & 1-\sqrt{2} \\ 1-\sqrt{2} & -\alpha \end{pmatrix}.$$

   où $\alpha = \sqrt{2\sqrt{2}-1}$.

   (b) Donner l'expression du contrôle optimal.

   (c) Montrer que la solution du système bouclé est

$$x(t) = \frac{2}{\beta} e^{-\frac{\alpha}{2}t} \sin\frac{\beta}{2}t,$$

   où $\beta = \sqrt{2\sqrt{2}+1}$.

   (d) Commenter brièvement les résultats et la méthode.

Exercice 4.4.5. Montrer que la solution de l'équation de Riccati stationnaire pour le problème LQ

$$\dot{x} = y,\ \dot{y} = u,\ \ C(u) = \int_0^{+\infty} x(t)^2 + y(t)^2 + u(t)^2\ dt,$$

est la matrice

$$E = \begin{pmatrix} -\sqrt{3} & -1 \\ -1 & -\sqrt{3} \end{pmatrix}.$$

Exercice 4.4.6. Résoudre le problème LQ

$$\dot{x} = y + u_1,\ \dot{y} = u_2,\ \ \min \int_0^{+\infty} x(t)^2 + y(t)^2 + u_1(t)^2 + u_2(t)^2\ dt.$$

Exercice 4.4.7. Déterminer la solution de $\dot{x} = -x + u$ minimisant le coût $\int_0^{\infty}(x(t)^2 + \alpha u(t)^2)dt$, avec $\alpha > 0$. Que se passe-t-il lorsque $\alpha \to +\infty$?

Solution numérique de l'équation de Riccati stationnaire On peut calculer numériquement la solution de l'équation de Riccati algébrique (4.27) en employant une méthode de Schur (voir [50, 51]). Ceci est implémenté en *Matlab* dans la fonction *lqr.m* (voir aussi *care.m*).

Ci-dessous, voici un exemple d'utilisation de lqr, en reprenant l'exemple 4.3.1.

```
function riccati2

% Systeme   dx/dt=y, dy/dt=z, dz/dt=u
% min int_0^T (x^2+y^2+z^2+u^2)

clc ; clear all ;

global A B W invU ;
% Systeme
A = [ 0 1 0
      0 0 1
      0 0 0 ] ;
B = [ 0
      0
      1 ] ;

% Matrices de ponderation
W = eye(3) ;
U = 1 ; invU = inv(U) ;

range = [0 : 0.01 : 10 ];

%% Utilisation de lqr

global K ;
[K,S,e] = lqr(A,B,W,U) ;
xinit = [ 1 ; 2 ; 3 ] ;
[t,X] = ode45(@systriccati,range,xinit) ;
plot(t,X(:,1));

%-------------------------------------------------------

function dXdt = systriccati(t,X)

global K ; u = -K*X ;

dXdt = [ X(2)
         X(3)
          u   ] ;
```

Le résultat est tracé sur la figure 4.2.

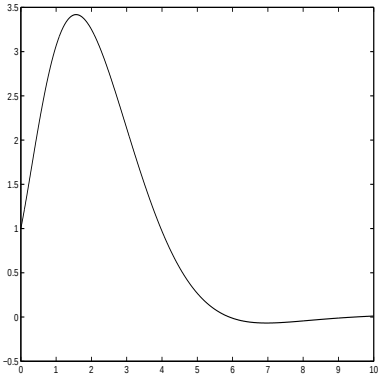


FIGURE 4.2 –

Deuxième partie

Théorie du contrôle optimal non linéaire

L'objectif de cette partie est de présenter des techniques d'analyse de problèmes de contrôle optimal non linéaires. On présente notamment le principe du maximum de Pontryagin et la théorie d'Hamilton-Jacobi. Un chapitre est consacré aux méthodes numériques en contrôle optimal.

D'un point de vue global, un problème de contrôle optimal se formule sur une variété M, mais notre point de vue est *local* et on travaille sur un ouvert V petit de $\mathbb{R}^n$. La problématique générale du contrôle optimal est la suivante. Considérons un système de contrôle général

$$\dot{x}(t) = f(t, x(t), u(t)),\ x(t_0) = x_0, \tag{4.28}$$

où f est une application de classe C^1 de $I \times V \times U$ dans $\mathbb{R}^n$, I est un intervalle de $\mathbb{R}$, V ouvert de $\mathbb{R}^n$, U un ouvert de $\mathbb{R}^m$, $(t_0, x_0) \in I \times V$. Par ailleurs on suppose que les contrôles $u(\cdot)$ appartiennent à un sous-ensemble de $L^\infty_{loc}(I, \mathbb{R}^m)$.

Ces hypothèses assurent, pour tout contrôle u, l'existence et l'unicité sur d'une solution maximale $x_u(t)$ sur un intervalle $J \subset I$, du problème de Cauchy (4.28) (voir section 11.3 en annexe).

Par commodité d'écriture on suppose dans toute la suite que $t_0 = 0$.

Pour tout contrôle $u \in L^\infty_{loc}(I, \mathbb{R}^m)$, la trajectoire associée $x_u(\cdot)$ est définie sur un intervalle maximal $[0, t_e(u)[$, où $t_e(u) \in \mathbb{R}^+ \cup \{+\infty\}$. Par exemple si $t_e(u) < +\infty$ alors la trajectoire explose en $t_e(u)$ (théorème d'échappement, ou d'explosion). Pour tout $T > 0$, $T \in I$, on note $\mathcal{U}_T$ l'ensemble des contrôles admissibles sur $[0, T]$, c'est-à-dire l'ensemble des contrôles tels que la trajectoire associée soit bien définie sur $[0, T]$, autrement dit $T < t_e(u)$.

Soient f^0 une fonction de classe C^1 sur $I \times V \times U$, et g une fonction continue sur V. Pour tout contrôle $u \in \mathcal{U}_T$ on définit le coût de la trajectoire associée $x_u(\cdot)$ sur l'intervalle $[0, T]$

$$C(T, u) = \int_0^T f^0(t, x_u(t), u(t))dt + g(T, x_u(T)). \tag{4.29}$$

Soient M_0 et M_1 deux sous-ensembles de V. Le problème de contrôle optimal est de déterminer les trajectoires $x_u(\cdot)$ solutions de

$$\dot{x}_u(t) = f(t, x_u(t), u(t)),$$

telles que $x_u(0) \in M_0$, $x_u(T) \in M_1$, et minimisant le coût $C(T, u)$. On dit que le problème de contrôle optimal est *à temps final non fixé* si le temps final T est libre, sinon on parle de problème *à temps final fixé*.

Chapitre 5

Définitions et préliminaires

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 le font en *coût minimal*.

Dans ce chapitre nous étudions le problème de contrôlabilité et rappelons quelques faits.

5.1 Application entrée-sortie

5.1.1 Définition

Considérons pour le système (4.28) le problème de *contrôle* suivant : étant donné un point $x_1 \in \mathbb{R}^n$, trouver un temps T et un contrôle u sur $[0, T]$ tel que la trajectoire x_u associée à u, solution de (4.28), vérifie

$$x_u(0) = x_0 \ , \ \ x_u(T) = x_1.$$

Ceci conduit à la définition suivante.

Définition 5.1.1. Soit $T > 0$. *L'application entrée-sortie* en temps T du système contrôlé (4.28) initialisé à x_0 est l'application

$$\begin{array}{rccc} E_T : \mathcal{U} & \longrightarrow & \mathbb{R}^n \\ u & \longmapsto & x_u(T) \end{array}$$

où $\mathcal{U}$ est l'ensemble des contrôles admissibles, *i.e.* l'ensemble de contrôles u tels que la trajectoire associée est bien définie sur $[0, T]$.

Autrement dit, l'application entrée-sortie en temps T associe à un contrôle u le point final de la trajectoire associée à u. Une question importante en théorie du contrôle est d'étudier cette application en décrivant son image, ses singularités, etc.

5.1.2 Régularité de l'application entrée-sortie

La régularité de E_T dépend bien entendu de l'espace de départ et de la forme du système.

Pour un système général

En toute généralité on a le résultat suivant (voir par exemple [13, 43, 64]).

Proposition 5.1.1. *Considérons le système (4.28) où f est C^p, $p \geqslant 1$, et soit $\mathcal{U} \subset L^\infty([0,T], I\!R^m)$ le domaine de définition de E_T, c'est-à-dire l'ensemble des contrôles dont la trajectoire associée est bien définie sur $[0,T]$. Alors $\mathcal{U}$ est un ouvert de $L^\infty([0,T], I\!R^m)$, et E_T est C^p au sens L^∞.*

De plus la différentielle (au sens de Fréchet) de E_T en un point $u \in \mathcal{U}$ est donnée par le système linéarisé en u de la manière suivante. Posons, pour tout $t \in [0,T]$,

$$A(t) = \frac{\partial f}{\partial x}(t, x_u(t), u(t)) \ , \ \ B(t) = \frac{\partial f}{\partial u}(t, x_u(t), u(t)).$$

Le système de contrôle linéaire

$$\begin{aligned} \dot{y}_v(t) &= A(t)y_v(t) + B(t)v(t) \\ y_v(0) &= 0 \end{aligned}$$

est appelé système linéarisé le long de la trajectoire x_u. La différentielle de Fréchet de E_T en u est alors l'application $dE_T(u)$ telle que, pour tout $v \in L^\infty([0,T], I\!R^m)$,

$$dE_T(u).v \ = \ y_v(T) \ = \ M(T) \int_0^T M^{-1}(s)B(s)v(s)ds \tag{5.1}$$

où $M(\cdot)$ est la résolvante du système linéarisé, i.e. la solution matricielle de $\dot{M}(t) = A(t)M(t)$, $M(0) = Id$.

Démonstration. Pour la démonstration du fait que $\mathcal{U}$ est ouvert, voir [64, 71, 72]. Par hypothèse $u(\cdot)$ et sa trajectoire associée $x(., x_0, u)$ sont définis sur $[0,T]$. L'ensemble des contrôles étant les applications mesurables et bornées muni de la norme L^∞, l'application E_T est de classe C^p sur un voisinage de $u(\cdot)$ en vertu des théorèmes de dépendance par rapport à un paramètre. Exprimons sa différentielle au sens de Fréchet. Soit $v(\cdot)$ un contrôle fixé, on note $x(\cdot) + \delta x(\cdot)$ la trajectoire associée à $u(\cdot) + v(\cdot)$, issue en $t = 0$ de x_0. Par un développement de Taylor, on obtient

$$\begin{aligned} \frac{d}{dt}(x + \delta x)(t) &= f(t, x(t) + \delta x(t), u(t) + v(t)) \\ &= f(t, x(t), u(t)) + \frac{\partial f}{\partial x}(t, x(t), u(t))\delta x(t) + \frac{\partial f}{\partial u}(t, x(t), u(t))v(t) \\ &\quad + \frac{\partial^2 f}{\partial x \partial u}(t, x(t), u(t))(\delta x(t), v(t)) + \cdots \end{aligned}$$

Par ailleurs, $\dot{x}(t) = f(t, x(t), u(t))$, donc

$$\frac{d}{dt}(\delta x)(t) = \frac{\partial f}{\partial x}(t, x(t), u(t))\delta x(t) + \frac{\partial f}{\partial u}(t, x(t), u(t))v(t) + \cdots$$

En écrivant $\delta x = \delta_1 x + \delta_2 x + \ldots$ où $\delta_1 x$ est la partie linéaire en v, $\delta_2 x$ la partie quadratique, etc, et en identifiant, il vient

$$\frac{d}{dt}(\delta_1 x)(t) = \frac{\partial f}{\partial x}(t, x(t), u(t))\delta_1 x(t) + \frac{\partial f}{\partial u}(t, x(t), u(t))v(t) = A(t)\delta_1 x(t) + B(t)v(t).$$

Or $x(0) + \delta x(0) = x_0 = x(0)$, donc $\delta x(0) = 0$ et la condition initiale de cette équation différentielle est $\delta_1 x(0) = 0$. En intégrant, on obtient

$$\delta_1 x(T) = M(T) \int_0^T M^{-1}(s)B(s)v(s)ds$$

où M est la résolvante du système homogène $\dfrac{d}{dt}(\delta_1 x)(t) = \dfrac{\partial f}{\partial x}(t, x(t), u(t))\delta_1 x(t)$, c'est-à-dire $\dot{M}(t) = A(t)M(t)$ avec $A(t) = \dfrac{\partial f}{\partial x}(t, x(t), u(t))$ et $M(0) = I_n$. On observe que $\delta_1 x(T)$ est linéaire et continu par rapport à $v(\cdot)$ en topologie L^∞. C'est donc la différentielle de Fréchet en $u(\cdot)$ de E_T. □

Remarque 5.1.1. En général E_T n'est pas définie sur $L^\infty([0, T], \mathbb{R}^m)$ tout entier à cause de phénomènes d'explosion. Par exemple si on considère le système scalaire $\dot{x} = x^2 + u, x(0) = 0$, on voit que pour $u = 1$ la trajectoire associée explose en $t = \frac{\pi}{2}$, et donc n'est pas définie sur $[0, T]$ si $T \geqslant \frac{\pi}{2}$.

Pour un système affine

Définition 5.1.2. On appelle *système affine contrôlé* un système de la forme

$$\dot{x}(t) = f_0(x(t)) + \sum_{i=1}^{m} u_i(t) f_i(x(t)),$$

où les f_i sont des champs de vecteurs de $\mathbb{R}^n$.

Pour un système affine on peut améliorer le résultat précédent (voir [64, 71, 72]).

Proposition 5.1.2. *Considérons un système affine lisse, et soit $\mathcal{U}$ le domaine de définition de E_T Alors $\mathcal{U}$ est un ouvert de $L^2([0, T], \mathbb{R}^m)$, et l'application entrée-sortie E_T est lisse au sens L^2, et est analytique si les champs de vecteurs sont analytiques.*

Il est très intéressant de considérer L^2 comme espace de contrôles. En effet dans cet espace on bénéficie d'une structure hilbertienne qui permet de faire une *théorie spectrale* de l'application entrée-sortie, et on bénéficie d'autre part de bonnes propriétés de *compacité faible* (voir [71, 72]).

5.2 Contrôlabilité

On veut répondre à la question suivante : étant donné le système (4.28), où peut-on aller en temps T en faisant varier le contrôle u ? On est tout d'abord amené à définir la notion d'ensemble accessible.

5.2.1 Ensemble accessible

Définition 5.2.1. *L'ensemble accessible en temps T* pour le système (4.28), noté $Acc(x_0, T)$, est l'ensemble des extrémités au temps T des solutions du système partant de x_0 au temps $t = 0$. Autrement dit, c'est l'image de l'application entrée-sortie en temps T.

Théorème 5.2.1. *Considérons le système de contrôle*

$$\dot{x} = f(t, x, u),\ x(0) = x_0,$$

où la fonction f est C^1 sur $\mathbb{R}^{1+n+m}$, et les contrôles u appartiennent à l'ensemble $\mathcal{U}$ des fonctions mesurables à valeurs dans un compact $\Omega \subset \mathbb{R}^m$. On suppose que

- *il existe un réel positif b tel que toute trajectoire associée est uniformément bornée par b sur $[0, T]$, i.e.*

$$\exists b > 0 \mid \forall u \in \mathcal{U} \quad \forall t \in [0, T] \quad \|x_u(t)\| \leqslant b, \tag{5.2}$$

- *pour tout (t, x), l'ensemble des vecteurs vitesses*

$$V(t, x) = \{f(t, x, u) \mid u \in \Omega\} \tag{5.3}$$

  est convexe.

Alors l'ensemble $Acc(x_0, t)$ est compact et varie continûment en t sur $[0, T]$.

Démonstration. Notons tout d'abord que puisque Ω est compact alors $V(t, x)$ est également compact. Montrons la compacité de $Acc(x_0, t)$. Cela revient à montrer que toute suite (x_n) de points de $Acc(x_0, t)$ admet une sous-suite convergente. Pour tout entier n soit u_n un contrôle reliant x_0 à x_n en temps t, et soit $x_n(\cdot)$ la trajectoire correspondante. On a donc

$$x_n = x_n(t) = x_0 + \int_0^t f(s, x_n(s), u_n(s))ds.$$

Posons, pour tout entier n et presque tout $s \in [0, t]$,

$$g_n(s) = f(s, x_n(s), u_n(s)).$$

D'après les hypothèses il s'ensuit que la suite de fonctions $(g_n(\cdot))_{n\in\mathbb{N}}$ est bornée dans $L^\infty([0, t], \mathbb{R}^n)$, et par conséquent à sous-suite près elle converge vers une fonction $g(\cdot)$ pour la topologie faible étoile de $L^\infty([0, t], \mathbb{R}^n)$. Posons alors, pour tout $\tau \in [0, t]$,

$$x(\tau) = x_0 + \int_0^\tau g(s)ds,$$

ce qui définit une application $x(\cdot)$ absolument continue sur $[0,t]$. De plus on a, pour tout $s \in [0,t]$,

$$\lim_{n\to+\infty} x_n(s) = x(s),$$

i.e. la suite de fonctions $(x_n(\cdot))_{n\in\mathbb{N}}$ converge simplement vers $x(\cdot)$. Le but est de montrer que la trajectoire $x(\cdot)$ est associée à un contrôle u à valeurs dans Ω, ce qui revient à montrer que pour presque tout $s \in [0,t]$ on a $g(s) = f(s, x(s), u(s))$.

Pour cela, définissons, pour tout entier n et presque tout $s \in [0,t]$,

$$h_n(s) = f(s, x(s), u_n(s)),$$

et introduisons l'ensemble

$$\mathcal{V} = \{h(\cdot) \in L^2([0,t], \mathbb{R}^n) \mid h(s) \in V(s, x(s)) \text{ pour presque tout } s \in [0,t]\},$$

de sorte que $h_n \in \mathcal{V}$ pour tout entier n. Pour tout (t,x) l'ensemble $V(t,x)$ est compact convexe, et, en utilisant le fait que de toute suite convergeant fortement dans L^2 on peut extraire une sous-suite convergeant presque partout, on montre que $\mathcal{V}$ est convexe fermé dans $L^2([0,t], \mathbb{R}^n)$ pour la topologie forte ; donc il est également fermé dans $L^2([0,t], \mathbb{R}^n)$ muni de la topologie faible (voir [19]).

Or, similairement à (g_n), la suite de fonctions (h_n) est bornée dans L^2, et donc à sous-suite près converge en topologie faible vers une fonction h, qui appartient nécessairement à $\mathcal{V}$ puisque ce sous-ensemble est fermé faible.

Enfin, montrons que $g = h$ presque partout. Pour cela, écrivons, pour toute fonction $\varphi \in L^2([0,t], \mathbb{R})$,

$$\int_0^t \varphi(s) g_n(s) ds = \int_0^t \varphi(s) h_n(s) ds + \int_0^t \varphi(s) \left(g_n(s) - h_n(s)\right) ds. \tag{5.4}$$

D'après les hypothèses, la fonction f est globalement lipschitzienne en x sur $[0,T] \times \bar{B}(0,b) \times \Omega$, et donc d'après le théorème des accroissements finis, il existe une constante $C > 0$ telle que, pour presque tout $s \in [0,t]$,

$$\|g_n(s) - h_n(s)\| \leqslant C \|x_n(s) - x(s)\|.$$

La suite de fonctions (x_n) converge simplement vers $x(\cdot)$, donc d'après le théorème de convergence dominée,

$$\int_0^t \varphi(s) \left(g_n(s) - h_n(s)\right) ds \underset{n\to+\infty}{\longrightarrow} 0.$$

Finalement en passant à la limite dans (5.4), il vient

$$\int_0^t \varphi(s) g(s) ds = \int_0^t \varphi(s) h(s) ds,$$

pour toute fonction $\varphi \in L^2([0,t], \mathbb{R})$, et par conséquent $g = h$ presque partout sur $[0,t]$.

En particulier $g \in \mathcal{V}$, et donc pour presque tout $s \in [0,t]$ il existe $u(s) \in \Omega$ tel que

$$g(s) = f(s, x(s), u(s)).$$

En appliquant un lemme de sélection mesurable de théorie de la mesure (notons que $g \in L^\infty([0,t], \mathbb{R}^n)$, on peut montrer que l'application $u(\cdot)$ peut être choisie mesurable sur $[0,T]$ (voir [52, Lem. 2A, 3A p. 161]).

Ainsi, la trajectoire $x(\cdot)$ est associée sur $[0,t]$ au contrôle u à valeurs dans Ω, et $x(t)$ est la limite des points x_n. Ceci montre la compacité de $Acc(x_0,t)$.

Il reste à établir la continuité par rapport à t de l'ensemble accessible. Soient t_1, t_2 deux réels tels que $0 < t_1 < t_2 \leqslant T$, et x_2 un point de $Acc(x_0,t_2)$. Par définition il existe un contrôle u à valeurs dans Ω, de trajectoire associée $x(\cdot)$, tel que

$$x_2 = x(t_2) = x_0 + \int_0^{t_2} f(t, x(t), u(t))dt.$$

Il est bien clair que le point

$$x_1 = x(t_1) = x_0 + \int_0^{t_1} f(t, x(t), u(t))dt$$

appartient à $Acc(x_0,t_1)$, et de plus d'après les hypothèses sur f on a

$$\|x_2 - x_1\| \leqslant C|t_2 - t_1|.$$

On conclut alors facilement. □

Remarque 5.2.1. L'hypothèse (5.2) est indispensable, elle n'est pas une conséquence des autres hypothèses. En effet considérons de nouveau le système de la remarque 5.1.1, *i.e.* $\dot{x} = x^2 + u, x(0) = 0$, où on suppose que $|u| \leqslant 1$ et que le temps final est $T = \frac{\pi}{2}$. Alors pour tout contrôle u constant égal à c, avec $0 < c < 1$, la trajectoire associée est $x_c(t) = \sqrt{c}\tan\sqrt{c}t$, donc est bien définie sur $[0,T]$, mais lorsque c tend vers 1 alors $x_c(T)$ tend vers $+\infty$. Par ailleurs il est facile de voir que sur cet exemple l'ensemble des contrôles admissibles, à valeurs dans $[-1,1]$, est l'ensemble des fonctions mesurables telles que $u(t) \in [-1,1[$.

Remarque 5.2.2. De même, l'hypothèse de convexité (5.3) est nécessaire (voir [52, Exemple 2 page 244]).

5.2.2 Résultats de contrôlabilité

Définition 5.2.2. Le système (4.28) est dit *contrôlable (en temps quelconque)* depuis x_0 si

$$\mathbb{R}^n = \bigcup_{T \geqslant 0} Acc(x_0, T).$$

Il est dit *contrôlable en temps* T si $\mathbb{R}^n = Acc(x_0,T)$.

Par des arguments du type théorème des fonctions implicites, l'étude de la contrôlabilité du système linéarisé (qui est plus simple), permet de déduire des résultats de *contrôlabilité locale* du système de départ (voir [13, 52]). Par exemple on déduit du théorème de contrôlabilité dans le cas linéaire la proposition suivante.

Proposition 5.2.2. *Considérons le système (4.28) où* $f(x_0, u_0) = 0$. *Notons* $A = \frac{\partial f}{\partial x}(x_0, u_0)$ *et* $B = \frac{\partial f}{\partial u}(x_0, u_0)$. *On suppose que*

$$rg\ \big(B|AB|\cdots|A^{n-1}B\big) = n.$$

Alors le système est localement contrôlable en x_0.

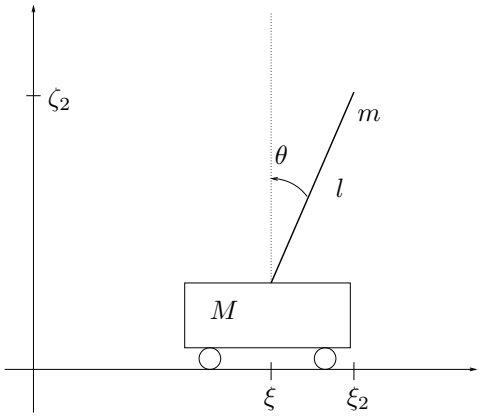


FIGURE 5.1 – Pendule inversé.

Exemple 5.2.1 (Pendule inversé)**.** Considérons un pendule inversé, de masse m, fixé à un chariot de masse M dont on contrôle l'accélération $u(t)$ (voir figure 5.1). Ecrivons les équations du mouvement en utilisant les équations d'Euler-Lagrange. L'énergie cinétique et l'énergie potentielle sont

$$E_c = \frac{1}{2}M\dot{\xi}^2 + \frac{1}{2}m(\dot{\xi}_2^2 + \dot{\zeta}_2^2),\ E_p = mg\zeta_2.$$

Par ailleurs, on a $\zeta_2 = l\cos\theta$ et $\xi_2 = \xi + l\sin\theta$. Donc le Lagrangien du système est

$$L = E_c - E_p = \frac{1}{2}(M+m)\dot{\xi}^2 + ml\dot{\xi}\dot{\theta}\cos\theta + \frac{1}{2}ml^2\dot{\theta}^2 - mgl\cos\theta.$$

D'après les équations d'Euler-Lagrange,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x} + F_{ext},$$

on obtient

$$\begin{cases} (M+m)\ddot{\xi} + ml\ddot{\theta}\cos\theta - ml\dot{\theta}^2\sin\theta = u, \\ ml\ddot{\xi}\cos\theta + ml^2\ddot{\theta} - mgl\sin\theta = 0, \end{cases}$$

d'où

$$\begin{cases} \ddot{\xi} = \dfrac{ml\dot{\theta}^2\sin\theta - mg\cos\theta\sin\theta + u}{M + m\sin^2\theta}, \\ \ddot{\theta} = \dfrac{-ml\dot{\theta}^2\sin\theta\cos\theta + (M+m)g\sin\theta - u\cos\theta}{M + m\sin^2\theta}. \end{cases}$$

On établit facilement que le système linéarisé au point d'équilibre ($\xi = \xi_c, \dot{\xi} = 0, \theta = 0, \dot{\theta} = 0$) est donné par les matrices

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{mg}{M} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \frac{(M+m)g}{lM} & 0 \end{pmatrix}, \text{ et } B = \begin{pmatrix} 0 \\ \frac{1}{M} \\ 0 \\ -\frac{1}{lM} \end{pmatrix}.$$

On vérifie aisément la condition de Kalman, ce qui établit que le pendule inversé est localement contrôlable en ce point d'équilibre (instable).

Le théorème de Chow relie la contrôlabilité à des propriétés de crochets de Lie du système. On a par exemple la conséquence suivante sur les systèmes dits *sous-Riemanniens*.

Proposition 5.2.3. *Considérons dans $\mathbb{R}^n$ le système sous-Riemannien lisse*

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x) \ , \ \ x(0) = x_0.$$

On suppose que l'algèbre de Lie engendrée par les champs de vecteurs f_i est de dimension n. Alors le système est contrôlable.

Démonstration. Pour simplifier, faisons la démonstration dans le cas $m = 2$ et $n = 3$. On suppose que $rg(f_1, f_2, [f_1, f_2])(x) = 3, \ \forall x \in \mathbb{R}^n$. Soit $\lambda \in \mathbb{R}$. On considère l'application

$$\varphi_\lambda : (t_1, t_2, t_3) \longmapsto (\exp \lambda f_1 \exp t_3 f_2 \exp\, - \lambda f_1)(\exp t_2 f_2)(\exp t_1 f_1)(x_0).$$

On a $\varphi_\lambda(0) = x_0$. Montrons que pour $\lambda \neq 0$ assez petit, φ_λ est une immersion en 0. En utilisant la formule de Campbell-Hausdorff, on obtient

$$\varphi_\lambda(t_1, t_2, t_3) = \exp(t_1 f_1 + (t_2 + t_3) f_2 + \lambda t_3 [f_1, f_2] + \ \ldots),$$

d'où

$$\frac{\partial \varphi_\lambda}{\partial t_1}(0) = f_1(x_0), \ \frac{\partial \varphi_\lambda}{\partial t_2}(0) = f_2(x_0), \ \frac{\partial \varphi_\lambda}{\partial t_3}(0) = f_2(x_0) + \lambda [f_1, f_2](x_0) + \mathrm{o}(\lambda).$$

Par hypothèse, les champs de vecteurs f_1, f_2, $[f_1, f_2]$ sont linéairement indépendants, donc la jacobienne de φ_λ est de rang 3 en 0. Le théorème d'inversion locale et un argument de connexité nous permettent de conclure. □

Remarque 5.2.3. En général, le problème de contrôlabilité est difficile. Il est lié à la question de savoir quand un semi-groupe opère transitivement. Il existe cependant des techniques pour montrer, dans certains cas, la contrôlabilité globale. L'une d'entre elles, importante, s'appelle la *technique d'élargissement* (voir [13, 43]).

5.3 Contrôles singuliers

5.3.1 Définition

Définition 5.3.1. Soit u un contrôle défini sur $[0,T]$ tel que sa trajectoire associée x_u issue de $x(0)=x_0$ est définie sur $[0,T]$. On dit que le contrôle u (ou la trajectoire x_u) est singulier sur $[0,T]$ si la différentielle de Fréchet $dE_T(u)$ de l'application entrée-sortie au point u n'est pas surjective. Sinon on dit qu'il est régulier.

Dans les résultats ci-dessous on suppose que les contrôles considérés sont à l'intérieur de l'ensemble des contrôles admissibles, sans quoi l'argument de fonctions implicites classique ne pourrait s'appliquer à cause de l'existence d'un bord.

Proposition 5.3.1. *Soient x_0 et T fixés. Si u est un contrôle régulier, alors E_T est ouverte dans un voisinage de u.*

Démonstration. Par hypothèse, il existe n contrôles v_i tels que $dE_T(u).v_i = e_i$ où $(e_1,\dots,e_n)$ est la base canonique de $\mathbb{R}^n$. On considère l'application

$$(\lambda_1,\dots,\lambda_n)\in\mathbb{R}^n \longmapsto E_T(u+\sum_{i=1}^{n}\lambda_i v_i).$$

Par construction, c'est un difféomorphisme local, et le résultat s'ensuit. □

Autrement dit en un point x_1 atteignable en temps T depuis x_0 par une trajectoire régulière $x(\cdot)$, l'ensemble accessible $Acc(x_0,T)$ est *localement ouvert*, *i.e.* est un voisinage du point x_1. En particulier cela implique que le système est *localement contrôlable* autour du point x_1. On parle aussi de *contrôlabilité le long de la trajectoire* $x(\cdot)$. On obtient ainsi la proposition suivante.

Proposition 5.3.2. *Si u est un contrôle régulier sur $[0,T]$, alors le système est localement contrôlable le long de la trajectoire associée à ce contrôle.*

Le corollaire suivant est immédiat.

Corollaire 5.3.3. *Soit u un contrôle défini sur $[0,T]$ tel que sa trajectoire associée x_u issue de $x(0)=x_0$ est définie sur $[0,T]$ et vérifie au temps T*

$$x(T)\in\partial Acc(x_0,T).$$

Alors le contrôle u est singulier sur $[0,T]$.

Remarque 5.3.1. Le système peut être localement contrôlable le long d'une trajectoire singulière. C'est le cas du système scalaire $\dot{x} = u^3$, où le contrôle $u = 0$ est singulier.

5.3.2 Caractérisation hamiltonienne des contrôles singuliers

Montrons qu'une trajectoire singulière peut se paramétrer comme la projection d'une solution d'un système hamiltonien contraint. Considérons de nouveau le système de contrôle général

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{5.5}$$

où f est une application de classe C^1 de $\mathbb{R}^{1+n+m}$ dans $\mathbb{R}^n$.

Définition 5.3.2. Le *Hamiltonien* du système (5.5) est la fonction

$$\begin{array}{rccl} H : \mathbb{R} \times \mathbb{R}^n \times (\mathbb{R}^n \setminus \{0\}) \times \mathbb{R}^m & \longrightarrow & \mathbb{R} \\ (t, x, p, u) & \longmapsto & H(t, x, p, u) = \langle p, f(t, x, u) \rangle \end{array}$$

où $\langle\ ,\ \rangle$ est le produit scalaire usuel de $\mathbb{R}^n$.

Remarque 5.3.2. Il est souvent commode de considérer p comme un vecteur ligne, et alors avec des notations matricielles on peut écrire

$$H(t, x, p, u) = pf(t, x, u).$$

Nous ferons toujours par la suite cette confusion, et le vecteur adjoint sera tantôt un vecteur ligne, tantôt un vecteur colonne, pour alléger les notations. Nous laissons au lecteur le soin de s'accommoder de cette volontaire ambiguïté.

Proposition 5.3.4. *Soit u un contrôle singulier sur $[0, T]$ pour le système de contrôle (5.5), et soit $x(\cdot)$ la trajectoire singulière associée. Alors il existe une application absolument continue $p : [0, T] \longrightarrow \mathbb{R}^n \setminus \{0\}$, appelée vecteur adjoint, telle que les équations suivantes sont vérifiées pour presque tout $t \in [0, T]$*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t, x(t), p(t), u(t)), \tag{5.6}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t), u(t)), \tag{5.7}$$

$$\frac{\partial H}{\partial u}(t, x(t), p(t), u(t)) = 0, \tag{5.8}$$

où H est le hamiltonien du système.

L'équation (5.8) est appelée *équation de contrainte*.

Démonstration. Par définition, le couple (x, u) est singulier sur $[0, T]$ si $dE_T(u)$ n'est pas surjective. Donc il existe un vecteur ligne $\psi \in \mathbb{R}^n \setminus \{0\}$ tel que pour tout contrôle v dans L^∞ on ait

$$\psi.dE_T(u).v = \psi \int_0^T M(T)M^{-1}(s)B(s)v(s)ds = 0$$

Par conséquent

$$\psi M(T)M^{-1}(s)B(s) = 0 \quad \text{p.p. sur [0,T].}$$

On pose $p(t) = \psi M(T)M^{-1}(t)$ pour tout $t \in [0, T]$. C'est un vecteur ligne de $\mathbb{R}^n \setminus \{0\}$, et $p(T) = \psi$. On a par dérivation

$$\dot{p}(t) = -p(t)\frac{\partial f}{\partial x}(t, x(t), u(t)).$$

En introduisant le Hamiltonien $H(t, x, p, u) = pf(t, x, u)$, on obtient

$$f(t, x(t), u(t)) = \frac{\partial H}{\partial p}(t, x(t), p(t), u(t)),$$

et

$$-p(t)\frac{\partial f}{\partial x}(t, x(t), u(t)) = -\frac{\partial H}{\partial x}(t, x(t), p(t), u(t)).$$

La dernière relation vient de $p(t)B(t) = 0$ car $B(t) = \dfrac{\partial f}{\partial u}(t, x(t), u(t))$. □

Remarque 5.3.3 (Interprétation géométrique du vecteur adjoint). Si u est un contrôle singulier sur $[0, T]$ alors u est aussi singulier sur $[0, t]$ pour tout $t \in]0, T]$, et de plus $p(t)$ est orthogonal à l'image de l'application linéaire $dE_t(u)$. En particulier Im $dE_t(u)$ est un sous-espace de $\mathbb{R}^n$ de codimension supérieure ou égale à 1.

En effet, on a pour tout contrôle $v \in L^\infty([0, t], \mathbb{R}^m)$

$$p(t)dE_t(u).v = p(t)M(t) \int_0^t M(s)^{-1}B(s)v(s)ds,$$

or $p(t) = \psi M(T)M(t)^{-1}$, d'où en prolongeant $v(s)$ par 0 sur $]t, T]$,

$$p(t)dE_t(u).v = \psi M(T) \int_0^T M(s)^{-1}B(s)v(s)ds = \psi dE_T(u).v = 0.$$

Remarque 5.3.4. La proposition 5.3.4 et les remarques précédentes sont les prémisses du principe du maximum de Pontryagin.

5.3.3 Calcul des contrôles singuliers

Considérons un point (t_0, x_0, p_0, u_0) appartenant à l'ensemble des contraintes

$$\Sigma = \left\{ (t, x, p, u) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \setminus \{0\} \times \mathbb{R}^m \mid \frac{\partial H}{\partial u}(t, x, p, u) = 0 \right\}.$$

Si la Hessienne $\left(\frac{\partial^2 H}{\partial u_i \partial u_j} \right)_{i,j}$ est inversible en ce point, alors d'après le théorème des fonctions implicites le contrôle singulier peut se calculer comme une fonction de (t, x, p) au voisinage de (t_0, x_0, p_0).

Exercice 5.3.1. Calculer les contrôles singuliers du système

$$\dot{x} = y + u, \ \dot{y} = -x + u^2.$$

Si le Hamiltonien est linéaire en le contrôle, la méthode consiste à dériver par rapport à t la contrainte (5.8). Considérons par exemple un *système affine mono-entrée* lisse

$$\dot{x} = f_0(x) + u f_1(x).$$

Il convient d'utiliser le formalisme Hamiltonien. Posons $h_i(x, p) = \langle p, f_i(x) \rangle, i = 0, 1$, et $z(t) = (x(t), p(t))$. En dérivant deux fois la contrainte on obtient

$$\{\{h_1, h_0\}, h_0\}(z(t)) + u(t)\{\{h_1, h_0\}, h_1\}(z(t)) = 0,$$

où $\{\ ,\ \}$ désigne le crochet de Poisson, et on en déduit donc le contrôle singulier pourvu que $\{\{h_1, h_0\}, h_1\}(z(t))$ ne s'annule pas (voir [16] pour plus de détails).

Chapitre 6

Contrôle optimal

6.1 Présentation du problème

Maintenant, en plus d'un problème de contrôle, on se donne un problème de minimisation : parmi toutes les solutions du système (4.28) reliant x_0 à x_1, trouver une trajectoire qui minimise une certaine fonction *coût* $C(T, u)$. Une telle trajectoire, si elle existe, est dite *optimale* pour ce coût. L'existence de trajectoires optimales dépend de la régularité du système et du coût (pour un énoncé général, voir [13, 43, 52]). Il se peut aussi qu'un contrôle optimal n'existe pas dans la classe de contrôles considérés, mais existe dans un espace plus gros : c'est le phénomène de Lavrentiev (voir [62]). En particulier on a intérêt à travailler dans un espace de contrôles complet et qui ait de bonnes propriétés de compacité.

6.2 Existence de trajectoires optimales

6.2.1 Pour des systèmes généraux

Théorème 6.2.1. *Considérons le système de contrôle*

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

où f est C^1 de $\mathbb{R}^{1+n+m}$ dans $\mathbb{R}^n$, les contrôles u sont à valeurs dans un compact $\Omega \subset \mathbb{R}^m$, et où éventuellement on a des contraintes sur l'état

$$c_1(x) \leqslant 0, \ldots, c_r(x) \leqslant 0,$$

où $c_1, \ldots c_r$ sont des fonctions continues sur $\mathbb{R}^n$. Soient M_0 et M_1 deux compacts de $\mathbb{R}^n$ tels que M_1 est accessible depuis M_0. Soit $\mathcal{U}$ l'ensemble des contrôles à valeurs dans Ω joignant M_0 à M_1. Soient f^0 une fonction de classe C^1 sur $\mathbb{R}^{1+n+m}$, et g une fonction continue sur $\mathbb{R}^n$. On considère le coût

$$C(u) = \int_0^{t(u)} f^0(t, x(t), u(t))dt + g(t(u), x(t(u))),$$

où $t(u) \geqslant 0$ est tel que $x(t(u)) \in M_1$. On suppose que

- *il existe un réel positif b tel que toute trajectoire associée à un contrôle $u \in \mathcal{U}$ est uniformément bornée par b sur $[0, t(u)]$, ainsi que le temps $t(u)$, i.e. ,*

$$\exists b > 0 \mid \forall u \in \mathcal{U} \quad \forall t \in [0, t(u)] \quad t(u) + \|x_u(t)\| \leqslant b, \tag{6.1}$$

- *pour tout $(t, x) \in \mathbb{R}^{1+n}$, l'ensemble*

$$\tilde{V}(t, x) = \left\{ \begin{pmatrix} f(t, x, u) \\ f^0(t, x, u) + \gamma \end{pmatrix} \;\middle|\; u \in \Omega,\ \gamma \geqslant 0 \right\} \tag{6.2}$$

  est convexe.

Alors il existe un contrôle optimal u sur $[0, t(u)]$ tel que la trajectoire associée joint M_0 à M_1 en temps $t(u)$ et en coût minimal.

Bien entendu pour un problème de contrôle optimal à temps final fixé on impose $t(u) = T$ (et en particulier on suppose que la cible M_1 est accessible depuis M_0 en temps T).

La preuve de ce théorème est semblable à celle du théorème 5.2.1. La prise en compte de contraintes sur l'état ne pose aucun problème. Notons que l'hypothèse (6.2) implique la convexité de l'ensemble des vecteurs vitesses, et aussi (terme $\gamma \geqslant 0$) une propriété de convexité d'épigraphe. Nous donnons tout de même cette preuve ci-dessous.

Remarque 6.2.1. On peut montrer un résultat plus général où l'ensemble de départ M_0 et la cible M_1 dépendent du temps t, ainsi que le domaine des contraintes Ω sur le contrôle (voir [52]).

Démonstration. Soit δ l'infimum des coûts $C(u)$ sur l'ensemble des contrôles admissibles $u \in L^\infty([0, t(u)], \Omega)$ engendrant des trajectoires telles que $x(0) \in M_0$, $x(t(u)) \in M_1$ et vérifiant les contraintes sur l'état $c_1(x(\cdot)) \leqslant 0, \ldots, c_r(x(\cdot)) \leqslant 0$.

Considérons une suite minimisante de trajectoires $x_n(\cdot)$ associées à des contrôles u_n, c'est-à-dire une suite de trajectoires vérifiant ces propriétés et telle que $C(u_n) \to \delta$ quand $n \to +\infty$. Pour tout n on note

$$\tilde{F}_n(t) = \begin{pmatrix} f(t, x_n(t), u_n(t)) \\ f^0(t, x_n(t), u_n(t)) \end{pmatrix} = \begin{pmatrix} F_n(t) \\ F_n^0(t) \end{pmatrix}$$

pour presque tout $t \in [0, t(u_n)]$. D'après les hypothèses, la suite de fonctions $(\tilde{F}_n(\cdot))_{n \in \mathbb{N}}$ (étendues par 0 sur $]t_n(u), b]$) est bornée dans $L^\infty([0, b], \mathbb{R}^n)$, et par conséquent à sous-suite près elle converge vers une fonction

$$\tilde{F}(\cdot) = \begin{pmatrix} F(\cdot) \\ F^0(\cdot) \end{pmatrix}$$

pour la topologie faible étoile de $L^\infty([0, b], \mathbb{R}^{n+1})$. A sous-suite près de même la suite $(t_n(u_n))_{n \in \mathbb{N}}$ converge vers $T \geqslant 0$, et on a $\tilde{F}(t) = 0$ pour $t \in]T, b]$. Enfin,

par compacité de M_0, à sous-suite près la suite $(x_n(0))_{n\in\mathbb{N}}$ converge vers un point $x_0 \in M_0$. Posons alors, pour tout $t \in [0,T]$,

$$x(t) = x_0 + \int_0^t F(s)ds,$$

ce qui construit une fonction $x(\cdot)$ absolument continue sur $[0,T]$. De plus on a, pour tout $t \in [0,T]$,

$$\lim_{n\to+\infty} x_n(t) = x(t),$$

i.e. la suite de fonctions $(x_n(\cdot))_{n\in\mathbb{N}}$ converge simplement vers $x(\cdot)$. Comme dans la preuve du théorème 5.2.1, le but est de montrer que la trajectoire $x(\cdot)$ est associée à un contrôle u à valeurs dans Ω, et que de plus ce contrôle u est optimal pour le problème considéré.

Pour tout entier n et presque tout $t \in [0, t(u_n)]$, on pose

$$\tilde{h}_n(t) = \begin{pmatrix} f(t,x(t),u_n(t)) \\ f^0(t,x(t),u_n(t)) \end{pmatrix}.$$

Si $T > t(u_n)$, on étend $\tilde{h}_n$ sur $[0,T]$ par

$$\tilde{h}_n(t) = \begin{pmatrix} f(t,x(t),v) \\ f^0(t,x(t),v) \end{pmatrix},$$

où $v \in \Omega$ est quelconque. Par ailleurs, on définit

$$\beta = \max\{|f^0(t,x,u)| \mid 0 \leqslant t \leqslant b,\ \|x\| \leqslant b,\ u \in \Omega\}.$$

Comme Ω est compact, $\beta > 0$ est bien défini. Pour tout $(t,x) \in \mathbb{R}^{1+n}$, on modifie alors légèrement la définition de $\tilde{V}(t,x)$ pour le rendre compact (tout en le gardant convexe), en posant

$$\tilde{V}_\beta(t,x) = \left\{ \begin{pmatrix} f(t,x,u) \\ f^0(t,x,u)+\gamma \end{pmatrix} \;\middle|\; u \in \Omega,\ \gamma \geqslant 0,\ |f^0(t,x,u)+\gamma| \leqslant \beta \right\}.$$

On définit alors

$$\tilde{\mathcal{V}} = \{\tilde{h}(\cdot) \in L^2([0,T],\mathbb{R}^{n+1}) \mid h(t) \in \tilde{V}_\beta(t,x(t)) \text{ pour presque tout } t \in [0,T]\}.$$

Par construction, on a $\tilde{h}_n \in \tilde{\mathcal{V}}$ pour tout entier n.

Lemme 6.2.2. *L'ensemble $\tilde{\mathcal{V}}$ est convexe fermé fort dans $L^2([0,T],\mathbb{R}^{n+1})$.*

Preuve du lemme 6.2.2. Montrons que $\tilde{\mathcal{V}}$ est convexe. Soient $\tilde{r}_1, \tilde{r}_2 \in \tilde{\mathcal{V}}$, et $\lambda \in [0,1]$. Par définition, pour presque tout $t \in [0,T]$ on a $\tilde{r}_1(t) \in \tilde{V}_\beta(t,x(t))$ et $\tilde{r}_2(t) \in \tilde{V}_\beta(t,x(t))$, or $\tilde{V}_\beta(t,x(t))$ est convexe donc $\lambda\tilde{r}_1(t) + (1-\lambda)\tilde{r}_2(t) \in \tilde{V}_\beta(t,x(t))$. Donc $\lambda\tilde{r}_1 + (1-\lambda)\tilde{r}_2 \in \tilde{\mathcal{V}}$.

Montrons que $\tilde{\mathcal{V}}$ est fermé fort dans $L^2([0,T],\mathbb{R}^n)$. Soit $(\tilde{r}_n)_{n\in\mathbb{N}}$ une suite de $\tilde{\mathcal{V}}$ convergeant vers $\tilde{r}$ pour la topologie forte de $L^2([0,T],\mathbb{R}^n)$. Montrons que $\tilde{r} \in \tilde{\mathcal{V}}$. A sous-suite près, $(\tilde{r}_n)_{n\in\mathbb{N}}$ converge presque partout vers $\tilde{r}$, or par définition, pour presque tout $t \in [0,T]$ on a $\tilde{r}_n(t) \in \tilde{V}_\beta(t,x(t))$, et $\tilde{V}_\beta(t,x(t))$ est compact, donc $\tilde{r}(t) \in \tilde{V}_\beta(t,x(t))$ pour presque tout $t \in [0,T]$. □

L'ensemble $\tilde{\mathcal{V}}$ est donc aussi convexe fermé faible dans $L^2([0,T],\mathbb{R}^{n+1})$. La suite de fonctions $(\tilde{h}_n)_{n\in\mathbb{N}}$ étant bornée dans $L^2([0,T],\mathbb{R}^{n+1})$, à sous-suite près elle converge faiblement vers une fonction $\tilde{h}$, qui appartient à $\tilde{\mathcal{V}}$ puisque ce sous-ensemble est fermé faible.

Montrons que $\tilde{F}=\tilde{h}$ presque partout. Pour cela, écrivons, pour toute fonction $\varphi\in L^2([0,T])$,

$$\int_0^T \varphi(t)\tilde{F}_n(t)\,dt=\int_0^T \varphi(t)\tilde{h}_n(t)\,dt+\int_0^T \varphi(t)\left(\tilde{F}_n(t)-\tilde{h}_n(t)\right)dt. \tag{6.3}$$

D'après les hypothèses, les fonctions f et f^0 sont globalement lipschitziennes en x sur $[0,T]\times\bar{B}(0,b)\times\Omega$, et donc d'après le théorème des accroissements finis, il existe une constante $C>0$ telle que, pour presque tout $t\in[0,T]$, on ait

$$\|\tilde{F}_n(t)-\tilde{h}_n(t)\|\leqslant C\|x_n(t)-x(t)\|.$$

La suite de fonctions $(x_n(\cdot))_{n\in\mathbb{N}}$ converge simplement vers $x(\cdot)$, donc d'après le théorème de convergence dominée,

$$\int_0^T \varphi(t)\left(\tilde{F}_n(t)-\tilde{h}_n(t)\right)dt\underset{n\to+\infty}{\longrightarrow}0.$$

Finalement en passant à la limite dans (6.3), il vient

$$\int_0^T \varphi(t)\tilde{F}(t)\,dt=\int_0^T \varphi(t)\tilde{h}(t)\,dt,$$

pour toute fonction $\varphi\in L^2([0,T])$, et par conséquent $\tilde{F}=\tilde{h}$ presque partout sur $[0,T]$.

En particulier, $\tilde{F}\in\tilde{\mathcal{V}}$, et donc pour presque tout $t\in[0,T]$ il existe $u(t)\in\Omega$ et $\gamma(t)\geqslant 0$ tels que

$$\tilde{F}(t)=\begin{pmatrix} f(t,x(t),u(t)) \\ f^0(t,x(t),u(t))+\gamma(t)\end{pmatrix}.$$

En appliquant un lemme de sélection mesurable de théorie de la mesure (notons que $\tilde{F}\in L^\infty([0,T],\mathbb{R}^{n+1})$, les fonctions $u(\cdot)$ et $\gamma(\cdot)$ peuvent de plus être choisies mesurables sur $[0,T]$ (voir [52, Lem. 2A, 3A p. 161]).

Il reste à montrer que le contrôle u ainsi défini est optimal pour le problème considéré. Tout d'abord, comme $x_n(t_n(u_n))\in M_1$, par compacité de M_1 et d'après les propriétés de convergence montrées précédemment, on obtient $x(T)\in M_1$. De même, clairement on obtient $c_1(x(\cdot))\leqslant 0,\dots,c_r(x(\cdot))\leqslant 0$. Par ailleurs, par définition $C(u_n)$ converge vers δ, et d'après les propriétés de convergence démontrées ci-dessus, $C(u_n)$ converge aussi vers $\int_0^T(f^0(t,x(t),u(t))+\gamma(t))\,dt+g(T,x(T))$. Comme γ est à valeurs positives, cela implique donc que

$$\begin{aligned}\int_0^T f^0(t,x(t),u(t))\,dt+g(T,x(T))&\\ \leqslant\int_0^T (f^0(t,x(t),u(t))+\gamma(t))\,dt+g(T,x(T))&\leqslant C(v),\end{aligned}$$

pour tout contrôle v admissible qui engendre une trajectoire reliant M_0 à M_1 et vérifiant les différentes contraintes. Autrement dit, le contrôle u est optimal. Notons d'ailleurs que la fonction γ est forcément nulle. □

6.2.2 Pour des systèmes affines

Le résultat précédent suppose des contraintes sur le contrôle. En l'absence de contraintes, on a par exemple, pour les systémes affines, le résultat suivant (des résultats plus généraux existent, voir par exemple [36]).

Proposition 6.2.3. *Considérons le système affine dans* $\mathbb{R}^n$

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x) \ , \ x(0) = x_0, x(T) = x_1, \tag{6.4}$$

avec le coût

$$C_T(u) = \int_0^T \sum_{i=1}^{m} u_i^2(t)dt, \tag{6.5}$$

où $T > 0$ *est fixé et la classe* $\mathcal{U}$ *des contrôles admissibles est le sous-ensemble de* $L^2([0,T], \mathbb{R}^m)$ *tel que*

1. $\forall u \in \mathcal{U} \quad x_u$ *est bien définie sur* $[0,T]$ *;*
2. $\exists B_T \mid \forall u \in \mathcal{U} \quad \forall t \in [0,T] \quad ||x_u(t)|| \leqslant B_T$.

Si x_1 *est accessible depuis* x_0 *en temps* T*, alors il existe un contrôle optimal reliant* x_0 *à* x_1.

Démonstration. Considérons une suite de contrôles $(u_i^{(n)}(t))_{n\in\mathbb{N}}$ transférant x_0 en x_1, telle que leur coût tend vers la borne inférieure des coûts des contrôles reliant x_0 à x_1. Soit $x^{(n)}$ la trajectoire associée au contrôle $u^{(n)}$, *i.e.*

$$x^{(n)}(t) = x_0 + \int_0^T \left(f_0(x^{(n)}(t)) + \sum_{i=1}^{m} u_i^{(n)}(t) f_i(x^{(n)}(t)) \right) dt.$$

Les $u_i^{(n)}$ sont bornés dans $L^2([0,T], \mathbb{R}^m)$, et par compacité faible,

$$\exists (n_k)_{k\in\mathbb{N}} \mid u_i^{(n_k)} \underset{k\to+\infty}{\rightharpoonup} v_i \in L^2([0,T], \mathbb{R}^m).$$

Il est par ailleurs facile de voir que la suite $\dot{x}^{(n_k)}$ est bornée dans $L^2([0,T], \mathbb{R}^n)$, et par conséquent $x^{(n_k)}$ est bornée dans $H^1([0,T], \mathbb{R}^n)$, et par réflexivité,

$$\exists (n_{k_p})_{p\in\mathbb{N}} \mid x^{(n_{k_p})} \overset{H^1}{\rightharpoonup} x \in H^1([0,T], \mathbb{R}^n)$$

Or $H^1 \overset{c}{\hookrightarrow} C^0$, donc $x^{(n_{k_p})} \xrightarrow{\text{uniformément}} x$ sur $[0,T]$. On conclut alors aisément par passage à la limite que

$$x(t) = x_0 + \int_0^t \left(f_0(x(t)) + \sum_{i=1}^{m} v_i(t) f_i(x(t)) \right) dt$$

et que $x(T) = x_1$. □

Chapitre 7

Principe du Maximum de Pontryagin

Dans cette section on donne une version générale du principe du maximum de Pontryagin. Ce théorème est difficile à démontrer. En revanche lorsqu'il n'y a pas de contrainte sur le contrôle, la preuve est simple, et on arrive au principe du maximum dit faible. C'est à cette version plus simple que nous allons d'abord nous intéresser. Puis nous passerons au cas général.

7.1 Cas sans contrainte sur le contrôle : principe du maximum faible

7.1.1 Le problème de Lagrange

Ce problème simplifié est le suivant. On cherche des conditions nécessaires d'optimalité pour le système

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{7.1}$$

où les contrôles $u(\cdot) \in \mathcal{U}$ sont définis sur $[0, T]$ et les trajectoires associées doivent vérifier $x(0) = x_0$ et $x(T) = x_1$; le problème est de minimiser un coût de la forme

$$C(u) = \int_0^T f^0(t, x(t), u(t))dt, \tag{7.2}$$

où T est fixé.

Associons au système (7.1) le *système augmenté* suivant

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \\ \dot{x}^0(t) &= f^0(t, x(t), u(t)), \end{aligned} \tag{7.3}$$

et notons $\tilde{x} = (x, x^0)$, $\tilde{f} = (f, f^0)$. Le problème revient donc à chercher une trajectoire solution de (7.3) joignant les points $\tilde{x}_0 = (x_0, 0)$ et $\tilde{x}_1 = (x_1, x^0(T))$, et minimisant la dernière coordonnée $x^0(T)$.

L'ensemble des états accessibles à partir de $\tilde{x}_0$ pour le système (7.3) est $\tilde{Acc}(\tilde{x}_0, T) = \bigcup_{u(\cdot)} \tilde{x}(T, \tilde{x}_0, u)$.

Le lemme crucial est alors le suivant.

Lemme 7.1.1. *Si le contrôle u associé au système de contrôle (7.1) est optimal pour le coût (7.2), alors il est singulier sur $[0, T]$ pour le système augmenté (7.3).*

Démonstration. Notons $\tilde{x}$ la trajectoire associée, solution du système augmenté (7.3), issue de $\tilde{x}_0 = (x_0, 0)$. Le contrôle u étant optimal pour le coût (7.2), il en résulte que le point $\tilde{x}(T)$ appartient à la frontière de l'ensemble $\tilde{Acc}(\tilde{x}_0, T)$ (voir figure 7.1). En effet sinon, il existerait un voisinage du point $\tilde{x}(T) = (x_1, x^0(T))$ dans $\tilde{Acc}(\tilde{x}_0, T)$ contenant un point $\tilde{y}(T)$ solution du système (7.3) et tel que l'on ait $y^0(T) < x^0(T)$, ce qui contredirait l'optimalité du contrôle u. Par conséquent, d'après la proposition 5.3.1, le contrôle u est un contrôle singulier pour le système augmenté (7.3) sur $[0, T]$. □

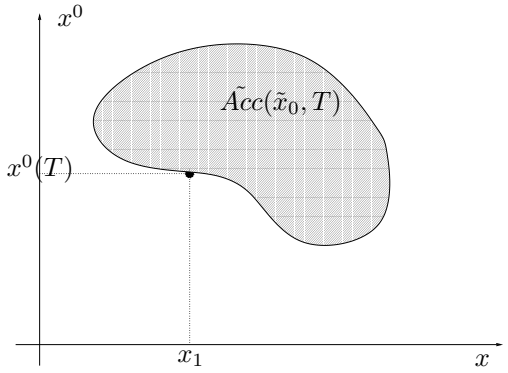


FIGURE 7.1 – Ensemble accessible augmenté.

Dans la situation du lemme, d'après la proposition 5.3.4, il existe une application $\tilde{p} : [0, T] \longrightarrow \mathbb{R}^{n+1} \setminus \{0\}$ telle que $(\tilde{x}, \tilde{p}, \tilde{u})$ soit solution du système hamiltonien

$$\dot{\tilde{x}}(t) = \frac{\partial \tilde{H}}{\partial \tilde{p}}(t, \tilde{x}(t), \tilde{p}(t), u(t)), \ \dot{\tilde{p}}(t) = -\frac{\partial \tilde{H}}{\partial \tilde{x}}(t, \tilde{x}(t), \tilde{p}(t), u(t)), \tag{7.4}$$

$$\frac{\partial \tilde{H}}{\partial u}(t, \tilde{x}(t), \tilde{p}(t), u(t)) = 0 \tag{7.5}$$

où $\tilde{H}(t, \tilde{x}, \tilde{p}, u) = \langle \tilde{p}, \tilde{f}(t, \tilde{x}, u) \rangle$.

En écrivant $\tilde{p} = (p, p^0) \in (\mathbb{R}^n \times \mathbb{R}) \setminus \{0\}$, où p^0 est appelée variable duale du coût, on obtient

$$(\dot{p}, \dot{p}^0) = -(p, p^0) \begin{pmatrix} \dfrac{\partial f}{\partial x} & 0 \\ \dfrac{\partial f^0}{\partial x} & 0 \end{pmatrix},$$

d'où en particulier $\dot{p}^0(t) = 0$, c'est-à-dire que $p^0(t)$ est constant sur $[0, T]$. Comme le vecteur $\tilde{p}(t)$ est défini à scalaire multiplicatif près, on choisit $p^0 \leqslant 0$. Par ailleurs, $\tilde{H} = \langle \tilde{p}, \tilde{f}(t, x, u) \rangle = pf + p^0 f$, donc

$$\frac{\partial \tilde{H}}{\partial u} = 0 = p \frac{\partial f}{\partial u} + p^0 \frac{\partial f^0}{\partial u}.$$

Finalement on a obtenu l'énoncé suivant.

Théorème 7.1.2 (Principe du maximum faible)**.** *Si le contrôle u associé au système de contrôle (7.1) est optimal pour le coût (7.2), alors il existe une application $p(\cdot)$ absolument continue sur $[0, T]$, à valeurs dans $\mathbb{R}^n$, appelée vecteur adjoint, et un réel $p^0 \leqslant 0$, tels que le couple $(p(\cdot), p^0)$ est non trivial, et les équations suivantes sont vérifiées pour presque tout $t \in [0, T]$*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t, x(t), p(t), p^0, u(t)), \tag{7.6}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t), p^0, u(t)), \tag{7.7}$$

$$\frac{\partial H}{\partial u}(t, x(t), p(t), p^0, u(t)) = 0, \tag{7.8}$$

où H est le Hamiltonien associé au système (7.1) et au coût (7.2)

$$H(t, x, p, p^0, u) = \langle p, f(t, x, u) \rangle + p^0 f^0(t, x, u). \tag{7.9}$$

7.1.2 Le problème de Mayer-Lagrange

On modifie le problème précédent en introduisant le coût

$$C(t, u) = \int_0^t f^0(s, x_u(s), u(s)) ds + g(t, x_u(t)), \tag{7.10}$$

et où le temps final t n'est pas fixé. Soit M_1 une variété de $\mathbb{R}^n$. Le problème de contrôle optimal est alors de déterminer une trajectoire solution de

$$\dot{x}(t) = f(t, x(t), u(t)), \ x(0) = x_0,$$

où les contrôles $u(\cdot)$ sont dans l'ensemble $\mathcal{U}$ des contrôles admissibles sur $[0, t_e(u)[$, telle que $x(T) \in M_1$, et de plus $x(\cdot)$ minimise sur $[0, T]$ le coût (7.10).

Supposons que la variété M_1 est donnée par

$$M_1 = \{x \in \mathbb{R}^n \mid F(x) = 0\},$$

où F est une fonction de classe C^1 de $\mathbb{R}^n$ dans $\mathbb{R}^p$ (submersive donc, puisque M_1 est une variété). En écrivant $F = (F_1, \ldots, F_p)$ où les fonctions F_i sont à valeurs réelles, il vient

$$M_1 = \{x \in \mathbb{R}^n \mid F_1(x) = \cdots = F_p(x) = 0\},$$

et de plus l'espace tangent à M_1 en un point $x \in M_1$ est

$$T_x M_1 = \{v \in \mathbb{R}^n \mid \nabla F_i(x).v = 0,\ i = 1, \ldots, p\}.$$

Introduisons alors l'application

$$h(t, u) = (F \circ E(t, u), C(t, u)).$$

Remarque 7.1.1. L'application h n'est pas forcément différentiable au sens de Fréchet. Cela dépend en effet de la régularité du contrôle u. Si par exemple u est continu en t, alors

$$\frac{\partial E}{\partial t}(t, u) = f(t, x(t), u(t)).$$

Dans les calculs qui suivent, on oublie cette difficulté et on suppose que h est différentiable.

Le fait suivant est une conséquence immédiate du théorème des fonctions implicites.

Lemme 7.1.3. *Si un contrôle u est optimal sur $[0, T]$ alors l'application h n'est pas submersive au point (T, u).*

Par conséquent dans ce cas l'application $dh(T, u)$ n'est pas surjective, et donc il existe un vecteur non trivial $\tilde{\psi}_1 = (\psi_1, \psi^0) \in \mathbb{R}^n \times \mathbb{R}$ qui est orthogonal dans $\mathbb{R}^{p+1}$ à Im $dh(T, u)$, *i.e.*

$$\tilde{\psi}_1 dh(T, u) = 0.$$

Ceci implique les deux égalités au point (T, u)

$$\psi_1 \frac{\partial}{\partial t} F \circ E + \psi^0 \frac{\partial C}{\partial t} = 0, \tag{7.11}$$

$$\psi_1 \frac{\partial}{\partial u} F \circ E + \psi^0 \frac{\partial C}{\partial u} = 0. \tag{7.12}$$

Posons

$$C_0(t, u) = \int_0^t f^0(s, x_u(s), u(s)) ds,$$

de sorte que

$$C(t, u) = C_0(t, u) + g(t, x_u(t)) = C_0(t, u) + g(t, E(t, u)).$$

Avec cette notation il vient, compte-tenu de $\frac{\partial C_0}{\partial t} = f^0$ et $\frac{\partial E}{\partial t} = f$,

$$\frac{\partial C}{\partial t} = f^0 + \frac{\partial g}{\partial t} + \frac{\partial g}{\partial x} f,$$

et

$$\frac{\partial C}{\partial u} = \frac{\partial C_0}{\partial u} + \frac{\partial g}{\partial x}\frac{\partial E}{\partial u},$$

au point (T, u). En reportant dans les relations (7.11) et (7.12) on obtient

$$\psi f + \psi^0 (f^0 + \frac{\partial g}{\partial t}) = 0, \tag{7.13}$$

$$\psi \frac{\partial E}{\partial u} + \psi^0 \frac{\partial C_0}{\partial u} = 0, \tag{7.14}$$

au point (T, u), où par définition

$$\psi = \psi_1 . \nabla F + \psi^0 \frac{\partial g}{\partial x}.$$

En particulier si on pose $\psi_1 = (\lambda_1, \ldots, \lambda_p)$, on obtient $\psi_1 . \nabla F = \sum_{i=1}^{p} \lambda_i \nabla F_i$.

Remarque 7.1.2. Si on envisage le problème de Mayer-Lagrange *à temps final fixé* T, alors on considère le coût

$$C_T(u) = \int_0^T f^0(s, x_u(s), u(s))ds + g(x_u(T)).$$

Le raisonnement précédent est quasiment inchangé, sauf que l'on raisonne sur l'application à temps fixé T

$$h_T(u) = (F \circ E_T(u), C_T(u)),$$

et on obtient de même la relation (7.14). En revanche on n'a plus l'équation (7.13).

Ainsi l'équation (7.13) traduit-elle le fait que le temps final n'est pas fixé.

Remarque 7.1.3. La relation (7.14) affirme exactement que le contrôle u est singulier sur $[0, T]$ pour le système $\dot{x} = f(t, x, u)$ affecté du coût $C_0(u)$. Autrement dit on s'est ramené à un problème de Lagrange à temps non fixé.

En particulier en appliquant la proposition 5.3.4, on obtient, similairement au paragraphe précédent, le résultat suivant.

Théorème 7.1.4 (Principe du Maximum faible, cas de Mayer-Lagrange)**.** *Si le contrôle u est optimal sur $[0, T]$ alors il existe une application $p : [0, T] \to \mathbb{R}^n \setminus \{0\}$ absolument continue, et un réel $p^0 \leqslant 0$, tels que le couple $(p(\cdot), p^0)$ est non trivial, et*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(t, x(t), p(t), p^0, u(t)),\ \ \dot{p}(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t), p^0, u(t)), \tag{7.15}$$

$$\frac{\partial H}{\partial u}(t, x(t), p(t), p^0, u(t)) = 0, \tag{7.16}$$

où $H(t, x, p, p^0, u) = \langle p, f(t, x, u)\rangle + p^0 f^0(t, x, u)$.

Si de plus la cible M_1 *est une sous-variété de* $\mathbb{R}^n$ *alors il existe des réels* $\lambda_1, \ldots, \lambda_p$, *tels que l'on ait au point final* (T, x_1)

$$p(T) = \sum_{i=1}^{p} \lambda_i \nabla F_i + p^0 \frac{\partial g}{\partial x}. \tag{7.17}$$

De plus si le temps final n'est pas fixé dans le problème de contrôle optimal, et si u *est continu au temps* T, *alors on a au temps final* T

$$H(T, x(T), p(T), p^0, u(T)) = -p^0 \frac{\partial g}{\partial t}(T, x(T)).$$

7.2 Principe du maximum de Pontryagin

La version forte suivante, beaucoup plus difficile à montrer, du théorème précédent (voir [60] pour une démonstration, voir aussi [13, 39, 52]), prend en compte les contraintes sur le contrôle, et affirme que cet extremum est un maximum. On a l'énoncé général suivant.

7.2.1 Enoncé général

Théorème 7.2.1. *On considère le système de contrôle dans* $\mathbb{R}^n$

$$\dot{x}(t) = f(t, x(t), u(t)), \tag{7.18}$$

où $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ *est de classe* C^1 *et où les contrôles sont des applications mesurables et bornées définies sur un intervalle* $[0, t_e(u)[$ *de* $\mathbb{R}^+$ *et à valeurs dans* $\Omega \subset \mathbb{R}^m$. *Soient* M_0 *et* M_1 *deux sous-ensembles de* $\mathbb{R}^n$. *On note* $\mathcal{U}$ *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* $t(u) < t_e(u)$.
Par ailleurs on définit le coût d'un contrôle u *sur* $[0, t]$

$$C(t, u) = \int_0^t f^0(s, x(s), u(s))ds + g(t, x(t)),$$

où $f^0 : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ *et* $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ *sont* C^1, *et* $x(\cdot)$ *est la trajectoire solution de (7.18) associée au contrôle* u.

On considère le problème de contrôle optimal suivant : déterminer une trajectoire reliant M_0 *à* M_1 *et minimisant le coût. Le temps final peut être fixé ou non.*

Si le contrôle $u \in \mathcal{U}$ *associé à la trajectoire* $x(\cdot)$ *est optimal sur* $[0, T]$, *alors il existe une application* $p(\cdot) : [0, T] \longrightarrow \mathbb{R}^n$ *absolument continue appelée vecteur*

adjoint, et un réel $p^0 \leqslant 0$, tels que le couple $(p(\cdot), p^0)$ est non trivial, et tels que, pour presque tout $t \in [0,T]$,

$$\begin{aligned} \dot{x}(t) &= \frac{\partial H}{\partial p}(t, x(t), p(t), p^0, u(t)), \\ \dot{p}(t) &= -\frac{\partial H}{\partial x}(t, x(t), p(t), p^0, u(t)), \end{aligned} \tag{7.19}$$

où $H(t,x,p,p^0,u) = \langle p, f(t,x,u)\rangle + p^0 f^0(t,x,u)$ est le Hamiltonien du système, et on a la condition de maximisation presque partout sur $[0,T]$

$$H(t, x(t), p(t), p^0, u(t)) = \max_{v\in\Omega} H(t, x(t), p(t), p^0, v). \tag{7.20}$$

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

$$\max_{v\in\Omega} H(T, x(T), p(T), p^0, v) = -p^0 \frac{\partial g}{\partial t}(T, x(T)). \tag{7.21}$$

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 $x(0) \in M_0$ et $x(T) \in M_1$, alors le vecteur adjoint peut être construit de manière à vérifier les conditions de transversalité aux deux extrémités (ou juste l'une des deux)

$$p(0) \perp T_{x(0)} M_0 \tag{7.22}$$

et

$$p(T) - p^0 \frac{\partial g}{\partial x}(T, x(T)) \perp T_{x(T)} M_1. \tag{7.23}$$

Remarque 7.2.1. Si le contrôle u est continu au temps T, la condition (7.21) peut s'écrire

$$H(T, x(T), p(T), p^0, u(T)) = -p^0 \frac{\partial g}{\partial t}(T, x(T)). \tag{7.24}$$

Remarque 7.2.2. Si la variété M_1 s'écrit sous la forme

$$M_1 = \{x \in \mathbb{R}^n \mid F_1(x) = \cdots = F_p(x) = 0\},$$

où les F_i sont des fonctions de classe C^1 sur $\mathbb{R}^n$ (indépendantes puisque M_1 est une variété), alors la condition (7.23) se met sous la forme

$$\exists \lambda_1, \ldots, \lambda_p \in \mathbb{R} \mid p(T) = \sum_{i=1}^{p} \lambda_i \nabla F_i(x(T)) + p^0 \frac{\partial g}{\partial x}(T, x(T)). \tag{7.25}$$

Remarque 7.2.3. Dans les conditions du théorème, on a de plus pour presque tout $t \in [0,T]$

$$\frac{d}{dt} H(t, x(t), p(t), p^0, u(t)) = \frac{\partial H}{\partial t}(t, x(t), p(t), p^0, u(t)). \tag{7.26}$$

En particulier si le système augmenté est *autonome*, *i.e.* si f et f^0 ne dépendent pas de t, alors H ne dépend pas de t, et on a

$$\forall t \in [0,T] \quad \max_{v\in\Omega} H(x(t),p(t),p^0,v) = \text{Cste}.$$

Notons que cette égalité est alors valable partout sur $[0,T]$ (en effet cette fonction de t est lipschitzienne).

Remarque 7.2.4. La convention $p^0 \leqslant 0$ conduit au principe du *maximum*. La convention $p^0 \geqslant 0$ conduirait au principe du *minimum*, *i.e.* la condition (7.20) serait une condition de minimum.

Remarque 7.2.5. Dans le cas où $\Omega = \mathbb{R}^m$, *i.e.* lorsqu'il n'y a pas de contrainte sur le contrôle, la condition de maximum (7.20) devient $\frac{\partial H}{\partial u} = 0$, et on retrouve le principe du maximum faible (théorème 7.1.2).

Définition 7.2.1. Une *extrémale* du problème de contrôle optimal est un quadruplet $(x(\cdot),p(\cdot),p^0,u(\cdot))$ solution des équations (7.19) et (7.20). Si $p_0 = 0$, on dit que l'extrémale est *anormale*, et si $p^0 \neq 0$ l'extrémale est dite *normale*.

Remarque 7.2.6. Lorsque $\Omega = \mathbb{R}^m$, *i.e.* lorsqu'il n'y a pas de contrainte sur le contrôle, alors la trajectoire $x(\cdot)$, associée au contrôle $u(\cdot)$, est une trajectoire singulière du système (7.1), si et seulement si elle est projection d'une extrémale anormale $(x(\cdot),p(\cdot),0,u(\cdot))$.

Ceci résulte en effet de la caractérisation hamiltonienne des trajectoires singulières, cf proposition 5.3.4. Remarquons que puisque $p^0 = 0$, ces trajectoires ne dépendent pas du coût. Elles sont intrinsèques au système. Le fait qu'elles puissent pourtant être optimales s'explique de la manière suivante : en général, une trajectoire singulière a une propriété de *rigidité*, *i.e.* c'est la seule trajectoire joignant ses extrémités, et donc en particulier elle est optimale, ceci indépendamment du critère d'optimisation choisi.

Ce lien entre extrémales anormales et trajectoires singulières, pour $\Omega = \mathbb{R}^m$, montre bien la difficulté liée à l'existence éventuelle de telles trajectoires.

Définition 7.2.2. Les conditions (7.22) et (7.23) sont appelées *conditions de transversalité sur le vecteur adjoint*. La condition (7.21) est appelée *condition de transversalité sur le Hamiltonien*. Elles sont ici écrites de manière très générale, et dans les deux paragraphes suivants nous allons les réécrire dans des cas plus simples.

Remarque 7.2.7. Le problème important du *temps minimal* correspond à $f^0 = 1$ et $g = 0$, ou bien à $f^0 = 0$ et $g(t,x) = t$. Dans les deux cas les conditions de transversalité obtenues sont bien les mêmes.

Remarque 7.2.8. Il existe des versions plus générales du principe du maximum, pour des dynamiques non lisses ou hybrides (voir par exemple [22, 69, 70] et leurs références, voir aussi plus loin pour le principe du maximum avec contraintes sur l'état).

7.2.2 Conditions de transversalité

Conditions de transversalité sur le vecteur adjoint

Dans ce paragraphe le temps final pour atteindre la cible peut être fixé ou non. Réécrivons les conditions (7.22) et (7.23) dans les deux cas importants suivants.

– **Problème de Lagrange.** Dans ce cas le coût s'écrit

$$C(t,u) = \int_0^t f^0(s,x(s),u(s))ds,$$

i.e. $g = 0$. Les conditions de transversalité (7.22) et (7.23) sur le vecteur adjoint s'écrivent alors

$$p(0) \perp T_{x(0)}M_0, \quad p(T) \perp T_{x(T)}M_1. \tag{7.27}$$

Remarque 7.2.9. Si par exemple $M_0 = \{x_0\}$, la condition (7.22) devient vide. Si au contraire $M_0 = \mathbb{R}^n$, *i.e.* si le point initial n'est pas fixé, on obtient $p(0) = 0$.
De même, si $M_1 = \mathbb{R}^n$, on obtient $p(T) = 0$. Autrement dit *si le point final est libre alors le vecteur adjoint au temps final est nul.*

– **Problème de Mayer.** Dans ce cas le coût s'écrit

$$C(t,u) = g(t,x(t)),$$

i.e. $f^0 = 0$. Les conditions de transversalité (7.22) et (7.23) (ou (7.25)) ne se simplifient pas a priori.
Mais dans le cas particulier important où $M_1 = \mathbb{R}^n$, autrement dit *le point final* $x(T)$ *est libre*, la condition (7.23) devient

$$p(T) = p^0 \frac{\partial g}{\partial x}(T,x(T)), \tag{7.28}$$

et alors forcément $p^0 \neq 0$ (on prend alors $p^0 = -1$). Si de plus g ne dépend pas du temps, on a coutume d'écrire $p(T) = -\nabla g(x(T))$.

Condition de transversalité sur le Hamiltonien

La condition (7.21) n'est valable que si le temps final pour atteindre la cible n'est pas fixé. Dans ce paragraphe nous nous plaçons donc dans ce cas.

La seule simplification notable de cette condition est le cas où la fonction g ne dépend pas du temps t (ce qui est vrai par exemple pour un problème de Lagrange), et la condition de transversalité (7.21) sur le Hamiltonien devient alors

$$\max_{v\in\Omega} H(T,x(T),p(T),p^0,v) = 0, \tag{7.29}$$

ou encore, si u est continu au temps T,

$$H(T,x(T),p(T),p^0,u(T)) = 0. \tag{7.30}$$

Autrement dit *le Hamiltonien s'annule au temps final.*

Remarque 7.2.10. Si le système augmenté est de plus autonome, *i.e.* si f et f^0 ne dépendent pas de t, alors d'après la remarque 7.2.3 on a le long d'une extrémale

$$\forall t \in [0,T] \quad \max_{v\in\Omega} H(x(t),p(t),p^0,v) = 0.$$

Généralisation des conditions de transversalité

Pour écrire les conditions de transversalité associées à un problème de contrôle plus général, il faut écrire les relations adéquates en termes de multiplicateurs de Lagrange.

Par exemple considérons un problème de Lagrange avec des conditions aux limites mélangées, *i.e.* on cherche une trajectoire solution de

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

minimisant le coût

$$C(T,u) = \int_0^T f^0(t,x(t),u(t))dt,$$

et vérifiant les conditions aux limites

$$(x(0),x(T)) \in M,$$

où M est une sous-variété de $\mathbb{R}^n \times \mathbb{R}^n$.

On peut alors montrer (voir [2]) que dans ce cas les conditions de transversalité (7.22) et (7.23) sur le vecteur adjoint s'écrivent

$$(-p(0),p(T)) \perp T_{(x(0),x(T))}M.$$

Un cas important de conditions mélangées est le cas des trajectoires périodiques, *i.e.* $x(0) = x(T)$ non fixé. Dans ce cas on a

$$M = \{(x,x) \mid x \in \mathbb{R}^n\},$$

et la condition de transversalité donne

$$p(0) = p(T).$$

Autrement dit, non seulement la trajectoire est périodique, mais aussi son relèvement extrémal.

7.2.3 Contraintes sur l'état

Principe du maximum avec contrainte sur l'état. Le principe du maximum tel qu'il vient d'être énoncé prend en compte des contraintes sur le contrôle, mais ne prend pas en compte d'éventuelles contraintes sur l'état. Ce problème est en effet beaucoup plus difficile. Il existe plusieurs versions du principe du maximum avec contraintes sur l'état (voir à ce sujet [21, 22, 36, 42, 55, 56]). La théorie est cependant beaucoup plus compliquée, et nous ne l'abordons pas

dans cet ouvrage. Une différence fondamentale avec le principe du maximum classique est que la présence de contraintes sur l'état peut rendre le vecteur adjoint discontinu. On rajoute alors des conditions de saut, ou de jonction.

En fait, lorsqu'il existe des contraintes sur l'état de la forme $c_i(x) \leqslant 0$, $i = 1, \ldots, p$, où les fonctions $c_i : \mathbb{R}^n \to \mathbb{R}$ sont de classe C^1, alors le vecteur adjoint $p(\cdot)$ est solution de l'équation intégrale

$$p(t) = p(T) + \int_t^T \frac{\partial H}{\partial x} dt - \sum_{i=1}^{p} \int_t^T \frac{\partial c_i}{\partial x} d\mu_i,$$

où les μ_i sont des mesures positives ou nulles dont le support est contenu dans $\{t \in [0, T] \mid c_i(x(t)) = 0\}$.

Dans la section 7.4, on traite complètement un exemple (simplifié) de problème de contrôle optimal où apparaissent des contraintes sur l'état (problème de rentrée atmosphérique d'une navette). Cependant on arrive à éviter l'usage d'un principe du maximum avec contraintes.

Méthode de pénalisation. Un moyen simple de manipuler des contraintes sur l'état est de résoudre un problème de contrôle optimal modifié, où, comme dans la théorie LQ, on pondère cette contrainte, de manière à la forcer à être vérifiée. Le principe général de cette méthode est le suivant. Supposons qu'on veuille imposer à l'état d'appartenir à un sous-ensemble $C \subset \mathbb{R}^n$. Donnons-nous une fonction g sur $\mathbb{R}^n$, nulle sur C et strictement positive ailleurs (il faut être capable de construire une telle fonction). Alors, en ajoutant au coût $C(t, u)$ le scalaire $\lambda \int_0^T g(x(t))dt$, où $\lambda > 0$ est un poids que l'on peut choisir assez grand, on espère que la résolution de ce problème de contrôle optimal modifié va forcer la trajectoire à rester dans l'ensemble C. En effet si $x(t)$ sort de l'ensemble C, et si λ est grand, alors le coût correspondant est grand, et probablement la trajectoire ne sera pas optimale.

La justification théorique de ce procédé réside dans la proposition générale suivante.

Proposition 7.2.2. *Soit (E, d) un espace métrique, C un sous-ensemble de E, et f une fonction k-lipschitzienne sur E. Pour tout $x \in E$, posons $g(x) = d(x, C)$. Supposons que la fonction f restreinte à C atteint son minimum en $x_0 \in C$, i.e.*

$$f(x_0) = \min_{x \in C} f(x).$$

Alors, pour tout réel $\lambda \geqslant k$, on a

$$f(x_0) + \lambda g(x_0) = \min_{x \in C} f(x) + \lambda g(x),$$

i.e. x_0 est aussi un point où $f + \lambda g$ atteint son minimum sur C. La réciproque est vraie si de plus $\lambda > k$ et si C est fermé.

Démonstration. Raisonnons par l'absurde, et supposons qu'il existe $y \in E$ et $\varepsilon > 0$ tels que $f(y) + \lambda d(y, C) < f(x_0) - \lambda\varepsilon$. Soit alors $z \in E$ tel que $d(y, z) \leqslant d(y, C) + \varepsilon$. On a

$$f(z) \leqslant f(y) + kd(y, z) \leqslant f(y) + \lambda d(y, C) + \lambda\varepsilon < f(x_0),$$

ce qui est une contradiction.

Pour la réciproque, supposons que $\lambda > k$ et que C est fermé. Soit $x_0 \in C$ un point où $f + \lambda g$ atteint son minimum sur C, et soit $\varepsilon > 0$. Il existe $z \in C$ tel que $d(x_0, z) < d(x_0, C) + \varepsilon/\lambda$. On a

$$\begin{aligned} f(z) &\leqslant f(x_0) + kd(x_0, z) \\ &\leqslant f(x_0) + kd(x_0, C) + \frac{k}{\lambda}\varepsilon \\ &< f(x_0) + \lambda d(x_0, C) - (\lambda - k)d(x_0, C) + \varepsilon \\ &< f(z) - (\lambda - k)d(x_0, C) + \varepsilon \end{aligned}$$

et donc $(\lambda - k)d(x_0, C) < \varepsilon$. Le réel $\varepsilon > 0$ étant arbitraire, on en déduit que $d(x_0, C) = 0$, et donc $x_0 \in C$ puisque C est fermé. On conclut que pour tout $z \in C$ on a $f(z) \geqslant f(x_0)$. □

7.3 Exemples et exercices

7.3.1 Contrôle optimal d'un ressort non linéaire

Reprenons l'exemple, leitmotiv de ce livre, du ressort non linéaire vu en introduction, modélisé par le système de contrôle

$$\begin{cases} \dot{x}(t) = y(t), \\ \dot{y}(t) = -x(t) - 2x(t)^3 + u(t), \end{cases}$$

où on autorise comme contrôles toutes les fonctions $u(t)$ continues par morceaux telles que $|u(t)| \leqslant 1$. L'objectif est d'amener le ressort d'une position initiale quelconque $(x_0, y_0 = \dot{x}_0)$ à sa position d'équilibre $(0, 0)$ *en temps minimal* t_*.

Application du Principe du Maximum

Le Hamiltonien du système précédent s'écrit

$$H(x, p, u) = p_x y + p_y(-x - 2x^3 + u),$$

et si (x, p, u) est une extrémale alors on doit avoir

$$\dot{p}_x = -\frac{\partial H}{\partial x} = p_y(1 + 6x^2), \text{ et } \dot{p}_y = -\frac{\partial H}{\partial y} = -p_x.$$

Notons que puisque le vecteur adjoint (p_x, p_y) doit être non trivial, p_y ne peut s'annuler sur un intervalle (sinon on aurait également $p_x = -\dot{p}_y = 0$). Par ailleurs la condition la condition de maximisation nous donne

$$p_y u = \max_{|v|\leqslant 1} p_y v.$$

Comme p_y ne s'annule sur aucun intervalle, on en déduit que, presque partout,

$$u(t) = \text{signe } p_y(t).$$

En particulier les contrôles optimaux sont successivement égaux à ± 1, c'est le principe *bang-bang* (voir [52]). Plus précisément, le vecteur adjoint au temps final t_* étant défini à scalaire multiplicatif près, on peut affirmer

$$u(t) = \text{signe}(p_y(t)) \text{ où } p_y \text{ est la solution de } \begin{cases} \ddot{p}_y(t) + p_y(t)\,(1 + 6x(t)^2) = 0, \\ p_y(t_*) = \cos\alpha,\ \dot{p}_y(t_*) = -\sin\alpha, \end{cases}$$

le paramètre $\alpha \in [0, 2\pi[$ étant indéterminé.

En inversant le temps $(t \mapsto -t)$, il est clair que notre problème est équivalent au problème du temps minimal pour le système

$$\begin{cases} \dot{x}(t) = -y(t) \\ \dot{y}(t) = x(t) + 2x(t)^3 - \text{signe}(p_y(t)) \\ \dot{p}_y(t) = p_x(t) \\ \dot{p}_x(t) = -p_y(t)(1 + 6x(t)^2) \end{cases} \tag{7.31}$$

avec

$$x(0) = y(0) = 0,\ x(t_*) = x_0,\ y(t_*) = y_0,\ p_y(0) = \cos\alpha,\ p_x(0) = \sin\alpha,$$

où $\alpha \in [0, 2\pi[$ est à déterminer.

Résolution numérique à l'aide de *Maple*

On suppose désormais que $x_0 = 0$ et $\dot{x}_0 = 6$.

Pour résoudre le problème on procède en 5 étapes.

Première étape. On saisit le système différentiel (7.31), puis on trace dans le plan de phase (x, y) les deux solutions respectivement associées à $\alpha = 1$ et $\alpha = 2.5$, avec $t \in [0, 10]$ (voir figure 7.2).

```
> eq1 := D(x)(t)=-y(t) :
  eq2 := D(y)(t)=x(t)+2*x(t)^3-signum(z(t)) :
  eq3 := D(z)(t)=w(t) :
  eq4 := D(w)(t)=-z(t)*(1+6*x(t)^2) :
  sys := eq1,eq2,eq3,eq4 :
> ic1 := [x(0)=0,y(0)=0,z(0)=cos(1),w(0)=sin(1)] :
```

```
  ic2 := [x(0)=0,y(0)=0,z(0)=cos(2.5),w(0)=sin(2.5)] :
  ic  := ic1,ic2 :
> DEplot([sys], [x(t),y(t),z(t),w(t)], t=0..10, [ic],
    stepsize=0.05, scene=[x(t),y(t)],linecolor=[blue,red]);
```

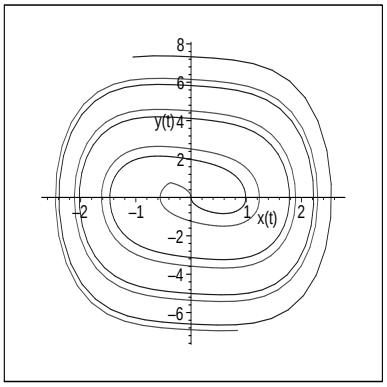


FIGURE 7.2 –

Deuxième étape. On pose $T = 10$, $N = 100$, $h = T/N$ et $t_n = nh, n = 0 \dots N$. Pour $\alpha = 1$, puis pour $\alpha = 2.5$, on écrit une boucle qui calcule le plus petit entier k tel que

$$x(t_k)x(t_{k+1}) \leqslant 0 \quad \text{et} \quad |y(t_{k+1}) - 6| < 0.5.$$

On affiche alors les valeurs de la solution aux temps t_k et t_{k+1}.

```
> sol1 := dsolve({sys,x(0)=0,y(0)=0,z(0)=cos(1),w(0)=sin(1)},
                          {x(t),y(t),z(t),w(t)}, type=numeric) :
  T:=10.0 : N:=100 : h:=T/N :
  xk:=0 :
  for k from 1 to N do
    solk := sol1(k*h) :
    xknew := subs(solk,x(t)) :
    yknew := subs(solk,y(t)) :
    if xk*xknew<=0 and abs(yknew-6)<0.5 then break fi:
    xk := xknew :
  od: sol1(k*h);
```

Troisième étape. On écrit une procédure *temps :=proc(alpha,eps)* qui calcule une approximation du temps t tel que $x(t) = 0$ et $|y(t) - 6| < 0.5$. Pour cela on localise tout d'abord ce temps comme à l'étape précédente, puis on effectue une dichotomie sur t entre t_k et t_{k+1} pour calculer le temps où $x(t)$ s'annule à *eps* près (c'est-à-dire $|x(t)| < eps$).

```
> temps := proc(alpha,eps)
  local sol,solk,T,N,h,k,xk,xknew,yknew,t0,t1,tm,x0,x1,xm :
  sol := dsolve({sys,x(0)=0,y(0)=0,z(0)=cos(alpha),
     w(0)=sin(alpha)}, {x(t),y(t),z(t),w(t)}, type=numeric) :
  T:=10.0 : N:=100 : h:=T/N :
  xk:=0 :
  for k from 1 to N do
    solk:=sol(k*h) :
    xknew := subs(solk,x(t)) :
    yknew := subs(solk,y(t)) :
    if xk*xknew<=0 and abs(yknew-6)<0.5 then break fi:
    xk := xknew :
  od:
  t0:=(k-1)*h : t1:=k*h :
  x0:=subs(sol(t0),x(t)) : x1:=subs(sol(t1),x(t)) :
     # remarque : x0 et x1 sont forcement de signes contraires
  while abs(x1-x0)>eps do
    tm:=(t0+t1)/2 :
    xm:=subs(sol(tm),x(t)) :
    if xm*x0<0 then x1:=xm : t1:=tm :
               else x0:=xm : t0:=tm :
    fi:
  od:
  RETURN(t0);
  end :
```

Quatrième étape. On écrit une procédure *dicho=proc(eps)* qui calcule par dichotomie sur α, entre $\alpha = 1$ et $\alpha = 2.5$, une approximation du réel α tel que la solution de (7.31) associée vérifie

$$\exists t_* \mid x(t_*) = 0,\ y(t_*) = 6,$$

le réel *eps* étant la précision, *i.e.* $|x(t_*)| < eps,\ |y(t_*) - 6| < eps$.
Plus précisément, on cherche le réel α par dichotomie de sorte que

$$|\, y(\alpha,\, temps(\alpha, eps)\,) - 6\,| < eps$$

où $\big(x(\alpha,\cdot), y(\alpha,\cdot), z(\alpha,\cdot), w(\alpha,\cdot)\big)$ est la solution de (7.31) (notons que la procédure *temps* assure déjà que $|x(\alpha, temps(\alpha, eps))| < eps$).

```
> dicho := proc(eps)
  local a,b,m,sola,solb,solm,ta,tb,tm,ya,yb,ym :
  a:=1 : b:=2.5 :
  sola := dsolve({sys,x(0)=0,y(0)=0,z(0)=cos(a),w(0)=sin(a)},
                       {x(t),y(t),z(t),w(t)}, type=numeric) :
  solb := dsolve({sys,x(0)=0,y(0)=0,z(0)=cos(b),w(0)=sin(b)},
                       {x(t),y(t),z(t),w(t)}, type=numeric) :
```

```
  ta:=temps(a,eps) : tb:=temps(b,eps) :
  ya:=subs(sola(ta),y(t)) : yb:=subs(solb(tb),y(t)) :
  while abs(yb-ya)>eps do
    m:=evalf((a+b)/2) :
    solm := dsolve({sys,x(0)=0,y(0)=0,z(0)=cos(m),w(0)=sin(m)},
                          {x(t),y(t),z(t),w(t)}, type=numeric) :
    tm:=temps(m,eps) :
    ym := subs(solm(tm),y(t)) :
    if (ym-6)*(ya-6)<0 then b:=m : yb:=ym :
                       else a:=m : ya:=ym :
    fi:
  od:
  RETURN(a);
  end:
```

Cinquième étape. On calcule une approximation de α pour $eps = 0.01$, et on trace dans le plan de phase la solution obtenue (voir figure 7.3).

```
> dicho(0.01);
                                2.136718750
> temps(2.136718750,0.01);
                                8.737500000
> DEplot([sys], [x(t),y(t),z(t),w(t)],t=0..8.7375,
   [[x(0)=0,y(0)=0,z(0)=cos(2.136718750),w(0)=sin(2.136718750)]],
              stepsize=0.05, scene=[x(t),y(t)],linecolor=[blue]);
```

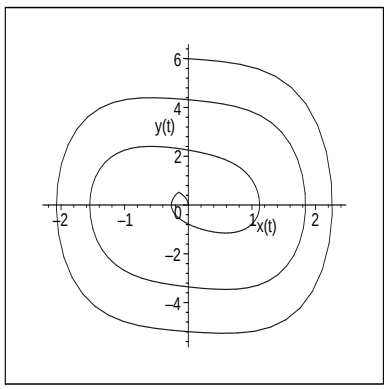


FIGURE 7.3 –

Le temps minimal pour amener le ressort de la position $(0, 6)$ à l'équilibre $(0, 0)$ est donc de 8.7375 s.

Remarque 7.3.1. Considérons le contrôle

$$u(t) = \text{signe } (y(t) - 0.1)/1.33.$$

On constate numériquement que la solution du système associée à ce contrôle passe bien par le point $(0, 6)$ au temps $t = 10.92$. Le temps qu'il faut à cette trajectoire pour aller de $(0, 0)$ au point $(0, 6)$ est bien supérieur au temps minimal calculé.

7.3.2 Exercices

Exercice 7.3.1 (Problème du temps minimal pour une fusée à mouvement rectiligne)**.** Considérons une modélisation simplifiée du mouvement rectiligne d'une fusée, *i.e.*

$$\dot{x}(t) = u(t),\ \dot{y}(t) = u(t)^2,$$

où $x(t)$ représente la vitesse et $y(t)$ est inversement proportionnelle à la masse de l'engin. Le contrôle $u(t)$ est la poussée et vérifie la contrainte $|u(t)| \leqslant 1$.

Résoudre le problème du temps minimal pour atteindre le point final (x_1, y_1), en partant de l'origine.

Indications : Le Hamiltonien est $H = p_x u + p_y u^2 + p^0$, où p_x et p_y sont constantes. Quelle que soit la valeur de p^0, il faut maximiser $p_x u + p_y u^2$, pour $-1 \leqslant u \leqslant 1$. Montrer que, selon les signes de p_x et p_y, le contrôle u est constant, et prend ses valeurs dans $\{-1, 1, -\frac{p_x}{2p_y}\}$.

Montrer que, pour aller en un point (x_1, x_2) tel que

- $0 < y_1 < x_1$, il existe un seul contrôle optimal, singulier et constant ;
- $y_1 = |x_1|$, il existe un seul contrôle optimal, constant, égal à 1 ou à -1 ;
- $y_1 > x_1$, il existe une infinité de contrôles optimaux, qui sont des successions d'arcs ± 1. Remarquer aussi que le temps minimal est $t_f = y_1$. En effet,

$$t_f = \int_0^{t_f} dt = \int_0^{y_1} \frac{dt}{dy} dy = \int_0^{y_1} \frac{dy}{u^2} = \int_0^{y_1} dy = y_1.$$

Noter qu'il n'y a pas unicité de la trajectoire optimale dans cette zone.

Exercice 7.3.2 (Problème de Zermelo)**.** Le mouvement d'une barque se déplaçant à vitesse constante sur une rivière où il y a un courant $c(y)$ est modélisé par

$$\begin{aligned} \dot{x}(t) &= v \cos u(t) + c(y(t)),\ x(0) = 0, \\ \dot{y}(t) &= v \sin u(t),\ y(0) = 0, \end{aligned}$$

où v est la vitesse et $u(t)$, l'angle de la barque par rapport à l'axe $(0x)$, est le contrôle.

1. Supposons que pour tout y on ait $c(y) > v$. Quelle est la loi optimale permettant de minimiser le déport $x(t_f)$ pour atteindre la berge opposée ?

2. Résoudre le problème de temps minimal pour atteindre la berge opposée.
3. Résoudre le problème de temps minimal pour atteindre un point M de la berge opposée.

Indications :

1. On a $H = p_x(v\cos u + c(y)) + p_y v \sin u$, et $p_x = -1$, $H(t_f) = 0$. On trouve
$$u = \text{Arccos}\left(-\frac{v}{c(y)}\right).$$
2. On a $H = p_x(v\cos u + c(y)) + p_y v \sin u + p^0$, et $p_x = 0$, $H(t_f) = 0$, puis $u = \frac{\pi}{2}$.
3. On a $H = p_x(v\cos u + c(y)) + p_y v \sin u + p^0$, et $p_x = \text{Cste}$, $H(t_f) = 0$, puis
$$u = \text{Arccos}\ \frac{p_x v}{1 - p_x c(y)},$$
où p_x doit être choisi de manière à atteindre M (cf méthode de tir), ou bien la solution avec $p^0 = 0$ qui est la solution de 1.

Exercice 7.3.3 (Transfert optimal de fichiers informatiques)**.** Un fichier de x_0 Mo doit être transféré par le réseau. A chaque temps t on peut choisir le taux de transmission $u(t) \in [0,1]$ Mo/s, mais il en coûte $u(t)f(t)$, où $f(\cdot)$ est une fonction connue. De plus au temps final on a un coût supplémentaire γt_f^2, où $\gamma > 0$. Le système est donc

$$\dot{x} = -u,\ x(0) = x_0, x(t_f) = 0,$$

et on veut minimiser le coût

$$C(t_f, u) = \int_0^{t_f} u(t)f(t)dt + \gamma t_f^2.$$

Quelle est la politique optimale ?

Indications : On pose $f^0 = uf$ et $g = \gamma t^2$. Le Hamiltonien est $H = -pu + p^0 fu$. Puisque $\dot{p} = 0$, on a $p(t) = \text{Cste} = p$. Par ailleurs, $u(t) = 0$ si $-p + p^0 f(t) < 0$, et $u(t) = 1$ si $-p + p^0 f(t) > 0$ (et $u(t)$ est indéterminé si $-p + p^0 f(t) = 0$ sur un sous-intervalle). Au temps final, on a
$$H(t_f) = -p^0 \frac{\partial g}{\partial t} = -2p^0 \gamma t_f,$$
d'où
$$-u(t_f)(p + f(t_f)) = -2p^0 \gamma t_f.$$
Si $p^0 = 0$, alors forcément $p \neq 0$, et $u(t)$ est constant, donc nécessairement $u(t) = 1$, mais alors la relation ci-dessus implique $p = 0$, ce qui est absurde. Donc $p^0 = -1$. Il est clair qu'au temps final t_f on a

$u(t_f) = 1$ (sinon u ne serait pas optimal, à cause du terme γt_f^2), et donc $p = -2\gamma t_f - f(t_f)$. Finalement,

$$u(t) = \begin{cases} 0 & \text{si } f(t) > -p, \\ 1 & \text{si } f(t) < -p, \end{cases}$$

avec $p = -2\gamma t_f - f(t_f)$. Notons que p et t_f sont nos deux degrés de liberté (paramétres de tir) déterminés par les équations $x(t_f) = 0$ et $p = -2\gamma t_f - f(t_f)$.
Noter qu'on aurait pu mettre le coût sous la forme

$$C(t_f, u) = \int_0^{t_f} (u(t)f(t) + 2\gamma t)dt.$$

Exercice 7.3.4 (Contrôle optimal du niveau d'un réservoir)**.** On veut ajouter de l'eau dans un réservoir, de façon à atteindre le niveau d'eau h_1, en tenant compte du fait qu'il faut compenser une perte d'eau linéaire en temps. La modélisation est

$$\dot{h}(t) = u(t) - t, \ h(0) = 0,$$

où $u(t)$ est le contrôle. Quelle est la loi optimale permettant d'atteindre l'objectif en minimisant $\int_0^{t_f} u(t)^2 dt$, le temps final t_f n'étant pas fixé ?

Indications : on trouve $u(t) = 2\sqrt{\frac{2h_1}{3}}$.

Exercice 7.3.5. Le mouvement d'un missile, décrit comme une particule de masse m soumise à la gravitation et à la résistance de l'air, est donné par les équations

$$\dot{x}_1 = x_3, \ \dot{x}_2 = x_4, \ \dot{x}_3 = \alpha \cos u, \ \dot{x}_4 = \alpha \sin u,$$

où $u(t) \in \mathbb{R}$ est le contrôle. Le but est de minimiser la quantité $t_f + g(x(t_f))$, où g est une fonction de classe C^1. Montrer que le contrôle doit vérifier

$$\tan u(t) = \frac{c_1 + c_2 t}{c_3 + c_4 t},$$

où $c_1, c_2, c_3, c_4 \in \mathbb{R}$.

Indications : les équations donnent $\tan u = \frac{p_3 c}{p_4 c}$, $\dot{p}_3 = -p_1$, $\dot{p}_4 = -p_2$, avec p_1, p_2 constantes.

Exercice 7.3.6 (Un problème de Bolzano en économie)**.** Un individu dispose d'un revenu $r(t)$, $0 \leqslant t \leqslant T$, qu'il peut dépenser ou placer à la banque avec un taux d'intérêt τ. Il veut réaliser un programme de dépense $u(t)$ sur $[0, T]$ de manière à maximiser la quantité

$$\int_0^T \ln u(t) e^{-at} dt.$$

L'évolution de son avoir $x(t)$ est alors donnée par

$$\dot{x}(t) = r(t) + \tau x(t) - u(t),$$

et de plus on impose $x(T) > 0$, *i.e.* l'avoir de l'individu est positif au temps final T. Quelle est la loi optimale ?

Remarque 7.3.2. De manière générale, on appelle *problème de Bolzano* un problème de contrôle optimal où on veut maximiser un coût du type

$$C_T(u) = \sum_{i=1}^{n} c_i x_i(T).$$

Indications : Pour avoir l'existence de trajectoires optimales, il faut relaxer la contrainte $x(T) > 0$ en $x(T) \geqslant 0$. Le cas $x(T) = 0$ est alors vu comme un cas limite. On distingue deux cas :
- si $x(T) > 0$, puisqu'il est non fixé, alors $p(T) = 0$. Or $\dot{p} = -pr$, d'où $p(t) = 0$, et $H = p^0 \ln u \; e^{-at}$. La condition de maximisation sur H conduit alors à une absurdité.
- si $x(T) = 0$, on n'a aucune condition sur $p(T)$. On peut prendre $p^0 = 1$ (pas d'anormale), et on trouve $u(t) = \frac{e^{-(a+r)t}}{p(0)}$. La condition initiale $p(0)$ est déterminée en calculant $x(t)$, et en imposant $x(T) = 0$ (cf méthode de tir).

Exercice 7.3.7 (Politique optimale de pêche)**.** L'évolution d'une population de poissons $x(t)$ est modélisée par

$$\dot{x}(t) = 0.08x(t)(1 - 10^{-6}x(t)) - u(t), \; x(0) = x_0,$$

où $u(t)$, le contrôle, représente le nombre de poissons pêchés. Déterminer une politique optimale de pêche, de manière à maximiser la quantité

$$\int_0^T e^{-0.03t} \ln u(t) \; dt,$$

et à avoir au temps final $x(T) > 0$.

Indications : même raisonnement qu'à l'exercice précédent.

Exercice 7.3.8 (Investissement optimal)**.** L'évolution du revenu $r(t)$ d'une entreprise est modélisée par le système contrôlé

$$\dot{r}(t) = -2r(t) + \frac{3}{2}u(t), \; r(0) = r_0,$$

où $u(t)$, le contrôle, représente l'investissement au temps t, et vérifie la contrainte $0 \leqslant u(t) \leqslant a$. Soit $T > \frac{1}{2}\ln 3$ un réel. Déterminer la politique optimale permettant de minimiser la quantité

$$-r(T) + \int_0^T (u(t) - r(t))dt.$$

Indications : Montrer qu'il n'y a pas d'anormale, puis que u dépend du signe de $\varphi = \frac{3}{2}p - 1$, où $\dot{p} = 2p - 1$, et $p(T) = 1$. Par intégration, montrer que $\varphi(t)$ s'annule en $t_c = T - \frac{1}{2}\ln 3$, et en déduire que la politique optimale est $u = 0$ sur $[0, t_c[$, puis $u = a$ sur $]t_c, T]$.

Exercice 7.3.9 (Contrôle optimal de population dans une ruche)**.** Considérons une population d'abeilles constituée au temps t de $w(t)$ travailleuses et de $q(t)$ reines. Soit $u(t)$ le contrôle, qui représente l'effort des abeilles pour fournir des reines à la ruche. La modélisation est

$$\dot{w}(t) = au(t)w(t) - bw(t),\ \dot{q}(t) = c(1 - u(t))w(t),\ 0 \leqslant u(t) \leqslant 1,$$

où a, b, c sont des réels strictement positifs tels que $a > b$. Quel doit être le contrôle $u(t)$ pour maximiser au temps T le nombre de reines ?

Indications : Le Hamiltonien est $H = p_1(auw - bw) + p_2c(1 - u)w$, où

$$\dot{p}_1 = -p_1(au - b) - p_2c(1 - u),\ \dot{p}_2 = 0.$$

Les conditions de transversalité donnent $p_1(T) = 0$ et $p_2(T) = 1$ (donc $p_2(t) = \text{Cste} = 1$), et selon la condition de maximisation on a, puisque $w > 0$,

$$u(t) = \begin{cases} 0 & \text{si } p_1(t)a - p_2c < 0, \\ 1 & \text{si } p_1(t)a - p_2c > 0. \end{cases}$$

Au temps final T on a donc $u(T) = 0$ puisque $p_1(T)a - p_2(T)c = -c < 0$. Par continuité de la fonction de commutation, le contrôle u est nul sur un intervalle $[t_1, T]$. Sur cet intervalle, on a alors $\dot{p}_1 = p_1b - c$, d'où

$$p_1(t) = \frac{c}{b}(1 - e^{b(t-T)}),$$

et p_1 est décroissant. En raisonnant en temps inverse, on a une commutation au temps t_1 tel que $p_1(t_1)a - c = 0$, soit

$$t_1 = T + \frac{1}{b}\ln(1 - \frac{b}{a}).$$

Pour $t < t_1$, on a $\dot{p}_1 = -p_1(a-b) < 0$, donc p_1 est encore décroissant. Il n'y a donc pas d'autre commutation.
Conclusion : la politique optimale est $u(t) = 1$ sur $[0, t_1]$, puis $u(t) = 0$.

Exercice 7.3.10 (Contrôle optimal d'une réaction chimique)**.** Une réaction chimique est modélisée par

$$\dot{x}_1 = -ux_1 + u^2x_2,\ x_1(0) = 1,$$
$$\dot{x}_2 = ux_1 - 3u^2x_2,\ x_2(0) = 0,$$

où x_1, x_2 sont les concentrations des réactifs, et le contrôle $u(t)$ vérifie la contrainte $0 < u(t) \leqslant 1$. Quelle est la politique optimale permettant de maximiser la quantité finale $x_2(1)$ du second réactif ?

Indications : Le Hamiltonien s'écrit $H = p_1(-ux_1+u^2x_2)+p_2(ux_1-3u^2x_2)$, et

$$\dot{p}_1 = (p_1 - p_2)u,\ \dot{p}_2 = (-p_1 + 3p_2)u^2,\ \ p_1(1) = 0,\ p_2(1) = 1.$$

Il faut maximiser sur $]0, 1]$ la fonction $\varphi = (p_2 - p_1)x_1u + (p_1 - 3p_2)x_2u^2$.

Montrer que, compte-tenu des conditions initiales, $x_2(t)$ ne reste pas nul pour $t > 0$ petit.

Montrer que le contrôle singulier s'écrit

$$u_s = \frac{(p_1 - p_2)x_1}{2(p_1 - 3p_2)x_2},$$

avec

$$\dot{p}_1 = \frac{(p_1 - p_2)^2x_1}{2(p_1 - 3p_2)x_2},\ \dot{p}_2 = \frac{(p_1 - p_2)^2x_1^2}{4(p_1 - 3p_2)x_2^2}.$$

En déduire que forcément $p_1(0) \neq p_2(0)$, et que $u_s(t) \sim +\infty$ pour $t > 0$ petit.

En déduire que la politique optimale consiste à prendre $u = +1$ au début, puis $u = u_s$.

Exercice 7.3.11 (contrôle optimal d'une épidémie par vaccination)**.** On considère une population de N individus soumis à une épidémie qu'on veut contrôler par vaccination. Par simplicité, on suppose qu'un individu qui a été malade et soigné peut à nouveau tomber malade.

Le modèle est le suivant. On note $\alpha > 0$ le taux de contamination, $u(t)$ (contrôle) le taux de vaccination, et $x(t)$ le nombre d'individus infectés. On a :

$$\dot{x}(t) = \alpha x(t)(N - x(t)) - u(t)x(t), \quad x(0) = x_0,$$

où $0 \leqslant x_0 \leqslant N$, et où le contrôle $u(t)$ vérifie la contrainte

$$0 \leqslant u(t) \leqslant C,$$

où $C > 0$ est une constante.

Soit $T > 0$ fixé. On cherche à minimiser le critère :

$$C_T(u) = \int_0^T (x(t) + \beta u(t))dt + \gamma x(T),$$

où $\beta > 0$ et $\gamma > 0$ sont des constantes de pondération (compromis entre économie de vaccins dépensés et minimisation du nombre d'infectés).

Décrire la structure du contrôle optimal, montrer qu'il est bang-bang.

Indications : En notant $p(t)$ le vecteur adjoint, on montre que la fonction $t \mapsto p(t)x(t) + \beta$ est strictement croissante.

Exercice 7.3.12 (Contrôle optimal d'une épidémie)**.** Considérons une population touchée par une épidémie que l'on cherche à enrayer par une vaccination. On note

- $I(t)$, le nombre d'individus infectieux, qui peuvent contaminer les autres ;
- $S(t)$, le nombre d'individus non infectieux, mais contaminables ;
- $R(t)$, le nombre d'individus infectés, et disparus, ou isolés du reste de la population.

Soit $r > 0$ le taux d'infection, $\gamma > 0$ le taux de disparition, et $u(t)$ le taux de vaccination. Le contrôle $u(t)$ vérifie la contrainte $0 \leqslant u(t) \leqslant a$. La modélisation est

$$\begin{aligned}\dot{S}(t) &= -rS(t)I(t) + u(t),\\ \dot{I}(t) &= rS(t)I(t) - \gamma I(t) - u(t),\\ \dot{R}(t) &= \gamma I(t),\end{aligned}$$

et le but est de déterminer une loi optimale de vaccination, de maniére à minimiser, en un temps T fixé, le coût

$$C(u) = \alpha I(T) + \int_0^T u(t)^2 dt,$$

où $\alpha > 0$ est donné.

Déterminer l'expression du contrôle optimal en fonction du vecteur adjoint. Que vaut le contrôle optimal au voisinage du temps final si $2a < \alpha$?

Indications : Le Hamiltonien est $H = p_S(-rSI+u)+p_I(rSI-\gamma I-u)+p_R\gamma I+p^0u^2$, et

$$\dot{p}_S = p_S rI - p_I rI,\ \dot{p}_I = p_S rS - p_I(rS-\gamma) - p_R\gamma,\ \dot{p}_R = 0.$$

Les conditions de transversalité sont $p_S(T) = 0$, $p_I(T) = p^0\alpha$, et $p_R(T) = 0$. On en déduit que $p^0 \neq 0$, et on choisit $p^0 = -1/2$. En remarquant que H est une fonction concave de u atteignant son maximum absolu en $u = p_S - p_I$, on en déduit que

$$u(t) = \begin{cases} 0 & \text{si} \quad p_S(t) - p_I(t) < 0,\\ p_S(t) - p_I(t) & \text{si} \quad 0 \leqslant p_S(t) - p_I(t) \leqslant a,\\ a & \text{si} \quad p_S(t) - p_I(t) > a.\end{cases}$$

Au temps final, $p_S(T) - p_I(T) = \alpha/2$, donc si $2a < \alpha$ alors $u(t) = a$ dans un voisinage du temps final.

Exercice 7.3.13 (Contrôle optimal d'un procédé de fermentation)**.** Considérons le procédé de fermentation

$$\begin{aligned}\dot{x}(t) &= -x(t) + u(t)(1-x(t)),\ x(0) = x_0,\\ \dot{y}(t) &= x(t) - u(t)y(t),\ y(0) = 0,\end{aligned}$$

où $x(t)$ représente la concentration de sucre, $y(t)$ la concentration d'éthanol, et $u(t)$, le contrôle, est le taux d'évaporation. On suppose $0 \leqslant u(t) \leqslant M$, et $0 < x_0 < 1$. Soit y_1 tel que $y_1 > 1/M$ et $y_1 > x_0$; on veut résoudre le probléme du temps minimal pour rejoindre $y(t_f) = y_1$.

1. Montrer que $x_0 \mathrm{e}^{-t} \leqslant x(t) < 1$, pour tout $t \in [0, t_f]$.
2. On note les variables adjointes (p_x, p_y) et p^0.
   (a) Ecrire le Hamiltonien du probléme de contrôle optimal et les équations des extrémales.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint et sur le temps. Montrer que $p_y(t_f) \neq 0$.
3. (a) Pour tout $t \in [0, t_f]$, soit $\varphi(t) = p_x(t)(1 - x(t)) - p_y(t)y(t)$. Calculer $\varphi'(t)$ et $\varphi''(t)$. Montrer que φ est strictement monotone.
   (b) En déduire que les contrôles optimaux sont bang-bang avec au plus une commutation, et préciser leur expression.
   (c) Montrer que $\dot{y}(t_f) \geqslant 0$.
   (d) En déduire qu'il existe $\varepsilon > 0$ tel que $u(t) = 0$, pour presque tout $t \in [t_f - \varepsilon, t_f]$.
   (e) Conclure sur la structure du contrôle optimal.

Corrigé :

1. $u \geqslant 0$, donc $\dot{x}(t) \geqslant -x(t)$ et donc $x(t) \geqslant x_0 \mathrm{e}^{-t} > 0$; puis $u \leqslant M$, donc $\dot{x} < M(1 - x)$ (tant que $x < 1$), avec $0 < x_0 < 1$, d'où $x(t) < 1$ (par raisonnement a priori et par comparaison avec la solution 1).
2. (a) Le Hamiltonien est $H = p_x(-x+u(1-x))+p_y(x-uy)+p^0$, et $\dot{p}_x = p_x(1 + u) - p_y$, $\dot{p}_y = up_y$.
   (b) On a $p_x(t_f) = 0$ et $H(t_f) = 0$. En particulier, $p_y(t_f)(x(t_f)-u(t_f)y(t_f)) + p^0 = 0$, donc forcément $p_y(t_f) \neq 0$ (sinon on aurait aussi $p^0 = 0$: absurde).
3. (a) On calcule $\varphi' = p_x - p_y$, puis $\varphi'' = (p_x - p_y)(1 + u) = (1 + u)\varphi'$. Si φ' s'annule en un temps t, alors $\varphi' \equiv 0$ sur $[0, t_f]$ tout entier d'aprés l'équation différentielle ci-dessus et par unicité de Cauchy. En particulier, $p_x(t_f) - p_y(t_f) = 0$, et donc $p_y(t_f) = 0$, ce qui est une contradiction. Donc φ' ne s'annule pas, et φ est strictement monotone.
   (b) Par le principe du maximum, on a $u(t) = 0$ si $\varphi(t) < 0$, et $u(t) = M$ si $\varphi(t) > 0$, et par le raisonnement ci-dessus φ ne s'annule qu'au plus une fois donc u est bien bang-bang avec au plus une commutation.
   (c) On veut passer de $y(0) = 0$ à $y(t_f) = y_1 > 0$ en temps minimal, donc forcément au temps minimal t_f on a $\dot{y}(t_f) \geqslant 0$.

En effet sinon, y serait strictement décroissante sur un intervalle $[t_f - \eta, t_f]$, et puisque $y(0) = 0$, d'aprés le théoréme des valeurs intermédiaires, il existerait $t_1 < t_f$ tel que $y(t_1) = y_1$, ce qui contredit le fait que t_f est le temps minimal.

(d) Montrons que $p_y(t_f) > 0$. Par l'absurde, si $p_y(t_f) < 0$, alors $\varphi(t_f) > 0$, et donc par continuité, $\varphi(t) > 0$ à la fin, donc $u = M$. Donc $\dot{y}(t_f) = x(t_f) - u(t_f)y(t_f) = x(t_f) - My_1 < 1 - My_1 < 0$ (par hypothése), ce qui contredit la question précédente.
Donc $p_y(t_f) > 0$, et $\varphi(t) < 0$ à la fin, i.e., $u = 0$ à la fin.

(e) Si u ne commute pas alors $u = 0$ sur tout $[0, t_f]$, donc on résout $\dot{x} = -x$ et $\dot{y} = x$, ce qui conduit à $y(t) = x_0(1-\mathrm{e}^{-t})$. En particulier, $y(t) < x_0$, et donc $y_1 > x_0$ est inatteignable.
Donc u commute une fois, et passe de M à 0.

Exercice 7.3.14 (Contrôle optimal d'un avion)**.** Considérons le mouvement d'un avion, modélisé par

$$\dot{x}(t) = v(t),\ \dot{v}(t) = \frac{u(t)}{mv(t)} - \mu g - \frac{c}{m}v(t)^2,$$

où $x(t)$ est la distance au sol parcourue, $v(t)$ est le module de la vitesse, le contrôle $u(t)$ est l'apport d'énergie, m est la masse, et μ, c sont des coefficients aérodynamiques. Le contrôle vérifie la contrainte

$$0 < a \leqslant u(t) \leqslant b,$$

et le but est de déterminer une trajectoire menant du point initial $x(0) = x_0, v(0) = v_0$, au point final $x(t_f) = x_f, v(t_f) = v_f$, et minimisant le coût $C(u) = \int_0^{t_f} u(t)dt$, le temps final t_f n'étant pas fixé.

Montrer qu'il n'existe aucune trajectoire singulière, puis expliquer comment mettre en oeuvre une méthode numérique pour résoudre ce problème.

Exercice 7.3.15 (Problème de Goddard simplifié)**.** Le décollage d'une fusée est modélisé par les équations

$$\begin{aligned}
\dot{h}(t) &= v(t),\ h(0) = 0,\\
\dot{v}(t) &= \frac{u(t)}{m(t)} - g,\ v(0) = 0,\\
\dot{m}(t) &= -bu(t),\ m(0) = m_0,
\end{aligned}$$

où $h(t)$ est l'altitude, $v(t)$ le module de la vitesse, $m(t)$ la masse, g l'accélération de la pesanteur, et $b > 0$ un réel. Le contrôle est la poussée $u(t)$, qui vérifie la contrainte $0 \leqslant u(t) \leqslant u_{max}$. Par ailleurs la masse de la fusée en l'absence de carburant est m_1, si bien que la masse $m(t)$ vérifie la contrainte $m_1 \leqslant m(t) \leqslant m_0$. Enfin, on suppose que $u_{max} > gm_0$.

Montrer que la politique optimale permettant de maximiser l'altitude finale est bang-bang, avec au plus une commutation, du type $u = u_{max}$ puis s'il y a commutation $u = 0$.

Indications : Montrer que les conditions de transversalité sont $p_h(t_f) = -p^0$, $p_v(t_f) = 0$ et $H(t_f) = 0$. Montrer que la fonction de commutation $\varphi(t) = \frac{p_v(t)}{m(t)} - bp_m(t)$ vérifie $\dot{\varphi} = -\frac{p_h}{m}$. Noter que, au début, on doit avoir $\dot{v} > 0$, *i.e.* $u > mg$, ce qui est possible puisque $u_{max} > gm_0$. En déduire que, au début, on a $u > 0$, et donc soit $\varphi > 0$, soit $\varphi \equiv 0$. Montrer alors par l'absurde que $p^0 \neq 0$, puis montrer que l'alternative $\varphi \equiv 0$ est impossible. En déduire que la politique optimale est bang-bang, avec au plus une commutation, du type $u = u_{max}$ puis s'il y a commutation $u = 0$.

Exercice 7.3.16 (Guidage d'un engin spatial)**.** Considérons le mouvement d'un engin spatial, modélisé par le système de contrôle (normalisé)

$$\begin{aligned}
\dot{r}(t) &= v(t),\\
\dot{v}(t) &= \frac{\theta(t)^2}{r(t)} - \frac{1}{r(t)^2} + u_1(t)\frac{c}{m(t)}\sin u_2(t),\\
\dot{\theta}(t) &= -\frac{v(t)\theta(t)}{r(t)} + u_1(t)\frac{c}{m(t)}\cos u_2(t),\\
\dot{m}(t) &= -u_1(t),
\end{aligned}$$

où $r(t)$ représente la distance de l'engin au centre de la Terre, $v(t)$ la vitesse radiale, $\theta(t)$ la vitesse angulaire, $m(t)$ la masse de l'engin. Les contrôles sont $u_1(t)$, la poussée, et $u_2(t)$, l'angle de gîte. Le contrôle u_1 vérifie la contrainte $0 \leqslant u_1 \leqslant \beta$. On considère les conditions aux limites

$$\begin{aligned}
&r(0) = 1,\ r(t_f) = r_f,\\
&v(0) = 0,\ v(t_f) = 0,\\
&\theta(0) = 1,\ \theta(t_f) = \frac{1}{\sqrt{r_f}},\\
&m(0) = 1.
\end{aligned}$$

Déterminer une trajectoire vérifiant ces conditions aux limites, et maximisant la masse finale $m(t_f)$, le temps final n'étant pas fixé.

Indications : raisonnement similaire à l'exercice 7.3.17.

Exercice 7.3.17 (Sujet d'examen)**.** Le problème est de maximiser le déport latéral d'une fusée dont le mouvement est plan et la poussée est limitée. Au temps t, on note $x(t) = (x_1(t), x_2(t))$ la position de la fusée, $v(t) = (v_1(t), v_2(t))$ sa vitesse, $m(t)$ sa masse, $\theta(t)$ l'angle de la direction de poussée, et $u(t)$ la variation de masse (proportionnelle à la force de poussée). Pour simplifier, on néglige

les forces aérodynamiques et on suppose que l'accélération de la pesanteur g est constante. Le système modélisant le mouvement de la fusée est alors le suivant :

$$\begin{aligned}
\dot{x}_1 &= v_1 \\
\dot{x}_2 &= v_2 \\
\dot{v}_1 &= \frac{c}{m} u \cos\theta \\
\dot{v}_2 &= \frac{c}{m} u \sin\theta - g \\
\dot{m} &= -u
\end{aligned}$$

où $c > 0$ est constante. Les contrôles sont $\theta(t)$ et $u(t)$. On suppose que

$$\theta \in \mathbb{R} \quad \text{et} \quad 0 \leqslant u \leqslant A.$$

Les données initiales sont :

$$x_1(0) = x_1^0,\ x_2(0) = x_2^0,\ v_1(0) = v_1^0,\ v_2(0) = v_2^0,\ m(0) = m_0.$$

La masse de la fusée lorsqu'il n'y a pas de carburant est m_1. Autrement dit $m(t)$ doit vérifier :

$$m_1 \leqslant m(t) \leqslant m_0.$$

On désire mener la fusée du point initial précédent à la variété terminale

$$x_2(t_f) = x_2^1,\ m(t_f) = m_1,$$

le temps final t_f étant libre, et on veut maximiser la quantité

$$x_1(t_f).$$

1. *Application du principe du maximum.*
   On introduit les variables adjointes $p_{x_1}, p_{x_2}, p_{v_1}, p_{v_2}, p_m$, et p^0. On pose de plus $\lambda = p_{x_2}(t_f)$.
   (a) Ecrire le Hamiltonien associé à ce problème de contrôle optimal, ainsi que le système différentiel extrémal.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Montrer que le Hamiltonien est nul le long de toute extrémale.
   (d) Calculer $p_{x_1}(t), p_{x_2}(t), p_{v_1}(t)$ et $p_{v_2}(t)$ en fonction de λ et p^0.
2. *Calcul des contrôles extrémaux.*
   (a) Montrer que l'on ne peut pas avoir simultanément $p^0 = 0$ et $\lambda = 0$. En déduire l'expression des contrôles extrémaux $\theta(t)$, montrer qu'ils sont constants, et préciser leur valeur θ_0 en fonction de λ et p^0.
   (b) On introduit la fonction φ sur $[0, t_f]$

$$\varphi(t) = \frac{c}{m(t)} \sqrt{(p^0)^2 + \lambda^2}(t_f - t) - p_m(t).$$

Montrer par l'absurde que la fonction φ ne s'annule sur aucun sous-intervalle de $[0, t_f]$. Préciser la monotonie de φ. En déduire que les contrôles extrémaux $u(t)$ commutent au plus une fois sur $[0, t_f]$, et passent dans ce cas de la valeur A à la valeur 0.

(c) On suppose que
$$At_f > m_0 - m_1.$$
Montrer que u commute exactement une fois au temps
$$t_c = \frac{m_0 - m_1}{A},$$
passe de la valeur A à la valeur 0 en ce temps t_c, et de plus $m(t_c) = m_1$.

3. *Calcul des contrôles en boucle fermée.*

(a) Montrer que
$$\theta_0 = -\arctan \frac{v_1(t_f)}{v_2(t_f)} \ .$$

(b) Montrer que
$$v_2(t_f)^2 = v_2(t_c)^2 - 2g(x_2^1 - x_2(t_c)) \ ,$$
et en déduire que
$$\tan^2 \theta_0 = \frac{v_1(t_c)^2}{v_2(t_c)^2 - 2g(x_2^1 - x_2(t_c))}.$$

(c) Montrer les trois formules suivantes :
$$\begin{aligned}
v_1(t_c) &= v_1^0 + c\cos\theta_0 \ln\frac{m_0}{m_1}, \\
v_2(t_c) &= v_2^0 + c\sin\theta_0 \ln\frac{m_0}{m_1} - gt_c, \\
x_2(t_c) &= x_2^0 + v_2^0 t_c - \frac{g}{2}t_c^2 - c\sin\theta_0 \left(t_c \ln\frac{m_1}{m_0} - t_c - \frac{m_0}{A}\ln\frac{m_1}{m_0} \right) \ .
\end{aligned}$$
En déduire que l'on peut exprimer θ_0 en fonction des données
$$x_2^0, v_1^0, v_2^0, m_0, m_1, A, c, g$$
(on ne cherchera pas une expression explicite). Montrer que l'on a ainsi exprimé les contrôles θ et u en boucle fermée. Quel est l'avantage de ce procédé ?

Exercice 7.3.18 (Sujet d'examen : politique d'investissement financier d'une banque)**.** Considérons une banque, qui gére une certaine quantité d'argent, et doit répondre aux besoins éventuels de ses clients en leur accordant un emprunt d'argent. Pour cela, la banque doit disposer d'argent immédiatement disponible,

qui lui rapporte moins d'intérêts que l'argent investi dans des titres financiers. La banque investit donc une partie du capital dans l'achat de titres. D'autre part, si la réserve d'argent est trop faible, la banque doit vendre des titres et pour cela doit payer une commission à un agent de change.

Le probléme est de déterminer une politique financiére qui réalise un compromis entre quantité d'argent disponible et argent investi, tout en maximisant le gain.

Notations :

$x(t)$: quantité d'argent disponible au temps t.
$y(t)$: quantité de titres financiers investis au temps t.
$d(t)$: taux instantané de demande d'emprunts par des clients
$u(t)$: taux de vente de titres ($u(t) > 0$ signifie que la banque vend des titres, et $u(t) < 0$ signifie que la banque achéte des titres).
$r_1(t)$: taux d'intérêt gagné sur l'argent disponible.
$r_2(t)$: taux d'intérêt gagné sur l'argent investi en titres (on suppose que $r_2(t) > r_1(t)$, pour tout temps t).
α : taux de commission prélevé par l'agent de change lors de la vente et de l'achat de titres (on suppose que $0 < \alpha < 1$).

Les équations modélisant le systéme sont

$$\begin{aligned} \dot{x}(t) &= r_1(t)x(t) - d(t) + u(t) - \alpha|u(t)|, \\ \dot{y}(t) &= r_2(t)y(t) - u(t), \end{aligned}$$

avec $x(0) = x_0$ et $y(0) = y_0$. Le contrôle $u(t)$ vérifie la contrainte

$$-U_2 \leqslant u(t) \leqslant U_1,$$

où $U_1, U_2 \geqslant 0$. On fixe un temps final T, et on veut maximiser la quantité

$$x(T) + y(T).$$

(dans l'étude qui suit, on ne tient pas compte du fait qu'il faut de plus imposer $x(t), y(t) \geqslant 0$, cette contrainte devant être vérifiée a posteriori).

1. Le principe du maximum classique ne peut pas s'appliquer à cause du terme $|u(t)|$. On propose donc de poser

$$u_1 = \max(u, 0) = \frac{u + |u|}{2}, \ u_2 = -\min(u, 0) = \frac{-u + |u|}{2}.$$

   (a) Avec ces notations, montrer que

$$-U_2 \leqslant u \leqslant U_1 \Leftrightarrow \begin{cases} u_1 \geqslant 0, \ u_2 \geqslant 0, \\ u_1 u_2 = 0, \\ -U_2 \leqslant u_1 - u_2 \leqslant U_1. \end{cases}$$

(b) Ecrire le nouveau probléme $\mathcal{P}$ de contrôle optimal, contrôlé par u_1 et u_2.

2. *Application du principe du maximum.*
On introduit les variables adjointes p_x, p_y, et p^0.
   (a) Ecrire le Hamiltonien associé au probléme de contrôle optimal $\mathcal{P}$, ainsi que le systéme différentiel extrémal.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Montrer que p^0 est forcément non nul. Dans la suite, on pose $p^0 = 1$.
   (d) Montrer que $p_x(t) > 0$ et $p_y(t) > 0$, pour tout $t \in [0, T]$.
3. *Calcul des contrôles extrémaux.*
Soient $u_1(t)$ et $u_2(t)$ les contrôles extrémaux au temps t.
   (a) Montrer que :
   – si $(1-\alpha)p_x(t) - p_y(t) > 0$, alors $u_1(t) = U_1$ et $u_2(t) = 0$;
   – si $(1+\alpha)p_x(t) - p_y(t) < 0$, alors $u_1(t) = 0$ et $u_2(t) = U_2$;
   – si $(1-\alpha)p_x(t) - p_y(t) < 0$ et $(1+\alpha)p_x(t) - p_y(t) > 0$ alors $u_1(t) = u_2(t) = 0$.
   (b) Montrer que les fonctions $t \mapsto (1-\alpha)p_x(t) - p_y(t)$ et $t \mapsto (1+\alpha)p_x(t) - p_y(t)$ ne s'annulent sur aucun sous-intervalle de $[0, T]$.
   (c) En déduire que les contrôles extrémaux sont bang-bang sur $[0, T]$. Décrire leur structure dans un graphe ayant p_x en abscisse et p_y en ordonnée.
   (d) Que valent $u_1(t)$ et $u_2(t)$ sur $[T - \eta, T]$, pour $\eta > 0$ assez petit ?
4. *Exemples explicites.* On pose $T = 1$ et $\alpha = 0.01$. Décrire la politique optimale de la banque (on donnera une approximation numérique à 0.01 prés des temps de commutation) dans chacun des cas suivants :
   (a) $r_1(t) = 1/3$ et $r_2(t) = 1/2$, pour tout $t \in [0, T]$.
   (b) $r_1(t) = 1/2$ et $r_2(t) = t/2$, pour tout $t \in [0, T]$.

Exercice 7.3.19 (Sujet d'examen : Contrôle optimal de la pollution par engrais)**.** On considére l'évolution de la quantité de pollution $x(t)$ dans un champ de céréales où l'on cherche, par ajout d'engrais, à optimiser le rendement tout en minimisant la pollution produite.

Le contrôle $u(t)$ est la quantité d'engrais ajouté. Il vérifie la contrainte

$$0 \leqslant u(t) \leqslant 3.$$

On note $\alpha > 0$ le taux de décroissance naturelle de la pollution. L'évolution de la pollution $x(t)$ est

$$\dot{x}(t) = u(t) - \alpha x(t),$$

avec $x(0) = x_0 > 0$.

D'une part, on cherche à minimiser la pollution engendrée par l'engrais, mais d'autre part, on cherche à optimiser le rendement de céréales par ajout d'engrais.

Cependant, un ajout excessif d'engrais a aussi un effet nocif sur les plantes, et donc sur le rendement. On fixe un temps final T, et on cherche à minimiser le critére

$$C_T(u) = \int_0^T \left(x(t)^2 - \sqrt{(3-u(t))(1+u(t))} \right) dt.$$

1. On introduit les variables adjointes p et p^0.
   (a) Ecrire le Hamiltonien de ce probléme de contrôle optimal, ainsi que les équations des extrémales.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Montrer que p^0 est forcément non nul. Dans la suite, on pose $p^0 = -1$.
2. (a) Montrer que $x(t) \geqslant x_0 \mathrm{e}^{-\alpha t}$, et en particulier $x(t) > 0$, pour tout $t \in [0, T]$.
   (b) En déduire que $p(t) < 0$, pour tout $t \in [0, T[$.
3. Soit $u(t)$ le contrôle extrémal au temps t.
   (a) Montrer que

$$u(t) = \begin{cases} 0 & \text{si} \quad p(t) \leqslant -1/\sqrt{3}, \\ 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}} & \text{si} \quad p(t) > -1/\sqrt{3}. \end{cases}$$

   (b) Montrer que $u(t) = 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}}$ sur $[T-\eta, T]$, pour $\eta > 0$ assez petit.
4. (a) Montrer que $p(t) \geqslant \left(p(0) + \frac{x_0}{\alpha} \right) \mathrm{e}^{\alpha t} - \frac{x_0}{\alpha} \mathrm{e}^{-\alpha t}$, pour tout $t \in [0, T]$.
   (b) En déduire que $p(0) \leqslant \frac{x_0}{\alpha} \left(\mathrm{e}^{-2\alpha T} - 1 \right)$.
5. On suppose désormais que $x_0(1 - \mathrm{e}^{-2\alpha T}) > \alpha/\sqrt{3}$.
   (a) Montrer que $u(t) = 0$ sur un intervalle du type $[0, t_1]$.
   (b) Que vaut $p(t)$ sur $[0, t_1]$?
   (c) Montrer que $p(0) + \frac{x_0}{\alpha} \geqslant 0$.
6. Montrer finalement que

$$u(t) = \begin{cases} 0 & \text{si} \quad 0 \leqslant t \leqslant t_1, \\ 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}} & \text{si} \quad t_1 < t \leqslant T. \end{cases}$$

   Caractériser le temps de commutation t_1 (sans chercher à le calculer explicitement).
7. Montrer que $\max \left(-\frac{x_0}{\alpha}, \frac{x_0}{\alpha} \left(\mathrm{e}^{-2\alpha T} - 1 \right) - \frac{1}{\sqrt{3}} \mathrm{e}^{-\alpha T} \right) \leqslant p(0) \leqslant \frac{x_0}{\alpha} \left(\mathrm{e}^{-2\alpha T} - 1 \right)$.
8. Décrire et discuter, critiquer, les méthodes numériques que l'on peut mettre en oeuvre pour résoudre numériquement ce probléme de contrôle optimal.

Corrigé :

1. (a) $H = p(u - \alpha x) + p^0(x^2 - \sqrt{(3-u(t))(1+u(t))})$, et $\dot{p} = \alpha p - 2p^0 x$.
   (b) $p(T) = 0$.
   (c) Donc $p^0 \neq 0$. Dans la suite, on pose $p^0 = -1$.
2. (a) $u(t) \geqslant 0$, donc $\dot{x} \geqslant -\alpha x$, d'où $x(t) > x_0 \mathrm{e}^{-\alpha t}$, et en particulier $x(t) > 0$, pour tout $t \in [0, T]$.
   (b) On a $\dot{p} = \alpha p + 2x$, avec $x > 0$. Donc, si en un temps $t_1 < T$, on a $p(t_1) \geqslant 0$, alors $p(t) > p(t_1) > 0$ pour $t > t_1$, ce qui contredit $p(T) = 0$. Et donc, $p(t) < 0$, pour tout $t \in [0, T]$.
3. Soit $u(t)$ le contrôle extrémal au temps t.
   (a) La condition de maximisation est $\max_{0 \leqslant u \leqslant 3} f(u)$ avec $f(u) = pu - \sqrt{(3-u)(1+u)}$. Etudions en fonction de p cette fonction $f(u)$ (sachant que $p < 0$), pour $0 \leqslant u \leqslant 3$. On trouve que f atteint son maximum sur l'intervalle $[-1, 3]$ lorsque $u = 1 + \frac{2p}{\sqrt{p^2+1}}$ (qui est bien toujours < 3). Par ailleurs, $1 + \frac{2p}{\sqrt{p^2+1}} = 0$ (avec $p < 0$) si et seulement si $p = -1/\sqrt{3}$. Donc, finalement,
   $$u(t) = \begin{cases} 0 & \text{si} \quad p(t) \leqslant -1/\sqrt{3}, \\ 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}} & \text{si} \quad p(t) > -1/\sqrt{3}. \end{cases}$$
   (b) A la fin, $p(T) = 0 > -1/\sqrt{3}$, donc $u(t) = 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}}$ sur $[T - \eta, T]$, pour $\eta > 0$ assez petit.
4. (a) D'aprés 2.a, on a $\dot{p}(t) \geqslant \alpha p(t) + 2x_0 \mathrm{e}^{-\alpha t}$, donc $\mathrm{e}^{\alpha t} \frac{d}{dt}(\mathrm{e}^{-\alpha t} p(t)) \geqslant 2x_0 \mathrm{e}^{-\alpha t}$, et en intégrant, $p(t) \geqslant \big(p(0) + \frac{x_0}{\alpha}\big) \mathrm{e}^{\alpha t} - \frac{x_0}{\alpha} \mathrm{e}^{-\alpha t}$, pour tout $t \in [0, T]$.
   (b) $p(T) = 0$, donc par l'inégalité précédente, $p(0) \leqslant \frac{x_0}{\alpha} \big(\mathrm{e}^{-2\alpha T} - 1\big)$.
5. (a) Sous l'hypothése $x_0(1 - \mathrm{e}^{-2\alpha T}) > \alpha/\sqrt{3}$, on obtient $p(0) \leqslant \frac{x_0}{\alpha} \big(\mathrm{e}^{-2\alpha T} - 1\big) < -1/\sqrt{3}$, et donc, $u(t) = 0$ sur un intervalle du type $[0, t_1]$.
   (b) Sur $[0, t_1]$, $u = 0$, donc $\dot{x} = -\alpha x$ et $\dot{p} = \alpha p + 2x$, avec $x(0) = x_0$, d'où en intégrant les équations, $p(t) = \big(p(0) + \frac{x_0}{\alpha}\big) \mathrm{e}^{\alpha t} - \frac{x_0}{\alpha} \mathrm{e}^{-\alpha t}$, pour tout $t \in [0, t_1]$.
   (c) Par l'absurde, si $p(0) + \frac{x_0}{\alpha} < 0$, alors d'aprés l'expression précédente de $p(t)$, on a $p(t) < 0$ pour tout $t \in [0, t_1]$, puis pour tout $t \in [0, T]$, ce qui contredit $p(T) = 0$. Donc, $p(0) + \frac{x_0}{\alpha} \geqslant 0$.

6. Tant que $u = 0$, sur $[0, t_1]$, $p(t)$ est donné par 5.b. En particulier, $\dot{p}(t) > 0$, et donc $p(t)$ est strictement croissante. D'aprés 5.a, $p(0) < -1/\sqrt{3}$, et d'autre part, $p(T) = 0$. Donc il existe bien un temps de commutation $t_1 < T$ pour lequel $p(t_1) = \left(p(0) + \frac{x_0}{\alpha}\right) \mathrm{e}^{\alpha t_1} - \frac{x_0}{\alpha}\mathrm{e}^{-\alpha t_1} = -1/\sqrt{3}$ (ce qui caractérise t_1). Ensuite, pour $t \geqslant t_1$, on sait d'aprés 4.a que $p(t) \geqslant g(t)$, où la fonction $g(t) = \left(p(0) + \frac{x_0}{\alpha}\right) \mathrm{e}^{\alpha t} - \frac{x_0}{\alpha}\mathrm{e}^{-\alpha t}$ est croissante (on a en effet $\dot{g}(t) > 0$ car $p(0) + \frac{x_0}{\alpha} \geqslant 0$), donc $p(t) > -1/\sqrt{3}$ pour $t_1 < t \leqslant T$. Finalement,

$$u(t) = \begin{cases} 0 & \text{si} \quad 0 \leqslant t \leqslant t_1, \\ 1 + \frac{2p(t)}{\sqrt{p(t)^2+1}} & \text{si} \quad t_1 < t \leqslant T. \end{cases}$$

7. Comme $t_1 < T$, et par croissance de la fonction $g(t)$, on a $-1/\sqrt{3} = g(t_1) < g(T)$, ce qui conduit à $p(0) > \frac{x_0}{\alpha}\left(\mathrm{e}^{-2\alpha T} - 1\right) - \frac{1}{\sqrt{3}}\mathrm{e}^{-\alpha T}$ et donc, d'aprés 4.b et 5.c,

$$\max\left(-\frac{x_0}{\alpha}, \frac{x_0}{\alpha}\left(\mathrm{e}^{-2\alpha T} - 1\right) - \frac{1}{\sqrt{3}}\mathrm{e}^{-\alpha T}\right) \leqslant p(0) \leqslant \frac{x_0}{\alpha}\left(\mathrm{e}^{-2\alpha T} - 1\right).$$

8. On peut implémenter la méthode de tir, où $p(0)$ est cherché dans l'intervalle ci-dessus, ou bien mettre en oeuvre une méthode directe. On peut en discuter les avantages et inconvénients.

Exercice 7.3.20 (Sujet d'examen : Commande optimale d'un réacteur chimique)**.** Un réacteur chimique industriel permet de fabriquer un produit à partir d'un réactif par une réaction irréversible du premier ordre avec dégagement de chaleur. Pour refroidir le réacteur, on fait circuler le contenu à travers un échangeur thermique ; la chaleur passe ainsi dans le liquide de refroidissement qui circule dans le circuit secondaire de l'échangeur avec un débit $u(t)$. Aprés diverses réductions de modéle, le systéme s'écrit sous la forme

$$\begin{aligned} \dot{x}_1(t) &= -a_1 x_1(t) - k x_1(t) e^{-\frac{a_2}{x_2(t)}} + r_1 \\ \dot{x}_2(t) &= a_3(a_4 - x_2(t)) + a_5 k x_1(t) e^{-\frac{a_2}{x_2(t)}} + a_6(u(t) - x_2(t)) - r_2 \end{aligned}$$

où $x_1(t)$ est la concentration du réactif au temps t, $x_2(t)$ est la température du réacteur au temps t, et r_1 et r_2 sont des réels strictement positifs. Par ailleurs, les coefficients k et a_i, $i = 1 \ldots 6$, sont des réels positifs. On suppose que le contrôle $u(t)$ vérifie la contrainte

$$|u(t)| \leqslant M,$$

où M est un réel positif. Soit $T > 0$ un temps final **fixé**. Dans ce qui suit, l'état initial est fixé :

$$x_1(0) = x_1^0, \quad x_2(0) = x_2^0,$$

et l'état final $(x_1(T), x_2(T))$ est libre.

1. Dans cette premiére question, on cherche à minimiser la quantité de réactif $x_1(T)$.
   On note les variables adjointes p et p^0.
   (a) Ecrire le Hamiltonien du probléme de contrôle optimal et les équations des extrémales.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Montrer que $p^0 \neq 0$. Que posez-vous pour la suite ?
   (d) Démontrer que les contrôles optimaux sont bang-bang, et préciser leur expression.
   (*indication :* démontrer, par l'absurde, que $p_2(t)$ ne peut s'annuler identiquement sur un sous-intervalle)
   (e) Montrer qu'il existe $\varepsilon > 0$ tel que $u(t) = M$, pour presque tout $t \in [T-\varepsilon, T]$ (autrement dit, le contrôle u vaut M à la fin).
   (f) On suppose que $a_5 = 0$. Démontrer que le contrôle optimal est constant sur $[0, T]$, égal à M.

2. Dans cette deuxiéme question, on cherche toujours à minimiser la quantité de réactif $x_1(T)$, mais en minimisant aussi la température $x_2(t)$ au cours de la réaction, et l'énergie fournie. Le compromis choisi et de chercher à minimiser le coût

$$C_T(u) = \int_0^T (u(t)^2 + \beta x_2(t)^2)dt + x_1(T),$$

   où $\beta \geqslant 0$ est fixé.
   (a) Ecrire le Hamiltonien du probléme de contrôle optimal et les équations des extrémales.
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Montrer que $p^0 \neq 0$. Que posez-vous pour la suite ?
   (d) Détailler la condition de maximisation du principe du maximum de Pontryagin, et donner l'expression des contrôles optimaux.
   (e) Montrer qu'il existe $\varepsilon > 0$ tel que $u(t) = \frac{1}{2}a_6 p_2(t)$, pour presque tout $t \in [T-\varepsilon, T]$.
   (f) On suppose que $a_5 = \beta = 0$. Démontrer que le contrôle optimal est strictement positif sur $[0, T]$, et préciser son expression.

   Corrigé :

   1. (a) Le Hamiltonien est $H = p_1(-a_1x_1 - kx_1e^{-\frac{a_2}{x_2}} + r_1) + p_2(a_3(a_4 - x_2) + a_5kx_1e^{-\frac{a_2}{x_2}} + a_6(u - x_2) - r_2)$. Les équations des extrémales sont

$$\dot{p}_1 = p_1(a_1 + ke^{-\frac{a_2}{x_2}}) - p_2a_5ke^{-\frac{a_2}{x_2}}$$
$$\dot{p}_2 = p_1kx_1\frac{a_2}{x_2^2}e^{-\frac{a_2}{x_2}} + p_2(a_3 + a_6 - a_5kx_1\frac{a_2}{x_2^2}e^{-\frac{a_2}{x_2}})$$

(b) Les conditions de transversalité sur le vecteur adjoint sont alors $p_1(T) = p^0$ et $p_2(T) = 0$, avec $p^0 \leqslant 0$.

(c) D'où il découle forcément que $p^0 \neq 0$. On pose alors $p^0 = -1$.

(d) Il résulte de la condition de maximisation que $u(t) = M * signe(p_2(t))$ (bang-bang), pourvu que p_2 ne s'annule pas identiquement sur un sous-intervalle.

Par l'absurde, si p_2 s'annule identiquement sur un sous-intervalle, alors, d'aprés l'équation de p_2, on obtient aussi $p_1 = 0$. Par unicité de Cauchy, on obtient alors $p_1 = p_2 = 0$ sur tout l'intervalle $[0, T]$, ce qui contredit $p_1(T) = -1$.

(e) A la fin $p_2(T) = 0$ et $p_1(T) = -1$, d'où, par l'équation de p_2, $\dot{p}_2(T) < 0$. Donc, à la fin p_2 est strictement décroissante, et comme $p_2(T) = 0$, on obtient $p_2(t) > 0$ sur un sous-intervalle, et donc, $u(t) = M$ à la fin.

(f) Si de plus $a_5 = 0$, alors, d'aprés l'équation de p_1, $p_1(t)$ ne peut s'annuler (par unicité de Cauchy), donc $p_1(t) < 0$ pour tout $t \in [0, T]$. Donc $\dot{p}_2 < (a_3+a_6)p_2$. Par conséquent, $\frac{d}{dt}e^{-(a_3+a_6)t}p_2(t) < 0$, d'où il résulte que $e^{-(a_3+a_6)t}p_2(t) > e^{-(a_3+a_6)T}p_2(T) = 0$, et donc, $p_2(t) > 0$, pour tout $t \in [0, T]$. Donc $u = M$ sur tout l'intervalle.

2. (a) Le Hamiltonien est $H = p_1(-a_1x_1 - kx_1e^{-\frac{a_2}{x_2}} + r_1) + p_2(a_3(a_4-x_2)+a_5kx_1e^{-\frac{a_2}{x_2}}+a_6(u-x_2)-r_2)+p^0(u^2+\beta x_2^2)$. Les équations des extrémales sont

$$\dot{p}_1 = p_1(a_1 + ke^{-\frac{a_2}{x_2}}) - p_2a_5ke^{-\frac{a_2}{x_2}}$$
$$\dot{p}_2 = p_1kx_1\frac{a_2}{x_2^2}e^{-\frac{a_2}{x_2}} + p_2(a_3 + a_6 - a_5kx_1\frac{a_2}{x_2^2}e^{-\frac{a_2}{x_2}}) - 2p^0\beta x_2$$

(b) Les conditions de transversalité sur le vecteur adjoint sont $p_1(T) = p^0$ et $p_2(T) = 0$, avec $p^0 \leqslant 0$.

(c) D'où il découle forcément que $p^0 \neq 0$. On pose alors $p^0 = -1$.

(d) La condition de maximisation est

$$\max_{-M \leqslant u \leqslant M}(a_6p_2(t)u - u^2).$$

La fonction à maximiser est quadratique, son maximum absolu (sans tenir compte des contraintes) et atteint pour $u = \frac{1}{2}a_6p_2(t)$, d'où il résulte que

$$u(t) = \begin{cases} -M & \text{si} \quad \frac{1}{2}a_6p_2(t) < -M, \\ \frac{1}{2}a_6p_2(t) & \text{si} \quad |\frac{1}{2}a_6p_2(t)| < M, \\ M & \text{si} \quad \frac{1}{2}a_6p_2(t) > M. \end{cases}$$

(e) A la fin, $p_2(T) = 0$, donc par continuité, $p_2(t)$ reste petit sur un intervalle du type $[T-\varepsilon, T]$, et donc $|\frac{1}{2}a_6 p_2(t)| < M$ sur cet intervalle, et donc, $u(t) = \frac{1}{2}a_6 p_2(t)$ à la fin.

(f) Si de plus $a_5 = \beta = 0$, alors, comme en 1.f, on montre que $p_2(t) > 0$, pour tout $t \in [0, T]$. Donc $u > 0$ sur tout l'intervalle, et vaut soit M soit $\frac{1}{2}a_6 p_2(t)$ comme ci-dessus.

Exercice 7.3.21 (Sujet d'examen : Troisième phase d'un lanceur.)**.** On considère un modèle simplifié de la troisième phase d'un lanceur, où la Terre est supposée plate et la gravité constante. Dans un repère cartésien, on note $(x_1(t), x_2(t))$ la position de la fusée au temps t ($x_2(t)$ étant l'altitude), et $(v_1(t), v_2(t))$ sa vitesse. Le contrôle u s'écrit

$$u(t) = T(t) \begin{pmatrix} \cos\theta(t) \\ \sin\theta(t) \end{pmatrix},$$

où $T(t)$ est la poussée et $\theta(t)$ est l'incidence. La poussée vérifie la contrainte $0 \leqslant T(t) \leqslant T_{max}$, où $T_{max} > 0$ est fixé. On note $m(t)$ la masse de la fusée au temps t. Le modèle s'écrit, sous forme complète ou sous forme vectorielle :

Forme complète :

$$\begin{aligned} \dot{x}_1 &= v_1 \\ \dot{x}_2 &= v_2 \\ \dot{v}_1 &= \frac{T}{m}\cos\theta \\ \dot{v}_2 &= \frac{T}{m}\sin\theta - g \\ \dot{m} &= -\beta T \end{aligned}$$

Forme vectorielle :

$$\begin{aligned} \dot{x} &= v \\ \dot{v} &= \frac{u}{m} - \vec{g} \\ \dot{m} &= -\beta\|u\| \end{aligned}$$

où β et g sont des constantes strictement positives, $\vec{g} = \begin{pmatrix} 0 \\ g \end{pmatrix}$, et avec les notations $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$. Les conditions initiales et finales sont :

$$\begin{array}{ll} x_1(0) \text{ libre} & x_1(t_f) \text{ libre} \\ x_2(0) = x_{20} & x_2(t_f) = x_{2f} \\ v_1(0) = v_{10} & v_1(t_f) = v_{1f} \\ v_2(0) = v_{20} & v_2(t_f) = 0 \\ m(0) = m_0 & m(t_f) \text{ libre} \end{array}$$

$$t_f \text{ libre}$$

avec $v_{1f} > v_{10}$ et $x_{2f} > x_{20} + \frac{v_{20}^2}{2g}$. On veut maximiser la masse finale ; autrement dit on considère le problème de minimisation

$$\min(-m(t_f)).$$

1. On note les variables adjointes $p = (p_{x_1}, p_{x_2}, p_{v_1}, p_{v_2}, p_m)$ et p^0. En notations vectorielles, $p_x = (p_{x_1}, p_{x_2})$ et $p_v = (p_{v_1}, p_{v_2})$.
   (a) Ecrire le Hamiltonien du problème de contrôle optimal et les équations des extrémales (dans les deux systèmes de notations).
   (b) Ecrire les conditions de transversalité sur le vecteur adjoint.
   (c) Dans la suite, on pose $\lambda = p_{x_2}$. Montrer que $p_{v_2}(t) = -\lambda t + p_{v_2}(0)$.
   (d) Montrer que $H = 0$ le long d'une extrémale.
2. On pose $\Phi(t) = \frac{\|p_v(t)\|}{m(t)} - p_m(t)\beta$. Montrer que
$$u(t) = T(t)\frac{p_v(t)}{\|p_v(t)\|} \quad \text{avec} \quad T(t) = \begin{cases} 0 & \text{si} \quad \Phi(t) < 0, \\ T_{max} & \text{si} \quad \Phi(t) > 0. \end{cases}$$
3. Montrer que p_m est une fonction croissante de t.
4. (a) Montrer que la fonction $t \mapsto p_v(t)$ ne s'annule identiquement sur aucun sous-intervalle.
   (b) Etablir que
$$\ddot{\Phi} = \frac{\beta T}{m}\dot{\Phi} - \frac{m}{\|p_v\|}\dot{\Phi}^2 + \frac{\|p_x\|^2}{m\|p_v\|}.$$
   (c) Montrer que $p_x = 0$ si et seulement si Φ est constante.
   (d) Montrer que la fonction Φ ne s'annule identiquement sur aucun sous-intervalle.
   (e) En déduire que si $p_x = 0$ alors la poussée T est constante sur $[0, t_f]$, égale à T_{max}.
   (f) Montrer que, si $p_x \neq 0$, alors :
   – soit Φ est strictement croissante sur $[0, t_f]$,
   – soit Φ est strictement décroissante sur $[0, t_f]$,
   – soit Φ admet un unique minimum sur $[0, t_f]$, et est strictement décroissante avant ce minimum, et strictement croissante ensuite.
   (g) En déduire les stratégies optimales pour la poussée $T(t)$.
5. Montrer que, si T a au moins une commutation, alors $p^0 \neq 0$.
6. Dans cette question, on se place dans le cas où T admet une seule commutation, et est du type T_{max} puis 0. On note t_1 le temps de commutation.
   (a) Montrer que $\lambda v_2(t) = g p_{v_2}(t) - T(t)\Phi(t)$, pour tout $t \in [0, t_f]$.
   (b) Montrer que $\lambda v_2(t_1) + \lambda g t_1 = g p_{v_2}(0)$.
   (c) Montrer que $t_f = \frac{p_{v_2}(0)}{\lambda}$.
   (d) Montrer que $\lambda > 0$, $p_{v_1} > 0$, $p_{v_2}(0) > 0$.
   (e) Expliquer comment simplifier la mise en oeuvre de la méthode de tir dans ce cas.
7. Dans cette question, on se place dans le cas où T est du type T_{max} puis 0 puis T_{max}. On note $t_1 < t_2$ les deux temps de commutation.

(a) Montrer que le minimum de Φ est atteint en $t = \frac{p_{v_2}(0)}{\lambda}$. En déduire que $0 < t_1 < \frac{p_{v_2}(0)}{\lambda} < t_2 < t_f$.

(b) Montrer que $t_2 = 2\frac{p_{v_2}(0)}{\lambda} - t_1$.

(c) Montrer que $\lambda v_2(t) = g p_{v_2}(t) - T(t)\Phi(t)$, pour tout $t \in [0, t_f]$.

(d) En déduire que $\lambda < 0$ et $p_{v_2}(0) < 0$.

(e) Montrer que $\frac{p_{v_2}(0)}{\lambda} < \frac{v_{20}}{g}$, puis que $v_2(t) \leqslant v_{20} - gt$ pour tout $t \in [0, \frac{v_{20}}{g}]$.

(f) En utilisant le fait que $v_2(t_f) = 0$, montrer qu'en fait ce cas n'arrive jamais.

Corrigé :

1. (a) On pose $f^0 = 0$ et $g = -m$. Le Hamiltonien est

$$\begin{aligned} H &= \langle p_x, v\rangle + \frac{1}{m}\langle p_v, u\rangle - \langle p_v, \vec{g}\rangle - p_m\beta\|u\| \\ &= p_{x_1}v_1 + p_{x_2}v_2 + p_{v_1}\frac{T}{m}\cos\theta + p_{v_2}(\frac{T}{m}\sin\theta - g) - p_m\beta T \end{aligned}$$

Les équations des extrémales sont

Forme complète :

$$\begin{aligned} &\dot{p}_{x_1} = 0 \\ &\dot{p}_{x_2} = 0 \\ &\dot{p}_{v_1} = -p_{x_1} \\ &\dot{p}_{v_2} = -p_{x_2} \\ &\dot{p}_m = \frac{T}{m^2}(p_{v_1}\cos\theta + p_{v_2}\sin\theta) \end{aligned}$$

Forme vectorielle :

$$\begin{aligned} &\dot{p}_x = 0 \\ &\dot{p}_v = -p_x \\ &\dot{p}_m = \frac{1}{m^2}\langle p_v, u\rangle \end{aligned}$$

(b) Les conditions de transversalité s'écrivent $p_{x_1}(0) = 0, p_{x_1}(t_f) = 0$, et $p_m(t_f) = -p^0$ (avec $p^0 \leqslant 0$).

(c) On en déduit que $p_{x_1} = 0$, $p_{x_2} = \text{Cste} = \lambda$, $p_{v_1} = \text{Cste}$, $p_{v_2}(t) = -\lambda t + p_{v_2}(0)$.

(d) Le temps final t_f est libre, donc $H(t_f) = 0$. Comme le système est autonome, on en déduit que $H = 0$ le long de toute extrémale.

2. La condition de maximisation s'écrit

$$\max_{u\in\mathbb{R}^2,\ \|u\|\leqslant T_{max}} \|u\|\left(\frac{\|p_v\|}{m}\left\langle\frac{p_v}{\|p_v\|}, \frac{u}{\|u\|}\right\rangle - p_m\beta\right),$$

d'où l'on déduit que soit $\|u\| = 0$, soit $\frac{u}{\|u\|} = \frac{p_v}{\|p_v\|}$ et la parenthèse doit être positive i.e. $\Phi = \frac{\|p_v\|}{m} - p_m\beta \geqslant 0$. De plus si $\Phi > 0$ alors nécessairement $\|u\| = T_{max}$. On en déduit :

$$\|u(t)\| = T(t) = \begin{cases} 0 & \text{si} \quad \Phi(t) < 0, \\ T_{max} & \text{si} \quad \Phi(t) > 0, \end{cases}$$

d'où la conclusion.

3. On a $\dot{p}_m = \frac{1}{m^2}\langle p_v, u\rangle = \frac{T}{m^2}\|p_v\|$, donc p_m est croissante. Plus précisément, p_m est strictement croissante lorsque $\Phi > 0$, et constante lorsque $\Phi < 0$.

4. (a) Par l'absurde, si $p_v = 0$ sur un intervalle I, alors en dérivant, $p_x = 0$, et par unicité de Cauchy, on a $p_v = p_x = 0$ sur tout $[0, t_f]$. On en déduit que $p_m = \text{Cste} = -p^0$. Par ailleurs, $H = 0$, donc $p^0\beta T = 0$ sur $[0, t_f]$. Comme $x_{2f} > x_{20}$, la poussée T ne peut pas être identiquement nulle sur $[0, t_f]$, d'où $p^0 = 0$. Autrement dit, $(p, p^0) = (0, 0)$, ce qui est une contradiction.

   (b) On dérive $\Phi = \frac{\|p_v\|}{m} - p_m\beta \geqslant 0$, et avec les équations des extrémales, on calcule

$$\dot{\Phi} = -\frac{\langle p_v, p_x\rangle}{m\|p_v\|},$$

   ce qui a bien un sens puisque p_v ne s'annule identiquement sur aucun sous-intervalle, puis

$$\ddot{\Phi} = \frac{\beta T}{m}\dot{\Phi} - \frac{m}{\|p_v\|}\dot{\Phi}^2 + \frac{\|p_x\|^2}{m\|p_v\|}.$$

   (c) D'après l'expression de $\dot{\Phi}$, si $p_x = 0$ alors $\dot{\Phi} = 0$ donc $\Phi = \text{Cste}$. Réciproquement si Φ est constante, alors $\dot{\Phi} = 0$, donc $\langle p_v, p_x\rangle = 0$; en redérivant, on obtient $p_x = 0$, car $\dot{p}_x = 0$ et $\dot{p}_v = -p_x$.

   (d) Si $\Phi = 0$ sur un sous-intervalle I, on est dans le cas singulier. D'abord, on en déduit que $p_x = 0$, et par ailleurs $p_v = \text{Cste}$. De la relation

$$0 = H = \frac{T}{m}\|p_v\| - \langle p_v, \vec{g}\rangle - p_m\beta T = T\Phi - \langle p_v, \vec{g}\rangle = -\langle p_v, \vec{g}\rangle$$

   on déduit que le vecteur (constant) p_v est alors orthogonal à $\vec{g}$, donc est horizontal. Donc $p_{v_2} = 0$ sur tout $[0, t_f]$. En particulier, l'incidence $\theta(t)$ est alors constante égale à 0 sur $[0, t_f]$. Alors $\dot{v}_2 = -g$, donc $v_2(t) = v_{20} - gt$. Comme $v_2(t_f) = 0$ on en déduit que $t_f = \frac{v_{20}}{g}$. Par ailleurs, on intègre $\dot{x}_2 = v_2$, donc $x_2(t) = x_{20} + v_{20}t - \frac{g}{2}t^2$, d'où $x_2(t_f) = x_{20} + \frac{v_{20}^2}{2g}$. Or, $x_2(t_f) = x_{2f}$, ce qui contredit l'hypothèse $x_{2f} > x_{20} + \frac{v_{20}^2}{2g}$.

   (e) Lorsque $p_x = 0$, on a $\Phi = \text{Cste}$, et d'après la question précédente, $\Phi > 0$ ou bien $\Phi < 0$.

– Si $\Phi < 0$, alors puisque Φ est constante on a $T = 0$ sur tout $[0, t_f]$: impossible puisque la poussée ne peut pas être toujours nulle ! (par exemple, parce que $x_{2f} > x_{20}$) Ce cas n'arrive donc pas.
– Si $\Phi > 0$, alors on a tout le temps $T = T_{max}$: la poussée est tout le temps maximale, on n'a aucune commutation.

Au final, si $p_x = 0$ alors la poussée T est constante sur $[0, t_f]$, égale à T_{max}.

(f) Si $p_x \neq 0$ alors Φ est non constante, donc $\dot{\Phi}$ n'est pas identiquement nulle. Si $\dot{\Phi}$ ne s'annule pas sur $[0, t_f]$, alors Φ est strictement monotone, ce qui donne les deux premiers cas. Si $\dot{\Phi}$ s'annule sur $[0, t_f]$, alors d'après l'équation de $\ddot{\Phi}$, là où $\dot{\Phi} = 0$ on a $\ddot{\Phi} > 0$ (car $p_x \neq 0$), et donc ce point est un minimum local. Ce raisonnement montre que tout extrémum de Φ est un minimum local. Par conséquent, la fonction $\dot{\Phi}$ ne peut s'annuler ailleurs, sinon on aurait un autre minimum local, et donc il existerait forcément un maximum local entre ces deux points : ce qui est absurde puisque tout extrémum de Φ est un minimum local. On en déduit donc que Φ admet un unique minimum, est strictement décroissante avant ce point, et strictement croissante ensuite.

(g) D'après les raisonnements précédents, la poussée optimale $T(t)$ est :
– soit constante égale à T_{max} (ce qui peut être impossible selon les données initiales et finales),
– soit du type T_{max} puis 0 (une seule commutation),
– soit du type 0 puis T_{max} (une seule commutation),
– soit du type T_{max} puis 0 puis T_{max} (deux commutations).

5. Raisonnons par l'absurde : si $p^0 = 0$, alors $p_m(t_f) = 0$. Par ailleurs, p_m est croissante, donc $p_m(t) \leqslant 0$ sur $[0, t_f]$. Comme p_v ne s'annule identiquement sur aucun sous-intervalle, on en déduit que $\Phi = \frac{\|p_v\|}{m} - p_m\beta > 0$ p.p. sur $[0, t_f]$, et donc $T = T_{max}$ sur $[0, t_f]$. Cela contredit l'hypothèse d'existence de commutation.

6. (a) En utilisant $H = 0$, $p_{x_1} = 0$ et $p_{x_2} = \lambda$, on a $\lambda v_2(t) + T(t)\Phi(t) - g p_{v_2}(t) = 0$.

(b) Sur $[t_1, t_f]$, on a $T = 0$, donc $\lambda v_2(t) = g p_{v_2}(t)$ d'après la relation précédente. Par ailleurs, pour tout $t \in [t_1, t_f]$, on a $\dot{v}_2 = -g$ donc $v_2(t) = v(t_1) - g(t - t_1)$; de même, on a $\dot{p}_{v_2} = -\lambda$, donc $p_{v_2}(t) = -\lambda t + p_{v_2}(0)$. Ce qui conduit à $\lambda v_2(t_1) + \lambda g t_1 = g p_{v_2}(0)$.

(c) En $t = t_f$, la relation de la question (a) donne $g p_{v_2}(t_f) = 0$, car $v_2(t_f) = 0$. Donc $p_{v_2}(t_f) = 0$. Comme $p_{v_2}(t_f) = -\lambda t_f +$

$p_{v_2}(0)$, on en déduit que $t_f = \frac{p_{v_2}(0)}{\lambda}$.

(d) Comme $t_f = \frac{p_{v_2}(0)}{\lambda}$, on en déduit que λ et $p_{v_2}(0)$ sont de même signe. Montrons en fait que $p_{v_2}(0) > 0$. Par l'absurde, si $p_{v_2}(0) < 0$ alors, comme p_{v_2} est affine et $p_{v_2}(t_f) = 0$, on a $p_{v_2}(t) \leqslant 0$ pour tout $t \in [0, t_f]$. Donc $\sin\theta(t) \leqslant 0$, et $\dot{v}_2 \leqslant -g$ sur $[0, t_f]$. En intégrant, on obtient $x_2(t) \leqslant x_{20} + v_{20}t - \frac{g}{2}t^2 \leqslant x_{20} + \frac{v_{20}^2}{2g}$. Au temps t_f, cela contredit l'hypothèse $x_{2f} > x_{20} + \frac{v_{20}^2}{2g}$. Et donc, $\lambda > 0$ et $p_{v_2}(0) > 0$.
Par ailleurs, pout tout $t \in [0, t_1[$, on a $\dot{v}_1 = \frac{T_{max}}{m}\cos\theta$ avec

$$\cos\theta = \frac{p_{v_1}}{\sqrt{p_{v_1}^2 + (-\lambda t + p_{v_2}(0))^2}},$$

puis, pour tout $t \in]t_1, t_f]$, $T = 0$ donc v_1 reste constante. Comme $v_{1f} > v_{10}$, cela impose $p_{v_1} > 0$.

(e) A priori, en mettant en oeuvre une méthode de tir, on a 5 inconnues, à savoir :
– le vecteur adjoint initial $(\lambda, p_{v_1}, p_{v_2}(0), p_m(0), p^0)$ (notons que $p^0 \neq 0$), défini à scalaire multiplicatif près,
– le temps final t_f,
et 5 équations :

$$x_2(t_f) = x_{2f},\ v_1(t_f) = v_{1f},\ v_2(t_f) = 0,\ p_m(t_f) = -p^0,\ H(t_f) = 0.$$

Au lieu de faire la normalisation habituelle $p^0 = -1$, le vecteur adjoint étant défini à scalaire mutliplicatif près, on choisit plutôt, comme $\lambda > 0$, de le normaliser de sorte que $\lambda = 1$. La variable p^0 ne sert qu'à ajuster l'équation $p_m(t_f) = -p^0$. Donc on peut oublier la variable p^0 et l'équation $p_m(t_f) = -p^0$.
Il reste alors 4 inconnues $(p_{v_1}, p_{v_2}(0), p_m(0))$ et t_f, pour 4 équations :

$$x_2(t_f) = x_{2f},\ v_1(t_f) = v_{1f},\ v_2(t_f) = 0,\ H(t_f) = 0.$$

Remarquons que la connaissance de $p_m(0)$ permet de déterminer la fonction de commutation $\Phi(t)$, et donc, le temps de commutation t_1. On peut donc remplacer la variable $p_m(0)$ par la variable t_1. On a alors les 4 inconnues $(p_{v_1}, p_{v_2}(0), t_1, t_f)$, et les 4 équations précédentes.
Par ailleurs, dans les calculs précédents, et avec $\lambda = 1$, on voit que le système d'équations

$$v_2(t_f) = 0,\ H(t_f) = 0$$

est équivalent au système d'équations

$$t_f = p_{v_2}(0),\ v_2(t_1) + gt_1 = gp_{v_2}(0).$$

On a alors 4 inconnues $(p_{v_1}, p_{v_2}(0), t_1, t_f)$ pour 4 équations :

$$x_2(t_f) = x_{2f},\ v_1(t_f) = v_{1f},\ v_2(t_1)+gt_1 = gp_{v_2}(0),\ t_f = p_{v_2}(0).$$

Le temps final t_f étant directement déterminé, on se ramène finalement à 3 inconnues $(p_{v_1}, p_{v_2}(0), t_1)$ pour 3 équations :

$$x_2(t_f) = x_{2f},\ v_1(t_f) = v_{1f},\ v_2(t_1) + gt_1 = gp_{v_2}(0).$$

La méthode de tir se réduit donc à la résolution de ce système de 3 équations, sachant de plus que $p_{v_1} > 0$ et $p_{v_2}(0) > 0$. Pour la programmation, on initialise de telles valeurs de p_{v_1} et $p_{v_2}(0)$. On résout numériquement l'équation différentielle pour déterminer $(x_2(t), v_1(t), v_2(t))$, et on arrête l'intégration au premier temps t_1 vérifiant $v_2(t_1) + gt_1 = gp_{v_2}(0)$ (en Matlab, on utilise un "events"). Sur l'intervalle $[t_1, t_f]$ (avec $t_f = p_{v_2}(0)$), on calcule explicitement $x_2(t)$ et $v_1(t)$:

$$v_1(t) = \text{Cste} = v_1(t_1),\ x_2(t) = x_2(t_1)+v_2(t_1)(t-t_1)-\frac{g}{2}(t-t_1)^2$$

et on résout le système d'équations

$$x_2(t_f) = x_{2f},\ v_1(t_f) = v_{1f},$$

par une méthode de Newton.

Notons qu'on peut calculer des expressions explicites de $x_2(t)$ et $v_1(t)$ sur tout l'intervalle $[0, t_f]$, mais numériquement il s'avère que cela ne fait pas gagner de temps.

7. (a) D'après l'expression de $\dot{\Phi}$, lorsque $\dot{\Phi}(t) = 0$, on a $\lambda p_{v_2}(t) = 0$. Comme Φ est non constante, on a $\lambda \neq 0$, donc $p_{v_2}(t) = 0$, d'où $t = \frac{p_{v_2}(0)}{\lambda}$. Ce minimum est atteint dans l'intervalle $]0, t_f[$ par définition.

   (b) Sur l'intervalle $[t_1, t_2]$, on a $T(t) = 0$, donc en particulier $m(t)$ et $p_m(t)$ restent constantes : $m(t_1) = m(t_2)$ et $p_m(t_1) = p_m(t_2)$. Or, la fonction $\Phi(t) = \frac{\sqrt{p_{v_1}^2+(p_{v_2}(0)-\lambda t)^2}}{m(t)} - p_m(t)\beta$ s'annule par définition en t_1 et t_2. On en déduit que

$$\sqrt{p_{v_1}^2 + (p_{v_2}(0) - \lambda t_1)^2} = \sqrt{p_{v_1}^2 + (p_{v_2}(0) - \lambda t_2)^2},$$

d'où $|p_{v_2}(0) - \lambda t_1| = |p_{v_2}(0) - \lambda t_2|$, puis, comme $t_1 \neq t_2$, on obtient $t_2 = 2\frac{p_{v_2}(0)}{\lambda} - t_1$.

Notons que cela illustre le fait que le graphe de Φ sur l'intervalle $[t_1, t_2]$ est symétrique par rapport au point $t = \frac{p_{v_2}(0)}{\lambda}$ où le minimum est atteint.

(c) Même raisonnement qu'en question 6.a.

(d) En prenant la relation de la question 7.c en $t = t_f$, et en remarquant que $\Phi(t_f) > 0$ et $v_2(t_f) = 0$, on en déduit que $p_{v_2}(t_f) > 0$. Comme $p_{v_2}(\cdot)$ est affine et s'annule en $\frac{p_{v_2}(0)}{\lambda}$, cela impose que $\lambda < 0$ et $p_{v_2}(0) < 0$.

(e) Sur l'intervalle $[0, t_1]$, on a $p_{v_2}(t) < 0$, donc $\dot{v}_2(t) < -g$, et donc, en intégrant, $v_2(t_1) < v_{20} - gt_1$. Par ailleurs, d'après la relation de la question 7.c., on a $v_2(t_1) = g(\frac{p_{v_2}(0)}{\lambda} - t_1)$. On en déduit que $\frac{p_{v_2}(0)}{\lambda} < \frac{v_{20}}{g}$.
Sur l'intervalle $[0, t_1]$, on a $\dot{v}_2(t) < -g$, et sur l'intervalle $[t_1, t_2]$ on a $\dot{v}_2(t) = -g$ et plus précisément $v_2(t) = g(\frac{p_{v_2}(0)}{\lambda} - t)$ (ce qui signifie en particulier que la fonction $v_2(\cdot)$ est strictement décroissante sur $[0, t_2]$ et s'annule en $\frac{p_{v_2}(0)}{\lambda}$; par ailleurs sur l'intervalle $[t_2, t_f]$ la fonction $v_2(\cdot)$ est soit croissante, soit décroissante puis croissante, et s'annule en t_f). On déduit en particulier de tout cela que $v_2(t) \leqslant v_{20} - gt$ pour tout $t \in [0, \frac{v_{20}}{g}]$. Notons que $\frac{v_{20}}{g} > \frac{p_{v_2}(0)}{\lambda}$.

(f) On déduit de la question précédente que $\dot{h}(t) \leqslant v_{20} - gt$ pour tout $t \in [0, \frac{v_{20}}{g}]$, et donc, en intégrant, $h(t) \leqslant h_0 + v_{20}t - \frac{g}{2}t^2$. Le minimum de ce trinôme étant $h_0 + \frac{v_{20}^2}{2g}$, on en déduit que $h(t) \leqslant h_0 + \frac{v_{20}^2}{2g}$ pour tout $t \in [0, \frac{v_{20}}{g}]$. Notons que, par hypothèse, $h_0 + \frac{v_{20}^2}{2g} < h_f$. On obtient donc une contradiction si $\frac{v_{20}}{g} \geqslant t_f$ (puisqu'on doit avoir $h(t_f) = h_f$).
Si $\frac{v_{20}}{g} < t_f$, vu que par ailleurs $\frac{v_{20}}{g} > \frac{p_{v_2}(0)}{\lambda}$, la fonction $v_2(\cdot)$ est négative sur l'intervalle $[\frac{v_{20}}{g}, t_f]$ et donc $h(\cdot)$ est décroissante sur cet intervalle ; donc $h(t_f) \leqslant h_0 + \frac{v_{20}^2}{2g} < h_f$ et on a également une contradiction.

Exercice 7.3.22 (Sujet d'examen : Contrôle optimal d'insectes nuisibles par des prédateurs.)**.** Pour traiter une population $x_0 > 0$ d'insectes nuisibles, on introduit dans l'écosystème une population $y_0 > 0$ d'insectes prédateurs (non nuisibles), se nourrissant des nuisibles.

1. Dans la première partie du problème, on suppose que les insectes prédateurs que l'on introduit sont stériles, et ne peuvent donc pas se reproduire. Le contrôle consiste en l'introduction régulière d'insectes prédateurs. Le modèle s'écrit

$$\begin{aligned} \dot{x}(t) &= x(t)(a - by(t)), \quad x(0) = x_0, \\ \dot{y}(t) &= -cy(t) + u(t), \qquad y(0) = y_0, \end{aligned}$$

où $a > 0$ est le taux de reproduction naturelle des nuisibles, $b > 0$ est un taux de prédation, $c > 0$ est le taux de disparition naturelle des prédateurs. Le contrôle $u(t)$ est le taux d'introduction de nouveaux prédateurs au temps t, il vérifie la contrainte

$$0 \leqslant u(t) \leqslant M,$$

où $M > 0$. On cherche à minimiser, au bout d'un temps $T > 0$ fixé, le nombre de nuisibles, tout en cherchant à minimiser la quantité globale de prédateurs introduits ; autrement dit on veut minimiser

$$x(T) + \int_0^T u(t)dt.$$

On note les variables adjointes $p = (p_x, p_y)$ et p^0.

(a) Démontrer que, pour tout contrôle u, $x(t) > 0$ et $y(t) > 0$ sur $[0, T]$.

(b) Ecrire le Hamiltonien du problème de contrôle optimal et les équations des extrémales.

(c) Ecrire les conditions de transversalité.

(d) Montrer que $p^0 \neq 0$. Que posez-vous pour la suite ?

(e) Démontrer que la fonction $t \mapsto x(t)p_x(t)$ est constante sur $[0, T]$. Exprimer cette constante en fonction de $x(T)$.

(f) En déduire une expression de $p_y(t)$, pour $t \in [0, T]$.

(g) Démontrer que les contrôles optimaux sont bang-bang, et préciser leur expression.
(*indication :* démontrer, par l'absurde, que la fonction $t \mapsto p_y(t) - 1$ ne peut s'annuler identiquement sur un sous-intervalle)

(h) Montrer qu'il existe $\varepsilon > 0$ tel que $u(t) = 0$, pour presque tout $t \in [T - \varepsilon, T]$ (autrement dit, le contrôle u vaut 0 à la fin).

(i) Montrer qu'en fait le contrôle optimal u admet au plus une commutation sur $[0, T]$. S'il y en a une, préciser en quel temps $t_1 \in [0, T]$ arrive cette commutation.
(on ne cherchera pas à établir des conditions sur les données initiales pour qu'il existe une telle commutation)

2. Dans la deuxième partie du problème, on suppose que les prédateurs que l'on introduit se reproduisent, de manière proportionnelle au nombre de nuisibles. Cette fois le contrôle est le taux de disparition des prédateurs. Pour simplifier l'écriture on normalise les variables de façon à ce que les autres taux soient égaux à 1. Le modèle s'écrit alors

$$\begin{aligned} \dot{x}(t) &= x(t)(1 - y(t)), & x(0) &= x_0, \\ \dot{y}(t) &= -y(t)(u(t) - x(t)), & y(0) &= y_0, \end{aligned}$$

où le contrôle $u(t)$ vérifie la contrainte

$$0 < \alpha \leqslant u(t) \leqslant \beta.$$

(a) Démontrer que, pour tout contrôle u, $x(t) > 0$ et $y(t) > 0$ sur $[0, T]$.

(b) On rappelle que, de manière générale, un point d'équilibre d'un système de contrôle $\dot{x}(t) = f(x(t), u(t))$ est un couple (x_e, u_e) tel que $f(x_e, u_e) = 0$.
Donner tous les points d'équilibre du système dans le quadrant $x > 0, y > 0$ (et les représenter sur un graphique dans ce quadrant).

(c) On cherche à résoudre le problème de joindre en temps minimal le point d'équilibre $x(t_f) = a$, $y(t_f) = 1$.

i. Ecrire le Hamiltonien de ce problème de contrôle optimal et les équations des extrémales.

ii. Ecrire les conditions de transversalité.

iii. Montrer que le Hamiltonien est égal à 0 le long de toute extrémale.

iv. Démontrer que les contrôles optimaux sont bang-bang, et préciser leur expression.

v. Montrer que, le long d'un arc où le contrôle u est égal à α (resp. β), la fonction

$$F_\alpha(x, y) = x + y - \alpha \ln x - \ln y$$

(resp. la fonction F_β, en remplaçant α par β dans la formule) reste constante le long de cet arc.

vi. Montrer que la fonction F_α admet un minimum global strict au point $(\alpha, 1)$.

vii. Calculer $\frac{d}{dt} F_\alpha(x(t), y(t))$, où la trajectoire $(x(\cdot), y(\cdot))$ est associée à un contrôle $u(\cdot)$ quelconque.

viii. Montrer qu'il existe $\varepsilon > 0$ tel que $u(t) = \alpha$, pour presque tout $t \in [t_f - \varepsilon, t_f]$ (autrement dit, le contrôle u vaut α à la fin).

ix. Supposons que les données initiales x_0 et y_0 sont telles que $\alpha < x_0 < \beta$ et $y_0 = 1$. En admettant que la trajectoire optimale admet une seule commutation, donner la structure du contrôle optimal et expliquer comment la construire géométriquement dans le plan (x, y).

x. En extrapolant la construction précédente, donner une stratégie de contrôle pour relier n'importe quel point (x_0, y_0) du quadrant au point $(\alpha, 1)$, et décrire comment mettre en oeuvre numériquement cette stratégie.

Corrigé :

1. (a) Comme $u(t) \geqslant 0$, on a $\dot{y}(t) \geqslant -cy(t)$, donc $y(t) \geqslant y_0 \mathrm{e}^{-ct} > 0$. Concernant $x(t)$, on raisonne par l'absurde : s'il existe $t_1 \in [0, T]$ tel que $x(t_1) = 0$, alors $x(t) = 0$ pour tout t, par unicité de Cauchy ; cela est absurde car $x(0) = x_0 > 0$.

(b) $H = p_x x(a - by) + p_y(-cy + u) + p^0 u$, et les équations adjointes sont $\dot{p}_x = -p_x(a - by)$, $\dot{p}_y = bp_x x + cp_y$.

(c) $p_x(T) = p^0$ et $p_y(T) = 0$.

(d) Si $p^0 = 0$ alors tout le vecteur adjoint est nul, ce qui est absurde. Dans la suite on pose $p^0 = -1$.

(e) $\frac{d}{dt}x(t)p_x(t) = x(t)p_x(t)(a-by(t))-x(t)p_x(t)(a-by(t)) = 0$, donc $x(t)p_x(t) = \text{Cste} = -x(T)$ car $p_x(T) = p^0 = -1$.

(f) L'équation en p_y devient alors $\dot{p}_y = -bx(T) + cp_y$. Comme $p_y(T) = 0$, on obtient, en intégrant,

$$p_y(t) = \frac{b}{c}x(T)(1 - \mathrm{e}^{c(t-T)}).$$

(g) La condition de maximisation s'écrit $\max_{0\leqslant u\leqslant M}(p_y - 1)u$, ce qui conduit à

$$u(t) = \begin{cases} 0 & \text{si} \quad p_y(t) - 1 < 0, \\ M & \text{si} \quad p_y(t) - 1 > 0, \end{cases}$$

sauf si la fonction $t \mapsto p_y(t) - 1$ s'annule identiquement sur un sous-intervalle. Supposons, par l'absurde, que ce soit le cas : $p_y(t) = 1$ pour tout $t \in I$. Cela contredit alors le résultat de la question précédente qui montre en particulier que la fonction p_y est strictement décroissante. Donc la fonction $t \mapsto p_y(t) - 1$ ne s'annule identiquement sur aucun sous-intervalle, et donc le contrôle optimal est bang-bang, donné par l'expression ci-dessus.

(h) A la fin, $p_y(T) - 1 = -1$, donc, par continuité, il existe $\varepsilon > 0$ tel que $p_y(t) - 1 < 0$ sur $[T - \varepsilon, T]$, et donc $u(t) = 0$.

(i) La fonction p_y est strictement décroissante (car $x(T) > 0$ par la première question), donc la fonction $t \mapsto p_y(t) - 1$, qui est égale à -1 en $t = T$, s'annule au plus une fois. Donc le contrôle optimal admet au plus une commutation sur $[0, T]$. S'il y a une commutation, elle doit avoir lieu en $t_1 \in [0, T]$ tel que $p_y(t_1) = 1$, ce qui conduit à

$$t_1 = T + \frac{1}{c}\ln\left(1 - \frac{c}{bx(T)}\right).$$

Notons que cette commutation ne peut avoir lieu que si $t_1 > 0$ (on a bien, par ailleurs, $t_1 < T$), donc, si $x(T) > \frac{c}{b}\frac{1}{1-\mathrm{e}^{-cT}}$. En intégrant en temps inverse les équations, on pourrait remonter à une condition implicite sur les données initiales pour que cette inégalité soit vraie, donc, pour qu'il y ait une commutation.

2. (a) Même raisonnement que pour $x(t)$ dans la question 1.a.

(b) Les points d'équilibre sont $x_e = u_e$, $y_e = 1$, pour tout $\alpha \leqslant u_e \leqslant \beta$. On a donc, dans le quadrant, un segment de points d'équilibres.

(c) i. $H = p_x x(1-y) - p_y y(u-x) + p^0$, et les équations adjointes sont $\dot{p}_x = -p_x(1-y) - p_y y$, $\dot{p}_y = p_x x + p_y(u-x)$.

ii. $H(t_f) = 0$.

iii. Le système étant autonome, le Hamiltonien est constant le long de toute extrémale, et cette constante est nulle puisque $H(t_f) = 0$.

iv. La condition de maximisation s'écrit $\max_{0\leqslant u\leqslant M}(-p_y y u)$, ce qui conduit, puisque $y(t) > 0$, à

$$u(t) = \begin{cases} \alpha & \text{si} \quad p_y(t) > 0, \\ \beta & \text{si} \quad p_y(t) < 0, \end{cases}$$

sauf si la fonction $t \mapsto p_y(t)$ s'annule identiquement sur un sous-intervalle. Supposons, par l'absurde, que ce soit le cas : $p_y(t) = 0$ pour tout $t \in I$. D'après l'équation différentielle en p_y, cela conduit à $x p_x = 0$ sur I, donc $p_x = 0$ sur I. Donc, sur I, on a $H = p^0$, et comme $H = 0$ d'après la question précédente, on en déduit $p^0 = 0$, d'où une contradiction car le vecteur adjoint (p_x, p_y, p^0) doit être non trivial. Donc la fonction $t \mapsto p_y(t)$ ne s'annule identiquement sur aucun sous-intervalle,et donc le contrôle optimal est bang-bang, donné par l'expression ci-dessus.

v. Le long d'un arc où $u = \alpha$, on calcule immédiatement $\frac{d}{dt} F_\alpha(x(t), y(t)) = 0$.
Notons que, formellement, on obtient cette intégrale première en calculant $\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{-y}{1-y}\frac{\alpha - x}{x}$ et en intégrant cette forme à variables séparées.

vi. On fait un développement limité à l'ordre 2 au point $(\alpha, 1)$:

$$F_\alpha(\alpha+h, 1+k) = \alpha - \alpha \ln \alpha + 1 + \frac{1}{2}\left(\frac{h^2}{\alpha} + k^2\right) + \mathrm{o}(h^2+k^2).$$

Cela montre que F_α admet un minimum local strict au point $(\alpha, 1)$. Pour montrer que ce minimum est global, il suffit de remarquer que la fonction F_α est (strictement) convexe, ce qui découle du fait que sa Hessienne

$$\begin{pmatrix} \frac{a}{x^2} & 0 \\ 0 & \frac{1}{y^2} \end{pmatrix}$$

est symétrique définie positive en tout point du quadrant $x > 0$, $y > 0$.

vii. On calcule $\frac{d}{dt}F_\alpha(x(t), y(t)) = (u(t) - \alpha)(1 - y(t))$.

viii. A la fin, on a soit $p_y(t_f) = 0$, soit $p_y(t_f) \neq 0$.

Si $p_y(t_f) = 0$, alors, d'après l'équation en p_y, on a $\dot{p}_y(t_f) = p_x(t_f)a$. Forcément $p_x(t_f) \neq 0$ (sinon, on obtient une contradiction, comme précédemment, en remarquant que $H(t_f) = p^0 = 0$). Donc $\dot{p}_y(t_f) \neq 0$, et par conséquent la fonction p_y est de signe fixe dans un intervalle du type $[t_f - \varepsilon, t_f[$. Donc, sur cet intervalle, le contrôle est constant, soit égal à α soit égal à β. Il ne peut être égal à α car sinon, d'après la question précédente, la fonction F_α serait constante le long de cet arc, et comme l'arc doit atteindre le point $(\alpha, 1)$, cette constante serait égale au minimum de F_α, ce qui imposerait donc que l'arc soit constant, égal au point $(\alpha, 1)$: cela est absurde car on doit avoir une trajectoire temps-minimale arrivant au point $(\alpha, 1)$.

Si $p_y(t_f) \neq 0$, alors la fonction p_y est de signe fixe dans un intervalle du type $[t_f - \varepsilon, t_f[$, et donc, sur cet intervalle, le contrôle est constant, soit égal à α soit égal à β. Le raisonnement précédent s'applique de nouveau, et $u = \alpha$.

ix. Au voisinage du point $(\alpha, 1)$, les courbes de niveau de la fonction F_α ressemblent à des cercles. En fait plus on s'éloigne de ce point, et plus les courbes de niveau ressemblent à des triangles rectangles, asymptotiques aux axes des abscisses et des ordonnées. Idem pour les courbes de niveau de la fonction F_β, relativement au point $(\beta, 1)$.

Partons du point (x_0, y_0), qui est situé sur le segment reliant les deux points $(\alpha, 1)$ et $(\beta, 1)$. On part avec le contrôle $u = \alpha$, et on reste sur une courbe de niveau de la fonction F_α (donc, "centrée" sur le point $(\alpha, 1)$). A un moment donné, on commute sur le contrôle $u = \beta$, et on reste sur la courbe de niveau de la fonction F_β (donc, "centrée" sur le point $(\alpha, 1)$) qui passe par le point final visé $(\alpha, 1)$.

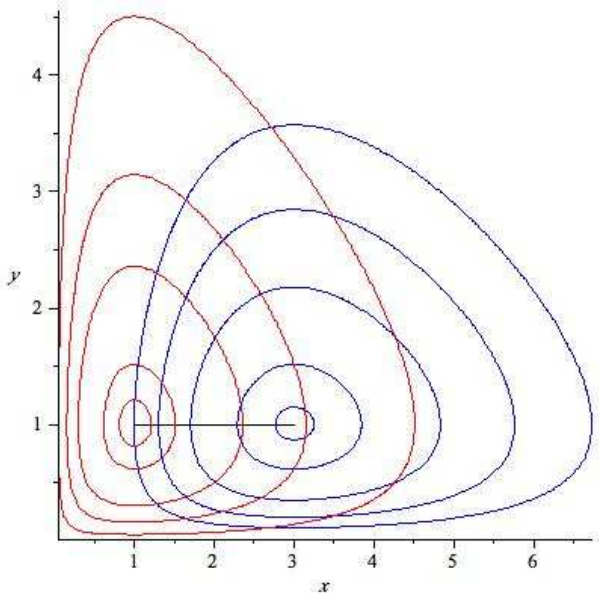


FIGURE 7.4 – Exemple avec $\alpha = 1$ et $\beta = 3$

x. Si on part de n'importe quel point, on détermine graphiquement une séquence d'arcs sur les courbes de niveau respectivement de F_α et F_β qui relie le point de départ au point d'arrivée.

Exercice 7.3.23 (Projet : transfert orbital d'un satellite en temps minimal.)**.** Un problème important en mécanique spatiale est de transférer un engin spatial soumis à l'attraction terrestre sur une ellipse Keplerienne ou en un point de cet ellipse, pour le problème de rendez-vous avec un autre engin. Ce type de problème classique a été réactualisé avec la technologie des moteurs à poussée faible et continue. L'objectif de ce projet est d'appliquer des techniques de contrôle optimal pour réaliser numériquement un problème de transfert en temps minimal, avec poussée faible, sur une orbite géostationnaire.

Modélisation du problème. Le satellite est assimilé à un point matériel de masse m, soumis à l'attraction terrestre et à une force de propulsion F. En première approximation, le système s'écrit

$$\ddot{q} = -\mu \frac{q}{\|q\|^3} + \frac{F}{m},$$

où q désigne le vecteur position du satellite dans un référentiel dont l'origine est le centre de la terre, μ la constante de gravitation de la planète. Le système libre $F = 0$ correspond aux équations de Kepler. Pratiquement, la poussée est limitée, *i.e.* $\|F\| \leqslant F_{\max}$ et on peut changer son orientation. La propulsion se

fait par éjection de matière, à vitesse v_e et il faut rajouter au système l'équation

$$\dot{m} = -\frac{F}{v_e},$$

et dans le problème à poussée faible, la force de poussée est petite comparée à la force d'attraction. L'état du système est $(q, \dot{q})$ et le problème de transfert orbital à résoudre est de transférer le système d'un état initial à une orbite géostationnaire en temps minimal. On contrôle la poussée de l'engin.

Ici, on considère le problème de transfert orbital à masse variable dans le plan, que l'on représente dans des coordonnées dites équinoxiales (p, e_x, e_y, L), où p est appelé le paramètre, (e_x, e_y) est appelé vecteur excentricité et L est la longitude. Le contrôle est décomposé dans le repère radial-orthoradial, ce qui conduit aux équations suivantes

$$\begin{aligned}
\dot{p} &= \frac{2}{W}\sqrt{\frac{p^3}{\mu}}\frac{T_{\max}}{m}u_{or} \\
\dot{e}_x &= \sqrt{\frac{p}{\mu}}\left(\frac{e_x+\cos L}{W}+\cos L\right)\frac{T_{\max}}{m}u_{or} + \sqrt{\frac{p}{\mu}}\sin L\frac{T_{\max}}{m}u_r \\
\dot{e}_y &= \sqrt{\frac{p}{\mu}}\left(\frac{e_y+\sin L}{W}+\sin L\right)\frac{T_{\max}}{m}u_{or} - \sqrt{\frac{p}{\mu}}\cos L\frac{T_{\max}}{m}u_r \\
\dot{L} &= \frac{W^2}{p}\sqrt{\frac{\mu}{p}} \\
\dot{m} &= -\delta T_{\max}|u|
\end{aligned}$$

o $W = 1 + e_x\cos L + e_y \sin L$, où $|u| = \sqrt{u_{or}^2 + u_o^2}$, et où $T_{\max}$ est la valeur maximale du module de poussée. Le problème consiste donc, en respectant la contrainte $u_r^2 + u_{or}^2 \leqslant 1$, à minimiser le temps de transfert d'une orbite basse définie par $p(0) = 11625$ km, $e_x(0) = 0.75$, $e_y(0) = 0$, $L(0) = \pi$, à une orbite géostationnaire définie par $p(t_f) = 42165$ km, $e_x(t_f) = 0$, $e_y(t_f) = 0$, la longitude finale étant libre.

Questions.

1. Pour des raisons numériques évidentes il est préférable de normaliser la variable p en posant $\bar{p} = \dfrac{p}{p(t_f)}$. En introduisant les constantes $\alpha = \sqrt{\dfrac{\mu}{p_f}}$ et $\varepsilon = \sqrt{\dfrac{p_f}{\mu}}T_{\max}$, montrer que les équations s'écrivent

$$\begin{aligned}
\dot{q} &= F_0(q) + u_r F_r(q) + u_{or}F_{or}(q), \\
\dot{m} &= -\delta T_{\max}|u|,
\end{aligned}$$

avec $q = (\bar{p}, e_x, e_y, l)$ et où les champs de vecteurs F_0, F_r et F_{or} sont définis par

$$F_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \alpha \dfrac{W^2}{\sqrt{\bar{p}^{3/2}}} \end{bmatrix}, \quad F_r = \begin{bmatrix} 0 \\ \dfrac{\varepsilon}{m}\sqrt{\bar{p}}\sin L \\ -\dfrac{\varepsilon}{m}\sqrt{\bar{p}}\cos L \\ 0 \end{bmatrix}, \quad F_{or} = \begin{bmatrix} 2\dfrac{\varepsilon}{m}\dfrac{\sqrt{\bar{p}^3}}{W} \\ \dfrac{\varepsilon}{m}\sqrt{\bar{p}}\left(\dfrac{e_x + \cos L}{W} + \cos L\right) \\ \dfrac{\varepsilon}{m}\sqrt{\bar{p}}\left(\dfrac{e_y + \sin L}{W} + \sin L\right) \\ 0 \end{bmatrix}$$

et écrire le Hamiltonien de ce système sous la forme

$$H = \langle \lambda, F_0 + u_r F_r(q) + u_{or} F_{or}(q) \rangle - \lambda_m \delta T_{\max} |u|,$$

où $(\lambda, \lambda_m) = (\lambda_{\bar{p}}, \lambda_{e_x}, \lambda_{e_y}, \lambda_l, \lambda_m)$ est le vecteur adjoint.

2. Montrer que les contrôles extrémaux vérifient $u_r^2 + u_{or}^2 = 1$ (pour cela, on essaiera de montrer que $\lambda_m(t)$ est toujours négatif). En déduire en particulier que $m(t) = m_0 - \delta T_{max} t$, et montrer que les contrôles extrémaux s'écrivent

$$u_r = \frac{\langle \lambda, F_r \rangle}{\sqrt{\langle \lambda, F_r \rangle^2 + \langle \lambda, F_{or} \rangle^2}}, \quad u_{or} = \frac{\langle \lambda, F_{or} \rangle}{\sqrt{\langle \lambda, F_r \rangle^2 + \langle \lambda, F_{or} \rangle^2}}.$$

3. En remarquant que l'on peut oublier la variable adjoint λ_m, écrire les équations du système extrémal données par le principe du maximum (faire les calculs à l'aide de *Maple*).

4 (application numérique). En utilisant une méthode de tir multiple, déterminer numériquement un vecteur adjoint initial $(\lambda_{\bar{p}}(0), \lambda_{e_x}(0), \lambda_{e_y}(0), \lambda_L(0))$ pour lequel la trajectoire extrémale vérifie les conditions initiales et finales imposées.

Comme données numériques, on prendra

$$m_0 = 1500 \text{ kg}, \ \delta = 0.05112 \text{ km}^{-1}.\text{s}, \ \mu = 398600.47 \text{ km}^3.\text{s}^{-2},$$

et on choisira une valeur T_{max} de plus en plus petite (problème à faible poussée). Par exemple on pourra prendre successivement

$$T_{max} = 60, \ 24, \ 12, \ 9, \ 6, \ 3, \ 2, \ 1.4, \ 1, \ 0.7, \ 0.5, \ 0.3 \quad \text{N}.$$

5. Pour $T_{max} = 0.3$, on se propose de stabiliser le système autour de la trajectoire construite dans la question précédente. Pour tenir compte de la contrainte sur le contrôle, on modifie la trajectoire obtenue selon la nouvelle contrainte $|u| \leqslant 1 - \varepsilon$, où ε est un petit paramètre. On choisit par exemple $\varepsilon = 0.1$.

5.1. Modifier la trajectoire précédente en tenant compte de cette nouvelle contrainte sur le contrôle.

5.2. Linéariser le système le long de la nouvelle trajectoire obtenue, et proposer une méthode de stabilisation LQ. Effectuer les simulations numériques.

7.4 Contrôle optimal et stabilisation d'une navette spatiale

Dans cette section, on traite en totalité un exemple d'application de la théorie du contrôle optimal.

On s'intéresse au problème de contrôle optimal d'une navette spatiale en phase de rentrée atmosphérique, où le contrôle est l'angle de gîte, et le coût est le flux thermique total (facteur d'usure de la navette). L'objectif est de déterminer une trajectoire optimale jusqu'à une cible donnée, puis de stabiliser le système autour de cette trajectoire nominale, sachant que la navette est de plus soumise à des contraintes sur l'état.

Le problème de rentrée atmosphérique présenté ici est simplifié. Le problème complet est difficile et a été complètement résolu dans une série d'articles [14, 15, 17].

On présente d'abord une modélisation du problème de rentrée atmosphérique et on pose un problème de contrôle optimal. Ensuite, on résout numériquement ce problème de contrôle optimal, et on détermine ainsi une trajectoire nominale (trajectoire de référence) pour la navette. Enfin, on utilise la théorie LQ pour stabiliser la navette autour de la trajectoire nominale précédemment déterminée.

7.4.1 Modélisation du problème de rentrée atmosphérique

Présentation du projet

Ce projet a été posé par le CNES, et est motivé par l'importance croissante de la théorie du contrôle, et du contrôle optimal, dans les techniques d'aérocapture :

- problèmes de guidage, transferts d'orbites aéroassistés,
- développement de lanceurs de satellites *récupérables* (l'enjeu financier très important),
- problèmes de rentrée atmosphérique : c'est l'objet du fameux projet *Mars Sample Return* développé par le CNES, qui consiste à envoyer une navette spatiale habitée vers la planète Mars, dans le but de ramener sur Terre des échantillons martiens.

En gros, le rôle de l'arc atmosphérique est

- de réduire suffisamment l'énergie cinétique, par frottement dans l'atmosphère ;
- d'amener l'engin spatial d'une position initiale précise à une cible donnée ;
- de plus il faut prendre en compte certaines contraintes sur l'état : contrainte sur le flux thermique (car il y a des gens à l'intérieur de la navette !), sur l'accélération normale (confort de vol), et sur la pression dynamique (contrainte technique de structure),
- enfin, on cherche de plus à minimiser un critère d'optimisation : le flux thermique total de la navette.

Une trajectoire optimale étant ainsi déterminée, il faut ensuite *stabiliser* la navette autour de cette trajectoire, de façon à prendre en compte de possibles perturbations.

Le contrôle est la configuration aérodynamique de la navette. La première question qui se pose est la suivante : les forces aérodynamiques peuvent-elles contribuer à freiner la navette de manière adéquate ? En fait si l'altitude est trop élevée (supérieure à 120 km), alors la densité atmosphérique est trop faible, et il est physiquement impossible de générer des forces aérodynamiques suffisammanent intenses. Au contraire, si l'altitude est trop basse (moins de 20 km), la densité atmosphérique est trop grande, et le seul emploi des forces aérodynamiques conduirait à un dépassement du seuil autorisé pour le flux thermique ou la pression dynamique. En effet la rentrée atmosphérique s'effectue à des vitesses très élevées. En revanche si l'altitude est comprise entre 20 et 120 km, on peut trouver un compromis. C'est ce qu'on appelle la *phase atmosphérique*.

Durant cette phase atmosphérique, la navette se comporte comme un *planeur*, c'est-à-dire que les moteurs sont coupés : il n'y a pas de force de poussée. L'engin est donc soumis uniquement à la force de gravité et aux forces aérodynamiques. Le contrôle est l'angle de gîte qui représente l'angle entre les ailes et un plan contenant la navette. Enfin, on choisit comme critère d'optimisation le flux thermique total de la navette.

La modélisation précise du problème a été effectuée dans [17]. Nous la rappelons maintenant.

Modélisation du problème

On utilise les lois de la mécanique classique, un modèle de densité atmosphérique et un modèle pour les forces s'exerçant sur la navette, la force gravitationnelle et la force aérodynamique qui se décompose en une composante dite de *traînée* et une composante dite de *portance*. Le système est mono-entrée et le contrôle est la *gîte cinématique* (l'angle d'attaque est fixé).

On donne un modèle général tenant compte de la rotation (uniforme) de la Terre autour le l'axe $K = NS$, à vitesse angulaire de module Ω. On note $E = (e_1, e_2, e_3)$ un repère galiléen dont l'origine est le centre O de la Terre, $R_1 = (I, J, K)$ un repère d'origine O en rotation à la vitesse Ω autour de l'axe K, et I l'intersection avec le méridien de Greenwich.

Soit R le rayon de la Terre et G le centre de masse de la navette. On note $R_1' = (e_r, e_l, e_L)$ le repère associé aux coordonnées sphériques de $G = (r, l, L)$, $r \geqslant R$ étant la distance OG, l la longitude et L la latitude (voir figure 7.5, (i)).

Le système de coordonnées sphériques présente une singularité au pôle Nord et au pôle Sud. Pour écrire la dynamique sous forme plus simple on introduit le repère mobile $R_2 = (i, j, k)$ dont l'origine est G de la manière suivante. Soit $\zeta : t \mapsto (x(t), y(t), z(t))$ la trajectoire de G mesurée dans le repère R_1 et $\vec{v}$ la vitesse relative $\vec{v} = \dot{x}I + \dot{y}J + \dot{z}K$. Pour définir i on pose $\vec{v} = |v|i$. Le vecteur j est un vecteur unitaire du plan (i, e_r) perpendiculaire à i et orienté par $j.e_r > 0$. On pose $k = i \wedge j$. La direction de la vitesse est paramétrisée dans le repère $R_1' = (e_r, e_l, e_L)$ par deux angles (voir figure 7.5, (ii)) :

- la pente γ, aussi appelée *angle de vol*, qui représente l'angle entre un plan horizontal et un plan contenant la navette,

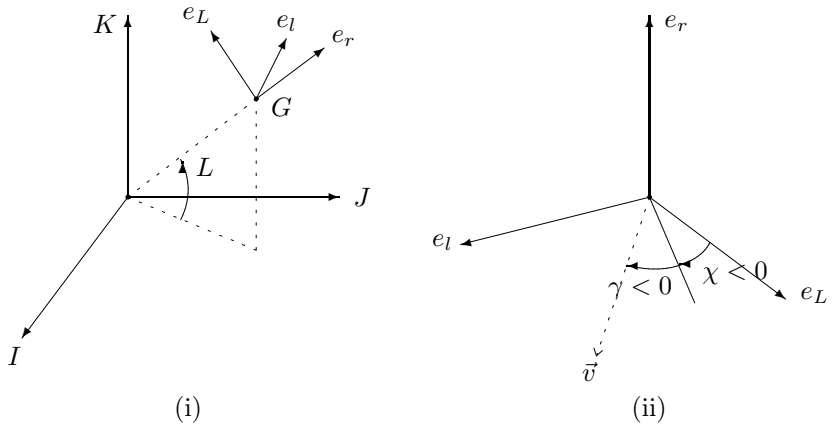


FIGURE 7.5 –

– l'azimut χ, qui est l'angle entre la projection de $\vec{v}$ dans un plan horizontal et le vecteur e_L (voir figure 7.5).

L'équation fondamentale de la mécanique, qui est une équation différentielle du second ordre sur $\mathbb{R}^3$, se traduit par un système dans les coordonnées $(r, l, L, v, \gamma, \chi)$.

Par ailleurs on fait les hypothèses suivantes, le long de l'arc atmosphérique.

Hypothèse 1 : la navette est un planeur, c'est-à-dire que *la poussée de la navette est nulle.*

Hypothèse 2 : on suppose que la vitesse de l'atmosphère est la vitesse de la Terre. La vitesse relative de la navette par rapport à la Terre est donc la vitesse relative $\vec{v}$.

Les forces

Les forces agissant sur la navette sont de deux types :

– **force de gravité :** pour simplifier on suppose que la Terre est sphérique et que la force de gravité est orientée selon e_r. Dans le repère R_2 elle s'écrit

$$\vec{P} = -mg(i \sin\gamma + j \cos\gamma),$$

où $g = g_0/r^2$.

– **force aérodynamique :** la force fluide due à l'atmosphère est une force $\vec{F}$ qui se décompose en
  – une composante dite de *traînée* opposée à la vitesse de la forme

$$\vec{D} = (\frac{1}{2}\rho S C_D v^2)i, \tag{7.32}$$

  – une force dite de *portance* perpendiculaire à $\vec{v}$ donnée par

$$\vec{L} = \frac{1}{2}\rho S C_L v^2 (j \cos\mu + k \sin\mu), \tag{7.33}$$

où μ est l'angle de gîte, $\rho = \rho(r)$ est la densité de l'atmosphère, et C_D, C_L sont respectivement les coefficients de traînée et de portance.

Hypothèse 3 : les coefficients C_D et C_L dépendent de l'angle d'attaque α qui est l'angle entre l'axe du planeur et le vecteur vitesse. C'est a priori un contrôle mais on suppose que durant l'arc atmosphérique il est fixé.

Notre seul contrôle est donc l'angle de gîte μ dont l'effet est double : modifier l'altitude mais aussi tourner à droite ou à gauche.

On choisit pour la densité atmosphérique un modèle exponentiel du type

$$\rho(r) = \rho_0 e^{-\beta r}, \tag{7.34}$$

et par ailleurs on suppose que

$$g(r) = \frac{g_0}{r^2}. \tag{7.35}$$

Le repère n'étant pas absolu, la navette est également soumise à la force de Coriolis $2m\overrightarrow{\Omega} \wedge \dot{q}$ et à la force d'entraînement $m\overrightarrow{\Omega} \wedge (\overrightarrow{\Omega} \wedge q)$.

Finalement, l'arc atmosphérique est décrit par le système

$$\begin{aligned}
\frac{dr}{dt} &= v \sin\gamma \\
\frac{dv}{dt} &= -g \sin\gamma - \frac{1}{2}\rho \frac{SC_D}{m} v^2 + \Omega^2 r \cos L(\sin\gamma \cos L - \cos\gamma \sin L \cos\chi) \\
\frac{d\gamma}{dt} &= \cos\gamma(-\frac{g}{v} + \frac{v}{r}) + \frac{1}{2}\rho \frac{SC_L}{m} v \cos\mu + 2\Omega \cos L \sin\chi \\
&\quad + \Omega^2 \frac{r}{v} \cos L(\cos\gamma \cos L + \sin\gamma \sin L \cos\chi) \\
\frac{dL}{dt} &= \frac{v}{r} \cos\gamma \cos\chi \\
\frac{dl}{dt} &= -\frac{v}{r} \frac{\cos\gamma \sin\chi}{\cos L} \\
\frac{d\chi}{dt} &= \frac{1}{2}\rho \frac{SC_L}{m} \frac{v}{\cos\gamma} \sin\mu + \frac{v}{r} \cos\gamma \tan L \sin\chi + 2\Omega(\sin L - \tan\gamma \cos L \cos\chi) \\
&\quad + \Omega^2 \frac{r}{v} \frac{\sin L \cos L \sin\chi}{\cos\gamma}
\end{aligned} \tag{7.36}$$

où l'état est $q = (r, v, \gamma, l, L, \chi)$ et le contrôle est l'angle de gîte μ.

Dans la suite on pose $r = r_T + h$, où r_T est le rayon de la Terre, et h est l'altitude de la navette.

Le problème de contrôle optimal

Le problème est d'amener l'engin spatial d'une variété initiale M_0 à une variété finale M_1, où le temps terminal t_f est libre, et les conditions aux limites sont données dans la table 7.1.

	Conditions initiales	Conditions finales
altitude (h)	119.82 km	15 km
vitesse (v)	7404.95 m/s	445 m/s
angle de vol (γ)	-1.84 deg	libre
latitude (L)	0	10.99 deg
longitude (l)	libre ou fixée à 116.59 deg	166.48 deg
azimut (χ)	libre	libre

TABLE 7.1 – Conditions aux limites

La navette est, au cours de la phase de rentrée atmosphérique, soumise à trois contraintes :

– Contrainte sur le *flux thermique*

$$\varphi = C_q \sqrt{\rho} v^3 \leqslant \varphi^{max}, \tag{7.37}$$

– Contrainte sur *l'accélération normale*

$$\gamma_n = \gamma_{n0}(\alpha) \rho v^2 \leqslant \gamma_n^{max}, \tag{7.38}$$

– Contrainte sur la *pression dynamique*

$$\frac{1}{2} \rho v^2 \leqslant P^{max}. \tag{7.39}$$

Elles sont représentées sur la figure 7.6 dans le domaine de vol, en fonction de l'accélération $d = \frac{1}{2} \frac{SC_D}{m} \rho v^2$ et de v.

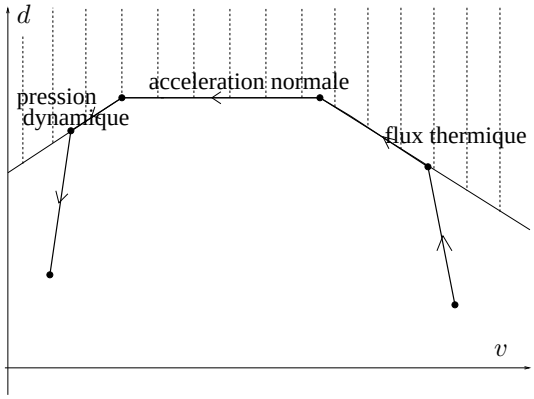


FIGURE 7.6 – Contraintes sur l'état, et stratégie de Harpold-Graves.

Le problème de contrôle optimal est de minimiser le flux thermique total

$$C(\mu) = \int_0^{t_f} C_q \sqrt{\rho} v^3 dt. \tag{7.40}$$

Remarque 7.4.1. Concernant ce critère d'optimisation, plusieurs choix sont en fait possibles et les critères à prendre en compte sont le facteur d'usure lié à l'intégrale du flux thermique et le confort de vol lié à l'intégrale de l'accélération normale. On choisit le premier critère, le temps final t_f étant libre.

Stratégie d'Harpold et Graves [35]

Si on fait l'approximation $\dot{v} \simeq -d$, le coût peut être écrit

$$C(\mu) = K \int_{v_0}^{v_f} \frac{v^2}{\sqrt{d}} dv,\ K > 0,$$

et la stratégie optimale consiste alors à maximiser l'accélération d pendant toute la durée du vol. C'est la politique décrite dans [35], qui réduit le problème à trouver une trajectoire suivant le bord du domaine d'états autorisés, dans l'ordre suivant : flux thermique maximal, puis accélération normale maximale, puis pression dynamique maximale (voir figure 7.6).

L'avantage de cette méthode est que le long d'un arc frontière le contrôle s'exprime sous forme de *boucle fermée*, c'est-à-dire en fonction de l'état. Cette forme est bien adaptée aux problèmes en temps réel, et se prête bien aux problèmes de stabilisation.

Cependant cette méthode *n'est pas optimale* pour notre critère, et notre but est tout d'abord de chercher une trajectoire optimale, puis de la stabiliser.

Données numériques

- Données générales :
  Rayon de la Terre : $r_T = 6378139$ m.
  Vitesse de rotation de la Terre : $\Omega = 7.292115853608596.10^{-5}$ rad.s^{-1}.
  Modèle de gravité : $g(r) = \frac{g_0}{r^2}$ avec $g_0 = 3.9800047.10^{14}$ m^3.s^{-2}.
- Modèle de densité atmosphérique :

$$\rho(r) = \rho_0 \exp\big(-\frac{1}{h_s}(r - r_T)\big)$$

  avec $\rho_0 = 1.225$ kg.m^{-3} et $h_s = 7143$ m.
- Modèle de vitesse du son : $v_{\text{son}}(r) = \sum_{i=0}^{5} a_i r^i$, avec

$$a_5 = -1.880235969632294.10^{-22},\quad a_4 = 6.074073670669046.10^{-15},$$
$$a_3 = -7.848681398343154.10^{-8},\quad a_2 = 5.070751841994340.10^{-1},$$
$$a_1 = -1.637974278710277.10^{6},\quad a_0 = 2.116366606415128.10^{12}.$$

- Nombre de Mach : $Mach(v, r) = v/v_{\text{son}}(r)$.
- Données sur la navette :
  Masse : $m = 7169.602$ kg.

Surface de référence : $S = 15.05$ m^2.
Coefficient de traînée : $k = \frac{1}{2}\frac{SC_D}{m}$.
Coefficient de portance : $k' = \frac{1}{2}\frac{SC_L}{m}$.

– Coefficients aérodynamiques :

Table de $C_D(Mach, incidence)$

	0.00	10.00	15.00	20.00	25.00	30.00	35.00	40.00	45.00	50.00	55.00 deg
0.00	0.231	0.231	0.269	0.326	0.404	0.500	0.613	0.738	0.868	0.994	1.245
2.00	0.231	0.231	0.269	0.326	0.404	0.500	0.613	0.738	0.868	0.994	1.245
2.30	0.199	0.199	0.236	0.292	0.366	0.458	0.566	0.688	0.818	0.948	1.220
2.96	0.159	0.159	0.195	0.248	0.318	0.405	0.509	0.628	0.757	0.892	1.019
3.95	0.133	0.133	0.169	0.220	0.288	0.373	0.475	0.592	0.721	0.857	0.990
4.62	0.125	0.125	0.160	0.211	0.279	0.363	0.465	0.581	0.710	0.846	0.981
10.00	0.105	0.105	0.148	0.200	0.269	0.355	0.458	0.576	0.704	0.838	0.968
20.00	0.101	0.101	0.144	0.205	0.275	0.363	0.467	0.586	0.714	0.846	0.970
30.00	0.101	0.101	0.144	0.208	0.278	0.367	0.472	0.591	0.719	0.849	0.972
50.00	0.101	0.101	0.144	0.208	0.278	0.367	0.472	0.591	0.719	0.849	0.972
Mach											

Table de $C_L(Mach, incidence)$

0.00	0.00	10.00	15.00	20.00	25.00	30.00	35.00	40.00	45.00	50.00	55.00 deg
0.00	0.000	0.185	0.291	0.394	0.491	0.578	0.649	0.700	0.729	0.734	0.756
2.00	0.000	0.185	0.291	0.394	0.491	0.578	0.649	0.700	0.729	0.734	0.756
2.30	0.000	0.172	0.269	0.363	0.454	0.535	0.604	0.657	0.689	0.698	0.723
2.96	0.000	0.154	0.238	0.322	0.404	0.481	0.549	0.603	0.639	0.655	0.649
3.95	0.000	0.139	0.215	0.292	0.370	0.445	0.513	0.569	0.609	0.628	0.626
4.62	0.000	0.133	0.206	0.281	0.358	0.433	0.502	0.559	0.600	0.620	0.618
10.00	0.000	0.103	0.184	0.259	0.337	0.414	0.487	0.547	0.591	0.612	0.609
20.00	0.000	0.091	0.172	0.257	0.336	0.416	0.490	0.552	0.596	0.616	0.612
30.00	0.000	0.087	0.169	0.258	0.338	0.418	0.493	0.555	0.598	0.619	0.613
50.00	0.000	0.087	0.169	0.258	0.338	0.418	0.493	0.555	0.598	0.619	0.613
Mach											

– Profil d'incidence imposé : si le nombre de Mach est plus grand que 10 alors l'incidence est égale à 40. Si le nombre de Mach est compris entre 2 et 10 alors l'incidence est une fonction linéaire du nombre de Mach, entre les valeurs 12 et 40. Si le nombre de Mach est plus petit que 2 alors l'incidence est égale à 12 (voir figure 7.7).

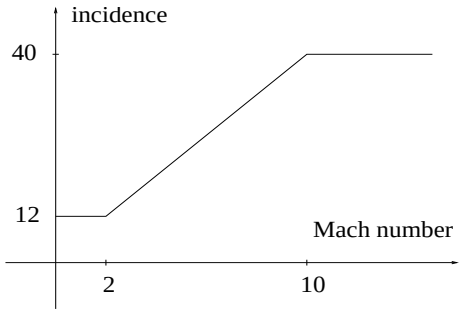


FIGURE 7.7 – Profil d'incidence imposé en fonction du nombre de Mach.

– Contraintes sur l'état :
Contrainte sur le flux thermique : $\varphi = C_q\sqrt{\rho}v^3 \leqslant \varphi^{\max}$, où

$$C_q = 1.705.10^{-4} \text{ S.I.} \quad \text{et} \quad \varphi^{\max} = 717300 \text{ W.m}^{-2}.$$

Contrainte sur l'accélération normale :

$$\gamma_n = \frac{S}{2m}\,\rho v^2 C_D \sqrt{1+\left(\frac{C_L}{C_D}\right)^2} \;\leqslant\; \gamma_n^{\max} = 29.34 \text{ m.s}^{-2}.$$

Contrainte sur la pression dynamique : $P = \dfrac{1}{2}\rho v^2 \;\leqslant\; P^{\max} = 25000\,\text{kPa}$.

– Conditions initiale et terminale : voir table 7.1.

Modèle simplifié en dimension 3

Ici, on se limite à un modèle simplifié en (r, v, γ), où le contrôle est $u = \cos\mu$, et où on suppose la force de Coriolis constante, ce qui conduit au modèle

$$\begin{aligned} \frac{dr}{dt} &= v\sin\gamma \\ \frac{dv}{dt} &= -g\sin\gamma - k\rho v^2 \\ \frac{d\gamma}{dt} &= \cos\gamma(-\frac{g}{v}+\frac{v}{r}) + k'\rho v u + 2\Omega \end{aligned} \tag{7.41}$$

où le contrôle u vérifie la contrainte $|u| \leqslant 1$.

Par ailleurs on prendra comme coefficients C_D et C_L les modèles simplifiés suivants, en fonction de la vitesse v :

$$C_D(v) = \begin{cases} 0.585 & \text{si} \quad v > 3000, \\ 0.075 + 1.7.10^{-4} v & \text{si} \quad 1000 < v \leqslant 3000, \\ 0.245 & \text{si} \quad v \leqslant 1000, \end{cases}$$

$$C_L(v) = \begin{cases} 0.55 & \text{si} \quad v > 3000, \\ 0.1732 + 1.256.10^{-4} v & \text{si} \quad v \leqslant 3000, \end{cases}$$

(voir figure 7.8).

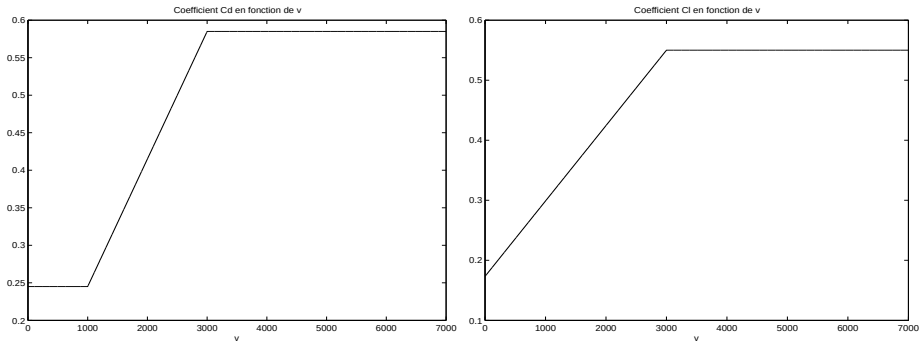


FIGURE 7.8 – Modèle simplifié des coefficients aérodynamiques.

Enfin, pour simplifier l'étude, on ne prend en compte que la contrainte sur le flux thermique

$$\varphi = C_q\sqrt{\rho}v^3 \leqslant \varphi^{max}.$$

7.4.2 Contrôle optimal de la navette spatiale

Dans cette section on résout numériquement le problème de contrôle optimal pour le système simplifié en dimension 3, d'abord en ne tenant pas compte de la contrainte sur le flux thermique, puis en la prenant en compte.

Le problème sans contrainte

Le système simplifié (7.41) en dimension 3 peut s'écrire comme un *système de contrôle affine mono-entrée*

$$\dot{x}(t) = X(x(t)) + u(t)Y(x(t)),\ |u| \leqslant 1, \tag{7.42}$$

où $x = (r, v, \gamma)$, et

$$X = v\sin\gamma\frac{\partial}{\partial r} - (g\sin\gamma + k\rho v^2)\frac{\partial}{\partial v} + \cos\gamma(-\frac{g}{v} + \frac{v}{r})\frac{\partial}{\partial\gamma},$$
$$Y = k'\rho v\frac{\partial}{\partial\gamma},$$

où $k = \frac{1}{2}\frac{SC_D}{m}, k' = \frac{1}{2}\frac{SC_L}{m}$. Le coût est toujours le flux thermique total

$$C(u) = \int_0^{t_f} \varphi dt,$$

avec $\varphi = C_q\sqrt{\rho(r)}v^3$.

Proposition 7.4.1. *Toute trajectoire optimale est bang-bang, i.e. est une succession d'arcs associés au contrôle $u = \pm 1$.*

Démonstration. Dans notre cas le Hamiltonien s'écrit

$$H(x, p, p^0, u) = \langle p, X(x) + uY(x)\rangle + p^0\varphi(x),$$

et la condition de maximisation implique que $u = \text{signe}(\langle p, Y\rangle)$ si $\langle p, Y\rangle \neq 0$. Il suffit donc de montrer que la fonction $t \mapsto \langle p(t), Y(x(t))\rangle$, appelée *fonction de commutation*, ne s'annule sur aucun sous-intervalle, le long d'une extrémale. Supposons le contraire, *i.e.*

$$\langle p(t), Y(x(t))\rangle = 0,$$

sur un intervalle I. En dérivant deux fois par rapport à t il vient

$$\langle p(t), [X, Y](x(t))\rangle = 0,$$
$$\langle p(t), [X, [X, Y]](x(t))\rangle + u(t)\langle p(t), [Y, [X, Y]](x(t))\rangle = 0,$$

où $[.,.]$ est le crochet de Lie de champs de vecteurs. Par conséquent sur l'intervalle I le vecteur $p(t)$ est orthogonal aux vecteurs $Y(x(t))$, $[X, Y](x(t))$, et $[X, [X, Y]](x(t)) + u(t)[Y, [X, Y]](x(t))$. Or on a le résultat suivant.

Lemme 7.4.2.

$$det(Y(x), [X,Y](x), [X,[X,Y]](x) + u[Y,[X,Y]](x)) \neq 0.$$

Preuve du lemme. A l'aide d'un logiciel de calcul formel comme *Maple*, on montre que $[Y,[X,Y]] \in \mathrm{Vect}(Y,[X,Y])$. Par ailleurs la quantité $\det(Y,[X,Y],[X,[X,Y]])$, calculée également avec *Maple*, n'est jamais nulle dans le domaine de vol. □

Il s'ensuit que $p(t) = 0$ sur I. Par ailleurs le Hamiltonien est identiquement nul le long de l'extrémale, et par conséquent $p^0\varphi(x(t)) = 0$ sur I. Comme $\varphi \neq 0$, on en déduit $p^0 = 0$. Donc le couple $(p(.), p^0)$ est nul sur I, ce qui est exclu par le principe du maximum. □

Le contrôle optimal $u(t)$ est donc une succession d'arcs $u = \pm 1$. Nous admettons le résultat suivant, qui découle d'une étude géométrique détaillée dans [15, 17].

Proposition 7.4.3. *La trajectoire optimale vérifiant les conditions initiale et finale (voir table 7.1) est constituée des deux arcs consécutifs $u = -1$ puis $u = +1$.*

Remarque 7.4.2. Cette stratégie consiste à faire tout d'abord "piquer" le plus possible la navette, puis à "redresser" au maximum.

Simulations numériques La trajectoire optimale est donc de la forme $\gamma_-\gamma_+$, où γ_- (resp. γ_+) représente un arc solution du système (7.41) associé au contrôle $u = -1$ (resp. $u = +1$). Il s'agit donc de déterminer numériquement le temps de commutation t_c, *i.e.* le temps auquel le contrôle $u(t)$ passe de la valeur -1 à la valeur $+1$.

Pour cela, on utilise le logiciel *Matlab*, et on détermine t_c par *dichotomie*, de la manière suivante. Etant donné un temps de commutation t_c, on intègre le système en (r, v, γ), jusqu'à ce que la vitesse v atteigne la valeur requise, soit 445 m/s (pour cela on utilise l'option *"events"* de *Matlab*, qui permet de stopper l'intégration numérique lorsqu'une fonction calculée le long de la solution s'annule). On effectue alors une dichotomie sur t_c de manière à ajuster l'altitude finale $r(t_f) = r_T + h(t_f)$ à la valeur souhaitée, soit 15 km.

Remarque 7.4.3. Il s'agit d'un cas particulier de méthode de tir, qui se ramène ici à une dichotomie, car le problème, rappelons-le, a été simplifié. Dans le cas général traité dans [14, 15], la mise en oeuvre d'une méthode de tir (multiple) est nécessaire.

Le programme permettant cette dichotomie, puis le tracé de la solution, sont donnés ci-dessous.

```
function [t,x]=simudim3
%% Fonction permettant le calcul du temps de commutation tc
%% et le trac\'e de la solution, pour le cas sans contrainte
%% sur l'\'etat.
```

```
clc ;
global g0 hs rt Cq Omega;
Omega=7.292115853608596e-005 ; g0=39800047e7 ; hs=7143 ;
rt=6378139 ; Cq = 1.705e-4 ;

range = [0 ; inf ];

% D\'ebut de la trajectoire (altitude 120 km) :
r0 = 0.64979590000E+07 ; v0 = 0.74049501953E+04 ;
gam0 = -0.32114058733E-01 ; flux0=0;

%% Dichotomie pour trouver le temps de commutation de sorte
%% que vf=445 ("events") et hf=15000.
global tc ; tc = -5 ; hf=0 ;
while hf<15000
    global tc ; tc = tc+5
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events);
    [t,x] = ode113(@systdim3,range,xinit,options);
    hf=x(length(t),1)-rt ;
end
a=tc-10 ; b=tc ; hfm=hf ;
while abs(hfm-15000)>1
    global tc ; tc=a;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    hfa=x(length(t),1)-rt;

    global tc ; tc=b;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    hfb=x(length(t),1)-rt;

    global tc ; tc=(a+b)/2 ;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    hfm=x(length(t),1)-rt ;

    if (hfa-15000)*(hfm-15000)<=0
        b=(a+b)/2
    else a=(a+b)/2
    end
```

```
end

%%%%%%%%%%% Resultats %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% tc pour le probleme sans contrainte : tc=242                          %
%       (\ie passage de -1 a +1)                                        %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

global tc ; tc = 242 ;
xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
options = odeset('events',@events,'RelTol',1e-6);
[t,x] = ode113(@systdim3,range,xinit,options);

disp(['altitude finale : ' num2str(x(length(t),1)-rt) ' m'])
disp(['vitesse finale : ' num2str(x(length(t),2)) ' m/s'])
disp(['gamma final : ' num2str(x(length(t),3)/pi*180) ' deg'])
disp(['flux total : ' num2str(x(length(t),4)) ' UI'])

for i=1:length(t)
    gee=g(x(i,1)) ; densite(i)=rho(x(i,1)) ;
    ck(i)=coef_k(x(i,2));
    cd(i)=CDsimple(x(i,2)) ; cl(i)=CLsimple(x(i,2)) ;
end

flux_thermique = Cq.*sqrt(densite(:)).*(x(:,2)).^3 ;
plot(t,flux_thermique)
hold on ; plot(t,717300,'red')
title('Flux thermique')

figure
subplot(311) ; plot(t,x(:,1)-rt) ; title('Altitude') ;
subplot(312) ; plot(t,x(:,2)) ; title('Vitesse') ;
subplot(313) ; plot(t,x(:,3)) ; title('Angle de vol') ;

%--------------------------------------------------------------------

function [value,isterminal,direction]=events(t,x)
global g0 hs Omega rt Cq;
%% Arret a vitesse 445 ou altitude 10000 (en cas d'accident...) :
value = (x(2)-445) * (x(1)-rt-10000) ;
isterminal=1;
direction=0;

%--------------------------------------------------------------------

function dXdt = systdim3(t,X,events)
% Syst\'eme simplifi\'e de la navette en r,v,gamma (dim 3 + flux)
```

```
global Omega g0 hs rt Cq ;
r=X(1) ; v=X(2) ; gam=X(3) ;
dXdt=[v*sin(gam)
      -g(r)*sin(gam)-coef_k(v)*rho(r)*(v)^2
      cos(gam)*(-g(r)/v+v/r)+2*Omega+coef_kp(v)*rho(r)*v*u(t,r,v,gam)
      Cq*sqrt(rho(r))*v^3] ;

%-----------------------------------------------------------------------

function controle=u(t,r,v,gam)
% Contr\^ole pour le probl\'eme sans contrainte : -1 puis +1.

global tc ;
if t<tc
    controle = -1 ;
else controle = 1 ;

%-----------------------------------------------------------------------

function locdensite = rho(r)

global hs rt ;
locdensite = 1.225*\mathrm{exp}(-1/hs.*(r-rt)) ;

%-----------------------------------------------------------------------

function ge=g(r)

global g0 ;
ge = g0./r.^2 ;

%-----------------------------------------------------------------------

function k = coef_k (v)

k = 0.5*15.05* CDsimple(v) /7169.602 ;

%-----------------------------------------------------------------------

function kp = coef_kp (v)

kp = 0.5*15.05* CLsimple(v) /7169.602 ;

%-----------------------------------------------------------------------
```

```
function cd=CDsimple(v)

if v > 3000
    cd=0.585 ;
elseif v>1000
    cd = 0.075+1.7e-4*v ;
else cd= 0.245 ;
end

%-------------------------------------------------------------------

function cl=CLsimple(v)

if v > 3000
    cl=0.55 ;
else cl = 0.1732+1.256e-4*v ;
end
```

Les résultats obtenus sont tracés sur les figures 7.9 et 7.10. On se rend compte que cette stratégie ne permet pas de respecter la contrainte sur le flux thermique, et n'est donc pas adaptée au problème. La prise en compte de cette contrainte sur l'état est donc indispensable

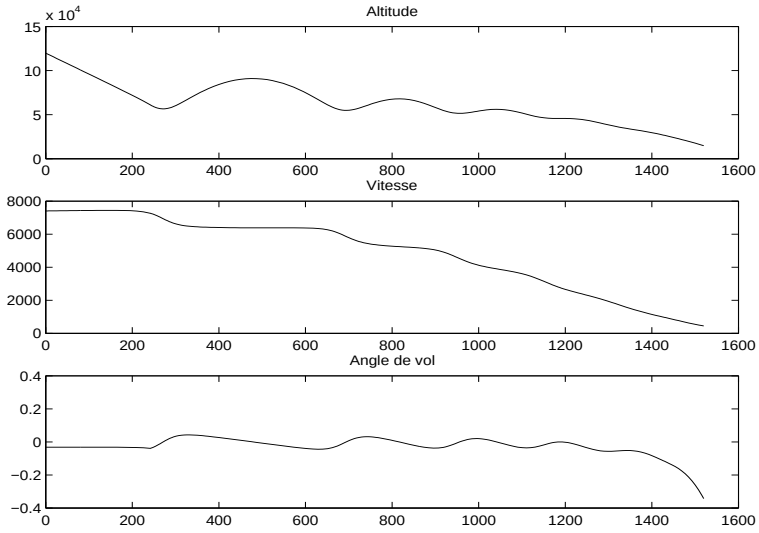


FIGURE 7.9 – Coordonnées d'état pour le problème sans contrainte.

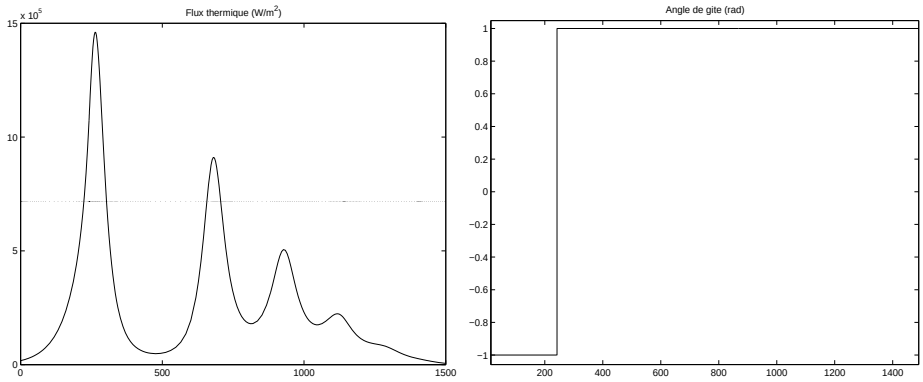


FIGURE 7.10 – Flux thermique, et angle de gîte (contrôle).

Le problème avec contrainte sur l'état

On tient maintenant compte de la contrainte sur le flux thermique. On admet le résultat suivant (voir [14, 15]).

Proposition 7.4.4. *La trajectoire optimale vérifiant les conditions initiale et finale requises est constituée des quatre arcs consécutifs : $u = -1$, $u = +1$, un arc frontière correspondant à un flux thermique maximal, puis $u = +1$.*

Comme pour le problème sans contrainte, on a trois temps de commutation à calculer numériquement :
- le temps de commutation t_1 de -1 à $+1$,
- le temps de commutation t_2 de $+1$ à u_s, où u_s est l'expression du contrôle permettant un flux thermique maximal,
- le temps de commutation t_3 de u_s à $+1$.

Calcul du contrôle iso-flux u_s Le long d'un arc frontière restant à flux thermique maximal, on doit avoir $\varphi = \varphi^{max}$. Par dérivation, on obtient

$$\begin{aligned}\dot{\varphi} &= \varphi(-\frac{1}{2}\frac{v}{h_s}\sin\gamma - \frac{3g_0}{r^2 v}\sin\gamma - 3k\rho v),\\ \ddot{\varphi} &= A + Bu,\end{aligned}$$

où les coefficients A et B sont calculés à l'aide de *Maple*. Le long de l'arc frontière iso-flux, on doit avoir

$$\varphi(t) = \varphi^{max},\ \dot{\varphi}(t) = \ddot{\varphi}(t) = 0,$$

d'où l'on déduit

$$u_s(t) = -\frac{A(t)}{B(t)}.$$

L'expression obtenue pour $u_s(t)$ est

$$u_s = \Big(\; g_0 r^2 v^2 + 7k\rho v^4 r^4 sin\gamma - r^3 v^4 \cos^2\gamma - 2\Omega r^4 v^3 \cos\gamma - 18 g_0 h_s r v^2 \cos^2\gamma$$
$$- 6 g_0^2 hs + 12 g_0^2 h_s \cos^2\gamma + 12 g_0 hsrv^2 - 12\Omega g_0 h_s r^2 v \cos\gamma + 6k^2 h_s \rho^2 r^4 v^4 \Big)$$
$$/(k' r^2 v^2 \rho (r^2 v^2 + 6 g_0 h_s)\cos\gamma).$$

Remarque 7.4.4. Les simulations à venir nous permettront de vérifier *a posteriori* que ce contrôle u_s est bien admissible, *i.e.* vérifie la contrainte $|u_s| \leqslant 1$, pendant la phase iso-flux.

Simulations numériques Le temps de commutation t_1 est calculé de la manière suivante. On intègre le système (7.41) jusqu'à ce que $\dot{\varphi} = 0$ (en utilisant l'option *"events"*). On calcule alors t_1 par dichotomie de façon à ajuster φ à sa valeur maximale φ^{max} en ce temps d'arrêt.

La boucle de dichotomie, qu'il faut insérer à la fonction "simudim3.m" du paragraphe précédent, est la suivante.

```
global t1 ; t1 = -5 ; flux=0 ;
while flux<717300
    global t1; t1 = t1+5
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events);
    [t,x] = ode113(@systdim3,range,xinit,options);
    flux=Cq*sqrt(rho(x(end,1)))*x(end,2)^3 ;
end
a=t1-10 ; b=t1 ; fluxm=flux ;
while abs(fluxm-717300)>50
    global t1; t1=a;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    fluxa=Cq*sqrt(rho(x(end,1)))*x(end,2)^3 ;

    global t1; t1=b;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    fluxb=Cq*sqrt(rho(x(end,1)))*x(end,2)^3 ;

    global t1; t1=(a+b)/2 ;
    xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
    options = odeset('events',@events,'RelTol',1e-6);
    [t,x] = ode113(@systdim3,range,xinit,options);
    fluxm=Cq*sqrt(rho(x(end,1)))*x(end,2)^3 ;
```

```
    if (fluxa-717300)*(fluxm-717300)<=0
        b=(a+b)/2
    else a=(a+b)/2
    end
end
```

Par ailleurs, la fonction *events* doit être modifiée ainsi.

```
function [value,isterminal,direction]=events(t,x)
global g0 hs Omega rt Cq;
%% Arret a derivee(flux)=0 :
value = -1/2*x(2)/hs*sin(x(3))-3/x(2)/x(1)^2*g0*sin(x(3))-...
         3*x(2)*coef_k(x(2))*rho(x(1)) ;
isterminal=1;
direction=0;
```

On détermine ainsi numériquement le premier temps de commutation $t_1 = 153.5$.

Le temps de sortie de la phase iso-flux est déterminé de manière complètement analogue. Finalement, on arrive aux résultats représentés sur les figures 7.11 et 7.12.

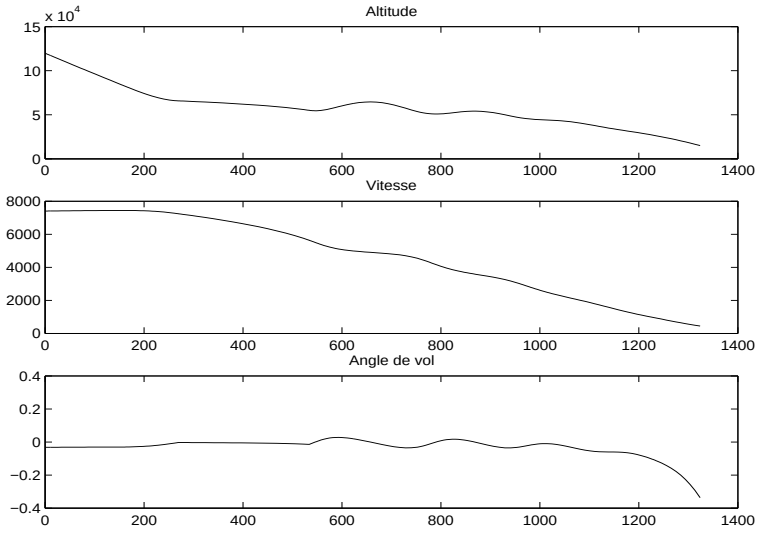


FIGURE 7.11 – Coordonnées d'état pour le problème avec contraintes.

On a donc ainsi déterminé numériquement une trajectoire optimale vérifiant les conditions aux limites souhaitées, et respectant les contraintes sur l'état.

Remarque 7.4.5. Pour le modèle non simplifié en dimension 6, ce n'est pas le cas : les contraintes sur le facteur de charge et sur la pression dynamique ne sont pas respectées, et il faut envisager une phase iso-accélération normale (voir [14, 15]).

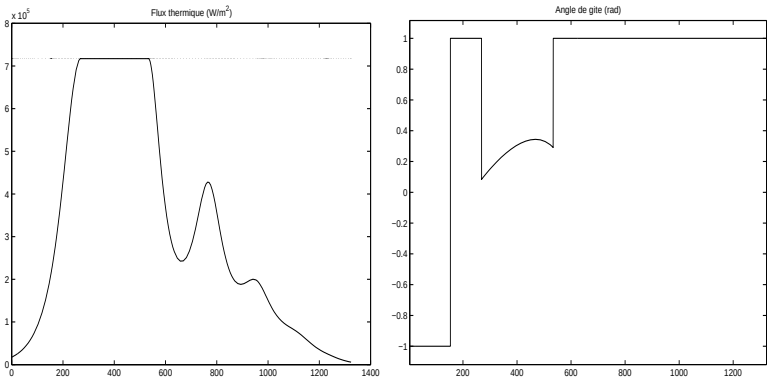


FIGURE 7.12 – Flux thermique et contrôle $u(t)$.

7.4.3 Stabilisation autour de la trajectoire nominale

On se propose maintenant de stabiliser le système simplifié autour de la trajectoire construite dans le paragraphe précédent, de façon à prendre en compte d'éventuelles perturbations, dues aux erreurs de modèles, aux perturbations atmosphériques, etc. Pour cela, on va utiliser la *théorie linéaire-quadratique* traitée précédemment dans cet ouvrage, qui permet d'exprimer le contrôle sous forme de boucle fermée, au voisinage de la trajectoire nomimale, de façon à la rendre stable.

Le système étudié est un système de contrôle non linéaire dans $\mathbb{R}^n$, du type

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

où $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ est C^1, et les contrôles admissibles u sont à valeurs dans $\Omega \subset \mathbb{R}^m$. Soit $(x_e(.), u_e(.))$ une trajectoire solution sur $[0, T]$, telle que pour tout $t \in [0, T]$ on ait $u(t) \in \overset{\circ}{\Omega}$.

Supposons que le système soit légèrement perturbé, ou bien que l'on parte d'une condition initiale proche de $x_e(0)$, et que l'on veuille suivre le plus possible la trajectoire nominale $x_e(\cdot)$. Posons alors $y(\cdot) = x(\cdot) - x_e(\cdot)$ et $v(\cdot) = u(\cdot) - u_e(\cdot)$. Au premier ordre, $y(\cdot)$ est solution du système linéarisé

$$\dot{y}(t) = A(t)y(t) + B(t)v(t),$$

où

$$A(t) = \frac{\partial f}{\partial x}(x_e(t), u_e(t)),\ B(t) = \frac{\partial f}{\partial u}(x_e(t), u_e(t)).$$

Le but est alors de rendre l'erreur $y(\cdot)$ la plus petite possible, ce qui nous amène à considérer, pour ce système linéaire, un coût quadratique du type (4.2), où les matrices de pondération Q, W, U sont à choisir en fonction des données du problème. Il s'agit, au premier ordre, d'un problème de poursuite avec $\xi = x_e$. En particulier on a $h = 0$ pour ce problème.

C'est cette stratégie que l'on adopte pour stabiliser la navette vers sa trajectoire de référence.

Pour tenir compte de la contrainte sur le contrôle, il faut d'abord modifier la trajectoire nominale $x_e(\cdot)$ obtenue précédemment de façon à ce qu'elle respecte la nouvelle contrainte sur le contrôle $|u_e| \leqslant 1 - \varepsilon$, où ε est un petit paramètre. On choisit par exemple $\varepsilon = 0.05$. On trouve alors de nouveaux temps de commutation, qui sont

$$t_1 = 143.59,\ t_2 = 272.05,\ t_3 = 613.37.$$

Dans le programme suivant, on implémente l'équation de Riccati. Celle-ci est intégrée en temps inverse puisqu'on se donne une condition finale. Il faut donc ensuite rétablir le cours normal du temps en symétrisant la matrice de discrétisation obtenue. Enfin, le contrôle bouclé obtenu est réinjecté dans le système initial. Les simulations sont effectuées en prenant des conditions initiales proches, mais différentes, de celles de la table 7.1.

```
function stabdim3
% Stabilisation de la navette, en dimension 3, par Riccati.

clc ;
global g0 hs rt Cq Omega S m;
Omega=7.292115853608596e-005 ; g0=39800047e7 ; hs=7143 ;
rt=6378139 ; Cq = 1.705e-4 ; S=15.05 ; m=7169.602 ;

range = [0 ; inf ] ;

%% Debut de la trajectoire (alt. 120000 km) :
r0 = 0.64979590000E+07 ; v0 = 0.74049501953E+04 ;
gam0 = -0.32114058733E-01 ; flux0 = 0 ;
%% Entree dans la phase iso-flux :
r0=6.443919913623549e+06 ; v0=7.243006878621867e+03 ;
gam0=-0.00319355995196 ;

global t1 t2 t3 ;
%t1=-1 ; t2=t1 ; t3=533.75-268.9198 ;
%t1=143.5908945796609 ; t2=272.0484928785223 ; t3=613.3766178785223 ;
t1=-1 ; t2=t1 ; t3=613.376617878522-272.0484928785223 ;
xinit = [ r0 ; v0 ; gam0 ; flux0 ] ;
options = odeset('events',@events,'RelTol',1e-6) ;
global te xe ;
%% trajectoire nominale
[te,xe] = ode113(@systdim3,range,xinit,options) ;

%% Definition des poids
global W invU ;
```

```
W = eye(3) ; W(1,1)=1e-6 ; W(2,2)=1e-3 ; W(3,3)=10 ;
invU = 1e-10 ;
global tricca ricca ;
minit = - [ 1e-6 ; 0 ; 0 ; 0 ; 0 ; 0 ] ;   %  E(T)=-Q
rangericca = fliplr(te) ;
[tricca,ricca] = ode113(@matriccati,rangericca,minit) ;
ricca = flipud(ricca);

xinit=[ r0 ; v0 ; gam0 ] ;
[t,x] = ode113(@systboucle,te,xinit+[ 1500 ; 40 ; -0.004 ] ) ;
close all
x(end,1)-rt
plot(t,x(:,1)-xe(:,1)) ;
figure ; plot(t,x(:,2)-xe(:,2)) ;
figure ; plot(t,x(:,3)-xe(:,3)) ;
for k=1:length(te)
    contfeed(k)=uboucle(t(k)) ;
    conte(k)=u(t(k),xe(k,1),xe(k,2),xe(k,3));
end
figure ; plot(t,contfeed-conte)

%-----------------------------------------------------------------

function [value,isterminal,direction]=events(t,x)
global g0 hs Omega rt Cq;
%% Arret a vitesse 445 ou altitude 10000 :
value = (x(2)-445) * (x(1)-rt-10000) ;
%% Arret a derivee(flux)=0 :
% value = -1/2*x(2)/hs*sin(x(3))-3/x(2)/x(1)^2*g0*sin(x(3))-...
%         3*x(2)*coef_k(x(2),x(1))*rho(x(1)) ;
isterminal=1;
direction=0;

%-----------------------------------------------------------------

function dXdt = matriccati(t,X)
% Eq de Riccati  dE/dt=W-A'E-EA-EBU^{-1}B'E, E(T)=-Q,
% en temps inverse

global W invU ;
E = [ X(1) X(2) X(3)
      X(2) X(4) X(5)         %% matrice de Riccati (symetrique)
      X(3) X(5) X(6) ];
[A,B] = matlinear(t) ;
mat = -W+A'*E+E*A+E*B*invU*B'*E ;
dXdt = [mat(1,1);mat(1,2);mat(1,3);mat(2,2);mat(2,3);mat(3,3) ] ;
```

```
%-----------------------------------------------------------------

function [matA,matB] = matlinear(t)

global te xe g0 hs S m ;
[val,k]=min(abs(te-t)) ; r=xe(k,1) ; v=xe(k,2) ; gam=xe(k,3) ;

if ((v<=1000)|(v>3000))
    derCD=0;
else
    derCD=1.7e-4;
end
if (v>3000)
    derCL=0;
else
    derCL=1.256e-4;
end

matA=zeros(3,3);
matA(1,2) = sin(gam) ;
matA(1,3) = v*cos(gam) ;
matA(2,1) = 2*g0/r^3*sin(gam)+coef_k(v)*v^2*rho(r)/hs ;
matA(2,2) = -rho(r)/(2*m)*S*derCD*v^2-coef_k(v)*rho(r)*2*v ;
matA(2,3) = -g(r)*cos(gam) ;
matA(3,1) = cos(gam)*(2*g0/(r^3*v)-v/r^2)-...
            coef_kp(v)*rho(r)/hs*v*u(t,r,v,gam) ;
matA(3,2) = cos(gam)*(g(r)/v^2+1/r)+rho(r)*u(t,r,v,gam)*...
            S/(2*m)*(derCL*v+CLsimple(v)) ;
matA(3,3) = -sin(gam)*(-g(r)/v+v/r) ;

matB = [ 0 ; 0 ; coef_kp(v)*rho(r)*v ] ;

%-----------------------------------------------------------------

function dXdt = systboucle(t,X)

% systeme tronque de la navette en r,v,gamma (dim 3 + flux)

global Omega g0 hs rt Cq ;

r=X(1) ; v=X(2) ; gam=X(3) ;

dXdt= [ v*sin(gam)
        -g(r)*sin(gam)-coef_k(v)*rho(r)*(v)^2
        cos(gam)*(-g(r)/v+v/r)+2*Omega+...
```

```
        coef_kp(v)*rho(r)*v*uboucle(t) ] ;

%-------------------------------------------------------------

function contfeedback = uboucle(t)

global te xe invU S m ricca ;
[A,B] = matlinear(t) ;
[val,k] = min(abs(te-t)) ; r=xe(k,1) ; v=xe(k,2) ; gam=xe(k,3) ;
contfeedback = u(te(k),r,v,gam)+invU*coef_kp(v)*rho(r)*v*...
               (ricca(k,3)*r+ricca(k,5)*v+ricca(k,6)*gam) ;
```

Quelques commentaires sur le programme. On effectue la procédure de stabilisation de Riccati à partir de l'entrée dans la phase iso-flux seulement, soit environ à une altitude de 65 km, une vitesse de 7200 m/s, et un angle de vol de -0.003 rad. En effet la phase iso-flux (flux thermique maximal) est la phase la plus dangereuse de la rentrée atmosphérique. Notons d'ailleurs que, récemment, la navette *Columbia* a explosé à une altitude d'environ 62 km, en pleine phase iso-flux (ce drame a eu lieu en mars 2003). C'est la phase où l'engin spatial s'échauffe le plus : les frottements avec l'atmosphère sont très intenses. Cette phase iso-flux est aussi assez longue, environ 350 secondes (la durée totale de la phase de rentrée atmosphérique est d'environ 1300 secondes).

Tout ceci justifie l'intérêt porté à la procédure de stabilisation de la navette, à partir du point d'entrée dans la phase iso-flux. Dans les simulations suivantes ce point d'entrée est donc notre condition initiale.

Notons $x_e(\cdot) = (r_e(\cdot), v_e(\cdot), \gamma_e(\cdot))$ la trajectoire nominale et $u_e(\cdot)$ son contrôle associé. Il vérifie $|u_e| \leqslant 0.95$. Notons par ailleurs $x = (r, v, \gamma)$ la trajectoire du système (7.41), partant d'un point $x(0)$ et associée au contrôle $u = u_e + v$, le contrôle v étant le correctif calculé par la procédure de Riccati. Il doit vérifier la contrainte $|v| \leqslant 0.05$. Aussi, dans le programme ci-dessus, on a forcé v à respecter cette contrainte.

Par ailleurs le choix des *poids* est très important. On obtient des poids adaptés par tâtonnements, et en tenant compte de l'ordre respectif des variables du système. Ici on a pris

$$W = \begin{pmatrix} 10^{-6} & 0 & 0 \\ 0 & 10^{-2} & 0 \\ 0 & 0 & 10 \end{pmatrix}, \quad Q = \begin{pmatrix} 10^{-6} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{et} \quad U = 10^{10}.$$

Bien entendu d'autres choix sont possibles. Ici notre choix de Q force l'altitude finale à être proche de l'altitude souhaitée. En revanche on laisse plus de liberté à la vitesse finale et à l'angle de vol final.

La trajectoire $x(\cdot)$ part d'un point $x(0)$ différent de $x_e(0)$. On a pris les données numériques suivantes :

– écart sur l'altitude initiale : 1500 m,

– écart sur la vitesse initiale : 40 m/s,
– écart sur l'angle de vol initial : -0.004 rad, soit -0.2292 deg.

Les résultats numériques obtenus sont assez satisfaisants : l'altitude finale obtenue est 15359 km, et la vitesse finale est 458 m/s. L'écart par rapport aux données souhaitées (altitude 15 km, vitesse 440 m/s) est donc assez faible.

Notons que l'écart sur l'angle de vol initial que nous avons pris ici est assez important. Cette pente initiale est en effet un paramètre très sensible dans les équations : si à l'entrée de la phase atmosphérique l'angle de vol est trop faible, alors la navette va rebondir sur l'atmosphère (phénomène bien connu, dit de *rebond*), et si au contraire il est trop important il sera impossible de redresser l'engin, qui va s'écraser au sol.

Les figures suivantes sont le résultat des simulations numériques. La figure 7.13 représente l'écart entre l'état nominal et l'état réel, et la figure 7.14 l'écart entre le contrôle nominal et le contrôle réel (*contrôle bouclé*, ou *contrôle feedback*). La figure 7.15 représente l'état, et la figure 7.16 le flux thermique. On constate que la contrainte sur le flux thermique est à peu près respectée. On peut conclure que la procédure de stabilisation ainsi réalisée est satisfaisante.

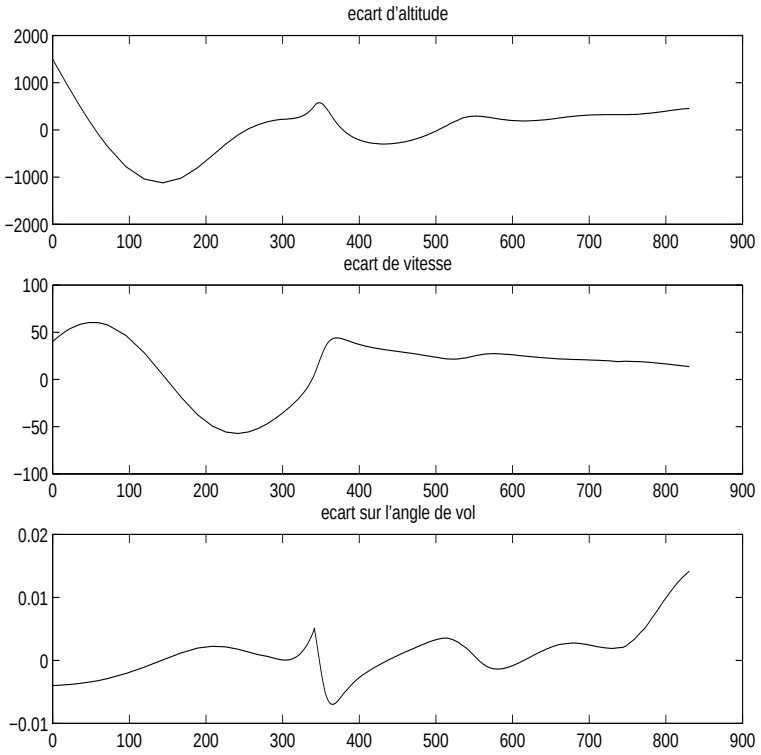


FIGURE 7.13 – Ecart entre l'état nominal et l'état réel.

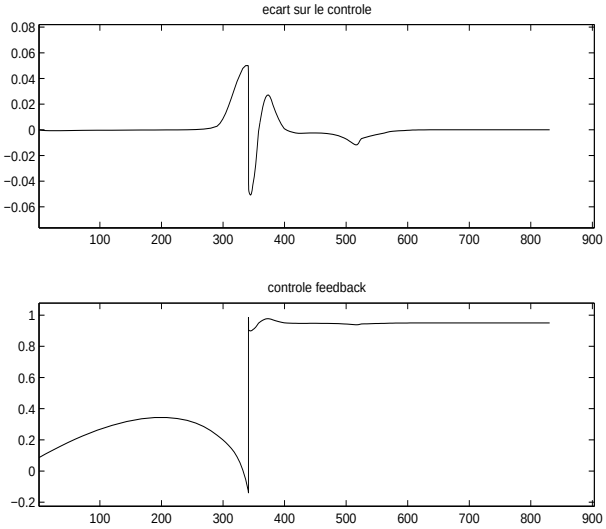


FIGURE 7.14 – Contrôle bouclé, et correction par rapport au contrôle nominal.

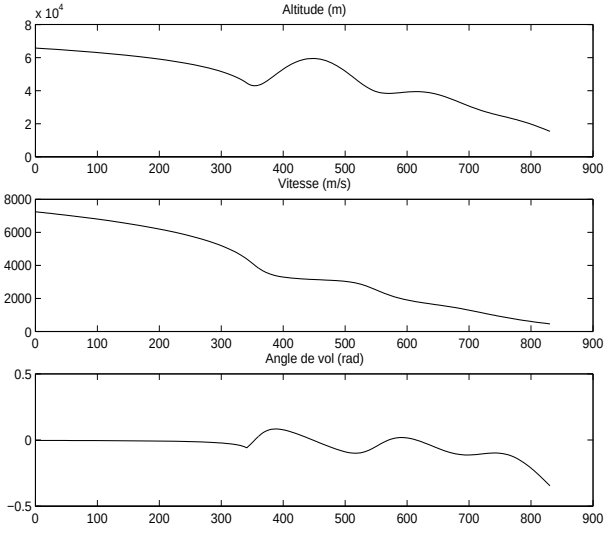


FIGURE 7.15 – Etat avec le contrôle feedback.

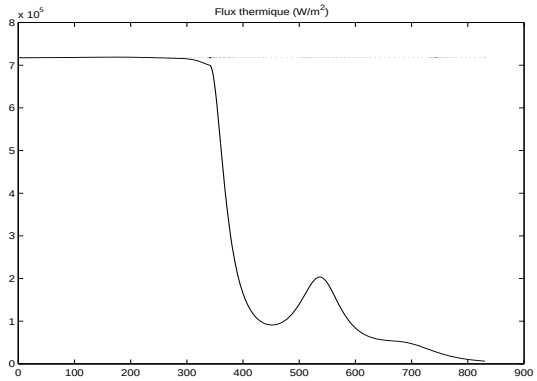


FIGURE 7.16 – Flux thermique avec le contrôle feedback.

Chapitre 8

Théorie d'Hamilton-Jacobi

8.1 Introduction

La théorie d'Hamilton-Jacobi est une branche du calcul des variations et de la mécanique analytique, dans laquelle trouver des extrémales se réduit à résoudre une équation aux dérivées partielles du premier ordre : l'équation d'Hamilton-Jacobi. Les fondements de la théorie ont été posé par Hamilton en 1820, concernant des problèmes d'optique ondulatoire et géométrique. En 1834, il étend ses idées à des problèmes de dynamique. Jacobi en 1837 applique la méthode à des problèmes généraux de calcul variationnel.

Le point de départ remonte cependant au 17è siècle, avec Fermat et Huygens en optique géométrique. Le principe de Fermat stipule que la lumière se propage d'un point à un autre dans un milieu inhomogène en temps minimal. Soit x_0 un point de départ, et $S(x)$ le temps minimal que met la lumière pour aller de x_0 à x. Cette fonction temps minimal est appelée *fonction Eikonal*, ou *longueur optique* du chemin. Soit $v(x)$ le module de la vitesse de la lumière en x. Supposons que la lumière parcourt la distance dx pendant la durée dt. Selon le principe d'Huygens, la lumière voyage le long de la normale à la surface de niveau de S. On obtient donc, au premier ordre,

$$S\left(x + \frac{\nabla S(x)}{\|\nabla S(x)\|} v(x)dt\right) = S(x) + dt,$$

d'où l'équation

$$\|\nabla S(x)\|^2 = \frac{1}{v(x)^2},$$

qui est l'équation d'Hamilton-Jacobi de l'optique géométrique, ou équation eikonale.

En mécanique analytique, on remplace la fonction Eikonal par l'*action*

$$S(t,x) = \int_\gamma L(s, x(s), \dot{x}(s))ds,$$

où γ est un chemin joignant (t_0, x_0) à (t, x), et L est le Lagrangien du système. Le principe de moindre action conduit aux *équations d'Euler-Lagrange*

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = \frac{\partial L}{\partial x}.$$

Si la transformation de Legendre $\mathcal{T}(x, \dot{x}) = (x, p)$, où $p = \frac{\partial L}{\partial \dot{x}}$, est un difféomorphisme, on définit le Hamiltonien du système $H(t, x, p) = p\dot{x} - L(t, x, \dot{x})$. Alors, le long d'une extrémale (*i.e.* une courbe vérifiant les équations d'Euler-Lagrange), on a $S(t, x(t), \dot{x}(t)) = \int_{t_0}^{t} L(s, x(s), \dot{x}(s))ds$, et par dérivation par rapport à t, on obtient $\frac{\partial S}{\partial t} + \frac{\partial S}{\partial x}\dot{x} = L$, d'où

$$\frac{\partial S}{\partial t} + H\left(t, x, \frac{\partial S}{\partial x}\right) = 0,$$

qui est l'équation d'Hamilton-Jacobi.

8.2 Solutions de viscosité

De manière générale, on étudie le problème de Dirichlet pour l'équation d'Hamilton-Jacobi

$$\begin{aligned} H(x, S(x), \nabla S(x)) &= 0 \text{ dans } \Omega, \\ S &= g \text{ sur } \partial\Omega, \end{aligned} \tag{8.1}$$

où Ω est un ouvert de $\mathbb{R}^n$, et H est une fonction sur $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$.

Remarque 8.2.1. Le cas d'une équation d'Hamilton-Jacobi d'évolution

$$\begin{aligned} \frac{\partial S}{\partial t} + H(x, \frac{\partial S}{\partial x}) &= 0 \text{ dans } \mathbb{R} \times \mathbb{R}^n, \\ S(0, x) &= g(x) \text{ sur } \mathbb{R}^n, \end{aligned} \tag{8.2}$$

est un cas particulier de (8.1). En effet il suffit de poser $\tilde{x} = (t, x)$, $\tilde{p} = (p_0, p)$, et $\tilde{H}(\tilde{x}, z, \tilde{p}) = p_0 + H(x, p)$.

Le but de cette section est de donner un cadre mathématique rigoureux à la définition d'une solution du problème (8.1). On va montrer que la notion classique de solution est insuffisante : la méthode des caractéristiques met en évidence l'apparition de singularités. On introduit alors la notion de solution de viscosité.

8.2.1 Méthode des caractéristiques

On introduit des chemins $x(s)$ dans Ω, partant de $\partial\Omega$, appelés *caractéristiques*, le long desquels on résout l'équation et on obtient les valeurs de S.

Posons $z(s) = S(x(s))$ et $p(s) = \nabla S(x(s))$, et cherchons une équation différentielle ordinaire décrivant l'évolution de z et p. On a

$$\dot{z}(s) = \nabla S(x(s)).\dot{x}(s) = p(s).\dot{x}(s), \quad \dot{p}(s) = d^2 S(x(s)).\dot{x}(s).$$

Or, en différentiant l'équation d'Hamilton-Jacobi $H(x, S(x), \nabla S(x)) = 0$ par rapport à x, on obtient

$$\frac{\partial H}{\partial x} + \frac{\partial H}{\partial z}\nabla S + \frac{\partial H}{\partial p}d^2 S = 0.$$

Choisissons alors le chemin $x(s)$ tel que $\dot{x} = \frac{\partial H}{\partial p}$. Il vient alors $\dot{z} = p.\frac{\partial H}{\partial p}$, et $\dot{p} = -\frac{\partial H}{\partial x} - \frac{\partial H}{\partial z}.p$.

Finalement, les équations

$$\begin{aligned}
\dot{x}(s) &= \frac{\partial H}{\partial p}(x(s), z(s), p(s)), \quad x(0) = \bar{x} \in \partial\Omega, \\
\dot{z}(s) &= p(s).\frac{\partial H}{\partial p}(x(s), z(s), p(s)), \quad z(0) = S(\bar{x}) = g(\bar{x}), \\
\dot{p}(s) &= -\frac{\partial H}{\partial x}(x(s), z(s), p(s)) - \frac{\partial H}{\partial z}(x(s), z(s), p(s)).p(s), \quad p(0) = \nabla S(\bar{x}),
\end{aligned}$$

sont appelées *équations caractéristiques*.

Remarque 8.2.2. Dans le cas d'évolution (8.2), on obtient en particulier $\frac{dt}{ds} = 1$, donc $t = s$, et $\dot{p}_0 = 0$. On a également

$$\dot{x} = \frac{\partial H}{\partial p},\ \dot{p} = -\frac{\partial H}{\partial x},$$

qui sont les équations de Hamilton. Enfin, on retrouve aussi

$$\dot{z} = p.\frac{\partial H}{\partial p} + p_0 = p\dot{x} + p_0,$$

avec $p_0 + H = \tilde{H} = 0$, d'où $\dot{z} = p\dot{x} - H = L$, et donc z est l'action. Autrement dit, *les caractéristiques sont les extrémales du problème de minimisation de l'action* (en tout cas si la transformation de Legrendre est un difféomorphisme).

Remarque 8.2.3. Discutons plus en détail la condition initiale $p(0) = \nabla S(\bar{x})$, dans le cas où $\partial\Omega$ est supposé être une sous-variété de $\mathbb{R}^n$. De la condition $S_{|\partial\Omega} = g_{|\partial\Omega}$, on déduit facilement que

$$\Pi_{T_x\partial\Omega}\nabla S(x) = \Pi_{T_x\partial\Omega}\nabla g(x),$$

pour tout $x \in \partial\Omega$, où $T_x\partial\Omega$ désigne l'espace tangent à la sous-variété $\partial\Omega$ au point x, et $\Pi_{T_x\partial\Omega}$ est la projection de $\mathbb{R}^n$ sur $T_x\partial\Omega$. Donc, cela signifie que $p(0)$ doit vérifier la condition, au point $\bar{x}$,

$$\Pi_{T_x\partial\Omega}p(0) = \Pi_{T_x\partial\Omega}\nabla g(\bar{x}),$$

ce qui détermine $n-1$ composantes de $p(0)$. La n-ème composante de $p(0)$ est déterminée en imposant de plus

$$H(p(0), g(\bar{x}), \bar{x}) = 0.$$

D'après le théorème des fonctions implicites, cela permet bien de déterminer la composante manquante de $p(0)$, pourvu que $\frac{\partial H}{\partial n} \neq 0$, où n est la normale à $\partial\Omega$.

Notons que, dans le cas d'évolution, cette dernière condition est toujours vérifiée. La condition initiale sur $p(0)$ est alors simplement

$$p(0) = \nabla g(\bar{x}).$$

Appliquons la méthode des caractéristiques à la construction d'une solution de (8.1) au voisinage de la frontière.

Pour tout $\bar{x} \in \partial\Omega$, notons $(x(\bar{x}, s), z(\bar{x}, s), p(\bar{x}, s))$ la solution des équations caractéristiques. Notons n la normale à $\partial\Omega$. Sous l'hypothèse $\frac{\partial H}{\partial n} \neq 0$, on montre facilement que, localement en $(\bar{x}, s)$, l'application $\varphi(\bar{x}, s) = x(\bar{x}, s)$ est inversible. On en déduit donc, localement, que

$$S(x) = z(\varphi^{-1}(x)).$$

Pour plus de détails sur la méthode des caractéristiques, et des preuves précises de tous ces faits, on se réfère à [25].

Remarque 8.2.4. Dans le cas d'évolution, l'hypothèse $\frac{\partial \tilde{H}}{\partial n} \neq 0$ est toujours vérifiée.

Remarque 8.2.5. Si g et H sont de classe C^2, alors localement la solution S est de classe C^2.

En faisant cette construction au voisinage de tout point de $\partial\Omega$, puis en recollant les voisinages, on obtient une solution S de (8.1) sur un voisinage de $\partial\Omega$ dans Ω. Dans le cas d'évolution (8.2), on obtient une solution pour t petit.

Mais en général, on ne peut pas prolonger S sur Ω tout entier, car des singularités se produisent lorsque des caractéristiques se croisent (voir figure 8.1).

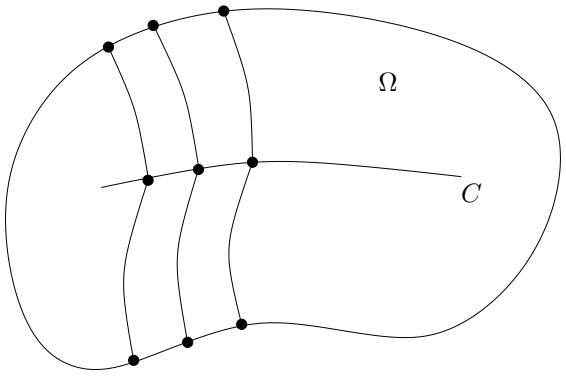


FIGURE 8.1 – Croisement des caractéristiques

Exemple 8.2.1. Un exemple simple de cette situation est donné par le problème de Dirichlet pour l'équation eikonale

$$\begin{aligned} \|S(x)\|^2 &= 1 \text{ dans } \Omega, \\ S &= 0 \text{ sur } \partial\Omega. \end{aligned} \tag{8.3}$$

Les équations caractéristiques sont

$$\dot{x} = 2p,\ \dot{p} = 0,\ \dot{z} = p.\dot{x} = 2.$$

Si $x(0) = \bar{x}$, $z(0) = 0$, et $p(0) = n$ normal à $\partial\Omega$, alors $x(s) = \bar{x} + 2sn$, et $z(s) = \|x(s) - \bar{x}\|$. Finalement, on trouve que la solution de (8.3) est

$$S(x) = d(x, \partial\Omega),$$

comme on pouvait s'y attendre. La fonction S n'est pas différentiable sur la courbe C où des caractéristiques se croisent, appelée *cut-locus*.

Par conséquent, en général il n'existe pas de solution globale de classe C^1 sur Ω. Il faut donc chercher un concept de solution généralisée.

On pense d'abord au théorème de Rademacher, selon lequel toute fonction lipschitzienne est différentiable presque partout. Il est donc tentant de définir une solution généralisée de (8.1) comme étant une fonction S lipschitzienne sur $\overline{\Omega}$, solution de l'équation d'Hamilton-Jacobi presque partout. Malheureusement ce concept est (de loin) trop faible pour avoir unicité et stabilité par passage à la limite dans L^∞, comme on peut le voir sur les deux exemples suivants.

Exemple 8.2.2. Le problème

$$\begin{aligned} \frac{\partial S}{\partial t} + \left(\frac{\partial S}{\partial x}\right)^2 &= 0 \quad \text{p.p. sur } \mathbb{R}\times]0,+\infty[, \\ S(0,\cdot) &= 0, \end{aligned}$$

a au moins deux solutions

1. $S(t,x) = 0$,
2. $S(t,x) = \begin{cases} 0 & \text{si} \quad |x| \geqslant t, \\ -t + |x| & \text{si} \quad |x| < t. \end{cases}$

Exemple 8.2.3. Le problème

$$\left|\frac{\partial S}{\partial x}\right| = 1 \quad \text{p.p. sur }]0,1[, \quad S(0) = S(1) = 0,$$

admet une infinité de solutions généralisées (voir figure 8.2).

En particulier, il existe une suite (S_n) de solutions convergeant uniformément vers 0, et pourtant 0 n'est pas solution.

Ce concept de solution est donc insuffisant.

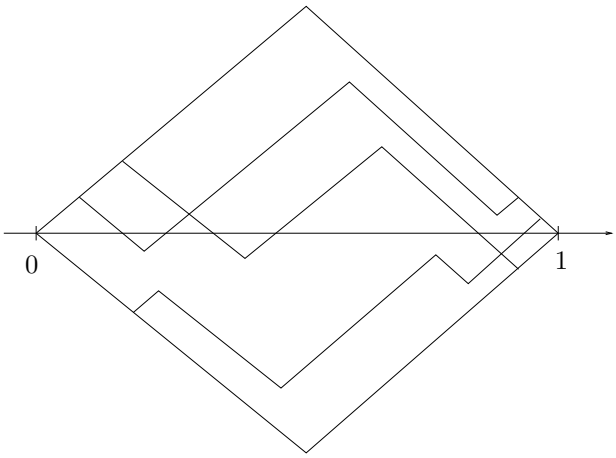


FIGURE 8.2 – Infinité de solutions

8.2.2 Définition d'une solution de viscosité

On cherche un concept de solution ayant les propriétés suivantes :

1. il existe une unique solution S de (8.1), dépendant continûment de g et H ;
2. on a stabilité par passage à la limite évanescente, *i.e.* si $H(x, S_\varepsilon, \nabla S_\varepsilon) = \varepsilon \triangle S_\varepsilon$ pour tout $\varepsilon > 0$ petit, alors $S_\varepsilon \to S$ lorsque ε tend vers 0 ;
3. si (8.1) est l'équation d'Hamilton-Jacobi d'une fonction valeur d'un problème de contrôle optimal, alors la fonction valeur est l'unique solution de (8.1).

L'idée de départ est en fait de régulariser l'équation (8.1) en lui ajoutant le terme $\varepsilon \triangle S$ (méthode de *viscosité évanescente*), car pour une EDP quasi-linéaire du second ordre on sait montrer qu'une solution régulière S_ε existe, et de plus on dispose d'estimations uniformes sur tout compact, ce qui permet les passages à la limite.

Le concept de solution qui convient est celui de *solution de viscosité*, introduit par [23] au début des années 80, et que l'on rappelle ici dans le cadre d'équations d'Hamilton-Jacobi du premier ordre.

Soit Ω un ouvert de $\mathbb{R}^n$, H une fonction continue sur $\Omega \times \mathbb{R} \times \mathbb{R}^n$, appelée *Hamiltonien*, et g une fonction continue sur $\partial\Omega$. Considérons l'équation d'Hamilton-Jacobi du premier ordre sur Ω

$$H(x, S(x), \nabla S(x)) = 0. \tag{8.4}$$

On rappelle tout d'abord la notion de sous- et sur-différentiel.

Définition 8.2.1. Soit S une fonction sur Ω. Le *sur-différentiel* en un point

$x \in \Omega$ est défini par

$$D^+S(x) = \{p \in \mathbb{R}^n \mid \limsup_{y \to x} \frac{S(y) - S(x) - \langle p, y - x\rangle}{\|y - x\|} \leqslant 0\}.$$

De même, le *sous-différentiel* en x est

$$D^-S(x) = \{p \in \mathbb{R}^n \mid \liminf_{y \to x} \frac{S(y) - S(x) - \langle p, y - x\rangle}{\|y - x\|} \geqslant 0\}.$$

Remarque 8.2.6. On a les propriétés suivantes.

- Soit S une fonction continue sur Ω.
  1. $p \in D^+S(x) \Leftrightarrow \exists \varphi \in C^1(\Omega) \mid \varphi \geqslant S,\ \varphi(x) = S(x),\ \nabla\varphi(x) = p.$
  2. $p \in D^-S(x) \Leftrightarrow \exists \varphi \in C^1(\Omega) \mid \varphi \leqslant S,\ \varphi(x) = S(x),\ \nabla\varphi(x) = p.$
- Si S est différentiable en x alors $D^+S(x) = D^-S(x) = \{\nabla S(x)\}$.
- Si $D^+S(x)$ et $D^-S(x)$ sont non vides, alors S est différentiable en x.
- L'ensemble des points de Ω tels que $D^+S(x)$ (resp. $D^-S(x)$) soit non vide est dense dans Ω.

Définition 8.2.2. Soit S une fonction continue sur Ω. La fonction S est dite *sous-solution de viscosité* de l'équation (8.4) si

$$\forall x \in \Omega \quad \forall p \in D^+v(x) \quad H(x, v(x), p) \leqslant 0.$$

De même, S est une *sur-solution de viscosité* de (8.4) si

$$\forall x \in \Omega \quad \forall p \in D^-v(x) \quad H(x, v(x), p) \geqslant 0.$$

Finalement, S est une *solution de viscosité* de (8.4) si elle est à la fois sous-solution et sur-solution.

Remarque 8.2.7. Si S est une solution de viscosité nulle part différentiable, on impose des conditions là où $D^{\pm}S(x) \neq \emptyset$, *i.e.* sur un ensemble dense.

Remarque 8.2.8.

- Si S est une solution de classe C^1, alors S est aussi solution de viscosité.
- Réciproquement, si S est solution de viscosité, alors en tout point x de Ω où S est différentiable, on a $H(x, S(x), \nabla S(x)) = 0$.

Ceci assure la cohérence avec la notion classique de solution. En particulier, si S est lipschitzienne, alors l'équation d'Hamilton-Jacobi (8.4) est vraie presque partout.

Exemple 8.2.4. La solution de viscosité du problème

$$\left|\frac{\partial S}{\partial x}\right| - 1 = 0 \ \text{ sur }]0,1[, \quad S(0) = S(1) = 0,$$

est

$$S(x) = \begin{cases} x & \text{si} \quad 0 \leqslant x \leqslant 1/2, \\ 1 - x & \text{si} \quad 1/2 \leqslant x \leqslant 1. \end{cases}$$

Remarquons toutefois que S n'est pas solution de viscosité de $1 - \left|\frac{\partial S}{\partial x}\right| = 0$. Notons aussi que cette solution est bien $S(x) = d(x, \partial\Omega)$. Enfin, remarquons que, parmi l'infinité de solutions de ce problème, S est la seule à pouvoir être obtenue comme limite de viscosité évanescente. En effet, toute autre solution admet au moins un minimum local strict dans $]0,1[$. Or si S_ε converge uniformément vers S, avec $|\nabla S_\varepsilon| - 1 = \varepsilon \triangle S_\varepsilon$, et si on note x_ε un minimum local strict de S_ε, alors $\nabla S_\varepsilon(x_\varepsilon) = 0$ et $\triangle S_\varepsilon(x_\varepsilon) \geqslant 0$, ce qui est absurde.

On a les résultats suivants (voir [23, 7, 8]).

Théorème 8.2.1. *Soient Ω un ouvert borné de $\mathbb{R}^n$, g une fonction continue sur $\partial\Omega$, et $H : \Omega \times \mathbb{R}^n \to \mathbb{R}$ une fonction continue, uniformément continue en x au sens où il existe une fonction $\omega : \mathbb{R}^+ \to \mathbb{R}^+$ continue et croissante, avec $\omega(0) = 0$ telle que*

$$|H(x,p) - H(y,p)| \leqslant \omega(\|x-y\|(1+\|p\|)).$$

Alors le problème de Dirichlet

$$\begin{aligned} S(x) + H(x, \nabla S(x)) = 0 \ \textit{dans}\ \Omega, \\ S_{|\partial\Omega} = g, \end{aligned}$$

admet au plus une solution de viscosité.

Théorème 8.2.2. *Soient g une fonction continue sur $\mathbb{R}^n$, et $H : [0,T] \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ telle que*

$$\begin{aligned} &|H(t,x,p) - H(s,y,p)| \leqslant C(|t-s| + \|x-y\|)(1+\|p\|), \\ &|H(t,x,p) - H(t,x,q)| \leqslant C\|p-q\|. \end{aligned}$$

Alors le problème de Cauchy

$$\begin{aligned} \frac{\partial S}{\partial t} + H(t,x,\frac{\partial S}{\partial x}) = 0 \ \textit{dans}\ ]0,T[\times\mathbb{R}^n, \\ S(0,\cdot) = g(\cdot), \end{aligned}$$

admet au plus une solution de viscosité bornée et uniformément continue.

Il existe beaucoup de théorèmes de ce type. Ce sont des résultats d'unicité, sous des conditions fortes.

Une méthode pour prouver l'existence d'une solution de viscosité est de régulariser par une viscosité évanescente, de prouver l'existence d'une solution régulière S_ε, puis de faire des estimations uniformes pour passer à la limite (voir [23]). Un autre moyen de prouver d'obtenir des résultats d'existence (moins général cependant) et d'utiliser la théorie du contrôle optimal, en montrant que la fonction valeur associée à un problème de contrôle optimal est solution de viscosité d'une équation d'Hamilton-Jacobi. C'est l'objet de la section suivante.

Exercice 8.2.1. Soient g une fonction continue sur $\mathbb{R}^n$, et $H : \mathbb{R}^n \to \mathbb{R}$ une fonction convexe telle que

$$\frac{H(p)}{\|p\|} \underset{\|p\|\to+\infty}{\longrightarrow} +\infty.$$

En montrant que les caractéristiques sont des droites, montrer que la solution de viscosité du problème de Cauchy

$$\begin{aligned} \frac{\partial S}{\partial t} + H(\frac{\partial S}{\partial x}) = 0 \text{ dans } \mathbb{R}^n \times]0, +\infty[, \\ S(0, \cdot) = g(\cdot) \text{ sur } \mathbb{R}^n, \end{aligned}$$

est, pour tout $t \neq 0$,

$$S(t, x) = \min_{y \in \mathbb{R}^n} \left(tL(\frac{x-y}{t}) + g(y) \right),$$

où L est le Lagrangien associé au Hamiltonien H, *i.e.* $L(v) = \sup_{p \in \mathbb{R}^n} (\langle p, v \rangle - H(p))$. Cette formule s'appelle *formule de Hopf-Lax*.

8.3 Equations d'Hamilton-Jacobi en contrôle optimal

8.3.1 Equations d'Hamilton-Jacobi d'évolution

Définition de la fonction valeur

Soit $T > 0$ fixé et $U \subset \mathbb{R}^m$ un compact non vide. Pour tout $t \in]0, T]$ et tout $x \in \mathbb{R}^n$, considérons le problème de contrôle optimal général suivant : déterminer une trajectoire solution sur $[0, t]$ du système de contrôle

$$\begin{aligned} \dot{x}_u(s) &= f(x_u(s), u(s)), \quad u(s) \in U, \\ x_u(t) &= x, \end{aligned} \tag{8.5}$$

qui minimise le coût

$$C(t, u) = \int_0^t f^0(x_u(s), u(s)) ds + g(x_u(0)), \tag{8.6}$$

le point initial $x(0)$ étant libre, et le temps final t étant fixé.

Définition 8.3.1. Soit $x \in \mathbb{R}^n$. Définissons la *fonction valeur* S sur $[0, T] \times \mathbb{R}^n$ par

$$S(t, x) = \inf\{C(t, u) \mid x_u(\cdot) \text{ solution de (8.5)}\}.$$

La fonction valeur est la généralisation du concept de distance. Par exemple en géométrie Riemannienne elle généralise le concept de distance Riemannienne.

Remarque 8.3.1. Il est bien clair que $S(0, x) = g(x)$.

L'équation d'Hamilton-Jacobi

Etablissons tout d'abord formellement l'équation d'Hamilton-Jacobi. Supposons que pour tout $t \in]0,T]$ et tout $x \in \mathbb{R}^n$ il existe une trajectoire optimale $x_u(\cdot)$ solution du problème de contrôle optimal (8.5), (8.6) (voir théorème 6.2.1). On a alors $x = x_u(t)$, et donc

$$S(t,x) = S(t,x_u(t)) = C(t,u) = \int_0^t f^0(x_u(s),u(s))ds + g(x_u(0)).$$

En dérivant formellement par rapport à t, on obtient

$$\frac{\partial S}{\partial t}(t,x_u(t)) + \frac{\partial S}{\partial x}(t,x_u(t))f(x_u(t),u(t)) = f^0(x_u(t),u(t)),$$

et donc

$$\frac{\partial S}{\partial t}(t,x) + \frac{\partial S}{\partial x}(t,x))f(x,u(t)) - f^0(x,u(t)) = 0.$$

Introduisons par ailleurs le Hamiltonien du problème de contrôle optimal

$$H(x,p,p^0,u) = \langle p, f(x,u)\rangle + p^0 f^0(x,u).$$

D'après le principe du maximum, le contrôle optimal $u(\cdot)$ doit vérifier

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

On obtient par conséquent

$$-p^0\frac{\partial S}{\partial t}(t,x) + \max_{v\in U} H(x,-p^0\frac{\partial S}{\partial x}(t,x),p^0,v) = 0. \tag{8.7}$$

L'équation (8.7) est l'équation générale dite de Hamilton-Jacobi-Bellman en contrôle optimal.

Remarque 8.3.2. S'il n'y a pas d'extrémale anormale, on peut supposer dans le calcul formel précédent que $p^0 = -1$, et on obtient alors l'équation usuelle

$$\frac{\partial S}{\partial t} + H_1(x,\frac{\partial S}{\partial x}) = 0,$$

où $H_1(x,p) = \max_{v\in U} H(x,p,-1,v)$.

Le calcul précédent est formel. En utilisant la notion de solution de viscosité, on a le résultat rigoureux suivant (voir [23, 7, 8]).

Théorème 8.3.1. *On suppose qu'il existe une constante $C > 0$ telle que, pour tous $x, y \in \mathbb{R}^n$ et tout $u \in U$, on ait*

$$\begin{aligned}
&\|f(x,u)\| \leqslant C,\ \|f^0(x,u)\| \leqslant C,\ \|g(x)\| \leqslant C,\\
&\|f(x,u) - f(y,u)\| \leqslant C\|x-y\|,\\
&\|f^0(x,u) - f^0(y,u)\| \leqslant C\|x-y\|,\\
&\|g(x) - g(y)\| \leqslant C\|x-y\|.
\end{aligned}$$

Alors la fonction valeur S est bornée, lipschitzienne en (t,x), et est l'unique solution de viscosité du problème de Dirichlet

$$\begin{aligned} \frac{\partial S}{\partial t} + H_1(x, \frac{\partial S}{\partial x}) = 0 \quad \textit{dans }]0,T[\times \mathbb{R}^n, \\ S(0,\cdot) = g(\cdot) \quad \textit{sur } \mathbb{R}^n, \end{aligned} \tag{8.8}$$

où $H_1(x,p) = \max_{v\in U} H(x,p,-1,v) = \max_{v\in U} \big(\langle p, f(x,v)\rangle - f^0(x,v)\big)$.

Remarque 8.3.3. En contrôle optimal, si on est capable de résoudre l'équation d'Hamilton-Jacobi alors on est capable d'exprimer les contrôles optimaux comme des feedbacks. En effet, rappelons que le principe du maximum permet d'exprimer les contrôles optimaux comme fonctions de (x,p). Or on vient de voir précédemment que $p(t) = -p^0 \frac{\partial S}{\partial x}(t,x(t))$ (au moins si S est différentiable en ce point). La connaissance de la solution S donne donc beaucoup plus que le principe du maximum, mais bien entendu cette équation d'Hamilton-Jacobi est aussi beaucoup plus difficile à résoudre. Pour les aspects numériques, voir le chapitre 9.

Remarque 8.3.4. Dans le cas de systèmes linéaires avec un coût quadratique, on retrouve l'équation de Riccati. En liaison avec la remarque précédente, on retrouve donc le fait que, dans le cadre LQ, l'équation de Riccati permet d'exprimer les contrôles optimaux comme des feedbacks.

Faisons enfin une dernière remarque qui fait le lien entre la théorie d'Hamilton-Jacobi et le principe du maximum.

Remarque 8.3.5. Au moins dans le cas où $\Omega = \mathbb{R}^m$, *i.e.* s'il n'y a pas de contrainte sur le contrôle, et si le contrôle s'exprime, par le PMP, comme une fonction lisse de (x,p), alors les extrémales du principe du maximum sont les courbes caractéristiques de l'équation d'Hamilton-Jacobi (8.8).

En effet, la méthode des caractéristiques consiste à résoudre, pour trouver une solution lisse de (8.8), le système d'équations

$$\dot{x} = \frac{\partial H_1}{\partial p},\ \dot{p} = -\frac{\partial H_1}{\partial x},\ x(0) = \bar{x} \in \Omega,\ p(0) = \nabla g(\bar{x}).$$

Notons $(x(t,\bar{x}), p(t,\bar{x}))$ la solution correspondante. Par construction, on a

$$p(t,\bar{x}) = \frac{\partial S}{\partial x}(t, x(t,\bar{x})),$$

d'où, en utilisant (8.8),

$$\frac{\partial p}{\partial t}(t,\bar{x}) = -\frac{\partial H_1}{\partial x}(x(t,\bar{x}), p(t,\bar{x})).$$

Par ailleurs, par hypothèse $H_1(x,p) = H(x,p,-1,u(x,p))$, avec de plus $\frac{\partial H}{\partial u}(x,p,-1,u(x,p)) = 0$ puisqu'il n'y a pas de contrainte sur le contrôle. Par conséquent

$$\frac{\partial H_1}{\partial x} = \frac{\partial H}{\partial x} + \frac{\partial H}{\partial u}\frac{\partial u}{\partial x} = \frac{\partial H}{\partial x},$$

et de même

$$\frac{\partial H_1}{\partial p} = \frac{\partial H}{\partial p}.$$

On retrouve donc les équations du principe du maximum

$$\frac{\partial x}{\partial t} = \frac{\partial H}{\partial p},\ \frac{\partial p}{\partial t} = -\frac{\partial H}{\partial x}.$$

Variante, à point initial fixé

Dans le problème précédent, le point initial était libre, ce qui a permis de résoudre un problème de Dirichlet. Fixons maintenant le point initial x_0, et considérons donc le problème de contrôle optimal

$$\begin{aligned} &\dot{x}_u(s) = f(x_u(s), u(s)), \quad u(s) \in U, \\ &x_u(0) = x_0,\ x_u(t) = x, \\ &C(t,u) = \int_0^t f^0(x_u(s), u(s))ds, \end{aligned} \tag{8.9}$$

le temps final t étant fixé. Pour tout $t \in [0,T]$ et tout $x \in \mathbb{R}^n$, la fonction valeur est définie par

$$S(t,x) = \inf\{C(t,u) \mid x_u(\cdot) \text{ solution de (8.9)}\}.$$

La différence par rapport au cas précédent réside dans la donnée initiale. Ici, il est clair que $S(0,x_0) = 0$, et par convention on pose $S(0,x) = +\infty$ si $x \neq x_0$.

Théorème 8.3.2. *On suppose qu'il existe une constante $C > 0$ telle que, pour tous $x, y \in \mathbb{R}^n$ et tout $u \in U$, on ait*

$$\begin{aligned} &\|f(x,u)\| \leqslant C,\ \|f^0(x,u)\| \leqslant C, \\ &\|f(x,u) - f(y,u)\| \leqslant C\|x-y\|, \\ &\|f^0(x,u) - f^0(y,u)\| \leqslant C\|x-y\|. \end{aligned}$$

Alors la fonction valeur S est solution de viscosité de

$$\begin{aligned} &\frac{\partial S}{\partial t} + H_1(x, \frac{\partial S}{\partial x}) = 0 \quad \textit{dans }]0,T[\times\mathbb{R}^n, \\ &S(0,x_0) = 0,\ S(0,x) = +\infty \textit{ si } x \neq x_0, \end{aligned} \tag{8.10}$$

où $H_1(x,p) = \max_{v\in U} H(x,p,-1,v) = \max_{v\in U}\big(\langle p, f(x,v)\rangle - f^0(x,v)\big)$.

8.3.2 Equations d'Hamilton-Jacobi stationnaires

On obtient des équations d'Hamilton-Jacobi stationnaires en laissant le temps final libre. Pour simplifier, on se limite ici au problème du temps minimal (voir [7] pour une généralisation). Considérons le problème de temps minimal

$$\begin{aligned} &\dot{x}_u(s) = f(x_u(s), u(s)), \quad u(s) \in U, \\ &x_u(0) = x_0,\ x_u(t) = x. \end{aligned} \tag{8.11}$$

Pour tout $x \in \mathbb{R}^n$, la fonction valeur, appelée fonction *temps minimal*, est définie par

$$T(x) = \inf\{t \mid x_u(\cdot) \text{ solution de (8.11)}\}.$$

Comme précédemment, on a $T(x_0) = 0$, et $T(x) = +\infty$ si $x \neq x_0$.

Etablissons tout d'abord formellement l'équation d'Hamilton-Jacobi vérifiée par T. Supposons que pour tout $x \in \mathbb{R}^n$ il existe une trajectoire temps minimale $x_u(\cdot)$ reliant x_0 à x. On a alors $x = x_u(t)$, et donc $T(x) = T(x_u(t)) = t$. En dérivant formellement par rapport à t, on obtient

$$\langle \nabla T(x_u(t)), f(x_u(t), u(t)) \rangle = 1.$$

On en déduit

$$\max_{v \in U} H(x, -p^0 \nabla T(x), p^0, v) = 0, \tag{8.12}$$

où $H(x, p, p^0, u) = \langle p, f(x, u) \rangle + p^0$ est le Hamiltonien du problème de temps minimal. S'il n'y a pas d'extrémale anormale, on peut supposer que $p^0 = -1$, et on obtient l'équation usuelle d'Hamilton-Jacobi pour le temps minimal.

En utilisant la notion de solution de viscosité, on a le résultat suivant (voir [23, 7, 8]).

Théorème 8.3.3. *On suppose qu'il existe une constante $C > 0$ telle que, pour tous $x, y \in \mathbb{R}^n$ et tout $u \in U$, on ait*

$$\|f(x, u)\| \leqslant C, \ \|f(x, u) - f(y, u)\| \leqslant C\|x - y\|.$$

Alors la fonction temps minimal T est solution de viscosité de

$$\begin{aligned} &\max_{v \in U} \langle \nabla T(x), f(x, v) \rangle = 1 \text{ dans } \mathbb{R}^n, \\ &T(x_0) = 0, \ T(x) = +\infty \text{ si } x \neq x_0. \end{aligned} \tag{8.13}$$

Remarque 8.3.6. Si $f(x, u) = u$ et U est la boule unité de $\mathbb{R}^n$, on retrouve l'équation Eikonale de l'introduction.

Exemple 8.3.1. Considérons le système de contrôle

$$\dot{x} = y, \ \dot{y} = z, \ \dot{z} = u, \text{ avec } |u| \leqslant 1.$$

La fonction temps minimal à partir d'un point fixé vérifie l'équation d'Hamilton-Jacobi

$$y \frac{\partial T}{\partial x} + z \frac{\partial T}{\partial y} + \left| \frac{\partial T}{\partial z} \right| = 1.$$

Chapitre 9

Méthodes numériques en contrôle optimal

On distingue deux types de méthodes numériques en contrôle optimal : les méthodes directes et les méthodes indirectes. Les méthodes directes consistent à discrétiser l'état et le contrôle, et réduisent le problème à un problème d'optimisation non linéaire (programmation non linéaire, ou "nonlinear programming"). Les méthodes indirectes consistent à résoudre numériquement, par une méthode de tir ("shooting method"), un problème aux valeurs limites obtenu par application du principe du maximum. Dans ce chapitre, on s'intéresse d'abord aux méthodes indirectes, puis aux méthodes directes. Dans une dernière section, on compare les méthodes, et on décrit les méthodes hybrides qui sont un mélange des deux approches.

9.1 Méthodes indirectes

9.1.1 Méthode de tir simple

Le principe est le suivant. Considérons le problème de contrôle optimal (4.28), (4.29), et supposons dans un premier temps que le temps final t_f est fixé. Le principe du maximum donne une condition nécessaire d'optimalité et affirme que toute trajectoire optimale est la projection d'une extrémale. Si l'on est capable, à partir de la condition de maximum, d'exprimer le contrôle extrémal en fonction de $(x(t), p(t))$, alors le système extrémal est un système différentiel de la forme $\dot{z}(t) = F(t, z(t))$, où $z(t) = (x(t), p(t))$, et les conditions initiales, finales, et les conditions de transversalité, se mettent sous la forme $R(z(0), z(t_f)) = 0$. Finalement, on obtient le *problème aux valeurs limites*

$$\begin{cases} \dot{z}(t) = F(t, z(t)), \\ R(z(0), z(t_f)) = 0. \end{cases} \tag{9.1}$$

Notons $z(t, z_0)$ la solution du problème de Cauchy

$$\dot{z}(t) = F(t, z(t)),\ z(0) = z_0,$$

et posons $G(z_0) = R(z_0, z(t_f, z_0))$. Le problème (9.1) aux valeurs limites est alors équivalent à

$$G(z_0) = 0,$$

i.e. il s'agit de *déterminer un zéro de la fonction G*.

Ceci peut se résoudre par une méthode de Newton (voir section 9.1.3).

Remarque 9.1.1. Si le temps final t_f est libre, on peut se ramener à la formulation précédente en considérant t_f comme une inconnue auxiliaire. On augmente alors la dimension de l'état en considérant l'équation supplémentaire $\frac{dt_f}{dt} = 0$. 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 le Hamiltonien.

9.1.2 Méthode de tir multiple

Par rapport à la méthode de tir simple, la méthode de tir multiple découpe l'intervalle $[0, t_f]$ 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. Une référence classique pour l'algorithme de tir multiple est [65].

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

$$\dot{z}(t) = F(t, z(t)) = \begin{cases} F_0(t, z(t)) & \text{si } \ t_0 \leqslant t < t_1 \\ F_1(t, z(t)) & \text{si } \ t_1 \leqslant t < t_2 \\ \quad \vdots & \\ F_s(t, z(t)) & \text{si } \ t_s \leqslant t \leqslant t_f \end{cases} \tag{9.2}$$

où $z = (x, p) \in \mathbb{R}^{2n}$ (p est le vecteur adjoint), et $t_1, t_2, \ldots, t_s \in [t_0, t_f]$ peuvent être des *temps de commutation* ; dans le cas où le problème inclut des contraintes sur l'état, ce peut être des *temps de jonction* avec un arc frontière, ou bien des *temps de contact* avec la frontière. On a de plus des conditions de continuité sur l'état et le vecteur adjoint aux points de commutation. Dans le cas de contraintes sur l'état, on a des conditions de saut sur le vecteur adjoint, et des conditions sur la contrainte c en des points de jonction ou de contact (voir à ce sujet [42, 55, 21, 56, 15, 14]). De plus on a des conditions aux limites sur l'état, le vecteur adjoint (conditions de transversalité), et sur le Hamiltonien si le temps final est libre.

Remarque 9.1.2. A priori le temps final t_f est inconnu. Par ailleurs dans la méthode de tir multiple le nombre s de commutations doit être fixé ; on le détermine lorsque c'est possible par une analyse géométrique du problème.

La méthode de tir multiple consiste à subdiviser l'intervalle $[t_0, t_f]$ en N sous-intervalles, la valeur de $z(t)$ au début de chaque sous-intervalle étant inconnue. Plus précisément, soit $t_0 < \sigma_1 < \cdots < \sigma_k < t_f$ une subdivision *fixée* de l'intervalle $[t_0, t_f]$. En tout point σ_j la fonction z est *continue*. On peut considérer σ_j comme un point de commutation fixe, en lequel on a

$$\begin{cases} z(\sigma_j^+) = z(\sigma_j^-), \\ \sigma_j = \sigma_j^* \quad \text{fixé.} \end{cases}$$

On définit maintenant les *noeuds*

$$\{\tau_1, \ldots, \tau_m\} \;=\; \{t_0, t_f\} \cup \{\sigma_1, \ldots, \sigma_k\} \cup \{t_1, \ldots, t_s\}. \tag{9.3}$$

Finalement on est conduit au *problème aux valeurs limites*

$$\begin{aligned} &\bullet\ \dot{z}(t) = F(t, z(t)) = \begin{cases} F_1(t, z(t)) & \text{si } \tau_1 \leqslant t < \tau_2 \\ F_2(t, z(t)) & \text{si } \tau_2 \leqslant t < \tau_3 \\ \quad\vdots & \\ F_{m-1}(t, z(t)) & \text{si } \tau_{m-1} \leqslant t \leqslant \tau_m \end{cases} \\ &\bullet\ \forall j \in \{2, \ldots, m-1\} \quad r_j(\tau_j, z(\tau_j^-), z(\tau_j^+)) = 0 \\ &\bullet\ r_m(\tau_m, z(\tau_1), z(\tau_m)) = 0 \end{aligned} \tag{9.4}$$

où $\tau_1 = t_0$ est fixé, $\tau_m = t_f$, et les r_j représentent les conditions intérieures ou limites précédentes.

Remarque 9.1.3. On améliore la stabilité de la méthode en augmentant le nombre de noeuds. C'est là en effet le principe de la méthode de tir multiple, par opposition à la méthode de tir simple où les erreurs par rapport à la condition initiale évoluent exponentiellement en fonction de t_f-t_0 (voir [65]). Bien sûr dans la méthode de tir multiple il y a beaucoup plus d'inconnues que dans la méthode de tir simple, mais éventuellement l'intégration du système (9.2) peut se paralléliser.

Posons $z_j^+ = z(\tau_j^+)$, et soit $z(t, \tau_{j-1}, z_{j-1}^+)$ la solution du problème de Cauchy

$$\dot{z}(t) = F(t, z(t)),\ z(\tau_{j-1}) = z_{j-1}^+.$$

On a

$$z(\tau_j^-) = z(\tau_j^-, \tau_{j-1}, z_{j-1}^+).$$

Les conditions intérieures et frontières s'écrivent

$$\begin{aligned} \forall j \in \{2, \ldots, m-1\} \quad r_j(\tau_j, z(\tau_j^-, \tau_{j-1}, z_{j-1}^+), z_j^+) = 0, \\ r_m(\tau_m, z_1^+, z(\tau_m^-, \tau_{m-1}, z_{m-1}^+)) = 0. \end{aligned} \tag{9.5}$$

Posons maintenant

$$Z = (z_1^+, \tau_m, z_2^+, \tau_2, \ldots, z_{m-1}^+, \tau_{m-1})^T \ \in \ \mathbb{R}^{(2n+1)(m-1)}$$

(où $z \in \mathbb{R}^{2n}$). Alors les conditions (9.5) sont vérifiées si

$$G(Z) \ = \ \begin{pmatrix} r_m(\tau_m, z_1^+, z(\tau_m^-, \tau_{m-1}, z_{m-1}^+)) \\ r_2(\tau_2, z(\tau_2^-, \tau_1, z_1^+), z_2^+) \\ \vdots \\ r_{m-1}(\tau_m, z(\tau_{m-1}^-, \tau_{m-2}, z_{m-2}^+), z_{m-1}^+) \end{pmatrix} \ = \ 0. \qquad (9.6)$$

On s'est donc ramené à déterminer un zéro de la fonction G, qui est définie sur un espace vectoriel dont la dimension est proportionnelle au nombre de points de commutation et de points de la subdivision. L'équation $G = 0$ peut alors être résolue itérativement par une méthode de type Newton (voir la section suivante).

9.1.3 Rappels sur les méthodes de Newton

Il s'agit de résoudre numériquement $G(z) = 0$, où $G : \mathbb{R}^p \to \mathbb{R}^p$ est une fonction de classe C^1. L'idée de base est la suivante. Si z_k est proche d'un zéro z de G, alors

$$0 = G(z) = G(z_k) + dG(z_k).(z - z_k) + \mathrm{o}(z - z_k).$$

On est alors amené à considérer la suite définie par récurrence

$$z_{k+1} = z_k - (dG(z_k))^{-1}.G(z_k),$$

un point initial $z_0 \in \mathbb{R}^p$ étant choisi, et on espère que z_k converge vers le zéro z. Ceci suppose donc le calcul de l'inverse de la matrice jacobienne de G, ce qui doit être évité numériquement. Il s'agit alors, à chaque étape, de résoudre l'équation

$$G(z_k) + dG(z_k).d_k = 0,$$

où d_k est appelé *direction de descente*, et on pose $z_{k+1} = z_k + d_k$.

Sous des hypothèses générales, l'algorithme de Newton converge, et la convergence est quadratique (voir par exemple [6, 61, 65]). Il existe de nombreuses variantes de la méthode Newton : méthode de descente, de quasi-Newton, de Newton quadratique, de Broyden, ... Cette méthode permet, en général, une détermination très précise d'un zéro. Son inconvénient principal est la petitesse du domaine de convergence. Pour faire converger la méthode, il faut que le point initial z_0 soit suffisamment proche de la solution recherchée z. Ceci suppose donc que pour déterminer le zéro z il faut avoir au préalable une idée approximative de la valeur de z.

Du point de vue du contrôle optimal, cela signifie que, pour appliquer une méthode de tir, il faut avoir une idée *a priori* de la trajectoire optimale cherchée.

Ceci peut sembler paradoxal, mais il existe des moyens de se donner une approximation, même grossière, de cette trajectoire optimale. Il s'agit là en tout cas d'une caractéristique majeure des méthodes de tir : elles sont très précises mais requièrent une connaissance a priori (plus ou moins grossière) de la trajectoire optimale cherchée.

9.2 Méthodes directes

Les méthodes directes consistent à transformer le problème de contrôle optimal en un *problème d'optimisation non linéaire en dimension finie.*

9.2.1 Discrétisation totale : tir direct

C'est la méthode la plus évidente 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 (ou problème de programmation non linéaire) de la forme

$$\min_{Z\in C} F(Z), \tag{9.7}$$

où $Z = (x_1, \ldots, x_N, u_1, \ldots, u_n)$, et

$$\begin{aligned} C \;=\; \{Z \mid\; & g_i(Z) = 0,\ i \in 1, \ldots, r, \\ & g_j(Z) \leqslant 0,\ j \in r+1, \ldots, m\}. \end{aligned} \tag{9.8}$$

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 $0 = t_0 < t_1 < \cdots < t_N = t_f$ de l'intervalle $[0, t_f]$. Réduisons l'espace des contrôles en considérant (par exemple) des 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 explicite. 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 9.2.1. Il existe une infinité de variantes. D'une part, on peut discrétiser l'ensemble des contrôles admissibles par des contrôles constants par morceaux, ou affines par morceaux, ou des splines, etc. D'autre part, il existe de nombreuses méthodes pour discrétiser une équation différentielle ordinaire : méthode d'Euler (explicite ou implicite), point milieu, Heun, Runge-Kutta, Adams-Moulton, etc (voir par exemple [24, 49, 61, 65]). Le choix de la méthode dépend du problème abordé.

La discrétisation précédente conduit donc au problème de programmation non linéaire

$$\begin{aligned} & x_{i+1} = x_i + h_i f(t_i, x_i, u_i),\ i = 0, \ldots, N-1, \\ & \min C(x_0, \ldots, x_N, u_0, \ldots, u_N), \\ & u_i \in \Omega,\ i = 0, \ldots, N-1, \end{aligned}$$

i.e. un problème du type (9.7).

Remarque 9.2.2. Cette méthode est très simple à mettre en oeuvre. De plus l'introduction d'éventuelles contraintes sur l'état ne pose aucun problème.

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

$$u \in \mathrm{Vect}(U_1, \ldots, U_N), \ \textit{i.e.} \ \ u(t) = \sum_{i=1}^{N} u_i U_i(t), \ u_i \in \mathbb{R},$$
$$x \in \mathrm{Vect}(X_1, \ldots, X_N), \ \textit{i.e.} \ \ x(t) = \sum_{i=1}^{N} x_i X_i(t), \ x_i \in \mathbb{R},$$

où les $U_i(t)$ et $X_i(t)$ représentent une base de Galerkin. 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 la discrétisation. On se ramène bien à un problème d'optimisation non linéaire en dimension finie de la forme (9.7).

La résolution numérique d'un problème de programmation non linéaire du type (9.7) est standard. Elle peut être effectuée, par exemple, par une méthode de pénalisation, ou par une méthode SQP (*sequential 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 (9.7), (9.8), 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

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

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 basé sur une approximation quadratique du Lagrangien. Pour plus de détails sur cette méthode, voir [29, 33].

Il y a une infinité de variantes des méthodes directes. L'approche décrite ci-dessus permet déjà de considérer de nombreuses variantes, mais il faut mentionner aussi les méthodes pseudo-spectrales, de collocation, etc. Pour un survey très complet sur les méthodes directes et leur mise en oeuvre numérique, nous renvoyons le lecteur à l'excellent livre [11].

9.2.2 Résolution numérique de l'équation d'Hamilton-Jacobi

Il existe de nombreuses méthodes numériques pour résoudre l'équation d'Hamilton-Jacobi

$$\frac{\partial S}{\partial t} + H_1(x, \frac{\partial S}{\partial x}) = 0,$$

où $H_1(x,p) = \max_{u\in U} \langle p, f(x,u)\rangle - f^0(x,u)$. Commençons par décrire une discrétisation simple de cette équation par différences finies. Il convient de remarquer, similairement aux équations de transport, que pour assurer la stabilité il faut décentrer le schéma. Ainsi, pour discrétiser

$$\left\langle \frac{\partial S}{\partial x}, f(x,u) \right\rangle = \sum_{p=1}^{n} \frac{\partial S}{\partial x_p} f_p(x,u),$$

on est amené à discrétiser $\frac{\partial S}{\partial x_p}$ par une différence divisée à droite ou à gauche, selon le signe de $f_p(x,u)$.

Considérons un maillage de l'espace $(x_{\bar{\imath}})$, où $\bar{\imath} = (i_1, \ldots, i_n) \in \mathbb{Z}^n$, supposé régulier pour simplifier, et une discrétisation régulière (t_j) de l'intervalle de temps. Notons $h = (h_1, \ldots, h_n)$ le pas d'espace, et et $k = t_{j+1} - t_j$ le pas de temps. Soit $S_{\bar{\imath},j}$ la valeur approchée de $S(t_j, x_{\bar{\imath}})$. Il convient d'approcher $\frac{\partial S}{\partial x_p}(x_{\bar{\imath}})$ par une différence divisée à gauche (resp. à droite) si $f_p(x_{\bar{\imath}}, u)$ est positif (resp. négatif). Pour tout réel a, on pose

$$a_+ = \max(a,0) = \frac{a+|a|}{2},\ a_- = \min(a,0) = \frac{a-|a|}{2}.$$

Pour tout $p \in \{1, \ldots, n\}$, on note $e_p = (0, \ldots, 1, \ldots, 0)$, le "1" étant en p-ème position. On obtient donc le schéma explicite

$$\begin{aligned} 0 \ &= \frac{S_{\bar{\imath},k+1} - S_{\bar{\imath},k}}{k} \\ &\quad + \max_{u\in U} \left(\sum_{p=1}^{n} \big(f_p(x,u)_+ \frac{S_{\bar{\imath},k} - S_{\bar{\imath}-e_p,k}}{h_p} + f_p(x,u)_- \frac{S_{\bar{\imath}+e_p,k} - S_{\bar{\imath},k}}{h_p} \big) - f^0(x_{\bar{\imath}}, u) \right). \end{aligned}$$

Il existe de nombreuses méthodes de discrétisation. Le schéma de discrétisation par différences finies proposé ci-dessus est le plus simple, mais on peut adopter des schémas d'ordre supérieur.

Il existe aussi les méthodes de front d'onde (voir [63]), qui consistent à calculer les ensembles de niveau de la fonction valeur S solution de l'équation d'Hamilton-Jacobi. Particulièrement efficaces en petite dimension, ces méthodes consistent à faire évoluer le front d'onde de la fonction valeur en partant d'un point ou d'un ensemble initial donné. La complexité algorithmique est linéaire en fonction du nombre de points de discrétisation. Ces méthodes ont été implémentées de manière très efficace sur des problèmes de dimension moyenne (typiquement 3). La construction de tels schémas n'est cependant pas immédiate, et en fonction de l'équation il faut être capable d'élaborer un schéma stable et consistant (voir [63] pour des exemples).

Remarque 9.2.3. Notons que, de même que précédemment, l'introduction de contraintes sur l'état ne pose aucun problème : il suffit en effet d'imposer à la fonction valeur d'être égale à $+\infty$ sur le domaine interdit. Numériquement, cela signifie qu'on impose une valeur assez grande à la fonction valeur, en les points du maillage qui sont dans le domaine interdit.

Remarque 9.2.4. Lorsqu'on a localisé les courbes de commutations, on peut éventuellement raffiner le maillage autour de ces courbes pour obtenir une meilleure précision.

9.3 Quelle méthode choisir ?

Les méthodes directes présentent les avantages suivants sur les méthodes indirectes :

- leur mise en oeuvre est plus simple car elles ne nécessitent pas une étude théorique préalable comme les méthodes indirectes ; en particulier, on n'a pas à étudier les variables adjointes, ou bien à connaître à l'avance la structure des commutations ;
- elles sont plus robustes ;
- elles sont peu sensibles au choix de la condition initiale (contrairement aux méthodes indirectes, cf ci-dessous) ;
- il est facile de tenir compte d'éventuelles contraintes sur l'état ;
- elles permettent de calculer les contrôles optimaux sous forme de feedback, *i.e.* en boucle fermée, ce qui est particulièrement adapté aux problèmes de stabilisation, et/ou à la mise en oeuvre de systèmes embarqués.

En revanche,

- les méthodes directes sont moins précises que les méthodes indirectes ; par exemple dans les problèmes de contrôle optimal issus de l'aéronautique, la précision des méthodes directes s'avère en général insuffisante, malgré l'augmentation du nombre de pas de la discrétisation ;
- la discrétisation directe d'un problème de contrôle optimal comporte souvent plusieurs minima (locaux), et les méthodes directes peuvent converger vers ces minima ; pourtant la solution ainsi déterminée peut s'avérer être très éloignée de la vraie solution optimale ;
- les méthodes directes sont gourmandes en mémoire, et de ce fait peuvent devenir inefficaces si la dimension d'espace est trop grande.

Remarque 9.3.1. Si la dynamique du système de contrôle est compliquée, le calcul du système extrémal, notamment des équations adjointes, peut être effectué avec un logiciel de calcul formel comme *Maple*.

Les avantages des méthodes indirectes sont

- l'extrême précision numérique ;
- la méthode de tir multiple est, par construction, parallélisable, et son implémentation peut donc être envisagée sur un réseau d'ordinateurs montés en parallèle.

Les inconvénients des méthodes indirectes sont les suivants :

- elles calculent les contrôles optimaux sous forme de boucle ouverte ;
- elles sont basées sur le principe du maximum qui est une condition nécessaire d'optimalité seulement, et donc il faut être capable de vérifier a posteriori l'optimalité de la trajectoire calculée ;
- rigidité de la méthode : la structure des commutations doit être connue à l'avance (par exemple par une étude géométrique du problème). De même, il n'est pas facile d'introduire des contraintes sur l'état, car d'une part cela requiert d'appliquer un principe du maximum tenant compte de ces contraintes (qui est beaucoup plus compliqué que le principe du maximum standard), d'autre part la présence de contraintes sur l'état peut rendre compliquée la structure de la trajectoire optimale, notamment la structure de ses commutations.
- Deuxièmement, il faut être capable de deviner de bonnes conditions initiales pour l'état et le vecteur adjoint, pour espérer faire converger la méthode de tir. En effet le domaine de convergence de la méthode de Newton peut être assez petit en fonction du problème de contrôle optimal.

Remarque 9.3.2. Que l'on ait utilisé une méthode directe ou une méthode indirecte, il faut être capable de vérifier, a posteriori, que l'on a bien obtenu la trajectoire optimale. Les causes sont cependant différentes selon la méthode.

- Si on a utilisé une méthode directe, il se peut qu'elle ait convergé vers un (pseudo)-minimum local, dû à la discrétisation du problème. Notons toutefois que l'équation d'Hamilton-Jacobi donne une condition nécessaire et suffisante d'optimalité, et conduit donc à des trajectoires globalement optimales.
- Les méthodes indirectes sont basées sur le principe du maximum qui donne une condition nécessaire d'optimalité locale. Une fois ces trajectoires déterminées, la théorie des points conjugués permet d'établir qu'une extrémale est localement optimale avant son premier temps conjugué (voir [13]). L'optimalité globale est beaucoup plus difficile à établir en général, et sur des exemples spécifiques on l'établit numériquement.

Remarque 9.3.3. Les méthodes directes donnent les contrôles extrémaux sous forme de boucle fermée, et les méthodes indirectes sous forme de boucle ouverte seulement. Cependant, une trajectoire optimale ayant été déterminée par une méthode indirecte, on peut stabiliser cette trajectoire en calculant, par une méthode LQ par exemple, un contrôle feedback localement autour ce la trajectoire.

Le tableau suivant résume les caractéristiques des méthodes directes et indirectes.

Méthodes directes	**Méthodes indirectes**
mise en oeuvre simple, sans connaissance a priori	connaissance a priori de la structure de la trajectoire optimale
peu sensibles au choix de la condition initiale	très sensibles au choix de la condition initiale
facilité de la prise en compte de contraintes sur l'état	difficulté théorique de la prise en compte de contraintes sur l'état
contrôles (globalement) optimaux en boucle fermée	contrôles (localement) optimaux en boucle ouverte
précision numérique basse ou moyenne	très grande précision numérique
efficaces en basse dimension	efficaces en toute dimension
gourmandise en mémoire	calculs parallélisables
problème des minima locaux	petit domaine de convergence

Pour pallier l'inconvénient majeur des méthodes indirectes, à savoir la sensibilité extrême par rapport à la condition initiale, on propose plusieurs solutions.

Une première solution raisonnable consiste à combiner les deux approches : méthodes directes et indirectes, de façon à obtenir ce qu'on appelle une *méthode hybride*. Quand on aborde un problème de contrôle optimal, on peut d'abord essayer de mettre en oeuvre une méthode directe. On peut ainsi espérer obtenir une idée assez précise de la structure des commutations, ainsi qu'une bonne approximation de la trajectoire optimale, et du vecteur adjoint associé. Si on souhaite plus de précision numérique, on met alors en oeuvre une méthode de tir, en espérant que le résultat fourni par la méthode directe donne une approximation suffisante de la trajectoire optimale cherchée, fournissant ainsi un point de départ appartenant au domaine de convergence de la méthode de tir. En combinant ainsi les deux approches (méthodes directes puis indirectes), on peut bénéficier de l'excellente précision numérique fournie par la méthode de tir tout en réduisant considérablement le désavantage dû à la petitesse de son domaine de convergence.

En appliquant d'abord une méthode directe, on peut obtenir une approximation de l'état adjoint. En effet, on a vu qu'une méthode directe consiste à résoudre numériquement un problème de programmation non linéaire avec contraintes. Les multiplicateurs de Lagrange associés au Lagrangien de ce problème de programmation non linéaire donnent une approximation de l'état adjoint (on a déjà vu que le vecteur adjoint n'est rien d'autre qu'un multiplicateur de Lagrange). A ce sujet, voir [21, 66, 34].

Une deuxième solution consiste à utiliser une *méthode d'homotopie* (ou *méthode de continuation*). Il s'agit de construire une famille de problèmes de contrôle optimal $(\mathcal{P}_\alpha)_{\alpha\in[0,1]}$ dépendant d'un paramètre $\alpha \in [0,1]$, où le problème initial correspond à $\mathcal{P}_0$. On doit s'arranger pour que le problème $\mathcal{P}_1$ soit plus simple à résoudre que $\mathcal{P}_0$. Une telle famille ne peut être construite que si l'on possède une bonne intuition et une bonne connaissance de la physique du problème. Par la méthode de tir, chaque problème de contrôle optimal $\mathcal{P}_\alpha$ se ramène à la détermination d'un zéro d'une fonction. On obtient donc une famille

à un paramètre d'équations non linéaires

$$G_\alpha(Z) = 0,\ \alpha \in [0,1].$$

Supposons avoir résolu numériquement le problème $\mathcal{P}_1$, et considérons une subdivision $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_p = 1$ de l'intervalle $[0,1]$. La solution de $\mathcal{P}_1$ peut alors être utilisée comme point de départ de la méthode de tir appliquée au problème $\mathcal{P}_{\alpha_{p-1}}$. Puis, par une procédure inductive finie, la solution du problème $\mathcal{P}_{\alpha_{i+1}}$ constitue une condition initiale pour appliquer la méthode de tir au problème $\mathcal{P}_{\alpha_i}$. Bien entendu il faut choisir judicieusement la subdivision (α_i), et éventuellement la raffiner.

Pour faciliter l'intuition, il est important que le paramètre α soit un paramètre naturel du problème. Par exemple si le problème de contrôle optimal comporte une contrainte forte sur l'état, du type $c(x) \leqslant 1$, une méthode d'homotopie peut consister à relaxer cette contrainte, en résolvant d'abord des problèmes où $c(x) \leqslant A$, avec $A > 0$ grand. Cela revient donc à résoudre une série de problèmes de contrôle optimal où l'on introduit petit à petit la contrainte sur l'état. Mathématiquement, pour pouvoir espérer la convergence de la méthode en passant d'un pas à un autre, il faut que la chaîne de problèmes de contrôle optimal introduite dépende continûment du paramètre α.

On peut généraliser cette approche par homotopie :

- chaque problème $\mathcal{P}_\alpha$ peut lui-même être résolu par homotopie, *i.e.* par la résolution de sous-problèmes (ce peut être le cas si par exemple le problème de contrôle optimal initial comporte plusieurs contraintes sur l'état fortement actives) ;
- la classe de problèmes considérés peut dépendre de plusieurs paramètres. Dans ce cas il faut choisir un chemin dans l'espace des paramètres, reliant le problème initial au problème plus simple à résoudre.

En conclusion, on utilisera plutôt une méthode directe si

- on n'a pas besoin d'une grande précision de calcul ;
- la dimension d'espace est assez petite ;
- on n'a aucune idée a priori de la trajectoire optimale recherchée, par exemple on ne sait rien sur la structure des commutations ;
- on veut introduire facilement des contraintes sur l'état.

On utilisera plutôt une méthode indirecte

- si la dimension d'espace est assez grande ;
- si on a besoin de calculer la trajectoire optimale de manière très précise ;
- dans un deuxième temps, après avoir appliqué une méthode directe qui a donné une première approximation de la solution optimale.

Cependant, pour des problèmes de contrôle optimal où le système de contrôle est *raide* (en anglais *stiff system*), en aéronautique notamment, l'application d'une méthode directe peut s'avérer insuffisante pour obtenir une bonne approximation de la solution optimale et du vecteur adjoint, *i.e.* cette approximation ne constitue pas une condition initiale assez précise pour faire converger la méthode de tir. En effet, le problème aux valeurs limites sous-jacent est mal conditionné, si bien que le domaine de convergence de la méthode est très petit,

inaccessible par une méthode directe. Dans ce cas, on peut avoir recours à une méthode d'homotopie pour tenter de faire converger la méthode de tir.

Par ailleurs, pour des problèmes de contrôle optimal complexes, comme par exemple des mission interstellaires, où la trajectoire optimale a une structure très compliquée, une précision extrême est requise pour calculer numériquement cette trajectoire. Dans ce cas l'application d'une méthode de tir s'avère indispensable.

En revanche, si le système extrémal est lui-même très complexe (c'est le cas dans certains problèmes de contrôle optimal de robots industriels, où l'écriture des équations adjointes peut requérir des centaines, voire des milliers de lignes), une méthode directe peut être préférable à la méthode de tir, avec toutefois les limitations précédemment décrites.

Il existe beaucoup d'autres méthodes pour contrebalancer les inconvénients respectifs des méthodes directes ou indirectes. Nous n'évoquons pas ici par exemple les techniques parfois pointues de contrôle optimal géométrique (voir par exemple [2, 13] pour le contrôle géométrique) qui peuvent améliorer notablement le champ d'applications des méthodes indirectes, comme cela a été montré de manière rapide dans le chapitre 7.4 sur le contrôle optimal de la rentrée atmosphérique d'une navette spatiale. Pour les méthodes directes, nous renvoyons le lecteur à [11].

Nous donnons ci-dessous un exemple très simple de mise en oeuvre numérique codé en *Matlab*, en utilisant des routines qui, attention, sont connues pour ne pas être efficaces (mais elles sont en revanche simples d'utilisation, dans un premier temps). Il faut noter que, dans la pratique, on utilise des outils plus évolués et efficaces, et pour gagner en vitesse d'exécution il est préférable de coder (dans un second temps) dans un langage compilé comme *Fortran* ou *C*, en utilisant des routines expertes qu'on peut trouver sur le web. Notons enfin qu'une manière simple mais efficace de coder des méthodes directes est l'utilisation du langage de modélisation mathématique *AMPL* combiné à une routine d'optimisation non linéaire comme *IPOPT*. Les calculs peuvent même être lancés en ligne sur le site web *NEOS Solvers*. On trouve sur le web de nombreux softwares "tout-en-un", qui implémentent des méthodes directes ou bien des méthodes indirectes. Faire une telle liste serait fastidieux, et une telle liste évolue sans arrêt. Le lecteur intéressé trouvera facilement sur le web ou pourra demander conseil à des experts.

Exemple 9.3.1. Comparons les méthodes décrites sur un exemple simple. Considérons le problème du temps minimal pour le système de contrôle

$$\begin{aligned} \dot{x}(t) &= y(t),\ x(0) = 0, \\ \dot{y}(t) &= u(t),\ y(0) = 0, \end{aligned} \tag{9.9}$$

avec $|u(t)| \leqslant 1$.

Solution exacte.
Commençons par calculer la solution exacte de ce problème. Le Hamiltonien est $H = p_x y + p_y u + p^0$, et les équations adjointes sont

$$\dot{p}_x = 0,\ \dot{p}_y = -p_x.$$

On en déduit que $p_x(t) = \text{cste} = \lambda$, et donc $p_y(t) = -\lambda t + \mu$. Par ailleurs la condition de maximum du principe du maximum donne $u(t) = \text{signe}(p_y(t))$. En particulier les contrôles extrémaux ont au plus une commutation.

Décrivons la trajectoire obtenue en prenant $u(t) = 1$ sur $[0, t_1[$, puis $u(t) = -1$ sur $]t_1, T]$. D'après les équations (9.9), on obtient

– si $0 \leqslant t \leqslant t_1$, alors $x(t) = \frac{t^2}{2}$ et $y(t) = t$;

– si $t_1 \leqslant t \leqslant T$, alors $x(t) = -\frac{t^2}{2} + 2t_1 t - t_1^2$ et $y(t) = -t + 2t_1$.

Les trajectoires obtenues en prenant d'abord $u = -1$, puis $u = 1$, sont les symétriques des précédentes par rapport à l'origine (voir figure 9.1).

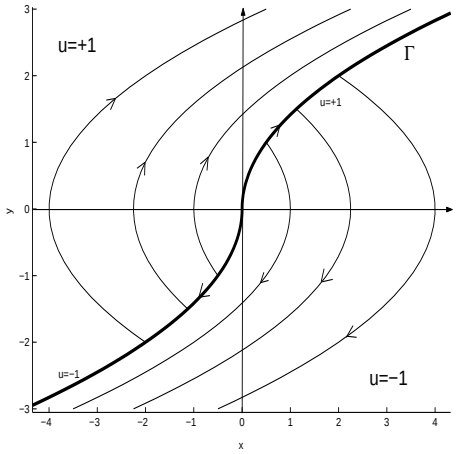


FIGURE 9.1 – Synthèse optimale

Il est clair que la courbe Γ définie par

$$\Gamma = \{(x, y) \in \mathbb{R}^2 \mid x = \frac{y^2}{2}\text{signe}(y)\}$$

est la courbe de commutation. Plus précisément, le contrôle temps minimal est donné par

– si $x > \frac{y^2}{2}\text{signe}(y)$ ou si $x = \frac{y^2}{2}$, alors $u(x, y) = +1$;

– si $x < \frac{y^2}{2}\text{signe}(y)$ ou si $x = -\frac{y^2}{2}$, alors $u(x, y) = -1$.

Calculons la fonction temps minimal $T(x, y)$ pour aller de $(0, 0)$ à (x, y). Supposons que le point (x, y) est en dessous de la courbe Γ. Ce point est atteint par la succession d'un arc $u = +1$, puis $u = -1$. On en déduit qu'il existe un unique couple (t_1, T) tel que $x = -\frac{T^2}{2} + 2t_1 T - t_1^2$, et $y = -T + 2t_1$. La résolution de ce système conduit facilement à $T = 2\sqrt{x + \frac{1}{2}y^2} - y$. De même, si le point (x, y) est au-dessus de la courbe Γ, on calcule $T = 2\sqrt{x + \frac{1}{2}y^2} + y$. Enfin, le long de la courbe Γ, on a clairement $T = |y|$. Finalement, la fonction temps minimal

est donnée par la formule

$$T(x,y) = \begin{cases} 2\sqrt{x+\frac{1}{2}y^2} - y & \text{si } \ x \geqslant \frac{y^2}{2}\text{signe}(y), \\ 2\sqrt{-x+\frac{1}{2}y^2} + y & \text{si } \ x < \frac{y^2}{2}\text{signe}(y). \end{cases} \tag{9.10}$$

Remarque 9.3.4. Notons que la fonction temps minimal (9.10) vérifie bien l'équation d'Hamilton-Jacobi associée au problème de contrôle optimal (9.9)

$$y\frac{\partial T}{\partial x} + \left|\frac{\partial T}{\partial y}\right| = 1, \quad T(0,0) = 0.$$

Les ensembles de niveau de la fonction temps minimal, *i.e.* les ensembles $T^{-1}(r) = \{(x,y) \in \mathbb{R}^2 \mid T(x,y) = r\}$, où $r > 0$, sont représentés sur la figure 9.2. Notons que $T^{-1}(r)$ est aussi le bord de l'ensemble accessible en temps r.

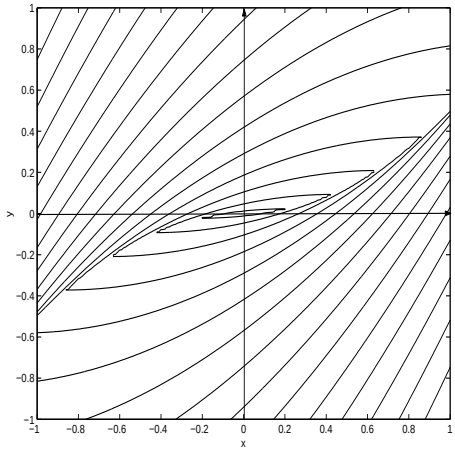


FIGURE 9.2 – Ensembles de niveau de la fonction temps minimal

Sur cet exemple, nous proposons trois méthodes numériques. Tout d'abord, nous résolvons numériquement l'équation d'Hamilton-Jacobi. Ensuite, nous mettons en oeuvre une méthode directe, puis une méthode indirecte, pour aller en temps minimal de $(0,0)$ à $(0,-1)$. Par commodité les programmes sont effectués sous *Matlab*, en utilisant certaines routines spécifiques.

Résolution numérique de l'équation d'Hamilton-Jacobi.
Notons h_x (resp. h_y) le pas de discrétisation en x (resp. en y). On discrétise de la maniere suivante :

– si $y(j) < 0$ alors

$$y(j)\frac{T(i+1,j) - T(i,j)}{h_x} + \max(0, \frac{T(i,j) - T(i,j+1)}{h_y}, \frac{T(i,j) - T(i,j-1)}{h_y}) = 1.$$

– si $y(j) > 0$ alors

$$y(j)\frac{T(i,j)-T(i-1,j)}{h_x}+\max(0,\frac{T(i,j)-T(i,j+1)}{h_y},\frac{T(i,j)-T(i,j-1)}{h_y})=1.$$

Notons que, en posant $m=\min(T(i,j-1),T(i,j+1))$, on a :

$$\max(0,\frac{T-T(i,j+1)}{h_y},\frac{T-T(i,j-1)}{h_y}=\left\{\begin{array}{ll} 0 & \text{si} \quad T<m, \\ \frac{T-m}{h_y} & \text{si} \quad T>m. \end{array}\right.$$

Avec des pas $h_x = h_y = 0.01$, et 200 itérations, on obtient le résultat de la figure 9.3.
Le programme est le suivant.

```
function hjb2

%% Discretisation de l'equation d'HJB :
%%          y dS/dx + |dS/dy| = 1, S(0,0)=0.
%% On discretise de la maniere suivante :
%% si y(j)<0 : y(j)*(S(i+1,j)-S(i,j))/hx
%%             + max(0,(S(i,j)-S(i,j+1))/hy,(S(i,j)-S(i,j-1))/hy) = 1
%% si y(j)>0 : y(j)*(S(i,j)-S(i-1,j))/hx
%%             + max(0,(S(i,j)-S(i,j+1))/hy,(S(i,j)-S(i,j-1))/hy) = 1
%% et on remarque que, si m=min(S(i,j-1),S(i,j+1)) :
%%  max(0,(S-S(i,j+1))/hy,(S-S(i,j-1))/hy) = 0          si S < m
%%                                         = (S-m)/hy   si S > m

clear all ; close all ; clc ;

hx = 0.01 ; hy = 0.01 ; big=1e6 ; Nit=200 ;
xmin = -1 ; xmax = 1 ; ymin = -1 ; ymax = 1 ;

x = [ xmin : hx : xmax ] ; y = [ ymin : hy : ymax ] ;
S = ones(length(x),length(y))*big ; Snew = S ;
i0 = find(x==0) ; j0 = find(y==0) ;
S(i0,j0) = 0 ;

for it=1:Nit
  for i=2:length(x)-1
    for j=2:length(y)-1
      m = min( S(i,j-1) , S(i,j+1) ) ;
        if y(j)>0
          if y(j)*(m-S(i-1,j)) > hx
            Snew(i,j) = S(i-1,j)+hx/y(j) ;
          else
            Snew(i,j) = (hx*hy+hx*m+hy*y(j)*S(i-1,j))/(hx+hy*y(j)) ;
          end
```

```
        else
          if y(j)*(S(i+1,j)-m) > hx
            Snew(i,j) = S(i+1,j)-hx/y(j) ;
          else
            Snew(i,j) = (hx*hy+hx*m-hy*y(j)*S(i+1,j))/(hx-hy*y(j)) ;
          end
        end
      end
  end
  S = Snew ;
  S(i0,j0) = 0 ;
end

[X,Y] = meshgrid(x,y) ; Nx = size(X,1)-2 ; Ny = size(X,2)-2 ;
figure ; contour(S(Nx/4:3*Nx/4,Ny/4:3*Ny/4),[0:0.1:1]) ;
```

Les résultats sont donnés sur la figure 9.3.

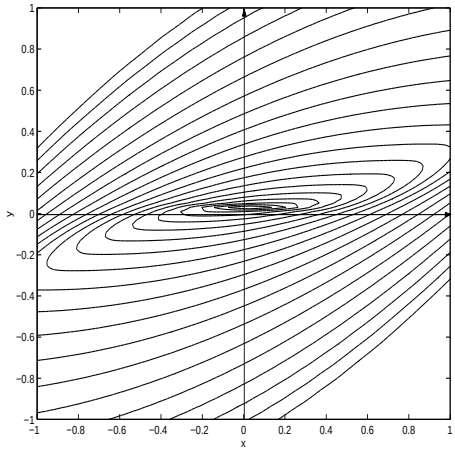


FIGURE 9.3 – Ensembles de niveau de la fonction temps minimal

Le problème est la lenteur de la convergence de cet algorithme. Ici, la convergence est en $\max(h_x, h_y)^{1/4}$. Les valeurs prises dans le programme précédent ne sont donc pas bonnes, et il faut prendre des valeurs h_x, h_y beaucoup plus petites. Mais alors le temps d'exécution du programme est très long. Il faut donc absolument avoir recours à un langage de programmation compilé comme le C++ pour implémenter cet algorithme, et avoir ainsi un temps d'exécution raisonnable. Ceci a été effectué, les résultats sont sur la figure 9.4. Notons qu'il faudrait en fait, vu la lenteur de la convergence, prendre $h_x = h_y = 10^{-5}$, mais dans ce cas même en C++ l'algorithme n'est pas efficace. Il faut donc envisager un algorithme plus fin, ou une méthode de front d'onde (voir [63]).

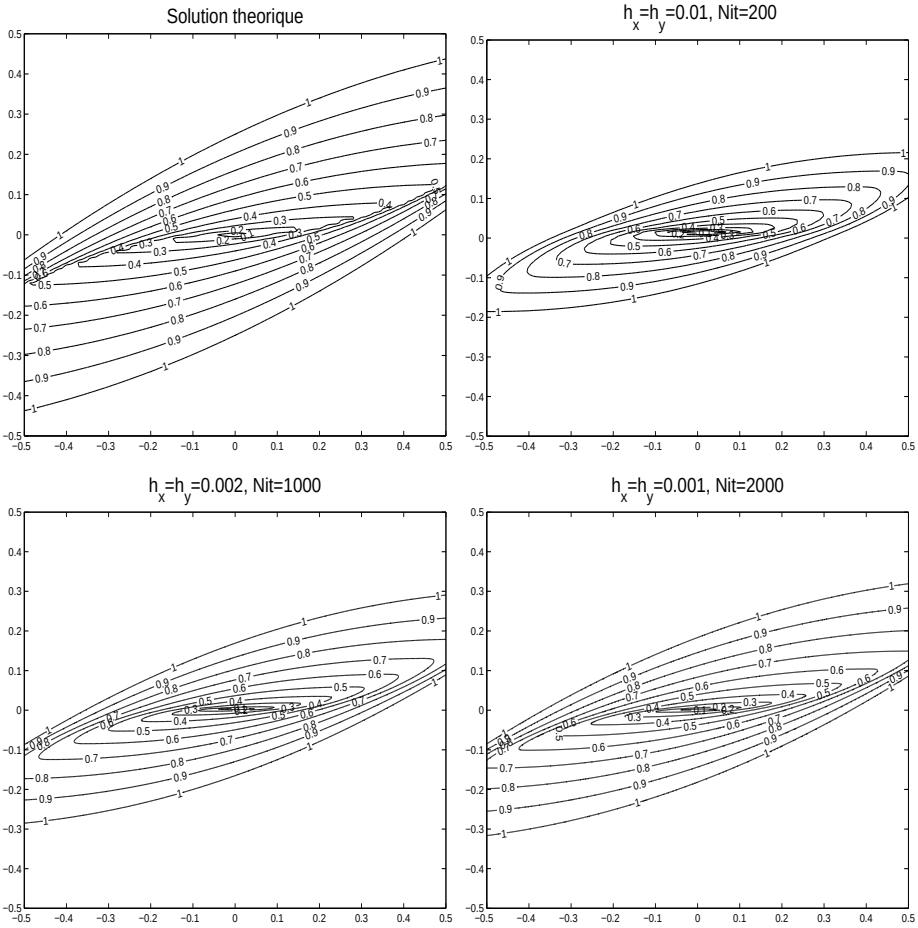


FIGURE 9.4 – Comparaison des résultats numériques

Mise en oeuvre d'une méthode directe. On se ramène à un problème de programmation non linéaire. Un tel problème se résout à l'aide d'une méthode SQP. Cet algorithme est implémenté dans la ToolBox *optim* de *Matlab*, il s'agit de la routine *fmincon.m*, que l'on utilise ici.

```
function direct

%% Discretisation directe (en utilisant fmincon.m)
%% du probleme de temps minimal
%% xdot=y, ydot=u, |u|<=1,
%% le probleme etant de joindre (0,0) a (0,-1) en temps minimal.

clear all ; close all ; clc ;
```

```
N = 20 ; % nombre de pas de discretisation
uinit = 2*rand(N,1)-1 ; % initialisation aleatoire du controle
tfinit = 1 ; xinit = [uinit ; tfinit] ;
%          point de depart pour fmincon
lb = -ones(N+1,1) ; lb(N+1) = 0 ; ub = ones(N+1,1) ; ub(N+1) = 20 ;
% contrainte sur le controle |u| <= 1, et  0 <= tf <= 20

[rep,Fval,exitflag] = fmincon(@tempsfinal,xinit,[],[],[],[],lb,ub,@cond) ;
exitflag

tf = rep(end) ; x(1)=0 ; y(1) = 0 ;
for i=1:N
    x(i+1) = x(i) + tf/N*y(i) ;
    y(i+1) = y(i) + tf/N*rep(i) ;
end      %  calcul de la trajectoire optimale

subplot(121) ; plot(x,y) ; axis square ; title('Trajectoire') ;
subplot(122) ; plot(linspace(0,tf,N),rep(1:N)) ;
axis square ; title('Controle') ;

%-----------------------------------------------------------------

function [c,ceq] = cond(x)

N = length(x)-1 ;
c = 0 ;
tf = x(end) ; xf = 0 ; yf = 0 ;
for i=1:N
    xf = xf + tf/N*yf ;   %  calcul du point final au temps tf
    yf = yf + tf/N*x(i) ; %  avec la methode d'Euler explicite
end
ceq = [ xf ; yf+1 ] ;     % on impose la condition finale xf=0, yf=-1

%-----------------------------------------------------------------

function val = tempsfinal(x)

val = x(end) ;  % x=[u;tf], ou u est le discretise du controle,
                %            et tf est le temps final
```

Les résultats sont tracés sur la figure 9.5.

Mise en oeuvre d'une méthode indirecte. On se ramène ici à un problème de tir simple. On utilise une méthode de Newton, implémentée dans la ToolBox *optim* de *Matlab*, à savoir la routine *fsolve.m*.

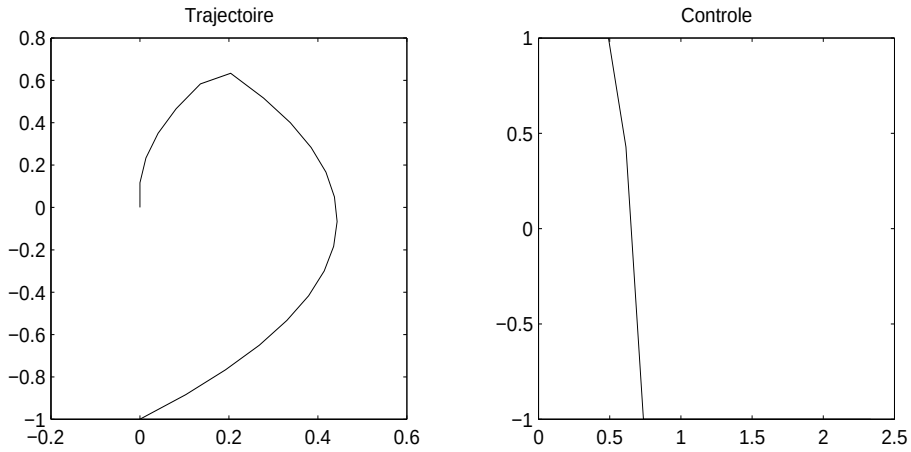


FIGURE 9.5 – Résultats de la méthode directe

```
function tirsimple

% Methode de tir simple, en utilisant fsolve.m,
% pour le systeme de controle
%         xdot=y, ydot=u,  |u|<=1.
% On veut aller de (0,0) \‘a (0,-1) en temps minimal.

clear all ; clf ; clc ; format long ;

global x0 ; x0=[0;0] ;
P0=[1;1] ; tf=5 ;

% Calcul de P0,tf
options=optimset('Display','iter','LargeScale','on');
[P0tf,FVAL,EXITFLAG]=fsolve(@F,[P0;tf],options);

EXITFLAG % 1 si la methode converge, -1 sinon

% Trace de la trajectoire optimale
options = odeset('AbsTol',1e-9,'RelTol',1e-9) ;
[t,z] = ode45(@sys,[0;P0tf(3)],[x0;P0tf(1);P0tf(2)],options) ;
subplot(121) ; plot(z(:,1),z(:,2)) ;
  axis square ; title('Trajectoire') ;
subplot(122) ; plot(t,sign(z(:,4))) ;
  axis square ; title('Controle') ;

%-------------------------------------------------------------
```

```
function Xzero=F(X)
% Definition de la fonction dont on cherche un zero

global x0 ;
options = odeset('AbsTol',1e-9,'RelTol',1e-9) ;
[t,z] = ode113(@sys,[0;X(3)],[x0;X(1);X(2)],options) ;
HamEnd = z(end,3)*z(end,2)+abs(z(end,4))-1 ;
Xzero = [ z(end,1)                  %  on impose xf=0
          z(end,2)+1                %  on impose yf=-1
              HamEnd   ] ;            %  tf libre donc H(tf)=0

%-------------------------------------------------------------

function zdot=sys(t,z)

u=sign(z(4)) ;
zdot = [ z(2)
         u
         0
         -z(3)   ] ;  % systeme extremal
```

Les résultats sont tracés sur la figure 9.6.

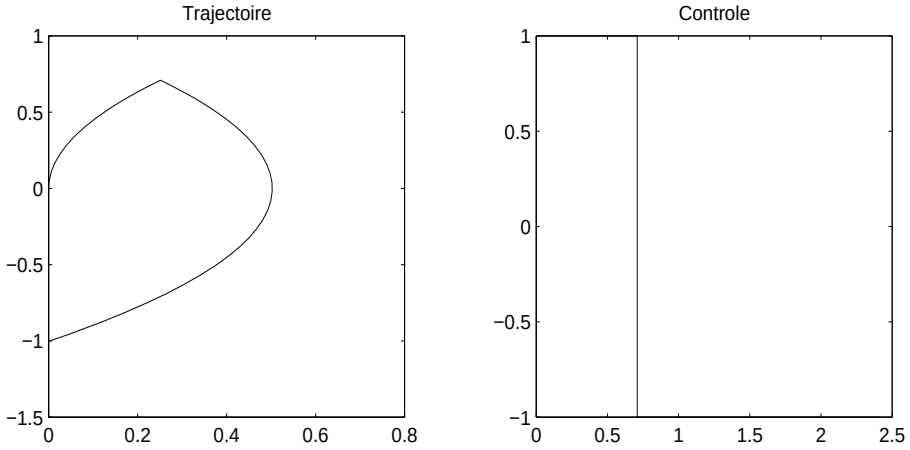


FIGURE 9.6 – Résultats de la méthode indirecte

Troisième partie

Annexe

Chapitre 10

Rappels d'algèbre linéaire

10.1 Exponentielle de matrice

Soit $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$, et soit $\|\cdot\|$ une norme multiplicative sur $\mathcal{M}_n(\mathbb{K})$ (*i.e.* $\|AB\| \leqslant \|A\|\,\|B\|$ pour toutes matrices $A, B \in \mathcal{M}_n(\mathbb{K})$; par exemple les normes d'opérateurs sont multiplicatives).

Définition 10.1.1. Soit $A \in \mathcal{M}_n(\mathbb{K})$. On définit l'exponentielle de la matrice A par

$$\exp(A) = e^A = \sum_{k=1}^{+\infty} \frac{A^k}{k!}.$$

C'est une série normalement convergente dans le Banach $\mathcal{M}_n(\mathbb{K})$, vu que

$$\left\| \sum_{k=p}^{q} \frac{A^k}{k!} \right\| \leqslant \sum_{k=p}^{q} \left\| \frac{A^k}{k!} \right\| \leqslant \sum_{k=p}^{q} \frac{\|A\|^k}{k!} \leqslant e^{\|A\|}.$$

Proposition 10.1.1.
- *Pour tout $A \in \mathcal{M}_n(\mathbb{K})$, on a $e^A \in GL_n(\mathbb{K})$, et $(e^A)^{-1} = e^{-A}$.*
- *L'application exponentielle est $\mathbb{K}$-analytique (et donc en particulier est de classe C^∞ sur le corps $\mathbb{K}$).*
- *La différentielle de Fréchet $d\exp(0)$ de l'application exponentielle en 0 est égale à l'identité sur $\mathcal{M}_n(\mathbb{K})$.*
- *Pour toutes matrices $A, B \in \mathcal{M}_n(\mathbb{K})$ qui commutent, i.e. $AB = BA$, on a*
$$e^{A+B} = e^A e^B.$$
- *Si $P \in GL_n(\mathbb{K})$, alors $Pe^A P^{-1} = e^{PAP^{-1}}$.*
- *Pour $A \in \mathcal{M}_n(\mathbb{K})$, l'application $f(t) = e^{tA}$ est dérivable, et $f'(t) = Ae^{tA} = e^{tA}A$.*

10.2 Réduction des endomorphismes

L'espace vectoriel $\mathcal{M}_n(\mathbb{K})$ est de dimension n^2 sur $\mathbb{K}$, donc les éléments $I, A, \ldots, A^{n^2}$ sont linéairement dépendants. Par conséquent il existe des polynômes P annulateurs de A, *i.e.* tels que $P(A) = 0$. L'anneau $\mathbb{K}[X]$ étant principal, l'idéal des polynômes annulateurs de A admet un unique générateur normalisé, *i.e.* un unique polynôme de plus petit degré, dont le coefficient dominant est égal à 1, annulant A ; on l'appelle *polynôme minimal* de la matrice A, noté π_A.

Par ailleurs, le *polynôme caractéristique* de A, noté χ_A, est défini par

$$\chi_A(X) = \det\,(XI - A).$$

Théorème 10.2.1 (Théorème de Hamilton-Cayley)**.** $\chi_A(A) = 0$.

En particulier, le polynôme minimal π_A divise le polynôme caractéristique χ_A. Notons que $\deg \chi_A = n$ et $\deg \pi_A \leqslant n$.

Exemple 10.2.1. Pour une matrice $N \in \mathcal{M}_n(\mathbb{K})$ nilpotente, *i.e.* il existe un entier $p \geqslant 1$ tel que $N^p = 0$, on a nécessairement $p \leqslant n$, $\pi_N(X) = X^p$ et $\chi_N(X) = X^n$.

Exemple 10.2.2. Pour une *matrice compagnon*, *i.e.* une matrice de la forme

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots 0 & 1 & \\ -a_n & -a_{n-1} & \cdots & -a_2 & -a_1 \end{pmatrix},$$

on a

$$\pi_A(X) = \chi_A(X) = X^n + a_1 X^{n-1} + \cdots + a_{n-1}X + a_n.$$

Le scalaire $\lambda \in \mathbb{K}$ est dit *valeur propre* s'il existe un vecteur non nul $v \in \mathbb{K}^n$, appelé *vecteur propre*, tel que $Av = \lambda v$. L'*espace propre* associé à la valeur propre λ est défini par

$$E(\lambda) = \ker(A - \lambda I)\ ;$$

c'est l'ensemble des vecteurs propres de A pour la valeur propre λ.

Lorsque $\mathbb{K} = \mathbb{C}$, les valeurs propres de A sont exactement les racines du polynôme caractéristique χ_A. En particulier on a

$$\chi_A(X) = \prod_{i=1}^{r} (X - \lambda_i)^{m_i} \quad \text{et} \quad \pi_A(X) = \prod_{i=1}^{r} (X - \lambda_i)^{s_i},$$

avec $s_i \leqslant m_i$. L'entier s_i (resp. m_i) est appelé *ordre de nilpotence* (resp. *multiplicité*) de la valeur propre λ_i. L'*espace caractéristique* de la valeur propre λ_i est défini par

$$N(\lambda_i) = \ker(X - \lambda_i)^{s_i}.$$

Théorème 10.2.2 (Théorème de décomposition des noyaux)**.** *Soient $A \in \mathcal{M}_n(\mathbb{K})$ et $P \in \mathbb{K}[X]$ un polynôme tel que*

$$P(X) = \prod_{i=1}^{r} P_i(X),$$

où les polynômes P_i sont premiers entre eux deux à deux. Alors

$$\ker P(A) = \bigoplus_{i=1}^{r} \ker P_i(A).$$

De plus, chaque sous-espace $\ker P_i(A)$ est invariant par A, et la projection p_i sur $\ker P_i(A)$ parallèlement à $\bigoplus_{j\neq i} \ker P_j(A)$ est un polynôme en A.

En appliquant ce théorème au polynôme minimal de A, on obtient, lorsque $\mathbb{K} = \mathbb{C}$,

$$\mathbb{C}^n = \bigoplus_{i=1}^{r} N(\lambda_i).$$

Notons que $N(\lambda_i) = \ker(X-\lambda_i)^{s_i} = \ker(X-\lambda_i)^{m_i}$.

La restriction de A à $N(\lambda_i)$ est de la forme $\lambda_i I + N_i$, où N_i est une matrice nilpotente d'ordre s_i. On peut alors montrer que toute matrice $A \in \mathcal{M}_n(\mathbb{K})$ admet une unique décomposition $A = D + N$, où D est diagonalisable sur $\mathbb{C}$, N est nilpotente, et de plus $DN = ND$ (décomposition $D+N$).

On peut préciser ce résultat avec la théorie de Jordan.

Théorème 10.2.3 (Décomposition de Jordan)**.** *Soit $A \in \mathcal{M}_n(\mathbb{K})$; on suppose que π_A est scindé sur $\mathbb{K}$ (ce qui est toujours le cas sur $\mathbb{C}$) tel que $\pi_A(X) = \prod_{i=1}^{r}(X-\lambda_i)^{s_i}$. Alors il existe $P \in GL_n(\mathbb{K})$ telle que*

$$P^{-1}AP = \begin{pmatrix} A_1 & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & A_r \end{pmatrix},$$

où les matrices A_i sont diagonales par blocs

$$A_i = \begin{pmatrix} J_{i,1} & & 0 \\ \vdots & \ddots & \vdots \\ 0 & & J_{i,e_i} \end{pmatrix},$$

et où les matrices $J_{i,k}$, $1 \leqslant i \leqslant r$, $1 \leqslant k \leqslant e_i$, sont des blocs de Jordan, i.e. des matrices carrées de la forme

$$J_{i,k} = \begin{pmatrix} \lambda_i & 1 & & 0 \\ 0 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix},$$

n'ayant pas forcément toutes le même ordre $|J_{i,k}|$. *Pour tout* $i \in \{1, \ldots, r\}$, *on a* $e_i = \dim E(\lambda_i)$, *et* $\max_{1 \leqslant k \leqslant e_i} |J_{i,k}| = s_i$.

Chapitre 11

Théorème de Cauchy-Lipschitz

Dans cette section nous rappelons une version générale du théorème de Cauchy-Lipschitz, adaptée à la théorie du contrôle, qui établit sous certaines conditions l'existence et l'unicité d'une solution d'une équation différentielle. Une bonne référence à ce sujet est [64, Appendix C].

11.1 Un énoncé général

Soit I un intervalle de $\mathbb{R}$ et V un ouvert de $\mathbb{R}^n$. Considérons le problème de Cauchy

$$\dot{x}(t) = f(t, x(t)),\ x(t_0) = x_0, \tag{11.1}$$

où f est une application de $I \times V$ dans $\mathbb{R}^n$, et $x_0 \in V$. Le théorème de Cauchy-Lipschitz usuel affirme l'existence et l'unicité d'une solution maximale pourvu que f soit continue, et localement lipschitzienne par rapport à x. Mais en théorie du contrôle ces hypothèses doivent être affaiblies car on est amené à considérer des contrôles non continus (au mieux, continus par morceaux), et par conséquent la continuité du second membre n'est plus assurée. En particulier la solution, si elle existe, n'est pas en général dérivable partout et il faut donc redéfinir de manière adéquate le concept de solution.

Définition 11.1.1. On suppose que pour tout $t \in I$ la fonction $x \mapsto f(t, x)$ est mesurable, et que pour tout $x \in U$ la fonction $t \mapsto f(t, x)$ est continue. On appelle solution du problème de Cauchy (11.1) tout couple $(J, x(\cdot))$, où J est un intervalle tel que $J \subset I$ et $t_0 \in J$, et où $x(\cdot)$ est une fonction absolument continue de J dans V telle que, pour tout $t \in J$,

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds,$$

ce qui est équivalent à

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)) \text{ p.p. sur } J, \\ x(t_0) &= x_0. \end{aligned}$$

Une solution $(J, x(\cdot))$ est dite *maximale* si, pour toute autre solution $(\bar{J}, \bar{x}(\cdot))$, on a $\bar{J} \subset J$ et $x(\cdot) = \bar{x}(\cdot)$ sur $\bar{J}$.

On a alors le théorème suivant.

Théorème 11.1.1 (Théorème de Cauchy-Lipschitz)**.** *On suppose que la fonction $f : I \times V \to V$ vérifie les deux hypothèses suivantes :*

1. *f est localement lipschitzienne par rapport à x au sens suivant :*

$$\begin{aligned} &\forall x \in V \quad \exists r > 0,\ B(x, r) \subset V, \quad \exists \alpha \in L^1_{loc}(I, \mathbb{R}^+) \\ &\forall t \in I \quad \forall y, z \in B(x, r) \quad \|f(t, y) - f(t, z)\| \leqslant \alpha(t)\|y - z\|, \end{aligned}$$

2. *f est localement intégrable par rapport à t, i.e.*

$$\forall x \in V \quad \exists \beta \in L^1_{loc}(I, \mathbb{R}^+) \quad \forall t \in I \quad \|f(t, x)\| \leqslant \beta(t).$$

Alors pour toute donnée initiale $(t_0, x_0) \in I \times V$, il existe une unique solution maximale $(J, x(\cdot))$ du problème de Cauchy (11.1).

Remarque 11.1.1. On n'a pas forcément $J = I$; par exemple considérons le problème de Cauchy $\dot{x}(t) = x(t)^2$, $x(0) = x_0$. Alors
- si $x_0 = 0$, on a $J = \mathbb{R}$ et $x(\cdot) \equiv 0$;
- si $x_0 > 0$, on a $J =]-\infty, 1/x_0[$ et $x(t) = x_0/(1 - x_0 t)$;
- si $x_0 < 0$, on a $J =]1/x_0, +\infty[$ et $x(t) = x_0/(1 - x_0 t)$.

Remarque 11.1.2. Si f est seulement continue on n'a pas unicité en général ; par exemple considérons le problème de Cauchy $\dot{x}(t) = \sqrt{|x(t)|}$, $x(0) = 0$. La fonction nulle est solution, ainsi que

$$x(t) = \begin{cases} 0 & \text{si } t \leqslant 0, \\ t^2/4 & \text{si } t > 0. \end{cases}$$

Théorème 11.1.2 (Théorème d'explosion)**.** *Sous les hypothèses du théorème de Cauchy-Lipschitz, soit $(]a, b[, x(\cdot))$ une solution maximale. Si $b < \sup I$, alors pour tout compact K contenu dans V, il existe un réel $\eta > 0$ tel que $x(t) \notin K$, pour tout $t \in]b - \eta, b[$.*

Remarque 11.1.3. En particulier si $V = \mathbb{R}^n$, alors $\|x(t)\| \underset{\substack{t \to b \\ t < b}}{\longrightarrow} +\infty$.

Remarque 11.1.4. On a une propriété semblable si $a > \inf I$.

Enonçons maintenant une version globale du théorème de Cauchy-Lipschitz.

Théorème 11.1.3. *Sous les hypothèses du théorème de Cauchy-Lipschitz, on suppose de plus que $V = \mathbb{R}^n$ et que f est globalement lipschitzienne par rapport à x, i.e.*

$$\exists \alpha \in L^1_{loc}(I, \mathbb{R}^+) \mid \forall t \in I \quad \forall y, z \in \mathbb{R}^n \quad \|f(t,y) - f(t,z)\| \leqslant \alpha(t)\|y - z\|.$$

Alors $J = I$.

Exercice 11.1.1 (Lemme de Gronwall)**.** Soient ψ et $y : [t_0, t_1] \to \mathbb{R}^+$ deux fonctions continues vérifiant

$$\exists c \geqslant 0 \ / \ \forall t \in [t_0, t_1] \quad y(t) \leqslant c + \int_{t_0}^{t} \psi(s)y(s)ds.$$

Montrer que

$$\forall t \in [t_0, t_1] \quad y(t) \leqslant c \exp\left(\int_{t_0}^{t} \psi(s)ds\right).$$

Indication : poser $F(t) = \displaystyle\int_{t_0}^{t} \psi(s)y(s)ds$, puis $G(t) = F(t)\exp\left(-\displaystyle\int_{t_0}^{t} \psi(s)ds\right)$.

Exercice 11.1.2. 1. Soit $y_0 \in \mathbb{R}$. Justifier qu'il existe une solution maximale $y(.)$ sur un intervalle $]a, b[$ du problème de Cauchy

$$(E) \qquad y'(t) = -y(t) + \frac{y(t)^4}{1+t^2}, \ y(0) = y_0.$$

2. Montrer que pout tout $t \in]a, b[$:

$$y(t) = e^{-t}y_0 + \int_0^t e^{s-t}\frac{y(s)^4}{1+s^2}ds$$

3. Soit T tel que $0 < T < b$. Supposons que pour tout $t \in [0, T]$ on ait $|y(t)| \leqslant 1$. A l'aide du lemme de Gronwall, montrer que

$$\forall t \in [0, T] \quad |y(t)| \leqslant |y_0|.$$

4. Supposons que $|y_0| < 1$. Montrer que $|y(t)| \leqslant 1$ pour tout $t \in [0, b[$, puis que $b = +\infty$.

   Indication : utiliser le lemme de Gronwall, et le théorème d'explosion.

Exercice 11.1.3. Soit $q : \mathbb{R} \to \mathbb{R}$ une fonction de classe C^1, strictement positive et croissante. Montrer l'existence et l'unicité d'une solution maximale définie sur un intervalle contenant $[0, +\infty[$, pour le problème de Cauchy $y''(t) + q(t)y(t) = 0$, $y(0) = y_0$, $y'(0) = y'_0$, puis que toutes les solutions de l'équation différentielle $y''(t) + q(t)y(t) = 0$ sont bornées sur $\mathbb{R}^+$.

Indication : dériver la fonction $V(t) = y(t)^2 + \frac{y'(t)^2}{q(t)}$.

Exercice 11.1.4 (Méthode des entonnoirs)**.** Soit (E) : $x'(t) = f(t, x(t))$ une équation différentielle, où $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est de classe C^1.

On dit que α (resp. β) est une *barrière inférieure* (resp. une *barrière supérieure*), si c'est une fonction de classe C^1 telle que pour tout $t \in \mathbb{R}$ on ait $\alpha'(t) < f(t, \alpha(t))$ (resp. $\beta'(t) > f(t, \beta(t))$). On appelle *entonnoir* l'ensemble

$$\mathcal{E} = \{(t, y) \in \mathbb{R} \times \mathbb{R} \mid \alpha(t) < y < \beta(t)\}.$$

Montrer que si $t \mapsto x(t)$ est une solution de (E) sur un intervalle $J \subset \mathbb{R}$ telle que $x(t_0) = x_0$ avec $(t_0, x_0) \in \mathcal{E}$, alors $(t, x(t)) \in \mathcal{E}$ pour tout $t \geqslant t_0, t \in J$.

Indication : raisonner par l'absurde, et poser $t_1 = \inf\{t \geqslant t_0 \,/\, (t, x(t)) \notin \mathcal{E}\}$. Montrer que $x(t_1) = \alpha(t_1)$ ou $\beta(t_1)$, et conclure.

Exercice 11.1.5 (Loi de Hubble)**.** Une des théories actuelles de l'univers (*théorie du big-bang*) admet que l'origine de l'univers est une gigantesque explosion à partir de laquelle la matière de l'univers a commencé à diverger à partir du point 0.

On considère que cette expansion est homogène et isotrope ; les positions successives se déduisent les unes des autres par une homothétie de centre 0 :

$$\overrightarrow{OM}(t) = \lambda(t, t_0)\overrightarrow{OM}(t_0).$$

1. Montrer que la vitesse du point M se met sous la forme

$$\vec{v}(M) = H(t)\overrightarrow{OM}(t) \qquad \textit{(loi de Hubble)}$$

   où on explicitera $H(t)$.
2. Montrer que la loi de Hubble est incompatible avec une valeur constante de H.
3. La valeur actuelle de H est $H \simeq 2,5.10^{-18}\ \ s^{-1}$. En prenant comme modèle $H(t) = \frac{\alpha}{t}$ (où α est à déterminer), trouver l'ordre de grandeur du rayon maximum de l'univers (on rappelle la vitesse de la lumière $c = 3.10^8\ \ m.s^{-1}$).

   En déduire l'âge de l'univers selon cette théorie.

11.2 Systèmes différentiels linéaires

Considérons le problème de Cauchy dans $\mathbb{R}^n$

$$\dot{x}(t) = A(t)x(t) + B(t),\ x(t_0) = x_0, \tag{11.2}$$

où les applications $t \mapsto A(t) \in \mathcal{M}_n(\mathbb{R})$ et $t \mapsto B(t) \in \mathbb{R}^n$ sont localement intégrables sur l'intervalle I considéré.

Définition 11.2.1. On appelle *résolvante* du problème (11.2) la solution du problème de Cauchy

$$\frac{\partial R}{\partial t}(t, t_0) = A(t)R(t, t_0),\ R(t_0, t_0) = Id,$$

où $R(t, t_0) \in \mathcal{M}_n(\mathbb{R})$.

Proposition 11.2.1. *La résolvante possède les propriétés suivantes :*
- $R(t_2, t_0) = R(t_2, t_1)R(t_1, t_0)$.
- *Si* $\Delta(t, t_0) = det\ R(t, t_0)$, *on a*

$$\frac{\partial \Delta}{\partial t}(t, t_0) = tr\ A(t).\delta(t, t_0),\ \Delta(t_0, t_0) = 1.$$

- *La solution du problème de Cauchy (11.2) est donnée par*

$$x(t) = R(t, t_0)x_0 + \int_{t_0}^{t} R(t, s)B(s)ds$$

(formule de variation de la constante).

Remarque 11.2.1. Lorsque $t_0 = 0$, on note plutôt $M(t) = R(t, 0)$. La formule de variation de la constante s'écrit alors

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

Remarque 11.2.2 (Expression de la résolvante). La résolvante admet le développement en série

$$R(b,a) = I + \int_{a\leqslant s_1\leqslant b} A(s_1)ds_1 + \int_{a\leqslant s_1\leqslant s_2\leqslant b} A(s_2)A(s_1)ds_1ds_2 + \cdots + \int_{a\leqslant s_1\leqslant\cdots\leqslant s_n\leqslant b} A(s_n)\cdots A(s_1)ds_1\cdots ds_n + \cdots .$$

De plus cette série est normalement convergente. C'est un *développement en série chronologique*.

Cas des systèmes autonomes. Considérons le problème de Cauchy dans $\mathbb{R}^n$

$$\dot{x}(t) = Ax(t),\ x(0) = x_0,$$

où $A \in \mathcal{M}_n(\mathbb{R})$. Alors, dans ce cas, la résolvante est $M : t \mapsto e^{tA}$, et la solution de ce problème est

$$x : t \mapsto e^{tA}x_0.$$

La décomposition de Jordan permet de préciser ce résultat. En effet, on a

$$e^{tA} = P\begin{pmatrix} e^{tJ_{1,e_1}} & & 0 \\ & \ddots & \\ 0 & & e^{tJ_{1,e_r}} \end{pmatrix}P^{-1}.$$

On calcule facilement

$$e^{tJ_{i,k}} = e^{t\lambda_i}\begin{pmatrix} 1 & t & \cdots & e^{t\lambda_i}\frac{t^{|J_{i,k}|-1}}{(|J_{i,k}|-1)!} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & & \ddots & e^{t\lambda_i}t \\ 0 & & 0 & e^{t\lambda_i} \end{pmatrix}.$$

On obtient donc le résultat suivant.

Proposition 11.2.2. *Toute solution du système $\dot{x}(t) = Ax(t)$ est de la forme*

$$x(t) = \sum_{\substack{1 \leqslant i \leqslant r \\ 0 \leqslant k \leqslant |J_{i,k}|}} e^{t\lambda_i} t^k v_{i,k},$$

où $v_{i,k} \in N(\lambda_i)$ (voir chapitre précédent pour la définition de l'espace caractéristique $N(\lambda_i)$).

Exercice 11.2.1. Résoudre le système différentiel

$$\left\{ \begin{array}{rcl} x'' & = & 2x - 3y, \\ y'' & = & x - 2y. \end{array} \right.$$

Indication : exprimer ce système comme un système différentiel d'ordre 1.

Exercice 11.2.2. On pose

$$A = \begin{pmatrix} 4 & 0 & 2 & 0 \\ 0 & 4 & 0 & 2 \\ 2 & 0 & 4 & 0 \\ 0 & 2 & 0 & 4 \end{pmatrix}.$$

Résoudre le système différentiel $X' = AX$. En déduire e^A.

Exercice 11.2.3. Soit $A :]0, +\infty[\to \mathcal{M}_n(\mathbb{R})$ une application continue. On considère le système différentiel $x'(t) = A(t)x(t)$ et l'on note $R(t, t_0)$ sa résolvante.

1. Montrer que la résolvante $S(t, t_0)$ du système $\dot{y}(t) = -A(t)^{\mathrm{T}} y(t)$ est $S(t, t_0) = R(t, t_0)^{\mathrm{T}}$.
2. On pose
$$A(t) = \begin{pmatrix} 2t + \frac{1}{t} & 0 & \frac{1}{t} - t \\ t - \frac{1}{t} & 3t & t - \frac{1}{t} \\ \frac{2}{t} - 2t & 0 & \frac{2}{t} + t \end{pmatrix}.$$
Montrer que $A(t)$ possède une base de vecteurs propres indépendante de t. En déduire la résolvante $R(t, t_0)$.

Exercice 11.2.4. On suppose que l'équation différentielle

$$X'(t) = AX(t) + B(t),$$

où $A \in \mathcal{M}_n(\mathbb{R})$ et $t \mapsto B(t)$ est une application continue de $\mathbb{R}^+$ dans $\mathbb{R}^n$, admet une solution X sur $\mathbb{R}^+$ qui vérifie

$$\int_0^\infty (\|X(t)\|^2 + \|B(t)\|^2) dt < +\infty.$$

Montrer que $X(t)$ tend vers 0 lorsque t tend vers $+\infty$.

Indication : montrer que $X(\cdot)$ est de Cauchy en $+\infty$.

Exercice 11.2.5. Soit A une matrice carrée réelle d'ordre n dont les valeurs propres (dans $\mathbb{C}$) sont distinctes de $2ik\pi,\ k \in \mathbb{Z}$. Soit d'autre part $B : \mathbb{R} \to \mathbb{R}^n$ une application continue et 1-périodique. Montrer que le système différentiel $x' = Ax + B(t)$ admet une et une seule solution 1-périodique.

Exercice 11.2.6. Soit $A : \mathbb{R}^+ \to \mathcal{M}_n(\mathbb{R})$ une application continue et périodique de période T. On considère le système différentiel dans $\mathbb{R}^n$

$$x'(t) = A(t)x(t),\ x(0) = x_0.$$

Soit $M(t) \in \mathcal{M}_n(\mathbb{R})$ la résolvante du système, c'est-à-dire la solution du problème de Cauchy $M'(t) = A(t)M(t),\ M(0) = Id$. Montrer pour tout $t \geqslant 0$ la relation

$$M(t+T) = M(t)M(T).$$

En déduire que l'origine est asymptotiquement stable pour le système si et seulement si les valeurs propres complexes de la matrice $M(T)$ (appelée *matrice de monodromie*) sont de module strictement plus petit que 1.

11.3 Applications en théorie du contrôle

11.3.1 Systèmes de contrôle linéaires

Considérons le système de contrôle linéaire

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(t_0) = x_0.$$

Les hypothèses du théorème 11.1.1 sont clairement vérifiées si les applications $t \mapsto A(t), B(t)u(t), r(t)$, sont localement intégrables sur l'intervalle I considéré. Supposons donc

- $A(\cdot) \in L^1_{loc}(I, \mathcal{M}_n(\mathbb{R}))$,
- $r(\cdot) \in L^1_{loc}(I, \mathbb{R}^n)$.

Par ailleurs, les hypothèses assurant l'intégrabilité locale de $B(\cdot)u(\cdot)$ dépendent de l'ensemble des contrôles considérés.

- Si $u(\cdot) \in L^\infty_{loc}(I, \mathbb{R}^m)$, alors on suppose que $B(\cdot) \in L^1_{loc}(I, \mathcal{M}_{n,m}(\mathbb{R}))$.
- Si $u(\cdot) \in L^2_{loc}(I, \mathbb{R}^m)$, alors on suppose que $B(\cdot) \in L^2_{loc}(I, \mathcal{M}_{n,m}(\mathbb{R}))$.
- De manière générale, si $u(\cdot) \in L^p_{loc}(I, \mathbb{R}^m)$, alors on suppose que $B(\cdot) \in L^q_{loc}(I, \mathcal{M}_{n,m}(\mathbb{R}))$ où $\frac{1}{p} + \frac{1}{q} = 1$.
- Si les contrôles sont des fonctions mesurables à valeurs dans un compact $\Omega \subset \mathbb{R}^m$, alors on suppose que $B(\cdot) \in L^1_{loc}(I, \mathcal{M}_{n,m}(\mathbb{R}))$.

11.3.2 Systèmes de contrôle généraux

Considérons le système de contrôle

$$\dot{x}(t) = f(t, x(t), u(t)),\ x(t_0) = x_0,$$

où f est une fonction de $I \times V \times U$, I est un intervalle de $\mathbb{R}$, V un ouvert de $\mathbb{R}^n$ et U un ouvert de $\mathbb{R}^m$.

Pour rester dans un cadre très général, il suffit de supposer que pour chaque contrôle u considéré, la fonction $F : (t, x) \mapsto f(t, x, u(t))$ vérifie les hypothèses du théorème 11.1.1. Bien entendu, en fonction de la classe de contrôles considérée, ces hypothèses peuvent être plus ou moins difficiles à vérifier.

On peut donner des hypothèses certes moins générales, mais qui suffisent dans la grande majorité des cas. Ces hypothèses sont les suivantes :

1. L'ensemble des contrôles considérés est inclus dans $L^\infty_{loc}(I, \mathbb{R}^m)$.
2. La fonction f est de classe C^1 sur $I \times V \times U$.

Il est facile de montrer qu'alors les hypothèses du théorème 11.1.1 sont vérifiées, et donc que, pour chaque contrôle fixé, il existe une unique solution maximale $(J, x(\cdot))$ du problème de Cauchy

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)) \text{ p.p. sur } J, \\ x(t_0) &= x_0. \end{aligned}$$

Chapitre 12

Modélisation d'un système de contrôle linéaire

12.1 Représentation interne des systèmes de contrôle linéaires

Considérons le système linéaire observé

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{cases} \tag{12.1}$$

où $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_{n,m}(\mathbb{R})$, et $C \in \mathcal{M}_{p,n}(\mathbb{R})$.

On appelle *représentation interne* ou *représentation d'état continue* l'expression de la sortie $y(t)$ sous la forme

$$y(t) = Ce^{tA}x(0) + Ce^{tA}\int_0^t e^{-sA}Bu(s)ds, \tag{12.2}$$

appelée en anglais *input-output relation*.

12.2 Représentation externe des systèmes de contrôle linéaires

Définition 12.2.1. La *réponse impulsionnelle* d'un système linéaire est la sortie de ce système (à conditions initiales nulles) quand on l'excite en entrée par une impulsion de Dirac.

Ici, la réponse impulsionnelle est donc la matrice

$$W(t) = \begin{cases} Ce^{tA}B & \text{si } t \geqslant 0, \\ 0 & \text{si } t < 0. \end{cases} \tag{12.3}$$

En effet, posons, pour tout $t \in [0, \varepsilon]$,

$$u(t) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1/\varepsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \frac{1}{\varepsilon} e_i,$$

et $u(t) = 0$ sinon. Alors

$$y(t) = Ce^{tA} \frac{1}{\varepsilon} \int_0^{\varepsilon} e^{-sA} B e_i ds \underset{\varepsilon \to 0}{\longrightarrow} Ce^{tA} B e_i.$$

Remarque 12.2.1. Puisque $x(0) = 0$, on a, pour $t \geqslant 0$,

$$y(t) = \int_0^t W(t-s)u(s)ds = \int_0^{+\infty} W(t-s)u(s)ds.$$

Autrement dit (on rappelle que $f * g(x) = \int_{\mathbb{R}} f(x-y)g(y)dy$), on a le résultat suivant.

Proposition 12.2.1. $\forall t \geqslant 0 \quad y(t) = (W * u)(t)$.

Cela incite à utiliser la *transformation de Laplace*, qui transforme un produit de convolution en un produit.

Définition 12.2.2. Soit $f \in L^1_{loc}([0, +\infty[, \mathbb{R})$. Il existe $a \in \mathbb{R} \cup \{\pm\infty\}$ tel que, pour tout complexe s, on ait
- si Re $s > a$ alors $\int_0^{+\infty} e^{-st}|f(t)|dt < +\infty$,
- si Re $s < a$ alors $\int_0^{+\infty} e^{-st}|f(t)|dt = +\infty$.

Pour tout complexe s tel que Re $s > a$, on définit la *transformée de Laplace* de f par

$$\mathcal{L}(f)(s) = \int_0^{+\infty} e^{-st} f(t)dt.$$

Remarque 12.2.2. La transformation de Laplace est linéaire (en faisant attention toutefois au problème du domaine de définition). De plus, on a

$$\mathcal{L}(f * g) = \mathcal{L}(f)\mathcal{L}(g).$$

Enfin, pour toute fonction f de classe C^1, on a

$$\mathcal{L}(f')(s) = s\mathcal{L}(f)(s) - f(0).$$

Posons alors $Y(s) = \mathcal{L}(y)(s)$ et $U(s) = \mathcal{L}(u)(s)$ (où, par convention, $y(t) = 0$ et $u(t) = 0$ si $t < 0$).

Définition 12.2.3. La *matrice de transfert* H est la transformée de Laplace de la matrice de réponse impulsionnelle, *i.e.*

$$H(s) = \mathcal{L}(W)(s) = \int_0^{+\infty} W(t)e^{-st}dt. \tag{12.4}$$

Proposition 12.2.2. $Y(s) = H(s)U(s)$.

Par ailleurs, en appliquant la transformation de Laplace au système (12.1), avec $x(0) = 0$ et $X(s) = \mathcal{L}(x)(s)$, on a

$$sX(s) = AX(s) + BU(s),\ Y(s) = CX(s),$$

d'où

$$Y(s) = C(sI - A)^{-1}BU(s).$$

La proposition suivante s'ensuit.

Proposition 12.2.3. $H(s) = C(sI - A)^{-1}B$.

Remarque 12.2.3. En particulier, on a $\mathcal{L}(Ce^{tA}B)(s) = C(sI - A)^{-1}B$.

Proposition 12.2.4. *Les coefficients de la matrice de transfert $H(s)$ sont des fractions rationnelles en s, avec un numérateur de degré strictement inférieur au degré du dénominateur.*

Démonstration. Il suffit de remarquer que

$$(sI - A)^{-1} = \frac{1}{\det(sI - A)}\mathrm{com}(sI - A)^{\mathrm{T}}.$$

□

Remarque 12.2.4. Si le système s'écrit

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases}$$

alors

$$H(s) = C(sI - A)^{-1}B + D.$$

Dans ce cas, il est clair que les coefficients de la matrice $H(s)$ sont des fractions rationnelles dont le numérateur et le dénominateur ont même degré.

Réciproquement, lorsqu'on dispose d'une matrice de transfert $H(s)$ pour représenter un système linéaire continu, on peut chercher à calculer un *modèle d'état* (*i.e.* déterminer des matrices A, B, C) tel que $H(s) = C(sI - A)^{-1}B$. Un tel triplet (A, B, C) est appelé *réalisation d'état continue* de la matrice de transfert $H(s)$.

Proposition 12.2.5. *La fonction de transfert (i.e. $m = p = 1$)*

$$H(s) = b_0 + \frac{b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n}$$

admet la réalisation

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & & 0 \\ 0 & 0 & 1 & & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \ddots & 1 & 0 \\ 0 & & \cdots & & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & \cdots & -a_2 & -a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

$$C = (b_n \ \cdots \ b_1), \quad D = b_0.$$

La démonstration se fait par un calcul facile. On peut de même montrer que tout matrice de transfert dont les coefficients sont des fractions rationnelles (le degré du numérateur étant inférieur ou égal à celui du dénominateur) admet une réalisation (voir par exemple [52]).

Remarque 12.2.5. Il n'y a pas unicité de la réalisation. En effet si (A, B, C) est une réalisation de $H(s)$, alors il est bien clair que (A_1, B_1, C_1) est aussi une réalisation de $H(s)$ avec

$$A_1 = \begin{pmatrix} A & 0 \\ * & * \end{pmatrix}, \quad B_1 = \begin{pmatrix} B \\ * \end{pmatrix}, \quad C_1 = \begin{pmatrix} C & 0 \end{pmatrix}.$$

Il existe un théorème d'unicité d'une *réalisation minimale* sous forme d'un système linéaire *contrôlable* et *observable* (voir par exemple [52]).

Exercice 12.2.1. Déterminer une réalisation de la matrice de transfert

$$H(s) = \begin{pmatrix} 1/(s^2 - 1) \\ s/(s^2 + s) \\ s/(s^2 - s) \end{pmatrix}.$$

Chapitre 13

Stabilisation des systèmes de contrôle

13.1 Systèmes linéaires autonomes

13.1.1 Rappels

Considérons le système différentiel $\dot{x}(t) = Ax(t)$, où $A \in \mathcal{M}_n(\mathbb{R})$ et $x(t) \in \mathbb{R}^n$. On note $x(\cdot, x_0)$ la solution telle que $x(0, x_0) = x_0$. On rappelle que le point origine 0, qui est un point d'équilibre, est *stable* si

$$\forall \varepsilon > 0 \quad \exists \eta > 0 \quad \forall x_0 \in \mathbb{R}^n \quad \|x_0\| \leqslant \delta \;\Rightarrow\; \forall t \geqslant 0 \quad \|x(t, x_0)\| \leqslant \varepsilon.$$

Le point 0 est *asymptotiquement stable* s'il est stable et de plus $x(t, x_0) \underset{t \to +\infty}{\longrightarrow} 0$. Pour un système linéaire, la stabilité locale est équivalente à la stabilité globale.

Théorème 13.1.1.
- *S'il existe une valeur propre λ de A telle que $Re\ \lambda > 0$, alors le point d'équilibre 0 est instable.*
- *Si toutes les valeurs propres de A sont à partie réelle strictement négative, alors le point d'équilibre 0 est asymptotiquement stable.*
- *Le point d'équilibre 0 est stable si et seulement si toute valeur propre de A est à partie réelle négative ou nulle, et si toute valeur propre à partie réelle nulle est simple.*

Remarque 13.1.1. Une valeur propre λ de A est simple si et seulement si λ est racine simple du polynôme minimal π_A. Ceci équivaut à dire que $N(\lambda) = E(\lambda)$, ou bien que $\ker(A - \lambda I) = \ker(A - \lambda I)^2$, ou encore que la décomposition de Jordan de A n'a pas de bloc de Jordan strict en λ.

La démonstration de ce théorème est claire d'après la proposition 11.2.2.

Définition 13.1.1. La matrice A est dite de *Hurwitz* si toutes ses valeurs propres sont à partie réelle strictement négative.

13.1.2 Critère de Routh, critère de Hurwitz

Dans cette section, on considère le polynôme complexe

$$P(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n,$$

et on cherche des conditions pour que ce polynôme ait toutes ses racines à partie réelle strictement négative, *i.e.* soit *de Hurwitz*.

Critère de Routh

Définition 13.1.2. La *table de Routh* est construite de la manière suivante :

$$\begin{array}{llll}
a_0 & a_2 & a_4 & a_6 \quad \cdots \quad \text{éventuellement complété par des 0} \\
a_1 & a_3 & a_5 & a_7 \quad \cdots \quad \text{éventuellement complété par des 0} \\
b_1 & b_2 & b_3 & b_4 \quad \cdots \quad \text{où } b_1 = \dfrac{a_1 a_2 - a_0 a_3}{a_1},\ b_2 = \dfrac{a_1 a_4 - a_0 a_5}{a_1}, \ldots \\
c_1 & c_2 & c_3 & c_4 \quad \cdots \quad \text{où } c_1 = \dfrac{b_1 a_3 - a_1 b_2}{b_1},\ c_2 = \dfrac{b_1 a_5 - a_1 b_3}{b_1}, \ldots \\
\vdots & \vdots & \vdots & \vdots
\end{array}$$

Le processus continue tant que le premier élément de la ligne est non nul.

La table de Routh est dit *complète* si elle possède $n+1$ lignes dont le premier coefficient est non nul.

Théorème 13.1.2. *Tous les zéros de P sont à partie réelle strictement négative si et seulement si la table complète existe, et les éléments de la première colonne sont de même signe.*

Théorème 13.1.3. *Si la table complète existe, alors P n'a aucun zéro imaginaire pur, et le nombre de zéros à partie réelle strictement positive est égal au nombre de changements de signes dans la première colonne.*

Critère de Hurwitz

On pose $a_{n+1} = a_{n+2} = \cdots = a_{2n-1} = 0$. On définit la matrice carrée d'ordre n

$$H = \begin{pmatrix}
a_1 & a_3 & a_5 & \cdots & \cdots & a_{2n-1} \\
a_0 & a_2 & a_4 & \cdots & \cdots & a_{2n-2} \\
0 & a_1 & a_3 & \cdots & \cdots & a_{2n-3} \\
0 & a_0 & a_2 & \cdots & \cdots & a_{2n-4} \\
0 & 0 & a_1 & \cdots & \cdots & a_{2n-5} \\
\vdots & \vdots & \ddots & & & \vdots \\
0 & 0 & 0 & * & \cdots & a_n
\end{pmatrix},$$

où $* = a_0$ ou a_1 selon la parité de n.

Soient $(H_i)_{i\in\{1,\dots,n\}}$ les mineurs principaux de H, *i.e.*

$$H_1 = a_1,\ H_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix},\ H_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix},\ \dots,\ H_n = \det H.$$

Théorème 13.1.4. *Si $a_0 > 0$, tout zéro de P est de partie réelle strictement négative si et seulement si $H_i > 0$, pour tout $i \in \{1,\dots,n\}$.*

Remarque 13.1.2. Supposons $a_0 > 0$.
- Si pour toute racine λ de P, on a Re $\lambda \leqslant 0$, alors $a_k \geqslant 0$ et $H_k \geqslant 0$, pour tout $k \in \{1,\dots,n\}$.
- Si $n \leqslant 3$ et si $a_k \geqslant 0$ et $H_k \geqslant 0$, pour tout $k \in \{1,2,3\}$, alors toute racine λ de P vérifie Re $\lambda \leqslant 0$.

Remarque 13.1.3. Une *condition nécessaire* de stabilité est donc, si $a_0 > 0$,

$$\forall k \in \{1,\dots,n\} \quad a_k \geqslant 0.$$

Mais cette condition n'est pas suffisante (poser $P(z) = z^4 + z^2 + 1$).

Exercice 13.1.1. Une condition nécessaire et suffisante pour qu'un polynôme de degré inférieur ou égal à 4, avec $a_0 > 0$, ait toutes ses racines à partie réelle strictement négative, est

$a_0z^2 + a_1z + a_2$	$a_1, a_2 > 0$
$a_0z^3 + a_1z^2 + a_2z + a_3$	$a_1, a_3 > 0$ et $a_1a_2 > a_0a_3$
$a_0z^4 + a_1z^3 + a_2z^2 + a_3z + a_4$	$a_1, a_2, a_4 > 0$ et $a_3(a_1a_2 - a_0a_3) > a_1^2a_4$

13.1.3 Stabilisation des systèmes de contrôle linéaires autonomes

Définition 13.1.3. Le système $\dot{x}(t) = Ax(t) + Bu(t)$, avec $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $A \in \mathcal{M}_n\mathbb{R}$, $B \in \mathcal{M}_{n,m}\mathbb{R}$, est dit *stabilisable* (par *retour d'état linéaire*, ou *feedback linéaire*), s'il existe $K \in \mathcal{M}_{m,n}\mathbb{R}$ tel que le système bouclé par le feedback $u(t) = Kx(t)$, *i.e.*

$$\dot{x}(t) = (A + BK)x(t),$$

soit asymptotiquement stable, *i.e.*

$$\forall \lambda \in \mathrm{Spec}(A + BK) \quad \mathrm{Re}\lambda < 0.$$

Remarque 13.1.4. Ce concept est invariant par similitude

$$A_1 = PAP^{-1},\ B_1 = PB,\ K_1 = KP^{-1}.$$

Théorème 13.1.5 (Théorème de placement de pôles (*pole-shifting theorem*))**.** *Si la paire (A, B) vérifie la condition de Kalman, alors pour tout polynôme réel P unitaire de degré n, il existe $K \in \mathcal{M}_{m,n}\mathbb{R}$ tel que $\chi_{A+BK} = P$, i.e. le polynôme caractéristique de $A + BK$ est égal à P.*

Corollaire 13.1.6. *Si le système de contrôle $\dot{x}(t) = Ax(t) + Bu(t)$ est contrôlable alors il est stabilisable.*

Démonstration du corollaire. Il suffit de prendre $P(X) = (X+1)^n$ et d'appliquer le théorème de placement de pôles. □

Démonstration du théorème de placement de pôles. Faisons d'abord la démonstration dans le cas $m = 1$ (on se ramènera ensuite à ce cas). Par théorème on sait que le système est semblable à la forme de Brunovski

$$A = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ -a_n & -a_{n-1} & \cdots & -a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

Posons alors $K = (k_1 \ \cdots \ k_n)$ et $u = Kx$. On a

$$A + BK = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ k_1 - a_n & k_2 - a_{n-1} & \cdots & k_n - a_1 \end{pmatrix},$$

et donc

$$\chi_{A+BK}(X) = X^n + (a_1 - k_n)X^{n-1} + \cdots + (a_n - k_1).$$

Donc, pour tout polynôme $P(X) = X^n + \alpha_1 X^{n-1} + \cdots + \alpha_n$, il suffit de choisir $k_1 = a_n - \alpha_n, \ldots, k_n = a_1 - \alpha_1$.

Dans le cas général où $m \geqslant 1$, montrons le lemme fondamental suivant.

Lemme 13.1.7. *Si la paire (A, B) vérifie la condition de Kalman, alors il existe $y \in \mathbb{R}^m$ et $C \in \mathcal{M}_{m,n}(\mathbb{R})$ tels que la paire $(A + BC, By)$ vérifie la condition de Kalman.*

D'après ce lemme, pour tout polynôme P unitaire de degré n, il existe $K_1 \in \mathcal{M}_{1,n}(\mathbb{R})$ tel que $\chi_{A+BC+ByK_1} = P$, et donc en posant $K = C + yK_1 \in \mathcal{M}_{m,n}(\mathbb{R})$, on a $\chi_{A+BK} = P$, ce qui prouve le théorème.

Preuve du lemme. Soit $y \in \mathbb{R}^m$ tel que $By \neq 0$. On pose $x_1 = By$. On a le fait suivant.

Fait 1 : Il existe $x_2 \in Ax_1 + \text{Im } B$ (et donc il existe $y_1 \in \mathbb{R}^m$ tel que $x_2 = Ax_1 + By_1$) tel que $\dim \text{Vect}\{x_1, x_2\} = 2$.

En effet sinon, on a $Ax_1 + \text{Im } B \subset \mathbb{R}x_1$, donc $Ax_1 \in \mathbb{R}x_1$ et $\text{Im } B \subset \mathbb{R}x_1$. D'où

$$\text{Im } AB = A\text{Im } B \subset \mathbb{R}Ax_1 \subset \mathbb{R}x_1,$$

et par récurrence immédiate

$$\forall k \in \mathbb{N} \quad \text{Im } A^k B \subset \mathbb{R}x_1.$$

On en déduit que

$$\text{Im } (B, AB, \ldots, A^{n-1}B) = \text{Im } B + \text{Im } AB + \cdots + \text{Im } A^{n-1}B \subset \mathbb{R}x_1,$$

ce qui contredit la condition de Kalman.

Fait 2 : Pour tout $k \leqslant n$, il existe $x_k \in Ax_{k-1} + \text{Im } B$ (et donc il existe $y_{k-1} \in \mathbb{R}^m$ tel que $x_k = Ax_{k-1} + By_{k-1}$) tel que dim $E_k = k$, où $E_k = \text{Vect}\{x_1, \ldots, x_k\}$.

En effet sinon, on a $Ax_{k-1} + \text{Im } B \subset E_{k-1}$, d'où $Ax_{k-1} \subset E_{k-1}$ et Im $B \subset E_{k-1}$. On en déduit que

$$AE_{k-1} \subset E_{k-1}.$$

En effet, on remarque que $Ax_1 = x_2 - By_1 \in E_{k-1} + \text{Im } B \subset E_{k-1}$, de même pour Ax_2, etc, $Ax_{k-2} = x_{k-1} - By_{k-1} \in E_{k-1} + \text{Im } B \subset E_{k-1}$, et enfin, $Ax_{k-1} \in E_{k-1}$.
Par conséquent

$$\text{Im } AB = A\text{Im } B \subset AE_{k-1} \subset E_{k-1},$$

et de même

$$\forall i \in \mathbb{N} \quad \text{Im } A^iB \subset E_{k-1}.$$

D'où

$$\text{Im } (B, AB, \ldots, A^{n-1}B) \subset E_{k-1},$$

ce qui contredit la condition de Kalman.

On a donc ainsi construit une base $(x_1, \ldots, x_n)$ de $\mathbb{R}^n$. On définit alors $C \in \mathcal{M}_{m,n}(\mathbb{R})$ par les relations

$$Cx_1 = y_1,\ Cx_2 = y_2, \ldots,\ Cx_{n-1} = y_{n-1},\ Cx_n \text{ quelconque.}$$

Alors la paire $(A + BC, x_1)$ vérifie la condition de Kalman, car

$$(A + BC)x_1 = Ax_1 + By_1 = x_2, \ldots, (A + BC)x_{n-1} = Ax_{n-1} + By_{n-1} = x_n.$$

□

Le théorème est prouvé. □

Remarque 13.1.5. Pour la mise en oeuvre numérique du placement de pôles, une première solution est, si la dimension d'espace n'est pas trop grande, d'appliquer les critères de Routh ou de Hurwitz de façon à déterminer une condition nécessaire et suffisante sur K pour stabiliser le système. En effet il suffit de calculer, par exemple à l'aide d'un logiciel de calcul formel comme *Maple*, le polynôme caractéristique de la matrice $A + BK$.

Une deuxième solution consiste à implémenter une méthode systématique réalisant un placement de pôles. Ce problème est essentiellement un problème inverse aux valeurs propres. Il existe beaucoup d'algorithmes mettant en oeuvre une méthode de placement de pôles. Parmi celles-ci, citons-en qui sont implémentées dans la *Control Toolbox* de *Matlab*. La première, *acker.m*, est basée sur la formule d'Ackermann (voir [44]), est limitée aux systèmes mono-entrée, mais

n'est pas fiable numériquement. Il vaut mieux utiliser *place.m*, qui est une méthode de placement de pôles robuste (voir [45]), basée sur des décompositions aux valeurs propres.

Dans l'exemple 13.3.1 traité plus loin, nous donnons un exemple d'utilisation de cette procédure.

Enfin, une troisième solution consiste à appliquer la théorie LQ (voir section 4.4.3).

13.2 Interprétation en termes de matrice de transfert

Tout d'abord, remarquons que les pôles de la matrice de transfert $H(s)$ sont exactement les valeurs propres de A. C'est pourquoi on parle des *pôles* de A (ou *modes propres*). Ainsi, le système est naturellement stable si les pôles sont à partie réelle strictement négative.

Définition 13.2.1. Le système est dit *EBSB-stable* (Entrée Bornée, Sortie Bornée) si pour toute entrée bornée, la sortie est bornée.

Proposition 13.2.1. *Si les pôles de A sont à partie réelle strictement négative alors le système est EBSB-stable (la réciproque est fausse).*

Remarque 13.2.1. La EBSB-stabilité peut donc se tester par les critères de Routh-Hurwitz.

Un feedback s'interprète de la manière suivante. Posons $C = I$. On a $H(s) = (sI - A)^{-1}B$, et $X(s) = H(s)U(s)$. On prend $u = Kx + v$, *i.e.* $U = KX + V$, d'où

$$X(s) = (I - H(s)K)^{-1}H(s)V(s).$$

Proposition 13.2.2. *Si les pôles de $I - H(s)K$ sont à partie réelle strictement négative, alors le système est EBSB-stable.*

Dans cette interprétation, la matrice de feedback K s'appelle le *gain*.

Remarque 13.2.2. La réponse impulsionnelle est $W(t) = e^{tA}B$, donc
- si $W(t) \underset{t\to+\infty}{\longrightarrow} 0$, alors le système est asymptotiquement stable ;
- si $\|W(t)\|$ est bornée quand $t \to +\infty$, alors le système est stable ;
- si $\|W(t)\|$ diverge, alors le système est instable.

13.3 Stabilisation des systèmes non linéaires

13.3.1 Rappels

Considérons le système différentiel dans $\mathbb{R}^n$

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

où $f : \mathbb{R}^n \to \mathbb{R}^n$ est C^1. On note $x(\cdot, x_0)$ la solution de ce système telle que $x(0, x_0) = x_0$.

Définition 13.3.1. – Le point $\bar{x}$ est un *point d'équilibre* si $f(\bar{x}) = 0$.
– Le point d'équilibre $\bar{x}$ est dit *stable* si

$$\forall \varepsilon > 0 \quad \exists \eta > 0 \mid \forall x_0 \in B(\bar{x}, \eta) \quad \forall t \geqslant 0 \quad \|x(t, x_0) - \bar{x}\| \leqslant \varepsilon.$$

– Le point d'équilibre $\bar{x}$ est dit *localement asymptotiquement stable* (LAS) si $\bar{x}$ est stable et si de plus $x(t, x_0) \underset{t \to +\infty}{\longrightarrow} \bar{x}$.

Théorème 13.3.1 (Théorème de linéarisation)**.** *Soit A la matrice jacobienne de f au point d'équilibre $\bar{x}$.*

1. *Si toutes les valeurs propres de A sont à partie réelle strictement négative, alors le point d'équilibre $\bar{x}$ est localement asymptotiquement stable.*
2. *S'il existe une valeur propre de A à partie réelle strictement positive, alors le point d'équilibre $\bar{x}$ est instable.*

Définition 13.3.2. Soit Ω un ouvert de $\mathbb{R}^n$ contenant le point d'équilibre $\bar{x}$. La fonction $V : \Omega \to \mathbb{R}$ est une *fonction de Lyapunov* en $\bar{x}$ sur Ω si
– V est C^1 sur Ω,
– $V(\bar{x}) = 0$, et $\forall x \in \Omega \setminus \{\bar{x}\} \quad V(x) > 0$,
– $\forall x \in \Omega \ \langle \nabla V(x), f(x) \rangle \leqslant 0$ (si l'inégalité est stricte, on dit que la fonction de Lyapunov est stricte).

Remarque 13.3.1. $\frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), f(x(t)) \rangle$.

Théorème 13.3.2 (Théorème de Lyapunov)**.** *S'il existe une fonction de Lyapunov au point d'équilibre $\bar{x}$ sur Ω, alors le point $\bar{x}$ est stable. Si la fonction de Lyapunov est stricte alors $\bar{x}$ est LAS. Si de plus V est propre sur Ω alors $\bar{x}$ est globalement asymptotiquement stable (GAS) sur Ω.*

Théorème 13.3.3 (Principe de Lasalle)**.** *Soit $V : \Omega \to \mathbb{R}^+$ une fonction de classe C^1 telle que*
– *V est propre, i.e. $\forall L \in V(\Omega) \quad V^{-1}([0, L])$ est compact dans Ω,*
– *$\forall x \in \Omega \quad \langle \nabla V(x), f(x) \rangle \leqslant 0$.*

Soit $\mathcal{I}$ le plus grand sous-ensemble de $\{x \in \Omega \mid \langle \nabla V(x), f(x) \rangle = 0\}$ invariant par le flot (en temps $t \geqslant 0$) de $\dot{x} = f(x)$. Alors toute solution $x(t)$ de $\dot{x} = f(x)$ tend vers $\mathcal{I}$, i.e.

$$d(x(t), \mathcal{I}) \underset{t \to +\infty}{\longrightarrow} 0.$$

Remarque 13.3.2. On peut énoncer le principe de Lasalle dans le cas particulier où l'ensemble invariant $\mathcal{I}$ se réduit au point $\bar{x}$. L'énoncé est alors le suivant.

Soit $\bar{x} \in \Omega$ un point d'équilibre, et $V : \Omega \to \mathbb{R}$ une fonction de classe C^1 telle que
– $V(\bar{x}) = 0$ et $\forall x \in \Omega \setminus \{\bar{x}\} \quad V(x) > 0$,
– V est propre,

– $\forall x \in \Omega \quad \langle \nabla V(x), f(x) \rangle \leqslant 0$, et de plus si $x(t)$ est une solution du système telle que $\langle \nabla V(x(t)), f(x(t)) \rangle = 0$ pour tout $t \geqslant 0$, alors $x(t) = \bar{x}$.

Alors $\bar{x}$ est GAS dans Ω.

Exercice 13.3.1. Déterminer les points d'équilibre du système différentiel

$$\begin{cases} x'(t) &= \sin\big(x(t) + y(t)\big) \\ y'(t) &= e^{x(t)} \ - \ 1, \end{cases}$$

puis étudier leur stabilité.

Exercice 13.3.2. Soit le système différentiel

$$\begin{cases} x'(t) = y(t)\big(1 + x(t) - y(t)^2\big), \\ y'(t) = x(t)\big(1 + y(t) - x(t)^2\big). \end{cases}$$

Trouver les points d'équilibre de ce système, et voir s'ils sont asymptotiquement stables (resp. instables).

Exercice 13.3.3. On considère le mouvement d'un solide rigide en rotation soumis à une force extérieure,

$$\begin{aligned} I_1\omega_1' &= (I_2 - I_3)\omega_2\omega_3 - \omega_1 \\ I_2\omega_2' &= (I_3 - I_1)\omega_3\omega_1 - \omega_2 \\ I_3\omega_3' &= (I_1 - I_2)\omega_1\omega_2 - \omega_3 \end{aligned}$$

où I_1, I_2, I_3 sont les moments d'inertie du solide, *i.e.* des constantes données. Construire une fonction de Lyapunov permettant de montrer que l'équilibre est asymptotiquement stable.

Exercice 13.3.4. Soit $g : \mathbb{R} \to \mathbb{R}$ de classe C^1 telle que $g(0) = 0$ et $xg(x) > 0$ si $x \neq 0$. Montrer que le point d'équilibre $x = 0, x' = 0$ est asymptotiquement stable pour l'équation différentielle $x'' + x' + g(x) = 0$.

Indication : on étudiera la fonction $F(x) = \int_0^x g(y)dy$ au voisinage de 0 et on introduira la fonction de Lyapunov $V(x, y) = F(x) + \frac{y^2}{2}$.

Exercice 13.3.5. Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ une fonction continue telle que toute solution de l'équation $y' = f(y),\ y(0) = y_0$, reste sur une sphère de $\mathbb{R}^n$, *i.e.*

$$\forall t \geqslant 0 \quad \|y(t)\| = \|y(0)\|.$$

1. Montrer que $f(0) = 0$.
2. Montrer que l'origine est asymptotiquement stable pour le système $x' = f(x) - x$.

Exercice 13.3.6 (Lemme de Lyapunov et applications)**.** 1. (a) Soit $A \in \mathcal{M}_n(\mathbb{R})$ dont les valeurs propres sont de partie réelle strictement négative. Montrer qu'il existe $B \in \mathcal{M}_n(\mathbb{R})$ symétrique définie positive telle que $A^T B + BA = -Id$.

(*Indication :* poser $B = \displaystyle\int_0^{+\infty} e^{tA^T} e^{tA} dt$)

(b) En déduire que $V(x) = \langle x, Bx\rangle$ est une fonction de Lyapunov pour l'équation différentielle $x' = Ax$, et que l'origine est asymptotiquement stable.

2. (a) Soit $q : \mathbb{R}^n \to \mathbb{R}^n$ une fonction continue telle que $q(x) = o(\|x\|)$. Montrer que la fonction V précédente est encore une fonction de Lyapunov pour le système $x' = Ax + q(x)$, et que l'équilibre 0 est asymptotiquement stable.

(b) Quel résultat peut-on en déduire sur la stabilité des points fixes d'un système autonome $x' = F(x)$, où $F : \mathbb{R}^n \to \mathbb{R}^n$ est de classe C^1 ?

13.3.2 Stabilisation locale d'un système de contrôle non linéaire

Considérons le système de contrôle non linéaire

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

et soit (x_e, u_e) un point d'équilibre, *i.e.* $f(x_e, u_e) = 0$. Le système linéarisé en ce point est

$$\dot{y}(t) = Ay(t) + Bv(t),$$

où

$$A = \frac{\partial f}{\partial x}(x_e, u_e),\ B = \frac{\partial f}{\partial u}(x_e, u_e).$$

Théorème 13.3.4. *Si le système linéarisé est stabilisable par le feedback $v = Ky$, alors le point d'équilibre (x_e, u_e) est LAS pour le système bouclé*

$$\dot{x}(t) = f(x(t), K(x(t) - x_e) + u_e).$$

Exemple 13.3.1. On établit qu'une condition nécessaire et suffisante sur K pour stabiliser le pendule inversé (cf exemple 5.2.1) localement autour du point d'équilibre $(\xi = \xi_c, \dot{\xi} = 0, \theta = 0, \dot{\theta} = 0)$ est

$$k_4 - k_2 L > 0, k_3 - k_1 L - (m + M)g > 0,\ k_1 > 0,$$
$$k_2((k_4 - k_2 L)(k_3 - k_1 L - (m + M)g) - MLgk_2) > k_1(k_4 - k_2 L)^2.$$

Mettons en oeuvre, en *Matlab*, sur cet exemple, différentes méthodes de stabilisation.

```
function placementpole

% L'exercice consiste \'a stabiliser le systeme de controle suivant
% (pendule invers\'e)
%
% xiddot = (m*L*thetadot^2*sin(theta)-m*g*cos(theta)*sin(theta)+u)/...
%              (M+m*sin(theta)^2)
% thetaddot = (-m*L*thetadot^2*sin(theta)*cos(theta)+...
```

```
%            (M+m)*g*sin(theta)-u*cos(theta))/(L*(M+m*sin(theta)^2))
% avec M masse du chariot,
%      m masse du pendule (tige sans masse),
%      L longueur de la tige,
% au voisinage de sa position d'equilibre instable
%                  xi = xidot = theta =thetadot = 0.
% Le systeme linearise en ce point est Xdot=AX+Bv avec
%     ( 0 1       0        0 )            (    0     )
% A = ( 0 0    -m*g/M      0 )     et B = (   1/M    ) .
%     ( 0 0       0        1 )            (    0     )
%     ( 0 0 (M+m)*g/(L*M)  0 )            ( -1/(L*M) )
% Pour cela on va poser u=K*X, o\'u K = ( k1 k2 k3 k4 ), et tester
% plusieurs m\'ethodes.
%
% 1. M\'ethode directe : calculer \'a l'aide de Maple le polynome
%    caract\'eristique de la matrice A+B*K, puis en appliquant
%    le crit\'ere de Routh d\'eterminer une CNS sur K
%    pour que la matrice A+B*K soit Hurwitz.
%    Tester diff\'erents feedbacks K.
%    Par identification des coefficients, d\'eterminer un feedback
%    K permettant de placer
%    les poles exactement aux valeurs (-1,-1,-1,-1).
%    Tester tous les feedbacks ainsi d\'etermin\'es sur le syst\'eme non
%    lin\'eaire initial.
%
% 2. Utilisation des outils Matlab (Control Toolbox).
% 2.a. Utiliser la fonction Matlab "acker" pour placer les poles
%      exactement aux valeurs (-1,-1,-1,-1).
% 2.b. Utiliser la fonction Matlab "place" pour placer les poles
%      exactement aux valeurs (-1,-2,-3,-4).
%      (attention : il faut forc\'ement prendre des poles distincts,
%       voir l'aide sur place)
% 2.c. Utiliser la fonction Matlab "lqr" pour stabiliser le syst\'eme.
% Tester ces diff\'erents outils pour stabiliser le syst\'eme non
% lin\'eaire initial.

% On prendra les valeurs num\'eriques suivantes :
% M = 10 ; m = 1 ; L = 1 ; g = 10 ;

clear all ; close all ; clc ;
global M m L g ; M = 10 ; m = 1 ; L = 1 ; g = 10 ;

% 1. M\'ethode directe. On calcule avec Maple :
% PolyCaract(A+B*K)(X) = X^4 + (k4-k2*L)/(M*L)*X^3 +
%                         (k3-k1*L-(m+M)*g)/(M*L)*X^2
%                          + k2*g/(M*L)*X + k1*g/(M*L)
```

```
% et d'apr\'es le crit\'ere de Routh on obtient la CNS suivante :
%   k4-k2*L>0, k3-k1*L-(m+M)*g>0, k1>0,
%   k2*((k4-k2*L)*(k3-k1*L-(m+M)*g)-M*L*g*k2) > k1*(k4-k2*L)^2.
% (on peut remarquer que n\'ecessairement k2>0)
% Avec les valeurs num\'eriques, cela donne :
% k1>0, k3>k1+110, k4>k2, k2*((k4-k2)*(k3-k1-110)-100*k2) > k1*(k4-k2)^2.
% Par exemple \c{c}a marche si k1=1, k2=1, k3=300, k4=2
%  (on peut fixer k1=1, k2=1, k4=2, et donner une in\'egalit\'e sur k3...)

k1=1 ; k2=1 ; k3=300 ; k4=2 ;
xinit = [0.5 0.2 0.4 1 ] ;
[t,x] = ode45(@systeme,[0 100],xinit,[],k1,k2,k3,k4) ;
figure ; subplot(2,2,1) ; plot(t,x(:,1)) ; title('xi') ;
         subplot(2,2,2) ; plot(t,x(:,2)) ; title('xidot') ;
         subplot(2,2,3) ; plot(t,x(:,3)) ; title('theta') ;
         subplot(2,2,4) ; plot(t,x(:,4)) ; title('thetadot') ;

% Pour avoir les poles exactement aux valeurs (-1,-1,-1,-1),
% on r\'esout le syst\'eme lin\'eaire
%          (k4-k2)/10 = 4
%      (k3-k1-110)/10 = 6
%                  k2 = 4
%                  k1 = 1
% d'o\'u k1=1, k2=4, k3=171, k4=44.

k1=1 ; k2=4 ; k3=171 ; k4=44 ;
xinit = [0.5 0.2 0.4 1 ] ;
[t,x] = ode45(@systeme,[0 20],xinit,[],k1,k2,k3,k4) ;
figure ; subplot(2,2,1) ; plot(t,x(:,1)) ; title('xi') ;
         subplot(2,2,2) ; plot(t,x(:,2)) ; title('xidot') ;
         subplot(2,2,3) ; plot(t,x(:,3)) ; title('theta') ;
         subplot(2,2,4) ; plot(t,x(:,4)) ; title('thetadot') ;


% 2. Utilisation des outils Matlab

% D\'efinition des matrices A et B
A = [ 0 1        0      0
      0 0     -m*g/M    0
      0 0        0      1
      0 0 (M+m)*g/(L*M) 0 ] ;
B = [ 0 ; 1/M ; 0 ; -1/(L*M) ] ;

% 2.a. Utilisation de acker
K = acker(A,B,[-1 -1 -1 -1]) ;
k1=-K(1) ; k2=-K(2) ; k3=-K(3) ; k4=-K(4) ;
```

```
xinit = [0.5 0.2 0.4 1 ] ;
[t,x] = ode45(@systeme,[0 20],xinit,[],k1,k2,k3,k4) ;
figure ; subplot(2,2,1) ; plot(t,x(:,1)) ; title('xi') ;
         subplot(2,2,2) ; plot(t,x(:,2)) ; title('xidot') ;
         subplot(2,2,3) ; plot(t,x(:,3)) ; title('theta') ;
         subplot(2,2,4) ; plot(t,x(:,4)) ; title('thetadot') ;

% 2.b. Utilisation de place
K = place(A,B,[-1 -2 -3 -4]) ;
k1=-K(1) ; k2=-K(2) ; k3=-K(3) ; k4=-K(4) ;
xinit = [0.5 0.2 0.4 1 ] ;
[t,x] = ode45(@systeme,[0 20],xinit,[],k1,k2,k3,k4) ;
figure ; subplot(2,2,1) ; plot(t,x(:,1)) ; title('xi') ;
         subplot(2,2,2) ; plot(t,x(:,2)) ; title('xidot') ;
         subplot(2,2,3) ; plot(t,x(:,3)) ; title('theta') ;
         subplot(2,2,4) ; plot(t,x(:,4)) ; title('thetadot') ;

% 2.c. Utilisation de lqr
[K,S,e] = lqr(A,B,eye(4),1) ;
k1=-K(1) ; k2=-K(2) ; k3=-K(3) ; k4=-K(4) ;
xinit = [0.5 0.2 0.4 1 ] ;
[t,x] = ode45(@systeme,[0 30],xinit,[],k1,k2,k3,k4) ;
figure ; subplot(2,2,1) ; plot(t,x(:,1)) ; title('xi') ;
         subplot(2,2,2) ; plot(t,x(:,2)) ; title('xidot') ;
         subplot(2,2,3) ; plot(t,x(:,3)) ; title('theta') ;
         subplot(2,2,4) ; plot(t,x(:,4)) ; title('thetadot') ;

% ------------------------------------------------------------------

function xdot = systeme(t,x,k1,k2,k3,k4)

global M m L g ;
xi = x(1) ; xidot = x(2) ; theta = x(3) ; thetadot = x(4) ;
u = k1*xi + k2*xidot + k3*theta + k4*thetadot ;
xdot = [ xidot
         (m*L*thetadot^2*sin(theta)-m*g*cos(theta)*sin(theta)+u)/...
            (M+m*sin(theta)^2)
         thetadot
         (-m*L*thetadot^2*sin(theta)*cos(theta)+...
            (M+m)*g*sin(theta)-u*cos(theta))/ (L*(M+m*sin(theta)^2)) ] ;
```

Les résultats sont représentés sur les figures 13.1 pour la méthode directe (question 1), et 13.2 pour l'utilisation de *place.m* (question 2.b). On peut constater l'efficacité de cette dernière procédure.

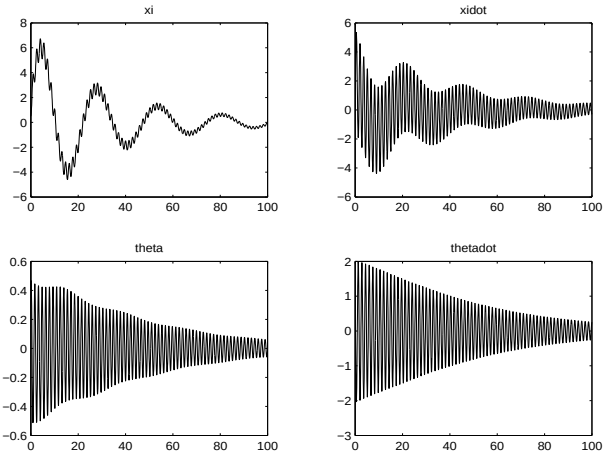


FIGURE 13.1 – Méthode directe

Exercice 13.3.7. On considère le système de contrôle

$$\dot{x}(t) = f(x(t)) + u(t)g(x(t)),$$

où l'état x et le contrôle u sont des réels, les fonctions f et g sont de classe C^∞, et $f(0) = 0$. On suppose que le point d'équilibre $(x = 0, u = 0)$ est globalement asymptotiquement stabilisable au sens suivant : il existe un contrôle feedback $u(x)$ de classe C^∞, avec $u(0) = 0$, et une fonction de Lyapunov globale stricte V pour le système bouclé $\dot{x} = f(x) + u(x)g(x)$ au point d'équilibre $x = 0$.

On considère alors le système *augmenté*

$$\begin{cases} \dot{x}(t) = f(x(t)) + y(t)g(x(t)), \\ \dot{y}(t) = v(t), \end{cases}$$

où v est le nouveau contrôle. En considérant la fonction

$$W(x, y) = V(x) + \frac{1}{2}(y - u(x))^2,$$

montrer que le feedback

$$v(x, y) = \frac{\partial u}{\partial x}(x)(f(x) + yg(x)) - \frac{\partial V}{\partial x}(x)g(x) - (y - u(x))$$

rend le point d'équilibre $(x = 0, y = 0, v = 0)$ asymptotiquement stable pour le système augmenté.

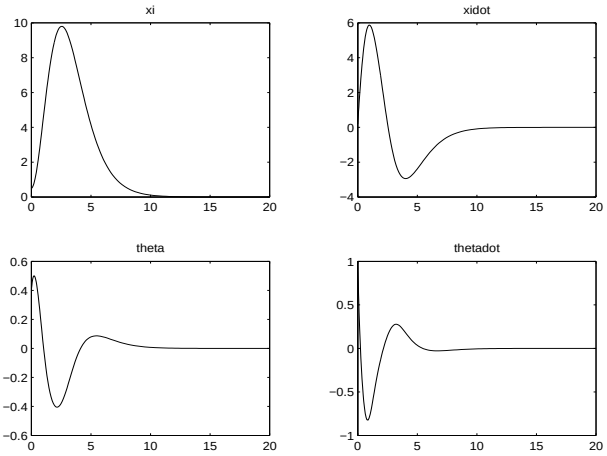


FIGURE 13.2 – Utilisation de *place.m*

13.3.3 Stabilisation asymptotique par la méthode de Jurdjevic-Quinn

Proposition 13.3.5. *On considère le système affine lisse dans* $\mathbb{R}^n$

$$\dot{x}(t) = f(x(t)) + \sum_{i=1}^{m} u_i(t) g_i(x(t)),$$

avec $f(\bar{x}) = 0$. *Supposons qu'il existe une fonction* $V : \mathbb{R}^n \to \mathbb{R}^+$ *telle que*

- $V(\bar{x}) = 0$ *et* $\forall x \neq \bar{x} \quad V(x) > 0$,
- V *est propre,*
- $\forall x \in \mathbb{R}^n \quad L_f V(x) = \langle \nabla V(x), f(x) \rangle \leqslant 0$,
- $\{x \in \mathbb{R}^n \mid L_f V(x) = 0 \text{ et } L_f^k L_{g_i} V(x) = 0, \ \forall i \in \{1, \ldots, n\}, \ \forall k \in \mathbb{N}\} = \{\bar{x}\}$.

Alors le feedback $u_i(x) = -L_{g_i} V(x), \ i = 1, \ldots, m$, *rend le point d'équilibre* $\bar{x}$ *globalement asymptotiquement stable.*

Démonstration. Soit $F(x) = f(x) - \sum_{i=1}^{m} L_{g_i} V(x) g_i(x)$ la dynamique du système bouclé. Notons tout d'abord que $F(\bar{x}) = 0$, *i.e.* $\bar{x}$ est un point d'équilibre pour le système bouclé. En effet, V est lisse et atteint son minimum en $\bar{x}$, donc $\nabla V(\bar{x}) = 0$, d'où $L_{g_i} V(\bar{x}) = 0$ pour $i = 1, \ldots, m$; de plus, $f(\bar{x}) = 0$. On a

$$L_F V(x) = \langle \nabla V(x), F(x) \rangle = L_f V(x) - \sum_{i=1}^{m} (L_{g_i} V(x))^2 \leqslant 0,$$

et si $L_F V(x(t)) = 0$ pour tout t, alors

$$L_f V(x(t)) = 0 \text{ et } L_{g_i} V(x(t)) = 0, \ i = 1, \ldots, m.$$

Par dérivation,

$$0 = \frac{d}{dt} L_{g_i} V(x(t)) = L_f L_{g_i} V(x(t)),$$

puisque $L_{g_i} V(x(t)) = 0$. D'où, clairement,

$$\forall i \in \{1, \ldots, m\} \quad \forall k \in \mathbb{N} \quad L_f^k L_{g_i} V(x(t)) = 0.$$

On en déduit que $x(t) = \bar{x}$, et la conclusion s'ensuit par le principe de Lasalle. □

Exercice 13.3.8 (Système prédateurs-proies)**.** Considérons le système prédateurs-proies contrôlé

$$\begin{aligned} \dot{x} &= x(1-y) + u, \\ \dot{y} &= -y(1-x). \end{aligned}$$

Pour le point d'équilibre $(x = 1, y = 1)$, montrer que la fonction $V(x, y) = \frac{1}{e^2} - xye^{-x-y}$ vérifie les hypothèses de la proposition précédente, et en déduire un feedback stabilisant globalement ce point d'équilibre.

Chapitre 14

Observabilité des systèmes de contrôle

Dans tout le chapitre, on se limite au cas linéaire autonome

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{14.1}$$

où $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, $A \in \mathcal{M}_n(\mathbb{R})$, $B \in \mathcal{M}_{n,m}(\mathbb{R})$, $C \in \mathcal{M}_{p,n}(\mathbb{R})$ et $D \in \mathcal{M}_{p,m}(\mathbb{R})$. Dans toute la suite, on peut supposer que $D = 0$, cela ne change rien aux résultats qui suivent.

14.1 Définition et critères d'observabilité

Notons $(x_u(t, x_0), y_u(t, x_0))$ la solution de (14.1) telle que $x_u(0, x_0) = x_0$.

Définition 14.1.1. Le système (14.1) est observable en temps T si

$$\forall x_1, x_2 \in \mathbb{R}^n \quad x_1 \neq x_2 \;\Rightarrow\; \exists u \in L^\infty([0,T], \mathbb{R}^m) \mid y_u(\cdot, x_1) \neq y_u(\cdot, x_2)$$

(dans ce cas on dit que x_1 et x_2 sont *distinguables*).

Autrement dit, si x_1 et x_2 sont distinguables s'il existe un contrôle tel que les trajectoires observées diffèrent. De manière équivalente, on peut dire

$$\forall x_1, x_2 \in \mathbb{R}^n \quad \forall u \in L^\infty([0,T], \mathbb{R}^m) \quad y_u(\cdot, x_1) = y_u(\cdot, x_2) \;\Rightarrow\; x_1 = x_2,$$

i.e. , la connaissance de la trajectoire observée détermine de manière univoque l'état initial.

L'intérêt de la notion d'observabilité est le suivant. Si on considère le système comme une boîte noire à laquelle on applique une entrée (*contrôle*, *input*) $u(t)$, et de laquelle émerge une sortie (*observable*, *output*) $y(t)$, la propriété d'être

distinguable signifie la possibilité de différentier par des expériences de type entrée-sortie.

On est aussi motivé par la stabilisation. En effet, on a vu comment stabiliser un système par retour d'état. Or il peut s'avérer coûteux de mesurer l'état complet d'un système. On peut alors se demander si la connaissance partielle de cet état permet de reconstituer l'état complet (c'est la propriété d'*observabilité*), et de stabiliser le système entier : c'est la *stabilisation par retour d'état dynamique*, ou *synthèse régulateur-observateur*.

Théorème 14.1.1. *Le système (14.1) est observable (en temps T quelconque) si et seulement si*

$$rang \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = n.$$

Démonstration. Faisons une démonstration directe de ce théorème. On montre d'abord le lemme fondamental suivant.

Lemme 14.1.2. *Le système (14.1) est observable en temps T si et seulement si, pour le système observé $\dot{x} = Ax$, $y = Cx$, $x(0) = x_0$, on a*

$$x_0 \neq 0 \ \Rightarrow y(\cdot) \not\equiv 0 \text{ sur } [0,T].$$

Preuve du lemme. Le système (14.1) est observable en temps T si et seulement si

$$\begin{aligned}
&\Big(x_1 \neq x_2 \ \Rightarrow \ \exists u \in L^\infty([0,T],\mathbb{R}^m) \mid y_u(\cdot,x_1) \neq y_u(\cdot,x_2) \text{ sur } [0,T]\Big) \\
&\Big(x_1 \neq x_2 \ \Rightarrow \exists u \in L^\infty([0,T],\mathbb{R}^m) \quad \exists t \in [0,T] \ \mid \\
\Longleftrightarrow \qquad & \qquad Ce^{tA}x_1 + Ce^{tA}\int_0^t e^{-sA}Bu(s)ds \ \neq \ Ce^{tA}x_2 + Ce^{tA}\int_0^t e^{-sA}Bu(s)ds\Big) \\
\Longleftrightarrow &\Big(x_0 = x_1 - x_2 \neq 0 \ \Rightarrow \ \exists t \in [0,T] \mid Ce^{tA}x_0 \neq 0\Big) \\
\Longleftrightarrow &\Big(x_0 \neq 0 \ \Rightarrow \ y(\cdot) \not\equiv 0 \text{ sur } [0,T] \text{ pour le système } \dot{x} = Ax, y = Cx, x(0) = x_0\Big)
\end{aligned}$$

□

On est maintenant en mesure de montrer le théorème.

Si (14.1) n'est pas observable en temps T, alors

$$\exists x_0 \neq 0 \mid \forall t \in [0,T] \quad y(t) = 0,$$

i.e.

$$\forall t \in [0,T] \quad Ce^{tA}x_0 = 0.$$

D'où, par dérivations sucessives, et en prenant $t = 0$,

$$Cx_0 = CAx_0 = \cdots = CA^{n-1}x_0 = 0,$$

i.e.

$$\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} x_0 = 0, \quad \text{et donc} \quad \text{rang} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} < n.$$

Réciproquement, si le rang de cette matrice est strictement inférieur à n, alors il existe $x_0 \neq 0$ tel que

$$Cx_0 = CAx_0 = \cdots = CA^{n-1}x_0 = 0,$$

et donc par le théorème d'Hamilton-Cayley,

$$\forall t \in \mathbb{R} \quad Ce^{tA}x_0 = 0,$$

et par conséquent le système (14.1) n'est pas observable. □

Remarque 14.1.1. Pour un système linéaire autonome, l'observabilité a lieu en temps quelconque si elle a lieu en temps T.

Remarque 14.1.2. La notion d'observabilité pour un système linéaire autonome ne dépend pas de la matrice B.

Remarque 14.1.3. On a

$$\text{rang} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = n \Leftrightarrow \text{rang} \begin{pmatrix} C^{\mathrm{T}} & A^{\mathrm{T}}C^{\mathrm{T}} & \cdots & A^{n-1\mathrm{T}}C^{\mathrm{T}} \end{pmatrix} = n,$$

et par conséquent, *le système $\dot{x} = Ax + Bu, y = Cx$ est observable si et seulement si le système $\dot{x} = A^{\mathrm{T}}x + C^{\mathrm{T}}u$ est contrôlable.* C'est la *dualité contrôlabilité/observabilité*. Ce fait, très important, permet de transférer aux systèmes observés tous les résultats établis sur les systèmes contrôlés.

On aurait pu prouver cette équivalence directement en utilisant l'application entrée-sortie, et en remarquant qu'une application linéaire $E : L^2 \to \mathbb{R}^n$ est surjective si et seulement si l'application adjointe $E^* : \mathbb{R}^n \to L^2$ est injective.

Corollaire 14.1.3. *Le système (14.1) est observable en temps T si et seulement si la matrice*

$$O(T) = \int_0^T e^{-sA^{\mathrm{T}}} C^{\mathrm{T}} C e^{-sA} ds$$

est inversible.

Définition 14.1.2 (Similitude)**.** Les systèmes

$$\begin{cases} \dot{x}_1 = A_1x_1 + B_1u_1 \\ y_1 = C_1x_1 \end{cases} \quad \text{et} \quad \begin{cases} \dot{x}_2 = A_2x_2 + B_2u_2 \\ y_2 = C_2x_2 \end{cases}$$

sont dits *semblables* s'il existe une matrice $P \in GL_n(\mathbb{R})$ telle que

$$A_2 = PA_1P^{-1},\ B_2 = PB_2,\ C_2 = C_1P^{-1}$$

(et dans ce cas on a $x_2 = Px_1, u_2 = u_1, y_2 = y_1$).

Proposition 14.1.4. *Tout système $\dot{x} = Ax + Bu$, $y = Cx$, est semblable à un système $\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u$, $y = \bar{C}\bar{x}$, avec*

$$\bar{A} = \begin{pmatrix} \bar{A}_1 & 0 \\ \bar{A}_2 & \bar{A}_3 \end{pmatrix},\ \bar{C} = (\bar{C}_1\ 0),$$

i.e.

$$\begin{cases} \dot{\bar{x}}_1 = \bar{A}_1\bar{x}_1 + \bar{B}_1u \\ \dot{\bar{x}}_2 = \bar{A}_2\bar{x}_1 + \bar{A}_3\bar{x}_2 + \bar{B}_2u \qquad \textit{partie non observable} \\ y_1 = \bar{C}_1\bar{x}_1 \end{cases}$$

et la paire $(\bar{A}_1, \bar{C}_1)$ est observable.

Démonstration. Il suffit d'appliquer le résultat vu en contrôlabilité au système $\dot{x} = A^{\mathrm{T}}x + C^{\mathrm{T}}u$. □

Définition 14.1.3. Dans cette décomposition, les valeurs propres de $\bar{A}_3$ sont appelées *modes propres inobservables* de A, et les valeurs propres de $\bar{A}_1$ sont dites *modes propres observables* de A.

Proposition 14.1.5 (Forme de Brunovski, cas $p = 1$)**.** *Dans le cas $p = 1$, le système $\dot{x} = Ax + Bu$, $y = Cx$, est observable si et seulement s'il est semblable au système $\dot{x}_1 = A_1x_1 + B_1u$, $y = C_1x_1$, avec*

$$A_1 = \begin{pmatrix} 0 & & \cdots & 0 & -a_n \\ 1 & 0 & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \\ 0 & \cdots & 0 & 1 & -a_1 \end{pmatrix},\quad C_1 = (0\ \cdots\ 0\ 1).$$

Exercice 14.1.1 (Ressort)**.** Le système $m\ddot{x} + kx = u$ est-il observable
- avec $y = x$?
- avec $y = \dot{x}$?

Exercice 14.1.2 (Amortisseurs d'une voiture)**.** Le système

$$\begin{cases} \ddot{x}_1 = -k_1x_1 - d_1\dot{x}_1 + l_1u, \\ \ddot{x}_2 = -k_2x_2 - d_2\dot{x}_2 + l_2u. \end{cases}$$

est-il observable
- avec $y_1 = x_1, y_2 = x_2$?

– avec $y = x_1$?
– avec $y_1 = x_1, y_2 = \dot{x}_2$?

Exercice 14.1.3. Le pendule inversé linéarisé (cf exemple 5.2.1) est-il observable
– avec $y = \xi$?
– avec $y = \theta$?
– avec $y_1 = \theta, y_2 = \dot{\theta}$?

14.2 Stabilisation par retour d'état statique

On peut se demander si, étant donné un système contrôlable et observable $\dot{x} = Ax + Bu$, $y = Cx$, il existe un feedback $u = Ky$ stabilisant le système, *i.e.* si la matrice $A + BKC$ est Hurwitz.

La réponse est *NON*. Pour le voir, considérons les matrices

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ C = (1\ 0).$$

Le système $\dot{x} = Ax + Bu$, $y = Cx$, est trivialement contrôlable et observable. Pourtant, pour toute matrice scalaire $K = (k)$, la matrice

$$A + BKC = \begin{pmatrix} 0 & 1 \\ k & 0 \end{pmatrix}$$

n'est pas Hurwitz.

En conclusion, un feedback par retour d'état statique ne suffit pas en général. C'est pourquoi, dans la suite, on va voir comment construire un retour d'état dynamique.

14.3 Observateur asymptotique de Luenberger

Motivation : supposons que le système $\dot{x} = Ax + Bu$, $y = Cx$, soit observable. Le but est de construire un *observateur asymptotique* $\hat{x}(\cdot)$ de $x(\cdot)$, *i.e.* une fonction dynamique $\hat{x}(\cdot)$ de l'observable $y(\cdot)$, telle que $\hat{x}(t) - x(t) \underset{t\to+\infty}{\longrightarrow} 0$. L'idée est de copier la dynamique du système observé et d'y ajouter un correctif en tenant compte de l'écart entre la prédiction et la réalité.

Définition 14.3.1. Un *observateur asymptotique* (ou *observateur de Luenberger*) $\hat{x}(\cdot)$ de $x(\cdot)$ est une solution d'un système du type

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(C\hat{x}(t) - y(t)),$$

où $L \in \mathcal{M}_{n,p}(\mathbb{R})$ est appelée *matrice de gain*, telle que

$$\forall x(0), \hat{x}(0) \in \mathbb{R}^n \quad \hat{x}(t) - x(t) \underset{t\to+\infty}{\longrightarrow} 0.$$

Remarque 14.3.1. Introduisons $e(t) = \hat{x}(t) - x(t)$, l'erreur entre la prédiction $\hat{x}(\cdot)$ et l'état réel $x(\cdot)$. On a

$$\dot{e}(t) = (A + LC)e(t),$$

et donc $e(t) \underset{t\to+\infty}{\longrightarrow} 0$ pour toute valeur initiale $e(0)$ si et seulement si la matrice $A + LC$ est Hurwitz. Construire un observateur asymptotique revient donc à déterminer une matrice de gain L telle que $A + LC$ soit Hurwitz. Ainsi, de manière duale au théorème de placement de pôles, on a le résultat suivant.

Théorème 14.3.1 (Théorème de placement des modes propres de l'observateur)**.** *Si la paire (A, C) est observable, alors le système admet un observateur asymptotique (i.e. on peut construire une matrice de gains L telle que $A + LC$ soit Hurwitz).*

Démonstration. La paire $(A^{\mathrm{T}}, C^{\mathrm{T}})$ étant contrôlable, d'après le théorème de placement de pôles il existe une matrice L^{T} telle que la matrice $A^{\mathrm{T}} + C^{\mathrm{T}}L^{\mathrm{T}}$ soit Hurwitz. □

14.4 Stabilisation par retour dynamique de sortie

On a vu comment construire
- un régulateur (feedback) pour un système contrôlable,
- un observateur asymptotique pour un système observable.

Il semble naturel, pour un système contrôlable et observable, de construire un régulateur en fonction de l'observateur asymptotique de l'état : c'est l'étape de *synthèse régulateur-observateur*.

Définition 14.4.1. On appelle *feedback dynamique de sortie*, ou *observateur-régulateur*, le feedback $u = K\hat{x}$, où

$$\dot{\hat{x}} = A\hat{x} + Bu + L(C\hat{x} - y).$$

Théorème 14.4.1 (Théorème de stabilisation par retour dynamique de sortie)**.** *Si le système $\dot{x} = Ax + Bu$, $y = Cx$, est contrôlable et observable, alors il est stabilisable par retour dynamique de sortie, i.e. il existe des matrices de gain $K \in \mathcal{M}_{m,n}(\mathbb{R})$ et $L \in \mathcal{M}_{n,p}(\mathbb{R})$ telles que les matrices $A + BK$ et $A + LC$ soient Hurwitz, et alors le système bouclé*

$$\begin{aligned}\dot{x} &= Ax + BK\hat{x}\\ \dot{\hat{x}} &= (A + BK + LC)\hat{x} - Ly\end{aligned}$$

est asymptotiquement stable.

Démonstration. Posons $e = \hat{x} - x$. Alors

$$\frac{d}{dt}\begin{pmatrix} x \\ e \end{pmatrix} = \begin{pmatrix} A + BK & BK \\ 0 & A + LC \end{pmatrix}\begin{pmatrix} x \\ e \end{pmatrix},$$

et donc ce système est asymptotiquement stable si et seulement si les matrices $A + BK$ et $A + LC$ sont Hurwitz, ce qui est possible avec les propriétés de contrôlabilité et d'observabilité. □

Définition 14.4.2. Les valeurs propres de $A + BK$ sont dites *modes propres du régulateur*, et les valeurs propres de $A + LC$ sont dites *modes propres de l'observateur.*

Application à la stabilisation locale d'un système non linéaire par retour dynamique de sortie.

Considérons le système non linéaire

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= g(x(t)) \end{aligned}$$

Soit (x_e, u_e) un point d'équilibre, *i.e.* $f(x_e, u_e) = 0$. Le système linéarisé en (x_e, u_e) s'écrit

$$\begin{aligned} \delta\dot{x}(t) &= A\delta x(t) + B\delta u(t) \\ \delta y(t) &= C\delta x(t) \end{aligned}$$

avec

$$A = \frac{\partial f}{\partial x}(x_e, u_e),\ B = \frac{\partial f}{\partial u}(x_e, u_e),\ C = \frac{\partial g}{\partial x}(x_e).$$

D'après le théorème de linéarisation, on obtient le résultat suivant.

Théorème 14.4.2. *Si le système linéarisé est contrôlable et observable, alors il existe des matrices de gains K et L telles que les matrices $A+BK$ et $A+LC$ soient Hurwitz, et alors le contrôle $u = u_e + K\delta\hat{x}$, où*

$$\delta\dot{\hat{x}} = (A + BK + LC)\delta\hat{x} - L(y - g(x_e)),$$

stabilise localement le système au voisinage du point d'équilibre (x_e, u_e).

Exercice 14.4.1 (Problème d'examen)**.** On considère un mélangeur dans lequel arrivent un même produit, par deux entrées différentes, avec des concentrations respectives c_1 et c_2 (constantes), et des débits $u_1(t)$ et $u_2(t)$. Le volume dans le mélangeur est noté $V(t)$ et la concentration du produit $c(t)$. Le débit en sortie est $d(t) = \gamma\sqrt{V(t)}$, où γ est une constante. Les contrôles sont $u_1(t)$ et $u_2(t)$.

1. Par un bilan volume-matière, établir que

$$\begin{cases} \dfrac{d}{dt}V(t) = u_1(t) + u_2(t) - d(t), \\ \dfrac{d}{dt}(c(t)V(t)) = c_1u_1(t) + c_2u_2(t) - c(t)d(t), \end{cases}$$

puis que

$$\begin{cases} \dot{V}(t) = u_1(t) + u_2(t) - \gamma\sqrt{V(t)}, \\ \dot{c}(t) = \dfrac{1}{V(t)}\left((c_1 - c(t))u_1(t) + (c_2 - c(t))u_2(t)\right). \end{cases}$$

Le but est de stabiliser le système à des débits constants en entrée et en sortie, à une concentration constante en sortie, et à un volume constant, *i.e.* on veut que, lorsque t tend vers $+\infty$,

$$u_1(t) \to u_1^0,\ u_2(t) \to u_2^0,\ d(t) \to d^0,\ c(t) \to c^0,\ V(t) \to V^0.$$

2. (a) Montrer que
$$\begin{cases} u_1^0 + u_2^0 = d^0 = \gamma\sqrt{V^0}, \\ c_1 u_1^0 + c_2 u_2^0 = c^0 d^0 = \gamma c^0 \sqrt{V^0}. \end{cases}$$

   (b) Montrer que le système linéarisé au point d'équilibre (V^0, c^0, u_1^0, u_2^0) est donné par les matrices
$$A = \begin{pmatrix} \alpha & 0 \\ 0 & 2\alpha \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ \beta_1 & \beta_2 \end{pmatrix},$$
   avec $\alpha = -\dfrac{\gamma}{2\sqrt{V^0}}$, $\beta_1 = \dfrac{c_1 - c^0}{V^0}$ et $\beta_2 = \dfrac{c_2 - c^0}{V^0}$.

   (c) Montrer que ce système linéarisé est contrôlable.

   (d) Enoncer et démontrer une condition suffisante sur $K = \begin{pmatrix} k_1 & k_2 \\ k_3 & k_4 \end{pmatrix}$ pour que le système bouclé par le feedback
$$u = \begin{pmatrix} u_1^0 \\ u_2^0 \end{pmatrix} + K \begin{pmatrix} V - V^0 \\ c - c^0 \end{pmatrix}$$
   soit localement asymptotiquement stable en ce point d'équilibre.

   (e) Construire un tel feedback plaçant les pôles en -1.

3. On observe en sortie la quantité $y(t) = c(t)V(t)$.

   (a) Ecrire le système linéarisé observé au point d'équilibre précédent, et montrer qu'il est observable.

   (b) Expliquer soigneusement comment effectuer une stabilisation locale en ce point par retour d'état dynamique, en utilisant l'observable précédente. On donnera notamment des conditions nécessaires et suffisantes sur les matrices de gain assurant la stabilisation.

Exercice 14.4.2 (Problème d'examen)**.** On considère un système mécanique plan formé d'un rail (représenté par un segment) et d'un chariot, assimilé à un point matériel de masse m, roulant sans frottement sur le rail. Le rail tourne autour du point O. Soit θ l'angle que fait le rail avec l'axe horizontal, et x l'abscisse du chariot sur le rail (distance entre O et le chariot). Soit J le moment d'inertie du rail par rapport à O, et g l'accélération de la pesanteur. Le contrôle est le couple u exercé sur le rail (voir figure 14.1).

1. (a) Montrer que le Lagrangien du système s'écrit
$$L(x, \dot{x}, \theta, \dot{\theta}) = \frac{1}{2} J\dot{\theta}^2 + \frac{1}{2} m(\dot{x}^2 + x^2\dot{\theta}^2) - mgx\sin\theta.$$

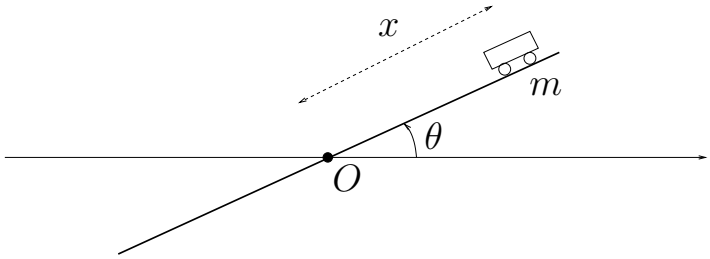


FIGURE 14.1 – Chariot sur rail

(b) En déduire que les équations du système mécanique sont

$$\begin{cases} \ddot{x}(t) = x(t)\dot{\theta}(t)^2 - g\sin\theta(t) \\ \ddot{\theta}(t) = \dfrac{1}{J + mx(t)^2}\Big(u(t) - 2mx(t)\dot{x}(t)\dot{\theta}(t) - mgx(t)\cos\theta(t)\Big). \end{cases}$$

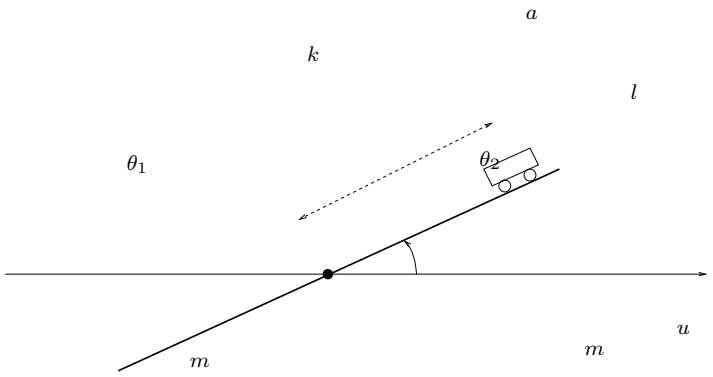
(c) En posant $y = \dot{x}$, $\dot{\theta} = \omega$, et $X = (x, y, \theta, \omega)$, mettre ce système sous forme d'un système de contrôle (S) de la forme $\dot{X}(t) = f(X(t), u(t))$.

2. Déterminer les points d'équilibre (X_e, u_e) du système de contrôle (S).
3. Ecrire le système linéarisé autour du point d'équilibre $(X_e, u_e) = (0, 0)$, et montrer qu'il est contrôlable.
4. (a) Enoncer et démontrer une condition suffisante sur $K = (k_1, k_2, k_3, k_4)$ pour que le système bouclé par le feedback $u = KX$ soit localement asymptotiquement stable en ce point d'équilibre.
   (b) Construire un tel feedback plaçant les pôles en -1.
5. (a) Le système linéarisé est-il observable si on observe $y = x$?
   (b) Est-il possible de stabiliser le système en $(0, 0)$ par le retour d'état statique $u = ky$?
6. Expliquer soigneusement comment effectuer une stabilisation en $(0, 0)$ par retour d'état dynamique, en utilisant l'observable précédente. On donnera notamment des conditions nécessaires et suffisantes sur les matrices de gain assurant la stabilisation.

Exercice 14.4.3 (Problème d'examen)**.** On considère un système mécanique plan formé de deux pendules de masse m, couplés par un ressort de raideur k, et ayant un angle θ_i, $i = 1, 2$, avec la verticale. Pour simplifier, on suppose que les tiges, de longueur l, des pendules, sont de masse nulle, et que, en approximation, l'axe du ressort reste horizontal au cours du mouvement. On note g l'accélération de la pesanteur. Le contrôle est une force u horizontale exercée sur le pendule de droite.

Avec l'approximation précédente, les équations du système mécanique sont

$$\begin{cases} ml^2\ddot{\theta}_1(t) = ka^2(\sin\theta_2(t) - \sin\theta_1(t))\cos\theta_1(t) - mgl\sin\theta_1(t) \\ ml^2\ddot{\theta}_2(t) = ka^2(\sin\theta_1(t) - \sin\theta_2(t))\cos\theta_2(t) - mgl\sin\theta_2(t) + u(t)\cos\theta_2(t) \end{cases}$$



1. En posant $\omega_i = \dot{\theta}_i$, $i = 1, 2$, et $X = (\theta_1, \omega_1, \theta_2, \omega_2)$, mettre ce système sous forme d'un système de contrôle (S) de la forme $\dot{X}(t) = f(X(t), u(t))$. Montrer que $(X_e, u_e) = (0, 0)$ est un point d'équilibre du système.

   Dans la suite du problème, on pose

$$\alpha = \frac{ka^2}{ml^2} + \frac{g}{l},\ \beta = \frac{ka^2}{ml^2},\ \gamma = \frac{1}{ml^2}.$$

2. Ecrire le système linéarisé autour du point d'équilibre $(X_e, u_e) = (0, 0)$, et montrer qu'il est contrôlable.
3. (a) Démontrer qu'une condition suffisante sur $K = (k_1, k_2, k_3, k_4)$ pour que le système (S) bouclé par le feedback $u = KX$ soit localement asymptotiquement stable en ce point d'équilibre est

$$k_4 < 0,\ k_3 < \frac{2\alpha}{\gamma},\ \beta\gamma k_1 + \alpha\gamma k_3 + \beta^2 < \alpha^2,$$

$$(\alpha k_4 + \beta k_2)(-\gamma k_4(2\alpha - \gamma k_3) + \gamma(\alpha k_4 + \beta k_2)) < \gamma k_4^2(\alpha\gamma k_3 + \beta^2 - \alpha^2 + \beta\gamma k_1).$$

   (b) Construire un tel feedback plaçant les pôles en -1.
4. (a) Le système linéarisé est-il observable si on observe $y = \theta_1$?
   (b) Est-il possible de stabiliser le système en $(0, 0)$ par le retour d'état statique $u = k_0 y$?
   (c) Expliquer soigneusement comment effectuer une stabilisation en $(0, 0)$ par retour d'état dynamique, en utilisant l'observable précédente. On donnera notamment des conditions suffisantes sur les matrices de gain assurant la stabilisation.

Exercice 14.4.4 (Probléme d'examen)**.** On considère un système de suspension magnétique, où une boule magnétique de masse m est maintenue en lévitation

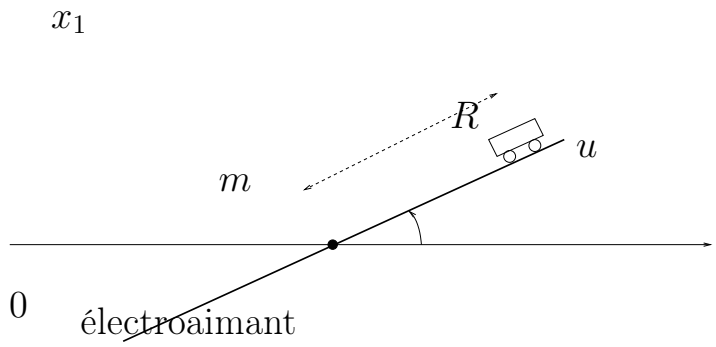
par un électroaimant relié à un circuit électrique, de résistance R. Le contrôle u est la tension aux bornes de ce circuit. On note $x_1 \geqslant 0$ la position verticale de la boule, avec, par convention, $x_1 = 0$ lorsque la boule repose sur l'électroaimant (en l'absence de courant). On note x_3 l'intensité traversant le circuit électrique. L'inductance électromagnétique est modélisée par

$$L(x_1) = L_1 + \frac{L_0}{1 + \frac{x_1}{a}},$$

où L_0, L_1 et a sont des constantes positives. La force électromagnétique (verticale) engendrée par l'électroaimant est alors

$$F(x_1, x_3) = -\frac{L_0 a x_3^2}{2(a + x_1)^2}.$$

Enfin, la boule est aussi soumise à une force de friction $-k\dot{x}_1$, où $k > 0$.



Les équations du système sont

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = g - \dfrac{k}{m} x_2 - \dfrac{L_0 a x_3^2}{2m(a + x_1)^2} \\ \dot{x}_3 = \dfrac{1}{L_1 + \frac{L_0}{1+\frac{x_1}{a}}} \left(-R x_3 + \dfrac{L_0 a x_2 x_3}{(a + x_1)^2} + u \right) \end{cases}$$

1. (a) En posant $x = (x_1, x_2, x_3)^T$, mettre ce système sous forme d'un système de contrôle (S) de la forme $\dot{x}(t) = f(x(t), u(t))$.

   (b) Montrer que les points d'équilibre (x_e, u_e) du système sont de la forme $((r, 0, i), u_e)$, où $r > 0$, et

$$i = (a + r)\sqrt{\frac{2mg}{L_0 a}}, \quad u_e = Ri.$$

Dans la suite du problème, on pose

$$\alpha = \frac{L_0 a i}{(a+r)^2}, \quad \beta = \frac{1}{L_1 + \frac{L_0}{1+\frac{r}{a}}}.$$

Le but est de stabiliser le système en un point d'équilibre $(x_e, u_e) = ((r, 0, i), u_e)$. On prendra les valeurs numériques suivantes :

$$m = 0.1\ kg, \quad k = 0.001\ N.m^{-1}.s^{-1}, \quad g = 9.81\ m.s^{-2}, \quad a = 0.05\ m,$$
$$L_0 = 0.01\ H, \quad L_1 = 0.02\ H, \quad R = 1\ \Omega, r = 0.05\ m.$$

2. (a) Montrer que le système linéarisé autour du point d'équilibre (x_e, u_e) s'écrit $\delta\dot{x}(t) = A\delta x(t) + B\delta u(t)$, avec

$$A = \begin{pmatrix} 0 & 1 & 0 \\ \frac{\alpha i}{m(a+r)} & -\frac{k}{m} & -\frac{\alpha}{m} \\ 0 & \alpha\beta & -\beta R \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ \beta \end{pmatrix},$$

et montrer qu'il est contrôlable (en temps quelconque).

(b) Ce système linéarisé est-il stable en l'absence de contrôle ? Qu'en déduire pour le système (S) ?

3. (a) Démontrer qu'une condition suffisante sur $K = (k_1, k_2, k_3)$ pour que le système (S) bouclé par le feedback $u = u_e + K(x - x_e)$ soit localement asymptotiquement stable en ce point d'équilibre est

$$k_1 > \frac{-i}{a+r}, \qquad k_3 < R + \frac{k}{\beta m},$$

$$\left(\beta R + \frac{k}{m} - \beta k_3\right)\left(Rk + \alpha^2 - \frac{\alpha i}{\beta(a+r)} + \alpha k_2 - k k_3\right) > \alpha k_1 + \frac{\alpha i}{a+r}.$$

(b) Construire un tel feedback plaçant les pôles en -1.

4. (a) Le système linéarisé est-il observable si on observe $y = x_1$?

(b) Est-il possible de stabiliser le système en (x_e, u_e) par le retour d'état statique $u = k_0 y$?

(c) Expliquer soigneusement comment stabiliser (S) en (x_e, u_e) par retour d'état dynamique, en utilisant l'observable précédente. On donnera notamment des conditions suffisantes sur les matrices de gain assurant la stabilisation.

Bibliographie

[1] H. Abou-Kandil, G. Freiling, V. Ionescu, G. Jank, Matrix Riccati equations, Control and systems theory, Systems & Control : Foundations & Applications, Birkhäuser Verlag, Basel, 2003.

[2] A. Agrachev, Y. Sachkov, Control theory from the geometric viewpoint, Encyclopaedia of Mathematical Sciences, 87, Control Theory and Optimization, II, Springer-Verlag, Berlin, 2004.

[3] B. D. Anderson, J. B. Moore, Optimal filtering, Prentice hall, Englewood Cliffs, 1979.

[4] B. d'Andréa-Novel, M. Cohen de Lara, Cours d'Automatique, commande linéaire des systèmes dynamiques, les Presses, Ecole des Mines de Paris, 2000.

[5] V. I. Arnold, Méthodes mathématiques pour la mécanique classique, Editions Mir, Moscou, 1976.

[6] A. Avez, Calcul différentiel, Masson, Paris, 1983.

[7] M. Bardi, I. Capuzzo-Dolcetta, Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations, Birkhäuser, Inc., Boston, 1997.

[8] G. Barles, Solutions de viscosité des équations de Hamilton-Jacobi, Math. & Appl. 17, Springer-Verlag, 1994.

[9] A. Bensoussan, Filtrage optimal des systèmes linéaires, Dunod, Paris, 1971.

[10] M. Bergounioux, Optimisation et contrôle des systèmes linéaires, Dunod, Collection Sciences Sup, 2001.

[11] J.T. Betts, *Practical methods for optimal control and estimation using nonlinear programming*, Second edition, Advances in Design and Control, 19, SIAM, Philadelphia, PA, 2010.

[12] O. Bolza, Calculus of variations, Chelsea Publishing Co., New York, 1973.

[13] B. Bonnard, M. Chyba, The role of singular trajectories in control theory, Math. Monograph, Springer-Verlag, 2003.

[14] B. Bonnard, L. Faubourg, E. Trélat, Optimal control of the atmospheric arc of a space shuttle and numerical simulations by multiple-shooting techniques, Math. Models Methods Applied Sci. 2, 15, 2005.

[15] B. Bonnard, L. Faubourg, G. Launay, E. Trélat, Optimal control with state constraints and the space shuttle re-entry problem, Journal of Dynamical and Control Systems, Vol. 9, no. 2, 2003, 155–199.

[16] B. Bonnard, I. Kupka, Generic properties of singular trajectories, Annales de l'IHP, Analyse non linéaire, Vol. 14, no. 2, 167–186, 1997.

[17] B. Bonnard, E. Trélat, Une approche géométrique du contrôle optimal de l'arc atmosphérique de la navette spatiale, ESAIM Cont. Opt. Calc. Var., Vol. 7, 2002, 179–222.

[18] U. Boscain, B. Piccoli, Optimal syntheses for control systems on 2-D Manifolds, Springer SMAI series, Vol. 43, 2004.

[19] H. Brezis, Analyse fonctionnelle, théorie et applications, Masson, Paris, 1983.

[20] R. Brockett, Finite dimensional linear systems, Wiley, New York, 1973.

[21] A. E. Bryson, Y. C. Ho, Applied optimal control, Hemisphere Publishing Corp. Washington, D.C., 1975.

[22] F. Clarke, Optimization and nonsmooth analysis, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley & Sons, Inc., New York, 1983.

[23] M. G. Crandall, P. L. Lions, Viscosity solutions of Hamilton-Jacobi equations, Trans. Amer. Math. Soc. 277, 1983, 1–42.

[24] M. Crouzeix, A. Mignot, Analyse numérique des équations différentielles, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1984.

[25] L. C. Evans, Partial differential equations, Amer. Math. Soc., Providence, RI, 1998.

[26] P. Faurre, M. Depeyrot, Eléments d'automatique, Dunod, 1974.

[27] P. Faurre, M. Robin, Eléments d'automatique, Dunod, 1984.

[28] H. Federer, Geometric measure theory, Die Grundlehren der mathematischen Wissenschaften, Band 153, Springer-Verlag, New York Inc., 1969.

[29] R. Fletcher, Practical Methods of Optimization, Vol. 1, Unconstrained Optimization, and Vol. 2, Constrained Optimization, John Wiley and Sons, 1980.

[30] B. Friedland, Control system design, Mac Graw-Hill, New York, 1986.

[31] R. V. Gamkrelidze, Discovery of the maximum principle, Journal of Dynamical and Control Systems, Vol. 5, no. 4, 1999, 437-451.

[32] J.-P. Gauthier, Y. Kupka, Deterministic observation theory and applications, Cambridge University Press, Cambridge, 2001.

[33] P. E. Gill, W. Murray, M. H. Wright, Practical Optimization, London, Academic Press, 1981.

[34] W. Grimm, A. Markl, Adjoint estimation from a direct multiple shooting method, J. Opt. Theory Appl. 92, no. 2, 1997, 262–283.

[35] J. Harpold, C. Graves, Shuttle entry guidance, Journal of Astronautical Sciences, Vol. 27, pp. 239–268, 1979.

[36] R. F. Hartl, S. P. Sethi, R. G. Vickson, A survey of the maximum principles for optimal control problems with state constraints, SIAM Review 37, no. 2, 1995, 181–218.

[37] H. Hermes, J.P. LaSalle, Functional analysis and time optimal control, Mathematics in Science and Engineering, Vol. 56, Academic Press, New York-London, 1969.

[38] L. M. Hocking, Optimal control, an introduction to the theory with applications, Oxford Applied Mathematics and Computing Science Series, 1991.

[39] A. D. Ioffe, V. M. Tihomirov, Theory of extremal problems, Studies in Mathematics and its Applications, 6, North-Holland Publishing Co., Amsterdam-New York, 1979.

[40] A. Isidori, Nonlinear control systems, Third edition, Communications and Control Engineering Series, Springer-Verlag, Berlin, 1995.

[41] A. Isidori, Nonlinear control systems, II, Communications and Control Engineering Series, Springer-Verlag London, Ltd., London, 1999.

[42] D. Jacobson, D. Lele, J. L. Speyer, New necessary conditions of optimality for control problems with state-variable inequality constraints, Journal of Mathematical Analysis and Applications, Vol. 35, pp. 255–284, 1971.

[43] V. Jurdjevic, Geometric control theory, Cambridge university press, 1997.

[44] T. Kailath, Linear Systems, Prentice-Hall, 1980.

[45] J. Kautsky, N. K. Nichols, Robust pole assignment in linear state feedback, Int. J. Control, 41, 1985, 1129–1155.

[46] H. K. Khalil, Nonlinear systems, Macmillan Publishing Company, New York, 1992.

[47] H. Kwakernaak, R. Sivan, Linear optimal control systems, John Wiley, New-York, 1972.

[48] J. Lafontaine, Introduction aux variétés différentielles, Presses universitaires, Grenoble, 1996.

[49] P. Lascaux, R. Théodor, Analyse numérique matricielle appliquée à l'art de l'ingénieur, Tomes 1 et 2, Masson, Paris.

[50] A. Laub, A Schur method for solving algebraic Riccati equations, IEEE Trans. Automat. Control, AC-24, 1979, 913–921.

[51] W. F. Arnold, A. J. Laub, Generalized eigenproblem algorithms and software for algebraic Riccati equations, Proc. IEEE 72, 1984,pp. 1746-1754.

[52] E. B. Lee, L. Markus, Foundations of optimal control theory, John Wiley, New York, 1967.

[53] G. Leitmann, An introduction to optimal control, McGraw-Hill Book Company, 1966.

[54] A. Locatelli, Optimal control, an introduction, Birkhäuser, Basel, 2001.

[55] H. Maurer, On optimal control problems with bounded state variables and control appearing linearly, SIAM Journal on Control and Optimization, Vol. 15, 3, pp. 345–362, 1977.

[56] A. Miele, Recent advances in the optimization and guidance of aeroassociated orbital transfers, Acta Astronautica, Vol. 38, 10, pp. 747–768, 1996.

[57] H. Nijmeijer, A. J. Van der Shaft, Nonlinear dynamical control systems, Springer Verlag, 1990.

[58] R. Pallu de la Barrière, Cours d'automatique théorique, Collection Universitaire de Mathématiques, No. 17, Dunod, Paris, 1966.

[59] F. Pham, Géométrie différentielle, 1992.

[60] L. Pontryagin, V. Boltyanski, R. Gamkrelidze, E. Michtchenko, Théorie mathématique des processus optimaux, Editions Mir, Moscou, 1974.

[61] J. Rappaz, M. Picasso, Introduction à l'analyse numérique, Presses Polytechniques et Universitaires Romandes, Lausanne, 1998.

[62] A. V. Sarychev, First- and second-order integral functionals of the calculus of variations which exhibit the Lavrentiev phenomenon, J. of Dynamical and Control Systems, Vol. 3, No. 4, 1997, 565-588.

[63] J. A. Sethian, Level set methods and fast marching methods. Evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science. Cambridge Monographs on Applied and Computational Mathematics, 3, Cambridge University Press, 1999.

[64] E. D. Sontag, Mathematical Control Theory, Deterministic Finite Dimensional Systems, Springer-Verlag, 2nd edition, 1998.

[65] J. Stoer, R. Bulirsch, Introduction to numerical analysis, Springer-Verlag, Berlin, 1980.

[66] O. von Stryk, R. Bulirsch, Direct and indirect methods for trajectory optimization, Annals of Operations Research 37, 1992, 357–373.

[67] A. Subbotin, Generalized solutions of first-order PDEs, The dynamical optimization perspective, Systems & Control : Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 1995.

[68] H. J. Sussmann, J. C. Willems, The brachistochrone problem and modern control theory, Contemporary trends in nonlinear geometric control theory and its applications (Mexico City, 2000), 113–166, World Sci. Publishing, River Edge, NJ, 2002.

[69] H. J. Sussmann, New theories of set-valued differentials and new versions of the maximum principle of optimal control theory, Nonlinear Control in the Year 2000, A. Isidori, F. Lamnabhi-Lagarrigue and W. Respondek Eds., Springer-Verlag, 2000, 487–526.

[70] H. J. Sussmann, A nonsmooth hybrid maximum principle, Stability and stabilization of nonlinear systems (Ghent, 1999), 325–354, Lecture Notes in Control and Inform. Sci., 246, Springer, London, 1999.

[71] E. Trélat, Some properties of the value function and its level sets for affine control systems with quadratic cost, Journal of Dynamical and Control Systems, Vol. 6, No. 4, 2000, 511–541.

[72] E. Trélat, Etude asymptotique et transcendance de la fonction valeur en contrôle optimal ; catégorie log-exp en géométrie sous-Riemannienne dans le cas Martinet. Thèse de doctorat, Univ. de Bourgogne, 2000.

[73] R. Vinter, Optimal control, Systems & Control : Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 2000.