ÈCOLE POLYTECHNIQUE
DÉPARTEMENT DE MATHÉMATIQUES APPLIQUÉES

Majeure

Mathématiques Appliquées

ANALYSE ET COMMANDE DE SYSTÈMES DYNAMIQUES

Frédéric Bonnans Pierre Rouchon

Édition 2003

Avant propos

Les mécanismes de régulation et d'adaptation sont largement répandus dans la nature. Derrière ces deux mécanismes se retrouvent souvent en filigrane la commandabilité, l'observabilité et l'optimalité. Ces mécanismes sont présents déjà chez les organismes vivants afin d'assurer le maintien de certaines variables essentielles comme le taux de sucre, la température, ... En ingénierie également les mécanismes d'asservissement et de recalage ont une longue histoire. Au temps des romains les niveaux d'eau dans les aqueducs étaient pilotés par un système complexe de vannes.

Les développements modernes ont débuté au 17ème siècle avec les travaux du savant hollandais Huyghens sur les horloges à pendules. Il était alors très important pour la marine de Louis XIV d'embarquer sur les bateaux des horloges les plus précises possible. La mesure du temps intervenait de façon cruciale dans les calculs de longitude. Huyghens s'est ainsi intéressé à la régulation en vitesse des horloges. Les idées élaborées par Huyghens et bien d'autres comme le savant anglais Robert Hooke furent utilisées dans la régulation en vitesse des moulins à vent. Une idée centrale fut alors d'utiliser un système mécanique à boules tournant autour d'un axe et dont la rotation était directement proportionnelle à celle du moulin. Plus les boules tournent vite et plus elles s'éloignent de l'axe. Elles actionnent alors par un système de renvois ingénieux les ailes du moulin de façon à réduire le couple dû au vent. En langage moderne, il s'agit d'un régulateur proportionnel.

La révolution industrielle vit l'adaptation par James Watt du régulateur à boules pour les machines à vapeur. Plus les boules tournent vite, plus elles ouvrent une soupape qui laisse s'échapper la vapeur. La pression de la chaudière baissant, la vitesse diminue. Le problème était alors de maintenir la vitesse de la machine constante malgré les variations de charge. Le mathématicien et astronome anglais Georges Airy fut le premier à tenter une analyse du régulateur à boules de Watt. Ce n'est qu'en 1868, que le physicien écossais James Clerk Maxwell publia une première analyse mathématique convaincante et expliqua ainsi certains comportement erratiques observés parmi les nombreux régulateurs en service à cet époque. Ses travaux furent le point de départ de nombreux autres sur la stabilité, sa caractérisation ayant été obtenue indépendamment par les mathématiciens A. Hurwitz et E.J. Routh.

Durant les années 1930, les recherches aux "Bell Telephone Laboratories" sur les amplificateurs sont à l'origine d'idées encore enseignées aujourd'hui. Citons par exemple les travaux de Nyquist et Bode caractérisant à partir de la réponse fréquentielle en boucle ouverte celle de la boucle fermée. Pendant la seconde guerre mondiale, ces techniques furent utilisées et très activement développées en particulier lors de la mise au point de batteries anti-aériennes. Le mathématicien Nobert Wiener a donné le nom de "cybernétique" à toutes ces techniques.

Tous ces développements se faisaient dans le cadre des systèmes linéaires avec une seule commande et une seule sortie : on disposait d'une mesure sous la forme d'un signal électrique. Cette dernière était alors entrée dans un amplificateur (un circuit électrique) qui restituait en sortie un autre signal électrique que l'on utilisait alors comme signal de contrôle. Ce n'est qu'après les années 50 que les développements théoriques et techno-

logiques (calculateurs numériques) permirent le traitement des systèmes multi-variables linéaires et non linéaires avec plusieurs entrées et plusieurs sorties. Citons comme contributions importantes dans les années 60 celles de Richard Bellmann avec la programmation dynamique, celles de Rudolf Kalman avec le filtrage et la commande linéaire quadratique et celles de L. Pontryagin avec la commande optimale.

Ces contributions continuent encore aujourd'hui à alimenter les recherches en théorie des systèmes. L'objectif de ce cours est double : d'une part présenter des notions et outils fondamentaux; d'autre part d'exposer des méthodes analytiques et numériques utiles pour les applications.

Nous vous serions reconnaissants de nous faire part de vos critiques et des erreurs que vous auriez découvertes par un message explicatif à `frederic.bonnans@inria.fr` ou à `pierre.rouchon@ensmp.fr` en identifiant votre message par "Poly X corrections".

Frédéric Bonnans et Pierre Rouchon
Septembre 2003

Table des matières

I Stabilité, Commandabilité et Observabilité 9

1 Introduction 11
1.1 Un exemple emprunté à la robotique . . . 11
1.2 Le plan . . . 16
1.3 Problème . . . 16

2 Étude de cas 19
2.1 Le bio-réacteur . . . 19
2.1.1 Étude à $D > 0$ fixé . . . 20
2.1.2 Stabilisation (globale) par feedback (borné) . . . 24
2.2 L'avion à décollage vertical . . . 26
2.2.1 Modèle de simulation . . . 27
2.2.2 Modèle de commande . . . 28
2.2.3 Commande linéaire . . . 28
2.2.4 Commande non-linéaire . . . 31
2.3 Pendule inversé sur un rail . . . 32
2.4 Moteur électrique à courant continu . . . 34
2.4.1 Stabilité en boucle ouverte . . . 35
2.4.2 Estimation de la vitesse et de la charge . . . 35
2.4.3 Le contrôleur . . . 36
2.4.4 L'observateur-contrôleur . . . 37
2.4.5 Robustesse par rapport à la dynamique rapide du courant . . . 37
2.4.6 Boucle rapide et contrainte de courant . . . 38

3 Systèmes dynamiques explicites 41
3.1 Espace d'état, champ de vecteurs et flot . . . 41
3.1.1 Un modèle élémentaire de population . . . 41
3.1.2 Existence, unicité, flot . . . 43
3.1.3 Remarque sur l'espace d'état . . . 51
3.1.4 Résolution numérique . . . 52
3.1.5 Comportements asymptotiques . . . 53
3.1.6 L'étude qualitative ou le contenu des modèles . . . 56
3.2 Points d'équilibre . . . 56
3.2.1 Stabilité et fonction de Lyapounov . . . 57
3.2.2 Les systèmes linéaires . . . 62
3.2.3 Lien avec le linéaire tangent . . . 65

3.3 Systèmes dynamiques discrets . . . 67
3.3.1 Point fixe et stabilité . . . 68
3.3.2 Les systèmes linéaires discrets . . . 68
3.4 Stabilité structurelle et robustesse . . . 69
3.5 Théorie des perturbations . . . 72
3.5.1 Les perturbations singulières . . . 73
3.5.2 Moyennisation . . . 77
3.6 Problèmes . . . 80

4 Commandabilité et observabilité 83
4.1 Commandabilité non linéaire . . . 84
4.1.1 Définition . . . 84
4.1.2 Intégrale première . . . 85
4.2 Commandabilité linéaire . . . 86
4.2.1 Matrice de commandabilité . . . 87
4.2.2 Invariance . . . 88
4.2.3 Un exemple . . . 90
4.2.4 Critère de Kalman et forme de Brunovsky . . . 91
4.2.5 Planification et suivi de trajectoires . . . 94
4.2.6 Linéarisation par bouclage . . . 96
4.3 Observabilité non linéaire . . . 100
4.3.1 Définition . . . 101
4.3.2 Critère . . . 101
4.3.3 Observateur, estimation, moindre carré . . . 103
4.4 Observabilité linéaire . . . 104
4.4.1 Le critère de Kalman . . . 104
4.4.2 Observateurs asymptotiques . . . 106
4.4.3 Observateur réduit de Luenberger . . . 107
4.5 Observateur-contrôleur linéaire . . . 107
4.6 Problèmes . . . 108

5 Annexe: Systèmes semi-implicites et inversion 115
5.1 Systèmes semi-implicites . . . 117
5.1.1 Un exemple . . . 117
5.1.2 Le cas général . . . 119
5.1.3 Linéaire tangent . . . 123
5.1.4 Résolution numérique . . . 124
5.2 Inversion et découplage . . . 125
5.2.1 Un exemple . . . 125
5.2.2 Le cas général . . . 127

Bibliographie commentée de la partie I 133

II Méthodes Numériques en Commande Optimale 137

1 Temps minimal : systèmes linéaires 139
1.1 Introduction . . . 139
1.2 Un problème d'alunissage . . . 139
1.3 Existence de solutions . . . 141
1.3.1 Position du problème . . . 141
1.3.2 Résultats d'existence . . . 142
1.4 Conditions d'optimalité . . . 143
1.4.1 Séparation de l'ensemble accessible de la cible . . . 143
1.4.2 Critère linéaire sur l'état final . . . 145
1.4.3 Etat adjoint et principe du minimum . . . 148
1.5 Exemples et classes particulières . . . 149
1.5.1 Contraintes de bornes sur la commande . . . 149
1.5.2 Cas de l'oscillateur harmonique . . . 151
1.5.3 Stabilisation d'un pendule inversé . . . 152
1.5.4 Cibles épaisses . . . 154

2 Temps minimal : systèmes non linéaires 157
2.1 Présentation du problème . . . 157
2.1.1 Un exemple . . . 157
2.1.2 Spécification du problème . . . 158
2.1.3 Existence de solutions . . . 158
2.2 Conditions d'optimalité . . . 159
2.2.1 Un résultat général . . . 159
2.2.2 Arc singulier . . . 160
2.3 Applications . . . 163
2.3.1 Pendule . . . 163
2.3.2 Avion à trajectoire horizontale . . . 164
2.4 Démonstration du résultat principal . . . 165
2.5 Notes . . . 171

3 Commande optimale : l'approche HJB 173
3.1 Cadre . . . 173
3.2 Valeur fonction de l'état . . . 174
3.2.1 Principe de programmation dynamique . . . 174
3.2.2 Equation de Hamilton-Jacobi-Bellman . . . 176
3.2.3 Continuité uniforme de la valeur . . . 178
3.3 Commande optimale . . . 179
3.4 Solution de viscosité . . . 181
3.4.1 Notion de solutions de viscosité . . . 181
3.4.2 Théorème de comparaison . . . 183
3.5 Temps d'arrêt et commande impulsionnelle . . . 186
3.5.1 Problèmes avec temps d'arrêt . . . 186
3.5.2 Commande impulsionnelle . . . 188
3.6 Notes . . . 190

4 Résolution numérique de l'équation HJB **191**
4.1 Motivation : problème continu . . . 191
4.2 Schémas décentrés et extensions . . . 192
4.2.1 Dimension d'espace $n = 1$ . . . 192
4.2.2 Forme de point fixe contractant . . . 193
4.2.3 Dimension d'espace quelconque . . . 195
4.2.4 Discrétisation par triangulation . . . 196
4.3 Convergence des schémas et essais numériques . . . 197
4.3.1 Un argument élémentaire de convergence . . . 197
4.3.2 Estimation d'erreur . . . 199
4.3.3 Equation eikonale . . . 201
4.3.4 Problème d'alunissage . . . 202
4.4 Notes . . . 203

5 Commande optimale stochastique **205**
5.1 Chaînes de Markov commandées . . . 205
5.1.1 Quelques exemples . . . 205
5.1.2 Chaînes de Markov et valeurs associées . . . 205
5.1.3 Quelques lemmes . . . 207
5.1.4 Principe de Programmation dynamique . . . 208
5.1.5 Problèmes à horizon infini . . . 209
5.1.6 Algorithmes numériques . . . 210
5.1.7 Problèmes de temps de sortie . . . 212
5.1.8 Problèmes avec décision d'arrêt . . . 213
5.1.9 Un algorithme implémentable . . . 214
5.2 Problèmes en temps et espace continus . . . 217
5.2.1 Position du problème . . . 217
5.2.2 Problème discrétisé en temps . . . 217
5.2.3 Schémas monotones : dimension 1 . . . 219
5.2.4 Différences finies classiques . . . 221
5.2.5 Différences finies généralisées . . . 223
5.2.6 Analyse de la condition de consistance forte . . . 225
5.3 Notes . . . 226

Bibliographie de la partie II **229**

Première partie

Stabilité, Commandabilité et Observabilité

Chapitre 1

Introduction

1.1 Un exemple emprunté à la robotique

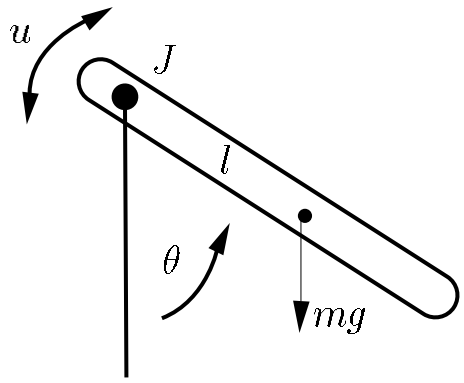


FIG. 1.1 – *un bras de robot tournant dans un plan vertical autour d'un axe horizontal motorisé.*

Modélisation Commençons par l'exemple de la figure 1.1 emprunté à la robotique. Il s'agit d'un bras rigide tournant dans un plan vertical autour d'un axe horizontal. Cet axe horizontal est équipé d'un moteur délivrant un couple variable u, que l'on peut choisir arbitrairement : u est la *commande* du système (on dit aussi *l'entrée*). La position géométrique du système est complètement décrite par un angle $\theta \in \mathbb{S}^1$ (l'espace des configurations géométriques du système est le cercle $\mathbb{S}^1$). La conservation du moment cinétique autour de l'axe horizontal permet de relier l'angle θ à la commande en couple u par l'équation différentielle du second ordre suivante :

$$J\ddot{\theta}(t) + mlg \sin\theta(t) = u(t) \tag{1.1}$$

où m est la masse du bras, J son moment d'inertie par rapport à l'axe, l la distance du centre de gravité à l'axe et g l'accélération due à la pesanteur.

Forme d'état Fixons un intervalle de temps $[0,T]$. La commande $[0,T] \ni t \mapsto u(t)$ étant fixée, nous obtenons la loi horaire $[0,T] \ni t \mapsto \theta(t)$ en intégrant cette équation du

second ordre à partir de conditions initiales en position $\theta(0) = \theta_0$ et en vitesse $\dot{\theta}(0) = \dot{\theta}_0$. L'ensemble des conditions initiales forme l'état du système (l'espace des phase en mécanique). Cela revient en fait à réécrire cette équation scalaire du second ordre en deux équations scalaires du premier ordre :

$$\begin{array}{rcl} \dot{\theta} & = & \omega \\ \dot{\omega} & = & u/J - (mgl/J)\sin\theta. \end{array} \tag{1.2}$$

Les variables (θ,ω) forment alors *l'état* du système; le triplé $t \mapsto (\theta(t),\omega(t),u(t))$ sera dit *trajectoire* du système s'il vérifie, pour tout t, les deux équations différentielles (1.2).

Commandabilité La *planification de trajectoires* consiste à trouver une trajectoire du système $t \mapsto (\theta(t),\omega(t),u(t))$ partant d'un état (θ_i,ω_i) en $t = 0$ et arrivant en $t = T$ à l'état final (θ_f,ω_f), ces deux états étant fixés par avance. Il s'agit du problème de base de la commandabilité : comment amener le système d'un endroit (d'un état) à un autre. Lorsque le système est commandable, on dispose, en général, d'une infinité de trajectoires et donc de commandes pour réaliser cette transition. Se pose alors le problème du choix entre ces diverses trajectoires : c'est en autre l'objet de la commande optimale qui sélectionne la trajectoire qui minimise un certain critère. Citons par exemple le temps minimum pour aller d'une position de repos $(\theta_i,\omega_i = 0)$ à une autre position de repos $(\theta_f,\omega_f = 0)$ sachant que la commande u reste bornée ($\forall t$, $|u(t)| \leq u_{max}$ où u_{max} est le couple maximum développé par le moteur). On en déduit ainsi une trajectoire de référence du système : $[0,T] \ni t \mapsto (\theta_r(t),\omega_r(t),u_r(t))$.

Bouclage Une autre question, directement liée à la première : étant donné que tout modèle est approximatif (les paramètres J et m et l sont connus avec une certaine précision), il convient d'ajuster la commande u en temps réel de façon à compenser les écarts à la trajectoire de référence, $\theta - \theta_r$ et $\omega - \omega_r$, qui peuvent apparaître. Il s'agit du *suivi de trajectoire* ("tracking" en anglais). Lorsque cette trajectoire est un point d'équilibre du système (comme, par exemple $(\theta,\omega,u) = 0$ ou $(\theta,\omega,u) = (\pi,0,0)$) on parle alors de *stabilisation*. Noter que la stabilisation du système autour d'une trajectoire $t \mapsto (\theta_r(t),\omega_r(t),u_r(t))$ qui n'est pas une trajectoire du système, i.e. qui ne vérifie pas les deux équations de (1.2) n'a aucun sens. En particulier, on ne peut pas parler de stabilisation autour d'un état qui n'est pas un état d'équilibre (comme, par exemple, $\theta_r = \pi/2$ et $\omega_r = 1$). Une démarche très naturelle consiste à corriger la commande de référence $u_r(t)$ par des termes du type $\theta - \theta_r(t)$ et $\omega - \omega_r(t)$. L'utilisation de ce type de terme correspond à un bouclage d'état, une boucle de rétro-action ("feedback" en anglais) qui l'on schématise souvent par le diagramme bloc de la figure 1.2. La mise en oeuvre de ce schéma revient, avec un calculateur temps-réel, à mettre à jour très rapidement (avec une période d'échantillonnage T_e bien plus rapide que les échelles de temps naturelles du système) la commande u en fonction de la trajectoire de référence et des mesures de θ et de ω.

Linéaire tangent Considérons, par exemple la stabilisation autour de l'équilibre instable $(\theta,\omega,u) = (\pi,0,0)$. Pour cela, linéarisons les équations (1.2) autour de ce point : nous faisons un développement limité des seconds membres en ne retenant que les termes

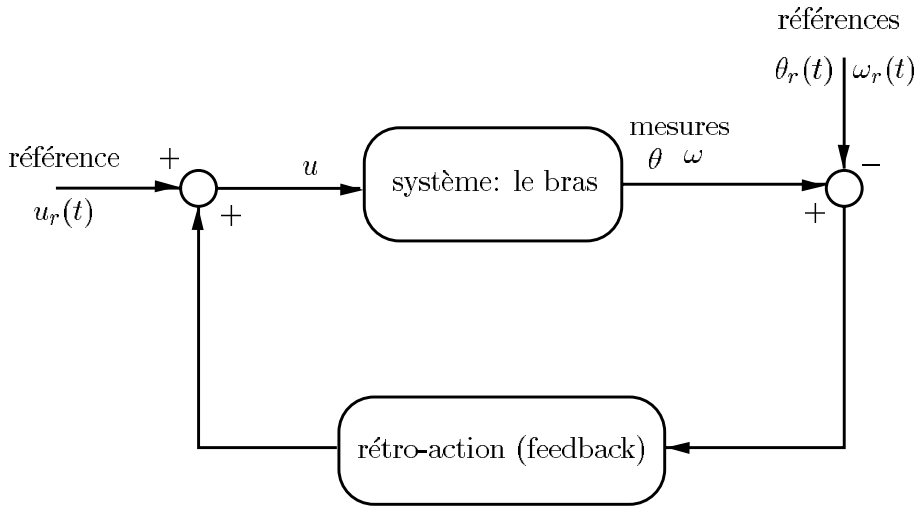


FIG. 1.2 – *schéma-bloc d'une loi de rétro-action, dit aussi retour d'état ou "feedback".*

d'ordre 1 (ceux d'ordre 0 sont nuls, car nous sommes autour d'un point d'équilibre). En notant $\tilde{\theta}$, $\tilde{\omega}$ et $\tilde{u}$ les écarts, nous obtenons les équations du système *linéarisé tangent*:

$$\begin{array}{rcl} \dot{\tilde{\theta}} & = & \tilde{\omega} \\ \dot{\tilde{\omega}} & = & \tilde{u}/J + (mgl/J)\tilde{\theta}. \end{array} \tag{1.3}$$

Stabilisation Si on pose, comme loi de commande, le retour statique d'état (feedback),

$$\tilde{u} = -(Jk_2 + mgl)\tilde{\theta} - Jk_1\tilde{\omega}, \tag{1.4}$$

avec k_1 et k_2 les gains du contrôleur (paramètres constants que nous choisirons ci-dessous), alors les équations du système linéaire tangent bouclé (c'est à dire avec sa boucle de rétro-action) sont

$$\begin{array}{rcl} \dot{\tilde{\theta}} & = & \tilde{\omega} \\ \dot{\tilde{\omega}} & = & -k_2\tilde{\theta} - k_1\tilde{\omega}. \end{array}$$

Avec $k_1 = (1/\tau_1 + 1/\tau_2)$ et , $k_2 = 1/(\tau_1\tau_2)$, où $0 < \tau_1 < \tau_2$ sont des temps caractéristiques, ce système devient *asymptotiquement stable*: pour toutes conditions initiales, ses solutions tendent vers zéros lorsque t tend vers l'infini. La convergence est même exponentielle: toute solution est une combinaison linéaire de $\exp(-t/\tau_1)$ et $\exp(-t/\tau_2)$: $-1/\tau_1$ et $-1/\tau_2$ sont appelés les *pôles* du système bouclé.

Robustesse et théorie des perturbations Ce bouclage a été réalisé sur une approximation au premier ordre du système. Se pose alors la question du comportement du système non linéaire (1.2) avec le bouclage linéaire $u = \tilde{u} = -(Jk_2 + mgl)(\theta - \pi) - Jk_1\omega$. Il est immédiat de voir que le linéarisé tangent autour de l'équilibre (π,0) du système non linéaire bouclé est identique au linéaire tangent bouclé. Un résultat classique sur la stabilité structurelle des points d'équilibres hyperboliques (les valeurs propres de la matrice jacobienne sont toutes à partie réelle non nulle) d'un système dynamique garantit alors la stabilité asymptotique locale du système non linéaire bouclé: cela veut dire simplement

que toute trajectoire du système (1.2) avec la commande (1.4) qui démarre assez près de (π,0) tends vers (π,0) lorsque t tends vers l'infini, la convergence étant exponentielle, comme pour le linéaire tangent.

Ètant donné que τ_1 et τ_2 sont deux constantes de temps arbitraires directement liées au taux de convergence, on aura tendance à les choisir aussi proches de zéro que possible. Cependant, il convient de ne pas les choisir trop proches de zéro : en effet, le modèle sur lequel la commande est synthétisée, n'est valable que pour une certaine gamme d'échelles de temps. Le modèle n'est pas valable pour des fréquences grandes. En effet, la dynamique du moteur est négligée: pour un moteur à courant continu, la commande physique est en fait la tension U_m appliquée au moteur. Elle est reliée au couple u par une équation différentielle du type :

$$L\dot{I}_m + RI_m = U_m, \quad u = K_c I_m \tag{1.5}$$

(L est l'inductance, R la résistance, K_c la constante de couple du moteur). En pratique la dynamique du moteur est souvent négligeable par rapport à la dynamique inertielle de la barre. Ainsi la constante de temps du moteur $\tau_m = L/R$ est bien inférieure au temps caractéristique du bras $\tau_b = \sqrt{J/(mlg)}$. Aussi, a-t-on l'approximation suivante dite quasi-statique :

$$RI_m = U_m, \quad u = K_c I_m = (K_c/R)U_m$$

qui relie directement le couple u à la tension U_m. Il convient de choisir τ_1 et τ_2 du même ordre de grandeur que τ_b, et donc très supérieur à τ_m, la constante de temps de la dynamique négligée.

D'autres phénomènes peuvent apparaître vers les hautes fréquences, comme la flexibilité du bras. Nous verrons dans le chapitre 3 un résultat asymptotique (théorie de perturbations, systèmes lents/rapides) assurant qu'avec des gains k_1 et k_2 pas trop grands (τ_1,$\tau_2 \gg \tau_m$), le système non linéaire avec la dynamique du moteur (1.5) et la commande en tension

$$U_m = -(K_c/R)\ ((Jk_1 - mgl)(\theta - \pi) + Jk_2\omega)$$

est localement asymptotiquement stable autour de (π,0), pour toute valeur > 0 de L assez faible.

Observabilité La loi de feedback précédente suppose que l'on mesure à chaque instant l'état complet du système θ et ω. Si nous connaissons uniquement la loi $t \mapsto \theta(t)$, nous obtenons $\omega(t)$ par simple dérivation : on dit que l'état du système est *observable* à partir de la sortie θ. D'une façon plus générale, l'état x d'un système sera dit observable à partir de la *sortie* y, si l'on peut reconstruire x à partir d'un nombre fini de dérivées de y.

Pour le bras, nous pouvons dériver numériquement le signal de mesure pour en déduire ω. Cette solution fonctionne correctement si la mesure de θ n'est pas trop bruitée. Sinon, l'opération de dérivation est à éviter. Pour cela, nous pouvons utiliser la dynamique du système pour construire un observateur asymptotique, c'est à dire, reconstruire la vitesse ω du système en intégrant (on peut dire aussi en filtrant) la position θ via une équation différentielle bien choisie. On obtient alors un filtre causal qui élimine les hautes fréquences à la fois sur la mesure et ses dérivées sans introduire de retard sur la partie basse fréquence des signaux.

Plaçons nous autour du point (π,0) et considérons le linéaire tangent (1.3) avec comme quantités connues la commande $\tilde{u}$ et l'angle $\tilde{\theta}$. L'objectif est de reconstruire à terme $\tilde{\omega}$

sans utiliser l'opération de dérivation très sensible au bruit. En revanche nous pouvons utiliser l'intégration et les changements de variables.

Nous allons montrer comment construire un *observateur asymptotique* (d'ordre réduit). Soit λ un paramètre que nous ajusterons plus tard. Considérons la variable $\xi = \tilde{\omega} + \lambda\tilde{\theta}$. Si l'on sait reconstruire ξ, on obtient $\tilde{\omega}$ avec $\tilde{\omega} = \xi - \lambda\tilde{\theta}(t)$. Or, grâce à (1.3), ξ vérifie

$$\dot{\xi} = \tilde{u}/J - (mgl/J)\tilde{\theta} + \lambda\tilde{\omega} = \tilde{u}/J - (mgl/J + \lambda^2)\tilde{\theta} + \lambda\xi.$$

Ainsi, en recopiant cette équation et en remplaçant la variable ξ non mesurée par $\hat{\xi}$, on obtient une équation différentielle du premier ordre dépendant des quantités connues $\tilde{u}$ et $\tilde{\theta}$ (un filtre d'ordre 1 d'une combinaison linéaire de la mesure $\tilde{\theta}$ et de la commande $\tilde{u}$) :

$$\dot{\hat{\xi}} = \tilde{u}(t)/J - (mgl/J + \lambda^2)\tilde{\theta}(t) + \lambda\hat{\xi} \tag{1.6}$$

Par soustraction avec l'équation différentielle satisfaite par le vrai ξ, les termes sources en $\tilde{u}$ et $\tilde{\theta}$ disparaissent. On obtient alors une dynamique de l'erreur $\hat{\xi} - \xi$ autonome

$$\frac{d}{dt}(\hat{\xi} - \xi) = \lambda(\hat{\xi} - \xi)$$

qui converge vers zéro, quelque soit la condition initiale sur $\hat{\xi}$, dès que le paramètre $\lambda = -1/\tau_f$ est choisi négatif (τ_f est la constante de temps de l'observateur (1.6)). Là encore, le gain λ de l'observateur (1.6) doit être choisi en fonction des niveaux de bruit sur θ et surtout des échelles de temps naturelles du système (prendre, par exemple, τ_f du même ordre de grandeur que $\tau_b = \sqrt{J/(mlg)}$).

Observateur-contrôleur, principe de séparation Ainsi, en combinant l'observateur (1.6) et la commande (1.4) où $\tilde{\omega}$ est remplacé par $\hat{\xi} - \lambda\tilde{\theta}$, nous obtenons un bouclage qui stabilise localement la position inverse du pendule. Ce bouclage est un *bouclage dynamique* sur la sortie $y = \tilde{\theta}$: le terme dynamique vient du fait que la commande u est une fonction de θ et de $\hat{\xi}$ qui est en fait une sorte "d'intégrale" de u et θ :

$$\begin{array}{rcl} \dot{\hat{\xi}} & = & \tilde{u}/J - (mgl/J + \lambda^2)\tilde{\theta} + \lambda\hat{\xi} \\ \tilde{u} & = & -(Jk_1 + mgl)\tilde{\theta} - Jk_2(\hat{\xi} - \lambda\tilde{\theta}). \end{array} \tag{1.7}$$

Il est alors très simple d'utiliser ces deux équations pour obtenir un algorithme temps-réel de stabilisation. Reprenons le schéma de la figure 1.2 et intéressons nous à la boucle de rétro-action. Notons T_e la période d'échantillonnage supposée petite, $\tilde{u}_n$ la valeur de la commande à $t = nT_e$, $\hat{\xi}_n$ la valeur de l'état interne du contrôleur et $\tilde{\theta}_n$ la mesure. Alors $\tilde{u}_{n+1}$ et $\hat{\xi}_{n+1}$ sont obtenus par récurrence en remplaçant $\dot{\hat{\xi}}$ dans (1.7) par $(\hat{\xi}_{n+1} - \hat{\xi}_n)/T_e$:

$$\begin{array}{rcl} \hat{\xi}_{n+1} & = & \hat{\xi}_n + T_e(\tilde{u}_n/J - (mgl/J + \lambda^2)\tilde{\theta}_n + \lambda\hat{\xi}_n) \\ \tilde{u}_{n+1} & = & -(Jk_1 + mgl)\tilde{\theta}_n - Jk_2(\hat{\xi}_n - \lambda\tilde{\theta}_n) \end{array}$$

Ainsi $\hat{\xi}_{n+1}$ est gardé en mémoire pour la commande suivante et $\tilde{u}_{n+1}$ est appliquée au système.

Nous n'abordons pas ici des problèmes liés à l'échantillonnage. Nous resterons au niveau continu, sachant que la mise en oeuvre est possible dès que la période T_e est très petite devant les échelles de temps du système et que les micro-processeurs sont suffisamment rapides pour calculer la nouvelle commande en un temps inférieur à T_e.

1.2 Le plan

Cet exemple permet de se faire une idée des techniques présentées dans cette première partie du cours. Nous allons maintenant reprendre de façon plus systématique et rigoureuse les divers points évoqués ci-dessus. Le chapitre 2 est constitué de 4 études de cas. Chaque cas reprend et applique les méthodes et notions fondamentales présentées dans leur généralité au niveau les chapitres 3 et 4.

Nous abordons dans le chapitre 3 les systèmes dynamiques explicites et, sans faire toutes les démonstrations, quelques résultats sur les équations différentielles ordinaires (problème de Cauchy, perturbation régulière, singulière, systèmes lents/rapides, stabilité au sens de Lyapounov) : ces résultats sont essentiels pour bien comprendre, entre autres, les liens entre le linéaire tangent et le système non linéaire associé, les questions de robustesse par rapport aux erreurs de modèle et aux dynamiques négligées.

Dans le chapitre 4, nous abordons la commandabilité et l'observabilité des systèmes explicites $\dot{x} = f(x,u)$. Après de courtes définitions, nous étudions les systèmes linéaires stationnaires. Nous mettons l'accent sur la forme canonique de Brunovsky, la planification de trajectoires, et la stabilisation par placement de pôle. Nous aborderons ensuite l'observabilité, qui peut être vue, pour les systèmes linéaires à coefficients constants, comme le problème dual de la commandabilité, la construction de bouclages stabilisants conduisant à celle d'observateurs asymptotiques. Enfin, nous terminons ce chapitre par le principe de séparation et la synthèse d'un bouclage dynamique de sortie (on dit aussi observateur-contrôleur ou commande modale).

La présentation s'appuie souvent sur des exemples. En général, ces exemples sont représentatifs de questions préoccupant les ingénieurs. Des exercices jalonnent également l'exposé. Ils sont souvent là pour suggérer au lecteur des extensions à des situations plus générales (non linéaire, dimension infinie, systèmes discrets, ...). Les parties écrites en petits caractères peuvent être ignorées dans une première lecture : il s'agit soit de compléments, soit de prolongements.

Dans l'annexe 5 nous étudions les systèmes semi-implicites. Ce chapitre peut être sauté dans une première lecture. En général, la modélisation d'un système dynamique complexe ne conduit pas directement à des équations différentielles explicites mais à un système mixte d'équations différentielles et d'équations algébriques. Des manipulations formelles sont alors nécessaires pour mettre le système sous forme explicite. Ces manipulations utilisent des dérivations, *l'index* étant alors le nombre minimal de dérivations nécessaires. Il s'avère que les techniques utilisées ici sont très proches de celles employées pour *l'inversion*, le *découplage* et la *linéarisation entrée/sortie* : tout repose sur un algorithme d'élimination différentielle dit *algorithme de structure*. La rédaction de ce chapitre s'appuie fortement sur deux exemples clés : (5.3) page 117 et (5.5) page 125. Leur compréhension implique pratiquement celle du cas général qui n'est guère plus compliqué.

1.3 Problème

On reprend ici, sous la forme d'un problème et pour l'étendre au non linéaire, l'observateur-contrôleur que nous avons construit avec le linéaire tangent du système (1.2). L'objectif de commande est d'aller du point d'équilibre $\theta = 0$ au point d'équilibre $\theta = \pi$ pendant le

temps $T > 0$ en ne mesurant que θ. Cette extension ne nécessite que très peu de calculs et reste à un niveau de complexité très élémentaire.

1. Donner une trajectoire du système $[0,t] \ni t \mapsto (\theta_r(t),\omega_r(t),u_r(t))$ qui assure cette transition.
2. Calculer le bouclage d'état qui stabilise la dynamique de l'erreur à la trajectoire $e = \theta - \theta_r(t)$ de la façon suivante :

$$\ddot{e} + \sigma_1 \dot{e} + \sigma_2 e = 0$$

   avec $\sigma_1,\sigma_2 > 0$.
3. On suppose que l'on ne mesure que θ. Montrer que l'observateur non linéaire

$$\dot{\hat{\xi}} = \lambda\hat{\xi} - \lambda^2\theta(t) + u(t)/J - (mgl/J)\sin\theta(t)$$

   permet de reconstruire asymptotiquement ω par $\hat{\omega} = \hat{\xi} - \lambda\theta$, dès que $\lambda < 0$.
4. Montrer la convergence de l'observateur-contrôleur où l'on a remplacé la mesure de vitesse ω dans la question 2 par l'estimée $\hat{\omega}$ de la question 3.
5. Faire des simulations de cette manoeuvre en $T = 5$ s en prenant comme paramètres $m = 1{,}0$ kg, $l = 0{,}2$ m, $J = 0{,}1$ kg m^2 et $g = 9{,}81$ m/s^2. Tester la robustesse de cette commande dynamique de sortie par rapport à des dynamiques négligées (rajouter une dynamique pour le moteur) et par rapport à des erreurs dans le modèle (1.2) (rajouter un petit frottement au niveau de l'axe du bras).

Chapitre 2

Étude de cas

A travers l'étude détaillée de plusieurs cas, un bio-réacteur, l'avion à décollage vertical, le pendule inversé et le moteur électrique, nous reprenons diverses notions fondamentales comme la stabilité, la commandabilité, l'observabilité ainsi que les techniques de base comme le bouclage (feedback), l'observateur asymptotique, la planification et le suivi de trajectoire. A chaque fois nous renvoyons le lecteur à une partie précise du cours où la formalisation et les définitions sont disponibles. Les réponses que nous apportons ici ne sont bien sûr pas les seules possibles. Elles ont cependant le mérite d'être simples, explicites et directement exploitables sur un calculateur temps-réel. Enfin certaines questions très naturelles et pourtant sans réponse systématique sont évoquées. En particulier, le respect de contraintes sur la commande et sur l'état est traité par des méthodes très spécifiques.

2.1 Le bio-réacteur

Nous avons choisi un bio-réacteur car ce système est représentatif d'un vaste domaine: les procédés de transformation de la matière. Leur modélisation dynamique s'appuie sur les lois de conservation matière et énergie, les lois cinétiques et la thermodynamique. Des secteurs industriels majeurs utilisent des installations de ce type: pétrole, pétro-chimie, plastique, chimie fine, pharmacie, biotechnologie, agro-alimentaire, ...

Nous reprenons avec cet exemple certaines notions importantes sur les systèmes dynamiques et leur stabilité (point d'équilibre hyperbolique [définition 10, page 66], fonction de Lyapounov [théorème 3, page 58]). Après une étude du comportement qualitatif (géométrie des courbes intégrales, bifurcation) en fonction de la commande prise comme paramètre, nous montrons qu'un feedback très simple permet de stabiliser globalement le système.

Les équations régissant la dynamique d'un bio-réateur fonctionnant en continu sont, pour un métabolisme simple, les suivantes :

$$\begin{aligned} \dot{X} &= (\mu(S) - D)X \\ \dot{S} &= D(S_e - S) - \mu(S)X \end{aligned}$$

où X est la biomasse (les bestioles), S le taux de substrat carboné (le sucre), $D > 0$ le taux de dilution ($D = L/V$, V volume du fermenteur, L débit liquide entrant égal au débit sortant), $\mu(S)X$ la production de biomasse par unité de volume correspondant au

métabolisme $X + S \to 2X$, S_e le taux de sucre dans l'alimentation. On supposera que $S_e > 0$ est fixe et que $D > 0$ est la variable de réglage (la commande).

On suppose que la fonction $\mu(S)$ est régulière et admet une forme en cloche (cf figure 2.1) : μ est strictement croissante pour $S \in [0,\bar{S}]$ $(0 < \bar{S} < S_e)$; $\mu(0) = 0$; $\mu(\bar{S}) = \bar{\mu}$; μ est strictement décroissante pour $S \in [\bar{S},S_e]$ avec $\mu(S_e) > 0$.

2.1.1 Étude à $D > 0$ fixé

Trajectoires, espace invariant, flot

Il s'agit de l'étude en boucle ouverte. L'espace des états est ici $(X,S) \in [0,+\infty[\times[0,+\infty[$. Montrons que le modèle conduit à des concentrations positives. Si $X = 0$ alors $\dot{X} = 0$ et X reste toujours nul (pas de génération spontanée). Si $S = 0$ alors $\dot{S} = DS_e > 0$ et donc S a tendance à croître. Ainsi le champ de vecteurs

$$x = \begin{bmatrix} X \\ S \end{bmatrix} \mapsto v(x) = \begin{bmatrix} (\mu(S) - D)X \\ D(S_e - S) - \mu(S)X \end{bmatrix}$$

définissant la dynamique est rentrant dans $[0, +\infty[\times[0, +\infty[$. Les trajectoires $t \mapsto (X(t),S(t))$ sont "positives".

Montrons maintenant qu'elles sont définies sur $[0, +\infty[$. Soit la variable $\xi = S + X$, il est évident que, $\dot{\xi} = D(S_e - \xi)$. Ainsi dès que $\xi \geq S_e$, $\dot{\xi} \leq 0$. Cela signifie que v est rentrant dans tout domaine triangulaire T_a défini par

$$T_a = \{(X,S) \in [0, +\infty[\times[0, +\infty[\mid X + S \leq a\}$$

avec $a > S_e$. Comme toute condition initiale (X_0,S_0) dans $[0, +\infty[\times[0, +\infty[$ appartient à T_a avec $a = \max(S_e,X_0 + S_0)$, la trajectoire démarrant en (X_0,S_0) ne peut quitter T_a: un tel T_a est ainsi positivement invariant [définition 5, page 53]. Ainsi le flot [définition 1, page 44] ϕ_t est défini pour tout $t \geq 0$.

Notons aussi que le segment

$$\Delta = \{(X,S) \mid X + S = S_e,\ X \geq 0,\ S \geq 0\}$$

est aussi positivement invariant. Cela résulte du fait que $\dot{\xi} \equiv 0$ dès que $\xi_0 = S_e$. Nous avons même plus, comme $\xi(t) = S_e + \xi_0 \exp(-Dt)$, $\xi(t) \mapsto S_e$ quand $t \mapsto +\infty$. Ainsi, les trajectoires du système convergent toutes vers Δ. On peut montrer que cela reste vrai même si D est variable. La convergence est assurée dès que $\int_0^t D(\tau)\ d\tau \mapsto +\infty$ quand $t \mapsto +\infty$.

Points d'équilibre et exposants caractéristiques

Etudions maintenant en fonction de D les points d'équilibre. Ils sont définies par $v(x) = 0$, i.e.

$$\begin{aligned} (\mu(S) - D)X &= 0 \\ D(S_e - S) - \mu(S)X &= 0. \end{aligned}$$

La première équation se scinde en deux $X = 0$ et $D = \mu(S)$.

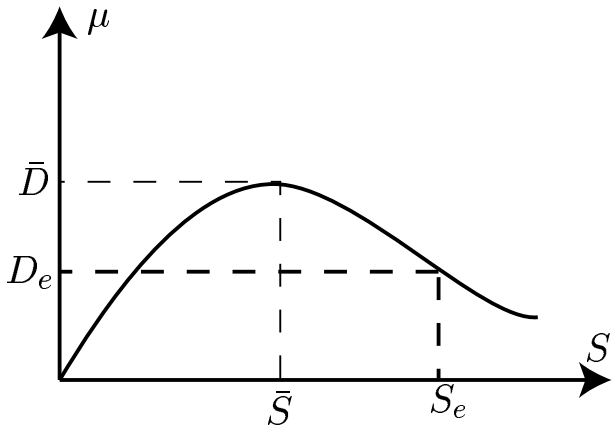


FIG. 2.1 – *La fonction* $\mu(S)$.

Le point d'équilibre $X = 0$ et $S = S_e$ correspond au lessivage du bio-réacteur. Aucune bio-masse n'est présente et le taux de sucre en sortie est celui de l'entrée.

Reste l'autre famille de solution. Il faut trouver S tel que $\mu(S) = D$. Comme le montre la figure 2.1, nous distinguons les trois cas suivants

1. $D \leq D_e = \mu(S_e)$: seule la racine $\leq \bar{S}$ de $\mu(S) = D$ donne un point d'équilibre physique avec $X > 0$ car $X = S_e - S$. Notons (X_s,S_s) ce point d'équilibre.
2. $D_e \leq D \leq \bar{D}$: deux points d'équilibre ayant un sens physique coexistent. Si $S_s \leq \bar{S}$ et $S_u \in [\bar{S},S_e]$ sont les deux racines de $\mu(S) = D$, on note $X_s = S_e - S_s$ et $X_u = S_e - S_u$.
3. $\bar{D} \leq D$: $\mu(S) = D$ n'admet pas de solution.

Etudions la stabilité de ces points d'équilibre. Nous savons [théorème 5, page 65] qu'elle est donnée par le signe de la partie réelle des valeurs propres du jacobien de v.

Pour $x = (0,S_e)$, nous avons,

$$Dv(x) = \begin{bmatrix} D_e - D & 0 \\ -D_e & -D \end{bmatrix}.$$

Pour $D < D_e$, $(0,S_e)$ admet une valeur propre stable $-D < 0$ et une valeur propre instable $D_e - D > 0$: c'est un col [figure 3.22, page 63]. L'équilibre est donc instable. Pour $D > D_e$, $(0,S_e)$ les deux valeurs propres sont stables $-D < 0$ et $D_e - D < 0$: c'est un noeud stable. L'équilibre est alors localement asymptotiquement stable.

Pour $D = D_e$, une valeur propre est stable $-D < 0$ l'autre est nulle : on ne peut pas conclure avec le linéaire tangent; ce n'est pas un point d'équilibre hyperbolique [définition 10, page 66].

Pour l'autre famille $x_s = (S_s,X_s)$ et $x_u = (S_u,X_u)$ de points d'équilibre, le jacobien de v vaut $(\alpha = s,u)$

$$Dv(x_\alpha) = \begin{bmatrix} 0 & \mu'(S_\alpha)X_\alpha \\ -D & -D - \mu'(S_\alpha)X_\alpha \end{bmatrix}.$$

Ces valeurs propres sont $-D$ et $-\mu'(S_\alpha)X_\alpha$. Ainsi l'équilibre $\alpha = s$ est toujours stable pour $D < \bar{D}$ (on suppose que μ' ne s'annule qu'en $S = \bar{S}$) : c'est un noeud stable. L'équilibre $\alpha = u$ est toujours instable pour $D_e \leq D < \bar{D}$: c'est un col.

Noter qu'en $D = \bar{D}$, ces deux branches d'équilibre se rejoignent avec comme valeur propre $-D$ et 0 : on ne peut rien dire sur la stabilité à partir du tangent. La valeur $D = \bar{D}$ est un valeur critique. Elle correspond à une bifurcation (bifurcation col-noeud classique), c'est à dire un changement qualitatif du portrait de phases. Dans les graphiques qui suivent nous avons, pour

$$\mu(S) = 2\bar{D}\frac{S/\bar{S}}{(S/\bar{S})^2 + 1}, \quad \bar{S} = 1, S_e = 3$$

tracé le champ de vecteurs ainsi que certaines trajectoires pour diverses valeurs de D.

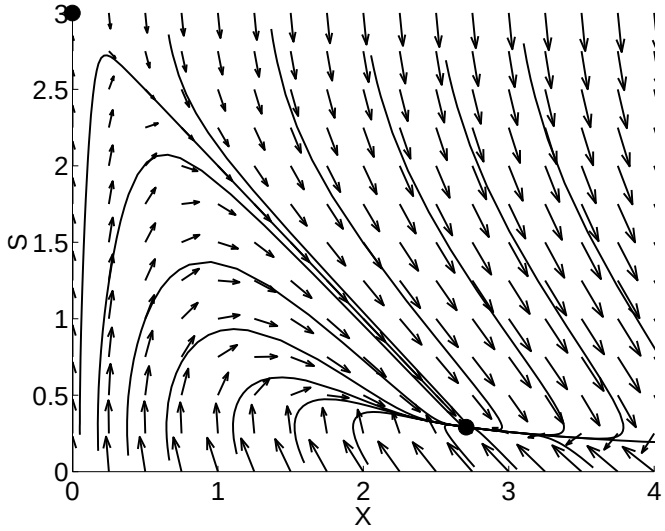


FIG. 2.2 – *portrait de phase en boucle ouverte $D < D_e$; deux points d'équilibres, un col et un noeud stable.*

Fonction de Lyapounov, stabilité asymptotique globale

Montrons que pour $D > \bar{D}$, les trajectoires convergent toutes vers le lessivage. A partir des valeurs propres du tangent nous savons déjà que cet équilibre est localement asymptotiquement stable.

Pour cela nous allons utiliser une méthode inventée par Lyapounov [théorème 3, page 58]. Considérons la fonction réelle

$$V(X,S) = \frac{1}{2}(X + S - S_e)^2 + \frac{1}{2}X^2.$$

Montrons que c'est une fonction de Lyapounov : V est infinie à l'infini dans l'orthant positif; V admet un seul minimum au lessivage $(0,S_e)$ et $\dot{V} \leq 0$ comme le montre ce qui suit.

En effet un calcul simple donne

$$\dot{V} = -D(X + S - S_e)^2 - (D - \mu(S))X^2.$$

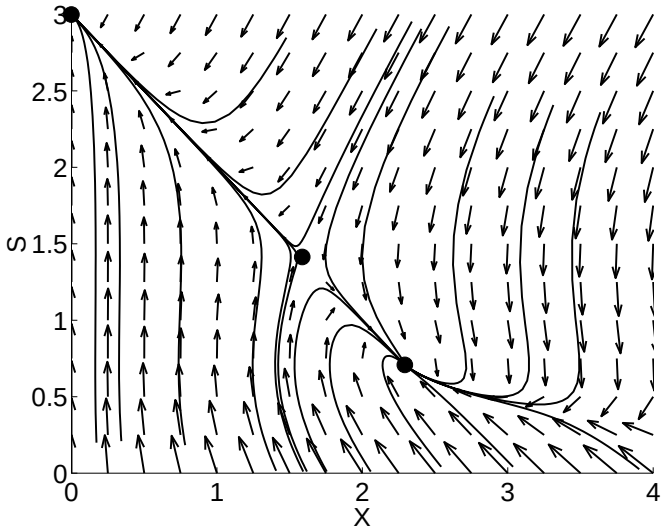


FIG. 2.3 – *portrait de phase en boucle ouverte $D_e < D < \bar{D}$; trois points d'équilibres, deux noeuds stables séparés par un col.*

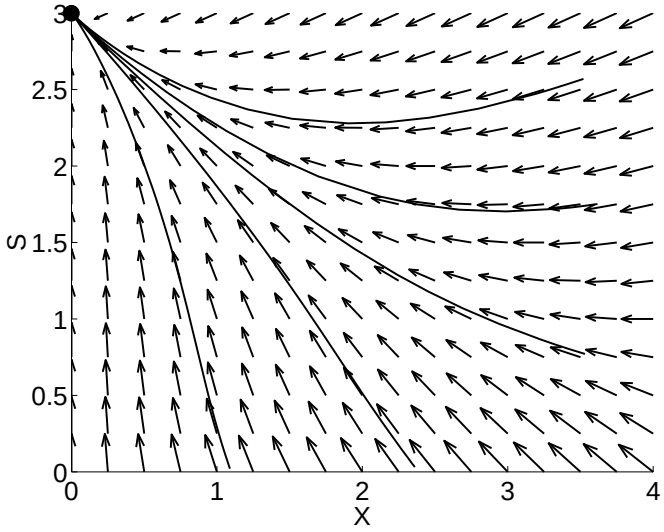


FIG. 2.4 – *portrait de phase en boucle ouverte $D > \bar{D}$; un seul point d'équilibre, le lessivage.*

Comme $\bar{D} = \sup(\mu)$, on a

$$\dot{V} \leq -D(X + S - S_e)^2 - (D - \bar{D})X^2.$$

Mais $D > \bar{D}$ donc $\dot{V} < 0$ dès que $(X,S) \neq (0,S_e)$. Ce qui montre la stabilité asymptotique globale de $(0,S_e)$. Noter l'interprétation géométrique de $\dot{V} \leq 0$. Le champ de vecteurs rentre dans les portions d'ellipses $V \leq cte$ du quart de plan positif (cf figure 2.5).

Pour $D = \bar{D}$, les calculs précédents restent valables : V reste une fonction de Lyapounov. Cependant $\dot{V}$ peut être nulle sans que nécessairement l'état soit $(0,S_e)$. Une étude plus fine à partir du principe d'invariance de Lasalle est nécessaire [théorème 3, page 58]. On sait que les trajectoires convergent alors vers le plus grand ensemble invariant contenu dans $\dot{V} = 0$. Ici cela donne donc le systèmes sur-déterminé suivant

$$\begin{aligned}
\dot{X} &= (\mu(S) - \bar{D})X \\
\dot{S} &= \bar{D}(S_e - S) - \mu(S)X \\
0 &= -(X + S - S_e)^2 - (\bar{D} - \mu(S))X^2.
\end{aligned}$$

Ses seules solutions sont les points d'équilibre $(0,S_e)$ et $(S_e - \bar{S},\bar{S})$. Ainsi les trajectoires convergent soit vers le lessivage soit vers le point d'équilibre interne correspondant au maximum de μ (cf, figure 2.6).

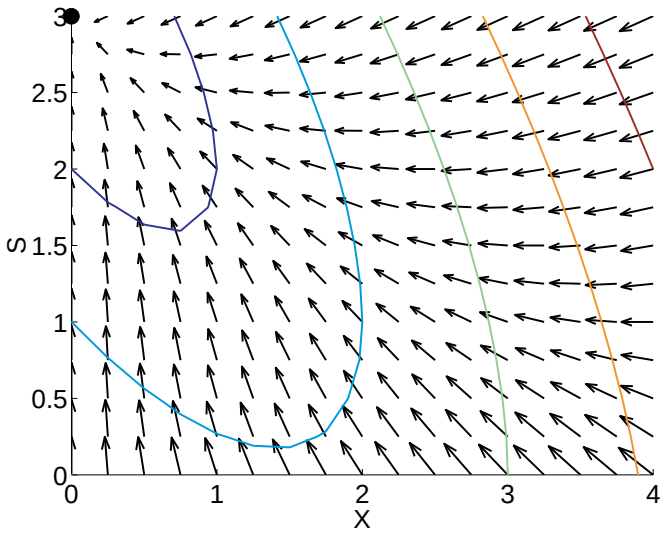


FIG. 2.5 – *V est une fonction de Lyapounov pour $D > \bar{D}$.*

2.1.2 Stabilisation (globale) par feedback (borné)

Le point d'équilibre double en $D = \bar{D}$ admet un intérêt pratique évident. C'est le seul point d'équilibre avec $X > 0$ et D le plus grand possible. Il correspond aussi au maximum du taux de croissant μ. Il est souvent intéressant de maintenir le système autour de ce régime.

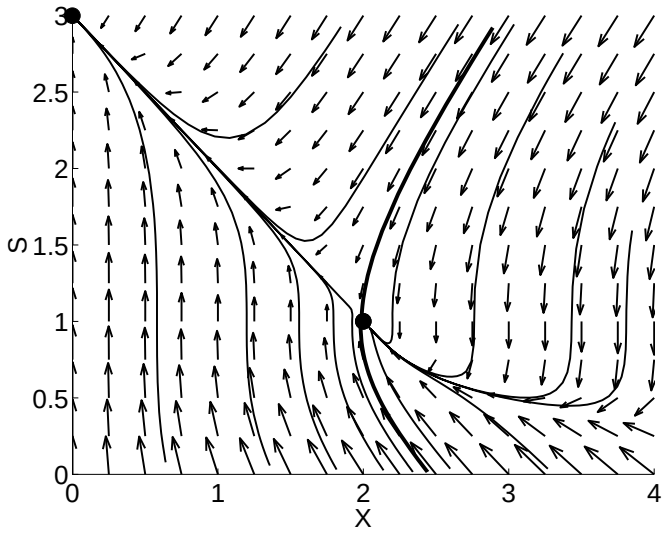


FIG. 2.6 – *portrait de phase en boucle ouverte pour $D = \bar{D}$; bifurcation col-noeud.*

Voyons si cela est possible sans rien faire, juste en maintenant D à $\bar{D}$. Alors le portrait de phase (figure 2.6) fait apparaître deux bassins d'attractions. Cependant, pour les trajectoires qui arrivent en $(S_e - \bar{S}, \bar{S})$, une petite perturbation suffit à les faire basculer dans le bassin d'attraction du lessivage $(0, S_e)$. Ainsi, même avec une condition initiale dans le bon bassin d'attraction, on aboutira toujours en pratique au lessivage. Une telle méthode n'est pas robuste (le point d'équilibre visé n'est pas asymptotiquement stable, le portrait de phase n'est pas structurellement stable [discussion de la section 3.4, page 69]). Il faut donc imaginer quelque chose pour maintenir les trajectoires autour de $(S_e - \bar{S}, \bar{S})$.

Nous allons voir que le simple régulateur proportionnel

$$D = \bar{D} - k(S - \bar{S})$$

avec un gain k bien choisi permet de stabiliser localement les trajectoires autour de $(S_e - \bar{S}, \bar{S})$. Il convient de bien comprendre la signification de $D = \bar{D} - k(S - \bar{S})$. Le taux de dilution (i.e., le débit d'entrée) varie en fonction de la valeur effective du taux de sucre dans le bio-réacteur selon une simple loi affine. Aussi les raisonnements en boucle ouverte qui précèdent ne sont plus valables. La dynamique a changé. Certes $(S_e - \bar{S}, \bar{S})$ reste un point stationnaire ainsi que $(0, S_e)$ mais beaucoup d'autres choses ont changé. En particulier les exposants caractéristiques [définition 10, page 66] autour de ces points sont affectés par cette loi de rétro-action.

Par exemple autour de $(S_e - \bar{S}, \bar{S})$, le jacobien du nouveau champ de vecteurs en boucle fermée est

$$\begin{bmatrix} 0 & k(S_e - \bar{S}) \\ -\bar{D} & -\bar{D} - k(S_e - \bar{S}) \end{bmatrix}.$$

Pour $k > 0$, cette matrice admet une trace < 0 et un déterminant > 0. Ses valeurs propres sont donc à partie réelle strictement négative. Ainsi un simple retour proportionnel avec $k > 0$ rend ce point d'équilibre hyperbolique et stable.

Notre analyse est locale. De plus pour des valeurs de S proches de $\bar{S}$, la commande D ainsi calculée reste positive. Elle est donc physiquement réalisable. Cependant pour des écarts $S - \bar{S}$ importants, D risque d'être négatif. Une première idée est alors de saturer D entre deux valeurs, disons $0 < \varepsilon << \bar{D}$ et $2\bar{D}$. La commande alors obtenue

$$D = \begin{cases} \varepsilon & \text{si } \bar{D} - k(S - \bar{S}) < \varepsilon \\ 2\bar{D} & \text{si } \bar{D} - k(S - \bar{S}) > 2\bar{D} \\ \bar{D} - k(S - \bar{S}) & \text{sinon.} \end{cases}$$

est non linéaire. Pour tout $k > 0$, elle stabilise localement l'équilibre $(S_e - \bar{S}, \bar{S})$. Montrons que même les trajectoires démarrant loin de $(S_e - \bar{S}, \bar{S})$ convergent vers $(S_e - \bar{S}, \bar{S})$ pour k assez grand.

Il est facile de voir que X et S restent toujours positifs. Comme $d/dt(X+S) = -D(S_e - X - S)$, les trajectoires sont bornées et puisque $D \geq \varepsilon$, elles convergent exponentiellement vers le segment $X + S = S_e$. On peut donc supposer que $X + S = S_e$. Mais alors $\dot{X} = (\mu(S_e - X) - D)X$ où D est une fonction de X à cause du bouclage $S = S_e - X$. Sur ce système de dimension 1, il est alors facile de montrer que, si k est choisi assez grand, ses seuls points d'équilibres sont $X = 0$ et $X = S_e - \bar{S}$. Le premier est alors instable et le second stable. Comme l'illustre le portrait de phase de la figure 2.7, cette loi de rétroaction élémentaire stabilise globalement le système au régime de croissance spécifique maximum.

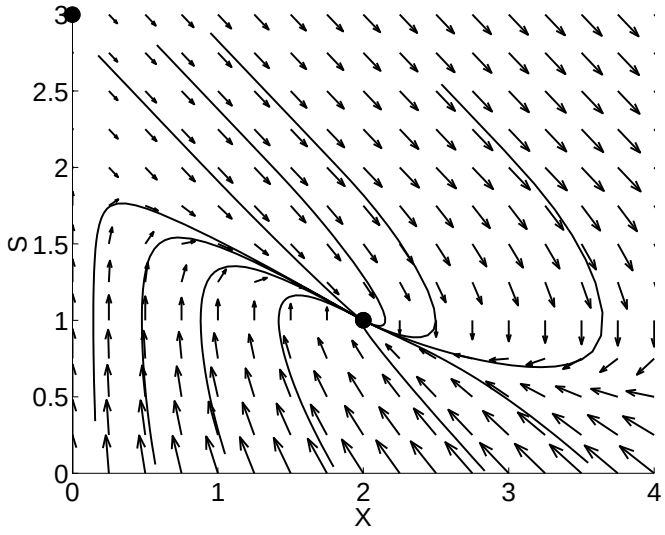


FIG. 2.7 – *stabilisation globale par un feedback $D = \bar{D} - k(S - \bar{S})$ avec saturation (à comparer avec la boucle ouverte, figure 2.6).*

2.2 L'avion à décollage vertical

Ce système est représentatif des problèmes de guidage et de pilote automatique d'engins volants, flottants ou spatiaux. Il s'agit souvent de systèmes mécaniques sous-actionnés,

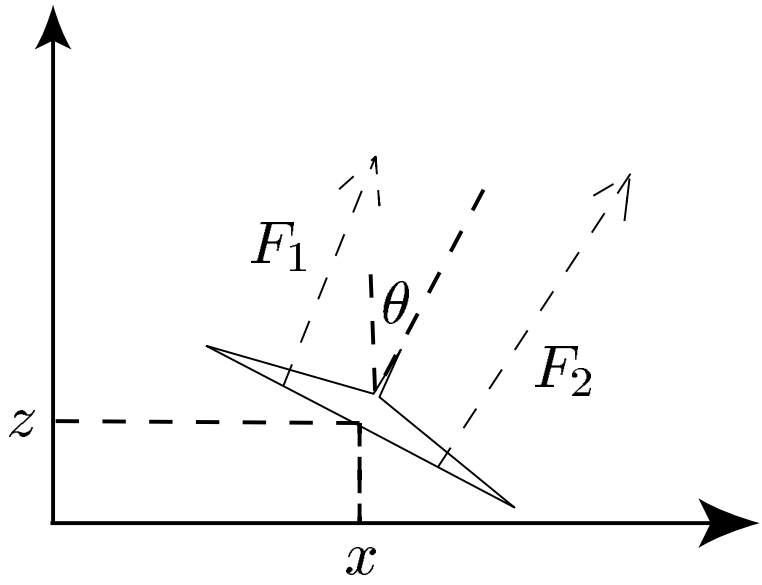


FIG. 2.8 – *le VTOL, l'avion à décollage vertical.*

i.e., avec moins de commandes que de degrés de liberté géométrique.

L'étude de cet exemple est l'occasion de revoir la commandabilité [définition 15, page 84], la planification et le suivi de trajectoires [section 4.2.5, page 94]. Bien que les résultats généraux du cours portent sur les systèmes linéaires, nous montrons comment des calculs comparables (avec les sorties plates, un analogue non linéaire des sorties de Brunovsky [théorème 11, page 91]) permettent de traiter le cas non linéaire. Nous introduisons ici une différence importance entre le modèle de simulation qui est aussi complet que possible et le modèle de commande de dimension réduite et ne prenant en compte que les effets dominants. Ces deux modèles sont proches l'un de l'autre au sens des perturbations singulières [section 3.5.1, page 73]. L'introduction d'un modèle de commande différent du modèle de simulation n'est pas gratuite : elle est fondamentalement liée aux questions de robustesse par rapport aux dynamiques négligées. Ces dynamiques sont d'une part mal connues et d'autre part très rapides (par rapport aux dynamiques à commander) et asymptotiquement stables.

On s'intéresse ici au pilotage d'un avion à décollage vertical en mode "hovering". En particulier on voudrait que l'avion soit en mesure de suivre une trajectoire horizontale (manœuvre-type d'approche à l'atterissage).

2.2.1 Modèle de simulation

Pour simplifier on considère que l'avion se déplace uniquement dans un plan vertical (modèle plan). Si de plus on néglige les effets aérodynamiques, qui sont très faibles en mode "hovering", le comportement dynamique est décrit par

$$\begin{aligned} m\ddot{x} &= (F_1 - F_2)\sin\alpha\cos\theta - (F_1 + F_2)\cos\alpha\sin\theta + f_x(\dot{x},\dot{z},\theta,\dot{\theta}) \\ m\ddot{z} &= (F_1 - F_2)\sin\alpha\sin\theta + (F_1 + F_2)\cos\alpha\cos\theta - mg + f_z(\dot{x},\dot{z},\theta,\dot{\theta}) \\ J\ddot{\theta} &= l(F_1 - F_2)\cos\alpha + f_\theta(\dot{x},\dot{z},\theta,\dot{\theta}), \end{aligned}$$

où (x,z) est la position du centre de masse, θ l'angle par rapport à l'horizontale, F_1,F_2 les poussées des réacteurs, l leur distance par rapport au centre de masse, α leur inclinaison; m est la masse de l'avion et J son moment d'inertie. Les fonctions f_x, f_z et f_θ représentent des petits effets aérodynamiques (en schématisant, ce sont des frottements s'opposant à la vitesse); ces fonctions s'annulent quand l'avion ne bouge pas.

Le transfert entre les poussées F_1,F_2 et les manettes de gaz u_1,u_2 est, grâce à des asservissements "rapides" de bas niveau, à peu près de la forme

$$\dot{F}_1 = \lambda(u_1 - F_1) \tag{2.1}$$
$$\dot{F}_2 = \lambda(u_2 - F_2) \tag{2.2}$$

[voir la section 3.5.1, page 73 où λ joue le rôle de $1/\varepsilon$].

Les valeurs numériques utilisées dans les simulations sont (en unités S.I.)

$$g = 10,\ m = 10000,\ J = 45000,\ l = 4.5,\ \tan\alpha = .1,\ \lambda = 15.$$

On mesure toutes les variables mécaniques, i.e., x,z,θ,$\dot{x}$,$\dot{z}$,$\dot{\theta}$ (un avion, surtout de combat, est toujours très bien instrumenté).

2.2.2 Modèle de commande

En fait la dynamique (2.1)-(2.2) des réacteurs n'est pas très bien connue (c'est un système très complexe). Par contre, on sait d'une part que cette dynamique est plutôt "rapide" et d'autre part qu'en régime établi on a vraiment $F_1 = u_1$ et $F_2 = u_2$.

Après avoir posé $\varepsilon = \dfrac{J}{ml}\tan\alpha$, $a = \dfrac{m}{\cos\alpha}$, $b = \dfrac{J}{l\cos\alpha}$, $v_1 = \dfrac{F_1 + F_2}{a}$ et $v_2 = \dfrac{F_1 - F_2}{b}$, on peut prendre comme *modèle de commande* le système

$$\begin{aligned} \ddot{x} &= \varepsilon v_2 \cos\theta - v_1 \sin\theta \\ \ddot{z} &= \varepsilon v_2 \sin\theta + v_1 \cos\theta - g \\ \ddot{\theta} &= v_2, \end{aligned}$$

ou v_1 et v_2 sont les commandes. De fait nous négligeons ici les effets aérodynamiques et la dynamique des réacteurs. Une justification dans le cadre des perturbations régulières et singulières est possible [théorème 7, page 74 et théorème 8, page 76].

Ce modèle possède une infinité d'état stationnaire : (x,z) arbitraire et $\theta = 0$ ou π. Le cas $\theta = \pi$ correspond à une poussée négative aussi nous l'excluons. Nous considérons toujours les équilibres à l'endroit $\theta = 0$ et (x,z) arbitraire.

2.2.3 Commande linéaire

Linéarisons les équations autour de $(x,z) = 0$ et $\theta = 0$. En notant δx, δz, ..., les écarts on obtient le système linéaire tangent suivant :

$$\begin{aligned} \ddot{\delta x} &= -g\ \delta\theta + \varepsilon\ \delta v_2 \\ \ddot{\delta z} &= \delta v_1 \\ \ddot{\delta\theta} &= \delta v_2. \end{aligned}$$

Le système se décompose donc en deux parties. La première partie

$$\ddot{\delta}z = \delta v_1$$

correspond à la dynamique verticale ne faisant intervenir que δv_1 lié à la poussée totale. La seconde partie

$$\ddot{\delta}x = -g\delta\theta + \varepsilon\delta v_2, \quad \ddot{\delta}\theta = \delta v_2$$

montre que les dynamiques horizontale et angulaire sont couplées et ne dépendent que de v_2, la différence des poussées.

Le suivi en position

La sortie de Brunovsky [théorème 11, page 91] de la dynamique en δz est $y_2 = \delta z$ car $\delta v_1 = \ddot{y}_2$. Ainsi le contrôleur

$$\delta v_1 = \ddot{y}_{2,r} + (p_1 + p_2)(\dot{\delta z} - \dot{y}_{2,r}) - p_1 p_2(\delta z - y_{2,r})$$

assure le suivi d'une référence $t \mapsto y_{2,r}(t)$ [section 4.2.5, page 94]. Les pôles de suivi sont p_1 et p_2 [théorème 12, page 94]. Ils doivent être choisis à partie réelle négative. Une premier choix est le suivant

$$p_1 = -\sqrt{\frac{g}{l}}(1+\imath), \quad p_2 = -\sqrt{\frac{g}{l}}(1-\imath).$$

Il suppose que l'échelle de temps des réacteurs est nettement inférieure à $\sqrt{\frac{l}{g}}$.

La sortie de Brunovsky pour la dynamique en δx et $\delta\theta$ est simplement $y_1 = \delta x - \varepsilon\delta\theta$. En effet on a

$$\delta x = y_1 + \varepsilon\frac{\ddot{y}_1}{g}, \quad \delta\theta = -\frac{\ddot{y}_1}{g}, \quad \delta u_1 = -\frac{y_1^{(4)}}{g}.$$

Le contrôleur

$$\delta u_1 = \left(\frac{-1}{g}\right)\left(y_{1,r}^{(4)} + s_1(-g\dot{\delta\theta} - y_{1,r}^{(3)}) - s_2(-g\delta\theta - \ddot{y}_{1,r}) + s_3(\dot{\delta x} - \varepsilon\dot{\delta\theta} - \dot{y}_{1,r}) - s_4(\delta x - \varepsilon\delta\theta - y_{1,r}\right)$$

assure le suivi de la référence $t \mapsto y_{1,r}(t)$. Les quantités s_i sont les fonctions symétriques homogènes de degré i des 4 pôles de suivi $r_1, \ldots, r_4$:

$$\begin{aligned}
s_1 &= r_1 + r_2 + r_3 + r_4 \\
s_2 &= r_1r_2 + r_1r_3 + r_1r_4 + r_2r_3 + r_2r_4 + r_3r_4 \\
s_3 &= r_1r_2r_3 + r_1r_2r_4 + r_1r_3r_4 + r_2r_3r_4 \\
s_4 &= r_1r_2r_3r_4.
\end{aligned}$$

Comme pour p_1 et p_2 on peut prendre

$$\begin{aligned}
r_1 &= -\sqrt{\frac{g}{l}}(1+\imath) \quad r_2 = -\sqrt{\frac{g}{l}}(1-\imath) \\
r_3 &= -\sqrt{\frac{g}{l}}(1/2+\imath/2) \quad r_4 = -\sqrt{\frac{g}{l}}(1/2-\imath/2).
\end{aligned}$$

Le suivi en vitesse

En pratique les commandes données au pilote, le manche à balai, correspondent à des vitesses, plutôt qu'à des positions. En effet, pour les manoeuvres d'atterrissage, de décollage ou de vol stationnaire, le pilote gère à vue la position. Une autre raison plus fondamentale est l'invariance par translation du modèle. Le fait qu'un pilote conduise naturellement un avion en vitesse vient en grande partie des symétries de translation : le comportement de l'avion est indépendant de sa position cartésienne (x,z). Aussi, un modèle réduit en vitesse à un sens. Il s'écrit

$$\begin{aligned} \dot{\delta u} &= -g\ \delta\theta + \varepsilon\ \delta v_2 \\ \dot{\delta w} &= \delta v_1 \\ \ddot{\delta\theta} &= \delta v_2 \end{aligned}$$

où $u = \dot{x}$ et $w = \dot{z}$.

Le suivi en vitesse est alors obtenu en tronquant le suivi en position. La sortie de Brunovsky de la dynamique en δw est $y_2 = \delta w$ car $\delta v_1 = \dot{y}_2$. Ainsi le contrôleur

$$\delta v_1 = \dot{y}_{2,r} + p(\delta w - y_{2,r})$$

assure le suivi d'une référence $t \mapsto y_{2,r}(t)$ de vitesse verticale. Le pôle p doit être choisi réel et négatif, par exemple $p = -\sqrt{\frac{g}{l}}$.

La sortie de Brunovsky pour la dynamique en δu et $\delta\theta$ est simplement $y_1 = \delta u - \varepsilon\dot{\delta\theta}$. En effet on a

$$\delta u = y_1 + \varepsilon\frac{\dot{y}_1}{g}, \quad \delta\theta = -\frac{\dot{y}_1}{g}, \quad \delta u_1 = -\frac{y_1^{(3)}}{g}.$$

Le contrôleur

$$\delta u_1 = \left(\frac{-1}{g}\right)\left(y_{1,r}^{(3)} + s_1(-g\dot{\delta\theta} - \ddot{y}_{1,r}) - s_2(-g\delta\theta - \dot{y}_{1,r}) + s_3(\delta u - \varepsilon\dot{\delta\theta} - y_{1,r})\right)$$

assure le suivi d'une référence $t \mapsto y_{1,r}(t)$ de vitesse horizontale. Les quantités s_i sont les fonctions symétriques homogènes de degré i des 3 pôles de suivi r_1, r_2 et r_3 :

$$\begin{aligned} s_1 &= r_1 + r_2 + r_3 \\ s_2 &= r_1r_2 + r_1r_3 + r_2r_3 \\ s_3 &= r_1r_2r_3. \end{aligned}$$

Comme pour p on peut prendre

$$r_1 = -\sqrt{\frac{g}{l}}(1+\imath), \quad r_2 = -\sqrt{\frac{g}{l}}(1-\imath), \quad r_3 = -\frac{1}{2}\sqrt{\frac{g}{l}}.$$

Les consignes en vitesse venant du pilote peuvent être très irrégulières s'il bouge rapidement le manche à balai. Il convient de les régulariser un peu. Notons les $y_{1,c}$ et $y_{2,c}$. On les supposera uniquement mesurables et bornées. La référence de vitesse verticale doit être au moins C^1 et horizontale au moins C^3. En fait il faut rajouter un ordre de régularité

car nous avons négligé la dynamique des réacteurs. Des discontinuité de poussée sont impossibles physiquement. Ainsi les références $y_{1,r}$ et $y_{2,r}$ doivent être respectivement C^4 et C^2. Elles correspondent donc à des valeurs lissées des consignes brutes issues du manche à balai, $y_{1,c}$ et $y_{2,c}$. Une simple convolution par un noyau régularisant h, positif, d'intégrale égale à 1, au moins C^4 et à support compact dans $]-\infty,0]$ assure ce lissage $(i = 1,2)$:

$$y_{i,r}(t) = \int_{-\infty}^{+\infty} h(t-\sigma)\ y_{i,c}(\sigma)\ d\sigma,$$

avec

$$y_{i,r}^{(\nu)}(t) = \int_{-\infty}^{+\infty} h^{(\nu)}(t-\sigma)\ y_{i,c}(\sigma)\ d\sigma, \quad \nu = 1,\ldots,4.$$

On peut aussi utiliser un filtre de dimension fini et d'ordre 4. Cela revient alors à une convolution avec un noyau h donc le support reste toujours dans $]-\infty,0]$ (causalité du filtre) mais qui n'est plus compact.

Notons enfin que ces méthodes sont locales et valables pour des vitesses pas trop grandes et une inclinaison réduite. De plus les ordres de poussées F_1 et F_2 donnés aux réacteurs doivent être entre deux bornes et en particulier positifs. Ici encore, il convient de s'assurer que les trajectoires suivis par l'avion vérifient ces contraintes en poussée, vitesse et inclinaison. Une façon de les garantir consiste à générer à partir des ordres du pilote $(y_{1,c},y_{2,c})$ des trajectoires de références $(y_{1,r},y_{2,r})$ suffisamment douces. On joue alors sur la forme h du noyau régularisant.

2.2.4 Commande non-linéaire

Le suivi de trajectoires élaboré à partir du linéaire tangent s'étend au modèle non linéaire de commande :

$$\begin{aligned} \ddot{x} &= \varepsilon v_2 \cos\theta - v_1 \sin\theta \\ \ddot{z} &= \varepsilon v_2 \sin\theta + v_1 \cos\theta - g \\ \ddot{\theta} &= v_2. \end{aligned}$$

En effet ce système admet une structure très particulière avec des "sorties de Brunovsky non linéaires", dites sorties plates :

$$y_1 = x - \varepsilon \sin\theta, \quad y_2 = z + \varepsilon \cos\theta.$$

Remplaçons x et z par y_1 et y_2 dans les équations du systèmes. Cela revient à étudier le même système mais avec un jeu de variables (y_1,y_2,θ,v_1,v_2) différentes de (x,z,θ,v_1,v_2). Le but est de faire des changements de variables qui simplifient les équations en les mettant sous une forme canonique dite forme normale. Dans ces nouvelles variables les équations du système deviennent

$$\begin{aligned} \ddot{y}_1 &= -(v_1 - \varepsilon\dot{\theta}^2)\sin\theta \\ \ddot{y}_2 &= (v_1 - \varepsilon\dot{\theta}^2)\cos\theta - g \\ \ddot{\theta} &= v_2. \end{aligned}$$

Ainsi nous savons que

$$\theta = \arctan(\ddot{y}_2 + g/\ddot{y}_1) \qquad \mod \pi$$

et donc

$$x = y_1 \pm \varepsilon \frac{\ddot{y}_1}{\sqrt{\ddot{y}_1^2 + (\ddot{y}_2 + g)^2}}$$
$$z = y_2 \pm \varepsilon \frac{(\ddot{y}_2 + 1)}{\sqrt{\ddot{y}_1^2 + (\ddot{y}_2 + g)^2}}.$$

Nous voyons donc que toutes les variable du systèmes s'expriment comme des fonctions, ici non linéaires, de y_1, y_2 et d'un nombre fini de leur dérivées. Noter que puisque nous avons deux commandes indépendantes, nous pouvons ainsi paramétrer toutes les trajectoires du système à partir de y_1 et y_2. C'est très comparable aux systèmes linéaires commandables et leurs sorties de Brunovsky.

Continuons avec un changement un peu plus général qui touche aux commandes v_1 et v_2. Partons des équations dans les variables (y_1,y_2,θ,v_1,v_2). Considérons les nouvelles commandes (u_1,u_2) définies à partir des anciennes commandes (v_1,v_2) par les équations (bouclage dynamique) :

$$\ddot{\xi} = -u_1 \sin\theta + u_2 \cos\theta + \xi\dot{\theta}^2$$
$$v_1 = \xi + \varepsilon\dot{\theta}^2$$
$$v_2 = \frac{-1}{\xi}(u_1 \cos\theta + u_2 \sin\theta + 2\dot{\xi}\dot{\theta}).$$

Des calculs simples montrent alors que

$$y_1^{(4)} = v_1, \qquad y_2^{(4)} = v_2.$$

C'est la forme normale de Brunovsky d'un système linéaire commandable à 8 états et 2 commandes. Ainsi par changement de variables et bouclage on se ramène à un système linéaire commandable. Dès lors, il suffit de planifier les trajectoires et de construire le suivi dans ces nouvelles variables où les équations sont particulièrement simples.

Une question se pose naturellement après ces quelques calculs : est-ce toujours possible de faire ainsi? La réponse est négative. Il n'est pas possible en général de "tuer" les non linéarités par des changements de variables et bouclages astucieux. Cependant, pour de nombreux systèmes physiques, c'est souvent le cas avec les changements de variables ayant un sens physique direct. Les systèmes rencontrés en pratique ne sont pas des systèmes génériques. Ils admettent souvent une structure particulière, liée à la physique, qui simplifie alors notablement leur contrôle. Pour l'exemple de l'avion, (y_1,y_2) sont les coordonnées cartésiennes du centre d'oscillation [1] de l'avion pour un axe de rotation orthogonal au plan Oxz et passant par le point d'intersection des deux directions de poussée (si les directions de poussées sont parallèles ($\alpha = 0$) alors le centre d'oscillation se confond avec le centre de gravité ($\varepsilon = 0$)). Pour plus de détail, voir le cours sur les systèmes plats téléchargeable à l'adresse *http://math.polytechnique.fr/xups/vol99.html*.

2.3 Pendule inversé sur un rail

Cet exemple est ultra-classique. Il permet néanmoins de se faire une idée des limitations des techniques non linéaires mêmes les plus récentes. En effet, son approximation linéaire tangente peut se traiter sans rien connaître. Il suffit juste d'utiliser les loi de la mécanique

1. Voir les travaux de Huygens sur les horloges à pendules.

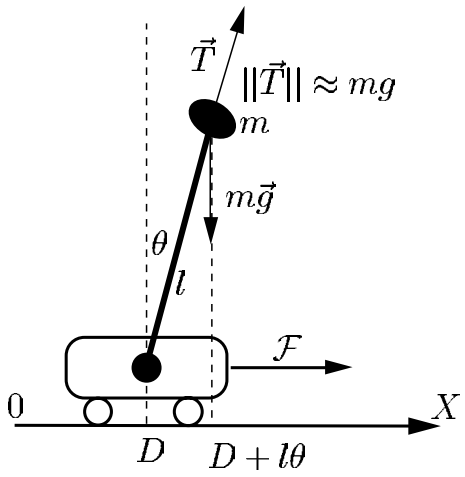


FIG. 2.9 – *pendule inversé sur un rail.*

et l'approximation des petits angles. En revanche, dès que les angles sont grands, des effets non linéaires intrinsèques (i.e., que l'on ne peut pas éliminer par changements de variables et bouclages comme pour l'avion à décollage vertical) apparaissent. Par des techniques non linéaires dites contrôle-Lyapounov, on sait calculer des commandes simples qui font passer de l'équilibre stable à l'équilibre instable. Cependant dès que l'on rajoute un second pendule (double pendule) on ne sait plus à l'heure actuelle élaborer des contrôles simples et mathématiquement prouvés qui retournent le double pendule. En revanche la stabilisation locale en position inverse ne pose pas de problème en utilisant le linéaire tangent (cf. le stand du double pendule inversé au musée des sciences et de l'industrie de la Villette à Paris, section mathématique).

Un pendule inversé sur un rail admet la dynamique suivante (approximation des petits angles)

$$\frac{d^2}{dt^2}(D + l\theta) = g\theta, \quad M\frac{d^2}{dt^2}D = -mg\theta + \mathcal{F}$$

où la commande est la force $\mathcal{F}$ appliquée au chariot et l est la distance du centre d'oscillation à l'axe de rotation du pendule. Il est clair que la sortie de Brunovsky est $y = D + l\theta$. En effet

$$\theta = \ddot{y}/g, \quad D = y - l\ddot{y}/g.$$

Un bouclage grand gain sur le chariot (u est la consigne de position du chariot)

$$\mathcal{F} = -Mk_1\dot{D} - Mk_2(D - u)$$

avec $k_1 \approx 10/\tau$, $k_2 \approx 10/\tau^2$ où $\tau = \sqrt{l/g}$ est le temps caractéristique du pendule, permet d'accélérer par la commande le porteur. On obtient ainsi une commande hiérarchisée avec un asservissement rapide en position du porteur et une stabilisation lente du pendule à partir du modèle lent

$$\frac{d^2}{dt^2}(y) = g(y - u)/l = \frac{y - u}{\tau^2}.$$

Le simple bouclage

$$u = -y - \tau^2\ddot{y}_r(t) + \tau(\dot{y} - \dot{y}_r(t)) + (y - y_r(t))$$

assurent le suivi d'une trajectoire de référence $t \mapsto y_r(t)$ pour l'abscisse du centre d'oscillation du pendule.

2.4 Moteur électrique à courant continu

Avec cet exemple nous abordons les capteurs logiciels, des traitements en temps-réel de l'information venant des capteurs pour en déduire des informations non bruitées sur des grandeurs mesurées ou non. Sur l'exemple choisi ici, il s'agit, à partir de la mesure des tensions et des courants qui traversent le moteur , d'estimer de façon causale sa vitesse mécanique et son couple de charge. L'intérêt pratique est évident: les informations électriques sont toujours disponibles car les capteurs sont simples et fiables. En revanche, les informations mécaniques nécessitent une instrumentation plus complexe, plus chère et moins fiable. Aussi pour des raisons de coût mais aussi de sécurité, déduire des courants et tensions, la vitesse de rotation est un enjeu technologique important en électro-technique. Dans d'autres domaines, on rencontre des problèmes très similaires. Pour les procédés, les débits, températures et pressions sont faciles à avoir par des capteurs simples et robustes alors que les qualités sont plus difficiles à mesurer rapidement (temps de retard de l'analyse, ...). Un traitement de l'information contenue dans les températures, pressions et débits permet souvent d'obtenir des estimations précieuses sur les compositions. Pour estimer l'orientation relative d'un mobile par rapport à un référentiel terrestre les mesures sont de deux types: les gyromètres donnent de façon précise les vitesses angulaires; des magnétomètres on déduit une mesure bruitée des cosinus directeurs de la direction du champ magnétique par rapport au mobile. Il faut en déduire grâce aux relations cinématiques une estimation robuste de l'orientation du mobile (ses trois angles d'Euler, i.e., une matrice de rotation). Ce problème est centrale pour la mise au point d'avion sans pilote.

L'exemple du moteur à courant continu permet de se faire une idée des questions soulevées et des techniques utiles pour apporter des solutions simples à ce type de question. Cet exemple illustre l'observabilité [définition 19, page 101], les observateurs asymptotiques [section 4.4.2, page 106], l'observateur-contrôleur [section 4.5, page 107] et les questions de robustesse en liaison avec les systèmes lents-rapides [fin de la sous-section 3.5.1, page 73].

Un premier modèle de moteur à courant continu est le suivant :

$$\begin{aligned} J\dot{\omega} &= k\,\iota - p \\ Li &= -k\omega - R\,\iota + u \end{aligned}$$

où ω est la vitesse de rotation du moteur, ι le courant, u la tension, $L > 0$ la self, $R > 0$ la résistance, k la constante de couple, p le couple de charge et J l'inertie de la partie tournante (moteur + charge).

Nous supposons les paramètres $J > 0$, $k > 0$, $L > 0$ et $R > 0$ connus et constants. En revanche seule l'intensité ι est mesurée. La charge p est une constante inconnue. Il nous faut concevoir un algorithme qui ajuste en temps réel la tension u de façon à suivre une vitesse de référence $\omega_r(t)$ variable dans le temps. Pour cela nous ne disposons que d'un capteur de courant. Beaucoup de variateurs de vitesse régulent la vitesse du moteur sans la mesurer. Rajouter un capteur de vitesse sur l'arbre du moteur est parfois délicat. En

revanche, brancher entre le moteur et son alimentation électrique une petit boite, i.e. le variateur de vitesse, est très simple à réaliser.

2.4.1 Stabilité en boucle ouverte

Il est facile de constater que la dynamique en boucle ouverte $u = cte$ est stable. On peut le voir directement en vérifiant que les exposants caractéristiques [définition 10, page 66] sont à partie réelle négative. Une autre façon de le voir, plus physique, consiste à remarquer que l'effet Joule $-R\iota^2$ est dissipatif. L'énergie du système

$$E = \frac{1}{2}J\omega^2 + \frac{1}{2}L\iota^2,$$

somme de l'énergie cinétique de la partie tournante et de l'énergie magnétique contenu dans les bobinages du moteur, vérifie

$$\frac{dE}{dt} = \omega p + ui - R\iota^2.$$

Or la stabilité du système pour u et p non nuls est équivalente ici (le système est linéaire) à celle du système avec u et p nuls (il suffit de faire une simple translation). Ainsi E est une fonction de Lyapounov pour $u = p = 0$: le système est stable et même asymptotiquement stable (invariance de Lasalle) [théorème 3, page 58].

2.4.2 Estimation de la vitesse et de la charge

Vérifions que le système

$$\begin{aligned} \dot{p} &= 0 \\ J\dot{\omega} &= k\ \iota - p \\ L\dot{\iota} &= -k\ \omega - R\ \iota + u \end{aligned}$$

avec comme commande u et comme sortie $y = \iota$ est observable [définition 21, page 101]. Cela revient à se pose la question suivante : connaissant $t \mapsto (\iota(t),u(t))$ et les équations du système, est-il possible de calculer ω et p. La réponse est positive et immédiate car

$$\omega = (u - L\dot{\iota} - R\iota)/k, \quad p = k\iota - (J/k)(\dot{u} - L\ddot{\iota} - R\dot{\iota}).$$

Le système est donc observable. On pourrait aussi reprendre le critère de Kalman [théorème 14, page 104]. Il est issu du même calcul que celui qui précède mais sur un système linéaire général.

Cependant les mesures de courant sont bruitées. Il est donc hors de question de dériver ce signal. Le fait d'être en théorie observable ne donne pas un algorithme d'estimation réaliste. Il nous faut concevoir un algorithme qui soit insensible au bruit, i.e., qui les filtre astucieusement sans introduire de déphasage. Ici apparaît une idée centrale celle d'observateur asymptotique. Elle consiste à copier la dynamique du système en lui rajoutant des termes correctifs liés à l'erreur entre la prédiction et la mesure. Cela donne ici l'observateur suivant

$$\begin{aligned} \dot{\hat{p}} &= L_p(\hat{\iota} - \iota) \\ J\dot{\hat{\omega}} &= k\ \hat{\iota} - \hat{p} + L_\omega(\hat{\iota} - \iota) \\ L\dot{\hat{\iota}} &= -k\hat{\omega} - R\ \hat{\iota} + u + L_\iota(\hat{\iota} - \iota) \end{aligned}$$

où il est d'usage de rajouter un chapeau sur les estimées. Noter que le paramètre inconnu p est rajouté à l'état avec comme équation $\dot{p} = 0$. On parle souvent pour p d'identification, $\hat{p}$ étant la valeur identifiée. En choisissant correctement les gains L_p, L_ω et L_ι les écarts entre les estimées et les vraies valeurs tendent vers zéro. En effet la dynamique de l'erreur (il est d'usage de noter avec un tilde les écarts entre les estimées et les grandeurs réelles)

$$\begin{aligned} \dot{\tilde{p}} &= L_p\tilde{\iota} \\ J\dot{\tilde{\omega}} &= -\tilde{p} + (L_\omega + k)\tilde{\iota} \\ L\dot{\tilde{\iota}} &= (L_\iota - R)\tilde{\iota} - k\tilde{\omega}. \end{aligned}$$

Il s'agit d'un système autonome donc les exposants caractéristiques sont les racines du polynôme de degré 3 suivant

$$X^3 - \frac{L_\iota - R}{L}X^2 + \frac{k(L_\omega + k)}{LJ}X - \frac{L_p k}{LJ}.$$

Étant donné qu'en jouant sur les L_p, L_ω et L_ι, on peut donner n'importe quelles valeurs aux fonctions symétriques des racines, il est possible de les choisir comme l'on veut. Si L_p, L_ω et L_ι vérifient

$$\begin{aligned} \frac{L_\iota - R}{L} &= r_1 + r_2 + r_3 \\ \frac{k(L_\omega + k)}{LJ} &= r_1r_2 + r_1r_3 + r_2r_3 \\ \frac{L_p k}{LJ} &= r_1r_2r_3 \end{aligned}$$

alors les racines seront r_1, r_2 et r_3. On peut choisir pour les pôles d'observation [théorème 15, page 106] les valeurs suivantes

$$r_1 = -\frac{R}{L}(1 + \sqrt{-1}), \quad r_2 = -\frac{R}{L}(1 - \sqrt{-1}), \quad r_3 = -\sqrt{\frac{k^2}{LJ}}.$$

Ce choix correspond aux échelles de temps caractéristiques du système en boucle ouverte.

Les valeurs $\hat{p}$, $\hat{\omega}$ et $\hat{\iota}$ ainsi calculées convergent vers p, ω et ι, quelque-soit la loi horaire $t \mapsto u(t)$. Noter aussi que même si la mesure de courant est bruitée (bruit haute fréquence et centré autour de 0), l'observateur nous donne une valeur filtrée, $\hat{\iota}$, sans déphasage et qui n'élimine que le bruit.

2.4.3 Le contrôleur

A cause de ce qui précède, nous pouvons supposer maintenant ω, p et ι connus et élaborer le suivi de trajectoire. La sortie de Brunovsky est ω (ici p n'est qu'un paramètre qui va intervenir dans le bouclage). On a

$$\ddot{\omega} = -\frac{k^2}{LJ}\omega - \frac{Rk}{LJ}\iota + \frac{k}{LJ}u.$$

Donc le bouclage assurant le suivi de ω_r est obtenu en ajustant u de sorte que

$$\ddot{\omega} = \ddot{\omega}_r + (p_1 + p_2)(\dot{\omega} - \dot{\omega}_r) - p_1p_2(\omega - \omega_r)$$

où p intervient car $\dot{\omega} = \frac{k\iota - p}{J}$. Les pôles de suivi [théorème 12, page 94] peuvent être choisis plus rapides que ceux de l'observateur (au aussi plus lents) :

$$p_1 = -2\sqrt{\frac{k^2}{LJ}}(1+\sqrt{-1}), \quad p_2 = -2\sqrt{\frac{k^2}{LJ}}(1-\sqrt{-1})$$

2.4.4 L'observateur-contrôleur

Partons maintenant des deux signaux, $t \mapsto y(t)$, la mesure de courant et $t \mapsto \omega_r(t)$, la référence de vitesse. Nous supposons pour simplifier que ω_r est C^2 par morceau (le moteur ne peut pas suivre des références de vitesse plus irrégulières)

Voyons quel algorithme causal nous avons pour calculer u. L'observateur est le système dynamique

$$\begin{aligned} \dot{\hat{p}} &= L_p(\hat{\imath} - y(t)) \\ J\dot{\hat{\omega}} &= k\ \hat{\imath} - \hat{p} + L_\omega(\hat{\imath} - y(t)) \\ L\dot{\hat{\imath}} &= -k\hat{\omega} - R\ \hat{\imath} + u(t) + L_\iota(\hat{\imath} - y(t)). \end{aligned}$$

La commande est alors obtenue en remplaçant dans le contrôleur les variables par leur estimées. Ainsi $u(t)$ est solution du système linéaire

$$-\frac{k^2}{LJ}\hat{\omega} - \frac{Rk}{LJ}\hat{\imath} + \frac{k}{LJ}u = \ddot{\omega}_r + (p_1+p_2)(\frac{k\hat{\imath} - p}{J} - \dot{\omega}_r) - p_1p_2(\hat{\omega} - \omega_r).$$

Le principe de séparation [2] [section 4.5, page 107] assure alors la convergence du système physique

$$\begin{aligned} J\dot{\omega} &= k\ \iota - p \\ Li &= -k\omega - R\ \iota + u \end{aligned}$$

couplé à l'observateur-contrôleur : les estimées tendent vers les vraies grandeurs et la vitesse ω converge vers sa référence ω_r même si cette dernière varie tout le temps.

2.4.5 Robustesse par rapport à la dynamique rapide du courant

Supposons, et c'est souvent le cas, que la dynamique électrique est nettement plus rapide que la dynamique mécanique. Cela revient à dire que la self L est très petite, positive mais mal connue. Ainsi tout ce que l'on sait c'est que $L \approx \varepsilon$ où ε est un petit paramètre positif inconnu. Il est alors facile de voir que les perturbations singulières [section 3.5.1, page 73] s'appliquent ici : le système reste stable en boucle ouverte avec une dynamique du courant convergeant immédiatement vers son régime quasi-statique d'équation [3]

$$\iota = (u - k\omega)/R$$

et la dynamique lente de la vitesse étant alors

$$J\dot{\omega} = -(k^2/R)\omega + (k/R)u - p.$$

2. On parle de principe de séparation car, pour les systèmes linéaires, il est possible de traiter séparément la construction de l'observateur et celle du contrôleur. Ce n'est plus le cas en général pour les systèmes non linéaires (c.f. la commande adaptative).

3. On parle souvent d'équation de la couche limite. Historiquement, les systèmes lents rapides ont été mis en évidence par Prandtl dans l'étude des fluides peu visqueux et de leur profils de vitesse près des parois, la dérivée en temps étant alors remplacée par la dérivée en espace.

Reprenons maintenant les calculs du contrôleur en repérant les petits diviseurs, i.e. les divisions par notre estimation $\hat{\varepsilon}$ de $L = \varepsilon$. La commande est solution de

$$-\frac{k^2}{\hat{\varepsilon}J}\omega - \frac{Rk}{\hat{\varepsilon}J}\iota + \frac{k}{\hat{\varepsilon}J}u = \ddot{\omega}_r + (p_1 + p_2)(\frac{k\iota - p}{J} - \dot{\omega}_r) - p_1p_2(\omega - \omega_r).$$

Avec comme dynamique réelle

$$\ddot{w} = -\frac{k^2}{\varepsilon J}\omega - \frac{Rk}{\varepsilon J}\iota + \frac{k}{\varepsilon J}u$$

où la vraie self $L = \varepsilon$ reste petite et positive, nous avons en boucle fermée

$$\ddot{w} = \frac{\hat{\varepsilon}}{\varepsilon}\ (\ddot{\omega}_r + (p_1 + p_2)(\dot{\omega} - \dot{\omega}_r) - p_1p_2(\omega - \omega_r))\,.$$

Mais une petite erreur en valeur absolue sur $\varepsilon = L$ peut être grande en valeur relative. Ici certaines divisions sont catastrophiques : $\frac{\hat{\varepsilon}}{\varepsilon}$ peut être très loin de 1. Aussi ce contrôleur n'est pas robuste à ce type d'incertitude sur $\varepsilon = L$. On peut effectuer la même analyse pour l'observateur et obtenir les mêmes conclusions.

Comment faire ? Il suffit de rependre la synthèse du contrôleur et de l'observateur sur le modèle lent. Nous avons alors un modèle de commande qui ne dépend plus de ε :

$$\iota = (u - k\omega)/R, \quad J\dot{\omega} = -(k^2/R)\omega + (k/R)u - p.$$

Tout se passera alors bien [fin de la section 3.5.1, page 73]. Il suffit de remarquer que puisque $\omega = (u - R\iota)/k$, la vitesse est indirectement connue en combinant la tension et la mesure de courant. L'observateur aura la forme suivante :

$$\begin{aligned} \dot{\hat{p}} &= L_p(\hat{\omega} - z(t)) \\ J\dot{\hat{\omega}} &= -(k^2/R)\hat{\omega} + (k/R)u - \hat{p} + L_\omega(\hat{\omega} - z(t)) \end{aligned}$$

où la mesure $z(t) = (u(t) - R\iota(t))/k$ correspond à la vitesse ω. Le suivi sera alors assuré par u solution de

$$-(k^2/R)\hat{\omega} + (k/R)u - \hat{p} = J(\dot{\omega}_r - (\hat{\omega} - \omega_r)/\tau).$$

La seule limitation est celle des gains L_p, L_ω et $1/\tau$: ils doivent respecter les échelles de temps du système et ne pas être trop grands. Le modèle de commande est un modèle approché. Il représente bien les dynamiques nettement plus lentes que celles du courant. Aussi les dynamiques d'observation et de suivi doivent être, elles aussi, nettement plus lentes que celles du courant.

2.4.6 Boucle rapide et contrainte de courant

Il est important pour des raisons de sécurité de garantir un courant ι borné. Ainsi nous avons comme contrainte

$$|\iota| \leq \iota_{max}$$

où $\iota_{max} > 0$ est le courant maximum supporté par le variateur et le moteur. Les algorithmes précédents ne garantissent pas le respect de cette contrainte d'état. Néanmoins, il

est possible de prendre en compte cette contrainte en ne modifiant que le contrôleur de la section précédente sans toucher à l'observateur. On considère donc un premier bouclage grand gain en courant :

$$u = -R\bar{\iota} + u - k\hat{\omega} + \frac{1}{\eta}(\bar{\iota} - \iota)$$

où $\bar{\iota}$ est une référence de courant et η tel que $\eta R \ll 1$. Un tel bouclage rend la dynamique du courant stable et bien plus rapide que celle de la vitesse. Remarquer que même si L est petit, ce bouclage ne fait que renforcer la rapidité du courant sans le déstabiliser. En effet on a

$$L\dot{\iota} = k\tilde{\omega} - (R + 1/\eta)(\iota - \bar{\iota})$$

et donc $\iota \approx \bar{\iota}$ est une très bonne approximation. Ainsi la dynamique de la vitesse se réduit à

$$J\dot{\omega} = k\bar{i} - p.$$

Une justification mathématique de cette réduction relève encore de la théorie des perturbations singulières.

La référence de courant peut alors être calculée ainsi :

$$\bar{\iota} = \frac{J\dot{\omega}_r - J(\hat{\omega} - \omega_r)/\tau + \hat{p}}{k}$$

où $\tau > 0$ est un temps nettement supérieur l'échelle de temps de la dynamique du courant, i.e., $\tau \gg L/(R+1/\eta)$. Si la référence de courant $\bar{\iota}$ ainsi calculée ne vérifie pas la contrainte alors il suffit de la saturer en valeur absolue à ι_{max} en préservant son signe. Il est facile de voir qu'une telle saturation ne peut pas déstabiliser la vitesse. Elle assure de fait le suivi au mieux de la référence ω_r.

Chapitre 3

Systèmes dynamiques explicites

Dans ce chapitre nous rappelons quelques résultats fondamentaux nécessaires à l'étude des équations différentielles ordinaires explicites, $\dot{x} = v(x)$: existence et unicité des solutions, comportements asymptotiques pour des temps grands. La théorie des équations différentielles ordinaires permet d'étudier de nombreux processus d'évolution *déterministes, finis* et *différentiables*. Nous exposons ici les principales notions indispensables à l'étude de tels systèmes. Ces notions sont à la base de la théorie des *systèmes dynamiques* dont l'objet principal est l'analyse *qualitative* des solutions sur de *longs* intervalles de temps. La difficulté principale vient du fait que, dans le cas général, nous ne connaissons pas la solution générale de tels systèmes.

Ce chapitre reprend des éléments du cours de tronc commun [2]. Pour une présentation intrinsèque nous renvoyons au cours de calcul des variations [6]. Sauf cas contraire, les démonstrations des résultats ci-dessous se trouvent dans [3] ou [12].

3.1 Espace d'état, champ de vecteurs et flot

Nous commençons par introduire les notions d'espace d'état (on parle aussi d'espace des phases), de champ de vitesse et de flot sur un exemple simple. Puis nous abordons le cas général avec les justifications mathématiques qui conviennent.

3.1.1 Un modèle élémentaire de population

Considérons une population de x micro-organismes (x grand) dans un milieu nutritif favorable (un fermenteur par exemple) et avec une vitesse de reproduction proportionnelle à x (cette condition est une bonne approximation tant que la nourriture est suffisante, tant que les micro-organismes ne meurent pas, ...).

Ce processus est décrit par l'équation différentielle de bilan suivante :

$$\frac{dx}{dt} = \mu x \tag{3.1}$$

où μ est la vitesse spécifique de reproduction (μ est une constante positive exprimée par exemple en 1/h). x est la grandeur caractéristique du système, son *état*: il appartient à l'ensemble des réels positifs $[0, +\infty[$, *l'espace d'état*, appelé aussi espace des phases

pour des raisons historiques[1]. μx est la vitesse d'évolution : elle résulte des hypothèses de modélisation.

Nous voyons qu'il n'est pas nécessaire de connaître les solutions de (3.1) pour connaître explicitement la vitesse d'évolution. Il suffit de connaître la position x dans l'espace d'état, i.e. l'état. Il est alors naturel d'introduire la notion de champ de vecteurs vitesse : le champ de vecteurs vitesse est l'application qui à chaque point x de l'espace d'état fait correspondre le vecteur vitesse $v(x)$ (ici $v(x) = \mu x$) d'évolution du phénomène.

La modélisation se décompose donc en deux étapes :

étape 1 : se donner un espace d'état convenable qui permette de caractériser le système à chaque instant par un point dans cet espace ;

étape 2 : décrire *quantitativement* l'évolution de proche en proche du système par un champ de vecteurs vitesse et en donner une expression calculable en fonction de la position x dans l'espace d'état.

A partir d'une population initiale x_0 au temps $t = 0$, la résolution explicite de (3.1) conduit à la loi horaire

$$x(t) = \exp(\mu t)x_0.$$

On constate alors les points suivants, également vrais pour les systèmes généraux différentiables :

- deux solutions ayant la même condition initiale x_0 sont identiques (unicité de la solution) ;
- par une condition initiale x_0 passe une solution (existence pour des intervalles de temps fini).

Pour chaque instant t, l'application

$$\begin{array}{rcl} \phi_t \ : [0, +\infty[& \longrightarrow & [0, +\infty[\\ x & \longrightarrow & \exp(\mu t)x \end{array}$$

est une bijection régulière (il est d'usage de dire difféomorphisme) de l'espace d'état ($[0, +\infty[$) dans lui même et $(\phi_t)^{-1} = \phi_{-t}$. Remarquons aussi que $\phi_t \circ \phi_s = \phi_{t+s}$ et $\phi_0 = I$ (I est l'identité). Ainsi, l'ensemble $G = (\phi_t)_{t\in\mathbb{R}}$ est un *groupe à un paramètre de difféomorphismes*. Il est d'usage d'appeler *flot*[2] ce groupe $\{\phi_t\}$. On appelle trajectoires, les courbes de l'espace d'état $\phi_t(x)$ paramétrées par le temps t.

La connaissance du flot $\{\phi_t\}$ entraîne la connaissance du champ de vecteurs vitesse. En effet

$$v(x) = \mu x = \left.\frac{d}{dt}\right|_{t=0} (\phi_t(x)).$$

1. L'espace des phases a été introduit par H. Poincaré en mécanique. Le mouvement d'un système mécanique autonome à n degrés de liberté (espace des configurations) est entièrement caractérisé par un point dans un espace des phases de dimension $2n$. Par exemple, la l'orientation d'un solide en rotation autour d'un point fixe est repérée par les 3 angles d'Euler (espace des configurations $SO(3)$, l'ensemble des rotations de l'espace euclidien physique $\mathbb{R}^3$), alors que sa dynamique, i.e. son évolution au cours du temps est caractérisée par sa position et sa vitesse initiale, i.e. les trois angles d'Euler et les trois vitesses instantanées de rotation. Dans cet exemple, l'espace des phases est de dimension 6 (le fibré tangent de $SO(3)$). Nous renvoyons le lecteur intéressé par ces notions au remarquable livre de V.I. Arnol'd [4] ainsi qu'au cours de calcul des variations [6].

2. Le terme "flot" vient d'une analogie cinématique avec l'écoulement stationnaire d'un fluide : si x_0 est la position d'un élément de fluide à l'instant $t = 0$, $\phi_t(x_0)$ est la position de l'élément de fluide à l'instant t lorsque chaque élément de fluide de position x est soumis à la vitesse d'écoulement $v(x)$.

Nous voyons donc qu'il est équivalent de se fixer le champ de vecteurs vitesse ou le flot : l'un permet la détermination de l'autre et réciproquement. Plus généralement, à tout groupe à un paramètre de difféomorphismes est associé une équation différentielle par la relation précédente.

Nous avons directement sur le flot $\{\phi_t\}$ les comportements des solutions de (3.1) lorsque t devient grand en valeur absolue :

- si $x_0 = 0$ alors $\phi_t(x_0) = 0$ pour tout $t \in \mathbb{R}$;
- si $x_0 > 0$ alors $\phi_t(x_0) \to +\infty$ quand t tend vers $+\infty$ et $\phi_t(x_0) \to 0$ quand t tend vers $-\infty$.

Ce type de renseignement sur le comportement asymptotique des solutions n'est pas évident a priori pour un champ de vecteurs vitesse $v(x)$ quelconque. Ici réside l'une des difficultés majeures : la modélisation fournit principalement l'espace d'état et le champ $v(x)$ sur cet espace alors que les réponses aux questions qualitatives sont fournies par le flot $\{\phi_t\}$. Dans le cas général, il est extraordinairement difficile de déduire, de la connaissance du champ de vecteurs vitesse, des propriétés globales relatives au flot et aux comportements asymptotiques d'ensembles de solutions.

3.1.2 Existence, unicité, flot

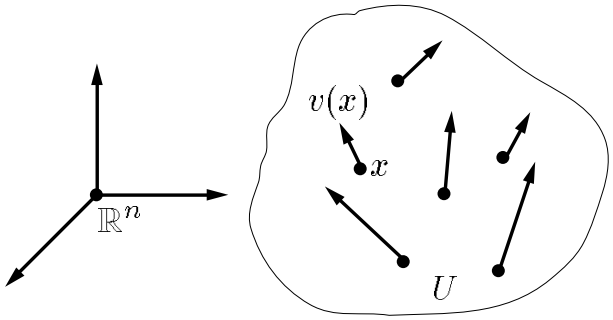


FIG. 3.1 – *champ de vecteurs vitesse $v(x)$ sur un domaine U contenu dans $\mathbb{R}^n$.*

Les résultats d'existence et d'unicité sont locaux en temps et en espace. Aussi peuvent-ils être énoncés pour un système différentiel du type

$$\frac{dx}{dt} = v(x) \tag{3.2}$$

où $x = x(t)$ appartient à un ouvert U (un ouvert de l'espace d'état paramétré par les coordonnées locales x) de $\mathbb{R}^n$ et v est une application régulière de U dans $\mathbb{R}^n$. L'application v est appelée *champ de vecteurs* (vitesse), c.f. figure 3.1.

Comme v ne dépend pas du temps, le système est dit autonome. Tout système non autonome $\dfrac{dx}{dt} = v(x,t)$ peut être vu comme une partie d'un système autonome de plus grande dimension. Il suffit de poser $\tilde{x} = (x,t)$ et $\tilde{v}(\tilde{x}) = (v(x),1)$ et de considérer le système étendu $\dfrac{d\tilde{x}}{dt} = \tilde{v}(\tilde{x})$.

Théorème 1 (existence et unicité) *Considérons le système (3.2) et supposons le champ de vecteurs v continûment dérivable sur U. Pour tout x_0 dans U, il existe $a < 0 < b$ réels et une unique solution*

$$\begin{array}{rcl} \phi_\cdot(x_0) \ :]a,b[& \longrightarrow & U \\ t & \longrightarrow & \phi_t(x_0) \end{array}$$

satisfaisant (3.2) avec $x(0) = x_0$ ($\phi_0(x_0) = x_0$).

L'hypothèse de dérivabilité de v peut être affaiblie en supposant v localement lipschitzienne ($\|v(x) - v(y)\| \leq K\|x - y\|$ avec K constante de Lipschitz). Cette hypothèse de régularité sur la variation de v est indispensable pour l'unicité. En effet, l'équation scalaire $\dfrac{dx}{dt} = x^{2/3}$ admet deux solutions distinctes ayant la même condition initiale 0 : $t \to 0$ et $t \to t^3/27$.

L'intervalle de temps $]a,b[$ dépend a priori de x_0. Si x_0 évolue dans un compact de U, il est possible de borner inférieurement $|a|$ et b. Ce qui permet de considérer tout un ensemble de conditions initiales et ainsi de définir le flot. Le théorème 1 assure l'existence d'une

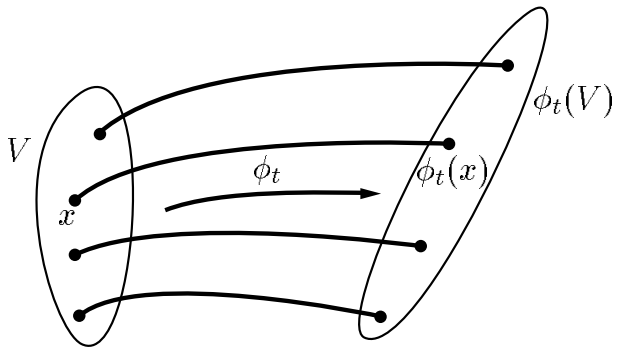


FIG. 3.2 – *transport d'un ensemble V par le flot ϕ_t*

solution sur un petit intervalle de temps autour de 0. En raison de l'unicité, deux solutions, qui coïncident au moins en un point, sont nécessairement égales. Comme l'illustre la figure 3.3 deux trajectoires distinctes d'un système autonome ne peuvent ni se recoller, ni se croiser. Cette propriété est tout à fait importante. Elle permet de définir la notion de

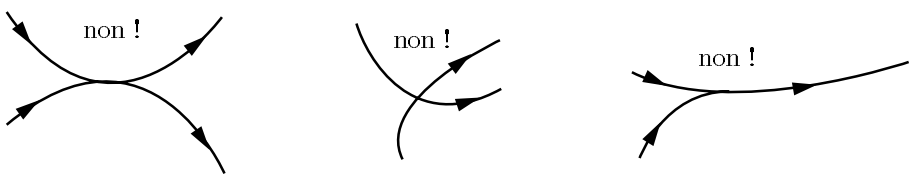


FIG. 3.3 – *jonction, tangence et intersection entre deux trajectoires différentes sont impossibles.*

courbe intégrale maximale, d'orbite (c.f. figure 3.5) et de flot.

Définition 1 (flot, trajectoire, orbite, portrait de phase) *Le champ de vecteurs v est appelé générateur infinitésimal du flot $\phi_t : U \to U$ défini par*

$$\frac{d}{dt}\left(\phi_t(x)\right)|_{t=\tau} = v(\phi_\tau(x)) \quad et \quad \phi_0(x) = x$$

pour $x \in U$ et τ entre 0 et t.

A $x \in U$ fixé, la courbe paramétrée $t \to \phi_t(x)$ est appelée trajectoire. Le lieu géométrique $\{\phi_t(x)\}_t$ est appelée orbite ou encore courbe intégrale passant par x. La partition de l'espace d'état en orbites est appelé portrait de phase.

Il faut noter que, pour x dans U, $\phi_t(x)$ est toujours défini pour t proche de 0. Le flot ϕ_t satisfait à des propriétés de groupe (lorsque les opérations sont définies) : $\phi_0 = I$ et $\phi_t \circ \phi_s = \phi_{t+s}$. Ainsi $x \to \phi_t(x)$ a pour réciproque $x \to \phi_{-t}(x)$.

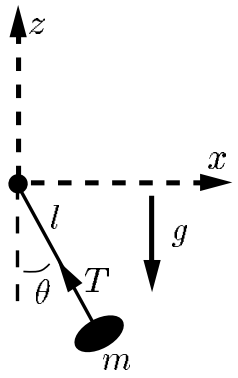


FIG. 3.4 – *le pendule ponctuel.*

Un pendule ponctuel de longueur l, soumis à une gravité g et d'angle θ par rapport à la verticale (c.f. figure 3.4) obéit à l'équation du second ordre

$$\frac{d^2\theta}{dt^2} = -\frac{g}{l}\ \sin\theta.$$

Cette équation différentielle se met sous la forme d'un système du premier ordre ayant deux équations

$$\frac{d\theta}{dt} = \omega, \quad \frac{d\omega}{dt} = -\frac{g}{l}\ \sin\theta \tag{3.3}$$

et deux inconnues $x = (\theta,\omega)$. L'espace engendré par x correspond au cylindre $\mathbb{S}^1 \times \mathbb{R}$, produit cartésien du cercle $\mathbb{S}^1$ et de la droite réelle $\mathbb{R}$ (l'angle θ est défini à 2π près et la vitesse angulaire varie de $-\infty$ à $+\infty$). $\phi_t(\theta,\omega) \in \mathbb{S}^1 \times \mathbb{R}$ admet donc deux composantes: elles correspondent à l'angle et à la vitesse du pendule à l'instant t sachant qu'à l'instant 0 l'angle était θ et la vitesse ω.

Définition 2 (intégrale maximale) *Pour une condition initiale x fixée, il est possible de choisir l'intervalle de temps $]a,b[$, sur lequel la solution $\phi_t(x)$ peut être définie, le plus grand possible : il correspond au prolongement maximal dans le passé ($t < 0$) et dans le futur ($t > 0$) de la solution passant par x à $t = 0$. On appelle intégrale maximale une telle solution. Une orbite correspond donc à l'ensemble des points de l'espace d'état décrit par une intégrale maximale. On porte habituellement sur le dessin d'un portrait de phases le sens de parcours des orbites.*

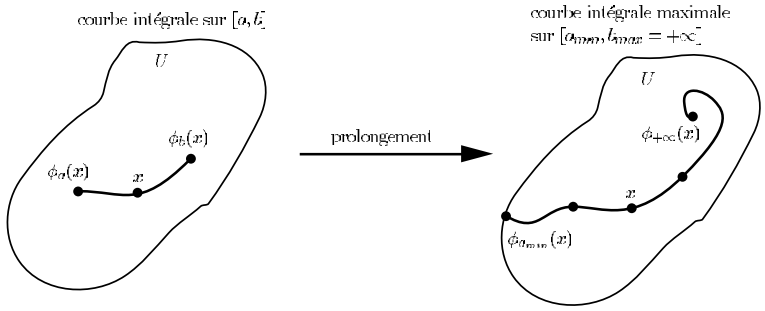


FIG. 3.5 – *exemple de prolongement maximum dans le passé et dans le futur d'une courbe intégrale.*

Les cas où l'intervalle $]a,b[$ de définition d'une intégrale maximale n'est pas $\mathbb{R}$ tout entier sont essentiellement les suivants (c.f. figure 3.6) :

- explosion en temps fini (la norme de la solution part vers l'infini) : l'exemple de base est le suivant $U = \mathbb{R}$ et $\dfrac{dx}{dt} = x^2 = v(x)$; $\phi_t(0) = 0$, $\phi_t(x) = \dfrac{-1}{t - 1/x}$ pour $x \neq 0$; si $x > 0$ alors l'intégrale maximale passant par x est définie sur $]-\infty, 1/x[$;
- la courbe intégrale arrive sur le bord du domaine U, en un temps fini, en un endroit où le vecteur vitesse $v(x)$ pointe soit vers l'extérieur de U (on dit que v est sortant) soit vers l'intérieur de U (on dit que v est rentrant) selon que l'on a $b \neq +\infty$ ou $a \neq -\infty$, respectivement.

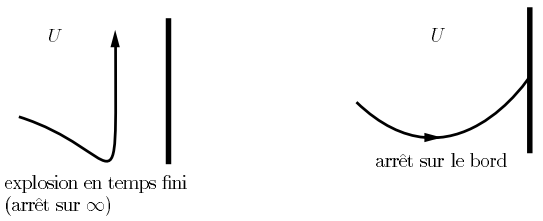


FIG. 3.6 – *les deux cas d'arrêt en temps fini d'une trajectoire.*

Les principaux cas où les courbes intégrales sont définies sur un intervalle de longueur infinie sont (figure 3.7) :

- soit $U = \mathbb{R}^n$ et Dv, la matrice jacobienne de v, est bornée sur $\mathbb{R}^n$ (évite les phénomènes d'explosion en temps fini) ; soit U est un domaine borné de $\mathbb{R}^n$ et le champ de vecteurs vitesse est tangent sur le bord de U (cas où v est nul sur le bord par exemple) ; dans les deux cas $a = -\infty$ et $b = +\infty$;
- si U est un domaine borné de $\mathbb{R}^n$ et si le champ de vecteurs est rentrant dans U, alors $b = +\infty$.
- U est une variété compacte.

Proposition 1 (dépendance régulière par rapport aux conditions initiales) *Soit le système (3.2) avec v continûment dérivable et $\{\phi_t\}$ le flot associé. Pour tout t, $x \mapsto \phi_t(x)$ est C^1. Sa dérivée, notée $D_x\phi_t(x)$ (la matrice $n \times n$ $\left(\dfrac{\partial[\phi_t]_i}{\partial x_j}\right)$), est solution de l'équation*

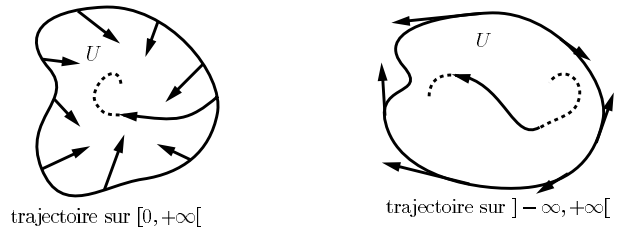


FIG. 3.7 – *courbes intégrales sur des intervalles de temps infinis.*

différentielle matricielle (première variation)

$$\frac{d}{dt}\left(D_x\phi_t(x)\right)_{t=\tau} = D_x v(\phi_\tau(x))\ D_x\phi_\tau(x)$$

avec comme condition initiale $D_x\phi_0(x) = I_n$. *De plus* $D_x\phi_t(x)\cdot v(x) = v(\phi_t(x))$.

Si v *dépend régulièrement d'un paramètre* λ ($v = v(x,\lambda)$) *alors le flot de* $v(\cdot,\lambda)$, $\{\phi_t^\lambda\}$ *dépend aussi régulièrement de* λ *et on a*

$$\frac{d}{dt}\left(D_\lambda\phi_t^\lambda(x)\right)_{t=\tau} = D_x v(\phi_\tau^\lambda(x),\lambda)\ D_\lambda\phi_\tau^\lambda(x) + D_\lambda v(\phi_\tau^\lambda(x),\lambda)$$

avec comme condition initiale $D_\lambda\phi_0^\lambda(x) = 0$.

Pour retrouver ces relations, il suffit de dériver par rapport à x et λ les relations définissant le flot,

$$\frac{d}{dt}\left(\phi_t^\lambda(x)\right)\big|_{t=\tau} = v(\phi_\tau^\lambda(x),\lambda), \quad \phi_0^\lambda(x) = x.$$

Montrons comment faire les calculs sur l'exemple (3.3) du pendule. Prenons des notation pour les deux composantes du flot :

$$\phi_t(\theta,\omega) = (\Theta_t(\theta,\omega),\Omega_t(\theta,\omega)).$$

On veut alors calculer les dérivées partielles de Θ et Ω par rapport à θ et ω, i.e., la matrice jacobienne :

$$D_{(\theta,\omega)}\varphi_t = \begin{pmatrix} \frac{\partial\Theta}{\partial\theta} & \frac{\partial\Theta}{\partial\omega} \\ \frac{\partial\Omega}{\partial\theta} & \frac{\partial\Omega}{\partial\omega} \end{pmatrix}.$$

Chaque colonne de cette matrice vérifie la même équation différentielle; seules les conditions initiales changent. Pour la première colonne, on obtient cette équation différentielle en dérivant (3.3) par rapport à θ :

$$\frac{d\left(\frac{\partial\Theta}{\partial\theta}\right)}{dt} = \frac{\partial\Omega}{\partial\theta}, \quad \frac{d\left(\frac{\partial\Omega}{\partial\theta}\right)}{dt} = -\frac{g}{l}\ \cos(\Theta_t(\theta,\omega))\ \frac{\partial\Theta}{\partial\theta}.$$

Les deux inconnues sont $\frac{\partial\Theta}{\partial\theta}$ et $\frac{\partial\Omega}{\partial\theta}$. Elles admettent comme conditions initiales 1 et 0 respectivement. Noter que les quantités θ et ω sont des paramètres fixes ici et que $\Theta_t(\theta,\omega)$ et $\Omega_t(\theta,\omega)$ sont considérées comme des fonctions du temps uniquement.

Ces notations sont assez lourdes. Pour faire les calculs, on leur préfère des notations moins rigoureuses mais bien plus commodes $\delta\theta$ et $\delta\omega$. Elles permettent d'obtenir rapidement l'équation différentielle ordinaire satisfaite par les dérivées partielles précédentes :

$$\frac{d(\delta\theta)}{dt} = \delta\omega, \quad \frac{d(\delta\omega)}{dt} = -\frac{g}{l}\ \cos(\theta(t))\ \delta\theta$$

où on a remplacé $\Theta_t(\theta,\omega)$ par $\theta(t)$ qui maintenant correspond à la valeur courante de l'angle. Cette équation correspond tout simplement au terme du 1er ordre en $\delta\theta$ et $\delta\omega$ dans (3.3) lorsque l'on remplace θ et ω par $\theta + \delta\theta$ et $\omega + \delta\omega$. Ce qui explique le nom de "première variation" donné à cette équation différentielle linéaire en $(\delta\theta,\delta\omega)$. Ainsi en même temps que l'on calcule une solution $(\theta(t),\omega(t))$ de (3.3) on peut ainsi calculer sa variation au premier ordre par rapport à une erreur de condition initiale $(\delta\theta_0,\delta\omega_0)$. Avec $(\delta\theta_0,\delta\omega_0) = (1,0)$ (resp. $(\delta\theta_0,\delta\omega_0) = (0,1)$ on obtient les dérivées partielles en θ (resp. ω).

Exercice 1 *Quelle est l'équation différentielle (avec sa condition initiale) vérifiée par la dérivée par rapport à l des solutions de (3.3).*

Nous voyons donc que deux trajectoires $\phi_t(x)$ et $\phi_t(y)$, ayant des conditions initiales voisines ($\|x - y\|$ petit), restent voisines l'une de l'autre sur un intervalle de temps *borné* ($\|\phi_t(x) - \phi_t(y)\|$ reste petit pour $0 \leq t \leq b$, $b < +\infty$).

Plus précisément, on a l'estimation a priori suivante (pour une démonstration voir [12]).

Proposition 2 *Si la norme de la matrice jacobienne $D_x v$ (norme matricielle issue de la norme sur les vecteurs de $\mathbb{R}^n$) est bornée sur U par une constante K alors, pour x et y dans U,*

$$\|\phi_t(x) - \phi_t(y)\| \leq \|x - y\| \exp(Kt).$$

Comme l'illustre la figure 3.8, rien ne dit que pour des temps grands, $t \gg 1/K$, les trajectoires restent encore voisines si elles le sont au départ. La majoration précédente peut très bien être une bonne approximation de l'écart entre deux trajectoires : la divergence est alors effectivement exponentielle sur des *temps longs*. Cette divergence peut être aussi interprétée comme une sensibilité importante par rapport aux conditions initiales. Cette propriété est l'une des caractéristiques des systèmes instables et dits chaotiques.

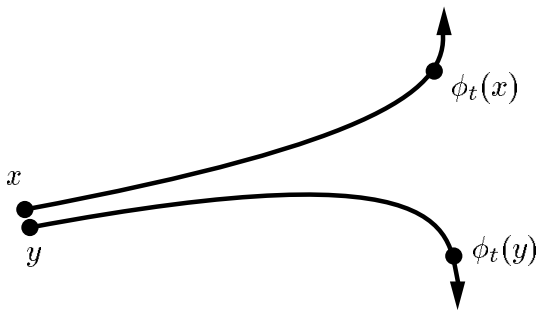


FIG. 3.8 – *sensibilité aux conditions initiales du flot $\{\phi_t\}$.*

Soit $y = f(x)$ un changement (local) de coordonnées sur U (par exemple le passage de coordonnées cartésiennes aux coordonnées polaires dans le plan). Autrement dit $f : x \to y =$

$f(x)$ est un difféomorphisme local. Alors, l'équation différentielle (3.2) devient dans les nouvelles variables y

$$\frac{dy}{dt} = \left(\frac{\partial f}{\partial x}\right)_{f^{-1}(y)} v(f^{-1}(y)) = w(y). \tag{3.4}$$

Il est alors clair que le flot $\psi_t(y)$, de générateur infinitésimal $w(y)$, est relié au flot $\phi_t(x)$, de générateur infinitésimal $v(x)$, par la relation

$$\psi_t(f(x)) = f(\phi_t(x)),$$

soit $\psi_t \circ f = f \circ \phi_t$ pour chaque t.

Nous allons voir que, autour d'un point où la vitesse v est non nulle, la structure locale du flot (i.e. des trajectoires) est particulièrement simple : comme l'illustre la figure 3.9, un changement de variables sur x (changement de coordonnées locales) permet de redresser le champ de vitesse en un champ constant arbitraire.

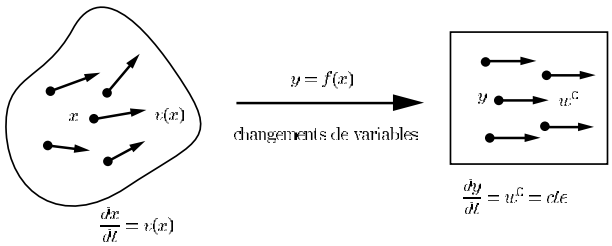


FIG. 3.9 – *structure locale du flot là où le champ des vitesses est non nul.*

On a le théorème suivant, dit théorème de redressement.

Théorème 2 (redressement) *Soit a dans U tel que $v(a) \neq 0$. Alors, il existe un difféomorphisme local f autour de a, $y = (y_1, \ldots, y_n) = f(x)$, qui transforme l'équation différentielle (3.2) dans la forme normale suivante :*

$$\frac{dy_1}{dt} = 0 \quad \ldots \quad \frac{dy_{n-1}}{dt} = 0 \quad \frac{dy_n}{dt} = 1.$$

La preuve de ce résultat est particulièrement simple si l'on fait une figure et que l'on raisonne géométriquement :

Preuve On peut supposer $a = 0$. Une fois que l'on a compris la construction de la figure 3.10, la preuve devient très simple. Il suffit d'introduire un hyperplan ne contenant pas $v(0)$, passant par 0 et dont la direction est définie par les vecteurs $(e_1, \ldots, e_{n-1})$ associés aux $n-1$ premières coordonnées de x (quitte à permuter des composantes de x, c'est toujours possible). Les nouvelles variables y, qui mettent le système localement sous la forme normale du théorème, sont alors données par $f = g^{-1}$ avec

$$g \ : \ y = (y_1, \ldots, y_{n-1}, y_n) \quad \longrightarrow \quad x = \phi_{y_n}(y_1, \ldots, y_{n-1}, 0)$$

où ϕ_t est le flot de v. Le fait que g soit un difféomorphisme local résulte aussitôt du théorème d'inversion locale. Par construction, la matrice jacobienne de g au point $y = 0$ est inversible. ■

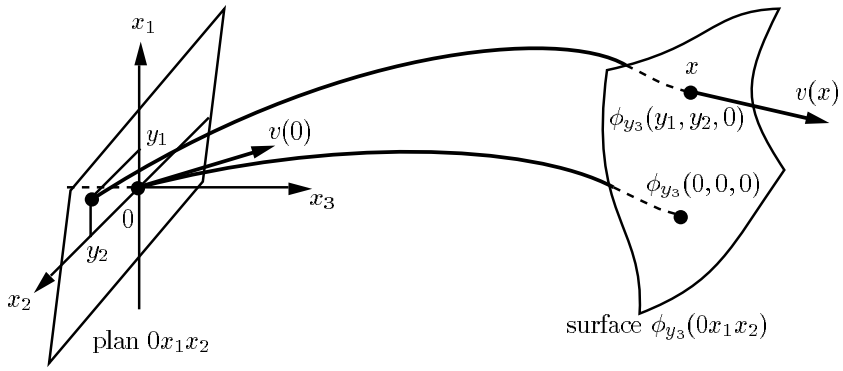


FIG. 3.10 – *preuve du théorème de redressement dans* $\mathbb{R}^3$.

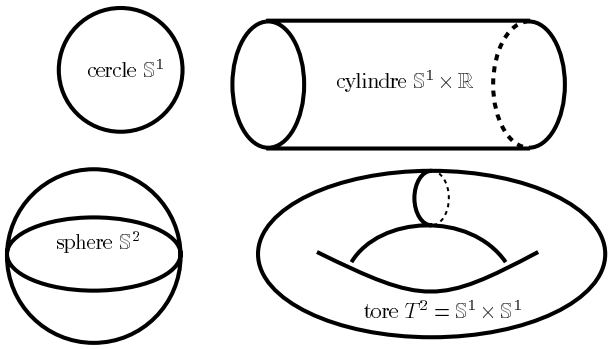


FIG. 3.11 – *espaces d'état les plus courants qui ne sont pas des ouverts sans trous (i.e. simplement connexes) de la droite* $\mathbb{R}$ *ou du plan* $\mathbb{R}^2$.

3.1.3 Remarque sur l'espace d'état

Dans les définitions précédentes, nous avons supposé que l'espace d'état est un ouvert U de $\mathbb{R}^n$. Or, pour une vision globale du flot, et en particulier des comportements sur de grands intervalles de temps du système, on est souvent obligé d'introduire la notion de variété d'état. Une variété abstraite peut être vue comme une mise bout à bout *globalement cohérente* d'ouverts de $\mathbb{R}^n$ correspondant, au moyen de coordonnées locales, à des petits morceaux (voisinages) de la variété (pour une définition mathématique d'une variété différentiable voir [3]). La notion de variété différentiable a pour origine l'étude des courbes (variété de dimension 1) et des surfaces (variété de dimension 2). La figure 3.11 rappelle les prototypes les plus classiques de variétés de dimension 1 et 2 :

- $\mathbb{S}^1$, le cercle, est le prototype des courbes fermées (orbite périodique) et donc apparaît très souvent au cours de l'étude de comportements périodiques.
- Le cylindre $\mathbb{S}^1 \times \mathbb{R}$ est la variété d'état naturelle du pendule plan : à chacun des points (θ,ω) du cylindre $\mathbb{S}^1 \times \mathbb{R}$ est associé un vecteur vitesse, tangent au cylindre (c.f. figure 3.12) au point considéré. La dynamique est alors déterminée par un champ de vecteurs tangents au cylindre.

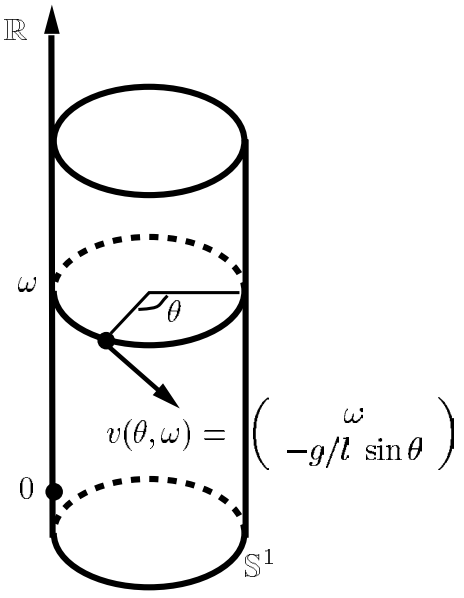


FIG. 3.12 – *l'espace d'état du pendule est le cylindre,* $(\theta,\dot{\theta} = \omega) \in S^1 \times \mathbb{R}$*, et le vecteur vitesse* $v(\theta,\omega)$ *est tangent au cylindre.*

- Le tore $T^2 = \mathbb{S}^1 \times \mathbb{S}^1$ apparaît naturellement lors de l'étude de deux oscillateurs.
- Bien que le tore T^2 ait la même dimension que la sphère $\mathbb{S}^2$, il n'est pas possible de faire correspondre *globalement* de façon régulière et biunivoque les points de T^2 et ceux de $\mathbb{S}^2$ (il est instructif d'essayer)[3]. Cette impossibilité est d'ordre topologique. Elle a des conséquences sur l'aspect global des champs de vecteurs tangents et sur les flots. Par exemple, il est possible de construire sur le tore T^2 un champ régulier de vecteurs tangents ne s'annulant jamais, alors que c'est impossible pour la sphère $\mathbb{S}^2$ (problème dit du hérisson).

3. Contrairement aux notations, ici trompeuses, $\mathbb{S}^1 \times \mathbb{S}^1$ n'est pas égal (difféomorphe) à $\mathbb{S}^2$.

3.1.4 Résolution numérique

Nous ne rappelons ici que des faits très élémentaires. La première idée qui vient à l'esprit est la récurrence suivante :

$$\frac{x^{n+1}_{\Delta t} - x^n_{\Delta t}}{\Delta t} = v(x^n_{\Delta t})$$

où $x^n_{\Delta t}$ serait une approximation de x à l'instant $t = n\ \Delta t$. Ce schéma est connu sous le nom de *schéma d'Euler explicite*. Il est d'ordre 1. Il est convergent. La convergence signifie ici la chose suivante : connaissant la condition initiale x^0 à $t = 0$, la solution $x(t)$ en $t = T > 0$ (quand elle existe) est alors la limite quand n tend vers $+\infty$ de $x^n_{\Delta t}$ ($x^0_{\Delta t} = x^0$) où $\Delta t = T/(n-1)$ dépend de n. La convergence n'est pas une propriété évidente à démontrer. Elle n'est pas directement reliée à l'ordre. Un schéma d'ordre 10 peut très bien être divergent et donc inutilisable. La difficulté vient du fait que plus le pas Δt est petit, plus le nombre d'itérations pour atteindre le temps final T est grand.

Le *schéma d'Euler implicite* correspond à la récurrence suivante

$$\frac{x^{n+1}_{\Delta t} - x^n_{\Delta t}}{\Delta t} = v(x^{n+1}_{\Delta t}).$$

Calculer $x^{n+1}_{\Delta t}$ nécessite la résolution d'une équation implicite et donc la mise en oeuvre de techniques type algorithme de Newton. Les calculs sont donc plus lourds. Ce schéma d'ordre 1 est convergent.

Les schémas implicites sont bien adaptés aux *systèmes raides*, c'est à dire, aux systèmes lents/rapides qui comportent une grande diversité d'échelles de temps, les échelles les plus rapides étant stables (c.f. la section sur la théorie des perturbations ci-dessous). En effet, il n'est nécessaire d'avoir un pas de temps Δt plus petit que l'échelle de temps la plus rapide comme c'est le cas pour les méthodes explicites. Aussi il peut être plus économique d'effectuer peu d'itérations avec un Δt assez grand (sachant que chaque itération coûte assez chère) plutôt que beaucoup d'itérations avec un Δt très petit.

Prenons un exemple :

$$\dot{x} = -x/\tau + y, \quad \dot{y} = -y/\varepsilon$$

($\tau >> \varepsilon$ sont deux paramètres positifs). Le schéma d'Euler explicite donne la récurrence

$$\begin{aligned} x^{n+1} &= (1 - \Delta t/\tau)\ x^n + \Delta t\ y^n \\ y^{n+1} &= (1 - \Delta t/\varepsilon)\ y^n. \end{aligned}$$

Cette récurrence est stable si $\Delta t < 2\varepsilon$. Le schéma implicite conduit à (la résolution est facile)

$$\begin{aligned} x^{n+1} &= \frac{1}{1 + \Delta t/\tau}\ \left(x^n + \frac{\Delta t}{1 + \Delta t/\varepsilon}\ y^n \right) \\ y^{n+1} &= \frac{1}{1 + \Delta t/\varepsilon}\ y^n, \end{aligned}$$

récurrence stable pour tout $\Delta t > 0$. Par exemple $\Delta t \approx \tau/10$ donne déjà une bonne approximation de la solution du système à des échelles de temps de l'ordre de τ. Nous

renvoyons le lecteur à [21] où sont présentées les méthodes numériques les plus classiques comme celle de Gear (prédicteur-correcteur) pour résoudre les systèmes raides.

Exercice 2 *Pour un système $\dot{x} = v(x)$ avec v régulier, montrer la convergence du schéma d'Euler explicite.*

3.1.5 Comportements asymptotiques

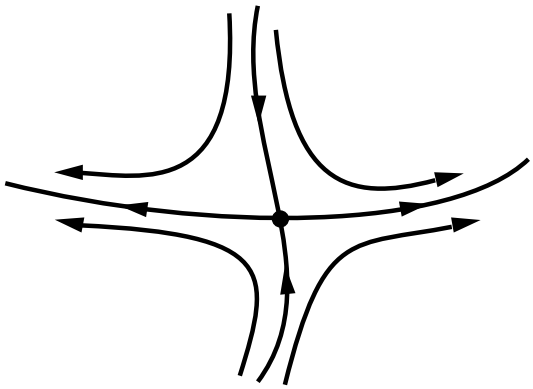


FIG. 3.13 – *exemple de point d'équilibre, le col.*

Là où la vitesse est non nulle, la structure locale du portrait de phases est très simple (théorème de redressement) : dans les bonnes coordonnées, les orbites sont des droites parallèles. En revanche, là où la vitesse s'annule, le portrait de phases peut être nettement plus compliqué. Pour s'en convaincre il suffit de comparer la figure 3.9, avec la figure 3.13. L'une des raisons essentielles de cette différence est que, pour étudier la structure des orbites autour d'un point où le vecteur vitesse s'annule, il faut considérer des intervalles de temps non bornés, contrairement au cas où la vitesse est non nulle.

Définition 3 (point d'équilibre) *Les points $\overline{x}$ où le champ de vitesse v s'annule sont appelés points critiques, ou points d'équilibre. Ils correspondent à des points fixes du flot : $\phi_t(\overline{x}) = \overline{x}$ pour tout t.*

Exercice 3 *Quels sont les points d'équilibre du pendule (3.3)?*

Un point d'équilibre est une trajectoire particulière. Une autre trajectoire particulière est la trajectoire qui se referme sur elle-même (c.f. figure 3.14).

Définition 4 (orbite périodique) *On appelle cycle, ou trajectoire périodique, ou encore orbite périodique, une trajectoire $\phi_t(x)$ qui n'est pas réduite à un point et telle qu'il existe $T > 0$ vérifiant $\phi_T(x) = x$. Le plus petit réel T strictement positif tel que $\phi_T(x) = x$ est appelé période. Elle est indépendante du point x pris sur la trajectoire.*

Les points d'équilibre et les orbites périodiques sont des exemples de sous-ensembles invariants dont la définition est donnée ci-dessous.

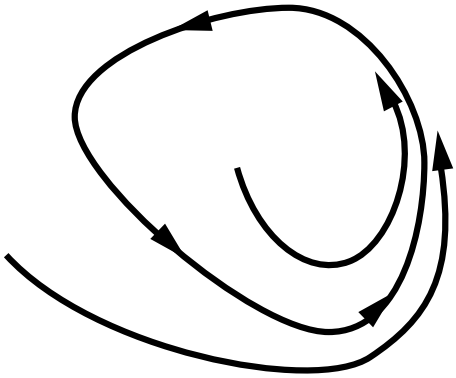


FIG. 3.14 – *le cycle limite attracteur.*

Définition 5 (ensemble invariant) *Soit A un sous-ensemble de l'espace d'état U. A est dit invariant (resp. positivement invariant) par le flot ϕ_t, si, pour tout t dans $\mathbb{R}$ (resp. dans $[0, +\infty[$), $\phi_t(A)$ est inclus dans A.*

D'autres exemples d'ensembles invariants sont fournis par les hypersurfaces de niveau d'une fonction réelle de l'espace d'état qui reste constante le long des trajectoires, i.e. une intégrale première.

Définition 6 (intégrale première) *On appelle intégrale première, une fonction C^1 $h : U \to \mathbb{R}$ telle que $\frac{d}{dt}[h(\phi_t(x))] = 0$ pour tout x dans U et pour tout t. Cette condition est équivalente à $D_x h(x) \cdot v(x) = 0$ pour tout x dans U (ce qui évite de connaître explicitement le flot). Ainsi les hypersurfaces de niveau, $\{x \in U \ : h(x) = c\}$ avec c constante réelle, sont invariantes par le flot.*

Géométriquement (c.f. figure 3.15) le champ de vitesses v est tangent aux hypersurfaces de niveau[4].

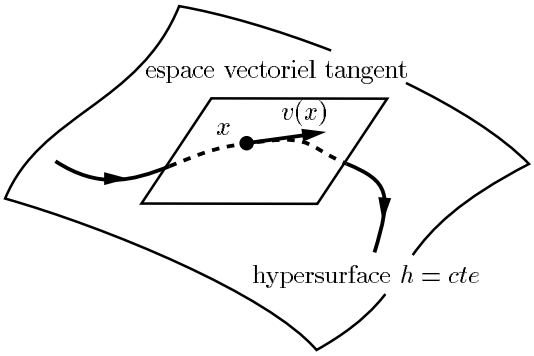


FIG. 3.15 – *h est une intégrale première de $\frac{dx}{dt} = v(x)$ lorsque le vecteur $v(x)$ est tangent aux hypersurfaces de niveau $h = cte$.*

4. En l'absence de point critique de h où ∇h s'annule.

Exercice 4 (intégrale première du pendule) *Montrer que (3.3) admet comme intégrale première* $\frac{1}{2}\omega^2 - g/l\cos\theta$. *En déduire l'équation des orbites. Dessiner l'allure du portrait de phase sur le cylindre* $\mathbb{S}^1 \times \mathbb{R}$.

Exercice 5 (intégrale première et énergie) *Les systèmes mécaniques holonomes parfaits (sans frottement) obéissent aux équations de Lagrange :*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{\partial L}{\partial q_i}, \quad i = 1,\ldots,n$$

où $q = (q_1,\ldots,q_n)$ *sont les coordonnées généralisées et* $L(q,\dot{q}) = T(q,\dot{q}) - U(q)$ *est le lagrangien, différence entre l'énergie cinétique* $T = \sum_{i,j} a_{i,j}(q)\dot{q}_i\dot{q}_j$ *et l'énergie potentielle* $U(q)$. *Montrer que l'énergie* $H = T + U$ *est une intégrale première du système.*

La notion d'intégrale première s'étend aux systèmes dynamiques régis par des équations aux dérivées partielles (EDP). Ces systèmes sont dits de dimension infinie car leur espace d'état, un espace fonctionnel, est de dimension infinie. L'exercice qui suit en est une illustration.

Exercice 6 (intégrale première pour les EDP) *La dynamique d'un fluide parfait incompressible dans une cavité* $\Omega \subset \mathbb{R}^3$ *obéit aux équations d'Euler*

$$\frac{\partial V_i}{\partial t} + \sum_{j=1}^{3} \frac{\partial V_i}{\partial x_j} V_j = -\frac{\partial p}{\partial x_i} \quad i = 1,2,3$$

$$\sum_{i=1}^{3} \frac{\partial V_i}{\partial x_i} = 0$$

$$\sum_{i=1}^{3} V_i n_i = 0 \text{ sur } \partial\Omega$$

où $x = (x_1,x_2,x_3)$ *sont des coordonnées cartésiennes,* $V(x,t) = (V_1,V_2,V_3)$ *le champ des vitesses (l'état du système),* p *la pression,* $n = (n_1,n_2,n_3)$ *la normale extérieure à la frontière* $\partial\Omega$. *Montrer que l'énergie cinétique*

$$T = \frac{1}{2}\int_\Omega (V_1^2 + V_2^2 + V_3^2)(x,t)\, dx$$

est constante si V *vérifie les équations d'Euler.*

Un prototype d'ensemble positivement invariant, très lié aux systèmes dits dissipatifs, est schématisé sur la figure 3.16. Soit K un sous-ensemble fermé et borné (compact) de U dont le bord ∂K est régulier (morceaux d'hypersurfaces). Si le champ de vitesse v est rentrant dans K, alors K est positivement invariant, i.e. $\phi_t(K)$ est contenu dans K pour tout $t \geq 0$. Il est alors naturel de considérer l'ensemble résiduel, lui aussi invariant, que l'on obtient par le flot à partir de K lorsque t tend vers $+\infty$: $A = \bigcap_{t\geq 0} \phi_t(K)$,

Ce cas type est à la base de la notion intuitive d'attracteur : ce vers quoi les trajectoires tendent lorsque le temps devient grand. Cette notion est difficile à définir d'une manière mathématiquement rigoureuse. Nous contenterons ici de la définition d'ensemble attracteur

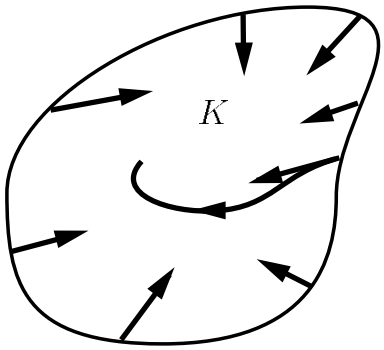


FIG. 3.16 – *exemple d'ensemble invariant K pour les temps positifs.*

Définition 7 (ensemble attracteur) *Un sous-ensemble fermé A de l'espace d'état est un ensemble attracteur s'il existe un ouvert V de l'espace d'état contenant A tel que, pour tout x dans V, $\phi_t(x) \in V$ pour $t \geq 0$ et $\phi_t(x) \to A$ lorsque $t \to +\infty$.*

3.1.6 L'étude qualitative ou le contenu des modèles

Les résultats précédents (existence et unicité des solutions, théorème de redressement) sont de nature locale en espace et en temps. Ils ne disent rien sur le comportement des solutions lorsque le temps devient grand. Les équations de Lorenz,

$$\left\{ \begin{array}{rcl} \dfrac{dx_1}{dt} & = & s(-x_1 + x_2) \\ \dfrac{dx_2}{dt} & = & rx_1 - x_2 - x_1x_3 \\ \dfrac{dx_3}{dt} & = & -bx_3 + x_1x_2. \end{array} \right.$$

sont d'apparence très simples, bien que, pour $s = 10$, $r = 28$ et $b = 8/3$, l'allure des solutions, sur de grands intervalles de temps, soit très irrégulière.

En général, l'objectif de la commande est d'éviter ce type d'instabilités et de comportements asymptotiques très irréguliers. Au contraire, nous cherchons à stabiliser le système autour d'un point ou d'une trajectoire. Nous abordons maintenant un cas élémentaire : les trajectoires convergent vers un point stationnaire (le régime limite du système est un point stationnaire).

Les développements qui suivent sont limités à quelques outils analytiques caractérisant la stabilité d'un point d'équilibre (valeurs propres du linéarisé tangent et fonction de Lyapounov). A cette occasion, on introduit la notion fondamentale d'hyperbolicité pour un point d'équilibre.

3.2 Points d'équilibre

Un point d'équilibre du système continu (3.2) correspond à ce que l'on appelle aussi un régime stationnaire. La question de la stabilité se pose alors en des termes très simples :

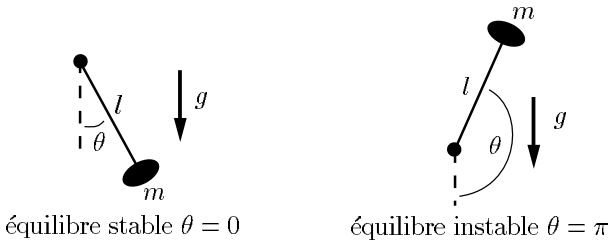


FIG. 3.17 – *les deux positions d'équilibre du pendule.*

si l'on écarte le système de l'équilibre, y reviendra-t-il ? Ou encore : une petite perturbation, qui éloigne légèrement le système de son régime stationnaire, peut-elle avoir des conséquences importantes et être amplifiée au cours du temps ?

3.2.1 Stabilité et fonction de Lyapounov

Prenons le pendule (3.3). Tout le monde connaît ses deux positions d'équilibre (figure 3.17) : celle du bas, $\theta = 0$, est stable (un petit écart n'entraîne que de petits effets) et celle du haut, $\theta = \pi$, est instable (un petit écart entraîne de grands effets). Si l'on tient compte du freinage de l'air, il est clair que l'équilibre du haut reste instable. L'équilibre du bas reste stable mais avec en plus un amortissement au cours du temps des petits écarts. On dit alors que l'équilibre du bas est *asymptotiquement stable* : au bout d'un certain temps, qui peut être grand si le freinage de l'air est faible, le pendule devient immobile (physiquement).

Ces questions de stabilité ont été étudiées par A.M. Lyapounov qui en a donné une définition assez générale englobant de nombreux systèmes physiques [16].

Définition 8 (stabilité locale) *Un point d'équilibre $\overline{x}$ de (3.2) est stable au sens de Lyapounov si, pour tout $\varepsilon > 0$, il existe $\eta > 0$ (dépendant de ε mais indépendant du temps t) tel que, pour tout x vérifiant $\|x - \overline{x}\| \leq \eta$, $\|\phi_t(x) - \overline{x}\| \leq \varepsilon$ pour tout $t > 0$.*

Dans un langage plus imagé : un petit déséquilibre initial n'entraine qu'un petit déséquilibre au cours du temps, déséquilibre qui peut très bien être permanent.

Définition 9 (stabilité asymptotique locale) *Un point d'équilibre $\overline{x}$ de (3.2) est asymptotiquement stable au sens de Lyapounov s'il est stable au sens de Lyapounov (c.f. définition 8) et si de plus, pour tout x suffisamment proche de $\overline{x}$, $\lim_{t\to+\infty} \phi_t(x) = \overline{x}$.*

Remarquons que ces définitions sont locales en espace : elles concernent uniquement les orbites voisines d'un point d'équilibre.

Revenons au pendule (3.3) et supposons que le freinage de l'air soit proportionnel à la vitesse angulaire $\omega = \dot{\theta}$. La dynamique du pendule est alors décrite par

$$\frac{d\theta}{dt} = \omega, \quad \frac{d\omega}{dt} = -\frac{g}{l} \sin\theta - \alpha\omega \tag{3.5}$$

où $\alpha > 0$ est le coefficient de frottement avec l'air divisé par la masse du pendule. L'énergie mécanique du pendule est proportionnelle à la fonction $V(\theta,\omega) = \omega^2/2 - (g/l)(\cos\theta - 1)$

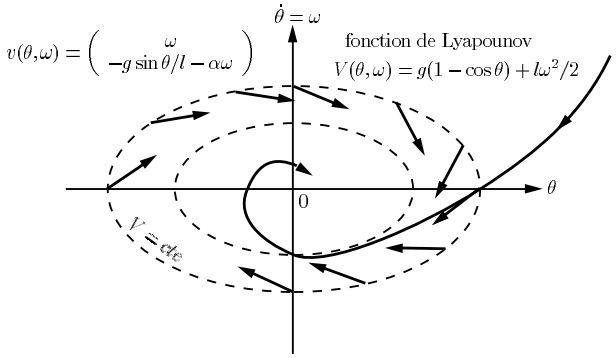


FIG. 3.18 – *stabilité asymptotique de l'équilibre du bas* ($\theta = 0$, $\omega = 0$) *du pendule en présence de frottement (portrait de phases local).*

(l'énergie mécanique de l'équilibre du bas $(\theta,\omega) = 0$ est prise égale à 0). La présence de frottement implique physiquement une dissipation d'énergie. Elle se traduit ici par le fait que V décroît le long des trajectoires,

$$\frac{dV}{dt} = -\alpha\omega^2 \leq 0,$$

le travail des forces de frottement est négatif.

V est appelé fonction de Lyapounov du système. L'intérêt d'une telle fonction est qu'il n'est pas nécessaire de résoudre explicitement l'équation différentielle (3.5) pour en déduire la stabilité de l'équilibre d'en bas. Contenu du fait que

$$V(\theta,\omega) \approx \omega^2 + (g/2l)\theta^2$$

pour (θ,ω) proche de 0, les ensembles $\{(\theta,\omega) : V(\theta,\omega) \leq \varepsilon\}$, avec $\varepsilon > 0$ petit, s'emboîtent les uns dans les autres autour de 0. La relation $\dot V \leq 0$ signifie géométriquement que le champ de vecteurs,

$$(\theta,\omega) \longrightarrow \begin{pmatrix} \omega \\ -g/l\sin\theta - \alpha\omega \end{pmatrix},$$

est rentrant dans cette famille d'ensembles emboîtés. On obtient ainsi la figure 3.18 qui montre clairement que l'équilibre inférieur est asymptotiquement stable au sens de Lyapounov.

Les développements des deux paragraphes précédents peuvent être rendus parfaitement rigoureux et correspondent à ce que l'on appelle *première méthode de Lyapounov* ou encore *méthode directe*. Ils s'appuient sur les deux résultats généraux suivants (démonstration dans [16]).

Théorème 3 (1ère méthode de Lyapounov, invariance de Lasalle) *Soient (3.2) avec $U = \mathbb{R}^n$ (pour simplifier) et une fonction C^1, $V : \mathbb{R}^n \to [0,+\infty[$, telle que :*

- *si $x \in \mathbb{R}^n$ tend vers l'infini en norme, $V(x)$ tend aussi vers l'infini ;*
- *V décroît le long de toutes les trajectoires, $\dfrac{dV}{dt} \leq 0$.*

Alors, toutes les trajectoires sont définies sur $[0, +\infty[$ *et convergent asymptotiquement vers le plus grand ensemble invariant (c.f. définition 5) contenu dans l'ensemble défini par* $D_x V \cdot v = 0$.

Une fonction V vérifiant les hypothèses du théorème 3 est appelée **fonction de Lyapounov** (globale). Le principe d'invariance consiste simplement à écrire le système sur-déterminé suivant,

$$\dot{x} = v(x), \quad D_x V(x) \cdot v(x) = 0,$$

système caractérisant le plus grand ensemble invariant contenu dans $\frac{dV}{dt} = 0$.

Pour le pendule avec frottement (3.5), $\dot{V} = 0$ s'écrit $\omega = 0$. Pour savoir vers quoi tendent les solutions, nous avons à résoudre le système sur-déterminé suivant :

$$\begin{aligned} \frac{d\theta}{dt} &= \omega \\ \frac{d\omega}{dt} &= -g/l \sin\theta - \alpha\omega \\ \omega &= 0. \end{aligned}$$

Les solutions sont $\theta = 0,\pi$ et $\omega = 0$. Ainsi quelque soit la condition initiale, les trajectoires tendent soit vers l'équilibre du haut soit vers l'équilibre du bas.

Exercice 7 *Dessiner, en utilisant les résultats qui précèdent et ceux de l'exercice 4, le portrait de phase de (3.5) avec* $\alpha > 0$ *petit. En déduire également le portrait de phase pour* $\alpha < 0$ *petit en valeur absolue.*

Ce théorème global admet aussi une version locale autour d'un point d'équilibre.

Théorème 4 *Si* $\overline{x}$ *est un point d'équilibre de (3.2) et si la fonction* C^1, $V : U \to [0,+\infty[$, *est telle que :*

- $V(\overline{x}) = 0$ *et* $V(x) > 0$ *pour* $x \neq \overline{x}$ *;*
- V *décroît le long de toutes les trajectoires (*$\frac{dV}{dt} \leq 0$*).*

Alors $\overline{x}$ *est stable au sens de Lyapounov. Si l'on suppose en plus que* $\frac{dV}{dt} < 0$ *si* $x \neq \overline{x}$, *alors* $\overline{x}$ *est asymptotiquement stable au sens de Lyapounov. Si l'on suppose encore en plus que* $V(x)$ *tend vers l'infini lorsque* $x \in \mathbb{R}^n$ *tend vers l'infini, toutes les trajectoires, même celles qui démarrent loin de* $\overline{x}$, *tendent vers* $\overline{x}$ *: on dit alors que le point* $\overline{x}$ *est globalement asymptotiquement stable.*

Ces deux théorèmes restent valables même si la fonction de Lyapounov V n'est pas aussi régulière. Par exemple V peut être supposée continue et uniquement dérivable par morceaux. Pour de plus amples détails voir [16].

Exercice 8 (stabilité pour des systèmes dépendant du temps) *Nous reprenons ici une idée très simple(cf. aussi l'exercice 11 pour des prolongements en dimension infinie).*

Soit le système dépendant du temps $\dot{x} = v(x,t)$, $x \in \mathbb{R}^n$ où à chaque instant la partie symétrique de la matrice jacobienne $D_x v = \left(\frac{\partial v_i}{\partial x_j}\right)$ est définie négative. Soient deux solutions distinctes $\alpha(t)$ et $\beta(t)$ de $\dot{x} = v(x,t)$.

1. *Montrer que la distance euclidienne entre α et β décroît au cours temps, i.e., que $r(t) = (\alpha(t) - \beta(t))^2$ décroît.*
2. *Que doit-on rajouter comme hypothèse sur la partie symétrique de $D_x v$ pour avoir la convergence exponentielle de α vers β, i.e., pour avoir l'existence de $a > 0$ (indépendant de α et β), tel que, pour tout $t > 0$, $r(t) \leq r(0)\exp(-at)$.*

Comme les intégrales premières, les fonctions de Lyapounov existent aussi pour les systèmes au dérivées partielles comme le montrent les exercices qui suivent.

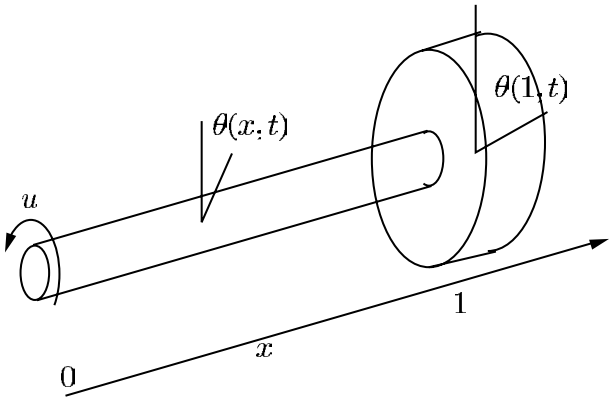


FIG. 3.19 – *poutre en torsion.*

Exercice 9 (poutre en torsion) *Une poutre en torsion (c.f. figure 3.19) autour d'un axe admet en élasticité linéaire le modèle suivant*

$$\begin{aligned}
&\partial_t^2\theta(x,t) = \partial_x^2\theta(x,t), \quad x \in [0,1] \\
&\partial_x\theta(0,t) = -u \\
&\partial_x\theta(1,t) = -\partial_t^2\theta(1,t),
\end{aligned}$$

Les conditions aux limites viennent du fait qu'en $x = 0$ la poutre est solidaire d'un moteur exerçant un couple u, qu'en $x = 1$ la poutre est solidaire d'une inertie.

1. *Calculer la dérivée le long des trajectoires de l'énergie mécanique*

$$T = \int_0^1 \frac{1}{2}\left([\partial_t\theta(x,t)]^2 + [\partial_x\theta(x,t)]^2\right)\, dx + \frac{1}{2}[\partial_t\theta(1,t)]^2.$$

   En déduire pour $u = 0$ (le système est libre) que T est une intégrale première.
2. *On dispose d'un capteur de vitesse en $x = 0$ (on connaît donc à chaque instant la vitesse du moteur $\partial_t\theta(0,t)$). Comment ajuster u de façon à faire décroître T (donner une interprétation physique).*

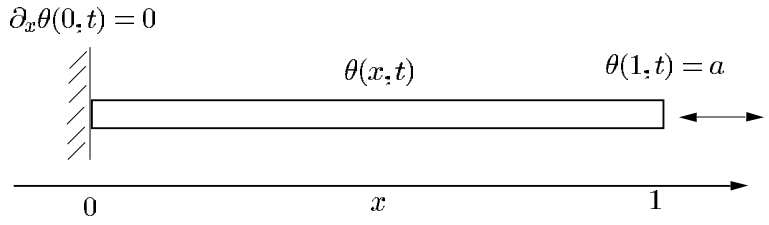


FIG. 3.20 – *équation de la chaleur pour une barre homogène.*

Exercice 10 (équation de la chaleur) *L'évolution du profil de température $\theta(x,t)$ (c.f. figure 3.20) dans une barre homogène isolée du coté $x=0$ et en contact avec un thermostat à la température constante a en $x=1$ est*

$$\begin{aligned}\partial_t\theta(x,t) &= \partial_x^2\theta(x,t), \quad x\in[0,1]\\ \partial_x\theta(0,t) &= 0\\ \theta(1,t) &= a\end{aligned}$$

Montrer que

$$\int_0^1 (\theta(x,t)-a)^2 \; dx$$

décroît au cours du temps. En déduire (formellement) que θ tend vers a.

Exercice 11 (réaction diffusion et entropie) *Un système enfermé dans le domaine $\Omega\subset\mathbb{R}^3$, isolé de l'extérieur, siège de diffusion et de réactions chimiques, peut être représenté par le modèle suivant*

$$\begin{aligned}\partial_t C(x,t) &= div\,(M(C)grad\,C)+v(C), \quad x\in\Omega\\ \frac{\partial C}{\partial\nu} &= 0 \text{ sur } \partial\Omega \text{ (ν normale extérieure)}\end{aligned}$$

avec $C=(C_1,\ldots,C_n)$ le vecteur des concentrations et n le nombre d'espèce chimiques. $M(C)$, la matrice des coefficients de diffusion, est symétrique définie positive (relations d'Onsager). $v=(v_1,\ldots,v_n)$ correspond aux cinétiques des diverses réactions chimiques. Nous supposons qu'il existe a tel que $v(a)=0$ et que la partie symétrique de la matrice jacobienne $\left(\frac{\partial v_i}{\partial C_j}\right)$ est symétrique définie négative. Montrer que

$$\int_\Omega (C(x,t)-a)^2 \; dx$$

décroît au cours du temps (cette quantité peut être interprétée comme l'opposée d'une entropie et sa décroissance comme la croissance de l'entropie). En déduire (formellement) que le profil de concentration C tend vers le profil homogène a lorsque t tend vers $+\infty$ (ce qui est bien conforme au second principe de la thermodynamique).

3.2.2 Les systèmes linéaires

Cette sous-section ne comporte que le strict minimum sur les systèmes linéaires. Pour un exposé complet avec démonstration, nous renvoyons à [12]. Nous considérons le système linéaire

$$\frac{dx}{dt} = Ax \tag{3.6}$$

avec $x \in \mathbb{R}^n$ et A une matrice $n \times n$ constante.

L'exponentielle d'une matrice

La matrice dépendant du temps $\exp(tA)$ est définie par la série absolument convergente

$$\exp(tA) = \left[I + tA + \frac{t^2}{2!}A^2 + \ldots + \frac{t^k}{k!}A^k + \ldots \right] \tag{3.7}$$

où I est la matrice identité. Toute solution de (3.6) passant par x à $t = 0$ s'exprime sous la forme

$$\exp(tA)\ x = \phi_t(x).$$

Voici les principales propriétés de l'exponentielle :

$$\begin{aligned}
\exp(tA)\ \exp(sA) &= \exp((t+s)A) \\
\frac{d}{dt}(\exp(tA)) &= \exp(tA)\, A \\
\exp(PAP^{-1}) &= P \exp(A) P^{-1} \\
\exp(A) &= \lim_{m \to +\infty} \left(I + \frac{A}{m} \right)^m \\
\det(\exp(A)) &= \exp(\mathrm{tr}(A))
\end{aligned}$$

où t et s sont des réels, P est une matrice inversible, "det" désigne le déterminant et "tr" désigne la trace.

Soient deux matrices carrées A_1 et A_2 de même taille. En général $\exp(A_1 + A_2) \neq \exp(A_1)\exp(A_2)$ car le produit de matrices n'est pas commutatif. Pour avoir l'égalité, on peut supposer que A_1 et A_2 commutent : $A_1 A_2 = A_2 A_1$. Ainsi, le flot d'un système linéaire dépendant du temps $\dot{x} = A(t)x$ n'admet pas d'expression simple avec des exponentielles : en général, $x(t) \neq \exp(\int_0^t A(\tau)d\tau)\ x(0)$. L'égalité a lieu si $A(t_1)$ et $A(t_2)$ commutent pour tout t_1, t_2. Pour s'en convaincre prendre l'équation $\ddot{x} = (a + bt)x$ qui n'admet pas de quadrature simple (fonction d'Airy).

Portrait de phases

Nous allons considérer maintenant les cas les plus intéressants, principalement les cas génériques (i.e. stables par petites perturbations des éléments de A), que l'on peut rencontrer en dimensions $n = 2$ et $n = 3$.

Dimension $n = 2$ Les principaux cas sont résumés sur les figures 3.21, et 3.22. λ_1 et λ_2 sont les valeurs propres de A (distinctes ou non, réelles ou complexes conjuguées), ξ_1 et ξ_2 sont les vecteurs propres réels associés quand ils existent.

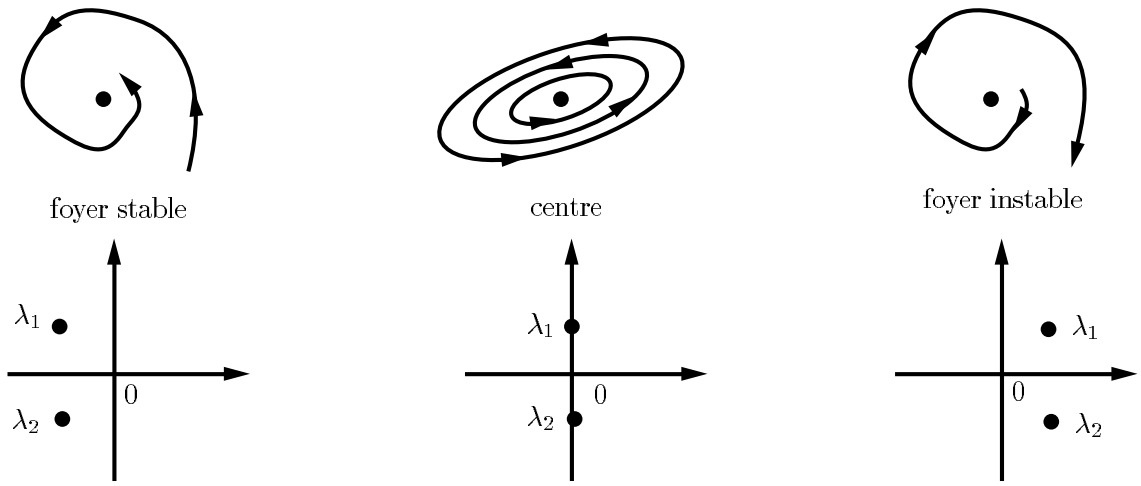


FIG. 3.21 – *portraits de phases plans et linéaires lorsque les deux exposants caractéristiques, λ_1 et λ_2, ont une partie imaginaire non nulle.*

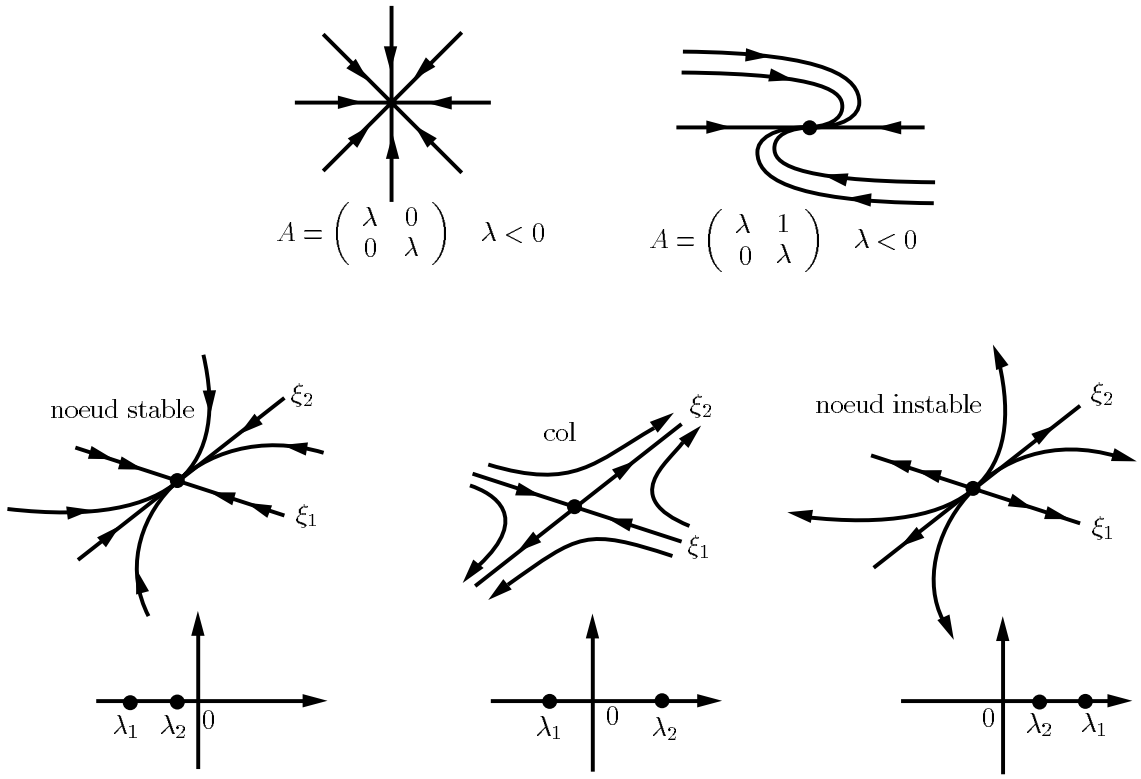


FIG. 3.22 – *portraits de phases plans et linéaires, $\dot{x} = Ax$, lorsque les exposants caractéristiques, λ_1 et λ_2, sont réels (ξ_1 et ξ_2 vecteurs propres de A, lorsqu'ils existent).*

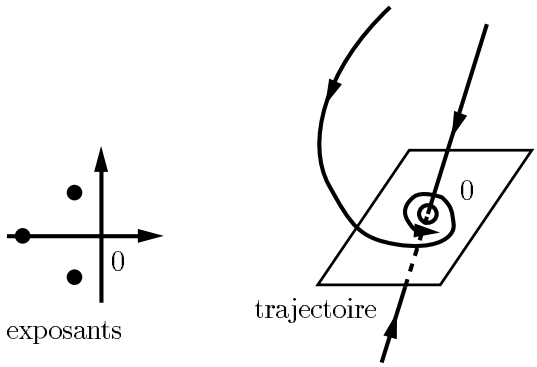


FIG. 3.23 – *exemple de portrait de phases d'un système linéaire de dimension* 3 *en fonction de ses exposants caractéristiques.*

Dimension $n = 3$ La figure 3.23, montre sur un exemple comment, à partir des portraits de phases en dimension 2, on construit, dans les cas génériques, le portrait de phases en dimension 3 : il suffit de décomposer $\mathbb{R}^3$ en somme d'espaces propres invariants de dimension 1 ou 2.

Forme de Jordan et calcul de l'exponentielle d'une matrice

Le calcul de $\exp(tA)$ peut être simplifié en faisant intervenir une transformation P inversible qui diagonalise A, lorsque c'est possible, ou qui transforme A en une matrice diagonale par blocs, dite matrice de Jordan (c.f. [12]). En dimension 2, on peut ainsi toujours se ramener aux trois formes normales de Jordan suivantes :

$$A = P \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} P^{-1} \quad \text{et} \quad \exp(tA) = P \begin{pmatrix} \exp(\lambda_1) & 0 \\ 0 & \exp(\lambda_2) \end{pmatrix} P^{-1}$$

$$A = P \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} P^{-1} \quad \text{et} \quad \exp(tA) = \exp(\alpha t)\, P \begin{pmatrix} \cos(\beta t) & -\sin(\beta t) \\ \sin(\beta t) & \cos(\beta t) \end{pmatrix} P^{-1}$$

$$A = P \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} P^{-1} \quad \text{et} \quad \exp(tA) = \exp(\lambda t)\, P \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} P^{-1}.$$

En dimension 3, une matrice A possède toujours une valeur propre réelle λ et un vecteur propre réel. Si l'on suppose que λ n'est pas une valeur propre de multiplicité 3, ce qui est très exceptionnel, on a

$$A = P \begin{pmatrix} \lambda & 0 \\ 0 & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{pmatrix} P^{-1}$$

avec P matrice d'ordre 3 inversible et λ, a, b, c et d réels. On se ramène ainsi à la dimension 2.

Exercice 12 (système linéaire dans le plan) *Pour un système linéaire $\dot{x} = Ax$ de dimension 2, établir en fonction de la trace et du déterminant de A, des différents portraits de phases possibles.*

3.2.3 Lien avec le linéaire tangent

Une méthode (dite indirecte ou seconde méthode de Lyapounov) pour analyser la stabilité autour d'un point d'équilibre $\overline{x}$ de $\dot{x} = v(x)$ consiste à étudier le système linéarisé tangent :

$$\frac{d(\Delta x)}{dt} = D_x v(\overline{x})\,\Delta x \quad \text{où} \quad D_x v(\overline{x}) = \left(\frac{\partial v_i}{\partial x_j}(\bar{x})\right)\Big|_{1\le i,j\le n}.$$

On a alors les deux résultats suivants.

Théorème 5 *Soit $\overline{x}$ un point d'équilibre de (3.2). Si les valeurs propres de $Dv(\overline{x})$ sont toutes à partie réelle* strictement *négative, alors $\overline{x}$ est un équilibre asymptotiquement stable au sens de Lyapounov.*

Cette condition suffisante sur les valeurs propres du linéarisé tangent n'est pas une condition nécessaire comme le montre l'équation scalaire $\dfrac{dx}{dt} = -x^3$ dont les solutions $t \to^+_- \sqrt{1/(t-a)}$ convergent toutes vers 0 quand t tend vers $+\infty$.

Preuve Elle consiste à construire une fonction de Lyapounov pour le linéaire tangent et à montrer que c'est aussi une fonction de Lyapounov locale pour le système non linéaire. Pour construire cette fonction de Lyapounov, nous pouvons utiliser la connaissance explicite du flot du linéaire tangent via l'exponentielle. La connaissance explicite du flot du système non linéaire étant hors de portée, c'est la seule manière de procéder.

Quitte à changer de notations, on suppose $\bar{x} = 0$. Notons $A = Dv(0)$. Comme les valeurs propres de A sont toutes à partie réelle négative, l'intégrale suivante est absolument convergente ($'$ signifie transpose),

$$Q = \int_0^{+\infty} \exp(tA')\exp(tA)\ dt,$$

sa valeur Q est une matrice symétrique strictement positive car pour tout t, $\exp(tA')\exp(tA)$ est symétrique définie positive ($\exp(tA)$ est inversible). Montrons que $V(x) = (1/2)x'Qx$ est une fonction de Lyapounov de $\dot{x} = v(x)$ autour de 0. Clairement V est positive, bornée inférieurement. On a, en développant v à l'ordre 1 en 0,

$$\dot{V} = x'Q\dot{x} = x'Q(Ax + o(\|x\|)).$$

Or

$$x'QAx = \int_0^{+\infty} x'\exp(tA')\exp(tA)Ax\ dt.$$

Comme $d/dt(\exp(tA)) = \exp(tA)A$,

$$\exp(tA')\exp(tA)A = (1/2)\frac{d}{dt}(\exp(tA')\exp(tA)).$$

Ainsi

$$\int_0^{+\infty} x'\exp(tA')\exp(tA)Ax\ dt = -x'x/2.$$

et donc

$$\dot{V} = x'Q\dot{x} = -\|x\|^2/2 + o(\|x\|^2)).$$

Ce qui montre que $\dot{V} < 0$ si $x \neq 0$ est proche de zéro. Le théorème 4 permet alors de conclure. ■

De façon très similaire, on a la condition suffisante (mais non nécessaire) d'instabilité suivante.

Théorème 6 *Soit $\overline{x}$ un point d'équilibre de (3.2). Si l'une des valeurs propres de $Dv(\overline{x})$ possède une partie réelle* strictement *positive alors $\overline{x}$ n'est pas un équilibre stable au sens de Lyapounov.*

Exercice 13 (équation de Lyapounov) *Montrer que $\dot{x} = Ax$ est asymptotiquement stable si, et seulement si, pour tout matrice Q symétrique définie positive, il existe une matrice P, symétrique définie positive, vérifiant l'équation dite de Lyapounov*

$$PA + A'P + Q = 0.$$

(Considérer la fonction de Lyapounov $V(x) = x'Px$ et l'intégrale $\int_0^{+\infty} exp(tA')Q\exp(tA)\,dt$).

Les valeurs propres du linéarisé tangent en $\overline{x}$ ne dépendent pas des coordonnées locales autour de $\overline{x}$ (ce qui est faux si $\bar{x}$ n'est pas un point d'équilibre). En effet, si $y = g(x)$ est un changement de variable local en $\overline{x}$, alors (3.2) s'écrit, dans les coordonnées y,

$$\frac{dy}{dt} = Dg(g^{-1}(y))\ v(g^{-1}(y)).$$

Un calcul simple montre que la matrice du linéarisé tangent en $\overline{y} = g(\overline{x})$ est semblable à $D_x v(\bar{x})$. Ainsi les valeurs propres sont les mêmes : ce sont des invariants par changement de variables.

Définition 10 (exposants caractéristiques, hyperbolicité) *Soit $\overline{x}$ un point stationnaire de (3.2), $v(\overline{x}) = 0$. Les valeurs propres de $Dv(\overline{x})$ sont appelées exposants caractéristiques du point d'équilibre $\overline{x}$. Le point d'équilibre $\overline{x}$ est dit hyperbolique si tous ses exposants caractéristiques sont à partie réelle non nulle.*

Pour un système commandé, $\dot{x} = f(x,u)$, et un point d'équilibre $(\bar{x},\bar{u})$ $(f(\bar{x},\bar{u}) = 0$, les pôles en boucle ouverte sont les exposants caractéristiques du système autonome $\dot{x} = v(x) = f(x,\bar{u})$ à l'équilibre $\bar{x}$.

Exercice 14 *Discuter en fonction du signe du coefficient α la stabilité et l'hyperbolicité des deux points d'équilibre du pendule (3.5).*

Noter (c.f. figure 3.24) que l'absence de stabilité au sens de Lyapounov n'implique nullement que les trajectoires, qui démarrent près de $\overline{x}$, ne convergent pas, quand t tend vers $+\infty$, vers $\overline{x}$.

Reprenons la signification du théorème de redressement dans une logique "développement limité". Ce théorème signifie que le premier terme $v(\bar{x})$ du développement en série de v autour de $\overline{x}$,

$$v(x) = v(\overline{x}) + Dv(\overline{x})\,(x - \overline{x}) + \ldots,$$

est suffisant pour avoir l'allure du portrait de phases autour de $\overline{x}$ si $v(\overline{x}) \neq 0$.

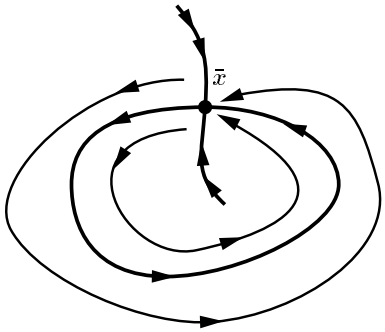


FIG. 3.24 – *exemple de point d'équilibre $\overline{x}$, instable au sens de Lyapounov, mais dont toutes les trajectoires, initialement proches de $\overline{x}$, convergent vers $\overline{x}$ lorsque $t \to +\infty$.*

Pour un point d'équilibre, le premier terme de cette série est nul ($v(\overline{x}) = 0$), il est alors naturel de considérer le second terme, c'est à dire le système linéarisé tangent au point stationnaire :

$$\frac{d\Delta x}{dt} = Dv(\overline{x})\,\Delta x$$

avec $\Delta x = x - \overline{x}$. Des résultats précédents sur la stabilité d'un point d'équilibre, il ressort que, même si la matrice $Dv(\overline{x})$ est inversible, on ne peut rien dire, en général sur le portrait de phases autour de $\overline{x}$ (i.e. des trajectoires qui démarrent près de $\overline{x}$).

Si le terme linéaire du développement limité n'est pas suffisant pour en déduire la stabilité locale (point d'équilibre non hyperbolique), il convient alors d'utiliser les termes d'ordre supérieur et des techniques vraiment non linéaires comme la variété centrale et l'éclatement de singularités [11, 4].

3.3 Systèmes dynamiques discrets

Un système dynamique discret (suite récurrente) est de la forme

$$x_{k+1} = G(x_k) \tag{3.8}$$

où G est une application régulière (un difféomorphisme en général) d'un ouvert U de $\mathbb{R}^n$ dans lui même. Le système continu (3.2) peut être étudié comme un système discret si, au lieu de considérer son flot continu ϕ_t, on considère $\tau > 0$ ("sorte" de période d'échantillonnage) et l'application associée

$$\begin{array}{rccc} G\ :\ & U & \to & U \\ & x & \to & G(x) = \phi_\tau(x) \end{array} .$$

Comme $\phi_\tau \circ \phi_\tau = \phi_{2\tau}$, il est clair que l'étude de ϕ_t lorsque $t \to +\infty$ et celle de

$$G^k = \underbrace{G \circ G \circ \ldots \circ G}_{k \text{ fois}}$$

lorsque l'entier k tend vers $+\infty$ doivent être très similaires.

Nous rappelons ici, succinctement, comment les notions et résultats précédents, introduits pour les systèmes continus, se transposent aux systèmes discrets.

3.3.1 Point fixe et stabilité

Définition 11 (point fixe, multiplicateurs, hyperbolicité) *Soit le système discret (3.8). Un point fixe $\overline{x}$ est défini par la relation $G(\overline{x}) = \overline{x}$. Les valeurs propres du jacobien de G en $\overline{x}$, $DG(\overline{x})$, sont appelées multiplicateurs caractéristiques de G en $\overline{x}$. Le point fixe $\overline{x}$ est dit hyperbolique si aucun de ses multiplicateurs caractéristiques n'est de module égal à 1.*

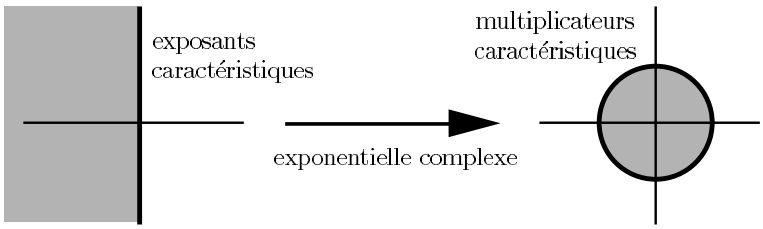


FIG. 3.25 – *la fonction exponentielle envoie le demi-plan des complexes à partie réelle négative dans l'intérieur du cercle unité.*

La figure 3.25 illustre comment l'exponentielle complexe permet de passer des exposants caractéristiques aux multiplicateurs caractéristiques, et justifie la terminologie.

Définition 12 (stabilité locale) *Un point fixe $\overline{x}$ de (3.8) est stable au sens de Lyapounov si, pour tout $\varepsilon > 0$, il existe $\eta > 0$ (dépendant de ε mais indépendant du nombre d'itérations k) tel que, pour tout x vérifiant $\|x - \overline{x}\| \leq \eta$, $\|G^k(x) - \overline{x}\| \leq \varepsilon$ pour tout entier $k > 0$.*

Définition 13 (stabilité asymptotique locale) *Un point fixe $\overline{x}$ de (3.8) est asymptotiquement stable au sens de Lyapounov s'il est stable au sens de Lyapounov et si, de plus, pour tout x suffisamment proche de $\overline{x}$, $\lim_{k\to+\infty} G^k(x) = \overline{x}$.*

Pour étudier la stabilité autour d'un point fixe $\overline{x}$, il est souvent utile d'étudier le système linéarisé tangent :

$$\Delta x_{k+1} = DG(\overline{x})\ \Delta x_k$$

où $\Delta x = x - \overline{x}$ correspond à un petit écart par rapport à $\overline{x}$. On a alors les deux résultats suivants.

Proposition 3 *Soit $\overline{x}$ un point fixe de (3.8). Si ses multiplicateurs caractéristiques sont tous de module strictement inférieur à 1, alors $\overline{x}$ est asymptotiquement stable au sens de Lyapounov.*

Cette condition de stabilité sur les multiplicateurs caractéristiques n'est pas nécessaire. Elle n'est que suffisante. On a aussi la condition suffisante (mais non nécessaire) d'instabilité suivante.

Proposition 4 *Soit $\overline{x}$ un point fixe de (3.8). Si l'un des multiplicateurs caractéristiques de $\overline{x}$ est de module strictement supérieur à 1, alors $\overline{x}$ est n'est pas stable au sens de Lyapounov.*

La preuve utilise simplement le fait que G est une contraction locale autour de $\bar{x}$.

3.3.2 Les systèmes linéaires discrets

Nous considérons ici le système linéaire discret suivant

$$x_{k+1} = Ax_k \tag{3.9}$$

avec $x \in \mathbb{R}^n$ et $A \in \mathbb{R}^n \times \mathbb{R}^n$ constant. L'étude des comportements asymptotiques de la suite récurrente x_k repose sur le calcul des puissances successives de A. Comme pour les équations différentielles linéaires à coefficients constants, il est commode d'utiliser la décomposition en blocs de Jordan.

En dimension 2, on a uniquement les 3 cas suivants ($(\lambda,\lambda_1,\lambda_2,\alpha,\theta)$ réels, P matrice 2×2 inversible) :

$$A = P \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} P^{-1} \quad \text{et} \quad A^k = P \begin{pmatrix} \lambda_1^k & 0 \\ 0 & \lambda_2^k \end{pmatrix} P^{-1}$$

$$A = \alpha\, P \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} P^{-1} \quad \text{et} \quad A^k = \alpha^k\, P \begin{pmatrix} \cos(k\theta) & -\sin(k\theta) \\ \sin(k\theta) & \cos(k\theta) \end{pmatrix} P^{-1}$$

$$A = P \begin{pmatrix} \lambda & 0 \\ 1 & \lambda \end{pmatrix} P^{-1} \quad \text{et} \quad A^k = P \begin{pmatrix} \lambda^k & 0 \\ k\lambda^{k-1} & \lambda^k \end{pmatrix} P^{-1}.$$

En dimension 3, on se ramène, sauf cas exceptionnel, à la dimension 2 par la décomposition de A suivante :

$$A = P \begin{pmatrix} \lambda & 0 \\ 0 & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{pmatrix} P^{-1}$$

avec P matrice d'ordre 3 inversible et λ, a, b, c et d réels.

A partir des calculs précédents, il est assez simple de dessiner l'allure des trajectoires, i.e. les portraits de phases, $(A^k(x))^{k\geq 0}$, dans $\mathbb{R}^2$ et $\mathbb{R}^3$. Les figures 3.26 et 3.27 en donnent quelques uns.

3.4 Stabilité structurelle et robustesse

Nous n'avons pas encore abordé une question centrale : la stabilité structurelle que l'on retrouve en automatique avec la notion de robustesse.

Un système dynamique est dit structurellement stable si, et seulement si, les portraits de phases de tous les systèmes "voisins" sont topologiquement équivalents. Deux systèmes sont dits topologiquement équivalents si et seulement s'il existe une homéomorphisme (bijection continue et d'inverse continue) entre les espaces d'état qui transforment les orbites de l'un en les orbites de l'autre, en préservant le sens de parcours des orbites[5]

Par exemple, si un système structurellement stable admet un seul point d'équilibre asymptotiquement stable $\overline{x}$ hyperbolique, alors, tout système "voisin" admet aussi un seul point d'équilibre, proche de $\overline{x}$, asymptotiquement stable et hyperbolique.

Il convient bien sûr de définir ce qu'est un système "voisin" : le plus simple consiste à perturber le champ de vitesse $v(x)$ par addition d'un champ $\delta v(x)$ petit en norme (la norme peut aussi porter sur les dérivées en x, $D_x(\delta v)$,...) et à considérer alors le système dynamique $v(x) + \delta v(x)$ comme système voisin.

Il est clair que cette question possède des motivations physiques importantes. En effet, toute modélisation est une approximation. Il est donc normal de s'intéresser aux

5. Il n'est pas possible de conserver la paramétrisation en temps car alors les périodes des orbites périodiques de deux systèmes topologiquement équivalents seraient rigoureusement égales. Ce qui est beaucoup trop contraignant.

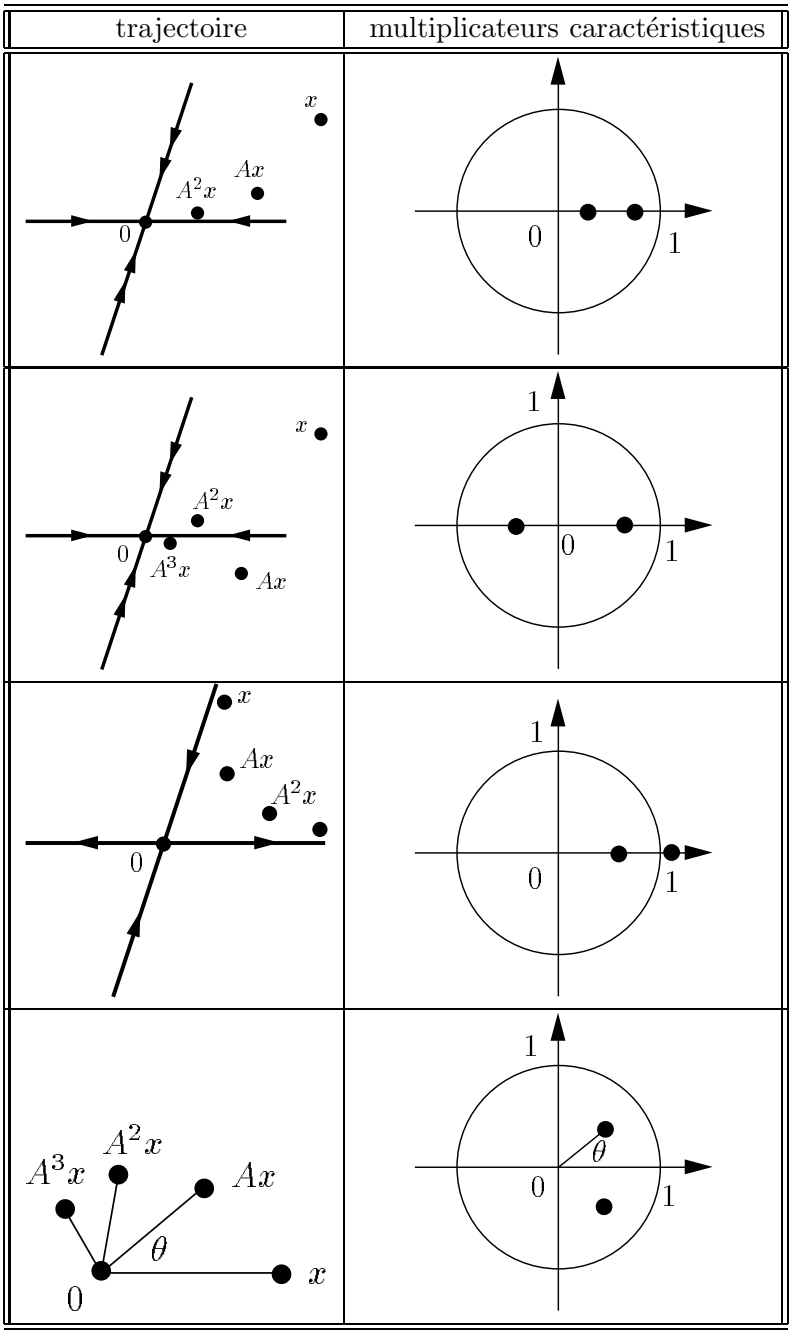


FIG. 3.26 – *portraits de phases de systèmes linéaires discrets dans le plan, $x_{k+1} = Ax_k$, en fonction de leurs multiplicateurs caractéristiques.*

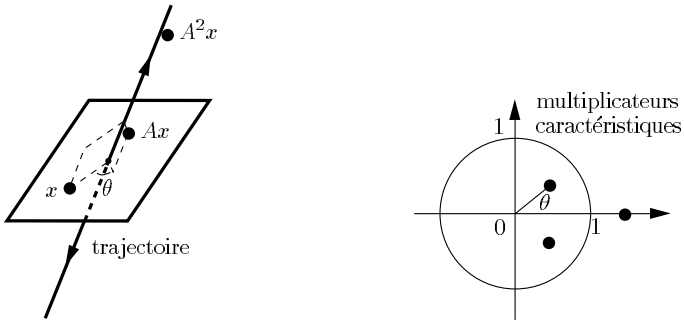


FIG. 3.27 – *exemple de système linéaire hyperbolique discret, $x_{k+1} = Ax_k$, de dimension 3.*

systèmes voisins du système dynamique de modélisation. En particulier, il apparaît important de savoir si les comportements asymptotiques contenus dans le système issu de la modélisation sont persistants et stables aux petites perturbations des équations, i.e. du champ des vitesses. D'où le nom de stabilité structurelle (à ne pas confondre avec la stabilité asymptotique) donné à ces questions.

Par exemple, un système $\dfrac{dx}{dt} = v(x)$ qui admet un point stationnaire $\overline{x}$ dont l'un des exposants caractéristiques est à partie réelle nulle, n'est pas structurellement stable. En effet, de petites perturbations δv du champ des vitesses v induisent sur la matrice $Dv(\overline{x})$, ainsi que sur ses valeurs propres (les exposants caractéristiques), des perturbations dans toutes les directions. Or, la stabilité asymptotique est une propriété invariante par équivalence topologique. Donc, nécessairement, un tel système ne peut pas être structurellement stable pour des perturbations aussi générales. En revanche, il peut très bien le rester pour des perturbations plus spécifiques, i.e. une topologie plus fine qui restreint la classe des systèmes voisins possibles.

La mise en forme des idées évoquées ci-dessus nécessite l'utilisation de notions mathématiques assez élaborées qui débordent largement le cadre de cet exposé d'introduction. Un lecteur intéressé pourra consulter d'abord [1], et pour en savoir plus [5, 9, 7].

En automatique, la notion de commande robuste est directement liée au problème suivant (cas linéaire). Etant donné $\dot{x} = Ax + Bu$, une incertitude sur la dynamique Mx de taille $\|M\| \leq \varepsilon$ (le vrai système est en fait $\dot{x} = (A + M)x + Bu$), calculer une borne ε sur $\|M\|$ pour qu'il existe un bouclage K stabilisant le système perturbé. Ainsi, il faut trouver K tel que $\dot{x} = (A + M + BK)x$ soit stable pour toute incertitude M vérifiant $\|M\| \leq \varepsilon$.

Ce problème est difficile : la dépendance des valeurs propres de la matrice $A+M+BK$ en fonction de M et de K est loin d'être triviale. De nombreuses méthodes existent et répondent en partie à cette question (méthodes LMI (Linear Matrix Inequalities), commande H^∞, marge de gain et marge de phase en fréquentiel, ...).

Pour des incertitudes $M(x)$ non linéaires, il est illusoire de vouloir donner des bornes en général. Tout au plus, peut-on espérer le résultat perturbatif suivant : si l'incertitude $M(x)$ est suffisamment petite et l'état initial assez proche de 0, alors le système restera asymptotiquement stable dès que le linéaire tangent bouclé est asymptotiquement stable $\dot{x} = (A+BK)x$ est stable. Dans ce cours, nous en resterons à ce niveau. Nous verrons donc la robustesse comme une conséquence de la stabilité structurelle des points d'équilibre

hyperboliquement stables.

3.5 Théorie des perturbations

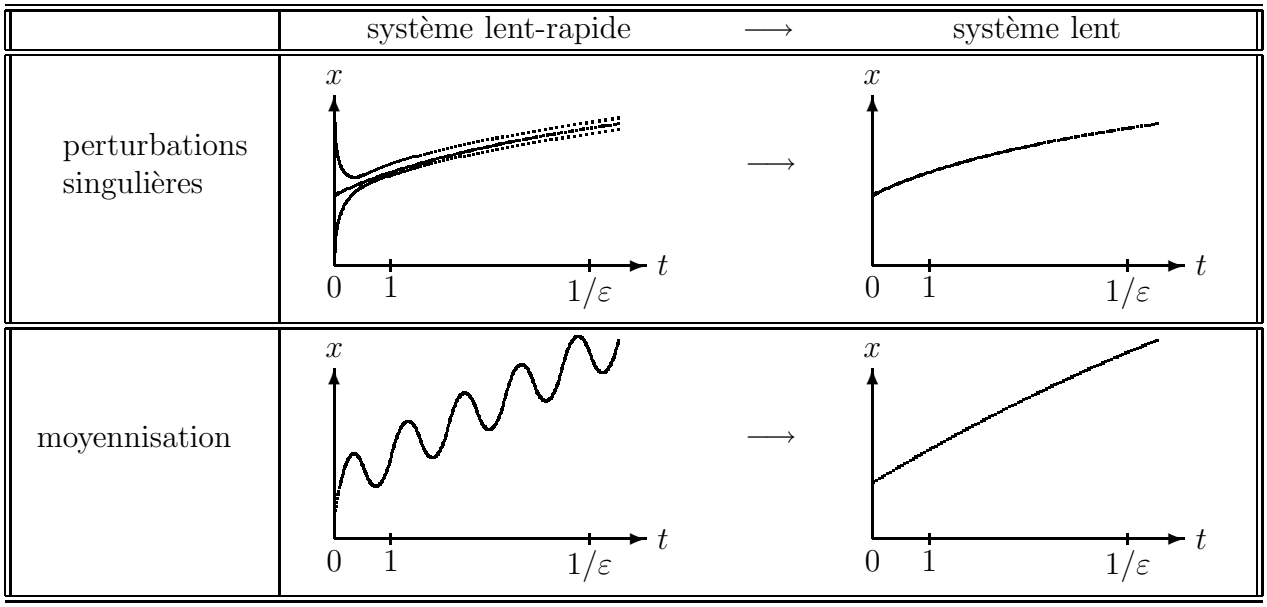


FIG. 3.28 – *la théorie des perturbations consiste à éliminer les effets à court terme, $t \sim 1$, qu'ils soient asymptotiquement stables ou oscillants, afin de ne conserver que les effets à long terme, $t \sim 1/\varepsilon$ $(0 < \varepsilon \ll 1)$.*

Par rapport à la stabilité structurelle, qui suppose un nombre fixe d'équations différentielles (on ne change pas d'espace d'état mais seulement le champ de vitesse), la théorie des perturbations permet de relier les trajectoires de deux systèmes ayant des espaces d'état de dimensions différentes : le système dit perturbé possède alors un nombre d'états plus grand que le système dit réduit. Plus précisément, cette théorie fournit un ensemble de techniques pour approximer un système perturbé, en éliminant les effets à court terme et en ne conservant que les effets à long terme. Ainsi, la théorie des perturbations constitue un outil précieux pour l'étude d'un système dynamique et de *son approximation par des systèmes lents de taille plus petite*, c'est à dire pour la construction de *modèles réduits* qui résument l'essentiel des comportements qualitatifs à long terme.

Classiquement, on distingue deux cas illustrés par la figure 3.28 :

- les effets rapides se stabilisent très vite et on parle alors de perturbations singulières et d'approximation quasi-statique ;
- les effets rapides ne sont pas asymptotiquement stables mais restent d'amplitude bornée ; ils sont donc oscillants et l'on parle alors de moyennisation.

Ces deux cas font l'objet des deux principales parties de cette section.

On considère les systèmes continus (une analyse similaire peut être conduite pour les

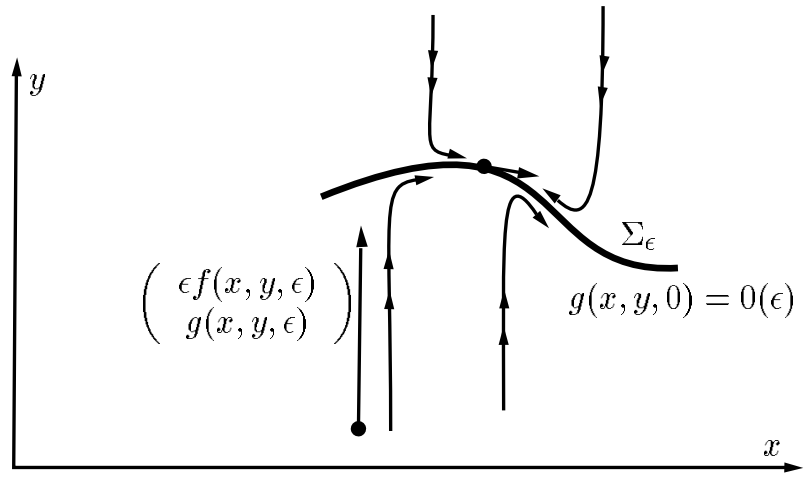


FIG. 3.29 – *Le champ de vecteur est quasi-vertical pour la forme normale de Tikhonov (3.10)*

systèmes discrets) du type :

$$\begin{cases} \dfrac{dx}{dt} &= \varepsilon f(x,y,\varepsilon) \\ \dfrac{dy}{dt} &= g(x,y,\varepsilon) \end{cases} \tag{3.10}$$

avec $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $0 < \varepsilon \ll 1$ un petit paramètre, f et g des fonctions régulières. L'état partiel x correspond aux variables dont l'évolution est lente (variation significative sur une durée en t de l'ordre $1/\varepsilon$) et y aux variables dont l'évolution est rapide (variation significative sur une durée en t de l'ordre de 1). Ainsi $t \approx 1$ correspond à l'échelle de temps rapide et $t \approx 1/\varepsilon$ à l'échelle de temps lente. Dans (3.10) l'échelle de temps est donc rapide. Avec le changement de temps $\tau = \varepsilon t$, on obtient les équations du système en échelle de temps τ lente :

$$\begin{cases} \dfrac{dx}{d\tau} &= f(x,y,\varepsilon) \\ \varepsilon\dfrac{dy}{d\tau} &= g(x,y,\varepsilon). \end{cases} \tag{3.11}$$

3.5.1 Les perturbations singulières

Considérons pour commencer l'exemple suivant

$$\begin{cases} \dfrac{dx}{dt} = \varepsilon y \\ \dfrac{dy}{dt} = x - y \end{cases} \tag{3.12}$$

avec $0 < \varepsilon \ll 1$. Intuitivement, on voit que x est une variable lente (sa vitesse est petite et d'ordre ε), tandis que y est une variable rapide (sa vitesse est d'ordre 1). On a donc envie

de dire que y atteint rapidement son point d'équilibre x et que x évolue selon $\dot{x} = \varepsilon x$. Cette idée est fondamentalement correcte, mais contient un certain nombre de subtilités comme par exemple celle de l'exercice qui suit.

Exercice 15 (chercher l'erreur) *On reprend (3.12) et le raisonnement intuitif précédent. Les arguments suivants conduisent à une contradiction : on fait l'hypothèse que y a atteint son point d'équilibre, i.e., $y = x$. Mais cela implique d'une part que $\dot{y} = 0$, car $\dot{y} = x - y$. D'une autre part $\dot{y} = \dot{x}$ puisque $y = x$; mais alors $\dot{y} = -\varepsilon y$ puisque $\dot{x} = \varepsilon x$. Ce qui contredit $\dot{y} = 0$. Où est l'erreur ?*

On suppose ici que les effets rapides sont asymptotiquement stables et hyperboliques. Comme exemple caractéristique citons la cinétique chimique où les constantes de vitesses de certaines réactions peuvent être nettement plus grandes que d'autres (réactions limitatives et réactions quasi-instantanées).

La situation géométrique est donnée par la figure 3.29 : grossièrement, pour $\varepsilon > 0$ assez petit et localement autour de $g(x,y,0) = 0$, les trajectoires du système sont quasi-verticales et convergent toutes vers une sous-variété de l'espace d'état, Σ_ε, invariante par la dynamique, et donnée à l'ordre 0 en ε par l'équation $g(x,y,0) = 0$.

Les résultats ci-dessous justifient alors, sous certaines hypothèses de stabilité du rapide et du lent, l'approximation des trajectoires du système perturbé (3.10) par celle du système semi-implicite lent :

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & \varepsilon f(x,y,\varepsilon) \\ 0 & = & g(x,y,\varepsilon) \end{array} \right. \tag{3.13}$$

(les dynamiques de convergence vers la sous-variété invariante Σ_ε sont négligées). Toute trajectoire du système (3.10) démarrant en (x,y) est proche, après une durée en t de l'ordre de 1, de la trajectoire du système lent démarrant avec le même x.

Nous voyons que cette approximation s'accompagne d'une diminution de la dimension de l'état. En fait, *la réduction n'est qu'une restriction à une sous-variété invariante* Σ_ε du champ de vecteurs donné par (3.10), les équations de cette sous-variété étant approximativement données par $g(x,y,0) = 0$. On a le premier résultat général suivant (démonstration dans [8])

Théorème 7 (Tikhonov) *Soit le système (3.10). Supposons que*

H1 *l'équation $g(x,y,0) = 0$ admet une solution, $y = h(x)$, avec h fonction régulière de x et*

$$\frac{\partial g}{\partial y}(x,h(x),0)$$

est une matrice dont toutes les valeurs propres sont à partie réelle strictement négative ;

H2 *le système réduit*

$$\left\{ \begin{array}{rcl} \dfrac{dx}{d\tau} & = & f(x,h(x),0) \\ x_{(\tau=0)} & = & x_0 \end{array} \right. \tag{3.14}$$

avec $\tau = \varepsilon t$ admet une solution $x^0(\tau)$ pour $\tau \in [0,T]$, $0 < T < +\infty$.

Alors, pour ε suffisamment proche de 0, le système complet

$$\begin{cases} \dfrac{dx}{d\tau} &= f(x,y,\varepsilon) \quad x(0) = x_0 \\ \varepsilon\dfrac{dy}{d\tau} &= g(x,y,\varepsilon) \quad y(0) = y_0 \end{cases}$$

admet une solution $(x^\varepsilon(\tau),y^\varepsilon(\tau))$ sur $[0,T]$ dès que y_0 appartient au bassin d'attraction du point d'équilibre $h(x_0)$ du sous-système rapide

$$\frac{d\xi}{dt} = g(x_0,\xi,0).$$

De plus on a

$$\lim_{\varepsilon\to 0^+} x^\varepsilon(\tau) = x^0(\tau) \quad et \quad \lim_{\varepsilon\to 0^+} y^\varepsilon(\tau) = y^0(\tau)$$

uniformément en temps sur tout intervalle fermé contenu dans $[0,T]$ et ne contenant pas 0.

L'hypothèse *H1* implique que, à x fixé, la dynamique de ξ

$$\frac{d\xi}{dt} = G(x,\xi,0).$$

est asymptotiquement stable autour du point d'équilibre $h(x)$.

Exercice 16 (réduction de schéma cinétiques) *Soit le schéma cinétique comportant 3 espèces chimiques X_1, X_2 et X_3 et mettant en jeu trois réactions chimiques indépendantes :*

$$X_1 \xrightarrow{k_1x_1} X_2, \quad X_2 \xrightarrow{k_2x_2} X_1, \quad X_1 + X_2 \xrightarrow{\varepsilon k x_1x_2} X_2 + X_3;$$

les x_i correspondent aux concentrations des espèces X_i; k_1, k_2 et εk sont les constantes cinétiques. Le petit paramètre $\varepsilon > 0$ indique que la troisième réaction est nettement plus lente que les deux premières. Les équations de conservation de chacune des espèces conduisent, pour un réacteur fermée homogène, aux équations différentielles suivantes :

$$\begin{aligned} \dot{x}_1 &= -k_1x_1 + k_2x_2 - \varepsilon k x_1x_2 \\ \dot{x}_2 &= k_1x_1 - k_2x_2 \\ \dot{x}_3 &= \varepsilon k x_1x_2. \end{aligned} \tag{3.15}$$

1. *Montrer que $\zeta = x_1 + x_2 + x_3$ est une intégrale première (les chimistes parlent d'invariant chimique). En déduire que seules les deux premières équations de (3.15) sont importantes.*
2. *Le modèle lent est-il obtenu en faisant brutalement $\dot{x}_2 = 0$, i.e., l'approximation $k_1x_1 = k_2x_2$ dans l'équation de $\dot{x}_1$? (indication : considérer le changement de variables $(x_1,x_2) \mapsto (x_1+x_2,x_2)$; établir les équations du modèle lent dans les nouvelles variables; repasser ensuite aux variables d'origine (x_1,x_2)).*

Sans hypothèses supplémentaires l'approximation du théorème 7 n'est valable, en général, que sur des intervalles de temps rapides t de longueur T/ε, i.e sur *des intervalles de temps lents τ de longueur bornée T*. L'hypothèse supplémentaire, qu'il convient

alors d'utiliser pour avoir une bonne approximation pour tous les temps positifs, concerne le comportement asymptotique du système réduit : si ce dernier admet un point d'équilibre hyperbolique et asymptotiquement stable, l'approximation est alors valable pour tous les temps positifs (pourvu que les conditions initiales soient proches de cet équilibre).

Théorème 8 *Supposons en plus des hypothèses du théorème 7 que le système réduit (3.14) admet un point d'équilibre hyperboliquement stable* $\bar{x}$ *:* $f(\bar{x},h(\bar{x}),0) = 0$ *et que les valeurs propres de*

$$\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\frac{\partial h}{\partial x}$$

en $x = \bar{x}$, $y = h(\bar{x})$ *sont à partie réelle strictement négative. Alors, pour tout* $\varepsilon \geq 0$ *assez proche de* 0*, le système perturbé (3.10) admet un point d'équilibre proche de* $(\bar{x},h(\bar{x}))$ *et hyperboliquement stable.*

Preuve L'existence du point stationnaire pour le système perturbé est laissée en exercice (il suffit d'utilise le théorème des fonctions implicites pour $g = 0$ et ensuite pour $f = 0$). Quitte à faire, pour chaque ε une translation, nous supposons que (0,0) est point stationnaire du système perturbé :

$$f(0,0,\varepsilon) = 0, \quad g(0,0,\varepsilon) = 0.$$

Notons, $y = h_\varepsilon(x)$, la solution, proche de $x = 0$, de $g(x,y,\varepsilon) = 0$. Suite à la translation précédente, on a $h_\varepsilon(0) = 0$. Considérons le changement de variables $(x,y) \mapsto (x,z = y - h_\varepsilon(x))$. Les équations du système perturbé dans les coordonnées (x,z) ont alors la forme suivante

$$\dot{x} = \varepsilon\tilde{f}(x,z,\varepsilon), \quad \dot{z} = \tilde{g}(x,z,\varepsilon)$$

avec $\tilde{f}(0,0,\varepsilon) = 0$, $\tilde{g}(x,0,\varepsilon) \equiv 0$. Le système réduit s'écrit alors, dans les coordonnées (x,z) :

$$\dot{x} = \varepsilon\tilde{f}(x,0,\varepsilon).$$

Ce changement de variable triangularise le jacobien du système perturbé :

$$\begin{pmatrix} \varepsilon\frac{\partial\tilde{f}}{\partial x}(0,0,\varepsilon) & \varepsilon\frac{\partial\tilde{f}}{\partial z}(0,0,\varepsilon) \\ 0 & \frac{\partial\tilde{g}}{\partial z}(0,0,\varepsilon) \end{pmatrix}$$

car $\dfrac{\partial\tilde{g}}{\partial x}(0,0,\varepsilon) = 0$. Comme les valeurs propres de $\frac{\partial\tilde{f}}{\partial x}(0,0,0)$ et de $\frac{\partial\tilde{g}}{\partial z}(0,0,0)$ sont toutes à parties réelles strictement négatives, les valeurs propres de $\frac{\partial\tilde{f}}{\partial x}(0,0,\varepsilon)$ et $\frac{\partial\tilde{g}}{\partial z}(0,0,\varepsilon)$ le sont aussi pour ε assez petit. Ce qui montre la stabilité asymptotique du système perturbé pour tout ε assez petit. ■

Cette preuve peut être améliorée pour montrer que l'approximation du théorème 7 devient valide, localement autour de $(\bar{x},h(\bar{x}))$ et pour tous les temps t positifs, dés que ε est assez petit (le caractère local étant alors indépendant de ε tendant vers zéro).

En automatique le théorème 8 est utilisé de la manière suivante. Rajoutons une commande u à (3.10) et supposons, à commande u fixée, que les hypothèses du théorème de

Tikhonov soient valables. Ainsi

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & \varepsilon f(x,y,u,\varepsilon) \\ \dfrac{dy}{dt} & = & g(x,y,u,\varepsilon) \end{array} \right. \tag{3.16}$$

avec $h(x,u)$ le point d'équilibre hyperbolique et stable de la partie rapide $\dfrac{d\xi}{dt} = g(x,\xi,u,0)$. Le système lent est alors $\dfrac{dx}{d\tau} = f(x,h(x,u),u,0)$ dans l'échelle de temps lente $\tau = \varepsilon t$.

Supposons que nous ayons un retour d'état lent $u = k(x)$ (en utilisant par exemple de linéaire tangent) tel que le système lent bouclé soit asymptotiquement stable autour du point d'équilibre $(\bar{x},\bar{u} = k(\bar{x}))$ hyperbolique. Alors pour tout $\varepsilon > 0$ assez petit, le système perturbé (3.16) avec le bouclage lent $u = k(x)$, admet un point d'équilibre hyperbolique proche de $(\bar{x},\bar{y} = h(\bar{x},\bar{u}))$. Cela veut simplement dire que l'on peut, pour la synthèse d'un bouclage, ignorer des dynamiques hyperboliquement stables et assez rapides. On parle alors de *robustesse par rapport aux dynamiques négligées*. Noter que le retour d'état ne porte que sur la partie lente, x. Un bouclage sur y risquerait de déstabiliser la partie rapide.

3.5.2 Moyennisation

On suppose ici que les effets rapides ont un caractère oscillant. La méthode de moyennisation a été utilisée en mécanique céleste depuis longtemps pour déterminer l'évolution des orbites planétaires sous l'influence des perturbations mutuelles entre les planètes et étudier la stabilité du système solaire. Gauss en donne la définition suivante qui est des plus intuitives : il convient de répartir la masse de chaque planète le long de son orbite proportionnellement au temps passé dans chaque partie de l'orbite et de remplacer l'attraction des planètes par celle des anneaux de matière ainsi définis.

Dans ce cadre, les équations non perturbées du mouvement de la terre sont celles qui ne prennent en compte que la force d'attraction due au soleil. L'orbite de la terre est alors une ellipse dont le soleil est l'un des foyers. Les équations perturbées sont celles où l'on rajoute les forces d'attraction entre la terre et les autres planètes en supposant que ces dernières décrivent toutes des orbites elliptiques selon les lois de Kepler. Le paramètre ε correspond au rapport de la masse du soleil à celles des planètes : $\varepsilon \approx 1/1000$. L'échelle de temps rapide est de l'ordre d'une période de révolution, quelques années. L'échelle de temps lente est de l'ordre de quelques millénaires. La question est alors de savoir si ces petites perturbations d'ordre ε peuvent entraîner à terme, i.e. à l'échelle du millénaire, une dérive systématique des longueurs du grand axe et du petit axe de la trajectoire de la terre, ce qui aurait des conséquences catastrophiques pour le climat. En fait, les calculs (moyennisation) montrent qu'il n'en est rien. En revanche, l'excentricité des orbites oscille lentement. Ces oscillations influencent le climat.

Revenons au système (3.10). Le régime oscillatoire le plus simple pour y est le régime périodique, de période T :

$$y = \left(\begin{array}{c} y_1 \\ y_2 \end{array} \right), \quad \left\{ \begin{array}{rcccl} \dfrac{dy_1}{dt} & = & y_2 & = & g_1(x,y,\varepsilon) \\ \dfrac{dy_2}{dt} & = & -\left[\dfrac{2\pi}{T}\right]^2 y_1 & = & g_2(x,y,\varepsilon), \end{array} \right.$$

Sans changer de notation on pose $f(x,y(t),\varepsilon) = f(x,t,\varepsilon)$: f est régulière en x et dépend de t de façon périodique (période T). Le système perturbé s'écrit alors

$$\frac{dx}{dt} = \varepsilon f(x,t,\varepsilon), \quad 0 \leq \varepsilon \ll 1. \tag{3.17}$$

Le système moyennisé (ou système lent) est alors

$$\frac{dz}{dt} = \varepsilon \frac{1}{T}\int_0^T f(z,t,0)\,dt \overset{\text{déf}}{=} \varepsilon \overline{f}(z). \tag{3.18}$$

Remplacer les trajectoires du système instationnaire (3.17) par celles du système stationnaire (3.18), revient alors à lisser les trajectoires de (3.17).

Le théorème suivant montre qu'à un point d'équilibre hyperbolique du système moyen correspond une petite orbite périodique du système perturbé (3.17) (démonstration dans [11]).

Théorème 9 (moyennisation à une fréquence) *Considérons le système perturbé (3.17) avec f régulière. Il existe un changement de variables, $x = z + \varepsilon w(z,t)$ avec w de période T en t, tel que (3.17) devienne*

$$\frac{dz}{dt} = \varepsilon \overline{f}(z) + \varepsilon^2 f_1(z,t,\varepsilon)$$

avec $\overline{f}$ définie par (3.18) et f_1 régulière de période T en t. De plus,

(i) *si $x(t)$ et $z(t)$ sont respectivement solutions de (3.17) et (3.18) avec comme conditions initiales x_0 et z_0 telles que $\|x_0 - z_0\| = O(\varepsilon)$, alors $\|x(t) - z(t)\| = O(\varepsilon)$ sur un intervalle de temps de l'ordre de $1/\varepsilon$.*

(ii) *Si $\overline{z}$ est un point fixe hyperbolique stable du système moyenné (3.18), alors il existe $\overline{\varepsilon} > 0$ tel que, pour tout $\varepsilon \in]0,\overline{\varepsilon}]$, le système perturbé (3.17) admet une unique orbite périodique $\gamma_\varepsilon(t)$, proche de $\overline{z}$ ($\gamma_\varepsilon(t) = \overline{z} + O(\varepsilon)$) et asymptotiquement stable (les trajectoires démarrant près de $\gamma_\varepsilon(t)$ ont tendance à s'enrouler autour de cette dernière). L'approximation, à $O(\varepsilon)$ près, des trajectoires du système perturbé (3.17) par celles du système moyenné (3.18) devient valable pour $t \in [0, +\infty[$.*

Il est instructif de voir comment est construit le changement de coordonnées $x = z + \varepsilon w(z,t)$ en enlevant à x des termes oscillants d'ordre ε (w de période T en t). On a, d'une part,

$$\frac{dx}{dt} = \frac{dz}{dt} + \varepsilon \frac{\partial w}{\partial z}(z,t)\frac{dz}{dt} + \varepsilon \frac{\partial w}{\partial t}(z,t)$$

et, d'autre part,

$$\frac{dx}{dt} = \varepsilon f(z + \varepsilon w(z,t),t,\varepsilon).$$

Ainsi

$$\begin{aligned} \frac{dz}{dt} &= \varepsilon \left(I + \varepsilon \frac{\partial w}{\partial z}(z,t)\right)^{-1} \left[f(z + \varepsilon w(z,t),t,\varepsilon) - \frac{\partial w}{\partial t}(z,t)\right] \\ &= \varepsilon \left[f(z,t,0) - \frac{\partial w}{\partial t}(z,t)\right] + O(\varepsilon^2). \end{aligned}$$

Comme la dépendance en t de w est T-périodique, il n'est pas possible d'annuler complètement le terme d'ordre 1 en ε car il n'y a aucune raison pour que la fonction définie par

$$\int_0^t f(z,s,0)\,ds$$

soit T-périodique en temps. En revanche, on peut éliminer la dépendance en temps du terme d'ordre 1 en ε. Il suffit de poser

$$w(z,t) = \int_0^t \big(f(z,s,0) - \overline{f}(z)\big)\ ds$$

(noter que w est bien de T-périodique) pour obtenir

$$\frac{dz}{dt} = \varepsilon \overline{f}(z) + O(\varepsilon^2).$$

Si cette approximation n'est pas suffisante, il faut prendre en compte les termes d'ordre 2 et éliminer leur dépendance en temps par un changement de variable du type $x = z + \varepsilon w_1(z,t) + \varepsilon^2 w_2(z,t)$ avec w_1 et w_2 T-périodique.

Terminons cette section par un exemple, l'équation du second ordre suivante :

$$\frac{d^2\theta}{dt^2} = -\theta + \varepsilon(1-\theta^2)\frac{d\theta}{dt}.$$

C'est l'équation d'un pendule pour lequel on a rajouté un petit frottement positif pour les grandes amplitudes ($\theta > 1$) et négatif pour les petites ($\theta < 1$). Mettons d'abord ce système sous la forme standard

$$\frac{dx}{dt} = \varepsilon f(x,t,\varepsilon).$$

Le terme oscillant vient du système non perturbé

$$\frac{d^2\theta}{dt^2} = -\theta$$

dont les orbites sont des cercles. Les phénomènes lents (échelle de temps $1/\varepsilon$) sont clairement relatifs aux rayons de ces cercles (i.e. les amplitudes des oscillations). C'est pourquoi il convient de passer en coordonnées polaires en posant $\theta = r\cos(\psi)$ et $\dot\theta = r\sin(\psi)$. Dans ces coordonnées, le système perturbé s'écrit :

$$\left\{\begin{array}{rcl} \dfrac{dr}{dt} &=& \varepsilon[1 - r^2\cos^2(\psi)]\sin^2(\psi) \\ \dfrac{d\psi}{dt} &=& -1 + \varepsilon\sin(\psi)\cos(\psi)[1 - r^2\cos^2(\psi)]. \end{array}\right.$$

ψ est quasiment égal, à une constante près, au temps $-t$. On peut écrire

$$\frac{dr}{d\psi} = \frac{dr}{dt}\frac{dt}{d\psi}.$$

Ainsi, on se ramène à la forme standard en prenant ψ comme variable de temps :

$$\frac{dr}{d\psi} = \varepsilon\frac{[1 - r^2\cos^2(\psi)]\sin^2(\psi)}{-1 + \varepsilon\sin(\psi)\cos(\psi)[1 - r^2\cos^2(\psi)]} = \varepsilon f(r,\psi,\varepsilon).$$

Le système moyennisé est alors

$$\frac{du}{d\psi} = -\frac{\varepsilon}{8}u(4 - u^2).$$

$\overline{u} = 2$ est un point d'équilibre hyperbolique attracteur pour $\psi \to -\infty$, i.e. $t \to +\infty$. Donc pour ε suffisamment petit, l'équation perturbée possède un cycle limite hyperbolique attracteur donc l'équation est approximativement $\theta^2 + \dot\theta^2 = 4 + O(\varepsilon)$.

L'inconvénient principal de la théorie des perturbations est qu'il faut, dès le départ, avoir une idée assez précise de ce que l'on cherche : il convient de trouver un petit paramètre ε et d'isoler la partie rapide du système. A ce niveau l'intuition physique joue un rôle important.

3.6 Problèmes

Problème 1 (des lapins et des renards) *Un modèle dit prédateur-proie, initialement introduit et étudié par le mathématicien italien Vito Volterra est le suivant. On considère deux espèces, l'espèce y, les renards, dévorant l'autre espèce x, les lapins. On se pose alors la question suivante. Quelles peuvent être les évolutions temporelles possibles du nombre de lapins $x(t)$ et du nombre de renards $y(t)$ si l'on émet les hypothèses grossières suivantes :*

H1 *lorsque les lapins sont peu nombreux et en l'absence de renards, ils ont suffisamment d'herbe (et de serpolet) à manger pour avoir un taux de reproduction spécifique constant ;*

H2 *toujours en l'absence de renards, si les lapins deviennent trop nombreux, ils ont des difficultés d'approvisionnement en herbe fraîche, ce qui fait chuter leur taux spécifique de reproduction ;*

H3 *un renard dévore d'autant plus de lapins qu'ils sont nombreux et faciles à rencontrer ;*

H4 *sans lapin, les renards sont obligés de faire un régime d'autant plus sévère et ravageur qu'ils sont plus nombreux ;*

H5 *plus il y a de lapins, plus les renards deviennent nombreux.*

1. *Montrer qu'un modèle simple (bilan sur les lapins et les renards) formalisant et quantifiant les 5 hypothèses précédentes est le suivant :*

$$\begin{array}{rcl} \dfrac{dx}{dt} & = & (a - by - \alpha x)x \\ \dfrac{dy}{dt} & = & (cx - d - \beta y)y \end{array} \tag{3.19}$$

   où a, b, c , d, α et β sont des paramètres positifs.

2. *Quel est l'espace d'état du système? Montrer que les solutions sont définies sur $[0, +\infty[$ (indication : considérer les deux isoclines, les états où le champ de vecteurs est parallèle à l'axe des x ou à l'axe des y).*

3. *Discuter, en fonction des valeurs des paramètres le nombre et la stabilité des points d'équilibre.*

Problème 2 (systèmes lents/rapides) *Pour certains systèmes lents/rapides de la forme (3.10), comme celui considéré dans cet exercice, l'approximation du théorème 7 est insuffisante, bien que le système comporte, de manière évidente physiquement, deux échelles de temps très distinctes.*

Considérons le système différentiel

$$\left\{ \begin{array}{rcl} \dfrac{dx_1}{dt} & = & -\varepsilon x_1 x_2 \\ \dfrac{dx_2}{dt} & = & \varepsilon x_1 x_2 - x_2 + \varepsilon x_3 \\ \dfrac{dx_3}{dt} & = & x_2 - \varepsilon x_3 \end{array} \right. \tag{3.20}$$

correspondant à un réacteur parfaitement agité fermé où les réactions chimiques suivantes apparaissent :

$$X_1 \overset{\varepsilon x_1 x_2}{\hookrightarrow} X_2 \quad et \quad X_2 \underset{\varepsilon x_3}{\overset{x_2}{\rightleftarrows}} X_3$$

(x_i est la concentration de l'espèce chimique X_i, $i = 1,2,3$).

1. *Peut-on appliquer le théorème de Tikhonov ?*
2. *En introduisant un changement de variable utilisant $\sigma = x_1 + x_2 + x_3$, montrer que l'on se ramène à la forme standard du théorème de Tikhonov.*
3. *Montrer que les conditions du théorème de Tikhonov sont alors remplies. En déduire le système lent.*
4. *Cette approximation est-elle valable pour tous les temps $t > 0$?*
5. *Comment faire pour obtenir une meilleure approximation (question difficile) ?*

Problème 3 (Colonne à distiller) *La dynamique d'une colonne à distiller séparant un mélange de deux composants (propane/butane par exemple) peut être représentée par le système suivant (c.f. figure 3.30)*

$$\begin{cases} H_1\dot{x}_1 = Vk(x_2) - Vx_1 \\ H_j\dot{x}_j = Lx_{j-1} + Vk(x_{j+1}) \\ \qquad\qquad -Lx_j - Vk(x_j),\ j = 2,\ldots,j_f - 1 \\ H_{j_f}\dot{x}_{j_f} = Lx_{j_f-1} + Vk(x_{j_f+1}) \\ \qquad\qquad -(L+F)x_{j_f} - Vk(x_{j_f}) + Fz_f \\ H_j\dot{x}_j = (L+F)x_{j-1} + Vk(x_{j+1}) \\ \qquad -(L+F)x_j - Vk(x_j),\ j = j_f + 1,\ldots,n-1 \\ H_n\dot{x}_n = (L+F)x_{n-1} - (L+F-V)x_n - Vk(x_n) \end{cases} \tag{3.21}$$

avec $x_i \in [0,1]$ la composition du liquide au plateau i ($i = 1,\ldots,n$); L et V sont des débits de réglage ($0 < L < V < L + F$); $F > 0$ et $z_F \in [0,1]$ sont le débit et la composition de l'alimentation; k est une bijection de $[0,1]$ sur $[0,1]$; $k(x)$ est la composition de la vapeur en fonction de celle du liquide. Le paramètre $H_i > 0$ (constant) correspond à la rétention totale du plateau i.

1. *Montrer que, pour toutes conditions initiales sur les x_i entre dans $[0,1]$, les composantes x_i de la solution $x(t)$ restent dans $[0,1]$ pour tout $t > 0$.*
2. *Montrer que*

$$\sum_{i=1}^{n} |H_i\dot{x}_i|$$

   est une fonction de Lyapounov (L,V,F,z_F) sont supposés constants). En déduire que les trajectoires convergent vers un point d'équilibre
3. *(facultatif) Montrer l'unicité du point d'équilibre.*

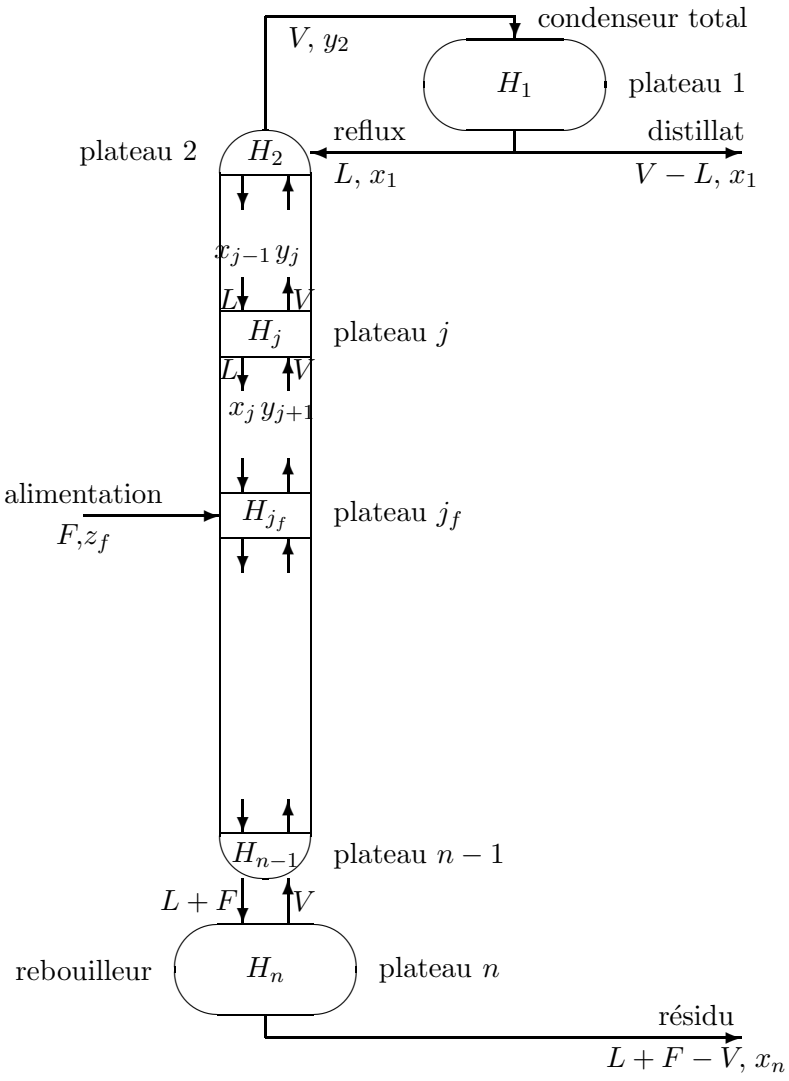


FIG. 3.30 – *colonne à distiller binaire.*

Chapitre 4

Commandabilité et observabilité

Un système commandé sous forme explicite, $\dot{x} = f(x,u)$, est un système sous-déterminé. La différence entre le nombre d'équations (indépendantes) et le nombre de variables donne le nombre de commandes indépendantes $m = \dim u$. Noter que le degré de sous-détermination est infini car il s'agit de m fonctions arbitraires du temps. Aussi l'étude des systèmes sous-déterminés d'équations différentielles ordinaires est d'une nature très différente de celle des systèmes déterminés, systèmes étudiés dans le chapitre précédent.

La première partie de ce chapitre aborde la commandabilité. Après de courtes définitions, nous étudions en détail les systèmes linéaires explicites $\dot{x} = Ax + Bu$. Leur commandabilité est caractérisée par le critère de Kalman et la forme normale dite de Brunovsky. Cette dernière permet un paramétrage explicite de toutes les trajectoires en fonctions de m fonctions scalaires arbitraires $t \mapsto y(t)$ et d'un nombre fini de leurs dérivées. Ces quantités y, dites sorties de Brunovsky, sont des combinaisons linéaires de x. Elles jouent d'une certaine façon le rôle d'un potentiel[1]. Elles permettent surtout de calculer très simplement les commandes u pour aller d'un état vers un autre (planification de trajectoire). Elles permettent également de construire le bouclage ("feedback") qui assure le suivi asymptotique d'une trajectoire de référence arbitraire (stabilisation par placement de pôles).

Le calcul de tels bouclages nécessite la connaissance à chaque instant de l'état x. Il est fréquent que seule une partie de l'état soit directement accessible à la mesure. Aussi est on confronté au problème suivant. Connaissant les équations du système (i.e., ayant un modèle), $\dot{x} = f(x,u)$, les relations entre les mesures y et l'état, $y = h(x)$, les entrées $t \mapsto u(t)$ et les mesures $t \mapsto y(t)$, estimer x. Cela revient à résoudre le problème suivant

$$\dot{x} = f(x,u), \quad y = h(x)$$

où x est l'inconnue (une fonction du temps) et où u et y sont des fonctions connues du temps. Il est clair que ce problème est sur-déterminé. L'unicité de la solution correspond à l'observabilité. L'existence au fait que y et u ne peuvent pas être des fonctions du temps indépendantes l'une de l'autre. Elles doivent vérifier des relations de compatibilité qui prennent la forme d'équations différentielles.

1. Il est classique en physique de paramétrer toutes les solutions du système sous-déterminé div $\overrightarrow{B} = 0$, par un potentiel vecteur arbitraire $\overrightarrow{A}$ avec la formule $\overrightarrow{B} = \text{rot } \overrightarrow{A}$. Le potentiel vecteur $\overrightarrow{A}$ est alors défini à partir du champs magnétique $\overrightarrow{B}$ à un champ de gradient près.

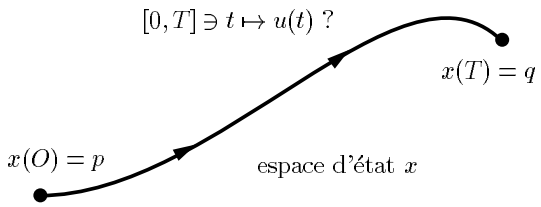


FIG. 4.1 – *la planification de trajectoire.*

La seconde partie de ce chapitre aborde cette question. Tout d'abord nous donnons les définitions et les critères assurant l'existence et l'unicité de la solution. Pour les systèmes linéaires nous présentons une méthode très économe en calculs pour obtenir x avec un observateur asymptotique.

En résumé, l'essentielle de ce chapitre porte sur les systèmes linéaires invariants en temps. Pour les systèmes non linéaires une référence classique est [15]. On trouvera aussi dans [2] des résultats sur la commandabilité non linéaire.

4.1 Commandabilité non linéaire

On considère le système explicite (f fonction régulière)

$$\frac{dx}{dt} = f(x,u), \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \tag{4.1}$$

4.1.1 Définition

Définition 14 (trajectoire) *On appelle trajectoire du système (4.1) toute fonction régulière $I \ni t \mapsto (x(t),u(t)) \in \mathbb{R}^n \times \mathbb{R}^m$ qui satisfait identiquement sur un intervalle d'intérieur non vide I de $\mathbb{R}$ les équations (4.1).*

Définition 15 (commandabilité) *Le système (4.1) est dit commandable en temps $T > 0$, si et seulement si, pour $p,q \in \mathbb{R}^n$, il existe une loi horaire $[0,T] \ni t \mapsto u(t) \in \mathbb{R}^m$, dite commande en boucle ouverte, qui amène le système de l'état $x(0) = p$ à l'état $x(T) = q$, c'est à dire, telle que la solution du problème de Cauchy*

$$\begin{aligned} \dot{x} &= f(x,u(t)) \quad \textit{pour } t \in [0,T] \\ x(0) &= p \end{aligned}$$

vérifie $x(T) = q$. Le système est dit simplement commandable lorsqu'il est commandable pour au moins un temps $T > 0$.

D'autres définitions sont possibles : elles correspondent toutes à des variantes plus ou moins subtiles de la définition 15. Comme l'illustre la figure 4.1, la commandabilité est une propriété topologique très naturelle. En général, la *commande en boucle ouverte* $[0,T] \ni t \mapsto u(t)$ n'est pas unique, il en existe une infinité. Cette étape s'appelle *planification de trajectoire* : calculer $t \mapsto u(t)$ à partir de la connaissance de f, p et q constitue

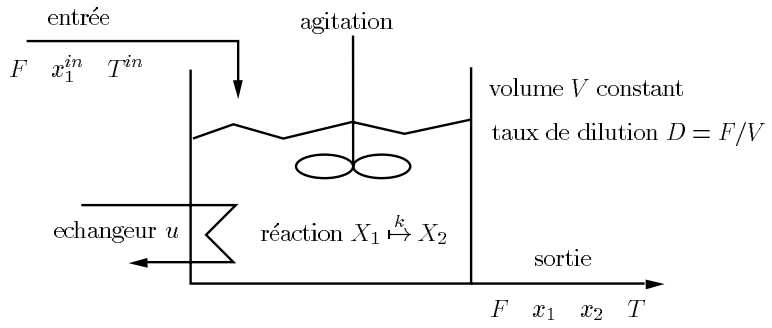


FIG. 4.2 – *un réacteur chimique exothermique où u correspond aux échanges thermiques avec l'extérieur.*

l'une des questions majeures de l'automatique. Cette question qui est loin d'être résolue actuellement.

Exercice 17 (commandabilité des systèmes discrets) *Donner une définition de la commandabilité pour le système discret défini par la récurrence en k suivante :*

$$x_{k+1} = f(x_k,u_k) \quad x_k \in \mathbb{R}^n, \quad u_k \in \mathbb{R}^m.$$

Très souvent, l'absence de commandabilité est due à l'existence d'intégrales premières non triviales. Ce sont des observables qui restent constantes le long de toute trajectoire.

4.1.2 Intégrale première

Considérons le réacteur exothermique de la figure 4.2. Les équations de bilan matière et énergie donnent alors les équations différentielles suivantes :

$$\begin{aligned} \dot{x}_1 &= D(x_1^{in} - x_1) - k_0 \exp(-E/RT)x_1 \\ \dot{x}_2 &= -Dx_2 + k_0 \exp(-E/RT)x_1 \\ \dot{T} &= D(T^{in} - T) + \alpha\Delta H \exp(-E/RT)x_1 + u. \end{aligned} \tag{4.2}$$

La cinétique est linéaire du premier ordre, les constantes physiques usuelles (D, x_1^{in}, k_0, E, T^{in}, α et ΔH) sont toutes positives, la commande u est proportionnelle à la puissance thermique échangée avec l'extérieur. x_i est la concentration de l'espèce chimique X_i, $i =$ 1,2. On reconnaît l'effet non linéaire essentiel de la loi d'Arrhenius $k = k_0 \exp(-E/RT)$ qui relie la constante de vitesse k à la température T. Il est assez facile de voir que ce système n'est pas commandable. En effet, le bilan global sur $X_1 + X_2$, élimine le terme non linéaire pour donner

$$\frac{d}{dt}(x_1 + x_2) = D(x_1^{in} - x_1 - x_2).$$

Ainsi donc la quantité $\xi = x_1 + x_2$ vérifie une équation différentielle autonome $\dot{\xi} = D(x_1^{in} - \xi)$. Donc $\xi = x_1^{in} + \xi_0 \exp(-Dt)$ où ξ_0 est la valeur initiale de ξ. Si, dans la définition 15, on prend l'état initial p tel que $\xi = x_1+x_2 = x_1^{in}$ et q tel que $\xi = x_1+x_2 = 0$, il n'existe pas de commande qui amène le système de p vers q. En effet, pour toute trajectoire démarrant en un tel p, la quantité $x_1 + x_2$ reste constante et égale à x_1^{in}. Cette partie non commandable du système représentée par la variable ξ admet ici un sens physique précis. Elle est bien connue des chimistes. C'est un invariant chimique.

L'exemple ci-dessus nous indique que l'absence de commandabilité peut-être liée à l'existence d'invariants, i.e., à des combinaisons des variables du système (on pourrait les appeler des observables) et éventuellement du temps, qui sont conservées le long de toute trajectoire. Pour (4.2), il s'agit de $(x_1+x_2-x_1^{in}) \exp(Dt)$ correspondant à ξ_0. Nous sommes donc conduits à prolonger la notion d'intégrale première pour les systèmes commandés.

Définition 16 (intégrale première) *Une fonction régulière $\mathbb{R} \times \mathbb{R}^n \ni (t,x) \mapsto h(t,x) \in \mathbb{R}$ est appelée intégrale première du système (4.1), si elle est constante le long de toute trajectoire du système. Une intégrale première est dite triviale si c'est une fonction constante sur $\mathbb{R} \times \mathbb{R}^n$.*

Si h est une intégrale première, sa dérivée le long d'une trajectoire arbitraire est nulle :

$$\frac{d}{dt}h = \frac{\partial h}{\partial t} + \frac{\partial h}{\partial x}\dot{x} \equiv 0$$

pour toute trajectoire $(t \mapsto (x(t),u(t))$ du système.

Exercice 18 (intégrale première en discret) *Donner une définition de la notion d'intégrale première pour le système dynamique discret défini par la récurrence en k : $x_{k+1} = f(x_k,u_k), \quad x_k \in \mathbb{R}^n, \quad u_k \in \mathbb{R}^m$.*

Si (4.1) admet une intégrale première non triviale $t \mapsto h(t,x)$ alors, (4.1) n'est pas commandable. Sinon, il existe $T > 0$, tel que pour tout $p,q \in \mathbb{R}^n$ et tout instant initial t $h(t,p) = h(t + T,q)$ (il existe une trajectoire reliant p à q sur $[t,t + T]$). Donc h est une fonction périodique du temps et indépendante de x. Mais alors la dérivée de h le long des trajectoires du système correspond à $\frac{\partial}{\partial t}h$. Comme elle est nulle, h est une constante, ce qui contredit l'hypothèse. Nous avons montré la proposition suivante

Proposition 5 *Si le système (4.1) est commandable, alors ses intégrales premières sont triviales.*

Il est possible de caractériser en termes finis (i.e., à partir de f et d'un nombre fini de ses dérivées partielles) l'existence d'intégrale première non triviale. Nous allons nous restreindre dans ce cours au cas linéaire. En effet, la démarche est la même pour le cas non linéaire. Elle conduit à des calculs plus lourds qui, pour être présentés de façon compacte, nécessitent le langage de la géométrie différentielle et les crochets de Lie.

4.2 Commandabilité linéaire

Nous considérons ici les systèmes linéaires stationnaires du type

$$\dot{x} = Ax + Bu \tag{4.3}$$

où l'état $x \in \mathbb{R}^n$, la commande (on dit aussi l'entrée) $u \in \mathbb{R}^m$ et les matrices A et B sont constantes et de tailles $n \times n$ et $n \times m$, respectivement.

4.2.1 Matrice de commandabilité

Supposons que (4.3) admette une intégrale première $h : \mathbb{R} \times \mathbb{R}^n \ni (t,x) \mapsto h(t,x) \in \mathbb{R}$. Soit le changement de variables sur x définit par $x = \exp(tA)z$. Avec les variables (z,u), (4.3) devient $\dot{z} = \exp(-tA)Bu$ et l'intégrale première devient $h(t, \exp(tA)z) = l(t,z)$. Comme la valeur de l est constante le long de toute trajectoire nous avons, en dérivant le long d'une trajectoire arbitraire $t \mapsto (z(t),u(t))$

$$\dot{l} = \frac{\partial l}{\partial t} + \frac{\partial l}{\partial z}\dot{z} = 0.$$

Comme $\dot{z} = \exp(-tA)Bu$, pour toute valeur de z et u on a l'identité suivante :

$$\frac{\partial l}{\partial t}(t,z) + \frac{\partial l}{\partial z}(t,z)\exp(-tA)Bu \equiv 0.$$

En prenant, $u = 0$, z et t arbitraires, on en déduit (prendre, e.g., la trajectoire du système qui passe par z à l'instant t et dont la commande u est nulle) :

$$\frac{\partial l}{\partial t}(t,z) \equiv 0.$$

Donc nécessairement l est uniquement fonction de z. Ainsi

$$\frac{\partial l}{\partial z}(z)\exp(-tA)B \equiv 0.$$

En dérivant cette relation par rapport à t, on a,

$$\frac{\partial l}{\partial z}(z)\exp(-tA)AB \equiv 0$$

car $\frac{d}{dt}(exp(-tA)) = -\exp(-tA)A$. Plus généralement, une dérivation à n'importe quel ordre $k \geq 0$ donne

$$\frac{\partial l}{\partial z}(z)\exp(-tA)A^kB \equiv 0.$$

En prenant $t = 0$ on obtient

$$\frac{\partial l}{\partial z}(z)A^kB = 0, \quad \forall k \geq 0.$$

Ainsi le vecteur $\frac{\partial l}{\partial z}(z)$ appartient à l'intersection des noyaux à gauche de la famille infinie de matrice $(A^kB)_{k\geq 0}$. Le noyau à gauche de A^kB n'est autre que $\mathrm{Im}(A^kB)^\perp$, l'orthogonal de l'image de A^kB. Donc

$$\frac{\partial l}{\partial z}(z) \in \bigcap_{k\geq 0} \mathrm{Im}(A^kB)^\perp.$$

Mais

$$\bigcap_{k\geq 0} \mathrm{Im}(A^kB)^\perp = \big(\mathrm{Im}(B) + \ldots + \mathrm{Im}(A^kB) + \ldots\big)^\perp.$$

La suite d'espace vectoriel $E_k = \mathrm{Im}(B) + \ldots + \mathrm{Im}(A^k B)$ est une suite croissante pour l'inclusion, $E_k \subset E_{k+1}$. Si pour un certain k, $E_k = E_{k+1}$, cela signifie que $\mathrm{Im}(A^{k+1}B) \subset E_k$, donc $A(E^k) \subset E_k$. Mais $\mathrm{Im}(A^{k+2}B) = \mathrm{Im}(AA^{k+1}B) \subset A(E^{k+1})$. Ainsi $\mathrm{Im}(A^{k+2}B) \subset E_k$. On voit donc que pour tout $r > 0$, $\mathrm{Im}(A^{k+r}B) \subset E_k$, d'où $E_{k+r} = E_k$. Ainsi la suite des E_k est une suite de sous-espaces vectoriels de $\mathbb{R}^n$ emboîtés les uns dans les autres. Cette suite stationne dès qu'elle n'est plus, pour un certain k, strictement croissante. Il suffit donc de ne considérer que ses n premiers termes soit $E_0, \ldots E_{n-1}$, car automatiquement $E_{n-1} = E_{n+r}$ pour tout $r > 0$.

En revenant à la suite des noyaux à gauche de $A^k B$, nous voyons que $\frac{\partial l}{\partial z}(z)$ dans le noyau à gauche de la suite infinie de matrices $(A^k B)_{k\geq 0}$, est équivalent à, $\frac{\partial l}{\partial z}(z)$ dans le noyau à gauche de la *suite finie* de matrices $(A^k B)_{0\leq k\leq n-1}$.[2]

Ainsi, pour tout z, $\frac{\partial l}{\partial z}(z)$ appartient au noyau à gauche de la matrice $n \times (nm)$,

$$\mathcal{C} = (B,AB,A^2B,\ldots,A^{n-1}B) \tag{4.4}$$

dite *matrice de commandabilité* de Kalman. Si $\mathcal{C}$ est de rang n, son noyau à gauche est nul, donc l ne dépend pas de z : l est alors une fonction constante et h également.

Réciproquement, si la matrice de commandabilité $\mathcal{C}$ n'est pas de rang maximal, alors il existe un vecteur $w \in \mathbb{R}^n/\{0\}$, dans le noyau à gauche de (4.4). En remontant les calculs avec $l(z,t) = w'z$ on voit que $\dot{\lambda} = 0$ le long des trajectoires. En passant aux variables (x,u), on obtient une intégrale première non triviale $= h(t,x) = w' \cdot \exp(-tA)x$. Toute trajectoire du système se situe dans un hyperplan orthogonal à w.

En résumé, nous avons démontré la

Proposition 6 *La matrice de commandabilité $\mathcal{C} = (B,AB,A^2B,\ldots,A^{n-1}B)$ est de rang n, si, et seulement si, les seules intégrales premières du système (4.3) sont triviales.*

Des propositions 5 et 6, il vient : si le système (4.3) est commandable, sa matrice de commandabilité est de rang n. Nous allons voir que la réciproque est vraie. Pour cela, nous avons besoin de certaines propriétés d'invariance.

4.2.2 Invariance

Définition 17 (changement d'état, bouclage statique régulier) *Un changement linéaire de coordonnées $x \mapsto \tilde{x}$ est défini par une matrice M inversible d'ordre n : $x = M\tilde{x}$. Un bouclage statique régulier $u \mapsto \tilde{u}$ est défini par une matrice N inversible d'ordre m et une autre matrice K, $m \times n$: $u = K\tilde{x} + N\tilde{u}$. C'est un changement de variables sur les commandes paramétré par l'état.*

2. On pourrait aussi utiliser le théorème de Cayley-Hamilton qui donne un résultat plus précis : toute matrice carrée est racine de son polynôme caractéristique. Cela veut dire, A étant de taille n, que A^n est une combinaison linéaire des $(A^k)_{0\leq k\leq n-1}$:

$$A^n = \sum_{k=0}^{n-1} p_k A^k$$

où les p_k sont définis par $\det(\lambda I_n - A) = \lambda^n - \sum_{k=0}^{n-1} p_k \lambda^k$. Nous avons préféré un argument plus simple avec la suite des E_k mais qui a l'avantage de passer au non linéaire et qui correspond au calcul de l'algèbre de Lie de commandabilité.

L'ensemble des transformations

$$\begin{pmatrix} \tilde{x} \\ \tilde{u} \end{pmatrix} \mapsto \begin{pmatrix} x \\ u \end{pmatrix} = \begin{pmatrix} M & 0 \\ K & N \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{u} \end{pmatrix} \tag{4.5}$$

forment un groupe lorsque les matrices M, N et K varient (M et N restant inversibles).

Exercice 19 *Si $\dot{x} = Ax + Bu$ est commandable (resp. n'admet pas d'intégrale première) montrer que $\dot{\widetilde{x}} = \widetilde{A}\widetilde{x}+\widetilde{B}\widetilde{u}$ obtenu avec (4.5) est commandable (resp. n'admet pas d'intégrale première).*

Les notions de commandabilité et d'intégrale première sont intrinsèques, c'est-à-dire, indépendantes des coordonnées avec lesquelles les équations du système sont établies. Si la matrice de commandabilité dans les coordonnées (x,u) est de rang n, la matrice de commandabilité dans les coordonnées $(\tilde{x},\tilde{u})$ sera aussi de rang n. Cette simple remarque conduit au résultat non évident suivant :

$$\text{rang}(B,AB,\dots A^{n-1}B) = n \quad \text{équivaut à} \quad \text{rang}(\tilde{B},\tilde{A}\tilde{B},\dots,\tilde{A}^{n-1}\tilde{B}) = n$$

où $\tilde{A}$ et $\tilde{B}$ s'obtiennent en écrivant $\dot{x} = Ax + Bu$ dans les coordonnées $(\tilde{x},\tilde{u})$:

$$\dot{\tilde{x}} = M^{-1}(AM + BK)\tilde{x} + M^{-1}BN\tilde{u}.$$

Soit $\tilde{A} = M^{-1}(AM + BK)$ et $\tilde{B} = M^{-1}BN$. En fait, il est possible d'aller beaucoup plus loin et de montrer que les indices de commandabilité définis ci-dessous sont aussi invariants.

Définition 18 (indices de commandabilité) *Pour tout entier k, on note σ_k le rang de la matrice $(B,AB,A^2B,\dots,A^kB)$. Les (σ_k) sont appelés indices de commandabilité du système linéaire (4.3),*

La suite σ_k est croissance, majorée par n. Ainsi, l'absence d'intégrale première est équivalente à $\sigma_{n-1} = n$.

Proposition 7 (invariance) *Les indices de commandabilité de $\dot{x} = Ax + Bu$ sont invariants par changement de variable sur x et bouclage statique régulier sur u.*

Nous laissons la preuve de ce résultat par récurrence sur n en exercice.

Il est important de comprendre la géométrie derrière cette invariance. Les transformations $(x,u) \mapsto (\tilde{x},\tilde{u})$ du type (4.5) forment un groupe. Ce groupe définit une relation d'équivalence entre deux systèmes ayant même nombre d'états et même nombre de commandes. La proposition précédente signifie simplement que les indices de commandabilité sont les même pour deux systèmes appartenant à la même classe d'équivalence, i.e, le même objet géométrique vu dans deux repères différents. En fait, on peut montrer que les indices de commandabilité sont les seuls invariants : il y a autant de classes d'équivalence que d'indices de commandabilité possibles. Nous ne montrerons pas en détail ce résultat. Tous les éléments nécessaires à cette preuve se trouvent dans la construction de la forme de Brunovsky ci-dessous (voir aussi [13, 17]).

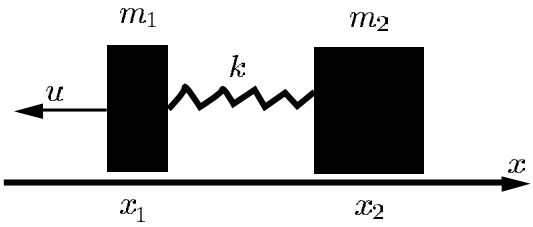


FIG. 4.3 – *deux masses couplées par un ressort, le tout piloté par une seule force u.*

4.2.3 Un exemple

Soit le système mécanique à deux degrés de liberté et une seule commande de la figure 4.3. Il s'agit d'un système mécanique sous actionné contrairement au bras motorisé étudié au chapitre 1 (un degré de liberté (l'angle θ) et un moteur). En négligeant les frottements et en supposant le ressort linéaire de raideur k, on est conduit au modèle suivant :

$$\begin{cases} m_1\ddot{x}_1 = k(x_2 - x_1) + u \\ m_2\ddot{x}_2 = k(x_1 - x_2). \end{cases} \tag{4.6}$$

Exercice 20 *Calculer la matrice de commandabilité de (4.6). Quel est son rang ?*

Montrons que ce système est commandable. Il suffit pour cela de remarquer que la quantité x_2, l'abscisse de la masse qui n'est pas directement soumise à la force u, joue un rôle très particulier (sortie de Brunovsky). Si au lieu de donner $t \mapsto u(t)$ et d'intégrer (4.6) à partir de positions et vitesses initiales, on fixe $t \mapsto x_2(t) = y(t)$. Alors $x_1 = \frac{m_2}{k}\ddot{y} + y$ et donc $u = m_1\ddot{x}_1 + m_2\ddot{x}_2 = \frac{m_1 m_2}{k} y^{(4)} + (m_1 + m_2)\ddot{y}$. Ainsi on peut écrire le système en faisant jouer à x_2 un rôle privilégié :

$$\begin{cases} x_1 = (m_2/k)\ \ddot{y} + y \\ x_2 = y \\ u = (m_1 m_2/k)\ y^{(4)} + m_1 + m_2)\ddot{y}. \end{cases}$$

On obtient ainsi un paramétrisation explicite de toutes les trajectoires du système. Les relations précédentes établissent une correspondance bi-univoque et régulière entre les trajectoires de (4.6) et les fonctions régulières $t \mapsto y(t)$.

Exercice 21 *Quel est l'index du système semi-implicite formé de (4.6) avec $x_2 = y(t)$ où $y(t)$ est une fonction connue du temps.*

Ainsi nous constatons que l'inverse du système (4.6) avec comme sortie $y = x_2$ est sans dynamique. Cela permet de calculer de la façon la plus élémentaire possible une commande $[0,T] \ni t \mapsto u(t)$ qui fait passer de l'état $p = (x_1^p, v_1^p, x_2^p, v_2^p)$ à l'état $q = (x_1^q, v_1^q, x_2^q, v_2^q)$ (v_i correspond à $\dot{x}_i$). Comme

$$\begin{cases} x_1 = (m_2/k)\ \ddot{y} + y \\ v_1 = (m_2/k)\ y^{(3)} + \dot{y} \\ x_2 = y \\ v_2 = \dot{y} \end{cases}$$

imposer p en $t = 0$ revient à imposer y et ses dérivées jusqu'à l'ordre 3 en 0. Il en est de même en $t = T$. Il suffit donc de trouver une fonction régulière $[0,T] \ni t \mapsto y(t)$ dont les dérivées jusqu'à l'ordre 3 sont données a priori en 0 et en T : un polynôme de degré 7 en temps répond à la question mais il existe bien d'autres possibilités.

Nous allons voir, avec la forme normale de Brunovsky, qu'une telle correspondance entre y et les trajectoires du système est générale. Il suffit que (4.3) soit commandable. Tout revient donc à trouver la sortie de Brunovsky y de même dimension que la commande u.

Exercice 22 *On veut transférer (4.6) de la configuration stationnaire $x_1 = x_2 = 0$ à la configuration stationnaire $x_1 = x_2 = D > 0$ durant le temps T. Calculer explicitement une commande $[0,T] \ni t \mapsto u(t)$ qui assure le transfert. On pourra supposer donnée une fonction C^4 $\phi : [0,1] \mapsto [0,1]$ telle que $\phi(0) = 0$, $\phi(1) = 1$ et $\frac{d^k\phi}{ds^k}(0) = \frac{d^k\phi}{ds^k}(1) = 0$ pour $k = 1,2,3$.*

4.2.4 Critère de Kalman et forme de Brunovsky

Théorème 10 (critère de Kalman) *Le système $\dot{x} = Ax + Bu$ est commandable si, et seulement si, la matrice de commandabilité $\mathcal{C} = (B,AB,\dots A^{n-1}B)$ est de rang $n = \dim(x)$.*

Pour abréger, on dit souvent que *la paire (A,B) est commandable*, pour dire que le rang de la matrice de commandabilité $\mathcal{C}$ est maximum.

La preuve que nous allons donner de ce résultat n'est pas la plus courte possible. Cependant, elle permet de décrire explicitement, pour toute durée $T > 0$ et pour $p,q \in \mathbb{R}^n$, les trajectoires du système qui partent de p et arrivent en q. Cette preuve utilise la forme dite de Brunovsky. Cette dernière se construit grâce à une méthode d'élimination, proche de celle très classique du pivot de Gauss. La même technique de calcul permet de traiter complètement la réalisation d'un transfert rationnel causal (c.f. problème 7).

Théorème 11 (forme de Brunovsky) *Si $(B,AB,\dots A^{n-1}B)$, la matrice de commandabilité de $\dot{x} = Ax + Bu$, est de rang $n = \dim(x)$ et si B est de rang $m = \dim(u)$, alors il existe un changement d'état $z = Mx$ (M matrice inversible $n \times n$) et un bouclage statique régulier $u = Kz + Nv$ (N matrice inversible $m \times m$), tels que les équations du système dans les variables (z,v) admettent la forme suivante (écriture sous la forme de m équations différentielles d'ordre ≥ 1) :*

$$y_1^{(\alpha_1)} = v_1, \quad \dots \quad , y_m^{(\alpha_m)} = v_m \tag{4.7}$$

avec comme état $z = (y_1,y_1^{(1)},\dots,y_1^{(\alpha_1-1)}, \quad \dots \quad ,y_m,y_m^{(1)},\dots,y_m^{(\alpha_m-1)})$, les α_i étant des entiers positifs.

Les m quantités y, qui sont des combinaisons linéaires de l'état x, sont appelées *sorties de Brunovsky.*

Exercice 23 (indices de commandabilité et forme de Brunovsky) *Relier, pour une paire (A,B) commandable, les indices de commandabilité σ_k, aux m entiers α_i de la forme de Brunovsky.*

Exercice 24 *On reprend les hypothèses du théorème 11. Les sorties de Brunovsky y sont des combinaisons linéaires de l'état x : y = Cx. Quel est l'index du système semi-implicite*

$$\dot{x} = Ax + Bu, \quad 0 = Cx.$$

Quelles sont ses solutions ?

Preuve du théorème 11. Elle repose sur

1. une mise sous forme triangulaire des équations d'état et l'élimination de u;
2. l'invariance du rang de $(B,AB,\dots A^{n-1}B)$ par rapport aux transformations (4.5);
3. une récurrence sur la dimension de l'état.

Mise sous forme triangulaire On suppose que B est de rang $m = \dim(u)$ (sinon, faire un regroupement des commandes en un nombre plus petit que m de façon à se ramener à ce cas). Alors, il existe une partition de l'état $x = (x_r,x_u)$ avec $\dim(x_r) = n - m$ et $\dim(x_u) = m$ telle que les équations (4.3) admettent la structure bloc suivante

$$\begin{array}{rcl} \dot{x}_r & = & A_{rr}x_r + A_{ru}x_u + B_r u \\ \dot{x}_u & = & A_{ur}x_r + A_{uu}x_u + B_u u \end{array}$$

où B_u est une matrice carrée inversible. Cette partition n'est pas unique, bien sûr. En tirant u de la seconde équation et en reportant dans la première, on obtient

$$\begin{array}{rcl} \dot{x}_r & = & A_{rr}x_r + A_{ru}x_u + B_r B_u^{-1}(\dot{x}_u - A_{ur}x_r - A_{uu}x_u) \\ \dot{x}_u & = & A_{ur}x_r + A_{uu}x_u + B_u u. \end{array}$$

En regroupant les dérivées dans la première équation, on a

$$\begin{array}{rcl} \dot{x}_r - B_r B_u^{-1}\dot{x}_u & = & (A_{rr} - B_r B_u^{-1}A_{ur})x_r + (A_{ru} - B_r B_u^{-1}A_{uu})x_u \\ \dot{x}_u & = & A_{ur}x_r + A_{uu}x_u + B_u u. \end{array}$$

Avec une transformation (4.5) définie par

$$\tilde{x}_r = x_r - B_r B_u^{-1}x_u, \quad \tilde{x}_u = x_u, \quad \tilde{u} = A_{ur}x_r + A_{uu}x_u + B_u u,$$

les équations $\dot{x} = Ax + Bu$ deviennent

$$\begin{array}{rcl} \dot{\tilde{x}}_r & = & \tilde{A}_r\tilde{x}_r + \tilde{A}_u\tilde{x}_u \\ \dot{\tilde{x}}_u & = & \tilde{u} \end{array}$$

où $\tilde{A}_r = (A_{rr} - B_r B_u^{-1}A_{ur})$ et $\tilde{A}_u = (A_{rr} - B_r B_u^{-1}A_{ur})B_r B_u^{-1} + (A_{ru} - B_r B_u^{-1}A_{uu})$. Dans cette structure triangulaire où la commande $\tilde{u}$ n'intervient pas dans la première équation, nous voyons apparaître un système plus petit d'état $\tilde{x}_r$ et de commande $\tilde{x}_u$. Cela nous permet de réduire la dimension de x et de raisonner par récurrence.

Invariance Un simple calcul par blocs nous montre que si $(B,AB,\dots A^{n-1}B)$ est de rang n alors $(A_u,A_rA_u,\dots A_r^{n-m-1}A_u)$ est de rang $n-m$. Du système de taille n on passe ainsi au système de taille réduite $n-m$, $\dot{\tilde{x}}_r = \tilde{A}_r\tilde{x}_r + \tilde{A}_u\tilde{x}_u$ ($\tilde{x}_r$ l'état, $\tilde{x}_u$ la commande).

Récurrence sur le nombre d'états Supposons donc, le résultat vrai pour toutes les dimensions d'état inférieures ou égales à $n-1$. Considérons un système $\dot{x} = Ax + Bu$ avec $n = \dim(x)$, sa matrice de commandabilité de rang n, et B de rang $m = dim(u) > 0$. L'élimination de u donne, après une transformation de type (4.5),

$$\begin{aligned} \dot{x_r} &= A_r x_r + A_u x_u \\ \dot{x_u} &= u \end{aligned}$$

où $\dim(u) = \dim(x_u) = m$ et $\dim(x_r) = n - m$ avec $(A_u, A_r A_u, \dots A_r^{n-m-1} A_u)$ de rang $n - m < n$ (les $\tilde{}$ ont été enlevés pour alléger les notations). Notons $\bar{m}$ le rang de A_u. Comme $\bar{m} \leq m$, un changement de variable sur x_u, $(\bar{x}_u, \tilde{x}_u) = Px_u$ avec P inversible, permet d'écrire le système sous la forme

$$\begin{aligned} \dot{x_r} &= A_r x_r + \bar{A}_u \bar{x}_u \\ \dot{\bar{x}}_u &= \bar{u} \\ \dot{\tilde{x}}_u &= \tilde{u} \end{aligned} \tag{4.8}$$

avec $(\bar{u}, \tilde{u}) = Pu$, $\dim(\bar{x}_u) = \bar{m}$ et $\bar{A}_u$ de rang $\bar{m}$. Comme le rang de la matrice de commandabilité de $\dot{x_r} = A_r x_r + \bar{A}_u \bar{x}_u$ (x_r est l'état et $\bar{x}_u$ la commande) est égal à $n - m = dim(x_r)$, l'hypothèse de récurrence assure l'existence d'un changement de variable $x_r = Mz$ et d'un bouclage statique régulier $\bar{x}_u = Kz + N\bar{v}$ ($\bar{v}$ est la nouvelle commande ici) mettant ce sous système sous forme de Brunovsky. Alors le changement d'état $(x_r, \bar{x}_u, \tilde{x}_u)$ défini par

$$\begin{pmatrix} x_r \\ \bar{x}_u \\ \tilde{x}_u \end{pmatrix} = \begin{pmatrix} M & 0 & 0 \\ K & N & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} z \\ \bar{v} \\ \tilde{x}_u \end{pmatrix}$$

et le bouclage statique régulier sur $(\bar{u}, \tilde{u})$

$$\bar{u} = KM^{-1}(A_r x_r + \bar{A}_u \bar{x}_u) + N\bar{v}, \quad \tilde{u} = \tilde{v}$$

transforme alors le système (4.8) sous forme de Brunovsky avec $v = (\bar{v}, \tilde{v})$ comme nouvelle commande. ∎

Preuve du théorème 10 La commandabilité est indépendante du choix des variables sur x et d'un bouclage statique régulier sur u. On peut donc supposer le système sous sa forme de Brunovsky. Dans ces coordonnées, aller d'un état à un autre est élémentaire. Il se ramène à étudier la commandabilité du système scalaire $y^{(\alpha)} = v$. L'état initial $(y_a, \dots, y_a^{(\alpha-1)})$ et l'état final $(y_b, \dots, y_b^{(\alpha-1)})$ ainsi que la durée T étant donnés, les lois horaires $t \mapsto v(t)$ assurant le passage entre ces deux états pendant la durée T correspondent alors à la dérivée α-ième de fonctions $[0,T] \ni t \mapsto \varphi(t) \in \mathbb{R}$, dont les dérivées jusqu'à l'ordre $\alpha - 1$ en 0 et T sont imposées par

$$\varphi^{(r)}(0) = y_a^{(r)}, \quad \varphi^{(r)}(T) = y_b^{(r)}, \quad r = 0, \dots, \alpha - 1.$$

Il existe bien sûr une infinité de telles fonctions φ (on peut prendre pour φ un polynôme de degré $2\alpha - 1$, par exemple). ∎

Exercice 25 (commandabilité les systèmes linéaires discrets) *Montrer que le système discret*

$$x_{k+1} = Ax_k + Bu_k, \quad x_k \in \mathbb{R}^n, \quad u_k \in \mathbb{R}^m$$

est commandable si, et seulement si, le rang de $(B,AB,A^2B,\ldots,A^{n-1}B)$ *vaut* n*. Quel est alors l'équivalent de la forme de Brunovsky.*

4.2.5 Planification et suivi de trajectoires

De la preuve des deux théorèmes précédents, il est important de retenir deux choses :

– Dire que le système $\dot{x} = Ax + Bu$ est commandable, est équivalent à l'existence d'un bouclage statique régulier $u = Kz + Nv$ et d'un changement d'état $x = Mz$ se ramenant à la forme de Brunovsky $y^{(\alpha)} = v$ et $z = (y,\ldots,y^{(\alpha-1)})$ (par abus de notation $y = (y_1,\ldots,y_m)$ et $y^{(\alpha)} = (y_1^{(\alpha_1)},\ldots,y_m^{(\alpha_m)})$). Ainsi

$$x = M(y,\ldots,y^{(\alpha-1)}), \quad u = L(y,\ldots,y^{(\alpha)})$$

où la matrice L est construite avec K, N et M. Lorsque l'on considère une fonction régulière arbitraire du temps $t \mapsto \varphi(t) \in \mathbb{R}^m$ et que l'on calcule $x(t)$ et $u(t)$ par les relations

$$x(t) = M(\varphi(t),\ldots,\varphi^{(\alpha-1)}(t)), \quad u(t) = L(\varphi(t),\ldots,\varphi^{(\alpha)}(t))$$

alors $t \mapsto (x(t),u(t))$ est une trajectoire du système : on a identiquement $\dot{x}(t) - Ax(t) - Bu(t) = 0$. Réciproquement, toutes les trajectoires régulières du système se paramétrisent de cette façon, grâce à m fonctions scalaires arbitraires $\varphi_1(t)$, …, $\varphi_m(t)$ et un nombre fini de leurs dérivées par les formules ci-dessus.

– La commandabilité de $\dot{x} = Ax + Bu$ implique la stabilisation par retour d'état. En effet, il suffit de considérer la forme de Brunovsky et dans la forme de Brunovsky, chacun des m sous-systèmes indépendants $y_i^{(\alpha_i)} = v_i$. Soient α_i valeurs propres, $\lambda_1,\ldots,\lambda_{\alpha_i}$, correspondant au spectre d'une matrice réelle de dimension α_i. Notons s_k les fonctions symétriques des λ_i (des quantités réelles donc) homogènes de degré k,

$$\prod_{k=1}^{\alpha_i}(X - \lambda_k) = X^{\alpha_i} - s_1X^{\alpha_i-1} + s_2X^{\alpha_i-2} + \ldots + (-1)^{\alpha_i}s_{\alpha_i}$$

Alors, dès que les λ_k sont à partie réelle strictement négative, le bouclage

$$v_i = s_1y_i^{(\alpha_i-1)} - s_2y_i^{(\alpha_i-2)} + \ldots + (-1)^{\alpha_i-1}s_{\alpha_i}y_i$$

assure la stabilité de $y_i^{(\alpha_i)} = v_i$: en effet, les exposants caractéristiques (on dit aussi les *pôles*) du système bouclé sont les λ_k.

Aussi de la forme de Brunovsky l'on déduit directement le résultat suivant :

Théorème 12 (placement de pôles) *Si la paire* (A,B) *est commandable alors, pour toute matrice réelle* F $n \times n$*, il existe une matrice* $m \times n$*,* K *(non nécessairement unique), telle que le spectre de* $A + BK$ *coïncide avec celui de* F*.*

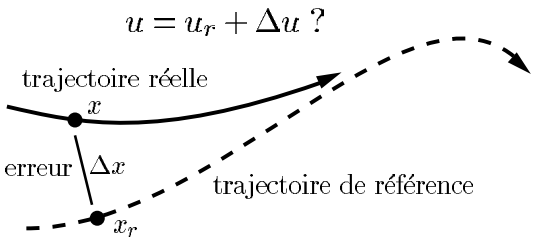


FIG. 4.4 – *le suivi de trajectoire.*

De retour dans les coordonnées de modélisation, $\dot{x} = Ax + Bu$, la planification de trajectoire nous donne une trajectoire du système (par exemple la trajectoire que doit suivre une fusée au décollage, la manoeuvre d'atterrissage d'un avion, ...). Nous la notons $t \mapsto (x_r(t),u_r(t))$ avec l'indice r pour référence. En pratique, et à cause des aléas de l'existence, il convient, comme l'illustre la figure 4.4, de corriger en fonction de l'écart Δx, la commande de référence u_r (il est rare de piloter un système en aveugle, uniquement en sachant d'où l'on part et où l'on veut aller). Le problème est donc de calculer la correction Δu à partir de Δx de façon à revenir sur la trajectoire de référence. On peut alors utiliser un bouclage stabilisant en plaçant les pôles sur la forme de Brunovsky.

D'une façon plus précise : comme $\dot{x}_r = Ax_r + Bu_r$, on obtient, par différence avec $\dot{x} = Ax + Bu$ l'équation d'erreur suivante

$$\frac{d(\Delta x)}{dt} = A\ \Delta x + B\ \Delta u$$

où $\Delta x = x - x_r$ et $\Delta u = u - u_r$; le système étant commandable, il existe K, matrice $m \times n$, telle que les valeurs propres de $A+BK$ soient à parties réelles strictement négatives (placement de pôles). Ainsi la correction

$$\Delta u = K\ \Delta x$$

assure le suivi asymptotique de la trajectoire de référence $t \mapsto x_r(t)$. La stabilité structurelle des points d'équilibres hyperboliques garantie que toute erreur assez faible (petite incertitude sur A et B, effets non linéaires faibles, erreurs de mesure, erreurs de troncature dues à la discrétisation de la loi de contrôle obtenue, ...) ne sera pas amplifiée au cours du temps : x restera ainsi proche de x_r.

Nous terminerons par une constatation d'ordre expérimental : lorsque le modèle dynamique $\dot{x} = Ax + Bu$ est d'origine physique, il n'est pas rare que sa partie non commandable, i.e., ses intégrales premières, ait une signification physique immédiate, tout comme les grandeurs y, fonction de x et intervenant dans la forme de Brunovsky (c.f. théorème 11) de sa partie commandable. Cet état de fait n'est vraisemblablement pas dû entièrement au hasard : en physique, les grandeurs qui admettent une signification intrinsèque, i.e., les grandeurs physiques, sont celles qui ne dépendent pas du repère de l'observateur. En automatique, le passage d'un repère à un autre correspond, entre autre, à une transformation de type (4.5). Il est alors clair que le "sous-espace" engendré par les sorties de Brunovsky est un invariant. Il a donc toutes les chances d'avoir un sens physique immédiat. De plus les sorties de Brunovsky admettent un équivalent non linéaire pour de nombreux systèmes physiques. On les appelle alors sorties plates (cf. exercices 26 et 27).

Exercice 26 *Soit le système de la figure 4.3. On suppose que le ressort est non linéaire. Dans (4.6) la raideur k est fonction de $x_1 - x_2$: $k = k_0 + a(x_1 - x_2)^2$ avec k_0 et $a > 0$. Montrer que le système reste commandable et calculer sa sortie "non linéaire" de Brunovsky (la sortie plate).*

Exercice 27 *Prenons l'exemple (4.2) en ne considérant que les deux équations différentielles relatives à x_1 et T (nous ne considérons que la partie commandable). Montrer (formellement) que ce sous-système à deux états et une commande est commandable (indication : la quantité $y = x_1$ joue le rôle de sortie "non linéaire" de Brunovsky (la sortie plate)) Calculer le bouclage statique qui linéarise le système.*

Exercice 28 *Pour le système (4.6) calculer explicitement un bouclage d'état qui place les pôles. Connaissant les paramètres m_1, m_2, et k que choisir comme pôles pour assurer la stabilité asymptotique du système bouclé ainsi que la robustesse par rapport à des dynamiques négligées.*

Exercice 29 *Soit le système de la figure 4.3. On rajoute un amortisseur linéaire entre les deux masses. Ainsi (4.6) devient ($a > 0$ est le coefficient de frottement)*

$$\begin{cases} m_1\ddot{x}_1 = k(x_2 - x_1) + a(\dot{x}_2 - \dot{x}_1) + u \\ m_2\ddot{x}_2 = k(x_1 - x_2) + a(\dot{x}_1 - \dot{x}_2). \end{cases}$$

Montrer que le système reste commandable et calculer sa sortie de Brunovsky.

4.2.6 Linéarisation par bouclage

Equivalence statique

La relation d'équivalence qui permet de mettre un système linéaire $\dot{x} = Ax + Bu$ commandable sous forme de Brunovsky peut être prolongée de la manière suivante. Au lieu de considérer des transformations du type

$$\begin{bmatrix} x \\ u \end{bmatrix} \mapsto \begin{bmatrix} Mx \\ Kx + Nu \end{bmatrix}$$

avec M et N matrices inversibles, considérons des transformations inversibles plus générales et non linéaires suivantes

$$\begin{bmatrix} x \\ u \end{bmatrix} \mapsto \begin{bmatrix} z = \phi(x) \\ v = k(x,u) \end{bmatrix}$$

où ϕ est un difféomorphisme et à x bloqué, $u \mapsto k(x,u)$ également. Il est donc logique de considérer maintenant les systèmes non linéaires de la forme $\dot{x} = f(x,u)$ et leur classification modulo le groupe de transformations ci-dessus. La relation d'équivalence qui en résulte est appelée équivalence par bouclage statique régulier et changement de coordonnées (d'une façon plus abrégée équivalence statique). Décider si deux systèmes avec les mêmes nombres d'états et des commandes, $\dot{x} = f(x,u)$ et $\dot{z} = g(z,v)$, (f, g régulières) sont équivalents, est un problème de géométrie très compliquée et largement ouvert. En revanche, il existe une caractérisation explicite des systèmes non linéaires équivalents aux systèmes linéaires commandables.

CNS de linéarisation statique

L'intérêt pratique est le suivant. Les équations issues de la physique $\dot{x} = f(x,u)$ sont en général non linéaires dans les coordonnées de modélisation x et u. La question " Existe-t-il des coordonnées, $z = \phi(x)$ et $v = k(x,u)$, qui rendent les équations linéaires, $\dot{z} = Az+Bv$ avec (A,B) commandable? " est alors d'importance. En effet, une réponse positive signifie que le système est faussement non linéaire: le système est alors dit linéarisable par bouclage statique. Il suffit de changer de "repère" pour que tout devienne linéaire.

A partir de maintenant, nous considérons le système

$$\dot{x} = f(x,u), \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m$$

avec f régulière et $f(0,0) = 0$. Notre point de vue sera local autour de l'équilibre $(x,u) = (0,0)$. Il peut être élargi à l'espace tout entier sans difficulté importante.

Lemme 1 *Les deux propositions suivantes sont équivalentes*

1. *Le système étendu*
$$\begin{cases} \dot{x} = f(x,u) \\ \dot{u} = \bar{u} \end{cases} \tag{4.9}$$
est linéarisable par bouclage statique ($\bar{u}$ est ici la commande)
2. *Le système*
$$\dot{x} = f(x,u) \tag{4.10}$$
est linéarisable par bouclage statique.

Preuve Si $x = \phi(z)$ et $u = k(z,v)$ transforment (4.10) en un système linéaire commandable $\dot{z} = Az + Bv$, alors $(x,u) = (\phi(z),k(z,v))$ et $\bar{u} = \frac{\partial k}{\partial z}(Az + Bv) + \frac{\partial k}{\partial v}\bar{v}$ transforment (4.9) en

$$\dot{z} = Az + Bv, \quad \dot{v} = \bar{v} \tag{4.11}$$

système linéaire commandable. Ainsi la seconde proposition implique la première.

Supposons maintenant la première proposition vraie. Comme tout système linéaire commandable peut s'écrire sous la forme (4.11) avec (A,B) commandable, (cf forme de Brunovsky) il existe une transformation $(x,u) = (\phi(z,v),\psi(z,v))$ et $\bar{u} = k(z,v,\bar{v})$ qui transforme (4.9) en (4.11) avec $\dim(z) = \dim(x)$. Cela veut dire que pour tout $(z,v,\bar{v})$

$$\frac{\partial \phi}{\partial z}(z,v)(Az + Bv) + \frac{\partial \phi}{\partial v}\bar{v} = f(\phi(z,v),\psi(z,v)).$$

Donc ϕ ne dépend pas de v et la transformation inversible $x = \phi(z)$, $u = \psi(z,v)$ transforme (4.10) en $\dot{z} = Az + Bv$. ∎

Ainsi, quitte à étendre l'état en posant $\dot{u} = \bar{u}$ et en prenant comme entrée $\bar{u}$, on peut toujours supposer que f est affine en u, i.e., que le système admet les équations

$$\dot{x} = f(x) + u_1 g_1(x) + \ldots + u_m g_m(x) \tag{4.12}$$

où f et les g_i sont des champs de vecteurs réguliers. Il est alors facile de voir que les transformations $x = \phi(z)$ et $u = k(z,v)$ qui rendent le système linéaire sont nécessairement affines en v, i.e., $k(x,v) = \alpha(x) + \beta(x)v$ avec β inversible pour tout x.

Prenons maintenant un changement régulier de variables: $x = \phi(z)$ d'inverse $\psi = \phi^{-1}$, $z = \psi(x)$. Considérons maintenant le système défini par (4.12) dans le repère x. Dans le repère z, nous avons les équations suivantes:

$$\dot{z} = (D\psi \cdot f + u_2 D\psi \cdot g_1 + \ldots + u_m D\psi \cdot g_m)_{x=\phi(z)} \tag{4.13}$$

où $D\psi$ est la matrice jacobienne de ψ: $\left(\frac{\partial \psi_i}{\partial x_j}\right)_{i,j}$. Ainsi f (resp. g_k) devient $D\psi \cdot f$ (resp. $D\psi \cdot g_k$).

A partir de ces champs de vecteurs définissant (4.12), on définit une suite croissantes d'espaces vectoriels indexés par x par la récurrence suivante

$$E_0 = \{g_1, \ldots, g_m\}, \qquad E_i = \{E_{i-1}, [f, E_{i-1}]\} \quad i \geq 1$$

où $[f,g]$ est le crochet de Lie de deux champs de vecteurs f et g et où $\{\ \}$ signifie espace vectoriel engendré par les vecteurs à l'intérieur des parenthèses. On rappelle que le crochet de deux champs de vecteurs f et g, de composantes $(f_1(x),...,f_n(x))$ et $(g_1(x),...,g_n(x))$ dans les coordonnées $(x_1,...,x_n)$, admet comme composantes dans les mêmes coordonnées x

$$[f,g]_i = \sum_{k=1}^{n} \frac{\partial f_i}{\partial x_k} g_k - \frac{\partial g_i}{\partial x_k} f_k.$$

Un simple calcul montrent que si $z = \psi(x)$ est un changement régulier de variables on obtient les composantes du crochet $[f,g]$ dans les coordonnées z par les mêmes formules que dans les coordonnées x. Cela veut dire que

$$D\psi.[f,g] = [D\psi.f, D\psi.g].$$

Ainsi on sait faire du calcul différentiel intrinsèque sans passer par un choix particulier de repère. Les E_k deviennennt, dans les coordonnées z, $D\psi.E_k$. On appelle ce type d'objet des distributions (rien à voir avec les distributions de Laurent Schwartz). Ce sont des objets intrinsèques car la méthode de construction de E_k ne dépend pas du système de coordonnées choisies pour faire les calculs. Le résultat suivant date des années 1980.

Théorème 13 CNS linéarisation statique *Autour de l'équilibre $(x,u) = (0,0)$, le système (4.12) est linéarisable par bouclage statique régulier si, et seulement si, les distributions E_i, $i = 1,\ldots,n-1$ définies ci-dessus sont involutives (stables par le crochet de Lie), de rang constant autour de $x = 0$ et le rang de E_{n-1} vaut n, la dimension de x.*

Une distribution E est dite involutive, si et seulement si, pour tous champs de vecteurs f et g dans E (pour tout x, $f(x)$ et $g(x)$ appartiennent à l'espace vectoriel $E(x)$), alors le crochet $[f,g]$ reste aussi dans E.

Preuve Il est évident que les distributions E_i restent également inchangées par bouclage statique $u = \alpha(x) + \beta(x)v$ avec $\beta(x)$ inversible. Comme pour un système linéaire commandable $\dot{x} = Ax + Bu$, E_i correspondent à l'image de $(B, AB, \ldots, A^iB)$, les conditions sur les E_i sont donc nécessaires.

Leur coté suffisant repose essentiellement sur le théorème de Frobenius [10]. Ce résultat classique de géometrie différentielle dit que toute distribution involutive E de rang constant m correspond, dans des coordonnées adaptées $w = (w_1,...,w_n)$, à l'espace vectoriel engendré par les m premières composantes. On a l'habitude de noté $\partial/\partial w_k$ le champ de vecteurs de composantes $(\delta_{i,k})_{1 \leq i \leq n}$ dans les coordonnées w. Alors $E = \{\partial/\partial w_1, ..., \partial/\partial w_m\}$.

Si les E_i vérifient les conditions du théorème, alors il existe un système de coordonnées locales $(x_1,\ldots,x_n)$ autour de 0 tel que

$$E_i = \left\{ \frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_{\sigma_i}} \right\}$$

où σ_i est le rang de E_i. Dans ces coordonnées locales, $\dot{x}_i$ pour $i > \sigma_0$ ne dépend pas de la commande u. Ainsi, en remplaçant u par $\alpha(x) + \beta(x)u$ avec une matrice inversible β bien choisie, la dynamique (4.12) s'écrit nécessairement ainsi

$$\begin{aligned} \dot{x}_i &= u_i, \quad i = 1,\ldots,\sigma_0 \\ \dot{x}_i &= f_i(x), \quad i = \sigma_0 + 1,\ldots,n. \end{aligned}$$

Un raisonnement simple montre que, pour $i > \sigma_1$, f_i ne dépend pas de $(x_1,\ldots,x_{\sigma_0})$ car E_1 involutive. Ainsi nous avons la structure suivante

$$\begin{aligned} \dot{x}_i &= u_i, \quad i = 1,\ldots,\sigma_0 \\ \dot{x}_i &= f_i(x_1,\ldots,x_n), \quad i = \sigma_0 + 1,\ldots,\sigma_1 \\ \dot{x}_i &= f_i(x_{\sigma_0+1},\ldots,x_n), \quad i = \sigma_1 + 1,\ldots,n. \end{aligned}$$

De plus le rang de $(f_{\sigma_0+1},\ldots,f_{\sigma_1})$ par rapport à $(x_1,\ldots,x_{\sigma_0})$ vaut $\sigma_1 - \sigma_0$. Donc $\sigma_0 \leq \sigma_1 - \sigma_0$. Quitte à faire des permutations sur les σ_0 premières composantes de x, on peut supposer que $(x_1,\ldots,x_{\sigma_1-\sigma_0}) \mapsto (f_{\sigma_0+1},\ldots,f_{\sigma_1})$ est inversible. Cela permet de définir un nouveau système de coordonnées en remplaçant les $\sigma_1 - \sigma_0$ premières composantes de x par $(f_{\sigma_0+1},\ldots,f_{\sigma_1})$. Dans ces nouvelles coordonnées et après bouclage statique régulier $u \mapsto \beta(x)u$ avec $\beta(x)$ inversible bien choisi, nous avons la structure suivante (les notations avec u, x et f sont conservées) :

$$\begin{aligned} \dot{x}_i &= u_i, \quad i = 1,\ldots,\sigma_0 \\ \dot{x}_i &= x_{i-\sigma_0}, \quad i = \sigma_0 + 1,\ldots,\sigma_1 \\ \dot{x}_i &= f_i(x_{\sigma_0+1},\ldots,x_n), \quad i = \sigma_1 + 1,\ldots,n. \end{aligned}$$

On sait que ce système est linéarisable si, et seulement si, le système réduit

$$\begin{aligned} \dot{x}_i &= x_{i-\sigma_0}, \quad i = \sigma_0 + 1,\ldots,\sigma_1 \\ \dot{x}_i &= f_i(x_{\sigma_0+1},\ldots,x_n), \quad i = \sigma_1 + 1,\ldots,n \end{aligned}$$

avec $(x_1,\ldots,x_{\sigma_1-\sigma_0})$ comme commande. Comme les distributions E_i associées à ce système réduit se déduisent simplement de celles du système étendu en éliminant les champs de vecteurs $\frac{\partial}{\partial x_1}$, $\ldots \frac{\partial}{\partial x_{\sigma_0}}$, on voit qu'elles vérifient, elles aussi, les conditions du théorème. Ainsi il est possible de réduire encore le système. A chaque étape, la linéarisation du système étendu est équivalente à celle du système réduit. Au bout de cette élimination (en au plus de $n-1$ étapes), la linéarisation du système de départ est alors équivalent à celle d'un système réduit de la forme

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

où le rang de f par rapport à u est égale à la dimension de x, linéarisation qui est alors triviale. ■

Bouclage dynamique

Le lemme 1 est trompeur. Il semble suggérer que le fait d'étendre un système en rajoutant des dérivées de la commande dans l'état ne rajoute rien pour la linéarisation. Ceci est vrai si on rajoute le même nombre d'intégrateurs sur toutes les commandes ("prolongation totale"). Par contre, des nombres différents peuvent permettre de gagner quelque chose. Par exemple le système

$$\ddot{x} = -u_1 \sin\theta, \quad \ddot{z} = u_1 \cos\theta - 1, \quad \ddot{\theta} = u_2$$

n'est pas linéarisable par bouclage statique bien que le système étendu

$$\ddot{x} = -u_1 \sin\theta, \quad \ddot{z} = u_1 \cos\theta - 1, \quad \ddot{u}_1 = \bar{u}_1, \quad \ddot{\theta} = u_2$$

de commande ($\bar{u}_1$,u_2) le soit. Ce fait n'est nullement contraire au lemme 1 puisque seule l'entrée u_1 a été prolongée deux fois. Pour un système à une seule commande, on ne gagne évidement rien.

Cette remarque est à l'origine de la linéarisation par bouclage dynamique. Un système $\dot{x} = f(x,u)$ est dit linéarisable par bouclage dynamique régulier, si, et seulement si, il existe un compensateur dynamique régulier

$$\dot{\xi} = a(x,\xi,v), \quad u = k(x,\xi,v),$$

tel que le système bouclé

$$\dot{x} = f(x,k(x,\xi,v)), \quad \dot{\xi} = a(x,\xi,v)$$

soit linéarisable par bouclage statique régulier. Noter que la dimension de ξ est libre. La dimension de l'espace dans lequel on doit travailler peut a priori être arbitrairement grande. Noter également que les compensateurs dynamiques qui consistent à ne prolonger que les entrées, sont des compensateurs particuliers. Ils ne permettent pas de linéariser certains systèmes comme celui ci :

$$\begin{aligned} \ddot{x} &= \varepsilon u_2 \cos\theta - u_1 \sin\theta \\ \ddot{z} &= \varepsilon u_2 \sin\theta + u_1 \cos\theta - g \\ \ddot{\theta} &= u_2. \end{aligned}$$

En effet, on peut montrer que, quelque soit le compensateur dynamique de la forme $u_1^{(\alpha_1)} = \bar{u}_1$, $u_2^{(\alpha_2)} = \bar{u}_2$ (α_1 et α_2 entiers arbitraires), le système étendu n'est pas linéarisable par bouclage statique. En revanche il est linéarisable par le bouclage dynamique endogène construit en 2.2.4.

Cette question est à l'origine des systèmes plats, les systèmes linéarisables par des bouclages dynamiques dits endogènes et auxquels est associée une relation d'équivalence (i.e., une géométrie). Pour en savoir plus voir [19].

4.3 Observabilité non linéaire

Nous considérons les systèmes non linéaires de la forme :

$$\begin{cases} \dfrac{dx}{dt} = f(x,u) \\ \quad y = h(x) \end{cases} \tag{4.14}$$

avec $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ et $y \in \mathbb{R}^p$. Les fonctions$f$ et h sont régulières.

4.3.1 Définition

Pour définir l'observabilité, il convient d'abord de définir la notion de distinguabilité.

Définition 19 (distinguabilité) *Deux états initiaux x et $\widetilde{x}$ sont dits indistinguables (notés $xI\widetilde{x}$) si pour tout $t \geq 0$, les sorties $y(t)$ et $\widetilde{y}(t)$ sont identiques pour toute entrée $u(t)$ admissible*[3]*. Ils sont dits distinguables sinon.*

L'indistinguabilité est une relation d'équivalence. Notons $I(x)$ la classe d'équivalence de x. L'observabilité est alors définie de la manière suivante :

Définition 20 (observabilité globale) *Le système (4.14) est dit observable en x si $I(x) = \{x\}$ et il est observable si $I(x) = \{x\}$ pour tout x.*

En fait, le système est observable si pour tous les états initiaux x et $\widetilde{x}$, il existe une entrée admissible u qui distingue x et $\widetilde{x}$, c'est à dire telle que $y(t) \neq \widetilde{y}(t)$ pour au moins un temps $t \geq 0$.

Il peut exister des entrées qui ne distinguent pas certains points. Cependant, le système peut être malgré tout observable. Par exemple

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & ux_2 \\ \dot{x}_2 & = & 0 \\ y & = & x_1 \end{array} \right.$$

est observable (pour $u = 1$ par exemple). Cependant l'entrée $u = 0$ ne distingue pas les point x et $\tilde{x}$ tel que $x_1 = \tilde{x}_1$ et $x_2 \neq \tilde{x}_2$. Notons que l'observabilité ne signifie pas que toute entrée distingue tous les états. L'observabilité est un concept global. Il peut être nécessaire d'aller très loin dans le temps et dans l'espace d'état pour distinguer deux états initiaux. Pour cela nous introduisons le concept plus fort :

Définition 21 (observabilité locale en temps et en espace) *L'état x de (4.14) est localement observable, si pour tout $\varepsilon > 0$ et pour tout voisinage U de x, il existe $\eta > 0$ plus petit que ε et un voisinage V de x contenu dans U, tel que pour tout $\widetilde{x} \in V$, il existe une entrée $[0,\eta] \ni t \mapsto u(t)$ qui distingue x et $\widetilde{x}$, i.e. telle que $y(\eta) \neq \widetilde{y}(\eta)$. Le système (4.14) est localement observable s'il l'est pour tout x.*

Intuitivement, le système (4.14) est localement observable si on peut instantanément distinguer chaque état de ses voisins en choisissant judicieusement l'entrée u.

4.3.2 Critère

La seule façon effective de tester l'observabilité d'un système est de considérer l'application qui à x associe y et ses dérivées en temps. Nous supposerons dans cette section que y et u sont des fonctions régulières du temps. Nous supposerons également que les rangs en x des fonctions de $(x,u,\dot{u},\ldots)$ qui apparaissent ci-dessous sont constants.

Considérons donc (4.14). On note $h_0(x) := h(x)$. En dérivant y par rapport au temps on a

$$\dot{y} = D_x h(x)\dot{x} = D_x h(x) \cdot f(x,u) := h_1(x,u).$$

3. $y(t)$ (resp. $\widetilde{y}(t)$) correspond à la sortie de (4.14) avec l'entrée $u(t)$ et la condition initiale x (resp. $\widetilde{x}$).

Des dérivations successives conduisent donc à une suite de fonctions $h_k(x,u,\ldots,u^{(k-1)})$ définie par la récurrence

$$h_{k+1} = \frac{d}{dt}(h_k), \quad h_0(x) = h(x).$$

Si pour un certain k, le rang en x du système

$$\begin{cases} h_0(x) = y \\ h_1(x,u) = \dot{y} \\ \vdots \\ h_k(x,u,\ldots,u^{(k-1)}) = y^{(k)} \end{cases}$$

vaut $n = \dim(x)$ alors le système est localement observable. Il suffit d'utiliser le théorème d'inversion locale pour calculer x en fonction de $(y,\ldots,y^{(k)})$ et $(u,\ldots,u^{(k-1)})$. Si à partir d'un certain k, h_{k+1} ne fait plus apparaître de nouvelle relation en x, i.e., si le rang en x de $(h_0,\ldots,h_k)'$ est identique à celui de $(h_0,\ldots,h_k,h_{k+1})'$, alors il en est de même pour $k+2$, $k+3$, ... Ainsi, il n'est pas nécessaire de dériver plus de $n-1$ fois y pour savoir si un système est localement observable ou non. Ce raisonnement est correct autour d'un état générique, nous ne traitons pas les singularités qui peuvent apparaître en des états et entrées particulières. Nous renvoyons à [18] pour les cas plus généraux avec singularités.

Ce calcul élémentaire montre aussi que y et u sont reliés par des équations différentielles. Elles correspondent aux relations de compatibilité associées au système sur-déterminé (4.14) où l'inconnue est x et les données sont u et y. On obtient toutes les relations possibles en éliminant x du système

$$\begin{cases} h_0(x) = y \\ h_1(x,u) = \dot{y} \\ \vdots \\ h_n(x,u,\ldots,u^{(n-1)}) = y^{(n)}. \end{cases}$$

On peut montrer que pour un système localement observable, u et y sont reliés par $p = \dim(y)$ équations différentielles indépendantes. Ces équations font intervenir y dérivé au plus n fois et u dérivé au plus $n-1$ fois.

La mise en forme des idées précédentes est assez fastidieuse mais néanmoins instructive. Nous nous contenterons de retenir qu'en général l'observabilité signifie que l'état peut être exprimé en fonction des sorties, des entrées et d'un nombre fini de leur dérivées en temps. Dans ce cas, y et u sont reliés par p équations différentielles d'ordre au plus n en y et $n-1$ en u.

Pour conclure, reprenons l'exemple du réacteur chimique (4.2) (page 85) afin d'illustrer l'analyse formelle précédente. Nous ne considérons que x_1 et T car l'invariant chimique x_1+x_2 est supposé égal à x_1^{in}. Nous supposons que la température T est mesurée (thermocouple) mais pas la concentration x_1. Nous avons donc à résoudre le système sur-déterminé (les quantités autres que (x_1,u,y,T) sont des constantes connues)

$$\begin{aligned} \dot{x}_1 &= D(x_1^{in} - x_1) - k_0 \exp(-E/RT)x_1 \\ \dot{T} &= D(T^{in} - T) + \alpha\Delta H \exp(-E/RT)x_1 + u \\ y(t) &= T. \end{aligned}$$

On a facilement x_1 en fonction de $(y,\dot{y})$ et u :

$$x_1 = \frac{\dot{y} - D(T^{in} - y) - u}{\alpha \Delta H \exp(-E/Ry)}. \tag{4.15}$$

Le système est donc observable. y et u sont reliés par une équation différentielle du second ordre en y et du premier ordre en u. On l'obtient en utilisant l'équation donnant $\dot{x}_1$:

$$\frac{d}{dt}\left(\frac{\dot{y} - D(T^{in} - y) - u}{\alpha \Delta H \exp(-E/Ry)}\right) = Dx_1^{in} - (D + k_0 \exp(-E/Ry)) \frac{\dot{y} - D(T^{in} - y) - u}{\alpha \Delta H \exp(-E/Ry)}. \tag{4.16}$$

Il s'agit d'une condition de compatibilité entre y et u. Si elle n'est pas satisfaite alors le système sur-déterminé de départ n'admet pas de solution. On conçoit très bien que ces relations de compatibilité sont à la base du *diagnostique et de la détection de panne*.

4.3.3 Observateur, estimation, moindre carré

Savoir que le système est observable est bien. Calculer x à partir de y et u est encore mieux. Cependant, la démarche formelle précédente ne répondre en pratique qu'à la première question. En effet, avoir x en fonction de dérivées des mesures s'avère d'une utilité fort limitée dès que l'ordre de dérivation dépasse 2 et/ou dès que les signaux sont bruités. Il convient en fait de calculer x en fonction d'intégrales de y et u. Dans ce cas, le bruit sur les signaux est beaucoup moins gênant. La synthèse d'observateur, c'est à dire estimer x sans utiliser les dérivées de y, pose des problèmes supplémentaires (et nettement plus difficiles en fait) que la caractérisation des systèmes observables.

Revenons à (4.14). Nous avons en fait un nombre infini d'équations en trop. En effet, puisque l'entrée u est connue, l'état est entièrement donné par sa condition initiale x grâce au flot ϕ_t^u de $\dot{\xi} = f(\xi,u(t))$. Ainsi x vérifie à chaque instant t, p équations, p étant donc le nombre de mesures :

$$y(t) = h(\phi_t^u(x)).$$

Il est très tentant de résoudre ce système par les moindres carrés, même si, pour un système non-linéaire cela n'a pas beaucoup de sens (dépend du choix des coordonnées et de la méthode utilisée pour mesurer les écarts). Fixons nous un intervalle d'observation $[0,T]$. x peut être calculé comme l'argument du minimum de

$$J(\xi) = \int_0^T (y(t) - h(\phi_t^u(x))^2 \, dt.$$

x est ainsi obtenu comme on obtient un paramètre à partir de données expérimentales et d'un modèle où ce paramètre intervient : en minimisant l'erreur quadratique entre l'observation $y(t)$ et la valeur prédite par le modèle $\phi_t^u(x)$. Ainsi les problèmes d'observateur sont fondamentalement proches des problèmes d'estimation pour lesquels l'optimisation joue un rôle important. Cependant les difficultés ne sont pas pour autant aplanies : le calcul du flot, i.e., la résolution de l'équation différentielle $\dot{x} = f(x,u)$ ne peut se faire que numériquement en général; la fonction J n'a aucune raison d'avoir les bonnes propriétés de convexité qui assure la convergence des principaux algorithmes d'optimisation (c.f.[14]). La synthèse d'observateur reste donc une question difficile en général bien que

très importante en pratique. Noter enfin que l'identification de paramètres θ sur un modèle $\dot{x} = f(x,u,\theta)$ est un sous-problème : l'identifiabilité correspond alors à l'observabilité du système

$$\dot{x} = f(x,u,\theta), \quad \dot{\theta} = 0, \quad y = x$$

d'état (x,θ) et de sortie $y = x$.

Dans le cas linéaire, $f = Ax + Bu$ et $h = Cx$, $\phi_t^u(x)$ est une fonction affine en x :

$$\phi_t^u(x) = \exp(tA)x + \int_0^t \exp((t-s)A)Bu(s)\ ds.$$

Avec un intervalle d'observation $[0,T]$, x peut être calculé comme l'argument du minimum de

$$J(\xi) = \int_0^T (z(t) - C\exp(tA)x)^2\ dt \tag{4.17}$$

où $z(t) = y(t) - C\int_0^t \exp((t-s)A)Bu(s)\ ds$. Nous voyons clairement que J est quadratique. On retrouve alors le filtre de Kalman dans le cadre déterministe et la commande LQG. Cet aspect étant traité par ailleurs, nous n'en parlerons pas. Nous allons maintenant aborder l'observabilité des systèmes linéaires avec un point de vue moins classique qui met l'accent sur les observateurs asymptotiques. Ces derniers fournissent, avec des calculs très économiques, x en fonction de y, u et leurs intégrales.

4.4 Observabilité linéaire

On considère ici le système, d'entrée u, d'état x et de sortie y suivant

$$\begin{array}{rcl} \dot{x} & = & Ax + Bu \\ y & = & Cx \end{array} \tag{4.18}$$

où A est une matrice $n \times n$, B une matrice $n \times m$ et C une matrice $p \times n$.

4.4.1 Le critère de Kalman

Théorème 14 (critère de Kalman) *Le système $\dot{x} = Ax + Bu$, $y = Cx$ est observable au sens de la définition 21 si, et seulement si, le rang de la matrice d'observabilité*

$$\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}$$

est égal à $n = \dim(x)$.

Pour abréger, on dit souvent que *la paire (A,C) est observable* lorsque le rang de la matrice d'observabilité $\mathcal{O}$ est maximum.

Preuve Dérivons y et d'utilisons l'équation d'état. Une première dérivation donne

$$\dot{y} = C\dot{x} = CAx + CBu.$$

Donc x est nécessairement solution du système (les fonctions y et u sont connues)

$$\begin{aligned} Cx &= y \\ CAx &= \dot{y} - CBu. \end{aligned}$$

A ce niveau, tout se passe comme si la quantité $\bar{y}_1 = \dot{y} - CBu$ était une nouvelle sortie. En la dérivant de nouveau, nous avons $CA^2x = \dot{\bar{y}}_1 - CABu$. Maintenant, x est nécessairement solution du système étendu

$$\begin{aligned} Cx &= \bar{y}_0 = y \\ CAx &= \bar{y}_1 = \dot{y} - CBu \\ CA^2x &= \bar{y}_2 = \dot{\bar{y}}_1 - CABu. \end{aligned}$$

Il est alors facile de voir que x sera nécessairement solution des équations

$$CA^k x = \bar{y}_k \tag{4.19}$$

où les quantités connues $\bar{y}_k$ sont définies par la récurrence $\bar{y}_k = \dot{\bar{y}}_{k-1} - CA^{k-1}Bu$ pour $k \geq 1$ et $\bar{y}_0 = y$.

Si le rang de la matrice d'observabilité est maximum et égal à n, elle admet un inverse à gauche (non nécessairement unique), P matrice $n \times pn$ vérifiant

$$P \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = 1_n.$$

Ainsi

$$x = P \begin{pmatrix} \bar{y}_0 \\ \vdots \\ \bar{y}_{n-1} \end{pmatrix}.$$

La condition de rang est donc suffisante.

Supposons maintenant que la matrice d'observabilité, de taille $pn \times n$, soit de rang $r < n$. Nous allons montrer qu'il existe, au moins, deux trajectoires différentes avec les mêmes commandes, donnant la même sortie. Cela montrera que la condition est aussi nécessaire.

Soit $w \in \mathbb{R}^n$ un élément non nul du noyau de la matrice de commandabilité. Pour $k = 0, \ldots, n-1$, $CA^k w = 0$. Par un raisonnement identique à celui fait lors de la preuve de la proposition 6 avec les noyaux à gauche de $A^k B$, on a nécessairement $CA^k w = 0$, pour toute $k \geq n$. Donc w est dans le noyau de toutes les matrices CA^k. Prenons comme première trajectoire $[0,T] \ni t \mapsto (x,u) = 0$. Alors, $y = 0$. Prenons maintenant comme seconde trajectoire, celle qui, à commande nulle, démarre en w : $[0,T] \ni t \mapsto (x,u) = (\exp(tA)w, 0)$. Sa sortie vaut

$$C \exp(tA)w = \sum_{i=0}^{+\infty} \frac{t^i}{i!} CA^i w = 0$$

car chaque terme de la série est nul. ■

Exercice 30 *Montrer que J définie par (4.17), page 104, est strictement convexe si, et seulement si, (A,C) est observable.*

Exercice 31 *Le système (4.6), page 90, est-il observable avec comme mesure $y = x_1$? L'est-il avec $y = \dot{x}_1$?*

Exercice 32 *Donner pour les systèmes discrets linéaires*

$$x_{k+1} = Ax_k + Bu_k, \quad y_k = Cx_k$$

une définition de l'observabilité et montrer que le critère de Kalman reste inchangé.

4.4.2 Observateurs asymptotiques

Il est classique de noter par $\widehat{x}$ une estimation de la quantité x. Nous cherchons ici à obtenir une estimation de l'état sans utiliser les dérivées de y et u. La première idée qui vient à l'esprit est de copier la dynamique du système. On intègre directement

$$\dot{\hat{x}} = A\hat{x} + Bu$$

à partir d'une condition initiale $\hat{x}_O$. Si la matrice A est stable, alors $\hat{x}$ peut être pris comme estimation de x car l'erreur $e_x = \hat{x} - x$ tend vers 0 puisque $\dot{e}_x = Ae_x$.

Si A est instable cette méthode ne marchera pas. En effet, une petite erreur initiale $e_x(0)$ sera amplifiée exponentiellement. Intuitivement, si l'erreur $\hat{x}-x$ devient grande alors, le système étant observable, l'erreur sur les sorties $\hat{y} - y$ deviendra grande également [4]. Comme y est connue, il est alors tentant de modifier $\dot{\hat{x}} = A\hat{x} + Bu$ par l'ajout d'un terme du type $L(\hat{y} - y)$ qu'on connaît et qui correspond à l'erreur d'observation. Ainsi, le problème suivant se pose, peut-on choisir la matrice L de façon à ce que la solution $\hat{x}$ du système

$$\dot{\hat{x}} = A\hat{x} + Bu(t) + L(\hat{y} - y(t)), \quad \hat{y} = C\hat{x}$$

converge vers x? Puis que $y = Cx$, la question se pose ainsi : peut-on ajuster la matrice L de façon à obtenir une équation différentielle d'erreur stable :

$$\dot{e}_x = (A + LC)e_x \ ?$$

Par un choix judicieux de L, peut-on imposer à $A + LC$ d'avoir toutes ses valeurs propres à partie réelle strictement négative?

Or, les valeurs propres restent inchangées par la transposition : $A + LC$ admet le même spectre que $A' + C'L'$. De plus la paire (A,C) est observable si, et seulement si, la paire (A',C') est commandable : on obtient le critère de Kalman de commandabilité en transposant celui de l'observabilité. Ainsi le théorème 12 se transpose de la manière suivante :

Théorème 15 (observateur asymptotique) *Si (A,C) est observable, il existe L, matrice $n \times p$, telle que le spectre de $A + LC$ soit le même que celui de n'importe quelle matrice réelle $n \times n$.*

4. On a noté $\hat{y} = C\hat{x}$.

Exercice 33 (forme canonique) *Donner pour un système linéaire observable la forme canonique duale de celle de Brunovsky. Quelle est la relation d'équivalence associée à cette forme canonique ?*

4.4.3 Observateur réduit de Luenberger

Supposons que C soit de rang maximum $p = \dim(y)$ et que la paire (A,C) soit observable. On peut toujours supposer, quitte à faire un changement de variable sur x, que y correspond aux p premières composantes de l'état x : $x = (y,x_r)$. L'équation d'état $\dot{x} = Ax + Bu$ s'écrit alors sous forme blocs :

$$\begin{aligned} \dot{y} &= A_{yy}y + A_{yr}x_r + B_y u \\ \dot{x}_r &= A_{ry}y + A_{rr}x_r + B_r u. \end{aligned}$$

Il est facile de montrer, en revenant, par exemple à la définition de l'observabilité, que (A,C) est observable si, et seulement si, (A_{rr},A_{yr}) l'est : en effet connaître y et u implique la connaissance de $A_{yr}x_r = \dot{y} - A_{yy}y - B_y u$, qui peut être vu comme une sortie du système $\dot{x}_r = A_{rr}x_r + (B_r u + A_{ry}y)$.

En ajustant correctement la matrice des gains d'observation L_r, le spectre de $A_{rr} + L_r A_{yr}$ coïncide avec celui de n'importe quelle matrice réelle carrée d'ordre $n-p = \dim(x_r)$. Considérons alors la variable $\xi = x_r + L_r y$ au lieu de x_r. Un simple calcul montre que

$$\dot{\xi} = (A_{rr} + L_r A_{yr})\xi + (A_{ry} + L_r A_{yy} - (A_{rr} + L_r A_{yr})L_r)y + (B_r + L_r B_y)u.$$

Ainsi en choisissant L_r, de façon à avoir $A_{rr} + L_r A_{yr}$ stable, nous obtenons un observateur d'ordre réduit $n - p$ pour ξ (donc pour $x_r = \xi - L_r y$) en recopiant cette équation différentielle

$$\dot{\hat{\xi}} = (A_{rr} + L_r A_{yr})\hat{\xi} + (A_{ry} + L_r A_{yy} - (A_{rr} + L_r A_{yr})L_r)y(t) + (B_r + L_r B_y)u(t).$$

En effet la dynamique de l'erreur sur ξ, $e_\xi = \hat{\xi} - \xi$ vérifie l'équation autonome stable

$$\dot{e}_\xi = (A_{rr} + L_r A_{yr})e_\xi.$$

Cet observateur réduit est intéressant lorsque $n-p$ est petit, typiquement $n-p = 1,2$: la stabilité d'un système de dimension 1 ou 2 est très simple à étudier.

Exercice 34 (observateur réduit non linéaire) *Construire pour le réacteur chimique (4.2), page 85, un observateur asymptotique réduit de la concentration x_1 à partir de la mesure de température T (considérer $\xi = x_1 + \lambda T$ avec λ bien choisi).*

4.5 Observateur-contrôleur linéaire

En regroupant les résultats sur la commandabilité et l'observabilité linéaires, nous savons comment résoudre de façon robuste par rapport à de petites erreurs de modèle et de mesures, le problème suivant : amener, à l'aide de la commande u, l'état x du système de p à q pendant le temps T en ne mesurant que y sachant que : $\dot{x} = Ax + Bu$, $y = Cx$, (A,B) commandable et (A,C) observable.

En effet, comme (A,B) est commandable, nous savons avec la forme de Brunovsky construire explicitement une trajectoire de référence $[0,T] \ni t \mapsto (x_r(t),u_r(t))$ pour aller de p à q. Le respect de certaines contraintes peut-être important à ce niveau et être un guide dans le choix de cette trajectoire de référence (définition du critère pour la commande optimale).

Toujours à cause de la commandabilité, nous savons construire un bouclage statique sur x, Kx, de façon à ce que la matrice $A + BK$ soit stable (placement de pôle). La matrice K est souvent appelée *matrice des gains de la commande*.

Grâce à l'observabilité, nous savons construire un observateur asymptotique sur x en choisissant les *gains d'observation* L de façon à avoir $A + LC$ stable.

Alors le bouclage dynamique de sortie

$$\begin{aligned} u(t) &= u_r(t) + K(\hat{x} - x_r(t)) \quad \text{contrôleur} \\ \dot{\hat{x}} &= A\hat{x} + Bu(t) + L(C\hat{x} - y(t)) \quad \text{observateur} \end{aligned}$$

assure le suivi asymptotique de la trajectoire de référence $[0,T] \ni t \mapsto (x_r(t),u_r(t))$. Avec ce bouclage, appelé *commande modale* ou encore *observateur-contrôleur*, les petites erreurs de conditions initiales sont amorties lorsque t croît et les petites erreurs de modèle et de mesures ne sont pas amplifiées au cours du temps.

En effet, comme $\dot{x} = Ax + Bu$ et $y = Cx$, on a pour la dynamique du système bouclé :

$$\begin{aligned} \dot{x} &= Ax + B(u_r(t) + K(\hat{x} - x_r(t))) \\ \dot{\hat{x}} &= A\hat{x} + B(u_r(t) + K(\hat{x} - x_r(t))) + L(C\hat{x} - Cx) \end{aligned}$$

où l'état est maintenant $(x,\hat{x})$. Comme (x_r,u_r) est une trajectoire du système, $\dot{x}_r = Ax_r + Bu_r$, on a en prenant comme variables d'état $(\Delta x = x - x_r(t), e_x = \hat{x} - x)$ au lieu de $(x,\hat{x})$, la forme triangulaire suivante :

$$\begin{aligned} \frac{de_x}{dt} &= (A + LC)\ e_x \\ \frac{d(\Delta x)}{dt} &= (A + BK)\ \Delta x + BK\ e_x. \end{aligned}$$

Ce qui montre que e_x et Δx tendent vers 0 exponentiellement en temps.

Exercice 35 *On dispose sur (4.6), page 90, de deux capteurs de position $y_1 = x_1$ et $y_2 = x_2$. Calculer une commande u, ne dépendant que des mesures y_1 et y_2 et de leurs "intégrales", qui stabilise asymptotiquement en 0 (indication pour avoir des calculs simples : utiliser un observateur réduit pour les vitesses; rajouter par la commande du frottement). Reprendre la question précédente dans le cas où le ressort est non linéaire (cf. exercice 26).*

4.6 Problèmes

Problème 4 (une classe de systèmes à retard) *Soit le système à retard suivant :*

$$\dot{x}(t) = Ax(t) + Bu(t-1), \quad \dim x = n, \quad \dim u = m.$$

1. *Montrer que si la paire (A,B) est commandable, alors ce système est commandable.*

2. *Montrer la relation suivante*

$$x(t+1) = \exp(A)x(t) + \int_{t-1}^{t} \exp((t-s)A)Bu(s)\ ds.$$

En déduire un bouclage à retards repartis du type

$$u(t) = Lx(t) + \int_{t-1}^{'} R(t-s)u(s)\ ds$$

qui stabilise le système (donner l'allure des matrices L et $R(t-s)$). Montrer que cette méthode de stabilisation est robuste à de petites incertitudes sur les matrices A et B.

3. *Ecrire explicitement le bouclage pour $\dot{x}(t) = x(t) + u(t-1)$. Tester en simulation la robustesse du bouclage pour une erreur de 10% sur les paramètres du modèle.*

Problème 5 (décomposition en partie commandable et non commandable) *Pour $\dot{x} = Ax + Bu$, on note $n-p$ le rang de la matrice de commandabilité ($n = \dim(x)$). Montrer qu'il existe un changement de variable sur x uniquement $x = M\tilde{x}$, $\tilde{x} = (\tilde{x}_1, \tilde{x}_2)$ avec $\dim(\tilde{x}_2) = p$, tel que l'équation d'état admette la structure bloc suivante :*

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{A}_{11}\tilde{x}_1 + \tilde{A}_{12}\tilde{x}_2 + \tilde{B}_1 u \\ \dot{\tilde{x}}_2 &= \tilde{A}_{22}\tilde{x}_2 \end{aligned}$$

où $(\tilde{A}_{11}, \tilde{B}_1)$ est commandable. La partie non commandable correspond ainsi à une équation différentielle autonome incluse dans le système. On pourra considérer la décomposition de l'espace d'état $\mathbb{R}^n$ en une somme directe faisant intervenir l'image de la matrice de commandabilité et un espace vectoriel complémentaire de rang p, les coordonnées $\tilde{x}_1$ et $\tilde{x}_2$ étant associées à cette somme directe.

Problème 6 (décomposition en partie observable et non observable) *En s'inspirant de l'exercice 5, montrer que tout système $\dot{x} = Ax + Bu$, $y = Cx$, se décompose, par changement de variables sur l'état uniquement, ainsi :*

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{A}_{11}\tilde{x}_1 + \tilde{A}_{12}\tilde{x}_2 + \tilde{B}_1 u \\ \dot{\tilde{x}}_2 &= \tilde{A}_{22}\tilde{x}_2 + \tilde{B}_2 u \\ y &= \tilde{C}_2\tilde{x}_2 \end{aligned}$$

où $(\tilde{A}_{22}, \tilde{C}_2)$ est observable.

Problème 7 (réalisation d'un transfert causal) *Considérons le transfert causal suivant*

$$y(s) = \frac{\sum_{i=0}^{p} b_i s^i}{\sum_{i=0}^{p} a_i s^i}\ u(s)$$

où s est la variable de Laplace et correspond à l'opérateur d/dt , $\dim y = \dim u = 1$, avec $a_p \neq 0$ Ainsi y et u sont reliés par une équation différentielle d'ordre p où des dérivées de la commande u apparaissent jusqu'à l'ordre p au plus :

$$a_p y^{(p)} + \ldots + a_0 y = b_p u^{(p)} + \ldots + b_0 u.$$

1. *Montrer que la forme d'état s'écrit de la manière suivante*

$$\dot{x} = Ax + B_0 u + B_1 \dot{u} + \ldots + B_p u^{(p)}, \quad y = Cx.$$

Expliciter x, A, B_0, …, B_p et C.

2. *Montrer que le changement de variables $\tilde{x} = x - B_p u^{(p-1)}$ donne*

$$\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{B}_0 u + B_1 \dot{u} + \ldots + \tilde{B}_{p-1} u^{(p-1)}, \quad y = \tilde{C}\tilde{x}$$

et permet d'éliminer $u^{(p)}$ (un peu comme pour l'élimination de u dans la preuve de la forme de Brunovsky). Expliciter $\tilde{A}$, $\tilde{B}_0$, …, $\tilde{B}_{p-1}$ et $\tilde{C}$.

3. *En déduire un algorithme en p étapes qui réalise le transfert entre y et u sous la forme*

$$\dot{z} = Fz + Gu, \quad y = Hz + Lu.$$

4. *Donner avec l'algorithme précédent la taille et les valeurs des matrices F, G, H et L, pour $p = 1$ et $p = 2$ en fonction des a_i et des b_i.*
5. *Etendre l'algorithme au cas multi-variable.*

Problème 8 (régulation de niveau) *Un débit liquide F variable au cours du temps (la perturbation) entre dans un réservoir contenant un volume V de liquide (l'état). Ce réservoir possède un soutirage liquide dont le débit L est ajustable avec une vanne (la commande). Sauf indication contraire, on suppose que l'alimentation $t \mapsto F(t)$ est connue et que le volume V est mesuré par l'intermédiaire d'une mesure de niveau. Le modèle élémentaire de ce système est*

$$\frac{dV}{dt} = F - L.$$

1. *On veut maintenir le niveau V à une consigne fixe V_c. Quelle loi de bouclage sur L proposez-vous?*
2. *La consigne est maintenant variable : $t \mapsto V_c(t)$ est une fonction C^1. Comment modifier la loi précédente de façon à suivre asymptotiquement la trajectoire $t \mapsto V_c(t)$, i.e., de façon à avoir $\lim_{t\mapsto\infty}(V(t) - V_c(t)) = 0$?*
3. *On suppose à partir de maintenant que F est fixe mais inconnue. La commande que vous proposez assure-t-elle $\lim_{t\mapsto\infty} V(t) = V_c$? Comment la modifier de façon à assurer la convergence vers V_c ? (rajouter un terme intégral).*
4. *Une façon de faire est de construire un observateur pour F.*
   (a) *Montrer que le système*

$$\dot{V} = F - L, \quad \dot{F} = 0$$

est observable avec comme sortie $y = V$, comme commande $u = L$ et comme état $x = (V,F)$.

   (b) *Montrer que, si la constante $\lambda < 0$, $\xi - \lambda V$ converge vers F où ξ est solution de*

$$\dot{\xi} = \lambda\xi - \lambda^2 V - \lambda L.$$

   (c) *Montrer alors que la commande de la question 1 où F est remplacé par $\xi - \lambda V$ assure le suivi asymptotique de V_c. Calculer le transfert du système en boucle fermée $V_c \mapsto V$. Que remarque-t-on sur les zéros du numérateur ?*

(d) *Que se passe-t-il si F varie selon une loi affine en temps $F(t) = F_0 + Qt$, (Q constante) ?*

5. *Reprendre la construction de l'observateur en prenant comme modèle de perturbation que $\ddot{F} = 0$, au lieu de $\dot{F} = 0$. Donner le transfert en boucle fermée entre $V_c \mapsto V$. Montrer que 0 annule à l'ordre 2 le numérateur. Que se passe-t-il maintenant si F varie selon une loi affine en temps ?*

Problème 9 (dynamique verticale d'une montgolfière) *Il s'agit de piloter la dynamique verticale d'une montgolfière, la dynamique horizontale étant très peu commandable comme chacun sait (c'est justement cette partie non commandable qui fait tout le charme de l'engin ...).*

On note θ l'écart de température par rapport à l'équilibre dans le ballon, v la vitesse ascensionnelle et h l'altitude. Un premier modèle simple est le suivant :

$$\begin{aligned} \dot{\theta} &= -\theta/\tau_1 + u \\ \dot{v} &= -v/\tau_2 + \sigma\theta + w/\tau_2 \\ \dot{h} &= v \end{aligned}$$

où $\tau_1 > 0$ et $\tau_2 > 0$ sont des constantes de temps fixes, σ est un paramètre de couplage correspondant à la poussée d'Archimède. w est la vitesse verticale du vent, considérée ici comme une perturbation. u est la commande proportionnelle à la chaleur fournie au ballon par le brûleur.

1. *On suppose dans cette question que la commande est w et que u est une perturbation. Le système est-il commandable ? Peut-on le stabiliser par un bouclage d'état ?*
2. *On suppose que u est la commande que w est une perturbation constante. Montrer que le système est commandable. Qu'elle est sa sortie de Brunovsky y ? Construire un contrôleur qui permet de suivre une trajectoire régulière $t \mapsto y_c(t)$ sur y.*
3. *On désire maintenant aller d'une altitude stabilisée h_0 vers une autre altitude stabilisée h_1. Comment faire, en sachant que la commande doit rester comprise entre deux bornes $-a^2 \leq u \leq b^2$?*
4. *On suppose que l'on dispose d'un altimètre donnant h. Peut-on en déduire v, θ et w en supposant qu'on connaisse u (c'est un minimum) et que w varie peu, i.e. $\dot{w} = 0$?*
5. *On suppose que u est la commande, h la mesure et que w est une perturbation constante. Construire l'observateur qui permet de reconstruire asymptotiquement l'état.*

Problème 10 (satellite en orbite) *On s'intéresse ici à la position (et non à l'orientation) d'un satellite de masse m tournant autour de la terre dont le référentiel est supposé galiléen. La position du centre de gravité est repérée avec les coordonnées sphériques (r,θ,φ).*

1. *Montrer que son énergie cinétique T est donnée par*

$$T = m(\dot{r}^2 + r^2\dot{\varphi}^2 + r^2\dot{\theta}^2\cos^2\varphi)/2$$

et que son énergie potentielle U vaut $U = -km/r$ où k est une constante. En déduire les équations du mouvement suivantes

$$\begin{aligned} \ddot{r} &= r\dot{\theta}^2\cos^2\varphi + r\dot{\varphi}^2 - k/r^2 + u_r/m \\ \ddot{\theta} &= -2\dot{r}\dot{\theta}/r + 2\dot{\theta}\dot{\varphi}\sin\varphi/\cos\varphi + u_\theta/(mr\cos\varphi) \\ \ddot{\varphi} &= -\dot{\theta}^2\cos\varphi\sin\varphi - 2\dot{r}\dot{\varphi}/r + u_\varphi/(mr) \end{aligned}$$

où $u = (u_r,u_\theta,u_\varphi)$ sont les composantes en sphériques des forces exercées par les trois moteurs sur le satellite (on peut retrouver aussi ces équations à partir des équations de Newton écrites en coordonnées sphériques).

2. *Montrer que la trajectoire équatoriale $r(t) = \bar{r}$, $\theta(t) = \bar{\omega}t$ et $\varphi(t) = 0$ est une trajectoire du système à commande $u = 0$ si $\bar{\omega}^2\bar{r}^3 = k$ (loi de Kepler). Vérifier que les équations linéarisées autour de cette trajectoire sont données par*

$$\begin{aligned} \ddot{\Delta}r &= 3\bar{\omega}^2\Delta r + 2\bar{\omega}\bar{r}\Delta\dot{\theta} + \Delta u_r/m \\ \ddot{\Delta}\theta &= -2(\bar{\omega}/\bar{r})\Delta\dot{r} + \Delta u_\theta/(m\bar{r}) \\ \ddot{\Delta}\varphi &= -\bar{\omega}^2\Delta\varphi + u_\varphi/(m\bar{r}). \end{aligned}$$

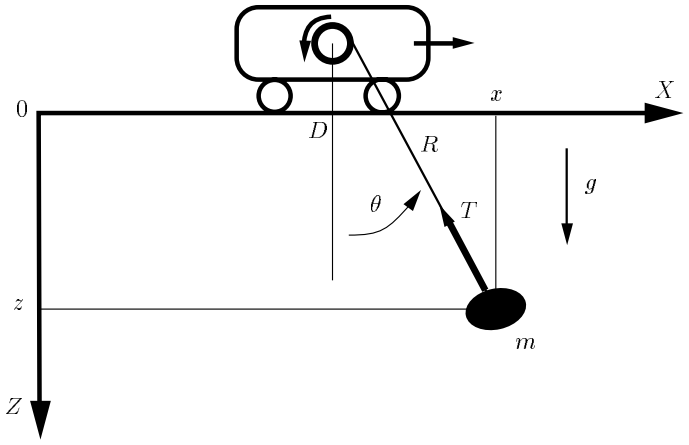
Que remarque-t-on ?

3. *Montrer que le système linéaire tangent est commandable et le mettre sous forme de Brunovsky. En déduire un bouclage qui stabilise le satellite autour de cette trajectoire. Comment placeriez-vous les pôles du système?*
4. *En partant directement du système non linéaire de départ construire un bouclage non linéaire cette fois-ci qui linéarise le système*

$$\ddot{r} = v_r, \quad \dot{\theta} = v_\theta, \quad \ddot{\varphi} = v_\varphi.$$

En déduire la commandabilité ainsi qu'un bouclage stabilisant autour de n'importe quelle trajectoire de référence C^2, $t \mapsto (r_c(t),\theta_c(t),\varphi_c(t))$.

5. *Généraliser encore au cas des systèmes mécaniques complètement commandés, c'est-à-dire ayant autant de degrés de liberté que de commandes indépendantes (comme les robots avec au moins un moteur par axe). En particulier montrer qu'ils sont commandables.*



Problème 11 (pont roulant) *On se propose ici de résoudre le problème suivant : comment déplacer une charge de masse m avec un pont roulant. Il s'agit de prendre la charge au repos à $t = 0$, de la déplacer et de la remettre au repos à $t = T$. Nous supposons, pour simplifier les calculs, le système dans un plan vertical fixe. La généralisation à une dimension horizontale supplémentaire ne pose pas de problème et peut être un excellent exercice. Pour effectuer cette manoeuvre, on dispose de deux commandes de haut niveau, la vitesse horizontale de déplacement du pont $\dot{D} = v_D$ et la vitesse d'enroulement du câble $\dot{R} = v_R$. En effet, des régulateurs (PI) de bas niveau dont la dynamique est rapide, assure la transformation des commandes de haut niveau (v_D,v_R), "lentement variables" pour les régulateurs, en efforts physiques développés par les moteurs électriques du pont.*

On note $0x$ l'axe horizontal sur lequel se déplace le chariot et Oz l'axe vertical descendant. g est l'accélération de la pesanteur. On suppose la charge m ponctuelle et le câble de masse négligeable. On note θ l'angle du câble par rapport à la verticale.

1. *Montrer que le système obéit à l'équation du second ordre suivante (Lagrange) :*

$$R\ddot{\theta} + 2\dot{R}\dot{\theta} + \ddot{D}\cos\theta + g\sin\theta = 0.$$

(indication : énergie cinétique de la masse $T = m/2(\dot{D}^2 + \dot{R}^2 + R^2\dot{\theta}^2 + 2R\dot{D}\dot{\theta}\cos\theta + 2\dot{D}\dot{R}\sin\theta)$; énergie potentielle $U = -mgR\cos\theta$). Montrer qu'en posant $p = R\dot{\theta} + \dot{D}\cos\theta$, on obtient les équations suivantes (Hamilton) :

$$\begin{aligned} \dot{\theta} &= (p - \dot{D}\cos\theta)/R \\ \dot{p} &= -\dot{R}\dot{\theta} - \dot{D}\dot{\theta}\sin\theta - g\sin\theta. \end{aligned}$$

Donner la forme d'état du linéaire tangent autour d'un point d'équilibre $\theta = 0$ et $R = \bar{R} > 0$.

2. *Montrer que ce système est commandable et donner sa forme de Brunovsky (la sortie de Brunovsky correspond aux deux coordonnées cartésiennes de la masse m).*
3. *A $t = 0$ la charge est au repos en $R = \bar{R}$ et $D = 0$. On désire la déplacer pendant le temps T à la position d'équilibre en $R = \bar{R}$ et $D = L$. Construire une trajectoire du linéaire tangent et une commande en boucle ouverte qui réalise ce déplacement. Que se passe-t-il si l'on utilise directement cette commande en boucle ouverte. Donner les équations du bouclage linéaire d'état qui stabilise le système autour de cette trajectoire.*
4. *On suppose que toutes les composantes de l'état ne sont pas disponibles.*
   (a) *On ne mesure que D et R. Le système est-il observable au premier ordre. Si l'on mesure en plus la vitesse angulaire $\dot{\theta}$, le système est-il observable ?*
   (b) *Nous supposons maintenant qu'on mesure (D,R,θ), la configuration du système. Montrer l'observabilité. Donner les équations de l'observateur réduit qui reconstruit asymptotiquement l'impulsion généralisée, p, et donc la vitesse angulaire $\dot{\theta}$.*
   (c) *Ecrire les équations de l'observateur-contrôleur qui assure le suivi asymptotique de la trajectoire de référence (question 3). Qualitativement, comment doit-on choisir le temps de transport T et les pôles du système bouclé de façon à obtenir une commande réaliste?*

5. *Simuler la commande modale précédente avec comme modèle de simulation les équations non linéaires sous forme de Hamilton. Tester en simulation, le domaine d'attraction et montrer l'intérêt d'une telle stratégie par rapport à de simples lois horaires sur D, allant de 0 à L, et ne prenant pas en compte les oscillations de la charge m générées par le déplacement.*

Chapitre 5

Annexe: Systèmes semi-implicites et inversion

Commençons par un exemple : le pendule sous forme semi-implicite. Une autre façon de représenter la dynamique du pendule (3.3) (figure 3.4, page 45) est d'écrire directement les lois de Newton en faisant intervenir les coordonnées cartésiennes du pendule (x,z) ainsi que la tension du fil $T = (T_x,T_z)$. On obtient alors le système (m est la masse du pendule) :

$$\begin{aligned} \frac{d^2x}{dt^2} &= T_x/m \\ \frac{d^2z}{dt^2} &= T_z/m - g \\ \frac{T_x}{x} &= \frac{T_z}{z} \\ x^2 + z^2 &= l^2. \end{aligned}$$

La troisième équation dit que T est co-linéaire à la direction du pendule, la quatrième que le pendule est le longueur constante l. Il est facile de mettre le système sous forme du premier ordre en rajoutant la vitesse (v_x,v_z) du pendule. On obtient un système déterminé avec (x,v_x,z,v_z) comme inconnues différentielles et (T_x,T_z) comme inconnues algébriques :

$$\begin{cases} \dfrac{dx}{dt} = v_x \\ \dfrac{dv_x}{dt} = T_x/m \\ \dfrac{dz}{dt} = v_z \\ \dfrac{dv_z}{dt} = T_z/m - g \\ \dfrac{T_x}{x} = \dfrac{T_z}{z} \\ x^2 + z^2 = l^2. \end{cases} \tag{5.1}$$

Très souvent les systèmes issus de la modélisation sont naturellement sous cette forme. Les mettre sous forme explicite nécessite alors des calculs compliqués et des changements de variables difficiles à manipuler. Pour un système mécanique élémentaire comme le pendule, les calculs sont assez simples. Pour des systèmes mécaniques comportant plusieurs corps, les calculs deviennent très vite inextricables.

Dans ce chapitre nous présentons les premiers résultats nécessaires à l'étude des systèmes différentiels dits implicites et comportant autant d'équations que d'inconnues. Ils se présentent sous deux formes.

La forme semi-implicite : on parle parfois de systèmes algébro-différentiels

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

où le vecteur des inconnues se décompose en deux, x les inconnues dites différentielles au nombre de n et u les inconnues dites algébriques au nombre de m ($m = \dim u = \dim h$).

La forme implicite :

$$f(x,\dot{x}) = 0$$

où toute les dérivées de x apparaissent implicitement dans les équations et où la matrice jacobienne $D_{\dot{x}} f$ est toujours singulière (de déterminant identiquement nul).

Ces systèmes sont caractérisés par leur index, un entier positif. L'index correspond au nombre de dérivations nécessaires pour écrire le système sous forme explicite. Son calcul repose sur l'algorithme de structure, algorithme qui fournit aussi les contraintes algébriques supplémentaires que doivent satisfaire les conditions initiales pour l'existence et l'unicité du problème de Cauchy. Ainsi nous verrons qu'un système différentiel implicite est en fait (en dehors des singularités) un système différentiel explicite de plus petite taille.

Nous ne traiterons pas la forme implicite générale. Nous nous contenterons d'étudier les systèmes semi-implicites. En effet $f(x,\dot{x}) = 0$ peut être vu comme une système semi-implicite de dimension double par le prolongement suivant :

$$\begin{aligned} \dot{x} &= u \\ 0 &= f(x,u). \end{aligned}$$

Noter que les systèmes explicites sont alors ceux pour lesquelles la partie algébrique $0 = h(x,u)$ n'existe pas : ce sont les équations différentielles explicites, objet du chapitre 3.

Exercice 36 *Quelles sont les relations entre (θ,ω) de (3.3) et (x,z,T_x,T_y) de (5.1) ?*

Résoudre

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

ou inverser le système dynamique

$$\left\{ \begin{array}{rcl} \dfrac{dx}{dt} & = & f(x,u) \\ y & = & h(x,u). \end{array} \right.$$

en imposant aux sorties y d'être nulles à chaque instant, revient exactement au même. Ce n'est qu'une question de vocabulaire. Les variables u sont interprétées comme des commandes, les variables $y = h(x,u)$ comme des sorties, les variables x comme l'état, la fonction $f(x,u)$ comme la dynamique en boucle ouverte, la fonction $h(x,u)$ comme la fonction de sortie. Le problème s'énonce ainsi : connaissant la loi horaire des sorties, calculer la loi horaire des commandes, $u(t)$ pour $t \geq 0$, sachant qu'elles agissent sur les sorties $h(x,u)$ par l'intermédiaire de l'équation différentielle $\dot{x} = f(x,u)$. Autrement dit, connaissant les sorties, calculer les entrées : ce problème d'inversion identique à la résolution des systèmes semi-implicites est en fait très proche du découplage et de la linéarisation entrée/sortie : tout repose encore sur l'algorithme de structure.

Ces questions feront l'objet du reste du chapitre avec la construction du bouclage dynamique qui découple[1] et linéarise la relation entre y et u .

La présentation ne faire appel qu'à un nombre limité d'outils mathématiques. De plus il suffit de traiter un exemple pour comprendre l'essentiel. Aussi nous conseillons le lecteur d'apporter toute son attention aux exemples traités dans les deux sous-sections 5.1.1 et 5.2.1. Le cas général sera alors très facile à comprendre ensuite.

5.1 Systèmes semi-implicites

Nous abordons ici l'existence et l'unicité des solutions pour un système semi-implicite,

$$\begin{cases} \dfrac{dx}{dt} = f(x,u), \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \\ 0 = h(x,u), \end{cases} \tag{5.2}$$

avec $f = (f_1, \ldots, f_n)$ et $h = (h_1, \ldots, h_m)$ fonctions régulières. Si la condition initiale (x^0,u^0) ne vérifie que $h(x^0,u^0) = 0$, la solution n'existe pas en générale, même si le système est correctement posé. La condition initiale doit vérifier d'autres équations algébriques, indépendantes de h et que l'on obtient en dérivant $\alpha - 1$ fois les équations, α étant l'index du système.

Nous présentons sur un exemple comment obtenir ces équations supplémentaires. Les calculs reposent sur un algorithme d'élimination différentielle, dit algorithme de structure. Cet algorithme est présenté en toute généralité après l'exemple.

5.1.1 Un exemple

Soit le système semi-implicite suivant :

$$\begin{cases} \dot{x}_1 &= x_1 + 2x_1u_1u_2 \\ \dot{x}_2 &= x_3 + x_1u_1u_2 \\ \dot{x}_3 &= x_3 + x_4 + x_3u_2 \\ \dot{x}_4 &= x_4 + u_1u_2 \\ 0 &= 1 + x_1 + x_1u_1u_2 \\ 0 &= x_2 + x_1u_1u_2. \end{cases} \tag{5.3}$$

On note $x = (x_1,x_2,x_3,x_4)$ et $u = (u_1,u_2)$. Nous nous posons la question suivante (problème de Cauchy). Soit (x^0,u^0) vérifiant les deux équations algébriques. Existe-t-il une solution de ce système semi-implicite ayant comme condition initiale (x^0,u^0). Nous allons voir que la réponse est non si (x^0,u^0) ne vérifie pas d'autres conditions, conditions dites de compatibilité et obtenues par l'algorithme de structure qui suit.

L'algorithme de structure

Etape 0 Il est clair que nous ne pouvons pas calculer u en fonction de x à partir de

$$\begin{cases} 0 &= 1 + x_1 + x_1u_1u_2 \\ 0 &= x_2 + x_1u_1u_2. \end{cases}$$

1. Ici, découpler signifier diagonaliser.

En effet, le rang de ce système par rapport à u est 1. Donc nécessairement, il contient implicitement une équation qui ne dépend que de x. Pour l'obtenir, il suffit ici de faire la différence entre les deux équations. On obtient alors le système,

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1 + x_1 u_1 u_2 \\ 0 & = & 1 + x_1 - x_2, \end{array} \right.$$

équivalent algébriquement au système de départ et qui se décompose en deux parties : une première partie (ici la première équation) dont la dépendance par rapport à u_1 et u_2 est maximum ; une seconde partie (ici la seconde équation) qui ne dépend que de x. Le nom d'élimination donné à cette méthode s'explique alors clairement. En effet, elle consiste à réécrire, de façon algébriquement équivalente, le système en éliminant au maximum la présence de u dans les équations.

Etape 1 On peut maintenant continuer en dérivant par rapport au temps la seconde équation [2]. En utilisant les équations relatives à $\dot{x}$, on obtient ainsi un nouveau système,

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1 + x_1 u_1 u_2 \\ 0 & = & x_1 - x_3 + x_1 u_1 u_2, \end{array} \right.$$

algébriquement indépendant du précédent. Son rang par rapport à u est toujours égal à 1. Par soustraction, on obtient le système,

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1 + x_1 u_1 u_2 \\ 0 & = & x_3 + 1, \end{array} \right.$$

algébriquement équivalent et en deux parties comme à l'étape précédente.

Etape 2 On dérive par rapport au temps la seconde équation et on obtient le système

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1 + x_1 u_1 u_2 \\ 0 & = & x_3 + x_4 + x_3 u_2. \end{array} \right.$$

Son rang par rapport à u est égal 2. Par inversion de ce système algébrique, nous obtenons u en fonction de x :

$$\left\{ \begin{array}{rcl} u_1 u_2 & = & -\dfrac{1 + x_1}{x_1} \\ u_2 & = & -\dfrac{x_3 + x_4}{x_3}. \end{array} \right.$$

Index et problème de Cauchy

En remplaçant u par sa valeur dans les équations donnant $\dot{x}$, on obtient

$$\left\{ \begin{array}{rcl} \dot{x}_1 & = & x_1 - 2(1 + x_1) \\ \dot{x}_2 & = & x_3 - (1 + x_1) \\ \dot{x}_3 & = & 0 \\ \dot{x}_4 & = & x_4 - \dfrac{1 + x_1}{x_1}. \end{array} \right.$$

2. Si nous avions directement dérivé l'une des deux équations de

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1 + x_1 u_1 u_2 \\ 0 & = & x_2 + x_1 u_1 u_2, \end{array} \right.$$

nous aurions obtenu des termes en $\dot{u}$ dont nous n'aurions eu que faire.

C'est un système différentiel ordinaire qui admet, localement au moins, une solution unique si l'on fixe la condition initiale x^0. Supposons que x^0 vérifie les deux équations ne dépendant que de x et obtenues lors des deux étapes 0 et 1 :

$$\left\{ \begin{array}{rcl} 1 + x_1^0 - x_2^0 & = & 0 \\ 1 + x_3^0 & = & 0. \end{array} \right.$$

Puisque $\dot{x}_3 = 0$, on a $x_3 = -1$ à chaque instant. Il est alors immédiat de voir que $\dot{x}_1 - \dot{x}_2 = 0$ et donc que $1 + x_1 - x_2 = 0$ à chaque instant.

Nous avons en fait montré que, pour qu'il existe une solution au système semi-implicite de départ ayant comme condition initiale x^0 et u^0 vérifiant les équations algébriques,

$$\left\{ \begin{array}{rcl} 0 & = & 1 + x_1^0 + x_1^0 u_1^0 u_2^0 \\ 0 & = & x_2^0 + x_1^0 u_1^0 u_2^0, \end{array} \right.$$

il faut et il suffit qu'en plus la condition initiale x^0 et u^0 vérifie deux autres équations, algébriquement indépendantes des deux premières, qui sont obtenues au cours des étapes 1 et 2 :

$$\left\{ \begin{array}{rcl} u_2^0 & = & -\dfrac{x_3^0 + x_4^0}{x_3^0} \\ 1 + x_3^0 & = & 0. \end{array} \right.$$

Le nombre de dérivation nécessaires pour obtenir ces conditions algébriques supplémentaires est ici 2. On dit alors que l'index est de 3, car une dérivation supplémentaire permet de calculer $\dot{u}$ en fonction de x et u et donc de mettre le système sous forme explicite on parle aussi de *forme involutive* ou *formellement intégrable*.

Remarquons enfin qu'intégrer ce système semi-implicite revient en fait à intégrer un système différentiel de taille inférieure à x, les variables différentielles. En effet, au cours des calculs précédents, nous avons obtenu deux types d'équations algébriques : les équations qui fournissent u en fonction de x et les équations ne portant que sur x

$$1 + x_1 + x_2 = 0, \quad 1 + x_3 = 0.$$

Aussi il suffit de connaître, par exemple, x_1 et x_4 pour en déduire les 4 autres variables à partir des 4 équations algébriques dont nous disposons. Mais (x_1,x_4) sont les solutions du système autonome

$$\frac{dx_1}{dt} = -x_1 - 2, \quad \frac{dx_4}{dt} = x_4 - \frac{1 - x_1}{x_1},$$

obtenues en remplaçant $u_1 u_2$ par $-(1 + x_1)/x_1$. La résolution du système de départ se ramène à celle de ce système explicite de dimension deux avec deux conditions initiales indépendantes.

Un système semi-implicite peut donc être vu comme un système différentiel explicite de taille inférieure. Les systèmes différentiels implicites sont en fait des systèmes différentiels explicites sur une sous-variété. Les équations de la sous-variété étant celles issues de l'algorithme de structure.

5.1.2 Le cas général

Nous revenons maintenant au système semi-implicite général (5.2).

L'algorithme de structure

Nous restons à un niveau structurel. Une présentation mathématiquement rigoureuse est possible en utilisant soit de la géométrie différentielle et la théorie des jets, soit l'algèbre différentielle. Le principal raccourci de cette présentation consiste à supposer, pour tout système dit algébrique $h(x,u) = 0$, que le rang de h par rapport à u, i.e., le rang μ de la matrice carré

$$\left(\frac{\partial h_i}{\partial u_j}\right) \begin{array}{l} 1 \leq i \leq m \\ 1 \leq j \leq m \end{array},$$

est constant et inférieur ou égal à m. Aussi les composantes de h se décompose en deux parties (sous des hypothèses convenables) :

– la première partie, notée $\overline{h}$, regroupe μ composantes de h; elle est telle que le rang de

$$\left(\frac{\partial \overline{h}_i}{\partial u_j}\right) \begin{array}{l} 1 \leq i \leq \mu \\ 1 \leq j \leq m \end{array}$$

soit μ

– la seconde partie, notée $\widetilde{h}$, regroupe les $m - \mu$ composantes restantes; $\widetilde{h}$ peut alors s'exprimer comme une fonction de x et de $\overline{h}$:

$$\widetilde{h}(x,u) = \Phi(x,\overline{h}(x,u)).$$

Une telle décomposition revient à éliminer u de $h(x,u) = 0$ pour obtenir $\Phi(x,0) = 0$, un système de plus petite taille $m - \mu$ ou seul x apparaît.

Exercice 37 *Donner des hypothèses qui assurent l'existence d'une telle décomposition de h en $\overline{h}$ et $\widetilde{h}$. Cette décomposition est-elle unique ?*

À chaque étape de l'algorithme, nous supposerons implicitement que les manipulations précédentes s'appliquent. Nous nous intéressons au cas générique. Les problèmes de singularité sont des problèmes difficiles qui relèvent de considérations topologiques et que nous ne voulons pas aborder ici.

On note $h_0(x,u)$ la fonction $h(x,u)$ du système (5.2). On définit par récurrence les fonctions $h_1(x,u)$, $h_2(x,u)$, ..., $h_k(x,u)$ à valeurs dans $\mathbb{R}^m$ comme suit.

Soit $k \geq 0$. Supposons définie h_k, fonction de x et u à valeurs dans $\mathbb{R}^m$. Soit μ_k le rang de h_k par rapport à u, i.e. le rang de la matrice

$$\frac{\partial h_k}{\partial u}.$$

Quitte à permuter les lignes de h_k, on peut supposer que ses μ_k premières lignes $\overline{h}_k = (h_k^1, \ldots, h_k^{\mu_k})$ sont telles que le rang de

$$\frac{\partial \overline{h}_k}{\partial u}$$

est maximum et égal à μ_k. Ainsi les $m - \mu_k$ dernières lignes de h_k, $\tilde{h}_k = (h_k^{\mu_k+1}, \ldots, h_k^m)$ ne dépendent de u que par l'intermédiaire de $\overline{h}_k$: il existe donc une fonction $\Phi_k(x,\cdot)$ telle que

$$\tilde{h}_k(x,u) \quad = \quad \Phi_k(x,\overline{h}_k(x,u)).$$

On définit h_{k+1} fonction de x et u à valeurs dans $\mathbb{R}^m$ par[3]

$$h_{k+1}(x,u) = \begin{pmatrix} \overline{h}_k(x,u) \\ \\ \dfrac{d}{dt}[\Phi_k(x,0)] = \left(\dfrac{\partial \Phi_k}{\partial x}\right)_{(x,0)} f(x,u) \end{pmatrix}.$$

A l'étape $k+1$, les μ_k premières composantes de $\overline{h}_{k+1}$ sont choisies de façon à former exactement les μ_k composantes du vecteur $\overline{h}_k$.

Index et problème de Cauchy

La suite μ_k est une suite croissante d'entiers inférieurs à m. Elle stationne donc à partir d'un certain rang. Il est intuitif, mais pas évident de démontrer sans faire appel à des outils mathématiques plus généraux, que la suite des entiers μ_k est en fait indépendante du choix des coordonnées sur x et du choix des commandes u : si $x = \Xi(\xi)$ et $u = V(\xi,v)$ sont des changements de variables sur x (Ξ est un difféomorphisme) et sur u ($V(\xi,\cdot)$ est un difféomorphisme), alors l'algorithme précédent donne la même suite μ_k pour le système (5.2) écrit avec ces nouvelles variables :

$$\left\{ \begin{array}{rcl} \dfrac{d\xi}{dt} & = & \left[\dfrac{\partial \Xi}{\partial \xi}(\xi))\right]^{-1} f(\Xi(\xi),V(\xi,v)) \\ 0 & = & h(\Xi(\xi),V(\xi,v)). \end{array} \right.$$

Définition 22 *Si la suite μ_k stationne à m alors, l'index du système semi-implicite est $\alpha+1$ où α est le plus petit entier k tel que $\mu_k = m$. Si la suite μ_k stationne à une valeur strictement inférieure à m alors l'index est infini.*

On peut démontrer le résultat suivant :

Lemme 2 *Si l'index $\alpha+1$ du système (5.2) est fini, alors $\alpha \leq n$ et le rang du jacobien*

$$\frac{\partial}{\partial x} \begin{pmatrix} \Phi_0(x,0) \\ \vdots \\ \Phi_{\alpha-1}(x,0) \end{pmatrix}$$

est égal au nombre de ses lignes : $\sum_{k=0}^{\alpha}(m-\mu_k)$.

Ce résultat implique donc que l'algorithme de structure comporte au plus n étapes. Ainsi pour mettre un système semi-implicite sous forme explicite il suffit de dériver au plus $n+1$ fois, n étant le nombre de variables différentielles.

La démonstration de ce résultat n'est qu'une mise en forme des deux remarques suivantes.

– Supposons que le passage de l'étape k à l'étape $k+1$ génère de nouvelles équations entre x et u : l'index étant fini, ces nouvelles équations sont alors nécessairement indépendantes de celles obtenues aux étapes précédentes; ainsi le rang des équations ne faisant intervenir que x,

$$\Phi_0(x,0) = 0, \quad \ldots, \quad \Phi_k(x,0) = 0,$$

est maximum, i.e., égal aux nombres d'équations $\sum_{i=0}^{k+1}(m-\mu_i)$.

3. En fait, $\left(\dfrac{\partial \Phi_k}{\partial x}\right)_{(x,0)} f(x,u)$ est égal à $\dfrac{d}{dt}\tilde{h}_k(x,u)$, car $\overline{h}_k$ est nul à chaque instant.

– Supposons que le passage de l'étape k à l'étape $k+1$ ne génère plus aucune équation nouvelle : alors l'étape $k+2$ sera identique à l'étape $k+1$; nous dériverons les mêmes équations, celles obtenues à l'étape k; il est inutile de dériver davantage; l'algorithme s'arrête car il ne fournit plus de nouvelle équation.

Nous avons ainsi le résultat suivant.

Théorème 16 *Supposons que le système semi-implicite (5.2) soit d'index fini $\alpha+1$. Prenons (x^0,u^0) qui vérifie $h_\alpha(x^0,u^0)=0$ et $\Phi_k(x^0,0)=0$ pour $k=1,\ldots,\alpha-1$. Alors il existe une unique solution de (5.2) ayant comme condition initiale (x^0,u^0).*

Le fait que la condition initiale vérifie $h(x^0,u^0)=0$ n'est pas suffisant pour assurer l'existence de la solution dès que l'index du système excède 1. En effet, dès l'index 2, des conditions supplémentaires et algébriquement indépendantes de h apparaissent.

Si l'index est infini, i.e. si les μ_k stationnent à une valeur $\mu_\infty < m$, il n'est pas possible de calculer la dérivée de u. Dans ce cas, les équations de départ sont liées entre elles. Elles ne sont pas *différentiellement algébriquement indépendantes* : la partie algébrique $h(x,u)=0$ ne comporte en fait que μ_∞ équations différentiellement algébriquement indépendantes. Alors pour une condition initiale fixée, le système admet soit aucune solution, soit une infinité si la condition initiale vérifie toutes les équations algébriques issue de l'algorithme de structure. En pratique, un index infini indique une modélisation incomplète.

Exercice 38 *Quel est l'index du système semi-implicite (5.1) ? Qu'elles sont les contraintes que doit satisfaire la condition initiale pour assurer l'existence de la solution.*

Les systèmes implicites apparaissent aussi pour les équations aux dérivées partielles mais alors les calculs et la théorie sont bien plus compliqués. Prenons cependant un exemple qui indique que la méthode reste cependant la même : enchaîner alternativement des dérivations et des éliminations. Prenons les équations d'Euler des fluides parfaits dans une cavité Ω (n normale extérieure) :

$$\begin{aligned} \frac{\partial V}{\partial t}+\frac{\partial V}{\partial x}V &= -\frac{\partial p}{\partial x} \\ \text{div } V &= 0 \\ V\cdot n &= 0 \text{ sur le bord } \partial\Omega. \end{aligned}$$

Par analogie avec le pendule (5.1) nous voyons que le champ de vitesse V correspond à la position (x,z) et que la pression p joue le rôle de la tension T. Ainsi les deux contraintes, div$V=0$ à l'intérieur et $V\cdot n=0$ sur le bord, correspondent aux deux équations algébriques de (5.1). Pour calculer p, on dérive donc ces deux contraintes par rapport à t :

$$0=\frac{\partial(\text{div } V)}{\partial t}=\text{div } \frac{\partial V}{\partial t}=-\Delta p-\text{div } \left(\frac{\partial V}{\partial x}V\right)$$

et

$$0=\frac{\partial V}{\partial t}\cdot n=-\frac{\partial p}{\partial n}-\left(\frac{\partial V}{\partial x}V\right)\cdot n.$$

Ainsi p dépend de V d'une façon "statique" (comme la tension T du pendule est une fonction de la position et de la vitesse du pendule). La pression est obtenue, à une constante près, en

fonction de V en résolvant le problème de Neuman suivant [14] :

$$\Delta p = -\sum_{i,j=1}^{3} \frac{\partial V_i}{\partial x_j}\frac{\partial V_j}{\partial x_i}$$
$$\frac{\partial p}{\partial n} = -\sum_{i,j=1}^{3} \frac{\partial V_i}{\partial x_j} V_j n_j \quad \text{sur } \partial\Omega.$$

5.1.3 Linéaire tangent

La méthode d'élimination ci-dessus est générale. Elle peut être mise en oeuvre (au moins formellement) pour n'importe quel système. Cependant, la difficulté essentielle n'est pas ici. Elle réside dans le fait que les calculs sont inextricables pour les systèmes complexes étudiés par les ingénieurs. Souvent deux à trois dérivations sont nécessaires pour rendre le système explicite : en pratique l'index est souvent plus petit que 3. Nous montrons ici comment calculer les exposants caractéristiques du linéarisé-tangent autour d'un point d'équilibre directement à partir de la forme semi-implicite sans passer par la forme explicite qui est de taille réduite.

Un point d'équilibre $(\bar{x},\bar{u})$ de (5.2) est caractérisé par

$$0 = f(\bar{x},\bar{u}), \quad 0 = h(\bar{x},\bar{u})$$

Le système tangent est alors obtenu en ne conservant que les termes d'ordre 1 :

$$\frac{d(\Delta x)}{dt} = \frac{\partial f}{\partial x}_{(\bar{x},\bar{u})} \Delta x + \frac{\partial f}{\partial u}_{(\bar{x},\bar{u})} \Delta u$$
$$0 = \frac{\partial h}{\partial x}_{(\bar{x},\bar{u})} \Delta x + \frac{\partial h}{\partial u}_{(\bar{x},\bar{u})} \Delta u$$

avec $\Delta x = x - \bar{x}$ et $\Delta u = u - \bar{u}$.

En supposant que le point d'équilibre est un point générique du système (absence de singularité...), nous pouvons avoir la dimension du système différentielle explicite de taille inférieur ainsi que les valeurs propres de son linéaire tangent directement sur le tangent implicite. Notons

$$E = \begin{pmatrix} 1_n & 0 \\ 0 & 0 \end{pmatrix}$$

et

$$A = \begin{pmatrix} \frac{\partial f}{\partial x}(\bar{x},\bar{u}) & \frac{\partial f}{\partial u}(\bar{x},\bar{u}) \\ \frac{\partial h}{\partial x}(\bar{x},\bar{u}) & \frac{\partial h}{\partial u}(\bar{x},\bar{u}) \end{pmatrix}.$$

Les exposants caractéristiques sont alors donnés par les solutions λ de l'équation polynômiale suivante

$$\det(\lambda E - A) = 0.$$

le degré de ce polynôme étant alors la dimension de la dynamique du système. Il s'agit d'une généralisation naturelle du cas explicite où E est la matrice identité. La démonstration de ce résultat repose sur la théorie des faisceaux de matrices. Nous renvoyons le lecteur intéressé à [22].

Exercice 39 *Calculer le point d'équilibre de (5.3) et écrire son linéarisé tangent. Vérifier que, pour ce système, le degré de* $\det(\lambda E - A) = 0$ *est bien égal à deux. Calculer les deux racines de ce polynôme en* λ*. Vérifier que l'on obtient bien les mêmes valeurs qu'en calculant le linéaire tangent sur la forme explicite réduite issue de l'algorithme de structure.*

5.1.4 Résolution numérique

Commençons par la résolution numérique des systèmes d'index 1, i.e. ceux pour lesquels la partie algébrique $0 = h(x,u)$ est de rang maximum en u et fournit donc u en fonction de x par inversion locale. Le premier schéma de discrétisation qui vient à l'esprit est le suivant :

$$\frac{x^{n+1} - x^n}{\Delta t} = f(x^n,u^n), \quad 0 = h(x^{n+1},u^{n+1})$$

où (x^n,u^n) serait une approximation de (x,u) à l'instant $n\Delta t$. Ce schéma est déjà implicite en u.Connaissant (x^n,u^n), il faut, pour calculer u^{n+1} résoudre $h = 0$. Un tel schéma correspond en fait au schéma d'Euler explicite. Son ordre est 1. Il est convergent dès que le pas de discrétisation Δt est choisi plus petit que les constantes de temps les plus rapides du système.

Un tel schéma ne peut pas convenir pour des systèmes d'index ≥ 2. En effet ll n'est plus possible de calculer u^{n+1} car h n'est plus inversible par rapport à u. D'une façon plus général, les méthodes de Gear [20] sont bien adaptées à la résolution des systèmes d'index 1. Pour des index supérieurs une adaptation du schéma est nécessaire.

Une façon de contourner le problème est de résoudre un système différentiel d'index 1 ou 0 dont on sait, par des considérations physiques de modélisation, que les solutions sont proches de celles du système de départ et d'index > 1. Très souvent un index > 1 résulte de dynamiques rapides, stables et négligées. Un bon sens physique permet de rajouter ces petites dynamiques. Cette façon de procéder admet une justification dans le cadre de la théorie des perturbations et les systèmes lents/rapides (c.f. section 3.5).

Nous traitons à titre d'exemple le pendule (5.1). Supposons que la barre qui soutient la masse soit légèrement élastique de raideur $1/\varepsilon$, $\varepsilon > 0$ (la tension dans la barre est alors $(\sqrt{x^2 + z^2} - l)/\varepsilon$). On peut également prendre en compte l'amortissement des vibrations hautes fréquences en rajoutant un terme opposé à la vitesse d'élongation dans le calcul de la tension de la barre $(x\dot{x} + z\dot{z})/\eta$, $\eta > 0$). Alors, les équations du pendule (5.1) deviennent :

$$\begin{cases} \dfrac{dx}{dt} = v_x \\ \dfrac{dv_x}{dt} = T_x/m \\ \dfrac{dz}{dt} = v_z \\ \dfrac{dv_z}{dt} = T_z/m - g \\ \dfrac{T_x}{x} = \dfrac{T_z}{z} \\ xT_x + zT_z = -\dfrac{\sqrt{x^2 + z^2} - l}{\varepsilon} - \dfrac{xv_x + zv_z}{\eta}. \end{cases}$$

Ce système est d'index 1. Ses trajectoires sont proches de celles de (5.1) si l'on choisit ε et η petits.

Exercice 40 *Étendre ce qui précède au mouvement d'une masse ponctuelle (x,y,z) dans l'espace à trois dimensions. Cette masse est soumise à un champ de force dérivant d'un potentiel $V(x,y,z)$. Elle glisse sans frotter sur une surface d'équation $h(x,y,z) = 0$.*

5.2 Inversion et découplage

Nous considérons le système suivant

$$\begin{cases} \dfrac{dx}{dt} = f(x,u) \\ \quad y = h(x,u) \end{cases} \tag{5.4}$$

avec l'état $x \in \mathbb{R}^n$, les commandes $u \in \mathbb{R}^m$, les sorties $y \in \mathbb{R}^m$. Les fonctions f et h sont supposées régulières. Nous présentons la démarche à suivre pour calculer un bouclage *quasi-statique* qui linéarise la relation entrée/sortie lorsque le système est inversible (i.e., lorsque c'est possible). On appelle *dynamique des zéros* le système semi-implicite issu de (5.4) en bloquant y à une valeur fixe. La stabilité locale du système bouclé est alors conditionnée par la stabilité locale de cette dynamique des zéros. Plus précisément, si elle est hyperboliquement stable, alors il est possible de stabiliser localement le système avec un bouclage linéarisant la relation entre y et u. Si la dynamique des zéros est instable (son linéaire tangent admet un pôle à partie réelle positive), alors un bouclage fondé sur la linéarisation entree/sortie déstabilise le système et n'a que peu d'intérêt. En linéaire, les systèmes qui admettent une dynamique des zéros asymptotiquement stable sont dits à *déphasage minimal*.

5.2.1 Un exemple

Soit le système

$$\begin{cases} \dfrac{dx_1}{dt} &= x_1x_2 + u_1 \\ \dfrac{dx_2}{dt} &= x_1x_2 + x_3 + u_1 \\ \dfrac{dx_3}{dt} &= x_3 + x_4 + u_2 \\ \dfrac{dx_4}{dt} &= x_3x_4 + \lambda x_4 + u_2 \\ y_1 &= x_1 \\ y_2 &= x_2 \end{cases} \tag{5.5}$$

avec λ un paramètre. On note $x = (x_1,x_2,x_3,x_4)$, $u = (u_1,u_2)$ et $y = (y_1,y_2)$. Nous voulons que y suive avec stabilité une loi horaire $t \mapsto y^r(t)$, la référence de sortie définie par avance. Il s'agit d'un problème typique de poursuite de trajectoire ("output tracking" en anglais). Pour cela nous disposons de la mesure de l'état x à chaque instant (nous savons où est le système) et nous connaissons les équations du système (nous disposons d'un modèle). Comment ajuster en temps-réel la commande u de façon à ce que l'erreur de suivi $y - y^r$ converge vers 0.

Nous allons voir qu'un bouclage du type

$$u = k(x,y^r(t),\dot{y}^r(t),\ddot{y}^r(t))$$

répond à la question. D'autres réponses sont possibles avec des techniques différentes. Celle que nous présentons est en faite élémentaire et reprend l'algorithme de structure que nous avons déjà vu.

En dérivant une fois y, on a :

$$\begin{cases} \dot{y}_1 &= x_1x_2 + u_1 \\ \dot{y}_2 &= x_1x_2 + x_3 + u_1. \end{cases}$$

Ce système est de rang 1 par rapport à u. L'élimination de u donne l'équation

$$\dot{y}_2 = \dot{y}_1 + x_3$$

que l'on dérive par rapport au temps pour obtenir

$$\ddot{y}_2 = \ddot{y}_1 + x_3 x_4 + u_2.$$

Ainsi on a

$$\begin{aligned}\dot{y}_1 &= x_1 x_2 + u_1\\ \ddot{y}_2 &= \ddot{y}_1 + x_3 x_4 + u_2.\end{aligned}$$

Nous pouvons donc imposer une vitesse arbitraire v_1 à y_1 et une accélération arbitraire v_2 à y_2 en choisissant u_1 et u_2 solution de

$$\begin{aligned}v_1 &= x_1 x_2 + u_1\\ v_2 &= \dot{v}_1 + x_3 x_4 + u_2.\end{aligned}$$

Nous aurions pu tout aussi bien imposer l'accélération de y_1 et la vitesse de y_2. Différents choix sont possibles à ce stade. Prenons

$$v_1 = \dot{y}_1^r(t) - a(y_1 - y_1^r(t))$$

et

$$v_2 = \ddot{y}_2^r(t) - b(\dot{y}_2 - \dot{y}_2^r(t)) - c(y_2 - y_2^r(t))$$

avec a,b,$c > 0$ (paramètres de réglage, les gains du suivi). Alors l'erreur de suivi $\Delta y = y - y^r$ obéit à

$$\frac{d(\Delta y_1)}{dt} = -a\Delta y_1$$

et à

$$\frac{d^2(\Delta y_2)}{dt^2} = -b\frac{d(\Delta y_2)}{dt} - c\Delta y_2.$$

Il s'agit de deux équations différentielles linéaires, découplées et asymptotiquement stables car les gains a, b et c sont positifs. Ce qui explique la terminologie découplage et linéarisation entrée/sortie.

Voyons maintenant l'allure du bouclage. Comme $u_1 = v_1 - x_1 x_2$, et $y_1 = x_1$, on a

$$u_1 = \dot{y}_1^r(t) - a(x_1 - y_1^r(t)) - x_1 x_2.$$

On sait aussi que $u_2 = v_2 - \dot{v}_1 - x_3 x_4$. Calculons donc v_2 et $\dot{v}_1$ en fonction de x, de la référence y^r et ses dérivées. On a

$$v_2 = \ddot{y}_2^r(t) - b(\dot{y}_1^r(t) - a(x_1 - y_1^r(t)) + x_3 - \dot{y}_2^r(t)) - c(x_2 - y_2^r(t))$$

puisque $y_2 = x_2$ et $\dot{y}_2 = \dot{y}_1 + x_3 = v_1 + x_3 = \dot{y}_1^r(t) - a(x_1 - y_1^r(t)) + x_3$. On obtient $\dot{v}_1$ en dérivant

$$v_1 = \dot{y}_1^r(t) - a(y_1 - y_1^r(t))$$

par rapport au temps, soit

$$\dot{v}_1 = \ddot{y}_1^r(t) - a(v_1 - \dot{y}_1^r(t)) = \ddot{y}_1^r(t) - a^2(y_1 - y_1^r(t))$$

car $\dot{y}_1 = v_1 = \dot{y}_1^r(t) - a(y_1 - y_1^r(t))$.

Ainsi le bouclage en u_2 est donné par

$$u_2 = \ddot{y}_2^r(t) - b(\dot{y}_1^r(t) - a(x_1 - y_1^r(t)) + x_3 - \dot{y}_2^r(t)) - c(x_2 - y_2^r(t)) - \ddot{y}_1^r(t) + a^2(y_1 - y_1^r(t)) - x_3 x_4.$$

Ce type de bouclage appelé *bouclage quasi-statique* assure donc la stabilisation des sorties y vers leur référence y^r. Cela ne signifie pas que le système en entier est stable. Supposons que la référence y^r soit constamment nulle. Alors y converge vers 0. Ainsi les trajectoires du système sont à terme proches des trajectoires du système semi-implicite obtenu en annulant y :

$$\left\{\begin{array}{rcl} \frac{dx_1}{dt} & = & x_1x_2 + u_1 \\ \frac{dx_2}{dt} & = & x_1x_2 + x_3 + u_1 \\ \frac{dx_3}{dt} & = & x_3 + x_4 + u_2 \\ \frac{dx_4}{dt} & = & x_3x_4 + \lambda x_4 + u_2 \\ 0 & = & x_1 \\ 0 & = & x_2. \end{array}\right.$$

L'index de ce système vaut 3 et il est alors facile de voir que $x_1 = x_2 = x_3 = 0$, que $u_1 = 0$ et $u_2 = -x_4$. La dynamique explicite dite *dynamique des zéros* est ainsi de dimension 1 avec $\dot{x}_4 = (\lambda - 1)x_4$. Si $\lambda > 1$ cette dynamique n'est pas asymptotiquement stable. Si $\lambda < 1$, cette dynamique est asymptotiquement stable.

On peut aisément montrer que lorsque la dynamique de zéros est hyperboliquement stable une telle méthode de commande stabilise localement le système tout entier (on ne peut rien dire globalement à cause de phénomènes de "picking" même si la dynamique des zéros est globalement asymptotiquement stable).

Exercice 41 *Le modèle dynamique d'un réacteur batch est le suivant :*

$$\left\{\begin{array}{rcl} \frac{dC_A}{dt} & = & -k_1(T)C_A^2 \\ \frac{dC_B}{dt} & = & k_1(T)C_A^2 - k_2(T)C_B \\ \frac{dT}{dt} & = & \gamma_1 k_1(T)C_A^2 + \gamma_2 k_2(T)C_B + (a_1 + a_2T) + (b_1 + b_2T)u \\ y & = & T \end{array}\right. \tag{5.6}$$

avec C_A et C_B les concentrations de A et B, T la température, u la variable de commande (apport ou extraction de chaleur), $k_1(T)$ et $k_2(T)$ des fonctions positives de T, γ_1, γ_2, a_1, a_2, b_1 et b_2 des paramètres constants. Le but de la commande est de suivre un profil de température $[0,\Theta] \ni t \mapsto T^r(t)$ durant toute la durée du batch Θ.

1. *Calculer le bouclage qui linéarise la dynamique de l'erreur $\Delta T = T - T^r$*
2. *Discuter en fonction de k_1 et de k_2 la stabilité de la dynamique des zéros.*

Exercice 42 *Calculer pour la colonne à distiller du problème 3, page 81 (système (3.21)) le bouclage linéarisant avec $u = (L,V)$ et $y = (x_1,x_n)$. Que dire des calculs autour d'un point stationnaire $\bar{x}$? Savez-vous montrer que la dynamique des zéros est stable ?*

5.2.2 Le cas général

Revenons au système général (5.4).

Inversion

Nous reprenons ici les calculs de l'algorithme de structure avec y dépendant du temps. Ainsi la donnée est $t \mapsto y(t)$ supposée suffisamment dérivable par rapport au temps. Les inconnues sont x et surtout u. La dérivée ν-ième en temps d'une variable ξ est notée $\xi^{(\nu)}$, ceci afin d'alléger les calculs qui suivent.

Étape $k = 0$ Notons $h_0(x,u)$ la fonction $h(x,u)$ du système (5.4). Par définition, $y = h_0(x,u)$. Soit μ_0 le rang de

$$\frac{\partial h_0}{\partial u}.$$

Quitte à permuter les lignes de h_0 et donc les composantes de y, on peut supposer que les μ_0 premières lignes $\overline{h}_0 = \kappa_0 = (h_0^1, \ldots, h_0^{\mu_0})$ sont telles que le rang de

$$\frac{\partial \overline{h}_0}{\partial u}$$

est maximum et égal à μ_0. Notons $\tilde{h}_0 = (h_0^{\mu_0+1}, \ldots, h_0^m)$ les $m - \mu_0$ dernières lignes de h_0. Ainsi $\tilde{h}_0$ ne dépend de u que par l'intermédaire de $\overline{h}_0$. Il existe donc une fonction $\Phi_0(x,\cdot)$ telle que

$$\tilde{h}_0(x,u) \quad = \quad \Phi_0(x,\overline{h}_0(x,u)).$$

Il est clair que $y = h_0(x,u)$ est algébriquement équivalent à

$$\left\{ \begin{array}{rcl} y_0 & = & \overline{h}_0(x,u) = \kappa_0(x,u) \\ \tilde{y}_1 & = & \Phi_0(x,y_0) \end{array} \right.$$

où $y = (y_0,\tilde{y}_1)$ avec $y_0 = \overline{y}_0$, les μ_0 premières composantes de y, et $\tilde{y}_1$ rassemblant les $m - \mu_0$ dernières composantes de y.

Étape $k \geq 0$ Supposons définies

- la suite croissante d'entiers $\mu_0, \ldots, \mu_k$;
- une partition des composantes de y en deux groupes, $y = (\overline{y}_k,\tilde{y}_{k+1})$; $\overline{y}_k = (y_0, \ldots, y_k)$ est de dimension μ_k, chaque y_i étant de dimension $\mu_i - \mu_{i-1}$ [4] ; $\tilde{y}_k$ est de dimension $m - \mu_k$;
- la fonction h_k à valeurs dans $\mathbb{R}^m$, dépendant de

$$(x,u,y,\ldots,y^{(k)}),$$

de rang μ_k par rapport à u et dont les composantes se divisent en deux : $h_k = (\overline{h}_k,\tilde{h}_k)$; $\overline{h}_k = (\kappa_0, \ldots, \kappa_k)$ est de dimension μ_k, chaque κ_i étant de dimension $\mu_i - \mu_{i-1}$; le rang de $\overline{h}_k$ par rapport à u est égal à μ_k ; $\tilde{h}_k$ est de dimension $m - \mu_k$; h_k vérifie

$$\left\{ \begin{array}{rcl} y_0 & = & \kappa_0(x,u) \\ y_1^{(1)} & = & \kappa_1(x,u,y_0,y_0^{(1)}) \\ & \vdots & \\ y_k^{(k)} & = & \kappa_k\left(x,u,\left(y_i^{(i)},\ldots,y_i^{(k)}\right)_{i=0,\ldots,k-1}\right) \\ \tilde{y}_{k+1}^{(k)} & = & \Phi_k\left(x,\left(y_i^{(i)},\ldots,y_i^{(k)}\right)_{i=0,\ldots,k-1},(y_0,y_1^{(1)},\ldots,y_k^{(k)})\right). \end{array} \right.$$

4. Avec la convention $\mu_{-1} = 0$.

On définit alors h_{k+1} en remplaçant la dernière équation du système précédent par sa dérivée rapport au temps :

$$\tilde{y}_{k+1}^{(k+1)} = \frac{\partial \Phi_k}{\partial x} f(x,u) + \Upsilon_k \left(x, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k} \right)$$

avec

$$\frac{\partial \Phi_k}{\partial x}\, f + \Upsilon_k = \frac{d}{dt} \left[\Phi_k \left(x, \left(y_i^{(i)}, \ldots, y_i^{(k)} \right)_{i=0,\ldots,k-1}, (y_0, y_1^{(1)}, \ldots, y_k^{(k)}) \right) \right] .$$

On définit alors h_{k+1} par

$$h_{k+1} = (\kappa_0, \ldots, \kappa_k, \frac{d\Phi_k}{dt} f + \Upsilon_k).$$

h_{k+1} est une fonction de $(x,u,y,\ldots,y^{(k+1)})$. Son rang par rapport à u est par définition μ_{k+1}.

Par construction de h_{k+1}

- $\mu_{k+1} \geq \mu_k$;
- on peut poser, quitte à permuter des lignes, que

$$\frac{\partial \Phi_k}{\partial x} f + \Upsilon_k = (\kappa_{k+1}, \tilde{h}_{k+1})$$

où κ_{k+1} est une fonction de

$$(x,u, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k})$$

à valeurs dans $\mathbb{R}^{\mu_{k+1}-\mu_k}$, où $\tilde{h}_{k+1}$ est aussi une fonction des mêmes variables mais à valeurs dans $\mathbb{R}^{m-\mu_{k+1}}$, et où le rang de $\overline{h}_{k+1} = (\overline{h}_k, \kappa_{k+1})$ par rapport à u est égal à μ_{k+1} ;

- $\tilde{y}_{k+1}$ se décompose comme $\tilde{h}_{k+1}$ en deux parties, $\tilde{y}_{k+1} = (y_{k+1}, \tilde{y}_{k+2})$ avec y_{k+1} de dimension $\mu_{k+1} - \mu_k$, $\tilde{y}_{k+2}$ de dimension $m - \mu_{k+1}$; on pose $y = (\overline{y}_{k+1}, \tilde{y}_{k+2})$ avec $\overline{y}_{k+1} = (\overline{y}_k, y_{k+1})$ de dimension μ_{k+1}.

Comme μ_{k+1} est le rang de $h_k = (\overline{h}_{k+1}, \tilde{h}_{k+1})$ et de $\overline{h}_{k+1}$ par rapport à u, il est clair que $\tilde{h}_{k+1}$ ne dépend de u que par l'intermédiaire de $\overline{h}_{k+1}$; autrement dit, il existe une fonction

$$\Phi_{k+1} \left[x, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k}, \quad . \quad \right]$$

telle que

$$\tilde{h}_{k+1} \left(x,u, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k} \right) = \Phi_{k+1} \left[x, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k}, \overline{h}_{k+1} \left(x,u, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k} \right) \right] .$$

Ainsi, on a

$$\left\{ \begin{array}{rcl} y_0 & = & \kappa_0(x,u) \\ y_1^{(1)} & = & \kappa_1(x,u,y_0,y_0^{(1)}) \\ & \vdots & \\ y_{k+1}^{(k+1)} & = & \kappa_{k+1} \left(x,u, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k} \right) \\ \tilde{y}_{k+2}^{(k+1)} & = & \Phi_{k+1} \left(x, \left(y_i^{(i)}, \ldots, y_i^{(k+1)} \right)_{i=0,\ldots,k}, (y_0, y_1^{(1)}, \ldots, y_{k+1}^{(k+1)}) \right) . \end{array} \right.$$

Ce qui permet de passer à l'étape suivante $k+1$.

Découplage et linéarisation entrée/sortie

Le découplage consiste à trouver une commande par retour d'état (ici quasi-statique), $u = K(x,v,\dot{v},\ldots,v^{(r)})$, telle que, sur le système bouclé,

$$\left\{\begin{array}{rcl} \dfrac{dx}{dt} & = & f(x,K(x,v,\dot{v},\ldots,v^{(r)})) \\ y & = & h(x,K(x,v,\dot{v},\ldots,v^{(r)})), \end{array}\right.$$

chaque composante de y vérifie une équation différentielle faisant intervenir uniquement cette composante, un nombre fini de ses dérivées et une seule composante de v. Ainsi, le découplage consiste à construire un changement de variable sur la commande $u \mapsto v$, changement de variable paramétré par l'état et s'interprétant comme un bouclage, tel que la relation entre la nouvelle commande v et la sortie y soit diagonale. Cela revient à compenser par bouclage les couplages non diagonaux entre u et y. Nous allons voir que l'on peut même aller un cran plus loin et en plus linéariser la relation entre la nouvelle entrée v et la sortie y.

Ce problème n'admet de solution (autour d'un point générique) que si le système est *inversible*, i.e., si la suite croissante des μ_k stationne à m. On note α alors l'unique entier tel que $\mu_\alpha = m$ et $\mu_{\alpha-1} < m$: α est appelé *ordre relatif du système entre u et y*.

Nous allons maintenant expliquer comment calculer formellement un tel bouclage. Pour cela nous reprenons l'algorithme d'inversion. La suite croissante d'entiers $\mu_0,\ldots,\mu_\alpha$ conduit à une partition des sorties en $\alpha+1$ groupes de composantes $y = (y_0,\ldots,y_\alpha)$ (de tailles respectives $(\mu_0,\mu_1-\mu_0,\ldots,\mu_\alpha-\mu_{\alpha-1})$)[5] associés aux $\alpha+1$ fonctions $(\kappa_0,\ldots,\kappa_\alpha)$. Il est possible de choisir $u = K(x,v,\dot{v},\ldots,v^{(\alpha)})$ tel que la dynamique en boucle fermée des sorties vérifie

$$\left\{\begin{array}{rcl} y_0 & = & v_0 \\ y_1^{(1)} & = & A_1(y_1) + v_1 \\ & \vdots & \\ y_\alpha^{(\alpha)} & = & A_\alpha(y_\alpha,\ldots,y_\alpha^{(\alpha-1)}) + v_\alpha \end{array}\right.$$

où

- les fonctions $(A_i)_{i=0,\ldots,\alpha}$ sont des fonctions arbitraires ;
- les nouvelles commandes $v \in \mathbb{R}^m$ se décomposent en $\alpha+1$ blocs de composantes $(v_0,\ldots,v_\alpha)$ de tailles respectives $(\mu_0,\mu_1-\mu_0,\ldots,\mu_\alpha-\mu_{\alpha-1})$.

La loi de commande est a priori une fonction de

$$x, \quad (v_0,\ldots,v_0^{(\alpha)}), \quad (v_1,\ldots,v_1^{(\alpha-1)}), \quad \ldots, \quad (v_{\alpha-1},v_{\alpha-1}^{(1)}), \quad \text{et} \quad v_\alpha.$$

En effet, u est obtenu à partir du système résultant de la dernière étape $k = \alpha-1$ de l'algorithme et où l'on a remplacé les y_i^i par $A_i + v_i$ $(i = 0,\ldots,\alpha)$:

$$\left\{\begin{array}{rcl} v_0 & = & \kappa_0(x,u) \\ A_1(y_1) + v_1 & = & \kappa_1(x,u,y_0,y_0^{(1)}) \\ & \vdots & \\ A_\alpha(y_\alpha,\ldots,y_\alpha^{(\alpha-1)}) + v_\alpha & = & \kappa_\alpha\left(x,u,\left(y_i^{(i)},\ldots,y_i^{(\alpha)}\right)_{i=0,\ldots,\alpha-1}\right) \end{array}\right.$$

Cependant, il convient d'exprimer les dérivées jusqu'à l'ordre α de y_0, …, y_α en fonction de x et des dérivées jusqu'à l'ordre α des nouvelles commandes v.

5. Noter que la composante y_k n'existe pas si $\mu_k - \mu_{k-1} = 0$.

Il est évident que $y_0^{(k)} = v_0^{(k)}$ pour $k = 0, \ldots, \alpha$. Pour $y_1^{(k)}$, nous distinguons deux cas :

- si $0 = k < 1$, alors par construction $y_1^{(0)}$ est donné par la fonction $\Phi_0(x, y_0)$ égale à $\tilde{y}_1^{(0)} = (y_1^{(0)}, y_2^{(0)}, \ldots, y_\alpha^{(0)})$ et obtenue à l'étape 0 de l'inversion ;
- si $k \geq 1$ il convient de dériver $k-1$ fois $y_1^{(1)} = A_1(y_1^{(0)}) + v_1$ pour obtenir $y_1^{(k)}$ explicitement en fonction de $y_1^{(0)}$ et $(v_1, \ldots, v_1^{(k-1)})$; comme $y_1^{(0)}$ est une fonction de x et y_0, on obtient en fin de compte $y_1^{(k)}$ en fonction de x, v_0, et $(v_1, \ldots, v_1^{(k-1)})$.

De proche en proche, on procède de même pour $y_2^{(k)}$, $y_3^{(k)}$, $\ldots, y_\alpha^{(k)}$ $(k = 0, \ldots, \alpha)$.

Il apparaît alors que v_0 doit être dérivé au plus α fois, v_1 au plus $\alpha - 1$ fois, $\ldots$, $v_{\alpha-1}$ au plus 1 fois et v_α au plus 0 fois. Ce qui explique pourquoi u dépend de x et uniquement de $(v_i, \ldots, v_i^{(\alpha-i)})_{i=0,\ldots,\alpha}$.

Bibliographie

[1] R.H. Abraham and C.D. Shaw. *Dynamics – The Geometry of Behavior : I-IV*. Aerial Press, Santa Cruz, California, 1981.
`La BD des systèmes dynamiques: sans equation et uniquement avec des dessins cette série de 4 livres donne un aperçu fidèle et assez vaste de divers types de comportements dynamiques allant des cycles limites vers l'accrochage de fréquences et les systèmes chaotiques comme le cheval de Smale.`

[2] C. Viterbo. *Systèmes dynamiques et équations différentielles*. Ecole Polytechnique, majeure de mathématiques, 2002.
`Les résultats de base sur les équations différentielles avec la moyennisation, la stabilité, et une ouverture vers le contrôle non-linéaire.`

[3] V. Arnold. *Equations Différentielles Ordinaires*. Mir Moscou, 1974.
`Un livre classique d'introduction; très géométrique sans trop de formalisme.`

[4] V. Arnold. *Méthodes Mathématiques de la Mécanique Classique*. Mir Moscou, 1976.
`Une excellente référence pour les systèmes dynamiques mécaniques (Lagrangien, Hamiltonnien, principes de moindre actions). D'un bon niveau avec de nombreuses annexes.`

[5] V. Arnold. *Chapitres Supplémentaires de la Théorie des Equations Différentielles Ordinaires*. Mir Moscou, 1980.
`De nombreux résultats sur les bifurcations et la théories de pertubations (moyennisation). D'un niveau élevé avec une rédaction parfois elliptique mais toujours très suggestive.`

[6] J.P. Bourguignon. *Calcul Variationnel*. Ecole Polytechnique, 1989.
`Un cours de géométrie différentielle intrinsèque (variétés, fibré tangent et co-tangent, champ de vecteurs, ...) avec de nombreuses informations historiques. Très complémentaire de [3, 4].`

[7] M. Demazure. *Géométrie, Catastrophes et Bifurcations*. Ecole Polytechnique, 1987.
`Une excellente introduction à l'étude qualitative des équations différentielles. Plus accessible que [5].`

[8] A. Tikhonov, A. Vasil'eva, and A. Sveshnikov. *Differential Equations*. Springer, New York, 1980.
`Sur les systèmes différentiels ordinaires lents-rapides et leur développements asymptotiques.`

[9] R Thom. *Stabilité Structurelle et Morphogénèse*. Inter-Édition, Paris, 1972.
Une discussion parfois philosophique sur les modèles et la robustesse par le père de la théorie des catastrophes.

[10] C. Godbillon. *Géométrie différentielle et mécanique analytique*. Hermann, Paris, 1969.
Un classique français très formel. Très différent de [4].

[11] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer, New York, 1983.
Traite dans le détail des exemples classiques de systèmes chaotiques (oscillations forcées, Van-der-Pol, Duffing, Lorenz,...). D'un bon niveau avec de nombreux résultats pointus. Très complémentaire de [1].

[12] M.W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems and Linear Algebra*. Acamedic Press: New-York, 1974.
Excellente indroduction avec des preuves détaillées. Moins complet que [3].

[13] T. Kailath. *Linear Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1980.
Classique et très complet sur le contrôle des systèmes linéaires.

[14] G. Allaire. *Analyse numérique et optimisation*. Ecole Polytechnique, mathématiques appliquées, 2002.
La partie sur l'optimisation traite la commande quatratique et les équations de Riccati.

[15] H.K. Khalil. *Nonlinear Systems*. MacMillan, 1992.
Manuel classique sur le contrôle non linéaire. De nombreux rappels sur les systèmes dynamiques, la théorie des perturbations et la stabilité. Assez mathématique.

[16] J.P. LaSalle and S. Lefschetz. *Stability by Liapounov's Direct Method With Applications*. Academic Press, New York, 1961.
Un classique sur la stabilité des systèmes dynamiques.

[17] E. Sontag. *Mathematical Control Theory*. Springer Verlag, 1990.
Une présentation abstraite des systèmes.

[18] J.P. Gauthier and I. Kupka. *Deterministic Observation Theory and Applications*. Cambridge University Press, 2001.
Une monographie récente sur l'observabilité en non-linéaire avec un accent mis sur les situations singulières.

[19] Ph. Martin, R. Murray and P. Rouchon. *Flat systems, equivalence and trajectory generation*. Technical report http://www.cds.caltech.edu/reports/, 2003.
Sur la linéarisation par bouclage, les systèmes plats avec extension aux équations aux dérivées partielles avec contrôle frontière. Catalogue de plusieurs dizaines d'exemples physiques.

[20] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. North-Holland, Amsterdam, 1989.
Un classique sur l'analyse numérique des systèmes mixtes ou dits implicites (équations différentielles ordinaires couplées à des équations algébriques).

[21] M. Crouzeix and A.L. Mignot. *Analyse Numérique des Equations Différentielles.* Masson, Paris, 1992.
Permet de s'orienter dans les divers schémas numériques proposés par Matlab ou Scilab pour la résolution des équations différentielles ordinaires.

[22] F.R. Gantmacher. *Théorie des Matrices: tome 1 et 2.* Dunod, Paris, 1966.
Un classique très complet sur les matrices, une mine de résultats...

Deuxième partie

Méthodes Numériques en Commande Optimale

Chapitre 1

Temps minimal : systèmes linéaires

1.1 Introduction

Lors de la conception du transfert d'un système dynamique commandé vers un point de l'espace d'état, il est nécessaire de prendre en compte plusieurs critères, en général en conflit les uns avec les autres, dont les principaux sont :

- Le temps de transfert,
- L'énergie dépensée,
- L'écart par rapport à une trajectoire de référence,
- La robustesse par rapport à des perturbations,
- La complexité du problème de calcul de la commande,
- La simplicité de mise en œuvre en temps réel.

Les poids respectifs de ces critères dépendent de chaque application. Dans les chapitres suivants, nous allons nous concentrer sur le problème de transfert en temps minimal.

Le plan du chapitre est le suivant. Nous discutons l'exemple du problème d'alunissage en section 1.2. L'existence de solutions est analysée en section 1.3, et les conditions d'optimalité en section 1.4. Enfin la théorie est appliquée à plusieurs exemples en section 1.5.

1.2 Un problème d'alunissage

Dans sa phase finale, et en négligeant la gravité, une manœuvre d'alunissage peut se modéliser par l'équation

$$\ddot{h}(t) = m^{-1}u(t), \quad t \geq 0, \tag{1.1}$$

où h est l'altitude, $m > 0$ la masse de l'engin, et u la poussée nette (après déduction de la pesanteur). On notera $v := \dot{h}$, et on impose la contrainte $u(t) \in [-1,1]$ à tout instant. Le problème est d'amener l'engin à vitesse et altitude nulle en un temps minimal.

La situation physique est celle où l'altitude initiale est positive. La solution intuitive est de fixer d'abord $u = -1$ jusqu'à atteindre un point où on commute à $u = 1$.

Nous allons résoudre graphiquement ce problème de transfert par une commande ne prenant que les valeurs ± 1, et changeant de signe au plus une fois. La théorie développée

ultérieurement permettra de montrer que que pour ce problème, ces commandes réalisent le transfert en temps minimal (voir la remarque 1.31).

Soit h_0, v_0 la condition initiale. Calculons d'abord les commandes permettant d'atteindre la cible avec une commande constante égale à ± 1. Si $u(t)$ vaut 1 pour tout $t \geq 0$, alors

$$h(t) = h_0 + tv_0 + \tfrac{1}{2}t^2, \quad v(t) = v_0 + t, \quad t \geq 0. \tag{1.2}$$

La trajectoire atteint la cible au temps $T > 0$ ssi $v_0 = -T$ et $h_0 = \frac{1}{2}T^2$. Si $u(t)$ vaut -1 pour tout $t \geq 0$, alors

$$h(t) = h_0 + tv_0 - \tfrac{1}{2}t^2, \quad v(t) = v_0 - t, \quad t \geq 0. \tag{1.3}$$

La trajectoire atteint la cible au temps $T > 0$ ssi $v_0 = T$ et $h_0 = -\frac{1}{2}T^2$. Les deux demi paraboles sont tracées en trait plein sur la figure 1.1.

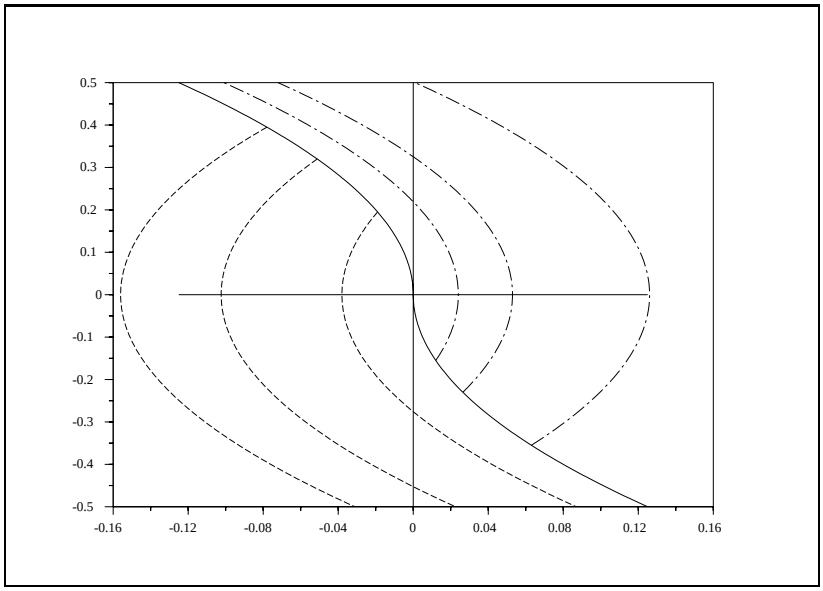


FIG. 1.1 – *Trajectoires en temps minimal*

Si la condition initiale se trouve sous la courbe en traits pleins la trajectoire obtenue avec $u = 1$ permet d'atteindre le lieu des point pouvant être transférés à 0 par une commande égale à -1; si la condition initiale se trouve au dessus, la trajectoire obtenue avec $u = -1$ permet d'atteindre le lieu des point pouvant être transférés à 0 par une commande égale à 1. Il est facile de vérifier que toutes les commandes égales à ± 1 et changeant de signe au plus une fois sont de ce type.

La courbe en traits pleins est le *lieu de changement de signe*; elle partage l'espace d'état en deux zones où la commande est constante. Nous avons réalisé (comme cela sera justifié ultérieurement) la *synthèse*, c'est à dire le calcul de la commande optimale en tout point de l'espace d'état : la commande s'exprime comme fonction de retour d'état, ou *feedback*

$$u(h,v) = \begin{cases} 1 & \text{si} \quad v \leq 0 \quad \text{et} \quad h \leq \frac{1}{2}v^2, \\ 1 & \text{si} \quad v > 0 \quad \text{et} \quad h < \frac{1}{2}v^2, \\ -1 & \text{sinon.} \end{cases} \tag{1.4}$$

1.3 Existence de solutions

1.3.1 Position du problème

Considérons le système dynamique linéaire

$$\dot{x}(t) = Ax(t) + Bu(t), \quad t \geq 0, \tag{1.5}$$

avec $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, et A et B de taille respectivement $n \times n$ et $n \times m$. La commande, fonction mesurable $\mathbb{R}_+ \longrightarrow \mathbb{R}^m$, doit respecter une contrainte du type

$$u(t) \in U, \quad \text{p.p.} \quad t \geq 0, \tag{1.6}$$

où U est un ensemble *convexe, compact* et tel que[1] $0 \in \operatorname{int} U$.

Soit C une partie convexe et fermée de l'espace d'état $\mathbb{R}^n$, appelée la *cible*. On considère le *problème de transfert en temps minimal* d'un état initial $x^0 \notin C$ à la cible :

$$\inf_{(x,u,T)} T; \quad x(T) \in C; \quad x(0) = x^0; \quad (x,u) \quad \text{satisfont (1.5)-(1.6).} \tag{1.7}$$

Remarque 1.1 La présence de la contrainte sur la commande est essentielle. En effet, si le système est commandable, le transfert de x^0 à un point quelconque de la cible est possible en un temps arbitrairement petit en l'absence de telles contraintes.

On dira que le problème en temps minimal (1.7) est *réalisable* s'il existe une commande transférant l'état initial à la cible. Le *temps minimal* noté $T(x^0)$ est la valeur de l'infimum dans (1.7), et vaut par définition $+\infty$ si le problème n'est pas réalisable.

On dit que la commande $\bar{u}$, fonction mesurable de $[0,T(x^0)]$ à valeurs dans U p.p., est une *commande en temps minimal* si elle réalise le transfert de l'état initial à la cible.

Rappelons la formule

$$x(t) = e^{tA}x^0 + \int_0^t e^{(t-s)A}Bu(s)ds, \quad t \geq 0, \tag{1.8}$$

où $e^A := \sum_{i=0}^{\infty} A^i/i!$. Etant donnés $t \geq 0$ et $x^0 \in \mathbb{R}^n$, on désigne par $\mathcal{R}(t,x^0)$ l'*ensemble des états accessibles* au temps t en partant de x^0 au temps $t = 0$. Autrement dit,

$$\mathcal{R}(t,x^0) = \left\{ e^{tA}x^0 + \int_0^t e^{(t-s)A}Bu(s)ds;\ u(s) \in U, \text{ p.p. } s \in [0,t] \right\}. \tag{1.9}$$

Soit $T > 0$. On vérifie facilement que $\cup_{0 \leq t \leq T} \mathcal{R}(t,x^0)$ est borné. Il est clair que $\mathcal{R}(T,x^0)$ est convexe; les propriétés de fermeture sont étudiées dans la section suivante à l'occasion de l'analyse de l'existence de solutions pour le problème (1.7).

1. On notera $\operatorname{int} U$ l'intérieur de U, défini comme l'ensemble des $u \in U$ tels que, pour $\rho > 0$ assez petit, la boule $B(u,\rho)$ de centre u et rayon ρ est contenue dans U.

1.3.2 Résultats d'existence

Théorème 1.2 *Si le problème en temps minimal* (1.7) *est réalisable, alors il existe une commande optimale.*

La démonstration du théorème nécessite un résultat d'analyse fonctionnelle que nous admettrons (voir Brézis [11]) :

Lemme 1.3 *Soit E une partie convexe fermée d'un espace de Hilbert F. De toute suite bornée $\{e_i\}$ dans E, on peut extraire une sous suite $\{e_j\}_{j\in J}$ qui converge faiblement vers un certain $\bar{e} \in E$, au sens où, pour toute forme linéaire continue L sur F, on a $\lim_{j\in J} L(e_j) = L(\bar{e})$.*

Lemme 1.4 *Soient $\tau > 0$, $\tau_k \longrightarrow \bar{\tau}$, et $x_k \in \mathcal{R}(\tau_k,x^0)$. Alors tout point d'adhérence x^d de $\{x_k\}$ appartient à $\mathcal{R}(\tau,x^0)$.*

Démonstration. On peut supposer que $x_k \longrightarrow x^d$. Notons u_k une commande à valeurs p.p. dans U telle que l'état associé noté x_{u_k} vérifie $x_{u_k}(\tau_k) = x_k$. Comme $\cup_{0\le t\le T_1}\mathcal{R}(t,x^0)$ est borné, l'équation d'état implique que $\|\dot{x}_{u_k}\|_{L^\infty(0,\tau_k,\mathbb{R}^n)}$ est uniformément bornée par $L > 0$. On en déduit que ces fonctions sont lipschitziennes de constante L, et donc $x_{u_k}(\bar{\tau}) \longrightarrow x^d$.

Par ailleurs la restriction de u_k à $[0,\tau]$ est bornée dans l'ensemble convexe fermé $L^2(0,\bar{\tau},U)$. Extrayant si nécessaire une sous suite on déduit du lemme 1.3 la convergence faible de cette restriction vers un certain $\bar{u} \in L^2(0,\bar{\tau},U)$. En particulier

$$x_k(\bar{\tau}) = \int_0^{\bar{\tau}} e^{(t-s)A}Bu_k(s)ds \longrightarrow \int_0^{\bar{\tau}} e^{(t-s)A}B\bar{u}(s)ds. \tag{1.10}$$

Comme $x^k(\bar{\tau}) \longrightarrow x^d$, ceci implique que x^d est la valeur de l'état associé à $\bar{u}$ à l'instant $\bar{\tau}$ d'où la conclusion. ■

Démonstration du théorème 1.2. Posons $\bar{T} := T(x^0)$. Par définition du temps minimal, il existe une suite décroissante $\{T_k\} \longrightarrow \bar{T}$ telle que $\mathcal{R}(T_k,x^0) \cap C \neq \emptyset$, et donc il existe des commandes u^k, fonctions mesurables de $[0,T_k]$ à valeurs dans U p.p., telles que les états associés x^k vérifient $x^k(T_k) \in C$. Extrayant une sous-suite, on peut supposer que la suite bornée $\{x^k(T_k)\}$ converge vers un point x^d; on conclut avec le lemme 1.4. ■

Notons l'ensemble des instants pour lesquels on peut atteindre la cible par

$$\mathcal{T}(x^0) := \left\{t > 0; \quad \mathcal{R}(t,x^0) \cap C \neq \emptyset\right\}. \tag{1.11}$$

Cet ensemble, fermé d'après le lemme 1.4, a une structure simple dans deux cas particuliers.

Définition 1.5 On dira que la cible C est *viable* si, pour tout $x^d \in C$, il existe une commande à valeur p.p. dans U telle que le système (1.5) avec état initial $x(0) = x^d$ vérifie $x(t) \in C$ pour tout $t \ge 0$.

La cible est viable si elle est réduite à 0, et plus généralement si, pour tout $x^d \in C$, il existe $u \in U$ tel que $Ax^d + Bu = 0$. On peut donner des caractérisations de la viabilité

basées sur la notion d'espace tangent à C, voir par exemple H. Frankowska [19, Section 1.3.5].

Proposition 1.6 *Si l'état initial est nul, ou si C est viable, alors $\mathcal{T}(x^0)$ est de la forme $[T(x^0),\infty[$.*

Démonstration. Notons d'abord que $T(x^0) \in \mathcal{T}(x^0)$ d'après le théorème 1.2. Si $x^0 = 0$, tout état accessible en temps t par une commande admissible $u = u(s)$ peut aussi être atteint en un temps $t' > t$ avec une commande u' nulle sur $[0,t'-t[$ et égale à $u'(s) = u(s-(t'-t))$ sur $[t'-t,t']$.

Si u transfère x^0 à $x^d \in C$ en un temps t, la viabilité de C implique l'existence d'une commande transférant x^0 à un point de C en tout temps $t' > t$. ■

Remarque 1.7 L'oscillateur harmonique, présenté dans la section 1.5.2, est un exemple de système pour lequel, en général, si C n'est pas réduit à $\{0\}$, alors $\mathcal{T}(x^0)$ n'est pas de la forme $[T(x^0),\infty[$.

1.4 Conditions d'optimalité

Cette section établit des conditions nécessaires d'optimalité pour un problème de transfert en temps minimal du type (1.7). Ces conditions, suffisantes dans certains cas, permettront de résoudre complètement un certain nombre d'exemples.

1.4.1 Séparation de l'ensemble accessible de la cible

Dans cette section, on notera $\bar{T} := T(x^0)$ le temps minimal de transfert de x^0 à C. On suppose que $x^0 \notin C$, et donc $\bar{T} > 0$. Les conditions d'optimalité sont fondées sur la notion de *séparation d'ensembles convexes*.

Définition 1.8 On dit qu'une forme linéaire q sur $\mathbb{R}^n$ sépare deux parties C_1 et C_2 de $\mathbb{R}^n$ si $q \neq 0$ et

$$q \cdot x_1 \leq q \cdot x_2, \quad \text{pour tout} \quad x_1 \in C_1,\, x_2 \in C_2. \tag{1.12}$$

Théorème 1.9 *Il existe une forme linéaire séparant C de $\mathcal{R}(\bar{T},x^0)$. Autrement dit, il existe $q \in \mathbb{R}^n$ non nulle telle que*

$$q \cdot y \leq q \cdot x, \quad \textit{pour tout} \quad y \in C \quad \textit{et} \quad x \in \mathcal{R}(\bar{T},x^0). \tag{1.13}$$

Démonstration. Soit $\{T_k\}$ une suite strictement croissante de limite $\bar{T}$, telle que $T_0 > 0$. Par définition du temps minimal, $\mathcal{R}(T_k,x^0) \cap C = \emptyset$. Nous allons séparer C de $\mathcal{R}(T_k,x^0)$, puis passer à la limite. Notons $\mathrm{dist}(\cdot,C)$ la distance (euclidienne) à l'ensemble C :

$$\mathrm{dist}(x,C) := \inf\{\|x-y\|; \quad y \in C\}. \tag{1.14}$$

Cette fonction continue atteint son minimum sur le compact $\mathcal{R}(T_k,x^0)$ en un point x^k. Puisque C est fermé, il existe $y^k \in C$ tel que $\mathrm{dist}(x^k,C) = \|y^k - x^k\|$. Posons

$$q^k := (x^k - y^k)/\|x^k - y^k\|. \tag{1.15}$$

Montrons que

$$q^k \cdot y \leq q^k \cdot y^k \leq q^k \cdot x^k \leq q^k \cdot x, \quad \text{pour tout} \quad y \in C,\ x \in \mathcal{R}(T_k,x^0). \tag{1.16}$$

La seconde inégalité est conséquence directe de (1.15). La première traduit le fait que y^k est la projection de x^k sur C. Enfin il est facile de vérifier que x^k est la projection de y^k sur $\mathcal{R}(T_k,x^0)$, ce que traduit la troisième inégalité.

Or $\{x^k\}$ est bornée, et $\{y^k\}$ l'est donc aussi. Extrayant une sous suite si nécessaire, on peut supposer que x^k converge vers x^d, avec $x^d \in \mathcal{R}(\bar{T},x^0)$ d'après le lemme 1.4, que y^k converge vers $\bar{y}$, avec $\bar{y} \in C$ puisque C est fermé, et enfin que q^k converge vers $\bar{q}$, forme linéaire de norme 1.

De plus, tout $x \in \mathcal{R}(\bar{T},x^0)$ est limite d'une suite de points de $\mathcal{R}(T_k,x^0)$: il suffit de prolonger la commande transférant à x en un temps $\bar{T}$ sur $[T_k,\bar{T}]$.

Passant à la limite dans (1.16), nous obtenons donc

$$\bar{q} \cdot y \leq \bar{q} \cdot \bar{y} \leq \bar{q} \cdot x^d \leq q \cdot x, \quad \text{pour tout} \quad y \in C \quad \text{et} \quad x \in \mathcal{R}(\bar{T},x^0), \tag{1.17}$$

d'où le résultat. ■

Remarque 1.10 On peut vérifier que les points x^d et $\bar{y}$ construits dans la démonstration précédente coïncident.

La *frontière* d'une partie K de $\mathbb{R}^n$ est notée $\partial K := K \setminus \operatorname{int} K$.

Remarque 1.11 La démonstration n'utilise pas le fait que $\bar{T}$ est le temps minimal de transfert, mais seulement l'existence d'une suite $\{T_k\}$ qui converge vers $\bar{T}$, et telle que $\mathcal{R}(T_k,x^0) \cap C = \emptyset$. La propriété de séparation est donc satisfaite par tout élément de la frontière $\partial\mathcal{T}(x^0)$ de $\mathcal{T}(x^0)$. Ce n'est donc pas une condition suffisante d'optimalité si $\partial\mathcal{T}(x^0) \neq \{T(x^0)\}$, autrement dit si $\mathcal{T}(x^0) \neq T(x^0,\infty[$. On verra dans le théorème 1.23 que sous certaines hypothèses supplémentaires ces conditions sont suffisantes.

Lemme 1.12 *Tout état final $x(\bar{T})$ associé à une commande en temps minimal appartient aux frontières des ensembles C et $\mathcal{R}(\bar{T},x^0)$.*

Démonstration. Il suffit de combiner le théorème 1.9 et le lemme qui suit [2]. ■

Lemme 1.13 *Soit une partie $\mathcal{C}$ convexe de $\mathbb{R}^n$ contenant y. Alors $y \in \operatorname{int}\mathcal{C}$ ssi il n'existe pas de forme linéaire séparant y de $\mathcal{C}$.*

Démonstration. Montrons d'abord que, si $y \in \operatorname{int}\mathcal{C}$, il n'existe pas de forme linéaire séparant y de $\mathcal{C}$. Soit $\rho > 0$ tel que $B(y,\rho) \subset \mathcal{C}$. S'il existe une forme linéaire q séparant y de $\mathcal{C}$, posons $\varepsilon := \rho/\|q\|$. Alors $y - \varepsilon q \in \mathcal{C}$, et donc avec (1.12), $0 \geq \varepsilon\|q\|^2$ ce qui donne la contradiction recherchée.

b) Soit maintenant $y \in \partial\mathcal{C}$; il faut construire une forme linéaire séparant y de $\mathcal{C}$.

Notons $\bar{\mathcal{C}}$ la fermeture de $\mathcal{C}$. Montrons que $y \in \partial\bar{\mathcal{C}}$. Dans le cas contraire, puisque $y \in \bar{\mathcal{C}}$, on aurait $y \in \operatorname{int}\bar{\mathcal{C}}$, ce qui, grâce à la convexité de $\mathcal{C}$, impliquerait $y \in \operatorname{int}\mathcal{C}$, contraire à l'hypothèse.

2. On peut admettre en première lecture ce lemme classique d'analyse convexe.

Il existe donc une suite $y^k \longrightarrow y$, avec $y^k \notin \bar{\mathcal{C}}$ pour tout k. Notons z^k la projection (orthogonale) de y^k sur $\bar{\mathcal{C}}$, et $q^k := z^k - y^k$. Puisque $y^k \notin \bar{\mathcal{C}}$, on a $q^k \neq 0$. Si $x \in \bar{\mathcal{C}}$ et $\alpha \in]0,1]$, on a $z^k + \alpha(x - z^k) \in \bar{\mathcal{C}}$, et donc

$$0 \leq \lim_{\alpha \downarrow 0} \frac{\|z^k + \alpha(x - z^k) - y^k\|^2 - \|z^k - y^k\|^2}{2\alpha} = q^k \cdot (x - z^k). \tag{1.18}$$

Or $q^k \cdot (z^k - y^k) = \|q^k\|^2 > 0$, donc $q^k \cdot (x - y^k) \geq 0$ pour tout $x \in \bar{\mathcal{C}}$. Ceci prouve que q^k sépare y^k de $\mathcal{C}$. Extrayant une sous suite si nécessaire, on peut supposer que $q^k/\|q^k\|$ converge vers $q \in I\!R^n$, de norme 1. Passant (à x fixé) à la limite dans la relation

$$\frac{q^k}{\|q^k\|} \cdot y^k \geq \frac{q^k}{\|q^k\|} \cdot x, \quad \text{pour tout} \quad x \in \mathcal{C}, \tag{1.19}$$

on obtient la relation désirée. ■

A vrai dire, l'appartenance à la frontière de l'ensemble accessible n'est une information utile que si le système est commandable. Dans le cas contraire, le lemme ci-dessous nous indique en effet que tout point accessible en temps T est un point frontière de $\mathcal{R}(T,x^0)$.

Lemme 1.14 *Pour tout $T > 0$, l'ensemble $\mathcal{R}(T,x^0)$ est d'intérieur non vide ssi le système est commandable.*

Démonstration. Si le système n'est pas commandable, soit $w \in I\!R^n$ un élément non nul du noyau à gauche de la matrice de commandabilité. Nous savons que la forme linéaire $x \longrightarrow w \cdot e^{-tA}x$ est une intégrale première; donc $\mathcal{R}(T)$ est d'intérieur vide.

Si le système est commandable, soit $\rho > 0$ tel que $B(0,\rho) \subset U$, et soit e_j un vecteur de base de $I\!R^n$. Il existe une commande continue u^j amenant l'état 0 à l'état e_j en un temps T. Posons $M := \max_j |u^j|_{L^\infty(0,T)}$. Alors $\pm\rho M^{-1}u_j$ est admissible pour tout j, et amène x^0 à $e^{TA}x_0 \pm \rho M^{-1}e_j$ en un temps T; donc $\mathcal{R}(T,x^0) \supset e^{TA}x_0 + \rho M^{-1}E$, où E désigne l'enveloppe convexe de $\{\pm e_1, \ldots, \pm e_n\}$. Puisque E est d'intérieur non vide, il en est de même pour $\mathcal{R}(T,x^0)$. ■

1.4.2 Critère linéaire sur l'état final

Dans cette section nous allons oublier (provisoirement) les problèmes de transfert en temps minimal, pour nous consacrer à l'étude du problème suivant :

$$\text{Inf } q \cdot x(T); \quad (x,u) \quad \text{satisfont (1.5)-(1.6)}, \tag{1.20}$$

où $q \in I\!R^n$ et l'horizon T sont donnés. La propriété de séparation (1.13) implique en effet qu'une commande en temps minimal est solution d'un tel problème, lorsque q est la forme linéaire séparante et $T = T(x^0)$.

Ce problème est convexe : il a un critère linéaire et des contraintes ponctuelles sur la commande. On peut caractériser ses solutions par un système d'optimalité faisant intervenir le *pseudo-hamiltonien* $H : I\!R^n \times I\!R^m \times I\!R^n \longrightarrow I\!R$ défini par

$$H(x,u,p) := p \cdot (Ax + Bu), \tag{1.21}$$

et l'*état adjoint* $p \in C([0,T],\mathbb{R}^n)$, solution de

$$\left\{ \begin{array}{ll} -\dot{p}(t) & = H_x(x(t),u(t),p(t)) = A^\top p(t), \quad t \in [0,T], \\ p(T) & = q. \end{array} \right. \tag{1.22}$$

On dira que la commande u, fonction mesurable de $[0,T]$ vers U, vérifie le *Principe du minimum* pour le problème (1.20) si elle satisfait la relation

$$H(x(t),u(t),p(t)) = \inf_{v \in U} H(x(t),v,p(t)), \quad \text{p.p.} \quad t \in [0,T]. \tag{1.23}$$

Noter que (1.23) équivaut à $p(t) \cdot B(v - u(t)) \geq 0$, pour tout $v \in U$, p.p. $t \in [0,T]$.

Théorème 1.15 *Une commande u, fonction mesurable de $[0,T]$ vers U, est solution de (1.20) ssi elle vérifie le principe du minimum.*

Démonstration. Soit u' une autre commande à valeur p.p. dans U. Posons

$$u''(t) := u(t) \text{ si } H(x(t),u(t),p(t)) \leq H(x(t),u'(t),p(t)), \quad u'(t) \quad \text{sinon.} \tag{1.24}$$

Alors u'' est mesurable, à valeur dans U p.p.; notons x'' l'état associé. Puisque $x(0) = x''(0) = x^0$, on a avec (1.5) et (1.22), après simplification,

$$\begin{array}{rl} 0 & \leq q \cdot (x''(T) - x(T)) \\ & = \displaystyle\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}[p(t) \cdot (x''(t) - x(t))] \, \mathrm{d}t = \int_0^T p(t) \cdot B(u''(t) - u(t)) \, \mathrm{d}t. \end{array} \tag{1.25}$$

Or $p(t) \cdot B(u''(t) - u(t)) \leq 0$ p.p., donc $p(t) \cdot Bu'(t) \geq p(t) \cdot Bu''(t) \geq p(t) \cdot Bu(t)$ p.p. comme on voulait le montrer. ■

On utilisera le lemme suivant dont la démonstration est immédiate.

Lemme 1.16 *Soient a et b deux fonction réelles d'une variable u. Alors*

$$\left| \inf_{u \in \mathcal{U}} a(u) - \inf_{u \in \mathcal{U}} b(u) \right| \quad \leq \quad \sup_{u \in \mathcal{U}} |a(u) - b(u)|. \tag{1.26}$$

et

$$\inf_{u \in \mathcal{U}} a(u) - \inf_{u \in \mathcal{U}} b(u) \quad \leq \quad \sup_{u \in \mathcal{U}} (a(u) - b(u)). \tag{1.27}$$

Proposition 1.17 *Si une commande u satisfait le principe du minimum sur $[0,T]$, alors l'application $t \longrightarrow H(x(t),u(t),p(t))$ est essentiellement constante*[3].

Démonstration. Posons

$$h(t) := \inf_{v \in U} H(x(t),u(t),p(t)). \tag{1.28}$$

Le lemme 1.16 implique

$$|h(t') - h(t)| \leq |p(t') \cdot Ax(t') - p(t) \cdot Ax(t)| + \sup_{v \in U} |(p(t') - p(t)) \cdot Bv| \, . \tag{1.29}$$

3. Autrement dit, constante à un ensemble de mesure nulle près.

On en déduit facilement l'existence de $M > 0$ tel que

$$|h(t') - h(t)| \leq M \left(\|x(t') - x(t)\| + \|p(t') - p(t)\| \right). \tag{1.30}$$

Or x et p sont lipschitziens, donc h l'est aussi, et a en conséquence une dérivée dans $L^\infty(0,T)$. De plus $h(t) = h(0) + \int_0^T \dot{h}(t)\mathrm{d}t$, pour tout $t \in [0,T]$. Montrons que $\dot{h}(t) = 0$ presque partout. Soit t_0 un point où h est dérivable. Le principe du minimum implique

$$\begin{aligned} \dot{h}(t_0) &\leq \lim_{t>t_0} \frac{H(x(t),u(t_0),p(t)) - H(x(t_0),u(t_0),p(t_0))}{t - t_0} \\ &= D_x H(x(t_0),u(t_0),p(t_0))\dot{x}(t_0) - D_p H(x(t_0),u(t_0),p(t_0))\dot{p}(t_0) = 0. \end{aligned} \tag{1.31}$$

De même, avec $t < t_0$ on montre que $\dot{h}(t_0) \geq 0$ et donc $\dot{h}$ est nulle p.p., de sorte que h est constante. Or $h(t) = H(x(t),u(t),p(t))$ p.p. d'après le principe du minimum, d'où la conclusion. ■

Soit $p(t)$ solution de (1.22). Alors $B^\top p(t)$ est une fonction analytique de t, donc soit est identiquement nulle, soit a un nombre fini de zéros sur $[0,T]$. Dans ce dernier cas on déduit du principe du minimum nombre de renseignements sur la commande en temps minimal.

Définition 1.18 On dit que U est *strictement convexe* si, étant donné deux points *distincts* u_1 et u_2 de U, le segment [4] $]u_1,u_2[$ appartient à l'intérieur de U.

Exemple 1.19 Dans $\mathbb{R}^n$, la boule unité fermée pour la norme ℓ^p est strictement convexe si $1 < p < \infty$, mais pas si $p = 1$ ou $p = \infty$.

Théorème 1.20 *Soit p solution de* (1.22)*, avec $q \neq 0$. Alors*
(i) *Si le système est commandable, l'application $t \longrightarrow B^\top p(t)$ n'est pas identiquement nulle.*
(ii) *Si $B^\top p(t)$ n'est pas identiquement nulle, toute solution u du problème à coût linéaire* (1.20) *est telle que $u(t) \in \partial U$ p.p. $t \in [0,T]$.*
(iii) *Si de plus l'ensemble U est strictement convexe, alors* (1.20) *a une solution unique, continue en tout instant t, sauf peut-être ceux (en nombre fini) où $B^\top p(t)$ est nul.*

Démonstration. (i) Supposons au contraire $B^\top p(t)$ identiquement nulle. Alors $0 = B^\top p(\bar{T}) = B^\top \dot{p}(\bar{T}) = \cdots$, d'où $q \cdot BA^i = 0$, pour $i = 1, \ldots, n-1$; autrement dit, q appartient au noyau à gauche de la matrice de commandabilité. Si le système est commandable, ceci implique $q = 0$, ce qui est impossible.
(ii) D'après le théorème 1.15, $u(t)$ doit minimiser la forme linéaire $v \longrightarrow p(t) \cdot Bv$ sur U à tout instant. Sauf en un nombre fini de points, cette forme linéaire est non nulle, ce qui implique que $u(t)$ est point frontière de U.
(iii) Le minimum d'une forme linéaire sur un ensemble strictement convexe compact existe et est unique. Il est facile de vérifier qu'il dépend continûment de la forme linéaire si cette dernière n'est pas nulle, ce qui assure le point (iii). ■

4. Ce segment est par définition $\{\alpha u_1 + (1-\alpha)u_2;\ \alpha \in]0,1[\}$.

1.4.3 Etat adjoint et principe du minimum

Revenons maintenant au problème de temps minimal (1.7). On dira que la commande u, fonction mesurable de $[0,T]$ vers U, vérifie le *Principe du minimum pour le problème* (1.7) si elle satisfait les relations suivantes :

$$\begin{cases} \dot{x}(t) &= Ax(t) + Bu(t), \quad t \geq 0, \\ x(0) &= x_0, \end{cases} \tag{1.32}$$

$$\begin{cases} -\dot{p}(t) &= A^\top p(t), \quad t \in [0,T], \\ p(T) &= q. \end{cases} \tag{1.33}$$

$$H(x(t),u(t),p(t)) = \inf_{v \in U} H(x(t),v,p(t)), \quad \text{p.p.} \quad t \in [0,T], \tag{1.34}$$

$$q \cdot y \leq q \cdot x(T), \quad \text{pour tout } \ y \in C; \quad x(T) \in C; \quad q \neq 0. \tag{1.35}$$

Le pseudo-hamiltonien dans (1.34) est toujours défini par (1.21). On reconnaît l'équation d'état et d'état adjoint, ainsi que la propriété de minimisation du pseudo-hamiltonien. Enfin (1.35) est conséquence de la propriété de séparation de la section 1.4.1. Définissons une *normale extérieure* à C en $z \in C$ comme un élément $q \in I\!R^n$ tel que

$$q \cdot y \leq q \cdot x(T), \quad \text{pour tout } \ y \in C. \tag{1.36}$$

Alors (1.35) dit que q est une normale extérieure *non nulle* à C en $x(T)$.

Des théorèmes 1.9 et 1.20, on déduit immédiatement le *résultat principal de ce chapitre*, qui exprime des conditions nécessaires d'optimalité :

Théorème 1.21
(i) *Toute solution u du problème de temps minimal* (1.7) *satisfait le principe du minimum* (1.32)-(1.35)*, avec $T = T(x^0)$, et $t \longrightarrow H(x(t),u(t),p(t))$ a une valeur constante p.p. le long de la trajectoire optimale.*
(ii) *Si le système est commandable, toute solution u de* (1.7) *satisfait p.p. $u(t) \in \partial U$.*
(iii) *Si le système est commandable, et U est strictement convexe, alors* (1.7) *a une solution unique, continue en tout instant t, sauf peut-être ceux (en nombre fini) où $B^\top p(t)$ est nul.*

Exemple 1.22 Etudions le cas où U est la boule unité euclidienne fermée, qui est strictement convexe. Le minimum de $v \longrightarrow r \cdot v$ sur U, pour $r \neq 0$, est atteint en $-r/\|r\|$. Donc si $B^\top p(t)$ n'est pas identiquement nulle, la commande en temps minimal vaut p.p. $u(t) = -B^\top p(t)/\|B^\top p(t)\|$.

Discutons maintenant la suffisance du principe du minimum.

Théorème 1.23 *On suppose U strictement convexe, le système commandable, et la cible viable. Alors une commande transférant le système de x^0 à C en en temps T réalise le transfert en temps minimal si et seulement elle satisfait le principe du minimum* (1.32)-(1.35).

Démonstration. D'après le théorème 1.20, ces conditions sont nécessaires. Réciproquement, supposons que la commande u satisfait (1.32)-(1.35). Le théorème 1.15 affirme que (1.32)-(1.34) caractérise les solutions du problème convexe (1.20). Soit u^* une autre

commande transférant x^0 à la cible en un temps $T^* < T$. Prolongeant u^* sur $[T^*,T]$ grâce à la viabilité de C, par une commande encore notée u^*. On obtient le transfert de x_0 en un point $x^* \in C$ avec la commande u^*. Alors (1.35) implique que u^* est aussi solution de (1.20). Comme ce dernier a une solution unique, $u = u^*$ comme il fallait le montrer. ■

Remarque 1.24 La démonstration n'exclut pas l'inégalité $T(x^0) < T$. Si une commande satisfait le principe du minimum, le temps de transfert est donc le premier instant où l'état associé appartient à la cible.

Remarque 1.25 La remarque 1.11 montre que, si la cible n'est pas viable, le principe du minimum n'est pas une condition suffisante d'optimalité.

1.5 Exemples et classes particulières

Nous allons voir que les résultats précédents permettent de donner une solution explicite au problème de commande en temps optimal dans quelques cas particuliers importants.

1.5.1 Contraintes de bornes sur la commande

Nous reprenons dans cette section le problème de temps minimal, dans le cas où l'ensemble U est la boule unité de $\mathbb{R}^m$ muni de la norme infinie :

$$U = \{u \in \mathbb{R}^m;\ \ |u_i| \leq 1,\, i = 1,\ldots,m\}. \tag{1.37}$$

Cet ensemble est convexe et compact, d'intérieur contenant 0. Il n'est en revanche pas strictement convexe si $m > 1$. Le principe du minimum implique

$$u_i(t) = \begin{cases} -1 & \text{si } (B^\top p(t))_i > 0, \\ 1 & \text{si } (B^\top p(t))_i < 0. \end{cases} \tag{1.38}$$

Si $(B^\top p(t))_i = 0$, le principe du minimum ne donne pas d'informations sur $u_i(t)$.

Puisque p est solution de l'équation linéaire homogène (sans second membre) (1.22) de dimension n, il est de la forme

$$\pi_1(t)e^{\alpha_1 t} + \cdots + \pi_r(t)e^{\alpha_r t}, \tag{1.39}$$

où $\alpha_1,\ldots,\alpha_r$ sont les valeurs propres *distinctes* de A (donc $r \leq n$) de multiplicité μ_i, et $\pi_i(t)$ est un polynôme de degré d_i, avec $d_i \leq \mu_i - 1$. Les fonctions $(B^\top p(t))_i$ sont également de la forme (1.39). Elles sont donc, sur $[0,\bar{T}]$, soit identiquement nulles, soit nulles en un nombre fini de points, et dans ce dernier cas le principe du minimum détermine u_i (sauf en ces points).

Lemme 1.26 *Soit u une commande amenant x^0 à x^d en un temps minimal $\bar{T}$, et p un état adjoint associé. Soit $i \in \{1,\ldots,m\}$. Alors, soit $(B^\top p(t))_i$ est identiquement nul, soit u_i change de signe un nombre fini de fois. Dans ce dernier cas, toutes les commandes transférant l'état de x^0 à x^d en temps minimal ont même composante i, sauf peut-être aux instants de changement de signe.*

Si les valeurs propres de A sont réelles, on peut donner une estimation du nombre des points de changement de signe :

Lemme 1.27 *Toute fonction $\psi(t)$ non nulle, de la forme (1.39), avec $\alpha_1,\dots,\alpha_r$ réels distincts et $\pi_i(t)$ polynôme réels de degré d_i, a au plus $d_1+\cdots+d_r+r-1$ zéros.*

Démonstration. Procédons par récurrence sur r. Si $r=1$, $\psi(t)=\pi_1(t)e^{\alpha_1 t}$ a les mêmes zéros que π_1; ce dernier étant un polynôme de degré d_1, au au plus $d_1=d_1+r-1$ racines sur $[0,T]$. Supposons maintenant le résultat vrai pour $r-1$. Alors $\psi(t)$ a les même zéros que la fonction

$$e^{-\alpha_1 t}\psi(t)=\pi_1(t)+\pi_2(t)e^{(\alpha_2-\alpha_1)t}+\cdots+\pi_r(t)e^{(\alpha_r-\alpha_r)t}. \tag{1.40}$$

La dérivée d'ordre d_1+1 de cette fonction est de la forme

$$\frac{\mathrm{d}^{(\mathrm{d}_1+1)}\psi(\mathrm{t})}{\mathrm{d}t^{(d_1+1)}}=\bar\pi_2(t)e^{(\alpha_2-\alpha_1)t}+\cdots+\bar\pi_r(t)e^{(\alpha_r-\alpha_1)t}, \tag{1.41}$$

avec $\bar\pi_2(t),\dots,\bar\pi_n(t)$ polynômes de degré d_i. D'après notre construction par récurrence, elle a au plus $d_2+\cdots+d_r+r-2$ zéros. Or, entre deux zéros d'une fonction se trouve au moins un zéro de sa dérivée. Si la fonction ψ avait plus de $d_1+\cdots+d_r+r-1$ zéros, sa dérivée d'ordre d_1+1 aurait donc plus de $d_2+\cdots+d_r+r-2$ zéros, d'où une contradiction. ■

Proposition 1.28 *Supposons les valeurs propres de A réelles. Soit u une commande amenant x^0 à x^d en un temps minimal T, et p un état adjoint associé. Soit $i\in\{1,\dots,m\}$. Alors, soit $(B^\top p(t))_i$ est identiquement nul, soit u_i change de signe au plus $n-1$ fois.*

Démonstration. Soient $\alpha_1,\dots,\alpha_r$ les valeurs propres *distinctes* de A de multiplicité μ_i. Alors $(B^\top p(t))_i$ est de la forme (1.39), avec $d_i\le\mu_i-1$, et a donc au plus $d_1+\cdots+d_r+r-1$ zéros. Mais

$$d_1+\cdots+d_r+r-1\le\mu_1+\cdots+\mu_r-1=n-1. \tag{1.42}$$

■

Discutons quelques exemples qui éclairciront les résultats ci-dessus.

Exemple 1.29 Considérons le problème de transfert en temps minimal de $x^0=(1,1)^\top$ à $x^d=0$, avec la dynamique

$$\dot x_1=u_1,\quad \dot x_2=2u_2, \tag{1.43}$$

et les contraintes $|u_i(t)|\le 1$, $t\in[0,T]$, $i=1,2$. Il est clair que le temps minimal de transfert est $T=1$; toute commande optimale u est telle que $u_1(t)=-1$ sur $[0,T]$; par contre on n'a pas d'unicité de $u_2(t)$. Comment cela se traduit-il sur le système d'optimalité ?

L'ensemble accessible au temps $T=1$ est $\mathcal{R}(T,x^0)=[0,2]\times[-1,3]$. Les formes linéaires séparant 0 de $\mathcal{R}(T,x^0)$ sont de la forme $q=(q_1,0)$ avec $q_1>0$. Les états adjoints associés sont $p(t)=q=(q_1,0)$. Le principe du minimum impose donc $u_1(t)=-1$ sur $[0,1]$, mais n'impose rien sur u_2, sinon d'être à valeurs dans $[-1,1]$, et tel que $x_2(T)=0$.

Exemple 1.30 Soit, pour $n \geq 1$, le système dynamique

$$\frac{\mathrm{d}^n}{\mathrm{d}\,t^n} z(t) = u(t), \quad t \in [0,T]. \tag{1.44}$$

Considérons le problème de transfert en temps minimal vers la position de repos ($z(t)$ nulle ainsi que ses dérivées jusqu'à l'ordre $n-1$) sous la contrainte $|u(t)| \leq 1$. Traduisons (1.44) en

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = x_{i+1}(t), \quad i = 1,\ldots,n-1, \quad \frac{\mathrm{d}}{\mathrm{d}t} x_n(t) = u(t), \quad t \in [0,T]. \tag{1.45}$$

La dynamique de l'état adjoint est

$$-\frac{\mathrm{d}}{\mathrm{d}t} p_i(t) = p_{i-1}(t), \quad i = 2,\ldots,n, \quad -\frac{\mathrm{d}}{\mathrm{d}t} p_1(t) = 0, \quad t \in [0,T]. \tag{1.46}$$

En particulier, $\mathrm{d}^n p_n(t)/\mathrm{d}\,t^n = 0$, donc $p_n(t)$ est un polynôme de degré au plus $n-1$.

Le système est commandable, et U est strictement convexe, donc (théorème 1.20) $B^\top p(t) = p_n(t)$ n'est pas identiquement nulle et la commande optimale est unique. La dynamique a pour seule valeur propre 0. La proposition 1.28 implique que cette commande optimale change de signe au plus $n-1$ fois.

On peut vérifier que la réciproque est vraie : toute commande amenant x^0 à 0 (en un temps T a priori quelconque) et changeant de signe au plus $n-1$ fois est optimale. En effet, soient $t_1,\ldots,t_r$ les instants de changement de signe, avec $r \leq n-1$. Posons $p(t) = \pm(t-t_1)\times\cdots\times(t-t_r)$. Alors p est un polynôme de degré $r \leq n-1$, donc satisfait l'équation de l'état adjoint, avec la condition finale $q = p(T)$, et (suivant le signe choisi dans $\pm$) la commande satisfait le principe du minimum. L'optimalité de la commande est alors conséquence du théorème 1.23.

Remarque 1.31 La discussion précédente montre que les commandes construites dans l'étude du problème d'alunissage (section 1.2) sont optimales. Le cas $n = 3$, nettement plus complexe, est traité dans Lee et Markus [22, Chapitre 2].

1.5.2 Cas de l'oscillateur harmonique

Considérons maintenant le problème de transfert en temps minimal de x^0 à $x^d = 0$ de l'oscillateur harmonique

$$\ddot{z}(t) + \omega^2 z(t) = u(t), \quad t \in [0,T], \tag{1.47}$$

où $\omega > 0$, sous la contrainte $|u(t)| \leq 1$. La dynamique avec une commande $u(t) = u_0$ constante est périodique, de la forme

$$z(t) = \omega^{-2} u_0 + r\cos(\omega t + \varphi), \quad t \in [0,T]. \tag{1.48}$$

La trajectoire décrit, dans l'espace d'état $(z,v = \dot{z})$ un cercle de rayon $(\omega^{-2}u_0,0)$ et de rayon r. Celui-ci, ainsi que la phase φ, sont déterminées par les conditions initiales. Le cercle est parcouru dans le sens des aiguilles d'une montre.

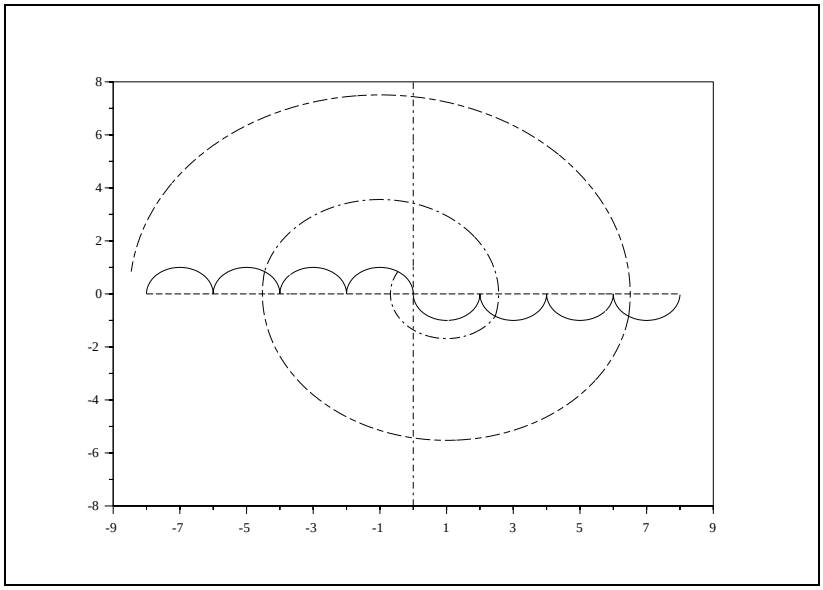


FIG. 1.2 – *Oscillateur harmonique : trajectoires en temps minimal*

L'équation de l'état adjoint $p = (p_z,p_v)$ est

$$-\dot{p}_z(t) = -\omega^2 p_v(t); \quad -\dot{p}_v(t) = p_z(t); \quad t \in [0,T], \tag{1.49}$$

et p_v est de la forme

$$p_v(t) = r' \cos(\omega t + \varphi'). \tag{1.50}$$

Les instants de changement de signe de la commande sont espacés de π/ω, et la trajectoire de transfert en temps minimal est une sucession de demi-tours (sauf le dernier qui s'arrête quand la cible est atteinte) autour des points (1,0) et (−1,0), sucessivement. Le lieu de changement de signe est marqué en traits pleins sur la figure 1.2; il est formé d'une union de demi-cercles de rayon ω^{-2}. On a représenté en pointillé une trajectoire en temps minimal dans le cas $\omega = 1$. La commande optimale est $u = 1$ en dessous du lieu de changement de signe, et $u = -1$ au dessus.

1.5.3 Stabilisation d'un pendule inversé

La linéarisation de l'équation d'un problème de stabilisation du pendule inversé conduit à l'équation

$$\ddot{z}(t) = z(t) - u(t), \quad t \in [0,T]. \tag{1.51}$$

On considère le problème d'atteinte du point de vitesse et position nulles en un temps minimal. Le système non commandé a pour valeurs propres ± 1 et n'est donc pas stable. Il faut déterminer à partir de quels points on peut atteindre la cible. Pour cela on peut s'appuyer sur les portraits de phase quand u est constant. Celui-ci est la translation de celui obtenu quand $u = 0$. (voir la figure 1.3).

D'après la proposition 1.28 une trajectoire optimale atteint la cible avec $u(t) = \pm 1$ et au plus un changement de signe.

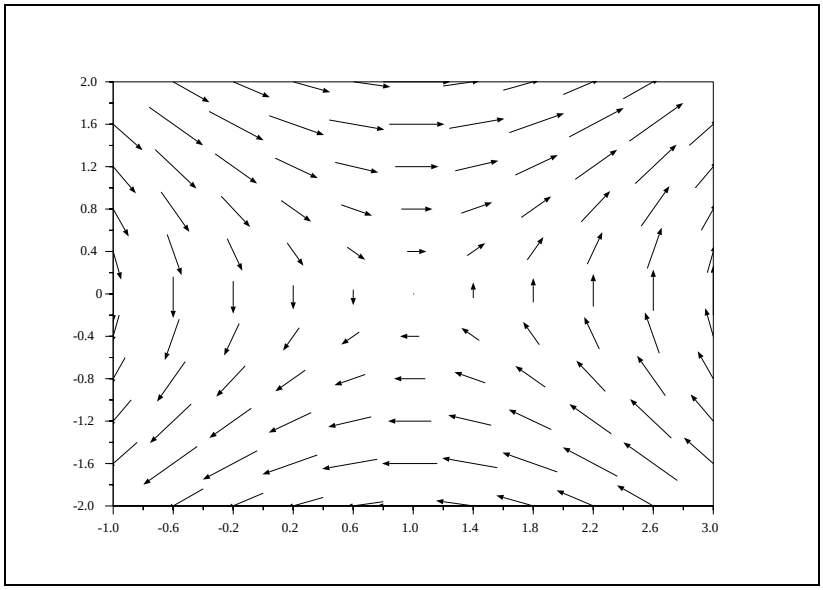


FIG. 1.3 – *Pendule inversé : portrait de phase, $u = 1$*

Points pouvant atteindre la cible avec $u = \pm 1$ constant Quand u est constant, $h(t)$ est de la forme

$$h(t) = \alpha e^t + \beta e^{-t} + u. \tag{1.52}$$

Atteindre la cible au temps T signifie que

$$\alpha e^T + \beta e^{-T} = -u; \quad \alpha e^T - \beta e^{-T} = 0. \tag{1.53}$$

De là $\alpha = -\frac{1}{2} u e^{-T}$ et $\beta = -\frac{1}{2} u e^T$. On en déduit l'expression du point initial :

$$\begin{array}{lcl} h(0) = \alpha + \beta + u & = & u - u \cosh T; \\ w(0) = \alpha - \beta & = & u \sinh T. \end{array} \tag{1.54}$$

Pour $u = \pm 1$ on obtient le lieu tracé en traits pleins sur la figure 1.4.

La courbe est tangente en 0 à l'axe vertical à la cible, et a pour asymptotes les droite $h + w = \pm 1$. En effet, on a $\cosh^2 T - \sinh^2 T = 1$, et donc $\cosh T - \sinh T = (\cosh T + \sinh T)^{-1} = o(1)$ pour T grand.

Points ne pouvant atteindre la cible Notons $v = \dot{z}$ et $\xi := h + v$. Alors $\dot{\xi} = \xi - u$. Donc si $|\xi| \geq 1$, $|\xi|$ ne peut diminuer au cours du temps. Ceci interdit d'atteindre la cible.

Points atteignant la cible avec un changement de signe de la commande On a déjà construit les trajectoires optimales sans changement de signe. Il suffit d'examiner quand les trajectoires obtenues avec $u = \pm 1$ rencontrent celles-ci. Or ces courbes sont obtenues par translation de $(u,0)$ de celles pour $u = 0$ (cf les portraits de phase, questions 3). La figure 1.4 donne la représentation des trajectoires optimales.

Remarque 1.32 Si $|h + w| < 1$, il est possible d'atteindre la cible mais il faut distinguer trois cas. Quand $|h - w| < 1$, le lieu des trajectoires en temps minimal sans changement

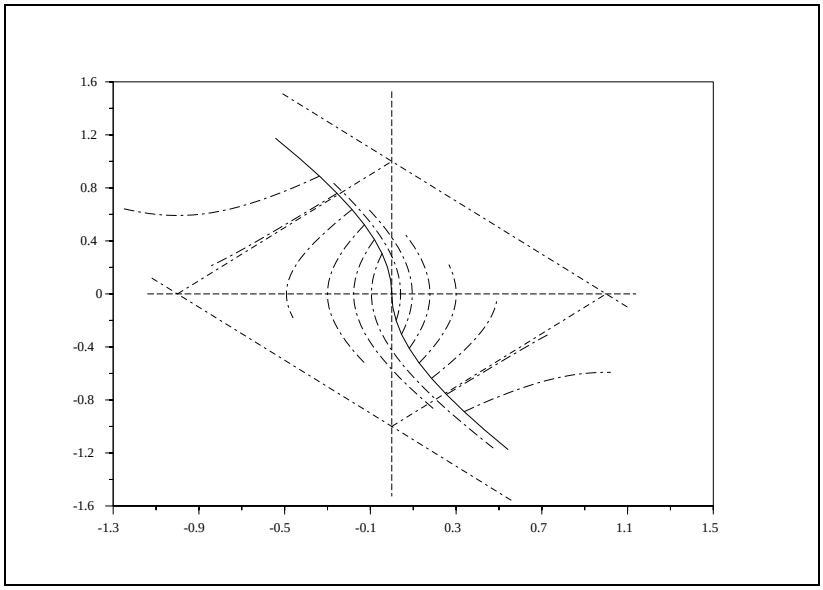


FIG. 1.4 – *Synthèse des trajectoires en temps minimal*

de signe de la commande est atteint par les trajectoires optimales en tournant dans le sens des aiguilles d'une montre. Au contraire, quand $|h-w| > 1$, les trajectoires optimales atteignent ce lieu en tournant dans le sens trigonométrique. Dans le cas limite $|h-w| = 1$, la première portion de la trajectoire optimale est rectiligne.

1.5.4 Cibles épaisses

Soit x^d l'état final d'une trajectoire en temps minimal $\bar{T}$. Dans les exemples précédents, la cible était réduite à un point et la condition de séparation (1.35) se réduisait donc à la séparation de x^d et $\mathcal{R}(\bar{T},x^0)$. Dans le cas dit de la cible épaisse, il faut prendre en compte le fait que q est une normale extérieure à C en x^d.

Exemple 1.33 Soit C égal à la boule unité fermée associée à la norme euclidienne. On sait (lemme 1.12) que $x^d \in \partial C$, soit $\|x^d\| = 1$. Toute normale extérieure à C en x^d est de la forme αx^d, avec $\alpha \in I\!R_+$. Or $q \neq 0$, et seule la direction de q importe et non son module. On peut donc supposer que $q = x(T)$. Le principe du minimum (1.32)-(1.35) équivaut alors à

$$\left\{ \begin{array}{ll} \dot{x}(t) & = Ax(t) + Bu(t), \quad t \geq 0, \\ x(0) & = x_0, \quad \|x(T)\| = 1, \end{array} \right. \tag{1.55}$$

$$\left\{ \begin{array}{ll} -\dot{p}(t) & = A^\top p(t), \quad t \in [0,T], \\ p(T) & = x(T). \end{array} \right. \tag{1.56}$$

$$H(x(t),u(t),p(t)) = \inf_{v \in U} H(x(t),v,p(t)), \quad \text{p.p.} \quad t \in [0,T]. \tag{1.57}$$

Exemple 1.34 Supposons encore C égal à la boule unité fermée associée à la norme euclidienne, l'équation d'état étant $\ddot{z} = u$, avec $U = [-1,1]$.

Notons $v = \dot{z}$, $x(T) = x^d = (z^d, v^d)$, et donc $q = (z^d, v^d)$. On a $-\dot{p}_z = 0$, $-\dot{p}_v = p_z$ et donc

$$p_z = z^d; \quad p_v = v^d + (\bar{T} - t)z^d. \tag{1.58}$$

La commande optimale vaut donc -1 et 1, respectivement, si (z^d, v^d) est dans le premier (resp. troisième) cadrant, et ne peut changer de signe que si v^d et z^d sont de signe différents.

Intégrant en temps rétrograde, à partir du temps T avec $x(T)$ quelconque de norme 1, on obtient le lieu de changement de signe. McCausland [26, Section 6.6] donne une étude détaillée de ce problème.

Chapitre 2

Temps minimal : systèmes non linéaires

Ce chapitre aborde le problèmes de transfert en temps minimal en présence d'une dynamique non linéaire. L'ensemble accessible n'est plus convexe. Une linéarisation non standard de la dynamique, basée sur des perturbations en aiguilles, permettra cependant une extension du principe du minimum.

Par ailleurs, dans le cas d'une dynamique linéaire, la commande optimale est, si le système est commandable, p.p. sur la frontière des commandes admissibles. Il n'en est plus de même quand la dynamique est non linéaire, même si la commande entre linéairement dans l'équation d'état, comme le montre l'exemple de la section 2.1.1. Ceci nous amènera à introduire la théorie des arcs singuliers.

2.1 Présentation du problème

2.1.1 Un exemple

Nous allons discuter le problème du transfert en temps minimal vers une position donnée d'un avion dont la trajectoire est horizontale et rectiligne.

Les variables d'état sont la position y, la vitesse v, et la masse m de l'engin. Les forces en jeu sont liées à la gravité g, supposée constante, la traînée D, et la portance L (drag et lift). La portance doit équilibrer la gravité, soit $L = mg$; la traînée est liée à la portance via l'incidence, et cette relation a pour expression

$$D = Av^2 + B\frac{L^2}{g^2v^2}, \tag{2.1}$$

où A et B sont deux constantes positives. Eliminant la portance, il vient

$$D = D(v,m) = Av^2 + B\frac{m^2}{v^2}. \tag{2.2}$$

La commande u est le débit d'éjection des gaz, et la poussée est $c\,u$ avec $c > 0$ constant. L'équation d'état est donc

$$\dot{y}(t) = v(t); \quad \dot{v}(t) = \frac{cu - D(v,m)}{m(t)}; \quad \dot{m}(t) = -u. \tag{2.3}$$

Nous mènerons autant que possible les calculs avec une traînée $D = D(v,m)$ sans utiliser l'expression (2.2) qui varie d'un avion à l'autre. L'état initial est noté (y^0,v^0,m^0) et la cible est

$$C = \{(y,v,m); \quad y \geq y^d; \quad m \geq m^d\}. \tag{2.4}$$

On suppose que $y^d > y^0$, $m^0 > m^d$ et que $v^0 > 0$ (si $v^0 < 0$ le problème n'a pas de sens).

Cet exemple permet d'illustrer un phénomène typique des systèmes non linéaires. Il n'est pas nécessairement optimal de rechercher des vitesses élevées en raison du terme de traînée. Il peut donc y avoir une phase du vol où la commande en temps minimal se trouvera hors des bornes. Nous allons vérifier qu'il en est ainsi, et montrer comment calculer la trajectoire optimale, en section 2.3.2.

2.1.2 Spécification du problème

Nous considérons le système dynamique non linéaire

$$\dot{x}(t) = f(t,x(t),u(t)), \quad t \geq 0; \quad x(0) = x^0, \tag{2.5}$$

avec $x(t) \in I\!R^n$, $u(t) \in I\!R^m$, et $f : I\!R \times I\!R^n \times I\!R^m \longrightarrow I\!R^n$. On supposera f lipschitzienne et dérivable, de dérivée lipschitzienne. Comme dans le chapitre précédent, on prendra en compte une contrainte sur la commande du type

$$u(t) \in U, \quad t \geq 0, \tag{2.6}$$

où U est un ensemble *convexe, compact* et tel que $0 \in \operatorname{int} U$. Le problème de transfert en temps minimal de l'état initial x^0 à un point de la cible C s'écrit

$$\inf_{(x,u,T)} T; \quad x(T) \in C; \quad (x,u) \quad \text{satisfont (2.5)-(2.6).} \tag{2.7}$$

2.1.3 Existence de solutions

Malheureusement, sous les hypothèses précédentes, il peut ne pas exister de solution au problème, comme le montre l'exemple suivant.

Exemple 2.1 Soit le système dynamique

$$\dot{x} = \sin 2\pi u, \quad \dot{y} = \cos 2\pi u, \quad \dot{z} = \pi^2(x^2 + y^2) - 1, \tag{2.8}$$

avec $x^0 = (0,0,1)$. On considère le problème du transfert en temps minimal à la cible $z = 0$, sous la contrainte $u \in [0,1]$ (qui se ramène au cas $0 \in \operatorname{int} U$ par translation). L'expression de la dérivée de z implique que le temps minimal de transfert ne peut être inférieur à 1, et que le transfert en temps $T = 1$ est impossible.

Soit k un entier positif. A la commande $u(t) = kt$ (modulo 1) est associé l'état

$$x(t) = \frac{1 - \cos 2\pi kt}{2\pi k}; \quad y(t) = \frac{\sin 2\pi kt}{2\pi k}; \tag{2.9}$$

et

$$z(t) = 1 + \int_0^t \left[\frac{1 - \cos 2\pi kt}{2k^2} - 1\right] \mathrm{d}t = 1 - t + \frac{t}{2k^2} - \frac{\sin 2\pi kt}{4\pi k^3}. \tag{2.10}$$

Cette expression permet de vérifier que $z(t_k) = 0$ pour un temps t_k tendant vers 1 quand $k \longrightarrow \infty$. L'infimum des temps de transfert est donc 1 et n'est jamais atteint.

Nous allons néanmoins donner un résultat d'existence pour la classe, importante dans les applications, des problèmes pour lesquels la *dynamique est affine par rapport à la commande*. Soient $g_1, \ldots, g_n$ des *champs de vecteurs*[1]. Supposant pour simplifier l'exposé que la dynamique est *autonome* (indépendante du temps), on se place donc dans le cas où f est de la forme

$$f(x,u) = g_0(x) + \sum_{i=1}^{n} u_i g_i(x). \tag{2.11}$$

Théorème 2.2 *On suppose la dynamique affine en la commande, les champs de vecteurs étant lipschitziens et bornés. Si le problème* (2.7) *est réalisable, il a au moins une solution.*

Démonstration. Soient u_k une suite minimisante et x_k les états associés. Le lemme 1.3 permet (extrayant une sous suite si nécessaire) d'affirmer que u_k a une limite faible $\bar{u}$, telle que $\bar{u}(t) \in U$ p.p. D'après les hypothèses sur g, la suite x_k est uniformément lipschitzienne, au sens où il existe $L > 0$ telle que

$$\|x_k(t') - x_k(t)\| \leq L|t' - t|, \quad \text{pour tout } t, t' \in [0,T]. \tag{2.12}$$

Extrayant une sous-suite si nécessaire, on en déduit[2] que x_k converge uniformément sur $[0,T]$ vers une fonction $\bar{x}$, lipschitzienne de constante L. De plus, $f(x_k(t),u_k(t))$ converge uniformément vers $f(\bar{x}(t),u_k(t))$. En conséquence, pour tout $t \in [0,T(x^0)]$,

$$\begin{aligned} \bar{x}(t) - x_0 &= \lim_k x_k(t) - x_0 = \lim_k \int_0^t f(x_k(s),u_k(s))\mathrm{d}s = \lim_k \int_0^t f(\bar{x}(s),u_k(s))\mathrm{d}s \\ &= \int_0^t g_0(\bar{x}(s))\mathrm{d}s + \sum_{i=1}^{n} \int_0^t (u_k)_i(s) g_i(\bar{x}(s))\mathrm{d}s = \int_0^t f(\bar{x}(s),\bar{u}(s))\mathrm{d}s \end{aligned} \tag{2.13}$$

où la dernière égalité est conséquence de la convergence faible. De plus $\bar{x}(\bar{T}) = \lim_k x_k(T_k)$ appartient à C puisque C est fermé. Donc $\bar{u}$ réalise le transfert en temps minimal. ■

2.2 Conditions d'optimalité

2.2.1 Un résultat général

Introduisons le *pseudo-hamiltonien* $H : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}$ défini par

$$H(t,x,u,p) := p \cdot f(t,x,u). \tag{2.14}$$

Dans le cas autonome on notera $f(x,u)$ la dynamique et $H(x,u,p)$ le pseudo-hamiltonien. On dit que la commande $u \in L^\infty(0,T,U)$ satisfait le *Principe du minimum pour le problème*

1. Un champ de vecteurs est une application de $\mathbb{R}^n$ dans lui même.

2. Par exemple en appliquant le théorème d'Ascoli-Arzela, concernant les familles équicontinues de fonctions.

(2.7) si elle satisfait les relations suivantes :

$$\begin{cases} \dot{x}(t) &= f(t,x(t),u(t)), \quad t \geq 0, \\ x(0) &= x_0, \end{cases} \tag{2.15}$$

$$\begin{cases} -\dot{p}(t) &= H_x(t,x(t),\bar{u}(t),p(t)), \quad t \in [0,T], \\ p(T) &= q, \end{cases} \tag{2.16}$$

$$H(t,x(t),u(t),p(t)) = \inf_{v \in U} H(t,x(t),v,p(t)),\ u(t) \in U, \text{ p.p. } t \in [0,T], \tag{2.17}$$

$$q \cdot y \leq q \cdot x(T), \quad \text{pour tout } \ y \in C; \quad x(T) \in C; \quad q \neq 0. \tag{2.18}$$

Théorème 2.3 *Toute solution du problème* (2.7)*. satisfait le principe du minimum.*

Démonstration. La démonstration étant technique, nous la reportons en section 2.4 pour discuter sans attendre les conséquences de ce résultat. ■

Remarque 2.4 Si la commande u satisfait le principe du minimum pour le problème (2.7), l'application $t \longrightarrow H(t,x(t),u(t),p(t))$ est essentiellement constante. On le vérifie facilement en étendant la démonstration de la proposition 1.17.

Exemple 2.5 Dans le cas de systèmes dynamiques autonomes affines en la commande, donc de dynamique donnée par (2.11), une commande u satisfaisant le principe du minimum vérifie presque partout en $t \in [0,T]$

$$\sum_{i=1}^{n} u_i(t) p(t) \cdot g_i(x(t)) \leq \sum_{i=1}^{n} v_i p(t) \cdot g_i(x(t)), \quad \text{pour tout } \ v \in U. \tag{2.19}$$

En particulier, on a p.p. $u(t) \in \partial U$ quand $p(t) \cdot g_i(x(t)) \neq 0$ pour au moins un i.

Remarque 2.6 Si la dynamique est linéaire et autonome on retrouve les résultats du chapitre précédent. Il en résulte que les conditions du théorème 2.3 ne sont pas des conditions suffisantes d'optimalité (remarque 1.25).

2.2.2 Arc singulier

Nous étudions dans cette section des systèmes dynamiques autonomes affines en la commande, dans le cas d'une seule commande :

$$\dot{x} = g_0(x(t)) + u(t) g_1(x(t)), \tag{2.20}$$

les champs de vecteurs g_0 et g_1 étant de classe C^∞, et en supposant $U = [-1,1]$. Le hamiltonien est fonction affine de la commande, de pente $p(t) \cdot g_1(x(t))$. Une commande satisfaisant le principe du minimum vérifie donc

$$\bar{u}(t) = \begin{cases} -1 & \text{si} \quad p(t) \cdot g_1(x(t)) > 0, \\ 1 & \text{si} \quad p(t) \cdot g_1(x(t)) < 0. \end{cases} \tag{2.21}$$

Nous avons vu dans l'exemple 2.1.1 que la commande en temps minimal peut se trouver hors des bornes sur un intervalle de temps $]\tau_1,\tau_2[$. Dans ce cas, l'application $v \longrightarrow H(x(t),v,p(t))$ est constante, et donc

$$p(t) \cdot g_1(x(t)) = 0, \tag{2.22}$$

de sorte que le principe du minimum ne semble donner aucune information sur la commande optimale. On appelle *arc singulier* la courbe $(x(t),u(t),p(t))$ sur $]\tau_1,\tau_2[$.

Dans la plupart des applications, on peut obtenir une expression de la commande en fonction de l'état et de l'état adjoint en dérivant autant de fois que nécessaire l'application $t \longrightarrow p(t) \cdot g_1(x(t))$.

Le calcul sera grandement simplifié par l'utilisation des *crochets de Lie* qui à une paire de champs de vecteurs (X,Y) différentiables associent un autre champ de vecteur

$$[X,Y] := X'Y - Y'X. \tag{2.23}$$

Autrement dit, $[X,Y]$ a pour composantes i, avec $1 \leq i \leq n$, la quantité

$$[X,Y]_i(x) = \sum_{j=1}^{n} X'_{ij}(x)Y_j(x) - Y'_{ij}(x)X_j(x). \tag{2.24}$$

On notera, pour $k \geq 1$,

$$adX.Y := ad^1X.Y := [X,Y]; \quad ad^{k+1}X.Y = [X,ad^kX.Y]. \tag{2.25}$$

On notera aussi $[g_0,g_1](t) = [g_0(x(t)),g_1(x(t))]$.

Théorème 2.7 *Soit une commande u vérifiant le principe du minimum. Alors, p.p. $t \in [0,T]$ on a*

$$\frac{\mathrm{d}}{\mathrm{d}t}H'_u(x(t),u(t),p(t)) = -p(t) \cdot [g_0,g_1](t), \tag{2.26}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}H'_u(x(t),u(t),p(t)) = p(t) \cdot (ad^2g_0.g_1(t) - u\, ad^2g_1.g_0(t)). \tag{2.27}$$

De plus, soit un instant t faisant partie d'un arc singulier, tel que

$$p(t) \cdot ad^2g_1.g_0(t) \neq 0.$$

Alors la commande est donné en fonction de l'état et de l'état adjoint par la formule

$$u(t) = \frac{p(t) \cdot ad^2g_0.g_1(t)}{p(t) \cdot ad^2g_1.g_0(t)}. \tag{2.28}$$

Démonstration. L'équation de l'état adjoint s'écrit ici

$$-\dot{p}(t) = (g'_0(x(t)) + u(t)g'_1(x(t)))^\top p(t). \tag{2.29}$$

Notons $\Delta^1 := \dfrac{\mathrm{d}}{\mathrm{d}t}H'_u(y(t),u(t),p(t))$ et $\Delta^2 = \dfrac{\mathrm{d}}{\mathrm{d}t}\Delta^1$. Pour simplifier les calculs, on omettra l'argument $x(t)$ des champs de vecteurs, et le temps en argument. Il vient

$$\begin{aligned}
\Delta^1 &= \frac{\mathrm{d}}{\mathrm{d}t}\langle p,g_1\rangle = \langle \dot{p},g_1\rangle + \langle p,g'_1\dot{x}\rangle \\
&= -\langle (g'_0 + ug'_1)^t p,g_1\rangle + \langle p,g'_1(g_0 + ug_1)\rangle \\
&= -\langle p,g'_0g_1 + ug'_1g_1\rangle + \langle p,g'_1g_0 + ug'_1g_1\rangle \\
&= -\langle p,g'_0g_1 - g'_1g_0\rangle = -\langle p,[g_0,g_1]\rangle,
\end{aligned}$$

d'où (2.26), et

$$\begin{aligned}\Delta^2 &= -\frac{\mathrm{d}}{\mathrm{d}t}\langle p,[g_0,g_1]\rangle = -\langle \dot{p},[g_0,g_1]\rangle - \langle p,\frac{\mathrm{d}}{\mathrm{d}t}[g_0,g_1]\rangle \\ &= \langle p,(g_0' + u\,g_1')[g_0,g_1]\rangle - \langle p,[g_0,g_1]'(g_0+ug_1)\rangle \\ &= \langle p,g_0'[g_0,g_1] - [g_0,g_1]'g_0 + u\,[g_1'[g_0,g_1] - [g_0,g_1]'g_1]\rangle \\ &= \langle p,[g_0,[g_0,g_1]] + u\,[g_1,[g_0,g_1]]\rangle.\end{aligned}$$

Mais $[g_0,g_1] = -[g_1,g_0]$, donc $[g_1,[g_0,g_1]] = -[g_1,[g_1,g_0]]$ d'où (2.27). Si t fait partie d'un arc singulier, les membres de (2.27) sont nuls, d'où (2.28). ■

Les formules (2.26)-(2.27) sont, dans certains exemples, contradictoires (quand $p(t) \neq 0$) ce qui permet alors d'exclure la présence d'arcs singuliers.

Proposition 2.8 *Supposons la dimension de l'espace d'état égale à 2 et, pour tout $x \in \mathbb{R}^2$, les champs g_1 et $[g_0,g_1]$ linéairement indépendants. Alors une trajectoire extrémale ne peut avoir d'arc singulier, et une commande en temps minimal change de signe un nombre fini de fois.*

Démonstration. Soit t un instant tel que $p(t)\cdot g_1(x(t))$, et appelons $\mathcal{T}$ l'ensemble de tels instants. On sait que $p(t) \neq 0$, sinon q serait nul, ce qui est impossible; comme $n = 2$, l'indépendance linéaire de g_1 et $[g_0,g_1]$ implique $p(t)\cdot[g_0,g_1](t) \neq 0$. D'après (2.27), t est donc un point isolé de $\mathcal{T}$. Or ce dernier est fermé, ce qui entraîne la conclusion. ■

Remarque 2.9 En d'autres termes, si $n = 2$, un arc singulier est contenu dans le *lieu singulier* de l'espace d'état défini par l'équation (on note '$\wedge$' le produit vectoriel)

$$G(x) := g_1(x) \wedge [g_0,g_1](x) = 0. \tag{2.30}$$

Sur un arc singulier, dérivant $G(x(t))$, il vient

$$G'(x(t))(g_0(x(t)) + u(t)g_1(x(t))) = 0. \tag{2.31}$$

Si $G'(x(t))g_1(x(t)) \neq 0$, on en tire une expression de la commande *en fonction de l'état.*

Remarque 2.10 Supposons n égal à 3, et notons encore $G(x) := g_1(x)\wedge[g_0,g_1](x)$. Si les vecteurs $g_1(x)$ et $[g_0,g_1](x)$ sont linéairement indépendants pour tout x, et si t appartient à un arc singulier, les relations (2.26)-(2.27) impliquent que $p(t)$ est colinéaire à $G(x(t))$. D'après le théorème 2.7, si

$$G(x(t))\cdot ad^2 g_1.g_0(t) \neq 0, \tag{2.32}$$

nous obtenons une expression de la commande *en fonction de l'état* sur un arc singulier :

$$u(t) = \frac{G(x(t))\cdot ad^2 g_0.g_1(t)}{G(x(t))\cdot ad^2 g_1.g_0(t)}. \tag{2.33}$$

Remarque 2.11 On trouvera d'autres aspects de la théorie des arcs singuliers dans Bryson et Ho [12], en particulier des conditions d'optimalité d'ordre élevé et des conditions dites de jonction, qui concernent les bords de l'arc singulier.

2.3 Applications

2.3.1 Pendule

Considérons le problème de commande du pendule

$$\ddot{\theta} + g \sin\theta = u, \tag{2.34}$$

avec $g > 0$ pesanteur, $\theta \in I\!R$ angle du pendule, et la contrainte $u \in U = [-1,1]$. Introduisant la vitesse angulaire ω, on obtient la forme suivante :

$$\dot{\theta} = \omega; \quad \dot{\omega} = u - g \sin\theta, \tag{2.35}$$

d'où l'expression du pseudo-hamiltonien

$$H(\theta,\omega,u,p_\theta,p_\omega) = p_\theta \omega + p_\omega (u - g \sin\theta) \tag{2.36}$$

et de la dynamique de l'état adjoint

$$-\dot{p}_\theta = -g p_\omega \cos\theta, \quad -\dot{p}_\omega = p_\theta. \tag{2.37}$$

Sur un arc singulier, $p_\omega = 0$, donc p_θ aussi, ce qui est impossible. Il n'existe donc pas d'arc singulier.

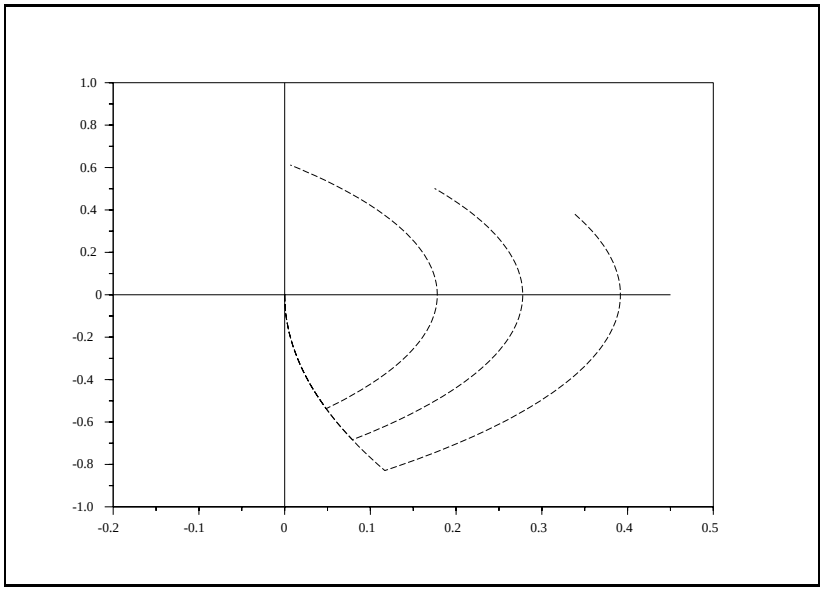


FIG. 2.1 – *Commande du pendule : quelques trajectoires en temps minimal*

La commande optimale est

$$u(t) = \begin{cases} -1 & \text{si} \quad p_\omega(t) > 0, \\ 1 & \text{si} \quad p_\omega(t) < 0. \end{cases} \tag{2.38}$$

Le long d'une trajectoire en temps minimal, on a donc

$$\begin{cases} \ddot{\theta} + g \sin\theta &= -\dfrac{p_\omega}{|p_\omega|}, \\ \ddot{p}_\omega + g p_\omega \cos\theta &= 0. \end{cases} \tag{2.39}$$

Si le temps minimal est assez petit, on obtient des trajectoires en temps minimal en intégrant l'équation d'état en temps rétrograde à partir de la cible, avec par exemple $u = 1$ sur un intervalle $[0,\tau]$, puis avec $u = -1$. Le tracé correspondant se trouve en figure 2.1.

Remarque 2.12 On trouvera dans Lee et Markus [22, Chapitre 7] une étude complète du problème (assez complexe!) de détermination de la commande optimale en des points éloignés de la cible.

D'un point de vue pratique, une heuristique consiste à prendre d'abord une commande réduisant le plus possible l'énergie mécanique, soit $u = -\dot{\omega}/|\dot{\omega}|$ puis, quand on est assez près de la cible, prendre la commande optimale (calculée ci-dessus). Enfin au voisinage immédiat de la cible on préferera un bouclage linéaire, pour éviter les oscillations rapides entre ± 1 qu'engendreraient inévitablement les bruits et vibrations diverses.

On voit sur cet exemple l'intérêt pratique de combiner différentes approches.

2.3.2 Avion à trajectoire horizontale

Nous reprenons le problème décrit dans la section 2.1 : transfert en temps minimal, vers une position donnée, d'un avion dont la trajectoire est horizontale et rectiligne. Nous allons calculer, sur un arc singulier, l'expression de la commande optimale, en fonction de l'état.

La dynamique est linéaire par rapport à la commande, et la théorie de l'arc singulier développée en section 2.2.2 s'applique donc. Cependant, plutôt que de calculer les crochets de Lie correspondants, *il est beaucoup plus simple d'effectuer des dérivations directes*[3].

On note (p_y,p_v,p_m) les coordonnées de l'état adjoint. Il vient avec (2.3), omettant les arguments quand on le peut :

$$H(y,v,m,u,p) = vp_y + p_v\frac{cu - D(v,m)}{m} - up_m, \tag{2.40}$$

et donc l'équation de l'état adjoint est

$$-\dot{p}_y = 0, \quad -\dot{p}_v = p_y - p_v\frac{D'_v}{m}, \quad -\dot{p}_m = p_v\frac{D - cu - mD'_m}{m^2}. \tag{2.41}$$

Notons que p_y est constant, donc $p_y(t) = q_y$. On a aussi

$$H'_u = c\frac{p_v}{m} - p_m. \tag{2.42}$$

Posons

$$\Delta(v,m) = \frac{mD'_m(v,m) - D(v,m) - cD'_v(v,m)}{c}. \tag{2.43}$$

Lemme 2.13 *Sur un arc singulier, la commande est solution de*

$$(\Delta + c\Delta'_v - m\Delta'_m)\,u = \Delta\,(\Delta - D'_v) + D\Delta'_v. \tag{2.44}$$

3. L'important est de comprendre le principe des calculs qui suivent, plus que le détail qui est quelque peu pénible. Dans la pratique on réalise les calculs avec des outils de calcul formel.

Démonstration. Sur l'arc singulier, on a avec (2.42), en omettant le temps en argument :

$$H'_u = c\frac{p_v}{m} - p_m = 0. \tag{2.45}$$

Dérivant en temps cette relation, il vient

$$0 = cu\frac{p_v}{m^2} - c\frac{p_y}{m} + c\frac{p_v D'_v}{m^2} + p_v\frac{D - cu - mD'_m}{m^2} = c\frac{mp_y - p_v\Delta}{m^2}. \tag{2.46}$$

Cette relation, qui conformément à la théorie ne dépend pas de u, équivaut à

$$mp_y - p_v\Delta = 0. \tag{2.47}$$

Dérivant cette relation par rapport au temps, il vient

$$-up_y + \Delta\left(p_y - p_v\frac{D'_v}{m}\right) - p_v\left(\Delta'_v\frac{cu - D}{m} - u\Delta'_m\right) = 0, \tag{2.48}$$

soit

$$\left(p_y + p_v\left(\Delta'_v\frac{c}{m} - \Delta'_m\right)\right)u = \Delta\left(p_y - p_v\frac{D'_v}{m}\right) + p_v\Delta'_v\frac{D}{m}. \tag{2.49}$$

Les relations (2.45) et (2.47) sont linéairement indépendantes (par rapport à p). Le long d'un arc singulier, p est donc proportionnel à la base du noyau des relations (2.42)-(2.46), d'expression

$$\mathcal{G}(x) = (\Delta, \quad m, \quad c)^\top. \tag{2.50}$$

Combinant avec (2.49), on obtient (2.44). ■

Remarque 2.14 Pour fournir une approximation numérique de la solution, on peut procéder comme suit. L'intuition physique suggère une première phase à débit maximum (si la vitesse initiale est faible) ou nul (si elle est élevée), suivie d'un arc singulier se terminant quand le réservoir est vide. Il suffit donc (dans chacun des deux cas) d'essayer différentes valeurs de l'instant d'entrée dans l'arc singulier. Dans les calculs on prendra garde aux bornes que doit respecter le débit de gaz dans l'arc singulier.

Remarque 2.15 On trouvera une analyse détaillée d'un problème similaire, mais un peu plus simple (on maximise la portée au lieu du temps de transfert, ce qui réduit à 2 la dimension de l'état) avec le tracé du lieu des trajectoires, dans Leitmann [23, Section 2.9].

2.4 Démonstration du résultat principal

Cette section est consacrée à la démonstration du théorème 2.3, dont la clé réside dans l'estimation de l'écart entre deux états associés à des commandes voisines, grâce à une linéarisation non standard de l'équation d'état.

On introduit la *distance d'Ekeland* sur l'espace $L^\infty(0,T,U)$:

$$\delta(u,v) := \mathrm{mes}(\{v(t) \neq u(t)\}). \tag{2.51}$$

Soient u, u_1 et u_2 dans $L^\infty(0,T,U)$, x, x_1 et x_2 leurs état associé. Posons $w := x_2 - x_1$. On note z la solution de la *linéarisation non standard* de l'équation d'état :

$$\left\{\begin{array}{rcl} \dot{z}(t) &=& f_x(t,x(t),u(t))z(t) + f(t,x(t),u_2(t)) - f(t,x(t),u_1(t)), \\ & & \text{p.p. } t \in [0,T], \\ z(0) &=& 0. \end{array}\right. \tag{2.52}$$

Lemme 2.16 *Soient u, u_1, u_2, x, x_1, x_2, w et z comme ci-dessus. Si $\delta(u_i,u) \longrightarrow 0$, $i = 1,2$, alors*

$$\text{(i) } \|w\|_{L^\infty(0,T,\mathbb{R}^n)} = O(\delta(u_2,u_1)), \quad \text{(ii) } \|w - z\|_{L^\infty(0,T,\mathbb{R}^n)} = o(\delta(u_2,u_1)). \tag{2.53}$$

Démonstration. L'application f est lipschitzienne, donc, p.p. $t \in [0,T]$:

$$\begin{array}{rcl} \|\dot{w}(t)\| &\leq& \|f(t,x_2(t),u_2(t)) - f(t,x_2(t),u_1(t))\| \\ & & + \|f(t,x_2(t),u_1(t)) - f(t,x_1(t),u_1(t))\| \\ &\leq& O(\|u_2(t) - u_1(t)\|) + O(\|w(t)\|). \end{array}$$

Comme U est compact, on a $\|u_2 - u_1\|_{L^1(0,T,U)} = O(\delta(u_2,u_1))$ et l'inégalité de Gronwall implique (2.53)(i).

Par ailleurs, on peut écrire

$$\dot{w}(t) = f(t,x(t),u_2(t)) - f(t,x(t),u_1(t)) + A_2(t) - A_1(t), \tag{2.54}$$

où pour $i = 1,2$, notant $\bar{x}_i := x_i - x$:

$$A_i(t) \;=\; f(t,x_i(t),u_i(t)) - f(t,x(t),u_i(t)) = \int_0^1 f_x(t,x(t) + \theta\bar{x}_i(t),u_i(t))\bar{x}_i(t)\mathrm{d}\theta,$$

et donc

$$\begin{array}{rl} A_2(t) - A_1(t) = & \displaystyle\int_0^1 f_x(t,x(t) + \theta\bar{x}_2(t),u_2(t))w(t)\mathrm{d}\theta + \\ & \displaystyle\int_0^1 \left[f_x(t,x(t) + \theta\bar{x}_2(t),u_2(t)) - f_x(t,x(t) + \theta\bar{x}_1(t),u_1(t))\right]\mathrm{d}\theta\bar{x}_1(t). \end{array}$$

Soient $A_3(t)$ et $A_4(t)$ les membres de droite de chaque ligne. La convergence uniforme de x_1 et x_2 vers x et l'estimation $\|w\|_{L^\infty(0,T,\mathbb{R}^n)} = O(\delta(u_2,u_1))$ impliquent :

$$A_3(t) \;=\; f_x(t,x(t),u(t))w(t) + o(\delta(u_2,u_1)) + o(\|u_2(t) - u_1(t)\|), \tag{2.55}$$

$$A_4(t) \;=\; o(\delta(u_2,u_1)) + o(\|u_2(t) - u_1(t)\|). \tag{2.56}$$

Au total, posant $y := z - w$, il vient

$$\dot{y}(t) = f_x(t,x(t),u(t))y(t) + o(\delta(u_2,u_1)) + o(\|u_2(t) - u_1(t)\|) \tag{2.57}$$

d'où (2.53)(ii) avec l'inégalité de Gronwall. ■

Définition 2.17 Soient $u \in L^\infty(0,T,U)$ et x l'état associé. On dit que $y \in \mathbb{R}^n$ est une *variation finale* associée à u, s'il existe une suite de commandes $u_k \in L^\infty(0,T,U)$, et une suite numérique $\varepsilon_k \downarrow 0$ telles que, notant x_k l'état associé à u_k, on a $(x_k(T) - x(T))/\varepsilon_k \longrightarrow y$. On note $\mathcal{C}_T(u)$ l'ensemble des variations finales.

Il est clair que $\mathcal{C}_T(u)$ est un cône fermé. Construisons un type particulier de variation admissible.

Définition 2.18 (i) La *perturbation en aiguille* associée à $t_0 \in]0,T[$ et $w \in U$, indicée par $\gamma > 0$, est la famille de commandes admissibles v_γ, d'état associé x_γ, définie par

$$v_\gamma(t) = w \ \text{ si } |t - t_0| \leq \gamma, \ u(t) \ \text{ sinon.} \tag{2.58}$$

(ii) Soit $z \in L^1(0,T,\mathbb{R}^n)$. On dit que $t_0 \in]0,T[$ est un *point de Lebesgue* de z si

$$z(t_0) = \lim_{\gamma \downarrow 0} \frac{1}{2\gamma} \int_{t_0-\gamma}^{t_0+\gamma} z(t)\mathrm{d}t. \tag{2.59}$$

On sait que (2.59) est satisfaite presque partout, voir par exemple Rudin [30, théorème 7.7]. En particulier, presque tout $t_0 \in]0,T[$ est un point de Lebesgue de $f(t,x(t),u(t))$.

Lemme 2.19 *Soient $u \in L^\infty(0,T,U)$, x l'état associé, et $t_0 \in]0,T[$ un point de Lebesgue de $f(t,x(t),u(t))$. Alors la perturbation en aiguille associée à $t_0 \in]0,T[$ et $w \in U$ est telle que la variation finale $y = \lim(x_\gamma(T) - x(T))/(2\gamma)$ existe. On l'appelle* variation en aiguille *associée à $t_0 \in]0,T[$ et $w \in U$. Si de plus p est solution de l'équation adjointe* (2.16) *(avec une condition terminale q quelconque), alors*

$$q \cdot y = H(t,x(t_0),w,p(t_0)) - H(t,x(t_0),u(t_0),p(t_0)). \tag{2.60}$$

Démonstration. On applique le lemme 2.16, avec $u_2 = v_\gamma$ et $u_1 = u$. Puisque $\delta(v_\gamma,u) \leq 2\gamma$, on a $x_\gamma(T) - x(T) = z_\gamma(T) + o(\gamma)$, où z_γ est solution de

$$\left\{ \begin{array}{rcl} \dot{z}_\gamma(t) & = & f_x(t,x(t),u(t))z_\gamma(t) + f(t,x(t),v_\gamma(t)) - f(t,x(t),u(t)), \\ & & \text{p.p. } t \in [0,T], \\ z(0) & = & 0, \end{array} \right. \tag{2.61}$$

et donc pour $q \in \mathbb{R}^n$ quelconque et p solution de l'équation adjointe (2.16),

$$\begin{array}{rcl} q \cdot z_\gamma(T) & = & p(T) \cdot z_\gamma(T) = \displaystyle\int_0^T \left[\dot{p}(t) \cdot z_\gamma(t) + p(t) \cdot \dot{z}_\gamma(t)\right] \mathrm{d}t \\ & = & \displaystyle\int_0^T p(t) \left(f(t,x(t),v_\gamma(t)) - f(t,x(t),u(t))\right) \mathrm{d}t. \end{array} \tag{2.62}$$

Revenant à la définition de v_γ et utilisant le fait que $t_0 \in]0,T[$ est point de Lebesgue de $f(t,x(t),u(t))$, on obtient (2.60) par passage à la limite, d'où la conclusion. ■

On note $\hat{\mathcal{C}}_T(u)$ le cône convexe engendré par les combinaisons linéaires positives de variations finales en aiguille. Autrement dit,

$$\hat{\mathcal{C}}_T(u) := \left\{ \sum_{i \in I} a_i z_i; \ I \text{ fini}, \ a_i \geq 0, \ z_i \in C_T(u), \ i \in I \right\}. \tag{2.63}$$

Lemme 2.20 *Les conditions suivantes sont équivalentes :*
(i) *La commande u satisfait le principe du minimum,*
(ii) *Il existe $q \neq 0$, normale extérieure à C en $x(T)$, telle que $q \cdot y \geq 0$, pour toute variation finale en aiguille y,*
(iii) $0 \notin \text{int}\left(x(T) + \hat{\mathcal{C}}_T(u) - C\right)$.

Démonstration. Le principe du minimum fournit une normale extérieure $q \neq 0$ à C en $x(T)$; si y est une variation finale en aiguille, alors $q \cdot y \geq 0$ d'après le lemme 2.19 combiné à (2.17), donc (i) implique (ii). Si (ii) est satisfait, soit p solution de (2.16) (cette équation différentielle linéaire rétrograde a une solution unique). Alors le lemme 2.19 implique le principe du minimum.

L'équivalence de (ii) et (iii) résulte du lemme 1.13, en notant que (ii) équivaut à la séparation de $\{0\}$ et de l'ensemble convexe $\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$. ■

Dans la suite on va prouver la nécessité du principe du minimum en montrant que, si u réalise le transfert en temps minimal, la condition (iii) du lemme 2.20 est satisfaite. Pour ceci il faut étudier l'ensemble $\hat{\mathcal{C}}_T(u)$.

Lemme 2.21 *Soit $u \in L^\infty(0,T,U)$. Alors $\hat{\mathcal{C}}_T(u) \subset \mathcal{C}_T(u)$.*

Démonstration. Soient $y = \sum_{i=1}^k a_i y_i$, avec $a_i > 0$ pour tout i, et y_i variation finale associée à la perturbation en aiguille associée à $t_i \in]0,T[$ et $w_i \in U$. Supposons d'abord les instants t_i distincts. On construit alors la perturbation de la commande de la manière suivante

$$v_\gamma(t) = w_i \ \text{ si } |t - t_i| \leq a_i\gamma, \ \ i = 1, \ldots, k; \ \ u(t) \ \text{ sinon.} \tag{2.64}$$

On conclut facilement avec le lemme 2.16, par des calculs similaires à ceux de la démonstration du lemme 2.19.

Donnons maintenant l'idée de la preuve du cas général en traitant le cas de deux points égaux $t_1 = t_2$, avec $k = 2$. On pose dans ce cas

$$v_\gamma(t) = \begin{cases} w_1 \ \text{ si } t \in [t_1 - 2a_1\gamma, t_1], \\ w_1 \ \text{ si } t \in]t_1, t_1 + 2a_2\gamma], \\ u(t) \ \text{ sinon.} \end{cases} \tag{2.65}$$

On conclut encore avec le lemme 2.16, par des calculs similaires à ceux de la démonstration du lemme 2.19. ■

On parlera encore de perturbation en aiguille associée à une variation finale en aiguille $y \in \hat{C} - T(u)$. Ces variation finales en aiguille notées encore v_γ, dont la démonstration donne le principe de construction, sont telles que $\delta(v_\gamma, u) = O(\gamma)$, et leur état associé x_γ vérifie $x_\gamma(T) = x(T) + 2\gamma y + o(\gamma)$.

Démontrons d'abord le théorème 2.3 dans le cas où C est d'intérieur non vide.

Lemme 2.22 *Soit u solution du problème (2.7). Si C est d'intérieur non vide, alors $0 \notin \operatorname{int}\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$, ce qui assure la conclusion du théorème 2.3 en raison du lemme 2.20.*

Démonstration. Supposons au contraire que

$$0 \in \operatorname{int}\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big). \tag{2.66}$$

Ceci implique que les convexes $x(T) + \hat{\mathcal{C}}_T(u)$ et $\operatorname{int} C$ ont une intersection non vide; sinon il existerait une forme linéaire les séparant, et séparant donc C et $x(T) + \hat{\mathcal{C}}_T(u)$, ce qui contredirait (2.66).

Il existe donc $y_0 \in \hat{\mathcal{C}}_T(u) \cap (\operatorname{int} C - x(T))$. La perturbation en aiguille v_γ correspondante est telle que son état associée x_γ vérifie $x_\gamma(T) = x(T) + 2\gamma y_0 + o(\gamma)$, donc $x_\gamma(T) \in \operatorname{int} C$ si $\gamma > 0$ est assez petit. Pour un tel $\gamma > 0$, quand $t < T$ est proche de T, on a encore $x_\gamma(t) \in \operatorname{int} C$, ce qui contredit l'optimalité de u. ■

Etudions maintenant le cas où la cible est réduite à un point.

Lemme 2.23 *Soit u solution du problème (2.7). Si C est réduit à un point, alors $0 \notin \operatorname{int}\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$, ce qui assure la conclusion du théorème 2.3 en raison du lemme 2.20.*

Démonstration. Notons $a_0, a_1, \ldots,$ diverses constantes positives. On peut supposer que $C = \{0\}$. Si la conclusion n'est pas satisfaite, de (2.66) on déduit qu'il existe r variations finales en aiguille y^1 à y^r telles que

$$2\varepsilon B \subset \operatorname{conv}\left\{y^1, \ldots, y^r\right\}. \tag{2.67}$$

Pour $\gamma > 0$, et $h \in I\!R^n_+$, on note $u_{\gamma,h}$ la perturbation en aiguille associée à la variation finale $y_h := \sum_{i=1}^r h_i y_i$, bien définie pour $\gamma < a_0$ et $\|h\| \leq a_1$, et $x_{\gamma,h}$ l'état associé. Notons aussi $S_1 := \{h \in I\!R^n_+;\ \|h\| \leq a_1\}$. On va montrer que, pour γ assez petit, si $\tau < T$ est proche de T, alors

$$0 \in \{x_{\gamma,h}(\tau); h \in S_1\}. \tag{2.68}$$

Bien entendu (2.68) contredit l'optimalité de u d'où la conclusion.

Montrons donc que (2.68) est satisfait. Pour $i = 1, \ldots, r$, notons $y_i(\tau)$ la variation au temps τ associée à la perturbation en aiguille associée à y_i (donc $y_i = y_i(T)$). Comme $y_i(\tau)$ est fonction continue de τ, (2.67) implique que pour τ proche de T on a

$$\varepsilon B \subset \operatorname{conv}\left\{y^1(\tau), \ldots, y^r(\tau)\right\}. \tag{2.69}$$

Etant donné $\gamma > 0$, essayons de résoudre en $h \in S_1$ l'équation $x_{\gamma,h}(\tau) = 0$ par l'"algorithme" de linéarisation suivant :

$$h_0 = 0; \quad \sum_{i=1}^r (h_i^{k+1} - h_i^k) y^i(\tau) = -x_{\gamma,h^k}(\tau),\ \ k = 1, \ldots. \tag{2.70}$$

D'après (2.67) cette équation a une solution telle que

$$\|h^{k+1} - h^k\| \leq a_2 \|x_{\gamma,h^k}(\tau)\|. \tag{2.71}$$

Notons u^k et x^k les commandes et états associés formés par l'algorithme. Pour que celui-ci soit bien défini pour tout k il faut, pour tout k, vérifier que $\|h^k\| \leq a_1$; c'est le cas pour $k = 0$.

Le lemme 2.16 montre que, étant donné $\varepsilon_1 > 0$, pour $\gamma > 0$ assez petit et réduisant a_1 si nécessaire, on a pour tout h et h' dans S_1 :

$$\left\| x_{\gamma,h'}(\tau) - x_{\gamma,h}(\tau) - \sum_{i=1}^r (h'_i - h_i) y_i(\tau) \right\| \leq \varepsilon_1 \|h' - h\|. \tag{2.72}$$

Donc tant que $h^k \in S_1$, pour $k \geq 2$, on a avec (2.71) et (2.72)

$$\|h^{k+1} - h^k\| \leq a_2 \|x_{\gamma,h^k}(\tau)\| \leq \varepsilon_1 a_2 \|h^k - h^{k-1}\|, \tag{2.73}$$

et donc prenant $\varepsilon_1 = \frac{1}{2}a_2$, tant que $h^{k+1} \in S_1$, pour $k \geq 2$

$$\|h^{k+1}\| \leq \sum_{i=1}^{k} \|h^{i+1} - h^i\| \leq 2\|h^1 - h^0\| \leq 2a_2 \|x(\tau)\|. \tag{2.74}$$

Si on prend τ tel que $2a_2\|x(\tau)\| < a_1$, on obtient par récurrence que $\|h^k\| \leq a_1$, donc la suite est bien définie; de plus $\|h^k - h^{k-1}\| \longrightarrow 0$, donc avec (2.73) $\|x_{\gamma,h^k}(\tau)\| \longrightarrow 0$. Posant $h^\infty := \lim_k h^k$, on obtient $x_{\gamma,h^\infty} = 0$ comme il fallait le montrer. ■

Traitons enfin le cas général, en commençant par un lemme préliminaire.

Lemme 2.24 *Si C est d'intérieur vide, moyennant si nécessaire un changement d'origine et de la base de $\mathbb{R}^n$ on peut supposer qu'il est de la forme*

$$C = \{x \in \mathbb{R}^n;\ x_i = 0,\ i = 1,\ldots,q;\ \ (x_{q+1},\ldots,x_n) \in \tilde{C}\}, \tag{2.75}$$

avec $\tilde{C}$ partie convexe de $\mathbb{R}^{n-q}$ d'intérieur non vide.

Démonstration. Le résultat est vrai si C est d'intérieur non vide. Sinon, comme C est convexe, ceci implique qu'il est contenu dans un hyperplan que par changement d'origine et de base on peut supposer de la forme $x_1 = 0$. Posons $C_1 := \{x' \in \mathbb{R}^{n-1};\ (0,x') \in C\}$. Procédant de même pour C_1, on arrive par récurrence au résultat cherché. ■

Lemme 2.25 *Soit u solution du problème (2.7). Alors $0 \notin \operatorname{int}\Big(x(T) + \hat{\mathcal{C}}_T(u) - C\Big)$, ce qui assure la conclusion du théorème 2.3 en raison du lemme 2.20.*

Démonstration. Procédons par l'absurde : supposons donc (2.66) satisfait. On peut supposer que $x(T) = 0$ et que C est de la forme (2.75). Notons $\tilde{y}$ le vecteur formé des composantes $q+1$ à n de $y \in \mathbb{R}^n$. Nous allons montrer qu'il existe une variation $y \in \hat{C}_T(u)$ telle que

$$y_i = 0,\ i = 1 \text{ à } q;\ \ \tilde{y} \in \operatorname{int} \tilde{C}. \tag{2.76}$$

En effet (2.66) implique que, pour tout $z \in \varepsilon B$ tel que $z_1 = \cdots = z_q = 0$, il existe $y \in \hat{C}_T(u)$ et $c \in C$ tels que $z = y - c$. Ceci assure que l'ensemble

$$\{(y_{q+1},\ldots,y_n);\ y \in \hat{C}_T(u); y_1 = \cdots = y_q = 0\} - \tilde{C}$$

est un voisinage de l'origine. Procédant comme dans la démonstration du lemme 2.22, on en déduit (2.76).

Soit v_γ la perturbation en aiguille associée à y et x_γ l'état correspondant. Alors $(x_\gamma)_i(T) = o(\gamma)$, $i = 1,\ldots,q$, et $(\tilde{x}_\gamma)_i(T) = 2\gamma y + o(\gamma)$.

Il existe donc $\alpha > 0$ tel que, pour tout $\varepsilon > 0$, si $\gamma > 0$ est assez petit, on a $|x_{\gamma,i}(T)| \leq \frac{1}{2}\varepsilon\gamma$, $i \leq q$, et $\tilde{x}_\gamma(T) + 2\alpha\gamma B(0,1) \subset \tilde{C}$. Pour $\tau < T$ assez proche de T, on aura donc

$$|x_{\gamma,i}(\tau)| \leq \varepsilon\gamma,\ \ i \leq q; \quad \tilde{x}_\gamma(\tau) + \alpha\gamma B_{n-q}(0,1) \subset \tilde{C}. \tag{2.77}$$

On procède alors comme dans le cas $C = \{0\}$ pour effectuer une correction assurant $x_i(\tau) = 0$, $i = 1$ à q. Comme cette correction modifie l'état à l'instant τ d'une quantité $O(\varepsilon\gamma)$, ceci assure (pour $\varepsilon > 0$ assez petit) $\tilde{x}(\tau) \in \tilde{C}$ et donc $x(\tau) \in C$ ce qui donne la contradiction recherchée. ■

2.5 Notes

On trouvera d'autres approches du principe du maximum dans l'école russe : Alexéev, V. Tikhomirov et Fomine [2], Ioffe and Tihomirov [20].

Pour les extensions au cadre non différentiable on consultera Clarke [13], Frankowska [19].

Chapitre 3

Commande optimale : l'approche HJB

3.1 Cadre

Dans ce chapitre nous étudions une classe de problèmes de commande optimale généralisant les problèmes de transfert en temps minimal. Cette classe est paramétrée par x, la condition initiale sur l'état. Nous montrerons que la valeur du problème est solution, en un sens généralisé, d'une équation aux dérivées partielles en la variable x, dite équation de Hamilton-Jacobi-Bellman (HJB). La commande optimale s'obtient alors en minimisant un hamiltonien faisant intervenir le gradient de la fonction valeur.

La classe de problèmes de commande optimale est la suivante :

$$(P_x) \qquad \begin{cases} \text{Min } \mathcal{V}(x,u,T) := \displaystyle\int_0^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t; \\ \\ \dot{y}_{x,u}(t) = f(y_{x,u}(t),u(t)), \quad t \in [0,+\infty[, \quad y_{x,u}(0) = x; \\ \\ y_{x,u}(T) \in C; \quad u(t) \in U, \quad \text{p.p.} \quad t \in [0,+\infty[. \end{cases}$$

Ici la *cible* C est une partie fermée (peut être vide ou non convexe) de $\mathbb{R}^n$, u est la *commande*, et doit appartenir à presque chaque instant à l'ensemble U, compact (convexe ou non) de $\mathbb{R}^m$; T est appelé *temps de transfert* de l'état initial x à la cible avec la commande u; il vaut par définition $+\infty$ si celle-ci n'est jamais atteinte; $y_{x,u}$ est l'état, $\lambda \geq 0$ est un *coefficient d'actualisation*, $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ est la *dynamique*, et $\ell : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ est le *coût distribué*. Nous faisons les hypothèses suivantes sur f et ℓ :

$$\begin{cases} f & \text{est lipschitzienne,} \\ \ell & \text{est lipschitzienne et bornée.} \end{cases} \tag{3.1}$$

On notera L_f et L_ℓ les constantes de Lipschitz. Ces hypothèses assurent que l'équation d'état admet, pour une commande $u \in L^\infty(0,T,U)$ donnée, une solution unique, et que le *critère*

$$\mathcal{V}(x,u,T) = \int_0^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t \tag{3.2}$$

est bien définie si T est fini ou si $\lambda > 0$.

On dit que (u,T) est admissible si $u(t) \in U$ p.p. t, $\int_0^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t$ est bien définie, et si de plus T est fini, alors $y_{x,u}(T) \in C$. On appelle *valeur* du problème (P_x) la quantité

$$V(x) := \inf\{\mathcal{V}(x,u,T);\ \ (u,T)\ \text{ admissibles}\}. \tag{3.3}$$

Si une commande (u,T) admissible atteint l'infimum, on l'appelle *commande optimale* et on dit qu'elle est *solution* du problème (P_x).

Remarque 3.1 La fonction valeur est, si $\lambda > 0$, une fonction bornée, positive si ℓ l'est. Dans ce dernier cas, $V(x) = 0$ si $x \in C$. Si de plus $\ell(x,u)$ est strictement positif pour tout $(x,u) \in \mathbb{R}^n \times U$, et si (u,T) est solution de (P_x), alors T est le premier instant où l'état atteint la cible.

Remarque 3.2 Nous retrouvons le cas particulier des problèmes de transfert en temps minimal dans le cas où ℓ vaut identiquement 1. En effet, le critère à minimiser vaut alors

$$\mathcal{V}(x,u,T) = \int_0^T e^{-\lambda t}\mathrm{d}t = \begin{cases} \lambda^{-1}(1-e^{-\lambda T}) & \text{si } \lambda > 0, \\ T & \text{si } \lambda = 0. \end{cases} \tag{3.4}$$

Minimiser ce critère équivaut bien à minimiser le temps de transfert. En particulier, si $\lambda = 0$, $V(x)$ est égal au temps minimal de transfert $T(x)$.

Un coefficient d'actualisation strictement positif permet de donner une valeur finie (égale à λ^{-1} dans le cas de problèmes de transfert en temps minimal) au critère si la cible n'est pas atteinte, ce qui facilite l'analyse mathématique ainsi que la discussion des procédés d'approximation numérique. Pour cette raison, nous *supposerons dans la suite* $\lambda > 0$.

3.2 Valeur fonction de l'état

3.2.1 Principe de programmation dynamique

Notons l'ensemble des commandes par

$$\mathcal{U} := \{u : [0,\infty[\longrightarrow \mathbb{R}^m \quad \text{mesurable};\quad u(t) \in U,\quad \text{p.p.}\quad t\}\,. \tag{3.5}$$

On dira que $(u,T) \in \mathcal{U} \times \mathbb{R}_+$ est *admissible*, relativement à la condition initiale $x \in \mathbb{R}^n$, si $y_{x,u}(T) \in C$.

Notons que, si $x \in \mathbb{R}^n \setminus C$, le fait que f soit lipschitzienne et que U soit compact assure que le temps minimal de transfert à C vérifie $T(x) > 0$.

Le théorème ci-dessous énonce le principe de programmation dynamique sous une forme un peu restrictive, mais qui suffit pour l'instant. Une forme plus complète est donnée dans le théorème 4.1.

Théorème 3.3 (Principe de Programmation Dynamique I) *Si $x \in \mathbb{R}^n \setminus C$ et $\tau \in]0,T(x)[$, alors la valeur $V(x)$ du problème (P_x) satisfait :*

$$V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}V(y_{x,u}(\tau)) \right\}. \tag{3.6}$$

Démonstration. Notons $v^*(x)$ le membre de droite de l'égalité ci-dessus. Rappelons que $\mathcal{V}(x,u,T)$ est le coût associé à l'état initial x et à une commande admissible (u,T). Alors $\tau < T(x) \leq T$, donc

$$\begin{aligned}
\mathcal{V}(x,u,T) &= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \int_\tau^T \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t, \\
&= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}\int_0^{T-\tau} \ell(y_{x,u}(t+\tau),u(t+\tau))e^{-\lambda t}\mathrm{d}t, \\
&= \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}\mathcal{V}(y_{x,u}(\tau),u(\cdot+\tau),T-\tau), \\
&\geq \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\tau}V(y_{x,u}(\tau)).
\end{aligned}$$

Minimisant chaque membre par rapport à u, il vient $V(x) \geq v^*(x)$. Pour montrer l'inégalité inverse, fixons $\varepsilon > 0$ et soit $\tilde{u}_\varepsilon$ une solution ε-optimale du problème de minimisation dans (3.6) et $\tilde{y}_\varepsilon$ l'état associé. On a donc

$$v^*(x) \geq \int_0^\tau \ell(\tilde{y}_\varepsilon(t),\tilde{u}_\varepsilon(t))e^{-\lambda t}\mathrm{d}t + V(\tilde{y}_\varepsilon(\tau))e^{-\lambda\tau} - \varepsilon.$$

Soit $(\hat{u}_\varepsilon,T)$ admissible et ε-optimal pour le problème $(P_{\tilde{y}_\varepsilon(\tau)})$ et $\hat{y}_\varepsilon$ l'état associé. Alors

$$\begin{aligned}
V(\tilde{y}_\varepsilon(\tau))e^{-\lambda\tau} + \varepsilon &\geq \int_0^T \ell(\hat{y}_\varepsilon(t),\hat{u}_\varepsilon(t))e^{-\lambda(t+\tau)}\mathrm{d}t &(3.7)\\
&= \int_\tau^{\tau+T} \ell(\hat{y}_\varepsilon(t-\tau),\hat{u}_\varepsilon(t-\tau))e^{-\lambda t}\mathrm{d}t. &(3.8)
\end{aligned}$$

Définissons la commande u_ε par

$$u_\varepsilon(t) = \begin{cases} \tilde{u}(t) & \text{si} \quad t \in [0,\tau], \\ \hat{u}_\varepsilon(t-\tau) & \text{si} \quad t \in]\tau,\infty], \end{cases} \tag{3.9}$$

et soit y_ε l'état associé. Alors

$$v^*(x) \geq \int_0^{\tau+T} \ell(y_\varepsilon(t),u_\varepsilon(t))e^{-\lambda t}\mathrm{d}t - 2\varepsilon = \mathcal{V}(x,u_\varepsilon,\tau+T) - 2\varepsilon \geq V(x) - 2\varepsilon.$$

Puisque ε peut être pris arbitrairement petit, ceci entraîne $v^*(x) \geq V(x)$, d'où le théorème. ■

Remarque 3.4 Le choix du poids exponentiel se traduit par une invariance de la valeur par rapport à l'instant initial : c'est la clé de la démonstration ci-dessus.

Remarque 3.5 Le principe de programmation dynamique peut se formuler ainsi : sur un horizon inférieur au temps minimal de transfert, la valeur optimale est égale à l'infimum de la somme du coût de transition entre les états aux instants 0 et τ et de la valeur actualisée en l'état à l'instant τ.

Exemple 3.6 Pour un problème de temps minimal de transfert, $\ell(x,u) = 1$, et le principe de programmation dynamique s'écrit donc :

$$\forall \quad \tau \in]0,T(x)[, \quad V(x) = \lambda^{-1}(1 - e^{-\lambda\tau}) + e^{-\lambda\tau} \inf_{u\in\mathcal{U}} V(y_{x,u}(\tau)). \tag{3.10}$$

3.2.2 Equation de Hamilton-Jacobi-Bellman

En vue de la discrétisation du principe de programmation dynamique, étudions le cas où $\tau \downarrow 0$ dans (3.6). Le lemme technique suivant sera utile à plusieurs reprises.

Lemme 3.7 *Soient $x \in \mathbb{R}^n \setminus C$ et $\tau \in]0,T(x)[$. Alors*

$$\tau\lambda V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(x,u(t))\mathrm{d}t + V(y_{x,u}(\tau)) - V(x) \right\} + o(\tau). \tag{3.11}$$

Démonstration. Le principe de programmation dynamique peut s'écrire

$$e^{\lambda\tau}V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t),u(t))e^{\lambda(\tau-t)}\mathrm{d}t + V(y_{x,u}(\tau)) \right\}. \tag{3.12}$$

Puisque f est lipschitzienne, on a $y_{x,u}(t) = x + O(\tau)$ (uniformément par rapport à la commande). Plus précisément, il existe $c > 0$, tel que, si $\tau > 0$ est assez petit, pour tout $t \in [0,\tau]$, on a

$$\|y_{x,u}(t) - x\| \le c\tau, \quad \text{pour tout} \quad u \in \mathcal{U}. \tag{3.13}$$

En conséquence,

$$\int_0^\tau \ell(y_{x,u}(t),u(t))e^{\lambda(\tau-t)}\mathrm{d}t = \int_0^\tau \ell(x,u(t))\mathrm{d}t + o(\tau), \tag{3.14}$$

là encore uniformément par rapport à la commande. De plus,

$$e^{\lambda\tau}V(x) = (1+\lambda\tau)V(x) + o(\tau), \tag{3.15}$$

Combinant avec (3.12) et (3.14), on obtient

$$\tau\lambda V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(x,u(t))\mathrm{d}t + V(y_{x,u}(\tau)) - V(x) + o(\tau) \right\} + o(\tau). \tag{3.16}$$

avec le premier $o(\tau)$ uniforme par rapport à la commande, et on conclut avec le lemme 1.16. ■

Introduisons le *hamiltonien* $\mathcal{H}$:

$$\mathcal{H}(x,p) \;:=\; \min_{u\in U}\{\ell(x,u) + p\cdot f(x,u)\}. \tag{3.17}$$

Remarque 3.8 Dans le cas de problèmes en temps optimal, on a introduit en (2.14) le pseudo hamiltonien $H(x,u,p) := p\cdot f(x,u)$ (dans le cas de données autonomes). Dans ce cas $\ell(x,u) = 1$, donc $\mathcal{H}(x,p) = 1 + \min_{u\in U} H(x,u,p)$.

Lemme 3.9 *Si V est différentiable en $x \in \mathbb{R}^n \setminus C$, alors*

$$\lambda V(x) = \mathcal{H}(x,DV(x)).$$

Démonstration. Puisque f est lipschitzienne, utilisant (3.13), il vient

$$y_{x,u}(\tau) = x + \int_0^\tau f(y_{x,u}(t),u(t))\mathrm{d}t = x + \int_0^\tau f(x,u(t))\mathrm{d}t + o(\tau), \tag{3.18}$$

avec $o(\tau)/\tau \longrightarrow 0$ quand $\tau \downarrow 0$, uniformément par rapport à la commande. Comme V est différentiable en x, on a

$$V(y_{x,u}(\tau)) = V(x) + \int_0^\tau DV(x) \cdot f(x,u(t))\mathrm{d}t + o(\tau), \tag{3.19}$$

avec encore un $o(\tau)$ uniforme. Combinant avec les lemmes 1.16 et 3.7, il vient

$$\tau\lambda V(x) = \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \left[\ell(x,u(t)) + DV(x)f(x,u(t))\right] \mathrm{d}t \right\} + o(\tau). \tag{3.20}$$

L'infimum ci-dessus est atteint en maximisant séparément pour chaque t; en conséquence,

$$\tau\lambda V(x) = \tau\mathcal{H}(x,DV(x)) + o(\tau), \tag{3.21}$$

d'où la conclusion en divisant par $\tau \downarrow 0$. ■

On appellera *équation de Hamilton-Jacobi-Bellman* (HJB), pour la famille de problèmes de commande optimale (P_x), l'équation aux dérivées partielles non linéaire du premier ordre sur $\mathbb{R}^n \setminus C$ avec conditions aux limites

$$\left\{ \begin{array}{lllll} \text{(i)} & \lambda v(x) & = & \mathcal{H}(x,Dv(x)), & x \in \mathbb{R}^n \setminus C, \\ \text{(ii)} & v(x) & = & 0, & x \in C, \end{array} \right. \tag{3.22}$$

dans laquelle l'inconnue est la fonction $v : \mathbb{R}^n \longrightarrow \mathbb{R}$.

Remarque 3.10 L'étude de cette équation aux dérivées partielles présente plusieurs difficultés :

(i) $V(x)$ n'est en général pas différentiable sur $\mathbb{R}^n \setminus C$. Il faut donc donner un sens à (3.22)(i) aux points où $V(x)$ n'est pas différentiable.

(ii) $V(x)$ n'est pas nécessairement continue sur C (voir l'exemple 3.11). Là encore il faut donner un sens à la condition aux limites.

(iii) Il peut y avoir plusieurs solutions continues sur $\mathbb{R}^n$, et différentiable sur $\mathbb{R}^n \setminus C$, de (3.22) (exemple 3.12).

Exemple 3.11 Soit le problème de transfert à 0, en dimension 1, avec la dynamique $\dot{x} = u$, $0 \le u \le 1$. Considérons la formulation actualisée avec $\lambda = 1$.

On sait que $V(x) = 1 - e^{-T(x)}$. Or $T(x)$ vaut $-x$ si $x \le 0$, et $+\infty$ sinon; donc

$$V(x) = \left\{ \begin{array}{ll} 1 - e^x & \text{si } x \le 0, \\ 1 & \text{sinon.} \end{array} \right. \tag{3.23}$$

La valeur est donc discontinue en 0.

Exemple 3.12 Soit le problème de transfert à 0, en dimension 1, avec la dynamique $\dot{x} = u$, $-1 \leq u \leq 1$. Considérons la formulation actualisée avec $\lambda = 1$. Alors $T(x) = |x|$, et donc $V(x) = 1 - e^{-|x|}$. La valeur est continue, et différentiable en tout point différent de la cible 0. Le hamiltonien a pour expression

$$\mathcal{H}(x,p) = \min_{u\in[-1,1]} \{1 + up\} = 1 - |p|, \tag{3.24}$$

et l'équation HJB s'écrit donc

$$\left\{ \begin{array}{rcl} v(x) & = & 1 - |Dv(x)|, \quad x \neq 0, \\ v(0) & = & 0. \end{array} \right. \tag{3.25}$$

La valeur est bien solution de cette équation. Mais les fonctions $w_1(x) = 1-e^x$ et $w_2(x) = 1-e^{-x}$ sont d'autres solutions continues et différentiables en tout point différent de 0 (elles sont même différentiables en 0). Notons cependant que ces solutions "parasites" sont non bornées alors que $V(x)$ l'est.

3.2.3 Continuité uniforme de la valeur

Une fonction est d'autant plus facile à approcher numériquement qu'elle est régulière. Montrons que, *si la cible est vide*, la fonction V est uniformément continue. On note alors $\mathcal{V}(x,u)$ le critère. Il vient donc

$$V(x) = \inf_{u\in\mathcal{U}} \mathcal{V}(x,u) \tag{3.26}$$

où $\mathcal{U}$ est défini en (3.5).

Lemme 3.13 *Si $C = \emptyset$, la fonction valeur $V(x)$ est hölderienne et bornée.*

Démonstration. Montrons que V est bornée. On a pour toute commande u

$$|\mathcal{V}(x,u)| \leq \int_0^\infty |\ell(y_{x,u}(t),u(t))| e^{-\lambda t} \mathrm{d}t \leq \lambda^{-1} \|\ell\|_\infty, \tag{3.27}$$

d'où $|V(x)| \leq \lambda^{-1}\|\ell\|_\infty$.

Montrons que V est uniformément continue. Puisque f est lipschitzien, la quantité

$$\lambda_0 := \sup_{\substack{u\in U \\ x\neq x'}} \frac{(f(x',u) - f(x,u)) \cdot (x' - x)}{|x' - x|^2} \tag{3.28}$$

est finie. Montrons que deux trajectoires associées à la même commande u satisfont la relation

$$|y_{x'}(t) - y_{x,u}(t)| \leq |x' - x| e^{\lambda_0 t}. \tag{3.29}$$

En effet, posant $z(t) := y_{x'}(t) - y_{x,u}(t)$, il vient

$$\tfrac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|z(t)|^2 = z(t) \cdot \dot{z}(t) \leq \lambda_0 |z(t)|^2,$$

et donc $|z(t)|^2 \leq e^{2\lambda_0 t}|x' - x|^2$, d'où (3.29). Par ailleurs, (1.26) implique

$$|V(x') - V(x)| \leq \sup_{u\in\mathcal{U}} \left\{ \int_0^\infty |\ell(y_{x'}(t),u(t)) - \ell(y_{x,u}(t),u(t))| e^{-\lambda t} \mathrm{d}t \right\}.$$

Soit $T > 0$. Notons

$$\begin{aligned}\Delta_1 &:= \sup_{u\in\mathcal{U}} \int_0^T |\ell(y_{x'}(t),u(t)) - \ell(y_{x,u}(t),u(t))| e^{-\lambda t}\mathrm{d}t,\\ \Delta_2 &:= \sup_{u\in\mathcal{U}} \int_T^\infty |\ell(y_{x'}(t),u(t)) - \ell(y_{x,u}(t),u(t))| e^{-\lambda t}\mathrm{d}t.\end{aligned}$$

Alors $|V(x') - V(x)| \leq \Delta_1 + \Delta_2$. Supposant sans perte de généralité $\lambda_0 > \lambda$ (il suffit que λ_0 majore le membre de droite de (3.27)), nous obtenons avec (3.29)

$$\begin{aligned}\Delta_1 &\leq L_\ell \int_0^T |x' - x| e^{(\lambda_0-\lambda)t}\mathrm{d}t = L_\ell \frac{e^{(\lambda_0-\lambda)T} - 1}{\lambda_0 - \lambda}|x' - x|,\\ \Delta_2 &\leq 2\int_T^\infty \|\ell\|_\infty e^{-\lambda t}\mathrm{d}t = \frac{2}{\lambda} e^{-\lambda T}\|\ell\|_\infty.\end{aligned}$$

Soit x' tel que $|x' - x| < 1$. Choisissons $T > 0$ tel que $e^{-T} = |x' - x|^{\frac{1}{\lambda_0}}$ (c'est possible !). Alors les quantités Δ_1 et Δ_2 se majorent ainsi :

$$\begin{aligned}\Delta_1 &\leq \frac{L_\ell}{\lambda_0 - \lambda}|x' - x|\left(|x' - x|^{\frac{\lambda}{\lambda_0}-1} - 1\right) \leq \frac{L_\ell}{\lambda_0 - \lambda}|x' - x|^{\frac{\lambda}{\lambda_0}},\\ \Delta_2 &\leq \frac{2}{\lambda}\|\ell\|_\infty |x' - x|^{\frac{\lambda}{\lambda_0}},\end{aligned}$$

et donc

$$|V(x') - V(x)| \leq \left(\frac{L_\ell}{\lambda_0 - \lambda} + \frac{2}{\lambda}\|\ell\|_\infty\right)|x' - x|^{\frac{\lambda}{\lambda_0}},$$

d'où la conclusion. ■

3.3 Commande optimale

Sous les hypothèses faites au début du chapitre, il n'existe pas en général de commande optimale (comme le montre l'exemple 2.1). Nous allons cependant, sous des hypothèses fortes, établir dans cette section comment obtenir la commande optimale à partir de la connaissance de la fonction valeur V.

Théorème 3.14 *Supposons la fonction valeur continûment différentiable sur $\mathbb{R}^n \setminus C$, et continue en tout point de C. Soit $x \in \mathbb{R}^n \setminus C$. Alors la commande u est optimale si et seulement si, p.p. $s \in [0,T]$, où T est le temps de transfert à C (éventuellement infini) avec la commande u, cette commande minimise le hamiltonien au sens suivant :*

$$\mathcal{H}(y_{x,u}(s),DV(y_{x,u}(s)) = \ell(y_{x,u}(s),u(s)) + f(y_{x,u}(s),u(s)) \cdot DV(y_{x,u}(s)). \tag{3.30}$$

Démonstration. Soient (u,T) une commande admissible, et $s \in]0,T[$; $y_{x,u}(s)$ n'appartient pas à C, donc V est dérivable en $y_{x,u}(s)$. Le lemme 3.9 et la définition du hamiltonien impliquent

$$\lambda V(y_{x,u}(s)) - f(y_{x,u}(s),u(s)) \cdot DV(y_{x,u}(s)) \leq \ell(y_{x,u}(s),u(s)), \tag{3.31}$$

avec égalité ssi (3.30) est satisfait.

Soit $\tau \in]0,T[$. La régularité de V permet d'écrire, compte-tenu de (3.31) :

$$\begin{aligned} V(x) - e^{-\lambda\tau}V(y_{x,u}(\tau)) &= \int_0^\tau \frac{\mathrm{d}}{\mathrm{d}t}\left[-e^{-\lambda t}V(y_{x,u}(t))\right]\mathrm{d}t \\ &= \int_0^\tau \left[\lambda V(y_{x,u}(t)) - f(y_{x,u}(t),u(t))\cdot DV(y_{x,u}(t))\right]e^{-\lambda t}\mathrm{d}t \\ &\leq \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t, \end{aligned} \tag{3.32}$$

avec égalité ssi (3.30) est satisfait.

Faisons maintenant tendre τ vers T. Si T est fini, de $V(y_{x,u}(T)) = 0$ on déduit que u est optimal ssi (3.30) est satisfait. Si $T = +\infty$, on a $e^{-\lambda\tau}V(y_{x,u}(\tau)) \longrightarrow 0$ puisque V est bornée, d'où la même conclusion. ■

Remarque 3.15 Le résultat précédent a plusieurs extensions utiles, par exemple au cas où la fonction valeur V est seulement dérivable en tout $y_{x,u}(s)$, $s \in]0,T[$, sauf peut-être en un nombre fini d'entre eux.

Le théorème précédent donne le moyen de vérifier si une commande *fonction du temps* est optimale. Voyons maintenant le résultat principal de la section, qui montre comment construire la commande optimale *en fonction de l'état (forme feedback)* :

Théorème 3.16 *Supposons* (i) *la fonction valeur continûment différentiable sur* $\mathbb{R}^n \setminus C$, *de dérivée localement lipschitzienne, et continue en tout point de* C, (ii) *le minimum dans la définition du hamiltonien* (3.17) *atteint en un point unique* $\Upsilon(x,p)$, *la fonction* Υ *étant localement lipschitzienne.*

Alors la commande ci-dessous, sous forme feedback, est optimale :

$$u(x) = \Upsilon(x,DV(x)). \tag{3.33}$$

Démonstration. L'équation différentielle

$$\dot{y}_{x,u}(t) = f(y_{x,u}(t),\Upsilon(y_{x,u}(t),DV(y_{x,u}(t)))) \tag{3.34}$$

a un second membre borné et localement lipschitzien, donc a une solution unique; l'optimalité de la commande découle du théorème 3.14. ■

Exemple 3.17 reprenons le problème de l'exemple 3.12. On a $V(x) = 1-e^{-|x|}$, et $V'(x) = -e^x$ si $x < 0$, $V'(x) = e^{-x}$ si $x > 0$. La commande réalisant le maximum dans la définition du hamiltonien est donc $u(x) = 1$ si $x < 0$, $u(x) = -1$ si $x > 0$. Chacun des deux théorèmes précédents peut être appliqué à ce problème.

Remarque 3.18 La vérification de l'hypothèse (ii) du théorème 3.16 se ramène à une analyse de stabilité de la solution d'un problème d'optimisation en dimension finie, voir [10, Section 4.4.1]. En pratique cette hypothèse se vérifie dans le cas (assez restrictif) où U est convexe fermé, f est affine par rapport à la commande, et ℓ est uniformément fortement convexe par rapport à la commande, pour tout x.

Remarque 3.19 La fonction valeur n'est en général pas continûment différentiable, même sous les hypothèses fortes de la remarque 3.18. Les théorèmes 3.14 et 3.16 ne peuvent donc être appliqués que dans un nombre limité de cas.
On retiendra néanmoins la règle heuristique suivante. Soit $x \in I\!R^n \setminus C$. Alors un "candidat sérieux", pour être commande optimale en x, est l'argument de la minimisation dans la définition du hamiltonien, évaluant celui-ci en $(x,DV(x))$.

3.4 Solution de viscosité

Cett section présente une notion permettant de donner un sens à l'équation (3.22), dite HJB :

$$\left\{ \begin{array}{lllll} \text{(i)} & \lambda v(x) & = & \mathcal{H}(x,Dv(x)), & x \in I\!R^n \setminus C, \\ \text{(ii)} & v(x) & = & 0, & x \in C, \end{array} \right. \tag{3.35}$$

avec C partie fermée de $I\!R^n$, *même quand la solution n'est pas différentiable*. On pourra passer les preuves en première lecture. Nous limiterons l'étude aux solutions continues sur $I\!R^n$. Le problème principal est de donner un sens à (3.35)(i), de manière à ce que la valeur soit l'unique solution de (3.35).

3.4.1 Notion de solutions de viscosité

Notons dans la suite

$$\Omega := I\!R^n \setminus C. \tag{3.36}$$

On peut définir une notion de solution généralisée de (3.35) grâce à l'observation suivante.

Lemme 3.20 *Soit Φ une fonction différentiable en $x \in \Omega$, telle que $V - \Phi$ a un maximum (resp. minimum) local en x. Alors*

$$\lambda V(x) - \mathcal{H}(x,D\Phi(x)) \quad \leq 0 \quad (\text{resp. } \geq 0). \tag{3.37}$$

Démonstration. Il suffit de donner la démonstration dans le cas où $V - \Phi$ a un maximum local en x. Alors, pour tout x' dans un voisinage $\mathcal{N}$ de x, on a

$$V(x') - V(x) \leq \Phi(x') - \Phi(x). \tag{3.38}$$

Pour τ assez petit, puisque f est bornée, $y_{x,u}(\tau) \in \mathcal{N}$, quelle que soit la commande appliquée. Combinant (3.38) et le lemme 3.7, il vient

$$\tau\lambda V(x) \leq \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(x,u(t))\mathrm{d}t + \Phi(y_{x,u}(\tau)) - \Phi(x) \right\} + o(\tau). \tag{3.39}$$

On procède alors comme dans la démonstration du lemme 3.9, en adaptant (3.19), (3.20) et (3.21) (changements d'égalités en inégalités, remplacement de V par Φ). ■

Formalisons ce qui précède en introduisant un vocabulaire adapté.

Définition 3.21 Une fonction $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ est dite *sous solution* (resp. *sur solution*) au *sens de viscosité* de (3.35)(i) si, pour tout $x_0 \in \Omega$, et $\Phi : \mathbb{R}^n \to \mathbb{R}$ de classe C^1, telle que x_0 est point de maximum (resp. minimum) local de $v - \Phi$, alors

$$\lambda v(x_0) - \mathcal{H}(x_0, D\Phi(x_0)) \quad \leq 0 \quad (\text{resp. } \geq 0). \tag{3.40}$$

On dit que v est *solution au sens de viscosité* de (3.35)(i) si elle est à la fois sur et sous solution au sens de viscosité.

Théorème 3.22 *La fonction valeur V est solution au sens de viscosité de* (3.35)(i).

Le théorème est conséquence immédiate du lemme 3.20. Ce dernier est apparemment plus fort, car il suppose seulement la fonction Φ dérivable en x. Nous allons voir que les deux énoncés sont équivalents, en introduisant un concept important.

Définition 3.23 Soit v une fonction $\mathbb{R}^n \longrightarrow \mathbb{R}$. On dit que $p \in \mathbb{R}^n$ est une *sous dérivée, ou sous gradient* (resp. *sur dérivée, ou sur gradient*) de v en x, si

$$v(x') - v(x) - p^\top (x' - x) \geq o(\|x' - x\|) \quad (\text{resp. } \leq o(\|x' - x\|)) \,. \tag{3.41}$$

On notera $D^- v(x)$ (resp. $D^+ v(x)$) l'ensemble des sous gradients (resp. sur gradients) de v en x.

Exemple 3.24 La fonction valeur absolue $v : \mathbb{R} \longrightarrow \mathbb{R}$, $v(x) := |x|$, est telle que $D^- v(0) = [-1,1]$ et $D^+ v(0) = \emptyset$.

Remarque 3.25 (i) Si v est dérivable en un point x, alors

$$D^- v(x) = D^+ v(x) = \{Dv(x)\}. \tag{3.42}$$

(ii) Si en un point x on a $D^- v(x) \neq \emptyset$ et $D^+ v(x) \neq \emptyset$, alors v est dérivable en x et (3.42) est satisfait.

Soit p un sur gradient de v en x. Posons

$$\Phi(x') = \max(v(x'), v(x) + p^\top (x' - x)).$$

Alors $v - \Phi$ atteint un maximum local en x et, par définition du sur gradient, la fonction Φ a pour dérivée p en x. Réciproquement, si une fonction Φ dérivable en x est telle que $v - \Phi$ atteint un maximum local en x, il est clair que $D\Phi(x)$ est un sur gradient de v en x. Nous avons montré que

$$D^+ v(x) = \{p;\ \exists \Phi : \mathbb{R}^n \longrightarrow \mathbb{R};\ p = D\Phi(x);\ v - \Phi \text{ a un maximum local en } x\}.$$

On peut montrer (voir par exemple Barles [5, Section 2.2]) que $D^+ v(x)$ est aussi l'ensemble des gradients de fonctions continûment dérivables, telles que $v - \Phi$ atteint un maximum local en x. Bien entendu on a un résultat similaire pour les sous gradients. Les conditions du lemme 3.20 et du théorème 3.22 coïncident donc. Ceci implique le résultat suivant :

Lemme 3.26 *Soit $x \in \Omega$ et $v : \mathbb{R}^n \longrightarrow \mathbb{R}$. Les énoncés suivants sont équivalents :*
(i) *On a $\lambda v(x) \leq \mathcal{H}(x,D\Phi(x))$, pour toute fonction Φ dérivable en x telle que $v-\Phi$ atteint un maximum local en x.*
(ii) *On a $\lambda v(x) \leq \mathcal{H}(x,D\Phi(x))$, pour toute fonction Φ continûment dérivable telle que $v-\Phi$ atteint un maximum local en x.*
(ii) *On a $\lambda v(x) \leq \mathcal{H}(x,p)$, pour tout $p \in D^+v(x)$.*

Nous laissons le lecteur énoncer le résultat correspondant concernant les sous gradients.

Remarque 3.27 (i) Soit v une sous solution de 3.37(i) au sens de viscosité, et $x \in \Omega$ tel que v soit dérivable en x. Combinant le lemme précédent et la remarque 3.25, il vient $\lambda v(x) \leq \mathcal{H}(x,Dv(x))$. De même pour les sur solutions.
(ii) Soit v dérivable sur Ω. Combinant le point (i) et le lemme précédent, on voit que v est sous solution de 3.37(i) au sens classique ssi elle est sous solution de viscosité. De même pour les sur solutions.

3.4.2 Théorème de comparaison

Nous avons noté que la fonction valeur V n'est pas toujours continue. Dans tous les cas, cette solution est solution de l'équation HJB au sens de viscosité.

Le résultat principal de cette section (théorème 3.31) implique, si la cible C est vide, l'unicité (autrement dit, l'existence d'au plus une) d'une solution hölderienne et bornée de l'équation HJB. Si V est hölderienne, on obtient donc l'existence et l'unicité de la solution, dans la classe des fonctions hölderiennes et bornées.

Pour l'étude de convergence des schémas numériques, nous avons besoin d'un résultat un peu plus fort que l'unicité : des résultats de comparaison entre les sous-solutions semi continues supérieurement (s.c.s.) et les sur solutions semi continues inférieurement (s.c.i.).

Définition 3.28 On dit que la fonction $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ est semi continue supérieurement (s.c.s.) (resp. semi continu inférieurement (s.c.i.)) si pour tout $x \in \mathbb{R}^n$ on a

$$v(x) \geq \limsup_{x' \longrightarrow x} v(x'), \quad \left(\text{resp. } v(x) \leq \liminf_{x' \longrightarrow x} v(x')\right).$$

Remarque 3.29 On peut exprimer les propriétés précédentes à l'aide de suites convergentes vers x. Ainsi, la fonction $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ est s.c.s. ssi, pour toute suite x_k convergeant vers x, on a $v(x) \geq \limsup_k v(x_k)$; ou encore, si pour tout point d'adhérence $v^* \in \mathbb{R} \cup \{\pm\infty\}$ de $v(x^k)$, on a $v(x) \geq v^*$. De même pour la semi continuité inférieure.

Définition 3.30 On appelle *principe d'unicité fort* pour l'équation (3.35) tout résultat du type suivant : Soient v (resp. w) une sous solution (resp. sur solution) de (3.35) (assorti éventuellement de conditions de régularité sur v et w satisfaites par la fonction valeur). Alors $\sup v \leq \inf w$.

Compte tenu de la difficulté de ce type de résultat, nous limiterons l'analyse au cas $C = \emptyset$. Pour l'extension aux problèmes avec temps d'arrêt, il est utile de considérer une équation aux dérivées partielles générale du premier ordre, notée

$$\overline{\mathbf{H}}(x,v(x),Dv(x)) = 0, \quad \text{pour tout } x \in \mathbb{R}^n. \tag{3.43}$$

On suppose que le "hamiltonien abstrait" $\overline{\mathbf{H}}$ vérifie les relations

$$\begin{aligned} |\overline{\mathbf{H}}(x,v,p') - \overline{\mathbf{H}}(x,v,p)| &\leq c_1\|p'-p\|; && (3.44)\\ |\overline{\mathbf{H}}(x',v,p) - \overline{\mathbf{H}}(x,v,p)| &\leq c_2\|x'-x\|(1+\|p\|); && (3.45)\\ \overline{\mathbf{H}}(x,v',p) - \overline{\mathbf{H}}(x,v,p) &\geq c_3(v'-v), && (3.46) \end{aligned}$$

avec $c_3 > 0$. Dans le cas de l'équation HJB on a

$$\overline{\mathbf{H}}(x,v,p) := \lambda v - \inf_{u\in U}\{\ell(x,u) + p\cdot f(x,u)\}, \tag{3.47}$$

et si la dynamique est bornée, on vérifie (3.44)-(3.46), avec

$$c_1 := \sup\{\|f(x,u)\|;\ (x,u)\in\Omega\times U\}; \quad c_2 := L_\ell + L_f; \quad c_3 := \lambda. \tag{3.48}$$

Une fonction $v : \mathbb{R}^n \longrightarrow \mathbb{R}$ est dite *sous solution* (resp. *sur solution*) au *sens de viscosité* de (3.43) si, pour tout $\Phi : \mathbb{R}^n \to \mathbb{R}$ de classe C^1, telle que x_0 est point de maximum (resp. minimum) local de $v - \Phi$, on a

$$\overline{\mathbf{H}}(x_0,v(x_0),D\Phi(x_0)) \quad \leq 0 \quad (\text{resp. } \geq 0). \tag{3.49}$$

On dit que v est *solution au sens de viscosité* de (3.43) si elle est à la fois sur et sous solution au sens de viscosité.

Théorème 3.31 (Principe d'unicité fort) *Sous les hypothèses* (3.44)-(3.46)*, si v est une sous solution s.c.s. bornée supérieurement de* (3.43)*, et w est une sur solution s.c.i. bornée inférieurement de* (3.43)*, une de ces deux fonction étant hölderienne, alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Corollaire 3.32 *Si la dynamique est bornée et $C = \emptyset$, la fonction valeur $V(x)$ du problème (P_x) est l'unique solution de viscosité continue et bornée sur $\mathbb{R}^n$ de l'équation HJB* (3.35)*.*

Démonstration. Le lemme 3.13 dit que la fonction valeur $V(x)$ est hölderienne et bornée. D'après le théorème 3.22, $V(x)$ est solution de viscosité dans $\mathbb{R}^n$ de (3.35)(i). Soit v une autre solution de viscosité continue et bornée. Le théorème 3.31 implique $v \leq V$ et $V \leq v$, d'où $v = V$. ■

Il reste à démontrer le théorème 3.31. La démonstration est quelque peu technique et le lecteur intéressé principalement par les méthodes numériques peut la sauter en première lecture. Donnons cependant un résultat de comparaison élémentaire (mais sous des hypothèses trop fortes) qui donnera une idée du principe de la démonstration.

Proposition 3.33 *On suppose que $C = \emptyset$. Soient v et w une sous et sur solution de* (3.49) *respectivement. Supposons le maximum de $v - w$ atteint en un point x_0 où v et w sont différentiables. Alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Démonstration. Puisque $v-w$ atteint son maximum en x_0, on a $Dv(x_0) = Dw(x_0)$. Or v est sous solution, et w est sous solution. Par une remarque analogue à 3.27(i), il vient

$$\overline{\mathbf{H}}(x_0,v(x_0),Dv(x_0)) \leq 0 \leq \overline{\mathbf{H}}(x_0,w(x_0),Dv(x_0)) \tag{3.50}$$

et on conclut avec (3.46). ■

Démonstration. (Démonstration du théorème 3.31). Supposons v hölderienne, l'autre cas se traitant d'une manière similaire. L'idée essentielle de la démonstration est le dédoublement des variables : Pour tout $\varepsilon > 0$, posons

$$\varphi(x,y) := v(x) - w(y) - \tfrac{1}{2}\varepsilon^{-2}\|x-y\|^2, \quad \text{pour tout } (x,y) \in I\!R^n \times I\!R^n. \tag{3.51}$$

Le rôle du dernier terme est d'obtenir des points x et y proches quand on considère des solutions approchées du problème de maximisation de φ.

Soit $\delta \in]0,1[$. On a

$$\sup \varphi \leq \sup v - \inf w < +\infty, \tag{3.52}$$

donc il existe (x_1,y_1) in $I\!R^{2n}$ tel que

$$\varphi(x_1,y_1) > \sup \varphi - \delta. \tag{3.53}$$

Il existe aussi une fonction ξ, de classe C^∞ à support compact, telle que

$$\xi(x_1,y_1) = 1, \qquad 0 \leq \xi \leq 1, \quad \sup_{x,y} \|D\xi(x,y)\| \leq 1. \tag{3.54}$$

Posons

$$\psi(x,y) = \varphi(x,y) + \delta\xi(x,y), \quad \text{pour tout} \quad (x,y) \in I\!R^{2n}. \tag{3.55}$$

Si (x,y) n'est pas dans le support de ξ, on a

$$\psi(x,y) = \varphi(x,y) \leq \sup \varphi < \psi(x_1,y_1). \tag{3.56}$$

Une suite maximisante pour ψ est donc, à partir d'un certain rang, incluse dans le support de ξ qui est compact; or ψ est s.c.s., donc atteint son maximum en un point (x_o,y_o). Autrement dit,

$$\psi(x_o,y_o) \geq \psi(x,y) \quad \text{pour tout } (x,y) \in I\!R^{2n}. \tag{3.57}$$

En particulier, la fonction $x \longrightarrow v(x) - \frac{1}{2}\varepsilon^{-2}\|x-y_0\|^2 + \delta\xi(x,y_0)$ atteint un maximum local en x_0. Par définition d'une sous solution de viscosité, on a donc

$$\overline{\mathbf{H}}(x_0,v(x_0),\varepsilon^{-2}(x_o-y_o) - \delta D_x\xi(x_o,y_o)) \leq 0. \tag{3.58}$$

De même, $y \longrightarrow w(y) + \frac{1}{2}\varepsilon^{-2}\|x_0-y\|^2 - \delta\xi(x_0,y)$ atteint un minimum local en x_0, donc par définition d'une sur solution de viscosité, on a

$$\overline{\mathbf{H}}(y_0,w(y_0),\varepsilon^{-2}(x_o-y_o) + \delta D_y\xi(x_o,y_o)) \geq 0. \tag{3.59}$$

Utilisant (3.54) et (3.58)-(3.59), il vient

$$\overline{\mathbf{H}}(x_0,v(x_0),\varepsilon^{-2}(x_o-y_o)) \leq \delta c_1; \quad \overline{\mathbf{H}}(y_0,w(y_0),\varepsilon^{-2}(x_o-y_o)) \geq -\delta c_1. \tag{3.60}$$

Soustrayant ces relations, nous obtenons avec (3.45) et (3.46),

$$\begin{aligned} c_3(v(x_0)-w(y_0)) &\leq \overline{\mathbf{H}}(x_0,v(x_0),\varepsilon^{-2}(x_o-y_o)) - \overline{\mathbf{H}}(x_0,w(y_0),\varepsilon^{-2}(x_o-y_o)) \\ &\leq \overline{\mathbf{H}}(y_0,w(y_0),\varepsilon^{-2}(x_o-y_o)) - \overline{\mathbf{H}}(x_0,w(y_0),\varepsilon^{-2}(x_o-y_o)) + 2\delta c_1 \\ &\leq c_2\varepsilon^{-2}\|x_o-y_o\|^2 + c_2\|x_o-y_o\| + 2\delta c_1. \end{aligned} \tag{3.61}$$

Estimons maintenant les membres de cette inégalité. On a, avec (3.57),

$$\sup\varphi - \delta \leq \psi(x_o,y_o) = v(x_0) - w(y_0) - \tfrac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2, \tag{3.62}$$

et donc

$$\tfrac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2 \leq \sup v - \inf w + \delta - \sup\varphi. \tag{3.63}$$

Ceci implique $\|x_o - y_o\| \longrightarrow 0$ quand $\varepsilon \downarrow 0$. Notons c_v la constante de Hölder de v. Prenant ε assez petit, comme v est hölderienne, il vient $v(x_0) - v(y_0) \leq c_v\|x_0 - y_0\|^\gamma$. Choisissant $x = y = y_0$ dans (3.57), il vient après simplification et usage de (3.54),

$$\begin{array}{rcl} \frac{1}{2}\varepsilon^{-2}\|x_o - y_o\|^2 & \leq & v(x_0) - v(y_0) + \delta(\xi(x_0,y_0) - \xi(y_0,y_0)) \\ & \leq & c_v\|x_0 - y_0\|^\gamma + \delta\|x_0 - y_0\|. \end{array} \tag{3.64}$$

On peut sans perte de généralité supposer γ dans $]0,1[$, et donc

$$c_v\|x_0 - y_0\|^\gamma + \delta\|x_0 - y_0\| \leq \tfrac{1}{2}K|x_o - y_o|^\gamma \tag{3.65}$$

pour une certaine constante K indépendante de ε et δ. Avec (3.64), nous obtenons $\varepsilon^{-2}\|x_o - y_o\|^2 \leq K\|x_o - y_o\|^\gamma$ soit $\|x_o - y_o\| \leq K\varepsilon^{\frac{2}{2-\gamma}}$. Combinant avec (3.61), il vient

$$\lambda(v(x_0) - w(y_0)) \leq c_2K'\varepsilon^{\frac{2\gamma}{2-\gamma}} + 2\delta c_1 \tag{3.66}$$

pour un certain K' indépendant de ε et δ. Or

$$\sup(v - w) \leq \sup\varphi \leq \varphi(x_0,y_0) + \delta \leq v(x_0) - w(y_0) + \delta. \tag{3.67}$$

Il vient donc $\sup(v - w) \leq O(\varepsilon^{\frac{2\gamma}{2-\gamma}} + \delta)$. Faisant tendre ε et δ vers 0, nous obtenons la conclusion. ■

Remarque 3.34 La preuve ci-dessus a l'intérêt d'être très proche de celle de l'estimation d'erreur du schéma de discrétisation : voir la section 4.3.2.

3.5 Temps d'arrêt et commande impulsionnelle

Les résultats principaux de cette section concernent les problèmes de commande impulsionnelle. Afin de préparer les outils nécessaires à leur étude, nous étudions d'abord les problèmes avec décision d'arrêt, qui ont leur propre intérêt.

3.5.1 Problèmes avec temps d'arrêt

Nous considérons un problème de commande optimale dans lequel on peut s'arrêter à tout instant en payant un coût φ actualisé :

$$(P_x) \qquad \left\{ \begin{array}{l} \text{Min } \mathcal{V}(x,u,\theta) := \displaystyle\int_0^\theta \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + e^{-\lambda\theta}\varphi(y_{x,u}(\theta)) \\ \\ \dot{y}_{x,u}(t) = f(y_{x,u}(t),u(t)), \quad t \in [0,\theta[, \quad y_{x,u}(0) = x; \\ \\ u(t) \in U \text{ p.p. } t \in [0,\theta[, \end{array} \right.$$

avec $\theta \geq 0$ et φ fonction bornée et lipschitzienne.

On note $a \wedge b := \min(a,b)$, et χ_s vaut 1 si s est vrai, et 0 sinon.

La démonstration du théorème ci-dessous ne présente pas de difficulté.

Théorème 3.35 (Principe de Programmation Dynamique)
La fonction valeur $V(x)$ satisfait, pour tout $\tau > 0$:

$$V(x) = \inf_{(u,\theta)} \left(\int_0^{\tau\wedge\theta} \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \chi_{\tau<\theta}e^{-\lambda\tau}V(y_{x,u}(\tau^-)) + \chi_{\tau\geq\theta}e^{-\lambda\theta}\varphi(y_{x,u}(\theta^-)) \right), \tag{3.68}$$

où le minimum s'entend sous les contraintes $u(t) \in U$, p.p. $t \in [0,\tau]$, et $\theta \geq 0$.

L'équation HJB de ce problème est dite inéquation variationnelle, par analogie avec les problèmes de contact en mécanique :

$$\max[\lambda v - \mathcal{H}(x,Dv), v - \varphi(x)] = 0, \quad \text{pour tout } x \in \mathbb{R}^n. \tag{3.69}$$

Théorème 3.36 *La fonction valeur V est solution au sens de viscosité de* (3.69)*, au sens où, pour tout $x \in \mathbb{R}^n$:*

$$\max[\lambda V(x) - \mathcal{H}(x,p), V(x) - \varphi(x)] \leq 0 \quad (\text{resp. } \geq 0), \tag{3.70}$$

pour tout $p \in D^+v(x)$ (resp. $p \in D^-v(x)$).

Démonstration. Il est clair que $V(x) \leq \varphi(x)$ pour tout x, puisqu'une impulsion à l'instant initial est possible. Distinguons deux cas.
a) Si $V(x) < \varphi(x)$, puisque V et φ sont continues, il existe $\varepsilon > 0$ tel que $V(x') + \varepsilon < \varphi(x')$, pour tout x' appartenant à un voisinage $\mathcal{N}$ de x. Puisque f est bornée, on déduit que pour τ assez petit, toute stratégie optimale à ε près ne comporte pas d'impulsion pour $t \in [0,\tau]$. Le principe de programmation dynamique (3.6) est donc valable pour τ assez petit. La démonstration du lemme 3.20 s'applique donc; elle montre que (3.37) est satisfaite en x si $V - \Phi$ a un maximum (resp. minimum) local en x. On en déduit (3.70) en combinant avec le lemme 3.26.
b) $V(x) = \varphi(x)$, le second cas de (3.70) est trivialement satisfait. Reste à montrer que si $p \in D^+v(x)$, alors $\lambda V(x) - \mathcal{H}(x,p) \leq 0$. Puisque les stratégies sans impulsions sont possibles, on a

$$V(x) \leq \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + V(y_{x,u}(\tau))e^{-\lambda\tau} \right\}. \tag{3.71}$$

Il suffit alors de reprendre les calculs des lemmes 3.7 et 3.20, en tenant compte de l'inégalité dans (3.71), pour vérifier que (3.37) est satisfaite, si Φ une fonction différentiable en x, telle que $V - \Phi$ a un maximum (resp. minimum) local en x. On conclut avec le lemme 3.26. ■

Théorème 3.37 (Unicité forte) *Soient v une sous solution s.c.s. de* (3.69) *bornée supérieurement, w une sur solution s.c.i. de* (3.69) *bornée inférieurement. Si une de ces deux fonctions est hölderienne, alors $v(x) \leq w(x)$, pour tout $x \in \mathbb{R}^n$.*

Démonstration. Il suffit d'appliquer le théorème 3.31; la vérification des hypothèses (3.44)-(3.46) se fait sans difficultés. ■

3.5.2 Commande impulsionnelle

Dans de nombreux problèmes de commande optimale, on a la possibilité de faire changer l'état de manière discontinue, en payant un prix associé. Un exemple typique est celui de la gestion de stock, dans lequel une commande a un coût fixe (déplacement du camion) et un coût proportionnel à la quantité livrée (n'excédant pas la capacité du camion). La modification de l'état peut ne s'effectuer qu'après un certain délai (temps de livraison).

Nous allons nous limiter ici à la discussion de problèmes de commande optimale impusionnelle sans délai. La dynamique du système est régie par les relations suivantes :

$$\begin{array}{lcll} \dot{y}_{x,u}(t) & = & f(y_{x,u}(t),u(t)), & t \in]\theta_i,\theta_{i+1}[, \\ y_{x,u}(\theta_i^+) & = & y_{x,u}(\theta_i^-) + \xi_i, & i = 1,\ldots,N, \\ y_{x,u}(0) & = & x. & \end{array} \tag{3.72}$$

La *dynamique* $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ est supposée *lipschitzienne* et *bornée*, ainsi que l'*ensemble des commandes* $U \subset \mathbb{R}^m$, supposé compact, et le coefficient d'actualisation $\lambda > 0$. On convient de noter L_f la constante de Lipschitz de f, et de même pour les autres fonctions. La suite $\{\theta_i\}$, $i = 1,\ldots,N$, de *temps d'arrêt* positifs, est finie (on pose alors $\theta_{N+1} = +\infty$) ou non (on a alors $N = +\infty$), croissante et sans points d'accumulation, et $\theta_0 = 0$. Les impulsions ξ_i appartiennent à $\Xi \subset \mathbb{R}^n$. Les suites θ et ξ font partie de la commande. Ainsi *l'état* $y_{x,u}(t)$ appartient à $\mathbb{R}^n$ et la commande, ou contrôle, $u(t)$ appartient à $\mathbb{R}^m$. Le critère à minimiser se décompose en une intégrale d'un *coût distribué* et une somme de *coûts de transition* :

$$\mathcal{V}(x,u,\theta,\xi) := \int_0^\infty \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + \sum_{i=1}^{N}(c_0 + c(\xi_i))e^{-\lambda\theta_i}. \tag{3.73}$$

Le coût distribué $\ell : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ est supposé *lipschitzien* et *borné*. Le coût de transition est $c_0 + c(\xi_i)$. La constante $c_0 > 0$ représente un coût fixe, et la fonction continue $c : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ est telle que $c(0) = 0$ et

$$c(\xi_1 + \xi_2) \leq c(\xi_1) + c(\xi_2), \quad \forall\, \xi_1,\xi_2 \in \mathbb{R}^n. \tag{3.74}$$

La stricte positivité de c_0 donne une borne sur le nombre d'impulsions d'une stratégie sous optimale sur un intervalle de temps fini, et la relation précédente implique qu'il n'est pas restrictif d'imposer que les instants θ_i soient tous différents. Le problème à résoudre est

$$(P_x) \qquad \operatorname*{Min}_{(u,\theta,\xi)} \mathcal{V}(x,u,\theta,\xi) \qquad \text{soumis à (3.72)}; \quad u(t) \in U, \text{ p.p. } t \in [0,+\infty[.$$

La *valeur* de ce problème (infimum du critère sur les commandes admissibles) est notée $V(x)$.

Nous allons dans un premier temps établir un résultat de régularité de la fonction valeur ainsi que le principe de programmation dynamique pour un problème sans impulsion.

Rappelons la notation $\mathcal{V}(x,u,\theta,\xi)$ du coût associé à une commande (voir (3.73)).

Proposition 3.38 *La fonction V appartient à $BUC(\mathbb{R}^n)$.*

Démonstration. a) Montrons que V est bornée. Soit la commande constante $u(t) = u_0$, où $u_0 \in U$, sans impulsion. Alors

$$V(x) \leq \int_0^\infty \ell(y_{x,u}(t),u_0)e^{-\lambda t}\mathrm{d}t \leq \lambda^{-1}\|\ell\|_\infty.$$

D'autre part, puisque le coût de transition est positif, on a pour tout commande (u,θ,ξ)

$$V(x) \geq \int_0^\infty \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t \geq -\lambda^{-1}\|\ell\|_\infty,$$

et donc $\|V\|_\infty \leq \lambda^{-1}\|\ell\|_\infty$.
b) Montrons que V est uniformément continue. On a

$$V(x') - V(x) \leq \sup_{(u,\theta,\xi)} \{\mathcal{V}(x',u,\theta,\xi) - \mathcal{V}(x,u,\theta,\xi)\},$$

donc après simplification des coûts de transition,

$$V(x') - V(x) \leq \sup_{(u,\theta,\xi)} \left\{ \int_0^\infty [\ell(y_{x'}(t),u(t)) - \ell(y_{x,u}(t),u(t))]e^{-\lambda t}\mathrm{d}t \right\}.$$

Soient $y_{x'}$ et $y_{x,u}$ deux trajectoires associées à la même commande (u,θ,ξ). Comme dans le cas sans impulsion, on a $|y_{x'} - y_{x,u}| \leq |x' - x|e^{\lambda_0 t}$, où λ_0 est défini par (3.27). On peut alors finir la démonstration de manière analogue à celle du lemme 3.13. ■

Théorème 3.39 (Principe de Programmation Dynamique)
La fonction valeur $V(x)$ satisfait, pour tout $\tau > 0$:

$$V(x) = \inf_{(u,\theta,\xi)} \left(\int_0^\tau \ell(y_{x,u}(t),u(t))\mathrm{d}t + \sum_{i=1}^{N} (c_0 + c(\xi_i))e^{-\lambda\theta_i} + e^{-\lambda\tau}V(y_{x,u}(\tau^-)) \right), \quad (3.75)$$

où le minimum s'entend sous les contrainte $u(t) \in U$, p.p. $t \in [0,\tau]$, les θ_i sont strictement croissants, et $\theta_N < \tau$.

Démonstration. La démonstration est similaire à celle du théorème 3.3. ■

Définissons l'opérateur qui à une fonction $w(\cdot)$ associe la valeur optimale après impulsion, noté

$$Mw(x) := \inf_{\xi \in \mathbb{R}^n} \{w(x+\xi) + c_0 + c(\xi)\}. \quad (3.76)$$

Lemme 3.40 *L'opérateur M est non expansif*[1] *pour la norme du max, et c'est une application de $BUC(\mathbb{R}^n)$ vers lui même.*

Démonstration. Soient w et w' dans $L_\infty(\mathbb{R}^n)$. Puisque $c(\cdot) \geq 0$ et $c(0) = 0$, on a

$$c_0 - \|w\|_\infty \leq Mw(x) \leq c_0 + w(x) \leq c_0 + \|w\|_\infty.$$

1. C'est à dire lipschitzien de constante 1.

De l'inégalité

$$Mw' - Mw \leq \sup_{\xi \in I\!R^n} \{w'(x+\xi) - w(x+\xi)\} = \|w' - w\|_\infty,$$

il résulte que M est non expansif pour la norme du max. En particulier, soit $w \in BUC(I\!R^n)$. Quand $y \longrightarrow 0$ dans $I\!R^n$, la fonction translatée $w_y(x) := w(x+y)$ tend uniformément vers w, donc $Mw_y - Mw \longrightarrow 0$ uniformément. Or

$$Mw_y(x) := \inf_{\xi \in I\!R^n} \{w(x+y+\xi) + c_0 + c(\xi)\} = Mw(x+y) = (Mw)_y(x),$$

donc $(Mw)_y - Mw$ tend uniformément vers 0. Ceci signifie que Mw est uniformément continue, d'où le lemme. ■

Théorème 3.41 *La fonction valeur V est solution au sens de viscosité de l'équation*

$$\max[\lambda V(x) - \mathcal{H}(x,DV(x)),V(x) - MV(x)] = 0, \tag{3.77}$$

au sens où

$$\max[\lambda V(x) - \mathcal{H}(x,p),V(x) - MV(x)] \leq 0 \quad (\text{resp. } \geq 0), \tag{3.78}$$

pour tout $p \in D^+v(x)$ (resp. $p \in D^-v(x)$).

Démonstration. La démonstration se réduit à celle du théorème 3.36, en identifiant la décision d'impulsion à une décision d'arrêt de coût $\varphi(x) := MV(x)$. ■

Pour le résultat d'unicité on se reportera à Barles [5, Section 3.2.2].

3.6 Notes

La référence classique sur la programmation dynamique est R. Bellman [7]. Une présentation simple, avec de nombreux exemples est donnée dans D. Bertsekas [8].

L'approche par solution de viscosité est dûe à Crandall et Lions [15]. Barles [5] fournit une introduction à ce sujet. Notons aussi l'ouvrage de Bardi et Capuzzo-Dolcetta [4].

Chapitre 4

Résolution numérique de l'équation HJB

Ce chapitre discute la résolution numérique du problème (P_x) du chapitre 3, en discrétisant l'équation HJB. Nous supposerons dans ce chapitre f et ℓ lipschitziennes et bornées, $\lambda > 0$, U compact non vide, et C fermé (supposé vide dans certains énoncés).

Introduisons deux espaces de fonctions, l'ensemble $B(\mathbb{R}^n)$ l'ensemble des fonctions bornées $\mathbb{R}^n \longrightarrow \mathbb{R}^n$, muni de la norme

$$\|v\|_\infty := \sup_{x\in\mathbb{R}^n} |v(x)| \tag{4.1}$$

qui en fait un espace de Banach, (à ne pas confondre avec $\mathcal{L}^\infty(\mathbb{R}^n)$, l'espace des fonction définies presque partout, essentiellement bornées) et l'espace

$$BUC(\mathbb{R}^n) := \{ \text{ Fonctions bornées, uniformément continues: } \mathbb{R}^n \longrightarrow \mathbb{R}\}\,. \tag{4.2}$$

On pose

$$a_+ := \max(a,0); \quad a_- := \min(a,0). \tag{4.3}$$

4.1 Motivation : problème continu

Le but de cette section est d'analyser une variante du principe de programmation dynamique qui se formule comme un opérateur de point fixe contractant, dit *itération sur les valeurs*. L'algorithme convergent qui en découle n'est pas implémentable sur ordinateur, puisqu'il s'applique au problème continu. Cependant, on obtiendra des algorithmes effectifs après discrétisation de l'espace d'état.

Rappelons l'expression du principe de programmation dynamique (théorème 3.3) : si $x \in \mathbb{R}^n \setminus C$, et $\tau \in]0,T(x)[$, alors

$$V(x) := \inf_{u\in\mathcal{U}} \left\{ \int_0^\tau \ell(y_{x,u}(t),u(t))e^{-\lambda t}\mathrm{d}t + V(y_{x,u}(\tau))e^{-\lambda\tau} \right\}. \tag{4.4}$$

Nous allons voir une variante de cette formulation qui permet de définir un opérateur dans tout l'espace. On note $t_1 \wedge t_2 := \min(t_1,t_2)$ et

Théorème 4.1 (Principe de Programmation Dynamique II) *Pour tout $x \in \mathbb{R}^n$ et $\tau > 0$, on a $V(x) = \mathcal{M}^\tau V(x)$, où*

$$\mathcal{M}^\tau v(x) := \inf_{(u,T)} \left\{ \int_0^{\tau \wedge T} \ell(y_{x,u}(t),u(t)) e^{-\lambda t} \mathrm{d}t + \chi_{\tau<T} v(y_{x,u}(\tau)) e^{-\lambda \tau} \right\}, \tag{4.5}$$

l'infimum portant sur les couples (u,T) admissibles.

Démonstration. La démonstration est similaire à celle du théorème 3.3. ■

Proposition 4.2 *Pour tout $\tau > 0$, l'opérateur $\mathcal{M}^\tau$ est monotone croissant de $B(\mathbb{R}^n)$ dans lui même, et c'est une contraction de rapport $e^{-\lambda\tau}$.*

Démonstration. Sachant que ℓ est borné, et donc

$$\left| \int_0^{\tau \wedge T} \ell(y_{x,u}(t),u(t)) e^{-\lambda t} \mathrm{d}t \right| \leq \lambda^{-1} \|\ell\|_\infty, \tag{4.6}$$

il est clair que $\mathcal{M}^\tau$ applique $B(\mathbb{R}^n)$ dans lui-même. La monotonie de $\mathcal{M}^\tau$ est immédiate. Soient v et v' deux fonctions de $B(\mathbb{R}^n)$. Par une majoration similaire à (1.26), il vient

$$|(Tv')(x) - (Tv)(x)| \leq \sup_{(u,T)} e^{-\lambda\tau} \chi_{\tau<T} \, |v'(y_{x,u}(\tau)) - v(y_{x,u}(\tau))| \leq e^{-\lambda\tau} \|v' - v\|_\infty,$$

d'où la conclusion. ■

On déduit du résultat précédent l'"algorithme" (en espace d'état continu) d'*itérations sur les valeurs* ci-dessous.

Corollaire 4.3 *On peut calculer V par l'algorithme de point fixe suivant : fixer $\tau > 0$, et former la suite $V_{k+1} = \mathcal{M}^\tau V_k$, en partant de $V_0 \in B(\mathbb{R}^n)$ quelconque. Cette suite vérifie*

$$\|V_{k+1} - V\|_\infty \leq e^{-k\lambda\tau} \|V_k - V\|_\infty. \tag{4.7}$$

Nous allons maintenant formuler des schémas numériques de discrétisation de l'équation HJB. Ces schémas se reformulent comme des points fixes d'opérateurs contractants qui s'interprètent comme des discrétisations de l'opérateur $\mathcal{M}^\tau$.

4.2 Schémas décentrés et extensions

4.2.1 Dimension d'espace $n = 1$

Nous allons chercher à discrétiser l'équation HJB en remplaçant la dérivée en espace par une différence finie. Soit $\Delta x > 0$ le pas d'espace. On note $x_j := j\Delta x$. L'espace discret est $\{x_j, j \in \mathbb{Z}\} = \Delta x \mathbb{Z}$.

On pose $\Omega := \mathbb{R}^n \setminus C$, et on notera $C_{\Delta x}$ et $\Omega_{\Delta x}$ les discrétisations des ensembles C et Ω, respectivement; ces deux ensembles forment une partition de $\Delta x \mathbb{Z}$. Nous supposerons que

$$C_{\Delta x} \text{ converge vers } C, \text{ au sens de la } \textit{distance de Hausdorff}. \tag{4.8}$$

on rappelle que, si C_1 et C_2 sont deux parties de $\mathbb{R}^n$, leur distance de Hausdorff est

$$\mathrm{dist}(C_1,C_2) := \max\left(\sup_{c_1\in C_1}\mathrm{dist}(c_1,C_2),\ \sup_{c_2\in C_2}\mathrm{dist}(c_2,C_1)\right). \tag{4.9}$$

On désire approcher $V(x_j)$ par la quantité v_j. Notons

$$D^d v_j = \frac{v_{j+1}-v_j}{\Delta x}, \qquad D^g v_j = \frac{v_j - v_{j-1}}{\Delta x}, \qquad D^0 v_j = \frac{v_{j+1}-v_{j-1}}{2\Delta x}, \tag{4.10}$$

les *différences divisées à droite, à gauche et centrées*, respectivement. Laquelle faut-il prendre pour discrétiser l'équation HJB ?

L'idée essentielle est de s'appuyer sur le principe de programmation dynamique, qui relie les valeurs de V en x et en les points voisins dans la direction de $f(x,u)$. *Il convient donc de décentrer à droite si $f(x,u)$ est positive, et à gauche sinon*[1]. On obtient ainsi le *schéma décentré*

$$\left\{\begin{array}{ll} \lambda v_j & = \inf\limits_{u\in U}\left\{\ell(x_j,u) + f(x_j,u)_+\dfrac{v_{j+1}-v_j}{\Delta x} + |f(x_j,u)_-|\dfrac{v_{j-1}-v_j}{\Delta x}\right\},\ j\in\Omega_{\Delta x},\\ v_j & = 0,\ j\in C_{\Delta x}. \end{array}\right. \tag{4.11}$$

Exemple 4.4 Soit le problème de transfert en temps minimal à 0, avec la dynamique $\dot{x}=u$, et la contrainte $-1\le u\le 1$. Il est naturel de prendre $C_{\Delta x}=\{0\}$. La fonction à minimiser est linéaire par morceaux : elle atteint son minimum en 0, -1 ou 1. Le schéma décentré s'écrit donc, pour $j\neq 0$,

$$\lambda v_j = 1 + \min\left\{0, \frac{v_{j-1}-v_j}{\Delta x}, \frac{v_{j+1}-v_j}{\Delta x}\right\}, \tag{4.12}$$

ou encore

$$(1+\lambda\Delta x)v_j = \Delta x + \min\left\{v_{j-1}, v_j, v_{j+1}\right\}, \tag{4.13}$$

formule à partir de laquelle on peut expliciter la valeur de v_j pour tout $j\in\mathbb{Z}$.

4.2.2 Forme de point fixe contractant

Nous allons réécrire le schéma (4.11) sous une forme de point fixe contractant, ce qui permettra de vérifier qu'il a une solution unique. Cette réécriture fait apparaître un pas de temps $\Delta t>0$ *fictif*.

Multipliant (4.11) par $\Delta t>0$, ajoutant v_j à chaque membre, et divisant par $(1+\lambda\Delta t)$, il vient

$$\begin{array}{rl} v_j = & (1+\lambda\Delta t)^{-1}\inf\limits_{u\in U}\left\{\Delta t\ell(x_j,u) + \left(1-\dfrac{\Delta t}{\Delta x}|f(x_j,u)|\right)v_j\right.\\ & \left. +\dfrac{\Delta t}{\Delta x}|f(x_j,u)_-|v_{j-1} + \dfrac{\Delta t}{\Delta x}f(x_j,u)_+ v_{j+1}\right\}. \end{array} \tag{4.14}$$

1. Ce qui traduit le fait que la prise de décision optimale nécessite d'envisager les conséquences de ses actes.

Nous allons vérifier que, pour Δt assez petit, (4.14) est une équation de point fixe monotone et contractant. Notons $N(f)$ la norme infinie de f restreinte à $\mathbb{R}^n \times U$,

$$N(f) := \sup_x \sup_{u\in U} |f(x,u)|, \tag{4.15}$$

et considérons la *condition de stabilité*

$$\frac{\Delta t}{\Delta x} N(f) \leq 1. \tag{4.16}$$

Remarque 4.5 Si (4.16) est satisfait, la combinaison linéaire de v_{j-1}, v_j, et v_{j+1} apparaissant dans (4.14) est tout simplement une formule d'*interpolation linéaire* de la valeur de v au point $x_j + \Delta t f(x_j,u)$. Ceci permet d'interpréter (4.14) comme une discrétisation du principe de programmation dynamique (4.5), le pas Δt correspondant à τ.

Proposition 4.6 (i) *Le schéma* (4.14) *possède une solution unique, telle que*

$$\|v\|_\infty \leq \lambda^{-1} \|\ell\|_\infty. \tag{4.17}$$

(ii) *Si* Δt *vérifie la condition de stabilité* (4.16)*, alors* (4.14) *est une équation de point fixe contractant pour la norme uniforme*

$$\|v_j\|_\infty := \sup\{|v_j|, \quad j \in \mathbb{Z}\}, \tag{4.18}$$

de rapport de contraction $(1+\lambda\Delta t)^{-1}$.

Démonstration. Soit $\mathcal{N}^{\Delta t}$ l'opérateur de point fixe du membre de droite de (4.14). Notons

$$\tilde f(x_j,u) := \frac{\Delta t}{\Delta x} f(x_j,u)$$

qui représente une mise à l'échelle de la dynamique. Utilisant (1.26), et le fait que (4.16) implique $1 - |\tilde f(x_j,u)| \geq 0$, il vient

$$\begin{aligned} |(\mathcal{N}^{\Delta t} v')_j - (\mathcal{N}^{\Delta t} v)_j| \leq \quad & (1+\lambda\Delta t)^{-1} \sup_{u\in\mathcal{U}} \Big\{ (1 - |\tilde f(x_j,u)|)|v'_j - v_j| \\ & + |\tilde f(x_j,u)_-||v'_{j-1} - v_{j-1}| + \tilde f(x_j,u)_+ |v'_{j+1} - v_{j+1}| \Big\}. \end{aligned} \tag{4.19}$$

Majorant $|v'_i - v_i|$, pour $i = j-1, j, j+1$, par $\|v' - v\|_\infty$ on obtient

$$|(\mathcal{N}^{\Delta t} v')_j - (\mathcal{N}^{\Delta t} v)_j| \leq (1+\lambda\Delta t)^{-1} \|v' - v\|_\infty \tag{4.20}$$

d'où (ii). L'existence et l'unicité sont conséquence directe de (ii). Enfin soit v la solution de (4.14); utilisant (4.14), pour tout $j \in \mathbb{Z}$, il vient

$$|v_j| \leq (1+\lambda\Delta t)^{-1} \left[\|\ell\|_\infty + \|v\|_\infty \right],$$

d'où (4.17). ■

Remarque 4.7 Rien n'empêche de considérer l'opérateur obtenu en prenant dans (4.14) un pas de temps Δt_j dépendant de l'indice d'espace; cela peut être avantageux d'un point de vue numérique. La condition de stabilité devient

$$\frac{\Delta t_j}{\Delta x} \sup_{u\in U} |f(x_j,u)| \leq 1, \quad \text{pour tout} \quad j \in \mathbb{Z}, \tag{4.21}$$

et le rapport de contraction est $(1 + \lambda \inf_j \Delta t_j)^{-1}$.

Remarque 4.8 (i) On appelle CFL (Courant-Friedrich-Levy) la quantité $\frac{\Delta t}{\Delta x}N(f)$. La condition de stabilité (4.16) peut donc s'énoncer ainsi : le CFL ne doit pas dépasser 1.
(ii) La condition de stabilité assure que, pendant le pas de temps Δt, le système dynamique varie au plus de Δx. Autrement dit, l'information se propage au moins aussi vite dans le schéma numérique que dans le problème d'origine.

Remarque 4.9 Les expressions (4.14) permettent, si la condition de stabilité (4.16) est satisfaite, d'interpréter le schéma décentré comme le principe de programmation dynamique pour le problème de commande optimale d'une chaîne de Markov : voir la remarque 5.19.

Remarque 4.10 Le coefficient de contraction, assurant la convergence de l'algorithme, est $(1+\lambda\Delta t)^{-1}$. Compte-tenu de la condition de stabilité, on voit que la constante optimale, obtenue pour $CFL = 1$, vaut $(1+\lambda\Delta x\, N(f)^{-1})^{-1}$. La convergence devient donc très lente quand $\Delta x \downarrow 0$.
D'autres algorithmes sont possibles, en particulier l'itération sur les politiques (voir la section 5.1).

4.2.3 Dimension d'espace quelconque

Le schéma décentré monodimensionnel peut se généraliser de multiples manières dans le cas où $n > 1$. Donnons seulement la plus naïve.

Soient $h_1,\ldots,h_n$ les pas d'espace, strictement positifs. A $j \in \mathbb{Z}^n$, on associe le point $x_j \in I\!R^n$ de coordonnées $j_i h_i$. Notons $e_1,\ldots,e_n$ la base naturelle de $I\!R^n$. Le décentrage se fait, pour chaque composante, suivant le signe de $f_i(x_j,u)$; on obtient le schéma suivant :

$$\begin{array}{ll} \lambda v_j = & \inf\limits_{u\in U}\left\{\ell(x_j,u)+\sum\limits_{i=1}^{n}\left(f_i(x_j,u)_+\frac{v_{j+e_i}-v_j}{h_i}+|f_i(x_j,u)_-|\frac{v_{j-e_i}-v_j}{h_i}\right)\right\}, \\ & j\in\Omega_\Delta; \\ v_j = & 0,\ i\in C_\Delta. \end{array} \tag{4.22}$$

Comme dans le cas monodimensionnel, il convient de multiplier (4.11) par un pas de temps fictif qu'on notera h_0, et d'ajouter v_j à chaque membre, ce qui donne

$$\begin{aligned} v_j = \ & (1+\lambda h_0)^{-1}\inf_{u\in U}\left\{h_0\ell(x_j,u)+\left(1-\sum_{i=1}^{n}\frac{h_0}{h_i}|f_i(x_j,u)|\right)v_j\right. \\ & \left.+\sum_{i=1}^{n}\frac{h_0}{h_i}f_i(x_j,u)_+v_{j+e_i}+\sum_{i=1}^{n}\frac{h_0}{h_i}|f_i(x_j,u)_-|v_{j-e_i}\right\}. \end{aligned} \tag{4.23}$$

Proposition 4.11 (i) *Le schéma* (4.22) *possède une solution unique, telle que*

$$\|v\|_\infty \le \lambda^{-1}\|\ell\|_\infty. \tag{4.24}$$

(ii) *Si* h_0 *vérifie la* condition de stabilité

$$h_0\sum_{i=1}^{n}\sup_x\sup_{u\in U}\frac{|f_i(x,u)|}{h_i}\le 1, \tag{4.25}$$

alors (4.23) *est une équation de point fixe contractant pour la norme uniforme, de rapport de contraction* $(1+\lambda h_0)^{-1}$.

Démonstration. La démonstration est similaire à celle de la proposition 4.6. ■

Exemple 4.12 Soit le système dynamique $\dot{x} = u$, avec $U = [-1,1]^n$. On a dans ce cas $\sup_x \sup_{u\in U} |f_i(x,u)| = 1$, pour $i = 1,\ldots,n$, et la condition de stabilité se réduit à

$$\frac{1}{h_0} \geq \sum_{i=1}^{n} \frac{1}{h_i}. \tag{4.26}$$

Autrement dit, le pas de temps maximal est dans ce cas la *moyenne harmonique* des pas d'espace.

Remarque 4.13 Le schéma aux différences finies (4.22) fait intervenir le point j et les 2^n points voisins de la grille, obtenus en changeant une seule coordonnée de j de ± 1. Si $n = 2$ on parle d'un schéma à 5 points.

Remarque 4.14 On peut étendre la remarque 4.5 : le schéma, sous la forme (4.23), est très proche du principe de programmation dynamique (4.5). Sous la condition de stabilité (4.25), les poids des $v_{j\pm e_i}$ s'interprètent comme les coordonnées barycentriques du point $x_j + \Delta t f(x_j,u)$.

Remarque 4.15 Si $|f(x,u)|$ peut prendre des valeurs élevées, la condition de stabilité oblige à prendre h_0 très petit. Pour éviter cela, on peut adopter des schémas faisant intervenir des points plus éloignés. Nous en donnons un exemple dans la section suivante.

4.2.4 Discrétisation par triangulation

Donnons maintenant un procédé de discrétisation spatiale qui constitue une alternative intéressante aux méthodes de différences finies. Un *simplexe* de $\mathbb{R}^n$ est un polyèdre formé par l'ensemble des combinaisons convexes de $k+1$ points (appelés sommets) non contenus dans un hyperplan. Autrement dit, un simplexe est de la forme

$$\left\{ \sum_{i=1}^{k+1} \alpha_i x_i; \quad \alpha \geq 0, \quad \sum_{i=1}^{k+1} \alpha_i = 1 \right\},$$

où $x_1,\ldots,x_{k+1}$, sont des points de $\mathbb{R}^n$ non contenus dans un hyperplan. On appelle *face* du simplexe l'ensemble des combinaisons convexes de n des points; la frontière du simplexe est l'union de ses $n+1$ faces.

Considérons une *triangulation régulière* de $\mathbb{R}^n$ réalisée par une famille de simplexes S_J, $J \in \mathbb{N}$. Autrement dit, l'union de ces simplexes est égale à $\mathbb{R}^n$, et l'intersection de deux simplexes est égale à une face de chacun des deux simplexes. On note $\mathcal{S}$ l'ensemble des simplexes, et $L_{\mathcal{S}}$ l'espace des fonctions linéaires sur chaque simplexes. Les fonctions de $L_{\mathcal{S}}$ sont déterminées par leur valeur aux sommets des simplexe. On a pour une telle fonction v

$$v(x_i + h_0 f(x_i,u)) = \sum_{j=1}^{k+1} \alpha_i(u) v(x_j), \tag{4.27}$$

où les $\alpha_i(u)$ sont les coefficients de la combinaison convexe représentant le point $x_i + h_0 f(x_i,u)$ dans un des simplexes auquel il appartient (coefficients barycentriques), tels que

$$\begin{array}{l} x_j + h_0 f(x_j,u) = \sum_{j=1}^{k+1} \alpha_i(u) x_j; \\ 0 \leq \alpha_i(u) \leq 1; \quad \sum_{j=1}^{k+1} \alpha_i(u) = 1. \end{array} \tag{4.28}$$

Le *schéma associé à la triangulation* est obtenu en écrivant une sorte de principe de programmation dynamique discret aux sommets de la triangulation :

$$\left\{ \begin{array}{lcll} v_j & = & (1+\lambda h_0)^{-1} \inf_{u\in U} \left\{ h_0 \ell(x_j,u) + v(x_j + h_0 f(x_j,u)) \right\}, & j \in \Omega_{\mathcal{S}}, \\ v_j & = & 0, & j \in C_{\mathcal{S}}, \end{array} \right. \tag{4.29}$$

où $\Omega_{\mathcal{S}}$ et $C_{\mathcal{S}}$ sont les ensembles de sommets considérés hors de et dans la cible.

On peut réécrire (4.29) sous la forme $v = \mathcal{M}^{\mathcal{S}} v$, avec

$$\left\{ \begin{array}{lcll} \mathcal{M}^{\mathcal{S}}_j & := & (1+\lambda h_0)^{-1} \inf_{u\in U} \left\{ h_0\, \ell(x_j,u) + v(x_j + h_0 f(x_j,u)) \right\}, & j \in \Omega_{\mathcal{S}}, \\ \mathcal{M}^{\mathcal{S}}_j & := & 0, & j \in C_{\mathcal{S}}. \end{array} \right. \tag{4.30}$$

L'opérateur $\mathcal{M}^{\mathcal{S}}$ est une contraction de rapport $(1+\lambda h_0)^{-1}$ pour la norme du max, ce qui permet de vérifier que (4.29) a un point fixe unique uniformément borné par $\lambda^{-1}\|\ell\|_\infty$.

Remarque 4.16 Ce schéma permet de raffiner la discrétisation dans une région donnée, ce qui n'est pas facile avec les différences finies. De plus il ne comporte pas de condition restrictive sur le pas de temps fictif, de type CFL.

En revanche son implémentation est plus complexe; un point délicat est de reconnaître rapidement dans quel triangle se trouve le point $x_j + h_0 f(x_j,u)$.

De plus, si un grand pas de temps permet une convergence rapide du point fixe, il suppose aussi un très grand nombre de triangles.

4.3 Convergence des schémas et essais numériques

Nous donnons deux résultats de convergence des schémas de différences finies. Celui de la section 4.3.1 établit la convergence uniforme sur les compacts. Celui de la section 4.3.2 fournit une estimation d'erreur dans le cas où la cible est vide.

On suppose dans l'ensemble de la section que la cible est vide. On sait alors que la valeur est uniformément continue (lemme 3.13).

4.3.1 Un argument élémentaire de convergence

Notons V la fonction valeur, et $v^{\Delta x}$ la solution obtenue pour un pas d'espace Δx. La démonstration utilise de façon essentielle les limites inférieure et supérieure

$$\bar{v}(x) := \limsup_{\substack{j\Delta x \to x \\ \Delta x \downarrow 0}} v_j^{\Delta x}, \qquad \underline{v} := \liminf_{\substack{j\Delta x \to x \\ \Delta x \downarrow 0}} v_j^{\Delta x}\ . \tag{4.31}$$

L'énoncé ci-dessous se limite au cas $n = 1$, mais la preuve s'étend facilement au schéma aux différences finies pour n quelconque, ainsi qu'à la discrétisation par triangulation.

Théorème 4.17 (Convergence du schéma décentré) *Si $C = \emptyset$, alors :*
(i) *Les fonctions $\bar{v}$ et $\underline{v}$ sont égales à la fonction valeur V du problème "standard" de commande optimale en horizon infini (P_x).*
(ii) *La convergence des valeurs discrètes est uniforme sur tout compact.*

Démonstration. Nous savons que les solutions discrètes sont uniformément bornées par $\lambda^{-1}\|\ell\|_\infty$. Les fonctions $\bar{v}$ et $\underline{v}$ sont donc bornées. Notons bien que ces fonctions sont définies en chaque point, et non presque partout. De la définition de $\bar{v}$ et $\underline{v}$, on déduit aisément que $\bar{v}$ est semi continu supérieurement (s.c.s.), et $\underline{v}$ est semi continu inférieurement (s.c.i.).

La définition de $\bar{v}$ et $\underline{v}$ implique $\bar{v} \geq \underline{v}$. Il suffit alors de montrer que $\bar{v}$ est sous solution, et que $\underline{v}$ est sur solution. En effet, d'après le principe d'unicité forte, ceci implique $\bar{v} \leq V \leq \underline{v}$, d'où l'égalité de ces trois fonctions. La convergence des valeurs discrètes uniforme sur tout compact se vérifie alors facilement avec une preuve par l'absurde.

On se contentera de montrer que $\bar{v}$ est sous solution, le fait que $\underline{v}$ soit sur solution se démontrant de manière analogue.

Soit x_0 un point de maximum local de $\bar{v} - \Phi$. Il existe donc $r > 0$ tel que $\bar{v} - \Phi$ atteint en x_0 son maximum sur $\bar{B}(x_0,r)$ (la boule fermée de centre x_0 et rayon r). Ajoutant $\|x - x_0\|^2$ à Φ si nécessaire, on peut supposer que

$$\bar{v}(x_0) = \Phi(x_0) \quad \text{et} \quad \bar{v}(x) < \Phi(x) \ \text{ si } \ x \neq x_0, \ x \in \bar{B}(x_0,r). \tag{4.32}$$

Par définition de $\bar{v}(x_0)$, il existe des suites $\Delta x_k \downarrow 0$ et $j_k \in \mathbb{Z}$ telles que

$$j_k \Delta x_k \longrightarrow x_0 \quad \text{et} \quad \bar{v}(x_0) = \lim v_{j_k}^{\Delta x_k}.$$

Soit $i_k \in \mathbb{Z}$ tel que

$$v_{j'}^{\Delta x_k} - \Phi(x_{j'}) \leq v_{i_k}^{\Delta x_k} - \Phi(i_k \Delta x_k), \ j' \neq j; \ \ x_{j'} \in \bar{B}(x_0,r). \tag{4.33}$$

Extrayant si nécessaire une sous suite, on peut supposer que $i_k \Delta x \longrightarrow \bar{x}$, et nécessairement $|\bar{x} - x_0| \leq r$. Par définition de $\bar{v}$, on a :

$$\bar{v}(x_0) - \Phi(x_0) = \lim_k v_{j_k}^{\Delta x_k} - \Phi(j_k \Delta x) \leq \limsup_k v_{i_k}^{\Delta x_k} - \Phi(i_k \Delta x) \leq \bar{v}(\bar{x}) - \Phi(\bar{x}). \tag{4.34}$$

Ceci, joint à (4.32), montre que $\bar{x} = x_0$ et aussi $\bar{v}(x_0) = \lim_k v_{i_k}^{\Delta x_k}$. Combinant (4.11) et (4.33), il vient

$$\lambda v_{i_k}^{\Delta x_k} \leq \inf_{u \in U} \left\{ \ell(x_{i_k},u) + f(x_{i_k},u)_+ \frac{\Phi((i_k+1)\Delta x) - \Phi(i_k \Delta x)}{\Delta x} + |f(x_{i_k},u)_-| \frac{\Phi((i_k-1)\Delta x) - \Phi(i_k \Delta x)}{\Delta x} \right\}.$$

Puisque $i_k \Delta x \longrightarrow \bar{x} = x_0$, passant à la limite quand $\Delta x \downarrow 0$, on obtient :

$$\lambda \bar{v}(x_0) + \mathcal{H}(x_0, D\Phi(x_0)) \leq 0, \tag{4.35}$$

ce qui prouve que $\bar{v}$ est sous solution. ■

4.3.2 Estimation d'erreur

Dans cette section on suppose que la cible est vide, et on donne une estimation de l'erreur de discrétisation. On note par v^h la fonction telle que $v^h(x_j) = v_j$ où $\{v_j; j \in \mathbb{Z}^n\}$ est la solution du schéma avec les pas $h_1,\ldots,h_n$. On remarquera le lien entre la démonstration ci-dessous et celle du théorème 3.31[2].

Théorème 4.18 *Soit $\gamma \in]0,1[$ une constante de Hölder de V. Alors il existe $C > 0$ tel que, pour tout $(h_1,\ldots,h_n) \in (\mathbb{R}_+^*)^n$, on a*

$$\sup_{x\in\mathbb{R}^n} |V(x) - v^h(x)| \le C \left(\max_{1\le i\le n} h_i\right)^{\gamma/2}. \tag{4.36}$$

Démonstration. Soit $0 < \varepsilon < 1$; posons

$$\beta_\varepsilon(x) := -\varepsilon^{-2}|x|^2, \quad x \in \mathbb{R}^n. \tag{4.37}$$

ainsi que

$$\varphi(x,y) := v^h(x) - V(y) + \beta_\varepsilon(x-y), \quad (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n.$$

Soit $\delta \in (0,1)$, et notons $\mathbb{R}_h^n = \{(j_1h_1,\ldots,j_nh_n); j \in \mathbb{Z}^n\}$.. Puisque V et v^h sont bornées, il existe (x_1,y_1) dans $\mathbb{R}_h^n \times \mathbb{R}^n$ tel que

$$\varphi(x_1,y_1) > \sup \varphi - \delta. \tag{4.38}$$

Soit $\xi \in C_0^\infty(\mathbb{R}^{2n})$ tel que

$$\xi(x_1,y_1) = 1, \qquad 0 \le \xi \le 1, \quad |D\xi| \le 1, \tag{4.39}$$

et posons

$$\psi(x,y) = \varphi(x,y) + \delta\xi(x,y), \quad (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n. \tag{4.40}$$

Alors ψ atteint son maximum sur $\mathbb{R}_h^n \times \mathbb{R}^n$ en un point (x_o,y_o) du support de ξ. Autrement dit,

$$\psi(x_o,y_o) \ge \psi(x,y), \quad \text{pour tout } (x,y) \in \mathbb{R}_h^n \times \mathbb{R}^n. \tag{4.41}$$

En particulier, $y \longrightarrow -\psi(x_o,y)$ atteint son minimum en y_o. Par définition d'une solution de viscosité, il existe $u^* \in U$ tel que

$$\lambda V(y_o) + f(y_o,u^*).(D\beta_\varepsilon(x_o - y_o) - \delta D_y\xi(x_o,y_o)) - \ell(y_o,u^*) \ge 0. \tag{4.42}$$

Puisque x_o appartient à $\mathbb{R}_h^n$, il existe $j \in \mathbb{Z}^n$ tel que $x_o = x_j$. On a avec (4.22)

$$\lambda v_j \le \ell(x_j,u^*) + \sum_i \left(f_i(x_j,u^*)_+ \frac{v_{j+e_i} - v_j}{h_i} + |f_i(x_j,u^*)| \frac{v_j - v_{j-e_i}}{h_i} \right). \tag{4.43}$$

Utilisant (4.41) avec $x = x_0 \pm h_ie_i$ et $y = y_o$, nous obtenons

$$\begin{aligned} v_{j\pm e_i} - v_j &\le \beta_\varepsilon(x_0 - y_0) + \delta\xi(x_0,y_0) \\ &\qquad -\beta_\varepsilon(x_0 \pm h_ie_i - y_0) - \delta\xi(x_0 \pm h_ie_i,y_0) \\ &\le -\beta'_\varepsilon(x_0 - y_0)(\pm h_ie_i) + \varepsilon^{-2}h_i^2 + \delta h_i. \end{aligned}$$

2. La démonstration du théorème étant technique, on pourra l'admettre en première lecture.

Multiplions cette inégalité (dans laquelle $\pm = -$) par $f_i(x_j,u^*)_+/h_i$; et (avec $\pm = +$)) par $|f_i(x_j,u^*)_-|/h_i$; ajoutons ces inégalités à (4.43); il vient

$$\lambda v_j \leq \ell(x_j,u^*) - \beta'_\varepsilon(x_0 - y_0)f(x_j,u^*) + \varepsilon^{-2}O(\max_i h_i) + O(\delta). \tag{4.44}$$

Soustrayant (4.42) de l'inégalité précédente, nous obtenons

$$\begin{aligned}\lambda(v_j - V(y_0)) \quad \leq \quad & (\ell(x_0,u^*) - \ell(y_0,u^*)) \\ & +\beta'_\varepsilon(x_0 - y_0)(f(y_0,u^*) - f(x_0,u^*)) + \varepsilon^{-2}O(\max_i h_i) + O(\delta).\end{aligned}$$

Combinant avec les relations

$$\ell(x_0,u^*) - \ell(y_0,u^*) = O(|x_0 - y_0|), \tag{4.45}$$
$$f(x_0,u^*) - f(y_o,u^*) = O(|x_0 - y_0|), \tag{4.46}$$

et prenant $\delta = O(h)$, il vient

$$v^h(x_o) - V(y_o) \leq C\left[|x_o - y_o| + \frac{|x_o - y_o|^2}{\varepsilon^2} + \frac{\max_i h_i}{\varepsilon^2}\right]. \tag{4.47}$$

De $ab \leq \frac{1}{2}(a^2 + b^2)$ on déduit que

$$|x_o - y_o| = \varepsilon\frac{|x_o - y_o|}{\varepsilon} \leq \tfrac{1}{2}(\varepsilon^2 + \frac{|x_o - y_o|^2}{\varepsilon^2}).$$

Avec (4.47), nous obtenons

$$v^h(x_o) - V(y_o) \leq C\left[\varepsilon^2 + \frac{|x_o - y_o|^2}{\varepsilon^2} + \frac{\max_i h_i}{\varepsilon^2}\right]. \tag{4.48}$$

Or

$$\sup\varphi - \delta \leq v(x_0) - w(y_0) - \frac{|x_o - y_o|^2}{\varepsilon^2},$$

donc

$$\frac{|x_o - y_o|^2}{\varepsilon^2} \leq \sup v - \inf w - \sup\varphi$$

ce qui prouve que $|x_0 - y_0| \longrightarrow 0$. Prenant $x = y = x_0$ dans (4.41), et utilisant le fait que V est hölderienne de constante γ, il vient

$$\frac{1}{\varepsilon^2}|x_o - y_o|^2 \leq V(x_0) - V(y_0) + \delta|x_o - y_o| \leq K|x_o - y_o|^\gamma,$$

pour un certain K indépendant de ε et h. De là

$$|x_o - y_o| \leq K\varepsilon^{\frac{2}{2-\gamma}}, \tag{4.49}$$

Donc, avec (4.48),

$$v^h(x_o) - V(y_o) \leq K\left[\varepsilon^{\frac{2\gamma}{2-\gamma}} + \frac{\max_i h_i}{\varepsilon^2}\right].$$

Prenant $\varepsilon = (\max_i h_i)^{(2-\gamma)/4}$, on obtient

$$v^h(x_o) - V(y_o) \leq K \left(\max_i h_i\right)^{\gamma/2}. \tag{4.50}$$

Prenant $\delta = O(\max_i h_i)$, il vient

$$\sup(v^h - V) \leq \sup \varphi \leq v^h(x_0) - V(y_0) + O(\max_i h_i) \leq O\left(\max_i h_i\right)^{\gamma/2} \tag{4.51}$$

d'où l'inégalité recherchée.

L'inégalité inverse se prouve de manière similaire, en maximisant la fonction

$$\varphi(x,y) := V(y) - v^h(x) + \beta_\varepsilon(x-y), \quad (x,y) \in I\!R^n_h \times I\!R^n.$$

Pour $\delta \in (0,1)$, on a encore l'existence de (x_1,y_1) dans $I\!R^n_h \times I\!R^n$ satisfaisant (4.38). Définissant ξ et ψ par (4.39) et (4.40) on obtient (4.41). Puisque x_o appartient à $I\!R^n_h$, il existe $j \in \mathbb{Z}^n$ tel que $x_o = x_j$. On poursuit de la même manière en faisant intervenir la commande $u^* \in U$ réalisant le minimum dans l'expression (4.22) du schéma. ■

Remarque 4.19 Si λ est assez grand, une variante de la démonstration du lemme 3.13 permet de montrer que V est lipschitzien. Dans ce cas on a une estimation d'erreur sur V de l'ordre de $O(t^{1/2})$.

Remarque 4.20 On trouvera la discussion d'autres schémas numériques dans l'annexe du livre [4], due à M. Falcone.

4.3.3 Equation eikonale

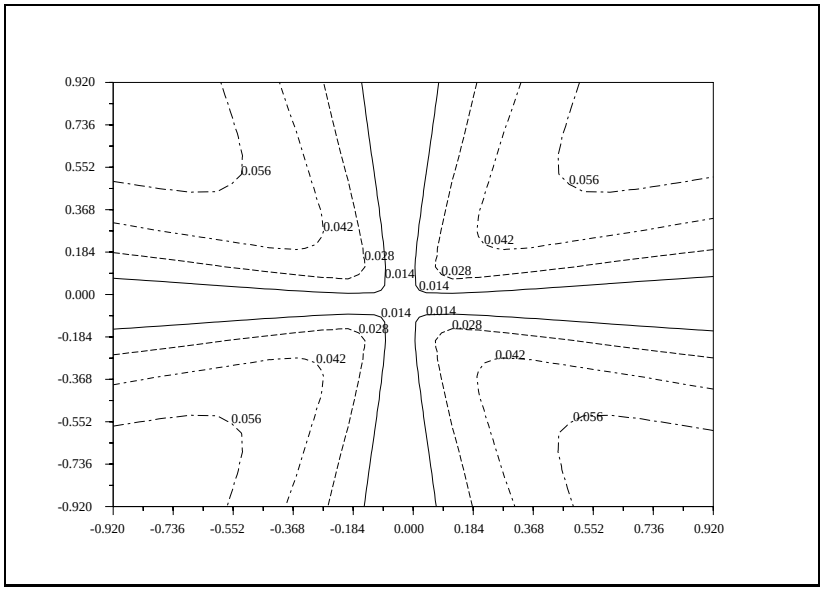


FIG. 4.1 – *Equation eikonale : erreur sur le temps minimal*

On considère le système dynamique suivant :

$$\dot{x} = F(x)u, \tag{4.52}$$

où $F : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ représente la vitesse du milieu. La commande u doit rester dans la boule unité pour la norme euclidienne. Pour $\lambda = 0$, l'équation HJB associée, dite équation eikonale, est de la forme

$$\begin{cases} 1 - F(x)\|Dv(x)\| &= 0 \quad \text{dans} \quad \Omega, \\ V(x) &= 0, \quad x \in C. \end{cases} \tag{4.53}$$

Dans l'exemple numérique, on a pris $C = \{0\}$, et $F(x) = 1$ pour tout x, de sorte que le temps de transfert est la distance euclidienne à 0.

Sur la figure 4.1, on a représenté les lignes de niveau de la différence entre solution calculée et valeur exacte, en limitant le domaine à [0,0.1] × [0,0.1]. La grille est de taille 25 × 25, et on a effectué 100 itérations sur les valeurs avec $\lambda = 0$. Comme on peut s'y attendre, on observe que les erreurs sont plus importantes dans les coins, en raison de l'accumulation d'erreurs inhérente à l'algorithme.

4.3.4 Problème d'alunissage

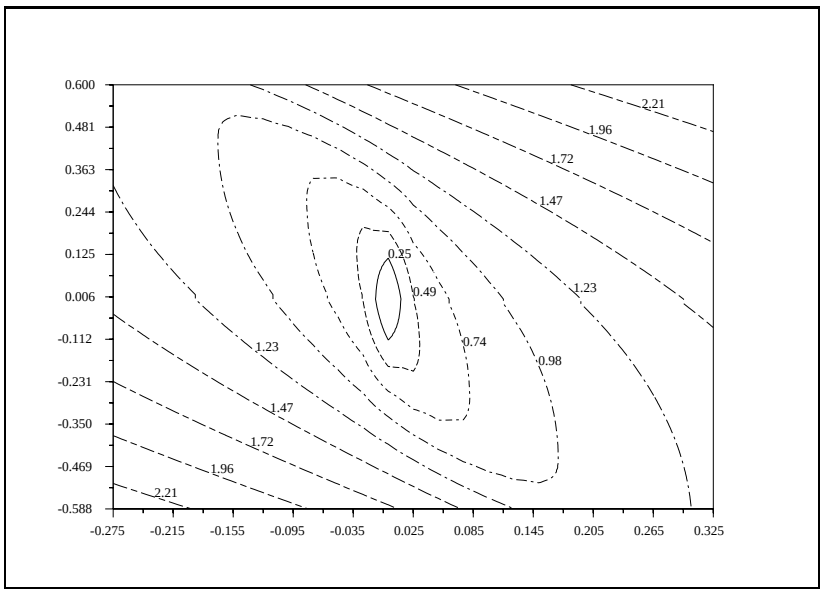


FIG. 4.2 – *Problème d'alunissage : isovaleurs du temps minimal*

Nous reprenons le problème d'alunissage discuté en section 1.2. Le problème discret est résolu sur le domaine $(z,\dot{z}) \in [-1,1] \times [-2,2]$. On prend 80 points de discrétisation pour z et on impose $\Delta t = \Delta x$. La condition de stabilité impose alors de prendre 320 points de discrétisation pour la vitesse. On fixe une condition aux limites artificielle égale à 100 sur le bord.

Les isovaleurs du temps minimal sont représentées en figure 4.2. La figure ne reprend que la partie centrale du domaine, pour éviter les effets dus au caractère borné du domaine.

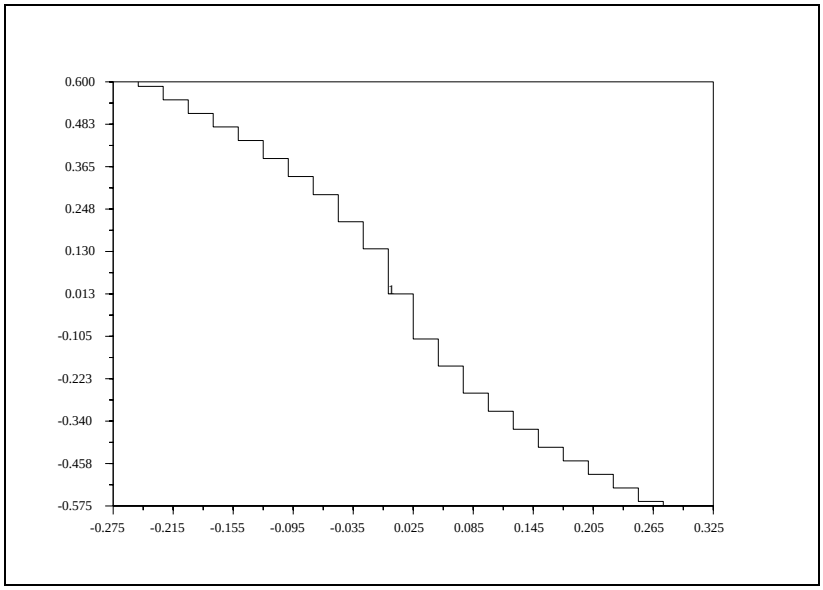


FIG. 4.3 – *Problème d'alunissage : lieu de changement de signe*

La figure 4.3 représente le lieu de changement de signe de l'estimation numérique de $\partial V/\partial \dot{z}$, qui en raison du théorème 3.14 détermine la statégie de feedback. Elle se relie bien aux isovaleurs de la figure 4.3. On la comparera à la figure 1.1 qui donne le lieu de changement de signe de la commande optimale.

4.4 Notes

La démonstration de convergence du théorème 4.17 reprend G. Barles and P. E. Souganidis [6]. L'estimation d'erreur du théorème 4.18 suit Capuzzo-Dolcetta et Ishii [16]. Une estimation analogue, dans le cas parabolique, se trouve dans M. G. Crandall and P.-L. Lions [14].

Chapitre 5

Commande optimale stochastique

5.1 Chaînes de Markov commandées

5.1.1 Quelques exemples

Un exemple classique de commande de chaînes de Markov est la *gestion de stock*: les achats des clients arrivent de manière aléatoire, et la commande consiste à réapprovisionner, avec paiement de pénalités pour tout achat non honoré. Autre exemple, la maintenance d'un parc d'outils de production. L'état du système est l'ensemble des outils en état de fonctionnement, et la commande consiste à effectuer les réparations des outils en panne. Il s'agit au fond de *conception de systèmes fiables*.

Enfin les problèmes de commande optimale (déterministes ou stochastiques) en espace continu (et temps continu ou discret) résolus en discrétisant l'équation HJB reviennent, comme on le verra, à résoudre un problème de commande d'une chaînes de Markov. En particulier, les problèmes d'évaluation d'options financière, d'identification de volatilité implicite, et de gestion de portefeuille sont de cette nature.

5.1.2 Chaînes de Markov et valeurs associées

Considérons un *système dynamique* dont l'état peut prendre un nombre fini ou dénombrable de valeurs, soit $1,\ldots,m$, avec m fini ou non. Il est utile de traiter le cas $m=\infty$ pour discuter le problème de discrétisation de systèmes continus.

On note x^k la valeur de l'état au temps k, où $k \in I\!N$. On suppose connue la probabilité M_{ij}^k de transition de l'état i au temps k, à l'état j au temps $k+1$. Autrement dit, notant $\mathcal{P}$ la loi de probabilité, on a

$$\mathcal{P}(x^{k+1}=j|x^k=i)=M_{ij}^k. \tag{5.1}$$

On supposera cette loi *markovienne*, c'est à dire

$$\mathcal{P}(x^{k+1}=j|x^k=i,x^{k-1}=i_{k-1},\ldots,x^0=i_0)=M_{ij}^k. \tag{5.2}$$

Ceci signifie que si on connaît la valeur de l'état au temps k, la connaissance des états passés n'apporte rien pour la prédiction du futur.

La "matrice" $M^k = \{M^k_{ij}\}$, où i et j varient de 1 à m, est le tableau (fini ou non) de valeur M^k_{ij} en ligne i et colonne j. Tous ses éléments sont positifs ou nuls, et la somme des éléments d'une ligne vaut 1. Une telle matrice est dite *stochastique*.

Si $m = \infty$, l'extension naturelle du calcul matriciel : produit de deux matrices, produit d'une matrice avec un vecteur (vertical) à droite ou (horizontal) à gauche, et produit de deux matrices, demande quelques précautions : il faut que les quantités en jeu soient sommables. Plus précisément, soient ℓ^1 et ℓ^∞, respectivement, l'espace des suites sommables et bornées, dont les éléments sont indicés de 1 à m, et représentés comme des vecteurs horizontaux (pour ℓ^1) et verticaux (pour ℓ^∞). Si $x \in \ell^1$ et $v \in \ell^\infty$, et si M est une matrice stochastique, on peut définir leur produit $xM \in \ell^1$ et $Mv \in \ell^\infty$ par

$$(xM)_j := \sum_{i=1}^{m} x_i M_{ij}; \quad (Mv)_i := \sum_{j=1}^{m} M_{ij} v_j.$$

On a en effet $\|xM\|_1 \leq \|x\|_1$ et $\|Mv\|_\infty \leq \|v\|_\infty$. Si M^1 et M^2 sont deux matrices stochastiques, on peut définir leur produit M^1M^2 par

$$(M^1M^2)_{ij} := \sum_{k=1}^{m} M^1_{ik} M^2_{kj}.$$

Il est facile de vérifier que le produit de deux matrices stochastiques est une matrice stochastique. On interprètera

$$\left\{ p \in \ell^1; \quad p_i \geq 0, \ i = 1, \ldots, m; \quad \sum_{i=1}^{m} p_i = 1 \right\}$$

comme l'*espace des probabilités* pour l'état du système à un temps donné, et ℓ^∞ comme un *espace de valeurs*. Cette dernière terminologie sera plus claire dans la suite.

Si l'état x^k du système à l'instant k est connu, la loi de probabilité de x^{k+1} est la ligne de M^k d'indice x^k. Si on dispose seulement d'une loi de probabilité pour x^k, notée $p^k = (p^k_1, \ldots, p^k_m)$, et considérée comme un vecteur horizontal, alors la loi de probabilité de x^{k+1} vérifie l'*équation de Kolmogorov avant*

$$p^{k+1} := \mathcal{P}(x^{k+1}|p^k) = \sum_i p^k_i M^k_{i,\cdot} = p^k M^k, \tag{5.3}$$

d'où on déduit par récurrence, si la probabilité initiale est p^0,

$$\mathcal{P}(x^{k+1}|p^0) = p^0 M^0 M^1 \ldots M^k. \tag{5.4}$$

Associons maintenant à ce processus la *fonction coût* $\{c^k_i\}$, $i = 1, \ldots, m$, $k \in \mathbb{N}$. On suppose que $c^k := \{c^k_i\}_{i=1,\ldots,m}$ appartient à ℓ^∞, ce qui veut dire que les coûts sont uniformément bornés en espace, et que c^k est représenté comme un vecteur vertical. Soit φ une application $\{1, \ldots, m\} \longrightarrow \ell^\infty$, appelée coût final. Définissons la *fonction valeur* du problème avec état initial i et instant initial k comme

$$V^k_i := \mathbb{E}\left(\sum_{\ell=k}^{N-1} c^\ell_{x^\ell} + \varphi(x^N) \mid x^k = i \right). \tag{5.5}$$

Ici $N > 0$ est l'*horizon*, et $\mathbb{E}$ représente l'*espérance mathématique*.

Proposition 5.1 *Pour tout $k = 0, \ldots, N$, la fonction valeur V^k est bien définie et appartient à ℓ^∞. De plus, la suite $\{V^k\}$ est solution de l'équation de récurrence de Kolmogorov arrière*

$$\begin{cases} V^k = c^k + M^k V^{k+1}, \quad k = 0, \ldots, N-1, \\ V^N = \varphi. \end{cases} \tag{5.6}$$

Démonstration. Si x^k a la valeur i, alors d'après l'équation de Kolmogorov avant

$$V_i^k = c_i^k + \sum_{j=1}^{m} M_{ij}^k V_j^{k+1},$$

d'où le résultat. ■

Considérons maintenant un problème avec $c^k = c \in \ell^\infty$ et $M^k = M$ indépendants du temps, *horizon infini*, et taux d'actualisation $\beta \in]0,1[$. La valeur de ce problème, c'est à dire

$$V_i := \mathbb{E}\left(\sum_{k=0}^{\infty} \beta^{k+1} c_{x^k} | x^0 = i \right), \tag{5.7}$$

est bien définie et appartient à ℓ^∞. En raison de l'équation de Kolmogorov avant, elle est solution de l'équation

$$V = \beta(c + MV). \tag{5.8}$$

Comme M est lipschitzienne de constante 1, cette équation est celle d'un opérateut de point fixe strictement contractant et a donc une solution unique.

5.1.3 Quelques lemmes

Commençons par le rappel du théorème de point fixe de Banach-Picard.

Lemme 5.2 *Soient X un espace de Banach et C une partie fermée de X. Soit T un opérateur contractant de C vers lui même. Autrement dit, il existe $c \in [0,1[$ tel que, si $x^i \in C$, $i = 1,2$, alors $Tx^i \in C$, $i = 1,2$, et*

$$\|Tx^2 - Tx^1\| \leq c\|x^2 - x^1\|. \tag{5.9}$$

Alors T a un unique point fixe $x^* \in C$ (c.a.d. l'équation $Tx = x$ a pour solution unique x^*). De plus, quel que soit $x^0 \in C$, la suite $\{x^k\}$ telle que $x^{k+1} = Tx^k$ converge vers x^*, et

$$\|x^k - x^*\| \leq c^k \|x^0 - x^*\|. \tag{5.10}$$

Voici un autre lemme, qui sera utile à plusieurs reprises.

Lemme 5.3 *Soit M une matrice stochastique, $\beta \in]0,1[$, $\varepsilon > 0$ et $w \in \ell^\infty$ tels que $w \leq \varepsilon \mathbf{1} + \beta M w$. Alors $w \leq (1-\beta)^{-1}\varepsilon \mathbf{1}$.*

Démonstration. On a $Mw \leq (\sup w)\mathbf{1}$ puisque M est une matrice stochastique, et donc $w \leq (\varepsilon + \beta \sup w)\mathbf{1}$. En conséquence, $\sup w \leq \varepsilon + \beta \sup w$, d'où la conclusion. ■

5.1.4 Principe de Programmation dynamique

Considérons maintenant une chaîne de Markov dont les probabilités de transition $M_{ij}(u)$ dépendent d'une variable de commande $u \in U_i$, où U_i est un ensemble quelconque dépendant de l'état i (certains résultats supposeront U_i *métrique compact*). Donnons nous des coûts dépendant de u et de l'état, soit $c_i^k(u) : U_i \longrightarrow I\!R$, telle que

$$\sup_u \sup_i \sup_k |c_i^k(u)| < \infty. \tag{5.11}$$

On considère le problème de minimisation du critère sur horizon fini

$$V_i^k(u) := I\!E\left(\sum_{\ell=k}^{N-1} c_{x^\ell}^\ell(u^k) + \varphi(x^N)|x^k = i\right). \tag{5.12}$$

Ici u^k est la valeur de la commande au temps k; pour donner un sens à ce problème, il faut spécifier l'information dont on dispose au temps k pour choisir la valeur de u^k. Nous allons nous limiter au cas de l'*observation complète*, dans lequel l'état x^k est connu. Ceci permet de choisir u^k fonction de l'état x^k, et bien sûr du temps k. Autrement dit, on choisit une stratégie de *retour d'état*. Posons

$$\mathcal{U} := \Pi_i U_i. \tag{5.13}$$

On notera u_i la commande adoptée (au temps k) par la stratégie feedback $u \in \mathcal{U}$ si l'état est i, et $M(u)$ la "matrice" de terme générique $M_{ij}(u_i)$. On considère donc le problème de calcul d'un retour d'état optimal

$$V_i^k := \inf_{u\in\mathcal{U}} V_i^k(u), \quad i = 1,\ldots,m. \quad k = 1,\ldots,N. \tag{5.14}$$

Proposition 5.4 *La fonction valeur V^k, solution du problème* (5.14) *avec observation complète, est solution du principe de programmation dynamique*

$$\begin{cases} V_i^k = \inf\limits_{u\in U_i}\left\{c_i^k(u) + \sum\limits_j M_{ij}^k(u)V^{k+1}\right\}, \quad i = 1,\ldots,m, \quad k = 0,\ldots,N-1, \\ V^N = \varphi. \end{cases} \tag{5.15}$$

De plus, l'ensemble $\bar{U}_i^k$ (éventuellement vide) des commandes optimales à l'instant k lorsque $x^k = i$ est

$$\bar{U}_i^k = \operatorname*{argmin}_{u\in U_i}\left\{c_i^k(u) + \sum_j M_{ij}^k(u)V^{k+1}\right\}. \tag{5.16}$$

Démonstration. On raisonne par récurrence. Il est clair que $V^N = \varphi$. Fixons $k < N$ et $i \in \{1,\ldots,m\}$. Si $x^k = i$, d'après l'équation de Kolmogorov arrière, le choix de la commande u à l'instant k donne la valeur $c_i^k(u) + \sum_j M_{ij}^k(u)V^{k+1}$. On obtient donc V_i^k en prenant l'infimum de cette quantité, et une commande est optimale si elle appartient à l'argument du minimum. De plus la quantité

$$\|V^k\|_\infty \leq \sup_u \|c^k(u)\| + \|V^{k+1}\|_\infty$$

est bien bornée. ■

5.1.5 Problèmes à horizon infini

Dans cette section, nous supposons la fonction coût et la matrice de transition indépendantes du temps, notées $c(u)$ et $M(u)$, et le coût actualisé avec un coefficient $\beta \in]0,1[$. Le théorème suivant caractérise les politiques optimales, et montre en particulier qu'on peut se limiter aux politiques feedback stationnaires (la commande ne dépend que de l'état mais pas du temps).

Théorème 5.5 (i) *Dans le cas de l'observation complète, la fonction valeur définie par*

$$V_i := \inf_{u\in\mathcal{U}} I\!E \left\{ \sum_{k=0}^{\infty} \beta^{k+1} c_{x^k}(u_k) | x^0 = i \right\}, \quad i = 1,\dots,m, \tag{5.17}$$

où $\beta \in]0,1[$, est solution unique de l'équation de programmation dynamique : trouver $v \in I\!R^m$ tel que

$$v_i = \beta \inf_{u\in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v \right\}, \quad i = 1,\dots,m. \tag{5.18}$$

(ii) *Soit $\varepsilon \geq 0$ et $u \in \mathcal{U}$ une politique telle que, pour tout i,*

$$\beta \left(c_i(u_i) + \sum_j M_{ij}(u_i) v^* \right) \leq v_i^* + \varepsilon \mathbf{1}. \tag{5.19}$$

Posons $\varepsilon' := (1-\beta)^{-1}\varepsilon$. Alors la politique u est ε' sous optimale, dans le sens où la valeur associée V satisfait

$$V \leq v^* + \varepsilon' \mathbf{1}. \tag{5.20}$$

(iii) *Supposons, pour tout i et j, U_i* métrique compact *et les fonctions $c_i(u)$ et $M_{ij}(u)$ continues. Alors il existe (au moins) une politique optimale.*

Démonstration. a) Montrons d'abord que (5.18) possède une solution unique. Cette équation est de la forme $v = Tv$, avec

$$(Tw)_i := \beta \inf_{u\in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) w \right\}. \tag{5.21}$$

Montrons que T est un opérateur contractant dans ℓ^∞. On a

$$\|Tw\|_\infty \leq \beta(\|c\|_\infty + \|w\|_\infty),$$

ce qui montre que T est un opérateur de ℓ^∞ dans lui même. Avec (1.26) et étant donnés w et w' dans ℓ^∞, utilisant le fait que la somme des éléments d'une ligne de $M(u)$ vaut 1, il vient :

$$|(Tw')_i - (Tw)_i| \leq \beta \sup_{u\in U_i} \sum_{j=1}^{m} |M_{ij}(u)(w'-w)_j| \leq \beta \|w'-w\|_\infty.$$

En conséquence, T est une contraction de rapport β dans ℓ^∞. Il découle alors du lemme 5.2 que l'équation (5.18) a une solution unique v^*.

b) Soit $u \in \mathcal{U}$ une politique et V la valeur associée, solution de $V = \beta(c(u) + M(u)V)$. Montrons que $v^* \leq V$. En effet, soit $i \in \{1,\ldots,m\}$. Utilisant

$$v^* \leq \beta(c(u) + M(u)v^*), \tag{5.22}$$

il vient

$$v^* - V \leq \beta M(u)(v^* - V). \tag{5.23}$$

Le lemme 5.3 assure que $v^* \leq V$, comme il fallait le démontrer. Nous avons montré (i).
c) Soit $\varepsilon \geq 0$. Si $\varepsilon > 0$, par définition de v^*, il existe une politique $\tilde{u}$ telle que $\tilde{u}_i$ satisfait (5.19) pour tout i. Revenons au cas général où $\varepsilon \geq 0$. Notons $\tilde{V}$ la valeur associée à la politique $\tilde{u}$. Utilisant $\tilde{V} = \beta(c(\tilde{u}) + M(\tilde{u})\tilde{V})$ et (5.19), il vient

$$\tilde{V} - v^* \leq \varepsilon \mathbf{1} + \beta M(\tilde{u})(\tilde{V} - v^*). \tag{5.24}$$

On en déduit (5.20) avec le lemme 5.3. D'autre part, on sait que $v^* \leq V$ pour toute valeur V associée à une politique.
(iii) D'après le point (ii), l'existence d'une politique optimale équivaut à la possibilité d'atteindre, pour tout état i, l'infimum dans (5.18). Montrons que ceci est conséquence des hypothèses du point (iii).

Pour i fixé, notons $\{u^q\}$ une suite minimisante. Puisque U est métrique compact, extrayant une sous-suite si nécessaire, on peut supposer que la suite converge vers $\bar{u} \in U$. A tout $\varepsilon \in]0,1[$, on peut associer une partition (I,J) de $\{1,\ldots,m\}$, telle que I est de cardinal fini et $\sum_{i\in I} M_{ij}(\bar{u}) \geq 1 - \frac{1}{2}\varepsilon$. Puisque I est fini, pour q assez grand, on a $\sum_{i\in I} M_{ij}(u^q) \geq 1 - \varepsilon$, et donc $\sum_{i\in J} M_{ij}(u^q) \leq \varepsilon$. De là

$$\begin{aligned}
\Delta &:= \left| \limsup_q (c_i(u_i^q) + \sum_j M_{ij}(u_i^q)V - c_j(\bar{u}_i) - \sum_j M_{ij}(\bar{u}_i)V) \right| \\
&= \left| \limsup_q \sum_{j\in J} (M_{ij}(u^q)V_j - M_{ij}(\bar{u})V_j) \right| \\
&\leq \limsup_q \sum_{j\in J} |M_{ij}(u^q) - M_{ij}(\bar{u})| \|V\|_\infty \leq 2\varepsilon \|V\|_\infty.
\end{aligned}$$

Ceci prouve que

$$(c(\bar{u}) + M(\bar{u})V)_i = \inf_{u\in U} (c(u) + M(u)V)_i\,, \tag{5.25}$$

d'où (iii).

■

5.1.6 Algorithmes numériques

Dans le cas de problèmes avec horizon infini, on peut mettre en œuvre un algorithme itératif de calcul de v à partir du principe de programmation dynamique. La méthode la plus simple est l'*itérations sur les valeurs*

$$v_i^{q+1} = \beta \inf_{u\in\mathcal{U}} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^q \right\}, \quad i = 1,\ldots,m, \quad q \in I\!N. \tag{5.26}$$

Ici v^q (à ne pas confondre avec la notation v^k employée dans le cas de l'horizon fini) représente la suite formée par l'algorithme.

Proposition 5.6 *L'algorithme d'itération sur les valeurs converge vers la solution unique v^* de* (5.18)*, et on a*

$$\|v^q - v^*\|_\infty \leq \beta^q \|v^0 - v^*\|_\infty. \tag{5.27}$$

Démonstration. Soit T l'opérateur construit en (5.21). Nous avons montré (démonstration du théorème 5.5) que T est contractant de rapport β dans la norme du max. L'algorithme d'itération sur les valeurs s'écrit $v^q = Tv^{q-1}$. On conclut avec le lemme 5.2. ■

Dans le cas assez fréquent où β est proche de 1, l'algorithme d'itération sur les valeurs peut être très lent. Une alternative intéressante est l'algorithme d'*itérations sur les stratégies*, ou *algorithme de Howard*. On fera l'hypothèse suivante :

$$\begin{cases} \text{U est métrique compact} \\ \text{Les fonctions } c_i(u) \text{ et } M_{ij}(u) \text{ sont continues pour tout } i \text{ et } j. \end{cases} \tag{5.28}$$

Chaque itération de l'algorithme comporte deux étapes :

– Etant donné une stratégie $u^q \in \mathcal{U}$, calculer la valeur v^q associée, solution de l'équation linéaire

$$v^q = \beta(c(u^q) + M(u^q)v^q). \tag{5.29}$$

– Calculer u^{q+1} solution de

$$u_i^{q+1} \in \arg\min_{u \in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^q \right\}, \quad i = 1, \ldots, m. \tag{5.30}$$

Proposition 5.7 *On suppose* (5.28)*. Alors l'algorithme d'itérations sur les politiques, initialisé avec une politique $u^0 \in \mathcal{U}$ quelconque, a les propriétés suivantes :*
(i) *Il est bien défini,*
(ii) *La suite v^q décroît,*
(iii) *Elle vérifie $\|v^{q+1} - v^*\| \leq \beta \|v^q - v^*\|$, où v^* est la fonction valeur, unique solution du principe de programmation dynamique* (5.18)*.*

Démonstration. (i) Vérifions que l'algorithme est bien défini. Le système linéaire (5.29) a une solution unique, car c'est l'équation de point fixe d'un opérateur contractant (lemme 5.2). Utilisant les arguments de la démonstration du théorème 5.5, on vérifie que le minimum dans la seconde étape est atteint en raison de (5.28).

Par ailleurs, la suite v^q est bornée dans ℓ^∞ car la relation

$$\|v^q\|_\infty \leq \beta(\|c(u^q)\|_\infty + \|M(u^q)v^q\|_\infty) \leq \beta(\|c(u^q)\|_\infty + \|v^q\|_\infty)$$

donne l'estimation

$$\|v^q\|_\infty \leq (1-\beta)^{-1}\beta \|c\|_\infty. \tag{5.31}$$

(ii) Les relations (5.29) et (5.30) impliquent

$$\begin{aligned} \beta^{-1}(v^{q+1} - v^q) &= c(u^{q+1}) + M(u^{q+1})v^{q+1} - c(u^q) - M(u^q)v^q, \\ &\leq c(u^{q+1}) + M(u^{q+1})v^{q+1} - c(u^{q+1}) - M(u^{q+1})v^q, \\ &= M(u^{q+1})(v^{q+1} - v^q), \end{aligned}$$

et donc $v^{q+1} - v^q \leq 0$ d'après le lemme 5.3.

(iii) Notons $\bar{v}^{q+1}$ la valeur calculée à partir de v^q, par l'itération sur les valeurs. On sait que $\|\bar{v}^{q+1} - v^*\| \leq \beta\|v^q - v^*\|$. Puisque $v^* \leq v^{q+1}$, il suffit d'établir que $v^{q+1} \leq \bar{v}^{q+1}$. Or

$$\begin{aligned} \beta^{-1}(v^{q+1} - \bar{v}^{q+1}) &= c(u^{q+1}) + M(u^{q+1})v^{q+1} - (c(u^{q+1}) - M(u^{q+1})v^q), \\ &= M(u^{q+1})(v^{q+1} - v^q). \end{aligned}$$

D'après le point (ii), $v^{q+1} \leq v^q$; donc $v^{q+1} \leq \bar{v}^{q+1}$. ■

Remarque 5.8 La démonstration précédente montre que l'itération sur les politiques converge au moins aussi vite que l'itération sur les valeurs.

5.1.7 Problèmes de temps de sortie

Soit Ω une partie de $\{1, \ldots, m\}$, et considérons une chaîne de Markov (sans commande) de matrice de transition M. Soit τ le premier instant de sortie de Ω :

$$\tau := \min\{k \in I\!N;\ x^k \notin \Omega\}. \tag{5.32}$$

Bien entendu, τ est une variable aléatoire. On considère la fonction valeur, où $i \in \{1, \ldots, m\}$:

$$V_i := I\!E\left(\sum_{k=0}^{\tau-1} \beta^{k+1} c_{x^k} + \beta^\tau \varphi_{x^\tau} | x^0 = i\right). \tag{5.33}$$

Proposition 5.9 *On suppose c et φ dans ℓ^∞. Alors l'espérance ci-dessus est bien définie, et La fonction valeur du problème de temps de sortie appartient aussi à ℓ^∞, et est solution unique de l'équation*

$$\begin{cases} v_i = \beta\left(c_i + \sum_j M_{ij} v_j\right), & i \in \Omega, \\ v_i = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.34}$$

Démonstration. Elle est similaire à celle des propositions précédentes. ■

Considérons maintenant le cas de la chaîne de Markov commandée de probabilité de transition $M_{ij}(u)$, avec $u \in U_i$, ensemble métrique compact, et les fonctions $c_i(u)$ et $M_{ij}(u)$ continues. On considère le problème de minimisation du critère avec temps de sortie

$$V_i := \inf_{u \in \mathcal{U}} I\!E\left\{\sum_{k=0}^{\tau-1} \beta^{k+1} c(u)_{x^k} + \beta^\tau \varphi_{x^\tau} | x^0 = i\right\}, \tag{5.35}$$

dans le cas de l'observation complète.

Remarque 5.10 Si c est le vecteur de coordonnées toutes égales à 1, et si φ est nul, alors le critère s'interprète comme une mesure du temps de sortie. Le problème est alors dit à temps minimal.

Proposition 5.11 *On suppose* $\sup_{u\in U}|c_i(u)|$ *fini et* φ *borné. Alors la fonction valeur du problème avec temps de sortie est solution unique de l'équation de la programmation dynamique*

$$\begin{cases} v_i = \beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j\right\}, & i \in \Omega, \\ v_i = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.36}$$

Démonstration. Elle est similaire à celle des propositions précédentes. ■

L'extension des algorithmes d'itérations sur les valeurs et sur les politiques à la situation étudiée ici ne présente pas de difficulté.

5.1.8 Problèmes avec décision d'arrêt

Nous étudions un problème de commande similaire à celui de la sous-section précédente, ajoutant la possibilité d'arrêt à tout instant, avec un coût d'arrêt $\psi \in I\!R^m$.

Soit Ω une partie de $\{1,\ldots,m\}$, et soient une chaîne de Markov commandée de matrice de transition $M_{ij}(u)$, avec $u \in U$, ensemble métrique compact, et les fonctions $c(u)$ et $M_{ij}(u)$ continues. On note τ le premier instant de sortie de Ω, et θ l'instant de décision d'arrêt. Posons

$$\chi_{\theta<\tau} = \begin{cases} 1 & \text{si } \theta < \tau, \\ 0 & \text{sinon }, \end{cases}$$

et adoptons une convention similaire pour $\chi_{\theta\geq\tau}$. On considère le problème de minimisation du critère avec temps d'arrêt

$$V_i := \inf_{u\in\mathcal{U}} I\!E\left\{\sum_{k=0}^{\theta\wedge\tau-1}\beta^{k+1}c(u)_{x^k} + \beta^\theta\chi_{\theta<\tau}\psi_{x^\theta} + \beta^\tau\chi_{\theta\geq\tau}\varphi_{x^\tau}|x^0 = i\right\}, \tag{5.37}$$

dans le cas de l'observation complète.

Remarque 5.12 Le cadre de cette section recouvre plusieurs situations intéressantes : (i) ensemble Ω égal à l'espace d'état, (ii) U_i réduit à un point pour tout i : la seule décision est d'arrêter ou non, (iii) stratégie optimale pouvant être de ne jamais arrêter le jeu.

Théorème 5.13 *On suppose* $\sup_{u\in U}|c_i(u)|$ *fini et* ψ *et* φ *borné. Alors la fonction valeur* v *du problème de temps d'arrêt est solution unique du système*

$$\begin{cases} \text{(i)} \quad v_i = \min\left(\beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j\right\}, \psi_i\right), & i \in \Omega, \\ \text{(ii)} \quad v_i = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.38}$$

Démonstration. La démonstration est similaire à celle des sections précédentes; contentons-nous de démontrer que l'équation 5.38 a une solution unique v^*. Définissons l'opérateur T de $I\!R^m$ dans lui même par

$$\begin{cases} (Tv)_i = \min\left(\beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j\right\}, \psi_i\right), & i \in \Omega, \\ (Tv)_i = \varphi_i, & i \notin \Omega, \end{cases} \tag{5.39}$$

alors pour la norme du max, T est une contraction stricte de rapport β, et a donc un unique point fixe v^*. Ceci établit l'existence et l'unicité de la solution de (5.38). ■

Les arguments qui précèdent assurent la convergence de l'*algorithme d'itérations sur les valeurs*, qui s'écrit, en reprenant les notations de (5.39),

$$v^{q+1} = T(v^q), \tag{5.40}$$

ou encore

$$\begin{cases} v_i^{q+1} = \min\left(\beta \inf\limits_{u\in U_i}\left\{c_i(u) + \sum\limits_j M_{ij}(u)v_j^q\right\}, \psi_i\right), & i \in \Omega, \\ v_i^{q+1} = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.41}$$

En ce qui concerne l'*algorithme d'itérations sur les politiques*, on peut écrire un algorithme de principe sous la forme suivante :

1. Choisir arbitrairement la stratégie initiale $u^0 \in \mathcal{U}$.
   Poser $q := 0$.
2. Etant donné une stratégie $u^q \in \mathcal{U}$, calculer v^q solution de

$$\begin{cases} v_i^q = \min\left(\beta\left\{c_i(u_i^q) + \sum\limits_j M_{ij}(u_i^q)v_j^q\right\}, \psi_i\right), & i \in \Omega, \\ v_i^q = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.42}$$

3. Calculer u^{q+1} solution, pour tout i, de

$$u_i^{q+1} \in \arg\min_{u\in U_i}\left\{c_i(u) + \sum_j M_{ij}(u)v_j^q\right\}. \tag{5.43}$$

4. $q := q + 1$, aller en 1.

Nous admettons la proposition suivante, dont la démonstration, extension de celle de la proposition 5.7, utilise (1.27).

Proposition 5.14 *L'algorithme ci-dessus, initialisé avec une politique $u^0 \in \mathcal{U}$ quelconque, est bien défini, et forme une suite de valeurs v^q décroissante, et qui vérifie $\|v^{q+1} - v^*\| \leq \beta\|v^q - v^*\|$, où v^* est solution unique de* (5.38).

5.1.9 Un algorithme implémentable

L'algorithme d'itérations sur les politiques que nous venons de présenter nécessite à chaque itération la résolution de l'équation non linéaire (5.42), ce qui peut être très coûteux. Nous allons formuler un autre algorithme, itérant sur les politiques, dans lequel on ne résout qu'une équation linéaire à chaque itération. L'idée est de calculer v^q solution de l'équation linéaire

$$\begin{cases} v_i^q = \beta\left(c_i(u_i^q) + \sum\limits_j M_{ij}(u_i^q)v_j^q\right), & i \in I^q, \\ v_i^q = \psi_i, & i \in \Omega \setminus I^q, \\ v_i^q = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.44}$$

L'ensemble I^q, inclus dans Ω, est une prédiction des états i pour lesquels la contrainte $v_i \leq \psi_i$ n'est pas active à l'optimum. Il doit être mis à jour. Ceci conduit à l'algorithme suivant :

1. Choisir arbitrairement la stratégie initiale $u^0 \in \mathcal{U}$.
   Calculer $\hat{v}^0$ solution de l'équation linéaire

$$\begin{cases} \hat{v}_i^0 = \beta \left(c_i(u_i^0) + \sum_j M_{ij}(u_i^0) \hat{v}_j^0 \right), & i \in \Omega, \\ \hat{v}_i^0 = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.45}$$

   Calculer v^0 comme suit :

$$\begin{cases} v_i^0 = \min(\hat{v}_i^0, \psi_i), & i \in \Omega, \\ \hat{v}_i^0 = \varphi_i, & i \notin \Omega. \end{cases} \tag{5.46}$$

   Poser $q := 0$ et

$$I^0 := \{ i \in \Omega ; \ \ v_i^0 < \psi_i \}. \tag{5.47}$$

2. Faire $q := q + 1$. Calculer u^q solution de

$$u_i^q \in \arg\min_{u \in U_i} \left\{ c_i(u) + \sum_j M_{ij}(u) v_j^{q-1} \right\}, \quad i = 1, \ldots, m. \tag{5.48}$$

3. Poser

$$I^q := I^{q-1} \cup \left\{ i \in \Omega ; \ \ \beta \left(c_i(u_i^q) + \sum_j M_{ij}(u_i^q) v_j^{q-1} \right) < \psi_i \right\}. \tag{5.49}$$

4. Calculer v^q, solution de l'équation linéaire (5.44).
   Aller en 2.

Proposition 5.15 *L'algorithme ci-dessus forme une suite de valeurs v^q décroissant vers la solution unique v^* de* (5.38).

Démonstration. a) Montrons la décroissance de v^q. S'il n'en est pas ainsi, soient $q \in \mathbb{N}$ et $i \in \Omega$ tels que $v_i^{q+1} - v_i^q > 0$. Etant donné $\varepsilon > 0$, on peut supposer que $(v^{q+1} - v^q)_i \geq \sup_j (v^{q+1} - v^q)_j - \varepsilon$. Par ailleurs, $i \in I^{q+1}$ (sinon v_i^{q+1} et v_i^q sont égaux à ψ_i). Donc

$$v_i^{q+1} = \beta \left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1}) v_j^{q+1} \right). \tag{5.50}$$

Posons $w := v^{q+1} - v^q$, et distinguons deux cas. Si $i \in I^q$, alors

$$v_i^q = \beta \left(c_i(u_i^q) + \sum_j M_{ij}(u_i^q) v_j^q \right), \tag{5.51}$$

et donc avec (5.48)

$$\begin{array}{rcl} w_i & = & \beta\left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1})v_j^{q+1} - c_i(u_i^q) - \sum_j M_{ij}(u_i^q)v_j^q\right), \\ & \leq & \beta\left(\sum_j M_{ij}(u^{q+1})w_j\right) \leq \beta(w_i+\varepsilon), \end{array} \tag{5.52}$$

ce qui donne la contradiction recherchée pour $\varepsilon > 0$ assez petit.

Si, au contraire, $i \notin I^q$, alors $v_i^q = \psi_i$ et, par définition de I^{q+1}, on a

$$\beta\left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1})v_j^q\right) < \psi_i = v_i^q. \tag{5.53}$$

Donc

$$\begin{array}{rcl} w_i & = & \beta\left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1})v_j^{q+1}\right)_i - \psi_i, \\ & \leq & \beta\left(c_i(u_i^{q+1}) + \sum_j M_{ij}(u_i^{q+1})v_j^{q+1} - c_i(u_i^{q+1}) - \sum_j M_{ij}(u_i^{q+1})v_j^q\right)_i, \end{array} \tag{5.54}$$

ce qui permet de conclure de la même manière.

b) On peut montrer, par des arguments déjà employés, que la suite v^q est bornée. Puisqu'elle est décroissante, elle converge vers une valeur $\hat{v}$. De même, I^q étant croissant, converge vers un certain I^*. Enfin par compacité on a la convergence de u^q vers $\hat{u} \in \mathcal{U}$ pour une sous suite. Passant à la limite dans (5.44)[1], il vient

$$\left\{\begin{array}{ll} \hat{v}_i = \beta\left(c_i(\hat{u}_i) + \sum_j M_{ij}(\hat{u}_i)\hat{v}_j\right), & i \in I^*, \\ \hat{v}_i = \psi_i, & i \in \Omega \setminus I^*, \\ \hat{v}_i = \varphi_i, & i \notin \Omega. \end{array}\right. \tag{5.55}$$

De plus la décroissance de v^q implique

$$\hat{v}_i \leq \psi_i, \quad i \in I^*, \tag{5.56}$$

et le passage à la limite dans (5.49) donne

$$\beta\left(c_i(\hat{u}_i) + \sum_j M_{ij}(\hat{u}_i)\hat{v}_j\right) \geq \psi_i, \quad i \in \Omega \setminus I^*. \tag{5.57}$$

Les trois relations ci-dessus impliquent que $\hat{v}$ est solution de (5.38), donc est égale à la fonction valeur v. ■

Remarque 5.16 L'algorithme présenté dans cette section peut s'avérer lent si la mise à jour de l'ensemble I^q n'est pas assez efficace. On peut y remédier, soit en introduisant quelques itérations sur les valeurs (peu coûteuses, comparées à la résolution du système (5.45)), soit en s'inspirant des algorithmes de résolution de problèmes de complémentarité linéaire, par exemple ceux basés sur les points intérieurs.

1. Par des arguments similaires à ceux employés dans la démonstration du théorème 5.5(iii).

5.2 Problèmes en temps et espace continus

5.2.1 Position du problème

Etudions le problème de commande optimale stochastique

$$(P_x) \qquad \begin{cases} \text{Min } I\!E \displaystyle\int_0^\infty \ell(y(t),u(t))e^{-\lambda t}\mathrm{d}t; \\ \\ \mathrm{d}y(t) = f(y(t),u(t))\mathrm{d}t + \sigma(y(t),u(t))\mathrm{d}w, \quad u(t) \in U, \quad t \in [0,\infty[, \\ \\ y_0 = x. \end{cases}$$

Dans ce problème nous retrouvons les ingrédients du problème de commande optimale déterministe : le taux d'actualisation $\lambda > 0$, les fonctions $\ell : I\!R^n \times I\!R^m \longrightarrow I\!R$ et $f : I\!R^n \times I\!R^m \longrightarrow I\!R^n$, tandis qu'apparaissent $\sigma(\cdot,\cdot)$, application de $I\!R^n \times I\!R^m$ vers l'espace des matrices de taille $n \times r$, et w, brownien standard de dimension r. On suppose dans la suite ℓ, f et σ lipschitziens et bornés.

Rappelons qu'un mouvement brownien standard (scalaire) sur l'intervalle de temps $I\!R_+$ est une variable aléatoire $I\!R_+ \longrightarrow I\!R$ telle que (i) ses accroissements sont indépendants, (ii) $w(0)$ est gaussien de moyenne nulle, et (iii) si $0 \leq s \leq t < \infty$, alors $w(t) - w(s)$ est gaussien de moyenne nulle et variance $t - s$. Un brownien standard de dimension r est un vecteur aléatoire dont les composantes sont des mouvement brownien standard scalaires indépendants.

L'étude de ce problème comporte deux phases : l'analyse mathématique, qui conduit à une équation HJB avec un opérateur différentiel du second ordre, et l'analyse numérique de cette équation HJB. Nous allons commencer par présenter une version en temps discret du problème, qui permettra une dérivation formelle de l'équation HJB.

5.2.2 Problème discrétisé en temps

Soit $h_0 > 0$ le pas de temps. Considérons le problème de commande optimale stochastique en temps discret et espace continu :

$$(P_x^{h_0}) \qquad \begin{cases} \text{Min } I\!E \left\{ h_0 \displaystyle\sum_{k=0}^{\infty} (1+\lambda h_0)^{-k-1} \ell(y_k,u_k) \right\}; \\ \\ y_{k+1} = y_k + h_0 f(y_k,u_k) + \sqrt{h_0}\, \sigma(y_k,u_k)\, \delta w_k\,, \quad u_k \in U, \quad k \in I\!N; \\ \\ y_0 = x. \end{cases}$$

Ici $\delta w_k \in I\!R^r$ est un vecteur aléatoire dont les coordonnées sont des tirages indépendants de ± 1 avec probabilités égales, donc de moyenne nulle et variance unité. Le terme $\sqrt{h_0}$ fait que, pour h_0 assez petit, si la $i$ème ligne de $\sigma(y_k,u_k)$ n'est pas nulle, alors l'essentiel de la variation de la $i$ème composante de l'état est due au bruit. Par ailleurs si $0 \leq s \leq t < \infty$, $s = k_0 h_0$ et $t = k_1 h_0$, alors $\sum_{k=k_0}^{k_1-1} \delta w_k$ est une variable asymptotiquement gaussienne, de moyenne nulle et variance $t - s$, ce qui est cohérent avec le problème continu.

A la différence du cas déterministe, il faut préciser quelle information est disponible quand on prend la décision u^k à l'instant k. Par exemple, si les tirages sont connus d'avance, on se retrouve dans une situation déterministe. En général le tirage δw_k n'est pas déterminé jusqu'à l'instant $k+1$; l'information sur ce tirage et sur l'état y_k peut être totale, partielle ou nulle. Il y a donc une variété de situations possibles.

Dans la suite nous supposerons que la décision u_k se fait en connaissant l'état y_k, mais pas les tirages δw_i, pour $i \geq k$: c'est le cas dit de l'*observation complète*. Compte tenu de l'invariance en temps du problème, ceci conduit à chercher une commande sous forme de retour d'état (feedback). Autrement dit l'ensemble $\mathcal{U}$ des commandes admissibles est celui des applications $u = u(y)$ de $\mathbb{R}^n$ vers U. A $u \in \mathcal{U}$ est associé un coût $\mathcal{V}^{h_0}(x,u)$ vérifiant la relation suivante (noter que l'espérance ci-dessous se réduit à la somme de deux termes)

$$\mathcal{V}^{h_0}(x,u) = (1+\lambda h_0)^{-1}\left(h_0\ell(x,u) + \mathbb{E}\left(V(x+h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0)\right)\right). \quad (5.58)$$

On pose

$$V^{h_0}(x) := \inf_{u\in U} \mathcal{V}^{h_0}(x,u). \quad (5.59)$$

Le *principe de programmation dynamique* s'écrit

$$V^{h_0}(x) = (1+\lambda h_0)^{-1} \inf_{u\in U}\left\{h_0\ell(x,u) + \mathbb{E}\left(V(x+h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0)\right)\right\}. \quad (5.60)$$

Supposons V^{h_0} de classe C^2, et de dérivée seconde uniformément bornées sur $\mathbb{R}^n$, uniformément par rapport à h_0 assez petit. Alors

$$\begin{array}{rcl} \Delta & := & V^{h_0}(x+h_0 f(x,u) + \sqrt{h_0}\sigma(x,u)\delta w_0), \\ & = & V^{h_0}(x) + h_0 DV^{h_0}(x) f(x,u) + \sqrt{h_0} DV^{h_0}(x)\sigma(x,u)\delta w_0 \\ & & + \frac{1}{2} h_0 D^2 V^{h_0}(x)(\sigma(x,u)\delta w_0, \sigma(x,u)\delta w_0) + o(h_0). \end{array} \quad (5.61)$$

Si A est une matrice $n\times n$ et $z \in \mathbb{R}^n$, on a $z^T A z = \mathrm{trace} A z z^T$. Utilisant cette relation, il vient

$$D^2 V^{h_0}(x)(\sigma(x,u)\delta w_0, \sigma(x,u)\delta w_0) = \mathrm{trace}\left(D^2 V^{h_0}(x)\sigma(x,u)\delta w_0 \delta w_0^T \sigma(x,u)^T\right). \quad (5.62)$$

Posons

$$a(x,u) := \tfrac{1}{2}\sigma(x,u)\sigma(x,u)^T. \quad (5.63)$$

La matrice $n\times n$ $a(x,u)$ est symétrique et semi définie positive. Puisque w est de moyenne nulle et variance unité, on a, avec la relation précédente :

$$\mathbb{E}(\Delta) = V^{h_0}(x) + h_0 DV^{h_0}(x) f(x,u) + h_0\,\mathrm{trace}\left(D^2 V^{h_0}(x) a(x,u)\right) + o(h_0). \quad (5.64)$$

Noter que

$$\mathrm{trace}\left(D^2 V^{h_0}(x) a(x,u)\right) = \sum_{i,j=1}^{n} a_{ij}(x,u) D^2_{x_i x_j} V^{h_0}(x). \quad (5.65)$$

Introduisons le hamiltonien $\mathcal{H}^\sigma$:

$$\mathcal{H}^\sigma(x,p,Q) := \inf_{u\in U}\{\ell(x,u) + p\cdot f(x,u) + \mathrm{trace}(a(x,u)Q)\}. \quad (5.66)$$

Ici $p \in I\!R^n$ et Q est une matrice symétrique $n \times n$. L'exposant σ fait réference au terme du deuxième ordre qui fait la différence avec le cas déterministe, voir (3.17).

Combinant avec le principe de programmation dynamique (5.60), il vient :

$$\lambda V^{h_0}(x) = \mathcal{H}^\sigma(x,DV^{h_0}(x),D^2V^{h_0}(x)) + o(1). \tag{5.67}$$

Passant à la limite quand $h_0 \downarrow 0$, on obtient formellement l'équation HJB du problème en temps continu :

$$\lambda V(x) = \mathcal{H}^\sigma(x,DV(x),D^2V(x)), \tag{5.68}$$

ou encore

$$\lambda V(x) = \inf_{u\in U} \left\{ \ell(x,u) + f(x,u) \cdot DV(x) + \mathrm{trace}(a(x,u)D^2V(x)) \right\}. \tag{5.69}$$

Lorsque $\sigma(x,u)$ est identiquement nul, on retrouve bien l'équation HJB (3.22) du cas déterministe (avec ici $C = \emptyset$).

Dans le cas d'un problème avec horizon fini T et sans terme d'actualisation, une discussion analogue à celle de l'horizon infini permet d'obtenir une équation de Hamilton-Jacobi-Bellman du problème continu, dont est solution la fonction valeur en temps rétrograde

$$W(x,s) := V(x,T-s).$$

Cette équation s'écrit :

$$\left\{ \begin{array}{ll} D_tW(x,t) = \mathcal{H}^\sigma(x,D_xW(x,t),D^2_{xx}W(x,t)), & (x,t) \in I\!R^n \times]0,T[, \\ W(x,0) = \varphi(x), & \forall\, x \in I\!R^n, \end{array} \right. \tag{5.70}$$

ou encore

$$\begin{array}{ll} D_tW(x,t) = & \inf_{u\in U} \left\{ \ell(x,u) + f(x,u) \cdot DW(x,t) + \mathrm{trace}(a(x,u)D^2W(x,t)) \right\}, \\ & (x,t) \in I\!R^n \times]0,T[, \\ W(x,0) = & \varphi(x), \quad \forall\, x \in I\!R^n. \end{array} \tag{5.71}$$

Nous allons étudier la résolution numérique de cette équation par des schémas aux différences finies, en commençant par le cas d'un état scalaire.

5.2.3 Schémas monotones : dimension 1

On note h_0, h_1, etc les pas de discrétisation en temps et suivants les variables d'espace x_1, etc. Nous discutons les schémas de résolution de problèmes à horizon infini.

Présentons une extension de l'algorithme décentré, dans lequel on approxime la dérivée seconde en espace (suivant la direction de x_i) par $(D^d w_j^k - D^g w_j^k)/h_i$, soit la différence divisée centrée

$$D^{2,0}w_j^k := \frac{1}{h_i^2}(w_{j+1}^k - 2w_j^k + w_{j-1}^k).$$

Le schéma décentré s'écrit alors

$$\begin{array}{rl} \lambda v_j = & \inf\limits_{u\in U} \left\{ \ell(x_j,u) + f(x_j,u)_+ \dfrac{v_{j+1}-v_j}{h_1} + |f(x_j,u)_-| \dfrac{v_{j-1}-v_j}{h_1} \right. \\ & \left. + a(x_j,u) \dfrac{v_{j+1} - 2v_j + v_{j-1}}{h_1^2} \right\}. \end{array} \tag{5.72}$$

Introduisons un *pas de temps fictif* $h_0 > 0$, par lequel on multiplie l'équation ci-dessus. Ajoutant v_j à chaque membre, et ordonnant les expressions suivant v_{j-1}, v_{j+1} et v_{j+1}, on obtient l'expression équivalente

$$\begin{aligned}\lambda v_j := \inf_{u\in U}\Bigg\{ & h_0\ell(x_j,u) + \left(1 - \frac{h_0}{h_1}|f(x_j,u)| - 2\frac{h_0}{h_1^2}a(x_j,u)\right) v_j \\ & + \left(\frac{h_0}{h_1}|f(x_j,u)_-| + \frac{h_0}{h_1^2}a(x_j,u)\right) v_{j-1} + \left(\frac{h_0}{h_1}f(x_j,u)_+ + \frac{h_0}{h_1^2}a(x_j,u)\right) v_{j+1}\Bigg\}.\end{aligned} \tag{5.73}$$

On pose

$$\|f\|_\infty := \sup_{(x,u)\in I\!R\times U} |f(x,u)|; \quad \|a\|_\infty := \sup_{(x,u)\in I\!R\times U} |a(x,u)|. \tag{5.74}$$

Proposition 5.17 (i) *Le schéma* (5.72) *possède une solution unique, telle que*

$$\|v\|_\infty \le \lambda^{-1}\|\ell\|_\infty. \tag{5.75}$$

(ii) *Si* h_0 *vérifie la* condition de stabilité

$$\frac{h_0}{h_1}\|f\|_\infty + \frac{2h_0}{h_1^2}\|a\|_\infty^2 \le 1, \tag{5.76}$$

alors (5.73) *est une équation de point fixe contractant pour la norme uniforme, de rapport de contraction* $(1+\lambda h_0)^{-1}$.

Démonstration. La démonstration est semblable à celle de la proposition 4.6. La condition de stabilité assure que, dans la formule (5.73), les poids de v_j et $v_{j\pm1}$ sont positif, ce qui permet d'établir que c'est une équation de point fixe contractant et d'obtenir l'estimation (5.75). ■

Remarque 5.18 Le terme dominant dans la condition de stabilité est lié à f si h_1 est grand par rapport à $2\|a\|_\infty/\|f\|_\infty$ (discrétisation spatiale grossière), et au terme de diffusion si h_1 est grand par rapport à $2\|a\|_\infty/\|f\|_\infty$ (discrétisation spatiale fine). Dans ce dernier cas, le pas de temps maximum respectant la condition de stabilité est de l'ordre de $\frac{1}{2}h_1^2/\|a\|_\infty$, donc beaucoup plus petit que dans le cas déterministe (où il vaut $h_1/\|f\|_\infty$).

Remarque 5.19 La condition de stabilité assure la positivité des poids de v_j et $v_{j\pm1}$ dans (5.73), ce qui permet de reconnaître dans cette expression le principe de programmation dynamique du problème de commande d'une chaîne de Markov dont les probabilités de transition sont précisément les poids de v_j et $v_{j\pm1}$.

Remarque 5.20 L'étude de la convergence de ce schéma est trop complexe pour être traitée ici. On se reportera aux notes de fin de chapitre.

Dans le cas de dimension d'espace supérieure à 1, on sait seulement donner des réponses *partielles* au problème de discrétisation par différence finie de l'équation HJB. Nous allons poser le problème et établir quelques résultats.

5.2.4 Différences finies classiques

Nous abordons l'études des schémas de discrétisation pour le cas de la dimension d'espace $n > 1$ par des schémas de différences finies. Notons D_i les dérivées par rapport à x_i, et on adopte le même type de convention pour les dérivées d'ordre supérieur. Pour approximer D_{ii} on utilise encore la formule centrée

$$D^2_{ii} v_j \approx \frac{v_{j+e_i} - 2v_j + v_{j-e_i}}{h_i^2}.$$

Pour alléger les formules il convient de noter $\delta_{\pm i}$, $\delta_{\pm,i\pm k}$, etc les opérateurs de translation de $\pm$ une coordonnée dans la direction i, k, etc; ainsi

$$\delta_i v_j = v_{j+e_i}, \quad \delta_{i,-k} v_j = v_{j+e_i-e_k}.$$

Avec cette notation l'approximation de D_{ii} est

$$D^2_{ii} \approx \frac{\delta_i - 2\delta_0 + \delta_{-i}}{h_i^2}.$$

Pour le calcul des dérivées croisées ($i \neq j$), plusieurs choix sont possibles. Par exemple, utilisant le développement, pour Φ régulier,

$$\begin{array}{rl} \Phi(x + h_i e_i + h_k e_k) = & \Phi(x) + D\Phi(x)(h_i e_i + h_k e_k) + \\ & \frac{1}{2} D^2\Phi(x)((h_i e_i + h_k e_k),(h_i e_i + h_k e_k)) + o(h_i^2 + h_k^2), \end{array} \tag{5.77}$$

et procédant de même pour $\Phi(x + h_i e_i)$ et $\Phi(x + h_k e_k)$, on déduit le choix

$$D^2_{ik} \approx \frac{\delta_{i,k} + \delta_0 - \delta_i - \delta_k}{h_i h_k},$$

qui fait intervenir les quatre points du "rectangle en haut à droite". On peut écrire une formule similaire faisant intervenir les points du rectangle opposé :

$$D^2_{ik} \approx \frac{\delta_{-i,-k} + \delta_0 - \delta_{-i} - \delta_{-k}}{h_i h_k}.$$

Il est classique de centrer l'estimation en prenant la moyenne des deux, ce qui donne

$$D^2_{ik} \approx \frac{\delta_{i,k} + \delta_{-i,-k} + 2\delta_0 - \delta_i - \delta_k - \delta_{-i} - \delta_{-k}}{2h_i h_k}. \tag{5.78}$$

Mais on peut aussi bien faire intervenir les estimations basées sur les deux autres rectangles :

$$D^2_{ik} \approx \frac{\delta_i + \delta_k + \delta_{-i} + \delta_{-k} - \delta_{i,-k} - \delta_{-i,k} - 2\delta_0}{2h_i h_k}. \tag{5.79}$$

Le point important est que ces deux formules font apparaître les points $\delta_{\pm i,\pm k}$ avec des poids positifs dans le premier cas, et négatifs dans le second. Soit $\hat{D}^{x,u}$ la matrice $n \times n$ d'opérateurs aux différences définie par

$$\hat{D}^{x,u}_{ik} = \begin{cases} \dfrac{\delta_i - 2\delta_0 + \delta_{-i}}{h_i^2} & \text{si } \ i = k, \\ \dfrac{\delta_{i,k} + \delta_{-i,-k} + 2\delta_0 - \delta_i - \delta_k - \delta_{-i} - \delta_{-k}}{2h_i h_k} & \text{si } \ a_{ik}(x,u) \geq 0, \\ \dfrac{\delta_i + \delta_k + \delta_{-i} + \delta_{-k} - \delta_{i,-k} - \delta_{-i,k} - 2\delta_0}{2h_i h_k} & \text{sinon.} \end{cases}$$

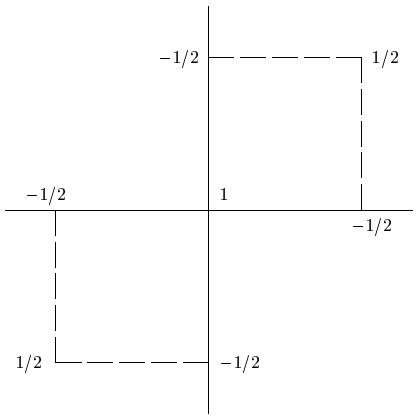


FIG. 5.1 – *Poids de l'approximation de D^2_{ij} : cas où $a_{ij} > 0$*

Pour les termes du premier ordre, on reprend le principe du décentrage exposé dans le cas de la commande optimale déterministe : à (x,u), associons $D^{\eta(x_j,u)} \in I\!R^n$ défini par

$$D^{\eta_i(x,u)} = \begin{cases} \dfrac{v_{j+e_i} - v_j}{h_i} & \text{si} \quad f_i(x,u) \geq 0, \\ \dfrac{v_j - v_{j-e_i}}{h_i} & \text{sinon.} \end{cases} \tag{5.80}$$

Considérons le schéma discret

$$\lambda v_j = \min_{u\in U} \left\{ \ell(x_j,u) + f(x_j,u) \cdot D^{\eta(x_j,u)} v_j + \sum_{i,k=1}^{n} a_{ik}(x_j,u) \hat{D}^{x,u}_{ik} v_j \right\}. \tag{5.81}$$

Multipliant l'équation par un pas de temps fictif h_0, ajoutant v_j à chaque membre, et réordonnant les expressions, il vient

$$\begin{aligned} \lambda v_j = \min_{u\in U} \{ & h_0 \ell(x_j,u) \\ & + \left(1 - \sum_{i=1}^{n} \frac{h_0}{h_i} |f_i(x_j,u)| - 2 \sum_{i=1}^{n} \frac{h_0}{h_i^2} |a_{ii}(x_j,u)| + \sum_{i\neq k} \frac{h_0}{h_i h_k} |a_{ik}(x_j,u)| \right) v_j \\ & + \sum_{i=1}^{n} \left(\frac{h_0}{h_i} |f_i(x_j,u)_-| + \frac{h_0}{h_i^2} a_{ii}(x_j,u) - \sum_{k\neq i} \frac{h_0}{h_i h_k} |a_{ik}(x_j,u)| \right) v_{j-e_i} \\ & + \sum_{i=1}^{n} \left(\frac{h_0}{h_i} f_i(x_j,u)_+ + \frac{h_0}{h_i^2} a_{ii}(x_j,u) - \sum_{k\neq i} \frac{h_0}{h_i h_k} |a_{ik}(x_j,u)| \right) v_{j+e_i} \\ & + \sum_{i>k} \frac{h_0}{h_i h_k} \left[a_{ik}(x_j,u)_+ (v_{j+e_i+e_k} + v_{j-e_i-e_k}) + |a_{ik}(x_j,u)_-| (v_{j+e_i-e_k} + v_{j-e_i+e_k}) \right] \Bigg\}. \end{aligned} \tag{5.82}$$

On peut introduire une mise à l'échelle de f et a :

$$f_i^h(x,u) := \frac{f_i(x,u)}{h_i}; \qquad a_{ij}^h(x,u) := \frac{a_{ij}(x,u)}{h_i h_j}; \tag{5.83}$$

d'où l'expression équivalente

$$\begin{aligned}
(1+\lambda h_0)v_j = \min_{u\in U}\{ & h_0\,\ell(x_j,u) \\
& + \left(1 - h_0\sum_{i=1}^n |f_i^h(x_j,u)| - 2h_0\sum_{i=1}^n |a_{ii}^h(x_j,u)| + h^0\sum_{i\neq k}|a_{ik}^h(x_j,u)|\right) v_j \\
& + h_0\sum_{i=1}^n \left(|f_i^h(x_j,u)_-| + a_{ii}^h(x_j,u) - \sum_{k\neq i}|a_{ik}^h(x_j,u)|\right) v_{j-e_i} \\
& + h_0\sum_{i=1}^n \left(f_i^h(x_j,u)_+ + a_{ii}^h(x_j,u) - \sum_{k\neq i}|a_{ik}^h(x_j,u)|\right) v_{j+e_i} \\
& + h_0\sum_{i>k}\left[a_{ik}^h(x_j,u)_+(v_{j+e_i+e_k} + v_{j-e_i-e_k}) + |a_{ik}^h(x_j,u)_-|(v_{j+e_i-e_k} + v_{j-e_i+e_k})\right]\Bigg\}.
\end{aligned} \tag{5.84}$$

Proposition 5.21 *On suppose que les pas d'espace $h_1,\ldots,h_n$ sont tels que, pour tout $(x,u)\in \mathbb{R}\times U$, la matrice de terme général $a_{ik}^h(x,u)$ est diagonale dominante. Alors*
(i) *Le schéma* (5.81) *possède une solution unique v, telle que*

$$\|v\|_\infty \leq \lambda^{-1}\|\ell\|_\infty. \tag{5.85}$$

(ii) *Si h_0 vérifie la* condition de stabilité

$$h_0\left[\sum_{i=1}^n \frac{|f_i(x_j,u)|}{h_i} + \sum_{i=1}^n\left(2\frac{|a_{ii}(x_j,u)|}{h_i^2} - \sum_{k\neq i}\frac{|a_{ik}(x_j,u)|}{h_i h_k}\right)\right] \leq 1, \tag{5.86}$$

alors (5.73) *est une équation de point fixe contractant pour la norme uniforme, de rapport de contraction $(1+\lambda h_0)^{-1}$.*

Démonstration. La démonstration est une extension simple de celle de cas monodimensionnel (proposition 5.17). ■

Si la matrice $a^h(x,u)$ n'est pas diagonale dominante, le schéma présenté ci-dessus ne convient pas. Une solution possible est de faire intervenir davantage de points dans le schéma.

Quand h tend vers 0 de manière à respecter la condition de diagonale dominante de a^h, on obtient la convergence des valeurs discrètes vers la valeur du problème continu : voir [21].

5.2.5 Différences finies généralisées

Dans cette approche, qui généralise la méthode usuelle de différences finies présentée dans la section précédente, le point de départ est l'approximation de la dérivée seconde de la fonction valeur suivant une direction quelconque.

Soit $\Phi : I\!R^n \longrightarrow I\!R$ de classe C^2. La dérivée seconde de Φ en $x \in I\!R^n$ dans la direction $d \in I\!R^n$ est par définition la quantité

$$D^2\Phi(x)(d,d) = \sum_{i,k=1}^{n} D^2_{x_i x_k}\Phi(x) d_i d_k.$$

Il vient avec la formule de Taylor

$$D^2\Phi(x)(d,d) = \lim_{t\downarrow 0} \frac{\Phi(x+td) - 2\Phi(x) + \Phi(x-td)}{t^2}.$$

En particulier, étant donné $\xi \in \mathbb{Z}^n$, notons

$$\Delta_\xi \Phi := \Phi(x_{j+\xi}) - 2\Phi(x_j) + \Phi(x_{j-\xi}).$$

Il vient, pour tout $j \in \mathbb{Z}^n$,

$$\Delta_\xi \Phi(x_j) = \sum_{i,k=1}^{n} h_i h_k \xi_i \xi_k D^2_{x_i x_k}\Phi(x_j) + o(\|h\|^2). \tag{5.87}$$

Ainsi on peut approcher la courbure de Φ, suivant une direction égale à la différence entre deux points de la grille discrète, par une combinaison des valeurs de Φ en trois points de la grille. On peut alors se poser le problème d'approcher la partie principale (du second ordre) de l'opérateur différentiel de l'équation HJB par une combinaison de tels termes. Il s'agit de trouver des coefficients $\alpha^u_{j,\xi}$ tels que :

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \Delta_\xi \Phi(x_j) = \sum_{i,k=1}^{n} a_{ik}(x_j,u) \Phi_{x_i x_k}(x_j) + o(1). \tag{5.88}$$

Ici $\mathcal{S}$ est une partie finie de $\mathbb{Z}^n$, qui représente (à la translation j près) les coordonnées des points entrant dans le schéma. Nous verrons qu'il convient de prendre les coefficients $\alpha^u_{j,\xi}$ positifs pour obtenir la monotonie du schéma.

Utilisant (5.87), on voit que ceci sera satisfait pour toute fonction Φ si

$$\alpha^u_{j,\xi} = O((\inf_i h_i)^{-2}), \tag{5.89}$$

et

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} h_i h_k \xi_i \xi_k = a^h_{ik}(x_j,u) + o(1), \qquad \text{pour tout} \quad i,k, \tag{5.90}$$

ou encore

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \xi\xi^T = a^h(x_j,u) + o(1). \tag{5.91}$$

Le schéma correspondant (de discrétisation de l'équation HJB) est

$$\lambda v_j = \inf_{u\in U} \left\{ \ell(x_j,u) + f(x_j,u)\cdot D^{\eta(x_j,u)} v_j + \sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \Delta_\xi v_j \right\}, \quad j \in \mathbb{Z}^n. \tag{5.92}$$

Définition 5.22 On dira que le schéma (5.92) est *consistant* si (5.91) est satisfait, et *fortement consistant* si

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \xi\xi^T = a^h(x_j,u). \tag{5.93}$$

La vérification de la condition de consistance (qui ne va pas de soi) fait l'objet de la section suivante.

Remarque 5.23 La relation ci-dessus donne une estimation de la taille des coefficients, qui implique (5.89). En effet, puisque ξ a des coordonnées entières, la matrice $\xi\xi^T$ a des éléments diagonaux supérieurs ou égaux à un. Un schéma fortement consistant satisfait donc

$$\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \leq \text{trace } a^h(x_j,u) = O((\inf_i h_i)^{-2}). \tag{5.94}$$

La forme de point fixe correspondante est (comme toujours) obtenue en multipliant la relation (5.92) par un pas de temps fictif h_0, puis en ajoutant v_j à chaque membre, et enfin en divisant par $1+h_0\lambda$. Reprenant la notation f^h définie en (5.83), on obtient l'expression suivante, à comparer à (4.23) dans le cas déterministe:

$$\begin{aligned} v_j = (1+\lambda h_0)^{-1} \inf_{u\in U} \Bigg\{ & h_0\ell(x_j,u) + \left(1 - h_0 \sum_{i=1}^n |f_i^h(x_j,u)| - 2h_0 \sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi}\right) v_j \\ & + h_0 \sum_{i=1}^n f_i^h(x_j,u)_+ v_{j+e_i} + h_0 \sum_{i=1}^n |f_i^h(x_j,u)_-| v_{j-e_i} + h_0 \sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi}(v_{j-\xi}+v_{j+\xi}) \Bigg\}. \end{aligned} \tag{5.95}$$

Comme dans le cas déterministe, il apparaît que le membre de droite représente une application contractante, de constante $(1+\lambda h_0)^{-1}$, si le coefficient de v_j est positif, ce qui est assuré si la condition de stabilité suivante est satisfaite :

$$h_0 \left(\sum_{i=1}^n \frac{\|f_i\|}{h_i} + 2 \sup_{j\in\mathbb{Z}^n, u\in U} \left(\sum_{\xi\in\mathcal{S}} \alpha^u_{j,\xi} \right) \right) \leq 1. \tag{5.96}$$

On peut combiner cette relation avec (5.94) pour en déduire une estimation du pas de temps : $h_0 = O((\inf_i h_i)^{-2})$.

5.2.6 Analyse de la condition de consistance forte

La condition de consistance forte (5.93) revient, puisque les coefficients $\alpha^u_{j,\xi}$ doivent être positifs, à vérifier que $a^h(x_j,u)$ appartient au cône engendré par l'ensemble $\{\xi\xi^T; \xi\in\mathcal{S}\}$. Nous allons caractériser ce cône dans quelques situations simples. Pour cela, quelques définitions s'imposent.

Définition 5.24 Soit $q\in\mathbb{N}$, $q>0$. (i) On dit que $C\subset\mathbb{R}^q$ est un *cône* si, pour tout $t>0$ et $c\in C$, on a $tc\in C$. (ii) Soient $c_1,\ldots,c_r$ dans $\mathbb{R}^q$. On appelle *cône convexe* C engendré par $c_1,\ldots,c_r$ l'ensemble des combinaisons linéaires positives de $c_1,\ldots,c_r$. On dit que $c_1,\ldots,c_r$ est un générateur de C. (iii) On appelle *générateur minimal* de C un générateur de C ne contenant pas strictement un générateur de C.

Définition 5.25 Soit C un cône convexe fermé de $\mathbb{R}^q$. On appelle *cône polaire* de C l'ensemble

$$C^+ := \{y\in\mathbb{R}^q;\ \ y\cdot x\geq 0,\ \ \text{pour tout}\ \ x\in C\}. \tag{5.97}$$

C'est un cône convexe fermé.

Voici un résultat important d'analyse convexe, que nous admettrons (voir par exemple [27].

Proposition 5.26 *Soit C un cône convexe fermé. Alors* (i) *il coïncide avec son cône bipolaire $(C^+)^+$,* (ii) *Si C a un générateur fini, il en est de même pour C^+.*

Il résulte de cette proposition que, si C est un cône convexe fermé de générateur fini, et il existe donc un générateur fini $c_1^*, \ldots, c_{r'}^*$ du cône polaire, alors C est caractérisé par les inégalités linéaires en nombre fini

$$C = \{x \in I\!R^q;\ \ c_1^* \cdot x \geq 0,\ \ i = 1, \ldots, r'\}. \tag{5.98}$$

On notera $C(\mathcal{S})$ le cône engendré par les $\{\xi\xi^T,\ \xi \in \mathcal{S}\}$. Considérons le cas où $\mathcal{S}$ est de la forme $\mathcal{S}_p^n$, avec

$$\mathcal{S}_p^n := \left\{ \varsigma \in \{-1,0,1\}^n; \sum_{i=1}^n |\varsigma_i| \leq p \right\}. \tag{5.99}$$

Autrement dit, on considère les transitions vers les points dont les coordonnées diffèrent d'au plus 1 (les voisins immédiats), avec au plus p coordonnées différentes.

Proposition 5.27 *On a les caractérisations suivantes :*
(i) *Pour tout $n > 0$, $C(\mathcal{S}_1^n)$ est l'ensemble des matrices diagonales semi définies positives.*
(ii) *Pour tout $n > 0$, $C(\mathcal{S}_2^n)$ est l'ensemble des matrices à diagonale dominante :*

$$\mathcal{C}(\mathcal{S}_2^n) = \left\{ A \in \mathcal{M}^{n\times n};\ \ A = A^T;\ \ A_{ii} \geq \sum_{j\neq i} |A_{ij}| \right\}. \tag{5.100}$$

(iii) *$A \in C(\mathcal{S}_3^3)$ si et seulement si, pour tout i,j dans $1, \ldots, n$ et p, q dans $\{0,1\}$:*

$$\left\{ \begin{array}{lcl} A_{ii} & \geq & |A_{ij}|, \\ A_{ii} + A_{jj} & \geq & (-1)^p A_{ik} + (-1)^q A_{jk} + 2(-1)^{p+q+1} A_{ij}. \end{array} \right. \tag{5.101}$$

Démonstration. Le point (i) est immédiat, et les points (ii) et (iii) résultent de l'analyse de [9]. ■

Remarque 5.28 Les résultats de cette section sont liés aux travaux récents de [9]. Une des questions ouvertes est le calcul rapide des coefficients $\alpha_{j,\xi}^u$, en particulier pour les dimensions 2 et 3.

5.3 Notes

La commande optimale de chaînes de Markov est discutée dans J. P. Quadrat [28]. E. Altman [3] étudie les problèmes avec contraintes en espérance. Le cas ergodique fait l'objet d'un chapitre de H.J. Kushner et P.G. Dupuis [21].

W. H. Fleming et R. Rishel [17] donnent une introduction générale à la théorie de la commande optimale déterministe et stochastique. L'approche par solutions de viscosité est introduite dans P.L. Lions [25]; on en trouvera une synthèse dans W.H. Fleming et

H.M. Soner [18]. J.L. Lions et A. Bensoussan [24] présentent l'approche de la commande stochastique par les techniques variationnelles d'équations aux dérivés partielles.

Les méthodes numériques pour la commande stochastique sont exposées dans H.J. Kushner et P.G. Dupuis [21]. On y trouvera en particulier une discussion d'une méthode d'approximation par chaîne de Markov qui inclut les différences finies généralisées. Pour les problèmes de très grande taille il peut être utile d'employer des méthodes multigrille, voir M. Akian [1]. De nombreuses méthodes numériques, dans un cadre de problèmes de finance, sont exposées dans L.C.G. Rogers et D. Talay [29].

Bibliographie

[1] M. Akian. Analyse de l'algorithme multigrille FMGH de résolution d'équations d'Hamilton-Jacobi-Bellman. In A. Bensoussan and J.-L. Lions, editors, *Analysis and optimization of systems (Antibes, 1990)*, volume 144 of *Lecture Notes in Control and Information Sciences*, pages 113–122. Springer Verlag, Berlin, 1990.

[2] V. Alexéev, V. Tikhomirov, and S. Fomine. *Commande optimale*. Mir, Moscow, 1982. Edition originale : Mir, Moscou, 1979.

[3] E. Altman. *Constrained Markov decision processes*. Chapman and Hall, Boca Raton, 1999.

[4] M. Bardi and I. Capuzzo-Dolcetta. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Systems and Control: Foundations and Applications. Birkhäuser, Boston, 1997.

[5] G. Barles. *Solutions de viscosité des équations de Hamilton-Jacobi*, volume 17 of *Mathématiques et Applications*. Springer, Paris, 1994.

[6] G. Barles and P. E. Souganidis. Convergence of approximation schemes for fully nonlinear second order equations. *Asymptotic Analysis*, 4:271–283, 1991.

[7] R. Bellman. *Dynamic programming*. Princeton University Press, Princeton, 1961.

[8] D. Bertsekas. *Dynamic programming and optimal control (2 volumes)*. Athena Scientific, Belmont, Massachusetts, 1995.

[9] J. F. Bonnans and H. Zidani. Consistency of generalized finite difference schemes for the stochastic HJB equation. *SIAM J. Numerical Analysis*, 41:1008–1021, 2003.

[10] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer-Verlag, New York, 2000.

[11] H. Brézis. *Analyse fonctionnelle*. Masson, Paris, 1983.

[12] A. E. Bryson and Y.-C. Ho. *Applied optimal control*. Hemisphere Publishing, New-York, 1975.

[13] F.H. Clarke. *Optimization and nonsmooth analysis*. Wiley, New York, 1983.

[14] M. G. Crandall and P.-L. Lions. Two approximations of solutions of Hamilton-Jacobi equations. *Mathematics of Computation*, 43(167):1–19, 1984.

[15] M.G. Crandall and P.-L. Lions. Viscosity solutions of Hamilton Jacobi equations. *Bull. American Mathematical Society*, 277:1–42, 1983.

[16] I. Capuzzo Dolcetta and H. Ishii. Approximate solutions of the Bellman equation of deterministic control theory. *Appl. Math. Optim.*, 11:161–181, 1984.

[17] W. H. Fleming and R. Rishel. *Deterministic and stochastic optimal control*, volume 1 of *Applications of mathematics*. Springer, New York, 1975.

[18] W. H. Fleming and H.M. Soner. *Controlled Markov processes and viscosity solutions*. Springer, New York, 1992.

[19] H. Frankowska. Value function in optimal control, 2001. Lecture notes, Summer School on Mathematical Control Theory, Trieste.

[20] A.D. Ioffe and V.M. Tihomirov. *Theory of Extremal Problems*. North-Holland Publishing Company, Amsterdam, 1979. Russian Edition: Nauka, Moscow, 1974.

[21] H. J. Kushner and P. G. Dupuis. *Numerical methods for stochastic control problems in continuous time*, volume 24 of *Applications of mathematics*. Springer, New York, 2001. Second edition.

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

[23] G. Leitmann. *An introduction to optimal control*. Mc Graw Hill, New York, 1966.

[24] J.-L. Lions and A. Bensoussan. *Application des inéquations variationnelles en contrôle stochastique*, volume 6 of *Méthodes mathématiques de l'informatique*. Dunod, Paris, 1978.

[25] P.-L. Lions. Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations. Part 2: viscosity solutions and uniqueness. *Communications in partial differential equations*, 8:1220–1276, 1983.

[26] I. McCausland. *Introduction to optimal control*. J. Wiley, New York, 1969.

[27] G.L. Nemhauser, A.H.G. Rinnoy Kan, and M.J. Todd, editors. *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*. North-Holland, Amsterdam, 1989.

[28] J.P. Quadrat. *Décision et commande en présence d'incertitude*. Ecole Polytechnique, Palaiseau, 1994. Polycopié de cours.

[29] L. C. G. Rogers and D. Talay, editors. *Numerical methods in finance*. Cambridge University Press, 1997.

[30] W. Rudin. *Real and complex analysis*. Mc Graw-Hill, New York, 1987.