MAP 561

Introduction à l'automatique

Introduction to Automatic Control

Ugo Boscain et Yacine Chitour

Notes de cours

Édition 2014/2015

Note : this text is divided in two parts. The French text and the English text. The French text was used in the previous 5 years. The English text is a draft, but it is more close to the contents of the course. Chapters 1,2,3,4 of the English text are translations of Chapters 1,2,3,4 of the French text. This is not the case for the other chapters. (However Chapter 8 of the English text is a rough translation of Chapter 5 of the French text.)

Table des matières

Avant-propos

Le cours « Introduction à l'automatique » a pour objectif de présenter les concepts de base de l'automatique linéaire. On utilise l'approche par représentation d'état, qui repose sur les équations différentielles ordinaires.

Ces notes de cours se composent de sept parties. La première est une rapide présentation de l'automatique à travers l'étude d'un exemple classique, celui de la commande d'un bras de robot et fait l'objet du chapitre 1. En particulier, on expliquera pourquoi les équations différentielles ordinaires (EDO) sont utilisées et la nécessité d'une bonne connaissance de leurs propriétés fondamentales avant d'aborder la résolution de problèmes en automatique. Il est à souligner que cette partie est relativement complète et que certains points abordés ont déja été vus. Dans ce cas, on pourra passer rapidement sur ces points.

La seconde partie, développée dans le chapitre 2, est quant à elle consacrée à l'étude des équations différentielles ordinaires (EDO) ainsi qu'à leur utilisation pour la modélisation en physique, mécanique, économie, biologie... L'accent est principalement mis sur deux points : tout d'abord la notion de stabilité dont l'importance, pour de nombreux problèmes pratiques, est comparable à celle de la connaissance effective des solutions ; et ensuite une description détaillée des solutions des EDO linéaires à coefficients constants.

Dans la troisième partie (chapitre 3), sont abordées les notions essentielles de l'automatique telles que la commandabilité, l'observabilité et la stabilisation des systèmes linéaires commandés. On établira en détail comment se caractérisent ces propriétes sous la forme de critères classiques tels que celui de Kalman ou le théorème de placement de pôles. On présentera aussi une solution effective à la question de commandabilité grâce à la sortie de Brunovski ainsi que l'observateur de Luenberger et le principe de séparation qui en résulte.

Dans le chapitre 4, on abordera l'étude de la commandabilité pour les systèmes non linéaires, en se contentant de ceux qui sont affines en la commande. On mettra en évidence la notion de crochet de Lie, centrale pour décrire l'ensemble atteignable. On donnera les principaux résultats généraux comme le théorème de Krener pour l'*accessibilité* et les théoèmes de Chow-Rashevski, Nagano et Sussman pour ce qui est de la commandabilité et la description de l'orbite.

Les trois dernière parties sont consacrées à des questions de commande optimale. La cinquième et sixième parties traitent respectivement de deux aspects classiques de la commande optimale : (a) la commande linéaire quadratique et son application la plus

fameuse, le filtre de Kalman ; (b) le principe du maximum de Pontryagin (PMP) appliqué à la minimisation du temps pour les systèmes linéaires.

Enfin, la septième et dernière partie traite de questions de contrôle optimal pour des systèmes non linéaires avec un énoncé général du PMP et ensuite son application pour caractériser les trajectoires optimales dans le cadre sous-riemannien et pour l'étude de la synthèse optimale pour des systèmes en dimension deux.

Les résultats sont parfois accompagnés de leur preuve. Lorsque celle-ci n'est pas utile à la compréhension du cours, elle est écrite en petits caractères (petits comme ceci) et est précédée du symbole *. Le même traitement (petits caractères et symbole *) est appliqué aux parties les plus avancées du document, qui ne seront pas traitées en cours. Il n'est cependant pas interdit de les lire. . .Le symbole " :=" signifie que ce qui est à gauche du symbole est défini par ce qui est à droite.

Enfin, le présent cours a été enseigné par Pierre Rouchon et Frédéric Bonnans entre 1994 et 2004 sous le titre « Commande et Optimisation de systèmes dynamiques» et fait l'objet de l'ouvrage $[Rou - Bo]$ extrêmement riche en exemples et qu'il est fortement conseillé de consulter. De plus les auteurs du présent document remercient Frédéric Jean et Emmanuel Trélat pour avoir autorisé de nombreux emprunts au très beau cours "Equations différentielles et fondements de l'automatique" ainsi qu'à l'excellent ouvrage "Contrôle optimal : théorie et applications" chez Vuibert.

Par ailleurs, ces notes de cours sont loin d'être parfaites et les auteurs seront grés à toute personne leur signalant des corrections à effectuer.

Chapitre 1

Introduction

L'automatique (ou théorie du contrôle) est la science qui traite des lois de régulation des systèmes commandés. Commander un objet (on dit aussi le contrôler ou l'asservir) signifie influencer son comportement pour lui faire effectuer une tâche définie à l'avance. Afin de réaliser en pratique cette "influence", les ingénieurs ont mis au point des mécanismes appropriés faisant appel à des principes théoriques généraux, eux-mêmes s'exprimant à l'aide de divers outils mathématiques. Ainsi, ces mécanismes vont du régulateur à boules de Watt (pour les moteurs à vapeur) aux microprocesseurs les plus sophistiqués que l'on peut trouver dans les CD players, les automobiles, ou encore dans les robots industriels ou les pilotes automatiques des avions. L'étude de ces mécanismes et leur interaction avec l'objet à commander est le sujet de ce cours.

Nous allons illustrer plus en détail notre propos à l'aide d'un exemple simple issu de la robotique. Nous reprenons la description qui en est faite dans $[Rou - Bo]$. Il s'agit d'un bras rigide tournant dans un plan vertical autour d'un axe horizontal (Ox) et ce dernier est équipé d'un moteur délivrant un couple variable $u \in \mathbb{R}$ que l'on peut choisir *arbitrairement* à chaque instant : u est la *commande* du système (ou encore *entrée*). La position géométrique du système est décrite par un angle $\theta \in S^1$, le cercle unité c'est-à-dire que S^1 est la "sphère" de dimension 1. La dynamique du système est obtenue à partir de la conservation du moment cinétique autour de l'axe (Ox) :

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

avec m la masse du bras, J son moment d'inertie par rapport à l'axe (Ox), l la distance du centre de gravité à (Ox) et g l'accélération dûe à la pesanteur. On aura reconnu l'équation d'un pendule pesant sans frottement. Bien que les constantes qui interviennent dans ce problème jouent un rôle prépondérant en pratique, on supposera dans la suite que $mlg = J = 1$. La dynamique du bras est donc

$$\ddot{\theta}(t) + \sin\theta(t) = u(t). \tag{1.2}$$

Pour un angle θ_r fixé, un objectif possible de commande est d'amener le bras à l'angle θ_r et de l'y maintenir ensuite. On dira alors que le bras est en position d'équilibre à $\theta = \theta_r$ et l'objectif de commande est de *stabiliser* le système en θ_r. Plus généralement, on pourra se donner comme objectif de *suivre une trajectoire* de référence $\theta_r(\cdot)$ qui vérifie la dynamique (1.2) pour une commande de référence $u_r(\cdot)$ avec les fonctions du temps θ_r, u_r définies sur un intervalle de temps $[0, T_r]$ (T_r pouvant être fini ou non).

Remarque. Remarquons que l'objectif initial est un cas particulier du suivi de trajectoire puisque amener le bras à un angle $\theta = \theta_1$ l'y maintenir ensuite est équivalent à suivre la trajectoire de (1.2) associée à $\theta_r(\cdot) \equiv \theta_r$ et $u_r(\cdot) \equiv \sin\theta_r$.

Nous allons maintenant décrire les étapes successives que suit l'automaticien pour résoudre le problème.

1.1 Mise sous forme d'état et définition du système commandé

Un intervalle de temps $[0, T]$ doit être fixé pour toutes les fonctions temporelles que nous envisagerons. Ici, il est naturel de prendre $T = T_r$, temps auquel on veut amener le bras en position déquilibre. Afin de connaitre l'évolution du mouvement lorsque l'on applique un couple $u(\cdot) : [0, T] \to \mathbb{R}$, il faut intégrer l'équation différentielle (1.2), qui est du second ordre donc. Pour cela, il faut connaitre, à l'instant $t = 0$, la position angulaire $\theta(0) = \theta_0$ et la vitesse angulaire $\dot\theta(0) = \omega_0$. La paire (θ_0, ω_0) représente les conditions initiales du système différentiel du premier ordre suivant, obtenu à partir de (1.2) :

$$(S) \left\{ \begin{array}{rcl} \dot\theta(t) & = & \omega, \\ \dot\omega(t) & = & -\sin\theta(t) + u. \end{array} \right. \tag{1.3}$$

A l'instant t, l'état du système est donc uniquement déterminé par la donnée de $(\theta(t), \omega(t))$ et $u(t)$. La variable $x = (\theta, \omega)$ forme l'état du système, qui est donc un point de $S^1 \times \mathbb{R}$, et le couple de fonctions $t \mapsto (x(t), u(t))$ avec $t \in [0, T]$ est appelé trajectoire du système.

De manière plus classique, une trajectoire est la *réponse* du système à la commande $u(\cdot)$ (appelée aussi entrée).

La dynamique (S) peut être réécrite en termes de x comme suit :

$$(S) \quad \dot x(t) = F(x(t), u(t)), \tag{1.4}$$

avec $F : S^1 \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^2$ est *le champ de vecteurs* qui, à tout point $(x, u) = (\theta, \omega) \in S^1 \times \mathbb{R} \times \mathbb{R}$ associe le vecteur $(\omega, -\sin\theta + u) \in \mathbb{R}^2$.

Enfin, on appèlera système commandé (associé au bras de robot) (Σ), l'ensemble des données suivantes : l'espace d'état $S^1 \times \mathbb{R}$, l'espace de commande $\mathbb{R}$, la dynamique

(1.4) et la classe des commandes admissibles Ad, c'est-à-dire l'ensemble des fonctions $u : [0, T] \to \mathbb{R}$ (continues, continues par morceaux, polynômes, fonctions bornées etc.).

Comme étude préliminaire à ce système (et aux objectifs de commande que l'on veut atteindre), il est impératif de comprendre ce qui se passe si "l'on ne fait rien", c'est-à-dire en mettant $u = 0$ dans (1.4). On est alors conduit à étudier l'équation différentielle ordinaire (EDO) définie sur $S^1 \times \mathbb{R}$ par

$$(S)_0 \begin{cases} \dot{\theta}(t) &= \omega, \\ \dot{\omega}(t) &= -\sin\theta(t), \end{cases} \tag{1.5}$$

ou encore $\dot{x} = F(x(t), 0)$. En automatique, cette étape est appelée *étude de la réponse libre*. Il est particulièrement important de savoir quel est le comportement des trajectoires libres lorsque t tend vers l'infini. On parle alors d'étude de la stabilité asymptotique de l'EDO (1.5). Par exemple, est-ce que les trajectoires peuvent converger vers un point de $S^1 \times \mathbb{R}^2$? De tels points sont appelés *points d'équilibre* du système libre et correspondent aux trajectoires constantes. Un simple calcul montre que les seuls points d'équilibre du bras sont $(0, 0)$ et $(\pi, 0)$. (Une trajectoire constante correspond à annuler le membre de droite.)

1.2 Commandabilité

Étant donnés deux états $x_0 = (\theta_0, \omega_0)$ et $x_1 = (\theta_1, \omega_1)$ dans l'espace d'état, le problème de commandabilité entre ces deux états consiste à trouver une trajectoire de (Σ) (c'est-à-dire un triplé $t \mapsto (\theta(t), \omega(t), u(t))$) partant de x_0 en $t = 0$ et arrivant en x_1 en $t = T$. En d'autres termes, il s'agit de trouver la commande $u(\cdot)$ pour amener le système d'un état à un autre. Si cela est possible, on dira que (Σ) est commandable entre x_0 et x_1 et qu'il est complètement commandable s'il est commandable pour toutes paires d'états. Il faut remarquer que la question de commandabilité peut se scinder en deux :

(Q1) étant donné (Σ), peut-on montrer qu'il est ou non complètement commandable, et ce sans explicitation des commandes ?

(Q2) Étant donnée une paire (x_0, x_1) d'états commandables par (Σ), donner une procédure effective pour déterminer une commande qui amène le système de x_0 à x_1.

La question $(Q1)$ est de nature théorique et est loin d'être résolue à l'heure actuelle. Nous l'étudierons dans le cas particulier des systèmes linéaires. On peut exprimer le problème de commandabilité en termes d'algèbre linéaire et on peut alors donner une condition nécessaire et suffisante sur la dynamique du système commandé qui caractérise la commandabilité.

La question $(Q2)$, appelée aussi *planification de trajectoires*, est encore plus difficile à résoudre que $(Q1)$. Nous proposerons, toujours pour les systèmes linéaires, une solution.

1.3 Bouclage

Pour le bras rigide, une manière assez naturelle de déterminer la commande $u(\cdot)$ qui doit réaliser notre objectif est de procéder comme suit : on sait d'où on part (du point x_0) et où on doit arriver (au point $x_r = (\theta_r, 0)$) en temps T. Supposons que, uniquement à partir de la connaissance de x_0, x_r et T, on soit maintenant capable de calculer une commande $u(\cdot)$ amenant le bras de x_0 à x_r. On dit alors que l'on commande *en boucle ouverte*. Cette façon de faire présente au moins deux défauts :

— à l'instant $t = 0$, on est censé calculer toute la loi de commande $u : t \in [0, T]$ puis, l'implémenter dans le système physique, pratiquement à l'instant $t = 0$. Cela suppose que le temps de calcul de la commande $u(\cdot)$ est négligeable par rapport à celui du système. Pour certaines applications, cela est irréaliste ;

— supposons qu'il arrive un incident sur l'intervalle $[0, T]$ qui ne soit pas pris en compte par la dynamique du système. La loi de commande ayant déja été calculée à l'avance, le système ne pourra pas réagir à l'incident imprévu. (Il faut remarquer que cet "imprévu" peut survenir constamment si le modèle qui est utilisé n'est qu'approximatif !)

C'est pour cela qu'il convient parfois de calculer $u(\cdot)$ de manière plus simple et de l'ajuster en temps réel de façon à compenser les écarts instantanés à la trajectoire de référence, $\theta - \theta_r$ et $\omega - \omega_r$ qui peuvent apparaître. Par exemple, on peut choisir, pour $t \in [0, T]$, $u(t)$ en terme de $\theta(t) - \theta_r$ et $\omega(t) - \omega_r$. L'utilisation de ce type de termes correspond à un *bouclage ou retour d'état*, appelée aussi *feedback*. Remarquons qu'une commande de ce type suppose la connaissance à tout instant $t \in [0, T]$ des quantités $\theta(t) - \theta_r$ et $\omega(t) - \omega_r$.

1.4 Stabilisation

Rappelons que notre objectif est de stabiliser le bras au point $x_r = (\theta_r, 0)$. Supposons que l'on atteigne x_r, on y reste avec la commande constante $u_r := \sin \theta_r$. De plus, il est clair que l'on peut s'approcher de θ_r en temps fini et rester dans un voisinage de cet angle avec une vitesse angulaire ω, elle aussi pas trop grande. Être capable alors de faire diminuer **en même temps** $\theta(t) - \theta_r$ et $\omega(t)$ vers zéro est un peu moins évident. Une façon de faire est alors de linéariser (Σ) au voisinage de x_r c'est-à-dire d'écrire

$$x = x_r + \delta x \text{ avec } \delta x = (\delta\theta, \delta\omega) \quad u = u_r + \delta u.$$

Ici, δx et δu sont petits. Pour écrire la dynamique de δx à partir de (1.3), on effectue un développement limité des seconds membres de (1.3) en ne retenant que les termes d'ordre 1 en $\delta x, \delta u$. On obtient alors *le linéarisé tangent* $(S)_L$ de (S) le long de la trajectoire constante x_r.

$$(S)_L \left\{ \begin{array}{rcl} \dot{\delta\theta}(t) & = & \delta\omega, \\ \dot{\delta\omega}(t) & = & \delta u(t) - cos(\theta_r)\delta\theta(t). \end{array} \right. \tag{1.6}$$

On peut aussi écrire ce système

$$(S)_L \quad \dot{\delta x} = A\delta x + b\delta u, \tag{1.7}$$

avec

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ et } b := \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

L'objectif de stabilisation consiste maintenant à amener tout point de $\mathbb{R}^2$ en $(0,0)$ le long de $(S)_L$. Remarquons que le membre de droite de $(S)_L$ est linéaire en $(\delta\theta, \delta\omega, \delta u)$. On dira alors que l'on a affaire à un système commandé linéaire stationnaire (c'est-à-dire ne dépendant pas explicitement du temps) et à coefficients constants. Comme nous le verrons, ces caractéristiques de linéarité permettent d'exprimer les trajectoires de ce système avec de manière explicite. Afin de stabiliser le système tout en conservant un caractère linéaire, il est naturel de choisir δu linéaire en δx,

$$\delta u = -k^T \delta x = -k_1 \delta\theta - k_2 \delta\omega, \tag{1.8}$$

avec $k = (k_1, k_2)^T$ vecteur constant appelé *gain du contrôleur*. La loi de feedback précédente est dénommée retour d'état ou feedback *statique*. Avec ce choix pour $u(\cdot)$, le système bouclé s'écrit

$$(Lin)_k \quad \dot{\delta x} = (A + bk^T)\delta x.$$

On pose

$$A(k) := A + bk^T.$$

Il s'agit donc de déterminer k de telle sorte que l'EDO linéaire $(Lin)_k$ soit *asymptotiquement stable* : pour toute condition initiale $\delta x_0 \in \mathbb{R}^2$, ses solutions tendent vers zéro lorsque t tend vers l'infini. Cela est équivalent au problème d'algèbre linéaire suivant : trouver un vecteur k tel que les valeurs propres de $A(k)$ soient de partie réelle strictement négative. Ces valeurs propres sont appelées les *pôles* du système bouclé $(Lin)_k$. On verra qu'il suffit de choisir $k_1, k_2 > 0$ et la convergence des solutions (vers zéro) est alors exponentielle. Par exemple, si les valeurs propres λ_1, λ_2 de $A(k)$ sont réelles, alors toute solution est combinaison linéaire de $\exp(\lambda_1 t)$ et $\exp(\lambda_2)t)$.

Enfin, on peut vouloir aussi contrôler la vitesse à laquelle on stabilise le bras autour de x_r ou encore demander à ce que'il n'y ait pas d'oscillation. Cela revient dans le premier cas à contrôler la vitesse de convergence de $(Lin)_k$ et dans le second à n'avoir que des valeurs propres réelles pour $A(k)$. On a deux instances particulières du problème plus général qui consiste à déterminer le vecteur k afin que $A(k)$ ait des valeurs propres vérifiant certaines conditions. On arrive donc à la question purement algébrique suivante : "étant donné une paire (A, B) avec A matrice 2×2 et b vecteur colonne, caractériser l'ensemble de toutes les valeurs propres de $A(k) = A + bk^T$ lorsque k est un vecteur quelconque. Le *théorème de*

placement de pôles dit essentiellement que si la paire (A, B) est commandable (c'est-à-dire que la matrice ayant pour colonnes b, Ab est inversible), alors l'ensemble précédent est $\mathbb{C}$ tout entier.

1.5 Observabilité

La loi de feedback donnée en (1.8) suppose que l'on mesure **à tout instant** $t \in [0, T]$ l'état complet du système $x = (\theta, \omega)$. En pratique, les capteurs de vitesse sont très onéreux. Il est donc souvent raisonnable de supposer que l'on ne mesure que la position et ici cela signifie que l'on ne connait de manière instantanée que θ. Les quantités mesurées constituent la *sortie* d'un système commandé. Celle-ci représente une information sur l'état, instantanée mais souvent partielle. Il est clair qu'une loi de feedback n'a de réalité pratique que si cette dernière ne s'obtient qu'à l'aide de la sortie.

Revenons au système commandé avec comme unique sortie la position angulaire θ. Si l'on essaye un feedback statique avec seulement une fonction de θ, on peut montrer qu'il est impossible de stabiliser le bras. Par exemple, si l'on prend un feedback linéaire en θ, on aboutit à une équation du type

$$\ddot{\delta\theta} + k\delta\theta = 0,$$

qui n'est pas asymptotiquement stable quelle que soit la valeur de k.

Cependant, on remarque que l'on peut obtenir $\omega(t)$ en dérivant $\theta(t)$. On dit alors que l'état x du système est *observable* à partir de la sortie θ. Plus généralement, on verra que l'état est observable à partir d'une 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 mesuré θ pour en déduire ω et ainsi construire une loi de feedback stabilisante. Cette solution fonctionne si la mesure de θ n'est pas trop bruitée. Dans le cas contraire, l'opération numérique de la dérivation est à éviter. Après linéarisation, l'idée est alors d'*estimer* l'état δx à partir de la seule connaissance de l'angle $\delta\theta$ **sans dériver** $\delta\theta$. Pour cela, il faut utiliser une autre information sur δx : il vérifie la dynamique $(S)_L$! On cherche alors à construire un état artificiel $\tilde{\delta x}$ tel que $\delta x - \tilde{\delta x}$ tende vers zéro lorsque t tend vers l'infini. Un tel $\tilde{\delta x}$ est appelé observateur asymptotique. Notons $\delta\theta = C\delta x$ avec C le vecteur ligne égal à (1 0). On choisit $\tilde{\delta x}$ comme trajectoire de

$$(S)_L \quad \dot{\tilde{\delta x}} = A\tilde{\delta x} + b\delta u - LC(\delta x - \tilde{\delta x}), \tag{1.9}$$

avec L un vecteur colonne à déterminer. Ici, on a bien sur $C(\delta x - \tilde{\delta x}) = (\delta\theta - \tilde{\delta\theta})$. Remarquons que la dynamique (1.9) est obtenue en ajoutant, à la dynamique linéarisée du bras, le terme $LC(\delta x - \tilde{\delta x})$ qui ne fait intervenir, mis à part des termes "artificiels", que la sortie $\delta\theta$. Lorsque l'on considère la dynamique de *l'erreur* $e := \delta x - \tilde{\delta x}$, on a

$$\dot{e} = (A + LC)e.$$

Posons $A(L) := A + LC$. Faire tendre e vers zéro lorsque t tend vers l'infini devient encore une fois un problème d'algèbre linéaire que l'on résout simplement. En appliquant alors la loi de feedback (1.8) obtenue pour $\tilde{\delta x}$, on obtiendra une loi de feedback qui stabilise (localement) le bras. Ce type de loi est appelée feedback dynamique puisque l'on stabilise le bras à l'aide de $\tilde{\delta x}$ qui est obtenu, à partir de la sortie $\delta\theta$, de manière dynamique (c'est-à-dire via une équation différentielle) . On remarquera aussi que l'action qui consiste à estimer δx est découplée de celle qui consiste à le stabiliser : les choix de L (estimation) et de k (stabilisation) sont indépendants l'un de l'autre. C'est le principe de séparation.

1.6 Commande optimale

Une fois la question de commandabilité comprise, on a vu que la détermination d'une loi de commande est effectuée en fonction de l'objectif de commande que l'on s'impose. Celui-ci peut être un but de stabilisation comme on l'a vu précédement. On peut aussi vouloir minimiser la commande nécessaire à la réalisation de l'objectif fixé. Par exemple, pour le bras de robot décrit par (1.1), on peut vouloir minimiser le travail de la force qui est nécessaire pour amener le système (1.1) d'un état x_0 à un autre x_1 en temps $T = 1$. Cela signifie qu'il faut minimiser $\int_0^1 |u|$ parmi toutes les lois de commandes qui amènent le système (1.1) de x_0 à x_1. De même, un autre type de minimisation est celui du temps lorsque l'amplitude de la commande est uniformément bornée : supposons que la commande u prenne ses valeurs dans $[-1, 1]$. Pour toutes paires d'états x_0 et x_1, il s'agit de minimiser le temps nécessaire pour amener le système (1.1) de x_0 à x_1.

Ainsi, lorsque l'on chercher à "optimiser" la commande, il faut se donner un critère d'optimisation qui sera appelé le coût. Le but est alors de montrer qu'il existe (ou non) une commande minimisante, appelée aussi commande optimale, et surtout de caractériser cette (ou ces) commande(s) optimale(s). Il existe plusieurs manières d'attaquer les questions de commande optimale, suivant le type de dynamiques que l'on considère. Notons d'ores et déja que ces méthodes peuvent être toutes regroupées sous un principe général, dit de Pontryagin.

1.7 Plan du cours

On va maintenant reprendre de manière plus brève les divers points évoqués dans l'exemple ci-dessus.

Le chapitre 2 est consacré aux équations différentielles ordinaires (EDO) autonomes en dimension finie. On définira les notions de champ de vecteurs, trajectoires et on présentera quelques résultats fondamentaux qui s'y rattachent : problème de Cauchy (existence et unicité de solutions avec le théorèreme de Cauchy-Lipschitz), stabilité au sens de Lyapunov pour des équilibres, etc... Le cas des EDO linéaires à coefficients constants sera soigneusement analysé.

Dans le chapitre 3, est entamée l'étude des systèmes linéaires stationnaires. Dans un premier temps, on mettra l'accent sur la forme de Brunovsky et son application à la planification de trajectoire ainsi que sur la stabilisation par placement de pôles. Nous aborderons ensuite l'observabilité comme problème dual de la commandabilité, la construction d'observateurs asymptotiques (ou de Luenberger) et la synthèse d'un bouclage dynamique de sortie (on dit aussi observateur-contrôleur). Le chapitre 4 constitue une brève introduction à la commandabilité des sytèmes non linéaire.

Les chapitres 5 et 6 font l'objet de (a) l'étude détaillée de la commande LQ avec son application au filtre de Kalman et (b) la minimisation du temps pour les systèmes linéaires. Enfin le chapitre 7 présente le principe du maximum de Pontryagin pour les systèmes affines en la commande ainsi qu'une introduction à la synthèse optimale en dimension deux.

Il faudra consulter [Rou-Bo] pour les nombreux exemples représentatifs de questions préoccupant les ingénieurs ainsi que pour les exercices qui permettent d'assimiler le cours.

Les parties écrites en petits caractères peuvent être ignorées sauf mention spéciale : il s'agit souvent de démonstrations.

Chapitre 2

Équations différentielles ordinaires et stabilité

2.1 Théorie générale des équations différentielles

Dans cette section, nous présentons la théorie générale des équations différentielles *autonomes*, qui sont de la forme

$$x'(t) = f\big(x(t)\big). \tag{2.1}$$

Cette théorie permet de modéliser et d'étudier de nombreux processus d'évolution déterministes, finis et *différentiables*.

Dans la formulation (2.1), les *données* sont :

— un ensemble ouvert $\Omega \subset \mathbb{R}^n$; x et Ω sont respectivement appelés l'état du système et *l'espace d'état* du système : à chaque instant le système est caractérisé par la donnée de x qui vit dans Ω.

— une application continue $f : \Omega \to \mathbb{R}^n$, (parfois appelée "membre de droite de l'équation différentielle").

(Les résultats que nous allons présenter restent valables quand on remplace $\mathbb{R}^n$ par n'importe quel espace vectoriel de dimension finie, par exemple $\mathbb{C}^n$, $M_n(\mathbb{R})$, ...) Une telle application $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est appelée un *champ de vecteurs* : à tout point x dans Ω, elle associe un vecteur $f(x)$ dans $\mathbb{R}^n$. En mécanique, f est aussi appelé *champ de vitesse*.

Exemple. En revenant à l'exemple du bras rigide (cf. chapitre précédent), la réponse libre correspond à l'EDO $x'(t) = f(x(t))$ avec $\Omega = S^1 \times \mathbb{R}$ et le champ de vecteur $f : S^1 \times \mathbb{R} \to \mathbb{R}^2$ défini par $f(x) = (w, -\sin\theta))^T$ si $x = (\theta, w)$.

Une *solution* de l'équation différentielle est une fonction dérivable $x(\cdot) : I \to \mathbb{R}^n$ telle que :

— I est un intervalle de $\mathbb{R}$;

— $x(\cdot)$ prend ses valeurs dans Ω, *i.e.* $x(I) \subset \Omega$;

— pour tout $t \in I$, $x'(t) = f\big(x(t)\big)$.

Une solution est donc en fait un couple $(x(\cdot), I)$: l'intervalle de définition I fait partie des inconnues. Nous verrons comment caractériser cet intervalle dans la section 2.1.2.

Notons enfin que, comme l'application f est supposée continue, toute solution $x(\cdot)$ de l'équation différentielle est automatiquement de classe C^1.

Remarque. Il peut sembler réducteur de ne considérer que les équations différentielles autonomes, alors que le cadre le plus général est celui des équations de la forme

$$x'(t) = f\big(t, x(t)\big), \qquad t \in J \subset \mathbb{R}, \tag{2.2}$$

qui dépendent explicitement du temps, et qui sont dites *non-autonomes*. Ce n'est en fait pas vraiment une restriction : toute équation non-autonome dans $\mathbb{R}^n$ peut être vue comme une équation autonome dans $\mathbb{R}^{n+1}$. En effet, définissons un champ de vecteur $F : J \times \Omega \subset \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ par $F(t,x) = (1, f(t,x))$. Il est alors clair que l'équation non-autonome (2.2) est équivalente à l'équation autonome

$$\begin{pmatrix} t \\ x \end{pmatrix}' = \begin{pmatrix} 1 \\ f\big(t,x\big) \end{pmatrix} = F\big(t, x(t)\big).$$

2.1.1 Existence et unicité

Toute équation différentielle n'a pas de solution. Pour s'en convaincre, considérons l'EDO définie sur $\mathbb{R}$ par

$$\left\{ \begin{array}{rcl} x'(t) & = & -sign\big(x(t)\big), \\ x(t_0) & = & 0, \end{array} \right.$$

avec $sign(x) = x/|x|$ si x est non nul et $sign(0) = 0$. Le lecteur essaiera de donner un argument (simple) montrant qu'il n'y a pas de solution pour l'EDO précédente sur n'importe quel voisinage de 0. A la lumière de cet exemple, il est nécessaire d'effectuer une hypothèse de régularité sur le membre de droite d'une EDO pour espérer avoir un "bout" de solution dans un voisinage ouvert du temps de départ.

Définition 2.1. On appelle problème de Cauchy, le système

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t)\big), \\ x(t_0) & = & x_0, \end{array} \right. \tag{2.3}$$

c'est-à-dire le système formé d'une EDO et d'une condition initiale (valeur de l'état donnée à un instant fixé). Rappelons que l'un des objectifs des EDO est de modéliser des processus physiques qui sont souvent déterministes : si on connait la dynamique d'un système et une condition initiale à $t = t_0$, alors l'évolution de ce système est unique pour $t \geq t_0$. Cette notion de déterminisme se traduit en termes mathématiques par le fait que tout problème de Cauchy a une solution et une seule pour $t \geq 0$. Avoir unicité des solutions d'une EDO est donc une nécessité pour un modèle réaliste.

Le propos du théorème ci-dessous est de répondre aux questions précédentes.

Théorème 2.1 (Théorème de Cauchy-Lipschitz). *Supposons f de classe C^1 sur Ω. Alors, pour tout point $x_0 \in \Omega$ et tout $t_0 \in \mathbb{R}$, il existe $\delta > 0$ tel que le problème de Cauchy défini en (2.3) possède une unique solution définie sur $]t_0 - \delta, t_0 + \delta[$.*

*Preuve.

▷ La démonstration de ce théorème repose sur le théorème du point fixe de Picard. Fixons un réel $\alpha > 0$ tel que la boule fermée $\overline{B}(x_0, \alpha)$ soit contenue dans Ω. Puisque f est C^1, il existe des constantes M et $K > 0$ telles que, sur $\overline{B}(x_0, \alpha)$, f est bornée en norme par M et est K-lipschitzienne (pourquoi ?) Posons en outre

$$\delta = \min(\frac{\alpha}{M}, \frac{1}{2K}).$$

▷ Définissons $\mathcal{E}$ comme étant l'ensemble des fonctions $x(\cdot)$ continues sur $]t_0 - \delta, t_0 + \delta[$ à valeurs dans $\overline{B}(x_0, \alpha)$ et telles que $x(t_0) = x_0$. Muni de la norme de la convergence uniforme $\| \cdot \|_0$, c'est un espace complet. L'application

$$\Phi\big(x(\cdot)\big) = x_0 + \int_{t_0}^{\cdot} f\big(x(s)\big) ds,$$

est une application de $\mathcal{E}$ dans lui même : en effet, pour $|t - t_0| \leq \delta$,

$$\|\Phi\big(x(t)\big) - x_0\| = \| \int_{t_0}^{t} f\big(x(s)\big) ds\| \leq \delta M \leq \alpha.$$

Cette application est en outre $\frac{1}{2}$-lipschitzienne puisque, pour $t \in]t_0 - \delta, t_0 + \delta[$,

$$\begin{aligned} \|\Phi\big(x(\cdot)\big) - \Phi\big(y(\cdot)\big)\|_0 &\leq \sup_{|t-t_0|<\delta} \Big(\int_{t_0}^{t} \|f\big(x(s)\big) - f\big(y(s)\big)\| \, ds \Big) \\ &\leq \sup_{|t-t_0|<\delta} \Big(\int_{t_0}^{t} K \|x(s) - y(s)\| \, ds \Big) \\ &\leq \delta K \|x(\cdot) - y(\cdot)\|_0 \leq \frac{1}{2} \|x(\cdot) - y(\cdot)\|_0. \end{aligned}$$

Le théorème du point fixe de Picard s'applique et montre que l'application Φ admet un unique point fixe dans $\mathcal{E}$, c'est-à-dire que le système (2.3) admet une unique solution $x(\cdot) :]t_0 - \delta, t_0 + \delta[\to \mathbb{R}^n$ à valeurs dans la boule $\overline{B}(x_0, \alpha)$.

▷ Il ne reste plus qu'à montrer que toute solution $x(\cdot) :]t_0 - \delta, t_0 + \delta[\to \mathbb{R}^n$ du système (2.3) est à valeurs dans la boule $\overline{B}(x_0, \alpha)$. Par l'absurde, supposons qu'une solution $x(\cdot)$ de (2.3) sorte de $\overline{B}(x_0, \alpha)$ en temps inférieur à δ, et notons t_1 le premier instant où $x(\cdot)$ sort de la boule ouverte $B(x_0, \alpha)$. D'après le théorème des accroissements finis,

$$\alpha = \|x(t_1) - x_0\| \leq \big(\sup_{t \in [t_0, t_1]} \|x'(t)\| \big) \, |t_1 - t_0| < M\delta,$$

ce qui contredit $\delta \leq \alpha / M$. Toute solution de (2.3) sur $]t_0 - \delta, t_0 + \delta[$ est donc à valeurs dans $\overline{B}(x_0, \alpha)$, ce qui montre le théorème.

□

Hypothèses plus faibles sur f

Nous avons énoncé le théorème de Cauchy-Lipschitz avec l'hypothèse que f est C^1 sur Ω car elle est simple à utiliser et fréquemment satisfaite dans les applications. Remarquons cependant que, dans la preuve, nous avons seulement besoin que f soit *localement lipschitzienne*, c'est-à-dire que pour tout $x_0 \in \Omega$, il existe un voisinage U_0 de x_0 dans Ω et une constante K tels que f soit K-lipschitzienne sur U_0. La conclusion du théorème de Cauchy-Lipschitz reste donc valable sous l'hypothèse que f est localement lipschitzienne. En particulier, elle est valable si f est (globalement) lipschitzienne sur Ω.

Que se passe-t-il si on affaiblit encore les hypothèses et que l'on suppose f seulement continue ? Un théorème de Peano affirme que, dans ce cas, le système (2.3) admet toujours une solution. En revanche, l'unicité n'est pas garantie. Par exemple le problème de Cauchy

$$\left\{ \begin{array}{rcl} y'(t) & = & \sqrt{|y(t)|} \\ y(0) & = & 0 \end{array} \right. , \qquad y \in \mathbb{R}$$

admet comme solutions les fonctions $y_1(t) = 0$ et $y_2(t) = \frac{t|t|}{4}$. Il en admet même une infinité puisque, pour tout $a \geq 0$, la fonction $y^a(\cdot)$ définie par

$$y^a(t) = 0 \quad \text{pour } t \leq a, \qquad y^a(t) = y_2(t-a) \quad \text{pour } t > a$$

est également solution.

Remarque. D'après la remarque faite en introduction, le théorème de Cauchy-Lipschitz est également valable pour une équation non-autonome $x' = f(t,x)$: si f est C^1 sur $J \times \Omega$, alors, pour $(t_0, x_0) \in J \times \Omega$, l'équation a une unique solution définie sur $]t_0 - \delta, t_0 + \delta[$ et valant x_0 en t_0.

On peut dans ce cas affaiblir nettement les hypothèses sur f : en effet, la conclusion du théorème restera valable si on suppose seulement que f est *localement lipschitzienne en la seconde variable* x, c'est-à-dire que, pour tout $(t_0, x_0) \in J \times \Omega$, il existe un voisinage J_0 de t_0 dans J, un voisinage U_0 de x_0 dans Ω et une constante K tels que, pour tout $t \in J_0$, l'application $f(t,\cdot)$ est K-lipschitzienne sur U_0. La preuve est une simple adaptation de celle que nous avons donnée ici.

2.1.2 Solutions maximales et durée de vie

Considérons l'équation différentielle

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

où le champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est supposé de classe C^1.

Nous avons défini au début de ce chapitre une solution de cette équation comme une fonction $x(\cdot)$ définie sur un certain intervalle I de $\mathbb{R}$. Cette section est consacrée à l'étude de cet intervalle de définition I. Nous rencontrerons deux types de problèmes.

— Étant donné un couple $(t_0, x_0) \in \mathbb{R} \times \Omega$, il existe une infinité de solutions de (2.4) satisfaisant la condition initiale $x(t_0) = x_0$: par exemple si $x(\cdot)$ est solution sur l'intervalle I, $t_0 \in I$, toute restriction de $x(\cdot)$ à un sous-intervalle de I contenant t_0 est une solution différente. Pour éviter de considérer comme solutions différentes la même fonction prise sur des sous-intervalles, nous chercherons à associer à une fonction un unique intervalle, le plus grand, sur lequel elle est solution : c'est la notion de *solution maximale*.

— Si on choisit l'intervalle I le plus grand possible, peut-on le prendre égal à $\mathbb{R}$ tout entier ? Si ce n'est pas le cas, que se passe-t-il pour la solution ? et quelle est la forme de I ? C'est le problème de la *durée de vie* des solutions.

Solutions maximales

Définition 2.2. On dit qu'une solution $x(\cdot) : I \to \Omega$ de (2.4) est une *solution maximale* si elle n'a pas de prolongement à un intervalle strictement plus grand, c'est-à-dire si elle n'est pas la restriction à I d'une solution définie sur un intervalle $I' \supsetneq I$.

Nous allons montrer qu'il existe une unique solution maximale de l'équation (2.4) satisfaisant une condition initiale donnée. Nous avons besoin pour cela d'un résultat d'unicité globale.

Proposition 2.2. *Si $x(\cdot)$ et $y(\cdot) : I \to \Omega$ sont deux solutions de (2.4) définies sur le même intervalle I qui coïncident en un point $t_0 \in I$, alors elles sont égales.*

Preuve.

▷ Considérons d'abord le dernier instant supérieurà t_0 pour lequel les solutions coïncident :

$$t_+ = \sup\{t \in I \ : \ t > t_0, \ x(s) = y(s) \text{ pour tout } s \in [t_0, t]\}.$$

Par l'absurde, supposons $t_+ < \sup I$. Les solutions étant continues, on a $x(t_+) = y(t_+)$. En appliquant le théorème de Cauchy-Lipschitz au couple $(t_+, x(t_+))$, on obtient que les deux solutions sont encore égales sur un intervalle $[t_+, t_+ + \delta[$, ce qui contredit la définition même de t_+. Donc $t_+ = \sup I$. Le même argument vaut pour l'infimum des instants où les deux solutions coïncident.

□

Théorème 2.3. *Pour toute donnée initiale $(t_0, x_0) \in \mathbb{R} \times \Omega$, il existe une unique solution maximale $x(\cdot) :]t_-, t_+[\to \Omega$ de (2.4) satisfaisant $x(t_0) = x_0$. Tout autre solution satisfaisant cette condition initiale est une restriction de $x(\cdot)$ à un sous-intervalle de $]t_-, t_+[$.*

Remarque. Insistons sur le fait que l'intervalle de définition d'une solution maximale est toujours un intervalle ouvert $]t_-, t_+[$. Les bornes t_+ et t_- de l'intervalle maximal sont des fonctions de (t_0, x_0) qui prennent leurs valeurs dans $\overline{\mathbb{R}}$: t_+ peut être soit un réel, soit $+\infty$, alors que t_- peut être soit un réel soit $-\infty$. Dans tous les cas, $t_- < t_0 < t_+$.

Preuve.

▷ Soit I la réunion de tous les intervalles contenant t_0 sur lesquels le système

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t)\big), \\ x(t_0) & = & x_0, \end{array} \right. \tag{2.3}$$

admet une solution. D'après le théorème de Cauchy-Lipschitz, cette réunion est un intervalle ouvert, c'est-à-dire de la forme $I =]t_-, t_+[$. Pour tout $t \in]t_-, t_+[$, définissons $x(t)$ comme la valeur en t de n'importe quelle solution de (2.3) définie sur $[t_0, t]$. La proposition précédente montre que la fonction $x(\cdot) :]t_-, t_+[\to \Omega$ ainsi définie est bien solution de (2.3). De plus, par construction, c'est un prolongement de toute autre solution.

□

Durée de vie

On s'intéresse maintenant à l'intervalle de définition $]t_-, t_+[$ d'une solution maximale $x(\cdot)$ de (2.4). Cet intervalle peut être différent de $\mathbb{R}$, même pour les équations les plus simples.

Exemple. Considérons l'équation $y'(t) = y^2(t)$ dans $\mathbb{R}$, dont la solution valant y_0 en t_0 est

$$y(t) = \frac{y_0}{(t - t_0)y_0 + 1}.$$

L'intervalle maximal de définition de cette solution est $]t_0 - \frac{1}{y_0}, +\infty[$ si $y_0 > 0$, $]-\infty, t_0 - \frac{1}{y_0}[$ si $y_0 < 0$, et $\mathbb{R}$ tout entier si $y_0 = 0$.

L'idée générale est que, si une solution ne peut être prolongée sur tout $\mathbb{R}$, c'est qu'elle s'approche en temps fini du bord de l'ensemble Ω. Formalisons cette idée pour la borne supérieure t_+ de l'intervalle (les résultats pour t_- sont similaires).

Proposition 2.4. *Soit $x(\cdot) :]t_-, t_+[\to \Omega$ une solution maximale de (2.4). Alors, si $t_+ < +\infty$, $x(t)$ sort définitivement de tout compact contenu dans Ω quand $t \to t_+$.*

*Preuve.

▷ À faire.

□

On rencontrera le cas $t_+ < +\infty$ essentiellement dans les deux situations suivantes :

— quand $\Omega = \mathbb{R}^n$ et $\lim_{t\to t_+} \|x(t)\| = +\infty$: c'est le phénomène *d'explosion en temps fini* dont nous avons donné un exemple ci-dessus ;
— quand le bord de Ω est borné et $x(t)$ converge vers un point du bord quand $t \to t_+$.

Inversement, retenons une condition suffisante pour que $]t_-, t_+[= \mathbb{R}$.

Corollaire 2.5. *Si toutes les valeurs d'une solution maximale $x(\cdot)$ sont contenues dans un compact inclus dans Ω, alors $x(\cdot)$ est définie sur tout $\mathbb{R}$.*

Champs de vecteurs complets

Définition 2.3. On dit que le champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est *complet* (ou que l'équation associée est *complète*) si toute solution maximale est définie sur $\mathbb{R}$ tout entier.

D'après le corollaire précédent, si toutes les solutions maximales sont contenues dans des compacts, le champ est complet.

Exemple. Si $\Omega = \mathbb{R}^n$, les champs linéaires $f(x) = Ax$ sont complets et, de façon plus générale, tous les champs admettant une majoration linéaire $\|f(x)\| \leq \alpha\|x\| + \beta$, avec $\alpha, \beta \geq 0$, sont complets (donc en particulier les champs bornés). C'est une conséquence du lemme de Gronwall.

2.1.3 Flots, portraits de phase

Considérons à nouveau une équation différentielle autonome

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

où le champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est supposé de classe C^1. Une des spécificités de cette équation est qu'elle ne dépend pas explicitement du temps (d'où le qualificatif autonome). En particulier, les solutions sont invariantes par translation du temps : si $x(\cdot)$ est solution, $x(t_0 + \cdot)$ aussi.

Proposition 2.6. *Soit $x(\cdot) :]t_-, t_+[\to \Omega$ une solution maximale de (2.4) et $t_0 \in \mathbb{R}$. Alors $\bar{x} : t \mapsto x(t + t_0)$, définie sur $]t_- - t_0, t_+ - t_0[$, est également une solution maximale de (2.4).*

Ainsi le temps n'a pas de rôle intrinsèque ici et on pourra se limiter aux données initiales en $t = 0$. Pour un point $x \in \Omega$, notons $\phi(\cdot, x)$ la solution maximale de (2.4) valant x en $t = 0$ et $I_x =]t_-, t_+[$ son intervalle de définition. Autrement dit, $\phi(\cdot, x)$ est la solution du système

$$\left\{ \begin{array}{rcl} \frac{\partial}{\partial t}\phi(t,x) & = & f\big(\phi(t,x)\big), \\ \phi(0,x) & = & x, \end{array} \right. \qquad \forall t \in I_x.$$

Définition 2.4. L'application $(t, x) \mapsto \phi(t, x)$ est appelée le *flot* du champ de vecteurs f (ou de l'équation $x' = f(x)$).

Par définition, l'application partielle à x fixé, $t \mapsto \phi(t, x)$, est une solution maximale de l'équation. Pour une étude qualitative de l'équation différentielle, il est important d'étudier plutôt l'autre application partielle, $\phi_t : x \mapsto \phi(t, x)$, pour t fixé. De façon imagée, $\phi_t(x)$ est la position à l'instant t d'un corps transporté par l'équation différentielle qui se trouvait à la position x en $t = 0$.

Exemple. Si f est linéaire, *i.e.* $f(x) = Ax$, $A \in M_n(\mathbb{R})$, le flot est donné par l'exponentielle de A :

$$\phi_t(x) = e^{tA}x, \quad \forall (t, x) \in \mathbb{R} \times \mathbb{R}^n.$$

Ainsi le flot est une généralisation de l'exponentielle de matrice. Il possède des propriétés similaires.

Proposition 2.7 (formule du flot)**.** *Si $t_1 \in I_x$ et $t_2 \in I_{\phi_{t_1}(x)}$, alors $t_1 + t_2 \in I_x$ et*

$$\phi_{t_1+t_2}(x) = \phi_{t_2}\big(\phi_{t_1}(x)\big).$$

En particulier, si $t \in I_x$,

$$\phi_{-t}\big(\phi_t(x)\big) = x$$

Preuve.

▷ D'après la proposition sur l'invariance par translation du temps, $t \mapsto \phi_{t_1+t}(x)$ est la solution maximale valant $\phi_{t_1}(x)$ en $t = 0$, ce qui est la définition de $t \mapsto \phi_t\big(\phi_{t_1}(x)\big)$. □

Remarque. La formule du flot peut aussi se lire de la façon suivante : si $x(\cdot)$ est une solution de (2.4), alors

$$x(t) = \phi_{t-t_0}\big(x(t_0)\big)$$

pour tous t_0 et t dans l'intervalle de définition de $x(\cdot)$.

Le domaine de définition du flot est l'ensemble

$$\mathcal{D} = \{(t, x) \in \mathbb{R} \times \Omega \ : \ t \in I_x\}.$$

Pour obtenir des propriétés intéressantes sur le flot (continuité, différentiabilité), il est nécessaire de montrer d'abord que son domaine de définition $\mathcal{D}$ est un ouvert de $\mathbb{R} \times \Omega$. Nous le verrons dans la section suivante.

Il y a cependant déjà un cas où cela est évident : si f est un champ de vecteurs complet sur Ω, c'est-à-dire si $I_x = \mathbb{R}$ pour tout $x \in \Omega$, le domaine de définition du flot est $\mathcal{D} = \mathbb{R} \times \Omega$. On peut alors réécrire les propriétés du flot de façon globale : pour tous $t, s \in \mathbb{R}$,

1. $\phi_t \circ \phi_s = \phi_{t+s}$;
2. $\phi_{-t} \circ \phi_t = \mathrm{id}$;
3. $\phi_0 = \mathrm{id}$;
4. $\frac{\partial}{\partial t}\phi_t = f \circ \phi_t$.

Les trois premières propriétés montrent en particulier que ϕ_t obéit à une loi de groupe.

Orbites et portraits de phase

Définition 2.5. On appelle *orbite* d'un point $x_0 \in \Omega$ (ou trajectoire passant par x_0) l'ensemble

$$\mathcal{O}_{x_0} = \{\phi_t(x_0) \ : \ t \in I_{x_0}\}.$$

Autrement dit, l'orbite de x_0 est la courbe tracée sur $\mathbb{R}^n$ par la solution maximale de l'équation (2.4) passant par x_0 en $t = 0$.

La propriété d'invariance par translation du temps implique que, pour tout point $x \in \mathcal{O}_{x_0}$, on a $\mathcal{O}_x = \mathcal{O}_{x_0}$. En effet, dans ce cas, il existe un instant t_0 tel que $x = \phi_{t_0}(x_0)$. Tout point y de $\mathcal{O}_x$ s'écrit alors $y = \phi_t(x) = \phi_{t+t_0}(x_0)$, c'est-à-dire $y \in \mathcal{O}_{x_0}$. En particulier, ceci implique que *deux orbites distinctes ne peuvent pas se croiser*. Chaque point de Ω appartient donc à une et une seule orbite.

La partition de Ω en orbites s'appelle le *portrait de phase* du champ de vecteurs. On y trouve trois sortes d'orbites :

— des points, *i.e.* $\mathcal{O}_{x_0} = \{x_0\}$: un tel point vérifie nécessairement $f(x_0) = 0$. C'est ce que l'on appelle un *point d'équilibre* (voir Définition 2.12). Remarquer qu'un point d'équilibre correspond à un point fixe de ϕ_t pour tout t : $\phi_t(x_0) = x_0$.

— des courbes fermées : il existe alors un point x dans l'orbite et un temps $T > 0$ tels que $\phi_T(x) = x$. Ceci implique $\phi_{t+T}(x) = \phi_t(x)$ pour tout $t \in \mathbb{R}$, c'est-à-dire que la solution maximale $\phi(\cdot, x)$ est T-périodique. On parlera dans ce cas d'*orbite périodique*.

— des courbes ouvertes : il n'y a alors aucun point double, *i.e.* si $t \neq s$, $\phi_t(x) \neq \phi_s(x)$.

On porte habituellement sur le dessin d'un portrait de phase le sens de parcours des orbites.

Exemple. Considérons le champ de vecteurs linéaire $f(x) = Ax$ dans $\mathbb{R}^2$, et supposons que la matrice $A \in M_2(\mathbb{R})$ a deux valeurs propres réelles et distinctes $\lambda_1 < \lambda_2$. L'étude réalisée dans la section 2.2.4 permet de déterminer la forme du portrait de phase en fonction de λ_1 et λ_2. Nous avons représenté les différentes possibilités dans la figure 2.1, où nous avons noté E_1 et E_2 les sous-espaces propres associés à λ_1 et λ_2.

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

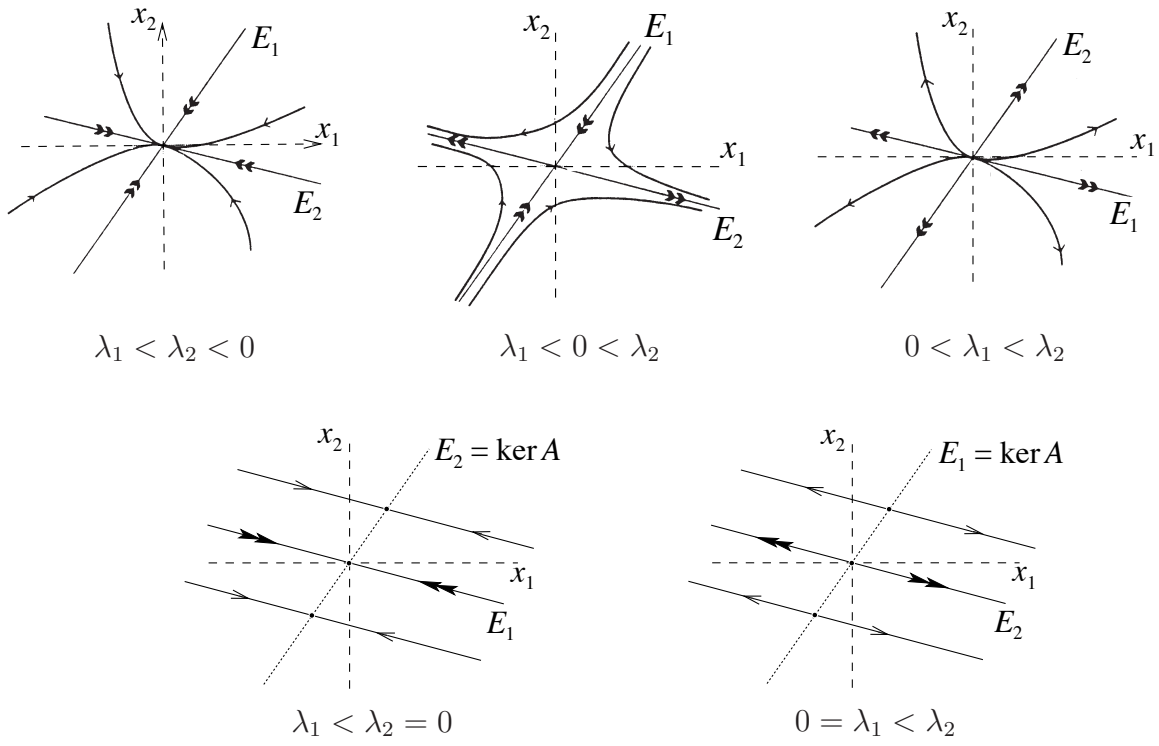


FIGURE 2.1 – Exemples de portraits de phase pour $f(x) = Ax$ dans $\mathbb{R}^2$.

Définition 2.6. Soit A un sous-ensemble de l'espace d'état Ω. On dit que A est *invariant* (respectivement *positivement invariant*) par le flot ϕ_t si, pour tout $t \in \mathbb{R}$ (respectivement dans $\mathbb{R}_+$) , $\phi_t(A)$ 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 2.7. On appelle *intégrale première* d'une EDO $\dot{x} = f(x)$, une fonction $h : \Omega \to \mathbb{R}$, de classe C^1, qui reste constante le long des trajectoires de l'EDO. Cela est vrai en particulier si, pour tout $x \in \Omega$ et t, $\frac{d}{dt}\big(h(\phi_t(x)\big) = 0$, condition qui est équivalente à

$$D_x h(x).f(x) = 0, \quad \text{pour tout } x \in \Omega.$$

(Noter que cette dernière condition ne demande pas une connaissance explicite du flot.) Ainsi, les hypersurfaces de niveau $H_c := \{x \in \Omega : \ \ h(x) = c\}$, $c \in \mathbb{R}$ sont invariantes par le flot.

Exemple. Montrer que (1.5) admet comme intégrale première $h := 1/2\omega^2 - \cos\theta$. En déduire l'équation des orbite. Dessiner l'allure du portrait de phases sur le cylindre $S^1 \times \mathbb{R}$ suivant la valeur (constante) de h. Afin d'avoir une représentation plane, on identifiera S^1 avec $\mathbb{R}/[0, 2\pi]$ et on montrera l'existence de trajectoires périodiques.

Pour d'autres exemples, voir [Rou-Bo] où est expliqué le fait que l'énergie totale (énergie cinétique plus énegrgie potentielle) d'un système mécanique holonome parfait (sans frottement) est une intégrale première du système.

2.1.4 Équations différentielles linéaires

On suppose que le membre de droite de (2.1) est linéaire par rapport à l'état x c'est-à-dire qu'il prend la forme

$$x'(t) = A(t)x(t), \qquad t \in I. \tag{2.5}$$

Précisons les notations. Les *données* sont :
- un intervalle I de $\mathbb{R}$;
- une application $A : I \to M_n(\mathbb{K})$ de classe C^k (k est un entier positif ou $k = \infty$) ; chaque valeur $A(t)$ est donc une matrice $(n \times n)$ à coefficients dans le corps $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$.

Une *solution* de (2.5) est une application dérivable $x : I \to \mathbb{K}^n$ telle que, pour tout $t \in I$, sa dérivée $x'(t) = \frac{dx}{dt}(t)$ vérifie $x'(t) = A(t)x(t)$. Noter qu'une solution est automatiquement de classe C^{k+1}.

Nous traiterons aussi le cas un peu plus général des équations différentielles *affines*,

$$x'(t) = A(t)x(t) + b(t), \qquad t \in I, \tag{2.6}$$

la donnée $b(\cdot)$ étant une application de I dans $\mathbb{K}^n$ de classe C^k. Nous verrons que l'étude de ces équations se déduit de celle des équations linéaires.

Remarque. Il est fréquent dans la littérature que l'expression « équation linéaire » soit utilisé pour les équations affines, les équations (2.5) étant alors appelées équations linéaires *homogènes*.

Existence et Unicité globales

Théorème 2.8 (Existence et Unicité globales)**.** *Soient $t_0 \in I$ et $x_0 \in \mathbb{K}^n$. Il existe une unique solution $x(\cdot)$ de l'équation (2.6) satisfaisant à la condition initiale*

$$x(t_0) = x_0.$$

Insistons sur le fait que ce théorème garantit l'existence de $x(\cdot)$ sur *tout* l'intervalle I. Ce phénomène est propre aux équations linéaires (on comparera avec le théorème 2.1 de Cauchy-Lipschitz concernant les équations non-linéaires).

La preuve de ce théorème repose sur la remarque suivante : $x(\cdot)$ est solution de (2.6) avec $x(t_0) = x_0$ si et seulement si $x(\cdot)$ est continue et vérifie pour tout $t \in I$,

$$x(t) = x_0 + \int_{t_0}^{t} \big(A(s)x(s) + b(s)\big)\, ds, \tag{2.7}$$

autrement dit, si $x(\cdot)$ est un point fixe de l'application

$$x(\cdot) \mapsto x_0 + \int_{t_0}^{\cdot} \big(A(s)x(s) + b(s)\big)\, ds.$$

Ainsi le théorème 2.8 est un résultat de point fixe et nous utiliserons donc le théorème du point fixe de Picard pour le montrer.

*Preuve.

▷ Supposons pour simplifier que I soit un intervalle compact de la forme $I = [a, b]$ et introduisons Φ l'application affine qui à toute fonction $x(\cdot) \in C^0(I, \mathbb{R}^n)$ associe la fonction

$$\Phi\big(x(\cdot)\big) = x_0 + \int_{t_0}^{\cdot} \big(A(s)x(s) + b(s)\big) ds,$$

qui est visiblement continue et à valeurs dans $\mathbb{R}^n$. Comme nous l'avons remarqué ci-dessus, il s'agit de montrer que Φ a un unique point fixe. Nous allons pour cela vérifier qu'un *itéré* de Φ est contractant et appliquer le théorème du point fixe de Picard à cet itéré. On vérifiera alors que les points fixes de Φ sont ceux de cet itéré.

▷ Pour $x(\cdot), y(\cdot) \in C^0(I, \mathbb{R}^n)$ et $t \in I$, on a

$$\begin{aligned}\|(\Phi \circ x)(t) - (\Phi \circ y)(t)\| &= \| \int_{t_0}^{t} (A(s)(x(s) - y(s)) ds\| \\ &\leq \|A\|_{C^0} \int_{t_0}^{t} \|(x - y)(s)\| ds \leq |t - t_0|\, \|A\|_{C^0}\, \|x - y\|_{C^0},\end{aligned}$$

où $\|f\|_{C^0} = \max_{t\in I} \|f(t)\|$ désigne la norme C^0. On voit en particulier que

$$\|\Phi \circ x - \Phi \circ y\|_{C^0} \leq (b - a)\, \|A\|_{C^0}\, \|x - y\|_{C^0},$$

ce qui prouve que Φ est continue de $C^0(I, \mathbb{R}^n)$ dans lui-même.

▷ Par le même calcul, on a, en notant $\Phi^2 = \Phi \circ \Phi$,

$$\begin{aligned}\|(\Phi^2 \circ x)(t) - (\Phi^2 \circ y)(t)\| &\leq \|A\|_{C^0} \int_{t_0}^{t} \|(\Phi \circ x - \Phi \circ y)(s)\| ds \\ &\leq \|A\|_{C^0} \|A\|_{C^0} \|x - y\|_{C^0} \int_{t_0}^{t} |s - t_0| ds \leq \|A\|_{C^0}^2\, \frac{1}{2}\, |t - t_0|^2\, \|x - y\|_{C^0},\end{aligned}$$

et on montre par récurrence que

$$\|(\Phi^N \circ x)(t) - (\Phi^N \circ y)(t)\| \leq \|A\|_{C^0}^N\, \frac{|t - t_0|^N}{N!}\, \|x - y\|_{C^0},$$

c'est-à-dire,

$$\|\Phi^N \circ x - \Phi^N \circ y\|_{C^0} \leq \frac{\big((b - a)\, \|A\|_{C^0}\big)^N}{N!}\, \|x - y\|_{C^0}.$$

▷ Choisissons N suffisamment grand pour que $\frac{((b-a)\|A\|_{C^0})^N}{N!} \leq \frac{1}{2}$ (ce qui est possible puisque le membre de gauche de cette inégalité tend vers 0 avec N). L'application continue $\Phi^N : C^0(I,\mathbb{R}^n) \to C^0(I,\mathbb{R}^n)$ est alors $\frac{1}{2}$-contractante et, d'après le théorème du point fixe de Picard, admet un unique point fixe $x(\cdot) \in C^0(I,\mathbb{R}^n)$.
Mais il n'est pas difficile de voir que les points fixes de Φ sont les point fixes de Φ^N : en effet si $x(\cdot)$ est point fixe de Φ on a évidemment

$$\Phi^N\big(x(\cdot)\big) = \Phi^{N-1} \circ \Phi\big(x(\cdot)\big) = \Phi^{N-1}\big(x(\cdot)\big),$$

et donc $\Phi^N\big(x(\cdot)\big) = x(\cdot)$; réciproquement, si $\Phi^N\big(x(\cdot)\big) = x(\cdot)$, alors

$$\Phi^N\big(\Phi\big(x(\cdot)\big)\big) = \Phi^{N+1}\big(x(\cdot)\big) = \Phi\big(\Phi^N\big(x(\cdot)\big)\big) = \Phi\big(x(\cdot)\big),$$

et donc $\Phi\big(x(\cdot)\big)$ et $x(\cdot)$ sont points fixes de Φ^N qui est *contractante* et admet de ce fait un *unique* point fixe : ils sont donc égaux. Ceci prouve le théorème pour I de la forme $I = [a,b]$.
□

La résolvante

Revenons à l'étude des équations linéaires dans $\mathbb{K}^n$

$$x'(t) = A(t)x(t), \tag{2.5}$$

et notons $\mathcal{E}$ l'ensemble des solutions de cette équation.

Proposition 2.9. *L'ensemble $\mathcal{E}$ est un espace vectoriel de dimension n.*

Preuve.

▷ Il est immédiat que $\mathcal{E}$ est un $\mathbb{K}$-espace vectoriel. Introduisons alors $L_{t_0} : \mathbb{K}^n \to \mathcal{E}$ l'application qui à $x_0 \in \mathbb{K}^n$ associe la solution $x(\cdot)$ de (2.5) telle que $x(t_0) = x_0$. C'est clairement une application linéaire et il résulte directement de l'existence et de l'unicité des solutions que L_{t_0} est un isomorphisme de $\mathbb{K}^n$ sur $\mathcal{E}$, ce qui prouve le résultat.
□

Définition 2.8. On appelle *résolvante* de l'équation (2.5) l'application $R_A(t,s) : \mathbb{K}^n \to \mathbb{K}^n$ qui à $x_0 \in \mathbb{K}^n$ associe $x(t)$, où $x(\cdot)$ est la solution de (2.5) qui satisfait $x(s) = x_0$.

Il résulte des théorèmes d'existence et d'unicité que la résolvante est linéaire et bijective ; $R_A(t,s)$ est donc une matrice *inversible* de $M_n(\mathbb{K})$. Elle permet d'exprimer toute solution $x(\cdot)$ de l'équation (2.5) en fonction d'une condition initiale :

$$x(t) = R_A(t,t_0)x(t_0).$$

En particulier, dans le cas autonome, c'est-à-dire quand $A(\cdot) \equiv A$ est constante, la résolvante est l'exponentielle de A : $R_A(t,s) = e^{(t-s)A}$.

La résolvante peut également être caractérisée par une équation différentielle.

Proposition 2.10.

1. *Pour tout $t_0 \in I$, $R_A(\cdot, t_0)$ est la solution de l'équation différentielle* matricielle
$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial t} R_A(t, t_0) & = & A(t) R_A(t, t_0), \\ R_A(t_0, t_0) & = & I, \end{array} \right. \tag{2.9}$$
2. *Pour tous t_0, t_1, t_2 dans I,*
$$R_A(t_2, t_0) = R_A(t_2, t_1) \times R_A(t_1, t_0).$$
3. *Si $A(\cdot)$ est de classe C^k, l'application $t \mapsto R_A(t, t_0)$ est de classe C^{k+1}.*

Preuve.

▷ Le premier point résulte du fait que, *pour tout* $x_0 \in \mathbb{K}^n$,
$$\frac{\partial}{\partial t}(R_A(t, t_0)x_0) = A(t)(R_A(t, t_0)x_0).$$
Le deuxième point résulte de la définition même de la résolvante. Quant à la dernière assertion, elle résulte de la première.

□

Remarques.

— L'équation différentielle satisfaite par la résolvante a la même forme que l'équation (2.5) mais a lieu dans $M_n(\mathbb{K})$ (au lieu de $\mathbb{K}^n$).

— Le point *2* (ou directement la définition) implique en particulier :
$$R_A(t, s)^{-1} = R_A(s, t).$$

— Les colonnes $x_1(t), \ldots, x_n(t)$ de la résolvante $R_A(t, t_0)$ sont les valeurs à l'instant t des solutions valant $e_1, \ldots, e_n$ (les vecteurs de la base canonique) à $t = t_0$, puisque par définition $x_j(t) = R_A(t, t_0)e_j$. Les fonctions $x_1(\cdot), \ldots, x_n(\cdot)$ forment donc une base de l'ensemble $\mathcal{E}$ des solutions.

Comme nous venons de le voir, pour résoudre une équation différentielle linéaire, il suffit de savoir calculer la résolvante (de même que pour les équations autonomes il suffit de savoir calculer l'exponentielle). Malheureusement, en dehors du cas autonome, **il est très rare de pouvoir donner une expression explicite de la résolvante**. Nous allons voir en revanche que l'on peut obtenir des informations *qualitatives* sur les solutions de l'équation grâce à l'étude de la résolvante.

Quelques propriétés de la résolvante

Proposition 2.11. *Soit $t_0 \in \mathbb{R}$. La fonction $\Delta(t) = \det R_A(t, t_0)$ vérifie l'équation différentielle*

$$\left\{ \begin{array}{rcl} \Delta'(t) & = & \mathrm{tr}(A(t))\,\Delta(t) \\ \Delta(t_0) & = & 1 \end{array} \right. ,$$

ce qui implique

$$\det R_A(t, t_0) = \exp\Big(\int_{t_0}^{t} \mathrm{tr}(A(u))\, du \Big).$$

Preuve.

▷ Rappelons que, si $A \in GL_n(\mathbb{K})$, $D\det(A) \cdot H = (\det A)\,\mathrm{tr}(A^{-1}H)$. On a donc

$$\begin{array}{rcl} \Delta'(t) & = & \big(\det R_A(t, t_0)\big)\,\mathrm{tr}\,\big(R_A^{-1}(t, t_0) R'_A(t, t_0)\big) \\ & = & \Delta(t)\,\mathrm{tr}(A(t)). \end{array}$$

La conclusion de la proposition suit.

□

Corollaire 2.12 (Liouville)**.** *Si pour tout $t \in \mathbb{R}$, $A(t)$ est de trace nulle, alors le déterminant de $R_A(t, s)$ est identiquement égal à 1.*

Il résulte de ce corollaire que, si $A(t)$ est de trace nulle, l'équation différentielle (2.5) préserve les volumes. En effet, si Γ est un domaine de $\mathbb{R}^n$, notons Γ_t son transport de t_0 à t par l'équation (2.5), c'est-à-dire

$$\Gamma_t = \{x(t) \ : \ x(\cdot) \text{ solution de (2.5) t.q. } x(t_0) \in \Gamma\},$$

ou encore $\Gamma_t = R_A(t, t_0)\Gamma$. Alors, en utilisant la formule de changement de variable dans les intégrales multiples, on obtient

$$\mathrm{vol}(\Gamma_t) = |\det R_A(t, t_0)|\,\mathrm{vol}(\Gamma),$$

et donc $\mathrm{vol}(\Gamma_t) = \mathrm{vol}(\Gamma)$ si $\mathrm{tr}\,A(t) \equiv 0$.

La préservation du volume a une conséquence sur le comportement asymptotique des solutions : il est en effet impossible dans ce cas que toutes les solutions de (2.5) tendent vers 0 quand $t \to \pm\infty$ (de même qu'il est impossible que $\|x(t)\|$ tende vers l'infini pour toute solution $x(\cdot)$).

Une classe particulièrement intéressante de matrices de trace nulle est l'ensemble des matrices antisymétriques, qui interviennent fréquemment dans les problèmes issus de la physique.

Proposition 2.13. *Si, pour tout $t \in \mathbb{R}$, $A(t)$ est une matrice réelle antisymétrique, la résolvante $R_A(t,s)$ est une rotation pour tous $t, s \in \mathbb{R}$.*

Rappelons qu'une rotation est une matrice $R \in M_n(\mathbb{R})$ *orthogonale* (c'est-à-dire $R^T R = I$) et de déterminant 1.

Preuve.

▷ On sait déjà d'après le corollaire précédent que $\det R_A(t,s) \equiv 1$. D'autre part

$$\begin{aligned}\frac{\partial}{\partial t}\big(R_A(t,s)^T R_A(t,s)\big) &= \frac{\partial}{\partial t}R_A(t,s)^T R_A(t,s) + R_A(t,s)^T \frac{\partial}{\partial t}R_A(t,s)\\ &= R_A(t,s)^T A(t)^T R_A(t,s) + R_A(t,s)^T A(t) R_A(t,s)\\ &= R_A(t,s)^T \big(A(t)^T + A(t)\big) R_A(t,s) = 0.\end{aligned}$$

Ainsi, $R_A(t,s)^T R_A(t,s)$ est constant. Comme $R_A(s,s) = I$, on a la conclusion.

□

Une conséquence de ce résultat est que, si $A(t)$ est une matrice réelle antisymétrique pour tout t, l'équation différentielle (2.5) préserve la norme. En effet, si $x(\cdot)$ est une solution de l'équation,

$$\|x(t)\| = \|R_A(t,t_0)x(t_0)\| = \|x(t_0)\|,$$

puisque $R_A(t,t_0)$ est une rotation. En particulier, toute solution est bornée. En revanche il est impossible qu'une solution tende vers 0 (sauf si $x(\cdot) \equiv 0$ bien sûr...). Nous verrons à la section 2.3.1 que l'on dit alors que 0 est un équilibre stable mais non asymptotiquement stable.

2.1.5 Linéarisation et perturbation du flot

Dans la pratique, on n'a quasiment jamais une connaissance exacte des conditions initiales (ni de l'équation elle-même, en fait). Il est donc primordial de savoir ce qui se passe pour la solution d'une équation différentielle lorsque la condition initiale est perturbée (ou quand l'équation elle-même est perturbée) : comment varie l'intervalle de définition et les valeurs de la solution, peut-on donner un ordre de grandeur de ces variations,... ? Les réponses à ces questions sont contenues dans le théorème ci-dessous : donnons d'abord le théorème et sa preuve, nous expliquerons ensuite pourquoi il permet de répondre aux questions précédentes.

Rappelons que nous considérons une équation différentielle autonome

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

où le champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est supposé de classe C^1.

Théorème 2.14. *Soit $\bar{x}(\cdot)$ une solution de l'équation (2.4) définie sur un intervalle $[a, b]$ contenant 0. Il existe alors un voisinage $\mathcal{V} \subset \Omega$ de $v_0 = \bar{x}(0)$ tel que, pour tout $v \in \mathcal{V}$, l'équation (2.4) admet une unique solution $x_v(\cdot)$ définie sur $[a, b]$ et vérifiant $x_v(0) = v$.*
De plus, l'application $v \mapsto x_v(\cdot)$ est de classe C^1 sur $\mathcal{V}$ et sa différentielle en v_0 est l'application qui à Δv associe la solution de

$$\left\{ \begin{array}{rcl} y'(t) & = & Df\big(\bar{x}(t)\big) \cdot y(t) \\ y(0) & = & \Delta v \end{array} \right. , \qquad t \in [a, b].$$

Remarque. La solution $x_v(\cdot)$ est simplement la restriction de la solution maximale $\phi(\cdot, v)$ à l'intervalle $[a, b]$.

Conséquences et signification du théorème 2.14

a. Domaine de définition du flot La première conséquence est que, pour toute condition initiale v dans le voisinage $\mathcal{V}$ de $\bar{x}(0)$, la solution maximale $\phi(\cdot, v)$ est définie sur tout l'intervalle $[a, b]$, *i.e.* $[a, b] \subset I_v$. Autrement dit, de façon informelle, si une solution est définie sur un temps « long », les solutions voisines sont également définies sur un temps « long ». Ceci se traduit par une propriété du domaine de définition du flot.

Corollaire 2.15. *Le flot ϕ est défini sur un ouvert $\mathcal{D}$ de $\mathbb{R} \times \Omega$.*
En particulier, si $(t, v_0) \in \mathcal{D}$, alors l'application ϕ_t est définie sur un voisinage de v_0.

Cette propriété est très importante pour l'étude du flot et de sa dépendance par rapport aux conditions initiales : en effet, ϕ et ϕ_t étant définies sur des ouverts, il est maintenant possible d'étudier leur continuité et leur différentiabilité.

Preuve.

▷ Rappelons que le domaine de définition du flot est

$$\mathcal{D} = \{(t, v) \in \mathbb{R} \times \Omega \ : \ t \in I_v\}.$$

Soit $(t_0, v_0) \in \mathcal{D}$. Puisque l'intervalle maximal I_{v_0} est ouvert (théorème 2.3), la solution maximale $\phi(\cdot, v_0)$ est définie sur un intervalle $[a, b] \subset I_{v_0}$ contenant t_0. Le théorème 2.14

implique alors que, pour tout v dans un voisinage $\mathcal{V}$ de v_0, on a encore $[a,b] \subset I_v$, c'est-à-dire que l'ensemble $]a,b[\times \mathcal{V}$, qui est un voisinage de (t_0, v_0) dans $\mathbb{R} \times \Omega$, est inclus dans $\mathcal{D}$.

□

b. Dépendance continue L'application $v \mapsto x_v(\cdot)$ définie dans le théorème 2.14 est l'application qui à une condition initiale dans $\mathcal{V}$ associe la solution correspondante de l'équation différentielle sur $[a,b]$. Cette application étant C^1, elle est en particulier continue, c'est-à-dire que
les solutions de l'équation différentielle (2.4) dépendent de façon continue de leur condition initiale.

C'est une propriété essentielle pour les applications (et d'un point de vue numérique) : en effet, elle signifie, grosso modo, que la solution calculée à partir d'une approximation de la condition initiale est une approximation de la vraie solution. Ceci justifie l'utilisation d'équations différentielles dans la modélisation de phénomènes réels, où on n'a qu'une connaissance approximative des données.

c. Équation linéarisée La dernière partie du théorème 2.14 affirme que les valeurs de la différentielle de l'application $\psi : v \mapsto x_v(\cdot)$ sont les solutions d'une certaine équation linéaire. Cette équation linéaire joue un rôle important dans la suite.

Définition 2.9. Soit $\bar{x}(\cdot) : [a,b] \to \Omega \subset \mathbb{R}^n$ une solution de (2.4). L'équation linéaire dans $\mathbb{R}^n$

$$y'(t) = Df\big(\bar{x}(t)\big) \cdot y(t), \qquad t \in [a,b],$$

est appelée *équation linéarisée de (2.4) autour de* $\bar{x}(\cdot)$.

Pour tout $\delta v \in \mathbb{R}^n$, $D\psi(\bar{x}(0)) \cdot \delta v$ est la solution de l'équation linéarisée autour de $\bar{x}(\cdot)$ valant δv en $t = 0$. En notant $R(t,s)$ la résolvante de l'équation linéarisée autour de $\bar{x}(\cdot)$, on obtient, pour tout $t \in [a,b]$,

$$\big(D\psi(\bar{x}(0)) \cdot \delta v\big)(t) = R(t,0)\delta v.$$

Intéressons-nous maintenant à l'application ϕ_t. Fixons un point v_0 de Ω et un instant $t \in I_{v_0}$. D'après le théorème 2.14, l'application ϕ_t est définie et de classe C^1 sur un voisinage $\mathcal{V}$ de v_0 dans Ω. Avec les notations ci-dessus, on a clairement $\phi_t(v) = x_v(t) = (\psi(v))(t)$ et donc

$$D\phi_t(v_0) \cdot \delta v = \big(D\psi(v_0) \cdot \delta v\big)(t).$$

On en déduit le résultat suivant.

Corollaire 2.16. *Soient $v_0 \in \Omega$ et $t \in I_{v_0}$. L'application ϕ_t est de classe C^1 sur un voisinage $\mathcal{V}$ de v_0 et*

$$D\phi_t(v_0) = R(t,0),$$

où R est la résolvante de l'équation linéarisée

$$y'(s) = Df\big(\phi_s(v_0)\big) \cdot y(s), \qquad s \in [0,t].$$

On peut donner une explication plus intuitive du rôle de l'équation linéarisée. On choisit une solution $\bar{x}(\cdot) : [a,b] \to \Omega$ de l'équation différentielle, de condition initiale $v_0 = \bar{x}(0)$. Considérons maintenant une perturbation $v_0 + \delta v$ de la condition initiale et écrivons la solution correspondante (*i.e.* $x_{v_0+\delta v}(\cdot)$) sous la forme d'une perturbation $\bar{x}(\cdot) + \delta x(\cdot)$ de la solution d'origine. Cette perturbation étant une solution, elle doit satisfaire l'équation différentielle :

$$\bar{x}'(t) + (\delta x)'(t) = f\big(\bar{x}(t) + \delta x(t)\big), \qquad t \in [a,b].$$

En utilisant un développement limité de f en $\bar{x}(t)$ (à t fixé) :

$$f\big(\bar{x}(t) + \delta x(t)\big) = f\big(\bar{x}(t)\big) + Df\big(\bar{x}(t)\big) \cdot \delta x(t) + \text{ reste},$$

et en tenant compte du fait que $\bar{x}'(t) = f\big(\bar{x}(t)\big)$, on obtient

$$(\delta x)'(t) = Df\big(\bar{x}(t)\big) \cdot \delta x(t) + \text{ reste}.$$

En ne conservant que les « termes du premier ordre » on retrouve l'équation linéarisée $(\delta x)'(t) = Df\big(\bar{x}(t)\big) \cdot \delta x(t)$. Autrement dit, la solution perturbée s'écrit $\bar{x}(\cdot) + \delta x(\cdot) +$ reste, où le « terme du premier ordre » $\delta x(\cdot)$ est la solution de

$$\begin{cases} (\delta x)'(t) &= Df\big(\bar{x}(t)\big) \cdot (\delta x)(t) \\ (\delta x)(0) &= \delta v \end{cases}.$$

L'équation linéarisée indique donc comment se propage au cours du temps une perturbation de la condition initiale.

Bien entendu ce qui précède n'est pas un raisonnement rigoureux (les restes posent évidemment quelques problèmes !), seulement une heuristique.

Remarque. L'équation linéarisée est en général une équation linéaire non-autonome, on ne sait donc pas a priori calculer ses solutions. Cependant, si v_0 est un point d'équilibre, la solution maximale $\bar{x}(\cdot) = \phi(\cdot, v_0)$ est la fonction constante $\bar{x}(\cdot) \equiv v_0$, définie sur tout $\mathbb{R}$, et dans ce cas l'équation linéarisée est autonome :

$$y'(t) = Df(v_0) \cdot y(t), \qquad t \in \mathbb{R}.$$

Exemple d'application : champs de vecteurs à divergence nulle. Rappelons d'abord que la divergence d'un champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = (f_1(x), \dots, f_n(x))$, est définie comme

$$\operatorname{div} f(x) = \frac{\partial f_1}{\partial x_1}(x) + \cdots + \frac{\partial f_n}{\partial x_n}(x) = \operatorname{tr} Df(x).$$

Considérons alors un temps t et un domaine Γ de $\mathbb{R}^n$, supposé inclus dans le domaine de définition de ϕ_t. Notons $\Gamma_t = \phi_t(\Gamma)$ le transport de Γ de 0 à t par l'équation (2.4). En utilisant la formule de changement de variable dans les intégrales multiples, on obtient

$$\operatorname{vol}(\Gamma_t) = \int_{\phi_t(\Gamma)} d\mu = \int_\Gamma |\det D\phi_t(x)| d\mu,$$

où, d'après le corollaire 2.16, $D\phi_t(x)$ est la résolvante de l'équation linéarisée.

Supposons maintenant que $\operatorname{div} f(x) = \operatorname{tr} Df(x) \equiv 0$. Le théorème de Liouville (corollaire 2.12) implique que le déterminant de la résolvante du linéarisé est égal à 1, et donc que $\det D\phi_t(x) = 1$. On a alors $\operatorname{vol}(\Gamma_t) = \operatorname{vol}(\Gamma)$, c'est-à-dire
si f est un champ de divergence nulle, le flot de f préserve le volume.

Dépendance par rapport à un paramètre

Considérons maintenant une famille d'équations différentielles dépendant d'un paramètre $\lambda \in \mathbb{R}^p$:

$$x'(t) = f_\lambda\big(x(t)\big), \tag{2.10}$$

où chaque f_λ est un champ de vecteurs sur $\Omega \subset \mathbb{R}^n$. Supposons également que $f(x, \lambda) = f_\lambda(x)$ est une application de classe C^1. On s'intéresse à la dépendance des solutions de ces équations différentielles par rapport au paramètre λ.

Remarquons d'abord que l'équation (2.10) est équivalente à

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t), \lambda\big) \\ \lambda'(t) & = & 0 \end{array} \right. ,$$

c'est-à-dire à l'équation différentielle dans $\mathbb{R}^{n+p}$ associée au champ de vecteurs $F(x, \lambda) = (f(x, \lambda), 0)$. Ainsi, les solutions de (2.10) dépendent du paramètre λ de la même façon que les solutions de l'équation différentielle $(x, \lambda)'(t) = F\big((x, \lambda)(t)\big)$ dépendent de leur condition initiale. D'après ce que nous avons vu précédemment dans cette section, nous avons donc les propriétés suivantes.

— Les solutions $\phi^\lambda(\cdot, v)$ de (2.10) dépendent de façon C^1, donc continue, du paramètre λ (et de la condition initiale v) ;

— La différentielle de l'application $(v, \lambda) \mapsto \phi^\lambda(\cdot, v)$ en un point (v_0, λ_0) est l'application qui à $(\delta v, \delta\lambda)$ associe la solution $y(\cdot)$ de l'équation différentielle affine

$$\left\{ \begin{array}{rcl} y'(t) & = & D_x f\big(\bar{x}(t), \lambda_0\big) \cdot y(t) + D_\lambda f\big(\bar{x}(t), \lambda_0\big) \cdot \delta\lambda \\ y(0) & = & \delta v \end{array} \right. ,$$

où $\bar{x}(\cdot) = \phi^{\lambda_0}(\cdot, v_0)$. Autrement dit, en utilisant la formule de variation de la constante,

$$y(t) = R(t, 0)\delta v + \int_0^t R(t, s) D_\lambda f\big(\bar{x}(s), \lambda_0\big) \cdot \delta\lambda \, ds,$$

où $R(t, s)$ est la résolvante associée au système linéarisé $y'(t) = D_x f\big(\bar{x}(t), \lambda_0\big) \cdot y(t)$.

***Application de Poincaré**

Supposons que l'équation (2.4) admette une solution T-périodique non triviale $x(\cdot)$ et notons $x(0) = x_0$. Traçons un hyperplan affine Σ passant par x_0 et *transverse* à $x(\cdot)$ en x_0, c'est-à-dire que $f(x_0)$ n'est pas parallèle à Σ (notez que $f(x_0) \neq 0$ puisque $x(\cdot)$ est non triviale). Nous ferons pour simplifier l'hypothèse que Σ est en fait un hyperplan affine parallèle à $\mathbb{R}^{n-1}$ et que $f(x_0) \in \mathbb{R}e_1$, où e_1 vérifie $\mathbb{R}^n = \mathbb{R}e_1 \oplus \mathbb{R}^{n-1}$: c'est une situation à laquelle on peut toujours se ramener par un changement linéaire de coordonnées.

Proposition 2.17. *Il existe un voisinage $\mathcal{V} \subset \mathbb{R}^n$ de x_0 et une fonction η de classe C^1 tels que, pour tout z dans $\Sigma \cap \mathcal{V}$, $\phi_{T+\eta(z)}(z)$ appartient à $\Sigma \cap \mathcal{V}$. L'application $G : \Sigma \cap \mathcal{V} \to \Sigma \cap \mathcal{V}$ ainsi définie est un C^1-difféomorphisme : c'est* l'application de premier retour de Poincaré.
On a en outre, pour tout $(\delta s, \delta z) \in \mathbb{R} \times \mathbb{R}^{n-1}$,

$$D\phi_T(x_0) \cdot \begin{pmatrix} \delta s \\ \delta z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & D_z G(x_0) \end{pmatrix} \cdot \begin{pmatrix} \delta s \\ \delta z \end{pmatrix}.$$

Remarque. Dans le cas des flots, l'application de Poincaré ne dépend que des sections transverses à l'orbite, Σ_0, Σ_1, et encore à conjugaison près : elle a donc un sens géométrique très fort. D'autre part cette construction est très utile dans l'étude du voisinage d'une orbite périodique. Enfin notez que la construction précédente permet de passer de l'étude d'un flot en dimension n à celui d'un difféomorphisme local en dimension $n-1$.

2.2 Équations différentielles linéaires autonomes

Nous abordons dans ce chapitre l'étude des équations différentielles les plus simples, les équations linéaires *autonomes* – aussi appelées équations linéaires à *coefficients constants* –, c'est-à-dire les équations de la forme

$$x'(t) = Ax(t). \tag{2.11}$$

Précisons les notations. La *donnée* $A \in M_n(\mathbb{K})$ est une matrice carrée $(n \times n)$ à coefficients dans $\mathbb{K}$, où $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. L'*inconnue* est une application dérivable $x(\cdot) : \mathbb{R} \to \mathbb{K}^n$. Résoudre l'équation (2.11) signifie trouver une application $x(\cdot)$ telle que, pour tout $t \in \mathbb{R}$, la dérivée $x'(t) = \frac{dx}{dt}(t)$ vérifie $x'(t) = Ax(t)$.

Cette équation est dite autonome parce que la donnée $A \in M_n(\mathbb{R})$ ne dépend pas du temps.

2.2.1 Approche élémentaire

Commençons par un cas connu, celui d'une équation scalaire

$$x'(t) = \alpha x(t),$$

où α est un réel et x une fonction de $\mathbb{R}$ dans $\mathbb{R}$. Une solution $x(\cdot)$ de cette équation décrit l'évolution en fonction du temps d'une quantité dont le taux de variation α est constant.

Rappelons (cela résulte également du théorème 2.19) que la seule solution de cette équation valant x_0 à l'instant t_0 est

$$x(t) = x_0 e^{\alpha(t-t_0)}.$$

Cette expression nous fournit toutes les informations que l'on peut souhaiter sur l'équation différentielle. Par exemple le comportement asymptotique de $x(t)$ quand $t \to +\infty$ est caractérisé par le signe de α :

— si $\alpha < 0$, $\lim_{t\to+\infty} x(t) = 0$,

— si $\alpha = 0$, $x(t)$ est constant,

— si $\alpha > 0$, $\lim_{t\to+\infty} |x(t)| = \begin{cases} +\infty & \text{si } x(t_0) \neq 0 \\ 0 & \text{si } x(t_0) = 0 \end{cases}$.

Considérons maintenant un système de deux équations différentielles

$$\begin{cases} x_1' &= \alpha_1 x_1 \\ x_2' &= \alpha_2 x_2 \end{cases}.$$

C'est un système très simple puisque les fonctions $x_1(t)$ et $x_2(t)$ sont découplées. La solution de ce système est bien évidemment

$$x_1(t) = x_1(t_0)e^{\alpha_1(t-t_0)}, \quad x_2(t) = x_2(t_0)e^{\alpha_2(t-t_0)}.$$

Comme dans le cas scalaire, on a une connaissance complète du comportement des solutions. Par exemple, si α_1 et α_2 sont strictement négatifs, toute solution $(x_1(t), x_2(t))$ du système d'équations tend vers l'origine quand $t \to +\infty$; si $\alpha_1 > 0$ et $x_1(t_0) \neq 0$, la norme de $(x_1(t), x_2(t))$ tend vers l'infini quand $t \to +\infty$...

Adoptons maintenant une écriture matricielle. En posant $x = (x_1, x_2)$, le système de deux équations ci-dessus apparaît comme le cas particulier $n = 2$ de l'équation différentielle dans $\mathbb{R}^n$ suivante :

$$x'(t) = \Delta x(t), \quad \text{où } \Delta = \begin{pmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_n \end{pmatrix} \text{ est diagonale.} \tag{2.12}$$

Cette équation étant en fait un système de n équations scalaires découplées, la solution est donnée par

$$x(t) = \begin{pmatrix} x_1(t_0)e^{\alpha_1(t-t_0)} \\ \vdots \\ x_n(t_0)e^{\alpha_n(t-t_0)} \end{pmatrix} = \begin{pmatrix} e^{\alpha_1(t-t_0)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\alpha_n(t-t_0)} \end{pmatrix} x(t_0),$$

avec $x = (x_1, \ldots, x_n)$. Nous verrons dans la section suivante que la matrice diagonale ci-dessus est l'exponentielle de la matrice Δ et nous la noterons donc $e^{(t-t_0)\Delta}$.

Ainsi, pour les équations différentielles de la forme (2.12), nous avons une connaissance parfaite des solutions. Nous sommes par exemple en mesure d'analyser le comportement asymptotique des solutions en fonction de $\alpha_1, \ldots, \alpha_n$ et de la condition initiale $x(t_0)$:

— si tous les α_i sont strictement négatifs, toute solution $x(t)$ converge vers l'origine quand $t \to +\infty$;
— si tous les α_i sont négatifs ou nuls, toute solution $x(t)$ est bornée quand $t \to +\infty$;
— si au moins un des α_i est strictement positif, alors $\lim_{t\to+\infty} \|x(t)\| = +\infty$ pour toute solution vérifiant $x_i(t_0) \neq 0$;
— etc...

L'équation différentielle $x'(t) = \Delta x(t)$ que nous venons de traiter est un cas très particulier, puisqu'il correspond à un système de n équations scalaires découplées. Beaucoup d'équations différentielles linéaires peuvent cependant s'y ramener. Considérons en effet le système $x'(t) = Ax(t)$ avec A diagonalisable dans $\mathbb{R}$: il existe donc une matrice inversible $P \in GL_n(\mathbb{R})$ et une matrice diagonale $\Delta \in M_n(\mathbb{R})$ telles que $A = P\Delta P^{-1}$.

Remarquons maintenant que, si $x(t)$ vérifie $x'(t) = Ax(t)$, alors $y(t) = P^{-1}x(t)$ vérifie $y'(t) = \Delta y(t)$. Autrement dit, *à un changement de coordonnées près*, l'équation différentielle est un système de n équations scalaires découplées. Connaissant $y(t_0) = P^{-1}x(t_0)$, on obtient alors $y(t) = e^{(t-t_0)\Delta}y(t_0)$, et

$$x(t) = Py(t) = Pe^{(t-t_0)\Delta}y(t_0) = Pe^{(t-t_0)\Delta}P^{-1}x(t_0).$$

On est donc encore capable de calculer les solutions de l'équation différentielle dans ce cas. Plus important, on voit que le comportement asymptotique des solutions est caractérisé par les éléments diagonaux de Δ, c'est-à-dire par les *valeurs propres* de A.

En résumé, cette première approche élémentaire fait apparaître les points clés que nous allons développer maintenant :
— les solutions se calculent à l'aide de l'exponentielle de matrice ;
— le comportement asymptotique des solutions est caractérisé par les valeurs propres de A.

2.2.2 Exponentielle de matrices

Définition 2.10. On appelle *exponentielle de matrice* l'application

$$\begin{array}{rccc} \exp : & M_n(\mathbb{K}) & \longrightarrow & M_n(\mathbb{K}) \\ & A & \longmapsto & \exp A = e^A = \displaystyle\sum_{k=0}^{\infty} \frac{A^k}{k!} \end{array} .$$

Notons que la série $\sum_{k=0}^{\infty} \frac{A^k}{k!}$ converge normalement. En effet

$$\sum_{k=0}^{\infty} \frac{\|A^k\|}{k!} \leq \sum_{k=0}^{\infty} \frac{\|A\|^k}{k!} = e^{\|A\|} < \infty,$$

où on a choisi pour $\|\cdot\|$ une norme multiplicative sur $M_n(\mathbb{K})$ (par exemple une norme d'opérateurs). L'application exponentielle est donc continue (elle est en fait C^∞). Rappelons ses propriétés principales (sans démonstrations).

Proposition 2.18.

1. *Pour $A \in M_n(\mathbb{K})$, l'application $t \mapsto e^{tA}$ est dérivable et*
$$\frac{d}{dt}e^{tA} = Ae^{tA} = e^{tA}A$$
2. *Pour tout $A \in M_n(\mathbb{K})$, $\exp(A)$ est inversible et $\exp(A)^{-1} = \exp(-A)$.*
3. *Si A et $B \in M_n(\mathbb{K})$* commutent, *i.e. $AB = BA$, on a*
$$\exp(A+B) = \exp(A)\exp(B).$$
4. *Si $P \in GL_n(\mathbb{K})$, alors $Pe^AP^{-1} = e^{PAP^{-1}}$.*
5. *Si Δ est une matrice diagonale d'éléments diagonaux $\lambda_1, \ldots, \lambda_n$, alors e^Δ est diagonale d'éléments diagonaux $e^{\lambda_1}, \ldots, e^{\lambda_n}$.*

Nous pouvons maintenant résoudre l'équation (2.11).

Théorème 2.19. *Soient $t_0 \in \mathbb{R}$ et $x_0 \in \mathbb{K}^n$. L'unique solution de l'équation $x'(t) = Ax(t)$ valant x_0 en t_0 est l'application $x(\cdot)$ définie par*
$$x(t) = e^{(t-t_0)A}x_0, \qquad \forall t \in \mathbb{R}.$$

Preuve.

▷ La première propriété de la proposition précédente implique
$$\frac{d}{dt}\left(e^{(t-t_0)A}x_0\right) = \left(\frac{d}{dt}e^{(t-t_0)A}\right)x_0 = Ae^{(t-t_0)A}x_0.$$
La fonction $x(t) = e^{(t-t_0)A}x_0$ est donc solution de $x'(t) = Ax(t)$, et $x(t_0) = x_0$ puisque $e^{0A} = I$.
Pour montrer l'unicité de la solution, considérons une autre solution $y(t)$ valant x_0 en t_0 et formons le produit $z(t) = e^{-(t-t_0)A}y(t)$. En utilisant encore la proposition 2.18, on obtient
$$z'(t) = -Ae^{-(t-t_0)A}y(t) + e^{-(t-t_0)A}y'(t) = -Ae^{-(t-t_0)A}y(t) + e^{-(t-t_0)A}Ay(t) = 0,$$
car A et e^{tA} commutent. Donc $z(t)$ est une constante, et puisque $z(t_0) = x_0$, on obtient $y(t) = e^{(t-t_0)A}z(t_0) = e^{(t-t_0)A}x_0$.

□

Ainsi, la résolution d'équations linéaires autonomes se ramène au calcul d'exponentielle de matrices. Remarquons en particulier que les deux derniers points de la proposition 2.18 permettent de retrouver le résultat de la section 2.2.1 : si A est diagonalisable dans $\mathbb{K}$, c'est-à-dire

$$A = P\Delta P^{-1}, \quad P \in GL_n(\mathbb{K}), \quad \Delta \in M_n(\mathbb{K}) \text{ diagonale},$$

alors $e^{tA} = Pe^{t\Delta}P^{-1}$ et la solution de l'équation (2.11) est

$$x(t) = Pe^{(t-t_0)\Delta}P^{-1}x(t_0).$$

Malheureusement toutes les matrices ne sont pas diagonalisables. La théorie de la réduction des endomorphismes permet cependant de mener à bien le calcul.

2.2.3 Calcul de l'exponentielle de matrices

Le but de cette section est de calculer l'exponentielle d'une matrice $A \in M_n(\mathbb{K})$ à un changement de base près, c'est-à-dire de calculer $P^{-1}e^{tA}P$ pour un $P \in GL_n(\mathbb{K})$ bien choisi.

Nous mènerons d'abord ce calcul dans le cas $\mathbb{K} = \mathbb{C}$, pour les matrices à coefficients complexes. Nous montrerons ensuite comment en déduire le cas $\mathbb{K} = \mathbb{R}$.

a. Matrices à coefficients complexes

Polynôme caractéristique. Considérons une matrice $A \in M_n(\mathbb{C})$ et notons $\lambda_1, \ldots, \lambda_r$ ses *valeurs propres*. Rappelons que ce sont les seuls nombres complexes pour lesquels l'équation

$$Av = \lambda_i v$$

admette une solution $v \in \mathbb{C}^n$ *non nulle* (les v_i correspondants s'appellent des *vecteurs propres*). Les valeurs propres s'obtiennent également comme les racines du *polynôme caractéristique* de A : $P_A(\lambda) = \det(\lambda I - A)$. Ce polynôme est donc de la forme

$$P_A(\lambda) = (\lambda - \lambda_1)^{p_1} \cdots (\lambda - \lambda_r)^{p_r},$$

où chaque entier p_i est strictement positif et $p_1 + \cdots + p_r = n$ (le polynôme caractéristique étant de degré n). On appelle p_i la *multiplicité algébrique* de la valeur propre λ_i.

Une propriété importante du polynôme caractéristique est qu'il s'annule en A.

Théorème 2.20 (Cayley-Hamilton)**.** *Toute matrice annule son polynôme caractéristique :*

$$P_A(A) = (A - \lambda_1 I)^{p_1} \cdots (A - \lambda_r I)^{p_r} = 0.$$

Sous-espaces propres, sous-espaces caractéristiques. À chaque valeur propre de A sont associés deux sous-espaces vectoriels de $\mathbb{C}^n$. Le premier est le *sous-espace propre* :

$$\Pi_i \text{ ou } \Pi_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I).$$

C'est l'ensemble des vecteurs propres associés à λ_i. L'entier $e_i = \dim \Pi_i$ est appelé la *multiplicité géométrique* de la valeur propre λ_i.

Le deuxième est le *sous-espaces caractéristique* :

$$\Gamma_i \text{ ou } \Gamma_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I)^{p_i}.$$

Il est clair que $\Pi_i \subset \Gamma_i$, mais ces deux espaces peuvent être différents. Le rôle des espaces caractéristiques est précisé dans le résultat suivant, que nous donnons sans démonstration.

Théorème 2.21 (de décomposition des noyaux)**.** *Avec les notations précédentes, on a la décomposition*

$$\mathbb{C}^n = \Gamma_1 \oplus \cdots \oplus \Gamma_r,$$

et les propriétés suivantes :

1. $\dim \Gamma_i = p_i$ *;*
2. *chacun des espaces* Γ_i *est invariant par* A *:* $x \in \Gamma_i \Rightarrow Ax \in \Gamma_i$ *;*
3. *la restriction* $A|_{\Gamma_i}$ *de* A *à* Γ_i *s'écrit*

$$A|_{\Gamma_i} = \lambda_i I_{\Gamma_i} + N_i,$$

où I_{Γ_i} *désigne l'identité de* Γ_i *et* $N_i \in \mathrm{End}(\Gamma_i)$ *est nilpotent d'ordre* $\leq p_i$*, i.e.* $N_i^{p_i} = 0$*.*

Ce résultat appelle un certain nombre de commentaires.

— Rappelons que $\mathrm{End}(\Gamma_i)$ désigne l'ensemble des *endomorphismes* de Γ_i, c'est-à-dire des applications linéaires de Γ_i dans lui-même. Dire que Γ_i est invariant par A est équivalent à dire que $A|_{\Gamma_i}$ appartient à $\mathrm{End}(\Gamma_i)$.

— L'opérateur N_i est défini comme $N_i = (A - \lambda_i I)|_{\Gamma_i}$. Le fait que N_i soit nilpotent d'ordre $\leq p_i$ n'est donc rien d'autre que la définition de Γ_i. Il est en revanche possible que l'ordre exact de nilpotence de N_i, c'est-à-dire le plus petit entier $m_i \leq p_i$ tel que $N_i^{m_i} = 0$, soit plus petit que p_i. Dans ce cas, on a

$$\ker(A - \lambda_i I)^{p_i} = \ker(A - \lambda_i I)^{m_i} \supsetneq \ker(A - \lambda_i I)^{m_i - 1}.$$

— Une matrice est *diagonalisable* si il existe une base de $\mathbb{C}^n$ formée de vecteurs propres, ce qui équivaut à

$$\mathbb{C}^n = \Pi_1 \oplus \cdots \oplus \Pi_r.$$

D'après le théorème de décomposition des noyaux, ceci n'est possible que si $\Pi_i = \Gamma_i$ pour tout i. Autrement dit :

A est diagonalisable si et seulement si pour toute valeur propre les multiplicités algébrique et géométrique coïncident, i.e. $\dim \Pi_i = p_i$ *pour* $i = 1, \dots, r$.

Réduction de Jordan dans $\mathbb{C}^n$. Choisissons une base $\mathcal{B}$ de $\mathbb{C}^n$ formée de la réunion d'une base de Γ_1, d'une base de Γ_2, ..., d'une base de Γ_r, et notons P la matrice de passage de cette base à la base canonique. D'après le théorème de décomposition des noyaux, l'application linéaire associée à A a pour matrice dans la base $\mathcal{B}$:

$$P^{-1}AP = \Delta + N,$$

où Δ est la matrice diagonale ayant pour éléments diagonaux λ_1 (p_1 fois), ..., λ_r (p_r fois), et N est la matrice nilpotente qui s'écrit par blocs

$$N = \begin{pmatrix} N_1 & & \\ & \ddots & \\ & & N_r \end{pmatrix}.$$

Il est en fait possible de choisir la base $\mathcal{B}$ de façon à mettre la matrice nilpotente N sous une forme relativement simple. On aboutit ainsi à la réduction de Jordan.

Théorème 2.22 (de Jordan)**.** *Pour toute matrice $A \in M_n(\mathbb{C})$, il existe $P \in GL_n(\mathbb{C})$ telle que $P^{-1}AP$ s'écrit sous forme de matrice diagonale par bloc*

$$P^{-1}AP = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix}, \tag{2.14}$$

où chaque J_i est une matrice $(p_i \times p_i)$ de la forme

$$J_i = \begin{pmatrix} J_{i,1} & & \\ & \ddots & \\ & & J_{i,e_i} \end{pmatrix}, \quad \textit{avec} \quad J_{i,k} = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

et $e_i = \dim \ker(A - \lambda_i I)$ est la multiplicité géométrique de λ_i.

On appelle la matrice $J = P^{-1}AP$ la *forme réduite de Jordan* de A et les matrices J_i des *blocs de Jordan*.

Remarques.

— Pour chaque i, la matrice J_i représente l'application linéaire $A|_{\Gamma_i}$, ce qui implique $J_i \in \mathrm{End}(\Gamma_i)$.

— Les matrices $J_{i,k}$ sont des matrices carrées dont la dimension $\dim(J_{i,k}) \leq p_i$ dépend de i et k. Dans le cas particulier où les multiplicités algébrique et géométrique de λ_i coïncident (*i.e.* $e_i = p_i$), la dimension $\dim(J_{i,k})$ est égale à 1 pour tout k. Chaque bloc $J_{i,k}$ est alors réduit au scalaire λ_i et $J_i = \lambda_i I_{p_i}$.
— D'après la remarque précédente, si pour chaque valeur propre les multiplicités algébrique et géométrique coïncident, la forme réduite de Jordan est diagonale. Ainsi la réduction de Jordan généralise la diagonalisation. Insistons cependant sur le fait que toute matrice de $M_n(\mathbb{C})$ admet une réduction de Jordan, mais n'est pas forcément diagonalisable.

Calcul de l'exponentielle. La réduction de Jordan permet de calculer e^{tA} à conjugaison près. En effet, commençons par le calcul de l'exponentielle d'un bloc $J_{i,k}$. Notons $n_{i,k} = \dim(J_{i,k})$ la dimension de ce bloc et écrivons-le $J_{i,k} = \lambda_i I + N_{i,k}$, où $N_{i,k}$ est la matrice ayant des 0 sur la diagonale et des 1 juste au-dessus. Comme $\lambda_i I$ et $N_{i,k}$ commutent,

$$e^{tJ_{i,k}} = e^{t\lambda_i I} e^{tN_{i,k}} = e^{t\lambda_i} e^{tN_{i,k}}.$$

De plus, $N_{i,k}$ étant nilpotente d'ordre $n_{i,k}$, on a :

$$e^{tN_{i,k}} = \sum_{l=0}^{\infty} \frac{(tN_{i,k})^l}{l!} = \sum_{l=0}^{n_{i,k}-1} \frac{(tN_{i,k})^l}{l!} = \begin{pmatrix} 1 & t & \dots & \frac{t^{n_{i,k}-1}}{(n_{i,k}-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}.$$

D'autre part, d'après les propriétés de l'exponentielle de matrice (proposition 2.18), on a $e^{tA} = e^{tPJP^{-1}} = Pe^{tJ}P^{-1}$, ce qui donne finalement l'expression de l'exponentielle de tA :

$$e^{tA} = P \begin{pmatrix} e^{tJ_{1,1}} & & \\ & \ddots & \\ & & e^{tJ_{r,e_r}} \end{pmatrix} P^{-1}, \tag{2.15}$$

$$\text{avec } e^{tJ_{i,k}} = e^{t\lambda_i} \begin{pmatrix} 1 & t & \dots & \frac{t^{n_{i,k}-1}}{(n_{i,k}-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}.$$

b. Matrices à coefficients réels

Considérons maintenant une matrice $A \in M_n(\mathbb{R})$, à coefficients *réels*. On peut bien entendu considérer A comme une matrice de $M_n(\mathbb{C})$; tout ce que nous venons de voir pour les matrices à coefficients complexes s'applique donc.

Désignons les valeurs propres réelles de A par $\lambda_1, \ldots, \lambda_s$, et ses valeurs propres non réelles par $\lambda_{s+1}, \bar{\lambda}_{s+1}, \ldots, \lambda_q, \bar{\lambda}_q$ (avec $2q - s = r$). Le polynôme caractéristique de A est donc le polynôme à coefficients réels

$$P_A(\lambda) = \prod_{i=1}^{s} (\lambda - \lambda_i)^{p_i} \prod_{i=s+1}^{q} [(\lambda - \lambda_i)(\lambda - \bar{\lambda}_i)]^{p_i}.$$

Les sous-espaces vectoriels $\Gamma_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I)^{p_i}$ de $\mathbb{C}^n$ sont maintenant appelés les sous-espaces caractéristiques *complexes*.

Remarque. Rappelons les liens qui existent entre les sous-espaces vectoriels de $\mathbb{C}^n$, qui sont des $\mathbb{C}$-espaces vectoriels, et ceux de $\mathbb{R}^n$, qui sont des $\mathbb{R}$-espaces vectoriels. On considère $\mathbb{R}^n$ comme un sous-ensemble de $\mathbb{C}^n$ et, pour un sous-espace vectoriel Γ de $\mathbb{C}^n$, on note $\Gamma \cap \mathbb{R}^n$ l'ensemble des vecteurs $v \in \Gamma$ qui sont réels. Il est facile de vérifier qu'un tel ensemble $\Gamma \cap \mathbb{R}^n$ est un sous-espace vectoriel de $\mathbb{R}^n$. De plus, si Γ est stable par conjugaison (*i.e.* $v \in \Gamma \Rightarrow \bar{v} \in \Gamma$), alors Γ et $\Gamma \cap \mathbb{R}^n$ ont même dimension en tant que sous-espaces respectivement de $\mathbb{C}^n$ et de $\mathbb{R}^n$ (en fait, dans ce cas, Γ est l'ensemble des combinaisons linéaires à coefficients complexes des éléments de $\Gamma \cap \mathbb{R}^n$; en conséquence, toute base du $\mathbb{R}$-espace vectoriel $\Gamma \cap \mathbb{R}^n$ est également une base du $\mathbb{C}$-espace vectoriel Γ).

Définissons maintenant les *sous-espaces caractéristiques réels* de A comme les sous-espaces vectoriels de $\mathbb{R}^n$:

$$\begin{aligned} E_i &= \Gamma_{\lambda_i} \cap \mathbb{R}^n, && 1 \leq i \leq s \\ E_i &= (\Gamma_{\lambda_i} \oplus \Gamma_{\bar{\lambda}_i}) \cap \mathbb{R}^n, && s+1 \leq i \leq q. \end{aligned}$$

Remarquons alors que $(A - \lambda_i I)^{p_i} v = 0$ implique $(A - \bar{\lambda}_i I)^{p_i} \bar{v} = 0$, ce qui signifie que Γ_{λ_i} pour λ_i réel et $\Gamma_{\lambda_i} \oplus \Gamma_{\bar{\lambda}_i}$ pour λ_i non réel sont stables par conjugaison. D'après la remarque précédente et le théorème de décomposition des noyaux dans $\mathbb{C}^n$, on a la décomposition

$$\mathbb{R}^n = E_1 \oplus \cdots \oplus E_q,$$

chaque sous-espace E_i étant invariant par A. À partir de cette décomposition, nous allons donner ci-dessous une forme réduite de Jordan réelle de A. Cette forme réduite n'est cependant pas indispensable à l'étude des équations différentielles : nous verrons dans la section suivante que les solutions de l'équation (2.11) dans $\mathbb{R}^n$ peuvent se déduire directement des solutions dans $\mathbb{C}^n$.

***Réduction de Jordan dans $\mathbb{R}^n$.** Donner une forme réduite de $A \in M_n(\mathbb{R})$ consiste à trouver pour chaque sous-espace caractéristique une base dans laquelle l'application linéaire associée à A a une expression simple (c'est ce que nous avons fait pour les matrices à coefficients complexes). Considérons donc un des sous-espaces caractéristiques réels E_k de A.

— Si λ_k est réelle, *i.e.* $1 \leq k \leq s$: dans ce cas $E_i = \ker_{\mathbb{R}}(A - \lambda_k I)^{p_k}$ et la restriction de A à E_k s'écrit simplement

$$A|_{E_k} = \lambda_k I|_{E_k} + N_k, \qquad \text{où } N_k \text{ nilpotente.}$$

On peut alors montrer comme dans le cas complexe que $A|_{E_k}$ est conjuguée au bloc de Jordan J_k.

— Si $\lambda_k = \alpha_k + i\beta_k$ n'est pas réelle, *i.e.* $s+1 \leq k \leq q$: choisissons une base $v_1, \dots, v_{p_k}$ de Γ_k dans laquelle $A|_{\Gamma_k}$ se met sous forme d'un bloc de Jordan J_k, c'est-à-dire, pour $j = 1, \dots, p_k$,

$$Av_j = \lambda_k v_j + \delta_j v_{j-1}, \qquad \text{où } \delta_j = (J_k)_{j-1,j} = 0 \text{ ou } 1,$$

en posant $v_0 = 0$. Par conjugaison, il est clair que $\bar{v}_1, \dots, \bar{v}_{p_k}$ est une base de $\Gamma_{\bar{\lambda}_k}$. Pour $j = 1, \dots, p_k$, posons $v_j = a_j + ib_j$, avec $a_j, b_j \in \mathbb{R}^n$. Les vecteurs $(a_1, b_1, \dots, a_{p_k}, b_{p_k})$ forment alors une base de E_k. En effet, les $a_j = \frac{1}{2}(v_j + \bar{v}_j)$ et $b_j = \frac{i}{2}(\bar{v}_j - v_j)$ appartiennent à $\Gamma_{\lambda_k} \oplus \Gamma_{\bar{\lambda}_k}$ et l'engendrent en tant que $\mathbb{C}$-espace vectoriel puisqu'ils engendrent la base des v_j, $\bar{v}_j$: ils forment donc une famille génératrice à $2p_k$ éléments du $\mathbb{R}$-espace vectoriel E_k de dimension $2p_k$ éléments, c'est-à-dire une base.
De plus, en identifiant les parties réelles et imaginaires dans l'expression

$$A(a_j + ib_j) = (\alpha_k + i\beta_k)(a_j + ib_j) + \delta_j(a_{j-1} + ib_{j-1}),$$

on obtient

$$A\begin{bmatrix} a_j \\ b_j \end{bmatrix} = C_k \begin{bmatrix} a_j \\ b_j \end{bmatrix} + \delta_j \begin{bmatrix} a_{j-1} \\ b_{j-1} \end{bmatrix} \qquad \text{où} \quad C_k = \begin{pmatrix} \alpha_k & -\beta_k \\ \beta_k & \alpha_k \end{pmatrix}.$$

La restriction de A à E_k est donc conjuguée dans la base $(a_1, b_1, \dots, a_{p_k}, b_{p_k})$ à la matrice

$$J'_k = \begin{pmatrix} C_k & \delta_2 I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \delta_{p_k} I_2 \\ & & & C_k \end{pmatrix}.$$

On obtient ainsi la réduction de Jordan dans $\mathbb{R}^n$ de A.

Théorème 2.23. *Pour toute matrice $A \in M_n(\mathbb{R})$, il existe $Q \in GL_n(\mathbb{R})$ telle que $Q^{-1}AQ$ s'écrit sous forme de matrice diagonale par bloc J', d'éléments diagonaux $J_1, \dots, J_s$, $J'_{s+1}, \dots, J'_q$, où, pour $i = s+1, \dots, q$, chaque J'_i est une matrice $(2p_i \times 2p_i)$ de la forme*

$$J'_i = \begin{pmatrix} J'_{i,1} & & \\ & \ddots & \\ & & J'_{i,2e_i} \end{pmatrix}, \quad \textit{avec} \quad J'_{i,k} = \begin{pmatrix} C_i & I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & I_2 \\ & & & C_i \end{pmatrix}$$

et, pour $i = 1, \dots, s$, les matrices J_i sont celles données dans le théorème 2.22.

Le calcul de l'exponentielle ne pose alors aucun problème, il suffit de savoir calculer e^{tC_i}. Écrivons C_i comme la somme de deux matrices qui commutent

$$C_i = \alpha_i I_2 + \beta_i B, \qquad B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

et donc son exponentielle comme un produit

$$e^{tC_i} = e^{\alpha_i t} e^{\beta_i t B}.$$

Remarquons que la matrice B est de carré égal à $-I_2$, ce qui implique $B^{2p} = (-1)^p I_2$ et $B^{2p+1} = (-1)^p B$, soit

$$e^{sB} = \sum_{k=0}^{\infty} \frac{1}{k!}(sB)^k = \sum_{p=0}^{\infty} \frac{1}{(2p)!}(-1)^p s^{2p} I_2 + \sum_{p=0}^{\infty} \frac{1}{(2p+1)!}(-1)^p s^{2p+1} B = \cos(s) I_2 + \sin(s) B.$$

On obtient ainsi l'exponentielle de C_i :

$$e^{tC_i} = e^{\alpha_i t} \begin{pmatrix} \cos(\beta_i t) & -\sin(\beta_i t) \\ \sin(\beta_i t) & \cos(\beta_i t) \end{pmatrix}.$$

2.2.4 Forme des solutions

Grâce au calcul de l'exponentielle que nous venons d'effectuer, nous sommes maintenant en mesure de préciser le théorème 2.19. Donnons d'abord la forme de la solution générale dans $\mathbb{C}^n$, qui s'obtient directement à partir de l'expression (2.15) de e^{tA}.

Théorème 2.24. *Soit $A \in M_n(\mathbb{C})$. Toute solution de $x'(t) = Ax(t)$ dans $\mathbb{C}^n$ s'écrit sous la forme*

$$x(t) = \sum_{1 \le i \le r} e^{t\lambda_i} \Big(\sum_{0 \le k \le m_i - 1} t^k \, v_{i,k} \Big), \qquad \text{où} \quad v_{i,k} \in \Gamma_i, \tag{2.17}$$

avec $m_i = \max_{1 \le k \le e_i} \dim(J_{i,k})$.

Remarques.

— Le terme en facteur de $e^{t\lambda_i}$ est polynômial en t quand $m_i > 1$, et constant quand $m_i = 1$. Rappelons que ce dernier cas a lieu si et seulement si les multiplicités algébrique et géométrique de λ_i coïncident.

— Il est également important de noter la dépendance de $x(t)$ par rapport à la condition initiale $x(0)$ dans la décomposition précédente. Si $x(0)$ s'écrit comme $x(0) = v_1 + \cdots + v_r$ dans $\mathbb{C}^n = \Gamma_1 \oplus \cdots \oplus \Gamma_r$, alors

$$v_{i,k} = \frac{1}{k!} N^k v_i,$$

où N est la matrice nilpotente de la décomposition $\Delta + N$ de $P^{-1}AP$. En particulier, $v_i = 0$ si et seulement si tous les vecteurs $v_{i,k}$ sont nuls. Ceci est en fait une conséquence de l'invariance des sous-espaces Γ_i par A.

Considérons maintenant une matrice A à coefficients réels et une solution $x(\cdot)$ dans $\mathbb{R}^n$ de $x' = Ax$, dont la condition initiale est $x(0) \in \mathbb{R}^n$. On peut bien entendu considérer A comme une matrice à coefficients complexes et $x(\cdot) = x(\cdot) + i0$ comme la solution dans $\mathbb{C}^n$ de $x' = Ax$ ayant pour condition initiale $x(0) + i0$. L'expression de $x(\cdot)$ est donc donnée par la formule (2.17). Puisque cette solution est réelle, elle est en fait égale à la partie réelle de la formule (2.17), la partie imaginaire devant être nulle. On obtient ainsi la forme générale suivante pour les solutions de $x' = Ax$ dans $\mathbb{R}^n$.

Théorème 2.25. *Soit $A \in M_n(\mathbb{R})$. Toute solution de $x'(t) = Ax(t)$ dans $\mathbb{R}^n$ s'écrit sous la forme*

$$x(t) = \sum_{1\le i\le q} e^{t\alpha_i}\Big(\sum_{0\le k\le m_i-1} t^k\big(\cos(\beta_i t)\, a_{i,k} + \sin(\beta_i t)\, b_{i,k}\big)\Big), \qquad (2.19)$$

où $\alpha_i = \Re e(\lambda_i)$, $\beta_i = \Im m(\lambda_i)$ et les vecteurs $a_{i,k}$, $b_{i,k}$ appartiennent à E_i.

Remarque. Comme dans le théorème 2.24, les vecteurs $a_{i,k}$, $b_{i,k}$ dépendent uniquement de la condition initiale $x(0)$. Si $x(0)$ s'écrit comme $x(0) = u_1 + \cdots + u_q$ dans $\mathbb{R}^n = E_1 \oplus \cdots \oplus E_q$, alors $u_i = 0$ si et seulement si tous les vecteurs $a_{i,k}$ et $b_{i,k}$ sont nuls.

Comportement asymptotique.

Les théorèmes 2.24 et 2.25 donnent toutes les informations que l'on peut souhaiter sur l'équation différentielle, généralisant les résultats de la section 2.2.1 sur les équations scalaires et les systèmes d'équations découplées. On constate en particulier que le comportement quand t tend vers l'infini des solutions $x(t)$ de $x'(t) = Ax(t)$ dépend essentiellement des signes des parties réelles des valeurs propres λ_i de A. Plus précisément, on peut décomposer le comportement des composantes de $x(t)$ sur chaque sous-espace caractéristique (on suppose ici A réelle) :

— si $\Re e(\lambda_i) < 0$ la projection sur E_i de $x(t)$ s'annule quand t tend vers $+\infty$ et croît de façon au moins exponentielle en $-\infty$;
— si $\Re e(\lambda_i) > 0$, c'est l'inverse, la projection sur E_i de $x(t)$ croît de façon au moins exponentielle en $+\infty$ et s'annule quand t tend vers $-\infty$;
— si $\Re e(\lambda_i) = 0$, la composante sur E_i de $x(t)$ croît de façon polynômiale en $\pm\infty$ quand $\dim \Pi_i < p_i$, et est bornée pour $t \in \mathbb{R}$ quand $\dim \Pi_i = p_i$.

Il est commode de regrouper les sous-espaces caractéristiques en fonction du signe de la partie réelle des valeurs propres correspondantes. Nous définissons ainsi, pour $A \in M_n(\mathbb{R})$,

– l'espace *stable* : $$E^s = \Big[\bigoplus_{\Re e(\lambda_i)<0} \Gamma_i\Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)<0} E_i,$$

– l'espace *instable* : $$E^u = \Big[\bigoplus_{\Re e(\lambda_i)>0} \Gamma_i\Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)>0} E_i,$$

– l'espace *indifférent* : $$E^c = \Big[\bigoplus_{\Re e(\lambda_i)=0} \Gamma_i\Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)=0} E_i,$$

si bien que $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$. De même, pour $A \in M_n(\mathbb{C})$, les espaces complexes *stable*, *instable* et *indifférent* sont définis respectivement comme

$$\Gamma^s = \bigoplus_{\Re e(\lambda_i)<0} \Gamma_i, \qquad \Gamma^u = \bigoplus_{\Re e(\lambda_i)>0} \Gamma_i, \qquad \Gamma^c = \bigoplus_{\Re e(\lambda_i)=0} \Gamma_i,$$

et on a la décomposition $\mathbb{C}^n = \Gamma^s \oplus \Gamma^u \oplus \Gamma^c$.

D'après le théorème de décomposition des noyaux, ces espaces ont la particularité d'être invariants par e^{tA} pour tout $t \in \mathbb{R}$: $e^{tA}E^s \subset E^s$, $e^{tA}E^u \subset E^u$, etc... Ce qui entraîne que, si une solution $x(\cdot)$ de $x'(t) = Ax(t)$ vérifie par exemple $x(0) \in E^s$, alors $x(t) \in E^s$ pour tout t ; si $x(0) \in E^c$, alors $x(t) \in E^c$ pour tout t, etc...

Les espaces stable, instable et indifférent correspondent chacun à un certain type de comportement asymptotique des solutions. Nous résumons ces comportements dans le théorème suivant, dont la démonstration est laissée en exercice (utiliser soit la forme générale des solutions, soit directement la réduction de Jordan).

Théorème 2.26. *Soit A une matrice $(n \times n)$ réelle (resp. complexe). Notons $x(\cdot)$ les solutions dans $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) de l'équation différentielle $x'(t) = Ax(t)$. Alors*

— *E^s (resp. Γ^s) est l'ensemble des $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) pour lesquels*

$$\lim_{t \to +\infty} \|x(t)\| = 0;$$

— *E^u (resp. Γ^u) est l'ensemble des $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) pour lesquels*

$$\lim_{t \to -\infty} \|x(t)\| = 0;$$

— *E^c (resp. Γ^c) est l'ensemble des $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) pour lesquels il existe un entier $M \geq 0$ et une constante $C > 0$ tels que, pour $|t|$ suffisamment grand,*

$$C^{-1} \|x(0)\| \leq \|x(t)\| \leq C\,|t|^M \|x(0)\|.$$

En outre, pour $0 < \alpha < \min_{\Re e(\lambda_i) \neq 0} |\Re e(\lambda_i)|$, il existe une constante $C > 0$ telle que :

— *si $x(0) \in E^s$ (ou Γ^s), alors, pour tout $t > 0$ assez grand,*

$$\|x(t)\| \leq Ce^{-\alpha t}\|x(0)\|, \qquad \|x(-t)\| \geq C^{-1}e^{\alpha t}\|x(0)\|;$$

— *si $x(0) \in E^u$ (ou Γ^u), alors, pour tout $t > 0$ assez grand,*

$$\|x(t)\| \geq C^{-1}e^{\alpha t}\|x(0)\|, \qquad \|x(-t)\| \leq Ce^{-\alpha t}\|x(0)\|.$$

Remarque. Dans la caractérisation de E^c, on peut prendre $M = 0$ si et seulement si A est diagonalisable (puisque dans ce cas, pour toute valeur propre, multiplicités algébrique et géométrique coïncident).

Pour obtenir le comportement asymptotique d'une solution particulière $x(\cdot)$, il suffit donc de décomposer sa condition initiale $x(0)$ en $x^s(0) + x^u(0) + x^c(0)$ dans $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$. On sait alors que la décomposition de $x(t)$ est $x(t) = x^s(t) + x^u(t) + x^c(t)$, où

$x^s(\cdot)$ (resp. $x^u(\cdot)$, $x^c(\cdot)$) est la solution de $x'(t) = Ax(t)$ ayant $x^s(0)$ (resp. $x^u(0)$, $x^c(0)$) pour condition initiale. En particulier, si on s'intéresse aux temps positifs, on a les critères suivants :

— si $x^u(0) \neq 0$, alors $\|x(t)\|$ tend vers l'infini quand $t \to +\infty$;
— si $x^c(0) \neq 0$, alors $\|x(t)\|$ ne tend pas vers 0 quand $t \to +\infty$ (mais $\|x(t)\|$ ne tend pas forcément vers l'infini).

Définition 2.11. Une matrice carrée A est dite **Hurwitz** si toutes ses valeurs propres ont une partie réelle strictement négatives (c'est-à-dire $E^s = \mathbb{R}^n$ ou $\Gamma^s = \mathbb{C}^n$).

Corollaire 2.27. *Toutes les solutions de $x'(t) = Ax(t)$ tendent vers 0 quand $t \to +\infty$ si et seulement si A est Hurwitz.*

Quand c'est le cas, on dit que 0 est un *équilibre asymptotiquement stable* de l'équation.

Nous dirons qu'une matrice A est *hyperbolique* si elle n'a aucune valeur propre de partie réelle nulle. Pour une telle matrice, on a $\Gamma^c = \{0\}$, c'est-à-dire que $\mathbb{C}^n = \Gamma^s \oplus \Gamma^u$ et $\mathbb{R}^n = E^s \oplus E^u$ si $A \in M_n(\mathbb{R})$.

La classe des matrices hyperboliques est d'un intérêt particulier puisqu'elle est en fait stable par perturbation.

Théorème 2.28. *Si A est une matrice hyperbolique, il existe un réel $\delta > 0$ (dépendant de A) tel que pour toute matrice F vérifiant*

$$\|F\| \leq \delta,$$

la matrice $A + F$ est encore hyperbolique. L'ensemble des matrices hyperboliques complexes (resp. réelles) est donc ouvert dans $M_n(\mathbb{C})$ (resp. $M_n(\mathbb{R})$).
En outre, les espaces stable et instable dépendent continûment de F (pour $\|F\| \leq \delta$).

*Preuve.

▷ Le fait que si F est assez petite $A+F$ est hyperbolique résulte de la continuité des valeurs propres des matrices : les valeurs propres de $A + F$ sont d'autant plus proches de celles de A que F est petite ; or celles-ci sont dans l'ouvert $\mathbb{C} - \{\Re e = 0\}$. Il en est donc de même de celles de F si elle est assez petite.

□

Remarque. Alors que la dépendance des espaces caractéristiques de $A + F$ n'est *pas* en général continue par rapport à F, celle des espaces stable et instable l'est.

Cas d'une matrice diagonalisable.

Considérons le cas particulier d'une matrice $A \in M_n(\mathbb{R})$ diagonalisable dans $\mathbb{C}$ (on dit aussi *semi-simple*). Comme nous l'avons vu dans la section 2.2.3, ceci signifie que A satisfait les conditions suivantes, qui sont équivalentes entre elles (nous utilisons les notations de la section 2.2.3) :

— il existe une base de $\mathbb{C}^n$ formée de vecteurs propres de A ;
— $\mathbb{C}^n = \Pi_1 \oplus \cdots \oplus \Pi_r$, où $\Pi_i = \ker_{\mathbb{C}}(A - \lambda_i I) \subset \mathbb{C}^n$ est le sous-espace propre associé à λ_i et $\lambda_1, \ldots, \lambda_r$ sont les valeurs propres (complexes) de A ;
— pour toute valeur propre λ_i, le sous-espace caractéristique (complexe) est égal au sous-espace propre : $\Gamma_i = \Pi_i$;
— pour toute valeur propre, les multiplicités géométrique et algébrique coïncident : $\dim \Pi_i = p_i$ pour $i = 1, \ldots, r$;
— dans la forme réduite de Jordan complexe de la matrice A, tous les blocs de Jordan sont des matrices 1×1 : $J_{i,k} = (\lambda_i)$.

Ce cas est en pratique très important, puisque l'ensemble des matrices de $M_n(\mathbb{R})$ diagonalisables dans $\mathbb{C}$ contient un ensemble ouvert et dense dans $M_n(\mathbb{R})$. Autrement dit, être diagonalisable dans $\mathbb{C}$ est une propriété générique sur $M_n(\mathbb{R})$.

Considérons donc une telle matrice A. La forme générale des solutions de $x' = Ax$ donnée par le théorème 2.25 se simplifie : tous les termes polynômiaux en t disparaissent, seuls restent les termes exponentiels et trigonométriques.

Corollaire 2.29. *Soit $A \in M_n(\mathbb{R})$ une matrice diagonalisable dans $\mathbb{C}$. Toute solution de $x'(t) = Ax(t)$ dans $\mathbb{R}^n$ s'écrit sous la forme*

$$x(t) = \sum_{1 \le j \le q} e^{t\alpha_j} \big(\cos(\beta_j t)\, a_j + \sin(\beta_j t)\, b_j \big),$$

où $\alpha_j = \Re e(\lambda_j)$, $\beta_j = \Im m(\lambda_j)$ et les vecteurs a_j et $b_j \in \mathbb{R}^n$ sont tels que $a_j + ib_j$ est un vecteur propre de A associé à λ_j (i.e. $a_j + ib_j \in \Pi_j$).

Les sous-espaces stables, instables et indifférents s'écrivent maintenant en fonction des sous-espaces propres (puisque ceux-ci sont égaux aux sous-espaces caractéristiques complexes) :

$$E^s = \Big[\bigoplus_{\Re e(\lambda_i)<0} \Pi_i \Big] \cap \mathbb{R}^n, \qquad E^u = \Big[\bigoplus_{\Re e(\lambda_i)>0} \Pi_i \Big] \cap \mathbb{R}^n, \qquad E^c = \Big[\bigoplus_{\Re e(\lambda_i)=0} \Pi_i \Big] \cap \mathbb{R}^n.$$

La caractérisation dynamique de ces espaces résulte du théorème 2.26.

Corollaire 2.30. *Soit $A \in M_n(\mathbb{R})$ une matrice diagonalisable dans $\mathbb{C}$. Notons $x(\cdot)$ les solutions dans $\mathbb{R}^n$ de l'équation différentielle $x'(t) = Ax(t)$. Alors*

- *E^s est l'ensemble des conditions initiales $x(0) \in \mathbb{R}^n$ correspondant à des solutions $x(t)$ qui tendent exponentiellement vers 0 quand $t \to +\infty$;*
- *E^u est l'ensemble des conditions initiales $x(0) \in \mathbb{R}^n$ correspondant à des solutions $x(t)$ qui tendent exponentiellement vers 0 quand $t \to -\infty$;*
- *E^c est l'ensemble des conditions initiales $x(0) \in \mathbb{R}^n$ correspondant à des solutions $x(t)$ périodiques, donc bornées sur $\mathbb{R}$.*

La seule différence par rapport au cas général concerne l'espace E^c : alors que dans le cas général le comportement asymptotique des solutions dans E^c était indéterminé (d'où le nom d'espace « indifférent »), on sait ici que toutes les solutions dans E^c sont bornées.

2.3 Stabilité

2.3.1 Équilibres et stabilité

Considérons l'équation différentielle autonome

$$x'(t) = f\big(x(t)\big), \tag{2.20}$$

où le champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ est supposé de classe C^1.

Définition 2.12. On dit qu'un point $x_0 \in \Omega$ est un *équilibre* de (2.20) si la fonction constante $x(\cdot) \equiv x_0$ est solution de (2.20) ou, de façon équivalente, si $f(x_0) = 0$ (vérifier que c'est bien équivalent !).

Autrement dit, $\phi_t(x_0) = x_0$ pour tout $t \in \mathbb{R}$, où ϕ est le flot du champ de vecteurs f (l'intervalle maximal associé à x_0 étant $I_{x_0} = \mathbb{R}$). L'orbite de x_0 est donc réduite à un point : $\mathcal{O}_{x_0} = \{x_0\}$.

Quand l'équation (2.20) modélise l'évolution d'un phénomène physique (mécanique, biologique, écologique, ...), un équilibre correspond bien à la notion habituelle « d'état d'équilibre » : si le système est dans l'état x_0, alors il y reste (et il y a toujours été). En pratique on sait cependant que seuls les états d'équilibre ayant certaines propriétés de stabilité sont significatifs.

Définition 2.13. Nous dirons qu'un équilibre x_0 est *stable* si, pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que

$$\|x - x_0\| < \delta \quad \text{et} \quad t > 0 \qquad \Longrightarrow \qquad \|\phi_t(x) - x_0\| < \epsilon.$$

Ainsi, toute solution proche de x_0 en reste proche.

Remarque. Toute solution dont la condition initiale est dans une boule $B(x_0, \delta)$ reste dans la boule $B(x_0, \epsilon)$, et donc dans un compact de Ω, pour $t > 0$ (on suppose ϵ suffisamment petit pour que $\bar{B}(x_0, \epsilon) \subset \Omega$). D'après la proposition 2.4, ces solutions sont donc définies pour tout $t > 0$.

Définition 2.14. Nous dirons qu'un équilibre x_0 est localement *asymptotiquement stable* (LAS) si il est stable et si il existe un voisinage V de x_0 tel que, pour tout $x \in V$,

$$\lim_{t\to\infty} \phi_t(x) = x_0.$$

Si V ést égal à tout l'espace d'état, on dira que x_0 est globalement *asymptotiquement stable* (GAS).

Dans le cas (LAS), toute solution proche de l'équilibre en reste proche et en plus converge vers lui.

Le cas linéaire

Considérons le cas particulier d'une équation différentielle autonome linéaire

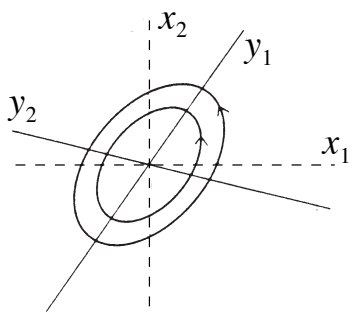
$$x'(t) = Ax(t), \qquad x \in \mathbb{R}^n.$$

L'origine est toujours un équilibre de cette équation (mais il peut y en avoir d'autres : tout élément de $\ker A$ est un équilibre). L'étude réalisée dans la section 2.2.4 permet de caractériser la stabilité de cet équilibre. Par ailleurs, due à la linéarité du système (son homogénéité suffit), il n'y a pas de distinction entre local ou global. En conséquence, dans les énoncés qui suivent, la stabilité, quand elle a lieu, est globale.

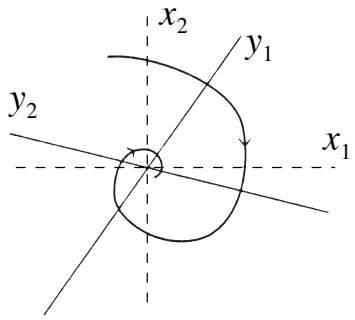
Proposition 2.31.
- *L'origine est un équilibre asymptotiquement stable de $x' = Ax$ si et seulement si toutes les valeurs propres de A sont de partie réelle strictement négative, i.e. $\mathbb{R}^n = E^s$.*
- *Si A a au moins une valeur propre de partie réelle strictement positive, alors l'origine n'est pas un équilibre stable de $x' = Ax$.*

Notons que l'origine peut être un équilibre stable mais non asymptotiquement stable. C'est une situation que l'on rencontre quand A a des valeurs propres de partie réelle nulle, par exemple quand A est antisymétrique (voir proposition 2.13 et la discussion qui suit). On a représenté figure 2.2 des portraits de phase dans $\mathbb{R}^2$ correspondant à une matrice antisymétrique (cas a) et une autre dont les valeurs propres ont une partie réelle < 0 (cas b).

Remarquons enfin que l'on peut donner une condition nécessaire et suffisante de stabilité :



Cas a. Équilibre stable mais non asymptotiquement stable



Cas b. Équilibre asymptotiquement stable

FIGURE 2.2 – Exemples de portraits de phase stables pour $f(x) = Ax$ avec $A \in M_2(\mathbb{R})$.

l'origine est un équilibre stable de $x' = Ax$ si et seulement si toutes les valeurs propres de A sont de partie réelle négative ou nulle et si pour toute valeur propre de partie réelle nulle, les multiplicités algébrique et géométrique coïncident, (c'est-à-dire $\mathbb{R}^n = E^s + E^c$ et $A|_{E^c}$ est diagonalisable dans $\mathbb{C}$).

Le cas affine

Considérons maintenant un champ de vecteurs affine $f(x) = Ax + b$ sur $\mathbb{R}^n$, où $A \in M_n(\mathbb{R})$ est une matrice et $b \in \mathbb{R}^n$ un vecteur. Un équilibre de l'équation

$$x'(t) = Ax(t) + b$$

est un point x_0 qui vérifie $Ax_0 + b = 0$ (noter qu'un tel point n'existe que si $b \in \operatorname{Im} A$). En remplaçant b par $-Ax_0$, on réécrit l'équation différentielle sous la forme

$$\frac{d}{dt}(x(t) - x_0) = A(x(t) - x_0).$$

Ainsi la stabilité et la stabilité asymptotique d'un équilibre de l'équation affine $x'(t) = Ax(t) + b$ sont équivalentes respectivement à celles de l'origine pour l'équation linéaire $y'(t) = Ay(t)$.

2.3.2 La stabilité par la linéarisation

Soit x_0 un équilibre de l'équation différentielle (2.20). Nous allons montrer dans les deux théorèmes suivants que l'étude des valeurs propres de la matrice $Df(x_0)$ permet souvent de caractériser la stabilité de l'équilibre.

Théorème 2.32. *Si toutes les valeurs propres de $Df(x_0)$ sont de partie réelle strictement négative, alors x_0 est un équilibre asymptotiquement stable.*

Remarque. Contrairement au cas des équations linéaires, la condition du théorème est suffisante mais pas nécessaire. Prenons par exemple l'équation $y'(t) = -y^3(t)$ dans $\mathbb{R}$. L'équilibre 0 ne satisfait pas la condition du théorème puisque $Df(0) = 0$. En revanche c'est un équilibre asymptotiquement stable puisque la solution valant $y_0 \neq 0$ en $t = 0$ est

$$y(t) = \frac{\text{signe}(y_0)}{\sqrt{2t + \frac{1}{y_0^2}}}, \qquad t \geq 0,$$

qui est décroissante et converge vers 0 quand $t \to +\infty$.

*Preuve.

▷ Pour simplifier, on se ramène par translation à $x_0 = 0$. D'après l'hypothèse, il existe $\alpha > 0$ tel que $-\alpha$ est strictement supérieur à la partie réelle de toute valeur propre de $Df(0)$. D'après un résultat classique d'algèbre linéaire, il existe alors un produit scalaire $\langle \cdot, \cdot \rangle_\alpha$ sur $\mathbb{R}^n$ tel que

$$\langle Df(0)x, x\rangle_\alpha \leq -\alpha \|x\|_\alpha^2, \qquad \forall x \in \mathbb{R}^n,$$

où $\| \cdot \|_\alpha$ est la norme associé au produit scalaire $\langle \cdot, \cdot \rangle_\alpha$ (le résultat est clair quand $Df(0)$ est diagonalisable, sinon il se montre en utilisant la décomposition de Jordan.
Or, par définition de la différentielle,

$$\langle f(x), x\rangle_\alpha = \langle Df(0)x, x\rangle_\alpha + o(\|x\|_\alpha^2).$$

Ainsi, pour x suffisamment proche de 0, disons $\|x\| < \delta$, on obtient

$$\langle f(x), x\rangle_\alpha \leq -\frac{\alpha}{2}\|x\|_\alpha^2.$$

▷ Prenons maintenant un point $v \neq 0$ vérifiant $\|v\| < \delta$ et notons $x(t) = \phi_t(v)$ la solution issue de v. On peut choisir un temps $t_0 > 0$ suffisamment petit pour que $\|x(t)\| < \delta$ pour tout $t \in [0, t_0]$. La fonction $t \mapsto \|x(t)\|_\alpha$ est dérivable (car $x(t) \neq 0\ \forall t$), et

$$\frac{d}{dt}\|x(t)\|_\alpha = \frac{\langle x'(t), x(t)\rangle_\alpha}{\|x(t)\|_\alpha} = \frac{\langle f(x(t)), x(t)\rangle_\alpha}{\|x(t)\|_\alpha} \leq -\frac{\alpha}{2}\|x(t)\|_\alpha.$$

Ceci implique d'abord que $\|x(t)\|_\alpha$ est décroissante : $x(t)$ reste donc confinée dans le compact $\|x\|_\alpha \leq \|v\|_\alpha$, ce qui entraîne que $x(\cdot)$ est définie pour tout $t > 0$ (proposition 2.4). D'autre part, d'après le lemme de Gronwall,

$$\|x(t)\|_\alpha \leq e^{-\frac{\alpha}{2}t}\|v\|_\alpha.$$

Finalement, on a montré que, si $\|v\| < \delta$, alors $\phi_t(v)$ reste dans $B(0, \delta)$ et tend vers 0, ce qui montre que 0 est un équilibre asymptotiquement stable.

□

Théorème 2.33. *Si x_0 est un équilibre stable, alors toutes les valeurs propres de $Df(x_0)$ sont de partie réelle négative ou nulle.*

On utilisera généralement la contraposée de ce théorème : *si $Df(x_0)$ a au moins une valeur propre de partie réelle strictement positive, alors l'équilibre x_0 n'est pas stable.*

Il est important de noter que les réciproques des théorèmes 2.32 et 2.33 sont fausses, comme le montre l'exemple ci-dessous. La stabilité d'un équilibre n'est donc pas forcément déterminée par le linéarisé. Nous allons ensuite introduire une classe d'équilibres pour lesquels les réciproques des théorèmes 2.32 et 2.33 sont vérifiées.

Exemple. Considérons deux équations différentielles dans $\mathbb{R}^2$,

$$x' = f(x) = \begin{pmatrix} x_2 - x_1(x_1^2 + x_2^2) \\ -x_1 - x_2(x_1^2 + x_2^2) \end{pmatrix} \qquad \text{et} \qquad x' = g(x) = \begin{pmatrix} x_2 + x_1(x_1^2 + x_2^2) \\ -x_1 + x_2(x_1^2 + x_2^2) \end{pmatrix},$$

où $x = (x_1, x_2)$. Ces deux équations ont pour unique équilibre 0. Leurs linéarisés en 0 sont égaux,

$$Df(0) = Dg(0) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

et ont pour valeurs propres $\pm i$, dont la partie réelle est évidemment nulle ! Cependant l'équilibre 0 est asymptotiquement stable dans le premier cas alors qu'il n'est pas stable dans le deuxième.

En effet, posons $\rho(x) = x_1^2 + x_2^2$. Si $x(\cdot)$ est une solution de l'équation $x' = f(x)$, alors

$$\frac{d}{dt}\rho\big(x(t)\big) = 2(x_1 x_1' + x_2 x_2') = -2\rho^2\big(x(t)\big).$$

Ainsi $\rho\big(x(t)\big) = \|x(t)\|^2$ est décroissant et tend vers 0 quand $t \to +\infty$, ce qui implique que 0 est asymptotiquement stable pour l'équation $x' = f(x)$.

De même, si $x(\cdot)$ est une solution de l'équation $x' = g(x)$, on obtient

$$\frac{d}{dt}\rho\big(x(t)\big) = 2\rho^2\big(x(t)\big).$$

Dans ce cas, $\rho\big(x(t)\big) = \|x(t)\|^2$ tend vers l'infini en temps fini (phénomène d'explosion), ce qui implique que l'équilibre 0 n'est pas stable pour l'équation $x' = g(x)$.

Équilibres hyperboliques

Définition 2.15. Un équilibre x_0 est dit *hyperbolique* si toutes les valeurs propres de $Df(x_0)$ ont une partie réelle non nulle.

Les équilibres hyperboliques jouent un rôle important en pratique puisque, comme nous l'avons vu à la fin de la section 2.2.4, la classe des matrices hyperboliques est ouverte et dense dans $M_n(\mathbb{R})$.

D'après les deux théorèmes précédents, la stabilité d'un équilibre hyperbolique x_0 est totalement caractérisée par le signe des parties réelles des valeurs propres de $Df(x_0)$.

Corollaire 2.34. *Un équilibre hyperbolique est soit asymptotiquement stable (si les valeurs propres de $Df(x_0)$ sont toutes de partie réelle négative), soit non stable.*

Ainsi, un équilibre hyperbolique x_0 est stable (resp. asymptotiquement stable) si et seulement si 0 est un équilibre stable (resp. asymptotiquement stable) pour l'équation linéarisée en x_0,

$$y'(t) = Df(x_0) \cdot y(t). \tag{2.21}$$

On a en fait beaucoup plus : les portraits de phase du système et de son linéarisé ont la même allure car ils sont topologiquement équivalents.

Théorème 2.35 (Théorème d'Hartman-Grobmann)**.** *Soit x_0 un équilibre hyperbolique. Notons $\phi_t^L : y \mapsto e^{tDf(x_0)}y$ le flot du linéarisé en x_0. Alors il existe un homéomorphisme $h : V_{x_0} \to V_0$, où V_{x_0} et V_0 sont des voisinages respectivement de x_0 et 0 dans $\mathbb{R}^n$, tel que*

$$\phi_t^L\big(h(x)\big) = h\big(\phi_t(x)\big),$$

partout où ces expressions ont un sens.

2.3.3 Fonctions de Lyapunov

Il existe une autre approche que la linéarisation pour obtenir des résultats de stabilité **qui ne nécessite pas une connaissance explicite du flot**. Commençons par donner un exemple qui illustre l'idée générale.

Champs de gradient

Un *champ de gradient* est un champ de vecteurs de la forme

$$f(x) = -\nabla V(x),$$

où $V : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ est une fonction de classe C^2 (de façon à ce que f soit de classe C^1). Rappelons que $\nabla V(x)$ désigne le *gradient de V en x*, c'est-à-dire l'unique vecteur de $\mathbb{R}^n$ vérifiant

$$DV(x) \cdot v = \langle \nabla V(x), v \rangle, \qquad \forall v \in \mathbb{R}^n.$$

Dans l'espace euclidien $\mathbb{R}^n$, en coordonnées, $\nabla V(x) = (\frac{\partial V}{\partial x_1}(x), \ldots, \frac{\partial V}{\partial x_n}(x))$.

Un équilibre x_0 de ce champ de vecteur est un point critique de V, *i.e.* $\nabla V(x_0) = 0$. Un équilibre peut donc être un minimum local, un maximum local, ou un point selle, mais nous allons voir que seuls les minima locaux peuvent être des équilibres stables.

La dynamique associée à un champ de gradient possède en effet une propriété qui la rend assez simple. Si $x(\cdot)$ est une solution de l'équation différentielle $x'(t) = -\nabla V\big(x(t)\big)$, alors

$$\frac{d}{dt}\Big[V\big(x(t)\big)\Big] = DV\big(x(t)\big) \cdot x'(t) = -\|\nabla V\big(x(t)\big)\|^2, \tag{2.22}$$

pour tout t dans l'intervalle de définition de $x(\cdot)$. Ainsi $V\big(x(t)\big)$ est soit constante, et dans ce cas $x(t) \equiv x_0$ est un point critique, soit strictement décroissante. Intuitivement,

ceci signifie que toute solution tend à se rapprocher d'un minimum, et donc (nous le montrerons plus loin) que :

— si un équilibre x_0 n'est pas un minimum local (*i.e.* x_0 est un maximum local ou un point selle), alors x_0 n'est pas un équilibre stable ;

— si x_0 est un minimum local strict, alors x_0 est un équilibre stable.

Fonctions de Lyapunov

Considérons maintenant une équation différentielle autonome quelconque

$$x'(t) = f\big(x(t)\big), \tag{2.20}$$

associée à un champ de vecteurs $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ de classe C^1. L'exemple des champs de gradient suggère d'introduire la définition suivante.

Définition 2.16. Soient x_0 un équilibre de (2.20), $U \subset \Omega$ un voisinage de x_0 et $L : U \to \mathbb{R}$ une fonction continue. On dit que L est une *fonction de Lyapunov locale pour (2.20) en* x_0 si :

(a) $L(x_0) = 0$ et $L(x) > 0$ pour $x \neq x_0$ (*i.e.* x_0 est un minimum strict de L sur U) ;

(b) la fonction $t \mapsto L\big(\phi_t(x)\big)$ est décroissante.

Si de plus L satisfait

(c) pour $x \neq x_0$, la fonction $t \mapsto L\big(\phi_t(x)\big)$ est strictement décroissante,

on dit que L est une fonction de Lyapunov locale *stricte* pour (2.20) en x_0.

Enfin, si U est égal à tout l'espace d'état, on peut remplacer dans ce qui précède "locale" par "globale".

Quand L est de classe C^1, on peut remplacer les conditions (b) et (c) respectivement par les conditions suivantes, qui sont plus fortes mais plus simples à vérifier que (b) et (c) :

$(b)'$ $\langle \nabla L(x), f(x) \rangle \leq 0$ pour tout $x \in U$;

$(c)'$ $\langle \nabla L(x), f(x) \rangle < 0$ pour tout $x \in U$, $x \neq x_0$.

Une fonction de Lyapunov est donc une sorte de fonction d'énergie qui décrit le long des trajectoires.

Théorème 2.36. *Supposons que l'équation différentielle (2.20) admette L comme fonction de Lyapunov locale en un équilibre x_0. Alors x_0 est un équilibre stable.*

— *(Loc) Si de plus L est stricte, alors x_0 est localement asymptotiquement stable.*

— *(Glob) Si de plus L est globale, stricte et L tend vers l'infini lorsque x tend vers l'infini, alors x_0 est globalement asymptotiquement stable.*

*Preuve.

▷ Soit $L : U \to \mathbb{R}$ la fonction de Lyapunov. Quitte à remplacer U par une boule fermée centrée en x_0, on le suppose compact. Pour tout $\varepsilon > 0$, l'ensemble $U_\alpha = \{x \in U \ : \ L(x) < \alpha\}$ est inclus dans la boule ouverte $B(x_0, \varepsilon)$ pour $\alpha > 0$ suffisamment petit (en effet, sinon il existe une suite de points x_n en dehors de $B(x_0, \varepsilon)$ vérifiant $\lim L(x_n) = 0$; U étant compact, x_n a alors un point d'accumulation $\bar{x} \neq x_0$, qui doit vérifier $L(\bar{x}) = 0$, ce qui est impossible d'après la propriété (a)).
Or, d'après la propriété (b) de L, pour tout $x \in U_\alpha$, la solution $\phi_t(x)$ est confinée dans U_α ; ceci montre la stabilité de l'équilibre x_0.

▷ Supposons maintenant que L est une fonction de Lyapunov locale et stricte. Considérons un point $x \in U_\alpha$ différent de x_0. La fonction $t \mapsto L\big(\phi_t(x)\big)$ étant strictement décroissante et minorée par 0, elle a une limite ℓ quand $t \to +\infty$. D'autre part, U étant compact, il existe une suite t_n, $t_n \to +\infty$, telle que $\phi_{t_n}(x)$ est convergente. Notons $\bar{x}$ la limite de cette dernière suite. Par continuité de L, $\bar{x}$ vérifie :

$$L(\bar{x}) = \lim_{t_n \to \infty} L\big(\phi_{t_n}(x)\big) = \lim_{t \to \infty} L\big(\phi_t(x)\big) = \ell.$$

De plus, pour tout $s > 0$, on a :

$$L\big(\phi_s(\bar{x})\big) = \lim_{t_n \to \infty} L\big(\phi_{s+t_n}(x)\big) = \ell,$$

ce qui montre que $s \mapsto L\big(\phi_s(\bar{x})\big) \equiv L(\bar{x})$ n'est pas décroissante, et donc, d'après la propriété (c) de L, que $\bar{x} = x_0$.
Ainsi, le seul point d'accumulation de $\phi_t(x)$ est x_0, ce qui montre que $\lim_{t\to\infty} \phi_t(x) = x_0$. L'équilibre x_0 est donc asymptotiquement stable.
On adapte sans peine cette preuve au cas (Glob) en remarquant que les ensembles $L_c := \{x \in \Omega \ \ L(x) \leq c\}$, $c > 0$, sont compacts.

□

Remarque. Pour un équilibre asymptotiquement stable x_0, on appelle *bassin d'attraction* l'ensemble des points $x \in \Omega$ tels que $\phi_t(x) \to x_0$ quand $t \to +\infty$. Par définition de la stabilité asymptotique, le bassin d'attraction contient un voisinage de x_0. Une question importante en pratique est de déterminer la taille de ce bassin, voire le bassin lui-même.

Le domaine de définition d'une fonction de Lyapunov stricte, si il en existe une, donne des éléments de réponse à cette question. Supposons par exemple que $\Omega = \mathbb{R}^n$ et que L soit une fonction de Lyapunov vérifiant les hypothèses du cas (Glob). Alors, le bassin d'attraction de x_0 est $\mathbb{R}^n$ tout entier.

Plus généralement, si $L : U \to \mathbb{R}$ une fonction de Lyapunov stricte en x_0 et $P \subset U$ un sous-ensemble fermé de $\mathbb{R}^n$ positivement invariant par le flot (*i.e.* $\phi_t(P) \subset P$ pour tout $t \geq 0$), alors P est inclus dans le bassin d'attraction de x_0. Par exemple, les ensembles $\{x \in U \ : \ L(x) \leq \alpha\}$ pour α suffisamment petit sont fermés et positivement invariants (cf. Définition 2.6), donc inclus dans le bassin d'attraction.

Dans de nombreuses applications, il n'est pas possible d'avoir une fonction de Lyapunov stricte c'est-à-dire vérifiant la condition $(c)'$. On a alors le résultat suivant, appelé *principe d'invariance de Lasalle.*

Théorème 2.37. *Supposons que l'équation différentielle (2.20) admette $L : U \to \mathbb{R}_+$ comme fonction de Lyapunov locale en un équilibre x_0. Notons D_U, le sous-ensemble de U défini par*

$$D_U := \{x \in U, \quad DL(x).f(x) = 0\}.$$

Alors,

- *— (Loc) toutes les trajectoires restant dans U convergent asymptotiquement vers le plus grand ensemble invariant (cf. Définition 2.6) contenu dans D_U.*
- *— (Glob) Si de plus L est globale (i.e. $U = \Omega$) et tend vers l'infini lorsque x tend vers l'infini, toutes les trajectoires sont définies sur $\mathbb{R}_+$ et convergent asymptotiquement vers le plus grand ensemble invariant contenu dans D_U.*

Le principe d'invariance consiste simplement à écrire le système surdéterminé suivant,

$$x' = f(x), \quad D_x L(x) = 0, \quad x \in U,$$

système caractérisant le plus grand ensemble invariant contenu dans l'intersection de U et D_U.

Exemples d'application

a. Champs de gradient Reprenons le cas d'un champ de gradients $f(x) = -\nabla V(x)$. Supposons que x_0 soit un minimum local strict de V, c'est-à-dire l'unique minimum de V sur un voisinage U de x_0. La fonction V restreinte à U est bien une fonction de Lyapunov en x_0, la propriété $(b)'$ étant toujours satisfaite d'après la formule (2.22). Donc x_0 est bien un équilibre stable.

b. Champ de force conservatif Considérons un objet de masse m soumis à une force dérivant d'un potentiel $V(x)$. L'évolution de l'état $x \in \mathbb{R}^n$ de l'objet au cours du temps est régi par le principe fondamental de la dynamique :

$$mx''(t) = -\nabla V\big(x(t)\big),$$

que l'on réécrit

$$\begin{pmatrix} x \\ x' \end{pmatrix}'(t) = \begin{pmatrix} x'(t) \\ -\frac{1}{m}\nabla V\big(x(t)\big) \end{pmatrix}.$$

Autrement dit, (x, x') est solution de l'équation différentielle du 1er ordre dans $\mathbb{R}^{2n}$

$$(x, v)'(t) = f\big((x, v)(t)\big), \qquad \text{où } f(x, v) = \big(v, -\frac{1}{m}\nabla V(x)\big).$$

Un équilibre de cette équation différentielle est un point $(x_0, 0) \in \mathbb{R}^{2n}$ où x_0 est un point critique du potentiel $V(x)$.

Supposons que x_0 soit un minimum local strict de V et cherchons une fonction de Lyapunov. Essayons à partir de l'énergie totale du système :

$$E(x, v) = \frac{1}{2} m\|v\|^2 + V(x).$$

En posant $L(x, v) = E(x, v) - V(x_0)$, on obtient une fonction qui satisfait la propriété (a) d'une fonction de Lyapunov. De plus, comme $\nabla L(x, v) = \big(\nabla V(x), mv\big)$, on obtient

$$\langle \nabla L(x, v), f(x, v) \rangle \equiv 0,$$

(c'est la conservation de l'énergie !), c'est-à-dire la propriété $(b)'$. La fonction L est donc bien une fonction de Lyapunov en $(x_0, 0)$, ce qui montre le résultat bien connu (théorème de Lagrange) :
si l'énergie potentielle $V(x)$ a un minimum local strict en x_0, l'équilibre $(x_0, 0)$ est stable.
Remarques.

- L'équilibre $(x_0, 0)$ ne peut pas être asymptotiquement stable : la fonction de Lyapunov $L(x, v)$ étant constante le long d'une solution $(x, v)(\cdot) \not\equiv (x_0, 0)$, elle ne peut tendre vers 0, ce qui implique que $(x, v)(t)$ ne peut pas tendre vers $(x_0, 0)$.
- L'approche par linéarisation n'aurait pas permis de conclure ici car $(x_0, 0)$ n'est pas un équilibre hyperbolique (exercice : le montrer).

c. Champ de vecteur linéaire Considérons une équation différentielle linéaire

$$x'(t) = Ax(t), \qquad x \in \mathbb{R}^n,$$

et supposons que toutes les valeurs propres de A sont de partie réelle strictement négative. On a vu (proposition 2.31) que dans ce cas l'origine est un équilibre (globalement) asymptotiquement stable. Cherchons une fonction de Lyapunov pour cette équation en 0. Le théorème suivant va nous en fournir.

Théorème 2.38. *Les propositions suivantes sont équivalentes.*

(i) *A est Hurwitz (i.e. toutes ses valeurs propres sont de partie réelle strictement négative) ;*

(ii) *il existe une matrice $P \in M_n(\mathbb{R})$ telle que $P = P^T > 0$ (c'est-à-dire que P est symmétrique réelle et définie positive) et*

$$A^T P + PA < 0;$$

(iii) *Pour toute matrice $Q = Q^T > 0$, il existe une unique matrice $P = P^T > 0$ solution de l'* **équation de Lyapunov**

$$A^T P + PA = -Q.$$

Preuve.

▷ Il suffit de montrer les implications $(ii) => (i)$ et $(i) => (iii)$.
Commençons par $(ii) => (i)$. Soit $P = P^T > 0$ telle que $A^TP + PA = -Q$. Notons $\beta = \frac{\lambda_m(Q)}{\lambda_M(P)} > 0$ avec $\lambda_m(Q)$ et $\lambda_M(P)$ respectivement la plus petite des valeurs propres de Q et la plus grande des valeurs propres de P. Considérons la fonction de Lyapunov $L(x) := x^TPx$. Alors,

$$\dot{L}(x) = 2x^TAPx = x^T(A^TP + PA)x = -x^TQx \leq -\lambda_m(Q)x^Tx \leq -\beta L(x).$$

On en conclut que $L(x(t)) \leq e^{-\beta t}L(x(0))$ (pourquoi ?) et donc (i).
Montrons maintenant $(i) => (iii)$. Supposons que A est Hurwitz. Pour $t \geq 0$, on pose

$$P(t) := \int_0^t e^{sA^T}Qe^{sA}ds.$$

Remarquons que $P(t) > 0$ pour $t > 0$ et $t \mapsto P(t)$ est une fonction strictement croissante. Comme A est Hurwitz, l'intégrale généralisée

$$P = \int_0^\infty e^{sA^T}Qe^{sA}ds,$$

est convergente et on en conclue que $P(\cdot)$ admet P comme limite lorsque t tend vers l'infini. De plus, $P(\cdot)$ vérifie l'EDO suivante

$$\dot{P}(t) = A^TP(t) + P(t)A + Q.$$

La limite P est nécessairement point d'équilibre de cette EDO (pourquoi ?) et donc satisfait l'équation de Lyapunov. Pour l'unicité, on procède comme suit. Soit $M > 0$ une solution de l'équation de Lyapunov. On multiplie alors l'équation de Lyapunov vérifiée par M à gauche par e^{sA^T} et à droite par e^{sA}. On remarque que

$$e^{sA^T}A^TMe^{sA} + e^{sA^T}MAe^{sA} = \frac{d}{ds}(e^{sA^T}Me^{sA}).$$

On intègre entre 0 et l'infini et on obtient $M = P$.

□

d. Critère de Hurwitz Bien qu'on ne puisse pas exprimer en général les racines d'un polynôme de degré $n \geq 5$ en fonction de ses coefficients, il existe des conditions algébriques ne portant **QUE** sur les coefficients d'un polynôme P qui sont nécessaires et suffisantes pour que toute racine de P soit de partie réelle strictement négative (par abus de language, un tel polynôme est dit Hurwitz). L'une de ces conditions est *le critère de Hurwitz*. On considère le polynôme complexe de degré n :

$$P(z) = a_0z^n + a_1z^{n-1} + \cdots + a_{n-1}z + a_n, \qquad (a_0 \neq 0).$$

On pose $a_{n+1} = a_{n+2} = \cdots = a_{2n-1} = 0$. On définit la matrice carrée d'ordre n :

$$H = \begin{pmatrix} a_1 & a_3 & a_5 & \cdots & \cdots & a_{2n-1} \\ a_0 & a_2 & a_4 & \cdots & \cdots & a_{2n-2} \\ 0 & a_1 & a_3 & \cdots & \cdots & a_{2n-3} \\ 0 & a_0 & a_2 & \cdots & \cdots & a_{2n-4} \\ 0 & 0 & a_1 & \cdots & \cdots & a_{2n-5} \\ \vdots & \vdots & \ddots & & & \vdots \\ 0 & 0 & 0 & * & \cdots & a_n \end{pmatrix},$$

où $* = a_0$ ou a_1 selon la parité de n.

Soient $(H_i)_{i\in\{1,\ldots,n\}}$ les mineurs principaux de H, i.e.

$$H_1 = a_1,\ H_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix},\ H_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix},\ \ldots, H_n = \det H.$$

Proposition 2.39. *Si $a_0 > 0$, toute racine de P est de partie réelle strictement négative si et seulement si $H_i > 0$, pour tout $i \in \{1, \ldots, n\}$.*

Remarque. Si $a_0 > 0$, on a :

— Si pour toute racine λ de P, on a $\mathrm{Re}\,\lambda \leq 0$, alors $a_k \geq 0$ et $H_k \geq 0$, pour tout $k \in \{1, \ldots, n\}$.

— Si $n \leq 3$ et si $a_k \geq 0$ et $H_k \geq 0$, pour tout $k \in \{1, 2, 3\}$, alors toute racine λ de P vérifie $\mathrm{Re}\,\lambda \leq 0$.

Remarque. Une *condition nécessaire* de stabilité est donc, si $a_0 > 0$:

$$\forall k \in \{1, \ldots, n\} \quad a_k \geq 0.$$

Mais cette condition n'est pas suffisante (donner un exemple).

e. Nouvelle preuve du théorème 2.32 Considérons un équilibre x_0 de l'équation différentielle

$$x'(t) = f\big(x(t)\big),$$

et supposons que toutes les valeurs propres de $Df(x_0)$ sont de partie réelle strictement négative. Nous allons montrer que l'équation admet une fonction de Lyapunov stricte en x_0, ce qui donnera une nouvelle preuve du théorème 2.32.

Quitte à faire une translation, on suppose $x_0 = 0$ et on choisit $P = P^T > 0$ solution de l'équation de Lyapunov

$$Df(0)^T P + P Df(0) = -I_n.$$

On considère $L(x) := x^T P x$ qui est une fonction de Lyapunov pour l'équation linéaire $y'(t) = Df(0) \cdot y(t))$ (à vérifier). Remarquons d'autre part que

$$f(x) = Df(0) \cdot x + o(\|x\|).$$

En utilisant le paragraphe précédent, on obtient

$$\begin{aligned} \langle \nabla L(x), f(x) \rangle &= \langle \nabla L(x), Df(0) \cdot x \rangle + \langle \nabla L(x), o(\|x\|) \rangle \\ &= -\|x\|^2 + 2 \int_0^\infty \langle e^{sDf(0)} x, e^{sDf(0)} o(\|x\|) \rangle \, ds. \end{aligned}$$

Le terme dans la dernière intégrale est un $o(\|x\|^2)$, donc pour $\|x\|$ suffisamment petit, $x \neq 0$, on obtient

$$\langle \nabla L(x), f(x) \rangle \leq -\frac{1}{2}\|x\|^2 < 0,$$

c'est-à-dire la propriété $(c)'$. La fonction L est donc une fonction de Lyapunov stricte en $x_0 = 0$, ce qui montre que cet équilibre est asymptotiquement stable.

Chapitre 3

Commandabilité et observabilité des systèmes linéaires

3.1 Systèmes de commande

Un large pan de l'automatique est basé sur les équations différentielles : c'est l'approche par représentation d'état des systèmes à temps continu déterministes (les systèmes stochastiques s'appuyant sur les équations différentielles stochastiques). La situation est la suivante : on considère un système physique (par exemple un satellite, une voiture,...), décrit par son état $x(t)$ au temps t (par exemple la position et la vitesse), sur lequel on peut agir à tout moment au moyen d'une *commande ou contrôle* u (par exemple la poussée des moteurs dans le cas du satellite). On représente l'état par un vecteur de $\mathbb{R}^n$, la commande par un vecteur de $\mathbb{R}^m$, et on modélise l'évolution du vecteur $x(t)$ au cours du temps par un *système de commande* (ou équation différentielle commandée)

$$(\Sigma) \ : \qquad x'(t) = f(t, x(t), u(t)), \qquad t \in [0, \tau],$$

où τ est un temps positif.

Que signifie cette expression ? La fonction $u(t)$, $t \in [0, \tau]$, appelée *loi de commande*, est notre moyen d'action sur le système (Σ) : on va la choisir en fonction des objectifs à réaliser. À une loi de commande $u(\cdot)$ est associée une équation différentielle ordinaire

$$(\Sigma_u) \ : \qquad x'(t) = f_u(t, x(t)), \qquad t \in [0, \tau],$$

où on a noté $f_u(t, x) = f(t, x, u(t))$. Ainsi, une fonction $x(\cdot)$ est solution du système (Σ) si il existe une loi de commande $u(\cdot)$ telle que $x(\cdot)$ est solution de l'équation différentielle (Σ_u).

Les principales questions que l'on est amené à se poser à propos du système (Σ) sont les suivantes.

Commandabilité Étant donnés un état de départ $x_0 \in \mathbb{R}^n$, un état cible $v \in \mathbb{R}^n$ et un temps $t = \tau > 0$, est-il possible de trouver une commande $u(\cdot)$ qui amène le

système initialement en $x(0)$ à $t = 0$ en l'état v au temps $t = \tau$? On formule aussi la question précédente comme suit : est-il possible de *commander* le système de x_0 à v en temps τ ?

Planification de trajectoires A la question structurelle précédente, correspond le problème plus concret d'établir une procédure effective qui associe, à une paire d'états $x_0, v \in \mathbb{R}^n$ et un temps τ, une commande $u(\cdot)$ qui amène le système de $x(0)$ à v en temps $t = \tau$.

Stabilisation Est-il possible de construire une commande $u(\cdot)$ qui *stabilise asymptotiquement* le système (Σ) autour d'un équilibre x_0, c'est-à-dire telle que, pour toutes conditions initiales $x(0)$, on ait

$$\lim_{t \to +\infty} x(t) = x_0.$$

Observabilité Afin de réaliser un objectif de commande (planification de trajectoire, stabilisation, etc...) et donc de choisir en conséquence une loi de commande appropriée, l'opérateur dispose d'une certaine information sur l'état du système x à l'instant t, celle-ci étant obtenue par le biais de mesures. Cependant, il n'est pas possible en général de mesurer (on dit *observer* en automatique) directement l'état $x(t)$ mais seulement une fonction $y(t)$ de l'état et de la commande :

$$y(t) = g(x(t), u(t), t).$$

Il s'agit alors de "reconstruire" l'état $x(\cdot)$ à partir de la *sortie* $y(\cdot)$. La question d'observabilité est alors la suivante : la connaissance de $y(t)$ et de $u(t)$ pour tout $t \in [0, \tau]$ permet-elle de déterminer l'état $x(\cdot)$ pour tout $t \in [0, \tau]$ (ou, ce qui est équivalent, l'état initial $x(0)$) ?

Les techniques que nous avons introduites lors de l'étude des équations différentielles linéaires autonomes vont nous permettre de répondre à ces questions dans le cadre de l'*automatique linéaire autonome* (ou stationnaire, ou encore invariante par rapport au temps).

Dans tout ce chapitre, nous supposerons donc que le système (Σ) est linéaire autonome (par rapport à (x, u)), c'est-à-dire de la forme :

$$(\Sigma) \ : \qquad x'(t) = Ax(t) + Bu(t), \qquad t \in [0, \tau], \tag{3.1}$$

où A est une matrice (carrée) de $M_n(\mathbb{R})$ et B une matrice (pas nécessairement carrée) de $M_{n,m}(\mathbb{R})$, avec n, m entiers positifs. Si $m = 1$, le système est dit *mono-entrée* et sinon *multi-entrée*.

Par ailleurs, lorsque nous aborderons les questions d'observabilité, nous supposerons que la sortie y est un vecteur de $\mathbb{R}^p$, p entier positif et que y est aussi linéaire par rapport à (x, u), plus simplement égale à

$$y(t) = Cx(t), \qquad t \in [0, \tau], \tag{3.2}$$

avec C une matrice $M_{p,n}(\mathbb{R})$. Dans ce cas, le système de commande linéaire autonomes sera défini par les deux équations (3.1) et (3.2).

Les lois de commande $u(\cdot)$ seront supposées *continues par morceaux*, définies sur l'intervalle $[0, \tau]$ et à valeurs dans $\mathbb{R}^m$. Nous aurions pu faire d'autres hypothèses sur la fonction $u(\cdot)$, la supposer seulement mesurable par exemple ou encore imposer que ses valeurs soient bornées (ce qui est raisonnable si u représente une force). Ces hypothèses interviennent naturellement dans des problèmes d'automatique, mais nous n'en parlerons pas ici : notre but est de montrer dans un cadre aussi simple que possible comment les techniques d'équations différentielles s'appliquent à ces problèmes.

Le point de départ de l'étude des systèmes linéaires autonomes est la formule dite *de variation de la constante* :

Proposition 3.1. *Soient $u(\cdot)$ une commande et $x_0 \in \mathbb{R}^n$. L'unique solution de $x'(t) = Ax(t) + Bu(t)$ valant x_0 à l'instant $t = 0$ est*

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)ds.$$

Preuve.

▷ Supposons qu'il existe une solution $x(\cdot)$ de $x'(t) = Ax(t) + Bu(t)$ telle que $x(0) = x_0$. Posons $y(t) = e^{-tA}x(t)$ et dérivons $y(\cdot)$ par rapport à t :

$$y'(t) = -Ae^{-tA}x(t) + e^{-tA}\big(Ax(t) + Bu(t)\big) = e^{-tA}Bu(t).$$

En intégrant entre 0 et t, on obtient

$$y(t) = y(0) + \int_0^t e^{-sA}Bu(s)ds,$$

et donc la conclusion, puisque $x(t) = e^{tA}y(t)$ et $y(0) = x(0) = x_0$.

□

Remarque. Notons en particulier que, si $x(0) = 0$,

$$x(t) = \int_0^t e^{(t-s)A}Bu(s)ds, \tag{3.3}$$

et que cette expression dépend linéairement de la loi de commande $u(\cdot)$.

Remarque. Pour ce qui est de l'observabilité des systèmes linéaires, la formule de la proposition 3.1 montre que la "reconstruction" d'une trajectoire $x(\cdot)$ avec la seule connaissance d'une sortie $y(\cdot)$ (et bien sur de la commande $u(\cdot)$) consiste en fait à retrouver la condition initiale x_0.

3.2 Commandabilité

Soit (Σ) le système de commande linéaire autonome (3.1). Étant donné $x_0 \in \mathbb{R}^n$, on dit qu'un état $v \in \mathbb{R}^n$ est *atteignable en temps* τ à partir de x_0 (par le système (Σ)) si il existe une loi de commande $u : [0, \tau] \to \mathbb{R}^m$ telle que $x(\tau) = v$, $x(\cdot)$ étant la solution de (Σ_u) satisfaisant $x(0) = x_0$. On note $\mathcal{A}(\tau, x_0)$ l'ensemble des états atteignables à partir de x_0 en temps τ, c'est-à-dire

$$\mathcal{A}(\tau, x_0) := \left\{ x(\tau) \ : \ \begin{array}{c} x(\cdot) \text{ solution de } (\Sigma) \\ \text{t.q. } x(0) = x_0 \end{array} \right\}.$$

Il résulte de la remarque ci-dessus et de la proposition 3.1 que l'ensemble $\mathcal{A}(\tau, 0)$ est un espace vectoriel, et que l'ensemble $\mathcal{A}(\tau, x_0)$ est l'espace affine $e^{\tau A}x_0 + \mathcal{A}(\tau, 0)$. L'ensemble des points atteignables à partir de x_0 est donc complètement caractérisé par l'ensemble $\mathcal{A}_\tau := \mathcal{A}(\tau, 0)$.

Définition 3.1. On dit que le système (Σ) est *commandable en temps* τ si $\mathcal{A}_\tau = \mathbb{R}^n$, ou, de façon équivalente, si tout état de $\mathbb{R}^n$ est atteignable en temps τ à partir de n'importe quel autre.

Nous allons chercher maintenant à caractériser algébriquement la commandabilité. Ceci passe par la détermination de l'ensemble $\mathcal{A}_\tau$.

Théorème 3.2. *L'espace $\mathcal{A}_\tau$ est égal à l'image de la matrice $(n \times nm)$*

$$\mathcal{C}(A, B) := \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix},$$

dite matrice de commandabilité.

Remarque. L'image de $\mathcal{C}(A, B)$ est l'espace vectoriel $\mathcal{R}(A, B) \subset \mathbb{R}^n$ engendré par les $A^i Bz$, $i \in \{0, \ldots, n-1\}$, $z \in \mathbb{R}^m$:

$$\mathcal{R}(A, B) = \text{Vect}\{A^i Bz \ : \ i = 0, \ldots, n-1, \ z \in \mathbb{R}^m\}.$$

La première conséquence de ce résultat est que $\mathcal{A}_\tau$ est indépendant de τ. Notons que ce ne serait évidemment pas le cas si nous avions choisi des commandes bornées. La deuxième conséquence est que la dimension de $\mathcal{A}_\tau$ est égale au rang de la matrice de commandabilité. On obtient ainsi un critère de commandabilité algébrique, et donc en général facile à vérifier.

Corollaire 3.3 (Critère de commandabilité de Kalman)**.** *Le système (Σ) est commandable si et seulement si la matrice de commandabilité $\mathcal{C}(A, B)$ est de rang n.*

Preuve.

▷ du Théorème 3.2 Montrons déjà que $\mathcal{A}_\tau \subset \mathcal{R}(A,B)$. Pour cela observons que, par définition, si v appartient à $\mathcal{A}_\tau$ il existe $u : [0,\tau] \to \mathbb{R}^m$ continue par morceaux telle que

$$v = \int_0^\tau e^{(\tau-s)A} B u(s) ds.$$

Le théorème de Cayley-Hamilton nous enseigne que le polynôme caractéristique de A annule A. Or ce polynôme est un polynôme normalisé (*i.e.* dont le coefficient de plus haut degré égale 1) de degré n. Ainsi A^n est combinaison linéaire de $I, \dots, A^{n-1}$ et par conséquent pour *tout* entier $i \geq 0$, A^i est combinaison linéaire de $I, \dots, A^{n-1}$. Par conséquent pour tout $i \geq 0$, A^i laisse invariant l'espace vectoriel

$$\mathcal{R}(A,B) = \mathrm{Vect}\{A^i B z \ : \ i = 0, \dots, n-1, \ z \in \mathbb{R}^m\}.$$

Mais pour tout $s \in [0,\tau]$, l'exponentielle $e^{(\tau-s)A}$ admet le développement

$$e^{(\tau-s)A} = I + (\tau-s)A + \cdots + \frac{(\tau-s)^k A^k}{k!} + \cdots,$$

et par conséquent $e^{(\tau-s)A}$ laisse également invariant l'espace $\mathcal{R}(A,B)$. Nous avons donc montré que $e^{(\tau-s)A} B u(s) \in \mathcal{R}(A,B)$ pour tout $s \in [0,\tau]$ et par conséquent

$$\int_0^\tau e^{(\tau-s)A} B u(s) ds \in \mathcal{R}(A,B).$$

Ainsi $\mathcal{A}_\tau \subset \mathcal{R}(A,B)$.

▷ Montrons l'inclusion réciproque. Il suffit pour cela de démontrer $\mathcal{A}_\tau^\perp \subset \mathcal{R}(A,B)^\perp$. Soit donc $w \in \mathbb{R}^n$ orthogonal à $\mathcal{A}_\tau$; le vecteur w est ainsi orthogonal à l'état $\tilde{w}$ que l'on peut atteindre au temps τ par la commande

$$u(t) = B^T (e^{(\tau-t)A})^T w.$$

La formule (3.3) montre que

$$\tilde{w} = \int_0^\tau e^{(\tau-s)A} B B^T (e^{(\tau-s)A})^T w \, ds,$$

et par conséquent, puisque $\langle \tilde{w}, w \rangle = 0$ on obtient

$$0 = \langle w, \int_0^\tau e^{(\tau-s)A} B B^T (e^{(\tau-s)A})^T w \, ds \rangle = \int_0^\tau \Big((e^{(\tau-s)A} B)^T w \Big)^T \Big((e^{(\tau-s)A} B)^T w \Big) ds,$$

ce qui est équivalent à

$$\forall s \in [0,\tau], \ (e^{(\tau-s)A} B)^T w = 0.$$

Dérivons une fois, puis deux fois,... cette égalité par rapport à t : il vient successivement

$$(e^{(\tau-s)A} AB)^T w = 0, \quad (e^{(\tau-s)A} A^2 B)^T w = 0, \quad \dots \quad (e^{(\tau-s)A} A^{n-1} B)^T w = 0,$$

soit, pour $s = \tau$,

$$B^T w = 0, \qquad \dots \qquad (A^{n-1}B)^T w = 0.$$

Ceci implique que, pour tout $j \in \{0, \dots, n-1\}$ et tout $z \in \mathbb{R}^m$,

$$0 = \langle z, (A^j B)^T w \rangle = \langle A^j B z, w \rangle,$$

c'est-à-dire $w \in \mathcal{R}(A,B)^\perp$. L'inclusion $\mathcal{R}(A,B) \subset \mathcal{A}_\tau$ est donc démontrée.

□

Nous allons maintenant décrire ce qui se passe lorsque le rang r de la matrice de commandabilité $\mathcal{C}(A,B)$ est quelconque. Pour cela, nous étudions l'effet d'un changement de coordonnées linéaire sur la commandabilité d'un système.

Définition 3.2. Les systèmes de commande linéaires $\dot{x}_1 = A_1 x_1 + B_1 u_1$ et $\dot{x}_2 = A_2 x_2 + B_2 u_2$ sont dits *linéairement équivalents* s'il existe $P \in GL_n(\mathbb{R})$ tel que $A_2 = P A_1 P^{-1}$ et $B_2 = P B_1$.

Remarque. On a alors $x_2 = P x_1$.

Proposition 3.4. *La propriété de Kalman est intrinsèque, i.e.*

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

et donc le rang de la matrice de Kalman est invariant par équivalence linéaire.

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

Théorème 3.5 (Décomposition de Kalman)**.** *La paire (A,B) est linéairement équivalente à une paire (A',B') de la forme*

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

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

Remarque. La décomposition précédente est aussi appelée décomposition du système en parties commandable et non commandable. Les valeurs propres de A sont appelées *pôles* du système, commandables pour les valeurs propres de A'_1 et non commandables pour celles de A'_3.

Preuve.

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

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

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

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

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

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

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

□

Dans le cas mono-entrée, il existe un changement de coordonnée très utile.

Théorème 3.6 (Forme compagne)**.** *Si $m = 1$ et si la paire (A, B) est contrôlable, alors elle est linéairement équivalente à la paire $(\tilde{A}, \tilde{B})$, où*

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

La matrice $\tilde{A}$ et la paire $(\tilde{A}, \tilde{B})$ sont respectivement appelées forme compagne de la matrice A et forme canonique commandable de la paire (A, B).

Remarque. Si une matrice carrée est sous forme compagne, alors les coefficients de son polynôme caractéristique se lisent sur la dernière ligne de la matrice.

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

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

Preuve.

▷ Considérons $P_A(X)$ le polynôme caractéristique de A

$$P_A(X) := X^n + a_1 X^{n_1} + \cdots + a_n.$$

Le changement de coordonnées sera donné par une matrice $n \times n$ inversible F de vecteur colonne $(f_1, \cdots, f_n)$. Ces vecteurs sont définis récursivement comme suit,

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

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

$$\begin{aligned} \text{Vect } \{f_n\} &= \text{Vect } \{b\}, \\ \text{Vect } \{f_n, f_{n-1}\} &= \text{Vect } \{b, Ab\}, \\ &\vdots \\ \text{Vect } \{f_n, \ldots, f_1\} &= \text{Vect } \{b, \ldots, A^{n-1}b\} = \mathbb{R}^n. \end{aligned}$$

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

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

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

□

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

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

les matrices $\tilde{A}_i$ étant des matrices sous forme compagne ; par ailleurs, il existe une matrice $G \in \mathcal{M}_{m,s}(\mathbb{R})$ telle que :

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

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

3.3 Planification de trajectoires

Nous allons maintenant proposer une stratégie pour résoudre le problème de planification de trajectoires pour un système de commande linéaire autonome commandable. Pour cela, nous allons effectuer des transformations linéaires sur le système afin de l'amener en une forme particulière simple et pour laquelle la planification de trajectoire sera aisée.

3.3.1 Exemple

Effectuons la planification de trajectoires pour le système commandé de dimension un, $\dot{x} = u$, évidemment commandable. Etant donnés deux réels arbitraires, x_0, x_1, il s'agit de trouver une commande u définie sur $[0, 1]$ telle que

$$x_1 = x_0 + \int_0^1 u(t)dt,$$

et donc plus simplement, de trouver pour tout $x \in \mathbb{R}$, une fonction u telle que $x = \int_0^1 u(t)dt$. Une fonction polynômiale (constante) suffit ici. La généralisation au système commandé $x^{(n)} = u$, avec x, u réels, est immédiate.

3.3.2 Forme de Brunovsky

Dans le cas général du système $\dot{x} = Ax + Bu$, l'idée est de transformer linéairement le système de telle sorte à le faire "ressembler" le plus possible à $x^{(n)} = u$.

Les transformations linéaires considérées ci-dessous sont plus générales que de simples changements de coordonées.

Définition 3.3 (bouclage statique régulier)**.** Soit $x_1 = Mx_2$, le changement de coordonnées défini sur $\mathbb{R}^n$ par M matrice carrée inversible $n \times n$. Un bouclage statique régulier est défini par $u_1 = Kx_2 + Nu_2$ avec N ne matrice carrée inversible $m \times m$ et K une matrice $m \times n$. C'est un changement de variables sur la commande paramétré par l'état. De manière matricielle, il s'écrit

$$\begin{pmatrix} x_1 \\ u_1 \end{pmatrix} := \begin{pmatrix} M & 0 \\ K & N \end{pmatrix} \begin{pmatrix} x_2 \\ u_2 \end{pmatrix}.$$

Les systèmes de commande linéaires $\dot{x}_1 = A_1x_1 + B_1u_1$ et $\dot{x}_2 = A_2x_2 + B_2u_2$ sont dits *équivalents par changement d'état avec bouclage statique régulier* s'il existe $M \in GL_n(\mathbb{R})$, $N \in GL_m(\mathbb{R})$ et $K \in M_{m,n}(\mathbb{R})$, tels que $A_2 = M^{-1}A_1M + M^{-1}B_1K$ et $B_2 = M^{-1}B_1N$.

Théorème 3.7 (Forme de Brunovsky d'un système commandable)**.** *Soit $\dot{x} = Ax + Bu$ un système linéaire commandable avec $x \in \mathbb{R}^n$ et $u \in \mathbb{R}^m$. Alors il existe un changement d'état avec bouclage statique régulier donné par $x = Mz$ et $u = Kz+Nv$ pour lequel il existe m coordonnées de z, notées $y_1, \cdots, y_m$ telles que*

(i) $z = \left(y_1, \cdots, y_1^{(\alpha_1-1)}, \cdots, y_m, \cdots, y_m^{(\alpha_m-1)}\right)^T$;

(ii) $\dot{x} = Ax + Bu$ *devient*

$$y_1^{(\alpha_1)} = v_1, \quad \cdots \quad y_m^{(\alpha_m)} = v_m,$$

où $y_l^{(j)}$ désigne la dérivée temporelle d'ordre j de la fonction scalaire y_l, $1 \leq l \leq m$.

Les coordonnées $y_1, \cdots, y_m$ sont appelées **sorties de Brunovsky** *du système.*

Ecrivons la forme de Brunovsky d'un système mono-entrée, c-à-d avec une seule commande scalaire ($m = 1$). Dans ce cas, $B = b$ est un vecteur colonne. Il existe donc un changement d'état avec bouclage statique régulier qui transforme $\dot{x} = Ax+bu$ en $\dot{z}_i = z_{i+1}$ pour $1 \leq i \leq n$ et $\dot{z}_n = v$. La sortie de Brunovsky est alors $y = z_1$ et le système commandé s'écrit simplement $y^{(n)} = v$.

3.3.3 Application à la planification de trajectoires

Supposons que le système de commande $\dot{x} = Ax+Bu$ soit commandable et sous forme de Brunovsky. Grâce à la structure bloc-diagonale, le problème de planification en temps $T > 0$ se décompose en m problèmes à une seule commande : pour $1 \leq i \leq m$,

$$\text{Aller de } \left(y_i(0), \cdots, y_i^{(\alpha_i-1)}(0)\right)^T \text{ à } \left(y_i(T), \cdots, y_i^{(\alpha_i-1)}(T)\right)^T \text{ le long de } y_i^{(\alpha_i)} = v_i.$$

Le $i$ème problème de planification défini ci-dessus se résout simplement puisque les conditions sur les points initial et final représentent $2\alpha_i$ contraintes à satisfaire. Par exemple, il suffit de choisir v_i comme un polynôme de degré $2\alpha_i - 1$ et de déterminer ses coefficients.

3.3.4 Preuve du théorème 3.7 pour le cas mono-entrée

Dans le reste de ce paragraphe, on supposera $m = 1$ et donc $B = b$ vecteur colonne. On note $P_A(X) = X^n + a_1X^{n_1} + \cdots + a_n$, le polynôme caractéristique de A et on suppose la matrice $n \times n$ inversible F met la paire (A, b) sous forme compagne, c-à-d que $(F^{-1}AF, F^{-1}b) = (\tilde{A}, \tilde{B})$, cf. Théorème 3.6. Si on a $Fy = x$, alors $\dot{y} = F^{-1}AFy + F^{-1}b$ avec

$$\dot{y}_1 = y_2, \ \cdots \ \dot{y}_{n-1} = y_n,$$

et $\dot{y}_n = -a_n y_1 - \cdots - a_1 y_n + u$. Si on définit une nouvelle commande $v := \dot{y}_n$, on a bien un bouclage statique régulier sur la commande u. Le système est donc sous forme de Brunovsky avec y_1 comme sortie de Brunovsky.

3.4 Stabilisation

Utiliser des lois de commande $u(t)$ dépendant en général du temps s'appelle *commander en boucle ouverte* : la loi de commande est fixée au départ, à $t = 0$, et est appliquée indépendamment du comportement du système pour $t > 0$. Les limitations de ce type de lois de commande sont assez évidentes : la moindre erreur sur les données (la condition initiale par exemple) ne pourra pas être prise en compte. Par exemple une commande en boucle ouverte sur une voiture donnerait ceci : pour suivre une ligne droite, positionnez vos roues dans l'axe, tenez bien votre volant, et fermez les yeux...

Pour réguler le système, il faut faire appel à un autre type de loi de commande $u(t) = K(t, x(t))$, dite *commande en boucle fermée* ou *par retour d'état* : à tout instant t, on tient compte de l'état du système à cet instant pour déterminer la commande. Notez que ces lois ont aussi leur inconvénient : elles nécessitent la connaissance (quasi instantanée) de l'état $x(t)$, ce qui peut être impossible, ou très coûteux.

Dans le problème de stabilisation, le but est de construire une loi de commande par retour d'état qui amène le système à l'origine, quel que soit le point de départ. Nous allons chercher le retour d'état sous la forme d'une fonction linéaire indépendante du temps, c'est-à-dire $u(t) = Kx(t)$ avec $K \in M_{m,n}(\mathbb{R})$ (une telle loi est appelée *loi proportionnelle*).

Définition 3.4. Un système de commande (Σ) est *asymptotiquement stabilisable* par retour d'état proportionnel s'il existe une loi de commande $u(t) = Kx(t)$, avec $K \in M_{m,n}(\mathbb{R})$, telle que l'équation (Σ_u) soit asymptotiquement stable, c'est-à-dire que, pour toute condition initiale $x(0)$, la solution $x(t)$ de (Σ_u) tende vers 0 quand $t \to +\infty$.

Noter que l'intervalle de temps considéré est maintenant infini, *i.e.* $\tau = +\infty$.

Nous nous plaçons comme précédemment dans le cadre des systèmes linéaires autonomes c-à-d que (Σ) est de la forme (3.1). Dans ce cas, pour un retour d'état proportionnel $u(t) = Kx(t)$, l'équation différentielle (Σ_u) s'écrit

$$x'(t) = Ax(t) + BKx(t) = (A + BK)\, x(t).$$

Or nous savons (corollaire 2.27) qu'une telle équation différentielle est asymptotiquement stable si et seulement si la matrice $A + BK$ a toutes ses valeurs propres de parties réelles strictement négatives. Le problème à résoudre est donc le suivant : existe-t-il $K \in M_{m,n}(\mathbb{R})$ telle que la matrice $A + BK$ satisfasse cette condition ?

Le but de cette section est de répondre à cette question. Tout d'abord, remarquons qu'un changement de coordonnées permet de remplacer (A, B) par toute paire qui lui est linéairement équivalente. En effet, considérons $\tilde{A} = P^{-1}AP$ et $\tilde{B} = P^{-1}B$ avec P la

matrice de changement de coordonnées $x = Py$ avec P matrice $n \times n$ inversible. On a alors $\tilde{K} = KP$ et $P^{-1}(A + BK)P = \tilde{A} + \tilde{B}\tilde{K}$.

Sans perte en généralité, nous pouvons donc supposer que (A, B) est décomposée en parties commandable et non commandable, cf. Théorème 3.5 :

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

avec $r \leq n$ le rang de $\mathcal{C}(A, B)$, $A_1 \in \mathcal{M}_r(\mathbb{R})$, $B_1 \in \mathcal{M}_{r,m}(\mathbb{R})$ et (A_1, B_1) commandable. Soit $K = (K_1 \;\; K_2)$ avec $K_1 \in \mathcal{M}_{m,r}(\mathbb{R})$ et $K_2 \in \mathcal{M}_{m,(n-r)}(\mathbb{R})$. On a donc

$$A + BK = \begin{pmatrix} A_1 + B_1K_1 & A_2 + B_1K_2 \\ 0 & A_3 \end{pmatrix}.$$

Donc, pour les polynômes caractéristiques, on a

$$P_{A+BK}(X) = P_{A_1+B_1K_1}(X)P_{A_3}(X).$$

Il faut donc nécessairement que les pôles non commandables (cf. Remarque 3.2) soient de partie réelle négative. On va voir que c'est aussi une condition suffisante pour stabiliser (Σ). Cela suggère la définition suivante.

Définition 3.5. Une paire de matrices (A, B) avec $A \in M_n(\mathbb{R})$ et $B \in M_{n,m}(\mathbb{R})$ est dite *stabilisable* si les pôles de sa partie non commandable sont de partie réelle négative.

Un polynôme P unitaire de $\mathbb{R}_n[X]$ est *assignable* pour (A, B) s'il existe $F \in M_{m,n}(\mathbb{R})$ tel que $P_{A+BF}(X) = P(X)$.

Le résultat principal de cette section est le suivant.

Théorème 3.8 (Théorème du placement de pôles - pole shifting theorem)**.** *Soit (A, B) une paire de matrices avec $A \in M_n(\mathbb{R})$, $B \in M_{n,m}(\mathbb{R})$ et r le rang de $\mathcal{C}(A, B)$. Les polynômes de $\mathbb{R}_n[X]$ assignables pour (A, B) sont exactement de la forme suivante*

$$P_{A+BF}(X) = Q(X)P_{nc}(X),$$

avec $Q(X)$ polynôme unitaire quelconque de $\mathbb{R}_r[X]$ et $P_{nc}(X)$ polynôme caractéristique de la partie non commandable de A. En particulier, (A, B) est commandable si et seulement si tout polynôme unitaire de $\mathbb{R}_n[X]$ est assignable pour (A, B).

Preuve.

▷ Si deux paires de matrices sont linéairement équivalentes, alors on peut leur assigner les mêmes polynômes. On peut donc supposer (A, B) décomposée selon Kalman et, d'après ce qui précède, il suffit désormais de se limiter au cas (A, B) commandable.

Faisons d'abord la démonstration dans le cas $m = 1$ (on se ramènera ensuite à ce cas). Par le théorème 3.6, le système est linéairement équivalent à

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

Posons alors $K = (k_1 \ \cdots \ k_n)$ et $u = Kx$. On a :

$$A + BK = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ k_1 - a_n & k_2 - a_{n-1} & \cdots & k_n - a_1 \end{pmatrix},$$

et donc

$$\chi_{A+BK}(X) = X^n + (a_1 - k_n)X^{n-1} + \cdots + (a_n - k_1).$$

Donc, pour tout polynôme $P(X) = X^n + \alpha_1 X^{n-1} + \cdots + \alpha_n$, il suffit de choisir $k_1 = a_n - \alpha_n, \ldots, k_n = a_1 - \alpha_1$.
Dans le cas général où $m \geq 1$, montrons le lemme fondamental suivant dû à Ackerman.

Lemme 3.9. *Si la paire (A, B) est comandable, alors il existe $y \in \mathbb{R}^m$ et $C \in \mathcal{M}_{m,n}(\mathbb{R})$ tels que la paire $(A + BC, By)$ est aussi commandable.*

D'après ce lemme, pour tout polynôme P unitaire de degré n, il existe $K_1 \in \mathcal{M}_{1,n}(\mathbb{R})$ tel que $\chi_{A+BC+ByK_1} = P$, et donc en posant $K = C + yK_1 \in \mathcal{M}_{m,n}(\mathbb{R})$, on a $\chi_{A+BK} = P$, ce qui prouve le théorème.

Preuve.

▷ [Preuve du lemme] Soit $y \in \mathbb{R}^m$ tel que $By \neq 0$. On pose $x_1 = By$. On a le fait suivant :
Fait 1 : Il existe $x_2 \in Ax_1 + \text{Im } B$ (et donc il existe $y_1 \in \mathbb{R}^m$ tel que $x_2 = Ax_1 + By_1$) tel que dim Vect$\{x_1, x_2\} = 2$.
En effet sinon, on a $Ax_1 + \text{Im } B \subset \mathbb{R}x_1$, donc $Ax_1 \in \mathbb{R}x_1$ et Im $B \subset \mathbb{R}x_1$. D'où

$$\text{Im } AB = A\text{Im } B \subset \mathbb{R}Ax_1 \subset \mathbb{R}x_1,$$

et par récurrence immédiate :

$$\forall k \in \mathbb{N} \quad \text{Im } A^k B \subset \mathbb{R}x_1.$$

On en déduit que

$$\text{Im } (B, AB, \ldots, A^{n-1}B) = \text{Im } B + \text{Im } AB + \cdots + \text{Im } A^{n-1}B \subset \mathbb{R}x_1,$$

ce qui contredit la condition de Kalman.

Fait 2 : Pour tout $k \leq n$, il existe $x_k \in Ax_{k-1} + \text{Im } B$ (et donc il existe $y_{k-1} \in \mathbb{R}^m$ tel que $x_k = Ax_{k-1} + By_{k-1}$) tel que dim $E_k = k$, où $E_k = \text{Vect}\{x_1, \ldots, x_k\}$.

En effet sinon, on a $Ax_{k-1} + \text{Im } B \subset E_{k-1}$, d'où $Ax_{k-1} \subset E_{k-1}$ et Im $B \subset E_{k-1}$. On en déduit que

$$AE_{k-1} \subset E_{k-1}.$$

En effet, on remarque que $Ax_1 = x_2 - By_1 \in E_{k-1} + \text{Im } B \subset E_{k-1}$, de même pour Ax_2, etc, $Ax_{k-2} = x_{k-1} - By_{k-1} \in E_{k-1} + \text{Im } B \subset E_{k-1}$, et enfin, $Ax_{k-1} \in E_{k-1}$.

Par conséquent :

$$\text{Im } AB = A\text{Im } B \subset AE_{k-1} \subset E_{k-1},$$

et de même :

$$\forall i \in \mathbb{N} \quad \text{Im } A^iB \subset E_{k-1}.$$

D'où :

$$\text{Im } (B, AB, \ldots, A^{n-1}B) \subset E_{k-1},$$

ce qui contredit la condition de Kalman.

On a donc ainsi construit une base $(x_1, \ldots, x_n)$ de $\mathbb{R}^n$. On définit alors $C \in \mathcal{M}_{m,n}(\mathbb{R})$ par les relations :

$$Cx_1 = y_1,\ Cx_2 = y_2, \ldots,\ Cx_{n-1} = y_{n-1},\ Cx_n \text{ quelconque.}$$

Alors la paire $(A + BC, x_1)$ vérifie la condition de Kalman, car :

$$(A + BC)x_1 = Ax_1 + By_1 = x_2, \ldots, (A + BC)x_{n-1} = Ax_{n-1} + By_{n-1} = x_n.$$

□

Le théorème est prouvé.

□

3.5 Observabilité

3.5.1 Définition et critère d'observabilité de Kalman

Considérons à nouveau un système de commande (Σ) linéaire autonome. En général, les mesures dont on dispose ne nous permettent pas d'observer directement l'état $x(t)$ mais seulement un vecteur $y(t) \in \mathbb{R}^p$, fonction de l'état et de la commande. Nous supposerons cette fonction linéaire et indépendante du temps, c'est-à-dire que le système est maintenant de la forme :

$$(\widetilde{\Sigma}) \ : \quad \left\{ \begin{array}{rcl} x'(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) + Du(t) \end{array} \right. , \qquad t \in [0, \tau],$$

où $C \in M_{p,n}(\mathbb{R})$ et $D \in M_{p,m}(\mathbb{R})$. Le problème de l'*observabilité* est le suivant : connaissant $y(t)$ et $u(t)$ pour tout $t \in [0, \tau]$ $(\tau > 0)$ est-il possible de déterminer la condition initiale $x(0)$? Remarquons :

— que la connaissance de $x(0)$ est équivalente à celle de $x(t)$ pour tout $t \in [0,\tau]$ puisque d'après la formule de variation de la constante

$$x(t) = e^{tA}x(0) + \int_0^\tau e^{(\tau-s)A}Bu(s)ds,$$

le deuxième terme du membre de droite de l'égalité précédente étant supposé connu ;

— que l'on peut supposer $D = 0$ et $B = 0$ puisque l'on connaît $u(\cdot)$.

Il suffit donc d'étudier le problème de l'observabilité pour le système réduit

$$(\widetilde{\Sigma}_0) \ : \quad \left\{ \begin{array}{rcl} x'(t) & = & Ax(t) \\ y(t) & = & Cx(t) \end{array} \right. , \qquad t \in [0,\tau],$$

c'est-à-dire que l'on se ramène à l'étude de $y(t) = Ce^{tA}x_0$. Appelons *espace d'inobservabilité* $\mathcal{I}_\tau$ du système $(\widetilde{\Sigma}_0)$ l'ensemble des conditions initiales $x(0) \in \mathbb{R}^n$ pour lesquelles la solution $y(t)$ est identiquement nulle sur $[0,\tau]$, *i.e.*

$$\mathcal{I}_\tau = \left\{ x_0 \in \mathbb{R}^n \ : \quad \begin{array}{c} \text{la solution de } (\widetilde{\Sigma}_0) \\ \text{avec } x(0) = x_0 \text{ vérifie } y(t) \equiv 0 \end{array} \right\}.$$

Définition 3.6. On dit que le système $(\widetilde{\Sigma}_0)$ est *observable* si son espace d'inobservabilité est réduit à $\{0\}$.

Le résultat élémentaire suivant montre que cette définition de l'observabilité correspond bien à la question que l'on s'était posée initialement

Proposition 3.10. *Si le système $(\widetilde{\Sigma}_0)$ est observable, la connaissance de $y(\cdot)$ sur $[0,\tau]$ détermine de façon univoque $x(0)$.*

Preuve.

▷ Si ce n'était pas le cas il existerait deux vecteurs distincts x_0 et $\tilde{x}_0$ dans $\mathbb{R}^n$ tels que

$$Ce^{tA}x_0 = Ce^{tA}\tilde{x}_0,$$

ce qui entraînerait $Ce^{tA}(x_0 - \tilde{x}_0) = 0$. Mais d'après la définition de l'observabilité ceci implique $x_0 = \tilde{x}_0$.

□

Il existe un critère très simple permettant de déterminer si un système est observable.

Théorème 3.11 (Critère d'observabilité de Kalman)**.** *L'espace d'inobservabilité du système* $(\widetilde{\Sigma}_0)$ *est le noyau de la matrice* $(np \times n)$

$$\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

Autrement dit, le système $(\widetilde{\Sigma}_0)$ *est observable si et seulement si* $\ker \mathcal{O} = \{0\}$.

Preuve.

▷ D'après le théorème de Cayley-Hamilton (c'est un argument que nous avons déjà utilisé dans la démonstration du théorème 3.2), pour tout $t \in [0, \tau]$

$$Ce^{tA} \in \mathrm{Vect}(C, CA, \ldots, CA^{n-1}).$$

Par conséquent, si

$$CA^j v = 0, \quad j = 0, \ldots, n-1, \tag{3.4}$$

on a $Ce^{tA}v = 0$, c'est-à-dire que v est dans l'espace d'inobservabilité de $(\widetilde{\Sigma}_0)$. Mais la condition (3.4) est équivalente au fait que $v \in \ker \mathcal{O}$. Nous avons donc démontré que $\ker \mathcal{O}$ est inclus dans l'espace d'inobservabilité de $(\widetilde{\Sigma}_0)$.

▷ Réciproquement, supposons que pour tout $t \in [0, \tau]$,

$$Ce^{tA}v = 0.$$

Alors, en dérivant j fois l'égalité précédente en $t = 0$ $(0 \leq j \leq n-1)$, il vient

$$\forall 0 \leq j \leq n-1,\ CA^j v = 0,$$

ce qui signifie que $v \in \ker \mathcal{O}$. L'inclusion réciproque est démontrée.

□

Remarque. Si l'on compare le théorème précédent avec le théorème 3.2 on s'aperçoit que le système $(\widetilde{\Sigma}_0)$ est observable si et seulement si le système dual (Σ) : $z'(t) = A^T z(t) + C^T u(t)$ est commandable (prendre la transposée de la matrice $\mathcal{O}$). C'est la *dualité contrôlabilité/observabilité*. Ce fait, très important, permet de transférer aux systèmes observés tous les résultats établis sur les systèmes contrôlés.

Définition 3.7 (Equivalence linéaire)**.** Les systèmes

$$\begin{cases} \dot{x}_1 = A_1 x_1 + B_1 u_1 \\ y_1 = C_1 x_1 \end{cases} \quad \text{et} \quad \begin{cases} \dot{x}_2 = A_2 x_2 + B_2 u_2 \\ y_2 = C_2 x_2 \end{cases}$$

sont dits *linéairement équivalents* s'il existe une matrice $P \in GL_n(\mathbb{R})$ telle que

$$A_2 = PA_1P^{-1},\ B_2 = PB_2,\ C_2 = C_1P^{-1}$$

(et dans ce cas on a $x_2 = Px_1, u_2 = u_1, y_2 = y_1$).

Proposition 3.12. *Tout système $\dot{x} = Ax+Bu$, $y = Cx$, est linéairement équivalent à un système $\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u$, $y = \bar{C}\bar{x}$, avec*

$$\bar{A} = \begin{pmatrix} \bar{A}_1 & 0 \\ \bar{A}_2 & \bar{A}_3 \end{pmatrix},\ \bar{C} = (\bar{C}_1\ 0),$$

i.e.

$$\begin{cases} \dot{\bar{x}}_1 = \bar{A}_1\bar{x}_1 + \bar{B}_1u \\ \dot{\bar{x}}_2 = \bar{A}_2\bar{x}_1 + \bar{A}_3\bar{x}_2 + \bar{B}_2u \qquad \textit{partie non observable} \\ y_1 = \bar{C}_1\bar{x}_1 \end{cases}$$

et la paire $(\bar{A}_1, \bar{C}_1)$ est observable.

Preuve.

▷ Il suffit d'appliquer le résultat vu en contrôlabilité au système $\dot{x} = A^Tx + C^Tu$.

□

Définition 3.8. Dans cette décomposition, les valeurs propres de $\bar{A}_3$ sont appelées *pôles inobservables* de A et les valeurs propres de $\bar{A}_1$ sont dites *pôles observables* de A.

Proposition 3.13 (Forme canonique d'observabilité, cas $p = 1$)**.** *Dans le cas $p = 1$, le système $\dot{x} = Ax+Bu$, $y = Cx$, est observable si et seulement si il est linéairement équivalent au système $\dot{x}_1 = A_1x_1 + B_1u$, $y = C_1x_1$, avec*

$$A_1 = \begin{pmatrix} 0 & & \cdots & 0 & -a_n \\ 1 & 0 & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \\ 0 & \cdots & 0 & 1 & -a_1 \end{pmatrix},\quad C_1 = (0\ \cdots\ 0\ 1).$$

3.5.2 Stabilisation par retour d'état statique

On peut se demander si, étant donné un système contrôlable et observable $\dot{x} = Ax + Bu$, $y = Cx$, il existe un feedback $u = Ky$ stabilisant le système, i.e. si la matrice $A + BKC$ est Hurwitz.

La réponse est *NON*. Pour le voir, considérons les matrices

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ B = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ C = (1\ 0).$$

Le système $\dot{x} = Ax + Bu$, $y = Cx$, est trivialement contrôlable et observable. Pourtant, pour toute matrice scalaire $K = (k)$, la matrice

$$A + BKC = \begin{pmatrix} 0 & 1 \\ k & 0 \end{pmatrix}$$

n'est pas Hurwitz.

En conclusion, un feedback par retour d'état statique ne suffit pas en général. C'est pourquoi, dans la suite, on va voir comment construire un retour d'état dynamique.

3.5.3 Observateur asymptotique de Luenberger

Motivation : supposons que le système $\dot{x} = Ax+Bu$, $y = Cx$, soit observable. Le but est de construire un *observateur asymptotique* $\hat{x}(\cdot)$ de $x(\cdot)$, i.e. une fonction dynamique $\hat{x}(\cdot)$ de l'observable $y(\cdot)$, telle que $\hat{x}(t) - x(t) \underset{t\to+\infty}{\longrightarrow} 0$. L'idée est de copier la dynamique du système observé et d'y ajouter un terme correctif qui tienne compte de l'écart entre la prédiction et la réalité.

Définition 3.9. Un *observateur asymptotique* (ou *observateur de Luenberger*) $\hat{x}(\cdot)$ de $x(\cdot)$ est une solution d'un système du type

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(C\hat{x}(t) - y(t)),$$

où $L \in \mathcal{M}_{n,p}(\mathbb{R})$ est appelée *matrice de gain*, telle que

$$\forall x(0), \hat{x}(0) \in \mathbb{R}^n \quad \hat{x}(t) - x(t) \underset{t\to+\infty}{\longrightarrow} 0.$$

Remarque. Introduisons $e(t) = \hat{x}(t) - x(t)$, l'erreur entre la prédiction $\hat{x}(\cdot)$ et l'état réel $x(\cdot)$. On a :

$$\dot{e}(t) = (A + LC)e(t),$$

et donc $e(t) \underset{t\to+\infty}{\longrightarrow} 0$ pour toute valeur initiale $e(0)$ si et seulement si la matrice $A+LC$ est Hurwitz. Construire un observateur asymptotique revient donc à déterminer une matrice de gain L telle que $A+LC$ soit Hurwitz. Ainsi, de manière duale au théorème de placement de pôles, on a :

Théorème 3.14 (Théorème de placement des pôles de l'observateur)**.** *Si la paire (A, C) est observable, alors le système admet un observateur asymptotique (i.e. on peut construire une matrice de gains L telle que $A + LC$ soit Hurwitz).*

Preuve.

▷ La paire (A^T, C^T) étant contrôlable, d'après le théorème de placement de pôles il existe une matrice L^T telle que la matrice $A^T + C^T L^T$ soit Hurwitz.

□

3.5.4 Stabilisation par retour dynamique de sortie

On a vu comment construire :
— un régulateur (feedback) pour un système contrôlable,
— un observateur asymptotique pour un système observable.
Il semble naturel, pour un système contrôlable et observable, de construire un régulateur en fonction de l'observateur asymptotique de l'état : c'est l'étape de *synthèse régulateur-observateur*.

Définition 3.10. On appelle *feedback dynamique de sortie*, ou *observateur-régulateur*, le feedback $u = K\hat{x}$, où

$$\dot{\hat{x}} = A\hat{x} + Bu + L(C\hat{x} - y).$$

Théorème 3.15 (Théorème de stabilisation par retour dynamique de sortie)**.** *Si le système $\dot{x} = Ax + Bu$, $y = Cx$, est contrôlable et observable, alors il est stabilisable par retour dynamique de sortie, i.e. il existe des matrices de gain $K \in \mathcal{M}_{m,n}(\mathbb{R})$ et $L \in \mathcal{M}_{n,p}(\mathbb{R})$ telles que les matrices $A + BK$ et $A + LC$ soient Hurwitz, et alors le système bouclé*

$$\begin{aligned} \dot{x} &= Ax + BK\hat{x} \\ \dot{\hat{x}} &= (A + BK)\hat{x} + LC(\hat{x} - x) \end{aligned}$$

est asymptotiquement stable.

Preuve.

▷ Posons $e = \hat{x} - x$. Alors :

$$\frac{d}{dt}\begin{pmatrix} x \\ e \end{pmatrix} = \begin{pmatrix} A + BK & BK \\ 0 & A + LC \end{pmatrix}\begin{pmatrix} x \\ e \end{pmatrix},$$

et donc ce système est asymptotiquement stable si et seulement si les matrices $A + BK$ et $A+LC$ sont Hurwitz, ce qui est possible avec les propriétés de contrôlabilité et d'observabilité.

□

Remarque. Le fait que la tâche de stabilisation se résolve indépendement de celle de reconstruction porte le nom de *principe de séparation*.

Chapitre 4

Commandabilité des systèmes non linéaires

4.1 Commandabilité locale et globale

Nous commençons ce chapitre en rappelant la notion de commandabilité pour un système de contrôle du type

$$\dot{x} = F(x, u), \qquad x \in \Omega \subset \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m, \tag{4.1}$$

où F est lisse ($\mathcal{C}^\infty$) par rapport à x. On supposera dans toute la suite que Ω est connexe.

Une classe spécialement importante de systèmes est celle dite des *systèmes affines en le contrôle*, du type

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad x \in \Omega \subset \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m.$$

On appelle f_0 la *dérive* du système et $f_1, \ldots, f_m$ les *champs contrôlés*.

Pour simplifier la discussion, nous considérons comme lois de commande admissibles les fonctions du temps qui sont constantes par morceaux et à valeur dans l'ensemble U. Nous ferons toujours l'hypothèse que toutes les trajectoires de (4.1) correspondantes à un contrôle constant sont définies et restent dans Ω pour tout temps (positif et négatif). On dira alors que chaque champ de vecteurs $F(\cdot, u)$ est *complet*.

Étant donné $x_0 \in \mathbb{R}^n$ et une loi de commande $u : [0, \tau] \to \mathbb{R}^m$, on note $x(t; x_0, u)$ la solution de (4.1) démarrant à x_0 à l'instant 0, associée à la commande $u(\cdot)$ et évaluée au temps t. On dit qu'un état $x_1 \in \mathbb{R}^n$ est *atteignable en temps* τ à partir de x_0 (par le système (4.1) s'il existe une loi de commande $u : [0, \tau] \to \mathbb{R}^m$ telle que $x(\tau; x_0, u) = x_1$.

On note $\mathcal{A}(\tau, x_0)$ l'ensemble des états atteignables à partir de x_0 en temps τ. On note aussi

$$\mathcal{A}(\leq \tau, x_0) = \cup_{t \in [0,\tau]} \mathcal{A}(t, x_0)$$

et

$$\mathcal{A}(x_0) = \cup_{t\in[0,+\infty[}\mathcal{A}(t, x_0).$$

Définition 4.1. Le système (4.1) est dit *complètement commandable* si $\mathcal{A}(x_0) = \Omega$ pour tout $x_0 \in \Omega$. Le système (4.1) est dit *localement commandable à x_0 en temps petit* si x_0 appartient à l'intérieur de $\mathcal{A}(\leq \tau, x_0)$ pour tout $\tau > 0$.

La linéarisation de (4.1) autour d'un équilibre (x_0, u_0) montre que si le système linéaire

$$\dot{y} = D_xF(x_0, u_0)y + D_uF(x_0, u_0)v, \qquad y \in \mathbb{R}^n, \qquad v \in \mathbb{R}^m,$$

est commandable, alors (4.1) est localement commandable à x_0 en temps petit.

Ce résultat ne nous suffit pas pour pouvoir nous prononcer sur la commandabilité locale, même dans des situations "non linéaires" relativement simples comme celle du robot E=M6 décrit par les équations

$$(E = M6) \quad \left\{ \begin{array}{lcl} x'(t) & = & u_1 \cos(\theta(t)), \\ y'(t) & = & u_1 \sin(\theta(t)), \\ \theta'(t) & = & u_2, \end{array} \right.$$

avec $(x, y, \theta) \in \mathbb{R}^2 \times S^1$ et $(u_1, u_2) \in \mathbb{R}^2$. Il a été vu en exercice que le linéarisé de $(E = M6)$ autour d'un équilibre formé d'un point $(\bar{x}, \bar{y}, \bar{\theta})$ quelconque et du contrôle $(0, 0)$ n'est jamais commandable bien que le système non linéaire $(E = M6)$ le soit.

Plutôt que la linéarisation, un outil plus adapté pour l'étude de la commandabilité (locale ou globale) des systèmes de contrôle non linéaires s'avère être le crochet de Lie, présenté dans la section suivante.

4.2 Crochets et algèbres de Lie

Soient f et g deux champs de vecteurs lisses et complets sur $\Omega \subset \mathbb{R}^n$. Le crochet de Lie entre f et g est défini comme le champ de vecteurs $[f, g]$ donné par la formule

$$[f, g](x) = Dg(x)f(x) - Df(x)g(x).$$

Remarquons que, pour tout $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$[f, \lambda_1 g_1 + \lambda_2 g_2] = \lambda_1[f, g_1] + \lambda_2[f, g_2]$$

et

$$[g, f] = -[f, g].$$

En particulier $[f, f] = 0$.

Pour comprendre le rôle du crochet de Lie dans l'analyse de commandabilité, associons à f et g, par analogie avec l'exponentielle matricielle, deux familles de transformations

de Ω, notés par e^{tf} et e^{tg}, donnés par les flots correspondants aux équations $\dot{x} = f(x)$ et $\dot{x} = g(x)$. Cela signifie que $e^{tf}(x_0)$, défini pour tout $x_0 \in \Omega$ et $t \in \mathbb{R}$, est l'évaluation au temps t de la solution du problème de Cauchy

$$\dot{x}(t) = f(x(t)), \qquad x(0) = x_0.$$

Remarquons que, pour tout $t, s \in \mathbb{R}$, $e^{(t+s)f} = e^{tf} \circ e^{sf}$. En particulier e^{tf} est inversible et $(e^{tf})^{-1} = e^{-tf}$.

Le lemme suivant relie la direction du champ de vecteurs $[f, g]$ aux propriétés de commutation des flots associés à f et g.

Lemme 4.1. *Pour tout $x \in \Omega$,*

$$e^{-tg} \circ e^{-tf} \circ e^{tg} \circ e^{tf}(x) = x + t^2[f, g](x) + o(t^2)$$

lorsque t tend vers 0.

Preuve.

▷ Il suffit de faire, pour chaque flot, un développement limité à l'ordre 3 lorsque t tend vers 0. Ainsi, on a

$$e^{tf}(x) = x + tf(x) + \frac{t^2}{2} Df(x)f(x) + O(t^3),$$

puis

$$e^{tg} \circ e^{tf}(x) = x + t\big(f(x) + g(x)\big) + \frac{t^2}{2} Df(x)f(x) + t^2 Dg(x)f(x) + \frac{t^2}{2} Dg(x)g(x) + O(t^3),$$

ce qui donne

$$e^{-tf} \circ e^{tg} \circ e^{tf}(x) = x + tg(x) + [f, g](x) + \frac{t^2}{2} Dg(x)g(x) + O(t^3),$$

et enfin le résultat.

□

Un corollaire très important du calcul précédent est le suivant.

Corollaire 4.2. *Les flots e^{tf} et e^{tg} correspondant à deux champs de vecteur f, g lisses et complets commutent pour tout temps t si et seulement si leur crochet de Lie $[f, g]$ est nul.*

Le résultat précédent nous fait interpréter une condition du type

$$[f, g](x) \notin \mathrm{Vect}(f(x), g(x)),$$

dans le sens suivant : la liberté de choisir à chaque instant une dynamique entre celle de f et celle de g nous permet d'atteindre, en démarrant de x, des points qui ne sont pas directement atteignables par une combinaison linéaire de $f(x)$ et $g(x)$. Ceci est à la base de la théorie de la commandabilité des systèmes non linéaires et est bien illustré par l'exemple du robot E=M6.

La dynamique ($E = M6$) peut s'écrire comme

$$\dot{z} = u_1 f_1(z) + u_2 f_2(z),$$

avec $z = (x, y, \theta)$ et

$$f_1(z) = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}, \quad f_2(z) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Le crochet de Lie $[f_1, f_2]$ correspond en pratique à effectuer un créneau (infinitésimal) et est égal à

$$[f_1, f_2](z) = \begin{pmatrix} \sin(\theta) \\ -\cos(\theta) \\ 0 \end{pmatrix}.$$

Il est clair que pour tout $z \in \mathbb{R}^2 \times S^1$, les vecteurs $f_1(z)$, $f_2(z)$ et $f_3(z)$ sont linéairement indépendants.

Une notion particulièrement importante qui généralise la situation décrite dans le cas du robot E=M6 est celle de la *génération par crochets de Lie*. Pour l'introduire, nous associons à une famille $\mathcal{F}$ de champs de vecteurs lisses et complets sur Ω, l'espace vectoriel $\mathrm{Lie}(\mathcal{F})$ engendré par $\mathcal{F}$ et tous les champs de vecteurs de la forme

$$[f_1, [f_2, [\ldots, [f_{k-2}, [f_{k-1}, f_k]]\cdots]]],$$

où $k \geq 2$ et $f_1, \ldots, f_k \in \mathcal{F}$.

Définition 4.2. On appelle $\mathrm{Lie}(\mathcal{F})$ l'*algèbre de Lie associée à* $\mathcal{F}$ et on dit que $\mathcal{F}$ *satisfait la condition de génération par crochets de Lie* à un point $x \in \Omega$ si la dimension de $\mathrm{Lie}_x(\mathcal{F}) := \{f(x) \mid f \in \mathrm{Lie}(\mathcal{F})\}$ est égale à la dimension de Ω. On dit que $\mathcal{F}$ satisfait la condition de génération par crochets de Lie si elle la satisfait pour tout $x \in \Omega$.

4.3 Accessibilité locale et conditions suffisantes pour la commandabilité globale

Considérons à nouveau le système (4.1) et notons $\mathcal{F}_U$ l'ensemble

$$\mathcal{F}_U = \{F(\cdot, u) \mid u \in U\}.$$

Les éléments de $\mathcal{F}_U$ sont dits *champs admissibles* de (4.1). La condition de génération par crochets de Lie exprime la notion suivante : en commutant entre les dynamiques de la famille $\mathcal{F}_U$, toutes les directions de l'espace deviennent atteignables par des trajectoires de (4.1).

Cela doit néanmoins tenir compte des contraintes de dérive : nous pouvons penser au cas où $\Omega = \mathbb{R}^2$ et $\mathcal{F}_U$ est donnée par la famille de champs de vecteurs constants $\{(1,1),(1,-1)\}$. Alors $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie en tout point de $\mathbb{R}^2$, même si le système $(\dot{x}_1, \dot{x}_2) = (1, \pm 1)$ n'est pas complètement commandable ni localement commandable en aucun point.

L'atteignabilité "dans toutes les directions" doit s'interpréter dans le sens du célèbre résultat suivant.

Théorème 4.3 (Krener)**.** *Si $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie à $x_0 \in \Omega$, alors, pour tout $\tau > 0$, x_0 appartient à la fermeture de l'intérieur de $\mathcal{A}(\leq \tau, x_0)$.*

Preuve.

▷ Remarquons d'abord que $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie en tous les points d'un voisinage de x_0. (Si n champs de vecteurs sont linéairement indépendants à x_0, ils le sont dans un voisinage.)

Remarquons aussi qu'il existe $f \in \mathcal{F}_U$ tel que $f(x_0) \neq 0$: on aurait sinon $\mathrm{Lie}_{x_0}(\mathcal{F}_U) = \{0\}$.

Si $\dim(\Omega) = 1$ la preuve est finie. Si $\dim(\Omega) > 1$ et si tout champ dans $\mathcal{F}_U$ est tangent à la courbe $t \mapsto e^{tf}(x_0)$, $0 < t < \varepsilon$, alors on déduit du Lemme 4.1 que $\mathrm{Lie}_{e^{tf}(x_0)}(\mathcal{F}_U)$ est aussi tangent à telle courbe et qu'il n'a donc pas la même dimension que Ω. Ceci contredit la première remarque faite dans cette preuve. Il existe donc $g \in \mathcal{F}_U$ et $0 < \bar{t} < \varepsilon$ tels que f et g sont linéairement indépendants dans un voisinage de $x_1 = e^{\bar{t}f}(x_0)$. Alors $(t,s) \mapsto e^{sg} \circ e^{tf}(x_0)$, $0 < s < \varepsilon'$, $\bar{t} - \varepsilon' < t < \bar{t} + \varepsilon'$, a comme image une surface de dimension deux.

Si $\dim(\Omega) = 2$ la preuve est finie. Sinon on recommence le même argument et on conclue par récurrence par rapport à la dimension de Ω.

□

Le théorème peut être interprété comme suit : sous la condition de génération par crochets de Lie, les trajectoires admissibles démarrant en un point x_0 peuvent accéder à un ensemble d'intérieur non vide en temps arbitrairement petit (on parle aussi, dans la littérature, d'accessibilité locale).

Dans le cas où l'ensemble des directions admissibles de (4.1) est convenablement symétrique, la conclusion du théorème de Krener devient plus explicite.

Théorème 4.4 (Chow). *Soient $f_1, \ldots, f_m$ des champs de vecteurs lisses et complets sur Ω. Si $U \subset \mathbb{R}^m$ contient 0 dans son intérieur et si $\{f_1, \ldots, f_k\}$ satisfait la condition de génération par crochets de Lie (en tout $x \in \Omega$), alors*

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x)$$

est complètement commandable.

Preuve.

▷ Sans perte de généralité, nous pouvons supposer que U est une boule de $\mathbb{R}^m$ centrée à l'origine.

Soit $x_0 \in \Omega$. Remarquons que $\mathcal{F}_U$, avec $F(x,u) = \sum_{i=1}^m u_i f_i(x)$, satisfait la condition de génération par crochets de Lie. Nous déduisons donc du théorème de Krener que $\mathcal{A}(x_0)$ contient un ouvert non vide ω.

Choisissons $u^1, \ldots, u^k \in U$ et $t_1, \ldots, t_k > 0$ tels que

$$e^{t_k \sum_{i=1}^m u_i^k f_i} \circ \cdots \circ e^{t_1 \sum_{i=1}^m u_i^1 f_i}(x_0) \in \omega.$$

Puisque $-u^1, \ldots, -u^k \in U$, nous avons que

$$e^{t_1 \sum_{i=1}^m (-u_i^1) f_i} \circ \cdots \circ e^{t_k \sum_{i=1}^m (-u_i^k) f_i}(\mathcal{A}(x_0)) \subset \mathcal{A}(x_0).$$

En particulier, $\mathcal{A}(x_0)$ contient

$$e^{t_1 \sum_{i=1}^m (-u_i^1) f_i} \circ \cdots \circ e^{t_k \sum_{i=1}^m (-u_i^k) f_i}(\omega)$$

qui est un voisinage de x_0.

Nous en déduisons que $\mathcal{A}(x_0)$ est ouvert. (En effet, si $x_1 \in \mathcal{A}(x_0)$, alors $x_1 \in \mathrm{int}(\mathcal{A}(x_1)) \subset \mathcal{A}(x_1) \subset \mathcal{A}(x_0)$.) De plus, la symétrie de U implique (comme nous venons de le voir) que $x_1 \in \mathcal{A}(x_0)$ si et seulement si $x_0 \in \mathcal{A}(x_1)$. Donc $\{\mathcal{A}(x_0) \mid x_0 \in \Omega\}$ est une partition de Ω en ouverts disjoints. Ω étant connexe, nous pouvons conclure que $\mathcal{A}(x_0) = \Omega$ pour tout $x_0 \in \Omega$. □

4.4 Champs compatibles

Une autre conséquence très importante du théorème de Krener est le corollaire suivant, qui sera la clé pour étendre le champ d'applications du théorème de Chow.

Corollaire 4.5. *Si $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie (à tout $x \in \Omega$) et si $\mathcal{A}(x_0)$ est dense dans Ω pour un certain $x_0 \in \Omega$, alors $\mathcal{A}(x_0) = \Omega$.*

Preuve.

▷ Soit $x_1 \in \Omega$ et considérons le système

$$\dot{x} = -F(x,u), \qquad x \in \Omega \subset \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m, \tag{4.2}$$

qui est obtenu à partir de (4.1) par renversement temporel. La famille de champs de vecteurs admissibles pour (4.2) étant $-\mathcal{F}_U$, elle satisfait la condition de génération par crochets de Lie.

L'ensemble atteignable pour (4.2) au départ de x_1, noté par $\mathcal{A}^-(x_1)$, contient donc un ouvert non vide (Krener). En particulier, $\mathcal{A}^-(x_1)$ a intersection non vide avec $\mathcal{A}(x_0)$. Cela signifie précisément que $x_1 \in \mathcal{A}(x_0)$. Puisque $x_1 \in \Omega$ était arbitraire, nous avons montré que $\mathcal{A}(x_0) = \Omega$.

□

Le corollaire 4.5 nous suggère la définition suivante.

Définition 4.3. Un champ de vecteurs lisse et complet g est dit *compatible avec* $\mathcal{F}_U$ si le système de contrôle

$$\dot{x} = \hat{F}(x,u), \qquad x \in \Omega, \qquad u \in \hat{U}, \tag{4.3}$$

avec $\hat{U} = U \cup \{\hat{u}\}$, $\hat{u} \notin U$, et

$$\hat{F}(x,u) = \begin{cases} F(x,u) & \text{si } u \in U, \\ g(x) & \text{si } u = \hat{u}, \end{cases}$$

satisfait à la condition suivante : pour tout $x_0 \in \Omega$, l'ensemble atteignable pour (4.3) au départ de x_0 est contenu dans la fermeture de $\mathcal{A}(x_0)$.

Nous allons maintenant présenter des critères permettant d'identifier des champs de vecteurs compatibles avec une famille $\mathcal{F}_U$ donnée.

Le premier critère est celui dit de *convexification* : il affirme que la combinaison convexe de champs de vecteurs de $\mathcal{F}_U$ est compatible avec $\mathcal{F}_U$. Il formalise l'intuition selon laquelle si on commute très rapidement entre la dynamique de deux champs de vecteurs f et g et si on reste à chaque fois pendant la même (très petite) quantité de temps sur chacune des deux dynamiques, alors la trajectoire résultante approchera celle du champ de vecteurs $\frac{f+g}{2}$.

Le lemme suivant formule précisément ce résultat, tout en y ajoutant la remarque que la multiplication d'une dynamique par une constante positive correspond simplement à une reparamétrisation du temps.

Lemme 4.6. *Pour tous $\lambda_1, \ldots, \lambda_k \geq 0$ et tous $f_1, \ldots, f_k \in \mathcal{F}_U$, le champ de vecteurs $\lambda_1 f_1 + \cdots + \lambda_k f_k$ est compatible avec $\mathcal{F}_U$.*

La preuve du lemme, assez technique, est basée sur l'inégalité de Gronwall.

Nous pouvons déduire du lemme précédent un précieux corollaire du théorème de Chow.

Corollaire 4.7. *Si, pour tout $x \in \Omega$, $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie et si 0 appartient à l'intérieur de l'enveloppe convexe de $\{F(x,u) \mid u \in U\}$, alors (4.1) est complètement commandable.*

Un deuxième critère garantissant la compatibilité est le suivant : si g est un champ de vecteurs lisse et complet qui est la limite uniforme sur tout compact de Ω d'une suite $\{f_n\}_{n\in\mathbb{N}}$, où chaque f_n est compatible avec $\mathcal{F}_U$, alors g est compatible avec $\mathcal{F}_U$. Ce critère peut facilement se déduire de la théorie générale des équations différentielles ordinaires, en n'affirmant rien d'autre qu'un principe de continuité des solutions d'une EDO par rapport au champ de vecteurs la régissant.

Nous déduisons de ce critère le résultat suivant, qui s'applique à la classe des systèmes affines dans le contrôle.

Corollaire 4.8. *Soient $f_0, f_1, \ldots, f_m$ des champs de vecteurs lisses et complets sur Ω. Si la famille $\mathcal{G} = \{f_1, \ldots, f_m\}$ satisfait la condition de génération par crochets de Lie, alors*

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \ldots, u_m) \in \mathbb{R}^m$$

est complètement commandable.

Preuve.

▷ Le théoème de Chow nous garantit que le système

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \ldots, u_m) \in \mathbb{R}^m$$

est complètement commandable.

Si nous montrons que chaque $\sum_{i=1}^{m} u_i f_i$ est compatible avec $\mathcal{F}_U$, alors la thèse suit du corollaire 4.5.

Il suffit donc de remarquer que

$$\sum_{i=1}^{m} u_i f_i = \lim_{n\to\infty} \frac{1}{n}(f_0 + \sum_{i=1}^{m} (nu_i) f_i)$$

et que chaque champ de vecteurs $\frac{1}{n}(f_0 + \sum_{i=1}^{m}(nu_i)f_i)$ est compatible avec $\mathcal{F}_U$.

□

Un dernier critère de compatibilité très utile peut être formulé en termes de récurrence d'un champ de vecteurs.

Définition 4.4 (Champ de vecteurs récurrent)**.** Un champ de vecteurs f lisse et complet sur Ω est dit *récurrent* si, pour tout point $x_0 \in \Omega$, tout voisinage V de x_0 et tout temps $t > 0$, il existe $t^* > t$ tel que $e^{t^* f}(x_0) \in V$.

Remarquons que si les trajectoires de f sont périodiques (éventuellement avec des périodes différentes selon les trajectoires) alors f est récurrent.

Lemme 4.9. *Si f est compatible avec $\mathcal{F}_U$ et est récurrent, alors $-f$ est aussi compatible avec $\mathcal{F}_U$.*

Preuve.

▷ Soient $x_0 \in \Omega$ et $t > 0$. Il suffit de montrer que $e^{-tf}(x_0)$ est la limite de points atteignables au départ de x_0.

La définition de champ récurrent implique l'existence d'une suite croissante non bornée $\{t_k\}_{k\in\mathbb{N}}$ de temps positifs telle que $e^{t_k f}(x_0) \to x_0$ pour $k \to \infty$.

Nous avons donc que $e^{(t_k - t)f}(x_0) \to e^{-tf}(x_0)$ pour $k \to \infty$ et $t_k > t$ pour k suffisamment grand.

Puisque f est compatible avec $\mathcal{F}_U$, alors $e^{(t_k-t)f}(x_0)$ appartient à la fermeture de $\mathcal{A}(x_0)$ pour tout k tel que $t_k > t$. Nous en déduisons que $e^{-tf}(x_0)$ appartient à la fermeture de $\mathcal{A}(x_0)$.

□

En conséquence du lemme précédent, nous pouvons parfois déterminer la commandabilité d'un système affine dans le contrôle même dans le cas où l'ensemble des contrôles est borné.

Corollaire 4.10. *Soient $f_0, f_1, \ldots, f_m$ des champs de vecteurs lisses et complets sur Ω et considérons le système*

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \ldots, u_m) \in U. \tag{4.5}$$

Supposons que 0 appartient à l'intérieur de l'enveloppe convexe de U, que $\{f_0, \ldots, f_m\}$ satisfait la condition de génération par crochets de Lie et que f_0 soit récurrent. Alors (4.5) est complètement commandable.

Preuve.

▷ Remarquons que f_0 est compatible avec $\mathcal{F}_U$ à cause du lemme 4.6. Lemme 4.9 affirme l'équivalence entre la commandabilité de (4.5) et celle de

$$\dot{x} = \sum_{i=0}^{m} u_i f_i(x), \qquad (u_0, \ldots, u_m) \in \{-1, 1\} \times U.$$

Cette dernière suit du corollaire 4.7.

□

Exemple. Soit $\Omega = S^2$, la sphère unité de $\mathbb{R}^3$, c'est-à-dire $\{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}$. Considérons les deux champs de vecteurs

$$f_0(x) = \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \qquad f_1(x) = \begin{pmatrix} 0 \\ -x_3 \\ x_2 \end{pmatrix},$$

dont les flots au temps t sont les rotations d'angle t par rapport, respectivement, à $(0, 0, 1)^T$ et $(1, 0, 0)^T$.

Le système commandé

$$\dot{x} = f_0(x) + u f_1(x), \qquad u \in (-1, 1), \qquad x \in S^2,$$

est complètement commandable puisque les trajectoires de f_0 sont périodiques (et donc f_0 est récurrent) et le crochet entre f_0 et f_1 est donné par

$$[f_0, f_1](x) = \begin{pmatrix} -x_3 \\ 0 \\ x_1 \end{pmatrix},$$

de telle sorte que $\mathrm{Lie}_x(\{f_0, f_1\})$ est de dimension deux pour tout $x \in S^2$.

4.5 Orbites et conditions nécessaires pour la commandabilité

Nous avons vu dans les sections précédentes plusieurs conditions suffisantes pour la commandabilité d'un système de contrôle non linéaire.

Cette section présente plutôt des conditions nécessaires, déduites d'un résultat profond de nature géométrique, le théorème de l'orbite. Ce théorème assure que, dans les cas non "pathologiques", l'algèbre de Lie associée à la famille $\mathcal{F}_U$ mesure de façon précise la taille de l'ensemble des directions auquelles on peut accéder à partir d'un point.

Nous définissons l'orbite à partir d'un point $x_0 \in \Omega$ pour le système (4.1) comme l'ensemble

$$\mathcal{O}(x_0) = \{e^{t_k f_k} \circ \cdots \circ e^{t_1 f_1}(x_0) \mid k \in \mathbb{N},\ t_1, \ldots, t_k \in \mathbb{R},\ f_1, \ldots, f_k \in \mathcal{F}_U\}.$$

Rappelons, par comparaison, que

$$\mathcal{A}(x_0) = \{e^{t_k f_k} \circ \cdots \circ e^{t_1 f_1}(x_0) \mid k \in \mathbb{N},\ t_1, \ldots, t_k \geq 0,\ f_1, \ldots, f_k \in \mathcal{F}_U\}.$$

Nou avons alors le théorème suivant.

Théorème 4.11 (Orbite)**.** *Pour chaque $x_0 \in \Omega$, l'ensemble $\mathcal{O}(x_0)$ a la structure d'une variété immergée. En particulier, celle-ci a même dimension en tout point. De plus, l'espace des directions tangentes à $\mathcal{O}(x_0)$ à un point $x \in \mathcal{O}(x_0)$ contient $\mathrm{Lie}_x(\mathcal{F}_U)$ et les deux espaces sont égaux si une des deux conditions suivantes est vérifiée : (i) chaque composante de chaque champ de vecteurs de $\mathcal{F}_U$ est une fonction analytique, ou (ii) la dimension de $\mathrm{Lie}_x(\mathcal{F}_U)$ est constante par rapport à $x \in \mathcal{O}(x_0)$.*

Le corollaire suivant retient du théorème de l'orbite les conséquences les plus directement exploitables pour l'analyse de commandabilité d'un système non linéaire.

Corollaire 4.12. *Si $\mathcal{F}_U$ ne satisfait pas la condition de génération par crochets de Lie et si (i) chaque composante de chaque champ de vecteurs dans $\mathcal{F}_U$ est une fonction analytique, ou (ii) la dimension de $\mathrm{Lie}_x(\mathcal{F}_U)$ est constante par rapport à $x \in \Omega$, alors (4.1) n'est pas complètement commandable.*

Chapitre 5

Théorie linéaire-quadratique

Dans ce chapitre, on s'intéresse aux systèmes de contrôle linéaires avec un coût quadratique. Ces systèmes sont d'une grande importance dans la pratique, comme on le verra en section 5.4. En effet un coût quadratique est souvent très naturel dans un problème, par exemple lorsqu'on veut minimiser l'écart au carré par rapport à une trajectoire nominale (problème de poursuite). Par ailleurs même si les systèmes de contrôle sont en général non linéaires, on est très souvent amené à linéariser le système le long d'une trajectoire, par exemple dans des problèmes de stabilisation.

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

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \tag{5.1}$$

muni d'un coût quadratique du type :

$$C(u) = {}^t x(T)Qx(T) + \int_0^T ({}^t x(t)W(t)x(t) + {}^t u(t)U(t)u(t))dt, \tag{5.2}$$

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

Le problème de contrôle optimal est alors le suivant, nous l'appelerons *problème LQ* (linéaire-quadratique) :

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

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

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

de sorte que

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

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

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

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

Remarque. Les résultats des sections 5.1 et 5.2 seront en fait valables pour des systèmes linéaires perturbés $\dot{x} = Ax + Bu + r$, et aussi avec une fonction g de $\mathbb{R}^n$ dans $\mathbb{R}$ continue ou C^1 pour laquelle on puisse montrer que $(g1)$ la fonction coût $C(u)$ soit bornée inférieurement ; $(g2)$ toute suite (u_n) convergente vers la borne inférieure du coût est bornée en norme L^2. Ceci a lieu lorsque que par exemple g vérifie l'hypothèse suivante : il existe un réel a tel que $\frac{g(x)}{||x||} + a$ est borné inférieurement pour $||x||$ assez grand. Nous préciserons pour chaque résultat les extensions possibles.

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

5.1 Existence de trajectoires optimales

Introduisons l'hypothèse de coercivité suivante sur U :

$$\exists \alpha > 0 \mid \forall u \in L^2([0, T], \mathbb{R}^m) \quad \int_0^T \|u(t)\|_U^2 dt \geq \alpha \int_0^T {}^t u(t) u(t) dt. \tag{5.3}$$

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

On a le théorème d'existence suivant :

Théorème 5.1. *Sous l'hypothèse (5.3), il existe une unique trajectoire minimisante pour le problème LQ.*

Preuve.

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

un contrôle u de L^2. Notons x_n (resp. x) la trajectoire associée au contrôle u_n (resp. u) sur $[0,T]$. D'après la formule de variation de la constante, on a, pour tout $t \in [0,T]$:

$$x_n(t) = M(t)x_0 + M(t)\int_0^t M(s)^{-1}B(s)u_n(s)ds \tag{5.4}$$

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

Passant maintenant à la limite dans (5.4), on obtient, pour tout $t \in [0,T]$:

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

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

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

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

Pour l'unicité on a besoin du lemme suivant.

Lemme 5.2. *La fonction C est strictement convexe.*

Preuve.

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

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

Par conséquent, l'application qui à un contrôle u associe $x_u(t)$ est convexe (pourquoi ?), ceci pour tout $t \in [0,T]$. Or la matrice $W(t)$ étant symétrique positive, ceci implique que l'application qui à un contrôle u associe ${}^tx(t)W(t)w(t)$ est convexe. On raisonne de même pour le terme ${}^tx(T)Qx(T)$. Enfin, l'intégration respectant la convexité, on a bien que le coût est strictement convexe en u.

□

L'unicité de la trajectoire optimale en résulte trivialement.

□

Remarque (Extension du théorème 5.1). Si la fonction g apparaissant dans le coût est une fonction continue quelconque de $\mathbb{R}^n$ dans $\mathbb{R}$ vérifiant les conditions $(g1)$ et $(g2)$ (cf. remarque5), et/ou si le système de contrôle est perturbé par une fonction $r(t)$, alors le théorème précédent reste vrai.

Remarque (Cas d'un intervalle infini). Le théorème est encore valable si $T = +\infty$, avec $g = 0$, pourvu que le système (5.1) soit contrôlable (en temps quelconque).

En effet il suffit juste de montrer qu'il existe des trajectoires solutions du système (5.1) sur $[0, +\infty[$ et de coût fini. Or si le système est contrôlable, alors il existe un contrôle u et un temps $T > 0$ tel que la trajectoire associée à u relie x_0 à 0 sur $[0, T]$. On étend alors le contrôle u par 0 sur $]T, +\infty[$, de sorte que la trajectoire reste en 0. On a ainsi construit une trajectoire solution du système sur $[0, +\infty[$ et de coût fini. Ceci permet d'affirmer l'existence d'une suite de contrôles minimisants. Les autres arguments de la preuve sont inchangés. On obtient donc le résultat suivant.

Proposition 5.3. *Considérons le problème de déterminer une trajectoire solution de*

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

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

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

Si le système est contrôlable en un temps $T > 0$*, et si l'hypothèse (5.3) est satisfaite sur* $[0, +\infty[$*, alors il existe une unique trajectoire minimisante.*

Remarque.

— Si on suppose de plus que les applications $A(\cdot)$ et $B(\cdot)$ sont L^2 sur $[0, +\infty[$, et si $W(\cdot)$ vérifie comme U une hypothèse de coercivité (5.3), alors la trajectoire minimisante tend vers 0 lorsque t tend vers l'infini.
En effet on montre facilement en utilisant l'inégalité de Cauchy-Schwarz que l'application $\dot{x}(\cdot)$ est dans L^1, et par conséquent $x(t)$ converge. Sa limite est alors forcément nulle.

— Dans le cas autonome (A et B sont constantes), si $W(\cdot)$ vérifie comme U une hypothèse de coercivité (5.3), alors la trajectoire minimisante tend vers 0 lorsque t tend vers l'infini.
En effet il suffit d'écrire l'inégalité :

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

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

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

Théorème 5.4. *La trajectoire x, associée au contrôle u, est optimale pour le problème LQ si et seulement s'il existe un vecteur adjoint $p(t)$ satisfaisant pour presque tout $t \in [0,T]$:*

$$\dot{p}(t) = -p(t)A(t) + {}^t x(t)W(t) \tag{5.8}$$

et la condition finale

$$p(T) = -{}^t x(T)Q. \tag{5.9}$$

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

$$u(t) = U(t)^{-1}\,{}^t B(t)\,{}^t p(t). \tag{5.10}$$

Preuve.

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

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

engendrant les trajectoires :

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

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

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

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

$$\delta x(t) = M(t) \int_0^t M(s)^{-1} B(s) \delta u(s) ds. \tag{5.11}$$

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

$$dC(u) = 0.$$

Or

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

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

$$\frac{1}{2}dC(u).\delta u = {}^t x(T)Q\delta x(T) + \int_0^T ({}^t x(t)W(t)\delta x(t) + {}^t u(t)U(t)\delta u(t))dt = 0, \tag{5.12}$$

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

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

La formule de variation de la constante nous conduit à :

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

pour tout $t \in [0, T]$, où :

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

Revenons alors à l'équation (5.12). Tout d'abord, en tenant compte de (5.11) puis en intégrant par parties, il vient :

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

Or

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

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

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

Injectons cette égalité dans (5.12), en tenant compte du fait que :

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

On trouve alors que :

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

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

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

ce qui est la conclusion souhaitée. Réciproquement s'il existe un vecteur adjoint $p(t)$ vérifiant (5.8) et (5.9) et si le contrôle u est donné par (5.10), alors il est bien clair d'après le raisonnement précédent que :

$$ dC(u) = 0. $$

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

□

Remarque. Si le système de contrôle est perturbé par une fonction $r(t)$, alors le théorème précédent reste vrai. Il le reste, de même, si la fonction g apparaissant dans le coût est une fonction C^1 de $\mathbb{R}^n$ dans $\mathbb{R}$ vérifiant les conditions $(g1)$ et $(g2)$ (cf. remarque5), sauf que la condition finale sur le vecteur adjoint (5.9) devient :

$$ p(T) = -\frac{1}{2}\nabla g(x(T)), \tag{5.13} $$

comme on le voit facilement dans la démonstration. Cette condition s'appelle *condition de transversalité.*

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

$$ \lim_{t\to+\infty} p(t) = 0. \tag{5.14} $$

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

$$ H(x,p,u) = p(Ax + Bu) - \frac{1}{2}({}^t xWx + {}^t uUu), $$

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

$$ \dot{x} = \frac{\partial H}{\partial p} = Ax + Bu, $$
$$ \dot{p} = -\frac{\partial H}{\partial x} = -pA + {}^t xW, $$

et

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

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

Remarque. L'approche développée dans la démonstration du théorème 5.4 est variationnelle. Nous avons ici préféré l'approche du calcul des variations classique, car elle permet une preuve plus rapide et élégante. L'autre approche est en fait plus générale et sera privilégiée dans le cas général (non linéaire) où elle conduit au principe du maximum de Pontryagin général.

5.3 Fonction valeur et équation de Riccati

5.3.1 Définition de la fonction valeur

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

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x, \tag{5.15}$$

minimisant le coût quadratique

$$C_T(u) = {}^t x(T)Qx(T) + \int_0^T (\|x(t)\|_W^2 + \|u(t)\|_U^2)dt. \tag{5.16}$$

On définit alors :

Définition 5.1. La *fonction valeur* S_T au point x est la borne inférieure des coûts pour le problème LQ. Autrement dit :

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

Remarque. Sous l'hypothèse (5.3) on a existence d'une unique trajectoire optimale d'après le théorème 5.1, et dans ce cas cette borne inférieure est un minimum.

5.3.2 Equation de Riccati

Théorème 5.5. *Sous l'hypothèse (5.3), pour tout $x \in \mathbb{R}^n$ il existe une unique trajectoire optimale x associée au contrôle u pour le problème (5.15), (5.16). Le contrôle optimal se met sous forme de boucle fermée :*

$$u(t) = U(t)^{-1}\,{}^t B(t)E(t)x(t), \tag{5.20}$$

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

$$\dot{E}(t) = W(t) - {}^t A(t)E(t) - E(t)A(t) - E(t)B(t)U(t)^{-1\,t}B(t)E(t),\ E(T) = -Q. \tag{5.21}$$

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

$$S_T(x) = -\,{}^t x E(0)x. \tag{5.22}$$

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

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

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

Preuve.

▷ D'après le théorème 5.1, il existe une unique trajectoire optimale qui, d'après le théorème 5.4, est caractérisée par le système d'équations :

$$\begin{aligned}\dot{x} &= Ax + BU^{-1}\,{}^tB\,{}^tp,\\ \dot{p} &= -pA + {}^txW,\end{aligned}$$

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

$$u = U^{-1}\,{}^tB\,{}^tp.$$

Il faut donc montrer que l'on peut écrire $p(t) = {}^tx(t)E(t)$, où $E(t)$ est solution de (5.21). Notons que si p s'écrit ainsi alors d'après l'équation satisfaite par le couple (x, p), on trouve facilement que $E(t)$ doit satisfaire l'équation (5.21). En utilisant l'unicité de la trajectoire optimale, on va maintenant montrer que p s'écrit effectivement ainsi. Soit $E(t)$ solution de l'équation :

$$\dot{E} = W - {}^tAE - EA - EBU^{-1}\,{}^tBE,\ E(T) = -Q.$$

Tout d'abord $E(t)$ est symétrique car le second membre de l'équation différentielle l'est, et la matrice Q est symétrique. A priori on ne sait pas cependant que la solution est bien définie sur $[0, T]$ tout entier. On montrera cela plus loin (lemme 5.6).
Posons maintenant $p_1(t) = {}^tx_1(t)E(t)$, où x_1 est solution de

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

et $u_1 = U^{-1}\,{}^tBEx_1$. On a alors :

$$\begin{aligned}\dot{p}_1 &= {}^t\dot{x}_1E + {}^tx_1\dot{E}\\ &= {}^t(Ax_1 + BU^{-1}\,{}^tBEx_1)E + {}^tx_1(W - {}^tAE - EA - EBU^{-1}\,{}^tBE)\\ &= -p_1A + {}^tx_1W.\end{aligned}$$

Autrement dit le triplet (x_1, p_1, u_1) vérifie exactement les équations du théorème 5.4. Par conséquent la trajectoire x_1 est optimale, et par unicité il vient $x_1 = x$, $u_1 = u$, puis $p_1 = p$. En particulier on a donc $p = {}^txE$, et $u = U^{-1}\,{}^tBEx$. Déduisons-en la formule (5.22). Pour cela calculons d'abord, le long de la trajectoire $x(t)$:

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

Par ailleurs de l'expression de u on déduit :

$$ {}^t u U u = {}^t(U^{-1}\,{}^t B E x) U U^{-1}\,{}^t B E x = {}^t x E B U^{-1}\,{}^t B E x = pBu. $$

Finalement on a l'égalité :

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

et par conséquent :

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

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

Lemme 5.6. *L'application $t \mapsto E(t)$ est bien définie sur $[0, T]$ tout entier.*

Preuve.

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

$$ |\,{}^t x_0 E(t_0) x_0| \geq \alpha. \tag{5.23} $$

D'après le théorème 5.1, il existe une unique trajectoire optimale $x(\cdot)$ pour le problème LQ sur $[t_0, T]$, telle que $x(t_0) = x_0$ (voir remarque 5). Cette trajectoire est caractérisée par le système d'équations :

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

Il résulte du théorème de dépendance continue des solutions d'une équation différentielle par rapport à la condition initiale que les extrémités $x(T)$ au temps T des trajectoires issues au temps t_0 de x_0, sont uniformément bornées lorsque $0 \leq t_0 < T$ et $\|x_0\| = 1$, et donc les solutions correspondantes $x(t), p(t)$ du système différentiel précédent sont uniformément bornées sur $[0, T]$. En particulier la quantité $p(t_0)x(t_0)$ doit être bornée indépendamment de t_0. Or on sait que $p(t) = {}^t x(t) E(t)$, donc :

$$ p(t_0) x(t_0) = {}^t x_0 E(t_0) x_0, $$

et on obtient une contradiction avec (5.23).

□

Ceci achève la preuve du théorème.

□

Remarque. Il est clair d'après l'expression (5.22) du coût minimal que la matrice $E(0)$ est symétrique négative. On peut améliorer ce résultat si la matrice Q est de plus définie :

Lemme 5.7. *Si la matrice Q est symétrique définie positive, ou bien si pour tout $t \in [0,T]$ la matrice $W(t)$ est symétrique définie positive, alors la matrice $E(0)$ est symétrique définie négative.*

Preuve.

▷ [Preuve du lemme 5.7.] Soit x_0 tel que ${}^t x_0 E(0) x_0 = 0$, et montrons que $x_0 = 0$. Pour cela on considère le problème LQ :

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

pour lequel, d'après le théorème 5.5, le coût minimal vaut $-{}^t x_0 E(0) x_0 = 0$. Par conséquent, puisque pour tout t la matrice $U(t)$ est définie positive, on a $u(t) = 0$ sur $[0,T]$. Si par ailleurs Q est définie positive on a aussi $x(T) = 0$. Donc la trajectoire $x(\cdot)$ est solution du problème de Cauchy $\dot{x} = Ax, x(T) = 0$, et par unicité $x(\cdot)$ est identiquement nulle. En particulier $x(0) = x_0 = 0$, ce qui achève la preuve. Dans le deuxième cas où $W(t)$ est définie positive, la conclusion est immédiate.

□

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

$$\dot{x} = Ax + Bu,\ x(t) = x, \tag{5.24}$$

minimisant le coût quadratique

$$C_T(t,u) = g(x(T)) + \int_t^T \lambda(\|x(t)\|_W^2 + \|u(t)\|_U^2 \rho) dt. \tag{5.25}$$

On définit alors :

Définition 5.2. La *fonction valeur* S au point (t,x) est la borne inférieure des coûts pour ce problème LQ. Autrement dit :

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

Théorème 5.8. *Sous l'hypothèse (5.3), pour tout $x \in \mathbb{R}^n$ et tout $t \in [0,T]$ il existe une unique trajectoire optimale x associée au contrôle u pour le problème (5.24), (5.25). Le contrôle optimal se met sous forme de boucle fermée :*

$$u(s) = U(s)^{-1}\,{}^tB(s)E(s)x(s), \tag{5.29}$$

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

$$\dot{E} = W - {}^tAE - EA - EBU^{-1}\,{}^tBE,\ E(T) = -Q. \tag{5.30}$$

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

$$S_T(t,x) = -{}^txE(t)x. \tag{5.31}$$

Preuve.

▷ La différence par rapport au cas précédent est que l'on paramétrise le temps initial. Le seul changement est donc la formule (5.31). Comme dans la démonstration précédente, on a :

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

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

□

Remarque. L'équation de Riccati étant fondamentale, notamment dans les problèmes de régulateur (voir section suivante), la question de son implémentation numérique se pose naturellement. On peut procéder de manière directe : il s'agit alors, en tenant compte du fait que $E(t)$ est symétrique, d'intégrer un système différentiel non linéaire de $n(n+1)/2$ équations.

5.3.3 Représentation linéaire de l'équation de Riccati

On a la propriété suivante.

Proposition 5.9. *Plaçons-nous dans le cadre du théorème 5.5. Soit*

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

la résolvante du système linéaire

$$\begin{aligned} \dot{x} &= Ax + BU^{-1}\,{}^tB\,{}^tp, \\ {}^t\dot{p} &= -{}^tA\,{}^tp + Wx, \end{aligned}$$

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

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

Preuve.

▷ Par définition de la résolvante on a :

$$\begin{aligned} x(t) &= R_1(t)x(T) + R_2(t)\,{}^tp(T), \\ {}^tp(t) &= R_3(t)x(T) + R_4(t)\,{}^tp(T). \end{aligned}$$

Or on sait que ${}^tp(T) = -Qx(T)$, donc :

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

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

Par conséquent pour résoudre l'équation de Riccati (5.21), il suffit d'intégrer un système linéaire (il faut calculer une résolvante), ce qui est très facile à programmer. Cette méthode (due à Kalman-Englar) est notamment préférable à la méthode directe dans le cas stationnaire.

5.4 Applications de la théorie LQ

5.4.1 Problèmes de régulation

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

Considérons le système de contrôle linéaire perturbé :

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0, \tag{5.32}$$

et soit $\xi(t)$ une certaine trajectoire de $\mathbb{R}^n$ sur $[0,T]$, partant d'un point ξ_0 (et qui n'est pas forcément solution du système (5.32)). Le but est de déterminer un contrôle tel que la trajectoire associée, solution de (5.32), suive le mieux possible la trajectoire de référence $\xi(t)$.

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

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

qui est solution du système de contrôle :

$$\dot{z}(t) = A(t)z(t) + B(t)u(t) + r_1(t),\ z(0) = z_0, \tag{5.33}$$

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

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

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

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

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

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

pour le système de contrôle

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

partant du point $z_1(0)$.

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

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

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

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

Posons :

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

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

$$\begin{array}{lll} \dot{E} &= W - {}^t AE - EA - EBU^{-1}\,{}^t BE, & E(T) = -Q, \\ \dot{h} &= -{}^t Ah - Er_1 - EBU^{-1}\,{}^t Bh, & h(T) = 0, \\ \dot{\alpha} &= -2\,{}^t r_1 h - {}^t hBU^{-1}\,{}^t Bh, & \alpha(T) = 0. \end{array} \tag{5.34}$$

Résumons tout ceci dans la proposition suivante.

Proposition 5.10. *Soit ξ une trajectoire de $\mathbb{R}^n$ sur $[0,T]$, et considérons le problème de poursuite pour le système de contrôle :*

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

où l'on veut minimiser le coût :

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

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

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

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

$$\begin{array}{ll} \dot{E} & = W - {}^tAE - EA - EBU^{-1}\,{}^tBE, \qquad E(T) = -Q, \\ \dot{h} & = -{}^tAh - E(A\xi - \dot{\xi} + r) - EBU^{-1}\,{}^tBh, \quad h(T) = 0, \end{array}$$

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

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

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

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

Remarque. Si $\dot{\xi} = A\xi + r$, i.e. la trajectoire de référence est solution du système sans contrôle, alors dans les notations précédentes on a $r_1 = 0$, et d'après les équations (5.34) on en déduit que $h(t)$ et $\alpha(t)$ sont identiquement nuls. On retrouve alors le cadre LQ de la section précédente. En fait :

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

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

On ajoute au problème précédent une variable de sortie :

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

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

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

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

Posons alors :

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

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

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

En raisonnant comme précédemment, on arrive au résultat suivant.

Proposition 5.11. *Soit ξ une trajectoire de $\mathbb{R}^p$ sur $[0,T]$, et considérons le problème de poursuite de la sortie r pour le système de contrôle avec sortie :*

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

où l'on veut minimiser le coût :

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

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

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

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

$$\begin{aligned}\dot{E} &= {}^tCWC - {}^tAE - EA - EBU^{-1}\,{}^tBE, \quad E(T) = -{}^tC(T)QC(T),\\ \dot{h} &= -{}^tCW\xi - {}^tAh - Er - EBU^{-1}\,{}^tBh, \quad h(T) = -{}^tC(T)Q\xi(T),\end{aligned}$$

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

$$-{}^tx(0)E(0)x(0) - 2{}^th(0)x(0) - \alpha(0),$$

où $\alpha(t)$ est solution de

$$\dot{\alpha} = {}^t\xi W\xi - 2{}^trh - {}^thBU^{-1}\,{}^tBh,\ \alpha(T) = {}^t\xi(T)Q\xi(T).$$

Remarque. Il existe d'autres variantes de ce problème, notamment le même problème que ci-dessus, sauf que le coût s'écrit :

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

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

Enfin, il y a aussi une autre variante du problème LQ, celle où la fonction g apparaissant dans le coût est linéaire en x. Nous laissons l'écriture de toutes ces variantes au lecteur, la méthode étant de toute façon la même que précédemment.

5.4.2 Filtre de Kalman déterministe

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

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

minimisant le coût

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

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

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

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

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

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

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

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

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

où :

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

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

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

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

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

Il faut donc déterminer un minimum de f. Notons tout d'abord que, la matrice Q étant par hypothèse définie positive, la matrice $\tilde{E}(0)$ est d'après le lemme 5.7 symétrique définie négative. En particulier la fonction f est strictement convexe et de ce fait admet un unique minimum. En un tel point on doit avoir $f'(x) = 0$, d'où $x = -\tilde{E}(0)^{-1}\tilde{h}(0)$.

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

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

on arrive au résultat suivant.

Proposition 5.12. *Soit $\xi(\cdot)$ une trajectoire définie sur $[0, T]$ à valeurs dans $\mathbb{R}^p$. On considère le problème de déterminer une trajectoire solution sur $[0, T]$ de*

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

minimisant le coût

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

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

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

et à la condition finale

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

où

$$\begin{array}{rcll} \dot{E} & = & {}^t CWC - {}^t AE - EA - EBU^{-1}\,{}^t BE, & E(0) = Q, \\ \dot{h} & = & -{}^t CW\xi - {}^t Ah - EBU^{-1}\,{}^t Bh, & h(0) = 0, \end{array}$$

et le coût minimal vaut alors :

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

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

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

$$\dot{F} = BU^{-1}\,{}^t B + AF + F\,{}^t A - F\,{}^t CWCF,\ F(0) = Q^{-1}.$$

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

$$\dot{z} = (A - F\,{}^t CWC)z + F\,{}^t CW\xi,\ z(0) = 0.$$

Finalement on arrive au résultat suivant.

Proposition 5.13. *Sous les hypothèses de la proposition 5.12, l'état final $x(T)$ de la solution optimale est égal à $z(T)$, où*

$$\begin{aligned} \dot{z} &= (A - F\,{}^tCWC)z + F\,{}^tCW\xi, & z(0) = 0, \\ \dot{F} &= BU^{-1}\,{}^tB + AF + F\,{}^tA - F\,{}^tCWCF, & F(0) = Q^{-1}. \end{aligned}$$

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

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

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

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

Il s'agit donc exactement du problème LQ suivant :

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

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

L'estimation optimale de l'état est donc égale à $z(T)$, voir proposition 5.13.

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

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

5.4.3 Régulation sur un intervalle infini et rapport avec la stabilisation

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

On cherche à déterminer une trajectoire solution de

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

minimisant le coût

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

où de même les matrices W et U sont constantes.

On a le résultat suivant.

Théorème 5.14. *On suppose que les matrices W et U sont symétriques définies positives, et que le système est contrôlable. Alors il existe une unique trajectoire minimisante pour ce problème, associée sur $[0, +\infty[$ au contrôle optimal :*

$$u(t) = U^{-1}\,{}^tBEx(t), \tag{5.37}$$

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

$${}^tAE + EA + EBU^{-1}\,{}^tBE = W. \tag{5.38}$$

De plus le coût minimal vaut $-{}^tx_0Ex_0$. Par ailleurs le système bouclé

$$\dot{x} = (A + BU^{-1}\,{}^tBE)x$$

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

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

Preuve.

▷ On sait déjà (voir proposition 5.3 et remarque 5.2) qu'il existe une unique trajectoire optimale, vérifiant les équations :

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

avec $u = U^{-1}\,{}^tB\,{}^tp$. De manière tout à fait similaire à la preuve du théorème 5.4 on montre, par un argument d'unicité, que $p(t) = {}^tx(t)E$, où E est solution, pourvu qu'elle existe, de l'équation (5.38). Il faut donc montrer l'existence d'une telle solution. C'est l'objet du lemme :

Lemme 5.15. *Il existe une unique matrice E symétrique définie négative solution de l'équation (5.38).*

Preuve.

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

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

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

$$\dot{E} = W - {}^tAE - EA - EBU^{-1}\,{}^tBE,\ E(T, T) = 0.$$

On sait que de plus le coût minimal est $C(T, u) = -{}^tx_0E(T, 0)^tx_0$. Posons alors $D(T, t) = -E(T, T - t)$. Il est bien clair que :

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

Cette équation étant en fait indépendante de T, on peut poser $D(t) = D(T, t)$, et $D(t)$ est solution de l'équation de Riccati ci-dessus sur $\mathbb{R}^+$. De plus pour tout $T > 0$ on a $D(T) = -E(T, 0)$, et comme la matrice W est symétrique définie positive on déduit du lemme 5.7 que $D(T)$ est symétrique définie positive.

Par ailleurs on a, pour tout $T > 0$: $C(T, u) = {}^tx_0D(T)x_0$. Il est clair que si $0 < t_1 \leq t_2$ alors $C(t_1, u) \leq C(t_2, u)$, et donc ${}^tx_0D(t_1)x_0 \leq {}^tx_0D(t_2)x_0$. Ceci est en fait indépendant de x_0, car l'équation de Riccati ne dépend nullement de la donnée initiale. Ainsi pour tout $x \in \mathbb{R}^n$ la fonction $t \mapsto {}^txD(t)x$ est croissante.

Montrons qu'elle est également majorée. Le système étant contrôlable, l'argument de la remarque 5.1 montre qu'il existe au moins un contrôle v sur $[0, +\infty[$ de coût fini. Comme le contrôle u est optimal, on en déduit que la fonction $t \mapsto C(t, u)$ est majorée (par $C(v)$). Pour tout $x \in \mathbb{R}^n$, la fonction $t \mapsto {}^txD(t)x$ étant croissante et majorée, on en déduit qu'elle converge. En appliquant cette conclusion aux éléments d'une base (e_i) de $\mathbb{R}^n$, on en déduit que chaque élément $d_{ij}(t)$ de la matrice $D(t)$ converge, car en effet :

$$d_{ij}(t) = {}^te_iD(t)e_j = \frac{1}{2}\,{}^te_i + e_jD(t)(e_i + e_j) - {}^te_iD(t)e_i - {}^te_jD(t)e_j.$$

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

Par ailleurs de l'équation différentielle vérifiée par D on déduit que $\dot{D}(t)$ converge, et cette limite est alors nécessairement nulle. En passant à la limite dans cette équation différentielle on obtient finalement l'équation de Riccati stationnaire (5.38).

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

□

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

$$\frac{d}{dt}V(x(t)) = -{}^t x(t)\left(W + EBU^{-1}\,{}^t BE\right)x(t).$$

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

□

Remarque. Le contrôle optimal s'écrit sous forme de boucle fermée $u = Kx$, avec $K = U^{-1}\,{}^t BE$. On retrouve le fait que si le système est contrôlable alors il est stabilisable par feedback linéaire (voir le théorème de placement de pôles). Cependant, alors que la méthode de stabilisation décrite par le théorème de placement de pôles consiste à réaliser un placement de pôles, ici la matrice K est choisie de manière à minimiser un certain critère. On parle de stabilisation par retour d'état optimal. C'est donc une méthode (parmi beaucoup d'autres) de stabilisation.

Remarque. En général l'équation (5.38) admet plusieurs solutions, mais elle n'admet qu'une seule solution symétrique définie négative.

Chapitre 6

Temps-optimalité pour les systèmes linéaires

6.1 Existence de trajectoires temps-optimales

Avant de traiter à proprement parler du sujet de ce chapitre, prouvons tout d'abord un résultat à propos de l'ensemble atteignable en temps fini pour le le système de contrôle dans $\mathbb{R}^n$:

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

où les contrôles u sont à valeurs dans un compact d'intérieur non vide $\Omega \subset \mathbb{R}^m$.

Théorème 6.1. *Considérons le système de contrôle linéaire dans $\mathbb{R}^n$:*

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

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

Remarque. La convexité de $\mathcal{A}(x_0, t)$ est facile à établir si Ω est convexe (petit exercice laissé au lecteur). Pourtant, et ce résultat est surprenant, la conclusion de ce théorème est encore vraie si Ω n'est pas convexe et nous admettrons ce résultat. Mentionnons juste le fait que la preuve de la convexité nécessite un lemme de Lyapunov en théorie de la mesure. Ce résultat implique en particulier le corollaire suivant très utile.

Corollaire 6.2. *Si on note $\mathcal{A}_\Omega(x_0, t)$ l'ensemble accessible depuis x_0 en temps t pour des contrôles à valeurs dans Ω, alors on a :*

$$\mathcal{A}_\Omega(x_0, t) = \mathcal{A}_{Conv(\Omega)}(x_0, t),$$

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

Démonstration du théorème 6.1. Il reste à montrer que $\mathcal{A}(x_0,t)$ est compact. Grâce au corollaire précédent, on peut donc supposer sans perdre en généralité que, dans la suite de la preuve, Ω est convexe. Cela simplifie grandement la preuve de la compacité.

Il s'agit de montrer que toute suite (x_n) de points de $\mathcal{A}(x_0,t)$ admet une sous-suite convergente. Pour tout entier n soit u_n un contrôle reliant x_0 à x_n en temps t, et soit $x_n(\cdot)$ la trajectoire correspondante. On a donc :

$$x_n = x_n(t) = M(t)x_0 + \int_0^t M(t)M(s)^{-1}(B(s)u_n(s) + r(s))ds. \tag{6.1}$$

Par définition les contrôles u_n sont à valeurs dans le compact convexe Ω, et par conséquent la suite (u_n) est bornée dans $L^\infty([0,t],\Omega)$ qui est compact pour la topologie faible-$*$. En conséquence (u_n) converge faiblement , à une sous-suite près, vers une fonction $u \in L^\infty([0,t],\Omega)$. Il est alors immédiat de montrer que (x_n) converge, à une sous-suite près, vers

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

ce qui prouve la compacité de $\mathcal{A}(x_0,t)$.

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

$$\forall t_1, t_2 \in [0,T] \quad |t_1 - t_2| \leq \delta \Rightarrow d(\mathcal{A}(t_1), \mathcal{A}(t_2)) \leq \varepsilon,$$

où on note pour simplifier $\mathcal{A}(t) = \mathcal{A}(x_0,t)$, et où :

$$d(\mathcal{A}(t_1), \mathcal{A}(t_2)) = \sup \lambda(\sup_{y\in\mathcal{A}(t_2)} d(y, \mathcal{A}(t_1)), \sup_{y\in\mathcal{A}(t_1)} d(y, \mathcal{A}(t_2))\rho).$$

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

1. $\forall y \in \mathcal{A}(t_2) \quad d(y, \mathcal{A}(t_1)) \leq \varepsilon$,
2. $\forall y \in \mathcal{A}(t_1) \quad d(y, \mathcal{A}(t_2)) \leq \varepsilon$.

Montrons juste le premier point (2. étant similaire). Soit $y \in \mathcal{A}(t_2)$. Il suffit de montrer que :

$$\exists z \in \mathcal{A}(t_1) \ / \ d(y,z) \leq \varepsilon.$$

Par définition de $\mathcal{A}(t_2)$, il existe un contrôle $u \in L^\infty([0,T],\Omega)$ tel que la trajectoire associée à u, partant de x_0, vérifie : $x(t_2) = y$, voir figure 6.1. On va voir que $z = x(t_1)$

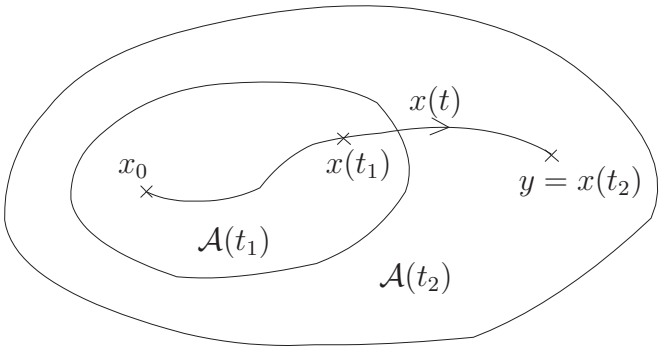


FIGURE 6.1 –

convient. En effet on a :

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

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

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

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

où les contrôles u sont à valeurs dans un compact d'intérieur non vide $\Omega \subset \mathbb{R}^m$. Soient $x_0, x_1 \in \mathbb{R}^n$. On suppose que x_1 est accessible depuis x_0, i.e. il existe au moins une trajectoire reliant x_0 à x_1. Parmi toutes les trajectoires reliant x_0 à x_1 on aimerait caractériser celles qui le font en temps minimal t^*, voir figure 6.2.

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

$$t^* = \inf\{t > 0 \ / \ x_1 \in \mathcal{A}(x_0, t)\}.$$

Ce temps t^* est bien défini car, d'après le théorème 6.1, $\mathcal{A}(x_0, t)$ varie continûment avec t, donc $\{t > 0 \ / \ x_1 \in \mathcal{A}(x_0, t)\}$ est fermé dans $\mathbb{R}$. En particulier cette borne inférieure est un minimum.

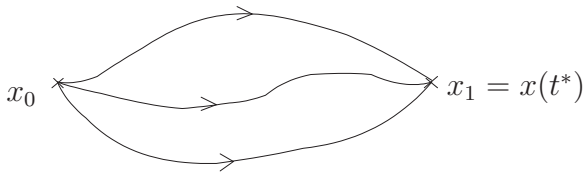


FIGURE 6.2 –

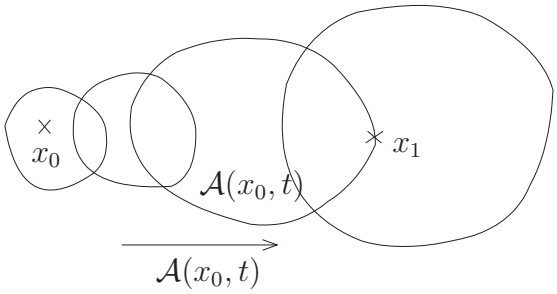


FIGURE 6.3 – Temps minimal

Le temps $t = t^*$ est le premier temps pour lequel $\mathcal{A}(x_0, t)$ contient x_1 (voir figure 6.3). D'autre part, on a nécessairement :

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

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

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

Théorème 6.3. *Si le point x_1 est accessible depuis x_0 alors il existe une trajectoire temps-minimal reliant x_0 à x_1.*

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

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

Définition 6.1. Le contrôle u est dit *extrémal* sur $[0,t]$ si la trajectoire du système $\dot{x}(t) = A(t)x(t) + b(t)u(t) + r(t)$ associée à u vérifie : $x(t) \in \partial\mathcal{A}(x_0,t)$.

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

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

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

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

Théorème 6.4. *Considérons le système de contrôle linéaire :*

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

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

$$p(t)B(t)u(t) = \max_{v\in\Omega} p(t)B(t)v \tag{6.3}$$

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

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

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

Preuve.

▷ On a vu que $\mathcal{A}_\Omega(x_0,T) = \mathcal{A}_{\mathrm{Conv}(\Omega)}(x_0,T)$, et par conséquent on peut supposer que Ω est convexe. Si u est extrémal sur $[0,T]$, soit x la trajectoire associée à u. On a $x(T) \in \partial\mathcal{A}(x_0,T)$. Par convexité de $\mathcal{A}(x_0,T)$, il existe d'après le théorème du convexe un hyperplan séparant au sens large $x(T)$ et $\mathcal{A}(x_0,T)$. Soit p_T un vecteur normal à cet hyperplan, voir figure 6.4.

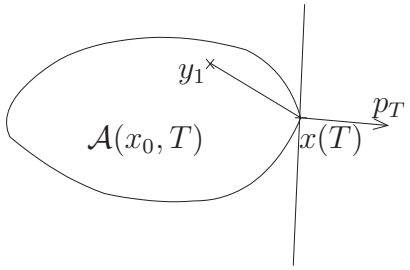


FIGURE 6.4 –

D'après le théorème du convexe :

$$\forall y_1 \in \mathcal{A}(x_0, T) \quad p_T(y_1 - x(T)) \leq 0. \tag{6.4}$$

Par définition de $\mathcal{A}(x_0, T)$, il existe un contrôle u_1 tel que la trajectoire associée $y(t)$ vérifie : $y_1 = y(T)$. L'inégalité (6.4) se réécrit :

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

D'où :

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

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

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

et donc :

$$\int_0^T p(s)B(s)u_1(s)ds \leq \int_0^T p(s)B(s)u(s)ds. \tag{6.5}$$

Si (6.3) n'est pas vraie alors

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

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

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

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

Comme u_1 est à valeurs dans Ω, l'inégalité (6.5) est vraie, alors que par ailleurs la définition de u_1 conduit immédiatement à l'inégalité stricte inverse, d'où la contradiction. Par conséquent (6.3) est vraie.

Réciproquement, supposons qu'il existe un vecteur adjoint tel que le contrôle u satisfait (6.3). Notons $x(\cdot)$ la trajectoire associée à u. On voit facilement en remontant le raisonnement précédent que :

$$\forall y_1 \in \mathcal{A}(x_0, T) \quad p(T)(y_1 - x(T)) \leq 0. \tag{6.6}$$

Raisonnons alors par l'absurde, et supposons que $x(T) \in \mathrm{Int}\mathcal{A}(\mathrm{x_0}, \mathrm{T})$. Alors il existe un point y_1 de $\mathcal{A}(x_0, T)$ qui est sur la demi-droite d'origine $x(T)$ et de direction $p(T)$, voir figure 6.5. Mais alors : $p(T)(y_1 - x(T)) > 0$, ce qui contredit (6.6). Donc $x(T) \in \partial\mathcal{A}(x_0, T)$, et donc u

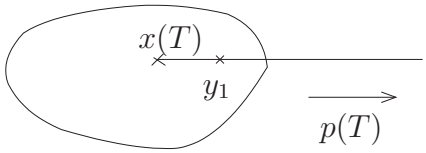


FIGURE 6.5 –

est extrémal.

□

Remarque. Si u est extrémal sur $[0, T]$ alors u est aussi extrémal sur $[0, t]$ pour tout $t \in [0, T]$, et de plus $p(t)$ est un vecteur normal extérieur à $\mathcal{A}(x_0, t)$. Cela découle facilement de la preuve et de la propriété (6.3).

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

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

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

Proposition 6.5. *Considérons dans $\mathbb{R}^n$ le système linéaire autonome $\dot{x}(t) = Ax(t) + Bu(t)$, où $B \in \mathbb{R}^n$ et $|u(t)| \leq 1$. Supposons que la paire (A, B) satisfait la condition de Kalman.*

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

PREUVE.

▷ Comme le système est commndable, le système peut s'écrire sous forme équivalente de Brunovski, et il est alors équivalent à une équation différentielle scalaire d'ordre n de la forme :

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

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

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

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

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

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

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

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

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

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

□

6.3 Exemple : Synthèse optimale pour le problème de l'oscillateur harmonique

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

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

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

6.3.1 Contrôlabilité du système

Le système s'écrit :

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

On a facilement $\mathrm{rg}(\mathrm{B}, \mathrm{AB}) = 2$, et par ailleurs les valeurs propres de A sont de partie réelle nulle, donc d'après le critère de Kalman, le système est contrôlable à 0, i.e. il existe des contrôles tels que les trajectoires associées relient X_0 à 0, ce qui répond à la première question.

6.3.2 Interprétation physique

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

6.3.3 Calcul du contrôle optimal

D'après le paragraphe précédent, il existe des contrôles permettant de relier X_0 à 0. On cherche maintenant à le faire en temps minimal. Pour cela, on applique le théorème 6.4 :

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

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

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

$$\left\{\begin{array}{l} \dot{x} = y \\ \dot{y} = -x - 1 \end{array}\right. \tag{6.7}$$

— Si $u = +1$:

$$\left\{\begin{array}{l} \dot{x} = y \\ \dot{y} = -x + 1 \end{array}\right. \tag{6.8}$$

La trajectoire optimale finale, reliant X_0 à 0, sera constituée d'arcs successifs, solutions de (6.7) et (6.8).

Solutions de (6.7). On obtient facilement $(x+1)^2+y^2 = \text{cste} = \mathrm{R}^2$, donc les courbes solutions de (6.7) sont des cercles centrés en $x=-1, y=0$, et de période 2π (en fait : $x(t) = -1 + R\cos t, y(t) = R\sin t$).

Solutions de (6.8). On obtient : $x(t) = 1 + R\cos t$, $y(t) = R\sin t$. Les solutions sont des cercles centrés en $x=1, y=0$, de période 2π.

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

Quitte à changer t en $-t$, nous allons raisonner à l'envers, et construire la trajectoire optimale menant de 0 à X_0. Pour cela, nous allons considérer toutes les trajectoires optimales partant de 0, et nous sélectionnerons celle qui passe par X_0.

C'est en faisant varier $p(0)$ que l'on change de trajectoire optimale : en effet, $p(0)$ détermine $p(t)$ pour tout t d'après le théorème de Cauchy-Lipschitz, ce qui détermine un contrôle optimal $u(t)$, et donc une trajectoire optimale.

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

— si $p_1(0) = 1, p_2(0) = 0$: alors $p_2(t) = -\sin t$, donc sur $]0,\pi[$ on a $u(t) = \text{signe}\mathrm{p_2(t)} = -1$. La trajectoire optimale correspondante, partant de 0, suit donc pendant un temps π l'arc de cercle Γ_- solution de (6.7), passant par 0, voir figure 6.6.

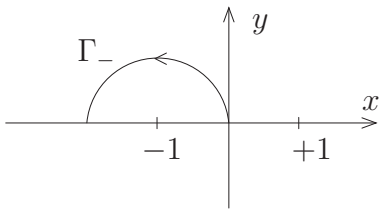


FIGURE 6.6 –

— si $p_1(0) = -1, p_2(0) = 0$: alors $p_2(t) = \sin t$, donc sur $]0,\pi[$ on a $u(t) = \text{sgn}\ \mathrm{p_2(t)} = +1$. La trajectoire optimale correspondante, partant de 0, suit donc pendant un temps π l'arc de cercle Γ_+ solution de (6.8), passant par 0, voir figure 6.7.

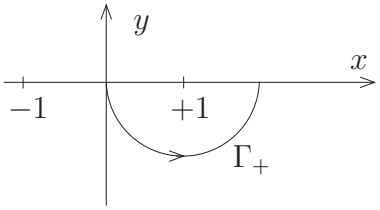


FIGURE 6.7 –

— Pour tout autre choix de $p(0)$ tel que $p_2(0) > 0$, la trajectoire optimale correspondante part de l'origine en suivant Γ_+ *jusqu'à ce que* $p_2(t) = 0$. *Au-delà de ce point,* $p_2(t)$ *change de signe, donc le contrôle commute et prend la valeur* -1, *pendant une durée* π *(i.e. jusqu'à ce que* $p_2(t)$ *change à nouveau de signe). La trajectoire optimale doit alors être solution de (6.7), en partant de ce point de commutation* M, *et doit donc suivre un arc de cercle centré en* $x = -1, y = 0$, *pendant un temps* π *(c'est donc un demi-cercle, vu la paramétrisation des courbes de (6.7)), voir figure 6.8.*

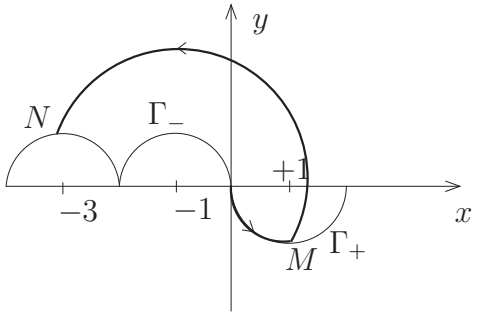


FIGURE 6.8 –

La trajectoire optimale rencontre un deuxième point de commutation N *lorsque à nouveau* $p_2(t)$ *change de signe. On remarque que* M *et* N *sont symétriques par rapport au point* $x = -1, y = 0$ *(en effet ce sont les extrémités d'un demi-cercle centré en ce point). Le point* M *appartenant au demi-cercle* Γ_+, *le point* N *appartient au demi-cercle image de* Γ_+ *par la symétrie par rapport au point* $x = -1, y = 0$ *qui est aussi, comme on le voit facilement, le translaté à gauche de* Γ_- *par la translation de vecteur* $(-2, 0)$.

Poursuivons alors notre raisonnement : on se rend compte que les points de commutation de cette trajectoire optimale partant de 0 sont situés sur la courbe W construite de la manière suivante : W est l'union de tous les translatés à gauche de Γ_- par la translation précédente, et aussi de tous les translatés à droite de Γ_+, voir figure 6.9.

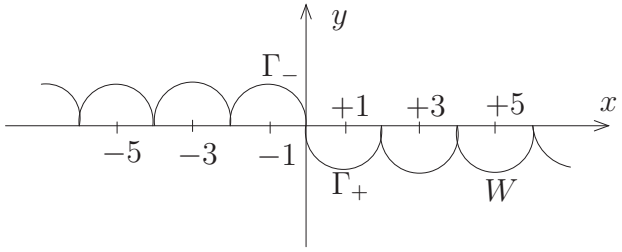


FIGURE 6.9 – Ensemble W

Les trajectoires optimales sont alors construites de la manière suivante : on part de 0 et on suit un morceau de Γ_+ ou Γ_-, jusqu'à un premier point de commutation. Si par exemple on était sur Γ_+, alors partant de ce point on suit un arc de cercle centré en $x = -1, y = 0$, au-dessus de W, jusqu'à ce qu'on rencontre W. De ce deuxième point de commutation, on suit un arc de cercle centré en $x = +1, y = 0$ jusqu'à rencontrer W en un troisième point de commutation, etc, voir figure 6.10.

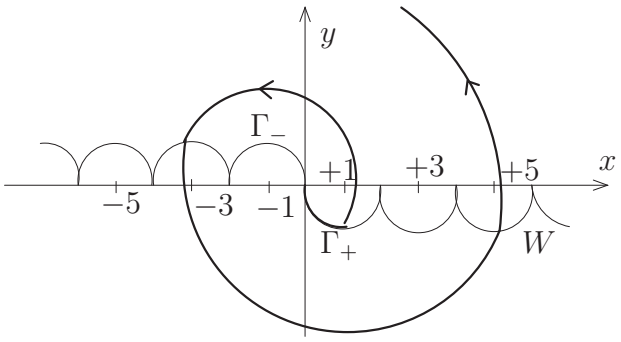


FIGURE 6.10 –

On est maintenant en mesure de répondre à la deuxième question, du moins graphiquement : le but est de relier 0 et X_0 par une trajectoire optimale. La théorie prévoit qu'on peut effectivement le faire. Une trajectoire partant de 0 est, comme on vient de le voir ci-dessus, déterminée par deux choix :

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

Si maintenant on se donne un point $X_0 = (x_0, y_0)$ du plan de phase, on peut déterminer graphiquement ces deux choix, et obtenir un tracé de la trajectoire optimale, voir figure 6.11. Dans la pratique il suffit d'inverser le temps, i.e. de partir du point final et d'atteindre le point initial.

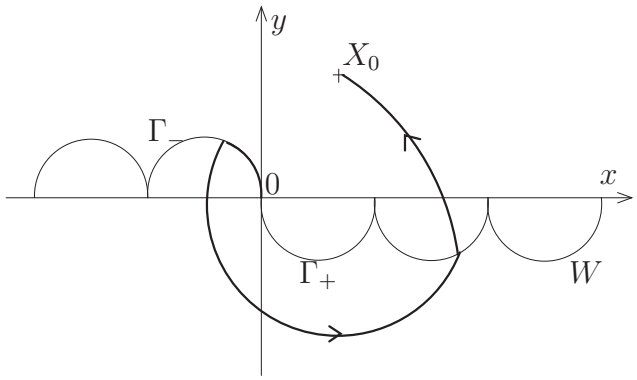


FIGURE 6.11 – Synthèse optimale

Chapitre 7

Contrôle optimal non-lineaire

Dans le chapitre suivant, nous allons énoncer une forme générale du Principe du Maximum de Pontryagin (en bref PMP) dans le cas non linéaire. Ce théorème constitue l'instrument fondamental pour calculer les trajectoires optimales. Nous l'appliquerons ensuite à deux classes de problèmes : le problème sous-riemannien et le problème de temps minimal pour des systèmes commandés affines avec contrôles bornés.

7.1 Enoncé général du Principe du maxminum de Pontryagin

On considère le système commandé défini sur $\mathbb{R}^n$ comme suit :

$$\dot{x}(t) = f(x(t), u(t)), \tag{7.1}$$

où $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ est de classe C^∞ et où les contrôles sont des applications mesurables, bornées, définies sur un intervalle de $\mathbb{R}^+$ et à valeurs dans $U \subset \mathbb{R}^m$. On note $\mathcal{U}$ l'ensemble des contrôles admissibles u dont les trajectoires associées relient un point initial de x_0 à un point final de x_1.
Par ailleurs on définit $C(t, u)$, le coût d'un contrôle u sur $[0, t]$, par :

$$C(t, u) = \int_0^t f^0(x(s), u(s))ds.$$

où $f^0 : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ est C^∞, et $x(\cdot)$ est la trajectoire solution de (7.1) associée au contrôle u.

On considère le problème de contrôle optimal suivant : déterminer une trajectoire reliant x_0 à x_1 et minimisant le coût C. Le temps final peut être fixé ou non. Si une telle trajectoire existe, on dira qu'elle est optimale et que le contrôle associé l'est aussi.

Théorème 7.1 (Principe du maximum de Pontryagin (PMP))**.** *Si le contrôle $u \in \mathcal{U}$ associé à la trajectoire $x(\cdot)$ est optimal sur $[0, T]$, alors il existe une application $p(\cdot)$:*

$[0,T] \longrightarrow \mathbb{R}^n$ *lipschitzienne appelée vecteur adjoint (ou covecteur), et un réel* $p^0 \leq 0$*, tels que le couple* $(p(\cdot), p^0)$ *est non trivial, et tels que les équations suivantes sont vérifiées pour presque tout* $t \in [0,T]$ *:*

$$\dot{x}(t) = \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)), \tag{7.2}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)), \tag{7.3}$$

où $H(x,p,p^0,u) = p \cdot f(t,x,u) + p^0 f^0(t,x,u)$ *est l'Hamiltonien du système.*

De plus, on a la condition de maximisation suivante, vraie presque partout sur $[0,T]$ *:*

$$H(x(t), p(t), p^0, u(t)) = \max_{v \in U} H(x(t), p(t), p^0, v). \tag{7.4}$$

Enfin,

$$H(x(t), p(t), p^0, u(t)) = const, \tag{7.5}$$

et cette constante est nulle si le temps final pour joindre la cible x_1 *n'est pas fixé.*

Il faut remarquer que le covecteur p est un vecteur ligne.

Remarque. Dans le cas où $U = \mathbb{R}^m$, i.e. lorsqu'il n'y a pas de contrainte sur le contrôle, la condition de maximum (7.4) peut être remplacée par la condition plus faible $\frac{\partial H}{\partial u} = 0$. On parle alors de principe du maximum faible (voir le paragraphe 7.4). Par ailleurs, le lecteur attentif de ce cours aura noté que, dans le cas de la commande LQ, les deux versions du PMP sont équivalentes.

Remarque. Il est important de remarquer que le PMP n'est qu'une condition *nécessaire* d'optimalité. Cela signifie pour obtenir en pratique les trajectoires optimales on doit d'abord montrer l'existence de ces trajectoires puis appliquer le PMP.

Définition 7.1. Une trajectoire *extrêmale* du problème de contrôle optimal est un solution des équations (7.2) et (7.4). Si $p_0 = 0$, on dit que l'extrêmale est *anormale*, et si $p^0 \neq 0$ l'extrêmale est dite *normale*.

Remarque. Dans le cas général et contrairement au cas linéaire, l'équation différentielle pour le covecteur dépend de la trajectoire. Appliquer le le PMP s'avère donc plus compliqué.

7.2 Le problème sous-riemannien

Une classe importante de problèmes de contrôle optimal est celle des problèmes sous-riemanniens, c'est-à-dire que la dynamique et le coût sont respectivement linéaire et quadratique par rapport à la commande. De plus, on suppose que le coût ne dépend que de la commande,

$$\dot{x} = \sum_{i=1}^{m} u_i \; F_i(x), \quad \int_0^T \sum u_i(t)^2 \; dt, \quad x(0) = x_0, \quad , x(T) = x_1. \tag{7.6}$$

Ici, $x \in \mathbb{R}^n$, F_i $(i = 1, \dots m)$ sont des champs de vecteurs réguliers, $m \geq 2$ et le temps final T est fixé.

On a alors,

$$H(x, p, p^0, u) = \sum_{i=1}^{m} u_i \; p \cdot F_i(x) + p^0 \sum u_i^2.$$

Dans de nombreuses applications intéressantes, il est possible de prouver qu'il n'y a pas d'extrêmales anormales et dans la suite de ce paragraphe, nous le supposerons. On peut alors normaliser l'Hamiltonien en choisissant $p^0 = -1/2$. D'après la condition de maximalité, on a

$$\frac{\partial H}{\partial u}(x(t), p(t), -1/2, u(t)) = 0,$$

et on en déduit $u_i(t) = p(t) \cdot F_i(t)$. En définissant $\mathcal{H}(x, p) := H(x, p, 1/2, p \cdot F_i(x)) = \frac{1}{2} \sum_{i=1}^{m} \big(p \cdot F_i(x)\big)^2$, on obtient les équations hamiltoniennes suivantes,

$$\dot{x}(t) \quad = \tfrac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)) = \tfrac{\partial \mathcal{H}}{\partial p}(x(t), p(t)) \tag{7.7}$$

$$\dot{p}(t) \quad = -\tfrac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)) = -\tfrac{\partial \mathcal{H}}{\partial x}(x(t), p(t)). \tag{7.8}$$

Une conséquence très importante des équations précédentes est que les trajectoires extrêmales (et donc les trajectoires optimales lorsqu'elles existent) sont de classe C^∞.

Remarquons aussi que $\mathcal{H}(x(t), p(t)) = \sum_{i=1}^{m} \big(u_i(t)\big)^2$ est une constante du mouvement.

Remarque. Soit $(x(\cdot), p(\cdot))$ une solution de (7.7), (7.8), correspondant aux contrôles $(u_i(\cdot))_{1 \leq i \leq m}$, telle que $x(0) = x_0$, $x(T) = x_1$ et $\mathcal{H}(x(t), p(t)) = \beta > 0$. Alors, pour tout $\alpha > 0$, la paire $(\bar{x}(t), \bar{p}(t)) := (x(\alpha t), \alpha p(\alpha t))$ est encore solution de (7.7), (7.8), correspondant à $(\alpha u_i(\cdot))_{1 \leq i \leq m}$ avec $x(0) = x_0$, $x(T/\alpha) = x_1$ et $\mathcal{H}(\bar{x}(t), \bar{p}(t)) = \alpha^2 \beta > 0$. En d'autres termes, la ligne de niveau de l'Hamiltonien est déterminée par le temps final T pour lequel la cible est atteinte.

7.3 Temps minimum pour un système affine bidimensionnel

Dans ce paragraphe, nous allons présenter la théorie du temps minimum pour les systemes commandés affines en la commande et définis sur $\mathbb{R}^2$:

$$\dot{x} = F(x) + uG(x), \quad x \in \mathbb{R}^2, \quad |u| \leq 1. \tag{7.9}$$

Il s'agit d'atteindre chaque point du plan en temps minimum en partant de l'origine.

Dans ce cas, l'Hamiltonien prend une forme particulièrement simple $H(x, p, p^0, u) = p \cdot F(x) + u \; p \cdot G(x) + p^0$. En appliquant le PMP, on obtient donc :

(i) $\dot{p}(t) = -p(t) \cdot (\nabla F + u(t) \nabla G)(x(t))$,

(ii) $p(t) \cdot F(x(t)) + u(t) \; p(t) \cdot G(x(t)) + p^0 = 0$,

(iii) $u(t)\ p(t)\cdot G(x(t)) = \max_{v\in[-1,1]} v\ p(t)\cdot G(x(t))$.

D'une manière analogue au cas des systèmes linéaires, nous allons définir la fonction de commutation.

Définition 7.2. (fonction de commutation) Soit $(x(\cdot),p(\cdot)) : [0,\tau] \to \mathbb{R}^2\times\mathbb{R}^2$ une trajectoire extrêmale avec le covecteur correspondant. La fonction de commutation associée est définie par $\phi(\cdot) := p(\cdot)\cdot G(x(\cdot)$. Notons que $\phi(\cdot)$ est absolument continue.

Le lemme suivant est une conséquence facile des équations du PMP.

Lemme 7.2. *Soit $(x(\cdot),p(\cdot)) : [0,\tau] \to \mathbb{R}^2\times\mathbb{R}^2$ une trajectoire extrêmale avec le covecteur correspondant et $\phi(\cdot)$ la fonction de commutation associée. Si ϕ n'a que des zéros isolés, alors $u(t) = sgn(\phi(t))$ dans $[0,\tau]$.*

Définition 7.3. (Trajectoire bang-bang et trajectoire singulière) Une trajectoire qui satisfait les hypothèses du lemme précédent est appelée une trajectoire *bang-bang*.

A l'opposé, on dira qu'une trajectoire extrêmale est *singulière* sur $[t_1,t_2]$ si la fonction de commutation associée est identiquement nulle sur $[t_1,t_2]$.

Comme ϕ est absolument continue, on peut toujours la dériver. On a alors :

Lemme 7.3. *Soit $(x(\cdot),p(\cdot)) : [0,\tau] \to \mathbb{R}^2\times\mathbb{R}^2$ une trajectoire extrêmale avec son covecteur correspondant et $\phi(\cdot)$ la fonction de commutation associée. Alors $\phi(\cdot)$ est de classe $\mathcal{C}^1$ et*

$$\dot\phi(t) = p(t)\cdot[F,G](x(t)).$$

Preuve.

▷ En utilisant le PMP, on a pour presque tout t :

$$\begin{aligned}\dot\phi(t) &= \frac{d}{dt}(p(t)\cdot G(x(t))) = \dot p(t)\cdot G(x(t)) + p\cdot\dot G(x(t))\\ &= -p(t)(\nabla F + u(t)\nabla G)(x(t))\cdot G(x(t)) + p\cdot\nabla G(x(t))(F+u(t)G)(x(t))\\ &= p(t)\cdot[F,G](x(t)). \qquad (7.10)\end{aligned}$$

□

L'étape suivante consiste à déterminer quand le contrôle change de signe ou peut prendre des valeurs en $]-1,+1[$.

7.3.1 Trajectoires singulières et détermination des commutations

Afin de déterminer les trajectoires singulières ainsi que les commutations, on définit $\Delta_A(\cdot)$ et $\Delta_B(\cdot)$, les fonctions suivantes :

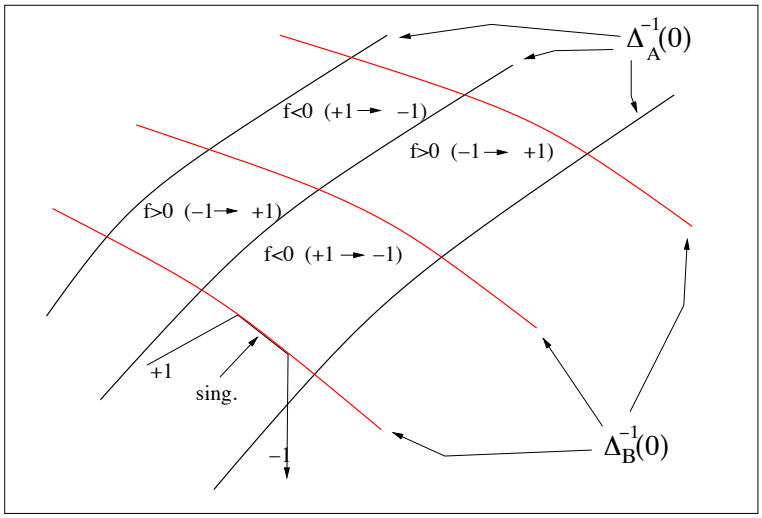


FIGURE : Possible commutations dans les composantes connexes de $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$ en relation avec le signe def_S, et un exemple de trajectoire extrêmale contenant un arc singulier.

$$\Delta_A(x) := \det(F(x), G(x)) = F_1(x)G_2(x) - F_2(x)G_1(x), \tag{7.11}$$
$$\Delta_B(x) := \det(G(x), [F,G](x)) = G_1(x)[F,G]_2(x) - G_2(x)[F,G]_1(x). \tag{7.12}$$

L'ensemble des zéros de ces fonctions $\Delta_A^{-1}(0), \Delta_B^{-1}(0)$, sont respectivement l'ensemble des points où F et G sont parallèles et l'ensemble des points où G est parallèle à $[F,G]$. Comme on va le voir, ces lieux sont cruciaux pour ce qui est de la construction des trajectoires optimales. En effet, en supposant que ces lieux sont inclus dans des courbes lisses, on montre que

— dans chaque composante connexe de $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$, toute trajectoire extrêmale est bang-bang avec au plus une commutation. De plus, le contrôle commute de -1 to $+1$ si $f_S := -\Delta_B \setminus \Delta_A > 0$ et de $+1$ to -1 si $f_S < 0$;

— le support des trajectoires singulières est contenu dans l'ensemble $\Delta_B^{-1}(0)$. (Remarquons que l'on ne sait rien encore sur les valeurs prises par le contrôle le long d'une trajectoire singulière.)

Afin de se convaincre des résultats précédents, on commence par étudier ce qui se passe sur $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$. Pour cela, on considère la définition suivante.

Définition 7.4. Les points de $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$ sont appelés *points ordinaires*. En un tel point, $F(x)$ et $G(x)$ sont linéairement indépendants et on définit les fonctions scalaires f_S et g_S comme les coefficients de la combinaison linéaire $[F,G](x) = f_S(x)F(x) + g_S(x)G(x)$.

On a alors :

Lemme 7.4. *En un point ordinaire x,*

$$f_S(x) = -\frac{\Delta_B(x)}{\Delta_A(x)}. \tag{7.13}$$

Preuve.

▷ Par un simple calcul,

$$\begin{aligned}\Delta_B(x) &= \det\big(G(x), [F,G](x)\big) = \det(G(x), f_S(x)F(x) + g_S(x)G(x)) \\ &= f_S(x)\det\big(G(x), F(x)\big) = -f_S(x)\Delta_A(x).\end{aligned}$$

□

Sur $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$, la structure des trajectoires optimales est simple :

Théorème 7.5. *Soit Ω un ouvert de $\mathbb{R}^2 \setminus (\Delta_A^{-1}(0) \cup \Delta_B^{-1}(0))$. Alors, toute trajectoire extrêmale $x(.)$ du système commandé (7.9) contenue dans Ω est bang-bang avec au plus une commutation. De plus, si $f_S > 0$ (resp. $f_S < 0$) sur Ω, alors la commande associée à $x(.)$ est ou bien constante égale à 1, ou -1, ou bien commute de -1 à 1 (resp. de 1 à -1).*

Preuve.

▷ On ne traitera que le cas où $f_S > 0$ sur Ω, l'autre cas étant similaire. Soit $(x(\cdot), p(\cdot))$ une trajectoire extrêmale avec le covecteur correspondant telle que $x(\cdot)$ est contenue dans Ω. Soit un temps $\bar{t}$ tel que $p(\bar{t}) \cdot G(x(\bar{t})) = 0$. Alors, on a

$$\dot{\phi}(\bar{t}) = p(\bar{t}) \cdot [F,G](x(\bar{t})) = p(\bar{t}) \cdot (fF + gG)(x(\bar{t})) = f(x(\bar{t}))\ p(\bar{t}) \cdot F(x(\bar{t})).$$

D'après le PMP, on obtient que $p(\bar{t}) \cdot F(x(\bar{t})) \geq 0$. Cela implique que $\dot{\phi} > 0$, puisque $F(x(\bar{t}))$ et $G(x(\bar{t}))$ sont linéairement indépendants. Cela prouve que ϕ a au plus un seul zéro et on conclue.

□

Enfin, le résultat annoncé sur les trajectoires singulières est une conséquence du lemme suivant.

Lemme 7.6. *Soit $(x(\cdot), p(\cdot))$ une trajectoire extrêmale avec le covecteur correspondant et $\phi(\cdot)$ la fonction de commutation associée. Supposons que $\phi(t) \equiv 0$ sur $[a,b]$. Alors, $x([a,b]) \subset \Delta_B^{-1}(0)$.*

Preuve.

▷ Comme $\phi(t) = p(t) \cdot G(x(t)) \equiv 0$ et $\dot{\phi}(t) = p(t) \cdot [F,G](x(t)) \equiv 0$, il résulte que G est parallèle à $[F,G]$ le long de $x(.)$ sur $[a,b]$. On conclue.

□

7.4 Principe du maximum de Pontryagin faible

Dans cette section, nous prouvons le Principe du maximum de Pontryagin faible, c'est-à-dire :

Théorème 7.7 (Principe du maximum de Pontryagin faible)**.** *On considère le système de contrôle optimal :*

$$\dot{x}(t) = f(x(t), u(t)),, \quad x(0) = x_0, \quad x(T) = x_1, \quad T \text{ fixé} \tag{7.14}$$

$$\int_0^T f^0(x(s), u(s))ds \to \min \tag{7.15}$$

où $f, f_0 : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ sont de classe C^∞ et où les contrôles sont des applications mesurables et bornées définies sur un intervalle de $\mathbb{R}^+$ et à valeurs dans $\mathbb{R}^m$.

Si le contrôle $u(\cdot)$ associé à la trajectoire $x(\cdot)$ est optimal sur $[0, T]$, alors il existe une application $p(\cdot) : [0, T] \longrightarrow \mathbb{R}^n$ absolument continue, et un réel $p^0 \leq 0$, tels que le couple $(p(\cdot), p^0)$ est non trivial, et tels que les équations suivantes sont vérifiées pour presque tout $t \in [0, T]$:

$$\dot{x}(t) = \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)), \tag{7.16}$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)), \tag{7.17}$$

$$\frac{\partial H}{\partial u}(x(t), p(t), p^0, u(t)) = 0, \tag{7.18}$$

où $H(x, p, p^0, u) = p \cdot f(t, x, u) + p^0 f^0(t, x, u)$.

Dans ce but, il faut introduire le concept d'application entrée-sortie et de contrôles singuliers.

Considérons de nouveau le système de contrôle général :

$$\dot{x}(t) = f(x(t), u(t)), \tag{7.19}$$

où f est une application de classe C^∞ de $\mathbb{R}^{n+m}$ dans $\mathbb{R}^n$, et où les contrôles sont des applications mesurables et bornés définies sur un intervalle de $\mathbb{R}^+$ et à valeurs dans $\mathbb{R}^m$.

Définition 7.5. Soit $T > 0$. *L'application entrée-sortie en temps T du système contrôlé (7.19) initialisé à 0 est l'application :*

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

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

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

7.4.1 Régularité de l'application entrée-sortie

En toute généralité on a le résultat suivant :

Proposition 7.8. *Considérons le système (7.19) où f est C^p, $p \geq 1$, et soit $\mathcal{U} \subset L^\infty([0,T],\mathbb{R}^m)$ le domaine de définition de E_T, c'est-à-dire l'ensemble des contrôles dont la trajectoire associée est bien définie sur $[0,T]$. Alors $\mathcal{U}$ est un ouvert de $L^\infty([0,T],\mathbb{R}^m)$, et E_T est C^p au sens L^∞.*

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

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

Le système de contrôle linéaire :

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

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

$$dE_T(u).v \ = \ y_v(T) \ = \ M(T)\int_0^T M^{-1}(s)B(s)v(s)ds \tag{7.20}$$

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

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

$$\begin{aligned} \frac{d}{dt}(x+\delta x)(t) &= f(x(t)+\delta x(t), u(t)+\delta u(t)) \\ &= f(x(t),u(t)) + \frac{\partial f}{\partial x}(x(t),u(t))\delta x(t) + \frac{\partial f}{\partial u}(x(t),u(t))\delta u(t) \\ &\quad + o(\delta x(t), \delta u(t)). \end{aligned}$$

Comme $\dot{x}(t) = f(x(t),u(t))$, on a donc :

$$\frac{d}{dt}(\delta x)(t) = \frac{\partial f}{\partial x}(x(t),u(t))\delta x(t) + \frac{\partial f}{\partial u}(x(t),u(t))\delta u(t) + o(\delta x(t), \delta u(t)).$$

Par ailleurs, on a aussi $x(0) + \delta x(0) = x_0 = x(0)$, donc $\delta x(0) = 0$.

En ne retenant que les termes du premier ordre (c-à-d en négligeant le "o"), on obtient le résultat. □

Définition 7.6. Soit $u(.)$ un contrôle défini sur $[0,T]$ tel que sa trajectoire associée $x(.)$ issue de $x(0)=x_0$ est définie sur $[0,T]$. On dit que le contrôle $u(.)$ (ou la trajectoire $x(.)$) est singulier sur $[0,T]$ si la différentielle de Fréchet $dE_T(u)$ de l'application entrée-sortie au point u n'est pas surjective. Sinon on dit qu'il est régulier.

Proposition 7.9. *Soient x_0 et T fixés. Si u est un contrôle régulier, alors E_T est ouverte dans un voisinage de u.*

Preuve.

▷ Par hypothèse, il existe n contrôles v_i tels que $dE_T(u).v_i = e_i$ où $(e_1,\dots,e_n)$ est la base canonique de $\mathbb{R}^n$. On considère l'application :

$$(\lambda_1,\dots,\lambda_n)\in\mathbb{R}^n \longmapsto E_T(u+\sum_{i=1}^{n}\lambda_i v_i).$$

Par construction, c'est un difféomorphisme local, et le résultat s'ensuit.

□

Autrement dit, en un point x_1 atteignable en temps T depuis x_0 par une trajectoire régulière $x(\cdot)$, l'ensemble accessible $\mathcal{A}(T,x_0)$ est *localement ouvert*, i.e. est un voisinage du point x_1. En particulier cela implique que le système est *localement contrôlable* autour du point x_1. On parle aussi de *contrôlabilité le long de la trajectoire $x(\cdot)$*. Ainsi :

Proposition 7.10. *Si u est un contrôle régulier sur $[0,T]$, alors le système est localement contrôlable le long de la trajectoire associée à ce contrôle.*

Le corollaire suivant est immédiat.

Corollaire 7.11. *Soit $u(.)$ un contrôle défini sur $[0,T]$ tel que sa trajectoire associée $x(.)$ issue de $x(0)=x_0$ est définie sur $[0,T]$ et vérifie au temps T :*

$$x(T)\in\partial\mathcal{A}(T,x_0).$$

Alors le contrôle u est singulier sur $[0,T]$.

7.4.2 Caractérisation hamiltonienne des contrôles singuliers

Définition 7.7. Le *Hamiltonien* du système (7.19) est la fonction :

$$\begin{array}{rcl}\mathbf{H}:\mathbb{R}^n\times(\mathbb{R}^n\setminus\{0\})\times\mathbb{R}^m & \longrightarrow & \mathbb{R}\\ (x,p,u) & \longmapsto & \mathbf{H}(x,p,u)=p\cdot f(x,u).\end{array}$$

Proposition 7.12. *Soit u un contrôle singulier sur $[0,T]$, et soit $x(\cdot)$ la trajectoire singulière associée. Alors il existe une application absolument continue $p:[0,T]\longrightarrow\mathbb{R}^n\setminus\{0\}$,*

appelée vecteur adjoint, telle que les équations suivantes sont vérifiées pour presque tout $t \in [0,T]$ *:*

$$\dot{x}(t) = \frac{\partial \mathbf{H}}{\partial p}(x(t), p(t), u(t)), \tag{7.21}$$

$$\dot{p}(t) = -\frac{\partial \mathbf{H}}{\partial x}(x(t), p(t), u(t)), \tag{7.22}$$

$$\frac{\partial \mathbf{H}}{\partial u}(x(t), p(t), u(t)) = 0, \tag{7.23}$$

où $\mathbf{H}$ *est le hamiltonien du système.*

L'équation (7.23) est appelée *équation de contrainte*. Il faut remarquer que $p(t) \neq 0$ pour tout $t \in [0,T]$.

Preuve.

▷ Par définition, le couple (x,u) est singulier sur $[0,T]$ si $dE_T(u)$ n'est pas surjective. Donc il existe un vecteur ligne $\psi \in \mathbb{R}^n \setminus \{0\}$ tel que pour tout contrôle v dans L^∞ on ait :

$$\psi.dE_T(u).v = \psi \int_0^T M(T)M^{-1}(s)B(s)v(s)ds = 0$$

Par conséquent :

$$\psi M(T)M^{-1}(s)B(s) = 0 \quad \text{p.p. sur [0,T].}$$

On pose $p(t) = \psi M(T)M^{-1}(t)$ pour tout $t \in [0,T]$. C'est un vecteur ligne de $\mathbb{R}^n \setminus \{0\}$, et $p(T) = \psi$. On a par dérivation :

$$\dot{p}(t) = -p(t)\frac{\partial f}{\partial x}(x(t), u(t)).$$

En introduisant le Hamiltonien $\mathbf{H}(x,p,u) = p \cdot f(x,u)$, on obtient :

$$f(x(t), u(t)) = \frac{\partial \mathbf{H}}{\partial p}(x(t), p(t), u(t)),$$

et

$$-p(t)\frac{\partial f}{\partial x}(x(t), u(t)) = -\frac{\partial \mathbf{H}}{\partial x}(x(t), p(t), u(t)).$$

La dernière relation vient de $p(t)B(t) = 0$ car $B(t) = \frac{\partial f}{\partial u}(x(t), u(t))$.

□

7.4.3 Démonstration du Théorème 7.7

Associons au système (7.14), (7.15) le *système augmenté* suivant :

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)), \\ \dot{x}^0(t) &= f^0(x(t), u(t)), \end{aligned} \tag{7.24}$$

et notons $\tilde{x} = (x, x^0)$, $\tilde{f} = (f, f^0)$. Le problème revient donc à chercher une trajectoire solution de (7.24) joignant les points $\tilde{x}_0 = (x_0, 0)$ et $\tilde{x}_1 = (x_1, x^0(T))$, et minimisant la dernière coordonnée $x^0(T)$.

L'ensemble des états accessibles à partir de $\tilde{x}_0$ pour le système (7.24) est $\tilde{\mathcal{A}}(\tilde{x}_0, T) = \bigcup_{u(\cdot)} \tilde{x}(T, \tilde{x}_0, u)$.

Le lemme crucial est alors le suivant.

Lemme 7.13. *Si le contrôle u associé au système de contrôle (7.14) est optimal pour le coût (7.15), alors il est singulier sur $[0, T]$ pour le système augmenté (7.24).*

Preuve.

▷ Notons $\tilde{x}$ la trajectoire associée, solution du système augmenté (7.24), issue de $\tilde{x}_0 = (x_0, 0)$. Le contrôle u étant optimal pour le coût (7.15), il en résulte que le point $\tilde{x}(T)$ appartient à la frontière de l'ensemble $\tilde{\mathcal{A}}(T, \tilde{x}_0)$, voir figure 7.1. En effet sinon, il existerait un voisinage du point $\tilde{x}(T) = (x_1, x^0(T))$ dans $\tilde{\mathcal{A}}(T, \tilde{x}_0)$ contenant un point $\tilde{y}(T)$ solution du système (7.24) et tel que l'on ait $y^0(T) < x^0(T)$, ce qui contredirait l'optimalité du contrôle u. Par conséquent, d'après la proposition 7.9, le contrôle u est un contrôle singulier pour le système augmenté (7.24) sur $[0, T]$.

□

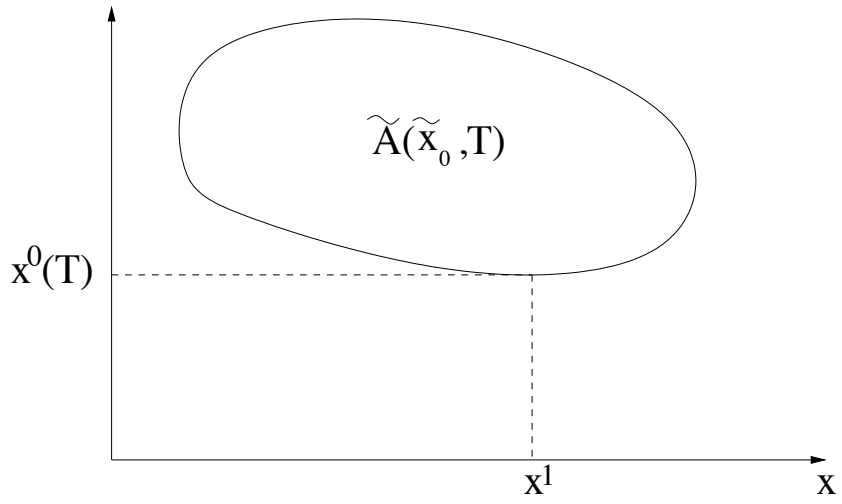


Figure 7.1 – Ensemble accessible augmenté.

Dans la situation du lemme, d'après la proposition 7.12, il existe une application $\tilde{p} : [0, T] \longrightarrow \mathbb{R}^{n+1} \setminus \{0\}$ telle que $(\tilde{x}, \tilde{p}, \tilde{u})$ soit solution du système hamiltonien :

$$\dot{\tilde{x}}(t) = \frac{\partial \tilde{H}}{\partial \tilde{p}}(\tilde{x}(t), \tilde{p}(t), u(t)),\ \dot{\tilde{p}}(t) = -\frac{\partial \tilde{H}}{\partial \tilde{x}}(\tilde{x}(t), \tilde{p}(t), u(t)), \tag{7.25}$$

$$\frac{\partial \tilde{H}}{\partial u}(\tilde{x}(t), \tilde{p}(t), u(t)) = 0, \tag{7.26}$$

où $\tilde{H}(\tilde{x}, \tilde{p}, u) = \tilde{p} \cdot \tilde{f}(t, \tilde{x}, u)$.

En écrivant $\tilde{p} = (p, p^0) \in (\mathbb{R}^n \times \mathbb{R}) \setminus \{0\}$, où p^0 est appelée variable duale du coût, on obtient :

$$(\dot{p}, \dot{p}^0) = -(p, p^0) \begin{pmatrix} \frac{\partial f}{\partial x} & 0 \\ \\ \frac{\partial f^0}{\partial x} & 0 \end{pmatrix}.$$

Or $\tilde{H} = \tilde{p} \cdot \tilde{f}(x, u) = pf + p^0 f^0$, donc :

$$\frac{\partial \tilde{H}}{\partial u} = 0 = p\frac{\partial f}{\partial u} + p^0 \frac{\partial f^0}{\partial u},$$

et donc $\dot{p}^0(t) = 0$, c'est-à-dire $p^0(t)$ est constant sur $[0, T]$. Comme le vecteur $\tilde{p}(t)$ est défini à scalaire multiplicatif près, on choisit $p^0 \leq 0$. Le théorème est démontré.

Bibliographie

[Agr-Sa] A. Agrachev, Y. Sachkov, Control Theory from the Geometric Viewpoint, vol. 87 of Encyclopaedia of Mathematical Sciences. Control Theory and Optimization, II. Springer, Berlin (2004)

[Rou-Bo] F. Bonnans, P. Rouchon, Commande et optimisation de systèmes dynamiques. Les Édition de L'École Polytechnique 2005.

[Bo-Pi] U. Boscain, B. Piccoli, Optimal Synthesis for Control Systems on 2-D Manifolds, Springer, SMAI, Vol.43, 2004.

[Jurd] Jurdjevic, V. : Geometric Control Theory, vol. 52 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1997)

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

[So] E. D. Sontag, Mathematical Control Theory : Deterministic Finite Dimensional Systems, no. 6 in Texts in Applied Mathematics, Springer-Verlag, New York/Heidelberg/Berlin, 2 ed., 1998.

MAP 561 ENGLISH DRAFT

Ugo Boscain et Yacine Chitour

Notes de cours

Édition 2014/2015

Table des matières

Foreward

In this course, the basic concepts of linear control theory are presented. We use a state space approach, which itself relies on ordinary differential equations (ODEs for short).

The lectures notes are divided in seven chapters. The first one 1 consists of a quick introduction to the main issues of control theory by going through a standard example, that of a robot arm. In particular, we will explain why ODEs are relevant in this context and why a good knowledge of their fundamental properties is necessary before addressing control theoretical problems.

The second chapter 2 is therefore devoted to the study of ODE and their use for modeling problems in physics, mechanics, economy, biology etc. Emphasis will be mainly put on two points. The first one refers to the notion of stability which in many practical situations should be compared to the effective knowledge of exact solutions of the ODEs. The second point consists of a detailed description of solutions for linear ODEs with constant coefficients. Note that this chapter is rather complete and several points may have been already seen and thus can be skipped over.

In the third chapter, are addressed basic issues of control theory such that controllability, observability and stabilization of linear control systems. We will show how these properties can be completely characterized with classical criteria such that Kalman criterion and the pole shifting theorem. We will also present a constructive solution to the motion planning problem based on the Brunovsky ouput, then the Luenberger (asymptotic) observer and the resulting separation principle.

Chapter 4 is devoted to the study of controllability of non linear control systems, more precisely for those which are affine in the control. We will put emphasis on the notion of Lie bracket which is instrumental in order to describe the structure of the reachable set. Basic results will be presented such as Krener theorem for accessibility as well as the theorems of Chow-Rashevski and Nagano-Sussmann for controllability and the structure of the orbit.

The last three chapters will be concerned with optimal control.

In these notes, results come sometimes with a proof. In the latter is not useful for other purposes in the course, it is written in small letters (small like that) and preceded by the symbol *. The same typography (small letters and symbol *) is used for the more advanced parts of the document which will not be treated in these notes. It is however not forbidden to read them. . .The symbol ” :=” means that the left-hand side is defined

by the right-hand one.

This course has been taught previously by Pierre Rouchon and Frédéric Bonnans between 1994 and 2004 under the title "Commande et Optimisation de systèmes dynamiques". It appeared as the book $[Rou - Bo]$ and contains a wealth of examples. Moreover, the authors of the present document want to thank Frédéric Jean and Emmanuel Trélat for allowing them to borrow to a large extent materials from the beautiful books "Equations différentielles et fondements de l'automatique" and "Contrôle optimal : théorie et applications".

Finally, these notes are by far not perfect and the authors welcome any comment and correction.

Chapitre 1

Introduction

Control Theory is the science which handles laws of regulation of controlled systems. Controlling an object means influencing its behavior to "oblige" it to make a task defined in advance. To realize in practice this "influence", engineers worked out appropriate mechanisms appealing to general theoretical principles, themselves formulated by means of mathematical tools. These mechanisms range from the Watt regulator (for the steam engines) to the most sophisticated microprocessors which we can find in players CD, automobiles, or still in the industrial robots or the automatic pilots of planes. The study of these mechanisms and their interaction with the corresponding object to control is the purpose of the course.

We are going to illustrate in more details our comment by means of a simple example stemming from robotics. We take back the description which is made in $[Rou - Bo]$. It is about a stiff arm turning in a vertical plan around a horizontal axis (Ox) and the latter is equipped with an engine delivering a variable couple $u \in \mathbb{R}$ which we can choose *arbitrarily* at every time : u is *command* of the system (or still *input*). The geometric position of the system is described by an angle $\theta \in S^1$, the unit circle i.e., that S^1 is the "sphere" of dimension 1. The dynamics of the system is obtained from the preservation of kinetic momentum around the axis (Ox) :

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

with m the mass of the arm, J its moment of inertia with respect to the axis (Ox), l the distance of the center of gravity to (Ox) and g the acceleration due to the gravity. We have recognized the equation of a pendulum weighing without friction. Although the constants which intervene in this problem play a dominant role in practice, we shall suppose for the rest of the chapter that $mlg = J = 1$. The dynamics of the arm is thus

$$\ddot{\theta}(t) + \sin\theta(t) = u(t). \tag{1.2}$$

For an angle θ_r fixed, a possible objective of command is to bring the arm to the angle θ_r and to maintain it there for later times. We shall say while the arm is in equilibrium

position in θ_r and the objective of command consists of *stabilizing* the system at θ_r. More generally, we can define as objective to *follow a trajectory* of reference $\theta_r(\cdot)$ which verifies the dynamics (1.2) for a reference control $u_r(\cdot)$ with the functions of time θ_r, u_r defined on an interval of time $[0, T_r]$ (T_r finite or not).

Remarque. Let us notice that the initial objective is a particular case of the follow-up of trajectory because to bring the arm to an angle θ_1 and to maintain it there, is equivalent to follow the trajectory of (1.2) associated to $\theta_r(\cdot) \equiv \theta_r$ and $u_r(\cdot) \equiv \sin\theta_r$.

We will now describe the successive steps that are typically followed by the control theorist in order to solve the problem.

1.1 Bringing the dynamics to state form and definition of the control system

An interval of time $[0, T]$ must be fixed for all the time functions which we shall envisage. Here, it is natural to set $T = T_r$, time to which we want to bring the arm in equilibrium position. To know the evolution of the movement when we apply a couple $u(\cdot) : [0, T] \to \mathbb{R}$, it is necessary to integrate the second order differential equation (1.2). For that purpose, it is necessary to know, at $t = 0$, the angular position $\theta(0) = \theta_0$ and the angular speed $\dot{\theta}(0) = \omega_0$. The pair (θ_0, ω_0) represents the initial condition of the following first order differential system, obtained from (1.2) :

$$(S) \left\{ \begin{array}{rcl} \dot{\theta}(t) & = & \omega, \\ \dot{\omega}(t) & = & -\sin\theta(t) + u. \end{array} \right. \tag{1.3}$$

At time t, the state of the system is thus only determined by the datum $(\theta(t), \omega(t))$ and $u(t)$. The variable $x = (\theta, \omega)$ forms *the state* of the system, which is thus a point of $S^1 \times \mathbb{R}$, and the couple of functions $t \mapsto (x(t), u(t))$ with $t \in [0, T]$ is called trajectory of the system.

In a more classical way, a trajectory is the *output* of the system with control $u(\cdot)$ (also called *input*).

The dynamics (S) can be rewritten in terms of x as follows :

$$(S) \quad \dot{x}(t) = F(x(t), u(t)), \tag{1.4}$$

with $F : S^1 \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^2$ the *the vector field* which associates to each point $(x, u) = (\theta, \omega) \in S^1 \times \mathbb{R} \times \mathbb{R}$ the vector $(\omega, -\sin\theta + u) \in \mathbb{R}^2$.

Finally, we call *control system* (associated with the arm of robot) (Σ), the following data : the space of state $S^1 \times \mathbb{R}$, the space of control values $\mathbb{R}$, the dynamics (1.4) and the class of the admissible control functions (or simply controls) Ad, i.e., all the functions $u : [0, T] \to \mathbb{R}$ (continuous, piecewise continuous, polynomials, etc.).

As preliminary study to this system (and in the objectives of command which we want to reach), it is imperative to understand what happens if "we do nothing", i.e., by putting $u = 0$ in (1.4). We are then driven to study the ordinary differential equation (ODE) defined on $S^1 \times \mathbb{R}$ by

$$(S)_0 \left\{ \begin{array}{rcl} \dot{\theta}(t) & = & \omega, \\ \dot{\omega}(t) & = & -\sin\theta(t), \end{array} \right. \tag{1.5}$$

or $\dot{x} = F(x(t), 0)$. In control theory, this stage is called *study of the free answer*. It is particularly important to know what is the behavior of the free trajectories when t tends infinity. We speak then of *study of the asymptotic stability* about the ODE (1.5). For example, do trajectories converge to a point of $S^1 \times \mathbb{R}^2$? Such points are called *equilibrium points* of the free system and correspond to constant trajectories. A simple calculation shows that the only equilibrium points points of the arm are $(0, 0)$ and $(\pi, 0)$. (A constant trajectory corresponds to set the right-hand side of the dynamics to be indentically equal to 0.)

1.2 Controllability

Given two states $x_0 = (\theta_0, \omega_0)$ and $x_1 = (\theta_1, \omega_1)$ in the state space, the controllability problem between these two states consists of finding a trajectory of (Σ) (i.e., $t \mapsto (\theta(t), \omega(t), u(t))$) starting from x_0 for $t = 0$ and ending at x_1 for $t = T$. In other words, one has to find a control $u(\cdot)$ which brings the system from one state to another one. If it is possible, we shall say that (Σ) is *controllable* between x_0 and x_1 and it is *completely controllable* if it is controllable for any pairs of states. It is necessary to notice that the question of controllability can split into two subproblems :

(Q1) given (Σ), can we show that it is completely controllable or not (most likely without explicitation of the controls) ?

(Q2) given a pair (x_0, x_1) of controllable states by (Σ), can one provide an effective procedure to determine a control which brings the system from x_0 to x_1 ?

The question $(Q1)$ is of theoretical nature and is far from being solved at the present moment. We shall study it in the particular case of linear systems. We can express the problem of controllability in linear algebraic terms and we can then give a necessary and sufficient condition onto the dynamics of the controlled system which characterizes controllability.

The question $(Q2)$, called also *trajectory planning* or *motion planning problem*, is even more difficult to solve than $(Q1)$. We shall propose a complete solution to that question in the case of linear systems.

1.3 Feedback

For the stiff arm, a way rather natural to determine the control $u(\cdot)$ which has to realize our objective is to proceed as follows : we know where we start from (the point x_0) and where we end up to (the point $x_r = \theta_r, 0)$) at time T. Let us suppose that, only from the knowledge of x_0, x_r and T, we are now capable of calculating a control $u(\cdot)$ bringing the arm from x_0 to x_r. We say that we command *in open loop*. This way of doing presents at least two defects :

— at time $t = 0$, we are supposed to calculate entirely the control law $u : t \in [0, T]$ then, to implement it into the physical system, practically at $t = 0$. It supposes that the calculation time of the control $u(\cdot)$ is negligible compared with that of the system. For certain applications, it is unrealistic ;

— let us suppose that an incident occurs on the interval $[0, T]$ which is not taken into account by the dynamics of the system. The law of command having already been calculated, the system cannot react to the unforeseen incident. (It is necessary to notice that this "unforeseen" event can arise constantly if the model which is used is only an approximation !).

That is why it is often advisable to calculate $u(\cdot)$ in a simpler way and to adjust it in real time in such a way it compensates the instantaneous errors with the reference trajectory, $\theta - \theta_r$ and $\omega - \omega_r$ which can appear. For example, we can choose, for $t \in [0, T]$, $u(t)$ in terms of $\theta(t) - \theta_r$ and $\omega(t) - \omega_r$. The use of this type of terms corresponds to a *feedback*. Let us notice that a control of this type supposes the knowledge for all times $t \in [0, T]$ of the quantities $\theta(t)$, θ_r and $\omega(t) - \omega_r$.

1.4 Stabilisation

Let us remind that our objective is to stabilize the arm in the point $x_r = \theta_r, 0)$. Let us suppose that we achieve x_r, we stay there with the constant control $u_r := \sin\theta_r$. Furthermore, it is clear that we can approach of θ_r in finite time stay in a neighborhood of this angle with an angular speed ω, not too big. Being able then of decreasing **at the same time** $\theta(t) - \theta_r$ and $\omega(t)$ towards zero is a little less obvious. One way to proceed is then to linearize (Σ) in the neighborhood of x_r i.e., to write

$$x = x_r + \delta x \text{ avec } \delta x = (\delta\theta, \delta\omega) \quad u = u_r + \delta u.$$

Here, δx and δu are small. To write the dynamics of δx from (1.3), we make a Taylor expansion of the second members of (1.3) and we retain only the terms of order 1 $\delta x, \delta u$. We obtain then *the linearized system* $(S)_L$ from (S) along the constant trajectory x_r.

$$(S)_L \left\{ \begin{array}{rcl} \dot{\delta\theta}(t) & = & \delta\omega, \\ \dot{\delta\omega}(t) & = & \delta u(t) - cos(\theta_r)\delta\theta(t). \end{array} \right. \tag{1.6}$$

We can also write this system as

$$(S)_L \quad \dot{\delta x} = A\delta x + b\delta u, \tag{1.7}$$

with

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ et } b := \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The objective of stabilization consists now in bringing any point of $\mathbb{R}^2$ at $(0,0)$ along $(S)_L$. Let us notice that the right-hand side of $(S)_L$ is linear in $(\delta\theta, \delta\omega, \delta U)$. We shall say while we have a linear control system constant in time i.e., not depending explicitly on time and with constant coefficients. As we shall see, these characteristics of linearity allow one to express the trajectories of this system in an explicit way. To stabilize the system while keeping a linear character, it is natural to choose δu linear in δx,

$$\delta u = -k^T \delta x = -k_1\delta\theta - k_2\delta\omega, \tag{1.8}$$

with $k = (k_1, k_2)^T$ constant vector called *gain of the controller*. The feedback law defined above is called *static feedback*. With this choice of $u(\cdot)$, the closed-looped system is written

$$(Lin)_k \quad \dot{\delta x} = (A + bk^T)\delta x.$$

Define

$$A(k) := A + bk^T.$$

It is thus a question of determining k so that the linear ODE $(Lin)_k$ is *asymptotically stable* : for any initial condition $\delta x_0 \in \mathbb{R}^2$, its solutions tend to when t tends to infinity. It is equivalent to the following problem of linear algebra : find a vector k such that the eigenvalues of $A(k)$ have strictly negative real part. These eigenvalues are called *poles* of the closed-loop system $(Lin)_k$. We shall see that it is enough to choose $k_1, k_2 > 0$ and the convergence of the solutions (towards zero) is then exponential. For example, if the eigenvalues λ_1, λ_2 of $A(k)$ are real and distinct, then any solution is a linear combination of $\exp(\lambda_1 t)$ and $\exp(\lambda_2 t)$.

Finally, we can want also to control the speed in which we stabilize the arm around x_r or still to ask there be no oscillation. It means in the first case controlling the speed of convergence $(Lin)_k$ and in the second case to have only real eigenvalues for $A(k)$. We have two particular situations of the more general problem which consists in determining the vector k so that $A(k)$ has eigenvalues verifying certain conditions. We thus arrive at the following purely algebraic question : " given a pair (A, b) with A a 2×2 matrix and b a vector column, characterize all the eigenvalues of $A(k) = A + bk^T$ when k is any vector. The *pole shifting theorem* says essentially the following : if the pair (A, b) is controllable (i.e., the vectors b and Ab are linearly independent), then any possible pair of eigenvalues in $\mathbb{C}^2$ can be realized by $A(k) = A + bk^T$ for some vector k.

1.5 Observability

The law of feedback given in (1.8) supposes that we measure **at every time** $t \in [0,T]$ the complete state of the system $x = (\theta, \omega)$. In practice, the sensors of speed are very expensive. It is thus often reasonable to suppose that we measure only the position and here it means that we know instantaneously only θ. The measured quantities constitute *the output* of a controlled system. They represent an information about the state, instantaneous but often partial. It is clear that a law of feedback has practical reality only if the latter is obtained only by means of the output.

Let us return to the control system with as only output the angular position θ. If we try a static feedback with only a function of θ, we can show that it is impossible to stabilize the arm. For example, if we take a feedback linear in θ, we end up with an equation of the type

$$\ddot{\delta\theta} + k\delta\theta = 0,$$

which is not asymptotically stable whatever the value of k.

However, we notice that we can obtain $\omega(t)$ by differentiating $\theta(t)$. We say that the state x of the system is *observable* from the output θ. More generally, we shall see that the state is observable from an output y if we can reconstruct x from a finite number of differentiations of y.

For the arm, we can differentiate numerically the measured signal θ to deduce it from ω and so build a stabilizing feedback law. This solution works if there is not too much "noise" in the measure of θ. Should the opposite occur, the digital operation of the differentiation must be avoided. After linearization, the idea is then of *estimate* the state δx from the only knowledge of the angle $\delta\theta$ **without deriving** $\delta\theta$. For that purpose, it is necessary to use another information on δx : the latter verifies the dynamics $(S)_L$! We try then to build an artificial state $\tilde{\delta x}$ such as $\delta x - \tilde{\delta x}$ tends to zero when t tends to infinity. Such a $\tilde{\delta x}$ is called *asymptotic observer*. Let us note $\delta\theta = C\delta x$ with C the line vector equal to (1 0). We choose $\tilde{\delta x}$ as the trajectory of

$$(S)_L \quad \dot{\tilde{\delta x}} = A\tilde{\delta x} + b\delta u - LC(\delta x - \tilde{\delta x}), \tag{1.9}$$

with L a vector column to be determined. Here, we indeed have $C(\delta x - \tilde{\delta x}) = (\delta\theta - \tilde{\delta\theta})$. Let us notice that the dynamics (1.9) is obtained by adding, in the linearized dynamics of the arm, the term $LC(\delta x - \tilde{\delta x})$ which brings in, set apart "artificial" terms, only the output $\delta\theta$. When we consider the dynamics of *the error* $e := \delta x - \tilde{\delta x}$, we have

$$\dot{e} = (A + LC)e.$$

Set $A(L) := A + LC$. Make aim e tend to zero when t tends to infinity becomes once again a problem of linear algebra which we solve simply. By applying then the law of feedback (1.8) obtained for $\tilde{\delta x}$, we shall obtain a law of feedback which stabilizes (locally) the arm. This type of law is called *dynamic feedback* because we stabilize the

arm by means of $\tilde{\delta x}$ which is obtained, from the output $\delta\theta$, in a dynamic way (i.e., via a differential equation). We shall also notice that the action which consists in estimating δx is decoupled by the one which consists in stabilizing it : the choices of L(estimation) and of k (stabilization) are independent one of the other one. It is the principle of separation.

1.6 Optimal Control

Once the question of controllability is understood, we saw that the determination of a law of control is made accordingly to the objective of command that we assign in advance. This one can be for instance stabilization as we saw above. We can also want to minimize the control which is necessary for the realization of the fixed objective. For example, for the arm of robot described by (1.3), we can want to minimize the work of the strength which is necessary to bring the system (1.3) of a state x_0 to another x_1 at time $T = 1$. It means that it is necessary to minimize $\int_0^1 |u|$ among all the control laws which bring the system (1.3) from x_0 to x_1. Also, another type of minimization is the one of time when the amplitude of the control is uniformly bounded : let us suppose that the control u takes its values in $(-1, 1)$. For any pairs of states x_0 and x_1, the problem consists in minimizing the necessary time to bring the system (1.3) from x_0 to x_1.

So, when one tries "to optimize" the control, it is necessary to define a criterion of optimization which will be called the *cost function* (or simply the *cost*). The purpose is then to show that there is (or not) a minimizing control, also called *optimal control*, and especially to characterize this (or these) optimal control(s). There are several manners to attack the questions of optimal control, depending on the type of dynamics we consider. Let us note that these methods can be included) under a general principle, called *Pontryagin maximum principle*.

1.7 Structure of the course

The chapter 2 is dedicated to the autonomous ordinary differential equations (ODE) in finite dimension. We shall define the notions of vector field, trajectories and we shall present some fundamental results which are connected with it : Cauchy problem (existence and uniqueness of solutions with the Cauchy-Lipschitz theorem), stability in the sense of Lyapunov for equilibrium points, etc. The case of the linear ODE with constant coefficients will be carefully analyzed.

In the chapter 3, is begun the study of the autonomous linear systems. At first, we shall emphasize on the Brunovsky form and its application to the motion planning problem and to the stabilization by pole shifting. We shall address then the observability as a dual problem to that of controllability, the construction of asymptotic observers (or of Luenberger) and finally the synthesis of a dynamic output feedback (we also say observer-controller). The chapter 4 constitutes a brief introduction in the controllability of non linear systems.

Chapters 5 and 6 are respectively the objects of the detailed study of the command LQ with its application in the Kalman filter and the minimization of time for linear systems. Finally the chapter 7 presents the Pontryagin maximum principle for the systems affine in the control as well as an introduction to the optimal synthesis in dimension two.

It will be necessary to consult $[Rou - Bo]$ for the numerous representative examples of questions worrying the engineers as well as for the exercises which allow to assimilate the course.

The parts written in small print can be ignored except special mention : it is often about demonstrations.

Chapitre 2

Ordinary differential equations and stability

2.1 General theory of differential equations

In this section, we present a general theory of *autonomous* differential equations of the form

$$x'(t) = f\big(x(t)\big). \tag{2.1}$$

This theory allows one to model and study numerous finite dimensional evolution processes which are deterministe and *differentiable*.

In the formulation (2.1), the data are :

— an open set $\Omega \subset \mathbb{R}^n$; x and Ω are respectively called *state* of the system and *state space* of the system : at each time the system is characterized by the value of x which lives in Ω.

— a continuous application $f : \Omega \to \mathbb{R}^n$, (sometimes called "right-hand side" of the equation).

(The results that we present remain valid when we replace $\mathbb{R}^n$ by any vector space of finite dimension, for example $\mathbb{C}^n$, $M_n(\mathbb{R}), \ldots.$) Such an application f : $\Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is called *vector field* : at any point $xin\Omega$, it associates a vector $f(x)$ in $\mathbb{R}^n$. In mechanics, f is also called *field of speed*.

Exemple. By returning to the stiff arm (cf. previous chapter), the free answer corresponds to the ODE $x'(t) = f(x(t))$ with $\Omega = S^1 \times \mathbb{R}$ and the vector field $f : S^1 \times \mathbb{R} \to \mathbb{R}^2$ defined by $f(x) = (w, -\sin\theta))^T$ if $x = (\theta, w)$.

A *solution* of the ODE is a differentiable function $x(\cdot) : I \to \mathbb{R}^n$ such that :

— I is an interval of $\mathbb{R}$;

— $x(\cdot)$ takes values in Ω, *i.e.* $x(I) \subset \Omega$;

— for every $t \in I$, $x'(t) = f\big(x(t)\big)$.

A solution is then a pair $(x(\cdot), I)$: the interval of definition I is part of the unknowns quantities. We shall see how to characterize that interval in Section 2.1.2.

Note that since the application f is supposed to be continuous, every solution $x(\cdot)$ of the ODE is necessarily of class C^1.

Remarque. It can seem restrictive to only consider autonomous differential equations, while the most general frame is the one of equations of the form

$$x'(t) = f\big(t, x(t)\big), \qquad t \in J \subset \mathbb{R}, \tag{2.2}$$

which depend explicitly on time, and which are said *non-autonomous*. It is in fact not really a limitation : any non-autonomous equation in $\mathbb{R}^n$ can be seen as an autonomous equation in $\mathbb{R}^{n+1}$. Indeed, let us define the vector field $F : J \times \Omega \subset \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ by $F(t,x) = (1, f(t,x))$. It is then clear that the non-autonomous equation (2.2) is equivalent to the autonomous equation

$$\begin{pmatrix} t \\ x \end{pmatrix}' = \begin{pmatrix} 1 \\ f\big(t,x\big) \end{pmatrix} = F\big(t, x(t)\big).$$

2.1.1 Existence and uniqueness

Any differential equation does not have necessarily a solution. For instance, Consider the ODE defined $\mathbb{R}$ by

$$\left\{ \begin{array}{rcl} x'(t) & = & -sign\big(x(t)\big), \\ x(t_0) & = & 0, \end{array} \right.$$

with $sign(x) = x/|x|$ if x is non zero and $sign(0) = 0$. (Try to give a (simple) argument showing that there is no solution for the previous ODE on any neighborhood of 0.) In the light of this example, it is necessary to make a hypothesis of regularity on the right-hand side of an ODE to hope to have a "piece" of solution in an open neighborhood of the initial time.

Définition 2.1. We call *Cauchy problem*, the system

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t)\big), \\ x(t_0) & = & x_0, \end{array} \right. \tag{2.3}$$

i.e., the system formed by an ODE and by an initial condition (value of the state given at a fixed moment). Let us remind that one of the objectives of the ODE is to model physical processes which are often determinist : if we know the dynamics of a system and an initial condition in $t = t_0$, then the evolution of this system is uniquely determined for $t \geq t_0$. This notion of determinism is translated in mathematical terms by the fact that any Cauchy problem has one and only a solution for $t \geq 0$. To have uniqueness of the solutions of an ODE is thus a necessity for a realistic model.

The purpose of the theorem below is to answer the previous questions.

Théorème 2.1 (Cauchy-Lipschitz theorem)**.** *Assume that f is of class C^1 on Ω. Then, for every point $x_0 \in \Omega$ and every $t_0 \in \mathbb{R}$, there exists $\delta > 0$ such that the Cauchy problem defined in (2.3) has a unique solution defined on $]t_0 - \delta, t_0 + \delta[$.*

*Preuve.

▷ The demonstration of this theorem is based on the fixed point theorem of Picard. Let us fix a real $\alpha > 0$ such as the closed ball $\overline{B}(x_0, \alpha)$ is contained in Ω. Because f is C^1, there are constants M and K >0 such as, on $\overline{B}(x_0, \alpha)$, f is bounded in norm by $M and is$K-Lipschitzian (why ?). In addition, set

$$\delta = \min(\frac{\alpha}{M}, \frac{1}{2K}).$$

▷ Define $\mathcal{E}$ as the set of functions $x(\cdot)$ continuous on $]t_0 - \delta, t_0 + \delta[$ with values in $\overline{B}(x_0, \alpha)$ and such that $x(t_0) = x_0$. Endowed with the norm of uniform convergence, $\| \cdot \|_0$ is a complete space. The mapping

$$\Phi\big(x(\cdot)\big) = x_0 + \int_{t_0}^{\cdot} f\big(x(s)\big) ds,$$

is an application from $\mathcal{E}$ into itself : for $|t - t_0| \leq \delta$,

$$\|\Phi\big(x(t)\big) - x_0\| = \| \int_{t_0}^{t} f\big(x(s)\big) ds\| \leq \delta M \leq \alpha.$$

This application is also $\frac{1}{2}$-Lipschitzian since for $t \in]t_0 - \delta, t_0 + \delta[$,

$$\begin{aligned} \|\Phi\big(x(\cdot)\big) - \Phi\big(y(\cdot)\big)\|_0 &\leq \sup_{|t-t_0|<\delta} \Big(\int_{t_0}^{t} \|f\big(x(s)\big) - f\big(y(s)\big)\| \, ds \Big) \\ &\leq \sup_{|t-t_0|<\delta} \Big(\int_{t_0}^{t} K\|x(s) - y(s)\| \, ds \Big) \\ &\leq \delta K \|x(\cdot) - y(\cdot)\|_0 \leq \frac{1}{2} \|x(\cdot) - y(\cdot)\|_0. \end{aligned}$$

The fixed point theorem of Picard can be applied and it shows that the application Φ admits a unique fixed point in $\mathcal{E}$, i.e., the system (2.3) admits a unique solution $x(\cdot) :]t_0 - \delta, t_0 + \delta[\to \mathbb{R}^n$ with values in $\overline{B}(x_0, \alpha)$.

▷ It remains to show that every solution $x(\cdot) :]t_0 - \delta, t_0 + \delta[\to \mathbb{R}^n$ of (2.3) takes values in $\overline{B}(x_0, \alpha)$. By contradiction, assume that a solution $x(\cdot)$ of (2.3) goes out of $\overline{B}(x_0, \alpha)$ in time less than δ, and let t_1 be the first time where $x(\cdot)$ goes out of $B(x_0, \alpha)$. By the mean value theorem,

$$\alpha = \|x(t_1) - x_0\| \leq \big(\sup_{t \in [t_0, t_1]} \|x'(t)\| \big) \, |t_1 - t_0| < M\delta,$$

contradicting $\delta \leq \alpha / M$. Every solution of (2.3) on $]t_0 - \delta, t_0 + \delta[$ takes then values in $\overline{B}(x_0, \alpha)$, qed.

□

Weaker hypotheses on f

We expressed the theorem of Cauchy-Lipschitz with the hypothesis that f is C^1 on Ω because it is simple to use and frequently satisfied in the applications. Let us notice however that, in the proof, we only need that f to be*locally Lipschitzian*, i.e., that for any $x_0 \in \Omega$, there is a neighborhood U_0 of x_0 in Ω and a constant K such that f is K-Lipschitzian on U_0. The conclusion of the Cauchy-Lipschitz theorem remains thus valid un ofr the hypothesis that f is locally Lipschitzian. In particular, it is valid if f is globally Lipschitzian on Ω.

What happens if we still weaken the hypotheses and we assume f to be only continuous ? A theorem of Peano says that System (2.3) has always solutions but uniqueness is not guaranteed. For instance, the Cauchy problem

$$\left\{ \begin{array}{rcl} y'(t) & = & \sqrt{|y(t)|} \\ y(0) & = & 0 \end{array} \right. , \qquad y \in \mathbb{R}$$

admits as solutions the functions $y_1(t) = 0$ and $y_2(t) = \frac{t|t|}{4}$. It actually admits an infinity since, for every $a \geq 0$, the function $y^a(\cdot)$ defined by

$$y^a(t) = 0 \quad \text{for } t \leq a, \qquad y^a(t) = y_2(t-a) \quad \text{for } t > a$$

is also solution.

Remarque. According to the remark made in introduction, the Cauchy-Lipschitz theorem is also valid for a non-autonomous equation $x' = f(t,x)$: if f is C^1 on $J \times \Omega$, then, for $(t_0, x_0) \in J \times \Omega$, the equation has a unique solution defined on $]+t_0 - \delta, t_0 + \delta[$ and equal to x_0 at t_0.

We can weaken in that case sharply the hypotheses on f : Indeed, the conclusion of the theorem will remain valid if we suppose only that f is *locally Lipschitzian in the second variable* x, i.e., that, for every $(t_0, x_0) \in J \times \Omega$, there is a neighborhood J_0 of t_0 in J, a neighborhood U_0 of x_0 in Ω and a constant K such that, for any $t \in J_0$, the application $f(t,\cdot)$ is K-Lipschitzian on U_0. The proof is a simple adaptation of the one given above.

2.1.2 Maximal solutions and time duration

Consider the ODE

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

where the vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is of class C^1.

We defined at the beginning of this chapter a solution of this equation as a function $x(\cdot)$ defined on a certain interval I of $\mathbb{R}$. This section is dedicated in the study of this interval of definition I. We shall meet two types of problems.

— Given a couple $(t_0, x_0) \in \mathbb{R} \times \Omega$, there is an infinity of solutions of (2.4) satisfying the initial condition $x(t_0) = x_0$: for example if $x(\cdot)$ is solution on the interval I, $t_0 \in I$, any restriction of $x(\cdot)$ in a subinterval of I containing t_0 is a different solution. To avoid considering as different solutions the same functions taken on subintervals, we shall try to associate with a function a unique interval, the biggest, on which it is solution : it is the notion of *maximal solution*.

— If we choose the interval I the biggest possible, can we take it equal in $\mathbb{R}$ entirely ? If it is not the case, what happens for the solution ? And which is the shape of I ? It is the problem of *time duration* for solutions.

Maximal solutions

Définition 2.2. We say that a solution $x(\cdot) : I \to \Omega$ of (2.4) is a *maximal solution* if it has no continuation in a strictly bigger interval, i.e., if it is not the restriction to I of a solution defined on an interval $I' \supsetneq I$.

We are going to show that there is a unique maximal solution of the equation (2.4) satisfying a given initial condition. We need for it of a result of global uniqueness.

Proposition 2.2. *If $x(\cdot)$ and $y(\cdot) : I \to \Omega$ are two solutions of (2.4) defined on the same interval I which coincide at one point $t_0 \in I$, then they are equal.*

Preuve.

▷ Consider first the last time larger than t_0 so that the solutions coincide :

$$t_+ = \sup\{t \in I \ : \ t > t_0, \ x(s) = y(s) \text{ for every } s \in [t_0, t]\}.$$

By contradiction, let us suppose $t_+ < \sup I$. The solutions being continuous, we have $x(t_+) = y(t_+)$. By applying the Cauchy-Lipschitz theorem to the couple $(t_+, x(t_+))$, we obtain that both solutions are still equal on an interval $[t_+, t_+ + \delta[$, which contradicts the definition of t_+. Thus $t_+ = \sup I$. The same argument works for the infimum of times where both solutions coincide.

□

Théorème 2.3. *For every initial data $(t_0, x_0) \in \mathbb{R} \times \Omega$, there exists a unique maximal solution $x(\cdot) :]t_-, t_+[\to \Omega$ of (2.4) verifying $x(t_0) = x_0$. Every other solution verifying the same initial condition is a restriction of $x(\cdot)$ to a subinterval of $]t_-, t_+[$.*

Remarque. Let us insist on the fact that the interval of definition of a maximal solution is always an open interval $]t_-, t_+[$. Borders t_+ and t_- of the maximal interval are functions of (t_0, x_0) which take their values in $\overline{\mathbb{R}}$: t_+ can be either real or $+\infty$, while t_- can be either real or $-\infty$. In every case, $t_- < t_0 < t_+$.

Preuve.

▷ Let I be the union of all the intervals containing t_0 where the system

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t)\big), \\ x(t_0) & = & x_0, \end{array} \right. \tag{2.3}$$

admits a solution. According to the Cauchy-Lipschitz theorem, this union is an open interval, $I =]t_-, t_+[$. For any $t \in]t_-, t_+[$, let us define $x(t)$ as the value t of any solution of (2.3) defined on $[t_0, t]$. The previous proposition shows that the function $x(\cdot) :]t_-, t_+[\to \Omega$ so defined is a solution of (2.3). Furthermore, by construction, it is a continuation of any other solution.

□

Time duration

We are now interested in the interval of definition $]t_-, t_+[$ of a maximal solution $x(\cdot)$ of (2.4). This interval can be different of $\mathbb{R}$, even for "simple" equations.

Exemple. Consider the ODE $y'(t) = y^2(t)$ in $\mathbb{R}$, whose solution taking the value y_0 at t_0 is

$$y(t) = \frac{y_0}{(t - t_0)y_0 + 1}.$$

The maximal interval of definition of that solution is $]t_0 - \frac{1}{y_0}, +\infty[$ if $y_0 > 0$, $]-\infty, t_0 - \frac{1}{y_0}[$ if $y_0 < 0$, and $\mathbb{R}$ if $y_0 = 0$.

The general idea is that, if a solution cannot be extended on $\mathbb{R}$, it is because it approaches in finite time the boundary of the set Ω. Let us formalize this idea for t_+ of the interval results for t_- are similar).

Proposition 2.4. *Let $x(\cdot) :]t_-, t_+[\to \Omega$ a maximal solution of (2.4). Then, if $t_+ < +\infty$, $x(t)$ exits once for all every compact set contained in Ω as $t \to t_+$.*

*Preuve.

▷ To be done.

□

We will meet the case $t_+ < +\infty$ essentially in two situations :
— when $\Omega = \mathbb{R}^n$ and $\lim_{t \to t_+} \|x(t)\| = +\infty$: it corresponds to the case of *explosion in finite time* (cf. above example) ;
— when the boundary of Ω is bounded and $x(t)$ converges to a boundary point as $t \to t_+$.

Conversely, notice that we have a sufficient condition for $]t_-, t_+[= \mathbb{R}$.

Corollaire 2.5. *If the values of a maximal solution $x(\cdot)$ are contained in a compact set of Ω, then $x(\cdot)$ is defined $\mathbb{R}$.*

Complete vector fields

Définition 2.3. We say that the vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is *complete* (or the associated equation is *complete*) if any maximal solution is defined on $\mathbb{R}$ entirely.

According to the previous corollary, if all the maximal solutions are contained in a compact, the vector field is complete.

Exemple. If $\Omega = \mathbb{R}^n$, the linear vector fields $f(x) = Ax$ are complete and, more generally, the vector fields admitting a linear upper bound $\|F(x)\| \leq \alpha\|x\| + \beta$, with $\alpha, \beta \geq 0$, are complete (thus in particular the bounded vector fields). It is a consequence of Gronwall's lemma.

2.1.3 Flows, phase portraits

Consider the autonomous ODE

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

where the vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is of class C^1. One of the specificities of this equation is that it does not depend explicitly on time (thus the adjective "autonomous"). In particular, the solutions are invariant by translation of time : if $x(\cdot)$ is solution, then $x(t_0 + \cdot)$ also.

Proposition 2.6. *Let $x(\cdot) :]t_-, t_+[\to \Omega$ be a maximal solution of (2.4) and $t_0 \in \mathbb{R}$. Then $\bar{x} : t \mapsto x(t + t_0)$, defined on $]t_- - t_0, t_+ - t_0[$, is also a maximal solution of (2.4).*

So the time does not play an intrinsic role here and we can limit ourselves to the initial data $t = 0$. For a point $x \in \Omega$, let us note $\phi(\cdot, x)$ the maximal solution of (2.4) equal to x at $t = 0$ and $I_x =]t_-, t_+[$ its interval of definition. In other words, $\phi(\cdot, x)$ is the solution of the system

$$\left\{ \begin{array}{rcl} \frac{\partial}{\partial t}\phi(t,x) & = & f\big(\phi(t,x)\big), \\ \phi(0,x) & = & x, \end{array} \right. \qquad \forall t \in I_x.$$

Définition 2.4. The application $(t, x) \mapsto \phi(t, x)$ is called the *flow* of the vector field f (orde the equation $x' = f(x)$).

By definition, the partial application with x fixed, $t \mapsto \phi(t, x)$, is a maximal solution of the equation. For a qualitative study of the differential equation, it is important to study the other partial application, $\phi_t : x \mapsto \phi(t, x)$, for t fixed. Clearly, $\phi_t(x)$ is the position at time t of a body transported by the differential equation which was in the position x at $t = 0$.

Exemple. If f is linear, *i.e.* $f(x) = Ax$, $A \in M_n(\mathbb{R})$, the flow is given by the exponential of A :

$$\phi_t(x) = e^{tA}x, \quad \forall (t,x) \in \mathbb{R} \times \mathbb{R}^n.$$

Then the flow is a generalization of the exponential of a matrix and it shares similar properties.

Proposition 2.7 (formule du flow)**.** *If $t_1 \in I_x$ and $t_2 \in I_{\phi_{t_1}(x)}$, then $t_1 + t_2 \in I_x$ and*

$$\phi_{t_1+t_2}(x) = \phi_{t_2}\big(\phi_{t_1}(x)\big).$$

In particular, if $t \in I_x$,

$$\phi_{-t}\big(\phi_t(x)\big) = x$$

Preuve.

▷ According to the proposition on the invariance by time translation, $t \mapsto \phi_{t_1+t}(x)$ is equal to the maximal solution $\phi_{t_1}(x)$ at $t = 0$, which is the definition of $t \mapsto \phi_t\big(\phi_{t_1}(x)\big)$.

□

Remarque. The flow formula can also be read as follows : if $x(\cdot)$ is a solution of (2.4), then

$$x(t) = \phi_{t-t_0}\big(x(t_0)\big)$$

for every t_0 and t in the interval of definition of $x(\cdot)$.

The domain of definition of the flow is the set

$$\mathcal{D} = \{(t,x) \in \mathbb{R} \times \Omega \ : \ t \in I_x\}.$$

To get interesting properties on the flow (continuity, differentiability), it is necessary to first establish that its domain of definition $\mathcal{D}$ is open in $\mathbb{R} \times \Omega$. We will see that in the next section.

There is however already a case where it is obvious : if f is a complete vector field on Ω, i.e., if $I_x\mathbb{R}$ for any $x \in \Omega$, the domain of definition of the flow is $\mathcal{D} = \mathbb{R} \times \Omega$. We can then rewrite the properties of the flow in a global way : for every $t, s \in \mathbb{R}$,

1. $\phi_t \circ \phi_s = \phi_{t+s}$;
2. $\phi_{-t} \circ \phi_t = \mathrm{id}$;
3. $\phi_0 = \mathrm{id}$;
4. $\frac{\partial}{\partial t}\phi_t = f \circ \phi_t$.

The first three properties show in particular that ϕ_t obeys to a group operation.

Orbits and phase portraits

Définition 2.5. We call *orbit* of a point $x_0 \in \Omega$ (or trajectory passing through x_0) the set

$$\mathcal{O}_{x_0} = \{\phi_t(x_0) \ : \ t \in I_{x_0}\}.$$

In other words, the orbit of x_0 is the curve drawn in $\mathbb{R}^n$ by the maximal solution of (2.4) passing through x_0 at $t = 0$.

The property of invariance by time translation implies that, for any point $x \in \mathcal{O}_{x_0}$, we have $\mathcal{O}_x = \mathcal{O}_{x_0}$. Indeed, in this case, there exists a time t_0 such that $x = \phi_{t_0}(x_0)$. Any point y of $\mathcal{O}_x$ can be written $y = \phi_t(x) = \phi_{t+t_0}(x_0)$, i.e., $y \in \mathcal{O}_{x_0}$. In particular, this implies that *two different orbits cannot cross themselves*. Every point of Ω thus belongs to a single orbit.

The partition of Ω in orbits is called *phase portrait* of the vector field. We find three sorts of orbits :

— points, *i.e.* $\mathcal{O}_{x_0} = \{x_0\}$: such a point verifies inevitably $f(x_0) = 0$. It is what we call an *equilibrium point* (see definition 2.12 below). Notice that an equilibrium point corresponds to a fixed point of ϕ_t for any t : $\phi_t(x_0) = x_0$.

— closed curves : there is then a point x in the orbit and a time $T > 0$ such that $\phi_T(x) = x$. This implies $\phi_{t+T}(x) = \phi_t(x)$ for any $t \in \mathbb{R}$, *i.e.* that the maximal solution $\phi(\cdot, x)$ is T-periodic. We shall speak in this case of a *periodic orbit*.

— open curves : there is not then any double point, *i.e.* if $t \neq s$, $\phi_t(x) \neq \phi_s(x)$.

Usually, one draws on the phase portrait the direction of the orbits.

Exemple. Let us consider the linear vector field $f(x) = Ax$ in $\mathbb{R}^2$, and let us suppose that the matrix $A \in M_2(\mathbb{R})$ has two real and different eigenvalues $\lambda_1 < \lambda_2$. The study carried out in Section 2.2.4 allows one to determine the shape of the phase portrait according to λ_1 and λ_2. We represented the various possibilities in the figure 2.1, where we noted E_1 and E_2 the eigenspaces associated in λ_1 and λ_2.

Equilibrium points and periodic orbits are examples of invariant subsets whose definition is given below.

Définition 2.6. Let A be a subset of the state space Ω. We say that A is *invariant* (respectively *positively invariant*) by the flow ϕ_t if, for any $t \in \mathbb{R}$ (respectively in $\mathbb{R}_+$), $\phi_t(A)$ is included in A.

Other examples of invariant sets are supplied by the level hypersurfaces of a real function of the state space which remains constant along trajectories, i.e. first integrals.

Définition 2.7. We call *first integral* of an ODE $\dot{x} = fx)$, a function $h : \Omega \to \mathbb{R}$, of class C^1, which remains constant along the trajectories of the ODE. It is true in particular if, for any $x \in \Omega$ and t, $\frac{d}{dt}\big(h(\phi_t(x))\big) = 0$, which is equivalent to

$$D_x h(x).f(x) = 0, \quad \text{for every } x \in \Omega.$$

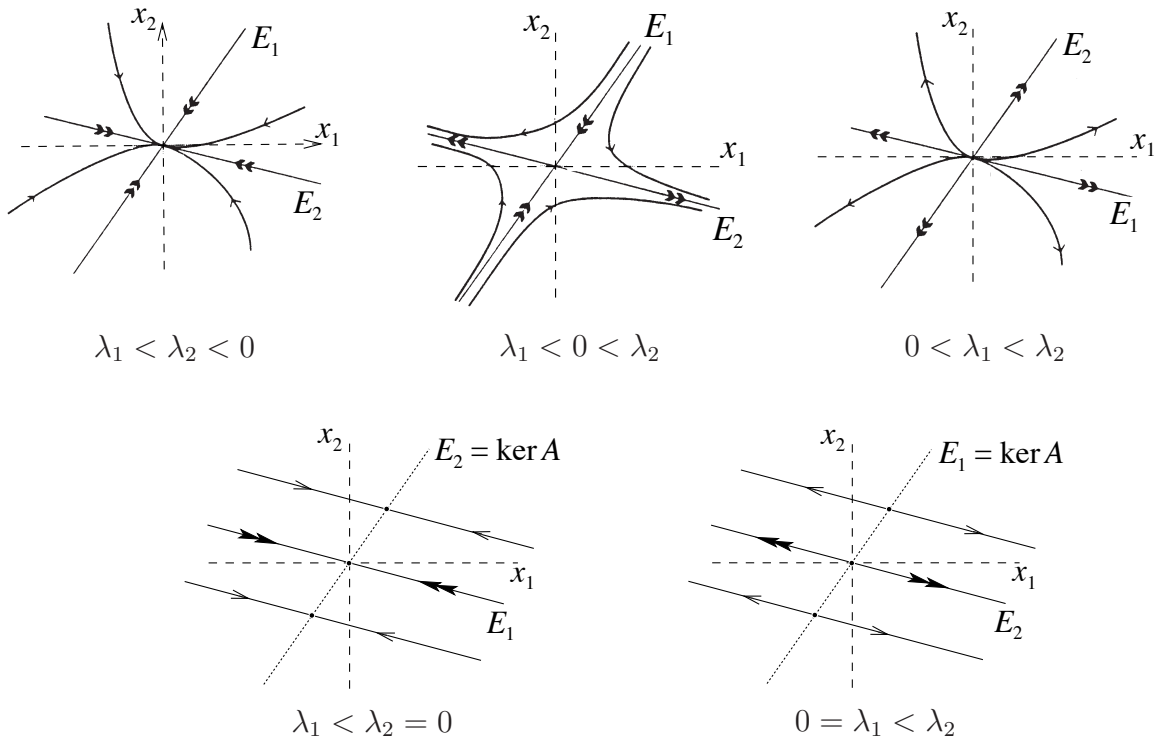


FIGURE 2.1 – Examples of phase portraits for $f(x) = Ax$ in $\mathbb{R}^2$.

(Note that the last condition does not require explicit knowledge of the flow.) The level sets of h given by $H_c := \{x \in \Omega : \quad h(x) = c\}$, $c \in \mathbb{R}$ are invariant by the flow.

Exemple. Show that (1.5) admits as first integrals $1/2\omega^2 - \cos\theta$. Deduce the equation of the orbits. Draw the the phase portrait on the cylinder $S^1 \times \mathbb{R}$ according to the value (constant) of h. To have a plane representation, we shall identify S^1 with $\mathbb{R}/[0, 2\pi]$ and we shall show the existence of periodic trajectories.

For other examples, to see [Rou-Bo] where it is explained the fact that the total energy (kinetic energy plus potential energy) of a perfect holonomic mechanical system (without friction) is a first integral of the system.

2.1.4 Linear differential equations

We suppose that the right-hand side of (2.1) is linear with regard to the state x *i.e.* it takes the shape

$$x'(t) = A(t)x(t), \qquad t \in I. \tag{2.5}$$

The *datas* are :
— an interval I of $\mathbb{R}$;

— an application $A : I \to M_n(\mathbb{K})$ of class C^k (k is a positive integer or $k = \infty$) ; each $A(t)$ is an $(n \times n)$ matrix with real or complex coefficients.

A *solution* of (2.5) is a differentiable application $x : I \to \mathbb{K}^n$ such that, for any $t \in I$, its time derivative $x'(t) = \frac{dx}{dt}(t)$ verifies $x'(t) = A(t)x(t)$. Note that a solution is automatically of class C^{k+1}.

We shall also handle the slightly more general case of *affine* differential equations

$$x'(t) = A(t)x(t) + b(t), \qquad t \in I, \tag{2.6}$$

with $b(\cdot)$ an application of I in $\mathbb{K}^n$ of class C^k. We shall see that the study of these equations can be deduced from that of the linear equations.

Remarque. It is frequent in the literature that the expression "linear equation" is used for the affine equations, equations (2.5) are then called *homogeneous* linear equations .

Global existence and uniqueness

Théorème 2.8 (Global existence and uniqueness)**.** *Let $t_0 \in I$ and $x_0 \in \mathbb{K}^n$. There exists a unique solution $x(\cdot)$ of equation (2.6) verifying*

$$x(t_0) = x_0.$$

This theorem insures the existence of $x(\cdot)$ on the whole interval. This is typical of linear equations (compare with Cauchy-Lipschitz theorem 2.1 regarding non-linear equations).

The proof of this theorem lies on the following remark : $x(\cdot)$ is solution of (2.6) with $x(t_0) = x_0$ if and only if $x(\cdot)$ is continuous and satisfies for every $t \in I$,

$$x(t) = x_0 + \int_{t_0}^{t} \big(A(s)x(s) + b(s)\big)\, ds, \tag{2.7}$$

i.e. if $x(\cdot)$ is a fixed point of the mapping

$$x(\cdot) \mapsto x_0 + \int_{t_0}^{\cdot} \big(A(s)x(s) + b(s)\big)\, ds.$$

Then theorem 2.8 is a fixed point result and one can use the Picard fixed point theorem to show it.

The resolvent

Let us come back to the study of linear equations in $\mathbb{K}^n$

$$x'(t) = A(t)x(t), \tag{2.5}$$

and denote $\mathcal{E}$ the solution set of this equation.

Proposition 2.9. *The set $\mathcal{E}$ is a vector space of dimension n.*

Preuve.

▷ Clearly, $\mathcal{E}$ is a $\mathbb{K}$ vector space. Let $L_{t_0} : \mathbb{K}^n \to \mathcal{E}$ be the mapping which associates to $x_0 \in \mathbb{K}^n$ the solution $x(\cdot)$ of (2.5) so that $x(t_0) = x_0$. This is a linear application and from the existence and uniqueness of solutions, L_{t_0} is an isomorphism from $\mathbb{K}^n$ onto $\mathcal{E}$.

□

Définition 2.8. We call *resolvent* of equation (2.5) the application $R_A(t,s) : \mathbb{K}^n \to \mathbb{K}^n$ which associates to $x_0 \in \mathbb{K}^n$ the vector $x(t)$, where $x(\cdot)$ is the solution of (2.5) satisfying $x(s) = x_0$.

From the theorems of existence and uniqueness of solutions, the resolvent is linear and bijective. Therefore, $R_A(t,s)$ is an *invertible* matrix of $M_n(\mathbb{K})$. Every solution $x(\cdot)$ of equation (2.5) is given by

$$x(t) = R_A(t,t_0)x(t_0).$$

In particular, in the autonomous case, *i.e.* when $A(\cdot) \equiv A$ is constant, one has $R_A(t,s) = e^{(t-s)A}$.

The resolvent can be characterized by an ODE.

Proposition 2.10.

1. *For every $t_0 \in I$, $R_A(\cdot,t_0)$ is the solution of the matricial ODE*

$$\left\{ \begin{array}{rcl} \dfrac{\partial}{\partial t} R_A(t,t_0) & = & A(t)R_A(t,t_0), \\ R_A(t_0,t_0) & = & I, \end{array} \right. \tag{2.9}$$

2. *For every t_0, t_1, t_2 in I,*

$$R_A(t_2,t_0) = R_A(t_2,t_1) \times R_A(t_1,t_0).$$

3. *If $A(\cdot)$ is of class C^k, then $t \mapsto R_A(t,t_0)$ is of class C^{k+1}.*

Remarques.

— The ODE verified by the resolvent has the same form as (2.5) but it lives in $M_n(\mathbb{K})$ (instead of in $\mathbb{K}^n$).

— The point *2* (or directly the definition) implies that

$$R_A(t,s)^{-1} = R_A(s,t).$$

— The columns $x_1(t), \dots, x_n(t)$ of the resolvent $R_A(t, t_0)$ are the values at t of the solutions equal to $e_1, \dots, e_n$ (canonical basis) at $t = t_0$, since $x_j(t) = R_A(t, t_0)e_j$. The functions $x_1(\cdot), \dots, x_n(\cdot)$ form a basis of the set $\mathcal{E}$ of solutions.

As we have just seen, to solve a linear differential equation, it is enough to know how to calculate the resolvent (as well as for the autonomous equations it is enough to know how to calculate the exponential). Unfortunately, except in the autonomous case, **it is very rare to be able to give an explicit expression of the resolvent**. We will see on the other hand that we can obtain *qualitative* information on the solutions of the equation thanks to the study of the resolvent.

Some properties of the resolvent

Proposition 2.11. *Let $t_0 \in \mathbb{R}$. The function $\Delta(t) = oftR_A(t, t_0)$ verifies the ODE*

$$\begin{cases} \Delta'(t) &= \operatorname{tr}(A(t))\,\Delta(t) \\ \Delta(t_0) &= 1 \end{cases},$$

implying that

$$oftR_A(t, t_0) = \exp\Big(\int_{t_0}^{t} \operatorname{tr}(A(u))\,du\Big).$$

Preuve.

▷ Recall that, if $A \in GL_n(\mathbb{K})$, $D\,oft(A)\cdot H = (\,oftA)\operatorname{tr}(A^{-1}H)$. Then

$$\begin{aligned} \Delta'(t) &= \big(\,oftR_A(t, t_0)\big)\operatorname{tr}\big(R_A^{-1}(t, t_0)R_A'(t, t_0)\big) \\ &= \Delta(t)\operatorname{tr}(A(t)). \end{aligned}$$

□

Corollaire 2.12 (Liouville)**.** *If for every $t \in \mathbb{R}$, $A(t)$ has null trace, then the determinant of $R_A(t, s)$ is identically equal to* 1.

Then, if the trace of $A(t)$ is zero, the ODE (2.5) preserves volumes. Indeed, if Γ is a domain of $\mathbb{R}^n$, let Γ_t be its transport from t_0 to t by the ODE (2.5), *i.e.*

$$\Gamma_t = \{x(t)\ :\ x(\cdot) \text{ solution of (2.5) s.t.. } x(t_0) \in \Gamma\},$$

i.e. $\Gamma_t = R_A(t,t_0)\Gamma$. Then, by using the formula of change of variables in multiple integrals, one gets

$$\mathrm{vol}(\Gamma_t) = |\ oftR_A(t,t_0)|\ \mathrm{vol}(\Gamma),$$

and then $\mathrm{vol}(\Gamma_t) = \mathrm{vol}(\Gamma)$ if $\mathrm{tr}\, A(t) \equiv 0$.

The conservation of the volume has a consequence on the asymptotic behavior of the solutions : it is indeed impossible in that case that all the solutions of (2.5) tend to 0 when $t \to \pm\infty$ (as well as it is impossible that $\|X(t)\|$ tends to infinity for every solution $x(\cdot)$).

A particularly interesting class of matrices of null trace is the skewsymmetric real matrices, which intervene frequently in the problems stemming from physics.

Proposition 2.13. *If, for every $t \in \mathbb{R}$, $A(t)$ is a skewsymmetric real matrix, the resolvent $R_A(t,s)$ is a rotation for every $t, s \in \mathbb{R}$.*

Recall that a rotation is a matrix $R \in M_n(\mathbb{R})$ *orthogonal* (*i.e.* $R^T R = I$) and of determinant equal to 1.

Preuve.

▷ By the previous corollary, one has $oftR_A(t,s) \equiv 1$. Moreover,

$$\begin{aligned}\frac{\partial}{\partial t}\big(R_A(t,s)^T R_A(t,s)\big) &= \frac{\partial}{\partial t}R_A(t,s)^T R_A(t,s) + R_A(t,s)^T \frac{\partial}{\partial t}R_A(t,s)\\ &= R_A(t,s)^T A(t)^T R_A(t,s) + R_A(t,s)^T A(t) R_A(t,s)\\ &= R_A(t,s)^T \big(A(t)^T + A(t)\big) R_A(t,s) = 0.\end{aligned}$$

Then $R_A(t,s)^T R_A(t,s)$ is constant. Since $R_A(s,s) = I$, one concludes.

□

A consequence of this result is that, if $A(t)$ is a skewsymmetric real matrix for any t, the differential equation (2.5) preserves the norm. Indeed, if $x(\cdot)$ is a solution of the equation,

$$\|x(t)\| = \|R_A(t,t_0)x(t_0)\| = \|x(t_0)\|,$$

because $R_A(t,t_0)$ is a rotation. In particular, any solution is bounded . On the other hand it is impossible that a solution tends to 0 (unless $x(\cdot) \equiv 0$). We shall see in the section 2.3.1 that we say 0 is a stable equilibrium point but not asymptotically stable.

2.1.5 Linearization and flow perturbation

In practice, we have almost never an exact knowledge of the initial conditions (or of the very equation, in fact). It is thus essential to know what takes place for the solution of a differential equation when the initial condition is perturbed (or when the equation is perturbed itself) : how varies the interval of definition and the values of the solution, can we give an order of magnitude of these variations, ... ? The answers to these questions are contained in the theorem below : let us give at first the theorem and its proof, we shall explain then why it allows to answer the previous questions.

Recall that we consider an autonomous ODE

$$x'(t) = f\big(x(t)\big), \tag{2.4}$$

the the vecto field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is of class C^1.

Théorème 2.14. *Let $\bar{x}(\cdot)$ be a solution equation (2.4) defined on $[a, b]$ containing 0. Then there exists a neighborhood $\mathcal{V} \subset \Omega$ of $v_0 = \bar{x}(0)$ such that, for everyv $\in \mathcal{V}$, equation (2.4) admits a unique solution $x_v(\cdot)$ defined on $[a, b]$ and verifying $x_v(0) = v$.*
Moreover, the application $v \mapsto x_v(\cdot)$ is of class C^1 on $\mathcal{V}$ and its differential at v_0 is the application which associates to Δv the solution of

$$\left\{ \begin{array}{rcl} y'(t) & = & Df\big(\bar{x}(t)\big) \cdot y(t) \\ y(0) & = & \Delta v \end{array} \right. , \qquad t \in [a, b].$$

Remarque. The solution $x_v(\cdot)$ is simply the restriction of the maximal solution $\phi(\cdot, v)$ to $[a, b]$.

Consequences and meaning of theorem 2.14

a. Domain of definition of the flow The first consequence is that, for any initial condition v in the neighborhood $\mathcal{V}$ of $\bar{x}(0)$, the maximal solution $\phi(\cdot, v)$ is defined on all the interval $[a, b]$, *i.e.* $[a, b] \subset I_v$. In other words, in a informal way, if a solution is defined for a large time, the nearby solutions are also defined for a large time. This is translated by a property of the domain of definition of the flow.

Corollaire 2.15. *The flow ϕ is defined on an open set $\mathcal{D}$ of $\mathbb{R} \times \Omega$.*
In particular, if $(t, v_0) \in \mathcal{D}$, then the application ϕ_t is defined on a neighborhood of v_0.

This property is very important for the study of the flow and its dependence with regard to the initial conditions : Indeed, ϕ and ϕ_t being defined on open, it is now possible to study their continuity and their differentiability.

Preuve.

▷ Recall that the domain of definition of the flow is

$$\mathcal{D} = \{(t, v) \in \mathbb{R} \times \Omega \ : \ t \in I_v\}.$$

Let $(t_0, v_0) \in \mathcal{D}$. Since the maximal interval I_{v_0} is open (theorem 2.3), the maximal solution $\phi(\cdot, v_0)$ in defined on $[a, b] \subset I_{v_0}$ containing t_0. The theorem 2.14 implies that, for every v in a neighborhood $\mathcal{V}$ of v_0, one still has $[a, b] \subset I_v$, *i.e.* the set $]a, b[\times \mathcal{V}$, which is a neighborhood of (t_0, v_0) in $\mathbb{R} \times \Omega$, is included in $\mathcal{D}$.

□

b. Continuous dependence The application $v \mapsto x_v(\cdot)$ defined in the theorem 2.14 is the application which to an initial condition in $\mathcal{V}$ associates the corresponding solution of the differential equation on $[a, b]$. This application being C^1, it is continuous, *i.e.* *the solutions of the differential equation (2.4) depend in a continuous way on their initial condition.*

It is an essential property for the applications (and from a digital point of view) : Indeed, it means, roughly speaking, that the solution calculated from an approximation of the initial condition is an approximation of the real solution. This justifies the use of differential equations in the modeling of real life phenomenons, where only an approximate knowledge of the data is available.

c. Linearized Equation The last part of theorem 2.14 says that the values of the differential of the application $\psi : v \mapsto x_v(\cdot)$ are solutions of a linear equation. This linear equation will play an important role.

Définition 2.9. Let $\bar{x}(\cdot) : [a, b] \to \Omega \subset \mathbb{R}^n$ be a solution of (2.4). The linear equation in $\mathbb{R}^n$

$$y'(t) = Df\big(\bar{x}(t)\big) \cdot y(t), \qquad t \in [a, b],$$

is called *linearized equation of (2.4) around* $\bar{x}(\cdot)$.

For every $\delta v \in \mathbb{R}^n$, $D\psi(\bar{x}(0)) \cdot \delta v$ is the solution of the linearized equation around $\bar{x}(\cdot)$ taking the value δv at $t = 0$. Denoting $R(t, s)$ the resolvent of the linearized equation around $\bar{x}(\cdot)$, one gets, for every $t \in [a, b]$,

$$\big(D\psi(\bar{x}(0)) \cdot \delta v\big)(t) = R(t, 0)\delta v.$$

Let us be now interested in the application ϕ_t. Let us fix a point v_0 of Ω and a time $t \in I_{v_0}$. According to the theorem 2.14, the application ϕ_t is defined and of class C^1 on a neighborhood $\mathcal{V}$ of v_0 in Ω. With the notations above, we have clearly $\phi_t(v) = x_v(t) = (\psi v))(t)$ and thus

$$D\phi_t(v_0) \cdot \delta v = \big(D\psi(v_0) \cdot \delta v\big)(t).$$

We deduce the following result.

Corollaire 2.16. *Let $v_0 \in \Omega$ and $t \in I_{v_0}$. The application ϕ_t is of class C^1 on a neighborhood $\mathcal{V}$ of v_0 and*

$$D\phi_t(v_0) = R(t,0),$$

where R is the resolvent of the linearized equation

$$y'(s) = Df\big(\phi_s(v_0)\big) \cdot y(s), \qquad s \in [0,t].$$

We can give a more intuitive explanation of the role of the linearized equation. We choose a solution $\bar{x}(\cdot) : [a,b] \to \Omega$ of the ODE, the initial condition $v_0 = \bar{x}(0)$. Let us consider now a disturbance $v_0 + \delta v$ of the initial condition and let us write the corresponding solution (*i.e.* $x_{v_0+\delta v}(\cdot)$) under the form of a disturbance $\bar{x}(\cdot) + \delta x(\cdot)$ of the first solution. This disturbance being a solution, it has to satisfy the differential equation :

$$\bar{x}'(t) + (\delta x)'(t) = f\big(\bar{x}(t) + \delta x(t)\big), \qquad t \in [a,b].$$

Using a Taylor expansion, of f at $\bar{x}(t)$ (at fixed t) :

$$f\big(\bar{x}(t) + \delta x(t)\big) = f\big(\bar{x}(t)\big) + Df\big(\bar{x}(t)\big) \cdot \delta x(t) + \text{ second} - \text{order},$$

and taking into account that $\bar{x}'(t) = f\big(\bar{x}(t)\big)$, one gets

$$(\delta x)'(t) = Df\big(\bar{x}(t)\big) \cdot \delta x(t) + \text{ second} - \text{order}.$$

Keeping only first order terms, one finds the linearized equation $(\delta x)'(t) = Df\big(\bar{x}(t)\big) \cdot \delta x(t)$. In other words, the perturbed solution writes $\bar{x}(\cdot) + \delta x(\cdot) + \text{ second} - \text{order}$, where the first order term $\delta x(\cdot)$ is solution of

$$\left\{ \begin{array}{rcl} (\delta x)'(t) & = & Df\big(\bar{x}(t)\big) \cdot (\delta x)(t) \\ (\delta x)(0) & = & \delta v \end{array} \right. .$$

The linearized equation thus indicates how propagates in time a disturbance of the initial condition.

Naturally what precedes is not a rigorous reasoning (the remainders raise obviously some problems !), only a heuristics.

Remarque. The linearized equation is generally a non-autonomous linear equation, we do not know how to calculate thus a priori its solutions. However, if v_0 is an equilibrium point, the maximal solution $\bar{x}(\cdot) = \phi(\cdot, v_0)$ is the constant function $\bar{x}(\cdot) \equiv v_0$, defined on any $\mathbb{R}$, and in that case the linearized equation is autonomous :

$$y'(t) = Df(v_0) \cdot y(t), \qquad t \in \mathbb{R}.$$

Example : divergence-free vector fields. First recall that the divergence of a vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$, $f(x) = (f_1(x), \dots, f_n(x))$, is defined as

$$\operatorname{div} f(x) = \frac{\partial f_1}{\partial x_1}(x) + \cdots + \frac{\partial f_n}{\partial x_n}(x) = \operatorname{tr} Df(x).$$

Consider then a time t and a domain Γ of $\mathbb{R}^n$, included in the domain of definition of ϕ_t. Let $\Gamma_t = \phi_t(\Gamma)$ be the transport of Γ from 0 to t by the equation (2.4). Using the formula of change of variables in multiple integrals, one gets

$$\operatorname{vol}(\Gamma_t) = \int_{\phi_t(\Gamma)} d\mu = \int_\Gamma |\det D\phi_t(x)| d\mu,$$

where, according to corollary 2.16, $D\phi_t(x)$ is the resolvent of the linearized equation.

Suppose now that $\operatorname{div} f(x) = \operatorname{tr} Df(x) \equiv 0$. Liouville theorem (corollary 2.12) implies that the determinant of the linearized system is equal to 1, and then $\det D\phi_t(x) = 1$. One has $\operatorname{vol}(\Gamma_t) = \operatorname{vol}(\Gamma)$, i.e.
if f is a divergence-free vector field, its flow preserves the volume.

Dependence with respect to a parameter

Consider now a family of ODE depending on a parameter $\lambda \in \mathbb{R}^p$:

$$x'(t) = f_\lambda\big(x(t)\big), \tag{2.10}$$

where each f_λ is a vector field on $\Omega \subset \mathbb{R}^n$. Suppose also that $f(x, \lambda) = f_\lambda(x)$ is an application of class C^1. We are interested in the dependence on the solutions of these ODE with respect to the parameter λ.

Notice first that equation (2.10) is equivalent to

$$\left\{ \begin{array}{rcl} x'(t) & = & f\big(x(t), \lambda\big) \\ \lambda'(t) & = & 0 \end{array} \right. ,$$

i.e. to the differential equation in $\mathbb{R}^{n+p}$ associated to the vector field $F(x, \lambda) = (f(x, \lambda), 0)$. The solutions of (2.10) depend of the parameter λ in the same way as the solutions of the differential equation $(x, \lambda)'(t) = F\big((x, \lambda)(t)\big)$ depend on their initial condition. From what precedes, we have the following properties.

- The solutions $\phi^\lambda(\cdot, v)$ of (2.10) depend in a C^1 way, therefore continuous, from the parameter λ (and of the initial condition v) ;
- The differential of the application $(v, \lambda) \mapsto \phi^\lambda(\cdot, v)$ at a point (v_0, λ_0) is the application which to $(\delta v, \delta\lambda)$ associates the solution $y(\cdot)$ of the affine differential equation

$$\left\{ \begin{array}{rcl} y'(t) & = & D_x f\big(\bar{x}(t), \lambda_0\big) \cdot y(t) + D_\lambda f\big(\bar{x}(t), \lambda_0\big) \cdot \delta\lambda \\ y(0) & = & \delta v \end{array} \right. ,$$

where $\bar{x}(\cdot) = \phi^{\lambda_0}(\cdot, v_0)$. In other words, using the formula of the variation of constant,

$$y(t) = R(t, 0)\delta v + \int_0^t R(t, s) D_\lambda f\big(\bar{x}(s), \lambda_0\big) \cdot \delta\lambda \, ds,$$

where $R(t, s)$ is the resolvent associated to the linearized system $y'(t) = D_x f\big(\bar{x}(t), \lambda_0\big) \cdot y(t)$.

*** Poincaré map**

Suppose equation (2.4) admits a non trivial T-periodic solution $x(0) = x_0$. Draw an affine hyperplane Σ passing through x_0 and *transverse* to $x(\cdot)$ at x_0, *i.e.* $f(x_0)$ is not parallel to Σ (note that $f(x_0) \neq 0$ since $x(\cdot)$ is non trivial). For simplicity, we assume that Σ is in fact an affine hyperplane parallel to $\mathbb{R}^{n-1}$ and that $f(x_0) \in \mathbb{R}e_1$, where e_1 verifies $\mathbb{R}^n = \mathbb{R}e_1 \oplus \mathbb{R}^{n-1}$: by a linear change of coordinates, we can always assume that situation holds.

Proposition 2.17. *There exists a neighborhood $\mathcal{V} \subset \mathbb{R}^n$ of x_0 and a function η of class C^1 such that, for every z in $\Sigma \cap \mathcal{V}$, $\phi_{T+\eta(z)}(z)$ belongs to $\Sigma \cap \mathcal{V}$. The application $G : \Sigma \cap \mathcal{V} \to \Sigma \cap \mathcal{V}$ thus defined is a C^1 diffeomorphism : this is* the first return Poincaré map.

Moreover, for every $(\delta s, \delta z) \in \mathbb{R} \times \mathbb{R}^{n-1}$,

$$D\phi_T(x_0) \cdot \begin{pmatrix} \delta s \\ \delta z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & D_z G(x_0) \end{pmatrix} \cdot \begin{pmatrix} \delta s \\ \delta z \end{pmatrix}.$$

Remarque. In the case of flows, the Poincaré map depends only on transverse sections in the orbit, Σ_0, Σ_1, and still up to conjugation : it has therefore a very strong geometrical sense. On the other hand this construction is very useful in the study of the neighborhood of a periodic orbit. Finally note that the previous construction allows one to pass from the study of a flow in dimension n in that of a local diffeomorphism in dimension $n-1$.

2.2 Autonomous linear differential equations

In this chapter, we study the simplest ODE namely the autonomous linear differential equations also called linear equations with constant coefficients, *i.e.* equations of the form

$$x'(t) = Ax(t). \tag{2.11}$$

The datum $A \in M_n(\mathbb{K})$ is an $(n \times n)$ matrix with coefficients in $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. The unknown is a differentiable application $x(\cdot) : \mathbb{R} \to \mathbb{K}^n$. Solving equation (2.11) means finding an application $x(\cdot)$ such that, for every $t \in \mathbb{R}$, the derivative $x'(t) = \frac{dx}{dt}(t)$ verifies $x'(t) = Ax(t)$.

This equation is said to be autonomous since $A \in M_n(\mathbb{R})$ does not depend on time.

2.2.1 Elementary approach

Consider a system of two differential equations

$$\begin{cases} x_1' &= \alpha_1 x_1 \\ x_2' &= \alpha_2 x_2 \end{cases}.$$

It is a simple system since the functions $x_1(t)$ et $x_2(t)$ are decoupled. The solution of the system is

$$x_1(t) = x_1(t_0)e^{\alpha_1(t-t_0)}, \quad x_2(t) = x_2(t_0)e^{\alpha_2(t-t_0)}.$$

One has a complete knowledge of the behavior of the solutions : if $\alpha_1 < 0$ and $\alpha_2 < 0$, every solution $(x_1(t), x_2(t))$ tends to the origin as $t \to +\infty$; if $\alpha_1 > 0$ and $x_1(t_0) \neq 0$, the norm of $(x_1(t), x_2(t))$ tends to infinity as $t \to +\infty \ldots$

Let us use a matrix notation. Set $x = (x_1, x_2)$. The above system of two ODEs apppears as a particular case $n = 2$ of the following differential equation in $\mathbb{R}^n$:

$$x'(t) = \Delta x(t), \quad \text{where } \Delta = \begin{pmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_n \end{pmatrix} \text{ is diagonal.} \tag{2.12}$$

This equation being a system of n decoupled scalar equations, its solution is given by

$$x(t) = \begin{pmatrix} x_1(t_0)e^{\alpha_1(t-t_0)} \\ \vdots \\ x_n(t_0)e^{\alpha_n(t-t_0)} \end{pmatrix} = \begin{pmatrix} e^{\alpha_1(t-t_0)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\alpha_n(t-t_0)} \end{pmatrix} x(t_0),$$

with $x = (x_1, \ldots, x_n)$. In the next section, that diagonal matrix is the exponential of Δ denoted $e^{(t-t_0)\Delta}$.

So, for the differential equations of the shape (2.12), we have a perfect knowledge of the solutions. We are for example in measure to analyze the asymptotic behavior of the solution functions of $\alpha_1, \ldots, \alpha_n$ and of initial condition $x(t_0)$:

— if all $\alpha_i < 0$, every solution $x(t)$ converges to the origin as $t \to +\infty$;
— if all $\alpha_i \leq 0$, every solution $x(t)$ is bounded as $t \to +\infty$;
— if at least one $\alpha_i > 0$, then $\lim_{t\to+\infty} \|x(t)\| = +\infty$ for every solution verifying $x_i(t_0) \neq 0$;
— andc...

The differential equation $x'(t) = \Delta x(t)$ that we have just dealt with is a particular case, because it corresponds to a system of n scalar decoupled equations. Many of the linear differential equations can however come down to it. Consider indeed the system $x'(t) = Ax(t)$ with A diagonalizable in $\mathbb{R}$: there exists therefore an invertible matrix $P \in GL_n(\mathbb{R})$ and a diagonal matrix $\Delta \in M_n(\mathbb{R})$ such that $A = P\Delta P^{-1}$.

Let us notice now that, if $x(t)$ verifies $x'(t) = Ax(t)$, then $y(t) = P^{-1}x(t)$ verifies $y'(t) = \Delta y(t)$. In other words, *up to a change of coordinates*, the differential equation is a system of n scalar decoupled equations. Knowing $y(t_0) = P^{-1}x(t_0)$, we obtain then $y(t) = e^{(t-t_0)\Delta}y(t_0)$, and

$$x(t) = Py(t) = Pe^{(t-t_0)\Delta}y(t_0) = Pe^{(t-t_0)\Delta}P^{-1}x(t_0).$$

We are therefore still able to compute the solutions of the ODE in this case. More importantly, the asymptotic behavior of the solutions is characterized by the diagonal elements of Δ, *i.e.* by the *eigenvalues* of A.

In summary, this elementary approach shows the key steps that we will develop :
— solutions are computed using the matrix exponential ;
— the asymptotic behavior of the solutions is characterized by the eigenvalues of A.

2.2.2 Exponential of matrices

Définition 2.10. We call *exponential of matrices* the application

$$\begin{array}{rccc} \exp: & M_n(\mathbb{K}) & \longrightarrow & M_n(\mathbb{K}) \\ & A & \longmapsto & \exp A = e^A = \displaystyle\sum_{k=0}^{\infty} \frac{A^k}{k!} \end{array} .$$

Note that the series $\sum_{k=0}^{\infty} \frac{A^k}{k!}$ converges normally. Indeed

$$\sum_{k=0}^{\infty} \frac{\|A^k\|}{k!} \leq \sum_{k=0}^{\infty} \frac{\|A\|^k}{k!} = e^{\|A\|} < \infty,$$

where we chose $\|\cdot\|$ as a multiplicative norm on $M_n(\mathbb{K})$ (for instance an operator norm). The application exponential is therefore continuous (in fact C^∞). Recall its main properties.

Proposition 2.18.

1. *For $A \in M_n(\mathbb{K})$, the application $t \mapsto e^{tA}$ is differentiable and*
$$\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA}A$$
2. *for every $A \in M_n(\mathbb{K})$, $\exp(A)$ is invertible and $\exp(A)^{-1} = \exp(-A)$.*
3. *If A and $B \in M_n(\mathbb{K})$* commute, *i.e. $AB = BA$, then*
$$\exp(A+B) = \exp(A)\exp(B).$$
4. *If $P \in GL_n(\mathbb{K})$, then $Pe^AP^{-1} = e^{PAP^{-1}}$.*
5. *If Δ is the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_n$, then e^Δ is diagonal with diagonal elements $e^{\lambda_1}, \ldots, e^{\lambda_n}$.*

Let us solve now equation (2.11).

Théorème 2.19. *Let $t_0 \in \mathbb{R}$ and $x_0 \in \mathbb{K}^n$. The unique solution of the equation $x'(t) = Ax(t)$ taking value x_0 at t_0 is the application $x(\cdot)$ defined by*

$$x(t) = e^{(t-t_0)A}x_0, \qquad \forall t \in \mathbb{R}.$$

Preuve.

▷ The first property of the previous proposition implies

$$\frac{d}{dt}\left(e^{(t-t_0)A}x_0\right) = \left(\frac{d}{dt}e^{(t-t_0)A}\right)x_0 = Ae^{(t-t_0)A}x_0.$$

The function $x(t) = e^{(t-t_0)A}x_0$ is therefore solution of $x'(t) = Ax(t)$, and $x(t_0) = x_0$ since $e^{0A} = I$.

To show uniqueness of the solution, consider another solution $y(t)$ taking value x_0 at t_0 and set $z(t) = e^{-(t-t_0)A}y(t)$. Using again proposition 2.18, one gets

$$z'(t) = -Ae^{-(t-t_0)A}y(t) + e^{-(t-t_0)A}y'(t) = -Ae^{-(t-t_0)A}y(t) + e^{-(t-t_0)A}Ay(t) = 0,$$

since A and e^{tA} commute. Therefore $z(t)$ is constant, and since $z(t_0) = x_0$, one gets $y(t) = e^{(t-t_0)A}z(t_0) = e^{(t-t_0)A}x_0$.

□

Hence solving autonomous linear differential equations reduces to computing matrix exponentials. Notice that the two last points of proposition 2.18 allow one to recover the result of section 2.2.1 : if A is diagonalizable in $\mathbb{K}$, i.e.

$$A = P\Delta P^{-1}, \quad P \in GL_n(\mathbb{K}), \quad \Delta \in M_n(\mathbb{K}) \text{ diagonal},$$

then $e^{tA} = Pe^{t\Delta}P^{-1}$ and the solution of equation (2.11) is

$$x(t) = Pe^{(t-t_0)\Delta}P^{-1}x(t_0).$$

Unfortunatly, matrices are not all diagonalizable. The theory of matrix reduction however overcomes that difficulty.

2.2.3 Computation of the exponential of matrices

The goal of this section consists of computing the exponential of a matrix $A \in M_n(\mathbb{K})$ to up to a change of coordinates, *i.e.* computing $P^{-1}e^{tA}P$ for an appropriate $P \in GL_n(\mathbb{K})$.

We first do that in the case $\mathbb{K} = \mathbb{C}$, *i.e.* for matrices with complex coefficients. Then, we will treat the case $\mathbb{K} = \mathbb{R}$.

a. Matrices with complex coefficients

Characteristic Polynomial. Consider a matrix $A \in M_n(\mathbb{C})$ and denote $\lambda_1, \ldots, \lambda_r$ its *eigenvalues*. Recall that they are the only complex numbers so that

$$Av = \lambda_i v$$

for a non zero $v \in \mathbb{C}^n$. The corresponding v_i are called *eigenvectors*). The eigenvalues are obtained as the roots of the *characteristic polynomial* of A : $P_A(\lambda) = \det(\lambda I - A)$. This polynomial is therefore of the form

$$P_A(\lambda) = (\lambda - \lambda_1)^{p_1} \cdots (\lambda - \lambda_r)^{p_r},$$

where each integer p_i is strictly positive and $p_1 + \cdots + p_r = n$ (the characteristic polynomial is of degree n). We call p_i the *algebraic multiplicity* of the eigenvalue λ_i.

An important property of the characteristic polynomial is the following.

Théorème 2.20 (Cayley-Hamilton)**.** *Every matrix annihilates its characteristic polynomial :*

$$P_A(A) = (A - \lambda_1 I)^{p_1} \cdots (A - \lambda_r I)^{p_r} = 0.$$

Eigenspaces, characteristic subspaces. To each eigenvalue of A are associated two vectorial subspaces of $\mathbb{C}^n$. The first one is the *eigenspace* :

$$\Pi_i \text{ ou } \Pi_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I).$$

It is the set of eigenvectors associated to λ_i. The integer $e_i = \dim \Pi_i$ is called the *geometric multiplicity* of the eigenvalue λ_i.

The second vectorial subspace is the *characteristic subspace* :

$$\Gamma_i \text{ ou } \Gamma_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I)^{p_i}.$$

Clearly, $\Pi_i \subset \Gamma_i$, but these two subspaces can be different. The role of these subspaces are explained next (with no proof).

Théorème 2.21 (of kernel decomposition)**.** *With the notations above, we have the decomposition*

$$\mathbb{C}^n = \Gamma_1 \oplus \cdots \oplus \Gamma_r,$$

and the following properties :

1. $\dim \Gamma_i = p_i$ *;*
2. *each* Γ_i *is invariant by* A *:* $x \in \Gamma_i \Rightarrow Ax \in \Gamma_i$ *;*
3. *the restriction* $A|_{\Gamma_i}$ *of* A *to* Γ_i *is written*

$$A|_{\Gamma_i} = \lambda_i I_{\Gamma_i} + N_i,$$

where I_{Γ_i} *is the identity map of* Γ_i *and* $N_i \in \mathrm{End}(\Gamma_i)$ *is nilpotent of order* $\leq p_i$*, i.e.* $N_i^{p_i} = 0$.

— Recall that $\mathrm{End}(\Gamma_i)$ is the set of the *endomorphisms* of Γ_i, *i.e.* of the linear applications of Γ_i in itself. Saying that Γ_i is invariant by A is equivalent to say that $A|_{\Gamma_i}$ belongs to $\mathrm{End}(\Gamma_i)$.
— The operator N_i is defined as $N_i = (A - \lambda_i I)|_{\Gamma_i}$. The fact that N_i is nilpotent of order $\leq p_i$ is exactly the definition of Γ_i. It is possible that the exact order of nilpotency of N_i, *i.e.* the smallest integer $m_i \leq p_i$ such that $N_i^{m_i} = 0$, is smaller than p_i. In that case, one has

$$\ker(A - \lambda_i I)^{p_i} = \ker(A - \lambda_i I)^{m_i} \supsetneq \ker(A - \lambda_i I)^{m_i - 1}.$$

— A matrix is *diagonalizable* if there exists a basis of $\mathbb{C}^n$ made of eigenvectors, *i.e.*

$$\mathbb{C}^n = \Pi_1 \oplus \cdots \oplus \Pi_r.$$

According to the theorem of kernel decomposition, it is possible only if $\Pi_i = \Gamma_i$ for every i, *i.e.*
A is diagonalizable if and only if for every eigenvalue the algebraic and geometric multiplicities coincide, i.e. $\dim \Pi_i = p_i$ *pour* $i = 1, \ldots, r$.

Jordan Reduction in $\mathbb{C}^n$. Let us choose a basis $\mathcal{B}$ of $\mathbb{C}^n$ formed of the union of a basis of Γ_1, of a basis of Γ_2, ..., of a basis of Γ_r, and let us denote P the matrix of passage of this basis to the canonical basis. According to the theorem of kernel decomposition, the linear application associated to A has for matrix in the basis $\mathcal{B}$:

$$P^{-1}AP = \Delta + N,$$

where Δ is the diagonal matrix with diagonal elements λ_1 (p_1 times), ..., λ_r (p_r times), and N is the nilpotent matrix which is written by blocks

$$N = \begin{pmatrix} N_1 & & \\ & \ddots & \\ & & N_r \end{pmatrix}.$$

Il is in fact possible to choose the basis $\mathcal{B}$ so that the nilpotent matrix N takes a simple form. This way, we end up to the Jordan reduction.

Théorème 2.22 (of Jordan)**.** *For every matrix $A \in M_n(\mathbb{C})$, there exists $P \in GL_n(\mathbb{C})$ such that $P^{-1}AP$ is written as a diagonal matrix by blocks*

$$P^{-1}AP = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix}, \tag{2.14}$$

where each J_i is a $(p_i \times p_i)$ matrix of the form

$$J_i = \begin{pmatrix} J_{i,1} & & \\ & \ddots & \\ & & J_{i,e_i} \end{pmatrix}, \quad \textit{avec} \quad J_{i,k} = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}$$

and $e_i = \dim \ker(A - \lambda_i I)$ is the geometric multiplicity of λ_i.

We call the matrix $J = P^{-1}AP$ the *reduced Jordan form* of A and the matrices J_i the *Jordan blocks.*

Remarques.

— For every i, the matrix J_i represents the linear application $A|_{\Gamma_i}$, implying that $J_i \in \mathrm{End}(\Gamma_i)$.

— The matrices $J_{i,k}$ are square matrices whose dimension $\dim(J_{i,k}) \leq p_i$ depends of i and k. In the particular case where the algebraic and geometrical multiplicities of λ_i coincide (*i.e.* $e_i = p_i$), the dimension $\dim(J_{i,k})$ is equal to 1 for every k. Every block $J_{i,k}$ is then reduced to the scalar λ_i and $J_i = \lambda_i I_{p_i}$.

— According to the preceding remark, if for every eigenvalue the algebraic and geometrical multiplicities coincide, the reduced Jordan form is diagonal. So the reduction of Jordan generalizes the diagonalization. Let us insist however on the fact that any matrix of $M_n(\mathbb{C})$ admits a reduced Jordan form but is not necessarily diagonalizable.

Computation of matrix exponential. The Jordan reduction allows one to compute e^{tA} up to conjugation. Indeed, let us start with the computation of the exponential of a block $J_{i,k}$. Denote $n_{i,k} = \dim(J_{i,k})$ the dimension of this block and write it $J_{i,k} = \lambda_i I + N_{i,k}$, where $N_{i,k}$ is the matrix having 0 on the main diagonal and 1 just above. As $\lambda_i I$ and $N_{i,k}$ commute,

$$e^{tJ_{i,k}} = e^{t\lambda_i I} e^{tN_{i,k}} = e^{t\lambda_i} e^{tN_{i,k}}.$$

and, $N_{i,k}$ being nilpotent of order $n_{i,k}$, one has :

$$e^{tN_{i,k}} = \sum_{l=0}^{\infty} \frac{(tN_{i,k})^l}{l!} = \sum_{l=0}^{n_{i,k}-1} \frac{(tN_{i,k})^l}{l!} = \begin{pmatrix} 1 & t & \dots & \frac{t^{n_{i,k}-1}}{(n_{i,k}-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}.$$

On the other hand, by the properties of the exponential of a matrix (proposition 2.18), one has $e^{tA} = e^{tPJP^{-1}} = Pe^{tJ}P^{-1}$, yielding finally

$$e^{tA} = P \begin{pmatrix} e^{tJ_{1,e_1}} & & \\ & \ddots & \\ & & e^{tJ_{r,e_r}} \end{pmatrix} P^{-1}, \tag{2.15}$$

$$\text{avec } e^{tJ_{i,k}} = e^{t\lambda_i} \begin{pmatrix} 1 & t & \dots & \frac{t^{n_{i,k}-1}}{(n_{i,k}-1)!} \\ & \ddots & \ddots & \vdots \\ & & \ddots & t \\ & & & 1 \end{pmatrix}.$$

b. Matrices with real coefficients

Consider now a matrix $A \in M_n(\mathbb{R})$, with real coefficients. We can naturally consider A as a matrix of $M_n(\mathbb{C})$; all that we have obtained for matrices with complex coefficients applies therefore.

Let us denote the real eigenvalues of A by $\lambda_1, \dots, \lambda_s$, and the non real eigenvalues by $\lambda_{s+1}, \bar{\lambda}_{s+1}, \dots, \lambda_q$ (with $2q - s = r$). The characteristic polynomial of A is therefore the polynomial with real coefficients

$$P_A(\lambda) = \prod_{i=1}^{s} (\lambda - \lambda_i)^{p_i} \prod_{i=s+1}^{q} [(\lambda - \lambda_i)(\lambda - \bar{\lambda}_i)]^{p_i}.$$

The subspaces $\Gamma_{\lambda_i} = \ker_{\mathbb{C}}(A - \lambda_i I)^{p_i}$ of $\mathbb{C}^n$ are now called *complex* characteristic subspaces.

Remarque. Recall the links which exist between the vectorial subspaces of $\mathbb{C}^n$, which are vectorial $\mathbb{C}$-spaces, and those of $\mathbb{R}^n$, which are vectorial $\mathbb{R}$-spaces. We consider $\mathbb{R}^n$ as a subset of $\mathbb{C}^n$ and, for a vectorial subspace Γ of $\mathbb{C}^n$, we note $\Gamma \cap \mathbb{R}^n$ the set of the vectors $v \in \Gamma$ which are real. It is easy to verify that such a set $\Gamma \cap \mathbb{R}^n$ is a vectorial subspace of $\mathbb{R}^n$. Moreover, if Γ is stable by conjugation (*i.e.* $v \in \Gamma \Rightarrow \bar{v} \in \Gamma$), then Γ and $\Gamma \cap \mathbb{R}^n$ have the same dimension as subspaces respectively of $\mathbb{C}^n$ and of $\mathbb{R}^n$ (in fact, in this case, Γ is the set of linear combinations with complex coefficients of elements of $\Gamma \cap \mathbb{R}^n$; and thus every basis of the $\mathbb{R}$-space $\Gamma \cap \mathbb{R}^n$ is also a basis of the $\mathbb{C}$-space Γ).

Define now the *real charateristic subspaces* of A as the subspaces of $\mathbb{R}^n$:

$$\begin{aligned} E_i &= \Gamma_{\lambda_i} \cap \mathbb{R}^n, & 1 \le i \le s \\ E_i &= (\Gamma_{\lambda_i} \oplus \Gamma_{\bar{\lambda}_i}) \cap \mathbb{R}^n, & s+1 \le i \le q. \end{aligned}$$

Let us notice while $(A - \lambda_i I)^{p_i} v = 0$ implies $(A - \bar{\lambda}_i I)^{p_i} \bar{v} = 0$, meaning that Γ_{λ_i} for λ_i real and $\Gamma_{\lambda_i} \oplus \Gamma_{\bar{\lambda}_i}$ for λ_i not real are stable by conjugation. According to the preceding remark and the theorem of kernel decomposition in $\mathbb{C}^n$, we have the decomposition

$$\mathbb{R}^n = E_1 \oplus \cdots \oplus E_q,$$

every sub-space E_i being invariant by A. From this decomposition, we give below a reduced real Jordan form of A. This reduced form is not however crucial to the study of ODE : we shall see in the following section that the solutions of the equation (2.11) in $\mathbb{R}^n$ can be deduced directly of the solutions in $\mathbb{C}^n$.

***Jordan Reduction in $\mathbb{R}^n$.** To give a reduced form of $A \in M_n(\mathbb{R})$ consists of finding for every characteristic subspace a basis in which the linear application associated to A has a simple expression (what we have done for matrices with complex coefficients). Consider therefore a real characteristic subspace E_k of A.

— If λ_k is real, *i.e.* $1 \le k \le s$: in this case $E_i = \ker_{\mathbb{R}}(A - \lambda_k I)^{p_k}$ and the restriction of A to E_k is simply given by

$$A|_{E_k} = \lambda_k I|_{E_k} + N_k, \qquad \text{where } N_k \text{ nilpotent.}$$

As in the complex case, one can show that $A|_{E_k}$ is conjugate to the Jordan block J_k.

— If $\lambda_k = \alpha_k + i\beta_k$ is not real, *i.e.* $s+1 \le k \le q$: choose a basis $v_1, \ldots, v_{p_k}$ of Γ_k in which $A|_{\Gamma_k}$ is put as a Jordan block J_k, i.e. , for $j = 1, \ldots, p_k$,

$$Av_j = \lambda_k v_j + \delta_j v_{j-1}, \qquad \text{where } \delta_j = (J_k)_{j-1,j} = 0 \text{ ou } 1,$$

by setting $v_0 = 0$. by conjugation, it is clear that $\bar{v}_1, \ldots, \bar{v}_{p_k}$ is a basis of $\Gamma_{\bar{\lambda}_k}$. For $j = 1, \ldots, p_k$, set $v_j = a_j + ib_j$, with $a_j, b_j \in \mathbb{R}^n$. The vectors $(a_1, b_1, \ldots, a_{p_k}, b_{p_k})$ form a basis of E_k. Indeed, the $a_j = \frac{1}{2}(v_j + \bar{v}_j)$ and $b_j = \frac{i}{2}(\bar{v}_j - v_j)$ belong to $\Gamma_{\lambda_k} \oplus \Gamma_{\bar{\lambda}_k}$ and span it as a $\mathbb{C}$-space since they span the basis of the v_j, $\bar{v}_j$: they therefore form a generating family with $2p_k$ elements of the $\mathbb{R}$-space E_k with dimension $2p_k$.
Identifying the real and imaginary parts of the expression below,

$$A(a_j + ib_j) = (\alpha_k + i\beta_k)(a_j + ib_j) + \delta_j(a_{j-1} + ib_{j-1}),$$

one gets

$$A \begin{bmatrix} a_j \\ b_j \end{bmatrix} = C_k \begin{bmatrix} a_j \\ b_j \end{bmatrix} + \delta_j \begin{bmatrix} a_{j-1} \\ b_{j-1} \end{bmatrix} \qquad \text{where} \quad C_k = \begin{pmatrix} \alpha_k & -\beta_k \\ \beta_k & \alpha_k \end{pmatrix}.$$

The restriction of A to E_k is therefore conjugate in the basis $(a_1, b_1, \ldots, a_{p_k}, b_{p_k})$ with the matrix

$$J'_k = \begin{pmatrix} C_k & \delta_2 I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \delta_{p_k} I_2 \\ & & & C_k \end{pmatrix}.$$

One hence gets the Jordan reduction of A in $\mathbb{R}^n$.

Théorème 2.23. *for every matrix $A \in M_n(\mathbb{R})$, there exists $Q \in GL_n(\mathbb{R})$ such that $Q^{-1}AQ$ is written as a diagonal matrix by blocks J', with diagonal elements $J_1, \ldots, J_s$, $J'_{s+1}, \ldots, J'_q$, where, for $i = s+1, \ldots, q$, each J'_i is a $(2p_i \times 2p_i)$ matrix of the form*

$$J'_i = \begin{pmatrix} J'_{i,1} & & \\ & \ddots & \\ & & J'_{i,2e_i} \end{pmatrix}, \quad \text{avec} \quad J'_{i,k} = \begin{pmatrix} C_i & I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & I_2 \\ & & & C_i \end{pmatrix}$$

and, for $i = 1, \ldots, s$, the matrices J_i are those given in theorem 2.22.

The computation of the exponential is now straitghforward : it is enough to compute e^{tC_i}. Write C_i as the sum of two commuting matrices

$$C_i = \alpha_i I_2 + \beta_i B, \qquad B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

and therefore its exponential as a product

$$e^{tC_i} = e^{\alpha_i t} e^{\beta_i t B}.$$

Note that the square of the matrix B is equal to $-I_2$, implying that $B^{2p} = (-1)^p I_2$ and $B^{2p+1} = (-1)^p B$, *i.e.*

$$e^{sB} = \sum_{k=0}^{\infty} \frac{1}{k!}(sB)^k = \sum_{p=0}^{\infty} \frac{1}{(2p)!}(-1)^p s^{2p} I_2 + \sum_{p=0}^{\infty} \frac{1}{(2p+1)!}(-1)^p s^{2p+1} B = \cos(s) I_2 + \sin(s) B.$$

One hence gets the exponential of C_i :

$$e^{tC_i} = e^{\alpha_i t} \begin{pmatrix} \cos(\beta_i t) & -\sin(\beta_i t) \\ \sin(\beta_i t) & \cos(\beta_i t) \end{pmatrix}.$$

2.2.4 Form of the solutions

Thanks to the calculation of the exponential which we have just made, we are now able to specify theorem 2.19. Let us give at first the form of the general solution in $\mathbb{C}^n$, which is obtained directly from the expression (2.15) of e^{tA}.

Théorème 2.24. *Soit $A \in M_n(\mathbb{C})$. Every solution of $x'(t) = Ax(t)$ in $\mathbb{C}^n$ is written as*

$$x(t) = \sum_{1 \le i \le r} e^{t\lambda_i} \left(\sum_{0 \le k \le m_i - 1} t^k \, v_{i,k} \right), \qquad \text{where} \quad v_{i,k} \in \Gamma_i, \tag{2.17}$$

avec $m_i = \max_{1 \le k \le e_i} \dim(J_{i,k})$.

Remarques.

— The quantity in factor of $e^{t\lambda_i}$ is polynomial in t if $m_i > 1$, and constant if $m_i = 1$. Recall that the latter occurs if and only if the algebraic and geometric multiplicities coincide.

— Note also the dependence of $x(t)$ with respect to the initial condition $x(0)$ in the previous decomposition. If $x(0)$ is written $x(0) = v_1 + \cdots + v_r$ in $\mathbb{C}^n = \Gamma_1 \oplus \cdots \oplus \Gamma_r$, then

$$v_{i,k} = \frac{1}{k!} N^k v_i,$$

where N is the nilpotent matrix of the decomposition $\Delta + N$ of $P^{-1}AP$. In particular, $v_i = 0$ if and only if all the vectors $v_{i,k}$ are zero. It follows from the invariance of the subspaces Γ_i by A.

Consider now a matrix A with real coefficients and a solution $x(\cdot)$ in $\mathbb{R}^n$ of $x' = Ax$, whise initial condition is $x(0) \in \mathbb{R}^n$. One can consider A as matrix with complex coefficients and $x(\cdot) = x(\cdot) + i0$ as the solution in $\mathbb{C}^n$ of $x' = Ax$ with initial condition $x(0) + i0$. The expression of $x(\cdot)$ is therefore given by the formula (2.17). Since this solution is real, it is in fact equal to the real part of formula (2.17), the imaginary part must equal to zero. One hence gets the general form of the solutions of $x' = Ax$ in $\mathbb{R}^n$.

Théorème 2.25. *Let $A \in M_n(\mathbb{R})$. Every solution of $x'(t) = Ax(t)$ in $\mathbb{R}^n$ is written*

$$x(t) = \sum_{1 \le i \le q} e^{t\alpha_i} \left(\sum_{0 \le k \le m_i - 1} t^k \big(\cos(\beta_i t)\, a_{i,k} + \sin(\beta_i t)\, b_{i,k} \big) \right), \tag{2.19}$$

where $\alpha_i = \Re e(\lambda_i)$, $\beta_i = \Im m(\lambda_i)$ and the vectors $a_{i,k}$, $b_{i,k}$ belong to E_i.

Remarque. As in theorem 2.24, the vectors $a_{i,k}$, $b_{i,k}$ only depend of the initial condition $x(0)$. If $x(0)$ writes $x(0) = u_1 + \cdots + u_q$ in $\mathbb{R}^n = E_1 \oplus \cdots \oplus E_q$,then $u_i = 0$ if and only if all the vectors $a_{i,k}$ and $b_{i,k}$ are zero.

Asymptotic Behavior.

Theorems 2.24 and 2.25 give all the necessary information on the differential equation, generalizing the results of the section 2.2.1 on the scalar equations and the systems of decoupled equations. We notice in particular that the behavior when t tends to infinity of the solutions $x(t)$ of $x'(t) = Ax(t)$ depends essentially of the signs of the real parts of the eigenvalues λ_i of A. More exactly, we can decompose the behavior of the components of $x(t)$ on every characteristic subspace (we suppose here A real) :

— if $\Re e(\lambda_i) < 0$ the projection on E_i of $x(t)$ tends to zero as t tends to $+\infty$ and increases exponentially at $-\infty$;

— if $\Re e(\lambda_i) > 0$, it is the contrary, the projection on E_i of $x(t)$ increases exponentially at $+\infty$ and tends to zero as t tends to $-\infty$;
— if $\Re e(\lambda_i) = 0$, the component on E_i of $x(t)$ increases polynomially at $\pm\infty$ when $\dim \Pi_i < p_i$, and is bounded for $t \in \mathbb{R}$ when $\dim \Pi_i = p_i$.

It is useful to gather the characteristic subspaces depending on the sign of the real part of the corresponding eigenvalues. For $A \in M_n(\mathbb{R})$, we define

– the *stable* space : $$E^s = \Big[\bigoplus_{\Re e(\lambda_i)<0} \Gamma_i \Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)<0} E_i,$$
– the *instable* space : $$E^u = \Big[\bigoplus_{\Re e(\lambda_i)>0} \Gamma_i \Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)>0} E_i,$$
– the *indifferent* space : $$E^c = \Big[\bigoplus_{\Re e(\lambda_i)=0} \Gamma_i \Big] \cap \mathbb{R}^n = \bigoplus_{\Re e(\lambda_i)=0} E_i,$$

so that $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$. Similarly, for $A \in M_n(\mathbb{C})$, lthe complex spaces *stable*, *instable* et *indifferent* are defined respectively as

$$\Gamma^s = \bigoplus_{\Re e(\lambda_i)<0} \Gamma_i, \qquad \Gamma^u = \bigoplus_{\Re e(\lambda_i)>0} \Gamma_i, \qquad \Gamma^c = \bigoplus_{\Re e(\lambda_i)=0} \Gamma_i,$$

and one has the decomposition $\mathbb{C}^n = \Gamma^s \oplus \Gamma^u \oplus \Gamma^c$.

According to the theorem of kernel decomposition, these subspaces are invariant by e^{tA} for every $t \in \mathbb{R}$: $e^{tA}E^s \subset E^s$, $e^{tA}E^u \subset E^u$, etc... which will imply that, if a solution$x(\cdot)$ of $x'(t) = Ax(t)$ verifies for instance $x(0) \in E^s$, then $x(t) \in E^s$ for every t ; if $x(0) \in E^c$, then $x(t) \in E^c$ for every t, etc...

Each of the stable, unstable and on different spaces corresponds to a certain type of asymptotic behavior of the solutions. We summarize this behavior in the following theorem, whose demonstration is left in exercise (use either the form of the solutions, or directly the Jordan reduction).

Théorème 2.26. *Let A be a real (resp. complex) $(n \times n)$ matrix. Denote by $x(\cdot)$ the solutions in $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) of the differential equation $x'(t) = Ax(t)$. Then*

— *E^s (resp. Γ^s) is the set of the $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) for which*

$$\lim_{t \to +\infty} \|x(t)\| = 0;$$

— *E^u (resp. Γ^u) is the set of the $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) for which*

$$\lim_{t \to -\infty} \|x(t)\| = 0;$$

— *E^c (resp. Γ^c) is the set of the $x(0) \in \mathbb{R}^n$ (resp. $x(0) \in \mathbb{C}^n$) for which there exists an integer $M \geq 0$ and a constant $C > 0$ such that, pour $|t|$ large enough,*

$$C^{-1} \|x(0)\| \leq \|x(t)\| \leq C\,|t|^M \|x(0)\|.$$

Moreover, for $0 < \alpha < \min_{\Re e(\lambda_i) \neq 0} |\Re e(\lambda_i)|$, there exists a constant $C > 0$ such that :

— *if $x(0) \in E^s$ (or Γ^s), then for every $t > 0$ large enough,*

$$\|x(t)\| \leq Ce^{-\alpha t}\|x(0)\|, \qquad \|x(-t)\| \geq C^{-1}e^{\alpha t}\|x(0)\|;$$

— *if $x(0) \in E^u$ (ou Γ^u), then for every $t > 0$ large enough,*

$$\|x(t)\| \geq C^{-1}e^{\alpha t}\|x(0)\|, \qquad \|x(-t)\| \leq Ce^{-\alpha t}\|x(0)\|.$$

Remarque. In the characterization of E^c, one can take $M = 0$ if and only if A is diagonalizable (since in that case, for every eigenvalue, the algebraic and geometric multiplicities coincide).

To get the asymptotic behavior of a particular solution $x(\cdot)$, it is enough to decompose its initial condition $x(0)$ as $x^s(0)+x^u(0)+x^c(0)$ in $\mathbb{R}^n = E^s \oplus E^u \oplus E^c$. We know that the decomposition of $x(t)$ is $x(t) = x^s(t) + x^u(t) + x^c(t)$, where $x^s(\cdot)$ (resp. $x^u(\cdot)$, $x^c(\cdot)$) is the solution of $x'(t) = Ax(t)$ with $x^s(0)$ (resp. $x^u(0)$, $x^c(0)$) as initial condition. In particular, if we consider positive times, we have the following :

— if $x^u(0) \neq 0$, then $\|x(t)\|$ tends to infinity as $t \to +\infty$;
— if $x^c(0) \neq 0$, then $\|x(t)\|$ does not tend to 0 as $t \to +\infty$ (but $\|x(t)\|$ does not tend necessarily to infinity).

Définition 2.11. A square matrix A is **Hurwitz** if all its eigenvalues have negative real part (*i.e.* $E^s = \mathbb{R}^n$ ou $\Gamma^s = \mathbb{C}^n$).

Corollaire 2.27. *All solutions of $x'(t) = Ax(t)$ tend to 0 as $t \to +\infty$ if and only if A is Hurwitz.*

In that case, we say that 0 is an *asymptotically stable equilibrium point* of the equation.

A matrix A is *hyperbolic* if it does not have an eigenvalue with zero real part. For such a matrix, $\Gamma^c = \{0\}$, *i.e.* $\mathbb{C}^n = \Gamma^s \oplus \Gamma^u$ and $\mathbb{R}^n = E^s \oplus E^u$ si $A \in M_n(\mathbb{R})$.

The set of hyperbolic matrices is important since it is stable by perturbation.

Théorème 2.28. *If A is hyperbolic, there exists $\delta > 0$ (dependent of A) such that for every matrix F verifying*

$$\|F\| \leq \delta,$$

the matrix $A + F$ is still hyperbolic. The set of hyperbolic matrices with complex coefficients (resp. real coefficient) is therefore open in $M_n(\mathbb{C})$ (resp. $M_n(\mathbb{R})$).
Moreover, the stable and instable spaces depend continuously of F (for $\|F\| \leq \delta$).

*Preuve.

▷ The fact that if F is small enough then $A + F$ is hyperbolic results of the continuity of the eigenvalues of the matrices : eigenvalues of $A + F$ is close enough to those of A as F small ; yet these lie in the open set $\mathbb{C} - \{\Re e 0\}$. It is the same for those of $A + F$ if F is small.

□

Remarque. While the dependence of the characteristic spaces of $A + F$ is *not* generally continuous with respect to F, that of the stable and unstable spaces is.

Case of a diagonalizable matrix.

Consider the particular case of a matrix $A \in M_n(\mathbb{R})$ diagonalizable in $\mathbb{C}$ (we also say *semi-simple*). As we saw in section 2.2.3, this means that A satisfies the following conditions, which are equivalent between them (we use the notations of section 2.2.3) :

— there exists a basis of $\mathbb{C}^n$ formed by eigenvectors of A ;
— $\mathbb{C}^n = \Pi_1 \oplus \cdots \oplus \Pi_r$, where $\Pi_i = \ker_{\mathbb{C}}(A - \lambda_i I) \subset \mathbb{C}^n$ is the eigenspace associated to λ_i and $\lambda_1, \ldots, \lambda_r$ are the (complex) eigenvalues of A ;
— for every eigenvalue λ_i, the characteristic subspace is equal to the eigenspace : $\Gamma_i = \Pi_i$;
— for every eigenvalue, the algebraic and geometric multiplicities coincide : $\dim \Pi_i = p_i$ pour $i = 1, \ldots, r$;
— in the complex reduced form of A, all the Jordan blocks are scalar matrices 1×1 : $J_{i,k} = (\lambda_i)$.

This case is in practice very important, since the set of matrices of $M_n(\mathbb{R})$ diagonalizable in $\mathbb{C}$ contains an open and dense set in $M_n(\mathbb{R})$. In other words, being diagonalizable in $\mathbb{C}$ is a generic property on $M_n(\mathbb{R})$.

Consider therefore such a matrix A. The general form of the solutions of $x' = Ax$ given by theorem 2.25 is simple : all the polynomial terms in t disappear, only remain the exponential and trigonometric terms.

Corollaire 2.29. *Let $A \in M_n(\mathbb{R})$ be a diagonalizable matrix in $\mathbb{C}$. Every solution of $x'(t) = Ax(t)$ in $\mathbb{R}^n$ is written as*

$$x(t) = \sum_{1 \le j \le q} e^{t\alpha_j} \big(\cos(\beta_j t)\, a_j + \sin(\beta_j t)\, b_j \big),$$

where $\alpha_j = \Re e(\lambda_j)$, $\beta_j = \Im m(\lambda_j)$ and the vectors a_j and $b_j \in \mathbb{R}^n$ verify $a_j + ib_j$ is an eigenvector of A associated to λ_j (i.e. $a_j + ib_j \in \Pi_j$).

The stable, instable and indifferent spaces are now given in terms of the eigenspaces (since the latter are equal to the characteristic subspaces) :

$$E^s = \Big[\bigoplus_{\Re e(\lambda_i)<0} \Pi_i \Big] \cap \mathbb{R}^n, \qquad E^u = \Big[\bigoplus_{\Re e(\lambda_i)>0} \Pi_i \Big] \cap \mathbb{R}^n, \qquad E^c = \Big[\bigoplus_{\Re e(\lambda_i)=0} \Pi_i \Big] \cap \mathbb{R}^n.$$

The dynamic characterization of these spaces results from theorem 2.26.

Corollaire 2.30. *Let $A \in M_n(\mathbb{R})$ be a diagonalizable matrix in $\mathbb{C}$. Denote by $x(\cdot)$ the real solutions of the differential equation $x'(t) = Ax(t)$. Then*

- *E^s is the set of initial conditions $x(0) \in \mathbb{R}^n$ corresponding to solutions $x(t)$ tending exponentially to 0 as $t \to +\infty$;*
- *E^u is the set of initial conditions $x(0) \in \mathbb{R}^n$ corresponding to solutions $x(t)$ tending exponentially to 0 as $t \to -\infty$;*
- *E^c is is the set of initial conditions $x(0) \in \mathbb{R}^n$ corresponding to periodic solutions $x(t)$, therefore bounded.*

The only difference with respect to the general case concerns the space E^c : while in the general case the asymptotic behavior of the solutions in E^c was indefinite (thus the name of "indifferent" space), we know here that all the solutions in E^c are bounded.

2.3 Stability

2.3.1 Equilibrium points and stability

Consider the autonomous differential equation

$$x'(t) = f\big(x(t)\big), \tag{2.20}$$

where the vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ is of class C^1.

Définition 2.12. We say that $x_0 \in \Omega$ is an *equilibrium point* of (2.20) if the constant function $x(\cdot) \equiv x_0$ is solution of (2.20) or, equivalently, if $f(x_0) = 0$ (verify it is equivalent !).

In other words, $\phi_t(x_0) = x_0$ for every $t \in \mathbb{R}$, where ϕ is the flow of the vector field f (the maximal interval associated to x_0 is $I_{x_0} = \mathbb{R}$). The orbit of x_0 is therefore reduced to a point : $\mathcal{O}_{x_0} = \{x_0\}$.

When the equation (2.20) models the evolution of a physical phenomenon (mechanical, biological, ecological, ...), an equilibrium point corresponds to the usual notion of equilibrium state : if the system is in the state x_0, then it stays there (and it was there always). In practice we know however that only states of equilibrium having certain properties of stability are significant.

Définition 2.13. We say that an equilibrium point x_0 is *stable* if, for every $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|x - x_0\| < \delta \quad \text{et} \quad t > 0 \qquad \Longrightarrow \qquad \|\phi_t(x) - x_0\| < \epsilon.$$

Then, every solution close to x_0 stays close.

Remarque. Any solution whose initial condition is in a ball $B(x_0, \delta)$ stays in the ball $B(x_0, \epsilon)$, and therefore in a compact of Ω, for $t > 0$ (we suppose ϵ small enough so that $\bar{B}(x_0, \epsilon) \subset \Omega$). According to proposition 2.4, these solutions are therefore defined for every $t > 0$.

Définition 2.14. We say that an equilibrium point x_0 is locally *asymptotically stable* (LAS) if it is stable and if there exists a neighborhood V of x_0 such that, for every $x \in V$,

$$\lim_{t \to \infty} \phi_t(x) = x_0.$$

If V is equal to the whole state space, we say that x_0 is globally *asymptotically stable* (GAS).

In the case (LAS), every solution close to the equilibrium point remains close to it and in addition converges to it.

Linear case

Consider the autonomous linear differential equation

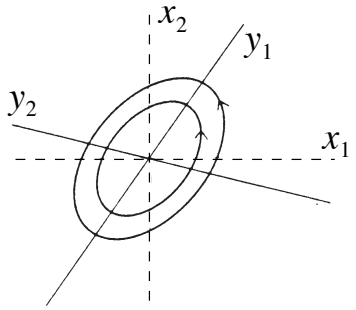
$$x'(t) = Ax(t), \qquad x \in \mathbb{R}^n.$$

The origin is always an equilibrium point of this equation (but there can be others : any element of $\ker A$ is an equilibrium point). According to section 2.2.4, we can characterize the stability of this equilibrium point. Owed to the linearity of the system (its homogeneity

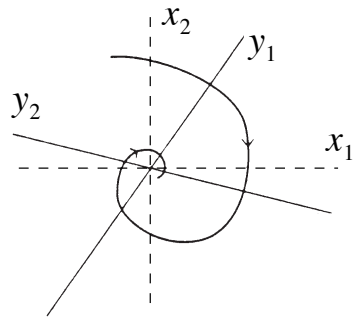
is enough), there is of no distinction between local or global. As a consequence, in the following statements, the stability, when it occurs, is global.

Proposition 2.31.
— *The origin is an asymptotically stable equilibrium point of $x' = Ax$ if and only if all the eigenvalues of A have negative real part i.e. $\mathbb{R}^n = E^s$.*
— *If A has at least one eigenvalue with positive real part, then the origin is not a stable equilibrium point of $x' = Ax$.*

Let us note that the origin can be a stable but not asymptotically stable equilibrium point. It occurs if A has eigenvalues with zero real part, for example when A is skew-symmetric (see proposition 2.13 and the discussion which follows). We represented in figure **??** the phase portraits in $\mathbb{R}^2$ corresponding to a skewsymmetric matrix (case A) and the other where eigenvalues has negative real part (case b).



Case a. stable but not asymptotically stable equilibrium point



Case b. asymptotically stable equilibrium point

FIGURE 2.2 – Examples of stable phase portrait for $f(x) = Ax$ with $A \in M_2(\mathbb{R})$.

Note that we can provide a necessary and sufficient condition for stability :
the origin is a stable equilibrium point of $x' = Ax$ if and only if all eigenvalues of A have non positive real part and for every eigenvalue with zero real part, the algebraic and geometric multiplicities coincide, (i.e. $\mathbb{R}^n = E^s + E^c$ and $A|_{E^c}$ is diagonalizable in $\mathbb{C}$).

Affine case

Consider now an affine vector field $f(x) = Ax + b$ on $\mathbb{R}^n$, where $A \in M_n(\mathbb{R})$ is a matrix and $b \in \mathbb{R}^n$ vector. An equilibrium point of the equation

$$x'(t) = Ax(t) + b$$

is a point x_0 verifying$Ax_0 + b = 0$ (it exists only if $b \in \operatorname{Im} A$). Replacing b by $-Ax_0$, the differential equation is written as

$$\frac{d}{dt}(x(t) - x_0) = A(x(t) - x_0).$$

Hence, stability issues of an equilibrium point of the affine equation $x'(t) = Ax(t) + b$ are equivalent to that of the origin of the linear equation $y'(t) = Ay(t)$.

2.3.2 Stability by linearization

Let x_0 be an equilibrium of the differential equation (2.20). We are going to show in the following two theorems that the study of the eigenvalues of the matrix $Df(x_0)$ often allows one to characterize the stability of the equilibrium point.

Théorème 2.32. *If all the eigenvalues of $Df(x_0)$ have negative real part, then x_0 is an asymptotically stable equilibrium point.*

Remarque. Contrary to the case of linear equation, the condition in the theorem is sufficient but not necessary. Let us take for example the equation $y'(t) = -y^3(t)$ in $\mathbb{R}$. The equilibrium point 0 does not satisfy the condition of the theorem since $Df(0) = 0$. On the other hand, it is an asymptotically stable equilibrium point since the solution equal to $y_0 \neq 0$ at $t = 0$ is

$$y(t) = \frac{\operatorname{signe}(y_0)}{\sqrt{2t + \frac{1}{y_0^2}}}, \qquad t \geq 0,$$

which is decreasing and converges to 0 as $t \to +\infty$.

*Preuve.

▷ To get matters simpler, we assume up to a translation that $x_0 = 0$. Then, there exists $\alpha > 0$ such that $-\alpha$ is strictly larger than the real part of every eigenvalue of $Df(0)$. By a classical result of linear algebra, there exists a scalar product $\langle \cdot, \cdot \rangle_\alpha$ on $\mathbb{R}^n$ such that

$$\langle Df(0)x, x\rangle_\alpha \leq -\alpha \|x\|_\alpha^2, \qquad \forall x \in \mathbb{R}^n,$$

where $\| \cdot \|_\alpha$ is the norm associated to the scalar product $\langle \cdot, \cdot \rangle_\alpha$.
Moreover, one has by definition of the differential,

$$\langle f(x), x\rangle_\alpha = \langle Df(0)x, x\rangle_\alpha + o(\|x\|_\alpha^2).$$

Hence, for x close enough to 0, let say $\|x\| < \delta$, one gets

$$\langle f(x), x\rangle_\alpha \leq -\frac{\alpha}{2}\|x\|_\alpha^2.$$

▷ Take now a point $v \neq 0$ verifying $\|v\| < \delta$ and denote $x(t) = \phi_t(v)$ the solution starting at v. We want to choose a time $t_0 > 0$ small enough so $\|x(t)\| < \delta$ for every $t \in [0, t_0]$. The function $t \mapsto \|x(t)\|_\alpha$ is derivable (since $x(t) \neq 0\ \forall t$), and

$$\frac{d}{dt}\|x(t)\|_\alpha = \frac{\langle x'(t), x(t)\rangle_\alpha}{\|x(t)\|_\alpha} = \frac{\langle f(x(t)), x(t)\rangle_\alpha}{\|x(t)\|_\alpha} \leq -\frac{\alpha}{2}\|x(t)\|_\alpha.$$

It implies that $\|x(t)\|_\alpha$ is decreasing : $x(t)$ stays in the compact $\|x\|_\alpha \leq \|v\|_\alpha$, implying that $x(\cdot)$ is defined for every $t > 0$ (proposition 2.4). Moreover, by Gronwall's lemma,

$$\|x(t)\|_\alpha \leq e^{-\frac{\alpha}{2}t}\|v\|_\alpha.$$

Finally, we showed that, if $\|v\| < \delta$, then $\phi_t(v)$ stays in $B(0, \delta)$ and tends to 0, *i.e.* 0 is a asymptotically stable equilibrium point.

□

Théorème 2.33. *If x_0 is a stable equilibrium point, then the eigenvalues of $Df(x_0)$ have non positive real part.*

We shall use generally the contraposition of this theorem : *if $Df(x_0)$ has at least one eigenvalue with strictly positive real part, then the equilibrium point x_0 not is not stable.*

Note that the converses of the theorems 2.32 and 2.33 are false, as showed in the example below. The stability of an equilibrium point is therefore not necessarily determined by the linearized. We are then going to introduce a class of equilibrium points for which converses of the theorems 2.32 and 2.33 are verified.

Exemple. Consider the two differential equations in $\mathbb{R}^2$,

$$x' = f(x) = \begin{pmatrix} x_2 - x_1(x_1^2 + x_2^2) \\ -x_1 - x_2(x_1^2 + x_2^2) \end{pmatrix} \qquad \text{et} \qquad x' = g(x) = \begin{pmatrix} x_2 + x_1(x_1^2 + x_2^2) \\ -x_1 + x_2(x_1^2 + x_2^2) \end{pmatrix},$$

where $x = (x_1, x_2)$. These two equations have a unique equilibrium point 0. Their linearizes at 0 are equal,

$$Df(0) = Dg(0) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

with eigenvalues $\pm i$. However, the equilibrium point 0 is asymptotically stable in the first case and not stable in the second.

Indeed, set $\rho(x) = x_1^2 + x_2^2$. Si $x(\cdot)$ is a solution of the equation $x' = f(x)$, then

$$\frac{d}{dt}\rho\big(x(t)\big) = 2(x_1 x_1' + x_2 x_2') = -2\rho^2\big(x(t)\big).$$

Hence $\rho\big(x(t)\big) = \|x(t)\|^2$ is decreasing and tends to 0 as $t \to +\infty$, implying that 0 is asymptotically stable for the equation $x' = f(x)$.

Similarly, if $x(\cdot)$ is a solution of the equation $x' = g(x)$, one gets

$$\frac{d}{dt}\rho\big(x(t)\big) = 2\rho^2\big(x(t)\big).$$

In that case, $\rho\big(x(t)\big) = \|x(t)\|^2$ tends to infinity in finite time (blow-up phenomenon), implying that the equilibrium point 0 is not stable for the equation $x' = g(x)$.

Hyperbolic equilibrium points

Définition 2.15. An equilibrium point x_0 is said to be *hyperbolic* if all the eigenvalues of $Df(x_0)$ have nonzero real part.

The hyperbolic equilibrium points are important in practice since, as seen in section 2.2.4, the class of the hyperbolic matrices is open and dense in $M_n(\mathbb{R})$.

By the two preceding theorems, the stability of an hyperbolic equilibrium point x_0 is completely characterized by the sign of the real parts of the eigenvalues of $Df(x_0)$.

Corollaire 2.34. *An hyperbolic equilibrium point is either asymptotically stable (if the eigenvalues of $Df(x_0)$ have negative real part), or stable.*

Hence, a hyperbolic equilibrium point x_0 is stable (resp. asymptotically stable) if and only if 0 is a stable equilibrium point stable (resp. asymptotically stable) for the linearized equation at x_0,

$$y'(t) = Df(x_0) \cdot y(t). \tag{2.21}$$

We have more : the phase portraits of the system and that of its linearized have same shape since they are topologically equivalent.

Théorème 2.35 (Hartman-Grobmann's theorem)**.** *Let x_0 be a hyperbolic equilibrium point. Denote $\phi_t^L : y \mapsto e^{tDf(x_0)}y$ the flow of the linearized at x_0. then there exists a homeomorphism $h : V_{x_0} \to V_0$, where V_{x_0} and V_0 are neighborhoods of x_0 and 0 in $\mathbb{R}^n$, such that*

$$\phi_t^L\big(h(x)\big) = h\big(\phi_t(x)\big),$$

everywhere these expressions have a meaning.

2.3.3 Lyapunov functions

There exists another approach than that of linearization to obtain stability results **which do not require explicit knowledge of the flow**. Let us start with an illustrating example.

Gradient fields

A *gradient fields* is a vector field of the type

$$f(x) = -\nabla V(x),$$

where $V : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is a function of class C^2 (so that f is of class C^1). Recall that $\nabla V(x)$ denotes the *gradient of V at x, i.e.* the unique vector of $\mathbb{R}^n$ verifying

$$DV(x) \cdot v = \langle \nabla V(x), v \rangle, \qquad \forall v \in \mathbb{R}^n.$$

in $\mathbb{R}^n$, in coordinates, $\nabla V(x) = (\frac{\partial V}{\partial x_1}(x), \dots, \frac{\partial V}{\partial x_n}(x))$.

An equilibrium point x_0 of this vector field is a critical point of V, *i.e.* $\nabla V(x_0) = 0$. An equilibrium point can therefore be a local minimum, a local maximum, or a saddle point, but we will see that only the local minima can be stable equilibrium points.

The associated dynamics to a gradient field has a property which makes it rather simple. If $x(\cdot)$ is a solution of the differential equation $x'(t) = -\nabla V\big(x(t)\big)$, then

$$\frac{d}{dt}\Big[V\big(x(t)\big)\Big] = DV\big(x(t)\big) \cdot x'(t) = -\|\nabla V\big(x(t)\big)\|^2, \tag{2.22}$$

for every t in the interval of definition of $x(\cdot)$. So $V\big(x(t)\big)$ is constant, and in this case, $x(t) \equiv x_0$ is a critical point, or strictly decreasing. Intuitively, it means that every solution tends to be close to a minimum and therefore (we will show that) :

— if an equilibrium point x_0 is not a local minimum (*i.e.* x_0 is a local maximum or a saddle point), then x_0 is not a stable equilibrium point ;

— if x_0 is un strict local minimum, then x_0 is a stable equilibrium point.

Lyapunov functions

Consider now an autonomous differential equation

$$x'(t) = f\big(x(t)\big), \tag{2.20}$$

associated to a vector field $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$ of class C^1. The example of gradient field suggests to introduce the following definition.

Définition 2.16. Let x_0 be an equilibrium point of (2.20), $U \subset \Omega$ a neighborhood of x_0 and $L : U \to \mathbb{R}$ a continuous function. We say that L is a *local Lyapunov function for (2.20) at* x_0 if :

(a) $L(x_0) = 0$ and $L(x) > 0$ for $x \neq x_0$ (*i.e.* x_0 is a strict minimum of L on U) ;

(b) the function $t \mapsto L\big(\phi_t(x)\big)$ is decreasing.

If moreover L satisfies

(c) for $x \neq x_0$, the function $t \mapsto L\big(\phi_t(x)\big)$ is strictly decreasing,

we say that L is a *strict* local Lyapunov function for (2.20) at x_0.

Finally, if U is equal to the whole state space, one can replace in what precedes "local" by "global".

If L is of class C^1, one can replace the conditions (b) and (c) respectively by the following conditions, which are more restrictive but simpler to verify than (b) and (c) :

(b)′ $\langle \nabla L(x), f(x) \rangle \leq 0$ for every $x \in U$;

(c)′ $\langle \nabla L(x), f(x) \rangle < 0$ for every $x \in U$, $x \neq x_0$.

A Lyapunov function is therefore a kind of energy function which decreases along trajectories.

Théorème 2.36. *Assume that the differential equation (2.20) admits L as local Lyapunov function at the equilibrium point x_0. Then x_0 is a stable equilibrium point.*
— *(Loc) If moreover L is strict, then x_0 is locally asymptotically stable.*
— *(Glob) If moreover L is global, strict and L tends to infinity as x tends to infinity, then x_0 is globally asymptotically stable.*

*Preuve.

▷ Let $L : U \to \mathbb{R}$ be the Lyapunov function. We can always assume that U is a closed ball centered at x_0. For every $\varepsilon > 0$, the set $U_\alpha = \{x \in U \ : \ L(x) < \alpha\}$ is included in $B(x_0, \varepsilon)$ for $\alpha > 0$ small enough (otherwise there exists a sequence of points x_n outside of $B(x_0, \varepsilon)$ verifying $\lim L(x_n) = 0$; U being compact, x_n has then an accumulation point $\bar{x} \neq x_0$, so that $L(\bar{x}) = 0$, which is impossible by (a)).
According to (b), for every $x \in U_\alpha$, the solution $\phi_t(x)$ stays in U_α ; showing the stability of the equilibrium point x_0.

▷ Let us suppose now that L is a strict local Lyapunov function. Consider a point $x \in U_\alpha$ different of x_0. The function $t \mapsto L\big(\phi_t(x)\big)$ being strictly decreasing and bounded from below by 0, it has a limit ℓ when $t \to +\infty$. Moreover, U being compact, there is a sequence t_n, tending to ∞, such that $\phi_{t_n}(x)$ is convergent to $\bar{x}$. By continuity of L, $\bar{x}$ verifies :

$$L(\bar{x}) = \lim_{t_n \to \infty} L\big(\phi_{t_n}(x)\big) = \lim_{t \to \infty} L\big(\phi_t(x)\big) = \ell.$$

Moreover, for every $s > 0$, one has

$$L\big(\phi_s(\bar{x})\big) = \lim_{t_n \to \infty} L\big(\phi_{s+t_n}(x)\big) = \ell,$$

showing that $s \mapsto L\big(\phi_s(\bar{x})\big) \equiv L(\bar{x})$ is not decreasing, and therefore, by (c), $\bar{x} = x_0$.
Hence the only accumulation point of $\phi_t(x)$ is x_0, showing that $\lim_{t\to\infty} \phi_t(x) = x_0$. The equilibrium point x_0 is therefore asymptotically stable.
We can adapt this argument to the case (Glob), noticing that the sets $L_c := \{x \in \Omega \ \ L(x) \leq c\}$, $c > 0$, are compact.

□

Remarque. For an asymptotically stable equilibrium point x_0, we call *basin of attraction* the set of points $x \in \Omega$ such as $\phi_t(x) \to x_0$ when $t \to +\infty$. By definition of asymptotic stability, the basin of attraction contains a neighborhood of x_0. An important question in practice is to determine the size of this basin, even the basin itself.

The domain of definition of a strict Lyapunov function, if there is one, gives some hints to answer that question. Let us suppose for example that $\Omega\mathbb{R}^n$ and that L is a Lyapunov function verifying the hypotheses of case (Glob). Then, the basin of attraction of x_0 is $\mathbb{R}^n$.

More generally, if $L : U \to \mathbb{R}$ is a strict Lyapunov function at x_0 and $P \subset U$ a closed subset of $\mathbb{R}^n$ positively invariant by the flow (*i.e.* $\phi_t(P) \subset P$ for every $t \geq 0$), then P is

included in the basin of attraction of x_0. For example, the sets $\{x \in U \;\; L(x) \leq \alpha\}$ for α small enough are closed and positively invariant (cf. definition 2.6), therefore included in the basin of attraction.

In numerous applications, it is not possible to have a strict Lyapunov function *i.e.* verifying the condition $(c)'$. We have then the following result, called *Lasalle invariance principle*.

Théorème 2.37. *Assume the differential equation (2.20) admits $L : U \to \mathbb{R}_+$ as local Lyapunov function at an equilibrium point x_0. Let D_U be the subset of U defined by*

$$D_U := \{x \in U, \quad DL(x).f(x) = 0\}.$$

then,

- *(Loc) all the trajectories staying in U converge asymptotically to the largest invariant set (cf. definition 2.6) contained in D_U.*
- *(Glob) If moreover L is global (i.e. $U = \Omega$) and tends to infinity as x tends to infinity, then all trajectories are defined on $\mathbb{R}_+$ and converge asymptotically to the largest invariant set contained in D_U.*

The invariance principe simply consists in writing the overdetermined system

$$x' = f(x), \quad D_x L(x) = 0, \quad x \in U,$$

which characterizes the largest invariant set contained in the intersection of U and D_U.

Examples

a. Gradient field Let $f(x) = -\nabla V(x)$ be a gradient field. Assume x_0 is a strict local minimum of V, *i.e.* the unique minimum of V in a neighborhood U of x_0. The function V restricted to U is a Lyapunov function at x_0, property $(b)'$ is always satisfied by formula (2.22). Therefore x_0 is a stable equilibrium point .

b. Conservative vector field Consider an objet of mass m subject to a force deriving from a potential $V(x)$. The time evolution of the state $x \in \mathbb{R}^n$ of the objet is given by the fundamental principle of dynamics :

$$mx''(t) = -\nabla V\big(x(t)\big),$$

which is written

$$\begin{pmatrix} x \\ x' \end{pmatrix}'(t) = \begin{pmatrix} x'(t) \\ -\frac{1}{m}\nabla V\big(x(t)\big) \end{pmatrix}.$$

In other words, (x, x') is solution of the first order differential equation in $\mathbb{R}^{2n}$

$$(x, v)'(t) = f\big((x, v)(t)\big), \qquad \text{where } f(x, v) = \big(v, -\frac{1}{m}\nabla V(x)\big).$$

An equilibrium point of this differential equation is a point $(x_0, 0) \in \mathbb{R}^{2n}$ where x_0 is a critical point of the potential $V(x)$.

Assume x_0 is a strict local minimum of V and let us look for a Lyapunov function. We first try the total energy of the system :

$$E(x, v) = \frac{1}{2}m\|v\|^2 + V(x).$$

Setting $L(x, v) = E(x, v) - V(x_0)$, one gets a function satisfying property (a) of a Lyapunov function. Moreover, since $\nabla L(x, v) = \big(\nabla V(x), mv\big)$, one gets

$$\langle \nabla L(x, v), f(x, v)\rangle \equiv 0,$$

(conservation of energy !), *i.e.* property $(b)'$. The function L is therefore a Lyapunov function at $(x_0, 0)$, showing the well-known result (Lagrange's theorem) :
if the potential energy $V(x)$ has a strict local minimum at x_0, then the equilibrium point $(x_0, 0)$ is stable.

Remarques.

— The equilibrium $(x_0, 0)$ cannot be asymptotically stable : the Lyapunov function $L(x, v)$ being constant along a solution $(x, v)(\cdot) \not\equiv (x_0, 0)$, it cannot converge to 0, which implies that $(x, v)(t)$ cannot converge to $(x_0, 0)$.

— The approach by linearization would not have allowed to conclude $(x_0, 0)$ is not a hyperbolic equilibrium point (show it).

c. Linear vector field Consider a linear differential equation

$$x'(t) = Ax(t), \qquad x \in \mathbb{R}^n,$$

and let us suppose that all the eigenvalues of A have strictly negative real part. We saw (proposition 2.31) that in this case the origin is a (globally) asymptotically stable equilibrium point. Let us look for a Lyapunov function for this equation at 0. The following theorem is going to supply us with it.

Théorème 2.38. *The following propositions are equivalent.*

— *(i) A is Hurwitz (i.e. all its eigenvalues have strictly negative real part) ;*

— *(ii) there exists a matrix $P \in M_n(\mathbb{R})$ such that $P = P^T > 0$ (i.e. P is real symmetric and positive definite) and*

$$A^T P + PA < 0;$$

— *(iii) for every matrix $Q = Q^T > 0$, there exists a unique matrix $P = P^T > 0$ solution of the* **Lyapunov equation**

$$A^T P + PA = -Q.$$

Preuve.

▷ It is enough to prove the implications $(ii) \Rightarrow (i)$ and $(i) \Rightarrow (iii)$.

We start by $(ii) \Rightarrow (i)$. Let $P = P^T > 0$ such that $A^T P + PA = -Q$. Denote $\alpha = \frac{\lambda_m(Q)}{\lambda_M(P)} > 0$ with $\lambda_m(Q)$ and $\lambda_M(P)$ respectively the smallest of the eigenvalues of Q and the largest of the eigenvalues of P. Consider the Lyapunov function $L(x) := x^T P x$. Then,

$$\dot{L}(x) = 2x^T APx = x^T(A^T P + PA)x = -x^T Qx \leq -\lambda_m(Q)x^T x \leq -\beta L(x).$$

One concludes that $L(x(t)) \leq e^{-\beta t} L(x(0))$ (why ?) and therefore (i).

Let us show now $(i) \Rightarrow (iii)$. Assume A is Hurwitz. For $t \geq 0$, set

$$P(t) := \int_0^t e^{sA^T} Q e^{sA} ds.$$

Notice that $P(t) > 0$ for $t > 0$ and $t \mapsto P(t)$ is a strictly increasing function. As A is Hurwitz, the generalized integral

$$P = \int_0^\infty e^{sA^T} Q e^{sA} ds,$$

is convergent and one concludes that $P(\cdot)$ admits P as limit when t tends to infinity. Moreover, $P(\cdot)$ verifies the following ODE

$$\dot{P}(t) = -(A^T P(t) + P(t)A + Q).$$

necessarily an equilibrium point of this ODE (why ?) and therefore satisfies the Lyapunov equation. For the uniqueness, one proceeds as follows. Let $M > 0$ be a solution of the Lyapunov equation. We multiply then the Lyapunov equation verified by M to the left by e^{sA^T} and to the right by e^{sA}. We notice that

$$e^{sA^T} A^T M e^{sA} + e^{sA^T} MA e^{sA} = \frac{d}{ds}(e^{sA^T} M e^{sA}).$$

Integrating between 0 and infinity, one gets $M = P$.

□

d. Hurwitz criterion Although we cannot express generally the roots of a polynomial of degree $n \geq 5$ as a function of its coefficients, there are algebraic conditions **only dependent** on the coefficients of a polynomial P which are necessary and sufficient conditions so that any root of P has strictly negative real part (by abuse of language, such a polynomial is said to be Hurwitz). One of these conditions is *the Hurwitz criterion*. We consider the complex polynomial of degree n :

$$P(z) = a_0 z^n + a_1 z^{n-1} + \cdots + a_{n-1} z + a_n, \qquad (a_0 \neq 0).$$

Set $a_{n+1} = a_{n+2} = \cdots = a_{2n-1} = 0$. Define the square matrix of order n :

$$H = \begin{pmatrix} a_1 & a_3 & a_5 & \cdots & \cdots & a_{2n-1} \\ a_0 & a_2 & a_4 & \cdots & \cdots & a_{2n-2} \\ 0 & a_1 & a_3 & \cdots & \cdots & a_{2n-3} \\ 0 & a_0 & a_2 & \cdots & \cdots & a_{2n-4} \\ 0 & 0 & a_1 & \cdots & \cdots & a_{2n-5} \\ \vdots & \vdots & \ddots & & & \vdots \\ 0 & 0 & 0 & * & \cdots & a_n \end{pmatrix},$$

où $* = a_0$ ou a_1 according to the parity of n.

Let $(H_i)_{i\in\{1,\ldots,n\}}$ be the principal minors of H, i.e.

$$H_1 = a_1,\ H_2 = \begin{vmatrix} a_1 & a_3 \\ a_0 & a_2 \end{vmatrix},\ H_3 = \begin{vmatrix} a_1 & a_3 & a_5 \\ a_0 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{vmatrix},\ \ldots, H_n = \det H.$$

Proposition 2.39. *If $a_0 > 0$, every root of P has strictly negative real part if and only if $H_i > 0$, for every $i \in \{1, \ldots, n\}$.*

Remarque. If $a_0 > 0$, one has :

— If for every racine λ of P, Re $\lambda \leq 0$, then $a_k \geq 0$ and $H_k \geq 0$, for every $k \in \{1, \ldots, n\}$.

— If $n \leq 3$ and if $a_k \geq 0$ and $H_k \geq 0$, for every $k \in \{1, 2, 3\}$, then every root λ of P verifies Re $\lambda \leq 0$.

Remarque. A *necessary condition* for stability is therefore, if $a_0 > 0$:

$$\forall k \in \{1, \ldots, n\} \quad a_k \geq 0.$$

But is is not sufficient (give an example).

e. New proof of theorem 2.32 Consider an equilibrium point x_0 of the differential equation

$$x'(t) = f\big(x(t)\big),$$

et assume all the eigenvalues of $Df(x_0)$ have strictly negative real part. We newt show that this ODE admits a strict Lyapunov function, thus giving a new proof of theorem 2.32.

Up to a translation, suppose $x_0 = 0$. Take $P = P^T > 0$ solution of the Lyapunov equation

$$Df(0)^T P + PDf(0) = -I_n.$$

Consider $L(x) := x^T Px$ which is a Lyapunov function for the linear equation $y'(t) = Df(0) \cdot y(t))$ (verify it). Notice that

$$f(x) = Df(0) \cdot x + o(\|x\|).$$

Using the previous paragraph, one gets

$$\begin{aligned} \langle \nabla L(x), f(x) \rangle &= \langle \nabla L(x), Df(0) \cdot x \rangle + \langle \nabla L(x), o(\|x\|) \rangle \\ &= -\|x\|^2 + 2 \int_0^\infty \langle e^{sDf(0)} x, e^{sDf(0)} o(\|x\|) \rangle \, ds. \end{aligned}$$

The term in the last integral is a $o(\|x\|^2)$, therefore for $\|x\|$ small enough, $x \neq 0$, one gets

$$\langle \nabla L(x), f(x) \rangle \leq -\frac{1}{2}\|x\|^2 < 0,$$

i.e. property $(c)'$. The function L is therefore a strict Lyapunov function at $x_0 = 0$, showing that this equilibrium point is asymptotically stable.

Chapitre 3

Controllability and observability of linear systems

3.1 Control Systems

A large part of control theory is based on differential equations : this is the so-called state space representation of deterministic systems in continuous time (versus stochastic systems using stochastic differential equations). It goes as follows : consider a physical system (e.g. a satellite, a car,...), described by its state $x(t)$ at time t (e.g. position and speed), on which one can act a every time by means of a *control* u (e.g. engine push for a satellite). We represent the state by a vector of $\mathbb{R}^n$, the control by a vector of $\mathbb{R}^m$, and we model evolution of the vector $x(t)$ by a *control system* (or controlled differential equation)

$$(\Sigma) \ : \qquad x'(t) = f(t, x(t), u(t)), \qquad t \in [0, \tau],$$

where $\tau > 0$.

What is the meaning of the latter expression ? The function $u(t)$, $t \in [0, \tau]$, called *control law* is the mean of action on the system (Σ) : it will be chosen in terms of the goals to be achieved. To a control law $u(\cdot)$, is associated an ordinary differential equation

$$(\Sigma_u) \ : \qquad x'(t) = f_u(t, x(t)), \qquad t \in [0, \tau],$$

where $f_u(t, x) := f(t, x, u(t))$. Hence, a function $x(\cdot)$ is solution of System (Σ) if there exists a control law $u(\cdot)$ such that $x(\cdot)$ is solution of (Σ_u).

The main issues to address are the following.

Controllability given an initial state $x_0 \in \mathbb{R}^n$, a final state $v \in \mathbb{R}^n$ and a time $t = \tau > 0$, is it possible to find a control law $u(\cdot)$ steering System (Σ) initially in $x(0)$ at $t = 0$ to the state v at time $t = \tau$? Equivalently, is it possible to *control* System (Σ) from x_0 to v in time τ ?

Planification of trajectories To the above structural question, corresponds the more practical problem of determining an effective procedure which associates, to a pair

of states $x_0, v \in \mathbb{R}^n$ and a time τ, a control law $u(\cdot)$ steering the system from $x(0)$ to v in time $t = \tau$.

Stabilization Is it possible to build a control law $u(\cdot)$ which *asymptotically stabilizes* System (Σ) at an equilibrium point x_0, i.e., such that, for every initial condition $x(0)$, one has

$$\lim_{t \to +\infty} x(t) = x_0 ?$$

Observability In order to achieve a control goal (planification of trajectories, stabilization, etc...) and therefore to choose the appropriate control law, a certain amount of information on the state x of the system is available at every time t. It is usually obtained by measurement. However, it is not possible to measure in general (one says *to observe* in control theory) directly the full state $x(t)$ but only a function $y(t)$ of the state and the control

$$y(t) = g(x(t), u(t), t).$$

One must then "reconstruct" the state $x(\cdot)$ from the *output* $y(\cdot)$. The observability issue resumes therefore to the following : does the knowledge of $y(t)$ and $u(t)$ for every $t \in [0, \tau]$ allow one to determine the state $x(\cdot)$ for every $t \in [0, \tau]$ (or, let say the initial state $x(0)$) ?

The technics introduces for the study of linear autonomous ODEs will allow us to answer to these questions in the framework of *linear autonomous control theory*.

In all this chapter, we will therefore assume that System (Σ) is linear autonomous (w.r.t. (x, u)), i.e.,

$$(\Sigma) \ : \qquad x'(t) = Ax(t) + Bu(t), \qquad t \in [0, \tau], \tag{3.1}$$

where $A \in M_n(\mathbb{R})$ and $B \in M_{n,m}(\mathbb{R})$, with n, m positive integers. If $m = 1$, the system is said to be *single-input* and otherwise *multi-input*.

Moreover, when we will consider observability, we will assume that the output $y \in \mathbb{R}^p$, p positive integer and that y is also a linear function of (x, u), or more simply equal to

$$y(t) = Cx(t), \qquad t \in [0, \tau], \tag{3.2}$$

with $C \in M_{p,n}(\mathbb{R})$. In that case, the control system is defined by the two equations (3.1) et (3.2).

The control laws $u(\cdot)$ are assumed to be *piecewise continuous*, defined on the interval $[0, \tau]$ and taking values in $\mathbb{R}^m$. One could have made other hypotheses on $u(\cdot)$, for instance assuming it is only measurable or bounded (which is reasonabe if u stands for a force). These hypotheses are natural in control theory but we will not consider them in these notes : our primary objective consists of illustrating in a framework as simple as possible how the ODEs technics apply to the control theoretic issues raised previously.

The starting point of the study of control linear systems is the *the variation of the constant formula* :

Proposition 3.1. *Let $u(\cdot)$ be a control law and $x_0 \in \mathbb{R}^n$. The unique solution of $x'(t) = Ax(t) + Bu(t)$ equal to x_0 at time $t = 0$ is*

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)ds.$$

Preuve.

▷ Assume there exists a solution $x(\cdot)$ of $x'(t) = Ax(t) + Bu(t)$ such that $x(0) = x_0$. Set $y(t) = e^{-tA}x(t)$ and take the time derivative of $y(\cdot)$:

$$y'(t) = -Ae^{-tA}x(t) + e^{-tA}\big(Ax(t) + Bu(t)\big) = e^{-tA}Bu(t).$$

Integrating between 0 et t, one gets

$$y(t) = y(0) + \int_0^t e^{-sA}Bu(s)ds,$$

and therefore the conclusion, since $x(t) = e^{tA}y(t)$ and $y(0) = x(0) = x_0$.

□

Remarque. In particular notice that, if $x(0) = 0$,

$$x(t) = \int_0^t e^{(t-s)A}Bu(s)ds, \tag{3.3}$$

and that expression depends linearly on the control law $u(\cdot)$.

Remarque. As regards observability of linear systems, the formula of Proposition 3.1 shows that the reconstruction of a trajectory $x(\cdot)$ with the sole knowledge of an ouput $y(\cdot)$ (and of course of the control law $u(\cdot)$) consists in fact of determining the initial condition x_0.

3.2 Controllability

Let (Σ) be the linear autonomous control system (3.1). Given $x_0 \in \mathbb{R}^n$, a state $v \in \mathbb{R}^n$ is *reachable in time* τ from x_0 (by (Σ)) if there exists a control law $u : [0, \tau] \to \mathbb{R}^m$ such that $x(\tau) = v$, $x(\cdot)$ being the solution of (Σ_u) satisfying $x(0) = x_0$. Let $\mathcal{A}(\tau, x_0)$ be the set of reachable states from x_0 in time τ, i.e.,

$$\mathcal{A}(\tau, x_0) := \left\{ x(\tau) \ : \ \begin{array}{c} x(\cdot) \text{ solution of } (\Sigma) \\ \text{t.q. } x(0) = x_0 \end{array} \right\}.$$

As a byproduct of the above remark and Proposition 3.1, one gets that the set $\mathcal{A}(\tau, 0)$ is a vector space, and the set $\mathcal{A}(\tau, x_0)$ is the affine space $e^{\tau A}x_0 + \mathcal{A}(\tau, 0)$. The set of reachable points from x_0 is therefore completely characterized by the set $\mathcal{A}_\tau := \mathcal{A}(\tau, 0)$.

Définition 3.1. System (Σ) is said to be *controllable in time* τ if $\mathcal{A}_\tau = \mathbb{R}^n$, or equivalently, if every state of $\mathbb{R}^n$ is reachable in time τ from any other state.

We will now characterize algebrically controllability. We must first determine the set $\mathcal{A}_\tau$ for $\tau > 0$.

Théorème 3.2. *For $\tau > 0$, the space $\mathcal{A}_\tau$ is equal to the image of the matrix $(n \times nm)$*

$$\mathcal{C}(A, B) := \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix},$$

called controllability matrix.

Remarque. The image of $\mathcal{C}(A, B)$ is the vector space $\mathcal{R}(A, B) \subset \mathbb{R}^n$ generated by $A^i Bz$, $i \in \{0, \ldots, n-1\}$, $z \in \mathbb{R}^m$:

$$\mathcal{R}(A, B) = \text{Vect}\{A^i Bz \ : \ i = 0, \ldots, n-1, \ z \in \mathbb{R}^m\}.$$

The first consequence of this result is that $\mathcal{A}_\tau$ is independent of τ. Note that this would be true if we would have chosen bounded control laws.The second consequence is that the dimension of $\mathcal{A}_\tau$ is equal to the rank of the controllability matrix. One gets an algebraic criterion of controllability, and therefore useful (in general) to verify.

Corollaire 3.3 (Kalman criterion for controllability)**.** *System (Σ) is controllable if and only if the rank of the controllability matrix $\mathcal{C}(A, B)$ is equal to n.*

Preuve.

▷ du Theoreme 3.2 Let us first show that $\mathcal{A}_\tau \subset \mathcal{R}(A, B)$. Observe that, by definition, if v belongs to $\mathcal{A}_\tau$ there exists $u : [0, \tau] \to \mathbb{R}^m$ piecewise continuous such that

$$v = \int_0^\tau e^{(\tau - s)A} Bu(s) ds.$$

Cayley-Hamilton's theorem says that the characteristic polynomial P_A of A annihilates A. Since P_A is of degree n and unitary (i.e., its coefficient of highest degree is equal to 1). Hence A^n is a linear combinaison of $I, \ldots, A^{n-1}$ and thus for *every* integer $i \geq n$, A^i is a linear combinaison of $I, \ldots, A^{n-1}$. Therefore, for every $i \geq 0$, A^i leaves invariant the vector space

$$\mathcal{R}(A, B) = \text{Vect}\{A^i Bz \ : \ i = 0, \ldots, n-1, \ z \in \mathbb{R}^m\}.$$

For every $s \in [0,\tau]$, the exponential $e^{(\tau-s)A}$ admits the development

$$e^{(\tau-s)A} = I + (\tau-s)A + \cdots + \frac{(\tau-s)^k A^k}{k!} + \cdots,$$

and hence $e^{(\tau-s)A}$ leaves also invariant the space $\mathcal{R}(A,B)$. We therefore have shown that $e^{(\tau-s)A}Bu(s) \in \mathcal{R}(A,B)$ for every $s \in [0,\tau]$ and

$$\int_0^\tau e^{(\tau-s)A}Bu(s)ds \in \mathcal{R}(A,B).$$

Hence $\mathcal{A}_\tau \subset \mathcal{R}(A,B)$.

▷ Let us show the other inclusion. It is enough to prove that $\mathcal{A}_\tau^\perp \subset \mathcal{R}(A,B)^\perp$. Let therefore $w \in \mathbb{R}^n$ be orthogonal to $\mathcal{A}_\tau$; the vector w is hence orthogonal to the state $\tilde{w}$ which can be reached in τ by the control law

$$u(t) = B^T(e^{(\tau-s)A})^T w.$$

The formula (3.3) shows that

$$\tilde{w} = \int_0^\tau e^{(\tau-s)A}BB^T(e^{(\tau-s)A})^T w\, ds,$$

and thus, since $\langle \tilde{w}, w\rangle = 0$ one gets

$$0 = \langle w, \int_0^\tau e^{(\tau-s)A}BB^T(e^{(\tau-s)A})^T w\, ds\rangle = \int_0^\tau \Big((e^{(\tau-s)A}B)^T w\Big)^T \Big((e^{(\tau-s)A}B)^T w\Big) ds,$$

which is equivalent to

$$\forall s \in [0,\tau],\ (e^{(\tau-s)A}B)^T w = 0.$$

Take all the time derivatives with respect to time the above equality. Then,

$$(e^{(\tau-s)A}AB)^T w = 0, \quad (e^{(\tau-s)A}A^2B)^T w = 0, \quad \dots \quad (e^{(\tau-s)A}A^{n-1}B)^T w = 0,$$

i.e., for $s = \tau$,

$$B^T w = 0, \qquad \dots \qquad (A^{n-1}B)^T w = 0.$$

It implies that, for every $j \in \{0,\dots,n-1\}$ and every $z \in \mathbb{R}^m$,

$$0 = \langle z, (A^jB)^T w\rangle = \langle A^jBz, w\rangle,$$

i.e., $w \in \mathcal{R}(A,B)^\perp$. The inclusion $\mathcal{R}(A,B) \subset \mathcal{A}_\tau$ is therefore established.

□

We now describe what happens when the rank r of the controllability matrix $\mathcal{C}(A,B)$ is arbitrary. To see that, we study the effect of a linear change of variable on the controllability of a system.

Définition 3.2. The linear control systems $\dot{x}_1 = A_1x_1 + B_1u_1$ and $\dot{x}_2 = A_2x_2 + B_2u_2$ are said to be *linearly equivalent* if there exists $P \in GL_n(\mathbb{R})$ such that $A_2 = PA_1P^{-1}$ et $B_2 = PB_1$.

Remarque. One has then $x_2 = Px_1$.

Proposition 3.4. *The Kalman property is intrinsic, i.e.,*

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

and therefore the rank of the Kalman matrix is invariant by linear equivalence.

Consider a pair (A, B) where $A \in \mathcal{M}_n(\mathbb{R})$ and $B \in \mathcal{M}_{n,m}(\mathbb{R})$.

Théorème 3.5 (Kalman Decomposition)**.** *The pair (A, B) is linearly equivalent to a pair (A', B') of the form*

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

where $A'_1 \in \mathcal{M}_r(\mathbb{R})$, $B'_1 \in \mathcal{M}_{r,m}(\mathbb{R})$, r being the rank of the Kalman matrix of the pair (A, B). Moreover, the pair (A'_1, B'_1) is controllable.

Remarque. The preceding decomposition is also called decomposition of the system in controllable and uncontrollable parts. The eigenvalues of A are called *poles* of the system, controllable for the eigenvalues of A'_1 and non controllable for those of A'_3.

Preuve.

▷ Assume that the rank r of the Kalman matrix of the pair (A, B) is stricly less than n (otherwise there is nothing to prove). The subspace

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

is of dimension r, and according to Cayley-Hamilton's theorem, it is clearly invariant by A. Let G be a supplementary space of F in $\mathbb{R}^n$, and let $(f_1, \ldots, f_r)$ be a basis of F, and $(f_{r+1}, \ldots, f_n)$ a basis of G. Let P be the matrix taking $(f_1, \ldots, f_n)$ to the canonical basis of $\mathbb{R}^n$. Then, since F is invariant by A, one has :

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

and, since Im $B \subset F$,

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

Finally, it is easy to see that the rank of the Kalman matrix of the pair (A_1', B_1') is equal to that of the pair (A, B).

□

In the single-input case, there exists a particularly useful change of coordinates.

Théorème 3.6 (Companion form)**.** *If $m = 1$ and if the pair (A, B) is controllable, then it is linearly equivalent to the pair $(\tilde{A}, \tilde{B})$, where*

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

The matrix $\tilde{A}$ and the pair $(\tilde{A}, \tilde{B})$ are respectively called companion form of the matrix A and canonical controllable form of the pair (A, B).

Remarque. If a square matrix is under companion form, then the coefficients of its characteristic polynomial form the last line of the matrix.

Remarque. In these new coordinates, the system is equivalent to the scalar ODE of order n :

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

Preuve.

▷ Consider $P_A(X)$ the characteristic polynomial of A,

$$P_A(X) := X^n + a_1 X^{n_1} + \cdots + a_n.$$

The change of coordinates will be given by F, an invertible matrix $n \times n$ with column vectors $(f_1, \cdots, f_n)$. These vectors are recursively defined as follows,

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

The family $(f_1, \ldots, f_n)$ is a basis of $\mathbb{R}^n$ since :

$$\begin{gathered} \text{Vect } \{f_n\} = \text{Vect } \{b\}, \\ \text{Vect } \{f_n, f_{n-1}\} = \text{Vect } \{b, Ab\}, \\ \vdots \\ \text{Vect } \{f_n, \ldots, f_1\} = \text{Vect } \{b, \ldots, A^{n-1}b\} = \mathbb{R}^n. \end{gathered}$$

One must finally check that $Af_1 = -a_n f_n$:

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

since according to Cayley-Hamilton's theorem, one has $A^n = -a_1 A^{n-1} - \cdots - a_n I$. In the basis, $(f_1, \ldots, f_n)$, the pair (A, b) is given by $(\tilde{A}, \tilde{b})$.

□

Remarque. This theorem admits the following generalization when $m > 1$. If the pair (A, B) is controllable, then one can conjugate it to a pair $(\tilde{A}, \tilde{B})$ such that :

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

the matrices $\tilde{A}_i$ being matrices in companion form ; moreover, there exists a matrix $G \in \mathcal{M}_{m,s}(\mathbb{R})$ such that :

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

where all the coefficients of each matrix $\tilde{B}_i$ are zero, except that of the last line, at the ith column, which is equal to 1.

3.3 Planification of trajectories

We now propose a method for solving the problem of planification of trajectories for a linear autonomous control system which is controllable. To do so, we will make linear transformations on the original system, so as to bring it under an extremely simple form for which planification of trajectories will become an easy task.

3.3.1 Exemple

The one dimensional control system $\dot{x} = u$ is clearly controllable. Its trajectory planification goes as follows. Given two arbitrary real numbers x_0, x_1, one has to find a control

law u defined on $[0,1]$ such that

$$x_1 = x_0 + \int_0^1 u(t)dt,$$

and therefore more simply, to find for every $x \in \mathbb{R}$, a function u such that $x = \int_0^1 u(t)dt$. A polynomial function (constant !) will be enough. The generalization of the system $x^{(n)} = u$ with x, u real scalars, is immediate.

3.3.2 Brunovsky Form

In the general case $\dot{x} = Ax + Bu$, the idea consists of transforming linearly the system to make it look like as much as possible as $x^{(n)} = u$.

The linear transformations considered below are more general than simple linear changes of coordinates.

Définition 3.3 (Regular static feedback)**.** Let $x_1 = Mx_2$ be the change of coordinates in $\mathbb{R}^n$ defined by a $n \times n$ square matrix M. A regular static feedback is defined by $u_1 = Kx_2 + Nu_2$ with N an invertible $m \times m$ matrix and K a matrix $m \times n$. This is a change of variable on the control parameterized by the state. It is expressed with matrices as

$$\begin{pmatrix} x_2 \\ u_2 \end{pmatrix} := \begin{pmatrix} M & 0 \\ K & N \end{pmatrix} \begin{pmatrix} x_2 \\ u_2 \end{pmatrix}.$$

The linear control systems $\dot{x}_1 = A_1x_1 + B_1u_1$ and $\dot{x}_2 = A_2x_2 + B_2u_2$ are said to be *equivalent by state change of variables with regular static feedback* if there exists $M \in GL_n(\mathbb{R})$, $N \in GL_m(\mathbb{R})$ and $K \in M_{m,n}(\mathbb{R})$, such that $A_2 = M^{-1}A_1M + M^{-1}BK$ et $B_2 = M^{-1}B_1N$.

Théorème 3.7 (Brunovsky Form of a controllable system)**.** *Let $\dot{x} = Ax + Bu$ be a linear controllable control system with $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. Then there exists a state change of variables with regular static feedback gibven by $x = Mz$ and $u = Kz + Nv$ for which there exists m coordinates of z, denoted $y_1, \cdots, y_m$ such that*

(i) $z = \left(y_1, \cdots, y_1^{(\alpha_1-1)}, \cdots, y_m, \cdots, y_m^{(\alpha_m-1)}\right)^T$;

(ii) $\dot{x} = Ax + Bu$ *becomes*

$$y_1^{(\alpha_1)} = v_1, \quad \cdots \quad y_m^{(\alpha_m)} = v_m,$$

where $y_l^{(j)}$ is the jth time derivative of the scalar function y_l, $1 \leq l \leq m$.

The coordinates $y_1, \cdots, y_m$ are called **Brunovsky ouputs** *of the system.*

Let us provide the Brunovsky form of a single-input system, i.e., with a unique scalar control ($m = 1$). In that case, $B = b$ is a column vector. There exists therefore a state change change of variables with regular static feedback which transforms $\dot{x} = Ax + bu$ in $\dot{z}_i = z_{i+1}$ for $1 \leq i \leq n$ and $\dot{z}_n = v$. Th Brunovsky output is then $y = z_1$ and the control system is simply $y^{(n)} = v$.

3.3.3 Application to the planification of trajectories

Assume that the control system $\dot{x} = Ax + Bu$ is controllable and under Brunovsky form. Thanks to the bloc-diagonal structure, the problem of planification in time $T > 0$ decomposes into m problems with a single control : for $1 \leq i \leq m$,

$$\text{Go to } \left(y_i(0), \cdots, y_i^{(\alpha_i-1)}(0)\right)^T \text{ à } \left(y_i(T), \cdots, y_i^{(\alpha_i-1)}(T)\right)^T \text{ along } y_i^{(\alpha_i)} = v_i.$$

The i-th planification problem defined above is easily solved since the conditions on the initial and final points represent $2\alpha_i$ constraints to be satisfied. For instance, it is enough to choose v_i as a polynomial of degree $2\alpha_i - 1$ and to determine its coefficients.

3.3.4 Proof of Theorem 3.7 for the single-input case

In the rest of this paragraph, we assume that $m = 1$ and therefore $B = b$. Note $P_A(X) = X^n + a_1X^{n_1} + \cdots + a_n$, the characteristic polynomial of A and we assume that the invertible matrix F takes the pair (A, b) under companion form, i.e., $(F^{-1}AF, F^{-1}b) = (\tilde{A}, \tilde{B})$, cf. Theorem 3.6. If $Fy = x$, then $\dot{y} = F^{-1}AFy + F^{-1}b$ with

$$\dot{y}_1 = y_2, \ \cdots \ \dot{y}_{n-1} = y_n,$$

and $\dot{y}_n = -a_ny_1 - \cdots - a_1y_n + u$. If one defines the new control $v := y_n$, then one has a regular static feedback on u. The sysem is therefore under Brunovsky form with y_1 as Brunovsky output.

3.4 Stabilization

Using control laws $u(\cdot)$ in general time-dependent is called *controlling in open loop* : the control law is fixed at the starting time $t = 0$ and is implemented independently of the system's behavior for $t > 0$. The limitations of such a control scheme are rather obvious : any error on the data (on the initial condition for instance) will not be corrected. Consider for example controlling a car in open loop : to follow a straight line, position the wheels along the axis, hold firmly on the wheel and close your eyes...

To regulate the system, one must therefore use another type of control laws, namely $u(t) = K(t, x(t))$, called *control in closed loop* or *state feedback* : at every time t, one takes into consideration the state at time t in order to determine the control law. Remark

that such control laws also have some disadvantages : they require the knowledge (almost instantaneous) of the state $x(t)$, which may be impossible or expensive.

In the stabilization problem, the goal consists of constructing a control law by state feedback which drives the system to the origin regardless of the initial condition. In the case of linear control systems, we will seek the state feedback as a linear time independent function, i.e., $u(t) = Kx(t)$ with $K \in M_{m,n}(\mathbb{R})$ (such a control law is called *proportional state feedback*).

Définition 3.4. A control system (Σ) is *asymptotically stabilizable* by proportional state feedback if there exists a control law $u(t) = Kx(t)$, with $K \in M_{m,n}(\mathbb{R})$, such that the equation (Σ_u) is asymptotically stable, i.e., for every initial condition $x(0)$, the solution $x(t)$ of (Σ_u) tends to 0 when $t \to +\infty$.

Note that the time interval considered now is infinite, *i.e.* $\tau = +\infty$.

For autonomous linear control systems, (Σ) is of the form (3.1). In that case, if $u(t) = Kx(t)$ is a proportional state feedback, the differential equation (Σ_u) is written

$$x'(t) = Ax(t) + BKx(t) = (A + BK)\, x(t).$$

We know from Corollary 2.27) that such a differential equation is asymptotically stable if and only if all the eigenvalues of the matrix $A + BK$ have negative real parts. The problem to be solved is therefore the following : does there exist $K \in M_{m,n}(\mathbb{R})$ such that the matrix $A + BK$ verifies that condition ?

The goal of this section consists of bringing an answer to that question. First of all, notice that a linear change of variables allows one to replace (A, B) by any other equivalent pair. Indeed, let $\tilde{A} = P^{-1}AP$ and $\tilde{B} = P^{-1}B$ with P the change of basis matrix $x = Py$ with P invertible matrix. One has then $\tilde{K} = KP$ and $P^{-1}(A + BK)P = \tilde{A} + \tilde{B}\tilde{K}$.

With no loss of generality, we can therefore assume that (A, B) is decomposed into controllable and uncontrollable parts, cf. Theorem 3.5 :

$$A = \begin{pmatrix} A_1 & A_2 \\ 0 & A_3 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

with $r \leq n$ the rang of $\mathcal{C}(A, B)$, $A_1 \in \mathcal{M}_r(\mathbb{R})$, $B_1 \in \mathcal{M}_{r,m}(\mathbb{R})$ and (A_1, B_1) controllable. Let $K = (K_1 \;\; K_2)$ with $K_1 \in \mathcal{M}_{m,r}(\mathbb{R})$ et $K_2 \in \mathcal{M}_{m,(n-r)}(\mathbb{R})$. One therefore gets

$$A + BK = \begin{pmatrix} A_1 + B_1K_1 & A_2 + B_1K_2 \\ 0 & A_3 \end{pmatrix}.$$

Hence, for the characteristic polynomials, one has

$$P_{A+BK}(X) = P_{A_1+B_1K_1}(X)P_{A_3}(X).$$

The uncontrollable poles must therefore have negative real part (cf. Remark 3.2). We will next see that this is also a sufficient condition to stabilize (Σ). It suggest the following definition.

Définition 3.5. A pair of matrices (A, B) with $A \in M_n(\mathbb{R})$ and $B \in M_{n,m}(\mathbb{R})$ is called *stabilizable* if the poles of its uncontrollable parts have negative real part.

A unitary polynomial P of $\mathbb{R}_n[X]$ is said to be *assignable* for (A, B) if there exists $F \in M_{m,n}(\mathbb{R})$ such that $P_{A+BF}(X) = P(X)$.

The principal resultat of this section is the following.

Théorème 3.8 (Pole shifting theorem)**.** *Let (A, B) be a pair of matrices with $A \in M_n(\mathbb{R})$, $B \in M_{n,m}(\mathbb{R})$ and r the rank of $\mathcal{C}(A, B)$. The polynomials of $\mathbb{R}_n[X]$ which are assignable for (A, B) are of the form*

$$P_{A+BF}(X) = Q(X)P_{nc}(X),$$

with $Q(X)$ any unitary polynomial of $\mathbb{R}_r[X]$ and $P_{nc}(X)$ the characteristic polynomial of the non controllable part of A. In particular, (A, B) is controllable if and only if every unitary polynomial of $\mathbb{R}_n[X]$ is assignable for (A, B).

Preuve I.

▷ f two pairs of matrices are linearly equivalent, then one can assign them the same polynomials. One can therefore suppose (A, B) decomposed along Kalman and, according to what precedes, it is enough to assume (A, B) controllable.

Let us first prove the theorem when $m = 1$. By Theorem 3.6, the system is linearly equivalent to

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

Set then $K = (k_1 \ \cdots \ k_n)$ et $u = Kx$. On a :

$$A + BK = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \\ k_1 - a_n & k_2 - a_{n-1} & \cdots & k_n - a_1 \end{pmatrix},$$

and therefore

$$\chi_{A+BK}(X) = X^n + (a_1 - k_n)X^{n-1} + \cdots + (a_n - k_1).$$

hence, for every polynomial $P(X) = X^n + \alpha_1 X^{n-1} + \cdots + \alpha_n$, it is enough to choose $k_1 = a_n - \alpha_n, \ldots, k_n = a_1 - \alpha_1$.

For the general case where $m \geq 1$, let us prove the following fundamental lemma due to Ackerman.

Lemme 3.9. *If the pair (A, B) is controllable, then there exists $y \in \mathbb{R}^m$ and $C \in \mathcal{M}_{m,n}(\mathbb{R})$ such that the pair $(A + BC, By)$ is also controllable.*

According to that lemma, for every unitary polynomial P of degree n, there exists $K_1 \in \mathcal{M}_{1,n}(\mathbb{R})$ such that $\chi_{A+BC+ByK_1} = P$, and therefore by setting $K = C + yK_1 \in \mathcal{M}_{m,n}(\mathbb{R})$, one has $\chi_{A+BK} = P$, i.e., the conclusion.

Preuve.

▷ [Proof of the lemma] Let $y \in \mathbb{R}^m$ such that $By \neq 0$. Set $x_1 = By$. The following claim holds true :

Claim 1 : there exists $x_2 \in Ax_1 + \text{Im } B$ (and therefore it exists $y_1 \in \mathbb{R}^m$ such that $x_2 = Ax_1 + By_1$) such that $\dim \text{Vect}\{x_1, x_2\} = 2$.

Indeed, otherwise, one would get $Ax_1 + \text{Im } B \subset \mathbb{R}x_1$, therefore $Ax_1 \in \mathbb{R}x_1$ et Im $B \subset \mathbb{R}x_1$. Thus

$$\text{Im } AB = A\text{Im } B \subset \mathbb{R}Ax_1 \subset \mathbb{R}x_1,$$

and by a trivial induction :

$$\forall k \in \mathbb{N} \quad \text{Im } A^k B \subset \mathbb{R}x_1.$$

One deduces that

$$\text{Im } (B, AB, \ldots, A^{n-1}B) = \text{Im } B + \text{Im } AB + \cdots + \text{Im } A^{n-1}B \subset \mathbb{R}x_1,$$

which contradicts Kalman's condition.

Claim 2 : for every $k \leq n$, there exists $x_k \in Ax_{k-1} + \text{Im } B$ (and therefore there exists $y_{k-1} \in \mathbb{R}^m$ such that $x_k = Ax_{k-1} + By_{k-1}$) such that $\dim E_k = k$, where $E_k = \text{Vect}\{x_1, \ldots, x_k\}$.

Indeed, otherwise, one would get $Ax_{k-1} + \text{Im } B \subset E_{k-1}$, then $Ax_{k-1} \subset E_{k-1}$ et Im $B \subset E_{k-1}$. One deduces

$$AE_{k-1} \subset E_{k-1}.$$

Indeed, notice that $Ax_1 = x_2 - By_1 \in E_{k-1} + \text{Im } B \subset E_{k-1}$, similarly for Ax_2, etc, $Ax_{k-2} = x_{k-1} - By_{k-1} \in E_{k-1} + \text{Im } B \subset E_{k-1}$, and finally, $Ax_{k-1} \in E_{k-1}$.

In consequence :

$$\text{Im } AB = A\text{Im } B \subset AE_{k-1} \subset E_{k-1},$$

and similarly

$$\forall i \in \mathbb{N} \quad \text{Im } A^i B \subset E_{k-1}.$$

One finally gets that

$$\text{Im } (B, AB, \ldots, A^{n-1}B) \subset E_{k-1},$$

which contradicts Kalman's condition.

One hence constructs a basis $(x_1, \ldots, x_n)$ de $\mathbb{R}^n$. Define $C \in \mathcal{M}_{m,n}(\mathbb{R})$ by the relations

$$Cx_1 = y_1,\ Cx_2 = y_2, \ldots,\ Cx_{n-1} = y_{n-1},\ Cx_n \text{ arbitrary.}$$

Then the pair $(A + BC, x_1)$ verifies Kalman's condition because :

$$(A + BC)x_1 = Ax_1 + By_1 = x_2, \ldots, (A + BC)x_{n-1} = Ax_{n-1} + By_{n-1} = x_n.$$

□

The theorem is established.

□

3.5 Observability

3.5.1 Definition and Kalman observability criterion

Consider again a autonomous linear control system (Σ). In general, the available measures do not allow us to directly observe the state $x(t)$ but only a vector $y(t) \in \mathbb{R}^p$, function of the state and the control. We will suppose that this function is linear and time independent, i.e., that the control system is now of the form

$$(\widetilde{\Sigma}) \ : \quad \left\{ \begin{array}{rcl} x'(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) + Du(t) \end{array} \right. , \qquad t \in [0, \tau],$$

where $C \in M_{p,n}(\mathbb{R})$ et $D \in M_{p,m}(\mathbb{R})$. The *observability* problem is the following : knowing $y(t)$ and $u(t)$ for every $t \in [0, \tau]$ $(\tau > 0)$ is it possible to determine the initial condition $x(0)$? Notice that

— the knowledge of $x(0)$ is equivalent to that of $x(t)$ for every $t \in [0, \tau]$ since, according to the variation of constant formula

$$x(t) = e^{tA}x(0) + \int_0^\tau e^{(\tau - s)A} Bu(s)ds,$$

the second term of the right-hand side of the above equality is supposed to be known ;

— one can assume $D = 0$ and $B = 0$ since $u(\cdot)$ is known.

It is therefore enough to study the observability problem for the reduced system

$$(\widetilde{\Sigma}_0) \ : \quad \left\{ \begin{array}{rcl} x'(t) & = & Ax(t) \\ y(t) & = & Cx(t) \end{array} \right. , \qquad t \in [0, \tau],$$

i.e., to study of $y(t) = Ce^{tA}x_0$. Let us call *unobservability space* $\mathcal{U}_\tau$ of the system $(\widetilde{\Sigma}_0)$ the set of initial conditions $x(0) \in \mathbb{R}^n$ for which the solution $y(t)$ is identically equal to zero on $[0, \tau]$, i.e.,

$$\mathcal{U}_\tau = \left\{ x_0 \in \mathbb{R}^n \ : \ \begin{array}{c} \text{la solution de } (\widetilde{\Sigma}_0) \\ \text{with } x(0) = x_0 \text{ verifie } y(t) \equiv 0 \end{array} \right\}.$$

Définition 3.6. The system $(\widetilde{\Sigma}_0)$ is said to be *observable* if its unobservability space reduces to $\{0\}$.

The following elementary result shows that this definition of observability corresponds to the question initially asked.

Proposition 3.10. *If the system* $(\widetilde{\Sigma}_0)$ *is observable, the knowledge of* $y(\cdot)$ *of* $[0,\tau]$ *determines uniquely* $x(0)$*.*

Preuve.

▷ If it were not the case, then there would exist two distinct vectors x_0 and $\tilde{x}_0$ in $\mathbb{R}^n$ such that

$$Ce^{tA}x_0 = Ce^{tA}\tilde{x}_0,$$

implying that $Ce^{tA}(x_0 - \tilde{x}_0) = 0$. According to the definition of observability one would get that $x_0 = \tilde{x}_0$. □

There exists a simple criterion allowing one to determine if a system is observable.

Théorème 3.11 (Kalman observability criterion)**.** *The unobservability space of the system* $(\widetilde{\Sigma}_0)$ *is the kernel of the matrix* $(np \times n)$

$$\mathcal{O} = \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

In other words, the system $(\widetilde{\Sigma}_0)$ *is observable if and only if* $\ker \mathcal{O} = \{0\}$*.*

Preuve.

▷ (The following argument was already used in the proof of Theoreme 3.2.) According to Cayley-Hamilton's theorem, for every $t \in [0,\tau]$

$$Ce^{tA} \in \mathrm{Vect}(C, CA, \ldots, CA^{n-1}).$$

It implies that

$$CA^j v = 0, \quad j = 0, \ldots, n-1, \tag{3.4}$$

which is equivalent to $Ce^{tA}v = 0$, i.e., v is in the unobservability space of $(\widetilde{\Sigma}_0)$. On the other hand, Condition (3.4) is equivalent to the fact that $v \in \ker \mathcal{O}$. We have therefore proved that $\ker \mathcal{O}$ is included in the unobservability space of $(\widetilde{\Sigma}_0)$.

▷ Conversely, assume that for every $t \in [0, \tau]$,

$$Ce^{tA}v = 0.$$

then, by taking j time derivatives of the above equality at $t = 0$ $(0 \leq j \leq n-1)$, one gets

$$\forall 0 \leq j \leq n-1,\ CA^j v = 0,$$

i.e., $v \in \ker \mathcal{O}$. The converse inclusion is shown.

□

Remarque. If one compares the above theorem with Theorem 3.2, then one notices that the system $(\widetilde{\Sigma}_0)$ is observable if and only if the dual system $(\Sigma) : z'(t) = A^T z(t) + C^T u(t)$ is controllable (take the transpose of the matrix $\mathcal{O}$). This is the *duality controllability/observability*. This important fact allows one to transfer to observed systems all the results obtained for controlled systems.

Définition 3.7 (Linear Equivalence)**.** The systems

$$\begin{cases} \dot{x}_1 = A_1 x_1 + B_1 u_1 \\ y_1 = C_1 x_1 \end{cases} \quad \text{and} \quad \begin{cases} \dot{x}_2 = A_2 x_2 + B_2 u_2 \\ y_2 = C_2 x_2 \end{cases}$$

are said to be *linearly equivalent* if there exists a matrix $P \in GL_n(\mathbb{R})$ such that

$$A_2 = PA_1P^{-1},\ B_2 = PB_2,\ C_2 = C_1P^{-1}$$

(and in this case $x_2 = Px_1, u_2 = u_1, y_2 = y_1$).

Proposition 3.12. *Every system $\dot{x} = Ax + Bu$, $y = Cx$, is linearly equivalent to a system $\dot{\bar{x}} = \bar{A}\bar{x} + \bar{B}u$, $y = \bar{C}\bar{x}$, with*

$$\bar{A} = \begin{pmatrix} \bar{A}_1 & 0 \\ \bar{A}_2 & \bar{A}_3 \end{pmatrix},\ \bar{C} = (\bar{C}_1\ 0),$$

i.e.

$$\begin{cases} \dot{\bar{x}}_1 = \bar{A}_1\bar{x}_1 + \bar{B}_1 u \\ \dot{\bar{x}}_2 = \bar{A}_2\bar{x}_1 + \bar{A}_3\bar{x}_2 + \bar{B}_2 u \qquad \textit{unobservable part} \\ y_1 = \bar{C}_1\bar{x}_1 \end{cases}$$

and the pair $(\bar{A}_1, \bar{C}_1)$ is observable.

Preuve.

▷ It is enough to apply the result seen for controllability to the system $\dot{x} = A^T x + C^T u$. □

Définition 3.8. In this decomposition, the eigenvalues of $\bar{A}_3$ are called *unobservable poles* of A and the eigenvalues of $\bar{A}_1$ are called *observable poles* of A.

Proposition 3.13 (Observability Canonical Form, case $p = 1$)**.** *If $p = 1$, the system $\dot{x} = Ax + Bu$, $y = Cx$, is observable if and only if it is linearly equivalent to the system $\dot{x}_1 = A_1 x_1 + B_1 u$, $y = C_1 x_1$, with*

$$A_1 = \begin{pmatrix} 0 & & \cdots & 0 & -a_n \\ 1 & 0 & & & \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \\ 0 & \cdots & 0 & 1 & -a_1 \end{pmatrix}, \quad C_1 = (0 \ \cdots \ 0 \ 1).$$

3.5.2 Stabilisation by Static-state feedback

Given a controllable and observable system $\dot{x} = Ax + Bu$, $y = Cx$, does there exists a feedback $u = Ky$ stabilizing the system, i.e., if the matrix $A + BKC$ is Hurwitz ?

The answer is *NO*. To see that, consider the matrices

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \ B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \ C = (1 \ 0).$$

The system $\dot{x} = Ax + Bu$, $y = Cx$, is trivially controllable and observable. However, for every scalar matrix $K = (k)$, the matrix

$$A + BKC = \begin{pmatrix} 0 & 1 \\ k & 0 \end{pmatrix}$$

is not Hurwitz.

In conclusion, a static-state feedback is not enough in general. This is why we will build in the sequel a dynamic-state feedback.

3.5.3 Luenberger asymptotic Observer

Motivation : assume that the system $\dot{x} = Ax + Bu$, $y = Cx$, is observable.The goal consists of building an *asymptotic observer* $\hat{x}(\cdot)$ de $x(\cdot)$, i.e. a dynamic function $\hat{x}(\cdot)$ of the observable $y(\cdot)$, such that $\hat{x}(t) - x(t) \underset{t \to +\infty}{\longrightarrow} 0$. The idea is to copy the dynamics of the observed system and to add a correction term which takes into account the error between the prediction and the exact state.

Définition 3.9. An *asymptotic observer* (or *Luenberger observer*) $\hat{x}(\cdot)$ of $x(\cdot)$ is a solution of a system of the type

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L(C\hat{x}(t) - y(t)),$$

where $L \in \mathcal{M}_{n,p}(\mathbb{R})$ is called *gain matrix*, such that

$$\forall x(0), \hat{x}(0) \in \mathbb{R}^n \quad \hat{x}(t) - x(t) \underset{t\to+\infty}{\longrightarrow} 0.$$

Remarque. Let $e(t) = \hat{x}(t) - x(t)$ be the error between the prediction $\hat{x}(\cdot)$ and the real state $x(\cdot)$. One has :

$$\dot{e}(t) = (A + LC)e(t),$$

and therefore $e(t) \underset{t\to+\infty}{\longrightarrow} 0$ for every initial value $e(0)$ if and only if the matrix $A + LC$ is Hurwitz. Constructing an asymptotic observer resumes therefore to determine a gain matrix L such that $A + LC$ is Hurwitz. Hence, proceeding in a dual manner with respect to the pole shifting theorem, one gets :

Théorème 3.14 (Pole shifting theorem for observability)**.** *If the pair (A, C) is observable, then the system admits an asymptotic observer (i.e., one can build a gain matrix L such that $A + LC$ is Hurwitz).*

Preuve.

▷ The pair (A^T, C^T) being controllable, according to the pole shifting theorem there exists a matrix L^T such that the matrix $A^T + C^T L^T$ is Hurwitz.

□

3.5.4 Stabilization by dynamic output feedback

We have seen how to build :
— a state feedback for a controllable system,
— an asymptotic observer for an observable system.
It seems therefore reasonable, for a controllable and observable system, to build a fedback in terms of the asymptotic observer of the state : it is the step of *feedback-observer synthesis*.

Définition 3.10. We call *dynamic output feedback*, or *feedback-observer*, the feedback $u = K\hat{x}$, where

$$\dot{\hat{x}} = A\hat{x} + Bu + L(C\hat{x} - y).$$

Théorème 3.15. *Stabilization theorem using a dynamic output feedback] If the system $\dot{x} = Ax + Bu$, $y = Cx$, is controllable and observable, then it is stabilizable by a dynamic output feedback, i.e., there exists gain matrices $K \in \mathcal{M}_{m,n}(\mathbb{R})$ and $L \in \mathcal{M}_{n,p}(\mathbb{R})$ such that the matrices $A + BK$ and $A + LC$ are Hurwitz, and the closed-loop system*

$$\begin{aligned} \dot{x} &= Ax + BK\hat{x} \\ \dot{\hat{x}} &= (A + BK)\hat{x} + LC(\hat{x} - x) \end{aligned}$$

is asymptotically stable.

Preuve.

▷ Set $e = \hat{x} - x$. then :

$$\frac{d}{dt}\begin{pmatrix} x \\ e \end{pmatrix} = \begin{pmatrix} A + BK & BK \\ 0 & A + LC \end{pmatrix}\begin{pmatrix} x \\ e \end{pmatrix},$$

and therefore this system is asymptotically stable if and only if the matrices $A + BK$ and $A + LC$ are Hurwitz, which is possible with the properties of controllability and observability. □

Remarque. The fact that the stabilization task is solved independently from the reconstruction task bears the name of *separation principe*.

Chapitre 4

Nonlinear controllability

In this Chapter we study the controllability of the nonlinear system

$$\dot{x} = F(x, u), \qquad x \in \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m, \tag{4.1}$$

where F is smooth function of its arguments.

To simplfy the discussion, we are going to consider only piecewise constant controls and we assume that the control system (4.1) is complete, i.e. for every choice of the control function and of the initial conditions, we assume that its solutions are defined for all $t > 0$.

Denote by $x(t; x_0, u(\cdot))$ the solution of (4.1) starting at time zero from x_0 and corresponding to a control function $u(\cdot)$. We recall the definitions of the *reachable* (or *attainable*) sets starting from x_0 :

$$\begin{aligned} \mathcal{A}(\tau, x_0) &= \{x_1 \in \mathbb{R}^n \mid \exists\ u(\cdot) : [0, \tau] \to U,\ x(\tau; x_0, u(\cdot)) = x_1\} \\ \mathcal{A}(\leq \tau, x_0) &= \cup_{t \in [0,\tau]} \mathcal{A}(t, x_0) \\ \mathcal{A}(x_0) &= \cup_{t \in [0,+\infty[} \mathcal{A}(t, x_0). \end{aligned}$$

Définition 4.1. The system (4.1) is said to be *completely controllable* if $\mathcal{A}(x_0) = \mathbb{R}^n$ for every $x_0 \in \mathbb{R}^n$. The system (4.1) is said to be *small time locally controllable at* x_0 if x_0 belongs to the interior of $\mathcal{A}(\leq \tau, x_0)$ for every $\tau > 0$.

The main question that we address in this chapter is : under which conditions (4.1) is completely controllable ?

In the following it will be useful to think to the system (4.1) as a *family of vector fields* parameterized by u. In other words we will often consider instead of the control system (4.1), the family [1]

$$\mathcal{F} = \{F(\cdot, u) \mid u \in U\}$$

To simplify the discussion, in this chapter, vector fields are considered smooth and complete.

1. In a differential equation $\dot{x} = f(x)$, usually f it is called a vector field, since it is a map that to every position x associate a velocity vector $f(x)$.

4.1 Lie brackets of vector fields

A crucial object in studying the controllability of $\mathcal{F}$ is the Lie algebra generated by the vector fields of $\mathcal{F}$. Let us first define the Lie brackets between two vector fields f and g, that is the vector field defined by

$$[f,g](x) = Dg(x)f(x) - Df(x)g(x).$$

Notice that for every $\lambda_1, \lambda_2 \in \mathbb{R}$,

$$[f, \lambda_1 g_1 + \lambda_2 g_2] = \lambda_1 [f, g_1] + \lambda_2 [f, g_2]$$

and that

$$[g,f] = -[f,g].$$

In particular $[f,f] = 0$. Notice moreover that the value of $[f,g]$ at a point x does not depend only from the values of f and g at x but from their values in a neighborhood of x. However if at x_0 we have $f(x_0) = g(x_0) = 0$ then $[f,g](x_0) = 0$.

To understand the geometric meaning of the Lie bracket, let us denote [2] by e^{tf} and e^{tg}, the flows corresponding to the differential equations $\dot{x} = f(x)$ et $\dot{x} = g(x)$. This means that $e^{tf}(x_0)$ is the evaluation at time t of the solution of the Cauchy problem

$$\dot{x}(t) = f(x(t)), \qquad x(0) = x_0.$$

Notice that for every $t, s \in \mathbb{R}$, $e^{(t+s)f} = e^{tf} \circ e^{sf}$. In particular e^{tf} is invertible and $(e^{tf})^{-1} = e^{-tf}$. Moreover $\frac{d}{dt} e^{tf} x = f(e^{tf} x)$. In particular $\frac{d}{dt}\big|_{t=0} e^{tf} x = f(x)$.

The following lemma show that $[f,g]$ is related to the properties of commutation of the flows associated to f et g.

Lemme 4.1. *For every $x \in \mathbb{R}^n$,*

$$e^{-tg} \circ e^{-tf} \circ e^{tg} \circ e^{tf}(x) = x + t^2 [f,g](x) + o(t^2),$$

for t that tends to zero.

Preuve.

▷ It is enough to compute for each flow the Taylor expansion at order 3. We have

$$e^{tf}(x) = x + tf(x) + \frac{t^2}{2} Df(x)f(x) + O(t^3),$$

and

$$e^{tg} \circ e^{tf}(x) = x + t\big(f(x) + g(x)\big) + \frac{t^2}{2} Df(x)f(x) + t^2 Dg(x)f(x) + \frac{t^2}{2} Dg(x)g(x) + O(t^3).$$

2. Remark that this is only a notation. It does not make sense to compute the exponential of a vector

Then

$$e^{-tf} \circ e^{tg} \circ e^{tf}(x) = x + tg(x) + [f,g](x) + \frac{t^2}{2} Dg(x)g(x) + O(t^3),$$

from which it follows the result.

□

An important corollary of the previous Lemma is

Corollaire 4.2. *The flows e^{tf} et e^{tg} (corresponding to the vector fields f and g) commute for every t if and only if their Lie bracket $[f,g]$ vanish.*

From the previous result it follows that if in the family $\mathcal{F}$ there are non commuting vector fields, then one can "gain" new directions.

Définition 4.2. Let $\mathcal{F}$ be a family of vector fields. We call $\mathrm{Lie}(\mathcal{F})$ the vector space spanned by all the vector fields of $\mathcal{F}$ and by the vector fields of the form

$$[f_1, [f_2, [\ldots, [f_{k-2}, [f_{k-1}, f_k]]\ldots]]],$$

where $k \geq 2$ et $f_1, \ldots, f_k \in \mathcal{F}$.

Définition 4.3. We say that the family $\mathcal{F}$ is *Lie bracket generated* at a point x if the dimension of $\mathrm{Lie}_x(\mathcal{F}) := \{f(x) \mid f \in \mathrm{Lie}(\mathcal{F})\}$ is equal to n. We say that the family $\mathcal{F}$ is *Lie bracket generated* if this condition is verified for every $x \in \mathbb{R}^n$.

4.2 The Krener Theorem : local accessibility

The fact that a control system is Lie Bracket generated does not permit in general to conclude that it is controllable. Consider for instance the family on the plane $\mathcal{F} = \{(1,1),(1,-1)\}$. It is Lie bracket generated but starting from the origin one cannot reach points having negative first coordinate.

The Lie Bracket generating condition permits to say that a system is *locally accessible* in the following sense

Théorème 4.3 (Krener)**.** *If $\mathcal{F}$ is Lie Bracket generated at x_0, then for every $\tau > 0$, x_0 belongs to the closure of the interior of $\mathcal{A}(\leq \tau, x_0)$.*

Preuve.

▷ Remarquons d'abord que $\mathcal{F}_U$ satisfait la condition de génération par crochets de Lie en tous les points d'un voisinage de x_0. (Si n champs de vecteurs sont linéairement indépendants à x_0, ils le sont dans un voisinage.)
Remarquons aussi qu'il existe $f \in \mathcal{F}_U$ tel que $f(x_0) \neq 0$: on aurait sinon $\mathrm{Lie}_{x_0}(\mathcal{F}_U) = \{0\}$.
Si $\dim(\Omega) = 1$ la preuve est finie. Si $\dim(\Omega) > 1$ et si tout champ dans $\mathcal{F}_U$ est tangent à la courbe $t \mapsto e^{tf}(x_0)$, $0 < t < \varepsilon$, alors on déduit du Lemme 4.1 que $\mathrm{Lie}_{e^{tf}(x_0)}(\mathcal{F}_U)$ est aussi tangent à telle courbe et qu'il n'a donc pas la même dimension que Ω. Ceci contredit la première remarque faite dans cette preuve. Il existe donc $g \in \mathcal{F}_U$ et $0 < \bar{t} < \varepsilon$ tels que f et g sont linéairement indépendants dans un voisinage de $x_1 = e^{\bar{t}f}(x_0)$. Alors $(t, s) \mapsto e^{sg} \circ e^{tf}(x_0)$, $0 < s < \varepsilon'$, $\bar{t} - \varepsilon' < t < \bar{t} + \varepsilon'$, a comme image une surface de dimension deux.
Si $\dim(\Omega) = 2$ la preuve est finie. Sinon on recommence le même argument et on conclue par récurrence par rapport à la dimension de Ω.

□

This theorem says that for a Lie bracket generated system, the trajectories starting from a point can reach (in an arbitrarily small time) a set having nonempty interior.

4.3 Symmetric systems

When the family $\mathcal{F}$ is Lie bracket generated and it is symmetric (ie. $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$), then one obtain that the system is completely controllable.

Théorème 4.4 (Chow). *If $\mathcal{F}$ is Lie bracket generated and syummetric, then $\mathcal{A}(x_0) = \mathbb{R}^n$*

Preuve.

▷ Soit $x_0 \in \Omega$. Remarquons que $\mathcal{F}_U$, avec $F(x, u) = \sum_{i=1}^m u_i f_i(x)$, satisfait la condition de génération par crochets de Lie. Nous déduisons donc du théorème de Krener que $\mathcal{A}(x_0)$ contient un ouvert non vide ω.
Choisissons $u^1, \ldots, u^k \in U$ et $t_1, \ldots, t_k > 0$ tels que

$$e^{t_k \sum_{i=1}^m u_i^k f_i} \circ \ldots \circ e^{t_1 \sum_{i=1}^m u_i^1 f_i}(x_0) \in \omega.$$

Puisque $-u^1, \ldots, -u^k \in U$, nous avons que

$$e^{t_1 \sum_{i=1}^m (-u_i^1) f_i} \circ \ldots \circ e^{t_k \sum_{i=1}^m (-u_i^k) f_i}(\mathcal{A}(x_0)) \subset \mathcal{A}(x_0).$$

En particulier, $\mathcal{A}(x_0)$ contient

$$e^{t_1 \sum_{i=1}^m (-u_i^1) f_i} \circ \ldots \circ e^{t_k \sum_{i=1}^m (-u_i^k) f_i}(\omega)$$

qui est un voisinage de x_0.

Nous en déduisons que $\mathcal{A}(x_0)$ est ouvert. (En effet, si $x_1 \in \mathcal{A}(x_0)$, alors $x_1 \in \text{int}(\mathcal{A}(x_1)) \subset \mathcal{A}(x_1) \subset \mathcal{A}(x_0)$.) De plus, la symétrie de U implique (comme nous venons de le voir) que $x_1 \in \mathcal{A}(x_0)$ si et seulement si $x_0 \in \mathcal{A}(x_1)$. Donc $\{\mathcal{A}(x_0) \mid x_0 \in \Omega\}$ est une partition de Ω en ouverts disjoints. Ω étant connexe, nous pouvons conclure que $\mathcal{A}(x_0) = \Omega$ pour tout $x_0 \in \Omega$. □

4.4 Compatible vector fields

When a family of vector fields is not symmetric, one can get controllability results with the technique of adding "compatible" vector fields. The idea is to look for vector fields that added to the family $\mathcal{F}$ change the reachable set only for its closure and such, that adding them, it is easy to prove controllability results.

The key point is the following corollary of the Krener Theorem

Corollaire 4.5. *If $\mathcal{F}$ is Lie bracket generated and $\mathcal{A}(x_0)$ is dense in $\mathbb{R}^n$ for some x_0, then $\mathcal{A}(x_0) = \mathbb{R}^n$.*

Preuve.

▷ Soit $x_1 \in \Omega$ et considérons le système

$$\dot{x} = -F(x, u), \qquad x \in \Omega \subset \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m, \tag{4.2}$$

qui est obtenu à partir de (4.1) par renversement temporel. La famille de champs de vecteurs admissibles pour (4.2) étant $-\mathcal{F}_U$, elle satisfait la condition de génération par crochets de Lie.

L'ensemble atteignable pour (4.2) au départ de x_1, noté par $\mathcal{A}^-(x_1)$, contient donc un ouvert non vide (Krener). En particulier, $\mathcal{A}^-(x_1)$ a intersection non vide avec $\mathcal{A}(x_0)$. Cela signifie précisément que $x_1 \in \mathcal{A}(x_0)$. Puisque $x_1 \in \Omega$ était arbitraire, nous avons montré que $\mathcal{A}(x_0) = \Omega$.

□

Corollary 4.5 suggest the following

Définition 4.4. A vector field g is said to be *compatible* with the family $\mathcal{F}$ if defining $\hat{\mathcal{F}} = \mathcal{F} \cup \{g\}$ we have the following. For every x_0, the reachable set $\hat{\mathcal{A}}(x_0)$ of $\hat{F}$ is contained in the closure of $\mathcal{A}_{x_0}$.

From Lemma 4.5 we have then immediately

Proposition 4.6. *If $\mathcal{F}$ is a Lie Bracket generated family of vector fields, g is compatible with $\mathcal{F}$ and $\mathcal{F} \cup \{g\}$ is controllable, then $\mathcal{F}$ is controllable as well.*

Now we are going to present some techniques to identify compatible vector fields. Most of these techniques works for *affine*control systems, namely of the form

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad x \in \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m.$$

Here f_0 is called the *drift* and $f_1, \ldots, f_m$ are called the *controlled vector fields*.

4.4.1 Recurrent drift

Définition 4.5 (Recurrent vector field)**.** A vector field f is said to be *recurrent* if for every point $x_0 \in \mathbb{R}^n$, every neighborhood V of x_0 and every time $t > 0$, there exists $t^* > t$ such that $e^{t^* f}(x_0) \in V$.

Notice that if the trajectories of f are periodic (maybe with the period that depends on the trajectory), then f is recurrent.

Lemme 4.7. *If f is recurrent and compatible with $\mathcal{F}$, then $-f$ is also compatible with $\mathcal{F}$.*

Preuve.

▷ Soient $x_0 \in \Omega$ et $t > 0$. Il suffit de montrer que $e^{-tf}(x_0)$ est la limite de points atteignables au départ de x_0.
La définition de champ récurrent implique l'existence d'une suite croissante non bornée $\{t_k\}_{k\in\mathbb{N}}$ de temps positifs telle que $e^{t_k f}(x_0) \to x_0$ pour $k \to \infty$.
Nous avons donc que $e^{(t_k - t)f}(x_0) \to e^{-tf}(x_0)$ pour $k \to \infty$ et $t_k > t$ pour k suffisamment grand.
Puisque f est compatible avec $\mathcal{F}_U$, alors $e^{(t_k-t)f}(x_0)$ appartient à la fermeture de $\mathcal{A}(x_0)$ pour tout k tel que $t_k > t$. Nous en déduisons que $e^{-tf}(x_0)$ appartient à la fermeture de $\mathcal{A}(x_0)$.
□

As a consequence of the previous Lemma we have the following

Corollaire 4.8. *Consider the control system*

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \ldots, u_m) \in U. \tag{4.4}$$

Assume that **(i)** 0 *belongs to the interior of* U, **(ii)** $\{f_0, \ldots, f_m\}$ *are Lie bracket generated,* **(iii)** f_0 *is recurrent. Then the system is completely controllable.*

Preuve.

▷ Notice that f_0 is compatible with $\mathcal{F}$ since 0 belongs to the interior of U. Lemme 4.7 states the equivalence between the controllability of (4.4) and that of

$$\dot{x} = \sum_{i=0}^{m} u_i f_i(x), \qquad (u_0, \dots, u_m) \in \{-1, 1\} \times U.$$

The controllability of this system follows from the Chow theorem since U contains a symmetric set.

□

Exemple. Soit $\Omega = S^2$, la sphère unité de $\mathbb{R}^3$, c'est-à-dire $\{x \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 1\}$. Considérons les deux champs de vecteurs

$$f_0(x) = \begin{pmatrix} -x_2 \\ x_1 \\ 0 \end{pmatrix}, \qquad f_1(x) = \begin{pmatrix} 0 \\ -x_3 \\ x_2 \end{pmatrix},$$

dont les flots au temps t sont les rotations d'angle t par rapport, respectivement, à $(0, 0, 1)^T$ et $(1, 0, 0)^T$.

Le système commandé

$$\dot{x} = f_0(x) + u f_1(x), \qquad u \in (-1, 1), \qquad x \in S^2,$$

est complètement commandable puisque les trajectoires de f_0 sont périodiques (et donc f_0 est récurrent) et le crochet entre f_0 et f_1 est donné par

$$[f_0, f_1](x) = \begin{pmatrix} -x_3 \\ 0 \\ x_1 \end{pmatrix},$$

de telle sorte que $\mathrm{Lie}_x(\{f_0, f_1\})$ est de dimension deux pour tout $x \in S^2$.

4.4.2 Non recurrent drift

When the drift is not recurrent one can still get complete controllability if the controls are unbounded and if it is not necessary to use the drift to get a Lie algebra of full dimension. More precisely we have the following

Proposition 4.9. *Consider the control system*

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \dots, u_m) \in \mathbb{R}^m \tag{4.6}$$

. If $\mathcal{G} = \{f_1, \dots, f_m\}$ is Lie bracket generated then (4.6) is completely controllable.

Preuve.

▷ Nous déduisons de ce critère le résultat suivant, qui s'applique à la classe des systèmes affines dans le contrôle.

The Chow Theorem guarantees that the system

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x), \qquad (u_1, \ldots, u_m) \in \mathbb{R}^m$$

is completely controllable.

We have then to show that $\sum_{i=1}^{m} u_i f_i$ is compatible with $\mathcal{F}$.

Remark that

$$\sum_{i=1}^{m} u_i f_i = \lim_{n\to\infty} \frac{1}{n}(f_0 + \sum_{i=1}^{m} (nu_i) f_i)$$

and that $\frac{1}{n}(f_0 + \sum_{i=1}^{m}(nu_i)f_i)$ is compatible with $\mathcal{F}_U$. The thesis follows from the fact that if a vector field is the uniform limit on all compacts of a sequence of compatible vector fields, then it is compatible as well (from the continuity of solutions of ODEs with respect to the vector field).

□

4.4.3 Convexification

A very useful criterium is the one that states that a convex combination of vector fields of $\mathcal{F}$ is compatible with $\mathcal{F}$. It formalize the intuition that if one commutes quickly between the dynamics of two vector fields, and one stays the same time on each dynamics, then the corresponding trajectory is close the trajectory of $\frac{f+g}{2}$ starting from the same point.

Lemme 4.10. *For every $\lambda_1, \ldots, \lambda_k \geq 0$ such that $\sum_{i=1}^{k} \lambda_k = 1$ and $f_1, \ldots, f_k \in \mathcal{F}$, the vector field $\lambda_1 f_1 + \cdots + \lambda_k f_k$ is compatible with $\mathcal{F}$.*

The proof of this Lemma is quite technical and it is based on the Gronwall inequality. From this Lemma one can get a useful corollary of the Chow Theorem.

Corollaire 4.11. *Consider the control system $\dot{x} = F(x,u)$, $x \in \mathbb{R}^n$, $u \in U \subset \mathbb{R}^m$. If the corresponding family $\mathcal{F}$ is Lie brackett generated and if 0 belongs to the interior of the convex hull of $\{F(x,u) \mid u \in U\}$, then (4.1) is it is competely controllable.*

Moreover in Corollary 4.8 one can relax **(i)** in the following hypothesis :

(ibis) 0 belongs to the interior of the convex hull of U.

4.5 Orbites et conditions nécessaires pour la commandabilité

Nous avons vu dans les sections précédentes plusieurs conditions suffisantes pour la commandabilité d'un système de contrôle non linéaire.

Cette section présente plutôt des conditions nécessaires, déduites d'un résultat profond de nature géométrique, le théorème de l'orbite. Ce théorème assure que, dans les cas non "pathologiques", l'algèbre de Lie associée à la famille $\mathcal{F}_U$ mesure de façon précise la taille de l'ensemble des directions auquelles on peut accéder à partir d'un point.

Nous définissons l'orbite à partir d'un point $x_0 \in \Omega$ pour le système (4.1) comme l'ensemble

$$\mathcal{O}(x_0) = \{e^{t_k f_k} \circ \cdots \circ e^{t_1 f_1}(x_0) \mid k \in \mathbb{N},\ t_1, \ldots, t_k \in \mathbb{R},\ f_1, \ldots, f_k \in \mathcal{F}_U\}.$$

Rappelons, par comparaison, que

$$\mathcal{A}(x_0) = \{e^{t_k f_k} \circ \cdots \circ e^{t_1 f_1}(x_0) \mid k \in \mathbb{N},\ t_1, \ldots, t_k \geq 0,\ f_1, \ldots, f_k \in \mathcal{F}_U\}.$$

Nou avons alors le théorème suivant.

Théorème 4.12 (Orbite)**.** *Pour chaque $x_0 \in \Omega$, l'ensemble $\mathcal{O}(x_0)$ a la structure d'une variété immergée. En particulier, celle-ci a même dimension en tout point. De plus, l'espace des directions tangentes à $\mathcal{O}(x_0)$ à un point $x \in \mathcal{O}(x_0)$ contient $\mathrm{Lie}_x(\mathcal{F}_U)$ et les deux espaces sont égaux si une des deux conditions suivantes est vérifiée : (i) chaque composante de chaque champ de vecteurs de $\mathcal{F}_U$ est une fonction analytique, ou (ii) la dimension de $\mathrm{Lie}_x(\mathcal{F}_U)$ est constante par rapport à $x \in \mathcal{O}(x_0)$.*

Le corollaire suivant retient du théorème de l'orbite les conséquences les plus directement exploitables pour l'analyse de commandabilité d'un système non linéaire.

Corollaire 4.13. *Si $\mathcal{F}_U$ ne satisfait pas la condition de génération par crochets de Lie et si (i) chaque composante de chaque champ de vecteurs dans $\mathcal{F}_U$ est une fonction analytique, ou (ii) la dimension de $\mathrm{Lie}_x(\mathcal{F}_U)$ est constante par rapport à $x \in \Omega$, alors (4.1) n'est pas complètement commandable.*

Chapitre 5

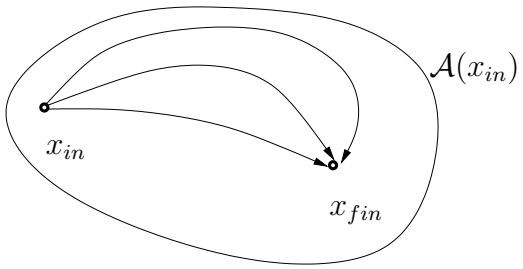
Introduction to Optimal Control

Consider the control problem

$$\dot{x} = F(x, u), \qquad x \in \mathbb{R}^n, \qquad u \in U \subset \mathbb{R}^m, \tag{5.1}$$

where F is a smooth function of its arguments.

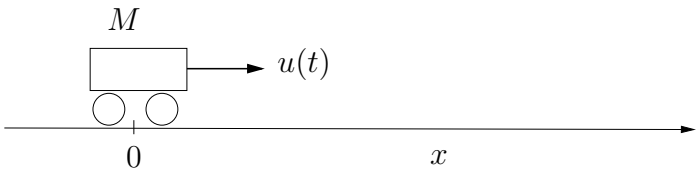
When $x_1 \in \mathcal{A}(x_0)$ usually there are many trajectories joining x_0 to x_1 (in most cases, an infinite number). In many applications it is very important to find the "best trajectory" for a given criterium. Let us see some examples.



Example 1 Consider a cart of mass M on which we act with an external force $u(t)$ such that $|u(t)| \leq F_{max} > 0$. The corresponding dynamics has the form $M\ddot{x} = u(t)$. Setting $x_1 = x$ and $x_2 = \dot{x}$ the control system becomes.

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{u}{M}, \quad |u| \leq F_{max} \end{aligned}$$

Find the trajectory joining $(x_1, x_2) = (0, 0)$ to $(x_1, x_2) = (a, 0)$ in minimum time. Notice that for this problem, initial and final points are fixed, but the final time is free. The criterium can be written in integral form since $T = \int_0^T 1\, dt$.



Example 2 Consider the robot $E = M6$:

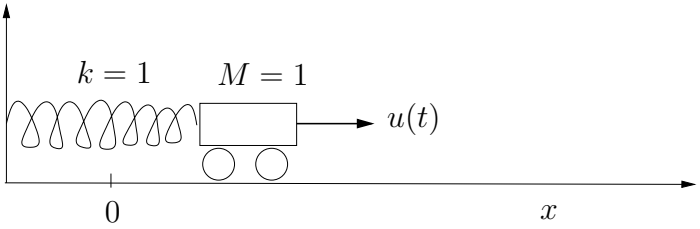
$$\begin{aligned} \dot{x} &= u_1 \cos(\theta) \\ \dot{y} &= u_1 \sin(\theta) \\ \dot{\theta} &= u_2, \qquad\qquad u_1, u_2 \in \mathbb{R}. \end{aligned} \tag{5.2}$$

Find the trajectory joining $(x, y, \theta) = (0, 0, 0)$ to $(x, y, \theta) = (\bar{x}, \bar{y}, \bar{\theta})$ minimizing an "energy-like cost" $\int_0^T (u_1^2 + u_2^2)\, dt$, where T is fixed. Notice that for this problem the final time and the initial and final points are fixed. A variant of this problem is to start from the point $(x, y, \theta) = (0, 0, 0)$ and to reach $(x, y) = (\bar{x}, \bar{y})$ with any orientation and minimizing the same criterium.

Example 3 Consider an harmonic oscillator of mass $M = 1$ and elastic constant $k = 1$, on which we act with an external force $u(t)$ such that $|u(t)| \leq 1$. The corresponding dynamics has the form $\ddot{x} = -x + u(t)$. Setting $x_1 = x$ and $x_2 = \dot{x}$ the control system becomes,

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + u, \quad |u| \leq 1 \end{aligned}$$

Find the trajectory starting from $(x_1, x_2) = (0, 0)$ and maximizing $x_1^2(1)$ (i.e. minimizing $-x_1^2(1)$). Notice that for this problem the final time and the initial point are fixed, but the final point is not. Moreover the cost is not in integral form (it only depends on the final position).



All these problems can be written in the following form

Optimal Control Problem

$$\dot{x} = F(x,u), \quad x \in \mathbf{R}^n, u \in U \subset \mathbf{R}^m,$$
$$x(0) = x_{in}, \quad x(T) \in \mathcal{T} \subset \mathbf{R}^n$$
$$\int_0^T L(x(t),u(t))\,dt + \psi(x(T) \to \min .$$

Here F, L, ψ are smooth functions of their arguments and the final time T can be fixed or free. Here we assume that $u(.)$ is a L^∞ function.

Observations

— $\mathcal{T}$ is called the target. When $\mathcal{T} = \{x_1\}$, for some $x_1 \in \mathbf{R}$ then the final position is fixed. In this case the term $\psi(x(T)$ does not play any role.

— Notice that this minimization problem is on a space that usually has infinite dimension (it is the space of curves joining the initial point to the target). The dynamics $\dot{x} = F(x,u)$ can be though as a "constraint". Notice that when the system is controllable this constraint is on the velocity (such kind of constraint is called nonholonomic) and not on the position.

— Here we assume that $u(\cdot)$ is an L^∞ function. The class of piecewise-constant controls is too small to guarantee existence of optimal controls under reasonable hypotheses.

5.1 Steps in solving an Optimal Control Problem

As for the problem of finding the minimum of a smooth function $L : \mathbf{R} \to \mathbf{R}$, the steps to find a solution of a minimization problem are the following ones.

— Find conditions which guarantee the existence of solutions. Recall that among smooth functions $L : \mathbf{R} \to \mathbf{R}$ it is easy to find examples not admitting a minimum (e.g. the function e^{-x} and the function x do not admit minima, for different reasons).

— Apply first order necessary conditions. For a smooth function $L : \mathbf{R} \to \mathbf{R}$ this means requiring that $L'(x) = 0$. This condition identifies local minima, local maxima and saddles.

— Apply second order conditions. For a smooth function $L : \mathbf{R} \to \mathbf{R}$ $L''(x) \geq 0$ is a necessary condition to have a minimum and $L''(x) > 0$ is a sufficient condition to have a local minimum.

— Once this is done, one has a family of candidates to optimality and one has to compare by hands their values.

Of course there are specific tests for very special class of functions (as the convexity test).

For an optimal control problem the steps are similar, but we have to bear in mind that we

are looking for minima in an infinite dimensional space (the space of all curves connecting the initial point to the final target) under constraints (the dynamics). The consequence is*Lagrange multipliers* and *anormal extremals* will appear.

Let us recall how to look for the minimum of a function of two variable $L(x_1, x_2)$ under the constraints $f(x_1, x_2) = 0$, with the method of Lagrange multipliers. Here L and f are assumed to be smooth.

The celebrated theorem of the Lagrange multipliers says the following..... `to be added`

5.2 Existence of Optimal Control for initial and final point fixeds and final time fixed

Let us start to discuss the problem of existence of optimal control for the problem

Optimal Control Problem (P1)

$$\begin{aligned} &\dot{x} = F(x,u), \quad x \in \mathbf{R}^n, u \in U \subset \mathbf{R}^m, \\ &x(0) = x_{in}, \quad x(T) = x_{fin} \\ &\int_0^T L(x(t), u(t))\, dt \to \min . \end{aligned}$$

Here F, L are smooth functions of their arguments and the final time T is fixed. We assume that $u(.)$ is a L^∞ function.

To attack the problem of existence let us define a new variable $x^0(t) = \int_0^t L(x(s), u(s))\, ds$ and let us define $\hat{x} = (x^0, x)$. The dynamics of this new variable in $\mathbb{R}^{n+1}$ is given by

$$\begin{aligned} \dot{\hat{x}} = \begin{pmatrix} \dot{x}^0 \\ \dot{x} \end{pmatrix} = \begin{pmatrix} L(x,u) \\ F(x,u) \end{pmatrix} =: \hat{F}(x,u) \\ \hat{x}(0) = (0, x_{in}), \quad \hat{x}(T) = (\text{free}, x_{fin}) \end{aligned}$$

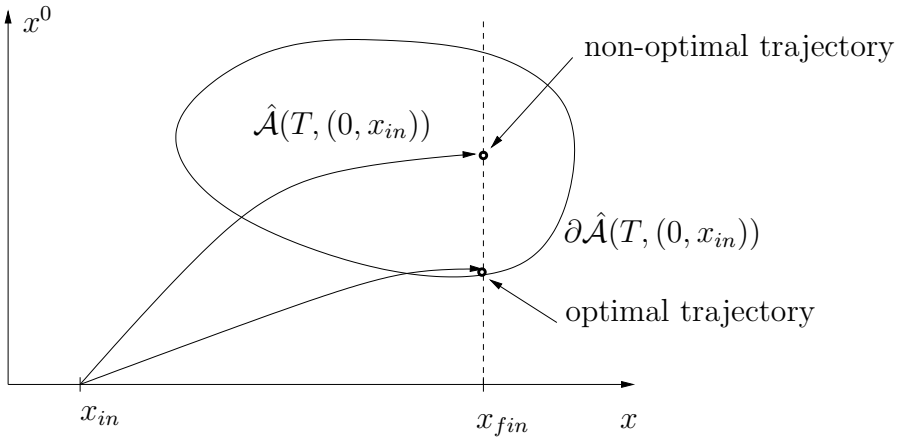
This control system is called the *augmented system*. Let us denote by $\hat{\mathcal{A}}(T, (0, x_{in}))$ its reachable set in time T, starting from $(0, x_{in})$.

Proposition 5.-1. *If $x(.)$ is an optimal trajectory for the problem* **(P1)**, *then $\hat{x}(T) \in \partial\hat{\mathcal{A}}(T, (0, x_{in}))$.*

Preuve.

▷ By contradiction if $\hat{x}(T) \in \text{int}\hat{\mathcal{A}}(T, (0, x_{in}))$ then there exists a trajectory reacing x_{fin} with a smaller cost. See Figure.

□



Corollaire 5.0. *If $\hat{\mathcal{A}}(T,(0,x_{in}))$ is compact then there exists a solution to the problem* **(P1)**.

Hence the compactness of $\hat{\mathcal{A}}(T,(0,x_{in}))$ is very important. A sufficient condition for the compactness of the reachable set is given by the following important theorem of which we omit the proof.

Théorème 5.1 (Filippov)**.** *Consider the control system $\dot{x} = F(x,u)$, $x \in \mathbf{R}^n$, $u \in U \subset \mathbf{R}^m$, where $u(.)$ is a L^∞ function.Assume the following conditions :*

- *the set U is compact,*
- *the set $\mathbf{F}(x) = \left\{F(x,u),\ u \in U\right\}$ is convex for every $x \in \mathbf{R}^n$,*
- *for every $x_0 \in \mathbf{R}^n$, the graphs of all solutions of $\dot{x} = F(x,u)$, $x(0) = x_0$ (in the interval $[0,T]$) are contained in a compact set of $\mathbb{R}^n$.*

Then for every $x_{in} \in \mathbf{R}^n$ and $T > 0$, the sets $\mathcal{A}(T,x_{in})$ and $\mathcal{A}(\leq T,x_{in})$ are compact.

By applying this theorem to the augmented system for the problem **(P1)**, one obtains

Proposition 5.2. *Assume that*

- *$x_{fin} \in \mathcal{A}(T,x_{in})$,*
- *the set U is compact,*
- *the set $\hat{\mathbf{F}}(x) = \{\begin{pmatrix} L(x,u) \\ F(x,u) \end{pmatrix},\ u \in U\}$ is convex for every $x \in \mathbf{R}^n$,*
- *for every $x_0 \in \mathbf{R}^n$, the graphs of all solutions of $\dot{\hat{x}} = \begin{pmatrix} L(x,u) \\ F(x,u) \end{pmatrix}$, $\hat{x}(0) = (0,x_0)$ (in the interval $[0,T]$) are contained in a compact set of $\mathbb{R}^{n+1}$.*

Then there exists a solution to the problem **(P1)**.

When the final time is free, it is more difficult to get the existence of optimal trajectories. However, the conclusion about the set $\mathcal{A}(\leq T,x_{in})$ in the Filippov theorem can be used to find conditions for the existence of optimal controls for minimum time.

Chapitre 6

Minimum time for linear systems

After having considered the problem of existence of optimal control, in this chapter we derive first order necessary condition for a simple, but very important class of problems namely the minimum time problem for linear systems. More precisely, we are going to study the following optimal control problem :

Problem P

$$\begin{aligned} &\dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in U, \quad A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m} \\ &x(0) = x_{in}, \quad x(T) = x_{fin}, \\ &T \to \min \end{aligned} \tag{6.2}$$

Here we assume that the set U is a compact and convex [a] subset of $\mathbb{R}^m$, and that the control $u(.)$ as function of the time belongs to L^∞.

a. Most of the results of this chapter can be obtained without the hypothesis of convexity of U. However, this extension is out of the purpose of these notes.

Later we are going to consider also the case in which the final condition belongs to a smooth target $\mathcal{T}$.

Notice that the quickest trajectories usually will try to use value of controls on the boundary of U. Hence, we expect optimal controls to be non-continuous and optimal trajectories to be non-smooth.

Problem **P** is particularly simple thanks to the linearity. Indeed given a control $u(.) : [0, T] \to U$, the corresponding trajectory satisfying $x(0) = x_{in}$ can be explicitly computed by the formula

$$x(t) = M(t)x_{in} + \int_0^t M(t)M(s)^{-1}Bu(s)ds. \tag{6.3}$$

where $M(t)$ is the solution of $\dot{M} = AM$ with $M(0) =$id (i.e. $M(t) = \exp(At)$).

6.1 Properties of the reachable set and existence of optimal controls

We have the following

Lemme 6.0. *Consider the control system*

$$\dot{x} = Ax + Bu, \quad x \in \mathbb{R}^n, \quad u \in U, \quad A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times m},$$

where U is a compact and convex subset of $\mathbb{R}^m$. For every $x_0 \in \mathbb{R}^n$ and $t > 0$ the reachable set $\mathcal{A}(t, x_0)$ is compact, convex and varies with continuity w.r.t t.

Preuve.

▷ The compactness is consequence of the Filippov Theorem. Indeed U is compact, the set of velocities is convex (since U is convex and the system is linear). The completeness of the systems follows from the linearity.

Let us prove convexity. Let $x_1 = x(t; x_0, u_1(\cdot))$ and $x_2 = x(t; x_0, u_2(\cdot))$. We want to prove that $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{A}(t, x_0)$ for every $\lambda \in [0, 1]$. By using formula (6.3) it follows that

$$\begin{aligned} \lambda x_1 + (1-\lambda)x_2 &= \lambda\left(M(t)x_0 + \int_0^t M(t)M(s)^{-1}Bu_1(s)ds\right) + \\ &\quad (1-\lambda)\left(M(t)x_0 + \int_0^t M(t)M(s)^{-1}Bu_2(s)ds\right) = \\ &\quad M(t)x_0 + \int_0^t M(t)M(s)^{-1}B(\lambda u_1(s) + (1-\lambda u_2(s))ds. \end{aligned}$$

Hence, the point $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{A}(t, x_0)$ is reached with control $\lambda u_1(s) + (1-\lambda u_2(s))$.

The continuity of $\mathcal{A}(t, x_0)$ w.r.t. t is with respect to the Hausdorff distance of sets. Namely if A and B are two subsets of $\mathbb{R}^n$ then $d(A, B) = \inf\{\varepsilon > 0 \mid \forall x_a \in A,\ \exists\, x_b \in B$ such that $\|x_a - x_b\| \leq \varepsilon$ and viceversa$\}$.

The continuity with respect to this distance is a simple consequence of the continuity of the map (6.3), w.r.t. t.

□

As in the previous chapter, applying the Filippov theorem to the augmented systems, we get,

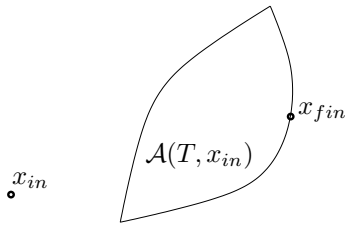
Lemme 6.1. *If $x_{fin} \in \mathcal{A}(x_{in})$ then problem* **P** *admits a solution.*

From Lemma 6.0 we easily have the following,

Corollaire 6.2. *Under the same hypotheses of Lemma 6.0, if $x \in \text{int}(\mathcal{A}(t, x_0))$, then there exists $\delta > 0$ such that for $|t - \bar{t}| < \delta$ we have $x \in \text{int}(\mathcal{A}(\bar{t}, x_0))$.*

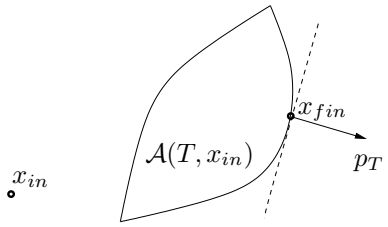
6.2 First order necessary conditions for optimality : the Pontryagin Maximum Principle in the linear case

Let $x_{fin} \in \mathcal{A}(x_{in})$. As a consequence of Lemma 6.1 $x_{fin} \in \mathcal{A}(T, x_{in})$ where T is the minimum time to go from x_{in} to x_{fin}. Thanks to Corollary 6.2, we have that $x_{fin} \in \partial\mathcal{A}(x_{in}, T)$, otherwise T would not be the minimum time. This is the crucial fact to get first order necessary conditions for optimality.

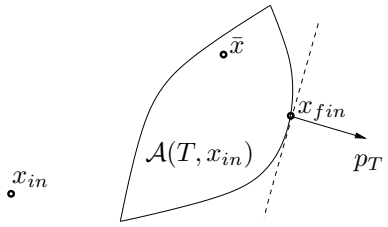


6.2.1 The maximum condition

Since $\mathcal{A}(x_{in}, T)$ is convex, there exists a plane separating x_{fin} and $\mathcal{A}(x_{in}, T)$. Let $\mathbb{R}^{n*} \ni p_T \neq 0$ be the linear form (covector) orthogonal to this plane chosen in such a way that it points to the opposite side w.r.t. $\mathcal{A}(T, x_{in})$.



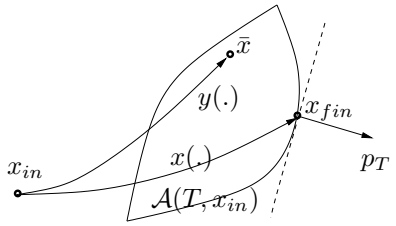
For every $\bar{x} \in \mathcal{A}(T, x_{in},)$ we have $p_T(x_{fin} - \bar{x}) \geq 0$.



Let now

-) $x(.)$ be the trajectory of the control system satisfying $x(0) = x_{in}$, $x(T) = x_{fin}$ and $u_x(.)$ be the corresponding control. This is a solution to problem **P**.

-) $y(.)$ be the trajectory of the control system satisfying $x(0) = x_{in}$, $x(T) = \bar{x}$ and $u_y(.)$ be the corresponding control.



From $p_T(x(T) - y(T)) \geq 0$, we have,

$$p_T\Big(M(T)x_{in} + \int_0^T M(T)M^{-1}(s)Bu_x(s)\,ds - M(T)x_{in} - \int_0^T M(T)M^{-1}(s)Bu_y(s)\,ds\Big) \geq 0.$$

Define now $p(s) := M(T)M^{-1}(s)B$. By direct computation it follows that p satisfies the equation $\dot{p} = -pA$. Equation (6.4) becomes

$$\int_0^T p(s)Bu_x(s)\,ds - \int_0^T p(s)Bu_y(s)\,ds \geq 0. \tag{6.4}$$

From this expression we have,

Lemme 6.3. *$p(s)Bu_x(s) = \max_{v\in U} p(s)Bv$ for a.e. $s \in [0,T]$.*

Preuve.

▷ Let $\bar{u}(s)$ be the control realizing the maximum of the second member and assume by contradiction that $p(s)Bu_x(s) < p(s)B\bar{u}(s)$ on a set $I \subseteq [0,T]$ of posive measure and $p(s)Bu_x(s) = p(s)B\bar{u}(s)$ on $[0,T] \setminus I$. By taking $u_y(s) = \bar{u}(s)$ we would get

$$\int_0^T p(s)Bu_x(s)\,ds - \int_0^T p(s)Bu_y(s)\,ds < 0$$

which contradicts (6.4).

□

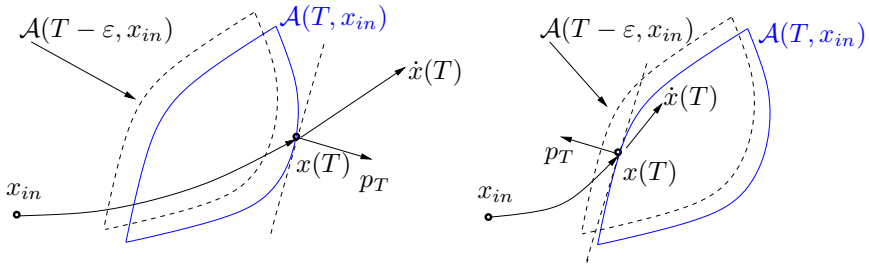
Then we have proved,

Proposition 6.4. *If $(x(.), u(.))$ is a solution to problem* **P** *then there exists $p : [0,T] \to \mathbb{R}^{n*}$ different from zero such that*
-) $\dot{p} = -pA$
-) $p(t)Bu(t) = \max_{v\in U} p(t)Bv$ for a.e. $t \in [0,T]$.

Notice that the equation for $p(.)$ is linear. Hence the condition $p(T) \neq 0$ or $p(t) \neq 0$ for every $t \in [0,T]$ are equivalent.

6.2.2 The relative position of p_T and $\dot{x}(T)$

The condition obtained in the previous section identifies both minimum-time and maximum-time trajectories. Assume that a trajectory $x(.)$, which satisfies the conditions given by Proposition 6.4, has $\dot{x}(T)$ defined. As shown in the following pictures



Locally time-minimal

Locally time-maximal

we have that

— if $p_T\dot{x}(T) < 0$ then $x(T)$ belongs to $\mathcal{A}(T-\varepsilon, x_{in})$ for some $\varepsilon > 0$ i.e. it is time-maximal (locally in time).

— if $p_T\dot{x}(T) > 0$ then $x(T)$ does not belong to $\mathcal{A}(T-\varepsilon, x_{in})$ for any sufficiently small $\varepsilon > 0$ i.e. it is time-minimal (locally in time).

Hence, we get the condition $p_T\dot{x}(T) = p_T(Ax(T) + Bu(T)) \geq 0$. Now since $x(.)$ is time optimal between x_{in} and $x(T)$ then $x(.)\Big|_{[0,t]}$ with $0 < t < T$ is time optimal between x_{in} and $x(t)$ we get :

Proposition 6.5. *With the same notations of Proposition 6.4 we have* $p(t)(Ax(t) + Bu(t)) \geq 0$ *for almost every* $t \in [0, T]$.

Notice that if $p_T\dot{x}(T) = 0$ then the corresponding trajectory is tangent to the boundary of the reachable set and one cannot say if it is either locally time-maximal or locally time-minimal or both, without higher order conditions. These trajectories are called *abnormal extremals* (see below).

6.2.3 The Pontryagin Maximum Principle in Hamiltonian formalism

The contents of the Pontryagin Maximum Principle is given by Propositions and 6.4 and 6.5. Next, we give an equivalent formulation that is much closer to the general one that we will see in the next chapters.

On $\mathbb{R}^n \times \mathbb{R}^{n*} \times \mathbb{R} \times U$ define the function (called *Hamiltonian*),

$$H(x, p, p^0, u) = p(Ax + Bu) + p^0 \tag{6.5}$$

where p^0 is a constant. With this function the equations satisfied by x and p (cfr. Proposition 6.4) can be written in Hamiltonian form as

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

The maximization condition given by proposition 6.4) can be written as $H(x(t), p(t), p^0, u(t)) = \max_{v \in U} H(x(t), p(t), p^0, v)$.

The condition given by Proposition 6.5 can be written as

$$H(x(t), p(t), p^0, u(t)) = 0, \quad \text{with } p^0 \leq 0$$

Finally, the condition $p(t) \neq 0$ can be rewritten as $(p(t), p^0) \neq 0$. Indeed, if we have $p(t) = 0$, equation (6.2.3) would give $p^0 = 0$. Then, we have obtained

Théorème 6.6 (Pontryagin Maximum Principle)**.** *Let $(x(.), u(.))$ be a solution to problem* **P**. *There exists a pair $(p, p^0) \neq (0, 0)$ where $p : [0, T] \to \mathbb{R}^{n*}$ and p^0 is a constant veryfying $p^0 \leq 0$ such that for a.e. $t \in [0, T]$ we have,*

$$\begin{aligned}
&\dot{x}(t) = \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)),\\
&\dot{p} = -\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)).\\
&H(x(t), p(t), p^0, u(t)) = \max_{v \in U} H(x(t), p(t), p^0, v)\\
&H(x(t), p(t), p^0, u(t)) = 0
\end{aligned}$$

where $H : \mathbb{R}^n \times \mathbb{R}^{n*} \times \mathbb{R} \times U$ is defined by $H(x, p, p^0, u) = p(Ax + Bu) + p^0$.

6.2.4 Comments on the Pontryagin Maximum Principle

The form that we have obtained for the Pontryagin Maximum Principle is very close to the one that we will get for a general optimal control problem (with non-linear dynamics $\dot{x} = F(x, u)$, arbitrary cost of the form $\int_0^T L(x(t), u(t))\, dt$ and fixed initial and final points).

Extremals

A trajectory (resp. a pair $(x(.), p(.))$) that is a solution to the equations of the Pontryagin Maximum Principle is called an *extremal trajectory* (resp. and *etremal pair*). Since the Pontryagin maximum Principle is only a necessary condition for optimality, extremal trajectories are trajectories only candidate to be optimal. Moreover, solutions are not

unique : we can have several optimal and non-optimal extremal trajectories going from x_{in} to x_{fin}.

Solutions to the Pontryagin Maximum Principle corresponding to $p^0 = 0$ (resp. $p^0 \neq 0$) are called abnormal extremals (resp. normal extremals). For normal extremals we can always normalize $p_0 = -1$ or any other negative value, since it does not enter inside the equations. Since the equation for p is linear, it follows that if $(x(.), p(.))$ is a solution, then $(x(.), \alpha p(.))$ with $\alpha > 0$ is a solution as well. Then, one can normalize $\|p(0)\| = 1$.

How to use it

The way in which this theorem should be used is the following.

Step 1 Using the maximum condition find "u as function of p".
In the linear case this is not difficult. Take for instance a case in which we have only one control satisfying $u \in [-1, 1]$. If we define the *switching function* as

$$\phi(t) = p(t)B,$$

then the maximization condition gives a.e.

$$u(t) = \begin{cases} 1 & \text{if } \phi(t) > 0 \\ -1 & \text{if } \phi(t) < 0 \end{cases}$$

Notice that since ϕ is an analytic function then either has only isolated zeros or it is always zero. The case in which ϕ is always zero corresponds to when the Kalman condition is not verified. This is the case in which trajectories cannot reach an open set and it is not particularly interesting.

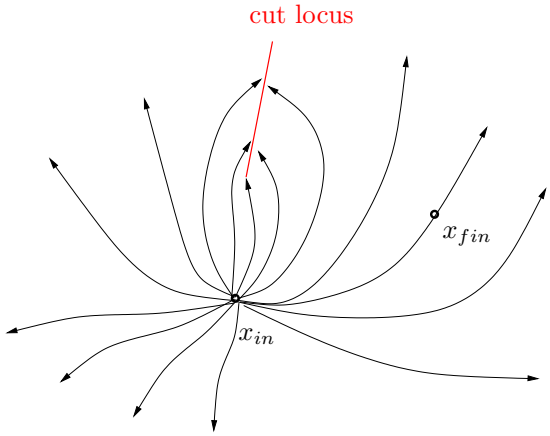
Step 2 By inserting "u as function of p" inside the Hamiltonian equations, find $x(t)$, and $p(t)$. Once that $p(.)$ is known, the control can be obtained with the result of the previous step. Notice that the equation for x and p have boundary conditions given in a nonstandard form. Indeed it is a system of $2n$ first order equations in which we have as boundary conditions $x(0)$ and $x(T)$ and no conditions on p. This, together with the fact that the Pontryagin maximum Principle is only a necessary condition for optimality, suggests that the best way to give a solution to an optimal control problem is an *optimal synthesis*, as discussed in the next paragraph.

6.2.5 Time Optimal Synthesis

Even for the linear case, the main difficulty in an optimal control problem is that the boundary conditions for the Hamiltonian equations are given half at the initial and half at the final point. Moreover, understanding if a trajectory is a global optimum or not is a global problem. In other words, for any initial covector $p(0)$ one has to understand which trajectories reach x_{fin} and find which ones are optimal. Since one is forced to compute all trajectories for all initial covectors, the best way to give a solution to the problem **P** is to forget about the final point x_{fin} and give a time optimal synthesis.

Définition 6.1. A *time optimal synthesis* for the problem **P** is the collection of all time optimal trajectories starting from x_{in}.

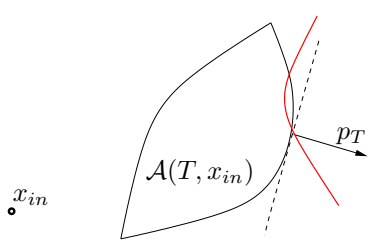
Notice that each trajectory $x(.)$ of a time optimal synthesis is defined on a different time interval $[0, T_{x(.)}]$ (with $T_{x(.)}$ possibly $+\infty$). The collection of all points, in which the trajectories loose optimality, is called the *cut locus*. The trajectories of a time optimal synthesis are parameterized by the initial covector. An example of time optimal synthesis is given below



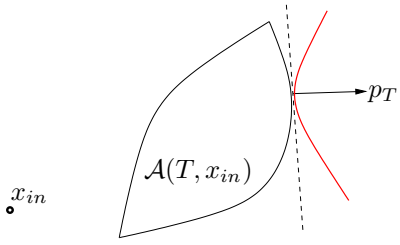
An example of optimal synthesis

6.2.6 The case of a smooth target

If the final condition $x(T) = x_{fin}$ in the problem **P** is changed in $x_{fin} \in \mathcal{T}$ where $\mathcal{T}$ (called *target*) is a smooth closed manifold, then one has the additional condition that $p(T)$ should be orthogonal to the tangent space of $\mathcal{T}$ at the arriving point. This is called *transversality condition*. It is clear why such a condition should hold. As shown by the following picture, if it is not satisfied then the target enters the interior of $\mathcal{A}(T, x_{in})$ and could be reached before.



Transversality conditions
that are not satisfied



Transversality conditions
that are satisfied

Moreover if the target separate the space in some internal and external part of the target then the covector should "enter" the target. Similar conditions should hold if the initial condition is changed in $x_{in} \in \mathcal{S}$ where $\mathcal{S}$, called the *source*, is a smooth closed manifold.

Notice that independently on the dimension of the source and of the target the number of boundary conditions that are given to the Hamiltonian equations is always $2n$.

These conditions can be easily generalized to the case of a convex closed target (possibly non-smooth).

Chapitre 7

Minimal energy for control affine systems

In this chapter we study a very important class of nonlinear optimal control problems, namely *control affine systems with quadratic cost*.

Problem P-AQ

$$\dot{x} = F_0(x) + \sum_{i=1}^{m} u_i F_i(x), \quad x \in \mathbb{R}^n, \quad u_i \in \mathbb{R}, \tag{7.3}$$

$$x(0) = x_{in}, \quad x(T) = x_{fin},$$

$$\int_0^T \sum_{i=1}^{m} u_i(t)^2 \, dt \to \min \tag{7.4}$$

Here we assume that the vector fields F_i, $i = 0, 1, \ldots, n$ are smooth and that the control $u(.)$, as function of the time, belongs to L^∞.

This problem is important since the control affine form (7.3) models most of the systems that one can find in applications and a cost of the type (7.4) represents the *energy* given by the controls to the systems.

Définition 7.1. When $F_0 = 0$ and the problem **P** is called *sub-Riemannian*.

7.0.7 Existence

The problem of existence of optimal controls for the problem **P-AQ** is difficult and has to be studied case by case. Indeed Filippov Theorem cannot be applied since the set of controls is not bounded. However in the sub-Riemannian case one can get easily the following result.

Théorème 7.0. *If $F_0 = 0$ and if the control system is complete*[1]*, then there exists a solution to the minimization problem.*

The proof is given in the PC.

7.1 The Pontryagin Maximum Principle for control affine systems with quadratic cost

In this section we prove the following Theorem

Théorème 7.1 (Pontryagin Maximum Principle for **P-AQ**)**.** *Let $(x(.), u(.))$ be a solution to the problem* **P-AQ**. *There exists a pair $(p, p^0) \neq (0, 0)$ where $p : [0, T] \to \mathbb{R}^{n*}$ and p^0 is a constant verifying $p^0 \leq 0$ such that for a.e. $t \in [0, T]$ we have,*

$$\begin{aligned}
\dot{x}(t) &= \frac{\partial H}{\partial p}(x(t), p(t), p^0, u(t)),\\
\dot{p} &= -\frac{\partial H}{\partial x}(x(t), p(t), p^0, u(t)).\\
&\frac{\partial H}{\partial u}(x(t), p(t), p^0, u(t)) = 0\\
&H(x(t), p(t), p^0, u(t)) = const(T)
\end{aligned}$$

where $H : \mathbb{R}^n \times \mathbb{R}^{n*} \times \mathbb{R} \times \mathbb{R}^m$ is defined by $H(x, p, p^0, u) = p(F_0(x) + \sum_{i=1}^m u_i F_i(x)) + p^0 \sum_{i=1}^m u_i(t)^2$.

This form of the PMP is very similar to the one we got in the previous chapter for the minimum time problems for linear systems. The only differences are the following :

— $H(x(t), p(t), p^0, u(t))$ is not required to be zero on every optimal trajectory. This is due to the fact that the final time is fixed for the problem **P-AQ**.

— The maximum condition is now written as $\frac{\partial H}{\partial u} = 0$. This condition is equivalent to the maximum condition since H is quadratic in the controls.

A trajectory (resp. a pair $(x(.), p(.))$) that is a solution to the equations of the Pontryagin Maximum Principle is called an *extremal trajectory* (resp. and *etremal pair*). Since the Pontryagin maximum Principle is only a necessary condition for optimality, extremal trajectories are trajectories only candidate to be optimal. Moreover, solutions are not unique : we can have several optimal and non-optimal extremal trajectories going from x_{in} to x_{fin}.

1. We recall that the control system (7.3) is said to be complete if for every $T > 0$, $x_0 \in \mathbb{R}^n$ and $u(.) \in L^\infty([0, T])$, all solutions of (7.3), in the interval [0,T], with $x(0) = x_0$ are contained in a compact set of $\mathbb{R}^n$.

Solutions to the Pontryagin Maximum Principle corresponding to $p^0 = 0$ (resp. $p^0 \neq 0$) are called abnormal extremals (resp. normal extremals). For normal extremals we can always normalize $p_0 = -1/2$ or any other negative value.

7.2 Proof of the PMP

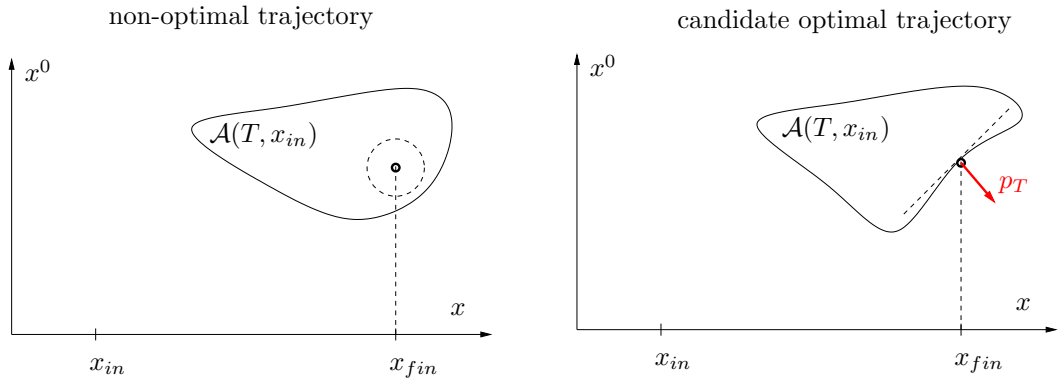
The idea of the proof is the following. We add a new variable $x^0(t) = \int_0^T \sum_{i=1}^m u(s)\, ds$ and we consider the dynamics for $\hat{x} = (x^0, x)^T$:

$$\dot{\hat{x}} = \begin{pmatrix} \dot{x}^0 \\ \dot{x} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m u_i^2 \\ F_0(x) + \sum_{i=1}^m u_i F_i(x) \end{pmatrix}$$

We define the *end-point mapping* as the map that to a control associates the end point of the corresponding trajectory starting from x_{in} :

$$L^\infty([0,T]) \ni u(.) = (u_1(.), \ldots, u_n(.)) \stackrel{End}{\longmapsto} \begin{pmatrix} x^0(T) \\ x(T) \end{pmatrix} \in \mathbb{R}^{n+1}$$

On an optimal control $\tilde{u}(.)$ this map cannot be a surjection (i.e. its differential cannot be invertible). Otherwise the image of a ball in L^∞ around the optimal control, would contain an open set in $\mathbb{R}^{n+1}$ and one could reach the same final point $x(T)$ with a smaller cost. Then it should exists a covector $\hat{p}_T \neq 0$ which is orthogonal to $\mathrm{Im}(D_{\tilde{u}}(End))$.



To avoid the difficulties of making the differential of a map on an infinite dimensional space, in the following we will make only variations in a finite dimensional set of controls. Moreover the condition $\hat{p}_T \neq 0$ orthogonal to $\mathrm{Im}(D_{\tilde{u}}(End))$ will be "transported back" the initial point. This permits to get an equation for the covector.

7.2.1 Notation

In the following, for simplicity of notation, we assume that the control system is complete. If it is not complete, the proof can be made similarly. We use the following notation.

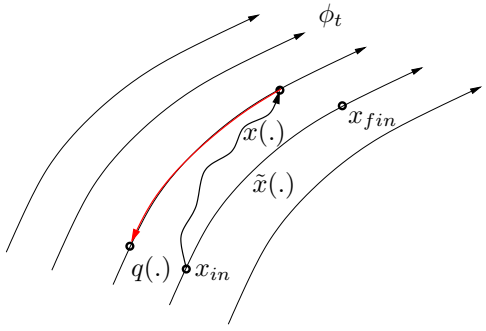
— We set $f_u(x) := F_0(x) + \sum_{i=1}^m u_i F_i(x)$
— We call $\tilde{u}(.)$ the optimal control and $\tilde{x}(.)$ the optimal trajectory starting from x_{in}.
— We call ϕ_t the *optimal flow* i.e. the flow associated to the differential equation $\dot{x} = f_{\tilde{u}(t)}(x)$. Notice that
  -) $\tilde{x}(t) = \phi_t(x_{in})$, but in general $\phi_t(\bar{x})$ is not an optimal trajectory for $\bar{x} \neq x_{in}$.
  -) $\frac{d}{dt}\phi_t = f_{\tilde{u}(t)} \circ \phi_t$
  -) Since the control system is complete, ϕ_t is a diffeomorphism.

7.2.2 The Variation

Let $x(.)$ be the trajectory starting from x_{in} and corresponding to a control $u(.) = \tilde{u}(.) + v(.)$. Let us define

$$q(t) = \phi_t^{-1}(x(t)).$$

This is the trajectory corresponding to the control $u(.)$ brought back with the optimal flow. Notice that $q(0) = x(0) = x_{in}$. Moreover if $v(.) \equiv 0$ then $q(.) \equiv x_{in}$.



Let us look for an equation for $q(t)$. Differentiating $x(t) = \phi_t(q(t))$, we get

$$\dot{x}(t) = \frac{d}{dt}\phi_t\Big|_{q(t)} + \frac{\partial \phi_t}{\partial x}\Big|_{q(t)} \dot{q}(t) = f_{\tilde{u}(t)}(\phi_t(q(t))) + \frac{\partial \phi_t}{\partial x}\Big|_{q(t)} \dot{q}(t)$$

Now $\dot{x}(t) = f_{u(t)}(x(t)) = f_{u(t)}(\phi_t(q(t)))$. Hence

$$\begin{aligned} \dot{q}(t) &= \left[\frac{\partial \phi_t}{\partial x}\Big|_{q(t)}\right]^{-1} \Big(f_{u(t)}(\phi_t(q(t))) - f_{\tilde{u}(t)}(\phi_t(q(t)))\Big) = \\ &\left[\frac{\partial \phi_t}{\partial x}\Big|_{q(t)}\right]^{-1} \sum_{i=1}^m v_i(t) F_i(\phi_t(q(t))) =: g_{v(t)}(q(t)) \end{aligned} \tag{7.5}$$

Notice that if we set $v \equiv 0$ in the Cauchy problem

$$\begin{cases} \dot{q} = g_{v(t)}(q) \\ q(0) = x_{in} \end{cases}$$

then $g_v \equiv 0$ and $q(.) \equiv x_{in}$. Notice moreover that $g_{sv} = s g_v$, for every $s \in \mathbb{R}$.

7.2.3 The crucial Lemma

Fix $v(.)$ and consider the map

$$s \mapsto \begin{pmatrix} x^0(T, \tilde{u}(.) + sv(.)) \\ q(T, \tilde{u}(.) + sv(.)) \end{pmatrix} \text{ starting from } \begin{pmatrix} 0 \\ x_{in} \end{pmatrix}$$

Lemme 7.2. *If $(\tilde{x}(.), \tilde{u}(.))$ is a solution to the problem* **P-AQ** *then there exists $\hat{p} \in (\mathbf{R}^{n+1})^*$, $\hat{p} \neq 0$, such that*

$$\hat{p}\Big(\frac{\partial x^0(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0}, \frac{\partial q(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0}\Big)^T = 0 \text{ for every } v(.).$$

Preuve.

▷ By contradiction there exists $n+1$ controls $v_0(.), v_1(.), \ldots, v_n(.)$ such that

$$\begin{pmatrix} \frac{\partial x^0(T,\tilde{u}(.)+sv_0(.))}{\partial s}|_{s=0} \\ \frac{\partial q(T,\tilde{u}(.)+sv_0(.))}{\partial s}|_{s=0} \end{pmatrix}, \ldots, \begin{pmatrix} \frac{\partial x^0(T,\tilde{u}(.)+sv_n(.))}{\partial s}|_{s=0} \\ \frac{\partial q(T,\tilde{u}(.)+sv_n(.))}{\partial s}|_{s=0} \end{pmatrix} \tag{7.6}$$

are linearly independent. It follows that the map

$$(s_0, \ldots, s_n) \mapsto \begin{pmatrix} x^0(T, \tilde{u}(.) + \sum_{i=0}^m s_i v_i(.)) \\ q(T, \tilde{u}(.) + \sum_{i=0}^m s_i v_i(.)) \end{pmatrix} \tag{7.7}$$

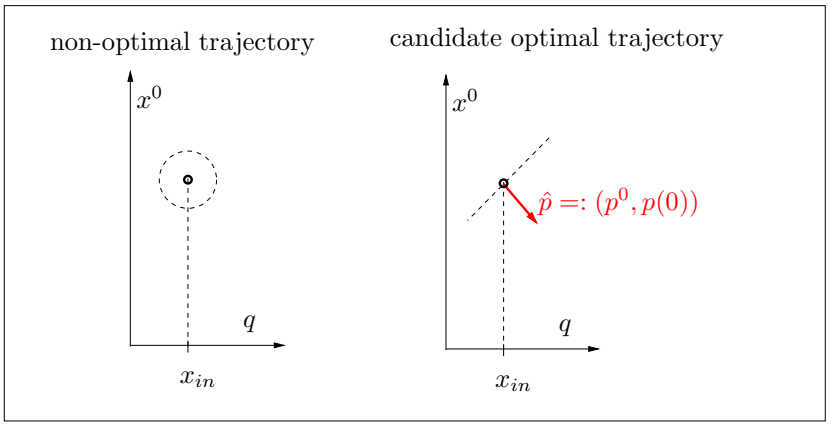
is a local diffeomorphism in a neighborhood of the origin of $\mathbf{R}^{n+1}$ (indeed the vectors (7.6) are the components of the differential of the map (7.7)). Notice that image of the origin through the map (7.7) is,

$$\begin{pmatrix} x^0(T, \tilde{u}(.)) \\ q(T, \tilde{u}(.)) \end{pmatrix} = \begin{pmatrix} x^0(T, \tilde{u}(.)) \\ x_{in} \end{pmatrix}.$$

But this is not possible, otherwise there exists $\tilde{v}(.)$ such that

$$x^0(T, \tilde{u}(.) + \tilde{v}(.)) < x^0(T, \tilde{u}(.))$$
$$x(T, \tilde{u}(.) + \tilde{v}(.)) = \phi_T(q(T, \tilde{u}(.) + \tilde{v}(.))) = \phi_T(q(T, \tilde{u}(.)) = \phi_T(x_{in}) = x_{in},$$

and $(\tilde{x}(.), \tilde{u}(.))$ would not be optimal.



□

Setting $\hat{p} = (p^0, p(0))$ we have that

$$p^0 \frac{\partial x^0(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0} + p(0)\frac{\partial q(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0} = 0 \text{ for every } v(.) \qquad (7.8)$$

Let us compute the different terms in the sum.

$$\begin{aligned}
\frac{\partial x^0(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0} &= \frac{\partial}{\partial s}\Big|_{s=0} \int_0^T (\tilde{u}(t) + sv(t))(\tilde{u}(t) + sv(t))\, dt \\
&= 2\int_0^T \sum_i \tilde{u}_i(s) v_i(t)\, dt \\
\frac{\partial q(T, \tilde{u}(.) + sv(.))}{\partial s}\Big|_{s=0} &= \frac{\partial}{\partial s}\Big|_{s=0} \int_0^T g_{sv(t)}(q(t, \tilde{u}(.) + sv(.)))\, dt \\
&= \int_0^T g_{v(t)}(q(t, \tilde{u}(.)))\, dt = \int_0^T \sum_{i=1}^m v_i D\phi^{-1}(x_{in}) F_i(\tilde{x}(t))\, dt
\end{aligned}$$

where we have used the notation $D\phi_t$ in place of $\frac{\partial \phi_t}{\partial x}$, the fact that $q(t, \tilde{u}(.)) = x_{in}$ and that $\phi_t(q(t, \tilde{u}(.))) = \tilde{x}(t)$. From (7.8) it follows that

$$\int_0^T \sum_{i=1}^m v_i \Big(p^0 2\tilde{u}_i(t) + p(0) D\phi^{-1}(x_{in}) F_i(\tilde{x}(t)) \Big)\, dt = 0$$

Defining $p(t) = p(0) D\phi^{-1}(x_{in}) F_i(\tilde{x}(t))$ and since $v_i(.)$ are arbitrary it follows that

$$p^0 2\tilde{u}_i(t) + p(t) F_i(\tilde{x}(t)) = 0 \text{ for almost every } t \in [0, T].$$

Now

— if $p^0 \neq 0$ we can normalize $p^0 = -1/2$ and we get $\tilde{u}(t) = p(t) F_i(\tilde{x}(t))$, for a.e. $t \in [0, T]$;

— if $p^0 = 0$ we get $p(t) F_i(\tilde{x}(t))$, for a.e. $t \in [0, T]$.

We have then proved the following

Proposition 7.3. *If $(\tilde{x}(.), \tilde{u}(.))$ is a solution to the problem* **P** *then there exists $p(.) : [0, T] \to (\mathbf{R}^n)^*$, such that*

-) $p(t) = p(0) D\phi_t^{-1}(x_{in})$, where ϕ_t is the flow corresponding to the optimal control $\tilde{u}(.)$;

-) $\tilde{u}_i(t) = p(t) F_i(\tilde{x}(t))$ or $p(t) F_i(\tilde{x}(t)) = 0$ for a.e. $t \in [0, T]$.

7.2.4 The Hamiltonian form

It remains to show that proposition 7.3 is equivalent to the PMP (with the change of notation $(x, u) \to (\tilde{x}, \tilde{u})$).

— The first equation of the PMP is equivalent to $\dot{\tilde{x}} = F_0(\tilde{x}(t)) + \sum_{i=1}^{m} \tilde{u}_i(t) F_i(\tilde{x}(t))$.

— The second equation of the PMP is equivalent to $\dot{p}(t) = -p(t)(DF_0 + \sum_{i=1}^{m} \tilde{u}_i(t) DF_i(\tilde{x}(t))$. A simple computation show that this is equivalent to $p(t) = p(0) D\phi_t^{-1}(x_{in})$.

— The third equation gives immediately that $\tilde{u}_i(t) = p(t) F_i(\tilde{x}(t))$ or $p(t) F_i(\tilde{x}(t)) = 0$ for a.e. $t \in [0, T]$.

Chapitre 8

Linear Quadratic Theory

In this chapter we study linear control systems with a quadratic cost. These systems are very important for practical applications, as we will see in Section 8.4. Indeed a quadratic cost is often very natural when one would like to minimize the error with respect to a reference trajectory (tracking problem). Moreover, even if control systems are in general nonlinear, one often linearize the system along a reference trajectory are in a neighborhood of a point (as for instance in a stabilization problem). We are then considering the linear control system in $\mathbb{R}^n$:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x_0, \tag{8.1}$$

with a quadratic cost of the type :

$$C(u) = {}^t x(T)Qx(T) + \int_0^T ({}^t x(t)W(t)x(t) + {}^t u(t)U(t)u(t))dt, \tag{8.2}$$

where $T > 0$ is fixed, for every t, $U(t) \in \mathcal{M}_m(\mathbb{R})$ is symmetric, positive definite, $W(t) \in \mathcal{M}_n(\mathbb{R})$ is symmetric positive and $Q \in \mathcal{M}_n(\mathbb{R})$ is a symmetric positive matrix. We assume that the dependence on t of A, B, W and U is L^∞ on $[0, T]$. Since the cost is quadratic, the natural functional space for the controls is $L^2([0, T], \mathbb{R}^m)$.

The problem of optimal control is then the following (called *LQ-problem*)

Problème LQ : Fix $x_0 \in \mathbb{R}^n$. Find the trajectory starting from x_0 that minimize the cost $C(u)$.

Notice that we do not impose any constraint on the final position. In the following we set :

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

in such a way that

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

The matrices Q, W, U are called *weight matrices*.

Remarque. By hypothesis the matrices Q and $W(t)$ are symmetric positive, but not necessarily definite. For instance if $Q = 0$ and $W = 0$ then the cost is always minimal for the control $u = 0$.

8.1 Existence of optimal controls

Let us introduce the following hypothesis of coercivity on U :

$$\exists \alpha > 0 \mid \forall u \in L^2([0,T],\mathbb{R}^m) \quad \int_0^T \|u(t)\|_U^2 dt \geq \alpha \int_0^T {}^t u(t)u(t)dt. \tag{8.3}$$

For instance this hypothesis is verified if the application $t \mapsto U(t)$ is continuous on $[0,T]$ and $T < +\infty$, or if there exists a constant $c > 0$ such that for all $t \in [0,T]$ and all vectors $v \in \mathbb{R}^m$ we have ${}^t vU(t)v \geq c\,{}^t vv$.

We have the following existence theorem :

Théorème 8.1. *Under the hypothesis (8.3), there exists a unique minimizing trajectory for the LQ-problem.*

Preuve.

▷ Let us first show the existence of such a trajectory. Consider a minimizing sequence $(u_n)_{n\in\mathbb{N}}$ of controls on $[0,T]$, i.e. the sequence $C(u_n)$ converge to the infimum of the cost. In particular this sequence is bounded. By hypothesis there exists a constant $\alpha > 0$ such that for every $u \in L^2([0,T],\mathbb{R}^m)$ we have $C(u) \geq \alpha \|u\|_{L^2}$. It follows that the sequence $(u_n)_{n\in\mathbb{N}}$ is bounded in $L^2([0,T],\mathbb{R}^m)$. As a consequence, up to a subsequence it converge weakly to a control u of L^2. Let x_n (resp. x) be the trajectory associated to the control u_n (resp. u) son $[0,T]$. Thanks to the formula of variation of constants, we have for every $t \in [0,T]$:

$$x_n(t) = M(t)x_0 + M(t)\int_0^t M(s)^{-1}B(s)u_n(s)ds \tag{8.4}$$

(and a similar formula for $x(t)$).

We have then easily that, up to a subsequence, the sequence (x_n) converge to x on $[0,T]$ (indeed one can also show that this convergence is uniform).

Passing to the limit in (8.4), we get for all $t \in [0,T]$:

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

and then x is a solution of the system associated to the control u. Let us show that it is minimizing. To this purpose we use the fact that since $u_n \rightharpoonup u$ in L^2, we have the inequality :

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

and then $C(u) \leq \liminf C(u_n)$. Since (u_n) is a minimizing sequence, $C(u)$ is equal to the lower bound of the cost, i.e. the control u is minimizing. This shows the existence of an optimal trajectory.

For the uniqueness we need the following Lemma

Lemme 8.2. *The function C is strictly convex.*

Preuve.

▷ [Preuve du lemme]

Let us first remark that for every $t \in [0,T]$, the function $f(u) = {}^t uU(t)u$ defined on $\mathbb{R}^m$ is strictly convex since by hypothesis the matrix $U(t)$ is symmetric positive. Then let $x_u(\cdot)$ be the trajectory associated to a control u. We hav for every $t \in [0,T]$:

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

As a consequence, the application that to a control u associates $x_u(t)$ is convex for every $t \in [0,T]$ (why ?). Now since the matrix $W(t)$ is symmetric positive, then the application $u \mapsto {}^t x(t)W(t)w(t)$ is convex. The same reasoning applies to the term ${}^t x(T)Qx(T)$. Finally since the integration respect the convexity, we have that the cost is strictly convex in u.

□

The uniqueness of the optimal trajectory follows immediately.

□

Remarque (The case $T = \infty$). The theorem still holds if $T = +\infty$, avec $g = 0$, under the hypothesis that (8.1) is controllable (in every time).

Proposition 8.3. *Consider the problem of finding a trajectory solution of*

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

on $[0, +\infty[$ and minimizing

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

If the system is controllable in time $T > 0$, and if the hypothesis (8.3) is satisfied $[0, +\infty[$, then there exists a unique minimizing trajectory.

8.2 Necessary and sufficient condition for optimality in the LQ case

Théorème 8.4. *The trajectory x, associated to a control u, is optimal for the LQ problem if and only if there exists and adjoint vector $p(t)$ satisfying for almost $t \in [0,T]$:*

$$\dot{p}(t) = -p(t)A(t) + {}^t x(t)W(t) \tag{8.8}$$

and the final condition

$$p(T) = -{}^t x(T)Q. \tag{8.9}$$

Moreover the optimal control u can be written for almost every $t \in [0,T]$:

$$u(t) = U(t)^{-1}\,{}^t B(t)\,{}^t p(t). \tag{8.10}$$

Preuve.

▷ Let u, defined on $[0,T]$, be an optimal control and x the corresponding trajectory. The cost is minimal among all trajectories of the system starting from x_0 (the final point is not fixed). Let us consider the variations of the control u in $L^2([0,T],\mathbb{R}^m)$:

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

The corresponding trajectory is

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

with $\delta x(0) = 0$. The trajectory x_{pert} is solution of $\dot{x}_{pert} = Ax_{pert} + Bu_{pert}$. Hence

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

and as a consequence we have for every $t \in [0,T]$:

$$\delta x(t) = M(t)\int_0^t M(s)^{-1}B(s)\delta u(s)ds. \tag{8.11}$$

Now the cost $C(\cdot)$ is a smooth function on $L^2([0,T],\mathbb{R}^m)$ (it is even analitic) in the sense of Fréchet. Since the control u is a minimizer, we have :

$$dC(u) = 0.$$

Now

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

and, since Q, $W(t)$ and $U(t)$ are symmetric, we have :

$$\frac{1}{2}dC(u).\delta u = {}^t x(T)Q\delta x(T) + \int_0^T ({}^t x(t)W(t)\delta x(t) + {}^t u(t)U(t)\delta u(t))dt = 0, \tag{8.12}$$

that should hold for every δu. This equation will give the expression of the optimal control u.

Let us introduce the covector $p(t)$ as the solution of the following Cauchy problem :

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

The formula of variation of constants gives :

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

for every $t \in [0,T]$, where :

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

Let us come back to the equation (8.12). Using (8.11) and integrating by parts we get,

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

Now

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

and using the expression of Λ we arrive to :

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

Injecting this equality in (8.12) and taking into account that,

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

we get :

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

this for every application $\delta u \in L^2([0,T],\mathbb{R}^m)$. This implies the equality (for almost every $t \in [0,T]$) :

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

that is the desired conclusion.

To prove the opposite, if there exists a covector $p(t)$ verifying (8.8) and (8.9) and if the control u is given by (8.10), then using the reasoning above we have that

$$dC(u) = 0.$$

Now since C is strictly convex, this implies that u is a global minimum of C.

□

Remarque. In the case in which $T = +\infty$, the condition become :

$$\lim_{t \to +\infty} p(t) = 0. \tag{8.13}$$

Remarque. Let us define the function $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ by :

$$H(x, p, u) = p(Ax + Bu) - \frac{1}{2}({}^t xWx + {}^t uUu),$$

where we use the convention that p is a row vector of $\mathbb{R}^n$. Then the equations given by the Pontryagin Maximum Principle in the linear quadratic case become :

$$\dot{x} = \frac{\partial H}{\partial p} = Ax + Bu,$$
$$\dot{p} = -\frac{\partial H}{\partial x} = -pA + {}^t xW,$$

and

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

since $pB - {}^t uU = 0$.

Notice that in the LQ case, the Pontryagin Maximum Principle is a necessary and sufficient condition for optimality. Moreover it is possible to write the control in the form of a feedbackthanks to the Riccati theory as it is explained in the next section.

8.3 Value function and Riccati equation

8.3.1 Definition of the value function

Let $T > 0$ be fixed and let $x \in \mathbb{R}^n$. Consider the LQ problem LQ of finding a trajectory solution of

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(0) = x, \tag{8.14}$$

minimizing the cost

$$C_T(u) = {}^t x(T)Qx(T) + \int_0^T (\|x(t)\|_W^2 + \|u(t)\|_U^2)dt. \tag{8.15}$$

Définition 8.1. The *value function* S_T at point x is the minimum of the cost for the LQ problem. In other words :

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

8.3.2 Riccati equation

Théorème 8.5. *Under the hypothesis (8.3), for every $x \in \mathbb{R}^n$ there exists a unique optimal trajectory x associated to a control u for the problem (8.14), (8.15). The optimal control can be written in the form of a feedback*

$$u(t) = U(t)^{-1}\,{}^tB(t)E(t)x(t), \tag{8.19}$$

where $E(t) \in \mathcal{M}_n(\mathbb{R})$ is the solution on $[0,T]$ of the matrix Riccati equation :

$$\dot{E}(t) = W(t) - {}^tA(t)E(t) - E(t)A(t) - E(t)B(t)U(t)^{-1}{}^tB(t)E(t),\ E(T) = -Q. \tag{8.20}$$

Moreover for every $t \in [0,T]$, the matrix $E(t)$ is symmetric and :

$$S_T(x) = -{}^txE(0)x. \tag{8.21}$$

Remarque. In particular the theorem says that the optimal control u can be written in the form of feedbacck :

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

où $K(t) = U(t)^{-1}\,{}^tB(t)E(t)$. This form is very useful for problems of stabilization as we wil see later.

Preuve.

▷ From Theorem 8.1, there exists a unique optimal trajectory that, thanks to Theorem il existe une unique trajectoire optimale qui, 8.4, is characterized by :

$$\begin{aligned}\dot{x} &= Ax + BU^{-1}\,{}^tB\,{}^tp,\\ \dot{p} &= -pA + {}^txW,\end{aligned}$$

with $x(0) = x$ and $p(T) = -{}^tx(T)Q$. Moreover the control is written as :

$$u = U^{-1}\,{}^tB\,{}^tp.$$

We have then to show that we can write $p(t) = {}^tx(t)E(t)$, where $E(t)$ is the solution of (8.20). Notice that if p is written like that, then thanks to the equation satisfied by (x,p), we find easily that $E(t)$ should satisfy (8.20). Using the uniqueness of the optimal trajectory , weare going to prove that p can be written like that. Let $E(t)$ the solution to the equation :

$$\dot{E} = W - {}^tAE - EA - EBU^{-1}\,{}^tBE,\ E(T) = -Q.$$

First $E(t)$ is symmetric since the second member of the equation is symmetric as well and the matrix Q is symmetric. A priori we do not know that the solution is well defined on $[0,T]$. We will show that later (Lemma 8.6).

Set $p_1(t) = {}^t x_1(t)E(t)$, where x_1 is solution of

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

et $u_1 = U^{-1}{}^t BEx_1$. On a alors :

$$\begin{aligned}\dot{p}_1 &= {}^t\dot{x}_1 E + {}^t x_1 \dot{E} \\ &= {}^t(Ax_1 + BU^{-1}{}^t BEx_1)E + {}^t x_1(W - {}^t AE - EA - EBU^{-1}{}^t BE) \\ &= -p_1 A + {}^t x_1 W.\end{aligned}$$

In other words the triplet (x_1, p_1, u_1) verifies exactly the equations of Theorem8.4. As a consequence the trajectory x_1 is optimal and by uniqueness we get $x_1 = x$, $u_1 = u$, and $p_1 = p$. In particular we have $p = {}^t xE$, and $u = U^{-1}{}^t BEx$. Let us prove formula (8.21). Let us compute along the trajectory $x(t)$:

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

From the expression of u we get :

$${}^t uUu = {}^t(U^{-1}{}^t BEx)UU^{-1}{}^t BEx = {}^t xEBU^{-1}{}^t BEx = pBu.$$

Finally we get :

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

et par conséquent :

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

Now since $E(T) = -Q$ and $x(0) = x$, we get $S_T(x) = -{}^t xE(0)x$.

Lemme 8.6. *The application $t \mapsto E(t)$ is well defined on $[0, T]$.*

Preuve.

▷ [Preuve du lemme] If the application $E(t)$ is not defined on $[0, T]$, then there exists $0 < t_* < T$ such that $\|E(t)\|$ tends to $+\infty$ as t tends to t_*. In particular for every $\alpha > 0$ there exists $t_0 \in]t_*, T]$ et $x_0 \in \mathbb{R}^n$, with $\|x_0\| = 1$, such that

$$|{}^t x_0 E(t_0)x_0| \geq \alpha. \tag{8.22}$$

Thanks to Theorem 8.1, it exists a unique $x(\cdot)$ for the LQ problem on $[t_0, T]$, such that $x(t_0) = x_0$ (see remark **??**). This trajectory is characterized by the system of equations :

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

Thanks to the theorem of continuous dependence of the solutions of a differential equation w.r.t the initial condition we get that $x(T)$ corresponding to trajectories starting at t_0 from x_0, are uniformely bounded when $0 \leq t_0 < T$ et $\|x_0\| = 1$. Then the solutions $x(t), p(t)$ of the previous system of differential equation are uniformly bounded on $[0, T]$. In particular the quantity $p(t_0)x(t_0)$ has to be bounded independnetly from t_0. Now we know that $p(t) = {}^t x(t)E(t)$, then :

$$p(t_0)x(t_0) = {}^t x_0 E(t_0)x_0,$$

and we get a contradiction with (8.22).

□

The theorem is proved

□

Remarque. Thanks to (8.21) it is clear that the matrix $E(0)$ is symmetric negative. We can get a better result if the matrix Q is definite.

Lemme 8.7. *If Q is symmetric positive definite, or if for every $t \in [0, T]$ the matrix $W(t)$ is symmetric positive definite, then the matrix $E(0)$ is symmetric definite negative.*

8.3.3 Linear representation of the Riccati equation

We have the following property

Proposition 8.8. *Consider the framework of Theorem 8.5. Let*

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

the resolvent of the linear system

$$\begin{aligned} \dot{x} &= Ax + BU^{-1}{}^t B {}^t p, \\ {}^t\dot{p} &= -{}^t A {}^t p + Wx, \end{aligned}$$

such that $R(T) = Id$. Then for every $t \in [0, T]$ we have :

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

Preuve.

▷ By definition of the resolvent we have

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

Now we know that ${}^t\!p(T) = -Qx(T)$, hence :

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

We conclude by remarking that ${}^t\!p(t) = E(t)x(t)$. Notice that the matrix $R_1(t) - R_2(t)Q$ is invertible on $[0, T]$ since the LQ problem is well posed as we saw above.

□

As a consequence, to solve the (8.20) equation it is sufficient to integrate a system of linear equation, which is easy to do numerically This method (due to Kalman-Englar) is better than the direct method in the stationary case.

8.4 Applications of the LQ theory

8.4.1 Tracking problem

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

Consider the linear control system (perturbed) :

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + r(t),\ x(0) = x_0, \tag{8.23}$$

and let $\xi(t)$ a trajectory of $\mathbb{R}^n$ on $[0, T]$, starting from a point ξ_0 (and that it is not necessarily a solution to (8.23)). The goal is to find a control such that the corresponding trajectory, solution to (8.23), follows as better as possible the reference trajectory $\xi(t)$.

We introduce the error on $[0, T]$:

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

which is solution of the control system

$$\dot{z}(t) = A(t)z(t) + B(t)u(t) + r_1(t),\ z(0) = z_0, \tag{8.24}$$

where $z_0 = x_0 - \xi_0$ et $r_1(t) = A(t)\xi(t) - \dot{\xi}(t) + r(t)$. It is then reasonable to minimize the cost :

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

where Q, W, U are "weight" matrices. To eliminate the perturbation r_1, we augment the system of one dimension by adding :

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

in such a way that we are left to minimize the cost :

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

for the system

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

starting from $z_1(0)$.

The LQ theory says that the optimal control exists, it is unique and is written as :

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

where $E_1(t)$ is solution to the Riccati equation :

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

Let us put

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

Substituting in the previous equation we get the equations :

$$\begin{array}{lll} \dot{E} & = W - {}^t A E - E A - E B U^{-1}\, {}^t B E, & E(T) = -Q, \\ \dot{h} & = -{}^t A h - E r_1 - E B U^{-1}\, {}^t B h, & h(T) = 0, \\ \dot{\alpha} & = -2\, {}^t r_1 h - {}^t h B U^{-1}\, {}^t B h, & \alpha(T) = 0. \end{array} \tag{8.25}$$

We have then obtained

Proposition 8.9. *Let ξ be a trajectory of $\mathbb{R}^n$ on $[0,T]$ and consider the problem of tracking for the control system :*

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

where we want to minimize the cost

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

The there exists a unique optimal control that is written as

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

where $E(t) \in \mathcal{M}_n(\mathbb{R})$ and $h(t) \in \mathbb{R}^n$ are solutions on $[0,T]$ of

$$\begin{array}{rll} \dot{E} & = W - {}^tAE - EA - EBU^{-1}\,{}^tBE, & E(T) = -Q, \\ \dot{h} & = -{}^tAh - E(A\xi - \dot{\xi} + r) - EBU^{-1}\,{}^tBh, & h(T) = 0. \end{array}$$

Moreover $E(t)$ is symmetric . The optimal cost is given by

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

Remarque. Notice that the optimal control is written in the form of a feedback :

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

Bibliographie

[Agr-Sa] A. Agrachev, Y. Sachkov, Control Theory from the Geometric Viewpoint, vol. 87 of Encyclopaedia of Mathematical Sciences. Control Theory and Optimization, II. Springer, Berlin (2004)

[Rou-Bo] F. Bonnans, P. Rouchon, Commande et optimisation de systèmes dynamiques. Les Édition de L'École Polytechnique 2005.

[Bo-Pi] U. Boscain, B. Piccoli, Optimal Synthesis for Control Systems on 2-D Manifolds, Springer, SMAI, Vol.43, 2004.

[Jurd] Jurdjevic, V. : Geometric Control Theory, vol. 52 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1997)

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

[So] E. D. Sontag, Mathematical Control Theory : Deterministic Finite Dimensional Systems, no. 6 in Texts in Applied Mathematics, Springer-Verlag, New York/Heidelberg/Berlin, 2 ed., 1998.