Mathématiques pour l'ingénieur

Jean-Pierre Richard, Hugues Mounier, Abdennebi Achour, Lotfi Belkoura, Selma Ben Attia, Michel Dambrine, Mekki Ksouri, Wilfrid Perruquetti, Joachim Rudolph, Salah Salhi, et al.

Mathématiques pour l'ingénieur

Abdennebi ACHOUR,
Lotfi BELKOURA,
Michel DAMBRINE,
Mekki KSOURI,
Hugues MOUNIER,
Wilfrid PERRUQUETTI,
Jean-Pierre RICHARD,
Joachim RUDOLPH,
Frank WOITTENNEK,
Salah SALHI,
Selma BEN ATTIA

Illustration du livre des procédés ingénieux (Kitâb al-Hiyal) publié en 850 par les trois frères Ahmed, Mohamed et Hasan bin Mûsa ibn Shâkir, travaillant dans la maison de la sagesse (Bayt al-Hikma) à Bagdad.

Table des matières

1 Introduction aux Distributions

Lotfi Belkoura[1]

[1]LAGIS & INRIA-ALIEN, Université des Sciences et Technologies de Lille, Bât. P2, 59650 Villeneuve d'Ascq, France. *E-mail* :
Lotfi.Belkoura@univ-lille1.fr

1.1 Introduction

La théorie des distributions permet, en se plaçant dans un cadre plus large que celui, classique, des équations différentielles ordinaires, de résoudre de nombreuses équations issues de la physique, de la mécanique des fluides ou du traitement du signal. Elle permet ainsi, par exemple, de dériver, même indéfiniment, en un certain sens, une fonction qui n'est dérivable au sens usuel, et des informations essentielles tels que les discontinuités des fonctions ne sont pas perdues par dérivation. Une des idées fondamentales de cette théorie consiste à définir les distributions au travers de leur action sur un espace de fonctions, dites fonctions tests.

Ce chapitre limite son ambition à l'acquisition rapide de techniques de calculs puissantes, et les aspects tels que ceux relatifs aux propriétés topologiques des différents espaces ne sont as abordés. Il ne faut donc pas hésiter à consulter les ouvrages tels que celui de Laurent Schwartz, auteur de cette théorie, pour des développements et démonstrations plus complets. Les exemples et énoncés sont pour la plus grande partie extraits des ouvrages cités en références [10, 7, 1, 9, 4, 2, 5, 12, 6, 11, 3, 8]. Bien que restreintes aux situations à une dimension (de la variable t), les représentations développées dans ce chapitre admettent généralement une extension naturelle aux dimensions d'ordre supérieur, permettant d'appréhender les problèmes d'équation aux dérivées partielles.

1.2 Espaces des fonctions tests-Espaces des distributions

Une distribution est une forme linéaire continue sur un espace vectoriel de fonctions, dites fonctions tests. Il existe différents types de distributions correspondant aux différents espaces de fonctions de test. Plus les conditions de régularité imposées aux fonctions tests sont sévères, plus les fonctionnelles ainsi définies seront générales. Les distributions, généralisant la notion de mesure, sont définies à partir de l'espace $\mathcal{D}(\Omega)$ définit ci-après. Dans tout ce qui suit, les définitions générales et exemples sont basés sur des fonctions tests notées $\varphi(t)$, en nous bornant, sauf mention contraire, au cas des fonctions définies sur $\mathbb{R}$;

Espace vectoriel $\mathcal{D}(\Omega)$

Définition 1.2.1. Soit Ω un ouvert de $\mathbb{R}$. On note $\mathcal{D}(\Omega)$ l'espace des fonctions indéfiniment dérivables à support compact dans Ω.

Rappelons au passage la définition du support d'une fonction ; soit A l'ensemble des t tels que $\varphi(t) \neq 0$. Le support de la fonction φ, noté $\operatorname{supp}\varphi$ est le sous ensemble fermé $\bar{A}$. Ainsi par exemple, pour la fonction "porte" χ_T de largeur $T > 0$, définie par :

$$\chi_T(t) = \begin{cases} 1 & |t| \leq \frac{T}{2} \\ 0 & |t| > \frac{T}{2} \end{cases}, \quad \text{nous aurons :} \quad \operatorname{supp}\chi_T = \left[-\frac{T}{2}, \frac{T}{2}\right]. \tag{1.1}$$

Des exemples de fonctions appartenant à l'espace vectoriel ainsi défini ne viennent pas immédiatement à l'esprit. Un exemple fréquent est fournit par la fonction suivante

$$\zeta(t) = \begin{cases} 0 & |t| \geq 1 \\ \exp\frac{1}{t^2-1} & |t| < 1 \end{cases}, \tag{1.2}$$

de support $[-1, +1]$. Plus généralement, toute fonction $\zeta_{ab}(t)$ définie par

$$\zeta_{ab}(t) = \begin{cases} 0 & t \notin]a, b[\\ \exp\frac{1}{2}\left[\frac{1}{t-b} - \frac{1}{t-a}\right] & t \in]a, b[\end{cases}, \tag{1.3}$$

est une fonction de $\mathcal{D}$ ayant pour support $[a, b]$. Enfin, le théorème suivant permet d'en construire bien d'autres :

Théorème 1.2.1. *Si $\varphi \in \mathcal{D}$ et si f est une fonction sommable à support borné, alors : $\psi(t) = \int f(\theta)\varphi(t-\theta)d\theta$ est une fonction de $\mathcal{D}$.*

Distributions

Définition 1.2.2. Une distribution sur un ouvert Ω de $\mathbb{R}$ est une forme linéaire continue sur l'espace $\mathcal{D}(\Omega)$. Les distributions forment un espace vectoriel noté $\mathcal{D}'(\Omega)$.

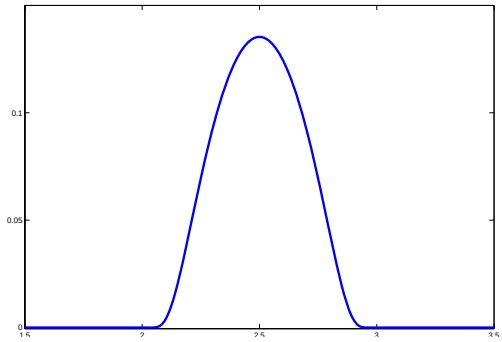


FIG. 1.1: Tracé de $\zeta_{23}(t)$.

Une distribution T est donc un application de $\mathcal{D}(\Omega)$ dans $\mathbb{C}$ faisant correspondre à une fonction test φ un nombre complexe noté $\langle T(t), \varphi(t)\rangle$, ou plus simplement $\langle T, \varphi\rangle$lorsqu'il n'y a pas ambiguïté sur la variable. C'est la valeur prise par la distribution sur la fonction φ. Les propriétés de linéarité et de continuité se traduisent respectivement par

$$\begin{cases} \forall \varphi_1, \varphi_2 \in \mathcal{D}(\Omega) : \langle T, \varphi_1 + \varphi_2\rangle = \langle T, \varphi_1\rangle + \langle T, \varphi_2\rangle \,, \\ \\ \forall \varphi \in \mathcal{D}(\Omega), \forall \lambda \in \mathbb{C} : \langle T, \lambda\varphi\rangle = \lambda \langle T, \varphi\rangle \,, \end{cases} \tag{1.4}$$

et pour la continuité par : Si φ_k converge dans $\mathcal{D}$ vers φ, la suite $\langle T, \varphi_k\rangle$ converge au sens usuel vers $\langle T, \varphi\rangle$, c'est à dire :

$$\forall \epsilon > 0 \; \exists N(\epsilon), \quad k \geq N \quad |\langle T, \varphi\rangle - \langle T, \varphi_k\rangle| \leq \epsilon. \tag{1.5}$$

Les distributions généralisent la notion de mesure définie par une fonctionnelle linéaire et continue sur l'espace $\mathcal{D}_0$ des fonctions continues à support borné.

Distributions régulières

On examine maintenant des distributions particulières, nommées distributions régulières, définies par une intégrale et qui permettent d'associer de manière univoque une fonction localement sommable (c'est à dire sommable sur tout ensemble borné) et la distribution qui lui est associée. Nous utiliserons pour simplifier la même notation pour désigner une fonction $f(t)$ et la distribution f qu'elle définit, le sens étant précisé par le contexte. Pour une fonction $f(t)$ localement sommable, on définit la distribution f par

$$\langle f, \varphi\rangle = \int f(t)\varphi(t)dt \quad \forall \varphi \in \mathcal{D}, \tag{1.6}$$

qui a toujours un sens puisque φ est à support borné. Notons que dans ce cas on ne peut attribuer à la distribution f une valeur pour chaque t, même si la

fonction $f(t)$ est une fonction régulière. En effet, deux distributions f et g seront considérées comme identiques si pour tout $\varphi \in \mathcal{D}$,

$$\langle f, \varphi \rangle = \langle g, \varphi \rangle . \tag{1.7}$$

Cela n'entraîne l'égalité des fonctions $f(t)$ et $g(t)$ dont elles sont issues que si ces dernières sont continues. Cependant, lorsque f et g sont des fonctions localement sommables quelconques, nous avons le théorème suivant ;

Théorème 1.2.2. *Deux fonctions localement sommables f et g définissent la même distribution si, et seulement si, elles sont égales presque partout.*

On voit ainsi que les distributions sont une extension non pas des fonctions localement sommables, mais des classes de fonctions sommables presque partout égales. Ceci provient du fait l'intégrale de Lebesgue n'est pas modifiée sur une ensemble de mesure nulle.

Distributions singulières

On appelle distribution singulière toute distribution qui n'est pas régulière. L'exemple le plus usuel est la distribution de Dirac δ définie en un point a quelconque par :

$$\langle \delta_a, \varphi \rangle = \varphi(a), \quad \forall \varphi \in \mathcal{D} . \tag{1.8}$$

Une telle distribution a été introduite initialement par Dirac pour les besoins de la mécanique quantique. Elle est parfois improprement appelée fonction de Dirac et manipulée comme une fonction en écrivant

$$\langle \delta_a, \varphi \rangle = \varphi(a) = \int \varphi(t)\delta(t-a)dt, \tag{1.9}$$

$$1 = \int \delta(t-a)dt. \tag{1.10}$$

Une telle fonction devrait être nulle pour $t \neq a$ et valoir ∞ au point $t = a$. D'après la théorie classique des fonctions, son intégrale serait nulle ce qui est en contradiction avec (1.10). La distribution de Dirac définit également une mesure, et c'est aussi l'exemple le plus simple de mesure qui ne soit pas une fonction. Plus généralement, toute combinaison linéaire, finie ou non, $\sum b_i \delta_{a_i}$ définit une distribution singulière, mais d'autres distributions singulières, qui ne sont plus des mesures, peuvent être définis. Ainsi, comme on le verra au paragraphe sur la dérivation, la dérivée d'un Dirac au point a admettra naturellement pour définition :

$$\langle \delta'_a, \varphi \rangle = -\varphi'(a), \quad \forall \varphi \in \mathcal{D} . \tag{1.11}$$

Une distribution peut parfois également être définie à partir d'une fonction qui n'est pas localement intégrable. Sa valeur sur une fonction φ est définie par la partie finie (notée pf) d'une intégrale divergente, notion introduite par Hadamard pour les besoins de la théorie des équations aux dérivées partielles. La

distribution associée est appelée pseudo-fonction, également notée pf et, f étant la fonction, on écrit :

$$\langle \text{pf}\, f, \varphi \rangle = \text{pf} \int_{-\infty}^{\infty} f(t)\varphi(t)dt. \tag{1.12}$$

Il faut dans chaque cas définir les conditions d'existence de la partie finie. Un exemple classique concernant la fonction $\log|x|$ et la pseudo-fonction $1/x$ est abordé au paragraphe portant sur la dérivation.

Support d'une distribution

La définition du support d'une distribution T, noté $\text{supp}\, T$ peut être envisagé au travers de celle de restriction d'une distribution à un ouvert définie ci-dessous :

Restriction d'une distribution à un ouvert Considérons deux ouverts $\Omega \subset \Omega'$ de $\mathbb{R}$ et soit $T \in \mathcal{D}'(\Omega')$. Nous pouvons associer à T une distribution T_Ω appelée restriction de T à Ω, définie pour toute $\varphi \in \mathcal{D}(\Omega)$ par :

$$\langle T_\Omega, \varphi \rangle = \langle T, \bar{\varphi} \rangle\,, \tag{1.13}$$

et dans laquelle $\bar{\varphi}$ est le prolongement par 0 de φ à Ω'. Il faut cependant prendre garde au fait que, contrairement aux fonctions, deux distributions distinctes sur Ω' peuvent avoir la même restriction à Ω. Ainsi par exemple, pour $a \in \Omega'$ et $\Omega = \Omega' - a$, la distribution nulle et la mesure de Dirac $\delta - a$ ont même restriction. Cela permet d'introduire la

Définition 1.2.3. Le support de T, noté $\text{supp}\, T$ est le complémentaire dans Ω du plus grand ouvert ω de Ω tel que la restriction de T à ω soit nulle.

Il y a cohérence entre la notion de support établie du point de vue de la théorie des fonctions et celle issue de la théorie des distributions. Le support d'une fonction coïncide avec celui de la distribution qu'elle définit. Pour des distribution singulières, et à titre d'exemple,

$$\text{supp}\, \delta_a = \{a\}\,. \tag{1.14}$$

Inversement, un théorème très utile est à notre disposition concernant les distributions à support ponctuel.

Théorème 1.2.3. *Toute distribution de support l'origine admet une décomposition unique comme combinaison linéaire finie de dérivées de la distribution de Dirac :*

$$T = \sum_{p \leq m} c_p\, \delta^{(p)}, \tag{1.15}$$

les c_p étant des constantes.

Deux classes de distributions sont appelées à jouer un rôle important en physique. Les distributions à support borné, et celles à support contenu dans $[0, \infty)$. Toutes deux forment des espaces vectoriels notés respectivement $\mathcal{E}'$ et $\mathcal{D}'_+$. Les distributions à support borné à gauche sont souvent appelée en physique distributions causales (la variable étant dans ce cas le temps). Elles représentent des phénomènes qui ne peuvent avoir lieu avant la cause qui les produit et par conséquent sont nulles pour $t < 0$.

Ordre d'une distribution

La notion d'ordre d'une distribution peut s'avérer utile dans certaines applications. On note $\mathcal{D}_K(\Omega)$ l'ensemble des fonction de $\mathcal{D}(\Omega)$ ayant leur support inclut dans $K \subset \Omega$.

Définition 1.2.4. On appelle distribution d'ordre fini toute distribution T de $\mathcal{D}'(\Omega)$ pour laquelle il existe $k \in \mathbb{N}$, tel que pour tout compact K inclus dans Ω, on ait :

$$\exists C_K > 0, \quad \forall \varphi \in \underset{K}{\mathcal{D}}(\Omega), \quad |\langle T, \varphi | \rangle \leq C_K \sup_{|\alpha| \leq k} \sup_{t \in \Omega} |D^\alpha \varphi(t)|. \tag{1.16}$$

L'entier k, qui ne dépend pas de K, est appelé ordre de la distribution T.

Ainsi par exemple, la distribution de Dirac au point a est d'ordre 0 car :

$$\forall \varphi \in \mathcal{D}(\Omega), |\langle \delta_a, \varphi | \rangle = |\varphi(a)| \leq \sup_{t \in \Omega} |\varphi(t)|. \tag{1.17}$$

Aussi, on établit que les fonctions localement sommables définissent des distributions d'ordre 0, tandis que les distributions singulières sous forme de somme finie ($\sum_0^r a_r \delta^{(r)}$ + fonctions) sont d'ordre r.

Sous espaces de $\mathcal{D}'(\Omega)$

Comme mentionné en introduction, si l'on prend des espaces de fonctions tests moins restreints que $\mathcal{D}$, on obtient des sous espaces de $\mathcal{D}'$. Les espaces de fonctions tests les plus couramment utilisés sont les espace $\mathcal{S}$ et $\mathcal{E}$ ci-après, vérifiant $\mathcal{D} \subset \mathcal{S} \subset \mathcal{E}$, et définis par :

- espace $\mathcal{S}$: espace des fonctions indéfiniment dérivables décroissant à l'infini, ainsi que toutes leurs dérivées, plus vite que toute puissance de $1/|x|$.
- espace $\mathcal{E}$: espace des fonctions indéfiniment dérivables quelconques.

Ils conduisent aux sous espaces de $\mathcal{D}'$ suivants :

- espace $\mathcal{S}'$: espace des distributions tempérées,
- espace $\mathcal{E}'$: espace des distributions à support borné.

avec les inclusions $\mathcal{D}' \supset \mathcal{S}' \supset \mathcal{E}'$. Les distributions tempérées jouent un rôle particulièrement important en physique, grâce notamment à la transformée de Fourier. Ainsi, toute distribution tempérée admet une une transformée de Fourier qui est elle même une distribution tempérée.

1.3 Opérations sur les distributions

Changement d'échelle et translation

Ces opérations peuvent être réalisées en considérant l'image d'une distribution par un opérateur affine. Soit v un opérateur linéaire continu dans $\mathcal{D}$. On définit par transposition dans l'espace dual $\mathcal{D}'$ l'opérateur ${}^t v$ qui à $T \in \mathcal{D}'$ associe ${}^t v(T)$ tel que :

$$\left\langle {}^t v(T), \varphi \right\rangle = \langle T, v(\varphi) \rangle \qquad \forall \varphi \in \mathcal{D}, \forall T \in \overset{\prime}{\mathcal{D}}. \tag{1.18}$$

Lorsque v est l'opérateur $\varphi(t) \xrightarrow{v} \varphi(at+b)$, et f une fonction localement sommable, une application classique du changement de variable dans l'intégrale permet d'obtenir :

$$\langle f(t), \varphi(at+b) \rangle = \frac{1}{|a|} \left\langle f(\frac{t-b}{a}), \varphi(t) \right\rangle. \tag{1.19}$$

Par analogie, pour une distribution T quelconque, on définit :

$${}^t v(T)(t) = \frac{1}{|a|} T(\frac{t-b}{a}). \tag{1.20}$$

Cette formule permet par exemple d'écrire, pour un changement d'échelle de temps, en prenant $b = 0$:

$$\delta(\frac{t}{a}) = |a| \delta(t). \tag{1.21}$$

De même, la parité d'une distribution peut être définie en prenant $a = -1$ et $b = 0$. Avec le notations $\check{T}(t) = T(-t)$, nous dirons qu'une distribution est paire (resp. impaire) lorsque $\check{T} = T$ (resp $\check{T} = -T$). Enfin, en prenant $a = 1$, on obtient l'image d'une distribution par translation d'amplitude b :

$$\langle T(t-b), \varphi(t) \rangle = \langle T, \varphi(t+b) \rangle, \tag{1.22}$$

et une distribution sera dite périodique de période $b > 0$ lorsque $T(t-b) = T(t)$.

Dérivation

La propriété essentielle des distributions est qu'elles sont indéfiniment dérivables. On introduit tout d'abord la dérivée d'une distribution de sorte que sa

définition coïncide avec avec la définition usuelle pour une fonction localement sommable f, à dérivée f' continue. Par intégration par partie, il vient pour cette dernière :

$$\left\langle f'(t), \varphi(t) \right\rangle = \int f'(t)\varphi(t)dt = -\int f(t)\varphi'(t)dt = -\left\langle f, \varphi' \right\rangle. \tag{1.23}$$

Ceci définit bien une distribution régulière puisque $\varphi' \in \mathcal{D}$. On est ainsi conduit à la définition :

$$\left\langle T', \varphi \right\rangle = -\left\langle T, \varphi' \right\rangle, \tag{1.24}$$

et plus généralement pour la dérivée d'ordre m :

$$\left\langle T^{(m)}, \varphi \right\rangle = (-1)^{(m)} \left\langle T, \varphi^{(m)} \right\rangle, \tag{1.25}$$

Exemples de dérivation

Dérivée de la distribution de Dirac δ. Un premier exemple de dérivation ne conduisant pas à une mesure est donnée par la relation (1.11) donnant l'expression de la dérivée d'une distribution de Dirac :

$$\left\langle \delta', \varphi \right\rangle = -\left\langle \delta, \varphi' \right\rangle = -\varphi'(0). \tag{1.26}$$

Dérivée de la fonction de Heaviside $H(t)$. La fonction de Heaviside, définie par :

$$H(t) = \begin{cases} +1 & t > 0 \\ 0 & t < 0 \end{cases}, \tag{1.27}$$

est une fonction localement sommable qui définit donc une distribution. A noter qu'en tant que distribution, la valeur de cette fonction en $t = 0$ n'a pas besoin d'être précisée. Sa dérivée s'écrit :

$$\begin{aligned} \left\langle H'(t), \varphi(t) \right\rangle &= -\left\langle H(t), \varphi'(t) \right\rangle = -\int_0^\infty \varphi'(t)dt \\ &= -\left[\varphi(t)\right]_0^\infty = \varphi(0). \end{aligned} \tag{1.28}$$

On en conclut que :

$$H' = \delta. \tag{1.29}$$

Fonctions régulières par morceaux. L'exemple précédent se généralise aisément aux fonctions f qui possèdent les caractéristiques suivantes : Soit $\{t_\nu\}_{\nu\in\mathbb{Z}}$ un suite de nombres réels dstincts tels que $\lim_{\nu\to+\infty} t_\nu = \infty$, et $\lim_{\nu\to-\infty} t_\nu = -\infty$. On suppose f indéfiniment dérivable au sens classique dans les intervalles $]t_\nu, t_{\nu+1}[$, et f, ainsi que toute ses dérivées d'ordre p au sens usuel, notées $f^{(p)}$, ont des discontinuités de première espèce (i.e. les limites à droite et à gauche de $f^{(p)}(t_\nu)$ existent). On note alors σ_ν^p les sauts de $f^{(p)}$ en t_ν :

$$\sigma_\nu^p = f^{(p)}(t_\nu + 0) - f^{(p)}(t_\nu - 0)). \tag{1.30}$$

On note enfin par la suite $D^p f$ la dérivée distribution d'ordre p de la fonction f, et ce afin de la distinguer de la dérivée usuelle $f^{(p)}$. La dérivée distribution Df de f est définie par :

$$\langle Df, \varphi \rangle = - \langle f, \varphi' \rangle = - \int f(t)\varphi'(t)dt = - \sum_\nu \int_{t_\nu}^{t_{\nu+1}} f(t)\varphi'(t)dt, \qquad (1.31)$$

pour toute $\varphi \in \mathcal{D}$. Une intégration par parties montre alors que :

$$\langle Df, \varphi \rangle = - \int f'(t)\varphi(t)dt + \sum_\nu \varphi(t_\nu)\sigma_\nu, \qquad (1.32)$$

ce qui s'écrit :

$$\langle Df, \varphi \rangle = f' + \sum_\nu \sigma_\nu \delta_\nu. \qquad (1.33)$$

Ainsi, les discontinuités de f apparaissent dans la dérivée Df sous forme de mesures de Dirac. Les informations concernant les discontinuités de la fonction f ne sont donc pas perdues par la dérivation. Cette formule se généralise de proche en proche pour les dérivations d'ordre supérieur, faisant apparaître les dérivées successives de la distribution de Dirac :

$$\langle D^p f, \varphi \rangle = f^{(p)} + \sum_\nu \sigma_\nu^{p-1}\delta_\nu + \sum_\nu \sigma_\nu^{p-2}\delta'_\nu + \cdots + \sum_\nu \sigma_\nu \delta_\nu^{(p-1)}. \qquad (1.34)$$

Dérivation de $\log|t|$ **et pseudo fonction** $\mathrm{pf}\, 1/t^n$**.** La dérivation de distributions régulières peut conduire à des distributions singulières. C'est le cas de la fonction $\log|t|$ qui définit une distribution régulière, mais dont la dérivée au sens des fonction $1/t$ n'est pas localement sommable. Elle ne peut donc représenter la distribution dérivée. Cette dernière s'écrit au sens des distributions :

$$\begin{aligned}
\langle D\log|t|, \varphi \rangle &= - \big\langle \log|t|, \varphi'(t)dt \big\rangle = - \int_{-\infty}^{\infty} \log|t|, \varphi'(t)dt \\
&= \lim_{\epsilon \to 0} - \int_{|t|\geq\epsilon} \log|t|, \varphi'(t)dt \\
&= \lim_{\epsilon \to 0} \left\{ \log\epsilon\,[\varphi(\epsilon) - \varphi(-\epsilon)] + \int_\epsilon^\infty \frac{\varphi(t) - \varphi(-t)}{t} dt \right\} \\
&= \lim_{\epsilon \to 0} \int_\epsilon^\infty \frac{\varphi(t) - \varphi(-t)}{t} dt, \qquad (1.35)
\end{aligned}$$

car $\log\epsilon\,[\varphi(\epsilon) - \varphi(-\epsilon)] \to 0$ grâce au théorème des accroissements finis. Ce dernier terme est noté alors partie finie de l'intégrale divergente $\int_{-\infty}^{\infty} \frac{\varphi(t)}{t} dt$ et définit une distribution notée $\mathrm{pf}\, \frac{1}{t}$. On obtient alors :

$$D\log|t| = \mathrm{pf}\, \frac{1}{t} \qquad (1.36)$$

Notons également que la dérivation de pf $\frac{1}{t}$ conduit à son tour à la formule générale :

$$D[\text{pf}\,\frac{1}{t^n}] = -n\,\text{pf}\,\frac{1}{t^{n+1}}. \tag{1.37}$$

Autrement dit, la règle de dérivation est celle classique de la théorie des fonctions. Ces résultats permettent de considérer des distributions basées sur des fraction rationnelle quelconques. Il suffit de décomposer en élément simples pour faire apparaître, outre les éléments qui sont indéfiniment dérivables, une somme de termes de la forme $\frac{A}{(t-a)^m}$ qui sont des pseudo fonctions translatées de a.

En restreignant le domaine d'étude à $t > 0$, nous pourrions de manière analogue introduire la pseudo-fonction pf $\frac{H(t)}{t}$ par :

$$\left\langle \text{pf}\,\frac{H(t)}{t}, \varphi \right\rangle = \lim_{\epsilon \to 0} \left\{ \int_\epsilon^\infty \frac{\varphi(t)}{t} dt + \varphi(0)\log(\epsilon) \right\}, \tag{1.38}$$

et pour laquelle une intégration par parties permet d'obtenir :

$$\left\langle \text{pf}\,\frac{H(t)}{t}, \varphi \right\rangle = \int_0^\infty \varphi'(t)\log t\, dt = \int_{-\infty}^\infty \varphi'(t)H(t)\log t\, dt. \tag{1.39}$$

La fonction $\varphi' \log t$ est intégrable, ce qui justifie l'existence de la partie finie. En outre, le dernier terme conduit, par définition de la dérivée d'une distribution, à :

$$D[H(t)\log t] = \text{pf}\,\frac{H(t)}{t}. \tag{1.40}$$

Dans ce cas, on montre que les dérivations successives conduisent à la formule suivante faisant intervenir les dérivées de la distribution de Dirac à l'origine :

$$D[\text{pf}\,\frac{H(t)}{t^n}] = -n\,\text{pf}\,\frac{H(t)}{t^{n+1}} + \frac{(-1)^n}{n!}\delta^{(n)}. \tag{1.41}$$

Multiplication des distributions

Si f et g sont deux fonctions localement sommables, et si T_f et T_g désignent les distributions correspondantes, nous souhaitons, comme pour le cas des fonctions, pouvoir écrire :

$$T_f.T_g = T_{fg}. \tag{1.42}$$

Malheureusement, l'existence de T_f et de T_g n'entraîne pas automatiquement celle de T_{fg}, car les deux fonctions peuvent être localement sommables sans que le produit le soit (l'exemple le plus courant étant $f(t) = g(t) = 1/\sqrt{t}$). Il semble donc ne pas y avoir de définition naturelle pour le produit de deux distributions quelconques.

Dans le cas du produit d'une distribution T et d'une fonction α indéfiniment dérivable, on définit le produit αT par :

$$\langle \alpha T, \varphi \rangle = \langle T, \alpha\varphi \rangle \,, \quad \forall \varphi \in \mathcal{D} \,. \tag{1.43}$$

La fonction $\alpha\varphi$ appartient bien à $\mathcal{D}$ car, comme φ, elle est à support borné, et indéfiniment dérivable (comme produit de deux fonctions indéfiniment dérivables). Donc $\forall T$, le produit αT a un sens. Pour certaines distributions, la définition du produit reste applicable même si la fonction α n'est pas indéfiniment dérivable. Il suffit par exemple que $\alpha(t)$ soit continue en 0 pour pouvoir définir $\alpha\delta$:

$$\begin{aligned} \langle \alpha(t)\delta(t), \varphi(t) \rangle &= \langle \delta(t), \alpha(t)\varphi(t) \rangle = \alpha(0)\varphi(0) \\ &= \langle \alpha(0)\delta(t), \varphi(t) \rangle \,. \end{aligned} \tag{1.44}$$

On obtient donc

$$\alpha(t)\delta(t) = \alpha(0)\delta(t). \tag{1.45}$$

On établit de même, en un point a quelconque où α est continue, $\alpha(t)\delta(t-a) = \alpha(a)\delta(t-a)$, et en particulier :

$$t\delta = 0. \tag{1.46}$$

Parmi les premières propriétés du produit multiplicatif, on note celle relative aux supports :

$$\operatorname{supp} \alpha T \subset \operatorname{supp} \alpha \cap \operatorname{supp} T. \tag{1.47}$$

D'une manière plus générale, nous avons le théorème suivant, aux conditions assez restrictives :

Théorème 1.3.1. *Le produit de plusieurs distributions, lorsque toutes, sauf une au plus, sont des fonctions indéfiniment dérivables au sens usuel, est associatif et commutatif.*

$$(\alpha_1 + \alpha_2)T = \alpha_1 T + \alpha_2 T, \tag{1.48}$$

$$(\alpha_1 \alpha_2)T = \alpha_1(\alpha_2 T), \tag{1.49}$$

$$\alpha(T_1 + T_2) = \alpha T_1 + \alpha T_2 \tag{1.50}$$

Si les conditions ne sont pas vérifiées, la multiplication n'est plus nécessairement associative, comme le montre l'exemple suivant avec δ, t, et pf $\frac{1}{t}$:

$$\begin{aligned} &(\delta\, t)\,\mathrm{pf}\,\tfrac{1}{t} = 0 \quad \text{car} \quad \delta\, t = 0, \\ &\delta\,(t\,\mathrm{pf}\,\tfrac{1}{t}) = \delta \quad \text{car} \quad t\,\mathrm{pf}\,\tfrac{1}{t} = 1. \end{aligned} \tag{1.51}$$

Dans ce dernier cas en effet, nous avons :

$$\left\langle t\,\mathrm{pf}\,\frac{1}{t}, \varphi \right\rangle = \left\langle \mathrm{pf}\,\frac{1}{t}, t\varphi \right\rangle = \mathrm{pf} \int_{-\infty}^{\infty} \frac{t\varphi}{t} dt = \int_{-\infty}^{\infty} \varphi(t) dt. \tag{1.52}$$

Toujours dans la mesure ou le le produit αT a un sens, la règle de dérivation s'obtient par la formule usuelle de Leibniz, avec $C_m^k = \frac{m!}{k!(m-k)!}$:

$$D^m(\alpha T) = \sum_{k \leq m} C_m^k (D^{(m-k)}\alpha) D^k T. \tag{1.53}$$

La démonstration pour la dérivée d'ordre 1 est simple et permet de revisiter les définitions précédemment introduites :

$$\begin{aligned} \langle (\alpha T)', \varphi \rangle &= -\langle \alpha T, \varphi' \rangle = -\langle T, \alpha\varphi' \rangle = -\langle T, (\alpha\varphi)' - \alpha'\varphi \rangle \\ &= -\langle T, (\alpha\varphi)' \rangle + \langle T, \alpha'\varphi \rangle = \langle T', \alpha\varphi \rangle + \langle \alpha' T, \varphi \rangle \\ &= \langle \alpha T', \varphi \rangle + \langle \alpha' T, \varphi \rangle = \langle \alpha T' + \alpha' T, \varphi \rangle . \end{aligned} \tag{1.54}$$

Le théorème ci-dessous peut s'avérer bien utile dans certaines applications :

Théorème 1.3.2. *Si T a un support compact K, et est d'ordre (nécessairement fini) m, αT est nulle toutes les fois que α et ses dérivées d'ordre $\leq m$ sont nulles sur K ; si T a un support quelconque et est d'ordre quelconque, fini ou infini, αT est nulle si α est nulle ainsi que toute ses dérivées sur le support de T.*

Compte tenu de ce qui précède, on établit ainsi que :

$$t^l\, \delta^{(n)} = \begin{cases} 0 & l > n, \\ (-1)^l \, \frac{n!}{(n-l)!} \, \delta^{(n-l)} & l \leq n. \end{cases} \tag{1.55}$$

et plus généralement :

$$\alpha\delta^{(n)} = \sum_{q \leq n} (-1)^{(n-q)} C_n^q \alpha^{(n-q)}(0) \delta^{(q)}. \tag{1.56}$$

Le produit de deux distributions peut également être étendu en faisant appel à la notion de support singulier :

Définition 1.3.1. Soit T une distribution sur un ouvert Ω de $\mathbb{R}$. Le support singulier de T est par définition le complémentaire du plus grand ouvert ω de Ω tel que la restriction de T à $\mathcal{D}(\omega)$ coïncide avec la distribution associée à une fonction de classe $\mathcal{C}^\infty$ sur ω.

Ainsi, par exemple, les distributions pf $\frac{H(t)}{t^n}$ ou plus simplement la fonction de Heaviside (en prenant $n = 0$) ont pour support $[0, \infty[$ et pour support singulier l'origine. On a alors la

Proposition 1.3.1. *Soient T_1 et T_2 deux distributions de supports singuliers disjoints. On peut alors donner un sens au produit $T_1 T_2$ en tant que distribution.*

Le problème de la division

La formule (1.46) (i.e. $t\delta = 0$) montre que le produit de deux distributions peut être nul sans qu'aucune d'elles le soit. Réciproquement, que peut-on dire d'une distribution T telle que $tT = 0$? Nous avons la

Proposition 1.3.2. *Les solutions de l'équation $tT = 0$ sont les distributions $T = c\delta$, où $c \in \mathbb{C}$.*

Plus généralement, on démontres que si $\alpha(t)$ est une fonction indéfiniment dérivable, n'admettant que des racines simples a_i à l'équation $\alpha = 0$, alors toutes les solutions de l'équations $\alpha T = 0$ sont de la forme :

$$T = \sum c_i \delta(t - a_i), \tag{1.57}$$

où les c_i sont des constantes arbitraires. Le problème de la division est assez délicat dans un cadre plus général, mais on notera que, avec la fonction α précédente, pour résoudre le problème

$$\alpha T = S, \tag{1.58}$$

où S est une distribution donnée, et en supposant connaître une solution T_0 particulière, nous avons immédiatement $\alpha(T - T_0) = 0$, et par suite,

$$T = T_0 + \sum c_i \delta(t - a_i). \tag{1.59}$$

Lorsque la fonction α admet des racines multiples, l'énoncé précédent n'est plus valable, mais nous disposons de la

Proposition 1.3.3. *Pour toue distribution S, il existe un infinité de distributions T vérifiant $t^k T = S$. Deux d'entres elles quelconques diffèrent d'une combinaison linéaire de dérivées $\delta^{(m)}$, $m \leq k$.*

Exemples :

1. Soit à résoudre $tT = 1$, où 1 est la distribution définie par la fonction constante. En notant que $\mathrm{pf}\,\frac{1}{t}$ constitue une solution particulière (car $t\,\mathrm{pf}\,\frac{1}{t} = 1$), T admet la forme générale $T = c\delta + \mathrm{pf}\,\frac{1}{t}$.
2. Si maintenant on cherche à résoudre $tT = \delta$, une astuce pour obtenir une solution particulière consiste à dériver $t\delta = 0$ pour obtenir $\delta + t\delta' = 0$, ce qui suggère comme solution particulière $T_0 = -\delta'$, et par suite la solution générale $T = \delta' + c\delta$.

1.4 Convolution des distributions

Le produit de convolution est une opération essentielle dans les mathématiques appliquées. La convolution possède un élément neutre, ce qui va permettre de résoudre certaines équations de convolution et d'obtenir des solutions élémentaires d'opérateurs différentiels.

Produit tensoriel de deux distributions

La définition du produit de convolution requiert celle de produit tensoriel (ou produit direct) dont l'existence est fondée sur le théorème suivant.

Théorème 1.4.1. *Soient deux distributions U et V sur $\mathbb{R}$ et deux fonctions g et h de $\mathcal{D}(\mathbb{R})$). Il existe une et une seule distribution sur $\mathbb{R}^2$, notée $U \otimes V$ et telle que :*

$$\langle U \otimes V, g(u)\, h(v)\rangle = \langle U, g(u)\rangle \, \langle V, h(v)\rangle \, . \tag{1.60}$$

C'est le produit tensoriel de U et V. Sa valeur sur une fonction $\varphi(u,v) \in \mathcal{D}(\mathbb{R}^2)$ est :

$$\langle U \otimes V, \varphi\rangle = \langle U(u), \langle V(v), \varphi(u,v)\rangle\rangle = \langle V(v), \langle U(u), \varphi(u,v)\rangle\rangle \, . \tag{1.61}$$

Les expressions $\langle V(v), \varphi(u,v)\rangle$ et $\langle U(u), \varphi(u,v)\rangle$ se calculent en laissant respectivement u et v fixes. Ainsi, si f et g sont deux fonctions localement sommables, le produit tensoriel des distributions associées s'écrit :

$$\langle f \otimes g, \varphi\rangle = \int\int f(u)g(v)\varphi(u,v)du\, dv), \tag{1.62}$$

$$\text{si bien que} \quad (f \otimes g)(u,v) = f(u)\, g(v). \tag{1.63}$$

Si maintenant on calcule le produit tensoriel de $\delta(u)$ et $\delta(v)$, il vient :

$$\langle \delta(u) \otimes \delta(v), \varphi(u,v)\rangle = \varphi(0,0) \tag{1.64}$$

$$\text{si bien que} \quad \delta(u) \otimes \delta(v) = \delta(u,v). \tag{1.65}$$

Définition et conditions d'existence

On commence par rappeler le produit de convolution de deux fonctions f et g, pour chercher ensuite à l'étendre au produit de convolution de distributions. Lorsqu'il existe le produit de convolution de f et g est la fonction h définie par :

$$h(t) = \int f(t-\theta)g(\theta)d\theta, \tag{1.66}$$

et on note $h = f * g$ ce produit. Lorsque f et g sont intégrable, h existe et est intégrable. Calculons alors la distribution associée à h, en définissant pour toute $\varphi \in \mathcal{D}$:

$$\begin{aligned} \langle f * g, \varphi\rangle &= \langle h, \varphi\rangle = \int h(t)\varphi(t)dt \\ &= \int\int f(t-\theta)g(\theta)\varphi(t)d\theta\, dt \\ &= \int\int f(u)g(v)\varphi(u+v)du\, dv, \end{aligned} \tag{1.67}$$

dans lequel nous avons posé le changement de variable $v = \theta$ et $u = t - \theta$. Ceci suggère de définir le produit de convolutions S et T par :

Définition 1.4.1. On appelle produit de convolution de deux distributions S et T, la distribution $S * T$, quand elle existe, telle que pour toute $\varphi \in \mathcal{D}$:

$$\langle S * T, \varphi \rangle = \langle S(u) \otimes T(v), \varphi(u+v) \rangle \,, \tag{1.68}$$

et dans lequel $S(u) \otimes T(v)$ est le produit direct de S et T.

Le produit de convolution de deux distributions quelconques n'est pas toujours défini car, si la fonction $t \mapsto \varphi(t)$ est à support borné dans $\mathbb{R}$, il n'en est pas de même de la fonction $(u, v) \mapsto \varphi(u+v)$. En effet, si (a, b) est le support de φ, le support de $(u, v) \mapsto \varphi(u+v)$ est la bande $a \leq u+v \leq b$, et par suite $\varphi(u+v)$ n'est pas un élément de $\mathcal{D}$. Néanmoins, on montre que le produit de convolution existe dans un des cas de figures suivant, que l'on rencontre très largement dans la pratique :

- l'une au moins des deux distributions est à support compact,
- les deux distributions sont à support limité à gauche (resp. à droite).

Il est clair que le produit de convolution est commutatif, $S * T = T * S$. L'extension au produit de plusieurs distributions (dans cet exemple 3) se généralise en :

$$\langle R * S * T, \varphi \rangle = \langle R(u) \otimes S(v) \otimes T(w), \varphi(u+v+w) \rangle \,, \tag{1.69}$$

et l'existence est assurée dans l'un ou l'autre des trois cas suivants :

- toutes sauf une au plus sont à support compact,
- toutes ont leur support limité à gauche (resp. à droite),
- les produit deux à deux sont bien définis.

De plus, lorsqu'il existe, ce produit est associatif. Il faut garder en mémoire qu'il s'agit ici de conditions suffisantes, le produit de convolution pouvant exister dans d'autres situations. Néanmoins, ne pas en tenir compte peut conduire à des conclusions erronées telles celle considérant sans précaution le produit $1 * \delta' * H$ pour aboutir aux deux différents résultats :

$$1 * (\delta' * H) = 1 * \delta = 1 \tag{1.70}$$

$$(1 * \delta') * H = 0 * H = 0. \tag{1.71}$$

Propriétés

Support du produit de convolution : Nous disposons de l'inclusion suivante :

$$\operatorname{supp} S * T \subset \operatorname{supp} S \, + \operatorname{supp} T, \tag{1.72}$$

et dans laquelle le membre de droite se lit (somme de Minkowski) :

$$\{x + y;\, x \in \operatorname{supp} S,\, y \in \operatorname{supp} T\} \,. \tag{1.73}$$

Ainsi, par exemple, si le support de S est inclut dans (a, ∞) et celui de T dans (b, ∞), celui de $S * T$ sera inclut dans $(a + b, \infty)$. A noter également un résultat plus précis, connu sous le nom de Théorème des supports, et qui s'énonce,

$$\text{conv}(\text{supp}\, S * T) = \text{conv}(\text{supp}\, S) + \text{conv}(\text{supp}\, T), \tag{1.74}$$

et dans laquelle $\text{conv}(\text{supp}\, X)$ désigne l'enveloppe convexe du support de X.

Convolution par δ : δ étant à support compact, $\delta * T$ existe pour toute distribution T, et, par définition :

$$\begin{aligned} \langle \delta * T, \varphi \rangle &= \langle \delta(u) T(v), \varphi(u+v) \rangle = \langle T(v), \langle \delta(u), \varphi(u+v) \rangle \rangle \\ &= \langle T(v), \varphi(v+0) \rangle = \langle T, \varphi \rangle \end{aligned} \tag{1.75}$$

Ce qui montre que

$$\delta * T = T, \tag{1.76}$$

c'est à dire que la distribution de Dirac joue le rôle de l'unité pour le produit de convolution.

Convolution par $\delta^{(m)}$: on considère tout d'abord le cas $m = 1$ qui conduit par définition à :

$$\begin{aligned} \left\langle \delta' * T, \varphi \right\rangle &= \left\langle \delta'(u) T(v), \varphi(u+v) \right\rangle = \left\langle T(v), \left\langle \delta'(u), \varphi(u+v) \right\rangle \right\rangle \\ &= -\left\langle T(v), \varphi'(v) \right\rangle = \left\langle T', \varphi \right\rangle, \end{aligned} \tag{1.77}$$

ce qui montre que la dérivation équivaut à un convolution avec δ'. On établirait de même que :

$$\delta^{(m)} * T = T^{(m)}. \tag{1.78}$$

Plus généralement encore, si $T = R * S$, le produit de convolution étant associatif, on obtiendrait :

$$(R * S)^{(m)} = R * S^{(m)} = R^{(m)} * S, \tag{1.79}$$

que l'on retient en disant que pour dériver un produit de convolution, il suffit de dériver l'un quelconque de ses termes.

Convolution par $\delta(t - a)$: Ici encore, l'application de la définition conduit à :

$$\begin{aligned} \langle \delta(t-a) * T(t), \varphi(t) \rangle &= \langle \delta(u-a) T(v), \varphi(u+v) \rangle \\ &= \langle T(v), \langle \delta(u-a), \varphi(u+v) \rangle \rangle \\ &= \langle T(v), \varphi(v+a) \rangle = \langle T(t-a), \varphi(t) \rangle, \end{aligned} \tag{1.80}$$

ce qui permet de conclure que pour translater une distribution de a, il suffit de la convoluer avec la translatée $\delta(t - a)$ de la distribution de Dirac δ.

Primitives d'ordre n dans $\mathcal{D}'_+$: Si f est une fonction de $\mathcal{D}'_+$, c'est à dire dont le support est contenu dans $(0, \infty)$, nous pouvons écrire :

$$\int_0^t f(\theta)\, d\theta = \int_0^\infty H(t-\theta) f(\theta)\, d\theta = (H * f)(t), \tag{1.81}$$

de sorte qu'intégrer f de 0 à t revient à convoluer f avec la fonction de Heaviside. Nous avons, en d'autres termes, obtenu une primitive de la fonction f. Plus généralement, la primitive d'ordre n s'écrit :

$$\int_0^t \int_0^{\theta_{n-1}} \cdots \int_0^{\theta_1} f(\theta)\, d\theta_1 \cdots d\theta_{n-1} d\theta = \frac{1}{(n-1)!} \int_0^t f(\theta)(t-\theta)^{n-1}\, d\theta. \tag{1.82}$$

Si on note H_n la fonction $t \mapsto \frac{H(t)t^{n-1}}{\Gamma(n)}$, la primitive d'ordre n de f s'écrit $H_n * f$. Ces considérations se généralisent également à l'obtention de la dérivée et la primitive d'ordre α avec α complexe.

Convolution de polynômes-exponentiels : Toujours dans le cas de fonctions sommables f_1 et f_2, le produit de convolution de deux distributions s'identifie au produit usuel de deux fonctions. En particulier, si elles appartiennent à $\mathcal{D}'_+$, nous aurons :

$$h(t) = (f_1 * f_2)(t) = H(t) \int_0^t f_1(\theta) f_2(t-\theta) d\theta. \tag{1.83}$$

A titre d'exemple, par application directe de cette relation, on obtient facilement :

$$H(t)e^{\lambda_1 t} * H(t)e^{\lambda_2 t} = H(t)\frac{e^{\lambda_1 t} - e^{\lambda_2 t}}{\lambda_1 - \lambda_2}. \tag{1.84}$$

Un autre exemple qui sera utile par la suite est celui où $f_n(t) = \frac{H(t)t^{n-1}}{\Gamma(n)} e^{\lambda t}$, et pour lequel on obtient que :

$$\frac{H(t)t^{n-1}}{\Gamma(n)} e^{\lambda t} * \frac{H(t)t^{m-1}}{\Gamma(m)} e^{\lambda t} = \frac{H(t)t^{n+m-1}}{\Gamma(n+m)} e^{\lambda t} \tag{1.85}$$

Il s'agit ici d'une application directe de (1.83), suivie du changement de variable $\theta = t\omega$, et utilisant la propriété :

$$\int_0^1 (1-\omega)^{(n-1)} \omega^{(m-1)} d\omega = \frac{\Gamma(n)\Gamma(m)}{\Gamma(n+m)}. \tag{1.86}$$

Multiplication par t^n, e^{at}, et convolution : La multiplication par un polynôme ou une exponentielle, combinée avec la convolution peut avoir des propriétés intéressantes dans certaines applications. On établit ainsi que si l'une des distibution S ou T est à support compact, alors :

$$t^n(S * T) = \sum_0^n C_n^k (t^k S) * (t^{n-k} T) \tag{1.87}$$

$$e^{at}(S * T) = (e^{at} S) * (e^{at} T) \tag{1.88}$$

Régularisation et continuité de la convolution : La convolution par une fonction a un effet régularisant (lissant), ce qui se traduit notamment par :

Théorème 1.4.2. *Lorsqu'il existe, le produit de convolution d'un distribution T et d'une fonction ψ indéfiniment dérivable est une fonction h donnée par*

$$h(t) = \langle T(u), \psi(t-u) \rangle . \tag{1.89}$$

Cette fonction est elle même indéfiniment dérivable lorsque ψ est à support borné, et de dérivée :

$$h^{(m)}(t) = \left\langle T(u), \psi^{(m)}(t-u) \right\rangle . \tag{1.90}$$

On établit de même que la convolution est une opération continue dans $\mathcal{D}'$ dans le sens où :

Théorème 1.4.3. *S étant une distribution fixée, $T_\alpha \to T$ entraîne $T_\alpha * S \to T * S$ dans chacune des circonstances suivantes : (a) les distributions T_α ont leur support contenu dans un ensemble compact fixe, indépendant de α, (b) S est à support compact, (c) T_α et S sont des éléments de $\mathcal{D}'_+$.*

On en déduit la propriété dite de densité de $\mathcal{D}$ dans $\mathcal{D}'$, à savoir que toute distribution est limite dans $\mathcal{D}'$ d'une suite de fonctions appartenant à $\mathcal{D}$.

Algèbres de convolution

On appelle algèbre sur le corps des nombres réels ou complexes, un ensemble A muni des trois opérations : somme, produit par un scalaire, et produit, ayant les propriétés :

- A muni de la somme et du produit par un scalaire est un espace vectoriel,
- le produit est une application bilinéaire de $A \times A$ dans A,
- le produit est associatif.

On s'intéresse ici aux algèbres pour les opérations : somme de deux distributions, produit d'une distribution par un scalaire, et convolution de deux distributions. On doit donc vérifier que le produit de convolution de deux distributions de A est encore un élément de A, et que la convolution est associative lorsqu'on la restreint aux éléments de A. L'intérêt de ces algèbres réside dans le fait que si $\delta \in A$ (A possède donc un élément unité), et si une distribution T admet une inverse (notée par la suite T^{*-1} ou T^{-1} lorsqu'il n'y a pas de risque de confusion), c'est à dire vérifie

$$T * T^{*-1} = \delta, \tag{1.91}$$

cette inverse est unique, et on saura résoudre toute les équations de convolutions (en X)

$$T * X = W, \tag{1.92}$$

où $W \in A$. On notera cependant que l'inverse de convolution n'existe pas toujours. C'est par exemple la cas où $T \in \mathcal{D}$ car dans ce cas nous avons vu que, quel que soit X, $T * X$ est une fonction indéfiniment dérivable qui ne saurait être égale à δ. L'associativité du produit de convolution permet aussi de montrer que si $T_1, \cdots, T_m$ admettent pour inverses $T_1^{-1}, \cdots, T_m^{-1}$, il vient :

$$(T_1 * \cdots * T_m)^{-1} = T_1^{-1} * \cdots * T_m^{-1}. \tag{1.93}$$

On distingue essentiellement deux algèbres que sont les :

- algèbre $\mathcal{E}'$ des distributions à support compact,
- algèbre $\mathcal{D}'_+$ des distributions à support contenu dans $[0, \infty)$.

Lorsqu'on s'intéresse particulièrement aux équations différentielles (ou aux dérivées partielles) à coefficients constants, l'algèbre $\mathcal{E}'$ présente un intérêt limité. On montre en effet que si $T \in \mathcal{E}'$ n'est pas de la forme $c\delta$, l'équation en X, $T * X = \delta$ n'admet aucune solution à support compact. On concentrera donc notre étude à l'algèbre $\mathcal{D}'_+$.

Exemples simples d'application :

- Soit $W \in \mathcal{D}'_+$ et soit à résoudre dans $\mathcal{D}'_+$ l'équation $X' = W$. Cette relation s'écrit encore :
$$X' = \delta' * X = W. \tag{1.94}$$
Sachant par ailleurs que $\delta' * H = \delta$, la fonction de Heaviside H est donc bien l'inverse de convolution de δ', si bien la solution unique X est donnée par $X = H * W$. En d'autres termes, la seule primitive de W est $H * W$.

- Il est aisé de vérifier que dans $\mathcal{D}'_+$, l'inverse de $\delta' - \lambda\delta$, $\lambda \in \mathbb{C}$, est donnée par $H(t)e^{\lambda t}$. Il est en effet simple de montrer que $(H(t)e^{\lambda t})' = \delta + \lambda H(t)e^{\lambda t}$, pour obtenir :
$$(\delta' - \lambda\delta) * H(t)e^{\lambda t} = \delta. \tag{1.95}$$

Equations différentielles Toute équation différentielle à coefficients constants admet la représentation équivalente :

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0 y = f, \tag{1.96}$$
$$P(\delta) * y = f, \tag{1.97}$$

où $P(\delta)$ est le polynôme symbolique en δ :

$$\begin{aligned} P(\delta) &= \delta^{(n)} + a_{n-1}\delta^{(n-1)} + \cdots + a_0\delta, \\ &= (\delta^{(1)} - \lambda_0\delta) * \cdots * (\delta^{(1)} - \lambda_n\delta), \end{aligned} \tag{1.98}$$

et où $\lambda_0, \cdots, \lambda_n$ sont les racines, distinctes ou non, du polynôme $z^n + a_{n-1}z^{n-1} + \cdots + a_0$. La résolution de l'équation différentielle se ramène donc à la recherche de l'inverse de convolution de $P(\delta)$, soit en utilisant (1.93), à la recherche de des inverses de $(\delta^{(1)} - \lambda_i\delta)$. Cette inverse a déjà été établit en (1.95), si bien que :

$$(P(\delta))^{*-1} = H(t)e^{\lambda_0 t} * \cdots * H(t)e^{\lambda_n t}. \tag{1.99}$$

On notera également que la présence de racines multiples $\lambda_0 = \cdots = \lambda_m = \lambda$ conduit, en accord avec la relation (1.85), à :

$$((\delta^{(1)} - \lambda\delta)^m)^{-1} = \frac{H(t)t^{m-1}}{(m-1)!}e^{\lambda t} \tag{1.100}$$

Exemple : filtre RC passe-bas. En supposant la capacité initialement non chargée, le courant $i(t)$ du circuit est lié à la force électromotrice de charge $e(t)$ par la relation :

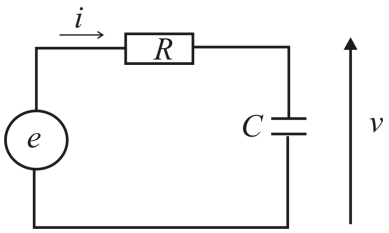


Fig. 1.2: Circuit RC.

$$R\,i(t) + \frac{1}{C}\int_0^t i(\theta)d\theta = e(t), \quad t \geq 0. \tag{1.101}$$

Au sens des distributions dans $\mathcal{D}'_+$, cela se traduit par :

$$(R\delta + \frac{1}{C}H) * i = e. \tag{1.102}$$

Si la mesure consiste en la tension $v(t) = \frac{1}{C}\int_0^t i(\theta)d\theta$ aux bornes de la capacité, cela s'écrit $v = \frac{1}{C}H * i$, soit $i = C\delta' * v$ et par suite :

$$(RC\delta' + \delta) * v = T * v = e. \tag{1.103}$$

Il suffit donc de savoir inverser la distribution $T = (RC\delta' + \delta)$ pour retrouver l'expression de $v(t)$ pour toute entrée $e(t)$. Il vient :

$$T^{-1} = \frac{1}{RC}He^{-\frac{t}{RC}}. \tag{1.104}$$

La distribution ainsi obtenue porte le nom de réponse impulsionelle du système, dans le sens ou elle correspond à la solution de (1.103) lorsque l'entrée e est

une distribution (impulsion) de Dirac. A titre d'exemple, si $e(t)$ consiste en la fonction de Heaviside, $e = H$, il vient immédiatement, compte tenu de (1.84) avec $(\lambda_1, \lambda_2) = (0, -\frac{1}{RC})$:

$$v = T^{-1} * H = \frac{1}{RC} H e^{-\frac{t}{RC}} * H = H(t) \left[1 - e^{-\frac{t}{RC}}\right]. \tag{1.105}$$

Prise en compte des conditions initiales Il arrive fréquemment que l'on cherche à résoudre, au sens des fonctions, une équation différentielle en y, d'ordre n, connaissant les conditions initiales $y(0), y^{(1)}(0), \cdots, y^{(n-1)}(0)$, notées par la suite $y_0, y_0^1, \cdots, y_0^{n-1}$. Il suffit pour cela multiplier l'équation par $H(t)$, de poser $z(t) = H(t)y(t)$, et de former l'équation différentielle vérifiée par z, en tenant compte de la formule des sauts présentée au paragraphe 1.3, ou plus directement par dérivation. Illustrons cette approche sur une équation différentielle du second ordre :

$$y^{(2)} + a_1 y^{(1)} + a_0 y = f. \tag{1.106}$$

Compte tenu que $z^{(1)} = Hy^{(1)} + y_0\delta$, $z^{(2)} = Hy^{(2)} + y_0\delta^{(1)} + y_0^1\delta$, la substitution dans la relation précédente (multipliée par H) conduit à :

$$z^{(2)} + a_1 z^{(1)} + a_0 z = Hf + y_0\delta^{(1)} + y_0^1\delta + a_1 y_0\delta. \tag{1.107}$$

Il suffit alors d'inverser l'opérateur $P(\delta) = \delta^{(2)} + a_1\delta^{(1)} + a_0\delta$ pour obtenir :

$$z = (P(\delta))^{-1} * \left[Hf + y_0\delta^{(1)} + y_0^1\delta + a_1 y_0\delta\right]. \tag{1.108}$$

Equations intégrales Dans certaines équations intégrales, nous pouvons être amené à inverser dans $\mathcal{D}'_+$ des éléments de le forme $\delta + K$ où K est une fonction de $\mathcal{D}'_+$. L'existence ainsi qu'une formulation de cette inverse sont données par la :

Proposition 1.4.1. *Si $K \in \mathcal{D}'_+$ est une fonction localement sommable et localement bornée, alors $\delta + K$ est inversible dans $\mathcal{D}'_+$ et $(\delta + K)^{*-1} = \delta + S$, où S est la somme de la série de terme général $(-1)^n K^{*n}$.*

Ainsi les équations de Volterra de la forme, pour $t \geq 0$:

$$y(t) + \int_0^t K(t-\theta)y(\theta)d\theta = g(t), \tag{1.109}$$

où K et g sont localement sommables, s'écrivent en terme de distributions :

$$(\delta + K) * y = g, \tag{1.110}$$

et la solution recherchée admet la forme $y = g + S * g$, soit, lorsque S est une fonction :

$$y(t) = g(t) + \int_0^t S(t-\theta)\, g(\theta)(d\theta. \tag{1.111}$$

Ce résultat peut également être appliqué aux équations différentielles avec retards, comme le montre l'exemple simple suivant :

$$y^{(1)}(t) + y(t-\tau) = g(t). \tag{1.112}$$

En faisant abstraction des conditions initiales, et en notant $H_\tau = H(t-\tau)$ et $\delta_\tau = \delta(t-\tau)$, cette équation se réecrit :

$$(\delta^{(1)} - \delta_\tau) * y = \delta^{(1)} * (\delta - H_\tau) * y = g, \tag{1.113}$$

et l'application de la proposition précédente conduit à la solution :

$$\begin{aligned} y &= (\delta - H_\tau)^{*-1} * H * g \\ &= (\delta - H_\tau + H_\tau^{*2} + \cdots + (-1)^n H_\tau^{*n} + \cdots) * H * g. \end{aligned} \tag{1.114}$$

A noter que chaque terme de la série admet le formulation explicite ci-dessous de support $(n\tau, \infty)$:

$$\begin{aligned} H_\tau^{*n} &= \underbrace{[(\delta_\tau * H) \cdots * (\delta_\tau * H)]}_{\text{n fois}} = \delta_{n\tau} * H^{*n} \\ &= H(t-n\tau)\frac{(t-n\tau)^{n-1}}{(n-1)!}. \end{aligned} \tag{1.115}$$

Equations matricielles La manipulation de systèmes d'équations de convolution se plie aux mêmes règles que celles du produit matriciel ordinaire, dans lequel la multiplication est remplacée par le produit de convolution. Dans ce cadre nous aurons la

Proposition 1.4.2. *Une matrice A $(n \times n)$ de convolution est inversible si et seulement si son déterminant Δ est inversible au sens de la convolution dans $\mathcal{D}'_+$. Son inverse s'obtient en convolant Δ^{*-1} par la matrices des cofacteurs.*

A titre d'exemple, reprenons l'exemple du second ordre décrit en (1.106), avec $f = b\,u$, b constant et u fonction de $\mathcal{D}'_+$ désignant la commande appliquée au système. Cette équation admet la représentation dite d'état, qui au sens des fonction s'écrit :

$$x^{(1)} = \begin{pmatrix} 0 & 1 \\ -a_0 & -a_1 \end{pmatrix} x + \begin{pmatrix} 0 \\ b \end{pmatrix} u, \qquad \text{avec} \quad x = \begin{pmatrix} y \\ y^{(1)} \end{pmatrix}. \tag{1.116}$$

Notant comme précédemment le vecteur $z = Hx$, il vient $z^{(1)} = Hx^{(1)} + x_0$ avec $x_0 = (y(0), y^{(1)}(0))^t$, et la substitution dans la représentation d'état se lit :

$$z^{(1)} = \begin{pmatrix} 0 & \delta \\ -a_0\delta & -a_1\delta \end{pmatrix} * z + \begin{pmatrix} 0 \\ b\delta \end{pmatrix} * u + x_0\delta \tag{1.117}$$

$$\text{soit} \qquad \underbrace{\begin{pmatrix} \delta^{(1)} & -\delta \\ a_0\delta & \delta^{(1)} + a_1\delta \end{pmatrix}}_{A} * z = \underbrace{\begin{pmatrix} 0 \\ b\delta \end{pmatrix} * u + x_0\delta}_{B} \tag{1.118}$$

Il suffit donc de savoir inverser la matrice A au sens de la convolution. Son déterminant est le polynôme de dérivation $\Delta = \delta^{(2)} + a_1\delta^{(1)} + a_0\delta$, déjà rencontré au paragraphe précédent. Il est toujours inversible et la solution z s'exprime par :

$$z = \Delta^{*-1} * \begin{pmatrix} \delta^{(1)} + a_1\delta & \delta \\ -a_0\delta & \delta^{(1)} \end{pmatrix} * \left[\begin{pmatrix} 0 \\ b\delta \end{pmatrix} * u + x_0\delta \right]. \tag{1.119}$$

Cette démarche ne se limite pas aux équations différentielles et peut s'étendre aux équations différentielles intégrales, faisant intervenir des termes de retard, ponctuels ou distribués. Considérons par exemple le système ci-dessous, pour lequel nous ferons abstraction des conditions initiales, un peu plus délicates dans le cadre général :

$$\begin{cases} \dot{y}_1(t) = y_1(t) + \int_{-1}^{0} y_2(t+\theta)d\theta \\ \\ \dot{y}_2(t) = y_1(t-1) + y_2(t) + \int_{-1}^{0} u(t+\theta)d\theta \end{cases} \tag{1.120}$$

En notant $\pi(t) = H(t) - H(t-1)$, il est aisé de constater que les termes faisant intervenir les intégrales se traduisent également sous forme de produit de convolution. On obtient alors la description matricielle :

$$\begin{pmatrix} \delta' - \delta & -\pi \\ -\delta_1 & \delta' - \delta \end{pmatrix} * \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -\pi \end{pmatrix} * u \tag{1.121}$$

1.5 Transformées de Fourier et de Laplace

Transformée de Fourier

Rappelons tout d'abord la définition de la transformée de Fourier de fonctions. Si f est une fonction Lebesgue intégrable sur $\mathbb{R}$, on définit sa transformée de Fourier, notée indifféremment $\mathcal{F}[f](\nu)$ ou $\hat{f}(\nu)$, par

$$\mathcal{F}[f](\nu) = \hat{f}(\nu) = \int f(t)\, e^{-2i\pi\nu t} dt. \tag{1.122}$$

Au vu de cette définition, nous sommes tentés de définir, pour toute $\varphi \in \mathcal{D}$, la transformée d'une distribution T par :

$$\langle \mathcal{F}[T], \varphi \rangle = \langle T, \mathcal{F}[\varphi] \rangle, \qquad \forall \varphi \in \mathcal{D}. \tag{1.123}$$

Cette définition ne peut pas être retenue car le membre de droite ne définit pas un distribution. Plus précisément, φ étant à support borné, on montre que $\mathcal{F}[\varphi]$ n'est jamais à support borné. Aussi sommes amené à considérer une classe plus restreinte de distributions (espace $\mathcal{S}' \subset \mathcal{D}'$), et, par conséquent, une classe plus large pour les fonctions tests φ (espace $\mathcal{S} \supset \mathcal{D}$). De tels espaces ont déjà été évoqués plus haut, rappelés plus précisément ici :

Définition 1.5.1 (Espaces $\mathcal{S}$)**.** On appelle $\mathcal{S}$ l'espace des fonctions indéfiniment dérivables, décroissant ainsi que toute leurs dérivées, quand $t \to \infty$, plus vite que toute puissance de $\frac{1}{t}$.

Les fonctions $\mathcal{S} \subset \mathcal{D}$ ne sont donc pas nécessairement à support compact, comme celle de $\mathcal{D}$. Leur décroissance à l'infini se traduit par, $\forall m, p$ entiers non négatifs,

$$\sup_t |t^m \varphi^{(p)}(t)| < \infty. \tag{1.124}$$

Définition 1.5.2 (Espaces $\mathcal{S}'$)**.** On appelle distribution tempérée toute forme linéaire continue sur $\mathcal{S}$.

Les distributions tempérées forment un espace vectoriel $\mathcal{S}'$ qui est un sous espace de $\mathcal{D}'$. La plupart des fonctions que l'on rencontre en physique sont des distributions tempérées. Il en est de même des distributions de Dirac et ses dérivées qui sont à support compact. Notons aussi que la dérivée d'une distribution tempérée, ainsi que la multiplication d'une distribution tempérée par un polynôme, forment des distributions tempérées. Des contre exemples sont donnés par des fonctions telles que e^t, e^{t^2}, qui elles n'appartiennent pas à $\mathcal{S}'$.

Définition 1.5.3. Si $T \in \mathcal{S}'$, on définit sa transformée de Fourier par :

$$\langle \mathcal{F}[T], \varphi \rangle = \langle T, \mathcal{F}[\varphi] \rangle \,, \qquad \forall \varphi \in \mathcal{S} \,. \tag{1.125}$$

On montre alors que $\mathcal{F}[T]$ est toujours une distribution tempérée si T l'est, et cette définition coïncide avec celle obtenue en (1.122) pour des distributions fonctions. Plus encore, dans le cas particulier où T est un distribution à support compact, on montre que sa transformée est la fonction indéfiniment dérivable :

$$\hat{T}(\nu) = \left\langle T(t), e^{-2i\pi\nu t} \right\rangle, \tag{1.126}$$

prolongeable, pour les valeurs complexes de ν en une une fonction holomorphe dans tout le plan complexe. Plus généralement, on notera la tendance selon laquelle plus la fonction $T(t)$ décroît à l'infini, et plus $\hat{T}(\nu)$ sera dérivable. Les tableaux ci-dessous fournissent quelques propriétés et exemples de transformées de Fourier de distributions singulières et régulières. Les propriétés relatives à la multiplication et au produit de convolution sont à considérer avec les réserves d'existence précédemment énoncés.

Dérivation	$\mathcal{F}[T'(t)] = 2i\pi\nu\hat{T}(\nu)$
Translation	$\mathcal{F}[T(t-a)] = e^{-2i\pi\nu\, a}\hat{T}(\nu)$
Changement d'échelle	$\mathcal{F}[T(at)] = \frac{1}{\lvert a\rvert}\hat{T}(\frac{\nu}{a})$
Convolution	$\mathcal{F}[S * T] = \mathcal{F}[S].\mathcal{F}[T]$
Produit	$\mathcal{F}[S.T] = \mathcal{F}[S] * \mathcal{F}[T]$

Quelques propriétés de la transformée de Fourier

$\mathcal{F}[1] = \delta$
$\mathcal{F}[e^{-2i\pi\,\nu_0\,t}] = \delta(\nu - \nu_0)$
$\mathcal{F}[\mathrm{sgn}(t)] = -\,\mathrm{pf}\,\frac{i}{\pi\nu}$
$\mathcal{F}[\delta] = 1$
$\mathcal{F}[\delta^{(n)}] = (2i\pi\nu)^n$
$\mathcal{F}[\delta(t-a)] = e^{-2i\pi\nu\,a}$

Exemples de transformées de Fourier

Transformée de Laplace

On procède ici aussi par analogie avec les transformées de Laplace des fonctions, et on se limite aux situations les plus fréquentes pour lesquelles la fonction $f(t)$ étudiée est nulle pour $t < 0$, donc appartient à $\mathcal{D}'_+$. On appelle transformée de Laplace de f la fonction $s \mapsto \mathcal{L}(s, f)$, notée aussi $\hat{f}(s)$ lorsqu'il n'y a pas de confusion possible avec la transformé de Fourier, définie dans $\mathbb{C}$ par :

$$\mathcal{L}(s, f) = \int_0^\infty f(t)\, e^{-st} dt. \tag{1.127}$$

Cette définition n'a pas toujours un sens. On appelle abscisse de sommabilité de $f(t)$ le réel a, borne inférieure de l'ensemble $\alpha = \mathrm{Re}(s)$ telle que $|f(t)e^{-\alpha t}|$ soit sommable. Dans ces conditions, $\mathcal{L}(s, f)$ est définie pour $\mathrm{Re}(s) > a$. On note plus précisément les cas de figures suivants :

- Si f est à support compact, $a = -\infty$,
- Si f est à décroissance rapide,$-\infty \leq a < 0$,
- Si f est tempérée, $a = 0$,
- Si f est à croissance rapide , $0 < a \leq \infty$.

Ce dernier cas montre que la transformée de Laplace de f peut exister tandis que sa transformée de Fourier n'est pas définie. Si a est fini, $s \mapsto \mathcal{L}(s, f)$ est une fonction holomorphe dans le demi plan $\mathrm{Re}(s) > a$. Cette définition est étendue aux distributions comme suit :

Définition 1.5.4. Soit T une distribution de $\mathcal{D}'_+$. On définit sa transformée de Laplace par la fonction $s \mapsto \mathcal{L}(s, T)$ définie dans $\mathbb{C}$ par :

$$\mathcal{L}(s, T) = \left\langle T(t), e^{-s\,t} \right\rangle. \tag{1.128}$$

Cette définition n'a de sens que si T vérifie également certaines conditions, en général satisfaites dans la majeure partie des applications. Plus précisément,

on montre que s'il existe un réel ζ tel la distribution $e^{-\zeta t}T$ soit tempérée, alors $\mathcal{L}(s,T)$ existe et définie une fonction analytique dans le demi plan $\mathrm{Re}(s) > \zeta$. On trouve dans certains ouvrages une autre définition de la transformée de Laplace d'une distribution T :

Définition 1.5.5. Soit T une distribution de $\mathcal{D}'_+$ qui est une dérivée $n^{\text{ième}}$ au sens des distributions d'une fonction continue $f \in \mathcal{D}'_+$, avec abscisse de sommabilité a. On définit sa transformée de Laplace par la fonction $\mathcal{L}(s,T)$ définie, pour $\mathrm{Re}(s) > a$, par :

$$\mathcal{L}(s,T) = s^n \, \mathcal{L}(s,f). \tag{1.129}$$

Les tableaux ci-dessous fournissent quelques propriétés et exemples de transformées de Laplace de distributions singulières et régulières. Comme pour la transformée de Fourier, Les propriétés relatives à la multiplication et au produit de convolution sont à considérer avec les réserves d'existence précédemment énoncés, de même que les abscisses de sommabilité doivent être précisées.

Dérivation	$\mathcal{L}(s,T') = s \; \mathcal{L}(s,T)$
Translation	$\mathcal{L}(s,T(t-a)) = e^{-a\,s} \, \mathcal{L}(s,T)$
Changement d'échelle	$\mathcal{L}(s,T(at)) = \frac{1}{\lvert a\rvert} \, \mathcal{L}(\frac{s}{a},T)$
Convolution	$\mathcal{L}(s,S*T) = \mathcal{L}(s,S).\,\mathcal{L}(s,T)$
Produit	$\mathcal{L}(s,e^{-at}T) = \mathcal{L}(s+a,T)$

Quelques propriétés de la transformée de Laplace

$$\begin{aligned}
&\mathcal{L}(s,\delta) = 1 \\
&\mathcal{L}(s,\delta^{(n)}) = s^n \\
&\mathcal{L}(s,\delta(t-a)) = e^{-a\,s} \\
&\mathcal{L}(s,H(t)) = \frac{1}{s} \\
&\mathcal{L}(s,H(t)e^{a\,t}\frac{t^{n-1}}{(n-1)!}) = \frac{1}{(s-a)^n}
\end{aligned}$$

Exemples de transformées de Laplace

On notera également ce cas particulier intéressant d'une multiplication d'une fonction f indéfiniment dérivable par un peigne de Dirac, conduisant à une série entière en la variable $z = e^{-s}$:

$$\begin{gathered}
f(t) \times \textstyle\sum_{n=0}^{\infty} \delta(t-n) = \sum_{n=0}^{\infty} f(n)\delta(t-n) \\
\downarrow \mathcal{L} \\
\textstyle\sum_{n=0}^{\infty} f(n)e^{-ns} = \sum_{n=0}^{\infty} f(n)z^n.
\end{gathered} \tag{1.130}$$

Cette série est appelée transformée en z de f et intervient fréquemment lorsqu'un signal est sujet à un échantillonnage .

Calcul symbolique Une application particulièrement pratique de la transformée de Laplace réside dans la résolution d'équations de convolutions qui se ramène au calcul d'inverse de transformées de Laplace. Sous réserve d'existence, cela se traduit par :

$$a * y = b \quad \Rightarrow \quad \hat{y}(s) = \frac{\hat{b}(s)}{\hat{a}(s)}. \tag{1.131}$$

En particulier, pour les équations différentielles à coefficients constants, $a = P(\delta) = \sum a_i \delta^{(i)}$ est un polynôme de dérivation dont l'inverse de convolution admet pour transformée une fraction rationnelle pouvant être décomposée en élément simples :

$$\mathcal{L}(s, a^{*-1}) = \frac{1}{P(s)} = \sum_k \frac{\beta_k}{(s-\lambda_k)^{\alpha_k}}. \tag{1.132}$$

Les inverses de chaque élément étant connus (cf. Exemples), il vient :

$$a^{*-1} = H(t) \sum_k \beta_k e^{-\lambda_k t} \frac{t^{\alpha_k - 1}}{(\alpha_k - 1)!}. \tag{1.133}$$

On obtient ainsi l'expression de la solution élémentaire de l'équation, c'est à dire celle de $a * y = \delta$. Cette décomposition en éléments simples peut également s'appliquer à la fraction $\hat{b}(s)/\hat{a}(s)$ lorsque le second membre b admet une transformée de Laplace analytique, conduisant ainsi directement à l'expression de la solution y. Ainsi, l'exemple traité en (1.119) avec $u = H$ et les valeurs numériques $a_1 = 2$, $a_0 = 1$, $x_0 = 0$, $b = 1$, se traduit par :

$$\hat{z}(s) = \begin{pmatrix} \dfrac{1}{s(s+1)^2} \\ \dfrac{1}{(s+1)^2} \end{pmatrix} \quad \Rightarrow \quad z = H(t) \begin{pmatrix} 1 - e^{-t} - te^{-t} \\ te^{-t} \end{pmatrix}. \tag{1.134}$$

Cette démarche peut aussi s'appliquer à certaines équation intégrales qui se traduisent sous forme d'équation de convolution, comme par exemple la recherche d'une solution y à support positif et vérifiant, pour $t \geq 0$:

$$\int_0^t y(\theta) \sin(t-\theta) d\theta = t^2 \quad \Rightarrow \quad (H(t) \sin t) * y = H(t) t^2 \tag{1.135}$$

et qui fournissent par transformée de Laplace et inversion

$$\frac{1}{s^2+1} \hat{y}(s) = \frac{2}{s^3}, \qquad \hat{y}(s) = \frac{2}{s} + \frac{2}{s^3}, \qquad y(t) = H(t)(2 + t^2). \tag{1.136}$$

1.6 Travaux Dirigés

Note : cette section a été rédigée par Kaouther Ibn Taarit et Lotfi Belkoura

Objectifs

Mettre en évidence les propriétés issues de la combinaison de la multiplication et du produit de convolution. Ces propriétés fournissent un cadre plus général aux approches du type intégration par parties et conduisent à des techniques simple d'estimation de paramètres.

Exercice 1 En s'inspirant de la preuve établie en annexe pour la multiplication d'un produit de convolution de deux distributions S et T par des fonctions exponentielles, établir la relation (1.137) et la généraliser à (1.138) :

$$t(S * T) = (tS) * (T) + (S) * (tT) \tag{1.137}$$

$$t^n(S * T) = \sum_0^n C_n^k (t^k S) * (t^{n-k} T), \tag{1.138}$$

où C_n^k désigne les coefficients du développement binomial. En déduire la relation :

$$t\, y^{(1)} = -z_0 + z_1^{(1)}, \qquad \text{avec} \qquad z_i = t^i\, y,$$

et donner l'expression correspondante au produit $t^2\, y^{(2)}$. En supposant $y \in \mathcal{D}'_+$ et en notant H l'échelon de Heaviside, à quelles simples manipulations se résument alors $H * (t\, y^{(1)})$ et $H * H * (t^2\, y^{(2)})$.

Exercice 2 On considère un système linéaire du premier ordre d'entrée $u(t)$, de sortie $y(t)$, d'état initial $y(0)$ et de fonction de transfert $F(s) = \frac{1}{s+a}$. Pour une entrée en échelon retardé $u(t) = H(t-\tau)$, et en considérant y dans $\mathcal{D}'_+$, donner l'équation différentielle décrivant ce processus au sens des distributions. En déduire que la réponse de ce système vérifie la relation :

$$t(t-\tau)\,[y^{(2)} + a\, y^{(1)}] = 0.$$

En supposant a connu, discuter de la possibilité d'obtenir une estimation du retard τ ne nécessitant pas la connaissance des dérivées de y. (On pourra s'inspirer des résultats de l'exercice précédent, et on examinera avec soin les supports des quantités formulées).

1.7 Travaux Pratiques

L'objectif de ces manipulation consiste à illustrer numériquement quelques propriétés fondamentales sur les distributions. Par la suite, le produit de convolution $a * b$ pourra être approché par la fonction $\text{conv}(a, b).t_e$ de Matlab, t_e désignant le pas d'échantillonnage choisi.

Objectif 1

Pour un produit de convolution de deux distributions, nous disposons de l'inclusion suivante dans laquelle le membre de droite (somme de Minkowsky) se lit :

$$\operatorname{supp} S * T \subset \operatorname{supp} S + \operatorname{supp} T = \{x + y;\, x \in \operatorname{supp} S,\, y \in \operatorname{supp} T\}\,. \quad (1.139)$$

Cette inclusion est en général au sens strict. Un résultat plus précis existe (Théorème des supports), dans lequel $\operatorname{conv}(\operatorname{supp} X)$ désigne l'enveloppe convexe du support de X.

$$\operatorname{conv}(\operatorname{supp} S * T) = \operatorname{conv}(\operatorname{supp} S) + \operatorname{conv}(\operatorname{supp} T).$$

Manipulation On note $\chi_{[a,b]}$ la fonction caractéristique de l'intervalle $[a, b]$. Effectuer le produit de convolution $\chi_{[2,3]} * \chi_{[4,5]}$. Représenter sur un même graphique les trois courbes et vérifier (graphiquement) la propriété sur les supports. Reprendre cette manipulation avec le produit de convolution $(\chi_{[0,0.5]} + \chi_{[2,3]}) * \chi_{[4,5]}$.

Objectif 2

Une suite de distribution T_n converge dans $\mathcal{D}'$ vers T lorsque la suite de nombres complexes $\langle T_n, \varphi\rangle$ converge dans $\mathbb{C}$ vers $\langle T, \varphi\rangle$ pour toute $\varphi \in \mathcal{D}$. En particulier, la distribution δ peut être aussi vue comme limite (dans $\mathcal{D}'$) de fonctions sommables. Ainsi, une suite de $f_k \geq 0$ localement sommables telles que $\int f_k(x)dx = 1$ et vérifiant $f_k(x) \to 0$ uniformément dans tout ensemble $0 < a < |x| < 1/a$, tend vers δ.

Manipulation Pour différentes valeurs de k entier, représenter graphiquement les fonctions γ_k et ζ_k ci-dessous :

$$\Pi(t) = \begin{cases} 0 & |t| \geq \frac{1}{2} \\ 1 & |t| < \frac{1}{2} \end{cases} \qquad \gamma_k(t) = k\Pi(kt) \qquad \zeta_k(t) = ke^{\frac{1}{k^2t^2-1}} \quad (1.140)$$

Réaliser ensuite les produits de convolutions $\gamma_k * y$ et $\zeta_k * y$ pour une fonction continue arbitraire et comparer les courbes obtenues avec la fonction y. Conclure.

Objectif 3

On établit que si une suite de distribution T_n converge dans $\mathcal{D}'$ vers T, alors les dérivées $T_n^{(m)}$ converge dans $\mathcal{D}'$ vers $T^{(m)}$. La dérivation est une opération linéaire et continue dans $\mathcal{D}'$ et on peut toujours permuter les signe de dérivation et de limite (propriété plus simple que pour les fonctions).

Manipulation Calculer au sens des distributions les dérivées $\dot{\gamma}_k$ et $\dot{\zeta}_k$ des fonctions γ_k et ζ_k précédentes. Réaliser les produits de convolutions $\dot{\gamma}_k * y$ et $\dot{\zeta}_k * y$ et comparer les courbes obtenues avec la fonction $\dot{y}$ (calculée ou approchée par les fonctions diff ou gradient de Matlab). Conclure.

Annexe

Rappel de formules sur la multiplication et la dérivation :

$$t^l\,\delta^{(n)} = 0 \text{ pour } l > n, \quad \text{et} \quad t^l\,\delta^{(n)} = (-1)^l\,\frac{n!}{(n-l)!}\,\delta^{(n-l)} \text{ pour } l \leq n.$$

$$\left\langle e^{at}(S*T)_t, \varphi(t)\right\rangle = \left\langle (S*T)_t, e^{at}\varphi(t)\right\rangle = \left\langle S_\zeta, \left\langle T_\eta, e^{a\zeta+a\eta}\varphi(\zeta+\eta)\right\rangle\right\rangle =$$
$$\left\langle S_\zeta, \left\langle e^{a\zeta+a\eta}T_\eta, \varphi(\zeta+\eta)\right\rangle\right\rangle = \left\langle e^{a\zeta}S_\zeta, \langle e^{a\eta}T_\eta, \varphi(\zeta+\eta)\rangle\right\rangle = \left\langle e^{at}S * e^{at}T, \varphi\right\rangle.$$

1.8 Bibliographie

[1] Boccara, N.: *Distributions.* Mathématiques pour l'ingénieur, Ellipses, 1997.

[2] Demengel, F. et G. Demengel: *Mesures et distributions. Théorie et illustration par les exemples.* Ellipses, 2000.

[3] Demengel, G.: *Transformation de Laplace, Théorie et illustration par les exemples.* Ellipses, 2002.

[4] Dupraz, J.: *La théorie des distributions et ses applications.* Cepadues-Editions, 1977.

[5] F. Hirsh, F. et G. Lacombe: *Elements d'analyse fonctionnelle.* Masson, 1997.

[6] Lacroix-Sonier, M T.: *Distributions, Espaces de Sobolev, Applications.* Ellipses, 1998.

[7] Petit, R.: *l'outil mathématique.* Enseignement de la physique, Masson, 1995.

[8] R. Dautray, R. et J L. Lions: *Analyse mathématique et calcul numérique, vol. 4, méthodes variationnelles.* Masson, 1988.

[9] Rodier, F.: *Distributions et Transformation de Fourier.* Ediscience international, 1993.

[10] Schwartz, L.: *Théorie des distributions, 2^{nd} ed.* Hermann, Paris, 1966.

[11] Yger, A.: *Analyse complexe et distributions.* Ellipses, 2001.

[12] Zuily, C.: *Elements de distributions et d'équations aux dérivées partielles.* Dunod, 2002.

2 | Optimisation et LMI

M. Dambrine[1]

[1]LAMIH, ENSIAME - Université de Valenciennes et du Hainaut-Cambrésis, 59313 Valenciennes cedex 9, France. *E-mail* :
`Michel.Dambrine@univ-valenciennes.fr`

2.1 Généralités

Notions de base et notations

Minimiser les coûts, maximiser le rendement, trouver le meilleur modèle possible, ... toutes ces expressions montrent à quel point l'optimisation a une place fondamentale dans les sciences de l'ingénieur.

Les données d'un *problème d'optimisation* de dimension finie sont les suivantes :

- un vecteur x composé des *variables de décision* $x_1, \ldots, x_n$;
- un ensemble K représentant l'ensemble des vecteurs x correspondant à des valeurs réalistes des variables de décision. Il peut être décrit par un ensemble d'égalités ou d'inégalités devant être vérifié par la solution. K sera appelé dans la suite *ensemble des contraintes* ou ensemble admissible ;
- enfin, une fonction J définie au moins sur K et à valeurs réelles représentant la quantité que l'on cherche à optimiser. La fonction J est appelée le *critère*, le *coût* ou la *fonction-objectif*

Le problème d'optimisation (formulé ici sous forme d'un problème de minimisation[1]) consiste à trouver — s'il en existe un — un élément $x_{\min}$ de K minimisant la fonction J sur K, c'est-à-dire, tel que

$$J(x_{\min}) \leq J(x) \quad \forall x \in K, \tag{2.1}$$

ou de manière équivalente

$$x_{\min} \in K \quad \text{et} \quad J(x_{\min}) = \inf_{x \in K} J(x).$$

[1]pour la recherche d'un maximum, il suffit de remplacer J par son opposé $-J$

Un tel élément est appelé *minimum global* du critère J sur K.

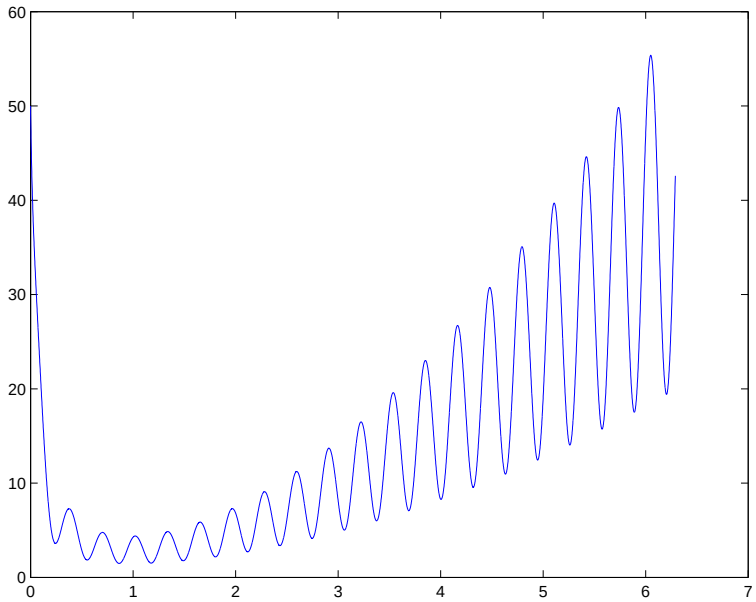


Fig. 2.1: une partie du graphe de $(\frac{1}{0.04+x} + 0.5x^2) * (2 + \sin(20x))$

La caractérisation d'un tel élément pouvant être délicate (cf. figure 2.1), on se contente souvent en pratique de la recherche d'un *minimum local*, c'est-à-dire, un élément $x_{\min}$ pour lequel l'inégalité (2.1) n'est vraie que pour des éléments $x \in K$ pris dans un voisinage V de $x_{\min}$:

$$J(x_{\min}) \leq J(x) \quad \forall x \in V \cap K, \tag{2.2}$$

Un minimum global ou local sera dit *strict* si l'inégalité large (2.1) ou (2.2) peut être remplacée par une inégalité stricte lorsque $x \neq x_{\min}$:

$$J(x_{\min}) < J(x) \quad \forall x \in V \cap K, x \neq x_{min}. \tag{2.3}$$

Le sous-ensemble K représente *l'ensemble des contraintes* du problème (ou ensemble admissible). Lorsque $K = \mathbb{R}^n$, on parle de problème sans contraintes, sinon de problème avec contraintes. Couramment, les contraintes sont définies par un système d'équations non linéaires

$$g_i(x) = 0 \quad \forall i \in \{1, \ldots, m\},$$

(contraintes de type égalité) ou un système d'inégalités

$$f_j(x) \leq 0 \quad \forall j \in \{1, \ldots, p\},$$

(contraintes de type inégalité) voire une combinaison de ces deux types.

Classification des problèmes d'optimisation

Suivant le type des variables de décision, la nature du critère J et de l'ensemble des contraintes K, on distingue différentes sortes de problèmes d'optimisation. Dans le cadre de ce cours, on supposera que les composantes de x sont réelles : on parle alors d'optimisation continue. Mais, on peut aussi considérer des problèmes pour lesquelles les variables sont entières, voire booléennes (0 ou 1) (optimisation combinatoire ou discrète) et d'autres où les variables sont des fonctions (optimisation en dimension infinie).

Dans le cadre même de l'optimisation continue, on catalogue les problèmes suivant les propriétés (linéarité, convexité, différentiabilité, ...) des fonctions J, g_i et f_i. On distingue ainsi

- les *problèmes linéaires* pour lesquels les fonctions J, g_i et f_j sont des fonctions affines de x ;
- les *problèmes non linéaires* (lorsqu'ils ne sont pas linéaires) ;
- les *problèmes quadratiques* lorsque J est une fonction quadratique, les fonctions g_i et f_j étant affines ;
- les *problèmes convexes* quand les fonction J et f_j sont convexes, les fonctions g_i étant affines ;
- les *problèmes différentiables* ou *non différentiables* suivant que les fonctions J, g_i et f_j sont différentiables (en un certain sens) ou non.

Existence d'un minimum

La première question que l'on peut se poser avant de rechercher la ou les solutions optimales d'un problème est de déterminer si le problème admet bien une solution optimale. Cette étape peut se révéler ardue même pour des problèmes catalogués comme simples tels les problèmes d'optimisation linéaire.

Le point de départ des critères d'existence d'un optimum repose sur un résultat classique d'Analyse : le *théorème de Weierstrass.*

Théorème 2.1.1 (Weierstrass)**.** *Toute fonction continue sur un compact non vide y admet un minimum.*

Rappelons qu'en dimension finie, les ensembles compacts sont les ensembles fermés et bornés. Il est possible de généraliser ce théorème au cas d'un sous-ensemble fermé $F \subset \mathbb{R}^n$ mais non nécessairement borné lorsque le critère à optimiser J est coercif : pour toute suite $(x_k)_{k\geq 0}$ d'éléments de F, on a

$$\lim_{k\to\infty} \|x_k\| = +\infty \Rightarrow \lim_{k\to\infty} J(x_k) = +\infty. \tag{2.4}$$

On convient ici que cette condition est systématiquement remplie lorsque l'ensemble F est borné. En effet, soit (x_k) une suite minimisante de J sur F, c'est-à-dire, une suite telle que $J(x_k)$ converge vers $\inf_{x\in F} J(x)$. La suite $J(x_k)$ étant majorée, la condition (2.4) (dite de coercivité) implique alors que (x_k) est une

suite bornée. D'après le théorème de Bolzano-Weierstrass, il est possible d'extraire de (x_k) une sous-suite (x_{n_k}) convergente vers un point $x_{\min}$ de $\mathbb{R}^n$. L'ensemble F étant fermé, $x_{\min}$ appartient à F. Si J est continue, alors on a :

$$J(x_{\min}) = \lim J(x_{n_k}) = \inf_{x \in F} J(x).$$

Théorème 2.1.2. *Soit F un ensemble fermé non vide de $\mathbb{R}^n$, et $J : F \to \mathbb{R}$ une fonction continue sur F et vérifiant la propriété de coercivité (2.4). Alors, la fonction J possède un minimum sur F.*

2.2 Minimisation sans contraintes

On suppose ici que $K = \mathbb{R}^n$ tout entier ou un ouvert $\mathcal{O}$ de $\mathbb{R}^n$.

On s'attaque dans cette partie au problème plus pratique de la caractérisation des points de minimum, c'est-à-dire établir des conditions qui sont automatiquement remplies dès lors qu'un point correspond à un minimum local de la fonction à minimiser. C'est à l'aide de ces conditions nécessaires que seront établis les algorithmes d'optimisation.

Notions préliminaires

Éléments de calcul différentiel

Soit $J : \mathbb{R}^n \to \mathbb{R}^m$, on dit que J *est différentiable en x* s'il existe une *application linéaire* L de $\mathbb{R}^n$ dans $\mathbb{R}^m$ telle que

$$J(x+h) = J(x) + L(h) + \|h\|\,\varepsilon(h), \qquad \lim_{h \to 0} \varepsilon(h) = 0$$

L'application L est *l'application dérivée de J en x* et est notée $J'(x)$.

Lorsque $m = 1$ (c-à-d, J à valeurs réelles), L est une forme linéaire (une application linéaire à valeurs réelles), il existe alors un vecteur noté $\nabla J(x) \in \mathbb{R}^n$ — *le vecteur gradient de J en x* — tel que

$$L(h) = J'(x)(h) = \nabla J(x)^T h, \quad \forall h \in \mathbb{R}^n$$

(produit scalaire de $\nabla J(x)$ et de h) et on a

$$\nabla J(x) = \begin{bmatrix} \frac{\partial J}{\partial x_1}(x) \\ \vdots \\ \frac{\partial J}{\partial x_n}(x) \end{bmatrix}.$$

La dérivée de J dans la direction h est alors donnée par

$$\lim_{\theta \to 0, \theta \in \mathbb{R}} \frac{J(x+\theta h) - J(x)}{\theta} = \nabla J(x)^T h. \tag{2.5}$$

Dans le cas général, la matrice associée à L (dans les bases canoniques de $\mathbb{R}^n$ et $\mathbb{R}^m$) est la Jacobienne de J en x que l'on notera encore $J'(x)$. Si

$$J(x) = \begin{pmatrix} J_1(x) \\ J_2(x) \\ \vdots \\ J_m(x) \end{pmatrix},$$

alors

$$J'(x) = \begin{pmatrix} \frac{\partial J_1}{\partial x_1}(x) & \dots & \frac{\partial J_1}{\partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial J_m}{\partial x_1}(x) & \dots & \frac{\partial J_m}{\partial x_n}(x) \end{pmatrix} = \begin{pmatrix} \nabla J_1(x)^T \\ \vdots \\ \nabla J_m(x)^T \end{pmatrix}$$

Pour une application $J : \mathbb{R}^n \to \mathbb{R}$ deux fois dérivable en un point x, on définit la matrice Hessienne $\nabla^2 J(x) \in \mathbb{R}^{n\times n}$ par :

$$\nabla^2 J(x) = \begin{pmatrix} \frac{\partial^2 J}{\partial {x_1}^2}(x) & \dots & \frac{\partial^2 J}{\partial x_1 \partial x_n}(x) \\ \vdots & & \vdots \\ \frac{\partial^2 J}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 J}{\partial {x_n}^2}(x) \end{pmatrix}.$$

C'est une matrice symétrique : $\partial^2 J(x)/\partial x_i \partial x_j = \partial^2 J(x)/\partial x_j \partial x_i$.

Le théorème suivant donne le développement limité à l'ordre 2 de J.

Théorème 2.2.1 (Formule de Taylor-Young)**.** *Si $J : \mathbb{R}^n \to \mathbb{R}$ est différentiable dans $\mathbb{R}^n$ et deux fois différentiable en x, alors*

$$J(x+h) = J(x) + \nabla J(x)^T h + \frac{1}{2} h^T \nabla^2 J(x) h + \|h\|^2 \varepsilon(h), \quad \lim_{h\to 0} \varepsilon(h) = 0.$$

Notion d'analyse convexe

Une partie C de $\mathbb{R}^n$ est dite *convexe* si tout segment d'extrémités prises dans C est inclus dans C, c'est-à-dire :

$$\boxed{\lambda x + (1-\lambda) y \in C \quad \forall x, y \in C, \forall \lambda \in [0,1],.}$$

Un point x est une combinaison convexe des points $x_1, \dots, x_k$ s'il existe des nombres réels $\lambda_1, \dots, \lambda_k$ tels que $\lambda_1, \dots, \lambda_k \geq 0$, $\lambda_1 + \dots + \lambda_k = 1$ et $x = \lambda_1 x_1 + \dots + \lambda_k x_k$. Par extension du résultat précédent, toute combinaison convexe d'éléments appartenant à un ensemble convexe C est également dans C.

Définition 2.2.1 (convexité, convexité stricte)**.** Une fonction $J : C \to \mathbb{R}$, où C est un ensemble convexe non vide de $\mathbb{R}^n$, est dite *convexe* si

$$\boxed{J(\lambda x + (1-\lambda)y) \leq \lambda J(x) + (1-\lambda)J(y) \quad \forall x, y \in C, \forall \lambda \in [0,1].}$$

J est dite *strictement convexe* si l'inégalité précédente est stricte lorsque $x \neq y$ et $\lambda \in]0,1[$.

Citons ici quelques propriétés des fonctions convexes :

Proposition 2.2.1.

- *L'ensemble des fonctions convexes sur $\mathbb{R}^n$ est un cône convexe : si J_1 et J_2 sont convexes alors la fonction $\lambda_1 J_1 + \lambda_2 J_2$ est également convexe quels que soient $\lambda_1, \lambda_2 \geq 0$.*
- *L'enveloppe supérieure $f(x) = \sup_{i \in I} f_i(x)$ d'une famille $(f_i)_{i \in I}$ de fonctions convexes est convexe.*
- *Si J est une fonction convexe sur un ensemble convexe C, tout point de minimum local de J sur C est un minimum global et l'ensemble des points de minimum est un ensemble convexe (éventuellement vide). Si de plus J est strictement convexe, alors il ne peut exister qu'au plus un point de minimum.*

Si la fonction J est différentiable, il est alors possible de caractériser les différentes notions de convexité à l'aide des propriétés suivantes :

1. J est convexe sur C si et seulement si

$$\boxed{J(x') \geq J(x) + \left\langle \nabla J(x), x' - x \right\rangle, \quad \forall x, x' \in C.} \tag{2.6}$$

2. J est strictement convexe sur C si et seulement si

$$\boxed{J(x') > J(x) + \left\langle \nabla J(x), x' - x \right\rangle, \quad \forall x, x' \in C,\ x' \neq x.}$$

Notons que la première propriété admet une interprétation géométrique simple : une fonction est convexe si et seulement si le plan tangent en chaque point est situé en dessous du graphe de la fonction.

Enfin, si la fonction J est deux fois différentiable, alors

3. J est convexe sur C si et seulement si $\nabla^2 J(x)$ est une matrice semi-définie positive, $\forall x \in C$.
4. J est strictement convexe sur C si et seulement si $\nabla^2 J(x)$ est une matrice définie positive, $\forall x \in C$.

Caractérisation d'un point de minimum

Conditions de minimalité du premier ordre

Soit $\mathcal{O}$ un ouvert de $\mathbb{R}^n$ et $x_{\min} \in \mathcal{O}$ un minimum local de la fonction J sur $\mathcal{O}$, la fonction J étant supposée différentiable au point $x_{\min}$. Le point $x_{\min}$ étant à l'intérieur de $\mathcal{O}$, pour tout vecteur $d \in \mathbb{R}^n$, on peut alors trouver un $\theta_{\max}$ tel que

$$x_{\min} + \theta d \in \mathcal{O} \quad \text{et} \quad J(x_{\min} + \theta d) - J(x_{\min}) \leq 0 \quad \forall \theta \in]-\theta_{\max}, \theta_{\max}[,$$

On a alors

$$\lim_{\theta \to 0} \frac{J(x_{\min} + \theta d) - J(x_{\min})}{\theta} \leq 0, \forall d \in \mathbb{R}^n,$$

c'est-à-dire, $\nabla J(x_{\min})^T d \leq 0$ pour tout vecteur $d \in \mathbb{R}^n$. On a donc

$$\nabla J(x_{\min}) = 0.$$

Théorème 2.2.2 (Condition nécessaire du premier ordre)**.** *Soit $J : \mathcal{O} \to \mathbb{R}$ une fonction différentiable au point $x_{\min} \in \mathcal{O}$. Si J admet un minimum local en $x_{\min}$, alors le gradient de J au point $x_{\min}$ est nul :*

$$\nabla J(x_{\min}) = 0 \qquad \textit{(équation d'Euler).}$$

Cette condition n'est bien sûr pas suffisante : le même raisonnement réalisé pour un maximum conduirait au même résultat. De plus, on sait que la dérivée de $J(x) = x^3$ (pour x réel) s'annule en $x = 0$ sans que ce point ne soit un minimum ou un maximum. Les points pour lesquels le gradient d'une fonction est nul sont appelés *points stationnaires* ou *points critiques* de cette fonction.

Comment détecter si un point stationnaire est bel et bien un minimum ? Il faut pour cela une étude plus poussée du comportement du critère J au voisinage de ce point. Ainsi, si J est convexe alors la CN de minimalité du premier ordre devient suffisante :

Théorème 2.2.3. *Soit $J : \mathcal{O} \to \mathbb{R}$ une fonction convexe différentiable sur l'ouvert convexe $\mathcal{O}$. Alors, J admet au point $x_{\min} \in \mathcal{O}$ un minimum local si et seulement si*

$$\nabla J(x_{\min}) = 0.$$

En effet, puisque J est convexe, on a

$$J(x) \geq J(x_{\min}) + \nabla J(x_{\min})^T (x - x_{\min}), \quad \forall x, x_{\min} \in \mathcal{O},$$

et donc, si $x_{\min}$ est un point stationnaire de J, on a

$$J(x) \geq J(x_{\min}) \quad \forall x \in \mathcal{O}.$$

Notons qu'alors, du fait de la convexité de J, $x_{\min}$ est *un minimum global* de J.

Si la fonction n'est pas convexe, on peut essayer le résultat suivant :

Théorème 2.2.4. *Si J est continue en x_{min} sur $\mathcal{O}$ et qu'il existe un voisinage V de ce point tel que :*

- *la fonction J est différentiable sur $V \setminus \{x_{min}\}$,*
- *l'inégalité $\nabla J(x)^T (x - x_{min}) \geq 0$ (resp. > 0) est vérifiée pour tout x dans $V \setminus \{x_{min}\}$,*

alors J admet an point $x_{min} \in \mathcal{O}$ un minimum local (resp. minimum local strict).

La démonstration de ce théorème repose sur l'utilisation du théorème des accroissements finis.

Conditions de minimalité du second ordre

On suppose, cette fois, la fonction $J : \mathcal{O} \to \mathbb{R}$ deux fois différentiable au point $x_{\min} \in \mathcal{O}$.

Si J admet en $x_{\min}$ un minimum local, alors on a $\nabla J(x_{\min}) = 0$ et la formule de Taylor-Young à l'ordre 2 s'écrit

$$0 \leq J(x_{\min} + \theta d) - J(x_{\min}) = \frac{\theta^2}{2} d^T \nabla^2 J(x_{\min}) d + \theta^2 \varepsilon(\theta),$$

avec $\lim_{\theta \to 0} \varepsilon(\theta) = 0$.

On a donc nécessairement

$$d^\top \nabla^2 J(x_{\min}) d \geq 0, \quad \forall d \in \mathbb{R}^n.$$

Théorème 2.2.5 (CN de minimalité du second ordre)**.** *Soit $J : \mathcal{O} \to \mathbb{R}$ une fonction deux fois différentiable au point $x_{\min} \in \mathcal{O}$. Si J admet un minimum local en $x_{\min}$, alors*

$$\nabla J(x_{\min}) = 0 \quad \textit{et} \quad \nabla^2 J(x_{\min}) \textit{ est semi-définie positive.}$$

Il est possible d'obtenir une condition suffisante de minimalité sous des hypothèses plus restrictives :

Théorème 2.2.6 (CS de minimalité du second ordre)**.** *Soient $J : \mathcal{O} \to \mathbb{R}$ et un point $x_{\min} \in \mathcal{O}$ où J est deux fois différentiable. Si*

$$\nabla J(x_{\min}) = 0 \quad \textit{et} \quad \nabla^2 J(x_{\min}) \textit{ est définie positive,}$$

alors J admet un minimum local strict au point $x_{\min}$.

En effet, soit $\lambda_{\min} > 0$ la plus petite valeur propre de $\nabla^2 J(x_{\min})$, alors, pour tout $d \in \mathbb{R}^n$, on a

$$d^T \nabla^2 J(x_{\min}) d \geq \lambda_{\min} \|d\|^2,$$

il suffit alors d'appliquer la formule de Taylor-Young à l'ordre 2 pour démontrer le résultat.

Algorithmes

Le but ici n'est pas de présenter un catalogue complet d'algorithmes mais quelques méthodes de base ainsi que leur principe-clé.

La plupart des algorithmes de résolution numérique des problèmes d'optimisation sont des procédés itératifs : on construit une suite (x_k) par une récurrence du type

$$x_{k+1} = x_k + \alpha_k h_k. \tag{2.7}$$

La direction h_k est choisie telle que

$$\boxed{\nabla J(x_k)^T h_k < 0}$$

(on dit alors que h_k est une *direction de descente*). Le coefficient α_k est pris tel que

$$J(x_k + \alpha_k h_k) < J(x_k),$$

la suite $(J(x_k))$ est décroissante, on parle d'*algorithmes de descente*.

Méthodes du gradient

On suppose ici la fonction J au moins une fois continûment différentiable sur $\mathbb{R}^n$. Dans le cas sans contraintes, un minimum ne peut être atteint qu'en un point stationnaire du critère J, c'est-à-dire en un point $x_{\min}$ tel que $\nabla J(x_{\min}) = 0$. Les techniques itératives de recherche des zéros d'une fonction peuvent donc être appliquées à la recherche d'un optimum : si la suite (x_k) vérifiant la récurrence

$$x_{k+1} = x_k - \alpha_k \nabla J(x_k), \quad \text{avec } \alpha_k \neq 0 \tag{2.8}$$

converge vers une limite $x_{\min}$, alors, par continuité de ∇J, on a nécessairement

$$\nabla J(x_{\min}) = 0.$$

Les méthodes du gradient reposent toutes sur le choix de l'opposé du gradient comme direction de descente : à chaque étape, on a

$$\boxed{h_k = -\nabla J(x_k).}$$

Suivant le choix du paramètre α_k, on distingue différentes méthodes :

- Méthode du gradient à pas fixe (α_k prend la même valeur à toutes les étapes).
- Méthode du gradient à pas optimal : à chaque étape, α_k est la solution du problème d'optimisation unidimensionnel

  $$\inf_{\alpha \in \mathbb{R}} f_k(\alpha), \tag{2.9}$$

  où $f_k(\alpha) = J(x_k + \alpha\, h_k)$.
  Remarquons que, de la condition d'optimalité $f_k'(\alpha_k) = 0$, l'on déduit la relation

  $$\nabla J(x_{k+1})^\top \nabla J(x_k) = 0,$$

  exprimant que deux directions de descente consécutives sont toujours orthogonales. La convergence de cet algorithme peut donc se révéler assez lente lorsque les valeurs propres de la matrice hessienne de J au point optimal ne sont pas du même ordre de grandeur (cf. Fig. 2.2),
- Méthodes du gradient à pas variable : en pratique, la résolution complète, à chaque étape, du problème (2.9) est trop coûteuse en temps de calcul. Pour éviter cela, la valeur du pas est choisie en considérant les éléments une suite géométrique de raison strictement inférieure à 1. La valeur du pas retenue étant la première à satisfaire certaines conditions (pour plus de détail, se référer à [8])

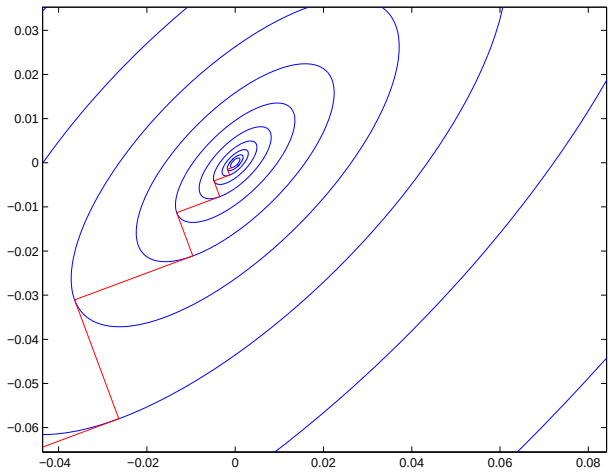


Fig. 2.2: Meth. du gradient à pas optimal

Méthode de Newton

Recherche des zéros d'une fonction : méthode de Newton-Raphson

Un algorithme couramment utilisé pour la recherche des zéros d'une fonction $F : \mathbb{R}^n \to \mathbb{R}$ est la méthode de Newton-Raphson. Le principe est le suivant : soit x_k un point donné de $\mathbb{R}^n$ constituant une estimation d'un zéro de F, on approche au voisinage de ce point, la fonction F par son développement limité au premier ordre :

$$\tilde{F}(x) = F(x_k) + F'(x_k)(x - x_k)$$

Le point suivant x_{k+1} est alors pris comme le zéro de $\tilde{F}$ (obtenu en résolvant le système d'équations linéaires $F'(x_k)(d_k) = -F(x_k)$, avec $x_{k+1} = x_k + d_k$) (la figure 2.3 illustre le cas d'une fonction F à une seule variable réelle x).

On peut montrer que si F est continûment différentiable au voisinage d'un de ces zéros x^* tel que $\det(F'(x^*)) \neq 0$, alors pour un point initial x_0 choisi suffisamment proche de x^*, la suite (x_k) converge vers x^*.

Application à l'optimisation : méthode de Newton

On suppose que la fonction J est deux fois continûment dérivable et que l'on peut calculer de manière relativement aisée la matrice Hessienne en tout point. Alors, la méthode de Newton-Raphson appliquée à la recherche d'un zéro de ∇J conduit à un algorithme itératif de la forme (2.7) où la direction de descente est

$$h_k = -(\nabla^2 J(x_k))^{-1} \nabla J(x_k). \tag{2.10}$$

Le paramètre α_k peut être pris égal à 1 (méthode de Newton pure) ou déterminé en résolvant le problème d'optimisation à une seule variable réelle (2.9).

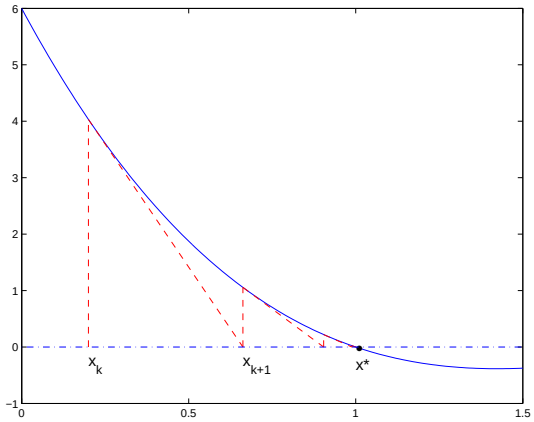


FIG. 2.3: Recherche d'un zéro de la fonction $F(x) = (1-x)(x-2)(x-3)$

Méthodes de quasi-Newton

La convergence de la méthode de Newton est rapide, mais la détermination de la matrice Hessienne de J et la résolution, à chaque étape, du système linéaire

$$\nabla^2 J(x_k)\, h_k = -\nabla J(x_k)$$

nécessitent un nombre important de calculs, ce qui rend cette technique inapplicable pour des problèmes de grande dimension. Un remède consiste alors à généraliser la formule (2.10) en choisissant pour direction de descente un vecteur de la forme

$$h_k = -H_k\, \nabla J(x_k),$$

où H_k est une matrice symétrique, définie positive. Les critères de sélection pour H_k sont :

1. Être calculable facilement et rapidement. On peut, pour cela, déterminer les matrices H_k à l'aide d'une récurrence de la forme

   $$H_{k+1} = H_k + \Delta_k,$$

   la matrice Δ_k pouvant être une fonction de H_k, x_{k+1}, x_k, ...

2. Converger vers l'inverse de la matrice Hessienne de J au point de minimum (au moins pour toute fonction J quadratique et elliptique) ou être telle que la direction de descente h_k converge vers la direction (2.10). Pour cela, il suffit que H_k vérifie l'égalité

   $$H_k\, (\nabla J(x_k) - \nabla J(x_{k-1})) = x_k - x_{k-1}.$$

Parmi les méthodes de quasi-Newton, l'*algorithme BFGS (Broyden, Fletcher, Goldfarb et Shanno)* est considéré comme le plus efficace. La mise à jour des matrices H_k se fait alors à l'aide de la relation de récurrence

$$H_{k+1} = H_k + (1 + \frac{g_k^T H_k g_k}{\delta_k^T g_k})\frac{\delta_k \delta_k^T}{\delta_k^T g_k} - \frac{\delta_k g_k^T H_k + H_k g_k \delta_k^T}{\delta_k^T g_k},$$

où $\delta_k = x_{k+1} - x_k$ et $g_k = \nabla J(x_{k+1}) - \nabla J(x_k)$.

2.3 Minimisation avec contraintes

Caractérisation du minimum

On considère ici le problème de la recherche du minimum d'une fonction $J : \mathbb{R}^n \to \mathbb{R}$ sur K, un sous-ensemble fermé de $\mathbb{R}^n$ représentant les contraintes.

Le raisonnement que l'on a effectué pour obtenir la condition nécessaire de minimalité dans le cas sans contraintes n'est plus valide ici : du fait des contraintes, le point de minimum peut être sur la frontière de K et seules les suites de la forme $x_{\min} + \theta_k d$ restant dans K pour θ_k suffisamment petit sont à considérer. On introduit alors la notion de *direction admissible.*

Définition 2.3.1. $h \in \mathbb{R}^n$ est une *direction admissible* de K au point $x \in K$ s'il existe une suite (x_k) d'éléments de K et une suite de réels strictement positifs (ε_k) telles que $\lim x_k = x$, $\lim \varepsilon_k = 0$ et $\lim(x_k - x)/\varepsilon_k = h$.

L'ensemble des directions admissibles au point $x \in K$ constitue un cône fermé de $\mathbb{R}^n$ noté $K(x)$.

Une autre façon d'interpréter le fait que d est une direction admissible de K en $x_{\min}$ est de dire qu'il existe une suite de vecteurs (d_k) convergeant vers d et une suite de réels positifs (θ_k) convergeant vers 0 telles que $x_{\min} + \theta_k d_k \in K$ pour tout K. Si $x_{\min}$ est un minimum local de J sur K, on a alors

$$\frac{J(x_{\min} + \theta_k d_k) - J(x_{\min})}{\theta_k} \geq 0$$

Si J est différentiable en $x_{\min}$, il suffit alors de passer à la limite lorsque $k \to \infty$ pour montrer le résultat suivant :

Théorème 2.3.1 (inéquation d'Euler)**.** *Soit K un sous-ensemble fermé de $\mathbb{R}^n$ et $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable au point $x_{\min} \in K$. Si J admet sur K un minimum local en $x_{\min}$, alors*

$$\nabla J(x_{\min})^\top d \geq 0 \qquad \forall d \in K(x_{\min}). \tag{2.11}$$

La difficulté dans l'application de ce théorème est de caractériser l'ensemble (ou un sous-ensemble) des directions admissibles. Avant de le faire pour des contraintes de type égalité et/ou inégalité, considérons deux cas particuliers simples à traiter.

– Si x est un point intérieur de K, alors $K(x) = \mathbb{R}^n$ et on retrouve la même condition nécessaire d'optimalité que dans le cas sans contrainte : $\nabla J(x_{\min}) = 0$.
– Si K est convexe, alors, étant donnés deux points x et y de K, le point $x + \varepsilon(y - x)$ est aussi dans K pour tout $\varepsilon \in [0, 1]$, autrement dit, $y - x$ est une direction admissible au point x. On a donc $K \setminus \{x\} \subset K(x)$. Il vient alors le corollaire suivant :

Corollaire 2.3.1. *Soit K un sous-ensemble convexe et fermé de $\mathbb{R}^n$ et $J : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable au point $x_{\min} \in K$. Si J admet sur K un minimum local en $x_{\min}$, alors*

$$\nabla J(x_{\min})^\top (x - x_{\min}) \geq 0 \qquad \forall x \in K.$$

Si de plus J est convexe, alors la réciproque est vraie. En effet, on a

$$J(x) \geq J(x_{\min}) + \nabla J(x_{\min})^\top (x - x_{\min}) \qquad \forall x \in K,$$

d'où

$$J(x) \geq J(x_{\min}).$$

La fonction J admet donc un minimum global en $x_{\min}$.

Cas des contraintes de type égalité

On suppose ici que l'ensemble des contraintes est de la forme suivante :

$$K = \{x \in \mathbb{R}^n : g_i(x) = 0 \quad \forall i \in \{1, \ldots, m\}\} \tag{2.12}$$

où $g_1, \ldots, g_m$ sont des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$.

Recherche des directions admissibles

Soit x un point de K — c-à-d. tel que $g_i(x) = 0 \quad \forall i \in \{1, \ldots, m\}$— et recherchons les directions admissibles à K en x.

Soit d une direction admissible à K en x, alors il existe $(d_k) \to d$ et $(\theta_k) \to 0$ telles que $x + \theta_k d_k \in K$, $\forall k$, c'est-à-dire :

$$g_i(x + \theta_k d_k) = 0 \quad \forall i \in \{1, \ldots, m\}$$

Si les fonctions g_i sont continûment différentiables en x, alors en développant au premier ordre et en prenant la limite lorsque $(d_k) \to d$, on montre que :

$$\nabla g_i(x)^\top d = 0 \quad \forall i \in \{1, \ldots, m\} \tag{2.13}$$

D'où

$$K(x) \subset T_K(x) \triangleq \left\{ d \in \mathbb{R}^n : \nabla g_i(x)^\top d = 0, \quad \forall i \in \{1, \ldots, m\} \right\} \tag{2.14}$$

Par extension, si les vecteurs $(g_i'(x))_{1 \leq i \leq m}$ sont *linéairement indépendants* — cas dit régulier — alors, à l'aide du théorème des fonctions implicites, il est possible de montrer que les deux ensembles $K(x)$ et $T_K(x)$ sont confondus :

$$\boxed{K(x) = \left\{ d \in \mathbb{R}^n : \nabla g_i(x)^\top d = 0, \quad \forall i \in \{1, \ldots, m\} \right\}}$$

Dans ce contexte, $K(x)$ est le *sous-espace tangent* à K en x et représente l'orthogonal du sous-espace vectoriel engendré par les $\nabla g_i(x)$.

Condition au premier ordre : règle de Lagrange

On se place dans le cas régulier. Si d est une direction admissible, alors $-d$ l'est aussi. La condition de minimalité (2.11) s'écrit alors

$$\nabla J(x_{\min})^\top d = 0, \quad \forall d \in K(x) = \left\{ h \in \mathbb{R}^n : \nabla g_i(x)^\top h = 0, \quad \forall i \in \{1, \ldots, m\} \right\}$$

On a donc $\nabla J(x_{\min}) \in K(x)^\perp$, puisque les vecteurs $(g_i'(x))$ pour $1 \leq i \leq m$ sont supposés linéairement indépendants, nécessairement $\nabla J(x_{\min})$ est un élément du sous-espace vectoriel engendré par les $\nabla g_i(x)$.

Théorème 2.3.2 (Règle de Lagrange)**.** *Soient J, $g_1, \ldots, g_m$ des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ et K le sous-ensemble de $\mathbb{R}^n$ défini par (2.12). On suppose que les fonctions J et g_i (pour $i = 1, \ldots, m$) sont différentiables en un point $x_{\min} \in K$ et que les vecteurs $(\nabla g_i(x_{\min}))_{1 \leq i \leq m}$ sont linéairement indépendants. Si J admet sur K un minimum local en $x_{\min}$, alors il existe $\overline{\lambda}_1, \ldots, \overline{\lambda}_m \in \mathbb{R}$ tels que*

$$\nabla J(x_{\min}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla g_i(x_{\min}) = 0. \tag{2.15}$$

Les coefficients $\overline{\lambda}_i$ sont appelés les *multiplicateurs de Lagrange* associés au point de minimum $x_{\min}$.

Définissons la fonction $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ par

$$\mathcal{L}(x, \lambda) = J(x) + \sum_{i=1}^{m} \lambda_i g_i(x),$$

où $\lambda = [\lambda_1, \ldots, \lambda_m]^\top$. Cette fonction est appelée *fonction de Lagrange* ou *Lagrangien* du problème de la minimisation de J sur K.

La règle de Lagrange dit alors que si $x_{\min}$ minimise J sur K, alors il existe un vecteur $\overline{\lambda} \in \mathbb{R}^m$ tel que le vecteur $\left(x_{\min}, \overline{\lambda}\right)$ soit un point stationnaire de $\mathcal{L}$.

Pour trouver la solution d'un problème d'optimisation avec des contraintes égalités, il suffit de résoudre le système à $n + m$ équations et $n + m$ inconnues

$$\nabla_x \mathcal{L}(x, \lambda) = 0 \quad \text{et} \quad \nabla_\lambda \mathcal{L}(x, \lambda) = 0$$

La règle de Lagrange ne donne qu'une condition nécessaire de minimalité (on obtiendrait de toute façon la même condition pour un maximum). Pour obtenir des conditions suffisantes, il faut rajouter des hypothèses, par exemple la convexité :

Théorème 2.3.3. *Soit J une fonction convexe différentiable sur $\mathbb{R}^n$. Et soient m fonctions g_i (pour $i = 1, \ldots, m$) supposées affines (de la forme $g_i(x) = a_i^\top x + b_i$). Alors, un élément $x_{\min} \in K = \{x \in \mathbb{R}^n : g_i(x) = 0, \forall i \in \{1, \ldots, m\}\}$ pour lequel il existe un vecteur $\lambda \in \mathbb{R}^m$ vérifiant (2.15) est un minimum (global) de J sur K.*

Conditions du second ordre

On suppose cette fois les fonctions J et g_i deux fois différentiables. On admettra les deux résultats suivants.

Théorème 2.3.4 (CN au 2nd ordre)**.**
Soient J, $g_1, \ldots, g_m$ des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ et K le sous-ensemble de $\mathbb{R}^n$ défini par (2.12). On suppose que J et les fonctions g_i sont deux fois différentiables en un point $x_{\min} \in K$ et que les différents vecteurs $(\nabla g_i(x_{\min}))_{1 \le i \le m}$ sont linéairement indépendants. Si J admet sur K un minimum local en $x_{\min}$, alors il existe $\overline{\lambda} = [\overline{\lambda}_1, \ldots, \overline{\lambda}_m] \in \mathbb{R}^m$ tels que

$$\nabla_x \mathcal{L}(x_{\min}, \overline{\lambda}) = 0. \tag{2.16}$$

$$d^\top \nabla_x^2 \mathcal{L}(x_{\min}, \overline{\lambda}) d \ge 0, \tag{2.17}$$

pour tout $d \in \left\{h \in \mathbb{R}^n : \nabla g_i(x)^\top h = 0, \quad \forall i \in \{1, \ldots, m\}\right\}$

Théorème 2.3.5 (CS au 2nd ordre)**.**
Sous les mêmes hypothèses qu'au théorème précédent, s'il existe $\overline{\lambda} \in \mathbb{R}^m$ tel que pour $x_{\min} \in K$

$$\nabla_x \mathcal{L}(x_{\min}, \overline{\lambda}) = 0. \tag{2.18}$$

$$d^\top \nabla_x^2 \mathcal{L}(x_{\min}, \overline{\lambda}) d > 0, \tag{2.19}$$

pour tout $d \neq 0$ dans $\left\{h \in \mathbb{R}^n : \nabla g_i(x)^\top h = 0, \quad \forall i \in \{1, \ldots, m\}\right\}$, alors, $x_{\min}$ est un minimum local strict de J sur K.

Cas des contraintes de type inégalité

On suppose maintenant que l'ensemble des contraintes est de la forme suivante :

$$K = \{x \in \mathbb{R}^n : f_j(x) \le 0 \quad \forall j \in \{1, \ldots, p\}\} \tag{2.20}$$

où $f_1, \ldots, f_p$ sont des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$.

Une contrainte $f_j(x) \le 0$ sera dite *active ou saturée* au point x si $f_j(x) = 0$. Seules ces contraintes ont un rôle dans le problème d'optimisation. En effet, si, à l'optimum, on a $f_j(x_{\min}) < 0$, $\forall j \in \{1, \ldots, p\}$, alors le point $x_{\min}$ est à l'intérieur de K et alors $K(x) = \mathbb{R}^n$.

Pour un point $x \in K$, on notera $I(x)$ l'*ensemble des indices des contraintes actives* défini par

$$I(x) = \{j \in \{1, \ldots, p\} : f_j(x) = 0\}.$$

Recherche des directions admissibles

Soit $d \in \mathbb{R}^n$ une direction admissible à K en x, alors il existe deux suites $(d_k) \to d$ et $(\theta_k) \to 0$ telles que $x + \theta_k d_k \in K$, $\forall k$, c'est-à-dire :

$$f_j(x + \theta_k d_k) \leq 0 \quad \forall j \in I(x)$$

Si les fonctions f_j sont continûment différentiables en x, alors en développant au premier ordre et en prenant la limite lorsque $(d_k) \to d$, on montre que :

$$\nabla f_j(x)^\top d \leq 0 \quad \forall j \in I(x) \tag{2.21}$$

En fait, on vient de montrer que

$$K(x) \subset K^*(x) = \{d \in \mathbb{R}^n : \nabla f_j(x)^\top d \leq 0, \quad \forall j \in I(x)\}$$

L'égalité n'est vraie que sous certaines hypothèses dites de *qualification des contraintes*. Il existe différentes conditions de qualification des contraintes, une des plus simples est la suivante :

Théorème 2.3.6. *Supposons les fonctions f_j différentiables, alors les contraintes sont* qualifiées *au point $x \in K$ s'il existe $h \in \mathbb{R}^n$ tel*

$$\nabla f_j(x)^\top h < 0 \quad \forall j \in I(x). \tag{2.22}$$

Lorsque f_j est une fonction affine, on peut remplacer l'inégalité stricte (2.22) par l'inégalité $\nabla f_j(x)^\top h \leq 0$.

Remarquons alors que si toutes les contraintes f_j sont affines, alors les contraintes sont systématiquement qualifiées : il suffit de choisir $h = 0$.

Géométriquement, l'ensemble $K^*(x)$ est un cône (cf. illustration dans le plan figure 2.4).

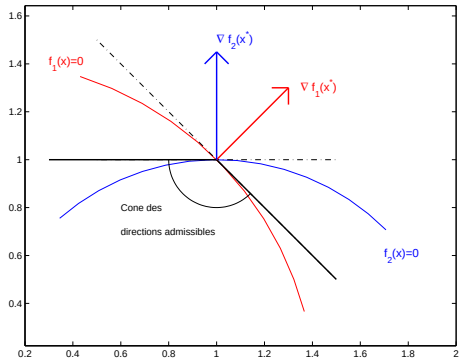


Fig. 2.4: cône des dir. admissibles

Condition au premier ordre

En supposant les contraintes qualifiées, la condition de minimalité (2.11) s'écrit ici :

$$\nabla J(x_{\min})^\top d \geq 0,$$

pour tout d dans $\{h \in \mathbb{R}^n : \nabla f_j(x)^\top h \leq 0, \quad \forall j \in I(x_{\min})\}$.

On peut alors montrer que cela est équivalent à l'existence d'un vecteur $\overline{\mu} \in \mathbb{R}^p$ à composante $\overline{\mu}_j \geq 0$ tel que

$$\nabla J(x_{\min}) + \sum_{j \in I(x_{\min})} \overline{\mu}_j \nabla f_j(x_{\min}) = 0$$

(cf. interprétation géométrique sur la figure 2.5)

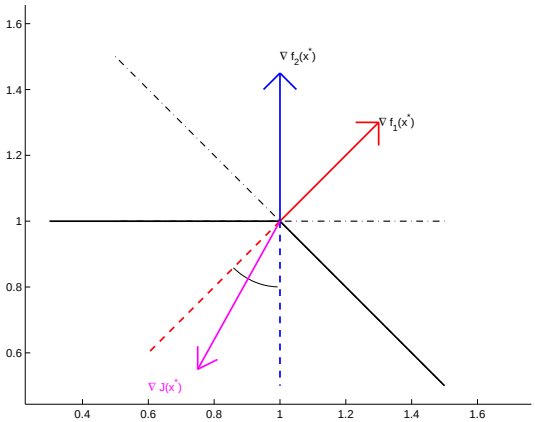


FIG. 2.5: CN de minimalité : $\nabla J(x_{\min})$ doit être dans le cône convexe engendré par les $-\nabla f_j(x_{\min})$ pour $j \in I(x_{\min})$

En résumé :

Théorème 2.3.7. *Soient $J, f_1, \ldots, f_p$ des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ et K le sous-ensemble de X défini par (2.20). On suppose que J et les fonctions f_j sont différentiables en un point $x_{\min} \in K$ et qu'en ce point l'hypothèse de qualification des contraintes est remplie. Si J admet sur K un minimum local en $x_{\min}$, alors il existe des réels $\mu_1, \ldots, \mu_p \geq 0$ tels que*

$$\nabla J(x_{\min}) + \sum_{j=1}^{p} \mu_j \nabla f_j(x_{\min}) = 0, \tag{2.23}$$

$$\mu_j f_j(x_{\min}) = 0, \quad \forall j \in \{1, \ldots, p\}. \tag{2.24}$$

Les coefficients μ_i jouent le même rôle que les multiplicateurs de Lagrange pour les problèmes avec contraintes égalités : on les appelle parfois "multiplicateurs de Lagrange généralisés". Attention, toutefois au signe positif imposé

dans le cadre de contraintes inégalités. La condition (2.24) est appelée condition de complémentarité, elle précise que si une contrainte est inactive au point de minimum, alors nécessairement le multiplicateur associé est nul.

Cas des contraintes de type mixte

CN de minimalité du premier ordre : conditions de KKT

Il est possible d'étendre sans difficulté majeure les résultats précédents de façon à prendre en compte des contraintes se présentant sous les deux formes égalités et inégalités :

$$K = \{x \in \mathbb{R}^n : f_j(x) \leq 0 \quad \forall j \in \{1, \ldots, p\} \text{ et } g_i(x) = 0 \quad \forall i \in \{1, \ldots, m\}\}. \tag{2.25}$$

Les fonctions f_j et g_i sont supposées différentiables sur $\mathbb{R}^n$.

Pour ce problème, les principales conditions de qualification des contraintes en un point x sont les suivantes :

QC1 les fonctions f_j (pour $j \in I(x)$) et g_i (pour $i = 1, \ldots, m$) sont affines.

QC2 les gradients $\nabla f_j(x)$, $\nabla g_i(x)$ (pour $j \in I(x)$ et $i = 1, \ldots, m$) sont linéairement indépendants.

QC3 si

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(x) + \sum_{j \in I(x)} \mu_j \nabla f_j(x) = 0 \quad \text{et} \quad \mu_j \geq 0, j \in I(x) \Rightarrow \lambda_i = 0, \quad \mu_j = 0$$

On a alors le résultat :

Théorème 2.3.8 (Karush-Kuhn-Tucker)**.** *Soient J, f_1, …, f_p, g_1, …, g_m des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ et K le sous-ensemble de $\mathbb{R}^n$ défini par (2.25). On suppose que J et les fonctions f_j, g_i sont différentiables en un point $x_{\min} \in K$ et qu'en ce point les contraintes sont qualifiées.*

Si J admet sur K un minimum local en $x_{\min}$, alors il existe des réels $\overline{\mu_1}, \ldots,$ $\overline{\mu}_p \geq 0$ et $\overline{\lambda}_1, \ldots, \overline{\lambda}_m$ tels que

$$\nabla J(x_{\min}) + \sum_{i=1}^{m} \overline{\lambda}_i \nabla g_i(x_{\min}) + \sum_{j=1}^{p} \overline{\mu}_j \nabla f_j(x_{\min}) = 0,$$
$$\mu_j f_j(x_{\min}) = 0, \quad \forall i \in \{1, \ldots, p\}.$$

Les coefficients λ_i, μ_j sont appelés coefficients de Lagrange (généralisés) ou de Karush-Kuhn-Tucker.

On peut reformuler les conditions de KKT en introduisant le Lagrangien du problème :

$$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_+$$

$$\mathcal{L}(x,\lambda,\mu) = J(x) + \sum_{1\le i\le m} \lambda_i g_i(x) + \sum_{1\le j\le p} \mu_i f_j(x)$$

Conditions de KKT :

$$\begin{cases} \nabla_x \mathcal{L}(x_{\min},\overline{\lambda},\overline{\mu}) = 0, & & \text{(i)}\,; \\ g_i(x_{\min}) = 0 & (1\le i\le m), & \text{(ii)}\,; \\ f_j(x_{\min}) \le 0 & (1\le j\le p), & \text{(iii)}\,; \\ \mu_j \ge 0 & (1\le j\le p), & \text{(iv)}\,; \\ \mu_j f_j(x_{\min}) = 0 & (1\le j\le p), & \text{(v)}. \end{cases}$$

2.4 Optimisation convexe

Optimisation convexe

La notion de convexité possède un rôle privilégié en théorie de l'optimisation car elle permet d'obtenir des conditions globales d'existence, mais aussi par l'existence de conditions nécessaires et suffisantes d'optimalité et cela même dans le cas non différentiable. Elle est également d'une grande importance pratique puisqu'il existe des algorithmes de résolution numérique très efficaces et, conséquence logique, de plus en plus de problèmes concrets sont traités au moyen de l'optimisation convexe.

Le problème

$$(P) \qquad \min_{x\in K} J(x),$$

est un *problème convexe* si la fonction objectif J et l'ensemble des contraintes K sont convexes. Lorsque l'ensemble K est défini par des égalités et des inégalités :

$$K = \{x\in\mathbb{R}^n : g_i(x) = 0,\ i\in\{1,\dots,m\} \text{ et } f_j(x)\le 0,\ j\in\{1,\dots,p\}\}, \quad (2.26)$$

alors il suffit que les fonctions g_i (pour $i = 1,\dots,m$) soient affines et les fonctions f_j (pour $j = 1,\dots,p$) convexes pour que K soit lui-même convexe. On supposera que c'est toujours le cas dans la suite.

Une propriété importante en optimisation convexe est la suivante

Théorème 2.4.1. *Si J est une fonction convexe sur un ensemble convexe C, tout point de minimum local de J sur C est un minimum global et l'ensemble des points de minimum est un ensemble convexe (éventuellement vide). Si de plus J est strictement convexe, alors il ne peut exister au plus qu'un point de minimum.*

Conditions d'optimalité

La plupart des conditions nécessaires d'optimalité données dans les sections précédentes sont également suffisantes dans le cas convexe. Il en est de même

pour les conditions d'optimalité de Karush, Kuhn et Tucker. Un autre avantage non négligeable obtenu à l'aide de la propriété de convexité est la possibilité d'utiliser une condition de qualification des contraintes (dite de Slater), plus simple d'emploi car valable non plus localement mais globalement :

QC4 Si, en écrivant les contraintes égalités sous la forme matricielle $Ax = b$, les m lignes de A sont linéairement indépendantes et s'il existe un point $\overline{x} \in \mathbb{R}^n$ tel que $A\overline{x} = b$ et

$$f_j(\overline{x}) < 0, \quad \forall j \in \{1, \ldots, p\}, \tag{2.27}$$

alors les contraintes sont qualifiées en tout x de K.

Pour toutes les fonctions f_j affines, l'inégalité stricte (2.27) peut être remplacée par l'inégalité large ($f_j(\overline{x}) \leq 0$).

Théorème 2.4.2 (Conditions de Karush, Kuhn et Tucker)**.** *Soient J, $f_1, \ldots$, f_p, $g_1, \ldots$, g_m des fonctions définies sur $\mathbb{R}^n$ et à valeurs dans $\mathbb{R}$ et K l'ensemble défini par (2.26). On suppose que les fonctions J et f_j ($1 \leq j \leq p$) sont convexes et différentiables en un point $x_{\min} \in K$ et que les fonctions g_i sont toutes affines.*

1. *Si J admet sur K un minimum en $x_{\min}$ et si les contraintes sont qualifiées, alors il existe des réels $\lambda_1, \ldots, \lambda_m$ (de signe quelconque) et $\mu_1, \ldots, \mu_p \geq 0$, tels que*

$$\nabla J(x_{\min}) + \sum_{j=1}^{p} \mu_j \nabla f_j(x_{\min}) + \sum_{i=1}^{m} \mu_j \nabla g_i(x_{\min}) = 0, \tag{2.28a}$$

$$\mu_j f_j(x_{\min}) = 0 \quad \forall j \in \{1, \ldots, p\}. \tag{2.28b}$$

2. *Réciproquement, s'il existe des réels $\lambda_1, \ldots, \lambda_m$ et $\mu_1, \ldots, \mu_p \geq 0$, vérifiant les relations* (2.28)*, alors J admet sur K un minimum en $x_{\min}$.*

Exemples de problèmes convexes

Il existe des familles de problèmes convexes ayant une portée très importante en pratique :

– Les problèmes d'optimisation linéaires pour laquelle la fonction objectif est linéaire par rapport aux variables de décision x_i, et les fonctions contraintes (égalités et inégalités) sont des fonctions affines des x_i :

$$(PL) \quad \begin{cases} \text{minimiser} & J(x) = c^\top x, \\ \textit{s.l.c.} \quad x \in \mathbb{R}^n : & a_i^\top x = b_i, \quad i = 1, \ldots, m \\ & d_j^\top x \leq e_j, \quad j = 1, \ldots, p \end{cases}$$

Ces problèmes seront étudiés au chapitre suivant.

– Les problèmes quadratiques où la fonction objectif est quadratique (convexe) et les contraintes affines :

$$(PQ) \quad \begin{cases} \text{minimiser} & J(x) = 1/2x^\top Qx + c^\top x + c_0, \\ s.l.c. \quad x \in \mathbb{R}^n : & a_i^\top x = b_i, \quad i = 1, \dots, m \\ & d_j^\top x \leq e_j, \quad j = 1, \dots, p \end{cases}$$

– Les problèmes d'optimisation quadratique avec contraintes quadratiques où la fonction objectif ainsi que les contraintes inégalités sont quadratiques (convexes) et les contraintes égalités affines :

$$(PQCQ) \quad \begin{cases} \text{minimiser} & J(x) = 1/2x^\top Q_0 x + p_0^\top x + r_0, \\ s.l.c. \quad x \in \mathbb{R}^n : & 1/2x^\top Q_j x + p_j^\top x + r_j \leq 0, \quad j = 1, \dots, p \\ & a_i^\top x = b_i, \quad i = 1, \dots, m \end{cases}$$

– Les problèmes d'optimisation semi-définie pouvant être mis sous la forme :

$$(SDP) \quad \begin{cases} \text{minimiser} & J(x) = c^\top x \\ s.l.c. \quad x \in \mathbb{R}^n : & F_0 + x_1 F_1 + \cdots + x_n F_n \succeq 0 \end{cases}$$

où les F_i, pour $i = 0, 1, \dots, n$, sont des matrices à coefficients réels, symétriques et de même dimension et l'inégalité $A \succeq 0$ signifie que la matrice symétrique A est semi-définie positive. Ces problèmes seront traités au chapitre 2.6.

2.5 Programmation linéaire

Définitions - formes inégalité et standard

On appelle *programme linéaire* un problème d'optimisation où le critère J et toutes les fonctions f_i, g_j intervenant dans les contraintes sont affines en x. Ce sont donc des problèmes pouvant s'écrire sous la forme

$$(P) \quad \begin{cases} \text{minimiser} & J(x) = c^\top x, \\ s.l.c. \quad x \in \mathbb{R}^n : & a_i^\top x = b_i, \quad i = 1, \dots, m \\ & d_j^\top x \leq e_j, \quad j = 1, \dots, p \end{cases}$$

(s.l.c. = sous les contraintes). Remarquons qu'il n'existe pas une façon unique d'écrire un programme linéaire :

– une contrainte égalité peut être transformée en deux contraintes inégalités :

$$a_i^\top x = b_i \Longleftrightarrow \begin{cases} a_i^\top x \leq b_i \\ -a_i^\top x \leq -b_i \end{cases}$$

– une contrainte inégalité peut être transformée en une contrainte de type égalité et une condition de signe sur une variable en introduisant une variable dite d'écart

$$d_j^\top x \leq e_j \Longleftrightarrow \exists z_j \in \mathbb{R} : \begin{cases} d_j^\top x + z_j = e_j \\ z_j \geq 0 \end{cases}$$

– une variable x_i de signe quelconque peut être remplacée par la différence de 2 nouvelles variables positives :

$$x_i = x_i^+ - x_i^-, \quad x_i^+ \geq 0,\ x_i^- \geq 0.$$

On peut ainsi écrire tout programme linéaire de manière équivalente sous :

– une forme **inégalité** :

$$\begin{aligned} &\text{minimiser} && J(x) = c^\top x \\ s.l.c. \quad &x \in \mathbb{R}^n : && a_i^\top x \leq b_i, \quad i = 1, \ldots, m \end{aligned}$$

où toutes les contraintes sont de type inégalité (on réécrira ces inégalités de manière plus condensée sous l'écriture matricielle $Ax \leq b$, où $A \in \mathbb{R}^{m \times m}$ est la matrice de i^e ligne $a_i^\top$, et $b = [b_1 \ldots b_m] \in \mathbb{R}^m$) ;

– une forme **inégalité à variables positives** :

$$\begin{aligned} &\text{minimiser} && J(x) = c^\top x, \\ s.l.c. \quad &x \in \mathbb{R}^n : && a_i^\top x \leq b_i, \quad i = 1, \ldots, m \\ & && x_j \geq 0, \quad j = 1, \ldots, n \end{aligned}$$

– une forme **standard :**

$$\begin{aligned} &\text{minimiser} && J(x) = c^\top x, \\ s.l.c. \quad &x \in \mathbb{R}^n : && a_i^\top x = b_i, \quad i = 1, \ldots, m \\ & && x_j \geq 0, \quad j = 1, \ldots, n \end{aligned}$$

où toutes les contraintes sont de type égalité, les variables d'optimisation x_i étant toutes de signe positif. On réécrira les contraintes sous l'écriture matricielle $Ax = b$, $x \geq 0$.

Remarque : dans ces trois formulations, le vecteur x, le nombre de variables n et le nombre de contraintes m ne sont pas nécessairement conservés.

Un exemple simple

Un laminoir peut produire deux types de bobines d'épaisseurs différentes, nommées B1 et B2. La première — plus épaisse — est produite à 200 tonnes par heure et est vendue avec un bénéfice de 25 euros/tonne, la seconde est produite à 140 tonnes/heure avec un bénéfice de 30 euros/tonne. Pour des raisons de saturation du marché, on ne veut pas produire cette semaine plus de 6000 tonnes de bobines B1 et de 4000 tonnes de B2. Il n'y a que 40 heures de production disponibles cette semaine, quelle quantité de chacune des bobines doit-on produire pour maximiser le profit ?

Si on note x_1 et x_2 les quantités de bobines B1 et B2 produites, alors il s'agit de résoudre le problème d'optimisation suivant :

$$\begin{aligned} \text{maximiser} \quad & 25x_1 + 30x_2 \\ s.l.c. \quad & x_1 \leq 6000 \\ & x_2 \leq 4000 \\ & x_1/200 + x_2/140 \leq 40 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Il est facile de résoudre graphiquement cet exemple :

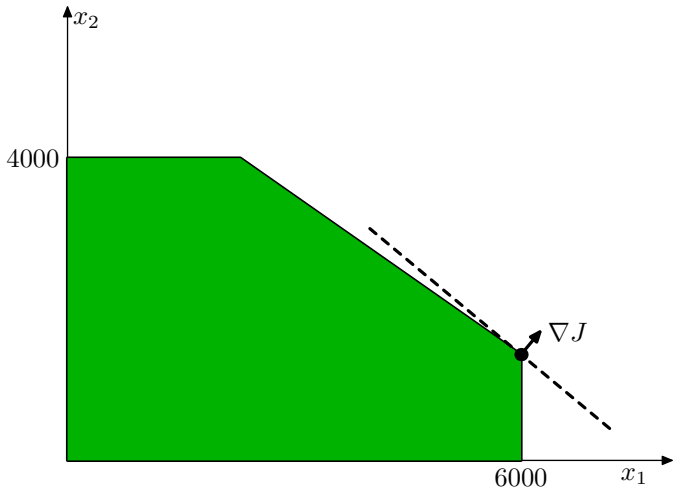


FIG. 2.6: Résolution graphique de l'exemple

On peut tirer de cet exemple simpliste quelques conclusions générales :
— le domaine des contraintes est un polyèdre convexe (ici un pentagone),
— la solution optimale est située sur un sommet de ce polyèdre.

Polyèdres convexes

Définition

Considérons un problème d'optimisation linéaire mis sous forme inégalité

$$\begin{aligned} &\text{minimiser} && J(x) = c^\top x \\ &s.l.c. \quad x \in \mathbb{R}^n && Ax \leq b \end{aligned}$$

où $A \in \mathbb{R}^{m \times n}$ et $b \in \mathbb{R}^m$.

L'ensemble des contraintes (appelé aussi ensemble admissible)

$$K = \{x \in \mathbb{R}^n \mid Ax \leq b\}$$

définit ce que l'on appelle un **polyèdre (convexe).** Il correspond à l'intersection des m demi-espaces

$$\{x \in \mathbb{R}^n \mid a_i^\top x \leq b_i\}$$

où les $a_i^\top$ correspondent aux lignes de la matrice A.

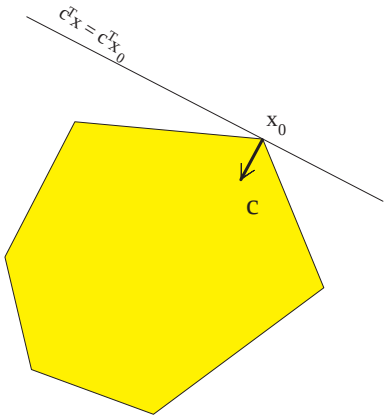
C'est un ensemble éventuellement vide, borné ou non et convexe : on peut facilement vérifier que

$$x, y \in K \Rightarrow \forall \theta \in [0,1], \quad \theta x + (1-\theta) y \in K.$$

Sommets d'un polyèdre

Définition 2.5.1 (sommet)**.** Un **sommet** d'un polyèdre convexe $K \subset \mathbb{R}^n$ est un point x_0 de K pour lequel il existe un vecteur $c \in \mathbb{R}^n$ tel que

$$c^\top x_0 < c^\top x, \quad \forall x \in K,\ x \neq x_0$$



On peut reformuler cette définition en disant qu'un point x_0 est un sommet du polyèdre convexe $K \subset \mathbb{R}^n$ s'il existe un vecteur $c \in \mathbb{R}^n$ tel que x_0 soit l'unique point de K minimisant la fonction $x \mapsto c^\top x$ sur K.

Caractérisation géométrique

Un sommet x_0 est également caractérisé géométriquement par le fait qu'il ne peut être situé strictement à l'intérieur d'un segment contenu dans K :

$$x_0 = \theta y + (1-\theta)z,\ \theta \in [0,1],\ y, z \in K \Rightarrow x_0 = y \text{ ou } x_0 = z$$

(on parle de point **extrémal** du convexe K).

Caractérisation algébrique

Intuitivement, on conçoit bien qu'un sommet est défini comme étant l'intersection de n hyperplans indépendants constituant la frontière du polyèdre. Pour formaliser cela, définissons pour un point x de K :

- $I(x)$ l'*ensemble des contraintes actives* (ou saturées) en x défini par

$$I(x) = \{i \in \{1, \dots, m\} \mid a_i^\top x = b_i\}.$$

  ($I(x)$ est un ensemble d'indices dont on notera k le cardinal et $i_1, \dots, i_k$ les éléments.)
- $\bar{A}(x)$ la sous-matrice de A obtenue en ne prenant que les lignes correspondant à des contraintes actives en x :

$$\bar{A}(x) = \begin{bmatrix} a_{i_1}^\top \\ \vdots \\ a_{i_k}^\top \end{bmatrix}, \quad \text{avec } I(x) = \{i_1, \dots, i_k\}.$$

  Alors, x_0 est un sommet de K si $x_0 \in K$ et vérifie $\text{rang}(\bar{A}(x_0)) = n$ (dans la littérature, un tel point est appelé une *solution basique admissible*).

Théorème 2.5.1. *Les 3 caractérisations d'un sommet sont équivalentes.*

Démonstration :

- Si x_0 est un sommet de K, c'est aussi un point extrémal de K.
  En effet, soient $y, z \in K$, $y \neq x_0$, $z \neq x_0$. x_0 est un sommet de K : il existe donc $c \in \mathbb{R}^n$ tel que :

$$c^\top x_0 < c^\top y, \quad c^\top x_0 < c^\top z.$$

  Pour $\theta \in [0,1]$, on a alors

$$c^\top x_0 < c^\top(\theta y + (1-\theta)z)$$

  d'où $x_0 \neq \theta y + (1-\theta)z$

– Si x_0 est un point extrémal de K, alors c'est aussi une solution basique admissible.
En effet, supposons qu'il existe un vecteur $d \neq 0$ tel que $\bar{A}(x_0)d = 0$, c'est-à-dire

$$a_i^\top d = 0, \quad i \in I(x_0).$$

Pour $\varepsilon > 0$ suffisamment petit, on a

$$y = x +_0 \varepsilon d \in K, \quad z = x_0 - \epsilon d \in K,$$

car
– pour $i \in I(x_0) : a_i^\top y = a_i^\top z = a_i^\top x_0 = b_i$,
– pour $i \notin I(x_0) : a_i^\top x_0 - b_i < 0$, prenons $\varepsilon < \left| \frac{a_j^\top x_0 - b_j}{d_j} \right|$ pour tous les $j \in \{1, \ldots, m\} \setminus I(x_0)$ tels que $d_j \neq 0$, alors $a_i^\top y < b_i$ et $a_i^\top z < b_i$.
On a alors $x_0 = (y + z)/2$: ce qui contredit le fait que x_0 est un point extrémal de K.
Conclusion : $\mathrm{Ker}(\bar{A}(x_0)) = \{0\} \Rightarrow \mathrm{rang}(\bar{A}(x_0)) = n$: x_0 est une solution basique admissible de K.
– Si x_0 est une solution basique admissible de K, alors c'est aussi un sommet.
Posons $c = -\sum_{i \in I(x_0)} a_i$. Alors, pour tout $y \in K$:

$$c^\top x_0 - c^\top y = \sum_{i \in I(x_0)} (a_i^\top y - b_i) \leq 0,$$

avec $c^\top y = c^\top x_0$ si et seulement si $a_i^\top y = b_i$ pour tout $i \in I(x_0)$. Or $\mathrm{rang}(\bar{A}(x_0)) = n$: il ne peut y avoir au plus qu'une seule solution à ce système d'équations, donc $y = x_0$.
Conclusion : $c^\top x_0 < c^\top y$ pour tout $y \in K$, $y \neq x_0$.

Propriétés des polyèdres sous forme standard

Soit K un polyèdre représenté sous forme standard :

$$K = \{x \in \mathbb{R}^n : Ax = b,\ x \geq 0\}.$$

Si $0 \in K$, alors c'est un sommet de K.

Un point x de K , $x \neq 0$, est un sommet si et seulement si les colonnes de A correspondant au composantes non nulles de x sont linéairement indépendantes, c-à-d.

$$\mathrm{rang}\,\{a_{j_1}, a_{j_2}, \ldots, a_{j_N}\} = N,$$

où

$$\{j_1, j_2, \ldots, j_N\} = I^+(x) = \{j \in \{1, 2, \ldots, n\} : x_j > 0\}$$

et a_j représente la $j^{ième}$ **colonne** de A.
Le sommet x de K est dit **dégénéré** s'il a plus que $n - m$ composantes nulles (ou encore l'entier N est strictement inférieur à m).

Remarque : on peut déduire de ce résultat qu'un polyèdre n'admet qu'un nombre fini de sommets.

Théorème 2.5.2. *Tout polyèdre non vide représenté sous forme standard*

$$K = \{x \in \mathbb{R}^n : Ax = b,\ x \geq 0\}$$

possède au moins un sommet.

Théorème fondamental de la programmation linéaire

On considère le problème d'optimisation linéaire mis sous forme standard

$$(P_s) : \begin{cases} \text{minimiser } c^\top x \\ s.l.c. \quad x \in K = \{x \in \mathbb{R}^n : Ax = b, \quad x \geq 0\} \end{cases}.$$

Notons

$$p^* = \inf_{x \in K} (c^\top x)$$

la valeur (optimale) du problème.
On conviendra que

– $p^* = +\infty$ si K est vide et

– $p^* = -\infty$ si l'ensemble $\{c^\top x : x \in K\}$ est non minoré

On appellera ensemble des solutions optimales l'ensemble des $x \in K$ tels que $c^\top x = p^*$.

Théorème 2.5.3. *Si p^* est fini, alors l'ensemble des solutions optimales contient au moins un sommet de K.*

Méthode du simplexe

La méthode du simplexe est due à George Dantzig (1947). Son principe consiste à construire une suite de sommets $\{x^k\}$ telle que, à chaque étape, on ait $J(x^{k+1}) \leq J(x^k)$. Le nombre de sommets d'un polyèdre étant fini, l'algorithme doit normalement converger en un nombre fini d'étapes.
Avant de formaliser l'algorithme du simplexe, on va voir sa mise en œuvre sur un exemple simple.

Exemple

On reprend l'exemple précédent (mis sous forme minimisation) :

$$\begin{aligned} \text{minimiser} \quad & J(x_1, x_2) = -25x_1 - 30x_2 \\ s.l.c. \quad & x_1 \leq 6000 \\ & x_2 \leq 4000 \\ & x_1/200 + x_2/140 \leq 40 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Reformulons ce problème sous forme standard en introduisant trois variables d'écart x_3, x_4, x_5 :

$$\begin{aligned} \text{minimiser} \quad & J(x_1, x_2, x_3, x_4, x_5) = -25x_1 - 30x_2 \\ s.l.c. \quad & x_1 + x_3 = 6000 \\ & x_2 + x_4 = 4000 \\ & x_1/200 + x_2/140 + x_5 = 40 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0 \end{aligned}$$

— **Phase I du simplexe : détermination d'un sommet initial**

Pour $x_1 = x_2 = 0$, on obtient $x_3 = 6000$, $x_4 = 4000$, $x_5 = 40$. On obtient ainsi une solution admissible puisque $x_1, x_2, x_3, x_4, x_5 \geq 0$. Le point obtenu constitue en fait un sommet de l'ensemble des contraintes[2]. Ce premier sommet a été obtenu ici relativement facilement, il n'en est pas toujours de même. On verra cependant au paragraphe 2.5 comment un premier sommet peut être obtenu systématiquement. Remarquons que pour ce premier sommet, la fonction J a pour valeur 0.

— **Première étape**

Test de l'optimalité du sommet courant L'ensemble des contraintes est donné par un ensemble de 3 contraintes égalités liant les 5 variables $x_1, \ldots, x_5$: on a donc 2 degrés de liberté. En choisissant d'exprimer la fonction coût ainsi que les variables x_3, x_4 et x_5 en fonction des variables x_1 et x_2 (qui lorsqu'elles sont nulles nous redonne le sommet courant), on obtient une formulation équivalente

[2] vérifier que les conditions données au paragraphe 2.5 sont remplies

du problème de départ

$$\begin{aligned} \text{minimiser} \quad & J(x_1, x_2, x_3, x_4, x_5) = -25x_1 - 30x_2 \\ s.l.c. \quad & x_3 = 6000 - x_1 \\ & x_4 = 4000 - x_2 \\ & x_5 = 40 - x_1/200 - x_2/140 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0 \end{aligned}$$

L'expression ci-dessus de la fonction J montre qu'il suffit d'augmenter x_1 ou x_2 pour faire décroître le coût : le sommet courant (correspondant à $x_1 = x_2 = 0$) n'est pas une solution optimale !

Passage à un nouveau sommet Pour faire décroître J, il suffit, par exemple, d'augmenter x_1 en laissant x_2 à zéro. La variable x_1 ne peut cependant pas augmenter indéfiniment : la première contrainte nous limite à une valeur de 6 000, tandis que la troisième, moins sévère, nous autorise une valeur maximale de 8 000. Pour la plus petite de ces deux valeurs, c-à-d. $x_1 = 6000$ (en laissant pour l'instant x_2 à 0), on obtient $x_3 = 0$ (normal !), $x_4 = 4000$, $x_5 = 10$. Ce point est un autre sommet de l'ensemble des contraintes. La valeur de la fonction coût J est passée de 0 à $-150\,000$.

— **Deuxième étape**

Test de l'optimalité Ce deuxième sommet est-il optimal ? Pour le voir, choisissons cette fois de tout exprimer en fonction des variables x_3, x_2. Les contraintes égalités s'écrivent alors :

$$\begin{aligned} x_1 &= 6000 - x_3 \\ x_4 &= 4000 - x_2 \\ x_5 &= 10 + \tfrac{1}{200}x_3 - \tfrac{1}{140}x_2 \end{aligned}$$

et la fonction coût a pour expression

$$J(x_1, x_2, x_3, x_4, x_5) = -150\,000 + 25x_3 - 30x_2$$

On voit donc que la solution n'est pas optimale puisque l'on peut faire encore diminuer J en augmentant x_2.

Recherche d'un nouveau sommet On augmente x_2 en laissant cette fois x_3 à zéro. La valeur maximale de x_2, égale à 1400, est maintenant fixée par la troisième inégalité. On obtient alors le nouveau sommet $x_1 = 6000$, $x_2 = 1400$, $x_4 = 2600$, $x_3 = x_5 = 0$. En ce sommet, J vaut $-192\,000$.

— **Dernière étape**

Cette solution est optimale. En effet, si l'on formule de nouveau le problème en choisissant comme variables indépendantes x_3 et x_5, on obtient le problème

$$\begin{array}{ll} \text{minimiser} & J(x_1,x_2,x_3,x_4,x_5) = -192\,000 + 4x_3 + 4200x_5 \\ s.l.c. & x_1 = 6000 - x_3 \\ & x_4 = 2600 - \frac{7}{10}x_3 + 140x_5 \\ & x_2 = 1400 + \frac{7}{10}x_3 - 140x_5 \\ & x_1, x_2, x_3, x_4, x_5 \geq 0 \end{array} .$$

Puisque $J(x_1,x_2,x_3,x_4,x_5) = -192\,000+4x_3+4200x_5$ et x_3, x_5 sont positifs, on a bien $J(x_1,x_2,x_3,x_4,x_5) \geq -192\,000$ et cette valeur est atteinte pour $x_3 = x_5 = 0$: c'est donc bien la valeur la plus faible possible. La solution optimale est donc obtenue en produisant 6000 bobines B1 et 1400 bobines B2.

Formalisation de l'algorithme

Hypothèses : On considère un problème linéaire **mis sous forme standard** :

$$(P_s) : \left\{ \begin{array}{lll} & \text{minimiser} & c^\top x \\ s.l.c. & x \in \mathbb{R}^n & Ax = b, \\ & & x \geq 0, \end{array} \right.$$

où $A \in \mathbb{R}^{m\times n}$.

On suppose dans cette section que A **est de rang** m (sinon, soit le système d'équations $Ax = b$ est incompatible, l'ensemble des contraintes est alors vide, soit il y a des équations redondantes que l'on peut supprimer).

Terminologie :

- Une **base** B est un ensemble de m indices pris dans $\{1, 2, \ldots, n\}$ — notée ci-dessous $\{j_1, j_2, \ldots, j_m\}$ — telles que les colonnes correspondantes de A (c'est-à-dire, $a_{j_1}, a_{j_2}, \ldots, a_{j_m}$) soient linéairement indépendantes.
  On notera N l'ensemble des autres indices (indices non basiques).
- Selon le choix de cette base, après permutation des composantes, on partitionnera un vecteur x de $\mathbb{R}^n$ sous la forme

$$x \to \left[\begin{array}{c} x_B \\ \\ x_N \end{array} \right],$$

où $x_B = [x_{j_1} \dots x_{j_m}]$ regroupe les **composantes basiques** (x_j pour $j \in B$) et
x_N les **composantes non basiques** (les autres composantes).

On utilisera la même partition pour le vecteur c et la matrice A :

$$c \to \begin{bmatrix} c_B \\ \\ c_N \end{bmatrix}, \quad A \to \begin{bmatrix} A_B & A_N \end{bmatrix}$$

On a alors

$$c^\top x = c_B^\top x_B + c_N^\top x_N$$

et, la matrice A_B étant inversible

$$Ax = b \Leftrightarrow A_B x_B + A_N x_N = b \Leftrightarrow x_B = A_B^{-1} b - A_B^{-1} A_N x_N.$$

– On appelle **point basique** une solution de l'équation $Ax = b$ ayant toute ses composantes non basiques nulles : on a alors

$$x_B = A_B^{-1} b, \quad x_N = 0.$$

– Si un point basique a toutes ses composantes basiques x_B positives, alors c'est un **sommet** (on rencontre également l'expression "point ou solution basique admissible")

$$x_B = A_B^{-1} b \geq 0, \quad x_N = 0$$

Si $x_B > 0$, on dira que x est un sommet **non dégénéré** : il ne peut correspondre qu'une seule base à un tel sommet. Inversement, si au moins une des composantes de x_B est nulle, on parle alors de sommet **dégénéré** (qui peut alors être associé à des bases différentes).

– Deux sommets x et $\tilde{x}$ sont dits **adjacents** si leurs bases ne diffèrent que d'un seul élément. Géométriquement, le segment $[x, \tilde{x}]$ est une arête du polyèdre K.

Une itération courante de la méthode du simplexe

On suppose que l'itération précédente nous a fourni un sommet $\hat{x}$ de $K = \{x \in \mathbb{R}^n : Ax = b \text{ et } x \geq 0\}$ et sa base B associée.

Le but de cette itération est de tester si $\hat{x}$ est une solution optimale du problème, ou alors, trouver un nouveau sommet $\hat{x}^+$ adjacent à $\hat{x}$ de coût inférieur ($J(\hat{x}^+) \leq J(\hat{x})$).

* *Reconnaître l'optimalité*

Soit $\hat{x}$ un sommet de base B et x un point quelconque de K

$$\begin{aligned} c^\top x &= c_B^\top x_B + c_N^\top x_N \\ &= c_B^\top A_B^{-1} b + \left[c_N^\top - c_B^\top A_B^{-1} A_N \right] x_N \end{aligned}$$

Si le vecteur $c_N^\top - c_B^\top A_B^{-1} A_N$ a toutes ses composantes ≥ 0. Alors, étant donné que $x \in K$, on a $x_N \geq 0$ et donc $c^\top x \geq c_B^\top A_B^{-1} b = c^\top \hat{x}$

* *Passage d'un sommet à un sommet adjacent*

Si le vecteur $c_N^\top - c_B^\top A_B^{-1} A_N$ admet au moins une composante $j^* < 0$, alors en faisant augmenter x_{j^*} la fonction objectif diminue (il ne faut pas toutefois sortir de l'ensemble des contraintes).

Soit $j^* \in N$ (un indice n'appartenant pas à B). On va rechercher un sommet $\hat{x}^+$ adjacent à $\hat{x}$ sous la forme

$$\hat{x}^+ = \hat{x} + td,$$

où $t \in \mathbb{R}^+$ et d est un vecteur tel que $d_{j^*} = 1$ et $d_j = 0$ pour $j \in N \setminus \{j^*\}$ (on ne veut modifier qu'une seule composante non basique de $\hat{x}$: la $j^* - ième$).
Choisissons le vecteur d tel que $Ad = 0$, car si $\hat{x}^+$ et $\hat{x}$ sont dans K, on a $0 = b - b = A\hat{x}^+ - A\hat{x} = tAd$.
On a donc

$$\sum_{j \in B} a_j d_j + a_{j^*} = 0 \Leftrightarrow d_B = -A_B^{-1} a_{j^*}.$$

Si $\hat{x}$ est un sommet non dégénéré, alors pour $t > 0$ suffisamment petit, le point $\hat{x} + td$ appartient aussi à K :
$A(\hat{x} + td) = b$,
$(\hat{x} + td)_j > 0$, pour $j \in B$ et $t > 0$ suffisamment petit,
$(\hat{x} + td)_{j^*} = t$, pour tout t
$(\hat{x} + td)_j = 0$, pour $j \in N \setminus \{j^*\}$ et tout t.
On recherche alors la plus grande valeur de t pour laquelle $(\hat{x} + td)$ reste dans K :

$$t_{\max} = \begin{cases} +\infty & \text{si} \quad d_B \geq 0 \\ \min\{-\frac{\hat{x}_j}{d_j} : j \in B,\ d_j < 0\} & \text{sinon} \end{cases}$$

Remarque : si x est dégénéré, on peut avoir $t_{\max} = 0$, sinon on a toujours $t_{\max} > 0$.

Si $t_{\max}$ est infini, alors K contient la demi-droite $\{\hat{x} + t_{\max} d,\ t \geq 0\}$.

Si $t_{\max}$ est fini et non nul, alors le point $\hat{x}^+ = \hat{x} + t_{\max} d$ est un nouveau sommet de K adjacent à $\hat{x}$ de base $B^+ = \{j^*\} \cup B \setminus \{k\}$[3], où $k \in B$ est tel que $-\frac{\hat{x}_k}{d_k} = t_{\max}$ (on dit alors qu'on a fait rentrer j^* dans la base et sortir k de la base).

Si $t_{\max}$ est nul, alors on reste sur le même sommet $\hat{x}^+ = \hat{x}$, mais avec la nouvelle base $B^+ = \{j^*\} \cup B \setminus \{k\}$.

* *Variation de la fonction coût*

[3]C'est bien une base, car sinon a_{j^*} serait une combinaison linéaire des a_j, $j \in J \setminus \{k\}$, c-à-d il existerait un vecteur u tel que $a_{j^*} = Bu$ avec $u_k = 0$, donc on aurait $u = -d_J$ ce qui contredit le fait que $d_k < 0$.

On a

$$\begin{aligned} J(x + t_{\max} d) &= c^\top (x + t_{\max} d) \\ &= c^\top x + t_{\max}(c_{j^*} + c_B^\top d_B) \\ &= J(x) + t_{\max}(c_{j^*} - c_B^\top A_B^{-1} a_{j^*}). \end{aligned}$$

On appelle $\bar{c}_j = c_j - c_B^\top A_B^{-1} a_j$ le **coût réduit** associé à l'indice non basique j.

Si $\bar{c}_j < 0$, alors le coût diminue si l'on fait entrer j dans la base. Si de plus $d_B \geq 0$ (c'est-à-dire $t_{\max} = +\infty$), alors le problème est non borné inférieurement.

Si $\bar{c}_j \geq 0$ pour tous les indices j non basiques ($j \in \{1, \ldots, n\} \setminus B$), alors le point $\hat{x}$ est une solution optimale du problème.

Algorithme du simplexe

On suppose que l'on dispose d'un sommet $\hat{x} \in \mathbb{R}^n$ de base B.

1. On calcule les coûts réduits $\bar{c}_j = c_j - c_B^\top A_B^{-1} a_j$ associés aux indices non basiques.
   Si $\bar{c}_j \geq 0$ pour tous les $j \notin B$, alors $\hat{x}$ est une solution du programme (fin).
   Sinon, on choisit un indice j^* tel que $\bar{c}_{j^*} < 0$.
2. On calcule la direction d :

   $$d_B = -A_B^{-1} a_{j^*}, d_{j^*} = 1 \text{ et } d_j = 0 \text{ pour les autres indices.}$$

   Si $d \geq 0$, alors le problème est non borné (fin), sinon on calcule

   $$t_{\max} = \min\{-\frac{\hat{x}_j}{d_j} : j \in B,\ d_j < 0\}$$

   On choisit l'indice k sortant de la base parmi ceux qui vérifient $t_{\max} = -\hat{x}_j/d_j$.
3. On remet à jour $\hat{x}$:

   $$\hat{x} := \hat{x} + t_{\max} d,$$

   ainsi que la base B : $B = (B \cup \{j^*\}) \setminus \{k\}$

Remarques :

- A la fin de la première étape, le choix de l'indice j^* peut se faire selon plusieurs critères :
  - minimiser $\bar{c}_j$
  - minimiser $t_{\max} \bar{c}_j$
  - plus petit indice tel que $\bar{c}_j < 0$.
    Pour une meilleure efficacité, ce sont les deux premières règles qui sont utilisées dans les algorithmes disponibles dans les bibliothèques spécialisées (netlib, ...)

– Si le point $\hat{x}$ est non dégénéré, alors soit il est optimal, soit le nouveau sommet obtenu à un coût strictement inférieur à $J(\hat{x})$. Si tous les sommets successifs sont non dégénérés, on est sûr que l'algorithme converge en un nombre fini d'étapes (il n'y a qu'un nombre fini de sommets). Si le sommet $\hat{x}$ est dégénéré, il se peut que $t_{\max} = 0$: le point suivant est encore $\hat{x}$ mais avec une autre base. Il y a alors risque de cyclage : l'algorithme reste collé au même sommet en reproduisant périodiquement les mêmes bases. Ce phénomène est très rarement rencontré en pratique et les algorithmes disponibles ne prennent pas en compte ce phénomène, mais il existe des remèdes à ce phénomène de cyclage, comme par exemple en choisissant comme indices entrant et sortant de la base les plus petits indices parmi ceux possibles (règle de Bland).
– Lorsque $t_{\max} = 0$, il y a deux situations possibles : soit il existe des valeurs de j permettant de faire progresser normalement l'algorithme (règle anti-cyclage), soit toutes les valeurs de j tel que $\bar{c}_j < 0$ conduisent au même cas ($t_{\max} = 0$) et la solution x est en fait optimale.

Initialisation de la méthode du simplexe (phase I)

Il reste à obtenir un premier sommet (et sa base associée) pour démarrer l'algorithme ou montrer qu'il n'en existe pas (K peut être vide).

Pour cela, on va en fait appliquer l'algorithme précédent au problème

$$(P_s) : \begin{cases} \text{minimiser} & \sum_{i=1}^{m} z_i \\ s.l.c. \quad x \in \mathbb{R}^n, z \in \mathbb{R}^m & Ax + Dz = b, \\ & x \geq 0,\ z \geq 0 \end{cases}$$

où D est une matrice diagonale avec $D_{ii} = 1$ si $b_i \geq 0$ et $D_{ii} = -1$ si $b_i < 0$. Pour ce problème, on dispose d'un sommet évident : $x = 0$, $z = Db$ de composantes $z_i = |b_i| \geq 0$.

Remarquons que le problème (P_s) a une valeur finie comprise entre 0 (car $z \geq 0$) et $\sum_{i=1}^{m} |b_i|$.

L'ensemble des contraintes du problème de départ (P) est non vide si et seulement si le problème (P_s) admet 0 comme valeur optimale. En effet, si l'ensemble des contraintes de (P) est non vide, il existe $x \geq 0$ tel que $Ax = b$. Il suffit de prendre pour z le vecteur nul : toute les contraintes de (P_s) sont vérifiées et $\sum_{i=1}^{m} z_i = 0$, ce qui est le minimum pour une somme de nombres positifs. Réciproquement, si (P_s) a 0 comme valeur optimale pour la solution optimale $(\hat{x}, \hat{z})$ alors

$$\sum_{i=1}^{m} \hat{z}_i = 0, \quad \hat{z}_i \geq 0 \Rightarrow \hat{z}_i = 0,\ i = 1, \ldots, m$$

et donc, $\hat{x}$ est un point vérifiant les contraintes du problème (P).

En résolvant le problème (P_s) avec la méthode précédente et en prenant comme premier sommet $x = 0$, $z_i = |b_i|$, on obtient alors deux cas possibles :

- la valeur optimale est strictement positive : il n' y a pas de point vérifiant les contraintes du problème (P) (fin de l'algorithme) ;
- la valeur optimale est obtenue en un sommet $(0,\hat{x})$. On peut alors vérifier que $\hat{x}$ est un sommet particulier de (P) : on peut alors démarrer la phase II.

Dualité

Motivation

Reprenons l'exemple introductif

$$\begin{aligned} \text{minimiser} \quad & J(x_1, x_2) = -25x_1 - 30x_2 \\ s.l.c. \quad & x_1 \leq 6000 \\ & x_2 \leq 4000 \\ & x_1/200 + x_2/140 \leq 40 \\ & x_1, x_2 \geq 0 \end{aligned}$$

Il est facile d'obtenir un majorant de la valeur p^* du problème si l'on connaît un point (x_1, x_2) vérifiant les contraintes, par exemple pour $x_1 = 2000$, $x_2 = 1400$, on obtient :

$$p^* \leq -25 \times 2000 - 30 \times 1400 = -92\,000.$$

Comment peut-on obtenir un minorant de p^* ?

Soient z_1, z_2 et z_3 des nombres positifs. Multiplions la première inégalité par $-z_1$, la seconde par $-z_2$ et la troisième par $-z_3$ et additionnons les inégalités obtenues, on obtient alors

$$-z_1\, x_1 - z_2\, x_2 - z_3\, (x_1/200 + x_2/140) \geq -(6000z_1 + 4000z_2 + 40z_3)$$

Si $-z_1 - z_3\,/200 \leq -25$ et $-z_2 - z_3/140 \leq -30$, alors, x_1, x_2 étant positifs, on a

$$J(x_1, x_2) = -25x_1 - 30x_2 \geq (-z_1 - z_3\,/200)\, x_1 + (-z_2 - z_3/140)\, x_2$$
$$(-z_1 - z_3\,/200)\, x_1 - (z_2 + z_3/140)\, x_2 \geq -(6000z_1 + 4000z_2 + 40z_3)$$

On a ainsi obtenu un minorant de p^* :

$$p^* \geq -(6000z_1 + 4000z_2 + 40z_3)$$

La recherche du plus grand de ces minorants conduit alors au nouveau problème

$$\begin{array}{rc} \text{maximiser} & -(6000z_1 + 4000z_2 + 40z_3) \\ s.l.c. & -z_1 - z_3/200 \leq -25 \\ & -z_2 - z_3/140 \leq -30 \\ & z_1, z_2, z_3 \geq 0 \end{array}$$

Si on choisit $z_1 = 4, z_2 = 0$ et $z_3 = 4200$, on obtient $p^* \geq -192000$ (et c'est le plus grand minorant possible puisque $p^* = -192000$).

Problème dual

Considérons le problème d'optimisation linéaire, appelé dans ce contexte le **primal**, admettant x^* comme solution optimale

$$(P) \quad \left\{ \begin{array}{lll} \text{minimiser} & & J(x) = c^\top x, \\ s.l.c. & x \in \mathbb{R}^n : & Ax = b, \quad b \in \mathbb{R}^m \\ & & x \geq 0, \end{array} \right.$$

Considérons la fonction

$$g(p) = \min_{x \geq 0} \left[c^\top x + p^\top (b - Ax) \right].$$

On a alors

$$g(p) \leq c^\top x^* + p^\top (b - Ax^*) = c^\top x^*.$$

$g(p)$ est donc un minorant de la valeur de (P). Cherchons le plus grand de ces minorants, c'est-à-dire : $\max g(p)$.

Avec quelques manipulations :

$$\begin{array}{rcl} g(p) & = & \min\limits_{x \geq 0} \left[c^\top x + p^\top (b - Ax) \right] \\ & = & p^\top b + \min\limits_{x \geq 0} (c^\top - p^\top A)x. \end{array}$$

et en remarquant que

$$\min_{x \geq 0} (c^\top - p^\top A)x = \left\{ \begin{array}{ll} 0 & \text{si} \quad c^\top - p^\top A \geq 0 \\ -\infty & \text{sinon} \end{array} \right.$$

on voit que maximiser $g(p)$ revient à résoudre le nouveau problème d'optimisation

$$(D) \quad \left\{ \begin{array}{lll} \text{maximiser} & & b^T p \\ s.l.c. & p \in \mathbb{R}^m : & A^T p \leq c \end{array} \right.$$

C'est le problème **dual** de (P).

De manière générale, on passe du primal au dual en utilisant les transformations suivantes

$$\begin{array}{llllll}
\min. & c^\top x & & \rightarrow & \max. & b^\top p \\
s.l.c. & a_i^\top x \geq b_i & i \in M_1 & & s.l.c. & p_i \geq 0 \quad i \in M_1 \\
& a_i^\top x \leq b_i & i \in M_2 & & & p_i \leq 0 \quad i \in M_2 \\
& a_i^\top x = b_i & i \in M_3 & & & p_i \text{ libre} \quad i \in M_3 \\
& x_j \geq 0 & j \in N_1 & & & p^\top A_j \leq c_j \quad j \in N_1 \\
& x_j \leq 0 & j \in N_2 & & & p^\top A_j \geq c_j \quad j \in N_2 \\
& x_j \text{ libre} & j \in N_3 & & & p^\top A_j = c_j \quad j \in N_3
\end{array}$$

où les A_j représente la j-ième colonne de la matrice A de i-ième ligne $a_i^\top$.
Remarque : le dual du dual est le primal.

Dualité faible

On considère de nouveau les problèmes (P) et (D) précédents :

$$(P) \quad \left\{ \begin{array}{ccc} \text{minimiser} & & c^\top x, \\ s.l.c. \quad x \in \mathbb{R}^n : & Ax = b, \quad b \in \mathbb{R}^m & \\ & x \geq 0, & \end{array} \right. \qquad (D) \quad \left\{ \begin{array}{cc} \text{maximiser} & b^\top p \\ s.l.c. \quad p \in \mathbb{R}^m : & A^\top p \leq c \end{array} \right.$$

On notera :
- $K = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ l'ensemble des solutions primales admissibles ;
- p^* la valeur du problème (P), c'est-à-dire $p^* = \inf(c^\top x : x \in K)$;
- $K^* = \left\{p \in \mathbb{R}^m : A^\top p \leq c\right\}$ l'ensemble des solutions duales admissibles ;
- d^* la valeur du problème (D), c'est-à-dire $d^* = \sup(b^\top p : p \in K^*)$.

On convient toujours que $p^* = -\infty$ si K est vide, $p^* = +\infty$ si l'ensemble $\left\{c^\top x : x \in K\right\}$ est non minoré (problème non borné). De même, $d^* = -\infty$ si K^* est vide, $d^* = +\infty$ si l'ensemble $\left\{b^\top p : p \in K^*\right\}$ est non majoré.

Si x est une solution primale admissible ($x \in K$) et z une solution duale admissible ($p \in K^*$), alors

$$b^\top p \leq c^\top x.$$

En effet,

$$b^\top p = (Ax)^\top p = x^\top A^\top p \leq c^\top x$$

Moralité : en minimisant par rapport à x, on voit que $b^\top p$ constitue un minorant de la valeur du problème primal :

$$p^* \geq b^\top p.$$

Dans cet esprit, le problème dual consiste à rechercher le plus grand de ces minorants, on a alors (en maximisant par rapport à p)

$$p^* \geq d^*.$$

et pour toute paire $(x, p) \in K \times K^*$, on a :

$$c^\top x \geq p^* \geq d^* \geq b^\top p.$$

On en déduit que si l'on trouve une paire (x, p) telle que x est solution primale admissible, p une solution duale admissible et $c^\top x = b^\top p$, alors x est une solution (primale) optimale (et p une solution duale optimale).

Dualité forte

En utilisant les propriétés des ensembles convexes, on peut montrer le lemme suivant

Lemme 2.5.1 (Farkas)**.** *Soit $A \in \mathbb{R}^{m\times n}$ et $b \in \mathbb{R}^m$, alors une et une seule des propositions suivantes est vraie :*
i) il existe un $x \in \mathbb{R}^n$, $x \geq 0$ tel que $Ax = b$,
ii) il existe un $y \in \mathbb{R}^m$ tel que $A^\top y \geq 0$ et $b^\top y < 0$.

A l'aide de ce résultat, on peut montrer qu'en fait les problèmes primal et dual ont la même valeur :

$$\boxed{p^* = d^*}$$

sauf si le problème primal est irréalisable (c'est-à-dire K est vide et donc $p^* = +\infty$), on peut alors avoir :
- $d^* = +\infty$ (problème dual non borné) ou
- $d^* = -\infty$ (problème dual irréalisable : $K^* = \emptyset$).

Démonstration :

i) On va d'abord démontrer le résultat, pour un problème mis sous forme inégalité

$$(P_i) : \begin{cases} \text{minimiser} & c^\top x, \\ s.l.c. \quad x \in \mathbb{R}^n : & Fx \leq g, \end{cases}$$

Le dual associé est

$$(D_i) : \begin{cases} \text{maximiser} & -g^\top z, \\ s.l.c. & F^\top z = -c, \quad z \geq 0 \end{cases}$$

La propriété de dualité faible est toujours valable : $\forall x \in \mathbb{R}^n : Fx \leq g$, $\forall z \geq 0 : F^\top z = -c$,

$$-g^\top z \leq d_i^* \leq p_i^* \leq c^\top x.$$

Supposons p_i^* fini et soit $\hat{x}$ une solution optimale de (P_i).

On note $I(\hat{x})$ l'*ensemble des contraintes actives* (ou saturées) en $\hat{x}$ défini par

$$I(\hat{x}) = \{i \in \{1, \ldots, m\} \mid f_i^\top \hat{x} = g_i\},$$

où les $f_i^\top$ représentent les lignes de F

Puisque $\hat{x}$ est une solution optimale, il n'existe pas de vecteur d tel que

$$f_i^\top d \leq 0, \ \forall i \in I(\hat{x}) \quad \text{et} \quad c^\top d < 0$$

sinon, pour $t > 0$ suffisamment petit, on aurait $f_i^\top (\hat{x} + td) \leq g_i$ pour $i = 1, \ldots, m$ et $c^\top (\hat{x} + td) < c^\top \hat{x}$.

D'après le lemme de Farkas, il existe donc des réels λ_i, $i \in I(\hat{x})$ tels que

$$\lambda_i \geq 0 \text{ et } \sum_{i \in I(\hat{x})} \lambda_i a_i = -c$$

Définissons le vecteur z par $z_i = \lambda_i$, $i \in I(\hat{x})$ et $z_i = 0$ sinon. Alors z est une solution duale admissible ($z \geq 0$, $A^\top z + c = 0$) et

$$-g^\top z = -\sum_{i \in I(\hat{x})} g_i z_i = -\sum_{i \in I(\hat{x})} (f_i^\top \hat{x}) z_i = -z^\top F \hat{x} = c^\top \hat{x}.$$

d'où z est une solution duale optimale et $p^* = d^*$.

ii) Le problème (P) peut être mis sous la forme (P_i) en posant

$$F = \begin{bmatrix} A \\ -A \\ -I \end{bmatrix}, \quad g = \begin{bmatrix} b \\ -b \\ 0 \end{bmatrix}.$$

D'après le résultat précédent, il existe $z \geq 0$ tel que $F^\top z + c = 0$ et $-g^\top z = c^\top \hat{x}$. Considérons la partition suivante du vecteur z :

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}, \quad \text{avec } z_1, z_2, z_3 \in \mathbb{R}^m \text{ et } \geq 0,$$

et posons $p = z_2 - z_1$. Alors,

$$A^\top z + c = 0 \Leftrightarrow -A^\top p - z_3 + c = 0 \Rightarrow A^T p \leq c,$$

et

$$-g^\top z = c^\top x^* \Leftrightarrow b^\top p = c^\top \hat{x}.$$

Lien entre la méthode du simplexe et la dualité

On suppose que le problème (P) admet une solution optimale et que l'algorithme du simplexe s'est terminé en donnant une base pour laquelle les coûts réduits $\bar{c}_j = c_j - c_B^\top A_B^{-1} a_j$ associés aux indices non basiques sont tous positifs. Le vecteur des coûts réduits $\bar{c}_N = c_N - A_N^\top A_B^{-T} c_B$ n'a donc que des composantes positives ou nulles (M^{-T} représente la transposée de l'inverse de M qui est aussi l'inverse de la transposée de M).

Posons $p = A_B^{-T} c_B$.

On a, d'une part,

$$A^\top p = \begin{bmatrix} A_B^\top \\ A_N^\top \end{bmatrix} \left(A_B^{-1}\right)^\top c_B = \begin{bmatrix} c_B \\ A_N^\top \left(A_B^{-1}\right)^\top c_B \end{bmatrix}$$

Puisque les coûts réduits associés aux composantes non basiques sont positives ou nulles, on a $A^\top p \leq c$: p est un vecteur admissible pour le dual.

D'autre part,

$$p^\top b = c_B^\top A_B^{-1} b,$$

c'est-à-dire la valeur optimale du problème primal. Par dualité forte, le vecteur $p = A_B^{-T} c_B$ est donc une solution optimale du problème dual.

Applications de la dualité

– **Conditions d'optimalité** : Une solution primale admissible x est optimale si, et seulement si, il existe un vecteur $p \in \mathbb{R}^m$ tel que $A^\top p \leq c$ tel que

$$x_j(c_j - p^\top A_j) = 0, \quad j = 1, \ldots, n \text{ (conditions de complémentarité)}$$

On peut reformuler ce résultat en disant qu'un point $x \in \mathbb{R}^n$ est une solution de (P) si et seulement si il existe $p \in \mathbb{R}^m$ et $s \in \mathbb{R}^n$ tels que

$$\begin{cases} A^\top p + s = c, \quad s \geq 0 \\ Ax = b, \quad x \geq 0 \\ x^\top s = 0 \end{cases}$$

Remarque : on retrouve les conditions d'optimalité de Karush, Kahn et Tucker données en 2.3, page 48.

– **Analyse de la sensibilité** : Que devient la valeur optimale d'un problème lorsqu'on modifie ses données (par exemple, les contraintes) ? Considérons le problème perturbé suivant :

$$(P_\varepsilon) : \begin{cases} \text{minimiser} & c^\top x \\ s.l.c. \quad x \in \mathbb{R}^n & Ax = b + \varepsilon d, \quad x \geq 0 \end{cases}$$

où le vecteur $d \in \mathbb{R}^m$ est donné.

– *Analyse globale.* Si $\hat{p}$ est la solution optimale du problème dual de (P), alors c'est aussi une solution duale admissible pour le problème dual de (P_ε). Par dualité faible, la valeur optimale $p^*(\varepsilon)$ du problème perturbé vérifie donc :

$$p^*(\varepsilon) \geq (b + \varepsilon d)^\top \hat{p} = p^* + \varepsilon d^\top \hat{p}.$$

– *Analyse locale.* On suppose que l'ensemble des solutions optimales de (P) contient un sommet non dégénéré $\hat{x}$, soit B la base associée. On a alors

$$\hat{x}_B = A_B^{-1} b > 0, \quad \hat{x}_N = 0.$$

La condition d'optimalité s'exprime alors sous la forme

$$c_N^\top - c_B^\top A_B^{-1} A_N \geq 0.$$

Notons que cette condition est indépendante du vecteur b.
Considérons alors la solution du problème perturbé (P_ε) en conservant la base B : définissons la solution $\hat{x}(\varepsilon)$ par ses composantes basiques et non basiques :

$$\hat{x}_B(\varepsilon) = A_B^{-1}(b + \varepsilon d), \quad \hat{x}_N(\varepsilon) = 0.$$

La condition d'optimalité étant remplie, il suffit de s'assurer que $\hat{x}(\varepsilon)$ est admissible (c-à-d $\hat{x}_B(\varepsilon) \geq 0$), ce qui est vrai si ε est suffisamment petit. La valeur du problème perturbé vaut alors

$$\begin{aligned} p^*(\varepsilon) &= c_B^\top A_B^{-1}(b + \varepsilon d) = p^* + \varepsilon c_B^\top A_B^{-1} d \\ &= p^* + \varepsilon \hat{p}^\top d \end{aligned}$$

Le nombre $\hat{p}_i$ est donc la sensibilité du coût par rapport au membre de droite de la $i^{ème}$ contrainte, on rencontre aussi l'expression de "coût marginal" associé à la $i^{ème}$ contrainte.

2.6 Programmation semi-définie - LMI

Inégalités matricielles linéaires (LMI)

Un programme semi-défini est un problème d'optimisation pouvant être mis sous la forme :

$$(SDP) \quad \begin{cases} \text{minimiser} & c^\top x \\ s.l.c. \quad x \in \mathbb{R}^n : & F_0 + x_1 F_1 + \cdots + x_n F_n \succeq 0 \end{cases}$$

où les F_i, pour $i = 0, 1, \ldots, n$, sont des matrices à coefficients réels, symétriques et de même dimension et où l'inégalité $A \preceq 0$ signifie que la matrice symétrique A est semi-définie positive.

L'ensemble des contraintes de ce problème est définie par une *LMI*, c'est-à-dire une inégalité linéaire matricielle :

$$F_0 + x_1 F_1 + \cdots + x_n F_n \succeq 0 \tag{2.29}$$

Dans la suite, on notera $x = [x_1, \ldots, x_n]^T$ et $F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n$. La LMI (2.29) s'écrit alors de manière plus compacte sous la forme $F(x) \succeq 0$.

Remarquons qu'un système de LMI ($F_1(x) \prec 0$, ..., $F_N(x) \prec 0$ peut se récrire sous la forme d'une seule LMI

$$\operatorname{diag}(F_1(x), \ldots, F_N(x)) \succeq 0$$

Dans la plupart des applications, on utilisera plutôt des variables matricielles. Par exemple, en théorie de la stabilité des systèmes linéaires, l'existence d'une matrice P symétrique et définie positive vérifiant l'inégalité

$$A^T P + PA \prec 0 \tag{2.30}$$

est une condition nécessaire et suffisante pour assurer la convergence vers 0 de toutes les solutions du système d'équations différentielles ordinaires

$$\dot{x}(t) = Ax(t) \tag{2.31}$$

où $x(t) \in \mathbb{R}^n$ et $A \in \mathbb{R}^{n \times n}$. L'inégalité 2.30 sera dite une LMI en la variable P. Pour obtenir une forme obéissant *stricto sensu* à la définition donnée précédemment, il suffit de décomposer la matrice P dans une base de l'espace vectoriel composé des matrices symétriques de dimension n. Par exemple, pour $n = 2$, on aura la décomposition

$$P = \begin{pmatrix} x_1 & x_2 \\ x_2 & x_3 \end{pmatrix} = x_1 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Deux types de question peuvent être envisagé concernant le problème (SDP) :
- Réalisabilité : Déterminer si l'ensemble des x vérifiant la LMI (2.29) est vide ou non (et s'il est non vide, donner un de ses éléments)
- Optimisation : résoudre le problème (SDP).

Remarque : on peut montrer facilement (en raisonnant sur la forme quadratique associée) que c'est bien un problème d'optimisation convexe, c'est-à-dire, l'ensemble des x tels que $F(x) \succeq 0$ est un sous-ensemble convexe de $\mathbb{R}^n$.

Quelques outils pour LMI

Transformation de congruence - Formule du complément de Schur

Deux matrices M, N sont dites congruentes s'il existe une matrice inversible T telle que $N = T^\top N T$. On alors la proposition suivante

Proposition 2.6.1. *Soit M et N deux matrices symétriques congruentes alors $M \succ 0$ si et seulement si $N \succ 0$.*

La démonstration est laissée à titre d'exercice au lecteur.

Soit M une matrice symétrique admettant la partition

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

où M_{11} est une matrice carrée alors, en effectuant une transformation de congruence avec la matrice partitionnée

$$T = \begin{bmatrix} I & -M_{11}^{-1}M_{12} \\ 0 & I \end{bmatrix}$$

on établit la proposition suivante

Proposition 2.6.2 (Complément de Schur)**.** *Sous les hypothèses précédentes, on a $M \succ 0$ si et seulement si*

$$\begin{cases} M_{11} \succ 0 \\ M_{22} - M_{21}M_{11}^{-1}M_{12} \succ 0 \end{cases}$$

En effectuant cette fois une transformation de congruence avec la matrice

$$T' = \begin{bmatrix} I & 0 \\ -M_{22}^{-1}M_{21} & I \end{bmatrix}$$

on obtient

Proposition 2.6.3 (Complément de Schur)**.** *Sous les hypothèses précédentes, on a $M \succ 0$ si et seulement si*

$$\begin{cases} M_{22} \succ 0 \\ M_{11} - M_{12}M_{22}^{-1}M_{21} \succ 0 \end{cases}$$

Lemme d'élimination

Lemme 2.6.1 (Lemme d'élimination)**.** *Pour des matrices réelles $W = W^T$, M, N de taille appropriée, les quatre propriétés suivantes sont équivalentes :*

1. *Il existe une matrice réelle K telle que :*

$$W + MKN^T + NK^TM^T < 0.$$

2. *Il existe deux matrices U, V telles que :*

$$W + MU + U^T M^T < 0 \quad \text{et} \quad W + NV + V^T N^T < 0.$$

3. *Il existe un scalaire σ tel que*

$$W < \sigma M M^T \quad \text{et} \quad W < \sigma N N^T,$$

4. *Les compléments orthogonaux $M_\perp$ et $N_\perp$ de M et N, respectivement, vérifient*

$$M_\perp^T W M_\perp < 0 \quad \text{et} \quad N_\perp^T W N_\perp < 0.$$

Rappelons que le *complément orthogonal* d'une matrice M est une matrice $M_\perp$ de rang maximal telle que $M_\perp^T M = 0$.

2.7 Bibliographie

[1] Ben Tal, A. et A. Nemirovski: *Optimization I-II : convex analysis, nonlinear programming theory, nonlinear programming algorithms.* rapport technique, Department ISYE, Georgia Institute of Technology, 2004. Lecture notes disponibles à URL : http ://www2.isye.gatech.eduñemirovs/.

[2] Bonnans, J.F., J.C. Gilbert, C. Lemaréchal et C.A. Sagastizábal: *Numerical Optimization. Theoretical and Practical Aspects.* Springer, 2nd édition, 2006.

[3] Boyd, S., L. El Ghaoui, E. Feron et V. Balakrishnan: *Linear Matrix Inequalities in Systems and Control Theory.* SIAM, 1994. Disponible sur internet à l'adresse `http://www.stanford.edu/~boyd/books.html`.

[4] Boyd, S. et L. Vandenberghe: *Convex Optimization.* Cambridge University Press, 2004. Disponible sur internet à l'adresse URL : http ://www.stanford.edu/ boyd/cvxbook/.

[5] Meinsma, G.: *Interior Point Methods.* rapport technique, Systems and Control Group, Dept. of applied Mathematics, Univ. of Twente, 1997. disponible à URL : http ://wwwhome.math.utwente.nl/ meinsmag/courses/.

[6] Nemirovski, A.: *Lectures on Modern Convex Optimization.* rapport technique, Department ISYE, Georgia Institute of Technology, 2005. Lecture notes disponibles à URL : http ://www2.isye.gatech.eduñemirovs/.

[7] Nesterov, Y. et A. Nemirovsky: *Interior point polynomial methods in convex programming : Theory and Applications.* SIAM, Philadelphie, 1994.

[8] Nocedal, J. et S. Wright: *Numerical Optimization.* Springer Series in Operation Research and Financial Engineering. Springer, 2nd édition, 2006.

[9] Scherer, C. et S. Weiland: *Disc Course on Linear Matrix Inequalities in Control 2004/2005.* rapport technique, DISC graduate course, 2004.

[10] Vanderbei, R.J.: *Linear Programming. Foundations and Extensions.* Kluwer's international series, 2nd édition, 2001.

3 | Systèmes stochastiques

A. Achour[1], M. Ksouri[2], S. Salhi[3] et S. Ben Attia[2]

[1]Faculté des Sciences de Tunis, FST Campus Universitaire, 2092 El Manar, Tunis, Tunisie. *E-mail* : Abdennebi.Achour@fst.rnu.tn
[2]ENIT, BP 37, Le Belvédère 1002, Tunis, Tunisie. *E-mail* : Mekki.Ksouri@insat.rnu.tn, benattiaselma@yahoo.fr
[3]Institut Supérieur d'Informatique de Tunis, 2 rue Abourraihan Al Bayrouni, 2080 Ariana, Tunis, Tunisie. *E-mail* : salhis@lycos.com

3.1 Introduction aux probabilités

Si on jette un dé, le résultat de l'*épreuve* est un élément de l'ensemble

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

L'espace Ω est appelé *univers* des possibles, ses sous-ensembles sont appelés *événements* et les événements qui ne contiennent qu'un seul élément sont appelés *événements élémentaires.*

On dit que l'événement A (A est donc une partie de Ω) est réalisé si le résultat du jet du dé appartient à A.

Le but du calcul des probabilités est d'associer à chaque événement un nombre qui mesure en quelque sorte la chance qu'a cet événement de se réaliser. Intuitivement, lorsqu'on dit que la probabilité de l'événement A vaut a, cela implique que si l'on répète l'expérience (si on jette le dé) un grand nombre n de fois, l'événement A va se réaliser, à peu-près, $n \times a$ fois. Cela justifie la définition suivante :

Une *probabilité* est une application $P : \mathcal{P}(\Omega) \to [0, 1]$ vérifiant les propriétés suivantes :

$$P(\phi) = 0\ , P(\Omega) = 1\, et\ P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

La probabilité la plus classique est celle qu'on appelle *probabilité uniforme :* elle

est définie par :

$$P(A) = \frac{\text{nombre d'éléments de } A}{\text{nombre d'éléments de } \Omega}$$

Les événements élémentaires ont alors tous la même probabilité de se réaliser et on dit qu'ils sont *équiprobables.* En l'absence d'indications contraires, c'est cette probabilité que l'on utilisera.

Exemple

On jette deux dés et on marque un point si le résultat total est supérieur ou égal à 10. On recommence 600 fois la même opération. Combien de points pensez-vous ainsi approximativement marquer ?

Le tableau ci-contre donne les différents totaux que l'on peut obtenir ainsi que la manière de les obtenir.

Le total 4 s'obtient de 3 façons :

$$4 = 3 + 1 = 2 + 2 = 1 + 3$$

et la probabilité d'avoir un total égal à 4 est donc

$$P(4) = 3/36 = 1/12.$$

	A	B	C	D	E	F
A	2	3	*4*	5	6	7
B	3	*4*	5	6	7	8
C	*4*	5	6	7	8	9
D	5	6	7	8	9	10
E	6	7	8	9	**10**	11
F	7	8	9	10	11	12

Notons T la somme des points obtenus, c'est une *variable aléatoire* dont *la loi de probabilité* est donnée par le tableau ci-dessous.

T	B	C	D	E	F	G	H	I	Š	‹	Œ
P_T	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Ce tableau permet de calculer la probabilité pour que la somme des points soit supérieure ou égale à 10 :

$$P(T \geq 10) = P(T = 10) + P(T = 11) + P(T = 12) = \frac{3 + 2 + 1}{36} = \frac{1}{6}$$

Pour 600 jets des 2 dés, on peut espérer marquer $600 \times \frac{1}{6} = 100$ points.

Le jeu de pile ou face

On dispose d'une pièce équilibrée dont les faces sont marquées 1 et 0. On la lance huit fois de suite et on marque chaque fois le chiffre obtenu. On peut obtenir par exemple le résultat 11010100. Il y a $2^8 = 256$ issues possibles qui sont toutes équiprobables puisqu'on a supposé que la pièce était équilibrée.

▶ *Quelle est la probabilité d'obtenir exactement trois fois le nombre* 1 ?

Déterminons le nombre de cas favorables : Pour avoir exactement trois fois le nombre 1, il nous faut choisir trois places parmi huit pour mettre les 1 et remplir ensuite le reste par des 0. Cela se fait de $\binom{8}{3} = \frac{8!}{3!5!} = 56$ façons différentes et la probabilité demandée vaut :

$$p = \binom{8}{3} 2^{-8} = 0.21875$$

▶ *Quelle est la probabilité d'obtenir exactement quatre séries* ?

Rappelons d'abord ce qu'est une série dans un jeu de pile ou face. Le résultat précédent 11010100 se compose de six séries :

$$11\lrcorner 0\lrcorner 1\lrcorner 0\lrcorner 1\lrcorner 00$$

On a utilisé 5 séparateurs $\lrcorner$, l'autre résultat correspondant à la même répartition des séparateurs est

$$00\lrcorner 1\lrcorner 0\lrcorner 1\lrcorner 0\lrcorner 11.$$

Pour compter les résultats comportant 4 séries, on compte le nombre de façons de mettre 3 séparateurs dans les 7 places séparant les huit chiffres et on multiplie ensuite par 2. Il y a donc $2 \times \binom{7}{3} = 70$ résultats comportant 4 séries et la probabilité demandée vaut

$$p = 70 \times 2^{-8} = 0.27344$$

▶ *Quelle est la probabilité d'obtenir au moins une série de deux* 1 ?

Nous allons calculer cette probabilité lorsqu'on lance n fois la pièce (la question initiale pour 8 lancers étant difficile, on préfère étudier le cas général pour pouvoir examiner d'abord les cas où n est petit).

Notons A_n l'ensemble de cas défavorables et a_n le nombre d'éléments de A_n.Pour $n = 1, 2, 3$ et 4 , on a le tableau ci-dessous :

n	A_n	a_n
1	0,1	2
2	00,10,01	3
3	000,100,010,001,101	5
4	0000,1000,0100,0010,0001,1010,1001,0101	8

Cela suggère que les a_n vérifient la relation $a_{n+2} = a_{n+1} + a_n$ et que A_{n+2}

s'obtient donc à partir de A_{n+1} et A_n. Le même tableau montre que l'on obtient les éléments de A_3 en faisant suivre les éléments de A_2 de 0 et les éléments de A_1 de 01. Il est maintenant facile de se convaincre que A_{n+2} est la réunion disjointe de l'ensemble $A_{n+1}0$ (dont les éléments s'obtiennent en faisant suivre ceux de A_{n+1} par 0) et de l'ensemble A_n01 (dont les éléments s'obtiennent en faisant suivre ceux de A_n de 01). Les probabilités cherchées sont donc données par :

$$a_1 = 2,\ a_2 = 3,\ a_{n+2} = a_{n+1} + a_n \ \textit{et}\ p_n = 1 - a_n 2^{-n} = 2^{-n}(2^n - a_n)$$

On a donc pour les premières valeurs de n :

n	1	2	3	4	5	6	7	8	9	10
a_n	2	3	5	8	13	21	34	55	89	144
2^n	2	4	8	16	32	64	128	256	512	1024
p_n	0	0.25	0.375	0.5	0.594	0.672	0.734	0.785	0.826	0.859

Probabilité conditionnelle

La probabilité d'un événement (la chance qu'a cet événement de se réaliser) change lorsque nos informations concernant l'expérience augmentent. Ainsi la probabilité d'obtenir un 6 en lançant un dé vaut à priori $\frac{1}{6}$, mais elle devient égale à $\frac{1}{3}$ si on nous dit que le nombre obtenu est pair et 0 si on nous dit qu'il est impair. Lorsque A et B sont deux événements quelconques, la probabilité que A se réalise sachant que B est réalisé est définie par

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Cela implique $P(A \cap B) = P(B) \times P(A \mid B)$ et c'est ce qu'on appelle le *principe des probabilités composées.*

Noter que l'application qui à A associe $P(A \mid B)$ est aussi une probabilité sur Ω.

Formule de Bayes

La formule de Bayes, quoique triviale, est d'un intérêt considérable puisqu'elle permet de modifier notre estimation des probabilités en fonction des informations nouvelles. Soit $E_1, E_2, \cdots, E_n$ un système complet d'événements (Cela veut dire qu'ils sont deux à deux disjoints et que leur réunion est égale à Ω tout entier.). Pour tout autre événement A , on peut écrire :

$$P(A \cap E_k) = P(A)P(E_k \mid A)\ \ ;\ \ P(A) = \sum_{i=1}^{n} P(A \cap E_i) = \sum_{i=1}^{n} P(E_i)P(A \mid E_i)$$

On en tire la formule de Bayes :

$$P(E_k \mid A) = \frac{P(A \cap E_k)}{P(A)} = \frac{P(E_k)P(A \mid E_k)}{P(E_1)P(A \mid E_1) + \cdots + P(E_n)P(A \mid E_n)}$$

Exemple

On mène l'enquête sur les causes d'une catastrophe aérienne. On peut émettre trois hypothèses H_1, H_2 et H_3 dont les probabilités à priori (avant le début de l'enquête) sont $P(H_1) = 0,7$, $P(H_2) = 0,2$ et $P(H_3) = 0,1$. L'enquête a révélé que le combustible a brûlé (notons A cet événement). Des données statistiques nous permettent d'écrire

$$P(A \mid H_1) = 0,1 \quad , \quad P(A \mid H_2) = 0,3 \quad et \quad P(A \mid H_3) = 1.$$

Déterminer l'hypothèse la plus probable de l'accident.

Les probabilités $P(H_i \mid A)$ s'obtiennent grâce à la formule de Bayes :

i	1	2	3
$P(H_i \mid A)$	0,304	0,261	0,435

et l'hypothèse la plus probable de l'accident est donc l'hypothèse H_3.

Exemple

Deux urnes d'apparence identique contiennent l'une 300 boules blanches et 700 boules noires et l'autre 700 boules blanches et 300 boules noires.

► *Je choisis une des deux urnes, êtes-vous prêt à parier* 10 *millimes contre* 10 *millimes que l'urne choisie est celle qui contient* 300 *boules blanches* ?

► *Je choisis une des deux urnes, je tire, avec remise*, 12 *boules de cette urne et j'obtiens* 4 *boules blanches et* 8 noires. *Etes-vous prêt à parier* 10 *millimes contre* 10 *que l'urne choisie contient* 700 *boules blanches* ? *Sinon, acceptez-vous de parier* 5 *millimes contre* 10 ?

Pour la première question, vous pouvez accepter le jeu puisque vous gagnerez 10 millimes avec une probabilité égale à 0,5 et perdrez 10 millimes avec la même probabilité et le jeu est donc équitable.

Pour la deuxième question, on est en droit de penser que l'urne choisie contient plus de boules noires que de boules blanches et qu'il devient imprudent de parier à 10 contre 10.Voyons vos chances de gagner si vous pariez à 5 contre 10. Notons U_1 l'urne qui contient 300 boules blanches, U_2 l'autre urne, B l'événement ‹‹tirer une boule blanche›› et N l'événement ‹‹tirer une boule noire››. Par hypothèse, on a :

$P(U_1) = 0,5$	$P(B \mid U_1) = 0,3$	$P(N \mid U_1) = 0,7$
$P(U_2) = 0,5$	$P(B \mid U_2) = 0,7$	$P(N \mid U_2) = 0,3$

Notons A l'événement ‹‹tirer 4 boules blanches et 8 boules noires››, on a :

$$P(A \mid U_1) = \binom{12}{4} \times 0,3^4 \times 0,7^8 \ \ ; \ \ P(A \mid U_2) = \binom{12}{4} \times 0,3^8 \times 0,7^4$$

La formule de Bayes permet maintenant de calculer $P(U_1 \mid A)$ et $P(U_2 \mid A)$:

$$P(U_1 \mid A) = \frac{P(U_1) \times P(A \mid U_1)}{P(U_1) \times P(A \mid U_1) + P(U_2) \times P(A \mid U_2)}$$

On trouve $P(U_1 \mid A) = 0,96737$ et donc $P(U_2 \mid A) = 1 - 0,96737 = 0,03263$. Cela implique, par exemple, que si vous jouez 100000 parties, vous allez, approximativement, perdre $5 \times 100000 \times 0,96737$ millimes et gagner $10 \times 100000 \times 0,03263$, soit en tout perdre 451045 millimes et vous ferez mieux de refuser de jouer.

Espérance et Ecart Type

Dans l'exemple précédent, on a vu qu'en acceptant de jouer 100000 parties à 5 contre 10, le gain probable était :

$$100000 \times (10 \times 0.03263 - 5 \times 0.96737) = -451045$$

Le gain moyen par partie est de $10 \times 0.03263 - 5 \times 0.96737 = -4,51045$ et on dit que l'espérance de gain est de $-4,51045$.

Plus généralement, l'espérance d'une variable aléatoire $X : \Omega \to R$ est le nombre $E(X)$ défini par :

$$E(X) = \sum_{\omega \in \Omega} X(\omega) \times P(\omega)$$

On définit aussi la variance $V(X)$ et l'écart type $\sigma(X)$ de X par :

$$V(X) = E(\,[X - E(X)]^2\,) \ , \ \sigma(X) = \sqrt{V(X)}$$

Noter que l'espérance de X représente le nombre moyen que prend la fonction X et que la variance de X n'est nulle que si X ne varie pas (X est donc constante sauf peut être sur un ensemble de probabilité nulle).

Exemple : La loi binomiale.

Si X prend les valeurs $k = 0, 1, ..., n$ avec les probabilités respectives

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

où p est compris entre 0 et 1 et si g est la fonction définie par

$$g(x) = \sum_{k=0}^{n} p_k x^k = (px + 1 - p)^n$$

alors on peut calculer $E(X)$ et $V(X)$ de la manière suivante :

★ $E(X) = \sum_{k=0}^{n} kp_k = g'(1) = np$.

★ $E(\,X(X-1)\,) = \sum_{k=0}^{n} k(k-1)p_k = g''(1) = n(n-1)p^2$.

★ $E(X^2) = E(\,X(X-1)\,) + E(X) = n(n-1)p^2 + np$.

★ $V(X) = E(X^2) - E(X)^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$

La loi des trois sigma

Dans la pratique, on adopte cette loi qui dit qu'il est fort improbable que la variable aléatoire X s'écarte de plus de $3\sigma(X)$ de sa valeur moyenne $E(X)$.

Exemple

On a demandé à un élève de lancer 50 fois un jeton équilibré dont les faces sont marquées 0 et 1 et de noter ce qu'il obtient. Il déclare avoir obtenu :

10101101001010100101100101011000101010011001010100

Qu'en pensez-vous ?

Notons X la variable aléatoire qui compte le nombre de 1 obtenus quand on lance 50 fois le jeton. X prend les valeurs $k = 0, 1, ..., 50$ avec les probabilités $p_k = \binom{50}{k} 2^{-50}$. On a donc $E(X) = 50 \times 2^{-1} = 25$ et $V(X) = 50 \times 2^{-1} \times 2^{-1} = 12,5$ si bien que $\sigma(X) = \sqrt{12.5} = 3,5355$.

L'élève, a obtenu 23 fois le nombre 1. On a donc $X(\omega) = 23$ et comme $|\, X(\omega) - E(X)\,| = 2 < 3 \times \sigma(X)$,on peut considérer que la suite obtenue par l'élève est plausible si on se contente de ce test.

Notons maintenant Y la variable aléatoire qui compte le nombre de *séries* obtenues quand on lance 50 fois le jeton. Y prend les valeurs $k+1$, où $k = 0, 1, ..., 49$, avec les probabilités $q_k = \binom{49}{k} 2^{-49}$. On a donc

$E(Y) = 49 \times 2^{-1} + 1 = 25,5$ et $V(Y) = 49 \times 2^{-2} = 12,25$ si bien que $\sigma(Y) = 3.5$

Cette fois, on a $Y(\omega) = 38$ et $|\, Y(\omega) - E(Y)\,| = 12,5 \geq 3 \times \sigma(Y) = 11,5$ et on est en droit de penser que la liste donnée par l'élève est fictive et n'a pas été obtenue en lançant 50 fois le jeton.

Bibliographie

Pour s'initier aux probabilités, on conseille les livres :

▷ Les probabilités.Par Albert Jacquard, Que sais-je ? n°1571

▷ L'enseignement des probabilités et de la statistique Vol 1, Arthur Engel. Collection : Cedic.

▷ Statistique et Probabilités pour aujourd'hui, Irving Adler. Edition O.C.D.L., Paris.

Chacun de ces trois livres constitue une très bonne introduction aux probabilités et donnera, sans aucun doute, satisfaction aux enseignants qui doivent initier les élèves au calcul des probabilités qui est devenu un des outils mathématiques les plus utilisés.

Et pour finir, on vous propose ce jeu "probabiliste" qu'on peut trouver dans le livre d'Engel :

Les faces des quatre dés, A, B, C, D sont marquées comme l'indique la figure ci-dessus :

	0			3			2			5				
4	0	4	3	3	3	2	2	2	1	1	1			
	4			3			6			5				
	4			3			6			5				
	A			B			C			D				

Il y a deux joueurs, l'un des deux choisit un dé, puis l'autre choisit un des trois dé restants. Chacun jette son dé. Celui qui obtient le nombre le plus grand est déclaré vainqueur. Ce qui étonne dans ce jeu, c'est que celui qui choisit le premier est désavantagé. Montrer qu'en choisissant judicieusement, le second joueur gagne avec une probabilité égale à $\frac{2}{3}$.

3.2 Probabilités

Tribus et probabilités

Soit Ω un ensemble non vide. Une **tribu** de Ω est un ensemble $\mathcal{T}$ de parties de Ω vérifiant les trois propriétés suivantes

- L'ensemble vide $\varnothing$ et l'ensemble Ω lui-même sont dans $\mathcal{T}$.
- Si A est dans $\mathcal{T}$, il en est de même de son complémentaire $A^c = \Omega \setminus A$.
- Si les ensembles A_n $(n \geq 0)$ sont dans $\mathcal{T}$, leur réunion $\cup_{n=0}^{\infty} A_n$ est dans $\mathcal{T}$.

Une **probabilité** sur une tribu $\mathcal{T}$ est une application $P : \mathcal{T} \longrightarrow [0,1]$ vérifiant

- $P(\varnothing) = 0$ et $P(\Omega) = 1$.
- Si les A_n sont dans $\mathcal{T}$ et s'ils sont deux à deux disjoints, alors

$$P\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=0}^{\infty} P(A_n)$$

Exemple. **Masse de Dirac**.

L'ensemble $\mathcal{P}(\Omega)$ de toutes les parties de Ω et une tribu et si a est un élément quelconque de Ω, l'application $\delta_a : \mathcal{P}(\Omega) \longrightarrow [0,1]$ définie par

$$\delta_a(A) = 1 \text{ si } a \in A \text{ et } \delta_a(A) = 0 \text{ sinon}$$

est une probabilité (appelée masse de Dirac au point a) sur $\mathcal{P}(\Omega)$.

De la même manière si a_n est une suite d'éléments dans Ω et si les réels positifs α_n sont de somme égale à 1, l'application

$$\sum_{n=0}^{\infty} \alpha_n \delta_{a_n} : \mathcal{P}(\Omega) \longrightarrow [0,1] \quad , \quad \left(\sum_{n=0}^{\infty} \alpha_n \delta_{a_n}\right)(A) = \sum_{n=0}^{\infty} \alpha_n . \delta_{a_n}(A)$$

est une probabilité sur $\mathcal{P}(\Omega)$.

Exemple. **La tribu des boréliens**.

- La tribu des boréliens de $\mathbf{R}$, notée $\mathcal{B}(\mathbf{R})$, est la plus petite tribu de $\mathbf{R}$ contenant tous les intervalles de $\mathbf{R}$. Si la fonction $f : \mathbf{R} \longrightarrow [0, \infty[$ vérifie

$$\int_{\mathbf{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1$$

alors l'application

$$P : \mathcal{B}(\mathbf{R}) \longrightarrow [0,1] \quad , \quad P(A) = \int_A f(x)dx$$

est une probabilité sur $\mathcal{B}(\mathbf{R})$ et on dit qu'elle est de densité f par rapport à la mesure de Lebesgue.

- De la même manière, la tribu des boréliens de $\mathbf{R}^n$, notée $\mathcal{B}(\mathbf{R}^n)$, est la plus petite tribu de $\mathbf{R}^n$ contenant tous les ensembles R de la forme

$$R = I_1 \times \cdots \times I_n \quad , \quad I_j \text{ est un intervalle de } \mathbf{R}$$

Si $f : \mathbf{R}^n \longrightarrow [0, \infty[$ vérifie $\int_{\mathbf{R}^n} f(x)dx = 1$, alors l'application

$$P : \mathcal{B}(\mathbf{R}^n) \longrightarrow [0,1] \quad , \quad P(A) = \int_A f(x)dx$$

est une probabilité sur $\mathcal{B}(\mathbf{R}^n)$ et on dit qu'elle est de densité f par rapport à la mesure de Lebesgue.

Exemple. **Tribu produit**.

- Soit $(\Omega_1, \mathcal{T}_1)$ et $(\Omega_2, \mathcal{T}_2)$ deux ensembles probabilisables (i.e. Ω_i est un ensemble et $\mathcal{T}_i$ est une tribu de Ω_i). Un rectangle de l'ensemble produit $\Omega_1 \times \Omega_2$ est un ensemble de la forme

$$A_1 \times A_2 \quad , \quad A_1 \in \mathcal{T}_1 \text{ et } A_2 \in \mathcal{T}_2$$

On notera $\mathcal{T}_1 \otimes \mathcal{T}_2$ (lire $\mathcal{T}_1$ tensoriel $\mathcal{T}_2$) la plus petite tribu de $\Omega_1 \times \Omega_2$ contenant tous les rectangles de $\Omega_1 \times \Omega_2$.

Si $P_1 : \mathcal{T}_1 \longrightarrow [0,1]$ et $P_2 : \mathcal{T}_2 \longrightarrow [0,1]$ sont deux probabilités, on notera

$$P_1 \otimes P_2 : \mathcal{T}_1 \otimes \mathcal{T}_2 \longrightarrow [0,1]$$

l'unique probabilité sur la tribu $\mathcal{T}_1 \otimes \mathcal{T}_2$ qui vérifie

$$P_1 \otimes P_2(A_1 \times A_2) = P_1(A_1).P_2(A_2)$$

- De la même manière, si $(\Omega_i, \mathcal{T}_i)$ $(1 \leq i \leq n)$ sont des espaces probabilisables, on notera $\otimes_{i=1}^n \mathcal{T}_i = \mathcal{T}_1 \otimes \cdots \otimes \mathcal{T}_n$ la plus petite tribu de l'espace produit $\prod_{i=1}^n \Omega_i = \Omega_1 \times \cdots \times \Omega_n$ contenant tous les rectangles

$$\prod_{i=1}^n A_i = A_1 \times \cdots \times A_n \quad , \quad A_i \in \mathcal{T}_i$$

Si $P_i : \mathcal{T}_i \longrightarrow [0,1]$ sont des probabilités, on notera $\otimes_{i=1}^n P_i = P_1 \otimes \cdots \otimes P_n$ l'unique probabilité sur $\mathcal{T}_1 \otimes \cdots \otimes \mathcal{T}_n$ qui vérifie

$$P_1 \otimes \cdots \otimes P_n(A_1 \times \cdots \times A_n) = P_1(A_1) \cdots P_n(A_n)$$

- Soit $(\Omega_n, \mathcal{T}_n)$ une suite infinie d'espaces probabilisables et $\Omega = \prod_{n=1}^\infty \Omega_n$ le produit des ensembles Ω_n.

Un élément ω de $\Omega = \prod_{n=1}^\infty \Omega_n$ s'écrit donc

$$\omega = (\omega(1), \omega(2), \ldots, \omega(n), \ldots) \quad , \quad \omega(n) \in \Omega_n$$

On appelle cylindre de $\Omega = \prod_{n=1}^\infty \Omega_n$, tout ensemble de la forme

$$C = A_1 \times \cdots \times A_m \times \Omega_{m+1} \times \Omega_{m+2} \times \cdots \quad , \quad m \geq 1 \ \text{ et } \ A_i \in \mathcal{T}_i$$

La plus petite tribu contenant ces cylindres est notée $\otimes_{i=1}^\infty \mathcal{T}_i$ et l'unique probabilité P vérifiant

$$P(A_1 \times \cdots \times A_m \times \Omega_{m+1} \times \Omega_{m+2} \times \cdots) = P_1(A_1). \ldots . P_m(A_m)$$

pour tout cylindre est notée $\otimes_{i=1}^\infty P_i$.

Exemple. **Image directe**.

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé (i.e. $\mathcal{T}$ est une tribu de Ω et P est une probabilité sur $\mathcal{T}$).

Soit $X : \Omega \longrightarrow Z$ une application quelconque de Ω dans un ensemble quelconque Z.

- L'ensemble

$$X_* \mathcal{T} = \{B \subset Z \ ; \ X^{-1}(B) \in \mathcal{T}\}$$

est une tribu de Z appelée **image directe** de la tribu $\mathcal{T}$ par l'application X.

- L'application $X_* P : X_* \mathcal{T} \longrightarrow [0,1]$ définie par

$$X_* P(B) = P(\ X^{-1}(B)\ )$$

est une probabilité sur $X_* \mathcal{T}$ appelée image directe de la probabilité P par X.

Variables aléatoires

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé.

Une **variable aléatoire** est une application $X : \Omega \longrightarrow [-\infty, +\infty]$ vérifiant

$$\{X < a\} = \{\omega \in \Omega \ ; \ X(\omega) < a\} \in \mathcal{T}$$

pour tout $a \in \mathbf{R}$.

Exemple. **Fonctions étagées**.

Si $A \in \mathcal{T}$, sa fonction caractéristique $\mathbf{1}_A$ qui est définie par

$$\mathbf{1}_A(\omega) = 1 \text{ si } \omega \in A \text{ et } \mathbf{1}_A(\omega) = 0 \text{ sinon}$$

est une variable aléatoire et on pose

$$\int_\Omega \mathbf{1}_A(\omega) dP(\omega) = \int_A dP = P(A)$$

Une **fonction étagée** est une application $g : \Omega \longrightarrow]-\infty, +\infty]$ de la forme

$$g = \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i} \quad , \quad A_i \in \mathcal{T} \quad \text{et} \quad -\infty < \alpha_i \leq +\infty$$

C'est une variable aléatoire et on pose, pour une telle fonction,

$$\int_\Omega g(\omega) dP(\omega) = \int_\Omega \sum_{i=1}^n \alpha_i \mathbf{1}_{A_i}(\omega) dP(\omega) = \sum_{i=1}^n \alpha_i \int_\Omega \mathbf{1}_{A_i}(\omega) dP(\omega) = \sum_{i=1}^n \alpha_i P(A_i)$$

où l'on convient que $\infty \times P(A_i) = \infty$ si $P(A_i) > 0$ et $\infty \times P(A_i) = 0$ si $P(A_i) = 0$.

Intégrale d'une variable aléatoire positive.

Si $X : \Omega \longrightarrow [0, +\infty]$ est une variable aléatoire positive, son intégrale est par définition l'élément de $[0, \infty]$ défini par

$$\int_\Omega X(\omega) dP(\omega) = \sup \left\{ \int_\Omega g(\omega) dP(\omega) \; ; \quad g \text{ étagée et } 0 \leq g \leq X \right\}$$

Cette définition implique aussitôt ces deux importants résultats :

Théorème de convergence monotone.

Si $X_n : \Omega \longrightarrow [0, +\infty]$ est une suite croissante de variables aléatoires positives, alors

$$\lim_{n \to \infty} \int_\Omega X_n(\omega) dP(\omega) = \int_\Omega \lim_{n \to \infty} X_n(\omega) dP(\omega)$$

Théorème de Beppo-Levy.

Si $X_n : \Omega \longrightarrow [0, +\infty]$ est une suite de variables aléatoires positives, alors

$$\int_\Omega \sum_{n=1}^\infty X_n(\omega) dP(\omega) = \sum_{n=1}^\infty \int_\Omega X_n(\omega) dP(\omega)$$

Une conséquence directe du théorème de convergence monotone est le lemme de Fatou qui est d'un usage fréquent

Lemme de Fatou.

Si $X_n : \Omega \longrightarrow [0, +\infty]$ est une suite de variables aléatoires positives, alors

$$\liminf \int_\Omega X_n(\omega) dP(\omega) \leq \int_\Omega \liminf X_n(\omega) dP(\omega)$$

Variables aléatoires intégrables.

Une variable aléatoire $X : \Omega \longrightarrow [-\infty, +\infty]$ est dite intégrable si elle vérifie

$$\int_\Omega |X(\omega)|\, dP(\omega) < \infty$$

et dans ce cas son intégrale est définie par

$$\int_\Omega X(\omega) dP(\omega) = \int_\Omega X(\omega) \mathbf{1}_{\{X \geq 0\}}(\omega) dP(\omega) - \int_\Omega -X(\omega) \mathbf{1}_{\{X < 0\}}(\omega) dP(\omega)$$

On montre, grâce au lemme de Fatou, cet important résultat

Théorème de convergence dominée.

Soit $X_n : \Omega \longrightarrow [-\infty, +\infty]$ une suite de variables aléatoires qui converge presque sûrement vers une variable aléatoire $X_\infty : \Omega \longrightarrow [-\infty, +\infty]$.

S'il existe une variable aléatoire positive $g : \Omega \longrightarrow [0, +\infty]$ vérifiant

$$|X_n(\omega)| \leq g(\omega) \quad , \quad \int_\Omega g(\omega) dP(\omega) < \infty$$

alors la limite X_∞ est intégrable et on a

$$\lim_{n\to\infty} \int_\Omega X_n(\omega) dP(\omega) = \int_\Omega \lim_{n\to\infty} X_n(\omega) dP(\omega) = \int_\Omega X_\infty(\omega) dP(\omega)$$

Corollaire.

Soit $X_n : \Omega \longrightarrow [-\infty, +\infty]$ une suite de variables aléatoires telle que

$$\sum_{n=1}^\infty \int_\Omega |X_n(\omega)|\, dP(\omega) < \infty$$

alors la série $\sum_{n=1}^\infty X_n(\omega)$ converge presque sûrement vers une variable aléatoire intégrable et on a

$$\int_\Omega \sum_{n=1}^\infty X_n(\omega).dP(\omega) = \sum_{n=1}^\infty \int_\Omega X_n(\omega).dP(\omega)$$

Notation.

Etant donné une variable aléatoire X positive ou intégrable et un élément A de la tribu $\mathcal{T}$, on pose

$$\int_A X(\omega) dP(\omega) = \int_\Omega X(\omega) \mathbf{1}_A(\omega) dP(\omega)$$

Théorème. Inégalité de Chebyshev.

Si $X : \Omega \longrightarrow [0, +\infty]$ est une variable aléatoire positive et si $a \in [0, \infty]$, alors

$$a \times P(X \geq a) \leq \int_\Omega X dP$$

Preuve. Cela résulte de

$$\int_\Omega X dP = \int_{\{X \geq a\}} X dP + \int_{\{X < a\}} X dP \geq \int_{\{X \geq a\}} X dP \geq a \times P(X \geq a)$$

Exemple d'application.

Soit A_n une suite d'éléments de $\mathcal{T}$, on appelle limite supérieure de ces A_n, et on note $\limsup A_n$, l'ensemble de ω dans Ω qui appartiennet à une infinité de A_n.

Considérons la variable aléatoire X définie par

$$X = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}$$

Elle est positive et vérifie $\limsup A_n = \{X = +\infty\}$; l'inégalité de Chebyshev implique

$$\infty \times P(\limsup A_n) \leq \int_\Omega X dP = \int_\Omega \sum_{n=1}^{\infty} \mathbf{1}_{A_n} dP = \sum_{n=1}^{\infty} \int_\Omega \mathbf{1}_{A_n} dP = \sum_{n=1}^{\infty} P(A_n)$$

(justifier chaque égalité). Il en résulte cette implication

$$\sum_{n=1}^{\infty} P(A_n) < \infty \Longrightarrow P(\limsup A_n) = 0$$

Remarques.

- Etant donné une variable aléatoire X positive ou intégrable, son intégrale $\int_\Omega X dP$ est aussi appelée **espérance** de X ou **valeur moyenne** de X et elle se note $\mathrm{E}(X)$. On a donc

$$\mathrm{E}(X) = \int_\Omega X dP = \int_\Omega X(\omega) dP(\omega)$$

- Une application $X : \Omega \longrightarrow \mathbf{C}$ est dite une variable aléatoire si les deux applications

$$\mathsf{Re}X : \Omega \longrightarrow \mathbf{R} \quad , \quad \mathsf{Im}X : \Omega \longrightarrow \mathbf{R}$$

sont des variables aléatoires. De plus, on pose

$$\mathrm{E}(X) = \int_\Omega X dP = \int_\Omega \mathsf{Re}X dP + i \int_\Omega \mathsf{Im}X dP = \mathrm{E}(\mathsf{Re}X) + i\mathrm{E}(\mathsf{Im}X)$$

lorsque les deux variables aléatoires $\mathsf{Re}X$ et $\mathsf{Im}X$ sont intégrables.

- La **variance** de la variable aléatoire X est l'élément de $[0, +\infty]$ défini par

$$\mathrm{V}(X) = \mathrm{E}\left(|X - \mathrm{E}(X)|^2\right) = \mathrm{E}(X^2) - \mathrm{E}^2(X)$$

Noter alors que $\mathrm{V}(X) = 0$ si et seulement si $X(\omega) = \mathrm{E}(X)$ presque sûrement.

Loi de probabilité d'une variable aléatoire

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé et $X : \Omega \longrightarrow [-\infty, \infty]$ une variable aléatoire.

On rappelle que l'image directe, par X, de la tribu $\mathcal{T}$ est la tribu $X_*\mathcal{T}$ de $[-\infty, \infty]$ définie par

$$X_*\mathcal{T} = \{A \subset [-\infty, \infty] \ ; \ X^{-1}(A) \in \mathcal{T}\}$$

Comme X est une variable aléatoire, l'image réciproque par X de tout intervalle de $\mathbf{R}$ est un élément de $X_*\mathcal{T}$ et par conséquent $X_*\mathcal{T}$ contient la tribu des boréliens $\mathcal{B}(\mathbf{R})$ de $\mathbf{R}$.

On a aussi défini l'image directe, par X, de la probabilité P comme étant la probabilité X_*P sur $X_*\mathcal{T}$ vérifiant

$$X_*P(A) = P(\ X^{-1}(A)\ ) = P(X \in A) \ \ , \ \ A \in X_*\mathcal{T}$$

Cette probabilité se note aussi P_X et elle est appelée **loi de probabilité** de X.

Le théorème suivant est une conséquence directe de la définition d'une intégrale.

Théorème de transfert.

Si $f : [-\infty, +\infty] \longrightarrow \mathbf{C}$, alors

$$\int_\Omega f(\ X(\omega)\ )dP(\omega) = \int_{[-\infty,+\infty]} f(x)dP_X(x)$$

chaque fois que le second membre a un sens.

Preuve. C'est vrai lorsque $f = \mathbf{1}_A$ (avec $A \in X_*\mathcal{T}$) par définition de P_X et c'est vrai dans tous les cas par définition de l'intégrale.

Exemple. **Loi géométrique**.

Considérons l'ensemble $\{0, 1\}$ que l'on munit de la probabilité P_1 définie par

$$P_1(\{1\}) = 1 \ \ , \ \ P_1(\{0\}) = 1 - p \ \ , \ \ 0 < p < 1$$

Considérons maintenant l'ensemble $\Omega = \{0, 1\}^{\mathbf{N}^*}$ dont les éléments ω s'écrivent

$$\omega = (\omega(1), \omega(2), \ldots, \omega(n), \ldots) \ \ , \ \ \omega(i) \in \{0, 1\}$$

et chaque $\omega \in \Omega$ est donc une suite $(\omega(n))_{n\geq 1}$ de 0 et 1.

On munit l'ensemble produit Ω de la probabilité P produit (tensoriel) des $(P_n)_{n\geq 1}$ où chaque P_n vaut P_1. Cela veut dire que si

$$C = A_1 \times \cdots \times A_m \times \{0, 1\} \times \{0, 1\} \times \cdots \ \ , \ \ A_i \subset \{0, 1\}$$

est un cylindre de Ω, alors

$$P(C) = P_1(A_1) \times \cdots \times P_m(A_m) = P_1(A_1) \times \cdots \times P_1(A_m)$$

Soit $X : \Omega \longrightarrow \{1, 2, \ldots, n, \ldots\} \cup \{+\infty\}$ la variable aléatoire définie par

$$X(\omega) = \begin{cases} +\infty & \text{si il n'existe aucun } n \text{ tel que } \omega(n) = 1 \\ \mathsf{inf}\{n \ ; \ \omega(n) = 1\} & \text{si il existe des } n \text{ tels que } \omega(n) = 1 \end{cases}$$

L'événement $\{X = 3\}$ est le cylindre

$$\{X = 3\} = \{0\} \times \{0\} \times \{1\} \times \{0, 1\} \times \{0, 1\} \times \cdots$$

et sa probabilité est donnée par

$$P(X = 3) = (1 - p) \times (1 - p) \times p = (1 - p)^2 \times p$$

De la même manière, on a

$$P(X = n) = (1 - p)^{n-1} \times p \quad , \quad n \geq 1$$

Il en résulte que

$$P(X < +\infty) = \sum_{1 \leq n < \infty} (1 - p)^{n-1} \times p = \frac{p}{1 - (1 - p)} = 1$$

si bien que $P(X = +\infty) = 1 - P(X < +\infty) = 1 - 1 = 0$.

- L'espérance de X est par définition

$$\begin{array}{rcl} \mathrm{E}(X) & = & \displaystyle\int_\Omega X dP = \sum_{n=1}^{\infty} \int_{\{X=n\}} X dP + \infty \times P(X = +\infty) \\ & = & \sum_{n=1}^{\infty} nP(X = n) + \infty \times 0 = \sum_{n=1}^{\infty} nP(X = n) \\ & = & \sum_{n=1}^{\infty} n(1 - p)^n p = p^{-1} \end{array}$$

On dit qu'une variable aléatoire suit une **loi de probabilité géométrique** de paramètre p si elle vérifie

$$P(X = n) = (1 - p)^{n-1} \times p \quad , \quad n \geq 1$$

L'espérance d'une telle variable vaut donc p^{-1}.

Interprétation.

On suppose que l'on dispose d'un jeton dont les faces sont marquées 0 et 1. On suppose aussi, qu'en le lançant, il retombe et la face marquée 1 apparaît avec la probabilité p et que l'autre face apparaît donc avec la probabilité $1 - p$.

On lance plusieurs fois ce jeton et on désigne par X la variable aléatoire qui compte le nombre de lancers nécessaires pour obtenir la face 1. Si par exemple le lanceur ω obtient la suite $(0, 0, 1, 0, 1, \ldots)$ alors $X(\omega) = 3$.

Si les lancers sont "indépendants" (c'est ce qu'on a supposé plus haut en introduisant la probabilité P produit tensoriel), alors X suit une loi géométrique de paramètre p et c'est pour cette raison qu'une loi géométrique s'appelle aussi loi du premier succés.

Fonction génératrice

Lorsqu'une variable aléatoire X prend ses valeurs dans $\mathbf{N} = \{0, 1, \ldots, n, \ldots\}$, on introduit la fonction génératrice

$$g(z) = g_X(z) = \sum_{n=0}^{\infty} P(X = n) z^n$$

C'est une série entière qui vérifie $g(1) = \sum_{n=0}^{\infty} P(X = n) = 1$ et dont la rayon de convergence est donc au moins égal à 1.

Elle permet de calculer l'espérance de X par dérivation, et on a

$$\mathrm{E}(X) = \sum_{n=0}^{\infty} nP(X = n) = g'_X(1) \in [0, +\infty]$$

et elle permet aussi de calculer la variance de X parce que

$$g''_X(1) = \sum_{n=0}^{\infty} n(n-1)P(X = n) = \mathrm{E}(X(X-1)) = \mathrm{E}(X^2) - \mathrm{E}(X)$$

d'où l'on tire $\mathrm{E}(X^2) = g''_X(1) + g'_X(1)$ et ensuite

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X) = g''_X(1) + g'_X(1) - \big(g'_X(1)\big)^2$$

Exemple. On dit qu'une variable aléatoire X suit **la loi de Poisson** de paramètre $a > 0$ si elle vérifie

$$P(X = n) = \frac{a^n}{n!} \mathsf{exp}(-a) \quad , \quad n \geq 0$$

Comme on a

$$\sum_{n=0}^{\infty} P(X = n) = \sum_{n=0}^{\infty} \frac{a^n}{n!} \mathsf{exp}(-a) = 1$$

il en résulte que X prend (preque sûrement) ses valeurs dans $\mathbf{N}$. La fonction génératrice d'une telle variable est donnée par

$$g_X(z) = \mathsf{exp}(-a)\mathsf{exp}(az)$$

et il en résulte que son espérance et sa variance son données par

$$\mathrm{E}(X) = a \quad , \quad \mathrm{V}(X) = a$$

Fonction caractéristique

Soit $X : \Omega \longrightarrow \mathbf{R}$ une variable aléatoire, sa **fonction caractéristique** est l'application

$$\varphi_X : \mathbf{R} \longrightarrow \mathbf{C} \quad , \quad \varphi_X(t) = \mathrm{E}(\mathsf{exp}(itX)) = \int_{\Omega} \mathsf{exp}(itX(\omega)) dP(\omega) = \int_{\mathbf{R}} e^{itx} dP_X(x)$$

Lorsque la variable aléatoire X est intégrable, on a

$$\varphi'_X(t) = \int_{\mathbf{R}} ixe^{itx} dP_X(x)$$

si bien que l'on a cette formule qui permet de calculer l'espérance de X

$$\varphi'_X(0) = \int_{\mathbf{R}} ixdP_X(x) = i\int_{\mathbf{R}} xdP_X(x) = i\mathrm{E}(X)$$

Lorsque la variable aléatoire est de carré intégrable, on a

$$\varphi''_X(t) = \int_{\mathbf{R}} -x^2 e^{itx} dP_X(x) \quad , \quad \mathrm{E}(X^2) = -\varphi''_X(0)$$

et il en résulte que

$$\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}^2(X) = (\varphi'_X(0))^2 - \varphi''_X(0)$$

Exemple. On dit qu'une variable X aléatoire suit la **loi normale réduite** si sa loi de probabilité P_X admet pour densité, par rapport à la mesure de Lebesgue, la fonction

$$G(x) = \frac{1}{\sqrt{2\pi}}\mathsf{exp}\left(-\frac{x^2}{2}\right)$$

Rappelons que cela veut dire que

$$dP_X(x) = G(x)dx = \frac{1}{\sqrt{2\pi}}\mathsf{exp}\left(-\frac{x^2}{2}\right)dx$$

La fonction caractéristique d'une telle variable aléatoire est donnée par

$$\varphi_X(t) = \mathsf{exp}\left(-\frac{t^2}{2}\right)$$

et il s'ensuit que

$$\mathrm{E}(X) = 0 \quad , \quad \mathrm{V}(X) = 1$$

Remarque.

- On dit d'une variable aléatoire qu'elle est **réduite** lorsque son espérance est nul et sa variance vaut 1.

- Lorsqu'une variable aléatoire est de carré intégrable et qu'elle n'est pas presque sûrement contante, sa réduite est la variable aléatoire X^* définie par

$$X^* = \frac{X - m}{\sigma}$$

où on a posé $m = \mathrm{E}(X)$ et $\sigma = \sqrt{\mathrm{V}(X)}$. Le réel m est donc la moyenne de X et le réel $\sigma > 0$ est l'**écart-type** de X.

Remarque.

La fonction caractéristique φ_X caractérise parfaitement la loi de probabilité P_X dans ce sens que

$$\varphi_X = \varphi_Y \Longrightarrow P_X = P_Y$$

Fonction de répartition

Etant donné une variable aléatoire $X : \Omega \longrightarrow \mathbf{R}$, sa fonction de répartition est l'application

$$F = F_X : \mathbf{R} \longrightarrow [0,1] \quad , \quad F_X(x) = P(X < x)$$

C'est une fonction croissante vérifiant

$$F_X(-\infty) = 0 \quad , \quad F_X(+\infty) = 1$$

Elle et continue à gauche ($\lim_{x \to a-} F_X(x) = F(a)$ pour tout a) et elle est discontinue aux points a vérifiant

$$P(X = a) = F_X(a+) - F_X(a) > 0$$

Exercice corrigé.

Soit X une variable aléatoire qui suit la loi normale réduite, m un réel quelconque et σ un nombre strictement positif.

1) Déterminer la loi de la variable aléatoire $Y = \sigma X + m$.

2) Déterminer la loi de la variable aléatoire $Z = |X|$.

Solution.

On peut procéder de la manière suivante

$$\begin{aligned} P(Y < x) &= P(\sigma X + m < x) \\ &= P(X < \sigma^{-1}(x-m)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sigma^{-1}(x-m)} \mathsf{exp}\left(-\frac{t^2}{2}\right) dt \end{aligned}$$

et alors la densité $dP_Y(x)$ de la loi de probabilité P_Y de Y est donnée par

$$dP_Y(x) = \frac{d}{dx} P(Y < x) = \frac{1}{\sigma\sqrt{2\pi}} \mathsf{exp}\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

De la même manière, on obtient

$$\begin{aligned} dP_Z(x) &= \frac{d}{dx} P(Z < x) = \frac{d}{dx} P(-x < X < x) = \frac{d}{dx} \frac{1}{\sqrt{2\pi}} \int_{-x}^{x} \mathsf{exp}\left(-\frac{t^2}{2}\right) dt \\ &= \sqrt{\frac{2}{\pi}} \mathsf{exp}\left(-\frac{x^2}{2}\right) \end{aligned}$$

Théorèmes de Fubini-Tonelli et de Fubini

Soit $(\Omega_1, \mathcal{T}_1, P_1)$ et $(\Omega_2, , \mathcal{T}_2, P_2)$ deux espaces probabilisés et considérons l'espace probabilisé produit

$$(\Omega_1 \times \Omega_2, \mathcal{T}_1 \otimes \mathcal{T}_1, P_1 \otimes P_2)$$

Les deux théorèmes suivants vont permettre d'intégrer les fonctions $f : \Omega_1 \times \Omega_2 \longrightarrow \mathbf{C}$ et ils disent que

Théorème de Fubini-Tonelli.

Si la fonction mesurable $f : \Omega_1 \times \Omega_2 \longrightarrow [0, \infty]$ est positive alors

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) dP_1 \otimes P_2(\omega_1, \omega_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2) dP_2(\omega_2) \right) dP_1(\omega_1)$$

Théorème de Fubini.

Si la fonction mesurable $f : \Omega_1 \times \Omega_2 \longrightarrow \mathbf{C}$ est intégrable alors

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) dP_1 \otimes P_2(\omega_1, \omega_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f(\omega_1, \omega_2) dP_2(\omega_2) \right) dP_1(\omega_1)$$

Variables aléatoires indépendantes

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé et $X : \Omega \longrightarrow \mathbf{R}^n$ une application.

- On dit que c'est une variable aléatoire si l'image directe $X_* \mathcal{T}$, par X, de la tribu $\mathcal{T}$ est incluse dans la tribu $\mathcal{B}(\mathbf{R}^n)$ des boréliens de $\mathbf{R}^n$.

- L'image directe $X_* P$, par X, de la probabilité P est une probabilité sur $X_* \mathcal{T}$ et donc sur $\mathcal{B}(\mathbf{R}^n)$ lorsque X est une variable aléatoire. On notera dorénavant P_X cette probabilité.

Etant donné deux variables aléatoires $X : \Omega \longrightarrow \mathbf{R}^n$ et $Y : \Omega \longrightarrow \mathbf{R}^m$, on a de façon naturelle une variable aléatoire

$$(X, Y) : \Omega \longrightarrow \mathbf{R}^n \times \mathbf{R}^m \quad , \quad \omega \longrightarrow (X, Y)(\omega) = (X(\omega), Y(\omega))$$

et on dit que ces deux variables sont **indépendantes** si

$$P_{(X,Y)} = P_X \otimes P_Y$$

Cela équivaut à dire que si $A \in \mathcal{B}(\mathbf{R}^n)$ et $B \in \mathcal{B}(\mathbf{R}^m)$ alors

$$P(X \in A, Y \in B) = P(X \in A) P(Y \in B)$$

Preuve. Cela résulte immédiatement des définitions puisque

$$\begin{aligned} P(X \in A, Y \in B) &= P((X, Y) \in A \times B) = P_{(X,Y)}(A \times B) = P_X \otimes P_Y(A \times B) \\ &= P_X(A) P_Y(B) = P(X \in A) P(Y \in B) \end{aligned}$$

De la même manière, on dit que des variables aléatoires $X_1, \ldots, X_m$ sont indépendantes si

$$P_{(X_1, \ldots, X_m)} = P_{X_1} \otimes \cdots \otimes P_{X_m}$$

et cela équivaut à dire que si $A_1, \ldots, A_m \in \mathcal{B}(\mathbf{R}^{n_i})$, alors

$$P(X_1 \in A_1, \ldots, X_m \in A_m) = P(X_1 \in A_1) \times \cdots \times P(X_m \in A_m)$$

On dit que la suite de variables aléatoires $(X_n)_n$ est indépendante si les v.a $X_1, \ldots, X_m$ sont indépendantes pour tout m.

Les deux propositions suivantes sont d'un usage fréquent.

Proposition.

Si les variables aléatoires $X : \Omega \longrightarrow \mathbf{R}^n$ et $Y : \Omega \longrightarrow \mathbf{R}^m$ sont indépendantes et si les fonctions $h : \mathbf{R}^n \longrightarrow \mathbf{R}^{n'}$ et $k : \mathbf{R}^m \longrightarrow \mathbf{R}^{m'}$ sont boréliennes (i.e. l'image réciproque, par h et k,d'un borélien est un borélien), alors les variables aléatoires $h \circ X : \Omega \longrightarrow \mathbf{R}^{n'}$ et $k \circ Y : \Omega \longrightarrow \mathbf{R}^{m'}$ sont aussi indépendantes.

Preuve. Si $A \in \mathcal{B}(\mathbf{R}^{n'})$ et $B \in \mathcal{B}(\mathbf{R}^{m\prime})$, alors

$$\begin{aligned} P(h \circ X \in A, k \circ Y \in B) &= P(X \in h^{-1}(A), Y \in k^{-1}(B)) \\ &= P(X \in h^{-1}(A))P(Y \in k^{-1}(B)) = P(h \circ X \in A)P(k \circ Y \in B) \end{aligned}$$

Proposition.

Si les variables aléatoires $X_1, \ldots, X_{n+m} : \Omega \longrightarrow \mathbf{R}$ sont indépendantes et si $h : \mathbf{R}^n \longrightarrow \mathbf{R}$ et $k : \mathbf{R}^m \longrightarrow \mathbf{R}$ sont boréliennes, alors

1) Les variables aléatoires $(X_1, \ldots, X_n) : \Omega \longrightarrow \mathbf{R}^n$ et $(X_{n+1}, \ldots, X_m) : \Omega \longrightarrow \mathbf{R}^m$ sont indépendantes.

2) Les variables aléatoires $h(X_1, \ldots, X_n) : \Omega \longrightarrow \mathbf{R}$ et $k(X_{n+1}, \ldots, X_m) : \Omega \longrightarrow \mathbf{R}$ sont indépendantes.

Les variables aléatoires simplifient énormément les calculs, comme on va le voir dans les théorèmes suivants.

Théorème.

1) Si les v.a. $X_1, \ldots, X_n$ sont indépendantes et intégrables, il en est de même de leur produit et

$$E\left(\prod_{i=1}^n X_i\right) = \prod_{i=1}^n E(X_i)$$

2) Si les v.a. $X_1, \ldots, X_n$ sont indépendantes et de carré intégrables, alors

$$V\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n V(X_i)$$

Preuve. On se contentera du cas de seulement deux variables aléatoires indépendantes X et Y. Considérons la fonction $m : \mathbf{R} \times \mathbf{R} \longrightarrow \mathbf{R}$ définie par $m(x, y) = xy$ de sorte que

$$XY = m \circ (X, Y)$$

Etudions d'abord le cas où les v.a. sont positives. La formule du transfert implique

$$\begin{aligned} E(XY) = \int_\Omega X(\omega)Y(\omega)dP(\omega) &= \int_{\mathbf{R}\times\mathbf{R}} m(x, y)dP_{(X,Y)}(x, y) \\ &= \int_{\mathbf{R}\times\mathbf{R}} xydP_X(x)dP_Y(y) \end{aligned}$$

et le théorème de Fubini-Tonelli implique

$$E(XY) = \int_{\mathbf{R}} x dP_X(x) \int_{\mathbf{R}} y dP_Y(y) = E(X)E(Y)$$

Le cas général s'obtient ainsi : Les v.a. $|X|$ et $|Y|$ sont positives et indépendantes et donc

$$E(|XY|) = E(|X|)E(|Y|) < \infty$$

Cela prouve que la v.a. XY est intégrable et il suffit de reprendre la démonstration ci-dessus en invoquant cette fois le théorème de Fubini à la place du théorème de Fubini-Tonelli.

La deuxième assertion du théorème résulte essentiellement du fait que

$$E([X - E(X)][Y - E(Y)]) = E(X - E(X))E(Y - E(Y)) = 0 \times 0 = 0$$

Théorème. Si les v.a. $X_1, \ldots, X_n : \Omega \longrightarrow \mathbf{R}$ sont indépendantes, alors la fonction caractéristique de leur somme est le produit de leurs fonctions caractéristiques

$$\varphi_{X_1+\cdots+X_n} = \prod_{i=1}^n \varphi_{X_i}$$

Preuve. Faisons la démonstration pour $n = 2$. Si X et Y sont indépendantes, alors il en est de même des v.a. $\mathsf{exp}(itX)$ et $\mathsf{exp}(itY)$ si bien que

$$\begin{aligned}\varphi_{X+Y}(t) &= \int_\Omega \mathsf{exp}(it(X(\omega) + Y(\omega))dP(\omega) = \int_\Omega \mathsf{exp}(it(X(\omega))\mathsf{exp}(itY(\omega))dP(\omega) \\ &= \int_\Omega \mathsf{exp}(it(X(\omega))dP(\omega) \int_\Omega \mathsf{exp}(it((\omega))dP(\omega) = \varphi_X(t)\varphi_Y(t)\end{aligned}$$

Produit de convolution de mesures finies

Notons $\sigma : \mathbf{R}^n \times \mathbf{R}^n \longrightarrow \mathbf{R}^n$ l'application "l'addition" définie par $\sigma(x, y) = x + y$.

Si maintenant μ et ν sont deux mesures finies sur les boréliens de $\mathbf{R}^n$, alors $\mu \otimes \nu$ est une mesure sur les boréliens de $\mathbf{R}^n \times \mathbf{R}^n$ et on pose

$$\mu * \nu = \sigma_*(\mu \otimes \nu)$$

Le produit de convolution de μ et ν est donc l'image directe de $\mu \otimes \nu$ par σ.

Exemples.

- Pour les masses de Dirac, la règle de calcul est

$$\delta_a * \delta_b = \delta_{a+b}$$

- Pour les mesures à densité, on a : Si $d\mu(x) = f(x)dx$ et $d\nu(x) = g(x)dx$, alors

$$d\mu * \nu(x) = h(x)dx \quad , \quad h(x) = \int_{\mathbf{R}^n} f(y)g(x - y)dy := f * g(x)$$

Proposition. Si $X, Y : \Omega \longrightarrow \mathbf{R}$ sont deux v.a. indépendantes alors la loi de probabilité de leur somme est égale au produit de convolution de leurs lois de probabilités :

$$P_{X+Y} = P_X * P_Y$$

Preuve. Cela résulte des égalités

$$P_{X+Y} = \sigma_* P_{(X,Y)} = \sigma_* P_X \otimes P_Y = P_X * P_Y$$

Le lemme de Borel-Cantelli

Soit $(\Omega, \mathcal{T}, P)$ est un espace probabilisé.

- Deux événements A et B (i.e. deux éléments de $\mathcal{T}$) sont **indépendants** si

$$P(A \cap B) = P(A)P(B)$$

Cela équivaut à dire que les v. a. $\mathbf{1}_A$ et $\mathbf{1}_B$ sont indépendantes.

- Par récurrence sur n, on dit que les événements $A_1, A_2, \ldots, A_n$ sont indépendants si

a) $n-1$ quelconques de ces événements sont indépendants.

b) $P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n)$.

Cela équivaut à dire que les v. a. $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_n}$ sont indépendantes.

Exemple.

Une urne contient quatre boules, trois de ces boules portent respectivement les numéros 1, 2, 3, et la quatrième porte ces trois numéros à la fois.

On tire une boule et on désigne par A_i l'événement " la boule tirée porte le numéro i ". On a

$$P(A_1) = P(A_2) = P(A_3) = \tfrac{1}{2}$$
$$P(A_1 \cap A_2) = P(A_2 \cap A_3) = P(A_3 \cap A_1) = \tfrac{1}{4}$$
$$P(A_1 \cap A_2 \cap A_3) = \tfrac{1}{4}$$

et ces trois événements sont donc deux à deux indépendants sans être indépendants.

- On dit qu'une suite A_n d'événements est une suite d'événements indépendants si, pour tout n, les événements $A_1, A_2, \ldots, A_n$ sont indépendants.

Cela équivaut à dire que la suite des v. a. $\mathbf{1}_{A_n}$ est une suite de v.a. indépendantes.

Soit A_n une suite d'événements, l'événement "une infinité d'événements A_k a lieu" s'appelle **limite supérieure** des A_n et s'écrit

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

Lemme de Borel-Cantelli

1) Si une suite d'événements A_n est telle que $\sum P(A_n) < \infty$, alors $P(\limsup A_n) = 0$.

2) Si une suite d'événements indépendants A_n est telle que $\sum P(A_n) = \infty$, alors $P(\limsup A_n) = 1$.

Preuve :

1) Soit $X : \Omega \to [0, \infty]$ **une variable aléatoire** (i.e. une fonction mesurable). Pour tout $\alpha \in [0, \infty]$, on a

$$\int_\Omega X dP = \int_{\{X < \alpha\}} X dP + \int_{\{X \geq \alpha\}} X dP \geq \alpha P(X \geq \alpha)$$

En appliquant **cette inégalité de Chebyshev** à la v. a. $X = \sum_{n \geq 1} 1_{A_n}$ et $\alpha = \infty$, on obtient

$$\sum_{n=1}^{\infty} P(A_n) = \int_\Omega X dP \geq \infty P(X \geq \infty) = \infty P(\limsup A_n)$$

et cela démontre la première assertion.

2) On a

$$P(\limsup A_n) = \lim_{n \to \infty} P(\bigcup_{k=n}^{\infty} A_k) = \lim_{n \to \infty} \left(1 - P(\bigcap_{k=n}^{\infty} A_k^c)\right)$$

et comme les événements A_n sont indépendants, on a on*et l'assertion 2 en résulte.

3.3 Le théorème central de la limite et les lois fortes des grands nombres

Le théorème central de la limite

Si μ est une mesure finie sur $(\mathbf{R}^n, \mathcal{B}(\mathbf{R}^n))$, sa transformée de Fourier $\hat{\mu} : \mathbf{R}^n \to \mathbf{C}$ est définie par

$$\hat{\mu}(y) = \int_{\mathbf{R}^n} \mathsf{exp}(i < x \mid y > d\mu(x)$$

La transformation de Fourier vérifie les deux propriétés suivantes :

$\triangleright$ Elle est injective : $\hat{\mu} = \hat{\nu} \Rightarrow \mu = \nu$.

$\triangleright$ Elle transforme le produit de convolution en produit ordinaire : $\widehat{\mu * \nu} = \hat{\mu}.\hat{\nu}$.

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé et X une variable aléatoire sur Ω, sa fonction caractéristique $\varphi_X : \mathbf{R} \to \mathbf{C}$ est définie par

$$\varphi_X(x) = \int_\Omega \mathsf{exp}(ixX(\omega)) dP(\omega) = \int_R \mathsf{exp}(ixy) dP_X(y)$$

La fonction caractéristique de X est donc la transformée de Fourier de sa loi de probabilité P_X.

Convergence de mesures.

Notons $C_c(\mathbf{R})$ l'espace des fonctions continues sur $\mathbf{R}$ qui sont à support compact et $C_b(\mathbf{R})$ celui des fonctions continues et bornées sur $\mathbf{R}$.

Soit μ_n , μ des mesures bornées sur $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$. On dit que la suite :

$\triangleright$ μ_n converge vaguement vers μ si $\mu_n(f)$ tend vers $\mu(f)$ pour tout $f \in C_c(\mathbf{R})$,

$\triangleright$ μ_n converge étroitement vers μ si $\mu_n(f)$ tend vers $\mu(f)$ pour tout $f \in C_b(\mathbf{R})$.

Lemme. *Si les mesures μ_n converge vaguement vers μ et si $\mu_n(\mathbf{R})$ tend vers $\mu(\mathbf{R})$ alors les mesures μ_n converge étroitement vers μ.*

Lemme. *Les assertions suivantes sont équivalentes* :

a) *μ_n converge étroitement vers μ.*

b) *$\widehat{\mu_n}$ converge simplement vers $\hat{\mu}$.*

Preuve de $b \Rightarrow a$. On aura besoin de l'espace de Schwartz $\mathcal{S}(\mathbf{R})$ qui est l'ensembles des fonctions $f : \mathbf{R} \longrightarrow \mathbf{C}$ indéfiniment dérivables et dont toutes les dérivées décroissent rapidement (i.e. $\lim_{x\to\pm\infty} x^m f^{(n)}(x) = 0$ pour tous les entiers naturels m et n). Cet espace est dense, pour la norme uniforme, dans $C_c(\mathbf{R})$ et il est invariant par la transformation de Fourier.

La suite $\mu_n(\mathbf{R}) = \widehat{\mu_n}(0)$ tend vers $\hat{\mu}(0) = \mu(\mathbf{R})$, elle est donc bornée.

Soit $g \in \mathcal{S}(\mathbf{R})$, il résulte du théorème de convergence dominée que :

$$\int \widehat{g}(t)d\mu_n(t) = \int g(x)\widehat{\mu_n}(x)dx \longrightarrow \int g(x)\widehat{\mu}(x)dx = \int \widehat{g}(t)d\mu(t)$$

Soit $h \in C_c(\mathbf{R})$ et $\varepsilon > 0$, il existe $g \in \mathcal{S}(\mathbf{R})$ telle que $\| h - g \|_\infty \leq \varepsilon$ et cela donne pour n assez grand :

$$\begin{aligned} |\int h d\mu_n - \int h d\mu| &\leq |\int (h-g)d\mu_n| + |\int (h-g)d\mu| + |\int g d\mu_n - \int g d\mu| \\ &\leq \varepsilon . \sup \mu_n(R) + \varepsilon\mu(R) + \varepsilon \end{aligned}$$

Lemme

Si μ_n *converge vaguement vers μ et si $\mu(\{a\}) = 0$ alors* $\lim \mu_n(\{a\}) = 0$.

Si μ_n *converge étroitement vers μ et si $\mu(\{a\}) = 0$ alors*

$$\lim \mu_n(]-\infty, a]) = \mu(]-\infty, a])$$

Le Théorème central de la limite.

Soit G la mesure (dite loi normale ou loi de Gauss) sur $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ définie par

$$dG(t) = \frac{1}{\sqrt{2\pi}} \mathsf{exp}(-\frac{1}{2}\, t^2)$$

C'est une probabilité vérifiant $\widehat{G}(x) = \mathsf{exp}(-\frac{1}{2}x^2)$.

Théorème central de la limite.

Soit λ une probabilité sur $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ et posons pour tout $A \in \mathcal{B}(\mathbf{R})$

$$\mu_n(A) = \lambda^{*n}(\sqrt{n}A) = \lambda * \cdots * \lambda(\sqrt{n}A) \ , \ \text{(n fois)}$$

Si $\int_{\mathbf{R}} td\lambda(t) = 0$ et $\int_{\mathbf{R}} t^2 d\lambda(t) = 1$ alors μ_n converge étroitement vers G.

Preuve : Posons $h(x) = \hat{\lambda}(x)$ si bien que

$$h(x) = 1 - \frac{1}{2}\,x^2 + o(x^2) \ , \ \widehat{\mu_n}(x) = h^n(x/\sqrt{n}) = [1 - x^2/n + o(n^{-3/2})]^n$$

On en déduit que $\lim_{n\to\infty} \widehat{\mu_n}(x) = \mathsf{exp}(-\frac{1}{2}\ x^2) = \widehat{G}(x)$.

Conséquence : **Théorème central de la limite**.

Soit X_n une suite de v.a. indépendantes, de même loi, d'espérance m et d'écart-type σ. Si $S_n = X_1 + \cdots X_n$, alors sa réduite

$$S_n^* = \frac{S_n - n.m}{\sqrt{n}.\sigma}$$

converge en loi vers la loi normale (i.e. les lois de probabilité de S_n^* convergent étroitement vers la loi normale de Gauss G).

Les lois fortes des grands nombres

On va énoncer les deux lois fortes de Kolmogorov qui vont résulter essentiellement d'une inégalité due à Kolmogorov laquelle généralise l'inégalité de Chebyshev.

Inégalité de Chebyshev.

Soit $(X, \mathcal{T}, \mu)$ en espace mesuré et $f : X \to [0, \infty]$ une fonction mesurable. Pour tout $a \in [0, \infty]$, on a

$$\int_X f d\mu = \int_{\{f \geq a\}} f d\mu + \int_{\{f < a\}} f d\mu \geq a.\mu(\{f \ \geq a\})$$

Soit $(\Omega, \mathcal{T}, P)$ un espace probabilisé et $X : \Omega \to R$ une variable aléatoire de carré intégrable. Si $a > 0$, l'inégalité ci-dessus implique cette inégalité de Chebyshev

$$P(\mid X - E(X) \mid \geq a) = P(\mid X - E(X) \mid^2 \geq a^2) \leq \ a^{-2}V(X)$$

Inégalité de Kolmogorov.

Soit S une v.a. de carré intégrable, l'inégalité de Chebyshev dit que :

$$P(\mid S - E(S) \mid \geq \varepsilon) \leq \varepsilon^{-2}V(S)$$

et l'inégalité de Kolmogorov améliore l'inégalité de Chebyshev lorsque S est une somme de v.a. indépendantes.

Inégalité de Kolmogorov.

Soit $X_1, \dots, X_n$ des v. a. indépendantes de carré intégrables et posons $S_m = X_1 + \cdots X_m$. On a

$$P(\ \{\ \max_{1\leq j\leq n} \mid S_j - E(S_j) \mid \geq \varepsilon\ \}\ ) \leq \varepsilon^{-2} V(S_n)$$

Preuve

En remplaçant X_i par $X_i - E(X_i)$, on se ramène aux cas où $E(X_i) = 0$. Posons

$$A = \{\ \max_{1\leq j\leq n} \mid S_j \mid \geq \varepsilon\ \} \text{ et } A_k = \{ \max_{1\leq j\leq k-1} \mid S_j \mid < \varepsilon\ ,\ \mid S_k \mid \geq \varepsilon\}$$

L'ensemble A est la réunion disjointe des A_k et alors

$$V(S_n) = E(S_n^2) \geq E(\mathbf{1}_A.S_n^2) = \sum_{k=1}^{n} E(\mathbf{1}_{A_k}.S_n^2)$$

Posons $Y_k = X_{k+1} + \cdots X_n$ si bien que $S_n = S_k + Y_k$ et donc

$$\mathbf{1}_{A_k}.S_n^2 = \mathbf{1}_{A_k}.S_k^2 + 2.\mathbf{1}_{A_k}.S_k.Y_k + \mathbf{1}_{A_k}.Y_k^2$$

Comme les v.a. Y_k et $1_{A_k}.S_k$ sont indépendantes et que $E(Y_k) = 0$, on a aussi $E(1_{A_k}.S_k.Y_k) = 0$ et alors

$$E(\mathbf{1}_{A_k}.S_n^2) = E(\mathbf{1}_{A_k}.S_k^2) + E(\mathbf{1}_{A_k}.Y_k^2) \geq E(\mathbf{1}_{A_k}.S_k^2) \geq \varepsilon^2 P(A_k)$$

On en déduit $V(S_n) \geq \sum_{k=1}^{n} E(1_{A_k}.S_n^2) \geq \sum_{k=1}^{n} \varepsilon^2 P(A_k) = \varepsilon^2 P(A)$.

Lois des grands nombres.

Soit X_n une suite de v.a. indépendantes et intégrables. On dit que

$\triangleright$ la suite X_n obéit à la loi forte des grands nombres si, presque sûrement, on a

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} [X_k - E(X_k)] = 0.$$

$\triangleright$ la suite X_n obéit à la loi faible des grands nombres si, pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} P(\ \mid \frac{1}{n} \sum_{k=1}^{n} [X_k - E(X_k)] \mid \geq \varepsilon\ ) = 0$$

Théorème de Markov

Si une suite de v.a. X_n vérifie $\lim_{n\to\infty} n^{-2} V(\sum_{k=1}^{n} X_n) = 0$ alors elle obéit à la loi faible des grands nombres.

Preuve :
C'est une conséquence directe de l'inégalité de Chebyshev.

Le lemme suivant sera utile pour la démonstration des deux lois fortes de Kolmogorov.

Lemme.
Soit X_n une suite de v.a. indépendantes. Si la série $\sum V(X_n)$ des variances des X_n converge alors la série $\sum[X_n - E(X_n)]$ converge presque sûrement.

Preuve. On se ramène au cas où $E(X_n) = 0$. Posons $S_m = X_1 + \cdots + X_m$, il suffit de montrer que pour tout $\varepsilon > 0$

$$\lim_{m\to\infty} P(\bigcup_{p=1}^{\infty}\{\mid S_{m+p} - S_m \mid \geq \varepsilon\}) = 0$$

Cela résulte de l'inégalité de Kolmogorov

$$\begin{aligned} P(\bigcup_{p=1}^{\infty}\{ \quad \mid S_{m+p} - S_m \mid \geq \varepsilon\}) &= \lim_n P(\bigcup_{p=1}^{n}\{\mid S_{m+p} - S_m \mid \geq \varepsilon\}) \\ &\leq \varepsilon^{-2} V(S_{m+n} - S_m) \leq \varepsilon^{-2} \sum_{p=m+1}^{\infty} V(X_p) \underset{m}{\to} 0 \end{aligned}$$

Premier théorème de Kolmogorov.

Soit X_n une suite de v.a. indépendantes. Si la série $\sum n^{-2}V(X_n)$ converge alors X_n obéit à la loi forte des grands nombres.

Preuve : Posons $Y_n = [X_n - E(X_n)]/n$. La série $\sum V(Y_n)$ converge et il résulte du lemme précédent que $\sum Y_n$ converge presque sûrement et cela implique que X_n obéit à la loi forte des grands nombres.

(Le dernier passage utilise le lemme de Kronecker qui dit que si la série $\sum z_n$ converge alors la suite $\frac{1}{n}\sum_{k=1}^{n} kz_k$ tend vers zéro : cela s'obtient en utilisant le procédé de sommation d'Abel.)

Deuxième théorème de Kolmogorov.

Soit X_n une suite de v.a. indépendantes ayant la même loi de probabilité. Si X_1 est intégrable alors X_n obéit à la loi forte des grands nombres.

Preuve : Posons $Y_n = X_n.1_{\{\mid X_n \mid < n\}}$, on a

$$\sum_{n=1}^{\infty} n^{-2}V(Y_n) \leq \sum_{n=1}^{\infty} n^{-2}E(Y_n^2) \leq \frac{\pi^2}{6}E(\mid X_1 \mid)$$

Il résulte du premier théorème de Kolmogorov que Y_n obéit à la loi forte des grands nombres et puisque $\lim E(Y_n) = E(X_1)$ on a

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \to E(X_1)$$

On a maintenant

$$\sum_{k=1}^{\infty} P(X_k \neq Y_k) = \sum_{k=1}^{\infty} P(\mid X_k \mid \geq k) = \sum_{k=1}^{\infty} P(\mid X_1 \mid \geq k) \leq E(\mid X_1 \mid) < \infty$$

et il résulte du lemme de Borel-Cantelli que $P(\limsup\{X_k \neq Y_k\}) = 0$. Cela implique que les deux suites $\frac{1}{n}\sum_{k=1}^{n} X_k$ et $\frac{1}{n}\sum_{k=1}^{n} Y_k$ sont presque sûrement de même nature.

Théorème de Glivenko-Cantelli.

Pour cet important théorème, on a besoin du lemme suivant qui généralise un lemme classique dû à Dini.

Lemme : *Soit F_n, F des fonctions croissantes, bornées et continues à droite et soit S une partie dense dans $\mathbf{R}$ contenant tous les points de discontinuité de la fonction F . Si*

$$\lim F_n(x) = F(x) \text{ et } \lim F_n(x^-) = F(x^-) \text{ pour tout } x \in S \cup \{\pm\infty\}$$

alors F_n converge uniformément vers F sur $\mathbf{R}$.

Théorème de Glivenko-Cantelli.
Si X_n est *une suite de v.a. indépendantes et de même loi de probabilité et si*

$$F(x) = P(X_1 \leq x) \text{ et } F_n(x,\omega) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{\{\, X_k \leq x \,\}}(\omega)$$

alors, presque sûrement, $F_n(\ ,\omega)$ converge uniformément vers F :

$$P(\ \{\, \omega\ ;\ \sup_{-\infty < x < \infty} \mid F_n(x,\omega) - F(x) \mid \to 0 \,\}\ ) = 1$$

Preuve. Posons, pour $x \in R$,

$$Y_k(x,\omega) = \mathbf{1}_{\{\, X_k \leq x \,\}}(\omega) \text{ et } Z_k(x,\omega) = \mathbf{1}_{\{\, X_k < x \,\}}(\omega)$$

On a $E(Y_k(x,.)) = F(x)$ et $E(Z_k(x,.)) = F(x^-)$ et il résulte du deuxième théorème de Kolmogorov, que les v.a. $Y_k(x,.)$ et $Z_k(x,.)$ obéissent à la loi forte des grands nombres et il existe alors un ensemble A_x de probabilité nulle tel que

$$\omega \notin A_x \ \Rightarrow\ \lim F_n(x,\omega) = F(x) \text{ et } \lim F_n(x^-,\omega) = F(x^-)$$

Soit S une partie dénombrable, dense dans $\mathbf{R}$ et contenant tous les points de discontinuité de F et posons $A = \cup_{x\in S} A_x$. Il résulte du lemme précédent que si $\omega \notin A$ alors $F_n(\ ,\omega)$ converge uniformément vers F sur $\mathbf{R}$.

Bibliographie.

Pour des livres de cours, on pourra consulter avec profit :
Gerald B. FOLLAND. Real analysis. Wiley-Interscience, 1984.
Robert B. ASH. Real analysis and probabilty. Academic Press, 1972.
Pour des livres d'exercices, on conseille :
G. LETAC. Intégration et probabilités, exercices. Masson, 1982.
M. COTTRELL, ... Exercices de probabilités. Belin, DIA, 1980.

3.4 Espérances conditionnelles

Soit (Ω, Σ, P) un espace probabilisé et $Y : \Omega \to \mathbf{R}$ une variable aléatoire intégrable.

Soit $\mathcal{A}$ une sous-tribu de Σ, l'application $\gamma : \mathcal{A} \to \mathbf{R}$ définie par

$$\gamma(A) = \int_A Y(\omega)\, dP(\omega)$$

est une mesure (non nécessairement positive) qui est absolument continue par rapport à la probabilité P (restreinte à $\mathcal{A}$) :

$$P(A) = 0 \Rightarrow \gamma(A) = 0$$

Il résulte alors d'un théorème dit de Radon-Nikodym qu'il existe une fonction $Z : \Omega \to \mathbf{R}$ vérifiant

a) Z est $\mathcal{A}$-mesurable.

b) Pour tout $A \in \mathcal{A}$, on a

$$\int_A Y(\omega)\, dP(\omega) = \gamma(A) = \int_A Z(\omega)\, dP(\omega)$$

La v. a. Z est appelée **espérance conditionnelle** de Y sachant $\mathcal{A}$ et est notée $E(\, Y \mid \mathcal{A}\,)$ ou $E^{\mathcal{A}}(Y)$.

Remarques.

1) L'espérance conditionnelle $E(\, Y \mid \mathcal{A}\,)$ est une variable aléatoire contrairement à l'espérance mathématique ordinaire $E(Y)$ qui est une constante (c'est donc aussi une fonction constante sur Ω).

2) Il résulte de la définition même de $E(\, Y \mid \mathcal{A}\,)$ que l'on a :

$$E\left(E(\, Y \mid \mathcal{A})\right) = E(Y)$$

Remarque.

Si la variable aléatoire $Y : \Omega \to \mathbf{R}$ est de carré intégrable, on sait que la fonction $h : \mathbf{R} \to \mathbf{R}$ définie par

$$h(a) = E\left(|Y - a|^2\right)$$

atteint son minimum pour $a = E(X)$.

Une propriété analogue a ction orthogonale de Y sur $L^2(\Omega, \mathcal{A}, P)$:

$$a \in L^2(\Omega, \mathcal{A}, P) \Rightarrow E\left(a \times (Y - E(\, Y \mid \mathcal{A}))\right) = 0$$

Cela provient de l'identité $aE(\, Y \mid \mathcal{A}) = E(\, aY \mid \mathcal{A})$ qui implique

$$E\left(a \times (Y - E(\, Y \mid \mathcal{A}))\right) = E(aY) - E(\, aE(\, Y \mid \mathcal{A})\,) = E(aY) - E(\, E(\, aY \mid \mathcal{A})\,) = 0$$

On vient de montrer que l'opérateur $E^{\mathcal{A}}$, quand on le restreint à $L^2(\Omega, \Sigma, P)$, représente la projection orthogonale sur le sous-espace fermé $L^2(\Omega, \mathcal{A}, P)$.

Exemples.

1) Si $\mathcal{A} = \{\varnothing\ ,\ \Omega\}$ alors toute fonction $\mathcal{A}$-mesurable est constante et

$$E(\,Y \mid \mathcal{A}) = E(Y)$$

Cela montre que la notion d'espérance conditionnelle généralise la notion d'espérance ordinaire.

En général, l'espérance conditionnelle $E(\,Y \mid \mathcal{A})$ est un meilleur "résumé", pour ceux qui s'intéressent aux éléments de $\mathcal{A}$, que l'espérance ordinaire $E(Y)$.

2) Si $\mathcal{A} = \{\varnothing, A, A^c, \Omega\}$ et si $P(A)$ et $P(A^c)$ sont non nuls, l'espérance conditionnelle $E(\,Y \mid \mathcal{A})$ est de la forme

$$E(\,Y \mid \mathcal{A}) = \alpha \mathbf{1}_A + \beta \mathbf{1}_{A^c}\ \ ,\ \ \alpha, \beta \in \mathbf{R}$$

et les constantes α et β, qui s'obtiennent en intégrant sur A et A^c, sont données par

$$\alpha = \frac{1}{P(A)} \int_A Y(\omega) dP(\omega)\ \ ,\ \ \beta = \frac{1}{P(A^c)} \int_{A^c} Y(\omega) dP(\omega)$$

Elles représentent donc les moyennes de la v. a. Y sur les éléments "non divisibles" A et A^c de $\mathcal{A}$.

3) Si la tribu $\mathcal{A}$ est engendrée par la partition A_n et si les $P(A_n)$ sont non nuls, on posera

$$E(\,Y \mid A_n) = \frac{1}{P(A_n)} \int_{A_n} Y(\omega) dP(\omega)$$

de sorte que $E(\,Y \mid A_n)$ représente la valeur moyenne de Y sur A_n et alors

$$E(\,Y \mid \mathcal{A}) = \sum_n E(\,Y \mid A_n) \times \mathbf{1}_{A_n}$$

et l'identité $E\,(E(\,Y \mid \mathcal{A})) = E(Y)$ donne

$$E(Y) = \sum_n E(\,Y \mid A_n) \times P(A_n)$$

Cela veut dire, tout simplement, que la moyenne $E(Y)$ est égale à la moyenne des moyennes $E(\,Y \mid A_n)$.

Exemple et notation.

Soit $X : \Omega \to \mathbf{R}$ une variable aléatoire et $\sigma(X)$ la tribu engendrée par X :

$$\sigma(X) = \{\ \sigma^{-1}(B)\ ,\ B \text{ est un borélien de } \mathbf{R}\ \}$$

La tribu $\sigma(X)$ est une sous-tribu de Σ et c'est la plus petite tribu qui rend X mesurable.

Pour simplifier l'écriture, on écrira $E(\,Y \mid X\,)$ au lieu de $E(\,Y \mid \sigma(X)\,)$.

Comme $E(\,Y \mid X\,)$ est une fonction $\sigma(X)$-mesurable, il existe une fonction $\varphi : \mathbf{R} \to \mathbf{R}$ borélienne telle que

$$E(\,Y \mid X\,) = \varphi \circ X$$

Par analogie au cas où la tribu $\mathcal{A}$ est engendrée par une partition dénombrable, nous poserons

$$E(\,Y \mid X = x\,) = \varphi(x) \ , \ x \in \mathbf{R}$$

et nous dirons que c'est la moyenne de Y sur l'ensemble $(X = x)$. Noter que cela permet de donner un sens à cette phrase qui ne veut rien dire lorsque $P(X = x) = 0$.

Exemple. On munit l'intervalle $[-1, 1]$ de la probabilité uniforme et on considère la v. a. X définie par

$$X(x) = x^2$$

Il est facile de vérifier que la tribu $\sigma(X)$ est formée des boréliens de $[-1, 1]$ qui sont symétriques par rapport à l'origine :

$$\sigma(X) = \{\ B \in \mathcal{B}([-1, 1])\ ;\ B = -B\ \}$$

et que si $Y : [-1, 1] \to \mathbf{R}$ est une v. a., alors son espérance conditionnelle sachant X est égale à la partie paire de Y :

$$E(Y \mid X)(x) = \frac{Y(x) + Y(-x)}{2} \ , \ x \in [-1, 1]$$

et on déduit que

$$E(Y \mid X = x) = \frac{Y(\sqrt{x}) + Y(-\sqrt{x})}{2} \ , \ x \in [0, 1]$$

Remarques et définitions.

1) Il résulte de la définition de l'espérance conditionnelle que l'on a

$$E(Y) = E(E(\,Y \mid X\,)) = \int_{\mathbf{R}} \varphi(x) dP_X(x) = \int_{\mathbf{R}} E(\,Y \mid X = x\,) dP_X(x)$$

et cela exprime encore une fois que la moyenne de Y est la moyenne des moyennes $E(\,Y \mid X = x\,)$.

2) Lorsque Y est la fonction caractéristique d'un ensemble $C \in \Sigma$, on posera

$$P(C \mid X) = E(\mathbf{1}_C \mid X)$$

et on dira que $P(C \mid X)$ est **la probabilité conditionnelle de C sachant X**. La formule ci-dessus devient dans ce cas

$$P(C \mid X) = \int_{\mathbf{R}} P(C \mid X = x) dP_X(x)$$

et cela généralise cette formule très utile :

$$P(C) = \sum_n P(C \mid A_n)P(A_n)$$

qui est valable chaque fois que les A_n forment une partition de Ω en événements de probabilités non nulles.

3) Soit $h : \mathbf{R} \to \mathbf{R}$ une fonction borélienne telle que la v. a. $h \circ Y$ soit intégrable, on sait qu'il existe une fonction borélienne $\varphi : \mathbf{R} \to \mathbf{R}$ telle que

$$E(\, h \circ Y \mid \, X\,) = \varphi \circ X$$

et que si A est $\sigma(X)$-mesurable

$$E(\mathbf{1}_A \times \varphi \circ X) = E(\mathbf{1}_A \times E(\, h \circ Y \mid \, X\,)) = E(\mathbf{1}_A \times h \circ Y)$$

L'ensemble A est de la forme $X^{-1}(B)$ où B est un borélien de $\mathbf{R}$ si bien que $1_A = 1_B \circ X$, et on montre facilement que si $k : \mathbf{R} \to \mathbf{R}$ est une fonction borélienne et positive, alors

$$E(k \circ X \times \varphi \circ X) = E(k \circ X \times h \circ Y)$$

Cela donne cette formule

$$\int_{\mathbf{R}} k(x)\varphi(x)dP_X(x) = \int_{\mathbf{R}\times\mathbf{R}} k(x)h(y)dP_{(X,Y)}(x,y)$$

qui permet souvent de déterminer l'application φ et donc l'espérance conditionnelle $E(\, h \circ Y \mid \, X\,)$.

Exemple.

Soit $X, Y : \Omega \to \mathbf{R}$ deux v. a. indépendantes de même loi de probabilité

$$dP_X(x) = dP_Y(x) = \mathsf{exp}(-x)\mathbf{1}_{[0,\infty[}(x)dx$$

Déterminer l'espérance conditionnelle $E(X \mid \max\{\, X, Y\,\}\,)$.

Solution. Posons $Z = \max\{\, X, Y\,\}$. On sait que, si $h : \mathbf{R} \to \mathbf{R}$ est borélienne et si la v. a. $h \circ X$ est intégrable, il existe une fonction borélienne $\varphi : \mathbf{R} \to \mathbf{R}$ telle que $E(h \circ X \mid Z\,) = \varphi \circ Z$ et que, pour toute fonction $k : \mathbf{R} \to \mathbf{R}$ borélienne positive, on a

$$E(k \circ Z \times \varphi \circ Z) = E(k \circ Z \times h \circ X)$$

Déterminons la loi de probabilité de Z : On a

$$P(Z < z) = P(\max\{\, X, Y\,\} < z) = P(X < z, Y < z) = (1 - e^{-\,z})^2$$

si bien que la loi de probabilité de Z est donnée par

$$dP_Z(z) = \frac{d}{dz}P(Z < a)dz = 2(1 - e^{-\,z})e^{-z}dz$$

Il en résulte que l'on a

$$\int_0^\infty k(z)\varphi(z)2(1-e^{-z})e^{-z}dz = \int_0^\infty \int_0^\infty k(\max\{x,y\})h(x)e^{-x}e^{-y}dxdy$$

Le second membre (notons-le S) vaut

$$\begin{aligned} S &= \int_0^\infty k(x)h(x)\left(\int_0^x e^{-y}dy\right)e^{-x}dx + \int_0^\infty k(y)\left(\int_0^y h(x)e^{-x}dx\right)e^{-y}dy \\ &= \int_0^\infty k(x)\left(h(x)(1-e^{-x}) + \int_0^x h(t)e^{-t}dt\right)e^{-x}dx \end{aligned}$$

Il en résulte que l'on a

$$\varphi(x)2(1-e^{-x}) = h(x)(1-e^{-x}) + \int_0^x h(t)e^{-t}dt$$

de sorte que

$$\varphi(x) = \frac{h(x)}{2} + \frac{\int_0^x h(t)e^{-t}dt}{2(1-e^{-x})}$$

Cette relation permet d'écrire, en faisant $h(x) = x$,

$$E(X \mid \max\{X,Y\} = x) = \frac{x}{2} + \frac{1}{2} - \frac{x}{2(\exp(x)-1)}$$

et de dire que **la loi conditionnelle de X sachant que** $\max\{X,Y\} = x$ est égale à

$$\frac{1}{2}\,\delta_x + \frac{1}{2(1-e^{-x})}\mathbf{1}_{[0,x]}(t)e^{-t}dt$$

Exercice.

1) Soit X la v. a. qui vaut 1 si le premier essai donne un succès et qui vaut 0 sinon.

1a) Calculer $E(Y_1 \mid X = 1)$ et montrer que $E(Y_1 \mid X = 0) = 1 + E(Y_1)$.

1b) Calculer $E(Y_1)$.

2) Montrer que $E(Y_{m+1} \mid Y_m) = Y_m + 1 + (1-p)E(Y_{m+1})$.

3) En déduire que

$$E(Y_m) = \sum_{n=1}^{m} p^{-n}$$

Exercice.

Soit X_n une suite de v. a. indépendantes, équidistribuées, de moyenne m et N une v. a. entière et indépendante des X_n. On pose

$$S_n(\omega) = \sum_{i=1}^{n} X_i(\omega) \quad , \quad S_N(\omega) = \sum_{i=1}^{N(\omega)} X_i(\omega)$$

1) Montrer que $E(S_N \mid N = n) = E(S_n)$.

2) Montrer que

$$E(S_N) = E(N)E(X_1)$$

3) On suppose que X_1 et N sont de carré intégrables, prouver que

$$V(S_N) = E(N)V(X_1) + V(N)E^2(X_1)$$

Exercice.

Déterminer l'espérance conditionnelle $E(X \mid X + Y)$.

Solution : Posons $T = X + Y$, la fonction $E(X \mid X + Y)$ est de la forme $\varphi \circ T$ où $\varphi : \mathbf{R} \to \mathbf{R}$ est borélienne et vérifie

$$\text{si } A \text{ est } \sigma(T)\text{-mesurable alors } E(\mathbf{1}_A \,.\, \varphi \circ T) = E(\mathbf{1}_A \,.\, X)$$

et donc

$$\text{si } g \text{ est borélienne positive alors } E(g \circ T \,.\, \varphi \circ T) = E(g \circ T \,.\, X)$$

On sait que T est une v. a. de densité $t\mathsf{exp}(-t)1_{[0,\infty[}(t)$ si bien que

$$E(g \circ T \,.\, \varphi \circ T) = \int_0^\infty g(t)\varphi(t)t\mathsf{exp}(-t)dt$$

Par ailleurs, on a

$$E(g \circ T \,.\, X) = \int_0^\infty \int_0^\infty g(x+y)x\mathsf{exp}(-x)\mathsf{exp}(-y)dxdy$$

Le changement de variables $(x, y) \to (t = x + y, s = x)$ donne

$$E(g \circ T \,.\, X) = \int_0^\infty \left(g(t)\mathsf{exp}(-t) \int_0^t sds) \right) dt$$

Il s'ensuit que $\varphi(t)t = \int_0^t sds$ et $\varphi(t) = t/2$. On a donc obtenu

$$E(X \mid X + Y) = \frac{X + Y}{2}$$

On montre de la même manière que si h est positive

$$E(h \circ X \mid X + Y) = \frac{1}{X + Y} \int_0^{X+Y} h(s)ds$$

et cela permet d'écrire

$$E(h \circ X \mid X + Y = t) = \frac{1}{t} \int_0^t h(s)ds = \int_0^t h(s) \,.\, t^{-1}ds$$

et on dira que la loi de X sachant $X+Y=t$ est la loi uniforme $t^{-1}ds$ sur $[0,t]$.

Remarque.

Lorsque la loi du couple (X,Y) est donnée par

$$dP_{(X,Y)}(x,y) = f(x,y)d\mu(x)d\gamma(y)$$

où μ et γ sont deux mesures σ-finies, on posera

$$f_2(x) = f_y(x) = \int_{\mathbf{R}} f(x,y)d\gamma(y)$$

de sorte que $f_y(x)d\mu(x)$ est exactement la loi de X.

Noter qu'il résulte du théorème de Fubini que l'on a le droit de dire

$$f \text{ est nulle sur l'ensemble } \{x\ ;\ f_y(x)=0\} \times \mathbf{R}$$

On posera ensuite

$$f(y \mid x) = \begin{cases} f(x,y)/f_y(x) & \text{si} \quad f_y(x) \neq 0 \\ 0 & \text{si} \quad f_y(x) = 0 \end{cases}$$

de sorte que l'on a $f(x,y) = f(y \mid x)f_y(x)$.

Soit h une fonction borélienne positive, l'espérance conditionnelle $E(h \circ Y \mid X)$ est de la forme $\varphi \circ X$ où φ est une fonction borélienne vérifiant

$$\text{si } g \text{ est borélienne positive alors } E(g \circ X \,.\, \varphi \circ X) = E(g \circ X \,.h \circ Y)$$

On va montrer que

$$\varphi(x) = \int_{\mathbf{R}} h(y)f(y \mid x)d\gamma(y)$$

et on dira que $f(y \mid x)d\gamma(y)$ est la loi conditionnelle de Y sachant que $X=x$.

Cela résulte des égalités

$$\begin{aligned} E(g \circ X \,.h \circ Y) &= \int_{\mathbf{R}}\int_{\mathbf{R}} g(x)h(y)f(x,y)d\mu(x)d\gamma(y) \\ &= \int_{\mathbf{R}}\int_{\mathbf{R}} g(x)h(y)f(y \mid x)f_y(x)d\mu(x)d\gamma(y) \\ &= \int_{\mathbf{R}} g(x)\left[\int_{\mathbf{R}} h(y)f(y \mid x)d\gamma(y)\right] f_y(x)d\mu(x) \end{aligned}$$

et de l'identité

$$E(g \circ X \,.\, \varphi \circ X) = \int_{\mathbf{R}} g(x)\varphi(x)f_y(x)d\mu(x)$$

3.5 Loi de Poisson et loi exponentielle

Introduction.

On observe, à partir de l'instant $t = 0$, un flux (une succession) d'événements et on désigne par N_t le nombre d'événements aperçus dans l'intervalle de temps $[0, t[$. C'est une variable aléatoire et on fait les hypothèses suivantes :

1) Le flux est **stationnaire** : La loi de la v. a. $N_{a+t} - N_a$, qui compte le nombre d'événements qui se réalisent dans l'intervalle de temps $[a, a + t[$, ne dépend que de t (et non de a). Cela permet de poser

$$p_n(t) = P(N_t = n) = P(N_{a+t} - N_a = n)$$

2) Le flux est sans **postaction** : Pour toute suite strictement croissante t_n, les v. a. $N_{t_{n+1}} - N_{t_n}$ sont indépendantes.

3) Il existe une constante λ telle que

$$p_1(t) = \lambda t + o(t) \ , \ \sum_{n=2}^{\infty} p_n(t) = o(t)$$

Il en résulte que

$$p_0(t) = 1 - \lambda t + o(t)$$

Calcul de $p_n(a)$, loi de probabilité de N_a

Pour $t = 0$: Il résulte des hypothèses que $p_0(0) = 1$ et donc $p_n(0) = 0$ pour tout $n \geq 1$.

• On a

$$\begin{aligned} p_0(a+t) &= P(N_{a+t} = 0) = P(N_a = 0 \ , \ N_{a+t} - N_a = 0) \\ &= P(N_a = 0)P(N_{a+t} - N_a = 0) \\ &= p_0(a)p_0(t) = p_0(a)(1 - \lambda t + o(t)) \end{aligned}$$

Cela implique $p_0'(a) = -\lambda p_0(a)$ et comme $p_0(0) = 1$, on en déduit que

$$p_0(a) = \mathsf{exp}(-\lambda a)$$

• On a

$$\begin{aligned} p_n(a+t) &= \sum_{i=0}^{n} P(N_a = n - i)P(N_{a+t} - N_a = i) \\ &= p_n(a)(1 - \lambda t + o(t)) + p_{n-1}(a)(\lambda t + o(t)) + o(t) \\ &= (1 - \lambda t)p_n(a) + \lambda t p_{n-1}(a) + o(t) \end{aligned}$$

Cela donne $p_n'(a) = -\lambda p_n(a) + \lambda p_{n-1}(a)$ et, par récurrence sur $n \geq 1$,

$$p_n(a) = \frac{(\lambda a)^n}{n!}\mathsf{exp}(-\lambda a) \ , \ n \geq 0$$

Il en résulte que N_a suit une loi de Poisson de paramètre λa.

La fonction génératrice de la v. a. N_a est

$$g_a(z) = \sum_{n=0}^{\infty} P(N_a = n)\, z^n = \sum_{n=0}^{\infty} p_n(a)\, z^n = \mathsf{exp}(-\lambda a + \lambda a z)$$

et en particulier

$$E(N_a) = \lambda a \ \ , \ \ V(N_a) = \lambda a$$

Il en résulte que la constante λ représente le nombre moyen d'événements aperçus par unité de temps.

La variable aléatoire $T = T^1 = T_1$.

Soit $T = T^1 = T_1$ la v. a. qui mesure le temps d'attente du premier événement.

Nous allons déterminer la densité de sa loi de probabilité : On a

$$P(T > t) = P(N_t = 0) = p_0(t) = \mathsf{exp}(-\lambda t)$$

et il en résulte que la densité de T est donnée par

$$-\frac{d}{dt} P(T > t) = \lambda \mathsf{exp}(-\lambda t)\ ,\ t > 0$$

La v. a. T suit donc la loi exponentielle de paramètre λ et en particulier

$$E(T) = \frac{1}{\lambda} = \sigma(T)$$

Cela donne cette autre interprétation de λ : $\frac{1}{\lambda}$ représente le temps moyen d'attente du premier événement.

Les variables aléatoires T^n , $n \geq 2$.

Soit T^n la v. a. qui mesure le temps d'attente du n-ième événement. On a

$$P(T^n > t) = P(N_t \leq n-1) = \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!} \mathsf{exp}(-\lambda t)$$

et on en déduit que la densité de T^n est donnée par

$$-\frac{d}{dt} P(T^n > t) = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} \mathsf{exp}(-\lambda t)$$

La v. a. T^n suit donc la loi gamma de paramètres (n, λ).

Rappelons que sa fonction caractéristique est donnée par

$$\int_\Omega \mathsf{exp}(ixT^n(\omega))\, dP(\omega) = \left(1 - \frac{ix}{\lambda}\right)^{-n}$$

si bien que son espérance et sa variance sont données par

$$E(T^n) = \frac{n}{\lambda} \ , \ V(T^n) = \frac{n}{\lambda^2}$$

Les variables aléatoires $T_n = T^n - T^{n-1}$, $n \geq 2$.

Notons T_n la v. a. qui mesure le temps qui sépare le n-ième événement du précédent.

Pour déterminer sa loi de probabilité, nous allons commencer par donner celle du couple (T^{n-1}, T^n). On a

$$\begin{aligned} P(T^{n-1} &< a \ , \ T^n > b) = P(N_a = n-1 \ , \ N_b - N_a = 0) \\ &= \frac{(\lambda a)^{n-1}}{(n-1)!}\mathsf{exp}(-\lambda a) \times \mathsf{exp}(-\lambda(b-a)) = \frac{(\lambda a)^{n-1}}{(n-1)!}\mathsf{exp}(-\lambda b) \end{aligned}$$

On en déduit que

$$P(T^{n-1} < a \ , \ T^n < b) = P(T^{n-1} < a) - P(T^{n-1} < a \ , \ T^n > b)$$

si bien que la densité du couple (T^{n-1}, T^n) est donnée par

$$\frac{\partial^2}{\partial a \partial b} P(T^{n-1} < a \ , \ T^n < b) = \lambda^2 \frac{(\lambda a)^{n-2}}{(n-2)!}\mathsf{exp}(-\lambda b) \ \ , \ \ 0 < a < b$$

On en déduit que

$$\begin{aligned} P(T_n &< b) = P(T^n - T^{n-1} < b) = \int_{\{y-x<b\}} \lambda^2 \frac{(\lambda x)^{n-2}}{(n-2)!}\mathsf{exp}(-\lambda y) dx dy \\ &= \int_0^\infty \lambda^2 \frac{(\lambda x)^{n-2}}{(n-2)!} \int_x^{x+b} \mathsf{exp}(-\lambda y) dy \, dx \end{aligned}$$

et la loi de T_n est donc donnée par

$$\frac{d}{db} P(T_n < b) = \int_0^\infty \lambda^2 \frac{(\lambda x)^{n-2}}{(n-2)!}\mathsf{exp}(-\lambda(x+b)) dx = \lambda \mathsf{exp}(-\lambda b)$$

En particulier les v. a. T_n , $n \geq 1$ suivent toutes la loi exponentielle de paramètre λ.

Indépendance des T_n , $n \geq 1$.

Commençons par montrer que T_1 et T_2 sont indépendantes.

On a $\{T_1 < x \ , \ T_2 < y\} = \{T^1 < x \ , \ T^2 - T^1 < y\}$ si bien que

$$P(T_1 < x \ , \ T_2 < y) = \int_0^x \left(\int_a^{a+y} \lambda^2 \mathsf{exp}(-\lambda b) db \right) da$$

et il en résulte que

$$\begin{aligned} \frac{\partial^2}{\partial x \partial y} P(T_1 < x \ , \ T_2 < y) &= \frac{\partial}{\partial x} \lambda^2 \int_0^x \mathsf{exp}(-\lambda(a+y)) da \\ &= \lambda \mathsf{exp}(-\lambda x) \times \lambda \mathsf{exp}(-\lambda y) \end{aligned}$$

et cette expression de la densité du couple (T_1, T_2) prouve que ces deux v. a. T_1 et T_1 sont indépendantes.

Montrons par exemple que T_1, T_2, T_3 et T_4 sont indépendantes.

Si $a < b < c < d$, alors l'événement

$$A = \{T^1 < a,\, T^2 > b,\, T^3 < c,\, T^4 > d\}$$

coïncide avec l'événement

$$\{N_a = 1,\, N_b - N_a = 0,\, N_c - N_b = 2,\, N_d - N_c = 0\}$$

si bien que

$$\begin{aligned} P(A) &= \lambda a e^{-\lambda a} \times e^{-\lambda(b-a)} \times \frac{\lambda^2 (c-b)^2}{2} e^{-\lambda(c-b)} \times e^{-\lambda(d-c)} \\ &= \frac{1}{2}\lambda^3 a (c-b)^2 e^{-\lambda d} \end{aligned}$$

La densité de la loi de (T_1, T_2, T_3, T_4) est donc donnée par

$$\left(\frac{\partial}{\partial a}\right)\left(\frac{-\partial}{\partial b}\right)\left(\frac{\partial}{\partial c}\right)\left(\frac{-\partial}{\partial d}\right) P(A) = \lambda^4 e^{-\lambda d} \mathbf{1}_{\{a<b<c<d\}}$$

On peut donc calculer la probabilité de l'événement

$$\begin{aligned} B &= \{T_1 < x,\, T_2 < y,\, T_3 < z,\, T_4 < t\} \\ &= \{T^1 < x,\, T^2 - T^1 < y,\, T^3 - T^2 < z,\, T^4 - T^3 < t\} \end{aligned}$$

qui vaut

$$\alpha = \int_0^x \Big(\int_{t_1}^{y+t_1} \Big(\int_{t_2}^{z+t_2} \Big(\int_{t_3}^{t+t_3} \lambda^4 e^{-\lambda t_4} dt_4\Big) dt_3\Big) dt_2\Big) dt_1$$

Cela permet de déterminer la densité de (T^1, T^2, T^3, T^4) ; elle vaut

$$\begin{aligned} \frac{\partial^4 \alpha}{\partial x y z t} &= \frac{\partial^3}{\partial y z t} \int_x^{y+x} \Big(\int_{t_2}^{z+t_2} \Big(\int_{t_3}^{t+t_3} \lambda^4 e^{-\lambda t_4} dt_4\Big) dt_3\Big) dt_2 \\ &= \frac{\partial^2}{\partial z t} \int_{y+x}^{z+y+x} \Big(\int_{t_3}^{t+t_3} \lambda^4 e^{-\lambda t_4} dt_4\Big) dt_3 \\ &= \frac{d}{dt} \int_{z+y+x}^{t+z+y+x} \lambda^4 e^{-\lambda t_4} dt_4 \\ &= \lambda^4 \mathsf{exp}(-\lambda(t+z+y+x)) \end{aligned}$$

et cette expression de la densité prouve que les v. a. T_1, T_2, T_3 et T_4 sont indépendantes.

Il devient clair maintenant que T_n $(n \geq 1)$ est une suite de v. a. indépendantes.

3.6 La loi du Chi deux

La loi gamma.

Une variable aléatoire X suit une loi gamma de paramètres (α, λ) ($\alpha > 0,\ \lambda > 0$) si sa densité f est donnée par

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\lambda x} \mathbf{1}_{[0,\infty[}(x)$$

La fonction caractéristique d'une telle v. a. est donnée par

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}$$

et il en résulte que son espérance et sa variance sont donnés par

$$E(X) = \frac{\alpha}{\lambda}\ ,\ \ V(X) = \frac{\alpha}{\lambda^2}$$

Pour $\alpha = 1$, la loi gamma se réduit à la loi exponentielle.

Pour $\alpha = n/2$ et $\lambda = 1/2$, la loi gamma est la loi du χ^2 à n degrés de liberté.

Exemples.

- Si la v. a. X suit une loi normale réduite, alors la v. a. $Y = X^2$ suit une loi gamma de paramètre $(1/2, 1/2)$.

Preuve.

$$\frac{d}{da}P(Y < a) = \frac{d}{da}\left(\frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{a}} \mathsf{exp}(-x^2/2)dx\right) = \frac{1}{\sqrt{2\pi}} a^{-1/2} e^{-a/2}$$

- Si les v. a. X_i sont indépendantes est suivent la loi normale réduite, la v. a.

$$V = X_1^2 + \cdots + X_n^2$$

suit la loi du χ^2 à n degrés de liberté.

Preuve. La fonction caractéristique de V est en effet donnée par

$$\varphi(t) = \left[(1 - 2it)^{-1/2}\right]^n$$

Théorème.

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes qui suivent la loi normale de paramètres $(m,\, \sigma^2)$ et posons

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i\ ,\ Y = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2\ ,\ Z = \sqrt{n}\frac{\overline{X} - m}{\sigma}$$

Les v. a. Y et Z sont indépendantes et Y suit la loi du χ^2 à $n-1$ degrés de liberté.

Preuve.

On se ramène au cas où $m = 0$ et $\sigma = 1$. Posons

$$Y_i = \frac{1}{\sqrt{i.(i+1)}}(X_1 + \cdots + X_i - iX_{i+1}) \text{ et } Y_n = \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$$

C'est une transformation "orthogonale" si bien que

$$X_1^2 + \cdots + X_n^2 = Y_1^2 + \cdots + Y_n^2$$

et comme $Y_n^2 = n\overline{X}^2$, il en résulte que

$$Y = \sum_{i=1}^{n} Y_i^2 - Y_n^2 = \sum_{i=1}^{n-1} Y_i^2$$

et cela implique que Y suit la loi du χ^2 à $n-1$ degrés de liberté.

Comme la transformation ci-dessus est orthogonale et que la loi de $(X_1, \ldots, X_n)$ est invariante par ces transformations, on en déduit que les v. a. $Y_1, \ldots, Y_n$ sont indépendantes et il en résulte que Y et Z sont indépendantes.

Convergence de la loi multinomiale.

Soit $p_1, \ldots, p_k$ des réels tels que

$$0 < p_i < 1 \text{ et } p_1 + \cdots + p_k = 1$$

Soit $Z = (Z_1, \ldots, Z_k)$ une variable aléatoire à valeurs dans $\mathbf{N}^k$. On dit que Z suit une loi multinomiale de paramètres $(p_1, \ldots, p_k\ ;\ n)$ si

$$P(Z_1 = n_1, \ldots, Z_k = n_k) = \frac{n!}{n_1! \times \cdots \times n_k!}\, p_1^{n_1} \times \cdots \times p_k^{n_k}$$

pour $n_1 + \cdots + n_k = n$.

Interprétation.

Si les événements $E_1, \ldots, E_k$ se produisent respectivement avec les probabilités $p_1, \ldots, p_k$, alors la probabilité pour que l'événement se produise n_1 fois, $\ldots$, E_k se produise n_k fois est donnée par

$$\frac{n!}{n_1! \times \cdots \times n_k!}\, p_1^{n_1} \times \cdots \times p_k^{n_k}$$

Cette loi est une généralisation de la loi binomiale et s'appelle loi multinomiale parce que les probabilités ci dessus sont les termes du développement de $(p_1 + \cdots + p_k)^n$.

Noter que chacune des v. a. Z_i sui une loi binomiale de paramètre (n, p_i)

Considérons maintenant la variable aléatoires

$$\chi^2 = \sum_{i=1}^{k} \frac{(Z_i - np_i)^2}{np_i}$$

Comme $E(Z_i - np_i)^2 = V(Z_i) = np_i(1 - p_i)$, il en résulte que

$$E(\chi^2) = k - 1 = f$$

La variance de χ^2 est plus difficile à calculer et vaut

$$V(\chi^2) = 2f + \frac{1}{n}\left(\frac{1}{p_1} + \cdots + \frac{1}{p_k} - k^2 - 2k + 2\right)$$

et il suffit de retenir que pour n assez grand, cette variance est voisine de $2f$.

Théorème.

Lorsque n tend vers l'infini, la variable aléatoire χ^2 tend en loi vers une loi de χ^2 à $k-1$ degrés de liberté.

Application.

On désire vérifier la régularité d'un dé, c'est-à-dire examiner si toutes ses faces sont équiprobables :

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = \frac{1}{6}$$

On a lancé ce dé 120 fois et on a obtenu

issue i	1	2	3	4	5	6
fréquence observée X_i	15	21	25	19	14	26
fréquence espérée np_i	20	20	20	20	20	20

Pour mesurer les écarts entre les X_i et les np_i, on calcule

$$\chi^2 = \sum_{i=1}^{6} \frac{(X_i - np_i)^2}{np_i} = 6,2$$

La valeur observée de χ^2 est $6,2$; elle est inférieure $f + \sqrt{2f} \simeq 8,16$ et on n'a pas de raison de douter du dé.

En règle générale, on adopte ces deux principes :

1) Lorsque la valeur observée est supérieure à $f + 2\sqrt{2f}$, on doit rejeter l'hypothèse.

2) Lorsque la valeur observée est trop petite, on est en droit de soupçonner quelqu'un d'avoir trafiqué les données pour démontrer quelque chose.

3.7 Exercices

Exercice. Variable de Bernoulli.

Soit $A \in \mathcal{T}$ un événement et $X = 1_A$. Cette v.a. ne prend que les valeurs 0 et 1 et sa loi de probabilité est donc portée par l'ensemble $\{0, 1\}$; elle est donnée par

$$P_X = P(X = 1)\delta_1 + P(X = 0)\delta_0$$

On dit qu'une v.a. X est de **Bernoulli** de paramètre p , $0 < p < 1$ si sa loi de probabilité est égale à

$$P_X = p\delta_1 + (1-p)\delta_0$$

Montrer que $E(X) = p$ et $V(X) = p(1-p)$.

Exercice. Fonction de répartition.

Soit $(\Omega\,,\,\mathcal{T}\,,\,P\,)$ un espace probabilisé et $X : \Omega \to \mathbf{R}$ une variable aléatoire. La fonction de répartition de X est la fonction $F : \mathbf{R} \to \mathbf{R}$ définie par

$$F(x) = P(\,X < x\,)$$

1) Montrer que F est une application croissante, continue à gauche et que

$$F(-\infty) = 0 \;\;,\;\; F(+\infty) = 1$$

2) Déterminer F lorsque la loi de probabilité de X est donnée par

$$a)\; P_X = 0.2\delta_0 + 0.8\delta_2 \qquad\qquad b)\; dP_X(x) = e^{-x}\mathbf{1}_{[0,\infty[}(x)dx$$

3) Montrer que $P(X = a) = F(a+) - F(a)$.

4) Montrer que si F est de classe $\mathcal{C}^1$ sur un intervalle $]a, b[$, alors

$$P(\,X \in A\,) = \int_A F'(x)dx$$

pour tout borélien A inclus dans $]a, b[$.

5) Montrer que si F est continue sur tout $\mathbf{R}$ et dérivable sauf en un nombre fini de points, alors

$$P(x < x) = F(x) = \int_{-\infty}^{x} F'(t)dt \;\;,\;\; x \in \mathbf{R}$$

Cela montre que la loi de X est une mesure à densité par rapport à la mesure de Lebesgue :

$$dP_X(x) = f(x)dx \;\;,\;\; f(x) = \frac{d}{dx}\,P(\,X < x\,)$$

6) Déterminer la loi de probabilité d'une v. a. X dont la fonction de répartition F vérifie

$$F(x) = \begin{cases} 0 & \text{si} \quad x \leq 0 \\ x/3 & \text{si} \quad x \leq 1 \\ 1/2 & \text{si} \quad 1 < x \leq 6 \\ (x-6)^2/16 + 3/4 & \text{si} \quad 6 < x \leq 8 \end{cases}$$

Exercice. Soit X une variable aléatoire dont la loi est donnée par

$$dP_X(x) = e^{-x}\mathbf{1}_{[0,\infty[}(x)dx$$

1) Déterminer les fonctions de répartitions puis les lois de probabilité des variables aléatoires

$$a)\ Y = 2X + 6 \qquad b)\ Z = |X| \qquad c)\ T = \ln X$$

2) Déterminer les espérances des v. a. X , Y , Z et T.

Exercice. Soit X et Y deux v. a. indépendantes et de même loi donnée

$$dP_X(t) = dP_Y(t) = t^{-2}\mathbf{1}_{[1,\infty[}(t)dt$$

1) Calculer les probabilités $P(X \leq -1)$, $P(2 < X \leq 3)$ et $P(2 < X < 3\ ,\ 4 < Y \leq 6)$.

2) Déterminer les lois de probabilité des v. a. X^2 et $\ln X$.

3) Calculer les probabilités $P(XY \leq 2)$, $P(Y/X < 1/2)$ et $P(Y/X \leq 2)$.

4) En déduire que les v. a. $U = XY$ et $V = Y/X$ ne sont pas indépendantes

5) Déterminer les lois de probabilité des v. a. $U = XY$ et $V = Y/X$.

Exercice. Soit X, Y deux variables aléatoires indépendantes et normales de paramètres $(1, 0)$:

$$dP_X(t) = dP_Y(t) = \frac{1}{\sqrt{2\pi}}\mathsf{exp}(-t^2/2)dt$$

1) Déterminer la loi de X^2.

2) Déterminer la loi de $X^2 + Y^2$.

3) Déterminer la loi de $\sqrt{X^2 + Y^2}$.

4) Déterminer la loi de Y/X.

Exercice. Soit X, Y deux variables aléatoires indépendantes et normales de paramètres (m, σ^2) :

$$dP_X(t) = dP_Y(t) = \frac{1}{\sqrt{2\pi\sigma}}\mathsf{exp}\left(-\frac{(t-m)^2}{2\sigma^2}\right)dt$$

1) Déterminer la fonction caractéristique de la v. a. X.

2) Déterminer l'espérance et la variance de X.

2) Que peut-on dire de la v.a. $(X - m)/\sigma$?

3) Que peut-on dire de la v. a. $X + Y$?

Remarque. La v.a. X prend avec une probabilité élevée des valeurs proches de son espérance :

k	$0,6745$	1	$1,645$	$1,96$	2	$2,58$	3
$P(\|X - m\| \leq k\sigma)$	$0,5$	$0,6827$	$0,9$	$0,95$	$0,9545$	$0,99$	$0,9973$

Deux de ces probabilités méritent d'être retenus

$$P(|X-m|>1,96\sigma)=0,05\ \ ,\ \ P(|X-m|>2,58\sigma)=0,01$$

Exercice.

1) Soit $X:\Omega\to\mathbf{N}$ une v. a. entière, prouver que

$$E(X)=\sum_{n=0}^{\infty}P(X>n)$$

2) Soit $X:\Omega\to[0,\infty[$ une variable aléatoire, prouver que

$$E(X)=\int_0^{\infty}P(X>x)dx$$

et montrer que si X est intégrable alors $\lim_{x\to\infty}xP(X>x)=0$.

Exercice. Loi géométrique.

On dit qu'une v. a. X à valeurs dans $\mathbf{N}^*$ suit une loi géométrique de paramètre p , $0<p<1$, si sa loi est donnée par

$$P(X=n)=p(1-p)^{n-1}\ \ ,\ \ n\in\mathbf{N}^*$$

1) Déterminer la fonction génératrice de X :

$$g(z)=E(z^X)=\sum_{n=1}^{\infty}P(X=n)z^n$$

2) Calculer $E(X)$, $E(X(X-1))$ et en déduire $V(X)$.

3) Montrer que cette v. a. vérifie

$$P(X>m+n\mid X>n)=P(X>m)$$

Application. On effectue des essais indépendants de probabilité de succès constante et égale à p , $0<p<1$, et on note X la v. a. qui compte le nombre de coups nécessaires pour obtenir le premier succès.

1) Calculer $P(X=1),\ P(X=2),\ P(X=3)$ et $P(X=n)$.

2) Que dire de X et que vaut son espérance ?

Exercice. Loi binomiale négative.

On dit qu'une v. a. X suit une loi binomiale négative de paramètres m et p, $m\in\mathbf{N}^*$ et $0<p<1$, si sa loi est donnée par

$$P(X=n)=\binom{n-1}{m-1}p^m(1-p)^{n-m}\ \ ,\ \ n\geq m$$

1) Que peut-on dire de X lorsque $m=1$?

2) Déterminer la fonction génératrice de X :

$$g(z) = E(z^X) = \sum_{n=1}^{\infty} P(X = n) z^n$$

3) Calculer $E(X)$, $E(X(X-1))$ et en déduire $V(X)$.

Application. On effectue des essais indépendants de probabilité de succès constante et égale à p , $0 < p < 1$, et on note X_m la v. a. qui compte le nombre de coups nécessaires pour obtenir le m-ième succès.

1) Montrer que si $n \geq m \geq 1$ alors

$$P(X_m = n) = \binom{m-1}{n-1} p^m (1-p)^{n-m}$$

2) Que dire de X_m et que vaut son espérance.

Exercice. Soit $X_1, \ldots, X_m$ des v. a. indépendantes suivant la loi géométrique de paramètre p et posons

$$S = X_1 + \cdots + X_m$$

1) Déterminer la fonction génératrice de S et en déduire que S suit la loi binomiale négative de paramètres m et p.

2) Que peut-on dire de la loi de la somme de deux v. a. indépendantes suivant des lois binomiales négatives de paramètres (m, p) et (n, p) ?

Exercice. On suppose que la v. a. X suit une loi binomiale négative de paramètre (m, p).

1) Déterminer le maximum de $p \to P(X = n)$ en fonction de n et m.

2) Montrer que si $m \geq 2$

$$E\left(\frac{m-1}{X-1}\right) = p$$

3) Montrer que si $m \geq 2$

$$E\left(\frac{m}{X}\right) > p$$

Exercice. Loi de Poisson.

On dit qu'une v. a. X suit une loi de Poisson de paramètre a si sa loi est donnée par

$$P(X = n) = e^{-a} \frac{a^n}{n!} \quad , \quad n \in \mathbf{N}$$

1) Déterminer la fonction génératrice de X

$$g(z) = E(\, z^X \,) = \sum_{n=0}^{\infty} P(X = n)\, z^n$$

2) En déduire que

$$E(X) = V(X) = a$$

3) Que peut-on dire de la somme de deux v. a. indépendantes suivant des loi de Poisson ?

Soit p_n une suite d'éléments de $]0,1[$ tels que la limite

$$a = \lim_{n\to\infty} np_n$$

existe. Montrer que, pour tout $k \in \mathbf{N}$, on a

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\,a}\,\frac{a^k}{k!}$$

Pour $n = 50$ et $p = 0.04$, on a $np = 2$

k	Binomiale	Poisson		k	Binomiale	Poisson
0	0.1299	0.1385		5	0.0346	0.0361
1	0.2706	0.2707		6	0.0108	0.0120
2	0.2762	0.2767		7	0.0028	0.0034
3	0.1842	0.1804		8	0.0006	0.0009
4	0.0902	0.0902		9	0.0001	0.0002

Exercice. Soit $X_1,\ X_2 : \Omega \longrightarrow R$ deux variables aléatoires indépendantes et posons

$$X_{(1)} = \min(X_1, X_2) \quad , \quad X_2 = \max(X_1, X_2)$$

1) Déterminer les fonctions de répartition de $X_{(1)}$ et $X_{(2)}$ en fonction de celles de X_1 et X_2.

2) On suppose que $X_1,\ X_2$ suivent des lois exponentielles de paramètres a et b

$$dP_{X_1}(x) = ae^{-ax}\mathbf{1}_{[0,\infty[}(x)dx \quad , \quad dP_{X_2}(x) = be^{-bx}\mathbf{1}_{[0,\infty[}(x)dx$$

Déterminer les densités de $X_{(1)}$ et $X_{(2)}$.

3.8 Processus stochastiques

Introduction

Les processus stochastiques décrivent l'évolution d'une grandeur aléatoire en fonction du temps ou bien de l'espace. Il existe de nombreuses applications des processus aléatoires notamment en physique statistique (par exemple le ferromagnétisme, les transitions de Phases), en biologie (l'évolution, la génétique et la génétique des populations), en médecine (croissance de tumeurs, épidémie),

et bien entendu les sciences de l'ingénieur. Dans ce dernier domaine, les applications principales sont pour l'administration des réseaux, de l'Internet, des télécommunications ainsi que dans le domaine de l'économie et des finances. Un processus stochastique tend à représenter l'évolution d'un système en fonction du temps. Il met en oeuvre des modèles probabilistes spécifiques dans le but de gérer l'incertitude et la manque d'information. Un processus stochastique (aléatoire) représente une évolution généralement dans le temps, d'une variable aléatoire.

Notations

Dans la suite, on adoptera les notations suivantes :

- $[\Omega, F, IP]$ un espace de probabilité
- (A, $\boldsymbol{A}$) un espace mesurable
- $\boldsymbol{E}$ l'espace où X(t) prend ses valeurs : l'espace d'état du processus stochastique X
- T l'espace du temps

Définition

- ***Processus stochastique*** : Un processus stochastique (ou processus aléatoire) représente une évolution, génèralement dans le temps, d'une variable aléatoire.

  On appelle processus aléatoire à valeurs dans (A, $\boldsymbol{A}$), un élément $((X_t(\omega))_{t\geq 0,\omega\in\Omega})$, où pour tout $t \in T$, X_t est une variable aléatoire à valeurs dans (A, $\boldsymbol{A}$). Si $(F_t)_t$ est une filtration, on appelle processus aléatoire adapté, à valeurs dans (A, $\boldsymbol{A}$), un élément X tel que X_t soit une variable aléatoire|–mesurable à valeurs dans (A, $\boldsymbol{A}$). Pour $\omega \in \Omega$ fixé, la fonction de $\Re_+$ dans A qui à t associe $X_t(\omega)$ est appelée la trajectoire associée à la réalisation ω. Un processus stochastique peut aussi être vu comme une fonction aléatoire : à chaque ω dans Ω, on associe la fonction de T dans E :$t \mapsto X_t(\omega)$, appelée la trajectoire associée à la réalisation ω.
  - Si T est dénombrable alors X est appelé un processus stochastique en temps discret.
  - Si T est non dénombrable alors X est appelé un processus stochastique en temps continu.
  - Si E est dénombrable alors X est appelé un processus stochastique à espace discret.
  - Si E est non dénombrable alors X est appelé un processus stochastique à espace continu.

  Exemple

  Soit le processus stochastique somme $S_t = S_0 + \sum_{i=1}^{t} X_i$

  Avec $X_1, X_2...$ une suite des variables aléatoires indépendantes, prenant chacune la valeur 1 avec probabilité p et la valeur -1 avec probabilité 1-p. La suite est une marche aléatoire partant de S_0 . S_t représente la fortune

d'un joueur ayant joué t parties, recevant 1 dinars s'il gagne payant 1 dinars s'il perd et ayant une richesse initiale de francs S_0 dinars.

- **Martingale**
  Définition : On se donne un espace de probabilité $[\Omega, F, IP]$ muni d'une filtration $(F_t)_t$. $(F_t)_t$ est donc une famille croissante de sous-tribus de F. Une famille de variable aléatoires $(X_t)_{t\geq 0}$ est une martingale par rapport à la filtration F_t si :

  - X_t est F_t-mesurable et intégrable pour tout t
  - $E[X_t/F_s] = X_s, \forall s \leq t$

Exemples des processus stochastiques

- **Processus stochastique en temps discret et à espace discret**
  X_n=nombre de programmes exécutés durant la n-ième heure de la journée.
  E=X_n,n=1,2,...,24 est un processus stochastique en temps discret et à espace discret.
  Avec : T=1,2,...,24 et E=N
- **Processus stochastique en temps discret et à espace continu**
  X_n=temps mis par un serveur Web pour traiter la n-ième requête de la journée.
  E=X_n,n=1,2,... est un processus stochastique en temps discret et à espace continu.
  Avec : T=N et E=$[0, \infty[$
- **Processus stochastique en temps continu et à espace discret**
  X(t) = nombre de bits qui traversent un routeur Internet donné dans [0,t].
  Alors,$\{X(t), t \geq 0\}$ est un processus stochastique en temps continu et à espace discret avec E=N,T= [0,t].
- **Processus stochastique en temps continu et à espace continu**
  X(t) = temps d'attente d'une requête (par exemple à un serveur Web) reçue à l'instant t.

3.9 Processus de Markov

Introduction

La notion d'une chaîne de Markov a été conçue, au début du vingtième siècle, par le Russe mathématicien A.A. Markov qui a étudié l'alternance des voyelles et des consonnes en poésie .Onegin de Pushkin. Markov a développé un modèle de probabilité dans lequel les résultats des épreuves successives dépendent les uns des autres et chaque épreuve dépend seulement de son prédécesseur immédiat. Ce modèle, étant la généralisation la plus simple de la probabilité des épreuves indépendantes et semble donner une excellente description de l'alternance des

voyelles et des consonnes. Markov a ainsi pu calculer avec une manière précise la fréquence à laquelle les consonnes se produisent en poésie de Pushkin. La théorie des processus de Markov apparaît dans de nombreuses applications telles que la biologie,l'informatique, la technologie et la recherche opérationnelle. Un processus de Markov permet de modéliser l'incertitude affectant plusieurs systèmes réels dont la dynamique varie au cours du temps. Les concepts de base d'un processus de Markov sont ceux d'un état et d'une transition d'état. La modélisation de certaines applications, consiste à trouver une description d'état tel que le processus stochastique associé vérifie la propriété markovienne c'est-à-dire que la connaissance de l'état actuel est suffisante pour prévoir le comportement futur stochastique. Un processus de Markov est une séquence aléatoire dont le comportement futur probabiliste du processus dépend seulement de l'état actuel du processus et non plus de son passé. On parle alors de la propriété Markovienne. Les processus de Markov constituent l'exemple le plus simple des processus stochastiques, lorsque dans l'étude d'une suite de variables aléatoires, on abandonne l'hypothèse d'indépendance On distingue deux types de processus de Markov :

- Les processus de Markov à temps discret appelés aussi chaîne de Markov à temps discret dans lesquels les transitions d'état se produisent dans des instants de temps fixes.

- Les processus de Markov à temps continu appelés aussi chaîne de Markov à temps continu ou encore processus Markoviens de sauts dans lesquels l'état peut changer en tout point de temps.

Processus de Markov à temps discret

Un processus de Markov à temps discret est un processus stochastique à temps discret représentant une séquence des variables aléatoires indépendantes.
Définitions
Soit E un espace dénombrable, c'est-à-dire soit fini, soit en bijection avec N .

- **Noyau** :
  on appelle noyau (ou matrice de probabilité) de transition sur E une famille $P(i,j), i,j \in E$ de réels telle que :
  (i) P(i,j) ≥ 0, pour tout i,j $\in E$
  (ii) Pour tout i $\in$ E, $\sum_{j \in E} P(i,j) = 1$

Formulation mathématique

- **Processus de Markov** : Soit $X_n (n \geq 0)$ une suite des variables aléatoires à valeurs dans l'ensemble E (E=N). On dit que cette suite est un processus de Markov à temps discret, si :
  Pour tout $n \geq 1$ et pour toute suite $(i_0, ..., i_n)$) d'élèments de E, telle que la probabilité $P(X_0 = i_0, ..., X_{n-1} = i_{n-1}, X_n = i_n)$ est strictement positive : On a la relation suivante entre les probabilités conditionnelles :

$P\{X_{n+1} = j \backslash X_0 = i_0, ..., X_{n-1} = i_{n-1}, X_n = i_n\} = P\{X_{n+1} = j \backslash X_n = i_n\}$
Autrement dit, dans l'évolution au cours du temps, l'état du processus à l'instant (n+1) ne dépend que de celui à l'instant n précèdent, mais non de ses états antérieurs. On dit que le processus est sans mémoire ou non héréditaire.

Remarque : Le processus de Markov à temps discret est dit homogène (dans le temps), si la probabilité précédente ne dépend pas de n.

- **Matrice de passage (ou de transition)** :la matrice P définie par :

$$P = \begin{pmatrix} p_{0,0} & p_{0,1} & p_{0,2} & \cdots \\ p_{0,0} & p_{0,0} & p_{0,0} & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

$p_{i,j} = P\{X_{n+1} = j \backslash X_n = i\}\ (n \geq 0)$
Cette probabilité est appellé la probabilité de passage de l'état i à l'état j, en une étape, ou une opération, ou encore, en une transition.

Processus de Markov à temps continu (Processus Markoviens à sauts)

Un processus de Markov à temps continu est un processus aléatoire $(X_t)_{t\geq 0}$ dont l'évolution future est indépendante du passé sachant l'état présent.Dans ce qui suit , on considère que le processus $(X_t)_{t\geq 0}$ prend ses valeurs dans un espace d'état E qui est dénombrable, typiquement E fini ou E=N.

Définitions

- **Processus de Markov à temps continu** :
  Le processus stochastique $(X_t)_{t\geq 0}$ est une chaîne de Markov à temps continu si
  $\forall 0 \leq t_0 < t_1 < ... < t_n < t\ P(X_t = x \backslash X_{t_0} = x_0, ..., X_{t_n} = x_n) = P(X_t = x \backslash X_{t_n} = x_n)$
  La loi de la variable aléatoire $(X_t)_{t\geq 0}$ est donnée par le vecteur ligne de probabilité suivant :
  $p(t) = [p_0(t), p_1(t), ...] = [P(X_t = 0), P(X_t = 1), ...]$
  Les probabilités de transition $p(t, t^{'})$ sont les matrices (éventuellement de taille infinie)
  $p_{i,j}(t, t^{'}) = P(X_t = j \backslash X_t = i)$
  Dans de nombreux modèles, les transitions entre deux instants t et $t^{'}$ ne dépendent pas des dates t et $t^{'}$, mais plutôt de la durée entre ces deux dates, c'est-à-dire de $(t^{'} - t)$. C'est ce qu'on appelle l'homogénéité du processus.
- **Processus de Markov homogène** :
  le processus de Markov $(X_t)_{t\geq 0}$ est homogène si les probabilités de transi-

tion $p(t,t^{'})$ ne dépendent que de $(t-t^{'})$. On note alors :
$p_{ij}(t) = p_{ij}(0,t) = P(X_t = j \backslash X_0 = i)$

– **Equations de Chapman-Kolmogorov** :
Les transitions de probabilités satisfont les équations suivantes, dites de Chapman-Kolmogorov :
Pour tous entiers i et j, pour tous réels positifs t et $t^{'}$, on a :
$p_{ij}(t+t^{'}) = \sum_{k=0}^{+\infty} p_{ik}(t)p_{kj}(t^{'})$

Preuve :
$\Omega = \coprod_{k=0}^{+\infty} \{X_t = k\}$ qui est une partition de l'éspace d'états.

$p_{ij}(t+t^{'}) = P(X_{t+t'} = j \backslash X_0 = i) = \frac{\mathrm{P(X_{t+t'}=j, X_0=i)}}{P(X_0=i)} = \sum_{\mathrm{k=0}}^{+\infty} \frac{\mathrm{P(X_{t+t'}=j, X_t=k, X_0=i)}}{P(X_0=i)}$

Et en reconditionnant :
$p_{ij}(t+t^{'}) = \sum_{k=0}^{+\infty} P(X_{t+t'} = j \backslash X_t = k, X_0 = i) P(X_t = k \backslash X_0 = i)$
Or la propriété de Markov assure que :
$P(X_{t+t'} = j \backslash X_t = k, X_0 = i) = P(X_{t+t'} = j \backslash X_t = k) = p_{ij}(t^{'})$
La dernière égalité venant de l'homogénéité de la chaîne.
D'où finalement :
$p_{ij}(t+t^{'}) = \sum_{k=0}^{+\infty} p_{kj}(t^{'})p_{ik}(t)$

Notation
Il est alors naturel d'introduire les matrices (infinies) de transition :
$(P(t))_{t\geq 0}$ pour tout $t \geq 0$ par :
$P(t) = [p_{ij}(t)]_{(i,j)\in IN^2}$
Les équations de Chapman-Kolmogorov se résument alors comme suit :
pour tous réels positifs t et $t^{'}$
$P(t+t^{'}) = P(t)P(t^{'})$
A tout instant t, la somme de chaque ligne de P (t) vaut 1 (c'est la somme d'une série). On peut donc voir une chaîne de Markov en temps continu comme une famille de matrice $(P(t))_{t\geq 0}$ satisfaisant l'équation ci-dessous. Néanmoins, il convient d'ajouter une hypothèse de régularité, à savoir :

$\forall (i,j) \in IN^2, p_{ij}(t) \to \delta_{ij}$

$\mathrm{t} \to 0$

c'est-à-dire : $\lim_{t\to 0} P(t) = I$
I est la matrice identité. On parle alors de processus de Markov standard.

– **Générateur infinitésimal** :
La matrice A définie par :
$A = P^{'}(0) = \lim_{t\to 0^+} \frac{P(t)-I}{t}$
est appelée générateur infinitésimal du processus de Markov à temps continu.

On a donc :

$a_{ij} = \lim_{t \to 0^+} \frac{p_{ij}(t)}{t}$ si i $\neq$ j

$a_{ii} = \lim_{t \to 0^+} \frac{p_{ii}(t)-1}{t}$ sinon

Et on peut écrire les développements limités à l'ordre 1 :

$p_{ij}(t) = a_{ij}t + o(t)$

Si le processus est dans l'état i initialement, la probabilité qu'elle l'ait quitté à l'instant t est environ a_{ii}t. Le coefficient positif -a_{ii} le taux instantané de départ de l'état i.

En reprenant les équations précédentes, on montre par ailleurs que :

$\forall i \in IN - \mathrm{a_{ii}} = \sum_{j \neq i} a_{ij}$

En d'autres termes, la somme de chaque ligne de A est nulle.

- **Processus de Poisson** :
  une chaîne de Markov en temps continu est complètement définie à partir de son générateur infinitésimal A et de la distribution initiale p(0). Un processus de Poisson est une chaîne de Markov en temps continu $(N(t))_{t \geq 0}$ à valeurs dans N de générateur A tel que :

  $a_{ii} = -\lambda$

  $a_{i,i+1} = \lambda$

  $a_{ij} = 0$ sinon

  On dit alors que le processus est de densité ou de taux λ .
- **Remarques** :
  - Concernant la condition initiale, on supposera généralement que $N_0 = 0$, i.e p(0) =δ_0
  - On note ce processus $(N(t))_{t \geq 0}$ plutôt que $(X(t))_{t \geq 0}$, car N_t correspond souvent au nombre d'évènements qui sont survenus entre l'instant 0 et l'instant t, par exemple le nombre de voitures arrivant à un péage, les appels à un central téléphonique, les émissions de particules radioactives. C'est pourquoi on l'appelle aussi processus de comptage.
- **Loi du nombre de points d'un processus de Poisson dans un intervalle donné** :
  Soit $P_n(t)$ la probabilité que exactement n événements arrivent dans un intervalle de longueur t,c'est-à-dire, $P_n(t) = P(N(t) = n)$.
  On a, pour tout

  $n \in 0, 1, 2, ... \geq 0$

  $P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$

- **Nombre moyen de points d'un processus de Poisson dans un intervalle de longueur t** :
  Pour tout $t \geq 0$ E [N(t)] $= \lambda t$
  Ainsi, le nombre moyen d'événements par unité de temps est donné par λ.
  Il y a en fait un lien très fort entre un processus de Poisson et la loi exponentielle.
  Pour cela, on considère le temps τ qui sépare l'occurrence de deux événements consécutifs d'un processus de Poisson.
- **Loi des interarrivées d'un processus de Poisson** :
  Pour chaque $x \geq 0, P(\tau \leq x) = 1 - e^{-\lambda x}$
  Ainsi, le temps qui sépare l'occurrence de deux points consécutifs d'un processus de Poisson d'intensité λ est distribué selon une loi exponentielle de paramètre λ.
- **Propriété sans mémoire de la loi exponentielle** :
  $P(X > x + y \backslash X > x) = P(X > y)$
- **Exemples de Processus Markoviens de sauts**

  - Une machine est soit en bon état de fonctionnement, état 1, ou bien en panne, état 0. Lorsqu'elle est en panne, on appelle un réparateur, qui met un temps aléatoire à la réparer. Dans ce cas E= 0,1 est un espace d'états finis.

  - On observe les arrivées à un péage d'autoroutes à partir d'un instant initial t = 0. Dans ce cas l'espace d.états est E =N et chaque trajectoire est croissante.

  - On étudie l'évolution d'une population au cours du temps : celle-ci peut augmenter (naissance) ou diminuer (mort). Ici encore l'espace d'états est E

  Dans ces situations, on s'intéresse surtout à la convergence en loi de processus $(X(t))_{t\geq 0}$, c'est-à-dire qu'on veut savoir s'il existe une loi de probabilité $\pi = [\pi_0, \pi_1, ...]$ sur E telle que :
  $\forall i \in E; \mathrm{P}(\mathrm{X_t} = i) \underset{t\to+\infty}{\to} \pi_i$

  Dans la figure 1, on observe quelques exemples de trajectoires de processus aléatoires.

File d'attente

Introduction

Une file d'attente est constituée de clients qui arrivent de l'extérieur pour rejoindre cette file, de guichets où les clients vont se faire servir par des serveurs.

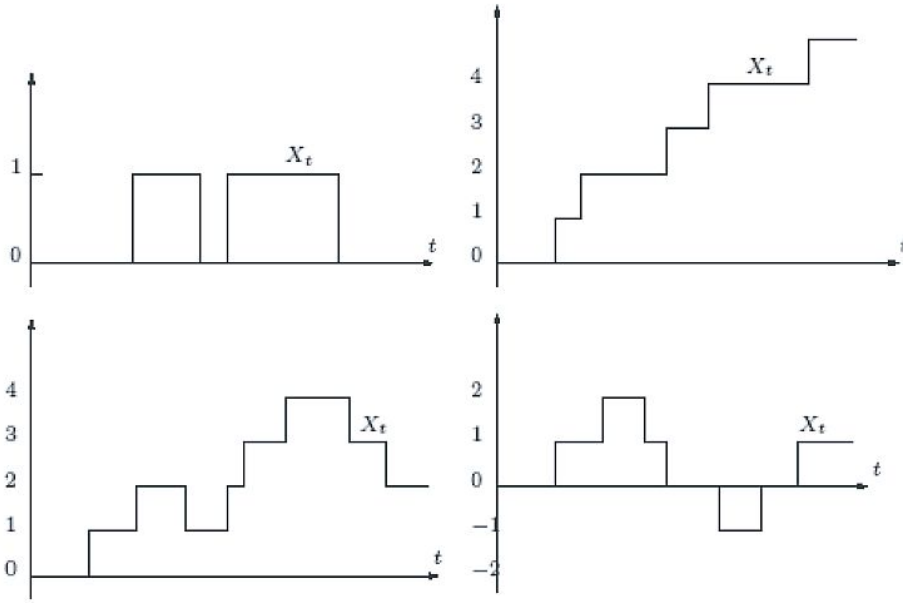


FIG. 3.1: Exemples des trajectoires de processus aléatoires

Dans certains cas les clients attendent dans une salle d'attente de capacité limitée. Un client servi disparaît. Les instants d'arrivée des clients et les temps de service sont aléatoires. On suppose que le premier arrivé est le premier servi (discipline FIFO : First In First Out).La théorie de ces files s'est développée pour la modélisation des centraux téléphoniques : un central recueille tous les appels d'une zone géographique donnée et les met en relation avec les correspondants. Les caisses d'un hypermarché donnent un exemple déjà assez compliqué de file d'attente. Un autre exemple important est donné par la file à l'entrée d'un élément d'un système informatique (Unité centrale CPU, imprimante, ...) lorsque les travaux qui arrivent se mettent en attente avant d'être traités par cet élément. Une file d'attente est décrite par la loi d'interarrivées des clients, la loi des temps de service, le nombre de serveurs, la taille maximale. La taille du système à un instant donné est le nombre de clients en train d'être servis et d'attendre. Nous supposerons toujours que les interarrivées sont des variables aléatoires indépendantes et de même loi, indépendantes des temps de service, eux mêmes indépendants et de même loi.
Pour les files simples, on utilise **les notations de Kendall :**
Loi d'interarrivée / Loi de service / Nombre de serveurs / Taille max.
Les lois sont notées symboliquement : M lorsqu'elles sont exponentielles (M pour Markov), G (G pour Général) sinon.
La file M/M/1
On considère une file M/M/1, donc d'après la notation de Kendall donnée en

haut, il s'agit d'une file à un serveur. Les clients arrivent à des instants successifs selon un processus de Poisson de paramètre λ . Ils se mettent en file d'attente et sont servis selon leur ordre d'arrivée. Le temps de service pour chaque client est une variable aléatoire de loi exponentielle de paramètre μ. Toutes les variables aléatoires qui interviennent sont indépendantes. On considère le processus $(X(t))_{t\geq 0}$ qui représente le nombre de clients en attente (y compris le client en train d'être servi) au temps t . C'est un processus de sauts à valeurs dans N. Quand un client arrive, le processus saute de +1 et quand un client s'en va à la fin de son service, le processus saute de -1. Si à un certain instant le processus saute et prend la valeur i (i $>$ 0), il va rester en i un temps aléatoire qui vaut inf (U_1, U_2) où U_1 est le temps nécessaire pour l'arrivée du prochain client et U_2 est le temps nécessaire pour la fin du service en cours. Or ces variables aléatoires sont indépendantes de lois exponentielles de paramètre λ et μ : Le temps de séjour dans l'état i sera donc une variable aléatoire de loi exponentielle de paramètre $\lambda + \mu$. La probabilité que le saut suivant soit de +1 est la probabilité que U_1 soit plus petite que U_2, c'est-à-dire, $\frac{\lambda}{\lambda+\mu}$, tandis que la probabilité pour que le saut soit de -1 vaut de la même manière $\frac{\mu}{\lambda+\mu}$. Si à un certain instant le processus saute et prend la valeur 0, il va rester en 0 un temps aléatoire qui vaut U_1 où U_1 est le temps nécessaire pour l'arrivée du prochain client, c'est-à-dire un temps aléatoire de loi exponentielle de paramètre λ. Le saut à l'issue de ce temps est nécessairement de +1.
On a bien la description d'un processus markovien de sauts.
Soit $\rho = \frac{\lambda}{\mu}$ l'intensité de trafic.
Ce rapport définit le taux d'utilisation ou d'occupation d'un serveur.
La distribution stationnaire du nombre de clients dans la file M/M/1 est donc géométrique.
La probabilité stationnaire d'avoir n clients dans le système (file d'attente + service) est donnée par l'expression suivante :
$P_n = (1-\rho)\rho^n, \forall t$
Soit L la moyenne de la distribution géométrique.
L représente ainsi le nombre moyen de clients dans le système.
$L = E(n) = \sum_{n=0}^{+\infty} nP_n = \sum_{n=0}^{+\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$
La file M/M/1 est décrite par un processus Markovien de saut.
Soit X_t le nombre total de clients dans le système à l'instant t, c'est-à-dire le nombre de clients dans la file plus le nombre de clients en train d'être servis.
Le processus X_t,$t \in IR^+$ à valeurs dans N est un processus de Markov de générateur :

$A(n, n+1) = \lambda, A(n, n-1) = \mu$ si n > 0

et A(n, m) = 0si $|$ n $-$ m$| \geq 2$

Analyse du comportement de la file M/M/1

On se propose d'utiliser une fonction qui permet de simuler un système avec file d'attente M/M/1. L'arrivée du système est déterminée par un processus de poisson d'intensité μ . Le temps de service poisson est d'intensité λ .

- Les entrées du système sont : n (nombre de sauts), λ (intensité d'arrivée) et μ (intensité de temps de service),
- Les sorties du système sont : le temps de saut cumulative ainsi que la longueur de système.

Pour étudier l'évolution du système, on définit Z_n la chaîne de matrice P associée au processus de saut X_t (nombre aléatoire de clients dans le système), et soient U_i des variables indépendantes équidistribuées à valeurs -1,1 telles que : $P\{U_i = 1\} = \frac{\lambda}{\lambda+\mu}$ et $P\{U_i = -1\} = \frac{\mu}{\lambda+\mu}$
La forme récurrente : $Z_{n+1} = |Z_n + U_{n+1}|$ définit une chaîne de Markov de matrice de transition P.

Trois cas peuvent se présenter.

- 1^{er} **cas :**$\lambda < \mu$
  Dans la figure 2, le débit moyen d'arrivée est strictement inférieur au débit maximum de service. Le nombre aléatoire X_t de personnes dans le système est un processus récurrent positif : il prend en presque sûr (p.s) une infinité de fois toute valeur arbitrairement grande et une infinité de fois la valeur zéro. Un régime stationnaire s'établit. Ainsi, en régime stationnaire, la probabilité de trouver n clients dans le serveur est P_n. L'expression régime stationnaire signifie qu'on a attendu assez longtemps pour oublier la situation initiale (file vide par exemple). On a donc convergence en loi $(X(t))_{t\geq 0}$ de vers une variable aléatoire X.
- 2^{me} **cas :**$\lambda > \mu$
  Dans la figure 3, la file sature. Le nombre de clients dans la file tend presque sûrement vers l'infini. C'est le cas transitoire.
- $3^{me}cas$ **:**$\lambda = \mu$
  Dans la figure 4, le nombre aléatoire X_t de personnes dans le système est un processus récurrent nul. La file se vide régulièrement mais le temps moyen entre deux retours à zéro est infini et il n'y a pas de convergence en loi du nombre de clients dans la file. Aucun régime stationnaire ne s'établit.

3.10 Processus de Wiener (ou mouvement brownien)

Introduction

Le mouvement brownien est le plus célèbre et le plus important des processus stochastiques et ceci pour plusieurs raisons. Historiquement, la découverte du processus qui s'est faite durant la période 1900–1930 et à laquelle sont attachés les noms de Bachelier, Einstein, Wiener et Kolmogorov, fut le premier cas où

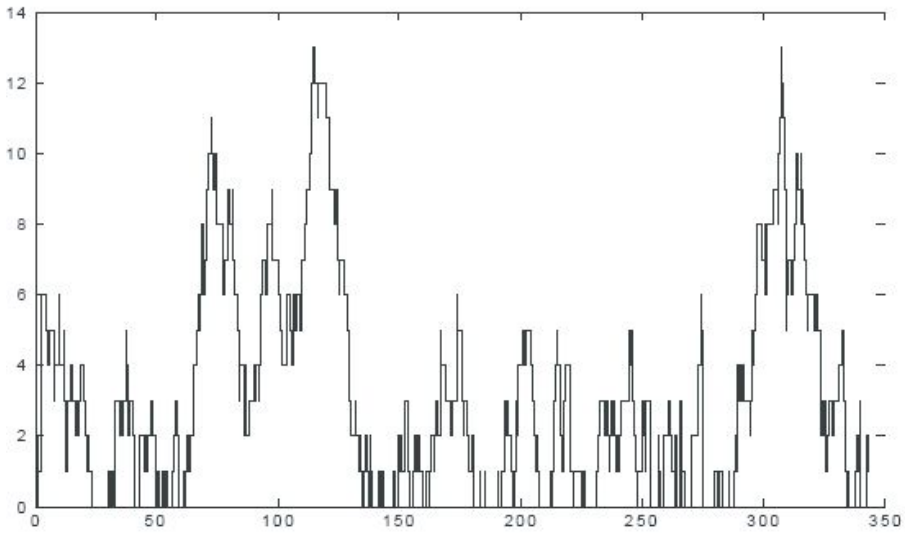


Fig. 3.2: Simulation pour λ=0.8 et μ=1

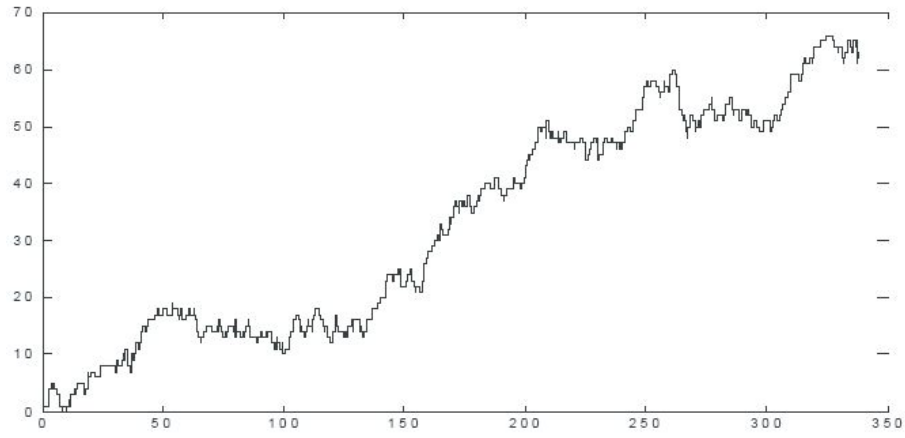


Fig. 3.3: Simulation pour λ=0.9 et μ=0.6

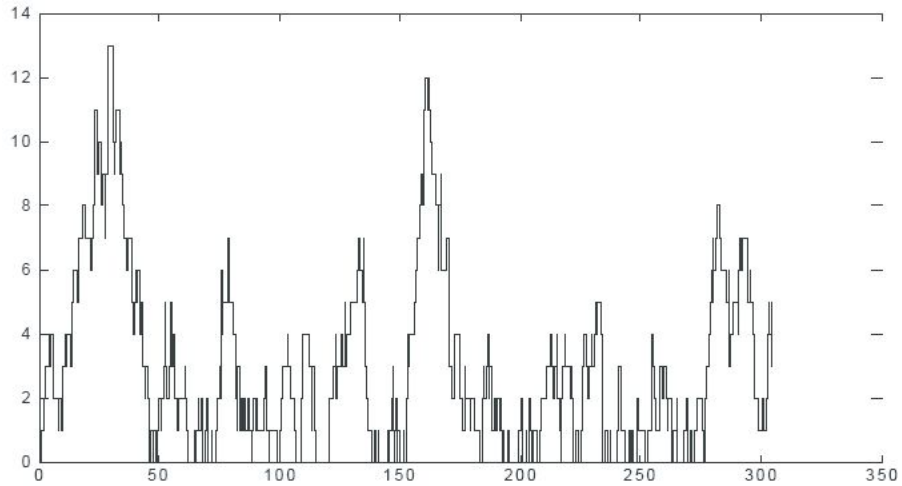


FIG. 3.4: Simulation pour λ=0.9 et μ=0.9

le calcul des probabilités s'appliquait à la description d'un phénomène physique indépendamment de tout jeu de hasard, à savoir la description du mouvement d'une petite particule dans un liquide ou gaz soumise à des chocs moléculaires dus à l'agitation thermique. Le champ d'application du mouvement brownien est beaucoup plus vaste que l'étude des particules microscopiques en suspension et inclut la modélisation du prix des actions boursières, du bruit thermique dans les circuits électroniques, du comportement limite des problèmes de files d'attente et des perturbations aléatoires dans un grand nombre de systèmes physiques, biologiques ou économiques.

Propriétés du Processus de Wiener

Soit le processus de Wiener normalisé W(t), défini sur (Ω, A, P), indexé sur R^+ à valeurs dans R_d. W(t) est un processus gaussien, du second ordre, centré, à trajectoires continus p.s. De plus, il vérifie les propriétés suivantes :

- Les processus $(W_1, ..., W_d)$ sont indépendants dans leur ensemble
- W(0)=0 p.s
- Pour tout $0 \leq t \leq t^{'}$, l'accroissement $\Delta W_{t,t'} = W_{t'} - W_t$ est une variable aléatoire gaussienne,centrée, de covariance $R_{\Delta W_{t,t'}} = E(\Delta W_{t,t'} \Delta W_{t,t'}^T)$
- Ses accroissements $\Delta W_{t,t'}$ sont indépendants et stationnaires Par ailleurs, la dérivée au sens des processus généralisé du processus de Wiener normalisé est le bruit blanc gaussien normalisé $\dot{W} = N_\infty$
  En conclusion, ce processus est un objet mathématique remarquable sur lequel s'accumulent un nombre considérable de propriétés :

- C'est le seul processus (à un changement de repère près) à accroissement

indépendants stationnaires à trajectoires continues,
- Il est relié à la fois à la théorie des processus stationnaires, au bruit blanc et à la théorie des fonctions harmoniques appelées aussi théorie du potentiel (fonctions réelles sur R_d telle que $\Delta f = 0$),
- Tous les processus de Markov à trajectoires continues peuvent se représenter à partir du mouvement brownien, c'est la théorie des équations différentielles stochastiques.

La figure 5 décrit l'évolution de la trajectoire du mouvement brownien.

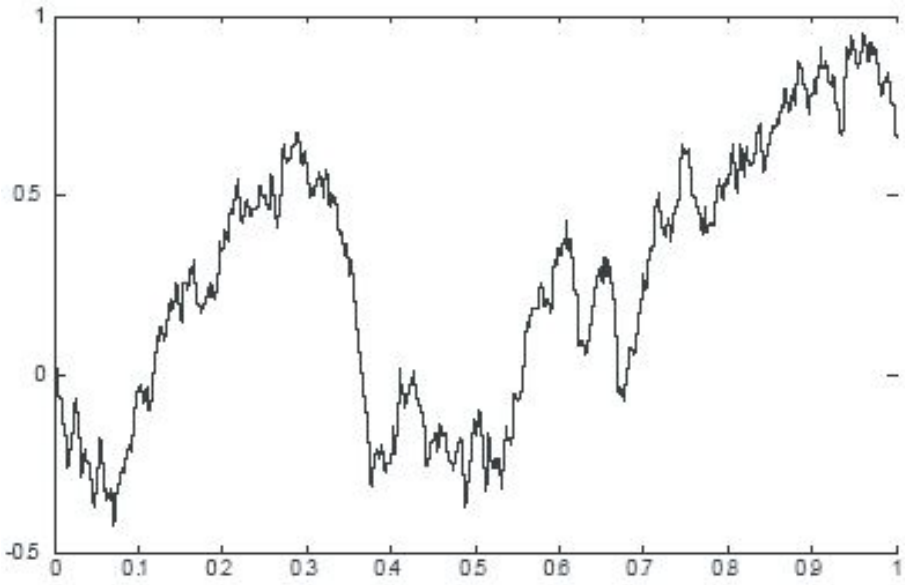


Fig. 3.5: Trajectoire d'un mouvement brownien

3.11 Problèmes et exercices pour l'Ingénieur

Espérance mathématique, variance

Exercice 3.11.1. On effectue la mesure d'une grandeur x à l'aide de deux capteurs différents. Soient $z_1$et z_2 les mesures obtenues. On admet qu'elles sont respectivement polluées par des bruits additifs $v_1$et v_2 supposés indépendants de moyennes nulles et de variances respectives σ_1^2 et σ_2^2. Faire la synthèse d'un estimateur optimal de la grandeur x.

Solution :
On écrit :

$$z_1 = x + v_1$$

et

$$z_2 = x + v_2.$$

Si $\widehat{x}$ est la valeur estimée de x, on pose

$$\widehat{x} = az_1 + bz_2.$$

a) On exprime que l'estimateur est sans biais soit :

$$E[\widehat{x}] = x,$$

d'où :

$$E[az_1 + bz_2] = aE[z_1] + bE[z_2] = ax + bx = (a+b)\,x \equiv x.$$

Soit :

$$a + b = 1 \Longrightarrow a = 1 - b\ .$$

et

$$\widehat{x} = (1-b)z_1 + bz_2.$$

a) On minimise e la variance de l'erreur d'estimation.

Soit e l'erreur d'estimation, on a

$$e = x - \widehat{x},$$

d'où la variance de e :

$$J = E[e^2] \Longrightarrow J = E[(x - \widehat{x})^2]$$

Calculons l'erreur d'estimation :

$$e = x - \widehat{x} = x - (1-b)z_1 - bz_2 = x - (1-b)\,(x + v_1) - b\,(x + v_2)\,,$$

d'où :

$$e = bv_2 - (1-b)\,v_1.$$

Il vient :

$$J = E\left[(bv_2 - (1-b)v_1)^2\right] \Longrightarrow J = (1-b)^2\sigma_1^2 + b^2\sigma_2^2.$$

$$\min_b J \Longleftrightarrow \tfrac{dJ}{db} = 0$$

d'où :

$$b = \tfrac{\sigma_1^2}{\sigma_1^2+\sigma_2^2};$$

on en déduit :

$$a = \tfrac{\sigma_2^2}{\sigma_1^2+\sigma_2^2};$$

et enfin :

$$\widehat{x} = \tfrac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} z_1 + \tfrac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} z_2$$

On peut calculer σ_e^2 la variance de l'erreur d'estimation :

$$\sigma_e^2 = J_{\min} = \tfrac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}.$$

On voit que σ_e est inférieur à σ_1 et à σ_2.

Remarque :

Ce résultat peut aussi être obtenu en utilisant la méthode des moindres carrés moyennant une normalisation des variances σ_1^2 et σ_2^2.

On obtient :

$$\tfrac{1}{\sigma_1} z_1 = \tfrac{1}{\sigma_1} x + e_1$$

$$\tfrac{1}{\sigma_2} z_2 = \tfrac{1}{\sigma_2} x + e_2$$

avec :

$$e_1 = \tfrac{v_1}{\sigma_1}$$

et

$$e_2 = \frac{v_2}{\sigma_2};$$

d'où

$$\sigma_{e_1} = \sigma_{e_2} = 1.$$

On pose

$$Y = \begin{pmatrix} \frac{z_1}{\sigma_1} \\ \frac{z_2}{\sigma_2} \end{pmatrix}, \quad H = \begin{pmatrix} \frac{1}{\sigma_1} \\ \frac{1}{\sigma_2} \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \text{ et } \theta = x.$$

d'où le modèle :

$$Y = H\theta + e.$$

et la solution :

$$\widehat{\theta} = (H^T H)^{-1} H^T Y = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} z_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} z_2.$$

Probabilité conditionnelle, maximum de vraisemblance, espérance mathématique conditionnelle

Exercice 3.11.2. Soit le modèle linéaire $Y = H\theta + e$, θ avec vecteur de paramètres à estimer, Y vecteur des observations, H une matrice déterministe connue et e vecteur de bruit supposé gaussien de moyenne nulle et de matrice de covariance $E\left[ee^T\right] = \sigma^2 I$, I étant la matrice unité. On considère la variable aléatoire Y/θ et la probalité conditionnelle $p(Y/\theta)$. Calculer la valeur de θ qui rend cette probabilté maximale. Etudier la statistique de cet estimateur.

Réponse

Le vecteur e étant gaussien, il s'en suit que Y/θ est aussi gaussien. Calculons sa moyenne et sa matrise de covariance P.

$$E\left[Y/\theta\right] = E\left[\left(H\theta + e\right)/\theta\right] = H\theta + E\left[e/\theta\right] = H\theta.$$
$$P = E\left[\left(Y - H\theta\right)\left(Y - H\theta\right)^T/\theta\right] = E\left[ee^T\right] = \sigma^2 I.$$

Y/θ étant gaussien sa densité de probabilité est donnée par :

$$p(Y/\theta) = Ke^{\left(-\frac{1}{2}\left((Y-H\theta)^T P^{-1}(Y-H\theta)\right)\right)} = Ke^{\left(-\frac{1}{2\sigma^2}\left((Y-H\theta)^T(Y-H\theta)\right)\right)}.$$
$$\underset{\theta}{Maxp}\left(Y/\theta\right) \iff \frac{\partial p(Y/\theta)}{\partial\theta} = 0 \iff \frac{\partial}{\partial\theta}\left(\ln\left(p\left(Y/\theta\right)\right)\right) = 0$$

d'où :

$$\frac{\partial}{\partial\theta}\left(\left(Y - H\theta\right)^T\left(Y - H\theta\right)\right) = 0$$

On obtient :

$$\widehat{\theta} = (H^T H)^{-1} H^T Y.$$

L'erreur d'estimation est donnée par :

$$\widetilde{e} = \theta - \widehat{\theta}.$$

Soit :

$$\widetilde{e} = \theta - (H^T H)^{-1} H^T Y = \theta - (H^T H)^{-1} H^T\left(H\theta + e\right) = (H^T H)^{-1} H^T e.$$

On calcule la moyenne de l'erreur d'estimation :

$$E\left[\widetilde{e}\right] = (H^T H)^{-1} H^T E\left[e\right] = 0.$$

Puis sa variance :

$$E\left[\widetilde{e}\,\widetilde{e}^T\right] = E\left[\left((H^TH)^{-1}H^Te\right)\left((H^TH)^{-1}H^Te\right)^T\right]$$
$$= (H^TH)^{-1}H^TE\left[ee^T\right]H(H^TH)^{-1},$$

D'où :

$$E\left[\widetilde{e}\,\widetilde{e}^T\right] = \sigma^2(H^TH)^{-1}.$$

Estimation d'un paramètre

Exercice 3.11.3. On procède à la mesure de la concentration z en sel d'un lac durant l'été (en absence de pluie). Une mesure est effectuée tous les jours à midi. Les mesures sont entachées de bruit v supposé gaussien ayant ses composantes indépendantes entre elles, une moyenne nulle et une variance σ^2. On donne la relation $z = z_0e^{\alpha t} + v$ où t est le temps de mesure exprimé en jours et α une constante positive assez faible.

Journée	6	7	8	9	10
$z(g/l)$	5.0457	5.0543	5.0629	5.0716	5.0802

1. On suppose α connu égal à 0.001 et on demande :
   1. 1. Estimer la valeur de la concentration initiale z_0.
   1. 2. Donner une estimation de σ^2.
   1. 3. Calculer la variance de l'erreur d'estimation.
2. En fait α est inconnu, proposer un estimateur de z_0 et de α.

Estimation de la variance de bruit

Exercice 3.11.4. On admet les hypothèses de l'exercice 2 et on demande de proposer un estimateur de la variance de bruit puis de calculer dans ces conditions la variance de l'erreur d'estimation.

Solution

Rappelons que dans l'exercice 2, nous avons établi l'expression de la variance de $\widetilde{e}$ (erreur d'estimation) :

$$E\left[\widetilde{e}\,\widetilde{e}^T\right] = \sigma^2(H^TH)^{-1}.$$

Cette expression s'exprime en fonction de σ^2, variance du bruit, souvent inconnue. Dans cet excercice il est proposé d'estimer cette variance.

Pour cela, considérons le terme

$$M = Y - H\widehat{\theta}$$

appelé terme résiduel :

$$M = Y - H\widehat{\theta} = (H\theta + e) - H\left(\left(H^TH\right)^{-1}H^TY\right)$$

d'où

$$M = H\theta + e - H\left(H^TH\right)^{-1}H^T(H\theta + e) = e - H\left(H^TH\right)^{-1}H^Te,$$

soit :

$$M = \left(I - H\left(H^T H\right)^{-1} H^T\right) e = Fe.$$

Calculons :

$$E\left[\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)\right] = E\left[(Fe)^T (Fe)\right] = E\left[e^T F^T Fe\right].$$

Remarquons au passage que la matrice F est symétrique et idempotente, d'où :

$$E\left[\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)\right] = E\left[e^T Fe\right] = E\left[e^T \left(I - H\left(H^T H\right)^{-1} H^T\right) e\right];$$

Soit :

$$E\left[\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)\right] = E\left[e^T e\right] - E\left[e^T H\left(H^T H\right)^{-1} H^T e\right].$$

Soit N la dimension du vecteur e, il vient :

$$E\left[\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)\right] = N\sigma^2 - E\left[e^T H\left(H^T H\right)^{-1} H^T e\right].$$

Le calcul du deuxième terme de l'expression précédente peut se faire en remarquant que l'expression $G = e^T H\left(H^T H\right)^{-1} H^T e$

est un scalaire et par suite :

$$E[G] = E\left[e^T H\left(H^T H\right)^{-1} H^T e\right] = E\left[trace\left(e^T H\left(H^T H\right)^{-1} H^T e\right)\right].$$

Or, on sait que si les produits AB et BA existent :

$$trace(AB) = trace(BA)$$

il vient :

$$trace\left(e^T H\left(H^T H\right)^{-1} H^T e\right) = trace\left(\left(H^T H\right)^{-1} H^T ee^T H\right).$$

D'où :

$$E[G] = E\left[trace\left(\left(H^T H\right)^{-1} H^T ee^T H\right)\right] = trace\left(E\left[\left(H^T H\right)^{-1} H^T ee^T H\right]\right)$$

Soit :

$$\begin{aligned} E[G] &= trace\left(\left(H^T H\right)^{-1} H^T E\left[ee^T\right] H\right) \\ &= trace\left(\left(H^T H\right)^{-1} H^T \sigma^2 I_N H\right) = n\sigma^2. \end{aligned}$$

n étant le nombre de colonne de H, c'est à dire le nombre de paramètres à estimer.

Il vient :

$$E\left[\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)\right] = (N - n)\,\sigma^2.$$

En définitive, nous pouvons choisir l'expression suivante comme estimateur de la variance du bruit :

$$\widehat{\sigma^2} = \frac{\left(Y - H\widehat{\theta}\right)^T \left(Y - H\widehat{\theta}\right)}{N - n}.$$

Processus gaussien markovien, expérience mathématique, Matrice de covariance

Exercice 3.11.5. Soit l'équation d'état $X_{k+1} = A_k\, X_k + \Gamma_k v_k$ et la condition initiale $X(0) = X_0$.Avec X_k l'état, u(k) la commande supposée déterministe, v_k un bruit blanc gaussien défini par $E[v_k] = 0$ et $E[v_i v_j] = Q\delta_{ij}$. X_0 l'état initial

est une variable aléatoire gaussienne donnée par $E[X_0] = m_0$ et $cov(X_0) = P_0$. On suppose que $E[X_0 v_k^T] = 0 \ \ \forall k$.

Q1 : Montrer que X_k est un processus gaussien markovien.

Q2 : Calculer l'éspérance mathmatique de X_{k+1}.

Q3 : Montrer que la matrice de covariance de X_{k+1} donnée par $P_{k+1} = A_k P_k A_k^T + \Gamma_k Q \Gamma_k^T$.

Q4 : Calculer $C(k+n, k)$ la matrice de covariance entre deux instants t_k et t_{k+n}.

Réponse :

Q1

On a :

$$X_k = A_{k-1}\ X_{k-1} + \Gamma_{k-1} v_{k-1},$$

soit en remplaçant X_k par son expression en fonction de X_{k-1} :

$$\begin{aligned} X_k &= A_{k-1}(A_{k-2}\ X_{k-2} + \Gamma_{k-2} v_{k-2}) + \Gamma_{k-1} v_{k-1} \\ &= A_{k-1} A_{k-2}\ X_{k-2} + A_{k-1}\Gamma_{k-2} v_{k-2} + \Gamma_{k-1} v_{k-1}, \end{aligned}$$

et en poursuivant le même processus :

$$X_k = X_0 + A_{k-1}....A_1 \Gamma_0 v_0 + A_{k-1}....A_2 \Gamma_1 v_1 + ...\Gamma_{k-1} v_{k-1}.$$

Si on pose :

$$A(k,i) = A_{k-1} A_{k-2} ... A_{k-i}$$

avec $A(k,k) = I$, il vient :

$$X_k = A(k,0) X_0 + A(k,1)\Gamma_0 v_0 + A(k,2)\Gamma_1 v_1 + ...\Gamma_{k-1} v_{k-1}.$$

D'où :

$$X_k = A(k,0) X_0 + \sum_{i=0}^{k-1} A(k, i+1)\Gamma_i v_i.$$

On en déduit que X_k est indépendant de v_k il en est de même pour X_{k-1}, X_{k-2}, ...

Dans ce cas, on peut écrire

$$P(X_{k+1}, X_k, X_{k-1}, ..., X_0) = P(X_{k+1}/X_k).$$

X_k est donc markovien car la connaissance de X_{k-1} suffit pour calculer X_k,

tout le passé est résumé dans l'état précédent.

X_k est aussi gaussien car il est ob tenu par une transformation linéaire de vecteurs aléatoires gaussiens.

Q2

On écrit :

$$E[X_{k+1}] = E[A_k X_k + \Gamma_k v_k] = A_k E[X_k] + \Gamma_k E[v_k] = A_k m_k\,;$$

On en déduit :

$$m_{k+1} = A_k m_k,$$

d'où :

$m_k = A_{k-1} m_{k-1} = A_{k-1}\,(A_{k-2} m_{k-2}) = A_{k-1}\,(A_{k-2}\,(A_{k-3}\,((...)))) = A(k,0) m_0$.

Q3

$$Cov(X_{k+1}) = P_{k+1} = E[(X_{k+1} - m_{k+1})(X_{k+1} - m_{k+1})^T];$$

Or :

$$X_{k+1} - m_{k+1} = A_k\,(X_k - m_k) + \Gamma_k v_k \Rightarrow$$
$$P_{k+1} = E[(A_k(X_m - m_k) + \Gamma_k v_k)((X_k - m_k)^T A_k^T + \Gamma_k^T v_k^T];$$

d'où :

$$P_{k+1} = A_k P_k A_k^T + \Gamma_k Q \Gamma_k^T$$

Q4

$$C(k+n,k) = E[(X_{k+n} - m_{k+n})(X_k - m_k)^T]$$

soit :

$$C(k+n,k) = E\Big[\big(A(k+n,k)(X_k - m_k) + \sum_{i=k}^{k+n-1} A(k+n,i+1)\Gamma_i v_i\big)(X_k - m_k)^T\Big]$$

D'où :

$$C(k+n,k) = A(k+n,k)P_k.$$

Un calcul analogue conduirait à :

$$C(k+n,k) = P_k A^T(k+n,k).$$

Espérance mathématique, Processus scalaire, Matrice de covariance, Processus stationnaire, Processus gaussien

Exercice 3.11.6. On considére le processus scalaire suivant $x_{k+1} = a\ x_k + v_k$ avec v_k bruit blanc gaussien tel que $E[v_k] = 0, E[v_i v_j] = q\delta_{ij}$ l'état initial x_0 est une variable aléatoire gaussienne tel que $E[x_0] = 0; E[x_0^2] = P_0$ et $E[x_0 v_k] = 0 \forall k$.

Q1 : Calculer la variance P_{k+1} de x_{k+1}en fonction de P_0, a et q.

Q2 : En déduire la limite quand k tend vers l'infini de P_{k+1}, on suppose que $|a| < 1$

Q3 : Montrer alors que le processus x_{k+1} est stationnaire.

Q4 : Comment choisir $E[x_0^2]$ pour que le processus devient stationnaire à l'instant initial.

Réponse : *Q1*

Calculons d'abord l'espérence mathématique de x_{k+1} :

$E[x_{k+1}] = E[a\ x_k + v_k] = aE[x_k] + 0 = aE[a\ x_{k-1} + v_{k-1}] = a^2 E[x_{k-1}] =$ ..... $= a^{k+1}E[x_0] = 0$.

D'où :

$$P_{k+1} = E[(x_{k+1} - E\,[x_{k+1}])^2] = E[x_{k+1}^2]$$

soit :

$$P_{k+1} = E[(a\ x_k + v_k)^2] = a^2 E[x_k^2] + E[v_k^2] = a^2\ P_k + q.$$

On en déduit (par exemple en utilisant la transformée en z) :

$$P_{k+1} = a^{2(k+1)} P_0 + \tfrac{q}{1-a^2}(1 - a^{2(k+1)})$$

Q2

Si $k \longrightarrow \infty$ et $|a| < 1$, ilvient :

$$P_{k+1} \longrightarrow P_\infty = \frac{q}{1-a^2}.$$

Q3

On vérifie

$E[x_{k+1}] = 0$;

et

$\lim_{k \longrightarrow \infty} E[x_{k+1}^2] = P_\infty = cst$,

lLe Processus est donc Stationnaire.

Q4 :

Il suffit de prendre

$$P_0 = \frac{q}{1-a^2}.$$

En effet, dans ce cas,

$$E[x_1^2] = E[(a\ x_0 + v_0)^2] = E\left[a^2 x_0 + 2a x_0 v_0 + v_0^2 = a^2 p_0 + q\right];$$

d'où :

$$E[x_1^2] = \frac{q}{1-a^2} \Longrightarrow E[x_i^2] = \frac{q}{1-a^2} \forall i.$$

Espérence mathématique, matrice de covariance, processus stationnaire

Exercice 3.11.7. On donne l'équation d'état $X_{k+1} = A_k\ X_k + \Gamma v_k$, même hypothèses que l'exercice 8 avec $A_k = A = cste$ et $\Gamma_k = cste$

Q1 : Calculer $m_k = E[x_k]$ ainsi que la matrice de covariance P_k de X_k en fonction de m_0 et P_0

Q2 : Quelles conditions doit-on imposer pour que le processus soit stationnaire

Réponse : *Q1 :*

$$m_k = E[x_k] = E[A\ X_{k-1} + \Gamma v_{k-1}] = A m_{k-1}$$
$$m_k = A m_{k-1} = A^2 m_{k-1} = ... = A^k m_0$$

$$\begin{aligned} P_k &= E[(x_k - m_k)(x_k - m_k)^T] \\ &= E[(A(x_{k-1} - m_{k-1}) + \Gamma v_k)(A(x_{k-1} - m_{k-1}) + \Gamma v_k)^T] \end{aligned}$$

$$P_k = A P_{k-1} A^T + \Gamma Q \Gamma^T$$
$$P_k = A^k P_0 (A^T)^k + \sum_{i=0}^{k-1} A^i \Gamma Q \Gamma^T (A^T)^i$$

Q2 :

Processus stationnaire impose m_k et P_k indépendants de m_0 et P_0 quand k tend vers l'infini.

Il vient alors :

1. $\lim_{k \longrightarrow \infty} A^k$ finie, $\Rightarrow A$ est stable (valeurs propres à l'intérieur du cercle unité).
2. $\lim_{k \longrightarrow \infty} P_k$ finie, $\Rightarrow P_\infty$ solution de $P_\infty = A P_\infty A^T + \Gamma Q \Gamma^T$

Variable aléatoire conditionnelle

Exercice 3.11.8. Soit le processus gaussien markovien : $X_{k+1} = A_k\ X_k + \Gamma_k v_k$.On admet les hypothéses de l'exercice 8.

Q1 : Calculer la moyenne de la variable aléatoire conditionnelle X_{k+1}/X_k.

Q2 : Calculer la matrice de covariance de cette variable conditionnelle.

Q3 : En déduire sa densité de probabilité $p\left(X_{k+1}/X_k\right)$.

Réponse : *Q1* :

$$E\left[X_{k+1}/X_k\right] = A_k\ X_k + \Gamma_k E[v_k] = A_k\ X_k.$$

Q2 :

$$E\left[\left(X_{k+1} - A_k X_k\right)\left(X_{k+1} - A_k X_k\right)^T / X_k\right] = E[\Gamma_k v_k v_k^T \Gamma_k^T / X_k] = \Gamma_k Q \Gamma_k^T.$$

Q3 :

X_{k+1}/X_k étant gaussien, on peut écrire :

$$p\left(X_{k+1}/X_k\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\left|\det(\Gamma_k Q \Gamma_k^T)\right|}} \exp(\frac{-1}{2}\left(X_{k+1} - A_k X_k\right)^T \left(\Gamma_k Q \Gamma_k^T\right)^{-1} \left(X_{k+1} - A_k X_k\right)).$$

Vecteur aléatoire gaussien, transformation linéaire de v.a.g.

Exercice 3.11.9. Soit X une variable aléatoire gaussienne , et Z obtenue par la transformation linéaire Z=AX+b, avec A et b déterministe. Montrer que Z est une variable aléatoire gaussienne.

Réponse : On écrit la fonction caractéristique de Z :

$$\varphi_Z(\nu) = E\left[e^{jv^T Z}\right] = E\left[e^{jv^T(AX+b)}\right] = e^{jv^T b} E\left[e^{jv^T AX}\right].$$

or X est une variable aléatoire gaussienne, sa fonction caractéristique s'écrit :

$$\varphi_X(u) = E\left[e^{ju^T X}\right] = e^{\left(ju^T E[X] - \frac{1}{2}u^T P u\right)},$$

avec P est la matrice de covariance de X.

On en déduit :

$$\varphi_Z(v) = e^{jv^T b}\varphi_X(A^T v) = e^{jv^T b} e^{\left(jv^T AE[X] - \frac{1}{2}v^T APA^T v\right)};$$

soit :

$$\varphi_Z(v) = e^{\left(jv^T(AE[X]+b) - \frac{1}{2}v^T(APA^T)v\right)}.$$

On en déduit que Z est une variable aléatoire gaussienne de moyenne

$$E[Z] = AE[X] + b$$

et de matrice de covariance

$$P_{ZZ} = APA^T.$$

Partition d'une v.a.g

Exercice 3.11.10. Soit J une variable aléatoire gaussienne et X, Y une partition sur J : $J = \begin{pmatrix} X \\ Y \end{pmatrix}$ montrer que X et$Y$ sont gaussiens.

Réponse : On peut écrire $X = (I : 0)J = AJ$ et $Y = (I : 0)J = BJ$.

X et Y s'écrivent comme combinaison linéaire de J, qui est une variable aléatoire gaussienne, X et$Y$ sont gaussiens.

Espérance mathématique conditionnelle, variable aléatoire gaussienne, matrice de covariance

Exercice 3.11.11. Soient X et Y deux vecteurs aléatoires gaussiens, on pose $W_1 = X + MY$ et $W_2 = Y$ avec $M = -P_{XY}P_{YY}^{-1}$ où P_{XY} et P_{YY} sont respectivement les matrices de covariances de X, Y et de Y. On forme le vecteur $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$.

Q1 : Montrer que W_1 et W_2 sont indépendants.

Q2 : Calculer la matrice de covariance de W.

Q3 : Exprimer X en fonction de W_1 et Y et en déduire la moyenne ainsi que la variance de la variable conditionnelle X/Y.

Q4 : Déduire que $E[X/Y]$ est une variable aléatoire gaussienne, fonction linéaire de Y et que $E[X/Y]$ est indépendante de Y .

Réponse : *Q1* :

$$P_{W_1W_2} = E\left[(W_1 - E[W_1])\,(W_2 - E[W_2])^T\right]$$

Soit :

$$\begin{aligned} P_{W_1W_2} &= E\left[(X + MY - E\left[X + MY\right])\,(Y - E[Y])^T\right] \\ &= E\left[(X - E\left[X\right] + M(Y - E[Y]))\,(Y - E[Y])^T\right]. \end{aligned}$$

D'où :

$$P_{W_1W_2} = P_{XY} + MP_{YY} = P_{XY} - P_{XY}P_{YY}^{-1}P_{YY} = 0.$$

Q2 :

La matrice de covariance de $W = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix}$ est donnée par :

$$P_{WW} = \begin{pmatrix} P_{W_1W_1} & P_{W_1W_2} \\ P_{W_2W_1} & P_{W_2W_2} \end{pmatrix}.$$

Or

$$\begin{aligned} P_{W_1W_1} &= E\left[(W_1 - E[W_1])\,(W_1 - E[W_1])^T\right] \\ &= P_{XX} + P_{XY}M^T + MP_{YX} + MP_{YY}M^T; \end{aligned}$$

soit

$$P_{W_1W_1} = P_{XX} - P_{XY}P_{YY}^{-1}P_{YX}.$$

Puis :

$$P_{W_1W_2} = 0; P_{W_2W_1} = 0;$$

et

$$P_{W_2W_2} = P_{YY}$$

D'où

$$P_{WW} = \begin{pmatrix} P_{XX} - P_{XY}P_{YY}^{-1}P_{YX} & 0 \\ 0 & P_{YY} \end{pmatrix}.$$

Q3 :

On a

$$W_1 = X + MY\,;$$

d'où :

$$X = W_1 - MY.$$

On en déduit :

$$E[X/Y] = E[W_1] - MY = E[W_1] + P_{XY}P_{YY}^{-1}Y.$$

Or :

$$E[W_1] = E[X + MY\ ] = E[X] + ME[Y] = E[X] - P_{XY}P_{YY}^{-1}E\,[Y]\,;$$

d'où :

$$E[X/Y] = E[X] + P_{XY}P_{YY}^{-1}(Y - E\,[Y]).$$

La variance de X/Y est donnée par :

$$P_{X/Y} = E\left[(X - E[X/Y])\,(X - E[X/Y])^T\,/Y\right].$$

Or :

$$X - E[X/Y] = X - E[X] - P_{XY}P_{YY}^{-1}(Y - E\,[Y]) = W_1 - E[W_1];$$

d'où

$$P_{X/Y} = P_{W_1W_1} = P_{XY} - P_{XY}P_{YY}^{-1}P_{YX}.$$

Q4 :

L'expression établie précédemment de $E[X/Y]$ montre cette dernière est obtenue par transformation linéaire de la variable aléatoire gaussienne Y.

Il s'en suit que X/Y est aussi une variable aléatoire gaussienne.

L'expression $X - E[X/Y] = W_1 - E[W_1]$ et l'indépendance entre W_1 et Y prouvent que $X - E[X]$ est indépendant de Y.

variable aléatoire conditionnelle, Estimation des paramétres dans un modéle statique, Variables gaussiennes, Probabilité conditionnelle, moyenne, matrice de covariance

Exercice 3.11.12. On considére le modéle linéaire $Y = H\theta + e$, avec Y un vecteur des observations, θ le vecteur des paramètres à estimer, e une variable aléatoire gaussienne caractérisée par $E[e] = 0$, $E\left[ee^T\right] = \sigma^2 I$ et H une matrice de transformation connue. On suppose que H est déterministe.

Q1 : Montrer que la variable $Y/\ \theta$ est gaussienne. Calculer sa moyenne et sa matrice de covariance.

Q2 : En déduire l'expression de la densité de probabilité conditionnelle $p(Y/\theta)$.

Q3 : Montrer que $p(Y/\theta)$ est maximale pour $\overset{\wedge}{\theta} = (H^T H)^{-1} H^T Y$.

Q4 : Montrer que cet estimateur n'est pas biaisé si H et e sont indépendants et E[e]=0.

Réponse : *Q1* :

$Y/\ \theta$ est une transformation linéaire de la variable aléatoire gaussienne e. Elle est aussi gaussienne.

$$E[Y/\theta] = E[H\theta/\theta] + E[e/\theta] = H\theta + E[e] = H\theta.$$
$$E\left[(Y - H\theta)(Y - H\theta)^T/\theta\right] = E\left[ee^T\right] = \sigma^2 I.$$

Q2 :

$Y/\ \theta$ est une variable aléatoire gaussienne, on en déduit :

$$p(Y/\theta) = K \exp(\tfrac{-1}{2}(Y - H\theta)^T \left(\sigma^2 I\right)^{-1} (Y - H\theta)).$$

Q3 :

$p(Y/\theta)$ est maximale si :

$$\tfrac{\partial p(Y/\theta)}{\partial \theta} = 0 \Longleftrightarrow \tfrac{\partial}{\partial \theta}\left(Log\left(p(Y/\theta)\right)\right) = 0$$

d'où :

$$\widehat{\theta} = (H^T H)^{-1} H^T Y.$$

Q4 :

$$E[\widehat{\theta}] = E[(H^T H)^{-1} H^T Y] = E[(H^T H)^{-1} H^T (H\theta + e)]$$

d'où :

$$E[\widehat{\theta}] = E[\theta + (H^T H)^{-1} H^T e] = E[\theta] + (H^T H)^{-1} H^T E[e]$$

D'où

$$E[\overset{\wedge}{\theta}] = \theta\,;$$

l'estimateur est donc sans biais.

Puissance spectrale, Extration de signal

Exercice 3.11.13. Un signal inconnu $f(t)$ est corrompu par un bruit additif $v(t)$. On mesure $x(t)$ et on a la relation $x(t) = f(t) + v(t)$. On admet que $f(t)$ et $v(t)$ sont stationnaires non corrélés et de puissance spectrale connue. On cherche

à réaliser un filtre linéaire dont la sortie $\widehat{f(t)}$ est une estimée de f(t) au sens du critère $J = E\left[\left(f(t) - \widehat{f(t)}\right)^2\right]$.

Solution :

Soit $h(t)$ la réponse impulsionnelle du filtre linéaire on a :

$$\widehat{f(t)} = \int_{-\infty}^{+\infty} x(t-\alpha)h(\alpha)d\alpha.$$

Le critère s'écrit alors sous la forme :

$$J = E\left[\left(f(t) - \int_{-\infty}^{+\infty} x(t-\alpha)h(\alpha)d\alpha\right)^2\right].$$

Il est minimal si :

$$\frac{\partial J}{\partial h} = 0$$

d'où :

$$E\left[\left(f(t) - \int_{-\infty}^{+\infty} x(t-\alpha)h(\alpha)d\alpha\right)x(t-\tau)\right] = 0.$$

On notera R_{fx} la fonction d'intercorrélation de $f(t)$ et $x(t)$:

$$R_{fx}(\tau) = E\left[f(t)x(t-\tau)\right],$$

et R_{xx} la fonction d'autocorrélation de $x(t)$:

$$R_{xx}(\tau) = E\left[x(t)x(t-\tau)\right].$$

Il s'en suit :

$$J = R_{fx}(\tau) - \int_{-\infty}^{+\infty} R_{xx}(\tau-\alpha)h(\alpha)d\alpha = 0.$$

Soit en passant à la transformée de Fourier :

$$S_{fx}(\omega) - S_{xx}(\omega)H(\omega) = 0.$$

D'où :

$$H(\omega) = \frac{S_{fx}(\omega)}{S_{xx}(\omega)}.$$

Or $f(t)$ et $v(t)$ sont orthogonaux :

$$S_{xx}(\omega) = S_{ff}(\omega) + S_{vv}(\omega);$$

d'où :

$$H(\omega) = \frac{S_{fx}(\omega)}{S_{ff}(\omega)+S_{vv}(\omega)}.$$

Intégrale stochastique, processus de Wiener

Exercice 3.11.14. On donne le processus de Wiener défini par l'équation stochastique $dx = A(t)xdt + G(t)dv(t)$, on demande de calculer la moyenne et la matrice de covariance de x(t). dv(t) est un accroissement infinitésimal du mouvement brownien, ou processus de Wiener. Envisager le cas stationnaire.

Réponse :

On doit rappeler les propriétés de l'intégrale stochastique $\int f(t)dv(t)$.

Si $\int E\left[|f(t,\omega)|^2\right] < \infty$ et $E[dv] = 0$ et $E\left[(dv)^2\right] = \sigma^2 dt$ alors :

1. $E\left[\int_a^b f(t,\omega)\,dv(t)\right] = 0.$
2. $E\left[\left(\int_a^b f_1(t,\omega)\,dv(t)\right)\left(\int_a^b f_2(t,\omega)\,dv(t)\right)^T\right] = \sigma^2 \int_a^b E\left[f_1(t,\omega)\,f_2(t,\omega)\,dt\right].$

L' intégration l'équation stochastique conduit à l'expression suivante qui contient une intégrale stochastique :

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu).$$

$\Phi(t, t_0)$ est la matrice de transition solution de l'équation différentielle matricielle :

$$\frac{d\Phi(t, t_0)}{dt} = A(t)\Phi(t, t_0).$$

Revenons à l'exercice et calculons l'expérence mathématique de $x(t)$.

$$m(t) = E\left[x(t)\right] = E\left[\Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu)\right]$$

Soit :

$$m(t) = \Phi(t, t_0)E\left[x(t_0)\right] + E\left[\int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu)\right].$$

Il vient alors :

$$m(t) = \Phi(t, t_0)m_0.$$

La matrice de covariance $P(t)$ est définie par :

$$P(t) = E\left[\left(x(t) - m(t)\right)\left(x(t) - m(t)\right)^T\right].$$

Le centrage de la variable aléatoire $x(t)$ donne :

$$x(t) - m(t) = \Phi(t, t_0)\left(x(t_0) - m_0\right) + \int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu);$$

d'où :

$$P(t) = E\Big[\left(\Phi(t, t_0)\left(x(t_0) - m_0\right) + \int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu)\right) \left(\Phi(t, t_0)\left(x(t_0) - m_0\right) + \int_{t_0}^{t} \Phi(t, \mu)G(\mu)dv(\mu)\right)^T\Big].$$

Soit :

$$P(t) = \Phi(t,t_0)P\left(t_0\right)\Phi^T(t,t_0)+ \\ E\left[\left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)dv(\mu)\right)\left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)dv(\mu)\right)^T\right].$$

Le deuxième terme se développe comme le montre la deuxième propiété de l'intégrale stochastique généralisée au cas vectoriel :

$$E\left[\left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)dv(\mu)\right)\left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)dv(\mu)\right)^T\right] = \\ \int_{t_0}^{t}\Phi(t,\mu)G(\mu)QG^T(\mu)\Phi^T(t,\mu)d\mu.$$

On obtient finalement :

$$P(t) = \Phi(t,t_0)P\left(t_0\right)\Phi^T(t,t_0) + \int_{t_0}^{t}\Phi(t,\mu)G(\mu)QG^T(\mu)\Phi^T(t,\mu)d\mu.$$

Si on dérive l'expression précédente, on obtient :

$$\frac{dP(t)}{dt} = A(t)\Phi(t,t_0)P\left(t_0\right)\Phi^T(t,t_0)+ \\ \Phi(t,t_0)P\left(t_0\right)\Phi^T(t,t_0)A^T(t) + G(t)QG^T(t)+ \\ A(t)\left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)QG^T(\mu)\Phi^T(t,\mu)d\mu\right)+ \\ \left(\int_{t_0}^{t}\Phi(t,\mu)G(\mu)QG^T(\mu)\Phi^T(t,\mu)d\mu\right)A^T(t).$$

Il vient :

$$\frac{dP(t)}{dt} = A(t)P(t) + P(t)A^T(t) + G(t)QG^T(t).$$

Dans le cas stationnaire :

$$A(t) = A$$

et

$$G(t) = G.$$

Il est facile détablir :

$$\Phi(t,t_0) = e^{A(t-t_0)}.$$
$$E\left[x(t)\right] = e^{A(t-t_0)}E\left[x\left(t_0\right)\right].$$

$$P(t) = e^{A(t-t_0)} P\,(t_0)\, e^{A^T(t-t_0)} + \int_{t_0}^{t} e^{A(t-\mu)} GQG^T e^{A^T(t-\mu)} d\mu.$$

$$\frac{dP(t)}{dt} = AP(t) + P(t)A^T + GQG^T.$$

Remarque :

Si A est stable (valeurs propres à parties réelles négatives),alors :

$\diamondsuit$la moyenne $m(t)$ tend vers zéro si $t \longrightarrow \infty$

$\diamondsuit$et la matrice de covariance $P(t)$ converge vers une valeur constante P_∞ solution de l'équation :

$$AP_\infty + P_\infty A^T + GQG^T = 0.$$

4 | EDO non-linéaire

Wilfrid Perruquetti[1]

[1]LAGIS & INRIA-ALIEN, Ecole Centrale de Lille, BP 48, 59651 Villeneuve d'Ascq cedex, France. *E-mail* : `Wilfrid.Perruquetti@ec-lille.fr`

4.1 Introduction

Le calcul infinitésimal (différentiel) a tout naturellement conduit les scientifiques à énoncer certaines lois de la physique en termes de relations entre, d'une part, des variables dépendantes d'une autre variable indépendante (en général le temps) et, d'autre part, leurs dérivées : il s'agit là d'**équations différentielles ordinaires** (en abrégé, **EDO**). L'un des précurseurs dans ce domaine fut Isaac Newton (1642-1727) qui, dans son mémoire de 1687 intitulé **Philosophiae naturalis principiae mathematica** (les lois mathématiques de la physique) écrit : ‹‹ Data aequatione quotcunque fluentes quantitae involvente fluxiones invenire et vice versa ›› (en langage moderne, il parle d'équations différentielles). Dès lors, de nombreux modèles de la physique ont étés énoncés par l'intermédiaire d'EDO (dont, au XVIIe siècle, les équations d'Euler-Lagrange pour les sytèmes mécaniques).

Nous donnons ci-dessous une liste non exhaustive de modèles par des EDO tirés de divers domaines des sciences pour l'ingénieur.

Biologie

Une boîte de Petri contient des bactéries qui se développent sur un substrat nutritif. En notant x le nombre de bactéries, un modèle simplifié, dit **modèle logistique**, est donné par :

$$\frac{dx}{dt} = ax(x_{\max} - x), \tag{4.1}$$

où a est une constante strictement positive et $x_{\max}$ est le nombre maximal de bactéries pouvant survivre dans la boîte de dimension donnée. En effet, lorsqu'il

y a peu de bactéries : $\dot{x} \sim ax$ (croissance exponentielle) et, lorsque x est proche de $x_{\max}$, la croissance est fortement ralentie puisque $\dot{x} \sim 0$. Une autre exemple célèbre, le modèle proie-prédateur de Volterra-Lotka, sera évoqué à l'exercice 4.6.12.

Chimie

Les différents bilans (de matière, thermodynamique) peuvent, sous leur forme simplifiée, s'exprimer par des EDO. On considère par exemple une cuve alimentée en produits chimiques A et B de concentrations respectives c_A et c_B par l'intermédiaire de deux pompes de débits volumiques respectifs u_1 et u_2. Dans cette cuve, un mélangeur homogénéise les deux produits, qui réagissent selon la réaction :

$$n_A A + n_B B \underset{k_2}{\overset{k_1}{\rightleftarrows}} n_C C,$$

où n_A, n_B et n_C sont les coefficients stochiométriques de chacun des composants. Le mélange est sous-tiré par un orifice de section s à la base de cette cuve de section $S = 1$ m^2. Le bilan de matière conduit, en utilisant la relation de Bernouilli, à :

$$S\frac{dh}{dt} = u_1 + u_2 - \sqrt{2sgh},$$

où h est la hauteur du mélange dans la cuve et g, l'accélération de la pesanteur (9.81 ms^{-2}). Les lois de la cinétique donnent la relation (sous l'hypothèse d'une cinétique de second ordre) :

$$v_{\text{cin}} = -k_1 c_A c_B + k_2 c_C^2.$$

Ainsi, les conservations de chacun des composants donnent :

$$\begin{aligned}
\frac{d(hc_A)}{dt} &= u_1 c_{A0} - \sqrt{2sgh}c_A - n_A v_{\text{cin}} h,\\
\frac{d(hc_B)}{dt} &= u_2 c_{B0} - \sqrt{2sgh}c_B - n_B v_{\text{cin}} h,\\
\frac{d(hc_C)}{dt} &= -\sqrt{2sgh}c_C + n_C v_{\text{cin}} h,
\end{aligned}$$

avec $c_{A0} = c_{A\ (\text{entrant})}$ et $c_{B0} = c_{B\ (\text{entrant})}$. En notant $x = (h, hc_A, hc_B, hc_C)^T$ le vecteur d'état, on obtient le modèle :

$$\left\{\begin{array}{l}
\dot{x}_1 = u_1 + u_2 - \sqrt{2sg}\sqrt{x_1},\\
\dot{x}_2 = u_1 c_{A0} - \sqrt{\frac{2sg}{x_1}}x_2 - n_A\frac{(-k_1 x_2 x_3 + k_2 x_4^2)}{x_1},\\
\dot{x}_3 = u_2 c_{B0} - \sqrt{\frac{2sg}{x_1}}x_3 - n_B\frac{(-k_1 x_2 x_3 + k_2 x_4^2)}{x_1},\\
\dot{x}_4 = -\sqrt{\frac{2sg}{x_1}}x_4 + n_C\frac{(-k_1 x_2 x_3 + k_2 x_4^2)}{x_1}.
\end{array}\right. \tag{4.2}$$

Electricité

On considère un système électrique constitué d'une résistance R, d'une inductance L et d'une capacité C montées en triangle. On note respectivement i_X et v_X le courant et la tension dans la branche comportant l'élément X. On suppose que les éléments L et C sont parfaitement linéaires et que l'élément résistif R obéit à la loi d'Ohm généralisée ($v_R = f(i_R)$). Les lois de Kirchoff permettent d'obtenir :

$$\begin{cases} L\frac{di_L}{dt} = v_L = v_C - f(i_L), \\ C\frac{dv_C}{dt} = i_C = -i_L. \end{cases} \tag{4.3}$$

Si dans cette EDO, dite **équation de Liénard**, on considère le cas particulier où $f(x) = (x^3 - 2\mu x)$, on abouti à l'**équation de Van der Pol** .

Electronique

Le montage ‹‹ PI ›› (proportionnel-plus-intégral) suivant :

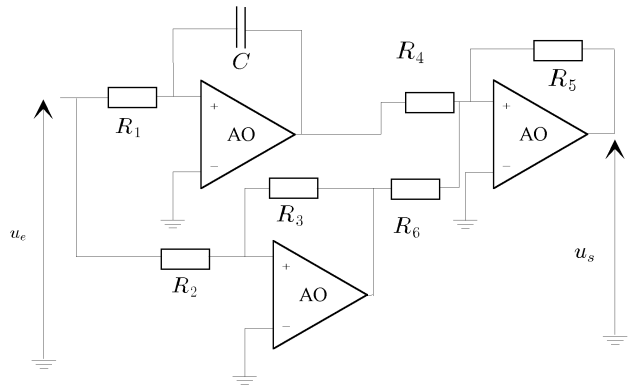


FIG. 4.1: Montage PI avec amplificateurs opérationnels.

a pour modèle, sous des hypothèses simplificatrices de linéarité :

$$T_i\frac{du_s}{dt} = k(u_e + T_i\frac{du_e}{dt}),$$
$$T_i = \frac{R_4R_3R_1C}{R_5R_2}, \quad k = \frac{R_6}{R_4R_1C}.$$

Electrotechnique

Pour un moteur pas-à-pas ayant au rotor n dents de polarité Nord et autant de polarité Sud, les différents bilans électromagnétiques (dans le repère dq dit

de Park) donnent :

$$\begin{aligned} L_d \frac{di_d}{dt} &= v_d - Ri_d + nL_q \omega i_q, \\ L_q \frac{di_q}{dt} &= v_q - Ri_q - nL_d \omega i_d - nmi_r \omega, \\ C_{\text{em}} &= n(L_d - L_q) i_d i_q + nmi_r i_q + K_d \sin(n\theta), \end{aligned}$$

où m et i_r sont l'inductance et le courant fictif du rotor, générant le flux (constant) mi_r dû à l'aimantation permanente ; i_d, i_q, v_d, v_q sont les courants et tensions du stator [9]. Le bilan mécanique s'écrit :

$$\begin{aligned} \frac{d\theta}{dt} &= \omega, \\ J \frac{d\omega}{dt} &= C_{\text{em}} - C_{\text{charge}}. \end{aligned}$$

Mécanique

Si un système mécanique est constitué de n éléments reliés entre eux par des liaisons parfaites (sans frottement), on aura la position du système qui dépendra de n paramètres indépendants (coordonnées généralisées notées $q_1, \ldots, q_n$). Pour écrire les **équations d'Euler-Lagrange**, il faut déterminer le **lagrangien** (différence entre l'énergie cinétique et l'énergie potentielle) :

$$\mathcal{L} = \mathcal{E}_c - \mathcal{E}_p, \tag{4.4}$$

le travail élémentaire de chaque force interne et externe D_i, ainsi que le travail des forces de frottement :

$$-\frac{\partial D}{\partial \dot{q}_i} dq_i,$$

donnant lieu à l'énergie dissipée D. On obtient alors le système d'équations d'Euler-Lagrange :

$$\frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) - \frac{\partial \mathcal{L}}{\partial q_i} + \frac{\partial D}{\partial \dot{q}_i} = D_i. \tag{4.5}$$

Notons que l'énergie cinétique $\mathcal{E}_c = \frac{1}{2}\dot{q}^{\mathrm{T}} M(q) \dot{q}$, où $M(q)$ est une matrice $n \times n$ symétrique définie positive, dépend des q_i et de leurs dérivées $\dot{q}_i$, alors que l'energie potentielle $\mathcal{E}_p$ ne dépend que des q_i. Pour un pendule pesant d'angle θ avec la verticale, on a $\mathcal{L} = \frac{1}{2} ml^2 \dot{\theta}^2 - mgl(1 - \cos(\theta))$. Si on néglige les frottements, ceci conduit à :

$$\ddot{\theta} = -\frac{g}{l} \sin(\theta). \tag{4.6}$$

Si de plus on tient compte d'un frottement sec, on peut voir apparaître une discontinuité sur $\ddot{\theta}$ (qui peut ne pas être définie à vitesse nulle). Aussi, pour ce type de modèle, on est amené à préciser la notion de solution et à donner des conditions suffisantes d'existence et/ou d'unicité de solution : ceci sera l'objet de

la section 4.3, qui développera plus particulièrement les EDO du premier ordre. Les EDO sous forme implicite seront abordées à la section 4.2. Au delà des conditions d'existence, il est important pour l'automaticien de pouvoir caractériser les comportements asymptotiques de ces solutions et de disposer d'outils d'analyse permettant de les localiser et de caractériser l'évolution temporelle des solutions vers ces ensembles. Enfin, de nombreux systèmes physiques font intervenir, dans la description de leurs dynamiques au moyen des EDO, des paramètres dont les variations peuvent conduire à des modifications qualitatives des solutions (bifurcations et, parfois, ‹‹ catastrophes ››).

4.2 Equations différentielles ordinaires sous forme implicite

Définitions

Une EDO sous **forme implicite** est une relation :

$$F\left(t, y, \frac{dy}{dt}, \ldots, \frac{d^k y}{dt^k}\right) = 0, \quad y \in \mathbb{R}^m, \tag{4.7}$$

où la fonction F est définie sur un ouvert de $\mathbb{R} \times \mathbb{R}^{m(k+1)}$ à valeur dans $\mathbb{R}^m$. L'**ordre** de l'EDO est l'entier k correspondant à la dérivée d'ordre le plus élevé apparaissant dans (4.7). Notons que (4.1), (4.2) et (4.6) sont des EDO d'ordres respectifs 1, 1 et 2. Le **théorème de la fonction implicite** garantit que ce système (4.7) de m équations peut se mettre (localement) sous **forme explicite** :

$$\frac{d^k y}{dt^k} = G\left(t, y, \frac{dy}{dt}, \ldots, \frac{d^{k-1} y}{dt^{k-1}}\right), \tag{4.8}$$

à condition que :

$$\det\left(J_F\right) \neq 0,$$

où J_F est la matrice jacobienne de F c'est-à-dire la matrice constituée des éléments $a_{ij} = \frac{\partial F_i}{\partial\left(\frac{d^k x_j}{dt^k}\right)}, (i,j) \in \{1, \ldots, m\}^2$.

En particulier, soit F est une fonction de classe $\mathcal{C}^1$ par rapport à chacune de ses variables, telle que $F(t, x, z) = 0$ admette p racines en la variable z. Si $\det(\frac{\partial F}{\partial z}\big|_{(t_0,x_0)}) \neq 0$, alors il existera p solutions locales dites **régulières**, solutions de $F(t, x, \dot{x}) = 0$, $x(t_0) = x_0$. Par contre si $\det(\frac{\partial F}{\partial z}\big|_{(t_0,x_0)}) = 0$, alors on ne peut rien affirmer sans faire une étude approfondie et toute solution sera dite **singulière**. Par exemple $y : t \mapsto t^2/4$, est une solution singulière de l'équation $\dot{y}^2 - t\dot{y} + y = 0$ au voisinage de $(y, \dot{y}) = (0, 0)$. Par ailleurs, $y : t \mapsto \sin(\operatorname{arcsinh}(t))$ est une solution régulière de $(1 + t^2)\dot{y}^2 + (-1 + y^2) = 0$ au voisinage de $(0, 0)$.

Quelques équations particulières

Equations de Lagrange et de Clairaut

Une EDO de la forme :

$$x + tf\left(\frac{dx}{dt}\right) + g\left(\frac{dx}{dt}\right) = 0, \tag{4.9}$$

est dite de **Lagrange**. Lorsque $f = Id$, elle sera dite de **Clairaut**. De façon à résoudre ce type d'équations, on cherche une représentation paramétrée des solutions : pour cela, on pose $\frac{dx}{dt} = m$ et $x = x(m), t = t(m)$, ce qui ramène la résolution de (4.9) à celle de :

$$(m + f(m))\frac{dt}{dm} + f'(m)t = -g'(m),$$

qui est linéaire du premier ordre en t.

Exemple 4.2.1. La résolution de l'équation $x + 2t\dot{x} + \dot{x}^3 = 0$ se ramène à celle de $3m\frac{dt}{dm} + 2t = -3m^2$, qui a pour solution $t\,(m) = \frac{1}{8}\frac{-3\left(\sqrt[3]{m}\right)^8 + 8C_1}{\left(\sqrt[3]{m}\right)^2}$. Les solutions paramétrées de la première équation sont donc :

$$x(m) = -\frac{1}{4}m^3 - 2\sqrt[3]{m}C_1, \quad t\,(m) = \frac{1}{8}\frac{-3\left(\sqrt[3]{m}\right)^8 + 8C_1}{\left(\sqrt[3]{m}\right)^2}.$$

Equation de Bernouilli

Une EDO de la forme :

$$\frac{dx}{dt} + f(t)x + g(t)x^r = 0, \tag{4.10}$$

est dite de **Bernouilli**. Les cas $r = 0$ ou $r = 1$ sont triviaux. Dans le cas général, on utilise le changement de variable $y = x^{1-r}$, conduisant à la résolution de l'EDO :

$$\frac{1}{1-r}\frac{dy}{dt} + f(t)y + g(t) = 0,$$

qui est linéaire du premier ordre en y.

Exemple 4.2.2. La résolution de l'équation $\frac{dx}{dt} + \sin(t)x + \sin(t)x^4 = 0$, en posant $y = x^{-3}$, se ramène à celle de $-\frac{1}{3}\frac{dy}{dt} + \sin(t)y + \sin(t) = 0$, qui admet pour solution $y\,(t) = \left(-e^{3\cos t} + C_1\right)e^{-3\cos t}$.

Equation de Riccati

Certaines équations ne sont pas intégrables (c'est-à-dire que l'on ne peut exprimer les solutions de façon explicite avec des fonctions de l'analyse). Par exemple l'**équation de Riccati** :

$$\frac{dx}{dt} + f(t)x + g(t)x^2 + h(t) = 0,$$

n'est pas intégrable si f, g, h (fonctions continues) ne satisfont pas certaines relations particulières.

4.3 Equations différentielles du premier ordre

Notons que, lorsque la variable y de l'équation implicite (4.7) appartient à un ensemble plus général qu'un produit cartésien d'ouvert de $\mathbb{R}$, alors en posant $x = \left(y, \frac{dy}{dt}, \ldots, \frac{d^{k-1}y}{dt^{k-1}}\right)^T$, l'équation explicite (4.8) peut se mettre sous la forme :

$$\frac{dx}{dt} = f(t, x), \quad t \in \mathcal{I}, \quad x \in \mathcal{X}. \tag{4.11}$$

Dans cette équation : $t \in \mathcal{I} \subset \mathbb{R}$ représente la **variable temporelle**, $\mathcal{X}$ est l'**espace d'état**[1]. En pratique, l'espace d'état peut être borné : il reflète les caractéristiques physiques du système (bornitude des performances). De façon générale, l'espace d'état est une variété différentiable . Lorsque le vecteur x s'exprime à l'aide d'une variable et de ses dérivées successives, $\mathcal{X}$ est aussi appelé **espace de phase**. Cependant, certains auteurs ([1] p. 11) emploient indifféremment les deux dénominations. Le vecteur $x \in \mathcal{X}$ correspondant est le **vecteur d'état** de (4.11) (ou de phase par abus de langage). En pratique, il est construit à partir des variables dont l'évolution régit celle du processus physique. $x(t)$ est l'état instantané à l'instant t. $f : \mathcal{I} \times \mathcal{X} \to T\mathcal{X}$ (espace tangent), $(t, x) \mapsto f(t, x)$, représente le **champ de vecteurs**. Afin de simplifier la suite de l'exposé, on se placera dans le cas où $\mathcal{I} \times \mathcal{X}$ est un ouvert de $\mathbb{R}^{n+1}$ et $T\mathcal{X}$ est $\mathbb{R}^n$.

Notion de solution

Lorsqu'on parle de solution, il faut préciser le problème associé : ici, pour les EDO, il existe le problème aux conditions limites ou frontières[2] et le **problème**

[1] vocabulaire de l'automatique.

[2] énoncé similaire au problème de Cauchy, mais pour lequel la condition initiale est remplacée par la donnée de n valeurs $\phi_{\sigma(i)}(t_i)$ aux instants t_i donnés, $i \in \mathbb{N} = \{1, \ldots, n\}$, $\sigma : \mathbb{N} \to \mathbb{N}$.

aux conditions initiales (dit **de Cauchy** , en abrégé PC) :

$$(\text{PC}) : \begin{cases} \text{‹‹ existe-t-il une fonction :} \\ \phi : \mathcal{I} \subset \mathbb{R} \to \mathcal{X} \subset \mathbb{R}^n, \\ t \mapsto \phi(t), \\ \text{satisfaisant à (4.11) et à la condition initiale suivante : } \phi(t_0) = x_0 \text{? ››} \end{cases}$$

On cherche une fonction du temps $\phi : t \mapsto \phi(t)$, qui soit suffisamment régulière et dont la dérivée soit égale à presque tout instant[3] à la valeur du champ à cet instant et en ce point $x = \phi(t)$. Si $f(u, \phi(u))$ est mesurable[4] , alors on peut exprimer $\phi(t)$ sous la forme :

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} f(v, \phi(v))dv, \tag{4.12}$$

l'intégration étant prise au sens de Lebesgue [27] et ce, même si $t \mapsto f(t, .)$ n'est pas continue en t (cas intéressant en automatique, car pour $\dot{x} = g(t, x, u)$ un retour $u = u(t)$ discontinu peut être envisagé). Ainsi, on cherchera des fonctions au moins absolument continues[5] par rapport au temps.

Solutions, portrait de phase

Définition 4.3.1. On appelle **solution** de (4.11) passant par x_0 à t_0, toute fonction ϕ absolument continue définie sur un intervalle non vide $\mathcal{I}(t_0, x_0) \subset \mathcal{I} \subset \mathbb{R}$ contenant t_0 :

$$\begin{aligned} &\phi : \mathcal{I}(t_0, x_0) \subset \mathcal{I} \subset \mathbb{R} \to \mathcal{X} \subset \mathbb{R}^n, \\ &t \mapsto \phi(t; t_0, x_0), \end{aligned}$$

notée plus simplement $\phi(t)$, satisfaisant (4.12) pour tout $t \in \mathcal{I}(t_0, x_0)$ (ou, de façon équivalente : $\dot{\phi} = f(t, \phi(t))$ presque partout sur $\mathcal{I}(t_0, x_0)$) et telle que $\phi(t_0) = x_0$.

Exemple 4.3.1. En séparant ses variables, l'équation du modèle logistique (4.1) devient :

$$\frac{dx}{ax(x_{\max} - x)} = dt,$$

[3] c'est-à-dire pour tous les temps $t \in \mathcal{T} \setminus \mathcal{M}$, $\mathcal{M}$ étant un ensemble de mesure nulle, avec la notation ensembliste $\mathcal{T} \setminus \mathcal{M} = \{x \in \mathcal{T} : x \notin \mathcal{M}\}$.

[4] cette condition est vérifiée si, pour x fixé, $t \mapsto f(t, x)$ est mesurable et pour t fixé, $x \mapsto f(t, x)$ est continue [27].

[5] $\phi : [\alpha, \beta] \mapsto \mathbb{R}^n$ est **absolument continue** si $\forall \varepsilon > 0, \exists \delta(\varepsilon) > 0 : \forall \{]\alpha_i, \beta_i[\}_{i \in \{1..n\}},]\alpha_i, \beta_i[\subset [\alpha, \beta], \sum_{i=1}^{n} (\beta_i - \alpha_i) \leq \delta(\varepsilon) \Rightarrow \sum_{i=1}^{n} \|\phi(\beta_i) - \phi(\alpha_i)\| \leq \varepsilon$. ϕ est absolument continue si et seulement si il existe une fonction (Lebesgue intégrable) qui soit presque partout égale à la dérivée de ϕ.

qui permet de construire la solution au PC (4.1), $x(0) = x_0$:

$$\begin{aligned} &\phi : \mathbb{R} \to \mathbb{R}, \\ &t \mapsto \phi(t; 0, x_0) = \frac{x_0 x_{\max}}{x_0 + e^{-a x_{\max} t}(x_{\max} - x_0)}. \end{aligned} \tag{4.13}$$

Définition 4.3.2. La solution de (4.11) peut être représentée dans deux espaces :

- soit dans l'espace d'état étendu $\mathcal{I} \times \mathcal{X}$ ou **espace du mouvement**, dans ce cas, on parle de mouvement ou de trajectoire ,
- soit dans l'espace d'état $\mathcal{X}$, dans ce cas, on parle d'**orbite**.

On appelle **portrait de phase** l'ensemble de toutes les orbites munies de leurs sens de parcours temporel.

Bien souvent, par commodité on ne représente que les ensembles de points d'accumulation vers lesquels les orbites convergent pour des temps très grands (positifs ou négatifs). Par exemple, pour le système :

$$\frac{dx}{dt} = \begin{pmatrix} 1 - x_1^2 - x_2^2 & -1 \\ 1 & 1 - x_1^2 - x_2^2 \end{pmatrix} x, \quad t \in \mathbb{R}, x \in \mathbb{R}^2, \tag{4.14}$$

les éléments importants du portrait de phase sont l'origine et le cercle unité $\mathbf{C}_1$: si la condition initiale est différente de l'origine, les orbites convergent vers $\mathbf{C}_1$; sinon, l'état reste figé à l'origine.

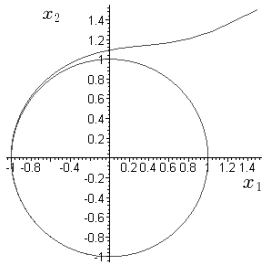


FIG. 4.2: Cercle unité : simulation de (4.14).

Existence

Le problème de Cauchy (PC) n'a pas forcément de solution et, parfois, peut en avoir plusieurs. En effet, le système :

$$\begin{aligned} \frac{dx}{dt} &= |x|^{\frac{1}{2}}, \quad x \in \mathbb{R}, \\ x(0) &= 0, \end{aligned} \tag{4.15}$$

admet une infinité de solutions définies par :

$$\varepsilon \in \mathbb{R}_+, \quad \phi_\varepsilon : \mathbb{R} \to \mathbb{R},$$
$$t \mapsto \phi_\varepsilon(t) = \begin{cases} 0 & \text{si} \quad t_0 - \varepsilon \leq t \leq t_0 + \varepsilon, \\ \frac{(t-t_0-\varepsilon)^2}{4} & \text{si} \quad t_0 + \varepsilon \leq t, \\ -\frac{(t-t_0+\varepsilon)^2}{4} & \text{si} \quad t \leq t_0 - \varepsilon. \end{cases} \tag{4.16}$$

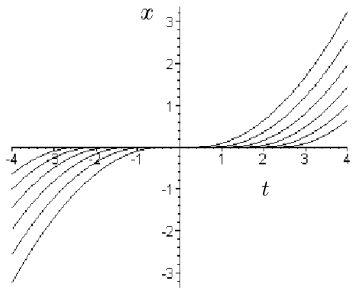


FIG. 4.3: Infinité de solutions au PC (4.15).

Ainsi, il se pose le problème de donner des conditions assurant l'existence d'une ou de plusieurs solutions au problème de Cauchy.

Selon la régularité de la fonction f on distingue les cinq cas A, B, C, D, E qui suivent.

Cas A) Si la fonction f est continue en x et éventuellement discontinue en t (mais mesurable), alors il y a existence de solutions absolument continues.

Théorème 4.3.1 (de Carathéodory, 1918)**.** *[11] Supposons que :*

A1) f soit définie pour presque tout t sur un tonneau :

$$\mathcal{T} = \{(t,x) \in \mathcal{I} \times \mathcal{X} : |t - t_0| \leq a, \|x - x_0\| \leq b\}, \tag{4.17}$$

A2) f soit mesurable en t pour tout x fixé, continue en x pour t fixé et telle que, sur $\mathcal{T}$, on ait $\|f(t,x)\| \leq m(t)$, où m est une fonction positive Lebesgue-intégrable sur $|t - t_0| \leq a$. Alors, il existe ***au moins une*** *solution (absolument continue) au problème de Cauchy définie sur au moins un intervalle $[t_0 - \alpha, t_0 + \alpha]$, $\alpha \leq a$.*

On peut même montrer l'existence de deux solutions, l'une dite **supérieure** et l'autre, **inférieure**, telles que tout autre solution soit comprise entre ces deux solutions [23] [11].

Cas B) Si la fonction f est continue en (t,x), alors il y a existence de solutions de classe $\mathcal{C}^1$.

Théorème 4.3.2 (de Peano, v.1886)**.** *[8] Supposons que :*

B1) f soit définie pour tout t sur un tonneau $\mathcal{T}$ défini par (4.17),

B2) f soit continue sur $\mathcal{T}$ défini par (4.17). Alors il existe ***au moins une*** *solution au problème de Cauchy de classe $\mathcal{C}^1$ définie sur au moins un intervalle $[t_0-\alpha, t_0+\alpha]$, $\alpha = \min(a, \frac{b}{\max_{\mathcal{T}}\|f(t,x)\|})$.*

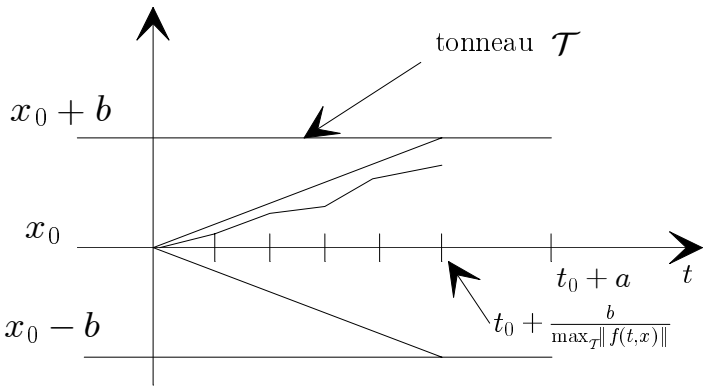


FIG. 4.4: Approximation d'Euler.

La preuve est basée sur les **approximations d'Euler**. Ce sont les lignes polygonales (voir la figure 4.4) définies par :

$$\begin{cases} \phi_0 = x_0, \\ \phi_n(t) = \phi_n(t_{i-1}) + f(t_{i-1}, \phi_n(t_{i-1}))(t - t_{i-1}), t_{i-1} < t \leq t_i, \\ t_i = t_0 + \frac{i}{n}\alpha, i = \{0, \ldots, n\}, \end{cases}$$

dont on montre qu'elles constituent une famille équicontinue de fonctions définies sur $[t_0-\alpha, t_0+\alpha]$, convergente, dont on peut extraire une sous-suite ϕ'_n uniformément convergente vers une fonction continue ϕ (lemme d'Ascoli-Arzela [27]) qui vérifie :

$$\phi(t) = \lim_{n\to+\infty} \phi'_n(t) = x_0 + \int_{t_0}^{t} \lim_{n\to+\infty} f(v, \phi'_n(v))dv$$
$$+ \lim_{n\to+\infty} \int_{t_0}^{t} \frac{d\phi'_n}{dt}(v) - f(v, \phi'_n(v))dv.$$

ϕ est donc solution de (4.12) puisque $\lim_{n\to+\infty} \frac{d\phi'_n}{dt}(v) - f(v, \phi'_n(v)) = 0$.

Prolongement de solution, unicité, solution globale

Evidement, pour l'exemple (4.15), il y a existence de solution ($f : x \mapsto \sqrt{|x|}$ est continue) mais pas unicité. Pour garantir l'unicité, il faut donc que la fonction f soit ‹‹ plus que continue ›› : par exemple il suffit qu'elle soit localement lipschitzienne en la seconde variable x, comme suit.

Définition 4.3.3. f est dite **localement lipschitzienne** sur $\mathcal{X}$ si :

$$\forall x_0 \in \mathcal{X}, \quad \exists \delta > 0 \text{ et } k(t) \text{ intégrable :}$$
$$\forall (x, y) \in \mathcal{B}_\delta(x_0) \Rightarrow \|f(t, x) - f(t, y)\| \leq k(t) \|x - y\| .$$

f est dite **globalement lipschitzienne** sur X si :

$$\exists k(t) \text{ intégrable } : \forall (x, y) \in \mathcal{X}^2,$$
$$\|f(t, x) - f(t, y)\| \leq k(t) \|x - y\| .$$

Ces propriétés sont dites **uniformes** si k ne dépend pas de t.

Proposition 4.3.1. *Toute fonction $\mathcal{C}^1(\mathcal{I} \times \mathcal{X})$ dont la norme de la jacobienne est bornée par une fonction intégrable, est localement lipschitzienne. De plus, si $\mathcal{X}$ est compact (fermé, borné), elle est globalement lipschitzienne.*

Sous l'hypothèse f localement lipschitzienne en x, il se peut qu'une solution ϕ définie sur $\mathcal{I}_1$ puisse être prolongée sur un intervalle $\mathcal{I}_2 \supset \mathcal{I}_1$, définissant ainsi une fonction $\widetilde{\phi}$ définie sur $\mathcal{I}_2 \supset \mathcal{I}_1$ et telle que $\widetilde{\phi} \mid \mathcal{I}_1 = \phi$. Aussi, afin de ne pas alourdir les notations, $\mathcal{I}(t_0, x_0) =]\alpha(t_0, x_0), \omega(t_0, x_0)[$ désignera par la suite le plus grand intervalle sur lequel on peut définir une solution passant à t_0 par x_0 et qui ne puisse être prolongée : la solution sera dite **maximale** .

Cas C) Si la fonction f est localement lipschitzienne en x et éventuellement discontinue en t (mais mesurable), alors il y a existence et unicité de solution (maximale) absolument continue.

Théorème 4.3.3. *Si dans le théorème 4.3.1, l'hypothèse de continuité dans A2) est remplacée par ‹‹ f localement lipschitzienne sur $\|x - x_0\| \leq b$ ››, alors il existe **une unique** solution (absolument continue) au problème de Cauchy définie sur*

$$\mathcal{I}(t_0, x_0) \supset \{t \in \mathcal{I} : |t - t_0| \leq \alpha\} .$$

De même, si f est continue en (t, x) et localement lipschitzienne en x, alors il existe une unique solution au problème de Cauchy de classe $\mathcal{C}^1$.

Les preuves de ces résultats sont basées sur les **approximations de Picard-Lindelöf** :

$$\begin{cases} \phi_0 = x_0, \\ \phi_{n+1}(t) = x_0 + \int_{t_0}^{t} f(v, \phi_n(v)) dv, \\ t \in [t_0 - \alpha, t_0 + \alpha], \alpha = \min(a, \frac{b}{\max_{\mathcal{T}} \|f(t,x)\|}), \end{cases}$$

dont on montre qu'elles convergent uniformément vers une solution. L'unicité de la solution se démontre ensuite par contradiction.

Cas D) Si la fonction f est bornée en norme par une fonction affine, c'est-à-dire que $\forall(t,x) \in (\mathcal{I} \times \mathcal{X}) : \|f(t,x)\| \leq c_1 \|x\| + c_2$ (éventuellement presque partout), alors en utilisant le lemme de Gronwall [27], on peut conclure que toute solution au PC est définie sur $\mathcal{I}$.

Cas E) Si le système possède une propriété de ‹‹ dissipativité ›› et si f est localement lipschitzienne, alors le PC admet une solution maximale unique définie pour tout $t \geq t_0$ ($\mathcal{I} = \mathbb{R}, \mathcal{X} = \mathbb{R}^n$).

La propriété de dissipativité peut être exprimée sous la forme : ‹‹ il existe $\alpha \geq 0, \beta \geq 0, v \in \mathbb{R}^n$ tels que pour tout $t \in \mathbb{R}$ et tout $x \in \mathbb{R}^n : < x-v, f(t,x) > \leq \alpha - \beta \|x\|^2$ ›› ou bien, à l'aide de fonctions de Liapounov[6], sous la forme ‹‹ il existe V et $W : \mathbb{R}^n \mapsto \mathbb{R}_+$ continues, définie positives sur un compact $\mathcal{A}$ (c'est-à-dire : $V(x) = 0 \Leftrightarrow x \in \mathcal{A}$), telles que pour tout $t \in \mathbb{R}$ et tout $x \in \mathbb{R}^n \backslash \mathcal{A}$[7] : $< \frac{\partial V}{\partial x}, f(t,x) > \leq -W(x)$. ››

Dépendance des conditions initiales

Théorème 4.3.4. *Sous les hypothèses du théorème 4.3.3, la solution au problème de Cauchy $t \mapsto \phi(t; t_0, x_0)$ définie sur $\mathcal{I}(t_0, x_0)$ est continue par rapport à chacun de ses arguments.*

En particulier, si t est suffisamment proche de t_0, alors $\phi(t; t_0, x_0)$ est proche de x_0. Cette proximité peut être étudiée pour des instants très grands : c'est la question de la stabilité (voir paragraphe 4.5).

Classification

Définition 4.3.4. L'équation (4.11) est dite **autonome** si la variable temporelle n'apparaît pas explicitement dans l'EDO :

$$\frac{dx}{dt} = g(x), \quad t \in \mathcal{I}, \quad x \in \mathcal{X}.$$

Dans le cas contraire, elle est dite **non autonome**.

Si on connaît les solutions d'une EDO autonome passant à un instant t donné par un point x donné, alors on obtient toutes les solutions passant par ce même

[6] Alexander Mikhaïlovich Liapounov, mathématicien et physicien russe. Après des études à l'Université de Saint-Pétersbourg (où il est élève de P.L. Tchebychev), il est assistant puis professeur à l'Université de Kharkov. En 1902, il est nommé professeur à l'Université de Saint-Pétersbourg.

[7] La notation $A \backslash B$ correspond à la différence ensembliste de A et B : $A \backslash B = \{x \in A : x \notin B\}$.

point à d'autres instants par simple translation temporelle des premières. Donc, une EDO autonome ne peut servir qu'à modéliser des phénomènes physiques indépendants du temps initial (chute d'un corps, etc.). Notons au passage que la longueur de $\mathcal{I}(t_0, x_0)$ ne dépend pas de l'instant initial.

Définition 4.3.5. On dit qu'un champ de vecteurs non linéaire non autonome $f(t, x)$ est $T-$**périodique** s'il existe un réel $T > 0$ tel que pour tout t et pour tout $x : f(t + T, x) = f(t, x)$.

Dans ce cas, si on connaît les solutions pour un intervalle de longueur T, on aura toutes les autres par translation temporelle.

Cas linéaire autonome

Lorsque (4.11) est de la forme :

$$\frac{dx}{dt} = Ax + b,$$

on dit qu'il s'agit d'une EDO **linéaire autonome**. Il existe alors une solution unique au PC (puisque $Ax + b$ est globalement uniformément lipschitzienne) de la forme :

$$x(t) = \mathrm{e}^{A(t-t_0)}x_0 + \mathrm{e}^{At}\left(\int_{t_0}^{t} \mathrm{e}^{-Av}dv\right)b,$$

ou encore $x(t) = \sum_{i=1}^{r} \mathrm{e}^{\lambda_i t}p_i(t) + c$, avec λ_i les valeurs propres de A et $p_i(t)$ des vecteurs de polynômes de degrés plus petits que l'ordre de multiplicité de la valeur propre correspondante λ_i.

Ce type de modèle caractérise des phénomènes pour lesquels :

1. l'instant initial n'a pas d'influence sur l'évolution temporelle du vecteur état (EDO autonome) ;
2. si $b = 0$ (respectivement, $b \neq 0$), alors une combinaison linéaire (respectivement, convexe) d'évolutions est encore une évolution possible : ceci traduit la **linéarité** du système.

Pour de tels systèmes, on peut noter que, au bout d'un temps infini, les différentes variables définissant le vecteur x :

1. soit convergent vers un vecteur (le vecteur x évoluant vers un vecteur fixe dit ‹‹ point d'équilibre ››),
2. soit divergent (la norme de x devient infiniment grande),
3. soit ont un comportement oscillatoire : lorsqu'on observe leurs évolutions les unes en fonction des autres, elles évoluent sur une courbe fermée (comme le cercle) : c'est ce qu'on appelle un cycle fermé (exemples : cycle économique, population cyclique, masse attachée à un ressort, etc.).

Enfin, si A et b dépendent du temps, le système est dit **linéaire non autonome** (ou **non stationnaire**) : en plus des comportements précités on retrouve la dépendance des solutions à l'instant initial . Notons que lorsque $A(t)$ est une fonction $T-$périodique continue sur $\mathbb{R}$, on peut étudier formellement les solutions grace à la **théorie de Floquet** [17] : il existe alors une transformation bijective $P(t)$ $T-$périodique et continue, $z(t) = P(t)x(t)$, telle que $\dot{z} = Mz+c(t)$, avec M une matrice constante vérifiant $M = \dot{P}(t) + P(t)A(t)P(t)^{-1}$ et $c(t) = P(t)b(t)$.

Cas non linéaire autonome

Lorsque (4.11) est de la forme :

$$\frac{dx}{dt} = g(x), \tag{4.18}$$

on dit qu'il s'agit d'une EDO **non linéaire autonome**. On ne peut lui donner de solution explicite, sauf dans des cas très particuliers. Outre les comportements précités dans le cas linéaire autonome, on peut mentionner :

1. **Cycles limites** : il s'agit de courbes fermées de $\mathcal{X}$ vers ou à partir desquelles les trajectoires du système se dirigent.
2. **Phénomène de chaos** : ces comportements, régis par des EDO (déterministes), sont en apparence aléatoires. Une des caractéristiques est la sensibilité aux conditions initiales : deux conditions initiales très proches donneront naissance à deux évolutions complètement différentes.
3. **Attracteur étrange** : il s'agit en général d'un ensemble de dimension non entière, ce qui traduit une certaine ‹‹ rugosité ›› de l'objet. Par exemple, une surface est de dimension 2, un volume est de dimension 3, alors qu'un flocon de neige ayant une infinité de ramifications est de dimension non entière comprise entre 2 et 3. Lorsque les trajectoires se dirigent vers (respectivement s'éloignent de) cet ensemble, il est dit ‹‹ attracteur (respectivement répulseur) étrange ››. Souvent la présence d'attracteurs ou répulseurs étranges est un signe du chaos, cependant dans certains cas le phénomène chaotique n'est que transitoire et disparaît au bout d'un temps suffisamment long.

Notion de flot

Dans cette partie, on suppose l'existence et l'unicité de la solution (maximale) au PC associé à (4.18), que l'on notera $\phi(t; t_0, x_0)$. Si le champ est **complet**, c'est-à-dire $\mathcal{I}(x_0) = \mathbb{R}$, et si on connaît une solution pour un couple (t_0, x_0), alors on aura toutes les autres (pour x_0 fixé) par translation temporelle. Considérons l'application qui, à toute condition initiale, associe sa solution maximale à l'instant t :

$$\begin{aligned} &\Phi_g^t : \mathcal{X} \to \mathcal{X}, \\ &x_0 \mapsto \phi(t; 0, x_0). \end{aligned}$$

Définition 4.3.6. Si le champ de vecteurs g de l'EDO (4.18) permet de générer une solution maximale unique pour tout (t_0, x_0) de $\mathbb{R} \times \mathcal{X}$, définie sur $\mathcal{I}(x_0) = \mathbb{R}$ (respectivement sur $[\alpha, \infty[$, sur $[\alpha, \omega]$ avec α et ω finis), alors l'application génératrice Φ_g^t est appelée un **flot** (respectivement un **semi-flot**, un **flot local**).

D'après les hypothèses, Φ_g^t est bijective, donc on a au moins un flot local. La justification du choix de la notation Φ_g^t devient évidente lorsqu'on calcule le flot d'une EDO linéaire autonome homogène : $\dot{x} = Ax$, $\Phi_g^t = e^{At}$. Dans le cas où g est de classe $\mathcal{C}^k$ (respectivement $\mathcal{C}^\infty$, analytique), le flot associé Φ_g^t est un difféomorphisme local de classe $\mathcal{C}^k$ (respectivement $\mathcal{C}^\infty$, analytique) pour tout instant t où il est défini. En particulier, si le flot Φ_g^t est défini pour tout $t \in \mathbb{R}$, alors il définit un groupe à un paramètre de difféomorphismes locaux de classe $\mathcal{C}^k$ (respectivement $\mathcal{C}^\infty$, analytique) (voir [1] p. 55 à 63) :

$$\Phi_g^t : x_0 \mapsto \Phi_g^t(x_0) \text{ est de classe } \mathcal{C}^\infty, \tag{4.19}$$

$$\Phi_g^t \circ \Phi_g^s = \Phi_g^{t+s}, \tag{4.20}$$

$$\Phi_g^0 = Id. \tag{4.21}$$

On en déduit, $\forall t \in \mathbb{R}, \forall x_0 \in \mathcal{X}$:

$$\Phi_g^t(x_0) = \Phi_{-g}^{-t}(x_0), \tag{4.22}$$

$$\Phi_g^t \circ \Phi_g^{-t} = \Phi_g^{t-t} = Id, \tag{4.23}$$

$$(\Phi_g^t)^{-1} = \Phi_g^{-t} = \Phi_{-g}^t. \tag{4.24}$$

La dualité caractérisée par (4.24) est importante puisque, si on connaît le portrait de phase du système dynamique (4.18) pour les temps positifs, son dual pour les temps négatifs s'obtient tout simplement en renversant le sens de parcours des orbites : cette propriété est utilisée dans la méthode du **renversement des trajectoires** (‹‹ trajectory-reversing method ››) permettant, en dimension deux (et parfois trois), de déterminer précisément la plupart des portraits de phase en alliant l'étude qualitative du champ de vecteurs non linéaire à des simulations (pour plus de détails, voir [5, 6, 13, 12, 25]).

Le **crochet de Lie** (ou commutateur, voir chapitre 4) défini par :

$$[g_1, g_2] = \left(\frac{\partial g_2}{\partial x} g_1 - \frac{\partial g_1}{\partial x} g_2 \right),$$

permet de calculer la condition de commutativité de deux flots $\Phi_{g_1}^t$ et $\Phi_{g_2}^s$.

Théorème 4.3.5. *Soient g_1 et g_2 des champs de vecteurs $\mathcal{C}^\infty$ complets, définis sur $\mathcal{X}$ (par exemple $\mathbb{R}^n$). Alors :*

$$\forall t, \forall s, \quad \Phi_{g_1}^t \circ \Phi_{g_2}^s = \Phi_{g_2}^s \circ \Phi_{g_1}^t \Leftrightarrow [g_1, g_2] = 0.$$

Démonstration : soient $x_0 \in \mathbb{R}^n$ et $t, s > 0$ donnés. Pour un champ de vecteurs analytique X, on a $\Phi_X^t(y) = y + tX(y) + R(t, y)$, où $R(t, y)$ représente un reste s'annulant pour $t \to 0$. On obtient donc :

$$\Phi_{g_1}^t \circ \Phi_{g_2}^s(x_0) = x_0 + (sg_2(x_0) + tg_1(x_0)) + st\frac{\partial g_1}{\partial x} g_2(x_0) + R_1(t, s, x_0),$$
$$\Phi_{g_2}^s \circ \Phi_{g_1}^t(x_0) = x_0 + (sg_2(x_0) + tg_1(x_0)) + st\frac{\partial g_2}{\partial x} g_1(x_0) + R_2(t, s, x_0),$$

et donc :

$$\Phi_{g_1}^t \circ \Phi_{g_2}^s(x_0) - \Phi_{g_2}^s \circ \Phi_{g_1}^t(x_0) = st[g_2, g_1](x_0) + R_3(t, s, x_0).$$

Prenons $t = s$, alors l'implication découle immédiatement. Pour la réciproque, $[g_1, g_2] = 0 \Rightarrow \forall x_0 \in \mathbb{R}^n : \lim_{t\to 0}(\frac{\Phi_{g_2}^{-t} \circ \Phi_{g_1}^s \circ \Phi_{g_2}^t(x_0) - \Phi_{g_1}^s(x_0)}{t}) = 0$. Soit la trajectoire $x(t) = \Phi_{g_2}^{-t} \circ \Phi_{g_1}^s \circ \Phi_{g_2}^t(x_0)$, alors $\dot{x}(t) = 0$, donc $\Phi_{g_2}^{-t} \circ \Phi_{g_1}^s \circ \Phi_{g_2}^t = \Phi_{g_1}^s$. ∎

En automatique, la non-commutativité des champs a une application très importante puisqu'elle permet de caractériser l'atteignabilité (version locale de la commandabilité) d'un système commandé du type $\dot{x} = g_1(x) + g_2(x)u$ (voir chapitre 4 et [20]).

Comparaison de solutions, inégalités différentielles

Dans cette partie, les inégalités vectorielles seront à comprendre composante à composante et D est un opérateur de dérivation (dit de Dini) :

$$Dx(t) = (Dx_1(t), \ldots, Dx_n(t)),$$

où les dérivées $Dx_i(t)$ sont toutes définies par l'une des quatre limites suivantes :

$$D_+x_i(t) = \lim_{\theta\to 0^+} \inf \frac{x_i(t+\theta) - x_i(t)}{\theta},\ D^+x_i(t) = \lim_{\theta\to 0^+} \sup \frac{x_i(t+\theta) - x_i(t)}{\theta},$$

$D_-x_i(t)$ et $D^-x_i(t)$ définies de façon similaire pour les limites par valeurs inférieures ($\theta \to 0^-$).

Soit $\mathcal{S} \subset \mathcal{I}$ un ensemble de mesure nulle. On considère les relations différentielles suivantes :

$$Dx(t) \leq f(t, x), \quad t \in \mathcal{I}\backslash\mathcal{S}, x \in \mathcal{X}, \tag{4.25}$$
$$\frac{dz}{dt} = f(t, z), \quad t \in \mathcal{I}\backslash\mathcal{S}, z \in \mathcal{X}. \tag{4.26}$$

On se pose le problème de savoir sous quelles conditions les solutions de (4.26) majorent celles de (4.25). Dans ce cas, (4.26) constituera un système majorant de (4.25), dont l'utilisation est intéressante pour l'analyse des comportements des solutions d'une EDO (voir paragraphe 4.5). La contribution de Ważewski [30] (précédée de [21] et complétée par celle de Lakshmikantham et Leela [23])

est certainement l'une des plus importantes puisqu'elle donne des conditions nécessaires et suffisantes pour que les solutions du problème de Cauchy associé à (4.25) avec x_0 donné à t_0, soient majorées par la solution supérieure de (4.26) démarrant de $z_0 \geq x_0$ à l'instant t_0.

Définition 4.3.7. La fonction $f : \mathcal{I} \times \mathcal{X} \to \mathcal{X} \subset \mathbb{R}^n$, $(t, x) \mapsto f(t, x)$ est **quasi-monotone non-décroissante en** x si :

$$\forall t \in \mathcal{I}, \forall (x, x') \in \mathcal{X}^2, \forall i \in \{1, \ldots, n\},$$
$$\big[(x_i = x'_i) \text{ et } (x \leq x')\big] \Rightarrow f_i(t, x) \leq f_i(t, x'). \tag{4.27}$$

Théorème 4.3.6. *Supposons que :*

1) $f(t, x)$ vérifie les hypothèses (A1-A2) du théorème 4.3.1 d'existence de solution,

2) $f(t, x)$ est quasi-monotone non-décroissante en x presque partout en t ($\forall t \in \mathcal{I} \backslash \mathcal{S}$).

Alors, pour tout $x_0 \in \mathcal{X}$, les solutions du problème de Cauchy associé à (4.26) vérifient :

$$x(t) \leq z_{\sup}(t), \tag{4.28}$$

pour tout instant où les quantités $x(t)$ et $z_{\sup}(t)$ ont un sens, avec $z_{\sup}(t)$ la solution supérieure de (4.25) démarrant de $z_0 \in \mathcal{X}$, avec $x_0 \leq z_0$.

4.4 EDO Linéaire : des comportements simplistes

Un peu de vocabulaire

Soit $x(t)$ un vecteur. Partant d'un système de n équations différentielles linéaires du premier ordre, on se ramène à une EDO de la forme

$$\frac{dx}{dt} = A(t)x + b(t). \tag{4.29}$$

Elle sera dite **linéaire non autonome non homogène** lorsque les fonctions $A(t), b(t)$ dépendent explicitement du temps (respectivement constantes) et **linéaire non autonome homogène** lorsque $b = 0$. Notons que toute équation linéaire d'ordre quelconque peut se mettre sous cette forme. En effet

$$a_0(t)y + a_1(t)\dot{y} + \ldots + a_n(t)y^{(n)} = b_0(t),$$

se met sous la forme (4.29) en posant $x = (y, \ldots, y^{(n-1)})^T$, $b = (0, \ldots, \frac{b_0(t)}{a_n(t)})^T$ et

$$A(t) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \\ -\frac{a_0(t)}{a_n(t)} & -\frac{a_1(t)}{a_n(t)} & \cdots & -\frac{a_{n-1}(t)}{a_n(t)} \end{pmatrix}.$$

Si on regarde l'équation homogène alors l'ensemble des solutions constituent un espace vectoriel : toute combinaison de solutions est solution. Ainsi le problème se ramène à la recherche d'une base de solutions fondamentales. Les solutions de l'équation non homogène se déduisent alors facilement. Par contre, en fixant une condition initiale le problème de Cauchy (PC) n'a qu'une et une seule solution :

Théorème 4.4.1. *Si b et A sont continues sur leur intervalle commun de définition (noté $\mathcal{I}$) alors le PC admet une solution unique, c'est-à-dire étant donné $t_0 \in \mathcal{I}$ et x_0 il n'existe qu'une seule fonction $x(t)$ vérifiant l'équation (4.29) et telle que $x(t_0) = x_0$.* ■

Lorsque l'EDO est de la forme

$$\frac{dx}{dt} = Ax + b, \tag{4.30}$$

on dit qu'il s'agit d'une EDO **linéaire autonome non homogène**. Il existe une solution unique au PC (utiliser théorème 4.4.1 ou alors cf. preuve directe qui suit). Cette solution est de la forme

$$x(t) = \exp(A(t - t_0))x_0 + \left(\int_{t_0}^{t} \exp(A\,(t - v))dv \right) b, \tag{4.31}$$

ou encore $x(t) = \sum_{i=1}^{r} \exp(\lambda_i t)p_i(t) + c$, avec λ_i les valeurs propres de A et $p_i(t)$ des vecteurs de polynômes de degré plus petit que l'ordre de multiplicité de la valeur propre correspondante λ_i. Notons que (4.31) peut s'obtenir facilement en utilisant la formule de la variation de la constante et peut être étendue au cas $b(t)$.

Démonstration. Existence : dériver (4.31). Unicité : si $x(t)$ et $y(t)$ sont deux solutions (de classe $\mathcal{C}^1$) alors $z = x - y$ est solution de $\dot{z} = Az, z(t_0) = 0$. Si $z(t)$ n'est pas identiquement nulle alors il existe un temps T tel que $z(t) = 0, \forall t < T$ et $z(T) \neq 0$. Notons i l'index d'une des composantes du vecteur z qui devient non nulle à $t = T$ alors $z_i(T) = z_i(T - \varepsilon) + \int_{T-\varepsilon}^{T} [Az(v)]_i \, dv = 0$ (contradiction). □

Les comportements linéaires

La solution au PC associé à

$$\dot{x} = Ax, x \in \mathbb{R}^n. \tag{4.32}$$

$(x(t_0) = x_0)$ peut s'exprimer sous la forme

$$x(t) = \exp(A(t - t_0))x_0. \tag{4.33}$$

Ce type de modèle caractérise les phénomènes suivants :

1. l'instant initial n'a pas d'influence sur l'évolution temporelle du vecteur état (EDO autonome) : si $x_1(t)$ et $x_2(t)$ sont des solutions de (4.32) telles que $x_1(t_1) = x_0$ et $x_2(t_2) = x_0$ alors $x_1(t - t_1) = x_2(t - t_2)$.
2. une combinaison linéaire d'évolutions est encore une évolution possible : ceci traduit la linéarité du système.

Pour de tels systèmes, on peut noter que, au bout d'un temps infini, le vecteur $x(t)$:

1. soit converge vers un vecteur fixe dit "point d'équilibre",
2. soit diverge (au moins une des composantes de $x(t)$ devient infiniment grande),
3. soit les composantes de $x(t)$ ont un comportement oscillatoire : lorsqu'on observe leurs évolutions les unes en fonction des autres, elles évoluent sur une courbe fermée (comme le cercle) : c'est ce qu'on appelle un cycle fermé (exemples : cycle économique, population cyclique, masse attachée à un ressort, etc...).

4.5 EDO Non linéaire

Un peu de vocabulaire

Dans cette partie, on s'intéresse à des EDO de la forme

$$\frac{dx}{dt} = g(x), \quad x \in \mathcal{X}. \tag{4.34}$$

Dans cette équation : $t \in \mathbb{R}$, représente la **variable temporelle**, $\mathcal{X}$ est l'**espace d'état**[8]. En pratique l'espace d'état peut être borné : il reflète les caractéristiques physiques du système (bornitude des performances). Lorsque le vecteur d'état s'exprime à l'aide d'une variable et de ses dérivées successives, l'espace d'état s'appelle aussi **espace de phase**. Cependant, certains auteurs ([1] p. 11) emploient indifféremment les deux dénominations. $x \in \mathcal{X}$, représente le **vecteur d'état** (ou de phase par abus de langage) construit à partir des variables dont l'évolution régit celle du processus physique. $x(t)$ le vecteur d'état instantané à l'instant t. $g : \mathcal{X} \to T\mathcal{X}$ (espace tangent), $x \mapsto g(x)$, représente le **champ de vecteurs**.

Condition 4.5.1. *Afin de clarifier la suite de l'exposé, on se placera dans le cadre où $\mathcal{X}$ est un ouvert de $\mathbb{R}^n$ et $T\mathcal{X}$ est $\mathbb{R}^n$.*

Lorsqu'on parle de solution, il faut préciser le problème associé : ici pour les EDO il existe le problème aux conditions limites ou frontières[9] et le **problème**

[8] vocabulaire de l'automatique.

[9] énoncé similaire au problème de Cauchy pour lequel la condition initiale est remplacée par la donnée de n valeurs $\phi_{\sigma(i)}(t_i)$ aux instants t_i donnés ($i \in N = \{1, \ldots, n\}, \sigma : N \to N$).

aux conditions initiales (dit Problème **de Cauchy** abrégé PC) : "existe-t-il une fonction

$$\begin{aligned}\phi &: \mathcal{I} \subset \mathbb{R} \to \mathcal{X} \subset \mathbb{R}^n,\\ t &\mapsto \phi(t),\end{aligned}$$

satisfaisant à (4.34) et à la condition initiale suivante : $\phi(t_0) = x_0$?" On cherche une fonction du temps $\phi : t \mapsto \phi(t)$, qui soit suffisamment régulière (par exemple de classe $\mathcal{C}^1$) telle que sa dérivée soit égale à la valeur du champ à cet instant et au point $x = \phi(t)$. Si g est intégrable, on peut alors exprimer $\phi(t)$ sous la forme

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} g(\phi(v))dv. \tag{4.35}$$

De nombreux résultats permettent de statuer quant à l'existence de solution (g continue) et l'unicité (g Lipschitzienne). On rappelle que le **portrait de phase** est l'ensemble de toutes les orbites munies de leur sens de parcours temporel. Bien souvent, par commodité on ne représente que les ensembles de points d'accumulation vers lesquels les orbites convergent pour des temps très grand ou très petit. Par exemple pour le système

$$\frac{dx}{dt} = \begin{pmatrix} 1 - x_1^2 - x_2^2 & -1 \\ 1 & 1 - x_1^2 - x_2^2 \end{pmatrix} x, \quad t \in \mathbb{R}, \quad x \in \mathbb{R}^2, \tag{4.36}$$

les éléments importants du portrait de phase sont l'origine et le cercle unité : si la condition initiale est différente de l'origine les orbites convergent vers le cercle unité sinon l'état reste figé à l'origine (cf. figure 4.5).

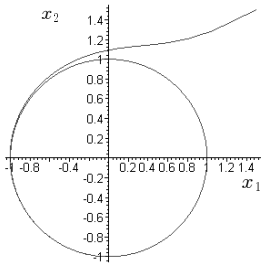


FIG. 4.5: Cercle unité : simulation de (4.36).

Equilibre

Pour certaines conditions initiales, le système reste "gelé", c'est-à-dire qu'il n'évolue plus : on parlera alors de points d'équilibre.

Définition 4.5.1. $x_e \in \mathcal{X}$ est un **point d'équilibre** pour le système (4.34) si les solutions $\phi(t;0,x_e)$ de (4.34) sont définies sur $[0,+\infty[$ et vérifient :

$$\phi(t;0,x_e) = x_e, \forall t \in [0,+\infty[. \tag{4.37}$$

Exemple 4.5.1. La solution au PC (4.1) $x(0) = x_0$ est donnée par

$$\begin{aligned} &\phi : \mathbb{R} \to \mathbb{R} \\ &t \mapsto \phi(t;0,x_0) = \frac{x_0 x_{\max}}{x_0 + e^{-at}(x_{\max} - x_0)}. \end{aligned} \tag{4.38}$$

Il est facile de vérifier que $x = 0$ ($\phi(t;0,x_0 = 0) = 0$) et $x = x_{\max}$ ($\phi(t;0,x_{\max}) = x_{\max}$) sont des points d'équilibres.

Si x_e est un point d'équilibre, alors pour que le système reste en ce point il faut que la vitesse soit nulle, c'est-à-dire que $g(x_e) = 0$. Cependant, cette condition seule n'est pas suffisante comme le montre l'étude de

$$\frac{dx}{dt} = |x|^{\frac{1}{2}}, \quad x \in \mathbb{R}, \tag{4.39}$$

on a bien $x_e = 0$ solution de $\sqrt{x} = 0$ mais il existe une infinité de solutions qui quittent ce point :

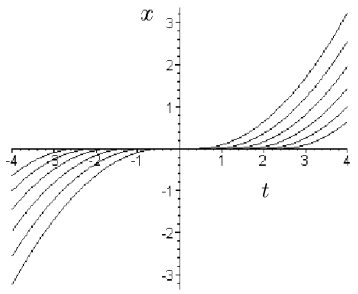


FIG. 4.6: Infinité de solutions de (4.39).

$$\begin{aligned} &\varepsilon \in \mathbb{R}_+, \phi_\varepsilon : \mathbb{R} \to \mathbb{R}, \\ &t \mapsto \phi_\varepsilon(t) = \begin{cases} 0 & \text{si} \quad t_0 - \varepsilon \le t \le t_0 + \varepsilon \\ \frac{(t-t_0-\varepsilon)^2}{4} & \text{si} \quad t_0 + \varepsilon \le t \\ -\frac{(t-t_0+\varepsilon)^2}{4} & \text{si} \quad t \le t_0 - \varepsilon \end{cases}. \end{aligned} \tag{4.40}$$

Pour lever ce problème

Théorème 4.5.1. *Si $g(x_e) = 0$ et g est localement Lipschitzienne en x_e alors x_e est un point d'équilibre pour le système (4.34).* ■

Par la suite, on considérera que le point d'équilibre est l'origine : en effet, l'étude de (4.34) au voisinage d'un point d'équilibre x_e se ramène, par changement de coordonnées $y = x - x_e$, à l'étude de $\dot{y} = g(y + x_e)$, ayant pour équilibre $(y = 0)$.

Définition 4.5.2. Dans la littérature, on classe les points d'équilibre de (4.34) en deux catégories :

1. les **points hyperboliques** (ou **non dégénérés**) : ce sont les points d'équilibres (x_e) pour lesquels la Jacobienne[10] correspondante $(J_g(x_e))$ ne comporte aucune valeur propre à partie réelle nulle.
2. les **points non hyperboliques** : ce sont les points d'équilibres (x_e) pour lesquels la Jacobienne correspondante $(J_g(x_e))$ possède au moins une valeur propre à partie réelle nulle.

Orbite périodique

L'étude des systèmes non linéaires a mis en évidence des orbites particulières :

1. les **orbites fermées** qui sont une extension des points fixes puisque si on laisse évoluer un système à partir d'une condition initiale appartenant à cette orbite, alors il continuera à évoluer sur cette orbite, (par exemple le cercle unité pour (4.36) cf. figure 4.5),
2. les **orbites homocliniques** et **hétérocliniques** qui relient des points d'équilibres (nous ne parlerons pas de ces dernières pour plus de détails voir [15])

Les définitions suivantes sont inspirées de [28] p. 87–88, [29] p. 8, [18] p. 113–117 et de [15].

Définition 4.5.3. La solution $\phi(t; t_0, x_0)$ est **T-périodique** (périodique de période T), si la solution est définie sur $\mathbb{R}$ et s'il existe un réel positif λ, tel que pour tout réel t on ait $\phi(t + \lambda; t_0, x_0) = \phi(t; t_0, x_0)$. Le plus petit réel positif λ noté T s'appelle la période de la solution. Dans ce cas l'orbite correspondante est une **orbite périodique** de période T (ou orbite T-périodique).

Définition 4.5.4. γ est une **orbite fermée** si γ est une orbite qui soit une courbe de Jordan, c'est-à-dire homéomorphe[11] à un cercle.

Toute orbite image d'une solution T-périodique non triviale (non identique à un point d'équilibre) est une orbite fermée.

[10] Si g est un champ de vecteurs sur $\mathbb{R}^n$ alors sa Jacobienne au point x est la matrice $\left(\frac{\partial g_i}{\partial x_j}(x)\right)$.

[11] Un homéomorphisme est un morphisme bijectif bicontinue. Ainsi, une courbe de Jordan c'est une courbe obtenue par transformation bijective bicontinue à partir du cercle.

Exemple 4.5.2. Si on reprend l'équation de Van der Pol (4.3) avec $f(x) = (x^3 - 2\mu x)$ en posant $i_L = -x_2, v_C = x_1, L = C = 1$, (4.3) devient :

$$\begin{aligned} \frac{dx_1}{dt} &= x_2, \\ \frac{dx_2}{dt} &= 2\mu x_2 - x_2^3 - x_1, \end{aligned} \tag{4.41}$$

ainsi, pour $\mu > 0$, on peut montrer (cf. [19] p. 211–227) l'existence d'une orbite périodique γ représentée sur la figure suivante

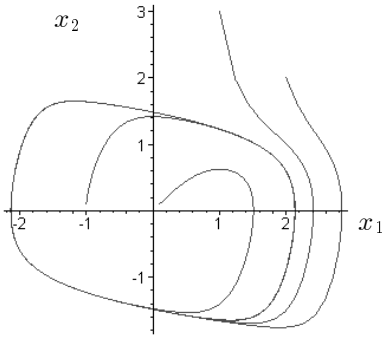


FIG. 4.7: Orbite fermée périodique pour l'oscillateur de Van der Pol (4.41).

Exemple 4.5.3. Le *système de Volterra-Lotka* est un modèle simple de lutte de deux espèces. En 1917, donc pendant la guerre, le biologiste Umberto d'Ancona constata une augmentation du nombre de sélaciens (requins) dans la partie nord de la mer Adriatique. Afin d'expliquer ce phénomène, il fit appel à son beau-père, le mathématicien Vito Volterra, qui expliqua ce phénomène de la façon suivante. Soit un volume d'eau infini (mer Adriatique par exemple), peuplé par deux espèces : l'une, carnivore (C : sélaciens), dévorant l'autre, herbivore (H : crevettes). Notons x et y les nombres respectifs d'individus des espèces (H) et (C). Si l'espèce (H) peuplait seule la mer, elle se développerait de façon exponentielle[12] et la vitesse de croissance de l'espèce (H) serait : $\frac{dx}{dt} = \alpha x$, avec $\alpha > 0$. Par contre, l'espèce (C) ne peut assurer seule son développement, ni même sa survie, donc sa vitesse de variation serait : $\frac{dy}{dt} = -\beta y$, avec $\beta > 0$. Lorsque les deux espèces cohabitent, les carnivores (C) dévorent les herbivores (H). En faisant l'hypothèse qu'à chaque rencontre d'un carnivore avec un herbivore, ce dernier est dévoré et que le nombre de rencontres est proportionnel au produit des densités volumiques des deux espèces (donc, aussi à xy), on peut conclure

[12]On fait ici l'hypothèse son développement n'est limité ni par l'espace ni par la quantité de nourriture.

que l'évolution des deux espèces est régie par le système différentiel :

$$\begin{cases} \frac{dx}{dt} = \alpha x - \gamma xy \quad \text{(herbivores)}, \\ \frac{dy}{dt} = -\beta y + \delta xy \quad \text{(carnivores)}, \end{cases} \tag{4.42}$$

avec $\alpha, \beta, \gamma, \delta$ des réels positifs. Dans ce cas, les variables d'état s'introduisent de façon naturelle : x, y. On peut a priori supposer que l'espace d'état est le quart de plan $\mathbb{R}^2_+$. Le théorème 4.3.3 permet de garantir l'existence et l'unicité des solutions. En séparant les variables selon : $\frac{dx}{x(\alpha-\gamma y)} = \frac{dy}{y(-\beta+\delta x)}$, on peut monter que $H(x,y) = [\alpha \ln(y) - \gamma y] + [\beta \ln(x) - \delta x]$ est une fonction constante le long des solutions de (4.42). On montre ainsi que, pour toute condition initiale strictement incluse dans le quart de plan strictement positif, les orbites du système sont fermées. De plus les solutions sont définies sur $\mathbb{R}$: on obtient un flot dont le portrait de phase est représenté sur la figure 4.8 (simulation pour $\alpha = \beta = \gamma = \delta = 1$).

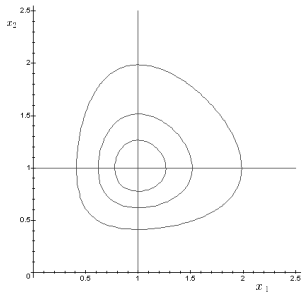


FIG. 4.8: Cycle pour (4.42).

Les orbites sont centrées autour du point d'equilibre $(\frac{\beta}{\delta}, \frac{\alpha}{\gamma})$. Avant la guerre, l'activité de la pêche était plus importante (on tient compte des prélèvements de la pêche "$-q_x x$" et "$-q_y y$" dans (4.42), avec q_x, q_y positifs) : c'est-à-dire que le couple de paramètres $(\alpha, -\beta)$ est remplacé par $(\alpha - q_x, -\beta - q_y)$, donc le point d'equilibre $(\frac{\beta}{\delta}, \frac{\alpha}{\gamma})$ est remplacé par $(\frac{\beta+q_y}{\delta}, \frac{\alpha-q_x}{\gamma})$. Ce qui explique un déplacement du cycle vers le haut pendant la guerre, donc une augmentation du nombre de célaciens.

Concepts de Stabilité et d'Attractivité

Dans cette section on se place dans le cadre d'une EDO

$$\frac{dx}{dt} = f(t,x), \quad x \in \mathcal{X}, \quad t \in \mathbb{R}. \tag{4.43}$$

Stabilité au sens de Liapunov d'un équilibre

Les ensembles remarquables (équilibres, orbites périodiques, etc...) peuvent caractériser des configurations à énergie minimale pour un système physique. Ces systèmes peuvent avoir tendance à rechercher une position de repos plutôt qu'une autre : c'est ce que les concepts de stabilité traduisent d'une certaine façon. Par exemple, un pendule pesant (voir introduction équation (4.6)), possède deux équilibres verticaux : l'un au dessus de l'horizontale, l'autre en dessous. Il est bien connu que la masse a naturellement tendance à se positionner en bas plutôt qu'en haut. La position d'équilibre basse est stable, l'autre instable. Par la suite nous ne présenterons que les concepts élémentaires pour les points d'équilibre et uniquement pour des EDO du type (4.34).

Définition 4.5.5. L'équilibre x_e est **stable au sens de Liapunov** si :

$$\begin{aligned} &\forall \varepsilon > 0, \exists \delta(t_0, \varepsilon) > 0 \text{ tel que :} \\ &\forall x_0 \in \mathcal{X} : \rho(x_0, x_e) \leq \delta(t_0, \varepsilon) \Rightarrow \rho(\phi(t; t_0, x_0), x_e) \leq \varepsilon, \forall t \geq t_0. \end{aligned} \tag{4.44}$$

Lorsque $\delta(t_0, \varepsilon) = \delta(\varepsilon)$ est indépendant de t_0 la propriété de stabilité sera dite **uniforme** . (ρ est une distance sur $\mathbb{R}^n$).

Cela revient à dire que pour tout voisinage $\mathcal{V}(x_e)$ de x_e, il existe un voisinage $\mathcal{W}(x_e)$ de x_e tel que : $x_0 \in \mathcal{W}(x_e) \Rightarrow \phi(t; t_0, x_0) \in \mathcal{V}(x_e), \forall t \geq t_0$ (voir [3] p. 58).

Exemple 4.5.4. L'origine $x = 0$ est un équilibre uniformément stable (par la suite, on omettra "au sens de Liapunov") pour l'EDO $\dot{x} = -x, x \in \mathbb{R}$: en effet $\forall x_0 \in \mathbb{R}, \phi(t; t_0, x_0) = \exp - (t - t_0)\, x_0$ donc $|\phi(t; t_0, x_0)| \leq |x_0| \leq \varepsilon, \forall t \geq t_0$, en prenant $|x_0| \leq \varepsilon$ ($\delta = \varepsilon$ dans la définition 4.5.5). Par contre, il n'est que stable pour l'EDO $\dot{x} = -\frac{2t}{1+t^2} x, x \in \mathbb{R}$: en effet : $\forall x_0 \in \mathbb{R}, \phi(t; t_0, x_0) = \frac{1+t_0^2}{1+t^2} x_0$ donc $|\phi(t; t_0, x_0)| \leq \left(1 + t_0^2\right) |x_0| \leq \varepsilon, \forall t \geq t_0$, en prenant $|x_0| \leq \delta = \frac{\varepsilon}{1+t_0^2}$.

Attractivité

Alors que la propriété de stabilité traduit le "non-éloignement" de solutions d'un ensemble de référence (par exemple un équilibre), la propriété d'attractivité elle traduit le rapprochement des solutions de cet ensemble et ce malgré d'éventuelles excursions pendant le régime transitoire.

Définition 4.5.6. L'équilibre x_e est **attractif** si :

$$\exists \delta > 0 \text{ tel que } \forall x_0 \in \mathcal{X} : \rho(x_0, x_e) \leq \delta \Rightarrow \lim_{t \to \infty} \phi(t; t_0, x_0) = x_e. \tag{4.45}$$

Le second membre de l'implication s'écrit aussi $\forall \varepsilon > 0, \exists T(x_0, \varepsilon) > 0$, $\rho(\phi(t; t_0, x_0), x_e) \leq \varepsilon$, $\forall t \geq t_0 + T(t_0, x_0, \varepsilon)$. Lorsque $T(\varepsilon, x_0) = T(\varepsilon)$ est indépendant de x_0, la propriété d'attractivité sera dite **uniforme**.

De même que pour la notion de stabilité, cette notion peut être formulée en termes de voisinages.

Stabilité asymptotique

La notion d'attractivité d'un ensemble assure que l'état va converger vers cet ensemble, mais il se peut que pendant le régime transitoire il y ait de grosses excursions des solutions ce qui peut être dommageable pour le système physique. Aussi, la stabilité limitant ces excursions on combine les deux notions précédentes pour donner naissance à la notion de **stabilité asymptotique**.

Il est important de noter qu'un ensemble peut être attractif sans être stable comme le montre l'exemple suivant

Exemple 4.5.5. Soit l'EDO

$$\frac{dx}{dt} = x\left(1-\sqrt{x^2+y^2}\right) - \frac{y}{2}\left(1-\frac{x}{\sqrt{x^2+y^2}}\right),$$
$$\frac{dy}{dt} = y\left(1-\sqrt{x^2+y^2}\right) + \frac{x}{2}\left(1-\frac{x}{\sqrt{x^2+y^2}}\right),$$

l'origine est un équilibre instable (pour le montrer, on pourra utiliser les résultats de la section 4.5) et l'équilibre $(1,0)$ est attractif mais instable : le portrait de phase est donné Figure 4.9.

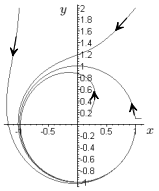


FIG. 4.9: Equilibre $(1,0)$ attractif et instable.

Exemple 4.5.6. La solution au PC (4.1) $x(0) = x_0$ est donnée par (4.38), il est facile de vérifier que l'équilibre $x = 0$ n'est pas attractif ($\lim_{t\to\infty}\phi(t;0,x_0) = \lim_{t\to\infty}\frac{x_0 x_{\max}}{x_0+e^{-ax_{\max}t}(x_{\max}-x_0)} = x_{\max}$) et que l'équilibre $x = x_{\max}$ est asymptotiquement stable. En effet, il est attractif ($\lim_{t\to\infty}\phi(t;0,x_0) = x_{\max}$) et stable puisque $\phi(t;0,x_0) - x_{\max} = \frac{x_{\max}(x_{\max}-x_0)e^{-ax_{\max}t}}{x_0+(x_{\max}-x_0)e^{-ax_{\max}t}}$ donc pour $\varepsilon > 0$, si $|x_0 - x_{\max}| < \varepsilon$, $|\phi(t;0,x_0) - x_{\max}| < x_{\max}\left|\frac{(x_{\max}-x_0)}{x_0+(x_{\max}-x_0)}\right| < \varepsilon$.

Pour cet exemple nous avons pu étudier la stabilité asymptotique à partir de l'expression analytique des solutions. Cependant, pour une EDO (4.34) dont en général on ne peut exprimer les solutions de façon explicite, il serait important de disposer de critères permettant d'étudier la question de la stabilité sans avoir à calculer les solutions : ce sont les résultats de la section 4.5 qui le permettent : première méthode de Liapunov et seconde méthode de Liapunov (Théorème 4.5.8) .

Stabilité d'équilibres : Résultats élémentaires

Les premiers travaux sur la stabilité ne retenaient des EDO que leur approximation linéaire du premier ordre. Il fallut attendre quelques années pour que H. Poincaré et A.M. Liapunov justifient et étendent les propriétés locales déduites du modèle linéarisé. L'un des résultats principaux est la première méthode de Liapunov (voir Théorème 4.5.10) : si l'origine est uniformément asymptotiquement stable pour le linéarisé alors il est localement uniformément asymptotiquement stable pour le système non linéaire. Cependant elle ne donne aucun renseignement quantitatif sur le domaine de conditions initiales conduisant à la stabilité asymptotique. Cette lacune fût contournée par l'introduction des célèbres fonctions de Liapunov : c'est la seconde méthode de Liapunov. D'une part, les fonctions de Liapunov sont analogues à des distances entre l'état du système le long de sa trajectoire et l'ensemble ou la trajectoire étudiée (point d'équilibre, etc... qui traduit une configuration d'énergie minimale). D'autre part, ces fonctions ont une relation directe avec la physique des systèmes puisque très souvent elles ne sont rien de plus que l'expression de l'énergie totale du système qui, s'il est dissipatif, décroît au cours du temps afin que le système rejoigne une configuration à énergie minimale (s'il n'y a pas d'apport d'énergie).

Par exemple, le pendule pesant dont un modèle est donné par

$$\ddot{\theta} = -\frac{\delta}{ml^2}\dot{\theta} - \frac{g}{l}\sin(\theta), \tag{4.46}$$

(l, m, g, δ positifs), a deux équilibres : la position haute et basse. Il est bien connue que seule la position basse est stable : elle correspond à une configuration à énergie minimale (Energie potentielle nulle si on prend cette position basse comme référence). En effet, l'énergie totale du système est :

$$V(\theta, \dot{\theta}) = \frac{1}{2}ml^2\dot{\theta}^2 + mgl(1 - \cos(\theta)), \tag{4.47}$$

(notons que $V(\theta = 0, \dot{\theta} = 0) = 0$ et que $V(\theta, \dot{\theta}) > 0$ pour $(\theta, \dot{\theta}) \neq (0, 0)$) ; ce qui conduit à

$$\frac{dV}{dt} = ml^2(-\frac{\delta}{ml^2}\dot{\theta} - \frac{g}{l}\sin(\theta))\dot{\theta} + mgl\sin(\theta)\dot{\theta} = -\delta\dot{\theta}^2 \leq 0, \tag{4.48}$$

ce qui montre que l'énergie du système décroît au cours du temps : le système tend à rejoindre une configuration à énergie minimale.

Résultat de stabilité pour un système linéaire

Pour un modèle linéaire (4.32), les solutions sont données par (4.33), ainsi le comportement des trajectoires est entièrement conditionné par les dilatations et contractions engendrées par l'exponentielle de la matrice A ; on en déduit :

Théorème 4.5.2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$, de spectre $\sigma(A) = \{\lambda_i \in \mathbb{C}, i = 1, \ldots, r \leq n : \det(\lambda_i Id - A) = 0$ et $\lambda_i \neq \lambda_j$ pour $i \neq j\}$ et $\nu(\lambda_i)$ le plus petit entier tel que $\ker(\lambda_i Id - A)^{\nu(\lambda_i)+1} = \ker(\lambda_i Id - A)^{\nu(\lambda_i)}$, $(i = 1, \ldots, r)$:*

1. *$\exists \lambda_i \in \sigma(A) : \mathrm{Re}(\lambda_i) > 0$ alors $\lim_{t\to+\infty} \|\exp(At)\| = +\infty$ et l'origine de (4.32) est instable,*
2. *$\exists \lambda_i \in \sigma(A) : \mathrm{Re}(\lambda_i) = 0$ et $\nu(\lambda_i) > 1$ alors $\lim_{t\to+\infty} \|\exp(At)\| = +\infty$ et l'origine de (4.32) est instable,*
3. *$\mathrm{Re}(\lambda_i) < 0$, $i = 1, \ldots, (r-1)$ et $\mathrm{Re}(\lambda_r) = 0$ avec $\nu(\lambda_i) = 1$ alors $\|\exp(At)\| < +\infty$ et l'origine de (4.32) est stable mais pas attractive,*
4. *$\forall \lambda_i \in \sigma(A) : \mathrm{Re}(\lambda_i) < 0$ alors $\lim_{t\to+\infty} \|\exp(At)\| = 0$ et l'origine de (4.32) est asymptôtiquement stable.* ■

Démonstration. Cela découle de la décomposition sous forme de Jordan de la matrice A : il existe une matrice régulière P telle que

$$A = PJP^{-1}, J = \mathrm{diag}(J(\lambda_i)), J(\lambda_i) = \begin{pmatrix} \lambda_i & 1 & 0 & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & \lambda_i \end{pmatrix}, \tag{4.49}$$

d'où $\exp(At) = P\exp(Jt)P^{-1}$ avec $\exp(Jt) = \mathrm{diag}(\exp(J(\lambda_i)t))$ et

$$\exp(J(\lambda_i)t) = \exp(\lambda_i t) \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \frac{t^{(k-1)}}{(k-1)!} \\ 0 & \ddots & \ddots & \frac{t^2}{2!} \\ \vdots & \ddots & \ddots & t \\ 0 & \cdots & 0 & 1 \end{pmatrix}, \tag{4.50}$$

les différents résultats en découlent tout naturellement. □

On en déduit la condition nécessaire et suffisante pour que l'origine soit asymptotiquement stable pour un système linéaire autonome :

Corollaire 4.5.1. *L'origine de (4.32) est asymptotiquement stable $\Leftrightarrow \forall \lambda_i \in \sigma(A) : \mathrm{Re}(\lambda_i) < 0$.* ■

Corollaire 4.5.2. *Si le polynôme caractéristique de A est de la forme $\pi_A(x) = x^n + \sum_{i=0}^{n-1} a_i x^i$, alors une condition nécessaire de stabilité de l'origine de (4.32) est que les a_i soient tous positifs.* ■

Démonstration. Si un $a_i < 0$, cela signifie qu'au moins un produit de valeurs propres est négatif donc qu'au moins deux valeurs propres sont de signes contraires. □

Structure locale des solutions au voisinage d'un point d'équilibre

Pour les points d'équilibre hyperboliques, les résultats suivants permettent d'éclaircir le comportement local des solutions de (4.34).

Théorème 4.5.3 (de Hartman-Grobman, 1964)**.** *Si la jacobienne $J_g(x_e) = A$ au point d'équilibre x_e n'a pas de valeur propre purement imaginaire ou nulle ($\sigma_c(A) = \emptyset$), alors il existe un homéomorphisme h défini dans un voisinage $\mathcal{V}(x_e)$ de x_e, envoyant localement les orbites du flot linéaire vers celles du flot non linéaire Φ_g^t de (4.34). De plus, h préserve le sens de parcours des orbites et peut être choisi de façon à préserver la paramétrisation du temps.*

A partir du voisinage $\mathcal{V}(x_e)$ où h est défini, on construit les **variétés locales stable** et **instable** :

$$W_{loc\ s}(x_e) = \{x \in \mathcal{V}(x_e) : \lim_{t\to+\infty} \Phi_g^t(x) = x_e \text{ et } \Phi_g^t(x) \in \mathcal{V}(x_e), \forall t > 0\},$$
$$W_{loc\ i}(x_e) = \{x \in \mathcal{V}(x_e) : \lim_{t\to-\infty} \Phi_g^t(x) = x_e \text{ et } \Phi_g^t(x) \in \mathcal{V}(x_e), \forall t > 0\},$$

à partir desquelles on définit les **variétés stable** et **instable** (relatives à x_e) :

$$W_s(x_e) = \cup_{t\geq 0} \Phi_g^t(W_{loc\ s}(x_e)),$$
$$W_i(x_e) = \cup_{t\leq 0} \Phi_g^t(W_{loc\ i}(x_e)).$$

Ces notions de variétés stable et instable exhibent donc des solutions de (4.34) qui sont respectivement ‹‹ contractantes ›› et ‹‹ dilatantes ››. Les variétés $W_s(x_e)$, $W_i(x_e)$ sont les images par h des sous-espaces correspondants sur le linéarisé : $W_s(x_e) = h[E_s(J_g(x_e))]$, $W_i(x_e) = h[E_i(J_g(x_e))]$.

Théorème 4.5.4 (de la variété stable)**.** *Si (4.34) a un point d'équilibre hyperbolique x_e, alors il existe $W_s(x_e)$ et $W_i(x_e)$:*

1. *de dimension n_s et n_i identiques à celles des espaces $E_s(J_g(x_e))$ et $E_i(J_g(x_e))$ du système linéarisé (avec $A = J_g(x_e)$),*
2. *tangentes à $E_s(J_g(x_e))$ et à $E_i(J_g(x_e))$ en x_e,*
3. *invariantes par le flot Φ_g^t.*

De plus, $W_s(x_e)$ et $W_i(x_e)$ sont des variétés aussi régulières que g (de même classe r que $g \in \mathcal{C}^r(R^n)$).

Dans le cas, dit critique, de points non hyperboliques (dégénérés), il a été montré le résultat suivant (voir [15] p. 127).

Théorème 4.5.5 (de la variété centre)**.** *(Kalley, 1967) Soit g un champ de vecteurs de classe $\mathcal{C}^r(\mathbb{R}^n)$, admettant un point d'équilibre dégénéré x_e. Soit $A = J_g(x_e)$. Alors, il existe :*

1. *$W_s(x_e)$ et $W_i(x_e)$ des variétés invariantes dites respectivement stable et instable de classe $\mathcal{C}^r$, tangentes à $E_s(J_g(x_e))$ et à $E_i(J_g(x_e))$ en x_e ;*
2. *$W_c(x_e)$ une variété centre de classe $\mathcal{C}^{(r-1)}$ tangente à $E_c(J_g(x_e))$ en x_e.*

Les variétés $W_s(x_e)$, $W_i(x_e)$ et $W_c(x_e)$ sont toutes invariantes par le flot Φ_g^t et de même dimension que les sous-espaces correspondants du système linéarisé ($E_s(J_g(x_e))$, $E_i(J_g(x_e))$ et $E_c(J_g(x_e))$). Les variétés stable $W_s(x_e)$ et instable $W_i(x_e)$ sont uniques, alors que $W_c(x_e)$ ne l'est pas forcément.

Cependant, de façon pratique, il est délicat d'obtenir ces variétés, même de façon numérique : souvent, le seul recours pour la détermination d'une variété centre est de faire un développement en série de Taylor de $W_c(x_e)$ au voisinage du point dégénéré x_e : cette méthode est connue depuis longtemps puisque A.M. Liapounov l'a utilisée en 1892 pour étudier les ‹‹ cas critiques ›› [24].

Pour des raisons de simplification, on effectue un changement de coordonnées sur le système initial (4.34) pour se ramener au cas où le point d'équilibre est l'origine. On va regarder ce qui se passe dans le cas le plus intéressant en pratique, c'est-à-dire $W_i(0)$ vide. Le théorème de la variété centre nous dit que le système initial (4.34) est topologiquement équivalent à :

$$\begin{cases} \frac{dx_c}{dt} = A_c x_c + g_1(x), \\ \frac{dx_s}{dt} = A_s x_s + g_2(x), \end{cases}$$

avec A_c de dimension n_c correspondant à $E_c(J_g(0))$ et qui a donc toutes ses valeurs propres à partie réelle nulle. A_s est de dimension n_s correspondant à $E_s(J_g(0))$, donc asymptotiquement stable. On peut exprimer $W_c(0)$ sous la forme d'une hypersurface :

$$W_c(0) = \{(x_c, x_s) \in \mathbb{R}^{n_c} \times \mathbb{R}^{n_s} : x_s = k(x_c)\}.$$

De plus, on sait que $W_c(0)$ contient 0 (donc $k(0) = 0$) et, en ce point, est tangent à $E_c(J_g(0))$ (donc $J_k(0) = 0$). On a :

$$x_s = k(x_c) \Rightarrow \frac{dx_s}{dt} = J_k(x_c)\frac{dx_c}{dt},$$

donc :

$$A_s x_s + g_2(x_c, k(x_c)) = J_k(x_c)\left(A_c x_c + g_1(x_c, k(x_c))\right), \tag{4.51}$$

$$k(0) = 0, \quad J_k(0) = 0. \tag{4.52}$$

On étudie la projection du champ de vecteurs de $x_s = k(x_c)$ sur $E_c(J_g(0))$:

$$\frac{dx_c}{dt} = A_c x_c + g_1(x_c, k(x_c)), \tag{4.53}$$

en tenant compte de (4.51) et de (4.52). Ce qui nous conduit au théorème suivant (voir [15] p. 131).

Théorème 4.5.6 (de Henry et Carr, 1981)**.** *Si :*

1. $W_i(0)$ *est vide,*
2. *l'équilibre* $x_{ec} = 0$ *de (4.53) est localement asymptotiquement stable (respectivement instable),*

alors l'équilibre x_e *de (4.34) est asymptotiquement stable (respectivement instable).*

La résolution de (4.53) étant en général impossible, le théorème suivant [15] permet d'étudier la stabilité locale de l'équilibre $x_{ec} = 0$ par approximation de k.

Théorème 4.5.7 (de Henry et Carr, 1981)**.** *S'il existe* $\psi : \mathbb{R}^{n_c} \to \mathbb{R}^{n_s}$ *avec* $\psi(0) = 0$ *et* $J_\psi(0) = 0$*, telle que, lorsque* $x \to 0$ *:*

$$J_\psi(x_c)[A_c x_c + g_1(x_c, \psi(x_c))] - A_s\psi(x_c) - g_2(x_c, \psi(x_c)) = o(x^r),\ r > 1, \tag{4.54}$$

alors $h(x_c) = \psi(x_c) + o(x^r)$*, lorsque* $x \to 0$*.*

Cette technique permet, dans beaucoup de cas, de conclure sur la stabilité asymptotique d'un équilibre dégénéré.

Exemple 4.5.7. Soit le système différentiel $(x, y) \in \mathbb{R}^2$:

$$\begin{aligned} \frac{dx}{dt} &= -x^2 + xy, \\ \frac{dy}{dt} &= -y + x^2. \end{aligned} \tag{4.55}$$

On a :

$$J_g(x, y) = \begin{pmatrix} -2x + y & x \\ 2x & -1 \end{pmatrix},$$

et le système présente deux points d'équilibre :

$$z_{e1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \text{dégénéré},\ J_g(z_{e1}) = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix},$$

$$z_{e2} = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\ \text{instable},\ J_g(z_{e2}) = \begin{pmatrix} -1 & 1 \\ 2 & -1 \end{pmatrix}.$$

Pour l'origine, les valeurs propres associées à la jacobienne sont 0 et -1 (on a $A_c = 0, A_s = -1$). On cherche alors la variété centre associée à ce point d'équilibre par son développement à l'ordre 3 : $k(x) = ax^2 + bx^3 + o(x^3)$, puisque $k(0) = J_k(0) = 0$. Ce développement doit vérifier (4.54), donc :

$$[2ax + 3bx^2 + o(x^2)][-x^2 + (ax^3 + bx^4 + o(x^4))] = [(1-a)x^2 - bx^3 + o(x^3)],$$

et, en égalant les termes de même degrés, on obtient $a = 1, b = 2$, soit : $k(x) = x^2 + 2x^3 + o(x^3)$. Donc (4.53) devient $\dot{x} = -x^2 + x^3 + o(x^3)$ et le théorème 4.5.6 permet de conclure à l'instabilité l'origine. Remarquons que le même résultat peut être obtenu plus intuitivement et sans trop de calcul, en notant que la seconde ligne (en y) de (4.55) converge beaucoup plus vite (exponentiellement) que la première (en x) : on peut donc considérer qu'après un transitoire, $\frac{dy}{dt} = 0 = -y + x^2$, soit $y = x^2$: on retrouve la variété centre $k(x) = x^2 + o(x^2)$.

Exemple 4.5.8. Soit le système différentiel :

$$\begin{aligned} \frac{dx}{dt} &= xy, \\ \frac{dy}{dt} &= -y - x^2, \end{aligned}$$

avec $(x, y) \in \mathbb{R}^2$. L'origine est le seul point d'équilibre. Les valeurs propres associées à la jacobienne sont 0 et -1. Un développement à l'ordre 3 de $k(x)$ est $-x^2 + o(x^3)$. Le théorème 4.5.6 nous permet de conclure que l'origine est asymptotiquement stable (mais non exponentiellement stable).

Remarque 4.5.1. Il existe une façon rapide de traiter ces deux exemples : au voisinage de l'origine, y converge exponentiellement, donc ‹‹ infinement plus rapidement ›› que x ne le ferait. On en déduit que $\frac{dy}{dt}$ s'annule ‹‹ infiniement ›› plus vite que $\frac{dx}{dt}$ soit, pour l'exemple 4.5.7, $y = x^2$, qui reporté dans $\frac{dx}{dt} = -x^2 + y$ donne bien l'équation approchée $\frac{dx}{dt} = -x^2 + x^3$. De même, l'exemple 4.5.8 conduit, quand $t \to \infty$, à avoir $y \to -x^2$, donc $\frac{dx}{dt} = -x^3$.

Méthodes de Liapunov

Comme nous l'avons vu en introduction dans cette section, les fonctions de Liapunov permettent de suivre l'évolution temporelle de l'état du système.

Définition 4.5.7. Une fonction $V : \mathcal{O} \subset \mathbb{R}^n \mapsto \mathbb{R}_+$ où $\mathcal{O}$ est un ouvert de $\mathbb{R}^n$ contenant l'équilibre $x_{\text{équ}}$, est dite de **Liapunov** pour un équilibre $x_{\text{équ}}$ ssi elle est continue, définie positive (i.e. $V(x) = 0 \Leftrightarrow x = x_{\text{équ}}$ et $V(\mathcal{O}) \subset \mathbb{R}_+$) et possède une dérivée. Une fonction de Liapunov est dite **radialement non bornée** si $\lim_{\|x\| \to +\infty} V(x) = +\infty$.

Par exemple, pour le pendule pesant (4.6), la fonction définie par (4.47) est bien une fonction de Liapunov radialement non bornée pour l'origine.

Il est alors concevable d'en conclure que si V décroît le long des trajectoires alors V va rejoindre un minimum qui correspond à ce que l'état ait rejoint l'équilibre : c'est ce que traduisent les résultats qui suivent.

Théorème 4.5.8. *Soit l'EDO (4.34), $\mathcal{O}$ un ouvert de $\mathbb{R}^n$ contenant l'origine, $V : \mathcal{O} \to \mathbb{R}_+$ de classe $\mathcal{C}^1$ telle que $V(x) = 0 \Leftrightarrow x = 0$ et les conditions suivantes C1) $\frac{dV}{dt}\big|_{(4.34)} = \frac{\partial V}{\partial x} g(x) \leq 0$, $\forall x \in \mathcal{O}$, C2) $\frac{\partial V}{\partial x} g(x) = 0 \Leftrightarrow x = 0$ Si C1) est vraie alors l'origine de (4.34) est localement stable. Si C1 et C2) sont vraies alors l'origine de (4.34) est localement asymptotiquement stable. Si dans les deux résultats précédents, $\mathcal{O} = \mathbb{R}^n$ et V est radialement non bornée alors les conclusions sont globales, en particulier pour le second résultat on pourra conclure que l'origine de (4.34) est globalement asymptotiquement stable.* ■

Démonstration. Soit $\mathbf{O} \subset \mathcal{O}$ un ouvert de l'origine et $\mathbf{C}$ un compact $\subset \mathcal{O}$ (contenant $\mathbf{O}$) alors $\mathbf{C} \cap (\mathcal{O} \backslash \mathbf{O})$ est compact et V étant continue elle y atteint un minimum $\varepsilon_{\min} : V_{\varepsilon_{\min}} \subset \mathbf{O}, V_\varepsilon = \{x \in \mathcal{O} : V(x) \leq \varepsilon\}$. De plus $\frac{dV(x(t))}{dt} \leq 0$ donc si $x_0 \in V_\varepsilon : \phi(t; t_0, x_0) \in V_\varepsilon$: l'origine est donc stable. Sur $V_{\varepsilon_{\min}} \subset \mathcal{O}$ compact, $V(t)$ est décroissante minorée par 0, donc elle admet une limite $l \leq \varepsilon_{\min}$. Si $l > 0$ et $\varepsilon > 0$ alors $V_{\varepsilon+l}$ est compact ($\subset \mathcal{O}$ pour ε suffisamment petit) donc $\dot{V}$ étant continue elle admet un maximum $-m$. Soit $x_0 \in V_{\varepsilon+l}, \phi(t; t_0, x_0) \in V_{\varepsilon+l}$

$$\lim_{t\to\infty} V(\phi(t; t_0, x_0)) = l = V(x_0) + \int_0^\infty \dot{V}(t)dt \leq V(x_0) - m \lim_{t\to\infty}(t) < 0. \quad (4.56)$$

D'où la contradiction. Le dernier point en découle. □

Exemple 4.5.9. Si on considère (4.36) dans laquelle le second membre de l'EDO est remplacé par son opposé, alors en prenant $V(x) = \frac{1}{2}(x_1^2 + x_2^2)$ on obtient $\dot{V} = (x_1^2 + x_2^2 - 1)(x_1^2 + x_2^2)$, on en conclut que l'origine est localement asymptotiquement stable.

Pour un système linéaire (4.32), on déduit du Théorème 4.5.8 et d'un autre résultat permettant de montrer sous certaines conditions (ici vérifiées) la réciproque du Théorème 4.5.8, le résultat suivant

Théorème 4.5.9. *Soit Q une matrice symétrique définie positive quelconque. L'origine de $\dot{x} = Ax, x \in \mathbb{R}^n$, est globalement asymptotiquement stable ssi il existe P une matrice symétrique définie positive solution de l'**équation de Liapunov***

$$PA + A^T P = -Q. \quad (4.57)$$

Dans ce cas, $V = x^T Px$ est une fonction de Liapunov radialement non bornée analytique telle que $\frac{dV}{dt}\big|_{(4.32)} = \frac{\partial V}{\partial x} Ax$ est définie négative. ■

A partir duquel on déduit :

Théorème 4.5.10. *Soit le système (4.34) tel que $g(0) = 0$, en notant $A = J_g(0)$: s'il existe P une matrice symétrique définie positive solution de (4.57) pour Q une matrice symétrique définie positive donnée, alors l'origine de (4.34) est localement asymptotiquement stable. Cela signifie que si l'origine est asymptotiquement stable (respectivement instable) pour le système $\dot{x} = J_g(0)x$ (dit linéarisé) alors il en est de même pour le système initial (4.34).* ■

Démonstration. Soit $V = x^T Px$, puisque $g(x) = Ax + o(\|x\|)$ on en déduit $\frac{\partial V}{\partial x} g(x) = -x^T Qx + o(\|x\|^2)$. Pour $\varepsilon < \lambda_{\min}(Q)$ donné, on peut trouver une boule de centre l'origine telle que $o(\|x\|^2) \leq \varepsilon \|x\|^2$ donc sur cette boule $\frac{\partial V}{\partial x} g(x) \leq (\varepsilon - \lambda_{\min}(Q)) \|x\|^2 \leq 0$. Le Théorème 4.5.8 permet de conclure. □

Exemple 4.5.10. Soit le système

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x - x^3 + xy - 2y. \end{aligned} \tag{4.58}$$

L'origine est un équilibre,

$$A = J_g(0) = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix}, \pi_A(x) = (x+1)^2,$$

$$PA + A^T P = -I, P = \frac{1}{2} \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}.$$

On en déduit que l'origine de (4.58) est asymptotiquement stable.

Principe d'invariance La Salle

Pour un pendule pesant amorti (4.46), la fonction (4.47) a pour dérivée (4.48), permettant seulement de conclure que l'origine est globalement stable. Cependant, nous savons que la position basse est aussi attractive : le pendule tend à rejoindre cette position sous l'effet d'une dissipation d'énergie par frottement. Pour le prouver mathématiquement, il nous faut une nouvelle fonction de Liapounov : c'est le principe d'invariance de La Salle qui nous permettra (sans changer de fonction de Liapounov) d'aboutir au résultat. Avant de donner ce résultat, examinons de plus près l'exemple du pendule pesant amorti. En posant $x = (x_1, x_2)^T = (\theta, \dot{\theta})^T$, (4.46) devient :

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -\frac{\delta}{ml^2} x_2 - \frac{g}{l} \sin(x_1), \end{cases}$$

et (4.47) donne $\frac{dV}{dt} = -\delta x_2^2$. Or, nous savons que V décroît sauf lorsque x_2 reste identiquement nulle (car $\dot{V} \equiv 0$). Mais, dans ce cas, $\dot{x}_2(t)$ reste identiquement

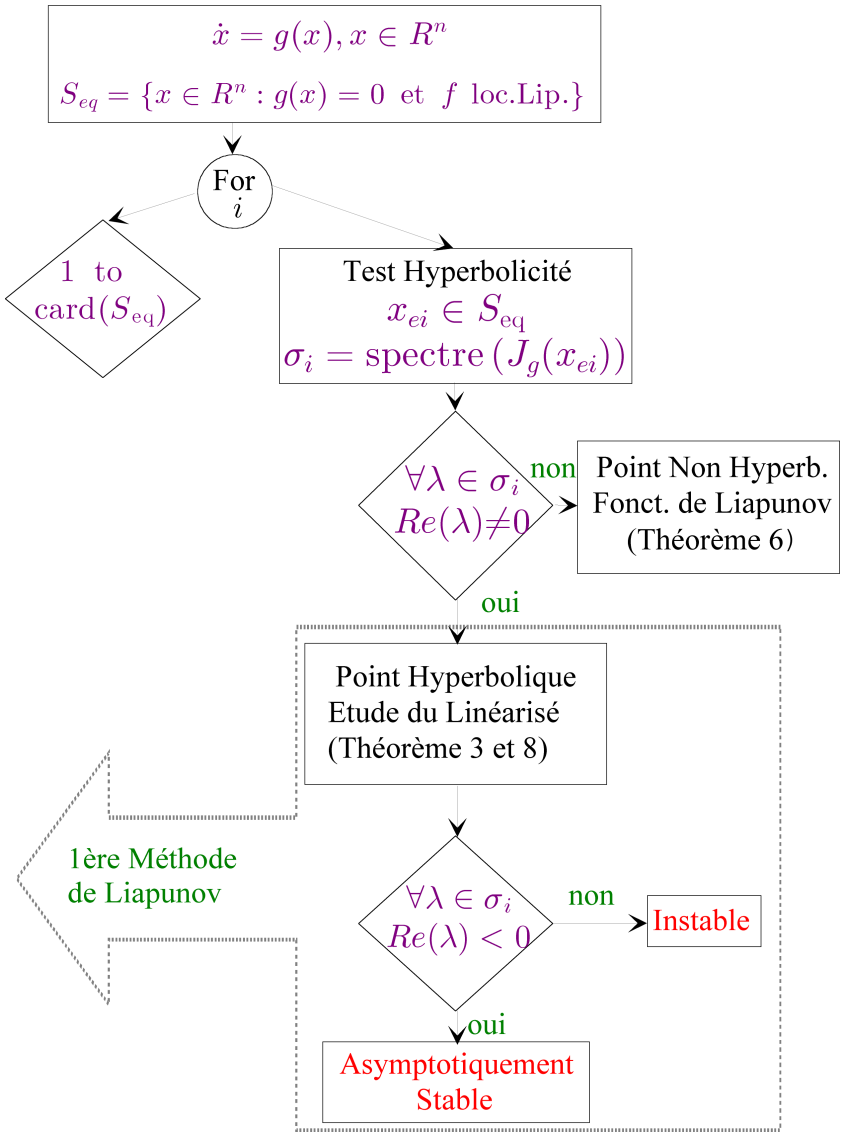


FIG. 4.10: Méthodologie

nulle, ce qui implique que $\sin(x_1)$ aussi. Ainsi $V(t)$ ne peut rejoindre sa valeur de repos (lorsque $\dot{V} \equiv 0$) que si le pendule est au repos en position basse $x = 0$ (si on démarre d'une position dans l'ensemble $\mathbf{C} = \{x : V(x) \leq 2mgl\}$ qui est positivement invariant et qui contient le segment $\{x_2 = 0, |x_1| < \pi\}$). Ce raisonnement, qui consiste à ne retenir, parmi les états annulant $\dot{V}$, que ceux qui sont invariants, est formalisé dans le théorème suivant.

Théorème 4.5.11 (Principe d'invariance de La Salle)**.** *Soit l'EDO (4.34),* $\mathbf{C} \subset \mathbb{R}^n$ *un ensemble compact positivement invariant et* $V : \mathbf{C} \subset \mathbb{R}^n \to \mathbb{R}$ *(non*

nécessairement définie positive) de classe $\mathcal{C}^1$ *telle que* $\forall x \in \mathbf{C} \ : \ \dot{V}(x) \leq 0$ *(négative mais non nécessairement définie). Alors toute solution initialisée dans* $\mathbf{C}$ *converge asymptotiquement vers* $\mathbf{I}$ *le plus grand ensemble invariant inclus dans* $\{x \in \Omega \ : \ \dot{V}(x) = 0\}$.

Démonstration : soit $x_0 \in \mathbf{C}$. Il faut montrer que $\Omega_g(x_0) \subset \mathbf{I}$ ($\mathbf{I} \subset \{x \in \mathbf{C} : \dot{V}(x) = 0\} \subset \mathbf{C}$). $\mathbf{C}$ étant invariant, $\Omega_g(x_0) \subset \mathbf{C}$ (car $\forall t : \Phi_g^t(x) \in \mathbf{C}$). $\dot{V}(x) \leq 0$ sur $\mathbf{C}$ donc V décroissante minorée (V étant continue et $\mathbf{C}$ compact elle admet un maximum et un minimum voir [27] chapitre 1) donc convergente vers une limite l : $\lim_{t\to\infty} V(\Phi_g^t(x_0)) = l$, d'où $\forall x \in \Omega_g(x_0) : V(x) = l$, donc $V(\Phi_g^t(x)) = \lim_{t_i\to\infty} V(\Phi_g^{t+t_i}(x_0)) = l$. On a ainsi $\dot{V} = 0$, d'où $\Omega_g(x_0) \subset \{x \in \mathbf{C} : \dot{V}(x) = 0\}$. De plus, cet ensemble est invariant, donc $\Omega_g(x_0) \subset \mathbf{I}$. Finalement, $\Phi_g^t(x_0)$ étant bornée, $\Phi_g^t(x_0)$ converge vers $\Omega_g(x_0) \subset \mathbf{I}$. ∎

Corollaire 4.5.3. *Si, dans (4.34), g est analytique, alors* $\mathbf{I} = \{0\} \Leftrightarrow \{\mathcal{L}_g^k V(x) = 0, k \in \mathbb{N}\} = \{0\}$ *(*$\mathbf{I}$ *étant défini dans le théorème 4.5.11).*

Théorèmes réciproques

Nous venons de donner des conditions suffisantes de stabilité (asymptotique ou non, globale ou locale) : il reste à savoir s'il existe des conditions nécessaires. Les théorèmes réciproques suivants donnent des réponses partielles.

Théorème 4.5.12. *[22] Soit l'EDO (4.11), avec f de classe* $\mathcal{C}^1$ *et de jacobienne bornée uniformement en t sur un compact contenant l'origine. Si l'origine est un point d'équilibre localement (respectivement globalement) exponentiellement stable, alors il existe un ouvert* $\mathcal{O}$ *de* $\mathbb{R}^n$ *contenant l'origine (respectivement* $\mathcal{O} = \mathbb{R}^n$*), une fonction* $V : [t_0, \infty[\times\mathcal{O} \to \mathbb{R}_+$ *de classe* $\mathcal{C}^1$ *et quatre constantes* $w_i > 0$ $(i \in \{1, ..., 4\})$ *telles que pour tout* $(t, x) \in [t_0, \infty[\times\mathcal{O}$ *:*

$$\begin{aligned} w_1 \|x\|^2 \leq V(t,x) &\leq w_2 \|x\|^2 , \\ \left.\frac{dV}{dt}\right|_{(4.34)} = \frac{\partial V}{\partial x} f(t,x) + \frac{\partial V}{\partial t} &\leq -w_3 \|x\|^2 , \\ \left\| \frac{\partial V}{\partial x} \right\| &\leq w_4 \|x\| . \end{aligned}$$

De même, il existe α_i *quatre* $\mathcal{K}-$*fonctions telles que pour tout* $(t, x) \in [t_0, \infty[\times\mathcal{O}$ *:*

$$\begin{aligned} \alpha_1(\|x\|) \leq V(t,x) &\leq \alpha_2(\|x\|) , \\ \left.\frac{dV}{dt}\right|_{(4.34)} = \frac{\partial V}{\partial x} f(t,x) + \frac{\partial V}{\partial t} &\leq -\alpha_3(\|x\|) , \\ \left\| \frac{\partial V}{\partial x} \right\| &\leq \alpha_4(\|x\|) . \end{aligned}$$

Si f est autonome (c'est-à-dire si on considère une EDO du type (4.34)), alors V peut être choisie indépendante du temps ($V = V(x)$*).*

Théorème 4.5.13. *[16, 17] Soit l'EDO (4.11), avec f de classe $\mathcal{C}^1$. Si l'origine est un point d'équilibre uniformément asymptotiquement stable, alors il existe un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ contenant l'origine et une fonction $V : [t_0, \infty[\times\mathcal{O} \to \mathbb{R}_+$ de classe $\mathcal{C}^1$, définie positive, convergeant uniformément en t vers zéro avec la norme de x et telle que sa dérivée soit définie négative. Si f est autonome (c'est-à-dire si on considère une EDO du type (4.34)), alors V peut être choisie indépendante du temps ($V = V(x)$).*

Théorèmes d'instabilité

Il existe de nombreux résultats donant des conditions suffisantes d'instabilité [16, 17].

Théorème 4.5.14 (de Liapounov)**.** *Soit l'EDO (4.11), admettant l'origine pour équilibre. S'il existe un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ contenant l'origine et une fonction $V : [t_0, \infty[\times\mathcal{O} \to \mathbb{R}$, $(t, x) \mapsto V(t, x)$, continue et convergeant uniformément en t vers zéro avec la norme de x et s'il existe un domaine non vide $\mathcal{D}$ contenant l'origine et sur lequel on a $V(t, x) < 0$ et $\dot{V}(t, x) \leq 0$, alors l'origine est instable.*

Exemple 4.5.11. Reprenons le modèle de Van der Pol (4.41) : l'équilibre $x = 0$ est instable pour $\mu > 0$. En effet, en prenant $V(x) = -\frac{1}{2}(x_1^2 + x_2^2)$, on obtient $\dot{V} = -2x_2^2(\mu - x_2^2)$ et le théorème 4.5.14 permet de conclure.

Corollaire 4.5.4 (de Liapounov)**.** *Soit l'EDO (4.34), admettant l'origine pour équilibre. S'il existe une fonction V continue dont la dérivée est définie négative et si $V(x)$ est définie négative ou indéfinie en signe, alors l'origine est instable.*

Exemple 4.5.12. Reprenons (4.1) : l'équilibre $x = 0$ est instable. En effet, en prenant $V(x) = -x^2$, on obtient $\dot{V} = -2ax^2(x_{\max} - x) < 0$ si $x \neq 0$. Le corollaire 4.5.4 permet de conclure.

Un résultat plus général est dû à N.G. Chetaev.

Théorème 4.5.15 (de Chetaev)**.** *Soit l'EDO (4.11), admettant l'origine pour équilibre. S'il existe un ouvert $\mathcal{O}$ de $\mathbb{R}^n$ contenant l'origine et une fonction $V : [t_0, \infty[\times\mathcal{O} \to \mathbb{R}$, $(t, x) \mapsto V(t, x)$ de classe $\mathcal{C}^1$ telle que :*

1. *$\forall \varepsilon > 0, \exists x \in \mathcal{B}_\varepsilon(0) : V(t, x) \leq 0, \forall t \geq t_0$ (on note $\mathcal{U}$ l'ensemble des points x pour lesquels $V(t, x) \leq 0, \forall t \geq t_0$),*
2. *$V(t, x)$ est minorée sur $\mathcal{U}'$ un sous-domaine de $\mathcal{U}$,*
3. *sur $\mathcal{U}'$, $\left.\frac{dV(t,x)}{dt}\right|_{(4.11)} < 0$ (en particulier, il existe α une $\mathcal{K}-$fonction telle que $\left.\frac{dV(t,x)}{dt}\right|_{(4.11)} \leq -\alpha(|V(t, x)|) < 0$),*

alors l'origine est instable.

Exemple 4.5.13. Soit l'EDO :

$$\dot{x} = x^3 + y^3,$$
$$\dot{y} = xy^2 + y^3.$$

L'origine est un point d'équilbre instable. En effet, prenons $V(x,y) = \frac{1}{2}(x^2 - y^2)$ et le domaine $\mathcal{U}$ défini par $\mathcal{U} = \{(x,y) \in \mathbb{R}^2 : -y \leq x \leq y \text{ ou } y \leq x \leq -y\}$ ($V(x) \leq 0$). En définissant $\mathcal{U}' = \mathcal{U} \cap \mathcal{B}_\varepsilon(0)$, V est minorée par $-\frac{\varepsilon^2}{2}$ et $\dot{V} = (x^4 - y^4) < 0$ sur $\mathcal{U}'$. Le théorème 4.5.15 permet alors de conclure à l'instabilité de l'origine. Notons enfin qu'en ce point, la jacobienne est nulle, ainsi les théorèmes 4.5.4 et 4.5.5 ne nous permettent pas de conclure.

Remarque 4.5.2. Ces résultats peuvent être facilement adaptés à la stabilité d'un ensemble $\mathcal{A}$. Enfin, on peut énoncer ces résultats de façon duale en remplaçant ‹‹ définie négative ›› par ‹‹ définie positive ›› et réciproquement.

Extensions vectorielles

Plus le système est complexe et de grande dimension et plus la construction de fonctions de Liapounov devient ‹‹ délicate ››. Face à ce genre de difficulté, il est d'usage :

R1) d'accepter de perdre un peu d'information pour se ramener à un problème connexe plus simple (c'est sur ce principe qu'est basé la première méthode de Liapounov, dite du linéarisé local) ;

R2) de décomposer le problème pour en comprendre plus facilement les parties constituantes, puis le re-composer (préceptes de Descartes).

Pour analyser un modèle de grande dimension, on peut ainsi tenter de le décomposer en plusieurs sous-systèmes de complexité et de dimension moindres. Les procédés R1 et R2 sont illustrés respectivement par les exemples 4.5.14 et 4.5.15.

Exemple 4.5.14. Soit le modèle :

$$\frac{dx}{dt} = 2x(-2 + \sin(t) + x), \quad t \in \mathbb{R}, \quad x \in \mathbb{R}. \tag{4.59}$$

En introduisant la variable $z = \text{sign}(x)x$, on obtient[13] :

$$\frac{dz}{dt} = 2z(-2 + \sin(t) + x), \text{ si } x \neq 0, \tag{4.60}$$
$$\frac{dz}{dt} \leq 2z(-1 + z), \tag{4.61}$$
$$0 \leq z(t) \leq \frac{z_0}{z_0 + (1 - z_0)\exp(2(t - t_0))}, \tag{4.62}$$

[13]Etant donné que la dérivée de la fonction signe au point $x = 0$ n'est pas définie au sens classique, nous exclurons ce cas pour la suite ($x \neq 0$). En fait, en utilisant une notion plus générale du gradient ou de la dérivée (voir [7, 27]), nous pourrions obtenir directement un résultat similaire à celui qui suit.

et il est alors évident que l'équilibre $x = 0$ de (4.59) est exponentiellement stable. Une estimation de $\mathcal{D}_{se}(0)$ est $]-\infty, 1[$. Par ailleurs, (4.62) reste valable pour $x = 0$.

Dans cet exemple 4.5.14, nous avons utilisé de façon implicite la notion de **système majorant (SM)** : en effet, les solutions de (4.61) sont majorées par celles de l'EDO : $\frac{dy}{dt} = 2y(-1+y)$ (pour des conditions initiales identiques). De tels systèmes majorants présentent les propriétés suivantes :

- leurs solutions permettent d'obtenir une estimation des comportements du système initial ;
- ils peuvent inférer une propriété qualitative $\mathcal{P}$ pour le système initial et, dans ce cas, le SM sera dit **système de comparaison (SC)** pour la propriété $\mathcal{P}$: c'est le cas dans l'exemple 4.5.14 où $\frac{dy}{dt} = 2y(-1+y)$ est un SC pour la propriété $\mathcal{P}$ de stabilité exponentielle pour le système (4.61) ;
- ils peuvent ne plus dépendre du temps ni d'éventuelles perturbations affectant le système initial, ce qui permet de simplifier l'étude de leurs solutions ;
- ils peuvent être de dimension réduite par rapport à celle du système initial (voir exemple 4.5.15).

Afin de formuler les concepts de SM et de SC, considérons les systèmes :

$$\dot{x} = f(t, x), \quad x \in \mathbb{R}^n, \tag{4.63}$$

$$\dot{z} = g(t, z), \quad z \in \mathbb{R}^n, \tag{4.64}$$

nous avons alors les définitions suivantes.

Définition 4.5.8. (4.64) est un **système majorant (SM)** de (4.63) sur $\mathcal{O} \subset \mathbb{R}^n$ si :

$$\forall (x_{01}, x_{02}) \in \mathcal{O}^2 \quad : \quad \mathcal{I} = \mathcal{I}_{(4.63)}(t_0, x_{01}) \cap \mathcal{I}_{(4.64)}(t_0, x_{02}) \neq \{t_0\},$$
$$x_{02} \geq x_{01} \Longrightarrow z(t; t_0, x_{02}) \geq x(t; t_0, x_{01}), \forall t \in \mathcal{I}.$$

De ce concept, nous pouvons déduire certaines propriétés qualitatives pour les solutions positives. Par exemple si $z(t; t_0, x_{02}) \geq x(t; t_0, x_{01}) \geq 0$ et si les solutions $z(t; t_0, x_{02})$ (pas forcément uniques) convergent vers l'origine, alors il en est de même pour les solutions $x(t; t_0, x_{01})$.

Définition 4.5.9. (4.64) est un **système de comparaison (SC)** de (4.63) pour la propriété $\mathcal{P}$ si [$\mathcal{P}$ vraie pour (4.64) $\Rightarrow$ $\mathcal{P}$ vraie pour (4.63)].

Exemple 4.5.15. Pour le système :

$$\frac{dx}{dt} = \begin{pmatrix} -2 + \sin t + \frac{1}{4}(x_1^2 + x_2^2) & -\sin t \\ \sin t & -2 + \sin t + \frac{1}{4}(x_1^2 + x_2^2) \end{pmatrix} x, \ t \in \mathbb{R}, x \in \mathbb{R}^2, \tag{4.65}$$

la variable $v = \frac{1}{4}(x_1^2 + x_2^2)$, est solution de :

$$\frac{dv}{dt} = 2v(-2 + \sin t + v), \quad t \in \mathbb{R}, v \in \mathbb{R}_+.$$

Ainsi, à l'aide de l'exemple 4.5.14, nous pouvons conclure que l'origine de (4.65) est exponentiellement stable et qu'une estimation de son domaine de stabilité exponentielle est $\{x \in \mathbb{R}^2 : (x_1^2 + x_2^2) < 4\}$.

L'analyse de l'exemple 4.5.15 nous permet de constater que l'utilisation de la fonction de Liapounov v permet de réduire la dimension de l'EDO étudiée, tout en conduisant à des conclusions significatives quant aux comportements des solutions de l'EDO initiale. Cette réduction de dimension entraîne une perte d'information sur les comportements liés au système initial. Il semble donc intéressant de réduire cette perte en utilisant non pas une seule fonction candidate à Liapounov, mais plusieurs fonctions regroupées dans un vecteur qui conduira à une **fonction vectorielle de Liapounov (FVL)**, en espérant que chacune d'elle apportera des informations provenant de différentes parties du système initial.

Définition 4.5.10. V est une **fonction vectorielle à Liapounov (FVCL)** si :

$$\begin{aligned} &V : \mathbb{R}^n \to \mathbb{R}^k, \\ &x \mapsto V(x) = [v_1(x), \ldots, v_k(x)]^T, \end{aligned}$$

où les fonctions $v_i(x)$ sont continues, semi-définies positives et telles que $[V(x) = 0 \Leftrightarrow x = 0]$.

Exemple 4.5.16. $V : \mathbb{R}^3 \to \mathbb{R}_+^2$, $x \mapsto V(x) = [x_1^2 + x_2^2, (x_2 - x_3)^2 + x_3^2]^T$ est une FVCL, alors que $V : x \mapsto [(x_1 + x_2)^2, (x_2 - x_3)^2]^T$ n'en est pas une.

Les normes vectorielles [2, 4, 14, 25, 26] consituent un cas particulier de FVCL, présentant l'avantage de permettre une construction systématique du système majorant. En particulier, en décomposant $\mathbb{R}^n$ en une somme directe :

$$\mathbb{R}^n = \bigoplus_{i=1}^{k} \mathbb{E}_i, \tag{4.66}$$

avec $\mathbb{E}_i$ sous-espace de $\mathbb{R}^n$, $\dim(\mathbb{E}_i) = n_i$ (isomorphe à $\mathbb{R}^{n_i}$), on construit des normes ‹‹ au sens usuel ›› $p_i(x_{[i]})$ sur $\mathbb{E}_i$ avec $x_{[i]} = \mathrm{Pr}_i(x)$ la projection de x sur $\mathbb{E}_i$. Ces différentes normes, regroupées dans un vecteur, permettent de définir une norme vectorielle régulière [14] :

$$\begin{aligned} &P : \mathbb{R}^n \to \mathbb{R}_+^k, \\ &x \mapsto P(x) = [p_1(x_{[i]}), \ldots, p_k(x_{[k]})]^T. \end{aligned}$$

[14] si la somme dans (4.66) n'est pas directe, alors P est une norme vectorielle non régulière

Ainsi, en décomposant (de façon non unique) le champ de vecteurs f suivant :

$$f(t,x,d) = A(t,x,d)x + b(t,x,d),$$

avec d un vecteur traduisant des incertitudes de modèle ou des perturbations, on obtient :

$$D^+P(x) \leq M(.)P(x) + q(.), \tag{4.67}$$

avec $(.) = (t,x,d)$, $M(.) = \{m_{ij}(.)\}$ une matrice $(k \times k)$ et $q(.) = [q_1(.), \ldots, q_k(.)]^T$ un k-vecteur, définis par :

$$m_{ij}(.) = \sup_{u \in \mathbb{R}^n} \left\{ \frac{\text{grad } p_i(u_{[i]})^T \,\text{Pr}_i\, A(.)\, \text{Pr}_j\, u_{[j]}}{p_j(u_{[j]})} \right\}, \tag{4.68}$$

$$q_i(.) = \left| (\text{grad } p_i(x_{[i]}))^T \,\text{Pr}_i\, b(.) \right|. \tag{4.69}$$

Pour certaines normes de Hölder , les expressions formelles de (4.68) et (4.69) peuvent aisément être obtenues [4, 14, 25, 26]. Par exemple, si $P(x) = [|x_1|, \ldots, |x_n|]^T$, alors M est la matrice A dont les éléments hors-diagonaux sont remplacés par leur valeur absolue (soit $m_{ij}(.) = |a_{ij}|$ si $i \neq j$ et $m_{ii}(.) = a_{ii}$) et $q = [|b_1|, \ldots, |b_n|]^T$. La fonction $M(.)z + q(.)$[15] est quasi-monotone non décroissante en z. De plus, on peut toujours trouver $g(z) = M(z)z + q(z) \geq M(.)z + q(.)$, qui soit quasi-monotone non décroissante en z (au moins localement). Ainsi, le théorème 4.3.6 permet de conclure que :

$$\dot{z} = M(z)z + q(z), \tag{4.70}$$

est un SM de (4.67). Ce qui conduit à divers résultats [25, 26], en particulier le suivant.

Théorème 4.5.16. *Considérons une des propriétés $\mathcal{P}$ définies aux paragraphes 4.5, 4.5, 4.5, par exemple : stabilité, attractivité, stabilité asymptotique, etc. Sous les hypothèses conduisant à la construction de (4.70), pour lequel z_e est un point d'équilibre positif de propriété $\mathcal{P}$ et ayant un domaine non vide associé $\mathcal{D}_\mathcal{P}(z_e)$, alors $\mathcal{A} = \{x \in \mathbb{R}^n : P(x) \leq z_e\}$ a la propriété $\mathcal{P}$, de domaine $\mathcal{D}_\mathcal{P}(\mathcal{A}) = \{x \in \mathbb{R}^n : \ P(x) \in \mathcal{D}_\mathcal{P}(z_e)\}$.*

Exemple 4.5.17. Soit le modèle :

$$\begin{cases} \dot{x}_1 = (1 - x_1^2 - x_2^2)x_1 + d_{12}(t)x_2, \\ \dot{x}_2 = (1 - x_1^2 - x_2^2)x_2 + d_{21}(t)x_1, \\ |d_{ij}(t)| \leq 1, \forall t \in \mathbb{R}, \end{cases} \tag{4.71}$$

[15] ainsi que $Mz + q$, avec la matrice M (respectivement le vecteur q) constituée des suprema sur (t,x,d) des coefficients de la matrice $M(.)$ définie par (4.68) (respectivement du vecteur $q(.)$ défini par (4.69))

où les fonctions d_{ij} sont continues par morceaux. Pour la norme vectorielle régulière $P(x) = [|x_1|, |x_2|]^T$, on obtient :

$$D_t P(x) \leq \begin{pmatrix} (1 - p_1^2(x) - p_2^2(x)) & 1 \\ 1 & (1 - p_1^2(x) - p_2^2(x)) \end{pmatrix} P(x),$$

auquel on associe le SM suivant :

$$\dot{z}(t) = g(z) = \begin{pmatrix} 1 - z_1^2 & 1 \\ 1 & 1 - z_2^2 \end{pmatrix} z(t),$$

(g est quasi-monotone non décroissante) ayant pour point d'équilibre positif :

$$z_e = \sqrt{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{4.72}$$

permettant de conclure que $\mathcal{A} = \left\{ x \in \mathbb{R}^n : P(x) \leq \sqrt{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$ est globalement asymptotiquement stable.

Enfin, notons que cette démarche peut être étendue au cas de matrices de fonctions de Liapounov [10].

4.6 Exercices

Comportements linéaires

Exercice 4.6.1. Déterminer le comportement des solutions des EDO suivantes :

$$\dot{x} = x, x \in \mathbb{R}^n, \tag{4.73}$$
$$\dot{x} = -2x, x \in \mathbb{R}^n, \tag{4.74}$$

$$\begin{cases} \dot{x} = y, \\ \dot{y} = -2x - y \end{cases}, (x, y) \in \mathbb{R}^2, \tag{4.75}$$

$$\begin{cases} \dot{x} = x + y, \\ \dot{y} = -x - 2y \end{cases}, (x, y) \in \mathbb{R}^2, \tag{4.76}$$

$$\begin{cases} \dot{x} = -x + 2y, \\ \dot{y} = -x + y \end{cases}, (x, y) \in \mathbb{R}^2, \tag{4.77}$$

$$\dot{x} = \begin{pmatrix} 1 & -2 & 0 \\ 2 & 0 & -1 \\ 4 & -2 & -1 \end{pmatrix} x, x \in \mathbb{R}^3, \text{ (nota 1 est valeur propre)} \tag{4.78}$$

$$\dot{x} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{pmatrix} x, x \in \mathbb{R}^3, \text{ (nota } -1 \text{ est valeur propre)} \tag{4.79}$$

Solution 4.6.1. *Pour ces équations $\dot{x} = Ax$, si A est inversible il n'y a qu'un seul point d'équilibre qui est l'origine sinon tout vecteur du noyau de A $\{y \in \mathbb{R}^n$ tq $Ay = 0\}$ est un équilibre. Enfin le comportement (cf. Théorème du cours) est entièrement déterminé par les parties réelles des valeurs propres de la matrice A. Pour (4.73), $A = Id$ donc il n'y a qu'un seul équilibre instable puisque 1 est une valeur propre positive (elle est d'ordre n). Pour (4.74), $A = -2Id$ donc il n'y a qu'un seul équilibre asymptôtiquement stable puisque toute les valeurs propres sont égales à -2 donc à parties réelles négatives. Pour (4.75),* $A = \begin{pmatrix} 0 & 1 \\ -2 & -1 \end{pmatrix}$, *dont les valeurs propres sont $-\frac{1}{2} + \frac{1}{2}i\sqrt{7}, -\frac{1}{2} - \frac{1}{2}i\sqrt{7}$ donc il n'y a qu'un seul équilibre asymptôtiquement stable puisque les valeur propres sont à parties réelles négatives. Pour (4.76),* $A = \begin{pmatrix} 1 & 1 \\ -1 & -2 \end{pmatrix}$, *dont les valeurs propres sont $\frac{1}{2}\sqrt{5} - \frac{1}{2}, -\frac{1}{2} - \frac{1}{2}\sqrt{5}$ donc il n'y a qu'un seul équilibre instable puisque une des valeurs propres est positive. Pour (4.77),* $A = \begin{pmatrix} -1 & 2 \\ -1 & 1 \end{pmatrix}$, *dont les valeurs propres sont $i, -i$ donc il n'y a qu'un seul équilibre stable (les espaces propres associé aux valeurs propres à parties réelles nulles sont tous les deux de dimension un). En fait, les solutions sont des ellipses. Pour (4.78),* $A = \begin{pmatrix} 1 & -2 & 0 \\ 2 & 0 & -1 \\ 4 & -2 & -1 \end{pmatrix}$, *dont les valeurs propres sont $1, -\frac{1}{2} + \frac{1}{2}i\sqrt{7}, -\frac{1}{2} - \frac{1}{2}i\sqrt{7}$ donc il n'y a qu'un seul équilibre instable. Pour (4.79),* $A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{pmatrix}$, *dont les valeurs propres sont $-1, -2, -3$ donc il n'y a qu'un seul équilibre asymptôtiquement stable.*

Equilibres

Exercice 4.6.2. Un pendule est constitué d'un fil de longueur l comportant une masse m en son extrémité :

$$\frac{d^2x}{dt^2} + \frac{g}{l}\sin(x) = 0, x \in \mathbb{R}. \tag{4.80}$$

Discuter de l'existence des solutions. Déterminer les points d'équilibre et leur nature.

Solution 4.6.2. *On pose $x_1 = x, x_2 = \dot{x}$, le système devient :*

$$\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -\frac{g}{l}\sin(x_1) \end{cases}, \tag{4.81}$$

Le second membre étant Lipchitzien, il y a bien existence et unicité de solution au PC (au moins localement). Donc les équilibres sont donnés par

$$\begin{cases} x_2 = 0, \\ x_1 \equiv 0 \bmod(\pi) \end{cases}$$

Pour chacun de ces équilibres on peut étudier le comportement local des solutions au voisinage de ces équilibres ($x_1 = k\pi, x_2 = 0$) à l'aide du système (dit linéarisé : $\dot{z} = Az, A$ étant la Jacobienne du second membre de (4.81))

$$\begin{cases} \dot{z}_1 = z_2, \\ \dot{z}_2 = (-1)^{k+1}\frac{g}{l}z_1 \end{cases}$$

Il suffit de déterminer les valeurs propres de la matrice $A = \begin{pmatrix} 0 & 1 \\ (-1)^{k+1}\frac{g}{l} & 0 \end{pmatrix}$, *dont le polynôme caractéristique est $\pi(A) = \lambda^2 + (-1)^k\frac{g}{l}$ donc si k est impair les valeurs propres sont $\pm\sqrt{\frac{g}{l}}$: une des valeurs propres étant positive l'équilibre correspondant est intable (position haute). En ce qui concerne le cas k pair les valeurs propres sont à parties réelle nulles : on ne peut pas conclure directement mais en utilisant la fonction $V(x) = \frac{1}{2}m\dot{x}^2 + mgl(1 - \cos(x))$ (Energie totale) sa dérivée est nulle le long des solutions de (4.81). Au voisinages des équilibres considérés V est définie positive et $\dot{V}$ nulle en utilisant le théorème 4.5.8 on conclut à la stabilité (mais pas asymptôtique).*

Etude de propriétés qualitatives des solutions d'EDO

Exercice 4.6.3. Déterminer et classer les points d'équilibres (point hyperbolique ou non) pour les systèmes décrits par les équations suivantes. Si le point est

hyperbolique on spécifiera s'il est asymptôtiquement stable, stable ou instable selon les valeurs des paramètres (ε) qui interviennent.

$$\frac{dx}{dt} = -|x|, x \in \mathbb{R}. \tag{4.82}$$

$$\frac{dx}{dt} = x^3(x-1), x \in \mathbb{R}. \tag{4.83}$$

$$\frac{d^2x}{dt^2} + \varepsilon\frac{dx}{dt} - x + x^n = 0, x \in \mathbb{R}. \tag{4.84}$$

$$\frac{d^2x}{dt^2} + \varepsilon\left(\frac{dx}{dt}\right)^2 + \sin(x) = 0, x \in \mathbb{R}. \tag{4.85}$$

$$\frac{d^2x}{dt^2} + \varepsilon\frac{dx}{dt}(x^2-1) + x = 0, x \in \mathbb{R}. \tag{4.86}$$

$$\frac{dx}{dt} = x(x-1), x \in \mathbb{R}. \tag{4.87}$$

$$\begin{cases} \frac{dx}{dt} = x(y-1) \\ \frac{dy}{dt} = y(x+1) \end{cases}. \tag{4.88}$$

$$\begin{cases} \frac{dx}{dt} = x(x-1) \\ \frac{dy}{dt} = x+y \end{cases}. \tag{4.89}$$

$$\begin{cases} \frac{dx}{dt} = -x+y \\ \frac{dy}{dt} = 2x-2y \end{cases}. \tag{4.90}$$

Solution 4.6.3. *Les résultats sont décrits dans le tableau suivant :*

EDO	*Unicité des sol.*	*équilibre*	*Hyp.*	*Val. Pr.*	*Stab. As.*
(4.82)	*? (pas Lipschitz)*	0	*?*	*?*	*non*
(4.83)	*oui (Lipschitz)*	0	*non*	0	*oui*
		1	*non*	1	*non*
(4.84)	*oui (Lipschitz)*	$(x=0,\dot{x}=0)$	ε *!*	$\frac{\varepsilon\pm i\sqrt{4-\varepsilon^2}}{2}$	ε *!*
		$(x=1,\dot{x}=0)$		$\frac{\varepsilon\pm\sqrt{\varepsilon^2-4(1-n)}}{2}$	
(4.85)	*oui (Lipschitz)*	$\left(x=\frac{k\pi}{2},\dot{x}=0\right)$	*non*	$(0,0)$	*non*
(4.86)	*oui (Lipschitz)*	$(x=0,\dot{x}=0)$	ε *!*	$\frac{\varepsilon\pm i\sqrt{4-\varepsilon^2}}{2}$	ε *!*
(4.87)	*oui (Lipschitz)*	0	*oui*	-1	*oui*
		1	*oui*	1	*non*
(4.88)	*oui (Lipschitz)*	$(x=0,y=0)$	*oui*	$(-1,1)$	*non*
		$(x=-1,y=1)$	*non*	$(i,-i)$	*? ?*
(4.89)	*oui (Lipschitz)*	$(x=0,y=0)$	*oui*	$(-1,1)$	*non*
		$(x=1,y=-1)$	*non*	$(1,1)$	*non*
(4.90)	*oui (Lipschitz)*	$(x=0,y=0)$	*non*	$(0,-3)$	*Stab.*

Exercice 4.6.4. Une masse m attachée à l'extrémité verticale d'un ressort dont le coefficient de rappel $k(x+x^3)$ est non linéaire :

$$\frac{d^2x}{dt^2}+\frac{k}{m}(x+x^3)=0, x\in\mathbb{R}. \tag{4.91}$$

Discuter de l'existence des solutions. Déterminer les points d'équilibres et leur natures.

Solution 4.6.4. *En posant $x=x_1,\dot{x}=x_2$, cette EDO devient*

$$\begin{aligned}\dot{x}_1&=x_2,\\ \dot{x}_2&=-\frac{k}{m}(1+x_1^2)x_1\end{aligned}$$

On en déduit l'existence et l'unicité des solutions au PC (le second membre étant $\mathcal{C}^1$). Du fait de l'unicité, les points d'équilibres sont les solutions de

$$\begin{aligned}\dot{x}_1&=x_2=0\\ \dot{x}_2&=-\frac{k}{m}(1+x_1^2)x_1=0\end{aligned}$$

il y a un seul point d'équilibre ($x = x_1 = \dot{x} = x_2 = 0$) qui est dégénéré (non hyperbolique) car

$$J_f = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{pmatrix}$$

mais stable : $V = \frac{1}{2}x_2^2 + \frac{k}{2m}(x_1^2 + \frac{1}{2}x_1^4), \dot{V} = -\frac{k}{m}x_1x_2\left(1+x_1^2\right) + \frac{k}{m}x_1x_2\left(1+x_1^2\right) = 0.$

Exercice 4.6.5. Monter que le champ de vecteur associé à l'équation différentielle donnée ci-dessous est bien à valeurs dans l'espace tangent au cercle.

$$\begin{cases} \frac{dx}{dt} = (x^2+y^2-1)x - \frac{2x^2y}{x^2+y^2} \\ \frac{dy}{dt} = (x^2+y^2-1)y + \frac{2x^3}{x^2+y^2} \end{cases}, (x,y) \in \text{cercle unité}, \tag{4.92}$$

Y-a-t-il existence et unicité des solutions (discuter) ? Déterminer, pour t_0 , x_0 et y_0 donnés, l'intervalle de définition $\mathcal{I}(t_0, x_0, y_0)$ des solutions passant par (x_0, y_0) à t_0. Expliciter les solutions. Indication : utiliser les coordonnées polaires.

Exercice 4.6.6. On considère

$$\dot{x} = A(t)x. \tag{4.93}$$

1. Montrer que les solutions forment une espace linéaire et que toute solution dépend linéairement de la condition initiale $x(t_0) = x_0$. En déduire l'existence d'une matrice dite "résolvante" $R(t, t_0)$ telle que $x(t, t_0, x_0) = R(t, t_0)x_0$. Montrer qu'elle vérifie les propriétés suivantes :

$$\begin{aligned} \frac{dR}{dt} &= A(t)R, \\ R(t_0, t_0) &= Id, \\ R(t_3, t_1) &= R(t_3, t_2)R(t_2, t_1) \\ R \text{ est inversible et } R^{-1}(t, t_0) &= R(t_0, t). \end{aligned}$$

2. Montrer que si $A(t) = A_0 + A_1(t)$ avec A_0 ayant toute ses valeurs propres à parties réelles strictement négatives et $\lim_{t\to\infty} A_1(t) = 0$, alors l'origine de (4.93) est asymptôtiquement stable.
3. On suppose cette fois que $A(t)$ est continue. Montrer que si $A(t)$ et l'intégrale $\int_{t_0}^{t} A(u)du$ ne commutent pas alors la résolvante $R(t, t_0)$ vérifie

$$R(t, t_0) = \exp\left(\int_{t_0}^{t} A(u)du\right).$$

Application : qu'en est-il pour $A(t) = \begin{pmatrix} t & 1 \\ 0 & 1 \end{pmatrix}$ puis pour

$A(t) = \begin{pmatrix} \cos(t) & -\sin(t) \\ \sin(t) & \cos(t) \end{pmatrix}$.

Exercice 4.6.7. Pour une machine à vapeur, ω la vitesse de rotation est liée à

$$\begin{aligned} \frac{d\omega}{dt} &= k\cos(\varphi+\theta) - F, \\ \frac{d^2\varphi}{dt^2} &= \omega^2 \sin\varphi\cos\varphi - g\sin\varphi - b\frac{d\varphi}{dt}, \end{aligned}$$

avec $b, g, k, F > 0$. Pour $\theta = 0$ déterminer le comportement de solutions.

Solution 4.6.5. *Pour $\theta = 0$, les équilibres vérifient (si $F < k$)*

$$\begin{aligned} \cos\varphi &= \frac{F}{k}, \\ \varphi_n^a &= a\arccos\left(\frac{F}{k}\right) + 2n\pi, a = \pm 1 \\ \omega^2 &= \frac{gk}{F}, \end{aligned}$$

Si on regarde une approximation au premier ordre $\cos(\varphi+\theta) = \cos\varphi = \frac{F}{k} - a\sqrt{1-\left(\frac{F}{k}\right)^2}(\varphi-\varphi_n)$, $\sin\varphi = a\sqrt{1-\left(\frac{F}{k}\right)^2} + \frac{F}{k}(\varphi-\varphi_n)$

$$\begin{aligned} \frac{d\omega}{dt} &= -a\sqrt{k^2-F^2}(\varphi-\varphi_n), \\ \frac{d^2\varphi}{dt^2} &= -\frac{gF}{k}(\varphi-\varphi_n) - b\frac{d\varphi}{dt}, \end{aligned}$$

En posant $x_1 = \omega - \sqrt{\frac{gk}{F}}, x_2 = \varphi - \varphi_n, x_3 = \frac{d\varphi}{dt}$:

$$\begin{aligned} \dot{x}_1 &= -a\sqrt{k^2-F^2}x_2, \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= -\frac{gF}{k}x_2 - bx_3, \end{aligned}$$

Soit en posant $x = (x_1, x_2, x_3)^T$:

$$\dot{x} = \begin{pmatrix} 0 & -a\sqrt{k^2-F^2} & 0 \\ 0 & 0 & 1 \\ 0 & -\frac{gF}{k} & -b \end{pmatrix} x$$

les valeur propres sont $0, \frac{1}{2}\left(-b \pm \sqrt{b^2 - 4\frac{gF}{k}}\right)$, *en résumé les équilibres* $\varphi_n^a = \pm \arccos\left(\frac{F}{k}\right) + 2n\pi$ *sont stables (pas attractifs).*

Exercice 4.6.8. Etudier les équilibres de

$$\begin{aligned} \dot{x} &= -xy^2 - 2y \\ \dot{y} &= x - x^2 y \end{aligned} \tag{4.94}$$

Montrer que l'origine est stable (Indication on utilisera la fonction $V = ax^2 + by^2, a > 0, b > 0$ et on choisira a, b pour que $\dot{V}$ soit définie négative ou nulle)

Solution 4.6.6. *Il n'y a qu'un seul équilibre : l'origine* $(x = y = 0)$

$$\dot{V} = 2(-axy(2 + xy) + bxy(1 - xy))$$

On a un choix trivial $-2a + b = 0$ *et* $a > 0$ *alors* $V > 0$ *et*

$$\dot{V} = -6ax^2y^2 \leq 0$$

le second théorème de Liapunov permet de conclure.

Exercice 4.6.9. On considère le système

$$\begin{aligned} \dot{x} &= -y + x(\epsilon - (x^2 + y^2)) \\ \dot{y} &= x + y(\epsilon - (x^2 + y^2)) \end{aligned} \tag{4.95}$$

1. Déterminer les points d'équilibres et leur nature (hyperbolicité, stabilité ?)
2. En utilisant la fonction $V(x) = \frac{1}{2}(x^2+y^2)$, montrer que l'origine est asymptôtiquement stable pour $\epsilon < 0$. Que se passe t-il pour $\epsilon = 0$?
3. Pour $\epsilon > 0$: qu'advient-il du comportement des solutions.
4. Retrouver ces résultats en intégrant cette équation (Indication : utiliser les coordonnées polaires).

Exercice 4.6.10. On considère le système suivant

$$\begin{cases} \frac{dx}{dt} = -x + y + f(x, y) \\ \frac{dy}{dt} = -x - y + g(x, y) \end{cases}. \tag{4.96}$$

1. On suppose dans cette question que $f(x, y) = g(x, y) = 0$. Déterminer les points d'équilibre.
2. Pour chacun de ces points d'équilibre, déterminer s'il est hyperbolique ou non, s'il est asymptotiquement stable, stable ou instable.

3. On supose dans cette question que $f(x,y) = -\frac{2x^2y}{x^2+y^2}, g(x,y) = \frac{2x^3}{x^2+y^2}$. On considère la fonction $V : (x,y) \mapsto \frac{1}{2}\left(x^2+y^2\right)$, évaluer $\frac{d}{dt}V(x(t),y(t))$, pour $x(t)$ et $y(t)$ solution de (4.96). Montrer que, pour tout couple (x_0,y_0) de conditions initiales de (4.96), la fonction V évaluée le long de trajectoires solutions de (4.96) c'est-à-dire pour des couples de points $(x(t),y(t))$ solution de (4.96) décroit exponentiellement vers 0.
4. En déduire le comportement qualitatif de (4.96) (c'est-à-dire comment se comportent les solutions au bout d'un temps infiniment grand et ce pour tout couple (x_0,y_0) de conditions initiales).

Solution 4.6.7. *On considère le système suivant*

$$\begin{cases} \frac{dx}{dt} = -x + y + f(x,y) \\ \frac{dy}{dt} = -x - y + g(x,y) \end{cases} . \tag{4.97}$$

1. On suppose dans cette question que $f(x,y) = g(x,y) = 0$. Déterminer les points d'équilibre.

Réponse : *Puisqu'il y a unicité des solutions (second membre loc. Lipschitz), on doit résoudre*

$$\begin{cases} 0 = -x + y \\ 0 = -x - y \end{cases} \tag{4.98}$$

$$A(x,y)^T = 0, A = \begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}$$

$$\det(A) = 2,$$

La matrice A étant non singulière, l'unique solution de (4.98) est :

$$x = 0, y = 0.$$

2.Pour chacun de ces points d'équilibre, déterminer s'il est hyperbolique ou non, s'il est asymptotiquement stable, stable ou instable.

Réponse : *Il faut calculer les valeurs propres de A : $-1 \pm i$ on en conclu qu'il s'agit d'un équilibre hyperbolique asymptôtiquement stable (cf. cours).*

3. On supose dans cette question que $f(x,y) = -\frac{2x^2y}{x^2+y^2}, g(x,y) = \frac{2x^3}{x^2+y^2}$. On considère la fonction $V : (x,y) \mapsto \frac{1}{2}\left(x^2+y^2\right)$, évaluer $\frac{d}{dt}V(x(t),y(t))$, pour $x(t)$ et $y(t)$ solution de (4.96). Montrer que, pour tout couple (x_0,y_0) de conditions initiales de (4.96), la fonction V évaluée le long de trajectoires solutions de (4.96) c'est-à-dire pour des couples de points $(x(t),y(t))$ solution de (4.96) décroit exponentiellement vers 0.

Réponse :

$$\begin{aligned}\frac{d}{dt}V(x(t),y(t)) &= \frac{\partial V}{\partial x}\frac{dx}{dt} + \frac{\partial V}{\partial y}\frac{dy}{dt}\\ &= x\dot{x} + y\dot{y}\\ &= x\left(-x + y - \frac{2x^2y}{x^2+y^2}\right) + y\left(-x - y + \frac{2x^3}{x^2+y^2}\right)\\ &= -2V + xy\left(1 - 1 - \frac{2x^2}{x^2+y^2} + \frac{2x^2}{x^2+y^2}\right)\\ &= -2V\end{aligned}$$

donc $V(x(t),y(t)) = \exp(-2t)V(x_0,y_0)$. *On en déduit que* $x^2(t)+y^2(t)$ *décroit exponentiellement vers 0 : les solutions décroissent exponentiellement vers 0 : l'origine est asymptotiquement stable (en fait on dit qu'il est exponentiellement stable).*

4. En déduire le comportement qualitatif de (4.96) (c'est-à-dire comment se comportent les solutions au bout d'un temps infiniment grand et ce pour tout couple (x_0,y_0) *de conditions initiales).*

Exercice 4.6.11. On considère le système

$$\begin{aligned}\dot{x} &= -y(1+x^2+y^2) + x(\epsilon - (x^2+y^2))\\ \dot{y} &= x(1+x^2+y^2) + y(\epsilon - (x^2+y^2))\end{aligned} \tag{4.99}$$

1. Déterminer les points d'équilibre et leur nature (hyperbolicité, stabilité ?)
2. En utilisant la fonction $V(x) = \frac{1}{2}(x^2+y^2)$, montrer que l'origine est asymptôtiquement stable pour $\epsilon < 0$. Que se passe t-il pour $\epsilon = 0$?
3. Pour $\epsilon > 0$: qu'advient-il du comportement des solutions.

Solution 4.6.8. *Quelques éléments de réponse :*

1. *Le second membre étant polynomiale on a bien existence et unicité locale des solutions. Les équilibre répondent à* $y(x^2+y^2) = x(\epsilon - (x^2+y^2))$ *et* $x(x^2+y^2) = -y(\epsilon-(x^2+y^2))$ *l'unique solution est donc l'origine* $x = y = 0$. *La Jacobienne en ce point vaut*

$$J = \begin{pmatrix} \epsilon & -1 \\ 1 & \epsilon \end{pmatrix}, \tag{4.100}$$

Les valeurs propres étant $-i+\epsilon, i+\epsilon$, *on en déduit que l'origine est :*
   - *Si* $\epsilon < 0$ *: Hyperbolique et asymptôtiquement stable,*
   - *Si* $\epsilon > 0$ *: Hyperbolique et instable,*
   - *Si* $\epsilon = 0$ *: Non Hyperbolique (pour la stabilité il faut une étude plus poussé : le* 1^{er} *théorème de Liapunov ne permet pas de conclure).*

2. *La fonction est bien définie positive :*

$$\begin{aligned}\frac{d}{dt}V(x(t),y(t)) &= x\dot{x}+y\dot{y}\\ &= (x^2+y^2)(\epsilon-(x^2+y^2))\\ &= 2V(\epsilon-2V))\end{aligned}$$

Si $\epsilon<0$ alors $\dot{V}$ est définie négative le 2^{nd} théorème de Liapunov permet de conclure. Pour $\epsilon=0:\dot{V}=-4V^2$ est aussi définie négative on en conclut que l'origine est asymptôtiquement stable en utilisant le second théorème de Liapunov (nota : le premier théorème de Liapunov ne nous avait pas permis de conclure).

3. *Il est clair que lorsque $\epsilon>0$, $\dot{V}$ est localement définie positive : l'origine est instable. Mais, Si on regarde l'EDO $\dot{V}=2V(\epsilon-2V)$ elle comporte deux équilibres $V=0$ et $V=\frac{\epsilon}{2}$: le premier étant instable et le second stable. Les solutions convergent donc vers l'ensemble $V=\frac{\epsilon}{2}$ c'est-à-dire le cercle de rayon $\sqrt{\epsilon}$ centré en l'origine.*

Exercice 4.6.12. Système de Volterra Lotka : modèle de lutte de deux espèces.

En 1917 (durant la première guerre mondiale), Umberto D'Ancona (biologiste) constata une augmentation du nombre de sélaciens (requins) dans la partie nord de la mer Adriatique. Afin d'expliquer ce phénomène, il fit appel à son beau-père Vito Volterra mathématicien de profession qui expliqua ce phénomène de la façon suivante. Soit un volume d'eau infini (mer Adriatique par exemple), peuplé par deux espèces : l'une carnivore (C : sélaciens) dévorant l'autre herbivore (H : crevettes). Notons x et y le nombre d'individus respectivement de l'espèce (H) et (C). Si l'espèce (H) peuplait seule la mer : elle se développerait de façon exponentielle (si l'on fait l'hypothèse que le développement de cette espèce n'est pas limité par l'espace et la quantité de nourriture). Donc dans ce cas, la vitesse de variation de l'espèce (H) serait : $\frac{dx}{dt}=ax$, avec a qui est un réel positif. Par contre, l'espèce (C) ne peut assurer seule ni son développement ni même sa survie, donc sa vitesse de variation serait : $\frac{dy}{dt}=-by$, avec b qui est un réel positif. Cependant, lorsque les deux espèces cohabitent, les carnivores dévorent les individus de l'espèce (H) : c'est leur fonction naturelle. En faisant l'hypothèse qu'à chaque rencontre d'un carnivore avec un herbivore, ce dernier est dévoré et que le nombre de rencontres est proportionnel au produit des densités volumiques des deux espèces (donc à xy), on peut conclure que les vitesses de variation des deux espèces sont régies par le système différentiel :

$$\begin{cases}\frac{dx}{dt}=ax-cxy\\ \frac{dy}{dt}=-by+dxy\end{cases}, \tag{4.101}$$

avec a,b,c,d qui sont des réels positifs.

Le problème pourrait se résumer à la question suivante :

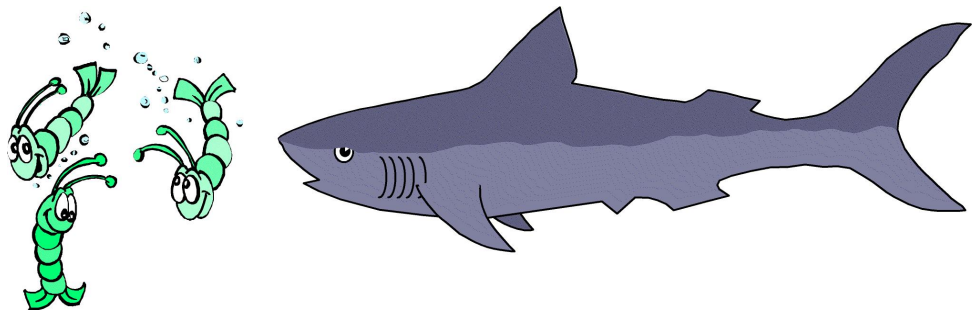


FIG. 4.11: Volterra-Lotka

"Après avoir expliqué le phénomène observé par Umberto d'Acona, quelle(s) politique(s) de pêche permet(tent) d'assurer un bon équilibre tout en maintenant une population de sélaciens suffisante ?"

A ces fins, on pourra répondre aux questions suivantes :

A) Reprendre l'étude d'une population unique :

1) Vérifier que pour une population modélisée par $\dot{y} = \tau y$ (loi de reproduction normale), celle-ci double en un temps indépendant de la population initiale.

2) Discuter des solutions de l'équation logistique.

B) Analyse du modèle dit de Volterra-Lotka :

1) Le problème de Cauchy associé admet il une ou plusieurs solutions ? Montrer que les solutions sont définies sur $\mathbb{R}$?

2) Le quart de plan positif est-il positivement invariant ? (interprétations).

3) Déterminer les points d'équilibres ainsi que le comportement des solutions au voisinage de ces points ?

4) Peut-on utiliser la méthode de séparation des variables ?

5) Montrer qu'il existe une fonction $H(x, y)$ constante le long des solutions de (4.101) :

$$\frac{dH(x,y)}{dt} = 0 = \frac{\partial H(x,y)}{\partial x}(ax - cxy) + \frac{\partial H(x,y)}{\partial y}(-by + dxy).$$

6) Montrer que les solutions non nulles de (4.101) rendent extremum la fonctionnelle suivante

$$J(x, y, \dot{x}, \dot{y}) = \int_{t_0}^{t_1} L(t, x(t), y(t), \dot{x}(t), \dot{y}(t))dt,$$
$$L = -H + \frac{1}{2xy}(x \ln x\, \dot{y} - (y \ln y\, \dot{x})$$

C) Analyse du problème

En 1917 (guerre oblige), l'activité de pêche est fortement ralentie, proposer un modèle tenant compte d'une certaine forme d'activité de pêche avant 1917.

Reprendre les questions de la partie B). Comparer alors les différents comportements. Conclusions.

D) Analyse dans un cadre plus général

Reprendre l'analyse dans un cadre plus général, quelles modifications doit-on apporter au modèle. Indication : la mer Adriatique est finie (loi Logistique).

4.7 Bibliographie

[1] ARNOLD, V.I.: *Equations Différentielles Ordinaires.* MIR, Moscou, 1988. 4ème édition traduit du russe.

[2] BELLMAN, R.: *Vector Lyapunov Functions.* J. SIAM Control, ser A, 1(1) :31–34., 1962.

[3] BHATIA, N.P. et G.P. SZEGÖ: *Stability Theory of Dynamical Systems.* Springer Verlag Berlin, 1970.

[4] BORNE, P.: *Contribution À l'Etude Des Sytsèmes Discrets Non-Linéaires de Grande Dimension. Application Aux Systèmes Interconnectés.* Thèse de doctorat, Université de Lille, 1976.

[5] CHIANG, H. D., M.W. HIRSCH et F.F. WU: *Stability Regions Of Nonlinear Autonomous Dynamical Systems.* IEEE Trans. Auto. Control., 33(1) :16–27, Janvier 1988.

[6] CHIANG, H.D. et J.S. THORP: *Stability Regions of Nonlinear Dynamical Systems : A Constructive Methodology.* IEEE Trans. Auto. Control, 34(12) :1229–1241, Décembre 1989.

[7] CLARKE, Frank. H.: *Optimization and Nonsmooth Analysis.* Wiley-Interscience Publication, 1983.

[8] CODDINGTON, E. et N. LEVINSON: *Theory of Ordinary Diffrential Equations.* Mc Graw-Hill, 1955.

[9] CORRIEU, P.L.: *Commande En Boucle Fermée D'un Actionneur Pas-À-Pas.* Mémoire de maîtrise, DEA d'Automatique, Université des Sciences et Technologies de Lille, 1999.

[10] DJORDJEVIC, M.Z.: *Stability Analysis of Nonlinear Systems by Matrix Lyapunov Method.* Dans *IMACS-IFACS, Modelling and Simulation for Control of Lumped and Distributed Parameter Systems*, pages 209–212, I.D.N 59 650 Villeneuve d'ASCQ (France), 3–6 Juin 1986.

[11] FILIPPOV, A. F.: *Differential Equations with Discontinuous Righthand Sides.* Kluwer Academic Publishers, 1988.

[12] GENESIO, R., M. TARTAGLIA et A. VICINO: *On Estimation of Asymptotic Stability Regions : State of Art and New Proposals.* IEEE Trans. Auto. Control, AC-30(8) :747–755, Août 1985.

[13] GENESIO, R. et A. VICINO: *New Techniques for Constructing Asymptotic Stability Regions for Nonlinear Systems*. IEEE Trans. Circuits and Sys., CAS-31(6) :574–581, Juin 1984.

[14] GRUJIĆ, Lj.T., J.C. GENTINA et P. BORNE: *"General Aggregation of Large-Scale Systems by Vector Lyapunov Functions and Vector Norms*. Int. J. Control,, 24(4) :529–550, 1977.

[15] GUCKENHEIMER, J. et P. HOLMES: *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer Verlag, 1983.

[16] HAHN, W.: *Theory and Application of Liapunov's Direct Method*. Prentice-Hall, Englewood Cliffs, N.J., 1963.

[17] HAHN, W.: *Stability of Motion*. Springer-Verlag N.Y., 1967.

[18] HALE, J. et H. KOÇAK: *Dynamics and Bifurcations*, tome 3 de *Text in Apllied Mathematics*. Springer-Verlag N.Y., 1991.

[19] HIRSH, M.W. et S. SMALE: *Differential Equations, Dynamical Systems, and Linear Algabra*. Academic Press, 1974.

[20] ISIDORI, A.: *Nonlinear Control Systems*, tome 1. Springer, 1989. 3e édition.

[21] KAMKE, E.: *Zur Theorie Gewöhnlicher Differentialgleichung II*. Acta Mathematica, 58 :57–87, 1932.

[22] KHALIL, H.K.: *Nonlinear Systems*. Prentice-Hall, 1996.

[23] LAKSHMIKANTHAM, V. et S. LEELA: *Differential and Integral Inequalities*, tome 1. Academic Press, New York, 1969.

[24] LIAPOUNOV, A.M.: *Stability of Motion : General Problem*. Int. J. Control, 55(3), Mars 1892 (1992). Lyapunov Centenary Issue.

[25] PERRUQUETTI, W.: *Sur la Stabilité et l'Estimation Des Comportements Non Linéaires, Non Stationnaires, Perturbés*. Thèse de doctorat, University of Sciences and Technology of Lille, France, 1994.

[26] PERRUQUETTI, W., J.P. RICHARD et P. BORNE: *Vector Lyapunov Functions : Recent Developments for Stability, Robustness, Practical Stability and Constrained Control*. Nonlinear Times & Digest, 2 :227–258, 1995.

[27] RICHARD, Jean Pierre: *Mathématiques Pour Les Systèmes Continus : Tome un*. Hermes, 2000.

[28] ROUCHE, N. et J. MAWHIN: *Equations Différentielles Ordinaires, Tome 1 : Théorie Générale*. Masson et Cie, Paris, 1973.

[29] ROUCHE, N. et J. MAWHIN: *Equations Différentielles Ordinaires, Tome 2 : Stabilité et Solutions Périodiques*. Masson et Cie, Paris, 1973.

[30] WAZEWSKI, T.: *Systèmes Des Equations et Des Inégalités Différentielles Ordinaires Aux Seconds Membres Monotones et Leurs Applications*. Ann. Soc. Polon. Math., 23 :112–166, 1950.

5 | Calcul des variations

Wilfrid Perruquetti[1]

[1]LAGIS & INRIA-ALIEN, Ecole Centrale de Lille, BP 48, 59651 Villeneuve d'Ascq cedex, France. *E-mail* : Wilfrid.Perruquetti@ec-lille.fr

5.1 Quelques exemples introductifs

De nombreux problèmes d'optimisation comportent un critère faisant intervenir une fonction que l'on cherche. Citons quelques exemples :

1. Soient A et B deux points donnés du plan. Déterminer la courbe rectifiable de longueur minimale reliant les deux points. Si on munit le plan d'un repère (Espace affine identifié à $\mathbb{R}^2$) dont l'origine est le point A, une courbe reliant ces deux points est la donnée d'une fonction $y = f(x)$ (telle que $f(0) = 0$ et $f(x_B) = y_B$). Si la courbe est rectifiable la longueur est donnée par
$$J_1(f) = \int_0^{x_B} \sqrt{1 + (f'(x))^2} dx, \tag{5.1}$$
(intuitivement on sent bien qu'il faut $f'(x) = 0$: c'est-à-dire le segment de droite reliant A à B est la réponse à notre problème).
2. On ne peut éviter l'exemple du brachistochrone ($\beta\rho\alpha\chi\iota\sigma\tau o\varsigma$ = le plus court, $\chi\rho o\nu o\varsigma$ = temps) qui fût le premier problème connu de ce genre. En juin 1696 dans un numéro d'"ACTA ERUDITORUM" Johann Bernouilli proposa le challenge suivant (reporté ici en termes modernes) : "Soient deux points donnés A, B dans un plan vertical. Quel est l'ensemble des trajectoires obtenues lorsqu'un point M part de A et arrive en B en un temps minimum sous l'unique influence de son poids. ...Afin d'éviter les conclusions farfelues, on peut remarquer que la ligne droite est évidemment la courbe de plus courte distance joignant les points A et B, mais ce n'est certes pas celle qui donne un temps de parcours minimum. Cependant, la courbe solution de ce problème - que je divulguerais si d'ici la fin de cette

année personne ne l'a trouvée- est fort bien connue des géomètres." Mise en équation du problème : prenons un repère (z, x) (avec l'axe des z orienté vers le bas) d'origine A. Une courbe reliant les deux points A et B est la donnée d'une fonction $x = f(z)$ (telle que $f(0) = 0$ et $f(z_B) = x_B$). La vitesse du point de masse unité s'exprime par $\left(\frac{dz}{dt}\right)^2 + \left(\frac{dx}{dt}\right)^2 = 2gz$ donc $dt = \frac{dz\sqrt{1+(f')^2}}{\sqrt{2gz}}$. Le temps mis pour aller de A à B est donc

$$T = J_2(f) = \int_0^{z_B} \frac{\sqrt{1+(f')^2}}{\sqrt{2gz}} dz. \tag{5.2}$$

On est donc amené à déterminer une fonction f qui minimise $T = J_2(f)$.

3. Problème de l'isopérimètre (Euler) : Déterminer la courbe embrassant l'aire maximale parmi l'ensemble des courbes fermées de classe $\mathcal{C}^1$ et de longueur donnée l. On peut sans perte de généralité, considérer que ces courbes sont issues de l'origine et que l'autre extrémité a pour abscisse x_B. La longueur est donc $l = \int_0^{x_B} \sqrt{1 + (f'(x))^2} dx$, quand à l'aire de révolution c'est
$$J_3(f) = 2\pi \int_0^{x_B} f(x)\sqrt{1 + (f'(x))^2} dx. \tag{5.3}$$
Ce problème est un peu plus complexe puisqu'il faut minimiser un critère sous une contrainte (la longueur doit être de l).

4. Parmi tous les arcs de courbes de classe $\mathcal{C}^1$ joignant A et B, deux points donnés du plan, et de longueur donnée l, déterminer celui qui délimite avec le segment AB une aire maximum. On peut sans perte de généralité, considérer que ces courbes sont issues de l'origine et que l'autre extrémité a pour abscisse x_B. La longueur est donc $l = \int_0^{x_B} \sqrt{1 + (f'(x))^2} dx$, quant à l'aire délimitée par la courbe et l'axe des abscisses c'est
$$J_4(f) = \int_0^{x_B} f(x) dx. \tag{5.4}$$
Ce problème fût posé et résolu (de façon intuitive) par la reine Dido de Carthage en 850 AV-JC. On retrouve un problème similaire à celui présenté en 3 (minimum sous contrainte). Une formulation plus générale consiste à déterminer la courbe qui définie la surface la plus grande parmi toutes les courbes fermées de périmètre donné (ce qui définit une contrainte).

5.2 Formulation du Problème

Soit

$$\begin{aligned} q &: [a, b] \subset \mathbb{R} \to \mathbb{R}^n \\ t &\mapsto q(t), \end{aligned} \tag{5.5}$$

les critères précédents (J_1 à J_4) se mettent sous la forme

$$J(q) = \int_a^b L(t, q(t), \dot{q}(t))dt, \tag{5.6}$$

avec L (appelé **Lagrangien**) :

$$\begin{aligned} L : [a,b] \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ (x,y,z) &\mapsto L(x,y,z) \end{aligned}$$

Problème : trouver toutes les fonctions q qui minimisent ou maximalisent $J(q)$ et satisfassent $q(a) = q_a, q(b) = q_b$ avec q_a, q_b donnés. $J(q)$ est une **fonctionnelle** : c'est une application qui à une fonction à valeur dans un espace vectoriel associe un élément d'un espace vectoriel. Donc on doit résoudre un problème d'optimisation sur un espace fonctionnel. Si on impose que les fonctions cherchées soient continûment différentiables, on travaillera avec l'espace fonctionnel

$$E = \mathcal{C}^1([a,b], \mathbb{R}^n). \tag{5.7}$$

Généralement on muni E de la **distance** de **Chebychev**

$$d_1^C(f,g) = \max_{t\in[a,b]} \|f(t) - g(t)\| + \sup_{t\in[a,b]} \left\|f'(t) - g'(t)\right\|,$$

induisant la norme n_1^C. Le sup étant atteint si on travaille sur E (on le remplace par un max), mais si on travaille sur E_s l'ensemble des fonctions $\mathcal{C}^1$ par morceaux (la dérivée peut être discontinue en un nombre fini de points) on conserve cette distance. Si on note $E_{a,b} = \{q \in E : q(a) = q_a, q(b) = q_b\}$, on peut donner la définition suivante

Définition 5.2.1. q_0 est un **minimum** (respectivement un **maximum**) **relatif** pour $J(q)$ dans l'ensemble $E_{a,b}$ muni de la norme n_1^C si

$$J(q) - J(q_0) \geq 0, \quad (\text{respectivement } \leq 0) \tag{5.8}$$

pour toute fonction $q \in E_{a,b}$ telle que $n_1^C(q - q_0) < \delta$, pour un δ donné strictement positif.

Intuitivement, l'extremum doit annuler une "dérivée". Ce qu'il faut bien noter c'est que J est une fonctionnelle : il faut adapter la notion de dérivée.

Définition 5.2.2. $J(q)$ est **différentiable** en q_0 (dans l'espace E muni de la norme n_1^C) s'il existe une **fonctionnelle linéaire continue**[1] $J'_{q_0} \in \mathcal{L}(E; \mathbb{R})$ telle que

$$J(q_0 + h) = J(q_0) + J'_{q_0}(h) + \varepsilon(h) \tag{5.9}$$

avec $\lim_{n_1^C(h)\to 0} \varepsilon(h) = 0$.

[1] Evidemment, linéaire signifie : $\forall(\alpha,\beta) \in \mathbb{R}^2, \forall(q_1,q_2) \in E^2 : J'_{q_0}(\alpha q_1 + \beta q_2) = \alpha J'_{q_0}(q_1) + \beta J'_{q_0}(q_2)$. Par exemple, la fonctionnelle $q \mapsto \int_a^b (q + \dot{q})dt$ est linéaire.

Pour notre fonctionnelle $J(q) = \int_a^b L(t,q,\dot{q})dt$, en supposant que L soit $\mathcal{C}^1$ par rapport à chacun de ses arguments :

Théorème 5.2.1. *La fonction réelle $J : q \mapsto J(q) = \int_a^b L(t,q,\dot{q})dt$ (avec L $\mathcal{C}^1$ par rapport à chacun de ses arguments), est continûment dérivable et sa dérivée $J'_{q_0} \in \mathcal{L}(E;\mathbb{R})$ en q_0 est donnée par*

$$\begin{aligned}\delta q \mapsto J'_{q_0}(\delta q) &= \int_a^b L'(t,q_0(t),\dot{q}_0(t))(0,\delta q,\delta q')dt \\ &= \int_a^b \left(\frac{\partial L}{\partial q}(t,q_0(t),\dot{q}_0(t))\delta q + \frac{\partial L}{\partial \dot{q}}(t,q_0(t),\dot{q}_0(t))\delta q'\right)dt\end{aligned}$$

■

Démonstration. Il suffit d'écrire la variation de l'intégrale

$$\Delta J = \int_a^b \Delta L dt, \quad \Delta L = L(t,q_0(t)+\delta q,\dot{q}_0(t)+\delta q') - L(t,q_0(t),\dot{q}_0(t))$$

or la formule des accroissements finis donne

$$\Delta L = \left(\frac{\partial L}{\partial q}(t,q_0(t),\dot{q}_0(t))\delta q + \frac{\partial L}{\partial \dot{q}}(t,q_0(t),\dot{q}_0(t))\delta q'\right) + R(t) \tag{5.10}$$

avec la majoration suivante du reste sur l'intervalle $[a,b]$:

$$\begin{aligned}|R(t)| &\leq M \times \max\left\{\sup_{t\in[a,b]} \|\delta q(t)\|, \sup_{t\in[a,b]} \|\delta q'(t)\|\right\}, \\ M &= \sup_{\substack{\|x\|\leq\|\delta q\| \\ \|x'\|\leq\|\delta q'\|}} \left\{\left\|L'(t,q_0(t)+x,\dot{q}_0(t)+x') - L'(t,q_0(t),\dot{q}_0(t))\right\|\right\}\end{aligned}$$

Ce qui nous sauve c'est la compacité de $[a,b]$ et la continuité de L' : en effet sur le compact $[a,b]$ L' est donc uniformément continue donc

$$\forall \varepsilon > 0, \exists \eta > 0 : \left\|(t,x,x')\right\| < \eta \Rightarrow \left\|L'(t,q_0(t)+x,\dot{q}_0(t)+x') - L'(t,q_0(t),\dot{q}_0(t))\right\| < \frac{\varepsilon}{|b-a|}.$$

Ainsi $\Delta J = J'(q_0)\delta q + \int_a^b R(t)dt$. Si $n_1^C(\delta q) < \eta$: $\left|\int_a^b R(t)dt\right| < \varepsilon n_1^C(\delta q)$. Clairement, ΔJ est continue par rapport à δq. Ainsi J est différentiable, de dérivée $\delta q \mapsto J'_{q_0}(\delta q)$. Montrons que cette application est continue. Soit q_0 et q_1 dans E :

$$J'_{q_0}(\delta q) - J'_{q_1}(\delta q) = \int_a^b \left(L'(t,q_0(t),\dot{q}_0(t)) - L'(t,q_1(t),\dot{q}_1(t))\right)(0,\delta q,\delta q')dt$$

or (5.10) implique que $d_c(q_0, q_1) < \delta \Longrightarrow \left|J'_{q_0}(\delta q) - J'_{q_1}(\delta q)\right| < \varepsilon n_1^C(\delta q)$ donc

$$\left\|J'(q_0)\delta q - J'(q_1)\right\| = \sup_{n_1^C(\delta q)=1} \left|J'(q_0)\delta q - J'(q_1)\delta q\right| < \varepsilon.$$

□

5.3 Condition Nécessaire : équations d'Euler

La fonction cherchée qui rend extremum (5.6) doit annuler la dérivée de J :

Théorème 5.3.1. *Pour que J (différentiable) possède un extremum en q_0, il est nécessaire que*

$$J'_{q_0} = 0. \tag{5.11}$$

■

Démonstration. Supposons que ce soit un maximum alors pour $\delta > 0$ suffisamment petit : $J(q_0 + h) - J(q_0) = J'_{q_0}(h) + \varepsilon(h)$ et ce pour tout h : $n_1^C(h) < \delta$. S'il existe h tel que $J'_{q_0}(h) < 0$, en prenant $-\eta h$, $\eta \in \mathbb{R}$ suffisamment petit, on obtient $J'_{q_0}(-\eta h) + \varepsilon(-\eta h) > 0$ ce qui contredit $J(q_0 + h) - J(q_0) \leq 0$ (maximum en q_0). L'autre éventualité se traitant de la même façon. □

Si on applique ce résultat à notre problème : il est nécessaire que q_0 vérifie

$$\begin{aligned} J'_{q_0}(\delta q) &= \int_a^b \left(\frac{\partial L}{\partial q}(t, q_0(t), \dot{q}_0(t))\delta q + \frac{\partial L}{\partial \dot{q}}(t, q_0(t), \dot{q}_0(t))\delta q' \right) dt \\ &= 0 \end{aligned}$$

et ce pour tout $\delta q \in \mathcal{C}^1([a, b], \mathbb{R}^n)$. Supposons que l'on cherche des solutions de classe $\mathcal{C}^2$ (dans ce cas $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)$ existe et on peut faire une intégration par parties) :

$$\begin{aligned} \int_a^b \left(\frac{\partial L}{\partial \dot{q}}(t, q_0(t), \dot{q}_0(t))\delta q' \right) dt &= \left[\frac{\partial L}{\partial \dot{q}}(t, q_0(t), \dot{q}_0(t))\delta q \right]_a^b - \\ &\quad \int_a^b \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}} \right)(t, q_0(t), \dot{q}_0(t))\delta q dt \\ &= -\int_a^b \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}} \right)(t, q_0(t), \dot{q}_0(t))\delta q dt \\ J'_{q_0}(\delta q) &= \int_a^b \left(\frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}} \right) \right)(t, q_0(t), \dot{q}_0(t))\delta q dt \\ 0 &= \left(\frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}} \right) \right)(t, q_0(t), \dot{q}_0(t)) \end{aligned}$$

Pour obtenir cette condition nécessaire sans supposer a priori que la solution cherchée est $\mathcal{C}^2$, nous avons besoin du lemme suivant (généralisation du lemme de Haar cf. lemme 5.5.2, en annexe) :

Lemme 5.3.1. *Soit E un espace vectoriel. Si f est une fonction continue sur un intervalle $[a,b]$ à valeur dans le dual de E (E^*) et si pour toute fonction $h \in \mathcal{C}^p([a,b],E)$ (donc à valeur dans E) telle que les $(p-1)-$ème dérivées successives s'annulent aux deux extrémités de $[a,b]$ ($h(a) = h(b) = h'(a) = h'(b) = \ldots = h^{(p-1)}(a) = h^{(p-1)}(b) = 0$) on a*

$$\int_a^b \left\langle f(t), h^{(p)}(t) \right\rangle dt = 0$$

alors $f^{(p)}$ est identiquement nulle sur $[a,b]$. ■

Remarque 5.3.1. En posant $\phi(x) = \int_a^x \frac{\partial L}{\partial q}(t, q_0(t), \dot{q}_0(t))dt$, on obtient

$$J'_{q_0}(\delta q) = [\phi(x)\delta q]_a^b + \int_a^b \left(\frac{\partial L}{\partial \dot{q}}(t, q_0(t), \dot{q}_0(t)) - \phi(t) \right) \delta q' dt, \tag{5.12}$$

comme $\delta q(a) = \delta q(b) = 0$, en utilisant le lemme 5.3.1 :

$$\frac{\partial L}{\partial \dot{q}}(t, q_0(t), \dot{q}_0(t)) - \phi(t) = C, \tag{5.13}$$

$$J'(q_0) = [\phi(x)\delta q]_a^b + [C\delta q]_a^b . \tag{5.14}$$

Ainsi en utilisant le lemme suivant

Lemme 5.3.2. *Soient f et g deux fonctions continues sur un intervalle $[a,b]$ à valeur dans $\mathbb{R}^n$ et si pour toute fonction h continûment différentiable s'annulant aux deux extrémités de $[a,b]$ ($h(a) = h(b) = 0$) on a*

$$\int_a^b \langle f(t), h(t) \rangle + \left\langle g(t), h'(t) \right\rangle dt = 0$$

alors g est différentiable sur $[a,b]$ et $g'(t) = f(t)$. ■

Démonstration. On pose $F(t) = \int_0^t f(x)dx$, donc $\int_a^b \langle f(t), h(t) \rangle = [F(t)h(t)]_{t=a}^{t=b} - \int_a^b \langle F(t), h'(t) \rangle dt = -\int_a^b \langle F(t), h'(t) \rangle dt$, ce qui conduit à la condition

$$\int_a^b \left\langle g(t) - F(t), h'(t) \right\rangle dt = 0$$

donc en utilisant le lemme 5.3.1 $[g(t) - F(t)]' = 0$ c'est-à-dire $g'(t) = f(t)$. □

On en déduit :

Théorème 5.3.2. *Soient $[a,b]$ un segment de $\mathbb{R}$ et Ω l'ensemble des fonctions q continûment dérivables telles que $q(a) = q_a, q(b) = q_b$. Pour que $q_0 \in \Omega \subset \mathcal{C}^2([a,b],E)$ rende extremum le critère (intégrale J stationnaire), il est nécessaire que q_0 soit solution de l'**équation d'Euler***

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \left(\frac{\partial L}{\partial q}\right), \tag{5.15}$$

■

Quelques cas simple d'intégration des équations d'Euler

1) L indépendant de q alors $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = 0$ donc $\frac{\partial L}{\partial \dot{q}_i} = cte \in E^*$ si $E = \mathbb{R}$: $\frac{\partial L}{\partial \dot{q}} = c \in \mathbb{R}, i = 1..n$)

2) L indépendant de t (application à la mécanique) : on obtient

$$\begin{aligned}
\frac{\partial L}{\partial q_i} - \frac{\partial^2 L}{\partial \dot{q}_i \partial t} - \sum_{j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial q_j}\dot{q}_j - \sum_{j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}\ddot{q}_j = 0, i = 1, \ldots, m \\
\frac{\partial L}{\partial q_i} - \sum_{j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial q_j}\dot{q}_j - \sum_{j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}\ddot{q}_j = 0, \\
\sum_{i=1}^{m} \frac{\partial L}{\partial q_i}\dot{q}_i - \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial q_j}\dot{q}_j\dot{q}_i - \sum_{i,j=1}^{m} \frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}\ddot{q}_j\dot{q}_i = 0, \\
\frac{d}{dt}\left(L - \sum_{j=1}^{m} \frac{\partial L}{\partial \dot{q}_j}\dot{q}_j\right) = 0, \\
L - \sum_{j=1}^{m} \frac{\partial L}{\partial \dot{q}_j}\dot{q}_j = cte \qquad (5.16)
\end{aligned}$$

3) J fait intervenir l'intégrale d'une fonction $f(x, y)$ par rapport à une abscisse curviligne

$$\begin{aligned}
J(q) = \int_a^b f(x, q)\sqrt{1 + q'^2}dx, \\
\frac{d}{dx}\left(\frac{\partial L}{\partial q'}\right) - \left(\frac{\partial L}{\partial q}\right) = \frac{1}{\sqrt{1 + q'^2}}\left[\left(\frac{\partial f}{\partial q}\right) - \left(\frac{\partial f}{\partial x}\right)q' - f\frac{q''}{1 + q'^2}\right] \qquad (5.17)
\end{aligned}$$

Condition nécessaire et suffisante.

L'application directe du lemme de Dubois-Reymond (cf. Annexe Lemme 5.5.1) :

Théorème 5.3.3. *Si on cherche q_0 parmi les fonctions $\mathcal{C}^2([a, b], \mathbb{R})$ qui rende extremum le critère J et vérifiant les conditions aux limites données, alors l'équation d'Euler et les conditions aux limites sont des CNS.* ■

Exemple 5.3.1. Si on reprend le problème 1 : la fonctionnelle à minimiser est $J_1(f) = \int_0^1 \sqrt{1 + (f'(x))^2}dx$ avec $f(0) = 0, f(1) = 0$: (5.15) s'écrit $\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$, soit $f''(x) = 0$, les solutions sont les droites !

Condition du second ordre pour un minimum : la Hessienne doit être définie positive

$$\text{cas scalaire } q \in \mathbb{R} : \frac{\partial^2 L}{\partial \dot{q}^2}(t,q,\dot{q}) \geq 0$$

$$\text{cas vectoriel } q \in \mathbb{R}^n : \left[\frac{\partial^2 L}{\partial \dot{q}_i \partial \dot{q}_j}(t,q,\dot{q})\right] \geq 0$$

Applications

Brachistochrone

La fonctionnelle à minimiser est (5.2) : $L = \frac{\sqrt{1+(f')^2}}{\sqrt{2gz}}$. La CN d'Euler s'écrit $\frac{d}{dz}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$ avec $\frac{\partial L}{\partial f'} = \frac{f'}{\sqrt{2gz}\sqrt{1+(f')^2}}$, $\frac{\partial L}{\partial f} = 0$. Il faut chercher $f(z)$ telle que $\frac{f'}{\sqrt{2gz}\sqrt{1+(f')^2}} = cte = \frac{\pm 1}{\sqrt{4gc}}$ (c constante arbitraire non nulle sinon $f' = 0$, $f = cte$ le point B est à la verticale du point A et dans ce cas le problème est trivial). L'équation devient (avec le signe $+$) $f'^2 = \frac{z}{2c}\left(1 + f'^2\right)$, c'est-à-dire :

$$f' = \pm\sqrt{\frac{z}{(2c - z)}} \tag{5.18}$$

en posant $z = c(1 - \cos(u))$ ($dz = c\sin(u)du$), (5.18) devient

$$\frac{df}{du} = \pm c\sin(u)\sqrt{\frac{1-\cos(u)}{1+\cos(u)}} = \pm c(1 - \cos(u))$$

les solutions sont les courbes paramétrées :

$$f = cte \pm c(u - \sin(u))$$
$$z = c(1 - \cos(u))$$

c'est une cycloïde.

Equations d'Euler-Lagrange en mécanique

Si un système mécanique est constitué de n éléments reliés entre eux par des liaisons parfaites (sans frottement), on aura la position du système qui dépendra de n paramètres indépendants (coordonnées généralisées notées $q_1, \ldots, q_n$). Ainsi, $\mathcal{E}_c$ l'énergie cinétique est une forme quadratique en les $\dot{q}_i$ et $\mathcal{E}_p$ l'énergie potentielle est une fonction des paramètres $q_1, \ldots, q_n$. Principe d'Hamilton (dit encore de moindre action)[2] : La trajectoire (c'est-à-dire le vecteur formé des fonctions $t \mapsto q_i(t)$) est donnée par les fonctions qui minimisent

$$J(q) = \int_a^b L(q(t), \dot{q}(t))dt, \tag{5.19}$$

[2]En fait, ce principe n'est vérifié que pour des temps suffisamment courts, cependant on peut le remplacer par le "principe de stationnarité" : La trajectoire est donnée par les fonctions qui rendent stationnaire $J(q)$ c'est-à-dire telles que $J'(q_0) = 0$.

avec $L(q,\dot{q}) = \mathcal{E}_c - \mathcal{E}_p$. Pour écrire les équations d'Euler-Lagrange il faut déterminer L le Lagrangien, le travail élémentaire de chaque forces interne et externe D_i, ainsi que le travail des forces de frottements $(-\frac{\partial D}{\partial \dot{q}_i} dq_i)$ donnant lieu à l'énergie dissipe D. On obtient alors le système d'équations d'Euler-Lagrange :

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} + \frac{\partial D}{\partial \dot{q}_i} = D_i. \tag{5.20}$$

Remarque 5.3.2. $\mathcal{E}_c$ dépend des q_i et de leurs dérivées $\dot{q}_i$ alors que $\mathcal{E}_p$ ne dépend que des q_i.

Remarque 5.3.3. $\mathcal{E}_c = \frac{1}{2}\dot{q}^{\mathrm{T}} M(q)\dot{q}$, où $M(q)$ est une matrice $n \times n$ symétrique définie positive.

Exemple 5.3.2. Une masse m est attachée à un pivot sans frottement à l'aide d'un fil non pesant de longueur l.

$$L = \frac{1}{2} m\dot{\theta}^2 - mgl\cos(\theta), D = 0. \tag{5.21}$$

En appliquant directement (5.20), on obtient :

$$(m\dot{\theta})(\ddot{\theta} + gl\sin(\theta)) = 0. \tag{5.22}$$

5.4 Que faire dans d'autres cadres

Parmi les problèmes mentionnés, certains font intervenir 1) des conditions terminales pouvant évoluer sur une contrainte, 2) des solutions $\mathcal{C}^1$ par morceaux, 3) des dérivées d'ordre supérieur et enfin 4) des contraintes diverses (égalité, inégalité, intégrale etc...). Nous allons compléter la CN d'Euler dans chacun de ces cas en donnant l'idée principale de la preuve.

1) Si $q(a)$ et $q(b)$ ne sont pas fixées : en utilisant (5.13) on obtient les conditions frontières

$$\phi(a) = 0 = \frac{\partial L}{\partial \dot{q}}(a, q_0(a), \dot{q}_0(a)) - C,$$
$$\phi(b) = \frac{\partial L}{\partial \dot{q}}(b, q_0(b), \dot{q}_0(b)) - C,$$

et la condition $J'(q_0) = 0$ (5.14) devient

$$\left(\frac{\partial L}{\partial \dot{q}}(b, q_0(b), \dot{q}_0(b))\right)\delta q(b) - \left(\frac{\partial L}{\partial \dot{q}}(a, q_0(a), \dot{q}_0(a))\right)\delta q(a) = 0$$

ce qui nous donne les conditions frontières suivantes (faire $\delta q(a) = 0$ et $\delta q(b)$ quelconque puis permuter)

$$\frac{\partial L}{\partial \dot{q}}(a, q_0(a), \dot{q}_0(a)) = 0, \tag{5.23}$$

$$\frac{\partial L}{\partial \dot{q}}(b, q_0(b), \dot{q}_0(b)) = 0. \tag{5.24}$$

Dans un cadre plus général, si les conditions aux frontières doivent vérifier des contraintes $q(a) = \Psi_1(a), q(b) = \Psi_2(b)$, les fonctions Ψ_i définissant des surfaces $(y = \Psi_i(x))$ auxquelles les points de départ et d'arrivée de la solution doivent appartenir ; on doit vérifier les conditions de transversalités suivantes

$$\left[L + \left(\dot{\Psi}_1(t) - \dot{q}(t)\right) \frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=a} = 0, \tag{5.25}$$

$$\left[L + \left(\dot{\Psi}_2(t) - \dot{q}(t)\right) \frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=b} = 0. \tag{5.26}$$

2) Aux points de discontinuité les conditions de "coins" dites de Weierstrass-Erdmann doivent être vérifiées (si $t = c$ est un point de 1ère espèce) :

$$\left[\frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=c^-} = \left[\frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=c^+}, \tag{5.27}$$

$$\left[L - \dot{q}(t)\frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=c^-} = \left[L - \dot{q}(t)\frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=c^+}. \tag{5.28}$$

3) On cherche à minimiser

$$J(q) = \int_a^b L(t, q(t), \dot{q}(t), \ldots, \frac{d^n q}{dt^n}(t))dt.$$

avec des conditions initiales et finales adéquates $(q(a), \dot{q}(a), \ldots, \frac{d^{n-1}q}{dt^{n-1}}(a)$ et $q(b)$, $\dot{q}(b)$, $\ldots$, $\frac{d^{n-1}q}{dt^{n-1}}(b)$ fixées). On travaille sur $E_n = \mathcal{C}^n([a, b], \mathbb{R})$ muni de

$$d_n(f, g) = \sum_{i=0}^{n} \left\{ \max_{t\in[a,b]} \left\| \left(\frac{\partial^i f}{\partial t^i}\right)_t - \left(\frac{\partial^i g}{\partial x^i}\right)_t \right\| \right\}.$$

On obtient

$$J'_{q_0}(\delta q) = \int_a^b \left(\sum_{i=0}^{n} \frac{\partial L}{\partial q^{(i)}}(t, q_0(t), \ldots, \frac{d^n q_0}{dt^n}(t))\delta q^{(i)} \right) dt.$$

En utilisant le lemme 5.5.3, on en déduit la condition d'Euler

$$\sum_{i=0}^{n} (-1)^i \frac{d^i}{dt^i} \left[\frac{\partial L}{\partial q^{(i)}} \left(t, q_0(t), \ldots, \frac{d^n q_0}{dt^n}(t) \right) \right] = 0 \tag{5.29}$$

4) On considère le problème parmi l'ensemble des fonctions continûment différentiables $q_1, \ldots, q_n$ qui satisfont les contraintes

$$\begin{aligned} g_j(t, q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n) = 0, j \in \{1, \ldots, m\}, \\ m < n, \\ q_i(a) = q_i^a, q_i(b) = q_i^b, i \in \{1, \ldots, n\}, \end{aligned}$$

trouver celles qui rendent extremum le critère

$$J(q_1, \ldots, q_n) = \int_a^b L(t, q(t), q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n) dt.$$

Si, rang $\left(\frac{\partial g_j}{\partial \dot{q}_i}\right) = m$, alors les q_i sont solutions des équations d'Euler en remplaçant L par

$$H(t, q(t), q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n) = L(t, q(t), q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n) + \tag{5.30}$$

$$\sum_{i=1}^{m} \lambda_i(t) g_i(t, q_1, \ldots, q_n, \dot{q}_1, \ldots, \dot{q}_n). \tag{5.31}$$

5.5 Quelques résultats annexes

Lemme 5.5.1 (Dubois-Reymond)**.** *Si f est une fonction continue sur un intervalle $[a, b]$ à valeur dans $\mathbb{R}^n$ et si pour toute fonction h continue s'annulant aux deux extrémités de $[a, b]$ ($h(a) = h(b) = 0$) on a*

$$\int_a^b \langle f(t), h(t) \rangle \, dt = 0$$

alors f est identiquement nulle sur $[a, b]$. ■

Ce résultat est un cas particulier du suivant :

Lemme 5.5.2 (Haar)**.** *Soit E un espace vectoriel. Si f est une fonction continue sur un intervalle $[a, b]$ à valeur dans le dual de E (E^*) et si pour toute fonction $h \in \mathcal{C}^p([a, b], E)$ (donc à valeur dans E) s'annulant aux deux extrémités de $[a, b]$ ($h(a) = h(b) = 0$) on a*

$$\int_a^b \langle f(t), h(t) \rangle \, dt = 0$$

alors f est identiquement nulle sur $[a, b]$. ■

Démonstration. Supposons que f ne soit pas identiquement nulle alors il existe un t_0 de $[a, b]$ tel que $f(t_0) \neq 0$. Alors dans $\mathcal{V} =]t_0 - \varepsilon, t_0 + \varepsilon[$ un voisinage de

ce point et par continuité de $\langle f(t), v\rangle$: on a $\langle f(t), v\rangle > 0$ (si négatif prendre l'opposé). En utilisant

$$\psi(t) = \begin{cases} 0 \text{ si } x \notin \mathcal{V} \\ (\varepsilon^2 - (t-t_0)^2)^{p+1} \text{ si } x \in \mathcal{V} \end{cases}$$

on a $\int_a^b \langle f(t), \psi(t)v\rangle\, dt = \int_{t_0-\varepsilon}^{t_0+\varepsilon} \langle f(t), \psi(t)v\rangle\, dt > 0$ et $\psi(t)v$ vérifie les hypothèses du lemme (contradiction). □

Lemme 5.5.3. *Soit E un espace vectoriel. Si les p fonctions f_i sont continue sur un intervalle $[a,b]$ à valeur dans le dual de E (E^*) et si pour toute fonction $h \in \mathcal{C}^p([a,b], E)$ (donc à valeur dans E) telle que les $(p-1)-$ème dérivées successives s'annulent aux deux extrémités de $[a,b]$ ($h(a) = h(b) = h'(a) = h'(b) = \ldots = h^{(p-1)}(a) = h^{(p-1)}(b) = 0$), on a*

$$\int_a^b \sum_{i=0}^{p} \left\langle f_i(t), h^{(i)}(t) \right\rangle dt = 0$$

alors $\sum_{i=0}^{p} (-1)^i \frac{d^i f_i}{dt^i}$ est identiquement nulle sur $[a,b]$. ■

5.6 Exercices

Fonctionnelles

Exercice 5.6.1. Etudier la continuité de la fonctionnelle

$$J : \mathcal{C}^0([a,b], \mathbb{R}^n) \to \mathbb{R}^n$$
$$f \mapsto J(f) = \sqrt{\int_a^b (f(x))^2\, dx}$$

on précisera la topologie utilisée.

Solution 5.6.1. *On considère l'espace*

$$E = \mathcal{C}^0([a,b], \mathbb{R}^n). \tag{5.32}$$

muni de la topologie induite par la distance

$$d_0(f,g) = \sup_{t\in[a,b]} \|f(t) - g(t)\|,$$

*dite **distance** de **Whitney** (au sens $\mathcal{C}^0$) induisant la norme*

$$n_0(f) = \sup_{t\in[a,b]} \|f(t)\|$$

La continuité de J découlera de celle de la fonctionnelle $f \mapsto \int_a^b (f(x))^2\,dx$ (par composition), or

$$\left|\int_a^b (f(x))^2\,dx - \int_a^b (g(x))^2\,dx\right| = \left|\int_a^b (f-g)(f+g)\,dx\right| \leq (b-a)d_0(f,g)n_0(f)n_0(g)$$

donc si on se fixe $\varepsilon > 0$ pour avoir $\left|\int_a^b (f(x))^2\,dx - \int_a^b (g(x))^2\,dx\right| < \varepsilon$ il suffit de prendre $d_0(f,g) \leq \delta$ avec $\delta = \frac{\varepsilon}{(b-a)n_0(f)n_0(g)}$ (fini).

Exercice 5.6.2. Soit $\mathcal{C}^p([a,b],\mathbb{R}^n)$ muni de la norme

$$\|f\|_{\mathcal{C}^p} = \max_{i\in\{0,\ldots,p\}} \left\{ \sup_{t\in[a,b]} \|f(t)\|, \sup_{t\in[a,b]} \left\|f'(t)\right\| \ldots, \sup_{t\in[a,b]} \left\|f^{(p)}(t)\right\| \right\}$$

ou de la norme

$$\|f\|^{\mathcal{C}^p} = \sum_{i=0}^{p} \sup_{t\in[a,b]} \left\|f^{(i)}(t)\right\|$$

Montrer que la continuité d'une fonctionnelle pour la topologie induite par la première norme est équivalente à la continuité pour la topologie induite par la seconde norme.

Solution 5.6.2. *Cela vient du fait que ces deux normes sont équivalentes :*

$$\|f\|_{\mathcal{C}^p} \leq \|f\|^{\mathcal{C}^p} \leq p\,\|f\|_{\mathcal{C}^p}$$

Exercice 5.6.3. Soit $J(q)$ une fonctionnelle différentiable, montrer que $H(q) = J(q)^2$ est différentiable calculer sa dérivée. Généraliser : soit $g : \mathbb{R} \mapsto \mathbb{R}$ dérivable, montrer que $g(J(q))$ est une fonctionnelle différentiable telle que

$$[g(J(q))]' = g'(J(q))J'_q$$

Solution 5.6.3. *$H(q+h) - H(q) = J(q+h)^2 - J(q)^2 = (J(q+h) - J(q))(J(q+h) + J(q))$ (en effet $J(q+h)$ est un réel). Par la différentiabilité de J on obtient :*

$$\begin{aligned} H(q+h) - H(q) &= \left(J'(q)h + \varepsilon(h)\right)\left(2J(q) + J'_q(h) + \varepsilon(h)\right), \\ &= 2J(q)J'_q(h) + E(h), \\ E(h) &= \varepsilon(h)\left[2J'_q(h) + 2J(q) + \varepsilon(h)\right] + \left(J'_q(h)\right)^2, \\ \lim_{n_1^C(h)\to 0} E(h) &= 0 \text{ puisque } \lim_{n_1^C(h)\to 0} \varepsilon(h) = 0 \end{aligned}$$

De façon plus générale

$$\begin{aligned} g(J(q+h)) - g(J(q)) &= g\left(J(q) + J'_q(h) + \varepsilon(h)\right) - g(J(q)), \\ &= g'(J(q))J'_q(h) + R(h) \\ R(h) &= \varepsilon(h)\left[g'(J(q) + J'_q(h))\right] + r(J'_q(h) + \varepsilon(h)), \\ \lim_{n_1^C(h)\to 0} R(h) = 0 \text{ } &\textit{puisque} \lim_{n_1^C(h)\to 0} \varepsilon(h) = 0 \text{ } \textit{et} \lim_{x\to 0} r(x) = 0 \end{aligned}$$

Problèmes variationnels simples

Exercice 5.6.4. Analyser les problèmes variationnels suivants

$$\begin{aligned} &\text{a)}\sqrt{\int_0^1 (f(x))^2\, dx}, \\ &\text{b)}\int_0^1 f'(x)dx, \\ &\text{c)}\int_0^1 f(x)f'(x)dx, \\ &\text{d)}\int_0^1 xf(x)f'(x)dx \end{aligned}$$

avec les conditions frontières $f(0) = 0, f(1) = 1$.

Solution 5.6.4. *a) idem à* $\int_0^1 (f(x))^2\, dx$ *pas de solution puisque la CN d'Euler* ($\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$ *avec* $L = f^2, \frac{\partial L}{\partial f'} = 0, \frac{\partial L}{\partial f} = 2f$) *s'écrit* $0 = 2f$ *donc* $f = 0$ *qui n'est pas compatible avec les conditions aux limites (il se peut que des solutions discontinues existent !). b)* ∞ *de solutions puisque la CN d'Euler* ($\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$ *avec* $L = f', \frac{\partial L}{\partial f'} = 1, \frac{\partial L}{\partial f} = 0$) *s'écrit* $0 = 0$ *elle est vérifiée (remarque :* $\int_0^1 f'(x)dx = f(1) - f(0) = 1$*). c)* ∞ *de solutions puisque la CN d'Euler* ($\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$ *avec* $L = ff', \frac{\partial L}{\partial f'} = f, \frac{\partial L}{\partial f} = f'$) *s'écrit* $f' = f'$ *elle est vérifiée (remarque :* $\int_0^1 ff'(x)dx = \frac{1}{2}(f^2(1) - f^2(0)) = 1$*) d) pas de solution puisque la CN d'Euler* ($\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) = \left(\frac{\partial L}{\partial f}\right)$ *avec* $L = xf(x)f'(x), \frac{\partial L}{\partial f'} = xf(x), \frac{\partial L}{\partial f} = xf'(x)$) *s'écrit* $f(x) + xf'(x) = xf'(x)$ *donc* $f = 0$ *qui n'est pas compatible avec les conditions aux limites. (mais il se peut que des solutions discontinues existent)*

Exercice 5.6.5. Déterminer les extremum des fonctionnelles suivantes

$$\begin{aligned} &\text{a)}\int_a^b \left(f(x)^2 + f'(x)^2 - 2f(x)\sin(x)\right) dx, \\ &\text{b)}\int_a^b \left(f(x)^2 + f'(x)^2 + 2f(x)\exp(x)\right) dx, \end{aligned}$$

Solution 5.6.5. *a) La CN d'Euler s'écrit*

$$f''(x) = f(x) - \sin(x)$$

donc

$$f(x) = \frac{1}{2}\sin(x) + \alpha\exp(x) + \beta\exp(-x)$$

qui est analytique donc on peut utiliser le second résultat et conclure que se sont les seules fonctions $\mathcal{C}^1$. *b) La CN d'Euler s'écrit*

$$f''(x) = f(x) + \exp(x)$$

donc

$$f(x) = \frac{1}{2}\sin(x) + \alpha\exp(x) + \beta\exp(-x)$$

qui est analytique donc on peut utiliser le second résultat et conclure que se sont les seules fonctions $\mathcal{C}^1$.

Exercice 5.6.6. On suppose que $q \in \mathcal{C}^1([a,b],\mathbb{R}^n)$ et qu'elle vérifie l'équation d'Euler suivante

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \left(\frac{\partial L}{\partial q}\right)$$

Montrer que si L est $\mathcal{C}^2$ par rapport à chacun de ses arguments et que $\left(\frac{\partial^2 L}{\partial \dot{q}^2}\right) \neq 0$ alors q est $\mathcal{C}^2$.

Solution 5.6.6. *Si* $\ddot{q}$ *existe on a*

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) = \frac{\partial}{\partial t}\left(\frac{\partial L}{\partial \dot{q}}\right) + \left[\frac{\partial}{\partial q}\left(\frac{\partial L}{\partial \dot{q}}\right)\right]\dot{q} + \left[\frac{\partial}{\partial \dot{q}}\left(\frac{\partial L}{\partial \dot{q}}\right)\right]\ddot{q},$$
$$\ddot{q} = \frac{1}{\left(\frac{\partial^2 L}{\partial \dot{q}^2}\right)}\left[\left(\frac{\partial L}{\partial q}\right) - \frac{\partial}{\partial t}\left(\frac{\partial L}{\partial \dot{q}}\right) - \left[\frac{\partial}{\partial q}\left(\frac{\partial L}{\partial \dot{q}}\right)\right]\dot{q}\right],$$

le membre de droite existe et est continue d'où le résultat (pour faire plus propre former les ratios adéquates).

Exercice 5.6.7. Parmi toutes les courbes joignant deux points (x_a, y_a) et (x_b, y_b), déterminer celle, qui par rotation autour de l'axe des x, engendre la surface d'air minimum.

Solution 5.6.7. *Par rotation autour de l'axe des* x, *la courbe* $y(x)$ *engendre une surface d'air*

$$A = 2\pi\int_{x_a}^{x_b} y(x)\sqrt{1+y'^2(x)}dx,$$

en utilisant le résultat du cours pour des fonctionelles faisant intervenir l'abscisse curviligne on obtient

$$\begin{aligned}
1 - \frac{yy''}{1+y'^2} &= 0, \\
2\int \frac{y'}{y} &= \int \frac{2y''y'}{1+y'^2}, \\
y^2 &= C^2\left(1+y'^2\right), \\
\int \frac{Cdy}{\sqrt{y^2-C^2}} &= \int dx, \\
\frac{x+D}{C} &= \ln\left(\frac{y+\sqrt{(y^2-C^2)}}{C}\right), \\
y &= C\cosh\left(\frac{x+D}{C}\right).
\end{aligned}$$

Exercice 5.6.8. En utilisant les équations d'Euler, déterminer les extremum de la fonctionnelle suivante :

$$\int_{t_0}^{t_1}\left(\frac{1}{2}m\dot{x}^2 - R(x)\right)dt, \tag{5.33}$$

lorsque $\operatorname{grad} R(x) = kx$, puis $\operatorname{grad} R(x) = kx(1-x^3)$. Interprétations ?

Solution 5.6.8. *On obtient*

$$m\ddot{x} = -\operatorname{grad} R(x)$$

C'est un ressort : dans le premier cas avec une raideur linéaire puis non linéaire.

Exercice 5.6.9. En utilisant les équations d'Euler, déterminer les extremum de la fonctionnelle suivante :

$$J(x) = \int_0^1 \sqrt{t^2+x(t)^2}\sqrt{1+\dot{x}(t)^2}dt, \tag{5.34}$$

Indication : utiliser le changement de variables

$$\begin{aligned}
t &= r\cos(\theta) \\
x &= r\sin(\theta)
\end{aligned}$$

Exercice 5.6.10. Quelle courbe minimise l'intégrale

$$\int_0^1\left(\frac{1}{2}f'^2(x) + f(x)f'(x) + f'(x) + f(x)\right)dx,$$

avec les conditions frontières $f(0) = 0, f(1) = 1$.

Exercice 5.6.11. Quelle courbe minimise l'intégrale

$$\int_0^1 \left(f'(x) + f(x)\right)^2 dx,$$

avec les conditions frontières $f(0) = 0, f(1) = 1$.

Solution 5.6.9. ***Réponse :*** *La CN d'Euler donne*

$$\begin{aligned}
L &= f^2 + 2ff' + f'^2 \\
\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) &= \left(\frac{\partial L}{\partial f}\right) \\
\left(\frac{\partial L}{\partial f}\right) &= 2f + 2f' \\
\left(\frac{\partial L}{\partial f'}\right) &= 2f + 2f' \\
\frac{d}{dx}\left(\frac{\partial L}{\partial f'}\right) &= 2f' + 2f'' \\
f'' &= f \\
f(x) &= a\exp(x) + b\exp(-x) \\
f(0) &= 0 = a + b \\
f(1) &= 1 = ae + \frac{b}{e} \\
b &= -a = \frac{e}{1 - e^2}
\end{aligned}$$

Problèmes variationnels complémentaires

Exercice 5.6.12. Déterminer l'extremum de

$$J(q) = \int_0^1 (1 + \ddot{q}^2)dt, \tag{5.35}$$

compatible avec les conditions $q(0) = 0, \dot{q}(0) = 1, q(1) = 1, \dot{q}(1) = 2$.

Solution 5.6.10. *On doit avoir* $\sum_{i=0}^{2}(-1)^i \frac{d^i}{dt^i}\left[\frac{\partial L}{\partial q^{(i)}}\left(t, q_0(t), \ldots, \frac{d^n q_0}{dt^n}(t)\right)\right] = 0$, *c'est-à-dire*

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} + \frac{d^2}{dt^2}\frac{\partial L}{\partial \ddot{q}} = 0$$

or $\frac{\partial L}{\partial q} = \frac{\partial L}{\partial \dot{q}} = 0$ *donc* $\frac{d^2}{dt^2}\frac{\partial L}{\partial \ddot{q}} = 2q^{(4)} = 0$. *On en déduit*

$$\begin{aligned}
q(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3, \\
q(0) &= a_0 = 0, \\
\dot{q}(0) &= a_1 = 1, \\
q(1) &= a_0 + a_1 + a_2 + a_3 = 1, \\
\dot{q}(1) &= a_1 + 2a_2 + 3a_3 = 2,
\end{aligned}$$

finalement $q(t) = t - t^2 + t^3$. *Attention, on peut être tenté de poser* $Q = \dot{q}$, *et alors on obtient* $\int_0^1 (1+\ddot{q}^2)dt = \int_0^1 (1+\dot{Q}^2)dt : \frac{\partial L}{\partial Q} = \frac{d}{dt}\frac{\partial L}{\partial \dot{Q}}, \ddot{Q} = 0 : q(t) = a_0 + a_1 t + a_2 t^2$. *Que se passe-t-il ?*

Exercice 5.6.13. Reprendre l'exercice précédent avec une extrémité libre $\dot{q}(1)$ non fixée.

Solution 5.6.11. *En reprenant la preuve du théorème*

$$J'_{q_0}(\delta q) = \int_a^b \left(\frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\delta \dot{q} + \frac{\partial L}{\partial \ddot{q}}\delta \ddot{q} \right) dt.$$

En posant $\phi(x) = \int_a^x \frac{\partial L}{\partial q} dt, \psi(x) = \int_a^x \left(\frac{\partial L}{\partial \dot{q}} - \phi(t) \right)$, *on obtient*

$$J'_{q_0}(\delta q) = [\phi(x)\delta q]_a^b + [\psi(x)\delta \dot{q}]_a^b + \int_a^b \left(\frac{\partial L}{\partial \ddot{q}} - \psi(x) \right) \delta \ddot{q} dt$$

Euler devient

$$\frac{\partial L}{\partial \ddot{q}}(t, q_0(t), \dot{q}_0(t)) - \psi(t) = C, \tag{5.36}$$

$$J'_{q_0}(\delta q) = 0 = [\phi(x)\delta q]_a^b + [\psi(t)\delta \dot{q}]_a^b + [C\delta \dot{q}]_a^b, \tag{5.37}$$

la condition frontière est donc

$$\begin{aligned} \frac{\partial L}{\partial \ddot{q}}(1, q_0(1), \dot{q}_0(1)) &= 0, \\ \ddot{q}(1) &= 0 = 2a_2 + 6a_3, \\ q(0) &= a_0 = 0, \\ \dot{q}(0) &= a_1 = 1, \\ q(1) &= a_0 + a_1 + a_2 + a_3 = 1, \end{aligned}$$

finalement $q(t) = t$.

Exercice 5.6.14. Déterminer les courbes qui rendent extremum la fonctionnelle

$$J(q) = \frac{1}{2}\int_0^1 (q^2 + \dot{q}^2)dt, \tag{5.38}$$

avec $q(0) = 0$ et le point terminal se trouvant :

1. sur la droite $t = 1$,
2. sur la courbe $y = t^2$.

Solution 5.6.12. *On se trouve dans un problème avec conditions frontières non fixées. L'équation d'Euler s'écrit*

$$\begin{aligned} \ddot{q} &= q, \\ q\,(t) &= C\sinh(t), \\ \dot{q}(t) &= C\cosh(t) \end{aligned}$$

1. *Il faut* $\frac{\partial L}{\partial \dot{q}}(1, q(1), \dot{q}(1)) = \dot{q}(1) = 0$, *donc* $C = 0$ *:* $q(t) = 0$. *(intuitivement ça colle :* $J = 0$ *!).*
2. *Cette fois la condition supplémentaire est*

$$\left[L + (2t - \dot{q}(t))\,\frac{\partial L}{\partial \dot{q}}(t, q(t), \dot{q}(t))\right]_{t=1} = \frac{1}{2}(q^2 + \dot{q}^2) + \dot{q}\,(2 - \dot{q}) = 0,$$

soit

$$\begin{aligned}(q^2 - \dot{q}^2) &= 4\dot{q},\\ C(\sinh(1)^2 - \cosh(1)^2) &= 4\cosh(1),\\ C &= -4\cosh(1),\\ q(t) &= -4\cosh(1)\sinh(t)\end{aligned}$$

Exercice 5.6.15. Déterminer les courbes qui rendent extremum la fonctionnelle

$$J(y) = \int_0^{x_2} \frac{\sqrt{1 + y'^2}}{y} dx, \tag{5.39}$$

avec $y(0) = 0$ et le point terminal se trouvant :

1. sur la droite $y = x + 1$,
2. sur le cercle $(x - 4)^2 + y^2 = 4$.

On commencera par examiner les conditions de transversalités du cours pour des fonctionnelles du type

$$\int_{x_1}^{x_2} f(x, y)\sqrt{1 + y'^2} dx,$$

Exercice 5.6.16. On cherche le ou les extremum(s) de

$$J(q) = \int_a^b L(t, q, \dot{q}) dt \tag{5.40}$$

tel que $q(a) = q_a, q(b) = q_b$ et $\int_a^b C(t, q, \dot{q}) dt = c$. Soit q_0 un telle fonction, montrer qu'il existe une constante λ telle que q_0 soit un extremum de la fonctionnelle suivante

$$\begin{aligned}J_H(q) &= \int_a^b H(t, q, \dot{q}) dt,\\ H(t, q, \dot{q}) &= L(t, q, \dot{q}) + \lambda C(t, q, \dot{q}).\end{aligned}$$

Indication : poser $y(t) = \int_a^t C(t, q, \dot{q}) dt$, on cherche un extremum de J sous la contrainte $\dot{y} = C(t, q, \dot{q})$, avec $y(b) = c$. Application : résoudre les problèmes 3 et 4 du cours : 3. Problème de l'isopérimètre (Euler) : Déterminer la courbe embrassant l'air maximale parmi l'ensemble des courbes fermées de classe $\mathcal{C}^1$

et de longueur donnée l. On peut sans perte de généralité, considérer que ces courbes sont issues de l'origine et que l'autre extrémité a pour abscisse x_B. La longueur est donc $l = \int_0^{x_B} \sqrt{1 + (f'(x))^2} dx$, quand à l'air de révolution c'est

$$J_3(f) = 2\pi \int_0^{x_B} f(x)\sqrt{1 + (f'(x))^2} dx. \tag{5.41}$$

Ce problème est un peu plus complexe puisqu'il faut minimiser un critère sous une contrainte (la longueur doit être de l). 4. Parmi tous les arcs de courbes de classe $\mathcal{C}^1$ joignant A et B deux points donnés du plan et de longueur donnée l, déterminer celui qui délimite avec le segment AB une aire maximum. On peut sans perte de généralité, considérer que ces courbes sont issues de l'origine et que l'autre extrémité a pour abscisse x_B. La longueur est donc $l = \int_0^{x_B} \sqrt{1 + (f'(x))^2} dx$, quand à l'air délimitée par la courbe et l'axe des abscisses c'est

$$J_4(f) = \int_0^{x_B} f(x) dx. \tag{5.42}$$

Ce problème fût posé et résolut (de façon intuitive) par la reine Dido de Carthage en 850 AV-JC. On retrouve un problème similaire à celui présenté en 3 (minimum sous contrainte). Une formulation plus générale consiste à déterminer la courbe qui définie la surface la plus grande parmi toutes les courbes fermées de périmètre donné (ce qui définit une contrainte).

Exercice 5.6.17. Un nageur partant de l'origine traverse une rivière de largeur b à vitesse constante c^2, la berge de départ coïncide avec l'axe des y, le courant de la rivière est $v(x)$ ($v^2 < c^2$). Déterminer la trajectoire que doit avoir le nageur pour rejoindre l'autre berge en un temps minimum. Application : $b = 1, v = c^2 x(1 - x)$.

Solution 5.6.13. *Le nageur ne peut définir sa trajectoire qu'en jouant sur l'angle d'attaque (α : l'angle que fait le nageur avec l'axe des x). On a*

$$\begin{aligned} \dot{x} &= c\cos\alpha, \\ \dot{y} &= c\sin\alpha + v, \end{aligned}$$

Donc le temps mis par le nageur pour traverser est :

$$t = \int_0^b \frac{dx}{\dot{x}} = \int_0^b \frac{dx}{c\cos\alpha}$$

Il ne reste plus qu'à exprimer $\cos(\alpha)$ *en fonction de* y, y' *:*

$$\begin{aligned} cy'\cos\alpha &= v \pm c\sqrt{1 - \cos^2\alpha}, \\ \left(cy'\cos\alpha - v\right)^2 &= c^2\left(1 - \cos^2\alpha\right) \end{aligned}$$

$$c^2\left(y'^2+1\right)\cos^2\alpha - 2vcy'\cos\alpha + \left(v^2-c^2\right) = 0$$

$$\begin{aligned}
\cos(\alpha) &= \frac{vy' \pm \sqrt{c^2y'^2-(v^2-c^2)}}{c\left(y'^2+1\right)},\\
t &= \int_0^b \frac{\left(y'^2+1\right)}{vy' \pm \sqrt{c^2y'^2-(v^2-c^2)}}dx\\
&= \int_0^b \frac{\left(y'^2+1\right)}{vy' \pm \sqrt{c^2y^{2\prime}-(v^2-c^2)}}dx\\
&= \int_0^b \frac{-vy' + \sqrt{c^2y'^2-(v^2-c^2)}}{(c^2-v^2)}dx
\end{aligned}$$

Or $L(x,y\prime)$ donc Euler Lagrange s'écrit : $\frac{\partial L}{\partial y'} = C$ de plus la condition frontière est $\left[\frac{\partial L}{\partial y'}\right]_{y=1} = 0$, on en déduit

$$\begin{aligned}
\frac{1}{(c^2-v^2)}\left[-v + \frac{c^2y'}{\sqrt{c^2y'^2-(v^2-c^2)}}\right] &= 0\\
c^2y' - v\sqrt{c^2y'^2-(v^2-c^2)} &= 0\\
cy' &= \pm v
\end{aligned}$$

Application $b=1, v=x(1-x)$

$$y = \pm\frac{1}{c}\left(\frac{x^2}{2} - \frac{x^3}{3}\right)$$

Nota $t = \int_0^1 \frac{y'}{v}dx = \frac{1}{c}$.

Exercice 5.6.18. James Bond à la poursuite de sa James Bond girl en voilier : (Tiré de "Principes variationnels et Mécanique analytique" par Jean-Louis Basdevant et Christoph Kopper, presse de l'X).

Un voilier évolu sur un plan d'eau à une vitesse v_b faisant un angle noté θ avec la vitesse du vent notée w. La vitesse du bateau v_b est proportionnelle à celle du vent et dépend de θ l'angle d'orientation du bateau (angle choisi par le capitaine du bateau). Cette vitesse est de la forme

$$v_b(\theta) = \frac{w}{\cos(\theta)h(\tan(\theta))}, \quad \text{avec } h(u) = \frac{1}{2}(u+\frac{1}{u}), \tag{5.43}$$

On s'intéresse la stratégie de "remontée au vent" du bateau, c'est-à-dire pour $\theta \leq \frac{\pi}{2}$, comme on le représente sur la figure (5.1). La vitesse v_b du bateau le long de l'axe Ox est opposée à celle du vent, et sa coordonnée x augmente toujours en fonction du temps. On suppose une côte rectiligne c'est-à-dire que la terre est le demi-plan $y<0$ alors que la mer est le demi-plan $y>0$. On suppose que le

vent est parallèle à la côte, de direction opposée à l'axe Ox, et que la norme de sa vitesse $w(y)$ ne dépend que de l'éloignement à la côte y. La vitesse du vent a la forme :

$$w(y) = w_0 - w_1 \frac{y_0}{y + y_0}, \tag{5.44}$$

où w_0 est la vitesse du vent loin de la côte, qui est supérieure à la vitesse $((w_0 - w_1) \geq 0)$ au bord de la côte $y = 0$.

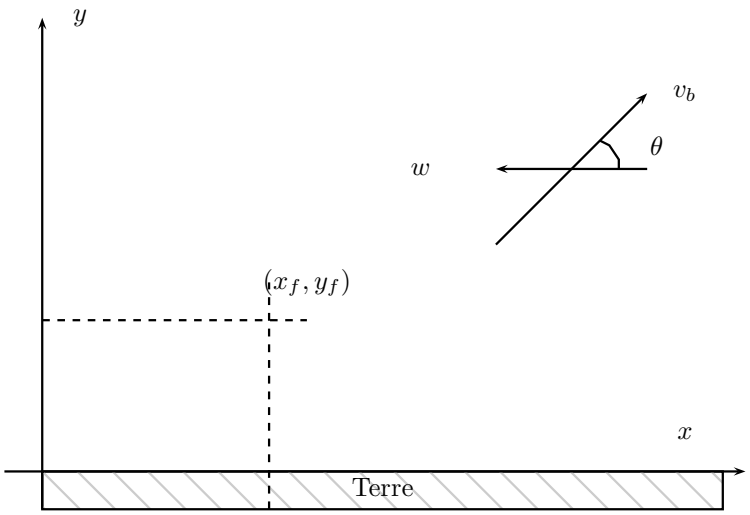


Fig. 5.1: Course de bateau

1. On note : $\dot{x} = \frac{dx}{dt}, \dot{y} = \frac{dy}{dt}$ montrer que $y' = \frac{dy}{dx} = \tan(\theta)$.
2. On suppose d'abord le vent uniforme $w = w_0$ $(w_1 = 0)$. Ecrire la vitesse du bateau suivant l'axe du vent $v_{bx} = \dot{x}$ en fonction de w et $h(\tan(\theta))$. Pour quelle valeur de θ et de y' cette vitesse est-elle maximum ? Quelle est alors sa valeur ?
3. On suppose maintenant que $w_1 \neq 0$. Le bateau va du point de départ, l'origine $(x = 0, y = 0)$, à un point d'arrivée au large $(x = x_f, y = y_f)$. On suppose que $\frac{dy}{dx} = y' \geq 0$ pour tout t (c'est-à-dire que le bateau ne vire jamais de bord). On veut déterminer la trajectoire $y(x)$ la plus rapide. Ecrire la valeur du temps total T pour aller du départ l'arrivée. En déduire l'équation qui détermine la trajectoire optimale. Montrer que l'invariance du problème par translation suivant Ox entraîne

$$\frac{h'(y')y' - h(y')}{w(y)} = A, \tag{5.45}$$

où A est une constante.

4. Utiliser le résultat précédent pour calculer la trajectoire sous la forme d'une fonction $x(y)$ (et non pas d'une fonction $y(x)$). Fixer la valeur de A.
5. Calculer la valeur de $\frac{dy}{dx} = y'$ en fonction de y. On suppose que $x_f >> y_f$ et $y_f << y_0$. Pensez-vous que le résultat obtenu corresponde effectivement à la meilleure stratégie ? Sinon, quelle modification doit-on apporter ?
6. Mister Bond (James de son prénom) part de l'origine avec son voilier. Il veut intercepter sa James Bond girl favorite. Celle-ci quitte le rivage à bord d'un bateau à moteur. Son point de départ est situé à une distance L de l'origine. Enfin, le bateau à moteur se déplace à une vitesse constante c en direction du large (selon l'axe des y). A quelles conditions James pourra retrouver sa belle ?

Exercice 5.6.1. On considère une pompe qui alimente un récipient dont le fond est équipé d'une vanne de fuite (cf. figure 5.2).

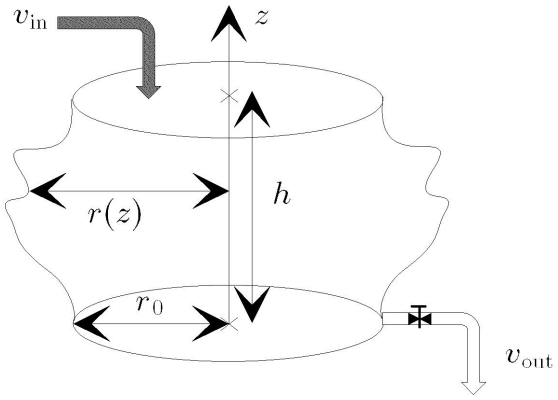


FIG. 5.2: Dipositif étudié

Partie I Etude du récipient

Pour construire le récipient, on dispose d'une tôle de surface S donnée et on façonne la tôle de sorte que le récipient ait une contenance maximale et sa forme soit un cône (cf. figure 5.3).

On rappelle que, pour un cône, le volume est $V = \frac{\pi}{3} h r^2$ et la surface est $S = \pi r \sqrt{h^2 + r^2}$.

Question a) Formuler ce problème. Donner le critère à optimiser et les contraintes éventuelles.

Question b) Donner les conditions nécessaires que doivent satisfaire les inconnues du problème.

Question c)Déterminer (r, h) en fonction de la surface de tôle S.

Partie II Etude de la vidange du récipient

Dans cette partie, le récipient est constitué d'un socle (disque de rayon r_0) et d'une tôle de surface S (donnée). Le profil de la tôle est engendré par rotation

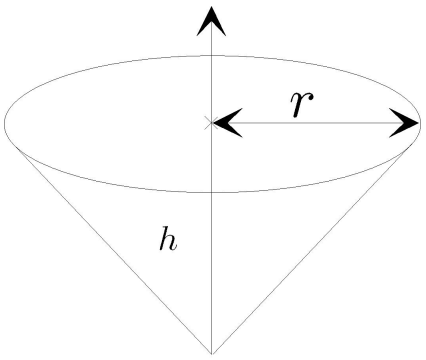


Fig. 5.3: Récipient conique

de la courbe $r(z)$ (qui est supposée $\mathcal{C}^1$) autour de l'axe orthogonal au socle et passant par le centre du socle (cf. figure 5.2). Si on note $z(t)$ la hauteur du liquide dans le récipient, le volume est

$$v(t) = \pi \int_0^{z(t)} r^2(z)dz. \tag{5.46}$$

Le bilan de matière donne

$$\frac{dv(t)}{dt} = v_{\text{in}} - v_{\text{out}},$$

le débit d'entrée (v_{in}) est constant (délivré par une pompe) et celui de sortie (v_{out}) est de la forme : $v_{\text{out}} = c\sqrt{P - P_a}$. Enfin l'équation de Bernouilli est $P = \rho g z + P_a$, avec ρ la masse volumique du liquide et g la constante gravitationnelle.

Question a) (1 point) Montrer que l'équation différentielle ordinaire que satisfait la hauteur $z(t)$ est :

$$\frac{dz(t)}{dt} = \frac{v_{\text{in}} - c\sqrt{\rho g}\sqrt{z(t)}}{\pi r^2(z(t))}$$

Question b) On cherche les extremums de

$$J(r) = \int_a^b L(z, r, r')dz \tag{5.47}$$

avec $r' = \frac{dr}{dz}$ et tel que $z(a) = z_a, z(b) = z_b$ et $\int_a^b C(z, r, r')dz = c$. Soit r_0 une telle fonction, montrer qu'il existe une constante λ telle que r_0 soit un extremum de la fonctionnelle suivante

$$J_H(r) = \int_a^b H(z, r, r')dz, \quad H(z, r, r') = L(z, r, r') + \lambda C(z, r, r').$$

Indication : poser $y(z) = \int_a^z C(z, r, r')dz$, on cherche un extremum de J sous la contrainte $y' = \frac{dy}{dz} = C(z, r, r')$, avec $y(b) = c$.

Question c) Pour une hauteur donnée h, l'aire est

$$A = 2\pi \int_0^h r(z)\sqrt{1 + r'^2}dz, \tag{5.48}$$

avec $r' = \frac{dr}{dz}$. On coupe l'alimentation de la pompe ($v_{\text{in}} = 0$). En partant d'une hauteur h donnée, le récipient se vide. Déterminer l'équation différentielle que doit satisfaire $r(z)$, le profil du récipient, pour que le temps de vidange du récipient soit le plus petit possible sous la contrainte que l'aire est fixée.

Question d)Déterminer en fonction du débit d'entrée constant, l'équilibre atteint noté z_{eq}. Discuter de sa nature et de sa stabilité.

Partie III Etude de la dynamique au voisinage d'un point de fonctionnement

Dans cette partie, on considère le récipient conique de la figure 2. On a la relation $r = z\tan(\theta_{\text{cône}})$, le volume (5.46) est donc

$$\begin{aligned} v(t) &= \pi \int_0^{z(t)} r^2(z)dz \\ &= \frac{\pi \tan^2(\theta_{\text{cône}})}{3} z^3(t) \end{aligned}$$

On considère un fonctionnement nominal autour duquel la pompe fonctionne ($v_{\text{in}} = v_{\text{in\_nominal}} + \delta v_{\text{in}}, u = u_{\text{nominal}} + \delta u$ avec $v_{\text{in\_nominal}}$ et u_{nominal} constants). Dans ce cas, la dynamique de la pompe est de la forme :

$$\tau \frac{d(\delta v_{\text{in}})}{dt} + \delta v_{\text{in}} = a\delta u$$

En posant $z = z_{\text{nominal}} + \delta z$, avec $z_{\text{nominal}} = \frac{v_{\text{in\_nominal}}^2}{c^2\rho g}$, on approximera $\frac{\sqrt{z}}{z^2}$.

Question a) Donner la dynamique linéarisée correspondante pour la variation de hauteur du liquide.

Question b) On applique un débit $\delta v_{\text{in}} = -k\delta z$, déterminer k tel que l'équilibre soit asymptotiquement stable.

Question c) On applique à la pompe une tension $\delta u = -\frac{2}{3}k\delta z$. On pose $x = (\delta v_{\text{in}}, \delta z)^T$. Mettre le système sous la forme $\frac{dx}{dt} = Ax$. Montrer que, pour les valeurs numériques suivantes : $\tau = 1, a = \frac{3}{2}, \theta_{\text{cône}} = 45°, z_{\text{nominal}} = 1, c = \frac{2}{3}, \rho = 0.1, g = 10$, on obtient

$$A = \begin{pmatrix} -1 & -k \\ \frac{1}{\pi} & -\frac{1}{3\pi} \end{pmatrix}.$$

Question d) Pour les valeurs numériques précédentes (Question c), calculer $\exp(At)$ par la méthode des matrices constituantes. En déduire les solutions du système $\frac{dx}{dt} = Ax$. Comment choisir k pour que l'équilibre soit asymptotiquement stable ?

6 | Systèmes à retard

J.P. Richard[1], M. Ksouri[2]

[1]LAGIS & INRIA-ALIEN, Ecole Centrale de Lille, BP 48, 59651 Villeneuve d'Ascq cedex, France. *E-mail* : Jean-Pierre.Richard@ec-lille.fr
[2]ENIT, BP 37, Le Belvédère 1002, Tunis, Tunisie. *E-mail* : Mekki.Ksouri@insat.rnu.tn

Cours rédigé par Jean-Pierre Richard

6.1 Introduction

Les systèmes à retards sont aussi appelés systèmes héréditaires, systèmes à post-effet, équations à argument différé[1] ou, encore, équations différentielles aux différences. Ils appartiennent à la classe des *équations différentielles fonctionnelles* (EDF) qui sont de dimension infinie, par opposition aux équations différentielles ordinaires (EDO, voir [118]). Ce chapitre ne peut détailler le grand nombre d'ouvrages dédiés aux systèmes à retards (plus de 30 monographies en anglais depuis 1963) mais le lecteur pourra se référer, par exemple, aux articles de synthèse [18, 60, 67, 82, 92, 103, 108, 116, 119, 138, 143], ou aux numéros spéciaux [23, 33, 84, 102, 122].

Quelles peuvent être les motivations d'une recherche si active et d'un intérêt si continu ? Les points suivants nous semblent donner quelques éléments de réponse.

[1]Traduction de l'anglais *deviating argument* en lien avec l'appellation *differential-difference equation*.

Un problème appliqué

Le retard est un phénomène physique qui se retyrouve dans une multitude d'applications : nombreux sont les systèmes réels dont l'évolution temporelle, contrairement à celle des systèmes « ordinaires », n'est pas définie à partir d'un simple vecteur d'*état* (exprimé au présent), mais dépend irréductiblement de l'histoire du système. Cette situation se rencontre dans les cas – nombreux – où un transport de matière, d'énergie ou d'information engendre un « temps mort » dans la réaction : en technologies de l'information et de la communication (réseaux de communication haut-débit [2, 10, 15, 56, 89, 95, 114, 130], contrôle des systèmes en réseau [15, 106, 120, 128], qualité de service dans les transmissions vidéo MPEG [91], systèmes télé-opérés [63, 79, 104, 105], calcul parallèle [1], calcul temps réel en robotique [5, 131]...), en dynamique des populations et en épidémiologie (temps de gestation ou d'incubation), en mécanique (viscoélasticité)...[2] Même si le processus ne contient pas intrinsèquement de post-effet, sa chaîne de commande peut introduire des retards (par exemple si les capteurs demandent un temps d'acquisition/transmission non négligeable). Pour ces raisons, il semble raisonnable de considérer le retard comme une caractéristique universelle de l'interaction entre l'homme et la nature (donc, des sciences pour l'ingénieur), au même titre que la non-linéarité, par exemple.

Un problème ancien mais encore ouvert

Les *équations différentielles fonctionnelles* (EDF) constituent un outil mathématique approprié à l'étude du phénomène d'hérédité, généralisant les équations différentielles ordinaires (EDO). Comme pour tous les systèmes dynamiques, leur investigation théorique inclut des sujets comme l'existence et l'unicité des solutions, leur périodicité, l'analyse des bifurcations, les problèmes aux limites, la commande et l'estimation, la caractérisation des comportements asymptotiques (stabilité, bornitude, moyennage, etc.), pour n'en citer que quelques-unes.

Les premières EDF ont été considérées par J. Bernoulli, L. Euler, J.L. Lagrange, P. Laplace, S. Poisson et d'autres, en lien avec certains problèmes géométriques posés au XVIII^{e} siècle. Au début du XX^{e} siècle, d'importantes applications en furent faites par V. Volterra [141]. La situation changea dans les années 30, avec l'apparition d'un grand nombre de problèmes techniques et scientifiques. La régulation sur base de modèles linéaires et stationnaires avec retard fut considérée par Y. Zypkin en 1941.

Les bases de la théorie moderne des EDF furent probablement posées par A.D. Myshkis en 1949 [97, 98]. En particulier, il fut le premier à formuler l'énoncé du problème de Cauchy pour des équations à retard arbitraire (ponctuel ou

[2] Des exemples plus détaillés pourront être trouvés dans [17, 66, 100].

distribué, fini ou infini). Les années suivantes[3] ont vu une explosion de la théorie des EDF et de leurs applications (voir par exemple [25, 66, 69, 72, 84, 99, 100, 116, 122] et les nombreuses références incluses). Dans ce chapitre, l'accent sera mis sur les grandes lignes de la théorie des équations différentielles à retards, permettant ainsi l'accès aux méthodes et résultats concrets qui en découlent.

Malgré le grand nombre d'outils d'analyse disponibles, la commande des systèmes à retards pose encore plusieurs problèmes ouverts [119]. On pourrait penser que les techniques de contrôle classiques (en dimension finie) seraient applicables après avoir remplacé les opérateurs de retard par des approximations rationnelles (fonction de transfert sans retard)[4]. Appliquées au calcul de lois de commande, de telles approximations peuvent donner des résultats probants dans le cas de systèmes linéaires à retards constants et connus. Cependant, elles trouvent rapidement leur limite, notamment parce qu'elles conduisent à des systèmes d'ordre élevé[5] dont la régulation n'est finalement pas plus simple que celle du modèle initial. En effet, dans ce cas des systèmes linéaires à retards constants et connus, plusieurs méthodes de synthèse de contrôleurs en boucle fermée donnent des résultats probants. Ces méthodes (placement de spectre, approches de type Lyapunov...) sont généralement basées sur un principe de prédicteur[6]. Le contrôleur intègre alors une anticipation des comportements futurs, possible si l'on dispose d'un bon modèle. Notons toutefois que, si la partie linéaire non retardée du modèle est d'ordre supérieur à 3 ou 4, le développement numérique et systématique de tels contrôleurs peut être difficile : en effet, dès que l'ordre et le nombre de paramètres de réglage augmentent, la phase d'analyse de stabilité doit être menée par l'intermédiaire de conditions suffisantes mais non nécessaires (nous verrons par la suite certaines de ces méthodes, de type Liapounov). La situation s'aggrave encore dans le cas des retards variables[7] ou

[3] En 1962, N.N. Krasovskii [74] publie une construction analytique de contrôle optimal en présence de retards, puis une généralisation de la seconde méthode de Liapounov [75].

[4] On pense ici aux approximations du type :

$$e^{-hs} \approx \frac{p(-hs)}{p(hs)}, \tag{6.1}$$

où $p \in \mathbb{R}[hs]$ est un polynôme dont les zéros sont tous dans le demi-plan complexe gauche $\mathrm{Re}\, s < 0$. Les approximants résultants sont ceux, classiques, de Padé au premier ordre $p(hs) = (1-\frac{hs}{2})$, mais aussi de Laguerre-Fourier $p(hs) = (1-\frac{hs}{2n})^n$, de Kautz $p(hs) = (1-\frac{hs}{2n}+\frac{h^2s^2}{8n^2})^n$, de Padé au second ordre $p(hs) = (1-\frac{hs}{2n}+\frac{h^2s^2}{12n^2})^n$ de Padé diagonal $p_n(hs) = \sum_{k=0}^{n} \frac{(2n-k)!(-hs)^k}{k!(n-k)!}$, $n \geq 3$.

[5] Une argumentation plus complète peut être trouvée dans [119].

[6] Le premier usage d'un prédicteur fut introduit par Smith [133] à la fin des années 1950. Il concernait des systèmes asymptotiquement stables en boucle ouverte et présentant un retard sur l'entrée.

[7] Même en dimension finie, on connaît les difficultés d'analyse rencontrées en non stationnaire : dans ce cas, la stabilité du modèle considéré comme « stationnaire à chaque instant »n'a aucun lien avec celle du système variant. Cette difficulté se transporte dans le cas retardé. Ainsi, l'équilibre $x = 0$ du système à retard variable, considéré dans [49] :

$$\dot{x}(t) = ax(t) + bx(t-h(t)), \quad h(t) = t-k \quad \forall t \in]k,(k+1)]\,,\ k \in \mathbb{N} \text{ (d'où } 0 \leq h(t) \leq 1), \tag{6.2}$$

mal connus.

Les particularités des systèmes à retards peuvent aussi être surprenantes : plusieurs études ont montré que l'introduction volontaire de retard peut améliorer la stabilisation d'équations ordinaires : amortissement et stabilisation [3, 121], résonateurs retardés [57], rejet de perturbation [58, 147], contrôle de cycle limite non linéaire [4], temps fini [143]...

Exercice 6.1.1. Simuler le système scalaire à retard (6.3) pour différentes conditions initiales. Montrer que $x(t) = 1$ est un équilibre et observer le comportement chaotique obtenu pour la condition initiale $x(t) = 2\ \forall t \in [-1, 0]$.

$$\dot{x}(t) = -5x(t) + 10\frac{x(t-1)}{1 + x(t-1)^8}. \tag{6.3}$$

Pour un système d'ordre 1 sans retard, les phénomènes observés (chaos, trajectoire interceptant le point d'équilibre) auraient-ils été possibles ?

Un outil de modélisation

Au delà des effets physiques de post-effet, l'utilisation d'un opérateur de retard peut être intéressante dans la phase de modélisation.

Ainsi, pour des systèmes d'ordre élevé (ou infini), on connaît le classique modèle de Strejć (voir par exemple [12]), très prisé en génie des procédés, où le phénomène (de diffusion, par exemple) est volontairement représenté par un transfert de pôle multiple τ (d'ordre n en général restreint à 1 ou 2) et un retard h :

$$F(s) = k\frac{e^{-ht}}{(1 + \tau s)^n}. \tag{6.4}$$

C'est que, malgré leur complexité, les systèmes à retards constituent une classe de modèles en dimension infinie relativement *simples* lorsqu'on les compare à la classe des systèmes d'équations aux dérivées partielles (EDP). Citons ici [65] : *It is usually not difficult to show that the appearance of delay in a differential equation results of some essential simplification of the model.* Les EDP hyperboliques peuvent être localement écrites en tant que systèmes à retards de type neutre [48, 69] grace à la transformation de d'Alembert. D'autres relations avec les équations à dérivées d'ordre fractionnaire ont aussi été établies [50].

Inversement, tout effet de retard $y(t) = u(t-h)$ peut être représenté par une classique équation de transport sur une distance l et à vitesse $c = lh^{-1}$:

$$\begin{gathered} h\tfrac{\partial}{\partial t}x(z,t) + \tfrac{\partial}{\partial z}x(z,t) = 0, \quad z \in [0,l], \\ x(0,t) = u(t), \quad y(t) = x(l,t). \end{gathered} \tag{6.5}$$

est-il instable pour $a = -3.5$ et $b = -4$ alors que les valeurs propres du système considéré à chaque instant ont toutes des parties réelles négatives. Pour $a = -1$, $b = 1.5$, il est asymptotiquement stable, alors que les conditions de stabilité ne seraient pas vérifiées si le retard était constant entre 0 et 1. D'autres contre-exemples sont donnés dans [86].

Signalons aussi que le retard permet de modéliser les effets de discrétisation temporelle tout en restant dans le domaine des équations différentielles en temps continu [30, 31]. En effet, une loi de commande $u(t)$ discrétisée à des instants $\{t_k\}$ (même sans périodicité de l'échantillonnage, c'est-à-dire pour $t_{k+1} - t_k \neq$ constante) peut être représentée par un effet de retard variable :

$$u_d(t) = u(t_k) = u(t-(t-t_k)) = u(t-\tau(t)),\ \ t_k \leq t < t_{k+1},\ \ \tau(t) = t-t_k, \tag{6.6}$$

où u_d est le signal échantillonné et bloqué à partir du signal $u(t)$. Le retard variable $\tau(t) = t - t_k$ est alors continu par morceaux, variant linéairement de 0 à $(t_{k+1} - t_k)$ sur l'intervalle $[t_k, t_{k+1}[$ et, ainsi, de dérivée $\dot{\tau}(t) = 1$ pour $t \neq t_k$ et, en théorie, $\dot{\tau}(t_k) = -\infty$. Cette écriture permet par exemple de considérer de façon homogène une situation de commande en réseau avec échantillonnage, retards de transmission et pertes de paquets [129].

6.2 Classes d'équations differentielles fonctionnelles

De nombreuses classes de modèles ont été proposées pour l'étude des systèmes à retards. La référence [119] en donne un tableau résumé et [116], un aperçu plus détaillé. Dans cet ouvrage, nous présenterons principalement la notation fonctionnelle, très générale, mais rappellerons ensuite quelques autres représentations dédiées aux systèmes linéaires (section 6.7).

Les équations differentielles fonctionnelles peuvent être considérées comme une combinaison d'équations différentielles ordinaires et d'équations fonctionnelles. Les valeurs de l'argument peuvent y être discrètes, continues ou mixtes : en correspondance, on définira les notions d'équations différentielles aux différences, d'équations intégrodifférentielles, ou mixtes[8].

Une EDF est dite *autonome* (ou *stationnaire*) si elle est invariante vis-à-vis de tout changement de variable $t \mapsto t + T$ (pour tout $T \in \mathbb{R}$). L'*ordre* d'une EDF est celui de la plus haute dérivée de la fonction inconnue régie par l'équation. Ainsi, les équations fonctionnelles peuvent être considérées comme des EDF d'ordre zéro et la notion d'EDF généralise les équations de l'analyse mathématique des fonctions d'un argument continu.

Equations à retards ponctuels

Considérons une EDF *à retards ponctuels*, de la forme :

$$z^{(m)}(t) = f_0\left(t, z^{(m_1)}\left(t - h_1\left(t\right)\right), ..., z^{(m_k)}\left(t - h_k\left(t\right)\right)\right), \tag{6.7}$$

dans laquelle $z\left(t\right) \in \mathbb{R}^q$, $z^{(m)}\left(t\right) = \frac{d^m}{dt^m} z\left(t\right)$, k et $m_i \in \mathbb{N}$, $h_i\left(t\right) \in \mathbb{R}^+$. Le membre de droite f_0 et les retards h_i sont donnés et z est une fonction inconnue

[8] Plus particulièrement, dans le cas des équations différentielles retardées (EDR) qui sera développé par la suite, on parlera de retards ponctuels, distribués ou mixtes.

de t. La propriété $h_i(t) \in \mathbb{R}^+$ (signifiant que toutes les déviations d'argument sont positives ou nulles) est cruciale pour la causalité de (6.7). Cette équation (6.7) est dite :

- équation différentielle fonctionnelle *de type retardé*, ou EDF retardée (en abrégé, *EDR*) , si :

$$m > \max\{m_1, ..., m_k\}; \tag{6.8}$$

- équation différentielle fonctionnelle *de type neutre*, ou EDF neutre (en abrégé, *EDN*), si :

$$m = \max\{m_1, ..., m_k\}; \tag{6.9}$$

- équation différentielle fonctionnelle *de type avancé*, ou EDF avancée (en abrégé, *EDA*), si :

$$m < \max\{m_1, ..., m_k\}. \tag{6.10}$$

Une EDR est ainsi caractérisée par le fait que la valeur de la dérivée d'ordre le plus élevé est définie, pour chaque valeur de l'argument t, par les valeurs des dérivées d'ordre plus faible prises en des arguments inférieurs ou égaux à t.

La pratique de la modélisation montre qu'à la quasi-unanimité, seules les équations de type retardé (6.8) ou neutre (6.9) sont utilisées pour représenter des processus réels. Comme dans le cas des équations différentielles ordinaires, l'équation (6.7) peut être réécrite sous la forme d'une équation différentielle du premier ordre (impliquant la dérivée $\dot{x} = \frac{dx}{dt}$) portant sur un vecteur $x \in \mathbb{R}^n$ de dimension plus grande ($n = (m-1)\,q$) en prenant comme nouvelles inconnues les dérivées successives de y. On aboutit ainsi aux EDR et EDN suivantes :

$$\dot{x}(t) = f\left(t, x\left(t - h_1\left(t\right)\right), ..., x\left(t - h_k\left(t\right)\right)\right), \tag{6.11}$$

$$\dot{x}(t) = f\left(t, x\left(t - h_1\left(t\right)\right), ..., x\left(t - h_k\left(t\right)\right), \dot{x}\left(t - g_k\left(t\right)\right), ..., \dot{x}\left(t - g_l\left(t\right)\right)\right). \tag{6.12}$$

Comme nous l'avons remarqué, toute EDF est une combinaison d'équations ordinaires et fonctionnelles et l'équation de type neutre (6.12) est équivalente au système *2-D*, ou *hybride*, suivant :

$$\begin{cases} \dot{x}(t) = y(t), \\ y(t) = f\left(t, x\left(t - h_1\left(t\right)\right), ..., x\left(t - h_k\left(t\right)\right),\ y\left(t - g_k\left(t\right)\right), ..., y\left(t - g_l\left(t\right)\right)\right). \end{cases}$$

Dans certains phénomènes, le retard peut dépendre d'une solution inconnue, c'est-à-dire avoir la forme $h_i\left(t, x\left(t\right)\right)$. De tels retards sont quelquefois dits *autoréglants*. Leur analyse est assez difficile [139]. Le retard peut aussi dépendre de l'entrée de commande et, dans ce cas, les techniques de contrôle sont rares [22, 109, 110].

Equations retardées générales, retards distribués

Une équation différentielle retardée générale (à retards non nécessairement ponctuels) se représente sous la forme :

$$\dot{x}(t) = f\left(t, x_t\right), \tag{6.13}$$

où, pour un certain t, $x\left(t\right) \in \mathbb{R}^n$ et l'*état* x_t est une fonction définie par :

$$\begin{cases} x_t : J_t \rightarrow \mathbb{R}^n, \quad x_t\left(\theta\right) \triangleq x\left(t+\theta\right), \\ J_t \subset \left]-\infty, 0\right], \quad \theta \in J_t. \end{cases}$$

Dans ce cas, J_t peut être un intervalle donné $[-h(t), -g(t)]$ ou $]-\infty, -g(t)]$. La fonction x_t peut être interprétée comme un fragment de la solution x à droite du point t, observé depuis ce point. Le membre de droite de (6.13) est une fonction de t et x_t : ainsi, à toute fonction $\psi : J_t \rightarrow \mathbb{R}^n$ d'une certaine famille de fonctions, correspond un vecteur $f\left(t, \psi\right) \in \mathbb{R}^n$.

Remarquons que l'équation à retards ponctuels (6.11) est un cas particulier de (6.13)[9]. Si un des intervalles J_t n'est pas de mesure nulle, l'équation différentielle fonctionnelle (6.13) est dite *retardée, à retards distribués.*

On peut définir de la même façon une équation différentielle fonctionnelle *neutre, à retards distribués* comme suit :

$$\dot{x}(t) = f\left(t, x_t, \dot{x}_t\right), \tag{6.14}$$

où la notation $\dot{x}_t$ correspond de façon similaire à $\dot{x}_t : J_t \rightarrow \mathbb{R}^n$, avec J_t de mesure non nulle et $\dot{x}_t\left(\theta\right) \triangleq \dot{x}\left(t+\theta\right)$.

Notations complémentaires

On utilisera par la suite les notations suivantes :
$J_t = [-h(t), -g(t)] \subset \left]-\infty, 0\right]$, $J = [\alpha, \beta] \subset \mathbb{R}$;
$\mathcal{C}\left(J_t\right)$ ensemble des fonctions continues de $J_t \rightarrow \mathbb{R}^n$;
$\mathcal{C}^1\left(J_t\right))$ ensemble des fonctions dérivables de $J_t \rightarrow \mathbb{R}^n$;
$x_t \in \mathcal{C}\left(J_t\right) : J_t \rightarrow \mathbb{R}^n, \quad \theta \mapsto x_t\left(\theta\right) \triangleq x\left(t+\theta\right)$;
$\mathcal{D} = [t_0, +\infty[\times \mathcal{C}\left[-h, 0\right], \left(t, \psi\right) \in \mathcal{D}$;
$\|.\|$ norme scalaire de vecteur : $\mathbb{R}^n \rightarrow \mathbb{R}^+$, $x \mapsto \|x\|$;
$\|.\|_{\mathcal{C}}$ norme de fonction, $\mathcal{C}\left[-h, 0\right] \rightarrow \mathbb{R}^+$, $\psi \mapsto \|\psi\|_{\mathcal{C}} \triangleq \sup_{\theta \in [-h,0]} \{\|\psi\left(\theta\right)\|\}$;
$\mathcal{B}_\delta \subset \mathcal{C}\left[-h, 0\right]$ boule fonctionnelle, $\mathcal{B}_\delta \triangleq \{\psi \in \mathcal{C}\left[-h, 0\right] ; \|\psi\|_{\mathcal{C}} < \delta\}$;
$\mathbb{R}[\nabla]$ l'anneau (commutatif) des polynômes en ∇ à coefficients réels ;
$\mathbb{R}(\nabla)$ le corps des fractions rationelles en ∇ à coefficients réels ;
$\mathbb{R}^n\left[\nabla\right]$ le module[10] de dimension n sur $\mathbb{R}[\nabla]$;

[9] Dans (6.11), l'ensemble J_t est de mesure nulle [117], réduit à un nombre fini de points.
[10] Equivalent d'un espace vectoriel mais sur un anneau, voir [117].

$\lambda_{\max}(Q)$ la plus grande valeur propre d'une matrice symétrique $Q \in \mathbb{R}^{n\times n}$; pour toute matrice réelle $M(t) = [m_{ij}(t)]$, on définit $|M(t)| = [\,|m_{ij}(t)|\,]$ et $M^+(t) = \left[m_{ij}^+(t)\right]$, $m_{ij}^+(t) = \begin{cases} |m_{ij}(t)| \text{ si } i \neq j, \\ m_{ii}(t) \text{ si } i = j. \end{cases}$

6.3 Le problème de Cauchy pour les EDR

Le problème de Cauchy consiste à montrer l'existence (et, si possible, l'unicité) de la solution de l'équation (6.13) correspondant à une certaine fonction initiale et à une certaine valeur initiale. Considérons l'équation différentielle retardée (6.13) et supposons que pour un certain $t_0 \in \mathbb{R}$, la fonction $f : (t,x) \mapsto f(t,x)$ est définie pour tout $t \in [t_0, +\infty[$ et $x \in \mathcal{C}(J_t)$, $J_t = [-h(t), -g(t)]$. Le point t_0 est appelé *point initial*[11] pour la solution. Nous supposerons également que $\overline{t_0} \triangleq \inf_{t\geq t_0}\{t - h(t)\} > -\infty$.

La *fonction initiale* ψ de l'équation (6.13) pour un point initial t_0 est prescrite sur l'*intervalle initial* $\left[\overline{t_0}, t_0\right[$. Si $\overline{t_0} = t_0$, alors cet intervalle initial est vide et on retrouve le problème de Cauchy classique pour les EDO (sans hérédité, sans fonction initiale). Cependant, dans tous les cas, la *valeur initiale* $x(t_0)$ de la solution doit être prescrite. Généralement (bien que cela ne soit pas nécessaire) la valeur initiale de la solution $x(t_0)$ fait partie de la fonction initiale, c'est-à-dire que cette dernière est prescrite sur l'intervalle fermé $\left[\overline{t_0}, t_0\right]$ avec $\psi(t_0) = x(t_0)$.

Soulignons que la solution $x(t)$ doit être construite dans le sens des t croissants, c'est-à-dire sur un intervalle J ayant comme extrémité gauche le point $t_0 \in J$. Ceci implique que x est à interpréter comme étant le prolongement de la fonction initiale, $x(t+\theta) \triangleq \psi(t+\theta)$ pour $t + \theta > t_0$.

Nous considérerons ici le problème de Cauchy pour des EDR à retard fini, et supposons que la solution appartient à $\mathcal{C}^1$ (c'est-à-dire, est une fonction continûment différentiable de t). Le problème étudié est donc :

$$\dot{x}(t) = f(t, x_t), \quad x_t(\theta) = x(t+\theta) \ \forall \theta \in [-h, 0], \tag{6.15}$$

$$x_{t_0} = \psi. \tag{6.16}$$

Ici, $h \geq 0$ est une constante (finie), $x(t) \in \mathbb{R}^n, t_0 \in \mathbb{R}$, et $\psi : [-h, 0] \to \mathbb{R}^n$. La solution $t \mapsto x(t)$ $(t \geq 0)$ du problème (6.15) (6.16) est le prolongement de la fonction intiale $t \mapsto x(t)$ $(t_0 - h \leq t \leq t_0)$.

Définition 6.3.1. Soit un intervalle J ayant t_0 comme borne gauche (incluse). Une fonction $x \in \mathcal{C}^1(J)$ est une *solution du problème de Cauchy* (6.15) (6.16) sur cet intervalle J si elle vérifie l'équation (6.15) avec les conditions initiales $x(t_0) = \psi(t_0)$ et (6.16) en tous les points de J (c'est-à-dire, $x_t(t+\theta) = \psi(t+\theta-t_0)$ $\forall t - \theta < t_0$).

[11] Ou *instant initial* si t représente le temps.

Théorème 6.3.1. *Soient $\psi \in \mathcal{C}[-h, 0]$ et une fonction vectorielle $f : \mathcal{D} \to \mathbb{R}^n$, continue et vérifiant dans le voisinage de tout couple $(t, \psi) \in \mathcal{D}$ une condition de Lipschitz par rapport à son deuxième argument ψ (la constante de Lispchitz correspondante dépendant, en général, de ce couple). Alors il existe un point t_ψ, $t_0 < t_\psi \leq +\infty$ dépendant de ψ, t_0, f, tel que :*

(*a*) *il existe une solution x du problème (6.15)(6.16) sur $J = [t_0, t_\psi[$;*

(*b*) *sur tout intervalle $[t_0, t_1] \subset [t_0, t_\psi[$, cette solution est unique ;*

(*c*) *si $t_\psi < +\infty$ alors $x(t)$ n'a pas de limite finie quand $t \to t_\psi$;*

(*d*) *la solution x dépend continûment de f et ψ.*

La dernière proposition (d) signifie que : $\forall t_1 \in [t_0, t_\psi]$ et $\forall \varepsilon > 0$, $\exists \delta > 0$ tel que si, dans (6.15) (6.16), f et ψ sont remplacées par $\overline{f}$ et $\overline{\psi}$ vérifiant les mêmes propriétés et avec :

$$\left\|\psi - \overline{\psi}\right\|_{\mathcal{C}} < \delta \text{ et } \left\|f(t, \psi) - \overline{f}(t, \psi)\right\| < \delta \text{ pour } t \in [t_0, t_1], \tag{6.17}$$

alors la solution $\overline{x}$ du problème transformé vérifie :

$$\|x(t) - \overline{x}(t)\| < \varepsilon \text{ pour } t \in [t_0, t_1]. \tag{6.18}$$

Par ailleurs, en prenant $t - t_0$ comme nouvelle variable indépendante, il est possible d'étudier de la même façon la dépendance de la solution envers le point initial t_0 de la même façon que sa dépendance envers f.
Démonstration : [points (a) et (b)] en intégrant les deux membres de l'équation (6.15), on constate que le problème (6.15)(6.16) est équivalent à l'existence d'une solution continue pour l'équation intégro-différentielle :

$$x(t) = \psi(0) + \int_{t_0}^{t} f(\tau, x_\tau)\, d\tau, \qquad t \in J_x. \tag{6.19}$$

Considérons le membre de droite de cette équation en tant qu'opérateur dans l'espace métrique $\{x \in \mathcal{C}([t_0, t_0 + \alpha]),\ [\psi(0) - \beta, \psi(0) + \beta]),\ x(t_0) = \psi(0)\}$, où $\alpha, \beta > 0$ sont des constantes suffisamment petites. Le theorème de Banach (principe d'application contractante [117]) garantit l'existence et l'unicité de la solution pour tout intervalle $[t_0, t_0 + \alpha]$ suffisamment petit. On en déduit l'unicité de la solution pour tout intervalle d'existence : en effet, s'il existe deux solutions x^1 et x^2, alors en décalant le point initial de t_0 à $\inf\{t > t_0 : x^1(t) \neq x^2(t)\}$, nous obtenons une contradiction. En effectuant l'union de tous les intervalles $[t_0, t_1], \{t_0 < t_1 < +\infty\}$ sur lesquels la solution existe, nous obtenons l'intervalle maximal d'existence. Cet intervalle, de type $[t_0, t_\psi[$, est ouvert à droite $(t_0 < t_\psi \leq +\infty)$. Ainsi, (a) et (b) sont démontrées.
[point (c)] Supposons que $t_\psi < +\infty$ et qu'il existe $x(t_\psi) = \lim_{t \to t_\psi^-}[x(t)]$. Alors, la fonction x complétée par $x(t_\psi)$ est continue sur $[t_0, t_\psi]$. Puisque f

est continue, l'équation (6.19) est également valable pour $t = t_\psi$, et la solution existe donc sur $[t_0, t_\psi]$. Or ceci contredit la définition de t_ψ. La limite finie $x(t_\psi)$ ne peut donc exister, ce qui prouve (c).
[point (d)] Supposons que $t \in]t_0, t_\psi[$ et $\varepsilon > 0$ sont fixés, avec ε assez petit pour que la fonction $(t, \psi) \mapsto f(t, \psi)$ soit bornée et lipschitzienne en ψ dans la bande $t_0 < t < t_1,\ \|\psi - x_t\|_{\mathcal{C}} < \varepsilon$ (un tel ε existe d'après les hypothèses sur f). Nous constatons alors que si l'équation (6.17) est vérifiée pour un $\delta > 0$ suffisamment faible alors l'égalité $\|x(t) - \overline{x}(t)\| = \varepsilon$ est impossible pour $t \in [t_0, t_1]$. Il s'en suit que $\overline{x}(t)$ est borné et, d'après (c), que l'équation (6.18) est vérifiée.
Inversement, supposons qu'il n'existe pas $\delta > 0$ vérifiant (6.17). Alors, il doit exister des suites $\delta_k \to 0$ $(\delta_k \in (0, \varepsilon))$, $\overline{f}_k$ et $\overline{\psi}_k$ vérifiant pour chaque k la propriété correspondante (6.17) et telles que pour des points $t_k \in]t_0, t_1]$ les solutions $\overline{x}_k$ du problème correspondant (6.15) (6.16) satisfassent :

$$\|x(t) - \overline{x}_k(t)\| < \varepsilon \ \ (t_0 \le t < t_k),\ \|x(t_k) - \overline{x}_k(t_k)\| = \varepsilon. \tag{6.20}$$

L'ensemble des fonctions $\overline{x}_k : [t_0, t_k] \to \mathbb{R}^n$ étant uniformément borné et équicontinu, le lemme d'Ascoli-Arzela [117] permet le passage à des sous-suites (uniformément convergentes sur chaque intervalle $\left[t_0, \overline{t}\right] \subset \left[t_0, \overline{t}\right[$), $t_k \to \overline{t} > t_0$, $\overline{x}_k \to \overline{x}$ pour $k \to +\infty$. La fonction $\overline{x}$ est uniformément continue sur $\left[t_0, \overline{t}\right[$ et peut donc être prolongée sur $\left[t_0, \overline{t}\right]$ avec $\left\|x\left(\overline{t}\right) - \overline{x}\left(\overline{t}\right)\right\| = \lim \|x(t_k) - \overline{x}_k(t_k)\| = \varepsilon$. La fonction $\overline{x}_k$ est la solution du problème :

$$\overline{x}_k(t) = \overline{\psi}_k(0) + \int_{t_0}^{t} \overline{f}_k(\tau, \overline{x}_{k_\tau})\, d\tau, \qquad t \in [t_0, t_k],$$
$$\overline{x}_k(t) = \overline{\psi}_k(t - t_0), \qquad t \in [t_0 - h, t_0].$$

Effectuons le passage à la limite $k \to +\infty$ sur tout intervalle ci-dessus $\left[t_0, \widetilde{t}\right]$ en utilisant la majoration suivante :

$$\left\| \int_{t_0}^{t} \overline{f}_k(\tau, \overline{x}_{k_\tau})\, d\tau - \int_{t_0}^{t} f(\tau, \overline{x}_\tau)\, d\tau \right\|$$
$$\le \int_{t_0}^{\widetilde{t}} \left\|\overline{f}_k(\tau, \overline{x}_{k_\tau}) - f(\tau, \overline{x}_\tau)\right\| d\tau - \int_{t_0}^{\widetilde{t}} \|f(\tau, \overline{x}_{k_\tau}) - f(\tau, \overline{x}_\tau)\|\, d\tau.$$

Les deux termes du membre de droite tendent vers 0 pour $k \to +\infty$: le premier par la convergence uniforme de $\overline{f}_k$ vers f dans la bande $t_0 < t < t_1, \|\psi - x_t\| < \varepsilon$; le second à cause des propriétés de Lipschitz de f en ψ (dans la même bande). Donc, $\overline{x}$ vérifie (6.19) sur $\left[t_0, \widetilde{t}\right]$ avec la condition initiale (6.16). D'après (c), $\overline{x}(t) = x(t), \forall t \in \left[t_0, \widetilde{t}\right]$. Ceci étant vrai pour tout $\widetilde{t} \in \left]t_0, \overline{t}\right[$, il vient $\overline{x}\left(\overline{t}\right) = x\left(\overline{t}\right)$, ce qui est impossible. Il doit donc exister $\delta > 0$ vérifiant (6.17), ce qui termine la preuve du point (d).

6.4 Méthode pas à pas

Il est rare de pouvoir obtenir l'expression analytique de la solution d'une équation différentielle fonctionnelle générale. Cependant, dans le cas d'équations retardées à retards ponctuels, il est quelquefois possible d'utiliser la méthode dite *pas à pas*. Nous la présenterons ici dans le cas scalaire :

$$\begin{aligned} \dot{x}(t) &= f\left(t, x\left(t\right), x\left(t-h\right)\right), \\ t &\geqslant t_0, \qquad h > 0 \text{ constant.} \end{aligned} \tag{6.21}$$

La fonction $f : [t_0 - h, +\infty[\times \mathbb{R}^2 \rightarrow \mathbb{R}$ est continue, lipschitzienne en son second argument. La fonction initiale $\psi\left(t\right)$ de l'équation (6.21) est continue, donnée sur l'intervalle $[t_0 - h, t_0]$.

Le « premier pas » correspond à l'intervalle $t \in [t_0, t_0 + h]$. Pour ces valeurs de t, l'équation (6.21) devient une équation différentielle ordinaire :

$$\dot{x}(t) = f\left(t, x\left(t\right), \psi\left(t-h\right)\right), \qquad t \in [t_0, t_0 + h],$$

qui peut généralement être résolue pour la condition initiale $x\left(t_0\right) = \psi\left(t_0\right)$, puisque nous sommes dans le cas scalaire. Le résultat donne la solution sur $[t_0, t_0 + h]$, qui à son tour conduit au « deuxième pas » de résolution pour $t \in [t_0 + h, t_0 + 2h]$, dans lequel la fonction $x\left(t-h\right)$ est connue, issue du pas précédent. Cette EDO est à son tour résolue pour la condition initiale $x\left(t_0 + h\right)$, et ainsi de suite. Considérons par exemple le système suivant, où α est une constante :

$$\begin{aligned} \dot{x}(t) &= \alpha x\left(t-h\right), \\ \psi\left(t\right) &\equiv \psi_0 \text{ (constante) } \forall t \in [t_0 - h, t_0], \end{aligned} \tag{6.22}$$

pour lequel la méthode pas à pas nous donne la solution, pour $t \in [t_0, +\infty[$:

$$x\left(t\right) = \psi_0 \sum_{k=0}^{+\infty} \frac{\alpha^k}{k!} \left[t - t_0 - (k-1)\, h\right]^k \omega\left(t - t_0 - (k-1)\, h\right), \tag{6.23}$$

avec la notation $\omega\left(\theta\right) \triangleq \left(1 + \frac{\text{signe}(\theta)}{2}\right)$.

La régularité de la solution croît donc avec le temps. Cette propriété de « lissage » est par ailleurs une caractéristique générale des équations différentielles de type *retardé* (voir [65] page 37 et [100] page 24).

La figure 6.1 en donne une illustration pour $\alpha = 1$ et $h = 1$. Pour une condition initiale constante sur $[-h, 0]$, la solution sur le premier intervalle $[0, h]$ est une droite (intégrale d'une constante). Sur $[h, 2h]$ c'est une parabole, puis une cubique sur $[2h, 3h]$, puis un polynôme en $\frac{1}{k}t^k$ sur $[kh, (k+1)h]$... La solution $x(t)$ est ainsi non dérivable en $t = 0$, dérivable une fois sur $]0, +\infty[$, deux fois sur $]h, +\infty[$ et k fois sur $]kh, +\infty[$. Pour une condition initiale polynomiale d'ordre 1, la technique pas à pas donne la solution représentée en pointillé sur cette même figure.

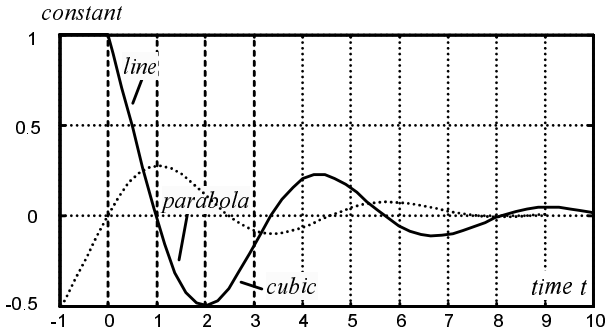


Fig. 6.1: $\dot{x}(t) = -x(t-1)$, fonctions initiales $\varphi \equiv 1$ (plein) et $\varphi = 0.5t$ (pointillé).

6.5 Stabilité des systèmes retardés

Le problème de la stabilité des EDR se pose très concrètement lors de la synthèse des asservissements, puisque la présence d'un retard dans un système bouclé conduit le plus souvent à des oscillations, voire des instabilités. Si la nature mathématique du phénomène de retard n'est pas prise en compte, la seule alternative de synthèse est de prévoir des marges de robustesse excessives, réduisant définitivement les performances dynamiques. Il est donc important de disposer d'outils spécifiques. Diverses méthodes sont disponibles (voir des bilans dans [19, 25, 69, 99]) : approches fréquentielles [9, 39, 135], méthodes de type Liapounov [14, 65, 69, 72] ou Popov [46], théorèmes de comparaison [78, 121], approches métriques [24, 26, 48], théorie des opérateurs [21], faisceaux matriciels [99], systèmes stochastiques [66, 72], etc. Nous ne présenterons ici que quelques uns de ces résultats : ils servent en général de base commune à toutes les méthodes d'investigation de la stabilité.

Notion d'équilibre pour une EDR

Considérons à nouveau le système (6.15)-(6.16), soit :

$$\begin{aligned} \dot{x}(t) &= f\left(t, x_t\right), \\ x_{t_0} &= \psi, \quad \psi \in \mathcal{C}\left[-h, 0\right]. \end{aligned} \tag{6.24}$$

Nous supposerons que $f(t, \varphi)$ est continue, bornée pour φ bornée, localement lipschitzienne en φ. La solution de (6.24) est notée $x\left(t, t_0, \psi\right)$.

Définition 6.5.1. La fonction $\varphi_e \in \mathcal{C}\left[-h, 0\right]$ est un *état d'équilibre* de (6.24) si pour tout $t_0 \in \mathbb{R}$, la solution $x\left(t, t_0, \varphi_e\right)$ existe et vérifie $x\left(t, t_0, \varphi_e\right) = \varphi_e$.

Théorème 6.5.1. [19] *La fonction $\varphi_e \in \mathcal{C}\left[-h, 0\right]$ est un état d'équilibre de (6.24) si, et seulement si, les trois conditions suivantes sont vérifiées :*

(*i*) $\forall t_0 \in \mathbb{R},\ x(t, t_0, \varphi_e)$ *existe et est unique ;*

(*ii*) $\forall t \in \mathbb{R},\ f(t, \varphi_e) = 0$ *;*

(*iii*) φ_e *est une fonction constante de* $\mathcal{C}[-h, 0] : \forall \theta \in [-h, 0], \varphi_e(\theta) = x_e$.

On parlera donc indifféremment d'état d'équilibre (φ_e) ou de *point d'équilibre* (x_e).

Définitions relatives à la stabilité des EDR

Nous faisons ici l'hypothèse que le système (6.24) possède un équilibre, placé à l'origine sans réduction de généralité, et donc que $f(t, 0) \equiv 0$.

Définition 6.5.2. L'équilibre $x = 0$ du système (6.24) est dit :

1. *stable* si $\forall \varepsilon > 0$, $\forall t_0$, $\exists \delta = \delta(t_0, \varepsilon) > 0$, $\psi \in \mathcal{B}_\delta \Rightarrow x(t, t_0, \psi) \in \mathcal{B}_\varepsilon$;
2. *uniformément stable par rapport à* t_0 si la propriété précédente est vérifiée avec $\delta = \delta(\varepsilon)$ (donc δ indépendant de t_0) ;
3. *asymptotiquement stable* s'il est stable et s'il existe $\eta = \eta(t_0) > 0$ tel que $[\psi \in \mathcal{B}_\eta] \Rightarrow [\lim_{t\to\infty} x(t, t_0, \psi) = 0]$;
4. *uniformément asymptotiquement stable* s'il est uniformément stable et si la limite de la propriété précédente est uniforme, c'est-à-dire si $\exists \eta > 0$: $\forall \gamma > 0, \exists T(\gamma) > 0 : [\psi \in \mathcal{B}_\eta$ et $t \geq T(\gamma)] \Rightarrow [x(t, t_0, \psi) \in \mathcal{B}_\gamma]\ \forall t_0$;
5. *globalement (uniformément) asymptotiquement stable* s'il est (uniformément) asymptotiquement stable avec $\eta = +\infty$;
6. *globalement exponentiellement stable* s'il existe deux nombres strictement positifs α (appelé *taux de convergence exponentielle*) et k tels que :

$$|x(t, t_0, \psi)| \leq k\|\psi\|_{\mathcal{C}}\ e^{-\alpha(t-t_0)}. \tag{6.25}$$

Stabilité des systèmes retardés linéaires stationnaires

Dans cette section, on parlera indifféremment de la stabilité asymptotique de l'équilibre ou du système : en effet dans le cas linéaire stationnaire, un point d'équilibre, s'il est asymptotiquement stable, est forcément unique (la propriété est globale et uniforme).

La stabilité d'un système linéaire stationnaire retardé est déterminée par la position des racines de son équation caractéristique par rapport à l'axe imaginaire.

Théorème 6.5.2. *Un système linéaire stationnaire de type retardé est globalement asymptotiquement stable si, et seulement si, toutes ses racines caractériqtiques sont dans le demi-plan complexe gauche (l'axe imaginaire étant exclu).*

On retrouve ici la même propriété que dans le cas des équations ordinaires : par contre, l'équation caractéristique étant ici un quasi-polynôme (fonction polynomiale en s et en e^{-s}, voir [117]), des tests simples de cette propriété (comme le critère de Routh-Hurwitz) ne sont plus disponibles.

Considérons l'équation générale[12] suivante à retard quelconque (fini ou infini, ponctuel ou distribué) :

$$\begin{aligned} \dot{x}(t) &= \int_{-\infty}^{0} [dK(\theta)]\, x(t+\theta), \qquad x(t) \in \mathbb{R}^n, t \geq 0, \\ x(\theta) &= \psi(\theta) \qquad \forall \theta \in]-\infty, 0], \end{aligned} \tag{6.26}$$

où l'intégrale est définie au sens de Stieljes [117], avec $K(\theta)$ une matrice $n \times n$ dont les coefficients k_{ij} sont des fonctions de $\theta \in]-\infty, 0]$ et à variations bornées. La transformation de Laplace appliquée à (6.26) conduit à :

$$\begin{gathered} \left[sI - \overline{K}(s)\right] \overline{x}(s) = \psi(0) + \overline{F}(s), \quad s \in \mathbb{C}, \\ \text{avec } F(t) = \int_{-\infty}^{-t} [dK(\theta)]\, \psi(t+\theta), \quad \overline{F}(s) = \int_{-\infty}^{0} e^{-s\theta} F(\theta)\, d\theta, \\ \overline{K}(s) = \int_{-\infty}^{0} e^{s\theta} dK(\theta), \quad \overline{x}(s) = \int_{-\infty}^{0} e^{-s\theta} x(\theta)\, d\theta. \end{gathered}$$

L'*équation caractéristique* correspondante est :

$$\Delta(s) = \det\left[sI - \overline{K}(s)\right] = 0. \tag{6.27}$$

Elle a généralement un nombre infini de solutions dans le plan complexe : dans le cas contraire (nombre fini de racines) on parle d'EDR *dégénérée*.

Théorème 6.5.3. *Le système (6.26) est asymptotiquement stable si les racines de (6.27) sont dans le demi-plan gauche strict (*Réel$(s) < 0$*) et si toutes les fonctions $k_{ij\ (i,j=1,\ldots,n)}$ vérifient :*

$$\int_{-\infty}^{0} |\theta|\ |dk_{ij}(\theta)| < +\infty.$$

De nombreuses méthodes ont été élaborées pour localiser les racines de (6.27) (voir par exemple [19, 69] ainsi que le Chapitre 10 de [117]). Le problème n'est pas simple dès l'instant où l'ordre n grandit, ou bien lorsque quelques paramètres de réglage (notamment le retard) sont conservés formellement.

[12] Le théorème de Riesz [117] assure l'existence de la fonction (dite canonique) K impliquée dans (6.26) pour toute fonctionnelle linéaire continue $g(x_t)$.

Exemple 6.5.1. Considérons l'équation $\dot{x}(t) = -x(t-1)$. Son équation caractéristique est $s + e^{-s} = 0$, dont les solutions $s = \alpha \pm j\beta$ sont en nombre infini. Le système n'est donc pas dégénéré. Ici, $s = -0.318 \pm 1.337j$ est une estimation de la paire de racines de plus grande partie réelle : il y a donc stabilité asymptotique[13]. Par contre, le cas suivant est dégénéré et instable :

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \\ 0 & -\frac{1}{2} & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} x(t-h),$$
$$\Delta(s) = s\left(s^2 - 1\right).$$

Exemple 6.5.2. Considérons le système (6.26) dans sa forme scalaire :

$$\dot{x}(t) = \int_{-\infty}^{0} x(t+\theta)\, dk(\theta),$$

et supposons que le noyau $k(s)$ est une fonction non croissante ($dk(\theta) \leq 0$), constante sur l'intervalle $\theta \leq -h < 0$ (par conséquent, l'effet de retard à l'instant t est limité aux instants $[t-h, t]$) :

$$\dot{x}(t) = \int_{-h}^{0} x(t+\theta)\, dk(\theta). \tag{6.28}$$

Alors, (6.28) est asymptotiquement stable si :

$$\gamma_0 = \int_{-h}^{0} dk(\theta) < 0 \text{ et } \gamma_1 = \int_{-h}^{0} |dk(\theta)| < \frac{\pi}{2h}.$$

Démonstration : pour $s = \alpha + j\beta$, l'équation (6.27) s'écrit :

$$\operatorname{Im}\Delta(s) = \beta - \int_{-h}^{0} e^{\alpha\theta} \sin\beta\theta \; dk(\theta) = 0, \tag{6.29}$$

$$\operatorname{Re}\Delta(s) = \alpha - \int_{-h}^{0} e^{\alpha\theta} \cos\beta\theta \; dk(\theta) = 0. \tag{6.30}$$

Or, ces deux équations ne peuvent être simultanément vérifiées pour $\alpha \geq 0$: si $|\beta| < \frac{\pi}{2h}$, alors (6.30) n'a pas de racine puisque $\alpha - \int_{-h}^{0} e^{\alpha\theta} \cos\beta\theta \; dk(\theta) \geq$

[13] On a en fait $\alpha \in]-0.3181, -0.3182[$. Les racines sont situées à l'intersection des courbes $\beta = \pm e^{-\alpha}\sqrt{(1-\alpha^2 e^{2\alpha})}$ et $\beta = \arccos(-\alpha e^{\alpha})$.

$-\int_{-h}^{0} e^{\alpha\theta}\cos\beta\theta\ dk\,(\theta)\ \geq\ -\gamma_0 e^{-\alpha h}\cos\beta h > 0$; si $|\beta| > \frac{\pi}{2h}$, alors (6.29) n'a pas de racine puisque $\left|\int_{-h}^{0} e^{\alpha\theta}\sin\beta\theta\ dk\,(\theta)\right|\ \leq\ \left|\int_{-h}^{0} dk\,(\theta)\right| < \frac{\pi}{2h}$. ∎

Première méthode de Liapounov

L'*approximation au premier ordre* (ou *approximation des petits mouvements*, ou *système linéarisé tangent*), bien connue pour les EDO, est encore valable dans le cas des systèmes retardés. Considérons :

$$\begin{aligned} \dot{x}(t) &= \sum_{i=0}^{k} A_i x(t-h_i) + q(t,x_t) \qquad (6.31)\\ q(t,x_t) &= q(t,x(t),x(t-\tau_1(t)),...,x(t-\tau_k(t)),\\ h_0 &= 0,\ h_i = \text{constantes},\ \tau_j(t)\in[0,\tau_i] \text{ continues},\\ \|u_i\| \leq \varepsilon &\Rightarrow \|q(t,u_0,...,u_k)\| \leq \beta_\varepsilon(\|u_0\|+...+\|u_k\|), \end{aligned}$$

avec β_ε = constante pour ε donné, β_ε uniformément décroissante vers 0 quand $\varepsilon \to 0$. L'approximation au premier ordre est définie par :

$$\dot{z}(t) = \sum_{i=0}^{k} A_i z(t-h_i). \qquad (6.32)$$

Théorème 6.5.4. [26] *Si le système linéarisé (6.32) est asymptotiquement stable, alors $z=0$ l'est aussi pour (6.31). Si (6.32) a au moins une racine caractéristique à partie réelle positive, alors $z=0$ est instable pour (6.31).*

Cas des retards faibles

Le résultat précédent peut être utilement complété par une *approximation des petits retards*, résultat de nature qualitative obtenu par continuité des racines caractéristiques de (6.32) vis-à-vis des retards h_i.

Théorème 6.5.5. [26] *Si $A = \sum_{i=0}^{k} A_i$ est de Hurwitz (respectivement, instable), alors pour des valeurs suffisamment faibles des retards h_i, la solution nulle $z=0$ est asymptotiquement stable (respectivement, instable) pour (6.32) et donc (6.31). Si, sur les n valeurs propres de A, $n-1$ ont des parties réelles strictement négatives et la $n-$ième est nulle, alors, pour des valeurs suffisamment faibles des h_i, $z=0$ est stable pour (6.32) et donc pour (6.31).*

Dans le cas d'un retard unique, une estimation quantitative des « petits » retards conservant la stabilité est donnée par le théorème suivant. Nous considérons ici le système linéaire à retard constant :

$$\frac{dz(t)}{dt} = A_0 z(t) + A_1 z(t-h), \qquad (6.33)$$

qui, pour un retard nul, devient :

$$\frac{dz(t)}{dt} = (A_0 + A_1)\, z(t). \tag{6.34}$$

Théorème 6.5.6. [40] *Si le système à retard nul (6.34) est asymptotiquement stable et si P est la matrice solution de l'équation de Liapounov (où Q est une matrice réelle définie positive [117]) :*

$$(A_0 + A_1)^T P + P(A_0 + A_1) = -Q^T Q, \tag{6.35}$$

alors (6.33) est asymptotiquement stable pour tout retard $h \in [0, h_{\max}]$:

$$h_{\max} = \frac{1}{2}\left[\lambda_{\max}(B^T B)\right]^{-\frac{1}{2}}, \quad \textit{avec } B = Q^{-T} A_1^T P (A_0 + A_1) Q^{-1}. \tag{6.36}$$

Démonstration : le principe de la démonstration sera donné dans la section suivante (exemple 4). ∎

Méthode directe de Liapounov

La méthode directe de Liapounov est un outil majeur pour étudier la stabilité des équations à retards. Comme en dimension finie, elle conduit généralement à des conditions seulement suffisantes. Par contre, vu la difficulté du calcul des racines caractéristiques, son emploi se justifie très souvent, même dans le cas linéaire.

Définition 6.5.3. Une fonction scalaire $\omega : [0, +\infty[\to [0, +\infty[$ est dite *définie positive* si elle est continue et vérifie $\omega(r) > 0$ pour $r > 0$, et $\omega(0) = 0$. Une matrice carrée réelle Q est définie positive si $\omega(x) = x^T Q x$ l'est (donc si, et seulement si, ses valeurs propres λ_i vérifient$\lambda_i > 0$).

Définition 6.5.4. Soit V une fonctionnelle vérifiant les propriétés suivantes :

(a) $V : \mathbb{R} \times \mathcal{B}_h \to \mathbb{R}$ $(h > 0)$ est continue, avec $V(t, 0) = 0$ pour tout t.

(b) il existe des fonctions scalaires ω_1, ω_2 définies positives, non décroissantes, telles que :

$$\omega_1(\varphi(0)) \leq V(t, \varphi) \leq \omega_2(\|\varphi\|_{\mathcal{C}}) \quad \forall t. \tag{6.37}$$

La *dérivée totale de la fonctionnelle* $V(t, \varphi)$ le long des solutions de (6.24) est alors définie par :

$$\dot{V}(t, \varphi) \triangleq \lim_{\varepsilon \to 0^+} \sup \frac{1}{\varepsilon}\left[V(t + \varepsilon, x(t + \varepsilon, t, \varphi)) - V(t, \varphi)\right].$$

Théorème 6.5.7. *S'il existe une fonctionnelle $V(t, \varphi)$ vérifiant les propriétés (a) et (b) ci-dessus et, pour tout t_0 et tout $t \geq t_0$:*

$$\dot{V}(t, \varphi) \leq -\omega_3(\varphi(0)), \tag{6.38}$$

où ω_3 est définie positive, non décroissante, alors l'équilibre $x = 0$ de l'EDR (6.24) est uniformément asymptotiquement stable.

Démonstration : soient $\varepsilon > 0$ et $\delta \leq h$ choisi tel que $\omega_2(\delta) \leq \omega_1(\varepsilon)$. Alors, pour toute fonction initiale $\varphi \in \mathcal{B}_\delta$, on a $\omega_1(\|x(t,t_0,\varphi)\|) \leq V(t,x_t) \leq V(t,\varphi) \leq \omega_2(\|\varphi\|_{\mathcal{C}}) \leq \omega_1(\varepsilon)$. Ainsi, $\|x(t,t_0,\varphi)\| \leq \varepsilon \quad \forall t \geq t_0$, prouvant la stabilité uniforme. Montrons maintenant que $\lim_{t\to+\infty} x(t) = 0$ pour toute fonction initiale $\varphi \in \mathcal{B}_\eta$ où η vérifie $0 < \eta \leq h$ et $\omega_2(\eta) \leq \omega_1(h)$. Comme pour la preuve de stabilité, on déduit $\|x(t,t_0,\varphi)\| \leq h \quad \forall \varphi \in \mathcal{B}_\eta$. Donc $\left\|\dot{x}(t,t_0,\varphi)\right\| \leq c < \infty$. Supposons que pour une condition initiale $\varphi \in \mathcal{B}_\eta$ la solution $x(t,t_0,\varphi)$ ne tende pas vers 0 quand $t \to +\infty$. Alors, il doit exister $\varepsilon > 0$ et une suite $\{t_i\}$, $\lim_{i\to+\infty} t_i \to +\infty$ tels que $\|x(t_i,t_0,\varphi)\| \geq \varepsilon$. Or, $\left\|\dot{x}(t,t_0,\varphi)\right\| \leq c < \infty$ et donc $t_{i+1} - t_i \geq 2\Delta$, $\Delta = \frac{\varepsilon}{2c}$, et $\|x(t_i+\tau,t_0,\varphi)\| \geq \frac{\varepsilon}{2}$ pour tout τ tel que $|\tau| \leq \Delta$. Pour ces instants τ, (6.38) implique que pour un $\alpha > 0$, $\dot{V}(t_i+\tau, x_{t_i+\tau}) \leq -\alpha$. Notons $V(t) = V(t,x_t)$ et $N(t)$ le nombre de points t_i tels que $t_0 + \Delta \leq t_i \leq t - \Delta$. Alors, $V(t) - V(t_0) \leq \sum_{t_0+\Delta\leq t_i\leq t-\Delta} [V(t_i+\Delta) - V(t_i-\Delta)] \leq -2\Delta\alpha N(t)$. Comme $N(t) \to +\infty$ pour $t \to +\infty$, on en déduit que $V(t) \to -\infty$ pour $t \to +\infty$, ce qui est impossible car $V(t) \geq 0$. ∎

Exemple 6.5.3. Considérons l'équation scalaire

$$\dot{x}(t) = -ax(t) + \int_0^{+\infty} x(t-\theta)\, dk(\theta), \quad t \geq 0, \tag{6.39}$$

où a est une constante positive et $k(t)$ est une fonction à variation bornée sur $[0,+\infty[$. Considérons la fonctionnelle de Liapounov suivante :

$$V(t,x_t) = x^2(t) + \int_0^{+\infty} |dk(\theta)| \int_{t-\theta}^{t} x^2(\tau)\, d\tau. \tag{6.40}$$

La dérivée de (6.40) le long de (6.39) est :

$$\dot{V}(t,x_t) = 2x(t)\left[-ax(t) + \int_0^{+\infty} x(t-\theta)\, dk(\theta)\right] + x^2(t)\int_0^{+\infty} |dk(\theta)| - \int_0^{+\infty} x^2(t-\theta)\,|dk(\theta)|.$$

Remarquons que :

$$2\left|x(t)\int_0^{+\infty} x(t\text{-}\theta)\, dk(\theta)\right| \leq x^2(t)\int_0^{+\infty} |dk(\theta)| + \int_0^{+\infty} x^2(t\text{-}\theta)\,|dk(\theta)|.$$

L'équilibre $x = 0$ est donc asymptotiquement stable pour (6.39) si :

$$a > \int_0^{+\infty} |dk(\theta)|, \quad \int_0^{+\infty} \theta\,|dk(\theta)| < +\infty,$$

puisque sous ces deux conditions les hypothèses (6.37) (6.38) sont validées.

Exemple 6.5.4. La démonstration du théorème 6 utilise la fonctionnelle de Liapounov $V = V_1 + V_2$, $V_1 = \alpha y(t)^T P y(t)$, $y(t) = \left[z(t) + \int_{t-\tau}^{t} A_1 z(\theta) d\theta\right]$, $V_2 = \int_{t-\tau}^{t}\left[\int_\theta^t z^T(v) Q^T Q z(v) dv\right] d\theta$. En remarquant que $\dot{y}(t) = (A_0 + A_1)\, z(t)$, on vérifiera que la dérivée $\dot{V}$ est négative sous la condition (6.36).

Quelques fonctionnelles de Liapounov-Krasovskii

Les extensions fonctionnelles des fonctions de Liapounov quadratiques ont été très largement étudiées dans le cadre des systèmes linéaires à retards. Ainsi, depuis une vingtaine d'années, diverses méthodes de construction de fonctionnelles de Liapounov-Krasovskii pour des équations particulières ont été proposées (voir par exemple [25, 65, 69, 99] et les références incluses). Ces travaux ont été soutenus par les progrès numériques de l'optimisation convexe : l'outil LMI (acronyme anglais de *inégalité matricielle linéaire*) est aujourd'hui intégré dans tous les logiciels dédiés à l'automatique.

Ainsi, pour le système simple :

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h), \tag{6.41}$$

et la fonctionnelle :

$$V(x_t) = x(t)^T P x(t) + \int_{-h}^{0} x(t+\theta)^T S x(t+\theta) d\theta, \tag{6.42}$$

on obtient des conditions suffisantes sous forme d'équations de Riccati : (6.41) est asymptotiquement stable pour tout $h \geq 0$ s'il existe des matrices P, S, R positives et symétriques telles que :

$$A_0^T P + P A_0 + P A_1 S^{-1} A_1^T P + S + R = 0. \tag{6.43}$$

Cette équation (6.43) est équivalente à la LMI suivante :

$$\begin{pmatrix} A_0^T P + P A_0 + S & P A_1 \\ A_1^T P & -S \end{pmatrix} < 0. \tag{6.44}$$

Bien sûr, pour $A_1 = 0$, (6.43) se réduit à l'équation de Liapounov $A_0^T P + P A_0 < 0$, CNS classique dans le cas ordinaire. Pourtant, dans le cas retardé, la condition suffisante (6.43)-(6.44) est loin d'être nécessaire. C'est pourquoi de très nombreuses généralisations de la fonctionnelle (6.42) ont été publiées dans les quinze dernières années. Elles mettent en jeu les termes variés suivants :

$$\begin{aligned}
V_1(x(t)) &= x^T(t) P x(t), \\
V_2(x_t) &= x^T(t) \int_{-h_i}^{0} Q_i x(t+\theta) d\theta, \\
V_3(x_t) &= \int_{-h_i}^{0} x^T(t+\theta) S_i x(t+\theta) d\theta, \\
V_4(x_t) &= \int_{-\tau_i}^{0} \int_{t+\theta}^{t} x^T(\theta) R_i x(\theta) d\theta \, ds, \\
V_5(x_t) &= x(t)^T \int_{-h_i}^{0} P_i(\eta) x(t+\eta) d\eta, \\
V_6(x_t) &= \int_{-h_i}^{0} \int_{-h_i}^{0} x(t+\eta)^T P_i(\eta,\theta) x(t+\theta) \, d\eta \, d\theta.
\end{aligned} \tag{6.45}$$

Pour faire simple, V_2, V_3 visent à montrer une stabilité indépendante du retard dans le cas de retards ponctuels ; V_4, à la stabilité dépendante du retard discret ou aux retards distribués. Par exemple, pour le système (6.41), $V(x_t) = V_1(x(t)) + V_4(x_t) + V_4(x_{t-h})$ constitue une application particulière de [68] qui conduit à la condition suivante (dépendant du retard), où $A = A_0 + A_1, R_1$ pour $V_4(x_t)$ et R_2 pour $V_4(x_{t-h})$:

$$\begin{pmatrix} A^T P + PA + hR_1 + hR_2 & hPA_1A_0 & hPA_1^2 \\ hPA_0^T A_1^T & -hR_1 & 0 \\ hA_1^{2T} P & 0 & -hR_2 \end{pmatrix} < 0. \tag{6.46}$$

V_5 et V_6 apparaissent, sous forme générale, dans des combinaisons visant des conditions *nécessaires et suffisantes* : [53] en linéaire pour le cas de retards ponctuels, [51] pour le cas distribué, [85] pour des retards variables. Cependant, pour généraliser ces techniques à des conditions de stabilité robuste, on se heurte au prolème du calcul de V_5 et V_6. Pour éviter ces limitation calculatoires, des formes plus particulières de V_5, V_6 ont été introduites [42, 43], mettant en jeu des fonctions constantes par morceaux $P_i(.)$ et conduisant à des *fonctionnelles discrétisées*. On peut alors choisir un compromis entre la réduction du conservatisme et l'effort de calcul. Un bon résumé de ces techniques est donné dans [100]. [62] a également proposé une façon d'éviter le calcul générique des matrices $P_i(\eta)$ et $P_i(\eta, \theta)$, en passant par les propriétés de la matrice fondamentale.

D'autres façons de régler le choix des fonctionnelles V_i reposent la reformulation préalable du modèle. Elles seront présentées dans la partie 6.8.

Stabilité et équations de Riccati

Les théorèmes qui suivent sont une application du théorème 7 aux systèmes linéaires, permettant de formuler des conditions de stabilité en terme d'existence d'une solution positive définie à certaines équations de Riccati (voir le chapitre 9 de [117]) auxiliaires. Pour ne pas alourdir la présentation, nous traiterons ici les seuls systèmes à retards ponctuels :

$$\dot{x}(t) = \sum_{i=1}^{m} A_i x(t - h_i). \tag{6.47}$$

Un cas plus général, incluant les modèles à retards distribués, est traité dans [70] [73]. De même, les conditions peuvent plus généralement concerner la stabilité dépendante de certains retards et indépendante des autres [68]. Notons que les équations de Riccati obtenues conduisent, à leur tour, à des conditions de type LMIs (voir [117] chapitre 12).

Nous utiliserons les notations suivantes :

$$A = \sum_{i=1}^{m} A_i, \quad A_{ij} = A_i A_j, \quad h_{ij} = h_i + h_j, \quad h = \sum_{i=1}^{m} h_i. \tag{6.48}$$

Théorème 6.5.8. *Le système (6.47) est asymptotiquement stable si, pour deux matrices symétriques et définies positives R, Q, il existe une matrice définie positive P solution de l'équation de Riccati :*

$$A^T P + PA + mRh + P \sum_{i,j=1}^{m} h_i A_{ij} R^{-1} A_{ij}^T P = -Q. \tag{6.49}$$

Démonstration : on choisit la fonctionnelle $V = V_1 + V_2$, $V_1 = x^T(t)Px(t)$, $V_2 = \sum_{i,j=1}^{m} \int_{h_j}^{h_{ij}} ds \int_{t-s}^{t} x^T(\tau)Rx(\tau)d\tau$, conduisant à $\dot{V} = -x^T(t)Qx(t) - \sum_{i,j=1}^{m} \int_{t-h_j}^{t-h_{ij}} [Rx(\theta) + A_{ij}^T Px(t)]R^{-1}[Rx(\theta) + A_{ij}^T Px(t)]^T d\theta$. ∎

Théorème 6.5.9. *Le système (6.47) est asymptotiquement stable si l'équation $x(t) + \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} x(s)ds = 0$ l'est et si, pour des matrices symétriques et définies positives $R_{i\ [i\in\{1,\dots,m\}]}$, Q, il existe une matrice définie positive P solution de l'équation de Riccati :*

$$A^T P + PA + \sum_{i=1}^{m} R_i h_i + \sum_{i,j=1}^{m} A^T P A_i R_i^{-1} A_i^T P A h_i = -Q. \tag{6.50}$$

Démonstration : basée sur la fonctionnelle $V(t, x_t) = V_1 + V_2$, $V_1 = [x(t) + \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} x(s)ds]^T P[x(t) + \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} x(s)ds]$,
$V_2 = \sum_{i=1}^{m} \int_{0}^{h_i} ds \int_{t-s}^{t} x^T(\tau)R_i x(\tau)d\tau$. ∎

Théorème 6.5.10. *Le système (6.47) est asymptotiquement stable si, pour deux matrices symétriques et définies positives R, Q, il existe une matrice définie positive P solution de l'équation de Riccati :*

$$A^T P + PA + \sum_{i=1}^{m} (h_i P A_i R^{-1} B_i^T P + mh A_i^T R A_i) = -Q. \tag{6.51}$$

Démonstration : la preuve est basée sur la fonctionnelle $V(t, x_t) = V_1 + V_2 + V_3$,
$V_1 = x^T(t)Px(t)$, $V_2 = \sum_{i=1}^{m} \int_{0}^{h_i} ds \int_{t-s}^{t} \dot{x}^T(\tau)R\dot{x}(\tau)d\tau$,
$V_3 = mh \sum_{i=1}^{m} \int_{t-h_i}^{t} x^T(s)A_i^T R A_i x(s)ds$. ∎

Remarque 6.5.1. Nous laissons au lecteur le soin de construire les trois fonctionnelles $V(t, x_t)$ des démonstrations ci-dessus en utilisant les trois transformations de la partie 6.8.

Remarque 6.5.2. Si dans les trois théorèmes précédents, les retards h_i sont tous nuls, les trois équations de Riccati coïncident avec l'équation de Liapounov du système linéaire ordinaire $\dot{x} = Ax$ et les conditions suffisantes présentées sont également nécessaires. Ceci donne à penser que ces conditions sont peu conservatives pour des retards faibles.

Exemple 6.5.5. Considérons le système du second ordre avec un nombre quelconque m de retards $h_i \geq 0$ (et α, β deux constantes) :

$$\dot{x}(t) = B \sum_{i=1}^{m} x(t-h_i), \quad B = \begin{pmatrix} -\alpha & \beta \\ -\beta & -\alpha \end{pmatrix}.$$

Les trois théorèmes donnent la même condition (suffisante) : $0 \leq h < \frac{\alpha}{\alpha^2+\beta^2}$.

Exemple 6.5.6. Considérons le système (6.47) avec $m = 2$:

$$\dot{x}(t) = \begin{pmatrix} -\alpha_1 & \beta_1 \\ -\beta_1 & -\alpha_1 \end{pmatrix} x(t-h_1) + \begin{pmatrix} -\alpha_2 & \beta_2 \\ -\beta_2 & -\alpha_2 \end{pmatrix} x(t-h_2).$$

En posant $\mu_i^2 = \alpha_i^2 + \beta_i^2 \ _{(i=1,2)}$, les équations (6.49) et (6.51) conduisent à :

$$(h_1+h_2)\left(h_1\mu_1^2 + h_2\mu_2^2\right) < \frac{(\alpha_1+\alpha_2)^2}{2\left(\mu_1^2+\mu_2^2\right)},$$

alors que (6.50) donne une condition moins contraignante, obtenue en choisissant $R_1 = |\mu_1| I$, $R_2 = |\mu_2| I$:

$$(h_1+h_2)\left(h_1\mu_1^2 + h_2\mu_2^2\right) < \frac{(\alpha_1+\alpha_2)^2}{2\left(\mu_1^2+\mu_2^2\right)}.$$

Ce dernier exemple montre que les conditions issues des trois équations de Riccati (6.49), (6.50) et (6.51) ne sont pas équivalentes.

Principe de comparaison

Le principe général de cette approche est de comparer les solutions des équations d'origine avec celles d'un système auxiliaire (sensé être plus simple) appelé *système de comparaison*. Celui-ci est en général obtenu à partir d'inégalités différentielles [78] vérifiées par le système d'origine. Le principe de comparaison s'applique à une classe très large de systèmes[14], ordinaires comme fonctionnels, et dans cette partie nous l'illustrerons principalement dans le cas linéaire non stationnaire :

$$\dot{x}(t) = A(t)\,x(t) + B(t)\,x(t-h(t)), \quad t \geq t_0, \tag{6.52}$$

$$x(t_0+\theta) = \varphi(\theta), \quad \forall \theta \leq 0. \tag{6.53}$$

Les coefficients des matrices $A(t) = (a_{ij}(t))$ et $B(t) = (b_{ij}(t))$, ainsi que le retard $h(t) \geq 0$, sont supposés continus. Nous emploierons les notations $A^+(t)$ et $|B(t)|, |x(t)|$ définies page 241.

[14] Voir des articles de synthèse comme [11, 121] qui concernent des cas présentant des nonlinéarités, discontinuités, retards multiples, systèmes neutres, etc.

Considérons une fonction de comparaison $z(t) \in \mathbb{R}^n$ vérifiant l'inégalité différentielle :

$$\dot{z}(t) \geq A^+(t) z(t) + |B(t)| z(t - h(t)), \quad \forall t \geq t_0, \tag{6.54}$$

$$|z(t_0 + \theta)| \geq |\varphi(\theta)|, \quad \forall \theta \leq 0. \tag{6.55}$$

Théorème 6.5.11. *Pour toute fonction $z(t)$ satisfaisant (6.54) (6.55), on a :*

$$z(t) \geq |x(t)| \geq 0 \qquad \forall t \in \mathbb{R},$$

où $x(t)$ est la solution du système (6.52) (6.53).

Démonstration : montrons tout d'abord que si $\varphi \neq 0$, alors $z(t) \geq 0$, $\forall t \geq t_0$[15]. D'après (6.55), ceci est vrai pour $t = t_0$. Par contradiction, notons $\tau > t_0$ le premier point où une composante de z s'annule, $z_j(\tau) = 0$. En ce point, d'après (6.54), $\dot{z}_j(\tau) \geq 0$ et $z_j(t)$ ne peut donc devenir négative.

Soit maintenant $\varepsilon \in]0, 1]$, et $x^\varepsilon(t)$ la solution du problème de Cauchy :

$$\dot{x}^\varepsilon(t) = [A(t) - \varepsilon I] x^\varepsilon(t) + B(t) x^\varepsilon(t\text{-}h(t)), \forall t \geq t_0, \tag{6.56}$$

$$x^\varepsilon(t_0 + \theta) = (1 - \varepsilon)\varphi(\theta), \quad \forall \theta \leq 0. \tag{6.57}$$

Montrons que :

$$|x^\varepsilon(t)| < z(t), \quad \forall t \geq t_0. \tag{6.58}$$

D'après (6.55) et (6.57), (6.58) est vraie pour $t = t_0$. Par contradiction, notons $\tau > t_0$ le premier point où l'inégalité stricte (6.58) devient une égalité pour une de ses composantes, $\left|x_j^\varepsilon(\tau)\right| = z_j(\tau)$. Considérons tout d'abord le cas $x_j^\varepsilon(\tau) > 0$. D'après (6.56) (6.55),

$$\begin{aligned}
\dot{x}_j^\varepsilon(\tau) - \dot{z}_j(\tau) &\leq -\varepsilon x_j^\varepsilon(\tau) + A(\tau) x^\varepsilon(\tau) - A^+(\tau) z(\tau) \\
&\quad + B(\tau) x^\varepsilon(\tau - h(\tau)) - |B(\tau)| z(\tau - h(\tau)) \\
&\leq -\varepsilon x_j^\varepsilon(\tau) + A^+(\tau) [|x^\varepsilon(\tau)| - z(\tau)] \\
&\quad + |B(\tau)| [|x^\varepsilon(\tau - h(\tau))| - z(\tau - h(\tau))] \\
&\leq -\varepsilon x_j^\varepsilon(\tau) < 0.
\end{aligned}$$

Ceci contredit la définition de τ. Le cas $x_j^\varepsilon(\tau) < 0$ se traite de même, conduisant à $-\dot{x}_j^\varepsilon(\tau) - \dot{z}_j(\tau) \leq \varepsilon x_j^\varepsilon(\tau) < 0$. La preuve est obtenue en passant à la limite, en notant que $\lim_{\varepsilon \to 0} x^\varepsilon(t) = x(t)$. ∎

Plusieurs résultats ont été obtenus à partir de l'utilisation du système de comparaison correspondant à l'égalité dans (6.54) et (6.55), soit :

$$\dot{z}(t) = \sup_t \left[A^+(t)\right] z(t) + \sup_t \left[|B(t)|\right] z(t - h(t)), \quad \forall t \geq t_0,$$

ainsi que du lemme suivant permettant de conclure à la stabilité des systèmes linéaires stationnaires obtenus par majoration.

[15] on peut plus strictement montrer $\|z(t)\| > 0$ en utilisant un passage à la limite analogue à celui de la deuxième partie de cette démonstration.

Lemme 6.5.1. [41] *Soient A, B_1 et B_2 des matrices $n \times n$ réelles, h_1 et h_2 des constantes positives ou nulles, et soit le système ($t \geq 0$) :*

$$\dot{z}(t) = A^+ z(t) + |B_1| \sup_{0 \leq \theta \leq h_1} z(t-\theta) + |B_2| \sup_{0 \leq \theta \leq h_2} z(t-\theta). \tag{6.59}$$

Si $(A^+ + |B_1| + |B_2|)$ est de Hurwitz[16]*, alors la solution $z = 0$ est asymptotiquement stable pour (6.59).*

Théorème 6.5.12. [41] *L'équilibre $x = 0$ du système linéaire perturbé :*

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bx(t - h(t)) \\ &\quad + f(x(t), t) + g(x(t - h(t), t) \\ |f(x,t)| &\leq F|x|, \quad |g(x,t)| \leq G|x|, \\ 0 &\leq h(t) \leq h_{\max}, \quad B = B' + B'', \end{aligned}$$

est asymptotiquement stable si la matrice

$$(A + B')^+ + \left|B''\right| + F + G + h_{\max}\left[\left|B'A\right| + \left|B'B\right| + \left|B'\right|(F + G)\right]$$

est de Hurwitz.

Si B' est choisie nulle, on obtient le corollaire suivant.

Corollaire 6.5.1. *L'équilibre $x = 0$ de (6.52) est asymptotiquement stable si $A^+ + |B|$ est une matrice de Hurwitz, avec $A^+ = \sup_t [A^+(t)] < \infty$ et $|B| = \sup_t [|B(t)|] < \infty$.*

Cette deuxième condition est indépendante du retard, mais nécessite que A^+ soit une matrice de Hurwitz. Elle ne permet donc pas d'étudier un éventuel effet stabilisant de la partie retardée $B(t)$. Par contre, la précédente, qui dépend de la valeur maximale du retard $h_{\max}$, nécessite la stabilité asymptotique de $(A + B')^+$ mais non celle de A.

Exemple d'application

Les modèles retardés sont souvent proposés en biologie pour décrire la lutte des espèces et leur dynamique de croissance. Considérons le *modèle logistique* suivant, correspondant au cas où une ressource en nourriture est limitée mais se renouvelle de façon autonome :

$$\dot{x}(t) = \gamma \left[1 - \frac{x(t-h)}{k}\right] x(t). \tag{6.60}$$

$x(t)$ est le nombre d'individus dans la population, le retard h est le temps de reproduction de la nourriture (le retard est quelquefois interprété comme l'âge

[16] Elle est alors l'opposée d'une M-matrice (ou matrice de Metzler, voir [117], chapitre 8).

moyen des reproducteurs). La constante γ est le coefficient de Malthus de croissance linéaire. La constante k est la population moyenne (d'équilibre) et est liée à la capacité de l'environnement à nourrir la population.

Le système (6.60) a deux points d'équilbre : $x = 0$ (mort de l'espèce) et $x = k$ (population moyenne). Pour étudier ce second équilibre, on introduit le changement de variable $x(t) = k[1 + y(t)]$, qui conduit au système d'équilibre $y = 0$ suivant :

$$\dot{y}(t) = -\gamma\, y(t-h)\,[1 + y(t)]\,. \tag{6.61}$$

Nous étudierons la stabilité de $y = 0$ sur le système linéarisé $\dot{y}(t) = -\gamma\, y(t-h)$ (théorème 4), qui est un cas particulier de l'intégrale de Stieljes (6.28) avec :

$$k(\theta) = \begin{cases} 0 \text{ si } \theta < -h, \\ \gamma \text{ si } \theta \geq -h. \end{cases}$$

D'après l'exemple 2, la stabilité asymptotique (locale) de $x = k$ pour (6.61) est garantie si $0 < \gamma < \frac{\pi}{2h}$.

6.6 Cas des systèmes de type neutre

Nous avons déjà présenté cette classe de systèmes dans le cas de retards ponctuels (6.7)(6.9) et distribués (6.14). Dans le cas général, un système de type neutre s'écrit :

$$\dot{x}(t) = f\left(x_t,\ t,\ \dot{x}_t,\ u_t\right), \tag{6.62}$$

Ainsi, dans un système de type neutre, le plus haut degré de dérivation touche à la fois certaines composantes de $x(t)$ et certaines de leurs valeurs passées. Ces systèmes, dont la complexité est un peu supérieure à celle des systèmes de type retardé, sont traités en détail dans [48, 69]. Ici, nous signalerons seulement certaines de leurs caractéristiques en termes de solutions, puis de stabilité.

On représente généralement les systèmes neutres sous la forme de Hale [48] :

$$\dot{Fx}_t = \frac{dFx_t}{dt} = f\left(x_t\ , t\ , u_t\right), \tag{6.63}$$

où $F : \mathcal{C} \to \mathbb{R}^n$ est un opérateur régulier (ce qui évite les systèmes implicites) à argument différé. Dans le cas linéaire, stationnaire et à retards ponctuels, un système neutre s'écrit :

$$\dot{x}(t) - \sum_{j=1}^{q} D_j\, \dot{x}(t-\omega_j) = \sum_{i=0}^{k} \left[A_i x(t-h_i) + B_i u(t-h_i)\right], \tag{6.64}$$

équation à laquelle on associe l'équation linéaire aux différences :

$$Fz_t = z(t) - \sum_{j=1}^{q} D_j z(t-\omega_j) = 0, \quad D_j \text{ matrices constantes.} \tag{6.65}$$

On notera que, dans les publications concernant les applications aux sciences pour l'ingénieur, le cas mono-retard est quasiment le seul représenté, sous la forme particulière suivante :

$$\dot{x}(t) - D\dot{x}(t-h_1) = A_0 x(t) + \sum_{i=1}^{k} \left[A_i x(t-h_i) + B_i u(t-h_i)\right], \tag{6.66}$$

dont l'équation aux différences associée est :

$$z(t) - Dz(t-h_1) = 0. \tag{6.67}$$

Nous avons vu que les solutions des systèmes *de type retardé* voient leur régularité augmenter avec le temps (voir page 245). Cette propriété de « lissage » n'est plus vérifiée pour les systèmes *de type neutre* : à cause de l'équation aux différences (6.65) impliquant $\dot{x}(t)$, la trajectoire peut « répliquer » toute irrégularité de la condition initiale $\varphi(t)$ même si, dans (6.63), f et F présentent des propriétés de régularité très fortes. Ceci peut poser des problèmes dans l'application des méthodes pas à pas [8].

La présence de ce même opérateur aux différences change également les caractéristiques de *stabilité* des systèmes neutres. En effet, à la différence d'un système de type retardé, un système linéaire neutre peut avoir une infinité de pôles instables et le Théorème 6.5.2 ne s'applique plus. Considérons le système linéaire (6.64). Son équation caractéristique s'écrit :

$$\det\left[sI - s\sum_{j=1}^{q} D_j e^{-s\omega_j} - \sum_{i=0}^{k}\left[A_i e^{-sh_i}\right]\right] = 0. \tag{6.68}$$

Dans le plan complexe, à cause de la présence du terme $-s\sum_{j=1}^{q} D_j e^{-s\omega_j}$ dans le déterminant, on peut obtenir des branches infinies de racines complexes tendant vers l'axe imaginaire tout en conservant des parties réelles strictement négatives. Des conditions basées sur le seul signe de la partie réelle doivent donc être considérées avec beaucoup de précaution [65].

Par contre, si on fait l'hypothèse de la stabilité asymptotique de l'équation aux différences (6.65) (ce que l'on nomme « stabilité formelle » du système neutre [16]), alors le nombre de racines instables devient fini [26]. La stabilité formelle est également appelée « f-stabilité » dans le cas non linéaire [69].

Rappelons que plusieurs conditions permettent d'analyser la stabilité asymptotique (donc, exponentielle) du système linéaire stationnaire aux différences (6.65) :

- Dans le cas mono-retard (6.67), une CNS (condition nécessaire et suffisante) est que D ait toutes ses valeurs propres dans le cercle unité ($\det(\lambda I - D) = 0 \Rightarrow |\lambda| < 1$) ou, autrement dit, que $\|D\| < 1$ (il s'agit là de la condition de stabilité usuelle des systèmes linéaires en temps discret).
- Dans le cas mono-retard (6.67) et avec un retard éventuellement variable ($h_1 = h_1(t)$), la condition précédente est suffisante [16, 66].

– Dans le cas multi-retardé (6.65), une CS est que $\sum_{j=1}^{q} \|D_j\| < 1$.
– Dans le cas scalaire multi-retardé ($D_j = d_j \in \mathbb{R}$) et à retards non commensurables[17], une CNS est que $\sum_{j=1}^{q} |d_j| < 1$ (voir [47]).

La stabilité formelle est une condition nécessaire pour la stabilité asymptotique du système neutre (6.64). Par ailleurs, si l'on veut montrer la stabilité *exponentielle* du système neutre, il faut alors vérifier la stabilité « formelle exponentielle », c'est à dire appliquer les conditions ci-dessus en remplaçant la valeur limite 1 par $\beta < 1$, par exemple : $\|D\| \leq \beta < 1$.

La stabilité formelle est également une propriété cruciale pour ce qui est de la stabilisation : il a été montré récemment que chercher à stabiliser un système neutre non formellement stable se heurte à d'importants problèmes de robustesse vis-à-vis des retards. Pour celà, un système scalaire à deux retards non commensurables ω_1, ω_2 a été considéré comme exemple dans [47] :

$$x(t) + d_1 x(t - \omega_1) + d_2 x(t - \omega_2) = u(t), \tag{6.69}$$

$$\text{avec} \qquad |d_1| + |d_2| \geq 1. \tag{6.70}$$

La condition (6.70) implique que (6.69) n'est pas pas exponentiellement stable pour $u = 0$. Pour stabiliser le système, on peut essayer la loi de commande $u(t) = -f_1 x(t - \omega_1) - f_2 x(t - \omega_2)$, qui stabilise (6.69) si, et seulement si, $|d_1 + f_1| + |d_2 + f_2| < 1$. Cependant, si le bouclage est appliqué avec une très légère erreur ϵ_1, ϵ_2 sur les retards, c'est-à-dire si on applique la commande :

$$u(t) = -f_1 x(t - \omega_1 - \epsilon_1) - f_2 x(t - \omega_2 - \epsilon_2), \tag{6.71}$$

alors il existe une suite $(\epsilon_1^j, \epsilon_2^j)$ *tendant vers zéro* telle que (6.69) bouclé par (6.71) soit exponentiellement instable, bien que le même bouclage soit exponentiellement stable pour $\epsilon_1 = \epsilon_2 = 0$. Ceci se montre en utilisant la dernière condition énoncée ci-dessus et le fait que $|d_1| + |f_1| + |d_2| + |f_2| > 1$.

Ainsi, si l'on veut stabiliser l'équation aux différences d'un système neutre non formellement stable, la moindre erreur sur les valeurs des retards de boucle peut être fatale, ce qui est caractéristique d'un manque de robustesse.

[17] C'est-à-dire en rapport irrationnel, voir page 263.

6.7 Modèles pour les systèmes linéaires stationnaires

Nous considérons dans cette partie le cas du système suivant, linéaire, à retards et à paramètres constants :

$$\begin{aligned}\dot{x}(t) = & \sum_{l=1}^{q} D_l\, \dot{x}(t-\omega_l) \\ & + \sum_{i=0}^{k} \left(A_i x(t-h_i) + B_i u(t-h_i)\right) \\ & + \sum_{j=1}^{r} \int_{t-\tau_j}^{t} \left(G_j(\theta)\, x(\theta) + H_j(\theta)\, u(\theta)\right) d\theta,\end{aligned} \tag{6.72}$$

$$y(t) = \sum_{i=0}^{k} C_i x(t-h_i) + \sum_{j=1}^{r} \int_{t-\tau_j}^{t} N_j(\theta) x(\theta) d\theta. \tag{6.73}$$

Ici, en posant $h_0 = 0$, $A_0 \in \mathbb{R}^{n\times n}$ (constante) represente la rétroaction *instantanée* ; les matrices $A_i \in \mathbb{R}^{n\times n}$, $i > 0$ (constantes), correspondent aux phénomènes de *retards ponctuels* ; la somme d'intégrales correspond aux *retards distribués*, pondérés par les G_j sur les intervalles temporels $[t-\tau_j,\, t]$; les matrices D_i constituent la partie neutre ; B_i et $H_j(s)$ sont les matrices d'entrée. Le retard maximal est $h = \max_{i,j,l} \{h_i, \tau_j, \omega_l\}$. L'équation (6.73), $y(t) \in \mathbb{R}^n$, définit l'équation de sortie avec, de même, des parties retardées de façon ponctuelle C_i et distribuée $N_j(\theta)$.

De nombreux systèmes physiques [96] peuvent être représentés (après linéarisation) par ce modèle. Dans la plupart des cas, un seul retard suffit pour la partie neutre (soit $q = 1$) correspondant d'ailleurs à l'un des retards de la partie retardée ($\tau_1 = h_1$). On remarquera que, dans (6.72), $G_j \equiv -G_k$ pour un couple (j,k) permet de représenter un effet de retard « ponctuel-plus-distribué » comme $\int_{t-\tau_j}^{t-\tau_k} G_j(\theta)x(\theta)d\theta$. Par ailleurs, une approximation suplémentaire peut permettre de ramener les retards distribués à une somme de retards ponctuels :

$$\int_{t-\tau}^{t} G(\theta)\, x(\theta) d\theta \approx \frac{\tau}{d} \sum_{i=1}^{d} \alpha_i G(\frac{i\tau}{d}) x(t - \frac{i\tau}{d}),$$

avec des coefficients constants $\alpha_i \in \mathbb{R}$. Cette simplification a motivé l'étude du cas particulier des systèmes à retards ponctuels multiples :

$$\begin{aligned}\dot{x}(t) &= \sum_{i=0}^{k} A_i x(t-h_i) + B_i u(t-h_i), \\ h_0 &= 0 < h_1 < \ldots < h_{k-1} < h_k,\end{aligned} \tag{6.74}$$

et, plus spécialement encore, des systèmes à *retards commensurables* (ou *rationellement dépendants*), pour lesquels les $h_i = i\delta$ sont tous des multiples entiers d'un même retard constant δ, soit :

$$\dot{x}(t) = \sum_{i=0}^{k} A_i x(t - i\delta) + B_i u(t - i\delta), \tag{6.75}$$

$$y(t) = \sum_{i=0}^{k} C_i x(t - i\delta), \quad k\delta = h. \tag{6.76}$$

Cette classe de modèles, finalement assez large[18], peut être représentée par un *système sur anneau*, qui permet d'utiliser les outils de l'algèbre [117]. Pour celà, on représente l'opérateur de retard $\nabla : x(t) \mapsto x(t - \delta)$ (ou bien $e^{-\delta s}$ en calcul opérationnele de Laplace) par la variable ∇ (lettre grecque « nabla »). On définit $\mathbb{R}[\nabla]$ comme l'anneau (commutatif) des polynômes en ∇ à coefficients réels. Un élément $\mathbf{M}(\nabla)$ de $\mathbb{R}[\nabla]^{m\times p}$ est alors une matrice $m \times p$ sur l'anneau $\mathbb{R}[\nabla]$ définie par $\mathbf{M}(\nabla) \triangleq \sum_{i=0}^{k} M_i \nabla^i$, où les matrices M_i sont dans $\mathbb{R}^{m\times p}$. Le système (6.75, 6.76) peut alors s'écrire sous la forme suivante :

$$\dot{x}(t) = \mathbf{A}(\nabla)x(t) + \mathbf{B}(\nabla)u(t), \tag{6.77}$$

$$y(t) = \mathbf{C}(\nabla)x(t), \tag{6.78}$$

$$\mathbf{A}(\nabla) \in \mathbb{R}^{n\times n}[\nabla], \quad \mathbf{B}(\nabla) \in \mathbb{R}^{n\times m}[\nabla], \quad \mathbf{C}(\nabla) \in \mathbb{R}^{p\times n}[\nabla].$$

Le formalisme opérationnel peut aussi être employé. En considérant que toutes les variables sont nulles avant l'instant initial $t = 0$, il conduit aussi à une formulation entrée-sortie classique basée sur les opérateurs s et e^{-s}. Ainsi, dans le cas plus général (6.72) (6.73) (retards distribués), et lorsque les noyaux (G_j, H_j, N_j) sont des matrices constantes (hypothèse que nous conserverons dans

[18] Les deux principales restrictions sont celle de linéarité et celle de retard constant. Elles s'obtiennent respectivement par linéarisation locale et moyennage des retards. Même si elle présente une importance en mathématiques (évitement de chaos, par exemple), la commensurabilité représente une moindre contrainte pour des systèmes de l'ingénieur, où la valeur numérique du retard provient d'une identification et laisse une marge d'appréciation. On peut ainsi choisir de prendre deux retards commensurables 1 et 1.4 plutôt que 1 et $\sqrt{2}$.

toute la fin de cette partie), on arrive à :

$$\begin{aligned}
\overline{y}(s) &= \mathrm{C}(s)(sI_n - \mathrm{A}(s))^{-1}\mathrm{B}(s)\,\overline{u}(s), \qquad (6.79)\\
\mathrm{C}(s) &= \sum_{i=0}^{k} C_i e^{-sh_i} + \sum_{j=1}^{r} N_j \frac{1-e^{-s\tau_j}}{s},\\
\mathrm{A}(s) &= \sum_{l=0}^{q} D_l\, s\, e^{-s\omega_l} + \sum_{i=0}^{k} A_i e^{-sh_i} + \sum_{j=1}^{r} G_j \frac{1-e^{-s\tau_j}}{s},\\
\mathrm{B}(s) &= \sum_{i=0}^{k} B_i e^{-sh_i} + \sum_{j=1}^{r} H_j \frac{1-e^{-s\tau_j}}{s},\\
\overline{u}\,(s) &= \mathcal{L}(u(t)) \triangleq \int_0^{\infty} e^{-st} u\,(t)\,dt, \quad \overline{y}\,(s) = \mathcal{L}(y(t)),
\end{aligned}$$

Ce formalisme fait apparaître, comme en dimension finie, la notion de *pôles* du système, solutions de l'*équation caractéristique* $\Delta(s) = 0$ avec :

$$\Delta(s) = \det(sI_n - \mathrm{A}(s)), \qquad (6.80)$$

$$\sigma(\mathrm{A}) = \{s \in \mathbb{C}, \Delta(s) = 0\}, \qquad (6.81)$$

et qui conditionnent les solutions de (6.72) à noyaux G_j constants. L'ensemble des pôles constitue le *spectre* $\sigma(\mathrm{A})$. Bien sûr, à part dans le cas dit *dégénéré*, l'équation transcendentale $\Delta(s) = 0$ a une inifinité de racines, autrement dit $\mathrm{card}\,\sigma(\mathrm{A}) = \infty$.

Le cas des systèmes à retards commensurables ($h_i = i\delta$, $\tau_j = j\delta$, $\omega_l = l\delta$), (6.79) peut être reformulé sous forme de matrice de transfert sur le corps $\mathbb{R}(s, e^{-\delta s})$ des fractions rationnelles en s et $e^{-\delta s}$, soit :

$$\mathrm{M}(s, e^{-\delta s}) = \mathrm{C}(s)(sI_n - \mathrm{A}(s))^{-1}\mathrm{B}(s). \qquad (6.82)$$

Le cadre *comportemental* [146] fournit une alternative à la théorie des transferts sur anneau. Un comportement $\mathcal{B}$ est défini comme $\mathcal{B} = \ker R$, où R est une matrice d'opérateurs différentiel et de retard $(s, \triangledown)$ agissant sur l'espace des fonctions. Si les deux approches sont voisines (comportement $\ker[D; -N]$ et transfert $D^{-1}N$), le cadre comportemental est cependant mieux adapté pour les questions de réalisation[19], puisqu'il intègre les éventuels modes non observables ou non contrôlables [37].

Une autre formulation générale (et, dans ce cas, le retard peut être infini) des systèmes linéaires (6.72) se base sur les *intégrales de Stieltjes*. Dans le cas

[19] Nous commenterons cette notion dans la partie 6.9.

retardé ($D_k = 0$)[20], on a :

$$\begin{aligned} \dot{x}(t) &= \int_{-\infty}^{0} [dK(\theta)]\, x(t+\theta), \\ x(t) &\in \mathbb{R}^n,\ t \geq 0, \\ x(\theta) &= \varphi(\theta) \quad \forall \theta \in]-\infty, 0], \end{aligned} \tag{6.83}$$

où tous les coefficients k_{ij} de la $(n \times n)-$matrice canonique $K(\theta)$ sont des fonctions à variation bornée. Avec quelques hypothèses[21] sur $K(\theta)$ et φ, la transformée de Laplace existe et, pour des valeurs de $\mathrm{Re}\, s$ suffisamment élevées, on a :

$$\left[sI - \overline{K}(s)\right] \overline{x}(s) = \varphi(0) + \overline{F}(s), \quad s \in \mathbb{C},$$

$$\begin{aligned} F(t) &= \textstyle\int_{-\infty}^{-t} [dK(\theta)]\, \varphi(t+\theta), & \overline{F}(s) &= \textstyle\int_0^\infty e^{-s\theta} F(\theta)\, d\theta, \\ \overline{K}(s) &= \textstyle\int_{-\infty}^{0} e^{-s\theta} dK(\theta), & \overline{x}(s) &= \textstyle\int_0^\infty e^{-s\theta} x(\theta)\, d\theta. \end{aligned}$$

L'équation caracteristique (6.80) de (6.83) est :

$$\Delta(s) = \det\left[sI_n - \overline{K}(s)\right] = 0. \tag{6.84}$$

6.8 Quelques liens entre modélisation et stabilité

Puisque l'analyse de stabilité est le plus souvent menée au moyen de conditions suffisantes mais non nécessaires, le choix du modèle de départ peut influencer les résultats obtenus. Cette partie présente donc quelques transformations de modèles qui peuvent améliorer l'étude de stabilité.

Formule de Leibniz-Newton

La plupart des résultats de stabilité dépendant du retard ont été obtenue par une re-formulation du modèle de départ, faisant apparaitre le retard dans les gains du modèle transformé. Nous ferons tout d'abord une constatation sur le système simple suivant :

$$\dot{x}(t) = A_1 x(t) + A_2 x(t-h), \quad x \in \mathbb{R}^n. \tag{6.85}$$

La stabilité indépendante du retard requiert que A_0 soit une matrice de Hurwitz (ce qui se retrouve de façon cohérente dans la condition LMI (6.44)) et que $A_1 + A_2$ le soit aussi (condition obtenue pour un retard nul, voir Théorème 6.5.5). A l'inverse, une condition assurant la stabilité pour un retard borné $h \in [0; h_M[$

[20] Pour le cas neutre ($D_k \neq 0$), on ajoute $\int_{-\infty}^{0} [dK_N(\theta)]\, \dot{x}(t+\theta)$ à droite de (6.83).

[21] $F(t)$ absolument convergente, $\int_{-\infty}^{0} |\theta|\, |dk_{ij}(\theta)| < +\infty$, $\|\varphi(0)\| + (\int_0^\infty \|\varphi(\theta)\|^2\, d\theta)^{\frac{1}{2}} < \infty$.

demande la stabilité de $A_1 + A_2$ (ce qui apparaît dans la condition (6.46) mais pas, en général, celle de A_1.

Plusieurs résultats[22] sur la stabilité dépendante du retard on été développés sur la base de la formule de Leibniz-Newton : $\int_{t-h}^{t} \dot{x}(s)ds = x(t) - x(t-h)$. Ainsi, en posant :

$$A_i x(t-h) = [A_i - L_i]\, x(t-h_i) + L_i \left[x(t) - \int_{t-h_i}^{t} \dot{x}(s)ds \right], \tag{6.86}$$

on obtient la transformation du système :

$$\dot{x}(t) = \sum_{i=1}^{m} A_i x(t-h_i) \text{ (avec, potentiellement, } h_1 = 0), \tag{6.87}$$

en un modèle à retard augmenté $h = \max(h_i + h_j)$:

$$\dot{x}(t) = \left[\sum_{i=1}^{m} L_i \right] x(t) + \sum_{i=1}^{m} [A_i - L_i]\, x(t-h_i) + \sum_{i=1,j=1}^{m} \int_{t-h_i}^{t} L_i A_j x(s-h_j)ds. \tag{6.88}$$

De cette façon, même si la partie non retardée $[A_1]$ de (6.85) est instable, elle peut être « remplacée » par une matrice stable, $[L_1]$ dans (6.88). Une telle décomposition peut être optimisée par le biais d'algorithmes LMI, pour relâcher les condiditons de stabilité. Il a cependant été remarqué dans [45] que cette transformation augmente le nombre de racines caractéristiques.

Formes de Kolmanovski-Richard

Le principe précédent peut être généralisé à d'autres transformations et d'autres systèmes[23]. En reprenant la notation (6.48), le système retardé (6.87) peut être ré-écrit des trois façons qui suivent :

$$\dot{x}(t) = Ax(t) - \sum_{i,j=1}^{m} A_{ij} \int_{t-h_{ij}}^{t-h_j} x(s)ds, \tag{6.89}$$

$$\dot{x}(t) = Ax(t) - \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} \dot{x}(s)ds, \tag{6.90}$$

$$\frac{d}{dt} \left[x(t) + \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} x(s)ds \right] = Ax(t). \tag{6.91}$$

[22] Les premiers travaux parus sont ceux de [41] dans le cas mono-retard, suivis de [101] pour des retards multiples. Les autres références figurent dans [45].

[23] Les résultats de cette partie sont issus de [71], où le cas des systèmes distribués est également considéré.

Par une procédure « en deux temps » (“*two-step procedure*” décrite dans [64]), chacune de ces écritures inspire ensuite des fonctions de Liapounov-Krasovskii adaptées, qui sont celles présentées plus haut dans les trois démonstrations de la partie 6.5.

Comme dans le cas de la transformation de Leibniz-Newton, les propriétés de stabilité du système original et de ses transformées ne sont pas équivaklentes. Alors que celle de (6.89) implique celle de (6.87) [73], le problème réciproque a été étudié dans [61], qui montre que la stabilité des premier et deuxième systèmes transformés n'est pas nécessaire à celle de l'initial, car des dynamiques additionnelles [45] sont introduites par la transformation, dynamiques dont le pôles sont absents du spectre initial[24]. La stabilité du troisième système est suffisante et devient nécessaire si on fait l'hypothèse que l'équation aux différences $x(t) + \sum_{i=1}^{m} A_i \int_{t-h_i}^{t} x(s)ds = 0$ est asymptotiquement stable (propriété de *stabilité formelle* pour les systèmes neutres, voir page 260).

La forme de Fridman (descripteur)

La « forme descripteur » (“*descriptor form*” en anglais) a été introduite par E. Fridman [29, 32]). Elle correspond à un modèle 2-D (voir page 240) et met en jeu la transformation (6.86), enrichie par des techniques liées aux systèmes singuliers. On ré-écrit le système linéaire retardé (6.72), ici considéré en régime libre ($u(t) \equiv 0$), sous la forme singulière suivante :

$$\begin{aligned} \dot{x}(t) &= z(t), \\ 0 \times \dot{z}(t) &= -z(t) + \sum_{l=1}^{q} D_l\, z(t-\omega_l) \\ &\quad + \sum_{i=0}^{k} A_i x(t-h_i) + \sum_{j=1}^{r} \int_{t-\tau_j}^{t} G_j(\theta)\, x(\theta) d\theta, \end{aligned} \tag{6.92}$$

puis on lui applique la transformation (6.86) avec $A_i = L_i$. En définissant le vecteur augmenté $X^T(t) = [x^T(t), z^T(t)]$, on peut alors mettre en oeuvre la

[24] Ainsi, la transformation (6.89) appliquée au système (6.85) donne :

$$\dot{x}(t) = (A_1 + A_2)x(t) - A_2 \int_{t-h}^{t} [A_1 x(s) + A_2 x(s-h)] ds,$$

dont les pôles sont les zéros de : $\det\left(I_n - \frac{1-e^{-sh}}{s} A_2\right) \det\left(sI_n - A_1 - A_2 e^{-sh}\right)$, alors que les pôles de (6.85) sont les zéros de $\det\left(sI_n - A_1 - A_2 e^{-sh}\right)$. Si $\|A_2\| > h^{-1}$, alors (6.89) est instable même si (6.85) est stable [44].

fonctionnelle de Liapounov-Krasovskii suivante :

$$V(x_t, z_t) = X^T(t)EPX(t) + V_3(x_t) + V_4(z_t), \tag{6.93}$$

$$E = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix},\ P = \begin{bmatrix} P_1 & 0 \\ P_2 & P_3 \end{bmatrix},\ P_1 = P_1^T. \tag{6.94}$$

La notation V_3, V_4 est celle de (6.45). Remarquons que, puisque la matrice E est singulière, la fonctionnelle est dégénérée (c'est-à-dire n'est pas définie positive), ce qui correspond aux techniques utilisées pour les systèmes singulièrement pertubés. Une optimisation LMI permet de calculer les matrices définissant les fonctionnelles V_3, V_4.

Les résultats concernent à la fois les systèmes de type retardé et neutre, avec incertitudes polytopiques. Sans être exhaustif, nous donnerons ici un résultat simple : le système (6.41) est asymptotiquement stable s'il existe $P_1 = P_1^T > 0$, P_2, P_3, $Q = Q^T, R = R^T$, telles que :

$$\begin{pmatrix} A^TP_2 + P_2^TA & P_1 - P_2^T + A^TP_3 & hP_2^TA_1 \\ P_1 - P_2 + P_3^TA & -P_3 - P_3^T + hR & hP_3^TA_1 \\ hA_1^TP_2 & hA_1^TP_3 & -hR \end{pmatrix} < 0.$$

Techniques de réductions par prédicteur

L'appellation *réduction d'Artstein* ("Artstein model reduction" en anglais) réfère à un article de 1982 [6] mais le principe de cette technique peut aussi être trouvé dans des travaux antérieurs [76], [80] et [132].

Son usage est assez simple si on considère des systèmes avec retard sur l'entrée seule :

$$\dot{x}(t) = Ax(t) + Bu(t-h), \quad x(t) \in \mathbb{R}^n. \tag{6.95}$$

En introduisant le nouveau vecteur : :

$$z(t) = x(t) + \int_{t-h}^{t} e^{A(t-h-\theta)}Bu(\theta)\,d\theta, \tag{6.96}$$

on réduit (6.95) à un système sans retard :

$$\dot{z}(t) = Az(t) + e^{-Ah}Bu(t), \quad z(t) \in \mathbb{R}^n. \tag{6.97}$$

On peut alors calculer facilement sur cette EDO un retour d'état classique, $u(t) = K_0z(t)$, à condition que la paire (A, B) soit stabilisable (ce qui garantit que la paire $(A, e^{-Ah}B)$ l'est aussi). En revenant à la notation du système initial, le contrôle résultant intègre donc un effet de retard distribué : $u(t) = K_0x(t) + \int_{t-h}^{t} K_0e^{A(t-h-\theta)}Bu(\theta)\,d\theta$.

Le problème est sensiblement plus compliqué si le système présente aussi un retard sur l'état. Le cas suivant a été considéré par Fiagbedzi et Pearson [27] :

$$\dot{x}(t) = A_0 x(t) + A_0 x(t-h) + B_0 u(t) + B_1 u(t-\tau), \tag{6.98}$$

qui aboutissent à la transformation :

$$\begin{aligned} z(t) = x(t) &+ \int_{t-h}^{t} e^{A(t-h-\theta)} A_1 x(\theta)\, d\theta \\ &+ \int_{t-\tau}^{t} e^{A(t-\tau-\theta)} B_1 u(\theta)\, d\theta. \end{aligned} \tag{6.99}$$

Le modèle réduit est alors :

$$\begin{aligned} \dot{z}(t) &= Az(t) + Bu(t), \quad z(t) \in \mathbb{R}^n, \\ A &= A_0 + e^{-Ah} A_1, \\ B &= B_0 + e^{-A\tau} B_1. \end{aligned} \tag{6.100}$$

Cependant, résoudre l'équation caractéristique matricielle (6.100) devient beaucoup moins simple. Dans [27], il est proposé de limiter le problème de calcul aux seuls vecteurs et valeurs propres instables (qui, pour un système retardé, sont en nombre fini). Une application (simulée) au controle de ralenti d'un moteur thermique est ptrésentée dans [35]. Plus généralelment encore, cette approche peut être considérée sur l'équation (6.83), mais demande alors l'étude de la structure propre de l'équation caratéristique $A = \int_{-h}^{0} e^{A\theta} dK(\theta)$.

Les contrôleurs obtenus par ces techniques de réduction contiennent des termes intégraux comme ceux de (6.96) et (6.99). A ce titre, ils font partie des contrôles de type prédicteur. Dans certains cas, ils peuvent donc s'avérer sensibles aux incertitudes paramétriques et, plus encore, aux erreurs d'identification du retard. La réduction d'Artstein constitue cependant un outil d'utilisation simple et très intéressant dans le cas de retard sur l'entrée seule.

Forme exponentielle de Seuret

La transformation qui suit vise à étudier la stabilité exponentielle d'un système à retard variable. Elle a été initialement introduite dans [126], où la stabilisation exponentielle robuste était considérée (voir également [125]). Nous n'en exposerons ici que le principe, à partir du cas simple suivant :

$$\dot{x}(t) = A_0 x(t) + A_1 x(t - \tau(t)), \tag{6.101}$$

où $\tau(t)$ est un retard variable borné, vérifiant les inégalités suivantes :

$$0 \leq h_1 \leq \tau(t) \leq h_2, \quad \forall t \geq 0, \tag{6.102}$$

Comme défini en (6.25), montrer la stabilité exponentielle à taux α signifie de prouver l'existence de deux réels α et k tels que : $|x(t, t_0, \psi)| \leq k \|\psi\|_{\mathcal{C}}\, e^{-\alpha(t-t_0)}$.

Celà revient aussi à montrer la convergence asymptotique du vecteur $e^{\alpha(t-t_0)}$ $x(t,t_0,\psi)$ vers zéro.

En choissant, sans restriction pour la suite, l'instant initial $t_0 = 0$, il est donc naturel d'introduire la nouvelle variable vectorielle $z = e^{\alpha t}x(t)$. Cette variable satisfait l'équation transformée suivante :

$$\dot{z}(t) = (A_0 + \alpha I_n)z(t) + e^{\alpha\tau(t)}A_1 z(t-\tau(t)), \tag{6.103}$$

dont la stabilité asymptotique pour un certain $\alpha > 0$ garantira la stabilité exponentielle de taux α pour le système initial (6.101). Cependant, une difficulté apparaît pour étudier le système transformé (6.103) : puisque $\tau(t)$ est variable, (6.103) est un système non stationnaire. Pour surmonter cet obstacle, [126] a proposé d'exprimer (6.103) sous une forme polytopique, basée sur l'existence de coefficients variables λ_1 et λ_2 conduisant à l'expression du terme $e^{\alpha\tau(t)}$ sous forme d'une somme convexe de ses bornes $e^{\alpha h_1}$ et $e^{\alpha h_2}$:

$$\begin{aligned} &e^{\alpha\tau(t)} = \lambda_1(t)e^{\alpha h_1} + \lambda_2(t)e^{\alpha h_2}, \\ &\lambda_1(t), \lambda_2(t) \geq 0 \text{ et } \lambda_1(t) + \lambda_2(t) = 1. \end{aligned} \tag{6.104}$$

Par ce biais, le système (6.103) est inclus dans la classe polytopique :

$$\dot{z}(t) = A_0 + \alpha I_n z(t) + \sum_{i=1}^{2} e^{\alpha h_i}\lambda_i(t)A_1 z(t-\tau(t)), \tag{6.105}$$

dont l'analyse est ensuite menée sous la forme descripteur :

$$\begin{cases} \dot{z}(t) = y(t), \\ 0 = -y(t) + (A_0 + \alpha I_n + \bar{A}_1(t))z(t) - A_1\displaystyle\int_{t-\tau}^{t} y(s)ds, \end{cases}$$

ou encore :

$$E\dot{\bar{z}}(t) = \begin{bmatrix} 0 & I_n \\ A_0 + \alpha I_n + \bar{A}_1(t) & -I_n \end{bmatrix} \bar{z}(t) - \begin{bmatrix} 0 \\ A_1 \end{bmatrix} \int_{t-\tau}^{t} y(s)ds, \tag{6.106}$$

avec $E = diag\{I_n, 0\}$, $\bar{z}(t) = col\{z(t), y(t)\}$ et $\bar{A}_1(t) = \sum_{i=1}^{2} e^{\alpha h_i}\lambda_i(t)A_1$.

Des conditions sous forme LMI sont alors obtenues, généralisables au cas de matrices perturbées [125, 126] et aux systèmes de type neutre [127].

Techniques de pseudo-retard

Ces techniques se placent dans le cadre fréquentiel et permettent de remplacer le retard par une fraction rationnelle. La stabilité du système (6.85) dépend des racines de l'équation caratéristique :

$$\det(sI - A_1 - A_2 e^{-sh}) = d(s) + n(s)e^{-hs} = 0. \tag{6.107}$$

Les valeurs critiques de h qui correspondent à des bifurcations de stabilité correspondent aux racines imaginaires pures $s = j\omega$ de (6.107), vérifiant donc $\left|\frac{d(j\omega)}{n(j\omega)}\right| = \left|e^{-hj\omega}\right| = 1$. De là, l'idée générale des techniques de pseudo-retard est, tout d'abord, de chercher les éventuelles intersections de la courbe $\frac{d(j\omega)}{n(j\omega)}$ avec le cercle unité. Ceci définit les « fréquences de croisement » ω_i.

Dans cette technique fréquentielle, le retard e^{-hs} peut être, de façon équivalente, remplacé par toute fonction de transfert $\frac{p_n(sT)}{p_d(sT)}$ ayant un module unitaire. Les fréquences de croisement ω_i sont alors obtenues en résolvant l'équation : $d(j\omega)p_n(j\omega T) + n(j\omega)p_d(j\omega T) = 0$ pour des valeurs croissantes du paramètre $T \in]0, +\infty[$. Cette équation est simple (polynomiale) et lorsqu'une paire (ω_i, T_i) donne une solution, le ou les retards correspondants en sont déduits. Le paramètre T est appelé le *pseudo-retard*. Z.V. Rekasius [115] a ainsi introduit la transformation $\frac{p_n(sT)}{p_d(sT)} = \frac{(1-sT)}{(1+sT)}$. A. Thowsen [136], en utilisant $\frac{(1-sT)^2}{(1+sT)^2}$, a obtenu l'équivalence suivante, correspondant au système (6.75) en régime libre.

Théorème 6.8.1. *$s = j\omega$, $\omega \geq 0$, est une racine du quasi-polynôme $\sum_{i=0}^{k} d_i(s)e^{-i\delta s}$ pour un $\delta > 0$ si, et seulement si, c'est aussi une racine de $\sum_{i=0}^{k} d_i(s)\,(1 - sT)^{2i}\,(1 + sT)^{2(k-i)}$ pour un $T \geq 0$.*

Ainsi, bien que ces techniques rappellent les approximations rationnelles (6.1), nous constatons que les pseudo-retard ne sont pas une technique d'approximation (voir aussi [100] p.137). Ils sont plutôt reliés aux transformations bilinéaires du type : $\mathbb{C} \to \mathbb{C}$, $z = e^{-j\omega h} \mapsto w = \frac{1-z}{1+z} = j\tan\frac{\omega h}{2}$, qui mettent en correspondance le cercle unité et l'axe imaginaire et sont bien connues en traitement du signal comme en commande numérique (« transformée homographique » ou « en w »).

Parmi les techniques associées aux pseudo-retards, on retiendra le résultat de K. Walton et J.E. Marshall [142], particulièrement pratique pour les systèmes mono-retard, c'est-à-dire dont l'équation caractéristique est (6.107). Il utilise l'équation $d(j\omega)d(-j\omega) - n(j\omega)n(-j\omega) = 0$ et, donc, le choix $\frac{p_n}{p_d} = -\frac{d(-s)}{n(-s)}$).

Modélisation issue des techniques de robustesse

En améliorant une technique de [34] (voir aussi [100] page 154), plusieurs auteurs [52, 54, 59] ont développé des approches inspirées de différentes techniques de commande robuste : techniques IQC, $\mu-$synthèse, approche de Kalman-Yakubovitch-Popov, théorie de Popov généralisée, etc.

Nous décrirons tout d'abord ici les grandes lignes de l'approche de M. Jun et M.G. Safonov [59]. On considère le système (6.41) pour $h \in [0, h_M]$, en supposant que $A = A_0 + A_1$ est asymptotiquement stable (ce qui est la condition de stabilité asymptotique de (6.41) à retard h nul). On associe à ce système un modèle opérationnel constitué de deux blocs interconnectés, décrit Figure 6.2. Cette représentation se prête à une ré-écriture des conditions de stabilité robuste sous forme de contrainte intégrale-quadratique (IQC en anglais, voir [90, 123]).

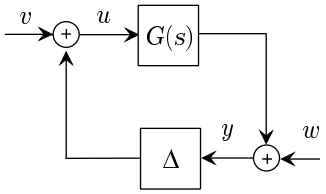


Fig. 6.2: Deux blocs interconnectés pour reformulation IQC.

Dans ce modèle, l'opérateur $\mathbf{\Delta}$ reprend les caractéristiques du retard. Sa transformée de Laplace est $\Delta(s) = \frac{e^{-hs}-1}{hs}$ et le système correspondant au bloc $G(s)$ est défini, dans le domaine temporel, comme suit :

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \\ u(t) = \mathbf{\Delta}(y(t)), \end{cases} \tag{6.108}$$

avec : $A = A_0 + A_1$,
$B = hH, \quad C = EA, \quad D = hEH,$
$A_1 = HE, \quad H \in \mathbb{R}^{n\times q},\ E \in \mathbb{R}^{q\times n},$
$q \leq n, \quad H$ et E sont des matrices de rang plein.

[59] montrent que la stabilité asymptotique du système (6.108) équivaut à celle du système (6.41). La preuve se fonde sur l'équivalence suivante :

$$(sI - A_0 - A_1 e^{-hs})x = 0 \quad \Leftrightarrow \quad (sI - (I - h\Delta(s)A_1)^{-1}(A_0 + A_1))x = 0.$$

Ainsi, (6.41) est asymptotiquement stable si, et seulement si, la matrice à gauche de l'équivalence est régulière dans le demi-plan droit $\operatorname{Re} s \geq 0$, c'est-à-dire si, et seulement si, la matrice à droite l'est aussi. Cette dernière correspond à (6.108).

Le théorème IQC requiert qu'un certain opérateur Φ satisfasse la contrainte intégrale-quadratique suivante pour tout α dans $[0, 1]$:

$$\int_{-\infty}^{+\infty} \left[y^T(t), \alpha\mathbf{\Delta}(y(t))\right] \mathbf{\Phi} \begin{bmatrix} y(t) \\ \alpha\mathbf{\Delta}(y(t)) \end{bmatrix} dt \geq 0.$$

On peut ainsi guarantir la stabilité asymptotique de (6.108) à la condition que A soit une matrice de Hurwitz et que :

$$\exists \varepsilon > 0,\ \forall \omega \in \mathbb{R},\ \left[G^T(-j\omega), I\right] \Phi \begin{bmatrix} G(j\omega) \\ I \end{bmatrix} \leq -\varepsilon I.$$

Cette condition, fréquentielle, peut ensuite être reformulée de façon équivalente sous une forme LMI grâce au lemme de Kalman-Yakubovich-Popov (lemme KYP, voir [145]).

Le choix de : $\Phi = \begin{bmatrix} Q & S \\ S^T & -Q \end{bmatrix}$ conduit au critère (suffisant) suivant :

Théorème 6.8.2. [59] *Le système (6.41) est asymptotiquement stable s'il existe deux matrices symétriques définies-positives $P > 0$, $Q > 0$ et une matrice anti-symétrique S telles que la matrice suivante soit définie-négative :*

$$\begin{bmatrix} A^TP + PA + C^TQC & PB + C^TS + C^TQD \\ B^TP + S^TC + D^TQC & -Q + D^TQC + S^TD + D^TS \end{bmatrix} < 0.$$

Une classe plus générale de systèmes a été considérée dans [52], incluant des retards multiples h_i dont, éventuellement, la borne inférieure pouvait être non nulle[25] : $h_i \in [h_{im}, h_{iM}]$). L'approche de robustesse est différente ($\mu-$synthèse, voir [137]). Pour traiter le cas $h_{im} > 0$, le sous-système $G(s)$ doit inclure des retards $e^{-h_{im}s}$. Cependant, les conditions obtenues peuvent s'avérer assez conservatives.

Parmi les techniques s'inspirant de la commande robuste, citons également l'utilisation de la théorie de Popov généralisée [55] dans le cas élargi des systèmes à retard sur l'état [54]. Ce résultat (voir également [100]) étudie des techniques de stabilisation par rétroaction et de commande H_∞ pour des systèmes à retards ponctuels ou distribués, formulées en termes de triplets de Popov $\Sigma = (A, B; P)$, $P = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}$, pour lesquels un « système de Kalman-Yakubovich-Popov en J » (c'est-à-dire uns système non linéaire à matrices inconnues X, V, W) doit être résolu :

$$\begin{aligned} R &= V^TJV, \\ L + XB &= W^TJV, \\ Q + A^TX + XA &= W^TJW. \end{aligned}$$

Avant de clore cette partie, mentionnons l'article [140] où différentes techniques de robustesse (lemme KYP, lemme Réel Strictement Borné, stabilité absolue, passivité) ont été appliquées à l'étude se stabilité des systèmes linéaires à retards ponctuels et *non-linéairement* pertubés.

[25] Ce type de condition correspond, dans la littérature en anglais, à l'appellation *nonsmall delay*. En français, on pourra parler de « retard borné » si $h_{im} \neq 0$, par distinction du « retard majoré » si $h_{im} = 0$ et $h_{iM} < \infty$.

Approximation LPV

Dans [7], S.P. Banks considère le système non linéaire (ici, sans entrée) :

$$\begin{aligned} \dot{x}(t) &= A\left(x(t), x(t-h)\right) x(t), \\ x(\theta) &= \varphi(\theta), \qquad -h \leq \theta \leq 0, \end{aligned} \tag{6.109}$$

qui est remplacé, comme suit, par une suite d'approximations linéaires à paramètres variants (LPV) :

$$\begin{aligned} \dot{x}^{[i]}(t) &= A\left(x^{[i-1]}(t), x^{[i-1]}(t-h)\right) x^{[i]}(t), \\ x^{[i]}(\theta) &= \varphi(\theta), \qquad -h \leq \theta \leq 0. \end{aligned} \tag{6.110}$$

Si $A(.,.)$ est localement lipschitzienne en ses deux variables, il est montré que la suite de fonctions $x^{[i]}(\theta)$ converge, dans $\mathcal{C}([0,T])$ pour un certain $T \in]0,h]$, vers la solution de (6.109). L'étude de stabilité fait intervenir une algèbre de Lie nilpotente associée à (6.110) et est menée grâce à une transformation de Liapounov particulière. La stabilisation est également étudiée.

6.9 Propriétés structurelles

La commandabilité des systèmes à retards présente trois principales différences par rapport au cas ordinaire :

1. La première différence vient la *nature fonctionnelle de l'état* : la notion de contrôlabilité (terme employé de préférence à « commandabiilté » lorsqu'on se place comme ici en dimension infinie) représente le fait de pouvoir rallier un état à un certain instant t_1. Pour un système à retard, il s'agit donc de rejoindre, à t_1, une fonction x_t de support $[-h, 0]$, ce qui demande de placer le vecteur $x(t)$ depuis l'instant $t_1 - h$ jusqu'à t_1. Pour une EDO, la commandabilité demande de rallier un *point* $x(t)$ à un instant t_1.
2. La deuxième différence (liée à la première) tient au *temps nécessaire au contrôle*. Pour un système linéaire et sans retard démarrant à l'instant t_0, tout point pouvant être atteint à l'instant $t_1 > t_0$ peut aussi l'être à l'instant $t_0+\alpha(t_1-t_0)$, $\alpha > 0$. Au contraire, la présence d'un retard entraîne généralement l'existence d'un *temps d'atteinte* minimum. . Pour donner un exemple, il est clair que le simple système $\dot{x}(t) = x(t) + u(t-1)$ ne peut pas être controllé en moins d'une seconde. Ainsi, en plus des notions de commandabilité usuelles (sous-espaces et indices de commandabilité), on devra ajouter un autre type d'index : la « classe » d'un système linéaire retardé correspond au nombre de fois le retard nécessaires pour atteindre l'objectif.
3. Enfin, la nature de la *réalisation du contrôle* à appliquer, constitue un enjeu important. Pour un système retardé, l'expression générique d'un retour

d'état est $u(t) = g(x_t)$, ce qui signifie que le contrôleur est lui aussi de dimension infinie (fonctionnel). Pour des raisons d'implantation numérique, on peut préférer restreindre sa réalisation à un retour instantanné (en anglais, "memoryless control") du type $u(t) = g(x(t))$ oui bien à un retour à retards ponctuels du type $u(t) = g(x(t), x(t-h_i))$.

La suite de cette partie reprendra quelques définitions relatives à la contrôlabilité des systèmes à retards. Des correspondances plus complètes, obtenues dans des contextes unificateurs, pourront être trouvées dans [28, 84] (contexte algébrique de la théorie des modules) ou [36] (approche comportementale).

Dans le cas des systèmes retardés, les notions de contrôlabilité peuvent être transposées à l'observabilité (voir [113] et les références incluses). Cependant, pour effectuer cette transposition dans le cas des systèmes neutres, l'hypothèse de stabilité formelle(voir page 260) est requise. Dans le cas contraire, le problème des observateurs asymptotiques est encore ouvert.

Définition 6.9.1. *(Contrôlabilité $\mathcal{M}_2$, ou $\mathcal{M}_2-$approchée [20])* On considère le système (6.13). L'état $\overline{x}_0$ est $\mathcal{M}_2$-contrôlable à l'instant t vers $\overline{x}_1 \in \mathcal{M}_2\big([-h,0];\mathbb{R}^n\big)$ s'il existe une suite de contrôles $\{u_i\}$ définis sur $\mathcal{L}_2\left([0,t];\mathbb{R}^m\right)$ telle que $\overline{x}(t;\overline{x}_0,u_i)$ converge vers $\overline{x}_1$ (au sens d'une norme sur $\mathcal{M}_2$). Le système est $\mathcal{M}_2$-contrôlable à t si tous les états $\overline{x}_0$ sont $\mathcal{M}_2$-contrôlables à t vers tout $\overline{x}_1 \in \mathcal{M}_2\left([-h,0];\mathbb{R}^n\right)$. Il est $\mathcal{M}_2$-strictement contrôlable si, dans la définition précédente, la suite $\{u_i\}$ est remplacée par un contrôle u.

Définition 6.9.2. *(Contrôlabilité absolue [107])* Le système linéaire à retards sur l'entrée $\dot{x}(t) = A_0x(t) + \sum_{i=0}^{k} B_iu(t-i\delta)$ est absolument contrôlable si, pour toute condition initiale $\big\{x(0),\ u(t)_{t\in[-k\delta,\ 0]}\big\}$, il existe un temps $t_1 > 0$ et un contrôle borné $u(t)$ tels que $x(t_1) = 0$ avec $u(\theta) = 0$ pour tout $\theta \in [t_1 - k\delta,\ t_1]$.

Une condition nécessaire et suffisante de contrôlabilité absolue peut être exprimée simplement : rang $\left[E, A_0E, ..., A_0^{n-1}E\right] = n$, avec $E = \sum_{i=0}^{k} e^{-i\delta A_0}B_i$. Cependant, la définition exige une fin de mouvement en régime libre ($u(\theta) = 0$ pour tout $\theta \in [t_1 - k\delta,\ t_1]$), ce qui représente une condition très forte. La définition qui suit s'affranchit de cette contrainte.

Définition 6.9.3. *($(\psi,\mathbb{R}^n)$-contrôlabilité [107, 144])* Le système linéaire (6.75) est $(\psi,\mathbb{R}^n)$-contrôlable (par rapport à une fonction $\psi \in \mathcal{C}$) si, pour toute condition intitiale $\varphi \in \mathcal{C}$, il existe un temps (fini) $t_1 > 0$ et une loi de commande $u(t) \in \mathcal{L}_2\left([0, t_1+h], \mathbb{R}^m\right)$ tels que $x\left(t;\varphi,u\right) = \psi(t-t_1-h)$ pour tout $t \in [t_1,\ t_1+h]$.

Pour des systèmes mono-retard, la $(0,\mathbb{R}^n)$-contrôlabilité peut être testée par des techniques de grammien [144].

Définition 6.9.4. *(Contrôlabilité spectrale [88])* Le système linéaire (6.75), considéré avec la notation (6.77), est spectralement contrôlable si :

$$\operatorname{rank}\Big[sI - \mathbf{A}(e^{-\delta s}), \mathbf{B}(e^{-\delta s})\Big] = n, \quad \forall s \in \mathbb{C}. \tag{6.111}$$

La contrôlabilité spectrale établit des bases très intéressantes pour une mise en œuvre effective de contrôles. Il s'agit bien là d'une propriété fonctionnelle, mais qui ne concerne que le contrôle (au sens du placement) du spectre $\sigma(\mathbf{A})$ défini en (6.81) : il a été montré[26] dans [13] et [36] que le système (6.75) est stabilisable si et seulement si (6.111) est vérifiée pour tout $s \in \mathbb{C}$, $\mathrm{Re}(s) \geq 0$. Par ailleurs, les propriétés spectrales s'étendent facilement à l'ensemble de l'analyse structurelle (stabilizabilité, détectabilité...) : toute matrice de transfert causale de $\mathbb{R}(s, e^{-\delta s})$, comme par exemple l'équation (6.82), admet une réalisation spectralement observable. Elle admet également une réalisation détectable et spectralement contrôlable. Il a néanmoins été remarqué que la notion de *réalisation minimale* (au sens de la contrôlabilité spectrale et de l'observabilité spectrale, comme au sens du nombre minimum d'opérateurs d'integration s^{-1} et de retard e^{-s} requis) ne fait pas toujours sens[27] : la fonction de transfert $\frac{1+e^{-2s}}{s+e^{-s}\pi/2}$ a été donnée comme exemple dans [81].

Définition 6.9.5. *(Contrôlabilité euclidienne, ou $\mathbb{R}^n$-contrôlabilité, $\mathbb{R}^n$-contrôlabilité forte [77])* Le système linéaire (6.75) est $\mathbb{R}^n$-contrôlable si, pour toute condition initiale $\varphi \in \mathcal{C}$ et tout vecteur $x_1 \in \mathbb{R}^n$, il existe un temps $t_1 > 0$ et une loi de commande $u(t) \in \mathcal{L}_2([0, t_1], \mathbb{R}^m)$ tels que $x(t_1; \varphi, u) = x_1$. Il est fortement $\mathbb{R}^n$-contrôlable si n'importe quel $t_1 > 0$ peut être pris. Si la propriété est restreinte à $x_1 = 0$, alors le système est $\mathbb{R}^n$-contrôlable vers l'origine.

La $\mathbb{R}^n$-contrôlabilité n'est pas une propriété fonctionnelle, puisqu'elle concerne le vecteur $x(t)$ et non l'état x_t. On notera trois différences avec le cas de la commandabilité d'une EDO :

- La trajectoire peut ne pas rester en x_1 après t_1.
- Sauf dans le cas rare de la contrôlabilité *forte*, l'instant t_1 ne peut pas être arbitrairement réduit.
- La $\mathbb{R}^n$-contrôlabilité n'est pas équivalente à la $\mathbb{R}^n$-contrôlabilité vers l'origine.

Définition 6.9.6. *(Contrôlabilité forte/faible, ou sur anneau/corps, ou $\mathbb{R}[\nabla]$/$\mathbb{R}(\nabla)$-contrôlabilité, voir [77])* Le système linéaire sur anneau (6.77) est contrôlable sur l'anneau $\mathbb{R}[\nabla]$ ou *fortement contrôlable*, s'il existe une loi de commande de type polynomial $u(t) = f(x, \nabla x, \nabla^2 x, ...)$, permettant de rejoindre tout élément du module $\mathbb{R}^n[\nabla]$ depuis tout état initial $x_0 \in \mathbb{R}^n[\nabla]$. Il est contrôlable sur le corps $\mathbb{R}(\nabla)$ ou *faiblement contrôlable*, s'il existe une loi de commande de type rationnel $u(t) = f(x, \nabla x, \nabla^2 x, ..., \nabla^{-1} x, \nabla^{-2} x,)$ permettant de rejoindre tout élément du module $\mathbb{R}^n[\nabla]$ depuis tout état initial $x_0 \in \mathbb{R}^n[\nabla]$.

La contrôlabilité forte n'est pas non plus de type fonctionnel. Elle met l'accent sur la complexité de la commande à appliquer. Une condition nécessaire

[26] Il s'agit d'une preuve constructive basée sur la propriété de domaine de Bézout.

[27] Comme cela a été montré dans [134], pour des systèmes sur un anneau, même des réalisations minimales peuvent ne pas être isomorphes, de telle sorte qu'il n'existe pas forcément de forme canonique de réalisation, contrairement aux systèmes sur un corps.

et suffisante peut être énoncée à partir du sous-module de contrôlabilité associé à la paire $(\mathbf{A}, \mathbf{B})$, *i.e.* $\langle \mathbf{A} / \operatorname{Im} \mathbf{B} \rangle = \operatorname{Im} \mathbf{B} + \mathbf{A}^2 \operatorname{Im} \mathbf{B} + ... + \mathbf{A}^{n-1} \operatorname{Im} \mathbf{B}$ ou, de façon équivalente, à partir de la matrice de contrôlabilité $\langle \mathbf{A} / \mathbf{B} \rangle = \left[\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, ..., \mathbf{A}^{n-1}\mathbf{B} \right]$. L'article de synthèse [77] donne, dans le cas du système linéaire invariant (6.74) avec retards commensurables, les implications suivantes (ainsi que d'autres qui utilisent la notion de sous-module de torsion) :

$$\begin{array}{c} \mathbb{R}[\nabla]\text{-contrôlabilité forte} \Rightarrow \text{Contrôlabilité absolue} \\ \Rightarrow \mathbb{R}(\nabla)\text{-contrôlabilité faible} \Rightarrow \mathbb{R}^n\text{-contrôlabilité.} \end{array}$$

$$\begin{array}{c} \text{Contrôlabilité approchée} \Rightarrow \text{Contrôlabilité spectrale} \\ \Rightarrow \mathbb{R}(\nabla)\text{-contrôlabilité faible.} \end{array}$$

Cela signifie que la $\mathbb{R}[\nabla]$-contrôlabilité forte est une propriété très exigeante. Effectivement, elle demande que le système soit contrôlable comme s'il ne comportait pas de retard.

Comme nous l'avons déjà signalé page 274, la notion usuelle d'indice de contrôlabilité a été étendue aux systèmes retardés [111], ajoutant, avec la notion de « classe », la prise en compte du temps d'atteinte, temps minimum t_1 nécessaire pour que les différentes composantes de l'état (sous-modules de contrôlabilité) puissent atteindre les valeurs voulues $x_1 \in \mathbb{R}^n$.

6.10 Compléments bibliographiques

Comme l'attestent les très nombreux travaux paraissant aujourd'hui au niveau international, les systèmes à retards constituent un champ d'étude important, pour l'automaticien comme pour l'ingénieur. En complément aux références déjà citées, mentionnons pour ce qui concerne d'autres aspects fondamentaux de l'automatique : modélisation [84], propriétés structurelles [84, 124], commande (par placement de spectre [84, 143], optimale [66, 72], par suivi de modèle [83, 112], par platitude [94], par modes glissants [40], par linéarisation [93]).

Parmi les ouvrages incluant des domaines d'application spécifiques, citons la commande en réseau [120], l'écologie [38], la biologie [87], la robotique [135]. De nombreux exemples sont également présentés dans [66, 84, 100, 122].

6.11 Bibliographie

[1] Abdallah, C., J.D. Birdwell, J. Chiasson, V. Chupryna, Z. Tang et T. Wang: *Load Balancing Instabilities Due to Time Delays in Parallel Computations.* Dans 3^{rd} *IFAC Workshop on Time Delay Systems*, Sante Fe, NM, Dec. 2001.

[2] Abdallah, C. et J. Chiasson: *Stability of Communication Networks in the Presence of Delays*. Dans 3^{rd} *IFAC Workshop on Time Delay Systems*, Sante Fe, NM, Dec. 2001.

[3] Abdallah, G., P. Dorato, J. Benitez-Read et R. Byrne: *Delayed Positive Feedback Can Stabilize Oscillatory Systems*. Dans *ACC93 (American Control Conf.)*, pages 3106–3107, 1993.

[4] Aernouts, W., D. Roose et R. Sepulchre: *Delayed Control of a Moore-Greitzer Axial Compressor Model*. Int. J. of Bifurcation and Chaos, 10(2), 2000.

[5] Ailon, A. et M.I. Gil: *Stability Analysis of a Rigid Robot with Output-Based Controller and Time-Delay*. Syst. and Control Letters, 40(1) :31–35, 2000.

[6] Artstein, Z.: *Linear Systems with Delayed Controls : A Reduction*. IEEE Trans. Aut. Control, 27(4) :869–879, 1982.

[7] Banks, S.P.: *Nonlinear Delay Systems, Lie Algebras and Lyapunov Transformations*. IMA J. Math. Control Information, 19(1-2) :59–72, 2002.

[8] Bellen, A. et M. Zennaro: *A Free Step-Size Implementation of Second Order Stable Methods for Neutral Delay Differential Equations*. Dans 3^{rd} *IFAC Workshop on Time Delay Systems*, Sante Fe, NM, Dec. 2001.

[9] Bellman, R. et K.L. Cooke: *Differential Difference Equations*. Academic Press, New York, 1963.

[10] Biberovic, E., A. Iftar et H. Ozbay: *A Solution to the Robust Flow Control Problem for Networks with Multiple Bottlenecks*. Dans 40^{th} *IEEE CDC01 (Conf. on Dec. and Control)*, pages 2303–2308, Orlando, FL, Dec. 2001.

[11] Borne, P., M. Dambrine, W. Perruquetti et J.P. Richard: *Vector Lyapunov Functions : Nonlinear, Time-Varying, Ordinary and Functional Differential Equations*, chapitre 2, pages 49–73. Stability and Control : Theory, Methods and Appl. Taylor and Francis, London, Martynyuk édition, 2002.

[12] Borne, P., Dauphin Tanguy G., J.P. Richard et I. Rotella F., Zambettakis: *Analyse et Régulation Des Processus Industriels*, tome 1 : régulation continue de *Collection Méthodes et Pratiques de l'Ingénieur, Série Automatique*. Technip, 1993.

[13] Bretheé, D.: *Contribution à l'Etude de la Stabilisation des Systèmes Linéaires à Retards*. (in French), IRCCyN, Univ. of Nantes, EC Nantes, France, Dec. 1997.

[14] Burton, T.A.: *Stability and Periodic Solutions of Ordinary and Functional Differential Equations*, tome 178. Academic Press, Orlando, 1985.

[15] Bushnell, L.: *Editorial : Networks and Control*. IEEE Control Syst. Magazine, 21(1) :22–99, Feb. 2001. Special section on Networks and Control.

[16] Byrnes, C.I., M.W. Spong et T.J. Tarn: *A Several Complex Variables Approach to Feedback Stabilization of Linear Neutral Delay-Differential Systems.* Math. Systems Theory, 17 :97–133, 1984.

[17] Chiasson, J. et J.J. Loiseau: *Applications of Time-delay Systems.* Numéro 352 dans *Lecture Notes in Control and Inform. Sc.* Springer Verlag, Berlin Heidelberg, 2007.

[18] Conte, G. et A.M. Perdon: *Systems over Rings : Theory and Applications.* Dans 1^{rst} *IFAC Workshop on Linear Time Delay Systems*, pages 223–234, Grenoble, France, July 1998. Plenary lecture.

[19] Dambrine, M.: *Contribution à L'étude de la Stabilité Des Systèmes à Retards.* Thèse de doctorat, Laboratoire d'Automatique et d'Informatique Industrielle de Lille, EC Lille, Univ. of Lille, France, Oct. 1994. (in French).

[20] Delfour, M. et S. Mitter: *Controllability, Observability and Optimal Feedback Control of Affine, Hereditary, Differential Systems.* SIAM J. Contr. Optim., 10 :298–328, 1972.

[21] Diekmann, O., S.A. Von Gils, S.M. Verduyn-Lunel et H.O. Walther: *Delay Equations, Functional, Complex and Nonlinear Analysis*, tome 110 de *Applied Math. Sciences.* Springer, 1995.

[22] Dieulot, J.Y. et J.P. Richard: *Tracking Control of a Nonlinear System with Input-Dependent Delay.* Dans 40^{th} *IEEE CDC01 (Conf. on Dec. and Control)*, Orlando, FL, Dec. 2001.

[23] Dion, J.M., L. Dugard et S.I. Niculescu: *Time Delay Systems.* Special issue of Kybernetica, 37(3-4), 2001.

[24] Driver, R.D.: *Ordinary and Delay Differential Equations.* Applied Math. Sciences. Springer, 1997.

[25] Dugard, L. et E.I. Verriest: *Stability and control of time-delay systems.* Numéro 228 dans *Lecture Notes in Control and Inform. Sc.* Springer Verlag, 1997.

[26] Elsgolts, L.E. et S.B. Norkin: *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*, tome 105 de *Mathematics in Sc. and Eng.* Academic Press, N.Y., 1973.

[27] Fiagbedzi, Y.A. et A.E. Pearson: *Feedback Stabilization of Linear Autonomous Time Lag Systems.* IEEE Trans. Aut. Control, 31 :847–855, 1986.

[28] Fliess, M. et H. Mounier: *Interpretation and Comparison of Various Types of Delay System Controllabilities.* Dans *IFAC Conf. System Structure and Control*, pages 330–335, Nantes, France, 1995.

[29] Fridman, E.: *New Lyapunov-Krasovskii Functionals for Stability of Linear Retarded and Neutral Type Systems.* Syst. and Control Letters, 43(4) :309–319, July 2001.

[30] Fridman, E., A. Seuret et J.P. Richard: *Robust Sampled-Data Stabilization of Linear Systems : An Input Delay Approach.* Automatica, 40(8) :1441–1446, Aug. 2004.

[31] Fridman, E., A. Seuret et J.P. Richard: *Robust Sampled-Data Control - An input delay approach*, pages 315–327. Numéro 352 dans *Lecture Notes in Control and Inform. Sc.* Springer Verlag, Berlin Heidelberg, Chiasson and Loiseau édition, 2007.

[32] Fridman, E. et U. Shaked: *A Descriptor System Approach to H_∞ Control of Linear Time-Delay Systems.* IEEE Trans. Aut. Control, 47(2) :253–270, Feb. 2002.

[33] Fridman, E. et U. Shaked: *Delay Systems.* Special issue of Int. J. Robust and Nonlinear Control, 13(9), July 2003.

[34] Fu, M., H. Li et S.I. Niculescu: *Robust Stability and Stabilization of Time-Delay Systems via Integral Quadratic Constraint Approach*, tome 228 de *LNCIS*, chapitre 4, pages 101–116. Springer, London, 1997.

[35] Glielmo, L., S. Santini et I. Cascella: *Stability of Linear Time-Delay Systems : A Delay-Dependent Criterion with a Tight Conservatism Bound.* Dans *ACC00 (American Control Conf.)*, pages 45–49, Chicago, IL, June 2000.

[36] Glüsing-Lüerßen, H.: *A Behavioral Approach to Delay-Differential Systems.* SIAM J. Contr. Optim., 35(2) :480–499, 1997.

[37] Glüsing-Lüerßen, H.: *Realization Behaviors Given by Delay-Differential Equations.* Dans *ECC97 (4^{th} European Control Conf.)*, Brussels, Belgium, July 1997. WE-M D6.

[38] Gopalsamy, K.: *Stability and Oscillations in Delay Differential Equations of Population Dynamics*, tome 74 de *Mathematics and Applications.* Kluwer Acad., 1992.

[39] Gorecki, H., S. Fuksa, P. Grabowski et A. Korytowski: *Analysis and Synthesis of Time Delay Systems.* John Wiley and Sons, 1989.

[40] Gouaisbaut, F., W. Perruquetti et J.P. Richard: *Sliding mode control for systems with time delay*, tome Sliding mode control in engineering de *Control Eng. Series*, chapitre 11. Marcel Dekker, Perruquetti and Barbot édition, 2002.

[41] Goubet-Bartholomeus, A., M. Dambrine et J.P. Richard: *Stability of Perturbed Systems with Time-Varying Delay.* Systems and Control Letters, 31 :155–163, 1997.

[42] Gu, K.: *A Generalized Discretization Scheme of Lyapunov Functional in the Stability Problem of Linear Uncertain Time-Delay Systems.* Int. J. Robust and Nonlinear Control, 9 :1–14, 1999.

[43] Gu, K.: *Discretization Schemes for Lyapunov-Krasovskii Functionals in Time Delay Systems.* Kybernetica, 37(4) :479–504, 2001.

[44] Gu, K. et S.I. Niculescu: *Additional Dynamics in Transformed Time-Delay Systems.* Dans 38^{th} *IEEE CDC99 (Conf. on Dec. and Control)*, pages 4673–4677, Phoenix, AZ, Dec. 1999.

[45] Gu, K. et S.I. Niculescu: *Further Remarks on Additional Dynamics in Various Model Transformations of Linear Delay Systems.* IEEE Trans. Aut. Control, 46(3) :497–500, 2001.

[46] Halanay, A.: *Differential Equations : Stability, Oscillations, Time Lags.* Academic Press, New York, 1966.

[47] Hale, J.K. et S. Verduyn-Lunel: *Strong Stabilization of Neutral Functional Differential Equations.* IMA J. Math. Control Information, 19(1-2) :5–24, 2002.

[48] Hale, J.K. et S.M. Verduyn-Lunel: *Introduction to Functional Differential Equations*, tome 99 de *Applied Math. Sciences.* Springer, NY, 1993.

[49] Hirai, K. et Y. Satoh: *Stability of a System with Variable Time-Delay.* IEEE Trans. Aut. Control, 25(3) :552–554, 1980.

[50] Hotzel, R. et M. Fliess: *On Linear Systems with a Fractional Derivation : Introductory Theory and Examples.* Math. and Computers in Simulation, 45(3-4) :385–395, Feb. 1998.

[51] Huang, W.: *Generalization of Lyapunov's Theorem in a Linear Delay System.* J. Math. Anal. Appl., 142 :83–94, 1989.

[52] Huang, Y.P. et K. Zhou: *Robust Stability of Uncertain Time Delay Systems.* IEEE Trans. Aut. Control, 45(11) :2169–2173, Nov. 2000.

[53] Infante, E.F. et W.B. Castelan: *A Lyapunov Functional for a Matrix Difference-Differential Equation.* J. Diff. Equations, 29 :439–451, 1978.

[54] Ionescu, V., S.I. Niculescu, J.M. Dion, L. Dugard et H. Li: *Generalized Popov Theory Applied to State-Delayed Systems.* Automatica, 37(1) :91–97, 2001.

[55] Ionescu, V., C. Oara et M. Weiss: *Generalized Riccati Theory.* John Wiley and Sons, 1998.

[56] Izmailov, R.: *Analysis and Optimization of Feedback Control Algorithms for Data Transfers in High-Speed Networks.* SIAM J. Contr. Optimiz., 34 :1767–1780, 1996.

[57] Jalili, N. et N. Olgac: *Optimum Delayed Feedback Vibration Absorber for MDOF Mechanical Structures.* Dans 37^{th} *IEEE CDC98 (Conf. on Dec. and Control)*, pages 4734–4739, Tampa, FL, Dec. 1998.

[58] Jeong, H.S. et C.W. Lee: *Time Delay Control with State Feedback for Azimuth Motion of the Frictionless Positioning Device.* IEEE-ASME Trans. Mechatronics, 2(3), Sept. 1997.

[59] Jun, M. et M.G. Safonov: *Stability Analysis of a System with Time-Delayed States.* Dans *ACC00 (American Control Conf.)*, pages 949–952, Chicago, IL, June 2000.

[60] Kharitonov, V.: *Robust Stability Analysis of Time Delay Systems : A Survey.* Dans 4^{th}*. IFAC Conf. on System Structure and Control*, pages 1–12, Nantes, France, July 8-10 1998. Penary lecture.

[61] Kharitonov, V.L. et D. Melchior-Aguliar: *On Delay-Dependent Stability Conditions.* Syst. and Control Letters, 40(1) :71–76, May 2000.

[62] Kharitonov, V.L. et A.P. Zhabko: *Lyapunov-Krasovski Approach to Robust Stability of Time Delay Systems.* Dans 1^{rst} *IFAC/IEEE Symp. on System Structure and Control*, Prague, Tch. Rep., Aug. 2001.

[63] Kim, W.S., B. Hannaford et A.K. Bejczy: *Force-Reflection and Shared Compliant Control in Operating Telemanipulators with Time-Delay.* IEEE Trans. Robotics and Automation, 8(2) :176–185, April 1992.

[64] Kolmanovskii, V.B.: *Stability of some nonlinear functional differential equations.* J. Nonlinear Differential Equations, 2 :185–198, 1995.

[65] Kolmanovskii, V.B. et A. Myshkis: *Applied theory of functional differential equations*, tome 85 de *Mathematics and Applications.* Kluwer Acad., 1992.

[66] Kolmanovskii, V.B. et A. Myshkis: *Introduction to the theory and applications of functional differential equations.* Kluwer Acad., Dordrecht, 1999.

[67] Kolmanovskii, V.B., S.I. Niculescu et K. Gu: *Delay Effects on Stability : A Survey.* Dans 38^{th}*IEEE CDC99 (Conf. on Dec. and Control)*, pages 1993–1998, Phoenix, AZ, Dec. 1999.

[68] Kolmanovskii, V.B., S.I. Niculescu et J.P. Richard: *On the Liapunov-Krasovskii Functionals for Stability Analysis of Linear Delay Systems.* Int. J. Control, 72(4) :374–384, Feb. 1999.

[69] Kolmanovskii, V.B. et V.R. Nosov: *Stability of functional differential equations.* Academic Press, London, 1986.

[70] Kolmanovskii, V.B. et J.P. Richard: *Stability of some Linear Systems with Delay.* IEEE Transactions on Automatic Control, 44(5) :984–989, May 1999.

[71] Kolmanovskii, V.B. et J.P. Richard: *Stability of some linear systems with delay.* IEEE Trans. Aut. Control, 44(5) :984–989, May 1999.

[72] Kolmanovskii, V.B. et L.E. Shaikhet: *Control of systems with aftereffect*, tome 157 de *Transl. of Mathematical Monographs*. American Math. Soc., 1996.

[73] Kolmanovskii, V.B., P.A. Tchangani et J.P. Richard: *Stability of linear systems with discrete-plus-distributed delay : application of some model transformations*. Dans *MTNS'98 (13^{th} Symp. Math. Theory of Networks and Systems)*, Padova, Italy, July 1998.

[74] Krasovskii, N.N.: *On the analytical construction of an optimal control in a system with time lags*. Prikl. Math. Mec., 26 :39–51, 1962. (English translation : J. Appl. Math. Mech. (1962), 50-67).

[75] Krasovskii, N.N.: *Stability of Motion*. Stanford Univ. Press, 1963. (translation by J. Brenner).

[76] Kwon, H.W. et A.E. Pearson: *Feedback Stabilization of Linear Systems with Delayed Control*. IEEE Trans. Aut. Control, 25(2) :266–269, 1980.

[77] Lafay, J.F., M. Fliess, H. Mounier et O. Sename: *Sur la commandabilité des systèmes linéaires à retards*. Dans *CNRS Conf. Analysis and Control of Systems with Delays*, pages 19–42, Nantes, France, 1996. (in French).

[78] Lakshmikantham, V. et S. Leela: *Differential and Integral Inequalities*, tome 2. Academic Press, New York, 1969.

[79] Lelevé, A., P. Fraisse et P. Dauchez: *Telerobotics over IP Networks : Towards a Low Level Real Time Architecture*. Dans *IROS01 (Int. Conf. On Intelligent Robots and Systems)*, Maui, Hawaii, Oct. 2001.

[80] Lewis, R.M.: *Control-Delayed System Properties Via an Ordinary Model*. Int. J. Control, 30(3) :477–490, 1979.

[81] Loiseau, J.J.: *A 2-D Transfert Without Minimal Realization*. Dans *Sprann'94, IMACS*, pages 97–100, Lille, France, 1994.

[82] Loiseau, J.J.: *Algebraic tools for the control and stabilization of time-delay systems*. Dans 1^{rst} *IFAC Workshop on Linear Time Delay Systems*, pages 234–249, Grenoble, France, July 1998. Plenary lecture.

[83] Loiseau, J.J. et D. Bretheé: *2-D Exact Model Matching with Stability, the Structural Approach*. Bulletin of the Polish Acad. of Sc. - Technical Sciences, 45(2) :309–317, 1997.

[84] Loiseau, J.J. et R. Rabah: *Analysis and Control of Time-Delay Systems*. Special issue of JESA, European J. of Aut. Systems, 31(6), 1997.

[85] Louisell, J.: *A stability analysis for a class of differential-delay equations having time-varying delay*, tome 1475 de *Lecture Notes in Math.*, chapitre Delay differential equations and dynamical systems, pages 225–242. Springer, Busenberg and Martelli édition, 1991.

[86] Louisell, J.: *Delay Differential Systems with Time-Varying Delay : New Directions for Stability Theory.* Kybernetika, 37(3) :239–252, 2001.

[87] MacDonald, N.: *Time Lags in Biological Models*, tome 27 de *Lecture Notes in Biomath.* Springer, 1978.

[88] Manitius, A. et A.W. Olbrot: *Finite spectrum assignment problem for systems with delays.* IEEE Trans. Aut. Control, 24(4) :541–553, 1979.

[89] Mascolo, S.: *Congestion Control in High Speed Communication Networks Using the Smith Principle.* Automatica, 35 :1921–1935, 1999.

[90] Megretski, A. et A. Rantzer: *System Analysis Via Integral Quadratic Constraints.* IEEE Trans. Aut. Control, 42(6) :819–830, June 1997.

[91] Mérigot, A. et H. Mounier: *Quality of Service and MPEG4 Video Transmission.* Dans *MTNS'00 (14^{th} Symp. Math. Theory of Networks and Systems)*, Perpignan, France, June 2000.

[92] Mirkin, L. et G. Tadmor: *H_∞ Control of Systems with I/O Delay : A Review of some Problem-Oriented Methods.* IMA J. Math. Control Information, 19(1-2) :185–200, 2002.

[93] Moog, C.H., R. Castro-Linares, M. Velasco-Villa et L.A. Marquez-Martinez: *The Disturbance Decoupling Problem for Time-Delay Nonlinear Systems.* IEEE Trans. Aut. Control, 45(2) :305–309, Feb. 2000.

[94] Mounier, H. et Rudolph J.: *Flatness Based Control of Nonlinear Delay Systems : A Chemical Reactor Example.* Int. J. Control, 71 :871–890, 1998.

[95] Mounier, H., M. Mboup, N. Petit, Rouchon P. et Seret D.: *High Speed Network Congestion Control with a Simplifed Time-Varying Delay Model.* Dans *IFAC Conf. System, Structure, Control*, Nantes, France, 1998.

[96] Mounier, H., P. Rouchon et J. Rudolph: *Some examples of linear systems with delays.* JESA, European J. of Aut. Systems, 31(6) :911–926, Oct. 1997.

[97] Myshkis, A.D.: *General theory of differential equations with delay.* Uspehi Mat. Naut (N.S.), 4(33) :99–141, 1949. (in Russian), English transl. in Transl. AMS, No. 55, p. 1-62, 1951.

[98] Myshkis, A.D.: *Lineare differentialgleichungen mit nacheilendem argument Linear differential equations with delay.* VEB Deutsch. Verlag, Berlin, 1955. (original edition 1951, German transl. 1955, English transl. 1972 Nauka).

[99] Niculescu, S.I.: *Systèmes à retard : aspects qualitatifs sur la stabilité et la stabilisation.* Nouveaux Essais. Diderot Multimedia, Paris, 1997. (in French).

[100] Niculescu, S.I.: *Delay Effects on Stability*, tome 269 de *LNCIS*. Springer, 2001.

[101] Niculescu, S.I. et J. Chen: *Frequency Sweeping Tests for Asymptotic Stability : A Model Transformation for Multiple Delays*. Dans 38^{th} *IEEE CDC99 (Conf. on Dec. and Control)*, pages 4678–4683, Phoenix, USA, Dec. 1999.

[102] Niculescu, S.I. et J.P. Richard: *Analysis and design of delay and propagation systems*. Special issue of IMA J. Math. Control Information, 19(1-2) :1–227, 2002.

[103] Niculescu, S.I., E.I. Verriest, L. Dugard et J.M. Dion: *Stability and Robust Stability of Time-Delay Systems : A Guided Tour*, tome 228 de *LNCIS*, chapitre 1, pages 1–71. Springer, London, 1997.

[104] Niemeyer, G.: *Using Wave Variables Intime Delayed Force Reflecting Teleoperation*. Thèse de doctorat, MIT, Cambridge, MA, Sept. 1996.

[105] Niemeyer, G. et J.J. Slotine: *Towards Force-Reflecting Teleoperation over the Internet*. Dans *IEEE Int. Conf. on Robotics and Automation*, pages 1909–1915, Leuven, Belgium, May 1998.

[106] Nilsson, J., B. Bernhardsson et B. Wittenmark: *Stochastic Analysis and Control of Real-Time Systems with Random Delays*. Automatica, 34(1) :57–64, 1998.

[107] Olbrot, A.W.: *Algebraic Criteria of Controllability to Zero Function for Linear Constant Time-Lag Systems*. Control and Cybernetics, 2(1/2), 1973.

[108] Olbrot, A.W.: *Finite spectrum property and predictors*. Dans 1^{rst} *IFAC Workshop on Linear Time Delay Systems*, pages 251–260, Grenoble, France, July 1998. Plenary lecture.

[109] Orlov, Y.V.: *Optimal Delay Control - Part I*. Automation and Remote Control, 49(12) :1591–1596, Dec. 1989. transl. from Avtomatika i Telemekhnika, No.12, 1988.

[110] Petit, N.: *Systèmes à Retards. Platitude en Génie des Procédés et Contrôle de Certaines Équations des Ondes*. (in French), Ecole des Mines de Paris, May 2000.

[111] Picard, P. et J.F. Lafay: *Further Results on Controllability of Linear Systems with Delay*. Dans *ECC95 (*3^{rd} *European Control Conf.)*, pages 3313–3318, 1995.

[112] Picard, P., J.F. Lafay et V. Kucera: *Model Matching for Linear Systems with Delays and 2-D Systems*. Automatica,, 34(2), 1998.

[113] Picard, P., O. Sename et J.F. Lafay: *Observers and observability indices for linear systems with delays*. Dans *CESA96 (IEEE-IMACS Conf. on*

Comp. Eng. In Syst. Applic.), pages 81–86, Lille, France, July 1996. Vol. 1.

[114] Quet, P., S. Ramakrishnan, H. Ozbay et S. Kalyanaraman: *On the H_∞ Controller Design for Congestion Control in Communication Networks with a Capacity Predictor.* Dans 40^{th} *IEEE CDC01 (Conf. on Dec. and Control)*, pages 598–603, Orlando, FL, Dec. 2001.

[115] Rekasius, Z.V.: *A Stability Test for Systems with Delays.* Dans *Proc. Joint Automatic Control Conf.*, pages TP9–A, San Francisco, CA, 1980.

[116] Richard, J.P.: *Some Trends and Tools for the Study of Time Delay Systems.* Dans *2nd Conf. IMACS-IEEE CESA'98, Computational Engineering in Systems Applications*, pages 27–43, Tunisia, April 1998. Plenary lecture.

[117] Richard, J.P.: *Algèbre et analyse pour l'automatique.* Traité IC2 : Information, Commande, Communication. Hermès-Lavoisier, 2001.

[118] Richard, J.P.: *Mathématiques pour les systèmes dynamiques.* Traité IC2 : Information, Commande, Communication. Hermès-Lavoisier, 2002.

[119] Richard, J.P.: *Time Delay Systems : An Overview of some Recent Advances and Open Problems.* Automatica, 39(10) :1667–1694, Oct. 2003.

[120] Richard, J.P. et T. Divoux: *Systèmes commandés en réseau.* Traité IC2 : Information, Commande, Communication. Hermès-Lavoisier, 2007.

[121] Richard, J.P., A. Goubet, P.A. Tchangani et M. Dambrine: *Nonlinear delay systems : tools for a quantitative approach to stabilization*, tome 228 de *Lecture Notes in Control and Inform. Sc.*, chapitre 10, pages 218–240. Springer Verlag, London, Verriest and Niculescu édition, 1997.

[122] Richard, J.P. et V. Kolmanovskii: *Delay Systems.* Special issue of Mathematics and Computers in Simulation, 45(3-4), Feb. 1998.

[123] Safonov, M.G.: *Stability and Robustness of Multivariable Feedback Systems.* MIT Press, 1980.

[124] Sename, O.: *Sur la commandabilité et le découplage des systèmes linéaires à retards.* (in French), Laboratoire d'Automatique de Nantes, Univ. of Nantes and EC Nantes, France, Oct. 1994.

[125] Seuret, A.: *Commande et observation des systè à retards variables : théorie et applications.* Thèse de doctorat, Ecole Centrale de Lille, LAGIS, Oct. 4th 2006.

[126] Seuret, A., M. Dambrine et J.P. Richard: *Robust exponential stabilization for systems with time-varying delays.* Dans *TDS04, 5th IFAC Workshop on Time Delay Systems*, Leuven, Belgium, Sept. 2004.

[127] Seuret, A., E. Fridman et J.P. Richard: *Sampled-data exponential stabilization of neutral systems with input and state delays.* Dans *IEEE MED 2005, 13th Mediterranean Conf. on Control and Automation*, Cyprus, January 22-24 2005.

[128] Seuret, A., F. Michaut, J.P. Richard et T. Divoux: *Networked control using GPS synchronization.* Dans *ACCŠ06, 25th IEEE American Control Conference*, Mineapolis, Minnesota, USA, June 14-16 2006.

[129] Seuret, A., M. Termens-Ballester, A. Toguyeni, S. El Khattabi et J.P. Richard: *Implementation of an Internet-controlled system under variable delays.* Dans *ETFH'06, 11th IEEE int. conf. on Emerging Technologies & Factory Automation*, Prague, Czech Republic, Sept. 2006. Inv. Track NeCST, "Networked Control Systems Tolerant to Faults".

[130] Shakkottai, S., R.T. Srikant et S. Meyn: *Boundedness of Utility Function Based Congestion Controllers in the Presence of Delay.* Dans 40^{th} *IEEE CDC01 (Conf. on Dec. and Control)*, pages 616–621, Orlando, FL, Dec. 2001.

[131] Shin, K.G. et X. Cui: *Computing Time Delay and its Effects on Real-Time Control Systems.* IEEE Trans. Control Syst. Technol., 3(2) :218–224, June 1995.

[132] Slater, G. L. et W. R. Wells: *On the Reduction of Optimal Time Delay Systems to Ordinary Ones.* IEEE Trans. Aut. Control, 17 :154–155, 1972.

[133] Smith, O.J.M.: *A Controller to Overcome Dead Time.* ISA, J. Instrument Society of America, 6 :28–33, 1959.

[134] Sontag, E. D.: *The Lattice of Minimal Realizations of Response Maps over Rings.* Math. Systems Theory, 11 :169–175, 1977.

[135] Stépan, G.: *Retarded Dynamical Systems : Stability and Characteristic Functions*, tome 210 de *Research Notes in Math. Series.* John Wiley and Sons, 1987.

[136] Thowsen, A.: *An Analytical Stability Test for a Class of Linear Time-Delay Systems.* IEEE Trans. Aut. Control, 25 :735–736, 1981.

[137] Tits, A.L. et V. Balakrishnan: *Small-μ Theorem with Frequency-Dependent Uncertainty Bounds.* Math. Contr., Signals, Syst., 11(3) :220–243, 1998.

[138] Tsoi, A.C.: *Recent advances in the algebraic system theory of delay differential equations*, tome Recent theoretical developments in control, chapitre 5, pages 67–127. Academic Press, Gregson édition, 1978.

[139] Verriest, E.I.: *Stability of Systems with State-Dependent and Random Delays.* IMA J. Math. Control Information, 19(1-2) :103–114, 2002.

[140] Verriest, E.I. et W. Aggoune: *Stability of Nonlinear Differential Delay Systems.* Math. and Computers in Simulation, 45(3-4) :257–268, Feb. 1998.

[141] Volterra, V.: *Sulle equazioni integrodiffenziali della teorie dell' elasticita.* Atti. Accad. Lincei, 18(295), 1909.

[142] Walton, K. et J.E. Marshall: *Direct Method for TDS Stability Analysis.* IEE Proc., 134(part D) :101–107, 1987.

[143] Watanabe, K., E. Nobuyama et K. Kojima: *Recent advances in control of time-delay systems A tutorial review.* Dans 35^{th}*IEEE CDC96 (Conf. on Dec. and Control)*, pages 2083–2089, Kobe, Japan, Dec. 1996.

[144] Weiss, L.: *On the Controllability of Delay-Differential Equations.* SIAM J. Cont. Optim., 5(4) :575–587, 1967.

[145] Willems, J.: *The Analysis of Feedback Systems.* MIT Press, 1971.

[146] Willems, J.: *Paradigms and Puzzles in the Theory of Dynamical Systems.* IEEE Trans. Aut. Control, 36 :259–294, 1991.

[147] Youcef-Toumi, K. et O. Ito: *A time delay controller design for systems with unknown dynamics.* ASME J. Dynamic Systems Measurement and Control, 112 :133–142, 1990.

7 Théorie algébrique de la commande des EDPs

H. Mounier[1], J. Rudolph[2] et F. Woittennek[2]

[1]Département AXIS, Institut d'Électronique Fondamentale, Bât. 220, Université Paris-Sud, 91405 Orsay, France. *E-mail* : Hugues.Mounier@u-psud.fr
[2]Institut für Regelungs- und Steuerungstheorie, Technische Universität Dresden, 01062 Dresden, Allemagne. *E-mail* : Joachim.Rudolph@tu-dresden.de, Frank.Woittennek@tu-dresden.de

7.1 Introduction

Une théorie algébrique pour la commande des systèmes à paramètres répartis commandés aux bords est présentée dans ce chapitre. Un système y est représenté par un module et les notions de commandabilités développées étendent la notion de π-liberté dégagée précédemment pour les systèmes à retards [11].

Le point de vue ici adopté trouve ses origines dans l'extension du cadre élaboré pour les systèmes à retards [14, 1, 12, 15, 16]. Des exposés détaillés se trouvent dans [42, 43]. Diverses voies possibles d'extension au cas non linéaire sont réalisées en [26, 35, 42]. Une vue plus proche de l'analyse est développée en [24, 25].

7.2 Motivations et méthodologie

Notre philosophie est guidée par deux attentions majeures : la première (*attention pratique*) est de dégager des propriétés structurelles qui surviennent fréquemment dans les applications. La deuxième (*attention de simplicité*), reliée à la première, consiste en l'obtention des propriétés les plus simples pour chaque classe d'applications. Ceci nous a conduit à la découverte d'une nouvelle notion,

nommée π-liberté, qui permet de résoudre le suivi d'une trajectoire de référence selon le même cheminement que celui emprunté pour les systèmes non linéaires de dimension finie différentiellement plats.

7.3 Notion de liberté

Le lien entre la notion de liberté des modules et la commandabilité a d'abord été introduite pour les systèmes linéaires de dimension finie par Michel Fliess [10], puis nous l'avons étendue aux systèmes à retards [30], [11], [31], [13], [34] et aux systèmes à paramètres répartis [12], [32].

Critères classiques de commandabilité

La vision classique de la commandabilité correspond en général à une accessibilité de l'espace d'état. Considérons un système de dimension finie Σ donné sous forme d'état comme suit

$$\dot{\boldsymbol{x}}(t) = A\boldsymbol{x}(t) + B\boldsymbol{u}(t) \tag{7.1}$$

avec $\boldsymbol{x}(t) = (x_1(t), \ldots, x_n(t))$ l'état, $\boldsymbol{u}(t) = (u_1(t), \ldots, u_m(t))$ la commande, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. La *commandabilité* du système Σ peut s'exprimer de diverses manières toutes équivalentes. En voici quelques-unes :

1. *Accessibilité de l'espace d'état*
   Pour tous états initial $\boldsymbol{x}_{\text{ini}}$ et final $\boldsymbol{x}_{\text{fin}}$ de $\mathbb{R}^n$, et tous temps initial t_{ini} et final t_{fin}, il existe une commande $u \in \mathcal{F}_u$, dans l'espace dit des *commandes admissibles* $\mathcal{F}_u$, amenant le système de l'état initial $\boldsymbol{x}(t_{\text{ini}}) = \boldsymbol{x}_{\text{ini}}$ à l'état final $\boldsymbol{x}(t_{\text{fin}}) = \boldsymbol{x}_{\text{fin}}$.
2. *Critère de Kalman*
   La matrice de commandabilité
   $$\mathscr{C} = (b, Ab, \ldots, A^{n-1}b)$$
   est de rang plein :
   $$\text{rg}\mathscr{C} = n$$
3. *Critère de Hautus-Popov-Belevitch*
   $$\forall s \in \mathbb{C}, \qquad \text{rg}_{\mathbb{C}}[sI - A \mid b] = n$$
4. *Forme canonique contrôleur*
   On se restreint dans cet alinéa au cas de systèmes mono-entrée. Il existe un changement d'état $\boldsymbol{z} = F\boldsymbol{x}$, $F \in \mathbb{R}^{n\times n}$ tel que la représentation de Σ

devienne :

$$\begin{aligned}
\dot{z}_1 &= z_2\\
\dot{z}_2 &= z_3\\
&\vdots\\
\dot{z}_{n-1} &= z_n\\
\dot{z}_n &= \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n + \beta u
\end{aligned}$$

Si, au lieu de vouloir imposer que l'état $\boldsymbol{z}$ aille de $\boldsymbol{z}_{\text{ini}}$ à $\boldsymbol{z}_{\text{fin}}$, nous voulons amener une sortie y d'une valeur initiale à une valeur finale (sachant que le nombre de commandes est en général bien moindre que celui des états), nous pouvons avoir un bien meilleur contrôle sur l'évolution temporelle de y. Considérons z_1 comme sortie de la forme canonique contrôleur :

$$\left\{\begin{aligned}
\dot{z}_1 &= z_2\\
\dot{z}_2 &= z_3\\
&\vdots\\
\dot{z}_{n-1} &= z_n\\
\dot{z}_n &= \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_n z_n + \beta u\\
y &= z_1
\end{aligned}\right.$$

Alors, connaissant y, toute autre variable est déterminée

$$\begin{aligned}
z_1 &= y\\
z_2 &= \dot{z}_1\\
z_3 &= \ddot{z}_1\\
&\vdots\\
z_n &= z_1^{(n-1)}\\
u &= -\frac{1}{\beta}\Big(\alpha_1 z_1 + \alpha_2 \dot{z}_1 + \cdots + \alpha_n z_1^{(n-1)} - z_1^{(n)}\Big)
\end{aligned}$$

Le modèle est alors différentiellement paramétrisé par y. En particulier, si l'on désire suivre une trajectoire $t \mapsto y_r(t)$ de y, la loi de commande en boucle ouverte assurant le suivi est donnée par

$$u_r(t) = -\frac{1}{\beta}\Big(\alpha_1 y_r(t) + \cdots + \alpha_n y_r^{(n-1)}(t) - y_r^{(n)}(t)\Big)$$

Remarques 7.3.1. 1. L'expression donnant u_r (de même que celles donnant les autres variables) ne recquière aucune intégration d'équation différentielle.

2. Le concepteur de la commande est libre de choisir la trajectoire de référence qu'il souhaite, pour autant qu'elle soit différentiable jusqu'à l'ordre n.
3. Cette étape en boucle ouverte, faisant une confiance aveugle en le modèle, suppose ce dernier parfait et suppose également que les conditions initiales soient parfaitement connues. Elle est en pratique complétée par une étape de stabilisation.
4. La fonction de transfert correspondante a un numérateur constant :

$$\frac{\hat{y}}{\hat{u}} = \frac{\beta}{s^n - \alpha_n s^{n-1} - \cdots - \alpha_1}$$

Exemple 7.3.1. Le modèle

$$M\ddot{y} + Ky = u$$

s'écrit sous forme canonique contrôleur

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{1}{M}\left(-Kx_1 + u\right) \\ y &= x_1 \end{aligned}$$

Nous avons donc :

$$\begin{aligned} x_1 &= y \\ x_2 &= \dot{y} \\ u &= Ky + M\ddot{y} \end{aligned}$$

Et la commande en boucle ouverte assurant le suivi de $t \mapsto y_r(t)$ est

$$u_r = Ky_r + M\ddot{y}_r$$

Système sans retard libre

Nous allons généraliser ici de manière d'abord informelle, puis en termes précis, la forme canonique contrôleur précédente. La notion obtenue, que nous nommerons liberté, conserve cette essentielle propriété de paramétrisation.

Un modèle général tel que (7.1) de dimension finie muni d'une entrée indépendante $\boldsymbol{u} = (u_1, \ldots, u_m)$ est dit *libre* (voir [10] et l'annexe 7.A pour une définition précise) s'il existe

$$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$$

que l'on nomme *base* (ou sortie basique), telle que :

1. Elle fait partie du système (caractère *endogène*) :
   Les ω_j, $j = 0, \ldots, m$ s'expriment comme combinaisons linéaires des variables du système (par ex. $\boldsymbol{x}, \boldsymbol{u}$) et de leurs dérivées :
   $$\omega_i = L\boldsymbol{x} + N_0\boldsymbol{u} + N_1 \tfrac{d}{dt}\boldsymbol{u} + \cdots + N_\alpha \tfrac{d^\alpha}{dt^\alpha}\boldsymbol{u}$$
   $L \in \mathbb{R}^{n\times n}$, $N_i \in \mathbb{R}^{n\times m}$, $i = 0, \ldots, \alpha$.
2. Ses composantes sont différentiellement indépendantes (*indépendance*) :
   Il n'existe aucune relation différentielle entre les ω_j, $j = 0, \ldots, m$:
   $$m_0\boldsymbol{\omega} + m_1 \tfrac{d}{dt}\boldsymbol{\omega} + \cdots + m_\beta \tfrac{d^\beta}{dt^\beta}\boldsymbol{\omega} = 0 \quad \Longrightarrow \quad m_i = 0$$
   $m_i \in \mathbb{R}^{1\times m}$, $i = 0, \ldots, \beta$.
3. Elle fournit une paramétrisation complète du système (*forme canonique de suivi*) :
   $$\boldsymbol{x} = P_0\boldsymbol{\omega} + P_1 \tfrac{d}{dt}\boldsymbol{\omega} + \ldots + P_\gamma \tfrac{d^\gamma}{dt^\gamma}\boldsymbol{\omega} \tag{7.3}$$
   $$\boldsymbol{u} = Q_0\boldsymbol{\omega} + Q_1 \tfrac{d}{dt}\boldsymbol{\omega} + \cdots + Q_\mu \tfrac{d^\mu}{dt^\mu}\boldsymbol{\omega} \tag{7.4}$$
   $P_i \in \mathbb{R}^{n\times m}$, $Q_j \in \mathbb{R}^{m\times m}$, $i = 0, \ldots, \gamma$, $j = 0, \ldots, \mu$,

La dernière propriété, et plus spécialement l'équation (7.4), fournit une solution simple et naturelle à la réalisation du suivi en boucle ouverte d'une trajectoire $t \mapsto \boldsymbol{\omega}(t)$.

Remarque 7.3.1. Ces propriétés (indépendance et paramétrisation du système) sont en correspondance directe avec les propriétés caractéristiques d'une base d'un espace vectoriel (ensemble maximalement indépendant et minimalement générateur).

Exemple de système libre

Considérons un modèle du premier mode souple d'un bras de robot à un degré de liberté, actionné par un moteur. Notons q_r le déplacement rigide, q_e le premier mode du déplacement élastique (par rapport au déplacement rigide) et u le couple moteur actionnant le bras.

Le bilan des couples donne :

$$\begin{pmatrix} J_{rr} & J_{re} \\ J_{er} & J_{ee} \end{pmatrix} \begin{pmatrix} \ddot{q}_r \\ \ddot{q}_e \end{pmatrix} + \begin{pmatrix} 0 \\ K_e q_e \end{pmatrix} = \begin{pmatrix} u \\ 0 \end{pmatrix}$$

où J_{xy} sont des inerties équivalentes et K_e une raideur élastique. Les équations du système sont :

$$J_{rr}\ddot{q}_r + J_{re}\ddot{q}_e = u \tag{7.5a}$$
$$J_{er}\ddot{q}_r + J_{ee}\ddot{q}_e = -K_e q_e \tag{7.5b}$$

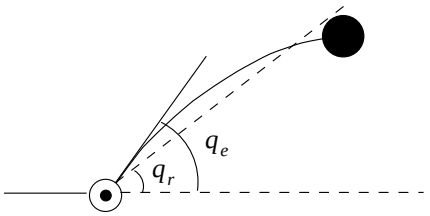


Fig. 7.1: Premier mode souple d'un bras de robot.

Ce système est libre, de base

$$\omega = J_{er} q_r + J_{ee} q_e$$

En effet, le premier membre de l'équation (7.5b) est la dérivée seconde de ω :

$$\ddot{\omega} = -K_e q_e \quad \text{ou encore} \quad q_e = -\frac{1}{K_e}\,\ddot{\omega}$$

puis, nous obtenons

$$q_r = \frac{1}{J_{er}}\,(\omega - J_{ee} q_e)$$

Et, finalement

$$q_r = \frac{1}{J_{er}}\,\omega + \frac{J_{ee}}{K_e J_{er}}\,\ddot{\omega} \tag{7.6a}$$

$$q_e = -\frac{1}{K_e}\,\ddot{\omega} \tag{7.6b}$$

$$u = \frac{J_{rr}}{J_{er}}\,\ddot{\omega} + \frac{1}{K_e}\left(\frac{J_{rr} J_{ee}}{J_{er}} - J_{re}\right)\omega^{(4)} \tag{7.6c}$$

Se donnant une trajectoire de référence $t \mapsto \omega_r(t)$, la dernière formule nous fournit une loi en boucle ouverte assurant un suivi exact sous l'hypothèse d'un modèle parfait et de conditions initiales connues. Pour une trajectoire désirée d'allure décrite en figure 7.2 à gauche, nous obtenons la loi en boucle ouverte illustrée à droite.

Les systèmes linéaires sont, dans notre approche, modélisés par des structures linéaires analogues aux espaces vectoriels : des modules (diverses notions d'algèbre sont rappelées en annexe 7.A, p. 324). Les axiomes de définition sont les mêmes pour les deux structures, mais les scalaires d'un espace vectoriel sont pris dans un corps, tel $\mathbb{R}$ ou $\mathbb{C}$, alors que ceux d'un module sont pris dans un anneau. Pour les systèmes sans retards, cet anneau sera celui des polynômes différentiels $\mathbb{R}[\frac{d}{dt}]$; pour les systèmes à retards, l'anneau sera $\mathbb{R}[\frac{d}{dt}, \boldsymbol{\delta}]$ où $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r)$, chaque δ_i étant un opérateur retard tel que, pour tout f, $(\delta_i f)(t) = f(t - h_i)$,

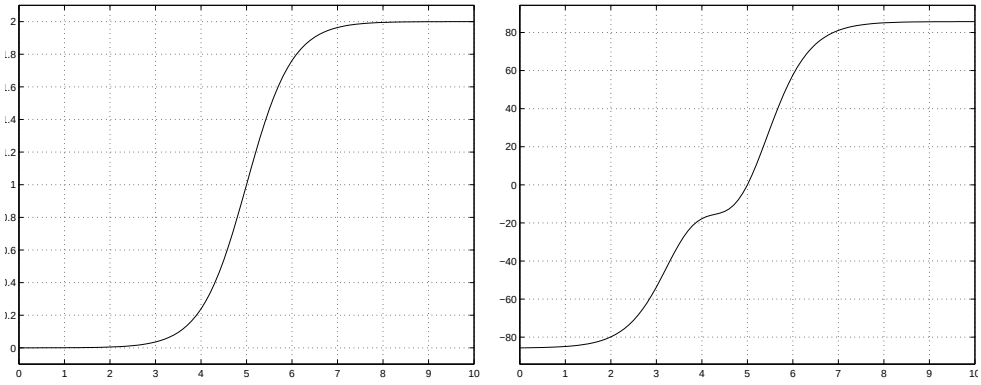


FIG. 7.2: Trajectoire de référence $t \mapsto \omega_r(t)$ (à gauche) et commande en boucle ouverte u_r (à droite).

où h_i est un réel positif, l'amplitude du retard ; pour des systèmes modélisés par des EDPs, les anneaux seront adaptés au système étudié.

Cette augmentation du nombre d'indéterminées dans l'anneau des opérateurs entraîne une bien plus grande complexité des modules associés, ce qui se retrouve dans les propriétés de commandabilité.

R-système, entrée

Définitions 7.3.1. Soit R un anneau commutatif[1], unifère et sans diviseurs de zéros.

- Un *R-système* Λ, ou un *système sur R* est un R-module.
- Une *R-dynamique*, ou une *dynamique sur R*, est un R-système Λ équippé d'une *entrée*, c.à.d., un sous-ensemble $\boldsymbol{u}$ de Λ tel que le R-module quotient $\Lambda/[\boldsymbol{u}]$ est de torsion.
- L'entrée $\boldsymbol{u}$ est *indépendante* si le R-module $[\boldsymbol{u}]$ est libre, de base $\boldsymbol{u}$.
- Une *sortie* $\boldsymbol{y}$ est un sous ensemble de Λ.
- Soit A une R-algèbre et Λ un R-système. Le A-module $A \otimes_R \Lambda$ est un A-système, qui *étend* Λ.

Remarques 7.3.2. 1. Pour un système de dimension finie, $R = k[\frac{d}{dt}]$ avec $k = \mathbb{R}$ ou $\mathbb{C}$. Pour un système à retards, $R = k[\frac{d}{dt}, \boldsymbol{\delta}]$, avec $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r)$, les δ_i étant des opérateurs retards algébriquement indépendants (c.à.d. les amplitudes des différents retards sont indépendantes sur $\mathbb{Q}$). On parle également d'incommensurabilité dans la littérature.

[1] On se restreint ici au cas commutatif. Pour le cas non commutatif, voir [9, 41] pour certaines extensions.

2. En termes informels, un système d'équation

$$a\,y = b\,u$$

où $a, b \in R = \mathbb{R}[\frac{d}{dt}]$ ou $\mathbb{R}[\frac{d}{dt}, \delta]$ sera représenté par une structure linéaire formée de toutes les combinaisons linéaires de y, u et de leurs dérivées, telles que la relation ci-dessus soit satisfaite

$$\Lambda = \left\{ p\,y + q\,u \;\middle|\; p, q \in R,\ a\,x = b\,u \right\}$$

3. Dans les mêmes termes informels, une entrée $\boldsymbol{u} = (u_1, \ldots, u_m)$ est telle que toute variable z du système satisfait une relation de la forme :

$$p\,z = \sum_{i=1}^{m} q_i\,u_i$$

où p, $q_i \in R$, $p \neq 0$.

Cas des systèmes à retards

Définition 7.3.1. Un *système linéaire stationnaire à retards* Λ est un $k[\frac{d}{dt}, \boldsymbol{\delta}]$-module finiment engendré.

Cette définition appelle quelques commentaires. Tout d'abord, un module finiment engendré sur $k[\frac{d}{dt}, \boldsymbol{\delta}]$ est entièrement déterminé par la donnée de générateurs et d'un ensemble de relations (ce dernier constituant lui-même un module) vérifiées par ces derniers. Le module est alors dit donné par *générateurs et relations* (voir l'annexe 7.A, p. 325 ou [8, 0.3 p. 17] pour une définition précise). Nous disposons donc de générateurs, $\Lambda = [w_1, \ldots, w_\alpha] = [\boldsymbol{w}]$ et de relations,

$$P_\Lambda(\tfrac{d}{dt}, \boldsymbol{\delta})\,\boldsymbol{w} = 0$$

où $P_\Lambda \in k[\frac{d}{dt}, \boldsymbol{\delta}]^{\beta\times\alpha}$ et $\mathrm{rg}_{k[\frac{d}{dt},\boldsymbol{\delta}]} P_\Lambda = \beta$.

Ici P_Λ se nomme une *matrice de présentation*[2] de Λ. Remarquons que cette matrice n'est pas unique.

Exemple 7.3.2. Prenons comme exemple le système d'équation

$$\dot{y}(t) - y(t) = u(t-1)$$

Soient $[U, Y]$ le $k[\frac{d}{dt}, \delta]$-module libre engendré par U et Y et $[E] = [\dot{Y} - Y - \delta U]$ un sous-module libre de $[U, Y]$. Posons alors $[u, y] = [U, Y]/[E]$; le système considéré est représenté par le module

$$\Lambda = [u, y]$$

[2] De manière rigoureuse, la matrice de présentation est P_Λ^T. Nous conserverons dans la suite cet abus de langage commode.

avec comme relation

$$\dot{y} - y = \delta u$$

qui peut alors s'écrire :

$$\left[-\delta \mid \tfrac{d}{dt} - 1 \right] \begin{bmatrix} u \\ y \end{bmatrix} = 0$$

avec comme matrice de présentation

$$P_\Lambda(\tfrac{d}{dt}, \boldsymbol{\delta}) = \left[-\delta \mid \tfrac{d}{dt} - 1 \right]$$

7.4 Notions de commandabilité

Sans torsion, liberté, projectivité

Explicitons une notion importante pour l'étude de la commandabilité : celle de torsion (voir également l'annexe 7.A, p. 324). Considérons un système linéaire (sans retards) Σ, modélisé par un module sur l'anneau $\mathbb{R}[\frac{d}{dt}]$. Un élément w de Σ est dit *de torsion* s'il existe un polynôme non nul p de $\mathbb{R}[\frac{d}{dt}]$ tel que

$$p(\tfrac{d}{dt})\, w = 0$$

Notons que ce phénomène ne peut se produire de manière non triviale dans un espace vectoriel ; en effet une relation de la forme $pw = 0$ dans un espace vectoriel implique $w = 0$, puisque p est inversible en tant qu'élément d'un corps. Un élément de torsion de Σ satisfait une équation différentielle à coefficients dans $\mathbb{R}$. Un module dont tous les éléments sont de torsion est dit *de torsion*. A l'inverse, un module dont aucun élément n'est de torsion est dit *sans torsion*. L'abscence de torsion signifie qu'il n'y a pas de variable satisfaisant à une équation (différentielle, différentielle aux différences, ..., selon l'anneau R) autonome, c.à.d. non influencée par l'entrée.

La liberté, notion en général plus forte que la précédente, signifie qu'il existe m (où m désigne le nombre d'entrées indépendantes) générateurs indépendants, fournissant une paramétrisation complète du système.

Une notion supplémentaire est celle de projectivité (voir l'annexe 7.A, p. 324).

Définition 7.4.1. Un R-système Λ est dit *R-commandable sans torsion* (resp. *R-commandable projectif*, *R-commandable libre*) si le R-module Λ est sans torsion (resp. projectif, libre).

Quelques considérations classiques d'algèbre homologique (voir, par exemple, [39]) conduisent au résultat suivant.

Proposition 7.4.1. *La R-commandabilité libre (resp. R-commandabilité projective) implique la R-commandabilité projective (resp. R-commandabilité sans torsion).*

Remarque 7.4.1. Pour des systèmes de dimension finie (cas où $R = k[\frac{d}{dt}]$, $k = \mathbb{R}$ ou $\mathbb{C}$), les trois notions précédentes coïncident. Ce n'est plus le cas généralement, sauf si R est un anneau principal (dans lequel tout idéal est principal, i.e. généré par un unique élément ; c'est le cas de $k[\frac{d}{dt}]$) ou s'il est un anneau de Bézout (dans lequel tout idéal finiment engendré est principal).

Critères de liberté et d'absence de torsion sur un anneau polynomial

Nous considérerons, tout au long de cette section un anneau R d'opérateurs qui est un anneau de polynômes du type $k[\boldsymbol{\xi}]$ où $\boldsymbol{\xi} = \{\xi_1, \ldots, \xi_r\}$ et $k = \mathbb{R}$ ou $\mathbb{C}$.

Critère de liberté

Notons que, sur $k[\boldsymbol{\xi}]$, les deux notions de commandabilité projective et commandabilité libre coïncident ; ceci constitue l'un des énoncés de la résolution de la "conjecture de Serre" ([36], [45]).

Proposition 7.4.2 (Quillen, Suslin)**.** *Un $k[\boldsymbol{\xi}]$-système est $k[\boldsymbol{\xi}]$-commandable libre si, et seulement s'il est $k[\boldsymbol{\xi}]$-commandable projectif.*

Cette proposition permet d'obtenir le critère suivant, où $\bar{k}$ désigne la *clôture algébrique*[3] de k :

Proposition 7.4.3. *Un $k[\boldsymbol{\xi}]$-système Λ est $k[\boldsymbol{\xi}]$-commandable libre si, et seulement si*

$$\forall (s_1, \ldots, s_r) \in \bar{k}^r, \quad rg_{\bar{k}} P_\Lambda(s_1, \ldots, s_r) = \beta$$

Ici β désigne le rang générique de la matrice de présentation P_Λ, c.à.d. $rg_{k[\boldsymbol{\xi}]} P_\Lambda$. Ce critère de rang équivaut à l'absence de zéros communs dans $\bar{k}^r$ des mineurs d'ordre β de P_Λ.

Démonstration. Indiquons simplement comment déduire le présent critère du résultat de Quillen et Suslin (proposition précédente). Il y a équivalence, pour le $k[\boldsymbol{\xi}]$-module Λ, entre projectivité et égalité à $k[\boldsymbol{\xi}]$ de l'idéal de Fitting $\mathfrak{I}$ engendré par les mineurs d'ordre β de P_Λ (voir [8, proposition 20.8 p. 495])

D'après le Nullstellensatz de Hilbert (voir par exemple [27] ou [8, corollaire 1.7 p. 34]) $\mathfrak{I} = k[\boldsymbol{\xi}]$ équivaut à ne pas avoir de zéros communs dans $\bar{k}$ entre les mineurs d'ordre β de P_Λ, ou encore :

$$\forall (s_1, \ldots, s_r) \in \bar{k}^r, \quad \mathrm{rg}_{\bar{k}} P_\Lambda(s_1, \ldots, s_r) = \beta$$

c'est-à-dire aucune chute de rang de la matrice de présentation, ce rang étant partout égal au rang générique $\mathrm{rg}_{k[\boldsymbol{\xi}]} P_\Lambda$. □

[3] Le corps contenant toutes les racines d'équations polynomiales à coefficients dans k.

Critère d'absence de torsion

Nous allons maintenant énoncer un critère de $k[\boldsymbol{\xi}]$-commandabilité sans torsion.

Proposition 7.4.4. *Un $k[\boldsymbol{\xi}]$-système Λ est $k[\boldsymbol{\xi}]$-commandable sans torsion si, et seulement si les mineurs $\beta \times \beta$ de P_Λ sont premiers*[4] *entre eux.*

Démonstration. Ce critère découle directement d'une proposition de [48]. Elle établit que les mineurs $\beta \times \beta$ de P_Λ sont premiers entre eux si, et seulement si pour tout i dans $\{0, \ldots, r\}$, il existe des matrices Q_i à coefficients dans $k[\boldsymbol{\xi}]$ telles que :

$$P_\Lambda Q_i = \psi_i I_\beta$$

où ψ_i est un polynôme de $k[\boldsymbol{\xi}]$ indépendant de la variable d'indice i.

L'équation $P_\Lambda Q_0 = \psi_0 I_\beta$ exprime la liberté de $\Lambda_0 \triangleq k(\xi_2, \ldots, \xi_r)[\xi_1] \otimes_{k[\boldsymbol{\xi}]} \Lambda$, l'équation $P_\Lambda Q_1 = \psi_1 I_\beta$ celle de $\Lambda_1 \triangleq k(\xi_1, \xi_3, \ldots, \xi_r)[\xi_2] \otimes_{k[\boldsymbol{\xi}]} \Lambda$ et ainsi de suite jusqu'à $P_\Lambda Q_r = \psi_r I_\beta$ exprimant la liberté de $\Lambda_r \triangleq k(\xi_1, \ldots, \xi_{r-1})[\xi_r] \otimes_{k[\boldsymbol{\xi}]} \Lambda$. Tous les anneaux qui viennent d'être cités étant principaux, la liberté des Λ_i équivaut à leur caractère sans torsion. Et de manière évidente, Λ est sans torsion si, et seulement si chacun des Λ_i l'est. Plus précisément, lorsqu'un élément w d'un $k[\boldsymbol{\xi}]$-module est de torsion, il existe un polynôme p de $k[\boldsymbol{\xi}]$ non nul et non inversible (c'est-à-dire non constant), tel que $p(\boldsymbol{\xi})\, w = 0$. Or p peut-être considéré comme un polynôme en ξ_1 à coefficients dans $k[\xi_2, \ldots, \xi_r]$, comme polynôme en ξ_2 à coefficients dans $k[\xi_1, \xi_3, \ldots, \xi_r]$ et ainsi de suite ; puisque p n'est pas constant, il sera nécessairement non inversible dans l'un des anneaux $k(\xi_2, \ldots, \xi_r)[\xi_1]$, $k(\xi_1, \xi_3, \ldots, \xi_r)[\xi_2]$, $\ldots$, $k(\xi_1, \ldots, \xi_{r-1})[\xi_r]$. □

π-liberté

Afin de recouvrer les avantages de la liberté pour un système qui n'est que sans torsion, nous disposons du résultat suivant, qui découle directement d'une proposition de [40].

Thèorème et définition 7.4.1. *Soit R un anneau, M un R-module finiment présenté et $\mathcal{S}$ une partie multiplicative de R. Supposons que $\mathcal{S}^{-1}M$ (le localisé de M en $\mathcal{S}$) soit un $\mathcal{S}^{-1}R$-module libre. Alors il existe un π dans S tel que $R[\pi^{-1}] \otimes_R M$ est*[5] *un $R[\pi^{-1}]$-module libre avec une base de même taille que celle de $\mathcal{S}^{-1}M$ sur $\mathcal{S}^{-1}R$.*

Un tel R-système est dit π-libre *(ou* π-commandable libre*) et toute base du module $R[\pi^{-1}] \otimes_R M$ est nommée une* π-base.

[4] Premiers entre eux signifiant que leur PGCD est un élément de k.

[5] La notation $R[\pi^{-1}]$ désigne le localisé de R en $\widetilde{\mathcal{S}} = \{\pi^i \mid i \in \mathbb{N}\}$, qui se note de manière précise $\widetilde{\mathcal{S}}^{-1}R = \{\pi^i \mid i \in \mathbb{N}\}^{-1}R$.

Exemples 7.4.1. 1. Le système $\dot{y} = (1+\delta)u$ n'est pas $\mathbb{R}[\frac{d}{dt}, \delta]$-commandable libre, mais il est $(1+\delta)$-libre. En effet, on a $u = (1+\delta^{-1})\frac{d}{dt}\, y$

2. Le système d'équation $\dot{y} = \delta u$ n'est pas $\mathbb{R}[\frac{d}{dt}, \delta]$-commandable libre, mais il est δ-libre.

Cas des systèmes à retards

L'unicité de la commandabilité de Kalman en dimension finie est perdue dans le cas avec retards. De multiples généralisations ont vu le jour dans la littérature, sans que les liens soient toujours explicites. Une explication possible de ce phénomène est la complexification de l'anneau des opérateurs R.

Pour les systèmes de dimension finie, c.à.d. sur $\mathbb{R}[\frac{d}{dt}]$, l'absence de torsion et la liberté se confondent. Pour les systèmes à retards, c.à.d. sur $\mathbb{R}[\frac{d}{dt}, \boldsymbol{\delta}]$, la situation est plus complexe, la liberté étant en particulier plus forte que le caractère sans torsion.

Exemples 7.4.2. 1. Le système

$$\begin{aligned} \dot{x}_1 &= \delta x_1 \\ \dot{x}_2 &= x_1 + \delta u \end{aligned}$$

n'est pas $\mathbb{R}[\frac{d}{dt}, \delta]$-sans torsion, puisque x_1 est un élément de torsion.

2. Un exemple moins trivial est le suivant :

$$\begin{aligned} \dot{x}_1 &= x_2 + u \\ \dot{x}_2 &= \delta^2 x_1 + \delta u \end{aligned}$$

n'est pas $\mathbb{R}[\frac{d}{dt}, \delta]$-sans torsion, puisque $z = \delta x_1 - x_2$ est un élément de torsion. Pour s'en convaincre, il suffit d'appliquer l'opérateur de retard δ à la première équation et de retrancher ce résultat de la deuxième (de façon à éliminer u). On obtien ainsi :

$$\frac{d}{dt}(\delta x_1 - x_2) = \delta(x_2 - \delta x_1)$$

3. Le système $\dot{y} = (1+\delta)u$ est $\mathbb{R}[\frac{d}{dt}, \delta]$-commandable sans torsion, mais non $\mathbb{R}[\frac{d}{dt}, \delta]$-commandable libre.

4. Le système $\dot{y} + \delta y = u$ est $\mathbb{R}[\frac{d}{dt}, \delta]$-commandable libre, de base y.

7.5 Des systèmes à retards aux systèmes à paramètres répartis

Exemple d'une équation des ondes

Un exemple de système régi par une équation aux dérivées partielles est celui d'une barre flexible en torsion avec un couple appliqué à l'une de ses extrémités

(voir [33]), et une charge J attachée à l'autre. Les équations du modèle suivent celles d'une équation des ondes unidimensionnelle. La commande est le couple appliqué à l'une des extrémités, un couple de réaction dû à la charge agissant à l'autre :

$$\sigma^2\partial_\tau^2 q(\tau,z) = \partial_z^2 q(\tau,z) \tag{7.7a}$$
$$\partial_z q(\tau,0) = -u(\tau), \qquad \partial_z q(\tau,L) = -J\partial_\tau^2 q(\tau,L) \tag{7.7b}$$
$$q(0,z) = 0, \qquad \partial_\tau q(0,z) = 0 \tag{7.7c}$$

Ici, $q(\tau,z)$ désigne le déplacement angulaire de la position au repos au point $z \in [0,L]$ et au temps $\tau \geqslant 0$; L est la longueur de la barre, σ est l'inverse de la vitesse de propagation des ondes, J le moment d'inertie de la charge, u le couple de commande.

Modèle à retards

Il est bien connu (voir [6]), que la solution générale de (7.7a) peut s'écrire sous la forme d'ondes progressive et rétrograde

$$q(\tau,z) = \phi(\tau+\sigma z) + \psi(\tau-\sigma z)$$

où ϕ et ψ désignent des fonctions d'une variable réelle arbitraire. Posant $\xi = \tau+\sigma z$, $\eta = \tau-\sigma z$, les conditions aux bords (7.7b) s'écrivent

$$\phi'(\tau) - \psi'(\tau) = -\frac{u(\tau)}{\sigma}$$
$$\phi'(\tau+T) - \psi'(\tau-T) = -\frac{J}{\sigma}\Big(\phi''(\tau+T) - \psi''(\tau-T)\Big)$$

où $T = \sigma L$. L'objectif de commande sera d'assigner une trajectoire $\tau \mapsto y(\tau) = q(\tau,L)$ (prescrite à l'avance) à la position angulaire de la charge ; la sortie est donc

$$y(\tau) = q(\tau,L)$$

Alors, considérant ϕ et ψ comme des fonctions du temps τ, on obtient

$$\phi(\tau+T) + \psi(\tau-T) = y(\tau) \tag{7.8a}$$
$$\phi'(\tau) - \psi'(\tau) = -\frac{1}{\sigma}u(\tau) \tag{7.8b}$$
$$\phi'(\tau+T) - \psi'(\tau-T) = -\frac{J}{\sigma}y''(\tau) \tag{7.8c}$$

D'après (7.8a) il vient $\phi(\tau) = y(\tau-T) - \psi(\tau-2T)$ ce qui conduit à

$$y'(\tau-T) - \psi'(\tau) - \psi'(\tau-2T) = -\frac{1}{\sigma}u(\tau)$$
$$y'(\tau) - 2\psi'(\tau-T) = -\frac{J}{\sigma}y''(\tau)$$

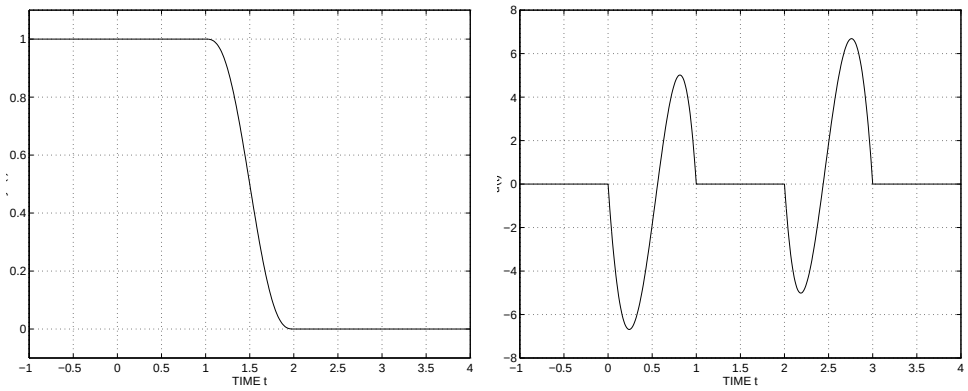


Fig. 7.3: Trajectoire de référence $t \mapsto y_r(t)$ (à gauche) et commande en boucle ouverte u_r.

L'élimination de ψ' donne

$$\frac{2}{\sigma}u(\tau - T) = \frac{J}{\sigma}\big(y''(\tau) + y''(\tau - 2T)\big) + y'(\tau) - y'(\tau - 2T)$$

Opérant le changement de temps $t = \kappa\tau$, avec κ constant, on obtient $y'(\tau) = \frac{dy}{d\tau}(\tau) = \kappa\frac{dy}{dt}(t) = \kappa\dot{y}(t)$ et

$$\frac{2}{\sigma\kappa}u(t - T) = \frac{J\kappa}{\sigma}\big(\ddot{y}(t) + \ddot{y}(t - 2T)\big) + \dot{y}(t) - \dot{y}(t - 2T) \tag{7.9}$$

Puis, posant $\kappa = \sigma/J$ et $v(t) = (2/\sigma\kappa)u(t)$, on obtient le système à retards suivant :

$$\ddot{y}(t) + \ddot{y}(t - 2T) + \dot{y}(t) - \dot{y}(t - 2T) = v(t - T) \tag{7.10}$$

Nous voyons sur l'équation ci-dessus qu'il suffit d'inverser l'opérateur de retard (c'est-à-dire s'autoriser des avances) pour obtenir un système libre :

$$v(t) = \ddot{y}(t + T) + \ddot{y}(t - T) + \dot{y}(t + T) - \dot{y}(t - T)$$

Ce système est δ_T-libre, de δ_T-base y (où δ_T, l'opérateur retard d'amplitude T est celui qu'on a inversé). On écrit ceci de la manière formelle suivante :

$$v = (\delta_T^{-1} + \delta_T)\ddot{y} + (\delta_T^{-1} - \delta_T)\dot{y}$$

Pour une trajectoire de référence $t \mapsto y_r(t)$ de la forme décrite en figure 7.3 à gauche, on obtient une loi de commande en boucle ouverte $u_r(t)$ reproduite en figure 7.3 à droite.

Il est intéressant de remarquer que le déplacement des autres points de la barre s'obtient par

$$q_r(z,t) = \frac{1}{2}\left[y_r(t - z + T) + \dot{y}_r(t - z + T) + y_r(t - T + z) - \dot{y}_r(t - T + z)\right]$$

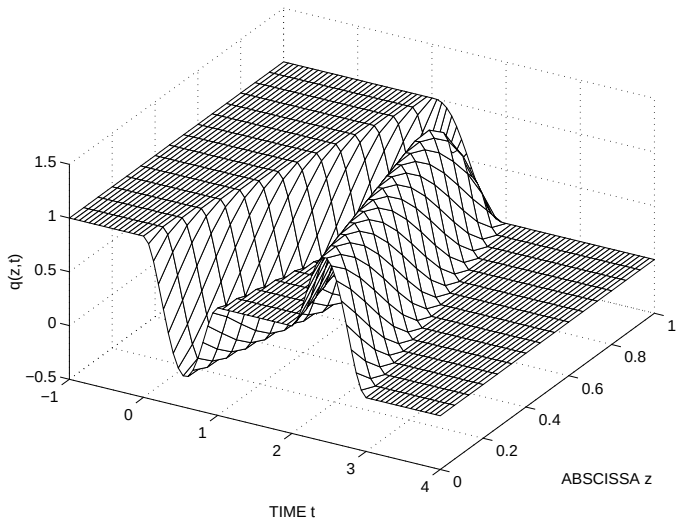


FIG. 7.4: Déplacements en boucle ouverte $q_r(z,t)$.

ou encore, formellement

$$q_r = \frac{1}{2}\left[\delta_{z-T}\, y_r + \delta_{z-T}\, \dot{y}_r + \delta_{T-z}\, y_r - \delta_{T-z}\, \dot{y}_r\right]$$

Ces déplacements sont reproduits en figure 7.4.

7.6 Généralisation de paramétrisation à d'autres systèmes à paramètres répartis : exemple de l'équation de la chaleur

Un système à paramètres répartis est composé d'une (ou plusieurs) équation(s) reliant les dérivées d'espace à celle de temps, ainsi que diverses conditions aux bords, faisant généralement intervenir les commandes. Nous verrons, comme dans le cas de la section précédente, qu'il est bon d'élargir la notion de paramétrisation reliée à la liberté.

Nous considérerons d'abord l'un des exemples mathématiquement les plus simples, l'équation de la chaleur unidimensionnelle, ce qui permettra de ne pas alourdir l'exposé par des calculs fastidieux.

Liberté pour l'équation de la chaleur

Considérons le système d'équations suivant :

$$\partial_x^2 w(x,t) = \partial_t w(x,t), \quad x \in [0,1],\ t \in [0,+\infty[\tag{7.11a}$$
$$\partial_x w(0,t) = 0 \tag{7.11b}$$
$$w(1,t) = u(t) \tag{7.11c}$$

Ces équations décrivent un processus de diffusion de chaleur dans une barre de longueur unité, $w(x,t)$ désignant la température de la barre à l'abscisse x et à l'instant t. La première des conditions aux limites indique qu'il n'y a pas de flux de chaleur en $x = 0$; la deuxième que la température est fixée par la commande $u(t)$ en $x = 1$.

Deux cadres opérationnels sont notamment possibles, développés respectivement par Mikusiński et par Komatsu. Mikusiński [29] utilise des anneaux de fonctions continues munis du produit de convolution ; Komatsu [22] se sert de transformées de Laplace d'hyperfonctions et d'ultradistributions Gevrey. Ces cadres sont reliés aux techniques de resommation développées par Écalle [7], Ramis [37], Balser [3], [4], ...

Perspective symbolique

Obtention de premières relations. Considérons la transformée de Laplace temporelle, ou bien l'équation opérationnelle de Mikusiński associée à l'équation (7.11a) :

$$s\widehat{w}(x,s) = \partial_x^2 \widehat{w}(x,s)$$

et considérons cette équation à s fixé ; nous obtenons l'équation différentielle ordinaire

$$\frac{d^2\widehat{w}}{dx^2}(x,s) - s\widehat{w}(x,s) = 0 \tag{7.12}$$

L'équation caractéristique associée à (7.12) s'écrit

$$\zeta^2 - s = 0, \quad \text{c.à.d.} \quad \zeta = \pm\sqrt{s}$$

La solution générale de (7.11a) peut alors se mettre sous la forme

$$\widehat{w}(x,s) = e^{x\sqrt{s}}\gamma_1(s) + e^{-x\sqrt{s}}\gamma_2(s)$$

ou bien

$$\widehat{w}(x,s) = \cosh(x\sqrt{s})\lambda_1(s) + \frac{\sinh(x\sqrt{s})}{\sqrt{s}}\lambda_2(s) \tag{7.13}$$

La deuxième forme est préférable, les dérivées des solutions fondamentales étant reliées de manière agréable. En effet, notant

$$C_x = \cosh(x\sqrt{s}), \quad S_x = \frac{\sinh(x\sqrt{s})}{\sqrt{s}}$$

nous disposons des relations, désignant la dérivation spatiale par un prime :

$$\begin{aligned} \partial_x C_x &= C_x' = sS_x \\ \partial_x S_x &= S_x' = C_x \end{aligned}$$

Nous pouvons déduire de ce qui précède les formes générales de la solution et de la première dérivée spatiale

$$\begin{aligned} \widehat{w}(x,s) &= C_x\lambda_1(s) + S_x\lambda_2(s) \\ \partial_x\widehat{w}(x,s) &= sS_x\lambda_1(s) + C_x\lambda_2(s) \end{aligned}$$

Les conditions aux limites (7.11b) et (7.11c) donnent alors successivement

$$\begin{aligned} \lambda_2(s) &= 0 \qquad (\partial_x\widehat{w}(0,s) = 0) \\ C_1\lambda_1(s) &= \widehat{u}(s) \end{aligned}$$

et nous obtenons l'équation

$$\cosh(\sqrt{s})\,\widehat{w}(x,s) = \cosh(x\sqrt{s})\,\widehat{u}(s)$$

ou encore

$$C_1\,\widehat{w}(x) = C_x\,\widehat{u}$$

Paramétrisation. Introduisant une nouvelle variable

$$\omega(t) = w(0,t)$$

Et ayant

$$\widehat{w}(x) = C_x\lambda_1 + S_x\lambda_2$$

Nous obtenons :

$$\begin{aligned} \lambda_2 &= 0 \\ C_1\lambda_1 &= \widehat{u} \\ \lambda_1 &= \widehat{\omega} \end{aligned}$$

et

$$\widehat{u} = C_1\widehat{\omega} \tag{7.14a}$$

$$\widehat{w}(x) = C_x\widehat{\omega} \tag{7.14b}$$

On recouvre une paramétrisation en termes de $\widehat{\omega}$ qui, dans le cadre de modules sur un anneau approprié, pourra correspondre à la liberté.

Expression dans le domaine temporel. Écrivons formellement

$$\cosh(\sqrt{s}) = \sum_{i\geqslant 0} \frac{s^i}{(2i)!}$$

Nous obtenons, dans le domaine temporel :

$$w(x,t) = \sum_{i\geqslant 0} \frac{x^{2i}}{(2i)!}\,\omega^{(i)}(t)$$
$$u(t) = \sum_{i\geqslant 0} \frac{1}{(2i)!}\,\omega^{(i)}(t)$$

Ces séries ont été écrites dans l'anneau des séries formelles $k\llbracket \frac{d}{dt} \rrbracket$.

Remarque 7.6.1. Notons au passage que l'anneau $k\llbracket \frac{d}{dt} \rrbracket$ est doté de très agréables propriétés algébriques : c'est un anneau principal (tout idéal y est principal, c.à.d. engendré par un unique élément ; sur la notion d'idéal, le lecteur pourra se reporter à l'annexe 7.A, p. 325). Les notions d'absence de torsion (cf. annexe 7.A, p. 324), de projectivité (cf. annexe 7.A, p. 327) et de liberté y coïncident, tout comme sur $k[\frac{d}{dt}]$ (anneau des systèmes de dimension finie). Remarquons également que $k[\frac{d}{dt}^{-1}]\llbracket \frac{d}{dt} \rrbracket$ est un corps. Ceci étant, rien ne garantit que les séries formelles mises en jeu soient convergentes.

Il est possible de donner un sens relativement général à ces séries par des procédés de resommation (voir l'annexe 7.B, p. 329).

Perspective temporelle

Une vue légèrement différente repose sur le théorème classique de Cauchy-Kovaleski. Le système

$$\partial_x^2 w(x,t) = \partial_t w(x,t), \qquad x \in [0,1], t \in [0,\infty[\tag{7.15a}$$
$$\partial_x w(0,t) = 0 \tag{7.15b}$$
$$w(0,t) = \omega(t) \tag{7.15c}$$

est effectivement sous forme de Cauchy-Kovaleski, nous permettant de chercher une solution formelle sous forme analytique

$$w(x,t) = \sum_{i\geqslant 0} a_i(t)\frac{x^i}{i!}$$

où les a_i sont des fonctions indéfiniment dérivables. Une verification formelle, utilisant (7.15), mène à

$$a_{i+2}(t) = \dot{a}_i(t), \quad i \geqslant 0$$
$$a_1(t) = 0$$
$$a_0(t) = \omega(t)$$

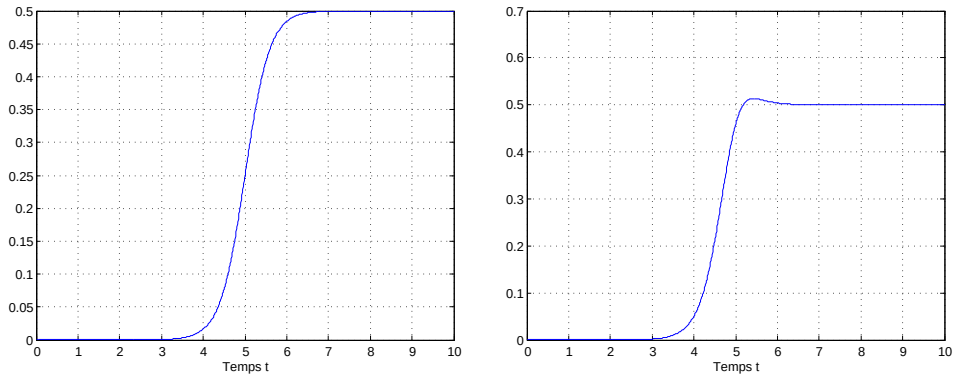


FIG. 7.5: Trajectoire de référence $t \mapsto \omega_r(t)$ (à gauche) et commande en boucle ouverte u (à droite).

D'où, pour $i \geqslant 0$

$$\begin{aligned} a_{2i}(t) &= \omega^{(i)}(t) \\ a_{2i+1}(t) &= 0 \end{aligned}$$

de sorte que nous retrouvons

$$w(x,t) = \sum_{i \geqslant 0} \frac{x^{2i}}{(2i)!}\, \omega^{(i)}(t) \tag{7.16a}$$

$$u(t) = \sum_{i \geqslant 0} \frac{\omega^{(i)}(t)}{(2i)!} \tag{7.16b}$$

Sur la figure 7.5 sont représentées la trajectoire de référence $t \mapsto \omega(t)$ et la commande en boucle ouverte $u(t)$. Sur la figure 7.6 est représenté $w(x,t)$. Nous avons pris pour $t \mapsto \omega(t)$ la trajectoire suivante :

$$\omega(t) = 0.25 * \big(1 + \tanh(1.7 * (t-5)\big)$$

Paramétrisation pour d'autres conditions aux limites

Considérons toujours l'équation de la chaleur, mais avec cette fois des conditions aux limites de Dirichlet :

$$\partial_x^2 w(x,t) = \partial_t w(x,t), \quad x \in [0,1],\ t \in [0,+\infty[\tag{7.17a}$$

$$w(0,t) = 0 \tag{7.17b}$$

$$w(1,t) = u(t) \tag{7.17c}$$

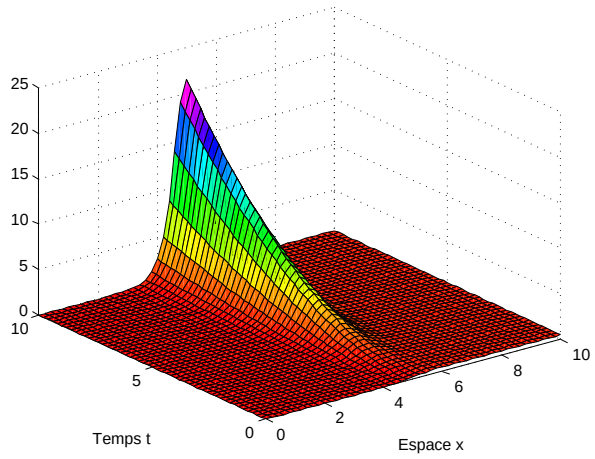


Fig. 7.6: Valeurs de $w(x,t)$.

La solution générale et sa dérivée spatiale s'expriment de la même manière :

$$\begin{aligned}\widehat{w}(x) &= C_x\lambda_1 + S_x\lambda_2\\ \partial_x\widehat{w}(x) &= sS_x\lambda_1 + C_x\lambda_2\end{aligned}$$

Seules les conditions aux limites se trouvent modifiées :

$$\begin{aligned}\lambda_1 &= 0 \quad (\widehat{w}(0) = 0)\\ S_1\lambda_2 &= \widehat{u}\end{aligned}$$

et la solution générale s'exprime comme

$$\widehat{w}(x) = S_x\lambda_2$$

Ainsi λ_2 apparaît comme le paramètre libre pouvant jouer le rôle d'une base. Remarquons que l'on a

$$\partial_x\widehat{w}(x) = sS_x\lambda_1 + C_x\lambda_2 = C_x\lambda_2$$

ce qui permet d'obtenir

$$\lambda_2 = \partial_x\widehat{w}(0)$$

Et $\partial_x\widehat{w}(0)$ peut jouer le rôle d'une base d'un module sur un anneau approprié.

Le système correspondant aux équations (7.17) doit alors contenir non seulement chacun des $\widehat{w}(x)$ pour $x \in [0,1]$, mais également toutes ses dérivées temporelles et spatiales.

Les divers exemples examinés nous ont mené à des structures linéaires sur des anneaux. Nous allons maintenant discuter des structures d'anneaux appropriés.

7.7 Calcul opérationnel utilisé

Nous utiliserons comme classe de fonctions généralisées servant de support à un calcul opérationnel les ultradistributions ou leur transformation de Laplace. Commençons par rappeler quelques éléments sur ces fonctions généralisées.

Bref panorama de classes de fonctions et d'opérateurs

Parmi les plus importantes classes de fonctions (généralisées ou non), on compte, du plus au moins régulier :

$\mathcal{O}$	Fonctions holomorphes
$\mathcal{A}$	Fonctions analytiques réelles
$\mathcal{E}^{(\rho)}$	Fonctions ultradifférentiables Gevrey de Beurling
$\mathcal{E}^{\{\rho\}}$	Fonctions ultradifférentiables Gevrey de Roumieu
$\mathcal{D}$	Fonctions indéfiniment différentiables à support compact
$\mathcal{D}^{(\rho)}$	Fonctions ultradifférentiables Gevrey de Beurling à support compact
$\mathcal{D}^{\{\rho\}}$	Fonctions ultradifférentiables Gevrey de Roumieu à support compact
$\mathcal{S}$	Fonctions indéfiniment différentiables à décroissance rapide
$\mathcal{E}$	Fonctions indéfiniment différentiables
C^n	Fonctions n fois différentiables
C	Fonctions continues
L^p	Fonctions p fois sommables
L^1	Fonctions sommables
M^1	Mesures complexes
$\mathcal{E}'$	Distributions à support compact
$\mathcal{S}'$	Distributions tempérées
$\mathcal{D}^{\{\rho\}'}$	Ultradistributions de Roumieu
$\mathcal{D}^{(\rho)'}$	Ultradistributions de Beurling
$\mathcal{D}'$	Distributions
$\mathcal{E}'^{\{\rho\}}$	Ultradistributions de Roumieu à support compact
$\mathcal{E}'^{(\rho)}$	Ultradistributions de Beurling à support compact
$\mathcal{B}$	Hyperfonctions

$\mathcal{O}'$ Fonctionnelles analytiques

Parmi toutes ces classes, la plus générale est celle des fonctionnelles analytiques. Malheureusement, elle manque d'un minimum de structure algébrique, en ce sens que $\mathcal{O}'(\Omega)$, l'ensemble de telles fonctionnelles définies sur un ouvert Ω de $\mathbb{R}^n$ ne constitue pas un faisceau ; ce dernier fait interdit de définir la notion de support pour ces fonctionnelles. C'est donc la classe des hyperfonctions, que l'on peut identifier au dual de $\mathcal{A}$, qui est, au sens de la classification précédente, la plus générale tout en conservant un minimum de propriétés algébriques agréables (voir [21]).

Ultradistributions

Étant donné un entier positif n, considérons le multi-indice $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ d'entiers positifs, on note $|\boldsymbol{\alpha}| = \alpha_1 + \alpha_2 + \cdots + \alpha_n$ et $\boldsymbol{\alpha}! = \alpha_1!\alpha_2! \cdots \alpha_n!$. Pour une fonction à valeurs complexe f indéfiniment différentiable définie sur un ouvert $\Omega \subset \mathbb{R}^n$, on note

$$D^{\boldsymbol{\alpha}} f(x) = \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_n^{\alpha_n}}$$

où $\boldsymbol{x} = (x_1, \ldots, x_n)$. Pour $\Omega \subset \mathbb{R}^n$ ouvert considérons M_p $(p = 0, 1, \ldots)$ une suite de nombres positifs soumise aux conditions suivantes :

(M.0) (Normalisation)

$$M_0 = M_1 = 1$$

(M.1) (Convexité logarithmique)

$$M_p^2 \leqslant M_{p-1} M_{p+1}, \quad p = 1, 2, \ldots$$

(M.2) (Stabilité par opérateurs différentiels)

$$\exists\, G, H, \text{ tels que } M_{p+1} \leqslant G H^p M_p, \quad p = 0, 1, \ldots$$

(M.3) (Non quasi-analycité)

$$\sum_{p=1}^{\infty} \frac{M_{p-1}}{M_p} < \infty$$

Définition 7.7.1. Soit M_p une suite de nombres positifs et $\Omega \subset \mathbb{R}^n$ un ouvert. Une fonction $f \in \mathcal{E} = C^\infty(\Omega)$ est dite ultradifférentiable de classe (M_p) (resp. $\{M_p\}$) si pour tout compact $K \subset \Omega$ et pour tout $h > 0$, il existe une constante C (resp. pour tout compact $K \subset \Omega$, il existe des constantes h et C) telle que

$$\sup_{x \in K} |D^{\boldsymbol{\alpha}} \phi(x)| \leqslant C h^{|\boldsymbol{\alpha}|} M_{|\boldsymbol{\alpha}|} \quad \text{pour tout } \boldsymbol{\alpha}$$

On désigne par $\mathcal{E}^*(\Omega)$ l'ensemble des *fonctions ultradifférentiables de classe* $*$ sur Ω (où $* = (M_p)$ ou $\{M_p\}$) et par $\mathcal{D}^*(\Omega)$ l'ensemble des fonctions de $\mathcal{E}^*(\Omega)$ à support compact dans Ω.

Notons que l'on peut remplacer $(M.3)$ par la condition de quasi-analycité suivante :

(M.3') (i) Il existe des constantes strictement positives L et C telles que, pour tous p

$$p\,! \leqslant CL^p M_p$$

(ii) La série écrite en (M.3) diverge

$$\sum_{p=1}^{\infty} \frac{M_{p-1}}{M_p} = \infty$$

De plus, si une suite M_p satisfaisant (M.3') est telle que

$$\liminf_{p\to\infty} \sqrt[p]{\frac{p\,!}{M_p}} > 0$$

alors $\mathcal{E}^{\{M_p\}}$ est la classe des fontions analytiques.

Un autre exemple est celui fourni par $M_p = (p\,!)^d$, auquel cas $\mathcal{E}^{\{M_p\}}$ est la classe des fonctions Gevrey d'ordre $d+1$.

Définition 7.7.2. Soit M_p une suite de nombres positifs satisfaisant (M.1) et (M.3). Pour tout ouvert $\Omega \subset \mathbb{R}^n$ le dual $\mathcal{D}^{*'}(\Omega)$ (resp. $\mathcal{E}^{*'}(\Omega)$) de $\mathcal{D}^*(\Omega)$ (resp. $\mathcal{E}^*(\Omega)$) est l'ensemble des *ultradistributions de classe* $*$ (resp. des *ultradistributions à support compact de classe* $*$) définies sur Ω où $*$ remplace $\{M_p\}$ ou bien (M_p).

La transformation de Laplace d'une ultradistribution $f \in \mathcal{E}'^*$ est donnée par $\widehat{f}(s) = f(g_s)$ avec $g_s(t) = \mathrm{e}^{-st}$. L'isomorphisme entre les anneaux de convolution d'ultradistributions à support compact et leur transformation de Laplace est donné par un théorème de type Paley-Wiener figurant par exemple dans [20].

Le calcul opérationnel que nous utiliserons est celui naturellement associé à la convolution des ultradistributions ou à leur transformation de Laplace.

7.8 Problèmes d'EDPs frontières comme systèmes de convolution

Classe de modèles considérés

Par souci de simplicité d'exposition, on se restreindra dans ce qui suit aux systèmes admettant le modèle suivant en les variables distribuées $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_l$

ainsi qu'en les variables concentrées $\boldsymbol{u} = (u_1, \ldots, u_m)$:

$$\begin{aligned} \partial_x \boldsymbol{w}_i = A_i \boldsymbol{w}_i + B_i \boldsymbol{u}, \quad \boldsymbol{w}_i : \Omega_i \to (\mathcal{D}'^*)^p, \quad \boldsymbol{u} \in (\mathcal{D}'^*)^m \\ A_i \in (\mathbb{R}[\partial_t])^{p_i \times p_i}, \quad B_i \in (\mathbb{R}[\partial_t])^{p_i \times m}, \quad i \in \{1, \ldots, l\} \end{aligned} \tag{7.18a}$$

où $\mathcal{D}'^*$ désigne un espace d'ultradistributions. Les intervalles $\Omega_1, \ldots, \Omega_l$ sont donné par un voisinage ouvert de $\tilde{\Omega}_i = [x_{i,0}, x_{i,1}]$. On supposera, sans perte de généralité, que $x_{i,0} = 0$. Nous ferons une hypothèse cruciale pour notre étude. Les polynômes caractéristiques des matrices $A_1, \ldots, A_l$ peuvent s'exprimer comme

$$P_i(\lambda) := \det(\lambda 1 - A_i) = \sum_{\nu=0}^{p_i} a_{i,\nu} \lambda^\nu, \quad a_{i,\nu} = \sum_{\nu+\mu \leq p_i} a_{i,\nu,\mu} s^\mu \tag{7.18b}$$

où $a_{i,\nu,\mu} \in \mathbb{R}$, $a_{i,p_i,0} = 1$. De plus, les parties principales de ces polynômes, données par $\sum_{\mu+\nu=p_i} a_{i,\mu,\nu} s^\mu \lambda^\nu$ sont hyperboliques relativement à l'axe temporel t, c.à.d. que les racines des polynômes $\sum_{\mu+\nu=p_i} a_{i,\mu,\nu} \lambda^j$ sont réelles.

Le modèle est complété par les conditions aux frontières

$$\sum_{i=1}^{l} L_i \boldsymbol{w}_i(0) + R_i \boldsymbol{w}_i(\ell_i) + D\boldsymbol{u} = 0 \tag{7.18c}$$

où $D \in (\mathbb{R}[\partial_t])^{q \times m}$ et $L_i, R_i \in (\mathbb{R}[\partial_t])^{q \times p_i}$.

Solution du problème de Cauchy

Rappelons ici quelques propriétés de la solution du problème de Cauchy de la forme (7.18a) avec les conditions initiales données par $x = \xi$, c.à.d.,

$$\partial_x \boldsymbol{w} = A\boldsymbol{w} + B\boldsymbol{u}, \quad \boldsymbol{w}(\xi) = \boldsymbol{w}_\xi \tag{7.19}$$

avec $A \in (\mathbb{R}[\partial_t])^{p \times p}$, $B \in (\mathbb{R}[\partial_t])^{p_i \times q}$ ayant les mêmes propriétés que A_i, B_i de la section précédente. (Tout au long de cette section on utilisera la notation de la section précédente en supprimant l'index $i \in \{1, \ldots, l\}$). Pour ce faire, considérons le problème aux valeurs initiales :

$$P(\partial_x) v(x) = 0, \quad (\partial_x^j v)(0) = v_j \in \mathcal{E}^*(\mathbb{R}), \quad j = 0, \ldots, p-1 \tag{7.20}$$

associé à l'équation caractéristique (7.18b). Sous la hypothèse la solution unique de (7.20) est donné par

$$v(x) = \sum_{j=0}^{p-1} C_j(x) v_j.$$

où $C_0, \ldots, C_{p-1} : \Omega \to \mathcal{E}'^*(\mathbb{R})$ sont des fonctions indéfiniment différentiables satisfaisant ($k, j \in \{0, \ldots, p-1\}$)

$$\partial_x^k C_j(0) = \begin{cases} 1, & k = j \\ 0, & k \neq j. \end{cases} \tag{7.21}$$

et la juxtaposition dénote la convolution. De plus, on peut vérifier les relations

$$\partial_x C_j = \partial_x C_{j-1} - a_j C_{p-1}, \quad j = 1, \ldots, p-1, \quad \partial_x C_0 = -a_0 C_{p-1}. \tag{7.22}$$

L'anneau $\mathcal{E}^*$ (resp. $\mathcal{D}'^*$) est un espace approprié de fonctions ultradifférentiables (resp. d'ultradistributions). Par ailleurs, $\mathcal{E}'^*$ désigne l'espace d'ultradistributions à support compact de même ordre Gevrey. En conformité avec les résultats donnés en [19, Thrm. 12.5.6] ou [38, Thrm 2.5.2,Prop. 2.5.6] on peut choisir l'espace $\mathcal{E}^{(p/(p-1))}$ (resp. $\mathcal{D}'^{(p/(p-1))}$) qui correspond aux fonctions ultradifférentiables (resp. aux ultradistributions) Gevrey de Beurling d'ordre $p/(p-1)$.

Exemple 7.8.1. Pour une équation de la chaleur, telle que (7.11) les expressions dans la domaine de Laplace sont

$$\widehat{C}_0(x) = \cosh(x\sqrt{s}), \quad \widehat{C}_1(x) = \frac{\sinh(x\sqrt{s})}{\sqrt{s}}.$$

Dans le domaine temporel, on obtient les expressions suivantes

$$C_0(x)v_0 = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}\partial_t^k v_0, \quad C_1(x)v_1 = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!}\partial_t^k v_1$$

Pour tout $x \in \Omega$ fixé, $C_0(x), C_1(x)$ sont des ultradistributions (éléments de $\mathcal{E}'^{(2)}$).

La solution unique $x \mapsto \Phi(x, \xi)$ du problème aux valeurs initiales

$$\partial_x \Phi(x, \xi) = A\Phi(x, \xi), \quad \Phi(\xi, \xi) = 1$$

avec 1 désignant l'identité de $\mathcal{E}'^*(\mathbb{R})^{p\times p}$, est alors donnée par

$$\Phi(x, \xi) = \sum_{j=0}^{p-1} A^j C_j(x - \xi) \tag{7.23}$$

De l'unicité de la solution, on déduit la formule de composition

$$\Phi(x, \xi)\Phi(\xi, \zeta) = \Phi(x, \zeta). \tag{7.24}$$

La solution du problème associé à l'équation inhomogène

$$\partial_x \Psi(x, \xi) = A\Psi(x, \xi) + B \tag{7.25}$$

avec des conditions initiales homogènes prescrites à $x = \xi$, s'obtient par variation des constantes. Ceci donne

$$\Psi(x, \xi) = \int_{\xi}^{x} \Phi(x, \zeta) d\zeta B \tag{7.26}$$

La solution générale du problème (7.19) est alors

$$\boldsymbol{w}(x) = \Phi(x,\xi)\boldsymbol{w}_\xi + \Psi(x,\xi)\boldsymbol{u}$$

ou, de manière équivalente

$$\boldsymbol{w}(x) = W(x,\xi)\boldsymbol{c},\ W(x,\xi) = \begin{pmatrix}\Phi(x,\xi) & \Psi(x,\psi)\end{pmatrix},\ \boldsymbol{c}_\xi = \begin{pmatrix}\boldsymbol{w}_\xi \\ \boldsymbol{u}\end{pmatrix}$$

Les composantes de la matrice $\Phi(x,\xi)$ sont combinaisons linéaires des $C_0(x-\xi),\dots,C_{p-1}(x-\xi)$ sur $\mathbb{C}[\partial_t]$. Au contraire, d'après (7.26), les composantes de Ψ peuvent contenir des intégrales de $C_0,\dots,C_{p-1}$. Nous allons montrer que ces intégrales peuvent s'exprimer comme combinaison linéaire des fonctions $C_0,\dots,C_{p-1}$ ayant des coefficients dans $\mathbb{C}(\partial_t)$. De telles représentations sont obtenues par l'intégration des équations (7.22) et par l'évaluation correspondante des «conditions initiale»(7.21). Pour obtenir les expressions générales, il faut supposer que les μ premiers coefficients $a_0,\dots,a_{\mu-1}$ sont égaux à zéro. Dans ce cas, les fonctions $C_0,\dots,C_{\mu-1}$ correspondent aux polynômes

$$C_j(x) = \frac{x^j}{j!},\quad j = 0,\dots,\mu-1 \tag{7.27a}$$

dont les intégrales en espace se déduisent facilement. Ensuite l'équation

$$\partial_x C_\mu(x) = \begin{cases} -a_0 C_{p-1}(x), & \mu = 0 \\ C_{\mu-1}(x) - a_\mu C_{p-1}(x), & \mu > 0 \end{cases}$$

donne

$$\int_0^x C_{p-1}(\zeta)d\zeta = \begin{cases} -\frac{1}{a_0}\left(C_\mu(x) - C_\mu(0)\right), & \mu = 0 \\ -\frac{1}{a_\mu}\left(C_\mu(x) - C_\mu(0) - \frac{x^\mu}{\mu!}\right), & \mu > 0 \end{cases} \tag{7.27b}$$

Finalement, des équations

$$\partial_x C_j(x) = C_{j-1}(x) - a_\mu C_{p-1}(x)$$

on déduit

$$\int_0^x C_{j-1}(\zeta)d\zeta = C_j(x) - C_j(x) + \int_0^x a_\mu C_{p-1}(\xi)d\xi,\quad j = \mu+1,\dots,p-2. \tag{7.27c}$$

Module du système

En utilisant les solutions du problème aux valeurs initiales dans les conditions aux bords, on obtient

$$\boldsymbol{w}(x) = W_{\boldsymbol{\xi}}(x)\boldsymbol{c}_{\boldsymbol{\xi}},\quad P_{\boldsymbol{\xi}}\boldsymbol{c}_{\boldsymbol{\xi}} = 0 \tag{7.28}$$

Ici, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ est *arbitraire mais fixé*, $\boldsymbol{c}_{\boldsymbol{\xi}}^T = (\boldsymbol{w}_1^T(\xi_1), \dots, \boldsymbol{w}_l^T(\xi_l), \boldsymbol{u}^T)$,

$$W_{\boldsymbol{\xi}} = \begin{pmatrix} \Phi_1(x,\xi_1) & 0 & 0 & \Psi_1(x,\xi_1) \\ 0 & \ddots & 0 & \vdots \\ 0 & \cdots & \Phi_l(x,\xi_l) & \Psi_l(x,\xi_l) \end{pmatrix}, \quad P_{\boldsymbol{\xi}} = \big(P_{\boldsymbol{\xi},1}, \dots, P_{\boldsymbol{\xi},l+1}\big)$$

avec

$$P_{\boldsymbol{\xi},i} = L_i\Phi_i(0,\xi_i) + R_i\Phi_i(\ell_i,\xi_i), \quad i = 1,\dots,l$$
$$P_{\boldsymbol{\xi},l+1} = D + \sum_{i=1}^{l} L_i\Psi_i(0,\xi_i) + R_i\Psi_i(\ell_i,\xi_i)$$

Nous allons représenter le système étudié par un module engendré par les $\boldsymbol{c}_\xi$, $\boldsymbol{u}$ avec la présentation donnée en (7.28) [32, 12, 11, 30]. L'anneau des coefficients doit contenir les composantes de $W_{\boldsymbol{\xi}}(x)$ et $P_{\boldsymbol{\xi}}$, qui sont constituées de valeurs des fonctions $C_{i,j}$, $(j = 1,\dots,p_i,\ i = 1,\dots,l)$ de $\mathbb{R}$ dans $\mathcal{E}'^*$. De plus, les matrices peuvent également contenir des valeurs des intégrales en espace de $C_{i,j}$. Un choix possible pour l'anneau des coefficients est alors $\mathcal{R}^I = \mathbb{C}[\partial_t, \mathfrak{S}, \mathfrak{S}^I] \subset \mathcal{E}'^*$ avec

$$\mathfrak{S} = \{C_{i,j}(x) | x \in \mathbb{R}; i = 1,\dots,l; j = 0,\dots,p_i - 1\},$$
$$\mathfrak{S}^I = \{C^I_{i,j}(x) | x \in \mathbb{R}; i = 1,\dots,l; j = 0,\dots,p_i - 1\}$$

et

$$C^I_{i,j}(x) = \int_0^x C_{i,j}(\zeta)d\zeta, \quad i = 1,\dots,l, \quad j = 0,\dots,p_i - 1.$$

Cet anneau est isomorphe à un sous anneau de l'anneau $\mathcal{O}$ des fonctions entières en s par transformation de Laplace.

Dans le même esprit qu'en [31, 5, 17], et afin de simplifier l'analyse des propriétés des modules, on utilisera, au lieu de $\mathcal{R}^I$, un anneau un peu plus grand, donné par $\mathcal{R} = \mathbb{C}(\partial_t)[\mathfrak{S}] \cap \mathcal{O}$. (On identifie l'anneau $\mathbb{C}(\partial_t)[\mathfrak{S}]$ avec son image par la transformation Laplace).

Définition 7.8.1. Le *système de convolution* Σ associé au problème frontière (7.18) est le module engendré par $\boldsymbol{c}_{\boldsymbol{\xi}}$ et $\boldsymbol{u}$ sur $\mathcal{R}$ avec $P_{\boldsymbol{\xi}}$ pour matrice de présentation.

On vérifie aisément que Σ ne dépend pas du choix de $\boldsymbol{\xi}$ (cf. [47, Section 3.3] et [44, Remark 4]).

Exemple 7.8.2. Nous allons considérer un exemple similaire à celui donné par les équations (7.11).

Modèle. Le modèle étudié est le suivant :

$$\partial_x^2 w(x,t) = \partial_t w(x,t), \quad x \in [0,\ell],\ t \in [0,+\infty[\tag{7.29a}$$
$$\partial_x w(0,t) = 0 \tag{7.29b}$$
$$\partial_x w(\ell,t) = u(t) \tag{7.29c}$$

Ce modèle peut se réecrire

$$\partial_x \begin{pmatrix} w(x,t) \\ \partial_x w(x,t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \partial_t & 0 \end{pmatrix} \begin{pmatrix} w(x,t) \\ \partial_x w(x,t) \end{pmatrix} \tag{7.30a}$$

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} w(0,t) \\ \partial_x w(0,t) \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} w(\ell,t) \\ \partial_x w(\ell,t) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t) \tag{7.30b}$$

Solutions fondamentales. Les solutions du problème sans frontière, c.à.d. de l'équation (7.30a) sont

$$\boldsymbol{w}(x) = \begin{pmatrix} C_{x-\xi} & S_{x-\xi} \\ \partial_t S_{x-\xi} & C_{x-\xi} \end{pmatrix} \boldsymbol{c}$$

Donc, nous avons

$$w(x) = C_{x-\xi}\, c_1 + S_{x-\xi}\, c_2$$

ce qui exprime le fait que (C_x, S_x) est une base de l'espace vectoriel des solutions de (7.30a) considérée comme une EDO par rapport à la variable x.

Conditions aux bords. Les conditions aux bords (7.30b) s'écrivent

$$L\Phi(0,\xi)\boldsymbol{c} + R\Phi(\ell,\xi)\boldsymbol{c} - Du = 0$$

ou bien, en forme matricielle explicite

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} C_{-\xi} & S_{-\xi} \\ \partial_t S_{-\xi} & C_{-\xi} \end{pmatrix} \boldsymbol{c} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} C_{\ell-\xi} & S_{\ell-\xi} \\ \partial_t S_{\ell-\xi} & C_{\ell-\xi} \end{pmatrix} \boldsymbol{c} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} u = 0$$

qui est équivalente à

$$\begin{pmatrix} -\partial_t S_\xi & C_\xi \\ \partial_t S_{\ell-\xi} & C_{\ell-\xi} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} u = 0$$

En résumé, nous obtenons

$$\begin{pmatrix} -\partial_t S_\xi & C_\xi & 0 \\ \partial_t S_{\ell-\xi} & C_{\ell-\xi} & -1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ u \end{pmatrix} = 0, \quad \boldsymbol{w}(x) = \begin{pmatrix} C_{x-\xi} & S_{x-\xi} & 0 \\ \partial_t S_{x-\xi} & C_{x-\xi} & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ u \end{pmatrix}$$

que nous désignerons par

$$P_\xi \begin{pmatrix} \boldsymbol{c} \\ u \end{pmatrix} = 0, \quad \boldsymbol{w}(x) = W_\xi(x) \begin{pmatrix} \boldsymbol{c} \\ u \end{pmatrix}$$

avec les notations

$$P_\xi = \begin{pmatrix} -\partial_t S_\xi & C_\xi & 0 \\ \partial_t S_{\ell-\xi} & C_{\ell-\xi} & -1 \end{pmatrix}, \qquad W_\xi = \begin{pmatrix} C_{x-\xi} & S_{x-\xi} \\ \partial_t S_{x-\xi} & C_{x-\xi} \end{pmatrix}$$

7.9 Commandabilités des systèmes à paramètres répartis du deuxième ordre

Dans cette section, le schéma obtenu en section 7.9 est appliqué à des systèmes constitués de réseaux de systèmes du second ordre couplés aux bords. Bien que ces systèmes ne soient relativement particuliers, une grande classe de systèmes physiques importants sont couverts. Par ailleurs, pour cette classe de systèmes, les résultats de commandabilité sont bien plus détaillés que dans le cas général.

Classe de modèles considérés

Nous supposons que $p_i = 2$, $i = 1, \ldots, l$ dans (7.18). De plus, toutes les matrices $A_1, \ldots, A_l$ donnent lieu au même polynôme caractéristique :

$$P_i(\lambda) = \lambda^2 - \sigma, \quad \sigma = as^2 + bs + c \neq 0, \quad a, b, c \in \mathbb{R}, \quad a \geqslant 0 \tag{7.31}$$

Par ailleurs, les intervalles $\Omega_1, \ldots, \Omega_l$ doivent avoir des longueurs rationnellement dépendantes. Plus précisément, nous supposerons que Ω_i $(i = 1, \ldots, l)$ est donné par un voisinage ouvert de

$$\tilde{\Omega}_i = [0, q_i \ell], \quad q_i \in \mathbb{Q}, \quad \ell \in \mathbb{R}. \tag{7.32}$$

Pour la matrice Φ, nous obtenons, en accord avec l'équation (7.23),

$$\Phi(x, \xi) = AS(x - \xi) + 1C(x - \xi)$$

où 1 désigne l'identité de $\mathcal{E}'^*(\mathbb{R})^{2\times 2}$. Nous remplaçons C_0 par C et C_1 par S par souci de simplicité des notations.

Le formule de composition (7.24) s'écrit dans le cas où A est la matrice compagnon du polynôme caractéristique, c.à.d.,

$$A = \begin{pmatrix} 0 & 1 \\ \sigma & 0 \end{pmatrix}, \quad \Phi(x, \xi) = \begin{pmatrix} C(x - \xi) & S(x - \xi) \\ \sigma S(x - \xi) & C(x - \xi) \end{pmatrix}, \tag{7.33}$$

sous la forme

$$\begin{pmatrix} C(x) & S(x) \\ \sigma S(x) & C(x) \end{pmatrix} \begin{pmatrix} C(y) & S(y) \\ \sigma S(y) & C(y) \end{pmatrix} = \begin{pmatrix} C(x+y) & S(x+y) \\ \sigma S(x+y) & C(x+y) \end{pmatrix}.$$

Cela donne en particulier les formules de composition pour les sinus et cosinus hyperboliques

$$C(x+y) = C(x)C(y) + \sigma S(x)S(y), \quad S(x+y) = C(x)S(y) + S(x)C(y). \tag{7.34}$$

Finalement, les intégrales de S et C sont obtenues à partir de (7.27) avec $a_0 = -\sigma$, $a_1 = 0$:

$$\int_0^x C(\zeta)dx = S(x), \quad \int_0^x S(\zeta)dx = (C(x) - 1)/\sigma.$$

Le module de système peut être défini comme en section 7.8. Ici, l'ensemble $\mathfrak{S}$ des générateurs de l'anneau des coefficients $\mathcal{R}$ consiste en les valeurs de S et de C uniquement. Pour l'analyse de la commandabilité, il est avantageux de spécifier quelles valeurs des fonctions génératrices devraient être utilisées comme éléments de l'anneau des coefficients $\mathfrak{S}$. Afin de se servir de la dépendance $\mathbb{Q}$-linéaire des longueurs ℓ_i de (7.32) il est utile de commencer avec un anneau engendré par les valeurs de C et S pris en des valeurs multiples rationnelles ou entières de ℓ. Ainsi, au lieu de $\mathcal{R}$, nous utilisons les anneaux $\mathcal{R}_{\mathbb{N}} = \mathbb{C}(\partial_t)[\mathfrak{S}_{\mathbb{N}}] \cap \mathcal{O}$ et $\mathcal{R}_{\mathbb{Q}} = \mathbb{C}(\partial_t)[\mathfrak{S}_{\mathbb{Q}}] \cap \mathcal{O}$, où pour tout $\mathbb{X} \subseteq \mathbb{R}$,

$$\mathfrak{S}_{\mathbb{X}} = \{C(z\ell), S(z\ell) | z \in \mathbb{X}\}$$

Afin de distinguer les modules obtenus pour divers anneaux de coefficients $\mathcal{R}_{\mathbb{N}} \subset \mathcal{R}_{\mathbb{Q}} \subset \mathcal{R}$, la notation $\Sigma_{\mathbb{X}}$ est utilisée.

Définition 7.9.1. Le *système de convolution* $\Sigma = \Sigma_{\mathbb{R}}$ associé au problème frontière (7.18) est le module engendré par $\boldsymbol{c_\xi}$ sur $\mathcal{R}_{\mathbb{R}}$ avec $P_{\boldsymbol{\xi}}$ pour matrice de présentation. Par $\Sigma_{\mathbb{Q}}$ on désigne le même système, mais vu comme module sur $\mathcal{R}_{\mathbb{Q}}$.

Commandabilités des systèmes à paramètres répartis commandés aux bords

Une notion supplémentaire de commandabilité est ici introduite, celle de commandabilité spectrale, généralisant le critère de Hautus des systèmes de dimension finie.

Définition et proposition 7.9.1. *Soit R un anneau isomorphe à un sous-anneau de l'anneau $\mathcal{O}$ des fonctions entières. Notons par $\mathscr{L}$ l'application $R \to \mathcal{O}$ de passage au domaine symbolique (la transformation de Laplace). Un R-système finiment présenté de matrice de présentation P est dit spectralement commandable si l'une des deux conditions équivalentes suivantes est vérifiée :*

(i) La matrice $\hat{P} = \mathscr{L}(P)$ *à coefficients dans* $\mathcal{O}$ *satisfait à la condition :* $\exists k \in \mathbb{N} : \forall \sigma \in \mathbb{C} : rk_{\mathbb{C}}\, \hat{P}(\sigma) = k$.

(ii) Le module $\Sigma_{\mathcal{O}} = \mathcal{O} \otimes_R \Sigma$ *est sans torsion.*

Démonstration. Ce résultat est une simple conséquence du fait que la matrice $\hat{P}$ admet une forme normale de Smith. □

Proposition 7.9.1. *Soit* R *un domaine de Bézout isomorphe à un sous anneau de* $\mathcal{O}$ *et* $\mathscr{L} : R \to \mathcal{O}$ *le passage au domaine symbolique. Alors les notions de commandabilité spectrale et de commandabilité sans torsion sont équivalentes si, et seulement si,* $\mathscr{L}$ *envoie les éléments non inversibles de* R *sur les éléments non inversibles de* $\mathcal{O}$.

Démonstration. Puisque R est un domaine de Bézout, le caractère sans torsion de Σ implique sa liberté. Par produit tensoriel avec le module libre $\mathcal{O}$ on obtient un autre module libre $\Sigma_{\mathcal{O}}$, et, par la Définition et Proposition 7.9.1, la commandabilité spectrale. À nouveau parce que R est un domaine de Bézout, toute matrice de présentation admet une forme normale de Hermite. Donc, le sous module de torsion $\mathrm{t}\Sigma$ de Σ peut être présenté par une matrice carrée triangulaire P^t de rang plein. Si Σ n'est pas sans torsion, au moins une composante de la diagonale de cette matrice est non inversible dans R. Si cette composante est envoyée sur un non inversible de $\Sigma_{\mathcal{O}}$ par $\mathscr{L}$, elle admet un zéro complexe σ_0. Donc, $\mathscr{L}(P^t)$ a une chute de rang en $\sigma = \sigma_0$. Inversement, s'il existe un élément non inversible $r \in R$ qui correspond à un élément inversible $\hat{r} \in \mathcal{O}$, considérons $\Sigma \cong [\tau]/[r\tau]$. De manière évidente, l'image de τ dans $\Sigma_{\mathcal{O}}$ est zéro. Donc le module trivial $\Sigma_{\mathcal{O}}$ est sans torsion. □

Nous ne détaillerons pas la preuve du résultat suivant :

Théorème 7.9.1. *L'anneau* $\mathcal{R}_{\mathbb{Q}}$ *est un domaine de Bézout, c.à.d., tout idéal finiment engendré est principal.*

Démonstration. (Esquisse de preuve). On montre que deux éléments quelconques $p, q \in \mathcal{R}_{\mathbb{Q}}$ possèdent un diviseur commun qui peut s'écrire comme combinaison linéaire de p, q. Pour cela, on s'appuie sur le fait que l'anneau $\mathbb{C}(s)[\mathfrak{S}_{\mathbb{Q}}]$ est également un domaine de Bézout. Ceci provient du fait que ce type d'anneau se construit typiquement comme le quotient $\widetilde{\mathcal{R}}_{\mathbb{X}} := k[\widetilde{C}_a, \widetilde{S}_a; a \in \mathbb{X}]/\mathfrak{a}$ avec l'idéal $\mathfrak{a}$ engendré par

$$\widetilde{C}_a\widetilde{C}_b \pm \sigma\widetilde{S}_a\widetilde{S}_b - \widetilde{C}_{a\pm b},\ \widetilde{S}_a\widetilde{C}_b \pm \widetilde{C}_a\widetilde{S}_b - \widetilde{S}_{a\pm b},\ \widetilde{C}_0 - 1,\ \widetilde{S}_0, \quad \sigma \in k, \quad a, b \in \mathbb{X}$$

Notant C_a et S_a les images canoniques de $\widetilde{C}_a$ et $\widetilde{S}_a$ dans $\widetilde{\mathcal{R}}_{\mathbb{X}}$, l'on déduit les relations

$$C_aC_b \pm \sigma S_aS_b = C_{a\pm b}, \quad S_aC_b \pm C_aS_b = S_{a\pm b} \tag{7.35a}$$

$$C_0 = 1, \quad S_0 = 0, \quad C_a = C_{-a}, \quad S_a = -S_{-a} \tag{7.35b}$$

$$2C_aC_b = C_{a+b} + C_{a-b},\ 2\sigma S_aS_b = C_{a+b} - C_{a-b},\ 2C_aS_b = S_{a+b} - S_{a-b} \tag{7.35c}$$

De plus, tout élément de $r \in \widetilde{\mathcal{R}}_{\mathbb{X}}$ peut s'écrire sous la forme

$$r = \sum_{i=0}^{n} a_{\alpha_i} C_{\alpha_i} + b_{\alpha_i} S_{\alpha_i}, \quad n \in \mathbb{N}, \quad a_{\alpha_i}, b_{\alpha_i} \in k, \quad \alpha_i \in \mathbb{X}^+ \tag{7.36}$$

où $\mathbb{X}^+ = \{|\alpha| : \alpha \in \mathbb{X}\}$. Sachant que $\mathbb{C}(s)[\mathfrak{S}_{\mathbb{Q}}]$ est un domaine de Bézout, on peut trouver des éléments $a, b \in \mathbb{C}[\partial_t, \mathfrak{S}_{\mathbb{Q}}]$ tels que

$$c = ap + bq \in \mathbb{C}[\partial_t, \mathfrak{S}_{\mathbb{Q}}] \tag{7.37}$$

est le PGCD dans $\mathbb{C}(s)[\mathfrak{S}_{\mathbb{Q}}]$. Donc, p/c et q/c appartiennent à $\mathbb{C}(s)[\mathfrak{S}_{\mathbb{Q}}]$. On en déduit un diviseur commun dans $\mathcal{R}_{\mathbb{Q}}$. □

Voici le principal résultat de cette section :

Théorème 7.9.2. *Le système de convolution Σ défini en 7.9.1 est libre, si et seulement s'il est sans torsion. Plus généralement $\Sigma = \mathrm{t}\Sigma \oplus \Sigma/\mathrm{t}\Sigma$, où $\mathrm{t}\Sigma$ est de torsion et $\Sigma/\mathrm{t}\Sigma$ est libre. De plus, Σ est spectralement commandable si, et seulement s'il est sans torsion.*

Démonstration. Rappelons que, selon la définition 7.9.1, $\Sigma \cong \mathcal{R} \otimes_{\mathcal{R}_{\mathbb{Q}}} \Sigma_{\mathbb{Q}}$ et $\mathcal{R}_{\mathbb{Q}}$ est un domaine de Bézout par la proposition 7.9.1. Puisque la première assertion est vraie pour les modules finiment présentés sur tout domaine de Bézout, elle est vraie pour $\Sigma_{\mathbb{Q}}$. La deuxième assertion découle le la proposition 7.9.1. (Le fait que la transformée de Laplace envoie tout élément non inversible de $\mathcal{R}_{\mathbb{Q}}$ sur un élément non inversible de $\mathcal{O}$ est évident). Clairement les deux résultats sont également valables pour Σ, qui est obtenu par extension de scalaires. □

7.10 Bibliographie

[1] Aoustin, Y., M. Fliess, H. Mounier, P. Rouchon et J. Rudolph: *Theory and practice in the motion planning control of a flexible robot arm using Mikusiǹski operators.* Dans *Proc. of 4^{th} Symposium on Robotics and Control*, pages 287–293, Nantes, 1997.

[2] Balser, W.: *From Divergent Power Series to Analytic Functions : Theory and Applications of Multisummable Power Series*, tome 1582 de *Lecture Notes in Mathematics*. Springer Verlag, 1994.

[3] Balser, W.: *Summability of formal power series solutions of ordinary and partial differential equations.* Functional Diff. Eq., 8 :11–24, 2001.

[4] Balser, W.: *Multisummability of formal power series solutions of partial differential equations with constant coeffcients.* J. Diff. Equations, 201 :63–74, 2004.

[5] Brethé, D. et J.J. Loiseau: *A result that could bear fruit for the control of delay-differential systems.* Dans Proc. *IEEE MSCA'96*, Chania, Crete, 1996.

[6] Courant, R. et D. Hilbert: *Methoden der mathematischen Physik*, tome 1. Julius Springer, Berlin, 1937. Traduction américaine : *Methods of mathematical physics.* Interscience Publishers, New York, 1953.

[7] Écalle, J.: *Introduction aux fontions analysables et preuve constructive de la conjecture, de Dulac.* Hermann, Paris, 1992.

[8] Eisenbud, D.: *Commutative Algebra with a view toward Algebraic Geometry.* Numéro 150 dans *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1995.

[9] Fliess, M.: *Generalized linear systems with lumped or distributed parameters and differential vector spaces.* Internat. J. Control, 49 :1989–1999, 1989.

[10] Fliess, M.: *Some basic structural properties of generalized linear systems.* Systems Control Lett., 15 :391–396, 1990.

[11] Fliess, M. et H. Mounier: *Controllability and observability of linear delay systems : an algebraic approach.* Control Optimization and Calculus of Variations, 3 :301–314, 1998.

[12] Fliess, M. et H. Mounier: *Tracking Control and π-Freeness of Infinite Dimensional Linear Systems.* Dans Picci, G. et D.S. Gilliam (rédacteurs) : *Dynamical systems, Control, Coding and Computer Vision*, tome 258, pages 41–68. Birkhäuser, Bâle, 1999.

[13] Fliess, M. et H. Mounier: *On a class of linear delay systems often arising in practice.* Kybernetica, 37 :295–308, 2001.

[14] Fliess, M., H. Mounier, P. Rouchon et J. Rudolph: *Systèmes linéaires sur les opérateurs de Mikusiǹski et commande d'une poutre flexible.* Dans *ESAIM Proc. "Élasticité, viscoélasticité et contrôle optimal", huitièmes entretiens du centre Jacques Cartier*, Lyon, 1996.

[15] Fliess, M., H. Mounier, P. Rouchon et J. Rudolph: *Controlling the transient of a chemical reactor : A distributed parameter approach.* Dans *Proc. of IEEE Conference on Computational Engineering in Systems Applications*, Nabeul-Hammamet, Tunisie, 1998.

[16] Fliess, M., H. Mounier, P. Rouchon et J. Rudolph: *A distributed parameter approach to the control of a tubular reactor : a multi-variable case.* Dans *Proc. of 37^{th} Conference on Decision and Control*, pages 439–442, Tampa, FL, États-Unis, 1998.

[17] Glüsing-Lüerßen, H.: *A behavioral approach to delay differential systems.* SIAM J. Contr. Opt., 35 :480–499, 1997.

[18] Grothendieck, A. et J.A. Dieudonné: *Eléments de géométrie algébrique I.* Numéro 166 dans *Grundlehren math. Wissensch.* Springer-Verlag, Berlin, 1971.

[19] Hörmander, L.: *The Analysis of Linear Partial Differential Operators II : Differential Operators with Constant Coefficients*, tome 257 de *Grundlehren der mathematischen Wissenschaften.* Berlin, Heidelberg, New York, 2. édition, 1990.

[20] Komatsu, H.: *Ultradistributions II : The kernel theorems and ultra distributions with supports in a submanifold.* J. Fac. Sci. Tokyo, 24 :607–624, 1977.

[21] Komatsu, H.: *Operational Calculus and Semi-groups of Operators*, tome 1540 de *Lect. Notes Math.*, pages 213–234. Springer, Berlin, 1991.

[22] Komatsu, H.: *Solution of differential equations by means of Laplace hyperfunctions*, pages 227–252. World Sci. Publishing, River Edge, NJ, 1996.

[23] Lafon, J.P.: *Algèbre commutative. Langages géométrique et algébrique.* Hermann, Paris, 1977.

[24] Laroche, B.: *Extension de la notion de platitude à des systèmes décrits par des équations aux dérivées partielles linéaires.* Tèse de doctorat, École Nationale Supérieure des Mines de Paris, 2000.

[25] Laroche, B., P. Martin et P. Rouchon: *Motion planning for the heat equation.* Int. J. Robust and Nonlinear Control, 10 :629–643, 2000.

[26] Lynch, A. et J. Rudolph: *Flatness based boundary control of a nonlinear parabolic equation modelling a tubular reactor.* Dans A. Isidori, F. Lamnabhi-Lagarrigue, W. Respondek (rédacteur) : *Nonlinear Control in the Year 2000, volume 2*, numéro 259 dans *Lecture Notes in Control and Inform. Sci.*, pages 45–54. Springer, Londres, 2000.

[27] Matsumura, H.: *Commutative Ring Theory.* Cambridge university press, Cambridge, 1990.

[28] McDonald, B.R.: *Linear Algebra over Commutative Rings.* Marcel Dekker, New York, 1984.

[29] Mikusiński, J.: *Operational Calculus.* Pergamon Press, Oxford, 1959.

[30] Mounier, H.: *Propriétés structurelles des systèmes linéaires à retards : aspects théoriques et pratiques.* Thèse de doctorat, Université Paris Sud, Orsay, 1995.

[31] Mounier, H.: *Algebraic interpretations of the spectral controllability of a linear delay system.* Forum Math., 10 :39–58, 1998.

[32] Mounier, H. et M. Fliess: *An algebraic framework for infinite dimensional linear systems.* e-sta, 1, 2004. URL : http ://www.e-sta.see.asso.fr.

[33] Mounier, H., P. Rouchon et J. Rudolph: *Some examples of linear systems with delays.* J. Europ. Syst. Autom., 31 :911–925, 1997.

[34] Mounier, H. et J. Rudolph: *Time delay systems.* Encyclopaedia of Life and Support Systems, 6.43.19.4, 2003.

[35] Ollivier, F. et A. Sedoglavic: *A generalization of flatness to nonlinear systems of partial differential equations. Application to the command of a flexible rod.* Dans *IFAC Symposium NOnLinear COntrol Systems*, pages 196–200, Saint-Petersbourg, 2001.

[36] Quillen, D.: *Projective modules over polynomial rings.* Inv. Math., 36 :167–171, 1976.

[37] Ramis, J.P.: *Séries divergentes et théories asymptotiques.* Soc. Math. France, Marseille, 1993.

[38] Rodino, L.: *Linear Partial Differential Operators in Gevrey Spaces.* World Scientific, Singapore, 1993.

[39] Rotman, J.: *An Introduction to Homological Algebra.* Academic Press, Orlando, 1979.

[40] Rowen, L.H.: *Ring Theory. Student Edition.* Academic Press, Boston, 1991.

[41] Rudolph, J.: *Beiträge zur flachheitsbasierten Folgeregelung linearer und nichtlinearer Systeme endlicher und unendlicher Dimension.* Shaker Verlag, 2003. ISBN 3-8322-1765-7. Thèse d'habilitation.

[42] Rudolph, J.: *Flatness Based Control of Distributed Parameter Systems.* Steuerungs- und Regelungstechnik. Shaker Verlag, Aachen, 2003.

[43] Rudolph, J., J. Winkler et F. Woittennek: *Flatness Based Control of Distributed Parameter Systems : Examples and Computer Exercices from Various Technological Domains.* Steuerungs- und Regelungstechnik. Shaker Verlag, Aachen, 2003.

[44] Rudolph, J. et F. Woittennek: *Motion planning and open loop control design for linear distributed parameter systems with lumped controls.* Int. J. Control, 81 :457–474, 2008.

[45] Suslin, A.A.: *Projective modules over a polynomial ring are free (in russian).* Dokl. Akad. Nauk. S.S.S.R., 229 :1063–1066, 1976. English translation : *Soviet Math. Dokl.*, **17**, p. 1160–1164.

[46] Vidyasagar, M.: *Control System Synthesis. A Factorization Approach.* MIT Press, Cambridge, Massachusetts, États-Unis, 1985.

[47] Woittennek, F.: *Beiträge zum Steuerungsentwurf für lineare, örtlich verteilte Systeme mit konzentrierten Stelleingriffen.* Berichte aus der Steuerungs- und Regelungstechnik. Shaker Verlag, Aachen, 2007.

[48] Youla, D.C. et G. Gnavi: *Notes on n-Dimensional System Theory.* IEEE Trans. Circuits Syst., 26 :105–111, 1979.

7.A Rappels d'algèbre

Module, liberté, torsion

On considère dans cette section un anneau R commutatif, unifère et sans diviseur de zéro.

Un *R–module* M est un groupe commutatif muni d'une action sur R, c'est-à-dire d'une application $R \times M \to M$, écrite $(r, m) \mapsto rm$, telle que, pour tous $r, s \in R$ et $m, n \in M$, l'on ait :

$$\begin{aligned} r(sm) &= (rs)m && \text{(associativité)} \\ r(m+n) &= rm + rn && \\ (r+s)m &= rm + sm && \text{(distributivité)} \\ 1m &= m && \text{(identité)} \end{aligned}$$

Notation 7.A.1. Le sous-module engendré par un sous ensemble S d'un R-module M est noté $[S]_R$ ou bien $[S]$ s'il n'y a pas d'ambiguïté.

Soit M un R–module, et S un sous-ensemble de M. Une *combinaison linéaire* d'éléments de S est une somme

$$\sum_{m \in S} a_m m$$

où $\{a_m\}$ est un ensemble d'éléments de R, presque tous nuls. Ces éléments sont les *coefficients* de la combinaison linéaire. L'ensemble N de toutes les combinaisons linéaires d'éléments de S est un sous-module de M, le sous-module *engendré* par S et l'on nomme S l'ensemble des *générateurs* de N.

Un module est dit *finiment engendré*, ou de *type fini* s'il ne possède qu'un nombre fini de générateurs.

Un sou-ensemble S d'un R–module M est dit *linéairement indépendant* (sur R) si, lorsque l'on a une combinaison linéaire de la forme

$$\sum_{m \in S} a_m m$$

qui est égale à zéro, alors $a_m = 0$ pour tout $m \in S$. Un sous-ensemble est dit *linéairement dépendant* s'il n'est pas linéairement indépendant.

Un module est dit *libre* s'il contient une *base*, c.à.d., un sous-ensemble indépendant et générateur.

Un élément m non nul d'un R–module M est dit de *torsion* s'il existe $a \in R$, $a \neq 0$ tel que

$$am = 0$$

En d'autres termes, l'ensemble $\{m\}$ est linéairement dépendant.

Un module M est dit de *torsion* si tous ses éléments le sont. Il est dit *sans torsion* si aucun de ses éléments non nul ne l'est.

Relations

Soit Λ un R-système. Il existe une suite exacte de R-modules [39]

$$0 \to N \to F \to \Lambda \to 0 \tag{7.38}$$

où F est libre. Le R-module N, parfois nommé *module des relations*, peut être envisagé comme un système d'équations définissant Λ.

Une *présentation libre* de Λ [39] est une suite exacte de R-modules

$$F_1 \to F_0 \to \Lambda \to 0$$

où F_0 et F_1 sont libres. Le R-module Λ est dit *finiment engendré*, ou de *type fini*, s'il existe une présentation libre où toute base de F_0 est finie. Il est dit *finiment présenté* s'il existe une présentation libre où toute base de F_0 et de F_1 est finie. La matrice correspondant à l'application $F_1 \to F_0$ est dite *matrice de présentation* de Λ ; nous la noterons P_Λ. Remarquons que cette matrice dépend des relations de Λ.

Exemple 7.A.1. Déterminons le R-module Λ correspondant au système d'équations R-linéaires

$$\sum_{\kappa=1}^{\mu} a_{\iota\kappa}\xi_\kappa = 0, \quad a_{\iota\kappa} \in A, \iota = 1, \dots, \nu$$

où $\xi_1, \dots, \xi_\mu$ sont les inconnues. Soit F le R-module type engendré par $f_1, \dots, f_\mu$. Soit $N \subseteq F$ le module des relations, c.à.d., le sous-module engendré par les combinaisons $\sum_{\kappa=1}^{\mu} a_{\iota\kappa} f_\kappa$, $\iota = 1, \dots, \nu$. Alors, $\Lambda = F/N$. Les ξ_κ sont les *résidus* des f_κ, c.à.d., les images canoniques des f_κ.

Considérons, par exemple, le système considéré à la remarque 7.3.2. Le $\mathbb{R}[s]$-module libre F est celui engendré par X et U (c.à.d. l'ensemble $\{p(s)\,X + q(s)\,U \mid p, q \in \mathbb{R}[s]\}$ équipé d'une structure linéaire). Le sous-module N des relations est engendré par l'élément $a(s)\,X - b(s)\,U$ (c.à.d. l'ensemble $\{r(s)(a(s)\,X - b(s)\,U) \mid r \in \mathbb{R}[s]\}$ équipé d'une structure linéaire). Alors $\Lambda = F/N = [X, U]/[a(s)\,X - b(s)\,U]$, avec x et u pour générateurs, les résidus de X et U, et pour relation $a(s)\,x = b(s)\,u$.

Idéal

Nous allons introduire un objet algébrique associé à un module M qui regroupe les mineurs de la matrice de présentation P_M et qui, contrairement à cette dernière, *ne dépend que du module M*. Pour la notion de matrice de présentation, matrice des équations du système, voir la sous-section 7.A, p. 325.

Un *idéal* $\mathfrak{a}$ de R est un sous-groupe additif de R tel que pour tout $\alpha \in \mathfrak{a}$ et $r \in R$, $r\,\alpha \in \mathfrak{a}$. C'est donc un sous-ensemble de R qui est également un R-module. Étant donné un idéal $\mathfrak{a}$ de R, s'il existe une famille d'éléments $(\alpha_i)_{i\in I}$ ($I \subseteq \mathbb{N}$) de $\mathfrak{a}$ tels que

$$\mathfrak{a} = \Big\{ \sum_{i\in I} \alpha_i p_i \mid p_i \in R \Big\}$$

l'idéal $\mathfrak{a}$ est dit *engendré par la famille* $(\alpha_i)_{i\in I}$. Lorsque I est fini, $\mathfrak{a}$ est dit *finiment engendré*.

Avec M donné par générateurs et relations comme décrit en section 7.A, p. 325, l'*idéal de Fitting d'ordre i* associé à M (voir [28] ou [8, définition 20.4 p. 493]), noté $\mathfrak{I}^i_M$, est défini comme l'idéal de R engendré par les déterminants de toutes les sous-matrices de taille $(\alpha-i)\times(\alpha-i)$ (les mineurs d'ordre $\alpha-i$) de P_M. Supposant que $\mathrm{rg}_R P_M = \gamma$, nous noterons $\mathfrak{I}_M$ l'idéal de Fitting associé aux mineurs d'ordre γ P_M. Cet idéal admet $C^\gamma_\alpha = \alpha!/(\gamma!(\alpha-\gamma)!)$ générateurs. Dans le cas où R est un anneau de polynômes en r indéterminées, nous considérerons en général – par abus de notation – les éléments de $\mathfrak{I}_M$ comme des fonctions polynomiales de $\mathbb{C}^r$.

Produit tensoriel

Le produit tensoriel est une technique qui permet, entre autres, d'étendre formellement l'anneau des scalaires agissant sur un module. Soit R un anneau, M et N deux R-modules. Notons F le groupe des combinaisons R-linéaires de toutes les paires ordonnées (m,n). Soit L le sous-groupe de F engendré par tous les éléments de la forme

$$\begin{aligned} &(m+m',n)-(m,n)-(m',n), \quad &&(am,n)-a(m,n)\\ &(m,n+n')-(m,n)-(m,n'), \quad &&(m,an)-a(m,n) \end{aligned}$$

où $m,m' \in M$, $n,n' \in N$ et $a \in R$. On pose alors $M \otimes_R N = F/L$; il s'agit d'un R-module que l'on nomme *produit tensoriel des modules M et N*. Les éléments de $M \otimes_R N$ s'écrivent $\sum_{\text{finie}} m_i \otimes_R n_j$ (ou $\sum_{\text{finie}} m_i \otimes n_j$ lorsqu'il n'y a pas de risque de confusion) et ce de manière non unique en général. Un des avantages de cette construction est de pouvoir réduire des applications bilinéaires sur $M \times N$ à des applications linéaires sur $M \otimes_R N$.

Exemple de changement d'anneau de base : extension des scalaires

Nous allons donner, dans cette section et la suivante, deux exemples d'une opération désormais classique en algèbre commutative et géométrie algébrique (voir [18] et par exemple [23]) : celle du changement d'anneau de base. Elle permet notamment de changer le point de vue que l'on a sur un objet algébrique. Un premier exemple est l'extension des scalaires. Soient R et S deux anneaux

avec $R \subseteq S$, S étant un R-module et M un R-module. Alors on dit que le S-module $S \otimes_R M$ est obtenu à partir de M par *extension des scalaires.*

Exemple de changement d'anneau de base : localisation

Un deuxième exemple de changement d'anneau de base est fourni par le procédé de localisation. Soit $\mathcal{S}$ un sous-ensemble multiplicativement clos de R (c'est-à-dire pour tous $\pi_1, \pi_2 \in \mathcal{S}$, $\pi_1\pi_2 \in \mathcal{S}$ et $1 \in \mathcal{S}$). Le *localisé en* $\mathcal{S}$ *de* R est un anneau, noté $\mathcal{S}^{-1}R$, muni d'un morphisme $\phi : R \longrightarrow \mathcal{S}^{-1}R$ tel que

(i) pour tout $\pi \in \mathcal{S}$, $\phi(\pi)$ est inversible dans $\mathcal{S}^{-1}R$;

(ii) pour tout $q \in \mathcal{S}^{-1}R$, il existe des $p \in R$ et $\pi \in \mathcal{S}$ tels que $q = \phi(p)\phi(\pi)^{-1}$. On note généralement l'élément q précédent par p/π.

Projectivité

Définition

Une caractérisation de la commandabilité des systèmes linéaires de dimension finie est la liberté du module sous-jacent (voir [10]), qui repose sur l'équivalence entre la liberté et l'absence de torsion pour un module sur l'anneau[6] $\mathbb{R}[\frac{d}{dt}]$. Dans le cas des systèmes à paramètres répartis, cette caractérisation n'est plus valable, *la liberté d'un R-module M étant une notion plus forte que le fait d'être sans torsion* (l'implication inverse de la suivante est fausse) :

$$M \text{ libre } \implies M \text{ sans torsion}$$

Il faut en fait distinguer deux propriétés : la liberté et la projectivité. Cette dernière recouvre la notion de "sous-espace" d'un module libre. D'une part, un sous-module M d'un module libre N peut-être libre mais pas nécessairement un terme en somme directe de N. D'autre part, il peut y avoir des termes en somme directe de N qui ne sont pas libres. C'est ce dernier concept qui fournit une généralisation naturelle de la notion de "sous-espace". Un module M sur un anneau commutatif R est donc dit *projectif*, si $N = M \oplus \widetilde{M}$ où N est un module libre (voir par exemple [28] ou [8, A3.2 p. 615]) ; $\widetilde{M}$ est alors également projectif pour la même raison.

Critères de projectivité

Nous avons le résultat suivant, sur un anneau commutatif R, intègre et sans diviseurs de zéros :

Proposition 7.A.1. *Un R-module M est projectif si, et seulement si son idéal de Fitting $\mathfrak{I}_M$ est égal à R.*

[6] Ceci est dû au caractère principal de l'anneau $\mathbb{R}[\frac{d}{dt}]$.

Nous avons également le critère local suivant (voir [28, théorème IV.32 p. 295] et [8, théorème 19.2 p. 471, théorème A3.2 p. 616 et exercices 4.11 et 4.12 p. 136]) :

Proposition 7.A.2 (Critère local de projectivité)**.** *Soit M un module* finiment présenté *sur un anneau commutatif R. Alors M est projectif si, et seulement s'il existe une famille finie d'éléments $x_1, \ldots, x_r$ de R qui génèrent l'idéal unité de R, telle que $M[x_i^{-1}]$ soit libre sur $R[x_i^{-1}]$ pour tout i.*

Dans le cas où R est un anneau de fonctions entières d'une variable complexe, et si *le Nullstellensatz est vrai sur R*, nous avons également

Proposition 7.A.3. *Pour un R-module M, $\mathfrak{I}_M = R$ si, et seulement si les générateurs de $\mathfrak{I}_M$ n'ont pas de zéro commun dans $\mathbb{C}$.*

Une caractérisation de la projectivité est alors la suivante : M est projectif si une matrice de présentation P_M ($P_M \in R^{\beta\times\alpha}, \mathrm{rg}_R P_M = \beta$) est inversible à droite (s'il existe $Q \in R^{\alpha\times\beta}$ tel que $P_M Q = I_\beta$).

Notons que cette caractérisation est directement liée aux équations de Bézout matricielles. Si l'on prend une dynamique Λ d'équations

$$\dot{\boldsymbol{x}} = F\boldsymbol{x} + G\boldsymbol{u}$$

avec F et G des matrices à coefficients dans R de tailles appropriées, la projectivité de Λ revient à l'existence de matrices $\overline{F}$ et $\overline{G}$ à coefficients dans R telles que

$$\left[\frac{d}{dt} I_n - F\right] \overline{F} + G\overline{G} = I_n.$$

Ce type d'équation peut servir à des fins de stabilisation par retour d'état dynamique (voir [46]).

Projectivité et liberté

Rappelons ici le célèbre théorème de Quillen et Suslin résolvant en 1976 une conjecture émise par Serre dans les années cinquante.

Théorème 7.A.1. *Tout module projectif sur un anneau de polynômes ($k[X_1, \ldots, X_n]$ où k est un corps) est libre.*

Définition 7.A.1. Un module est *stablement libre* s'il devient libre après ajout d'un module libre.

Proposition 7.A.4. *Tout module stablement libre n'est pas nécessairement libre ; un contre exemple classique est sur $k[X, Y, Z]/(1 - X^2 - Y^2 - Z^2)$.*

7.B Rappels sur les séries divergentes et les fonctions Gevrey

Sommabilité et multi-sommabilité

Voir par exemple [37], [2], [3], [4]. Considérons une série formelle en x

$$w(x,\tau) = \sum_{i\geqslant 0} \alpha_i(\tau)\, x^i$$

à coefficients des fonctions $a_i(\tau)$, holomorphes dans un disque

$$D_\rho = \{\tau \in \mathbb{C} \mid |\tau| < \rho\}$$

Cette série est élément de $C^\omega(\mathbb{C})[[x]]$.
Lorsque ces séries seront utilisées avec des $a_i(t)$ fonctions du temps, les coefficients seront des fonctions analytiques réelles.

Sommabilité

Définition 7.B.1. On dira que la série $w(x,\tau)$ est *k-sommable dans la direction d* ($k > 0$, $d \in \mathbb{R}$) s'il est possible de trouver un rayon de convergence $r \in]0,\rho[$ tel que les deux propriétés suivantes soient satisfaites :

- La transformée de Borel en x d'ordre k, c.à.d. la série

$$\widetilde{w}(z,\tau) = \mathscr{B}_k(w(x,\tau)) = \sum_{i\geqslant 0} a_i(\tau)\, \frac{z^i}{\Gamma(1+i/k)} \tag{7.39}$$

  converge absolument pour $|\tau| \leqslant r$ et $|z| < R$, avec R dépendant de r mais indépendant de τ.
- Il existe δ tel que, pour tout $\tau \in \bar{D}_r$, la fonction $\widetilde{w}(z,\tau)$ peut être prolongée analytiquement en x dans le secteur $S_{d,\delta} = \{x : |d - \arg z| < \delta\}$, prolongement que l'on note $\mathscr{C}_{S_{d,\delta}}(\widetilde{w})(z,\tau)$. Ce prolongement est de plus borné par une exponentielle d'ordre k dans un tout sous secteur : pour tout $\delta_1 < \delta$, il existe des constantes $C, K > 0$ telles que

$$\forall x \in S_{d,\delta_1}, \quad \sup_{|\tau|\leqslant r} |\mathscr{C}_{S_{d,\delta}}(\widetilde{w})(z,\tau)| \leqslant C \exp\left(K|z|^k\right)$$

Dans ce cas, la transformée de Laplace $\mathscr{L}_k$ d'ordre k de $\mathscr{C}_{S_{d,\delta}}(\widetilde{w})(z,\tau)$, c.à.d. la fonction

$$\mathscr{S}_{k,d}(w(x,\tau)) = \mathscr{L}_k(\mathscr{C}_{S_{d,\delta}}(\mathscr{B}_k(w(x,\tau)) = x^{-k} \int_0^{\infty(\gamma)} w(\xi,\tau) e^{-(\xi/x)^k} d\xi$$

intégrant le long du rayon $\arg \xi = \gamma$ avec $|d-\gamma| < \delta$ est nommée *k-somme de la série formelle* $w(x,\tau)$, et notée

$$\mathsf{w}(x,\tau) = \mathscr{S}_{k,d}\, w(x,\tau)$$

Le procédé de k-sommation est donc réalisé en trois étapes :

$$\mathscr{S}_{k,d} = \mathscr{L}_k \circ \mathscr{C}_{S_{d,\delta}} \circ \mathscr{B}_k$$

Remarque 7.B.1. Notons que la transformée de Laplace d'ordre 1 de z^n est

$$\mathscr{L}_1(z^n) = x^{-1} \int_0^{\infty(\gamma)} \xi^n e^{-\xi/x} d\xi = n! x^n$$

qui apparaît donc comme un opérateur accélérant la divergence d'une série formelle. À l'inverse, la transformée de Borel accélère la convergence de cette même série formelle. Le procédé de sommation de Borel consiste donc à accélérer la convergence, ce qui permet d'obtenir une fonction analytique, dont on peut construire un prolongement analytique. La transformée de Borel inverse, c.à.d. la transformée de Laplace, permet de revenir dans le domaine initial.

Exemple 7.B.1. Ce procédé permet de sommer des séries de la forme $\sum_0^\infty i! x^i$. La transformée de Borel correspondante est $1/(1-z)$ qui possède les propriétés requises dans toutes les directions d, sauf l'axe réel positif. Cette série est donc 1-sommable dans toute direction $d \not\equiv 0$ modulo 2π.

Multi-sommabilité

Pour les solutions de certaines EDOs et de certaines EDPs (comme $(\partial_t - \partial_z^2)(\partial_t - \partial_z^3)$), le procédé de sommabilité ci-dessus ne suffit pas. La première des conditions de la définition précédente peut être relaxée en imposant que la transformée de Borel (7.39) ne soit plus convergente, mais sommable en un certain sens. Nous utiliserons les notations suivantes : soit $q \geqslant 2$ un entier naturel ; un q-uple $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)$ où les κ_i sont des réels positifs sera nommé un *type de multisommabilité* ; le q-uple $\boldsymbol{d} = (d_1, \ldots, d_q)$ où les d_i sont des réels sera nommé une *multidirection admissible* si

$$2\kappa_j |d_j - d_{j-1}| \leqslant \pi, \quad 2 \leqslant j \leqslant q$$

On itère donc la définition précédente de la manière suivante :

Définition 7.B.2. Donnons nous un type de multisommabilité $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_q)$ et une multidirection admissible $\boldsymbol{d} = (d_1, \ldots, d_q)$ et considérons une série formelle $w(x,\tau) = \sum_{i \geqslant 0} \alpha_i(\tau)\, x^i$. On dira que $w(x,\tau)$ est $\boldsymbol{\kappa}$*-sommable dans la multidirection* $\boldsymbol{d}$ si :

– La série

$$\sum_{i \geqslant 0} a_i(\tau)\, \frac{z^i}{\Gamma(1 + i/\kappa_1)}$$

est $(\kappa_2, \ldots, \kappa_q)$-sommable dans la multidirection $(d_2, \ldots, d_q)$ et l'on notera $\widetilde{w}(z,\tau)$ sa somme.

– Il existe δ tel que, pour tout $\tau \in \bar{D}_r$, la fonction $\widetilde{w}(z,\tau)$ peut être prolongée analytiquement en x dans le secteur $|d_1 - \arg z| < \delta$, et vérifie pour des constantes $C, K > 0$ suffisamment grandes

$$|\widetilde{w}(z,\tau)| \leqslant C \exp\left(K|z|^k\right)$$

pour tout z comme ci-dessus.

Dans ce cas, la fonction

$$x^{-\kappa_1} \int_0^{\infty(\gamma)} w(\xi,\tau) e^{-(\xi/x)^{\kappa_1}} d\xi$$

est nommée $\boldsymbol{\kappa}$*-somme de la série formelle* $w(x,\tau)$, *dans la multidirection* $\boldsymbol{d}$ et notée

$$\mathsf{w}(x,\tau) = \mathscr{S}_{\boldsymbol{\kappa},\boldsymbol{d}}\, w(x,\tau)$$

Fonctions Gevrey

Voir par exemple [37], [2], [3], [4]. Une fonction de classe C^∞ $f(t)$ de $[0,T]$ dans $\mathbb{R}$ est dite *Gevrey* d'indice d si elle vérifie les estimées suivantes

$$\sup_{t\in[0,T]} |f^{(i)}(t)| \leqslant C K^i \Gamma(1+(d+1)i), \quad \forall i \geqslant 0$$

avec des constantes C et $K > 0$.

De manière analogue, on dira qu'une série formelle

$$F(x,s) = \sum_{i\leqslant 0} f_i(x) \frac{s^i}{i!}, \quad x \in [0,r)$$

est Gevrey d'ordre d s'il existe des constantes $\rho \in [0,r)$, C, $K > 0$ telles que

$$|f_i(x)|(t) \leqslant C K^i \Gamma(1+(d+1)i), \quad \forall i \geqslant 0, |x| < \rho$$

7.C Représentation des opérateurs $S(x)$ et $C(x)$

Si $a > 0$ dans l'équation (7.18b) on peut réécrire σ comme

$$\sigma = \tau^2 \left((s+\alpha)^2 - \beta^2\right), \quad \tau = \sqrt{a}, \quad \alpha = \frac{b}{2a}, \quad \beta = \sqrt{\frac{b^2}{4a^2} - \frac{c}{a}}.$$

L'opérateur $S(x)$ correspond à la fonction à support compact

$$S(x,t) = \big(H(t+x\tau) - H(t-x\tau)\big) \frac{\mathrm{e}^{-\alpha t}}{2\tau} J_0(\beta\sqrt{\tau^2x^2 - t^2}),$$

où J_0 désigne la fonction de Bessel d'ordre zéro et H la distribution de Heaviside. Par contre, si $a = 0$ dans (7.18b), $S(x)$ peut s'écrire

$$S(x) = \sum_{k=0}^{\infty} \frac{(bs+c)^k x^{2k+1}}{(2k+1)!}.$$

8 Platitude différentielle et commande de systèmes non linéaires : une introduction à l'approche par l'algèbre différentielle

J. Rudolph[1]

[1]Institut für Regelungs- und Steuerungstheorie, Technische Universität Dresden, 01062 Dresden, Allemagne. *E-mail* : Joachim.Rudolph@tu-dresden.de

La platitude différentielle des systèmes non linéaires est caractérisée par la possibilité de paramétrer la solution par un ensemble fini de trajectoires indépendantes, pour la sortie plate [26, 40, 25]. Il en résulte des méthodes simples et efficaces pour la planification de trajectoires et pour leur stabilisation. On parle aussi de « poursuite » dans ce dernier cas. (Pour une introduction on peut aussi consulter [26, 40, 25, 52, 53, 45, 46, 47, 50, 57], par exemple.)

Le cadre algébrique de la théorie des corps différentiels se prête très bien à l'étude des systèmes non linéaires (de dimension finie, algébriques) et en particulier à celle des systèmes plats.

Ces systèmes sont décrits par des systèmes d'équations différentielles ordinaires, explicites ou non. Le cadre algébrique permet de les traiter sans introduction explicite de coordonnées, en les comprenant comme extensions de corps différentiels (ordinaires). L'emploi de cette approche aux systèmes non linéaires fut introduit en théorie du contrôle par M. Fliess au milieu des années 1980 (voir par ex. [16, 21]). La base mathématique est fournie par l'algèbre différentielle au sens de J. F. Ritt [44] — voir aussi [38, 36, 56, 1] par exemple.

8.1 Systèmes plats

Le point de départ de nos considérations est un système (implicite) d'équations différentielles ordinaires (é.d.o.). On les supposera algébriques, c'est-à-dire polynômiales (voir toutefois la remarque à la page 335 à ce sujet).

Si le système comprend s variables, $w_1, \dots, w_s$, ces équations s'écrivent donc

$$P_i(w_1, w_2, \dots, w_s, \dot{w}_1, \dots, w_j^{(l)}, \dots, w_s^{(\alpha)}) = 0, \quad i = 1, \dots, q,$$

ou bien, plus brièvement,

$$P_i\left(w, \dots, w^{(\alpha)}\right) = 0, \quad i = 1, \dots, q. \tag{8.1}$$

Ici les expressions $P_i, i = 1, \dots, q$, sont des polynômes en $w_i, i = 1, \dots, s$, et leurs dérivées (ordinaires) ; les coefficients appartiennent à un corps différentiel approprié, k. Ces polynômes sont donc des éléments d'un anneau différentiel $k\{w\}$ engendré sur k par la famille $w = (w_1, \dots, w_s)$ (voir #1). Pour simplifier on suppose que le corps de base k soit un corps de constantes comprenant $\mathbb{Q}$.

Système : Un *système algébrique (continu et de dimension finie)* est une extension de corps différentielle, de type fini, Σ/k.

On observera que cette définition ne nécessite l'introduction ni de variables (ou coordonnées) ni d'équations. Toutefois, comme il s'agit d'une extension de corps différentielle *de type fini* (voir #14), on sait que des équations algébriques (donc polynômiales) en un nombre fini de variables existent. Autrement dit, on peut choisir dans Σ un nombre fini de *variables du système*, $w_1, \dots, w_s$, de façon à pouvoir représenter tous les éléments de Σ comme expressions rationnelles en les w_i, $i = 1, \dots, s$, et leurs dérivées (à coefficients appartenant au corps de base k). On peut alors faire appel explicitement aux éléments choisis, qu'on collectera dans la famille $w = (w_1, \dots, w_s)$ (une famille génératrice (différentielle) de Σ/k (#14)), et écrire $\Sigma = k\langle w \rangle$.

Remarque 8.1.1. Pour préciser la nature des coefficients, il convient parfois de parler de k-système (algébrique) [10].

Comment construire l'extension Σ/k à partir d'un système d'é.d.o. (8.1) ? Une première approche consiste à considérer le corps différentiel $\Sigma = k\langle w \rangle$, et à supposer qu'il soit défini de telle façon à ce que les relations entre les w_i, $i = 1, \dots, s$, soient juste données par les équations en considération. Ces relations existent dans le corps Σ si les éléments appropriés sont nuls. Les éléments de Σ étant des expressions rationnelles en w_i, $i = 1, \dots, s$, et leurs dérivées (à coefficients dans k) on obtient (par multiplications par les polynômes dénominateurs) les équations de la forme (8.1).

Un tel corps différentiel ne peut être associé à tout système d'é.d.o. polynômial, une condition supplémentaire étant requise : Σ étant un corps il ne

peut contenir des diviseurs de zéro, une condition de primalité en résulte. Pour construire le système $k\langle w\rangle/k$ supposons donné des équations de la forme (8.1), et prenons une famille $W = (W_1, \ldots, W_s)$ (de cardinal égal à celui de w) et l'anneau différentiel $k\{W\}$ (libre, c'est-à-dire sans relations non triviales) engendré sur k par W. Les éléments de $k\{W\}$ sont les polynômes en les indéterminées W_i, $i = 1, \ldots, s$, et leurs dérivées, à coefficients dans k. (Les indéterminées sont différentiellement algébriquement indépendantes sur k (#17).) Soient alors $P_j, j = 1, \ldots, q$, les polynômes de l'anneau différentiel $k\{W\}$ que l'on obtient en remplaçant dans les membres gauches des é.d.o. (8.1) les variables du système, $w_i, i = 1, \ldots, s$, par les indéterminées W_i. Si, et seulement si, l'idéal différentiel I dans $k\{W\}$ engendré par les $P_j, j = 1, \ldots, q$ est premier, l'anneau des fractions de l'anneau quotient (des classes de résidus) $k\{W\}/I$ ne contient aucun diviseur de zéro, et forme donc un corps différentiel, extension de k (comparer à #16 et voir aussi la remarque à la page 337). Or, comme les images canoniques w_i résultent des W_i en localisant, on a $\Sigma = k\langle w\rangle$; et les équations satisfaites par les éléments de Σ correspondent à (8.1). On se rend compte par cette construction que l'extension de corps différentielle Σ/k est indépendante du choix des générateurs w et de celui des équations de départ. Ce qui est important, c'est l'idéal différentiel I, et non pas ses générateurs.

Exemple 8.1.1. La construction de l'extension de corps différentielle d'un modèle de grue sera détaillée dans 8.12 (p. 357).

Exemple 8.1.2. L'oscillation d'un pendule mathématique est décrite par l'é.d.o.

$$\ddot{\varphi} + \sin\varphi = 0.$$

Au lieu de cette représentation non algébrique (explicite) on peut aussi utiliser une représentation en coordonnées cartésiennes, qui elle est algébrique (mais implicite). Avec $x = \sin\varphi$ et $y = \cos\varphi$ il vient (pour $y \neq 0$, donc $\varphi \neq \pm\pi/2$) :

$$\begin{aligned} y^2\ddot{x} + xy^3 + x\dot{x}^2 &= 0 \\ x^2 + y^2 - 1 &= 0. \end{aligned}$$

Finalement on peut aussi (pour $\varphi \neq \pm\pi$) introduire $z = \tan(\varphi/2)$, et obtenir la représentation algébrique (implicite)

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

Remarque 8.1.2. Comme les systèmes (algébriques) différentiels continus discutés ici, on peut aussi définir les systèmes (algébriques) discrets, décrits par des systèmes (polynômiaux) d'équations aux différences (ordinaires) [17, 19, 20]. On se sert alors des anneaux et corps aux différences.

8.2 Platitude différentielle

Soit Σ/k un système algébrique.

Platitude différentielle : Un système Σ/k est dit *(différentiellement) plat*, si à une clôture algébrique (non différentielle) près, Σ est un corps d'extension purement transcendant de k (#17). Autrement dit, Σ/k est (différentiellement) plat s'il existe une base de transcendance différentielle $y = (y_1, \ldots, y_m)$ de $\overline{\Sigma}/k$, pour laquelle $\overline{\Sigma} = \overline{k\langle y\rangle}$. Une telle famille[1] y est appelée *sortie plate* de Σ/k.

Ici, $\overline{\Sigma}$ désigne la clôture algébrique (#5) de Σ, et $\overline{k\langle y\rangle}$ celle de $k\langle y\rangle$. Le nombre, m, des composantes d'une sortie plate est (visiblement) égal au degré de transcendance différentiel de l'extension Σ/k.

Considérons les relations entre les variables du système et la sortie plate y — soit $\Sigma = k\langle w\rangle$ à cette fin. On déduit alors de la définition de y :

1. Les composantes $y_i, i = 1, \ldots, m$, d'une sortie plate y satisfont des relations de la forme

$$Q_i(y_i, w, \dot{w}, \ldots, w^{(\alpha_i)}) = 0, \quad i = 1, \ldots, m,$$

avec $Q_i \in k\{w\}[y_i]$ des polynômes, qu'on peut résoudre (localement)[2] par rapport aux y_i :

$$y_i = \phi_i(w, \dot{w}, \ldots, w^{(\alpha_i)}), \quad i = 1, \ldots, m.$$

Ceci découle directement de $y_i \in \overline{k\langle w\rangle}, i = 1, \ldots, m$.

2. Il n'existe aucune relation (non triviale) de la forme

$$R(y, \ldots, y^{(\beta)}) = 0,$$

où R est un polynôme de l'anneau différentiel $k\{y\}$. Ceci équivaut à deg tr diff $k\langle y\rangle/k = m$.

3. Toute variable du système, $z \in \Sigma$, s'exprime en fonction de y et ses dérivées, car $z \in \overline{\Sigma} = \overline{k\langle y\rangle}$ implique

$$S(z, y, \ldots, y^{(\gamma)}) = 0,$$

avec un polynôme $S \in k\{w\}[z]$, d'où la relation locale

$$z = \psi(y, \ldots, y^{(\gamma)}).$$

[1] On observera que y joue un rôle analogue à celui d'une base d'un module libre (correspondant à un système commandable dans la théorie de [18, 52]). Un nom plus approprié pour y serait peut-être « paramètre fondamental » pour ne pas parler de base dans ce contexte non linéaire, mais le nom « sortie plate » (et le symbole y) sont bien établis dans la littérature.

[2] Comprendre le terme « locale » ici (et dans la suite) dans le sens que $\partial Q_i/\partial y_i \neq 0$, de sorte que l'on peut appliquer le théorème des fonctions implicites dans un contexte mathématique adapté.

Par la seconde de ces propriétés, il est garanti que les trajectoires des composantes de la sortie plate y peuvent être choisies indépendamment et librement (dans le sens qu'ils ne doivent satisfaire aucune équation différentielle particulière). La troisième propriété implique la possibilité de calculer les trajectoires de toutes les variables à partir de celles de y, sans être obligé d'intégrer une équation différentielle ; il suffit de les dériver. On peut donc résumer : *Un système (différentiellement) plat est complètement, finiment et librement différentiellement paramétrisable.*

Exemple 8.2.1. La platitude du modèle d'une grue sera discuté en section 8.12 (p. 358). Cet exemple illustrera aussi l'utilité de la clôture algébrique dans la définition de la platitude (voir la remarque à la page 359).

Remarques 8.2.1. 1. Observons que la définition de la « platitude » est basée directement sur le corps Σ, sans en distinguer les générateurs : en termes de théorie du contrôle, elle ne fait pas appel au concepts d'entrée ou d'état.
2. Il convient parfois de généraliser la définition en considérant un autre corps de base. On définit ainsi la $\mathcal{D}$-platitude [10] en demandant à ce que $\overline{\mathcal{D}\langle y\rangle} = \overline{\Sigma}$, avec $\mathcal{D}$ un sous-corps différentiel de Σ.
3. On peut généraliser le concept de platitude en admettant une transformation du temps, qui peut dépendre des variables du système ; on parle alors de systèmes *orbitalement plats* [24].
4. Une autre possibilité de généraliser la définition discutée ci-dessus consiste à lever la restriction aux fonctions algébriques, une approche par la géométrie différentielle est alors adaptée [25] (ou aussi [43]).
5. Esquissons une explication pour l'emploi du terme « plat ». Pour ceci, supposons (sans perte de généralité) que les composantes y_i, $i = 1, \dots, m$, de la sortie plate soient les m premières composantes de la famille w des variables du système, et supposons donné $q = s - m$ équations indépendantes $P_i(w, \dot{w}, \dots, w^{(\alpha)}) = 0$, $i = 1, \dots, q$. Introduisons un espace de dimension infinie de coordonnées $z_i^{(j)}$, $i = 1, \dots, s$, $j \geq 0$, et considérons (localement autour de points « réguliers ») les transformations entre les variables du système, w, et de leurs dérivées et des variables $z_i, i = 1, \dots, m + q$ données par

$$\begin{aligned} z_i &= y_i = w_i, && i = 1, \dots, m \\ z_i &= P_{i-m}(w, \dot{w}, \dots, w^{(\alpha)}), && i = m+1, \dots, s. \end{aligned}$$

Alors on obtient les équations des dérivées des $z_i, i = 1, \dots, m+q$, par dérivation. La transformation définie ainsi peut être inversée sans intégration [30] ; car des relations

$$w_i = \psi_i(y, \dot{y}, \dots, y^{(\gamma_i)}), \quad i = 1, \dots, s,$$

pour les sorties plates on déduit $\tilde{z} = (z_1, \dots, z_m)$

$$w_i = \psi_i(\tilde{z}, \dot{\tilde{z}}, \dots, \tilde{z}^{(\gamma_i)}), \quad i = 1, \dots, s.$$

Or, les q dernières composantes de z (et toutes leurs dérivées) sont égales à zéro (par les équations du système). En revanche, les m premières composantes et leurs dérivées, c'est-à-dire les composantes de $\tilde{z}$ et leurs dérivées à eux sont indépendantes. Ainsi les équations du système représentent un sous-espace (linéaire, de dimension infinie), avec les coordonnées $z_i^{(j)}$, $i = 1, \ldots, m$, $j \geq 0$, de l'espace en considération. Cet espace (linéaire) peut être interprété comme un hyper-plan, il est donc plat.

Les équations d'un système différentiellement plat peuvent toujours être données dans la forme particulière $S_i(w_i, y, \ldots, y^{(\nu_i)}) = 0$, $i = 1, \ldots, s$. Il en découle une conséquence intéressante pour la construction de son extension de corps à partir de l'idéal différentiel $[S]$. Étant donné un système de s équations de ce type, les membres gauches des équations (algébriques) $S_i = 0$ peuvent être vus comme des polynômes non différentiels en les w_i, $i = 1, \ldots, s$, avec des coefficients appartenant au corps différentiel $k\langle y\rangle$. Ceci simplifie l'analyse de l'idéal différentiel engendré par les polynômes $S = (S_1, \ldots, S_s)$ correspondants, mais en les indéterminées $W = (W_1, \ldots, W_s)$. En particulier, si S_j est de degré 1 en w_j, dans l'anneau différentiel localisé $k\langle y\rangle\{W_1, \ldots, W_s\}/[S]$ on a $w_j^{(l)} \in k\langle y\rangle$, $l \geq 0$, avec w_j l'image de W_j dans cet anneau. De même on a $w_j^{(l)} \in \overline{k\langle y\rangle}$, $l \geq 0$, si S_j peut s'écrire dans la forme $S_j = W_j^2 + c^2$, avec un $c \in k\langle y\rangle$, ou dans la forme $S_j = W_j^2 - d$, avec $d \in k\langle y\rangle$, qui ne peut être représenté comme $d = e^2$ avec $e \in k\langle y\rangle$. Si tous les membres gauches des équations du système peuvent être représentés dans une de ces formes, il est clair que l'anneau des résidus ne comprend aucun diviseur de zéro, et l'idéal $[S]$ est donc premier (#3). Des simplifications similaires peuvent être discutées pour les polynômes de degré supérieur, l'analyse étant réduite à la discussion de polynômes non différentiels en une seule indéterminée.

Exemple 8.2.2. On discutera la construction de l'extension des corps à partir d'un idéal différentiel pour l'exemple de la grue dans la section 8.12 (p. 357).

8.3 Entrées et dynamiques

Soit Σ/k un système algébrique.

Entrée : Une famille $u = (u_1, \ldots, u_m)$ d'éléments de Σ est appelée une *entrée* si l'extension de corps différentielle $\Sigma/k\langle u\rangle$ est différentiellement algébrique (#17). Une entrée u est dite *indépendante* si elle forme une base de transcendance différentielle (#20) de $k\langle u\rangle/k$.

De manière équivalente, et peut-être plus élégante, on peut dire : Une entrée indépendante est une base de transcendance différentielle de Σ/k ; alors on a $m = \text{deg tr diff } \Sigma/k$ (voir #20), où $m = \text{card}\, u$.

Si u est une entrée de Σ/k, alors pour tout élément de Σ, c'est-à-dire pour toute variable z du système, il existe une équation différentielle de la forme

$$Q(z, \ldots, z^{(\beta)}) = 0,$$

où Q est un polynôme de $k\langle u\rangle\{z\}$, c'est-à-dire un polynôme en z et ses dérivées, à coefficients dans $k\langle u\rangle$. Ceci équivaut à l'existence d'une relation

$$R(z, \ldots, z^{(\beta)}, u, \ldots, u^{(\gamma)}) = 0,$$

avec $R \in k\{z, u\}$ un polynôme (différentiel) à coefficients dans k. Réciproquement, la famille u doit former une base de transcendance différentielle pour que de telles relations existent sans qu'il n'y ait une relation (non triviale) du type

$$R(u, \ldots, u^{(\gamma)}) = 0,$$

avec $R \in k\{u\}$.

Remarques 8.3.1. 1. Souvent les entrées correspondent aux variables de commande du processus dont le système est un modèle, qui en général sont libres (ne doivent satisfaire aucune contrainte différentielle (é.d.o.) particulière). Ainsi il convient souvent de les supposer indépendantes.
2. L'existence d'équations $R = 0$ (ou $Q = 0$) montre que, une fois une trajectoire pour u fixée, celles des autres variables s'en suivent par la solution d'é.d.o.
3. Le degré de transcendance différentiel deg tr diff Σ/k, et ainsi le nombre de composantes de toute entrée, correspond au degré d'indétermination du système d'é.d.o. Il correspond à la différence entre le nombre de variables indépendantes et des équations (différentielles) indépendantes.

Dynamique : Une extension de corps différentielle $\Sigma/k\langle u\rangle$, où u forme une entrée de Σ/k, est appelée une *dynamique* d'entrée u.

Une dynamique $\Sigma/k\langle u\rangle$ est donc une extension de corps qui est différentiellement algébrique. Ainsi, le degré de transcendance différentiel de $\Sigma/k\langle u\rangle$ est nul.

Ceci est équivalent à ce que le degré de transcendance non différentiel (#12) de l'extension de corps $\Sigma/k\langle u\rangle$, que l'on désigne par deg tr $\Sigma/k\langle u\rangle$, est fini (#22). On peut s'en servir pour vérifier si une famille u particulière forme une entrée. Un cas spécial est donné si deg tr $\Sigma/k\langle u\rangle$ n'est pas seulement fini, mais égal à zéro :

Dynamique triviale : Une dynamique $\Sigma/k\langle u\rangle$ est dite *triviale*, si son degré de transcendance différentiel est égal à zéro : deg tr $\Sigma/k\langle u\rangle = 0$.

Exemple 8.3.1. Une entrée pour le système de la grue sera choisie dans la section 8.12 (p. 357).

8.4 Systèmes entrée-sortie

Soit Σ/k un système algébrique.

Sortie : Une *sortie* $y = (y_1, \ldots, y_p)$ est une famille d'éléments de Σ.

Système entrée-sortie : Si $\Sigma = k\langle u, y\rangle$ et u est une entrée, alors on appelle Σ/k un *système entrée-sortie*, avec sortie y. Si $\Sigma \supsetneq k\langle u, y\rangle$, on appelle $k\langle u, y\rangle/k$ un *sous-système entrée-sortie* de Σ/k.

Le système entrée-sortie $k\langle u, y\rangle/k$ avec l'entrée u peut alors être représenté par p é.d.o. de la forme

$$R_j(y_j, \ldots, y_j^{(\beta_j)}, u, \ldots, u^{(\gamma_j)}) = 0, \quad j = 1, \ldots, p,$$

avec R_j, $j = 1, \ldots, p$, des polynômes à coefficients dans k.

Inversion

On définit des notions d'inversibilité, qui dépendent du choix du corps $k\langle y\rangle$ (on parle d'inversibilité et d'inversion en analogie au linéaire) — mais non de celui d'une entrée u. Soit pour ceci Σ/k un système, avec deg tr diff $\Sigma/k = m$, et soit $y = (y_1, \ldots, y_p)$ une sortie.

Rang de sortie : Le degré de transcendance différentiel de $k\langle y\rangle/k$ est appelé le *rang (différentiel) de sortie* du système Σ/k (par rapport à $k\langle y\rangle/k$, ou à y), que l'on note $\rho_y =$ deg tr diff $k\langle y\rangle/k$.

Inversibilité : Le système Σ/k avec la sortie y est dit *inversible à droite* si le rang de sortie est égal au nombre de composantes de la sortie : $\rho_y = p$. Il est dit *inversible à gauche* si le rang de sortie est égal au nombre de composantes indépendantes d'une entrée : $\rho_y = m =$ deg tr diff Σ/k. Le système est dit *inversible* s'il est inversible à droite et à gauche.

Remarque 8.4.1. De façon plus exacte, mais encombrante, on parlerait d'inversibilité par rapport à un sous-corps $k\langle y\rangle$ de Σ (ou par rapport à une sortie y).

Si le système Σ/k est inversible à droite par rapport à une sortie y, alors les composantes de y sont « découplées » : il n'existe aucune relation (non triviale) $E(y,\dots,y^{(\gamma)})=0$, avec $E\in k\{y\}$, entre ses composantes. Si Σ/k est inversible à gauche par rapport à une sortie y, alors l'extension $\Sigma/k\langle y\rangle$ est différentiellement algébrique. De même on a l'assertion réciproque, car avec

$$\text{deg tr diff } \Sigma/k = \text{deg tr diff } \Sigma/k\langle y\rangle + \text{deg tr diff } k\langle y\rangle/k$$

(voir #27) le degré de transcendance différentiel deg tr diff $\Sigma/k\langle y\rangle$ est égal à zéro si, et seulement si, deg tr diff $k\langle y\rangle/k=$ deg tr diff $\Sigma/k=m$. Il s'ensuit que dans un système inversible à gauche la sortie y peut jouer le rôle d'une entrée (non nécessairement indépendante, car $\rho_y\le\min(m,p)$). On observe qu'un système inversible est « quadratique » : $p=m$.

Soit $\Sigma=k\langle u,y\rangle$, et soit u une entrée de Σ/k. Le système entrée-sortie avec l'entrée y et la sortie u est alors le *système (entrée-sortie) inverse* (de celui d'entrée u et de sortie y).

Remarque 8.4.2. On trouvera une « formule » pour le rang de sortie et un algorithme pour son calcul dans [15].

Dynamique inverse : Soit y la sortie d'un système inversible à gauche Σ/k, c'est-à-dire $\rho_y=$ deg tr diff $k\langle y\rangle/k=$ deg tr diff $\Sigma/k=m$. Alors $\Sigma/k\langle y\rangle$ forme une dynamique, avec entrée y. On parle de *dynamique inverse*. L'entrée y de la dynamique inverse est indépendante si Σ/k est inversible ($\rho_y=p=m$).

Inversion et platitude

Un système plat est inversible par rapport à toute sortie plate y, car $\overline{\Sigma}=\overline{k\langle y\rangle}$, et ainsi deg tr diff $\overline{\Sigma}/k\langle y\rangle=0$. Et, qui plus est, le degré de transcendance non différentiel deg tr $\overline{\Sigma}/k\langle y\rangle=0$: La dynamique $\overline{\Sigma}/k\langle y\rangle$ est triviale (dans le sens de la section 8.3). Ceci permet de donner une autre caractérisation de la propriété de platitude :

Sortie plate comme entrée : Un système Σ/k est plat si, et seulement si, il existe une entrée indépendante y du système étendu $\overline{\Sigma}/k$, tel que la dynamique correspondante est triviale (c'est-à-dire deg tr $\overline{\Sigma}/k\langle y\rangle=0$). Interprétant y comme une sortie (plate), on a $\overline{\Sigma}/k\langle y\rangle$ pour la dynamique inverse par rapport à y associée à $\overline{\Sigma}/k$, et cette dernière est triviale.

Pour l'entrée d'un système plat on a donc

$$0=A_i(u_i,y,\dot y,\dots,y^{(\gamma)}),\quad i=1,\dots,m,$$

avec A_i, $i = 1, \ldots, m$, des polynômes à coefficients dans k. Il en découle que le calcul de u à partir d'une sortie plate y et ses dérivées peut se faire sans résoudre une é.d.o. : le système inverse n'a pas de dynamique, elle est « triviale ».

En partant de cette caractérisation de la platitude, on peut classifier les systèmes par rapport à leur « défaut » (par rapport à la platitude) [26] :

Défaut : Soit $\Gamma \subseteq \overline{\Sigma}$ un sur-corps différentiel de k, tel que, à une clôture algébrique près, l'extension Γ/k soit différentiellement purement transcendante et le degré de transcendance (non différentiel) deg tr $\overline{\Sigma}/\Gamma$ soit minimal (entre ces extensions). Alors on appelle deg tr $\overline{\Sigma}/\Gamma$ le *défaut* du système Σ/k.

Le défaut est égal à zéro si, et seulement si, le système Σ/k est plat. Il est toujours fini, car pour n'importe quelle base de transcendance différentielle v de Σ/k l'extension $\Sigma/k\langle v\rangle$ est différentiellement algébrique.

Remarque 8.4.3. Si z est une base de transcendance différentielle de Γ/k, le défaut est égal à la dimension d'état (voir la section 8.5) de la dynamique inverse par rapport à z. Pour une paramétrisation complète du système on doit spécifier, en plus des m trajectoires pour une sortie plate, au maximum deg tr $\overline{\Sigma}/\Gamma$ trajectoires supplémentaires. Toutefois, une telle paramétrisation n'est pas libre !

Une condition nécessaire pour la platitude peut être donnée en remplaçant dans les équations du système les dérivées par des variables non différentielles indépendantes entre elles [26, 48].

Condition nécessaire de platitude (Critère des surfaces réglées) : Soit $\Sigma = k\langle w\rangle$, avec $w = (w_1, \ldots, w_m)$ et les équations

$$P_i(w, \ldots, w^{(\alpha)}) = 0, \quad i = 1, \ldots, q,$$

où P_i, $i = 1, \ldots, q$, polynômes à coefficients dans k. Alors, si le système Σ/k est plat, il existe une famille $a = (a_1, \ldots, a_s)$, $a_i \in k(\xi_0, \ldots, \xi_{\alpha-1}), i = 1, \ldots, s$, $a \neq 0$, telle que, pour des paramètres $\lambda \in k$ arbitraires on a

$$P_i(\xi_0, \ldots, \xi_{\alpha-1}, \xi_\alpha + \lambda a) = 0, \quad i = 1, \ldots, q.$$

On peut interpréter la solution de cette équation comme une surface dans le k-espace avec les coordonnées ξ_i, $i = 0, \ldots, \alpha$. La condition signifie alors que (dans des point réguliers) il existe une droite en direction des ξ_α qui appartient à la surface. Autrement dit, la projection de la surface définie par $S = 0$ sur le sous-espace avec les coordonnées ξ_α est une surface réglée, c'est-à-dire engendrée par des droites.

Démonstration (voir [26]) : Si Σ/k est plat, et y est une sortie plate, alors il existe un nombre entier $\mu \geq 0$ tel que les dérivées $w^{(i)}$, $i = 0, \ldots, \alpha - 1$, dépendent de

dérivées de y jusqu'à un certain ordre μ, que nous pouvons supposer minimal (tel que $\mu - 1$ ne suffit pas) :

$$w^{(j)} = S_j(y, \dot{y}, \ldots, y^{(\mu)}), \quad j = 0, \ldots, \alpha - 1.$$

En remplaçant les $w^{(j)}$, $j = 0, \ldots, \alpha$, dans l'équation $P = 0$ on obtient, à cause de

$$w^{(\alpha)} = \sum_{i=0}^{\mu} \frac{\partial S_{\alpha-1}}{\partial y^{(i)}}\Big|_{(y,\dot{y},\ldots,y^{(\mu)})} y^{(i+1)} =: A(Y_\mu) + B(Y_\mu)\, y^{(\mu+1)},$$

une équation que l'on peut réécrire comme

$$\tilde{P}(Y_\mu, A(Y_\mu) + B(Y_\mu)\, y^{(\mu+1)}) = 0$$

en rassemblant les dérivées de y jusqu'à l'ordre μ dans Y_μ. Maintenant on fixe la valeur de Y_μ et considère les composantes de $y^{(\mu+1)}$ comme paramètres libres. Alors, si a appartient à l'image de l'application $B(Y_\mu)$ la propriété énoncée est donnée. □

Exemple 8.4.1. Le critère des surfaces réglées pour démontrer la non-platitude est discuté dans la section 8.12 (p. 367) pour l'exemple d'un pendule sur un chariot.

Planification de trajectoires

La propriété de platitude permet une synthèse systématique de trajectoires, en particulier pour des transitions entre régimes stationnaires. Souvent, au début d'une telle transition on part d'un point d'équilibre. On peut décrire ces points par les équations explicites (locales)

$$z = \psi(y, \ldots, y^{(\gamma)}),$$

en substituant $y^{(j)} = 0, j > 0$. Les points d'équilibre sont alors paramétrés par les valeurs constants de y.

La transition est entièrement spécifiée par le choix d'une trajectoire de référence pour une sortie plate, et comme y est différentiellement indépendant, les trajectoires pour ses composantes peuvent être choisies indépendamment. *A priori* toutes les trajectoires de référence $t \mapsto y_{\text{réf},i}(t)$ sont loisibles ; des restrictions apparaissent toutefois par des singularités rencontrées lors de la mise sous forme explicite des équations.

Le calcul des trajectoires est particulièrement simple si l'on choisit des trajectoires polynômiales, car alors leurs coefficients résultent des valeurs de y du début (pour $t = 0$) et de la fin de la transition (pour $t = t_*$), comme solution d'un système d'équations linéaires. Plus généralement, on obtient pour $t = 0$ et $t = t_*$ des conditions pour les y_i et toutes leurs dérivées jusqu'à un certain ordre λ_i. En choisissant une paramétrisation polynômiale pour la trajectoire de y_i le

degré du polynôme nécessaire résulte directement du nombre de ces conditions : Dans le cas de λ conditions il faut un polynôme de degré $\lambda - 1$.

Si, par exemple, les dérivées de y_i jusqu'à l'ordre 2 interviennent dans le calcul de u, et la trajectoire de u doit être continue, alors on a 6 conditions, que l'on peut satisfaire avec un polynôme de degré 5 (ou plus), soit

$$y_{\text{réf},i}(t) = c_{i,0} + c_{i,1}t + c_{i,2}t^2 + c_{i,3}t^3 + c_{i,4}t^4 + c_{i,5}t^5.$$

La valeur $y_{\text{réf},i}(0)$ donne $c_{i,0} = y_{\text{réf},i}(0)$. Les conditions $\dot{y}_{\text{réf},i}(0) = 0$ et $\ddot{y}_{\text{réf},i}(0) = 0$ mènent à $c_{i,1} = c_{i,2} = 0$, et les autres coefficients ($c_{i,3}, c_{i,4}$ et $c_{i,5}$) satisfont le système linéaire

$$\begin{aligned} y_{\text{réf},i}(t_*) - y_{\text{réf},i}(0) &= c_{i,3}t_*^3 + c_{i,4}t_*^4 + c_{i,5}t_*^5 \\ 0 &= 3c_{i,3}t_*^2 + 4c_{i,4}t_*^3 + 5c_{i,5}t_*^4 \\ 0 &= 6c_{i,3}t_* + 12c_{i,4}t_*^2 + 20c_{i,5}t_*^3. \end{aligned} \tag{8.3}$$

Il convient d'introduire une reparamétrisation par

$$a_{j-3} = \frac{c_{i,j}t_*^j}{y_{\text{réf},i}(t_*) - y_{\text{réf},i}(0)}, \quad j = 3, 4, 5.$$

Ainsi on a

$$y_{\text{réf},i}(t) = y_{\text{réf},i}(0) + (y_{\text{réf},i}(t_*) - y_{\text{réf},i}(0)) \frac{t^3}{t_*^3} \left(a_0 + a_1 \frac{t}{t_*} + a_2 \frac{t^2}{t_*^2} \right). \tag{8.4}$$

Les nouveaux coefficients a_0, a_1 et a_2 résultent (indépendamment des valeurs initiales et finales et l'index i) du système linéaire

$$\begin{aligned} 1 &= a_0 + a_1 + a_2 \\ 0 &= 3a_0 + 4a_1 + 5a_2 \\ 0 &= 6a_0 + 12a_1 + 20a_2, \end{aligned}$$

que l'on obtient ou de (8.3) ou des conditions finales ; il vient

$$a_0 = 10, \qquad a_1 = -15, \qquad a_2 = 6. \tag{8.5}$$

L'avantage de la paramétrisation (8.4) est que les coefficients a_0, a_1, a_2 sont indépendants des valeurs initiale et finale de la transition.

Le calcul de trajectoires de référence est simple si le problème consiste en une transition sur un intervalle de temps fini, même si ce ne sont pas des points d'équilibre ; on peut utiliser (8.3) pour des valeurs arbitraires des dérivées initiales et finales. On peut, bien sûr, aussi choisir d'autres paramétrisations.

Exemple 8.4.2. Une plus ample discussion de la planification de trajectoires se trouve, pour l'exemple de la grue, dans la section 8.12 (p. 359).

8.5 États généralisés

État généralisé : Un *état (généralisé)* d'une dynamique (algébrique) $\Sigma/k\langle u\rangle$ est une base de transcendance (non différentielle) $\xi = (\xi_1, \ldots, \xi_n)$ de $\Sigma/k\langle u\rangle$. Comme $\Sigma/k\langle u\rangle$ est différentiellement algébrique, $n = \deg \operatorname{tr} \Sigma/k\langle u\rangle$ est fini (#22) ; on l'appelle la *dimension d'état* de la dynamique $\Sigma/k\langle u\rangle$.

L'état ξ étant une base de transcendance de $\Sigma/k\langle u\rangle$, tout élément z de Σ satisfait une équation de la forme

$$\Psi(z, \xi) = 0,$$

avec Ψ un polynôme appartenant à l'anneau $k\langle u\rangle[z, \xi]$. Ces relations s'écrivent, avec un polynôme en z, ξ, ainsi que u et ses dérivées et à coefficients dans k, dans la forme

$$\psi(z, \xi, u, \ldots, u^{(\gamma)}) = 0.$$

En particulier, on a pour la dérivée des composantes de ξ

$$A_i(\dot{\xi}_i, \xi, u, \dot{u}, \ldots, u^{(\alpha_i)}) = 0, \quad i = 1, \ldots, n,$$

ou bien, localement :

$$\dot{\xi}_i = F_i(\xi, u, \dot{u}, \ldots, u^{(\alpha_i)}), \quad i = 1, \ldots, n.$$

De la même façon on obtient, pour des sorties $y = (y_1, \ldots, y_m)$, localement des relations de la forme

$$y_j = H_j(\xi, u, \dot{u}, \ldots, u^{(\beta_j)}), \quad j = 1, \ldots, p.$$

Si dans cette *représentation d'état* aucune dérivée de u n'apparaît on parle d'un *état classique* et d'une *représentation d'état classique*.

Chaque base de transcendance de $\Sigma/k\langle u\rangle$ forme un état (généralisé). Si $\bar{\xi}$ est un autre état (généralisé) de $\Sigma/k\langle u\rangle$, alors chaque composante de $\bar{\xi}$ satisfait une équation de la forme $P(\bar{\xi}_i, \xi) = 0$, avec les coefficients de P dans $k\langle u\rangle$. Deux états ξ et $\bar{\xi}$ d'une dynamique sont donc reliés par une *transformation d'état (généralisée)* du type

$$\begin{aligned} &\varphi_j(\bar{\xi}_i, \xi, u, \ldots, u^{(\gamma_i)}) = 0 \\ &\bar{\varphi}_i(\xi_i, \bar{\xi}, u, \ldots, u^{(\bar{\gamma}_i)}) = 0, \quad i = 1, \ldots, n. \end{aligned}$$

Pour les rendre explicites, ces équations peuvent encore être résolues (localement) par rapport à $\bar{\xi}_i$ et ξ_i respectivement. On les appelle *classiques* si elle ne font pas intervenir l'entrée et ses dérivées.

Notons que, contrairement au cas linéaire, il n'est pas possible, en général, d'éliminer les dérivées de l'entrée des représentations d'état généralisées en changeant d'état. Autrement dit, il n'existe pas toujours une représentation d'état

classique. On le comprend facilement à partir d'une condition nécessaire simple, que nous dérivons.

Si x est un état classique, alors $\tilde{x}$ est classique si, et seulement si, la transformation entre x et $\tilde{x}$ est classique : De $\tilde{x} = \varphi(x)$ et $\dot{x} = f(x,u)$ il résulte

$$\dot{\tilde{x}} = \left.\frac{\partial \varphi}{\partial x}\right|_{x=\tilde{\varphi}(\tilde{x})} f(\tilde{\varphi}(\tilde{x}), u),$$

et de $\tilde{x} = \varphi(x,u)$ et $\dot{x} = f(x,u)$ il découle que

$$\dot{\tilde{x}} = \left.\frac{\partial \varphi}{\partial x}\right|_{x=\tilde{\varphi}(\tilde{x},u),u} f(\tilde{\varphi}(\tilde{x},u),u) + \left.\frac{\partial \varphi}{\partial u}\right|_{x=\tilde{\varphi}(\tilde{x},u),u} \dot{u}\,,$$

qui dépend de $\dot{u}$ si φ dépend de u.

En généralisant ce calcul on obtient une condition nécessaire simple pour la réduction (de 1) de l'ordre maximal des dérivées $u_i^{(\alpha_i)}$ de u_i. Pour simplifier la démarche supposons que l'ordre maximal soit égal à $\alpha_i = \alpha, i = 1, \ldots, m$ pour toutes les composantes de u. Alors, de

$$\tilde{x} = \varphi(x, u, \ldots, u^{(\alpha-1)})$$

et

$$\dot{x} = f(x, u, \ldots, u^{(\alpha)}) \tag{8.6}$$

on déduit

$$\begin{aligned}\dot{\tilde{x}} = {} & \left.\frac{\partial \varphi}{\partial x}\right|_{x=\tilde{\varphi}(\tilde{x},u,\ldots,u^{(\alpha-1)}),u,\ldots,u^{(\alpha-1)}} f(\tilde{\varphi}(\tilde{x},u,\ldots,u^{(\alpha-1)}),u,\ldots,u^{(\alpha)}) + \\ & \sum_{i=0}^{\alpha-1} \left.\frac{\partial \varphi}{\partial u^{(i)}}\right|_{x=\tilde{\varphi}(\tilde{x},u,\ldots,u^{(\alpha-1)}),u,\ldots,u^{(\alpha-1)}} u^{(i+1)}.\end{aligned}$$

Par un choix approprié de $\partial\varphi/\partial u^{(\alpha-1)}$ on peut éliminer $u^{(\alpha)}$ du membre droit de cette équation si la fonction f est affine en cette variable, c'est-à-dire l'équation (8.6) a la forme

$$\dot{x} = f_1(x, u, \ldots, u^{(\alpha-1)}) + f_2(x, u, \ldots, u^{(\alpha-1)})\, u^{(\alpha)}.$$

Il est évident qu'une condition nécessaire similaire existe pour la possibilité d'éliminer les dérivées d'ordre maximal individuel des composantes de u. On trouve une discussion complète, avec des conditions nécessaires et suffisantes[3] dans [9].

[3]Les conditions suffisantes correspondent à l'intégrabilité d'é.d.p. linéaires d'ordre 1 pour les transformations d'état et sont donc basées sur le théorème de Frobenius.

Remarque 8.5.1. Dans [13] (voir aussi [41]) il a été démontré que l'idéal différentiel correspondant à une représentation d'état (généralisée) avec une fraction rationnelle dans le membre droit, c'est-à-dire

$$\dot{x}_i = \frac{p_i(x, u, \dots, u^{(\alpha)})}{q_i(x, u, \dots, u^{(\alpha)})}, \quad i = 1, \dots, n,$$

avec $p_i, q_i \in k[x, u, \dots, u^{(\alpha)}]$, $i = 1, \dots, n$, est un idéal premier (#3). Ainsi pour une telle représentation il existe toujours une extension de corps différentielle $k\langle x, u\rangle/k$.

Exemple 8.5.1. Un exemple d'une dynamique n'admettant pas de représentation d'état classique est le sous-système « pendule » du modèle de grue [7] (voir section 8.12, p. 364).

8.6 État de Brunovský et forme de commande généralisée

Pour les systèmes plats on peut définir des états particuliers, de Brunovský, et une forme de commande généralisée correspondante, comme en linéaire.

Dynamique plate : Une dynamique $\Sigma/k\langle u\rangle$ est appelée *(différentiellement) plate* si le système Σ/k est (différentiellement) plat.

État de Brunovský : Soit $y = (y_1, \dots, y_m)$ une sortie plate d'une dynamique plate $\Sigma/k\langle u\rangle$. Alors il existe une famille $\kappa = (\kappa_1, \dots, \kappa_m) \in \mathbb{N}^m$ telle que (avec la convention que $y_i^{(-1)} = \emptyset$)

$$x = (y_1, \dot{y}_1, \dots, y_1^{(\kappa_1 - 1)}, y_2, \dots, y_m^{(\kappa_m - 1)}) \tag{8.7}$$

forme un état (généralisé) de la dynamique $\Sigma/k\langle u\rangle$; un tel x est appelé *état de Brunovský* de $\Sigma/k\langle u\rangle$.

Démonstration [10] : On choisit une base de transcendance $\bar{y}^0 \subseteq y$ de l'extension $k\langle u\rangle(y)/k\langle u\rangle$, et puis pour tout $r \geq 0$ une base de transcendance $\bar{y}^r \subseteq \dot{\bar{y}}^{r-1}$ de l'extension de corps $k\langle u\rangle(y, \dots, y^{(r)})/k\langle u\rangle(y, \dots, y^{(r-1)})$. Ceci est possible, car on a l'égalité $k\langle u\rangle(y, \dots, y^{(r)}) = k\langle u\rangle(y, \dots, y^{(r-1)}, \dot{\bar{y}}^{r-1})$. Le degré de transcendance $\deg \operatorname{tr} \overline{\Sigma}/k\langle u\rangle$ étant fini, la réunion des $\bar{y}^r$, $r \geq 0$, forme un état de $\overline{\Sigma}/k\langle u\rangle$. On peut l'écrire comme x dans (8.7). Ceci détermine les $\kappa_i = \min\{r \in \mathbb{N} \mid y_i^{(r)} \notin \bar{y}^r\}$, $i = 1, \dots, m$. □

Si au moins un des κ_i est non nul, on peut renuméroter les composantes de y tel que

$$\kappa_i > 0,\ i = 1, \dots, \bar{m} \leq m, \quad \kappa_i = 0,\ i = \bar{m} + 1, \dots, m.$$

Si $\kappa_i = 0$ pour tout i, choisissons $\bar{m} = 0$ et $x = \emptyset$. Une telle numérotation des composantes de y sera supposée dans la suite dans les cas où l'index $\bar{m}$ est utilisé. En plus, on définit $n_i = \sum_{j=1}^{i} \kappa_j$, $i = 1, \ldots, \bar{m}$.

Forme de commande généralisée : Un état de Brunovský x d'une dynamique plate $\Sigma/k\langle u\rangle$ donne lieu à une *forme de commande généralisée* :

$$\begin{aligned} \dot{x}_j &= x_{j+1}, && j \in \{1, \ldots, n\} \backslash \{n_1, \ldots, n_{\bar{m}}\}, \\ 0 &= \Phi_j(\dot{x}_{n_j}, x, u, \ldots, u^{(\alpha_j)}), && j = 1, \ldots, \bar{m}, \\ 0 &= \Phi_j(y_j, x, u, \ldots, u^{(\alpha_j)}), && j = \bar{m}+1, \ldots, m, \end{aligned}$$

avec $\Phi_j, j = 1, \ldots, m$, des polynômes à coefficients dans k, où $\partial\Phi_j/\partial\dot{x}_{n_j} \neq 0, j = 1, \ldots, \bar{m}$, et $\partial\Phi_j/\partial y_j \neq 0, j = \bar{m}+1, \ldots, m$.

Exemple 8.6.1. Un état de Brunovský et une forme de commande généralisée pour le sous-système « pendule » de la grue seront construits en section 8.12 (p. 364).

8.7 Équivalence par bouclages d'états quasi statiques

Pour la définitions de bouclages d'états on se sert de filtrations de l'extension de corps différentielle Σ/k définissant le système [8] (cf. #28).

Filtration entrée-état : Pour une dynamique $\Sigma/k\langle u\rangle$ et un état (généralisé) correspondant x, la *filtration entrée-état* $\mathcal{U} = (\mathcal{U}_r)_{r\in\mathbb{Z}}$ de Σ/k est définie comme suit :

$$\begin{aligned} \mathcal{U}_r &= k && \text{pour} \quad r \leq -2 \\ \mathcal{U}_{-1} &= \overline{k(x)} && \text{pour} \quad r = -1 \\ \mathcal{U}_r &= \overline{k(x, u, \dot{u}, \ldots, u^{(r)})} && \text{pour} \quad r \geq 0. \end{aligned}$$

Une filtration entrée-état $\tilde{\mathcal{U}} = (\tilde{\mathcal{U}}_r)_{r\in\mathbb{Z}}$ pour une dynamique $\tilde{\Sigma}/k\langle\tilde{u}\rangle$ est définie de façon analogue. Les deux filtrations ont une *différence bornée* s'il existe un entier r_0, tel que $\tilde{\mathcal{U}}_r \subset \mathcal{U}_{r+r_0}$ et $\mathcal{U}_r \subset \tilde{\mathcal{U}}_{r+r_0}$ pour tout r (cf. #28). Il est clair qu'alors $\overline{\Sigma} = \overline{\tilde{\Sigma}}$.

Équivalence par bouclages d'états : Deux dynamiques $\Sigma/k\langle u\rangle$ et $\tilde{\Sigma}/k\langle\tilde{u}\rangle$ sont dites *équivalentes par bouclages quasi statiques d'un état dans X* s'il existe des états de $\Sigma/k\langle u\rangle$ et $\tilde{\Sigma}/k\langle\tilde{u}\rangle$ tels que les filtrations entrée-état $\mathcal{U}$ de $\Sigma/k\langle u\rangle$ et $\tilde{\mathcal{U}}$ de $\tilde{\Sigma}/k\langle\tilde{u}\rangle$ ont une différence bornée, et en plus $\mathcal{U}_{-1} = \tilde{\mathcal{U}}_{-1} = X$.

Les deux dynamiques sont dites *équivalentes par bouclages statiques d'un état dans X* si les filtrations entrée-état correspondantes coïncident, c'est-à-dire si pour tout $r \in \mathbb{Z}$ on a $\mathcal{U}_r = \tilde{\mathcal{U}}_r$, et en plus $\mathcal{U}_{-1} = \tilde{\mathcal{U}}_{-1} = X$.

Les relations définies sont en effet des relations d'équivalence. La symétrie et la reflexivité en sont évidentes. La transitivité est évidente pour le cas statique. Pour le cas quasi statique, soit $\hat{\Sigma}/k\langle\hat{u}\rangle$ une troisième dynamique, avec un état $\hat{x}$ et $X = \overline{k\langle\hat{x}\rangle}$. Soit $\hat{\mathcal{U}} = (\hat{\mathcal{U}}_r)_{r\in\mathbb{Z}}$ la filtration entrée-état correspondant à $\hat{x}$, pour $\hat{\Sigma}/k\langle\hat{u}\rangle$. Pour l'équivalence par bouclages quasi statiques d'un état dans X il vient alors : $\mathcal{U}_r \subset \tilde{\mathcal{U}}_{r+r_0}$ et $\tilde{\mathcal{U}}_r \subset \mathcal{U}_{r+r_0}$, ainsi que $\tilde{\mathcal{U}}_r \subset \hat{\mathcal{U}}_{r+r_1}$ et $\hat{\mathcal{U}}_r \subset \tilde{\mathcal{U}}_{r+r_1}$, pour tout r. Avec $\mathcal{U}_r \subset \hat{\mathcal{U}}_{r+r_0+r_1}$ et $\hat{\mathcal{U}}_r \subset \mathcal{U}_{r+r_0+r_1}$, pour tout r, la transitivité est démontrée.

L'équivalence de deux dynamiques $\Sigma/k\langle u\rangle$ et $\tilde{\Sigma}/k\langle\tilde{u}\rangle$, avec des états respectifs x et $\tilde{x}$, par bouclages statiques d'un état dans X implique l'existence de relations du type

$$\begin{aligned}&\phi_{i,0}(u_i, x, \tilde{u}, \dot{\tilde{u}}, \ldots, \tilde{u}^{(r_0)}) = 0,\\ &\tilde{\phi}_{i,0}(\tilde{u}_i, x, u, \dot{u}, \ldots, u^{(r_0)}) = 0, \quad i = 1, \ldots, m,\end{aligned}$$

et des relations analogues existent pour les dérivées d'ordres supérieurs :

$$\begin{aligned}&\phi_{i,r}(u_i^{(r)}, x, \tilde{u}, \dot{\tilde{u}}, \ldots, \tilde{u}^{(r+r_0)}) = 0,\\ &\tilde{\phi}_{i,r}(\tilde{u}_i^{(r)}, x, u, \dot{u}, \ldots, u^{(r+r_0)}) = 0, \quad i = 1, \ldots, m, \quad r > 0.\end{aligned}$$

Dans ces relations, $\phi_{i,r}, \tilde{\phi}_{i,r}$, $i = 1, \ldots, m$, $r \geq 0$, sont des polynômes à coefficients dans k. Dans le cas statique aucune dérivée (d'ordre supérieur ou égal à 1) des entrées n'apparaît.

Remarque 8.7.1. Les filtrations entrée-état sont discrètes, excellentes et exhaustives (#28). Pour le démontrer il faut prendre en compte que, pour les indices larges, dans la construction de $\mathcal{U}_{r+1}$ à partir de $\mathcal{U}_r$ on ajoute des éléments $\frac{d}{dt}z$ avec $z \in \mathcal{U}_r$, ce qui pour $u^{(r+1)}$ est une conséquence de la construction de la filtration, et pour $\dot{x}$ une conséquence de la représentation d'état généralisée.

Pour les états $\tilde{x}$ de la dynamique $\overline{\tilde{\Sigma}}/k\langle\tilde{u}\rangle$, la condition $\overline{k(x)} = \overline{k(\tilde{x})}$ implique l'existence de relations du type

$$\begin{aligned}&\psi_j(x_j, \tilde{x}) = 0,\\ &\tilde{\psi}_j(\tilde{x}_j, x) = 0, \quad j = 1, \ldots, n,\end{aligned}$$

où les $\psi_j, \tilde{\psi}_j$, $j = 1, \ldots, n$, sont encore des polynômes à coefficients dans k. Ainsi, contrairement aux représentations d'état généralisées, dans les transformations d'état admises dans les bouclages d'état aucune entrée, ni ses dérivées, ne peuvent apparaître : Les transformations doivent être classiques.

L'utilité des bouclages d'état quasi statiques résulte du fait que l'état (ou, plus exactement, le corps $\overline{k\langle x\rangle}$) est invariant par ce genre de bouclages (par rapport à l'invariance voir aussi [51]). D'un point de vue pratique, ceci a l'avantage qu'aucune « dynamique supplémentaire » n'est introduite dans la boucle fermée.

Exemple 8.7.1. La construction d'un bouclage d'état quasi statique pour la grue est discutée en section 8.12 (p. 366).

Remarques 8.7.1. 1. Soulignons que l'apparition des dérivées dans les équations des bouclages ne signifie pas la nécessité de dérivations (numériques) de signaux d'entrée (voir la section 8.9). Des discussions détaillées et des exemples se trouvent aussi dans [11, 12, 10, 54].
2. Les bouclages d'état quasi statiques forment une classe spéciale des *bouclages endogènes* de [39, 26], que l'on peut définir comme suit. Deux systèmes Σ_1/k et Σ_2/k sont dits *équivalents par bouclages endogènes* si $\overline{\Sigma_1} = \overline{\Sigma_2}$. Contrairement aux bouclages d'état quasi statiques la dimension d'état n'est pas préservée sous bouclages endogènes. (En fait, la définition des bouclages endogènes ne fait pas appel aux notions d'entrée et d'état.) La synthèse de bouclages endogènes pour la poursuite de trajectoires stable mène, en général, à des lois de commandes dynamiques, nécessitant l'intégration d'équations différentielles (voir par ex. [39, 26]).
3. Soit $m = 1$, et soit l'état utilisé pour construire les filtrations entrée-état $\mathcal{U}$ et $\tilde{\mathcal{U}}$ un état classique. Alors la différence bornée des deux filtrations $\mathcal{U}$ et $\tilde{\mathcal{U}}$ implique leur égalité, c'est-à-dire le bouclage quasi statique avec l'état x est un bouclage statique. On peut s'en rendre compte en essayant, pour un système $\dot{x} = f(x, u)$ avec $m = 1$ et $u = \phi_0(x, \tilde{u}, \dot{\tilde{u}})$, de trouver des équations de la forme $\tilde{u}^{(i)} = \tilde{\phi}_i(x, u, \ldots, u^{(i+r_0)})$, $i \geq 0$. Ceci est impossible. Il s'ensuit, pour le cas $m = 1$: Si x est un état généralisé de $\Sigma/k\langle u\rangle$ qui n'est pas classique, alors il ne peut pas non plus être un état classique de $\tilde{\Sigma}/k\langle\tilde{u}\rangle$, car un bouclage statique d'un état $\overline{k\langle x\rangle}$ ne modifie pas l'ordre minimal α des dérivées de l'entrée apparaissant dans la représentation d'état [10]. Ceci est une conséquence du fait que $\dot{x} \in \mathcal{U}_\alpha, \dot{x} \notin \mathcal{U}_{\alpha-1}$ et $\mathcal{U}_r = \tilde{\mathcal{U}}_r, r \in \mathbb{Z}$ impliquent $\dot{x} \in \tilde{\mathcal{U}}_\alpha, \dot{x} \notin \tilde{\mathcal{U}}_{\alpha-1}$. Des résultats supplémentaires sur l'équivalence par bouclages d'état quasi statiques pour des représentations d'état classiques se trouvent dans [51].

8.8 Linéarisabilité par bouclages d'état quasi statiques

Forme de Brunovský : Pour des nombres non négatifs arbitraires $\kappa_1, \ldots, \kappa_m$ la dynamique $k\langle y, v\rangle/k\langle v\rangle$ avec $v_i = y_i^{(\kappa_i)}$, $i = 1, \ldots, m$, est appelée une *forme de Brunovský.*

Il est évident qu'une forme de Brunovský est une dynamique plate, dont y est une sortie plate.

Linéarisabilité par bouclages d'état quasi statiques : Une dynamique $\Sigma/k\langle u\rangle$ est dite *linéarisable par bouclages d'état quasi statiques* s'il existe un état x de $\overline{\Sigma}/k\langle u\rangle$ et une forme de Brunovský, tels que cette dernière et $\Sigma/k\langle u\rangle$ sont équivalentes par bouclages quasi statiques d'un état dans $\overline{k(x)}$.

Remarque 8.8.1. Le problème de la linéarisabilité par bouclages d'état quasi statiques est une généralisation naturelle du problème bien établi de la linéarisabilité par bouclages d'état statiques pour les représentations d'état classiques, qui admet une solution complète dans le cadre de la géométrie différentielle [33, 31, 59]. D'autres extensions sont la linéarisabilité par bouclages dynamiques dans le sens de [2, 3], et, plus particulièrement, par les bouclages endogènes [26]. Le concept de la platitude à été introduit dans ce contexte, et dans [23] on trouve la conjecture que les systèmes linéarisables par bouclages dynamiques sont juste les systèmes plats. (Que les systèmes plats soient linéarisables par bouclages dynamiques est évident, à partir de leur linéarisabilité par bouclages d'états quasi statiques ou endogènes.) Pour plus de détails sur ces questions voir [39, 45], par exemple.

Platitude et linéarisabilité : Une dynamique $\Sigma/k\langle u\rangle$ est linéarisable par bouclages d'état quasi statiques si, et seulement si, elle est plate.

Preuve [10] : On peut considérer un état de Brunovský, de $v_i = y_i^{(\kappa_i)}$, $i = 1, \ldots, m$, et l'utiliser pour construire la « filtration état-sortie » pour l'état x et la sortie y, $\mathcal{Y} = (\mathcal{Y}_r)_{r\in\mathbb{Z}}$, avec

$$\begin{aligned} \mathcal{Y}_r &= k && \text{pour} \quad r \leq -2 \\ \mathcal{Y}_{-1} &= \overline{k(x)} && \text{pour} \quad r = -1 \\ \mathcal{Y}_r &= \overline{k(x, y, \dot{y}, \ldots, y^{(r)})} && \text{pour} \quad r \geq 0. \end{aligned}$$

Comme y est une sortie plate, la filtration $\mathcal{Y}$ est exhaustive dans $\overline{\Sigma}$, et ainsi $\overline{k\langle v\rangle(x)} = \overline{\Sigma}$. En plus, pour tout $r \in \mathbb{Z}$, $(x, v, \dot{v}, \ldots, v^{(r)})$ est une famille k-algébriquement indépendante dans $\overline{\Sigma}$. Or, v est une entrée du système $\overline{\Sigma}/k$, et x est un état de la dynamique $\overline{\Sigma}/k\langle v\rangle$. Désignant la filtration entrée-état, pour x et v, par $\mathcal{V}$, les filtration $\mathcal{U}$ et $\mathcal{V}$ ont une différence bornée (#28), car les filtrations entrée-état sont discrètes, excellentes et exhaustives dans $\overline{\Sigma}$ (voir aussi la remarque p. 349). Comme x est un état des deux dynamiques, $\Sigma/k\langle u\rangle$ et $\overline{\Sigma}/k\langle v\rangle$, elles sont équivalentes par un bouclage quasi statique de x, ou de n'importe quelle autre base de transcendance de $\overline{k(x)}/k$.

Pour démontrer la nécessité il suffit de constater que l'équivalence par bouclages quasi statiques d'un état dans $\overline{k(x)}$ d'une dynamique $\overline{\Sigma}/k\langle u\rangle$ et une forme de Brunovský $k\langle y, v\rangle/k\langle v\rangle$ implique $\overline{\Sigma} = \overline{k\langle y, v\rangle}$. Ainsi, toute forme de Brunovský étant plate, $\overline{\Sigma}/k$ l'est aussi. □

8.9 Poursuite de trajectoires pour des systèmes plats

Une méthode systématique pour la poursuite de trajectoires pour les systèmes plats (non linéaires) s'appuie sur leur linéarisabilité par bouclages d'état quasi statiques. Elle permet une stabilisation exponentielle le long de trajectoires (non singulières).

Les bouclages d'état linéarisants transforment le système de telle façon que

$$y_i^{(\kappa_i)} = v_i, \qquad i = 1, \dots, m.$$

Pour cette dynamique, linéaire commandable, une poursuite de trajectoires par un bouclage stabilisant (exponentiellement) est aisée, il suffit de prendre

$$v_i = y_{\mathrm{r},i}^{(\kappa_i)} + \sum_{j=0}^{\kappa_i - 1} \lambda_{i,j} (y_i^{(j)} - y_{\mathrm{r},i}^{(j)}), \qquad i = 1, \dots, m.$$

Les $y_{\mathrm{r},i}$, $i = 1, \dots, m$, seront ensuite remplacés par les trajectoires de référence. La dynamique de la boucle fermée est assignée en choisissant les paramètres $\lambda_{i,j}$. Pour le calcul des dérivées de v on substitue $v_i = y_i^{(\kappa_i)}$, $i = 1, \dots, m$, et l'on obtient ainsi, successivement, ces dérivées en fonction de l'état x et les dérivées des trajectoires de référence (pour $i = 1, \dots, m$ et $l \geq 0$) :

$$v_i^{(l)} = y_{\mathrm{r},i}^{(\kappa_i + l)} + \sum_{j=0}^{\kappa_i - 1 - l} \lambda_{i,j} (y_i^{(j+l)} - y_{\mathrm{r},i}^{(j+l)}) + \sum_{r=\kappa_i - l}^{\kappa_i - 1} \lambda_{i,r} \, (v_i^{(r - \kappa_i + l)} - y_{\mathrm{r},i}^{(j+l)}).$$

Exemple 8.9.1. La poursuite de trajectoires est illustrée sur l'exemple de la grue en section 8.12 (p. 366).

8.10 Les systèmes linéaires tangents

Les systèmes linéaires à coefficients constants résultent, le plus souvent, d'une linéarisation d'un système non linéaire, modèle d'un processus, autour d'un point d'équilibre (dans l'espace des variables du système). En considérant, au lieu d'un seul point, une trajectoire, la linéarisation donne lieu à un système instationnaire. Pour le système d'équations

$$P_i \left(w, \dots, w^{(\alpha)} \right) = 0, \quad i = 1, \dots, q,$$

on obtient

$$\sum_{j=1}^{s} \sum_{l=0}^{\alpha} \left. \frac{\partial P_i}{\partial w_j^{(l)}} \right|_{(w_{\mathrm{réf}}(t), \dots, w_{\mathrm{réf}}^{(\alpha)}(t))} dw_j^{(l)} = 0, \quad i = 1, \dots, q.$$

Les équations du système ainsi linéarisé sont donc les équations des « petits écarts » $dw_j^{(l)}$, qui prennent leurs coefficients dans le corps Σ. Ces coefficients

dépendront donc du temps, une fois les trajectoires $t \mapsto w_{\text{réf}}(t)$ substituées. Il est clair que l'on peut écrire ces équations comme des relations $\Sigma[\frac{d}{dt}]$-linéaires entre les dw_j.

Dans le cadre de l'algèbre différentielle cette linéarisation se traite à l'aide des différentielles de Kähler (#24) :

Système linéaire tangent : Soit Σ/k un système algébrique. Le $\Sigma[\frac{d}{dt}]$-module $\Omega_{\Sigma/k}$ des différentielles de Kähler est appelé le *système linéaire tangent* associé.

Le système linéaire tangent est défini en employant la k-dérivation $d_{\Sigma/k} : \Sigma \to \Omega_{\Sigma/k}$ (voir #24). Soit $\Sigma = k\langle w\rangle$, avec $w = (w_1, \ldots, w_s)$; alors $\Omega_{\Sigma/k}$ est le $\Sigma[\frac{d}{dt}]$-module engendré par les différentielles de Kähler $d_{\Sigma/k}w_i, i = 1, \ldots, s$ (voir #24). Pour chaque relation $R(w, \ldots, w^{(\gamma)}) = 0$ dans Σ il existe une relation de dépendance $\Sigma[\frac{d}{dt}]$-linéaire dans $\Omega_{\Sigma/k}$:

$$\sum_{j=1}^{s}\sum_{l=0}^{\gamma}\frac{\partial R}{\partial w_j^{(l)}}\left(\frac{d}{dt}\right)^l d_{\Omega/k}\, w_j = 0.$$

Ceci est évident. Ce qui est plus intéressant, mais plus difficile à démontrer [35], c'est l'inverse : Si une famille $d_{\Sigma/k}z$ d'éléments de $\Omega_{\Sigma/k}$ est $\Sigma[\frac{d}{dt}]$-linéairement dépendante (resp. indépendante), alors z est différentiellement k-algébriquement dépendant (resp. indépendante). Ce fait offre une possibilité immédiate pour le calcul des degrés de transcendance (différentielles ou non) importants dans les systèmes algébriques [15]. En plus, elle établie un lien directe entre la platitude et la commandabilité du système linéaire tangent [26] (à propos de l'emploi de $\overline{\Sigma}$ voir #25) :

Platitude et commandabilité du système linéaire tangent : Si le système (algébrique) Σ/k est plat, alors le système linéaire tangent de $\overline{\Sigma}/k$, c'est-à-dire le $\overline{\Sigma}[\frac{d}{dt}]$-module des différentielles de Kähler $\Omega_{\overline{\Sigma}/k}$, est libre. Si y est une sortie plate de Σ/k, alors $d_{\overline{\Sigma}/k}y$ est une base de $\Omega_{\overline{\Sigma}/k}$.

On associe ainsi des modules libres à des systèmes plats. La commandabilité (liberté) du système linéaire tangent est donc une condition *nécessaire* pour la platitude du système algébrique. La réciproque, par contre, n'est pas vraie en générale : Le système linéaire tangent associé à un système non plat peut être libre (commandable). De même on à le résultat suivant :

Défaut et sous-système non commandable : Le défaut d'un système algébrique Σ/k est au moins aussi large que la dimension d'état du sous-système non commandable du système linéarisé tangent associé.

Démonstration : Soit $\Gamma \subseteq \overline{\Sigma}$, à la clôture algébrique près, un corps d'extension différentiellement purement transcendant de k, tel que le degré de transcendance (non différentiel) deg tr $\overline{\Sigma}/\Gamma$ est minimal. Alors le défaut du système est égal à deg tr $\overline{\Sigma}/\Gamma$, et ce dernier est égal à la dimension du $\overline{\Sigma}$-espace vectoriel des différentielles de Kähler[4] $\Omega_{\overline{\Sigma}/\Gamma}$. L'image $d_{\overline{\Sigma}/k}\Gamma$ de Γ sous $d_{\overline{\Sigma}/k}$ est un $\overline{\Sigma}[\frac{d}{dt}]$-sous-module libre de $\Omega_{\overline{\Sigma}/k}$, donc un sous-système du sous-système commandable du système linéaire tangent $\Omega_{\overline{\Sigma}/k}$. Ainsi $\Omega_{\overline{\Sigma}/\Gamma}$ est isomorphe à $\Omega_{\overline{\Sigma}/k}/d_{\overline{\Sigma}/k}\Gamma$, et ce dernier contient un sous-module qui est isomorphe à $\mathrm{t}\Omega_{\overline{\Sigma}/k}$, le sous-module de torsion (c'est-à-dire la partie non commandable) de $\Omega_{\overline{\Sigma}/k}$. La dimension du $\overline{\Sigma}$-espace vectoriel $\Omega_{\overline{\Sigma}/\Gamma}$ est alors au moins aussi large que celle du $\overline{\Sigma}$-espace vectoriel $\mathrm{t}\Omega_{\overline{\Sigma}/k}$. □

Exemple 8.10.1. Un exemple d'un système non plat, un pendule sur un chariot, sera discuté en section 8.12 (p. 367).

8.11 Observabilité

Soit Σ/k un système algébrique. Un système $\tilde{\Sigma}/k$ est appelé sous-système de Σ/k dans le cas $\tilde{\Sigma} \subseteq \Sigma$.

Observabilité [14] : Un sous-système $\tilde{\Sigma}/k$ de Σ/k est dit *observable par une famille* z si $\overline{\tilde{\Sigma}} = \overline{k\langle z\rangle}$. Un système Σ est dit *observable* si $\overline{\Sigma} = \overline{k\langle y, u\rangle}$.

Autrement dit, un ensemble de variables du système est observable par z si ses éléments peuvent être « reconstruits » de z (considéré comme ensemble de mesures), c'est-à-dire calculé de z et ses dérivées sans intégration. Pour l'observabilité du système (tout court) on suppose que ce sont les u et y qui sont connus. On peut considérer le sous-système $\overline{k\langle u, y\rangle}/k$ de $\overline{\Sigma}/k$ comme son *sous-système observable*.

Remarque 8.11.1. Une notion d'observabilité plus stricte (rationnelle) est obtenue en se passant de la clôture algébrique [14].

Pour un système plat, avec une sortie plate y, on obtient immédiatement de $\overline{\Sigma} = \overline{k\langle y\rangle}$ l'affirmation suivante. (Notons que l'on n'y a pas besoin de faire appel aux entrées.)

Observabilité par une sortie plate : Un système plat Σ/k est observable par n'importe quelle sortie plate.

Finalement, on peut établir le lien entre l'observabilité d'un système algébrique et le système linéaire tangent associé.

[4] L'application $d_{\overline{\Sigma}/\Gamma} : \overline{\Sigma} \to \Omega_{\overline{\Sigma}/\Gamma}$ envoie les éléments de Γ à 0 (voir #24).

Observabilité des systèmes algébrique et linéaire tangent : Un sous-système $\tilde{\Sigma}/k$ de Σ/k est observable par une famille $z \in \overline{\tilde{\Sigma}}$ si, et seulement si, le système linéaire tangent $\Omega_{\overline{\tilde{\Sigma}}/k}$ est observable par $d_{\overline{\tilde{\Sigma}}/k}z$. Le système Σ/k est observable si, et seulement si, le système linéaire tangent $\Omega_{\overline{\Sigma}/k}$ associé à $\tilde{\Sigma}/k$ est observable.

Démonstration : Il s'agit d'une conséquence directe de l'équivalence entre la dépendance $k\langle z\rangle$-algébrique d'une famille $\zeta = (\zeta_1, \ldots, \zeta_r)$ d'éléments de Σ et de la dépendance Σ-linéaire des différentielles de Kähler $d_{\Sigma/k\langle z\rangle}\zeta = (d_{\Sigma/k\langle z\rangle}\zeta_1, \ldots, d_{\Sigma/k\langle z\rangle}\zeta_r)$ (cf. #24 et #25). □

Remarque 8.11.2. La réalisation des bouclages d'état stabilisants nécessite la connaissance des états. Si ces derniers ne sont pas mesurés directement, on peut, si le système est observable par les mesures, construire un observateur. C'est un problème non trivial pour les systèmes non linéaires. Toutefois, pour les systèmes plats (observables) on peut construire un observateur pour le système linéaire tangent autour des trajectoires de référence. Une commande stabilisante étant réalisée pour la poursuite de ces trajectoires, on obtient une boucle fermée localement stable. On parle d'observateurs de poursuite dans ce cas [28, 29, 45].

Alternativement, des méthodes de calcul numérique rapide des dérivées des mesures proposées plus récemment forment une approche intéressante plus directement basée sur l'observabilité [22].

8.12 Exemple : Une grue

Nous considérons une grue comme esquissée sur la figure 8.1. Pour simplifier la discussion, supposons que le chariot et la charge sont toujours situés dans un même plan vertical ; la généralisation spatiale en est simple (voir section 8.12). Le modèle mathématique avec les variables X, Y, D_x, R, T, C, F, θ et ω s'écrit :

$$m\ddot{X} = -T\sin\theta \tag{8.8a}$$
$$m\ddot{Y} = -T\cos\theta + mg \tag{8.8b}$$
$$X = R\sin\theta + D_x \tag{8.8c}$$
$$Y = R\cos\theta \tag{8.8d}$$
$$M\ddot{D}_x = F + T\sin\theta - c_{\mathrm{d}}\dot{D}_x \tag{8.8e}$$
$$J\dot{\omega} = C - \rho T - c_{\mathrm{r}}\omega \tag{8.8f}$$
$$\dot{R} = -\rho\omega. \tag{8.8g}$$

Ici (X, Y) désigne les coordonnés cartésiennes de la charge (dans un système fixe), D_x celle de la charge. Les autres variables sont T pour la force dans le câble,

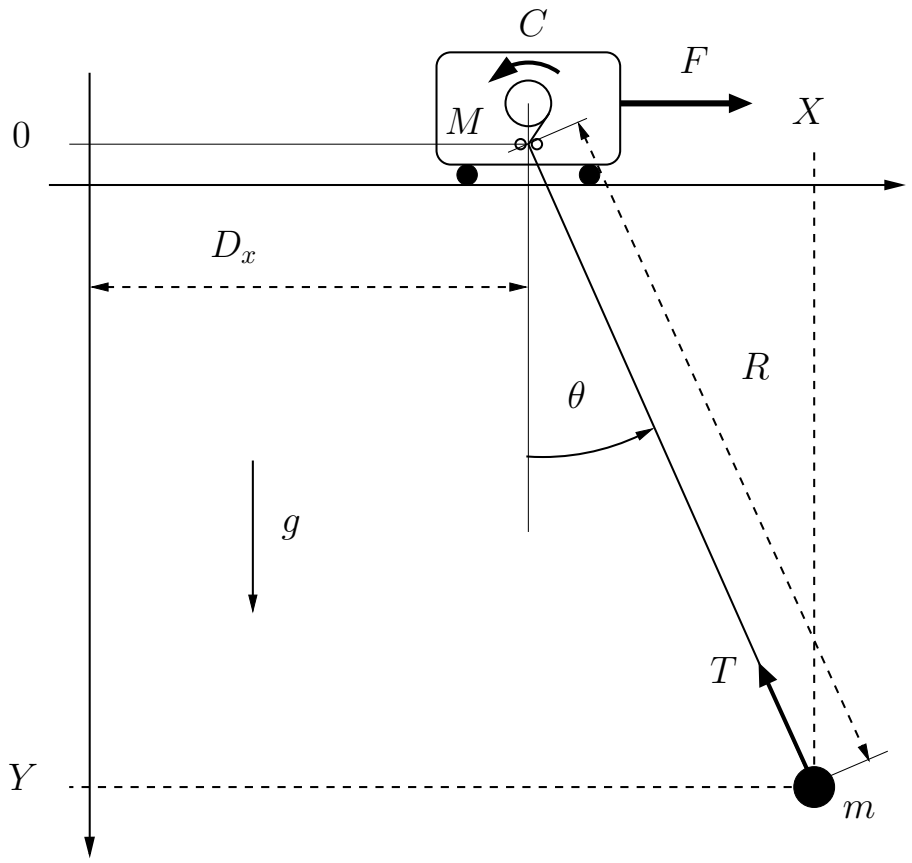


Fig. 8.1: Schéma d'une grue

R pour sa longueur, θ pour son angle avec la verticale, C pour le couple exercé au tambour du câble, ω sa vitesse angulaire, et F la force horizontale exercée au chariot. Les paramètres sont l'accélération par la gravité g, les masses m, de la charge, et M, du chariot, ainsi que le moment d'inertie J du tambour, son rayon ρ, et les coefficients de frottement visqueux c_r et c_d. Tous ces paramètres sont constants.

En éliminant l'angle θ entre le câble et la verticale, la force T et la vitesse angulaire ω on obtient une représentation algébrique (implicite) :

$$(\ddot{Y} - g)(X - D_x) = \ddot{X}Y \tag{8.9a}$$

$$(X - D_x)^2 + Y^2 = R^2 \tag{8.9b}$$

$$M\ddot{D}_x = F - c_\mathrm{d}\dot{D}_x - m\ddot{X} \tag{8.9c}$$

$$JY\ddot{R} = -Y(\rho C + c_\mathrm{r}\dot{R}) + \rho^2 mR(g - \ddot{Y}). \tag{8.9d}$$

Ce système d'équations comprend deux parties : la première décrit le mouvement pendulaire de la charge fixée au câble (avec les équations (8.9a) et (8.9b)), la

seconde décrit la cinétique du chariot et du tambour ((8.9c) et (8.9d)).

Pour la synthèse d'une loi de commande il conviendra de considérer ces deux parties séparément, en commençant par la synthèse d'une commande en boucle ouverte (ou fermée) pour le sous-système « pendule ». Elle servira pour calculer les références des boucles PID sous-jacentes [27, 26] — voir aussi la remarque à la p. 361.

Définition de l'extension de corps différentielle

Soit le corps différentiel de base $k = \mathbb{Q}\langle g, M, J, c_\mathrm{d}, m, \rho, c_\mathrm{r}\rangle$. Comme tous les paramètres sont supposés constants on a $k = \mathbb{Q}\langle g, M, J, c_\mathrm{d}, m, \rho, c_\mathrm{r}\rangle = \mathbb{Q}(g, M, J, c_\mathrm{d}, m, \rho, c_\mathrm{r})$, un corps des fractions rationnelles à coefficients rationnelles. Ensuite, soit le corps différentiel d'extension $\Sigma = k\langle X, Y, R, D_x, F, C\rangle$, dans lequel les relations non triviales sont données par (8.9). Il contient les fractions rationnelles en X, Y, R, D_x, F et C, avec des coefficients appartenant à k. Ainsi le système est défini comme l'extension de corps différentielle Σ/k.

Pour la construction de l'extension de corps différentielle Σ/k à partir de l'idéal différentiel correspondant aux équations (8.9) (voir #16) on peux s'appuyer sur une remarque dans la section 8.2. La structure particulière de (8.10) peut être exploitée. On travaille dans l'anneau différentiel $k\langle X, Y\rangle\{W_1, W_2\}$ des polynômes différentiels en les indéterminées W_1, W_2, à coefficients dans $k\langle X, Y\rangle$. Dans cet anneau on considère les deux polynômes différentiels

$$P_1 = (\ddot{Y} - g)(X - W_1) - \ddot{X}Y$$
$$P_2 = (X - W_1)^2 + Y^2 - W_2^2,$$

qu'on extrait des membres gauches des équations (8.9a) et (8.9b), en remplaçant D_x par W_1 et R par W_2. La structure de P_1 implique directement $W_1 \in k\langle X, Y\rangle$, donc $k\langle X, Y, W_1\rangle = k\langle X, Y\rangle$. Ensuite P_2 peut être considéré comme un polynôme appartenant à l'anneau différentiel $k\langle X, Y\rangle\{W_2\}$. Soit $\xi = \sqrt{(X - W_1)^2 + Y^2}$, pour simplifier. Alors, la structure de P_2 implique $W_2 \in k\langle X, Y\rangle\{\xi\}$. Comme $\xi \notin k\langle X, Y\rangle$ l'anneau différentiel $k\langle X, Y\rangle\{\xi\}$ ne contient pas de diviseur de zéro. Son corps de fractions est le corps différentiel $k\langle X, Y, R, D_x\rangle = k\langle X, Y, R\rangle$. Pour les variables F et C on obtient alors, de (8.9c) et (8.9d), que $F \in k\langle X, Y, R\rangle$ et $C \in k\langle X, Y, R\rangle$, et ainsi $\Sigma = k\langle X, Y, R\rangle$.

Une inspection des équations du système, (8.9), suffit ici pour constater que (F, C) est une base de transcendance différentielle de l'extension de corps différentielle Σ/k. Les équations (8.9) sont indépendantes, ce qui suit du fait que, en plus de X et Y, (8.9a) ne fait intervenir que D_x, (8.9b) fait intervenir R, (8.9c) fait intervenir F, et (8.9d) fait intervenir C. Somme tout il y a 6 variables du système dans ces 4 équations indépendantes, et on a donc $m = \deg\mathrm{tr}\,\mathrm{diff}\,\Sigma/k = 2$.

On peut choisir (F, C) comme entrée, ce qui donne la dynamique $\Sigma/k\langle F, C\rangle$. Pour le vérifier, on pourrait montrer que les équations (8.9) peuvent être réécrites tel que pour chacune des variables X, Y, R et D_x on obtient une é.d.o.

algébrique à coefficients dans $k\langle F, C\rangle$ (correspondant aux équations $Q = 0$ en section 8.3). Mais c'est compliqué, et peu utile. Il suffit, plutôt, de se servir du fait que le degré de transcendance non différentiel deg tr $\Sigma/k\langle F, C\rangle$ est fini si, et seulement si, deg tr diff $\Sigma/k\langle F, C\rangle = 0$, ce qui veut dire (F, C) est une entrée (#22). En reprenant encore (8.9), on observe qu'une base de transcendance non différentielle de $\Sigma/k\langle F, C\rangle$ est contenue dans $z = (X, Y, R, D_x, \dot{X}, \dot{Y}, \dot{D}_x, \ddot{Y})$. (Observons que z n'est pas algébriquement indépendante sur k, suite à (8.9b), et ne forme donc pas une base de transcendance de $\Sigma/k\langle F, C\rangle$.) Il en découle qu'une base de transcendance non différentielle de $\Sigma/k\langle F, C\rangle$ est finie : De (8.9a) on déduit directement $\ddot{X} \in k(z)$, et ainsi, de l'équation obtenue en dérivant (8.9b), $\dot{R} \in k(z)$, de (8.9c) il résulte $\ddot{D}_x \in k\langle F, C\rangle(z)$, et $\ddot{X} \in k(z)$ ainsi que $\ddot{Y} \in k\langle F, C\rangle(z)$ découlent de (8.9d). Par conséquent, aussi toutes les dérivées supérieures de X, R, D_x et Y appartiennent à $k\langle F, C\rangle(z)$.

Une seconde dynamique, importante dans la synthèse de la commande, résulte du choix de $u = (R, D_x)$ comme entrée. Le fait que l'extension de corps différentielle $\Sigma/k\langle u\rangle$ définissant cette dynamique est différentiellement algébrique, c'est-à-dire que deg tr diff $\Sigma/k\langle u\rangle = 0$, est une conséquence de $\Sigma = k\langle u\rangle(X, \dot{X}, Y, \dot{Y})$. (On poursuivra ceci en section 8.12.) Pour représenter cette dynamique $\Sigma/k\langle u\rangle$ il suffit des équations (8.9a) et (8.9b). Avec cette interprétation, F et C sont définis par (8.9c) et (8.9d) comme éléments de $k\langle u\rangle(X, \dot{X}, Y, \dot{Y}) = \Sigma$. Comme deg tr diff $\Sigma/k = 2 = \operatorname{card} u$, l'entrée $u = (R, D_x)$ est également indépendante. L'interprétation physique de la dynamique $\Sigma/k\langle u\rangle$ résulte des équations (8.9a) et (8.9b). Elle décrit le mouvement de la charge induite par les modifications de la longueur R du câble et de la position D_x du chariot.

Platitude

Le système Σ/k est plat, et les coordonnés $y = (X, Y)$ de la charge en forment une sortie plate [6, 27, 26, 39]. La discussion dans la section précédente nous a fait comprendre que D_x et R satisfont des équations algébriques (implicites) en y. Par une substitution de $(X - D_x)$ dans (8.9b) on obtient les relations

$$D_x = X - \frac{\ddot{X}Y}{\ddot{Y} - g} \tag{8.10a}$$

$$R^2 = \left(\frac{\ddot{X}Y}{\ddot{Y} - g}\right)^2 + Y^2. \tag{8.10b}$$

Ainsi, on pourrait se servir de (8.10), (8.9c) et (8.9d) pour calculer des expressions de F et C en y et ses dérivées. Il en découle, pour les clôtures algébriques des corps différentiels, $\overline{k\langle X, Y\rangle} = \overline{k\langle X, Y, R\rangle} = \overline{\Sigma}$.

On observe également que des deux relations (8.10) on ne peut pas éliminer R et D_x en même temps, ni X et Y, afin d'obtenir une é.d.o. uniquement en $y = (X, Y)$, ou uniquement en $u = (R, D_x)$: Les 2 équations en les 4 indéterminées sont indépendantes. Ceci correspond au fait que y comme u sont

différentiellement algébriquement indépendants sur k. (Comme l'on a vu dans la section précédente, le degré de transcendance différentiel deg tr diff $\Sigma/k = 2$, et avec ceci deg tr diff $\overline{\Sigma}/k =$ deg tr diff $\overline{k\langle X, Y\rangle}/k = 2$.)

Pour l'extension de corps différentielle Σ/k il en découle $\overline{k\langle y\rangle} = \overline{\Sigma}$, et $y = (X, Y)$ est donc différentiellement k-algébriquement indépendant. Par conséquent, le système Σ/k est différentiellement plat, et y, la position de la charge, en forme une sortie plate. Ceci simplifie la synthèse de commandes, en boucles ouvertes ou fermées, pour le transport de la charge. On observe, sur les équations (8.10), que la dynamique inverse $\Sigma/k\langle y\rangle$ par rapport à la sortie plate y est triviale : Le degré de transcendance non différentiel de $\Sigma/k\langle y\rangle$ est nul, $\Sigma/k\langle y\rangle$ donc une extension de corps algébrique.

Remarque 8.12.1. Cet exemple souligne qu'il convient d'introduire la clôture algébrique du corps Σ dans la définition de la platitude : La variable R n'est pas contenue dans le corps $k\langle X, Y\rangle$, mais dans $\overline{k\langle X, Y\rangle}$ si. Ceci correspond à la prise en compte de relation implicites (quadratiques ici).

Pour une grue dont le chariot peut être déplacé sur un plan horizontal, au lieu d'une seule droite, on peut procéder de façon analogue [26]. Avec une seconde coordonnée horizontale Z et une position D_z du chariot dans cette direction, on obtient pour le sous-système « pendule »

$$\begin{aligned}(\ddot{Y} - g)(X - D_x) &= \ddot{X}Y\\ (\ddot{Y} - g)(Z - D_z) &= \ddot{Z}Y\\ (X - D_x)^2 + (Z - D_z)^2 + Y^2 &= R^2.\end{aligned}$$

Encore les coordonnées de la charge (X, Y, Z) forment une sortie plate :

$$\begin{aligned}D_x &= X - \frac{\ddot{X}Y}{\ddot{Y} - g}\\ D_z &= Z - \frac{\ddot{Z}Y}{\ddot{Y} - g}\\ R^2 &= Y^2 + \left(\frac{\ddot{X}Y}{\ddot{Y} - g}\right)^2 + \left(\frac{\ddot{Z}Y}{\ddot{Y} - g}\right)^2.\end{aligned}$$

Comme dans le cas du mouvement dans le plan vertical, ce système peut aisément être étendu par des équations pour la dynamique du chariot et du tambour.

Planification de trajectoires

La grue est utilisée pour transporter des charges entre deux positions de repos. Pendant un transport rapide, des mouvements pendulaires importants peuvent se produire. En prenant en compte leur dynamique (non linéaire) on peut admettre ces mouvements, au lieu d'essayer de les éviter, tel qu'il serait

nécessaire pour rester dans un domaine d'utilité d'un modèle linéarisé. Ainsi, il convient de planifier une commande sur la base du sous-système « pendule », qui est suffisamment lent pour ne pas exciter les mouvements rapides non modélisés, tout en étant assez rapide à ce que des mouvements pendulaires importants de la charge se produisent. Le but sera alors d'arriver à la position finale, de repos, sans qu'il y ait des oscillations de la charge. En se servant de la sortie plate $y = (X, Y)$ la synthèse d'une telle commande est largement simplifiée, car le problème est défini en ces coordonnées.

La commande s'ensuit directement d'une trajectoire de référence pour la sortie plate, $t \mapsto y_{\text{réf}}(t) = \big(X_{\text{réf}}(t), Y_{\text{réf}}(t)\big)$, avec les équations (8.10) :

$$D_{x,\text{réf}}(t) = X_{\text{réf}}(t) - \frac{\ddot{X}_{\text{réf}}(t) Y_{\text{réf}}(t)}{\ddot{Y}_{\text{réf}}(t) - g} \tag{8.11a}$$

$$R_{\text{réf}}(t) = \sqrt{(Y_{\text{réf}}(t))^2 + \left(\frac{\ddot{X}_{\text{réf}}(t) Y_{\text{réf}}(t)}{\ddot{Y}_{\text{réf}}(t) - g}\right)^2}. \tag{8.11b}$$

Notons qu'aucun paramètre n'apparaît dans ces équations : Ainsi la commande est indépendante de la masse (inconnue) de la charge.

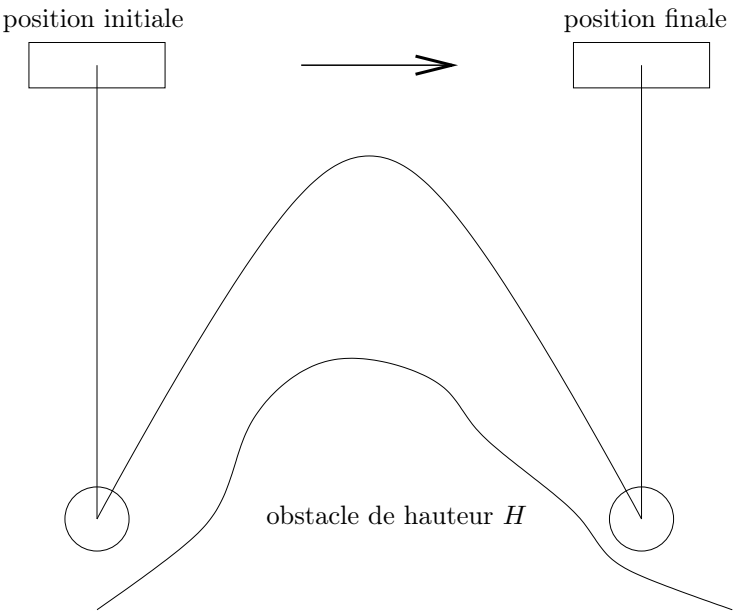


Fig. 8.2: Transport d'une charge

Il faut donc choisir des trajectoires de référence pour X et Y, qui seront ensuite substituées dans les équations ci-dessus. On y voit que les trajectoires

doivent être au moins deux fois différentiables. Toutefois, ayant négligé les dynamiques du chariot et du tambour, il convient de demander plus de régularité, par exemple des trajectoires trois fois différentiables, afin de pouvoir les utiliser comme références pour les contrôleurs PD ou PID sous-jacents des moteurs électriques (voir la remarque à la p. 361).

Les positions initiale et finale déterminent les valeurs initiale et finale des trajectoires de référence, $\big(X_{\text{réf}}(0), Y_{\text{réf}}(0)\big)$ et $\big(X_{\text{réf}}(t_*), Y_{\text{réf}}(t_*)\big)$. Les positions initiale et finale devant être des positions de repos, les valeurs initiales et finales des dérivées doivent être nulles. En même temps, il faut faire attention à la condition $\ddot{Y}_{\text{réf}}(t) < g$, sinon les valeurs de fractions intervenant dans les calculs ne seront pas bornées. Cette condition pour l'accélération correspond en même temps au fait que le câble doit toujours être tendu, ce qui est une hypothèse pour la dérivation du modèle.

Le chemin parcouru par la charge dans le plan vertical, entre les deux positions de repos, est encore libre ; ça permet d'éviter des obstacles de position et taille connues. On peut ainsi choisir, par exemple, un chemin parabolique contournant un obstacle de hauteur H (marges comprises) (voir la figure 8.2) :

$$Y_{\text{réf}}(X_{\text{réf}}) = Y(0) - H + H\left(\frac{2X_{\text{réf}} - X(0) - X(t_*)}{X(0) - X(t_*)}\right)^2. \tag{8.12}$$

Les deux trajectoires pour X et Y sont ainsi découplées, et il suffit donc de choisir celle de la position horizontale X, par exemple. En choisissant celle de (8.12), les positions verticales de la charge au début et à la fin du transfert seront égales : $Y_{\text{réf}}(0) = Y_{\text{réf}}(t_*)$.

On peut choisir une trajectoire polynômiale pour X, ce qui mène à des équations linéaires pour les paramètres. Ayant, somme tout, six conditions initiales et finales on a besoin d'un polynôme de degré 5 pour $X_{\text{réf}}$:

$$X_{\text{réf}}(t) = X_{\text{réf}}(0) + \big(X_{\text{réf}}(t_*) - X_{\text{réf}}(0)\big)\frac{t^3}{t_*^3}\left(10 - 15\frac{t}{t_*} + 6\frac{t^2}{t_*^2}\right). \tag{8.13}$$

Le choix de $t \mapsto X_{\text{réf}}(t)$ fixant ainsi le mouvement horizontal, le mouvement vertical suit du chemin parabolique calculé avec (8.12). Le choix d'un temps de transport t_* suffisamment grand garantit que la condition sur l'accelération, $\ddot{Y} < g$, sera respectée. Le temps minimal requis avec une trajectoire polynômiale de degré 5 peut être calculé. Pour réduire ce temps, on peut choisir d'autres paramétrisations de la trajectoire de référence $t \mapsto X_{\text{réf}}(t)$. On fera toutefois attention à ne pas choisir des mouvements trop rapides, afin de pouvoir négliger les sous-systèmes du chariot et du tambour avec leurs contrôleurs (voir la remarque suivante).

Remarques 8.12.1. Pour la réalisation de la commande on peut se servir de contrôleurs PD ou PID au niveau du chariot et du tambour. Ils déterminent d'une part la force F, ceci sur la base de l'écart $\Delta D_x = D_x - D_{x,\text{réf}}$ entre la

position D_x du chariot et sa référence $D_{x,\text{réf}}$ résultant de la boucle extérieure, d'autre part le couple C, à partir de l'écart $\Delta R = R - R_{\text{réf}}$ entre la longueur du câble et sa référence $R_{\text{réf}}$:

$$\begin{aligned} F &= F_{\text{réf}}(t) + \Delta F \\ &= M\ddot{D}_{x,\text{réf}}(t) + c_{\text{d}}\dot{D}_{x,\text{réf}}(t) + m\ddot{X}_{\text{réf}}(t) - k_{\text{P}}\Delta D_x - k_{\text{D}}\Delta\dot{D}_x \qquad (8.14a) \\ C &= C_{\text{réf}}(t) + \Delta C \\ &= -\frac{J}{\rho}\ddot{R}_{\text{réf}}(t) - \frac{c_{\text{r}}}{\rho}\dot{R}_{\text{réf}}(t) + \rho\frac{mR_{\text{réf}}(t)(g - \ddot{Y}_{\text{réf}}(t))}{Y_{\text{réf}}(t)} - l_{\text{P}}\Delta R - l_{\text{D}}\Delta\dot{R}. \qquad (8.14b) \end{aligned}$$

Pour le choix des paramètres $k_{\text{P}}, k_{\text{D}}, l_{\text{P}}$ et l_{D} on peut regarder le comportement autour d'un point stationnaire. Dans ces points le câble est en position verticale, avec des valeurs constantes (arbitraires) $X_{\text{r}} = D_{x,\text{r}}$ et $Y_{\text{r}} = R_{\text{r}}$ et la force $T_{\text{r}} = mg$ dans le câble. En linéarisant le système (implicite) (8.10) du sous-système pendulaire, et les équations (8.8e) et (8.8f) du chariot et du tambour avec (8.8a), (8.8b) et (8.8g) on obtient

$$\Delta\ddot{X} + \frac{g}{R_{\text{r}}}\Delta X = \frac{g}{R_{\text{r}}}\Delta D_x \qquad (8.15a)$$

$$M\Delta\ddot{D}_x = \Delta F - m\Delta\ddot{X} - c_{\text{d}}\Delta\dot{D}_x \qquad (8.15b)$$

$$-\frac{J}{\rho}\Delta\ddot{R} = \Delta C + \rho m\Delta\ddot{R} + \frac{c_{\text{r}}}{\rho}\Delta\dot{R} \qquad (8.15c)$$

avec $\Delta X = X - X_{\text{réf}}$. Bien sur, l'équation (8.15a) décrit les oscillations d'un pendule mathématique de longueur R_{r}.

Les paramètres l_{P} et l_{D} des contrôleurs PD (8.14b) pour le chariot peuvent être choisis en spécifiant les valeurs propres du système linéarisé (8.15c). La paramétrisation du contrôleur PD (8.14a) du chariot est légèrement plus délicate, car les équations (8.15b) et (8.15a) sont couplées : Un mouvement de la charge induit un déplacement du chariot. On choisira donc les paramètres du contrôleur PD pour réduire cet effet.

Pour ceci on peut, en se servant de (8.15a), substituer la partie du contrôleur (8.14a) qui est intéressante pour les petits mouvements, à savoir $\Delta F = -k_{\text{P}}\Delta D_x - k_{\text{D}}\Delta\dot{D}_x$ et puis $\Delta\ddot{X}$, dans (8.15b). On obtient ainsi

$$M\Delta\ddot{D}_x = -k_{\text{P}}\Delta D_x - k_{\text{D}}\Delta\dot{D}_x - m\left(\frac{g}{R_{\text{r}}}\Delta D_x - \frac{g}{R_{\text{r}}}\Delta X\right) - c_{\text{d}}\Delta\dot{D}_x,$$

ou encore

$$\Delta\ddot{D}_x + \left(\frac{k_{\text{D}} + c_{\text{d}}}{M}\right)\Delta\dot{D}_x + \left(\frac{k_{\text{P}}}{M} + \frac{mg}{MR_{\text{r}}}\right)\Delta D_x = \frac{mg}{MR_{\text{r}}}\Delta X.$$

On choisit ensuite les coefficients dans ce système selon

$$\left(\frac{k_{\text{D}} + c_{\text{d}}}{M}\right) = \frac{\lambda_1 + \lambda_2}{\varepsilon} \qquad \text{et} \qquad \left(\frac{k_{\text{P}}}{M} + \frac{mg}{MR_{\text{r}}}\right) = \frac{\lambda_1\lambda_2}{\varepsilon^2},$$

où λ_1 et λ_2 sont de l'ordre de grandeur de la constante de temps du pendule, $\sqrt{g/R_{\mathrm{r}}}$, et $\varepsilon = 1/10$. Il en résultent les gains k_{P} et k_{D}. Avec ce choix des paramètres, le déplacement ΔD_x du chariot oscille (à peu près) dix fois plus rapidement que la charge, qui, quant à elle, n'exerce que peu d'influence sur le mouvement du chariot. La base de ce découplage entre mouvements lents et rapides est un argument de perturbations singulières [37, 60].

Afin d'éviter un écart permanent entre la position finale et sa consigne dans le cas d'incertitude sur la masse m de la charge, il convient de compléter le contrôleur pour la longueur R du câble par une partie intégrale. En plus, on peut introduire un petit filtre pour réaliser la partie dérivateur. En outre une simplification peut également être envisagé : On peut utiliser $F_{\mathrm{réf}} = 0$ et $C_{\mathrm{réf}} = \rho m g$, au lieu des références variables $F_{\mathrm{réf}}(t)$ et $C_{\mathrm{réf}}(t)$, dans (8.14).

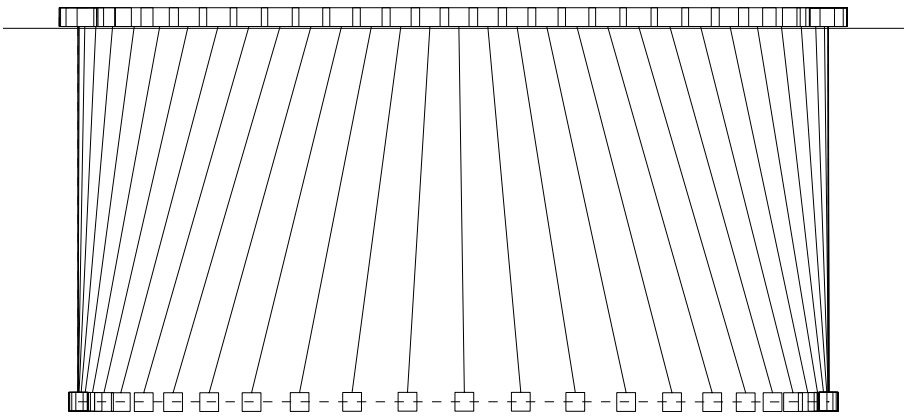


FIG. 8.3: Simulation de la commande de la grue : Transfert horizontal avec une commande en boucle ouverte extérieure et des contrôleurs PID en cascade.

Le résultat d'une simulation d'un transfert horizontal de la charge, avec la commande en boucle ouverte basée sur la platitude, est reporté sur la figure 8.3 ; les trajectoires de référence sont celles avec $Y_{\mathrm{réf}}(t) = Y_{\mathrm{réf}}(0)$ et le polynôme de degré 5 dans (8.13) pour $X_{\mathrm{réf}}(t)$. Pour comparer, la figure 8.4 montre les trajectoires obtenues sans la planification basée sur la platitude, mais avec des consignes constantes pour la longueur du câble, $R_{\mathrm{réf}}(t) = R(0)$, et pour la position du chariot, $D_{x,\mathrm{réf}}(t) = X_{\mathrm{réf}}(t_*)$ (avec les mêmes paramètres comme pour la figure 8.3). Enfin le résultat d'une simulation d'un transport le long d'un chemin parabolique, tel que discuté ci-dessus, est illustré dans la figure 8.5. Pour toutes les simulations les contrôleurs PID discutés ci-dessus ont été utilisés.

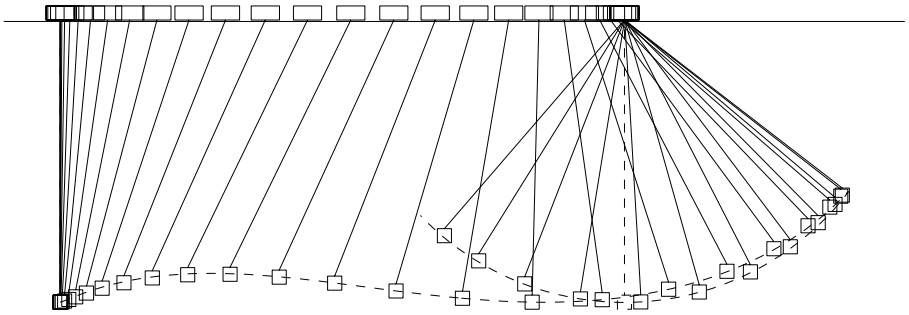


Fig. 8.4: Simulation de la commande de la grue : Transfert horizontal dans le même intervalle que dans la figure 8.3, mais avec les consignes fixes $R_{\text{réf}}(t) = R(0)$ et $D_{x,\text{réf}}(t) = X_{\text{réf}}(t_*)$ pour les contrôleurs PID.

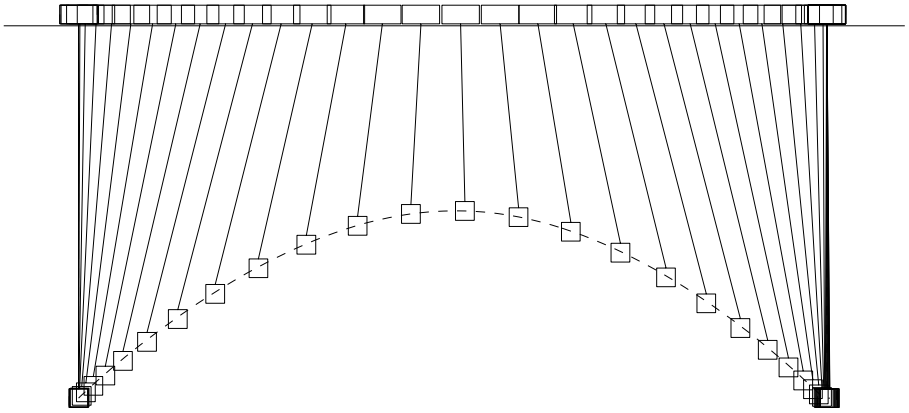


Fig. 8.5: Simulation de la commande par platitude : Transport d'une charge le long d'un chemin parabolique.

Représentations d'état généralisées

La dimension d'état de la dynamique $\Sigma/k\langle R, D_x\rangle$ est égale à 2. En choisissant l'état $x = (x_1, x_2)$ avec

$$x_1 = \frac{Y}{X - D_x}, \quad x_2 = \dot{Y}(X - D_x) - (\dot{X} - \dot{D}_x)Y$$

on obtient une représentation d'état (généralisée) dans laquelle aucune dérivée de la longueur de câble R n'apparaît. Si l'on remplace X, Y et leurs dérivées en

se servant de (8.9a) et de (8.9b)

$$\dot{x}_1 = \frac{x_2(1+x_1^2)}{R^2} \tag{8.16a}$$

$$\dot{x}_2 = \frac{(\ddot{D}_x x_1 + g)R}{\sqrt{1+x_1^2}}. \tag{8.16b}$$

(Notons que $\sqrt{1+x_1^2} = R/(X - D_x)$ est un élément de Σ.) Cette représentation pourrait être utile si l'on considère l'accélération $\ddot{D}_x$ du chariot avec la longueur R du câble comme entrée. Dans ce cas il s'agit d'une représentation d'état classique pour la dynamique $\Sigma/k\langle R, \ddot{D}_x\rangle$

On peut éliminer $\ddot{D}_x$ de cette représentation d'état (généralisée) de $\Sigma/k\langle R, D_x\rangle$ en choisissant un état (généralisé) $\bar{x}$ avec les composants $\bar{x}_1 = x_1$ et $\bar{x}_2 = x_2\sqrt{1+x_1^2} - \dot{D}_x x_1 R$. Dans la représentation d'état (généralisée) résultante on voit apparaître les dérivées premières $\dot{D}_x$ et $\dot{R}$:

$$\dot{\bar{x}}_1 = \frac{(\bar{x}_2 + \dot{D}_x R\bar{x}_1)\sqrt{1+\bar{x}_1^2}}{R^2} \tag{8.17a}$$

$$\dot{\bar{x}}_2 = gR + \frac{(\bar{x}_2 + \dot{D}_x R\bar{x}_1)^2\bar{x}_1}{R^2\sqrt{1+\bar{x}_1^2}} - \frac{\dot{D}_x(\bar{x}_2 + \dot{D}_x R\bar{x}_1)\sqrt{1+\bar{x}_1^2}}{R} - \bar{x}_1\dot{R}\dot{D}_x. \tag{8.17b}$$

Il n'est donc pas possible d'éliminer à la fois toutes les dérivées, c'est-à-dire la dynamique $\Sigma/k\langle R, D_x\rangle$ n'admet aucune représentation d'état classique [7].

Remarque 8.12.2. Notons qu'il y a des singularités pour $R = 0$ et pour $Y = 0$ (à cause d'une division par zéro). Toutefois, ceci n'a pas d'importance pratique, car le câble ne sera jamais complètement enroulé et l'on fera attention à ce qu'il soit toujours pendant.

La forme de commande généralisée associé à la dynamique $\Sigma/k\langle R, D_x\rangle$ et son état généralisé $\xi = (X, \dot{X})$ peuvent également être déduits des équations du système, (8.9a) et (8.9b) :

$$\dot{\xi}_1 = \xi_2 \tag{8.18a}$$

$$\dot{\xi}_2 = -\frac{\xi_1 - D_x}{R^2}\Big[g\sqrt{R^2-(\xi_1-D_x)^2} + (\xi_2 - \dot{D}_x)^2 - R\ddot{R} - \dot{R}^2 + \frac{(R\dot{R} - (\xi_1 - D_x)(\xi_2 - \dot{D}_x))^2}{R^2-(\xi_1-D_x)^2} - \ddot{D}_x(\xi_1 - D_x)\Big] \tag{8.18b}$$

avec

$$X = \xi_1, \qquad Y = \sqrt{R^2-(\xi_1-D_x)^2}. \tag{8.18c}$$

(Notons encore que la racine $\sqrt{R^2-(\xi_1-D_x)^2}$ appartient au corps $k\langle X, Y\rangle$.)

Synthèse d'une commande pour la poursuite de trajectoires

La forme de commande (généralisée) trouvée ci-dessus peut servir dans la synthèse de la commande. En choisissant les entrées $v_1 = \ddot{X}$ et $v_2 = Y$ ont obtient une forme de Brunovský. Ce choix de v définit un bouclage quasi statique de l'état ξ. On voit sur (8.18b) et (8.18c) que $v = (v_1, v_2)$ dépend de ξ, u, $\dot{u}$ et $\ddot{u}$. Pour trouver la loi de commande à implémenter il faut toutefois exprimer u en fonction de ξ, de v et les dérivées de ce dernier. Si l'on se sert de la représentation d'état (8.18) pour retrouver les relations requises les calculs sont compliqués. En revanche il est simple de les déduire directement des équations implicites (8.9) du système. Il suffit de remplacer, dans (8.10a) et (8.10b), les variables $\ddot{X}$ par v_1, Y par v_2, $\ddot{Y}$ par $\ddot{v}_2$ et X par ξ_1 :

$$D_x = \xi_1 - \frac{v_1 v_2}{\ddot{v}_2 - g} \tag{8.19a}$$

$$R^2 = \left(\frac{v_1 v_2}{\ddot{v}_2 - g}\right)^2 + v_2^2. \tag{8.19b}$$

On s'est servi ici de la représentation d'état explicite (8.18), et l'on résoudra aussi (8.19b) en utilisant la racine positive, car la longueur R du câble est toujours positive. Ainsi, aussi pour la commande, les équations implicites font apparaître une singularité ($\ddot{v}_2 = g$) que l'on devra surveiller dans le calcul des commandes.

On voit cette singularité également dans la définition du bouclage d'état : Les équations du sous-système « pendule » sont données par (8.9a) et (8.9b), celle de la commande par (8.19). Utilisant (8.19a) et $\xi_1 = X$ dans (8.19b) et soustrayant (8.19b) de (8.9b) on obtient $Y^2 = v_2^2$, c'est-à-dire Y est ou bien égal à $+v_2$, ou bien à $-v_2$. Mais la commande a été obtenue en choisissant $Y = v_2$. Qu'est-ce qui c'est passé ? Pour la commande on utilise (8.19). Le système d'équations résultant de la loi de commande implicite et des équations implicites du modèle admet deux branches de solution. Toutefois, d'un point de vue pratique, il est clair que l'on choisira $v_2 > 0$, $R > 0$ (et la position de référence verticale positive, pour éviter le voisinage de la singularité à zéro). La branche de la solution correspondante est $Y = v_2$.

Remarque 8.12.3. Dans le cadre algébrique les deux branches de la solution correspondent à la construction de deux idéaux différentiels premiers, I_1 et I_2, tels que leur intersection correspond à l'idéal différentiel I engendré par les polynômes dans les équations (8.18) et (8.19) ; ce dernier n'est pas premier (#3). Ainsi I_1 contient l'antécédent de $(Y - v_2)$, et I_2 celui de $(Y + v_2)$, alors que I ne contient que l'antécédent de $(Y^2 - v_2^2)$ mais ni celui de $(Y - v_2)$ ni celui de $(Y + v_2)$. Ce n'est qu'en choisissant un des idéaux premiers, I_1 ou I_2, que l'on peut construire le corps représentant le système de la boucle fermée. La construction de bouclages d'états par le choix d'une nouvelle entrée v dans le corps Σ du système est donc avantageux. Le fait de rester dans le corps correspond au choix d'une branche de la solution.

La stabilisation autour de la trajectoire de référence est acquise en employant le bouclage d'état supplémentaire (linéaire)

$$v_1 = \ddot{X}_{\text{réf}}(t) + k_1(\xi_2 - \dot{X}_{\text{réf}}(t)) + k_0(\xi_1 - X_{\text{réf}}(t)) \tag{8.20a}$$
$$v_2 = Y_{\text{réf}}(t) \tag{8.20b}$$

avec des paramètres (gains) constants $\mathbb{R} \ni k_0, k_1 < 0$. Ainsi on obtient une dynamique linéaire et stable pour l'erreur de poursuite $(X - X_{\text{réf}}(t), Y - Y_{\text{réf}}(t))$. La dérivée $\ddot{v}_2$ apparaissant dans (8.19) est remplacée par $\ddot{Y}_{\text{réf}}(t)$.

Remarques 8.12.2. 1. La sortie plate y complète ne peut être inclue dans l'état de la dynamique $\Sigma/k\langle R, D_x\rangle$, car avec (8.9b) les deux composantes sont reliées par une équation polynômiale à coefficients dans $k\langle R, D_x\rangle$.
2. On pourrait aussi choisir une dynamique d'ordre deux pour l'erreur dans la direction verticale, en utilisant l'état $\bar{\xi} = (Y, \dot{Y})$ de $\Sigma/k\langle R, D_x\rangle$. Toutefois, en inspectant les symétries du problème tridimensionnel (au lieu du mouvement dans un plan vertical, voir section 8.12) on comprend que le choix fait est « plus naturel » : Les coordonnées dans les directions des x et des z sont sur le même pied. On peut déplacer ou tourner le système des coordonnées du plan horizontal, sans qu'il n'en résulte une modification des équations. Il convient ainsi d'associer des chaînes d'intégrateurs de la même longueur, 2, aux coordonnées X et Z de la charge. On trouvera plus de résultats sur les symétries et les bouclages invariants correspondants dans [49, 40].

Un pendule sur un chariot : exemple d'un système non plat

En remplaçant le câble de la grue par une barre de longueur fixe de masse négligeable on obtient un pendule avec un point de suspension bougeant sur une droite horizontale (avec le chariot). (On peut, bien sûr, également représenter n'importe quel pendule physique planaire par un pendule mathématique équivalent.) On obtient ainsi un modèle du mouvement en substituant $\dot{R} = 0$ dans (8.9) :

$$\begin{aligned}(\ddot{Y} - g)(X - D_x) &= \ddot{X}Y \\ (X - D_x)^2 + Y^2 &= R^2 \\ M\ddot{D}_x &= F - c_{\text{d}}\dot{D}_x - m\ddot{X},\end{aligned}$$

la longueur R étant maintenant un paramètre constant. En analogie avec la démarche pour la grue on peut asservir la position D_x du chariot, avec une structure cascadée, tel que le système en considération se réduit à

$$(\ddot{Y} - g)(X - D_x) = \ddot{X}Y \tag{8.21a}$$
$$(X - D_x)^2 + Y^2 = R^2. \tag{8.21b}$$

Alors les variables du système sont les coordonnées X, Y de la charge et celle du chariot D_x, et le corps différentiel du système est $k\langle X, Y, D_x\rangle$, car $g, R \in k$.

Nous vérifions que le système $k\langle X, Y, D_x\rangle/k$ n'est pas plat. Une première possibilité est donnée par l'analyse du système linéaire tangent $\Omega_{k\langle X,Y,D_x\rangle/k}$. Ces équations résultent de (8.21) :

$$(X - D_x)\, d\ddot{Y} + (\ddot{Y} - g)\,(dX - dD_x) - \ddot{X}\, dY - Y\, d\ddot{X} = 0 \tag{8.22a}$$

$$(X - D_x)\,(dX - dD_x) + Y\, dY = 0. \tag{8.22b}$$

L'analyse est simplifié par le choix d'un point d'équilibre particulier. (Les équations valables autour d'un point particulier résultent de celles de $\Omega_{k\langle X,Y,D_x\rangle/k}$, bien que ce ne soit pas le même système.) S'il existe un point d'équilibre autour duquel le linéarisé est commandable, alors $\Omega_{k\langle X,Y,D_x\rangle/k}$ est également commandable. Par contre, on ne peut pas déduire la non-commandabilité de $\Omega_{k\langle X,Y,D_x\rangle/k}$ de la non-commandabilité autour d'un point particulier.

Considérons donc d'abord des points d'équilibre, avec le pendule dans une de ces positions verticales, où $X = D_x$. Alors, de (8.22) on obtient, bien sûr, l'équation de l'oscillateur linéaire

$$-g\,(dX - dD_x) - Y\, d\ddot{X} = 0,$$

ou encore

$$d\ddot{X} + \frac{g}{Y}\, dX = \frac{g}{Y}\, dD_x.$$

Ce système linéaire constant admet une base (sortie plate) dX. Il est donc commandable, et $\Omega_{k\langle X,Y,D_x\rangle/k}$ aussi. L'analyse de la commandabilité du système linéarisé ne permet donc pas de conclure sur la platitude du système $k\langle X, Y, D_x\rangle/k$.

Remarque 8.12.4. Démontrer la commandabilité du système linéaire tangent (instationnaire) $\Omega_{k\langle X,Y,D_x\rangle/k}$, obtenu en linéarisant autour de trajectoires, en exhibant une sortie plate (base) est un peu plus compliqué. On peut d'abord éliminer dD_x avec (8.22b) :

$$-(X - D_x)^2\, d\ddot{Y} + (Y(\ddot{Y} - g) + (X - D_x)\ddot{X})\, dY + (X - D_x)Y\, d\ddot{X} = 0.$$

Cette équation a la forme

$$d\ddot{X} - \gamma\, d\ddot{Y} = \beta\, dY$$

avec

$$\beta = \frac{\ddot{Y} - g}{(X - D_x)} + \frac{\ddot{X}}{Y} \quad \text{et} \quad \gamma = \frac{X - D_x}{Y}.$$

On montre alors que

$$b = (\beta + \ddot{\gamma})\,(dX - \gamma\, dY) + 2\dot{\gamma}(d\dot{X} - \gamma\, d\dot{Y} + \dot{\gamma}\, dY)$$

est une base du module $\Omega_{k\langle X,Y,D_x\rangle/k}$.

Nous avons vu que nous ne pouvons pas répondre à la question de la platitude du système $k\langle X, Y, D_x\rangle/k$ uniquement sur la base du système linéaire tangent. Considérons donc une autre condition nécessaire, le critère des surfaces réglées de la section 8.4).

Essayons d'abord de l'appliquer aux équations (8.21). Pour ceci considérons les variables du système et leurs dérivées comme des grandeurs indépendantes, et notons $\xi_{1,0}$ pour X, $\xi_{1,1}$ pour $\dot{X}$, $\xi_{1,2}$ pour $\ddot{X}$ et $\xi_{2,i}$, $i = 0, 1, 2$, pour Y, $\dot{Y}$ et $\ddot{Y}$, ainsi que $\xi_{3,i}$, $i = 0, 1, 2$, pour D_x, $\dot{D}_x$ et $\ddot{D}_x$. Avec ces notations, le système d'équations algébrique à étudier s'écrit

$$\begin{aligned}(\xi_{2,2} - g)(\xi_{1,0} - \xi_{3,0}) &= \xi_{1,2}\xi_{2,0}\\ (\xi_{1,0} - \xi_{3,0})^2 + \xi_{2,0}^2 &= R^2.\end{aligned}$$

Si ce système est plat, il existe $a_1, a_2, a_3 \in k(\xi_{1,0}, \xi_{2,0}, \xi_{3,0}, \xi_{1,1}, \xi_{2,1}, \xi_{3,1})$ tels que, pour des $\lambda \in k$ arbitraires on ait

$$\begin{aligned}(\xi_{2,2} + \lambda a_2 - g)(\xi_{1,0} - \xi_{3,0}) &= (\xi_{1,2} + \lambda a_1)\xi_{2,0}\\ (\xi_{1,0} - \xi_{3,0})^2 + \xi_{2,0}^2 &= R^2.\end{aligned}$$

On voit que cette condition est satisfaite avec $a = (0, 0, a_3)$ pour des a_3 arbitraires. Appliquer le critère des surfaces réglées aux équations (8.21) ne permet donc pas non plus de répondre à la question de la platitude du système $k\langle X, Y, D_x\rangle/k$. Toutefois, cette condition nécessaire dépend du choix des équations considérées. On le verra tout de suite.

En posant $\dot{R} = 0$ dans la représentation d'état (généralisée) (8.17) pour la grue (p. 365) on obtient une représentation d'état pour le mouvement horizontal du pendule dans la forme

$$\dot{\bar{x}}_1 = \frac{(\bar{x}_2 + \dot{D}_x R\bar{x}_1)\sqrt{1+\bar{x}_1^2}}{R^2} \tag{8.23a}$$

$$\dot{\bar{x}}_2 = gR + \frac{(\bar{x}_2 + \dot{D}_x R\bar{x}_1)^2\bar{x}_1}{R^2\sqrt{1+\bar{x}_1^2}} - \frac{\dot{D}_x(\bar{x}_2 + \dot{D}_x R\bar{x}_1)\sqrt{1+\bar{x}_1^2}}{R}, \tag{8.23b}$$

avec $\bar{x}_1 = \frac{Y}{X - D_x}$ et $\bar{x}_2 = x_2\sqrt{1+x_1^2} - \dot{D}_x x_1 R$. Afin d'appliquer le critère des surfaces réglées de la section 8.4, introduisons $\bar{\xi}_{1,0} = \bar{x}_1$, $\bar{\xi}_{1,1} = \dot{\bar{x}}_1$, $\bar{\xi}_{2,0} = \bar{x}_2$, $\bar{\xi}_{2,1} = \dot{\bar{x}}_2$, $\bar{\xi}_{3,0} = D_x$ et $\bar{\xi}_{3,1} = \dot{D}_x$. Il vient

$$\bar{\xi}_{1,1} = \frac{(\bar{\xi}_{2,0} + \bar{\xi}_{3,1} R\bar{\xi}_{1,0})\sqrt{1+\bar{\xi}_{1,0}^2}}{R^2} \tag{8.24a}$$

$$\bar{\xi}_{2,1} = gR + \frac{(\bar{\xi}_{2,0} + \bar{\xi}_{3,1} R\bar{\xi}_{1,0})^2\bar{\xi}_{1,0}}{R^2\sqrt{1+\bar{\xi}_{1,0}^2}} - \frac{\bar{\xi}_{3,1}(\bar{\xi}_{2,0} + \bar{\xi}_{3,1} R\bar{\xi}_{1,0})\sqrt{1+\bar{\xi}_{1,0}^2}}{R}. \tag{8.24b}$$

Si le système est plat, alors il est possible de trouver $a_1, a_2, a_3 \in k(\bar{\xi}_{1,0}, \bar{\xi}_{2,0}, \bar{\xi}_{3,0})$ tels que, pour des $\lambda \in k$ arbitraires, on ait

$$\begin{aligned}
\bar{\xi}_{1,1} + \lambda a_1 &= \frac{(\bar{\xi}_{2,0} + (\bar{\xi}_{3,1} + \lambda a_3) R \bar{\xi}_{1,0}) \sqrt{1 + \bar{\xi}_{1,0}^2}}{R^2} \\
\bar{\xi}_{2,1} + \lambda a_2 &= gR + \frac{(\bar{\xi}_{2,0} + (\bar{\xi}_{3,1} + \lambda a_3) R \bar{\xi}_{1,0})^2 \bar{\xi}_{1,0}}{R^2 \sqrt{1 + \bar{\xi}_{1,0}^2}} - \\
&\qquad \frac{(\bar{\xi}_{3,1} + \lambda a_3)(\bar{\xi}_{2,0} + (\bar{\xi}_{3,1} + \lambda a_3) R \bar{\xi}_{1,0}) \sqrt{1 + \bar{\xi}_{1,0}^2}}{R}.
\end{aligned}$$

En prenant en compte (8.24), qui correspondent au cas $\lambda = 0$, la première de ces équations donne

$$a_1 = a_3 \frac{\bar{\xi}_{1,0} \sqrt{1 + \bar{\xi}_{1,0}^2}}{R},$$

alors que la seconde mène à une équation quadratique pour a_3. Pour des $\lambda \in k$ arbitraires, celle-ci ne peut être satisfaite qu'avec $a_3 = 0$. Mais ceci implique nécessairement $a_1 = 0$ et $a_2 = 0$, et le système $k\langle X, Y, D_x\rangle/k$ n'est donc pas plat.

Une autre possibilité pour démontrer que ce système n'est pas plat s'appuie sur le fait qu'il s'agit d'un système admettant une seule entrée indépendante. Comme discuté en section 8.7 (voir la remarque à la p. 350), dans le cas $m = 1$ un bouclage d'état quasi statique d'un état classique est nécessairement un bouclage d'état statique. L'équivalence d'un système plat à une forme de Brunovský, ayant une représentation d'état classique, implique que la forme de commande du système est également classique. Si elle est plate, la dynamique $k\langle X, Y, D_x\rangle/k\langle D_x\rangle$ doit donc admettre une représentation d'état classique. Le membre droit de la représentation d'état généralisée (8.23) de la dynamique $k\langle X, Y, D_x\rangle/k\langle D_x\rangle$ n'est pas affine par rapport à la dérivée d'ordre maximal $\dot{D}_x$ de D_x. Ceci implique, avec la condition nécessaire de la section 8.5, que l'on ne peut pas éliminer $\dot{D}_x$ par le choix d'un autre état généralisé : Aucune représentation classique n'existe, et le système $k\langle X, Y, D_x\rangle/k$ ne peut donc pas être plat[5].

8.13 Bibliographie

[1] Bass, H., A. Buium et P. J. Cassidy (rédacteurs): *Selected works of Ellis Kolchin with commentary.* Amer. Math. Soc., 1999.

[5] Finalement, on peut aussi démontrer que $k\langle X, Y, D_x\rangle/k$ n'est pas plat en introduisant une représentation d'état classique (par ex. (8.16), en interprétant $\ddot{D}_x$ comme entrée, ou encore en incluant les équations différentielles du chariot et considérant la force F comme entrée). Alors on peut se servir des conditions de linéarisabilité par bouclage d'état statique (régulier) issu de la géométrie différentielle [33, 31, 59, 32, 42, 58]. Elle ne sont pas satisfaites. Comme il s'agit d'un système avec une seule entrée, affine, ceci implique la non-platitude du système.

[2] Charlet, B., J. Lévine et R. Marino: *On dynamic feedback linearization.* Systems Control Lett., 13 :143–151, 1989.

[3] Charlet, B., J. Lévine et R. Marino: *Sufficient conditions for dynamic state feedback linearization.* SIAM J. Control Optim., 29 :38–57, 1991.

[4] Cohn, P. M.: *Algebra*, tome 2. John Wiley & Sons, Chichester, 2"a édition, 1989.

[5] Cohn, P. M.: *Algebra*, tome 3. John Wiley & Sons, Chichester, 2"a édition, 1991.

[6] d'Andréa-Novel, B. et J. Lévine: *Modelling and nonlinear control of an overhead crane.* Dans Kaashoek, M. A., J. H. van Schuppen et A. C. M. Ran (rédacteurs) : *Robust Control of Linear and Nonlinear Systems, MTNS–89*, tome 2, pages 523–529. Birkhäuser, Boston, 1990.

[7] Delaleau, E.: *Lowering orders of input derivatives in generalized state representations of nonlinear systems.* Dans Fliess, M. (rédacteur) : *Nonlinear Control Systems Design 1992, Selected papers from the 2nd IFAC symposium*, pages 347–351. Pergamon Press, Oxford, 1993.

[8] Delaleau, E. et M. Fliess: *Algorithme de structure, filtrations et découplage.* C. R. Acad. Sci. Paris Sér. I Math., 315 :101–106, 1992.

[9] Delaleau, E. et W. Respondek: *Lowering the orders of derivatives of controls in generalized state space systems.* J. Math. Systems Estim. Control, 5 :1–27, 1995. (Summary : 375–378).

[10] Delaleau, E. et J. Rudolph: *Control of Flat Systems by Quasi-Static Feedback of Generalized States.* Internat. J. Control, 71 :745–765, 1998.

[11] Delaleau, E. et P. S. Pereira da Silva: *Filtrations in feedback synthesis : Part I — Systems and feedbacks.* Forum Math., 10 :147–174, 1998.

[12] Delaleau, E. et P. S. Pereira da Silva: *Filtrations in feedback synthesis : Part II — Input-output decoupling and disturbance decoupling.* Forum Math., 10 :259–275, 1998.

[13] Diop, S.: *Differential-algebraic decision methods and some applications to system theory.* Theor. Computer Sci., 98 :137–161, 1992.

[14] Diop, S. et M. Fliess: *On nonlinear observability.* Dans *Proc. 1st European Control Conference*, pages 152–157, Grenoble, France, 1991.

[15] El Asmi, S. et M. Fliess: *Formules d'inversion.* Dans Bonnard, B., B. Bride, J. P. Gauthier et I. Kupka (rédacteurs) : *Controlled Dynamical Systems*, tome 8 de *Progr. Systems Control Theory*, pages 201–210. Birkhäuser, Boston, 1991.

[16] Fliess, M.: *Automatique et corps différentiels.* Forum Math., 1 :227–238, 1989.

[17] Fliess, M.: *Automatique en temps discret et algèbre aux différences.* Forum Math., 2 :213–232, 1990.

[18] Fliess, M.: *Some basic structural properties of generalized linear systems.* Systems Control Lett., 15 :391–396, 1990.

[19] Fliess, M.: *Reversible linear and nonlinear discrete-time dynamics.* IEEE Trans. Automat. Control, AC–37 :1144–1153, 1992.

[20] Fliess, M.: *Invertibility of causal discrete time dynamical systems.* J. of Pure and Applied Algebra, 86 :173–179, 1993.

[21] Fliess, M. et S. T. Glad: *An algebraic approach to linear and nonlinear control.* Dans Trentelman, H. L. et J. C. Willems (rédacteurs) : *Essays on Control : Perspectives in the Theory and its Applications*, tome 14 de *Progr. Systems Control Theory*, pages 223–267. Birkhäuser, Boston, 1993.

[22] Fliess, M., C. Join et H. Sira-Ramírez: *Non-linear estimation is easy.* 2007.

[23] Fliess, M., J. Lévine, P. Martin, F. Ollivier et P. Rouchon: *Flatness and dynamic feedback linearizability : two approaches.* Dans Isidori, A., S. Bittanti, E. Mosca, A. De Luca, M. D. Di Benedetto et G. Oriolo (rédacteurs) : *Proc. 3rd European Control Conference*, pages 649–654, 1995.

[24] Fliess, M., J. Lévine, P. Martin et P. Rouchon: *Linéarisation par bouclage dynamique et transformations de Lie-Bäcklund.* C. R. Acad. Sci. Paris Sér. I Math., 317 :981–986, 1993.

[25] Fliess, M., J. Lévine, P. Martin et P. Rouchon: *A Lie-Bäcklund approach to equivalence and flatness of nonlinear systems.* IEEE Trans. Automat. Control, AC–44 :922–937, 1999.

[26] Fliess, M., J. Lévine, Ph. Martin et P. Rouchon: *Flatness and defect of non-linear systems : introductory theory and examples.* Internat. J. Control, 61 :1327–1361, 1995.

[27] Fliess, M., J. Lévine et P. Rouchon: *Generalized state variable representation for a simplified crane description.* Internat. J. Control, 58 :277–283, 1993.

[28] Fliess, M. et J. Rudolph: *Local "tracking observers" for flat systems.* Dans *Proc. Symposium on Control, Optimization and Supervision, Computational Engineering in Systems Application, IMACS Multiconference, Lille, July 9–12 1996 (CESA'96)*, pages 213–217, 1996.

[29] Fliess, M. et J. Rudolph: *Corps de Hardy et observateurs asymptotiques locaux pour systèmes différentiellement plats.* C. R. Acad. Sci. Paris Sér. IIb, 324 :513–519, 1997.

[30] Hilbert, D.: *Über den Begriff der Klasse von Differentialgleichungen.* Math. Ann., 73 :95–108, 1912.

[31] Hunt, L. R., R. Su et G. Meyer: *Design for multi-input nonlinear systems.* Dans Brockett, E. W. (rédacteur) : *Differential Geometric Control Theory*, pages 258–298. Birkhäuser, Boston, 1983.

[32] Isidori, A.: *Nonlinear Control Systems.* Springer-Verlag, Berlin, 3"a édition, 1995.

[33] Jakubczyk, B. et W. Respondek: *On linearization of control systems.* Bull. Acad. Polonaise Sci. Sér. Sci. Math., 28 :517–522, 1980.

[34] Johnson, J.: *Differential dimension polynomials and a fundamental theorem on differential modules.* Amer. J. Math., 91 :239–248, 1969.

[35] Johnson, J.: *Kähler differentials and differential algebra.* Ann. of Math., 89 :92–98, 1969.

[36] Kaplansky, I.: *An introduction to differential algebra.* Hermann, Paris, 2"a édition, 1976.

[37] Kokotović, P., H. K. Khalil et J. O'Reilly: *Singular perturbation methods in control : analysis and design.* Academic Press, London, 1986.

[38] Kolchin, E. R.: *Differential Algebra and Algebraic Groups.* Academic Press, New York, 1973.

[39] Martin, P.: *Contribution à l'étude des systèmes différentiellement plats.* Thèse de Doctorat, École Nationale Supérieure des Mines de Paris, 1992.

[40] Martin, P., R. M. Murray et P. Rouchon: *Flat systems.* Dans Bastin, G. et M. Gevers (rédacteurs) : *Plenary Lectures and Mini-Courses, 4th European Control Conference, Brussels, Belgium*, pages 211–264. 1997.

[41] Moog, C. H., J. Perraud, P. Bentz et Q. T. Vo: *Prime differential ideals in nonlinear rational control systems.* Dans *Proc. Symposium on Nonlinear Control Systems Design (NOLCOS'89)*, pages 178–182, Capri, Italy, 1989.

[42] Nijmeijer, H. et A. J. van der Schaft: *Nonlinear Dynamical Control Systems.* Springer-Verlag, New York, 1990.

[43] Pomet, J. B.: *A differential geometric setting for dynamic equivalence and dynamic linearization.* Dans Jakubczyk, B., W. Respondek et T. Rzeżuchowski (rédacteurs) : *Geometry in Nonlinear Control and Differential Inclusions*, tome 32 de *Banach Center Publ.*, pages 319–339. Banach Center, Warszawa, 1995.

[44] Ritt, J. F.: *Differential Algebra.* American Mathematical Society, New York, 1950.

[45] Rothfuß, R.: *Anwendung der flachheitsbasierten Analyse und Regelung nichtlinearer Mehrgrößensysteme.* Fortschritt-Berichte, Reihe 8, Nr. 664. VDI-Verlag, Düsseldorf, 1997.

[46] Rothfuß, R., J. Rudolph et M. Zeitz: *Flatness Based Control of a Nonlinear Chemical Reactor Model.* Automatica J. IFAC, 32 :1433–1439, 1996.

[47] Rouchon, P.: *Flatness and robust control of pendulum mechanical systems.* Technical report 469, Centre Automatique et Systèmes, École des mines de Paris, June 1994.

[48] Rouchon, P.: *Necessary condition and genericity of dynamic feedback linearization.* J. Math. Systems Estim. Control, 5 :345–358, 1995.

[49] Rouchon, P. et J. Rudolph: *Invariant tracking and stabilization.* Dans Aeyels, D., F. Lamnabhi-Lagarrigue et A. van der Schaft (rédacteurs) : *Stability and Stabilization of Nonlinear Systems*, tome 246 de *Lecture Notes in Control and Inform. Sci.*, chapitre 14, pages 261–273. Springer-Verlag, 1999.

[50] Rouchon, P. et J. Rudolph: *Réacteurs chimiques différentiellement plats : planification et suivi de trajectoires.* Dans Corriou, J. P. (rédacteur) : *Commande de procédés chimiques — Réacteurs et colonnes de distillation*, chapitre 5, pages 163–200. Hermès Science Publications, 2001.

[51] Rudolph, J.: *Well-formed dynamics under quasi-static state feedback.* Dans Jakubczyk, B., W. Respondek et T. Rzeżuchowski (rédacteurs) : *Geometry in Nonlinear Control and Differential Inclusions*, tome 32 de *Banach Center Publ.*, pages 349–360. Banach Center, Warszawa, 1995.

[52] Rudolph, J.: *Beiträge zur flachheitsbasierten Folgeregelung linearer und nichtlinearer Systeme endlicher und unendlicher Dimension.* Shaker Verlag, 2003.

[53] Rudolph, J.: *Flatness based control of distributed parameter systems.* Berichte aus der Steuerungs- und Regelungstechnik. Shaker Verlag, Aachen, 2003.

[54] Rudolph, J. et E. Delaleau: *Some examples and remarks on quasi-static feedback of generalized states.* Automatica J. IFAC, 34 :993–999, 1998.

[55] Rudolph, J. et S. El Asmi: *Filtrations and Hilbert polynomials in control theory.* Dans Helmke, U., R. Mennicken et J. Saurer (rédacteurs) : *Systems and Networks : Mathematical Theory and Applications, (MTNS'93, Invited and Contributed Papers)*, tome 2, pages 449–452. Akademie Verlag, 1994.

[56] Seidenberg, A.: *Some basic theorems in differential algebra.* Trans. Amer. Math. Soc., 73 :174–190, 1952.

[57] Sira-Ramírez, H. et S. K. Agrawal: *Differentially Flat Systems.* Marcel Dekker, 2004.

[58] Slotine, J. J. E. et J. W. Li: *Applied Nonlinear Control.* Prentice-Hall, Englewood Cliffs, 1991.

[59] Sommer, R.: *Control design for multivariable non-linear time-varying systems.* Internat. J. Control, 31 :883–891, 1980.

[60] Tikhonov, A., A. Vasil'eva et A. Sveshnikov: *Differential equations.* Springer-Verlag, Berlin, 1980.

8.A Bases mathématiques

1 : Dans l'*algèbre différentielle* on étudie des structures algébriques avec des dérivations [44, 38]. Soit R un anneau *commutatif*, $1 \in R$ et $\mathbb{Q} \subset R$. Une *dérivation* dans (ou de) R est une application $\partial : R \to R$, telle que, $\forall\ a, b \in R$:

(i) $\partial\,(a+b) = \partial\, a + \partial\, b$ (linéarité)

(ii) $\partial\,(ab) = (\partial\, a)\, b + a\, \partial\, b$ (règle de Leibniz).

Un *anneau différentiel ordinaire* R est un anneau muni d'une seule dérivation ∂ tel que $a \in R \ \Rightarrow\ \partial a \in R$.

Notation : On écrit aussi $\frac{d}{dt}$ pour la dérivation d'un anneau différentiel ordinaire R, et $\dot{a}$ pour $\frac{d}{dt}a$, ainsi que $\frac{d}{dt}(\frac{d}{dt}a) = \frac{d}{dt}\dot{a} = \ddot{a}$, et plus généralement $(\frac{d}{dt})^i a = a^{(i)}, i > 0$.

Rem. : On parle d'anneau différentiel ordinaire car une relation $a = 0$ dans R peut être interprétée comme une équation différentielle ordinaire. Avec plusieurs dérivations, commutant entre elles, on traite des é.d.p..

Un *corps différentiel* K est un anneau différentiel qui forme un corps (commutatif). La règle de dérivation de fractions $a/b \in K, a, b \in R, b \neq 0$ se déduit de celles de R : $\partial(b(a/b)) = \partial a$, $(\partial b)\,(a/b) + b\,(\partial(a/b)) = \partial a$, $b^2(\partial(a/b)) = b\partial a - (\partial b)\, a$, d'où $\partial(a/b) = (b\partial a - (\partial b)\, a)/(b^2)$.

Une *constante* dans un corps différentiel K est un élément $c \in K$, tel que $\dot{c} = 0$. Un *corps de constantes* C est un corps différentiel dans lequel $\forall c \in C, \dot{c} = 0$. (Il est aisé de vérifier le propriétés d'un corps.)

Ex. : 1. Les corps de nombres $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ sont des exemples simples de corps de constantes.

2. Le corps $\mathbb{R}(t)$ des fractions rationnelles en t avec des coefficients appartenant (ou dans) $\mathbb{R}$ est un corps différentiel ordinaire (avec la dérivation $\frac{d}{dt}$).

2 : Une *extension de corps* F/E consiste en deux corps E et F commutatifs tels que E est un sous-corps de F (c'est-à-dire en restreignant la multiplication et l'addition de F aux éléments de E ce dernier forme un corps). Le corps F est aussi appelé corps d'extension de E.

Soit X un sous-ensemble de F. Le corps engendré par X sur E est le plus petit (par rapport à l'inclusion (#7)) sous-corps de F contenant à la fois E et X. On le note $E(X)$. L'extension de corps F/E est dite *simple*, si X ne contient qu'un seul élément, elle est dite *finiment engendrée (ou de type fini)* si X contient un nombre fini d'éléments.

Ex. : L'extension de corps $\mathbb{C}/\mathbb{R}$ est finiment engendrée, car $\mathbb{C} = \mathbb{R}(\sqrt{-1})$.

3 : Un *idéal premier* d'un anneau commutatif R est un idéal I dans R pour lequel $a, b \in R$ et $ab \in I$ implique qu'au moins un des éléments a et b appartient à l'idéal I.

L'anneau des résidus $S = R/I$ ne contient pas de diviseur de zéro si, et seulement si, I est un idéal premier. En introduisant les inverses des éléments non nuls (par rapport à la multiplication), c'est-à-dire en localisant en $S\backslash\{0\}$, on obtient le corps des fractions (non commutatif) S.

Soit donné un système d'équations algébriques

$$P_i(x_1, \ldots, x_n) = 0, \quad i = 1, \ldots, q,$$

avec P_i, $i = 1, \ldots, q$, des polynômes en x_i, $i = 1, \ldots, n$, à coefficients dans un corps. Soit $P_i(X_1, \ldots, X_n) \in K[X_1, \ldots, X_n]$, $i = 1, \ldots, q$, avec $K[X_1, \ldots, X_n] = K[X]$ l'anneau polynômial engendré librement par $X = (X_1, \ldots, X_n)$ (c'est-à-dire les X_i ne satisfont aucune équation non triviale, il s'agit d'indéterminées), et soit l'idéal I engendré par la famille $P = (P_1, \ldots, P_q)$ dans $K[X]$ un idéal premier. Alors on obtient le corps des fractions de l'anneau des résidus $K[X]/I$ par la construction décrite. Si les x_j, $j = 1, \ldots, n$ sont les éléments du corps des fractions, issus des classes $X_i + I$, alors ils satisfont aux équations $P_i(x) = 0$, $i = 1, \ldots, q$.

Rem. : Si I n'est pas premier, on peut trouver des idéaux premiers dans $K[X]$ dont l'intersection est I. Pour chacun des ces idéaux contenant I on peut construire un corps. Ceci correspond à une réunion des ensembles des solutions des systèmes d'équations correpondant à ces idéaux, c'est-à-dire aux branches de solutions.

Ex. : Considérons l'équation $P(x_1, x_2) = x_1^2 - x_2^2 = 0$. Avec $P = X_1^2 - X_2^2 \in \mathbb{Q}[X]$ l'idéal I dans $\mathbb{Q}[X]$ engendré par P n'est pas premier, car $P = (X_1 + X_2)(X_1 - X_2)$. En revanche, les deux idéaux I_1 et I_2, engendrés par $P_1 = X_1 + X_2$ et $P_2 = X_1 - X_2$ respectivement, sont premiers. On a $I = I_1 \cap I_2$. La solution de l'équation en considération est donnée par la réunion de deux branches, les solutions respectives de $x_1 + x_2 = 0$ et $x_1 - x_2 = 0$.

4 : Un élément a de F est dit *algébrique sur* E s'il existe un polynôme (non trivial) $P \in E[Z], P \neq 0$ tel que $P(a) = 0$. Sinon a est dit *transcendant sur* E.

Si *tout* élément de F est algébrique sur E l'extension de corps F/E est dite *algébrique*, elle est dite *transcendante* sinon.
Si l'extension de corps $E(a)/E$ est transcendante, $E(a)$ est isomorphe au corps $E(X)$ des fractions rationnelles en une indéterminée X à coefficients dans E.

Ex. : 1. La racine $a = \sqrt{2} \in \mathbb{R}$ est algébrique sur $\mathbb{Q}$, car avec $P(Z) = Z^2 - 2$ on a $P(a) = 0$.
2. On sait que e et π sont transcendants sur $\mathbb{Q}$.

5 : Soient K un corps et E/K une extension de corps algébrique. Alors si tout polynôme $P \in K[x]$ s'écrit

$$P = a_0(x - \alpha_1) \cdots (x - \alpha_n), \qquad a_0 \in k, \alpha_i \in E$$

on appelle E une clôture algébrique de K. Si on peut choisir $E = K$ le corps K est dit algébriquement clos.
Chaque corps admet une clôture algébrique, qui est déterminée à un isomorphisme sur K près. On écrit $\overline{K}$ pour cette clôture algébrique unique (à un isomorphisme près) [4].

6 : Une famille z d'éléments de L est dite *algébriquement indépendante* sur K s'il n'existe aucun polynôme $P(Z) \in K[Z]$, $P \neq 0$, tel que $P(z)=0$. Dans le cas contraire on dit que z est *K-algébriquement dépendant* (ou bien *algébriquement dépendant sur K*).

7 : Un ensemble A est un *sous-ensemble minimal* (par rapport à l'inclusion) de B avec une certaine propriété E, si A possède la propriété E, mais aucun de ses sous-ensemble $A\backslash\{a\}$, $a \in A$ ne la partage. De même, A est un *sous-ensemble maximal* (par rapport à l'inclusion) de B avec une certaine propriété E, si A a la propriété E, mais pour des $b \in B$ arbitraires, la réunion $A \cup \{b\}$ ne la partage pas.

8 : Soit l une loi qui à chaque sous-ensemble fini X d'un ensemble S associe certains éléments de S, que l'on appelle *dépendant de X*. Une famille $x = \{x_i \mid i \in I\}$ d'éléments de S est appelé *indépendante*, si aucun des x_i n'est dépendant de $\{x_j \mid j \neq i; i, j \in I\}$, sinon elle est dite *dépendante*. Supposons en plus que les conditions suivantes soient remplies :

(i) Il suit de $x = \{x_1, \ldots, x_n\}$ que chacun des x_i, $i = 1, \ldots, n$ est dépendant de x.

(ii) Si un élément $z \in S$ est dépendant d'un ensemble Y et chaque élément de Y est dépendant de X, alors z est également dépendant de X.

(iii) Si un élément $z \in S$ est dépendant de $x = \{x_1, \ldots, x_n\}$, main non de $\{x_2, \ldots, x_n\}$, alors x_1 est dépendant de $\{z, x_2, \ldots, x_n\}$.

Dans ce cas l est appelée une *loi de dépendance*. La troisième propriété est appelée *propriété d'échange*, la seconde *transitivité* [5].

9 : Soit encore X un sous-ensemble d'un ensemble S avec une loi de dépendance. On dit que X *engendre* S si tous les éléments de S sont dépendant de X. L'ensemble X est appelé *base* de S si X est indépendant et engendre S.

L'ensemble X est un sous-ensemble de S, qui est maximal (par rapport à l'inclusion (#7))), si, et seulement si, X forme une base de S. En outre X est une base de S si, et seulement si, X est un sous-ensemble minimal de S engendrant S.

Pour une preuve voir [5, Prop. 1.4.1] ou [52].

10 : Soit S un ensemble avec une loi de dépendance admettant une base finie. Alors tout sous-ensemble indépendant fait partie d'une base, et deux bases ont le même cardinal.

Pour une preuve voir [5, Lemma 1.4.2] ou [52].

11 : La dépendance algébrique sur K est une relation de dépendance au sens de #8. On dit $y \in L$ est algébriquement dépendant sur K d'une famille $z = (z_1, \ldots, z_s)$ d'éléments de L si $(y, z_1, \ldots, z_s)$ est K-algébriquement dépendant.

Pour une preuve voir [4, 5] ou [52].

12 : Soit F/E une extension de corps. Une famille z d'éléments de F qui est algébriquement indépendante sur E, et qui, en plus, est maximale (par rapport à l'inclusion (#7)), est appelée une *base de transcendance* de F/E. Le cardinal d'une telle famille z est appelé le *degré de transcendance* de F/E ; on le note deg tr F/E. (L'unicité du degré de transcendance d'une extension de corps de type finie suit de #11 et #10.) L'extension de corps $F/E(z)$ est alors appelée algébrique.

On observe que si F/E est algébrique (#4) deg tr $F/E = 0$, et inversement, deg tr $F/E = 0$ implique que F/E est algébrique.

Rem. : Pour les extensions de corps finies il existe toujours une base de transcendance (finie), que l'on peux prendre dans une famille génératrice.

13 : Si G/F et F/E sont deux extensions de corps il en est de même pour G/E. En outre deg tr $G/E =$ deg tr $G/F +$ deg tr F/E. Si X et Y sont des bases de transcendance de F/E et de G/F respectivement, alors $X \cup Y$ est une base de transcendance de G/E.

Pour une preuve voir [52] ou [5, p. 170].

14 : Une *extension de corps différentielle* L/K consiste en deux corps différentiels K et L (voir #1) tels que K est un sous-corps de L (#2), et en plus la dérivation de K coïncide avec celle de L. Alors en appelle aussi le corps différentiel L un *corps différentiel d'extension* de K.

Si X est un sous-ensemble de L, le corps différentiel engendré (différentiellement) par X sur K est le plus petit (par rapport à l'inclusion (#7)) sous-corps différentiel de L qui contient et K et X. On le note $K\langle X\rangle$.

Une extension de corps différentielle L/K est dite *(différentiellement) finiment engendrée* (ou de type différentiel fini), si $L = K\langle x\rangle$ avec une famille finie $x = (x_1, \ldots, x_n)$ d'éléments de L.

Rem. : Nous supposons partout que les extensions de corps différentielles soient différentiellement finiment engendrées à une clôture algébrique près. En outre, comme déjà ramarqué dans #1, on ne considère que des corps différentiels ordinaires. Les extensions de corps (non différentielles), quant à elles, ne sont pas nécessairement de type fini. Par exemple, toute extension différentielle $K\langle z\rangle/K$ peut être interprétée comme une extension de corps non différentielle. Soit le cardinal de z égal à 1, pour simplifier. Alors les générateurs sont juste toutes les dérivées de z, et si z est différentiellement transcendant sur K (voir #17), alors on ne peut se passer d'aucun des éléments du système de générateurs $Z = \{z, \dot{z}, \ddot{z}, \ldots\}$, c'est-à-dire aucun sous-ensemble de Z ne forme un système de générateurs de $K\langle z\rangle/K$.

15 : Un *polynôme différentiel* en une indéterminée Z (ou en la famille d'indéterminées $Z = (Z_1, \ldots, Z_s)$) à coefficients dans un corps différentiel K, que l'on note $P(Z, \dot{Z}, \ldots, Z^{(\alpha)})$ (avec $\alpha \in \mathbb{N}$), est un élément de l'anneau différentiel $K\{Z\}$ (voir #1). On obtient ce dernier, par exemple, en équipant l'anneau des polynômes $K[Z]$ d'une dérivation dont la restriction à K coïncide avec la dérivation de K. Ainsi $P \in K\{Z\}$ est un polynôme à coefficients dans K, en les éléments de Z et un nombre fini de ces dérivées.

16 : Un *idéal différentiel* I d'un anneau différentiel R est un idéal dans R qui est clos sous la dérivation de R : $a \in I \Rightarrow \dot{a} \in I$. Si I est un idéal premier, alors on obtient, de manière analogue à la construction dans #3, un corps de fractions de l'anneau (différentiel) R/I, qui est un corps différentiel.

Ainsi la construction d'un corps différentiel peut se faire à partir d'un système d'é.d.o. algébriques menant à un idéal premier, comme dans le cas non différentiel dans #3.

17 : Un élément $a \in L$ est dit *différentiellement algébrique sur* K s'il existe un polynôme différentiel (non trivial) $P(Z, \dot{Z}, \ldots, Z^{(\alpha)})$ à coefficients dans K (c'est-à-dire un polynôme $P \in K\{Z\}, P \neq 0$), tel que $P(a, \dot{a}, \ldots, a^{(\alpha)}) = 0$. Dans le cas contraire on dit que a est *différentiellement transcendant sur* K.

Soit r le plus petit entier tel qu'il existe un polynôme P dont l'ordre maximal de dérivées apparaissant effectivement est égal à r (c'est-à-dire $\partial P/\partial Z^{(i)} \neq 0, i = r$ et $\partial P/\partial Z^{(i)} = 0, i > r$). Alors le degré de transcen-

dance non différentiel de l'extension de corps (simple) $K\langle a\rangle/K$ est égal à r, car la famille $\overline{a} = (a, \dot{a}, \ldots, a^{(r-1)})$ en forme une base de transcendance (non différentielle) de $K\langle a\rangle/K$.

Démonstration : Visiblement $\overline{a}$ est algébriquement indépendant sur K, mais $a^{(r)}$ est $K(\overline{a})$-algébrique. Par dérivation on obtient $\frac{d}{dt}P = 0$, et $a^{(r+1)}$ est donc $K(\overline{a}, a^{(r)})$-algébrique, et $K(\overline{a})$-algébrique. On peut procéder de la même manière pour toutes les dérivées supérieures. □

Si *tout* élément de L est différentiellement algébrique sur K, l'extension de corps différentielle L/K est dite *différentiellement algébrique* ; elle est dite *différentiellement transcendante* sinon. Si *tout* élément de L qui n'appartient pas à K est différentiellement transcendant sur K l'extension de corps différentielle L/K est dite *différentiellement purement transcendante*.

Ex. : On voit que $a = \sin t \in \mathbb{Q}\langle \sin t\rangle$ est différentiellement algébrique sur $\mathbb{Q}$, car avec $P(Z, \dot{Z}) = Z^2 + \dot{Z}^2 - 1$ on a $P(a, \dot{a}) = 0$.

18 : Une famille z d'éléments de L est dite *différentiellement algébriquement indépendante* sur K s'il n'existe aucun polynôme différentiel $P(Z, \dot{Z}, \ddot{Z}, \ldots, Z^{(\gamma)}) \in K\{Z\}$, $P \neq 0$ tel que $P(z, \dot{z}, \ddot{z}, \ldots, z^{(\gamma)}) = 0$. Dans le cas contraire on dit que z est *différentiellement K-algébriquement dépendant* (ou *différentiellement algébriquement dépendant sur K*).

19 : La dépendance algébrique différentielle sur K est une relation de dépendance au sens de #8. On dit pour ceci que $y \in L$ est différentiellement algébriquement dépendant sur K d'une famille $z = (z_1, \ldots, z_s)$ d'éléments de L si $(y, z_1, \ldots, z_s)$ est différentiellement K-algébriquement dépendant. (Comparer au cas non différentiel dans #11.)

Pour une preuve voir [52].

20 : Soit L/K une extension de corps différentielle. Une famille z d'éléments de L qui est différentiellement algébriquement indépendante sur K et maximale (par rapport à l'inclusion) est appelée *base de transcendance différentielle* de L/K. Pour les extensions de corps différentielles de type fini il existe toujours une base de transcendance différentielle (finie). (Ceci est évident, car on peut la choisir dans une famille génératrice.)

Le cardinal d'une base de transcendance différentielle de L/K est appelé le *degré de transcendance différentiel* de L/K ; on le note deg tr diff L/K. (L'unicité du degré de transcendance différentiel d'une extension de corps différentielle (de type fini) suit de #11 et #10.) Si z est une base de transcendance différentielle d'une extension de corps différentielle L/K, alors l'extension de corps différentielle $L/K\langle z\rangle$ est différentiellement algébrique.

21 : Une extension de corps différentielle L/K est différentiellement algébrique (#17) si, et seulement si, deg tr diff $L/K = 0$. Si L/K est différentiellement purement transcendant alors une famille génératrice minimale (par rapport à l'inclusion) est aussi une base de transcendance différentielle, et réciproquement, dans ce cas, une base de transcendance différentielle est une famille génératrice. Comme une base de transcendance différentielle est différentiellement K-algébriquement indépendante, elle est aussi minimale. Ainsi une extension de corps différentielle L/K qui est différentiellement purement transcendante est caractérisée par le fait qu'il existe une famille $z = (z_1, \ldots, z_m)$ dans L telle que $L = K\langle z\rangle$ et deg tr diff $L/K = m$.

22 : Il existe une base de transcendance (non différentielle) *finie* d'une extension de corps différentielle L/K (c'est-à-dire une base de transcendance finie pour l'extension non différentielle des corps avec les éléments de L, respectivement de K) si, et seulement si, L/K est différentiellement algébrique, c'est-à-dire deg tr diff $L/K = 0$.

Preuve : Soit $z = (z_1, \ldots, z_m)$ une famille génératrice de L/K, et soit deg tr diff $L/K = 0$. Le degré de transcendance de l'extension de corps (simple) $K\langle z_1\rangle/K$ est fini (#17). Il en est de même pour chacune des extensions de corps (simples) $K\langle z_1, \ldots, z_j\rangle/K\langle z_1, \ldots, z_{j-1}\rangle$, $j = 2, \ldots, m$. Avec #13 on sait que deg tr L/K est également fini. Comme dans #17 il est aussi clair que, pour toutes les extensions de corps (simples) $K\langle z_1\rangle/K$ et $K\langle z_1, \ldots, z_j\rangle/K\langle z_1, \ldots, z_{j-1}\rangle$, $j = 2, \ldots, m$, il existe des bases de transcendance de la forme $\overline{z_i} = (z_i, \dot{z}_i, \ldots, z_i^{(r_i)})$.

Soit, pour la réciproque, $\zeta \in L$ différentiellement transcendant sur K. Alors il n'y a pas de dépendance K-algébrique entre le nombre infini d'éléments $\zeta^{(j)}, j \geq 0$ de L, et ainsi le degré de transcendance (non différentiel) ne peut être fini.

Rem. : Si l'extension de corps différentielle n'est finiment engendrée qu'a une clôture algébrique près, on obtient le même résultat. Il suffit de prendre la clôture algébrique à la fin de la démonstration, ceci ne change pas le degré de transcendance (non différentiel).

23 : Soit k un corps différentiel, et soi $\frac{d}{dt}$ la dérivation de k. Dans l'anneau $k[\frac{d}{dt}]$ des polynômes en $\frac{d}{dt}$ à coefficients dans k nous définissons la multiplication par

$$\forall a \in k : \frac{d}{dt}\, a := \dot{a} + a\, \frac{d}{dt}.$$

Les éléments de $k[\frac{d}{dt}]$ sont alors de la forme $\sum_{i=1}^{n} a_i \left(\frac{d}{dt}\right)^i$, avec $a_i \in k, i = 1, \ldots, n$, que l'on peut interpréter (formalement) comme des opérateurs différentiels.

Visiblement, l'anneau $k[\frac{d}{dt}]$ est alors commutatif si, et seulement si, k est un corps de constantes.

24 : Soit L/K une extension de corps différentielle, finiment engendrée par $z = (z_1, \ldots, z_m)$, c'est-à-dire $L = K\langle z\rangle$, et soit $\Omega_{L/K}$ un $L[\frac{d}{dt}]$-module (gauche). Supposons que l'application $d_{L/K} : L \to \Omega_{L/K}$ satisfasse, pour tout $x, y \in L$ et $a \in K$,

$$\begin{aligned}\frac{d}{dt}(d_{L/K}(x)) &= d_{L/K}(\dot{x})\\ d_{L/K}(x+y) &= d_{L/K}(x) + d_{L/K}(y)\\ d_{L/K}(xy) &= y\, d_{L/K}(x) + x\, d_{L/K}(y)\\ d_{L/K}(a) &= 0.\end{aligned}$$

Il s'agit donc d'une K-dérivation de L dans $\Omega_{L/K}$. (Les images des éléments de K sont nuls.) On écrit $d_{L/K}\, x$ pour $d_{L/K}(x)$.

Le module $\Omega_{L/K}$ est le $L[\frac{d}{dt}]$-module (gauche) engendré par l'image suivante $d_{L/K}\, z = (d_{L/K}\, z_1, \ldots, d_{L/K}\, z_m)$ de la famille des générateurs z de L/K. Le module $\Omega_{L/K}$ est appelé le *module des différentielles de Kähler* de L/K, ses éléments sont appelés *différentielles de Kähler*.

Rem. : 1. Si L/K est clair par le contexte, on écrit aussi $d\, z$ au lieu de $d_{L/K}\, z$, pour simplifier.

2. La construction de $\Omega_{L/K}$ est universelle : Si M est un $L[\frac{d}{dt}]$-module quelconque et D une K-dérivation de L dans M, il existe un homomorphisme de $L[\frac{d}{dt}]$-modules, et un seul, $\varphi : \Omega_{L/K} \to M$, pour lequel $\varphi \circ d_{L/K} = D$ (voir [35]).

3. À partir d'une extension de corps (non différentielle) F/E on obtient (comme cas spécial avec $L = F$, $K = E$ et $\frac{d}{dt}$ la dérivation triviale) un F-espace vectoriel de différentielles de Kähler.

25 : Un élément $w \in \overline{L} = \overline{K\langle z\rangle}$ satisfait une équation de la forme

$$P(w, z, \dot{z}, \ldots, z^{(\alpha)}) = 0,$$

avec $P \in K\{Z\}[W]$, un polynôme en W et Z, et un nombre fini de dérivées de Z, à coefficients dans K. L'application $d_{\overline{L}/K}$ envoie P sur $d_{\overline{L}/K}\, P = d\, P$, avec

$$d\, P = \tfrac{\partial P}{\partial W}\, d\, w + \sum_{i=1}^{m}\sum_{j=0}^{\alpha} \tfrac{\partial P}{\partial Z_i^{(j)}}\, d\, z_i^{(j)} = \tfrac{\partial P}{\partial W}\, d\, w + \sum_{i=1}^{m} Q_i\left(\frac{d}{dt}\right)\, d\, z_i$$

et $Q_i(\frac{d}{dt}) = \sum_{j=0}^{\alpha} \frac{\partial P}{\partial Z_i^{(j)}} \frac{d}{dt}^j$, $i = 1, \ldots, m$. L'équation $d\, P = 0$ définit donc une relation de dépendance $\overline{L}[\frac{d}{dt}]$-linéaire pour $(d\, w, d\, z_1, \ldots, d\, z_m)$. Il suit de $w \notin L = K\langle z\rangle$ que $\frac{\partial P}{\partial W} \neq 0$. Or, on peur diviser dans cette relation de

dépendance par ce coefficient, un élément du corps $\overline{L}$. Par conséquent $d\,w$ est un élément dans le $\overline{L}[\frac{d}{dt}]$-module engendré par $d\,z$. Si $L = \overline{K\langle z\rangle}$, le module des différentielles de Kähler est le $\overline{L}[\frac{d}{dt}]$-module $\Omega_{\overline{L}/K}$, engendré par la famille des différentielles de Kähler $d_{\overline{L}/K}\,z = (d_{\overline{L}/K}\,z_1, \ldots, d_{\overline{L}/K}\,z_m)$, où $\Omega_{\overline{L}/K} \cong \overline{L}[\frac{d}{dt}] \otimes_{L[\frac{d}{dt}]} \Omega_{L/K}$.

\# 26 : Soit L/K une extension de corps différentielle, et soit y une famille d'éléments de L. Alors y est différentiellement algébriquement dépendant sur K si, et seulement si, $d_{L/K}\,y$ est une famille $L[\frac{d}{dt}]$-linéairement indépendante dans $\Omega_{L/K}$. De même, si y est (non différentiellement) algébriquement indépendant sur K, alors $d_{L/K}\,y$ est une famille L-linéairement indépendante dans $\Omega_{L/K}$ (voir [35]). On en déduit

$$\text{deg tr diff } L/K = \text{Rang}_{L[\frac{d}{dt}]}\,\Omega_{L/K}.$$

Pour une preuve voir [35] ou [52].

\# 27 : Si M/L et L/K sont deux extensions de corps différentielles, alors il en est de même de M/K, et l'on a deg tr diff M/K = deg tr diff M/L + deg tr diff L/K. Si X et Y sont des bases de transcendance différentielles de M/L, respectivement de L/K, alors $X \cup Y$ est une base de transcendance différentielle de M/K.

La preuve suit les lignes de celle de #13.

\# 28 : Une *filtration* (différentielle) d'une extension de corps différentielle L/K est une suite non décroissante $\mathcal{L} := (\mathcal{L}_r)_{r\in\mathbb{Z}}$ de corps (non différentiels) $\mathcal{L}_r$, qui sont algébriquement clos, telle que $K \subseteq \mathcal{L}_r \subseteq \overline{L}$ et $\mathcal{L}_r \subseteq \mathcal{L}_{r+1}$ pour tout $r \in \mathbb{Z}$. Une filtration $\mathcal{L}$ de L/K est appelée

(i) *exhaustive* (dans $\overline{L}$) si $\cup_{r\in\mathbb{Z}}\mathcal{L}_r = \overline{L}$;

(ii) *discrète* si $\mathcal{L}_r = K$ pour $r \in \mathbb{Z}$ suffisamment petit ;

(iii) *finie* si, à une clôture algébrique près, tous les $\mathcal{L}_r$ sont finiment engendré sur K ;

(iv) *bonne* si, avec un $r' \in \mathbb{Z}$ approprié, $\mathcal{L}_{s+1} = \overline{\mathcal{L}_s(\frac{d}{dt}\mathcal{L}_s)}$ pour $s > r'$ (où pour un sous-ensemble Z d'un corps différentiel K on a défini l'ensemble $\frac{d}{dt}A = \{x \in K \mid \exists z \in Z,\ x = \frac{d}{dt}z\}$), c'est-à-dire si pour des s assez larges les corps successifs $\mathcal{L}_s$ de la filtration sont engendrés par dérivation, et clôture algébrique.

Enfin on dit que la filtration est *excellente* si elle est et finie et bonne [34, 55]*.

*Dans [34] les corps $\mathcal{L}_r$ ne sont pas pris algébriquement clos. En plus on y suppose $\forall a \in \mathcal{L}_r$, $\dot{a} \in \mathcal{L}_r$ pour toutes les filtrations $\mathcal{L}$. Ainsi la définition des filtrations bonnes n'est pas exactement la même.

Deux filtrations $\mathcal{L}$ et $\tilde{\mathcal{L}}$ ont une *différence bornée* (ou *finie*) s'il existe un $r_0 \in \mathbb{Z}$ tel que $\mathcal{L}_r \subseteq \tilde{\mathcal{L}}_{r+r_0}$ et $\tilde{\mathcal{L}}_r \subseteq \mathcal{L}_{r+r_0}$, pour tout $r \in \mathbb{Z}$. S'il existe un tel r_0 on l'appelle la *différence* de deux filtrations.

La différence finie définit une relation d'équivalence sur l'ensemble des filtrations de L/K. Deux filtrations de L/K qui sont discrètes, excellentes et exhaustives ont une différence finie.

Preuve [10] : Soient $\mathcal{L}$ et $\tilde{\mathcal{L}}$ deux filtrations de L/K qui sont discrètes et excellentes, ainsi que exhaustives (dans $\overline{L}$). Du fait que les filtrations sont discrètes il découle que $\mathcal{L}_r = \tilde{\mathcal{L}}_r = K$ pour r suffisamment petit. Comme elles sont finies et exhaustives (dans $\overline{L}$) on a $\mathcal{L}_r \subseteq \tilde{\mathcal{L}}_{s_r+r}$ et $\tilde{\mathcal{L}}_r \subseteq \mathcal{L}_{\tilde{s}_r+r}$ pour tout $r \in \mathbb{Z}$ et pour des $s_r, \tilde{s}_r \in \mathbb{Z}$ assez large, dépendants de r. Pour des r suffisamment larges, $\mathcal{L}_{r+1} = \overline{\mathcal{L}_r(\frac{d}{dt}\mathcal{L}_r)} \subseteq \overline{\tilde{\mathcal{L}}_{r+s_r}(\frac{d}{dt}\tilde{\mathcal{L}}_{r+s_r})} = \tilde{\mathcal{L}}_{r+s_r+1}$ et $\tilde{\mathcal{L}}_{r+1} = \overline{\tilde{\mathcal{L}}_r(\frac{d}{dt}\tilde{\mathcal{L}}_r)} \subseteq \overline{\mathcal{L}_{r+\tilde{s}_r}(\frac{d}{dt}\mathcal{L}_{r+\tilde{s}_r})} = \mathcal{L}_{r+\tilde{s}_r+1}$, et en plus les s_r et $\tilde{s}_r$ ne dépendent plus de r. La différence finie s'ensuit par le choix de $r_0 = \max(s_r, \tilde{s}_r)$ pour des r larges. □

Institut Supérieur des Systèmes
Industriels de Gabès

Institut Supérieur des Études
Technologiques de Djerba

Association Tunisienne
d'Automatique et de Numérisation

Il est bien admis et depuis longtemps que les mathématiques constituent un passage obligatoire pour tout développement en matière de recherche et développement. Plusieurs outils sont actuellement disponibles sous forme de programmes permettant à l'ingénieur et au chercheur de résoudre le problème qui se pose à eux sans s'attarder sur l'écriture des procédures mathématiques. Malheureusement ceci ne se passe pas toujours sans incidents. En effet, une certaine ignorance des théories mathématiques a été fréquemment à l'origine de mésaventures des utilisateurs potentiels de ces boites à outils mathématiques. C'est justement pour combler ce déficit de connaissances des théories mathématiques les plus exploitées aussi bien en recherche qu'en ingénierie que cet ouvrage est proposé à un prix symbolique aux étudiants, ingénieurs, enseignants et chercheurs.

Editeurs : Ridha Ben Abdennour
Kamel Abderrahim
Hugues Mounier

Institut National de Recherche en
Informatique et en Automatique