UNIVERSITE ABOUBAKR BELKAID

TLEMCEN

FACULTE DES SCIENCES DE L'INGENIEUR

DEPARTEMENT D'ELECTRONIQUE

MEMOIRE

présenté pour l'obtention du diplôme de

MAGISTER

Option : Signaux et Systèmes

par

M^elle^ BOUZEGAOUI Naïma

Thème

SYNTHESE D'OBSERVATEURS NON LINEAIRES PAR INTERPOLATION ET DERIVATION NUMERIQUE

Soutenu publiquement en Octobre 1999 devant le jury :

Président :	Mr N. GHOUALI	Professeur, Université de Tlemcen.
Examinateur :	Mr F. ABI AYAD	Maître Assistant, Université de Tlemcen.
Examinateur :	Mr M. F. KHELFI	Chargé de Cours, Université d'Oran.
Rapporteur :	Mr B. CHERKI	Maître de Conférence, Université deTlemcen.

Remerciements

Au début de ce mémoire, je tiens à remercier Monsieur B. CHERKI pour m'avoir encadrée. Sa disponibilité, son soutien et son aide permanente m'ont beaucoup servi dans la réalisation de ce travail.

Je tiens à exprimer toute ma reconnaissance à Monsieur N. GHOUALI qui m'a fait l'honneur d'accepter la présidence du jury de ce projet.

Mes sincères remerciements vont également à Messieurs F. ABI AYAD et M.F. KHELFI d'avoir accepté de faire partie du jury.

Dédicaces

Je dédie ce modeste travail à ma chère mère , à la mémoire de mon père , à mes frères et à ma sœur ,

aux enseignants de l'institut d'électronique ,

à mes collègues ,

et à tout le personnel de l'institut.

Table des matières

Chapitre 1

Introduction générale

Dans des applications diverses, les processus physiques sont souvent représentés par des modèles mathématiques non linéaires. Pendant une longue période, des méthodes développées pour les systèmes linéaires ont été appliquées par approximation à ces processus. Par souci de précision et d'efficacité, il était intéressant de développer des techniques propres aux systèmes non linéaires.

Généralement, les lois utilisées pour commander une telle classe de systèmes nécessite la connaissance d'informations sur le système qui, quelques fois, ne sont pas disponibles à la mesure ou sont très affectées par le bruit. Ceci exige alors deux solutions : synthétiser des commandes ne dépendant que des mesures, ou alors reconstituer les informations manquantes en utilisant les mesures données par le capteur. L'algorithme de reconstruction en temps réel porte le nom d'*observateur*. La qualité du résultat dépend notamment de la propriété d'*observabilité* du système (i.e., la possibilité de déduire de façon unique l'état initial à partir des observations).

Pour les systèmes linéaires le problème de la synthèse d'un observateur est considéré résolu depuis les années 60 par l'observateur de Luenberger. Il faut, toutefois, noter que l'observabilité d'un système linéaire est une condition nécessaire et suffisante à l'existence d'un observateur.

Quant aux systèmes non linéaires ce problème est plus complexe, puisque, l'obser-

vabilité du système n'est qu'une condition nécessaire pour l'existence d'un observateur. Aussi, elle est fortement liée aux entrées qui leur sont appliquées. Enfin, parceque le principe de séparation n'est plus valable dans ce cas.

La synthèse d'observateurs non linéaires fait encore l'objet de recherches intenses. Parmi les solutions on peut citer les observateurs à grands gains [10] [16], par modes glissants [4] [9], à dynamique d'erreur linéaire [6] [12], par minimisation d'un critère [17],...etc..

Dans les dernières années, les calculateurs numériques sont de plus en plus impliqués dans la commande des processus industriels. Ceci exige donc l'implémentation d'algorithmes discrets. Par conséquent, les observateurs cités ci-dessus ne peuvent être utilisés à moins que s'ils soient discrétisés. Malheureusement, le processus de discrétisation n'est pas toujours trivial. Ceci est dû aux problèmes numériques qui peuvent en résulter ou encore à la complexité du modèle, surtout, s'il est de dimension importante.

L'approche proposée dans ce mémoire suppose, dés le départ, que les données disponibles sont discrètes. Elle part d'une idée apparue dans [8]. Son principe est le suivant : on suppose que les différents états d'un système peuvent être exprimés par la sortie et un certain nombre de ses dérivées. A cet effet, on fera subir à la sortie du système considéré un échantillonnage et un suréchantillonnage et des mesures sont prélevées aux instants de suréchantillonnage. Ensuite, en utilisant une méthode d'approximation appropriée, elles sont interpolées pour donner une fonction approchant au mieux la sortie. Ainsi, pour déterminer les différents états, les dérivées considérées seront celles de la fonction approchante. Il est certain que la qualité de l'approximation dépend de la méthode adoptée.

Notre mémoire comporte 5 chapitres :

Le premier rappelle les notions fondamentales d'observabilité et d'observateurs des systèmes non linéaires. Un rapide aperçu des observateurs linéaires sera également inclus.

Le deuxième chapitre recouvre quelques résultats de la théorie de l'approximation. Tout d'abord, on définira ce qu'est un problème d'approximation. Ensuite, on présentera les conditions nécessaires et suffisantes pour l'existence et l'unicité du meilleur approxi-

mant, l'unique à garantir, relativement à une norme donnée, une faible distance entre la fonction et l'espace d'approximation. A titre d'exemple, on abordera un type d'approximation très répandu qu'est l'interpolation polynômiale de Lagrange, et on s'intéressera aux expressions des erreurs d'approximation commises par cette méthode.

Dans le troisième chapitre, on montrera, à travers un exemple, qu'à cause du phénomène de Runge [7] l'interpolation polynômiale fournira de mauvais résultats. On recourira, alors, à un autre type de fonctions d'approximation dites *Splines*. On s'intéressera, tout au long du chapitre, à leurs propriétés qui nous semblent si importantes qu'on les utilisera dans notre processus d'approximation.

Dans le quatrième chapitre, on présentera, en détail, le principe de l'*observateur par interpolation et dérivation numérique*. Les algorithmes d'observation et de commande adoptés seront mis en œuvre à travers deux exemples. On abordera, par la suite, le problème du bruit et on montrera que le bouclage d'un système utilisant notre observateur pourra jouer, avec efficacité, le rôle d'un filtre. Enfin, on discutera l'effet de l'utilisation des splines ainsi que deux autres méthodes sur la *dérivation numérique*.

Dans le dernier chapitre, on utilisera notre observateur pour commander deux processus industriels : le robot à articulation flexible à un axe et le robot rigide à deux axes. Dans un premier point, des modèles d'état non linéaires leur sont établis. Ensuite, on étudiera leur observabilité. Enfin, les résultats des différents simulations seront présentés et discutés.

Chapitre 2

Observabilité et observateurs

2.1 Introduction

La connaissance de l'état d'un système est d'une grande importance, que ce soit pour la synthèse d'une loi de commande ou pour l'élaboration d'une stratégie de diagnostic ou de détection de défaillance. En général, pour des raisons de réalisabilité technique, il n'est pas possible de mesurer toutes les grandeurs du système. Parfois, on préfère utiliser un nombre réduit de capteurs afin de minimiser le coût. Aussi les grandeurs peuvent ne pas avoir de signification physique et par conséquent on ne peut parler de mesure. Ceci entraîne que l'état x(t) ne peut pas être déduit de la sortie y(t) à l'instant courant 't'. Il est alors nécessaire de penser à un moyen permettant la détermination de x(t) en se basant sur la connaissance des entrées et sorties sur un intervalle de temps passé.

L'observateur est un système dynamique [10] qui, sous certaines conditions, fournit une "estimation" de la valeur courante de l'état en fonction des entrées et sorties passées. Le schéma d'un tel observateur est illustré sur la figure (2.1.).

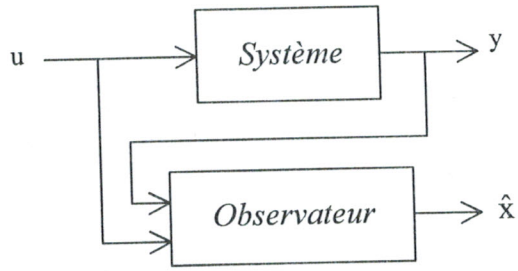


Figure 2.1. Observateur.

La construction d'observateurs pour les systèmes linéaires est un problème résolu depuis les années 60 par Luenberger. Dans ce cas l'existence d'un observateur ne dépend que de l'observabilité du système. Quant aux systèmes non linéaires, la synthèse d'un observateur reste un domaine de recherche très ouvert. Leur observabilité, contrairement aux systèmes linéaires, dépend de l'entrée appliquée. Depuis quelques temps, quelques recherches ont permis de mettre au point des approches résolvant le problème de synthèse d'observateurs pour certaines classes de systèmes non linéaires. On cite:

- Observateurs à grands gains [15]

  Ce type d'observateurs repose sur le principe suivant: le système est décomposé en une partie linéaire ou affine en l'état et une partie non linéaire. A l'aide de gains élevés on augmente l'influence des dynamiques linéaires par rapport aux dynamiques non linéaires. Enfin, l'observateur sera réalisé sur la base de la partie linéaire. L'inconvénient de ce type d'observateurs est sa sensibilité aux bruits à cause des gains élevés.

- Observateurs à modes glissants [15]

  La technique de ces observateurs consiste à faire tendre, à l'aide de fonctions discontinues, les dynamiques de l'erreur d'estimation vers une surface de glissement correspondant à une erreur d'estimation nulle. L'avantage que présentent ces observateurs est leur grande robustesse face aux bruits de mesure et aux erreurs de modèle.

- Observateurs à dynamique d'erreur linéaire [14]

  Ils sont conçus pour des systèmes non linéaires écrits, à l'aide d'une transformation d'état, sous la forme d'une partie linéaire plus une injection de sortie. Dans ce cas la dynamique de l'erreur d'estimation est linéarisable et la synthèse de l'observateur se fait selon une théorie de synthèse linéaire.

- Observateurs par linéarisation approximative [5]

  Ils sont synthétisés sur la base du linéarisé tangent du système non linéaire.

- Observateurs basés sur les techniques de Lyapunov [5]

  Ces observateurs restent d'ordre théorique puisqu'on n'arrive pas jusqu'à ce jour à trouver de méthodes pour la détermination d'une fonction de Lyapunov.

- Observateurs basés sur les méthodes de minimisation [14]

  Ils sont obtenus par minimisation d'une fonctionnelle représentant l'écart entre la sortie mesurée et la sortie calculée. Ce type d'observateurs est généralement de dimension infinie.

- Observateurs de Newton [14]

  Ces observateurs sont conçus essentiellement pour les systèmes non linéaires discrets. Leur principe est basé sur une méthode itérative nécessitant, à chaque étape, une évaluation explicite du Hessien. En pratique cette méthode est coûteuse de point de vue calcul. C'est pour cette raison qu'on l'utilise sous une forme modifiée (la méthode de Broyden par exemple) dans laquelle le Hessien est approché par des méthodes de sécantes.

Ce chapitre comprend une première partie consacrée à des rappels sur la notion d'observabilité, les observateurs et le principe de séparation dans le cas des systèmes linéaires. Celui des systèmes non linéaires sera traité dans la deuxième partie. Dans un troisième

point, on présentera l'effet des entrées sur l'observabilité dans le cas non linéaire, on parlera plus particulièrement des entrées universelles. La quatrième partie traitera du cas des systèmes non linéaires dotés d'une propriété forte d'observabilité: c'est l'observabilité uniforme, ou observabilité pour toute entrée. On abordera, ensuite, la mise en équation des systèmes non linéaires sous forme canonique d'observabilité. Enfin, on présentera le principe de synthèse d'un observateur non linéaire : l'observateur à grand gain.

2.2 Observabilité et observateurs des systèmes linéaires

Considérons le système linéaire invariant défini par :

$$\begin{aligned} \dot{x}(t) &= A.x(t) + B.u(t) \\ y(t) &= C.x(t) \end{aligned} \tag{2.1}$$

avec $x \in \Re^n$, $u \in \Re^m$, $y \in \Re^p$.

2.2.1 Définitions

Définition 1 *Le système (2.1) est observable si pour tout $t_0 \geq 0$ et $T \geq t_0$, la connaissance de $y(t_0, T)$ et $u(t_0, T)$ permette de déterminer de manière unique l'état $x(t_0) = x_0$ quelle que soit l'entrée appliquée.*.

$y(t_0, T)$ et $u(t_0, T)$ désignent respectivement les valeurs de y et u sur l'intervalle $[t_0, T]$. [5]

La notion d'observabilité caractérise la possibilité de reconstruire, exactement ou asymptotiquement, le vecteur d'état connaissant les entrées et les sorties mesurées.

L'observabilité d'un système linéaire peut être vérifiée en testant le rang de la matrice suivante:

$$\Theta = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

Théorème 2 *Le système (2.1) est observable si et seulement si le rang de la matrice* Θ *est* n. *On dit alors que la paire* (A, C) *est observable. [13]*

Si la paire (A, C) est observable, il est possible de construire un observateur [15] pour le système (2.1). Sa forme est celle proposée par Luenberger en 1966. Elle est donnée par:

$$\dot{\hat{x}} = A\hat{x} + K(y - C\hat{x}) + Bu$$

Définissant l'erreur par $e = x - \hat{x}$, ceci donne $\dot{e} = (A - KC)e$ qui peut être asymptotiquement stable si le vecteur K est choisi tel que $VP(A - KC) \in C^-$. Par conséquent, $\hat{x}(t) \rightarrow x(t)$ quand $t \rightarrow \infty$.

2.2.2 Observateurs en boucle fermée – Principe de séparation

Pour les systèmes linéaires, la synthèse d'une loi de commande s'effectue indépendamment de celle d'un observateur. Cela est dû au principe de séparation [15]. En effet, supposons une commande par retour linéaire $u = Fx$. Si le vecteur d'état n'est pas accessible en entier on utilise un observateur et la commande considérée sera $u = F\hat{x}$. Dans ce cas, on obtient:

$$\begin{aligned} \dot{x} &= Ax + BF\hat{x} \\ \dot{\hat{x}} &= KCx + (A - KC + BF)\hat{x} \end{aligned}$$

ce qui équivaut à

$$\begin{pmatrix} \dot{x} \\ \dot{\hat{x}} \end{pmatrix} = \begin{pmatrix} A & BF \\ KC & A - KC + BF \end{pmatrix} \begin{pmatrix} x \\ \hat{x} \end{pmatrix}$$

On a $\dot{x} - \dot{\hat{x}} = A(x - \hat{x}) - KC(x - \hat{x}) = (A - KC)(x - \hat{x})$. Ceci donne

$$\begin{pmatrix} \dot{x} \\ \dot{x} - \dot{\hat{x}} \end{pmatrix} = \underbrace{\begin{pmatrix} A + BF & -BF \\ 0 & A - KC \end{pmatrix}}_{D} \begin{pmatrix} x \\ x - \hat{x} \end{pmatrix}$$

Les valeurs propres de la matrice D sont constituées de $VP(A+BF) \cup VP(A-KC)$. Donc la stabilité du système s'obtient si l'observateur et le système bouclé sur l'état réel sont stables à la fois : c'est le principe de séparation.

Cette propriété ne s'applique pas aux systèmes non linéaires. On peut constater cela à travers l'exemple suivant (exemple de Kokotovic) [5] :

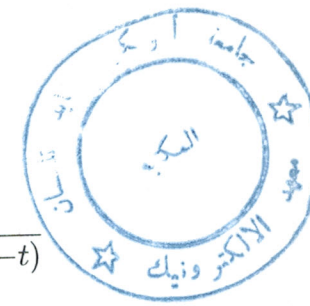
soit le système

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_2 x_1^2 + u \\ \dot{x}_2 &= -x_2 + x_1^2 \\ y &= x_1 \end{aligned}$$

Supposons tout d'abord que les deux états x_1 et x_2 sont disponibles. Dans ce cas $u = -x_2 x_1^2$ rendrait stable le point d'équilibre $x_1 = x_2 = 0$. On suppose maintenant que x_2 n'est pas mesurable et on définit un observateur réduit à convergence exponentielle $\dot{\hat{x}}_2 = -\hat{x}_2 + x_1^2$. Dans ce cas la loi de commande devient $u = -\hat{x}_2 x_1^2$, et l'équation de $\dot{x}_1$ s'écrit : $\dot{x}_1 = -x_1 + x_1^2(x_2 - \hat{x}_2)$. En notant l'erreur d'estimation $e_2 = x_2 - \hat{x}_2$, on a $\dot{e}_2(t) = -e_2(t)$. La dynamique de l'erreur est décrite donc par la fonction $e_2(t) = e_2(0)\exp(-t)$,

et par conséquent $x_1(t)$ s'écrit

$$x_1(t) = \frac{2x_1(0)}{(2 - x_1(0)e_2(0))\exp(t) + x_1(0)e_2(0)\exp(-t)}$$



Pour $x_1(0)e_2(0) > 2$, ce système est instable. Donc un observateur stable et convergent et une loi de commande stabilisante n'impliquent pas nécessairement la stabilité du système bouclé.

2.3 Observabilité des systèmes non linéaires

Contrairement aux systèmes linéaires, le problème d'observation d'un système non linéaire est plus délicat. L'observabilité dans ce cas n'est qu'une condition nécessaire pour l'existence d'un observateur.

2.3.1 Définitions

À travers ce paragraphe, on se propose de présenter les différentes définitions [13] [11] de la notion d'observabilité qui est en liaison avec celle d'indiscernabilité.

Le but d'un observateur est d'estimer l'état. Ceci suppose que la connaissance de l'entrée et de la sortie sur un intervalle de temps $[0, t[$ avec $t > 0$, permette de discerner tout couple d'états initiaux.

Considérons le système non linéaire suivant:

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= h(x(t)) \end{aligned} \tag{2.2}$$

avec $u \in \Omega \subset \Re^m$, $x \in M$ une variété C^∞ connexe de dimension n , $y \in \Re^p$. f et h sont deux fonctions analytiques.

Définition 3 *Deux états initiaux x_1 et x_2 sont dits indiscernables $(x_1 I x_2)$ si, pour toute fonction d'entrée $u(t)$ et pour tout $t \geq 0$, les sorties $y(t,x_1)$ et $y(t,x_2)$ qui en résultent*

sont égales.

L'indiscernabilité est une relation d'équivalence sur M. On note par $I\{x_0\}$ la classe d'équivalence d'un état x_0 quelconque.

Si le système (2.2) ne possède pas de couples d'états initiaux distincts $\{x_1, x_2\}$ indiscernables on dit qu'il est observable.

Définition 4 *L'état x_0 est observable si l'ensemble des points indiscernables de x_0 se réduit à x_0, i.e., $I\{x_0\} = \{x_0\}$. Le système (2.2) est observable si $I\{x\} = \{x\}$ $\forall x \in M$.*

La notion d'observabilité est globale : chaque point est discernable de tous les autres, même s'ils sont très éloignés. Il existe un concept d'observabilité plus fort, c'est celui d'observabilité locale.

Définition 5 *Soit U un sous ensemble de M et x_1,$x_2 \in U$. On dit que x_1 est U-indiscernable de x_2 $(x_1 I_u x_2)$ si, pour chaque entrée $u(t)$, les trajectoires initiées en x_1 et x_2 restent dans U et les sorties correspondantes sont identiques.*

L'U-indiscernabilité n'est, en général, pas une relation d'équivalence car elle n'est pas transitive. Notons par $I_u(x_0)$ l'ensemble des états U-indiscernables de x_0 quelconque.

Définition 6 *L'état x_0 est localement observable si, pour tout voisinage ouvert U de x_0, $I_u(x_0) = \{x_0\}$. Le système (2.2) est localement observable si, pour tout $x \in M$, $I_u(x) = \{x\}$.*

Si un état est le seul indiscernable dans son voisinage alors il est faiblement observable. Ceci affaiblit le concept d'observabilité globale.

Définition 7 *L'état x_0 est faiblement observable s'il existe un voisinage ouvert V de x_0 tel que $I(x_0) \cap V = \{x_0\}$. Le système (2.2) est faiblement observable si $I(x) \cap V = \{x\}$ $\forall x \in M$.*

Le concept d'observabilité locale peut être affaibli également.

Définition 8 *L'état x_0 est localement faiblement observable s'il existe un voisinage ouvert V de x_0 tel que, pour tout voisinage U de x_0 contenant dans V, $I_u(x_0) = \{x_0\}$. Le système (2.2) est localement faiblement observable si $I_u(x) = \{x\}\ \forall x \in M$.*

Les définitions ci-dessus peuvent être schématisées par le diagramme [13] suivant :

$$\begin{array}{ccc} \text{(2.2) localement observable} & \Rightarrow & \text{(2.2) observable} \\ \Downarrow & & \Downarrow \\ \text{(2.2) localement faiblement observable} & \Rightarrow & \text{(2.2) faiblement observable} \end{array}$$

2.3.2 Condition de rang d'observabilité

Pour exprimer une condition de rang pour les systèmes non linéaires, on définit l'espace d'observation.

Définition 9 *Soit le système (2.2). L'espace d'observation noté H est le plus petit sous espace vectoriel de fonctions de $\Re^n$ à valeurs dans l'espace de sortie qui contienne $h_1, \ldots, h_p$ (h_i, $i = 1, \ldots, p$ sont les composantes de la fonction h) et qui soit fermé pour la dérivation de Lie par rapport à tous les champs de vecteurs de type $f(x,u)$, $u \in \Omega$, fixé.*

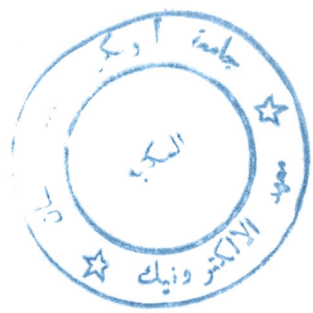
Soit dH l'espace des différentielles des éléments de H. Cet espace est défini pour une valeur de l'entrée fixée. L'espace $dH(x_0)$ (i.e. évalué en x_0) caractérise l'observabilité faible locale de (2.2) en x_0.

Définition 10 *Le système (2.2) est dit satisfaisant la condition de rang d'observabilité en x_0 si la dimension de $dH(x_0)$ est n.*

Théorème 11 *Si le système (2.2) satisfait la condition de rang d'observabilité en x_0, alors (2.2) est localement faiblement observable en x_0. Si le système (2.2) satisfait la condition de rang d'observabilité en tout point, alors il est localement faiblement observable.*

Définition 12 *Le système*

$$\dot{x} = f(x) \qquad x \in \Re^n$$
$$y = h(x) \qquad y \in \Re$$



vérifie la condition de rang d'observabilité si

$$Rg \frac{\partial}{\partial x} \begin{bmatrix} h(x) \\ L_f h(x) \\ \vdots \\ L_f^{n-1} h(x) \end{bmatrix} = n$$

2.4 Rôle des entrées dans l'observabilité

Si un système linéaire (2.1) est observable, pour une entrée $u(t)$, on peut reconstruire l'état initial. Cette propriété n'est en général pas vraie pour les systèmes non linéaires. Il se peut que certaines entrées ne permettent pas de discerner tout couple d'états initiaux distincts comme le montre l'exemple suivant :

$$\begin{aligned} \dot{x} &= u \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x \\ \dot{y} &= x_1 \end{aligned} \tag{2.3}$$

Il est aisé de vérifier que, pour $u \equiv 0$, ce système n'est pas observable : deux valeurs différentes de x_2 sont indiscernables. Ceci nous conduit à définir la notion d'entrée universelle [10].

Définition 13 *Une fonction d'entrée u est dite universelle pour le système (2.2) sur l'intervalle $[0, t]$ si tout couple d'états initiaux $\{x_1, x_2\}$ peut être discerné par les sorties sur l'intervalle $[0, t]$, le système étant excité par u, c'est à dire s'il existe $\tau \in [0, t]$ tel que $h(x(\tau, x_1)) \neq h(x(\tau, x_2))$.*

- Une entrée universelle sur $\Re^+$ est dite universelle;

- Une entrée non universelle est dite singulière;

- Si $0 < t_1 < t_2$, alors une entrée universelle sur $[0, t_1]$ est aussi universelle sur $[0, t_2]$.

Pour les systèmes affines en l'état, on sait définir un indice d'universalité de l'entrée u.

Soit pour le système

$$\begin{aligned} \dot{x}(t) &= A(u(t))x(t) + B(u(t)) \\ y &= C(u(t)) \end{aligned} \tag{2.4}$$

le grammien d'observabilité

$$\Gamma_u(t, t_0) = \int_{t_0}^{t} \Phi_u^T(\tau, t) C^T(u(\tau)) C(u(\tau)) \Phi_u(\tau, t) d\tau$$

$\Phi_u(\tau, t)$ étant la matrice de transition du système (2.4) obtenue par l'application de la fonction $u(t)$ au système sans le terme $B(u(t))$.

L'indice d'universalité $\gamma_u(t, t_0)$ est la plus petite valeur singulière de $\Gamma_u(t, t_0)$.

Pour les systèmes affines en l'état (2.4) on a la définition suivante :

Définition 14 *Une entrée u est universelle sur $[0, t]$ pour (2.4) si $\gamma_u(t, t_0) > 0$.*

2.5 Systèmes uniformément observables

La notion d'entrée universelle permet de définir une classe intéressante de systèmes : les systèmes uniformément observables [10].

Définition 15 *Un système dont toutes les entrées sont universelles est dit uniformément observable, ou encore, par abus, observable pour toute entrée. Si pour $t > 0$, toutes les entrées sont universelles sur $[0, t]$, le système est dit uniformément localement observable.*

- Le concept de l'observabilité locale ne distingue pas de l'observabilité simple pour les systèmes analytiques;

- Un système linéaire observable est uniformément localement observable.

On peut donner, pour les systèmes mono sortie affines en l'entrée, une condition nécessaire et suffisante d'observabilité locale uniforme.

Soit le système affine en l'entrée suivant :

$$\begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= h(x) \in \Re \end{aligned} \tag{2.5}$$

On suppose qu'il existe un domaine physique (ouvert relativement compact) $\Omega \subset \Re^n$ d'évolution de l'état, qui est le domaine d'intérêt du système. La première opération va consister à écrire (2.5) sous une forme normale. Supposons que (2.5) est observable, que $u \equiv 0$ est une entrée universelle et que le jacobien de $\{\zeta_1,\ldots,\zeta_n\} = \{h,\ldots,L_f^{n-1}h\}$ par rapport à x est de rang n en un point x_0. Ceci détermine, sur un voisinage de x_0, un système local de coordonnées dans lesquels le système (2.5) s'écrit :

$$\begin{aligned} \dot{\zeta}_1 &= \zeta_2 + \varphi_1(\zeta)u \\ &\vdots \\ \dot{\zeta}_{n-1} &= \zeta_n + \varphi_{n-1}(\zeta)u \\ \dot{\zeta}_n &= \tilde{\varphi}_n(\zeta) + \varphi_n(\zeta)u \\ y &= \zeta_1 \end{aligned} \tag{2.6}$$

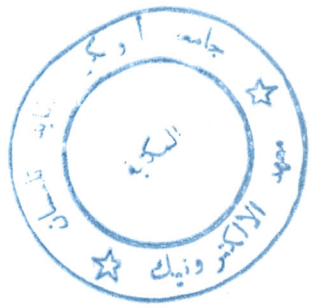
On fait alors les hypothèses suivantes :

- La fonction ζ choisie est un difféomorphisme de Ω sur $\zeta(\Omega)$;
- ζ et φ peuvent être étendues de Ω à $\Re^n$ par une fonction C^∞;
- Le système (2.6) est complet pour toutes les fonctions d'entrée admissibles (mesurables bornées) à valeurs dans $U \subset \Re^m$, de sorte que (2.6) définit globalement un système qui coïncide avec le système (2.5) sur Ω.

On peut maintenant énoncer le résultat suivant :

Théorème 16 *Supposons que le système (2.5) puisse être mis sous la forme (2.6) et satisfasse les hypothèses qui précèdent. Alors il est uniformément localement observable si et seulement si la fonction φ est de la forme*

$$\begin{aligned}
&\varphi_1(\zeta) = \varphi_1(\zeta_1)\\
&\varphi_2(\zeta) = \varphi_2(\zeta_1, \zeta_2)\\
&\vdots\\
&\varphi_{n-1}(\zeta) = \varphi_{n-1}(\zeta_1, \zeta_2, \ldots, \zeta_{n-1})
\end{aligned}$$



2.6 Entrées régulièrement persistantes

Pour le système (2.3), l'entrée $u \equiv 1$ est universelle. Au contraire l'entrée $u \equiv 0$ est singulière. La fonction $\tilde{u}$ qui vaut 1 jusqu'à $t_0 > 0$ et qui est nulle ensuite est aussi une entrée universelle. Supposons qu'on dispose d'un observateur pour ce système, et appliquons l'entrée $\tilde{u}$. Si une perturbation se produit sur x_2 à un temps $t_1 > t_0$, elle n'influence pas la sortie. Par conséquent, l'observateur ne peut pas réagir à une telle perturbation.

Les entrées universelles ne suffisent donc pas à garantir à un observateur de bonnes propriétés en présence de perturbations. Ceci amène à introduire la notion plus forte de persistance régulière qu'on sait définir pour les systèmes affines en l'état [10].

Définition 17 *Une fonction d'entrée u est dite régulièrement persistante pour le système affine en l'état (2.4) s'il existe $T > 0$, $\alpha > 0$ et $t_0 > 0$ tels que $\gamma_u(t+T, t) \geq \alpha$ pour tout $t \geq t_0$.*

Remarque 18 *Une entrée u régulièrement persistante est universelle. Non seulement sa translatée $u\delta(t) = u\delta(t+\delta)$ reste universelle pour δ arbitrairement grand (persistance), mais elle le reste avec une qualité garantie (régularité).*

2.7 Formes canoniques d'observabilité

Ces formes sont souvent utilisées dans la synthèse d'observateurs non linéaires. Elles sont obtenues grâce à une transformation de coordonnées d'état. Dans cette partie, seront proposées les formes canoniques d'observabilité de systèmes mono et multi sortie sans commande [5] [15].

Considérons le système non linéaire décrit par

$$\begin{aligned} \dot{x} &= f(x(t), u(t)) \\ y &= h(x(t), u(t)) \end{aligned} \tag{2.7}$$

$x \in M \subset \Re^n$, $u \in \Omega \subseteq \Re^m$, $y \in \Re^p$.

Les fonctions $f(.,.)$ et $h(.)$ sont supposées analytiques.

Le système (2.7) étant supposé observable, le but est de trouver une transformation de coordonnées d'état $z = \Phi(x, u, \dot{u}, \ldots, u^{(j-1)})$, $z \in \Re^n$, permettant d'écrire (2.7) sous la forme :

$$\begin{aligned} \dot{z} &= Az + \varphi(y, u, \ldots, u^{(j)}) \\ y &= Cz \end{aligned} \tag{2.8}$$

Il est facile de remarquer que le système

$$\dot{\hat{z}} = A\hat{z} + \varphi(y, u, \ldots, u^{(j)}) - KC(\hat{z} - z) \tag{2.9}$$

constitue un observateur pour (2.8) si K est choisi tel que $(A - KC)$ soit stable.

Si z est un difféomorphisme, un observateur pour le système(2.7) pourra être déduit en écrivant (2.9) dans l'espace x avec $x = \Phi^{-1}(z, u, \dot{u}, \ldots, u^{(j-1)})$.

2.7.1 Cas général

Définissons le vecteur

$$\bar{u} = \begin{bmatrix} u \\ \dot{u} \\ \vdots \\ u^{(n-1)} \end{bmatrix} \in \bar{\Omega} \subseteq \Re^{m.n}$$

et l'opérateur différentiel linéaire $\pounds_f$ tel que

$$\pounds_f(h_i) = \frac{\partial h_i}{\partial x} f(x,u) + \frac{\partial h_i}{\partial \bar{u}} \dot{\bar{u}}$$

$$\pounds_f(dh_i) = dh_i \frac{\partial f}{\partial x} + f^T \frac{\partial}{\partial x}(dh_i)^T + \dot{\bar{u}}^T \left(\frac{\partial}{\partial \bar{u}}(dh_i)^T\right)^T$$

Cet opérateur vérifie $d\pounds_f(h_i) = \pounds_f(dh_i)$.

Le système (2.7) étant supposé observable alors il existe une transformation d'état :

$$z = \begin{pmatrix} z_1(x,\bar{u}) \\ \vdots \\ z_p(x,\bar{u}) \end{pmatrix} \tag{2.10}$$

avec (pour $1 < i < n$)

$$z_i = \begin{pmatrix} y_i \\ \dot{y}_i \\ \vdots \\ y_i^{(k_i-1)} \end{pmatrix} = \begin{pmatrix} h_i \\ \pounds_f h_i \\ \vdots \\ \pounds_f^{(k_i-1)} h_i \end{pmatrix}$$

tel que le système (2.7) soit équivalent à

$$\begin{aligned} \dot{z} &= Az + \varphi(y, u, \ldots, u^{(n-1)}) \\ &= Az + \begin{pmatrix} \varphi_1(z, \bar{u}) \\ \vdots \\ \varphi_p(z, \bar{u}) \end{pmatrix} \\ y &= Cz \end{aligned} \tag{2.11}$$

où

$$A = \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{bmatrix}, \; A_i = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \; C = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_p \end{bmatrix}, \; C_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

$$\varphi_i = \left. \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \pounds_f^{k_i} h_i \end{pmatrix} \right|_{(x,\bar{u})}$$

La taille de chaque bloc i est k_i, $i = 1, \ldots, p$, avec $\sum_{i=1}^{p} k_i = n$. Les k_i sont appelés indices d'observabilité. La représentation (2.11) est appelée forme canonique d'observabilité.

Remarque 19 *La transformation de coordonnées d'état (2.10) est unique pour un vecteur de sortie, i.e., pour des indices d'observabilité donnés, la forme canonique d'observabilité est unique.*

2.7.2 Cas d'un système multi sortie sans commande

Dans ce cas l'opérateur différentiel se réduit à la dérivation de Lie. La forme canonique est donnée par (2.11). Les φ_i ne dépendront que de z.

2.7.3 Cas d'un système mono sortie sans commande

Dans ce cas également $\pounds_f$ se réduit à la dérivation de Lie. Ayant $p = 1$, on obtient

$$z = \begin{pmatrix} h(x) \\ L_f h(x) \\ \vdots \\ L_f^{(n-1)} h(x) \end{pmatrix}$$

et la forme canonique est exprimée par

$$\begin{aligned} \dot{z} &= \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix} z + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ L_f^{(n-1)} h \end{bmatrix} \\ y &= \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} z \end{aligned}$$

2.8 Observateurs non linéaires

Considérons de nouveau le système (2.2)

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)) \\ y(t) &= h(x(t)) \end{aligned}$$

Définition 20 *On appelle observateur [10], ou reconstructeur d'état, du système dynamique (2.2) un système dynamique dont les entrées sont constituées des vecteurs d'entrée et de sortie du système à observer et dont le vecteur de sortie, noté $\hat{x}$, est l'état estimé*

$$\begin{aligned} \dot{z} &= \hat{f}(z, y, u) \\ \hat{x} &= \hat{h}(z, y, u) \end{aligned} \tag{2.12}$$

tel que :

- $\|e(t)\| = \|\hat{x}(t) - x(t)\| \to 0$ *quand* $t \to \infty$;
- *Si à* $t = t_0$ *on a* $\hat{x}(t_0) = x(t_0)$ *alors pour tout* $t \geq t_0$ *on a* $\hat{x}(t) = x(t)$.

Remarque 21 *Le rôle de l'observateur présenté ci-dessus laisse penser que le problème posé est celui de la méconnaissance de l'état initial. En fait, en pratique, on recherche plutôt l'estimation de l'état à chaque instant, afin de calculer par exemple une loi de commande performante. Par contre, chaque perturbation non mesurée est perçue comme une réinitialisation par l'observateur qui doit estimer un état dépendant de cette perturbation.*

Remarque 22 *Le lien entre la commande et l'observateur est très fort, vu que les informations nécessaires à la synthèse d'une loi de commande sont issues de l'observateur. Le schéma de la figure 2.1. peut donc être complété par un bouclage système observateur (figure 2.2.)*

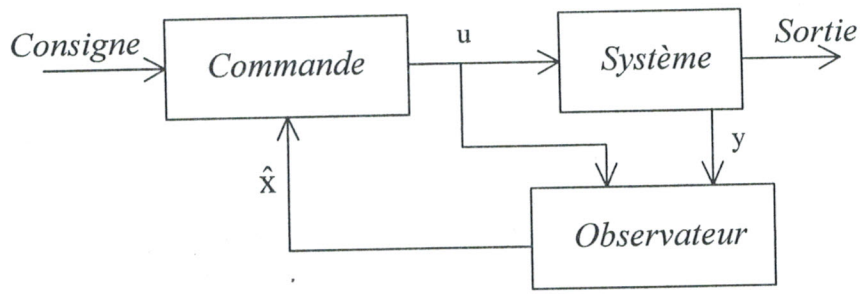


Figure 2.2. Bouclage système-observateur.

Hypothèse 23 *Il existe un voisinage* $U \subset M$ *de l'origine tel que, si* $e(t_0) \in U$, *alors* $\|e(t)\| \leq Ke^{-ct}$, *où* K *et* c *sont des constantes positives [15].*

Définition 24 *S'il existe un observateur (2.12) pour le système (2.2) qui vérifie la définition (20) et l'hypothèse (23), alors l'observateur est dit (localement) exponentiel. Si* $U \equiv M$, *l'observateur est dit globalement exponentiel [15].*

Tout au long de ce paragraphe, on s'intéresse à des classes particulières de systèmes non linéaires car il est très difficile d'effectuer la synthèse pour des classes plus générales.

2.8.1 Observateurs à grands gains

Cet observateur est formée d'une recopie des dynamiques du système non linéaire et d'un terme dépendant linéairement de l'erreur d'estimation de la mesure. l'idée est d'augmenter l'influence des dynamiques linéaires de l'observateur par rapport aux dynamiques non linéaires du système à l'aide de gains élevés.

Cette classe s'applique aux systèmes uniformément observables [5]. Considérons le système affine en l'entrée suivant :

$$\begin{aligned} \dot{x} &= f(x) + g(x)u \\ y &= h(x) \end{aligned} \tag{2.13}$$

$u(t)$ appartient à l'ensemble $U \subset \Re^m$ des valeurs admissibles de l'entrée. On suppose de plus qu'il existe un domaine physique (ouvert, borné) $\Omega \subset \Re^n$. L'entrée $u \equiv 0$ est supposée universelle. Le jacobien de $\{h_1, L_f h_1, \ldots, L_f^{n-1} h_p\}$ par rapport à x est de rang n presque partout sur $\Re^n$. Au voisinage d'un point régulier on peut sélectionner un sous ensemble de rang plein :

$$\{z_1, \ldots, z_n\} = \{h_1, L_f h_1, \ldots, L_f^{\eta_1} h_1, \ldots, h_p, \ldots, L_f^{\eta_p} h_p\}$$

Ces hypothèses conduisent à un système local de coordonnées dans lequel (2.13) s'écrit

$$\begin{aligned} \dot{z} &= Az + \tilde{\varphi}(z) + \bar{\varphi}(z)u \\ y &= Cz \end{aligned}$$

avec

$$A = \begin{bmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_p \end{bmatrix}, \; A_k = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & 0 \end{bmatrix}, \; C = \begin{bmatrix} C_1 & & \\ & \ddots & \\ & & C_p \end{bmatrix}, \; C_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$$

$$\tilde{\varphi}(z) = \begin{bmatrix} \tilde{\varphi}_1(z) \\ \vdots \\ \tilde{\varphi}_p(z) \end{bmatrix} \quad , \quad \tilde{\varphi}_k(z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \breve{\varphi}_k(z) \end{bmatrix}$$

La dimension des A_k est η_k, $k = 1, \ldots, p$, avec $\sum_{i=1}^{p} \eta_i = n$ et $\mu_1 = 1$, $\mu_k = \mu_{k-1} + \eta_{k-1}$, $k = 2, \ldots, p$.

En négligeant la linéarité en u dans ce qui suit, les systèmes considérés seront de la forme

$$\begin{aligned} \dot{z} &= Az + \varphi(z, u) \\ y &= Cz \end{aligned} \tag{2.14}$$

On fait alors les hypothèses suivantes :

- La fonction z choisie est un difféomorphisme de Ω sur $z(\Omega)$;
- z et φ peuvent être étendues de Ω vers $\Re^n$ par une fonction C^∞;
- Le système (2.14) est complet pour toutes les fonctions d'entrée admissibles (mesurables, bornées) à valeurs dans $U \subset \Re^m$, de sorte que (2.14) définit globalement un système qui coïncide avec le système (2.13) sur Ω.

Systèmes mono sortie

On s'intéresse dans un premier temps au système (2.14) à une seule sortie.

Théorème 25 *On considère le système (2.14) avec $p = 1$, dans lequel :*

i. la fonction φ est globalement lipschitzienne par rapport à x, uniformément par rapport à u.

Soit K une matrice $n \times 1$ telle que la matrice $(A - KC)$ ait toutes ses valeurs propres à partie réelle négative, et $\Lambda(T) = \begin{bmatrix} T & & & \\ & T^2 & & \\ & & \ddots & \\ & & & T^n \end{bmatrix}$.

Supposons que :

ii. $\dfrac{\partial \varphi_i}{\partial x_j} \equiv 0$ *pour* $i = 1, \ldots, n-1$; $j = i+1, \ldots, n$.

Alors le système (2.14) est uniformément localement observable, et il existe T_0 tel que, pour tout T qui satisfait $0 < T \leq T_0$, le système

$$\dot{\hat{z}} = A\hat{z} + \varphi(\hat{z}, u) + \Lambda^{-1}(T)K(y - C\hat{z})$$

est un observateur pour le système (2.14). De plus la norme de l'erreur d'observation est bornée par une exponentielle dont la vitesse de décroissance peut être choisie arbitrairement grande.

Systèmes multi sortie

Considérons maintenant le système (2.14) avec plusieurs sorties. Appelons σ_i, $i = 1, \ldots, n$, les puissances de T dans Λ.

Théorème 26 *On considère le système (2.14) dans lequel :*

i. la fonction φ est globalement lipschitzienne par rapport à x, uniformément par rapport à u.

Soit K une matrice de dimension adéquate telle que, pour chaque bloc k, la matrice $(A_k - K_k C_k)$ ait toutes ses valeurs propres à partie réelle négative.

Supposons qu'il existe deux ensembles d'entiers $\sigma = \{\sigma_1, \ldots, \sigma_n\}$ et $\delta = \{\delta_1, \ldots, \delta_p\}$ avec

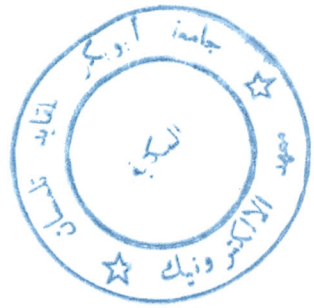
ii. $\sigma_{\mu_k+l} = \sigma_{\mu_k+l-1} + \delta_k$, $k = 1, \ldots, p$, $l = 1, \ldots, \eta_k - 1$;

iii. $\dfrac{\partial \varphi_i}{\partial x_j} \not\equiv 0 \Rightarrow \sigma_i \geq \sigma_j$, $i, j = 1, \ldots, n$, $j \neq \mu_k$, $k = 1, \ldots, p$.

Alors le système (2.14) est uniformément localement observable, et il existe T_0 *tel que, pour tout*

T, $0 < T \leq T_0$, *le système suivant*

$$\dot{\hat{z}} = A\hat{z} + \varphi(\hat{z}, u) + \Lambda^{-1}(T, \delta)K(y - C\hat{z})$$



avec $\hat{z}_{\mu_k} = y_k$, $\hat{z}_j = \hat{z}_j$, $j \neq \mu_k$

$$\Lambda(T, \delta) = \begin{bmatrix} T^{\delta_1}\Delta(T^{\delta_1}) & & \\ & \ddots & \\ & & T^{\delta_p}\Delta(T^{\delta_p}) \end{bmatrix}, \quad \Delta_k(T) = \begin{bmatrix} 1 & & \\ & T & \\ & & \ddots \\ & & & T^{\eta_k - 1} \end{bmatrix}$$

est un observateur pour (2.14). De plus la norme de l'erreur d'observation est bornée par une exponentielle dont la vitesse de décroissance peut être choisie arbitrairement grande.

Remarque 27 *La terminologie de 'grand gain' est utilisée car l'emploi de* T *suffisamment petit (*T^{-1} *suffisamment grand) permet de cacher l'effet de la partie non linéaire de la synthèse de l'observateur.*

Donc comme on l'a constaté ce type d'observateur nécessite des conditions pour sa synthèse. Aussi, il présente un inconvénient majeur : c'est sa grande sensibilité aux bruits de mesure, ceci est dû à la présence de gains élevés.

2.9 Conclusion

Dans ce chapitre, nous avons défini la notion d'observabilité ainsi que certaines propriétés des entrées appliquées aux systèmes considérés. Ces deux points fournissent des conditions

nécessaires à la synthèse d'un observateur. Ceci est important dans la mesure où il faut d'abord s'assurer de l'existence d'une solution avant de se lancer dans la recherche. À cet égard, nous avons mis en évidence la dépendance de l'observabilité vis-à-vis de l'entrée car il existe, en général, des entrées singulières qui entraînent une perte d'observabilité.

Nous avons souligné, également, le principe de séparation, un concept qui facilite considérablement la synthèse mais qui n'est, généralement, pas applicable dans le cas non linéaire. Toutefois, il a été montré que certaines classes de systèmes non linéaires vérifient un principe de séparation affaibli.

On a, ensuite, parlé de la mise des systèmes non linéaires sous formes canoniques d'observabilité : ceci sert pour le calcul de l'équation différentielle entrée-sortie sur laquelle se fondent quelques méthodes de synthèse. Dans un dernier point, on a présenté le principe de l'observateur à grand gain.

Chapitre 3

Eléments de la théorie de l'approximation

3.1 Introduction

Le contexte du problème d'approximation peut être des plus divers. Les techniques proposées sont si nombreuses qu'on se demande, parfois, si cette diversité n'est pas due à l'absence d'une approche scientifique. Si cette approche avait existé, peut être, aurait elle permis de dégager une méthode d'approximation à la fois optimale et universelle. Les différents chapitres de la théorie des approximations, et notamment l'interpolation, peuvent être assimilés à l'étude des modèles abstraits de certaines classes concrètes de problèmes.

1. Ci-dessous un problème simple conduisant à l'approximation des fonctions [3]. On observe aux instants discrets $t_1, \ldots, t_n$ les valeurs d'une fonction $f(t)$; on demande de restituer ses valeurs pour d'autres t. Il arrive qu'on sache, certaines considérations supplémentaires indiquant que la fonction d'approximation doit être de la forme

$$f(t) \approx g(t; a_1, \ldots, a_n)$$

Si les paramètres $a_1, \ldots, a_n$ sont définis par les conditions de coïncidence de $f(t)$ et de la fonction approximante aux points $t_1, \ldots, t_n$

$$g(t_j; a_1, \ldots, a_n) = f(t_j) \quad j = 1, \cdots, n$$

on est en présence d'une méthode d'approximation appelée interpolation. Si le point n'appartient pas au segment $[\min t_i, \max t_i]$, on parle d'extrapolation.

2. On sait souvent que la fonction est bien approchée par des fonctions d'une certaine forme, par exemple, par des polynômes, mais on ignore la façon de choisir au mieux le degré du polynôme approximant. Le problème se complique singulièrement si les valeurs données de la fonctions sont entachées de grosses erreurs.

3. On connait la forme d'un bon approximant de la fonction, mettons que ce soit un polynôme du deuxième degré. D'autre part, les valeurs de la fonction sont mesurées avec une grande erreur. On demande d'obtenir une meilleure approximation relativement à une certaine norme avec un nombre minimal de mesures. Ce problème se rencontre en biologie, chimie, physique, géographie et art militaire.

L'approximation des fonctions est un problème qui est susceptible de se résoudre aussi bien séparément que dans le cadre d'autres problèmes plus vastes. Le présent chapitre est consacré à ce procédé d'approximation très répandu qu'est l'interpolation. L'interpolation est un outil auxiliaire de premier plan de l'analyse numérique, notamment dans l'intégration et la dérivation numériques, la résolution d'équations différentielles.

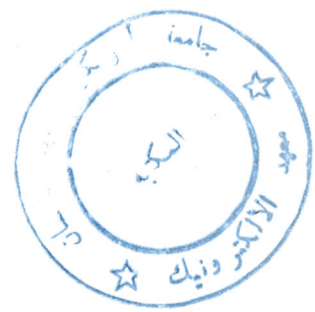
Le présent chapitre comprend un paragraphe dans lequel on rappelle certains concepts des espaces vectoriels. La partie qui suit expose quelques notions fondamentales de la théorie concernant l'approximation de fonctions. On évoquera ensuite l'interpolation polynômiale. On s'intéressera en particulier à l'interpolation de Lagrange, et dans le but d'établir une estimation des erreurs de cette méthode on définira ce qu'est une différence divisée.

On va au préalable rappeler certaines définitions [3].

3.2 Définitions

Un ensemble d'éléments E constitue un espace vectoriel s'il est muni des opérations internes d'addition et de multiplication par un nombre appartenant à un corps K satisfaisant les conditions

- l'addition est associative: $(x+y)+z = x+(y+z) \quad \forall x, y, z \in E$;
- l'addition est commutative: $x+y = y+x \quad \forall x, y \in E$;
- il existe un élément nul tel que: $x+0 = x \ \ \forall x \in E$;
- $0.x = 0$ pour tout $x \in E$;
- $(\alpha+\beta).x = \alpha.x + \beta.x \ \ \forall x \in E, \forall \alpha, \beta \in K$;
- $\alpha.(x+y) = \alpha.x + \alpha.y \quad \forall x, y \in E, \forall \alpha \in K$;
- $\alpha.(\beta.x) = (\alpha.\beta).x \ \ \forall x \in E, \forall \alpha, \beta \in K$;
- $1.x = x$ pour tout $x \in E$.



On introduit dans l'espace vectoriel la notion de dépendance et d'indépendance linéaires d'éléments. Un ensemble d'éléments $x_1, \ldots, x_n$ d'un espace vectoriel E est dit linéairement dépendant s'il existe des $c_1, \ldots, c_n$ non tous nuls tels que

$$c_1 x_1 + \ldots + c_n x_n = 0$$

Dans le cas contraire le système est dit linéairement indépendant.

Un sous ensemble F d'un sous espace vectoriel est un sous espace vectoriel si $x, y \in F$ implique $\alpha.x + \beta.y \in F \ \forall \alpha, \beta \in K$.

Un espace E est dit métrique si l'on peut faire correspondre à deux quelconques de ses éléments la distance $d(x,y)$ vérifiant les conditions :

- $d(x,y) \geq 0$, $d(x,y) = 0$ ssi $x = y$;
- $d(x,y) = d(y,x)$;
- $d(x,y) \leq d(x,z) + d(z,y)$ $\forall x, y, z \in E$.

L'espace E est dit espace vectoriel normé si :

a. il est vectoriel;

b. on peut faire correspondre à tout élément $f \in E$ un nombre réel $\|f\|$ appelé norme de f, tel que

1. $\|f\| \geq 0$, $\|f\| = 0$;
2. $\|\alpha f\| = |\alpha| \, . \, \|f\|$ pour tout $\alpha \in K$;
3. $\|f_1 + f_2\| \leq \|f_1\| + \|f_2\|$;

Enfin, on dit que l'espace E est strictement normé si on a

- $\|f_1 + f_2\| = \|f_1\| + \|f_2\|$ si et seulement si $\exists \alpha \geq 0$ tel que $f_2 = \alpha f_1$.

3.3 Approximation des fonctions continues

De manière schématique, le problème type dans la théorie de l'approximation de fonctions est le suivant : approcher les éléments d'un espace fonctionnel E par les éléments d'un sous ensemble F donné.

Supposons que E est un espace normé. On peut espérer, pour le problème considéré, deux types de solutions:

- F est dense dans E

Autrement dit tout élément de E peut être approché d'aussi près que l'on veut par des éléments de F.

- F n'est pas dense dans E

Si $x \in E$, alors la distance de x à F

$$d(x, F) = \inf_{y \in F} \|x - y\|$$

est en général non nulle, est le problème intéressant est alors le suivant:

$$\begin{cases} \text{trouver } \widehat{x} \in F \text{ tel que} \\ \|x - \widehat{x}\| = d(x, F) \end{cases} \tag{3.1}$$

Définition 28 *Dans les conditions qui précèdent, si $\widehat{x}$ vérifie (3.1), on dit que $\widehat{x}$ est un meilleur approximant de x dans F.*

3.3.1 Meilleure approximation dans un espace vectoriel normé

Soit f un élément de l'espace vectoriel normé E. Le but est de trouver une meilleure approximation pour f sous la forme d'une combinaison linéaire $\sum_{i=1}^{n} \lambda_i \varphi_i$ d'éléments linéairement indépendants $\varphi_1, \ldots, \varphi_n \in E$. Cela équivaut à trouver un élément $\sum_{i=1}^{n} \lambda_i^* \varphi_i$ (meilleur approximant) tel que

$$\left\| f - \sum_{i=1}^{n} \lambda_i^* \varphi_i \right\| = \inf_{\lambda_1, \ldots, \lambda_n} \left\| f - \sum_{i=1}^{n} \lambda_i \varphi_i \right\|$$

Théorème 29 *Le meilleur approximant existe.*

Preuve. Etant donné que

$$\left| \left\| f - \sum_{i=1}^{n} \lambda_i^1 \varphi_i \right\| - \left\| f - \sum_{i=1}^{n} \lambda_i^2 \varphi_i \right\| \right| \leq \left\| \sum_{i=1}^{n} (\lambda_i^1 - \lambda_i^2) \varphi_i \right\| \leq \sum_{i=1}^{n} |\lambda_i^1 - \lambda_i^2| \, \|\varphi\|$$

la fonction

$$F_f(\lambda_1, \ldots, \lambda_n) = \left\| f - \sum_{i=1}^{n} \lambda_i \varphi_i \right\|$$

est une fonction continue en ses arguments λ_j quelle que soit $f \in \Re$. Soit $\|\lambda\|$ la norme euclidienne du vecteur $\lambda = (\lambda_1, \ldots, \lambda_n)$. La fonction

$$F_0(\lambda_1, \ldots, \lambda_n) = \|\lambda_1 \varphi_1 + \cdots + \lambda_n \varphi_n\|$$

est continue sur la sphère unité $\|\lambda\| = 1$ et un point $(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ de celle-ci réalise donc sa borne inférieure $\tilde{F}$ sur la sphère. $\tilde{F} \neq 0$ puisque l'égalité $\tilde{F} = \left\| \tilde{\lambda}_1 \varphi_1 + \cdots + \tilde{\lambda}_n \varphi_n \right\| = 0$ contredit l'indépendance linéaire des éléments $\varphi_1, \ldots, \varphi_n$. Pour tout $\lambda = (\lambda_1, \ldots, \lambda_n) \neq (0, \ldots, 0)$, on a l'estimation suivante

$$\begin{aligned} \|\lambda_1 \varphi_1 + \cdots + \lambda_n \varphi_n\| &= F_0(\lambda_1, \ldots, \lambda_n) \\ &= \|\lambda\| \, F_0\left(\frac{\lambda_1}{\|\lambda\|}, \cdots, \frac{\lambda_n}{\|\lambda\|}\right) \geq \|\lambda\| \, \tilde{F} \end{aligned}$$

Posons $\gamma = \dfrac{2\|f\|}{\tilde{F}}$. La fonction $F_f(\lambda_1, \ldots, \lambda_n)$ est continue dans la boule $\|\lambda\| \leq \gamma$ et atteint donc en un point $(\lambda_1^*, \ldots, \lambda_n^*)$ sa borne inférieure F^* sur la boule. On a $F^* \leq F_f(0, \ldots, 0) = \|f\|$. A l'extérieur de cette boule

$$F_f(\lambda_1, \ldots, \lambda_n) \geq \|\lambda_1 \varphi_1 + \cdots + \lambda_n \varphi_n\| - \|f\| > \tilde{F} \frac{2\|f\|}{\tilde{F}} - \|f\| = \|f\| \geq F^*$$

Ainsi, $F_f(\lambda_1, \ldots, \lambda_n) \geq F^* = F_f(\lambda_1^*, \ldots, \lambda_n^*)$ quels que soient $\lambda_1, \ldots, \lambda_n$. ■

Il existe en général plusieurs meilleurs approximants. Le théorème suivant donne une condition nécessaire pour l'unicité du meilleur approximant.

Théorème 30 *Si l'espace E est strictement normé, le meilleur approximant est unique.*

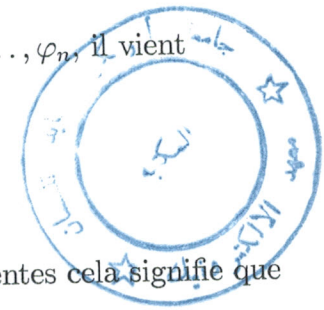
Preuve. On raisonne par l'absurde et suppose qu'il existe deux éléments $f_1 \neq f_2$, $f_j = \sum_{i=1}^{n} \lambda_{ij} \varphi_i$ tels que $\|f - f_1\| = \|f - f_2\| = \Delta$. Il est clair que $\Delta \neq 0$, sinon $f_1 = f =$

f_2. De plus

$$\left\| f - \frac{f_1 + f_2}{2} \right\| = \left\| \frac{f - f_1}{2} + \frac{f - f_2}{2} \right\| \leq \left\| \frac{f - f_1}{2} \right\| + \left\| \frac{f - f_2}{2} \right\| = \Delta$$

Puisque $\dfrac{f_1 + f_2}{2}$ est combinaison linéaire des éléments $\varphi_1, \ldots, \varphi_n$, il vient

$$\left\| f - \frac{f_1 + f_2}{2} \right\| \geq \Delta$$



selon la définition de Δ. Compte tenu des relations précédentes cela signifie que

$$\left\| \frac{f - f_1}{2} \right\| + \left\| \frac{f - f_2}{2} \right\| = \left\| \frac{f - f_1}{2} + \frac{f - f_2}{2} \right\|$$

Du moment que l'espace est supposé strictement normé, on a

$$\frac{f - f_1}{2} = \alpha \frac{f - f_2}{2}$$

Si $\alpha \neq 1$, alors $f = \dfrac{f_1 - \alpha f_2}{1 - \alpha}$ est une combinaison linéaire des $\varphi_1, \ldots, \varphi_n$ et, partant, $\Delta = 0$. Quand $\alpha = 1$, on a $f_1 = f_2$ ce qui contredit l'hypothèse $f_1 \neq f_2$. ■

3.3.2 Meilleure approximation uniforme

Si dans un espace vectoriel la norme n'est pas déterminée à l'aide d'un produit scalaire, la recherche du meilleur approximant se complique sensiblement.

Soit E un espace de fonctions réelles bornées définies sur $[a, b] \in \Re$ avec pour norme

$$\|f\| = \sup_{[a,b]} |f(x)|$$

On cherche la meilleure approximation de la forme

$$P_n(x) = \sum_{j=0}^{n} a_j x^j$$

Conformément au théorème (29), la meilleure approximation existe, i.e., un polynôme $P_n^*(x)$ tel que

$$E_n(f) = \|f - P_n^*\| \leq \|f - P_n\|$$

pour n'importe quel P_n de degré n. Ce polynôme s'appelle polynôme de meilleure approximation uniforme. Dans la suite, on établira les conditions nécessaires et suffisantes pour qu'un polynôme soit la meilleure approximation pour une fonction continue.

Théorème 31 *S'il existe $n+2$ points $x_0 < \cdots < x_{n+1}$ de l'intervalle $[a,b]$ tels que $sg((f(x_i) - P_n(x_i))(-1)^i) = const$, i.e., en passant d'un point x_i au point suivant x_{i+1}, la quantité $f(x) - P_n(x)$ change de signe, alors*

$$E_n(f) \geq \mu = \min_{i=0,\ldots,n+1} |f(x_i) - P_n(x_i)|$$

Preuve. La proposition est évidente pour $\mu = 0$. Soit $\mu > 0$ et supposons le contraire, i.e., que le polynôme de meilleure approximation $P_n^*(x)$ vérifie

$$\|P_n^* - f\| = E_n(f) < \mu$$

On a $sg(P_n(x) - P_n^*(x)) = sg((P_n(x) - f(x)) - (P_n^*(x) - f(x)))$. Aux points x_i, le premier terme est supérieur en module au deuxième, aussi $sg(P_n(x_i) - P_n^*(x_i)) = sg(P_n(x_i) - f(x_i))$. Par conséquent, le polynôme $P_n(x) - P_n^*(x)$ de degré n change de signe $n+1$ fois. On aboutit à une contradiction. ■

Le résultat de ce théorème est utilisé pour démontrer les deux qui suivent.

Théorème 32 *Pour qu'un polynôme $P_n(x)$ soit le polynôme de meilleure approximation de la fonction continue $f(x)$, il faut et il suffit que l'intervalle $[a,b]$ contienne au moins*

$n+2$ *points* $x_0 < \cdots < x_{n+1}$ *tels que*

$$f(x_i) - P_n(x_i) = \alpha(-1)^i \, \|f - P_n\| \qquad i = 0, \ldots, n+1$$

$\alpha = 1$ *(ou* $\alpha = -1$*) simultanément pour tous les* i*.*

On dit que les points $x_0, \ldots, x_{n+1}$ satisfaisant aux conditions du théorème qu'ils forment le support de Tchebycheff.

Une démonstration détaillée de ce théorème est donnée en [3].

Théorème 33 *Il existe un seul polynôme, meilleure approximation de la fonction continue.*

Preuve. Supposons qu'il en existe deux de degré n

$$P_n^1(x) \neq P_n^2(x) \ , \ \|f - P_n^1\| = \|f - P_n^2\| = E_n(f)$$

Alors

$$\left\| f - \frac{P_n^1 + P_n^2}{2} \right\| \leq \left\| \frac{f - P_n^1}{2} \right\| + \left\| \frac{f - P_n^2}{2} \right\| = E_n(f)$$

Le polynôme $\dfrac{P_n^1(x) + P_n^2(x)}{2}$ est donc lui aussi une meilleure approximation uniforme. Soient $x_0, \ldots, x_{n+1}$ les points correspondants du support de Tchebycheff. Alors

$$\left| \frac{P_n^1(x_i) + P_n^2(x_i)}{2} - f(x_i) \right| = E_n(f) \qquad i = 0, \ldots, n+1$$

ou

$$\left| (P_n^1(x_i) - f(x_i)) + (P_n^2(x_i) - f(x_i)) \right| = 2E_n(f)$$

Puisque $\left| P_n^k(x_i) - f(x_i) \right| \leq E_n(f)$, $k = 1, 2$, la dernière relation n'a lieu que si

$$P_n^1(x_i) - f(x_i) = P_n^2(x_i) - f(x_i)$$

Deux polynômes distincts $P_n^1(x)$ et $P_n^2(x)$ de degré n coïncident en $n+2$ points, ce qui est contraire à l'hypothèse. ■

Donc, comme on a pu le constater, le but de cette partie est de démontrer que si le meilleur approximant existe alors il est unique. On va aborder maintenant un type d'approximation très répandu : c'est celui de l'interpolation polynômiale.

3.4 Interpolation polynômiale

Pour l'approximation des fonctions, on utilise souvent des polynômes car ils peuvent être évalués, différentiés et intégrés facilement.

Un polynôme d'ordre n ou de degré $< n$ est une fonction de la forme

$$P_n(x) = a_1 + a_2 x + \cdots + a_n x^{n-1} = \sum_{j=1}^{n} a_j x^{j-1}$$

L'ensemble ou l'espace de tous les polynômes d'ordre n sera noté $\mathbb{P}_n$.

Etant données des valeurs f_i d'une fonction $f(x)$ aux points x_i, le problème d'interpolation consiste à déterminer les paramètres a_i tels que les paires (x_i, f_i), $i = 1, \ldots, n$, avec $x_i \neq x_k$ pour $i \neq k$ satisfont

$$P(x_i) = f_i \qquad i = 1, \ldots, n$$

Les paires (x_i, f_i) sont appelées points de support (ou d'appui) [2].

On a le théorème suivant

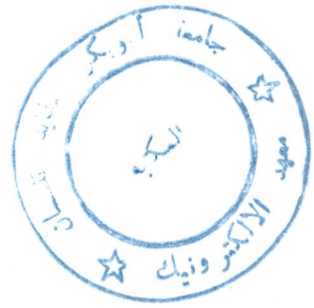
Théorème 34 *Il existe un polynôme et un seul de degré au plus $n-1$ tel que*

$$P(x_i) = f_i \qquad i = 1, \ldots, n$$

Preuve. Pour la preuve [2], il suffit de démontrer que le déterminant de Vandermonde

$$V_{n+1} = \begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & & x_1^n \\ \vdots & & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix}$$

ne s'annule pas. Comme $x_i \neq x_k$ pour $i \neq k$ on a $V_{n+1} \neq 0$. ■



3.4.1 Formule d'interpolation de Lagrange

Considérons une séquence $x = (x_i)_1^n$ de n points distincts et posons

$$l_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{t - x_j}{x_i - x_j}$$

l_i est appelé i$^{\text{ème}}$ polynôme de Lagrange pour x [7]. C'est un polynôme d'ordre n, s'annulant en tout x_j et valant 1 pour $t = x_i$. A l'aide du symbole de Kronecker, ceci s'écrit

$$l_i(x_j) = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

Ainsi pour une fonction f donnée

$$P = \sum_{i=1}^{n} f(x_i) l_i \tag{3.2}$$

est un élément de $\mathbb{P}_n$ et satisfait

$$P(x_i) = f(x_i) \qquad i = 1, \ldots, n$$

Théorème 35 *Pour n points d'appui arbitraires (x_i, f_i), $i = 1, \ldots, n$, $x_i \neq x_k$ pour $i \neq k$, il existe un polynôme $P \in \mathbb{P}_n$ tel que $P(x_i) = f(x_i)$, $i = 1, \ldots, n$. Sous la forme*

de Lagrange ce polynôme est donné par (3.2).

Dans le but de démontrer certains résultats, on va maintenant définir la notion de différence divisée [7].

Définition 36 *La $k^{ème}$ différence divisée d'une fonction f aux points $\tau_i, \ldots, \tau_{i+k}$ est le coefficient de x^k du polynôme d'ordre $k+1$ qui coïncide avec f aux points $\tau_i, \ldots, \tau_{i+k}$. Elle sera notée*

$$[\tau_i, \ldots, \tau_{i+k}\,]f$$

Cette définition a les conséquences suivantes [7] :

i. si $P_i \in \mathbb{P}_i$ coïncide avec f aux points $\tau_1, \ldots, \tau_i$ pour $i = k$ et $k+1$ alors

$$P_{k+1}(x) = P_k(x) + (x - \tau_1) \cdots (x - \tau_k)[\tau_1, \ldots, \tau_{k+1}\,]f$$

Le polynôme $P_{k+1} - P_k$ est d'ordre $k+1$ et s'annule aux points $\tau_1, \ldots, \tau_k$ et son coefficient principal est $[\tau_1, \ldots, \tau_{k+1}\,]f$. Ceci implique que

$$P_{k+1}(x) - P_k(x) = c(x - \tau_1) \cdots (x - \tau_k) \tag{3.3}$$

avec $c = [\tau_1, \ldots, \tau_{k+1}\,]f$.

Un polynôme $P_n(x)$ peut s'écrire sous la forme

$$P_n(x) = P_1(x) + (P_2(x) - P_1(x)) + (P_3(x) - P_2(x)) + \cdots + (P_n(x) - P_{n-1}(x))$$

En utilisant (3.3), $P_n(x)$ s'exprimera par

$$P_n(x) = [\tau_1]f + (x - \tau_1)\,[\tau_1, \tau_2]f + \cdots + (x - \tau_1) \cdots (x - \tau_{n-1})[\tau_1, \ldots, \tau_n\,]f$$

Cette forme est appelée forme de Newton.

ii. si $f \in C^{(k)}$, i.e., f a k dérivées continues, alors il existe un point ξ appartenant au plus petit intervalle contenant $\tau_i, \ldots, \tau_{i+k}$ tel que

$$[\tau_i, \ldots, \tau_{i+k}\,]f = f^{(k)}(\xi)/k!$$

iii. pour les calculs il est important de noter que

$$[\tau_i, \ldots, \tau_{i+k}]\,f = \begin{cases} f^{(k)}(\tau_i)/k! & \text{si } \tau_i = \ldots = \tau_{i+k} \text{ et } f \in C^{(k)} \\ \dfrac{[\tau_{i+1}, \ldots, \tau_{i+k}]\,f - [\tau_i, \ldots, \tau_{i+k-1}]\,f}{\tau_{i+k} - \tau_i} & \text{si } \tau_i \neq \tau_{i+k} \end{cases}$$

3.5 L'erreur d'interpolation

Si f est connue uniquement à travers les valeurs qu'elle prend aux points $(\tau_i)_1^n$ alors il est souhaitable d'évaluer l'erreur entre la fonction et son polynôme d'interpolation. Pour ce fait on utilisera quelques propriétés des différences divisées et on aura besoin du théorème suivant

Théorème 37 (Rolle) *Si la fonction f est dérivable dans l'intervalle fermé $[a, b]$ et s'annule aux extrémités de cet intervalle, alors f' s'annule au moins une fois à l'intérieur de cet intervalle.*

D'après (3.3) si on prend $k = n$ on obtient

$$P_{n+1}(x) = P_n(x) + c(x - \tau_1) \cdots (x - \tau_n)$$

avec $c = [\tau_1, \ldots, \tau_{n+1}\,]f$.

3.5.1 L'erreur sur la fonction

Posons $E(x) = P_{n+1}(x) - P_n(x)$ et $F(x) = f(x) - P_{n+1}(x)$. Fixons x et prenons par exemple $x = \tau_{n+1}$. Dans ce cas $f(x)$ s'écrira

$$f(x) = P_n(x) + c(x - \tau_1) \cdots (x - \tau_n)$$

On pose $\Pi(x) = (x - \tau_1) \cdots (x - \tau_n)$ et on obtient

$$c = \frac{f(x) - P_n(x)}{\Pi(x)}$$

On considère maintenant le polynôme d'ordre $n + 1$ en t

$$P_{n+1}(t) = P_n(t) + \Pi(t)\frac{f(x) - P_n(x)}{\Pi(x)}$$

Aux points d'interpolation on a

$$E(\tau_i) = P_{n+1}(\tau_i) - P_n(\tau_i) = 0$$

et

$$F(\tau_i) = f(\tau_i) - P_n(\tau_i) - \Pi(\tau_i)\frac{f(x) - P_n(x)}{\Pi(x)} = 0$$

D'après le théorème de Rolle (entre deux zéros consécutifs de F il existe au moins un zéro de F') et sachant que F possède $(n + 1)$ zéros sur $[\tau_1, x]$, F' admettra au moins n zéros. Ceci implique que $F^{(2)}$ a $(n - 1)$ zéros, $\ldots$, $F^{(n)}$ possède un zéro qu'on notera ξ_x et il appartient au plus petit intervalle contenant $x, \tau_1, \ldots, \tau_n$. Nous avons donc

$$0 = F^{(n)}(\xi_x) = f^{(n)}(\xi_x) - P_{n+1}^{(n)}(\xi_x)$$

P_{n+1} est un polynôme de degré n: son terme de plus haut degré est

$$t^n \frac{f(x) - P_n(x)}{\Pi(x)}$$

Ceci donne

$$f^{(n)}(\xi_x) = P_{n+1}^{(n)}(\xi_x) = n! \frac{f(x) - P_n(x)}{\Pi(x)}$$

et par conséquent

$$e(x) = f(x) - P_n(x) = \frac{1}{n!} \Pi(x) f^{(n)}(\xi_x)$$

3.5.2 Les erreurs sur les dérivées

Etant données les valeurs d'une fonction f aux points $\tau_1, \ldots, \tau_n$, on demande de calçuler $f^{(k)}(x)$. Les formules les plus simples de dérivation numériques sont fournies par la dérivation des fonctions d'interpolation.

Formons le polynôme d'interpolation $P_n(x)$ et évaluons l'erreur dans les formules de dérivation numérique [3]. On a

$$f(x) - P_n(x) = \Pi(x) \, [\tau_1, \ldots, \tau_n, x] \, f$$

d'où, d'après la formule de Leibniz

$$f^{(k)}(x) - P_n^{(k)}(x) = \sum_{j=0}^{k} C_k^j ([\tau_1, \ldots, \tau_n, x] \, f)^{(j)} \Pi^{(k-j)}(x)$$

En vertu de la propriété (ii) des différences divisées: si g est une fonction continûment dérivable, alors

$$[x - q\varepsilon, x - (q-1)\varepsilon, \ldots, x - \varepsilon, x] g = g^{(q)}(\xi\varepsilon)/q!$$

avec $x - q\varepsilon \leq \xi\varepsilon \leq x$; doù

$$g^{(q)}(x) = q! \lim_{\varepsilon \to 0} [x - q\varepsilon, x - (q-1)\varepsilon, \ldots, x - \varepsilon, x]g$$

Par conséquent

$$([\tau_1, \ldots, \tau_n, x]\, f)^{(j)} = j!(\lim_{\varepsilon \to 0} [\tau_1, \ldots, \tau_n, x - q\varepsilon, x - (q-1)\varepsilon, \ldots, x - \varepsilon, x]f)$$

L'expression entre parenthèses devient

$$\left[\tau_1, \ldots, \tau_n, \underbrace{x, \ldots, x}_{(j+1) \text{ fois}}\right] f$$

Toujours en utilisant (ii), on obtient

$$[\tau_1, \ldots, \tau_n, x, \ldots, x]\, f = f^{(n+j)}(\xi_x)/(n+j)!$$

Donc

$$f^{(k)}(x) - P_n^{(k)}(x) = \sum_{j=0}^{k} \frac{k!}{(k-j)!(n+j)!} f^{(n+j)}(\xi_x)\Pi^{(k-j)}(x)$$

Généralement, on n'a d'informations ni sur ξ_x ni sur $f^{(n+k)}$. Ceci nous conduit à imposer une borne sur l'erreur. Posons $\Delta = \max_{1 \leq i \leq n-1} (\tau_{i+1} - \tau_i)$, on a alors

$$\|e\|_\infty \leq \frac{\Delta^{n-1}}{n!} \left\|f^{(n)}\right\|_\infty$$
$$\left\|e^{(k)}\right\|_\infty \leq \theta \Delta^{n-k} \max_{0 \leq i \leq k} \left\|f^{(n+i)}\right\|_\infty$$

θ est une constante indépendante de f et de la séquence τ.

3.6 Conclusion

Dans ce chapitre, on a abordé quelques concepts essentiels de la théorie de l'approximation, à savoir, la meilleure approximation, l'interpolation polynômiale et l'évaluation des erreurs dans le cas de l'interpolation de Lagrange.

Comme on l'a vu une fonction peut avoir plusieurs approximants mais le meilleur garantie une faible distance entre la fonction et son approximant et par conséquent l'approximation résultante est meilleure. On a évoqué ensuite une des techniques les plus répandues d'interpolation : la méthode de Lagrange, vu sa simplicité. Ce dernier avantage se voit eclipser devant un inconvénient sérieux : le phénomène de Runge. Enfin, on a présenté des estimations des erreurs commises par cette méthode, car, la connaissance de la grandeur de l'erreur nous permet de savoir à quel niveau on peut relaxer certaines contraintes imposées à la précision de la méthode.

Chapitre 4

Les fonctions splines

4.1 Introduction

On a parlé dans le chapitre précédent de l'interpolation polynômiale. Ce type d'approximation de fonctions est très sensible au choix des points d'interpolation. En effet, soit à approcher la fonction $f(x) = 1/1+25x^2$ dans l'intervalle $[a,b] = [-1,1]$. Considérons des points d'interpolation équidistants, i.e., $\tau_i = (i-1)h-1$, $i = 1,\ldots,n$, avec $h = 2/(n-1)$. Les figures ci-dessous représentent la fonction $f(x)$ ainsi que ses polynômes de Lagrange pour différentes valeurs de n.

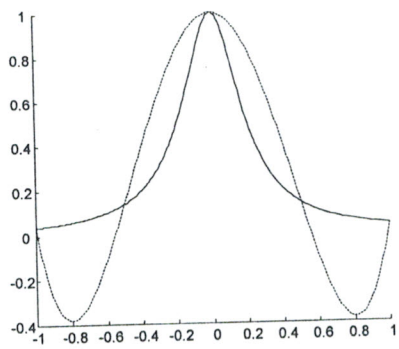


Figure 4.1. n=5.

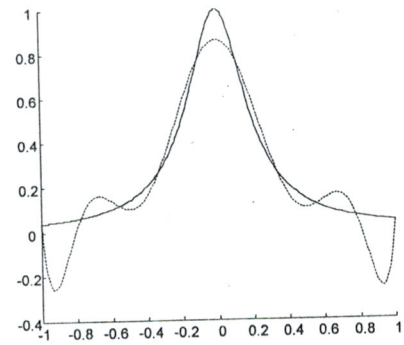


Figure 4.2. n=10.

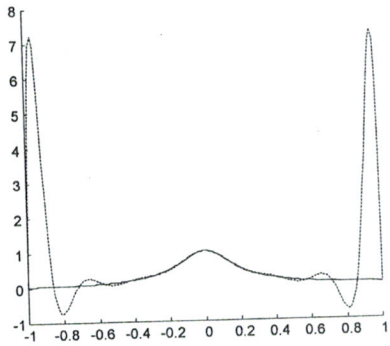


Figure 4.3. n=15.

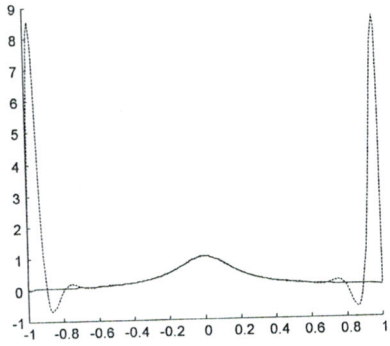


Figure 4.4. n=20.

On constate que plus n augmente, plus l'intervalle où l'approximation est bonne est grand, ceci d'une part. D'autre part, on remarque que l'amplitude des oscillations se produisant aux extrémités de l'intervalle augmente avec n. Ce phénomène est appelé phénomène de Runge [7].

Un autre choix des points d'interpolation peut remédier à ce problème. Considérons, à titre d'exemple, les points de Tchebycheff [7] donnés par: $\tau_i = (a+b-(a-b)-\cos((2i-1)\pi/2n))/2$, $i = 1, \ldots, n$. Pour les mêmes valeurs de n que précédemment on a obtenu les tracés illustrés sur les figures ci-dessous:

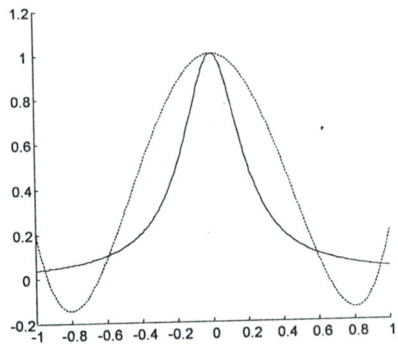


Figure 4.5. n=5.

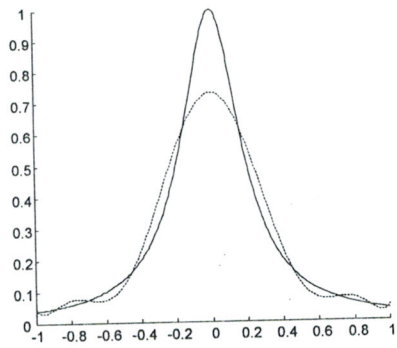


Figure 4.6. n=10.

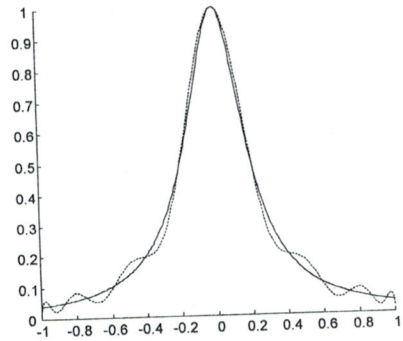


Figure 4.7. n=15.

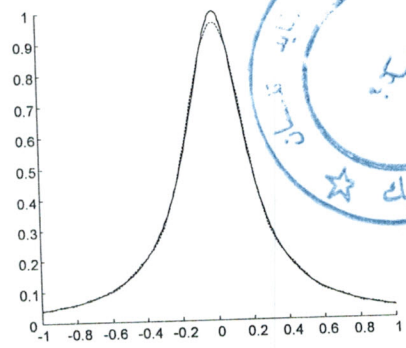


Figure 4.8. n=20.

Remarquons bien que l'approximation est meilleure pour n plus grand.

Malheureusement, cette solution n'est pas toujours la bonne car dans les problèmes d'automatique, une période d'échantillonnage (i.e., le pas entre deux points successifs) constante est toujours imposée. L'expérience a montré que l'interpolation polynômiale par morceaux fournit des résultats meilleurs. Ce procédé consiste à déterminer, sur chaque sous intervalle de l'intervalle d'interpolation, un polynôme (d'ordre approprié) coïncidant avec le morceau de la fonction défini sur ce sous intervalle.

L'idée de ce type d'approximation [1] doit son origine à la latte du dessinateur. Il s'agissait d'un problème d'interpolation : pour tracer une courbe passant par des points donnés, les dessinateurs utilisaient des lattes (en anglais "splines") flexibles. Ces lattes étaient maintenues en place par des poids de plomb, appelés "ducks". En jouant d'une part sur les points où les ducks étaient attachés à la latte, d'autre part sur la position de la latte et des points par rapport à la surface, on arrivait à faire passer la latte par les points imposés. Si on appelle Γ la courbe dessinée par l'axe déformé de la latte, la "fonction spline" mathématique sera une approximation de la courbe Γ.

La théorie mathématique des fonctions splines est récente. Cette méthode d'approximation est apparue sous sa forme actuelle pour la première fois dans un papier de Schoenberg (1946).

Ce chapitre fait un usage intensif à la référence [7]. Il sera organisé comme suit : la première partie sera consacrée à certains types de splines ainsi qu'à leurs propriétés. On abordera ensuite le concept des B-splines. La partie qui suit incluera l'interpolation par

des fonctions splines définies comme étant une combinaison linéaire de B-splines. Enfin, on évoquera, le problème d'interpolation en présence de données imprécises.

4.2 Approximation linéaire par morceaux ou splines linéaires

Ce type d'approximants n'est pas aussi pratique que les splines cubiques ou les splines d'ordre supérieur, mais il permettra d'exposer les points essentiels de l'approximation polynômiale par morceaux de la façon la plus simple qu'elle soit.

4.2.1 Définitions

Soient les points d'interpolation $\tau_1, \ldots, \tau_n$ avec $a = \tau_1 < \ldots < \tau_n = b$. L'interpolant pár lignes brisées d'une fonction f sera noté $I_2 f$. L'indice 2 représente l'ordre des morceaux polynômiaux (des lignes) formant l'interpolant. Ce dernier est défini par

$$I_2 f(x) = f(\tau_i) + (x - \tau_i)\,[\tau_i, \tau_{i+1}]\,f \qquad \text{sur } \tau_i \leq x \leq \tau_{i+1} \qquad i = 1, \ldots, n-1$$

Etant donné que $f(x) = f(\tau_i) + (x - \tau_i)\,[\tau_i, \tau_{i+1}]\,f + (x - \tau_i)(x - \tau_{i+1})\,[\tau_i, \tau_{i+1}, x]\,f$, on a pour $\tau_i \leq x \leq \tau_{i+1}$

$$f(x) - I_2 f(x) = (x - \tau_i)(x - \tau_{i+1})\,[\tau_i, \tau_{i+1}, x]\,f$$

Si on pose $\Delta\tau_i = \tau_{i+1} - \tau_i$, on obtiendra une estimation de l'erreur d'interpolation

$$|f(x) - I_2 f(x)| \leq (\Delta\tau_i/2)^2 \max_{\tau_i \leq \zeta \leq \tau_{i+1}} \left| f''(\zeta)/2 \right|$$

si f est deux fois continûment dérivable. Par conséquent

$$\|f(x) - I_2 f(x)\| \leq \frac{1}{8} |\tau|^2 \left\| f'' \right\| \qquad \text{avec } |\tau| = \max_i \Delta\tau_i \tag{4.1}$$

4.2.2 L'interpolant par lignes brisées est unique

Soit $\$_2$ l'espace linéaire de toutes les lignes continues et brisées sur $[\tau_1, \tau_n]$ avec des cassures aux points $\tau_2, \ldots, \tau_{n-1}$ ($\$_2$ indique que les splines sont d'ordre 2). On note que

$$I_2 f = f \quad \text{pour toute } f \in \$_2 \tag{4.2}$$

En effet sur chaque intervalle $[\tau_i, \tau_{i+1}]$, $I_2 f$ coïncide avec la droite interpolant f aux points τ_i et τ_{i+1}. Sachant que $f \in \$_2$, alors f est elle même une droite sur $[\tau_i, \tau_{i+1}]$ et par conséquent $I_2 f$ n'est que f. Ceci par l'unicité de l'interpolant polynômial.

4.2.3 Approximation au sens des moindres carrés par lignes brisées

Le théorème énoncé à la fin de ce paragraphe est un outil important pour la démonstration de certains résultats ultérieurement.

Tout d'abord, on va définir une base pour $\$_2$. Soient $\tau_0 = \tau_1$, $\tau_{n+1} = \tau_n$ et posons

$$H_i(x) = \begin{cases} (x - \tau_{i-1})/(\tau_i - \tau_{i-1}) & \tau_{i-1} < x \leq \tau_i \\ (\tau_{i+1} - x)/(\tau_{i+1} - \tau_i) & \tau_i \leq x < \tau_{i+1} \\ 0 \quad . & \text{ailleurs} \end{cases}$$

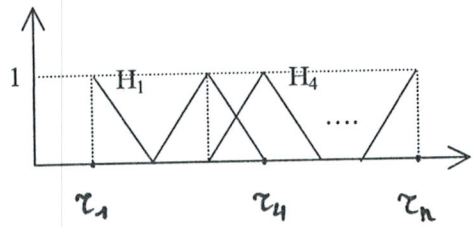


Figure 4.9. Les fonctions 'chapeau'.

Ces fonctions sont appelées fonctions "Chapeau" [7]. Il est clair que $H_i \in \$_2$, tout i,

et comme l'indique la figure

$$H_i(\tau_j) = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

Ceci montre que $\sum_1^n f(\tau_i)H_i$ est un élément de $\$_2$ coïncidant avec f en $\tau_1, \ldots, \tau_n$. En utilisant (4.2), on établit le résultat suivant

$$I_2 f = \sum_{i=1}^{n} f(\tau_i)H_i$$

Ceci implique, également, que $(Hi)_1^n$ est une base pour $\$_2$, i.e., chaque ligne brisée sur $[\tau_1, \tau_n]$ avec des cassures aux points $\tau_2, \ldots, \tau_{n-1}$ peut être écrite de façon unique sous la forme d'une combinaison linéaire des H_i. Par rapport à la base $(Hi)_1^n$, les coordonnées d'une fonction $g \in \$_2$ sont ses valeurs $(g(\tau_1), \ldots, g(\tau_n))$ aux points de cassure,i.e., $g = \sum_{i=1}^{n} g(\tau_i)H_i$, $g \in \$_2$.

Notons par $L_2 f$ l'approximation au sens des moindres carrés de f dans $\$_2$, i.e.,

$$\int |f(x) - L_2 f(x)|^2 \, dx = \min_{g \in \$_2} \int |f(x) - g(x)|^2 \, dx$$

et $L_2 f \in \$_2$. Pour obtenir $L_2 f$, il suffit de déterminer le minimum de $\int \left| f(x) - \sum_{j=1}^{n} \alpha_i H_j(x) \right|^2 dx$, en annulant sa première dérivée partielle par rapport à $\alpha_1, \ldots, \alpha_n$. Ceci permet d'obtenir le système linéaire suivant :

$$\sum_{j=1}^{n} \left[\int H_i(x) H_j(x) dx \right] \hat{\alpha}_j = \int H_i(x) f(x) dx \quad i = 1, \ldots, n$$

et donc $L_2 f$ sera exprimée par

$$L_2 f = \sum_{j=1}^{n} \hat{\alpha}_j H_j$$

Plus explicitement, on obtient :

$$(\Delta\tau_{i-1}/6)\hat{\alpha}_{i-1} + (\tau_{i+1} - \tau_{i-1})\hat{\alpha}_i/3 + (\Delta\tau_i/6)\hat{\alpha}_{i+1} = \beta_i = \int H_i(x)f(x)dx \quad i = 1, \ldots, n \tag{4.3}$$

La matrice des coefficients est tridiagonale, et strictement diagonalement dominante. La solution du système est donc unique.

Théorème 38 *L'approximation $L_2 f$ d'une fonction $f \in [a, b]$, par des éléments de $\$_2$ satisfait*

$$\|L_2 f\| \leq 3\,\|f\|$$

Sachant que L_2 est additive et que $L_2 g = g$ pour $g \in \$_2$, on a

$$\|f - L_2 f\| \leq 4\,dist(f, \$_2)$$

Preuve. On a $\|L_2 f\| = \max_i |(L_2 f)(\tau_i)| = \max_i |\hat{\alpha}_i|$. Multiplions chaque terme de (4.3) par $6/(\tau_{i+1} - \tau_{i-1})$. Ceci donne

$$\frac{\Delta\tau_{i-1}}{\tau_{i+1} - \tau_{i-1}}\hat{\alpha}_{i-1} + 2\hat{\alpha}_i + \frac{\Delta\tau_i}{\tau_{i+1} - \tau_{i-1}}\hat{\alpha}_{i+1} = 3\hat{\beta}_i \quad i = 1, \ldots, n \tag{4.4}$$

Si j est tel que $|\hat{\alpha}_j| = \|\hat{\alpha}\| = \max_i |\hat{\alpha}_i|$, alors en utilisant (4.4), on obtient

$$\begin{aligned} |2\hat{\alpha}_j| &= \left|3\hat{\beta}_j - (\hat{\alpha}_{j-1}\Delta\tau_{j-1} + \hat{\alpha}_{j+1}\Delta\tau_j)/(\Delta\tau_{j-1} + \Delta\tau_j)\right| \\ &\leq 3\left|\hat{\beta}_j\right| + |\hat{\alpha}_j| \end{aligned}$$

Ce qui implique $\|\hat{\alpha}\| \leq 3\left|\hat{\beta}_j\right| \leq 3\left\|\hat{\beta}\right\|$. D'autre part on a $\hat{\beta}_i = \int \hat{H}_i(x)f(x)dx$, avec $\hat{H}_i(x) = (2/(\tau_{i+1} - \tau_{i-1}))H_i(x)$ positive pour $\tau_{i-1} < x < \tau_{i+1}$ et nulle ailleurs, et $\int \hat{H}_i(x)dx = 1$. Ainsi, $\left|\hat{\beta}_i\right| = \left|\int \hat{H}_i(x)f(x)dx\right| \leq \int \hat{H}_i(x)dx * \max\{|f(x)| : \tau_{i-1} \leq x \leq \tau_{i+1}\} \leq \|f\|$. Par conséquent,

$$\|L_2 f\| \leq 3\,\|f\|$$

On a $\|f - L_2 f\| = \|f - g - L_2(f-g)\| \leq \|f-g\| + \|L_2(f-g)\| \leq \|f-g\| + 3\|f-g\|$. Et donc $\|f - L_2 f\| \leq 4\|f-g\| = 4\text{dist}(f, \$_2)$. ■

4.3 Interpolation cubique par morceaux

Les approximants par lignes brisés manquent de lissage et d'efficacité. Pour améliorer ces deux propriétés, on recourt à une approximation polynômiale par morceaux d'ordre plus élevé. le choix le plus connu est celui des fonctions d'approximation cubiques par morceaux.

Soient $f(\tau_1), \ldots, f(\tau_n)$ les valeurs d'une fonction f aux points $\tau_1, \ldots, \tau_n$ avec $a = \tau_1$ et $b = \tau_n$. On construit g un interpolant cubique par morceaux comme suit : sur chaque intervalle $[\tau_i, \tau_{i+1}]$, g doit coïncider avec un polynôme P_i d'ordre 4,

$$g(x) = P_i(x) \quad \text{pour } \tau_i \leq x \leq \tau_{i+1} \ , \ P_i \in \mathbb{P}_4, \ i = 1, \ldots, n$$

Le i$^{\text{ème}}$ morceau polynômial P_i doit satisfaire les conditions suivantes :

$$\begin{aligned} P_i(\tau_i) &= f(\tau_i) \ , \ P_i(\tau_{i+1}) = f(\tau_{i+1}) \\ P_i'(\tau_i) &= s_i \quad , \ P_i'(\tau_{i+1}) = s_{i+1}, \ i = 1, \ldots, n-1 \end{aligned}$$

$s_1, \ldots, s_n$ sont des paramètres aux choix. La fonction cubique par morceaux g coïncide avec f aux points $\tau_1, \ldots, \tau_n$ et est $C^{(1)}[a, b]$, i.e., continue et a une première dérivée continue sur $[a, b]$. Pour déterminer les coefficients du i$^{\text{ème}}$ morceau polynômial P_i, on utilise la formule de Newton :

$$P_i(x) = P_i(\tau_i) + (x-\tau_i)[\tau_i, \tau_i] P_i + (x-\tau_i)^2 [\tau_i, \tau_i, \tau_{i+1}] P_i + (x-\tau_i)^2 (x-\tau_{i+1}) [\tau_i, \tau_i, \tau_{i+1}, \tau_{i+1}] P_i$$

Les différents coefficients sont calculés à partir de la table des différences divisées

suivante :

	$[\,]\,P_i$	$[\,,\,]\,P_i$	$[\,,\,,\,]\,P_i$	$[,\,,\,,]\,P_i$
τ_i	$f(\tau_i)$			
		s_i		
τ_i	$f(\tau_i)$		$([\tau_i, \tau_{i+1}]\,f - s_i)/\Delta\tau_i$	
		$[\tau_i, \tau_{i+1}]\,f$		$(s_{i+1} + 2s_i - 2\,[\tau_i, \tau_{i+1}]\,f)/(\Delta\tau_i)^2$
τ_{i+1}	$f(\tau_{i+1})$		$(s_{i+1} - [\tau_i, \tau_{i+1}]\,f)/\Delta\tau_i$	
		s_{i+1}		
τ_{i+1}	$f(\tau_{i+1})$			

Le polynôme P_i s'écrit

$$P_i(x) = c_{1,i} + c_{2,i}(x - \tau_i) + c_{3,i}(x - \tau_i)^2 + c_{4,i}(x - \tau_i)^3$$

avec : $c_{1,i} = P_i(\tau_i) = f(\tau_i)$,

$c_{2,i} = P_i'(\tau_i) = s_i$,

$c_{3,i} = P_i''(\tau_i)/2 = ([\tau_i, \tau_{i+1}]\,f - s_i)/\Delta\tau_i - c_{4,i}\Delta\tau_i$,

$c_{4,i} = P_i'''(\tau_i)/6 = (s_i + s_{i+1} - 2\,[\tau_i, \tau_{i+1}]\,f)/(\Delta\tau_i)^2$.

Selon le choix des paramètres $(s_i)_1^n$, on retrouve différentes méthodes d'interpolation cubique par morceaux [7], on peut citer :

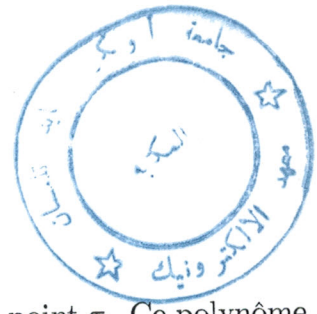
- l'interpolation cubique d'Hermite : ici, on prend $s_i = f'(\tau_i)$, tout i. On a pour $\tau_i \leq x \leq \tau_{i+1}$

$$\begin{aligned} |f(x) - g(x)| &= \left|(x - \tau_i)^2(x - \tau_{i+1})^2\,[\tau_i, \tau_i, \tau_{i+1}, \tau_{i+1}, x]\,f\right| \\ &\leq \left(\frac{\Delta\tau_i}{2}\right)^4 \max_{\tau_i \leq \xi \leq \tau_{i+1}} \left|f^{(4)}(\xi)\right|/4! \end{aligned}$$

En posant $|\tau| = \max_i \Delta\tau_i$, on peut obtenir une estimation pour l'erreur de cette mé-

thode

$$\|f - g\| \leq (1/384)\,|\tau|^4\,\|f^{(4)}\|$$



- l'interpolation cubique de Bessel

Dans ce cas s_i représente la pente d'un polynôme d'ordre 3 au point τ_i. Ce polynôme coïncide avec f aux points $\tau_{i-1}, \tau_i, \tau_{i+1}$. Après calcul, on trouve

$$s_i = \frac{\Delta\tau_i\,[\tau_{i-1}, \tau_i]\,f + \Delta\tau_{i-1}\,[\tau_i, \tau_{i+1}]\,f}{\Delta\tau_i + \Delta\tau_{i-1}}$$

- l'interpolation par splines cubiques

Les paramètres $s_2, \ldots, s_{n-1}$ sont déterminés de la condition que g doit être deux fois continûment dérivable. Pour $i = 2, \ldots, n-1$, cette condition est donnée par $P'_{i-1}(\tau_i) = P'_i(\tau_i)$ et qui peut être exprimée par

$$s_{i-1}\Delta\tau_i + s_i 2(\Delta\tau_{i-1} + \Delta\tau_i) + s_{i+1}\Delta\tau_{i-1} = b_i \tag{4.5}$$

avec $b_i = 3(\Delta\tau_i\,[\tau_{i-1}, \tau_i]\,f + \Delta\tau_{i-1}\,[\tau_i, \tau_{i+1}]\,f)$, $i = 2, \ldots, n-1$. Ceci suppose que les paramètres s_1 et s_n sont déterminés autrement.

4.3.1 Les conditions aux limites

Dans le cas de l'interpolation par splines cubiques le choix de s_1 et s_n est varié :

- si f' est connue en τ_1 et τ_n, on prendra $s_1 = f'(\tau_1)$ et $s_n = f'(\tau_n)$. La fonction spline résultante $g = I_4 f$ coïncide avec f aux points $\tau_0, \ldots, \tau_{n+1}$ (avec $\tau_0 = \tau_1$ et $\tau_{n+1} = \tau_n$) et s'appelle "une spline cubique complète".

- si f'' est connue aux limites, on peut donc imposer $g'' = f''$ en ces points en ajoutant à (4.5) les équations suivantes :

$$2s_1 + s_2 = 3\,[\tau_1, \tau_2]\,f - (\Delta\tau_1) f''(\tau_1)/2$$
$$s_{n-1} + 2s_n = 3\,[\tau_{n-1}, \tau_n]\,f - (\Delta\tau_{n-1}) f''(\tau_n)/2$$

- si on impose $g''(\tau_1) = g''(\tau_n) = 0$, on obtient les splines appelées "naturelles".
- si on ne dispose d'aucune information sur les valeurs des dérivées aux limites on utilisera alors la condition "n'est pas un nœud" (not a knot). Dans ce cas s_1 et s_n sont choisis tels que $P_1 = P_2$ et $P_{n-2} = P_{n-1}$ (i.e. le premier et le dernier nœuds intérieurs ne sont pas actifs). Ceci nécessite la continuité de g''' en τ_2 et τ_{n-1}.
- on note enfin qu'on peut faire un choix libre pour s_1 et s_n.

4.3.2 Les propriétés de meilleure approximation de l'interpolation par spline cubique complète et ses erreurs

On montrera à travers cette partie que parmi toutes les fonctions dérivables deux fois et coïncidant avec f aux points $\tau_0, \ldots, \tau_{n+1}$ la spline cubique complète $g = I_4 f$ construite précédemment minimise $\int_a^b [g''(x)]^2 dx$ [7]. Considérons les points $a = \tau_0 = \tau_1 < \ldots < \tau_n = \tau_{n+1} = b$. On rappelle que la fonction $I_4 f$ coïncidant avec f aux points $(\tau_i)_0^{n+1}$ est une spline cubique dans l'intervalle $[a, b]$ avec comme nœuds intérieurs $\tau_2, \ldots, \tau_{n-1}$, i.e., 2 fois continûment dérivable dans $[a, b]$ et ayant les points de cassures $\tau_2, \ldots, \tau_{n-1}$.

On notera par $\$_4$ l'espace linéaire de toutes les splines cubiques sur $[a, b]$ avec $\tau_2, \ldots, \tau_{n-1}$ comme nœuds intérieurs.

Le lemme suivant nous permettra de déterminer les propriétés de meilleure approximation des splines cubiques complètes.

Lemme 39 *Si f est deux fois continûment dérivable, alors la seconde dérivée de l'erreur d'interpolation $e = f - I_4 f$ est orthogonale à $\$_2$, i.e.,*

$$\int_a^b e''(x)\phi(x)dx = 0 \qquad \phi \in \$_2$$

Preuve. Le développement en série de Taylor avec reste intégral d'une fonction h (deux fois dérivable) est donné par

$$h(x) = h(a) + (x-a)h'(a) + \int_a^x (x-t)h''(t)dt$$

Puisque la seconde différence divisée d'une droite est nulle, on obtient

$$[\tau_{i-1}, \tau_i, \tau_{i+1}]\, h = [\tau_{i-1}, \tau_i, \tau_{i+1}] \int_a^x (x-t)h''(t)dt \tag{4.6}$$

Pour simplifier cette écriture, introduisons la 'fonction tronquée'

$$(x-t)_+ = \max\{0, x-t\}$$

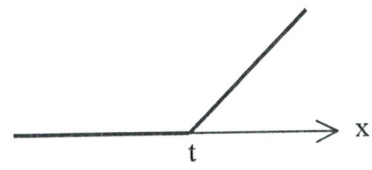


Figure 4.10. La fonction tronquée.

On peut écrire

$$\int_a^x (x-t)h''(t)dt = \int_a^b (x-t)_+ h''(t)dt \quad \text{pour } a \le x \le b$$

et donc (4.6) devient

$$\begin{aligned} [\tau_{i-1}, \tau_i, \tau_{i+1}]\, h &= \int_a^b [\tau_{i-1}, \tau_i, \tau_{i+1}]\, (x-t)_+ h''(t)dt \\ &= \int_a^b \hat{H}_i(t)h''(t)dt/2 \end{aligned}$$

où $\hat{H}_i$ est la fonction chapeau définie précédemment

$$\begin{aligned}\hat{H}_i(t) &= 2\left[\tau_{i-1},\tau_i,\tau_{i+1}\right](.-t)_+ \\ &= \frac{2}{\tau_{i+1}-\tau_{i-1}}\begin{cases}(t-\tau_{i-1})/\Delta\tau_{i-1} & \tau_{i-1}<t\leq\tau_i \\ (\tau_{i+1}-t)/\Delta\tau_i & \tau_i\leq t<\tau_{i+1} \\ 0 & \text{ailleurs}\end{cases}\end{aligned}$$

Aux points d'interpolation τ_i, l'erreur $e = f - I_4 f$ est nulle et par conséquent ses secondes différences divisées en ces points sont nulles, i.e.,

$$0 = \left[\tau_{i-1},\tau_i,\tau_{i+1}\right]e = \int_a^b \hat{H}_i(t)e''(t)dt \quad i = 1,\ldots,n$$

Ceci montre que e'' est orthogonale à chaque $\hat{H}_i$ $(i = 1,\ldots,n)$, et puisqu'elles forment une base pour $\$_2$ le lemme est démontré. ■

Théorème 40 *Pour toute fonction f 2 fois continûment dérivable sur $[a,b]$*

$$\int_a^b \left[f''(x)\right]^2 dx = \int_a^b \left[(I_4 f)''(x)\right]^2 dx + \int_a^b \left[(f - I_4 f)''(x)\right]^2 dx$$

Preuve. On a

$$\begin{aligned}\int_a^b \left[f''(t)\right]^2 dt &= \int_a^b \left[(I_4 f)''(t) + e''(t)\right]^2 dt \\ &= \int_a^b \left[(I_4 f)''(t)\right]^2 dt + 2\int_a^b (I_4 f)''(t)e''(t)dt + \int_a^b \left[e''(t)\right]^2 dt\end{aligned}$$

Mais d'après le lemme (39), $2\int_a^b (I_4 f)''(t)e''(t)dt = 0$ $((I_4 f)'' \in \$_2)$. ■

Corollaire 41 *Parmi toutes les fonctions ayant deux dérivées continues et coïncidant*

avec la fonction f aux points $\tau_0, \ldots, \tau_{n+1}$, $I_4 f$ est l'unique fonction qui minimise $\int_a^b [g''(t)]^2\, dt$.

Preuve. Pour une telle fonction g on doit avoir $I_4 g = I_4 f$ et par le théorème (40) on a

$$\begin{aligned}\int_a^b [g''(t)]^2\, dt &= \int_a^b [(I_4 f)''(t)]^2\, dt + \int_a^b [g''(t) - (I_4 f)''(t)]^2\, dt \\ &\geq \int_a^b [(I_4 f)''(t)]^2\, dt\end{aligned}$$

Cette inégalité se réduit à une égalité ssi $g'' = (I_4 f)''$ qui, par les conditions d'interpolation, est équivalente à $g = (I_4 f)$. ■

Corollaire 42 *On a*

$$(I_4 f)'' = L_2(f'')$$

Preuve. Soit $s \in \$_2$ et prenons $h(x) = \int_a^b (x-t)_+ s(t)dt$. Alors $h'' = s$ et $h \in \$_4$ et par conséquent $I_4 h = h$, $(I_4 h)'' = h'' = s$ et $h - I_4 h = 0$. Remplaçons dans le théorème (40) f par $f - h$, on obtient

$$\begin{aligned}\int_a^b [(f''(x) - s(x))]^2\, dx &= \int_a^b [((I_4 f)'' - s)(x)]^2\, dx + \int_a^b [(f - I_4 f)''(x)]^2\, dx \\ &\geq \int_a^b [f''(x) - (I_4 f)''(x)]^2\, dx\end{aligned}$$

L'égalité s'obtient ssi $(I_4 f)'' = s$. Donc $L_2(f'') = (I_4 f)''$. ■

On a établit ci-dessus les propriétés de meilleure approximation par les splines cubiques complètes. De plus, à partir du théorème (38), on a $\|(I_4 f)''\| \leq 3\|f''\|$ et $\|e''\| = \|f'' - (I_4 f)''\| \leq 4\text{dist}(f'', \$_2) \leq \frac{1}{2}|\tau|^2 \left\|f^{(4)}\right\|$. En utilisant ce résultat, on peut déterminer

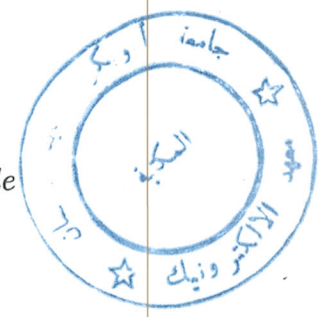
une borne sur l'erreur maximale d'interpolation $\|e\|$. Si $\tau_i \leq x \leq \tau_{i+1}$

$$\begin{aligned} e(x) &= e(\tau_i) + (x-\tau_i)\,[\tau_i, \tau_{i+1}]\,e + (x-\tau_i)(x-\tau_{i+1})\,[\tau_i, \tau_{i+1}, x]\,e \\ &= 0 + 0 + (x-\tau_i)(x-\tau_{i+1})e''(\xi_x)/2 \end{aligned}$$

pour $\xi_x \in [\tau_i, \tau_{i+1}]$. On a donc $|e(x)| \leq (\Delta\tau_i/2)^2 \max_{\tau_i \leq \xi_x \leq \tau_{i+1}} |e''(\xi_x)|/2$, et par conséquent

$$\|e\| \leq \frac{1}{8}|\tau|^2 \|e''\| \leq \frac{1}{8}|\tau|^2 4\mathrm{dist}(e'', \$_2)$$

Ceci donne



Corollaire 43 *Pour une fonction deux fois continûment dérivable*

$$\|f - I_4 f\| \leq \frac{1}{2}|\tau|^2 \, dist(f'', \$_2)$$

et en utilisant (4.1), on obtient

$$\|f - I_4 f\| \leq \frac{1}{16}|\tau|^4 \left\|f^{(4)}\right\|$$

dans le cas où f est 4 fois continûment dérivable.

Dans le cadre de la dérivation numérique, l'interpolation par splines cubiques semble fournir des résultats acceptables. On a déjà vu que

$$\|f'' - (I_4 f)''\| \leq \frac{1}{2}|\tau|^2 \left\|f^{(4)}\right\|$$

On a également pour une séquence τ uniforme

$$\|f' - (I_4 f)'\| \leq \frac{1}{24}|\tau|^3 \left\|f^{(4)}\right\|$$

Enfin, la qualité de l'approximation de la troisième dérivée dépend de τ. Hall et Meyer

[7] ont montré que

$$\left\| f^{(3)} - (I_4 f)^{(3)} \right\| \leq \frac{1}{2}(M\tau + 1/M\tau)\,|\tau|\,\left\| f^{(4)} \right\|$$

avec $M\tau = \dfrac{|\tau|}{\min\limits_{i=1,\ldots,n-1} \Delta\tau_i}$.

4.3.3 Exemple d'interpolation par splines cubiques complètes

Soit à approcher les valeurs f_i de la fonction $f(x) = \sin(x)$ aux points $\tau_i = 2\pi(i-1)/14$, $i = 1, \ldots, 15$. Soit g la fonction d'interpolation de f. Les figures ci-dessus représentent les écarts $e1(x) = \sin(x) - g(x)$, $e2(x) = \cos(x) - g'(x)$ et $e3(x) = (-\sin(x)) - g''(x)$ respectivement.

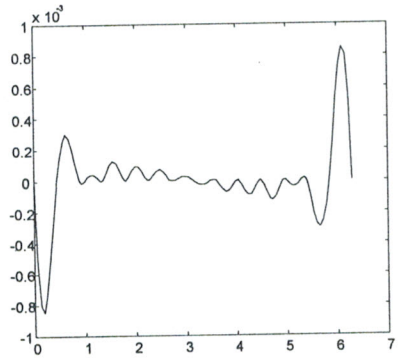


Figure 4.11. e1.

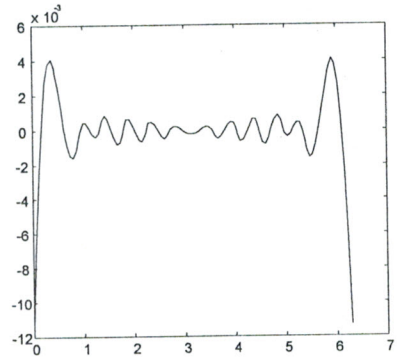


Figure 4.12. e2.

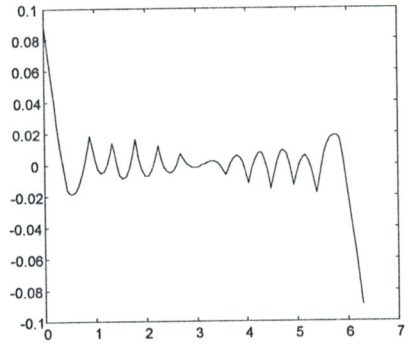


Figure 4.13. e3.

Selon ces résultats, on peut conclure que les splines cubiques sont de bons approximants. En effet l'erreur sur la deuxième dérivée reste faible.

4.4 Fonctions polynômiales par morceaux

Comme on l'a vu, les splines cubiques fournissent des résultats acceptables, mais si on a besoin de dérivées d'ordre plus élevé (> 3) on a recourt à des fonctions polynômiales par morceaux d'ordre supérieur à 4.

4.4.1 Définitions

Pour comprendre ce type d'approximation, on a besoin de définir quelques concepts.

Définition 44 *Soit $\xi = (\xi_i)_1^{l+1}$ une séquence de points strictement croissante, et soit k un entier positif. Si $P_1, \ldots, P_l$ est une séquence de polynômes, d'ordre k chacun, i.e., de degré $< k$, alors la fonction polynômiale par morceaux g d'ordre k est définie par*

$$g(x) = P_i(x) \qquad si\ \xi_i < x < \xi_{i+1}\ ,\ i = 1, \ldots, l$$

Les points ξ_i sont appelés points de cassure de g. Cette fonction peut être définie sur $\Re$ entier par l'extension du premier et du dernier morceaux, i.e.,

$$g(x) = \begin{cases} P_1(x) & \text{si } x \leq \xi_1 \\ P_l(x) & \text{si } x \geq \xi_{l+1} \end{cases}$$

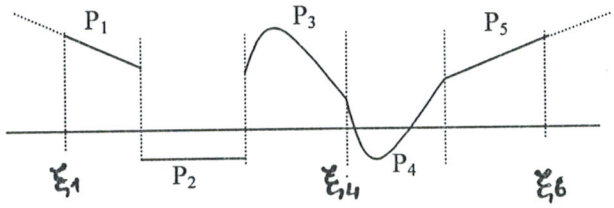


Figure 4.14. Une fonction polynômiale par morceaux (l=5).

Aux points de cassure $\xi_2, \ldots, \xi_l$ la fonction g est indéfinie puisqu'elle prend deux valeurs : $g(\xi_i^-) = P_{i-1}(\xi_i)$ (à gauche) et $g(\xi_i^+) = P_i(\xi_i)$ (à droite). Pour éviter cette situation, on supposera par la suite que g est continue à droite,i.e., $g(\xi_i) = g(\xi_i^+)$, $i = 2, \ldots, l$.

Soit $D^j g$ la jème dérivée de la fonction polynômiale par morceaux g. C'est une fonction polynômiale par morceaux d'ordre $k - j$, ayant les mêmes points de cassure que g et est constituée des j$^{\text{ème}}$ dérivées des morceaux polynômiaux constituant g. Notons par $\mathbb{P}_{k,\xi}$ l'ensemble de toutes les fonctions polynômiales par morceaux d'ordre k et de points de cassure $(\xi_i)_1^{l+1}$. Il est clair que $\mathbb{P}_{k,\xi}$ est un espace linéaire de dimension kl. A toute fonction $g \in \mathbb{P}_{k,\xi}$ correspond une représentation polynômiale par morceaux dont la valeur de la j$^{\text{ème}}$ dérivée $D^j g$ en un point x est donnée par

$$D^j g(x) = \sum_{m=j}^{k-1} c_{m+1,i}(x - \xi_i)^{m-j}/(m-j)!$$

avec : $c_{ji} = D^j g(\xi_i^+)$, $j = 1, \ldots, k$, $i = 1, \ldots, l$;
$i = 1$ et $x < \xi_2$, ou $1 < i < l$ et $\xi_i \leq x < \xi_{i+1}$, ou $i = l$ et $x \geq \xi_l$.

4.4.2 Un espace pour les fonctions polynômiales par morceaux

Etant donnée une fonction f, on désire construire une fonction polynômiale par morceaux $g \in \mathbb{P}_{k,\xi}$ satisfaisant les mêmes conditions que f. De plus, g doit avoir un certain nombre de dérivées continues. Les conditions de continuité [7] de ces dérivées sont exprimées par

$$saut_{\xi_i} D^j g = 0 \qquad j = 1, \ldots, v_i,\ i = 2, \ldots, l \tag{4.7}$$

avec $v = (v_i)_2^l$: un vecteur d'entiers non négatifs et où v_i représente le nombre de conditions de continuité nécessaires en un point ξ_i. La fonction $saut_\alpha f$ est définie par

$$saut_\alpha f = f(\alpha^+) - f(\alpha^-)$$

Le sous ensemble de toutes les fonctions $g \in \mathbb{P}_{k,\xi}$ satisfaisant (4.7), pour un vecteur v donné, est un sous espace linéaire de $\mathbb{P}_{k,\xi}$ et il sera noté $\mathbb{P}_{k,\xi,v}$. Dans cet espace tout élément $g \in \mathbb{P}_{k,\xi,v}$ sera écrit de façon unique sous la forme

$$\sum_i \alpha_i \phi_i$$

où $\phi_1, \phi_2, \ldots$ sont linéairement indépendants (donc constituent une base pour $\mathbb{P}_{k,\xi,v}$) et les coefficients α_i sont déduits des informations sur la fonction f.

4.5 Représentation des fonctions polynômiales par morceaux par les B-splines

A travers ce paragraphe, on va définir des fonctions qui, formant une base pour $\mathbb{P}_{k,\xi,v}$, permettront de représenter tout élément $g \in \mathbb{P}_{k,\xi,v}$.

Définition 45 *Soit $t = (t_i)$ une séquence de points non décroissante. La $i^{ème}$ B-spline (normalisée) d'ordre k pour la séquence de nœuds t est notée $B_{i,k,t}$ et est définie par*

$$B_{i,k,t}(x) = (t_{i+k} - t_i)\,[t_i, \ldots, t_{i+k}]\,(. - x)_+^{k-1} \quad \forall x \in \Re$$

On utilisera souvent la notation B_i au lieu de $B_{i,k,t}$.

Ces fonctions possèdent certaines propriétés :

i. La i^{ème} B-spline B_i a un petit support, i.e.,

$$B_i(x) = 0 \quad \text{pour } x \notin [t_i, t_{i+k}]$$

En effet, si $x \notin [t_i, t_{i+k}]$, alors $g(t) = (t - x)_+^{k-1}$ est un polynôme de degré $< k$ sur $[t_i, t_{i+k}]$ et par conséquent $[t_i, \ldots, t_{i+k}]\, g = 0$.

ii. On a

$$\sum_i B_i(x) = \sum_{i=r+1-k}^{s-1} B_i(x) = 1 \quad \text{pour tout } t_r < x < t_s$$

iii. Chaque B_i est positive sur son support, i.e., $B_i(x) > 0$ pour $t_i < x < t_{i+k}$.

Exemple. Une séquence de B-splines paraboliques ($k = 3$). La figure ci-dessous indique les cinq B-splines pour la séquence de nœuds $t = [0, 1, 1, 3, 4, 6, 6, 6]$.

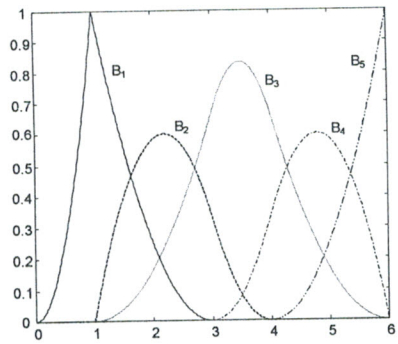


Figure 4.15. Une séquence de B-splines.

L'ensemble de ces fonctions constitue effectivement une base pour l'espace $\mathbb{P}_{k,\xi,v}$ et permettra d'élargir le concept des splines. On a

Définition 46 *Une spline d'ordre k pour la séquence de nœuds t est une combinaison linéaire de B-splines d'ordre k pour la séquence t.*

L'ensemble de toutes ces fonctions sera noté

$$\$_{k,t} = \left\{ \sum_i \alpha_i B_{i,k,t} : \alpha_i \text{ réel, tout } i \right\} = \mathbb{P}_{k,\xi,v}$$

La valeur d'une fonction g en un point $x \in [t_k, t_{n+1}]$ sera donnée par

$$g(x) = \sum_{i=1}^{n} \alpha_i B_i(x)$$

En particulier si $t_j \leq x \leq t_{j+1}$, $j \in [k, n]$, alors

$$g(x) = \sum_{i=j-k+1}^{j} \alpha_i B_i(x)$$

On revient maintenant à notre problème de dérivation numérique. A travers le paragraphe suivant on établira une relation qui permet d'obtenir la j$^{\text{ème}}$ dérivée d'une fonction polynômiale par morceaux définie dans une base de B-splines.

La première dérivée de la fonction $g(x) = (t-x)_+^{k-1}$ par rapport à x est donnée par

$$D_x(t-x)_+^{k-1} = -(k-1)(t-x)_+^{k-2}$$

Ceci permettra d'écrire

$$DB_{i,k}(x) = (k-1)\left(\frac{-B_{i+1,k-1}(x)}{t_{i+k} - t_{i+1}} + \frac{B_{i,k-1}(x)}{t_{i+k-1} - t_i}\right)$$

Si on s'intéresse plus particulièrement à l'intervalle $[t_r, t_s]$, on obtient alors

$$D\left(\sum_i \alpha_i B_{i,k}\right) = \sum_{i=r-k+2}^{s-1} (k-1)\frac{\alpha_i - \alpha_{i-1}}{t_{i+k-1} - t_i} B_{i,k-1}$$

Ainsi pour la j$^{\text{ème}}$ dérivée on a .

$$D^j\left(\sum_i \alpha_i B_{i,k}\right) = \sum_i \alpha_i^{(j+1)} B_{i,k-j}$$

avec

$$\alpha_r^{(j+1)} = \begin{cases} \alpha_r & \text{si } j = 0 \\ \dfrac{\alpha_r^{(j)} - \alpha_{r-1}^{(j)}}{(t_{r+k-j} - t_r)/(k-j)} & \text{si } j > 0 \end{cases}$$

4.6 L'interpolation spline

On discutera dans cette partie de l'interpolation par des splines d'ordre arbitraire. Soit à nouveau la séquence non décroissante de nœuds $t = (t_i)_1^{n+k}$ avec $t_i < t_{i+k}$ tout i, et $(B_i)_1^n$ la séquence correspondante de B-splines d'ordre k. La séquence strictement croissante $\tau = (\tau_i)_1^n$ représente les points d'interpolation. Alors pour une fonction f donnée la spline $g = \sum_i^n \alpha_i B_i$ coïncide avec f si et seulement si

$$\sum_{j=1}^{n} \alpha_j B_j(\tau_i) = f(\tau_i) \tag{4.8}$$

On s'intéresse au cas garantissant l'unicité de la solution du système (4.8). Ce dernier admet une solution si et seulement si la matrice des coefficients $(B_j(\tau_i))$ est inversible. On a alors le théorème suivant :

Théorème 47 *Soit τ une séquence strictement croissante telle que $t_i = \cdots = t_{i+r} = \tau_j$, $r < k$. Alors la matrice $(B_j(\tau_i))$ du système linéaire (4.8) est inversible si et seulement si*

$$B_i(\tau_i) \neq 0 \quad i = 1, \ldots, n$$

i.e., si et seulement si $t_i < \tau_i < t_{i+k}$, tout i.

Le problème maintenant est le suivant : étant donné un ensemble de points d'interpolation $(\tau_i)_1^n$, comment peut-on construire une séquence de nœuds $(t_i)_1^{n+k}$?

Soit

$$\|f - S_f\| \leq const_S \left\|f^{(k)}\right\|$$

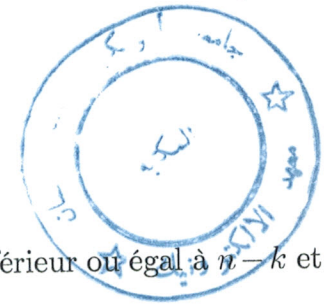
une estimation de l'erreur d'un schéma d'approximation S.

Micchelli, Rivlin et Winograd [7] ont déterminé une méthode permettant de construire une séquence de nœuds telle que $const_S$ soit la plus petite possible. Posons

$$[a, b] = [\tau_1, \tau_n]$$

On choisira $t_1 = \cdots = t_k = a$, $t_{n+1} = \cdots = t_{n+k} = b$ et les $n-k$ points restants $t_{k+1}, \ldots, t_n \in [a,b]$ formeront des points de cassure pour la fonction unité h telle que

$$\begin{aligned} |h(x)| &= 1 \ \forall x \in [a,b] \\ h(a^+) &= 1 \end{aligned}$$



Dans l'intervalle $[a,b]$, le nombre de changements de signe est inférieur ou égal à $n-k$ et

$$\int_a^b g(x)h(x)dx = 0 \qquad \forall g \in \$_{k,\tau}$$

Pour τ donnée, la séquence t est déterminée par la méthode de Newton [7].

4.6.1 L'erreur d'interpolation

Soit $\|f\| = \max_{a \le x \le b} |f(x)|$, $[a,b]$ étant un intervalle contenant les points d'interpolation $\tau_1, \ldots, \tau_n$. Considérons la séquence de nœuds $t = (t_i)_1^{n+k}$ avec $a = t_1 = \cdots = t_k < t_{k+1} \le \cdots \le t_n < t_{n+1} = \cdots = t_{n+k} = b$. On supposera $t_i < \tau_i < t_{i+k}$, $i = 1, \ldots, n$, et notons la fonction spline $g = \sum_i \alpha_i B_i$ par I_f.

Lemme 48 *Pour toute fonction f continue sur $[a,b]$, l'erreur d'interpolation est*

$$\|f - I_f\| \le (1 + \|I\|) dist(f, \$_{k,t}) \tag{4.9}$$

avec : $dist(f, \$_{k,t}) = \min \{\|f-h\|, h \in \$_{k,t}\}$, $\|I\| = \max \{\|I_f\| / \|f\| : f \in C[a,b] \setminus \{0\}\}$.

Bien sûr, une bonne approximation est obtenue si cette erreur est faible. On s'intéressera, alors, aux quantités $\text{dist}(f, \$_{k,t})$ et $(1 + \|I\|)$.

$\text{dist}(f, \$_{k,t})$ détermine la qualité avec laquelle une fonction d'une certaine régularité peut être approchée par des splines d'ordre donné. Définissons le module de continuité par

$$\omega(f,h) \triangleq \max \{|f(x) - f(y)| : |x-y| \le h, x, y \in [a,b]\}$$

et énonçons le théorème suivant

Théorème 49 *Pour $j = 0, \ldots, k-1$, il existe $const_{k,j}$ (une constante dépendant de k et j) telle que, pour tout $t = (t_i)_1^{n+k}$ avec*

$$a = t_1 = \cdots = t_k < t_{k+1} \leq \cdots \leq t_n < t_{n+1} = \cdots = t_{n+k} = b \tag{4.10}$$

et pour toute $f \in C^{(j)}[a,b]$

$$dist(f, \$_{k,t}) \leq const_{k,j} \, |t|^j \, \omega(f^{(j)}, h)$$

en particulier, pour $j = k-1$, on a

$$dist(f, \$_{k,t}) \leq const_{k,j} \, |t|^k \, \left\| f^{(k)} \right\| \tag{4.11}$$

dans le cas où f a k dérivées continues.

Ce théorème montre que la distance entre une fonction régulière et l'espace $\$_{k,t}$ tend vers zéro au moins aussi rapidement que $|t|^k$.

Le problème, maintenant, est comment choisit-on t satisfaisant (4.10), pour n fixé, tel que $\text{dist}(f, \$_{k,t})$ soit la plus petite possible? Il est certain qu'on ne peut pas placer chaque nœud de façon optimale mais on peut obtenir une distribution optimale de l'ensemble des nœuds. Ci-dessous, sera présentée brièvement une méthode [7] répondant à ce besoin.

La relation (4.11) permet d'exprimer l'erreur entre la fonction f et son interpolant spline d'ordre k, A_f, dans l'intervalle $[t_j, t_{j+1}]$ par

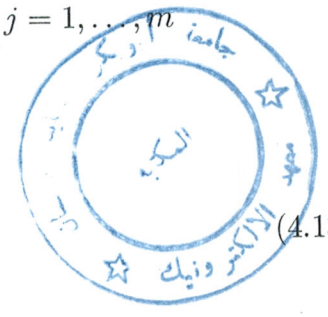
$$\|f - A_f\|_{[t_j, t_{j+1}]} \leq const_k \, |I_j|^k \, \left\| f^{(k)} \right\|_{I_j} \tag{4.12}$$

avec $I_j = [t_{j+2-k}, t_{j+k-1}]$ et $|I_j|$ sa longueur.

La technique proposée consiste à rassembler les nœuds $(t_i)_1^{n+k}$, dans l'intervalle $]a,b[$, en des groupes de $k-1$ nœuds chacun et faire de sorte que chaque groupe corresponde

à un nœud de multiplicité $k-1$. Si le nombre des nœuds n'est pas un multiple de $k-1$, alors on apportera d'autres. Soient $\xi_2 < \cdots < \xi_m$ des points distincts de l'intervalle $]a,b[$ et prenons $\xi_1 = a$, $\xi_{m+1} = b$, alors dans $[a,b]$, $\$_{k,t}$ coïncidera avec $\mathbb{P}_{k,\xi} \cap C[a,b]$. Dans ce cas (4.12) devient

$$\|f - A_f\|_{[\xi_j,\xi_{j+1}]} \leq const_k \left\|f^{(k)}\right\|_{[\xi_j,\xi_{j+1}]} |\Delta\xi_j|^k \qquad j = 1,\ldots,m$$



On désire placer $\xi_2,\ldots,\xi_m$ afin de minimiser

$$\max_j \left\|f^{(k)}\right\|_{[\xi_j,\xi_{j+1}]} |\Delta\xi_j|^k \tag{4.13}$$

Sachant que

$$s(\alpha,\beta) = \left\|f^{(k)}\right\|_{[\alpha,\beta]} |\beta - \alpha|^k$$

est une fonction continue en α et β(si $f^{(k)}$ est continue) et monotone, croissante en β et décroissante en α, alors pour m fixé, (4.13) est minimisée si $\xi_2,\ldots,\xi_m$ sont choisis tels que

$$\left\|f^{(k)}\right\|_{[\xi_j,\xi_{j+1}]} |\Delta\xi_j|^k = \text{constante} \qquad j = 1,\ldots,m$$

Ceci est équivalent à déterminer $\xi_2,\ldots,\xi_m$ tels que

$$\left(\left\|f^{(k)}\right\|_{[\xi_j,\xi_{j+1}]}\right)^{1/k} \Delta\xi_j = \text{constante} \qquad j = 1,\ldots,m \tag{4.14}$$

(4.14) donne, asymptotiquement, la même distribution des ξ_j que celle fournie par le problème de déterminer $\xi_2,\ldots,\xi_m$ tels que

$$\int_{\xi_j}^{\xi_{j+1}} \left|f^{(k)}(x)\right|^{1/k} dx = \frac{1}{m} \int_a^b \left|f^{(k)}(x)\right|^{1/k} dx \qquad j = 1,\ldots,m$$

Ce problème devient simple à résoudre si on remplace $\left|f^{(k)}\right|$ par une fonction constante par morceaux

$$h \sim \left|f^{(k)}\right|$$

On a alors

$$F(x) = \int_a^x (h(s))^{1/k}\, ds$$

Soit, maintenant, g une fonction polynômiale par morceaux interpolant f. Alors la valeur de la fonction h dans l'intervalle $[\xi_j, \xi_{j+1}]$ sera celle de la dérivée, au point $(\xi_{i-1} + 3\xi_i + 3\xi_{i+1} + \xi_{i+2})/8$, de la parabole interpolant la fonction

$$\int_a^x \left|g^{(k)}(s)\right| ds$$

aux points $\xi_{i-1/2}$, $\xi_{i+1/2}$, $\xi_{i+3/2}$. (on a $\tau_{i+1/2} = (\tau_i + \tau_{i+1})/2$)

Remarque 50 *On note que si une première approximation ne donne pas des résultats satisfaisants, alors quelques itérations de l'algorithme présenté ci-dessus peuvent apporter des améliorations. A chaque itération, la fonction à considérer est l'interpolant résultant de l'itération précédente.*

Passons maintenant à la quantité $\|I\|$ et pour laquelle on se contente de déterminer une borne inférieure, d'où

Lemme 51 *Il existe une constante positive $const_k$ telle que la norme $\|I\|$ soit bornée inférieurement, i.e.,*

$$\|I\| \geq const_k \max_i \frac{\min\{t_{j+k-1} - t_j : (t_j, t_{j+k-1}) \cap (\tau_i, \tau_{i+1}) \neq \emptyset\}}{\Delta\tau_i}$$

On remarque d'après ce lemme que si deux points d'interpolation sont proches l'un de l'autre alors $\|I\|$ devient arbitrairement grande.

Tous les résultats présentés ci-dessus supposent que les valeurs de la fonction à interpoler sont précises. Mais que devient ce type d'approximation si les données sont bruitées?

4.7 Cas de données bruitées

Malheureusement, une bonne approximation ne peut être obtenue si les valeurs disponibles de la fonction ne sont pas très précises.

Supposons que pour des points de données $\tau = (\tau_i)_1^N$ on dispose des valeurs f_i d'une fonction f telles que

$$f_i = f(\tau_i) + \delta_i$$

δ_i est la variation commise sur f_i.

L'approximation au sens des moindres carrés est une technique très convenable pour la reconstitution d'une fonction régulière à partir de données bruitées. Pour ce fait, on va définir une norme déduite à partir du produit scalaire. On a

$$\langle f, h \rangle = \sum_{i=1}^{N} f(\tau_i) h(\tau_i) w_i$$

On supposera la séquence $\tau = (\tau_i)_1^n \in [a, b]$ non décroissante. Les poids w_i, $i = 1, \ldots, N$, peuvent être choisis tels que

$$w_i = \begin{cases} \Delta\tau_i/2 & i = 1 \\ (\Delta\tau_{i-1} + \Delta\tau_i)/2 & i = 2, \ldots, N-1 \\ \Delta\tau_{N-1}/2 & i = N \end{cases}$$

La norme induite par le produit scalaire $\langle ., . \rangle$ sera notée par

$$\|h\|_2 = \langle h, h \rangle^{1/2}$$

Soit \$ l'espace linéaire de, dimension finie, de toutes les fonctions définies sur $[a, b]$. On cherche une meilleure approximation de \$ pour f par rapport à la norme $\|.\|_2$, i.e., une fonction $g^* \in \$$ telle que

$$\|f - g^*\| = \min_{g \in \$} \|f - g\|_2 \tag{4.15}$$

La dimension finie de \$ garantie l'existence de g^*.

Lemme 52 *La fonction g^* est une meilleure approximation de f dans \$ par rapport à $\|.\|_2$ si et seulement si $g^* \in \$$ et la fonction $f - g^*$, i.e., l'erreur, est orthogonale à \$, i.e., pour toute $g \in \$$*

$$\langle g, f - g^* \rangle = 0$$

Ceci implique que f a une meilleure approximation et une seule dans \$ si et seulement si $\|.\|_2$ est une norme sur \$, i.e., si l'unique fonction $g \in \$$ vérifiant $\|g\|_2 = 0$ n'est autre que $g = 0$.

Supposons, en effet, que $\|.\|_2$ est une norme sur \$ et que $(\phi_i)_1^n$ est une base pour \$. Donc (4.15) a une solution unique $g^* \in \$$ et elle est équivalente à

$$\langle \phi_i, f - g^* \rangle = 0 \qquad i = 1, \ldots, n$$

Aussi, g^* a une représentation unique $\sum_1^n \alpha_i \phi_i$ dans la base $\phi = (\phi_i)_1^n$. Ainsi, le système linéaire

$$\left\langle \phi_i, f - \sum_{j=1}^{n} \alpha_j \phi_j \right\rangle = 0 \qquad i = 1, \ldots, n \tag{4.16}$$

a une solution unique (α_i).

L'équation (4.16) est équivalente à

$$\sum_{j=1}^{n} \langle \phi_i, \phi_j \rangle \alpha_j = \langle \phi_i, f \rangle \qquad i = 1, \ldots, n$$

Si on choisit les (ϕ_i) telles que

$$\langle \phi_i, \phi_j \rangle = 0 \quad \text{pour } i \neq j$$

alors la solution est donnée par

$$\alpha_j = \langle \phi_j, f \rangle \,/\, \langle \phi_j, \phi_j \rangle$$

On peut éviter le problème du mauvais conditionnement de la base (ϕ_i) si on choisit $(\phi_i)_1^n = (B_i)_1^n$ (une base de B-splines).

Prenons $\$ = \$_{k,t}$ avec $t = (t_i)_1^n$, on obtient donc

$$\sum_{j=1}^{n} \langle B_i, B_j \rangle \alpha_j = \langle B_i, f \rangle \qquad i = 1, \ldots, n \tag{4.17}$$

C'est un système linéaire qui peut être résolu par la méthode de Gauss sans pivotation.

L'unicité de la solution de (4.17) est garantie si notre norme est bien une norme sur $\$_{k,t}$. Le théorème (47) donne

Lemme 53 *La norme* $\|g\|_2 = \left(\sum_i w_i (g(\tau_i))^2\right)^{1/2}$ *avec* $w_i > 0$, *tout* i, *et* (τ_i) *non décroissante, est une norme sur* $\$_{k,t}$ *si et seulement si pour* $1 \leq j_1 < \cdots < j_n \leq N$

$$t_i < \tau_{j_i} < t_{i+k} \quad i = 1, \ldots, n$$

4.7.1 Exemple

Soit à interpoler les valeurs $f(\tau_i)$ de la fonction $f(x) = \sin(x)$, $x \in [0, 2\pi]$, noyées dans un bruit. Ce dernier est représenté par une séquence de nombres aléatoire, soit

$$\begin{aligned} f_i &= f(\tau_i) + \delta_i \\ \tau_i &= 2\pi(i-1)/14 \quad i = 1, \ldots, 15 \end{aligned}$$

On désire déterminer les trois premières dérivées de $f(x)$ à partir de l'interpolation des valeurs f_i.

La figure ci-dessous montre l'allure du signal bruité.

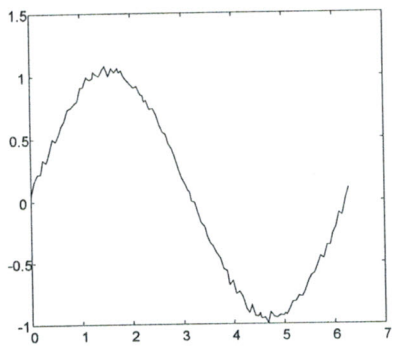


Figure 4.16. Signal bruité.

Les 4 figures ci-dessous représentent l'interpolant de la fonction $f(x)$, qu'on notera $I_f(x)$, ainsi que ses trois premières dérivées, respectivement. L'interpolation utilisée étant du type spline(l'ordre $k = 6$) au sens des moindres carrés.

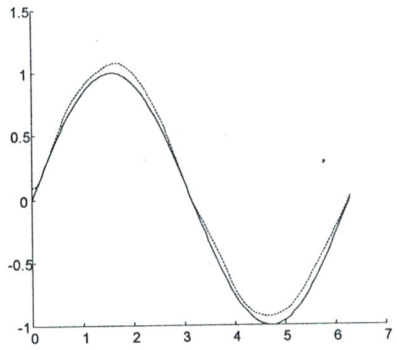


Figure 4.17. f et I_f.

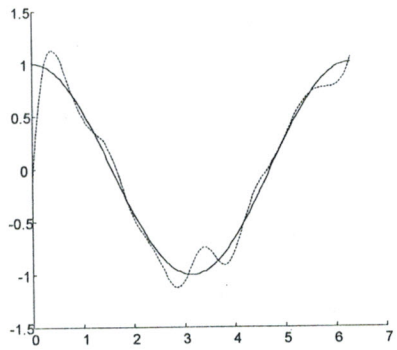


Figure 4.18. f' et $(I_f)'$.

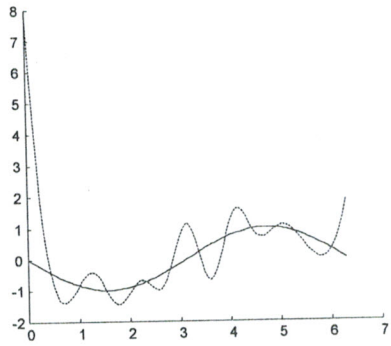


Figure 4.19. f" et (I_f)".

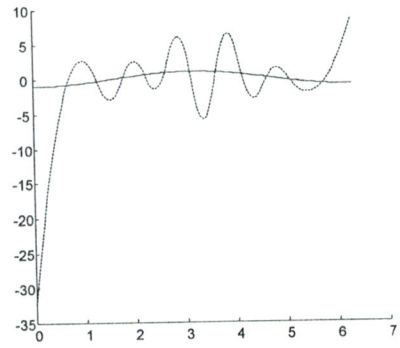


Figure 4.20. $f^{(3)}$ et $(I_f)^{(3)}$.

La ligne continue indique le signal net, la discontinue le signal estimé par les moindres carrés.

On remarque bien que plus l'ordre de dérivation augmente, plus la qualité de l'approximation se dégrade. Pour remédier à ce problème ou au moins alléger cette dégradation, on procédera à un placement optimal des nœuds considérés initialement. Quatre itérations de l'algorithme décrit ci-dessus permettent d'obtenir les résultats suivants

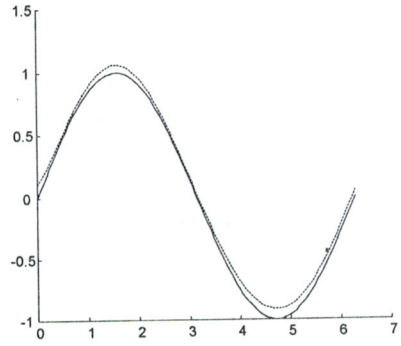


Figure 4.21. f et I_f.

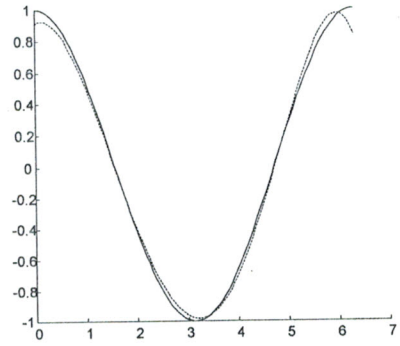


Figure 4.22. f' et (I_f)'.

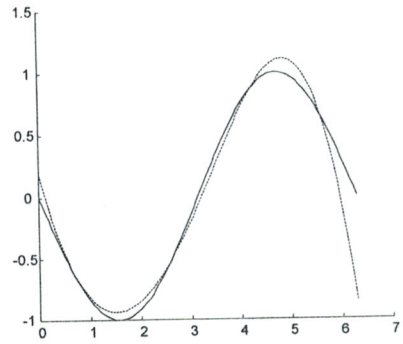


Figure 4.23. f" et (I_f)".

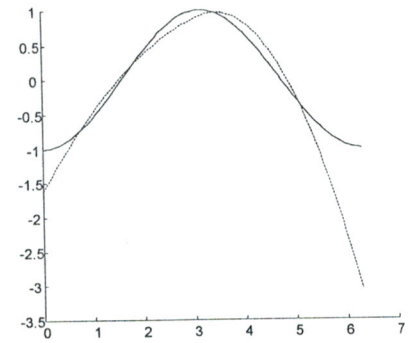


Figure 4.24. $f^{(3)}$ et $(I_f)^{(3)}$.

Ces figures montrent que l'approximation s'est considérablement améliorée.

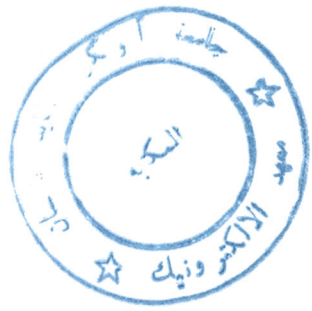
A travers un exemple du chapitre suivant, on mettra en œuvre cette méthode, i.e., des splines au sens des moindres carrés plus un placement optimal des nœuds, pour l'estimation des dérivées entrant dans l'expression de la commande d'un système non linéaire.

4.8 Conclusion

Pour faciliter la compréhension de certains concepts, on a commencé ce chapitre par les splines linéaires. On a également souligné les splines cubiques car, comme c'est déjà mentionné, elles minimisent l'énergie de déformation de la latte du dessinateur. Dans les problèmes d'approximation, cette propriété est dite de meilleure approximation et elle nous a permis d'établir certains résultats concernant les erreurs de ce type de splines.

Vu la nécessité de dériver une fonction jusqu'à un ordre supérieur à 3, nous étions amenés à introduire le concept de B-spline. Partant d'un certain nombre de propriétés, on a constaté q'une combinaison linéaire de ces fonctions permettait de définir une spline d'ordre k. Evidemment, comme il est toujours le cas dans les problèmes d'interpolation, il est intéressant d'évaluer l'erreur de la méthode. Dans ce cadre on a discuté brièvement les points garantissant la minimalité de cette grandeur. Enfin, pour s'approcher plus de la réalité, on a considéré une situation fréquente : le cas où on dispose de valeurs imprécises de la fonction à interpoler. On a présenté, en l'occurence, la technique la plus conseillée

dans ce genre de problèmes : l'approximation au sens des moindres carrés mais dans un contexte de splines.



Chapitre 5

L'observateur par interpolation et dérivation numérique

5.1 Introduction

La synthèse d'observateurs est un problème fondamental dans la théorie du contrôle et ses applications. En effet, un grand nombre de techniques de commande des systèmes non linéaires, en particulier, suppose la totale disponibilité du vecteur d'état pour l'élaboration d'une loi de commande donnée. Le problème exposé ici considère la situation où on cherche à estimer les variables d'un système continu à partir de données observées à des instants discrets. L'idée de l'observateur par interpolation et dérivation numérique [8] est fondée sur un point essentiel dans la théorie du système : obtenir un certain nombre de dérivées de la sortie du système défini par des équations différentielles ou à travers les valeurs qu'il prend aux instants d'échantillonage. Ceci peut être illustré à travers l'exemple suivant. Un système non linéaire est donné par

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = \varphi(x, u) \\ y = x_1 \end{cases} \tag{5.1}$$

le but est de construire un observateur pour (5.1) sur la base des mesures disponibles. Il est clair qu'on n'a pas besoin d'estimer x_1 puisqu'il est déduit directement de la mesure : $x_1 = y$. On remarque également que $x_2 = \dot{y}$. Si, comme il est souvent le cas, on ne peut pas accéder à $\dot{y}$, alors il faut calculer $\dot{y}$ à partir de la mesure. C'est ici où intervient la dérivation numérique. Dans le cas où y est connue à travers des échantillons, le procédé de dérivation numérique devra fournir une approximation de $\dot{y}$ à un instant t_k soit $\widehat{\dot{y}}(t_k)$, d'où la nécessité de mettre au point un algorithme calculant le nombre $\widehat{\dot{y}}(t_k)$ et donnant une estimation de l'erreur d'approximation. Souvent, non seulement $\dot{y}$ mais un nombre fini de dérivées supérieures de y est utile pour la détermination des états du système. Sur le plan théorique, on peut trouver un nombre considérable de méthodes d'approximation de fonctions. Les plus connues sont celles basées sur les approximants polynômiaux. Dans notre travail on s'intéressera aux fonctions splines, car comme on l'a vu précédemment, elles possédant des propriétés intéressantes.

A travers ce chapitre, on va présenter dans un premier lieu la notion d'observabilité adoptée dans notre travail. La partie suivante contient une description de l'algorithme d'observation et de commande utilisé dans les différentes applications. Ils seront, ensuite, mis en œuvre à travers deux exemples académiques. Suivra, ensuite une autre partie dans laquelle certaines méthodes soulignées dans le chapitre précédent seront exploitées pour appliquer le processus d'interpolation et de dérivation numérique dans le cas où les mesures sont bruitées. Enfin, dans un dernier point, on comparera les résultats obtenus par dérivation de l'interpolant spline, avec ceux que fournissent deux autres méthodes de dérivation numérique.

Dans le but de synthétiser un observateur il faut s'assurer, tout d'abord, de l'observabilité du système d'intérêt.

5.2 Définition de l'observabilité

Considérons un système décrit par

$$\begin{cases} \dot{x} = f(x,u) \\ y = h(x,u) \end{cases} \tag{5.2}$$

$x \in \Re^n$, $y \in \Re^p$.

Le concept d'observabilité adopté [8] pour notre problème est donné par

Définition 54 *Le système (5.2) est observable s'il existe un entier N tel que l'application $H(u, \dot{u}, \ldots, u^{(N-2)}, .)$ définie par*

$$H(u, \dot{u}, \ldots, u^{(N-1)}, x) = \begin{pmatrix} y \\ \dot{y} \\ \vdots \\ y^{(N-1)} \end{pmatrix}$$

soit injective pour toute entrée universelle fixée.

Pour un système (5.2) observable, on peut écrire

$$x = L(u, \dot{u}, \ldots, u^{(N-2)}, y, \dot{y}, \ldots, y^{(N-1)})$$

L'existence de l'application L est garantie par notre définition de l'observabilité [8]. Pour de tels systèmes, le problème de synthèse d'observateur peut être vu comme un problème de dérivation numérique : une fois les estimées des dérivées de y et de u déterminées des mesures disponibles, alors x peut être déduit de L. Il est parfois impossible de déterminer une expression explicite de la fonction L, on utilise donc une méthode itérative telle que la méthode de Newton pour en déduire une.

5.3 Algorithme d'observation

Le but de notre travail est de commander des systèmes non linéaires à travers des lois exprimées par un certain nombre d'états du système.

Notre observateur a la sortie (la mesure) du système considéré comme entrée. A l'instant initial $t = 0$, le système est soumis à l'action d'une commande initiale . Sur chaque fenêtre $I_k = [t_k, t_k + T]$, $T > 0$ (instant d'échantillonnage), les mesures prélevées aux instants de suréchantillonnage $t_k + n\frac{T}{l}$, $l \in \mathbb{N}_+^*, n = 0, \ldots, l$, sont interpolées par des fonctions splines. L'interpolant résultant est donc dérivé jusqu'à un ordre approprié. Ainsi, à l'instant $t_k + T$, l'observateur doit fournir des estimations des dérivées de la sortie. En même temps, la commande doit accomplir deux tâches :

- à un instant t_k, une commande u_{k+1} est calculée au moyen des états estimés $\hat{x}_k$ est gardée constante sur l'intervalle I_k,
- à l'instant $t_{k+1} = t_k + T$, une nouvelle observation $\hat{x}_{k+1}$ est déterminée utilisant la commande u_{k+1} et la sortie et ses dérivées estimées à t_{k+1}.

5.4 Exemples

Dans ce paragraphe, on va illustrer l'algorithme cité ci-dessus à travers deux exemples.

Exemple 1

Les équations d'un pendule inversé sont données par

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -\sin(x_1) + u \\ y = x_2 \end{cases} \tag{5.3}$$

Le but est de stabiliser (5.3) par la commande

$$u = \sin x_1 - 0.25x_1 - x_2$$

L'état x_2 est disponible et on supposera qu'on ne peut pas accéder à x_1. On doit alors construire un observateur donnant une estimation de l'état x_1 utilisant la mesure et l'entrée.

A partir de (5.3), on a

$$\begin{cases} x_2 = y \\ x_1 = \arcsin(\tilde{u} - \dot{\hat{y}}) \\ \tilde{u} = \sin \hat{x}_1 - 0.25\hat{x}_1 - x_2 \end{cases}$$

L'état initial est $\begin{pmatrix} x_{10} \\ x_{20} \end{pmatrix} = \begin{pmatrix} \pi/2 \\ 0 \end{pmatrix}$ et $u(t=0) = 0$, $T = 0.1s$, $l = 10$.
Les figures ci-dessous représentent les états du système commandé par $\tilde{u}$..

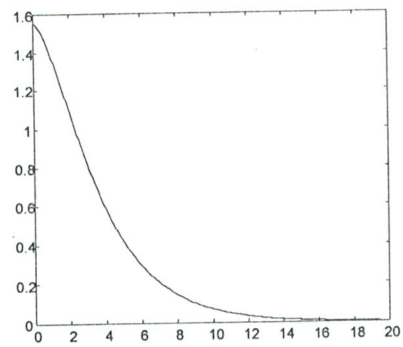


Figure 5.1. x_1.

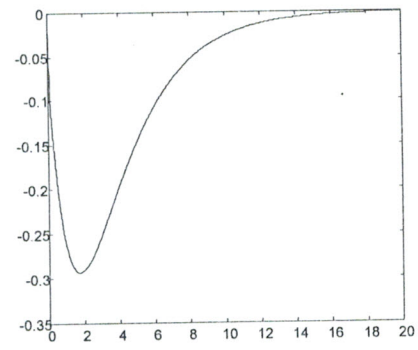


Figure 5.2. x_2.

Les erreurs d'estimation $e_1 = x_1 - \hat{x}_1$, $e_2 = x_2 - \hat{x}_2$ aux instants d'échantillonnage

sont illustrées sur les figures (5.3.) et (5.4.).

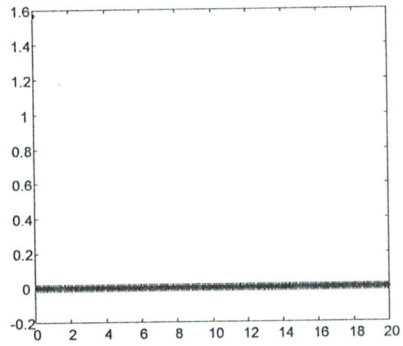


Figure 5.3. e_1.

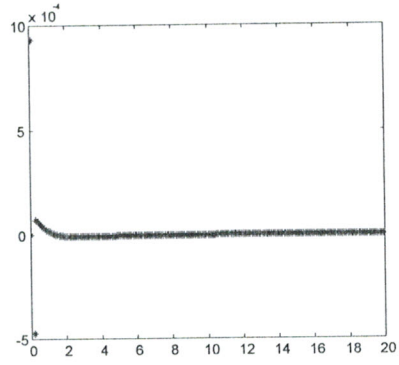


Figure 5.4. e_2.

D'après ces figures on peut conclure que le processus d'observation est performant vu les valeurs faibles des erreurs d'estimation. Aussi, la commande calculée est efficace puisqu'on arrive à stabiliser le système (5.3). Il est important de noter que la convergence de l'erreur vers zéro n'est pas influencée par le choix de la commande initiale.

Pour le système considéré, on a synthétisé un observateur à grand gain. Dans une première étape, on a mis les équations du système sous une forme canonique d'observation. Le système (5.3) devient donc

$$\dot{\xi} = A\xi + \varphi(\xi, u)$$
$$y = C\xi$$

où : $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\varphi(\xi, u) = \begin{bmatrix} u \\ -\xi_1\sqrt{1-\xi_2^2} \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 \end{bmatrix}$

La fonction φ vérifie les conditions de synthèse d'un observateur à grand gain indiquées dans le chapitre 2.

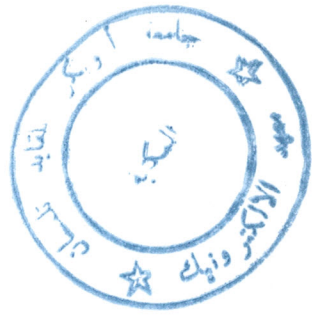
Donc, pour ce système, l'observateur est donné par

$$\dot{\hat{\xi}} = A\hat{\xi} + \varphi(\hat{\xi}, u) + \Lambda^{-1}(\tau)K(y - C\hat{\xi})$$

avec : $\Lambda(\tau) = \begin{bmatrix} \tau & 0 \\ 0 & \tau^2 \end{bmatrix}$ et K est choisi tel que $VP(A - KC) \in C^-$. Ainsi l'observa-

teur dans l'espace ξ est

$$\dot{\hat{\xi}}_1 = \hat{\xi}_2 + u + \frac{3}{\tau}(y - \hat{\xi}_1)$$
$$\dot{\hat{\xi}}_2 = -\hat{\xi}_1\sqrt{1 - \hat{\xi}_2^2} + \frac{2}{\tau^2}(y - \hat{\xi}_1)$$



Les états x_1 et x_2 sont déduits de

$$\hat{x}_2 = \hat{\xi}_1$$
$$\hat{x}_1 = -\arcsin(\hat{\xi}_2)$$

Pour être en la mesure de comparer les résultats des deux méthodes, on a procédé à une discrétisation de l'observateur continu. Pour $T = 0.1s$, $\tau = 0.2$, on a obtenu les courbes illustrées sur les figures ci-dessous.

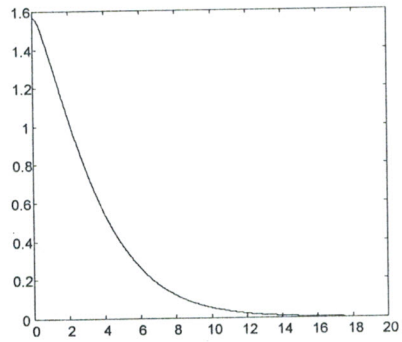


Figure 5.5. x_1.

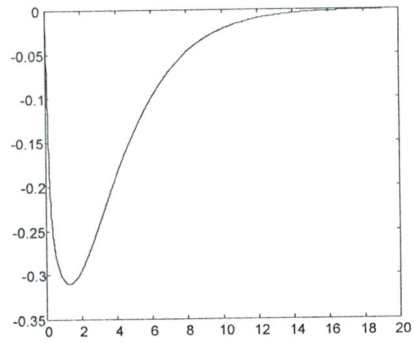


Figure 5.6. x_2.

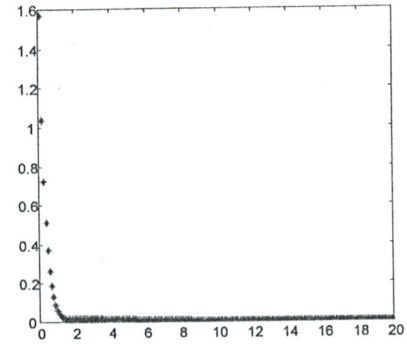


Figure 5.7. e_1.

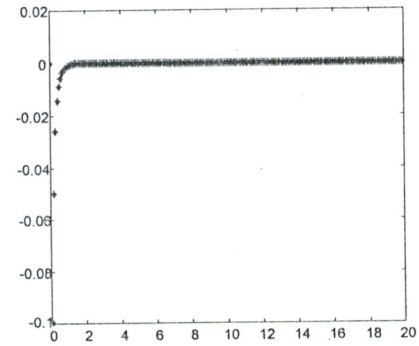


Figure 5.8. e_2.

Suite à ces résultats, on a remarqué que les erreurs d'estimation de l'observateur à

grand gain sont du même ordre de grandeur que celles de l'observateur par interpolation et dérivation numérique. On constate, également, que par le premier observateur, les deux états x_1 et x_2 convergent vers zéro dans un temps un peu plus court que par le deuxième.

Exemple2

Soit le système non linéaire suivant :

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1^3 + u \\ y = x_2 \end{cases}$$

A travers cet exemple, notre méthode est exploitée pour réaliser un suivi de trajectoire par l'état x_1. Le signal désiré est $x_d(t) = \sin(\omega t)$, $\omega = 1rd/s$. La condition initiale est $x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $T = 0.1$, $l = 10$ et $u(t = 0) = 0$.

La commande u est donnée par

$$u(t) = \hat{x}_1^3(t) + \omega^2 \sin(\omega t) + k_1(\hat{x}_1(t) - \sin(\omega t)) + k_2(x_2(t) - \omega \cos(\omega t))$$

Les gains k_1 et k_2 sont choisis tels que $e_{d1}(t) = x_1(t) - x_d(t)$ tende vers zéro quand $t \to \infty$. On a pris $k_1 = -2$, $k_2 = -3$.

Utilisant le fait que $x_2 = y$, $x_1 = \sqrt[3]{u - \dot{y}}$, on a implémenté un observateur par interpolation et dérivation numérique pour estimer l'état x_1. Les erreurs de suivi de trajectoires $e_{d1} = x_1 - x_d$, $e_{d2} = x_2 - \dot{x}_d$ et d'estimation $e_1 = x_1 - \hat{x}_1$, $e_2 = x_2 - \hat{x}_2$ sont

représentées ci-dessous.

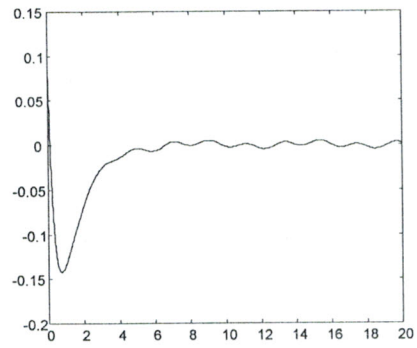


Figure 5.9. e_{d1}.

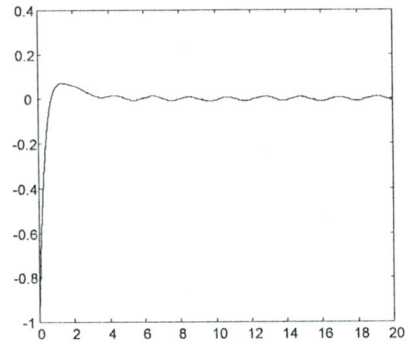


Figure 5.10. e_{d2}.

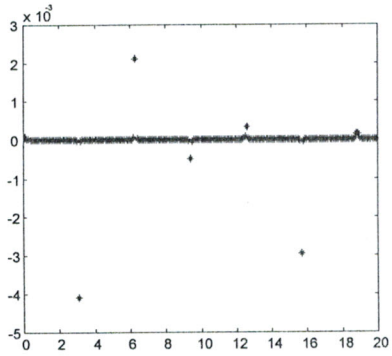


Figure 5.11. e_1.

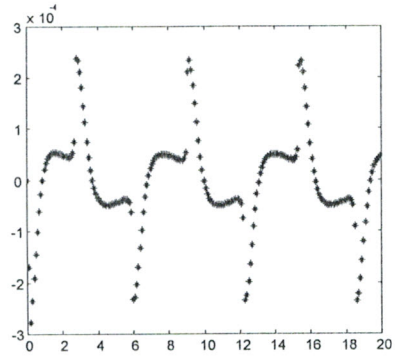


Figure 5.12. e_2.

Vu les faibles valeurs des erreurs de suivi de trajectoire, on peut conclure que la trajectoire désirée est bien réalisée. Ceci est bien confirmé par les résultats illustrés sur les figures (5.11.) et (5.12.), parcequ'une bonne estimation permet à la commande calculée d'être efficace.

On note que dans les deux cas l'observateur était initialisé en $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

5.5 Cas de mesures bruitées

On va supposer, maintenant, que les mesures disponibles sont entachées d'un bruit de capteur.

Dans un premier temps, on considérera un système linéaire non commandé, soit

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 \\ y = x_1 \end{cases}$$

Le but est d'utiliser l'interpolation et la dérivation numérique pour déterminer $\dot{y}$. Pour ce fait, on utilisera des splines au sens des moindres carrés ainsi que l'algorithme plaçant les nœuds optimalement. On choisira une période d'échantillonnage de $100ms$ et de suréchantillonnage de $10ms$. Sur la figure ci-dessous on peut visualiser, aux instants d'échantillonnage, les échantillons nets de la mesure(+), bruités (o) et estimés (*).

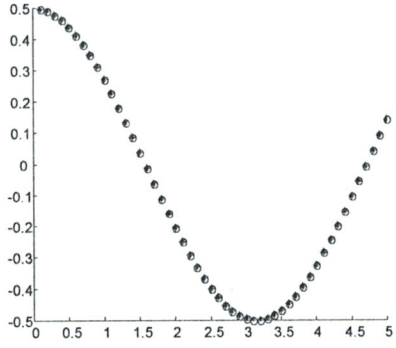


Figure 5.13. x_1.

On note que le signal bruité est

$$y(kT) = x_1(kT) + \delta$$

avec δ un bruit aléatoire tel que

$$\delta = 0.001 * rand(1, 10)$$

et $rand(1, 10)$ dix nombres aléatoires compris entre 0 et 1.

Sur cette figure, on a pu constater que l'estimation représente la moyenne du signal

bruité mais reste différente du signal propre, malgré la faible intensité du bruit. Ceci a une mauvaise conséquence sur l'estimation de la dérivée comme on peut le voir à travers les figures 5.14. et 5.15..

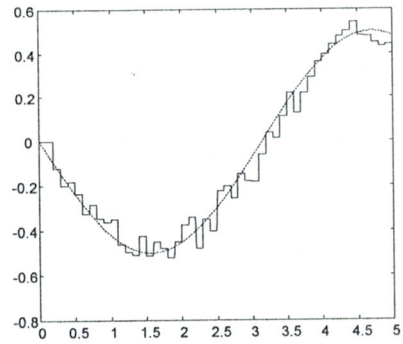


Figure 5.14. x_2 et $\hat{x}_2$.

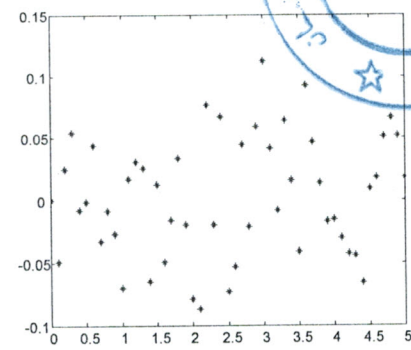


Figure 5.15. Erreur d'estimation.

Evidemment, les erreurs sur les dérivées supérieures seront plus importantes.

Considérons à nouveau le système non linéaire et commandé (5.3). Notre but, maintenant, est de le stabiliser tout en supposant que les mesures sont superposées à un bruit de capteur. Pour des périodes d'échantillonnage de $100ms$ et de suréchantillonnage de $10ms$, le comportement du système est comme l'indiquent les figures ci-dessous

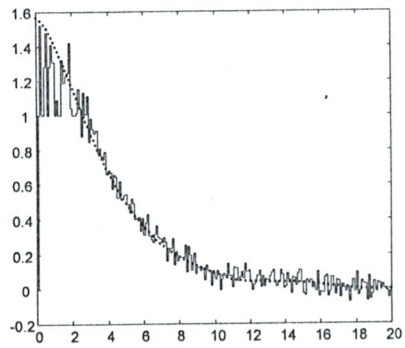


Figure 5.16. x_1.

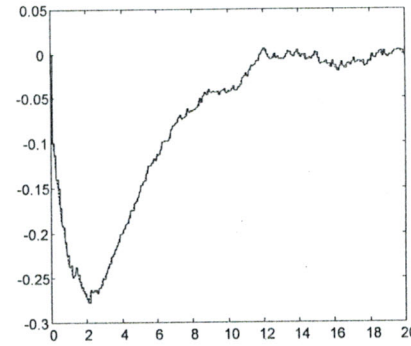


Figure 5.17. x_2.

La remarque essentielle qu'on puisse tirer de ces figures est que malgré la mauvaise estimation de la dérivée, intervenant dans l'expression de la commande, cette dernière arrive à stabiliser le système. Ceci est dû à un point important est que la boucle fermée permet de diminuer l'effet du bruit, i.e., le filtrer.

5.6 Autres méthodes de dérivation numérique

Cette partie a pour motivation de comparer les résultats fournis par trois méthodes de dérivation numérique. Pour des mesures prélevées aux instants de suréchantillonnage, on veut déterminer la première dérivée à chaque instant d'échantillonnage. Pour cela, on se servira des splines, de la dérivation numérique sur deux points et de cette dernière méthode mais après un filtrage des données.

On va considérer, à nouveau, le système

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -x_1 \\ y = x_1 \end{cases}$$

L'état x_1 est déduit directement de la mesure et il reste à déterminer $x_2 = \dot{y}$. Les périodes d'échantillonnage et de suréchantillonnage sont fixées à $100ms$ et $10ms$ respectivement.

En dérivant l'interpolant spline, l'erreur d'estimation, $e_s = x_2 - \hat{x}_2$, obtenue est illustrée sur la figure ci-dessous.

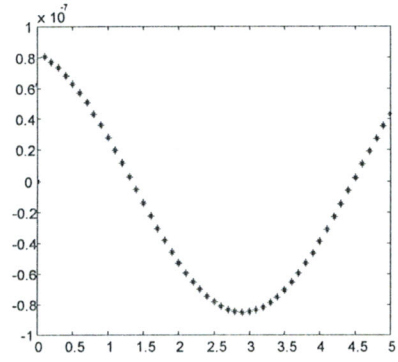


Figure 5.18. e_s.

Par dérivation numérique sur deux points, i.e., en utilisant

$$\dot{y}_{kT} = \frac{y_{(k+1)T} - y_{kT}}{T} \tag{5.4}$$

on obtient

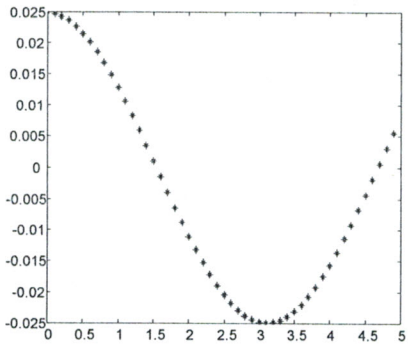


Figure 5.19. e_s.

Enfin, en utilisant la même formule (5.4), mais avec

$$y_{kT} = \sum_{j=1}^{10} b_j x_{1j}$$

les b_j sont les coefficients du filtre (FIR) et x_{1j} sont les mesures aux instants de suréchantillonnage, on obtient

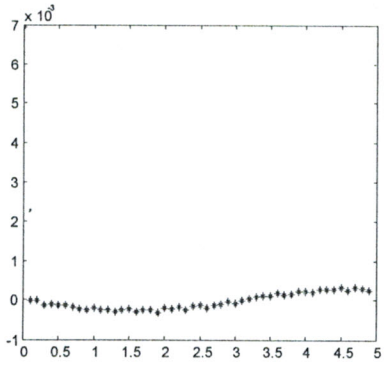


Figure 5.20. e_s.

On remarque, à travers les trois cas, que c'est par la dérivation de l'interpolant spline que la plus faible erreur est obtenue.

5.7 Conclusion

Dans ce chapitre, nous avons présenté l'observateur par interpolation et dérivation numérique. On a mis en évidence l'algorithme d'observation et de commande à travers deux exemples. Suite à la synthèse d'un observateur à grand gain, on a pu constater que le notre peut fonctionner aussi bien que le précédent. De plus il offre l'avantage de ne pas exiger la condition de lipschitzité. Une autre partie de ce chapitre a été consacrée à un problème délicat dans les processus physiques : c'est celui du filtrage du bruit. Malgré l'utilisation d'algorithmes censés en réduire l'effet, on a vu que les estimations délivrées par notre observateur étaient de mauvaise qualité. On a constaté, par contre, que la boucle fermée joue un rôle efficace dans l'atténuation de l'effet du bruit sur le système. Enfin, on a procédé à une comparaison de notre méthode de dérivation numérique avec deux autres. Il était bien clair, à travers l'exemple, que la dérivation de l'interpolant spline fournit la plus faible erreur d'estimation. La dérivation numérique sur deux points est moins efficace mais celle effectuée après un filtrage par un FIR permet d'améliorer nettement les résultats.

Chapitre 6

Applications

6.1 Introduction

L'objectif de l'automatique non linéaire est de résoudre des problèmes concrets de commande. Une de ses particularités est de s'appliquer à des domaines physiques très divers. C'est dans cet état d'esprit que sont exposés, dans ce chapitre, deux exemples d'observation et de commande de systèmes non linéaires.

L'application traitée appartient à un domaine privilégié de la théorie de la commande non linéaire, la robotique, et est représentative de l'applicabilité des résultats obtenus. Il s'agit de procédés issus d'un contexte industriel : le robot rigide et le robot à articulation flexible. On leur associera des modèles d'état non linéaires et leur commande nécessitera un observateur. En effet, pour ces systèmes, il est intéressant de minimiser le nombre de capteurs, afin de diminuer le coût de l'installation.

Pour la commande de tels systèmes, différentes classes d'observateurs ont été proposées. La sélection d'une méthode ou d'une autre dépend du choix des variables mesurées. Les lois de commande mises au point ces dernières années pour commander de tels robots utilisent souvent les positions et vitesses des bras et des axes moteurs, et parfois même les dérivées supérieures de ces grandeurs.

Dans notre travail, le schéma de commande proposé est basé sur un observateur par

interpolation et dérivation numérique plus une loi de commande découplante et linéarisante pour le suivi de trajectoire.

La répartition de ce chapitre sera faite comme suit : la première partie concernera le bras rigide. Dans un premier point, on présentera un modèle pour ce système. On discutera, ensuite, de son observabilité. La dernière partie inclut les résultats des différentes simulations ainsi que des commentaires. La même démarche sera suivie pour le bras à articulation flexible.

6.2 Le robot rigide

6.2.1 Modèle dynamique

Il représente la relation entre les couples appliqués aux actionneurs et les positions, vitesses et accélérations articulaires. C'est donc une relation de la forme

$$\Gamma = f(q, \dot{q}, \ddot{q})$$

Γ étant le vecteur des couples, q, $\dot{q}$, $\ddot{q}$ les vecteurs des positions, vitesses et accélérations articulaires [5]. Généralement, cette relation est donnée par

$$A(q)\ddot{q} + C(q, \dot{q})\dot{q} + F_v\dot{q} + F_s sign(\dot{q}) + G(q) = \Gamma$$

ou

$$A(q)\ddot{q} + H(q, \dot{q}) = \Gamma$$

avec

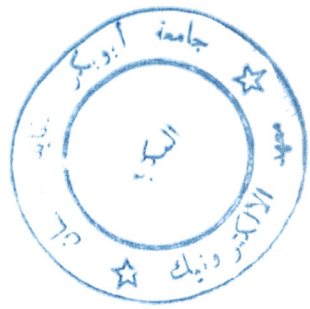
$A(q)$: matrice d'inertie du robot;

$C(q, \dot{q})$: forces centrifuges et de Coriolis;

F_v : vecteur des coefficients de frottement visqueux;

F_s : vecteur des coefficients de frottement secs;

$G(q)$: forces de gravité.



6.2.2 Observabilité du robot rigide

Si on pose $x_1 = q$ et $x_2 = \dot{q}$, le modèle du robot s'écrira alors

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= A^{-1}(x_1)(\Gamma - H(x_1, x_2)) \end{aligned}$$

Cette écriture peut être mise sous une autre forme, soit

$$\dot{x} = f(x) + g(x).u$$

avec : $f(x) = \begin{bmatrix} x_2 \\ -A^{-1}(x_1)H(x_1, x_2) \end{bmatrix}$, $g(x) = \begin{bmatrix} 0 \\ +A^{-1}(x_1) \end{bmatrix}$ et $u = \Gamma$.

Supposons que les positions articulaires sont mesurées, donc

$$y = h(x) = x_1$$

La dimension du système est $2n$, n étant le nombre de degré de liberté.

Ce système est uniformément obseravble car on peut déduire l'état de la sortie et de sa première dérivée. La condition de rang est vérifiée. En effet,

$$Rang\left(\frac{\partial}{\partial x}\left([h(x), L_f h(x)]^T\right)\right) = 2n$$

et par conséquent le système est localement faiblement observable.

Pour notre système, on prendra $n = 2$.

6.2.3 Etude en simulation

Le robot considéré est plan à deux axes rotoïdes (Figure 6.1.). Les actionneurs du robot sont des moteurs à courant continu et à aimants permanents [5].

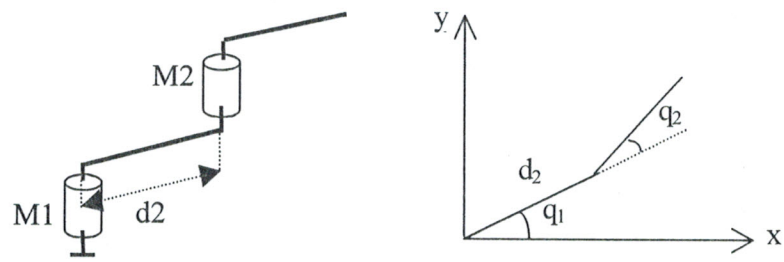


Figure 6.1. Le robot rigide.

Le modèle de ce robot est donné par

$$A(q)\ddot{q} + H(q,\dot{q}) = \Gamma$$

$A(q)$ (2×2) : matrice d'inertie, symétrique et définie positive;

$H(q,\dot{q})$ (2×1) : correspond aux forces centrifuges et de Coriolis ;

Γ : le couple moteur.

Les éléments des deux matrices sont

$A_{11} = zzr1 + zzr2 + 2mxr2 * d2 * \cos(q_2) + 2my2 * d2 * \sin(q_2)$

$A_{12} = A_{21} = zzr2 + mxr2 * d2 * \cos(q_2) + my2 * d2 * \sin(q_2)$

$A_{22} = zzr2$

$H_1 = (-mxr2 * d2 * \sin(q_2) + my2 * d2 * \cos(q_2)) * \dot{q}_2^2 - (2mxr2 * d2 * \sin(q_2))\dot{q}_1\dot{q}_2 + f_{v1} * \dot{q}_1 + f_{s1} * sign(\dot{q}_1)$

$H_2 = (mxr2 * d2 * \sin(q_2) - my2 * d2 * \cos(q_2)) * \dot{q}_1^2)) + f_{v2} * \dot{q}_2 + f_{s2} * sign(\dot{q}_2)$

avec : $zzr1 = zz1 + m_2 d_2^2 + I_{a_1}$, $zzr2 = zz2 + I_{a_2}$

$zz1$ et $zz2$ sont les moments d'inertie des deux bras, m_2 est la masse du bras 2, I_{a_1} et I_{a_2}sont les inerties des actionneurs, $mxr2$ et $my2$ sont les projections du moment du bras 2 selon les axes x et y du repère associé.

Les valeurs numériques ci-dessous correspondent aux paramètres du bras rigide construit au Laboratoire d'Automatique de Nantes [5].

$zzr1 = 3.7kg.m^2$, $zzr2 = 0.08kg.m^2$, $mxr2 = 0.284kgf.m^2$, $my2 = 0.022kgf.m^2$, $f_{v1} = 0.07Nm/(rd/s)$, $f_{v2} = 0.013Nm/(rd/s)$, $f_{s1} = 0.62Nm$, $f_{s2} = 0.17Nm$.

Dans le but d'implémenter notre observateur, on a réécrit le modèle du système sous la forme

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= f_1(x,\Gamma)\\ \dot{x}_3 &= x_4\\ \dot{x}_4 &= f_2(x,\Gamma)\end{aligned}$$

avec :

$$x = [q_1, \dot{q}_1, q_2, \dot{q}_2]^T$$

$$f_1(x,\Gamma) = \frac{1}{\det(A)}\left[A_{22}\left(\Gamma_1 - H_1\right) - A_{12}\left(\Gamma_2 - H_2\right)\right]$$

$$f_2(x,\Gamma) = \frac{1}{\det(A)}\left[-A_{21}\left(\Gamma_1 - H_1\right) + A_{11}\left(\Gamma_2 - H_2\right)\right]$$

Les variables mesurées sont les angles de rotation des deux bras $q_1$et q_2.

Le suivi de trajectoire sera accompli à l'aide d'une loi de commande découplante et linéarisante de la forme

$$\Gamma = A(q)w + H(q,\dot{q})$$

et w est donnée par

$$w = \ddot{q}_d - k_d(\dot{q} - \dot{q}_d) - k_p(q - q_d)$$

q_d étant la trajectoire désirée. L'observateur doit donc fournir, aux instants d'échantillonnage, des estimations des dérivées premières $\dot{q}_1$et $\dot{q}_2$.

Pour les deux bras on a pris $q_{1d} = q_{2d} = q_d = \sin(\omega t)$, $\omega = 2rd/s$. Les gains de la commande sont choisis de façon à avoir une boucle à amortissement critique, la bande passante du premier axe est fixée à $\omega_n = 5rd/s$ ce qui correspond à $k_p = 25$ et $k_d = 10$. Pour le deuxième axe $\omega_n = 25rd/s$ ce qui donne $k_p = 625$ et $k_d = 50$.

La période d'échantillonnage est fixée à $T = 1ms$, et de suréchantillonnage à $t_{sh} = 100\mu s$. Quant aux conditions initiales on a choisi $x = [0,2,0,2]^T$ pour le système et pour

l'observateur, i.e., l'erreur initiale d'estimation est nulle.

Passons maintenant aux résultats de simulation. Les couples moteurs sont représentés sur les figures ci-dessous.

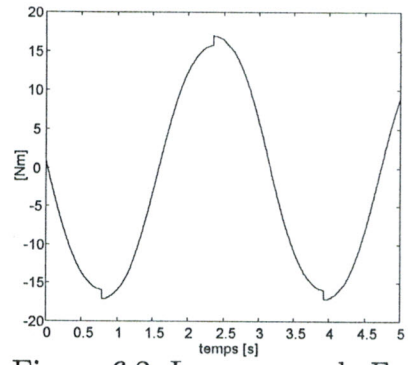


Figure 6.2. La commande Γ_1.

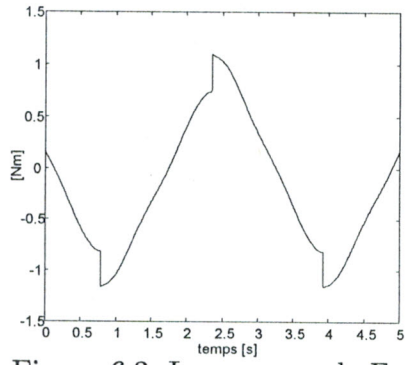


Figure 6.3. La commande Γ_2.

Ils sont compris entre $\pm 17Nm$ pour l'axe 1 et $\pm 1.18Nm$ pour l'axe 2, sachant que les valeurs maxima acceptables sont $18.23Nm$ et $10.3Nm$ pour les deux axes respectivement. La discontinuité qu'on peut remarquer au niveau des deux couples est due à la fonction 'signe' du modèle.

Sur les figures suivantes sont illustrées les positions et vitesses des deux axes. On montre également les erreurs de suivi de trajectoire qui, étant faibles, prouvent que les signaux désirés sont bien reproduits par le système.

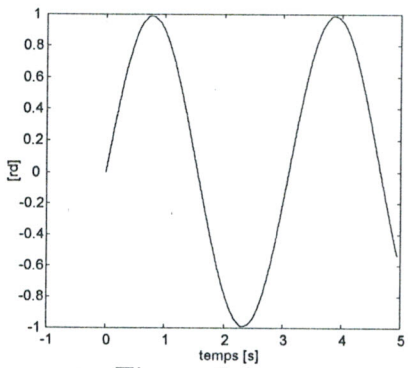


Figure 6.4. x_1.

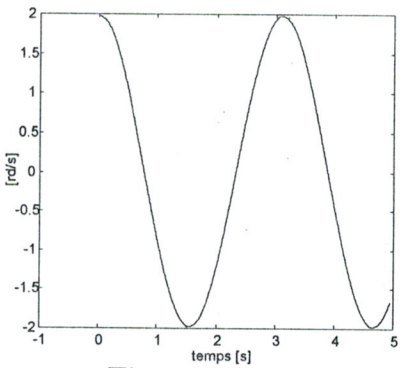


Figure 6.5. x_2.

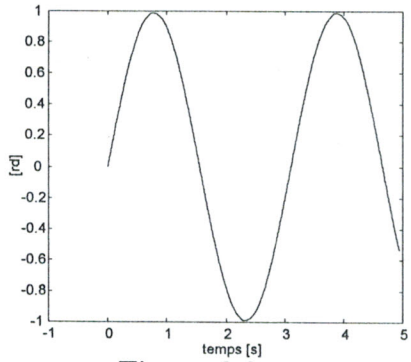


Figure 6.6. x_3.

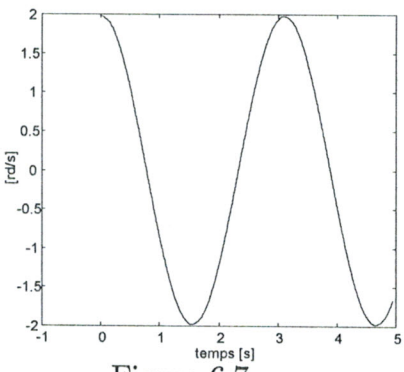


Figure 6.7. x_4.

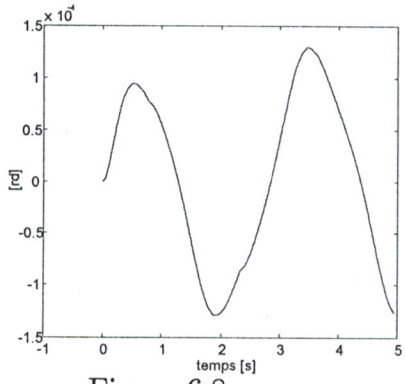


Figure 6.8. x_1-q_d .

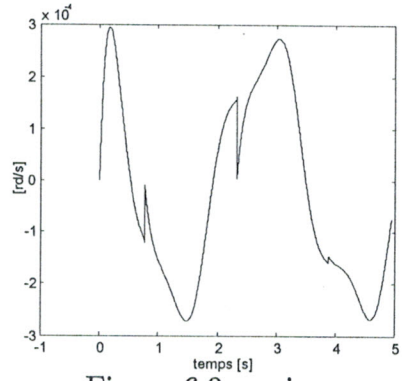


Figure 6.9. x_2-$\dot{q}_d$.

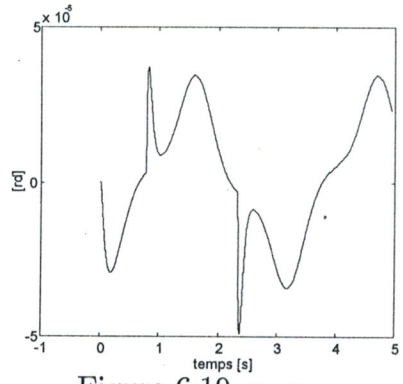


Figure 6.10. x_3-q_d .

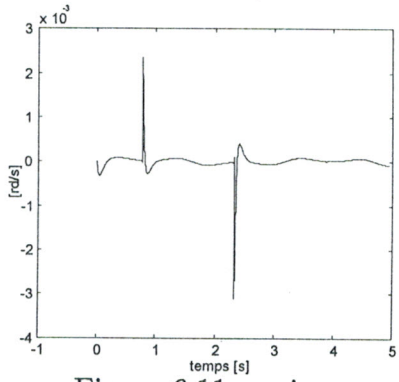


Figure 6.11. x_4-$\dot{q}_d$.

Enfin, d'après les figures ci-dessous, on peut conclure que le processus d'estimation

fonctionne bien.

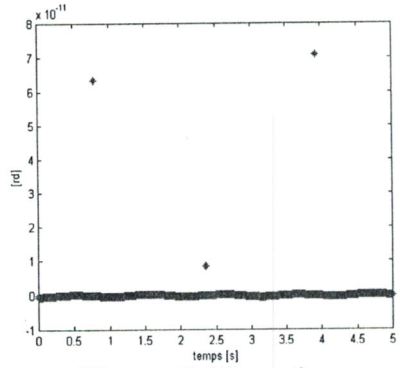


Figure 6.12. x_1-$\hat{x}_1$.

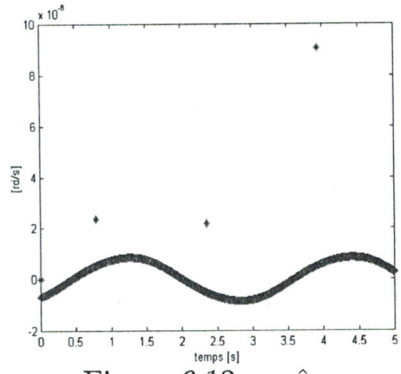


Figure 6.13. x_2-$\hat{x}_2$.

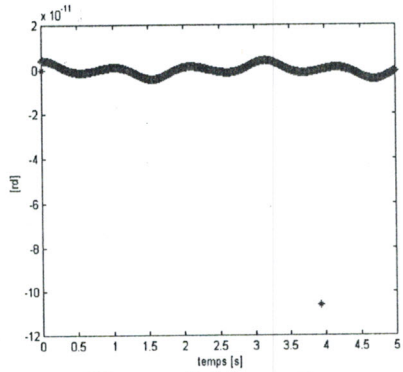


Figure 6.14. x_3-$\hat{x}_3$.

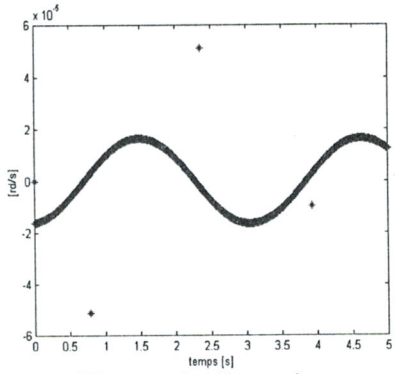


Figure 6.15. x_4-$\hat{x}_4$.

On note que les mêmes simulations ont été exécutées pour une période d'échantillonnage de $T = 10ms$, et de suréchantillonnage de $t_{sh} = 1ms$ mais les erreurs étaient plus importantes.

Donc, comme on l'a constaté notre observateur estime bien la première dérivée. A travers l'exemple qui suit on verra que même jusqu'à une troisième dérivation notre processus d'interpolation et dérivation numérique reste performant.

6.3 Le robot à articulation flexible

6.3.1 Modèle dynamique

On considère un robot à un axe commandé par un moteur à courant continu. L'élasticité au niveau de l'articulation est modélisée par un ressort de torsion linéaire de raideur k [5]. Le schéma d'un tel robot est illustré par la figure (6.16.).

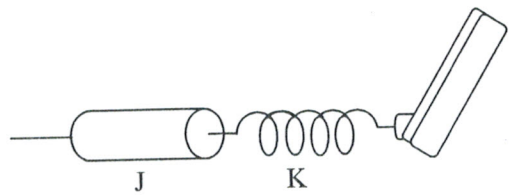


Figure 6.16. Le robot à articulation flexible.

On note q_l l'angle de rotation du bras, et q_m l'angle de rotation du moteur. Pour des raisons de simplification on tiendra compte des deux hypothèses suivantes [15] :

- H1: l'énergie cinétique du rotor est principalement due à sa rotation;
- H2: l'inertie du rotor est symétrique par rapport à l'axe de rotation du rotor de telle sorte que l'énergie potentielle du système et la vitesse du centre de masse du moteur soient indépendants de la position du rotor.

L'énergie cinétique du moteur E_c s'écrit

$$E_c = \frac{1}{2}mv^2 + \frac{1}{2}J_m w^2$$

m est la masse du rotor, v la vitesse du centre de masse du rotor, w la vitesse de rotation et J_m l'inertie du moteur.

A cause de l'hypothèse H1, E_c peut s'écrire [15]

$$E_c = \frac{1}{2}J_l\dot{q}_l^2 + \frac{1}{2}J_m\dot{q}_m^2$$

avec J_l l'inertie du bras.

L'énergie potentielle du bras s'écrit

$$E_p = \frac{1}{2}K(q_l - q_m)^2 + mgl(1 - \cos(q_l))$$

g étant la constante gravitationnelle et l la distance entre l'axe de rotation et le centre de masse du bras.

Les forces de frottement visqueux peuvent être introduites par la fonction de dissipation de Rayleigh [15]

$$\phi = \frac{1}{2}f_1\dot{q}_l^2 + \frac{1}{2}f_2\dot{q}_m^2$$

Les équations du mouvement de ce système sont données par

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} + \frac{\partial \phi}{\partial \dot{q}} = \Gamma$$

avec $q = (q_l, q_m)^T$, $\Gamma = (0, u)^T$ où u est le couple externe appliqué au moteur et $L = E_c - E_p$ est le Lagrangien.

Les équations résultantes sont

$$\begin{aligned} J_l\ddot{q}_l + mgl\sin(q_l) + K(q_l - q_m) + f_1\dot{q}_l &= 0 \\ J_m\ddot{q}_m - K(q_l - q_m) + f_2\dot{q}_m &= u \end{aligned}$$

En choisissant le vecteur d'état $x = (q_l, \dot{q}_l, q_m, \dot{q}_m)^T$ la représentation du système dans l'espace d'état s'écrira alors

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{-mgl}{J_l}\sin(x_1) - \frac{K}{J_l}(x_1 - x_3) - \frac{f_1}{J_l}x_2 = \varphi(x) \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= \frac{K}{J_m}(x_1 - x_3) - \frac{f_2}{J_m}x_4 + \frac{1}{J_m}u \end{aligned} \tag{6.1}$$

6.3.2 Observabilité du robot à articulation flexible

Sous une forme plus compacte, (6.1) s'écrira

$$\dot{x} = f(x) + g(x).u$$

où : $f(x) = \begin{bmatrix} x_2 \\ \frac{-mgl}{J_l}\sin(x_1) - \frac{K}{J_l}(x_1 - x_3) - \frac{f_1}{J_l}x_2 \\ x_4 \\ \frac{K}{J_m}(x_1 - x_3) - \frac{f_2}{J_m}x_4 \end{bmatrix}$, $g(x) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{1}{J_m} \end{bmatrix}$

Si on choisit comme mesure la position du bras, on obtient

$$y = x_1 = h(x)$$

Les dérivées de Lie du vecteur de sortie sont données par

$L_f h(x) = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} f(x) = x_2;$

$L_f^2 h(x) = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} f(x) = \varphi(x);$

$L_f^3 h(x) = \begin{bmatrix} \frac{\partial \varphi}{\partial x_1} & \frac{\partial \varphi}{\partial x_2} & \frac{\partial \varphi}{\partial x_3} & \frac{\partial \varphi}{\partial x_4} \end{bmatrix} f(x) = \psi(x).$

La matrice d'observabilité $\mathcal{O} = \frac{d}{dx}\begin{bmatrix} h(x) & L_f h(x) & L_f^2 h(x) & L_f^3 h(x) \end{bmatrix}$ est donnée par

$$\mathcal{O} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ * & * & \frac{K}{J_l} & 0 \\ * & * & * & \frac{K}{J_l} \end{bmatrix}$$

Les termes représentés par des étoiles sont non nuls et il n'est pas nécessaire de les évaluer pour conclure quant à l'observabilité du système. On a $\det(\mathcal{O}) = \left(\frac{K}{J_l}\right)^2 \neq 0$ d'où l'observabilité faible locale du bras à articulation flexible étant données les mesures des positions bras.

6.3.3 Etude en simulation

Dans cet exemple l'observateur par interpolation et dérivation numérique sera utilisé pour déterminer un certain nombre de dérivées de la sortie permettant l'implémentation d'une loi commandant la position du bras. Tout d'abord il faut déduire l'équation différentielle entrée-sortie du système (6.1). Elle est donnée par

$$y^{(4)} = f(y, \dot{y}, \ddot{y}, y^{(3)}) + \frac{K}{J_l J_m} u$$

avec

$$f(y, \dot{y}, \ddot{y}, y^{(3)}) = -(\frac{f_1}{J_l} + \frac{f_2}{J_m}) y^{(3)} - \left(\frac{K}{J_l} + \frac{K}{J_m} + \frac{f_1 f_2}{J_l J_m} + \frac{mgl}{J_l} \cos(y) \right) \ddot{y} + \left(\frac{mgl}{J_l} \sin(y) \right) \dot{y}^2 +$$
$$- \left(f_2 \frac{mgl}{J_l J_m} \cos(y) + (f_1 + f_2) \frac{K}{J_l J_m} \right) \dot{y} - \frac{K mgl}{J_l J_m} \sin(y)$$

La mesure étant la position du bras, on a $y = q_l$.

La commande linéarisante et découplante employée sera de la forme

$$u = \frac{J_l J_m}{K} (v_u - f(y, \dot{y}, \ddot{y}, y^{(3)}))$$

où : $v_u = q_d^{(4)} - k_4(y^{(3)} - q_d^{(3)}) - k_3(\ddot{y} - \ddot{q}_d) - k_2(\dot{y} - \dot{q}_d) - k_1(y - q_q)$

Les gains k_i, $i = 1, \ldots, 4$, sont choisis tels que l'erreur $e = q_l - q_d$ tende vers zéro quand $t \to \infty$, q_d étant la trajectoire désirée. Cette fois-ci on a choisi $q_d = \sin(\omega t)$, $\omega = 1 rd/s$ et $k_1 = 10000$, $k_2 = 4000$, $k_3 = 600$, $k_4 = 40$ (i.e. les quatres pôles de la dynamique d'erreur sont placés en -10). Les valeurs numériques des différents paramètres du robot sont : $f_1 = 0.1 N/rad.s^{-1}$, $f_2 = 0.2 N/rad.s^{-1}$, $J_l = 1.125 kg.m^2$, $J_m = 1.44 kg.m^2$, $m = 15.5 kg$, $K = 100 N/rad$, $g = 10 m.s^{-2}$.

Les simulations ci-dessous étaient obtenues pour une période d'échantillonnage de $T = 1 ms$, et de suréchantillonnage de $t_{sh} = 100 \mu s$. Les conditions initiales sont choisies telles que les erreurs d'estimation initiales sur la position q_l et la vitesse $\dot{q}_l$ soient nulles.

La figure ci-dessous montre la commande appliquée au système. Elle reste dans les

limites acceptables ($\pm 120 Nm$).

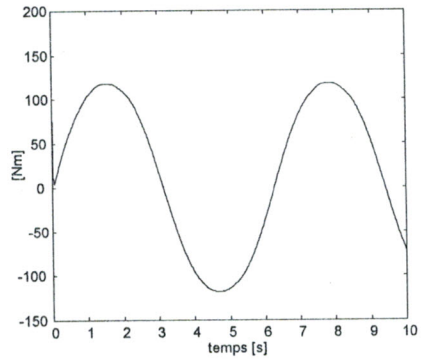


Figure 6.17. Le couple u.

Sur les figures 6.18., 6.19., 6.20., 6.21., est illustré le comportement du système. La trajectoire imposée au bras est bien reproduite. En effet, les erreurs de suivi (figures 6.22., 6.23.), après 3 secondes, rentrent dans des tubes de rayons 0.004 rd et 0.00426 rd/s respectivement.

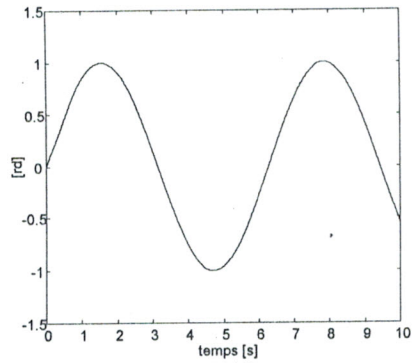


Figure 6.18. x_1 .

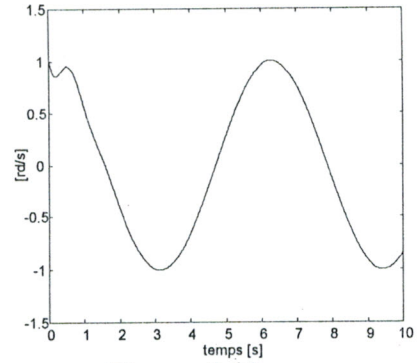


Figure 6.19. x_2 .

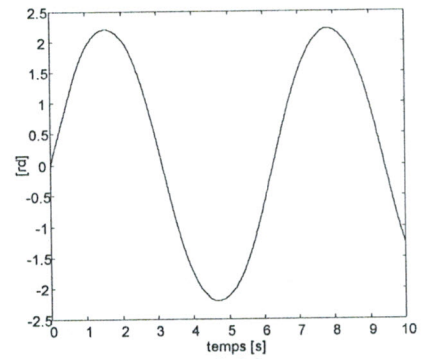


Figure 6.20. x_3 .

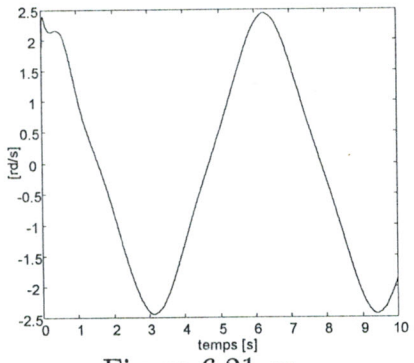


Figure 6.21. x_4 .

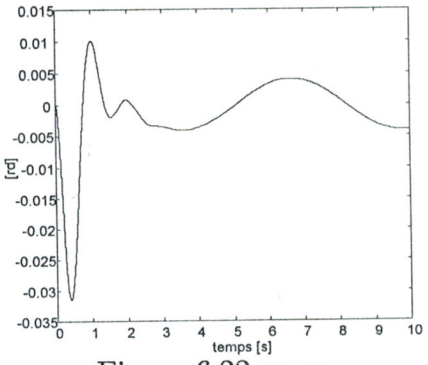


Figure 6.22. x_1-q_d .

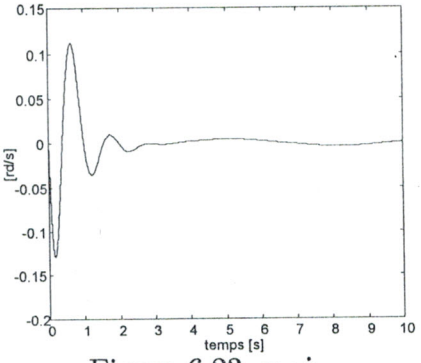


Figure 6.23. x_2-$\dot{q}_d$.

Les erreurs d'estimation déterminées aux instants d'échantillonnage et montéres ci-dessous confirment le bon fonctionnement de notre observateur.

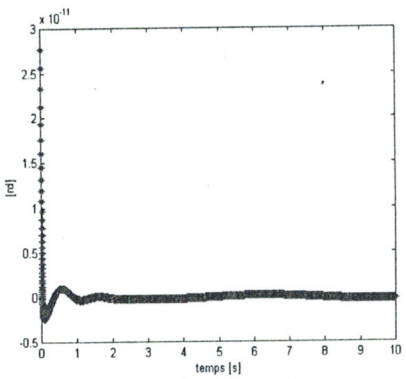


Figure 6.24. x_1-$\hat{x}_1$.

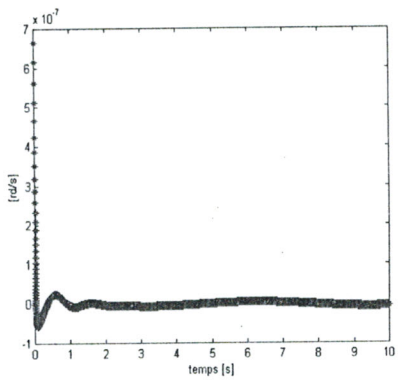


Figure 6.25. x_2-$\hat{x}_2$.

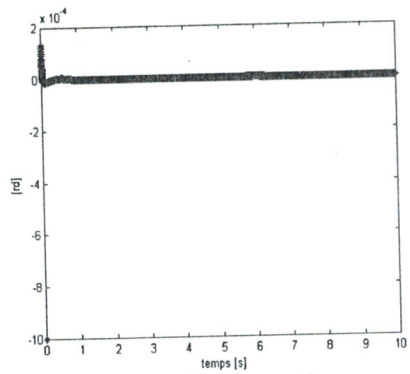


Figure 6.26. x_3-$\hat{x}_3$.

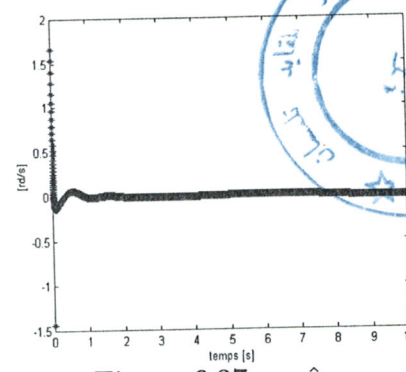


Figure 6.27. x_4-$\hat{x}_4$.

Dans cet exemple également, la simulation a été exécutée pour $T = 10ms$ et $t_{sh} = 1ms$, mais des résultats aussi satisfaisants que les précédents n'ont pu être obtenus, surtout, au niveau des erreurs d'estimation.

Remarque 55 *Les ordres des fonctions splines ont été choisis supérieurs à l'ordre le plus élevé des dérivées nécessaires à l'implémentation de la commande.*

6.4 Conclusion

Tout d'abord, il faut noter que les différentes simulations étaient exécutées sous Simulink à l'aide de programmes écrits sous Matlab. Les essais ont permis de conclure quant à l'applicabilité de l'interpolation et la dérivation numérique à la synthèse d'observateurs et de lois de commande des systèmes considérés.

Pour l'exécution des différentes simulations, il a fallu tenir compte d'un certain nombre de contraintes telles que le choix de la période d'échantillonnage et de suréchantillonnage, du pas d'intégration de la méthode numérique adoptée ainsi que la méthode elle même.

La période de suréchantillonnage ne peut être choisie n'importe comment, car, elle dépend essentiellement des dynamiques du système considéré. On sait qu'au niveau de chaque instant de suréchantillonnage on dispose d'une information concernant l'évolution du système, donc si au cours d'une durée, les dynamiques varient rapidement, le risque de perdre des informations importantes serait certain et par conséquent la commande appliquée ne pouvait jouer son rôle avec efficacité, ceci d'une part. D'autre part, la valeur de

la période d'échantillonnage ne devait pas être très élevée, car pour un pas de suréchantillonnage faible, ceci entraînerait un nombre de points d'interpolation important. Donc, pour déterminer la spline correspondante, son ordre étant fixé au préalable, il fallait évaluer un nombre similaire de B-splines . Il est évident que cette opération exigerait plus de temps et donc influerait la qualité de la commande, sachant qu'elle devait être effectuée en temps réel. A ce niveau, on peut citer l'exemple du robot à articulation flexible où un pas d'échantillonnage de $100ms$ et de suréchantillonnage de $10ms$ ont causé la divergence de la commande, et donc du système, en quelques instants. La remède était, bien sûr, de diminuer les valeurs considérées des deux pas.

On note, aussi, qu'il était plus simple pour nous de choisir une période d'échantillonnage un multiple de celle de suréchantillonnage, ceci pour tenir compte, dans le processus d'interpolation, de l'information présente à la fin de chaque fenêtre d'échantillonnage.

Quant au pas d'intégration de la méthode, on a constaté que pour des mauvais choix, certains instants n'étaient pas pris en compte, alors s'ils correspondaient, par exemple à un instant d'échantillonnage ou de suréchantillonnage, notre processus d'estimation ne pourrait fonctionner correctement ou ne fonctionnerait même pas.

Enfin, on n'oublie pas de signaler que tout le travail effectué ci-dessus suppose que les mesures sont prélevées avec une précision parfaite, i.e., pas de bruit. En conclusion, même dans la réalité, notre observateur peut fournir des résultats très satisfaisants si, à notre connaissance, un processus de filtrage efficace est utilisé.

Chapitre 7

Conclusion générale

Notre objectif, à travers ce travail, était de construire un observateur par interpolation et dérivation numérique, ceci afin de situer les limites de son applicabilité.

Comme on l'a déjà mentionné, les états non accessibles d'un système peuvent être déterminés si l'on dispose de la sortie et d'un certain nombre de ses dérivées. Il était montré, par ailleurs, que les dérivées d'ordres supérieurs des fonctions splines présentaient des précisions acceptables. Il nous a semblé, alors, très intéressant d'exploiter leurs propriétés dans le but de fournir, en temps réel, une estimation des dérivées de la sortie utiles à la détermination des états d'un système donné. En se basant sur ces points, on a synthétisé un observateur et l'utilisé pour la commande de deux systèmes.

Notre but était de réaliser un suivi de trajectoire. On a remarqué, à travers les différents résultats de simulation, que les trajectoires désirées étaient bien reproduites. Aussi, les erreurs d'estimation déterminées aux instants d'échantillonnage étaient très faibles.

On n'oublie pas de noter que les périodes d'échantillonnage et de suréchantillonnage devaient être choisies correctement afin de disposer d'un maximum d'informations sur l'évolution des dynamiques du système.

Les dimensions des systèmes considérés étaient assez réduites ($n = 4$) et donc rien ne garantit le bon fonctionnement de notre observateur si elles étaient importantes, car dans ce cas, il serait nécessaire d'estimer des dérivées d'ordres élevés.

Dans notre cas, le problème de prouver la convergence ne se pose pas, car étant donné le principe utilisé, la bonne estimation est fournie dès le premier instant.

On a constaté, également, que notre processus de dérivation numérique fournit des résultats meilleurs, comparés à ceux obtenus par dérivation numérique sur deux points avec et sans filtrage.

En présence du bruit, il suffit qu'un bon filtrage soit effectué sur les données pour que notre observateur soit utilisé avec succés. En réalité, ce filtrage existe toujours dans les systèmes commandés par calculateurs pour éviter le repliement spectral.

Ce type d'observateurs est une bonne alternative à ceux existants quand les conditions théoriques de leur applicabilité ne sont pas vérifiées. L'estimation des dérivées de la sortie est parfois nécessaire pour la construction d'observateurs à dynamique d'erreur linéaire [15].

Enfin, il serait intéressant, à notre connaissance, de coupler cette technique avec le modèle du processus pour améliorer les propriétés vis-à-vis du bruit.

Bibliographie

[1] J. H. Ahlberg, E. N. Nilson, J. L. Walsh, "The theory of splines and their applications", Academic Press, Inc., London, 1967.

[2] K. Arbenz, A. Wohlhauser, "Analyse numérique", Presses Polytechniques et Universitaires Romandes, Lausanne, 1996.

[3] N. Bakhvalov, "Méthodes numériques", Editions Mir, Moscou, 1976.

[4] G. Bartolini, A. Ferrara, E. Usai, "Real time output derivatives estimation by means of higher order sliding modes", IMACS Multiconference, CESA'98, Tunisia, April 1-4, 1998.

[5] B. Cherki, "Commande des robots manipulateurs par retour d'état estimé", Thèse de Doctorat, Ecole Centrale de Nantes, Université de Nantes, 1996.

[6] S. T. Chung, J. W. Grizzle, "Sampled data observer error linearization", Automatica, vol.26, No 6, pp 997-1007, 1990.

[7] C. De Boor, "A practical guide to splines", Springer Verlag, New York, 1978.

[8] S. Diop, J. W. Grizzle, P.E. Moraal, A.G. Stefanopoulou, "Interpolation and numerical differentiation for observer design", in Proceedings of the American Control Conference, American Automatic Control Council, Evanston, IL, pp. 1329-1333, 1994.

[9] N. Djemai, J. Hernandez, J.P. Barbot, "Nonlinear control with flux observer for a singularly pertubed induction motor", IEEE Conference on Decision and Control, CDC'93, USA, 1993.

[10] A. J., Fossard, D. Normand-Cyrot, "Systèmes non linéaires - Commande", tome 3, pp. 177-219, Masson, 1993.

[11] J. P. Gauthier, G. Bornard, "Observability for any u(t) of a class of nonlinear systems", IEEE Transaction Automatic Control, vol AC-26, No. 4, August 1981.

[12] Y. Hanlong, S. Mehrdad, "Monitoring and diagnostics of a class of nonlinear systems using a nonlinear unknown input observer", Proceedings of the 1996 IEEE International Conference on Control Applications, MI, September 15-18, 1996.

[13] R. Hermann, A. J. Krener, "Nonlinear controllability and observability", IEEE Transaction Automatic Control, vol AC-22, No. 5, October 1977.

[14] P. E. Moraal, "Nonlinear observer design: Theory and applications to automotive control", PHD Thesis, Michigan, 1995.

[15] F. Plestan, "Linéarisation par injection d'entrée-sortie généralisée et synthèse d'observateurs", Thèse de Doctorat, Ecole Centrale de Nantes, Université de Nantes, 1995.

[16] A. Tornambe, "High gain observers for non linear systems", Int. J. Systems SCI., vol.23, No 9, pp 1475-1489, 1992.

[17] G. Zimmer, "State observation by on-line minimization", Int. J. Control, vol.60, No 4, pp 595-606, 1994.