Analyse numérique de quelques problèmes liés au traitement de signaux

Jacques Wolf

THESE

présentée à

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

par

pour obtenir le grade de

Docteur es sciences mathématiques

Jacques WOLF

ANALYSE NUMERIQUE DE QUELQUES PROBLEMES LIES AU TRAITEMENT DE SIGNAUX

Thèse soutenue le 18 octobre 1974 devant la Commission d'Examen :

JURY :

Monsieur CARRE
Monsieur COATMELEC
Monsieur GASTINEL
Monsieur LAURENT
Monsieur PERENNOU

UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

LISTE DES PROFESSEURS

Président : Monsieur Michel SOUTIF
Vice-Président : Monsieur Gabriel CAU

PROFESSEURS TITULAIRES

MM.	ANGLES D'AURIAC Paul	Mécanique des fluides
	ARNAUD Georges	Clinique des maladies infectieuses
	ARNAUD Paul	Chimie
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOUD Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale
	BENOIT Jean	Radioélectricité
	BERNARD Alain	Mathématiques Pures
	BESSON Jean	Electrochimie
	BEZES Henri	Chirurgie générale
	BLAMBERT Maurice	Mathématiques Pures
	BOLLIET Louis	Informatique (IUT B)
	BONNET Georges	Electrotechnique
	BONNET Jean-Louis	Clinique ophtalmologique
	BONNET-EYMARD Joseph	Pathologie médicale
	BONNIER Etienne	Electrochimie Electrométallurgie
	BOUCHERLE André	Chimie et Toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques Appliquées
	BRAVARD Yves	Géographie
	BRISSONNEAU Pierre	Physique du solide
	BUYLE-BODIN Maurice	Electronique
	CABANAC Jean	Pathologie chirurgicale
	CABANEL Jean	Clinique rhumatologique et hydrologie
	CALAS François	Anatomie
	CARRAZ Gilbert	Biologie animale et pharmacodynamie
	CAU Gabriel	Médecine légale et Toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques Pures
	CHARACHON Robert	Oto-Rhino-Laryngologie
	CHATEAU Robert	Thérapeutique
	CHENE Marcel	Chimie papetière
	COEUR André	Pharmacie chimique
	CONTAMIN Robert	Clinique gynécologique
	COUDERC Pierre	Anatomie Pathologique
	CRAYA Antoine	Mécanique

Mme	DEBELMAS Anne-Marie	Matière médicale
MM.	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DESRE Pierre	Métallurgie
	DESSAUX Georges	Physiologie animale
	DODU Jacques	Mécanique appliquée
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	DUGOIS Pierre	Clinique de Dermatologie et Syphiligraphie
	FAU René	Clinique neuro-psychiatrique
	FELICI Noël	Electrostatique
	GAGNAIRE Didier	Chimie physique
	GALLISSOT François	Mathématiques Pures
	GALVANI Octave	Mathématiques Pures
	GASTINEL Noël	Analyse numérique
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques Pures
	GIRAUD Pierre	Géologie
	KLEIN Joseph	Mathématiques Pures
Mme	KOFLER Lucie	Botanique et Physilogie végétale
MM.	KOSZUL Jean-Louis	Mathématiques Pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre-Jean	Mathématiques appliquées
	LEDRU Jean	Clinique médicale B
	LLIBOUTRY Louis	Géophysique
	LOUP Jean	Géographie
Mle	LUTZ Elisabeth	Mathématiques Pures
MM.	MALGRANGE Bernard	Mathématiques Pures
	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Seméiologie médicale
	MASSEPORT Jean	Géographie
	MAZARE Yves	Clinique médicale A
	MICHEL Robert	Minéralogie et Pétrographie
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NEEL Louis	Physique du solide
	OZENDA Paul	Botanique
	PAUTHENET René	Electrotechnique
	PAYAN Jean-Jacques	Mathématiques Pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET René	Servomécanismes
	PILLET Emile	Physique industrielle
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REULOS René	Physique industrielle
	RINALDI Renaud	Physique
	ROGET Jean	Clinique de pédiatrie et de puériculture
	SANTON Lucien	Mécanique
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SILBERT Robert	Mécanique des fluides
	SOUTIF Michel	Physique générale

MM.	TANCHE Maurice	Physiologie
	TRAYNARD Philippe	Chimie générale
	VAILLAND François	Zoologie
	VALENTIN Jacques	Physique nucléaire
	VAUQUOIS Bernard	Calcul électronique
Mme	VERAIN Alice	Pharmacie galénique
M.	VERAIN André	Physique
Mme	VEYRET Germaine	Géographie
MM.	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale
	YOCCOZ Jean	Physique nucléaire théorique

PROFESSEURS ASSOCIES

MM.	BULLEMER Bernhard	Physique
	HANO JUN-ICHI	Mathématiques Pures
	STEPHENS Michaël	Mathématiques appliquées

PROFESSEURS SANS CHAIRE

MM.	BEAUDOING André	Pédiatrie
Mme	BERTRANDIAS Françoise	Mathématiques Pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques appliquées
	BIAREZ Jean-Pierre	Mécanique
	BONNETAIN Lucien	Chimie minérale
Mme	BONNIER Jane	Chimie générale
MM.	CARLIER Georges	Biologie végétale
	COHEN Joseph	Electrotechnique
	COUMES André	Radioélectricité
	DEPASSEL Roger	Mécanique des fluides
	DEPORTES Charles	Chimie minérale
	GAUTHIER Yves	Sciences biologiques
	GAVEND Michel	Pharmacologie
	GERMAIN Jean-Pierre	Mécanique
	GIDON Paul	Géologie et Minéralogie
	GLENAT René	Chimie organique
	HACQUES Gérard	Calcul numérique
	JANIN Bernard	Géographie
Mme	KAHANE Josette	Physique
MM.	MULLER Jean-Michel	Thérapeutique
	PERRIAUX Jean-Jacques	Géologie et Minéralogie
	POULOUJADOFF Michel	Electrotechnique
	REBECQ Jacques	Biologie (CUS)
	REVOL Michel	Urologie
	REYMOND Jean-Charles	Chirurgie générale
	ROBERT André	Chimie papetière
	DE ROUGEMONT Jacques	Neurochirurgie
	SARRAZIN Roger	Anatomie et chirurgie
	SARROT-REYNAULD Jean	Géologie
	SIBILLE Robert	Construction mécanique
	SIROT Louis	Chirurgie générale
Mme	SOUTIF Jeanne	Physique générale

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

Mle	AGNIUS-DELORD Claudine	Physique pharmaceutique
	ALARY Josette	Chimie analytique
MM.	AMBLARD Pierre	Dermatologie
	AMBROISE-THOMAS Pierre	Parasitologie
	ARMAND Yves	Chimie
	BEGUIN Claude	Chimie organique
	BELORIZKY Elie	Physique
	BENZAKEN Claude	Mathématiques appliquées
	BILLET Jean	Géographie
	BLIMAN Samuel	Electronique (EIE)
	BLOCH Daniel	Electrotechnique
Mme	BOUCHE Liane	Mathématiques (CUS)
MM.	BOUCHET Yves	Anatomie
	BOUVARD Maurice	Mécanique des fluides
	BRODEAU François	Mathématiques (IUT B)
	BRUGEL Lucien	Energétique
	BUISSON Roger	Physique
	BUTEL Jean	Orthopédie
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHIAVERINA Jean	Biologie appliquée (EFP)
	CHIBON Pierre	Biologie animale
	COHEN-ADDAD Jean-Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie médicale
	CONTE René	Physique
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie
	DURAND Francis	Métallurgie
	DUSSAUD René	Mathématiques (CUS)
Mme	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	GENSAC Pierre	Botanique
	GIDON Maurice	Géologie
	GRIFFITHS Michaël	Mathématiques appliquées
	GROULADE Joseph	Biochimie médicale
	HOLLARD Daniel	Hématologie
	HUGONOT Robert	Hygiène et Médecine préventive
	IDELMAN Simon	Physiologie animale
	IVANES Marcel	Electricité
	JALBERT Pierre	Histologie
	JOLY Jean-René	Mathématiques Pures
	JOUBERT Jean-Claude	Physique du solide
	JULLIEN Pierre	Mathématiques Pures
	KAHANE André	Physique générale
	KUHN Gérard	Physique
	LACOUME Jean-Louis	Physique
Mme	LAJZEROWICZ Jeannine	Physique
MM.	LANCIA Roland	Physique atomique
	LE JUNTER Noël	Electronique
	LEROY Philippe	Mathématiques
	LOISEAUX Jean-Marie	Physique nucléaire
	LONGEQUEUE Jean-Pierre	Physique nucléaire
	LUU DUC Cuong	Chimie organique
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et Médecine préventive
	MARECHAL Jean	Mécanique
	MARTIN-BOUYER Michel	Chimie (CUS)

MM.	MAYNARD Roger	Physique du solide
	MICHOULIER Jean	Physique (IUT A)
	MICOUD Max	Maladies infectieuses
	MOREAU René	Hydraulique (INP)
	NEGRE Robert	Mécanique
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (IUT B)
	PEFFEN René	Métallurgie
	PELMONT Jean	Physiologie animale
	PERRET Jean	Neurologie
	PERRIN Louis	Pathologie expérimentale
	PFISTER Jean-Claude	Physique du solide
	PHELIP Xavier	Rhumatologie
Mlle	RIERY Yvette	Biologie animale
MM.	RACHAIL Michel	Médecine interne
	RACINET Claude	Gynécologie et obstétrique
	RENAUD Maurice	Chimie
	RICHARD Lucien	Botanique
Mme	RINAUDO Marquerite	Chimie macromoléculaire
MM.	ROMIER Guy	Mathématiques (IUT B)
	SHOM Jean-Claude	Chimie générale
	STIEGLITZ Paul	Anesthésiologie
	STOEBNER Pierre	Anatomie pathologique
	VAN CUTSEM Bernard	Mathématiques appliquées
	VEILLON Gérard	Mathématiques appliquées (INP)
	VIALON Pierre	Géologie
	VOOG Robert	Médecine interne
	VROUSSOS Constantin	Radiologie
	ZADWORNY François	Electronique

MAITRES DE CONFERENCES ASSOCIES

MM.	BOUDOURIS Georges	Radioélectricité
	CHEEKE John	Thermodynamique
	GOLDSCHMIDT Hubert	Mathématiques
	SIDNEY STUARD	Mathématiques Pures
	YACOUD Mahmoud	Médecine légale

CHARGES DE FONCTIONS DE MAITRES DE CONFERENCES

Mme	BERIEL Hélène	Physilogie
Mme	RENAUDET Jacqueline	Microbiologie

Fait le 30 mai 1972.

Je tiens à exprimer ma profonde reconnaissance à :

Monsieur le Professeur GASTINEL qui m'a appris les difficultés de l'analyse numérique, qui s'est vivement intéressé à ce travail et qui a dirigé cette thèse. Qu'il me soit permis - en cette occasion - de le remercier, en particulier des longues et nombreuses et ... fructueuses discussions que nous avons eues dans le calme du samedi matin.

Monsieur le Professeur COATMELEC qui, lorsque j'étais étudiant à Rennes, a su m'initier et m'orienter vers ce qu'on appelait le "calcul automatique", et qui se trouve aujourd'hui dans ce jury.

Monsieur le Professeur LAURENT qui, par ses remarques à divers séminaires a permis d'améliorer certaines démonstrations.

Monsieur PERENNOU, Directeur du Centre Interuniversitaire de Calcul de Toulouse, et Monsieur CARRE, chargé de recherche au C.N.R.S., de l'intérêt qu'ils témoignent à ce travail en acceptant de faire partie du jury.

Je tiens à remercier également toute l'équipe d'analyse numérique, et en particulier Messieurs DELLA-DORA et EBERHARD dont les remarques tout au long de ce travail ont été précieuses.

Je tiens à remercier Mademoiselle PAYERNE pour sa patience à frapper de tels textes, à remercier le service perforation et tous les opérateurs sans qui les essais de tous les programmes n'auraient pu être menés à bien, et à remercier le service de reproduction pour la qualité de ce document.

P L A N

INTRODUCTION

Le traitement des "signaux" n'est pas un domaine nouveau de la science. Mais ce qui est sans doute nouveau c'est la précision et la rapidité que l'on attend des méthodes traitant du signal. Le traitement analogique (en continu) du signal a fait la force de l'électronique. Cette électronique, on la retrouve encore comme support du traitement digital de l'information, mais l'avènement des ordinateurs a amené de nombreuses études sur le traitement digital du signal, certaines étant la transposition des méthodes analogiques, d'autres entièrement nouvelles et dues à la nature différente du cas discret.

Dans un sens peu précis on sait que l'on appelle signal une quantité qui varie avec le temps et par là même un signal est considéré comme un support d'information, par exemple :

- pression acoustique au niveau de l'oreille
- hauteur d'eau en un point de la côte (étude des marées)
- excitation d'une cellule photo-électrique.

En fait un "signal" sera représentable par x : $A \subset \mathbb{R} \to \mathbb{R}$ (ou $\mathbb{C}$). On peut alors réunir des signaux possédant une même propriété :

Cas continu :

signaux périodiques : $\{x / x(t+T) = x(t) \ \forall t\}$

signaux d'énergie finie : $\{x / \int_{-\infty}^{+\infty} |x(t)|^2 dt < +\infty\}$

Cas discret :

signaux d'énergie finie : $\{\{x_i\} / \sum_{-\infty}^{+\infty} |x_i|^2 < +\infty\}$

signaux bornés : $\{\{x_i\} / \operatorname{Max}_{i=-\infty,+\infty} |x_i| < M\}$

En considérant des applications de tels ensembles dans d'autres, on définira, par exemple un filtre linéaire comme une application de ℓ_∞ dans ℓ_∞ .

La notion de fonctionnelle (application d'un ensemble de signaux dans $\mathbb{R}$) permet, en général, de calculer une caractéristique du signal traité

Exemple : $$x \in S \to \ell(x) = \int_{-\infty}^{+\infty} |x(t)|^2 dt$$

qui permet de mesurer l'énergie transportée par x(t).

Une grande partie de cette étude sera essentiellement consacrée à une application linéaire - appelée filtre - d'un ensemble de signaux discrets dans un autre ensemble de signaux discrets. A cette application on ajoute, une fois les "espaces" précisés, les propriétés de continuité, stationnarité. On verra alors que cette application linéaire est caractérisée par sa fonction de transfert.

ESPACES DE SIGNAUX

En définissant un certain nombre de lois sur les éléments d'un ensemble de signaux, on peut définir des structures : espaces vectoriels de signaux, sous-espaces vectoriels,... . On a souvent la nécessité de définir une norme, ce qui entraine une notion de convergence, et donc une structure topologique.

PLAN DE TRAVAIL

Dans ce travail on ne considère que l'ensemble des signaux, représentés par des applications de $\mathbb{R}$ dans $\mathbb{R}$:

$$S = \{x \ / \ x(t) = \sum_{k=1}^{n} h_k(t)\sin(\omega_k t + \varphi_k) \ ; \ h_k \in C(-\infty,\infty); \ h_k(t) \equiv 0 \ t \notin [a,b]\}.$$

La première partie (chapitre I) est consacrée à l'étude de signaux périodiques, dans laquelle on met en évidence une méthode de calcul de la période par "passage à zéros", méthode convergente lorsque le pas de discrétisation tend vers 0.

Le chapitre suivant (chapitre II) se propose de calculer quelques caractéristiques de signaux transitoires ($\in S$), ce sera l'occasion d'aborder et de montrer les difficultés numériques d'un calcul direct de la transformée de Hilbert.

Les chapitres suivants sont plus précisément orientés vers le calcul de filtres digitaux. Après avoir rappelé les principales propriétés et constructions de tels filtres (chapitre III), le chapitre IV permet de démontrer l'existence de filtres optimaux (au sens de la norme de L^2) et des propriétés d'alternance de ces filtres optimaux. Le chapitre V explicite des moyens numériques d'obtenir les filtres précédemment étudiés, et permet de définir des procédés économiques de construction de filtres optimaux ou quasi-optimaux à partir de quelques filtres de base.

Un élément de L^2 n'ayant pas forcément un meilleur approximant rationnel (généralisé) dans L^2, on montre l'existence d'un tel meilleur approximant après régularisation. Ce travail se termine (chapitre VII) par une application concrète (analyse des sons) qui met en oeuvre toutes les techniques étudiées dans ce travail.

Il faut cependant souligner que tous les problèmes reliés au filtrage numérique à une dimension n'ont pas été étudiés, non parce que ces problèmes n'ont pas été jugés importants — bien au contraire — mais parce que le temps, le volume de cette étude auraient été au moins doublé (ou plus). Il faut ainsi signaler :

les problèmes de quantification [29] [7] ,

les problèmes d'échantillonnage,

les filtres non récursifs, [5][41][46]

Plus généralement pour faire le point sur tous ces problèmes et leurs applications on se reportera avec intérêt aux revues spécialisées [a] [b] , et à [59] .

CHAPITRE I

PARTIE STATIONNAIRE D'UN SIGNAL

I.1. - INTRODUCTION

Dans ce chapitre on étudie la partie stationnaire d'un signal discrétisé sur un intervalle de temps borné.

On s'attachera à calculer certains paramètres de deux types de signaux:

- signaux périodiques
- signaux dont les partiels ne sont pas des harmoniques du fondamental (signaux presque périodiques).

Parmi les méthodes existantes pour calculer la période (ou le fondamental et les partiels) d'une fonction, la plus connue et la plus utilisée est la méthode utilisant des approximations du spectre de la fonction. Les méthodes de Fourier ont supplanté toutes les autres avec la mise au point d'algorithmes très rapides comme la "Fast Fourier Transform" [13] . On a voulu, ici, éviter le recours à l'analyse spectrale et préféré "travailler" directement sur le signal discrétisé, voie qui n'est d'ailleurs pas nouvelle, afin d'étudier, directement, l'influence de divers paramètres (longueur d'enregistrement, pas d'échantillonage) sur la détermination de la période. Dans une première partie on se propose d'utiliser une méthode basée sur les passages par zéro de la fonction périodique étudiée, ce qui nous conduira à des algorithmes du premier ordre deuxième ordre,... ou plus. Cependant dans tout traitement du signal il y a deux paramètres délicats à choisir (de telle manière que le "travail" — c'est-à-dire le volume de calculs — à effectuer soit le plus faible possible), ce sont le pas d'échantillonage et la longueur d'enregistrement

L'idéal est de réunir — afin d'avoir des résultats précis — les deux qualités :

* pas d'échantillonage petit
* grande longueur d'enregistrement (ceci par rapport au fondamental étudié).

Une deuxième partie (où le signal n'est pas périodique) avait l'ambition de répondre au problème : qu'elle est la longueur d'enregistrement nécessaire pour avoir des résultats valables. Mais il semble jusqu'à présent très difficile d'utiliser cette méthode directement parce qu'on utilise un signal tronqué alors que la transformée de Hilbert utilise dans notre théorie un signal défini sur $[-\infty,+\infty]$.

I.2. - SIGNAL PERIODIQUE

Soient N, entier positif, et T, réel positif, deux nombres fixés. On considère l'ensemble V des fonctions qui peuvent s'écrire :

$$\sum_{k=1}^{N} a_k \cos(k\Omega t + \varphi_k) \qquad (\Omega = 2\pi/T)$$

V est un espace vectoriel, de fonctions périodiques de période T

PROPOSITION 1 :

Sur l'intervalle $[pT,(p+1)T]$, $\forall p$, tout élément de V possède un nombre (> 0) fini de racines.

Soit $f \in V$:

a) f possède au moins une racine sur $[pT,(p+1)T]$

$$f \in V \iff f(t) = \sum_{k=1}^{N} a_k \cos(k\Omega t + \varphi_k)$$

$$\int_{pT}^{(p+1)T} \sum_{1}^{N} a_k \cos(k\Omega t + \varphi_k)dt = \sum_{1}^{N} a_k \int_{pT}^{(p+1)T} \cos(k\Omega t + \varphi_k)dt = 0$$

donc

$$\int_{pT}^{(p+1)T} f(t)dt = 0 \quad \forall p \ ,$$

et puisque f(t) est continue :

$$\exists \xi \in]pT,(p+1)T[\quad \text{tel que } f(\xi) = 0 \quad \text{(théorème de la moyenne).}$$

b) f possède un nombre fini de racines sur $[pT,(p+1)T]$

f est une fonction analytique. Puisque [9] toute fonction analytique possède un nombre fini de zéros sur un compact, la proposition 1 est ainsi démontrée.

Soit un intervalle $[a,b]$ et un maillage de cet intervalle, c'est-à-dire la donnée de (figure I.1)

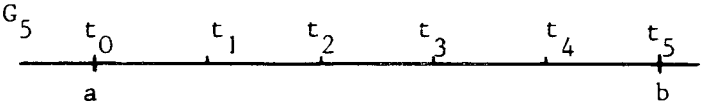


Figure I.1

$$G_n = \{t_i : t_i = a+ih \ , \ i=0,1,\ldots,n \ ; \ h = \frac{b-a}{n}\}$$

Soit $f \in V$, on suppose :

a) f(t) donnée, pour tout $t \in G_n$

b) (b-a) >> T,c'est-à-dire que l'intervalle $[a,b]$ contient un nombre "assez grand" de périodes.

2.1. CARACTERISATION DE LA PERIODE T

Sur l'intervalle $[a,b]$, soient $\rho_0,\rho_1,\ldots,\rho_P$ les racines — en nombre fini d'après la propriété 1 — de f(t) sur $[a,b]$

Il est immédiat de voir que :

$$\exists m_k \ (>0) \text{ tel que } \rho_{j+m_k} - \rho_j = kT \quad \forall j=0,1,\ldots,P-m_k \qquad (0)$$

Soit $m = m_1$, c'est-à-dire m est l'entier qui donne la période.

La réciproque n'est pas vraie :

$$\forall j \qquad \rho_{j+m} - \rho_j = u \not\Rightarrow u = kT$$

Contre exemple : il suffit de prendre :

$$f(t) = \cos\Omega t \in V \ , \ [a,b] \equiv [\tfrac{\pi}{2},5\pi]$$

pour trouver $u = T/2$.

DEFINITION

$$\text{Soit } E(t,f) = \{\tfrac{u}{T} \ / \ f(t+ju) = f(t) \quad \forall j \in \mathbb{Z} \ , \ u \neq kT \quad \forall k \in \mathbb{Z}\}$$

$$\text{et} \quad S(f) = \{t \ / \ E(t,f) \neq \emptyset\} \ .$$

S(f) est appelé <u>support</u> de f.
(Une étude détaillée de ce support sera faite dans le chapitre II).
On suppose alors que :

$$\forall t \in G_n \ , \ t \notin S(f) \quad \text{ou que} \quad S(f) = \emptyset \ .$$

2.2. ZEROS NUMERIQUES

Puisque f n'est connue qu'en un certain nombre de points, il est en général impossible de connaître les racines ρ_i (i=0,...,P) exactement. Cependant nous dirons que t_k est un zéro numérique de f si :

* $f(t_k) = 0$

* ou bien $f(t_k)f(t_{k+1}) < 0$.

Soit G_{n_p} un maillage donné, et h_p le "pas de discrétisation" correspondant, dans ce cas f possède les zéros numériques, sur l'intervalle $[a,b]$:

$$\rho_o^{(p)},\rho_1^{(p)},\ldots,\rho_{N_p}^{(p)} \quad \text{avec} \quad \begin{matrix} \tau_o^{(p)},\ldots,\tau_{L_p}^{(p)} & \text{zéros impairs} \\ \eta_o^{(p)},\ldots,\eta_{M_p}^{(p)} & \text{zéros pairs} \end{matrix}$$

Soient τ_i i=0,1,...,L les zéros d'ordre impair de f sur $[a,b]$

η_i i=0,1,...,M les zéros d'ordre pair de f sur $[a,b]$

PROPOSITION 2

Si $f \in V$, et si $\forall t \in G_{n_p}$, $t \notin S(f)$ alors :

i) $\exists h$, tel que $\forall h_p \leq h$, $L_p = L$

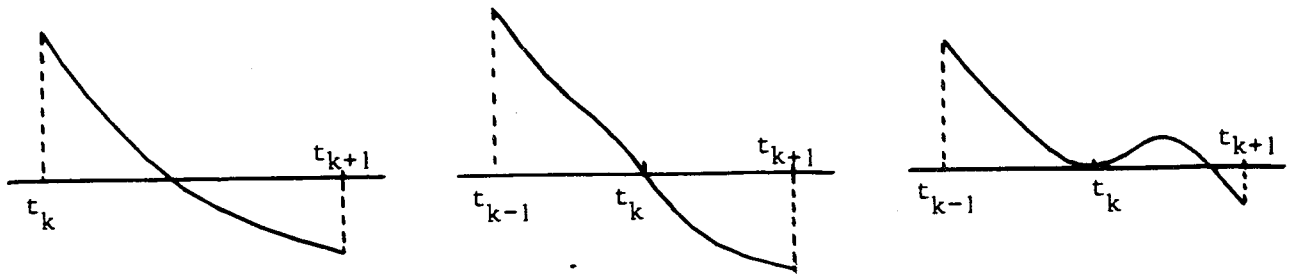
ii) on peut construire une suite $\{\nu^{(p)}\}$ telle que

si $h_p \downarrow 0$ $p \to \infty$ $\lim_{p\to\infty} \nu^{(p)} = T$

On pose :

$$\delta = \text{Inf } (\tau_{i+1} - \tau_i) > 0 \qquad (1)$$

On dira que t_k est un zéro numérique d'ordre <u>impair</u> de f dans les trois cas suivants : (figure I.2)



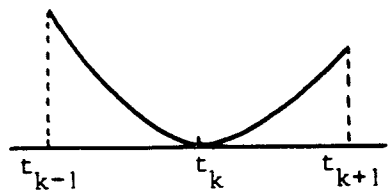
$f(t_k)f(t_{k+1}) < 0$

$f(t_k) = 0$
$f(t_{k-1})f(t_{k+1}) < 0$

$f(t_k) = 0$
$f(t_{k-1})f(t_{k+1}) < 0$

Figure I.2

Le seul cas possible d'avoir un zéro d'ordre <u>pair</u> est : (figure I.3)



$f(t_{k-1})f(t_{k+1}) < 0$

$f(t_k) = 0$
$f(t_{k-1})f(t_{k+1}) > 0$

Figure I.3

Si $h_p < \delta$, entre τ_i et τ_{i+1} ($\forall i=0,1,\ldots,L-1$) il existe nécessairement deux points de discrétisation, ce qui signifie qu'il existe deux points $\in G_{n_p}$ à encadrer chaque zéro impair de f (cas de la figure I.2)

Il est alors clair que tous les zéros impairs de f sont trouvés, aux points t_{k_i} ($i=0,\ldots,L_p$) donc si $h_p < \delta$, $L_p = L$.

Puisqu'il est possible aussi d'avoir trouvé des zéros pairs de f, ceci montre que nécessairement $N_p \geq L$.

On a de plus en posant :

$$\tau_j^{(p)} = t_{k_j} \qquad (j=0,\ldots,L) : \ |\tau_j^{(p)} - \tau_j| < h_p \qquad (2)$$

et si t_{i_j} est un zéro pair de f(t) on a immédiatement :

$$\eta_j = t_{i_j} \qquad (j=0,\ldots,M_1) \qquad M_1 \leq M \qquad (3)$$

2.3. CALCUL DE LA PERIODE T

Soit un pas h_p de discrétisation fixé, satisfaisant $h_p < \delta$, on rappelle que $\tau_i^{(p)}$ (i=0,...L) sont les zéros numériques impairs de f. On se propose alors d'étudier la distance mutuelle de ces zéros numériques en posant :

$$z_{ij}^{(p)} = \tau_{i+j}^{(p)} - \tau_j^{(p)} \qquad j=0,\ldots,N_p-i$$

avec $i \in I$ (ensemble d'indices que l'on précisera plus tard).

Soient, pour i fixé, $\nu_i^{(p)}$ la moyenne de ces distances de zéros et $\sigma_i^{(p)}$ l'écart type correspondant :

$$\nu_i^{(p)} = \sum_{j=0}^{L-i} z_{ij}^{(p)} \ / \ (L-i+1) \qquad i \in I$$

$$[\sigma_i^{(p)}]^2 = \sum_{j=0}^{L-i} [z_{ij}^{(p)} - \nu_i^{(p)}]^2 \ / \ (L-i+1) \qquad i \in I$$

Si nous connaissons une borne inférieure de $\lfloor \frac{b-a}{T} \rfloor$ soit R on pose

$$B = \lfloor L/R \rfloor$$

et on pose :

$$I = \{i : 1,2,\ldots,B\} \ .$$

PROPOSITION 3

$$Si\ h_p < \delta = \inf_{i=0,\ldots,L} (\tau_{i+1} - \tau_i)$$

$$\exists k \ tel\ que \lim_{p\to\infty} \sigma_k^{(p)} = 0$$

Posons : $J = \{j : 0,1,\ldots,L-i\}$

Evaluons la variance :

$$[\sigma_i^{(p)}]^2 = \sum_{j\in J} [z_{ij} - \frac{\sum_{k\in J} z_{ik}^{(p)}}{L+1-i}]^2 / (L-i+1)$$

$$= \sum_{j\in J} [\sum_{k\in J} \frac{z_{ij}^{(p)} - z_{ik}^{(p)}}{L-i+1}]^2 / (L-i+1)$$

Evaluons les quantités :

$$z_{ij}^{(p)} - z_{ik}^{(p)} = (\tau_{i+j}^{(p)} - \tau_j^{(p)}) - (\tau_{i+k}^{(p)} - \tau_k^{(p)})$$

$$z_{ij}^{(p)} = (\tau_{i+j}^{(p)} - \tau_{j+i}) + (\tau_j - \tau_j^{(p)}) + (\tau_{j+i} - \tau_j) \qquad (3')$$

D'après (2) on a :

$$\tau_k^{(p)} - \tau_k = \xi_k^{(p)} h_p \qquad |\xi_k^{(p)}| < 1$$

On rappelle la définition de m , voir (0) :

$$\exists m \ tel\ que \ \tau_{j+m} - \tau_j = T \qquad \forall j=0,\ldots,L-m$$

On a donc :

$$z_{ij}^{(p)} = (\xi_{j+i}^{(p)} - \xi_j^{(p)}) h_p + \tau_{j+i} + T - \tau_{j+m} \qquad (4)$$

D'où :

$$z_{ij}^{(p)} - z_{ik}^{(p)} = \underbrace{(\tau_{j+i} - \tau_{j+m} - \tau_{k+i} + \tau_{k+m})}_{u_{jk}(i)} + \underbrace{(\xi_{j+i}^{(p)} - \xi_{k+i}^{(p)} + \xi_{k}^{(p)} - \xi_{j}^{(p)})h_p}_{\delta_{jk}^{(p)}(i)}$$

On a immédiatement :

$$|\delta_{jk}^{(p)}(i)| < 4 \qquad \forall i \qquad (5)$$

et

* si $i = m$ $\qquad u_{jk}(m) = 0$,

Donc :

$$\left[\sigma_m^{(p)}\right]^2 = h_p^2 \sum_{j\in J} \left[\sum_{k\in J} \delta_{jk}^{(p)}(m)\right]^2 / (L-m+1) = A(m)h_p^2$$

ce qui entraine d'après (5)

$$\sigma_m^{(p)} < 4\, h_p \ . \qquad (6)$$

* Si $i \neq m$ (et $i \neq m_k$) :

$$\left[\sigma_i^{(p)}\right]^2 = \left[U(i) + 2\Delta(i)h_p + V(i)h_p^2\right] / (L-i+1)$$

U(i) étant une constante ne dépendant pas de h (on a supposé $i \neq m_k$, car on aurait aussi pu écrire :

$$z_{ij}^{(p)} = (\xi_{j+i}^{(p)} - \xi_{j}^{(p)})h_p + \tau_{j+i} + kT - \tau_{j+m_k}$$

Ce qui montre, d'après (6), que :

lorsque $h_p \to 0$, quand $p \to \infty$, $\lim_{p\to\infty} \sigma_m^{(p)} = 0$.

Pour cet indice m calculons alors $\nu_m^{(p)}$

$$\nu_m^{(p)} = \sum_{j=0}^{L-m} z_{mj}^{(p)} \; / \; (L-m+1)$$

et d'après (4)

$$z_{mj}^{(p)} = T + (\xi_{j+m}^{(p)} - \xi_j^{(p)})h_p$$

donc :

$$\nu_m^{(p)} = T + \Lambda(m)h_p \qquad (\Lambda(m) < 2) \tag{7}$$

Ce qui démontre, lorsque $h_p \searrow 0$, la proposition 2.

2.4. CALCUL DE LA SUITE $\nu_i^{(p)}$, p FIXE

Si on pose :

$$\mu_i^{(p)} = \sum_{j \in J} [\tau_{j+i}^{(p)} - \tau_j^{(p)}]$$

il est immédiat de voir que :

$$\mu_i^{(p)} = \mu_{i-1}^{(p)} + \rho_{L-(i-1)}^{(p)} - \rho_{i-1}^{(p)}$$

$$\nu_i^{(p)} = \frac{\mu_i(p)}{L-i+1} = \frac{L-i+2}{L-i+1} \frac{\mu_{i-1}^{(p)}}{L-(i-1)+1} + \frac{\rho_{L-(i-1)}^{(p)} - \rho_{i-1}^{(p)}}{L-i+1}$$

c'est-à-dire :

PROPOSITION 4 :

La suite des $\nu_i^{(p)}$ se calcule par la récurrence : $\nu_o^{(p)} = 0$,

$$\nu_i^{(p)} = \frac{L-i+2}{L-i+1} \nu_{i-1}^{(p)} + \frac{\tau_{L-(i-1)}^{(p)} - \tau_{i-1}^{(p)}}{L-i+1} \qquad \forall i > 0$$

2.5. ORDRE DE LA METHODE - ACCELERATION

2.5.1. Interpolation

La méthode est dite d'ordre ℓ si :

$$\nu_m^{(p)} - T = \eta h_p^{\ell} \qquad (\eta \text{ borné lorsque } p \to \infty) .$$

D'après (7), la méthode est au moins d'ordre 1. Il est aussi clair de voir que plus l'ordre est élevé, plus vite la période T sera obtenue. (Il ne faut cependant pas oublier que ceci est valable pour $h_p < \delta$) .

On peut accélérer le procédé en construisant des procédés d'ordre supérieur :

Deux cas sont à envisager :

Premier cas :

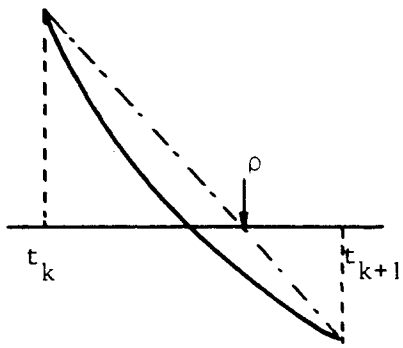


Figure I.4

Au lieu de prendre $\rho = t_k$, on prendra pour ρ l'intersection de la droite passant par $f(t_k)$ et $f(t_{k+1})$ et de l'axe des t — interpolation linéaire — soit : (Figure I.4)

$$\rho = \frac{t_k f_{k+1} - t_{k+1} f_k}{f_{k+1} - f_k}$$

Deuxième cas :

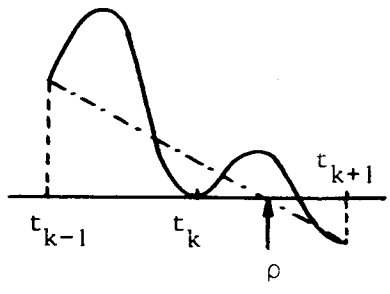


Figure I.5

Ici, toujours par interpolation, on prendra :

$$\rho = \frac{t_{k+1} f_{k-1} - t_{k-1} f_{k+1}}{f_{k-1} - f_{k+1}}$$

comme zéro impair, et la valeur de t_k comme zéro d'ordre pair (figure I.5).

En choisissant ces valeurs pour ρ il est facile de voir — résultat classique — que (2) est remplacée par :

$$|\tau_j^{(p)} - \tau_j| \leq c\ h_p^2$$

et la relation (4) s'écrira maintenant :

$$z_{mj}^{(p)} = (\xi_{j+m}^{(p)} - \xi_j^{(p)})h_p^2 + T \qquad |\xi_j^{(p)}| \leq c$$

ce qui nous donne un procédé du second ordre, donc a priori, plus rapide.

2.5.2. Extrapolation

Soit $G(h) = \nu_m(h)$.

On a immédiatement d'après (4) : $G(0) = \nu_m(0) = T$.

Supposons la valeur de G calculée pour $h = h_1,\ldots,h_n$, on choisit alors des coefficients B_k^n afin de prendre pour approximation de G(0)

$$M_n(G) = \sum_{k=1}^{n} B_k^n\ G(h_k)$$

On a [32] : une condition nécessaire et suffisante, pour que $\lim_{n\to\infty} M_n(G) = G(0)$ pour toute fonction continue en 0^+ , est qu'il existe une constante $\alpha > 1$ telle que $h_i / h_{i+1} \geq \alpha$.

Il nous faut donc vérifier que G est continue à droite :

$$z_{mj}(h) = T + (\tau_j - \tau_j(h)) + (\tau_{i+j}(h) - \tau_{i+j})$$

On représente $\tau_j - \tau_j(h)$ (figure I.6) fonction continue en 0^+ . Puisque $\nu_m(h)$ est une combinaison linéaire finie de $z_{mj}(h)$ $\nu_m(h)$ est donc continue en 0^+ .

Donc si h_i / h_{i+1} est choisi $\geq \alpha > 1$ les conditions du théorème précédent sont remplies et le procédé d'extrapolation est applicable.

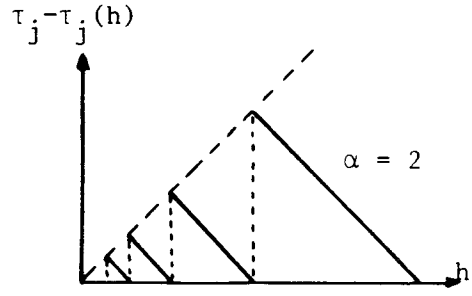


Figure I.6

2.6. EXEMPLE NUMERIQUE

Soit :

$$f(t) = \sum_{k=1}^{10} a_k \sin\left(\frac{2\pi}{T} kt + 2\pi\varphi_k\right)$$

k	1	2	3	4	5	6	7	8	9	10
a_k	1	0.89	0.56	0.12	0.25	0.32	0.54	0.21	0.09	0.05
φ_k	0.02	0.35	0.56	0.12	0.23	0.51	0.65	0	0.2	0.35

Période simulée T = 0.0117 seconde

$$[a,b] \equiv [0,0.062]$$

Estimation de $\left[\frac{b-a}{T}\right] = 3$, donnée a priori.

Tableau de résultats : (procédé du premier ordre)

h	$4\ 10^{-3}$	$2\ 10^{-3}$	10^{-3}	$5\ 10^{-4}$	$2.5\ 10^{-4}$	$1.25\ 10^{-4}$
Nombre de racines sur $[a,b]$	10	12	18	30	40	44
T calculée	$1.77\ 10^{-2}$	$2.17\ 10^{-2}$	$3.41\ 10^{-3}$	$2.08\ 10^{-3}$	$1.54\ 10^{-3}$	$1.1698\ 10^{-3}$

fréquence la plus basse dans le signal : $f_b = 85$ Hz

fréquence la plus haute dans le signal : $f_m = 850$ Hz

On peut remarquer que, dès que tous les zéros sont trouvés, les résultats sont excellents, ce que confirment les autres exemples traités.

2.7. QUELQUES GENERALISATIONS DE LA METHODE

2.7.1. Signal particulier

On suppose que le signal f(t) vérifie sur l'intervalle $[a,b]$:

i) $f(t + T + \varepsilon(t)) = f(t)$ avec $|\varepsilon(t)| \leq \varepsilon << T \quad \forall t \in [a,b]$

ii) f(t) possède un nombre fini de zéros sur $[a,b]$

Soit encore :

$$E_\varepsilon(t,g) = \{u \,/\, g(t+ju+\varepsilon(t)) = g(t) \quad \forall j \in \mathbb{Z}\ ,\ u \neq kT \quad \forall k \in \mathbb{Z}\}$$

et $$S(g) = \{t \,/\, E_\varepsilon(t,g) \neq \emptyset\}$$

S est encore appelé support de f.

On suppose aussi que, G_n étant un maillage

iii) $\forall t \in G_n$, $t \notin S(g)$, ou que $S(g) = \emptyset$

On a alors le résultat suivant :

PROPOSITION 5 :

Si le signal g(t) satisfait les conditions i), ii), iii) alors $\exists h$ *tel que , si* $h_p < h$, *et si* $h_p \searrow 0$ *lorsque* $p \to \infty$, *on peut construire une suite*

$$\eta^{(p)} \text{ telle que } \lim_{p\to\infty} |\eta^{(p)} - T| \leq \varepsilon$$

Les démonstrations sont de même nature que celles utilisées précédemment, en voici les grandes lignes [67].

Soient $\alpha_0,\ldots,\alpha_p$ les zéros de g sur [a,b]

$\exists m$ tel que $|(\alpha_{j+m} - \alpha_j) - T| \leq \varepsilon$

On prend à nouveau $h < \inf_i (\tau_{i+1} - \tau_i)$ (τ_i zéros impairs de g)

On construit alors les suites :

$$\eta_i^{(p)} = \sum_j (\alpha_{i+j}^{(p)} - \alpha_j^{(p)}) \; / \; (L-i+1)$$

$$\xi_i^{(p)} = \sum_j [(\alpha_{i+j}^{(p)} - \alpha_j^{(p)}) - \eta_i^{(p)}]^2 \; / \; (L-i+1)$$

On montre alors que

$$* \qquad [\xi_i^{(p)}]^2 = S_i(\alpha) + R_i\,\varepsilon^2 + V_i\,h\varepsilon + \ldots$$

avec $S_i(\alpha)\begin{cases} = 0 & i = m \\ \neq 0 & i \neq m \; (\neq m_k) \end{cases}$

* $\lim_{p\to\infty} \eta_m^{(p)} = T + W\varepsilon$ ($|W| < 1$) ce qui conduit au résultat.

2.7.2. Signal modulé

Soit une fonction $k(t) > 0 \quad \forall t \in [a,b]$, k est appelé modulation du signal k(t), — h(t) étant appelée porteuse et étant : soit le signal précédent f soit g — . On considère le signal :

$$H(t) = k(t)\, h(t)$$

On ne peut évidemment pas parler de la période de H(t), mais on peut chercher à calculer la période de h(t).

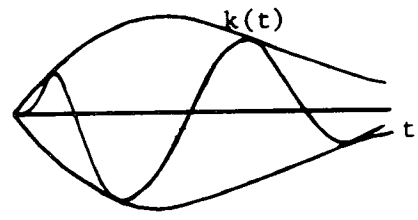


Figure I.7

Il est à remarquer, puisque $k(t) > 0$, que k(t) n'intervient pas dans la détermination des zéros de H(t) qui sont ceux de h(t). Donc utilisant la méthode précédente, on est capable de calculer la période de la porteuse.

2.8. CONCLUSION SUR CETTE METHODE

En général, en théorie du signal [38] , le pas d'échantillonage utilisé est choisi en fonction du théorème de Shannon, c'est-à-dire h doit satisfaire :

$$h \leq \frac{1}{2f_m} \tag{9}$$

f_m étant la fréquence la plus haute apparaissant dans le signal étudié. Mais cette condition (9) est une condition nécessaire et suffisante afin de reconstituer entièrement ce signal à partir des points de discrétisation obtenus, mais ce n'est pas une condition nécessaire pour connaître la période. En effet : prenons l'exemple : (figure I.8)

$$f(t) = \sin t + a \sin 2t$$

* $|2a| < 1$, avec le théorème de Shannon il est nécessaire d'avoir au moins quatre points sur $[0,2\pi]$

avec la méthode précédente, on peut prendre deux points, 2, 5 , 3 et une extrapolation à la limite donnera de très bons résultats.

* $|2a| \geq 1$, on voit qu'on peut construire, en choisissant bien a, deux zéros impairs de f aussi voisins que l'on veut l'un de l'autre, dans ce cas cette méthode semble défavorable.

Figure I.8

Cette méthode peut donc être très "puissante", mais elle peut aussi exiger un pas très fin.

I.3. - SIGNAL PERIODIQUE OU NON PERIODIQUE

Soit la fonction définie sur $(-\infty,+\infty)$ par

$$f(t) = \sum_{k=1}^{\infty} a_k \cos(\omega_k t + \varphi_k) \qquad (10)$$

telle que

i) $\{a_k\} \in \ell_1$

ii) $1 \leq \Omega \leq \omega_1 < \omega_2 \ldots < \omega_p < \ldots$

A cause de l'hypothèse i), f(t) est une fonction presque périodique [2] et appartient au sous espace vectoriel des fonctions presque périodiques développables en série de Fourier généralisée.

On peut écrire à cause de l'absolue convergence des séries :

$$f(t) = \sum_{1}^{\infty} a_k \cos \varphi_k \cos \omega_k t - \sum_{1}^{\infty} a_k \sin \varphi_k \sin \omega_k t$$

Posant $\alpha_k = a_k \cos \varphi_k$, on préfère considérer la partie "paire" du signal (10) soit le signal :

$$g(t) = \frac{f(t) + f(-t)}{2} = \sum_{1}^{\infty} \alpha_k \cos \omega_k t \qquad (11)$$

Il est évident que les paramètres de (11) vérifient aussi les propriétés i) et ii).

3.1. TRANSFORMEE DE HILBERT

Soit une fonction s(t) définie ∀t. On appelle transformée de Hilbert de s(t) [8], la fonction $\hat{s}(t)$ — lorsqu'elle existe — définie par la valeur principale de Cauchy de l'intégrale :

$$\hat{s}(t) = \frac{1}{\pi} \text{v.p.} \int_{-\infty}^{+\infty} \frac{s(\tau)}{t-\tau} d\tau \qquad (12)$$

On peut montrer [60] le résultat suivant :

Si $s(t) \in L_p(-\infty,+\infty)$ $(p > 1)$ alors la formule (12) définit presque partout une fonction $\hat{s}(t)$, qui appartient aussi à L_p , dont la transformée de Hilbert, est :

$$\hat{\hat{s}}(t) = -s(t)$$

Exemple de transformée de Hilbert

$s(t) = \cos \omega t \qquad (\omega > 0)$

$$\hat{s}(t) = \frac{1}{\pi} \text{v.p} \int_{-\infty}^{+\infty} \frac{\cos\omega(t-\tau)}{\tau} d\tau = \frac{1}{\pi} \text{v.p} \left[\int_{-\infty}^{+\infty} \frac{\cos \omega t \cos \omega\tau}{\tau} d\tau + \int_{-\infty}^{+\infty} \frac{\sin \omega t \sin \omega\tau}{\tau} d\tau \right]$$

Les intégrales $\int_{\varepsilon A}^{+\varepsilon\infty} \frac{\sin \omega t}{t} dt$ et $\int_{\varepsilon A}^{+\varepsilon\infty} \frac{\cos \omega t}{t} dt$ sont convergentes ($\varepsilon = \pm 1$)

et $\text{v.p.} \int_{-\infty}^{+\infty} \frac{\sin \omega\tau}{\tau} d\tau = \int_{-\infty}^{+\infty} \frac{\sin \omega\tau}{\tau} d\tau$ puisque $\lim_{x \to 0} \frac{\sin x}{x} = 1$

donc :

$$\pi\hat{s}(t) = \cos\omega t \ \text{v.p.} \int_{-\infty}^{+\infty} \frac{\cos\omega\tau}{\tau} d\tau + \sin \omega t \int_{-\infty}^{+\infty} \frac{\sin \omega\tau}{\tau} d\tau$$

$$\text{v.p.} \int_{-\infty}^{+\infty} \frac{\cos \omega\tau}{\tau} d\tau = \lim_{\alpha \to 0} \int_{-\infty}^{-\alpha} \frac{\cos \omega\tau}{\tau} d\tau + \lim_{\alpha \to 0} \int_{\alpha}^{+\infty} \frac{\cos \omega\tau}{\tau} d\tau = 0$$

soit :

$$\hat{s}(t) = \frac{1}{\pi} \sin \omega t \int_{-\infty}^{+\infty} \frac{\sin \omega u}{u} du$$

puisque $\int_{0}^{\infty} \frac{\sin \omega u}{u} du = \int_{0}^{\infty} \frac{\sin u}{u} du = \frac{\pi}{2} \qquad (\omega > 0)$

$$\hat{s}(t) = \sin \omega t$$

On pourrait montrer de même que :

$$s(t) = \sin \omega t \longleftrightarrow \hat{s}(t) = - \cos \omega t$$

3.2. SIGNAL ANALYTIQUE - SIGNAL EN QUADRATURE

Soit $\Phi(z)$ une fonction analytique de la variable complexe z, régulière dans le demi-plan $Im(z) > 0$

On pose $\Phi(x+io) = u(x) + iv(x)$

On a alors sous certaines conditions, voir [60]

$$u(x) = \frac{1}{\pi} \mathbf{v.p} \int_{-\infty}^{+\infty} \frac{v(t)}{t-x} dt$$

$$v(x) = -\frac{1}{\pi} v.p \int_{-\infty}^{+\infty} \frac{u(t)}{t-x} dt$$

Donc connaissant u(x) , on lui associe une autre fonction v, de telle manière que u+iv soit la valeur d'une fonction analytique sur la droite réelle.

La valeur de cette fonction $\Phi(z)$ sur la droite réelle est appelée par Ville [63] : le signal analytique . Ce signal analytique permettant des développements intéressants en particulier définition d'une enveloppe d'un signal, dans certains cas d'une fréquence instantanée... Le signal v(x) est appelé signal en quadrature de u(x) , la signification de cette appellation est évidente en se reportant à l'exemple $s(t) = \sin \omega t$.

3.3. ALGORITHME DE CALCUL DU FONDAMENTAL

Soit V l'ensemble des fonctions f telles que :

a) f admet un développement $\sum_{1}^{\infty} a_k \cos \omega_k t$

i) $\{a_k\} \in \ell_1$

ii) $1 \leq \Omega \leq \omega_1 < \omega_2 \ldots \omega_p < \omega_{p+1} \ldots$

On peut définir de même V_1, ensemble des fonctions f telles que :

b) f admet un développement $\sum_1^\infty a_k \sin \omega_k t$

i) $\{a_k\} \in \ell_1$

ii) $1 \leq \Omega \leq \omega_1 < \omega_2 \ldots$

Soit P l'opérateur défini sur V par

$$h \in V \xrightarrow{P} Ph$$

avec $$Ph(t) = \int_0^t h(x)dx$$

et soit H l'opérateur défini sur V_1 par :

$$k \in V_1 \xrightarrow{H} Hk$$

avec :

$$Hk(t) = \frac{1}{\pi} v.p \int_{-\infty}^{+\infty} \frac{k(u)}{t-u} du$$

On se propose d'étudier quelques propriétés de la transformation, si elle est définie, HP .

Soit $h \in V$, on a :

$$h(t) = \sum_1^\infty a_k \cos \omega_k t$$

et à cause de l'absolue convergence de la série $\{a_k\}$

$$\int_0^t \sum_1^\infty a_k \cos \omega_k t \, dt = \sum_1^\infty a_k \int_0^t \cos \omega_k t \, dt = \sum_1^\infty \frac{a_k}{\omega_k} \sin \omega_k t$$

Puisque :

$$\forall k \quad \left|\frac{a_k}{\omega_k}\right| \leq \left|\frac{a_k}{\Omega}\right| \leq |a_k|$$

donc $\forall h(t) \in V \Rightarrow Ph(t) \in V_1$

On peut donc définir HPh $\forall h \in V$

On a alors :

$$HPh(t) = \frac{1}{\pi} v.p \int_{-\infty}^{+\infty} \frac{\sum_1^\infty \frac{a_k}{\omega_k} \sin \omega_k u}{t-u} du$$

La série $\{\frac{a_k}{\omega_k}\}$ étant absolument convergente, alors $\Sigma \frac{a_k}{\omega_k} \sin \omega_k u$ est continue et la valeur principale de Cauchy existe $\forall t$, et on peut aussi intervertir signe Σ et $\int$, donc :

$$HPh(t) = \sum_1^\infty \frac{a_k}{\omega_k} \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{\sin \omega_k u}{t-u} du = \sum_1^\infty \frac{a_k}{\omega_k} \cos \omega_k t \qquad (13)$$

Donc HP est une application de V dans V.

On pose :

$$G^{(0)}(t) = \sum_1^\infty a_k \cos \omega_k t$$

$$G^{(p)}(t) = HP\left[G^{(p-1)}(t)\right] \qquad (p=1,2,\ldots) \qquad (14)$$

On a alors :

THEOREME :

Presque partout en t, $\lim_{p\to\infty} \frac{G^{(p)}(t)}{G^{(p+1)}(t)} = -\omega_1$

D'après (14) on peut écrire :

$$G^{(p)}(t) = HP\left[G^{(p-1)}(t)\right] = HP^2\left[G^{(p-2)}(t)\right]$$

ce qui entraîne que :

$$G^{(p)}(t) = HP^i\left[G^{(p-i)}(t)\right] \qquad \forall i$$

donc :

$$G^{(p)}(t) = HP^p(G^{(0)}(t))$$

D'après le résultat (13)

$$HP\left(\sum_1^\infty a_k \cos \omega_k t\right) = - \sum \frac{a_k}{\omega_k} \cos \omega_k t$$

en itérant on obtient immédiatement :

$$HP^p\left(\sum a_k \cos \omega_k t\right) = (-1)^p \sum_1^\infty \frac{a_k}{\omega_k{}^p} \cos \omega_k t$$

Formons :

$$\frac{G^{(p)}(t)}{G^{(p+1)}(t)} = \frac{\sum_1^\infty (-1)^p \frac{a_k}{\omega_k^p} \cos \omega_k t}{\sum_1 (-1)^{p+1} \frac{a_k}{\omega_k{}^{p+1}} \cos \omega_k t} = -\omega_1 \frac{a_1 \cos\omega_1 t + \sum_2^\infty a_k \left(\frac{\omega_1}{\omega_k}\right)^p \cos \omega_k t}{a_1 \cos\omega_1 t + \sum_2^\infty a_k \left(\frac{\omega_1}{\omega_k}\right)^{p+1} \cos \omega_k t}$$

puisque d'après ii) :

$$\frac{\omega_1}{\omega_k} < 1 \qquad \forall k > 1$$

on obtient si $\omega_1 t \neq (2j+1)\frac{\pi}{2}$ $\qquad \forall j \in \mathbb{Z}$

$$\lim_{p \to \infty} \frac{G^{(p)}(t)}{G^{(p+1)}(t)} = - \omega_1$$

ce qui démontre le résultat.

3.4. UTILISATION PRATIQUE

Nous allons voir que la difficulté essentielle de la transformée de Hilbert réside dans le fait que le signal g(t) est connu uniquement sur un intervalle fini $[-T,T]$.

On pose :

$$h(t) = \begin{cases} 1 & t \in [-T,T] \\ 0 & t \notin [-T,T] \end{cases} \qquad (T > 0)$$

et soit l'exemple simple suivant :

$$g(t) = \sin \omega t \qquad t \in (-\infty,\infty)$$

le signal réel à traiter étant :

$$f(t) = h(t)g(t)$$

On voit immédiatement que :

$$Hg(\pm T) = -\cos \omega T$$

$$Hf(\pm T) = \int_{-T}^{T} \frac{\sin \omega t}{t-T}\, dt = \begin{cases} \text{si } \omega T \neq k\pi \text{ ,n'est pas définie} \\ \text{valeur finie si } \omega T = k\pi \end{cases}$$

Comme les ω_k du signal initial sont inconnus , il est en général impossible de choisir T multiple de $\frac{\pi}{\omega_k}$ (dans la mesure où ces $\frac{\pi}{\omega_k}$ admettent un multiple commun).

Aussi on peut se poser le problème : T étant fixé, sur quel intervalle peut-on espérer

$$\max_{x \in [-T,T]} |H(g-f)(x)| < \varepsilon \quad ?$$

Soit :

$$G(x) = \pi Hg(x) = \int_{-\infty}^{+\infty} \frac{\sin \omega t}{x-t}\, dt$$

$$G_T(x) = \pi Hf(x) = \int_{-T}^{T} \frac{\sin \omega t}{x-t}\, dt$$

Posons :

$$R_T(x) = |G(x)-G_T(x)| = \left|\int_T^{\infty} \sin \omega t \left[\frac{1}{t-x} + \frac{1}{t+x}\right]dt\right|$$

Ce qui donne immédiatement :

$$\omega R_T(x) = \left|\cos \omega T \left(\frac{1}{T-x} + \frac{1}{T+x}\right) - \int_T^{\infty} \cos \omega t \left[-\frac{1}{(t-x)^2} + \frac{1}{(t+x)^2}\right]dt\right|$$

donc :

$$\omega R_T(x) \leq \left|\frac{2T}{T^2-x^2} + \frac{1}{T-x} + \frac{1}{T+x}\right|$$

soit encore :

$$R_T(x) \leq \frac{1}{\Omega} \quad \frac{4T}{|T^2-x^2|}$$

Si nous désirons $R_T(x) \leq \varepsilon$, il suffit que :

$$\frac{1}{\Omega} \left|\frac{4T}{T^2-x^2}\right| \leq \varepsilon \text{ (figure I.9)}$$

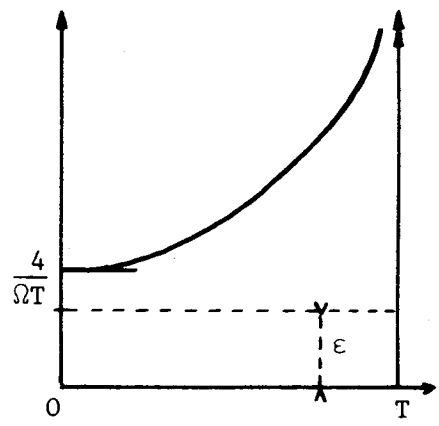


Figure I.9

Puisque nous supposons $0 < x < T$, traçons :

$$y = \frac{4T}{\Omega(T^2-x^2)}$$

On voit immédiatement que

$|y| < \varepsilon$ n'est pas réalisé $\forall T$ fixé

On doit avoir (figure I.10)

$\frac{4}{\Omega T} < \varepsilon$ dans ce cas on peut effectivement calculer l'intégrale de Hilbert sur un intervalle $[0,x']$ avec une précision ε . La méthode proposée étant itérative, sur la transformée de Hilbert, il sera en général impossible de trouver T (à moins de prendre $T = +\infty$) de telle sorte que $HP^{(p)}(g(t))$ soit calculable précisément.

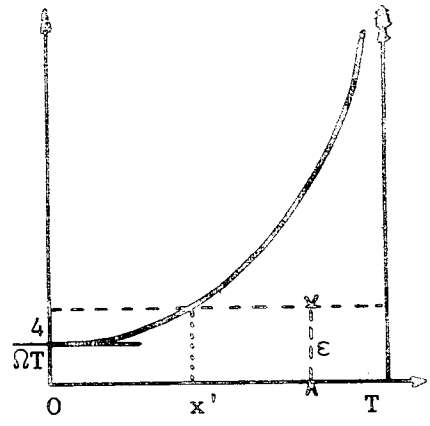


Figure I.10

3.5. CALCUL DE LA TRANSFORMEE DE HILBERT D'UNE FONCTION PERIODIQUE

On peut remarquer, que, si g(t) est un signal périodique de période connue, en choisissant "correctement" T, on peut rendre la différence entre la transformée de Hilbert et celle du signal tronqué arbitrairement petite.

3.5.1. Cas où $g(t) = \cos \omega t$

Posant à nouveau :

$$G(x) = \int_{-\infty}^{+\infty} \frac{\cos \omega t}{x-t}\, dt \quad \text{et} \quad G_T(x) = \int_{-T}^{T} \frac{\cos \omega t}{x-t}\, dt$$

on a :

$$R_T(x) = |G(x)-G_T(x)| = \left|\int_T^{\infty} \cos \omega t \left[\frac{1}{t-x} - \frac{1}{t+x}\right] dt\right| = \left|\int_T^{\infty} \frac{\cos \omega t}{t^2-x^2}\, dt\right| 2x$$

Posons $T = \frac{2\pi}{\omega}$ période du phénomène étudié :

et supposons $T > T$, et $0 \leq x \leq T$

alors : $R_T(x) \leq 2x \log\left|\frac{T+x}{T-x}\right|$

Si on désire trouver un intervalle $[0,x']$, tel que $R_T(x) \leq \varepsilon$
$\forall x \in [0,x']$
il suffit donc de prendre $x' = X$
(figure I.11).

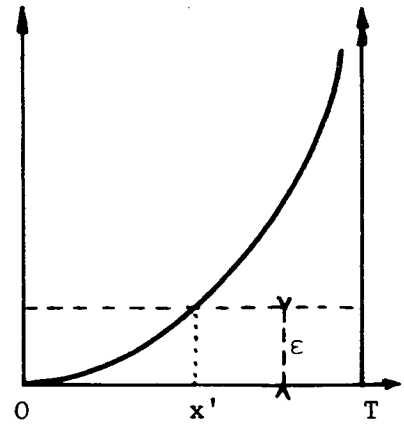


Figure I.11

Puisque la fonction — transformée de Hilbert de $\cos\omega t$ — est périodique, il suffit d'obtenir $X \geq T$ donc il suffit de choisir T assez grand.

3.5.2. Cas où $g(t) = \sin \omega t$

En posant : $$G(x) = \int_{-\infty}^{+\infty} \frac{\sin \omega t}{x-t}\, dt \quad \text{et} \quad G_T(x) = \int_{-T}^{T} \frac{\sin \omega t}{x-t}\, dt$$

Cette fonction g(t) étant périodique, et sa transformée de Hilbert, aussi, il suffit d'après la figure I.10 d'avoir $x' > T$ mais puisque x' est donné par la solution de :

$$\frac{4T}{\Omega} \frac{1}{T^2-x^2} \leq \varepsilon$$

il suffit de choisir T tel que $T^2 - \frac{4T}{\varepsilon\omega} \geq T^2$

On peut donc, un signal périodique étant donné,

$$g(t) = \sum_{1}^{\infty} a_k \cos(k\omega t + \varphi_k) \qquad \{a_k\} \in \ell_1 ,$$

calculer le signal en quadrature correspondant.

REMARQUE :

Lorsqu'un signal périodique est de période connue (2π), on utilise plutôt les formules suivantes de Hilbert :

$$\cos\theta = \frac{1}{2\pi}\, v.p \int_{-\infty}^{\infty} \operatorname{cotg}\frac{\phi-\theta}{2} \sin\phi \, d\phi$$

$$\sin\theta = \frac{1}{2\pi}\, v.p \int_{-\pi}^{\pi} \operatorname{cotg}\frac{\phi-\theta}{2} \cos\phi \, d\phi$$

pour un calcul numérique de ces formules voir [12] .

3.6. CONCLUSIONS SUR CETTE METHODE

L'écueil essentiel est évidemment son impossibilité actuelle de l'essayer numériquement. Avant de terminer, il faut souligner l'analogie avec la méthode de la puissance en analyse numérique. Il semble naturel que — dans l'exemple de signaux audibles — des fréquences seront d'autant plus vite repérées qu'elles seront mieux séparées, ce que montre cette méthode de la puissance. Il eut, dans ce cas, été fort intéressant d'étudier l'influence de la longueur d'enregistrement sur la détermination du fondamental au point de vue vitesse, précision... .

CHAPITRE II

PARTIE TRANSITOIRE DE QUELQUES SIGNAUX

Certains phénomènes physiques peuvent être uniquement constitués de transitoires, (exemple : étude d'un choc, mouvement d'une corde excitée par une impulsion très brève...) d'autres peuvent avoir une partie transitoire suivie d'un régime stationnaire (exemple : établissement d'un courant dans un circuit,...).

On se propose, dans ce chapitre, de donner des méthodes de calcul de "fondamentaux" quoique les signaux ne soient pas, à proprement parler, périodiques.

II.1 - NATURE DES SIGNAUX TRAITES

Soit E un ensemble de fonctions donné : (exemple: espace vectoriel des polynômes,...). Soit :

$$\Omega = \{\omega/\omega_1,\omega_2,\ldots,\omega_n,\ldots \; ; \; \omega_p > \omega_{p-1} > 0 \quad \forall p > 0\}$$

On considère l'espace vectoriel V des signaux de la forme [17][31] pour $t \in [a,b]$

$$f(t) = \sum_{k=1}^{\infty} h_k(t) \cos(\omega_k t + \varphi_k)$$

La fréquence $f_1 = \frac{\omega_1}{2\pi}$ est appelée fréquence fondamentale.

Soit encore :

$$\Omega_H = \{\omega_p \; / \; \omega_p = p\omega \qquad p=1,2,\ldots\}$$

.On étudiera les trois cas suivants :

a) E ensemble des polynomes de degré 0 et $\Omega = \Omega_H$.

Dans ce cas, V est l'espace vectoriel des fonctions périodiques (Il est alors à remarquer que dans ce cas la partie étudiée du signal n'est pas transitoire, et cette étude aurait dû être placée dans le chapitre précédent).

b) E ensemble des polynomes de degré $\leq M$ et $\Omega = \Omega_H$.

c) E ensemble des fonctions qui sont une combinaison linéaire finie d'exponentielles (réelles), et Ω est un ensemble fini.

II.2 - ALGORITHME DE CALCUL DU FONDAMENTAL DANS LES CAS a) ET b)

Soit $f \in V$, donc , puisque $\Omega = \Omega_H$,

$$f(t) = \sum_{k=1}^{\infty} h_k(t) \cos(k\omega t + \varphi_k)$$

Puisque, pour $t \in [a,b]$, $h_k(t)$ est un polynome de degré $\leq M$

$$h_k(t) = \sum_{i=0}^{M} a_{ki}\, t^{M-i} \qquad k=1,2,\ldots$$

On suppose désormais que les suites :

$$\{a_{ki}\} \in \ell_1 \qquad \forall i = 0,1,2,\ldots,M .$$

On a alors :

$$f(t) = \sum_{i=0}^{M} \Big(\sum_{k=1}^{\infty} a_{ki} \cos(k\omega t + \varphi_k) \Big) t^{M-i}$$

Si on pose :

$$\psi_i(t) = \sum_{k=1}^{\infty} a_{ki} \cos(k\omega t + \varphi_k) \qquad i=0,\ldots,M$$

on obtient :

$$f(t) = \sum_{i=0}^{M} \psi_i(t)\, t^{M-i} \quad .$$

(Avec ψ_i fonction de période $T = 2\pi/\omega$) .

2.1. ALGORITHME DE CALCUL DE LA PERIODE T DES FONCTIONS ψ_i [69]

Soit G_n le maillage de $[a,b]$ défini par :

$$G_n = \{t \;/\; t_i = a+ih \;,\; i=0,1,\ldots,n \;;\; h = (b-a)/n\}$$

Soit $\tilde{t} \in G_n$, point quelconque, mais fixé, et soient les quantités

$$\xi_j(u) = f(\tilde{t}+ju) \qquad j=0,1,\ldots,N_1 = \left[\frac{b-\tilde{t}}{u}\right] \quad , \qquad (1)$$

où u désigne un paramètre quelconque ($\in \mathbb{R}^+$)

et $[x]$ = plus petit entier contenu dans x.

PROPOSITION 1 :

Si $u = kT$ ($\forall k \in \mathbb{Z}$), il existe un polynome de degré M passant par les points (j,ξ_j) $j=0,1,\ldots,N_1$ $(N_1 > M)$

Posons :

$$P_M(x) = \sum_{i=0}^{M} \psi_i(\tilde{t})(\tilde{t}+x)^{M-i} = \sum_{i=0}^{M} \beta_i\, x^{M-i} \qquad (2)$$

Calculons :

$$\xi_j(kT) = f(\tilde{t}+jkT) = \sum_{i=0}^{M} \psi_i(\tilde{t}+jkT)(\tilde{t}+jkT)^{M-i}$$

Puisque les fonctions ψ_i sont périodiques et de période T

$$\psi_i(\tilde{t+jkT}) = \psi_i(\tilde{t}) \qquad \forall \tilde{t}$$

donc :

$$\xi_j(kT) = \sum_{i=0}^{M} \psi_i(\tilde{t})(\tilde{t+jkT})^{M-i} = P_M(jkT) \qquad \forall j=0,\ldots,N_1$$

ce qui démontre la proposition.

On en déduit immédiatement un moyen de calcul de T :
Soit la fonction de u :

$$R^2(u) = \sum_{j=0}^{N_1} (\xi_j(u) - P_M(ju))^2 \geq 0 \qquad \forall u \qquad (3)$$

La proposition 1 entraine immédiatement que :

$$R^2(kT) = 0 \ .$$

Pour chercher la période des fonctions ψ_i , c'est-à-dire T , on peut penser faire une tabulation de $R^2(u)$ pour différentes valeurs de u , et chercher les minima de $R^2(u)$.

(Il est à remarquer que P_M se calcule "relativement" simplement en effet, c'est le polynome d'interpolation de degré M passant par les points $(j,\xi_j(u) \quad j=0,\ldots,M < N_1)$. Ces calculs — à faire pour chaque u sont en général longs, aussi pratiquement on prendra une autre méthode, basée encore sur la proposition 1.

On sait, en effet, que, si toutes les différences d'ordre M+1 construites sur les points $(x_i,y_i) \quad i=0,\ldots,N_1$ $(N_1 > M)$ sont nulles, alors les valeurs y_i sont les valeurs que prend un polynome de degré M aux points x_i (et réciproquement) [55] .

On considère alors la quantité :

$$L^2(u) = \frac{1}{N_1 - M} \sum_{j=0}^{N_1-M-1} \left[\Delta_j^{M+1}(u)\right]^2$$

où Δ_j^k sont les différences non divisées progressives d'ordre k, c'est-à-dire :

$$\begin{cases} \Delta_j^k(u) = \Delta_{j+1}^{k-1}(u) - \Delta_j^{k-1}(u) \\ \Delta_j^0(u) = f(\tilde{t}+ju) \end{cases}$$

On a évidemment $L^2(kT) = 0$. Et on cherche T (ou ses multiples) au moyen d'une tabulation.

PROPOSITION 2 :

Une condition nécessaire d'utilisation de la méthode précédente est $(b-a) > MT$ *M degré du polynome.*

Puisque $N_1 > M$, on a en utilisant la définition de N_1 :

$N_1 = \left[\frac{b-\tilde{t}}{u}\right]$, avec $\tilde{t} = a$, et $u = T$

on obtient :

$$\left[\frac{b-a}{T}\right] = N_1 > M$$

d'où le résultat, qui signifie que plus le signal est complexe (M grand) plus la longueur d'enregistrement devra être grande.

2.2. RESULTATS NUMERIQUES

On a représenté ici deux exemples avec $M = 4$.
Sur la figure II.1 sont représentés les polynomes $h_k(t)$, $k=1,\ldots,4$.
Les deux signaux traités $f(t)$ apparaissent sur les figures II.2 et II.4, signaux qui sont la somme des quatre partiels $h_k(t) \cos (k\omega t + \varphi_k)$, lorsque $T = 2\pi/\omega = 5\ 10^{-2}$ seconde .

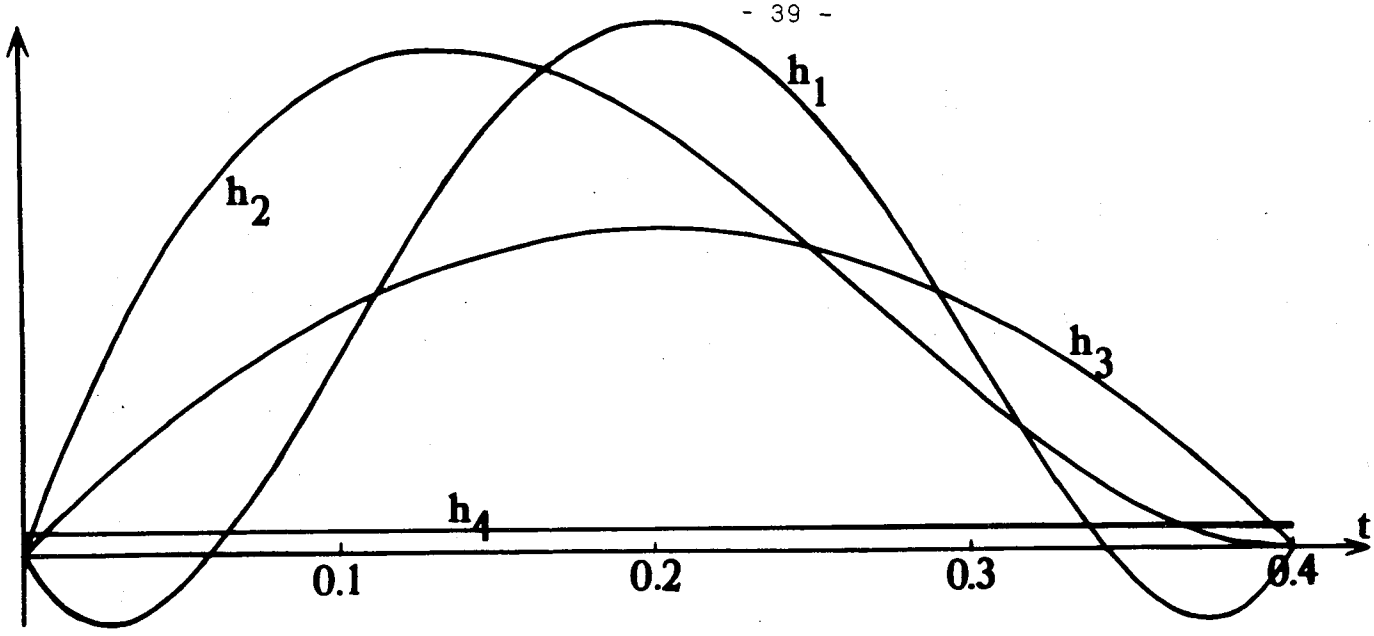


figure II .1. Polynomes $h_k(t)$ $(k=1,..,4)$

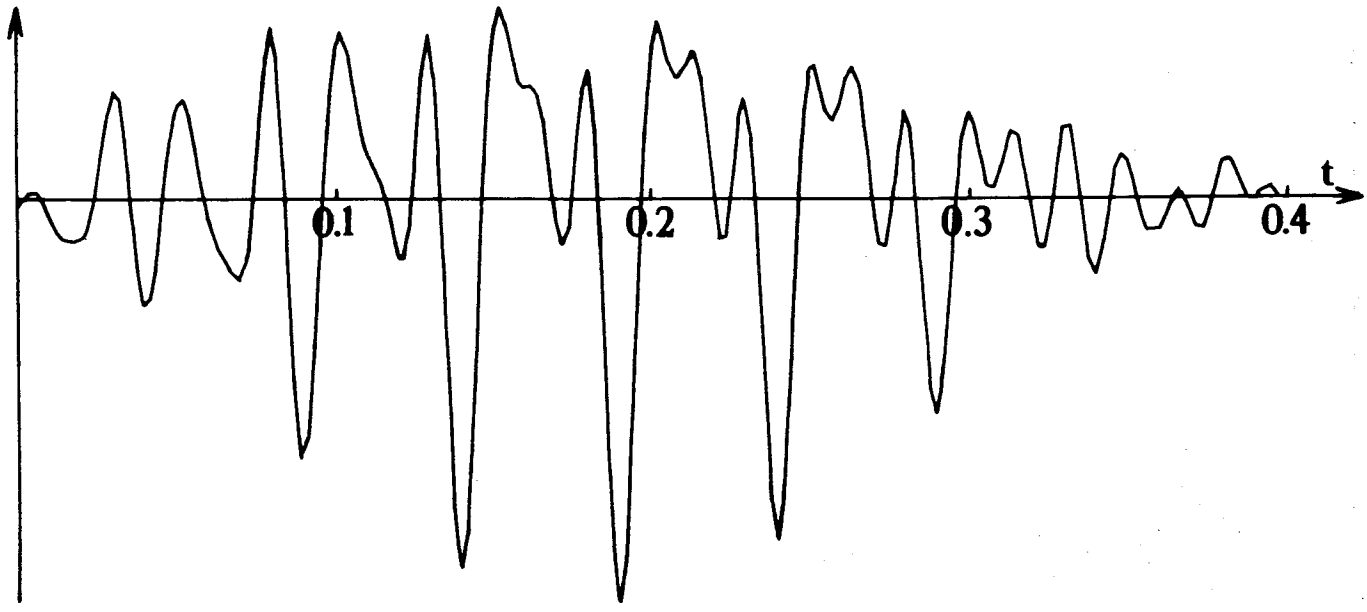


figure II .2. Signal: f(t)

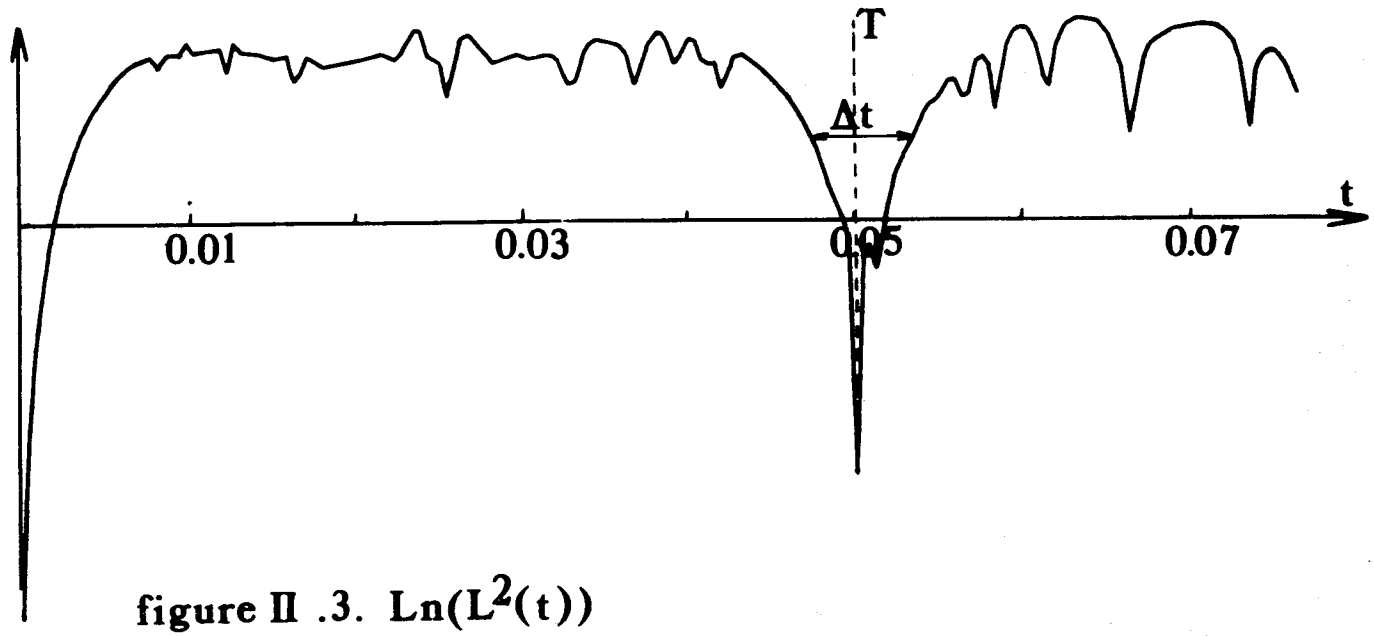


figure II .3. $Ln(L^2(t))$

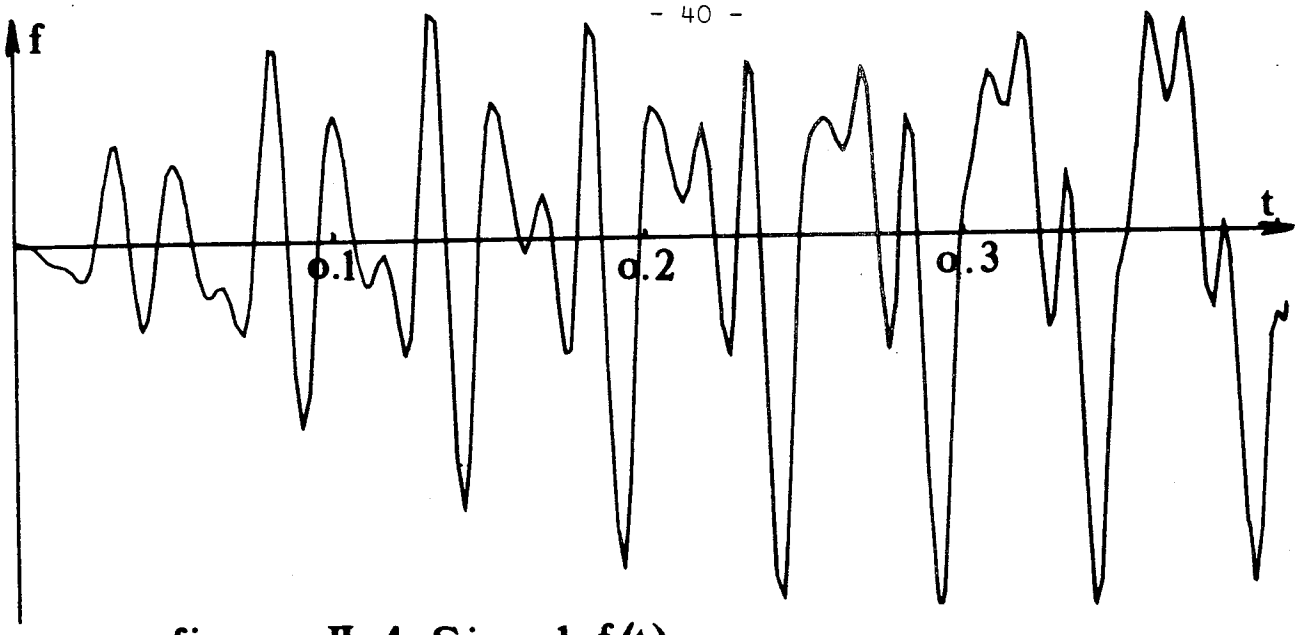


figure II.4. Signal f(t)

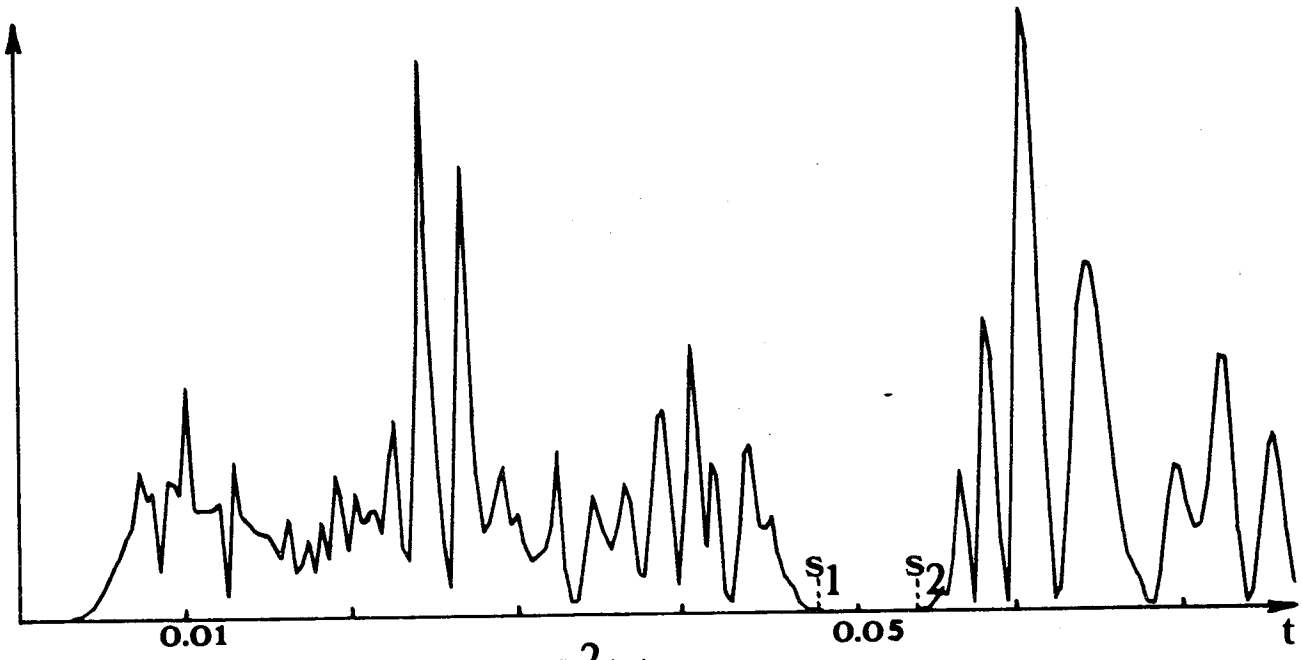


figure II.5. $L^2(t)$

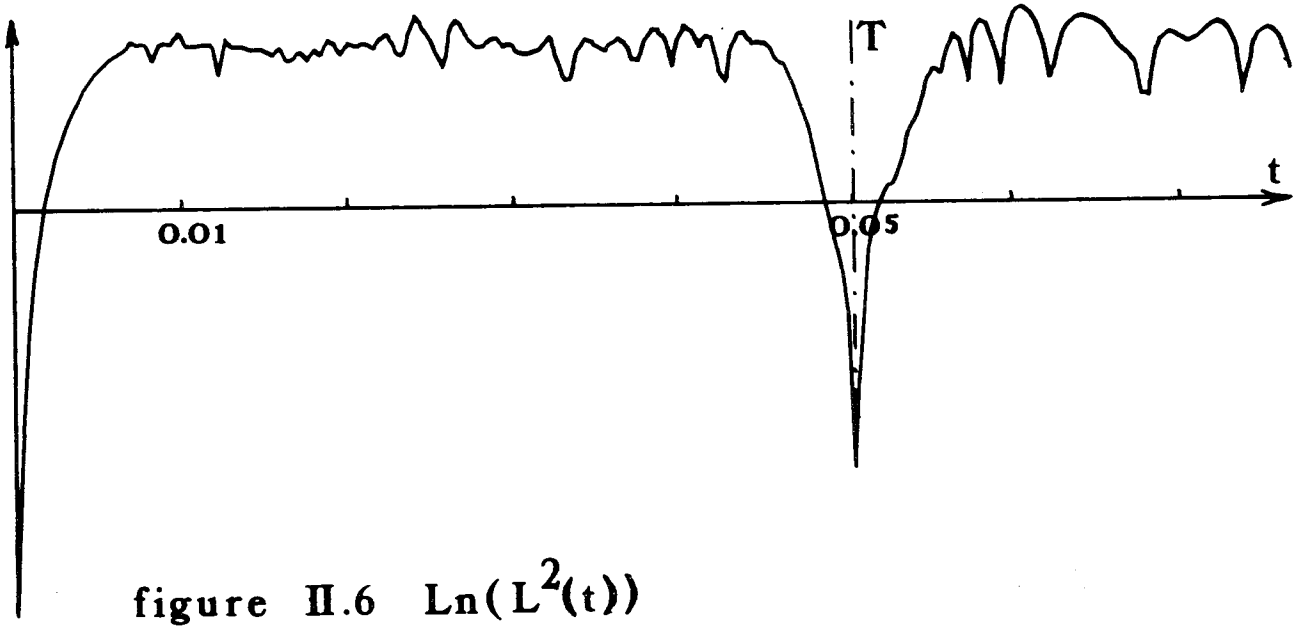


figure II.6 $Ln(L^2(t))$

La figure II.5 montre le tracé de $L^2(u)$, il faut remarquer que la période T n'est pas du tout "lisible", on peut juste dire que T appartient à l'intervalle : $[s_1,s_2]$. Au lieu de $L^2(u)$, qui en général présente de fortes variations, on a porté sur les figures II.3 et II.6 une quantité proportionnelle à Ln $(L^2(t))$ qui fait immédiatement ressortir le minimum qui apparaît en $t = 5\ 10^{-2}$ seconde $(t \neq 0)$.

(Tous ces dessins sortent automatiquement sur un traceur Benson).

REMARQUES :

* Pas d'échantillonage : Il peut être choisi convenablement dès que l'on possède une certaine expérience de la méthode. Dans le premier exemple une analyse "a posteriori" montre qu'en choisissant $h \leq \Delta t$ on est sûr d'avoir un minimum significatif.

* Nombre de périodes à considérer : la période T n'étant pas connue, ce nombre est difficile à évaluer, mais il semble que la précision sur T sera d'autant plus grande que ce nombre sera plus élevé.

* Intervalle d'étude : les figures II.3 et II.6 ont été tracées pour $t \in [0,0.07]$, mais on peut — et on doit — se limiter à un intervalle plus petit si des informations supplémentaires sont fournies. Exemple, si on sait que $T \in [T_m,T_M]$, il suffit d'étudier $L_n(L^2(u))$ pour $u \in [T_m,T_M]$.

2.3. CALCUL DES POLYNOMES $h_k(t)$

La période de T étant, désormais, supposée calculée, on connait la valeur du polynome $P_M(x)$ aux points $x = jT$ $j=0,\ldots,M$, ce qui nous permet de connaitre les coefficients β_i (2) (en utilisant la formulation de Lagrange, ou de Newton du polynome d'interpolation).

On choisit $\tilde{t} = t_o$. On a alors d'après (2)

$$\sum_{i=0}^{M} \psi_i(t_o)(t_o+x)^{M-i} = \sum_{0}^{M} \beta_i \, x^{M-i} \qquad (2)$$

Ce qui nous permet de calculer $\psi_i(t_o)$ pour $i=0,1,\ldots,M$

Dans (2) posons $(t_o+x) = X$

$$\sum_{i=0}^{M} \psi_i(t_o)X^{M-i} = \sum_{i=0}^{M} \beta_i(X-t_o)^{M-i} = \sum_{i=0}^{M} \beta_i \sum_{i=0}^{M-i} (-t_o)^k \, X^{M-i-k}$$

donc en identifiant :

$$\begin{cases} \psi_o(t_o) = \beta_o \\ \psi_1(t_o) = -t_o \, C_M^1 \, \beta_o + \beta_1 \\ \vdots \\ \psi_k(t_o) = \sum_{i=0}^{k} \beta_i \, C_{M-i}^{k-i} \, (-t_o)^{k-i} \\ \psi_M(t_o) = (-t_o)^M \, \beta_o + (-t_o)^{M-1} \, \beta_1 + \ldots + \beta_M \end{cases}$$

Le même procédé est alors appliqué en prenant une nouvelle valeur initiale t_k , ce qui permet alors de calculer

$$\psi_i(t_k) \qquad i=0,1,\ldots,M \quad , \quad k=1,2,\ldots$$

Connaissant ainsi $\psi_i(t_j)$ $\begin{cases} i=0,\ldots,M \\ j=0,\ldots,N_2 \end{cases}$ une méthode quelconque (analyse de Fourier, moindres carrés,...) permet alors d'avoir une estimation des a_{ki} et des φ_k et par là même une approximation du polynome $h_k(t)$, $k=1,2,\ldots$.

2.4. PROBLEME INVERSE

On peut poser le problème suivant :

si $L^2(u^*) = 0$ a-t-on $u^* = kT$? (k entier)

On peut répondre négativement. En effet, il existe un ensemble S de points $\tilde{t}$ tel que $L^2(u) = 0$, pour $u \neq kT$.

Contre-exemple :

$$f(t) = \sin t \qquad [a,b] \equiv (0,+\infty)$$

soit $\tilde{t} = a = 0$.

On a alors $L^2(u) = 0$ pour $u = k2\pi$, $\forall k = 0,1,\ldots$
mais aussi $L^2(u) = 0$ pour $u = k\pi$, $\forall k$ entier

mais $T = \pi$ n'est évidemment pas la période.

Les paragraphes suivants constituent une étude de l'ensemble S des points $\tilde{t}$ singuliers.

Si $L^2(u) = 0$, ou, ce qui est équivalent :

$$R^2(u) = \sum_{j=0}^{N_1} (\xi_j(u) - P_M(ju))^2 = 0$$

on doit avoir :

$$\xi_j(u) - P_M(ju) = 0 \qquad \forall j \text{ tel que } ju \in [a,b] \qquad (4)$$

On étudie séparément les cas : $M = 0$, et $M \neq 0$.

2.4.1. Etude du cas M = 0 , $(a,b) \equiv (-\infty,+\infty)$

$$\text{On pose } \psi(t) = f(t) = \sum_{k=1}^{\infty} a_k \cos(k\omega t + \varphi_k) \qquad \{a_k\} \in \ell_1$$

$\psi(t)$ est donc une fonction périodique, de période T, continue.

Soit t un nombre arbitraire mais fixé, on s'intéresse alors au système infini suivant (en u) (ceci résulte de (4))

$$\psi(t+ju) - \psi(t) = 0 \qquad \forall j \in \mathbb{Z} \tag{5}$$

On pose, pour t donné :

$$E(t,\psi) = \{\frac{u}{T} \;/\; \psi(t+ju) - \psi(t) = 0 \quad \forall j \in \mathbb{Z} \text{ et } u \neq kT \quad \forall k \in \mathbb{Z}\}$$

$$S(\psi) = \{t \;/\; t \in \mathbb{R} \;,\; E(t,\psi) \neq \emptyset\}$$

$S(\psi)$ sera appelé support de la solution du système (5), ou support de $E(t,\psi)$.

LEMME 1 :

Pour t donné, $E(t,\psi) \subseteq Q$, si ψ n'est pas une constante.

Supposons $\xi = \frac{u}{T}$ irrationnel et soit :

la fraction continue :

$$x = a_o + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ldots}} = [a_o, a_1, a_2, \ldots]$$

$x_n = [a_o, a_1, \ldots, a_n]$ est le $n^{\text{ième}}$ convergent ou la réduite d'ordre n de la fraction continue.

Si $\{a_i\}$ est une suite d'entiers satisfaisant : $a_i > 0 \quad \forall i$ [64]

alors $x_n \to x$

On peut encore écrire :

$$x_n = \frac{p_n}{q_n} \qquad (p_n,\ q_n \text{ entiers})$$

avec :

$$\begin{cases} p_o = a_o \\ q_o = 1 \end{cases} \qquad \begin{cases} p_1 = a_o\, a_1 + 1 \\ q_1 = a_1 \end{cases}$$

$$\begin{cases} p_n = a_n\, p_{n-1} + p_{n-2} & n \geq 2 \\ q_n = a_n\, q_{n-1} + q_{n-2} & n \geq 2 \end{cases}$$

On a le résultat suivant [23].

$\forall \eta$ irrationnel (ou non) $\exists r_n$, s_n entiers ($\frac{r_n}{s_n}$ $n^{\text{ième}}$ convergent d'une fraction continue) tels que :

$$0 < \left|\eta - \frac{r_n}{s_n}\right| < \frac{1}{s_n\, s_{n+1}} \qquad \text{avec} \qquad s_j \geq j \qquad \forall j$$

ce qui s'écrit encore :

$$0 < \left|\eta\, s_n - r_n\right| < \frac{1}{n+1}$$

Remplaçant η par $\xi = \frac{u}{T}$

$$0 < \left|u\, s_n - T\, r_n\right| < T \,/\, n+1 \qquad (6)$$

PROPOSITION 3 :

L'ensemble $A = \{x \ / \ u,T \text{ donnés } ; \quad x = qu + pT \ ; \ p,q \in \mathbb{Z}\}$

est dense dans $\mathbb{R}$.

Soit $Z \in [0,T]$ d'après (6) , $|us_n - T\, r_n|$ est aussi petit que l'on veut mais positif

Donc $\forall\, \varepsilon > 0$, $\exists$ M,N entiers tels que :

$$Z - \varepsilon \leq M|u\, s_N - T\, r_N| \leq Z + \varepsilon$$

soit :

$$|Z - |M\, s_N u - M\, r_N\, T|| \leq \varepsilon$$

si $s_N u - r_N T > 0$ on pose $q = M\, s_N$ $\quad p = -\, r_N\, M$

si $s_N u - r_N T < 0$ on pose $q = -Ms_N$ $\quad p = r_N\, M$

Si $Z \in [kT,(k+1)T]$ il suffit de prendre p = p+k. Donc la proposition 3 est démontrée dans le cas où $Z \in [kT,(k+1)T]$, donc $\forall k \in \mathbb{Z}$.

Donc l'ensemble $\{qu + pT : \frac{u}{T}$ irrationnel, q, p entiers$\}$ est dense dans $\mathbb{R}$.

La fonction ψ étant de période T, on peut écrire (système (5))

$$\psi(t_o + ju + kT) = \psi(t_o) \qquad \forall j,\, k \in \mathbb{Z}$$

D'après la proposition 3 :

$\forall\, \alpha > 0$, $\forall x \in \mathbb{R}$, $\exists$ p,q tel que $|x - (pu + qT)| < \alpha$

donc $\exists\, |\theta| \leq 1$ tel que $x - \theta\alpha = pu + qT$

Puisque $\psi(t)$ est continue :

$$\forall \varepsilon > 0 \quad \exists \eta \quad \text{tel que} \quad |h| < \eta \Rightarrow |\psi(t+h) - \psi(t)| < \varepsilon$$

On a aussi :

$$\psi(t_o + pu + qT) - \psi(t_o) = 0$$

Soit :

$$\psi(t_o + x - \theta\alpha) - \psi(t_o) = 0$$

$$[\psi(t_o + x - \theta\alpha) - \psi(t_o + x)] + [\psi(t_o + x) - \psi(t_o)] = 0$$

$$|\psi(t_o + x) - \psi(t_o)| = |\psi(t_o + x - \theta\alpha) - \psi(t_o + x)| \quad ,$$

si on choisit $\alpha = \eta$, d'après la continuité :

pour tout $\varepsilon > 0$, $|\psi(t_o + x) - \psi(t_o)| \leq \varepsilon \qquad \forall x$

On doit donc avoir $\psi(t)$ = constante.

Donc si $\psi(t) \neq c$, nous aurions nécessairement,

si $t \in S(\psi)$, ξ = un rationnel

soit $u = \dfrac{p_o}{q_o} T$ $\quad p_o$, q_o premiers entre eux ce qui démontre le lemme 1.

LEMME 2 :

Si $\dfrac{p_o}{q_o} \in E(t,\psi)$ *alors* $\dfrac{1}{q_o} \in E(t,\psi)$

Le système (5) s'écrit d'après le lemme 1 :

$$\psi(t_o + j \frac{p_o}{q_o} T + kT) - \psi(t_o) = 0 \qquad \forall j,k$$

p_o, q_o étant premiers entre eux d'après le théorème de Bezout $\exists$ j',k' tels que $j'p_o + k'q_o = 1$

Donc en posant $j = j'r$ $\qquad k = k'r$

$$\psi(t_o + r \frac{j'p_o + k'q_o}{q_o} T) = \psi(t_o) \qquad \forall r$$

donc :

$$\psi(t_o + \frac{r}{q_o} T) = \psi(t_o) \qquad \forall r$$

ce qui démontre le lemme 2.

Il est aussi intéressant de connaitre les propriétés du support $S(\psi)$ que celles de $E(t,\psi)$.

On pose :

$$F(u_i,\psi) = \{t : \psi(t+ju_i) - \psi(t) = 0 \quad \forall j \in \mathbb{Z}, \; u_i \neq kT \quad \forall k \in \mathbb{Z}\}$$

On a immédiatement :

$$S(\psi) = \bigcup_{i \in I} F(u_i,\psi) \qquad I \subseteq N \qquad \text{(d'après le lemme 2)}$$

Si $\forall u_i$ $\quad F(u_i,\psi)$ est dénombrable alors $S(\psi)$ est dénombrable.

On pose :

$$S_k(\psi) = \{t \;/\; t \in S(\psi) \;,\; t \in [kT,(k+1)T]\}$$

et $\qquad F_k(u_i,\psi) = \{t \;/\; t \in F(u_i,\psi) \;,\; t \in [kT,(k+1)T]\}$.

LEMME 3 :

On suppose de plus $\psi'(t)$ existe et est continue.
Si $F_k(u_i,\psi)$ est un ensemble infini , alors tout point d'accumulation de $F_k(u_i,\psi)$ appartient à $F_k(u_i,\psi')$.

Supposons $F_k(u_i,\psi)$ infini.

Puisque tout ensemble borné contenant une infinité de points possède au moins un point d'accumulation (Bolzano-Weierstrass), $F_k(u_i,\psi)$ possède au moins un point d'accumulation T .

Montrons que $T \in F_k(u_i,\psi)$

ψ étant continue $\forall x_1$

$\forall \varepsilon > 0 \quad \exists \eta$ tel que $|x_2-x_1| < \eta \Rightarrow |\psi(x_2) - \psi(x_1)| \leq \varepsilon$

Puisque T point d'accumulation :

$\forall \varepsilon' > 0 \quad \exists N$ tel que $\forall n > N \Rightarrow |T-t_n| \leq \varepsilon' \qquad (t_n \in F_k(u_i,\psi))$

prenant $\varepsilon' = \eta$

$$|\psi(t_n + ju_i) - \psi(T + ju_i)| \leq \varepsilon$$

$$|\psi(t_n) - \psi(T)| \leq \varepsilon$$

$$|\psi(t_n+ju_i)-\psi(T+ju_i)-\psi(t_n)+\psi(T)| \leq |\psi(t_n+ju_i)-\psi(T+ju_i)| + |\psi(t_n)-\psi(T)| \leq 2\varepsilon$$

donc $\forall \varepsilon \quad |\psi(T) - \psi(T + ju_i)| \leq 2\varepsilon \Rightarrow T \in F_k(u_i,\psi)$

Or a donc :

$$\psi(T + ju_i) - \psi(T) = 0$$

$$\psi(t_p + ju_i) - \psi(t_p) = 0 \qquad \forall t_p \in F_k(u_i,\psi)$$

$$\Rightarrow \quad \psi(t_p + ju_i) - \psi(T + ju_i) = \psi(t_p) - \psi(T)$$

$$(t_p - T)\psi'(T + ju_i + \theta_1(t_p - T)) = (t_p - T)\psi'(T + \theta_1'(t_p - T))$$

$$\theta_1 , \theta_1' \leq 1$$

si $p > N(\varepsilon')$ nous obtenons alors, puisque ψ' est supposée continue

$$\psi'(T + ju_i) = \psi'(T) \qquad \forall j$$

d'où le lemme 3 .

LEMME 4 :

Si $S(\psi) \cap S(\psi') = \emptyset$ alors $S(\psi)$ est dénombrable (ψ' supposée continue).

Si $F_k(u_i,\psi) \cap F_k(u_i,\psi') = \emptyset$

alors d'après le lemme 3 $F_k(u_i,k)$ fini ou encore

si $S_k(\psi) \cap S_k(\psi') = \emptyset \Rightarrow F_k(u_i,k)$ fini

Puisque $S_k(\psi) = \bigcup_{i\in I} F_k(u_i,\psi)$

on en déduit que $S_k(u_i,\psi)$, union dénombrable d'ensembles dénombrables, est dénombrable.

Puisque les $S_k(\psi)$ sont les translatés de $S_0(\psi)$

$$S_k(\psi) \cap S'_k(\psi) = \emptyset \iff S(\psi) \cap S(\psi') = \emptyset$$

puisque :

$$S(\psi) = \bigcup_{k=-\infty}^{+\infty} S_k(\psi)$$

On en déduit le lemme 4 .

La condition du lemme 4 est une condition suffisante, mais non nécessaire pour que $S(\psi)$ soit dénombrable.

Exemple :

$$y(t) = \sin \Omega t - \frac{1}{5} \sin 5\Omega t$$

On verra que S(y) est dénombrable (Théorème 1). Il est immédiat de voir que :

$$S(\psi) \cap S(\psi') \supseteq \{0, \frac{T}{2}\} \neq \emptyset$$

THEOREME 1 :

Si $\psi(t) = \sum_{k=1}^{N} \alpha_k \cos(k\omega t + \varphi_k)$ *N fixé*

alors le support du système

$$\psi(t + ju) - \psi(t) = 0$$

est dénombrable.

Soit $u = \bar{u} \in E(t,\psi)$

$$Y_{kj} = \{t \ / \ \psi(t+j\bar{u}) = \psi(t) \qquad t \in [kT,(k+1)T] \qquad j \text{ fixé}\}$$

et $$Y_k = \bigcap_{j=-\infty}^{+\infty} Y_{kj} \subseteq Y_{kp} \qquad \forall p$$

$H(t) = \psi(t+j\bar{u}) - \psi(t)$ est analytique sur $[kT,(k+1)T]$ donc [9]

$H(t)$ possède un nombre fini de racines sur le compact $[kT,(k+1)T]$ $\forall k$

Donc $S_{\bar{u}}(\psi) = \bigcup_{k=-\infty}^{+\infty} Y_k$ est dénombrable (union dénombrable d'ensembles dénombrables).

On a encore :

$$S(\psi) = \bigcup_{\bar{u} \in E} S_{\bar{u}}(\psi)$$

puisque E est dénombrable, $S(\psi)$ est aussi dénombrable.

2.4.2. Etude du cas $M \neq 0$, (a,b) fermé, borné

On pose ici :

$$\psi_i(t) = \sum_{k=1}^{N_i} a_{ki} \cos(k\Omega t + \varphi_k) \qquad (\text{avec } N_i \leq N \ , \ \forall i = 0,1,\ldots,M)$$

Soit $\psi^T = [\psi_0,\ldots,\psi_M]$, $\psi \in \mathbb{R}^{M+1}$

Pour t fixé, on pose :

$$Q_t(x) = \sum_{i=0}^{M} \psi_i(t)(t+x)^{M-i}$$

en se servant de (4)

et $\quad E(t,\psi) = \{u : f(t+ju) = Q_t(ju) \ ; \ u \neq kT \ ; \ ju \in [a,b] \ ; \ \forall k,j \in \mathbb{Z}\}$

On appelle à nouveau support de E, l'ensemble :

$$S(\psi) = \{t : E(t,\psi) \neq \emptyset\}.$$

On a alors :

THEOREME 2 :

$$Si \ \psi_i(t) = \sum_{k=1}^{N_i} a_{ki} \cos(k\omega t + \varphi_k) \quad , \quad N_i \leq N \quad \forall i \ ,$$

le support $S(\psi)$ est dénombrable.

Soit $Z_{jk}(t) = \{u : f(t+ju) = Q_t(ju) \ ; \ u \neq pT \ ; \ ju \in [kT,(k+1)T]\}$

Soient k_1, k_2 tels que : $\quad a \in [k_1T,(k_1+1)T]$

$$b \in [k_2T,(k_2+1)T]$$

et $\quad Z_j(t) = \bigcup_{k_1}^{k_2} Z_{jk}(t)$

On a alors immédiatement :

$$E(t,\psi) \subseteq \bigcap_{j\in J} Z_j(t) \subseteq Z_p(t) \qquad \forall p \qquad (J \subseteq \mathbb{Z})$$

Si on montre que $Z_{1k}(t)$ est dénombrable, il en résultera que, $Z_1(t)$, union dénombrable d'ensembles dénombrables est aussi dénombrable.

Soit $\quad G(u) = \sum_{i=0}^{M} (\psi_i(t+u) - \psi_i(t)) \ (t+u)^{M-i}$

Cette fonction G(u) est évidemment analytique sur $[kT,(k+1)T]$ $\forall k$. Puisque toute fonction non identiquement nulle, analytique, sur un compact possède un nombre fini de zéros, Z_{1k} possède donc un nombre fini d'éléments $\forall k$. $Z_1(t)$ est alors dénombrable et il en est de même de $E(t,\psi)$.

Il reste à démontrer que $S(\psi)$ est aussi dénombrable.

Soit $u = \bar{u} \in E(t,\psi)$

$$Y_{kj} = \{t : f(t+j\bar{u}) = Q_t(j\bar{u}) \ , \ t \in [kT,(k+1)T], t+j\bar{u} \in [a,b] \ j \text{ fixé}\}$$

$$Y_k = \bigcap_{j \in J} Y_{kj} \subset Y_{kp} \qquad \forall p$$

Puisque $H(t) = \sum_{i=0}^{M} [\psi_i(t+u) - \psi_i(t)](t+u)^{M-i}$ est analytique, non identiquement nulle sur $[kT,(k+1)T]$, $H(t)$ possède un nombre fini de zéros sur le compact.

Y_{k1} est donc dénombrable.

On a encore :

$$S_{\bar{u}}(\psi) = \bigcup_{k_1}^{k_2} Y_k \text{ qui est dénombrable donc :}$$

$$S(\psi) = \bigcup_{\bar{u} \in E} S_{\bar{u}}(\psi)$$

Puisque E est dénombrable, $S(\psi)$ est alors dénombrable.

2.5. CONCLUSION :

Il s'ensuit immédiatement, que l'ensemble $S(\psi)$ est un ensemble de mesure nulle dans $[a,b]$. Donc que les points t, tels que $L^2(u) = 0$ pour $u \neq kT$, ont "très peu de chance" d'apparaitre. Ce qui fait que cette méthode est pratiquement utilisable, les expériences — numériques — ayant montré d'ailleurs qu'il était difficile d'avoir un tel point t (A moins, évidemment, de construire un exemple de circonstance).

Le problème le plus délicat dans cette méthode étant d'ailleurs de choisir correctement la valeur de M. (Cette difficulté est d'ailleurs commune à de nombreux problèmes).

II.3 - CAS OU $h_k(t) = \sum_{i=1}^{M} a_{ki}\, e^{\nu_{ki} t}$ $\quad k=1,2,\ldots,N$

Le signal observé f(t) est alors de la forme :

$$f(t) = \sum_{k=1}^{N} \sum_{i=1}^{M} a_{ki}\, e^{\nu_{ki} t} \cos(\omega_k t + \varphi_k) \qquad t \in [a,b]$$

On supposera — pour l'instant — que le produit M × N est connu, et f(t) connue $\forall t \in G_n$

$$G_n = \{t \;/\; t_i = a+ih \;;\; i=0,1,\ldots,n \;;\; h = (b-a)/n\}$$

Afin de simplifier les notations, on pose $f_k = f(a+kh)$ et $a = 0$

En posant :

$$\rho_{kj} = e^{\nu_{kj} + i\omega_k} \qquad k=1,\ldots,N$$

$$D_{kj} = a_{kj}\, e^{i\varphi_k} / 2 \qquad j=1,\ldots,M$$

On peut écrire :

$$f(t) = \sum_{k,j} (D_{kj}\, \rho_{kj}^{t} + \bar{D}_{kj}\, \bar{\rho}_{kj}^{t})$$

On pose à nouveau :

$$\left.\begin{array}{l} \rho_{kj} = R_{N(j-1)+k} \;,\; \bar{\rho}_{kj} = R_{NM+N(j-1)+k} \\ D_{kj} = C_{N(j-1)+k} \;,\; \bar{D}_{kj} = C_{N(j-1)+k+NM} \end{array}\right\} \begin{array}{l} k=1,\ldots,N \\ j=1,\ldots,M \end{array}$$

Le signal peut maintenant s'écrire :

$$f(t) = \sum_{k=1}^{2M \times N} C_k (R_k)^t \qquad (p = 2 \times M \times N)$$

Utilisant une méthode due à Prony [24] , on se propose de calculer R_k $(k=1,\ldots,p)$, c'est-à-dire ω_k puisque :

$$\omega_k = \mathrm{Arg}\,(R_{N(j-1+k)}) \qquad (\forall j) \;.$$

Puisque f(t) est connue aux instants t_i $(\in G_n)$, on doit résoudre le système :

$$\sum_{k=1}^{p} C_k R_k^{jh} = f_j \qquad j=0,1,\ldots,n$$

aux 2p inconnues C_k , $Y_k = R_k^h$. Ce système est évidemment non linéaire. Ecrivons ce système :

$$\left\{\begin{array}{lllll} C_1 + C_2 + \ldots\ldots + C_p = f_0 & & & & S(0) \\ C_1 Y_1 + C_2 Y_2 + \ldots\ldots + C_p Y_p = f_1 & & & & S(1) \\ \vdots & & & & \\ C_1 Y_1^{2p-1} + \ldots\ldots + C_p Y_p^{2p-1} = f_{2p-1} & & & & S(2p-1) \end{array}\right.$$

d'où la nécessité d'avoir $n \geq 4 \times (M \times N) - 1$.

On écrit alors que $Y_1,\ldots,Y_p$ sont les racines de l'équation algébrique $Y^p - C_1 Y^{p-1} - \ldots - C_p = 0$ (7)

On obtient en faisant des combinaisons linéaires des lignes S(i) que les coefficients C_k sont solution du système linéaire :

$$\begin{cases} f_{p-1} C_1 + f_{p-2} C_2 + \ldots + f_0 C_p = f_p \\ \vdots \\ f_{2p-2} C_1 + f_{2p-3} C_2 + \ldots + f_{p-1} C_p = f_{2p-1} \end{cases} \quad (8)$$

La matrice de ce système linéaire est une matrice de Hankel :

$$H_p = \begin{bmatrix} f_0 & f_1 & \ldots & f_{p-1} \\ \vdots & & & \\ f_{p-1} & f_p & \ldots & f_{2p-2} \end{bmatrix}$$

On a d'ailleurs :

rang $H_p = p$ puisque le nombre p est supposé être connu :

ce qui veut dire que :

$$f(t) = \sum_{k=1}^{p} C_k (R_k)^t$$

est tel que $C_1 C_p \neq 0$, ou encore que les f_i vérifient une relation de récurrence linéaire (à coefficients constants) d'ordre p.

Les coefficients C_i (i=1,...,p) étant désormais supposés connus par la résolution de (8), il suffit donc de déterminer les racines de (7) pour connaître les fréquences $2\pi\ \omega_k$ (k=1,...,M) .

Si, de plus, on cherche la courbe d'établissement de ces fréquences, on calcule les coefficients a_{ki} par la résolution d'un système linéaire.

Choix du paramètre P

Soit une équation aux différences d'ordre P

$$\sum_{i=0}^{P} \alpha_i \, y_{i+j} = 0$$

qui admet donc pour solution générale

$$y_i = \sum_{k=1}^{P} \theta_k \, r_k^i$$

avec r_k racines de $\sum_{i=0}^{P} \alpha_k \, r_k^i = 0$

On peut montrer alors [22] le résultat très intéressant suivant, si :

$$H_m = \begin{bmatrix} y_o & y_1 & \cdots & y_{m-1} \\ y_1 & y_2 & \cdots & \vdots \\ \vdots & & & \vdots \\ y_{m-1} & y_m & & y_{2m-2} \end{bmatrix}$$

alors :

$$\text{rang } H_m = \begin{cases} m & \text{si} \quad m \leq P \\ P & \text{si} \quad m > P \end{cases}$$

Ceci montre que, si le système linéaire (8) a un déterminant nul, c'est que le nombre P a été choisi trop grand. Si ce déterminant est différent de 0 , P a été ou bien choisi , ou sa valeur est trop petite.

CHAPITRE III

FONCTIONS D'APPROXIMATION EN FILTRAGE DIGITAL

CHAPITRE III

III.1 - INTRODUCTION

Il ne sera pas fait un rappel du filtrage linéaire continu (se reporter à un cours d'électronique [30]). Mais on peut dire que le filtrage linéaire digital est la transposition du filtrage linéaire continu au cas discret, et que le modèle mathématique d'un filtre digital est tout naturellement une équation linéaire aux différences à coefficients constants (comme les équations différentielles linéaires à coefficients constants sont les modèles des filtres continus ou analogiques

Soit un "signal" x(t) — dit d'entrée — échantillonné tous les ΔT et le signal de sortie y(t). (Figure III.1)

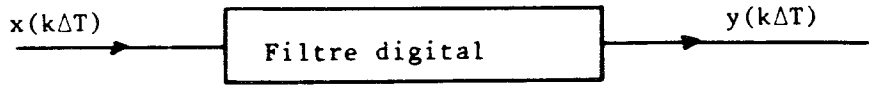


Figure III.1

Se donner un filtre, c'est se donner deux jeux de coefficients

$(\alpha_o,\ldots,\alpha_n)$, $(\beta_o,\ldots,\beta_m)$

de telle manière que "entrée" et "sortie" soient liées par la relation

$$\begin{cases} \sum_{k=0}^{n} \alpha_k \, y(p\Delta T - k\Delta T) = \sum_{k=-m_1}^{m} \beta_k x(p\Delta T - k\Delta T) & p=n,n+1,\ldots \\ \\ y(j\Delta T) \text{ données } \quad j=0,\ldots,n-1 & (\alpha_o \neq 0) \end{cases} \qquad (1)$$

L'ordre du filtre est l'ordre de l'équation aux différences (1), c'est-à-dire n . On suppose de plus que les polynomes $\sum_{k}^{n} \alpha_k \, u^{n-k}$ et $\sum_{k}^{m} \beta_k \, u^{m-k}$ sont premiers entre eux , (lorsque $m_1 = 0$).

Un filtre digital peut être utilisé dans plusieurs buts différents :

a) simuler un filtre continu

b) écrire un programme qui traite un signal discrétisé

c) réaliser un calculateur spécialisé dans le filtrage numérique.

Seul le point b) retiendra notre attention.

III.2 - CLASSIFICATION

2.1. FILTRE EN TEMPS REEL

Il existe plusieurs définitions :

2.1.1.

Un filtre est dit travailler en temps réel si et seulement si le temps nécessaire pour calculer une nouvelle valeur filtrée est inférieure à ΔT [47] .

2.1.2.

Un filtre est dit en temps réel si et seulement si $m_1 \geq 0$ [48] .

REMARQUES :

Il est certain que la première définition n'est pas une propriété du filtre, mais plutôt du calculateur sur lequel est activé le programme.

La deuxième définition veut dire que seules les informations connues à l'instant t sont prises en considération pour calculer la sortie à t . Dans tout ce qui suit, nous supposerons $m_1 = 0$.

2.2. RECURSIVITE

2.2.1.

Un filtre est dit récursif si :

$\exists\, k \in [1,n]$ tel que $\alpha_k \neq 0$ $\qquad (\alpha_o \neq 0)$

(Seuls les filtres récursifs seront étudiés dans la suite).

2.2.2.

Un filtre est dit non récursif si :

$\forall\, k \in [1,n] \qquad \alpha_k = 0 \qquad (\alpha_o \neq 0)$

2.3. CAUSALITE

Plus généralement, un système physique est dit causal si l'"effet" ne peut précéder la cause, c'est-à-dire ici :

$$x(n\Delta T) = 0 \ (n < 0) \Rightarrow y(n\Delta T) = 0 \quad (\forall n < 0)$$

III.3 - TRANSFORMEE EN z

Un des outils essentiels dans l'étude des filtres digitaux est la transformée en z, comme la transformée de Laplace est l'outil de base pour les filtres analogiques.

3.1. DEFINITION [14]

Soit une fonction f de la variable réelle t telle que $f(t) = 0$ $(t < 0)$., On considère la fonction en escalier :

$$\tilde{f}(t) = f(n\Delta T) \quad n\Delta T \leq t < (n+1)\Delta T \qquad n=0,1,\ldots$$

Calculons la transformée de Laplace — si elle existe — de $\tilde{f}(t)$:

$$L[\tilde{f}] = \int_0^\infty e^{-st}\tilde{f}(t)dt = \frac{1-e^{-\Delta Ts}}{s} \sum_{n=0}^{\infty} f(n\Delta T)e^{-sn\Delta T}$$

Posons $e^{\Delta Ts} = z$. La quantité :

$$F(z) = \sum_{n=0}^{\infty} f_n z^{-n}$$

est appelée transformée en z de la suite f_n .

3.2. PRINCIPALES PROPRIETES

On emploiera le signe de correspondance $\circ\!\!-\!\!\bullet$ pour montrer que F(z) est la transformée en z de la suite f_n

$$f_n \circ\!\!-\!\!\bullet F(z)$$

3.2.1.

Linéarité

3.2.2.

Déplacement

$$f_{n-k} \circ\!\!-\!\!\bullet z^{-k}F(z) \qquad k=0,1,\ldots$$

$$f_{n+k} \circ\!\!-\!\!\bullet z^{k}\left[F(z)-\sum_{j=0}^{k-1} f_j z^{-j}\right] \qquad k=0,1,\ldots$$

III .4 - FONCTION DE TRANSFERT - REPONSE EN FREQUENCE

4.1.

La transformée en z est un outil commode pour intégrer une équation aux différences (et à coefficients constants).

Soit l'équation aux différences :

$$\begin{cases} \sum_{0}^{n} \alpha_k y_{p-k} = \sum_{0}^{m} \beta_k x_{p-k} & p=n,n+1,\ldots \\ y_i \text{ données} & i=0,1,\ldots,n-1 \end{cases} \qquad (2)$$

La suite x_n a pour transformée en z:X(z), et Y(z) est celle de la suite y_n . Dans le cas où $n \geq m$, on peut montrer que, en prenant la transformée en z des deux membres de (2)

$$\sum_{k=0}^{n} \alpha_k z^{-k} \left[Y(z) - \sum_{j=0}^{n-k-1} y_j z^{-j} \right] = \sum_{k=0}^{m} \beta_k z^{-k} \left[X(z) - \sum_{j=0}^{m-k-1} x_j z^{-j} \right]$$

d'où :

$$Y(z) = \frac{\sum_{k=0}^{m} \beta_k z^{-k} \left[X(z) - \sum_{j=0}^{m-k-1} x_j z^{-j} \right] + \sum_{k=0}^{n} \alpha_k z^{-k} \sum_{j=0}^{n-k-1} y_j z^{-j}}{\sum_{k=0}^{n} \alpha_k z^{-k}}$$

(si $n < m$, on trouve un résultat analogue).

Pour obtenir alors la solution de l'équation aux différences (2) satisfaisant aux conditions initiales données, il suffit de chercher la suite y_n qui a pour transformée $Y(z)$. On peut se servir de tables [28] . L'avantage de cette méthode est qu'elle tient compte des conditions initiales.

La quantité

$$H(z) = \frac{\sum_{0}^{m} \beta_k z^{-k}}{\sum_{0}^{n} \alpha_k z^{-k}}$$

est appelée <u>fonction de transfert</u>.

4.2.

La solution générale de l'équation aux différences (2) s'obtient en ajoutant la solution générale de l'équation sans second membre à une solution particulière de l'équation avec second membre.

- Solution de l'équation homogène on pose $y_n = r^n$.

On obtient alors :

$$y_n = \sum_{i=1}^{n} c_i r_i^n$$

avec r_i racines (supposées simples) de $\sum_{0}^{n} \alpha_k r^{n-k} = 0$, équation caractéristique de l'équation aux différences.

- Supposons que $x_n = x(n\Delta T) = e^{in\omega\Delta T}$

Cherchons une solution particulière de l'équation (2) sous la forme

$$y_n = G(\omega)e^{in\omega\Delta T}$$

On obtient :

$$G(\omega) = \frac{\sum_{}^{m} \beta_k e^{-ik\omega\Delta T}}{\sum_{}^{n} \alpha_k e^{-ik\omega\Delta T}}$$

Donc la solution de (2) s'écrit :

$$y(p\Delta T) = \sum_{1}^{n} c_i r_i^p + H(e^{i\omega\Delta T})e^{i\omega p\Delta T}$$

car $G(\omega) = H(e^{i\omega\Delta T})$.

Les coefficients c_i sont tels que y satisfait aux conditions initiales.

$H(e^{i2\pi f\Delta T})$ est appelée la réponse en fréquence.

4.3. CAS D'UN SYSTEME CAUSAL

Le système étant causal, les conditions initiales sont telles que :

$$\begin{cases} x_p = x(p\Delta T) = 0 & p < 0 \\ y_p = y(p\Delta T) = 0 & p < 0 \end{cases}$$

Prenons la transformée en z des deux membres :

$$\sum_{p=0}^{\infty} \sum_{0}^{n} \alpha_k y_{p-k} z^{-p} = \sum_{0}^{\infty} \sum_{0}^{m} \beta_p x_{p-k} z^{-p}$$

$$\sum_{0}^{n} \alpha_k z^{-k} \left[\sum_{0}^{\infty} y_j z^{-j} + \sum_{-k}^{-1} y_j z^{-j}\right] = \sum_{0}^{m} \beta_p z^{-k} \left[\sum_{0}^{\infty} x_j z^{-j} + \sum_{-k}^{-1} x_j z^{-j}\right]$$

donc :

$$\frac{Y(z)}{X(z)} = H(z) = \frac{\sum_{0}^{m} \beta_k z^{-k}}{\sum_{0}^{n} \alpha_k z^{-k}}$$

THEOREME 1 :

La solution d'une équation aux différences linéaires représentant un système causal peut s'écrire :

$$y_k = \sum_{i=n-m}^{k} A_i x_{k-i} \qquad n-m \geq 0$$

où les A_i sont les coefficients du développement formel

$$\frac{\sum_{0}^{m} \beta_k u^k}{\sum_{0}^{n} \alpha_k u^k} = \sum_{i=n-m}^{\infty} A_i u^i$$

Si le système est causal, on a en effet :

$$Y(z) = \frac{\sum_{0}^{m} \beta_k z^{-k}}{\sum_{0}^{n} \alpha_k z^{-k}} X(z) = \sum_{i=n-m}^{\infty} A_i z^{-i} X(z)$$

de la propriété 3.2.2. on tire immédiatement que :

$$y_k = \sum_{i=n-m}^{\infty} A_i x_{k-i} = \sum_{i=n-m}^{k} A_i x_{k-i} \qquad (x_k = 0 \ , \ k < 0) \ .$$

III.5 - STABILITE

Il existe plusieurs définitions de la stabilité d'un système physique. Nous en donnerons deux définitions, et dans la suite nous nous placerons dans le cas où ces deux notions sont équivalentes.

5.1. STABILITE AU SENS DE LYAPUNOV [20]

Soit un système ayant une entrée $x(t)$ et une sortie $y(t)$.

Supposons que sa position d'équilibre soit :

$$x(t) = 0 \qquad y(t) = 0 \qquad \forall t$$

On dit qu'on a une position d'équilibre stable, si (t_o) désignant l'instant initial :

étant donné $\varepsilon > 0$, $\exists \eta > 0$ tel que si $|x(t_o)| < \eta \Rightarrow$

$\exists T_o$ tel que $\forall t > T_o \quad |y(t)| < \varepsilon$

(Il est à remarquer que cette notion de stabilité ne requiert pas $y(t) \xrightarrow[t \to \infty]{} 0$) . Si de plus $y(t) \to 0$ quand $t \to \infty$, le système est dit

asymptotiquement stable.

Dans le cas d'un filtre linéaire à coefficients constants, il est aisé de montrer qu'un filtre est asymptotiquement stable si et seulement si $|r_i| < 1 \quad \forall i=1,\ldots,n$

(r_i racines de l'équation caractéristique $\sum_{0}^{n} \alpha_k r^{n-k} = 0$)

5.2. STABILITE AU SENS DE JAMES [20]

"Toute entrée bornée donne une sortie bornée".

On peut alors montrer que cette définition de la stabilité est équivalente à la stabilité asymptotique de Lyapunov.

5.3. ANALYSE FONCTIONNELLE

On considère un filtre comme une application (cf. Introduction) de l'espace des signaux S_1 dans l'espace des signaux S_2 .

5.3.1.

La stabilité asymptotique (ou de James) entraîne immédiatement que le système physique considéré, peut être considéré comme une application de :

L_∞ dans L_∞ (cas continu)

ou

ℓ_∞ dans ℓ_∞ (cas discret)

5.3.2.

Si l'espace des signaux d'entrée est L_2 , et si la réponse en fréquence du filtre est bornée, alors le filtrage linéaire est une application de L_2 dans L_2 .

En effet soit :

$x(t)$ le signal d'entrée de transformée de Fourier $X(\omega)$

$y(t)$ le signal de sortie de transformée de Fourier $Y(\omega)$

D'après la formule de Parseval :

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |X(\omega)|^2 d\omega$$

En utilisant la fonction de transfert d'un système continu (linéaire, à coefficients constants), on obtient immédiatement :

$$Y(\omega) = H(\omega)\, X(\omega)$$

utilisant à nouveau la relation de Parseval

$$\int_{-\infty}^{+\infty} |y(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |H(\omega)|^2 |X(\omega)|^2 d\omega \leq \frac{K}{2\pi} \int_{-\infty}^{+\infty} |X(\omega)|^2 d\omega = K \int_{-\infty}^{+\infty} |x(t)|^2 dt$$

(Ceci est bien naturel, car les systèmes étudiés ici sont de type passif et ne peuvent donc qu'absorber de l'énergie).

5.4. PARTIE TRANSITOIRE DE LA SORTIE D'UN SYSTEME STABLE

Nous avons vu dans 4.2. que la solution générale de l'équation (2) avec second membre est :

$$y_p = \sum_{1}^{n} c_i r_i^p + H(e^{i\omega\Delta T}) e^{i\omega p \Delta T}$$

(Dans le cas d'une entrée sinusoïdale, et dans le cas de racines simples. Si il existe une (ou des) racine(s) multiple(s), y_p est un peu plus compliqué à écrire mais le résultat sera inchangé.)

* $\sum_{1}^{n} c_i r_i^p$ correspond au régime transitoire (car $|r_i| < 1$) qui sera d'autant plus bref que $\underset{i=1,\ldots,n}{\text{Max}} |r_i|$ sera plus petit.

* $H(e^{i\omega\Delta T})e^{ip\omega\Delta T}$ correspond au régime permanent.

III.6 - FONCTION D'APPROXIMATION

Ayant restreint notre étude à des équations aux différences linéaires à coefficients constants, nous savons que H(z) est une fraction rationnelle de z^{-1} . Cette étude est basée essentiellement sur les propriétés des fonctions de transfert (analogiques ou digitales).

Le problème est le suivant : on se donne un modèle de filtre par une fonction C(ω) (complexe ou non de la variable réelle ω), on se propose alors par un choix des (α_k, β_k) d'ajuster "au mieux"

$$C(\omega) \quad \text{par} \quad H(e^{i\omega\Delta T}) \ .$$

Dans cette étude, nous nous bornerons, en général, à ajuster le carré du module (pour des raisons de simplifications) de C(ω) soit G (ω) par $|H(e^{i\omega\Delta T})|^2$ qui est quelquefois appelé fonction d'approximation.

Il est immédiat de voir que la fonction d'approximation possède les deux propriétés :

* périodicité de période $2\pi/\Delta T$

* parité.

Nous demandons donc que G(ω) possède ces deux propriétés.

6.1. PROPRIETE DE LA FONCTION D'APPROXIMATION [25]

6.1.1. Fraction rationnelle en cos ωΔT

$$|H(e^{i\omega\Delta T})|^2 = \frac{|\sum_0^m \beta_k e^{ik\omega\Delta T}|\,|\sum_0^m \beta_k e^{-ik\omega\Delta T}|}{|\sum_0^n \alpha_k e^{ik\omega\Delta T}|\,|\sum_0^n \alpha_k e^{-ik\omega\Delta T}|} = \frac{\sum_0^m \eta_j \cos j\omega\Delta T}{\sum_0^n \theta_j \cos j\omega\Delta T}$$

D'après les relations des polynomes de Tchebycheff, on sait que :

$$\cos j\omega\Delta T = \sum_{p=0}^{j} \nu_{jp} \cos^p \omega\Delta T$$

donc :

$$\sum_0^m \eta_j \cos j\omega\Delta T = \sum_0^m \eta_j \sum_0^j \nu_{jp} \cos^p \omega\Delta T = \sum_0^m b_k \cos^k \omega\Delta T$$

On a alors :

$$|H(e^{i\omega\Delta T})|^2 = \frac{\sum_0^m b_k \cos^k \omega\Delta T}{\sum_0^n a_k \cos^k \omega\Delta T} = \frac{C(\cos\omega\Delta T)}{D(\cos\omega\Delta T)} .$$

On a évidemment $C(\cos\omega\Delta T) \geq 0 \qquad \forall\omega \in [0,\pi/\Delta T]$

et

$$D(\cos\omega\Delta T) > 0 \qquad \forall\omega \in [0,\pi/\Delta T]$$

car le filtre considéré a été supposé asymptotiquement stable donc $|r_i| \neq 1$.

6.1.2. Problème inverse

A quelle condition un polynome P_n en cos x à coefficients réels peut-il s'écrire

$$P_n(\cos x) = S_n(e^{ix})\ S_n(e^{-ix})$$

où $S_n(y)$ est un polynome — en y — de degré n à coefficients réels.

THEOREME 2 :

Une condition nécessaire et suffisante pour que :

$$P_n(\cos x) = \sum_{j=0}^{n} b_j \cos^j x \qquad (b_j \text{ réels})$$

puisse s'écrire :

$$P_n(\cos x) = S_n(e^{ix})\ S_n(e^{-ix})$$

$$\text{avec } S_n(y) = \sum_{k=0}^{n} \xi_k y^k \qquad (\xi_k \text{ réels})$$

est que :

$$P_n(\cos x) \geq 0 \qquad \forall x \in [0,\pi]$$

La condition nécessaire est évidente.

Démonstration de la condition suffisante :

Soit :

$$P_n(u) = \sum_{0}^{n} b_j u^j \geq 0 \qquad \forall u \in [-1,1] \quad (u = \cos x)$$

c'est un polynome à coefficients réels, donc :

$$P_n(u) = b_n(u-u_1),\ldots,(u-u_n)\ .$$

Groupons les racines de $P_n(u)$ en trois classes :

$$C_1 = \{u_p : P_n(u_p) = 0 \qquad -1 < u_p < 1\}$$

$$C_2 = \{u_p : P_n(u_p) = 0 \qquad u_p \geq 1 \text{ ou } u_p \leq -1\}$$

$$C_3 = \{u_p : P_n(u_p) = 0 \qquad u_p \text{ complexe}\}$$

Examinons chaque classe C_i , et faisons le changement de variable

$$u = (z+z^{-1}) / 2$$

a) Soit $u_p \in C_1$. Si u_p est une racine d'ordre impair, alors $P_n(u)$ change de signe sur $[-1,1]$, ce qui est contraire à l'hypothèse. Donc u_p est une racine d'ordre pair.

Analysons le terme $(u-u_p)^2$

$$r(z) = \left(\frac{z+z^{-1}}{2} - u_p\right)^2 = \frac{1}{4}[z^2-2u_pz+1][z^{-2}-2u_pz^{-1}+1]$$

b) Soit $u_p \in C_2$. $(u-u_p)$ a un signe constant sur $[-1,1]$.

$$v(z) = \frac{z+z^{-1}}{2} - u_p = -\frac{1}{2z_1}(z-z_1)(z^{-1}-z_1)$$

($z_1 = u_p + \sqrt{u_p^2-1}$ est un zéro réel de v(z) , l'autre étant $z_2 = 1/z1$ qui est aussi réel).

c) Soit $u_p \in C_3$. $P_n(u)$ étant un polynome à coefficients réels, $\bar{u}_p \in C_3$

$$w(z) = \left(\frac{z+z^{-1}}{2} - u_p\right)\left(\frac{z+z^{-1}}{2} - \bar{u}_p\right) = \frac{1}{4z^2}[z^2-2u_pz+1][z^2-2\bar{u}_pz+1]$$

Si z_1 désigne une racine de $z^2-2u_p z+1$

$$w(z) = \frac{1}{4z^2}(z-z_1)(z-\frac{1}{z_1})(z-\bar{z}_1)(z-\frac{1}{\bar{z}_1})$$

$$= \frac{1}{4|z_1|^2}[z^2-(z_1+\bar{z}_1)z+|z_1|^2][z^{-2}-(z_1+\bar{z}_1)z^{-1}+|z_1|^2]$$

Donc finalement $P_n(\frac{z+z^{-1}}{2})$ s'écrit comme le produit d'un polynome $S_n(z)$, à coefficients réels, de degré n, par $S_n(z^{-1})$.

Puisque $\cos x = \frac{e^{ix}+e^{-ix}}{2}$ le théorème est ainsi démontré :

COROLLAIRE 1 :

Les racines du polynome $S_n(z)$ peuvent être choisies telles que

$|z_i| \leq 1 \qquad i=1,\ldots,n$

* si $u_p \in C_1$, on a immédiatement $|z_p| = 1$.

* si $u_p \in C_2 \cup C_3$. On a choisi pour racine de $Z(z)$, z_p.
  Si $|z_p| \leq 1$, le corollaire est montré ; sinon, au lieu de choisir z_p, on prend la deuxième racine $1/z_p$ qui est de module inférieur à 1.
  Dans ce cas

$$v(z) = -\frac{z_1}{2}(z-\frac{1}{z_1})(z^{-1}-\frac{1}{\bar{z}_1})$$

$$w(z) = \frac{|z_1|^2}{4}[z^2-(\frac{1}{z_1}+\frac{1}{\bar{z}_1})z+\frac{1}{|z_1|^2}][z^{-2}-(\frac{1}{z_1}+\frac{1}{\bar{z}_1})z^{-1}+\frac{1}{|z_1|^2}]$$

On arrive donc au résultat très intéressant suivant :

Se donner une fonction d'approximation $\frac{C(\cos\omega\Delta T)}{D(\cos\omega\Delta T)}$
$(C(u) \geq 0 \ , \ D(u) > 0 \qquad \forall u \in [-1,1]$

revient donc à se donner une famille de filtres, c'est-à-dire (α,β))
(car la décomposition du théorème 2 n'est évidemment pas unique).

COROLLAIRE 2 :

La fonction d'approximation $\frac{C(\cos\omega\Delta T)}{D(\cos\omega\Delta T)}$ étant donnée, il existe un filtre stable qui a pour fonction d'approximation $\frac{C(\cos\omega\Delta T)}{D(\cos\omega\Delta T)}$.

C'est une application immédiate du théorème 2 et du corollaire 1. Les coefficients du filtre sont obtenus en mettant $S_{C_m}(z)$ et $S_{D_n}(z)$ sous forme canonique.

6.2. FILTRE IDEAL

6.2.1. Filtre passe-bas (Figure III.2)

$$G_b(\omega) = \begin{cases} 1 & 0 < \omega \leq \omega_c \\ 0 & \omega_c < \omega < \pi/\Delta T \end{cases}$$

ω_c fréquence de coupure.

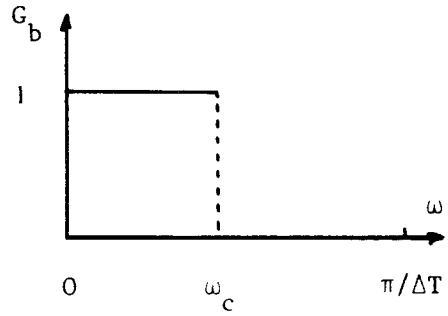


Figure III.2

6.2.2. Filtre passe-haut

$$G_h(\omega) = 1-G_b(\omega) = \begin{cases} 0 & 0 < \omega \leq \omega_c \\ 1 & \omega_c < \omega \leq \pi/\Delta T \end{cases}$$

6.2.3. Filtre passe-bande

$$G_{pb}(\omega) = G_{b,\omega_2}(\omega)[1-G_{b,\omega_1}(\omega)] = \begin{cases} 1 & \omega_1 \leq \omega \leq \omega_2 \\ 0 & \omega \notin [\omega_1,\omega_2] \end{cases}$$

(ω_1,ω_2) est la bande passante.

6.2.4. Filtre rejeteur de bande

(Figure III.3)

$$G_{rb}(\omega) = \begin{cases} 0 & \omega \in [\omega_1, \omega_2] \\ 1 & \omega \notin [\omega_1, \omega_2] \end{cases}$$

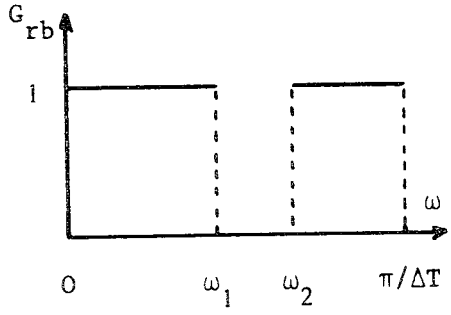


Figure III.3

Ces filtres sont considérés dans un sens idéal, car il est impossible de faire coïncider une fonction d'approximation avec ces modèles en tous les points de l'intervalle $[0,\pi/\Delta T]$.

6.3. CONSTRUCTIONS DIVERSES

En général, seul le filtre passe-bas idéal est considéré, car on peut construire les fonctions d'approximation des autres types à partir de celui-ci.

6.3.1. Changement de variable

Posons $u = \cos\omega\Delta T$, donc $u \in [-1,1]$ quand $\omega \in [0,\pi/\Delta T]$.

Soit G(u) un modèle de filtre passe-bas.

On a donc :

$$G(u) = \begin{cases} 1 & u \in [u_o, 1] \\ 0 & u \in [-1, u_o[\end{cases} \qquad (u_o = \cos\omega_c \Delta T)$$

Soit X(u) une fonction définie sur $[-1,1]$ ($|X(u)| \leq 1$).

On a alors :

$$G(X(u)) = \begin{cases} 1 & X(u) \in [u_o, 1] \\ 0 & X(u) \in [-1, u_o] \end{cases}$$

Une méthode basée sur un principe semblable est utilisée en filtrage analogique [57].

La figure III.4 montre la construction d'un filtre passe-bande (u_1, u_2) à partir d'un passe-bas (u_o).

Le moyen pratique d'utilisation de cette méthode est le suivant :

* trouver une fonction d'approximation du filtre passe-bas idéal
* changer $\cos\omega\Delta T$ en $X(\cos\omega\Delta T)$. X(u) doit donc être une fraction rationnelle de u.

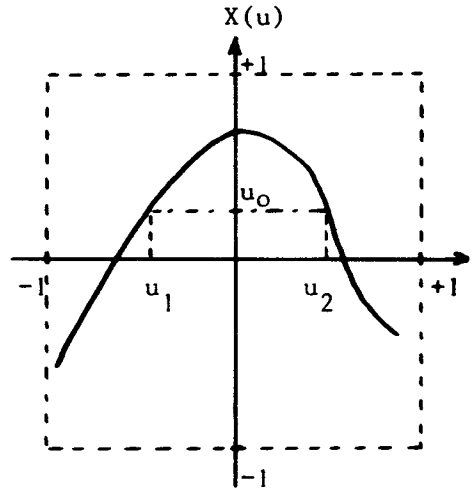


Figure III.4

6.3.2. Méthode directe :

Soit $G(\omega)$ un modèle de filtre. La fonction d'approximation $\frac{C(\omega)}{D(\omega)}$ est choisie telle que : [56]

$$\| G - \frac{C}{D} \|$$

soit minimum ($\| \ \|$ étant une norme convenablement choisie).

Cette propriété sera exploitée plus tard (cf. chapitre IV).

III.7 - CONSTRUCTION DE FONCTIONS D'APPROXIMATION A PARTIR DE FILTRES ANALOGIQUES

Le point de départ est en général la fonction de transfert d'un filtre analogique que nous noterons Y(s).
(Y(s) étant supposée être une fraction rationnelle).

7.1. INVARIANCE DE LA REPONSE IMPULSIONNELLE [15]

Un système est caractérisé par sa réponse impulsionnelle ou encore par sa fonction de transfert.

Nous pouvons écrire (cas de racines simples) Y(s) sous la forme

$$Y(s) = \sum_{i=1}^{n} \frac{A_i}{s+s_i}$$

La réponse impulsionnelle (réponse à la fonction $\delta(t)$ de Dirac) est alors l'inverse de la Transformée de Laplace de la fonction de transfert Y(s) donc :

$$y(t) = L^{-1}\left(\sum_{1}^{n} \frac{A_i}{s+s_i}\right) = \sum_{1}^{n} A_i e^{-s_i t}$$

Soit $h(k\Delta T)$ la réponse impulsionnelle d'un filtre digital (réponse à $\delta_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases}$), on désire donc que :

$$h(k\Delta T) = y(k\Delta T) = \sum_{1}^{n} A_i e^{-s_i k\Delta T}$$

Si on prend la transformée en z de $\{h_n\}$, on obtient :

$$H(z) = \sum_{0}^{\infty} h(k\Delta T) z^{-n} = \sum A_i \sum e^{-s_i k\Delta T} z^{-n} = \sum_{1}^{n} \frac{A_i}{1-e^{-s_i \Delta T} z^{-1}}$$

Donc, si on veut avoir identité des réponses impulsionnelles, on transforme :

$$Y(s) = \sum_{1}^{n} \frac{A_i}{s+s_i} \Rightarrow H(z) = \sum_{1}^{n} \frac{A_i}{1-e^{-s_i \Delta T} z^{-1}}$$

(Dans le cas où les poles de Y(s) sont multiples, on obtient un résultat du même type, mais plus compliqué à établir).

EXEMPLE :

$$Y(s) = \frac{s+a}{(s+a)^2+b^2} \Rightarrow H(z) = \frac{1-e^{-a\Delta T}\cos(b\Delta T)z^{-1}}{1-2e^{-a\Delta T}\cos(b\Delta T)z^{-1}+e^{-2a\Delta T}z^{-2}}$$

L'ennui avec ces méthodes qui ajustent la sortie temporelle avec celle d'un filtre analogique est que la réponse en fréquence peut être grandement modifiée. Une autre technique est d'ajuster plutôt la réponse en fréquence.

7.2. FONCTIONS TRIGONOMETRIQUES

Il est aisé de voir que la fonction d'approximation d'un filtre analogique est une fraction rationnelle de ω^2

$$\left(\frac{C_1(\omega^2)}{D_1(\omega^2)} \quad \text{avec} \quad C_1(\omega^2) \geq 0 \quad , \quad D_1(\omega^2) > 0 \quad \forall\omega \right) .$$

On obtient une fonction d'approximation digitale en remplaçant ω^2 par

$$\text{trig}^2 \frac{\omega\Delta T}{2}$$

donc :

$$\frac{C(\cos(\omega\Delta T))}{D(\cos(\omega\Delta T))} = \frac{C_1(\text{trig}^2 \frac{\omega\Delta T}{2})}{D_1(\text{trig}^2 \frac{\omega\Delta T}{2})}$$

où trig ( ) représente une fonction trigonométrique élémentaire quelconque :

$$\sin(\ ),\ \cos(\ ),\ \mathrm{tg}(\ ),\ \mathrm{cotg}(\ )\ .$$

En effet, alors $\mathrm{trig}^2\ (\frac{\omega\Delta T}{2})$ s'exprime rationnellement en fonction de $\cos\omega\Delta T$

$$\cos\omega\Delta T = \frac{1-\mathrm{tg}^2\ \frac{\omega\Delta T}{2}}{1+\mathrm{tg}^2\ \frac{\omega\Delta T}{2}} \quad ; \quad \cos\omega\Delta T = 2\cos^2\frac{\omega\Delta T}{2} - 1 = 1-2\sin^2\frac{\omega\Delta T}{2}$$

7.2.1. Filtres de Butterworth

Soit $G_n(\omega)$ la fonction d'approximation d'un filtre analogique d'ordre n , d'un filtre passe-bas idéal ($G(\omega)$).

Si $G_n(\omega)$ satisfait :

* $G_n(0) = 1$
* $G_n(\omega)$ monotone (sur $(0,\infty)$)
* $\lim_{n\to\infty} G_n(\omega) = G(\omega)$

Le filtre correspondant est appelé filtre de Butterworth.

EXEMPLE :

$$G_n(\omega) = \frac{1}{1+(\frac{\omega}{\omega_c})^{2n}}$$

Cas où l'on prend :

trig : tg. On a alors :

$$H_n(\omega) = \frac{1}{1+\left(\frac{\mathrm{tg}^2\ \frac{\omega\Delta T}{2}}{\mathrm{tg}^2\ \frac{\omega_c\Delta T}{2}}\right)^n}$$

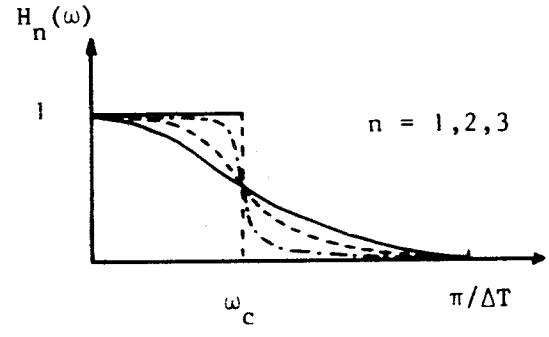


Figure III.5

ce qui s'écrit (Figure III.5)

$$H_n(\omega) = \frac{tg^{2n}\frac{\omega_c \Delta T}{2}}{tg^{2n}\frac{\omega_c \Delta T}{2} + \left(\frac{1-\cos\omega\Delta T}{1+\cos\omega\Delta T}\right)^n}$$

Une étude systématique des cas où trig ≡ (cos, sin, tg) est faite dans [25] .

7.2.2. Filtres de Tchebycheff

La fonction d'approximation d'un tel filtre (ordre n) est : (Figure III.6)

$$G_n(\omega) = \frac{1}{1+\varepsilon^2 T_n^2\left(\frac{\omega}{\omega_c}\right)}$$

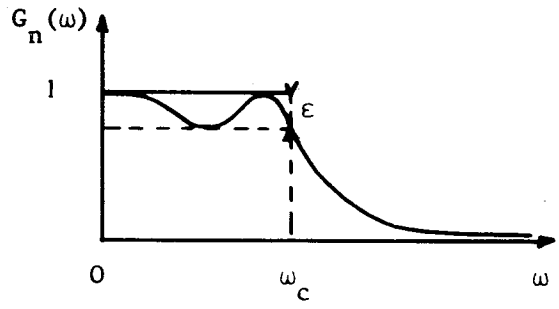


Figure III.6

où $T_n(x)$ est le polynome de Tchebycheff d'ordre n.

On sait que

$T_{2p}(x)$ est un polynome pair

donc $T_n^2(x)$ est une fonction paire

T_{2p+1} est un polynome impair

Donc pour avoir le filtre digital passe-bas correspondant, on change

$$\frac{\omega^2}{\omega_c^2} \longrightarrow \frac{tg^2\frac{\omega\Delta T}{2}}{tg^2\frac{\omega_c \Delta T}{2}}$$

7.2.3. Filtre elliptique [21]

Sa construction est basée sur les propriétés des fonctions elliptiques de Jacobi.

Le carré du module de sa fonction de transfert est : (cas continu) :

$$J^2(u) = \frac{1}{1+\varepsilon^2\, sn^2(u,k_1)}$$

avec

$$k_1 = \varepsilon / \sqrt{A^2 - 1}$$

(Figure III.7)

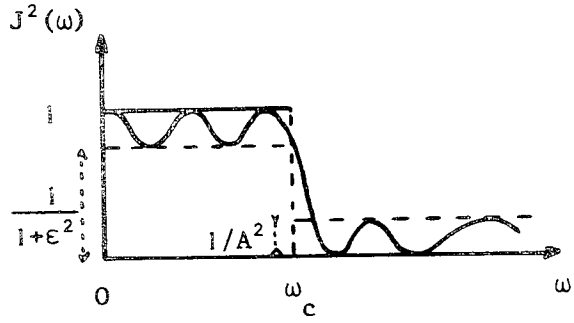


Figure III.7

$$\mu = sn^{-1}y = \int_0^y \frac{dt}{\sqrt{(1-t^2)(1-k_1^2t^2)}}$$

Il existe encore bien d'autres types de filtres qui ont des expressions analytiques simples. L'avantage de chacun est en général fixé par son auteur.

7.3. "TRANSFORMATION BILINEAIRE"

On remarque que pour définir un filtre analogique, il suffit de se donner la valeur d'une fraction rationnelle sur l'axe imaginaire, et que, pour définir un filtre digital, il suffit de se donner la valeur d'une fraction rationnelle sur le cercle unité. Aussi on peut rechercher une transformation

$$s = f(z)$$

qui transforme l'axe imaginaire en le cercle unité. De plus, partant d'un filtre analogique, on désire un filtre digital stable, on devra donc transformer la partie gauche du plan s en l'intérieur du cercle unité.

Soit Y(s) la fonction de transfert du filtre analogique de départ, et H(z) celle du filtre digital d'arrivée. On a le résultat suivant [43]

THEOREME 3 :

> *On peut calculer H(z) en introduisant* $s = f(z)$ *dans Y(s) seulement si* $f(z)$ *est rationnelle en* z *et si pour cette transformation le demi-plan gauche des* s *se transforme uniquement dans l'intérieur du cercle unité des* z .

La transformation définie par :

$$s = \frac{2}{\Delta T}\ \frac{z-1}{z+1}$$

satisfait les conditions ci-dessus.

EXEMPLE :

$$Y(s) = \frac{1}{1 + \frac{s}{\omega_c}} \Rightarrow H(z) = \frac{1 + z^{-1}}{(1 + \frac{2}{\omega_c \Delta T}) + (1 - \frac{2}{\omega_c \Delta T})z^{-1}}$$

Par définition de la transformée en z , on obtient immédiatement le filtre numérique :

$$(1 + \frac{2}{\omega_c \Delta T})y(n\Delta T) + (1 - \frac{2}{\omega_c \Delta T})y((n-1)\Delta T) = x(n\Delta T)+x((n-1)\Delta T) \qquad (3)$$

Il est cependant assez gênant de travailler avec le filtre (3). Car il fait intervenir ω_c (fréquence de coupure du filtre analogique, c'est-à-dire $|Y(j\omega_c)|^2 = 1/2$) alors que ω_c n'a pas la même signification dans le filtre numérique.

Notons ω_A une "pulsation analogique"
ω_D une "pulsation digitale".

On a :

$$Y(i\omega_A) = H(e^{i\omega_D \Delta T}) ,$$

si ω_A et ω_D dont tels que :

$$i\omega_A = \frac{2}{\Delta T} \frac{e^{i\omega_D \Delta T} - 1}{e^{i\omega_D \Delta T} + 1} = i \, tg \frac{\omega_D \Delta T}{2} \frac{2}{\Delta T}$$

Donc si :

$$\omega_A = \frac{2}{\Delta T} tg \frac{\omega_D \Delta T}{2}$$

On a donc une distorsion quand on passe par la transformée bilinéaire, d'un filtre analogique à un filtre digital.

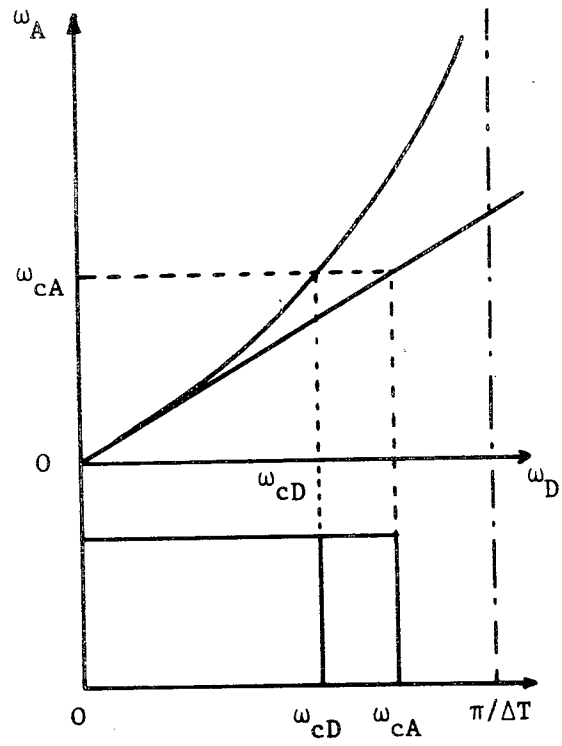


Figure III.8

La figure III.8 montre comment est transformée le filtre passe-bas analogique (ω_c) de l'exemple en un filtre passe-bas digital.

Donc au lieu d'utiliser le filtre digital (3) on utilisera

$$\left(1 + \frac{1}{tg \frac{\omega_{CD} \Delta T}{2}}\right) y(n\Delta T) + \left(1 - \frac{1}{tg \frac{\omega_{CD} \Delta T}{2}}\right) y((n-1)\Delta T) = x(n\Delta T) + x((n-1)\Delta T)$$

qui est un filtre passe-bas ayant pour pulsation de coupure ω_{CD} .

7.4. (ρ,σ) METHODES

Cette méthode est basée sur l'intégration numérique de l'équation différentielle associée à la fonction de transfert Y(s) [68] .

Soit l'équation différentielle :

$$\begin{cases} \sum_{0}^{n} \alpha_k y^{(k)}(t) = \sum_{0}^{n-1} \beta_k x^{(k)}(t) \\ \\ y^{(k)}(0) = 0 \qquad\qquad k=0,1,\ldots,n-1 \ . \end{cases} \tag{4}$$

Soient deux polynomes à coefficients réels

$$\rho(\lambda) = \sum_{0}^{p} \nu_k \lambda^k \quad , \quad \sigma(\lambda) = \sum_{0}^{r} \mu_k \lambda^k$$

On désigne par E l'opérateur d'avancement, c'est-à-dire

$$Ey_k = y_{k+1}$$

On remplace alors le problème différentiel (4) par le problème aux différences :

$$\begin{cases} \sum_{0}^{n} \alpha_k \dfrac{\rho^k(E)\sigma^{n-k}(E)}{\Delta T^k} y_j = \sum_{0}^{n-1} \beta_k \dfrac{\rho^k(E)\sigma^{n-k}(E)}{\Delta T^k} x_j \qquad j=0,1,\ldots \\ \\ y_j = \text{condition initiale donnée} \qquad j=0,\ldots,n-1 \end{cases} \tag{5}$$

Cette méthode fait un peu la synthèse des méthodes précédentes. En effet, si on est intéressé par :

* l'ajustement de la réponse temporelle, à celle de la réponse du filtre analogique, on choisit une méthode consistante et stable, ce qui entraîne la convergence, en un point fixé, vers la solution de l'équation différentielle.

EXEMPLE : Méthode de Milne

$$\rho(z) = z^2-1 \qquad \sigma(z) = \frac{1}{3}(z^2+4z+1) .$$

* l'ajustement en fréquence,
On devra alors choisir

$$s = \frac{\rho(z)}{\Delta T \sigma(z)}$$

satisfaisant au théorème 3 de 7.3.

EXEMPLE, règle du trapèze

$$\rho(z) = z-1 \qquad \sigma(z) = \frac{z+1}{2}$$

On retrouve la transformation bilinéaire.

Dans le cas où on utilise une (ρ,σ) méthode, il se produit les mêmes phénomènes de distorsion, que celui évoqué à propos de la transformation bilinéaire, mais qui doivent être formulés pour chaque ρ et σ choisis.

CHAPITRE IV

FILTRES OPTIMAUX.

APPROXIMATION PAR FRACTIONS RATIONNELLES

IV.1. INTRODUCTION - DEFINITION

On a montré (III.1) que, construire un filtre revenait à calculer deux jeux de coefficients

$$(\alpha_o,\ldots,\alpha_n) \quad \text{et} \quad (\beta_o,\ldots,\beta_m)$$

de telle manière que la réponse en fréquence $H(e^{i\omega\Delta T})$ s'ajuste le mieux possible au modèle donné $H(\omega)$.

On se limitera ici, en raison de la complexité du problème, à l'ajustement du carré du module du modèle (CMM).

On a alors (III.6.1.1.) en posant $x = \omega\Delta T$

$$|H(e^{ix})|^2 = \frac{\sum_0^m b_k \cos^k x}{\sum_0^n a_k \cos^k x} = g(x)$$

où g(x) est une fonction paire et périodique (2π) : propriétés que le C.M.M. doit aussi satisfaire :

Soit V l'ensemble des fonctions définies sur $[0,\pi]$ de la forme :

$$g(x) = \frac{\sum_0^m b_k \cos^k x}{\sum_0^n a_k \cos^k x}$$

où les coefficients a_k et b_k sont :

i) tels que $\sum_0^m b_k \cos^k x \geq 0 \qquad \forall x \in [0,\pi]$

ii) tels que $\sum_0^n a_k \cos^k x > 0 \qquad \forall x \in [0,\pi]$

iii) normalisés (Cette normalisation sera précisée ultérieurement).

On cherchera — s'il existe — un meilleur approximant $g^* \in V$ tel que si f est le carré du module du modèle

$$\|g^*-f\| = \inf_{g\in V} \|g-f\|$$

On a montré, (III.6.1.2.), que, si $g^* \in V$, il est alors possible de construire, c'est-à-dire trouver (α,β), un filtre stable. Ce filtre stable sera alors dit optimal.

IV.2. EXISTENCE D'UN FILTRE OPTIMAL.

2.1. CHOIX D'UN ESPACE, CHOIX D'UNE NORME.

On note par $L^p_w[-1,1]$, $(p \geq 1)$ — ou plus simplement $L^p[-1,1]$ — l'espace vectoriel (des classes d'équivalence) des fonctions de puissance $p^{ième}$ intégrable (avec la fonction poids w(x) telle que $w(x) > 0$, $\forall x \in [-1,1]$), et on pose :

$$\|h\|_p = \left(\int_{-1}^{+1} w(x)|h(x)|^p dx\right)^{1/p}$$

pour tout $h \in L^p[-1,1]$.

Soient : P_n = {polynomes de degré $\leq n$}

$$P_n^+ = \{p \ / \ p \in P_n \ , \ p(x) > 0 \quad \forall x \in [-1,1]\}$$

$$P_n^o = \{p \ / \ p \in P_n \ , \ p(x) \geq 0 \quad \forall x \in [-1,1]\}$$

et soit alors :

$$R_n^m = \{p/q \ / \ p \in P_m^o \ , \ q \in P_n^+\} \ .$$

On se propose de déterminer $r^* \in R_n^m$, s'il existe, tel que :

$$\|r^*-f\|_p \leq \|r-f\|_p \qquad \forall r \in R_n^m$$

(Pour obtenir g^* — défini en IV.1 — il suffit de faire le changement de variable $x = \cos u$).

2.2. EXISTENCE DE r^* .

(Une partie de la démonstration de l'existence de r^* est empruntée à l'article de Cheney - Goldstein [10]).

Soit g une fonction définie sur $[-1,1]$, et soit l'ensemble défini par

$$S(g) = \{x \;/\; x \in [-1,1] \quad , \quad g(x) \neq 0\}$$

Soit

$$f_T(x) = \begin{cases} f(x) & x \in T \\ 0 & x \notin T \end{cases}$$

T étant un sous ensemble quelconque fermé de $[-1,1]$, et soient P,Q deux espaces vectoriels sur $\mathbb{R}$ de fonctions définies sur $[-1,1]$ et inclus dans $L^p[-1,1]$. (Au lieu de P on pourra considérer P^o le cône des éléments positifs ou nuls de P). On note :

$$Q^+ = \{q \;/\; q \in Q \;,\; q(x) > 0 \quad \forall x \in [-1,1]\}$$

On dira que :

P (ou P^o), Q et la norme $\|\ \|_p$ possèdent la <u>propriété C_p</u> si

pour tout $f \in L_w^p[-1,1]$ et $p \in P$ (ou P^o) , $q \in Q$: $\|(f-p/q)_T\|_p \leq \lambda$ pour tout $T \subset S(q)$

$\Rightarrow$ $\exists\ p_o \in P$ (ou P^o) , $q_o \in Q^+$ tel que $\|f-p_o/q_o\|_p \leq \lambda$

Posons $r(x) = \frac{p(x)}{q(x)}$, pour $x \in [-1,1]$ avec $p \in P_m$ (ou P_m^o) , $q \in P_n$

$r : x \to r(x)$ est définie sauf en les zéros de q. On a :

PROPOSITION 1 :

r est un élément de $L^p_w[-1,1]$, $(w > 0)$, si et seulement si l'ensemble des zéros de q dans $[-1,1]$, est contenu dans l'ensemble des zéros de p dans $[-1,1]$. (Les zéros de p identiques à ceux de q ayant une mutiplicité supérieure ou égale à ceux de q).

Condition suffisante :

Soient $\bar{x}_{j_i}$ $i=1,\ldots,k$ $(k \leq n)$ les racines de q sur $[-1,1]$.

On peut écrire :

$$r(x) = \frac{p(x)}{q(x)} = \frac{p_1(x)}{q_1(x)} \; \frac{\prod_{i=1}^{k} (x-\bar{x}_{j_i})}{\prod_{i=1}^{k} (x-\bar{x}_{j_i})}$$

avec $q_1(x) > 0$ (ou < 0) $\forall x \in [-1,1]$, donc $r \in L^p_w[-1,1]$.

Condition nécessaire :

Si q admet un zéro $\bar{x}$ dans $[-1,1]$, (d'ordre k , $1 \leq k \leq n$) (que l'on suppose unique pour faciliter l'écriture), on peut écrire :

$$q(x) = (x-\bar{x})^k q_1(x) \qquad q_1(x) \neq 0 \qquad \forall x \in [-1,1]$$

Il est immédiat de voir que, puisque $p \geq 1$,

$$w(x) \left| \frac{p(x)}{(x-\bar{x})^k q_1(x)} \right|^p$$

n'est sommable sur $[-1,1]$ que, si $p(x) = (x-\bar{x})^k p_1(x)$.

Si, donc $r \in L^p_w[-1,1]$, c'est que :

$$r = \frac{p(x)}{q(x)} = \frac{p_1(x)}{q_1(x)}$$

avec $q_1(x) \neq 0 \qquad \forall x \in [-1,1]$.

PROPOSITION 2 :

La propriété (C_p) est vérifiée lorsque :

$$P \equiv P_m \quad (ou\ P^o \equiv P_m^o\ )$$

$$Q \equiv P_n$$

Soit $\varphi(x) = \left|f(x) - \frac{p(x)}{q(x)}\right|^p w(x)$ avec $p \in P_m^o$.

Lorsque $q \in P_n$ on pose

$Z(q) = \{x_i \ / \ q(x_i) = 0$, distincts, appartenant à $[-1,1]\}$.

Soit $R_k(q) = \{x \ / \ x \in U\ ]x_i - \frac{1}{k}\ ,\ x_i + \frac{1}{k}[\quad (x_i \in Z(q))\}$ et

soit T_k la famille (infinie) de fermés tels que :

$$T_k = \complement R_k(q) \cap [-1,1] \qquad \text{pour } k \geq K > 2 \ / \ \min_{i,j} |x_i - x_j|$$

on a bien alors :

$T_k \subset S(q) \quad \forall\ k \geq K$, et

On pose :

$$\varphi_{T_k}(x) = \begin{cases} \left|f(x) - \frac{p(x)}{q(x)}\right|^p w(x) & x \in T_k \\ 0 & x \notin T_k \end{cases}$$

Ces fonctions φ_{T_k} sont sommables sur $[-1,1]$, convergent en mesure sur $[-1,1]$ vers φ , et la limite inférieure de leurs intégrales n'est pas $+\infty$, d'après l'hypothèse de la condition (C_p) ; alors, puisque $\varphi_{T_k}(x) \geq 0$ pour tout k , et tout $x \in [-1,1]$, d'après un corollaire du lemme de Fatou [39]

$$\int_{-1}^{+1} \varphi(x)\, dx \leq \lim_{k\to\infty} \inf \int_{-1}^{+1} \varphi_{T_k}(x)\, dx \ .$$

Utilisant la condition (C_p) on obtient :

$$\left|\int_{-1}^{+1} \varphi(x)dx\right|^{1/p} = \left|\int_{-1}^{+1} w(x)\left|f(x) - \frac{p(x)}{q(x)}\right|^p dx\right|^{1/p} \leq \lambda$$

φ est donc sommable, donc $(f - \frac{p}{q}) \in L^p_w[-1,1]$, c'est-à-dire puisque $p \in P^o_m$, et d'après la proposition 1

$$\exists\ p_o \in P^o_m \ , \ q_o \in P^+_n \ \text{ tel que } \ \left\|f - \frac{p_o}{q_o}\right\|_p \leq \lambda$$

THEOREME 1 :

Tout élément f de $L^p[-1,1]$ possède un meilleur approximant r^* dans R^m_n , c'est-à-dire :

$$\exists\ r^* \in R^m_n \ \text{ tel que } \ \|f-r^*\|_p \leq \|f-r\|_p \qquad \forall r \in R^m_n$$

Au cours de la démonstration, on utilisera les notations

$$\|g\|_T = \max_{x\in T} |g(x)| \quad \text{et} \quad \|g\|_\infty = \max_{x\in[-1,1]} |g(x)| \quad \text{pour } g \in C[-1,1]$$

(Si $p \in P^o_m$, on notera encore par $\|p\|$ la trace sur P^o_m de la norme $\|\ \|$ définie sur P_m) .

On commence par établir l'existence de deux constantes positives α et β , telles que :

(1) $\|g_T\|_p \leq \alpha\|g\|_T$ $\quad g \in C[-1,1]$, pour tout fermé $T \subset [-1,1]$

(2) $\left\|\frac{p}{q}\right\|_p \geq \beta\|p\|_\infty \ / \ .\|q\|_\infty \qquad p/q \in R^m_n$.

La norme $\| \ \|_p$ est une norme monotone , c'est-à-dire

si $\quad |f(x)| \leq |g(x)| \qquad \forall x \in [-1,1]$

on a $\quad \|f\|_p \leq \|g\|_p$

Pour montrer (1), on a, puisque $g \in C[-1,1]$:

$$|g(x)| \leq \|g\|_T \qquad \text{pour } x \in T$$

donc :

$$\|g_T\|_p \leq \|(\|g\|_T)\|_p = \|1\|_p \, \|g\|_T \qquad \text{donc} \quad \alpha = \|1\|_p$$

Pour montrer (2), on utilise le fait que toutes les normes sur P_m, de dimension finie, sont topologiquement équivalentes [4] donc pour tout $p \in P_m$, $\exists\ \beta > 0$ tel que $\|p\|_p \geq \beta \|p\|_\infty$. Inégalité qui est donc vraie en particulier pour $p \in P_m^o \subset P_m$. De plus pour tout $p/q \in R_n^m$, on a :

$$\left|\frac{p(x)}{q(x)}\right| \geq \frac{p(x)}{\|q\|_\infty}$$

donc, à cause de la monotonie

$$\left\|\frac{p}{q}\right\|_p \geq \left\|\frac{p}{\|q\|_\infty}\right\|_p = \frac{1}{\|q\|_\infty} \|p\|_p \geq \beta \frac{\|p\|_\infty}{\|q\|_\infty}$$

Prouvons maintenant le théorème 1. On pose :

$$\lambda = \inf_{r \in R_n^m} \|f-r\|_p$$

D'après la définition de la borne inférieure, il existe une suite d'éléments $r_k \in R_n^m$ telle que la suite des nombres réels $\lambda_k = \|f-r_k\|_p$ converge vers λ en décroissant.

Posons $r_k = p_k / q_k$ $(p_k \in P_m^o$, $q_k \in P_n^+)$. Il n'y a aucune perte de généralité à assurer que $\|q_k\|_\infty = 1$. Puisque la suite $\|f-r_k\|_p$ est bornée,

la suite $\|r_k\|_p$ est bornée.

Par l'inégalité (2), la suite $\|p_k\|_\infty$ est bornée. Alors en passant aux sous suites, si nécessaire, on peut assurer que les suites p_k et q_k convergent en norme $\|\ \|_\infty$, c'est-à-dire uniformément ; soit $p_k \to p \in P_m^o$ et $q_k \to q$.

Si $q \in P_n^+$ la fonction $r_o = p/q$ a la propriété $\|f-r_o\|_p = \lambda$.

En effet $\|r_k-r_o\|_\infty \to 0$ donc par l'inégalité (1) $\|r_k-r_o\|_p \to 0$ et de l'inégalité

$$\lambda \leq \|f-r_o\|_p \leq \|f-r_k\|_p + \|r_k-r_o\|_p$$

on tire la conclusion indiquée.

Si $q \notin P_n^+$ soit T un ensemble fermé quelconque de $[-1,1]$ tel que $q(x) > 0$ sur T. Alors $r_k \to r = p/q$ uniformément sur T.

Donc $\|(r_k-r)_T\|_p \to 0$ et $\|(f-r)_T\|_p \leq \lambda$ comme précédemment.

Par la propriété (C_p), il existe $p^* \in P_m^o$ et $q^* \in P_n^+$ tels que

$$\|f-p^*/q^*\|_p \leq \lambda$$

IV.3. PROPRIETES DU MEILLEUR APPROXIMANT DANS $L^2[-1,1]$

3.1. r^* EST UN ELEMENT NORMAL

On note le degré d'un polynome p par ∂p. Un élément $r = p/q$ de R_n^m sera dit normal si $\partial p = m$ ou $\partial q = n$, et non normal sinon.

LEMME 1 :

Soit $f \in L^2[-1,1]$

si $\int_{-1}^{+1} w(x)\, \frac{f(x)}{x-\alpha}\, dx = 0 \quad \forall \alpha \in [-2-\eta,-1-\eta]$ (η *nombre donné* > 0)

alors $f(x) = 0$ *p.p sur* $[-1,1]$

On peut écrire :

$$I = \int_{-1}^{+1} w(x) \frac{f(x)}{x-\alpha} dx = -\frac{1}{\alpha} \int_{-1}^{+1} w(x) \frac{f(x)}{(1-\frac{x}{\alpha})} dx = -\frac{1}{\alpha} \int_{-1}^{+1} w(x) \sum_{0}^{\infty} (\frac{x}{\alpha})^n dx$$

puisque $|\frac{x}{\alpha}| < 1$, on a en posant $\beta = \frac{1}{\alpha}$

$$I = -\beta \sum_{n=0}^{\infty} [\int_{-1}^{+1} w(x)\, x^n f(x)dx]\, \beta^n$$

puisque $I = 0 \quad \forall \beta \in [-\frac{1}{1+\eta}, \frac{-1}{2+\eta}]$, on a nécessairement

$$\int_{-1}^{+1} w(x)\, x^n f(x)dx = 0 \qquad n=0,1,2,\ldots$$

Donc $f(x) = 0$ p.p. sur $[-1,1]$

THEOREME 2 :

Soit $f \in L^2[-1,1]$, $(f \notin R_n^m)$

Un élément non-normal de R_n^m ne peut pas être un meilleur approximant de f.

Soit $r^* = \frac{p}{q}$ un meilleur approximant de f, supposons que r^* soit non-normal, c'est-à-dire, $\partial p < m$ et $\partial q < n$.

Soit η un nombre (> 0) fixé, et soit :

$$s(x) = x-\alpha \qquad -2-\eta \leq \alpha \leq -1-\eta$$

Pour $\lambda \in [-\eta,1]$ on considère

$$r_\lambda = \frac{p}{q} (1 + \frac{\lambda}{s(x)})$$

$r_\lambda \in R_n^m$, en effet $qs \in P_n^+$

et $p(x-\alpha+\lambda) \geq 0 \quad \forall x \in [-1,1]$ donc $\in P_m^o$.

Soit alors :

$$\phi(\lambda) = \int_{-1}^{+1} w(x)[r_\lambda - f]^2 \, dx$$

puisque $r = r^*$ est un meilleur approximant on doit avoir $\phi(\lambda)$ minimum pour $\lambda = 0$ donc :

$$\phi'(0) = \int_{-1}^{+1} w(x) \frac{p(x)}{q(x)} [r(x)-f(x)] \frac{1}{s(x)} \, dx = 0 \qquad \forall \alpha$$

C'est-à-dire d'après le lemme 1 on doit avoir :

$$\frac{p(x)}{q(x)} [r(x)-f(x)] = 0 \qquad \text{p.p. sur } [-1,1]$$

Comme $p(x)$ n'est pas le polynome identiquement nul, on a nécessairement

$$f(x) = r(x) \text{ p.p.}$$

ce qui est contraire à l'hypothèse, donc un élément non normal de R_n^m ne peut être un meilleur approximant.

3.2. PROPRIETE DE r^*.

Soit $r^* = \frac{p^*}{q^*}$ un point de R_n^m où $\|r-f\|$ ($\| \|$ sera désormais utilisée pour désigner $\| \|_2$ (c'est-à-dire la norme dans $L^2[-1,1]$)) est minimum. Puisque $p^*(x) \geq 0 \quad \forall x \in [-1,1]$ on peut écrire :

$$p^*(x) = \tilde{p}(x).p_o(x)$$

avec :

$$\partial\tilde{p} = k \qquad 0 \leq k \leq m \qquad \tilde{p} \text{ possède k racines sur } [-1,1]$$
$$\text{et } \tilde{p}(x) \geq 0 \qquad \forall x \in [-1,1]$$

et : $\partial p_o = m-k$ où $p_o(x) > 0 \quad \forall x \in [-1,1]$

PROPOSITION 3 :

Si $r^* = \tilde{p}\,\frac{p_o}{q^*} \in R_n^m$ *est un minimum (global) de* $\|r-f\|$ *alors :*

$$\int_{-1}^{+1} w(x)\ \frac{\tilde{p}(x)}{(q^*(x))}\,(r^*(x)-f(x))(p_o(x)q(x)+p(x)q^*(x))dx = 0$$

pour tout $p \in P_{m-k}$ *et* $q \in P_n$

Avant de montrer la proposition 3, montrons :

LEMME 2 :

Il existe $\eta_1 > 0$, $\eta_2 > 0$ *tels que :*

$$\forall\lambda \in [-\eta_1,\eta_1] \qquad p_o(x) + \lambda x^i \geq 0 \qquad \forall x \in [-1,1]$$

$$\forall\mu \in [-\eta_2,\eta_2] \qquad q^*(x) - \mu x^j > 0 \qquad \forall x \in [-1,1]$$

pour tout i , $(i=0,1,\ldots,m-k)$ *et pour tout* j, $(j=0,\ldots,n)$.

Par hypothèse, $p_o(x) > 0 \quad \forall x \in [-1,1]$, il suffit alors de prendre $\eta_1 = \min_{x\in[-1,1]} p_o(x)$, pour montrer la première partie du lemme.

Pour la seconde partie, $q^*(x)$ étant aussi strictement positif il suffit de prendre :

$$\eta_2 = \min_{x\in[-1,1]} q^*(x) / 2 \ .$$

On considère alors :

$$\phi(\lambda) = \int_{-1}^{+1} w(x)\left[\frac{\tilde{p}(x)}{q^*(x)}\,(p_o(x)+\lambda x^i)-f(x)\right]^2 dx \quad \text{pour} \quad \lambda \in [-\eta_1,\eta_1]$$

et

$$\Psi(\mu) = \int_{-1}^{+1} w(x)\left[\frac{p^*(x)}{q^*(x)-\mu x^j} - f(x)\right]^2 dx \qquad \text{pour} \quad \mu \in [-\eta_2,\eta_2]$$

puisque p^* / q^* est un meilleur approximant on doit nécessairement avoir :

$$\phi'(0) = 0 \qquad i=0,1,\ldots,m-k$$

$$\Psi'(0) = 0 \qquad j=0,1,\ldots,n$$

c'est-à-dire :

$$\int_{-1}^{+1} w\,\frac{\tilde{p}(x)}{(q^*(x))^2}\,q^*(x)x^i[r^*(x)-f(x)]dx = 0 \qquad i=0,1,\ldots,m-k$$

et

$$\int_{-1}^{+1} w\,\frac{\tilde{p}(x)}{(q^*(x))^2}\,p_o(x)x^j[r^*(x)-f(x)]dx = 0 \qquad j=0,1,\ldots,n$$

ce qui, après des combinaisons linéaires de ces relations, montre la proposition 3.

PROPOSITION 4 :

L'ensemble des polynomes

$p_o(x)q(x) + p(x)q^*(x)$

engendre P_{n+m-k} *, lorsque* p *et* q *varient dans* P_{m-k} *et* P_n *.*

D'après le théorème 2, les polynomes $p^*(x)$ et $q^*(x)$ sont premiers entre eux donc il en est de même de $p_o(x)$ et $q^*(x)$. La démonstration se trouvant entièrement dans [10], ne sera pas transcrite ici.

THEOREME 3 :

Si $r^* = \tilde{p}\,\frac{p_o}{q^*}$ $(\in R_n^m)$ *est un meilleur approximant de* $f \in L^2$ *alors :*

$$\int_{-1}^{+1} w(x)\,\frac{\tilde{p}(x)}{(q^*(x))^2}\,[r^*(x)-f(x)]\,x^j\,dx = 0 \qquad \forall j = 0,1,\ldots,m+n-k$$

La démonstration de ce théorème est évidente, en se servant des propositions 3 et 4 précédentes.

3.3. ALTERNANCES DE $e(x) = r^*(x)-f(x)$

THEOREME 4 :

Soit r^* *l'élément de* R_n^m *qui réalise la (ou une) meilleure approximation de f, dans* $L^2[-1,1]$ *, avec* $f \notin R_n^m$.

Si f(x) est continue par morceaux, et
si $r^*(x)$ *possède k zéros* $(0 \leq k \leq m)$, *sur* $[-1,1]$, *alors :*

$e(x) = r^*(x)-f(x)$ *possède au moins* $n+m-k+1$ *changements de signe sur* $]-1,1[$.

Supposons que $r^*(x)$ possède k zéros sur $[-1,1]$, $0 \leq k \leq m$, écrivant à nouveau $r^*(x) = \tilde{p}(x)p_o(x)/q^*(x)$, on pose :

$$\omega(x) = w(x)\tilde{p}(x) \;/\; (q^*(x))^2 \geq 0 \qquad \forall x \in [-1,1]$$

D'après le théorème 3 on doit avoir :

$$I_j = \int_{-1}^{+1} \omega(x)e(x)x^j dx = 0 \qquad j=0,1,\ldots,n+m-k = \ell$$

a) e(x) possède au moins un changement de signe sur $[-1,1]$:

en effet :

$$I_o = \int_{-1}^{+1} \omega(x)e(x)dx = 0$$

ce qui entraine, d'après le théorème de la moyenne, que e(x) ($\not\equiv 0$) possède au moins un changement de signe sur $]-1,1[$, puisque $\omega(x) \geq 0$ ($\not\equiv 0$) $\forall x \in [-1,1]$.

b) e(x) est continue sur

(θ_j,θ_{j+1}) $\quad j=0,1,\ldots,r$

en posant $\theta_o = -1$ et $\theta_{r+1} = +1$

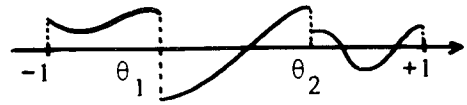


Figure IV.1

Montrons alors que l'hypothèse :

"e(x) possède moins de $(\ell+1)$ changements de de signe sur $[-1,1]$" est absurde.

Par hypothèse $I_j = 0$ $j=0,1,\ldots,\ell-1,\ell$.

Supposons que e(x) change de signe uniquement en

$$\eta_1, \eta_2, \ldots, \eta_s \qquad 1 \leq s \leq \ell$$

et posons

$$\eta_o = -1 \quad , \quad \text{et} \quad \eta_{s+1} = +1 \quad , \quad \text{et}$$

$$z(x) = (x-\eta_1),\ldots,(x-\eta_s) \; .$$

On remarque que l'on peut écrire :

$$x^{\ell} = (x-1)^{\ell-s}(x-\eta_1).....(x-\eta_s) + Q_{\ell-1}(x) \qquad \partial^{\circ}Q_{\ell-1} \leq \ell-1$$

donc :

$$I_{\ell} = \int_{-1}^{+1} \omega(x)(x-1)^{\ell-s} z(x)e(x)dx + \int_{-1}^{+1} \omega(x)Q_{\ell-1}(x)e(x)dx$$

le dernier terme étant nul par hypothèse : on a donc :

$$I_{\ell} = \sum_{i=0}^{s} \int_{\eta_i}^{\eta_{i+1}} \omega(x)(x-1)^{\ell-s} z(x)e(x)ds$$

et le tableau suivant :

si $x \in$	$[\eta_o,\eta_1]$	$[\eta_1,\eta_2]$	....	$[\eta_j,\eta_{j+1}]$	....	$[\eta_s,\eta_{s+1}]$
signe $z(x)$	-	+		$(-1)^{j+1}$		$(-1)^{s+1}$
si signe $e(x)$	+	-		$(-1)^{j}$		$(-1)^{s}$
alors signe de $z(x)$ $e(x)$	-	-		-		-
si signe $e(x)$	-	+		$(-1)^{j+1}$		$(-1)^{s+1}$
alors signe $z(x)$ $e(x)$	+	+		+		+

(Dans chaque cas il faut lire sous entendu "ou nul").

Puisque $e(x) \not\equiv 0$, $\exists p$ tel que sur $[\eta_p,\eta_{p+1}]$ $e(x) \not\equiv 0$, ce qui permet de voir immédiatement que :

$$I_{\ell} \neq 0$$

ce qui est contraire à l'hypothèse.

Donc nécessairement $e(x)$ possède au moins $(\ell+1)$ changements de signe.

IV.4. REMARQUES SUR L'UNICITE

Le problème de l'unicité n'est pas encore résolu, et la difficulté semble dépasser les connaissances actuelles. Ce problème est abordé par l'exhibition de quelques contre-exemples remarquables [34] [35] . (Dans le cas où il n'y a pas de contrainte sur le numérateur). On va cependant démontrer :

4.1. CAS OU m ET n SONT IMPAIRS

THEOREME 5 :

Si $f \in L^2[-1,1]$, et si de plus $f(x)$ et $w(x)$ sont deux fonctions paires, alors il existe au moins deux meilleurs approximants (distincts) de f dans $R^{2k+1}_{2\ell+1}$.

Soit $r^* = \dfrac{b_o+b_1x+\ldots+b_{2k+1}x^{2k+1}}{a_o+\ldots+a_{2\ell+1}x^{2\ell+1}} \in R^{2k+1}_{2\ell+1}$ un meilleur approximant de f. D'après le théorème 2 on a nécessairement

$$b^2_{2k+1} + a^2_{2\ell+1} \neq 0 \qquad (*)$$

La quantité $\|r-f\|$ est minimum pour $r=r^*$. On a alors :

$$\|r^*-f\|^2 = \int_{-1}^{+1} w(x) \left[\frac{b_o+b_1x+\ldots+b_{2k+1}x^{2k+1}}{a_o+\ldots+a_{2\ell+1}x^{2\ell+1}} - f(x)\right]^2 dx$$

En faisant le changement de variable $x = -u$ on obtient :

$$\|r^*-f\|^2 = \int_{-1}^{+1} w(u) \left[\frac{b_o-b_1u+\ldots-b_{2k+1}u^{2k+1}}{a_o-a_{2\ell+1}u^{2\ell+1}} - f(u)\right]^2 du = \|\bar{r}-f\|$$

avec $\bar{r}(u) = \dfrac{b_o-b_1u+\ldots-b_{2k+1}u^{2k+1}}{a_o-a_1u+..-a_{2\ell+1}u^{2\ell+1}}$.

On a $r^* \neq \bar{r}$

En effet si $r^*(u) \equiv \bar{r}(u) \quad \forall u \in [-1,1]$ on doit nécessairement avoir $b_{2k+1} = a_{2\ell+1} = 0$, ce qui est contraire à l'hypothèse (*)

EXEMPLE :

$$f(x) = \begin{cases} 1 & x \in [-\frac{1}{2}, \frac{1}{2}] \\ 0 & \text{sinon} \end{cases}$$

Cherchons dans R_1^1 un meilleur approximant de f.
On trouve :

$$r^*(u) = \frac{0.561\ (1 + u)}{1 + 0.828\ u}$$

et aussi :

$$\bar{r}(u) = \frac{0.561\ (1 - u)}{1 - 0.828\ u}$$

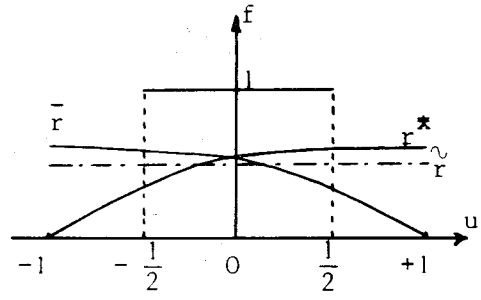


Figure IV.2

On remarque :

COROLLAIRE :

Une fonction paire, la fonction poids étant paire aussi, n'a pas dans $R_{2\ell+1}^{2k+1}$ un meilleur approximant pair.

Le meilleur approximant pair de f dans l'exemple précédent est

$\tilde{r} = 0.5$

On a en effet :

$$\|\bar{r}-f\|^2 = \|r^*-f\|^2 = 0.44662 < 0.5 = \|\tilde{r}-f\|^2 \ .$$

4.2. CAS GENERAL R_n^m

Ici on ne peut absolument rien dire. Une étude systématique dans R_2^o de fonctions f(x) telles que :

$$f(x) = \begin{cases} 1 & x \in [\alpha,\beta] \subset [-1,1] \\ 0 & \text{sinon} \end{cases}$$

a montré que pour chaque exemple, il existait un meilleur approximant unique, qui dans le cas d'une fonction paire est un meilleur approximant pair.

Commentaires sur les figures IV.3 à IV.6

Dans le cas de R_2^o , on doit minimiser :

$$\Psi = \int_{-1}^{+1} \left[f(x) - \frac{\alpha}{a_o+a_1x+a_2x^2}\right]^2 dx$$

la normalisation des coefficients de r étant telle que

$$a_o^2 + a_1^2 + a_2^2 = 1$$

c'est-à-dire, en conservant les variables indépendantes a_2 (axe horizontal) et a_1 (axe vertical) on doit avoir :

$$a_1^2 + a_2^2 \leq 1 \qquad \text{(figure IV.3 et IV.5)}$$

De plus, on doit avoir aussi :

$$\sqrt{1-a_1^2-a_2^2} + a_1x + a_2x^2 > 0 \qquad \forall x \in [-1,1]$$

(le complémentaire de cette partie est représenté par la partie hachurée de la figure IV.3).

Ayant dans un premier temps minimisé Ψ par rapport à α, on trace quelques courbes de niveaux de la surface

$$\phi(a_1,a_2) = \operatorname{Min}_{\alpha} \Psi(a_1,a_2,\alpha) \text{ (figure IV.3 et IV.5)} .$$

Les figures IV.4 et IV.6 permettent d'avoir des agrandissements des zones de voisinage des minima de chaque exemple, et de voir ainsi que f possède un seul meilleur approximant.

INSTITUT DE MATHEMATIQUES APPLIQUEES DE GRENOBLE
CALCUL DE COURBES DE NIVEAUX A L'AIDE DE FONCTIONS SPLINES

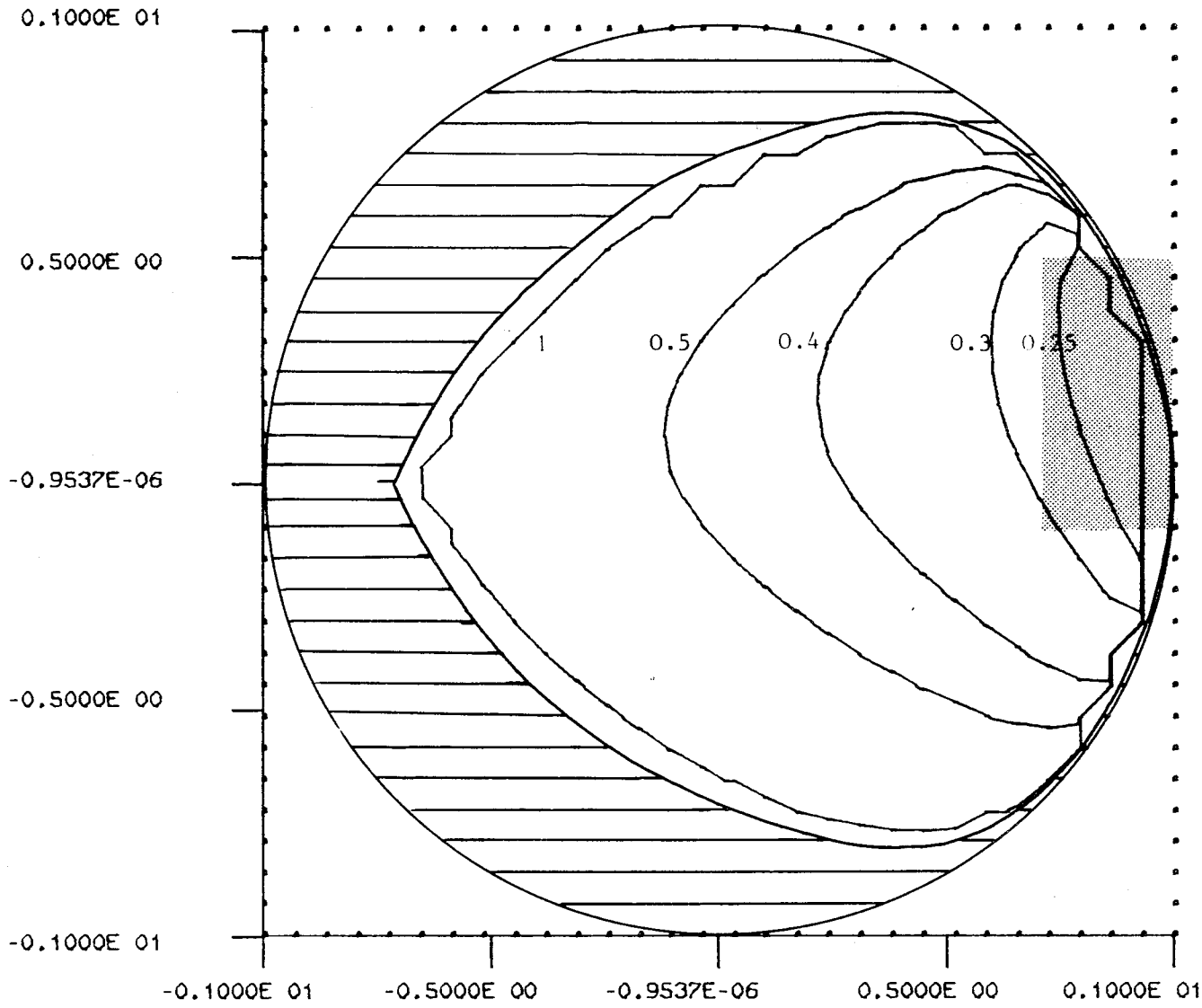


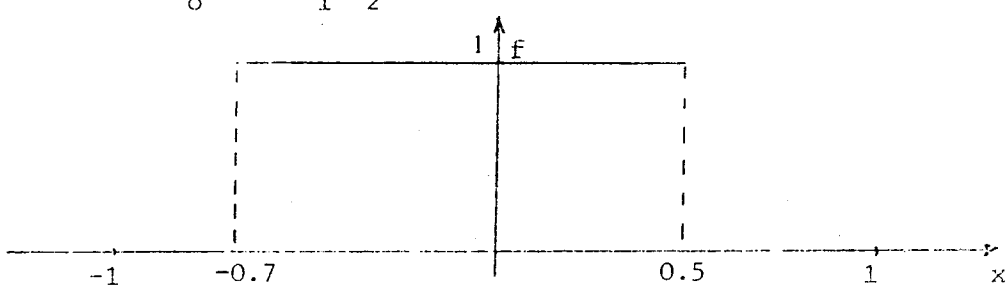
IMAGE DISPLAY NO 008 ECHELLE: 06/10

Figure IV.3

courbes de niveaux de la surface

$$\phi(a_1,a_2) = \min_{\alpha} \int_{-1}^{+1} \left[f(x) - \frac{\alpha}{a_o+a_1x+a_2x^2} \right]^2 dx = \int_{-1}^{+1} f^2(x)dx - \frac{\left(\int_{-1}^{+1} \frac{f(x)dx}{a_o+a_1x+a_2x^2} \right)^2}{\int_{-1}^{+1} \frac{dx}{(a_o+a_1x+a_2x^2)^2}}$$

avec $a_o = \sqrt{1-a_1^2-a_2^2}$, lorsque



INSTITUT DE MATHEMATIQUES APPLIQUEES DE GRENOBLE
CALCUL DE COURBES DE NIVEAUX A L'AIDE DE FONCTIONS SPLINES

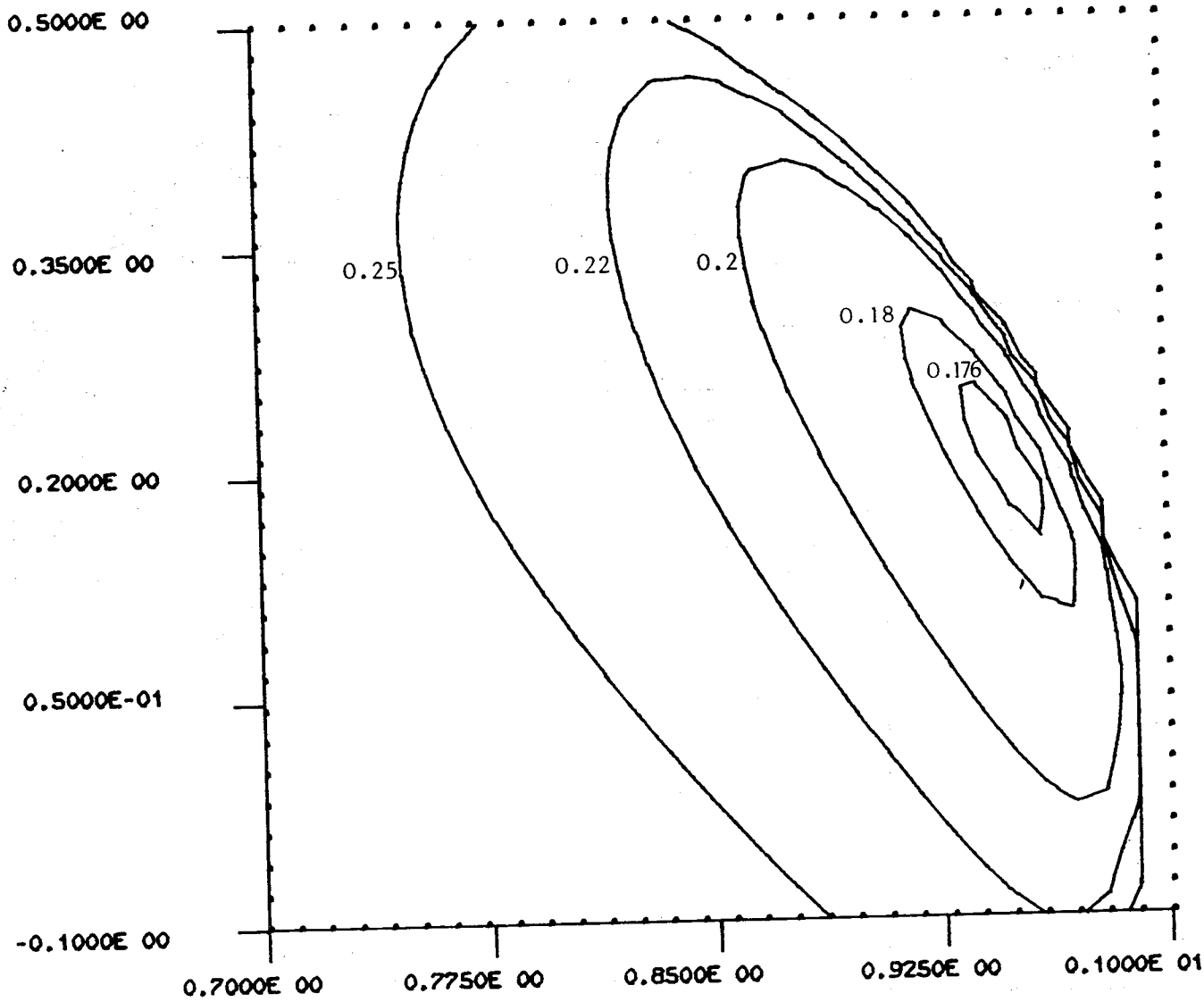


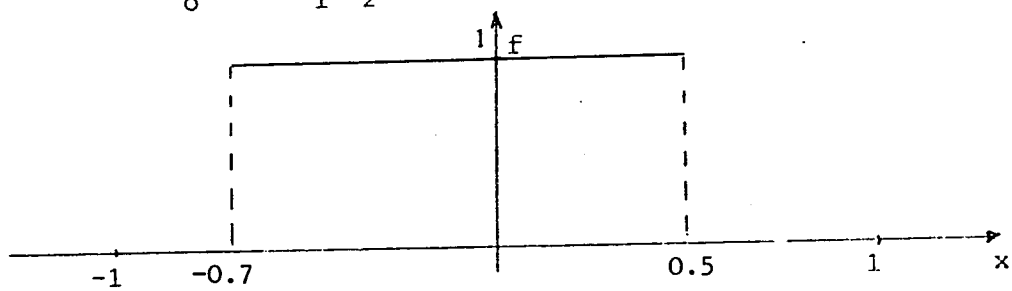
Figure IV.4

PAGE DISPLAY NO 009 ECHELLE: 06/10

courbes de niveaux de la surface (*)

$$\phi(a_1,a_2) = \min_{\alpha} \int_{-1}^{+1} \left[f(x) - \frac{\alpha}{a_o+a_1x+a_2x^2} \right]^2 dx = \int_{-1}^{+1} f^2(x)dx - \frac{\left(\int_{-1}^{+1} \frac{f(x)dx}{a_o+a_1x+a_2x^2} \right)^2}{\int_{-1}^{+1} \frac{dx}{(a_o+a_1x+a_2x^2)^2}}$$

avec $a_o = \sqrt{1-a_1^2-a_2^2}$, lorsque



(*) grossissement de la partie ombrée de la figure IV.3

INSTITUT DE MATHEMATIQUES APPLIQUEES DE GRENOBLE
CALCUL DE COURBES DE NIVEAUX A L'AIDE DE FONCTIONS SPLINES

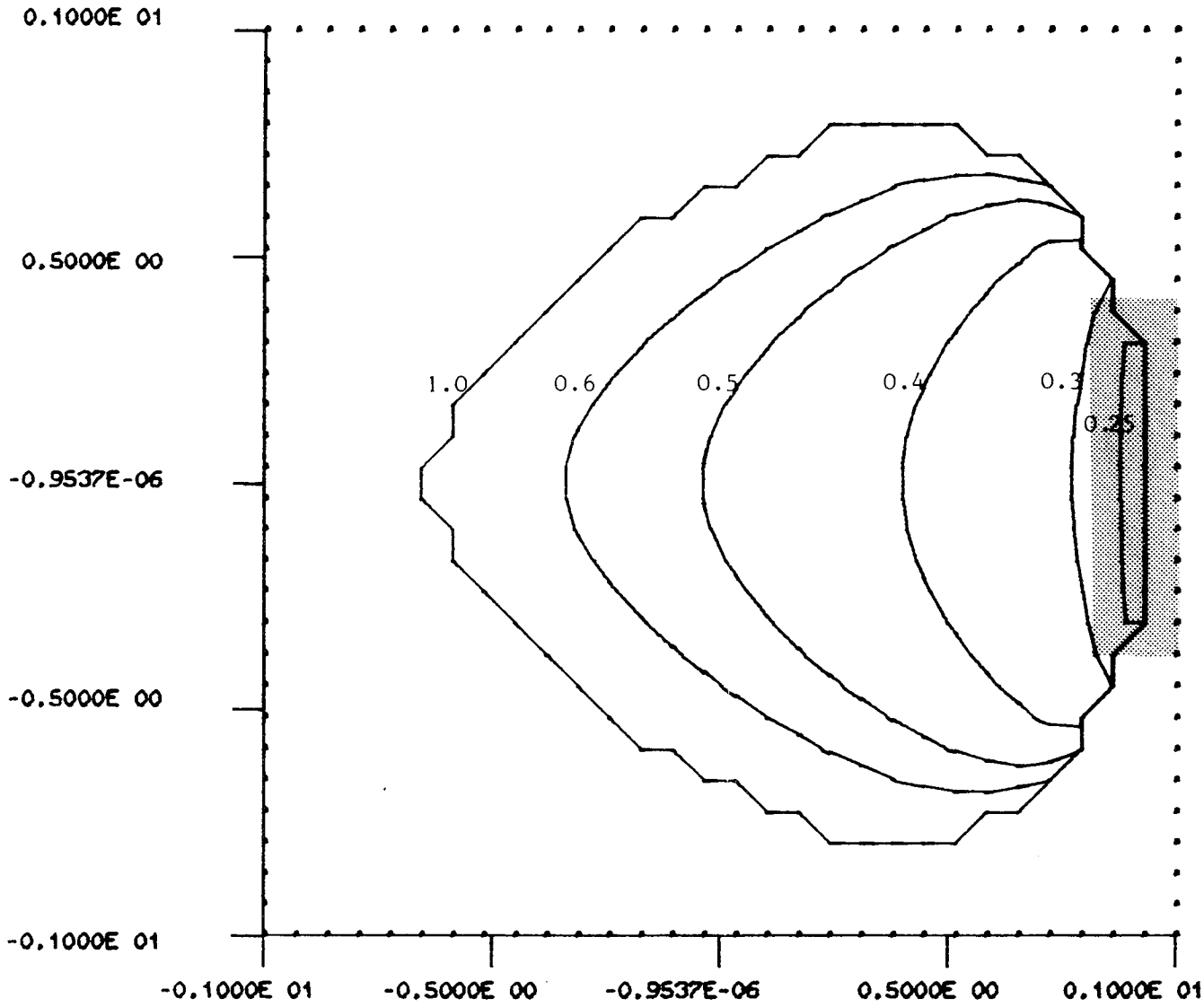


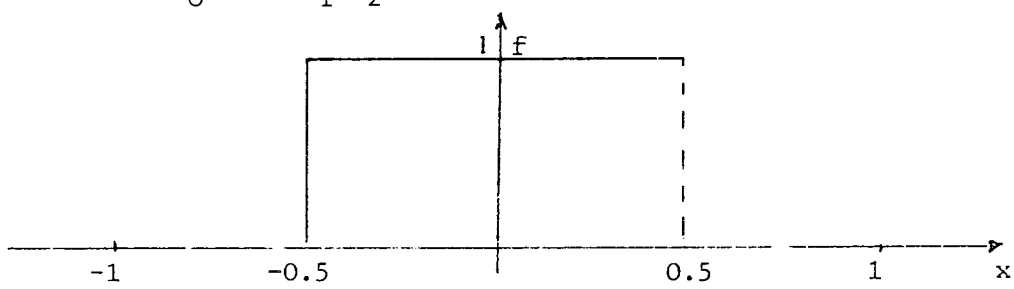
Figure IV.5

IMAGE DISPLAY NO 010 ECHELLE: 06/10

courbes de niveaux de la surface

$$\phi(a_1,a_2) = \min_{\alpha} \int_{-1}^{+1} \left[f(x) - \frac{\alpha}{a_o+a_1x+a_2x^2} \right]^2 dx = \int_{-1}^{+1} f^2(x)dx - \frac{\left(\int_{-1}^{+1} \frac{f(x)dx}{a_o+a_1x+a_2x^2} \right)^2}{\int_{-1}^{+1} \frac{dx}{(a_o+a_1x+a_2x^2)^2}}$$

avec $a_o = \sqrt{1-a_1^2-a_2^2}$, lorsque



INSTITUT DE MATHEMATIQUES APPLIQUEES DE GRENOBLE
CALCUL DE COURBES DE NIVEAUX A L'AIDE DE FONCTIONS SPLINES

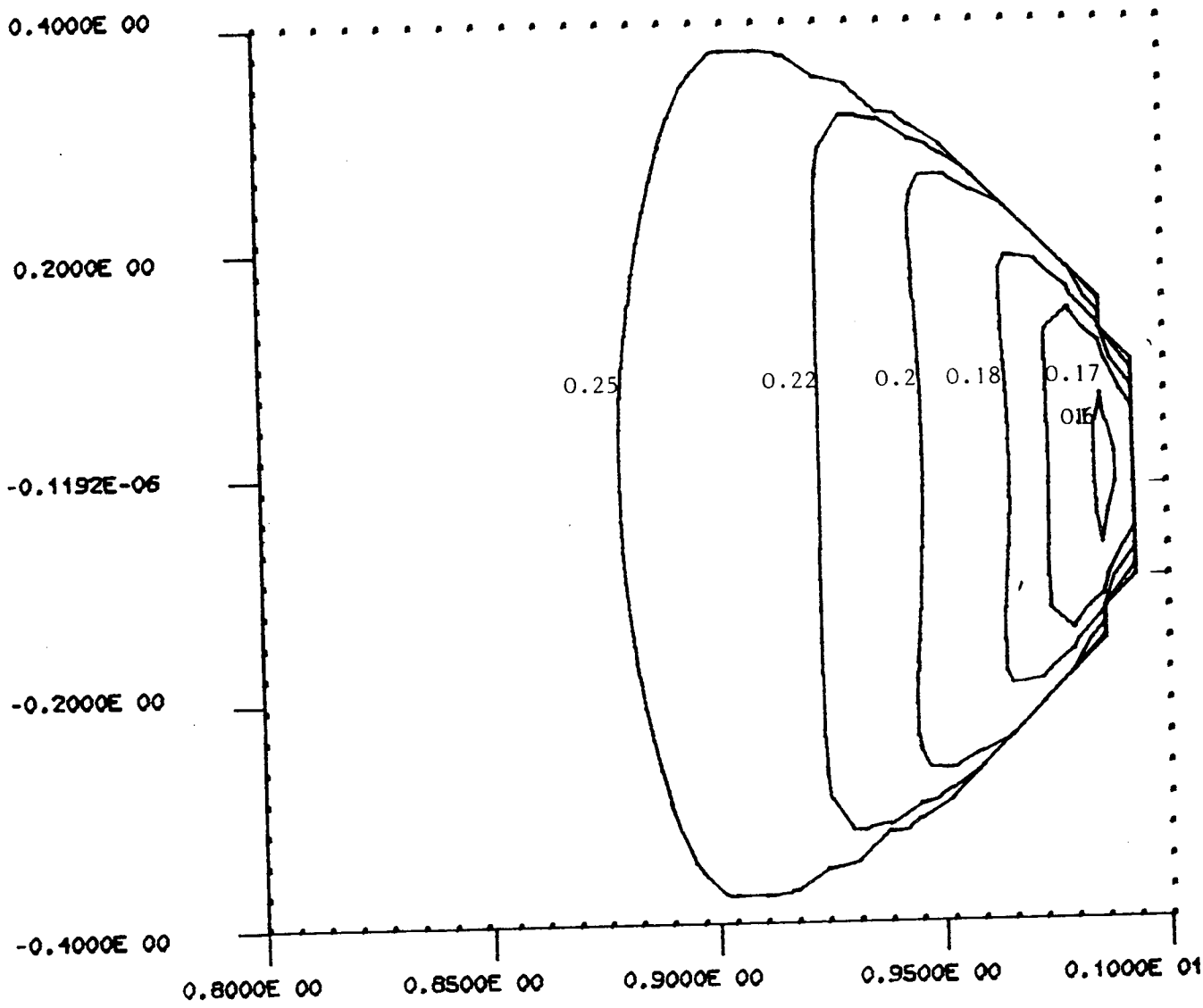


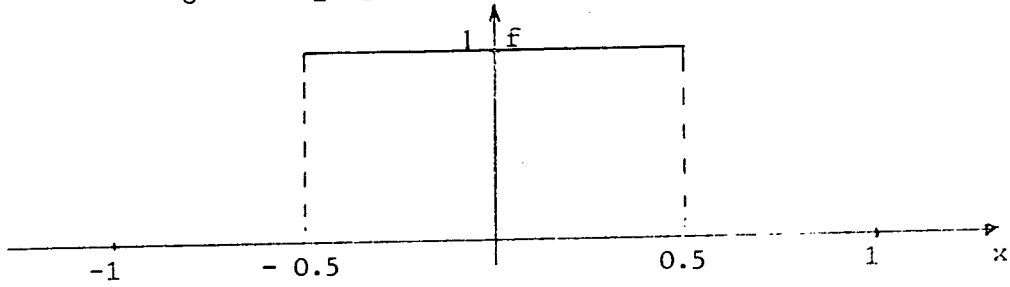
Figure IV.6

IMAGE DISPLAY NO 011 ECHELLE: 06/10

courbes de niveaux de la surface (*)

$$\phi(a_1,a_2) = \min_{\alpha} \int_{-1}^{+1} \left[f(x) - \frac{\alpha}{a_o+a_1x+a_2x^2} \right]^2 dx = \int_{-1}^{+1} f^2(x)dx - \frac{\left(\int_{-1}^{+1} \frac{f(x)dx}{a_o+a_1x+a_2x^2} \right)^2}{\int_{-1}^{+1} \frac{dx}{(a_o+a_1x+a_2x^2)^2}}$$

avec $a_o = \sqrt{1-a_1^2-a_2^2}$, lorsque



(*) grossissement de la partie ombrée de la figure IV.5

IV.5. PROPRIETE DU MEILLEUR APPROXIMANT (CAS PARTICULIERS)

On supposera que f est telle que :

$$f(x) = \begin{cases} 1 & x \in [\alpha,\beta] \\ 0 & x \in [-1,\alpha[\cup]\beta,1] \end{cases} \qquad \alpha<\beta$$

et on montre une propriété du meilleur approximant de f dans R^{o}_{2n} .

THEOREME 6 :

Soit $r^{*}(x) = 1 / q^{*}_{2n}(x)$ le meilleur approximant de f dans R^{o}_{2n} alors toutes les racines de $q^{*}_{2n}(x)$ sont complexes.

Montrons que la juxtaposition des hypothèses H_1 et H_2 est absurde

H1) $e(x) = f(x)-r^{*}(x)$ change de signe au moins (2n+1) fois sur $[-1,1]$ (théorème 4)

H2) $q^{*}(x)$ possède 2p racines réelles $(p \geq 1)$.

Soient $x_1,\ldots,x_\ell$ $(x_1 < \ldots < x_\ell)$ les racines réelles de q^{*} , x_i ayant pour multiplicité ℓ_i avec :

$$\sum_{i=1}^{\ell} \ell_i = 2p$$

* $\ell = 1$ $\quad q^{*}(x) = (x-x_1)^{2p} q_{2n-2p}(x)$

$$q'(x) = (x-x_1)^{2p-1}[2p\,q_{2n-2p}(x)+(x-x_1)q'_{2n-2p}(x)]$$

q' possède sur $(-\infty,x_1[$ (ou sur $]x_1,+\infty)$) au maximum 2n-2p racines.

* $\ell \neq 1$ sur $]x_i,x_{i+1}[$, $q'(x)$ possède au moins une racine et au maximum $(2n-1)-(\ell-1) - \sum_{i=1}^{\ell} (\ell_i-1)+1 = 2n-2p+1$ racines

sur $(-\infty,x_1[$ (ou sur $]x ,+\infty)$) , $q'(x)$ possède au maximum $(2n-1)-(\ell-1) - \sum_{i=1}^{\ell} (\ell_i-1) = 2n-2p$.

CAS 1

$$[-1,1] \subset]x_i, x_{i+1}[\text{ , donc } q(x) > 0 \quad \forall x \in]x_i, x_{i+1}[$$

Supposons $\exists k$ racines de q' dans $]x_i, \alpha] \cup [\beta, x_{i+1}[$

$k \geq 2$ il y a au maximum 2n-2p+1-2 racines de q' sur $]\alpha,\beta[$ donc le nombre de changement de signe de e(x) est au maximum

$$N \leq (2n-2p-1)+1+2 = 2n-2p+2 \leq 2n$$

puisque $p \geq 1$.
La figure IV.7 (ainsi que les suivantes) montre la configuration donnant le plus de changements de signe de e(x).

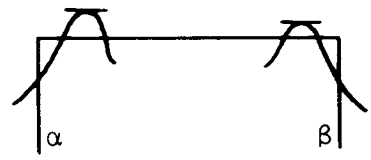


Figure IV.7

Donc ce résultat $N \leq 2n$ est incompatible avec H1.

$k = 1$ il y a au maximum 2n-2p+1-1 racines de q' sur $]\alpha,\beta[$ le nombre de changements de signe N est au maximum (figure IV.8)

$$N \leq (2n-2p)+1+1 \leq 2n$$

(puisque $x \xrightarrow{\text{si}} x_i^+$ (ou x_{i+1}^-) $q(x) \to +\infty$).

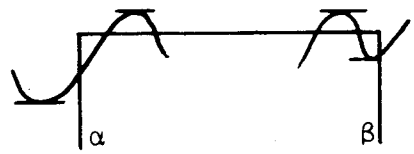


Figure IV.8

$k = 0$ il y a au maximum 2n-2p+1 racines de q' sur $]\alpha,\beta[$ dans ce cas (figure IV.9)

$$N \leq (2n-2p+1)+1 \leq 2n$$

.

ce qui est encore incompatible avec H1.

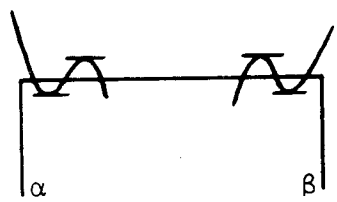


Figure IV.9

CAS 2

$$[-1,1] \not\subset]x_i, x_{i+1}[\qquad \forall i = 1, \ldots, -1$$

On montre encore et de la même manière que le nombre de changements de signe de e(x) sur $]\alpha,\beta[$ est encore au maximum 2n.

Ce qui est à nouveau incompatible avec H1.
Les cas 1 et 2 sont les seuls à envisager car q(x) ne peut s'annuler sur $[-1,1]$.

Les hypothèses H1 et H2 étant incompatibles le théorème 6 est ainsi montré.

REMARQUE :

Il n'est pas possible d'avoir un théorème identique lorsque le meilleur approximant de f appartient à R_{2n}^{2m} , en se servant de la seule hypothèse H1 (modifiée évidemment par le résultat du théorème 4). Il semble cependant qu'on doit pouvoir montrer que $q_{2n}(x)$ possède au maximum deux racines réelles.

Contre exemple :

$$r(x) = 8\,\frac{1-x^2}{7-x^2}$$

Les propriétés d'alternance et d'annulation de p_{2m} sont satisfaites (figure IV.10) et cependant q(x) possède deux racines réelles

$(\alpha = -3/4 \ , \ \beta = 3/4)$.

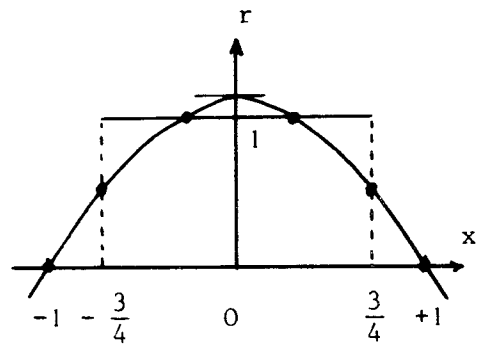


Figure IV.10

IV.6. CAS DE L'APPROXIMATION DISCRETE [44] PAR FRACTION RATIONNELLE.

Soit X = {N abscisses distinctes x_i i=1,2,...,N} , et les N couples donnés, à partir de X, (x_i,f_i) (i=1,2,...,N). On pose :

$$S_n^m = \{p/q \; ; \; p \in P_m \; , \; q \in P_n \text{ et } q(x) \neq 0 \; \forall x \in X\}$$

On se propose alors de chercher $r^* \in S_n^m$, si il existe , tel que :

$$F(r) = \sum_{i=1}^{N} w_i[r(x_i)-f_i]^2$$

soit minimum.

Suivant la répartition des couples (x_i,f_i), des degrés n et m, on ne pourra assurer l'existence d'un tel minimum, comme va le montrer ce qui va suivre.

EXEMPLE :

Choisissons $S_n^m \equiv S_1$ et les données suivantes

x_i	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1
f_i	1	0	0	0	0

$$r(x) = \frac{ax+b}{cx+d} = \frac{ax+b}{x+d} \quad (c=1)$$

$$Q(a,b,d) = \sum_{i=1}^{5} w_i\left[\frac{ax_i+b}{x_i+d} - f\right]^2$$

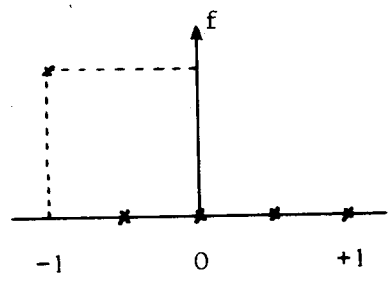


Figure IV.11

Il est immédiat de voir que

$$\lim_{n\to\infty} Q(\frac{1}{n}, \frac{2}{n}, 1+\frac{1}{n}) = 0$$

mais $\lim_{p\to\infty} (x+1+\frac{1}{p}) = x+1$, ce qui montre que la fraction rationnelle correspondante n'appartient pas à S !.

CHAPITRE V

ALGORITHMES DE CALCULS

THEOREME 7 :

Si les couples (x_{i_j}, f_{i_j}) $(j=1,2,\ldots p)$ appartiennent au graphe d'un polynome de degré $m-n$, et si $0 < N-p \leq n$ alors il n'existe pas r^* dans S_n^m $(n \geq 1)$ qui minimise $F(r)$.

Soit $t(x) = \sum_{k=0}^{m-n} \alpha_k x^k$ le polynome dont le graphe contient les couples (x_{i_j}, f_{i_j}) $(j=1,2,\ldots,p)$, on note :

$$\bar{X} = \{x \ / \ x_{i_j} \quad j=1,2,\ldots p\}$$

On note θ_j $(j=1,..,N-p)$ les abscisses de X n'appartenant pas à $\bar{X}$.
Soit alors :

$$r_\ell(x) = \sum_{k=0}^{m-n} (\alpha_k + \frac{1}{\ell})x^k \prod_{j=1}^{N-p} \frac{(x - \theta_j + \frac{\xi_j}{\ell})}{(x - \theta_j + \frac{1}{\ell})}$$

où $\xi_j = f_j \ / \ \sum_{k=0}^{m-n} \alpha_k \theta_j^k$ (on peut toujours supposer $t(\theta_j) \neq 0 \quad \forall j$)

$r_\ell(x) \in S_n^m$ car $\xi_j \neq 1$, sinon θ_j appartiendrait à $\bar{X}$.
Posons :

$$p_j(\ell,x) = \frac{x - \theta_j + \xi_j/\ell}{x - \theta_j + 1/\ell} = 1 - \frac{-\xi_j + 1}{\ell(x-\theta_j+1/\ell)}$$

on a :

$x \in \bar{X}$	$p_j(\ell,x) \to 1$	$\ell \to \infty$
$x = \theta_s \in X-\bar{X}$	$j \neq s \quad p_j(\ell,\theta_s) \to 1$	$\ell \to \infty$
	$j = s \quad p_s(\ell,\theta_s) = +\xi_s$	$\forall \ell$

On a :

$$r_\ell(x) = \sum_{k=1}^{m-n} \alpha_k x^k \prod p_j(\ell,x) + \frac{1}{\ell} \sum_{k=0}^{m-n} x^k \prod p_j(\ell,x)$$

et

$$Q_\ell = \sum_1^N w_i[r_\ell(x_i)-f_i]^2 = \sum_{x\in X} w_i[r_\ell(x_i)-f_i]^2 + \sum_{j=1}^{N-p} w_j[r_\ell(\theta_j)-f_j]^2$$

D'après les remarques précédentes et le choix de ξ_j, il est immédiat de voir que :

$$Q_\ell \to 0 \qquad \text{lorsque} \qquad \ell \to \infty$$

cependant le dénominateur :

$$\lim_{\ell\to\infty} \sum_{j=1}^{N-p} (x - \theta_j + \frac{1}{\ell}) = \prod_{j=1}^{N-p} (x - \theta_j)$$

s'annule en des points de X donc $\lim_{\ell\to\infty} r_\ell(x) \notin S_n^m$.

V .1. COMPLEXITE DU PROBLEME.

D'après le chapitre IV on sait qu'il existe $b \in \mathbb{R}^m$, $a \in \mathbb{R}^n$ tels que :

$$\varphi(a,b) = \int_0^\pi w(x)\left[\frac{b_o+b_1 \cos x + \ldots + b_m \cos^m x}{a_o+a_1 \cos x + \ldots + a_n \cos^n x} - f(x)\right]^2 dx$$

soit minimum lorsque :

$$b_o + \ldots + b_m \cos^m x \geq 0 \qquad \forall x \in [0,\pi] \qquad (A)$$

$$a_o + \ldots + a_n \cos^n x > 0 \qquad \forall x \in [0,\pi] \qquad (B)$$

A cause de (A) , la normalisation des coefficients sera choisie -- dans tout ce qui va suivre — telle que :

$$a_o = 1$$

(En effet pour $x = \pi/2$ on doit avoir $a_o > 0$).

REMARQUES :

1) Cette normalisation ne préjuge en rien de celle qui peut être éventuellement faite sur les coefficients du filtre, c'est-à-dire sur (α,β) .

2) Dans tous les exemples qui suivront f sera choisie continue par morceaux, c'est-à-dire dans les conditions d'application du théorème IV.4.

Le calcul de la meilleure approximation, c'est-à-dire du filtre optimal, est difficile pour deux raisons essentielles :

* La fonction $\varphi(a,b)$ à minimiser ne possède pas de "bonnes propriétés", qu'on espère trouver dans les problèmes de minimisation par exemple fonction convexe.

* On doit minimiser $\varphi(a,b)$ sous un ensemble de contraintes en nombre infini. Les domaines A et B ne sont d'ailleurs pas faciles à construire.

V .2. METHODES NUMERIQUES.

2.1. METHODE DE PENALISATION [26]

2.1.1.

On rappelle, brièvement, ce qu'est une méthode de pénalisation intérieure [45] :

On pose le problème: trouver z^o tel que :

$$f^o(z^o) = \min \{f^o(z) \mid z \in C \subset \mathbb{R}^n\} \qquad (1)$$

On suppose que :

* $\exists z' \in C$ tel que $Z' = \{z \mid f^o(z) \leq f^o(z')\}$ est compact.
* C est fermé, et possède un intérieur non vide.

Une suite de fonctions continues p_i ($C^o \to \mathbb{R}$) $i=0,1,2,\ldots$ (figure V .1.) est dite être une suite de fonctions de pénalisations intérieures pour l'ensemble C si

* $0 < p_{i+1}(z) < p_i(z) \qquad \forall z \in C^o \qquad i=0,1,\ldots$
* $p_i(z) \to 0$ quand $i \to \infty \qquad \forall z \in C^o$
* $p_i(z_j) \to \infty$ quand $j \to \infty$, pour toute suite $\{z^j\} \in C_o$ telle que $z^j \to z^* \in \partial C$, quand $j \to \infty$, $i=0,1,\ldots$

On remplace alors le problème (1) par la suite de problèmes :

Trouver z_i qui réalise :

(2_i) $\min\{f^o(z)+p_i(z) \mid z \in C^o\}$

Ce qui ramène donc le problème initial à une suite de problèmes sans contrainte.

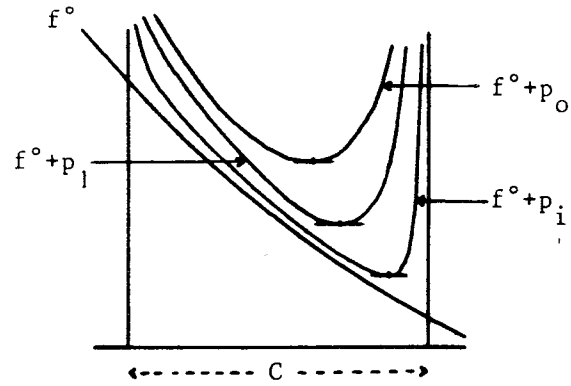


Figure V .1

Polak montre alors que :
Soit z_i le "point optimal" de (2_i) alors tout point d'accumulation $\hat{z}$ de la suite $\{z_i\}$ $i=0,\ldots,\infty$ est optimal pour le problème (1) avec contraintes.

2.1.2. Résultats numériques :

Les résultats obtenus ont été décevants, malgré la complexité du problème. Pour minimiser les problèmes (2_i) la méthode utilisée a été la méthode du gradient. Les pénalisations intérieures étant :

$$p_i(b) = \varepsilon_i \int_0^{\pi} \frac{dx}{b_o+\ldots+b_m \cos^m x} \quad \text{avec} \quad \varepsilon_{k+1} < \varepsilon_k \quad \forall k$$

et $\lim_{k\to\infty} \varepsilon_k = 0$. (Les contraintes sur a n'étant pas actives).

Cette méthode a été testée très en détail [62] , puis ensuite abandonnée pour diverses raisons :

- pour chaque i, lenteur de la convergence, lorsque l'on se rapproche de la solution, dûe au choix de la méthode du gradient,
- lorsque le programme était arrêté, il y avait peu de moyens de savoir si on était "près" ou "loin" de la solution optimale.

La figure V .2. en est un exemple frappant :

On cherche à approcher un filtre passe-bas (fréquence de coupure en $\pi/2$) par une fraction rationnelle dont les degrés des numérateur et dénominateur sont égaux à 4 .

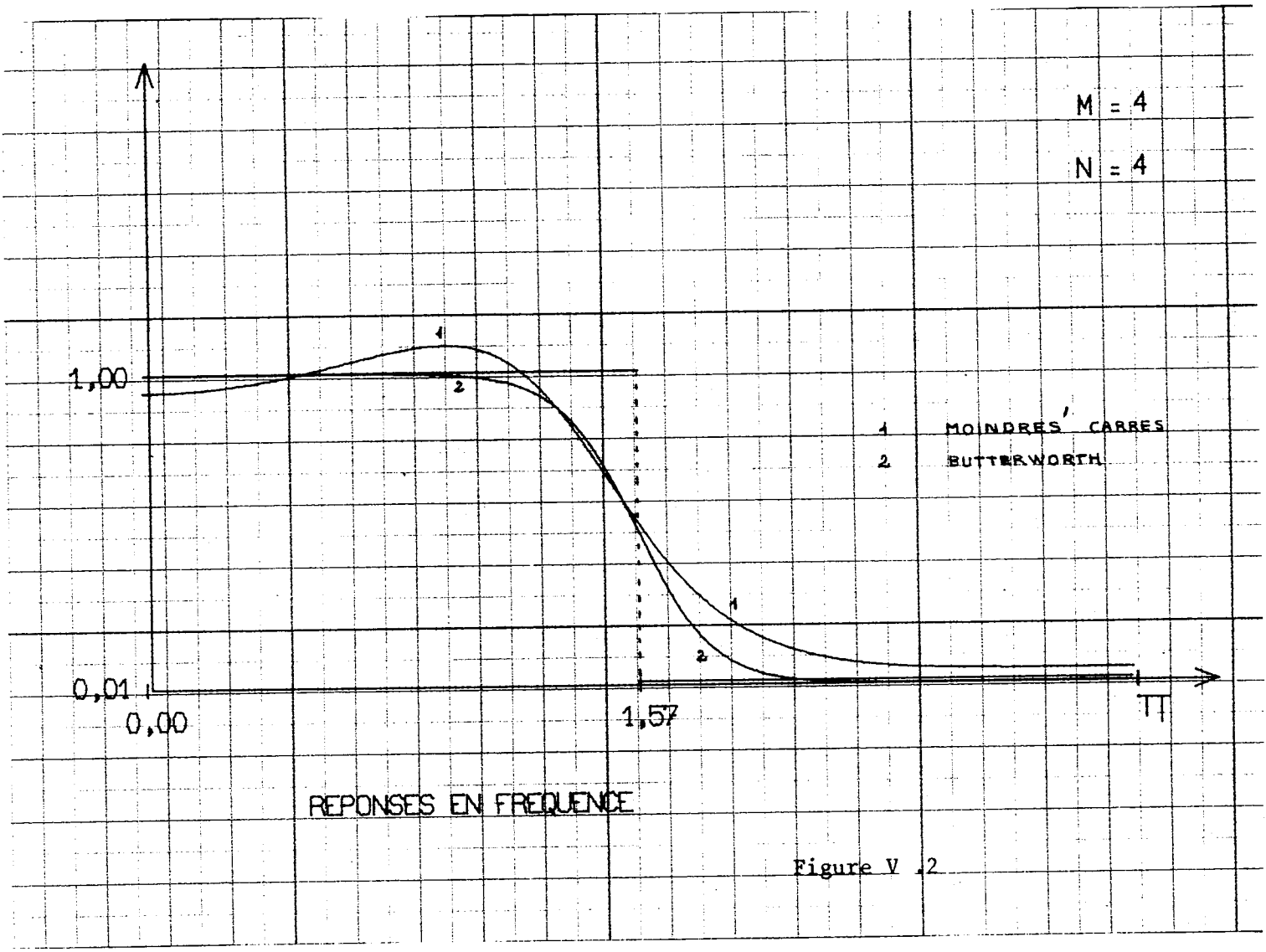


Figure V .2

D'après le théorème IV.4, la fonction erreur doit posséder $p = m+n+1 = 8$ changements de signe ou zéros. On voit sur la figure que ce nombre est ici égal à 3, ce qui montre que la fonction d'approximation obtenue est loin d'être optimale.

2.2. METHODE MIXTE, SANS CONTRAINTE

2.2.1.

Il est a remarquer que les contraintes sur les $a_1,\ldots,a_n$ ne sont pas actives.

Examinons le numérateur $P_m(\cos x) = \sum_{j=0}^{m} b_j \cos j\,x$,

l'ensemble des contraintes est :

$$B = \{b \;/\; P_m(\cos x) \geq 0 \quad \forall x \in [0,\pi]\}$$

Mais d'après le théorème 2 (III.6.1.2), ceci veut dire

$$\exists\ \beta_o,\beta_1,\ldots,\beta_m \text{ tel que } P_m(\cos x) = \Big(\sum_{0}^{m} \beta_k e^{ikx}\Big)\Big(\sum_{0}^{m} \beta_k e^{-ikx}\Big)$$

le polynome P_m s'écrit alors :

$$P_m(x) = (\beta_o^2+\ldots+\beta_m^2) + 2(\beta_o\beta_1+\ldots+\beta_{m-1}\beta_m)\cos x +\ldots+ 2\beta_o\beta_m \cos m\,x$$

On pose alors :

$$\phi(a,\beta) = \int_0^{\pi} w(x)\Big[\frac{(\beta_o^2+\ldots+\beta_m^2)+\ldots+2\beta_o\beta_m \cos m\,x}{1+a_1\cos x+\ldots+a_n\cos^n x} - f(x)\Big]^2 dx$$

et on minimise ϕ par rapport aux variables $\{a_1,\ldots,a_n,\beta_o,\ldots,\beta_m\}$, cette minimisation étant désormais une minimisation sans contrainte.

REMARQUE :

Il ne faut pas oublier que l'on recherche un filtre, c'est-à-dire les coefficients $\{\alpha_o,\ldots,\alpha_n\}$, $\{\beta_o,\ldots,\beta_m\}$. Le moyen précédent donne donc immédiatement les coefficients β_i , i=0,...,m , sans autre calcul comme il eut fallu en faire si les coefficients b_i avaient été calculés.

Cependant, est-il possible de choisir pour le dénominateur cette forme résolue ? On peut répondre non, car le minimum de φ ayant lieu au point $\{\alpha_o^*,\alpha_1^*,\ldots,\alpha_n^*,\beta_o^*,\ldots,\beta_m^*\}$, , cas de la forme résolue, il se peut alors que le polynome $\sum_{i=0}^{n} \alpha_i z^i$ soit instable, c'est-à-dire possède une (ou plusieurs) racine(s) de module supérieur ou égal à 1.

2.2.2. Résultats numériques

La méthode choisie pour minimiser est la méthode de Fletcher et Powell [16] . Le sous programme utilisé étant celui de la bibliothèque FORTRAN de I.B.M. [53].

Les résultats des essais expliqués ici se rapportent au modèle dont le carré du module est :

$$f(x) = \begin{cases} 1 & x \in [0,\frac{\pi}{2}] \\ 0 & \text{sinon} \end{cases}$$

La figure V .3 présente les approximations successives de f lorsque m et n varient (de telle manière que m+n = 4).

On appelle coût d'un filtre le nombre d'opérations nécessaires pour passer de $y(p\Delta T)$ à $y((p+1)\Delta T)$. Aussi à titre de comparaison on a tracé de plus le filtre de Butterwoth de même coût que les précédents.

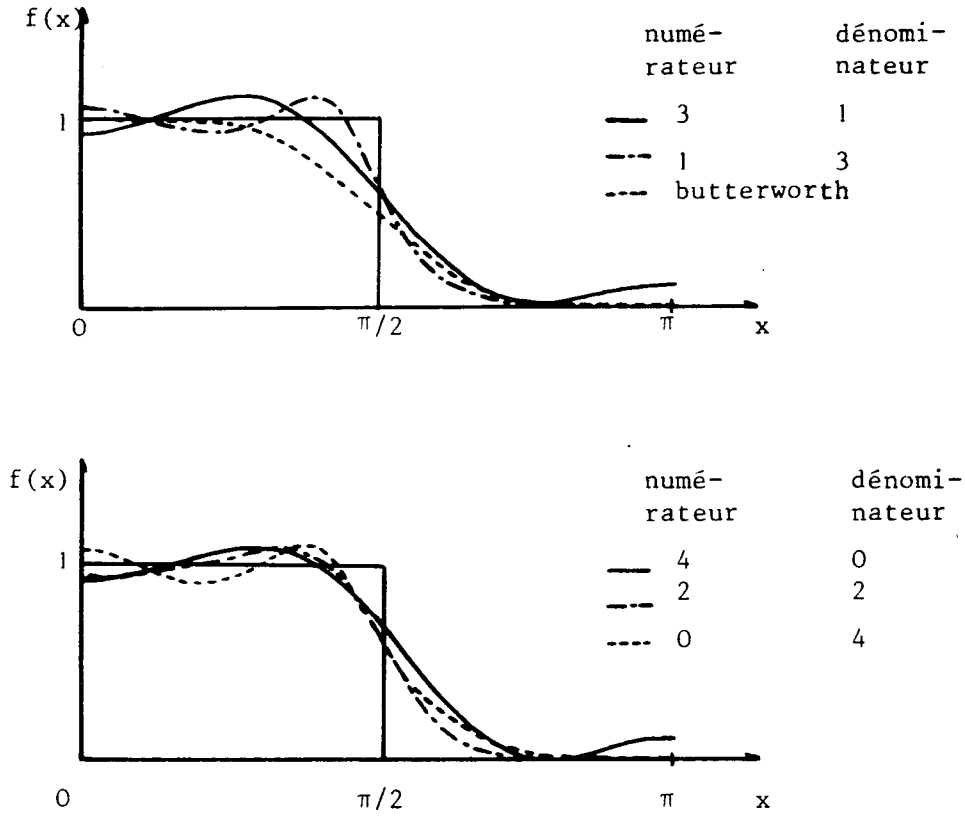


Figure V .3

On notera que les coefficients (α,β) des filtres étudiés n'ont pas été indiqués, en effet ici on est surtout intéressé par le côté théorie de l'approximation.

V .3. MINIMISATION AVEC CONTRAINTES

3.1. MINIMISATION D'UNE FORME QUADRATIQUE AVEC UNE INFINITE DE CONTRAINTES [70]

Soit $H = L^2_W(a,b)$ l'espace de Hilbert des fonctions de carré sommable par rapport à la fonction poids W $(W(x) \geq 0 \ \forall x \in [a,b])$, dont le produit scalaire est noté, f et g étant deux éléments quelconques de $L^2_W[a,b]$:

$$< f,g > = \int_a^b W(x)f(x)g(x)dx$$

Soit V le sous-espace vectoriel de dimension (n+1) de H engendré par $\varphi_o, \varphi_1, \ldots, \varphi_n$.

Soit (x,y) le produit scalaire usuel de $\mathbb{R}^{n+1}$

$$(x,y) = \sum_{i=0}^{n} x_i y_i$$

Soit C_1 le convexe de $\mathbb{R}^{n+1}$ défini par :

$$C_1 = \{x \in \mathbb{R}^{n+1} \ / \ \forall u \in [\alpha,\beta] \ , \ (x-\hat{x} \ , P(u)) \geq 0\}$$

avec :

$$P^T(u) = [P_o(u), \ldots, P_n(u)]$$

et tel que :

i) $P_i(u) \in C^1[\alpha,\beta]$

ii) $P(\alpha), P(\beta)$ soient linéairement indépendants

iii) $P(u), P'(u)$ soient linéairement indépendants $\forall u \in [\alpha,\beta]$

Il est immédiat de voir que C_1 est un cône de sommet $\hat{x}$. On considère alors le domaine convexe :

$$D = \{g \in L^2_W[a,b] ; g(t) = \sum_{i=0}^{n} x_i \varphi_i(t) \ , \ x \in C_1\} \quad .$$

Etant donné $f \in L^2_W[a,b]$ ($\notin D$), on veut caractériser $\bar{g} \in D$ tel que :

$$\|\bar{g}-f\|_H = \min_{g\in D} \|g-f\|_H$$

On a le résultat suivant [33] .

THEOREME 1 :

Si $\overset{c}{g}$ désigne la meilleure approximation de f dans V alors

$\bar{g}$ réalise $\min_{g\in D} \|g-\overset{c}{g}\|_V$

3.1.1. Calcul de la meilleure approximation de f dans V .

Rechercher la meilleure approximation $\bar{g} \in D$ de f revient donc à minimiser, pour $x \in C_1$

$$\varphi(x) = \|g-f\|^2_H = \int_a^b W(t)\Big[\sum_{i=0}^{n} x_i\varphi_i(t)-f(t)\Big]^2 dt$$

ce qui s'écrit encore :

$$\varphi(x) = x^T Ax - 2b^T x + c$$

avec :

$$a_{ij} = <\varphi_i,\varphi_j> \quad , \quad b_i = <f,\varphi_i> \quad , \quad c = <f,f>$$

La matrice A, matrice de Gram formée avec les φ_i (i=0,...,n) linéairement indépendants, est symétrique, définie positive.

On obtient la meilleure approximation $\overset{o}{g}$ de f dans V (c'est-à-dire dans $\mathbb{R}^{n+1}$) en résolvant le système linéaire

$$Ax = b \ .$$

Soit $\overset{\circ}{x}$ la solution de ce système. D'après le théorème 1, le problème est ramené au calcul de la "distance" de $\overset{\circ}{x}$ à C_1. (Si $\overset{\circ}{x} \in C_1$ le problème est évidemment terminé, ce que l'on ne supposera pas).

3.1.2. Changement de variable.

D'après le théorème de Cholesky, on peut écrire $A = R^T R$ où R est une matrice triangulaire supérieure et

$$\varphi(x) = \|R(x-\overset{\circ}{x})\|^2 + c - \overset{\circ}{x}^T A \overset{\circ}{x} .$$

En posant

$$y = R(x-\overset{\circ}{x})$$

on doit minimiser :

$$\phi(y) = (y,y) \qquad y \in C$$

$$C = \{y \in \mathbb{R}^{n+1} \ / \ \forall u \in [\alpha,\beta] \ , \ (R^{-1}y,P(u)) + (\overset{\circ}{x}-\hat{x},P(u)) \geq 0\}$$

c'est-à-dire que désormais on cherche uniquement la distance euclidienne de l'origine (minimum de $\Phi(y)$ dans V) à C . C est un cone convexe de sommet $y = -R\tilde{x}$ en posant $\tilde{x} = \overset{\circ}{x}-\hat{x}$.

La projection de 0 sur un convexe est unique. Soit alors y^* le point de C qui minimise la distance de 0 à C.

3.1.3. Cône associé à C.

On pose afin de faciliter les notations :

$$h(u) = (R^{-1}y+\tilde{x},P(u)) = (y+\tilde{z},Q(u)) \quad \text{où} \quad \begin{cases} Q(u) = (R^{-1})^T P(u) \\ \tilde{z} = R\tilde{x} \end{cases}$$

Soit ∂C la frontière de C, (la "surface" de C)

$$\partial C = \{y \in \mathbb{R}^{n+1} \ / \ \exists u_o \in [\alpha,\beta] \text{ tel que } h(u_o) = 0, \text{ et}$$

$$\forall u \in [\alpha,\beta] \ h(u) \geq 0\}$$

On pose :

$$\delta = \{y \in R^{n+1} \text{ / toute racine (il en existe au moins une) de } h(u) = 0$$

appartenant à $[\alpha,\beta]$ est d'ordre pair, et si $h(u) \neq 0$, $h(u) > 0$ $\forall u \in [\alpha,\beta]\}$

$$\delta_\varepsilon = \{y \in R^{n+1} \text{ / } h(u) \geq 0 \ \forall u \in [\alpha,\beta] \text{ et } h(\varepsilon) = 0, \varepsilon \text{ racine d'ordre impair}\}$$

On a alors

$$\partial C = \delta \cup \delta_\alpha \cup \delta_\beta$$

La figure V .4 montre la structure géométrique du cône C lorsque n=1, et 2.

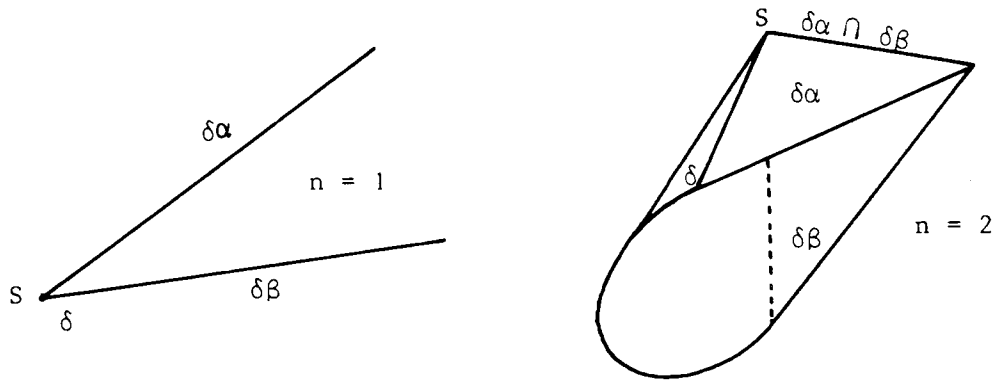


Figure V .4

On considère maintenant le cône C engendré par :

* les "génératrices" $\begin{cases} h(u) = (y+\tilde{z}, Q(u)) = 0 \\ h'(u) = (y+\tilde{z}, Q'(u)) = 0 \end{cases}$

lorsque u varie entre α et β

* les "génératrices" appartenant aux hyperplans :

$$h(\alpha) = (y+\tilde{z}, Q(\alpha)) = 0 \qquad (Q\alpha)$$

et

$$h(\beta) = (y+\tilde{z}, Q(\beta)) = 0 \qquad (Q\beta)$$

On pose :

$$\delta' = \{y \in \mathbb{R}^{n+1} \ / \ \exists u_o \in]\alpha,\beta[, \text{ racine d'ordre impair } (\geq 3) \text{ de } h(u)\}$$

$$\delta'_\varepsilon = \{y \in \mathbb{R}^{n+1} \ / \ h(\varepsilon) = 0 \ , \ \text{et} \ \ h(u) \not\geq 0 \ \ \text{sur} \ [\alpha,\beta]\}$$

On a alors :

$$C = \delta \cup \delta' \cup \delta_\alpha \cup \delta'_\alpha \cup \delta_\beta \cup \delta'_\beta = \partial C \cup (\delta' \cup \delta'_\alpha \cup \delta'_\beta)$$

THEOREME 2 :

Pour que y $(\in C)$ appartienne à C, il faut et il suffit que

$$(y+\tilde{z} \ , \ Q(u)) \geq 0 \qquad \forall u \in [\alpha,\beta]$$

Ce théorème est évident car $\partial C \cap (\delta \cup \delta'_\alpha \cup \delta'_\beta) = \emptyset$

On pose :

$$\bar{C} = \delta \cup \delta' \qquad \bar{C}_{\alpha\beta} = (\delta_\alpha \cap \delta_\beta) \cup (\delta'_\alpha \cap \delta'_\beta)$$

$$\bar{C}_\alpha = \delta_\alpha \cup \delta'_\alpha \qquad \bar{C}_\beta = \delta_\beta \cup \delta'_\beta$$

On a évidemment :

$$C = \bar{C} \cup \bar{C}_\alpha \cup \bar{C}_\beta \cup \bar{C}_{\alpha\beta}$$

3.1.4. Etude des normales issues de O aux diverses parties de C

3.1.4.1. Normale à la surface {S}

Puisque {S} est un point, cette normale est OS, et on pose

$$d_S = d^2(O,S) = \tilde{x}^T A \tilde{x} = \tilde{z}^T \tilde{z}$$

3.1.4.2. Normale à la surface $\{\bar{C}_\alpha\}$

Puisque $\bar{C}_\alpha$ est un hyperplan, on construit la normale à cet hyperplan :

$$(y,Q(\alpha)) = (-\tilde{z},Q(\alpha)) \qquad (Q\alpha)$$

le pied de la normale à $(Q\alpha)$ issue de O est :

$$y_\alpha = - \frac{(z,Q(\alpha))}{(Q(\alpha),Q(\alpha))} Q(\alpha)$$

D'après le théorème 2, si :

$$(R^{-1} y + \tilde{x}, P(u)) \geq 0$$

alors $y_\alpha \in \partial C$, et on pose :

$$d_\alpha = y_\alpha^T y_\alpha = (\tilde{z},Q(\alpha))^2 \; / \; (Q(\alpha),Q(\alpha))$$

sinon on pose :

$$d_\alpha = +\infty$$

3.1.4.3. Normale à la surface $\bar{C}_\beta$

On a de même en permuttant α et β :

$$d_\beta = \begin{cases} (\tilde{z},Q(\beta))^2 \; / \; (Q(\beta),Q(\beta)) & \text{si} \quad y_\beta \in \partial C \\ +\infty & \text{si} \quad y_\beta \notin \partial C \end{cases}$$

3.1.4.4. Normale à la surface $\bar{C}_{\alpha\beta}$

On recherche donc la projection de O sur $\bar{C}_{\alpha\beta}$ soit sur :

$$\begin{cases} (y,\ Q(\alpha)) = -(\tilde{z},Q(\alpha)) \\ (y,\ Q(\beta)) = -(\tilde{z},Q(\beta)) \end{cases}$$

On a le résultat suivant [3]

On veut minimiser (x,x) avec x satisfaisant à

$$(x,a_i) = b_i \qquad i=1,\ldots,k \qquad (k < n)$$

En supposant les vecteurs a_i linéairement indépendants, le minimum de (x,x) est alors obtenu au point

$$x^* = \sum_{j=1}^{k} c_j a_j$$

avec les coefficients c_j solution du système linéaire :

$$\sum_{j=1}^{k} c_j(a_j,a_i) = b_i \qquad i=1,\ldots,k$$

Utilisant ce résultat, d'après (3), on a donc ici :

$$y_{\alpha\beta} = C_1 Q(\alpha) + C_2 Q(\beta)$$

avec :

C_1 et C_2 solution de

$$(Q(\alpha),Q(\alpha))C_1 + (Q(\beta),Q(\alpha))C_2 = -(\tilde{z},Q(\alpha))$$
$$(Q(\alpha),Q(\beta))C_1 + (Q(\beta),Q(\beta))C_2 = -(\tilde{z},Q(\beta))$$

soit :

$$C_1 = \frac{(Q(\beta),Q(\alpha))(\tilde{z},Q(\beta)) - (Q(\beta),Q(\beta))(\tilde{z},Q(\alpha))}{(Q(\alpha),Q(\alpha))(Q(\beta),Q(\beta)) - (Q(\beta),Q(\alpha))^2}$$

$$C_2 = \frac{(Q(\alpha),Q(\beta))(\tilde{z},Q(\alpha)) - (Q(\alpha),Q(\alpha))(z,Q(\beta))}{(Q(\alpha),Q(\alpha))(Q(\beta),Q(\beta)) - (Q(\beta),Q(\alpha))^2}$$

On pose alors :

$$d_{\alpha\beta} = \begin{cases} y_{\alpha\beta}^T y_{\alpha\beta} = -C_1(\tilde{z},Q(\alpha)) - C_2(\tilde{z},Q(\beta)) = -(\tilde{z},y_{\alpha\beta}) & \text{si } y_{\alpha\beta} \in \partial C \\ +\infty & \text{si } y_{\alpha\beta} \notin \partial C \end{cases}$$

3.1.4.5. Normale à la surface $\bar{C}$

Soit une "génératrice" fixée de $\bar{C}$, ($u_o \in [\alpha,\beta]$, fixé)

$$\begin{cases} (y+\tilde{z},Q(u_o)) = 0 \\ (y+\tilde{z},Q'(u_o)) = 0 \end{cases} \qquad (5)$$

La projection orthogonale de l'origine sur (5) est obtenue comme précédemment

$$\bar{y} = d_1 Q(u_o) + d_2 Q'(u_o)$$

avec d_1 et d_2 solution du système linéaire :

$$\begin{cases} (Q(u_o),Q(u_o))d_1 + (Q'(u_o),Q(u_o))d_2 = -(\tilde{z},Q(u_o)) \\ (Q'(u_o),Q(u_o))d_1 + (Q'(u_o),Q'(u_o))d_2 = -(\tilde{z},Q'(u_o)) \end{cases} \qquad (6)$$

Pour que $\bar{y}$ ($\in \bar{C}$) soit le pied d'une normale passant par l'origine 0, il faut et il suffit que $\bar{y}$ soit la projection orthogonale de 0 sur le plan tangent à $\bar{C}$.

Ce plan tangent ayant pour équation :

$$(y,Q(u_o)) = -(\tilde{z},Q(u_o))$$

$\bar{y}$ est la projection orthogonale de 0 sur $\bar{C}$ si et seulement si :

$$\exists u_o \text{ tel que } \bar{y} = -\frac{(\tilde{z},Q(u_o))}{(Q(u_o),Q(u_o)} Q(u_o) \qquad (7)$$

En résolvant le système linéaire (6), on obtient :

$$\bar{y} = -\frac{[(\tilde{z},Q)(Q',Q')-(\tilde{z},Q')(Q',Q)]Q + [(\tilde{z},Q')(Q,Q)-(\tilde{z},Q)(Q',Q)]Q'}{(Q,Q)(Q',Q') - (Q,Q')^2}$$

En portant dans (7), on obtient immédiatement puisque $Q(u_o)$ et $Q'(u_o)$ sont linéairement indépendants :

$$(\tilde{z},Q')(Q,Q) - (\tilde{z},Q)(Q',Q) = 0 \qquad \text{(coefficients de } Q' = 0)$$

et

$$\frac{(Q',Q)\left[(\tilde{z},Q')(Q,Q) - (\tilde{z},Q)(Q',Q)\right]}{(Q,Q)\left[(Q,Q)(Q',Q') - (Q,Q')^2\right]} = 0 \qquad \text{(coefficients de } Q = 0\text{)}$$

Pour que $\bar{y}$ soit le pied de la normale à $\bar{C}$ passant par 0, u_o doit satisfaire :

$$(\tilde{z},Q'(u_o))(Q(u_o),Q(u_o)) - (\tilde{z},Q(u_o))(Q'(u_o),Q(u_o)) = 0$$

Soient p le nombre de racines de cette équation en u_o, qui sont $v_1, v_2, \ldots, v_p$, appartenant à l'intervalle $]\alpha,\beta[$.

Si $u_o = v_k$ on pose $y_k = \bar{y}$ et on a alors :

$$d^2(0,y_k) = d_k = (\tilde{z},Q(v_k))^2 \;/\; (Q(v_k),Q(v_k))$$

Soit $\quad d(u) = (\tilde{z},Q(u))^2 \;/\; (Q(u),Q(u))$

On a $\quad d'(u) = \dfrac{2(\tilde{z},Q(u))\left[(\tilde{z},Q'(u))(Q(u),Q(u)) - (\tilde{z},Q(u))(Q(u),Q'(u)\right]}{\left[(Q(u),Q(u))\right]^2}$

d(u) possède des extrema pour

$$v(u) = (\tilde{z},Q(u)) = 0$$

$$t(u) = (\tilde{z},Q'(u))(Q(u),Q(u)) - (\tilde{z},Q(u))(Q(u),Q'(u)) = 0$$

Dire que $v(\bar{u}) = 0$, signifie qu'il existe un plan tangent passant par l'origine, ce qui est sans intérêt à moins que de plus $t(\bar{u}) = 0$.

Dans ce cas, $(v(\bar{u}) = t(\bar{u}) = 0)$, et $y = y(\bar{u})$ appartient à $\bar{C}$

- si $y \notin \delta$, cette normale est ignorée car son pied $\notin \partial C$
- si $y \in \delta$, ceci voudrait dire que $0 \in \partial C$ car $d(0,y) = 0$

ce qui est contraire à notre hypothèse de départ.

Différents cas sont à envisager suivant le nombre p de racines sur $[\alpha,\beta]$ de l'équation $t(u) = 0$

a) $p = 0$

Il n'existe aucune normale de $\bar{C}$ passant par O, ceci veut dire donc, qu'il n'existe aucune normale à δ en particulier. C'est-à-dire se reportant au cône C , la meilleure approximation cherchée appartient ou à δ_α ou à δ_β ou à $\delta_\alpha \cap \delta_\beta$ ou à $\{S\}$.

b) $p \neq 0$

On sait que si la meilleure approximation n'appartient pas à δ_α ou ou δ_β ou $\delta_\alpha \cap \delta_\beta$ ou $\{S\}$, elle appartient nécessairement à δ, et dans ce cas :

* il existe une normale (au moins) à δ
* la distance de O au plan tangent contenant la génératrice passant par le pied de cette normale présente un maximum (figure V .5) [51]
  Ceci veut donc dire que, au point $y_j = y(v_j)$, si $d(v_j)$ n'est pas maximum le point $y(v_j)$ n'est pas à conserver car il ne donnera pas la meilleure approximation sur C (figure V .6).

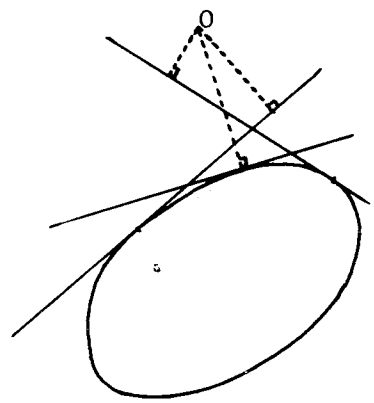
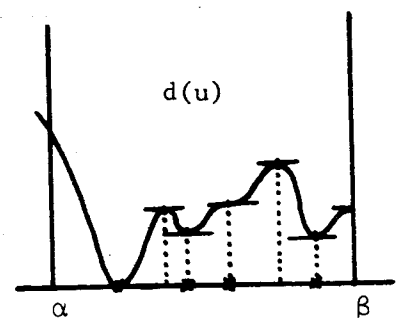


Figure V .5

REMARQUE :

Si $d'(\alpha) = 0$ (et/ou $d'(\beta) = 0$) Il n'est pas utile d'étudier les projections sur δ_α (et/ou δ_β), car on aurait les mêmes calculs à refaire qu'ici.



(* points a supprimer)

Figure V .6

Donc les seuls cas à envisager sont ceux où $d(u) \neq 0$.

Soient alors les abscisses

$$\xi_1, \xi_2, \ldots, \xi_m \qquad m \leq k \quad (\xi_i \in [\alpha, \beta])$$

où $d(u)$ est maximum, avec $y_i = y(\xi_i)$.

$$y_i = - \frac{(\tilde{z}, Q(\xi_i))Q(\xi_i)}{(Q(\xi_i), Q(\xi_i))} \qquad i=1,2,\ldots,m$$

On pose :

$$d_i = \begin{cases} d^2(0, y_i) = y_i^T y_i & \text{si} \quad y_i \in \partial C \\ +\infty & \text{si} \quad y_i \notin \partial C \end{cases}$$

$i=1,2,\ldots,m$.

THEOREME 3 :

Le point le plus proche de l'origine sur C est le point

$$y^* = y_\eta \qquad \eta \text{ étant tel que}$$

$$d_\eta = min(d_S, d_\alpha, d_\beta, d_{\alpha\beta}, d_1, d_2, \ldots, d_m)$$

3.2. APPLICATION : CALCUL D'UN FILTRE NON RECURSIF

Les filtres non récursifs sont des cas particuliers des filtres récursifs, c'est-à-dire on a ici $n=0$. On se propose donc de minimiser :

$$\int_0^\pi w(u) \Big[\sum_{i=0}^{m} x_i \cos iu - f(u) \Big]^2 du$$

avec :

$$\sum_0^m x_i \cos iu \geq 0 \qquad \forall u \in [0, \pi]$$

Il est immédiat de voir que la méthode précédente est directement utilisable en choisissant :

$$\varphi_p(u) = \cos pu \qquad p=0,\ldots,m$$

et

$$P^T(u) = [\varphi_o(u),\ldots,\varphi_m(u)]$$

EXEMPLE :

$$m = 4$$

$$f(x) = \begin{cases} 0 & x \in [0, \frac{\pi}{4}] \\ \frac{4}{\pi}(x - \frac{\pi}{2})+1 & x \in [\frac{\pi}{4},\frac{\pi}{2}] \end{cases}$$

f(x) est complètée par symétrie par rapport à la droite $\pi/2$.
De plus $w(x) = 1 \quad \forall x$.

Le meilleur approximant est, après calculs : figure V .7

- sans contrainte :

$$y_1(x) = \frac{1}{\pi}(\frac{\pi}{4} - \frac{4}{\pi}\cos 2x)$$

- avec contraintes :

$$y_2(x) = 0.39119 - 0.37291 \cos 2x - 0.01828 \cos 4x$$

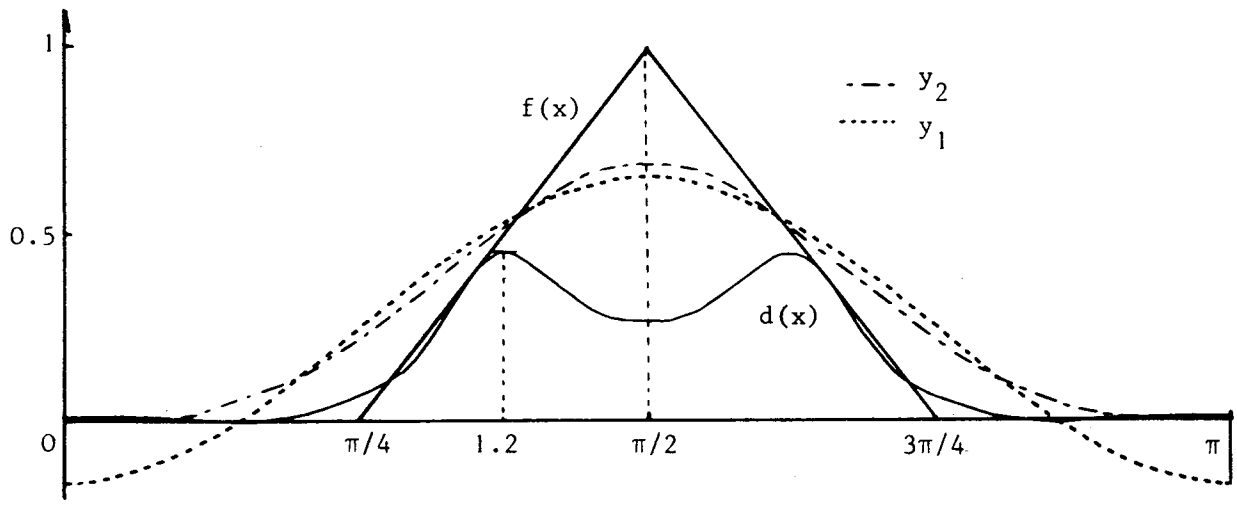


Figure V .7

On a tracé de plus d(x) sur la figure, ce qui permet de localiser rapidement les maxima en 0 et 1.2. La projection de 0 sur (P_o) appartient à ∂C, c'est donc le point de C le plus proche de 0.
En effet, l'autre projection serait telle que $d(1.2) > d(0)$.

3.3. CALCUL D'UN FILTRE RECURSIF

On rappelle que l'on veut minimiser :

$$Q = \int_0^{\pi} w(x) \left[\frac{\sum_0^m b_i \cos^i x}{1+\sum_1^n a_i \cos^i x} - f(x) \right]^2 dx = \int_{-1}^{+1} \eta(u) \left[\frac{\sum_0^m b_i u^i}{1+\sum_1^n a_i u^i} - g(u) \right]^2 du$$

On pourra utiliser la méthode précédente dans le cas où n est faible. Par exemple si n=2, pour différentes valeurs de a_1, a_2, on minimise Q, et on trace les lignes de niveau (figure V .8) ce qui permet de localiser le point (a^*, b^*) tel que :

$$Q(a^*, b^*) = \min_a (\min_b Q(a,b))$$

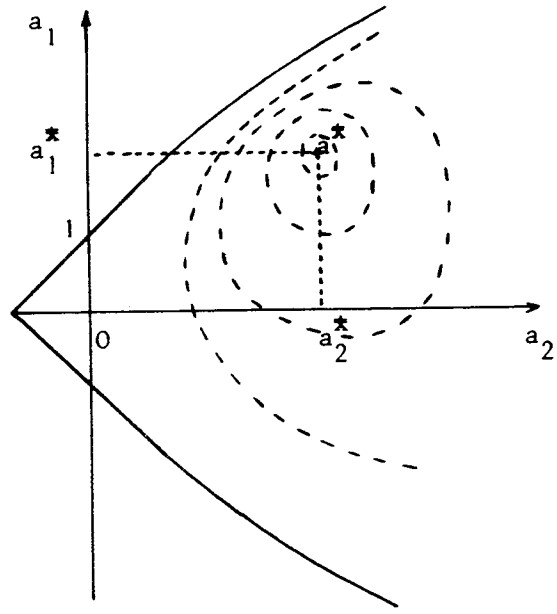


Figure V .8

V .4. FILTRES PASSE-BANDE OPTIMAUX PONDERES

4.1. RAPPELS ET DEFINITIONS

Construire un filtre numérique consiste à trouver deux jeux de coefficients réels $(\beta_o,\ldots,\beta_m)$ et $(\alpha_o,\ldots,\alpha_n)$ tels que la réponse en fréquence (à vrai dire en pulsation) du filtre :

$$H(e^{i\omega\Delta T}) = \frac{\sum_{0}^{m} \beta_p \, e^{-ip\omega\Delta T}}{\sum_{0}^{n} \alpha_p \, e^{-ip\omega\Delta T}}$$

satisfasse un certain nombre de conditions fixées à l'avance.
Le filtre tel que :

$$\text{i)} \quad \int_0^{\pi} w(x) \left| \left| \frac{\sum_{0}^{m} \beta_p \, e^{-ipx}}{\sum_{0}^{n} \alpha_p \, e^{-ipx}} \right|^2 - f(x) \right|^2 dx \quad \text{est minimum}$$

($f \in L^2$ étant un modèle donné à l'avance)

ii) le polynome $\sum_{0}^{n} \alpha_k z^{n-k}$ possède toutes ses racines de module inférieur à 1 (condition de stabilité)

sera dit filtre optimal au sens de la norme $L^2_w[0,\pi]$.

<u>degré</u> Le filtre construit sera dit de degré n si et seulement si n = m .

4.2. FILTRES PASSE-BANDE. FONCTION POIDS

Un filtre sera dit passe-bande au sens large, si son modèle est tel que :

$$f(x) = \begin{cases} 1 & x \in [x_1, x_2] \\ 0 & x \notin [x_1, x_2] \end{cases}$$

$x_1 < x_2$

- si $x_1 = 0$ il est dit passe-bas
- si $x_2 = \pi$ il est dit passe-haut

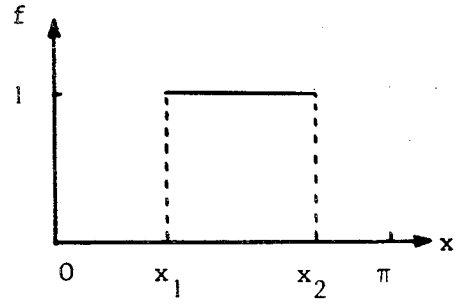


Figure V .9

On se propose de montrer quelques propriétés remarquables de ces filtres passe-bande lorsque la fonction poids est choisie telle que :

$$w(x) = \frac{1}{\operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2}} (1 + \operatorname{tg}^2 \frac{x}{2})(1 + \frac{\operatorname{tg}\frac{x_1}{2} \operatorname{tg}\frac{x_2}{2}}{\operatorname{tg}^2 \frac{x}{2}})$$

Dans toute la suite seule cette fonction poids, désormais notée $\Psi(x)$ sera utilisée.

4.3. FILTRES PASSE-BAS OPTIMAUX

Soit $F_{\omega 1}$ le filtre optimal, c'est-à-dire les vecteurs α' et β', dont le modèle est :

$$f_1(x) = \begin{cases} 1 & x \in [0, \omega_1] \\ 0 & x \in]\omega_1, \pi] \end{cases}$$

Dans ce cas la fonction poids est égale à

$$\Psi(x) = \frac{1}{\operatorname{tg}\frac{\omega_1}{2}} (1 + \operatorname{tg}^2 \frac{x}{2}) .$$

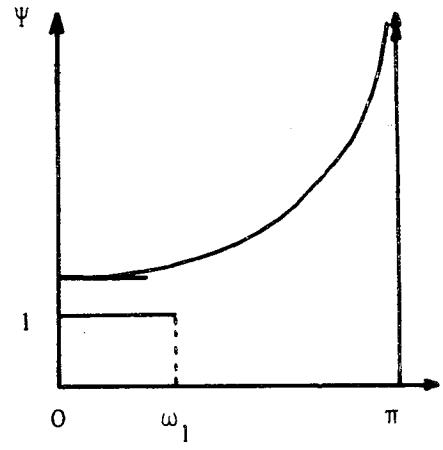


Figure V .10

THEOREME 4 [71]

Soient $\beta^{1T} = [\beta_o^1, \ldots, \beta_n^1]$ *et* $\alpha^{1T} = [\alpha_o^1, \ldots, \alpha_n^1]$ *les coefficients du filtre optimal passe-bas* F_{ω_1} *de degré n par rapport à la fonction poids*

$$\Psi(x) = \frac{1}{tg \frac{\omega_1}{2}} (1 + tg^2 \frac{x}{2})$$

Et soient β^{2T} *et* α^{2T} *les coefficients du filtre passe-bas optimal* F_{ω_2} *de degré n par rapport à la fonction poids*

$$\Psi(x) = \frac{1}{tg \frac{\omega_2}{2}} (1 + tg^2 \frac{x}{2})$$

Alors il existe une matrice M à (n+1) lignes et (n+1) colonnes (aisément calculable) telle que :

$$\beta^2 = M \beta^1 \quad et \quad \alpha^2 = M \alpha^1$$

Le filtre optimal F_{ω_2} est le filtre tel que I soit minimum :

$$I = \frac{1}{tg \frac{\omega_2}{2}} \int_0^{\pi} (1+tg^2\frac{x}{2}) \left| \frac{\sum_0^n \beta_p^2 e^{-ipx}}{\sum_0^n \alpha_p^2 e^{-ipx}} \cdot \frac{\sum_0^n \beta_p^2 e^{ipx}}{\sum_0^n \alpha_p^2 e^{ipx}} - f_2(x) \right|^2 dx$$

Dans I effectuons le changement de variable défini par

$$tg \frac{x}{2} = \frac{tg\ \omega_2/2}{tg\ \omega_1/2}\ tg \frac{\theta}{2}$$

ce qui entraine :

$$(T) \qquad e^{ix} = \frac{(1-\eta)+(1+\eta)e^{i\theta}}{(1+\eta)+(1-\eta)e^{i\theta}} \qquad (\eta = tg\ \omega_2/2 \ / \ tg\ \omega_1/2)$$

et

$$f_2(x) = g(\theta) = \begin{cases} 1 & x \in [0,\omega_1] \\ 0 & x \notin [0,\omega_1] \end{cases}$$

C'est-à-dire $g(\theta) \equiv f_1(\theta)$ le modèle de F_{ω_1} .

On a donc :

$$I = \frac{1}{\operatorname{tg}\frac{\omega_1}{2}} \int_0^{\pi} (1+\operatorname{tg}\ \frac{\theta}{2}) \left[\frac{\sum_0^n \beta_p^2 \left(\frac{(1-\eta)+(1+\eta)e^{-i\theta}}{(1+\eta)+(1-\eta)e^{-i\theta}}\right)^p \sum_0^n \beta_p^2 \left(\frac{(1-\eta)+(1+\eta)e^{i\theta}}{(1+\eta)+(1-\eta)e^{i\theta}}\right)^p}{\sum_0^n \alpha_p^2 \left(\frac{(1-\eta)+(1+\eta)e^{-i\theta}}{(1+\eta)+(1-\eta)e^{-i\theta}}\right)^p \sum_0^n \alpha_p^2 \left(\frac{(1-\eta)+(1+\eta)e^{i\theta}}{(1+\eta)+(1-\eta)e^{i\theta}}\right)^p} - f_1(\theta) \right]^2 d$$

ce qui s'écrit encore (degré du numérateur égal degré du dénominateur) :

$$I = \frac{1}{\operatorname{tg}\frac{\omega_1}{2}} \int_0^{\pi} (1 + \operatorname{tg}^2 \frac{\theta}{2}) \left[\frac{\sum_0^n \beta_p e^{-ip\theta} \sum_0^n \beta_p e^{ip\theta}}{\sum_0^n \alpha_p e^{-ip\theta} \sum_0^n \alpha_p e^{ip\theta}} - f_2(\theta) \right]^2 d\theta \qquad (8)$$

qui est donc minimum, lorsque $\beta_p = \beta_p^1$, $\alpha_p = \alpha_p^1$ $p=0,\ldots,n$, soit en notant par ξ_p^i : α_p^i ou β_p^i $(i=1,2)$

On a :

$$\sum_0^n \xi_p^2 [(1-\eta)+(1+\eta)z]^p [(1+\eta)+(1-\eta)z]^{n-p} \equiv \sum_0^n \xi_p^1 z^p$$

En refaisant dans (8) le changement de variable inverse :

$$\operatorname{tg}\frac{\theta}{2} = \nu \operatorname{tg}\frac{x}{2} \qquad \nu = \operatorname{tg}(\omega_1/2) \ / \ \operatorname{tg}(\omega_2/2)$$

on obtient :

$$\sum_0^n \xi_p^1 [(1-\nu)+(1+\nu)z]^p [(1+\nu)+(1-\nu)z]^{n-p} \equiv \lambda \sum_0^n \xi_p^2 z^p$$

(λ étant pris tel que, par exemple, $\alpha_o^1 = 1$).

Ce qui montre en développant le premier membre le théorème 4, en identifiant les termes en z^p.

EXEMPLE :

$n = 2$

$$M = \lambda \begin{bmatrix} p^2 & pd & d^2 \\ 2pd & p^2+d^2 & 2pd \\ d^2 & pd & p^2 \end{bmatrix} \quad \text{où} \quad \begin{matrix} p=1+\nu \\ d=1-\nu \end{matrix}$$

REMARQUES :

1) Pour que $\left|\frac{\Sigma\beta_p e^{-ipx}}{\Sigma\alpha_p e^{-ipx}}\right|^2 \in L^2_\Psi[0,\pi]$, il faut et il suffit que :

$$\Sigma\beta_p e^{-ipx} = (1+\cos x)\ (\text{polynome de degré } (n-1) \text{ en } \cos x)$$

2) *PROPRIETE :*

Le filtre F_{ω_1} étant supposé stable, F_{ω_2} est aussi un filtre stable.

En effet, l'intérieur du cercle unité est conservé par la transformation (T) [61]

De plus z_k^1 (k=1,...,n) étant les poles de F_{ω_1}, ceux de F_{ω_2} sont alors :

$$z_k^2 = \frac{(\eta-1)+(\eta+1)z_k^1}{(\eta+1)+(\eta-1)z_k^1} \qquad (k=1,\ldots,n)$$

4.4. FILTRES PASSE-HAUT OPTIMAUX

Un filtre optimal H_1 passe-haut par rapport à la fonction poids

$$\Psi(x) = \frac{1}{\cotg \frac{\omega_1}{2}} (1 + \cotg \frac{x}{2})$$

est le filtre qui a pour modèle :

$$f(x) \begin{cases} 1 & x \in [\omega_1,\pi] \\ 0 & x \notin [\omega_1,\pi] \end{cases}$$

Pour obtenir H_{ω_1}, il suffit de considérer le filtre optimal passe-bas $F_{\pi-\omega_1}$ qui a pour coefficients $(\beta_o,\ldots,\beta_n)$, $(\alpha_o,\ldots,\alpha_n)$ et de faire une symétrie par rapport à la droite $\pi/2$.

THEOREME 5 :

Soient $(\beta_o,\ldots,\beta_n)$ *et* $(\alpha_o,\ldots,\alpha_n)$ *les coefficients du filtre passe-bas optimal* $F_{\pi-\omega_1}$ *. Les coefficients du filtre optimal* H_{ω_1} *passe-haut sont alors :*

$$(\beta_o,-\beta_1,\ldots,(-1)^n\beta_n)$$
$$(\alpha_o,-\alpha_1,\ldots,(-1)^n\alpha_n)$$

De plus ce filtre est stable, car les poles (et les zéros) de H_{ω_1} *sont symétriques de ceux de* $F_{\pi-\omega_1}$ *par rapport à l'origine.*

4.5. CALCUL PRATIQUE DES FILTRES PASSE-BAS ET PASSE-HAUT

Avant d'indiquer comment calculer de tels filtres on définit le filtre de base de degré n :

Le filtre de base de degré n est le filtre de degré n $F\pi/2$ optimal par rapport à la fonction poids :

$$\Psi(x) = (1 + tg^2 \frac{x}{2}) .$$

On dresse ainsi une table des coefficients (et des pôles de ces filtres de base pour n=1,2,3,... , (cf. page 139). Un filtre passe-bas optimal quelconque de degré k est alors calculé immédiatement par les formules du théorème 4, et en se servant du filtre de base de degré k donné par la table. Tenant compte du théorème 5, un filtre passe -haut peut être ainsi construit.

La table des filtres de base a été construite à partir des méthodes numériques exposées dans les paragraphes précédents. Il faut souligner l'importance de cette table qui permettra de construire, aussi, rapidement des filtres passe-bande.

LE FILTRE PASSE-BAS (PI/2) (DEGRE DU NUMERATEUR ET DENOMINATEUR = 1) A POUR POLES :

PARTIE REELLE	PARTIE IMAGINAIRE
0.2733006E 00	0.0

LES COEFFICIENTS DU FILTRE SONT :

ALPHA

0 0.1000000E 01
1-0.2733005E 00

BETA

0 0.3999347E 00
1 0.3999347E 00

LE FILTRE PASSE-BAS (PI/2) (DEGRE DU NUMERATEUR ET DENOMINATEUR = 2) A POUR POLES :

PARTIE REELLE	PARTIE IMAGNINAIRE
0.2086750E 00	-0.5558982E 00
0.2086750E 00	0.5558982E 00

LES COEFFICIENTS DU FILTRE SONT :

ALPHA

0 0.1000000E 01
1-0.4173499E 00
2 0.3525679E 00

BETA

0 0.2161070E 00
1 0.4350651E 00
2 0.2189581E 00

CE FILTRE PASSE-BAS (PI/2) (DEGRE DU NUMERATEUR ET DENOMINATEUR = 3) A POUR POLES :

PARTIE REELLE	PARTIE IMAGINAIRE
-.9830761E-01	0.0
0.6348693E-01	-0.7312412E 00
0.6348693E-01	0.7312412E 00

LES COEFFICIENTS DU FILTRE SONT :

ALPHA

0 0.1000000E 01
1-0.2866625E-01
2 0.5262613E 00
3 0.5296263E-01

BETA

0 0.2698397E 00
1 0.4818042E 00
2 0.4785778E 00
3 0.2664132E 00

4.6. FILTRES PASSE-BANDE QUASI-OPTIMAUX

4.6.1. Filtre passe-bande symétrique

On considère le filtre passe-bande dont le modèle est

$$f_{x_1,\pi-x_1}(x) = \begin{cases} 1 & x \in [x_1,\pi-x_1] \\ 0 & \text{sinon} \end{cases}$$

On peut remarquer alors que :

$$\Psi(x) = \frac{1}{\operatorname{tg}\frac{\pi-x_1}{2} - \operatorname{tg}\frac{x_1}{1}}\left(1+\operatorname{tg}^2\frac{x}{2}\right)\left(1+\frac{\operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{\pi-x_1}{2}}{\operatorname{tg}^2\frac{x}{2}}\right) = \Psi(\pi-x)$$

On pose $T_s^j = \left\{r = p/q \; ; \; r(x) = \left|\frac{\sum_{0}^{j} \beta_\ell e^{-i\ell x}}{\sum_{0}^{s} \alpha_\ell e^{-i\ell x}}\right|^2 \quad , q > 0 \quad \forall x \in [0,\pi]\right\}$

Rechercher le filtre optimal de degré k, dont le modèle est f_1, revient à minimiser :

$$\|r-f_1\|^2 \qquad \text{pour} \qquad r \in T_k^k$$

On pose :

$$\tilde{T}_p^j = T_p^j \cap \{r;r(x) = r(\pi-x) \quad \forall x \in [0,\pi]\}$$

On dira que le filtre passe-bande $F_{x_1,\pi-x_1}$ de degré k est quasi-optimal, si sa fonction d'approximation $r^*(x)$ est telle que :

$$\|r^*-f_{x_1,\pi-x_1}\| \leq \|r - f_{x_1,\pi-x_1}\| \qquad \forall r \in \tilde{T}_k^k$$

(la fonction poids étant $\Psi(x)$).

4.6.1. Filtre passe-bande symétrique quasi-optimal

Soit $F_{\omega,\pi-\omega}$ le filtre quasi-optimal passe-bande symétrique de degré 2n de fréquences de coupure ω et $\pi-\omega$, il est tel que

$$J = \frac{1}{\operatorname{cotg}\frac{\omega}{2} - \operatorname{tg}\frac{\omega}{2}} \int_0^{\pi} (1 + \operatorname{tg}^2\frac{x}{2})(1 + \operatorname{cotg}^2\frac{x}{2}) \left| \left| \frac{\sum_0^{2n} \beta_\ell e^{-i\ell x}}{\sum_0^{2n} \alpha_\ell e^{-i\ell x}} \right|^2 - f_{\omega,\pi-\omega}(x) \right|^2 dx$$

$$J_1 = \frac{J}{2} = 2 \operatorname{tg}\omega \int_{\pi/2}^{\pi} \frac{1}{\sin^2 x} \left| \left| \frac{\sum_0^{2n} \beta_\ell e^{-i\ell x}}{\sum_0^{2n} \alpha_\ell e^{-i\ell x}} \right|^2 - f_{\omega,\pi-\omega}(x) \right|^2 dx$$

soit minimum.

Faisons dans J_1 le changement de variable défini par :

$$\operatorname{tg}\omega \operatorname{cotg} x = - \operatorname{tg}\frac{u}{2}$$

soit :

$$-2\operatorname{tg}\omega \frac{dx}{\sin^2 x} = - (1 + \operatorname{tg}^2\frac{x}{2})du$$

On remarque alors que :

a) $$f_{\omega,\pi-\omega}(x) = g(u) = f_{\pi/2}(u) = \begin{cases} 1 & u \in [0,\pi/2] \\ 0 & u \notin [0,\pi/2] \end{cases}$$

b) $$r = \left| \frac{\sum_0^{2n} \beta_\ell e^{-i\ell x}}{\sum_0^{2n} \alpha_\ell e^{-i\ell x}} \right|^2 = \frac{\sum_0^{n} b_{2i} \cos^{2i} x}{\sum_0^{n} a_{2i} \cos^{2i} x} \quad \text{puisque} \quad r \in \tilde{T}_{2n}^{2n}$$

Le changement de variable (9) entraine :

$$\cos^2 x = \frac{1 - \cos u}{(1 - \cos u) + \mathrm{tg}^2\,\omega(1 + \cos u)}$$

donc :

$$r(x) = \tilde{r}(u) = \frac{\sum_{0}^{n} \tilde{b}_i \cos^i u}{\sum_{0}^{n} \tilde{a}_i \cos^i u} = \left| \frac{\sum_{0}^{n} \tilde{\beta}_\ell\, e^{-i\ell u}}{\sum_{0}^{n} \tilde{\alpha}_\ell\, e^{-i\,u}} \right|^2$$

On a donc :

$$J_1 = \int_0^{\pi} (1 + \mathrm{tg}^2 \frac{u}{2}) \left[\left| \frac{\sum_{0}^{n} \tilde{\beta}_\ell e^{-i\ell u}}{\sum_{0}^{n} \tilde{\alpha}_\ell e^{-i\ell u}} \right|^2 - f_{\pi/2}(u) \right]^2 du$$

qui est minimum lorsque les coefficients $(\tilde{\alpha},\tilde{\beta})$ sont les coefficients (α^*,β^*) du filtre optimal $F_{\pi/2}$ de base de degré n.

4.6.2. Filtre passe-bande quasi-optimal.

Soit le filtre passe-bande dont le modèle est

$$f_{x_1,x_2}(x) = \begin{cases} 1 & x \in [x_1,x_2] \\ 0 & x \notin [x_1,x_2] \end{cases} \qquad (x_1 < x_2)$$

F_{x_1,x_2} est alors optimal de degré 2n si :

$$K = \frac{1}{\mathrm{tg}\frac{x_2}{2} - \mathrm{tg}\frac{x_1}{2}} \int_0^{\pi} (1 + \mathrm{tg}^2 \frac{\theta}{2})(1 + \frac{\mathrm{tg}\frac{x_1}{2}\,\mathrm{tg}\frac{x_2}{2}}{\mathrm{tg}^2\frac{\theta}{2}}) \left[\left| \frac{\sum_{0}^{2n} \bar{\beta}_\ell\, e^{-i\ell\theta}}{\sum_{0}^{2n} \bar{\alpha}_\ell\, e^{-i\ell\theta}} \right|^2 - f_{x_1,x_2}(\theta) \right]^2$$

est minimum.

Dans K faisons le changement de variable

$$\operatorname{tg}\frac{\theta}{2} = \sqrt{\operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}}\ \operatorname{tg}\frac{x}{2}$$

$$K = \frac{\sqrt{\operatorname{tg}\frac{x_1}{2}\ \operatorname{tg}\frac{x_2}{2}}}{\operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2}} \int_0^{\pi} (1 + \operatorname{tg}^2\frac{x}{2})(1 + \frac{1}{\operatorname{tg}^2\frac{x}{2}}) \left[\left| \frac{\sum_0^{2n} \beta'_\ell\, e^{-i\ell x}}{\sum_0^{2n} \alpha'_\ell\, e^{-i\ell x}} \right|^2 - f_{\omega,\pi-\omega}(x) \right]^2 dx$$

où ω est tel que $\operatorname{tg}\omega = 2\sqrt{\operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}} \ /\ (\operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2})$

Donc :

$$K = \frac{1}{\operatorname{cotg}\frac{\omega}{2} - \operatorname{tg}\frac{\omega}{2}} \int_0^{\pi} (1 + \operatorname{tg}^2\frac{x}{2})(1 + \frac{1}{\operatorname{tg}^2\frac{x}{2}}) \left[\left| \frac{\sum_0^{2n} \beta'_\ell\, e^{-i\ell x}}{\sum_0^{2n} \alpha'_\ell\, e^{-i\ell x}} \right|^2 - f_{\omega,\pi-\omega}(x) \right]^2 dx$$

F_{x_1,x_2} sera dit filtre passe-bande quasi-optimal si les coefficients (α',β') sont ceux du filtre quasi-optimal $F_{\omega,\pi-\omega}$.

4.6.3. Coefficients du filtre passe-bande quasi-optimal.

THEOREME 6 :

Soient (β^*,α^*) les vecteurs des coefficients du filtre optimal passe-bas $F_{\pi/2}$ de degré n par rapport à la fonction poids

$$\Psi(x) = (1 + \operatorname{tg}^2\frac{x}{2})$$

et soient (β,α) ceux du filtre quasi-optimal passe-bande F_{x_1,x_2} de degré 2n par rapport à la fonction poids

$$\Psi(x) = \frac{1}{\operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2}} (1 + \operatorname{tg}^2\frac{x}{2})(1 + \frac{\operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}}{\operatorname{tg}^2\frac{x}{2}})$$

Alors il existe une matrice N à (2n+1) lignes et (n+1) colonnes telle que :

$$\beta = N\beta^* \qquad et \qquad \alpha = N\alpha^*$$

On est passé de F_{x_1,x_2} à $F_{\pi/2}$ suivant le schéma suivant :

$$F_{x_1,x_2} \longrightarrow F_{\omega,\pi-\omega} \longrightarrow F_{\pi/2}$$

$$\operatorname{tg}\frac{\theta}{2} = \sqrt{\operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}}\,\operatorname{tg}\frac{x}{2} \qquad \operatorname{tg}\omega\,\operatorname{cotg}\,x = -\operatorname{tg}\frac{u}{2}$$

$$\operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2}\;\;\frac{\operatorname{tg}^2\frac{\theta}{2} - \operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}}{\operatorname{tg}\frac{\theta}{2}} = \operatorname{tg}\frac{u}{2}$$

Changement de variable qui est tel que :

$$e^{iu} = \frac{d(1-e^{2i\theta}) - [(p+1) + 2(p-1)e^{i\theta} + (p+1)e^{2i\theta}]}{d(1-e^{2i\theta}) + [(p+1) + 2(p-1)e^{i\theta} + (p+1)e^{2i\theta}]}$$

en ayant posé :

$$d = \operatorname{tg}\frac{x_2}{2} - \operatorname{tg}\frac{x_1}{2}$$

$$p = \operatorname{tg}\frac{x_1}{2}\operatorname{tg}\frac{x_2}{2}$$

Comme dans la démonstration du théorème 4 on a en remplaçant e^{iu} par sa valeur en fonction de $e^{i\theta}$

$$\sum_{0}^{n} \xi_j^{*}[(d-1-p)-2(p-1)z-(d+1+p)z^2]^j[(d+1+p)+2(p-1)z-(d-1-p)z^2]^{m-j} \equiv \lambda \sum_{0}^{2n} \xi$$

en notant par ξ_j^{*} les coefficients (α_j^{*} ou β_j^{*}) de $F_{\pi/2}$ et par ξ_j (resp. α_j ou β_j) ceux du filtre passe-bande F_{x_1,x_2} (λ pouvant être choisi, par exemple, tel que $\alpha_o^{*} = 1$).

L'identité précédente montre l'existence de la matrice N et aussi un moyen de la calculer.

REMARQUES :

1) Pour pouvoir parler de l'existence du filtre F_{x_1,x_2} il est nécessaire d'avoir $n \geq 1$, et pour assurer la convergence de l'intégrale K il est nécessaire et suffisant que

$$\left|\sum_{0}^{2n} \beta_j e^{-jix}\right|^2 = (1-\cos^2 x)(\text{polynome de degré } 2n-2 \text{ en } \cos x)$$

soit que :

$$\begin{cases} \beta_o+\beta_1+\ldots+\beta_{2n} = 0 \\ \beta_o-\beta_1+\ldots+\beta_{2n} = 0 \end{cases}$$

2) *PROPRIETE :*

Le filtre $F_{\pi/2}$ étant supposé stable, F_{x_1,x_2} est aussi un filtre stable.

En effet si X est un pole de $F_{\pi/2}$ et Y un pole de F_{x_1,x_2}

$$X = \frac{d(1-Y^2) - [(1+p) + 2(p-1)Y + (1+p)Y^2]}{d(1-Y^2) + [(1+p) + 2(p-1)Y + (1+p)Y^2]}$$

Ce qui entraine, puisque $|X| < 1$, et en posant $Y = \rho e^{i\varphi}$

$$(p+1) - 2\rho\cos\varphi(p-1)(1-\rho^2)-\rho^4(p+1) > 0$$

soit

$$(1-\rho^2)[(p+1)\rho^2+2(p-1)\cos\varphi\rho + (p+1)] > 0$$

$$\delta = -4p-(p+1)\sin^2\varphi < 0 \quad \forall p > 0$$

On a donc nécessairement $1 - \rho^2 > 0$ soit $\rho < 1$.

3) Pôles de F_{x_1,x_2}

x_k (k=1,...,n) étant les pôles de $F_{\pi/2}$, ceux de F_{x_1,x_2} sont

$$y_k = \frac{-(p-1) \pm \sqrt{a_k^2 - 4p}}{a_k+p+1} \qquad k=1,2,\ldots,n \text{ (le signe - donnant les n autres pôles)}$$

où $a_k = d(1-x_k) / (1+x_k)$.

4.7. OPTIMALITE ET QUASI-OPTIMALITE

La notion de quasi-optimalité est intervenue dans la construction de $F_{\omega,\pi-\omega}$. En effet, il n'a pas été possible de montrer que :

Si la fonction poids est symétrique (par rapport à $\pi/2$), parmi le(s meilleur(s) approximant(s) de f (x) dans T_{2n}^{2n} il en existe un qui soit symétrique par rapport à $\pi/2$.

Il est encore facile de montrer que :

Si le meilleur approximant est unique alors ce meilleur approximant est symétrique.

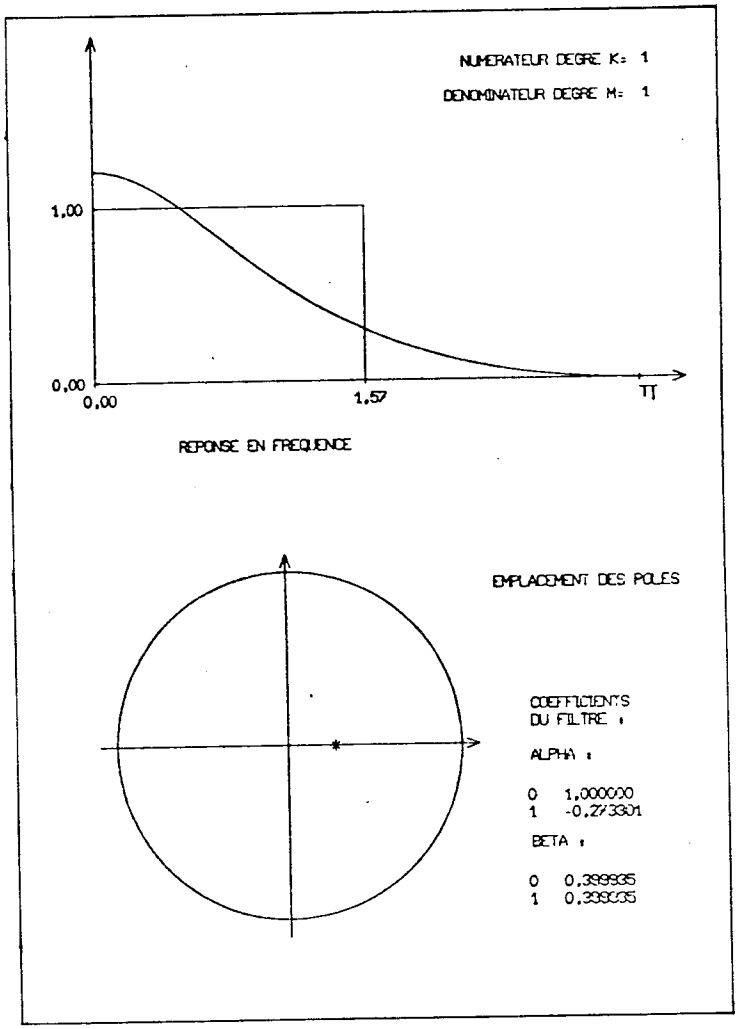
Si un meilleur approximant symétrique existe alors la notion de quasi-optimalité disparait et doit être remplacée par l'optimalité.

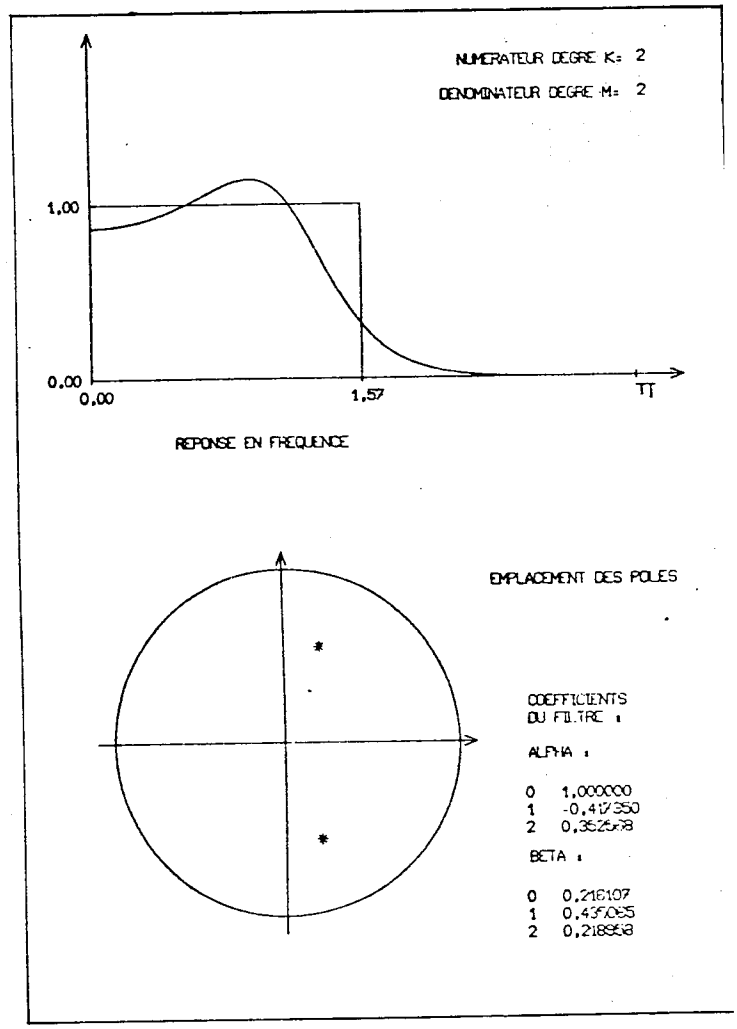
4.8. EXEMPLES DE CONSTRUCTION DE FILTRES.

La figure V .11 est la traduction graphique de la table de la page 139. La "sortie" des programmes de calcul de filtres, comme ceux de la figure V .12 aussi, se font sur deux supports :

- listings présentation comme page 139
- traceur Benson, qui permet de
  - tracer la fonction d'approximation, c'est-à-dire le carré du module de la réponse en fréquence
  - localiser les pôles du filtre étudié, ce qui donne des indications sur les transitoires, en sortie, dûs au filtre
  - indiquer les coefficients du filtre

Les filtres passe-bas, passe-haut, et passe-bande de la figure V .12, ont été construits, à partir des données sur le filtre de base de degré 3, suivant les méthodes indiquées par les théorèmes 4, 5 et 6.





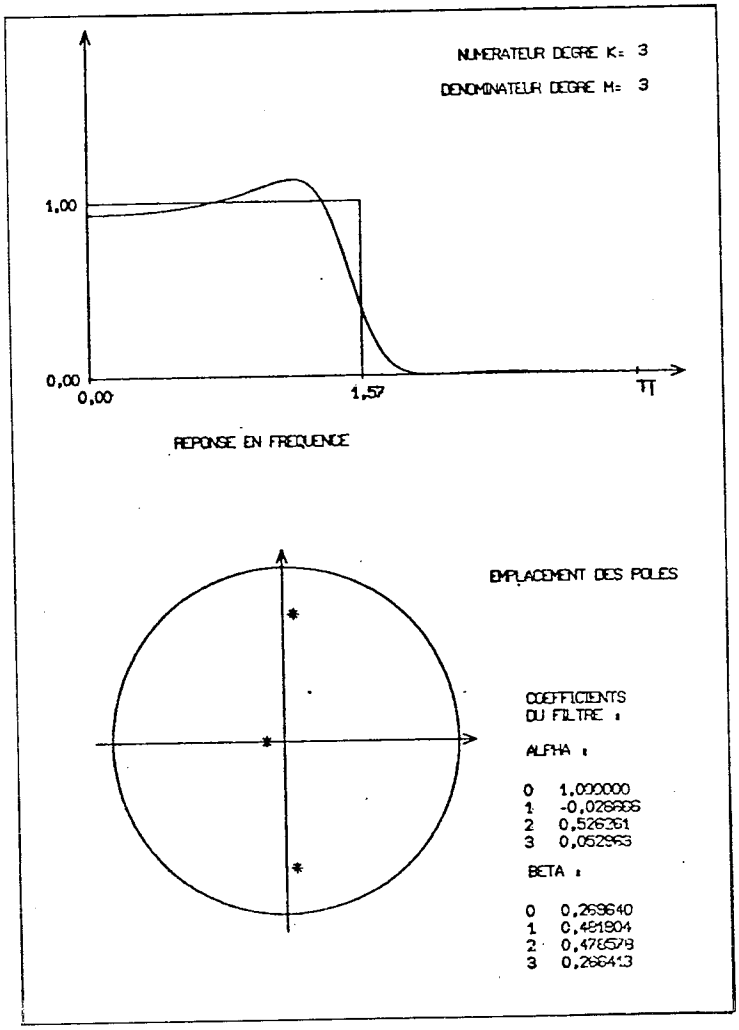
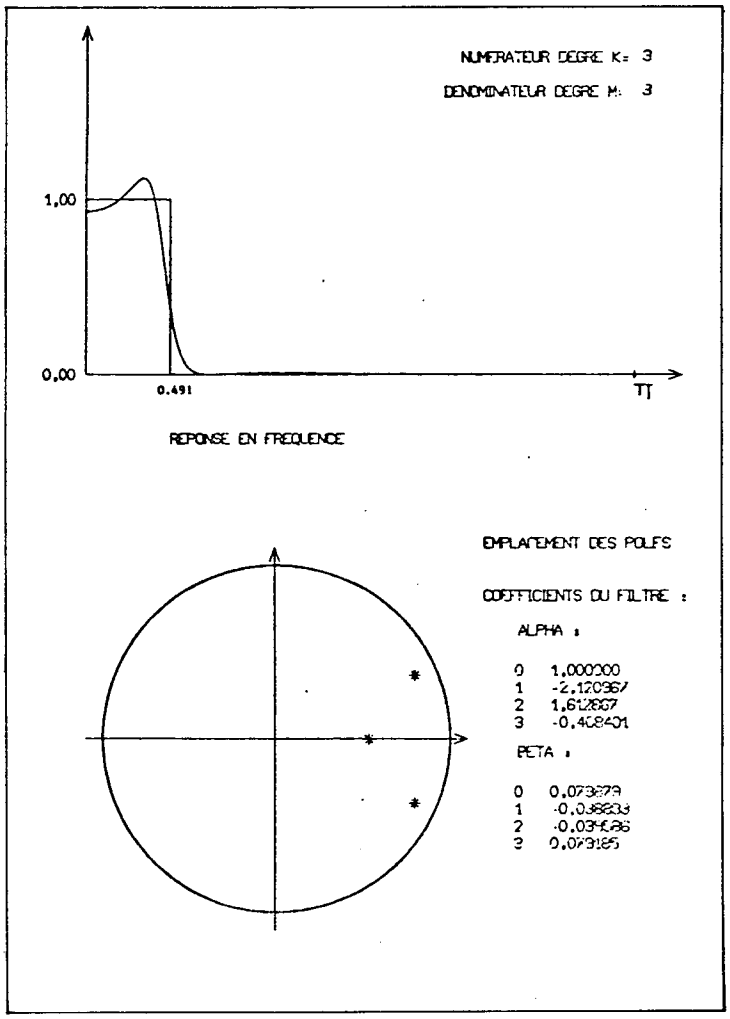
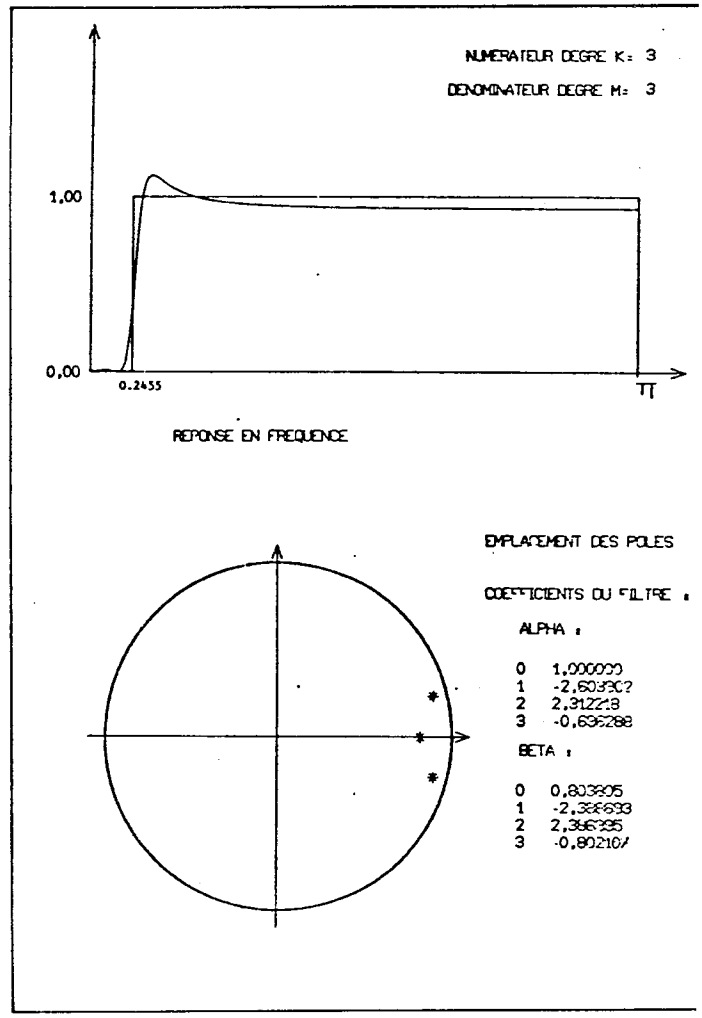


Figure V.11 a | b / c

Filtres de base de degré 1 (a), 2 (b), 3 (c).





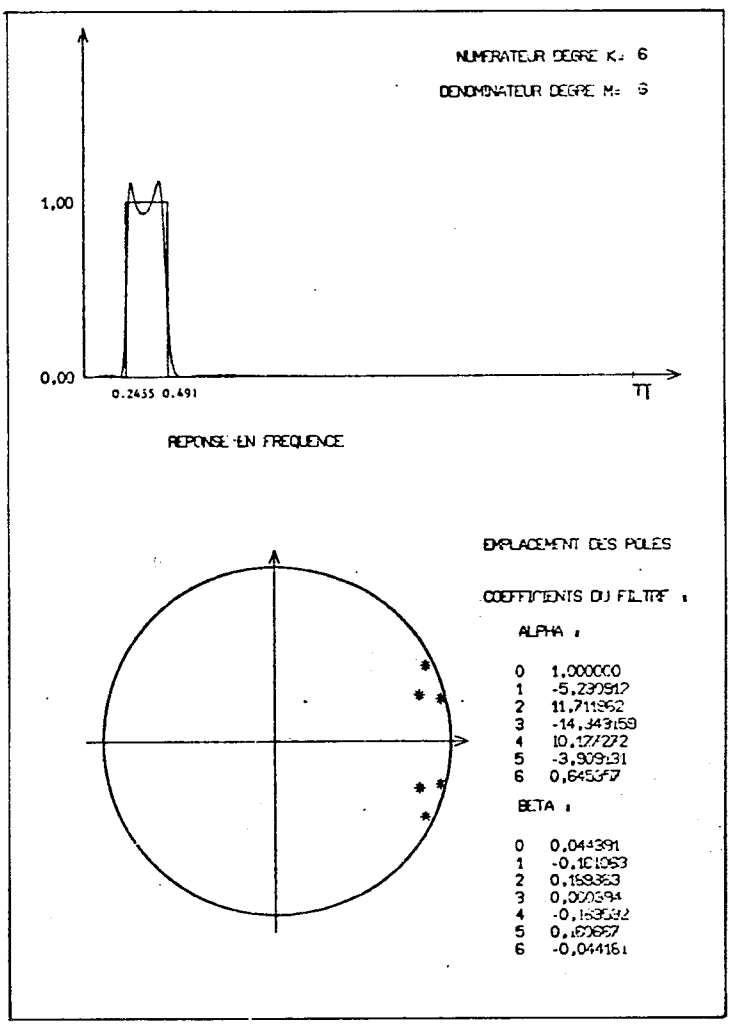


Figure V.12 $\begin{array}{c|c} a & b \\ \hline c & \end{array}$

Filtres passe-bas a
passe-haut b
passe-bande c

CHAPITRE VI

APPROXIMATION PAR FRACTIONS RATIONNELLES GENERALISEES

CHAPITRE VI

VI -1 - FRACTIONS RATIONNELLES GENERALISEES

Le problème de l'existence d'un meilleur approximant (rationnel généralisé) d'un élément de $L^2_{[-1,1]}$ en norme L^2 n'est pas simple, et n'est pas encore résolu. Les études sur ce sujet sont rares et seuls quelques articles y sont consacrés. Cheney-Goldstein [10] montrent l'existence d'un tel approximant lorsque l'élément à approcher est continu, et lorsque l'ensemble des approximants vérifie une certaine condition (C). Le problème de l'unicité, lorsqu'il y a existence, n'est pas abordé si ce n'est par quelques contre-exemples remarquables [34], [35].

1.1. NOTATIONS - DEFINITION

Sur $L^2[-1\ 1]$, on considère la norme

$$\|f\| = \left[\int_{-1}^{+1} w(x)\ f^2(x)dx\right]^{1/2}$$

où $w(x)$ désigne une fonction poids positive définie sur $[-1,1]$, et appartenant à $L^\infty[-1,1]$.

Dans $L^2[-1,1]$, soient :

* P_1 l'espace vectoriel engendré par

$\varphi_o, \varphi_1, \ldots, \varphi_m$ supposés linéairement indépendants

* Q_1 l'espace vectoriel engendré par :

$\Psi_o \equiv 1, \Psi_1, \ldots, \Psi_n$ supposés linéairement indépendants,

soit aussi

$$Q_1^+ = \{q : q \in Q_1, q(x) > 0 \quad \forall x \in [-1,1]\} .$$

Les fonctions de l'ensemble :

$$R(P_1,Q_1) = \{p/q \ ; \ p \in P_1 \ , \ q \in Q_1^+\}$$

sont appelées <u>fractions rationnelles généralisées</u>.

Un élément $r^* \in R(P_1,Q_1)$ est dit meilleur approximant de $f \in L^2[-1,1]$ si :

$$\|r^*-f\| \leq \|r-f\| \qquad \forall r \in R(P_1,Q_1)$$

Dans des conditions aussi générales sur P_1, Q_1 il n'est pas possible de montrer l'existence d'un tel meilleur approximant à moins que P_1 et Q_1 ne soient ou des polynomes ou des polynomes trigonométriques [52], [72].

On considère alors, dans $L^2[-1,1]$

* P espace vectoriel engendré par

$\varphi_o, \varphi_1, \ldots, \varphi_m$ supposés linéairement indépendants, de plus on suppose : P contient un élément , $\bar{p}$, dont le support est dense dans $[-1,1]$.

Le support d'une fonction g est la fermeture de l'ensemble.

$\{x / x \in [-1,1] , g(x) \neq 0\}$.

* Q espace vectoriel engendré par :

$\Psi_o \equiv 1$, $\Psi_1, \ldots, \Psi_n$ supposés linéairement indépendants, de plus on suppose que :

* ce système est un ensemble de Haar [40] (H2)

* il existe un élément $\bar{q}(x) > 0 \quad \forall x \in [-1,1]$ tel que

$\bar{q}^o, \bar{q}^1, \ldots, \bar{q}^k, \ldots$ soit fondamental dans $L^2[-1,1]$ [54] (H3)

Et soit encore

$$Q^+ = \{q : q \in Q , q(x) > 0 \quad \forall x \in [-1,1]\} .$$

On pose :

$$R_n^m = \{p/q ; p \in P , q \in Q^+\}$$

Dans la suite on supposera $n \geq 1$ (le cas $n = 0$ étant un problème sur un ensemble d'approximants formant une variété linéaire, sans intérêt ici).

Supposons $\|f\| = 0$ il est alors évident que :

$$\|r-f\| = \|r\|$$

et le meilleur approximant existe c'est $r(x) \equiv 0$ soit donc

$$p(x) \equiv 0 .$$

PROPOSITION 1 :

Si $\|f\| \neq 0$ *, il existe un élément* $\tilde{r}$ $(\neq 0)$ *de* R_n^m *tel que*

$$\|\tilde{r}-f\| < \|f\|$$

Si $r = \frac{\bar{p}}{q}$ $(\in R_n^m)$, on doit minimiser

$$\int_{-1}^{+1} w(x)\left[f(x)-\alpha \frac{\bar{p}(x)}{q(x)}\right]^2 dx$$

En posant :

$$F(q) = \operatorname*{Min}_{\alpha} \int_{-1}^{+1} w(x)\left[f(x) - \frac{\alpha \bar{p}(x)}{q(x)}\right]^2 dx = \|f\|^2 - \frac{\left(\int_{-1}^{+1} w(x) \frac{\bar{p}(x)f(x)}{q(x)} dx\right)^2}{\int_{-1}^{+1} w(x) \frac{\bar{p}^2(x)}{(q(x))^2} dx} ,$$

on a évidemment le dénominateur positif et il suffit alors de montrer que :

$$\exists\ q \in Q^+ \text{ tel que } \int_{-1}^{+1} w(x)\bar{p}(x) \frac{f(x)}{q(x)} dx \neq 0$$

Soit $\bar{q}(x)$ un élément de Q^+ satisfaisant (H3) et de plus tel que

$$\operatorname*{Max}_{x\in[-1,1]} \bar{q}(x) < 1 \qquad (1)$$

Considérons alors, pour $0 \leq \lambda \leq 1$ (2)

$$q(x) = 1 + \lambda\, \bar{q}(x) \in Q^+$$

et

$$\theta(\lambda) = \int_{-1}^{+1} w(x)\bar{p}(x)\frac{f(x)}{1+\lambda\bar{q}(x)}\,dx = \int_{-1}^{+1} w(x)\bar{p}(x)f(x)\Big[\sum_{0}^{\infty}(-1)^k\bar{q}^k(x)\lambda^k\Big]dx$$

$$\theta(\lambda) = \sum_{0}^{\infty}(-1)^k\Big[\int_{-1}^{+1} w(x)\bar{p}(x)(\bar{q}(x))^k f(x)dx\Big]\lambda^k$$

puisque d'après (1) et (2) $|\bar{q}(x)\lambda| < 1 \quad (\forall x \in [-1,1])$, la série converge uniformément et absolument.

$$\theta(\lambda) \equiv 0 \quad \forall\ \lambda \in [0,1] \iff f(x) = 0 \quad \text{p.p.} \quad x \in [-1,1]$$

De droite à gauche la proposition est évidente. Dans l'autre sens, puisque

$$\theta(\lambda) \equiv 0 \quad \forall\ \lambda \in [0,1]$$

$$\int_{-1}^{+1} w(x)\bar{p}(x)(\bar{q}(x))^k f(x)dx = 0 \qquad k=0,1,\ldots$$

donc puisque l'ensemble $\bar{q}^0,\bar{q}^1,\ldots,\bar{q}^k$, est fondamental ceci entraine que $\bar{p}(x)f(x) = 0$ p.p. sur $[-1,1]$, et donc à cause de (H1) que $f(x) = 0$ p.p. sur $[-1,1]$.

1.2. NORMALISATION DES COEFFICIENTS

Soient $p(x) = b_0\varphi_0(x)+\ldots+b_m\varphi_m(x)$

$$q(x) = a_0\Psi_0(x)+\ldots+a_n\Psi_n(x)$$

On désigne par $b \in \mathbb{R}^{m+1}$, et $a \in \mathbb{R}^{n+1}$ les vecteurs des coefficients de ces éléments de P, et de Q^+ et par

$$E(a,b) = \|r-f\|^2 \geq 0 \quad , \quad \rho = \inf_{a,b} E(a,b)$$

Puisque $p(x)$, pour la meilleure approximation, si elle existe, ne peut être identiquement nul ($\|f\| \neq 0$), on peut s'en tenir au cas où b est tel que

$$\|b\| = \Big(\sum_{0}^{m} b_i^2\Big)^{1/2} = 1 \quad .$$

VI .2 - REGULARISATION DE L'ENSEMBLE D'APPROXIMANTS.

Soient $\{x_i\}$ $i=0,1,\ldots,n$; n+1 points distincts de $[-1,1]$ et

soit $\Gamma_\theta = \{q : q \in Q ; q(x) \geq \theta\ \emptyset(q)\ ,\ \forall x \in [-1,1] ; q \not\equiv 0\}$

où

$$\emptyset(q) = \sum_{i=0}^{n} \lambda_i q(x_i) \quad \text{avec} \quad \lambda_i > 0 \ \ \forall i\ , \quad \sum_{i=0}^{n} \lambda_i = 1$$

PROPOSITION 2 :

$\exists\ \theta_o$ $(0 < \theta_o < 1)$ tel que, pour tout θ $(0 < \theta \leq \theta_o)$, $\Gamma_\theta \subset Q^+$

Si $q \in \Gamma_\theta$, on a en particulier $q(x_i) \geq \theta\ \emptyset(q)$ $\quad i=0,1,\ldots,n$

donc $\sum_{i=0}^{n} \lambda_i q(x_i) \geq (\sum \lambda_i)\ \theta\ \emptyset(q)$

c'est-à-dire $\quad \emptyset(q) \geq \theta\ \emptyset(q)$

ce qui entraine, en supposant $0 < \theta < 1$, que $\emptyset(q) \geq 0$

donc que $q(x) \geq 0$ $\forall x \in [-1,1]$ et cela pour tout $q \in \Gamma_\theta$.

On a $\quad \emptyset(q) = 0$ si et seulement si $\quad q \equiv 0$.

En effet :

$$\emptyset(q) = \sum_{i=0}^{n} \lambda_i q(x_i) = 0$$

$\Rightarrow \quad q(x_i) = 0$ (puisque $q(x) \geq 0$ $\forall x \in [-1,1]$) $\quad i=0,1,\ldots,n$.

Il existe donc un élément de Q qui possède (n+1) racines sur $[-1,1]$, ce qui est en contradiction avec l'hypothèse (H2). Ce ne peut être que l'élément identiquement nul qui lui n'appartient pas à Γ_θ .

Donc si $q \in \Gamma_\theta$ alors $q > 0$.

Il reste à montrer que dans Γ_θ il existe $\tilde{q} > 0$, tel que :

$\tilde{q}^o, \tilde{q}^1, \ldots, \tilde{q}^k, \ldots$ soit fondamental dans $L^2[-1,1]$.

Soit $\bar{q} > 0$ un élément de Q, tel que :

* $\bar{q}^o, \bar{q}^1, \ldots, \bar{q}^k, \ldots$ soit fondamental

* $\underset{x\in[-1,1]}{\text{Max}} \bar{q}(x) < 1$.

On choisit alors θ_o tel que :

$$\theta_o = \underset{x\in[-1,1]}{\text{Min}} \bar{q}(x) \Big/ \sum_{0}^{n} \lambda_i \bar{q}(x_i)$$

nombre qui est bien inférieur à 1.

Il est alors évident que $\tilde{q} = \bar{q} \in \Gamma_\theta$ et donc à Q^+ .

Ce qui montre la proposition 2.

2.1. PROPRIETE DES ELEMENTS DE L'ENSEMBLE Γ_θ

PROPOSITION 3 :

Si $q \in \Gamma_\theta$ *alors :*

$\forall$ Λ *norme définie sur* Q , $q(x) \geq \theta' \Lambda(q)$

$$N(q) = \sum_{i=0}^{n} \lambda_i |q(x_i)| \qquad \text{avec } \lambda_i > 0 \text{ et } \Sigma \lambda_i = 1$$

$$\text{et } x_i \in [-1,1] \text{ distincts}$$

est une norme définie sur Q — à cause la condition de Haar —.

Pour $q \geq 0$, $N(q) = \emptyset(q)$,
on en déduit que pour tout $q \in \Gamma_\theta$ on a :

$$q(x) \geq \theta \, \emptyset(q) = \theta \, N(q)$$

Et si Λ est une norme quelconque (sur un espace vectoriel de dimension finie) définie sur Q :

$$Q(x) \geq \theta' \Lambda(q)$$

On pose maintenant, avec $0 < \theta \leq \theta_o < 1$

$$R_n^m(\theta) = \{p/q \; ; \; p \in P \; , \; \|b\| = 1 \; ; \; q \in \Gamma_\theta\}$$

D'après les propositions précédentes on a évidemment :

+ si $0 < \theta_1 < \theta_2 < \theta_o$ $\qquad \Gamma_{\theta_2} \subset \Gamma_{\theta_1} \subset Q^+$

+ $\bigcup_{\theta \in]0,\theta_o]} \Gamma_\theta = Q^+$

* et donc $\qquad R_n^m(\theta) \subset R_n^m \qquad \forall \; \theta \in]0,\theta_o]$.

2.2. EXISTENCE D'UN MEILLEUR APPROXIMANT DANS $R_n^m(\theta)$

PROPOSITION 4 :

Soit une suite $(a^{(p)}, b^{(p)})$ *telle que* $\|b^{(p)}\| = 1$

et $\lim_{p\to\infty} E(a^{(p)}, b^{(p)}) = \rho$. *Si* $r^{(p)} \in R_n^m(\theta)$ *alors nécessairement la suite* $\Psi(a^{(p)})$ *est bornée.* (Ψ *désignant une norme quelconque définie sur* $\mathbb{R}^{n+1}$) .

De la proposition 3 on tire immédiatement que :

$$\exists\ \theta_1 \text{ tel que si } q \in \Gamma_\theta \qquad q(x) \geq \theta_1 \Psi(a) \tag{3}$$

a étant le vecteur des coefficients de q(x) , et Ψ une norme définie sur $\mathbb{R}^{n+1}$.

Supposons que dans $\mathbb{R}^{n+1}$ la suite $a^{(p)}$ ne soit pas bornée, il existe donc nécessairement une sous-suite $a^{(p')}$ telle que :

$$\Psi(a^{(p')}) \to \infty \quad \text{si} \quad p' \to \infty$$

$$E(a^{(p')}, b^{(p')}) = \int_{-1}^{+1} w(x)\left[\frac{p(b^{(p')}, x)}{q(a^{(p')}, x)} - f(x)\right]^2 dx$$

$$= \int_{-1}^{+1} w(x)\,\frac{p^2(x)}{q^2(x)}\,dx - 2\int_{-1}^{+1} w(x)\,\frac{p(x)}{q(x)}\,f(x)dx + \|f\|^2$$

puisque $p/q \in R_n^m(\theta)$, et d'après (3)

$$\forall x \in [-1,1] \qquad \frac{p^2(x)}{q^2(x)} \leq \frac{p^2(x)}{\theta_1 \Psi^2(a^{(p')})} \leq (m+1)^2\ M^2\ /\ \theta_1^2 \Psi^2(a^{(p')})$$

Notant par $M = \max_{i=0,\ldots,n}\ \max_{x\in[-1,1]} |\varphi_i(x)|$

On a donc $\left\|\frac{p(b^{(p')})}{q(a^{(p')})}\right\| \to 0$, lorsque $p' \to \infty$

ainsi que $\lim_{p'\to\infty} \left|\int_{-1}^{+1} w(x) \frac{p}{q} f(x)dx\right| \leq \lim_{p'\to\infty} \left\|\frac{p}{q}\right\| \|f\| = 0$

Donc :

$$\rho = \lim_{p'\to\infty} E(a^{(p')},b^{(p')}) = \|f\|^2$$

ce qui est impossible d'après la proposition 1.

Donc la sous-suite $(a^{(p)},b^{(p)})$ est bornée dans $\mathbb{R}^{n+1} \times \mathbb{R}^{m+1}$, on peut donc en extraire une sous-suite convergente vers (a^*,b^*) ,

le polynome $p(b^{(p)},x)$ converge uniformément vers $p(b^*,x)$

le polynome $q(a^{(p)},x)$ converge uniformément vers $q(a^*,x) \in \Gamma_\theta$

On a évidemment $r^* = \frac{p(b^*,x)}{q(a^*,x)} \in R_n^m(\theta)$ d'où

THEOREME 1 :

Pour toute fonction f de carré sommable dans [-1,1] la meilleure approximation de f dans $R_n^m(\theta)$ existe.

2.3. CONTRE-EXEMPLE ET EXEMPLE.

Soit [50] $f(x) = (1+x) \qquad x \in [-1,1]$

Soient $P \equiv \{\varphi_o(x) = (1+x)^2\}$

$Q \equiv \{\Psi_o(x) = 1 \ , \ \Psi_1(x) = x\}$

Q satisfait bien les hypothèses (H2) et (H3)

Il est immédiat de voir que :

Il n'y a pas existence d'un meilleur approximant de f dans R_n^m .

En effet, on se propose de minimiser :

$$\int_{-1}^{+1} w(x)\left[\frac{(1+x)^2}{a_o+a_1x} - f(x)\right]^2 = R(a_o,a_1)$$

En choisissant la suite $a_o^{(p)} = 1 + \frac{1}{p} = a_1^{(p)}$

$$\lim_{p\to\infty} R(a_o^{(p)},a_1^{(p)}) = 0$$

mais

$$\lim_{p\to\infty} (a_o^{(p)}+a_1^{(p)}x) = 1+x$$

n'appartient pas à Q^+ .

La meilleure approximation de f(x) dans $R_1^o(\theta)$ existe d'après le théorème précédent, si $w(x) = 1 \quad \forall x \in [-1,1]$ et si on choisit pour $\phi(q)$

$$\emptyset(q) = \frac{1}{2}\left[q(1)+q(-1)\right]$$

On obtient pour meilleur approximant :

$$r_j^*(x) = \frac{(1+x)^2}{a^{(j)*}\left[1+(1-\theta^{(j)})x\right]}$$

$a^{(j)*}$ étant donné en fonction de $\theta^{(j)}$ par le tableau :

j	0	1	2	3
$\theta^{(j)}$	0.1	0.01	0.001	0.0001
$a^{(j)*}$	1.0283	1.0025	1.0002	1.0000

La figure montre l'erreur (figure VI .1)

$$e(x) = f(x) - r_j^*(x) \qquad x \in [-1,1]$$

Lorsque $\theta = 0.1$, 0.01

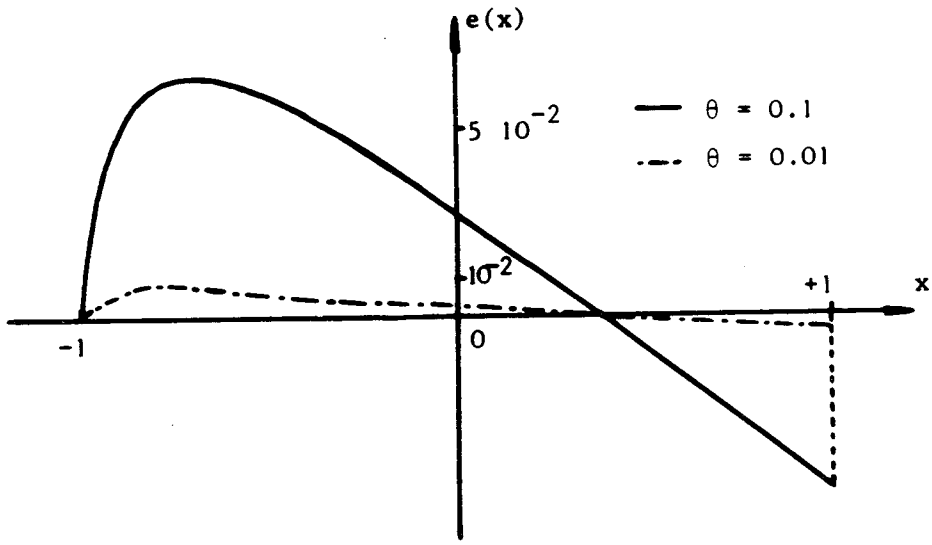


Figure VI .1

VI .3 - POINT INTERIEUR DE L'ENVELOPPE CONVEXE D'UNE COURBE (Ω) $(\in \mathbb{R}^n)$

3.1. COURBE DE HAAR

Soit (Ω) la courbe de $\mathbb{R}^n$, lieu du point $M(x) \begin{vmatrix} \Psi_1(x) \\ \vdots \\ \Psi_n(x) \end{vmatrix}$

lorsque x varie sur l'intervalle $[-1,1]$, où 1 et les $\Psi_i(x)$ $(i=1,\ldots,n)$ sont supposées linéairement indépendantes.

Soit C le point

$$C = \sum_{i=0}^{n} \lambda_i M(x_i)$$

avec $\lambda_i > 0 \quad \forall i$, $\sum_{i=0}^{n} \lambda_i = 1$ et $x_i \in [-1,1]$ distincts.

Soit encore $H(\Omega)$ l'homothétique de (Ω) dans l'homothétie de centre C et de rapport $\rho = 1 + \varepsilon$ $(\varepsilon > 0)$.

PROPOSITION 5 :

Si $Co(\Omega) \subset C\mathring{o}(H(\Omega))$ *alors* $C \in C\mathring{o}(\Omega)$.

(Co(A) signifiant enveloppe convexe de A).

$\alpha)$ $C \in \partial Co(\Omega)$

Il est alors immédiat de voir que $C \in Co(H(\Omega))$ donc dans ce cas

$Co(\Omega) \not\subset C\mathring{o}(H(\Omega))$.

β) $C \notin Co(\Omega)$,

Soit une droite issue de C contenant deux points M et M_1 distincts de $\partial\, Co(\Omega)$. Pour fixer les idées, de ces deux points soit M le plus proche de C. (Figure VI .2)

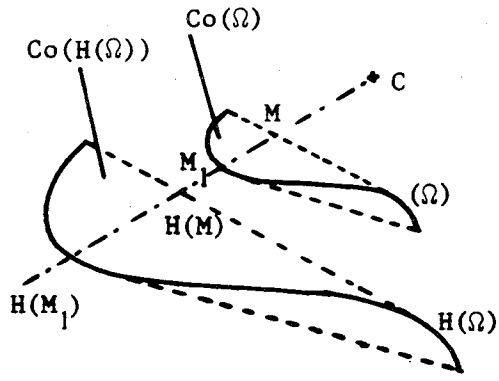


Figure VI .2

Dans l'homothétie de centre C et de rapport ρ (> 1)

$M \to H(M)$ et $M_1 \to H(M_1)$,

les seuls points de la droite CM qui appartiennent à $Co(H(\Omega))$ sont les points du segment $H(M)H(M_1)$. Donc M qui appartient à $Co(\Omega)$ ne peut appartenir à $Co(H(\Omega))$, et dans ce cas encore :

$$Co(\Omega) \not\subset \overset{\circ}{Co}(H(\Omega))$$

Ce qui montre la proposition puisque l'on vient de montrer que :

$$C \notin \overset{\circ}{Co}(\Omega) \Rightarrow Co(\Omega) \not\subset \overset{\circ}{Co}(H(\Omega)) .$$

La courbe (Ω) $(\in \mathbb{R}^n)$ est dite courbe de Haar si l'ensemble $\{1, \Psi_1, \ldots, \Psi_n\}$ vérifie la condition de Haar.

3.2. PROPRIETE DE L'ENVELOPPE CONVEXE D'UNE COURBE DE HAAR

PROPOSITION 6 :

Si l'ensemble $\{1, \Psi_1, \ldots, \Psi_n\}$ est un ensemble de Haar alors :

$$Co(\Omega) \subset \overset{\circ}{Co}(H(\Omega))$$

Montrons que $(\Omega) \subset Co(H(\Omega))$.

Le point C combinaison linéaire convexe de points de (Ω) appartient à $Co(\Omega)$, mais aussi à $Co(H(\Omega))$, en effet :

$$C = \sum_{i=0}^{n} \lambda_i M(x_i) = \sum_{j=0}^{n} \lambda_j \left[(1+\varepsilon)M(x_j) - \varepsilon \sum_{i=0}^{n} \lambda_i M(x_i)\right]$$

puisque $\sum_{i=0}^{n} \lambda_i = 1$.

Le point $N_j = (1+\varepsilon)M_j - \varepsilon \sum_{i=0}^{n} \lambda_i M_i$ est l'homothétique du point M_j dans l'homothétie de centre C et de rapport $(1+\varepsilon)$.
Donc C combinaison linéaire convexe de $(n+1)$ $(N_j, j=0,\ldots,n)$ de $H(\Omega)$ appartient à $Co(H(\Omega))$.

Soit $M_k \in (\Omega)$on peut écrire, définition de l'homothétique :

$$M_k = \frac{\varepsilon}{1+\varepsilon} C + \frac{1}{1+\varepsilon} H(M_k)$$

M_k combinaison linéaire convexe de deux points de $Co(H(\Omega))$ appartient aussi à $Co(H(\Omega))$.

Ce qui montre que $Co(H(\Omega)) \supseteq (\Omega)$, $Co(\Omega)$ est le plus petit convexe contenant (Ω) donc :

$$Co(\Omega) \subset Co(H(\Omega)).$$

Montrons maintenant que :

si $M \in \partial\, Co(H(\Omega))$ alors $M \notin Co(\Omega)$

$M \in \partial\, Co(H(\Omega)) \Leftrightarrow$ $\exists$ un hyperplan de $\mathbb{R}^n$ d'équation

$$\pi(x) = a_o + a_1 x_1 + \ldots + a_n x_n = 0$$

tel que :

$$\pi(M) = 0$$

$$\pi(P) \geq 0 \qquad \forall P \in Co(H(\Omega)).$$

Donc si $P \in H(\Omega)$, on a :

$$a_o + a_1[\Psi_1(x)(1+\epsilon) - \sum_0^n \lambda_i \Psi_1(x_i)] + \ldots + a_n[\Psi_n(x)(1+\epsilon) - \epsilon \sum_0^n \lambda_i \Psi_n(x_i)] \geq 0$$

$$\forall x \in [-1,1]$$

Soit encore en utilisant la définition de Γ_θ

$$q(x) \geq \frac{\epsilon}{1+\epsilon} \; \emptyset(q)$$

en choisissant $\theta = \frac{\epsilon}{1+\epsilon}$ (< 1) , ce qui entraine d'après la proposition 2 que :

$$q(x) > 0 \qquad \forall x \in [-1,1]$$

Puisque $M \in Co(\Omega)$, d'après [33]

$$\exists\, M_1, M_2, \ldots, M_k \in (\Omega) \quad (k \leq n+1) \quad \rho_i \geq 0 \quad \forall i \; , \quad \sum_1^k \rho_i = 1$$

tel que :

$$M = \sum_{i=1}^k \rho_i M_i$$

donc

$$\pi(M) = \pi(\Sigma \; \rho_i M_i) = \sum_1^k \rho_i \pi(M_i)$$

$$\pi(M_i) = a_o + a_1\Psi_1(x_i) + \ldots + a_n\Psi_n(x_i) = q(x_i)$$

soit donc :

$$\pi(M) = \sum_1^k \rho_i \; q(x_i) > 0$$

Ce qui montre donc que, si $M \in \partial\, Co(H(\Omega))$, $M \notin Co(\Omega)$ et ainsi la proposition 6.

THEOREME 2 : [73]

Le point $\sum_{i=0}^{n} \lambda_i M(x_i)$ $(\lambda_i > 0 \ \forall i \ , \ \sum_{0}^{n} \lambda_i = 1$ *et* $x_i \in [-1,1]$ *distincts) est un point intérieur de l'enveloppe convexe de la courbe lieu de* $M(x) \begin{bmatrix} \Psi_1(x) \\ \vdots \\ \Psi_n(x) \end{bmatrix}$ *lorsque x varie entre -1 et 1, lorsque l'ensemble* $\{1, \Psi_1(x), \ldots, \Psi_n(x)\}$ *forme un ensemble de Haar sur* $[-1,1]$.

Ce théorème est évident en utilisant la proposition 6 puis la proposition 5.

3.3. EXEMPLE ET CONTRE-EXEMPLE :

a) $\Psi_1(x) = x$

$\Psi_2(x) = x^4 - x^2 + 1$

L'ensemble $\{1, \Psi_1, \Psi_2\}$ ne vérifie pas la condition de Haar sur $[-1,1]$.

Il est immédiat de voir que le point :

$$C = \lambda_o M(-1) + \lambda_1 M(0) + \lambda_2 M(1) \quad \forall \lambda_i > 0 \quad \sum_{0}^{2} \lambda_i = 1$$

n'est pas un point intérieur de $Co(\Omega)$ (Figure VI .3).

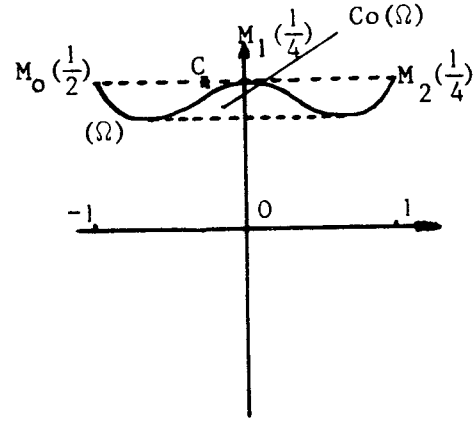


Figure VI .3

b) $\Psi_1(x) = \sin \pi x$

$\Psi_2(x) = \cos \pi x$

L'ensemble $\{1,\Psi_1,\Psi_2\}$ ne vérifie pas la condition de Haar sur $[-1,1]$.

Cependant le point $\sum_{i=0}^{2} \lambda_i M_i$ est un point intérieur à $Co(\Omega)$ (Figure VI .4)

La condition de Haar n'est donc pas une condition nécessaire.

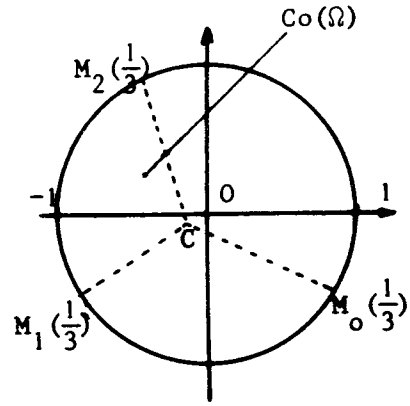


Figure VI .4

c) $\Psi_1(x) = x$

$\Psi_2(x) = x^2$

L'ensemble $\{1,\Psi_1,\Psi_2\}$ vérifie la condition de Haar sur $[-1,1]$.

Donc :

$$\sum_{0}^{2} \lambda_i M_i$$

est un point intérieur, ce que l'on voit sur la figure VI .5.

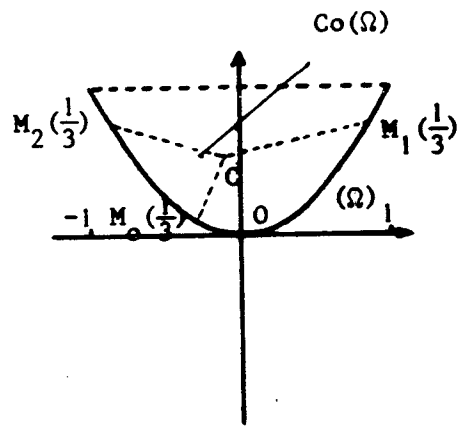


Figure VI .5

CHAPITRE VII

APPLICATION A L'ANALYSE DES SONS

VII.1 - MODELE ETUDIE

1.1. SONS ET FREQUENCES

On se propose de reconnaitre une mélodie jouée par un instrument de musique donné.

Cet instrument (correctement accordé) peut êmettre un certain nombre de sons complexes — appelés notes — formés par la superposition de plusieurs sons élémentaires (partiels). A chaque note correspond sa hauteur caractérisée par un nombre appelée fréquence fondamentale, les partiels ayant des hauteurs dont les fréquences sont supérieures à la fréquence fondamentale [49] .

Si la fréquence de chaque partiel est un multiple de la fréquence fondamentale on dira que la note a une répartition (de partiels) harmonique.

Les notes jouées par une grande partie des instruments usuels ont une répartition harmonique (sauf, surtout,les percussions).

Soit un instrument émettant à l'instant t_o, jusqu'à l'instant t_1, la note de fréquence f, on admet [18] que le signal émis par cet instrument sur l'intervalle de temps $[t_o, t_1]$ est :

$$g(t) = \sum_{k=1}^{N_f} a_k(t) \cos (2\pi kft + \varphi_k)$$

Sur l'intervalle où $a_k(t)$ est constant $\forall k$ la note est dite dans un état stationnaire. Cette partie peut exister ou ne pas exister sur un instrument donné
- sur le piano, la guitare, elle n'existe pas car les cordes sont frappées ou pincées
- sur un violon, un instrument à vent elle existe, ou peut être supprimée au gré de l'instrumentiste.

Pour des renseignements complets sur la répartition des a_k (valeur relative de l'un par rapport à l'autre, on se reportera avec intérêt au livre [42]).

Sur l'intervalle où $a_k(t)$ est effectivement fonction du temps — partie transitoire — de nombreux modèles ont été proposés [18], [6], [58], essayant de tenir compte et de l'instrument et de l'instrumentiste.
Le modèle du chapitre II en est un, qui peut être illustré par la courbe d'établissement d'un son de flûte à bec indiqué sur la figure VII. 8.

1.2. ECHELLE MUSICALE

Il ne sera pas fait ici un résumé de la théorie de la musique, ni de l'histoire de la musique qui trouvent leurs places dans des ouvrages spécialisés. Cependant on rappellera que dans la musique occidentale on a eu connaissance au cours des ages de trois échelles musicales [37] :

- gamme de Pythagore
- gamme de Zarlin
- gamme tempéré de J.S. Bach. Une telle gamme est divisée en douze intervalles égaux. C'est-à-dire que si f est la fréquence de la première note do on obtient :

$do^{\#} \equiv re^{\flat}$	fréquence	$f\ 2^{1/12}$
ré	fréquence	$f\ 2^{2/12}$
⋮		
si	fréquence	$f\ 2^{11/12}$
→ do	fréquence	$f\ 2^{12/12} = 2\ f$

La dernière note étant à nouveau un do à l'octave du précédent et ainsi de suite...

C'est cette gamme tempérée, qui, de nos jours, dans le monde occidental, est la plus utilisée dans la musique dite classique (en omettant de parler de toutes les recherches actuelles [1] , [66]) et dont nous nous serviront pour reconnaitre une mélodie jouée par un instrument donné.

Une mélodie étant une succession de notes, appartenant à une gamme tempérée, dont les durées respectives de l'une par rapport à l'autre indiqueront le rythme.

VII.2 - DETECTEUR DE MELODIE

Soit I l'instrument dont on se propose de reconnaitre une mélodie, cet instrument a alors un "registre" bien fixé c'est-à-dire que les notes qu'il peut émettre sont déterminées par sa structure physique :

violon $196 \times 2^{1/12}$ $i=0,1,\ldots,45$

piano $275 \times 2^{1/12}$ $i=0,1,\ldots,88$

Soit N_I le nombre de notes pouvant être jouées par l'instrument de musique I , et soit λ la durée minimum d'une note jouée par cet instrument — en éliminant dans un premier temps les jeux particuliers : pizzicati, glissandi...

Le détecteur idéal pourrait être un détecteur composé de N_I filtres passe-bande, "centrés" en f_i ($= 196.2^{i/12}$, par exemple pour un violon), et de "rapports de bande" $\theta^2 = 2^{1/12}$. Par rapport de bande on signifie que le rapport des fréquences de coupure est — dans ce cas — égal à $\theta^2 = 2^{1/12}$, et par filtre centré on signifie un filtre dont les fréquences de coupure sont f_i / θ et $f_i \ \theta$.

Cependant il est à peu près impossible de se servir de tels filtres car leur temps de réponse est trop long par rapport à λ , comme on pouvait déjà s'y attendre en examinant les diverses relations d'incertitude existantes [19] , [36] .

2.1. CONSTANTE DE TEMPS D'UN FILTRE PASSE-BANDE

Soit un filtre passe-bande idéal de rapport de bande θ^2 centré en F, auquel on associe le filtre quasi-optimal F de degré n (chapitre V), auquel est associé l'équation réccurrente — reliant entrée et sortie —

$$(1)\quad \begin{cases} \sum_{i=0}^{n} \alpha_i \, y_{n-i+k} = \sum_{i=0}^{n} \beta_i \, x_{n-i+k} \\ y_i \quad \text{donnés pour} \quad i=0,\ldots,n-1 \end{cases}$$

Si les conditions initiales sont telles que $x_n = 0 \quad (n < 0) \Rightarrow y_n = 0 \quad (n < 0)$ (chapitre III), le filtre sera dit causal ; sinon les conditions initiales seront telles que $y_i = 0 \quad i=0,\ldots,n-$ et le filtre sera dit non causal.

2.1.1. Courbe de réponse du filtre F

On peut écrire la solution de (1) — cas de racines simples de l'équation caractéristique — lorsque le signal d'entrée est :

$$x_p = \sin(2\pi f p \Delta t) \qquad p=0,1,\ldots$$

sous la forme :

$$y_p = \sum_{i=1}^{n} C_i r_i^p + |H(2\pi f)| \sin(2\pi f p \Delta t + \Psi_f) \qquad (2)$$

où on écrit la réponse en fréquence du système :

$$H(2\pi f) = |H(2\pi f)| e^{i\Psi_f}$$

Soient (i_k, y_{i_k}) $(k=0,1,\ldots)$ les points où $|y_p|$ est maximum. La courbe constituée des segments de droites joignant $(i_k, |y_{i_k}|)$ à $(i_{k+1}, |y_{i_{k+1}}|)$ sera appelée courbe de réponse du filtre au signal d'entrée de fréquence f.

2.1.2. Rapport de bande

Deux notes de musique successives (gamme tempérée) ont leurs fréquences telles que :

$$\text{fréquence note } 2 = 2^{1/12} \text{ fréquence note } 1$$

donc pour les séparer on fixera une fréquence de coupure du filtre à utiliser au "centre" de ces deux notes soit en

$$2^{1/24} \text{ fréquence note } 1$$

Soient $n_1, n_2, \ldots, n_{N_I}$ les notes pouvant être émises par I, dont le les fréquences sont

$$\gamma_1, \gamma_2, \ldots, \gamma_{N_I} \qquad (\gamma_{i+1} = 2^{1/12}\ \gamma_i)$$

Si on désire séparer les notes $n_k, n_{k+1}, \ldots, n_{k+p}$ des autres on construira un filtre tel que :

$$F/\theta = 2^{-1/24}\ \gamma_k$$

$$F\ \theta = 2^{1/24}\ \gamma_{k+p}$$

On a donc :

$$F = \sqrt{\gamma_k\ \gamma_{k+p}} = 2^{k/12 + p/24}\ \gamma_1$$

$$\theta = 2^{(p+1)/24}$$

Supposons que les valeurs du signal d'entrée sont échantillonnées tous les ΔT. La réponse en amplitude d'un filtre passe-bande idéal est caractérisée par ses valeurs sur $[0, 1/2\Delta T]$, c'est-à-dire par la donnée de deux nombres qui peuvent être par exemple :

- les deux fréquences de coupure
- ou le rapport de bande θ^2 et le centre de bande F.

L'emplacement de la bande passante dans l'intervalle $[0,1/2\Delta T]$ est donc — le rapport de bande étant fixé — fonction de F ou encore ce qui est équivalent du nombre $\nu = 1/(F\Delta T)$ qui est en réalité le nombre de points par période (correspondant à la fréquence centrale).

La figure VII.1 montre des réponses en amplitude de filtres passe-bande de fréquences centrales F^1, F^2, F^3 (soient $\nu_i = 20,10,5$) lorsque $\theta = \sqrt{2}$.

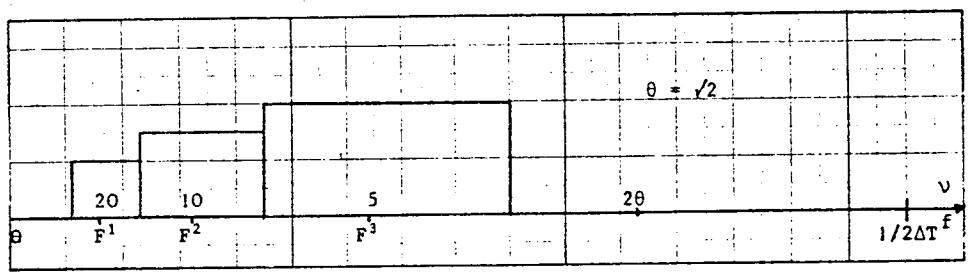


Figure VII.1

On peut remarquer que, pour que le filtre garde sont caractère de passe-bande on a nécessairement

$$F\theta < 1/2\Delta T \qquad \text{soit} \qquad \nu > 2\theta$$

2.1.3. Tracé de quelques courbes de réponse

Un filtre passe-bande idéal étant donné, on calcule les coefficients du filtre quasi-optimal correspondant de degré 2n. (ici on choisit 2n = 6) (cf. chapitre V), on trace les courbes de réponse de ce filtre F (θ,ν fixés) pour diverses fréquences du signal d'entrée (figure VII.2), qui sont telles que, si on désigne par F la fréquence centrale :

$$f_1 = F/2 \quad , \quad f_2 = \frac{F}{\theta}\, 2^{-1/24} \quad , \quad f_3 = \frac{F}{\theta}\, 2^{1/24} \quad , \quad f_4 = F$$

$$f_7 = 2F \quad , \quad f_6 = F\theta\ 2^{1/24} \quad , \quad f_5 = F\theta\ /\ 2^{1/24}$$

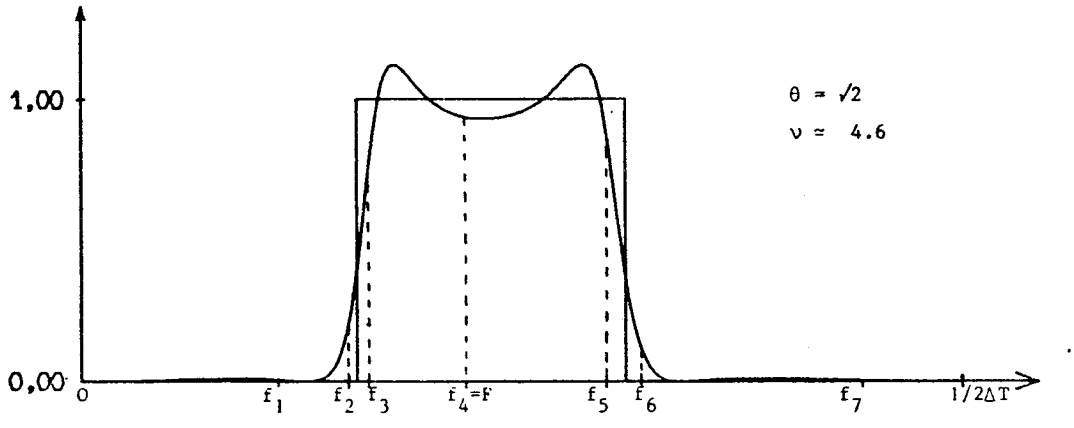


Figure VII.2

Diverses courbes de réponse se trouvent sur les figures VII.3 et VII.4 (filtres non causal) et VII.5 causal.

2.1.4. Détermination de la constante de temps de F

Sur les figures VII.3 à VII.5, on remarque que la réponse "se stabilise" au bout d'un certain temps, remarque qui aurait déjà pu être faite en considérant la relation (2) du paragraphe 2.1.1., car le filtre construit étant stable le terme :

$$\sum_{i=1}^{n} C_i r_i^p \to 0 \qquad \text{lorsque} \qquad p \to \infty$$

Puisque $y_p = \sum_{i=1}^{n} C_i r_i^p + |H(2\pi F)| \sin\left(\frac{2\pi}{\nu}\, p + \Psi\right)$, étudions la suite des maxima de $\left|\sin\left(\frac{2\pi}{\nu}\, p + \Psi\right)\right|$.

THEOREME :

Si ν est rationnel, la suite des maxima de $|\sin(\frac{2\pi}{\nu} p + \Psi)|$ ($p \geq 0$) est une suite périodique de période $k'\pi$ où k' est le plus petit entier (> 0) tel que $k'\nu = 2e$ (e étant un entier le plus petit possible).

$\sin|(\frac{2\pi}{\nu} p + \Psi)|$ est maximum pour :

$$p = p_k = \lfloor (2k+1)\frac{\nu}{4} - \frac{\Psi}{2\pi}\nu \rfloor$$

($\lfloor x \rfloor$ entier le plus proche de x).

On a :

$$p_{k+k'} = \lfloor (2k+1)\frac{\nu}{4} - \frac{\Psi}{2\pi}\nu + \frac{k'\nu}{2} \rfloor = p_k + e$$

si $\frac{k'\nu}{2}$ = e entier (choisi le plus petit possible), donc

$$|\sin(\frac{2\pi}{\nu} p_{k+k'} + \Psi)| = |\sin(\frac{2\pi}{\nu} p_k + \Psi + k'\pi)| = |\sin(\frac{2\pi}{\nu} p_k + \Psi)|$$

ce qui montre le résultat.

On appellera constante de temps à ε près le nombre $\tau_\varepsilon = p^*/F$ tel que :

si C désigne la limite, lorsque $p \to \infty$, de la moyenne arithmétique de k' points successifs de la courbe de réponse de F pour le signal d'entrée de fréquence F :

$$\left| \sum_{j=0}^{k'-1} |y_{i_j}| / k' - C \right| \leq \varepsilon C \qquad \forall\ i_o \geq p^*$$

On a évidemment :

$$C = \lfloor \sum_{k=0}^{k'-1} \sin(\frac{2\pi}{\nu} \lfloor (2k+1)\frac{\nu}{4} - \frac{\Psi}{2\pi}\nu \rfloor + \Psi) \rfloor \frac{|H(2\pi F)|}{k'} .$$

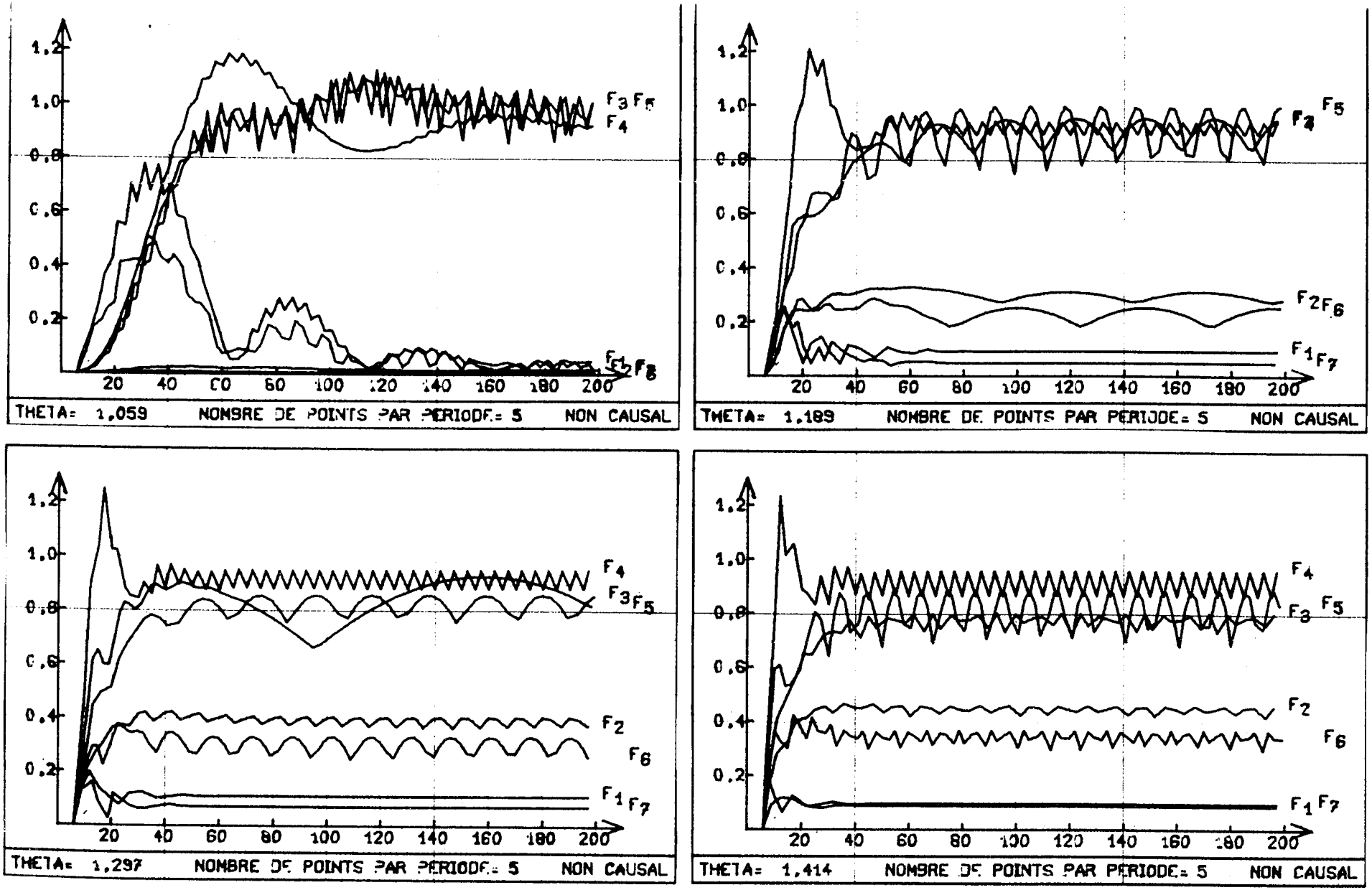


Figure VII.3 $\begin{array}{c|c} a & b \\ \hline c & d \end{array}$

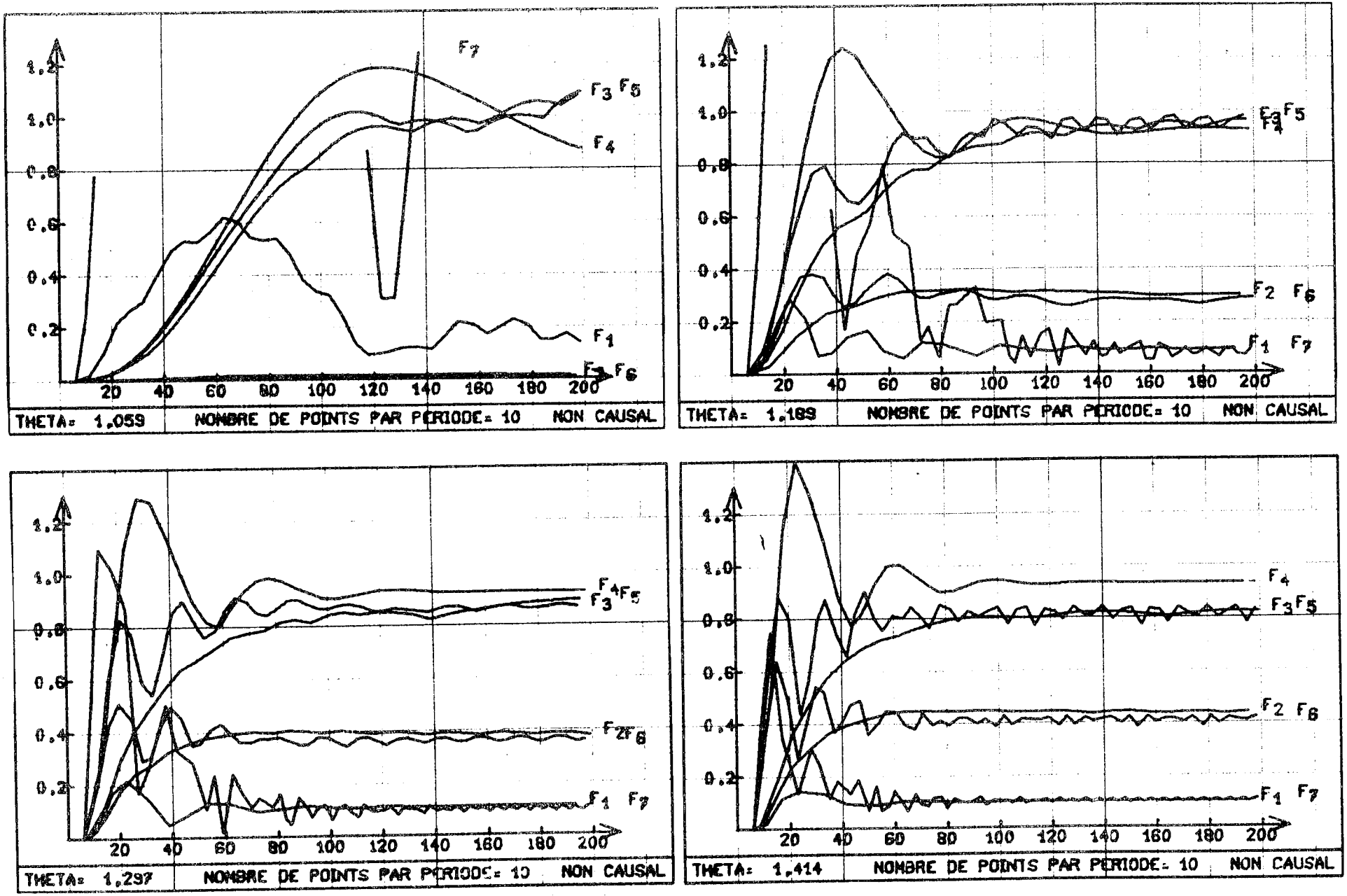
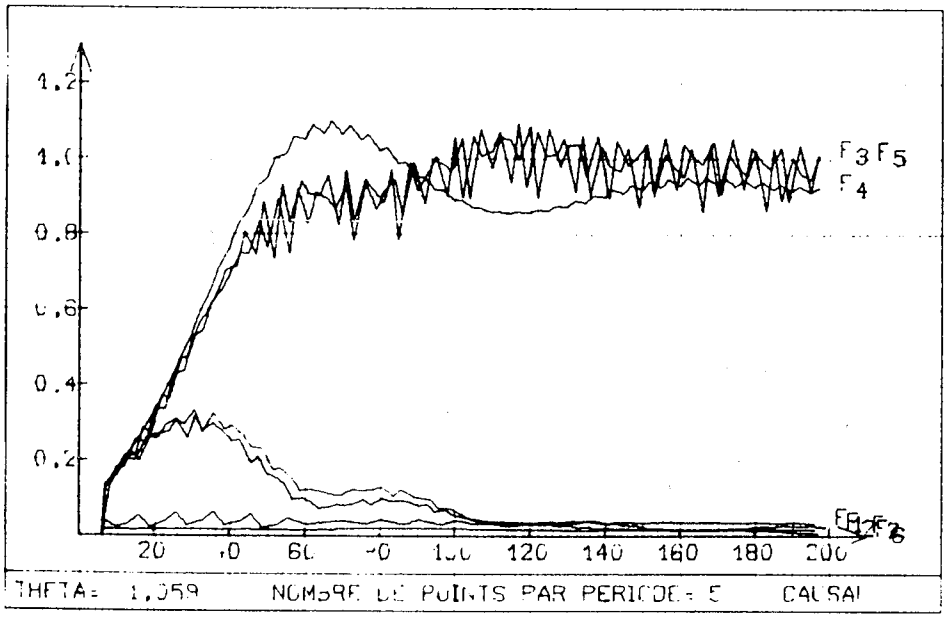
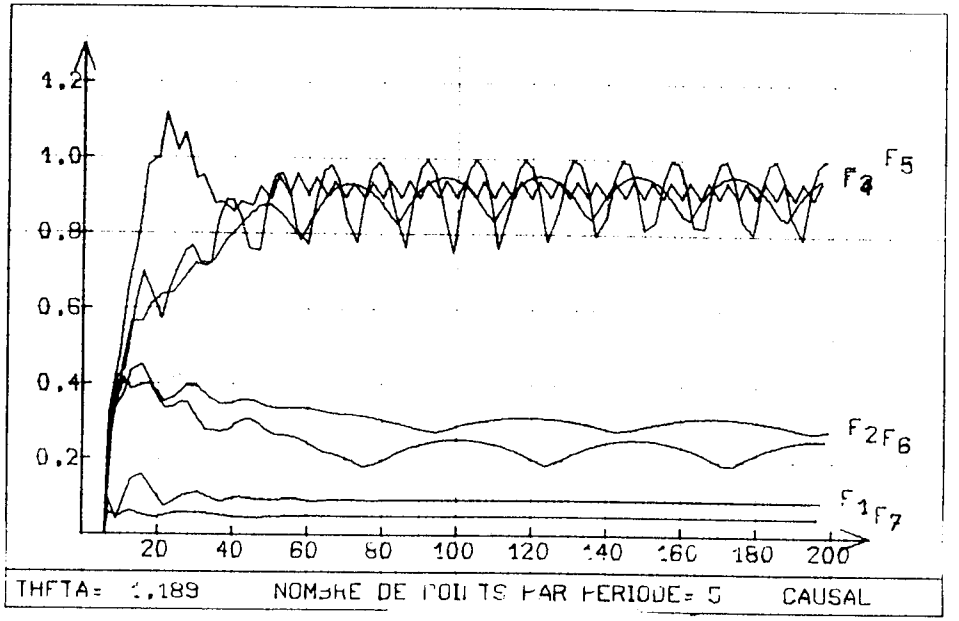
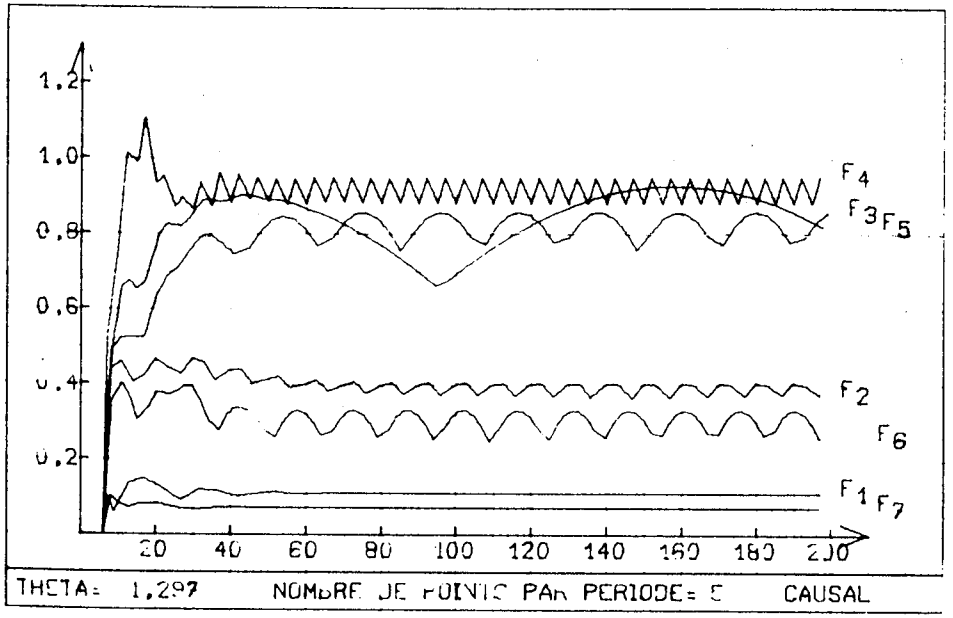


Figure VII.4 $\frac{a|b}{c|d}$







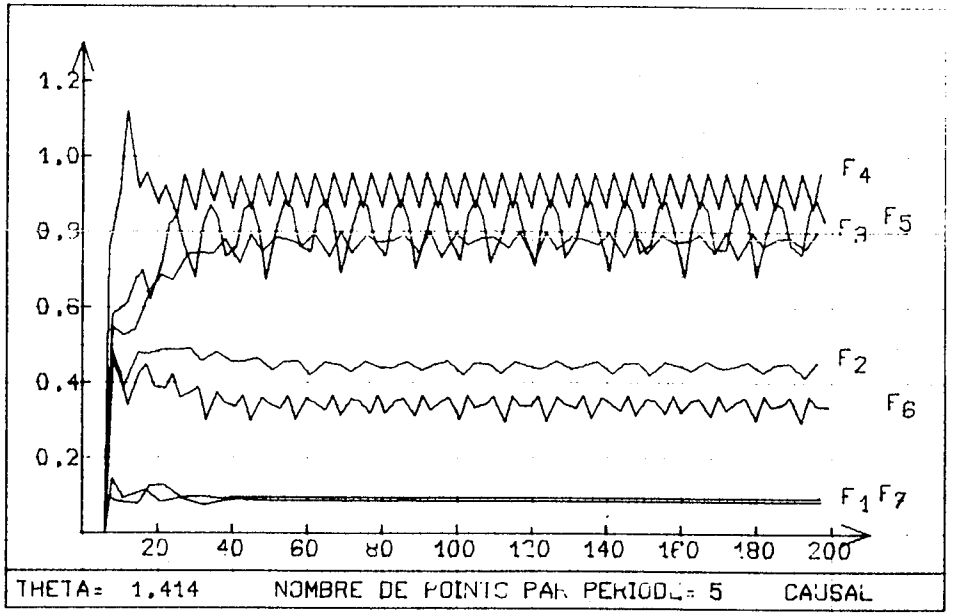


Figure VII.5

a	b
c	d

On remarque, lors du calcul de τ_ε (ici on a choisi $\varepsilon = 0.05$, et $\tau_{0.05}$ sera systèmatiquement notée par τ), que

τ ne dépend pratiquement pas de ν

τ est essentiellement fonction de θ

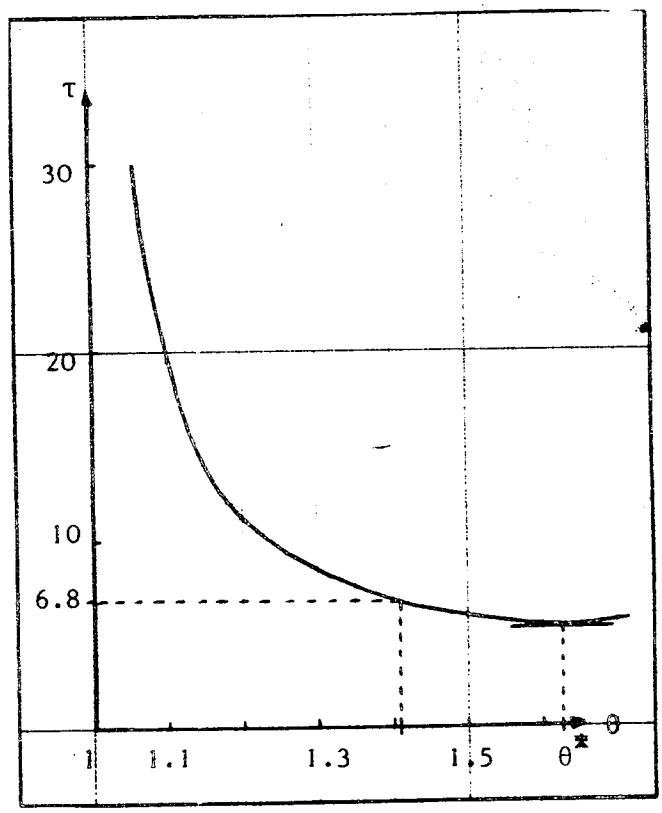


Figure VII.6.

Aussi on a tracé (figure VII.6), expérimentalement, τ en fonction de θ, ce qui montre :

La constante τ n'est pas fonction de la largeur de bande mais du rapport de bande.
Il existe un θ^* optimal tel que pour cette valeur la constante de temps τ est minimum.

2.2. MODELE D'UN DETECTEUR DE MELODIE

L'étude de $\tau = \tau(\theta)$, indique que pour avoir une réponse stabilisée rapide à une entrée sinusoïdale d'un filtre F le rapport de bande de ce filtre doit être voisin de θ^{*2} . Cependant dans l'analyse du son émis par un instrument il n'est pas possible d'utiliser un tel rapport : en effet, si la note de fréquence f est émise cette note possède le partiel de fréquence 2f, fréquence qui peut aussi être celle d'une note émise par l'instrument. Il est donc nécessaire que le rapport de bande soit tel que si f appartient à la bande passante, 2f ne puisse y appartenir. C'est ainsi que l'on choisira pour rapport de bande $\theta^2 = 2$ — les filtres associés sont appelés filtres d'octave —

A ces filtres permettant de discerner un paquet de douze notes, on adjoindra, un compteur de passage par zéro (cpz) du signal de sortie. En effet, la sortie étant stabilisée, le signal est,en première approximation, une sinusoïde pure (les intensités des différents harmoniques sont négligeables par rapport à celle du fondamental), dont il est alors rapide de trouver la période ou fréquence (cf. chapitre I).

Soit un instrument I dont le registre est $[n_1, n_I]$
(n_1 et n_I notes la plus basse et la plus haute pouvant être émises par I.)
Son registre s'étend donc sur

$$N_I = \frac{\lfloor I+11 \rfloor}{12} \quad \text{octaves}$$

($\lfloor x \rfloor$ signifiant le plus grand entier inférieur ou égal à x).

Le détecteur de mélodie est ainsi composé de N filtres d'octaves (figure VII.7). Soit λ la durée minimale d'une note — pour être plus précis il faudrait parler de $\lambda(f)$ — émise par l'instrument considéré. Soit h le temps nécessaire pour obtenir la "sortie" des filtres d'octave (et du cpz utilisé). On pose :

$$\rho = \lambda / h$$

Si entre les instants t_i et $t_{i+p} = t_i + ph$ où $p \geq \rho$, on constate que la fréquence f_k est émise et uniquement celle-là — f_k pouvant cependant varier dans un intervalle de rapport $2^{1/12}$ — .

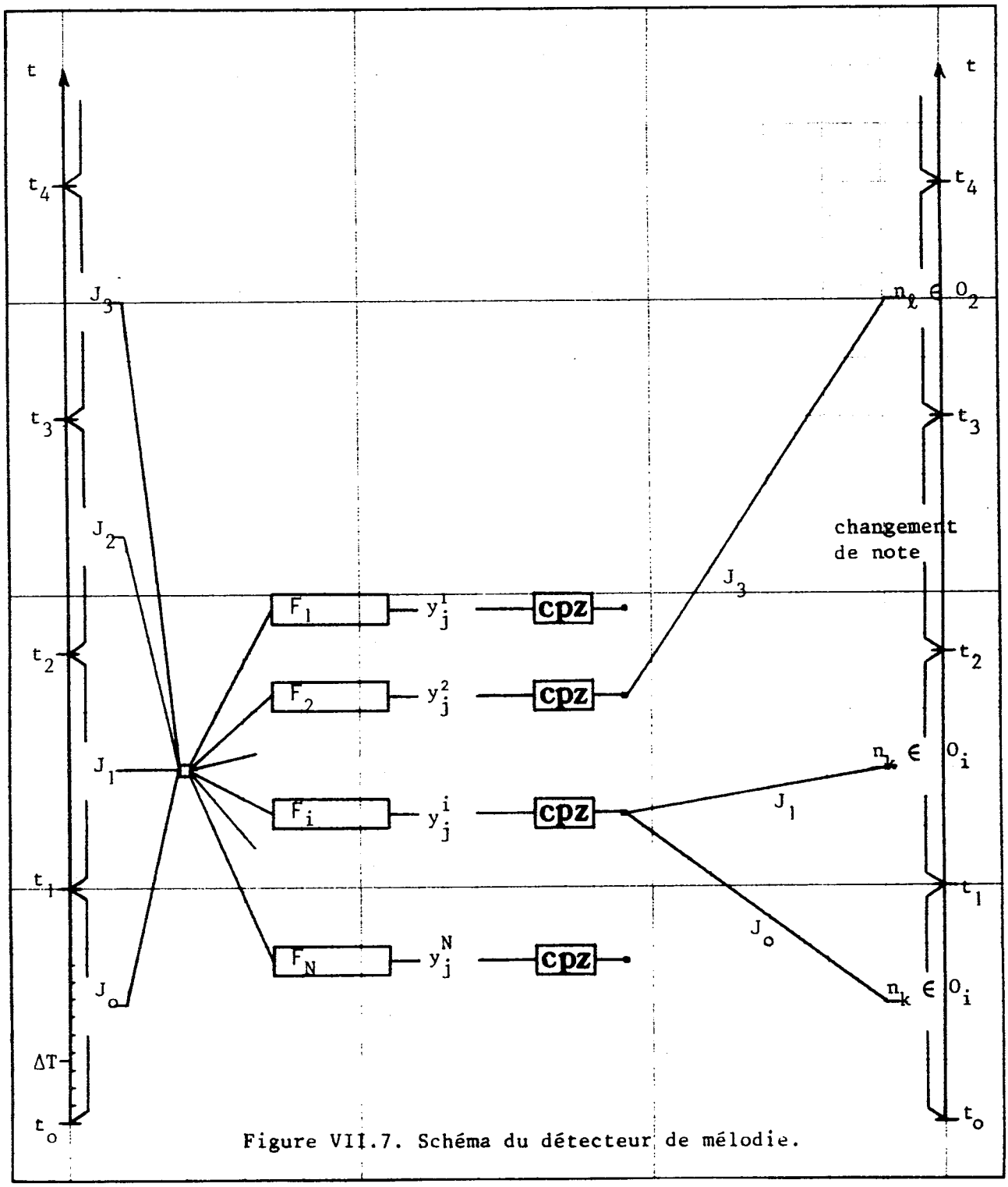


Figure VII.7. Schéma du détecteur de mélodie.

On dira que la note n_k a été émise à partir de l'instant t_i , pendant une durée ph.

2.3. MELODIE JOUEE PAR UNE FLUTE A BEC

A titre documentaire la figure VII.8 représente un morceau de l'"attaque" — transitoire — d'une note par une flûte à bec.

Figure VII.8. Etablissement d'un son de flûte.

Les données techniques du programme sont alors les suivantes

λ = 1/10 seconde (calculée à partir des courbes de réponses)

ρ = 4

Le nombre de filtres d'octaves utilisé est 3 — dû au registre relativement peu étendu de cet instrument. La mélodie jouée dans l'exemple est le suivant

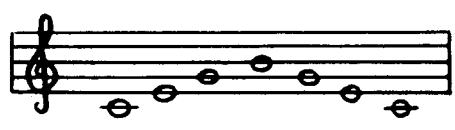


Figure VII.9

Le pas de discrétisation est T = 1/6500 seconde . Le programme d'analyse des sons est alors activé sur le signal discrétisé et les résultats (sortie ordinateur) sont :

note (fréquence)		instant initial	durée (sec.)
do#	554.8	0	0.258
do#	545.6	0.301	0.474
fa	689.0	1.246	0.948
sol#	822.0	2.323	0.904
do	1055	3.40	0.818
sol#	819.2	4.381	0.990
fa	683.3	5.430	0.206
do#	541.8	6.545	1.034

Pour illustrer cette sortie de résultats assez peu parlante musicalement et mathématiquement, on a représenté (figure VII.10), sur un même graphique la sortie du filtre excité et de son cpz (courbe en pointillé) ainsi que les résultats du tableau précédent.

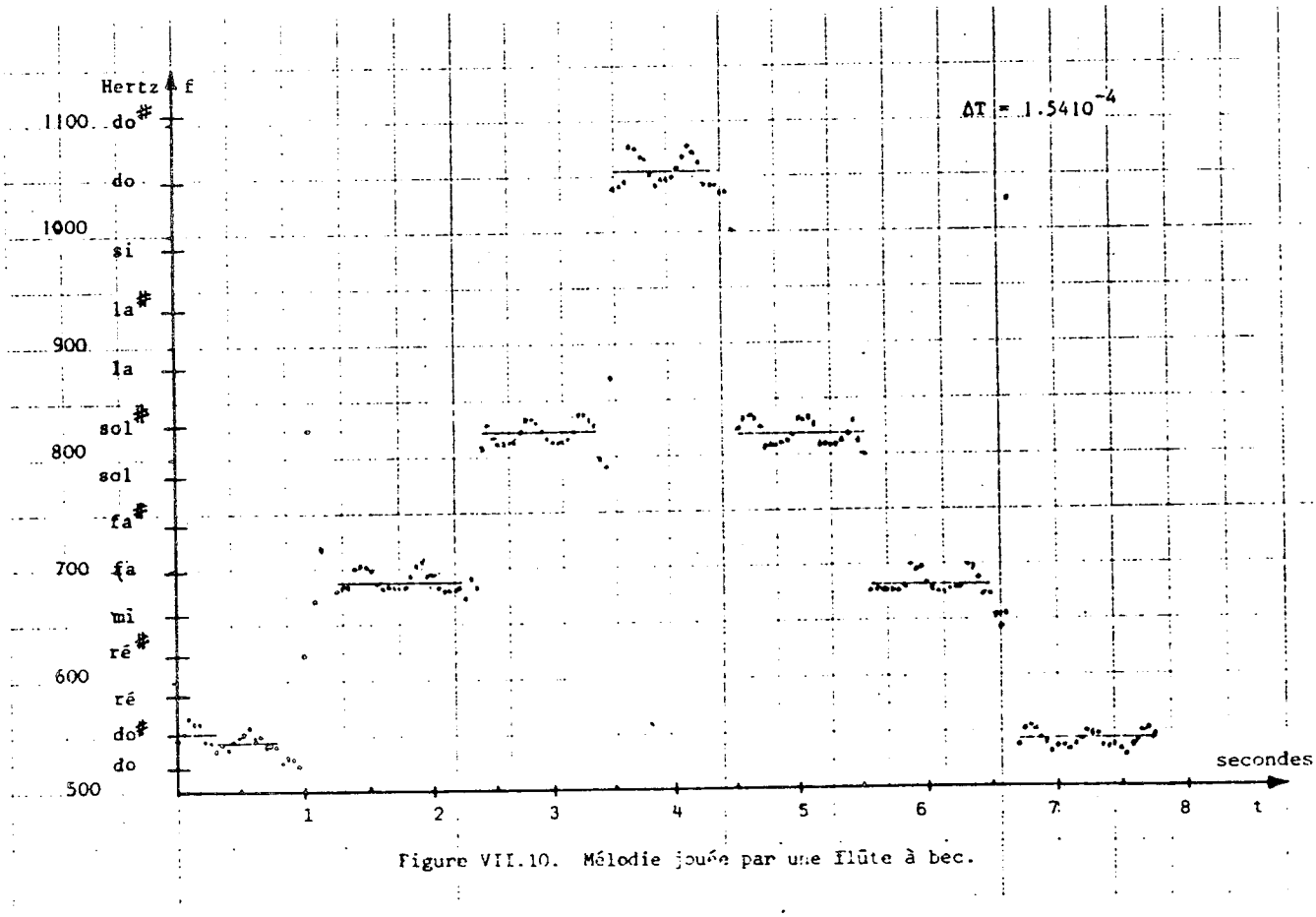


Figure VII.10. Mélodie jouée par une flûte à bec.

REMARQUES :

- Les deux types de sorties proposées sont assez peu satisfaisantes pour un musicien, aussi il est envisageable et envisagé de sortir [27] la mélodie sur le support habituel du musicien — c'est-à-dire une portée et ceci aussi à l'aide d'un ordinateur .

- Il faut cependant voir l'avantage du tableau qui permet de voir si l'instrumentiste joue... juste ou si l'instrument est parfaitement accordé, ce qui n'a pas été le cas de l'exemple précédent où la flûte jouait en réalité presque un demi-ton au-dessus des notes fixées.

VII.3. - CONCLUSION

Les programmes écrits pour l'analyse des sons ont montré qu'ils utilisaient des techniques valables et correctement mises au point. Cependant, actuellement, il faudra les améliorer au point de vue temps de calcul — point qui, il faut l'avouer, a été un peu négligé — afin d'arriver à travailler en temps réel, et nous pensons que cela est possible grâce aux filtres de faible degré utilisés. Il faut remarquer que le temps passé au diagnostic automatique des seuils est sans doute plus élevé que celui passé au filtrage proprement dit, et souligner que ce diagnostic se fait d'une manière dynamique, ce qui est d'ailleurs la seule voie admissible pour un tel problème.

Ces méthodes de calcul ont trouvé ici une application particulière — l'analyse des sons — mais il va de soi qu'elles sont utilisables dans bien d'autres domaines où la notion de fréquence est appelée à jouer un grand rôle.

BIBLIOGRAPHIE

[a] IEEE Transactions on circuit theory

[b] IEEE Transactions on audio and electroacoustic

[1] BARBAUD : La musique, discipline scientifique.
Dunod.

[2] BASS J. : Cours de Mathématiques, Tome II.
Masson 1971.

[3] BELLMANN R. : Introduction to matrix analysis.
Mc Graw Hill (1960).

[4] BERTRANDIAS J.P. : Analyse fonctionnelle.
Armand Colin (1970).

[5] BLACKMAN R.B., TUKEY : The measurement of power spectra from the point of view of communication engineering.
Dover (1959).

[6] BLADIER B. : Sur les phénomènes transitoires des cordes vibrantes.
Acustica, vol.14 (1964).

[7] BONZANIGO F, PELLANDINI F. : Problèmes de réalisation des filtres digitaux. Revue de l'AGEN n° 9, vol. 1 (Juillet 1969).

[8] BUTZER P.L., NESSEL R.J. : Fourier Analysis and Approximation.
Birkauser Verlag, vol. I, (1971).

[9] CARTAN H. : Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables. Hermann (1961).

[10] CHENEY E.W., GOLDSTEIN : Mean square approximation by generalized rational functions.
Math. Zeitschr. 95, 232-241 (1967).

[11] CHENEY E.W., LOEB H.L. : Generalized rational approximation.
J. SIAM Numer. Anal. Ser. B, Vol. 1, (1964) pp 11-25.

[12] CISEK : Discrete Hilbert Transform.
IEEE Trans. Audio and Electro. (Dec. 1970).

[13] COOLEY, TUKEY : An algorithm for the Machine Calculation of Complex Fourier Series.
Math. of Comput., Vol. 19, pp 297-301 (April 1965).

[14] DOETSCH : Guide to the applications of the Laplace and Z-transforms.
Van Nostrand Reinhold (1971).

[15] EL MALLAWANY : Le filtrage numérique.
Annales dés télécommunications, t. 24, n° 3-4 (1969).

[16] FLETCHER, POWELL : A rapidly convergent descent method for minimisation.
Computer Journal, vol. 6 (1963).

[17] FREEDMAN M.D. : Analysis of musical instrument tones.
J. Acoust. Soc. Ame. 41 793 (1967).

[18] FREEDMAN M.D. : A technique for the analysis of musical instrument tones.
U.S. Public Health Grant GM 10718 (03)
Techn. Report n° 6, (1965).

[19] GABOR D. : Theory of information.
J. Inst. Elec. Engrs. pt III, vol. 93, pp 429-457 (1946).

[20] GILLE, DECAULNE, PELEGRIN : Méthodes d'étude des systèmes asservis non linéaires. Dunod (1967).

[21] GOLD and RADER : Digital Processing of Signals.
Lincoln Laboratory Publications. Mac-Graw Hill (1969).

[22] GOODMAN : Determination of the number of terms necessary for a class of approximation procedures.
in "Data Representation".
University of Queensland Press (Anot) (1970).

[23] HARDY G, Mc WRIGHT E.M. : An introduction to the theory of numbers.
Clarendon Press (1960).

[24] HILDEBRAND F.B. : Introduction to numerical analysis.
Mac Graw Hill (1956).

[25] HOLTZ, LEONDES : The synthesis of recursive digital filters.
Journal of ACM. Vol. 13, n° 2, (1966) pp. 262-280.

[26] HUARD : Tour d'horizon en programmation non linéaire.
Bulletin E.D.F., série n° 1 (1971).

[27] JAEGER D. : Un périphérique d'ordinateur à l'usage des musiciens.
Thèse troisième cycle, Grenoble (1974).Univ. Scient. et Méd.

[28] JURY E.I. : Theory and application of the Z-transform method.
Wiley (1964).

[29] KAISER : Digital Filters.
Systems Analysis by Digital Computer Wiley (1967).

[30] LAGASSE J : Etude des circuits électriques.
Tome I. Eyrolles (1963).

[31] LAGASSE J : Etude des circuits électriques.
Tome II. Eyrolles (1963).

[32] LAURENT P.J. : Etude de procédés d'extrapolation en analyse.
Thèse, Grenoble (1965).

[33] LAURENT P.J. : Approximation et Optimisation.
Hermann (1972).

[34] LAMPRECHT G. : Zur Mehrdentigkeit bei der Approximation in der Lp norm. Computing J., 349-355 (1970).

[35] LAMPRECHT G. : Abhängigkeit der Fehlerfunktion bei der rationalen Lp-Approximation.
Num. Math. 15, 392-403 (1970).

[36] LERNER R.M. : Representation of Signals in "Lectures on Communication Theory". Baghdady Editor.

[37] MARTIN H. : Les mathématiques et la musique dans les grands courants de la pensée mathématique. Blackard Editor (1962).

[38] MAX J. : Traitement du Signal.
Tome I, Masson (1972).

[39] McSHANE E.J. : Integration.
Princeton (1944).

[40] MEINARDUS G. : Approximation of functions : theory and numerical methods.
Springer-Verlag, Berlin-Heidelberg, New-York (1967).

[41] MONROE A.J. : Digital processes for sampled data systems.
John Wiley and Sons Inc. New-York (1962).

[42] OLSON H.F. : Music, Physics and engineering.
Dover Publication (1967).

[43] PELLANDINI : Synthèse des filtres digitaux avec contre réaction dans le domaine des fréquences.
Revue de l'AGEN (Juillet 1969) pp. 30-40.

[44] POMENTALE T. : On discrete rational least squares approximation.
Num. Math. 12, 40-46 (1968).

[45] POLAK E. : Computational methods in optimization.
Academic Press (1971).

[46] RABINER : Techniques for designing finite derivation impulse reponse digital filters.
IEEE Trans. Commun. Techniol. Vol. CoM. 19 (April 1971).

[47] RADER C., GOLD B. : Digital filter design techniques in the frequency domain.
Proceedings of the IEEE Vol. 55 n° 2, (1967).

[48] RADIX J.C. : Introduction au filtrage numérique.
Eyrolles (1971).

[49] RAYLEIGH, LORD : Theory of sound.
Mac Millan (1877) London.

[50] RICE J.R. : The approximation of functions.
Vol. II, Addison-Wesley, Pub. Comp. (1969).

[51] ROCKAFELLAR R.T. : Convex analysis.
Princeton (1970).

[52] SCHAEFFER : Traité des objets musicaux (essai interdisciplines)
Edit. du Seuil (1966).

[53] SCIENTIFIC SUBROUTINE PACKAGE (IBM 360).

[54] SHAPIRO H.S. : Topics in Approximation Theory.
187 Springer-Verlag, Berlin-Heidelberg, New-York (1971).

[55] STEFFENSEN J.F. : Interpolation.
Chelsea Publishing Co (1965).

[56] STEIGLITZ : Computer aided design of recursive digital filter.
IEEE Trans. Audio-Electroacoust. vol. Au 18 (Juillet 1970).

[57] STORER : Passive network synthesis.
Chap. 14, Mac Graw Hill (1957).

[58] STRONG W., CLARK U. : Synthesis of wind instrument tones.
J. Acoust. Soc. Am. 41, 39-52, (1967).

[59] SZENTIRMAI (edited by) : Computer-aided filter design.
IEEE Press (1973).

[60] TRICOMI : Integral Equations.
Pure and Applied Mathematic,vol V, Interscience (1957).

[61] VALIRON : Theorie des fonctions.
Masson et Cie (1948).

[62] VIGOUROUX, WOLF : Filtrage numérique.
Rapport de stage Institut de Programmation, Grenoble (1972

[63] VILLE : Théorie et applications de la notion de signal analytique.
C. et T. deuxième année, n° 1 (1948).

[64] WALL H : Analytic theory of continued fractions.
Van Nostrand Comp. Inc. (1948).

[65] WALSH J.C. : Interpolation and approximation.
American Math. Soc. Colloq. Publ. Vol. XX (1960).

[66] WINCKEL : Vues Nouvellessur le monde des sons.
Dunod (1960).

[67] WOLF J. : Calcul de la période d'une fonction périodique, par une méthode statistique.
Séminaire IMAG (1969).

[68] WOLF J. : Construction de filtres digitaux par la résolution d'équations différentielles.
R.A.I.R.O., R1, (1972).

[69] WOLF J. : Etude de quelques signaux transitoires.
R.A.I.R.O. (avril 1973) R.1, 101-106.

[70] WOLF J. : Minimisation d'une forme quadratique avec une infinité de contraintes.
Colloque C.N.R.S. La Colle Sur Loup (1973).

[71] WOLF J. : Approximation par des fractions rationnelles en norme L^2 .
Colloque sur la Théorie constructive des fonctions
CLUJ (1973) (à paraître).

[72] WOLF J. : Approximation d'un élément de L^2 pour une fraction rationnelle.
Séminaire I.M.A. Grenoble (1974).

[73] WOLF J. : Approximation d'un élément de L^2, par une fraction rationnelle généralisée, en norme L^2 , et régularisation d'un ensemble d'approximation.
C.R. Acad. Sc. Paris t. 278, (22 avril 1974) Série A.

TABLE DES MATIERES