Méthodes numériques de recherche de la meilleure approximation

Jean-Louis Ville

N° d'ordre

T H E S E

présentée à la FACULTE DES

SCIENCES de l'université de GRENOBLE

pour obtenir

le titre de DOCTEUR DE TROISIEME CYCLE

"MATHEMATIQUES APPLIQUEES"

par

Jean - Louis V I L L E

METHODES NUMERIQUES DE RECHERCHE DE LA MEILLEURE APPROXIMATION

Thèse soutenue le devant la commission d'examen

Président : M. KUNTZMANN

Examinateurs: MM . GASTINEL
HACQUES
LAURENT

N° d'ordre

T H E S E

présentée à la FACULTE DES

SCIENCES de l'université de GRENOBLE

pour obtenir

le titre de DOCTEUR DE TROISIEME CYCLE

"MATHEMATIQUES APPLIQUEES"

par

Jean - Louis V I L L E

METHODES NUMERIQUES DE RECHERCHE DE LA MEILLEURE APPROXIMATION

Thèse soutenue le devant la commission d'examen

Président : M. KUNTZMANN

Examinateurs: MM . GASTINEL
HACQUES
LAURENT

N° d'ordre

T H E S E

présentée à la FACULTE DES

SCIENCES de l'université de GRENOBLE

pour obtenir

le titre de DOCTEUR DE TROISIEME CYCLE

"MATHEMATIQUES APPLIQUEES"

par

Jean - Louis V I L L E

METHODES NUMERIQUES DE RECHERCHE DE LA MEILLEURE APPROXIMATION

Thèse soutenue le devant la commission d'examen

Président : M. KUNTZMANN

Examinateurs: MM . GASTINEL
HACQUES
LAURENT

FACULTE DES SCIENCES

L I S T E D E S P R O F E S S E U R S

DOYENS HONORAIRES — M. FORTRAT P.

M. MORET L.

DOYEN — M. WEIL L.

PROFESSEURS TITULAIRES

MM.	NEEL L.	MAGNETISME ET PHYSIQUE DU SOLIDE
	DORIER A.	ZOOLOGIE
	HEILMANN R.	CHIMIE ORGANIQUE
	KRAVTCHENKO J.	MECANIQUE RATIONNELLE
	CHABAUTY C.	CALCUL DIFFERENTIEL ET INTEGRAL
	PARDE M.	POTAMOLOGIE
	BENOIT J.	RADIOELECTRICITE
	CHENE M.	CHIMIE PAPETIERE
	BESSON J.	ELECTROCHIMIE
	WEIL L.	THERMODYNAMIQUE
	FELICI N.	ELECTROSTATIQUE
	KUNTZMANN J.	MATHEMATIQUES APPLIQUEES
	BARBIER R.	GEOLOGIE APPLIQUEE
	SANTON L.	MECANIQUE DES FLUIDES
	OZENDA P.	BOTANIQUE
	FALLOT M.	PHYSIQUE INDUSTRIELLE
	GALVANI O.	MATHEMATIQUES
	MOUSSA A.	CHIMIE NUCLEAIRE
	TRAYNARD P.	CHIMIE
	SOUTIF M.	PHYSIQUE
	CRAYA A.	HYDRODYNAMIQUE
	REULOS R.	THEORIE DES CHAMPS
	AYANT Y.	PHYSIQUE APPROFONDIE
	GALLISSOT F.	MATHEMATIQUES APPLIQUEES
Melle	LUTZ E.	MATHEMATIQUES
MM.	BLAMBERT M.	MATHEMATIQUES
	BOUCHEZ R.	PHYSIQUE NUCLEAIRE
	ILLIBOUTRY L.	GEOPHYSIQUE
	MICHEL R.	GEOLOGIE ET MINERALOGIE
	BONNIER E.	ELECTROCHIMIE
	DESSAUX G.	PHYSIQUE ANIMALE
	PILLET E.	ELECTROCHIMIE
	DEBELMAS J.	GEOLOGIE
	GERBER R.	MATHEMATIQUES
	PAUTHENET R.	ELECTROTECHNIQUE
	VAUQUOIS B.	MATHEMATIQUES APPLIQUEES
	SILBER R.	MECANIQUE DES FLUIDES
	MOUSSIEGT J.	ELECTRONIQUE
	BARBIER J. C.	PHYSIQUE
	KPSZUL J. L.	MATHEMATIQUES

PROFESSEURS SANS CHAIRE

M.	LACASE A.	THERMODYNAMIQUE
Mme	KOFLER L.	BOTANIQUE
MM.	DREYFUS B.	THERMODYNAMIQUE
	VAILLANT F.	ZOOLOGIE ET HYDROBIOLOGIE
	GIRAUD P.	GEOLOGIE
	GIDON P.	GEOLOGIE ET MINERALOGIE
	ARNAUD P.	CHIMIE
	PERRET R.	SERVOMECANISMES
Mme	LUMER L.	MATHEMATIQUES
Mme	BARBIER M.J.	ELECTROCHIMIE
Mme	SOUTIF J.	PHYSIQUE
MM.	BRISSONNEAU P.	PHYSIQUE
	COHEN J.	ELECTROCHIMIE
	DEPASSEL R.	MECANIQUE
	GASTINEL N.	MATHEMATIQUES APPLIQUEES

PROFESSEURS ASSOCIES

MM.	LUMER G.	MATHEMATIQUES
	HIGUCHI	BIOSYNTHESE DE LA CELLULOSE
	WAGNER	BOTANIQUE

MAITRES DE CONFERENCES

MM.	ROBERT A.	CHIMIE PAPETIERE
	ANGLES D'AURIAC	MECANIQUE DES FLUIDES
	BIAREZ J. P.	MECANIQUE PHYSIQUE
	COUMES A.	ELECTRONIQUE
	DODU J.	MECANIQUE DES FLUIDES
	DUCROS P.	MINERALOGIE ET CRISTALLOGRAPHIE
	CLENAT P.	CHIMIE
	HACQUES G.	CALCUL NUMERIQUE
	LANCIA R.	PHYSIQUE AUTOMATIQUE
	PEBAY-PEROULA	PHYSIQUE
	KAHANE	PHYSIQUE GENERALE
	DOLIQUE	ELECTRONIQUE
Mme	KAHANE J.	PHYSIQUE
MM.	DEGRANGE C.	ZOOLOGIE
	GAGNAIRE D.	CHIMIE PAPETIERE
	RASSAT A.	CHIMIE SYSTEMATIQUE
	KLEIN J.	MATHEMATIQUES
	BETHOUX P.	MATHEMATIQUES APPLIQUEES
	POULOUJADOFF M.	ELECTROTECHNIQUE
	DEPOMMIER P.	PHYSIQUE NUCLEAIRE
	DEPORTES C.	CHIMIE
	BARRA J.	MATHEMATIQUES APPLIQUEES
Mme	BOUCHE L.	MATHEMATIQUES
MM.	PERRIAUX J.	GEOLOGIE
	SARROT-REYNAULD	GEOLOGIE
	CAUQUIS G.	CHIMIE GENERALE
	LABBE A.	BOTANIQUE
	BONNET G.	PHYSIQUE GENERALE
	BARNOUD F.	BIOSYNTHESE DE LA CELLULOSE
Mme	BONNIER M. J.	CHIMIE

MAITRES DE CONFERENCES ASSOCIES

MM.	ISHIKAWA Y.	MAGNETISME
	QUATTROPANI	THERMODYNAMIQUE

Je remercie Monsieur le Professeur KUNTZMANN, Directeur de l'Institut de Mathématiques Appliquées de Grenoble, qui me fait l'honneur de présider le Jury de cette thèse.

J'adresse le témoignage de ma vive gratitude à Monsieur le Professeur GASTINEL qui a bien voulu diriger ce travail et m'aider de ses conseils et de ses encouragements.

Je remercie Monsieur LAURENT, Maître de Conférences, qui a accepté de s'intéresser à cette thèse et de me faire part de ses suggestions.

Je remercie Monsieur HACQUES, Maître de Conférences, de bien vouloir faire partie du jury de cette thèse.

Je remercie également tous les membres du Laboratoire de Calcul de l'Université de Grenoble, qui ont pris part à la réalisation de ce travail.

INTRODUCTION

Etant donnés une fonction continue f(t), définie sur un intervalle $[a,b]$ et un sous-espace vectoriel V de l'espace des fonctions continues sur l'intervalle $[a,b]$ si g (t) est un élément de V, on se propose d'étudier les méthodes numériques qui permettent de déterminer un élément $g^*(t) \in V$ (s'il existe) qui approche le mieux la fonction f(t), sur tout l'intervalle $[a,b]$, au sens de Tschebyscheff, c'est-à-dire tel que :

$$\max_{t \in [a,b]} |f(t) - g^*(t)| = \inf_{g \in V} \left(\max_{t \in [a,b]} |f(t) - g(t)| \right) = \rho_V(f).$$

On étudiera également le cas où l'on considère, non plus l'intervalle $[a,b]$, mais un certain nombre de points t_i (i = 1, ..., m) dans l'intervalle $[a,b]$. Le problème est alors de déterminer $g^*(t)$ (s'il existe) tel que :

$$\max_{i=1,..,m} |f(t_i) - g^*(t_i)| = \inf_{g \in V} \max_{i=1,..,m} |f(t_i) - g(t_i)|$$

Dans le premier chapitre, nous rappellerons les théorèmes fondamentaux de l'approximation au sens de Tschebyscheff, concernant l'existence, l'unicité, et la caractérisation de la meilleure approximation, dans l'intervalle des fonctions continues sur un intervalle $[a,b]$.

Nous étudierons dans le chapitre II une méthode dérivée du deuxième algorithme de Remez, méthode qui permet d'atteindre la meilleure approximation en faisant croître une borne inférieure de la déviation maximum $\rho_V(f)$.

Le chapitre III et le chapitre IV seront consacrés respectivement à la méthode de décomposition de la norme et au premier algorithme de Remez. Ces deux méthodes ont pour principes la correction de la fonction g(t) afin de faire décroître la quantité $\max_{t \in [a,b]} |f(t) - g(t)|$.

Dans les chapitres V et VI on étudiera les méthodes de recherche de la meilleure approximation sur un ensemble discret de points (Méthode de Stiefel et méthodes de programmation linéaire).

D'autre part, il semble utilse de citer les travaux de Mahely [6] et de Herz [4] . Les méthodes qui en découlent permettent d'atteindre la meilleure approximation en partant d'une fonction g(t) suffisament voisine de la solution.

Afin de pouvoir confronter l'efficacité des méthodes étudiées, nous avons choisi des exemples numériques simples qui sont traités successivement par chaque méthode. Toutes les expériences numériques ont été réalisées à l'aide de la machine à calculer IBM 7044, à l'Institut de Mathématiques Appliquées de Grenoble.

CHAPITRE I

THEOREMES GENERAUX SUR L'APPROXIMATION (1)

1) Théorème I Théorème d'existence d'une meilleure approximation

Soit E un espace vectoriel sur R (ou (C)), normé à l'aide d'une norme notée $\| \ \|$ et soient $f_1, f_2, \ldots, f_n$ n éléments de E qui engendrent une variété linéaire V finie. Pour $g \in V$ on pose :

$$g = \lambda_1 f_1 + \ldots + \lambda_n f_n$$

Etant donné $f \in E$, on pose :

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

On se propose alors de trouver, s'il existe, un élément $g^* \in V$ tel que

$$\| f-g^* \| = \rho$$

Soit $\varphi(g) = \| f-g \|$. $\varphi(g)$ est une fonction continue de $g \in V$

Montrons que le minimum, s'il existe, est atteint pour un g qui se trouve à l'intérieur d'une certaine boule compacte.

Soit g_o un élément de V

$$\mu = \| f-g_o \| \geqslant \rho$$

Alors pour tout $g \in V$ tel que $\| g - g_o \| > 2\mu$

on a :

$$\| f-g \| \geqslant \left| \| f-g_o \| - \| g_o-g \| \right| > \mu \geqslant \rho$$

Donc un élément g^* (s'il existe) tel que $\| f-g^* \| = \rho$ ne peut être qu'à l'intérieur de la boule fermée de centre g_o et de rayon 2μ

(1) Voir Achieser [1]

La fonction $\varphi(g)$ étant continue sur la boule compacte

$\|g-g_o\| \leqslant 2\mu$ $\varphi(g)$ atteint son minimum.

Alors il existe $g^* \in V$ tel que

$$\|f-g^*\| = \rho$$

$g^* = \lambda^*_1 f_1 + ... + \lambda^*_n f_n$ est appelée la meilleure approximation de f par des éléments de V.

2) Théorème II : Condition d'unicité de la meilleure approximation dans l'espace des fonctions continues, à valeurs réelles, définies sur un compact $\mathcal{M}$ de R^m

On considère l'espace vectoriel des fonctions continues, à valeurs réelles, définies sur un compact $\mathcal{M}$ de R^m.

Soient $f_1(P)$, ..., $f_n(P)$ $(P \in \mathcal{M})$ n fonctions de cet espace, linéairemen indépendantes.

On considère le polynôme généralisé

$$F(P;x) = x_1 f_1(P) + ... + x_n f_n(P)$$

On appelle

$$L(x) = L(x;f) = \max_{P \in \mathcal{M}} |f(P) - F(P;x)|$$

L'espace des fonctions continues, à valeurs réelles, définies sur $\mathcal{M}$ est un espace vectoriel normé avec la norme

$$\|f\| = \max_{P \in \mathcal{M}} |f(P)|$$

D'après le théorème I, il existe au moins un x* tel que :

$$L(x^*;f) = \inf_{x \in R^n} L(x;f)$$

Condition de Haar :(1)

Une condition nécessaire et suffisante pour que la solution $F(P,x^*)$ soit unique, est que tout polynôme généralisé $F(P,x)$ non identiquement nul n'ait pas plus de n-1 zéros dans $\mathcal{M}$.

(1) voir ACHIESER [1].

On va rappeler les démonstrations dont les principes jouent un rôle dans les méthodes numériques.

Condition nécessaire :

On montre que si $F(P;\alpha)$ a n zéros dans $\mathcal{M}$, soit $P_1, \ldots, P_n$, alors on peut construire une fonction qui possède une infinité de meilleures approximations.

Si $P_1, \ldots P_n$ sont n zéros distincts de $F(P;\alpha)$

$$\alpha_1 f_1(P_1) + \ldots + \alpha_n f_n(P_1) = 0$$

$$\alpha_1 f_1(P_n) + \ldots + \alpha_n f_n(P_n) = 0$$

Alors

$$\det \begin{vmatrix} f_1(P_1) & \ldots & f_n(P_1) \\ & & \\ f_1(P_n) & \ldots & f_n(P_n) \end{vmatrix} = 0$$

De là, il existe n scalaires $c_1, \ldots, c_n$ tels que

$$c_1 f_k(P_1) + \ldots + c_n f_k(P_n) = 0 \quad \text{pour } k = 1, \ldots n .$$

En faisant une combinaison linéaire des équations précédentes, pour tout $x \in R^n$

$$c_1 F(P_1;x) + c_2 F(P_2;x) + \ldots + c_n F(P_n;x) = 0$$

Soit λ un scalaire tel que

$$\max_{P \in \mathcal{M}} |\lambda F(P;\alpha)| < 1$$

On considère alors g (P) une fonction continue sur $\mathcal{M}$ telle que

$$|g(P)| \leqslant 1 \quad \text{pour } P \in \mathcal{M}$$

et $g(P_i) = \text{signe}(c_i)$ pour $c_i \neq 0$ $(i = 1, \ldots, n)$

Considérons alors la fonction

$$f(P) = g(P) \left[1 - |\lambda F(P;\alpha)| \right] .$$

Elle est continue sur $\mathcal{M}$

En raison du choix de λ, $|f(P)| \leqslant 1$ pour $P \in \mathcal{M}$

$$f(P_i) = g(P_i) = \text{signe}(C_i)$$

Si :

$$L(x,f) = \operatorname{Max}_{P \in \mathcal{M}} |f(P) - F(P,x)| \quad \text{alors}$$

$L(x,f) \geqslant 1$. En effet si $L(x,f) < 1$ alors pour tout $P \in \mathcal{M}$

$$|f(P) - F(P;x)| < 1 \quad \text{et en particulier}$$

$$|f(P_i) - F(P_i;x)| < 1 \quad i = 1, \ldots, n$$

comme $f(P_i) = \text{signe}(C_i)$, $C_i \neq 0$

$$|\text{signe}(C_i) - F(P_i,x)| < 1$$

alors signe $F(P_i;x) = \text{signe}(C_i) = \text{signe } f(P_i)$ pour $C_i \neq 0$

Or $C_1 F(P_1 ; x) + C_2 F(P_2 ; x) + \ldots + C_n F(P_n ; x) = 0$.

Donc $L(x, f) \geqslant 1$ et en particulier pour la meilleure approximation de f.

Soit ϵ un nombre tel que $|\epsilon| \leqslant 1$. Alors :

$$|f(P) - \epsilon\lambda F(P;\alpha)| \leqslant |f(P)| + |\epsilon \lambda F(P;\alpha)| =$$

$$= |g(P)| \left\{1 - |\lambda F(P;\alpha)|\right\} + |\epsilon \lambda F(P ; \alpha)| \leqslant$$

$$\leqslant 1 - |\lambda F(P;\alpha)| + |\epsilon \lambda F(P, \alpha)| =$$

$$= 1 - (1-|\epsilon|)(|\lambda F(P;\alpha)|) \leqslant 1$$

Pour tout ϵ avec $-1 \leqslant \epsilon \leqslant 1$, $\epsilon \lambda F(P ; \alpha)$ est un polynôme de meilleure approximation pour $f(P)$.

$f(P)$ a une infinité de meilleures approximations.

Condition suffisante :

Pour la démonstration de la condition suffisante, on rappelle les énoncés de deux lemmes (Achieser [1]).

Lemme 1 : si

$$(1) \qquad \begin{vmatrix} f_i(P_i) & f_{i+1}(P_i) & \cdots\cdots & f_k(P_i) \\ & & & \\ f_i(P_k) & f_{i+1}(P_k) & \cdots\cdots & f_k(P_k) \end{vmatrix} \neq 0 \qquad (1 \leqslant i < k < n)$$

Alors à tout entier q ($k < q \leqslant n$) on peut faire correspondre des points $P_{k+1}, \ldots, P_q$ tels que :

$$\begin{vmatrix} f_i(P_i) & f_{i+1}(P_i) & \cdots\cdots & f_q(P_i) \\ & & & \\ f_i(P_q) & f_{i+1}(P_q) & \cdots\cdots & f_q(P_q) \end{vmatrix} \neq 0$$

Lemme 2 :

Si les points $P_1, P_2, \ldots, P_k$ ($k < n$) sont tous distincts, alors un au moins des déterminants d'ordre k de la matrice

$$\begin{vmatrix} f_1(P_1) & \cdots\cdots & f_n(P_1) \\ & & \\ f_1(P_k) & \cdots\cdots & f_n(P_k) \end{vmatrix}$$

est différent de zéro.

Lemme 3 :

Si le nombre de points de $\mathcal{M}$ où

$$|f(P) - F(P,x)| = L(x) = L(x ; f)$$

est inférieur à n, alors F (P ; x) n'est pas la meilleure approximation de f (P).

Soient $P_1, \ldots, P_m$ ($m < n$) m points dans $\mathcal{M}$ pour lesquels on a

$$|f(P) - F(P, x)| = L(x)$$

En raison du lemme 2 le système

$$f_1(P_k)\,\xi_1 + f_2(P_k)\,\xi_2 + \ldots f_n(P_k)\,\xi_n = f(P_k) - F(P_k, x)$$
$$(k = 1, 2, \ldots, m)$$

peut être résolu en $\xi_1 \ldots \xi_n$.

Soit $R(P) = f(P) - F(P,x)$

On choisit alors au voisinage de chacun des points $P_k (k = 1,2,\ldots,m)$ un ensemble $\mathcal{M}_k$ tel que

$$\mu_k = \min_{P \in \mathcal{M}_k} |R(P)| > 0$$

et

$$\min_{P \in \mathcal{M}_k} |F(P, \xi)| \geqslant \frac{L(x)}{2}$$

De plus soit $M_k = \max_{P \in \mathcal{M}_k} |F(P ; \xi)|$

$$M = \max_{P \in \mathcal{M}^*} F(P;\xi) \quad , \quad L^*(x) = \max_{P \in \mathcal{M}^*} |R(P)|$$

avec $\mathcal{M}^* = \mathcal{M} - \mathcal{M}_1 - \mathcal{M}_2 - \ldots\ldots - \mathcal{M}_m$

Alors $\mu = L(x) - L^*(x) > 0$.

Soit ϵ un nombre tel que

$$0 < \epsilon < \min \left\{ \frac{\mu}{M} , \frac{\mu_1}{M_1} , \ldots, \frac{\mu_m}{M_m} \right\}$$

On pose $x'_i = x_i + \epsilon \xi_i \qquad (i = 1, \ldots, n)$

Alors :

$$|f(P) - F(P ; x')| = |f(P) - F(P;x) - \epsilon F(P;\xi)|$$

$$= |R(P) - \epsilon F(P;\xi)|$$

$$|f(P) - F(P ; x')| \leqslant |R(P)| \left\{ 1 - \epsilon \frac{F(P;\xi)}{R(P)} \right\}$$

$$\leqslant L(x) \left\{ 1 - \frac{\epsilon}{2} \right\}$$

pour $P \in \mathcal{M}_k \qquad (k = 1,2, \ldots, m)$

et

$$|f(P) - F(P ; x')| \leqslant |R(P)| + \epsilon |F(P ; \xi)|$$

$$\leqslant L^*(x) + \epsilon M < L(x)$$

pour $P \in \mathcal{M}^*$.

$$\text{Alors } L(x') = \max_{P \in \mathcal{M}} |f(P) - F(P ; x')| < L(x)$$

ce qui démontre le lemme 3. A partir de ce lemme on pourrait développer une méthode numérique pour se rapprocher de la meilleure approximation tant que le nombre des points de $\mathcal{M}$ tels que $|f(P) - F(P ; x)| = L(x)$ est inférieur à n.

Montrons alors que la condition de Haar est suffisante. Supposons qu'il existe deux meilleures approximations $F(P ; x)$, $F(P ; y)$

Alors $F(P ; \frac{x+y}{2})$ est aussi une meilleure approximation puisque :

$$|F(P ; \tfrac{x+y}{2}) - f(P)| \leqslant \tfrac{1}{2} |F(P ; x) - f(P)| + \tfrac{1}{2} |F(P;y) - f(P)|$$

D'après le lemme 3, si on pose :

$$L = L(\tfrac{x+y}{2}) = L(x) = L(y)$$

alors l'équation

$$|f(P) - F(P ; \tfrac{x+y}{2})| = L \text{ possède au moins n solutions}$$

$P_1, \ldots, P_n$ dans $\mathcal{M}$.

Pour que $|f(P_i) - F(P_i ; \frac{x+y}{2})| = L$

il faut que

$$f(P_i) - F(P_i ; x) = f(P_i) - F(P_i ; y) = L$$

puisque :

$$L = |f(P_i) - F(P_i ; \tfrac{x+y}{2})| = |\tfrac{1}{2} f(P_i) - F(P_i ; \tfrac{x}{2}) + \tfrac{1}{2} f(P_i) - F(P_i ; \tfrac{y}{2})|$$

$$\leqslant \tfrac{1}{2} |f(P_i) - F(P_i ; x)| + \tfrac{1}{2} |f(P_i) - F(P_i ; y)|$$

et que

$$|f(P_i) - F(P_i ; x)| \leqslant L \quad |f(P_i) - F(P_i ; y)| \leqslant L$$

Alors le polynôme $F(P ; x-y)$ non identiquement nul à n zéros distincts dans $\mathcal{M}$ ce qui est en contradiction avec la condition de Haar.

3) Système de fonctions de Tschebyscheff

Définition

On appelle système de fonctions de Tschebyscheff relatif à un intervalle borné fermé $[a,b]$ de la droite réelle un ensemble de n fonctions continues, à valeurs réelles, $f_1(x)$, ..., $f_n(x)$ définies sur $[a,b]$ et qui satisfait la condition de Haar.

La condition de Haar est équivalente à la propriété "d'interpolation" :
Si l'on se donne n abscisses x_1, ..., x_n dans $[a,b]$ et les valeurs correspondantes de la fonction f :
$f(x_1)$ ....., $f(x_n)$ il existe un polynôme et un seul $\lambda_1 f_1(x) + \ldots + \lambda_n f_n(x)$ tel que pour i = 1, ...,n.

$$\lambda_1 f_j(x_i) + \ldots + \lambda_n f_n(x_i) = f(x_i)$$

On considère l'espace des fonctions continues à valeurs réelles, définies sur $[a,b]$, muni de la norme

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

La propriété suivante sera utile au paragraphe 5, pour la démonstration du théorème de Tschebyscheff.

Lemme : Si $x_1, x_2, \ldots, x_{n-1}$ sont n - 1 points distincts de l'intervalle $[a,b]$ il existe un polynôme (et un seul à un facteur près) non identiquement nul

$$F(x, \lambda) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \ldots + \lambda_n f_n(x)$$

dont les zéros sont $x_1, \ldots, x_{n-1}$

Si x_k est à l'intérieur de l'intervalle $[a,b]$, le polynôme change de signe en x_k.

Le polynôme

$$D(x;x_1,x_2,..,x_{n-1}) = \begin{vmatrix} f_1(x) & f_2(x) & \ldots & f_n(x) \\ f_1(x_1) & f_2(x_1) & \ldots & f_n(x_1) \\ & & & \\ f_1(x_{n-1}) & f_2(x_{n-1}) & \ldots & f_n(x_{n-1}) \end{vmatrix}$$

qui est non identiquement nul (théorème II, lemme 2) s'annule aux points x_1, x_2,..,x_{n-1} et n'a pas d'autres zéros, puisque f_1, ..., f_n forment un système de Tschebyscheff.

$$F(x ; \lambda) \equiv CD (x ; x_1, x_2, \ldots, x_{n-1})$$

C étant un facteur constant.

Si F $(x ; \alpha)$ et F $(x ; \beta)$ non proportionnels admettent x_1, ..., x_{n-1} comme zéros alors on peut trouver λ et μ tels que :

$$\lambda F(x; \alpha) + \mu F(x ; \beta) \text{ s'annule en } x_n \in [a,b] \quad x_n \neq x_1, \ldots, x_{n-1}.$$

Alors le polynôme $\lambda F (x;\alpha) + \mu F (x ; \beta)$ admettrait n zéros, ce qui est en contradiction avec l'hypothèse que les f_1, ..., f_n forment un système de Tschebyscheff. $D (x ; x_1, x_2, \ldots, x_{n-1}) \neq 0$ toutes les fois que x n'est pas un point x_1, ..., x_{n-1}.

Alors $D (x ; x_1, \ldots, x_{n-1})$ change de signe en x_k si x_k est à l'intérieur du segment $[a,b]$, et en aucun autre point.

4) Généralisation du théorème de De La Vallée Poussin

Soit $f_1(x)$, ..., $f_n(x)$ un système de Tschebyscheff relatif à l'intervalle $[a,b]$ $f(x) \in \mathcal{C}[a,b]$ espace des fonctions continues à valeurs réelles définies sur l'intervalle $[a,b]$

Si en n + 1 points $x_1 < x_2 < \ldots. < x_{n+1}$

$$f(x_i) - F (x_i, \alpha) = \pm (-1)^i \lambda_i \text{ pour } i = 1, \ldots, n+1 \text{ avec } \lambda_i > 0.$$

Alors

$$\rho = \inf_{\alpha \in R^n} [L (\alpha)] = \inf_{\alpha \in R^n} \left[\max_{x \in [a,b]} |f(x) - F (x, \alpha)| \right]$$

est tel que $\rho \geqslant \min [\lambda_1, \ldots, \lambda_{n+1}]$;

En raisonnant par l'absurde, supposons que F $(x ; \alpha^*)$ soit la meilleure approximation de f et que

$$\| F (x ; \alpha^*) - f \| < \min [\lambda_1, \ldots, \lambda_{n+1}]$$

Considérons :

$$d(x) = F(x;\alpha^*)-F(x;\alpha) = F(x;\alpha^*)-f - (F(x;\alpha) - f)$$

$d(x_i)$ est la différence entre $(-1)^i \lambda_i$ et un nombre plus petit en valeur absolue.

Alors les $d(x_i)$ ont des signes alternés pour $i = 1, \ldots, n+1$ et le polynôme $F(x ; \alpha^* - \alpha) = d(x)$ aurait n zéros dans $[a,b]$ ce qui est contraire à la condition de Haar.

5) Théorème III :

Théorème de Tschebyscheff pour la caractérisation de la meilleure approximation de $\mathcal{C}[a,b]$.

Si (S) est un système de Tschebyscheff relatif à l'intervalle $[a,b]$ et si $f(x) \in \mathcal{C}[a,b]$ alors la meilleure approximation de $f(x)$ dans $[a,b]$ est caractérisée par le fait que :

$|f(x) - F(x ; \alpha)|$ atteint sa valeur maximale en au moins $n + 1$ points de l'intervalle $[a,b]$, soient $x_1, \ldots, x_{n+1}$ et que tout $i = 1, \ldots, n$

$$f(x_{i+1}) - F(x_{i+1} ; \alpha) = - (f(x_i) - F(x_i ; \alpha))$$

-Condition nécessaire

Supposons que $f(x) - F(x ; \alpha)$ prenne la valeur $L = \max_{a \leq x \leq b} |f(x) - F(x ; \alpha)|$ avec des signes alternés en au plus q points consécutifs $q \leq n$:

$$y_1 < y_2 < \ldots < y_q \qquad \text{dans } [a,b]$$

Alors $[a,b]$ peut être divisé en q intervalles

$$(1)\ [a,x_1]\ ,[x_1, x_2]\ , \ldots\ldots,[x_{q-1}, b]$$

tels que $q \leq y_1 < x_1 < \ldots < y_{q-1} < x_{q-1} < y_q < b$

et tels que alternativement dans les intervalles (1) une des deux inégalités suivantes soit vérifiée avec :

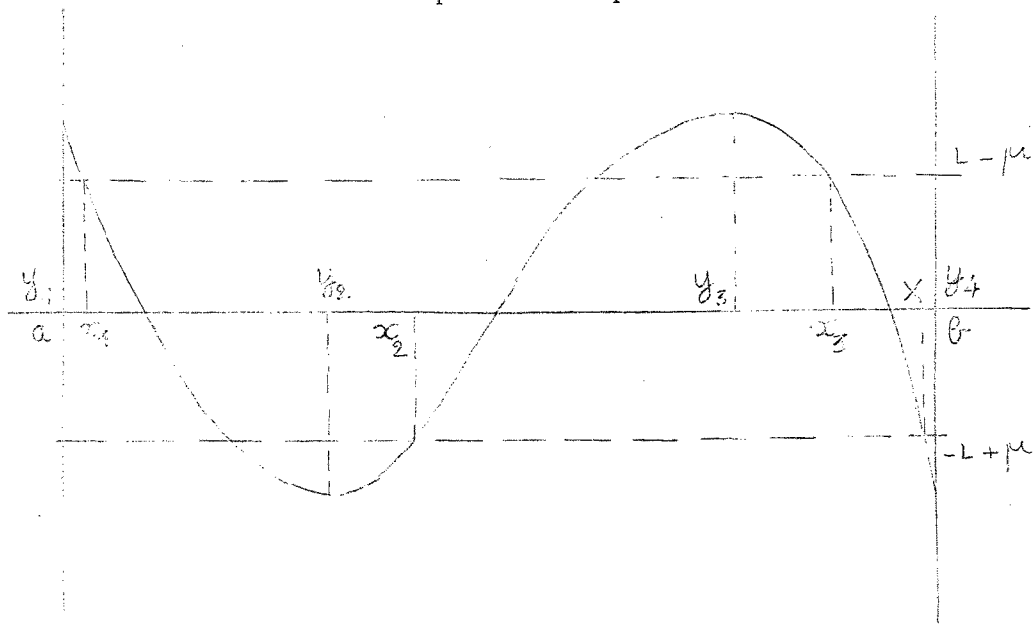
$$\mu < \frac{L}{2} \qquad \mu > 0 :$$

$$-L \leq f(x) - F(x ; \alpha) < L - \mu$$

ou

$$-L + \mu < f(x) - F(x ; \alpha) \leq L$$

De plus, on suppose $f(x_{q-1}) = F(x_{q-1} ; \alpha)$



On choisit X entre x_{q-1} et y_q de façon telle que dans l'intervalle $[x_{q-1}, X]$

$$- L + \mu < f(x) - F(x ; \alpha) < L - \mu$$

et dans $[x_{q-1}, X]$ on choisit

$$x_q < x_{q+1} < \ldots < x_{q+2m-1}$$

avec m = partie entière $\left[\frac{n-q}{2} \right]$

q+2 m-1 est égal soit à n - 1 soit à n - 2

On construit :

$$F(x;\beta) = F(x;\alpha) + \epsilon\, D(x;x_1, \ldots, x_n)$$

avec :

$$D(x;x_1,\ldots,x_n) = \begin{vmatrix} f_1(x) & f_2(x) & \ldots & f_n(x) \\ f_1(x_1) & f_2(x_1) & \ldots & f_n(x_1) \\ & & & \\ f_1(x_{n-1}) & \ldots\ldots & & f_n(x_{n-1}) \end{vmatrix}$$

avec $x_{n-1} = b$ si $q + 2m - 1 = n-2$

On choisit ϵ tel que

$$\max_{a \leqslant x \leqslant b} |\epsilon\, D(x;x_1,\ldots,x_n)| < \mu$$

ϵ peut être choisi tel que

$$\text{signe}\left\{f(y_1) - F(y_1;\alpha)\right\} = \text{signe}\left\{\epsilon\, D(x;x_1,\ldots,x_{n-1})\right\}$$

dans l'intervalle $[a,x_1]$

D'après le lemme précédent

$$\text{signe}\left\{\epsilon\, D(x;x_1,\ldots,x_n)\right\} = \text{signe}\left\{f(y_k) - F(y_k;\alpha)\right\}$$

avec $x_{k-1} \leqslant x \leqslant x_k$ pour $k = 1,2,\ldots q-1$ et $x_o = a$.

Alors $|f(x) - F(x\,;\beta)| < L$ dans $[a,\ x_{q-1}]$

Si $q + 2m - 1 = n - 1$ cette inégalité est encore valable dans $[x_{q-1},\ b]$

Si $q + 2m - 1 = n - 2$ alors pour $a \leqslant x < b$ nous avons $|f(x) - F(x\,;\beta)| < L$.

On sait que $|f(b) - F(b\,;\beta)| \leqslant L$

Si l'inégalité est stricte alors $F(x\,;\beta)$ est une meilleure approximation que $F(x\,;\alpha)$.

Si $|f(b) - F(b\,;\beta| = L$ alors on choisit $F(x\,;\gamma)$ tel que :

$$F(b\,;\gamma)\,[f(b) - F(b\,;\beta)] \geqslant 0 \text{ de façon telle que pour } \delta > 0$$

$$|f(x) - F(x\,;\beta) - \delta\, F(x\,;\gamma)| < L \quad \text{pour } a \leqslant x \leqslant b.$$

Condition suffisante.

Si $f(x) - F(x\,;\alpha)$ prend la valeur

$$L = \max_{a \leqslant x \leqslant b} |f(x) - F(x\,,\alpha)| \text{ en } n+1 \text{ points avec des signes alternés,}$$

d'après la généralisation du théorème de De La Vallée Poussin

$$L \geqslant \rho \geqslant L$$

donc $\rho = L$ et $F(x\,;\alpha)$ est la meilleure approximation.

CHAPITRE II

METHODE RELATIVE AU DEUXIEME ALGORITHME DE REMEZ (1)

1) Généralités

Soient E un espace vectoriel normé muni d'une norme notée $\| \ \|$ et V un sous-espace de E de dimension finie n.

Soit f un élément de E qui n'appartient pas à V. On pose :

$$\rho_V (f) = \inf_{g \in V} \| f - g \|$$

D'après le théorème d'existence, il existe $g^* \in V$ tel que

$$\| f - g^* \| = \rho_V (f)$$

Propriété :

Si l'on a une fonctionnelle linéaire continue de $E \longrightarrow R$ telle que

1er) $\| L \| \leqslant 1$ avec $\| L \| = \sup_{f \in E} \frac{|L (f)|}{\| f \|}$

2em) $L(g) = 0$ pour tout $g \in V$

alors $|L (f)| \leqslant \rho_V (f)$

en effet :

$$\frac{|L (f)|}{\|f - g^*\|} = \frac{|L (f-g^*)|}{\|f - g^* \|} \leqslant 1$$

d'où :

$$|L (f)| \leqslant \|f - g^*\| = \rho_V (f)$$

(1) Voir MEINARDUS [7] , REMEZ [9] , FRASER AND HART [3]

Théorème

Il existe une fonctionnelle linéaire continue L de E dans R ayant les 3 propriétés suivantes :

1) $\|L\| = 1$

2) $L(g) = 0$ pour tout $g \in V$

3) $L(f) = \rho_V(f)$

Considérons l'espace vectoriel V* à n + 1 dimensions engendré par V et f. Tout élément h de V* peut s'écrire d'une façon unique, sous la forme

$$h = \beta f + g \qquad \text{avec } g \in V$$

Soit $\tilde{L}(h)$ la fonctionnelle linéaire continue de V* dans R telle que $\tilde{L}(h) = \beta \rho_V(f)$, elle a les propriétés :

1) $\tilde{L}(f) = \rho_V(f)$

2) $\tilde{L}(g) = 0$ pour tout $g \in V$

3) $\|\tilde{L}\|^* = 1$ si l'on pose

$$\|\tilde{L}\|^* = \sup_{\substack{h \in V^* \\ h \neq o}} \frac{|\tilde{L}(h)|}{\|h\|}$$

en effet pour $\beta \neq 0$

$$\|\beta f + g\| = |\beta| \, \|f + \frac{g}{\beta}\| \geqslant |\beta| \, \rho_V(f) = |\tilde{L}(\beta f + g)|$$

et donc $\|\tilde{L}\|^* \leqslant 1$

Pour $h = f - g^*$

$$\tilde{L}(f - g^*) = \tilde{L}(f) = \rho_V(f) = \|f - g^*\|$$

donc $\|\tilde{L}\|^* = 1$.

D'après le théorème de Hahn Banach, il existe une fonctionnelle linéaire L, de E dans R, continue, telle que

$$\|L\| = \|\tilde{L}\|^*$$

et pour tout $h \in V^*$ $\quad L(h) = \tilde{L}(h)$.

Remarque :

Si l'on a une fonctionnelle linéaire L vérifiant les conditions 2) et 3) du théorème et telle que $\|L\| \leqslant 1$, alors, on a l'égalité $\|L\| = 1$.

En effet : Supposons que l'on ait $\|L\| < 1$ alors, il existe $\lambda > 1$ tel que $L^* = \lambda L$ soit tel que $\|L^*\| \leqslant 1$. Du fait de la propriété énoncée avant le théorème, on aurait :

$$|L^*(f)| \leqslant \rho_v(f)$$

Or comme $L(f) = \rho_v(f)$, $L^*(f) = \lambda \times \rho_v(f) > \rho_v(f)$

ce qui est en contradiction avec la propriété.

2) Application à l'approximation des fonctions continues sur un intervalle [a,b]

Considérons pour E l'espace vectoriel normé $\mathcal{C}[a,b]$ des fonctions continues, à valeurs réelles, définies sur un intervalle fini $[a,b]$, muni de la norme :

$$\|f\| = \max_{t \in [a,b]} |f(t)|$$

Soient $f_\nu(t)$ $(\nu = 1, \ldots, n)$ n fonctions appartenant à $\mathcal{C}[a,b]$, formant un système de Tschebyscheff (possèdant la propriété d'interpolation) et qui engendrent un sous espace $V \subset \mathcal{C}[a,b]$.

Soit $g^*(t)$ la meilleure approximation de f par des éléments de V :

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

On sait d'après le théorème de caractérisation de la meilleure approximation (chapitre I, théorème III) que si $g^*(t)$ est la meilleure approximation alors il existe au moins n + 1 points extrémaux alternés $\xi_\mu \in [a,b]$ $(\mu = 1, \ldots, n+1)$ tels que :

$$f(\xi_\mu) - g^*(\xi_\mu) = -\left[f(\xi_{\mu+1}) - g^*(\xi_{\mu+1})\right] \quad (\mu = 1, \ldots, n)$$

$$\left|f(\xi_\mu) - g^*(\xi_\mu)\right| = \rho_v(f) \quad (\mu = 1, \ldots, n+1)$$

En se restreignant aux points ξ_μ on considère les vecteurs de R^{n+1}

$$\overline{f_1} \begin{vmatrix} f_1(\xi_1) \\ \vdots \\ f_1(\xi_{n+1}) \end{vmatrix} \quad \overline{f_2} \begin{vmatrix} f_2(\xi_1) \\ \vdots \\ f_2(\xi_{n+1}) \end{vmatrix} \quad \overline{f_n} \begin{vmatrix} f_n(\xi_1) \\ \vdots \\ f_n(\xi_{n+1}) \end{vmatrix} \quad \text{et } \overline{f} \begin{vmatrix} f(\xi_1) \\ \vdots \\ f(\xi_{n+1}) \end{vmatrix}$$

Puisque le sous espace V possède par hypothèse la propriété d'interpolation sur l'intervalle $[a,b]$, les vecteurs $\bar{f}_1, \ldots, \bar{f}_n \in R^{n+1}$ satisfont la condition nécessaire et suffisante pour que la meilleure approximation du vecteur $\bar{f}$ par des combinaisons linéaires des vecteurs $\bar{f}_1, \ldots, \bar{f}_n$ soit unique (voir chapitre V) avec la norme $\|\bar{f}\| = \underset{\mu=1,..,n+1}{\text{Max}} |f(\xi_\mu)|$).

Cette meilleure approximation correspond à la meilleure approximation de la fonction f (t) , par des éléments de V, sur l'ensemble discret des n + 1 points ξ_μ.

Montrons que si g* (t) est la meilleure approximation de f (t) sur l'intervalle $[a,b]$, alors g* (t) est aussi la meilleure approximation de f (t) sur l'ensemble discret des n + 1 points extrémaux alternés ξ_μ. (qui est unique).

En raisonnant par l'absurde si g* (t) n'est pas la meilleure approximation de f(t) sur l'ensemble des points ξ_μ, alors il existe $g_1(t) \in V$ tel que $\mu = 1, \ldots, n+1$

$$(1) \quad |f(\xi_\mu) - g_1(\xi_\mu)| < |f(\xi_\mu) - g^*(\xi_\mu)|$$

Si on pose $d(t) = (f(t) - g^*(t)) - (f(t) - g_1(t) = g_1(t) - g^*(t)$ alors $d(t) \in V$ et pour $\mu = 1, \ldots, n+1$

$$|d(\xi_\mu)| > 0 \text{ et signe } d(\xi_\mu) = \text{signe } (f(\xi_\mu) - g^*(\xi_\mu)).$$

d'après l'inégalité (1)).

Alors d(t) s'annule au moins n fois dans l'intervalle $[a,b]$, ce qui est en contradiction avec la condition de Haar sur l'intervalle $[a,b]$ (équivalente à la propriété d'interpolation sur l'intervalle $[a,b]$).

Sur R^{n+1} toutes les fonctionnelles linéaires continues sont de la forme :

$$L(y) = \sum_{\mu=1}^{n+1} \lambda_\mu y_\mu \qquad y \in R^{n+1}$$

Choisissons les λ_μ tels que :

1) $L(\bar{f}_\nu) = \sum_{\mu=1}^{n+1} \lambda_\mu f_\nu(\xi_\mu) = 0$ pour $\nu = 1, \ldots, n$

2) $\|L\| = 1$ (norme pour L fonctionnelle de $R^{n+1} \longrightarrow R$)

soit $\sum |\lambda_\mu| = 1$

Alors les λ_μ sont déterminés de façon unique au signe près et sont $\neq 0$.

La fonctionnelle L est unique au signe près.

En appliquant le théorème précédent (E étant R^{n+1} et V le sous espace de dimension n engendré par les $\bar{f}_\nu$). On sait qu'il existe une telle fonctionnelle (vérifiant les conditions 1 et 2) mais en plus telle que sa valeur pour $\bar{f}$ soit $\rho_v(f)$. C'est donc celle là au signe près. Alors :

$$L(\bar{f}) = \chi \rho_v(f). \text{ avec } \chi = +1 \text{ ou } -1.$$

Or :

$$\sum_{\mu=1}^{n+1} |\lambda_\mu| = 1 \qquad \text{et } |f(\xi_\mu) - g^*(\xi_\mu)| = \rho_v(f).$$

Donc :

$$L(\bar{f}) = \sum_{\mu=1}^{n+1} \lambda_\mu (f(\xi_\mu) - g^*(\xi_\mu)) = \rho_v(f) \sum_{\mu=1}^{n+1} \lambda_\mu \text{ signe } (f(\xi_\mu) - g^*(\xi_\mu))$$

$$= \chi \rho_v(f)$$

$$\sum_{\mu=1}^{n+1} \chi \lambda_\mu \text{ signe } (f(\xi_\mu) - g^*(\xi_\mu)) = 1$$

or

$$\sum_{\mu=1}^{n+1} |\lambda_\mu| = 1$$

en retranchant les 2 égalités, il vient :

$$\sum_{\mu=1}^{n+1} \left[\chi \text{ signe } (\lambda_\mu) \text{ signe } (f(\xi_\mu) - g^*(\xi_\mu)) - 1\right] |\lambda_\mu| = 0$$

Or χ signe (λ_μ) signe $(f(\xi_\mu) - g^*(\xi_\mu))$ est égal à -1 ou $+1$.

Pour que l'expression précédente soit nulle, il faut que :

$$\chi \text{ signe } (\lambda_\mu) \text{ signe } (f(\xi_\mu) - g^*(\xi_\mu)) = 1$$

Soit :

$$\text{signe } (\lambda_\mu) = \chi \text{ signe } (f(\xi_\mu) - g^*(\xi_\mu)).$$

comme $f(\xi_\mu) - g^*(\xi_\mu)$ alterne en signe avec μ alors :

$$\text{signe } \lambda_\mu = \chi' \times (-1)^\mu \qquad \text{avec } \chi' = +1 \text{ ou } -1$$

Or cette propriété n'est pas liée à un choix particulier des abscisses ξ_μ : en effet, si l'on prend $n+1$ abscisses t_μ, quelconques, en ordre croissant (distinctes) on pourra toujours trouver une fonction continue $f(t)$ et la meilleure approximation $g^*(t)$ sur l'intervalle $[a,b]$, tels que les points t_μ soient les points extrémaux de $f(t) - g^*(t)$.

Donc quels que soient les abscisses t_μ, les λ_μ associés vérifient :

$$\text{signe } \lambda_\mu = \chi \times (-1)^\mu \qquad (\mu = 1, \ldots, n+1).$$

$$\text{avec } \chi = +1 \text{ ou } -1$$

<u>1ère conséquence pour la meilleure approximation sur un ensemble discret</u> :

Etant donnés $n+1$ points t_μ ($\mu = 1, \ldots, n+1$) contenus dans un intervalle $[a,b]$ une fonction continue $f(t)$ sur cet intervalle et n fonctions $f_\nu(t)$ ($\nu = 1, \ldots, n$) possèdant la propriété de Haar sur l'intervalle $[a,b]$ soient :

$$\bar{f} \left| \begin{matrix} f(t_1) \\ \vdots \\ f(t_{n+1}) \end{matrix} \right. \qquad \text{le vecteur correspondant}$$

de R^{n+1} et $\bar{f}_\nu \left| \begin{matrix} f_\nu(t_1) \\ \vdots \\ f_\nu(t_{n+1}) \end{matrix} \right.$ qui engendrent un sous espace vectoriel V

de R^{n+1} alors la meilleure approximation de $\bar{f}$ par des éléments de V

(avec la norme $\|\bar{f}\| = \underset{\mu=1,\ldots,n+1}{\text{Max}} |f(t_\mu)|$) soit $\bar{g}^* \in R^{n+1}$ existe, est unique et possède la propriété que pour $\mu = 1,\ldots, n+1$ les $f(t_\mu) - g^*(t_\mu)$ sont alternés (et égaux en valeur absolue à la déviation maximum).

2ème Conséquence.

Etant donnés $n+1$ abscisses t_μ quelconques (on peut construire les λ_μ correspondants) on considère g appartenant au sous espace engendré par $f_1(t), \ldots, f_n(t)$ tel que :

$$g(t_\mu) + (-1)^\mu \lambda = f(t_\mu) \quad \text{pour } \mu = 1, \ldots, n+1$$

avec :

$$g(t) = \sum_{\nu=1}^{n} x_\nu f_\nu(t)$$

On a $n+1$ équation à $n+1$ inconnues : les x_ν ($\nu = 1, \ldots, n$) et λ.

Comme d'après la définition des λ_μ $\quad \sum_{\mu=1}^{n+1} \lambda_\mu g(t_\mu) = 0$

si $\quad \bar{f} \begin{vmatrix} f(t_1) \\ \vdots \\ f(t_{n+1}) \end{vmatrix} \in R^{n+1}$

alors :

$$L(\bar{f}) = \sum_{\mu=1}^{n+1} \lambda_\mu f(t_\mu) = \lambda \sum_{\mu=1}^{n+1} (-1)^\mu \lambda_\mu$$

Comme les λ_μ alternent en signe avec μ et que $\sum_{\mu=1}^{n+1} |\lambda_\mu| = 1$

alors

$$|L(\bar{f})| = |\lambda|$$

d'où la double inégalité :

$$|L(f)| = |\lambda| \leq \rho_v(f) \leq \|f-g\|$$

3) Méthode de construction de g*

Soit $M_o = \{t^o_\mu\}$ $\mu = 1, \ldots, n+1$ un ensemble de $n+1$ points

$$a \leqslant t^o_1 < t^o_2 < \ldots < t^o_{n+1} \leqslant b$$

Déterminons $g_o(t)$ tel que :

$$\text{(I)} \quad \begin{cases} g_o(t^o_\mu) + (-1)^\mu \lambda^o = f(t^o_\mu) \text{ pour } \mu = 1, \ldots, n+1 \\ \text{avec :} \\ g_o(t) = \sum_{\nu=1}^{n} x_\nu f_\nu(t) \end{cases}$$

où $f_\nu(t)$ $(\nu = 1, \ldots, n)$ est une base pour le sous espace vectoriel V.

C'est un système linéaire de $n+1$ équations à $n+1$ inconnues : les x_ν et λ.

On considère $L_o(\bar{f}) = \sum_{\mu=1}^{n+1} \lambda^o_\mu f(t^o_\mu)$ telle que :

$L_o(\bar{g}) = 0$ pour tout $\bar{g} \in V$ et $\|L\| = 1$.

Comme au paragraphe précédent les λ^o_μ sont déterminés à un signe près.

Alors :

$$|\lambda^o| = |L_o(\bar{f})| \leqslant \rho_\nu(f) \leqslant \|f - g_o\|$$

Nous cherchons alors un nouvel ensemble $M_1 = \{t^1_\mu\}$ tel que la fonctionnelle linéaire correspondante vérifie :

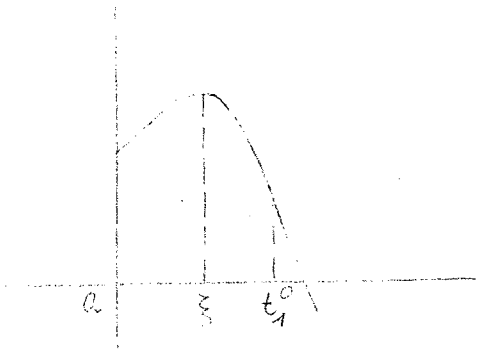
$$|L_o(\bar{f})| < |L_1(\bar{f})|$$

Considérons la fonction $\varphi_o(t) = f(t) - g_o(t)$ et soit ξ un point de $[a,b]$ tel que $|\varphi_o(\xi)| > |L_o(\bar{f})|$.

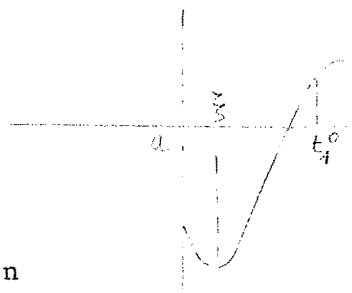
On pourra prendre par exemple ξ tel que $|\varphi_o(\xi)| = \underset{t \in [a,b]}{\text{Max}} |\varphi_o(t)|$.

Alors on peut échanger un point de M_o avec ξ de façon que $|L_1(\bar{f})| > |L_o(\bar{f})|$, à condition que M_o ne soit pas déjà l'ensemble des points extrémaux alternés et dans ce cas, il n'existe pas de ξ tel que $|\varphi_o(\xi)| > |L_o(\bar{f})|$

Si $a \leqslant \xi < t^o_1$ et signe $\varphi_o(\xi)$ = signe $\varphi_o(t^o_1)$ on prend ξ à la place de t^o_1.

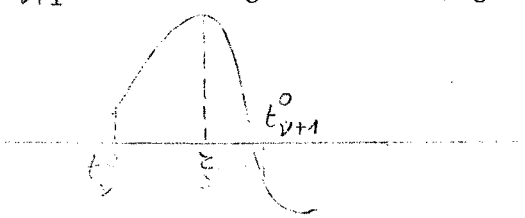


Si $a \leqslant \xi < t^o_1$ et signe $\varphi_o(\xi)$ = - signe $\varphi_o(t^o_1)$ on prend ξ à la place de t^o_{n+1}.

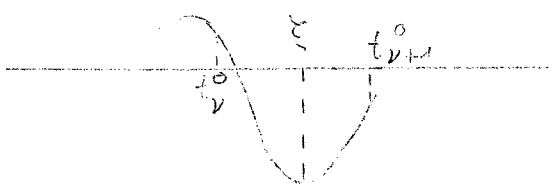


Pour $1 \leqslant \nu \leqslant n$

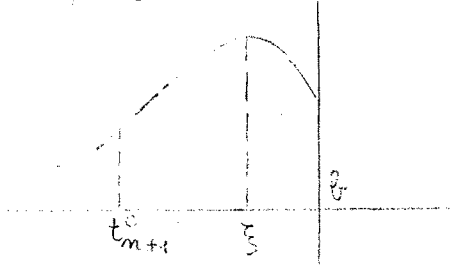
Si $t^o_\nu < \xi < t^o_{\nu+1}$ et signe $\varphi_o(\xi)$ = signe $\varphi_o(t^o_\nu)$ on prend ξ à la place de t^o_ν.



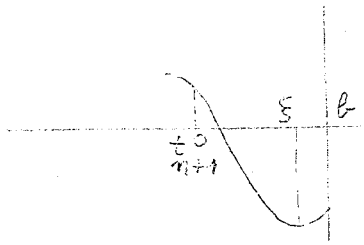
Si $t^o_\nu < \xi < t^o_{\nu+1}$ et signe $\varphi_o(\xi)$ = signe $\varphi_o(t^o_{\nu+1})$ on prend ξ à la place de $t^o_{\nu+1}$



Si $t^o_{n+1} < \xi \leqslant b$ et signe $\varphi_o(\xi)$ = signe $\varphi_o(t^o_{n+1})$ on prend ξ à la place de t^o_{n+1}



Si $t^o_{n+1} < \xi \leqslant b$ et signe $\varphi_o(\xi)$ = - signe $\varphi_o(t^o_{n+1})$ on prend ξ à la place de t^o_1



Montrons que pour ce choix de l'ensemble M_1 on a :

$$|L_1(\bar{f})| > |L_o(\bar{f})|$$

Soit $\varepsilon = +1$ ou $\varepsilon = -1$.

Pour $\mu = 1, \ldots, n+1$ $|\varphi_o(t^1_\mu)| \geqslant |L_o(\bar{f})|$ est si μ_o est l'indice tel que :

$\xi = t^1_{\mu_o}$ pour $\mu = \mu_o$.

$$|\varphi_o(t^1_{\mu_o})| > |L_o(\bar{f})|$$

De plus pour $\mu = 1, \ldots, n+1$

$$\text{signe } \varphi_o(t^1_\mu) = \text{signe } (\varepsilon\, \varphi_o(t^o_\mu))$$

Alors :

$$L_1(\bar{f}) = \sum_{\mu=1}^{n+1} \lambda^1_\mu\, \varphi_o(t^1_\mu) = \sum_{\mu=1}^{n+1} \lambda^1_\mu\, |\varphi_o(t^1_\mu)| \text{ signe } (\varepsilon\, \varphi_o(t^o_\mu))$$

En raison des signes de λ^1_μ

$$|L_1(\bar{f})| = \sum_{\mu=1}^{n+1} |\lambda^1_\mu|\, |\varphi_o(t^1_\mu)|$$

Alors :

$$|L_1(\bar{f})| = |L_o(\bar{f})| + \sum_{\mu=1}^{n+1} |\lambda^1_\mu| \left\{ |\varphi_o(t^1_\mu)| - |L_o(\bar{f})| \right\}$$

$$\boxed{|L_1(\bar{f})| > |L_o(\bar{f})|}$$

De plus pour $\mu \neq \mu_o$ $|\varphi_o(t^1_\mu)| = |L_o(\bar{f})|$ donc

$$|L_1(\bar{f})| = |L_o(\bar{f})| + |\lambda^1_{\mu_o}| \times \left(|\varphi_o(t^1_{\mu_o})| - |L_o(\bar{f})| \right)$$

4) Etude de la convergence du deuxième algorithme de Remez

A la $p^{ème}$ itération, l'ensemble M_p est remplacé par un ensemble M_{p+1} de la même manière.

La suite $|L_p(\bar{f})|$ forme ainsi une suite strictement croissante et bornée supérieurement par $\rho_v(f)$, donc c'est une suite convergente. (Sauf si pour un indice p on atteint exactement la meilleure approximation). Montrons que si cette suite admettait une limite β différente de $\rho_v(f)$ ($\beta < \rho_v(f)$) on aboutirait à une contradiction.

Considérons la suite $(t^o_1, \ldots, t^o_{n+1}), \ldots, (t^p_1, \ldots, t^p_{n+1}), \ldots$

Puisque pour tout $p = o,1,2, \ldots$ et pour tout $\mu = 1, \ldots, n+1$ ou a $a \leqslant t^p_\mu \leqslant b$, la suite précédente est prise dans un compact de R^{n+1}.

Montrons que pour μ fixé ($\mu = 1, \ldots, n+1$), il existe un $C(\mu) > 0$ tel que $(t^p_{\mu+1} - t^p_\mu) \geqslant C(\mu)$ pour tout p.

En effet, en raisonnant par l'absurde, supposons que pour tout $\epsilon > 0$, il existe un rang p tel que :

$$t^p_{\mu+1} - t^p_\mu < \epsilon$$

Alors, on pourrait extraire une sous suite
$(t^{q_1}_1, \ldots, t^{q_1}_{n+1}), \ldots, (t^{q_n}_1, \ldots, t^{q_n}_{n+1}), \ldots$ convergente vers :

$(\bar{t}_1, \ldots, \bar{t}_{n+1})$ et telle que pour l'indice μ fixé précédemment on ait :

$$\bar{t}_{\mu+1} = \bar{t}_\mu.$$

Alors, on pourrait trouver $\bar{g} \in V$ tel que $\bar{g}(\bar{t}_j) = f(\bar{t}_j)$ pour $j = 1, \ldots, \mu, \mu+2, \ldots n+1$

La fonction $f(t) - \bar{g}(t)$ étant continue, pour $\epsilon < \lambda^1$ il existe $\eta > o$ tel que pour tout t tel que $|t - \bar{t}_j| < \eta$, on ait :

$$|\bar{g}(t) - f(t)| < \epsilon \quad \text{(pour tout j)}.$$

De plus, on peut trouver N tel que pour $q_n \geqslant N$, on ait $|t^{q_n}_j - \bar{t}_j| < \eta$ pour tout j.

La différence

$$\bar{g}\,(t^{q_n}_j) - g^{q_n}\,(t^{q_n}_j) = (f\,(t^{q_n}_j) - g^{q_n}\,(t^{q_n}_j)) - (f(t^{q_n}_j) - \bar{g}\,(t^{q_n}_j))$$

est du signe de $f\,(t^{q_n}_j) - g^{q_n}\,(t^{q_n}_j)$. Or cette différence alterne en signe pour les $n+1$ points $t^{q_n}_j$.

Alors la fonction $\bar{g} - g^{q_n} \in V$ aurait n zéros dans l'intervalle $[a,b]$, ce qui est en contradiction avec la condition de Haar.

Donc pour μ fixé ($\mu = 1, \ldots, n+1$) il existe $C(\mu) > 0$ tel que $(t^p_{\mu+1} - t^p_\mu) \geqslant C(\mu)$ pour tout p.

L'indice μ parcourant un ensemble fini si $C = \min_{\mu=1,..,n+1} (C(\mu)) > 0$, alors il existe $C > o$ tel que pour tout μ et pour tout p :

$$(t^p_{\mu+1} - t^p_\mu) \geqslant C.$$

Montrons que les λ^p_μ dépendent de façon continue des t^p_μ.

Les λ^p_μ sont obtenus en résolvant le système :

$$\tilde{\lambda}^p_1\, f_1\,(t^p_1) + \tilde{\lambda}^p_2\, f_1\,(t^p_2) + \ldots + \tilde{\lambda}^p_n\, f_1\,(t^p_n) = -\, f_1\,(t^p_{n+1})$$

$$\tilde{\lambda}^p_1\, f_n\,(t^p_1) + \tilde{\lambda}^p_2\, f_n\,(t^p_2) + \ldots + \tilde{\lambda}^p_n\, f_n\,(t^p_n) = -\, f_n\,(t^p_{n+1})$$

et en normalisant :

$$\lambda^p_\mu = \tilde{\lambda}^p_\mu \;/\; \sum_{\mu=1}^{n+1} |\tilde{\lambda}^p_\mu| \qquad (\text{avec } \tilde{\lambda}^p_{n+1} = 1)$$

les coefficients de la matrice et le second membre dépendent de façon continue des t^p_μ. De plus le déterminant de la matrice reste supérieur en valeur absolue à un certain nombre $\delta > 0$. (Il ne peut s'annuler puisque pour tout $\mu = 1, \ldots. n$ et pour tout p $t^p_{\mu+1} - t^p_\mu > C$). Il en résulte que les coefficients de la matrice inverse et donc aussi les $\tilde{\lambda}^p_\mu$ dépendent de façon continue des t^p_μ.

Alors il existe un nombre $d > o$ tel que pour tout p et pour tout μ

$$|\lambda^p_\mu| \geqslant d > 0$$

En effet, pour μ_1 fixé, il existe $d(\mu_1) > 0$ tel que $|\lambda^{p}_{\mu_1}| \geqslant d(\mu_1)$ sinon on pourrait extraire une sous suite $|\lambda^{p_n}_{\mu_1}|$ convergente vers zéro lorsque $p_n \longrightarrow \infty$ et strictement décroissante $|\lambda^{p_{n+1}}_{\mu_1}| < |\lambda^{p_n}_{\mu_1}|$. De cette sous-suite, on pourrait extraire une nouvelle sous-suite telle que (pour $\mu = 1, \ldots, n+1$) $t^{q_n}_{\mu}$ converge vers une limite $\bar{t}_\mu$ et $\lambda^{q_n}_{\mu_1}$ converge vers 0. Les $\lambda^{q_n}_{\mu}$ étant des fonctions continues des points $t^{q_n}_{\mu}$ on a $\lim_{q_n \to \infty} \lambda^{q_n}_{\mu} = \bar{\lambda}_\mu$ et ces $\bar{\lambda}_\mu$ devraient être tous différents de zéro car il correspondent aux $\bar{t}_\mu$ qui sont tels que :

$$\bar{t}_{\mu+1} - \bar{t}_\mu \geqslant C \quad \text{pour } \mu = 1, \ldots, n$$

ce qui est en contradiction avec :

$$\lim_{q_n \to \infty} |\lambda^{q_n}_{\mu_1}| = 0$$

Cette propriété étant valable quel que soit μ_1, il existe $d = \min_{\mu=1,..,n+1} d(\mu), d>0$ tel que pour tout μ et pour tout p :

$$|\lambda^{p}_{\mu}| \geqslant d > 0.$$

Or d'après le § précédent

$$|L_{p+1}(\bar{f})| - |L_p(\bar{f})| \geqslant |\lambda^{p+1}_{\mu_o}| \times (\rho_v(f) - |L_p(\bar{f})|)$$

$$\geqslant d \times (\rho_v(f) - |L_p(\bar{f})|)$$

Soit :

$$(\rho_v(f) - |L_p(\bar{f})|) - (\rho_v(f) - |L_{p+1}(\bar{f})|) \geqslant d \times (\rho_v(f) - |L_p(\bar{f})|)$$

$$(1-d) \times (\rho_v(f) - |L_p(\bar{f})|) \geqslant \rho_v(f) - L_{p+1}(\bar{f})$$

Il est toujours possible de prendre $d < 1$ puisque c'est une borne inférieure

$$k = 1 - d < 1$$

Alors $\rho_v(f) - |L_{p+1}(\bar{f})| \leq k \times (\rho_v(f) - L_p(\bar{f}))$.

La suite $|L_p(\bar{f})|$ converge vers la valeur $\rho_v(f)$. La rapidité de convergence est supérieure à celle d'une progression géométrique de raison k.

5. Application numérique

On considère la fonction $\frac{\sin(t)}{t}$ dans l'intervalle $[o, \pi/2]$ et on se propose de trouver un polynôme pair de degré ≤ 6 tel que :

$$\max_{t \in [o, \frac{\pi}{2}]} \left| \frac{\sin(t)}{t} - (x_1 t^6 + x_2 t^4 + x_3 t^2 + x_4) \right| \text{ soit minimal, en employant}$$

la méthode précédente.

Prenons comme points de départ les points $t^o_i = \frac{\pi}{2 \times 20} \times i$

pour i = o, 1, 2, 3, 4 (suffisamment éloignés des points extrémaux de f(t) - g* (t).

Pour les applications pratiques, on peut prendre comme points de départ les points de Tschebyscheff. Dans le cas présent, si l'on prend pour t^o_i les points de Tschebyscheff relatifs à l'intervalle $[0, \pi/2]$ la fonction g_o est telle que $f - g_o$ est de l'ordre de 10^{-6}.

Avec $t^o_i = \frac{\pi}{2 \times 20} \times i$ $(i = o, 1, \ldots, 4)$ la fonction g_o correspondant à ces points est telle que :

$$\|f - g_o\| = o,\ 1596 \times 10^{-4}$$

A chaque itération le système (I) est résolu par une méthode de Gauss sans division et le maximum de l'erreur

$|f(t) - (x_1 t^6 + x_2 f^4 + x_3 t^2 + x_4)|$ est obtenu en prenant le maximum de :

$$|f(t_i) - (x_1 t^6_i + x_2 t^4_i + x_3 t^2_i + x_4)| \quad \text{avec} \quad t_i = \frac{\pi}{2 \times 100} \times i$$

pour i = 0, ..., 100 et en opérant par dichotomie à partir de ce maximum.

On trouve pour la meilleure approximation

$$x^*_1 = -0,\ 000\ 185\ 24693$$

$$x^*_2 = \ \ 0,\ 008\ 313\ 2721$$

$$x^*_3 = -0,\ 166\ 656\ 84$$

$$x^*_4 = \ \ 0,\ 999\ 999\ 24$$

On a pour résultat :

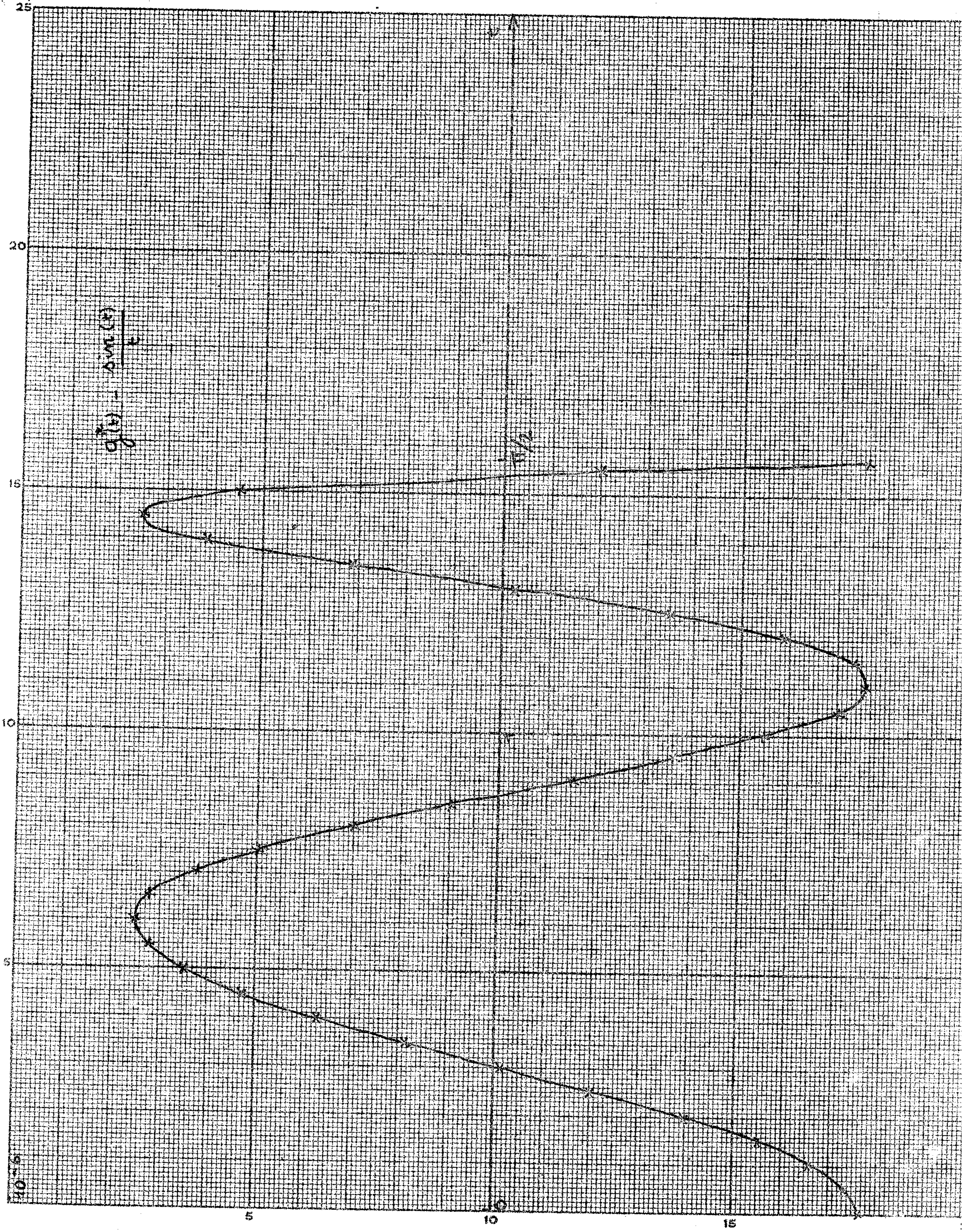
$$\lambda = 0,\ 75 \times 10^{-6} \leqslant \| f - g^* \| \leqslant \ 0,\ 76 \times 10^{-6}$$

En faisant la tabulation de l'erreur :

$$x^*_1\ t^6 + x^*_2\ t^4 + x^*_3\ t^2 + x^*_4 - \frac{\sin t}{t}$$

On obtient les points de déviation maximum

Points de déviation maximum	Valeur de l'erreur
0	- 0, 000 000 75996
0, 6	0, 000 000 75996
1, 12	- 0, 000 000 75251
1, 44	0, 000 000 75996
$\frac{\pi}{2}$ = 1, 5707963	- 0, 000 000 76



$g(t) = \frac{\sin(t)}{t}$
t
π/2
25
20
15
10
5
5
10
15

Le programme ALGOL qui suit permet d'obtenir le polynôme de meilleure approximation de degré $\leq n$, pour une fonction $f(t)$ continue sur l'intervalle $[a,b]$.

Au départ les points t^o_i $(i = 0, \ldots, n+1)$ sont pris équidistants sur l'intervalle $[a,b]$.

$$t^o_i = i \times (b - a) / (n + 1)$$

La résolution du système linéaire sera faite par une méthode d'élimination de Gauss sans division et afin de faciliter cette résolution dans le cas où $b > a \geq o$ on écrira la matrice du système :

$$\begin{pmatrix} (t_{n+1})^n & (t_{n+1})^{n+1} & \ldots\ldots\ldots\ldots & 1 & -1 \\ (t_n)^n & (t_n)^{n-1} & \ldots\ldots\ldots\ldots & 1 & 1 \\ (t_o)^n & (t_o)^{n-1} & \ldots\ldots\ldots\ldots & 1 & (-1)^{n+2} \end{pmatrix}$$

afin que les plus grands coefficients se trouvent au dessus de la diagonale principale.

Pour déterminer le maximum de l'erreur en valeur absolue $\varphi(t) = f(t) - p_n(t)$ on peut chercher le maximum sur un ensemble de $m + 1$ points $t^o_i = i \times (b - a) / m$ $(i = o, \ldots, m)$ et à partir de l'abscisse t_{i_o} ainsi trouvé, on peut opérer à l'aide d'une méthode d'interpolation pour les 3 points

$$t_{i_o - 1}, \quad t_{i_o}, \quad t_{i_o+1}$$

A l'expérience, il est apparu plus efficace de procéder par dichotomie à partir de t_{i_o}.

```
Procédure REMEZ (FCTF,N,PØ,PI,X,EPSI1) ;
    Réel procédure FCTF ; entier N ; réel PØ, PI,EPSI1 ;
    Réel tableau X ;
    Commentaire  CETTE PROCEDURE CALCULE LE POLYNOME DE MEILLEURE APPRØXIMATIØN DE
    DEGRE ≤ N, X1 Y^N + Y^(N-1) + ... + X(N+1), DE LA FONCTION FCTF (Y) SUR L'INTERVALLE
    (PØ, PI) PO < PI, PAR LA METHODE DU DEUXIEME ALGORITHME DE REMEZ ;
    Début entier M ; M : = 100 ; N . = N+2 ;
     Début réel tableau A [1:N, 1:N+1], XØ [0:N];
      entier I,J,K,CØMPT ;
      Réel H,T,INTE,XIP,U ;
    Réel procédure PHI (Y) ; valeur Y ; réel Y ;
      début réel S ; entier P ; S : = 0 ;
      pour P:=1 pas 1 jusqu a N-1  faire
      S:=S + X [P] x (si P = N-1 alors 1.0 sinon (Y^P(N-P-1))) ;
      PHI : = S - FCTF (y) ;
      fin ;
    Procédure RESOL ;
    Commentaire RESOLUTION DU SYSTEME LINEAIRE
      PAR ELIMINATION DE GAUSS ;
      début réel S ;
         pour K:=1 pas 1  jusqu a N-1  faire
         début pour I:=K+1 pas 1 jusqu a N faire
         début si A [K,K]  =0 alors allera  DSØRTIE
            sinon pour J:=K+1 pas 1 jusqu a N+1 faire
            A[I,J]:=A[I,J]x A[K,K]- A[I,K]x A[K,J]
            fin
            fin ;
         X[N]:= A[N,N+1] / A[N,N];
         pour I:=N-1 pas -1 jusqu a 1 faire
         début S:=0 ;
           pour J:=N pas -1 jusqu'a I+1 faire
           S:=S-A[I,J]x X[J];
           X[I]: = (A [I,N+1] + S)/ A[I,I]
         fin
       fin ;
```

```
Procédure TRANS (Z,E) ; réel Z,E ;
    début réel Z4 ;
       WIEDER : Z4 : = PHI (XIP + E x U) ;
         si ABS (Z4) > ABS (Z) alors
         Début XIP:=XIP + E x U ; Z:=Z4 ;
               allera WIEDER fin
         sinon début si E > 0. 001 alors
             début E:=E/10 ; allera  WIEDER
             fin
         fin
     fin ;
Procédure DIC ;
    début réel YO,Y1,Y2 ;
       si (XIP < PI - 0. 0001) et (XIP ⩾ PØ) alors
         début
         Y1:=PHI (XIP) ; Y2 := PHI (XIP + U) ;
         YO:=PHI (XIP - U) ;
         si Y1 x(Y2 - YO) > 0 alors TRANS (Y1, 0.1) ;
         si Y1 x(Y2 - YO) < 0 alors TRANS (Y1, -0.1) ;
         fin
     fin ;

Procédure  ECHANGE ;
    début réel Z,MU,TØ,INT ;
       entier L ; L:=0 ;
    MAX:MU:=0 ;
         TØ:=(PI-PØ)/M ;
         pour L:=0 pas 1 jusqu a M faire
         début INT:=ABS (PHI (LxTØ+PØ)) ;
         si INT > MU alors
         début MU:=INT ; K:=L
         fin
         fin ;
```

```
        XIP:=KxTØ+PØ ;
        Z:=PHI(XIP) ; DIC ;
ECH:si ABS (Z) - ABS (X[N]) < 0 alors aller a SØRTIE ;
        I:=0 ;
    si XI[I]- XIP ⩾ 0 alors
       début si Z x PHI (XI[I] ) ⩾ 0 alors
         début XI[I] := XIP ; aller a SUITE
         fin
         sinon début pour J:=N-1 pas -1 jusqu a 1 faire
         XI[J] :=XI[J-1] ;
         XI[0] :=XIP ; aller a SUITE
         fin
       fin ;
       RIT:I:I+1 ;
         si XI[I] ⩾ XIP alors
         début si Z x PHI (XI[I-1]) ⩾ 0 alors
            début XI[I-1] :=XIP ; aller a SUITE
                fin
         sinon début XI[I] := XIP ; aller a SUITE
               fin
            fin ;
    si I < N-1 alors aller a RIT
    I:=N-1 ;
    si Z x PHI (XI[N-1]) > 0 alors
       début XI[N-1] : = XIP ; aller a SUITE
       fin
    sinon début pour J:=0 pas 1 jusqu a N-2
       faire XI[J] := XI[J+1] ;
             XI[N-1] := XIP ; aller a SUITE
       fin
    fin ;
```

```
INITIAL :
  U:= (PI-PØ)/M ;
  pour I:=o pas 1 jusqu a N-1 faire
    XI[I] := (I x (PI - PØ)) / N-1 ;
  pour I:=N pas -1 jusqu a 1 faire
   début INTE:=XI[N-I];
      pour K:=1 pas 1 jusqu a N-1 faire
      A[I,K]:= si K=N-1 alors 1.0 sinon INTE↑(N-1-K) ;
      K:=N ; A[I,K] := (-1)↑(I+1) ; K:=N+1 ;
      A[I,K] :=FCTF (INTE)
   fin ;
   RESØL ; COMPT:=O ;
ITERATIØN:H:=X[N]; CØMPT:=COMPT+1 ;
   pour I:=1 pas 1 jusqu a N-1 faire
      XØ[I]:= X[I]; ECHANGE ;
   pour I:=N pas -1 jusqu a 1 faire
      début INTE := XI[N-I] ;
       pour K:=1 pas 1 jusqu a N-1 faire
      A[I,K]: = si K=N-1 alors 1.0 sinon INTE↑(N-1-K) ;
      K:=N ; A[I,K]:= (-1)↑(I+1) ; K := N+1 ;
      A[I,K] := FCTF (INTE)
      fin ;
RESØL ; si (ABS (X[N]) - ABS (H) ⩾ EPSI 1)
      alors aller a ITERATIØN ;
DSØRTIE :
 SØRTIE :
      fin
   fin ;
```

Résultats numériques

1) Approximation de e^t par des polynômes de degré $\leqslant n$ dans l'intervalle $[0,1]$

$$x_1 t^n + \ldots + x_n t + x_{n+1}$$

n = 3

$x^*_1 = 0,\ 279\ \ 979\ 03$

$x^*_2 = 0,\ 421\ \ 699\ 46$

$x^*_3 = 1,\ 016\ \ 603\ 4$

$x^*_4 = 0,\ 999\ \ 415\ 26$

$|\lambda| = 0,\ 000\ 544\ 72 \leqslant \rho \leqslant 0,\ 000\ 544\ 74$

n = 4

$x^*_1 = 0,\ 069\ \ 704\ 983$

$x^*_2 = 0,\ 139\ \ 696\ 84$

$x^*_3 = 0,\ 510\ \ 140\ 51$

$x^*_4 = 0,\ 998\ \ 685\ 26$

$x^*_5 = 1,\ 000\ \ 027\ 1$

$|\lambda| = 0,\ 000\ \ 027\ 10 \leqslant \rho \leqslant 0,000\ \ 027\ 21$

n = 5

$x^*_1 = 0,\ 013\ \ 903\ 832$

$x^*_2 = 0,\ 034\ \ 800\ 498$

$x^*_3 = 0,\ 170\ \ 401\ 79$

$x^*_4 = 0,\ 499\ \ 096\ 34$

$x^*_5 = 1,\ 000\ \ 079\ 4$

$x^*_6 = 0,\ 999\ \ 998\ 86$

$|\lambda| = 0,\ 000\ 001\ 13 \leqslant \rho \leqslant 0,\ 000\ 001\ 16$

(La procédure EXP (t) donnant les valeurs de e^t avec une erreur $\leqslant 10^{-8}$ est sans doute calculée à l'aide d'un polynôme de meilleure approximation. Pour le calcul d'une meilleure approximation de l'ordre de 10^{-6} l'utilisation de la procédure EXP (t) est encore correcte).

2) Approximation de $\Gamma(1+t)$ par des polynômes de degré $\leq n$ dans l'intervalle $[0,1]$.

$$x_1 t^n + ... + x_n t + x_{n+1}.$$

Les valeurs de $\Gamma(1+t)$ ont été calculées à l'aide du développement en série de Tschebyscheff donné par Clenshaw (1) (en tenant compte des erreurs d'arrondi $\Gamma(1+t)$ est ainsi calculée avec une erreur $\leq 10^{-7}$)

n = 3

$x^*_1 = -0,\ 141\ 304\ 54$

$x^*_2 = \ \ 0,\ 666\ 832\ 01$

$x^*_3 = -0,\ 525\ 527\ 45$

$x^*_4 = \ \ 0,\ 998\ 645\ 03$

$|\lambda| = 0,\ 001\ 354\ 97 \leq \rho \leq 0,\ 001\ 355\ 00.$

n = 4

$x^*_1 = \ \ 0,\ 174\ 579\ 99$

$x^*_2 = -0,\ 484\ 168\ 62$

$x^*_3 = \ \ 0,\ 875\ 541\ 90$

$x^*_4 = -0,\ 565\ 560\ 64$

$x^*_5 = \ \ 0,\ 999\ 803\ 69$

$|\lambda| = 0,\ 000\ 196\ 30 \leq \rho \leq 0,\ 000\ 196\ 33$

Voir CLENSHAW [1]

n = 5

$x^*_1 = -\ 0,\ 100\ 133\ 43$

$x^*_2 = \ \ 0,\ 420\ 169\ 25$

$x^*_3 = -\ 0,\ 693\ 747\ 50$

$x^*_4 = \ \ 0,\ 947\ 878\ 90$

$x^*_5 = -\ 0,\ 574\ 167\ 17$

$x^*_6 = \ \ 0,\ 999\ 962\ 88$

$|\lambda| = 0,\ 000\ 037\ 10 \leqslant \rho \leqslant 0,\ 000\ 037\ 13$

n = 6

$x^*_1 = \ \ 0,\ 878\ 978\ 81$

$x^*_2 = -\ 0,\ 325\ 203\ 63$

$x^*_3 = \ \ 0,\ 668\ 940\ 59$

$x^*_4 = -\ 0,\ 819\ 867\ 61$

$x^*_5 = \ \ 0,\ 976\ 573\ 29$

$x^*_6 = -\ 0,\ 576\ 528\ 02$

$x^*_7 = \ \ 0,\ 999\ 993\ 77$

$|\lambda| = 0,\ 000\ 006\ 22 \leqslant \rho \leqslant 0,\ 000\ 006\ 25$

n = 7

$x^*_1 = -\ 0,\ 0511\ 451\ 00$

$x^*_2 = \ \ 0,\ 252\ \ 893\ 28$

$x^*_3 = -\ 0,\ 564\ \ 709\ 19$

$x^*_4 = \ \ 0,\ 829\ \ 310\ 43$

$x^*_5 = -\ 0,\ 874\ \ 746\ 10$

$x^*_6 = \ \ 0,\ 985\ \ 458\ 83$

$x^*_7 = -\ 0,\ 577\ \ 062\ 11$

$x^*_8 = \ \ 0,\ 999\ \ 998\ 92$

$|\lambda| = 0,\ 000\ 001\ 08 \leqslant \rho \leqslant 0,\ 000\ 001\ 11$

3) Approximation de $e^t \int_t^\infty \frac{e^{-x}}{x} dx$

dans l'intervalle $[4,10]$ par un polynôme de degré $\leq n$:

$$x_1 t^n + x_2 t^{n-1} + \ldots + x_n t + x_{n+1}$$

Les valeurs de la fonction $e^t \int_t^\infty \frac{e^{-x}}{x} dx$ ont été calculées à l'aide du développement en série de Tschebyscheff donné par Clenschaw [1]

n = 3

$x^*_1 = -0,\ 000\ 353\ 654\ 78$

$x^*_2 = \ \ 0,\ 000\ 904\ 104\ 5$

$x^*_3 = -0,\ 102\ 619\ 14$

$x^*_4 = \ \ 0,\ 480\ 516\ 28$

$|\lambda| = 0,\ 000\ 494\ 15 \leq \rho \leq 0,\ 000\ 494\ 17$

(1) Voir CLENSHAW [2]

4) Application de $f(t) = -\int_t^{\infty} \frac{e^{-x}}{x}\,dx - \log(|t|)$

dans l'intervalle $[-4, +4]$ par un polynôme de degré $\leqslant 4$. $x_1 t^4+x_2 t^3+x_3 t^2+x_4 t+x_5$.

La fonction f (t) a été calculée à l'aide du développement donné par Clenshaw (1). En divisant tous les coefficients du système par 10^2 le programme donne les résultats suivants :

$x^*_1 = 0,\ 019\ 089\ 561$

$x^*_2 = 0,\ 100\ 177\ 56$

$x^*_3 = 0,\ 179\ 504\ 85$

$x^*_4 = 0,\ 805\ 734\ 88$

$x^*_5 = 0,\ 668\ 247\ 92$

$|\lambda| = 0,\ 183\ 023\ 30 \leqslant \rho \leqslant 0,\ 183\ 024\ 20$

(1) Voir CLENSHAW [2]

CHAPITRE III

METHODE PAR DECOMPOSITION DE LA NORME (1)

Soit B un espace vectoriel normé complet sur le corps des réels, muni d'une norme φ ; f, g, h, ... désignent les élements de B.

Soit L (x) une application linéaire de R^n dans B. Elle est complètement déterminée lorsque l'on connait les images de la base fondamentale $\{e_i\}$ de R^n. Soient

$$f_1 = L(e_1), \ldots, f_n = L(e_n)$$

$$x \xrightarrow{L} x_1 f_1 + x_2 f_2 + \ldots + x_n f_n$$

Soit $\bar{B}$ l'ensemble des fonctionnelles linéaires continues de B dans R. C'est le dual de B. Soient $\bar{f}, \bar{g}, \bar{h}, \ldots \in \bar{B}$

$$f \xrightarrow{\bar{f}} \langle \bar{f}, f \rangle = \bar{f}(f) \in R.$$

On sait que $\bar{B}$ est un espace de Banach relativement à la norme $\bar{\varphi}$ définie par :

$$\bar{\varphi}(\bar{f}) = \sup_{\substack{f \neq o \\ f \in B}} \frac{|\langle \bar{f}, f \rangle|}{\varphi(f)}$$

Soit L une application linéaire de R^n dans B :

Pour $\bar{f} \in \bar{B}$ et $x \in R^n$ $\quad \langle \bar{f}, L(x) \rangle \in R$

Si $\bar{f}$ est fixé, on considère l'application linéaire de R^n dans R qui à x fait correspondre le nombre :

$$x_1 \langle \bar{f}, f_1 \rangle + \ldots + x_n \langle \bar{f}, f_n \rangle.$$

Ainsi à $\bar{f}$ on peut faire correspondre une application linéaire de R^n dans R, c'est à dire un élément du dual de R^n qui peut être identifié à R^n lui-même.

$$\bar{f} \xrightarrow{L^T} L^T(\bar{f}) = \begin{vmatrix} \langle \bar{f}, f_1 \rangle \\ \vdots \\ \langle \bar{f}, f_n \rangle \end{vmatrix} = z_{\bar{f}} \in R^n$$

(1) Le principe de cette méthode a été exposé par le professeur N. GASTINEL au Congrès sur l'approximation à GALESBURG (Nov. 1963).

On a $< \bar{f}, L(x) > = < L^T(\bar{f}), x >$ pour tout $x \in R^n$ et pour tout $\bar{f} \in B$.
L^T est l'application transposée, on dit aussi adjointe de L.

Décomposition de la norme φ.

Définition

On dit que la norme φ est décomposée si pour tout $f \in B$ il existe $\bar{z}_f \in \bar{B}$ tel que :

$$< \bar{z}_f, f > = \varphi(f).$$

Application à la recherche de la meilleure approximation dans l'ensemble $\mathcal{C}$ des fonctions continues, à valeurs réelles définies sur un compact K de R^n.

1. Décomposition de la norme.

Soit K un compact de R^n et $\mathcal{C}$ l'ensemble de fonctions à valeurs réelles, continues sur K. $\mathcal{C}$ est un espace de Banach relativement à la norme

$$\varphi(x) = \max_{t \in K} |x(t)| .$$

On sait que l'espace dual $\bar{\mathcal{C}}$ est isomorphe à l'espace $\mathcal{B}$ des mesures μ, avec signes, régulières sur K :

Si u est une application linéaire de $\mathcal{C}$ dans R, continue, alors il existe $\mu_u \in \mathcal{B}$ telle que pour tout $x \in \mathcal{C}$

$$u(x) = \int_K x \, d\mu_u$$

La norme sur $\mathcal{B}$ est $\bar{\varphi}(u) = \int_K |d\mu_u| .$

On a (Inégalité de Hölder)

$$|u(x)| \leq \varphi(x) \int_K |d\mu_u| = \varphi(x) \, \bar{\varphi}(u).$$

On sait que l'égalité à lieu si et seulement si :

1) La mesure μ_u a le signe de x c'est à dire
$\int_E x \, d\mu_u \geq 0$ pour tout $E \subset K$ et μ_u mesurable

2) Le support de μ_u est l'ensemble des points t tels que

$$|x(t)| = \varphi(x).$$

2. Décomposition naturelle de la norme de $\mathcal{C}$.

On considère $f \in \mathcal{C}$, $\varphi(f) = \underset{t \in K}{\text{Max}} |f(t)|$.

Soient $e^+ = \{ t ; \quad f(t) = \varphi(f) \}$, $e^- = \{ t ; f(t) = -\varphi(f) \}$.

$\mu^-_{z_f}$ étant définie par des masses de poids total $\alpha > 0$ placées aux points e^+ et par des masses de poids total $-\beta < 0$ en e^- et des masses nulles dans $K - e^+ - e^-$.

$$\int_K f \, d\mu^-_{z_f} = \varphi(f) \left[\int_{e^+} d\mu^-_{z_f} - \int_{e^-} d\mu^-_{z_f} \right]$$

$$= \left[\alpha + \beta \right] \varphi(f)$$

Si on pose $\alpha + \beta = 1$

$$\int_K f \, d\mu^-_{z_f} = \varphi(f) \quad \text{et} \quad \int_K |d\mu^-_{z_f}| = \int_{e^+} |d\mu^-_{z_f}| + \int_{e^-} |d\mu^-_{z_f}|$$

$$\int_K |d\mu^-_{z_f}| = \alpha + \beta = 1$$

L'application de $\mathcal{C}$ dans $\tilde{\mathcal{C}}$ qui à f fait correspondre $\bar{z}_f$ est une décomposition de norme telle que $\bar{\varphi}(\bar{z}_f) = 1$.

Une telle décomposition est appelée décomposition naturelle dans $\mathcal{C}$.

Soit $f \in \mathcal{C}$. V un sous espace vectoriel de $\mathcal{C}$ ayant pour base $f_1, \ldots, f_n$. Soit L l'application linéaire de R^n sur V définie par :

$$x \xrightarrow{L} L(x) = x_1 f_1 + \ldots + x_n f_n$$

On suppose que f n'appartient pas à V.

Soit $r = L(x) - f$. Le problème de meilleure approximation revient à chercher x^* tel que :

$$\varphi(f - L(x^*)) = \underset{x \in R^n}{\text{Inf}} \quad \varphi(f - L(x))$$

3. Théorème I : Caractérisation de la meilleure approximation dans $\mathcal{C}$.

Pour que $x^* \in R^n$ soit tel que $L(x^*)$ soit une meilleure approximation d'une fonction $f \in \mathcal{C}$ relativement au sous espace vectoriel V, il faut et il suffit que :

$$L^T(\bar{z}_{r^*}) = 0 \quad \in R^n \quad \text{c'est-à-dire :}$$

$$< \bar{z}_r, L(x) > = 0 \quad \text{pour tout } x \in R^n.$$

Condition nécessaire.

Soient $r^* = L(x^*) - f$ et $\rho_v(f) = \varphi(r^*)$

D'après le théorème I du chapitre II, il existe une fonctionnelle linéaire que l'on note ici $\bar{v} \in \bar{\mathcal{C}}$ telle que :

$$\begin{cases} < \bar{v}, L(x) > = 0 & \text{pour tout } x \in R^n \\ < \bar{v}, f > \; = \bar{v}(f) = \rho_v(f) \\ \bar{\varphi}(\bar{v}) \; = 1 \end{cases}$$

Alors $\bar{v} = -\bar{z}_{r^*}$

En effet, puisque :

$$\rho_v(f) = <\bar{v}, f> = <\bar{v}, f - L(x^*)> \leq \varphi(\bar{v}) \times \varphi(-r^*) = \rho_v(f)$$

l'inégalité de Hölder devient une égalité ce qui entraine que $\mu_{\bar{v}}$ est une mesure de même signe que $-r^*$ et a pour support les points extrémaux de r^*

$$\bar{v} = -\bar{z}_{r^*}.$$

Condition suffisante.

Si dans R^n il existe x^* tel que pour tout $x \in R^n$

$$< \bar{z}_{r^*}, L(x) > = 0$$

alors, d'après l'inégalité de Hölder :

$$|< \bar{z}_{r^*}, f >| = |< \bar{z}_{r^*}, f - L(x) >| \leq \bar{\varphi}(\bar{z}_{r^*}) \times \varphi(-r).$$

Or $\bar{\varphi}(\bar{z}_{r^*}) = 1$,

$$< \bar{z}_{r^*},\ f \geqslant = < \bar{z}_{r^*},\ f - L(x^*) > = -\varphi(r^*)$$

Alors d'après l'inégalité précédente

$$\varphi(r^*) \leqslant \varphi(r) \text{ pour tout } x \text{ dans } R^n$$

x^* fournit donc une meilleure approximation.

4. Algorithme du rapprochement maximum en norme φ, sur les directions $L^T(\bar{z}_r)$.

Si x est donné dans R^n on se propose de trouver x' tel que $\varphi(r') < \varphi(r)$.

Soit x, donc $L^T(\bar{z}_r)$, donné avec $L^T(\bar{z}_r) \neq 0$. On considère la fonction de λ :

$$\mu(\lambda) = \varphi(L(x - \lambda L^T(\bar{z}_r)) - f)$$
$$= \varphi(r - \lambda L L^T(\bar{z}_r)).$$

Lorsque λ varie de 0 à $+\infty$ la fonction $\mu(\lambda)$ est une fonction continue de λ, telle que $\mu(o) = \varphi(r)$ et $\mu(\lambda)$ tend vers $+\infty$ si λ tend vers $+\infty$.

Soit $\mathscr{C}_x$ le lieu des points ξ de R^n, tels que si $\rho = L(\xi) - g$

$$\varphi(\rho) \leqslant \varphi(r)$$

$\mathscr{C}_x$ est un ensemble convexe de R^n contenant x^* dans son intérieur.

Théorème II :

Pour tout x différent d'une meilleure approximation x^* , $L^T(\bar{z}_r)$ n'est pas nul (d'après le théorème I chapitre 3).

La direction $L^T(\bar{z}_r)$ est perpendiculaire à un hyperplan d'appui en x, à $\mathscr{C}_x$ et dirigée vers l'extérieur de $\mathscr{C}_x$.

Si $\xi \in \mathcal{C}_x$, le produit scalaire

$$< L^T(\bar{z}_r),\ \xi - x > = < \bar{z}_r,\ L(\xi - x) > = < \bar{z}_r,\ L(\xi) - f > - <\bar{z}_r,\ r >$$

$$= < \bar{z}_r,\ \rho > - \varphi(r) \ .$$

D'après la décomposition naturelle de la norme φ

$$|< \bar{z}_r,\ \rho >| \leqslant \varphi(\rho).$$

Alors le produit scalaire $< L^T(\bar{z}_r),\ \xi - x >$ est $\leqslant 0$ pour tout $\xi \in \mathcal{C}_x$.

$L^T(\bar{z}_r)$ est perpendiculaire à un plan d'appui en x à $\mathcal{C}_x$.

On considère $x - \lambda\ L^T(\bar{z}_r)$.

Deux cas se présentent :

a) Il existe $\lambda_o > o$ tel que pour $o \leqslant \lambda \leqslant \lambda_o$
$x - \lambda\ L^T(\bar{z}_r) \in \mathcal{C}_x$, λ_o correspondant au point x_o où l'on sort de $\mathcal{C}_x$.

b) Seul $x \in \mathcal{C}_x$ pour tout λ.

Dans le premier cas $\mu(\lambda)$ est telle que

$\mu(0) = \varphi(r)$ $\quad$ $\mu(\lambda_o) = \varphi(r)$ $\quad$ et pour

$0 < \lambda < \lambda_o$ $\quad$ $\mu(\lambda) \leqslant \varphi(r)$.

Définition :

Pour tout $x \neq x^*$ avec $L^T(\bar{z}_r) \neq o$ s'il existe une valeur de $\lambda \neq 0$, soit m telle que pour $x' = x - m\ L^T(\bar{z}_r)$

$$\varphi(r') = \inf_{\lambda} \varphi\ (r - \lambda\ L\ L^T(\bar{z}_r)).$$

On dira que la décomposition $\bar{z}_r$ est adaptée à L.

$\mathcal{M}$ est alors la meilleure valeur de λ pour approcher x^* à partir de x, dans la direction de $L^T(\bar{z}_r)$, au sens de la norme φ.

5. Application à la recherche de la meilleure approximation dans l'espace $\mathcal{C}(a,b)$ des fonctions continues à valeurs réelles, définies sur un intervalle $[a,b]$.

On considère sur l'espace vectoriel $\mathcal{C}[a,b]$ la norme

$$\varphi(f) = \max_{t \in [a,b]} |f(t)| .$$

Soit V un sous espace vectoriel de $\mathcal{C}[a,b]$ ayant pour base $f_1, \ldots, f_n$ et possédant la propriété d'interpolation (condition de Haar). Soit $f \in \mathcal{C}[a,b]$ mais n'appartenant pas à V. On cherche la meilleure approximation de f par des éléments de V.

Posons $r(t) = f(t) - (x_1 f_1(t) + \ldots + x_n f_n(t))$

et $M = \max_{t \in [a,b]} |r(t)| = \varphi(r)$.

On considère l'ensemble des points t_i $(i = o, 1, \ldots, p)$ $t_i \in [a,b]$ tels que :

$$|r(t_i)| = M.$$

Soit :

$$\bar{z}_r = \alpha_o \times \text{signe}(r(t_o)) \times \delta_o + \alpha_1 \times \text{signe}(r(t_1)) \times \delta_1 + \ldots + \alpha_p \times \text{signe}(r(t_p)) \times S_p$$

où $\delta_1, \ldots, \delta_p$ représentent les distributions de Dirac aux points $t_o, \ldots t_p$.
De plus $\alpha_i \geqslant 0$ pour $i = 1, \ldots p$ et $\alpha_o + \ldots + \alpha_p = 1$.

On obtient ainsi une décomposition naturelle de la norme φ.

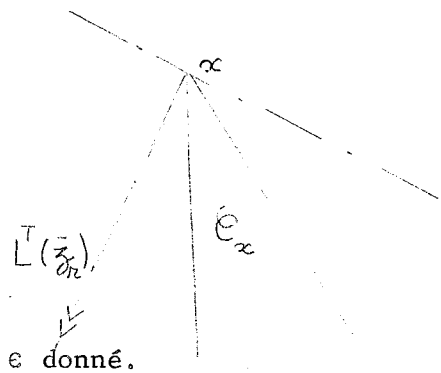
$$L^T(\bar{z}_r) = \left| \begin{array}{l} \langle \bar{z}_r, f_1 \rangle = \langle \alpha_o \times \text{signe}(r(t_o)) \times \delta_o + \ldots\ldots\ldots\ldots \\ \qquad + \alpha_p \times \text{signe}(r(t_p)) \times \delta_p, f_1 \rangle \\ \langle \bar{z}_r, f_n \rangle = \langle \alpha_o \times \text{signe}(r(t_o)) \times \delta_o + \ldots\ldots\ldots\ldots \\ \qquad + \alpha_p \times \text{signe}(r(t_p)) \times \delta_p, f_n \rangle \end{array} \right.$$

Itération de l'algorithme de § 4.

Soit x_o donné. On calcule $L^T(\bar{z}_{r_o})$ et par un procédé de dichotomie, on se déplace suivant la direction $L^T(\bar{z}_{r_o})$ vers l'intérieur de $\mathscr{C}_{x_o}$ jusqu'au point x_1 ou $\varphi(r_1)$ est minimum. Puis on itère à partir de x_1.

Remarque :

Dès que r (t) possède plus d'un point t_i tel que $|r(t_i)| = M$, la surface de $\mathscr{C}_x$ au point x peut avoir "une pointe". On n'est plus assuré qu'il existe $\lambda_o > o$ tel que pour $o \leqslant \lambda \leqslant \lambda_o$ $\quad x - \lambda L^T(\bar{z}_r) \in \mathscr{C}_x$



pour ε donné.

Pour tenter d'éviter cette difficulté, on considèrera non pas l'ensemble des points t_i tels que $|r(t_i)| = M$, mais l'ensemble des maxima de $|r(t)|$ (et s'il y a lieu les points a et b) tels que :

$$M-\varepsilon \leqslant |f(t_i)-(x_1 f_1(t_i)+ \ldots+x_n f_n(t_i))| \leqslant M+\varepsilon$$

6. Application numérique

Approximation de sin (t) | t dans l'intervalle $[o, \pi/2]$ par un polynôme de degré 6 pair :

$$x_1 t^6 + x_2 t^4 + x_3 t^2 + x_4$$

Dans ce cas :

$$L^T(\bar{z}_r) = \begin{vmatrix} \alpha_o \text{ signe } (r(t_o)) \times t_o^6 + \alpha_1 \text{ signe } (r(t_1)) x t_1^6 + \ldots + \alpha_p \text{ signe } (r(t_p)) x t_p^6 \\ \alpha_o \text{ signe } (r(t_o)) \times t_o^4 + \ldots\ldots\ldots\ldots\ldots\ldots + \alpha_p \text{ signe } (r(t_p)) x t_p^4 \\ \\ \alpha_o \text{ signe } (r(t_o)) + \ldots\ldots\ldots\ldots\ldots\ldots + a_p \text{ signe } (r(t_p)) \end{vmatrix}$$

Pour i = 1, ..., p on prend

$$\alpha_i = \frac{1}{p} \times \frac{1}{\sqrt{t_i^{12} + t_i^8 + t_i^4 + 1}}$$

afin que chacun des vecteurs $\begin{vmatrix} t_i^6 \\ t_i^4 \\ \vdots \\ t_i^2 \\ 1 \end{vmatrix}$ intervienne

dans le calcul de $L^T(\bar{z}_r)$ avec une norme euclidienne égale à 1.

On prend comme vecteur de départ :

$$\begin{vmatrix} x^o_1 = -0,001 \\ x^o_2 = 0,01 \\ x^o_3 = -0,15 \\ x^o_4 = 1 \end{vmatrix}$$

Le maximum de l'erreur est :

$$M_o = \underset{t \in [0, \frac{\pi}{2}]}{\text{Max}} |r_o(t)| = 0,03912901$$

Pour ce vecteur de départ, il existe un seul point :

$$t = \pi/2 \quad \text{tel que} \quad |r_o(t)| = M_o.$$

On étudie la méthode, en faisant varier ϵ qui permet de déterminer les points t^o_i $i = 1, \ldots p$ tels que :

$$M - \epsilon \leq |f(t_i) - (x_1 f_1(t_i) + \ldots + x_n f_n(t_i))| \leq M + \epsilon$$

Soient à la $k^{ème}$ itération le vecteur $\begin{vmatrix} x^k_1 \\ \\ x^k_4 \end{vmatrix}$

et $M_k = \underset{t \in [0,\pi/2]}{\text{Max}} \left| x_1^k t^6 + \ldots + x^k_4 - \frac{\sin(t)}{t} \right|$

1) On prend $\epsilon = M_k/5$

Le procédé s'arrête après 8 itérations

$M_8 = 0,\ 004\ 279\ 548\ 7$

$x^8_1 = -0,\ 000\ 876\ 704\ 43$

$x^8_2 = 0,\ 005\ 533\ 284\ 8$

$x^8_3 = -0,\ 155\ 350\ 52$

$x^8_4 = 0,\ 995\ 784\ 10$

Le maximum de l'erreur est atteint en 3 points avec des signes alternés

$t_o = 0$

$t_1 = 1,\ 159\ 68$

$t_2 = \pi/2$

2) On prend $\varepsilon = M_k / 10$

Le procédé s'arrête après 9 itérations

$$M_9 = 0,\ 004\ 184\ 700\ 5$$

$$x^9_1 = -0,\ 000\ 903\ 827\ 29$$

$$x^9_2 = 0,\ 005\ 511\ 876\ 8$$

$$x^9_3 = -0,\ 155\ 367\ 03$$

$$x^9_4 = 0,\ 995\ 815\ 30$$

Le maximum de l'erreur est atteint en 3 points avec des signes alternés.

$$t_o = 0$$

$$t_1 = 1,\ 151\ 81$$

$$t_2 = \pi/2$$

Soient à la $k^{ème}$ itération le vecteur $\begin{vmatrix} x^k_1 \\ \vdots \\ x^k_4 \end{vmatrix}$

et $M_k = \underset{t\in[o,\frac{\pi}{2}]}{\text{Max}} \left| x^k_1 t^6 + x^k_2 t^4 + x^k_3 t^2 + x^k_4 - \frac{\sin(t)}{t} \right|$

1) On prend $\varepsilon = M_k/5$

Le procédé s'arrête après 8 itérations.

$$M_8 = 0,\ 004\ 279\ 548\ 7$$

$$x^8_1 = -0,\ 000\ 876\ 704\ 43$$

$$x^8_2 = 0,\ 005\ 533\ 284\ 8$$

$$x^8_3 = -0,\ 155\ 350\ 62$$

$$x^8_4 = 0,\ 995\ 784\ 10$$

Le maximum de l'erreur en valeur absolue est atteint en 3 points, avec des signes alternés.

$t_o = 0$

$t_1 = 1,\ 159\ 68$

$t_2 = \pi/2.$

2) On prend $\varepsilon = M_k/10$

Le procédé s'arrête après 9 itérations

$M_9 = 0,\ 004\ 184\ 700\ 5$

$x^9_1 = -0,\ 000\ 903\ 827\ 29$

$x^9_2 = 0,\ 005\ 511\ 876\ 8$

$x^9_3 = -0,\ 155\ 367\ 03$

$x^9_4 = 0,\ 995\ 815\ 30$

Le maximum de l'erreur en valeur absolue est atteint en 3 points avec des signes alternés.

$t_o = 0$

$t_1 = 1,\ 151\ 81$

$t_2 = \pi/2.$

3) On prend $\varepsilon = M_k/100$

Le procédé s'arrête après 7 itérations

$M_7 = 0,\ 004\ 303\ 917\ 2$

$x^7_1 = -0,\ 000\ 972\ 186\ 35$

$x^7_2 = 0,\ 005\ 618\ 420\ 9$

$x^7_3 = -0,\ 155\ 225\ 79$

$x^7_4 = 0,\ 995\ 718\ 78$

Le maximum de l'erreur en valeur absolue est atteint en 3 points avec des signes alternés.

$t_o = 0$

$t_1 = 1,\ 153\ 85$

$t_2 = \pi/2$

Dans les trois cas précédents, lorsque la valeur absolue de l'erreur r(t) atteint son maximum en trois points t_o, t_1, t_2, sur la direction $L^T(\bar{z}_r)$ telle qu'elle a été calculée, il n'existe pas toujours un λ_o tel que pour $0 \leqslant \lambda \leqslant \lambda_o$ $x - \lambda (L^T(\bar{z}_r))$ appartient à $\mathcal{C}_x$ et seul x appartient à $\mathcal{C}_x$.

7. Méthode dérivée de la méthode du gradient conjugué.

a) On va rappeler le principe d'une méthode qui permet de minimiser une expression F (x), x étant un vecteur de R^n, à l'intérieur du domaine convexe $F(x) \leqslant$ constante dans le cas où l'on peut calculer le gradient.

Exemple : résolution de $Ax - \lambda B\ x = o$ avec A symétrique B symétrique définie positive.

On considère $F(x) = \dfrac{x^T A\ x}{x^T B\ x}$ le minimum de F (x) fournira la plus petite valeur propre.

Soit x_o donné. On peut calculer le gradient en x_o soit r_o normal à la surface $F(x) = F(x_o)$. Soit x_1 un point sur r_o avec $r_1^T\ r_o = o$. On suppose que les vecteurs

r_1 et r_o sont de longueur 1.

On considère :

$u_1 = r_o + r_1$

$v_1 = -r_o + r_1$

On cherche par dichotomie

sur la direction u_1 le point x_{11} tel que $F(x_{11}) = F(x_1)$

sur la direction r_1 le point x_{21} tel que $F(x_{21}) = F(x_1)$

sur la direction v_1 le point x_{31} tel que $F(x_{31}) = F(x_1)$

On considère l'ellipse passant par x_{11}, x_{21}, x_{31}, x_1 et tangente en x_1 à r_o. Alors le centre de l'ellipse est pris comme point x_2. $F(x_2) < F(x_1)$. On itère le procédé à partir de x_2, en calculant r_2, le gradient en x_2 et en utilisant les vecteurs

$$u_2 = r_1 + r_2$$
$$v_2 = -r_1 + r_2$$

b) Pour la recherche de la meilleure approximation on ne peut plus utiliser une ellipse puisque le vecteur $L^T(\bar{z}_r)$ n'est plus normal à la surface $\mathcal{E}_x$, lorsque $\mathcal{E}_x$ présente un angle aigu au point x.

On se propose alors d'utiliser le procédé suivant : Soit x_o un vecteur donné. On calcule $L^T(\bar{z}_{r_o})$ et on détermine x_1 comme un paragraphe précédent.

A partir de x_1 on calcule $L^T(\bar{z}_{r_1})$. Si l'on note $\|L^T(\bar{z}_r)\|$ la norme euclidienne du vecteur $L^T(\bar{z}_r) \in R^n$ pour $\alpha = 5$ on considère les trois vecteurs.

$$v_1 = \frac{L^T(\bar{z}_{r1}}{\|L^T(\bar{z}_{r_1})\|}$$

$$v_2 = \frac{L^T(\bar{z}_{r_1}}{\|L^T(\bar{z}_{r_1})\|} + 1/\alpha \quad \frac{L^T(\bar{z}_{r_o})}{\|L^T(\bar{z}_{r_o})\|}$$

et $$v_3 = \frac{L^T(\bar{z}_{r1})}{\|L^T(\bar{z}_{r_1})\|} - 1/\alpha \quad \frac{L^T(\bar{z}_{r_o})}{\|L^T(\bar{z}_{r_o})\|}$$

Soit $M_1 = \underset{t\in[o,\ \pi/2]}{\text{Max}} |r_1(t)|$

On se déplace à partir de x_1 suivant les trois directions $-v_1$, $-v_2$, $-v_3$ et on cherche par dichotomie m_1, m_2, m_3 trois scalaires tels que aux points $x_1 - m_1 v_1$, $x_1 - m_2 v_2$, $x_1 - m_3 v_3$ les erreurs correspondantes reprennent en norme la valeur M_1.

On considère :

$m_k = \text{Max}\ (m_1, m_2, m_3)$ et on prend

$x_2 = x_1 - (m_k/2) \times v_k$

Alors $\varphi(r_2) \leqslant \varphi(r_1)$.

On itère le procédé à partir de x_2.

On prend comme précédemment pour vecteur de départ

$x^o_1 = -0,001$

$x^o_2 = 0,01$

$x^o_3 = -0,15$

$x^o_4 = 1$

$M_o = \underset{t \in [o, \pi/2]}{\text{Max}} |r_o(t)| = 0,039\ 129\ 01$

Le procédé s'arrête après 7 itérations

$M_7 = 0,003\ 087\ 595\ 1$

$x^7_1 = -0,000\ 227\ 775\ 1$

$x^7_2 = 0,004\ 491\ 445\ 1$

$x^7_3 = -0,156\ 941\ 27$

$x^7_4 = 0,996\ 997\ 83$

Le maximum de l'erreur est atteint en 3 points avec des signes alternés

$t_o = o$

$t_1 = 1,113\ 324\ 8$

$t_2 = \pi/2$

Dans cet exemple ϵ a été pris égal à la $i^{\text{ème}}$ itération à $M_i/10$

CHAPITRE IV

LE PREMIER ALGORITHME DE REMEZ (1)

APPLICATION AU CAS DES POLYNOMES

1. Soit f (t) une fonction appartenant à l'espace $\mathcal{C}[a,b]$ des fonctions continues, à valeurs réelles définies sur un intervalle fermé [a,b].

On cherche le polynôme de degré inférieur ou égal à n soit :

$$p^*_n(t) = x^*_1 t^n + \ldots\ldots + x^*_{n+1}$$

tel que $p^*_n(t)$ soit la meilleure approximation de f(t) sur l'intervalle [a,b] au sens de Tschebyscheff.

On prend la norme :

$$\|f\| = \max_{t\in[a,b]} |f(t)|.$$

Soit $p_n(t)$ un polynôme donné. On pose :

$$\varphi(t) = f(t) - p_n(t) \quad \text{et} \quad L = \|\varphi\| = \max_{t\in[a,b]} |\varphi(t)| .$$

On se propose de trouver un polynôme $\omega(t)$ de $d^o \leqslant n$ tel que :

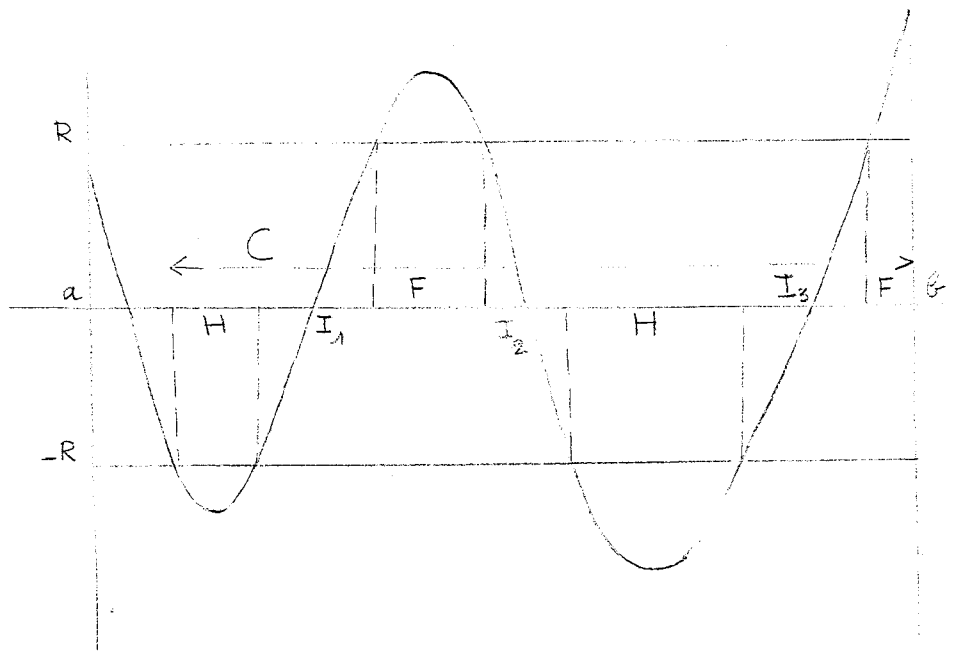
$$\|\varphi - \omega\| < \|\varphi\| .$$

On pose $\rho = \|f - p^*_n\|$,

(1) Voir REMEZ [9]

Soit R un nombre positif donné $o < R < L$.

On considère les ensembles $F \subset [a,b]$ et $H \subset [a,b]$ tels que respectivement $\varphi(t) \geqslant R$ et $\varphi(t) \leqslant -R$



Soit C le plus petit segment qui peut contenir FUH. Dans C - (FUH) on considère l'ensemble des intervalles I_1, I_2, ..., I_m qui séparent un segment de F et un segment de H.

Si $m > n$ alors d'après le théorème de De La Vallée Poussin la déviation maximum ρ (pour la meilleure approximation) est telle que $\rho \geqslant R$. On obtient ainsi une borne inférieure de la déviation maximum.

On appelle A la valeur maximum de la borne inférieure R que l'on peut ainsi obtenir.

$$A \leqslant \rho \leqslant L$$

Si le nombre m des intervalles $I_1, \ldots, I_m$ est $< n + 1$ pour tout R aussi petit qu'on veut, alors on prendra $A = 0$.

Après avoir déterminé L et A on prend :

$$R_1 = A + \theta_1 (L-A), \quad R_2 = A + \theta_2 (L-A)$$

avec $o < \theta_1 < \theta_2 < 1$.

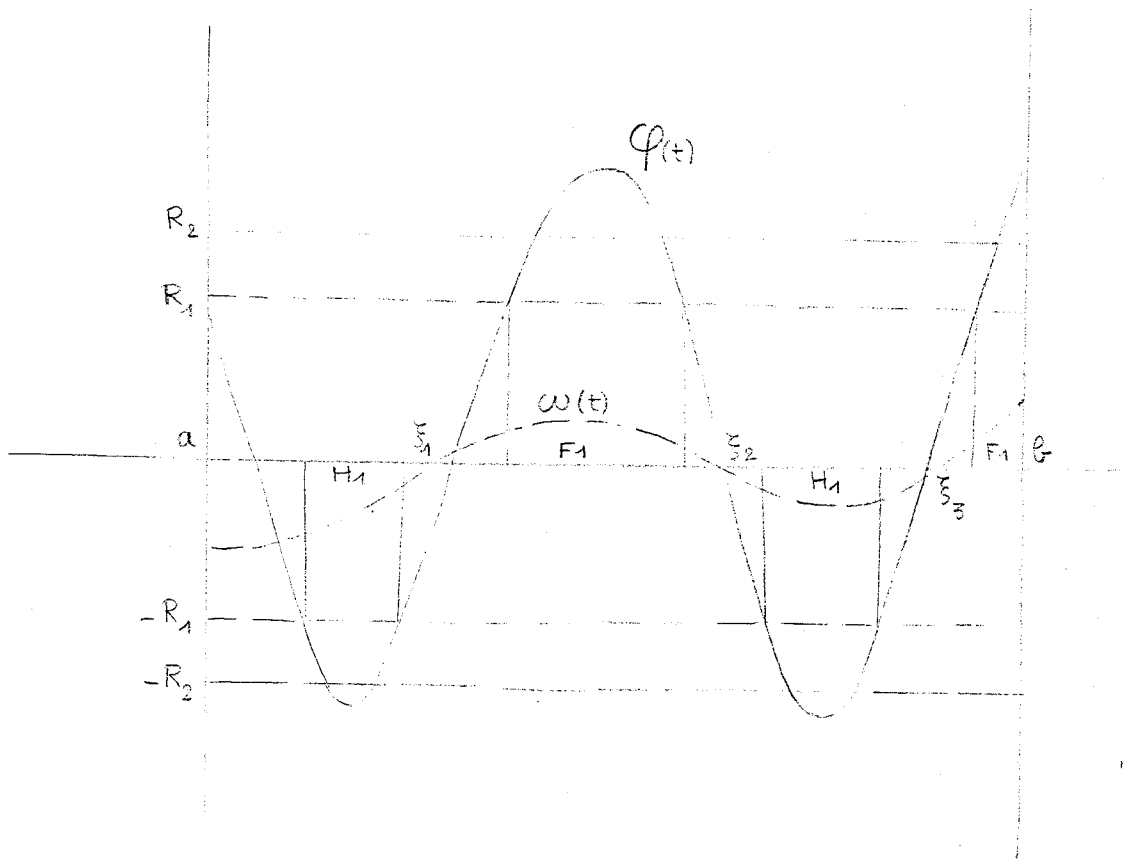
Soient F_1, H_1, F_2, H_2 les F et H correspondant à R_1 et R_2.

Soient $I_1, I_2, \ldots, I_m$ $m \leqslant n$ les intervalles correspondant à R_1.

On cherche un polynôme de correction $\omega(t)$ tel que :

$$\max_{t \in [a,b]} (|f(t)-p_n(t)-\omega(t)|) - \rho \leqslant \Theta \times (\max_{t \in [a,b]} (|f(t)-p_n(t)|) - \rho)$$

avec $\Theta < 1$.



On détermine un polynôme ω (t) qui soit $\geq$ o sur F_1 et négatif sur H_1. On peut trouver des points ξ_1, ..., ξ_m tels que respectivement ξ_1, ..., ξ_m appartiennent aux intervalles I_1, ..., I_m et tels que la distance minimum à l'ensemble $F_2 \cup H_2$ soit supérieure à un nombre $\delta > 0$. (Cela est possible en prenant par exemple ξ_i au milieu de l'intervalle I_i, bien que cette détermination ne soit pas toujours la meilleure pour l'utilisation de l'algorithme).

Un polynôme de degré $\leq n$ qui s'annule en ξ_1, ..., ξ_m est de la forme $\alpha \times q_{n-m}(t) \times (t - \xi_1) \times \ldots \times (t - \xi_m)$, α étant un scalaire et $q_{n-m}(t)$ un polynôme de degré $\leq n - m$.

(Dans le cas où il n'y a aucun intervalle I alors on prendra :

$$\alpha \times q_n(t)).$$

On considère comme polynôme $q_{n-m}(t)$ (de degré $\leq$ n-m) un polynôme qui garde un signe constant sur l'ensemble $F_1 \cup H_1$ et tel que :

$$\min_{t \in F_2 \cup H_2} | q_{n-m}(t) | \; / \; \max_{t \in [a,b]} | q_{n-m}(t) | \geq r > 0.$$

(Un tel polynôme existe puisque l'on peut prendre $q_{n-m}(t) = 1$)

Alors le polynôme $q_{n-m}(t) \times (t - \xi_1) \times .... \times (t - \xi_m)$ est un polynôme de degré $\leq n$ qui garde sur F_1 un signe constant et sur H_1 le signe opposé.

On considère un polynôme de la forme $\omega(t) = \alpha \times q_{n-m}(t) \times (t-\xi) \times \ldots \times (t-\xi_m)$ le signe de α étant tel que $\omega(t) \geq 0$ pour $t \in F_1$ et par suite $\omega(t) < 0$ pour $t \in H_1$. On pose ϵ = signe (α) . Alors il existe $\alpha_o > 0$ tel que pour $|\alpha| = \alpha_o$ si

$$\omega(t) = \epsilon \times \alpha_o \times q_{n-m}(t) \times (t-\xi_1) \times \ldots \times (t-\xi_m)$$

On ait :

$$\max_{t \in [a,b]} (|f(t) - p_n(t) - \omega(t)|) - \rho \leq \theta \times (\max_{t \in [a,b]} (|f(t) - p_n(t)|) - \rho)$$

avec $\theta < 1$.

En effet, considérons $\alpha_o > 0$ tel que $\max_{t\in[a,b]} |\omega(t)| = R_2 - R_1$

Alors :

$$|\alpha_o| \geqslant \frac{R_2 - R_1}{(b-a)^m \times \max_{t\in[a,b]} |q_{n-m}(t)|} > 0$$ et pour tout

$t \in [a,b] - F_1 UH_1$ on a $|\varphi(t)| < R_1$ donc :

$$|\varphi(t) - \omega(t)| \leqslant |\varphi(t)| + |\omega(t)| < R_1 + R_2 - R_1 = R_2.$$

c'est-à-dire :

$$\max_{t\in[a,b]-F_1UH_1} |\varphi(t) - \omega(t)| < R_2.$$

Si on ajoute à $p_n(t)$ le polynôme $\omega(t)$ et si on considère :

$$L_1 = \|\varphi - \omega\| = \max_{t\in[a,b]} |\varphi(t) - \omega(t)|$$

alors :

Sur l'ensemble $[a,b] - F_1 \cup H_1$

$$L - \max_{t\in[a,b]-F_1UH_1} |\varphi(t)-\omega(t)| > L-R_2=(1-\theta_2)(L-A) \geqslant (1-\theta_2)(L-\rho).$$

Sur l'ensemble $F_1UH_1-F_2UH_2$, $\varphi(t)$ et $\omega(t)$ sont de même signe et l'on a :

$$R_1 \leqslant |\varphi(t)| < R_2 \quad \text{et} \quad |\omega(t)| \leqslant R_2 - R_1$$

Si $\varphi(t) > 0$

$$R_2 \geqslant \varphi(t) - \omega(t) \geqslant \varphi(t) - R_2 + R_1 \geqslant -R_2$$

Si $\varphi(t) < 0$

$$-R_2 < \varphi(t) - \omega(t) \leqslant \varphi(t) + R_2 - R_1 < R_2$$

$$L - \max_{t\in F_1UH_1-F_2UH_2} |\varphi(t) - \omega(t)| > L-R_2 = (1 - \theta_2)(L-A) \geqslant (1-\theta_2)(L-\rho).$$

Sur l'ensemble F_2UH_2, $|\omega(t)| \leqslant R_2-R_1$ et $|\varphi(t)| \geqslant R_2$

$$|\varphi(t)| - |\omega(t)| \geqslant R_1 > 0 \quad \text{donc} \quad |\varphi(t)| > |\omega(t)|$$

et $\varphi(t)$ et $\omega(t)$ sont de même signe

$$\text{L-max}_{t\in F_2UH_2} |\varphi(t)-\omega(t)| > |\alpha_o| \times \text{Min}_{t\in F_2UH_2} |q_{n-m}(t)| \times \text{Min}_{t\in F_2UH_2} |(t-\xi_1)\ldots(t-\xi_m)|$$

$$\geqslant \frac{R_2-R_1}{(b-a)^m \max_{t\in[a,b]} |q_{n-m}(t)|} \times r \times \max_{t\in[a,b]} |q_{n-m}(t)| \times \delta^m$$

$$\geqslant \frac{R_2-R_1}{(b-a)^m} \times r \times \delta^m \geqslant (\theta_2-\theta_1)(L-A)r \left(\frac{\delta}{b-a}\right)^m$$

$$\geqslant (\theta_2-\theta_1) \times r \times \left(\frac{\delta}{b-a}\right)^n \quad (L-\rho).$$

Alors on peut écrire :

$$L_1 - A < \theta \times (L-A)$$

$$L_1 - \rho < \theta \times (L-\rho)$$

avec :

$$\theta = \text{Max}\left(\theta_2,\ 1 - (\theta_2-\theta_1)\, r \left(\frac{\delta}{b-a}\right)^m\right).$$

2. Etude de la convergence du premier algorithme de Remez.

Montrons alors que par itération du processus pour θ_1 et θ_2 fixés la méthode est convergente.

Propriété :

La longueur des intervalles I_m est bornée inférieurement quelle que soit l'itération.

En effet, soit $]\alpha,\beta[$ un intervalle ouvert I_m.

($\varphi(\alpha) = R_1$ et $\varphi(\beta) = -R_1$ ou inversement). Alors il existe 1 dépendant de la fonction f, du degré n, de l'intervalle $[a,b]$, de R_1 et de L, tel que $\beta-\alpha \geqslant l \geqslant 0$.

La fonction f étant uniformément continue sur $[a,b]$, pour tout $\varepsilon > 0$, il existe $\eta(\varepsilon)$ tel que si $|t_1 - t_2| < \eta(\varepsilon)$

$$|f(t_1) - f(t_2)| < \varepsilon.$$

Puisque $p_n(t) = f(t) - \varphi(t)$, on a :

$$\|p_n\| \leqslant \|f\| + \|\varphi\|$$

et si on pose $m = \|f\|$

$$\|p_n\| \leqslant m + L.$$

D'après un théorème de Markov on sait que :

$$\|p'_n\| \leqslant \frac{2 \times (m+L)}{(b-a)} \times n^2$$

et par suite :

$$|p_n(\alpha) - p_n(\beta)| \leqslant 2(m+L)\, n^2 \frac{(\beta-\alpha)}{(b-a)}$$

Alors on ne peut avoir à la fois

$$\beta - \alpha < \eta(R_1)$$

et

$$\beta - \alpha < \frac{(b-a)\, R_1}{2(m+L)\, n^2}$$

sinon d'après la première inégalité $|f(\alpha) - f(\beta)| < R$, d'après la seconde inégalité

$$| p_n(\alpha) - p_n(\beta) | < R_1.$$

Ce qui entraine $|\varphi(\alpha) - \varphi(\beta)| < 2 R_1$ qui est en contradiction avec l'hypothèse.

Alors $\beta - \alpha \geqslant 1 = \min \left(\eta(R_1), \dfrac{(b-a) \times R_1}{2 (m+L) n^2}\right)$

Or $\eta(R_1)$ ne peut tendre vers zéro que si $R_1 \longrightarrow o$.

1 ne peut tendre vers zéro que si $R_1 \longrightarrow o$.

Les longueurs des intervalles I_m sont donc bornées inférieurement.

Il est alors possible de choisir $\delta > o$ de façon indépendante de l'itération.

Si à la $p^{ème}$ itération $L_p = \max_{t \in [a,b]} |\varphi_p(t)|$, L_p converge vers ρ et la rapidité de la convergence est supérieure à celle d'une progression géométrique de raison $\theta < 1$.

Remarque :

Dans le cas général de l'approximation d'une fonction f par des éléments d'un sous espace V (autre que celui des polynômes) le principe du premier algorithme reste valable mais la détermination des points ξ_i par exemple conduit à des difficultés pratiques.

3. Relation avec la méthode de décomposition de la norme.

Soit $p_n(t)$ une approximation de la fonction f(t) qui ne soit pas la meilleure approximation

$$p_n(t) = x^o_1 \; t^n + ... + x^o_{n+1}$$

avec :

$$\max_{t \in [a,b]} |f(t) - p_n(t)| = L$$

Soit x_o le vecteur de R^{n+1} $\begin{vmatrix} x^o_1 \\ \vdots \\ x^o_{n+1} \end{vmatrix}$

Comme pour le chapitre III on considère l'ensemble $\mathcal{C}_{x_o}$ des vecteurs

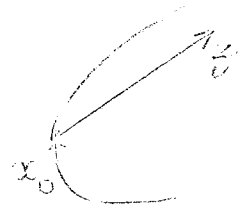
$$\eta \in R^{n+1} \qquad \eta \left| \begin{array}{l} \eta_1 \\ \vdots \\ \eta_{n+1} \end{array} \right. \qquad \text{tels que}$$

$\| f(t) \ (\eta_1 \ t^n + ... + \eta_{n+1}) \| \leqslant L$. Nous avons vu que $\mathcal{C}_{x_o}$ est un ensemble convexe.

Si l'on pose :

$$z_1 \ t^n + ... + z_{n+1} = \epsilon \ x \ q_{n-m} \ (t) \ (t - \xi_1) \ x \ .... \ x \ (t - \xi_m)$$

alors $\left| \begin{array}{l} z_1 \\ \vdots \\ z_{n+1} \end{array} \right.$ fournit une direction qui pénètre dans l'ensemble $\mathcal{C}_{x_o}$



Pour un $\lambda = \lambda_1 \geqslant o$ on a :

$$\underset{t \in [a,b]}{\text{Max}} \ \| f(t) - p_n(t) - \lambda_1 (z_1 \ t^n + ... + z_{n+1}) \| = L$$

Alors il existe un λ_o tel que :

$$\| f(t) - p_n(t) - \lambda_o (z_1 \ t^n + ... + z_{n+1}) \| = \inf_{\lambda \in [o, \lambda_1]} \| f(t) - p_n(t) - \lambda (z_1 t^n + .. + z_{n+1}) \|$$

On peut alors déterminer λ_o par dichotomie.

4. Application numérique

Lorsque $m < n$ on a :

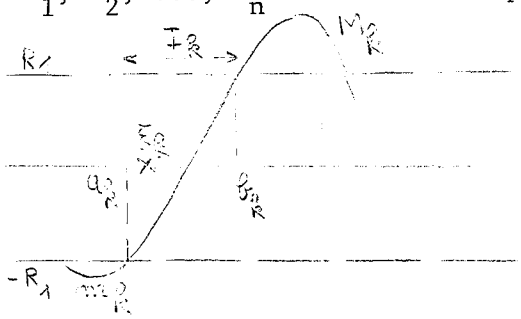
$\omega(t) = \alpha \ q_{n-m} \ (t) \ (t-\xi_1) \ x \ ... \ x \ (t-\xi_m)$ On peut alors prendre les points $\xi_1, \ ..., \ \xi_m$ au milieu des intervalles $I_1, \ ..., \ I_m$.

Pour $q_{n-m}(t)$ on pourrait poser $q_{n-m}(t) = 1$, mais la correction porterait alors uniquement sur les $m + 1$ coefficients de $p_n(t)$ de plus bas degré. On essaiera alors de trouver un polynôme q_{n-m} de degré le plus élevé possible ($d^o\ q_{n-m}(t) \leqslant n-m$) et qui emplifie la correction dans le voisinage du maximum de l'erreur en valeur absolue.

Lorsque $m = n$ on a :

$$\omega(t) = \alpha\,(t - \xi_1) \times \ldots \times (t - \xi_n)$$ et la direction $\begin{vmatrix} z_1 \\ \vdots \\ z_{n+1} \end{vmatrix}$

ne dépend que de la position des points $\xi_1, \ldots, \xi_n$. On tentera alors d'améliorer cette direction en ne prenant plus les points $\xi_1, \ldots, \xi_n$ au milieu des intervalles $I_1, I_2, \ldots, I_n$ mais en les déplaçant de la manière suivante :



Soit I_k le $k^{ème}$ intervalle a_k et b_k ses bornes ($a_k < b_k$). On prend alors ξ_k du coté du plus maximum m_k, par rapport au centre de l'intervalle I_k.

Exemple :

Approximation de e^t par un polynôme de degré 5

$x_1 t^5 + x_2 t^4 + x_3 t^3 + x_4 t^2 + x_5 t + x_6$ dans l'intervalle $[0,1]$

Vecteur de départ

$x_1^o = 0,\ 012\ 992\ 350$

$x_2^o = 0,\ 036\ 803\ 726$

$x_3^o = 0,\ 168\ 888\ 69$

$x_4^o = 0,\ 499\ 556\ 87$

$x_5^o = 1,\ 000\ 030\ 3$

$x_6^o = 0,\ 999\ 999\ 67$

$L_o = \| p^o_5(t) - e^t \| = \underset{t\in[0,1]}{\text{Max}} \left| p^o_5(t) - e^t \right| = \underline{0,\ 000\ 013\ 411}$

Soit $\varphi_o(t) = e^t - (x^o_1\, t^5 + x^o_2\, t^4 + x_3^o\, t^3 + x^o_4\, t^2 + x^o_5 t + x^o_6)$

On a pour les points a et b et pour les maximums de l'erreurs

$\varphi_o(o) = 0,\ 000\ 000\ 33$

$\varphi_o(o,o5) = -0,\ 000\ 000\ 313$

$\varphi_o(o,19) = 0,\ 000\ 000\ 596$

$\varphi_o(o,43) = -0,\ 000\ 000\ 299$

$\varphi_o(o,62) = 0,\ 000\ 000\ 536$

$\varphi_o(o,857) = -0,\ 000\ 003\ 78$

$\varphi_o(1) = 0,\ 000\ 013\ 41$

$A_o = 0,\ 000\ 000\ 3$

Si l'on prend $\theta_1 = 0,\ 1$

$R_1 = 0,\ 000\ 000\ 3 + 0,\ 1\ (0,\ 000\ 013\ 4 - 0,\ 000\ 000\ 3)$

$\simeq 0,\ 000\ 001\ 6$

On a :

$\varphi_o(0,\ 96) \simeq R_1$

$\varphi_o(0,\ 93) \simeq -R_1$

$\varphi_o(0,\ 76) \simeq -R_1$

On prend $\xi_1 = \dfrac{0,\ 96 + 0,\ 93}{2} = 0,\ 945$

$q_{n-m}(t) = t^3 \times (t - 0,\ 685) \qquad (\varphi_o(0,\ 685) \simeq 0)$

Alors

$\omega(t) = \alpha\, t^3\ (t - 0,\ 685)\ (t - 0,\ 945)$

$= \alpha\ (t^5 - 1,63\ t^4 + 0,\ 65\ t^3)$

avec $\alpha > 0$.

Comme au chapitre précédent on se déplace par dichotomie à partir du

point $\left|\begin{array}{l} x^o_1 \\ \cdot \\ \cdot \\ \cdot \\ x^o_6 \end{array}\right.$ suivant la direction $\left|\begin{array}{l} 1 \\ -1,63 \\ 0,65 \\ 0 \\ 0 \\ 0 \end{array}\right.$

On trouve :

$\alpha = 0,\ 000\ 4$

$x^1_1 = 0,\ 013\ 392\ 352$

$x^1_2 = 0,\ 036\ 151\ 722$

$x^1_3 = 0,\ 169\ 148\ 68$

$x^1_4 = 0,\ 499\ 556\ 87$

$x^1_5 = 1,\ 000\ 030\ 3$

$x^1_6 = 0,\ 999\ 999\ 67$

$L_1 = \| P^1_5(t) - e^t \| = 0,\ 000\ 004\ 64$

$A_1 = 0 \qquad \theta_1 = 0,\ 4$

$R_1 = \theta_1 \times L_1 \simeq 0,\ 000\ 001\ 8$

$\varphi_1(0,27) = R_1$

$\varphi_1(0,61) = R_1$

$\varphi_1(0,99) = -R_1$

alors on prend $\xi_1 = \dfrac{0,61 + 0,99}{2} = 0,\ 80$

$q_{n-m}(t) = t^3 \times (t - 0,\ 125)$, avec $\alpha > o$ $\qquad (\varphi_1(0,125) \simeq o)$

on se déplace par dichotimie suivant la direction

1
- 0, 925
0,1
0
0
0

On trouve $\alpha = 0,\ 000\ 02$

$x^2_1 = 0,\ 013\ 414\ 351$

$x^2_2 = 0,\ 036\ 131\ 371$

$x^2_3 = 0,\ 169\ 150\ 88$

$x^2_4 = 0,\ 499\ 556\ 87$

$x^2_5 = 1,\ 000\ 030\ 3$

$x^2_6 = 0,\ 999\ 999\ 67$

$\underline{L_2 = 0,\ 000\ 004\ 43}$

$A_2 = 0$ $\qquad\qquad$ $\theta_1 \simeq 0,\ 2$

$R_1 = 0,\ 000\ 000\ 8$

$\varphi_2\ (0,\ 22) = R_1$

$\varphi_2\ (0,\ 64) = R_1$

$\varphi_2\ (0,\ 72) = -R_1$

$\varphi_2\ (0,\ 77) = -R_1$

$\varphi_2\ (0,\ 84) = R_1$

On prend $\xi_1 = 0,\ 70$ $\qquad$ $\xi_2 = 0,\ 78$ $\qquad$ et $q_{n-m}\ (t) = t - 0,2$

Alors :

$$\omega(t) = \alpha\ (t-0,2) \times (t - 0,70) \times (t - 0,\ 78) \qquad \text{avec } \alpha < 0$$

on se déplace par dichotomie suivant la direction

$$\begin{vmatrix} 0 \\ 0 \\ -1 \\ 1,68 \\ -0,84 \\ 0,11 \end{vmatrix}$$

$x^3_1 = 0,\ 013\ 414\ 353$

$x^3_2 = 0,\ 036\ 131\ 367$

$x^3_3 = 0,\ 169\ 114\ 58$

$x^3_4 = 0,\ 499\ 617\ 82$

$x^3_5 = 0,\ 999\ 999\ 73$

$x^3_6 = 1,\ 000\ 003\ 6$

$L_3 = 0,\ 000\ 003\ 57$

Une simple translation suivant le vecteur $\begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -1 \end{vmatrix}$ permet d'obtenir

$x^4_1 = 0,\ 013\ 414\ 353$

$x^4_2 = 0,\ 036\ 131\ 367$

$x^4_3 = 0,\ 169\ 114\ 58$

$x^4_4 = 0,\ 499\ 617\ 82$

$x^4_5 = 0,\ 999\ 999\ 73$

$x^4_6 = 1,\ 000\ 002\ 6$

$L_4 = 0,\ 000\ 002\ 57$

$A_4 = 0$

$R_1 = 0,\ 000\ 001 \qquad \theta_1 \simeq 0,\ 4$

$\varphi_4\ (0,07) = R_1$

$\varphi_4\ (0,33) = R_1$

$\varphi_4\ (0,58) = R_1$

$\varphi_4\ (0,67) = -R_1$

$\varphi_4\ (0,82) = -R_1$

$\varphi_4\ (0,90) = R_1$

$\varphi_4\ (0,96) = R_1$

$\varphi_4\ (0,99) = -R_1$

On prend $\xi_1 = 0,\ 625 \quad \xi_2 = 0,86 \quad \xi_3 = 0,\ 975$

et $q_{n-m}\ (t) = (t - 0,2)^2$

$$\omega(t) = \alpha\ (x - 0,2)^2\ (x - 0,\ 675)\ (x - 0,86)\ (x - 0,975)$$

on se déplace par dichotimie suivant la direction :

1
-2,86
3,01
-1,41
0,29
-0,021

$x^5_1 = 0,\ 013\ 634\ 353$

$x^5_2 = 0,\ 035\ 502\ 158$

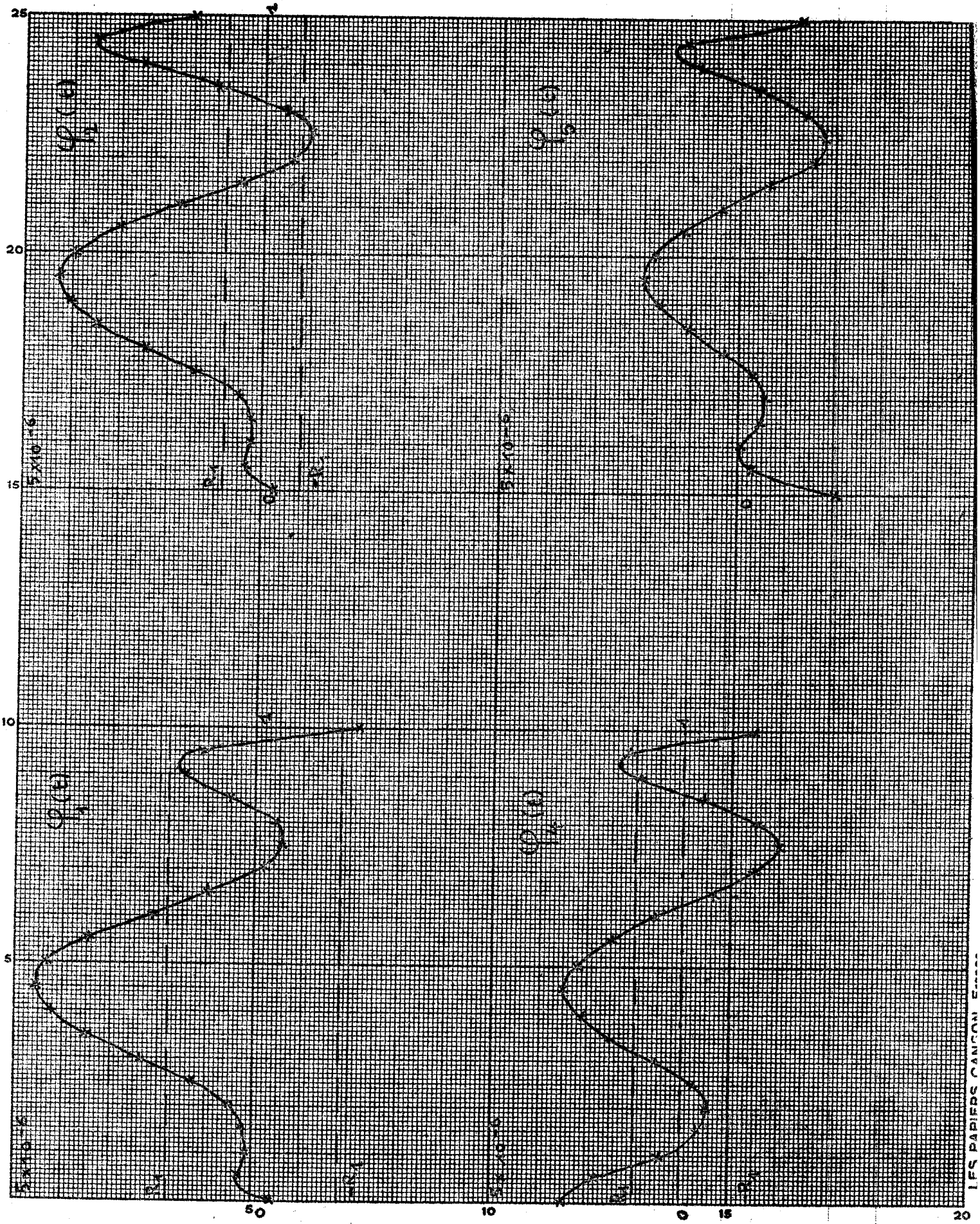
$x^5_3 = 0,\ 169\ 776\ 64$

$x^5_4 = 0,\ 499\ 306\ 17$

$x^5_5 = 1,\ 000\ 063\ 3$

$x^5_6 = 0,\ 999\ 997\ 97$

$L_5 = 0,\ 000\ 002\ 0$


25
20
15
10
5
0
5
10
15
20
5x10-3
5x10-4
5x10-5
5x10-6
LES PAPIERS CANSON

CHAPITRE V

-=-=-=-

ETUDE DU PROBLEME DISCRET

METHODE DE STIEFEL DANS LE CAS DES POLYNOMES

On rappelle les théorèmes démontrés par Stiefel (1) qui par l'introduction de la notion de référence permettent la caractérisation de la meilleure approximation dans le cas d'un ensemble de points, discret.

De plus ces résultats conduisent à une méthode numérique pour l'approximation par des polynômes de degré $\leqslant n$ qui permet la comparaison avec les résultats numériques obtenus à l'aide des autres méthodes et à l'introduction des méthodes de programmation linéaire.

Soit f(t) une fonction continue à valeurs réelles, définie sur un intervalle $[a,b]$.

Soient m points ($i = 1, \ldots, m$), $t_i \in [a,b]$, $t_i \neq t_j$ pour $i \neq j$ et $m \geqslant n+1$.

On considère n fonctions continues $f_j(t)$ ($j = 1, \ldots, n$) sur l'intervalle $[a,b]$ qui engendrent une variété linéaire V de $\mathcal{C}(a,b)$.

Le problème est de trouver un ensemble de n paramètres $x_1, \ldots, x_n$ (s'il existe) tel que :

$$\max_{i=1,\ldots,n} | x_1 f_1(t_i) + \ldots + x_n f_n(t_i) - f(t_i) | \text{ soit minimal.}$$

D'après le théorème I du chapitre I on sait qu'un tel ensemble $x_1, \ldots, x_n$ existe.

1. Caractérisation et unicité de la solution.

On suppose que les fonctions $f_1(t), \ldots, f_n(t)$ vérifient la condition d'interpolation sur l'ensemble discret des m points t_i c'est-à-dire :

Pour tout ensemble de n abscisses t_σ (distinctes) prises parmi les m points t_i, si

$$g(t) = x_1 f_1(t) + \ldots + x_n f_n(t)$$

alors il existe un g (t) et un seul tel que $f(t_\sigma) = g(t_\sigma)$ pour $\sigma = 1, \ldots, n$.

(1) Voir STIEFEL [12]

Définition

On appelle référence $[t_\sigma]$ un ensemble de n + 1 abscisses distinctes parmi les m points t_i.

Théorème I :

Il existe $\lambda_1, \ldots, \lambda_{n+1}$ tels que :

$$\lambda_1 g(t_1) + \ldots + \lambda_{n+1} g(t_{n+1}) = 0 \quad \text{pour } g \in V.$$

On considère la matrice A telle que :

$$a_{ij} = f_j(t_i) \text{ pour } j = 1, \ldots, n \; ; \; i = 1, \ldots, n+1.$$

n lignes quelconques de la matrice A sont linéairement indépendantes d'après la propriété d'interpolation.

Alors il existe un vecteur $\begin{vmatrix} \lambda_1 \\ \vdots \\ \lambda_{n+1} \end{vmatrix}$ tel que

$$(\lambda_1, \ldots, \lambda_{n+1}) \times A = 0 \qquad \text{d'où :}$$

$$\lambda_1 g(t_1) + \ldots + \lambda_{n+1} g(t_{n+1}) = 0 \qquad \text{pour tout } g \in V.$$

C'est la relation caractéristique.

Définition :

$g(t) \in V$ est appelée fonction de référence relative à la référence $[t_\sigma]$ si :

$$\text{signe } h_\sigma = \text{signe } \lambda_\sigma \quad \text{ou bien}$$

$$\text{signe } h_\sigma = - \text{ signe } \lambda_\sigma \quad \text{pour tout } \sigma.$$

Si l'on pose $h_\sigma = g(t_\sigma) - f(t_\sigma)$

Alors d'après la relation caractéristique :

$$\lambda_1 [f(t_1) + h_1] + \lambda_2 [f(t_2) + h_2] + \ldots + \lambda_{n+1} [f(t_{n+1}) + h_{n+1}] = 0.$$

$$\sum_{\sigma=1}^{n+1} \lambda_\sigma h_\sigma = - \sum_{\sigma=1}^{n+1} \lambda_\sigma f(t_\sigma) \qquad (1)$$

$$\sum_{\sigma=1}^{n+1} |\lambda_\sigma| |h_\sigma| = \pm \sum_{\sigma=1}^{n+1} \lambda_\sigma f(t_\sigma)$$

Définition :

g(t) est appelée "meilleure fonction de référence" relative à une référence $[t_\sigma]$ si pour tout σ :

$$h_\sigma = h \operatorname{sgn} \lambda_\sigma$$

$|h|$ est appelée "la meilleure déviation de référence".

$$h = - \frac{\sum \lambda_\sigma f_\sigma}{\sum |\lambda_\sigma|}$$

Si g(t) est une fonction de référence quelconque :

$$h_\sigma = g(t_\sigma) - f(t_\sigma)$$

$$h = - \frac{\sum \lambda_\sigma [g(t_\sigma) - h_\sigma]}{\sum |\lambda_\sigma|} = - \frac{\sum \lambda_\sigma h_\sigma}{\sum |\lambda_\sigma|}$$

Comme signe h_σ = signe λ_σ ou signe h_σ = - signe λ_σ pour tout σ

$$h = \frac{\sum |\lambda_\sigma| |h_\sigma|}{\sum |\lambda_\sigma|}$$

Théorème II .

La meilleure déviation de référence relative à la référence $[t_\sigma]$ est la moyenne pondérée des erreurs $|h_\sigma|$ de toute fonction de référence avec pour coefficients des nombres positifs (tous non nuls) alors :

$$\min_\sigma |h_\sigma| \leqslant |h| \leqslant \max_\sigma |h_\sigma| .$$

Dans le cas où g (t) n'est plus une fonction de référence alors :

$$|h| \leqslant \max_\sigma |h_\sigma| .$$

Corollaire :

g(t) étant une fonction de référence si l'on a l'égalité $\max_\sigma |h_\sigma| = |h|$ alors g(t) coïncide avec la meilleure fonction de référence à $[t_\sigma]$.

La meilleure déviation de référence relative à la référence $[t_\sigma]$ est telle que :

$$|h| = \frac{\sum_\sigma |\lambda_\sigma| |h_\sigma|}{\sum_\sigma |\lambda_\sigma|}$$

On a :

$$0 = \operatorname{Max}_{\sigma} \; |h_\sigma| - |h| = \frac{\sum_{\sigma} |\lambda_\sigma| \times (\max_{\sigma} |h_\sigma| - |h_\sigma|)}{\sum_{\sigma} |\lambda_\sigma|}$$

donc :

$$|h_\sigma| = \max_{\sigma} |h_\sigma| = |h| \qquad \text{pour tout } \sigma.$$

Soient m abscisses $t_1 < t_2 < \ldots. < t_m$. $(n + 1 < m)$.

Soit g(t) une approximation avec $h_i = g(t_i) - f(t_i)$ $(i = 1, \ldots, m)$.

Théorème III. Théorème d'échange.

Soit une référence $[t_\sigma]$ et une fonction de référence liée à $[t_\sigma]$.

Soit t_i un point appartenant à l'ensemble des m abscisses $(t_1, \ldots, t_m)$ mais n'appartenant pas à la référence $[t_\sigma]$.

Alors il existe t_ρ appartenant à $[t_\sigma]$ tel que g (t) est une fonction de référence relative à la référence formée par les points de $[t_\sigma]$ tels que $t_\sigma \neq t_\rho$ et le point t_i.

La méthode d'échange

On considère une référence $[t_\sigma]$. On construit g (t) la meilleure fonction de référence liée à $[t_\sigma]$.

Pour i = 1, ..., m les h_i sont tels que :

$$\mu = \max_{i=1}^{m} |h_i| \geqslant h.$$

Si $\mu = |h|$ alors on arrête le procédé.

Sinon, soit t_i un point pour lequel l'erreur prend la valeur μ en valeur absolue. Alors d'après le théorème précédent, il existe une nouvelle référence $[t^*_\sigma]$ contenant le point t_i et telle que g(t) soit encore une fonction de référence relative à $[t^*_\sigma]$.

Pour les erreurs h^*_σ, en n points on a $|h^*_\sigma| = |h|$ et en un point l'erreur est égale à μ.

On construit alors la meilleure fonction de référence relative à $[t^*_\sigma]$; soit h^* la déviation correspondante.

D'après le théorème II (h^* étant la moyenne pondérée) on a :

$$|h| < |h^*| \leqslant \mu.$$

Si $|h^*| < \mu$ on itère le procédé en construisant une nouvelle référence.

On ne peut prendre deux fois la même référence du fait que l'on a l'inégalité stricte $|h| < |h^*|$ pour deux références successives.

La méthode s'arrête après un nombre fini d'échanges puisque le nombre de références $[t_\sigma]$ que l'on peut prendre dans l'ensemble $t_1, \ldots, t_m$ est un nombre fini.

Supposons que la méthode s'arrête pour une référence $[t_\sigma]$ et pour la meilleure fonction de référence g (t).

Soit H la déviation de référence et H_i (i = 1, ..., m) les erreurs aux points t_i.

Alors Max $|H_i| = H$.

La fonction g(t) ainsi trouvée est une solution du problème de Tschebyscheff. En effet :

Soit $\tilde{g}$ (t) une fonction quelconque appartenant au sous espace engendré par $f_1, \ldots, f_n$ ayant pour erreur $\tilde{h}_i$ aux points t_i. Alors :

$$\operatorname*{Max}_{i=1,\ldots,m} |\tilde{h}_i| \geqslant \operatorname{Max} |\tilde{h}_\sigma| \geqslant H = \operatorname{Max} |H_i|$$

De plus Stiefel a démontré (1) que la méthode d'échange est équivalente à un problème de programmation linéaire (cf chapitre VI). Cette méthode permet d'obtenir la meilleure approximation sur un ensemble discret lorsque l'on a pas l'alternance.

(Dans le cas où il y a alternance la méthode d'échange se ramène à la méthode développée au chapitre II adaptée au cas discret).

(1) Voir Stiefel [11]

Théorème IV

Si l'ensemble des n fonctions $f_1, \ldots, f_n$ possède la propriété d'interpolation sur l'ensemble des m points t_i, la meilleure approximation est unique. Elle est caractérisée par le fait que c'est la meilleure fonction de référence liée à une certaine référence $[t_\sigma]$, telle que le maximum de la valeur absolue de l'erreur (sur l'ensemble des m points t_i) soit égal à la déviation de référence.

(On peut montrer d'une façon analogue au théorème II du chapitre I que la condition d'interpolation sur l'ensemble des m points t_i est une condition nécessaire et suffisante pour que la solution soit unique).

Remarque 1 :

On voit que dans le cas général la meilleure approximation sur un ensemble discret de m points (même lorsqu'elle est unique) peut ne pas avoir la propriété d'alternance.

Par contre si l'on a m points dans un intervalle $[a,b]$ et que les fonctions $f_1, \ldots, f_n$ vérifient la condition de Haar sur tout l'intervalle $[a,b]$ d'après le § 3 du chapitre II, la meilleure approximation a la propriété d'alternance.

Remarque 2 :

On peut démontrer quela fonction g(t) telle que $\sum_1^m h^2_i$ soit minimal est une fonction de référence (1). Elle peut donc fournir une fonction de départ pour la méthode.

2. Etude de l'approximation discrète lorsqu'il y a alternance.

On suppose que l'on a m points $(t_1, \ldots, t_m)$ dans un intervalle $[a,b]$ et que les fonctions $f_1, \ldots, f_n$ vérifient la condition de Haar sur tout l'intervalle $[a,b]$. On sait alors que la meilleure approximation a la propriété d'alternance.

On peut alors employer la méthode donnée au chapitre II adaptée au cas discret.

Exemple : Approximation de $\frac{\sin t}{t}$ par un polynôme pair de degré 6 sur l'ensemble des points $t_i = i \times \frac{\pi}{20}$ pour $i = 0, 1, \ldots, 20$.

Soit $x^*_1 t^6 + x^*_2 t^4 + \ldots + x^*_4$ la meilleure approximation.

(1) Voir STIEFEL [10]

On trouve :

$$\lambda = 0,\ 000\ 000\ 73 \leqslant \rho \leqslant 0,\ 000\ 000\ 76$$

$$x^*_1 = -\ 0,\ 000\ 185\ 304\ 47$$

$$x^*_2 = \quad 0,\ 008\ 313\ 477\ 6$$

$$x^*_3 = -\ 0,\ 166\ 656\ 99$$

$$x^*_4 = \quad 0,\ 999\ 999\ 24$$

3. Approximation par un polynôme de degré $\leqslant n - 1$

Si on prend pour fonctions de base $f_1, \ldots, f_n$ les puissances de t : $t^{n-1},..,t,1$ la relation caractéristique peut s'écrire :

$$\sum_{\sigma} \frac{g\ (t_\sigma)}{\prod_{\tau \neq \sigma} (x_\sigma - x_\tau)} = 0 \qquad \lambda_\sigma = \frac{1}{\prod_{\tau \neq \sigma} (x_\sigma - x_\tau)}$$

Si la suite des abscisses $[t_\sigma]$ est prise de gauche à droite par ordre croissant alors on retrouve le résultat que les λ_σ ont des signes alternés.

Dans le cas des polynômes, pour le calcul numérique de la meilleure fonction de référence relative à une référence $[t_\sigma]$ on peut utiliser la propriété que la différence divisée d'ordre n d'un polynôme de degré n - 1 est nulle.

On pose $h_\sigma = (-1)^\sigma\ h$

$$g\ (t_\sigma) = f\ (t_\sigma) + (-1)^\sigma\ h$$

Si l'on note $\delta_\sigma^{\,n}\ (f)$ la différence divisée d'ordre n de la fonction f sur l'ensemble des abscisses $[t_\sigma]$

$$\delta_\sigma^{\,n}\ (g) = 0 = \delta_\sigma^{\,n}\ (f) + h \times \delta^n_{\ \sigma}\ (-1)^\sigma$$

De là on tire :

$$h = -\ \frac{\delta^n_{\ \sigma}\ (f)}{\delta^n_{\ \sigma} (-1)^\sigma}$$

Connaissant h et les différences divisées d'ordre inférieur à n pour les fonctions f et $(-1)^{\sigma}$ on peut calculer les coefficients de g(ou utiliser la formule de Newton).

```
'PROCEDURE' CHEBFIT(M,N,X,Y,A) FF 'ENTIER' M,N FF
'REEL' 'TABLEAU'  X,Y,A FF
 'COMMENTAIRE' MEILLEURE APPROX D'UNE FCT Y(X)
 DONNEE PAR M COUPLES DE POINTS X,Y PAR UN
 POLYNOME DEGRE N FF
 'DEBUT' 'REEL' 'TABLEAU' T.(1FM).,AX.(1FN+2).,
 AY.(1FN+2).,AH.(1FN+2).,BY.(1FN+2).,BH.(1FN+2). FF
 'ENTIER' 'TABLEAU' IN.(1FN+2). FF 'REEL' TMAX,H FF
 'ENTIER' I,J,IMAX,K FF
 INITIALIZE F
 KF=(M-1)/(N+1) FF
 'POUR' IF=1 'PAS' 1 'JUSQUA' N+1 'FAIRE'
   IN.(I).F=(I-1)*K+1 FF IN.(N+2).F=M FF
 STARTF 'COMMENTAIRE' DEBUT ITERATION FF
 'POUR' IF=1 'PAS' 1 'JUSQUA' N+2 'FAIRE'
 'DEBUT' AX.(I).F=X.(IN.(I).). FF
    AY.(I).F=Y.(IN.(I).). FF
    AH.(I).F=(-1)**(I-1)
 'FIN' FF
 DIFFERENCEF 'COMMENTAIRE' DIFFERENCES DIVISEES FF
 'POUR' IF=2 'PAS' 1 'JUSQUA' N+2 'FAIRE'
 'DEBUT'
 'POUR' JF=I-1 'PAS' 1 'JUSQUA' N+2 'FAIRE'
 'DEBUT' BY.(J).F=AY.(J). FF
    BH.(J).F=AH.(J).
 'FIN' FF
 'POUR' JF=I 'PAS' 1 'JUSQUA' N+2 'FAIRE'
 'DEBUT' AY.(J).F=(BY.(J).-BY.(J-1).)/(AX.(J).-AX.(J-I+1).) FF
 AH.(J).F=(BH.(J).-BH.(J-1).)/(AX.(J).-AX.(J-I+1).)
 'FIN'
 'FIN' FF
 HF=-AY.(N+2)./AH.(N+2). FF
 POLYF 'COMMENTAIRE' COEF DU POLYNOME FF
 'POUR' IF=0 'PAS' 1 'JUSQUA' N 'FAIRE'
 'DEBUT' A.(I).F= AY.(I+1).+AH.(I+1).*H FF
   BY.(I+1).F=0 FF
 'FIN' FF
 BY.(1).F=1 FF TMAXF=ABS(H) FF IMAXF=IN.(1). FF
 'POUR' IF=1 'PAS' 1 'JUSQUA' N 'FAIRE'
 'DEBUT'
 'POUR' JF=0 'PAS' 1 'JUSQUA' I-1 'FAIRE'
 'DEBUT'
    BY.(I+1-J).F=BY.(I+1-J).-BY.(I-J).*X.(IN.(I).). FF
    A.(J).F=A.(J).+A.(I).*BY.(I+1-J).
    'FIN'
 'FIN' FF
 ERREURF 'COMMENTAIRE' CALCUL DEVIATION FF
 'POUR' IF=1 'PAS' 1 'JUSQUA' M 'FAIRE'
 'DEBUT' T.(I).F=A.(N). FF
    'POUR' JF=1 'PAS' 1 'JUSQUA' N 'FAIRE'
    T.(I).F=T.(I).*X.(I).+A.(N-J). FF
    T.(I).F=T.(I).-Y.(I). FF
 'SI' ABS(T.(I).)'INFEG'TMAX 'ALORS' 'ALLERA' L1 FF
 TMAXF=ABS(T.(I).) FF
 IMAXF=I                FF
```

```
 L1F 'FIN' FF
     'POUR' IF=1 'PAS' 1 'JUSQUA' N+2 'FAIRE'
'DEBUT'
'SI' IMAX'INFER'IN.(I). 'ALORS' 'ALLERA' L2 FF
'SI' IMAX=IN.(I). 'ALORS' 'ALLERA' FIT
'FIN' FF
L2F 'SI' T.(IMAX).*T.(IN.(I).).'INFER'0 'ALORS' 'ALLERA' L3 FF
   IN.(I).F=IMAX FF 'ALLERA' START FF
L3F 'SI' IN.(1).'INFER'IMAX 'ALORS' 'ALLERA' L4 FF
    'POUR' IF=1 'PAS' 1 'JUSQUA' N+1 'FAIRE'
    IN.(N+3-I).F=IN.(N+2-I). FF
    IN.(1).F=IMAX FF
    'ALLERA' START FF
L4F 'SI' IN.(N+2).'INFEG'IMAX 'ALORS' 'ALLERA' L5 FF
    IN.(I-1).F=IMAX FF
    'ALLERA' START FF
L5F 'POUR' IF=1 'PAS' 1 'JUSQUA' N+1 'FAIRE'
    IN.(I).F=IN.(I+1). FF
    IN.(N+2).F=IMAX FF
    'ALLERA' START FF
FITF
 'FIN' FF
```

* Source : Communications of the ACM

vol 5, N 5, Mai 1962

4. Application numérique.

Approximation de la fonction e^t (1) sur l'ensemble des points

$t_i = i/20$ pour $i = 0, \ldots, 20$ par un polynôme de degré $\leqslant 5$

$$p_5(t) = x_1 t^5 + x_2 t^4 + x_3 t^3 + x_4 t^2 + x_5 t + x_6.$$

On trouve :

$$|H| = 0,\ 000\ 001\ 090$$

$$x^*_1 = 0,\ 013\ 910\ 789$$

$$x^*_2 = 0,\ 034\ 779\ 367$$

$$x^*_3 = 0,\ 170\ 424\ 55$$

$$x^*_4 = 0,\ 499\ 086\ 37$$

$$x^*_5 = 1,\ 000\ 080\ 7$$

$$x^*_6 = 0,\ 999\ 998\ 90$$

D'après la tabulation on voit que :

$$\max_{i=0,\ldots,20} |e^{t_i} - p^*_5(t_i)| = 0,\ 000\ 001\ 192.$$

Pour la meilleure approximation ρ

$$0,\ 000\ 001\ 09 \leqslant \rho \leqslant 0,\ 000\ 001\ 12$$

(1) Pour e^t on a utilisé ici la fonction standard EXP (T) sur IBM 7044.

Tabulation de l'erreur.

t	$p^*_5(t) - e^t$
0	- 0, 000 001 095
0, 05	0, 000 001 073
0, 1	0, 000 000 969
0, 15	0, 000 000 045
0, 2	- 0, 000 000 760
0, 25	- 0, 000 001 118
0, 3	- 0, 000 000 939
0, 35	- 0, 000 000 387
0, 40	0, 000 000 268
0, 45	0, 000 000 820
0, 5	0, 000 001 058
0, 55	0, 000 000 909
0, 60	0, 000 000 402
0, 65	- 0, 000 000 253
0, 70	- 0, 000 000 864
0, 75	- 0, 000 001 162
0, 8	- 0, 000 000 924
0, 85	- 0, 000 000 149
0, 90	0, 000 000 775
0, 95	0, 000 000 983
1	- 0, 000 001 192

Autres exemples numériques:

Approximation de e^t sur l'ensemble des points

$t_i = i/20$ pour $i = 0, \ldots, 20$ par un polynôme :

degré $\leqslant 3$ $\quad p_3(t) = x_1 t^3 + x_2 t^2 + x_3 t + x_4$

On trouve :

$0,\ 000\ 543\ 18 \leqslant \rho \leqslant 0,\ 000\ 543\ 24$

$x^*_1 = 0,\ 279\ 940\ 50$
$x^*_2 = 0,\ 421\ 767\ 80$
$x^*_3 = 1,\ 016\ 573\ 5$
$x^*_4 = 0,\ 999\ 456\ 81$

degré $\leqslant 4$ $\quad p_4(t) = x_1 t^4 + x_2 t^3 + x_3 t^2 + x_4 t + x_5$

$0,\ 000\ 027\ 07 \leqslant \rho \leqslant 0,\ 000\ 027\ 15$

$x^*_1 = 0,\ 069\ 693\ 393$
$x^*_2 = 0,\ 139\ 721\ 84$
$x^*_3 = 0,\ 510\ 124\ 09$
$x^*_4 = 0,\ 998\ 688\ 30$
$x^*_5 = 1,\ 000\ 027\ 1$

Approximation de $\Gamma(1+t)$ par un polynôme de degré $\leqslant 5$ sur l'ensemble des points

$t_i = i/20$ pour $i = 0, \ldots, 20.$ (1)

$p_5(t) = x_1 t^5 + x_2 t^4 + x_3 t^3 + x_4 t^2 + x_5 t + x_6$

$H = 0,\ 000\ 035\ 40 \leqslant \rho \leqslant 0,\ 000\ 035\ 41$
$x^*_1 = -\ 0,\ 100\ 744\ 65$
$x^*_2 = \ \ 0,\ 421\ 670\ 63$
$x^*_3 = -\ 0,\ 695\ 037\ 38$
$x^*_4 = \ \ 0,\ 948\ 342\ 49$
$x^*_5 = -\ 0,\ 574\ 231\ 09$
$x^*_6 = \ \ 0,\ 999\ 964\ 61$

(1) $\Gamma(1+t)$ est calculé à l'aide du développement donné par CLENSHAW [2]

COMPARAISON AVEC LES RESULTATS OBTENUS AU CHAPITRE II

	Problème continu : intervalle 0,1	Problème discret : ensemble $t_i = i/20 (i=0,\ldots 20)$
	Fonction e^t	
n=3	$0,000\ 544\ 72 \leq \rho_c \leq 0,000\ 544\ 74$	$0,000\ 543\ 18 \leq \rho_d \leq 0,000\ 543\ 24$
n=4	$0,000\ 027\ 10 \leq \rho_c \leq 0,000\ 027\ 21$	$0,000\ 027\ 07 \leq \rho_d \leq 0,000\ 027\ 15$
n=5	$0,000\ 001\ 13 \leq \rho_c \leq 0,000\ 001\ 16$	$0,000\ 001\ 09 \leq \rho_d \leq 0,000\ 001\ 20$
	fonction $\Gamma(1+t)$	
n=5	$0,000\ 037\ 10 \leq \rho_c \leq 0,000\ 037\ 13$	$0,000\ 035\ 40 \leq \rho_d \leq 0,000\ 035\ 41$

C H A P I T R E IV

-=-=-=-

METHODES DE PROGRAMMATION LINEAIRE POUR LA RECHERCHE DE LA SOLUTION DU PROBLEME DISCRET

A) Méthode par abaissement de la borne supérieure de l'erreur.

1) Enoncé du problème

Soient m points $t_i (i = 1, \ldots, m)$ pris dans un intervalle $[a,b]$. Soit $g(t)$ une fonction continue à valeurs réelles, définie sur l'intervalle $[a,b]$. On considère n fonctions $\varphi_j(t)$ $(j = 1, \ldots, n)$ appartenant à $\mathcal{C}[a,b]$, avec $m \geq n + 1$.

Soit g le vecteur à m composantes $g \left| \begin{matrix} g(t_1) \\ \vdots \\ g(t_m) \end{matrix} \right.$

et V l'espace vectoriel engendré par les

n vecteurs $\varphi_j \left| \begin{matrix} \varphi_j(t_1) \\ \vdots \\ \varphi_j(t_m) \end{matrix} \right.$ pour $j = 1, \ldots, n$.

On considère la norme

$$\| g \|_\infty = \operatorname{Max}_{i=1,..,m} | g(t_i) | .$$

Le problème est alors de trouver un ensemble de n scalaires $x_1, \ldots, x_n$ tels que :

$$\operatorname*{Max}_{i=1,..,m} | x_1 \varphi_1 (t_i) + \ldots + x_n \varphi_n(t_i) - g (t_i) | \text{ soit minimum.}$$

On pose :

$$\rho = \operatorname*{Inf}_{\varphi \in V} \| \varphi - g \|_\infty .$$

D'après le théorème I du chapitre I on sait qu'il existe $g^* \in V$ tel que :

$$\| g^* - g \|_\infty = \rho.$$

On supposera par la suite que le sous espace V vérifie la condition d'interpolation, ce qui assure l'unicité de la solution g^*.

2) Le problème borné des moindres carrés (1)

Soit $R \geqslant \rho$. On considère l'ensemble $V_R \subset V$

$$V_R = \{ \varphi \in V : \| \varphi - g \|_\infty \leqslant R \}.$$

V_R n'est pas vide puisque $g^* \in V_R$.

Soit la norme :

$$\| g \|_2 = \sqrt{\sum_{k=1}^{m} g^2_k}$$

On pose $\rho_2(R) = \operatorname*{Inf}_{\varphi \in V_R} \| \varphi - g \|_2$

Le problème borné des moindres carrés consiste à chercher $\varphi_R \in R_V$ tel que :

$$\| \varphi_R - g \|_2 = \rho_2 (R).$$

On appelle φ^* la solution du problème non borné des moindres carrés c'est-à-dire φ^* tel que :

$$\| \varphi^* - g \|_2 = \operatorname*{Inf}_{\varphi \in V} \| \varphi - g \|_2 = \rho_2.$$

On pose $L = \| \varphi^* - g \|_\infty$.

(1) Voir KRABS [5]

On va rappeler les théorèmes principaux démontrés par W. Krabs (1) qui permettent de mettre le problème précédent sous la forme d'un problème de programmation linéaire.

Théorème 1 :

Le problème borné des moindres carrés a une solution unique.

Théorème 2 :

Une condition nécessaire et suffisante pour que $\varphi_R \in V_R$ soit solution du problème borné des moindres carrés est qu'il existe $v \in R^m$ tel que les produits scalaires :

$$(g - \varphi_R - v, \varphi) = 0 \qquad \text{pour tout } \varphi \in V$$

$$(v, \varphi_R - \varphi) \geqslant 0 \qquad \text{pour tout } \varphi \in V_R$$

Théorème 3 :

Si $\rho \leqslant \bar{R} < R \leqslant L$ alors $\rho_2(R) < \rho_2(\bar{R})$

Théorème 4 :

Une condition suffisante pour que $\varphi_R \in V_R$ soit solution du problème borné des moindres carrés est qu'il existe $v \in R^m$ tel que :

$$\begin{cases} (g - \varphi_R - v, \varphi) = 0 & \text{pour tout } \varphi \in V \\ v_k \geqslant 0 \text{ entraine } \varphi_{R_k} - g_k = R \\ v_k < 0 \text{ entraine } \varphi_{R_k} - g_k = -R \end{cases}$$

En effet :

$$(v, \varphi_R - \varphi) = (v, \varphi_R - g) - (v, \varphi - g)$$

$$= \sum_{v_k > o} v_k (R - (\varphi_k - g_k)) + \sum_{v_k < o} v_k (-R - (\varphi_k - g_k))$$

pour tout $\varphi \in V$.

(1) Voir KRABS [5]

Comme $\varphi \in V_R$ est équivalent à

$$R\,(\varphi_k - g_k) \geqslant 0,\ R - (\varphi_k - g_k) \leqslant 0 \text{ pour tout } k \text{ alors } (v, \varphi_k - \varphi) \geqslant 0$$

Les conditions du théorème 2 sont bien vérifiées.

3) Condition nécessaire et suffisante pour que $\varphi^o \in V_R$ soit solution du problème borné des moindres carrés correspondant à R.

Nous allons montrer que la condition suffisante donnée au théorème 4 est en fait également nécessaire.

Si le sous-espace V à n dimensions est représenté par les vecteurs colonnes à m composantes, linéairement indépendants $a_1, \ldots, a_j, \ldots, a_n$ avec $a_{ij} = \varphi_j(t_i)$, tout vecteur $\varphi \in V$ peut s'écrire $\varphi = Ax$. $x \in R^n$. A étant la matrice (m,n) ayant pour élément a_{ij}.

La matrice $A^T A$ de dimensions (n,n) n'est pas singulière et on pose $D\,(A^T A)^{-1}$; soient $e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ vecteur à m composantes et :

$$y^+ = R\,e + g - \varphi = R\,e + g - A\,x$$

$$y^- = R\,e - g + \varphi = R\,e - g + A\,x.$$

Soit y le vecteur à 2 m composantes $y = \begin{pmatrix} y^+ \\ y^- \end{pmatrix}$

les vecteurs $\varphi \in V_R$ sont caractérisés par le fait que $y_k \geqslant 0$ pour $k = 1, \ldots, 2\,m$.

Soit H la matrice (m,m), $H = A\,D\,A^T$ et I la matrice unitaire d'ordre (m,m).

Alors on peut montrer que pour R fixé $\rho \leqslant R \leqslant L$ le problème borné des moindres carrés est équivalent au problème suivant de programmation quadratique :

si b est un vecteur connu ($b = R\,(I + H)\,e - (I - H)\,g$) sous les conditions

$$(H,\ I)\ y = b$$
$$y \geqslant 0$$

minimiser la forme quadratique :

$$\varphi\ (R,\ y) = \frac{1}{2}\ y^T\ y - R\ e^T\ y.$$

(où (H,I) représente la matrice à m lignes et 2 m colonnes formée par la juxtaposition de la matrice H (m,n) et de la matrice unité I (m,m)).

D'après un théorème de P. Wolfe, $y \geqslant o \quad y \in R^{2m}$ est solution de ce problème quadratique, s'il existe $W \geqslant 0$, $W \in R^{2m}$ et u, $u \in R^m$ tels que :

$$(2)\begin{cases} W^T\ y = 0 \\ y - W + \binom{H}{I}\ u - R\ e = 0 \end{cases}$$

De là en posant $W = \begin{pmatrix} W^+ \\ W^- \end{pmatrix}$ W^+ représentant les m premières composantes de W et W^- les autres et soit $v = \frac{1}{2}\ (W^+ - W^-)$ on obtient le <u>théorème 5</u> :

Pour R donné une condition nécessaire et suffisante pour que $\varphi^o \in V_R$ soit solution du problème borné des moindres carrés, est qu'il existe $v \in R^m$ tel que :

$$\begin{cases} (g - \varphi^o - v)^T\ \varphi = 0 \quad \text{pour tout } \varphi \in V \\ v_k \geqslant 0 \quad \text{entraine } \varphi^o{}_k - g_k = R \\ v_k < 0 \quad \text{entraine } \varphi^o{}_k - g_k = - R \end{cases}$$

En tenant compte des notations matricielles données au début du paragraphe, le théorème de caractérisation de la solution du problème borné des moindres carrés peut s'écrire :

Pour R donné $\varphi = Ax$ est solution du problème borné des moindres carrés correspondants s'il existe $v \in R^m$ tel que :

$$3\ \begin{cases} A^T\ (g - Ax - v) = 0 \\ v_k \geqslant 0 \text{ entraine } (A\ x)_k - g_k = R \\ v_k < 0 \text{ entraine } (A\ x)_k - g_k = - R. \end{cases}$$

Si l'on pose comme en (1)

$$y^+ = R\,e + g - A\,x$$

$$y^- = R\,e - g + A\,x$$

et de plus si on décompose $v = v^+ - v^-$ avec :

$v^+_k = \max(v_k, 0)$ $\qquad$ $v^-_k = \max(-v_k, o)$ les deux dernières conditions de (3) s'expriment :

$$v^+_k \geqslant 0 \longrightarrow y^+_k = 0$$

$$v^-_k \geqslant 0 \longrightarrow y^-_k = 0.$$

Alors si l'on change les notations en écrivant :

$$y = \begin{pmatrix} y^+ \\ y^- \end{pmatrix} \quad ; \quad v = \begin{pmatrix} v^+ \\ v^- \end{pmatrix} \quad ; \quad h = \begin{pmatrix} g - ADA^T\,g) \\ -g + ADA^T\,g) \end{pmatrix};$$

$$H = \begin{pmatrix} ADA^T & -ADA^T \\ -ADA^T & ADA^T \end{pmatrix} \quad ; \quad f = DA^T g \; ; \; F = (-DA^T,\ DA^T)$$

D'après le théorème 5 pour $R \in [\rho, L]$, $\varphi = Ax$ est solution du problème borné des moindres carrés correspondant s'il existe $v \in R^{2m}$ avec :

$$\begin{cases} x = f + F\,v \\ y = Re + h + Hv \\ y \geqslant 0 \quad v \geqslant 0 \; R \geqslant 0. \\ v_k \geqslant 0 \quad \text{entraine } y_k = 0 \end{cases}$$

4) Relation avec le problème discret de Tschebyscheff.

On considère le problème suivant :

(I) Sous les conditions

$$y = Re + h + H\,v$$

$$y \geqslant o, v \geqslant o, \; R \geqslant 0$$

$$v_k \geqslant o \longrightarrow y_k = 0 \quad , \text{ minimiser } R \; (y \in R^{2m},\ v \in R^{2m})$$

Pour résoudre ce problème, on peut employer une méthode du simplexe utilisant les pas de Jordan, telle que à chaque pas la troisième condition soit vérifiée.

Cette méthode donne le minimum $R = \rho = \min_{\varphi \in V} \| \varphi - g \|_\infty$

En abandonnant la condition $v_k \geqslant 0 \longrightarrow y_k = 0$ on obtient le problème (II) de programmation linéaire :

Sous les conditions :

$$(E, -H, -e) \begin{pmatrix} y \\ v \\ R \end{pmatrix} = h \qquad \begin{pmatrix} y \\ v \\ R \end{pmatrix} \geqslant 0$$

minimiser la forme linéaire :

$$(0, 0, 1) \begin{pmatrix} y \\ v \\ R \end{pmatrix} = R$$

(0 étant un vecteur à 2 m composantes nulles)

W. Krabs a démontré que le problème (II) est équivalent au problème de Tschebyscheff.

5) Résolution du problème II

Sous les conditions

$$(E, -H, e) \begin{pmatrix} y \\ v \\ R \end{pmatrix} = h \qquad \begin{pmatrix} y \\ v \\ R \end{pmatrix} \geqslant 0$$

minimiser la forme linéaire $(0, 0, 1) \begin{pmatrix} y \\ v \\ R \end{pmatrix} = R$

On a :

$$x = f + F v \qquad x \in R^n$$

$$y = h + H v + R e$$

Pour résoudre ce problème, on utilise une méthode du simplexe. Au départ, on considère la solution du problème non borné des moindres carrés qui correspond à x=f.

$$R = L = \min_k (-h_k).$$

La méthode consiste à échanger les variables à l'aide de pas de Jordan (1) en diminuant R à chaque pas.

Si $-h_{k_o} = \max_k (-h_k)$ alors on peut choisir au début comme variables de base y_k pour $k \neq k_o$ et R.

Après un certain nombre d'itérations, on note y_1, v_1 les variables qui ont été échangées et y_2, v_2 celles qui restent.

$$(4)\begin{cases} x = a + A^1 y_1 + A^2 v_2 + d\,R \\ v_1 = b_1 + B^{11} y_1 + B^{12} v_2 + c_1 R \\ y_2 = b_2 + B^{21} y_1 + B^{22} v_2 + c_2 R. \end{cases}$$

On obtient la dernière solution de base en posant :

$$y_1,\ v_2 = 0$$

Alors on doit avoir d'après les conditions (4) $C_k = 0 \longrightarrow b_k \geqslant 0$

et $0 < \max\limits_{C_k \geqslant o} \left(-\frac{b_k}{C_k}\right) \leqslant R \leqslant \min\limits_{C_k < 0} \left(-\frac{b_k}{C_k}\right)$

$\max\limits_{C_k > 0} \left(-\frac{b_k}{C_k}\right) > 0$ sinon R pourrait atteindre la valeur zéro contrairement à l'hypothèse.

Le plus petit R possible correspondant à la solution de base $y_1,\ v_2 = 0$ est $R = \max\limits_{C_k \geqslant 0} \left(-\frac{b_k}{C_k}\right)$.

Soit par exemple :

$$-\frac{b_{2k_o}}{C_{2k_o}} = \max_{C_k > 0} \left(-\frac{b_k}{C_k}\right).$$

$R = -\frac{b_{2k_o}}{C_{2k_o}}$ entraine pous la solution de base $y_{2k_o} = 0$.

On a $y_{2k_o} = b_{2k_o} + \sum\limits_k B^{21}_{k_o k}\, y_{1k} + \sum\limits_k B^{22}_{k_o k}\, v_{2k} + c_{2k_o} R.$

(1) Voir STIEFEL [11]

Si on annule y_{2k_o} alors :

$$R = - \frac{b_{2k_o}}{C_{2k_o}} - \sum_k \frac{B^{21}_{k_o k}}{C_{2k_o}} y_{1k} - \sum_k \frac{B^{22}_{k_o k}}{C_{2k_o}} v_{2k}$$

Il y a alors deux cas :

1) Tous les $B^{21}_{k_o k}$ et $B^{22}_{k_o k} \leqslant 0$ et dans ce cas le R obtenu par la solution de base est minimal et en faisant $y_1 = v_2 = 0$ on a la solution du problème.

2) Ou bien il existe par exemple $B^{22}_{k_o k_1} > 0$ et alors on peut choisir $v_{2k_1} > 0$, et $v_{2k} = 0$ pour $k \neq k_1$ et $y_1 = 0$ pour avoir R plus petit.

Choix de v_{2k_1}

On doit avoir :

$$(5) \quad \begin{cases} v_{1k} = b_{1k} + B^{12}_{kk_1} v_{2\,k_1} + C_{1k} R \geqslant 0 \\ y_{2k} = b_{2k} + B^{22}_{kk_2} v_{2\,k_1} + C_{2k} R \geqslant 0 \quad \text{pour tout } k. \end{cases}$$

comme $C_{2k_o} > 0$ d'après (4)

$$R = - \frac{b_{2k_o}}{C_{2k_o}} - \frac{B^{22}_{k_o k_1}}{C_{2k_o}} v_{2k_1} + \frac{1}{C_{2k_o}} y_{2k_o}$$

et avec $v_{2k_1} > 0$, R diminue surement lorsque l'on prend :

$y_{2k_o} = 0$. En remplaçant R par sa valeur dans (5), il vient :

$$b_{1k} - \frac{b_{2k_o}}{C_{2k_o}} C_{1k} + (B^{12}_{kk_1} - \frac{B^{12}_{k_o k_1}}{C_{2k_o}} C_{1k}) v_{2k_1} \geqslant 0$$

$$b_{2k} - \frac{b_{2k_o}}{C_{2k_o}} C_{2k} + (B^{22}_{kk_1} - \frac{B^{22}_{k_o k_1}}{C_{2k_o}} C_{2k}) v_{2k_1} \geqslant 0$$

On peut les écrire :

$\bar{b}_k + \bar{B}_k v_{2k_1} \geqslant 0$. qui est équivalent à : $\bar{B}_k = 0 \longrightarrow \bar{b}_k \geqslant 0$

$$\max_{\bar{B}_k > 0} - \frac{\bar{b}_k}{\bar{B}_k} \leqslant v_{2k_1} \leqslant \min_{\bar{B}_k < o} - \frac{\bar{b}_k}{\bar{B}_k}$$

$$\text{soit } \sigma = \min_{\bar{B}_k < 0} - \frac{b_k}{B_k}$$

si $\sigma > 0$ alors on peut prendre $v_{2k_1} = \sigma$.

D'après le choix de $v_{2k_1} = \sigma$ une autre composante que y_{2k_o} de v_1 ou y_2 s'annule.

En échangeant par un pas de Jordan

y_{2k_o} et v_{2k_1} on obtient :

$$x = a' + A'^{1} y_1 + A'^{2} v_2 + d'R$$

$$v_1 = b'_1 + B'^{11} y_1 + B'^{12} v_2 + C'_1 R$$

$$y_1 = b'_2 + B'^{21} y_1 + B'^{22} v_2 + C'_2 R$$

v_1 ayant une composante de plus et y_2 une composante de moins.

On appelle "cas normal" le cas décrit précedement où une composante de y est échangée avec une composante de v.

On a programmé la méthode, proposée par W. Krabs pour résoudre ce problème dans le cas normal. Cette méthode utilise le fait que à chaque pas de la méthode si $y = \begin{pmatrix} y^+ \\ y^- \end{pmatrix}$ et $v = \begin{pmatrix} v^+ \\ v^- \end{pmatrix}$, à la place des composantes de y, des composantes de v deviennent variables de base. Comme les composantes de y correspondantes s'annulent, si à un pas de la méthode Ax - g atteint son maximum pour des composantes d'indices $k_1, \ldots, k_r$ la nouvelle solution $\bar{x}$ est telle que $A\bar{x} - g$ atteint son maximum aux points $k_1, \ldots, k_r$, et k_{r+1}.

Méthode simplifiée dans le cas normal

En posant $h_1 = g - Af$ $\qquad H_1 = A D A^T$

le problème II peut s'écrire avec $y = \begin{pmatrix} y^+ \\ y^1 \end{pmatrix}$ $\qquad v = \begin{pmatrix} v^+ \\ v^- \end{pmatrix}$

Sous les conditions :

$$\begin{cases} y^+ = R\,e + h_1 + H_1 v^+ - H_1 v^- \\ y^- = R\,e - h_1 - H_1 v^+ + H_1 v^- \\ y^+, y^-, v^+, v^- \geq 0, \; R \geq 0 \qquad \text{minimiser } R. \end{cases}$$

Si $F_1 = -DA^T$

$$x = f + F_1 v^+ - F_1 v^-$$

$$y^+ = R\,e + g - A\,x$$

$$y^- = R\,e - g + A\,x.$$

En changeant les notations, on pose maintenant :

$$h = h_1,\; H = H_1,\; F = F_1,\; y = y^+,\; v = v^+ - v^-.$$

Le problème s'écrit alors :

sous les conditions

$$\begin{cases} y = R\,e + h + H\,v & y \in R^m \\ 0 \leqslant y \leqslant 2\,R\,e & \text{minimiser } R. \end{cases}$$

avec $x = f + F\,v$, $y = R\,e + g - A\,x$.

De là on tire :

$$g - A\,x = h + H\,v.$$

Soit v tel que $0 \leqslant y = Re + h + Hv \leqslant 2\,R\,e$ on pose $\bar{h} = H\,v = g - A\,x$ qui représente le vecteur erreur.

Soient $k_1, \ldots, k_r$ les indices pour lesquels

$$\left|\bar{h}_{k_1}\right| = \ldots = \left|\bar{h}_{k_r}\right| = \left|\bar{h}\right|_{\mathscr{O}} = R > 0.$$

On cherche alors un vecteur $\bar{v}$ tel que :

$$\begin{cases} \bar{v}_k = 0 \quad \text{pour } k \neq k_1, \ldots, k_r \\ (1) \qquad y_{k_\sigma} = \bar{R} + h_{k_\sigma} + \sum_{\rho=1}^{r} H_{k_\sigma k_\rho}\,\bar{v}_{k_\rho} = \begin{cases} 2\,\bar{R} \text{ si } h_{k_\sigma} > 0 \\ 0 \quad \text{si } \bar{h}_{k_\sigma} < 0 \end{cases} \\ \qquad \text{pour } \sigma = 1, \ldots, r \\ (2) \qquad 0 \leqslant y_k = \bar{R} + \bar{h}_k + \sum_{\rho=1}^{r} H_{kk_\rho}\,\bar{v}_{k_\rho} \leqslant 2\,\bar{R} \\ \qquad \text{pour } k \neq k_\sigma \qquad \sigma = 1, \ldots, r. \end{cases}$$

et de plus $\bar{R} \leqslant R$.

En posant $\bar{x} = f + Fv + F\,\bar{v} = x + F\,\bar{v}$

$$\bar{y} = \bar{R}\,e + \bar{h} + H\,\bar{v}$$

La condition (1) assure que pour $\sigma = 1, \ldots, r$

$$\left| \bar{h}_{k_\sigma} + \sum_{\rho=1}^{r} H_{k_\sigma k_\rho} \bar{v}_{k_\rho} \right| = \bar{R}.$$

Calcul des $\bar{v}_{k_\sigma}$ $\quad \sigma = 1, \ldots, r.$

La condition (1) peut s'écrire :

$$\sum_{\rho=1}^{r} (H_{k_\sigma k_\rho}) (\text{signe } \bar{h}_{k_\sigma}) \bar{v}_{k_\rho} = \bar{R} - \|\bar{h}\|$$

$(\sigma = 1, \ldots, r).$

Deux cas se présentent :

1) si $r = 1$ on pose $K = H_{k_1 k_1}$ signe $\bar{h}_{k_1}$

2) si $r \geqslant 2$

Dans la condition (1) en retranchant la rème égalité des r premières, il vient :

$$\sum_{\rho=1}^{r} (H_{k_\sigma k_\rho} \text{ signe } \bar{h}_{k_\sigma} - H_{k_r k_\rho} \text{ signe } \bar{h}_{k_r}) \bar{v}_{k_\rho} = 0$$

$(\sigma = 1, \ldots, r - 1).$

Si le système à une solution alors :

$$\bar{v}_{k_\rho} = a_\rho \lambda \qquad \lambda \text{ scalaire réel } (\rho = 1, \ldots, r)$$

si a_ρ est la solution prise pour $a_r = 1$. Les a_ρ sont solutions du système linéaire

$$\sum_{\rho=1}^{r-1} (H_{k_\rho k_\sigma} \text{ signe } \bar{h}_{k_\sigma} - H_{k_r k_\rho} \text{ signe } \bar{h}_{k_r}) a_\rho$$

$$= - (H_{k_\sigma k_r} \text{ signe } \bar{h}_{k_\sigma} - H_{k_r k_r} \text{ signe } \bar{h}_{k_r})$$

pour $\sigma = 1, \ldots, r - 1.$

Alors on pose :

$$K = \sum_{\rho=1}^{r-1} (H_{k_r k_\sigma} \text{ signe } \bar{h}_{k_r})\, a_\rho + H_{k_r k_r} \text{ signe } \bar{h}_{k_r} .$$

Condition pour que $\bar{R} < R$

Il faut que $\lambda \neq 0$.

On peut écrire $K\lambda = \bar{R} - R$ pour $\bar{R} < R$ on doit avoir :

$K \neq 0$ et signe λ = - signe K.

On pose λ = - u signe K. avec $u > 0$ $\quad \bar{R} - R = - u \ |K|$

Si on pose :

$$\lambda_k = \begin{cases} \sum_{\rho=1}^{r} H_{k\, k_\rho}\, a_\rho + H_{k\, k_r} & \text{pour } r \geqslant 2 \\ H_{k\, k_1} & \text{pour } r = 1 \end{cases}$$

la condition (2) s'exprime :

$$0 \leqslant R - |K|\, u + \bar{h}_k - K_k \text{ (signe K)} u \leqslant 2R - 2|K|u$$

$$k \neq k_\rho \quad (\rho = 1, \ldots r)$$

et si :

$$r_k = - |K| - K_k \text{ signe } k \qquad R_k = |K| - K_k \text{ signe K}$$

$$u \leqslant \min_{\substack{k \neq k_\rho \\ r_k < 0}} - \frac{R + \bar{h}_k}{r_k} = W_1$$

$$u \leqslant \min_{\substack{k \neq k_\rho \\ R_k > 0}} \frac{R - \bar{h}_k}{R_k} = W_2$$

Comme $|\bar{h}_k| < R$ $\quad W_1 > 0,\ W_2 > 0$

Alors en posant $u = \min (W_1, W_2)$ la condition (2) est vérifiée et $\bar{R} = R - |K|\, u > 0$.

Alors :

$$\bar{v}_{k_\rho} = -\,u \quad (\text{signe } K)\; a_\rho \quad (\rho = 1, \ldots, r) \quad \text{si } r \geqslant 2$$

$$\bar{v}_{k_1} = -\,u \quad (\text{signe } K) \quad \text{si } r = 1$$

si K = 0 alors on arrête la méthode.

D'après le choix de $u = \min (W_1, W_2)$ pour au moins un $k \neq k_1, \ldots, k_r$ soit $k = k_{r+1}$ on peut écrire :

$$\left| \bar{h}_{k_{r+1}} + \sum_{\rho=1}^{r} H_{k_{r+1}\, k_\rho}\; \bar{v}_{k_\rho} \right| = \left| g_{k_{r+1}} - (A\,\bar{x})_{k_{r+1}} \right| = \bar{R}.$$

Le vecteur erreur atteint son maximum pour les composantes $k_1, \ldots, k_r$ et pour une nouvelle composante k_{r+1}.

6) Résultats numériques.

a) Approximation de $\frac{\sin t}{t}$ par un polynôme de degré 6 $x_1 t^6 + x_2 t^4 + x_3 t^2 + x_4$ sur l'ensemble des points t_i

$$t_i = i \times \pi/20 \quad \text{pour } i = 0, \ldots, 20.$$

On constate que, contrairement à ce qui était prévu, l'erreur n'atteint pas à chaque pas son maximum en au moins un point de plus. Cependant la borne supérieure de la valeur absolue de l'erreur est diminuée à chaque pas.

On obtient :

$$\rho \leqslant 0,\ 000\ 000\ 738$$

$$x^*_1 = -\,0,\ 000\ 185\ 293\ 86$$

$$x^*_2 = \quad 0,\ 008\ 313\ 447\ 6$$

$$x^*_3 = -\,0,\ 166\ 656\ 99$$

$$x^*_4 = \quad 0,\ 999\ 999\ 27$$

On rappelle que par la méthode du deuxième algorithme de Remez adaptée au problème discret (chap. V § 4) on avait obtenu :

$$0,\ 000\ 000\ 73 \leqslant \rho \leqslant 0,\ 000\ 000\ 76.$$

b) Approximation de e^t par un polynôme de degré $\leqslant 3$

$x_1 t^3 + x_2 t^2 + x_3 t + x_4$ sur l'ensemble des points t_i

$t_i = i/20 \quad (i = 0, \ldots, 20)$

$\rho \leqslant 0,\ 000\ 543\ 193$

$x^*_1 = 0,\ 279\ 940\ 41$

$x^*_2 = 0,\ 421\ 768\ 00$

$x^*_3 = 1,\ 016\ 573\ 4$

$x^*_4 = 0,\ 999\ 456\ 81$

On rappelle que la méthode de Stiefel (ch. V, § 3) donnait comme résultat

$$0,\ 000\ 543\ 18 \leqslant \rho \leqslant 0,\ 000\ 543\ 24.$$

Dans le cas de e^t par un polynôme de degré $\leqslant 4$ sur l'ensemble des points $t_i = i/20$ $(i = 0, \ldots, 20)$ après 160 itérations, on obtient $\rho \leqslant 0,\ 000\ 079$ alors que la méthode de Stiefel donnait comme résultat $0,\ 000\ 027\ 0 \leqslant 0,\ 000\ 027\ 1$.

Pour e^t par un polynôme de degré $\leqslant 5$ sur le même ensemble de point l'écart est encore plus important.

B- LE PROBLEME DUAL ET SON RAPPORT AVEC LA METHODE D'ECHANGE DE STIEFEL

1) Etude du problème dual.

Comme on l'a vu au paragraphe précédent le problème de Tschebyscheff est équivalent au problème de programmation linéaire suivant :

Sous les conditions :

$$(I, -H, -e) \begin{pmatrix} y \\ v \\ R \end{pmatrix} = h , \qquad \begin{pmatrix} y \\ v \\ R \end{pmatrix} \geqslant 0$$

minimiser $R = (0, 0, 1) \begin{pmatrix} y \\ v \\ R \end{pmatrix}$

Le problème dual est alors :

sous les conditions

$$\begin{cases} z \leqslant 0 \\ -H^T z \leqslant 0 \\ -e^T z \leqslant 1 \end{cases}$$

rendre maximum la forme linéaire $L(z) = h^T z$.

D'après une propriété de la dualité le minimum ρ de R est égal au maximum de L (z) et pour tout z, $L(z) < \rho$.

En changeant z en -z et en tenant compte , d'après les définitions de h, H que $H^T = H$, $e^T H = 0$, $h^T H = 0$, $H H = H$

il vient :

sous les conditions :

$$Hz = 0$$
$$z \geqslant 0$$
$$e^T z \leqslant 1$$

rendre la forme linéaire $L(z) = -h^T z$ maximum.

On peut démontrer que la condition $e^T z \leqslant 1$ peut - être remplacée, sans changer la valeur du maximum L (z) par $e^T z = 1$.

La condtion $H\,z = 0$ si l'on pose $z = \begin{pmatrix} z^+ \\ z^- \end{pmatrix}$ est équivalente à

$ADA^T z^+ - ADA^T z^- = ADA^T (z^+ - z^-) = 0$ ou encore $A^T (z^+ - z^-) = 0$.

De plus :

$$-h^T z = (ADA^T g - g)^T z^+ - (ADA^T g - g)^T z^-$$

$$= g^T ADA^T (z^+ - z^-) - g^T (z^+ - z^-) \text{ et comme } Hz = 0$$

$$- h^T z = g^T (z^+ - z^-)$$

Le problème s'écrit : sous les conditions

$$\begin{cases} A^T (z^+ - z^-) = 0 \\ z^+,\ z^- \geqslant 0 \\ e^T z^+ + e^T z^- = 1 \end{cases} \quad \text{rendre maximum la forme } L\begin{pmatrix} z^+ \\ z^- \end{pmatrix} = - g^T (z^+ - z^-)$$

En posant $u = z^+ - z^-$ et $\|u\|_1 = \sum_{k=1}^{m} |u_k|$ on a :

Sous les condition :

$$\begin{cases} A^T u = 0 \\ \|u\|_1 = 1 \end{cases} \quad \text{rendre maximum la forme linéaire } L^T (u) = - g^T u$$

Ce problème pourrait être résolu par la méthode d'échange développée par Stiefel dans le cas de l'unicité de la solution.

Nous avons programmé une méthode équivalente exposée par W. Krabs. Cette méthode est basée sur les deux théorèmes suivants :

Théorème 1

Soit E un espace vectoriel normé sur le corps des réels (C), V un sous espace vectoriel et $g \in E$ avec $g \notin V$. Alors :

$$\rho = \inf_{\varphi \in V} \|\varphi - g\| = \max_{L \in \mathcal{L}} \frac{1}{\|L\|_E}$$

avec :

$$\|L\|_E = \sup_{\substack{z \in E \\ z \neq 0}} \frac{|L(z)|}{\|z\|}$$

et $\mathcal{L} = \{L \in E^* : L(\varphi) = 0 \text{ pour } \varphi \in V \text{ et } L(g) = 1\}$

Le calcul de ρ revient alors à calculer le minimum de $\|L\|_E$ $L \in \mathcal{L}$.

Théorème 2

On considère $E = R^m$ avec $\|f\|_\infty = \max_k |f_k|$

et V un sous espace engendré par les vecteurs linéairement indépendants

$a_1, \ldots, a_n$ $n < m$ et $g \notin V$

On sait que pour $L \in E^*$ il existe $y \in E$ avec :

$$L(x) = y^T x = \sum_1^m y_k x_k \quad \text{pour tout } x \in E.$$

et :

$$\|L\|_E = \|y\|_1 = \sum_{k=1}^m |y_k|.$$

Alors :

$$\rho = \inf_{\varphi \in V} \|\varphi - g\|_\infty = \max_{y \in \mathcal{L}} \frac{1}{\|y\|_1} \quad \text{avec}$$

$$\mathcal{L} = \left\{ y \in R^m : y^T a_j = \sigma \ (j = 1, \ldots, n) \text{ et } y^T g = 1 \right\}.$$

Le calcul de la borne minimale ρ est ainsi équivalent à la solution du problème suivant :

On cherche un vecteur $y \in R^m$ tel que sous les conditions

$$(1) \quad \sum_{k=1}^m a_{jk} y_k = b_j \quad \begin{cases} 0 & \text{pour } j = 1, \ldots, n \\ 1 & \text{pour } j = n+1 \end{cases}$$

(en convenant de poser : $a_{n+1,k} = g_k$)

On ait : $\|y\|_1 = \sum_{k=1}^m |y_k|$ minimum.

Pour la résolution de ce problème, on utilise une méthode du simplexe modifiée qui donne la solution du problème de Tschebyscheff.

Comme les vecteurs $a_1, \ldots, a_n$ sont linéairement indépendants et g non combinaison linéaire des a_j la matrice ayant pour élément (a_{jk}) $j = 1, \ldots, n+1$ $k = 1, \ldots, m$ est de rang $n + 1$.

Il existe une solution du système (1) $y \in R^m$ avec $y_k = 0$ pour $k \neq k_1, \ldots, k_{n+1}$ si $\det (a_{jk_\mu})_{(j=1, \ldots, n+1 \; ; \mu = 1, \ldots, n+1)} \neq 0$

un tel y est appelé solution de base.

A partir du système (1) on peut écrire :

$$y_{k\mu} = d_{\mu n+1} + \sum_{\nu=1}^{m-n-1} C_{\mu\nu} \; (-y_{k_{n+1+\nu}}) \quad \text{pour } \mu = 1, \ldots, n+1$$

avec $C_{\mu\nu} = \sum_{j=1}^{n+1} d_{\mu j} \; a_{jk_{n+1+\nu}}$ $\qquad \nu = 1, \ldots, m - n - 1$

et $\sum_{j=1}^{n} d_{\mu j} \; a_{jk_\nu} = \delta_{\mu\nu} = \begin{cases} 0 \text{ pour } \mu \neq \nu \\ 1 \text{ pour } \mu = \nu \end{cases}$

$(\mu \; , \nu = 1; \ldots, n+1)$

Pour la solution de base

$$y_{k\mu} = d_{\mu, n+1}$$

$$\|y\|_1 = \sum_{\mu} | d_{\mu \; n+1} |$$

Cette méthode permet de traiter le cas ou il y a dégénérescence (la solution du problème n'était pas unique). Si tous les $d_{\mu \; n+1} \neq 0$ alors il n'y a pas dégénérescence.

On pose :

$$C = \operatorname*{Max}_{\nu} \left\{ \left| \sum_{\substack{\mu=1 \\ d_{\mu n+1} \neq 0}}^{n+1} C_{\mu\nu} \; \text{signe } d_{\mu n+1} \right| - \sum_{d_{\mu n+1}=0}^{n+1} | C_{\mu\nu} | \right\}$$

On peut montrer que si $C > 1$ alors il existe $\bar{y}$ avec :

$$\| \bar{y} \|_1 < \| y \|_1$$

Si cette condition n'est pas vérifiée $C \leqslant 1$ et si la solution n'est pas dégénérée (ce qui est la cas puisque l'on suppose la condition d'interpolation vérifiée) alors on obtient la solution du problème de Tschebyscheff en résolvant le système :

$$\sum_{j=1}^{n+1} a_{jk_\mu} x_j = \text{signe } d_{\mu\, n+1} \qquad \mu = 1, \ldots, n+1$$

c'est à dire :

$$x_j = \sum_{\mu=1}^{n+1} d_{\mu j} \text{ signe } d_{\mu\, n+1} \qquad (j = 1, \ldots, n+1)$$

et en particulier :

$$x_{n+1} = \sum_{\mu=1}^{n+1} |d_{\mu\, n+1}| = \| g \|_1$$

Dans l'application numérique pour l'approximation de $\frac{\sin(t)}{t}$ par un polynôme pair de degré 6 sur l'ensemble des 21 points $t_i = i \times \pi/20$ $(i = 0, \ldots, 20)$ en prenant comme départ la solution des moindres carrés correspondante dès la première itération , on obtient :

$C = 1,\ 000\ 9$

Cette méthode qui permet de résoudre le problème de Tschebyscheff dans le cas où le nombre des points t_i n'est pas trop grand ne parait pas adaptée au cas où le nombre des points est suffisamment grand pour pouvoir comparer les résultats avec ceux des méthodes donnant la solution du problème continu.

BIBLIOGRAPHIE

[1] N.I. ACHIESER : Théory of Approximation Frederic Ungar Publishung (New-York, 1956).

[2] C.W. CLENSHAW : Mathematical Tabels Volume 5 National Physical laboratory 1962.

[3] W. FRASER et J.F. HART : Approximations rationelles des fonctions continues (Revue ACM, Volume 5, nb 7, juillet 1962).

[4] J.C HERZ : Sur la détermination effective de la meilleure approximation rationelle d'une fonction réelle sur un intervalle. Premier congrès de l'association française de calcul (Grenoble 1960).

[5] W. KRABS : Einige Methoden zur Lösung des diskreten linearen Tschebyscheff - Problems. (Thèse : Hamburg 1963).

[6] MAHELY : Rational approximation for transcendental functions Congres IFIP 1959.

[7] G. MEINARDUS : Über Tschebyscheffsche Approximationen "Archive for rational mechanichs and Analysis" (Vol 9/nb 4/ 1962, p 329-351).

[8] NOVODVORSKI et PINSKER : Procédé d'égalisation des maximums (Russe : Uspehi Math. Nauk. 6, 1951).

[9] E. REMEZ : Méthodes générales de calcul pour l'approximation de Tschebyscheff (Russe 1957).

[10] E. STIEFEL : Über discrete und lineare Tschebyscheff Approximationen Numerische Mathematik 1, 1959.

[11] E. STIEFEL : Note on Jordan Elimination, lienear programming and Tschebyscheff Approximation Numerische Mathematik 2, 1960.

[12] E. STIEFEL Numerical méthods of Tschebyscheff Approximation. On Numerical approximation Congrè LANGER (1959).

TABLE DES MATIERES

-=-=-=-=-=-=-

Vu,

Grenoble, le

Le Président de la Thèse.

Vu,

Grenoble, le

Le Doyen de la Faculté des Sciences

Vu et permis d'imprimer,

Le Recteur de l'Académie de GRENOBLE