Méthodes de localisation du maximum global et des zéros d'une fonction sur un intervalle de la droite numérique

Patricio Basso

THESE

présentée à

Université Scientifique et Médicale de Grenoble

pour obtenir le grade de

DOCTEUR INGENIEUR
Mathématiques Appliquées

par

Patricio BASSO

METHODES DE LOCALISATION DU MAXIMUM GLOBAL ET DES ZEROS D'UNE FONCTION SUR UN INTERVALLE DE LA DROITE NUMERIQUE

Thèse soutenue le 15 juin 1978 devant la Commission d'Examen :

Président : N. GASTINEL

Examinateurs : M. DUC JACQUET
P.J. LAURENT
B. MARTINET

UNIVERSITE SCIENTIFIQUE
ET MEDICALE DE GRENOBLE

Monsieur Gabriel CAU : Président
Monsieur Pierre JULLIEN : Vice Président

MEMBRES DU CORPS ENSEIGNANT DE L'U.S.M.G.

PROFESSEURS TITULAIRES

MM.	AMBLARD Pierre	Clinique de dermatologie
	ARNAUD Paul	Chimie
	ARVIEU Robert	I.S.N.
	AUBERT Guy	Physique
	AYANT Yves	Physique approfondie
Mme.	BARBIER Marie-Jeanne	Electrochimie
MM.	BARBIER Jean-Claude	Physique expérimentale
	BARBIER Reynold	Géologie appliquée
	BARJON Robert	Physique nucléaire
	BARNOU Fernand	Biosynthèse de la cellulose
	BARRA Jean-René	Statistiques
	BARRIE Joseph	Clinique chirurgicale
	BEAUDOING André	Clinique de pédiatrie et puériculture
	BELORIZKY Elie	Physique
	BERNARD Alain	Mathématiques pures
Mme.	BERTRANDIAS Françoise	Mathématiques pures
MM.	BERTRANDIAS Jean-Paul	Mathématiques pures
	BEZEZ Henri	Pathologie chirurgicale
	BLAMBERT Maurice	Mathématiques pures
	BOLLIET Louis	Informatique (IUT B)
	BONNET Jean-Louis	Clinique ophtalmologique
	BONNET-EYMARD Joseph	Clinique gastro-entérologique
Mme.	BONNIER Marie-Jeanne	Chimie générale
MM.	BOUCHERLE André	Chimie et toxicologie
	BOUCHEZ Robert	Physique nucléaire
	BOUSSARD Jean-Claude	Mathématiques appliquées
	BOUTET DE MONVEL Louis	Mathématiques pures
	BRAVARD Yves	Géographie
	CABANEL Guy	Clinique rhumatologique et hydrologique
	CALAS François	Anatomie
	CARLIER Georges	Biologie végétale
	CARRAZ Gilbert	Biologie animale et pharmocodynamie
	CAU Gabriel	Médecine légale et toxicologie
	CAUQUIS Georges	Chimie organique
	CHABAUTY Claude	Mathématiques pures
	CHARACHON Robert	Clinique oto-rhino-laryngologique
	CHATEAU Robert	Clinique de neurologie
	CHIBON Pierre	Biologie animale
	COEUR André	Pharmacie chimique et chimie analytique
	CONTAMIN Robert	Clinique gynécologique
	COUDERC Pierre	Anatomie pathologique

Mme.	DEBELMAS Anne-Marie	Matière médicale
MM.	DEBELMAS Jacques	Géologie générale
	DEGRANGE Charles	Zoologie
	DELORMAS Pierre	Pneumophtisiologie
	DEPORTES Charles	Chimie minérale
	DESRE Pierre	Métallurgie
	DESSAUX Georges	Physiologie animale
	DODU Jacques	Mécanique appliquée (IUT I)
	DOLIQUE Jean-Michel	Physique des plasmas
	DREYFUS Bernard	Thermodynamique
	DUCROS Pierre	Cristallographie
	GAGNAIRE Didier	Chimie physique
	GALVANI Octave	Mathématiques pures
	GASTINEL Noël	Analyse numérique
	GAVEND Michel	Pharmacologie
	GEINDRE Michel	Electroradiologie
	GERBER Robert	Mathématiques pures
	GERMAIN Jean-Pierre	Mécanique
	GIRAUD Pierre	Géologie
	JANIN Bernard	Géographie
	KAHANE André	Physique générale
	KOSZUL Jean-Louis	Mathématiques pures
	KLEIN Joseph	Mathématiques pures
	KRAVTCHENKO Julien	Mécanique
	KUNTZMANN Jean	Mathématiques appliquées
	LACAZE Albert	Thermodynamique
	LACHARME Jean	Biologie végétale
Mme.	LAJZEROWICZ Janine	Physique
MM.	LAJZEROWICZ Joseph	Physique
	LATREILLE René	Chirurgie générale
	LATURAZE Jean	Biochimie pharmaceutique
	LAURENT Pierre-Jean	Mathématiques Appliquées
	LEDRU Jean	Clinique médicale B
	LE ROY Philippe	Mécanique (IUT I)
	LLIBOUTRY Louis	Géophysique
	LOISEAUX Pierre	Sciences nucléaires
	LONGEQUEUE Jean-Pierre	Physique nucléaire
	LOUP Jean	Géographie
Melle	LUTZ Elisabeth	Mathématiques pures
MM.	MALINAS Yves	Clinique obstétricale
	MARTIN-NOEL Pierre	Clinique cardiologique
	MAZARE Yves	Clinique médicale A
	MICHEL Robert	Minéralogie et pétrographie
	MICOUD Max	Clinique maladies infectieuses
	MOURIQUAND Claude	Histologie
	MOUSSA André	Chimie nucléaire
	NOZIERES Philippe	Spectrométrie physique
	OZENDA Paul	Botanique
	PAYAN Jean-Jacques	Mathématiques pures
	PEBAY-PEYROULA Jean-Claude	Physique
	PERRET Jean	Semeiologie médicale (Neurologie)
	RASSAT André	Chimie systématique
	RENARD Michel	Thermodynamique
	REVOL Michel	Urologie
	RINALDI Renaud	Physique
	DE ROUGEMONT Jacques	Neuro-chirurgie
	SEIGNEURIN Raymond	Microbiologie et Hygiène
	SENGEL Philippe	Zoologie
	SIBILLE Robert	Construction mécanique (IUT I)

MM.	SOUTIF Michel	Physique générale
	TANCHE Maurice	Physiologie
	TRAYNARD Philippe	Chimie générale
	VAILLANT François	Zoologie
	VALENTIN Jacques	Physique nucléaire
	VAUQUOIS Bernard	Calcul électronique
Mme.	VERAIN Alice	Pharmacie galénique
MM.	VERAIN André	Physique
	VEYRET Paul	Géographie
	VIGNAIS Pierre	Biochimie médicale

PROFESSEURS ASSOCIES

MM.	CRABBE Pierre	CERMO
	DEMBICKI Eugéniuz	Mécanique
	JOHNSON Thomas	Mathématiques appliquées
	PENNEY Thomas	Physique

PROFESSEURS SANS CHAIRE

Melle	AGNIUS-DELORD Claudine	Physique pharmaceutique
	ALARY Josette	Chimie analytique
MM.	AMBROISE-THOMAS Pierre	Parasitologie
	ARMAND Gilbert	Géographie
	BENZAKEN Claude	Mathématiques appliquées
	BIAREZ Jean-Pierre	Mécanique
	BILLET Jean	Géographie
	BOUCHET Yves	Anatomie
	BRUGEL Lucien	Energétique (IUT I)
	BUISSON René	Physique (IUT I)
	BUTEL Jean	Orthopédie
	COHEN ADDAD Pierre	Spectrométrie physique
	COLOMB Maurice	Biochimie
	CONTE René	Physique (IUT I)
	DELOBEL Claude	M.I.A.G.
	DEPASSEL Roger	Mécanique des fluides
	FONTAINE Jean-Marc	Mathématiques pures
	GAUTRON René	Chimie
	GIDON Paul	Géologie et minéralogie
	GLENAT René	Chimie organique
	GROULADE Joseph	Biologie médicale
	HACQUES Gérard	Calcul numérique
	HOLLARD Daniel	Hématclogie
	HUGONOT Robert	Hygiène et médecine préventive
	IDELMAN Simon	Physiologie animale
	JOLY Jean-René	Mathématiques pures
	JULLIEN Pierre	Mathématiques appliquées
Mme.	KAHANE Josette	Physique
MM.	KRAKOWIACK Sacha	Mathématiques appliquées
	KUHN Gérard	Physique (IUT I)
	LUU DUC Cuong	Chimie organique
	MAYNARD Roger	Physique du solide
Mme.	MINIER Colette	Physique (IUT I)
MM.	PELMONT Jean	Biochimie
	PERRIAUX Jean-Jacques	Géologie et minéralogie
	PFISTER Jean-Claude	Physique du solide
Melle	PIERY Yvette	Physiologie animale

MM.	RAYNAUD Hervé	M.I.A.G.
	REBECQ Jacques	Biologie (CUS)
	REYMOND Jean-Charles	Chirurgie générale
	RICHARD Lucien	Biologie végétale
Mme.	RINAUDO Marguerite	Chimie macromoléculaire
MM.	ROBERT André	Chimie papetière
	SARRAZIN Roger	Anatomie et chirurgie
	SARROT-REYNAULD Jean	Géologie
	SIROT Louis	Chirurgie générale
Mme.	SOUTIF Jeanne	Physique générale
MM.	STIEGLITZ Paul	Anesthésiologie
	VIALON Pierre	Géologie
	VAN CUTSEM Bernard	Mathématiques appliquées

MAITRES DE CONFERENCES ET MAITRES DE CONFERENCES AGREGES

MM.	ARMAND Yves	Chimie (IUT I)
	BACHELOT Yvan	Endocrinologie
	BARGE Michel	Neuro-chirurgie
	BEGUIN Claude	Chimie organique
Mme	BERIEL Hélène	Pharmacodynamie
MM.	BOST Michel	Pédiatrie
	BOUCHARLAT Jacques	Psychiatrie adultes
Mme.	BOUCHE Liane	Mathématiques (CUS)
MM.	BRODEAU François	Mathématiques (IUT B) (Personne étrangère habilitée à être directeur de thèse)
	CHAMBAZ Edmond	Biochimie médicale
	CHAMPETIER Jean	Anatomie et organogénèse
	CHARDON Michel	Géographie
	CHERADAME Hervé	Chimie papetière
	CHIAVERINA Jean	Biologie appliquée (EFP)
	CONTAMIN Charles	Chirurgie thoracique et cardio-vasculaire
	CORDONNIER Daniel	Néphrologie
	COULOMB Max	Radiologie
	CROUZET Guy	Radiologie
	CYROT Michel	Physique du solide
	DENIS Bernard	Cardiologie
	DOUCE Roland	Physiologie végétale
	DUSSAUD René	Mathématiques (CUS)
Mme.	ETERRADOSSI Jacqueline	Physiologie
MM.	FAURE Jacques	Médecine légale
	FAURE Gilbert	Urologie
	GAUTIER Robert	Chirurgie générale
	GIDON Maurice	Géologie
	GROS Yves	Physique (IUT I)
	GUIGNIER Michel	Thérapeutique
	GUITTON Jacques	Chimie
	HICTER Pierre	Chimie
	JALBERT Pierre	Histologie
	JULIEN-LAVILLAVROY Claude	O.R.L.
	KOLODIE Lucien	Hématologie
	LE NOC Pierre	Bactériologie-virologie
	MACHE Régis	Physiologie végétale
	MAGNIN Robert	Hygiène et médecine préventive
	MALLION Jean-Michel	Médecine du travail
	MARECHAL Jean	Mécanique (IUT I)
	MARTIN-BOUYER Michel	Chimie (CUS)
	MICHOULIER Jean	Physique (IUT I)

MM.	NEGRE Robert	Mécanique (IUT I)
	NEMOZ Alain	Thermodynamique
	NOUGARET Marcel	Automatique (IUT I)
	PARAMELLE Bernard	Pneumologie
	PECCOUD François	Analyse (IUT B) (Personnalité étrangère habilité à être directeur de thèse)
	PEFFEN René	Métallurgie (IUT I)
	PERRIER Guy	Géophysique-Glaciologie
	PHELIP Xavier	Rhumatologie
	RACHAIL Michel	Médecine interne
	RACINET Claude	Gynécologie et obstétrique
	RAMBAUD André	Hygiène et hydrologie (Pharmacie)
	RAMBAUD Pierre	Pédiatrie
	RAPHAEL Bernard	Stomatologie
Mme.	RENAUDET Jacqueline	Bactériologie (Pharmacie)
MM.	ROBERT Jean-Bernard	Chimie physique
	Romier Guy	Mathématiques (IUT B) (Personnalité étrangère habilité à être directeur de thèse)
	SCHAERER René	Cancérologie
	SHOM Jean-Claude	Chimie générale
	STOEBNER Pierre	Anatomie pathologie
	VROUSOS Constantin	Radiologie

MAITRES DE CONFERENCES ASSOCIES

MM.	DEVINE Roderick	Spectro physique
	HODGES Christopher	Transition de phases

Fait à SAINT MARTIN D'HERES, NOVEMBRE 1976.

A Myriam

Les travaux qui sont l'objet de cette thèse ont été effectués au Laboratoire de Mathématiques Appliquées de Grenoble. Je tiens à exprimer ma reconnaissance à tous ses membres pour leur accueil et leurs encouragements.

Monsieur M. DUC-JACQUET, Maître de Conférences, a dirigé ce travail. Pour son appui constant, pour ses conseils judicieux et pour tout le temps qu'il a bien voulu me consacrer, je lui exprime ma très vive reconnaissance.

Monsieur le Professeur N. GASTINEL a bien voulu s'intéresser à ce travail et accepter de présider le Jury. Je le prie de trouver ici l'expression de ma respectueuse gratitude.

Je suis reconnaissant à Messieurs les Professeurs P.J. LAURENT et B. MARTINET d'avoir accepter de participer au Jury.

Je tiens aussi à remercier mon collègue, D. BUSSONAIS, pour les fructueuses conversations et l'intérêt qu'il a témoigné pour mon travail.

J'adresse enfin mes vifs remerciements à Madame BICAIS pour le soin apporter à la dactylographie de ce document ainsi qu'à l'équipe du Service de Reproduction pour sa réalisation matérielle.

TABLE DES MATIERES

INTRODUCTION

CHAPITRE - I - UNE CLASSE DE METHODES DE LOCALISATION

CHAPITRE-II - METHODES DE LOCALISATION DANS H^1 ET H^2

CHAPITRE - III : ESSAIS NUMERIQUES ET ANALYSE DE RESULTATS

INTRODUCTION

Si f est une fonction numérique réelle définie sur un intervalle [a,b], on s'intéresse aux deux problèmes suivants :

(I) "Trouver $E = \{x \in [a,b] / f(x) = \alpha = \max_{x \in [a,b]} f(x)\}$"

(II) "Trouver $E = \{x \in [a,b] / f(x) = 0\}$".

Pour tout ε positif, on considère le problème suivant associé à l'un ou l'autre des problèmes (I) ou (II).

(III) "Trouver L_ε, réunion finie d'intervalles, tel que $(E \setminus L_\varepsilon) \leq \varepsilon$ " (*)

En général les méthodes existantes, qui permettent de résoudre III, consistent à construire une suite de localisations $\{L_n\}$, $n \geq 0$ telle que $E \subseteq L_{n+1} \subseteq L_n$ quelque soit $n \in \mathbb{N}$.

Mis à part les méthodes concernant les fonctions polynomiales et certaine concernant les fonctions lipshitzs continues, toutes les méthodes font, de façon explicite ou non, l'hypothèse que E ne contient qu'un seul point.

Par exemple, l'unicité intervient dans la constructibilité de la suite $\{L_n\}$, $n \geq 0$, dans la méthode de dichotomie, celle de GROSS et JOHNSON [14] (localisation de racines), du nombre d'or, de la suite de Fibonnaci de KIEFER [19],[20], de BERMAN [2] (maximum de fonctions unimodales). Une autre méthode particulièrement simple et performante a été proposée par CEA [5] dans le cadre du choix du pas dans les méthodes de descente (minimum des fonctions convexes).

(*) $(E \setminus L_\varepsilon)$ désigne la mesure de l'ensemble différence de E et L_ε.

Par ailleurs, l'hypothèse d'unicité est nécessaire pour garantir la convergence des méthodes de MICHELLI et MIRANKER [22] et celle de ČERNOUSKO [
(localisation de racines de fonctions lipshitzs-continues monotones).

Il existe également des méthodes qui sans l'hypothèse d'unicité calculent un point $\bar{x}$, appartenant à [a,b], tel que $f(\bar{x})$ approche le maximum global à ε-près mais qui ne localisent pas l'ensemble solution (cf. problème III). De telles méthodes ont été développées par BRENT [3],[4] et EVTUSHENKO [9] (maximum global de fonctions lipshitz-continues).

Dans le cas des polynômes l'intérêt s'est porté vers la localisation des racines dans le plan complexe et les méthodes sont souvent employées en conjonction avec un procédé de deflation. Suivant une idée de WEYL [34], les algorithmes proposés sont basés sur un test, appelés test d'exclusion par HENRICI et GARGANTINI [18] ou de proximité par HENRICI [15], lequel peut être appliqué à un polynôme et une certaine région du plan complexe. Les régions pour lesquelles le test est positif sont des régions dont on peut assurer qu'elles ne contiennent aucun zéro du polynôme. Les algorithmes consistent à partager à chaque pas une région qui contient toutes les racines et à tester chacune de ses parties pour détecter les ensembles susceptibles d'en contenir une. La première de ces méthodes a été proposée par LEHMER [21] en utilisant comme test l'algorithme de SCHUR-COHN [32],[7] sur des cercles, puis généralis par HENRICI [15] pour d'autres types de tests. HENRICI et GARGANTINI [18] ont étudié des méthodes pour des carrés et TOURNIER [33] a proposé une méthode qui utilise des rectangles et un test basé sur la théorie des caractéristiques de Kronecker. Cette dernière méthode a été programmée en langage formel. GARGANTINI et HENRICI [12] et HENRICI [16] ont étudié la question d'accélération de la convergence des localisations une fois qu'une partie ou la totalité des racines du polynôme ont été isolées dans des cercles. Finalement, HENRICI [15] et FRIEDLI [10] ont étudié des stratégies de recouvrement optimal qui minimisent le nombre de tests nécessaires pour atteindre une précision donnée. Pour une revue des principales méthodes et références additionnelles on peut consulter HENRICI [17].

Dans le cas des fonctions lipshitz-continues, une méthode a été développée par PIYAVSKII [26] et par SHUBERT [28]. Cette méthode, qui utilise la connaissance d'une constante de Lipshitz de la fonction, ne fait pas l'hypothèse d'unicité.

Nous nous proposons d'étudier une classe de méthodes de localisation pour résoudre le problème III et pour lesquelles aucune hypothèse d'unicité n'est nécessaire.

En supposant que la fonction appartienne à l'espace de Sovolev $H^q(a,b)$, nous utiliserons sa semi-norme pour déterminer des majorants et des minorants à l'aide desquels nous construirons une suite d'ensembles permettant de résoudre le problème III.

Dans le chapitre I, après avoir proposé un formalisme général pour les méthodes de localisation basées sur des suites de majorants et de minorants, nous étudions la méthode de Shubert-Piyavskii [28], [26]. Nous donnons ensuite une méthode pour majorer et minorer une fonction de $H^q(a,b)$ sur un intervalle fermé I de [a,b] et nous étudions son comportement lorsque la mesure de I converge vers zéro. Finalement, nous utilisons ces majorants et minorants pour définir des méthodes de localisation permettant de résoudre le problème III.

Au chapitre II nous étudions ces méthodes du point de vue de sa mise en oeuvre en machine. Nous nous restreignons aux cas $q = 1$ et $q = 2$. Dans le cas $q = 2$ nous introduisons une légère modification afin de faciliter les calculs. Motivés par une certaine instabilité numérique observée dans les essais numériques nous proposons une méthode de correction de certaines valeurs calculées en machine. Pour cela nous utilisons les règles des opérations arithmétiques en virgule flottante en machine (WILKINSON [35], PICHAT [25]).

Au chapitre III nous utilisons les méthodes étudiées pour la localisation du maximum global et des racines d'une fonction sur un intervalle donné. Le comportement des méthodes est étudié pour différentes fonctions présentant différentes difficultés. Pour le problème de localisation du maximum global nous comparons les résultats avec la méthode de Shubert-Piyavskii. Finalement, des conclusions sont établies et certaines lignes de développement possibles sont proposées.

CHAPITRE - I

UNE CLASSE DE METHODES DE LOCALISATION

I - INTRODUCTION

Nous allons étudier dans ce chapitre une classe de méthodes de localisation, pour résoudre le problème III, basées sur des majorants et des minorants de la fonction. Au paragraphe 2 nous proposons un formalisme général pour cette classe de méthodes.

Au paragraphe 3, nous utilisons ce contexte pour décrire la méthode de SHUBERT-PIYAVSKII [28],[26] pour la localisation du maximum global.

Au paragraphe 4 nous donnons une façon de construire des majorants et des minorants d'une fonction appartenant à $H^q(a,b)$ sur un sous-intervalle de [a,b] et nous étudions son comportement lorsque la mesure de l'intervalle converge vers zéro.

Finalement, au paragraphe 5, nous proposons une méthode de construction des suites des majorants et des minorants qui convergent ponctuellement vers la fonction dans [a,b]. A partir de ces majorants et minorants nous définissons deux classes de méthodes, une première qui utilise les racines des majorants et minorants et une deuxième basée sur un test qui est appliqué à des sous-intervalles de [a,b]. La convergence de ces deux classes de méthodes découle des résultats généraux du paragraphe 2

II - METHODES DE LOCALISATION PAR MAJORATION ET MINORATION DE LA FONCTION

Soit $f \in C(a,b)$ et k la limite commune de deux suites de nombres réels $\{k_n^-\}$ et $\{k_n^+\}$ $n \geq 0$, tels que

$(H.1)^-$ $k_n^- \leq k \quad \forall n \in \mathbb{N}$ $\qquad$ $(H.1)^+$ $k \leq k_n^+ \quad \forall n \in \mathbb{N}$

$(H.2)^-$ $\lim_{n\to\infty} k_n^- = k$ $\qquad$ $(H.2)^+$ $\lim_{n\to\infty} k_n^+ = k$

et considérons le problème de la localisation de l'ensemble des solutions de l'équation :

$$f(x) = k \qquad \text{sur } [a,b] \tag{1.1}$$

Si la quantité k est connue explicitement, le problème est donc un problème de localisation des racines de $g(x) = f(x)-k$ dans $[a,b]$. Par contre, si k est le maximum global de f dans $[a,b]$ on a le problème de localisation du maximum global.

Soit E l'ensemble des solutions du problème (1.1)

$$E = \{x \in [a,b] / f(x) = k\}$$

De façon évidente l'ensemble E peut s'écrire comme

$$E = E^+ \cap E^-$$

avec

$$E^+ = \{x \in [a,b] / f(x) \geq k\}$$

et

$$E^- = \{x \in [a,b] / f(x) \leq k\}$$

Il suffit donc d'étudier une technique de localisation de l'ensemble E^+ laquelle, avec certaines modifications, sera valable pour l'ensemble E^-.

Soit $\{F_n^+\}$ une suite de fonctions telle que

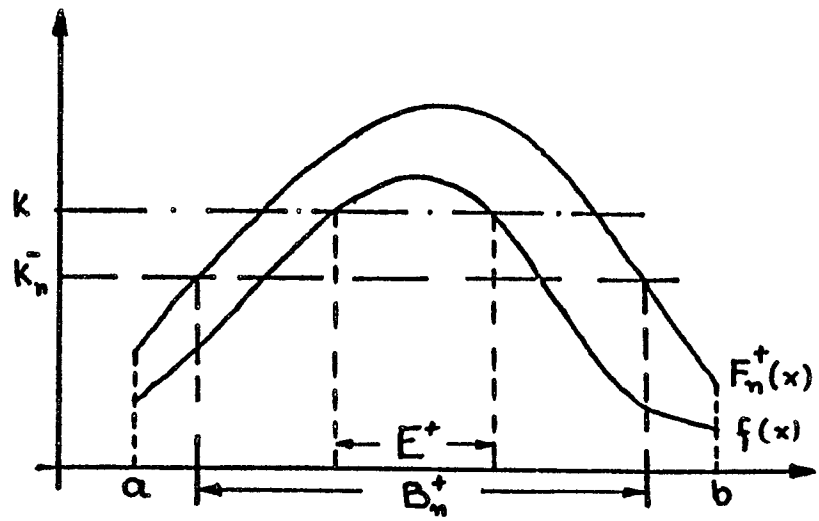
H.3) $\quad F_n^+(x) \geq f(x) \qquad \forall x \in [a,b]$

H.4) $\quad \lim_{n\to\infty} F_n^+(x) = f(x) \qquad \forall x \in [a,b]$

Il est évident que l'ensemble

$$B_n^+ = \{x \in [a,b] / F_n^+(x) \geq k_n^-\}$$

est une localisation de l'ensemble E^+ car $\forall x \in [a,b]$ $F_n^+(x) \geq k \geq k_n^-$.
L'ensemble B_n^+ est indiqué schématiquement dans la figure ci-dessous :



PROPOSITION 1.1

Sous les hypothèses $(H.1)^-$, $(H.2)^-$, (H.3) et (H.4) la suite d'ensembles

$$L_0^+ = [a,b]$$

$$L_{n+1}^+ = \{x \in L_n^+ / F_n^+(x) \geq k_n^-\} \qquad n \geq 0$$

et telle que

$$L_\infty^+ = \bigcap_{n\geq 0} L_n^+ = E^+$$

Démonstration :

1°) si $k > \max_{x\in[a,b]} f(x)$ alors $E^+ = \emptyset$

si $L_\infty^+ \neq \emptyset$, il existe $x \in L_\infty^+$ tel que $x \in L_n^+$ $\forall n \in \mathbb{N}$, d'où

$$F_n^+(x) \geq k_n^- \qquad \forall n \in \mathbb{N}$$

ce qui entraîne d'après $(H.2)^-$ et (H.4)

$$f(x) \geq k$$

ce qui est impossible et montre donc que $L_\infty^+ = \emptyset = E^+$.

2°) Soit $k \leq \max_{x\in[a,b]} f(x)$ et soit $\{B_n^+\}$, $n \geq 0$ la suite d'ensembles :

$$B_o^+ = [a,b]$$

$$B_n^+ = \{x \in [a,b] / F_n^+(x) \geq k_n^- \} \qquad n \geq 1$$

Soit $x \in \bigcap_{n\geq 0} B_n^+$, alors $x \in B_n^+$ quelque soit $n \in \mathbb{N}$, d'où

$$F_n^+(x) \geq k_n^- \qquad \forall n \in \mathbb{N}$$

D'après les hypothèses $(H.2)^-$ et (H.4) cela entraîne

$$f(x) \geq k$$

donc $x \in E^+$.

Inversement si $x \in E^+$, d'après les hypothèses $(H.1)^-$ et (H.3) on a :

$$F_n^+(x) \geq f(x) \geq k \geq k_n^- \qquad \forall n \in \mathbb{N}$$

d'où $x \in B_n^+ \quad \forall n \in \mathbb{N}$

On a donc

$$E^+ = \bigcap_{n\geq 0} B_n^+$$

Finalement, comme $L_m^+ = \bigcap_{n=0}^{m} B_m^+$ car $L_m^+ = L_{m-1}^+ \cap B_{m-1}^+$ et $L_o^+ = B_o^+ = [a,b]$, on a :

$$E^+ = \bigcap_{n\geq 0} B_n^+ = \bigcap_{m\geq 0} \left(\bigcap_{n=0}^{m} B_m^+ \right) = \bigcap_{m\geq 0} L_m^+$$

≡ ≡ ≡

Evidemment, on peut démontrer une proposition similaire par rapport à l'ensemble E^-, le problème de la localisation de l'ensemble E se réduit donc à la construction des suites $\{F_n^-\}$, $\{F_n^+\}$, $\{k_n^-\}$ et $\{k_n^+\}$. En particulier, pour le problème de localisation des racines de f on peut prendre $k_n^- = k = k_n^+ = 0$.

III - METHODES DE LOCALISATION POUR DES FONCTIONS LIPSHITZIENNES

Nous allons décrire dans cette section une méthode de localisation qui utilise une constante de Lipshitz L de la fonction. Cette méthode est à notre connaissance la seule capable de résoudre le problème posé dans l'introduction (problème III) pour des fonctions qui ne soient pas des polynômes et sans l'hypothèse d'unicité. Elle a été développée indépendamment par SHUBERT [28] et par PIYAVSKII [26] bien que ce dernier l'ait fait dans un cadre un peu plus général. Nous décrirons d'abord la méthode telle qu'elle a été proposée par Shubert et nous indiquerons ensuite les différences avec la présentation faite par Piyavskii.

La méthode de Shubert consiste à construire une suite $\{x_n\}$, $n \geq 0$ x_o arbitraire, le passage de la nième à la (n+1)ème itération étant définie par :

1°) $$F_n^+(x) = \min_{j=0,..,n} \{f(x_j)+L|x-x_j|\}$$

2°) $$M_n = \max_{x\in[a,b]} F_n^+(x)$$

3°) x_{n+1} est un point tel que $F_n(x_{n+1}) = M_n$ (il peut y avoir plusieurs)

≡ ≡ ≡

Si l'on définit la suite $\{k_n^-\}$, $n \geq 0$ par

i) $$k_o^- = f(x_o)$$

ii) $$k_{n+1}^- = \max \{k_n^-, f(x_{n+1})\}$$

Shubert établit les propriétés de convergence suivantes :

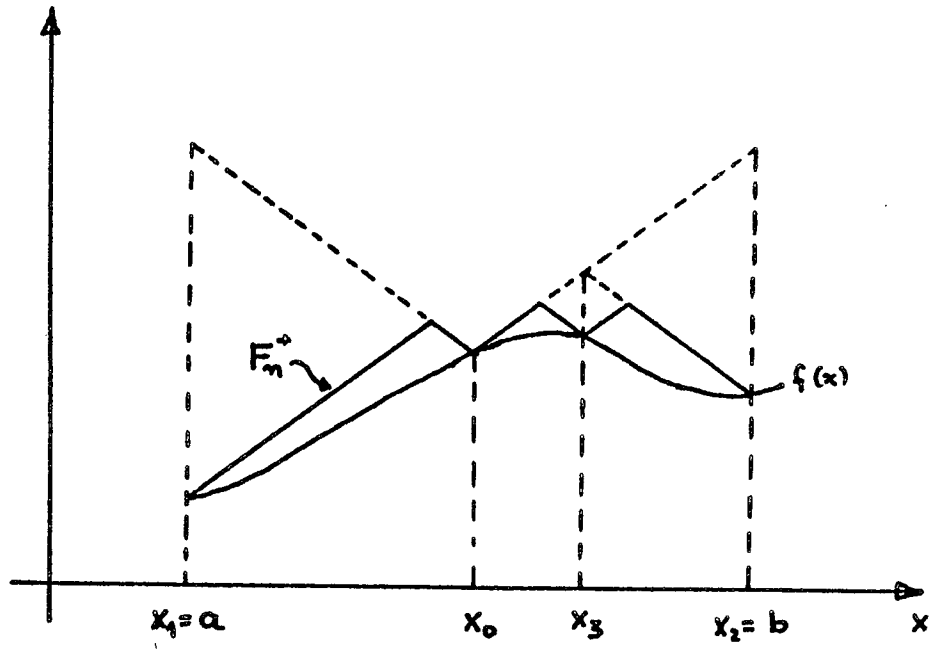
a) $$k_n^- \uparrow k = \max_{x\in[a,b]} f(x)$$

b) $$M_n \downarrow k$$

c) $$\inf_{x\in E} |x-x_n| \to 0 \quad \text{avec } E = \{x \in [a,b]/f(x) = k\}$$

La suite d'ensembles $B_n^+ = \{x \in [a,b]/F_n^+(x) \geq k_n^-\}$ est telle que $E \subseteq B_n^+$ pour tout $n \geq 0$. L'ensemble $B_\infty^+ = \bigcap_{n\geq 0} B_n^+$ diffère de l'ensemble E, au pl en une quantité dénombrable de points isolés ([28]).

La suite de majorants $\{F_n^+\}$ est donc construite en ajoutant un nouveau point à chaque pas. La fonction F_n^+, que nous schématisons ci-après, est une fonction linéaire par morceaux qui interpole f sur l'ensemble de points $\{x_j, j = 0,...,n\}$. Chaque morceau de droite est tracé à partir du point $(x_j, f(x_j))$ avec une pente égale à +L ou -L.



D'après les propriétés (a) et (b), le montant k du maximum global est encadré par $k_n^- \leq k \leq M_n$. Shubert étudie le problème de l'optimalité sous l'angle de la minimisation, en n évaluations de f, de la quantité $M_n - k_n^-$, sur la classe de fonctions qui ont la même constante de Lipshitz. Il démontre que si $f(x_j) = y_o = cte$ $(j = 0,\ldots,n)$, alors $M_n - k_n^-$ est de l'ordre de $L\delta_n/2$ avec δ_n la distance maximale entre deux points consécutifs de l'ensemble $\{x_j\}$ $(j = 0,\ldots,n)$. En particulier si $\delta_n = (b-a)/n$, alors $M_n - k_n^-$ est de l'ordre de $L(b-a)/n$.

La question d'optimalité pour des méthodes de localisation qui utilisent une constante de Lipshitz a été étudiée par SUKHAREV [29],[30],[31].

PIYAVSKII [26] pour sa part se place dans le cas d'une fonction définie sur un compact X d'un espace métrique et il définit cette fois des minorants de f à l'aide d'une fonction $g : X^2 \to \mathbb{R}$ telle que

1°) $g(x,y) \leq f(x) \qquad \forall x,y \in X$

2°) $g(x,x) = f(x) \qquad \forall x \in X$

Le minorant est maintenant défini par la donnée d'un ensemble de points $\{y_j,\ j = 0,\ldots,n\} \subseteq X$ par

$$G(x/y_o,\ldots,y_n) = \max_{j=0,\ldots,n} g(x,y_j)$$

Le reste suit le développement fait par Shubert dont sa méthode devient un cas particulier en prenant la fonction $g(x,y) = f(y)-L|x-y|$ qui est, d'ailleurs, le seul exemple de choix de g donné dans ce papier.

Nous avons programmé cette méthode et nous ferons au chapitre III, la comparaison avec la méthode que nous allons proposer.

IV - UNE CLASSE DE MAJORANTS ET DE MINORANTS

La possibilité de pouvoir déterminer effectivement la localisation L_n^+ de la proposition 1.1 réside fondamentalement dans le choix du majorant F_n^+. Il faut le choisir de sorte que l'on puisse calculer les racines de l'équation : $F_n^+(x) = k_n^-$ et de sorte que L_n^+ puisse être exprimé en fonction de ces racines.

D'autre part, la "vitesse de convergence" de la méthode dépendra de la convergence de F_n^+ vers f sur [a,b]. Ces deux facteurs nous amènent à chercher des majorants qui soient simples et qui convergent "rapidement" vers f.

Comme nous l'avons dit dans l'introduction, nous allons faire l'hypothèse

(H5) $\qquad f \in H^q(a,b)$

et on supposera que l'on connait la semi-norme de f

$$|f|^2 = \int_a^b [f^{(q)}(s)]^2\, ds$$

Au paragraphe 4 du chapitre II nous analyserons cette hypothèse en détail et au chapitre III nous la comparerons avec l'hypothèse faite par

Shubert et Piyavskii.

Pour toute fonction $\varphi \in H^q(a,b)$ on peut écrire :

$$\varphi(x) = \sum_{i=0}^{q-1} \frac{(x-a)^i}{i!} \varphi^{(i)}(a) + \int_a^b \frac{(x-s)_+^{q-1}}{(q-1)!} \varphi^{(q)}(s)\, ds \tag{1.2}$$

et en utilisant l'inégalité de Schwartz on obtient :

$$\left|\varphi(x) - \sum_{i=0}^{q-1} \frac{(x-a)^i}{i!} \varphi^{(i)}(a)\right| \leq \left(\int_a^b \frac{(x-s)_+^{2q-2}}{[(q-1)!]^2}\, ds\right)^{1/2} |\varphi|$$

ou encore

$$\left|\varphi(x) - \sum_{i=0}^{q-1} \frac{(x-a)^i}{i!} \varphi^{(i)}(a)\right| \leq C(x-a)^{\frac{2q-1}{2}} \tag{1.3}$$

avec

$$C = \frac{|\varphi|}{(2q-1)^{1/2}(q-1)!}$$

Si l'on prend $\varphi = f - P_a$, avec P_a polynôme de degré inférieur ou égal à q,

$$P_a(x) = \sum_{i=0}^{q} \alpha_i (x-a)^i, \tag{1.4}$$

et si l'on choisit α_i $(i = 0,\ldots,q-1)$ tels que

$$P_a^{(j)}(a) = f^{(j)}(a) \qquad (j = 0,\ldots,q-1)$$

autrement dit, si l'on choisit α_i $(i = 0,\ldots,q-1)$ comme

$$\alpha_i = \frac{f^{(i)}(a)}{i!} \qquad (i = 0,\ldots,q-1) \tag{1.5}$$

alors l'équation (1.3) s'écrit :

$$(1.6) \qquad |f(x)-P_a(x)| \leq \frac{|f-P_a|}{(2q-1)^{1/2}(q-1)!}(x-a)^{\frac{2q-1}{2}}$$

avec

$$|f-P_a|^2 = \int_a^b \{f^{(q)}(s)-q!\alpha_q\}^2 ds$$

Comme nous voulons construire des majorants les plus "fins" possibles, on a intérêt à choisir α_q de façon à obtenir le "plus petit" majorant dans l'équation (1.6), autrement dit, on choisira α_q qui minimise $|f-P_a|$. Cela est assuré si l'on prend

$$\alpha_q = \frac{f^{(q-1)}(b)-f^{(q-1)}(a)}{(b-a)q!}$$

Avec ce choix de α_q on a

$$|f-P_a|^2 = |f|^2 - \frac{\{f^{(q-1)}(b)-f^{(q-1)}(a)\}^2}{b-a}$$

Remarque - Il en résulte que P_a est la projection orthogonale de f sur la variété linéaire affine des polynômes de degré inférieur ou égal à q qui interpolent f et ses dérivées jusqu'à l'ordre (q-1) en a, et cela dans $H^q(a,b)$ muni de la norme

$$\|\varphi\|^2 = |\varphi|^2 + \sum_{i=0}^{q-1} [\varphi^{(i)}(a)]^2.$$

≡ ≡ ≡

Nous avons donc obtenu la majoration

$$\boxed{f(x) \leq F_a^+(x) = P_a(x)+W(x-a)^{\frac{2q-1}{2}}} \quad \forall x \in [a,b]$$

avec

$$(1.7)\quad \left\{ \begin{array}{l} P_a(x) = \sum_{i=0}^{q} \alpha_i (x-a)^i \\ \\ \alpha_i = \dfrac{f^{(i)}(a)}{i!} \qquad (i = 0,\ldots,q-1) \\ \\ \alpha_q = \dfrac{\{f^{(q-1)}(b) - f^{(q-1)}(a)\}}{(b-a)\,q!} \\ \\ W^2 = \dfrac{|f|^2 - \dfrac{\{f^{(q-1)}(b) - f^{(q-1)}(a)\}^2}{b-a}}{(2q-1)\quad [\,(q-1)!\,]^2} \end{array} \right.$$

Dans [1] nous avions déjà trouvé cette majoration en utilisant les propriétés d'un espace de Hilbert à noyau reproduisant qui avait été préalablement étudié par M. DUC-JACQUET [8] . Il faut remarquer que l'utilisation des propriétés du noyau, ainsi que les exemples numériques que nous avons développés dans [1], nous ont servi à mieux comprendre le problème, bien que, comme il découle de la construction que nous venons de faire, l'utilisation du noyau de l'espace n'était pas essentiel (au moins dans le cas à une dimension).

D'une façon analogue, en partant de la relation

$$\varphi(x) = \sum_{i=0}^{q-1} \frac{(x-b)^i}{i!}\, \varphi^{(i)}(b) + \int_a^b \frac{(s-x)_+^{q-1}}{(q-1)!}\, \varphi^{(q)}(s)\, ds$$

on obtient une nouvelle majoration

$$\boxed{f(x) \le F_b^+(x) = P_b(x) + W(b-x)^{\frac{2q-1}{2}}} \qquad \forall x \in [a,b]$$

avec

$$(1.8)\quad \begin{cases} P_b(x) = \sum_{i=0}^{q} \beta_i (x-b)^i \\ \beta_i = \dfrac{f^{(i)}(b)}{i!} \quad (i = 0,\ldots,q-1) \\ \beta_q = \alpha_q \end{cases}$$

et W défini comme dans (1.7)

Finalement on peut définir un majorant unique de f sur [a,b] par

$$(1.9)\quad \boxed{F^+(x) = \min(F_a^+(x),\ F_b^+(x))}$$

D'après la construction de F_a^+ et F_b^+, le majorant F^+ interpole la fonction f et ses dérivées jusqu'à l'ordre (q-1) aux points a et b mais il n'appartient pas nécessairement à $H^q(a,b)$.

Considérons maintenant un ensemble ordonné de points $\{x_j\} \in [a,b]$ $(j = 0,\ldots,n+1)$ $a = x_0 < x_1 < \ldots < x_{n+1} = b$ et soit F_j^+ le majorant de f construit sur l'intervalle $[x_j,x_{j+1}]$ $(j = 0,\ldots,n)$ suivant la formule (1.9). Alors la fonction définie sur [a,b] qui coïncide avec F_j^+ sur le j-ième intervalle est encore un majorant de f. Pour pouvoir utiliser ce majorant dans la proposition 1.1 il faut montrer qu'il satisfait l'hypothèse H.4, autrement dit, il faut montrer qu'il converge ponctuellement vers f dans [a,b].

Nous allons étudier cette convergence quand l'ensemble $\{x_j\}$ $(j \geq 0)$ devient dense dans [a,b]. Cela nous amène à étudier le comportement de F_j^+ quand $x_{j+1}-x_j$ converge vers zéro. Pour ne pas compliquer les notations nous allons étudier le comportement de F^+ défini par (1.9) quand (b-a) converge vers zéro mais auparavant nous allons calculer une borne supérieure de $\Delta^+(x) = F^+(x)-f(x)$ dans [a,b].

PROPOSITION 1.2

$$\Delta^+(x) \leq \left(\frac{b-a}{2}\right)^{q-1/2} \left\{ |f| \left\{ \frac{1}{\sqrt{2q!}} + \frac{1}{(2q-1)^{1/2}(q-1)!} \right\} + W \right\} \quad \forall x \in [a,b]$$

Démonstration :

Soit $\Delta_a^+(x) = F_a^+(x) - f(x)$. D'après les formules (1.7), on a

$$\Delta_a^+(x) = P_a(x) - f(x) + W(x-a)^{\frac{2q-1}{2}}$$

$$\Delta_a^+(x) = \alpha_q (x-a)^q + \sum_{i=0}^{q-1} \alpha_i (x-a)^i - f(x) + W(x-a)^{\frac{2q-1}{2}}$$

$$\Delta_a^+(x) \leq |\alpha_q| (x-a)^q + \left| \sum_{i=0}^{q-1} \alpha_i (x-a)^i - f(x) \right| + W(x-a)^{\frac{2q-1}{2}} \quad \forall x \in [a,b]$$

Nous allons majorer chacun des termes sur $[a, \frac{a+b}{2}]$ pour obtenir une borne en fonction de $(b-a)/2$.

$$|\alpha_q| (x-a)^q = \left| \frac{f^{(q-1)}(b) - f^{(q-1)}(a)}{(b-a)q!} \right| (x-a)^q$$

$$= \left| \frac{f^{(q-1)}(b) - f^{(q-1)}(a)}{q!} \right| \left(\frac{x-a}{b-a}\right) (x-a)^{q-1}$$

d'où

$$|\alpha_q| (x-a)^q \leq \left| \frac{f^{(q-1)}(b) - f^{(q-1)}(a)}{2q!} \right| \left(\frac{b-a}{2}\right)^{q-1} \quad \forall x \in [a, \frac{a+b}{2}]$$

D'autre part comme $f^{(q-1)} \in H^1(a,b)$ on a :

$$|f^{(q-1)}(b) - f^{(q-1)}(a)| \leq (b-a)^{1/2} |f|$$

d'où

$$|\alpha_q|(x-a)^q \leq \frac{|f|}{\sqrt{2}\,q!}\left(\frac{b-a}{2}\right)^{q-1/2} \qquad \forall x \in [a, \tfrac{a+b}{2}]$$

Pour majorer le deuxième terme on utilise l'équation (1.3) avec $\varphi = f$ et compte tenu de la définition des coefficients α_i $(i = 0,\ldots,q-1)$ (eq. 1.5) on obtient :

$$\left|\sum_{i=0}^{q-1} \alpha_i (x-a)^i - f(x)\right| \leq \frac{|f|}{(2q-1)^{1/2}(q-1)!}\left(\frac{b-a}{2}\right)^{q-1/2} \qquad \forall x \in [a, \tfrac{a+b}{2}]$$

Finalement pour le troisième terme de la majoration de Δ_a^+ on a :

$$W(x-a)^{\frac{2q-1}{2}} \leq W\left(\frac{b-a}{2}\right)^{q-1/2} \qquad \forall x \in [a, \tfrac{a+b}{2}]$$

En remplaçant ces trois majorations on obtient la borne suivante de Δ_a^+

$$\Delta_a^+(x) \leq \left(\frac{b-a}{2}\right)^{q-1/2}\left\{|f|\left\{\frac{1}{\sqrt{2}\,q!} + \frac{1}{(2q-1)^{1/2}(q-1)!}\right\}+W\right\} \qquad \forall x \in [a, \tfrac{a+b}{2}]$$

De façon analogue on montre que la fonction $\Delta_b^+(x) = F_b^+(x) - f(x)$ est bornée par cette même quantité mais sur $[\frac{a+b}{2}, b]$, d'où le résultat énoncé.

≡ ≡ ≡

Dans l'expression de Δ^+ interviennent les quantités W et $|f|$. Comme la semi-norme de f est calculée sur l'intervalle [a,b], elle converge vers zéro quand (b-a) → 0. De même, la constante W définie par l'équation (1.9) converge vers zéro quand (b-a) → 0 car comme $f^{(q-1)} \in H^1(a,b)$ on a (cf. [8]).

$$\lim_{b-a\to 0} \frac{[f^{(q-1)}(b)-f^{(q-1)}(a)]^2}{b-a} = 0$$

Nous avons donc établi le théorème de convergence suivant :

PROPOSITION 1.3

Soit $f \in H^q(x_1,x_2)$ et soit F^+ le majorant de f sur $[a,b] \subseteq [x_1,x_2]$ défini par l'équation (1.9). Alors

$$\lim_{b-a\to 0} \frac{F^+(x)-f(x)}{(\frac{b-a}{2})^{q-1/2}} = 0$$

≡ ≡ ≡

Si l'on fait l'hypothèse additionnelle que $f \in C^q(x_1,x_2)$ alors

$$|f|^2 \leq (b-a) \max_{x\in[a,b]} |f^{(q)}(x)|^2$$

et comme

$$W^2 \leq |f|^2$$

alors $|f|$ et W convergent vers zéro avec $(b-a)^{1/2}$ d'où la proposition.

PROPOSITION 1.4

Soit $f \in C^q(x_1,x_2)$ et soit F^+ la majorant de f sur $[a,b] \subseteq [x_1,x_2]$ défini par l'équation (1.9) Alors :

$$F^+(x)-f(x) = O((\frac{b-a}{2})^q)$$

≡ ≡ ≡

Nous allons utiliser la proposition 1.3 pour construire une suite de majorants de f convergeant vers f sur [a,b]. Nous définissons chaque majorant F_n^+ en fonction d'une famille T_n de sous-intervalles de [a,b] et nous démontrerons que si la suite $\{T_n\}$, $n \geq 0$ satisfait certaines hypothèses, la suite des majorants $\{F_n^+\}$ converge ponctuellement vers f sur [a,b].

PROPOSITION 1.5

Soit $f \in H^q(a,b)$ et soit $T_n = \{I_{n,j},\ j = 1,\ldots,m_n\}$, $m_n \in \mathbb{N}$, $n \geq 1$, une famille de sous-intervalles fermés de [a,b] tels que l'intersection de deux intervalles quelconques de la famille est au plus réduite à un point.

Soit $S_n = \bigcup_{j=1}^{m_n} I_{n,j}$ et $\lambda_n = \max_{j=1,..,m_n} \text{mes}(I_{n,j})$.

Soit $F_{n,j}^+$ le majorant de f construit sur $I_{n,j}$ et soit F_n^+ définie par

$$F_n^+(x) = \begin{cases} f(x) \text{ si } x \notin S_n \\ F_{n,j}^+(x) \text{ si } x \in I_{n,j} \end{cases}$$

Alors, sous les hypothèses

1°) $S_{n+1} \subseteq S_n \qquad \forall n \geq 1$

2°) $\lim_{n\to\infty} \lambda_n = 0$

la suite $\{F_n^+\}$, $n \geq 0$ converge ponctuellement vers f dans [a,b].

Démonstration

Remarquons d'abord que la fonction F_n^+ est bien définie et continue car chaque fonction $F_{n,j}^+$ interpole f aux extrêmités de l'intervalle $I_{n,j}$.

Soit $S_\infty = \bigcap_{n\geq 0} S_n$, alors quelque soit $x \in [a,b]$ on a deux cas possibles

a) si $x \notin S_\infty$, alors, étant donné que $S_{n+1} \subseteq S_n \quad \forall n \geq 0$, il existe $\hat{n} \in \mathbb{N}$ tel que $x \notin S_n \quad \forall n \geq \hat{n}$, d'où

$$F_n^+(x) = f(x) \qquad \forall n \geq \hat{n}$$

b) si $x \in S_\infty$, alors $x \in S_n \quad \forall n \geq 0$ et pour chaque n il existe un intervalle I_{n,j_n} contenant le point x d'où

$$F_n^+(x) = F_{n,j_n}^+(x) \qquad \forall n \geq 0$$

D'après la proposition 1.3 on a sur I_{n,j_n}

$$|F_n^+(x)-f(x)| = |F_{n,j_n}^+(x)-f(x)| \leq \left(\frac{mes(I_{n,j_n})}{2}\right)^{q-1/2} O(1) \qquad \forall n \geq 0$$

d'où

$$|F_n^+(x)-f(x)| \leq \left(\frac{\lambda_n}{2}\right)^{q-1/2} O(1) \qquad \forall n \geq 0$$

ce que, d'après l'hypothèse (2°), montre la convergence de $F_n^+(x)$ vers $f(x)$.

≡ ≡ ≡

Le cas le plus simple d'une suite $\{T_n\}$ satisfaisant les hypothèses de la proposition 1.5 est sans doute le cas où T_n est défini par un recouvrement de [a,b] par n sous-intervalles d'égale mesure (b-a)/n. Néanmoins ce cas ne présente aucun intérêt pratique car chaque majorant est indépendant des autres.

Dans la proposition 1.5 on suppose la suite $\{T_n\}$ $n \geq 0$ donnée et sur chaque famille T_n on construit le majorant F_n^+. Néanmoins la suite $\{T_n\}$ $n \geq 0$ peut être construite récursivement en utilisant T_n et F_n^+ dans la détermination de T_{n+1}. Au paragraphe suivant, nous définirons plus précisément un tel procédé.

V - METHODES DE LOCALISATION DANS $H^q(a,b)$

La proposition 1.5 donne un moyen de construire une suite de majorar convergeants vers f sur [a,b]. Nous allons construire dans ce paragraphe deux types de méthodes de localisation dont la convergence découlera de la propositi 1.5 et de la proposition générale 1.1 Dans la première classe de méthodes, les extrémités des intervalles de la famille T_{n+1} sont déterminés en fonction de F_n^+ tandis que dans la deuxième classe ils seront choisis dans une suite d'abscisse dense dans [a,b], préalablement fixé.

V.1 - PREMIERE CLASSE DE METHODES

La méthode que nous allons proposer est basée sur l'application itérée d'un procédé de base permettant de définir une suite T_n de familles de sous-ensembles satisfaisant aux hypothèses de la proposition 1.5. Le procédé est défini par la donnée d'un nombre entier $r \geq 2$ et un nombre réel θ arbitraire et il est tel qu'appliqué à une famille T de sous-intervalles d'un intervalle [a,b] il produit une nouvelle famille $\hat{T}$.

PROCEDE DE BASE P.B.

Soit $f \in H^q(a,b)$ et soit $T = \{I_j,\ j = 1,\ldots,m\}$ une famille de sous-intervalles de [a,b] tel que

$$\max \{x \in I_j\} \leq \min \{x \in I_{j+1}\} \quad (j = 1,\ldots,m-1).$$

Soit F_j^+ le majorant de f construit sur I_j $(j = 1,\ldots,m)$ et soit F^+ défini par :

$$F^+(x) = \begin{cases} f(x) \text{ si } x \notin \bigcup_{j=1}^{m} I_j \\ F_j^+(x) \text{ si } x \in I_j \end{cases}$$

Etant donné $\theta \in \mathbb{R}$, chaque ensemble $L_j = \{x \in I_j / F_j^+(x) \geq \theta\}$ est formé par un nombre minimal m_j d'intervalles disjoints, éventuellement réduits à un point. Nous obtenons ainsi $p = \sum_{j=1}^{m} m_j$ intervalles et on définit la nouvelle famille $\hat{T}$ comme la famille d'intervalles obtenue en recouvrant chacun des p intervalles par r sous-intervalles d'égale longueur. Il est clair que $\hat{T}$ est donc formé de $m = p.r$ intervalles. Nous noterons $\hat{T} = PB_f(T,\theta,r)$.

≡ ≡ ≡

Nous avons supposé les m_j intervalles constituant L_j disjoints, mais comme les intervalles de la famille T peuvent se toucher en un point il se peut que le sous-intervalle de I_j le plus à droite touche le sous-intervalle le plus à gauche de I_{j+1}. Nous ne raccordons pas ces intervalles de sorte que si l'on note $\hat{I}_i$ $(i = 1,\ldots,\hat{m})$ les intervalles de $\hat{T} = PB_f(T,\theta,r)$ on aura toujours que chaque intervalle $\hat{I}_i \in \hat{T}$ sera complètement inclus dans un intervalle de $I_j \in T$ et on a la relation :

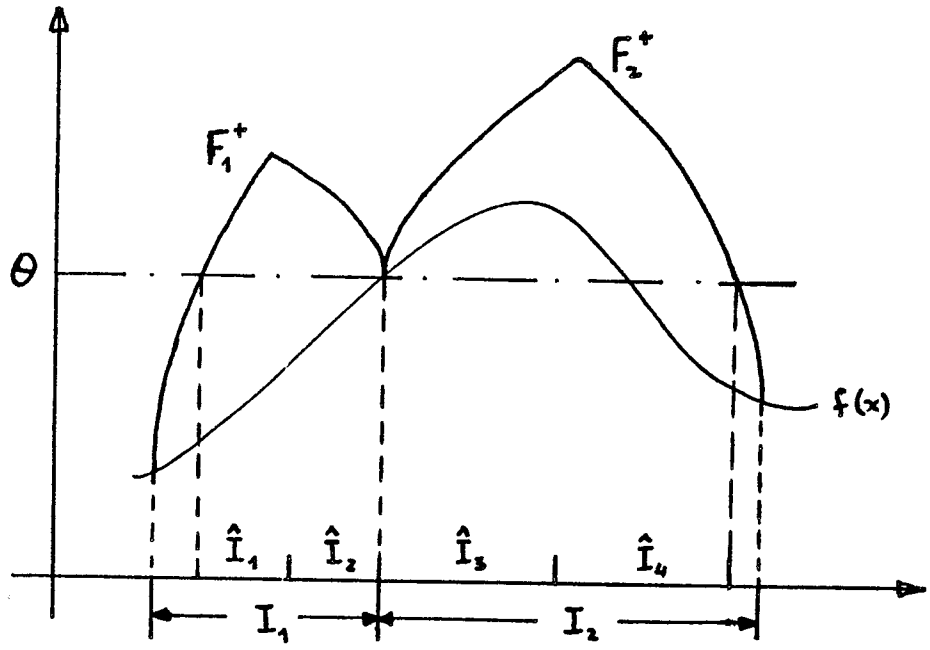
$$\text{mes}(\hat{I}_i) \leq \frac{\text{mes}(I_j)}{r} \qquad \forall I_i \subseteq I_j,\ \hat{I}_i \in \hat{T},\ I_j \in T$$

l'égalité n'ayant lieu que si $I_j = L_j$.
Nous obtenons donc la relation

$$\hat{\lambda} = \max_{i=1,..m} \text{mes}(\hat{I}_i) \leq \frac{1}{r} \max_{j=1,..,m} \text{mes}(I_j) = \frac{\lambda}{r} \tag{1.10}$$

Le procédé PB nous assure donc, pour $r \geq 2$, que la mesure des intervalles de PB(T) est toujours inférieure à la moitié de la mesure des intervalles de T.

Dans la figure ci-dessous, nous schématisons, pour r = 2 le résultat du procédé PB dans le cas de deux intervalles qui se touchent.



Le nombre entier $r \geq 2$ et le nombre réel θ étant arbitraires, nous pouvons construire à partir d'une famille T_1 une suite $\{T_n\}$ $n \geq 1$ de familles d'intervalles par itération du procédé P.B en associant à chaque T_n un entier $r_n \geq 0$ et un nombre réel θ_n. Plus précisément nous pouvons construire une suite $\{T_n\}$, à partir d'une suite de nombres entiers $\{r_n\}$, $r_n \geq 2$ et d'une suite de nombres réels $\{\theta_n\}$, $n \geq 0$ par l'algorithme simple.

(1.11)

1°) $T_1 = \{I_{1,j},\ j=1,\ldots r_o\}$ avec $I_{1,j}=[x_j,x_{j+1}]$, $x_1=a$, $x_j=x_1+jh$, $h=$

2°) $T_{n+1} = PB_f(T_n,\theta_n,r_n) = \{I_{n+1,j} \quad (j = 1,\ldots,m_{n+1})\}$

Nous allons montrer que la suite $\{T_n\}$ satisfait aux hypothèses de la proposition 1.5 et alors la suite $\{F_n^+\}$ de fonctions associées à $\{T_n\}$ sera une suite convergente de majorants de f.

PROPOSITION 1.6

La suite $\{T_n\}$ définie par l'algorithme 1.11 satisfait aux hypothèses de la proposition 1.5 quelque soit $\{r_n\}$, $r_n \geq 2$ $r_n \in \mathbb{N}$ et quelque soit $\{\theta_n\}$, $\theta_n \in \mathbb{R}$, $n \geq 1$.

Démonstration

Soit $T_n = \{I_{n,j} \quad (j = 1,\ldots,m_n)\}$ et soit

$$S_n = \bigcup_{j=1}^{m_n} I_{n,j}$$

$$\lambda_n = \max_{j=1,\ldots,m_n} \text{mes}(I_{n,j})$$

La condition $S_{n+1} \subseteq S_n \quad \forall n \geq 1$ est satisfaite par construction de T_{n+1} à partir de T_n, car chaque intervalle de T_{n+1} est inclus dans un intervalle de T_n.

Quant à la condition $\lim_{n\to\infty} \lambda_n = 0$ nous avons d'après (1.10) que

$$\lambda_n \leq \frac{\lambda_{n-1}}{r_n} \leq \frac{\lambda_{n-1}}{2} \leq \frac{\lambda_1}{2^n} \qquad \forall n \geq 2$$

et comme $\lambda_1 = \frac{b-a}{r_o} \leq \frac{b-a}{2}$ on a :

$$\lambda_n \leq \left(\frac{b-a}{2^{n+1}}\right) \qquad \forall n \geq 1$$

d'où

$$\lim_{n\to\infty} \lambda_n = 0$$

≡ ≡ ≡

Nous pouvons donc utiliser la proposition 1.5 pour démontrer finalement la proposition :

PROPOSITION 1.7

Soit $f \in H^q(a,b)$, $\{r_n\}$, $r_n \geq 2$ une suite de nombres entiers et soit $\{\theta_n\}$ une suite de nombres réels. Soit $\{T_n\}$ la suite définie par l'algorithme 1.11, $T_n = \{I_{n,j},\ (j = 1,\ldots,m_n)\}$. Alors la suite de fonctions

$$F_n^+(x) = \begin{cases} f(x) & \text{si } x \notin \overset{m_n}{\underset{j=1}{\cup}} I_{n,j} \\ F_{n,j}^+(x) & \text{si } x \in I_{n,j} \end{cases}$$

converge ponctuellement vers f dans [a,b] indépendamment du choix des suites $\{r_n\}$et $\{\theta_n\}$.

≡ ≡ ≡

Si l'on dispose d'une suite $\{k_n^-\}$ satisfaisant aux hypothèse $(H.1)^-$ et $(H.2)^-$ nous pouvons construire une suite convergente de localisations $\{L_n^+\}$ et l'ensemble $E^+ = \{x \in [a,b]/\ f(x) \geq k\}$ en utilisant la proposition 1.1 En principe la suite $\{\theta_n\}$ qui sert à définir le procédé $P.B_f$ peut être choisie indépendante de la suite $\{k_n^-\}$ mais cela entraînerait un double calcul, d'abord avec k_n^- pour déterminer la localisation L_{n+1}^+ et puis avec θ_n pour déterminer la nouvelle famille T_{n+1}. En prenant comme $\{\theta_n\}$ précisément la suite $\{k_n^-\}$ nous déterminons simultanément la localisation L_{n+1}^+ et la nouvelle famille T_{n+1}

Il faut remarquer, néanmoins, que l'ensemble S_n qui est la réunion des intervalles de T_n peut différer de L_n^+ car dans la construction de T_n nous avons délaissé les éventuels points isolés qui apparaissent.

Ces points isolés peuvent être repérés à condition que dans chaque intervalle $I_{n,j} \in T_n$ on puisse expliciter l'ensemble $\{x \in I_{n,j}/F_{n,j}^+(x) \geq k_n^-\}$ en fonction des racines de l'équation

$$F_{n,j}^+(x) = k_n^-$$

Si l'on suppose que le procédé P.B appliqué à l'ensemble T_n permet de déterminer en plus de la nouvelle famille d'intervalles T_{n+1} les points isolés, alors la localisation L_n^+ peut être parfaitement définie. Comme L_n^+ sera la réunion de S_n et de l'ensemble de points isolés provenant de l'itération précédente, on aura en général :

$$L_n^+ = S_n \cup \{t_j^n,\ j = 1,\ldots,p_n\} \quad p_n \geq 0 \tag{1.12}$$

et

$$L_{n+1}^+ = \{x \in S_n \ / \ F_n^+(x) \geq k_n^-\} \cup \{t_j^n \ / \ F_n^+(t_j^n) \geq k_n^-\}$$

ce qui compte tenu que $F_n^+(x) = f(x)$ pour $x \notin S_n$ s'écrit

$$L_{n+1}^+ = \{x \in S_n \ / \ F_n^+(x) \geq k_n^-\} \cup \{t_j^n \ / \ f(t_j^n) \geq k_n^-\}$$

Nous savons donc que les points isolés de L_{n+1}^+ seront les points isolés de L_n^+ qui "passent le test" : $f(t_j^n) \geq k_n^-$ et les nouveaux points isolés qui apparaitront dans le passage de T_n à T_{n+1}.

≡ ≡ ≡

Tout ce que nous avons dit relativement à E^+ peut être facilement répété pour l'ensemble E^-. Si l'on veut, comme c'est notre cas, localiser l'ensemble $E = E^+ \cap E^-$, le procédé de base PB_f doit être redéfini pour tenir compte de la présence des minorants et de la nouvelle suite $\{k_n^+\}$. Cela se fait facilement, en considérant pour chaque intervalle $I_{n,j} \in T_n$, au lieu de l'ensemble $\{x \in I_{n,j} \ / \ F_{n,j}^+(x) \geq k_n^-\}$, son intersection avec l'ensemble $\{x \in I_{n,j} / F_{n,j}^-(x) \leq k_n^+\}$.

Nous pouvons ainsi définir l'algorithme suivant :

ALGORITHME 1

Etant donné $\{r_n\}$, $r_n \geq 2$, $r_n \in \mathbb{N}$ et deux suites $\{k_n^-\}$ et $\{k_n^+\}$ satisfaisant aux hypothèses (H.1) et (H.2) on construit une suite $\{L_n\}$ de localisation de $E = \{x \in [a,b]/f(x) = k\}$ par l'algorithme suivant :

0) Initialisation : $L_1 = [a,b]$

$$T_1 = \{[x_i, x_{i+1}],\ x_1 = a_1,\ x_i = x_{1+ih} (i=0,..,r_o),\ h = \frac{b-a}{r_o}\}$$

$n = 1$

1) $T_{n+1} = L_{n+1} = \emptyset$

2) Pour chaque intervalle $I_{n,j} \in T_n$ faire

2.1) Calculer $F^+_{n,j}$ et $F^-_{n,j}$

2.2) Déterminer $L_{n+1,j} = \{x \in I_{n,j} / F^+_{n,j}(x) \geq k_n^- \text{ et } F^-_{n,j}(x) \leq k_n^+\}$ et $T_{n+1,j}$ famille d'intervalles obtenues en recouvrant chaque intervalle propre de $L_{n+1,j}$ (non réduit à un point) par r_n intervalles d'égale longueur.

2.3) Faire $L_{n+1} = L_{n+1} \cup L_{n+1,j}$ et $T_{n+1} = T_{n+1} \cup T_{n+1,j}$

3) Test d'arrêt : si le test est positif aller à (5)

4) $n := n+1$;

5) Arrêt.

≡ ≡ ≡

LES SUITES $\{k_n^-\}$ ET $\{k_n^+\}$

Nous avons montré dans la proposition 1.7 que la suite de majorants $\{F_n^+\}$ converge vers f dans [a,b] indépendamment du choix de la suite $\{\theta_n\}$ qui sert à définir T_n $(n \geq 0)$. Dans le cas où k est connu nous prendrons évidemment $k_n^- = k_n^+ = k \quad \forall n \geq 0$. D'autre part, il se peut que les suites soient le résultat d'une autre méthode qui sert à calculer itérativement la valeur k et dans ce cas l'algorithme 1 représente, en conjonction avec l'algorithme qui sert à encadrer k, $k_n^- \leq k \leq k_n^+ \quad \forall n \geq 0$, une méthode doublement infinie et convergente. Cette méthode peut être particulièrement intéressante pour le calcul, avec une précision donnée, des racines de $f(x) = k$ sur [a,b] quand chaque itération pour encadrer k est d'un coût très élevée.

Dans le problème de localisation du maximum global, le problème se réduit à la localisation de E^+ et par conséquence il suffit de déterminer la suite $\{k_n^-\}$. Nous pouvons calculer la suite $\{k_n^-\}$ récursivement en prenant $k_1 = \max_{j=1,..,r_o+1} f(x_j)$ et, une fois la famille T_n déterminée, définir k_n^- comme le maximum des nombres k_{n-1}^- et le maximum discret de f pris aux extrémités des intervalles de T_n et aux points isolés de L_n^+. Cette façon de procéder ne demande aucune évaluation additionnelle de la fonction car toutes ces valeurs servent dans la détermination de L_{n+1}^+.

De façon explicite, si $T_n = \{I_{n,j} = [x_{n,j}^{(1)}, x_{n,j}^{(2)}] \ , \ j = 1,\ldots,m_n\}$ et $L_n^+ = S_n \cup \{t_j^n, \ j = 1,\ldots,p_n\}$ (eq. 1.12), on définit la suite $\{k_n^-\}$ par

$$(1.13) \qquad \begin{cases} k_o^- = -\infty \\ k_n^- = \max\{k_{n-1}^-, \max\limits_{\substack{j=1,..,m_n \\ i=1,2}} f(x_{n,j}^{(i)}), \max\limits_{j=1,\ldots,p_n} f(t_j^n)\} \end{cases}$$

Nous allons démontrer que cette suite converge en croissant vers k.

PROPOSITION 1.8

La suite $\{k_n^-\}$ définie par l'équation (1.13) est telle que

$$\lim_{n\to\infty} \uparrow k_n^- = k = \max_{x\in[a,b]} f(x)$$

Démonstration

La suite $\{k_n^-\}$ est une suite monotone et bornée par k ; elle converge donc vers une limite $\hat{k} \leq k$. En utilisant l'algorithme 1 la suite $\{L_n^+\}$ est telle que $\bigcap_{n\geq 0} L_n^+ = \hat{E}^+ = \{x \in [a,b]/f(x) \geq \hat{k}\}$. Soit $\bar{x} \in a,b$ un point tel que $f(x) = \hat{k}$ alors $\bar{x} \in \hat{E}^+$, d'ou $\bar{x} \in L_n^+ \quad \forall n \geq 0$.

Si pour $n_o \in \mathbb{N}$, $\bar{x}$ est un point isolé de $L_{n_o}^+$ on a, d'après la définit 1.13, que $k_{n_o}^- = k$ et $k_n^- = k \quad \forall n \geq n$.

Si $\bar{x}$ n'est point isolé d'aucune des localisations L_n^+, alors pour chaq $n \in \mathbb{N}$ il existe un intervalle $I_{n,j_n} \in T_n$ tel que $\bar{x} \in I_{n,j_n} \subseteq L_n^+$. Notons $[\alpha_n, \beta_n]$ l'intervalle I_{n,j_n}, alors

$$k_n^- \geq \max \{f(\alpha_n), f(\beta_n)\}$$

et comme $\text{mes}(I_{n,j_n}) = \beta_n - \alpha_n$ converge vers zéro quand $n\to\infty$ on a

$$\hat{k} = \lim_{n\to\infty} k_n^- \geq \lim_{n\to\infty} \max \{f(\alpha_n), f(\beta_n)\} = f(\bar{x}) = k$$

d'où $\hat{k} \geq k$ ce qui démontre que $\hat{k} = k$.

≡ ≡ ≡

V.2 - DEUXIEME CLASSE DE METHODES DE LOCALISATION

La méthode définie par l'algorithme 1 repose sur la détermination pratique des ensembles du type $\{x \in [\alpha,\beta]/F^+(x) \geq \theta\}$ où F^+ est un majorant de f construit sur [a,b] qui dépend des valeurs de f et de certaines de ses dérivées en α et β . En général, telle qu'a été définie la méthode, les points α et β ne sont pas nécessairement extrémités des nouveaux intervalles de sorte que une fois L_n^+ calculée, les valeurs de f et ses dérivées aux points où elles ont été évaluées ne servent plus dans la détermination des localisations suivantes.

D'autre part, dans certains problèmes f et ses dérivées peuvent être données au moyen d'une table. Nous aurions donc intérêt à construire une méthode qui n'utilise que des abscisses fixées à l'avance et dans laquelle au moins une partie des évaluations de f et ses dérivées puisse être réutilisée.

Nous construirons une telle méthode par une modification de celle que nous avons développées au paragraphe précédent. L'idée de cette nouvelle classe de méthodes est simple : si une partie d'un des intervalles de T_n appartient à L_{n+1}^+ on gardera l'intervalle tout entier. Cela revient à tester chacun des intervalles de T_n et à garder ceux qui sont suspects de contenir la solution. Cette technique a été employée, comme nous l'avons remarqué dans l'introduction, pour la localisation des racines de polynômes dans le plan complexe.

Le test sur chaque intervalle est défini en utilisant le maximum du majorant sur l'intervalle de sorte que si ce maximum est plus petit que k_n^- on est assuré que le maximum n'appartient pas à cet intervalle. Dans le cas contraire on n'est pas assuré que le maximum global soit atteint dans l'intervalle, lequel n'est que suspect de le contenir. Il s'agit donc, avec la nomenclature employée par HENRICI et GARGANTINI [8] d'un test d'exclusion.

Dans la même ligne d'idée de la méthode de la section précédente nous construirons une suite $\{T_n\}$ de familles d'intervalles de sorte que chaque intervalle de T_{n+1} soit complètement inclus dans un intervalle de T_n.

Soit $f \in H^q(a,b)$ et soit I un intervalle contenu dans [a,b]. Nous noterons $M_f^+(I)$ la quantité

$$M_f^+(I) = \max_{x \in I} F^+(x) \tag{1.14}$$

où F^+ est le majorant de f construit sur I (même définition pour $M_f^-(I)$ en changeant F^+ par F^-).

Considérons le problème de localiser l'ensemble $E^+ = \{x \in [a,b]/f(x)$ et définissons, quelque soit $\theta \leq k$ et pour tout intervalle $I \subseteq [a,b]$ le test : $M_f^+(I) \geq \theta$. Si l'intervalle I "ne passe pas" le test, autrement dit si $M_f(I) < \theta$ on est assuré, d'après la définition 1.14, que $I \cap E^+ = \emptyset$ car

$$f(x) \leq F^+(x) \leq M_f^+(I) < \theta \leq k \qquad \forall x \in I$$

Considérons maintenant une suite $\{k_n^-\}$ satisfaisant les hypothèses $(H.1)^-$ et $(H.2)^-$ et soit $\{x_j\}$ $j \geq 0$, une suite d'abscisses dense dans [a,b] et telle que $x_0 = a$, $x_1 = b$, $x_i \neq x_j$ si $i \neq j$. Pour chaque entier $n \geq 2$ on notera $\{p_j^n\}_{j=0}^n$ l'ensemble ordonné qui coïncide avec les n+1 premiers terme de la suite $\{x_j\}$, $j \geq 0$. Etant donné une suite $\{r_n\}$, $n \geq 0$ de nombres entiers, $r_{n+1} > r_n$, nous définissons une suite $\{T_n\}$ de familles d'intervalles par

1°) $$T_o = \{[p_{j-1}^{r_o}, p_j^{r_o}], \ j = 1,\ldots,r_o\}$$

2°) Etant donné $T_n = \{I_{n,j}, \ j = 1,\ldots,m_n\}$ on construit T_{n+1} par le procédé suivant :

2.1 On construit l'ensemble $S_{n+1} = \bigcup_{j=1}^{m_n} \{I_{n,j}/M_f^+(I_{n,j}) \geq k_n^-\}$

2.2 $T_{n+1} = \{I_{n+1,j} = [p_{j-1}^{r_{n+1}}, p_j^{r_{n+1}}] \ j = 1,\ldots,r_{n+1} / I_{n+1,j} \subseteq S_{n+1}\}$

} 1.

≡ ≡ ≡

Le procédé consiste donc à tester chaque ensemble $I_{n,j} \in T_n$ et recouvrir les intervalles qui "passent le test" par des sous-intervalles formées à partir des points de $\{p_j^{r_{n+1}}\}$ $(j = 0,\ldots,r_{n+1})$ qui appartiennent à l'intervalle testé positivement.

Le procédé est bien défini car les extrémités de chaque intervalle $I_{n,j}$ sont des abscisses de $\{p_j^{r_n}\}$ $(j = 0,\ldots,r_n)$ et

$$\{p_j^{r_{n+1}}\} = \{p_j^{r_n}\} \cup \{x_j\}_{j=r_n+1}^{r_{n+1}}$$

Comme les points de $\{x_j\}$ $j \geq 0$ qui sont en dehors de S_{n+1} n'interviennent plus dans la construction des T_j, $j \geq n+1$, il suffit de définir la suite $\{x_j\}$, $j \geq 0$ au fur et à mesure, si elle n'est pas donnée à l'avance, par exemple en recouvrant chaque intervalle $I_{n,j} \in T_n$ qui passe le test, par un certain nombre de sous-intervalles d'égale longueur.

Dans ce procédé, il n'y a pas de points isolés et la localisation L_{n+1}^+ coïncide avec l'ensemble S_{n+1}. On a donc

$$L_{n+1}^+ = \cup \{I_{n,j} \in T_n \,/\, M_f^+(I_{n,j}) \geq k_n^-\} \tag{1.16}$$

Nous allons démontrer que cette suite de localisation converge vers E^+.

PROPOSITION 2.9

Soit $f \in H^q(a,b)$, $\{k_n^-\}$ une suite satisfaisant les hypothèses $(H.1)^-$ et $(H.2)^-$, $\{x_j\}$ $j \geq 0$ une suite d'abscisses dense dans $[a,b]$, $x_0 = a$, $x_1 = b$, $x_i \neq x_j$ si $i \neq j$ et $\{r_n\}$ une suite de nombres entiers tel que $r_{n+1} > r_n$. Alors la suite d'ensembles définie par 1.15 et 1.16 est une suite convergente de localisations de E^+.

Démonstration

1°) Soit $x \in L_\infty^+ = \bigcap_{n\geq 0} L_n^+$. Alors $\forall n \geq 0$, il existe j_n tel que

$$x \in I_{n,j_n} \qquad \forall n \geq 0$$

et

$$M_f^+(I_{n,j_n}) \geq k_n^-$$

D'après la proposition 1.3, si F_{n,j_n}^+ est le majorant de f construit sur I_{n,j_n} on a :

$$F_{n,j_n}^+(t) = \left(\frac{mes(I_{n,j_n})}{2}\right)^{q-1/2} O(1) + f(t)$$

d'où

$$k_n^- \leq M_f^+(I_{n,j_n}) = \max_{t\in I_{n,j_n}} F_{n,j_n}^+(t) \leq \left(\frac{mes(I_{n,j_n})}{2}\right)^{q-1/2} O(1) + \max_{t\in I_n}$$

Comme $\{x_j\}$ $j \geq 0$ est dense dans [a,b], la $mes(I_{n,j_n})$ converge vers zéro quand $n\to\infty$ et comme $\lim_{n\to\infty} k_n^- = k$, on obtient :

$$k = \lim_{n\to\infty} k_n^- \leq \lim_{n\to\infty} \max_{t\in I_{n,j_n}} f(t)$$

Finalement, comme $f \in H^q(a,b)$, $q \geq 1$ et $x \in I_{n,j_n}$ $\forall n \geq 0$ on a

$$\lim_{n\to\infty} \max_{t\in I_{n,j_n}} f(t) = f(x)$$

d'où $f(x) \geq k$ ce qui entraîne $x \in E^+$

On a donc : $L_\infty^+ \subseteq E^+$.

2°) Soit $x \in E^+$, alors $x \in L_0^+ = [a,b]$. Supposons que $x \in L_n^+$ $n \geq 1$ et soit $I_{n,j} \subseteq L_n^+$ un intervalle le contenant. Comme $f(x) \geq k$ on a :

$$M_f^+(I_{n,j}) \geq F_{n,j}^+(x) \geq f(x) \geq k \geq k_n^-$$

L'intervalle $I_{n,j}$ contenant x passe donc le test et on a :

$$x \in I_{n,j} \subseteq L^+_{n+1}$$

Alors, pour tout $n \in \mathbb{N}$, $x \in L^+_n$, d'où $x \in L^+_\infty$ ce qui montre que

$$E^+ \subseteq L^+_\infty$$

≡ ≡ ≡

La suite de localisation $\{L^+_n\}$ $n \geq 0$ définie par l'équation 1.15 est donc une suite qui converge vers E^+. Ci-dessous nous donnons l'algorithme qui permet la construction de cette suite.

ALGORITHME 2

0) Initialisation $L^+_o = [a,b]$

$$T_o = \{[p^{r_o}_{j-1}, p^{r_o}_j], \quad j = 1,\ldots,r_o\}$$

$$n = 0$$

1) $T_{n+1} = L^+_{n+1} = \emptyset$

2) Pour chaque intervalle $I_{n,j} \in T_n$ faire

2.1) Calculer $M^+_f(I_{n,j})$

2.2) si $M^+_f(I_{n,j}) \geq k^-_n$ alors faire

2.2.1) $L^+_{n+1} = L^+_{n+1} \cup I_{n,j}$

2.2.2) $T_{n+1} = T_{n+1} \cup \{I_{n+1,k}=[p^{r_{n+1}}_{k-1}, p^{r_{n+1}}_k] / I_{n+1,k} \subseteq I_{n,j}\ k=1,..r_{n+1}\}$

3) Test d'arrêt : si le test est positif aller à (5)

4) n := n+1 et aller à (1)

5) arrêt

≡ ≡ ≡

Remarque

Si la suite $\{x_j\}$, $j \geq 0$ n'est pas donnée à l'avance, on peut la construire en recouvrant chaque intervalle $I_{n,j}$ par r_n intervalles d'égale longueur. Cela modifie le point 2.2.2 de l'algorithme mais ne change pas la convergence de la méthode car on préserve la convergence de $mes(I_{n,j})$ vers zéro.

≡ ≡ ≡

Les algorithmes 1 et 2 demeurent théoriques. La mise en oeuvre requiert encore la résolution de plusieurs problèmes comme par exemple la "détermination" des ensembles du type $\{x \in [\alpha,\beta]/F^+(x) \geq \theta\}$, le calcul de $M_f^+(I)$ ou le calcul de la semi-norme de f intervenant dans l'expression de F^+. Nous ne pouvons pas donner une réponse à ces problèmes dans le cadre général des espaces H^q, mais nous le pouvons dans le cas de q = 1 et avec certaines modifications dans le cas q = 2.

L'étude de la mise en oeuvre effective des algorithmes que nous avons décrits dans ces deux cas sera l'objet du prochain chapitre.

LA SUITE $\{k_n^-\}$

Les remarques que nous avons faites à propos des suites $\{k_n^-\}$ et $\{k_n^+\}$ restent évidemment valables pour le nouvel algorithme à l'exception du calcul de la suite $\{k_n^-\}$ dans le cas du problème du calcul du maximum global.

Dans ce cas nous définissons pour chaque $n \in \mathbb{N}$

$$k_n^- = \max_{j=0,\ldots,r_n} f(x_j)$$

et comme la suite $\{x_j\}$, $j \geq 0$ devient dense dans $[a,b]$ on a

$$\lim_{n\to\infty} \uparrow k_n^- = k = \max_{x\in[a,b]} f(x)$$

Cette façon de construire la suite $\{k_n^-\}$ ne sera intéressante que si pour le calcul de k_n^- on peut se restreindre aux points x_j $(j = 0,\ldots,r_n)$ qui appartiennent à la localisation L_n^+ et cela est établie dans la proposition suivante :

PROPOSITION 1.11

$$k_n^- = \max_{j=0,\ldots,r_n} \{f(x_j) \;/\; x_j \in L_n^+\}$$

Démonstration

Nous allons démontrer par induction que pour tout $x \in [a,b]$ tel que $x \notin L_n^+$ on a $f(x) < k_n^-$.

Pour $n = 1$ soit $x \notin L_1^+$, alors il existe j_1 tel que $x \in I_{1,j_1} \in T_o$ et l'intervalle I_{1,j_1} ne passe pas le test d'où

$$f(x) \leq M_f(I_{1,j_1}) < k_1^-$$

Supposons que si $x \notin L_n^+$ alors $f(x) < k_n^-$ et soit $\bar{x} \notin L_{n+1}^+$, alors ou bien $\bar{x} \notin L_n^+$ d'où $f(\bar{x}) < k_n^- \leq k_{n+1}^-$, ou bien il existe $j_n \in \mathbb{N}$ tel que $I_{n,j_n} \in T_n$ et l'intervalle I_{n,j_n} ne passe pas le test, d'où

$$f(\bar{x}) \leq M_f(I_{n,j_n}) < k_n^- \leq k_{n+1}^-$$

≡ ≡ ≡

ENCADREMENT DU MONTANT DU MAXIMUM GLOBAL

Dans le problème de localisation du maximum global, l'algorithme 2, et avec un petit calcul additionnel l'algorithme 1, fournissent un moyen simple d'encadrer le montant $\alpha = \max_{x \in [a,b]} f(x)$.

En effet, si dans l'algorithme 1, on calcule à chaque itération la quantité $M_f^+(I_{n,j})$, on a

PROPOSITION 1.12

Soit $\lambda_n^+ = \max_{j=1,..m_n} M_f^+(I_{n,j})$, alors

$$k_n^- \leq \alpha \leq \lambda_n^+$$

≡ ≡ ≡

CHAPITRE - II

METHODES DE LOCALISATION DANS H^1 ET H^2

I - INTRODUCTION

Dans le chapitre I nous avons donné un procédé pour, étant donné une fonction $f \in H^q(a,b)$, construire un majorant et un minorant de f sur un sous-intervalle donné de [a,b]. L'utilisation de ces majorants et minorants et d'un procédé de base PB_f permettant la construction d'une suite $\{T_n\}$ de familles de sous-intervalles de [a,b] nous a permis de construire deux méthodes de localisation (algorithmes 1 et 2).

La mise en oeuvre de ces deux algorithmes repose sur la question préalable de pouvoir construire les majorants et minorants de f sur un intervalle donné, et cela dépend, d'après les formules 1.8, 1.3 et 1.7, de la connaissance des valeurs de f et ses dérivées jusqu'à l'ordre q-1 aux extrémités de l'intervalle donné mais surtout, de la connaissance explicite de la constante W ou d'une majoration de cette constante. Nous analyserons cette hypothèse dans le paragraphe 4 de ce chapitre.

En supposant connue la constante W, la mise en oeuvre de l'algorithme 1, dépend de la possibilité de pouvoir codifier en mémoire de l'ordinateur les ensembles du type $L^+ = \{x \in [a,b] / F^+(x) \geq k\}$. Nous dirons qu'un ensemble L^+ est "calculé" ou "déterminé" si l'on peut mettre en évidence les points extrêmes des intervalles qui le composent et l'ensemble de points isolés qui éventuellement lui appartiennent.

Quant à l'algorithme 2, sa mise en oeuvre requiert le calcul de la constante $M_f^+(I)$ associée à l'intervalle I donné, ou une majoration de cette constante. Etant donné la construction du majorant F^+ construit sur un intervalle donné I, la "détermination" des ensembles de type de L^+ et le calcul de $M_f^+(I)$ ne sont simples que pour q = 1 et q = 2, et même dans ce deuxième cas il faudra faire quelques modifications.

Dans le paragraphe 2, nous étudierons le problème du "calcul" de L^+ et de la constante $M_f^+(I)$ pour le cas q = 1. Le cas q = 2 sera l'objet de la troisième section.

II - METHODE DE LOCALISATION DANS H^1

Soit $f \in H^1(\alpha,\beta)$ et soit $I = [a,b] \subseteq [\alpha,\beta]$. Au chapitre I nous avons construit des majorants de f sur [a,b] et nous avons remarqué que la construction des minorants pouvait se faire de façon analogue. D'après les équations 1.7, 1.8 et 1.9 et en considérant les relations correspondantes aux minorants nous avons

$$P_a(x)-W(x-a)^{1/2} = F_a^-(x) \le f(x) \le F_a^+(x) = P_a(x)+W(x-a)^{1/2}$$

$$P_a(x)-W(b-x)^{1/2} = F_b^-(x) \le f(x) \le F_b^+(x) \le P_a(x)+W(b-x)^{1/2}$$

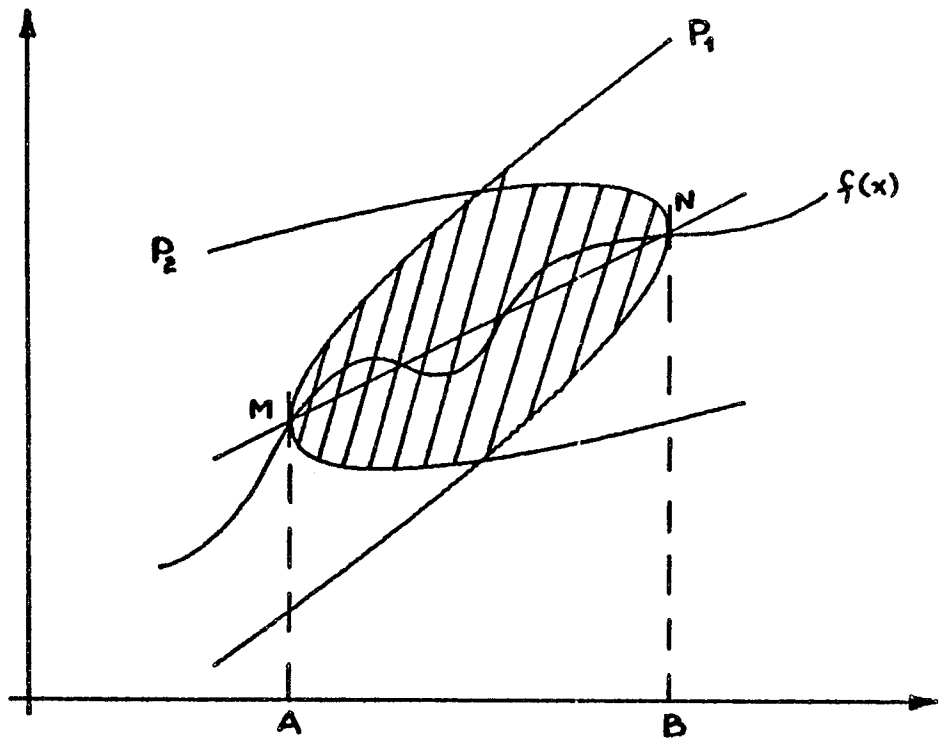
avec

$$W^2 = |f|^2 - \frac{\{f(b)-f(a)\}^2}{b-a} = |f|^2 - \alpha_1^2\,(b-a)$$

Ces relations nous permettent de définir le majorant et le minorant pour f:

$$F^-(x) = \max\{F_a^-(x),\ F_b^-(x)\} \le f(x) \le \min\{F_a^+(x),\ F_b^+(x)\} = F^+(x).$$

$P_a(x)$ est la droite qui interpole f aux noeuds a et b (points M et N du graphe ci- joint). Les fonctions F_a^+ et F_a^- sont les branches de la parabole P_1 : $(y-P_a(x))^2 = W^2(x-a)$ tandis que F_b^+ et F_b^- sont les branches de la parabole P_2 : $(y-P_a(x))^2 = W^2(b-x)$



La droite AM est tangente à la parabole P_1 en M et MN est la direction conjuguée. De même, BN est tangente à P_2 en N et MN est aussi la direction conjuguée de la direction BN.

La constante W^2 définit "l'ouverture" des deux paraboles et elle est telle que toutes les fonctions de $H^1(a,b)$ qui passent par M et N et qui ont une semi-norme inférieure ou égale à $|f|$ ont leur graphe entièrement situé dans la région hachurée.

Si on note $\bar{x} = \frac{a+b}{2}$ on vérifie que :

$$F_a^+(\bar{x}) = F_b^+(\bar{x})$$

$$F_a^+(x) \leq F_b^+(x) \qquad \forall x \in [a,\bar{x}]$$

$$F_a^+(x) \geq F_b^+(x) \qquad \forall x \in [\bar{x},b]$$

d'où

$$F^+(x) = \begin{cases} F_a^+(x) & \text{si } x \in [a,\bar{x}] \\ F_b^+(x) & \text{si } x \in [\bar{x}.b] \end{cases} \tag{2.6}$$

Des relations semblables peuvent être établies pour F^-.

Soit $k \in \mathbb{R}$. Nous allons expliciter l'ensemble $L^+ = \{x \in [a,b]/F^+(x) \geq k\}$ en fonction des racines de l'équation $F^+(x) = k$ et nous allons en même temps déterminer la quantité :

$$M^+ = \max_{x \in [a,b]} F^+(x)$$

Il faut distinguer les trois cas $\alpha_1 > 0$, $\alpha_1 = 0$ et $\alpha_1 < 0$. Nous ne développerons que le premier cas, les autres cas étant semblables.

Si $\alpha_1 > 0$, alors comme

$$(F_a^+)'(x) = \alpha_1 + \frac{W}{2(x-a)^{1/2}}$$

la fonction F_a^+ est strictement monotone croissante pour $x \geq a$.

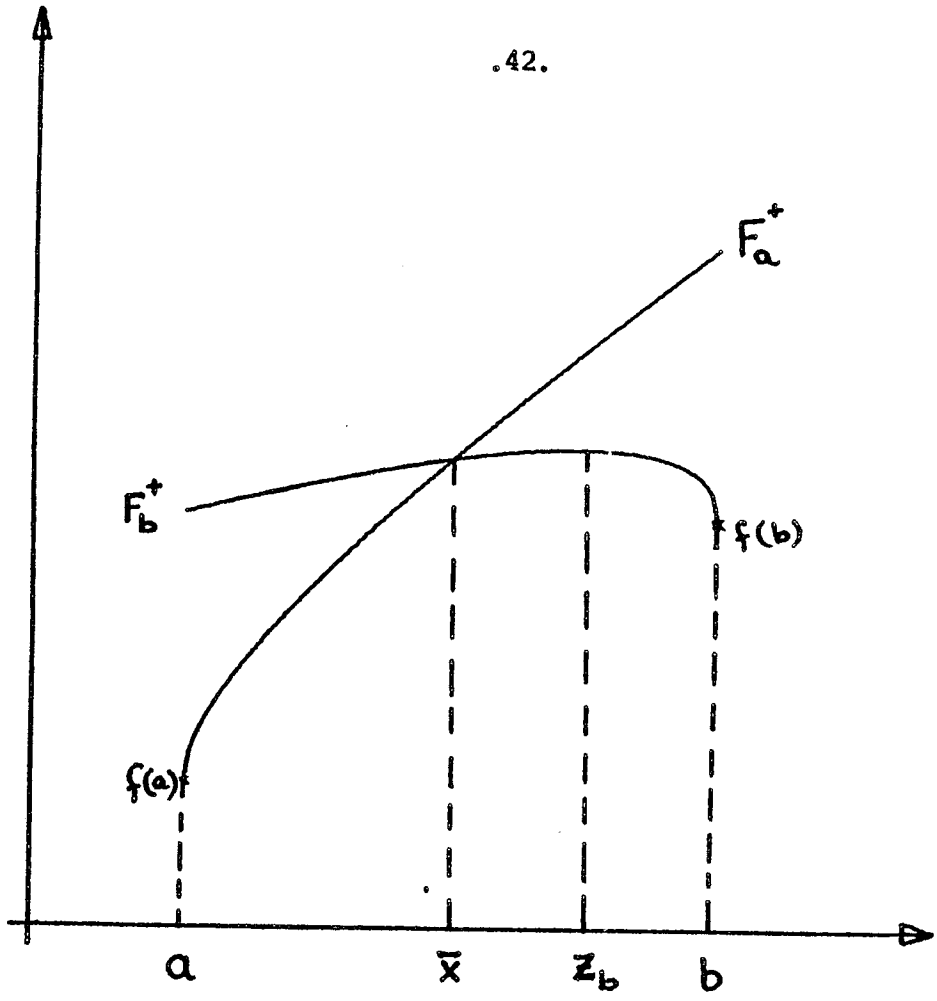
D'autre part :

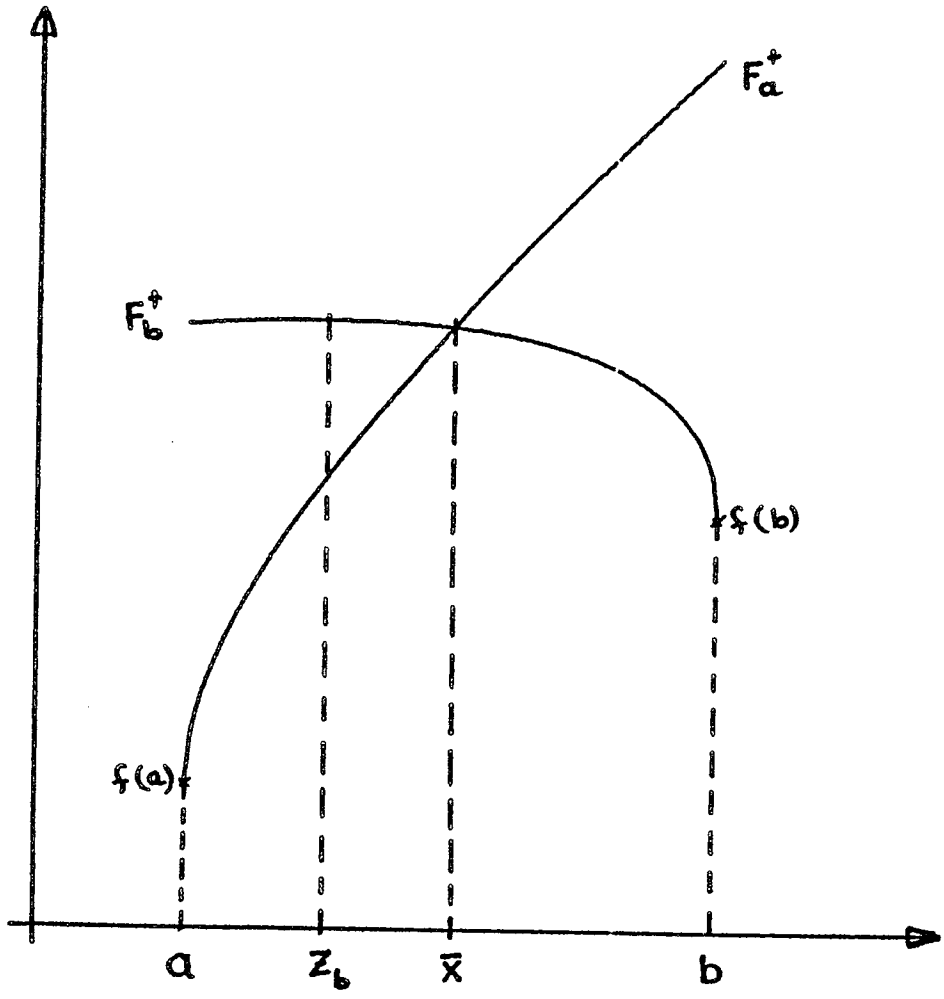
$$(F_b^+)'(x) = \alpha_1 - \frac{W}{2(b-x)^{1/2}}$$

alors, F_b^+ a un maximum au point.

$$Z_b = b - \left(\frac{W}{2\alpha_1}\right)^2$$

Selon la position relative de Z_b par rapport au point $\bar{x}$ on peut distinguer deux cas : si $\bar{x} \leq Z_b$, alors F_b^+ atteint son maximum dans $[\bar{x},b]$; si $Z_b \leq \bar{x}$, F_b^+ atteint son maximum en dehors de $[\bar{x},b]$ et elle est donc strictement monotone décroissante dans $[\bar{x},b]$. Dans le premier cas F^+ atteint son maximum au point Z_b et dans le deuxième cas elle l'atteint au point $\bar{x}$. Les deux cas possibles sont schématisés dans la figure ci-après :


F_a^+
F_b^+
f(b)
f(a)
a
$\bar{x}$
z_b
b


F_a^+
F_b^+
f(b)
f(a)
a
z_b
$\bar{x}$
b

La constante M^+ est donc donnée par

$$M^+ = \begin{cases} F_b^+(Z_b) & \text{si } \bar{x} \leq Z_b \\ F_a^+(\bar{x}) = F_b^+(\bar{x}) & \text{si } \bar{x} \geq Z_b \end{cases}$$

ou encore explicitant les valeurs des fonctions

$$M^+ = \begin{cases} F(b) + \dfrac{W^2}{4\alpha_1} & \text{si } \bar{x} \leq Z_b \\ \dfrac{f(a)+f(b)}{2} + W\left(\dfrac{b-a}{2}\right)^{1/2} & \text{si } Z_b \leq \bar{x} \end{cases}$$

Quant à l'ensemble L^+, il est vide si $k > M^+$ et il est toujours réduit à un seul intervalle si $k \leq M^+$. Pour $k = M^+$ l'intervalle est réduit au point Z_b si $\bar{x} \leq Z_b$ et au point $\bar{x}$ si $Z_b \leq \bar{x}$. Pour expliciter l'ensemble L^+ nous aurons à résoudre les équations :

$$F_a^+(x) = k \tag{2.7}$$

$$F_b^+(x) = k \tag{2.8}$$

ou plus précisément les équations

$$\alpha_1 (x-a)^2 + W(x-a) + f(a) - k = 0 \tag{2.9}$$

$$\alpha_1 (x-b)^2 - W(x-b) + f(b) - k = 0 \tag{2.10}$$

Notons $(x_a^+)^-$ et $(x_a^+)^+$ les racines de (2.9) et $(x_b^+)^-$, $(x_b^+)^+$ les racines de 2.10 avec, si elles sont réelles, $(x_a^+)^- \leq (x_a^+)^+$ et $(x_b^+)^- \leq (x_b^+)^+$. L'équation 2.7 a une seule racine réelle pour $k \geq f(a)$ et elle est $(x_a^+)^-$ tandis que l'équation 2.9 peut avoir une, deux ou aucune racine réelle suivant $k < f(b)$, $f(b) \leq k \leq \max_{x \in [a,b]} F_b^+(x)$ ou $k > \max_{x \leq [a,b]} F_b^+(x)$. Avec cette relation l'ensemble L^+ s'exprime selon les cas comme :

1°) $k \leq f(a)$, alors $L^+ = [a,b]$

2°) $f(x) \leq k \leq f(b)$ alors $L^+ = [\max\{(x_a^+)^-, (x_b^+)^-\}, b]$

3°) $k > f(b)$ alors $L^+ = [\max\{(x_a^+)^-, (x_b^+)^-\}, (x_b^+)^+]^*$

où l'étoile doit s'interpréter dans le sens que si $(x_b^+)^+ < (x_a^+)^-$ ou si $(x_b^+)^-$ et $(x_b^+)^+$ sont des nombres complexes alors $L^+ = \emptyset$.

Comme nous avons dit, les autres cas, $\alpha_1 = 0$ et $\alpha_1 < 0$ s'analysent de façon analogue. Le résultat complet comprenant tous les cas possibles aussi bien pour le calcul de M^+ comme par celui de L^+ sont résumés dans les deux propositions suivantes :

PROPOSITION 2.1

Soit $f \in H^1(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et F^+ le majorant de f construit sur I. Alors, la constante $M^+ = M_f^+(I) = \max_{x \in [a,b]} F^+(x)$ s'exprime :

$$M^+ = \begin{cases} \underline{\text{Si}}\ 2\alpha_1^2(b-a) \leq W^2 \ \ \underline{\text{alors}}\ \dfrac{f(a)+f(b)}{2} + W\left(\dfrac{b-a}{2}\right)^{1/2} \\ \underline{\text{sinon}} \begin{cases} f(a) - \dfrac{W^2}{4\alpha_1} & \text{si } \alpha_1 < 0 \\ f(b) + \dfrac{W^2}{4\alpha_1} & \underline{\text{si}}\ \alpha_1 > 0 \end{cases} \end{cases}$$

≡ ≡ ≡

Remarque :

Nous avions déjà trouvé ce résultat antérieurement, voir [1], mais en utilisant des propriétés d'un certain espace de Hilbert à noyau reproduisant préalablement étudié par M. DUC-JACQUET [8]. L'expression de M^+ a été légèrement modifiée pour des raisons de programmation, compte tenu du fait que la quantité $\alpha_1^2(b-a)$ est une valeur intermédiaire dans le calcul de W^2.

≡ ≡ ≡

PROPOSITION 2.2

Soit $f \in H^1(\alpha,\beta)$, $I = [a,b] \in [\alpha,\beta]$ et F^+ le majorant de f construit sur I. Soient :

$$D_a^+ = W^2 - 4\alpha_1 (f(a)-k)$$

$$D_b^+ = W^2 + 4\alpha_1 (F(b)-k)$$

Alors l'ensemble $L^+ = \{x \in [a,b] / F^+(x) \geq k\}$ s'écrit

1°) Si $\alpha_1 > 0$ alors

1.1 si $k \leq f(a)$ alors $L^+ = [a,b]$

1.2 si $f(a) < k < f(b)$ alors $L^+ = [\max\{(x_a^+)^-, (x_b^+)^-\}, b]$

1.3 si $k \geq f(a)$ alors $D_b^+ \geq 0$ $L^+ = [\max\{(x_a^+)^-, (x_b^+)^-\}, (x_b^+)^+]^*$

sinon $L^+ = \emptyset$

2°) Si $\alpha_1 < 0$ alors

2.1 si $k \leq f(b)$ alors $L^+ = [a,b]$

2.2 si $f(b) < k < f(a)$ alors $L^+ = [a, \min\{(x_a^+)^+, (x_b^+)^+\}]$

2.3 si $k \geq f(b)$ alors si $D_a^+ \geq 0$ $L^+ = [(x_a^+)^-, \min\{(x_a^+)^+, (x_b^+)^+\}]^*$

sinon $L^+ = \emptyset$

3°) si $\alpha_1 = 0$ alors

3.1 si $k \leq f(a)$ alors $L^+ = [a,b]$

3.2 si $k > f(b)$ alors $L^+ = [t_a^+, t_b^+]^*$

avec $t_a^+ = a + (\frac{k-f(a)}{W})^2$ et $t_b^+ = b - (\frac{k-f(b)}{W})^2$

≡ ≡ ≡

Nous rappellons que dans cette proposition $(x_a^+)^{\pm}$ et $(x_b^+)^{\pm}$ sont des racines des équations 2.9 et 2.10 respectivement et qu'un intervalle avec une étoile est un intervalle qui doit être considéré comme l'ensemble vide si l'une de ses bornes n'est pas définie (min ou max d'un nombre complexe) ou si l'ordre des extrêmités est inversé.

Les propositions que nous venons d'énoncer n'ont qu'un intérêt théorique car, comme il est bien connu (N. GASTINEL [13]), il n'existe pas de méthode effective pour déterminer dans laquelle des situations de la proposition 2.2 ou 2.1 on se trouve (même en considérant que α_1 est un nombre calculable). Néanmoins, les formules sont stables dans le sens que aussi bien la valeur de M^+ que les extrêmités des intervalles changent continuellement en passant d'une situation à l'autre ($\alpha_1 < 0$, $\alpha_1 = 0$, $\alpha_1 > 0$) dans les propositions 2.1 et 2.2 .

Les formules données dans les deux propositions précédentes servent à localiser l'ensemble $E^+ = \{x \in [a,b]/f(x) \geq k\}$. Pour localiser l'ensemble $E = \{x \in [a,b]/f(x) = k\}$ nous devons utiliser, en plus de M^+, la quantité $M^- = M_f^-(I) = \max_{x \in [a,b]} F^-(x)$ dans un cas et savoir "calculer" ensembles du type $L = \{x \in [a,b] / F^+(x) \geq k_1, F^-(x) \leq k_2\}$ pour $k_1 \leq k \leq k_2$. Les résultats correspondants sont donnés dans les deux propositions suivantes.

PROPOSITION 2.3

Soit $f \in H(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et F^- le minorant de f construit sur I alors, la constante $M^- = M_f^-(I) = \min_{x \in [a,b]} F^-(x)$ s'exprime :

$$M^- = \begin{cases} \underline{\text{si}}\ 2\alpha_1^2(b-a) \leq W^2 \ \ \underline{\text{alors}}\ \dfrac{f(a)-f(b)}{2} - W\left(\dfrac{b-a}{2}\right)^{1/2} \\ \underline{\text{sinon}} \begin{cases} f(a) - \dfrac{W^2}{4\alpha_1} & \text{si } \alpha_1 > 0 \\ f(b) + \dfrac{W^2}{4\alpha_1} & \text{si } \alpha_1 < 0 \end{cases} \end{cases}$$

≡ ≡ ≡

PROPOSITION 2.4

Soit $f \in H^1(\alpha,\beta)$, $I = [a,b] \in [\alpha,\beta]$ et F^+, F^- le majorant et le minorant de f construits sur I. Soient :

$$D_a^+ = W^2 - 4\alpha_1\{f(a)-k_1\}$$

$$D_a^- = W^2 - 4\alpha_1\{f(a)-k_2\}$$

$$D_b^+ = W^2 + 4\alpha_1\{f(b)-k_1\}$$

$$D_b^- = W^2 + 4\alpha_1\{f(b)-k_2\}$$

$$(X_a^+)^{\pm} = a + \left(\frac{(D_a^+)^{1/2} \pm W}{2\alpha_1}\right)^2 \qquad \text{si } \alpha_1 \neq 0$$

$$(X_b^+)^{\pm} = b - \left(\frac{(D_b^+)^{1/2} \mp W}{2\alpha_1}\right)^2 \qquad \text{si } \alpha_1 \neq 0$$

$$t_a^+ = a + \left(\frac{k_1 - f(a)}{W}\right)^2$$

$$t_b^+ = b - \left(\frac{k_1 - f(b)}{W}\right)^2$$

$$(X_a^-)^{\pm} = a + \left(\frac{(D_a^-)^{1/2} \pm W}{2\alpha_1}\right)^2 \qquad \text{si} \quad \alpha_1 \neq 0$$

$$(X_b^-)^{\pm} = b - \left(\frac{(D_a^-)^{1/2} \mp W}{2\alpha_1}\right)^2 \qquad \text{si} \quad \alpha_1 \neq 0$$

Alors, l'ensemble

$L = \{x \in [a,b] / F^+(x) \geq k_1 \text{ et } F^-(x) \leq k_2\}$ s'écrit :

1°) Si $\alpha_1 > 0$ alors

1.1) si $k_2 < f(a)$ alors si $D_a^- \geq 0$ $\quad L = [(x_a^-)^-, \min\{(x_b^-)^+, (x_a^-)^+\}]^*$

sinon $L = \emptyset$

1.2) si $k_1 \leq f(a) < k_2 < f(b)$ alors $L = [a, \min\{(x_b^-)^+, (x_a^-)^+\}]$

1.3) si $k_1 \leq f(a) < f(b) \leq k_2$ alors $L = [a,b]$

1.4) si $f(a) < k_1 \leq k_2 < f(b)$ alors $L = [\max\{(x_a^+)^-, (x_b^+)^-\}, \min\{(x_a^-)^+, (x_b^-)^+\}]^*$

1.5) si $f(a) < k_1 < f(b) \leq k_2$ alors $L = [\max\{(x_a^+)^-, (x_b^+)^-\}, b]$

1.6) si $f(b) \leq k_1$ alors si $D_b^+ \geq 0$ $\quad L = [\max\{(x_a^+)^-, (x_b^+)^-\}, (x_b^+)^+]^*$

sinon $L = \emptyset$

2°) Si $\alpha_1 < 0$ alors

2.1) si $k_2 \leq f(b)$ alors si $D_b^- \geq 0$ $L = [\max\{(x_a^-)^-, (x_b^-)^-\}, (x_b^-)^+]^*$

sinon $L = \emptyset$

2.2) si $k_1 \leq f(b) < k_2 < f(a)$ alors $L = [\max\{(x_a^-)^-, (x_b^-)^-\}, b]$

2.3) si $k_1 \leq f(b) < f(a) \leq k_2$ alors $L = [a,b]$

2.4) si $f(b) < k_1 \leq k_2 < f(a)$ alors $L = [\max\{(x_a^-)^-, (x_b^-)^-\}, \min\{(x_a^+)^+, (x_b^+)^+\}$

2.5) si $f(b) < k_1 < f(a) \leq k_2$ alors $L = [a, \min\{(x_a^+)^+, (x_b^+)^+\}]$

2.6) si $f(a) \leq k_1$ alors si $D_a^+ \geq 0$, $L = [(x_a^+)^-, \min\{(x_a^+)^+, (x_b^+)^+\}]^*$

sinon $L = \emptyset$

3°) Si $\alpha_1 = 0$ alors

3.1) si $k_1 \leq f(a) \leq k_2$ alors $L = [a,b]$

3.2) si $f(a) < k_1$ alors $L = [t_a^+, t_b^+]^*$

3.3) si $k_2 < f(a)$ alors $L = [t_a^-, t_b^-]^*$

≡ ≡ ≡

Les quantités $(x_a^+)^{\pm}$, $(x_b^+)^{\pm}$, $(x_a^-)^{\pm}$ et $(x_b^-)^{\pm}$ sont les racines des équations :

$$(2.11)\quad \begin{cases} \alpha_1 (x-a)^2+W(x-a)+f(a)-k_1 = 0 \\ \alpha_1 (x-b)^2-W(x-b)+f(b)-k_1 = 0 \\ \alpha_1 (x-a)^2-W(x-a)+f(a)-k_2 = 0 \\ \alpha_1 (x-b)^2+W(x-b)+f(b)-k_2 = 0 \end{cases}$$

et elles sont telles que $(x_a^+)^- \leq (x_a^+)^+$, $(x_b^+)^- \leq (x_b^+)^+$ etc... dans le cas où elles sont réelles. De même t_a^+, t_b^+, t_a^- et t_b^- sont les racines de ces équations dans le cas où le discriminant est nul.

Si $k_1 = k_2 = k$, cas qui arrive si le nombre k est connu explicitement, les formules peuvent se simplifier en remarquant que $D_a^+ = D_a^-$ et que $D_b^+ = D_b^-$ d'où

$$(2.12)\quad \begin{cases} (x_a^+)^- = (x_a^-)^- = x_a^- \\ (x_a^+)^+ = (x_a^-)^+ = x_a^+ \\ (x_b^+)^- = (x_b^-)^- = x_b^- \\ (x_b^+)^+ = (x_b^-)^+ = x_b^+ \\ t_a^+ = t_a^- = t_a \\ t_b^+ = t_b^- = t_b \end{cases}$$

D'autre part, si $k_1 = k_2 = k$, les cas 1.3, 1.5, 2.3, 2.5 de la proposition 2.4 ne peuvent pas avoir lieu tandis que les cas du (3°) se réduisent au cas $f(a) = k$ et $f(a) \neq k$. Avec ces simplifications, nous obtenons le cas particulier important :

PROPOSITION 2.5

Soit $f \in H^1(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et F^+, F^- le majorant et le minorant de f construit sur I. Alors l'ensemble $L = \{x \in [a,b] / F_a^- \leq k \leq F_a^+(x)\}$ s'écrit, en fonction des quantités définies par l'équation 2.12 avec $k_1 = k_2 =$

1°) si $\alpha_1 > 0$ alors

1.1) si $k \leq f(a)$ alors si $D_a \geq 0$ $\quad L = [x_a^-, \min\{x_a^+, x_b^+\}]^*$

sinon $\quad L = \emptyset$

1.2) si $f(a) < k < f(b)$ alors $\quad L = [\max\{x_a^-, x_b^-\}, \min\{x_a^+, x_b^+\}]^*$

1.3) si $f(b) \leq k$ alors si $D_b \geq 0$ $\quad L = [\max\{x_a^-, x_b^-\}, x_b^+]^*$

sinon $\quad L = \emptyset$

2°) si $\alpha_1 < 0$ alors

2.1) si $k \leq f(b)$ alors si $D_b \geq 0$ $\quad L = [\max\{x_a^-, x_b^-\}, x_b^+]^*$

sinon $\quad L = \emptyset$

2.2) si $f(b) < k < f(a)$ alors $\quad L = [\max\{x_a^-, x_b^-\}, \min\{x_a^+, x_b^+\}]^*$

2.3) si $f(a) \leq k$ alors si $D_a \geq 0$ $\quad L = [x_a^-, \{\min x_a^+, x_b^+\}]^*$

sinon $\quad L = \emptyset$

3°) si $\alpha_1 = 0$ alors

3.1) si $f(a) = k$ alors $\quad L = [a,b]$

3.2) si $f(a) \neq k$ alors $\quad L = [t_a, t_b]^*$

III - METHODES DE LOCALISATION DANS H^2

D'après les propositions 1.2 et 1.3, le majorant pour des fonctions de H^1 converge vers la fonction ponctuellement, mais cette convergence est de l'ordre de $(b-a)^{1/2}$. D'autre part, le majorant construit a une dérivée non bornée au voisinage des points d'évaluations de la fonction. Nous allons étudier la majoration correspondante à H^2 qui bien que valable dans une classe plus restreinte de fonctions, a l'avantage de converger avec un ordre $(b-a)^{3/2}$ et d'avoir une dérivée bornée.

Soit $f \in H^2(\alpha,\beta)$ et soit $I = [a,b] \subseteq [\alpha,\beta]$. Soit F^+ le majorant de f construit sur I suivant les formules du chapitre I, paragraphe 4.
Les expressions de F_a^+ et F_b^+ s'écrivent (équations 1.7 et 1.9) comme :

$$F_a^+(x) = \alpha_2(x-a)^2+f'(a)(x-a)+f(a)+W(x-a)^{3/2}$$

$$F_b^+(x) = \alpha_2(x-b)^2+f'(b)(x-b)+f(b)+W(b-x)^{3/2}$$

avec

$$\alpha_2 = \frac{f'(b)-f'(a)}{2(b-a)}$$

et

$$W^2 = \frac{1}{3}\{\|f\|^2-\alpha_2^2(b-a)\}$$

Nous avons vu dans la section précédente, que pour pouvoir expliciter l'ensemble $L^+ = \{x \in [a,b]/F^+(x) \geq k\}$ il faut calculer les racines des équations

$$F_a^+(x) = k = F_b^+(x)$$

Dans le cas présent, cela consiste à résoudre les équations polynomiales de degré 4. Pour rendre les calculs plus aisés nous allons définir des nouveaux majorants, un peu plus "lâches", en utilisant les relations :

$$
\begin{aligned}
&W(x-a)^{3/2} \leq W(b-a)^{1/2}(x-a) \qquad \forall x \in [a,b] \\
&W(b-x)^{3/2} \leq W(b-a)^{1/2}(b-x) \qquad \forall x \in [a,b]
\end{aligned}
\tag{2.13}
$$

Nous allons encore noter par W la nouvelle constante $W(b-a)^{1/2}$ et de même nous noterons encore F_a^+ et F_b^+ les nouveaux majorants.Avec ces conventions nous obtenons :

$$
\boxed{
\begin{aligned}
&F_a^+(x) = \alpha_2(x-a)^2+\{f'(a)+W\}(x-a)+f(a) \\
&F_b^+(x) = \alpha_2(b-x)^2+\{W-f'(b)\}(b-x)+f(b) \\
&\text{avec} \\
&\alpha_2 = \frac{f'(b)-f'(a)}{2(b-a)} \\
&\text{et} \\
&W^2 = \frac{b-a}{3}\{|f|^2-\alpha_2^2(b-a)\}
\end{aligned}
}
\tag{2.14}
$$

Ces formules remplaçent celles données par les équations 1.7 et 1.8 dans le cas q = 2. Comme nous avons seulement fait une majoration de W, tout en préservant son comportement par rapport à (b-a), toutes les propositions que nous avons démontrées en utilisant 1.7 et 1.8 restent valables.

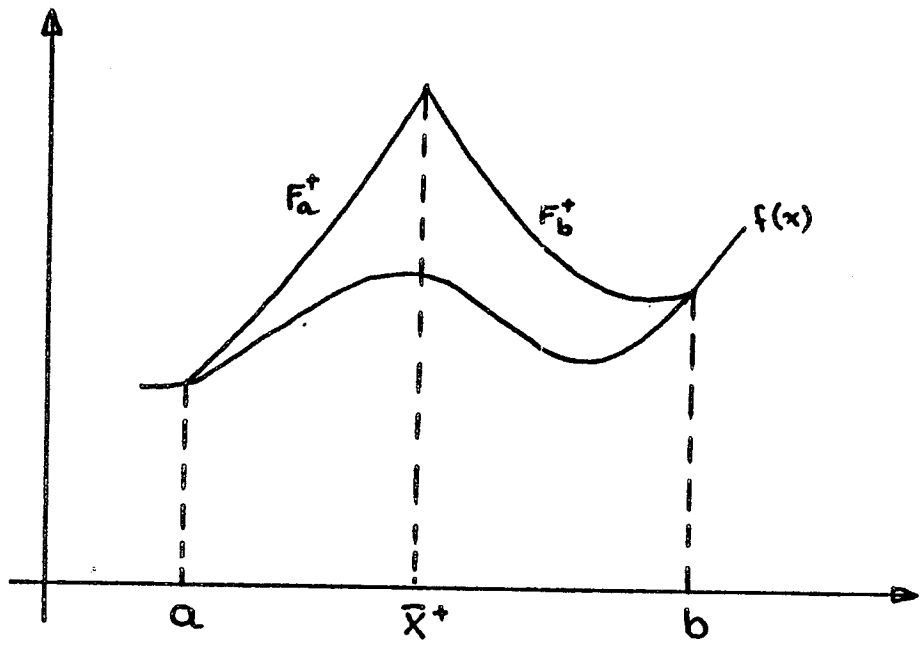
Le nouveau majorant $F^+(x) = \min(F_a^+(x), F_b^+(x))$, diffère de celui obtenu en utilisant 1.7 et 1.8. Il n'interpole plus la dérivée f' aux points a et b. La majoration 2.13 introduit une perturbation car $(F^+)'(a) = f'(a)+W$ et $(F^+)'(b) = f'(b)-W$. Néanmoins, quand (b-a) converge vers zéro, la constante

W converge vers zéro (proposition 1.3) et $(F^+)'$ interpole "asymptotiquement" f' en a et b.

Les fonctions F_a^+ et F_b^+ sont toutes les deux soient concaves, soient convexes. L'équation $F_a^+(x) = F_b^+(x)$ a une seule racine réelle $\bar{x}^+$ qui se trouve dans [a,b]

$$\bar{x}^+ = \frac{a+b}{2} + \frac{b-a}{W}\left\{\frac{f(b)-f(a)}{b-a} - \frac{f'(a)+f'(b)}{2}\right\}$$

Ci-dessous, nous montrons schématiquement un des cas possibles quand $\alpha_2 > 0$



Le calcul de la constante M^+ ainsi que la "détermination" de l'ensemble L^+ ne sont plus aisés et nous ne donnerons que quelques idées sur la façon dont les calculs ont été menés pour donner ensuite, comme dans le cas q = 1, les résultats sans démonstration.

Pour le calcul de la constante M^+ et de l'ensemble L^+ il faut considérer comme dans q = 1, les trois cas $\alpha_2 > 0$, $\alpha_2 < 0$ et $\alpha_2 = 0$. Pour le calcul de M^+ nous considérons les signes de :

$$(F_a^+)'(a) = f'(a)+W$$

$$(F_a^+)'(b) = f'(b)+W$$

$$(F_b^+)'(a) = f'(a)-W$$

$$(F_b^+)'(b) = f'(b)-W$$

La comparaison des quatre quantités f'(a), f'(b), W et -W nous conduit à six cas pour $\alpha_2 > 0$, six cas pour $\alpha_2 < 0$ et trois cas pour $\alpha_2 = 0$ (car $\alpha_2 = 0$ entraîne que f'(a) = f'(b)). Au total il faut donc analyser 15 situations possibles.

Dans le cas $\alpha_2 > 0$, les majorants F_a^+ et F_b^+ sont convexes, ce qui entraîne une convexité par morceaux de F^+. Le maximum M^+ est donc atteint en un des trois points a,b ou $\bar{x}$.

Dans le cas $\alpha_2 < 0$, les majorants sont concaves et il faut tenir compte de la position relative des maxima de F_a^+, F_b^+ et du point $\bar{x}$. Nous donnons à continuation les résultats relatifs à M^+ et à M^-.

PROPOSITION 2.6

Soit $f \in H^2(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et F^+ le majorant de f construit sur I. Soit :

$$\bar{x}^{+} = \frac{a+b}{2} + \frac{b-a}{2W}\left\{\frac{f(b)-f(a)}{b-a} - \frac{f'(a)+f'(b)}{2}\right\}$$

$$x_a^{+} = \max\left\{a, a - \frac{f'(a)+W}{2\alpha_2}\right\} \qquad \text{si} \quad \alpha_2 \neq 0$$

$$x_b^{+} = \min\left\{b, b + \frac{W-f'(b)}{2\alpha_2}\right\} \qquad \text{si} \quad \alpha_2 \neq 0$$

Alors, $M^{+} = \max_{x \in I} F^{+}(x)$ s'exprime

$$M^{+} = \begin{cases} F_b^{+}(\min\{\bar{x}^{+}, x_b^{+}\}) & \text{si } \bar{x}^{+} \leq x_a^{+} \\ F_a^{+}(x_a^{+}) & \text{si } x_a^{+} \leq \bar{x}^{+} \end{cases}$$

≡ ≡ ≡

PROPOSITION 2.7

Soit $f \in H^2(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et F^{-} le minorant de f construit sur I. Soient :

$$\bar{x}^{-} = \frac{a+b}{2} - \frac{b-a}{2}\left\{\frac{f(b)-f(a)}{b-a} - \frac{f'(a)+f'(b)}{2}\right\}$$

$$x_a^{-} = \max\left\{a, a + \frac{W-f'(a)}{2\alpha_2}\right\} \qquad \text{si} \quad \alpha_2 \neq 0$$

$$x_b^{-} = \min\left\{b, b - \frac{W+f'(b)}{2\alpha_2}\right\} \qquad \text{si} \quad \alpha_2 \neq 0$$

Alors $M^{-} = \min_{x \in I} F^{-}(x)$ s'écrit

$$M^- = \begin{cases} F_b^-(x_b^-) & \text{si } \bar{x}^- \leq x_b^- \\ F_a^-(\min\{\bar{x}^-, x_a^-\}) & \text{si } x_b^- \leq \bar{x}^- \end{cases}$$

≡ ≡ ≡

Pour le "calcul" de l'ensemble $L^+ = \{x \in [a,b]/F^+(x) \geq k\}$, nous devons résoudre les équations :

$$(2.15) \quad \begin{cases} F_a^+(x) = k \\ F_b^+(x) = k \end{cases}$$

avec F_a^+ et F_b^+ donnés par les formules 2.14.

Si $\alpha_2 < 0$, F_a^+ et F_b^+ sont concaves et on démontre que L^+ est réduit au seul intervalle $L^+ = [x_1, x_2]^*$ avec x_1 et x_2 exprimés en fonction des racines des équations 2.15.

Pour étudier le cas $\alpha_2 > 0$, comme F_a^+ et F_b^+ sont maintenant convexes, on détermine séparément les ensembles $L_a^+ = \{x \in [a,b]/F_a^+(x) \leq k\}$ et $L_b^+ = \{x \in [a,b]/F_b^+(x) \geq k\}$ et on calcule L^+ comme $L_a^+ \cap L_b^+$. Le résultat est donné comme quatre intervalles dont au moins un est toujours vide.

Le cas $\alpha_2 = 0$ est le plus simple car F_a^+ et F_b^+ sont linéaires.

Pour le calcul de L^+ dans H^1 nous avons utilisé le fait que l'ordre des racines dans les équations 2.11 (équations qui servent à calculer $(x_a^-)^+$ et $(x_a^+)^+$) était indépendant du signe de α_1. Malheureusement, dans le cas présent, cet ordre dépend du signe de α_2 de sorte que les formules pour le calcul de $(x_a^+)^\pm$ et $(x_b^+)^\pm$ doivent être élaborées dans les deux cas $\alpha_2 < 0$ et $\alpha_2 > 0$. Nous donnons par le suite les propositions relatives au calcul de L^+ et de L.

PROPOSITION 2.8.

Soit $f \in H^2(\alpha,\beta)$, $I = [a,b] \subseteq [\alpha,\beta]$ et soient F_a^+ et F_b^+ les majorants de f donnés par l'équation 2.14. Soient $(x_a^+)^+$ et $(x_a^+)^-$ les racines de l'équation :

$$F_a^+(x) = k$$

avec $(x_a^+)^- \leq (x_a^+)^+$ quand elles sont réelles. Soient $(x_b^+)^-$ et $(x_b^+)^+$ les racines de l'équation

$$F_b^+(x) = k$$

avec $(x_b^+)^- \leq (x_b^+)^+$ quand elles sont réelles.

Soient $D_a^+ = (f'(a)+W)^2 - 4\alpha_2(f(a)-k)$

$$D_b^+ = (W-f'(b))^2 - 4\alpha_2(f(b)-k)$$

alors l'ensemble $L^+ = \{x \in [a,b] / F^+(x) = \min(F_a^+(x), F_b^+(x)) \geq k\}$ est défini par :

1°) <u>si</u> $\alpha_2 < 0$ <u>alors</u>

1.1) <u>Si</u> $D_a^+ < 0$ ou $D_b^+ < 0$ <u>alors</u> $L^+ = \emptyset$

1.2) <u>Si</u> $D_a^+, D_b^+ \geq 0$ <u>alors</u> $L^+ = [\max\{a,(x_a^+)^-,(x_b^+)^-\}, \min\{b,(x_a^+)^+,(x_b^+)^+\}]^*$

2°) <u>si</u> $\alpha_2 > 0$ <u>alors</u>

2.1) <u>Si</u> $D_a^+ < 0$ et $D_b^+ \geq 0$ <u>alors</u>

$$L^+ = [a,\{\min b,(x_b^+)^-\}]^* \cup [\max\{a,(x_b^+)^+\},b]^*$$

2.2) Si $D_a^+ \geq 0$ et $D_b^+ < 0$ alors

$$L^+ = [a, \{\min b, (x_a^+)^-\}]^* \cup [\max\{a, (x_a^+)^+\}, b]^*$$

2.3) Si $D_a^+, D_b^+ \geq 0$ alors

$$L^+ = [a, \min\{b, (x_a^+)^-, (x_b^+)^-\}]^* \cup [\max\{a, (x_b^+)^+\}, \min\{b, (x_a^+)^-\}]^*$$

$$\cup [\max\{a, (x_a^+)^+\}, \min\{b, (x_b^+)^-\}]^* \cup [\max\{a, (x_a^+)^+, (x_b^+)^+\}, b]^*$$

2.4) Si $D_a^+, D_b^+ < 0$ alors $L^+ = \emptyset$

3°) Si $\alpha_2 = 0$ alors, avec

$$t_a^+ = a - \frac{f(a)-k}{f'(a)+W} \qquad \text{si } f'(a)+W \neq 0$$

$$t_b^+ = b + \frac{f(b)-k}{W-f'(b)} \qquad \text{si } W-f'(b) \neq 0$$

On a :

3.1) si $f'(a) > W$ alors $L^+ = [\max\{a, t_a^+, t_b^+\}, b]^*$

3.2) si $f'(a) = W$ alors si $k \leq f(b)$ alors $L^+ = [\max\{a, t_a^+\}, b]^*$

sinon $L^+ = \emptyset$

3.3) si $-W < f'(a) < W$ alors $L^+ = [\max\{a, t_a^+\}, \min\{b, t_b^+\}]^*$

3.4) si $f'(a) = -W$ alors si $k \leq f(a)$ alors $L^+ = [a, \min\{b, t_b^+\}]^*$

sinon $L^+ = \emptyset$

3.5) si $f'(a) < -W$ alors $L^+ = [a, \min\{b, t_a^+, t_b^+\}]^*$

≡ ≡ ≡

Pour le calcul de L nous aurons besoin de F_a^- et F_b^- lesquelles sont définies comme F_a^+ et F_b^+ mais en changeant W par -W dans les formules 2.14

PROPOSITION 2.9

Soit $f \in H^2(a,b)$, $I = [a,b] \subseteq [\alpha,\beta]$, F_a^+ et F_b^+ les majorants de f et F_a^- et F_b^- les minorants, construits sur [a,b]. Soient :

$(x_a^+)^-$, $(x_a^+)^+$ les racines de $F_a^+(x) = k_1$, $(x_a^+)^- \leq (x_a^+)^+$ si $(x_a^+)^-$, $(x_a^+)^+ \in \mathbb{R}$

$(x_b^+)^-$, $(x_b^+)^+$ les racines de $F_b^+(x) = k_1$, $(x_b^+)^- \leq (x_b^+)^+$ si $(x_b^+)^-$, $(x_b^+)^+ \in \mathbb{R}$

$(x_a^-)^-$, $(x_a^-)^+$ les racines de $F_a^-(x) = k_2$, $(x_a^-)^- \leq (x_a^-)^+$ si $(x_a^-)^-$, $(x_a^-)^+ \in \mathbb{R}$

$(x_b^-)^-$, $(x_b^-)^+$ les racines de $F_b^-(x) = k_2$, $(x_b^-)^- \leq (x_b^-)^+$ si $(x_b^-)^-$, $(x_b^-)^+ \in \mathbb{R}$

Soient : $D_a^+ = (W+f'(a))^2 - 4\alpha_2(f(a)-k_1)$

$D_b^+ = (W-f'(b))^2 - 4\alpha_2(f(b)-k_1)$

$D_a^- = (W-f'(a))^2 - 4\alpha_2(f(a)-k_2)$

$D_b^- = (W+f'(b))^2 - 4\alpha_2(f(b)-k_2)$

alors l'ensemble $L = \{x \in [a,b] / F^+(x) \geq k_1 \text{ et } F^-(x) \leq k_2\}$ s'exprime :

1°) Si $\alpha_2 < 0$ soient

$$x_1 = \max \{a, (x_a^+)^-, (x_b^+)^-\}$$

$$x_2 = \min \{b, (x_a^+)^+, (x_b^+)^+\}$$

$$p_1 = \max \{(x_b^-)^+, x_1\}$$

$$p_2 = \min \{(x_a^-)^-, x_2\}$$

$$p_3 = \max \{(x_a^-)^+, x_1\}$$

$$p_4 = \min \{(x_b^-)^-, x_3\}$$

alors si

1.1) $D_a^+ < 0$ ou $D_b^+ < 0$ alors $L = \emptyset$

1.2) si $D_a^+, D_b^+ \geq 0$ alors

1.2.1) si $D_a^- < 0, D_b^- \geq 0$ alors $L = [x_1, p_4]^* \cup [p_1, x_2]^*$

1.2.2) si $D_a^- \geq 0, D_b^- < 0$ alors $L = [x_1, p_2]^* \cup [p_3, x_2]^*$

1.2.3) si $D_a^-, D_b^- \geq 0$ alors

$$L = [x_1, \min\{p_2, p_4\}]^* \cup [p_1, p_2]^* \cup [p_3, p_4]^* [\max\{p_1, p_3\},$$

1.2.4) si $D_a^-, D_b^- < 0$ alors $L = [x_1, x_2]$

2°) si $\alpha_2 < 0$ soient

$$x_1 = \max\{a, (x_a^-)^-, (x_b^-)^-\}$$

$$x_2 = \min\{b, (x_a^-)^+, (x_b^-)^+\}$$

$$p_1 = \max\{(x_b^+)^+, x_1\}$$

$$p_2 = \min\{(x_a^+)^-, x_2\}$$

$$p_3 = \max\{(x_a^+)^+, x_2\}$$

$$p_4 = \min\{(x_b^+)^-, x_2\}$$

alors si

2.1) $D_a^- < 0$ ou $D_b^- < 0$ alors $L = \emptyset$

2.2) $D_a^-, D_b^- \geq 0$ alors

2.21) si $D_a^+ < 0$ et $D_b^+ \geq 0$ alors $L = [x_1, p_4]^* \cup [p_1, x_2]^*$

2.2.2) si $D_a^+ \geq 0$ et $D_b^+ < 0$ alors $L = [x_1, p_2]^* \cup [p_3, x_2]^*$

2.2.3) si $D_a^+, D_b^+ \geq 0$ alors

$$L = [x_1, \min\{p_2, p_4\}]^* \cup [p_1, p_2]^* \cup [p_3, p_4]^* \cup [\max\{p_1, p_3\}, x_2]^*$$

2.2.4) si $D_a^+, D_b^+ < 0$ alors $L = [x_1, x_2]$

3°) <u>si $\alpha_2 = 0$ soient</u>

$$t_a^+ = a - \frac{f(a)-k_1}{f'(a)+W} \qquad \text{si } f'(a)+W \neq 0$$

$$t_b^+ = b + \frac{f(b)-k_1}{W-f'(b)} \qquad \text{si } W-f'(b) \neq 0$$

$$t_a^- = a - \frac{f(a)-k_2}{f'(a)-W} \qquad \text{si } f'(a)-W \neq 0$$

$$t_b^- = b - \frac{f(a)-k_2}{W+f'(b)} \qquad \text{si } W+f'(b) \neq 0$$

alors si

3.1) $f'(a) > W$ alors $L = [\max\{a, t_a^+, t_b^+\}, \min\{b, t_a^-, t_b^-\}]^*$

3.2) $f'(a) = W$ alors si $f(a) > k_2$ ou $f(b) < k_1$ alors $L = \emptyset$

sinon $L = [\max\{a, t_a^+\}, \min\{b, t_b^-\}]^*$

3.3) $-W < f'(a) < W$ alors $L = [\max\{a, t_a^+, t_a^-\}, \min\{b, t_b^+, t_b^-\}]^*$

3.4) $f'(a) = -W$ alors si $f(a) < k_1$ ou $f(b) > k_2$ alors $L = \emptyset$

sinon $L = [\max\{a, t_a^-\}, \min\{b, t_b^+\}]^*$

3.5) $f'(a) < -W$ alors $L = [\max\{a, t_a^-, t_b^-\}, \min\{b, t_a^+, t_b^+\}]^*$

≡ ≡ ≡

Remarque

Dans la proposition précédente les calculs du max et du min sont définis formellement. Si l'un de ses arguments est un nombre complexe, alors l'intervalle correspondant doit être considéré comme étant vide.

Toutes les formules données dans cette section ainsi que dans la précédente, ont été programmées et utilisées pour la mise en oeuvre des algorithmes 1 et 2 du chapitre I (paragraphe 5.1 et 5.2 respectivement). Des résultats numériques seront présentés dans le prochain chapitre.

IV - CALCUL DE LA CONSTANTE W

Le problème du calcul de M^+, M^- et de la "détermination" de L^+ et L étant résolus, il reste comme dernier obstacle à la mise en oeuvre des algorithmes de localisation développés au deuxième chapitre, le calcul de la constante W.

Pour q = 1 la constante W est donnée par l'équation 1.7 tandis que pour q = 2 il faut considérer la formule 2.14. Les deux cas peuvent s'exprimer comme :

$$(2.16) \qquad W^2 = \left(\frac{b-a}{3}\right)^{q-1} \{ \|f\|^2 - \alpha_q^2 (b-a) \} \qquad q = 1,2$$

Le calcul de W^2 est donc lié au problème du calcul des quantités

$$(2.17) \qquad I_f(a,b) = \|f\|^2 = \int_a^b [f^{(q)}(s)]^2 ds$$

et

$$(2.18) \qquad S_f(a,b) = \alpha_q^2 (b-a) = \frac{[f^{(q-1)}(b) - f^{(q-1)}(a)]^2}{b-a}$$

Les algorithmes que nous avons développés reposent essentiellement sur la détermination d'un majorant de f sur [a,b]. Nous avons pu remarquer sur quelques exemples que le calcul de W^2 présente une instabilité numérique. La quantité W^2 étant fortement sensible aux erreurs d'arrondis, sa valeur en machine peut être sensiblement inférieure au valeur théorique et devenir même négative quand (b-a) devient petit (du fait de la différence entre les valeurs machine de 2.17 et 2.18, valeurs elles-mêmes très sensibles aux erreurs de calcul.).

Cette instabilité dont nous montrerons des exemples dans le chapitre III, produit, soit une erreur d'exécution du programme ($W^2 < 0$) soit, ce qui est plus grave une falsification des résultats.

Dans cette section, nous étudierons le problème du calcul de $I_f(a,b)$ d'une part et nous analyserons le problème des erreurs générées et propagées sur l'ordinateur d'autre part. Cette analyse des erreurs nous permettra d'introduire une correction de la valeur calculée de W^2 de façon à garantir une majoration effective de f.

IV.1 - CALCUL DE $I_f(a,b)$

Si l'on connait une primitive

$$I_f(x) = \int_0^x [f^{(q)}(s)]^2 ds, \tag{2.19}$$

on pourrait obtenir, en théorie, la quantité $I_f(a,b)$ comme

$$I_f(a,b) = I_f(b) - I_f(a) \tag{2.20}$$

Malheureusement, quand b-a devient petit, la valeur machine de $I_f(a,b)$ calculée par 2.20, l'est avec une erreur relative de plus en plus grande.

La question est donc de savoir quand on peut dire que (b-a) est petit et rend alors inutilisable 2.20. Pour pouvoir répondre à cette question, il nous faut un moyen "numériquement sûr" pour calculer $I_f(a,b)$. Nous allons utiliser ici la théorie d'intervalles de MOORE [23] pour étudier les instabilités numériques de la formule 2.20.

La théorie des intervalles de Moore fournit un moyen d'encadrer l'intégr définie d'une fonction rationnelle sur un intervalle donné. Comme cet encadrement peut être rendu aussi fin que l'on désire, nous pourrons déterminer la valeur de (b-a) pour laquelle la valeur-machine de $I_f(a,b)$ sort de cet encadrement. Par ailleurs, la borne supérieure de l'encadrement nous fournit une majoration de $I_f(a,b)$.

IV.1.1 - Encadrement de $I_f(a,b)$

R.E. MOORE [23] a développé une arithmétique d'intervalles qui permet d'expliciter des intervalles contenant les valeurs d'une fonction rationnelle quand la variable indépendante parcourt un domaine donné.

L'arithmétique d'intervalles est définie pour les opérations arithmétiqu usuelles par :

$$[a,b] * [c,d] = \{x * y \;/\; x \in [a,b] \text{ et } y \in [c,d]\} \quad * \in \{+, -, ., \div\}$$

Le résultat d'une opération entre deux intervalles est toujours un nouvel intervalle, lequel peut être défini explicitement en fonction des extrêmités a, b, c et d des opérandes par des opérations arithmétiques réelles. Cette arithmétique est consistante avec l'arithmétique réelle si l'on considère des intervalles de la forme [x,x]. Nous renvoyons le lecteur au livre de Moore pour une étude approfondie du sujet.

Nous nous contenterons de rappeler ici qu'étant donné une fonction rationnelle définie sur [a,b], une fonction F définie sur la famille $\mathbb{I}_{[a,b]}$ de sous-intervalles fermés de [a,b] et à valeur dans $\mathbb{I}_{\mathbb{R}}$ est une extension par intervalles de f si elle satisfait les conditions :

a) $\quad F([x,x]) = f(x) \qquad \forall x \in [a,b]$

b) $\quad \{f(x),\ x \in I\} \subseteq F(I) \qquad \forall I \in \mathbb{I}_{[a,b]}$

A partir d'une extension par intervalles d'une fonction rationnelle f on peut construire des intervalles contenant les valeurs de l'intégrale définie de f sur [a,b]. Cette proposition découle de la relation évidente

$$\int_a^b f(s)ds \in F([a,b]).[b-a,\ b-a] \tag{2.21}$$

qui n'est qu'une conséquence du théorème de la moyenne.

De façon générale si $[a,b] = \bigcup_{i=1}^{n} [a_i,b_i]$ avec $a_1 = a$, $a_{i+1} = b_i$ $i = 1,\ldots,n-1$ $b_n = b$, on obtient la relation

$$\int_a^b f(s)ds \in \sum_{i=1}^{n} F([a_i,b_i]).[b_i-a_i,\ b_i-a_i] \tag{2.22}$$

et on peut démontrer (MOORE [23]), qu'il existe un nombre réel positif K, indépendant de n, tel que :

$$\text{mes}\left(\sum_{i=1}^{n} F([a_i,b_i]).[b_i-a_i,\ b_i-a_i]\right) \leq K\,\frac{(b-a)^2}{n}$$

ce qui montre que la formule 2.22 est une méthode d'encadrement de premier ordre. Si (b-a) n'est pas petit, cette méthode a besoin d'un très grand nombre de sous-intervalles, donc d'évaluations de f, pour pouvoir borner raisonnablement l'intégrale.

Si $f \in C^k[a,b]$ est une fonction rationnelle et si $F^{(j)}$ est une extensic par intervalles de $f^{(j)}$ $(j = 0,\ldots,k)$, alors on peut définir une méthode de quadrature par intervalles par la formule :

$$(2.23)\qquad I_{n,k} = \sum_{i=1}^{n}\sum_{j=0}^{k-1} \frac{F^{(j)}([a_{i-1},a_{i-1}])}{(j+1)!}\,(b_i-a_i)^{j+1} +$$

$$+\ \frac{1}{(k+1)!}\sum_{i=1}^{n} F^{(k)}([a_i,b_i])\ (b_i-a_i)^{k+1}$$

MOORE, STROTHER et YAND [24] ont démontré que quelque soit $n,k \geq 1$ on a

$$\int_a^b f(s)\,ds \in I_{n,k}$$

et que pour chaque $k \geq 1$ il existe une constante C_k, indépendante de n, tel que

$$\text{mes}(I_{n,k}) \leq C_k \max_{i=1,..,n} (b_i-a_i)^{k+1}$$

La méthode 2.23 est particulièrement intéressante car MOORE [23] donne une façon de déterminer $F^{(j+1)}$ une fois connues $F^{(o)}$, $F^{(j)}$ dont le coût est linéaire avec j.

La formule 2.23 fournit une méthode pour encadrer une intégrale définie d'une fonction rationnelle qui peut servir à borner l'intégrale $I_f(a,b)$.

Cela pourraît être utilisé pour majorer la constante W^2 en utilisant seulement des valeurs ponctuelles de f et ses dérivées.

L'application de ce type de méthodes est restreinte aux fonctions rationnelles. Cette hypothèse n'est qu'une limitation apparente car sur l'ordinateur on n'utilise que des approximations rationnelles.

Il existe néanmoins une limitation pratique importante qui est la nécessité d'un langage orienté vers des opérations d'intervalles. Un tel langage a été développé, par exemple, par REITER [27] a l'Université de Wisconsin (USA). Nous nous proposons d'étudier dans le futur l'utilisation de ces méthodes de quadrature comme une réponse globale à la question du calcul de la constante W^2, et par conséquent, au problème de la validité de la méthode de localisation que nous proposons.

Nous nous limiterons ici à l'utilisation de cette technique pour mettre en évidence l'instabilité du calcul de $I_f(a,b)$ (équation 2.20) sur un exemple concret. Pour se placer dans le cadre réel d'utilisation de la quantité $I_f(a,b)$, dans les algorithmes 1 et 2, nous allons simuler une suite T_n, en prenant une suite emboîtée d'intervalles $[a_n,b_n]$.

Soit f le polynôme

$$f(x) = 1009\ x^5-278.7\ x^4+282.5\ x^3-1280\ x^2+25.0\ x$$

et définissons la suite d'intervalles par les relations :

$$\left.\begin{aligned} H_o &= 1 \\ a_n &= 0.5-H_n/2 \\ b_n &= 0.5+H_n/2 \\ H_{n+1} &= H_n/10 \end{aligned}\right\} \quad n \geq 0$$

La suite d'intervalles est centrée en x = 0.5, et la simulation corresp donc à ce qui se passerait approximativement si l'on utilisait une méthode de localisation dans H^1 pour localiser les racines de f(x) = f(0.5) dans [0,1

Pour chaque intervalle $[a_n,b_n]$ $n \geq 0$ nous avons calculé $I_f(a,b)$ selon la formule 2.20 et nous avons calculé un encadrement par intervalles. Dans le tableau ci-dessous, on montre les résultats obtenus avec la formule de premier ordre 2.21. Dans la première colonne on donne le pas H_n, $n \geq 0$, dans la deuxième et quatrième colonne les bornes inférieures et supérieures obtenues avec la méthode de quadrature par intervalles (équation 2.20) et finalement dans la troisième colonne la valeur de l'intégrale en utilisant la primitive.

APPROCHE DE LA SEMI-NORME PAR INTERVALLES

H	BORNE INF.	DEFINITION	BORNE SUP.
1.0000'+00	-3.0720'+05	2.7249'+01	3.8007'+05
1.0000'-01	-7.1500'+01	1.0498'-01	7.2319'+01
1.0000'-02	-6.0597'-02	1.1150'-02	1.3809'-01
1.0000'-03	6.2215'-04	1.1157'-03	1.7516'-03
1.0000'-04	1.0599'-04	1.1157'-04	1.1729'-04
1.0000'-05	1.1100'-05	1.1157'-05	1.1213'-05
1.0000'-06	1.1151'-06	1.1157'-06	1.1162'-06
1.0000'-07	1.1156'-07	1.1157'-07	1.1157'-07
1.0000'-08	1.1157'-08	1.1157'-08	1.1157'-08
1.0000'-09	1.1157'-09	1.1151'-09	1.1157'-09
1.0000'-10	1.1157'-10	1.1206'-10	1.1157'-10
1.0000'-11	1.1157'-11	1.0839'-11	1.1157'-11
1.0000'-12	1.1157'-12	9.9476'-13	1.1157'-12
1.0000'-13	1.1157'-13	-4.6185'-14	1.1157'-13
1.0000'-14	1.1157'-14	1.1688'-12	1.1157'-14
1.0000'-15	1.1157'-15	-2.6290'-13	1.1157'-15
1.0000'-16	1.1157'-16	6.2883'-13	1.1157'-16

On peut remarquer que pour de grandes valeurs de H l'encadrement est grossier. Il ne devient du même ordre que l'intégrale, que pour des pas inférieurs à 10^{-3}.

Pour des pas inférieurs à 10^{-9} la valeur de $I_f(a,b)$ n'est plus dans l'intervalle obtenu avec l'équation 2.21.

Dans le calcul des bornes on utilise des valeurs de $[f'(x)]^2$ aux extrémités de l'intervalle $[a_n,b_n]$ tandis que le calcul de $I_f(a,b)$ entraîne une différence entre deux nombres assez proches et différents de zéro. Nous en concluons que la valeur de $I_f(a,b)$ n'a plus de signification pour des pas inférieurs à 10^{-9}.

En résumé, nous pouvons conclure à partir de cet exemple que l'on ne peut pas utiliser la primitive de l'intégrale (équation 2.20) pour des pas "trop petits" et d'autre part, qu'il n'est pas convenable d'utiliser la méthode d'intervalles pour des pas "trop grands" (au moins la méthode du premier ordre). Evidemment, les notions de "trop petit" et "trop grand" sont liées au problème particulier.

Les résultats présentés suggèrent, évidemment, l'emploi d'une stratégie mixte, mais laquelle devrait se faire toutefois avec les formules de quadrature par intervalle d'ordre supérieur.

IV.1.2 - Quadrature approchée de $I_f(a,b)$

Si l'on ne dispose pas de la primitive $I_f(x)$, on peut penser à utiliser une méthode de quadrature approchée. Néanmoins, le fait de remplacer $I_f(a,b)$ par une approximation, si bonne soit-elle, peut se traduire en une sous-estimation de la semi-norme $I_f(a,b)$ avec la conséquence d'une perte de fiabilité de la méthode.

Néanmoins, dans la pratique on peut délaisser la condition d'avoir une majoration de W quand la mesure de la localisation et la valeur de $I_f(a,b)$ sont encore grands, à condition de pouvoir assurer des majorations effectives quand l'une ou l'autre deviennent petites. Cela tient au fait que pour de grands intervalles, ou pour de grandes valeurs de $I_f(a,b)$, le majorant F^+ est suffisamment exagéré pour absorber l'erreur commise et ce n'est que lorsque $F^+(x)$ s'approche de $f(x)$ que la sous-estimation de $I_f(a,b)$ peut fausser le résultat.

Nous pouvons utiliser, dans la pratique, n'importe quelle formule de quadrature approchée à condition qu'elle soit convergente et à condition qu'elle ne souffre pas de l'instabilité de l'équation 2.20, ou encore à condition de pouvoir effectuer une correction pour obtenir asymptotiquement des majorations qui ne soient pas trop grossières.

Nous avons employé une méthode de quadrature assez simple qui a l'avantage de ne pas sortir du cadre, que nous nous sommes imposé, de travailler dans H^q.

Etant donné $f \in H^q[a,b]$ et une sous-division de $[a,b]$ $a = p_o < p_1 < \ldots < p_n = b$, nous allons approcher $I_f(a,b)$ par

$$\hat{I}_f(a,b) = \sum_{j=1}^{n} \frac{\{f^{(q-1)}(p_j) - f^{(q-1)}(p_{j-1})\}^2}{p_j - p_{j-1}} \tag{2.24}$$

Cette formule converge vers $|f|^2$ quand $\max(p_{j+1}-p_j)$ converge vers zéro. En considérant que $f^{(q-1)} \in H^1(a,b)$, la convergence découle du fait que $\hat{I}_f$ est la semi-norme de la fonction linéaire par morceaux qui interpole a $f^{(q-1)}$ aux points p_j (DUC-JACQUET [8]).

Cette méthode est particulièrement intéressante dans le cas de l'algorithme 2 puisque quelques unes des quantités qui apparaissent dans la somme pourront être réutilisées dans les itérations suivantes.

IV.2 - CORRECTION DE W^2 EN MACHINE

Nous avons constaté sur un exemple, que le calcul de $I_f(a,b)$ est très instable quand $(b-a)$ devient petit. Cette instabilité est aussi présente dans le calcul de $S_f(a,b)$ (équation 2.18). Nous allons essayer de borner cette erreur de calcul en fonction de la précision de la machine, de sorte à pouvoir introduire une correction dans la valeur de W^2, correction qui nous permette d'obtenir une majoration de la vraie valeur de W^2. Nous étudierons en même temps cette correction pour le cas où $I_f(a,b)$ est approché par $\hat{I}_f(a,b)$.

Suivant la notation employée par WILKINSON [35], on notera fl(y) la valeur machine du nombre réel y en arithmétique à virgule flottante. Si $y \neq 0$, la valeur fl(y) est associée à la valeur exacte y, par une relation du type

$$fl(y) = y(1+\theta_y) \qquad y \neq 0 \tag{2.25}$$

Si l'on note $\mathcal{M}$ l'ensemble des nombres-machine pratiquement tous les systèmes d'arrondi en arithmétique à virgule flottante utilisés à l'heure actuelle, peuvent être approximés par les équations suivantes : (WILKINSON [35]).

$$\left\{ \begin{array}{ll} fl(x \pm y) = x(1+\varepsilon_1) \pm y(1+\varepsilon_2) & \\ fl(xy) = xy(1+\varepsilon) & \forall x,y \in \mathcal{M} \\ fl(x/y) = (x/y)(1+\varepsilon) & \end{array} \right. \tag{2.26}$$

et les quantités ε, ε_1 et ε_2 sont bornées par une quantité E, $|\varepsilon|, |\varepsilon_1|, |\varepsilon_2| \leq E$, qui dépend de la base de représentation des nombres et de l'arithmétique employée. Une étude approfondie du sujet et des différents systèmes arithmétiques employés a été faite par PICHAT [25]

Pour le calcul de W^2, nous devons calculer les quantités $I_f(a,b)$ (ou $\hat{I}_f(a,b)$ dans le cas approché) et la quantité $S_f(a,b)$. Nous allons borner ces quantités en fonction des erreurs relatives des quantités $fl(f^{(q-1)}(a))$, $fl(f^{(q-1)}(b))$, $fl(b-a)$, $fl(I_f(a))$, $fl(I_f(b))$ et de la précision machine E.

IV.2.1 - Correction de $S_f(a,b)$

Pour simplifier les notations, soient

$$(2.27)\quad \begin{cases} x_3 = f^{(q-1)}(b) \\ x_4 = f^{(q-1)}(a) \\ x_5 = b-a \end{cases}$$

alors $S_f(a,b)$ donné par l'équation 2.18, s'écrit

$$(2.28)\qquad S_f(a,b) = \frac{(x_3-x_4)^2}{x_5}$$

Supposons qu'au moment de commencer le calcul de $S_f(a,b)$ on dispose, en machine, de quantités $fl(x_j)$ $(j = 3,4,5)$ qui satisfont les relations :

$$(2.29)\qquad fl(x_j) = x_j(1+\theta_j) \qquad (j = 3,4,5)$$

Les valeurs θ_j sont le résultat d'une série d'erreurs successives. D'abord, si $f^{(q-1)}$ est une fonction non-rationnelle. On commence par l'approcher par une fonction rationnelle $f_R^{(q-1)}$. Cette fonction dépendra d'un ensemble de coefficients a_j $(j = 1,\ldots,N)$. Pour évaluer f^{q-1} dans un point x, on approche x et les coefficients a_j par des nombres machines $fl(x)$ et $fl(a_j)$ et on exécute ensuite un programme P qui donne en sortie la valeur $fl(x_j)$. L'exécution du programme P ajoute les erreurs générées par la machine du fait de l'utilisation d'une arithmétique à virgule flottante.

En suivant les règles de l'arithmétique à virgule flottante données par les équations 2.26, le calcul de $S_f(a,b)$ se fait en machine de la façon suivante (quand x_5, $fl(x_5 \neq 0)$.

$$fl(S_f(a,b)) = fl\left(\frac{(x_3-x_4)^2}{x_5}\right)$$

$$= \frac{fl((x_3-x_4)^2)}{fl(x_5)}\ (1+\varepsilon_1)$$

$$= \frac{fl^2(x_3-x_4)\ (1+\varepsilon_1)\ (1+\varepsilon_2)}{fl(x_5)}$$

$$= \frac{\{fl(x_3)\ (1+\varepsilon_3)-fl(x_4)\ (1+\varepsilon_4)\}^2\ (1+\varepsilon_1)\ (1+\varepsilon_2)}{fl(x_5)}$$

Remplaçant les valeurs de $fl(x_j)$ $(j = 3,4,5)$ données par 2.29 on obtient :

$$fl(S_f(a,b) = \frac{\{x_3(1+\theta_3)\ (1+\varepsilon_3)-x_4(1+\theta_4)\ (1+\varepsilon_4)\}^2\ (1+\varepsilon_1)\ (1+\varepsilon_2)}{x_5(1+\theta_5)}$$

qui peut s'exprimer comme

$$fl(S_f(a,b)) = \frac{x_3^2(1+\eta_1)-2x_3x_4(1+\eta_2)+x_4(1+\eta_3)}{x_5} \tag{2.30}$$

avec

$$\begin{cases} 1+\eta_1 = (1+\theta_3)^2(1+\varepsilon_3)^2(1+\varepsilon_2)\ (1+\varepsilon_1)\ (1+\beta) \\ 1+\eta_2 = (1+\theta_3)\ (1+\theta_4)\ (1+\varepsilon_3)\ (1+\varepsilon_4)\ (1+\varepsilon_2)\ (1+\varepsilon_1)\ (1+\beta) \\ 1+\eta_3 = (1+\theta_4)^2(1+\varepsilon_4)^2(1+\varepsilon_2)\ (1+\varepsilon_1)\ (1+\beta) \\ 1+\beta = \frac{1}{1+\theta_5} \end{cases} \tag{2.31}$$

Avec ces définitions on obtient :

$$(2.32) \qquad fl(S_f(a,b)) = S_f(a,b) + \frac{x_3^2\eta_1 - 2x_3x_4\eta_2 + x_4\eta_3}{x_5}$$

Cette relation exprime donc la relation entre la valeur exacte $S_f(a,b)$ et la valeur $fl(S_f(a,b))$ calculée en machine avec une arithmétique à virgule flottante à partir des valeurs $fl(x_j)$ $(j = 3,4,5)$ quand $fl(x_5)$, $x_5 \neq 0$.

Comme les quantités θ_j $(j = 3,4,5)$ et ε_j $(j = 1,\ldots,4)$ peuvent être positives négatives ou nulles, il en est de même pour les quantités η_j $(j = 1,2,3)$. Donc, il est possible que $fl(S_f(a,b))$ soit supérieur, inférieur ou égal à $S_f(a,b)$. En particulier, quand $x_5 = b-a$ converge vers zéro, la quantité théorique $S_f(a,b)$ converge vers zéro tandis que $fl(S_f(a,b))$ peut rester non bornée.

Nous avons déjà insisté sur la nécessité d'avoir des majorations de W^2, donc d'avoir une minoration de $S_f(a,b)$ (équations 2.16 et 2.18). Pour obtenir un minorant nous allons borner $|fl(S_f(a,b)) - S_f(a,b)|$ dans 2.32

$$|fl(S_f(a,b)) - S_f(a,b)| \leq \frac{x_3^2|\eta_1| + 2|x_3||x_4||\eta_2| + x_4^2|\eta_3|}{|x_5|}$$

avec égalité si les signes des η_j $(j = 3,4,5)$ sont choisis de façon approprié

Soit $E_f = \max\{|\theta_3|, |\theta_4|, |\theta_5|\}$. Comme $|\varepsilon_j| \leq E$ $(j = 1,\ldots,4)$, alors, d'après les équations 2.14, on peut écrire

$$\frac{(1-E_f)^2(1-E)^4}{1+E_f} \leq 1 + \eta_j \leq \frac{(1+E_f)^2(1+E)^4}{1-E_f} \qquad (j = 1,2,3)$$

Nous allons supposer l'erreur E_f supérieure à E, hypothèse raisonnable si l'on considère qu'elle résulte d'une série d'erreurs, en particulier, aux produits par les opérations arithmétiques du programme qui évalue $f^{(q-1)}$. Nou pouvons donc faire la majoration :

$$\frac{(1-E_f)^6}{1+E_f} \leq 1+\eta_j \leq \frac{(1+E_f)^6}{1-E_f} \qquad (j = 1,2,3)$$

d'où

$$1-5E_f+O(E_f^2) \leq 1+\eta_j \leq 1+5E_f+O(E_f^2) \qquad (j = 1,2,3)$$

et finalement

$$|\eta_j| \leq 5E_f+O(E_f^2) \qquad (j = 1,2,3)$$

Avec cette borne nous obtenons

$$|fl(S_f(a,b))-S_f(a,b)| \leq 5 \frac{(|x_3|+|x_4|)^2}{|x_5|} E_f+O(E_f^2) \tag{2.33}$$

Si nous remplaçons x_j par $fl(x_j)$ nous introduisons, d'après les équations 2.25, une erreur qui est encore de deuxième ordre, d'où, si x_5, $fl(x_5) \neq 0$ on obtient :

$$|fl(S_f(a,b))-S_f(a,b)| \leq 5 \frac{(|fl(x_3)|+|fl(x_4)|)^2}{|fl(x_5)|} E_f+O(E_f^2) \tag{2.34}$$

Si les valeurs de $|fl(x_3)|$ et $|fl(x_4)|$ sont grandes, la borne donnée par l'équation 2.34 peut être trop exagérée. Nous allons éviter cette restriction en calculant une majoration alternative qui permette de remplacer la borne 2.34 quand $|fl(x_3)|$ et $|fl(x_4)|$ sont grands.

Il faut remarquer que les quantités ε_j qui apparaissent dans les calculs précédents dépendent de l'ordre dans lequel sont exécutées les opérations. Néanmoins, la borne 2.34 est indépendante de cet ordre.

Définissons les nouvelles variables :

$$(2.35)\qquad \begin{cases} z_3 = x_3-\alpha \\ z_4 = x_4-\alpha \end{cases}$$

alors on a :

$$(2.36)\qquad S_f(a,b) = \frac{(z_4-z_3)^2}{x_5}$$

et la formule 2.34 reste encore valable en remplaçant E_f par $\hat{E}_f$, borne de l'erreur relative de z_3, z_4 et x_5. Or, il faut considérer aussi que $fl(S_f(a,b))$ représente le résultat de l'évaluation numérique de $S_f(a,b)$ sous cette nouvelle forme. L'équation 2.33 s'écrit donc :

$$(2.37)\qquad |fl(S_f(a,b))-S_f(a,b)| \leq 5\,\frac{(|x_3-\alpha|+|x_4-\alpha|)^2}{|x_5|}\,\hat{E}_f+O(\hat{E}_f)$$

Pour déterminer $\hat{E}_f$ nous allons considérer les nouvelles opérations arithmétiques introduites. En supposant que :

$$fl(\alpha) = \alpha(1+r)$$

et

$$fl(z_j) = z_j(1+\gamma_j) \qquad (j = 3,4)$$

on obtient d'après la définition des z_j (équations 2.35) et selon les règles du calcul en virgule flottante :

$$fl(z_j) = x_j(1+\theta_j)(1+\varepsilon_1)-\alpha(1+r)(1+\varepsilon_2) \qquad (j = 3,4)$$

$$fl(z_j) = z_j+x_j(\theta_j+\varepsilon_1+\varepsilon_1\theta_j)-\alpha(r+\varepsilon_2+r\varepsilon_2) \qquad (j = 3,4)$$

d'où

$$\gamma_j = \frac{x_j(\theta_j+\varepsilon_1+\varepsilon_1\theta_j)-\alpha(r+\varepsilon_2+r\varepsilon_2)}{x_j-\alpha} \qquad \text{si } x_j \neq \alpha$$

ou encore

$$\gamma_j = (\varepsilon_j+\varepsilon_1+\varepsilon_1\theta_j) + \frac{\alpha(\theta_j+\varepsilon_1+\varepsilon_1\theta_j-r-\varepsilon_2-r\varepsilon_2)}{x_j-\alpha}$$

Soit $E_z = \max\{|\theta_3|, |\theta_4|, |r|\}$ et $T = \min\{|x_3-\alpha|, |x_4-\alpha|\}$, alors pour $T \neq 0$ on a :

$$|\gamma_j| \leq 2(1+2\frac{|\alpha|}{T})\ E_z + O(E_z^2) = \hat{E}_f$$

Remplaçant cette valeur de $\hat{E}_f$ dans 2.37, on peut démontrer que la borne est minimale pour $\alpha = \frac{x_3+x_4}{2}$ si $x_3 \neq x_4$. Pour cette valeur de α nous pouvons maintenant expliciter E_z

$$fl(\alpha) = \frac{fl(x_3+x_4)}{2}\ (1+\varepsilon_1)$$

$$fl(\alpha) = \frac{fl(x_3)\ (1+\varepsilon_3)+fl(x_4)\ (1+\varepsilon_4)}{2}\ (1+\varepsilon_1)$$

$$fl(\alpha) = \frac{x_3(1+\theta_3)\ (1+\varepsilon_3)+x_4(1+\theta_4)\ (1+\varepsilon_4)}{2}\ (1+\varepsilon_1)$$

d'où en supposant toujours $E_f \geq E$ on obtient :

$$|r| = \left|\frac{fl(\alpha)-\alpha}{\alpha}\right| \leq 3(1+2\frac{\min\{|x_3|, |x_4|\}}{|x_3+x_4|})\ E_f$$

Comme cette quantité est plus grande que E_f, on obtient pour E_z l'expression

$$E_z = \max\{|\theta_3|, |\theta_4|, |r|\} = 3(1+2\frac{\min\{|x_3|, |x_4|\}}{|x_3+x_4|})E_f$$

Pour ce choix de α on a

$$z_3 = \frac{x_3-x_4}{2} = -z_4$$

$$T = |z_3|$$

et $$\hat{E}_f = 6(1+2\frac{|x_3+x_4|}{|x_3-x_4|})(1+2\frac{\min\{|x_3|,|x_4|\}}{|x_3+x_4|})E_f$$

En remplaçant ces valeurs dans l'équation 3.37 on obtient

(2.38)

$$|fl(S_f(a,b))-S_f(a,b)| \leq 30\frac{|x_3-x_4|^2}{|x_5|}(1+2\frac{|x_3+x_4|}{|x_3-x_4|})(1+2\frac{\min\{|x_3|,|x_4|\}}{|x_3+x_4|})E_f+O(E_f^2)$$

Cette nouvelle borne doit être comparée à celle donnée par l'équatio 2.33, pour obtenir une condition permettant de décider laquelle des deux utiliser.

En supposant que $|x_4| \leq |x_3|$, et en effectuant la comparaison on trouve que si

(2.39) $$\frac{x_4}{x_3} > 1-t$$

avec t la racine unique dans [0,1] de l'équation

(2.40) $$19t^3-102t^2+108t-8 = 0, \qquad (t \doteq 0.08...)$$

alors, la borne donnée par la relation 2.38 est meilleure que celle de l'équation 2.33. Evidemment, la situation inverse se produit si $\frac{x_4}{x_3} < 1-t$ et les deux bornes sont égales si $\frac{x_3}{x_4} = 1-t$.

Ce résultat permet de choisir la borne appropriée dans la résolution du problème de localisation des racines de l'équation $f(x) = k$ sur $[a,b]$.

Pour $q = 1$, en considérant un intervalle $[a,b] \subseteq [\alpha,\beta]$ suffisamment petit contenant une solution de l'équation, on a pour le cas $k > 0$ que $x_4 = \min\{f(a), f(b)\}$ et $x_3 = \max\{f(a), f(b)\}$. Alors le quotient $\frac{x_4}{x_3}$ converg vers +1 quand la localisation devient petite.
Nous devons donc utiliser dans ce cas la borne 2.38. Cela sera le cas également lorsque $k < 0$. Si (b-a) est trop grand il se peut néanmoins que $\frac{x_4}{x_3}$ soit plus petit que 1-t, mais comme la correction ne se fait sentir que quand b-a devient petit, l'importance relative de la correction par rapport

à $fl(S_f(a,b))$ sera négligeable aussi longtemps que (b-a) reste grand.

Si k = 0 et la racine contenue dans [a,b] est d'ordre impaire, alors f(a) et f(b) auront des signes différents quand b-a devient petit et par conséquent le quotient x_4/x_3 deviendra négatif. Nous devons donc dans ce cas utiliser la borne 2.33.

Une analyse du même type peut être faite pour le cas q = 2. Néanmoins nous pensons qu'il est préférable de choisir la borne à employer de façon automatique en comparant pour chaque intervalle à l'étude le rapport x_4/x_3 Cela rend le programme plus "flexible" et c'est ainsi que nous avons programmé les algorithmes.

Si $fl(x_3) \neq fl(x_4)$ et si $fl(x_5) \neq 0$ nous pouvons remplacer les valeurs x_3, x_4 et x_5 dans l'équation 2.38 par ses valeurs approchées $fl(x_3)$, $fl(x_4)$ et $fl(x_5)$ en introduisant une erreur qui est encore de deuxième ordre.

Dans le cas où $fl(x_5) = 0$, nous avons évidemment atteint le point limite de l'algorithme.

Si $fl(x_5) \neq 0$ mais $fl(x_3) = fl(x_4)$, alors $fl(S_f(a,b)) = 0$ et pour borner $S_f(a,b)$ on peut partir directement de la relation

$$fl(x_3) = x_3(1+\theta_3) = x_4(1+\theta_4) = fl(x_4)$$

d'où

$$x_3 - x_4 = x_4\theta_4 - x_3\theta_3$$

$$S_f(a,b) = \frac{(x_3-x_4)^2}{x_5} = \frac{(x_4\theta_4-x_3\theta_3)^2}{x_5}$$

$$S_f(a,b) = \frac{\{(x_4-x_3)\theta_4+x_3(\theta_4-\theta_3)\}^2}{x_5}$$

$$S_f(a,b) = \frac{(x_4-x_3)^2\theta_4^2}{x_5} + \frac{2x_3(x_4-x_3)\theta_4(\theta_4-\theta_3)}{x_5} + \frac{x_3^2(\theta_4-\theta_3)^2}{x_5}$$

$$S_f(a,b) \leq \frac{(x_4-x_3)^2}{|x_5|}E_f^2 + 4\frac{|x_3||x_4-x_3|}{|x_5|}E_f^2 + 4\frac{x_3^2}{|x_5|}E_f^2$$

et comme

$$|x_3-x_4| \leq (|x_3|+|x_4|)E_f$$

on a

$$S_f(a,b) \leq 4\frac{x_3^2}{|x_5|}E_f^2+O(E_f^3) \tag{2.41}$$

En remplaçant les valeurs de x_j par leur approximations $fl(x_j)$ $(j = 3,4,5)$ dans les équations 2.38 et 2.41 nous obtenons le résultat :

$$|fl(S_f(a,b))-S_f(a,b)| \leq \begin{cases} S_f^c(a,b)E_f+O(E_f^2) & \text{si } fl(x_3) \neq fl(x_4) \\ S_f^c(a,b)E_f^2+O(E_f^3) & \text{si } fl(x_3) = fl(x_4) \end{cases} \tag{2.42}$$

(2.43)

$$\text{avec } S_f^c(a,b) = \begin{cases} 30\frac{|fl(x_3)-fl(x_4)|}{|fl(x_5)|}(1+2\frac{|fl(x_3)+fl(x_4)|}{|fl(x_3)-fl(x_4)|})(1+2\frac{\min\{|fl(x_3)|,|$$

$$\text{si } fl(x_3) \neq fl(x_4)$$

$$4\frac{fl^2(x_3)}{|fl(x_5)|} \qquad \text{si } fl(x_3) = fl(x_4)$$

Nous retiendrons donc les résultats 2.34 à utiliser quand $x_4/x_3 \leq 1-t$ et le résultat 2.42 et 2.43 à utiliser quand $x_4/x_3 > 1-t$ (en supposant $|x_4| \leq |x_3|$) avec t la racine unique dans [0,1] de l'équation 2.40.

≡ ≡ ≡

Pour pouvoir utiliser les expressions que nous venons de développer, il faut pouvoir donner une estimation de E_f en tenant compte que E_f est une borne de l'erreur relative des quantités $fl(x_3)$, $fl(x_4)$ et $fl(x_5)$. Autrement dit, E_f est une borne de l'erreur relative des quantités $fl(f^{(q-1)}(a))$, $fl(f^{(q-1)}(b))$ et $fl(b-a)$ par rapport aux valeurs exactes $f^{(q-1)}(a)$, $f^{(q-1)}(b)$ et b-a respectivement.

La précision relative de $fl(f^{(q-1)}(a))$ et de $fl(f^{(q-1)}(b))$ est liée, comme nous l'avons dit, à la précision de l'approximation rationnelle $f_R^{(q-1)}$ de $f^{(q-1)}$ et au programme qui l'évalue. En général si $f^{(q-1)}(a)$ et $f^{(q-1)}(b)$ sont différents de zéro, on peut espérer que l'évaluation en machine de ces quantités aura une erreur relative raisonnablement petite.

Si $f^{(q-1)}(a) = 0$, mais $f^{(q-1)}(b) \neq 0$, on peut démontrer que les majorations trouvées restent encore valables mais avec E_f définie cette fois en remplaçant l'erreur relative de $fl(f^{(q-1)}(a))$ par la valeur absolue de $fl(f^{(q-1)}(a))$.

Si $f^{(q-1)}(a) = f^{(q-1)}(b) = 0$, alors la valeur théorique $S_f(a,b)$ est zéro et n'importe quelle valeur de E_f donnera une borne.

Quand x_5 converge vers zéro, la valeur $S_f(a,b)$ converge elle aussi vers zéro et une sous-estimation de E_f est compensée par la quantité $1/|fl(x_5)|$ qui elle devient de plus en plus grande.

L'analyse que nous venons de faire montre qu'une estimation de E_f raisonnable pour le cas "normal" ou $f^{(q-1)}(x) \neq 0$ sert aussi dans les autres cas. Nos expériences numériques montrent que des valeurs de l'ordre de $10^{-(p-2)}$ ou $10^{-(p-4)}$ pour une machine à p-digits servent largement pour des fonctions qui ne présentent pas d'instabilités numériques dans leur évaluation en machine. Néanmoins, le choix de E_f reste toujours un problème ouvert qui doit être résolu dans chaque cas particulier en tenant compte des propriétés de la fonction f.

≡ ≡ ≡

Pour finir cette section, nous allons montrer sur un exemple le comportement des bornes que nous avons calculées. En prenant encore la simulation qui nous a servi pour l'étude de la stabilité dans le calcul de $I_f(a,b)$ (paragraphe IV.1.1 p. 68), nous obtenons les résultats suivants :

COMPARAISON DES BORNES DE S(A,B)

H	FL(S(A,B))	C1	C2
1.0000'+00	2.8900'+C0	2.89CC'+C0	2.8900'+00
1.0000'-01	1.0005'-C1	1.C0C5'-C1	1.0005'-01
1.0000'-02	1.1145'-02	1.1145'-02	1.1145'-02
1.0000'-03	1.1157'-C3	1.1157'-03	1.1157'-03
1.0000'-04	1.1157'-C4	1.1157'-C4	1.1157'-04
1.0000'-05	1.1157'-C5	1.1161'-05	1.1157'-05
1.0000'-06	1.1157'-C6	1.1635'-C6	1.1157'-06
1.0000'-07	1.1157'-C7	5.9013'-C7	1.1157'-07
1.0000'-08	1.1157'-C8	4.7968'-06	1.1157'-08
1.0000'-09	1.1157'-C9	4.7858'-05	1.1161'-09
1.0000'-10	1.1157'-10	4.7856'-C4	1.1196'-10
1.0000'-11	1.1157'-11	4.7856'-C3	1.1549'-11
1.0000'-12	1.1152'-12	4.7856'-C2	1.5073'-12
1.0000'-13	1.1031'-13	4.7856'-C1	5.0022'-13
1.0000'-14	9.1163'-15	4.7856'+C0	3.6358'-13
1.0000'-15	1.2326'-15	4.7856'+C1	4.1340'-13
1.0000'-16	4.9304'-16	4.7856'+C2	8.2483'-13

La colonne C1 correspond à une majoration de $S_f(a,b)$ selon la formul 2.34, tandis que la colonne C2 correspond à la formule 2.42. Dans cet exemple a et b convergent vers 0.5 et f(0.5) est de l'ordre de 1.5, donc le rapport x_4/x_3 converge vers +1 et c'est, d'après ce que nous avons montré, la formule 2.42 (colonne C2) qui donne asymptotiquement le meilleur résultat.

On constate en effet qu'à partir de $H = (b-a) = 10^{-5}$ les valeurs dans C1 commencent à donner une borne de plus en plus mauvaise arrivant à être près de 10^{15} fois la valeur de C2. D'autre part, les valeurs de la colonne C2 coïncident avec $fl(S_f(a,b))$, pour les digits imprimés, jusqu'à un pas de l'ordre de 10^{-8} et elle n'est que plus grande que $fl(S_f(a,b))$ par

un facteur de l'ordre de 10^3 pour $H = 10^{-16}$. Dans cet exemple, nous avons pris $E_f = 10^{-15}$.

Cet exemple montre donc l'intérêt que l'on a à choisir la correction de $S_f(a,b)$ selon les critères que nous avons développés.

IV.2.2. - Correction de $I_f(a,b)$ et de $\hat{I}_f(a,b)$

Dans la section précédente, nous avons donné une méthode pour corriger en machine la quantité $S_f(a,b)$ et cela avec le but de corriger le calcul de la constante W^2 qui nous l'avons vu, est fortement instable quand (b-a) converge à zéro.

Pour la même raison nous devons corriger la valeur calculée en machine de $I_f(a,b)$ ou celle de $\hat{I}_f(a,b)$ si l'on approche l'intégrale par la méthode proposée dans la section IV.1.2 (équation 2.24).

Correction de $I_f(a,b)$

Supposons que l'on connaisse explicitement la fonction

$$I_f(x) = \int_0^x [f^{(q)}(s)]^2 ds$$

et que l'on veuille calculer, en machine, la quantité

$$I_f(a,b) = I_f(b) - I_f(a).$$

En utilisant un programme P qui évalue I_f, ou une approximation rationnelle de I_f, si I_f est non-rationnel, on peut disposer en machine des valeurs :

$$fl(I_f(b)) = I_f(b)(1+\theta_1)$$

$$fl(I_f(a)) = I_f(a)(1+\theta_2)$$

d'où $\quad fl(I_f(a,b)) = I_f(b)(1+\theta_1)(1+\varepsilon_1) - I_f(a)(1+\theta_2)(1+\varepsilon_2)$

Soit $E_I = \max\{|\theta_1|, |\theta_2|\}$, alors

$$|fl(I_f(a,b)) - I_f(a,b)| \leq 2(|I_f(a)| + |I_f(b)|)E_I + O(E_I^2)$$

Remplaçant les valeurs exactes par ces approximations en virgule flottante, on obtient finalement :

$$|fl(I_f(a,b)) - I_f(a,b)| \leq 2\{|fl(I_f(a))| + |fl(I_f(b))|\}E_I + O(E_I^2) \qquad (2.44)$$

Malheureusement dans ce cas une modification comme celle que nous avons introduite dans le cas du calcul de $S_f(a,b)$ n'a pas d'effet.

Pour l'estimation de la borne E_I les commentaires que nous fait à propos de E_f restent valables.

Correction de $\hat{I}_f(a,b)$

Dans la section IV.1.2 nous avons analysé la possibilité d'approcher $I_f(a,b)$ et nous avions proposé une formule adaptée à la classe des fonctions étudiées autrement dit, une formule qui pour $f \in H^q(a,b)$ n'utilise que des valeurs de $f^{(q-1)}$. Nous reprendrons ici cette approche, pour étudier une correction de cette formule pour tenir compte des erreurs générées et propagées en machine.

La méthode proposée (équation 2.24) utilise un ensemble $\{p_j\}$ $(j = 0,\ldots,n)$ de points de $[a,b]$, $a = p_o < p_1 <\ldots< p_n = b$

$$\hat{I}_f(a,b) = \sum_{j=1}^{n} \frac{\{f^{(q-1)}(p_j) - f^{(q-1)}(p_{j-1})\}^2}{p_j - p_{j-1}}$$

Pour simplifier les notations nous allons définir les quantités

$$S_f^j = \frac{\{f^{(q-1)}(p_j) - f^{(q-1)}(p_{j-1})\}^2}{p_j - p_{j-1}} \tag{2.45}$$

et la formule s'écrit de façon abrégée

$$\hat{I}_f(a,b) = \sum_{j=1}^{n} S_f^j \tag{2.46}$$

Chacune des quantités S_f^j est de la même forme que $S_f(a,b)$ et nous pouvons donc écrire pour chacune d'elles une relation du type

$$|fl(S_f^j) - S_f^j| \leq B_f^j \qquad (j = 1,\ldots,n) \tag{2.47}$$

avec les B_f^j calculés suivant les formules développées dans la section IV.2.1.

Pour borner la quantité 2.46 nous aurons besoin du lemme suivant :

Lemme

Soit $S = \sum_{j=1}^{n} u_j$, $d_j = fl(u_j) - u_j$ $(j = 1,\ldots,n)$ et soit E une borne des erreurs relatives aux opérations arithmétiques en virgule flottante. Alors :

$$|fl(S) - S| \leq \left(\sum_{j=1}^{n} j\,|fl(u_j)|\right)E + \sum_{j=1}^{n} |d_j|$$

≡ ≡ ≡

La démonstration se fait facilement en utilisant les règles de l'arithmétique à virgule flottante (équation 2.26).

En utilisant le lemme avec $u_j = S_f^j$, $S = \hat{I}_f$ et en prenant compte du fait que $|d_j| \leq B_f^j$ on obtient :

$$(2.48)\qquad |fl(\hat{I}_f(a,b)-\hat{I}_f(a,b)| \leq \left(\sum_{j=1}^{n} j|fl(S_f^j)|\right)E+\sum_{j=1}^{n} B_f^j$$

Evidemment, on peut arranger la somme $\sum_{j=1}^{n} S_f^j$ de façon à minimiser la borne 2.48.

La formule 2.48 permet de calculer une majoration de $\hat{I}_f(a,b)$. Nous allons étudier, sur la simulation que nous avons employée dans les secti précédentes, la variation de l'écart de cette majoration par rapport à l'inté grale exacte quand (b-a) converge vers zéro. D'après le résultat de la sectic IV1.1 nous considèrerons comme valeur exacte la quantité $I_f(a,b)$ pour des grandes valeurs de h et la valeur calculée avec une quadrature par intervalle pour les h petits.

Dans la table suivante, nous avons recopié, par commodité, les valeu de $I_f(a,b)$ dans la deuxième colonne et ceux de la borne supérieure de la méth de quadrature par intervalles dans la troisième. Dans la colonne 4, sous le titre C1, nous avond donné les valeurs obtenues en corrigeant la valeur de $I_f(a,b)$ à l'ordinateur (équation 2.44) et dans la cinquième colonne nous avon listé les valeurs corrigées de $\hat{I}_f(a,b)$ (équation 2.48) pour n = 2. Dans tous les cas nous avons pris $E_f = E_I = 10^{-14}$.

ETUDE COMPARATIVE DES DIFFERENTES APPROCHES DE LA SEMI-NORME

H	DEFINITION	BORNE SUP.	C1	C2
1.0000'+00	2.7249'+01	3.8007'+05	2.7249'+01	4.8325'+00
1.0000'-01	1.0498'-01	7.2319'+01	1.0498'-01	1.0358'-01
1.0000'-02	1.1150'-02	1.3809'-01	1.1150'-02	1.1148'-02
1.0000'-03	1.1157'-03	1.7516'-03	1.1157'-03	1.1157'-03
1.0000'-04	1.1157'-04	1.1729'-04	1.1157'-04	1.1157'-04
1.0000'-05	1.1157'-05	1.1213'-05	1.1157'-05	1.1157'-05
1.0000'-06	1.1157'-06	1.1162'-06	1.1157'-06	1.1157'-06
1.0000'-07	1.1157'-07	1.1157'-07	1.1157'-07	1.1157'-07
1.0000'-08	1.1157'-08	1.1157'-08	1.1158'-08	1.1158'-08
1.0000'-09	1.1151'-09	1.1157'-09	1.1161'-09	1.1172'-09
1.0000'-10	1.1206'-10	1.1157'-10	1.1310'-10	1.1314'-10
1.0000'-11	1.0839'-11	1.1157'-11	1.1876'-11	1.2726'-11
1.0000'-12	9.9476'-13	1.1157'-12	2.0316'-12	2.6843'-12
1.0000'-13	-4.6185'-14	1.1157'-13	9.9067'-13	1.6700'-12
1.0000'-14	1.1688'-12	1.1157'-14	2.2057'-12	1.4270'-12
1.0000'-15	-2.6290'-13	1.1157'-15	7.7396'-13	2.9729'-12
1.0000'-16	6.2883'-13	1.1157'-16	1.6657'-12	2.2441'-11

Nous pouvons remarquer que les valeurs de $I_f(a,b)$ (colonne 2) et de sa borne C1 coïncident pour des pas h inférieurs à 10^{-7} (au moins dans les 5 premiers digits) et qu'à partir de $h = 10^{-8}$ la borne C1 commence à s'écarter. Elle reste néanmoins du même ordre que la valeur "exacte" calculée en utilisant une technique d'intervalles jusqu'à $h = 10^{-12}$ après quoi, elle croît jusqu'à devenir plus grande que la borne supérieure par un facteur de 10^4 pour $h = 10^{-16}$.

Ce phénomène de croissance pour des pas inférieurs à 10^{-12} est dûe, comme nous l'avons remarqué au facteur $1/fl(b-a)$ qui apparait dans les corrections. Il aura comme résultat un ralentissement de la convergence des méthodes de localisation mais empêchera l'apparition des instabilités de calcul qui apparaissent dans la valeur de $I_f(a,b)$ à partir de $h = 10^{-9}$.

Quant à la valeur approchée C2, on voit que, comme on pouvait s'y attendre, elle sous-estime la valeur de la semi-norme quand h est trop grand et ne devient du même ordre que pour des pas h de l'ordre de 10^{-2}. A partir de $h = 10^{-9}$ elle commence aussi à croître mais la croissance est moins importante que celle de C1.

Nous reviendrons sur ces résultats à l'occasion de l'analyse de quelques exemples numériques d'application des méthodes de localisation, que nous ferons dans le prochain chapitre. Pour l'instant, nous nous contenterons de remarquer que l'emploi des corrections des valeurs numériques obtenues en machine est très important, si l'on veut atteindre des localisations de mesure suffisamment petite.

IV.2.3 - Correction de W^2

D'après les résultats des sections précédentes, nous sommes maintenant en mesure de corriger la valeur machine de W^2. Pour cela il suffit de rassembler les résultats concernant $I_f(a,b)$ (ou $\hat{I}_f(a,b)$) et $S_f(a,b)$ et de tenir compte de la définition de W^2. En utilisant les définitions de $I_f(a,b)$, $S_f(a,b)$ et l'équation 2.16 qui définit W^2 on a :

$$(2.49) \qquad W^2 = \left(\frac{b-a}{3}\right)^{q-1} \{I_f(a,b) - S_f(a,b)\}$$

ou bien, si l'on utilise une quadrature approchée pour le calcul de la semi-norme

$$(2.50) \qquad \hat{W}^2 = \left(\frac{b-a}{3}\right)^{q-1} \{\hat{I}_f(a,b) - S_f(a,b)\}$$

La partie la plus importante de l'erreur dans le calcul de W^2 (ou de $\hat{W}^2$) est produite par la différence $I_f(a,b) - S_f(a,b)$ (ou $\hat{I}_f(a,b) - S_f(a,b)$) car, comme nous l'avons vu, quand (b-a) devient petit, l'erreur relative de $I_f(a,b)$ et de $S_f(a,b)$ croît de façon très sensible. En négligeant l'erreur de $fl[(\frac{b-a}{3})^{q-1}]$ par rapport à l'erreur de la différence, quand $q = 2$, on peut écrire :

$$(2.51) \qquad fl(W^2) = \left(\frac{b-a}{3}\right)^{q-1} fl\{I_f(a,b) - S_f(a,b)\}$$

$$(2.52) \qquad fl(\hat{W}^2) = \left(\frac{b-a}{3}\right)^{q-1} fl\{\hat{I}_f(a,b) - S_f(a,b)\}$$

Prenons d'abord l'équation 2.51 et développons les calculs suivant les règles de l'arithmétique en virgule flottante

$$fl(W^2) = \left(\frac{b-a}{3}\right)^{q-1} \{fl(I_f(a,b))(1+\varepsilon_1) - fl(S_f(a,b))(1+\varepsilon_2)\}$$

$$fl(W^2) = \left(\frac{b-a}{3}\right)^{q-1} \{fl(I_f(a,b)) - fl(S_f(a,b)) + \varepsilon_1\, fl(I_f(a,b)) - \varepsilon_2\, fl(S_f(a,b))\}$$

Soit $C_1 = fl(I_f(a,b)) - I_f(a,b)$

et

$$C_2 = fl(S_f(a,b)) - S_f(a,b)$$

alors

$$fl(W^2) = (\frac{b-a}{3})^{q-1}\{I_f(a,b)-S_f(a,b)+C_1-C_2+\varepsilon_1\ fl(I_f(a,b))-\varepsilon_2\ fl(S_f(a,b))\}$$

d'où

$$fl(W^2) = W^2+(\frac{b-a}{3})^{q-1}\ \{C_1-C_2+\varepsilon_1\ fl(I_f(a,b))-\varepsilon_2\ fl(S_f(a,b))\}$$

et finalement

$$(2.53)\qquad |fl(W^2)-W^2| \leq (\frac{b-a}{3})^{q-1}\{|C_1|+|C_2|+(|fl(I_f(a,b))|+|fl(S_f(a,b))|)E\}$$

Etant donné les définitions de C_1 et C_2, on peut borner l'expression 2.53 en utilisant les bornes de $|fl(I_f(a,b))|$ et de $|fl(S_f(a,b)-S_f(a,b)|$ que nous avons calculées dans les paragraphes IV.2.2 et IV.2.1 respectivement. Dans le cas de $|C_2|$, sa borne doit être calculée suivant les critères donnés à la section IV.2.1 soit par l'expression 2.34, soit par l'expression 2.42, 2.43.

Un développement tout à fait analogue peut être fait pour l'équation 2.52, et avec

$$\hat{C}_1 = fl(\hat{I}_f(a,b))-\hat{I}_f(a,b)$$

on obtient l'expression

$$(2.54)\qquad |fl(\hat{W}^2)-\hat{W}^2| \leq (\frac{b-a}{3})^{q-1}\{|\hat{C}_1|+|C_2|+(|fl(\hat{I}_f(a,b))|+|fl(S_f(a,b))|)E\}$$

Dans ce cas la quantité $|\hat{C}_1|$ doit être bornée en utilisant la relation 2.48.

Pour finir, remarquons que l'analyse d'erreur que nous venons de faire permet d'améliorer le comportement des méthodes par rapport aux erreurs générées et propagées par la machine. Néanmoins, le problème fondamental de l'emploi de fonctions machine à la place des fonctions de H^q reste non résolu.

CHAPITRE - III

ESSAIS NUMERIQUES ET ANALYSE DE RESULTATS

I - INTRODUCTION

Dans ce chapitre nous allons étudier sur quelques exemples les problèmes de localisation du maximum global et celui de la localisation des racines d'une fonction sur un intervalle. Dans le cas du maximum global, nous comparerons nos résultats avec ceux de la méthode de SHUBERT-PIYAVSKII ([28],[26]), que nous avons brièvement décrite au paragraphe 3 du premier chapitre.

Nous avons considéré les méthodes correspondantes aux algorithmes 1 et 2 pour q = 1 et q = 2 et des versions approchées qui utilisent l'approche de la semi-norme étudiée au paragraphe IV.1.2. du chapitre II. Ci-dessous, nous définissons la notation que nous allons employer par la suite.

	ALGORITHME 1		ALGORITHME 2	
	SEMI-NORME EXACTE	SEMI-NORME APPROCHEE	SEMI-NORME EXACTE	SEMI-NORME APPROCHEE
q=1	M1	-	M3	A3
q=2	M2	A2	M4	-

Toutes les méthodes ont été élaboréesen utilisant la valeur corrigée de la constante W (chapitre II). Pour cela il faut définir, en entrée de chaque programme, un paramètre E, estimation de l'erreur relative des valeurs calculées de f, f' et I_f.

Quant aux suites $\{T_n\}$ de familles d'intervalles nous avons pris en général l'intervalle tout entier au départ pour les méthodes exactes et un découpage en 4 intervalles pour les méthodes approchées.

A chaque itération la famille T_{n+1} a été construite en doublant le nombre d'intervalles ce qui correspond à prendre r = 2 dans le procédé de base P.B. (cf. paragraphe V.1, chapitre I) et à prendre la suite $\{x_i\} = \{a,b,a+(b-a)/2,\ a+(b-a)/4,\ a+3(b-a)/4,\ldots\}$ dans l'algorithme 2.

II - LOCALISATION DU MAXIMUM GLOBAL

Soit $f \in H^q\ [a,b]$ (q = 1,2) et considérons le problème (I) du chapitre I,

"Trouver $E = \{x \in [a,b] / f(x) = \alpha = \max_{x \in [a,b]} f(x)\}$"

Pour $\varepsilon > 0$ nous cherchons à construire un ensemble L_ε, localisation de E, tel que $mes(E-L_\varepsilon) < \varepsilon$ et un intervalle I_α, localisation de α, tel que $\alpha \in I_\alpha$.

Par abus de langage nous dirons que L_ε est une localisation du maximu global et nous réserverons l'expression "encadrement du montant" pour I_α.

Comme nous l'avons remarqué à la fin du premier chapitre (Prop. 1.12) l'encadrement du montant est fait par le montant discret k_n^- et le maximum global du majorant, noté λ_n^+, c'est-à-dire :

$$I_\alpha^{(n)} = [k_n^-, \lambda_n^+]$$

Nous allons présenter d'abord un exemple qui nous permettra de décrir les caractéristiques générales des différentes méthodes pour étudier ensuite d'autres exemples présentant quelques aspects particuliers qui diffèrent de ce comportement général.

EXEMPLE n° 1 $f(x) = (-3x+1.4) \sin(18x)$ sur $[0,1]$

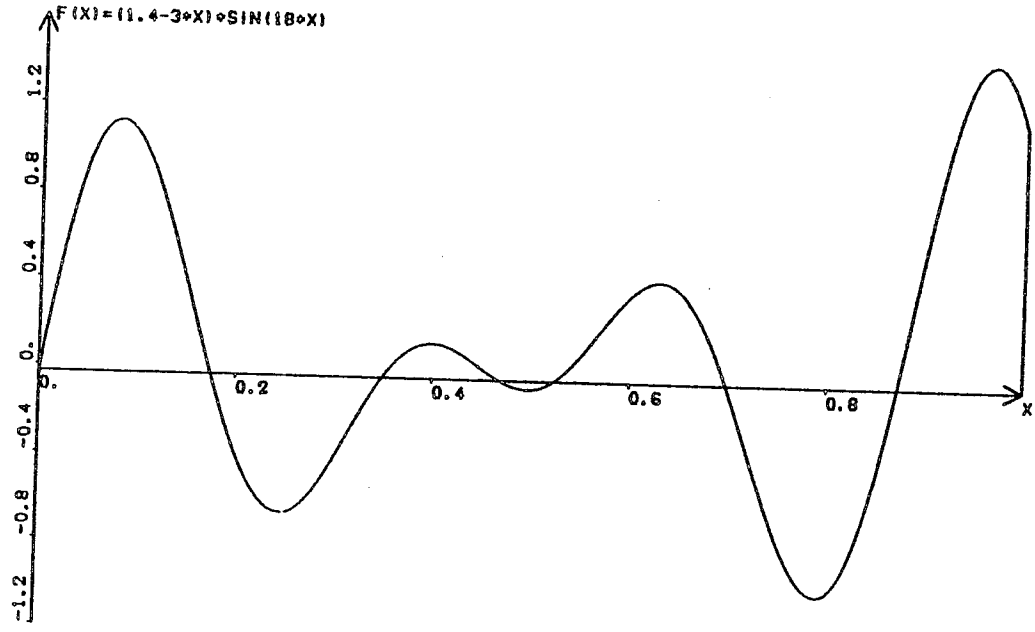
Il s'agit d'une fonction dont le maximum global

$$\alpha = 1.48907253868961$$

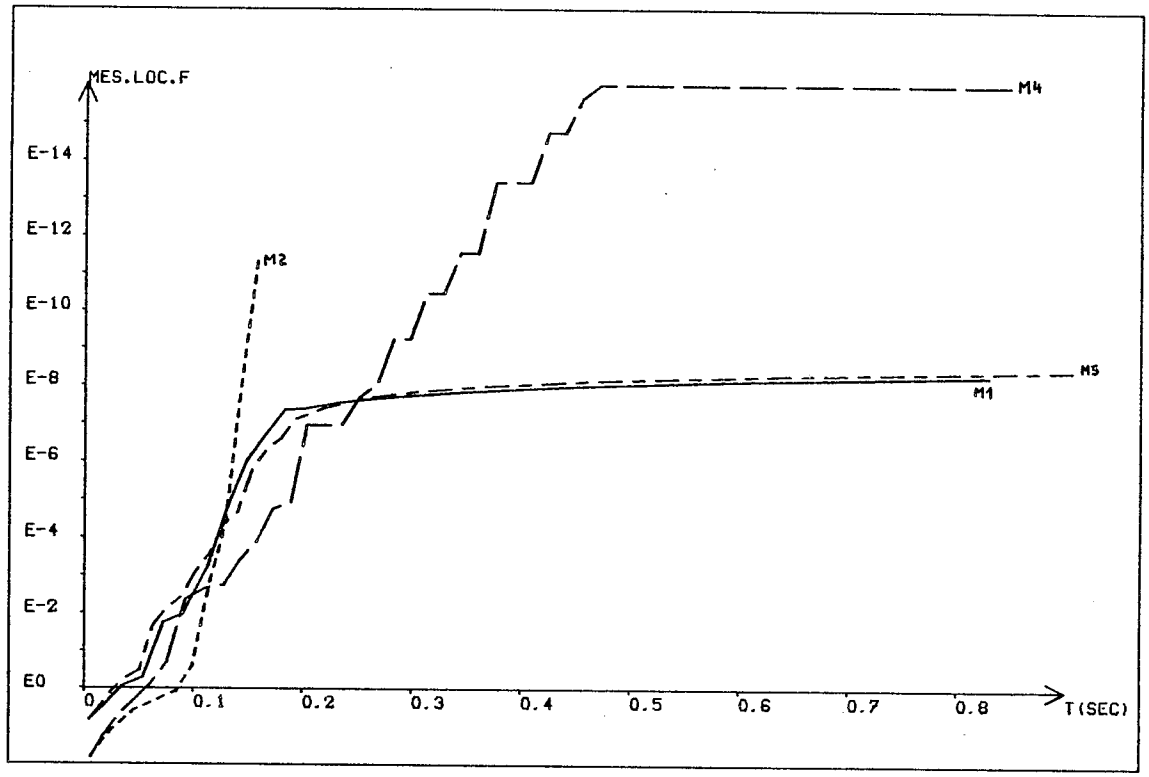
est atteint en un seul point

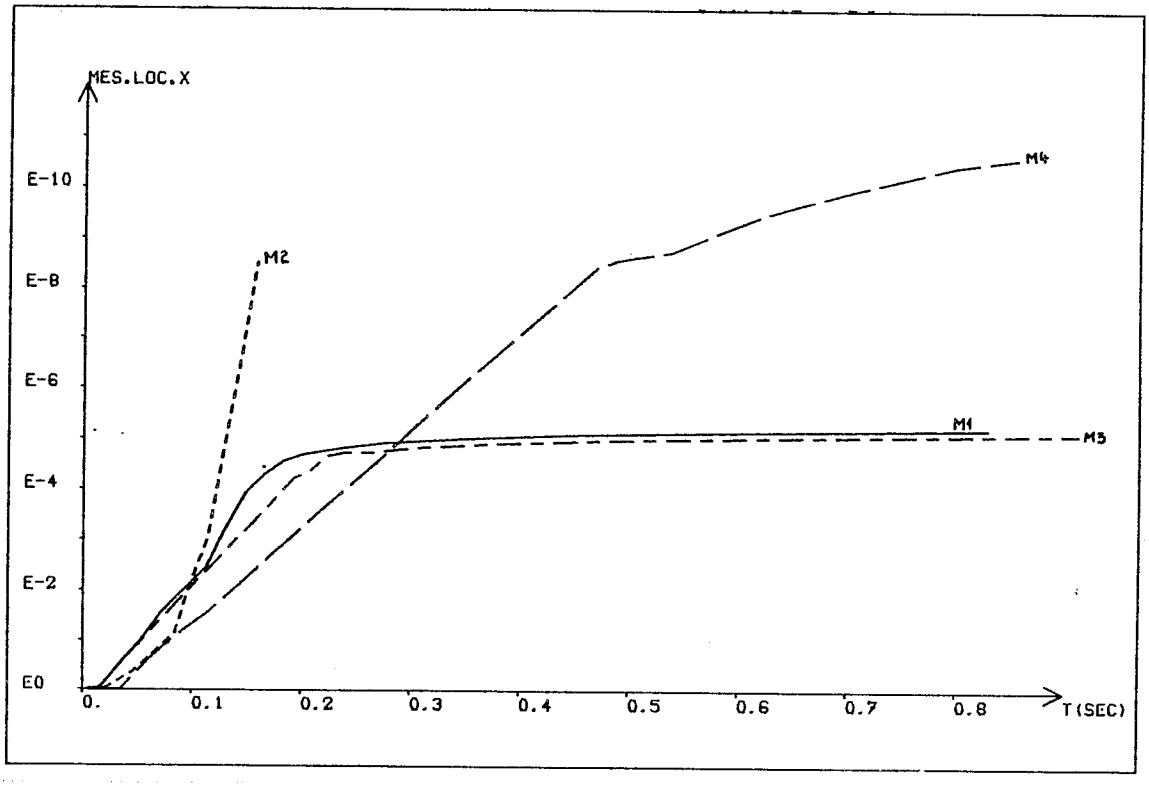
$$x_1 = 0.966085803826851$$

De plus, les autres maxima locaux sont assez éloignés de x_1 et leur montant est nettement plus petit que α .



Pour comparer les méthodes entre elles nous avons représenté d'une part la mesure de la localisation (MES.LOCX) et d'autre part la mesure de l'encadrement du montant (MES.LOCF) en fonction du temps. Le temps est mesuré en secondes correspond au temps d'exécution sur un ordinateur IBM 360/67, le langage employé étant Algol W.





Sur chaque courbe on peut distinguer une première région de croissance lente. Cette situation étant plus marquée dans les méthodes M2 et M4. Ce comportement est dû au fait que les majorants sont encore trop exagérés.

Après avoir franchi cette première région et quand la mesure des intervalles de la famille T_n devient suffisamment petite (donc les majorants suffisamment fins), on rentre dans une région de croissance rapide. La pente de la courbe dans cette région est fonction du coût d'évaluation de f, f' et I_f, de l'espace H^q utilisé et de la forme de la courbe autour du maximum. Nous analyserons ce dernier facteur dans l'exemple n° 3 de cette section.

Dans les méthodes M1 et M3 et dans une moindre mesure dans M4, il apparaît une troisième région de ralentissement de la convergence. Ce ralentissement apparaît précisément au moment où commence à se faire sentir une instabilité dans le calcul de W. La correction que nous avons été amené à introduire (paragraphe IV.2 chapitre II) a pour résultat que la majoration devient relativement de plus en plus grossière quand la mesure de la localisation devient petite. Le phénomène est moins marqué pour les méthodes M2 et M4 du fait du facteur (b-a)/3 qui apparaît dans le calcul de W (cf. eq. 2.14).

Il faut remarquer que si l'on n'utilise pas la correction de la constante W^2, ou bien elle devient négative produisant une erreur dans le programme, ou bien la localisation est fausse car elle ne contient plus la solution.

En pratique, le ralentissement de la convergence apparaît sous la forme d'une croissance du nombre d'intervalles appartenant à la famille T_n. Comme chaque nouvelle localisation est construite en évaluant f et I_f aux extrêmes de ces intervalles (et f' s'il s'agit de M2 et M4) chaque itération devient de plus en plus coûteuse.

Nous utiliserons par la suite le nombre d'intervalles de T_n comme une mesure de la vitesse de convergence des méthodes. Comme chaque localisation est une réunion des intervalles de T_n (et éventuellement de points isolés) et comme ces intervalles peuvent être contigüs (cf. procédé de base P.B., pagraphe V.1, chapitre I), le nombre d'intervalles disjoints composant la localisation sera toujours inférieur ou égal au nombre d'intervalles de la famille T_n. Quand il n'y aura pas de confusion possible, nous employerons le terme "intervalles de la localisation" pour nous référer aux éléments de T_n.

Dans le cas de la méthode M2, malgré la correction de W, il subsiste une certaine instabilité pour des localisations de petite mesure. Cette instabilité est dûe aux erreurs dans le calcul des racines des équations qui définissent la localisation (cf. proposition 2.9). Ces erreurs peuvent inverser l'ordre des racines et certains intervalles sont éliminés à tort. Pour éviter cette erreur, nous avons arrêté l'algorithme dès que la mesure relative de la localisation est inférieure à 10^{-7}.

Cette instabilité pourrait être contrôlée par une étude de l'erreur semblable à celui que nous avons faite pour la constante W ou bien par l'utilisation d'une méthode itérative pour le calcul des racines qui tire partie du fait qu'il s'agit d'équations de deuxième degré pour déterminer une localisation un peu élargie. Dans une méthode proposée par BRENT [3], [4] pour le calcul du montant à ε-près, on utilise une constante de Lipshitz de f' et ce même type d'instabilité apparaît. Elle a été surmontée précisément par une analyse d'erreur.

Un autre aspect dont il faut tenir compte est une limitation intrinsèque à toutes les méthodes qui cherchent le maximum global, en utilisant les valeurs de f. Si α_M désigne le montant-machine et f_M la fonction-machine, alors l'ensemble E_M des nombres-machine tels que $f_M(x) = \alpha_M$ est en général beaucoup plus large que l'ensemble solution E. Dans le cas présent, l'ensemble E_M contient l'intervalle de nombres-machine [0.966085798, 0.966085808], c'est à-dire E_M est plus large qu'un intervalle de longueur 10^{-8}.

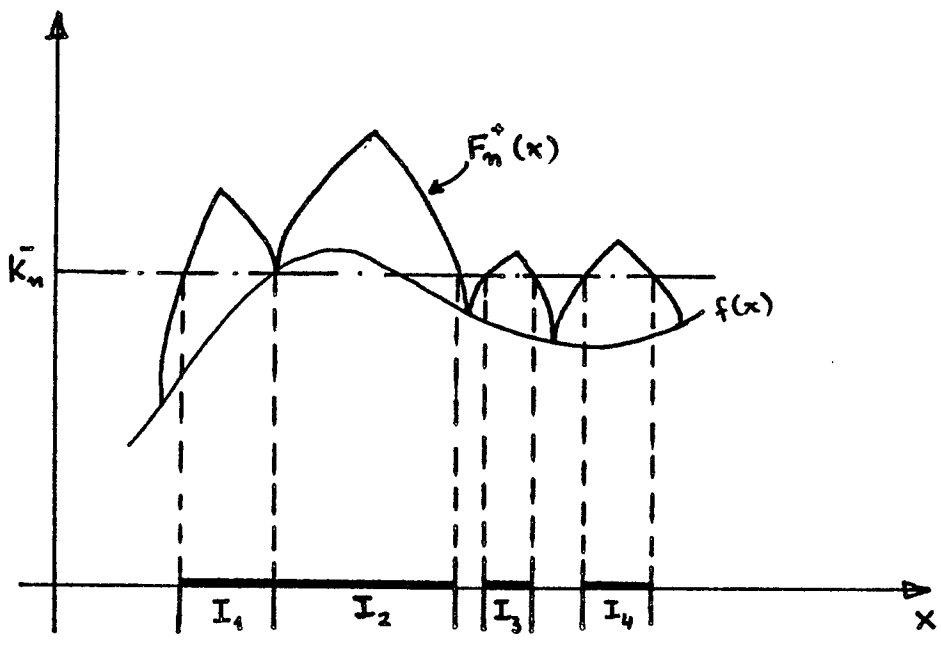
Dans les courbes correspondantes à l'encadrement du montant, on peut observer le même comportement que dans celles de la localisation du maximum. Néanmoins, la précision atteinte est dans ce cas plus grande. D'autre part, on observe, spécialement pour M4, des paliers dans les courbes qui sont dûs au comportement de k_n^- qui peut être le même d'une itération à une autre.

Ci-dessous nous donnons quelques résultats pour chacune des méthodes. Les algorithmes ont été arrêtés soit parce que la mesure relative est inférieure à la précision demandée (10^{-7} pour M2 et A2 et 10^{-12} pour les autres), soit par un dépassement du nombre maximal d'intervalles appartenant à T_n (300). Les deux premières lignes dans chaque méthode correspondent, à peu près, aux itérations limitant la région de convergence rapide tandis que la troisième correspond à la dernière itération. Dans la deuxième colonne (INT) nous donnons le nombre d'intervalles (de la famille T_n) composant la localisation et dans les trois colonnes suivantes le nombre d'évaluations de la fonction f (NEF), de la dérivée f' (NED) et de la primitive I_f (NEI).

	IT	INT	NEF	NED	NEI	TEMPS	MES.LOCX	MES.LOCF
M1	5	2	23	-	23	0.072	2.75D-C2	1.80D-02
	11	2	56	-	56	C.183	2.79D-C5	4.49D-08
	20	174	818	-	818	1.914	4.42D-C6	2.68D-09
M2	4	1	23	23	23	0.084	7.33D-C2	1.18D+00
	9	1	38	38	38	0.158	3.C2D-C9	5.42D-12
A2	2	1	11	19	-	0.035	7.27D-C2	1.21D+00
	6	1	23	39	-	C.C91	9.C5D-C8	6.26D-09
M3	5	2	9	-	9	0.050	1.25D-C1	3.36D-01
	16	2	31	-	31	C.193	6.1CD-C5	7.96D-08
	26	217	348	-	348	1.374	6.47D-C6	2.60D-09
A3	3	1	17	-	-	0.042	6.25D-C2	9.82D-02
	15	2	53	-	-	C.206	3.C5D-C5	2.79D-08
	24	173	561	-	-	1.594	5.16D-C6	1.63D-09
M4	5	2	11	11	11	0.075	1.25D-C1	2.07D-01
	30	2	61	61	61	C.470	3.72D-C9	********
	40	14	125	125	125	C.856	2.55D-11	********

Dans la méthode M4 l'encadrement du montant est réduit à un seul point à partir de IT = 27, c'est-à-dire le montant a été calculé avec certitude à la précision machine. Cela est représenté par les étoiles qui apparaissent dans la table.

Dans la méthode M1 le maximum global a été isolé à la cinquième itératio Puis jusqu'à IT = 11, la localisation est réduite à un seul intervalle compos soit par un, soit par deux intervalles contigüs de T_n. On observe en général que le montant discret k_n^- est une meilleure approximation du montant α que le maximum λ_n^+ du majorant. A partir de IT = 12 le nombre d'intervalles de la localisation (éléments de T_n) commence à croître rapidement jusqu'à IT = 20 o on a 174 intervalles. Ces intervalles sont disjoints à l'exception de deux d'entre eux (forcément contigü) dont sa réunion contient précisément le maximum global. L'extrême commun de ces deux intervalles est précisement un point dont la valeur de f définit le montant discret k_n^-. Le fait que les autres intervalles soient disjoints s'explique parce que la dérivée du majora est non-bornée dans tout voisinage des points extrêmes des intervalles de T_n qui servent à définir la nouvelle localisation. Cette situation est schématisée dans la figure ci-dessous :



La méthode M2 a un démarrage plus lent que M1 car, le majorant dans H^2 est plus grand que celui de H^1 quand l'intervalle est grand. Ainsi pour IT = 3 la localisation est composée de quatre intervalles disjoints correspondant au quatre maximum locaux de f et la mesure totale est 0.44. A partir de IT = 4 le maximum global est isolé et la localisation est réduite à un seul intervalle. Dans cette méthode ne se produit pas la croissance du nombre d'intervalles composant la localisation car la convergence de la suite de majorants vers f est suffisamment rapide, et surtout, parce que dans ce cas on a aussi une convergence de la dérivée du majorant vers la dérivée de la fonction. Pour IT = 9 la mesure de l'encadrement du montant est de 5.42'-12 mais le montant discret k_n^- coïncide avec le montant α à la précision machine.

La méthode A2 converge plus rapidement que M2 car, la semi-norme étant approchée par défaut, le majorant est plus petit que celui de M2.

La méthode M3 conçue à partir d'un test sur chaque intervalle de T_n a l'avantage par rapport à M1 d'effectuer une sorte de "fermeture" de la localisation en remplissant les trous entre les intervalles contigüs de T_n ce qui a pour résultat une diminution du nombre d'évaluation. Or, la localisation étant plus large, elle converge moins rapidement. D'autre part, la localisation contient dans le meilleur des cas deux intervalles de T_n et comme l'algorithme opère par sous-division des intervalles, son comportement optimal ne peut être que dichotomique. Le rapport entre M1 et M3 dépend donc de la vitesse de convergence de M1 (celle de M3 étant limitée) et du coût d'évaluation de la fonction f et la primitive I_f. Dans l'exemple considéré, la localisation obtenue avec M3 est toujours réduite à un seul intervalle (composé de plusieurs intervalles contigüs de T_n). Pour IT = 11, M3 a fait 21 évaluations contre 56 de MI mais la mesure de la localisation est de 1.95'-03 contre 2.79'-05 de M1. Pour IT = 21 la méthode M3 a fait 54 évaluations et la mesure est de 1.53'-05 (donc légèrement inférieure à celle de M1 pour IT = 11). Par ailleurs, sur les courbes de la page 94 , on voit que ces deux méthodes sont, pour cet exemple, comparables.

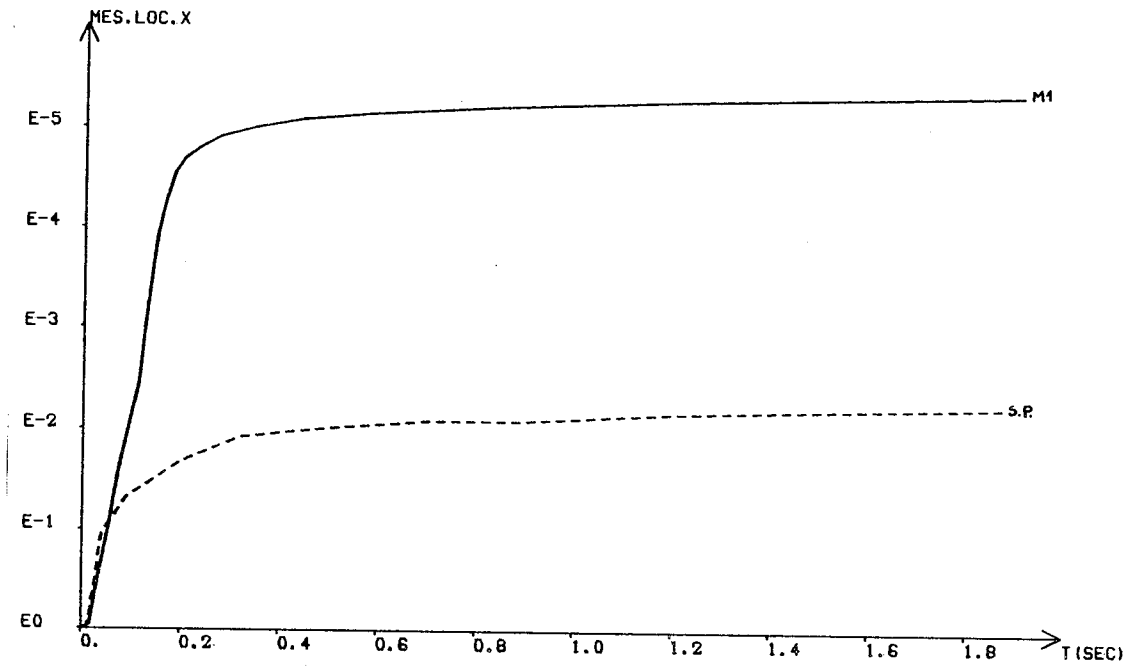
La conduite de A3 par rapport à M3 est semblable à celle de A2 par rapport à M2. Du fait de la sous-estimation de la semi-norme, elle converge légèrement plus vite. Cette méthode, n'utilise que des valeurs de la fonction

mais par contre, elle n'est pas capable d'assurer une localisation effective de la solution du moment et l'on n'est pas assuré d'avoir un majorant. Néanmoins, dans cet exemple ainsi que dans presque tous ceux que nous avons étudiés, cette méthode réussit bien.

La méthode de M4 est du même type que M3, mais cette fois les majorants utilisés sont ceux construits dans H^2. La conduite est donc liée au comportement de M2 et à celle de M3. Comme pour M2, cette méthode a un démarrage lent, accentué par la propriété de fermeture, ce qui fait que pour IT = 3, la localisation est encore l'intervalle [0,1]. A partir de IT = 5, elle développe sa vitesse de convergence maximale, c'est-à-dire chaque localisation est composée de deux intervalles contigüs de T_n. A partir de IT = 27 l'encadrement du montant est réduit à un seul point. A partir de IT = 31 il se produit une croissance du nombre d'intervalles de T_n composants la localisation, mais il est nettement moins accentué que celui observé dans M1 ou M3.

≡ ≡ ≡

Pour finir avec l'analyse de cet exemple, nous allons comparer la méthode M1 avec la méthode de Shubert-Piyavskii. Ci-dessous on montre le graphe de la mesure de la localisation en fonction du temps pour ces deux méthodes.



On peut observer qu'au début la méthode de Shubert-Piyabskii a un léger avantage par rapport à M1. Cela est du au fait que notre majorant a une dérivée non-bornée aux extrêmités des intervalles. Néanmoins, à partir d'une mesure de l'ordre de 0.1, la méthode de Shubert-Piyabskii commence à fléchir, ne pouvant atteindre que des mesures de l'ordre de 10^{-2}. Cela est du au fait que dans la méthode de Shubert-Piyabskii on utilise la même constante de Lipshitz tout au long de l'algorithme, tandis que nous recalculons notre constante W sur chaque nouvelle localisation. L'effet de croissance du nombre d'intervalles que nous avons remarqué dans M1 se produit aussi dans la méthode de Shubert-Piyavskii, mais dans ce cas, il apparait beaucoup plus tôt.

Dans la table ci-dessous, nous donnons quelques résultats obtenus avec la méthode de Shubert-Piyavskii. Dans cette méthode, le nombre d'évaluations de la fonction coïncide avec le nombre ditérations. La constante de Lipshitz a été calculée en majorant la dérivée de f sur [0,1] et entrée comme une donnée du programme. Les temps d'exécution ne considèrent donc pas le temps de calcul de cette constante.

METHODE DE SHUBERT-PIYAVSKII

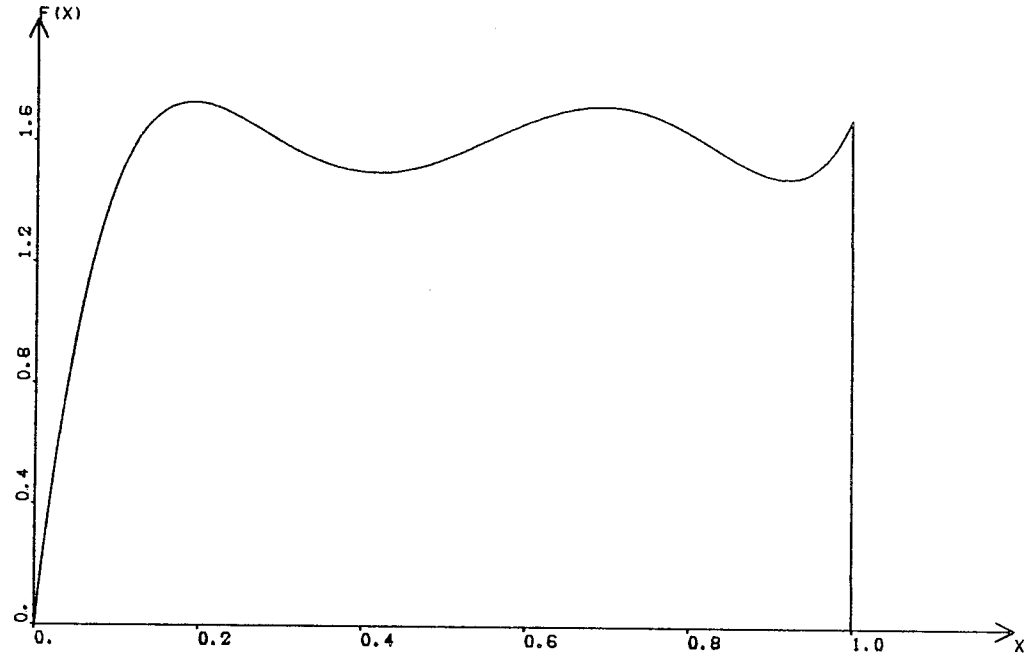
IT	INT	TEMPS	MES.LOCX	MES.LCCF
20	7	0.04	1.08 D-01	2.28 D-01
40	21	0.09	4.81 D-02	2.59 D-02
60	31	0.20	2.09 D-02	1.01 D-02
80	41	0.32	1.23 D-02	5.20 D-03
100	53	0.49	9.51 D-03	3.02 D-03
500	317	11.86	2.35 D-03	9.27 D-05

Telle qu'elle a été proposée par Shubert-Piyabskii, la méthode est fortement dépendante du fait de l'utilisation d'une même constante de Lipshitz. D'autre part, un éventuel calcul d'une nouvelle constante de Lipshitz à chaque itération apparaît difficile pour une fonction quelconque. Nous discuterons cette possibilité dans le cas des polynômes dans l'exemple suivant.

EXEMPLE n° 2 $\quad f(x) = \sum_{i=0}^{5} a_i x^i \quad$ sur $[0,1]$

avec $a_o = 0.0$, $\quad a_1 = 25.0$, $\quad a_2 = -128.0$, $\quad a_3 = 282.5$, $\quad a_4 = -278.7$

et $\quad a_5 = 100.9$



Cette fonction a trois maximum locaux dans [0,1] dont les montants sont voisin

x	f(x)
0.189168220683086	1.72866096364692
0.689273618554688	1.72080674452856
1.0	1.7

Nous donnons ci-dessous quelques résultats des diverses méthodes. La première ligne de chaque méthode correspond à l'itération à partir de laquelle le maximum global a été isolé. La deuxième ligne de M1, M3, A3 et M4 correspond à la dernière itération avant que ne commence la croissance du nombre d'intervalles, c'est-à-dire, au moment ou la courbe de la mesure de la localisation en fonction du temps commence à fléchir. Pour M2 et A2 la deuxième ligne correspond à la dernière itération (mesure relative inférieure à 10^{-7}). Pour M1, M3 et A3 la troisième ligne de la table correspond à la dernière itération avant de dépasser 300 intervalles dans T_n, tandis que pour M4 elle correspond au nombre maximal d'itérations qui avait été prévu.

	IT	INT	NEF	NED	NEI	TEMPS	MES.LOCX	MES.LOCF
M1	5	2	31	-	31	0.077	1.9C⁰-C2	2.81⁰-03
	10	2	56	-	56	C.154	8.32⁰-C5	3.56⁰-08
	19	182	908	-	908	1.7C9	1.36⁰-05	2.16⁰-09
M2	4	2	21	20	20	0.069	6.C1⁰-C2	2.60⁰-01
	10	1	41	40	40	0.153	2.8E⁰-C9	2.00⁰-14
A2	3	1	21	34	-	0.063	1.12⁰-C2	1.93⁰-02
	7	1	33	54	-	0.11S	9.36⁰-C9	1.74⁰-10
M3	5	2	12	-	12	0.058	1.25⁰-C1	1.18⁰-02
	15	2	32	-	32	0.181	1.22⁰-C4	1.51⁰-07
	24	184	300	-	300	1.C71	2.20⁰-C5	2.47⁰-09
A3	3	2	23	-	-	0.058	1.25⁰-C1	8.76⁰-03
	13	2	55	-	-	0.200	1.22⁰-C4	6.49⁰-08
	22	218	687	-	-	1.S15	2.6C⁰-C5	3.51⁰-09
M4	5	2	11	11	11	0.067	1.25⁰-C1	3.40⁰-03
	28	2	57	57	57	0.401	1.4S⁰-C8	********
	40	6	104	104	104	0.676	1.CS⁰-11	********

Le comportement des méthodes est semblable à celui de l'exemple précéden
La méthode M4 permet le calcul du montant avec certitude, à la précision machine, à partir de IT = 23. Cela est exprimé sur la table par les étoiles dans la dernière colonne.

Dans la table ci-dessous, on donne les résultats obtenus avec la méthode de Shubert-Piyavskii. Ici encore la constante de Lipshitz a été introduite comme une donnée (L = 25.0).

METHODE DE SHUBERT-PIYAVSKII

IT	INT	TEMPS	MES.LOCX	MES.LCCF
20	19	0.05	0.95	9.16'-01
40	39	0.14	0.94	4.26'-01
60	59	0.28	0.94	2.43'-01
80	79	0.45	0.94	1.72'-01
100	87	0.78	0.80	1.12'-01
200	147	2.72	0.54	3.40'-02
300	195	5.67	0.33	1.53'-02
400	191	9.98	0.19	6.55'-02
500	203	14.12	0.14	3.74'-03

La méthode de Shubert-Piyavskii est plus coûteuse en temps et en nombre d'évaluations de la fonction que les méthodes que nous proposons. En plus, au moins dans le cas des polynômes, le calcul de la semi-norme de la fonction est très simple, tandis que le calcul de la constante de Lipshitz est plus compliqué. GANSHIN [11] a étudié le problème du calcul de la constante de Lipshitz dans des méthodes qui permettent le calcul du montant à ε-près, notamment pour la méthode de EVTUSHENKO [9]. Pour le cas des polynômes, il propose l'utilisation d'une décomposition en deux polynômes monotones, pour sur-estimer la constante de Lipshitz. En particulier, Ganshin étudie une stratégie "presque-optimale" pour fixer le nombre d'évaluations du polynôme (points équidistants), qu'il convient de faire pour le calcul d'une constante de Lipshitz de façon à minimiser les évaluations du polynôme nécessaires pour approcher le montant à ε-près.

Suivant Ganshin, si l'on note P le polynôme et si R et Q sont deux polynômes à coefficients positifs tels que $P(x) = R(x)-Q(x)$, alors on peut calculer une constante de Lipshitz en évaluant P dans (K+1) points équidistants par la formule :

$$L = \max \{v_i \ / \ i = 1,\ldots,k\}$$

$$v_i = \max \{R'(t_i)-Q'(t_{i-1}), | R'(t_{i-1})-Q'(t_i) |\}$$

$$t_i = t_{i-1}+(b-a)/K \quad i = 1,\ldots,k$$

$$t_o = a$$

Dans la table ci-dessous, nous donnons les valeurs de la constante L pour différentes valeurs de K.

ESTIMATION DE L

K	L
50	76.79
100	40.72
150	28.71
170	25.89
180	25.03
190	25.02
240	25.01

Il faudrait donc de l'ordre de 180 évaluations du polynôme pour obtenir une bonne estimation de L. Malheureusement, la structure de la méthode de Shubert-Piyabskii ne permet pas l'utilisation de ces évaluations par la suite, sauf évidemment, pour avoir une première approximation du montant discret.

Par contre, la méthode que nous proposons est très adaptée pour l'utilisation de n'importe quel majorant M_f^+ a condition qu'il satisfasse une condition du type

$$\lim_{b \to a \to 0} \max_{x \in [a,b]} \{M_f^+(x) - f(x)\} = 0$$

Comme évidemment, le majorant utilisé par Shubert-Piyavskii satisfait cette condition, on pourrait utiliser notre méthode avec des majorants construits à partir d'une constante de Lipshitz calculée sur chaque intervalle de la famille de T_n en suivant le procédé proposé par Ganshin. Nous pensons qu'une telle méthode pourrait éliminer le principal désavantage de la méthode de Shubert-Piyavskii qui est l'utilisation de la constante de Lipshitz calculée sur tout l'intervalle.

Nous remarquerons pour terminer, que Ganshin généralise cette façon de sur-estimer la constante de Lipshitz à d'autres types de fonctions, mais cette généralisation n'est utilisable que pour une classe trop restreinte de fonctions, car elle suppose une décomposition de la fonction en des fonctions dont on peut calculer facilement les racines.

≡ ≡ ≡

EXEMPLE n° 3

Il s'agit du même polynôme que dans l'exemple précédent à l'exception du coefficient a_5 qui est changé en $a_5 = 101.0$. Cette modification renverse l'ordre des maximums locaux, le maximum étant atteint maintenant au point $x = 1.0$ avec $f(1.0) = 1.8$.

Le comportement des méthodes M1, M3 et A3 est semblable à celui observé dans l'exemple 2, sauf que maintenant, dès que le maximum global est isolé, il l'est dans une localisation composée par un seul intervalle. D'autre part, la croissance du nombre d'intervalles est dans ce cas beaucoup moins importante ce qui permet d'atteindre une meilleure précision.

Dans les méthodes M2 et A2 on obtient la solution comme un point isolé en 4 et 2 itérations respectivement. Cela s'explique par le fait que pour W suffisamment petit, le majorant devient convexe par morceaux sur la localisation.

Dans la méthode M4 se produit à partir de IT = 27 une croissance rapide du nombre d'intervalles de la localisation. Cette croissance, que l'on n'avait pas dans les exemples précédents, est dûe à une convergence plus lente du majorant car la dérivée n'est pas nulle dans le maximum.

	IT	INT	NEF	NED	NEI	TEMPS	MES.LOCX
M1	3	1	11	-	11	0.031	1.99D-01
	8	1	26	-	26	0.087	2.46D-11
	13	7	53	-	53	0.167	3.76D-12
M3	4	1	7	-	7	0.038	1.25D-01
	36	1	39	-	39	0.329	2.91D-11
	40	4	47	-	47	0.377	7.28D-12
A3	2	1	13	-	-	0.028	1.25D-01
	33	1	75	-	-	0.376	5.82D-11
	39	9	113	-	-	0.511	8.19D-12
M4	4	1	8	8	8	0.055	1.25D-01
	26	1	30	30	30	0.273	2.98D-08
	40	123	311	311	311	1.515	2.42D-10

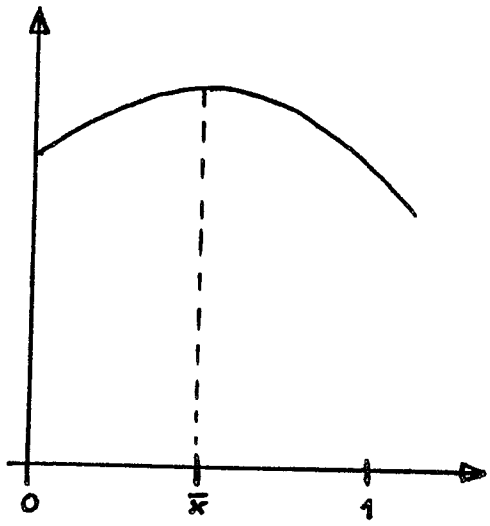
La méthode de Shubert-Piyavskii, dont nous donnons quelques résultats ci-dessous, converge d'abord lentement jusqu'à IT = 140, puis très rapidement jusqu'à localiser le maximum à précision machine pour IT = 280. Ce comportement est dû au fait que le maximum de f' est atteint précisément dans x = 1.0.

METHODE DE SHUBERT-PIYAVSKII

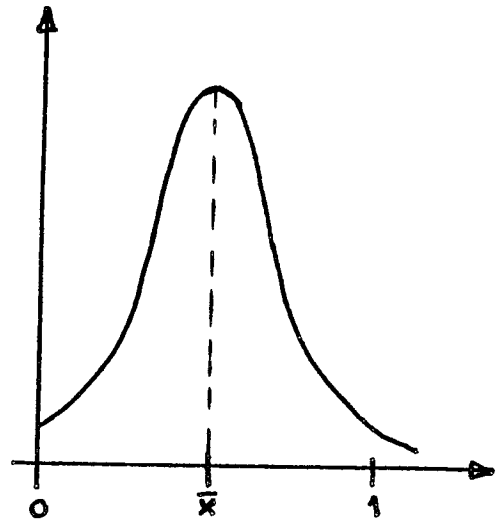
IT	INT	TEMPS	MES.LOCX
20	19	0.05	9.43D-01
140	21	1.44	1.10D-01
260	9	1.80	2.70D-13

EXEMPLE n° 4 $\quad f(x) = \dfrac{1}{1+[A(x-\bar{x})]^2}$, $\bar{x} = \dfrac{4}{9}$ sur [0,1]

Nous avons étudié le comportement des différentes méthodes en fonction du paramètre A. Pour des petites valeurs de A on obtient une fonction de plus en plus plate de sorte que pour $A = 10^{-8}$ l'approximation machine de f est presque la fonction constante égale à 1. Pour des grandes valeurs de A on obtient une fonction présentant un pic autour de $\bar{x}$ de plus en plus accusé.



A < 1



A > 1

Dans la table ci-après, nous donnons les résultats obtenus, dans un même nombre d'itérations, pour $A = 10^{-4}$, 1 et 10^{4}.

	IT	A	INT	NEF	NED	NEI	TEMPS	MES.LOCX	MES.LOCF
M1	11	1.0-4	82	436	-	436	0.766	2.55-02	2.10-12
		1.0+0	2	49	-	49	0.154	2.12-05	9.31-11
		1.0+4	2	73	-	73	0.183	1.74-03	1.59+00
	20	1.0-4	+	+	+	+	+	+	+
		1.0+0	127	673	-	673	1.256	2.08-06	1.41-12
		1.0+4	2	121	-	121	0.319	2.46-08	1.30-08
M2	4	1.0-4	1	14	12	12	0.048	4.21-12	*
		1.0+0	1	12	12	12	0.045	2.63-04	2.37-04
		1.0+4	8	20	20	20	0.075	1.00+00	1.96+04
	18	1.0-4	*	*	*	*	*	*	*
		1.0+0	*	*	*	*	*	*	*
		1.0+4	1	420	420	420	1.257	2.03-11	2.50-07
A2	4	1.0-4	1	14	24	-	0.048	5.97-12	*******
		1.0+0	1	14	24	-	0.050	3.51-08	2.14-08
		1.0+4	1	17	29	-	0.063	2.67-02	(1)
	14	1.0-4	*	*	*	*	*	*	*
		1.0+0	*	*	*	*	*	*	*
		1.0+4	1	50	84	-	0.204	5.77-09	4.96-05
M3	13	1.0-4	147	254	-	254	0.755	3.59-02	1.54-12
		1.0+0	2	25	-	25	0.139	4.88-04	2.10-08
		1.0+4	2	33	-	33	0.158	4.88-04	8.28-01
	27	1.0-4	+	+	+	+	+	+	+
		1.0+0	205	379	-	379	1.247	3.05-06	1.41-12
		1.0+4	2	61	-	61	0.318	2.58-08	3.39-09
A3	7	1.0-4	256	513	-	-	1.325	1.00+00	4.84-07
		1.0+0	2	29	-	-	0.095	7.81-03	1.73-06
		1.0+4	2	29	-	-	0.096	7.81-03	(1)
	19	1.0-4	+	+	-	-	+	+	+
		1.0+0	182	567	-	-	1.491	1.74-04	7.56-09
		1.0+4	2	69	-	-	0.272	1.91-06	1.02-05
M4	7	1.0-4	3	13	13	13	0.085	4.69-02	7.00-15
		1.0+0	2	13	13	13	0.092	3.13-02	4.82-05
		1.0+4	64	65	65	67	0.254	1.00+00	8.65+02
	25	1.0-4	+	+	+	+	+	+	+
		1.0+0	13	65	65	65	0.429	7.75-07	1.00-15
		1.0+4	2	163	163	163	0.760	1.19-07	7.01-08

Dans cette table, le signe + indique que l'on a dépassé dans une itération précédente, les 300 intervalles dans la localisation. Le symbole * indique que la précision demandée a été atteinte dans une itération précédente.

Le symbole ******* indique que le montant a été encadré à la précision machine Finalement, les valeurs marquées (1) correspondent à des instabilités numérique dans le calcul du maximum du majorant qui empêchent le calcul de l'encadrement du montant.

A mesure que A devient petit, les méthodes de H^1 ont de plus en plus de difficulté à réduire la mesure de la localisation, tandis que le montant est encadré de façon de plus en plus précise. Pour A inférieur à 10^{-8} elles ne sont même pas capable de réduire la localisation initiale [0,1].

A l'inverse, les méthodes M2 et A2 n'ont aucune difficulté à localiser la solution même pour $A = 10^{-12}$. Ces deux comportements différents sont évidemment dûs à la convergence de la dérivée du majorant dans ce deuxième type de méthodes.

La méthode M4, si bien elle utilise les majorants en H^2, est limitée par l'utilisation d'un test sur chaque intervalle, de sorte que son comportement est semblable à la méthode M3.

Pour des grandes valeurs de A, les méthodes M1 et M3 convergent de plus en plus vite, mais le montant est encadré cette fois avec plus de difficulté. Néanmoins, pour des valeurs de A trop grandes (plus grande que 10^8) on retrouve à nouveau un ralentissement dû a ce que la semi-norme, et donc le majorant, est exagérément grand. Par exemple, pour $A = 10^{+8}$ et pour des intervalles de longueur de l'ordre de 10^{-3} autour du maximum global, le montant du majorant est de l'ordre de 5.5×10^2.

Les méthodes M2 et M4 ont un comportement irrégulier pour de grandes valeurs de A. D'abord, elles convergent très lentement avec une croissance du nombre d'intervalles composant la localisation, puis, dès que les intervalles sont suffisament petits, elles retrouvent leur convergence rapide. En réalité, cela n'est autre chose qu'une amplification du phénomène observé dans l'exemple n° 1 où ces méthodes avaient un démarrage lent (première région de la courbe) à cause d'un majorant trop grand.

Pour des valeurs de A supérieures à 10^8 il apparaît une instabilité dans M2 qui fait disparaître la localisation, tandis que M4 atteint plus de 300 intervalles dans la localisation, sans pouvoir sortir de la première phase de convergence lente.

Quant aux méthodes approchées, elles ont une tendance à trop sous-estimer la semi-norme quand A devient grand. Cela est dû au fait qu'en utilisant trop peu de points dans le calcul de la semi-norme, ces méthodes "sautent" le pic. Le maximum du majorant est par conséquent inférieur au montant du maximum global dans les premières itérations ce qui fait perdre toute signification à l'encadrement du montant (d'où les valeurs marquées (1) dans la table). Une fois la mesure des intervalles devenue suffisament petite pour "attraper" le pic, tout rendre dans l'ordre et ces méthodes réussissent à localiser la solution même pour $A = 10^8$.

≡ ≡ ≡

La méthode de Shubert-Piyavskii pour sa part, à une tendance à rejoindre le comportement de M1 pour des petites valeurs de A, tout en restant plus coûteuse, tandis que pour de grandes valeurs de A sa convergence se ralenti rapidement. Ainsi par exemple pour $A = 10^4$, avec 500 évaluations de la fonction et 16.42 secondes d'exécution, la mesure de localisation est de 0.999995967.

≡ ≡ ≡

EXEMPLE n° 5 $\quad f(x) = \sum_{k=1}^{5} k \sin\{(k+1)x+k\} \quad$ sur $[-10,10]$

Cet exemple sert à étudier le cas de fonctions dont le maximum est atteint en plusieurs points et en même temps, il sert à étudier le comportement des méthodes pour des fonctions très oscillantes.

Le maximum global (machine) de cette fonction est :

$$\alpha = 12.0312494421671$$

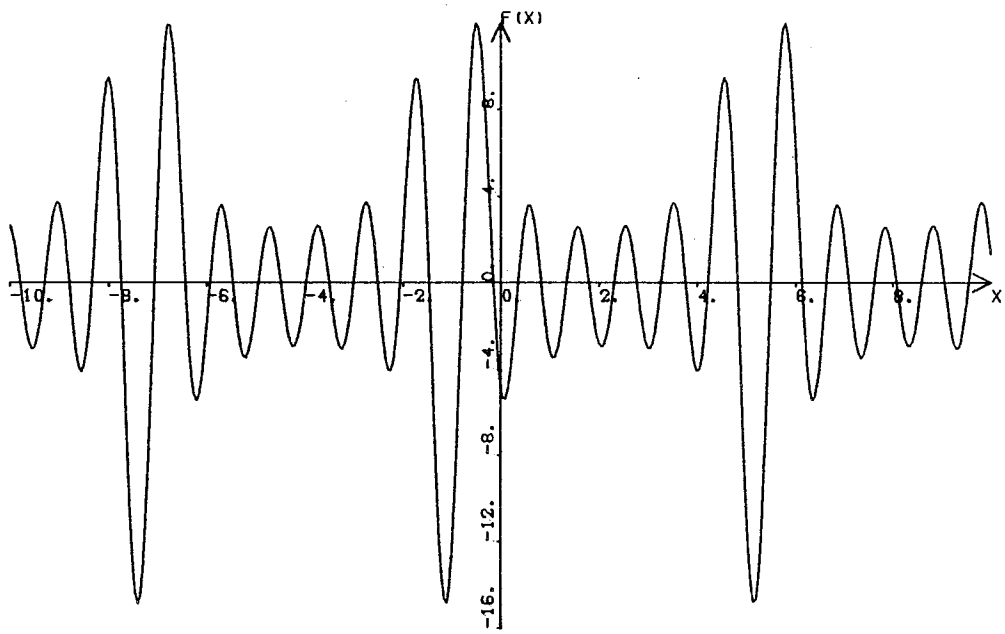
et il est atteint en trois points :

$x_1 \doteq -6.77457614343890$

$x_2 \doteq -0.491390836259315$

$x_3 \doteq 5.79179447092027$

Comme on peut le voir dans la figure ci-dessous, cette fonction a seize autres maximum locaux dont trois d'un montant supérieur à 9.0



La méthode M1 isole les trois maximum en sept itérations dans la localisation de mesure 2.95'-02, 9.01'-02 et 2.77'-02 tandis que le montant est encadré dans un intervalle de longueur 0.4. Le temps est de 0.821 secondes et il a fallu 116 évaluations (de f et de I_f). Après le comportement par rapport à chaque maximum est le même et il correspond étant donné la forme du maximum, à celui que nous avons observé dans l'exemple précédent pour $A > 1$.

La méthode M2 isole les trois maximum, elle aussi, en sept itérations dans des intervalles de mesure 5.82'-02, 5.73'-02 et 6.01'-02. Le nombre d'évaluations est aussi égal à celui de M1 mais il faut y ajouter les évaluations de f' ce qui produit un temps d'exécution un peu plus grand : 1.019 secondes. Le comportement relatif de M2 par rapport à M1 ainsi que les résultats obtenus à partir de IT= 8 sont conséquent avec ceux que nous avons remarqués dans l'exemple précédent.

La méthode M3 a aussi un comportement conséquent avec les observations de l'exemple précédent. Le seul fait qu'il faut remarquer est que cette fois elle est plus efficace que M1 ainsi pour IT = 22 cette méthode produit des localisations du même ordre que celles produites par M1 pour IT = 13 (9.54'-05, 5.72'-05 et 8.58'-05) mais elle fait 130 évaluations contre 197 de M1 et elle emploie 1.184 secondes contre 1.409 de M1.

Le seul fait remarquable dans la méthode M4 est l'apparition d'une instabilité à la 31-ème itération qui fait disparaître les localisations de x_2 et x_3 ne laissant que celle de x_1. Cette instabilité est dûe à une mauvaise estimation de l'erreur E que nous avons prise égale à 10^{-12}. En prenant $E = 10^{-10}$, la méthode reprend sa conduite normale avec croissance d'intervalles à partir de IT = 27.

Quant aux méthodes approchées A2 et A3, elles ne localisent pas la solution si l'on prend 4 intervalles au départ, à cause de la sous-évaluation de la semi-norme. En prenant 40 intervalles au départ, elles réussissent à localiser la solution, mais pour A2 il a fallu aussi prendre $E = 10^{-10}$ pour éviter des instabilités numériques.

Finalement, la méthode de Shubert-Piyavskii a un comportement semblable à celui du premier exemple, de sorte que, avec 500 évaluations de f, on obtient une localisation de mesure totale 0.089 en 10.69 secondes.

III - LOCALISATION DE RACINES

Etant donné $f \in H^q(a,b)$, $q = 1,2$, nous considèrerons dans ce paragraphe le problème :

(II) "Trouver $E = \{x \in [a,b]/f(x) = 0\}$"

En particulier, étant donné ε positif, nous cherchons une localisation L_ε, réunion d'intervalles, telle que $mes(E-L_\varepsilon) < \varepsilon$

Evidemment, il faut maintenant utiliser aussi une suite de minorants et nous allons prendre $k_n^- = k_n^+ = 0$ dans les algorithmes 1 et 2. Remarquons néanmoins, que telle que les méthodes ont été développées, on peut aussi envisager le problème de localisation des racines de l'équation $f(x) = k$, à condition de disposer de deux suites $\{k_n^-\}$ et $\{k_n^+\}$ satisfaisant aux hypothèse (H.1) et (H.2) du paragraphe 2 du premier chapitre.

EXEMPLE n° 6 $\quad f(x) = \sum_{i=0}^{5} a_i x^i$

avec $a_o = -1.6$, $a_1 = 25.0$, $a_2 = -128.0$, $a_3 = 282.5$, $a_4 = -278,7$ et $a_5 = 100.9$

Il s'agit de la même fonction que dans l'exemple n° 2, mais translatée en 1.6 (graphe page 102). Ce polynôme a ses cinq racines dans [0,1] et il nous sert à caractériser le comportement typique des différentes méthodes.

D'une façon générale, le comportement observé dans cet exemple est semblable au comportement des méthodes dans le calcul du maximum global. Nous nous contenterons donc de donner quelques résultats et de remarquer les aspects dans lesquels on a pu observer quelques différences.

METHODE M1

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
4	23	-	23	0.057	1.98'-01	8.11'-02 1.20'-02 5.33'-03 1.25'-02 8.74'-02	1 1 1 1 1
8	98	-	98	0.175	1.31'-09	5.08'-10 4.47'-11 4.30'-11 2.51'-11 6.91'-10	1 1 1 1 1
11	155	-	155	0.316	3.36'-11	2.62'-12 9.64'-12 1.15'-11 7.00'-12 2.87'-12	2 4 5 5 2

On peut observer que les mesures atteintes par M1 sont maintenant plus petites que celles produites par la même méthode dans le calcul du maximum global. Cela est dû au fait que l'ensemble des solutions-machine est dans ce cas beaucoup plus petit que dans le cas du maximum global.

Le nombre d'intervalles composant la localisation est pour la même raison plus petit dans l'autre cas. Remarquons néanmoins que les intervalles sont maintenant tous disjoints.

Nous profiterons de cet exemple pour mettre en relief l'importance de la correction de la constante W^2. Pour cela nous avons programmé une version de M1 sans correction de W^2. Cette version s'arrête au cours de IT = 6 car W^2 devient négatif. Pour IT = 5 la mesure de chaque localisation est légèrement inférieure à celle de la méthode avec correction (une différence de l'ordre de 10^{-11}). La mesure totale de la localisation est, pour la méthode sans correction, de 2.45101'-02 et celle de M1 est de 2.45102'-02.

Nous pouvons donc conclure que l'utilisation de la correction de W^2 ne change pas, de manière importante, les résultats au cours des premières itérations. D'autre part, cette correction permet à la méthode d'atteindre des précisions beaucoup plus grandes. Dans l'exemple présent, la mesure totale de la localisation continue à décroître jusqu'à atteindre 1.38'-11 pour IT = 15.

METHODE M3

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
5	15	-	15	0.066	3.75'-01	1.25'-01 6.25'-02 6.25'-02 6.25'-02 6.25'-02	2 1 1 1 1
33	156	-	156	0.600	1.16'-09	2.33'-10 2.33'-10 2.33'-10 2.33'-10 2.33'-10	1 1 1 1 1
38	192	-	192	0.730	9.46'-11	7.28'-12 2.18'-11 3.64'-11 2.18'-11 7.28'-12	1 3 5 3 1

Dans la méthode M3 on observe aussi un gain dans la précision qui peut être par rapport au calcul du maximum global, à la différence du cas du maximum global où les résultats de M2 étaient comparables à ceux de M1, ici on observe une très nette supériorité de M1 aussi bien dans le nombre d'évaluation que dans le temps d'exécution.

METHODE A3

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
5	49	-	-	0.125	7.81D-02	1.56D-02 1.56D-02 1.56D-02 1.56D-02 1.56D-02	1 1 1 1
33	377	-	-	1.057	1.16D-09	1.75D-10 2.91D-10 3.49D-10 2.33D-10 1.16D-10	3 5 6 4 2
38	821	-	-	2.195	2.05D-10	2.73D-11 4.73D-11 5.46D-11 4.73D-11 2.91D-11	15 26 30 26 16

Dans la table ci-dessus, nous donnons les résultats obtenus avec A3 pour les mêmes valeurs de IT présentées pour M3. On peut observer que le nombre d'intervalles croît beaucoup plus vite en comparaison avec M3. Cela entraîne un plus grand nombre d'évaluations (si bien que l'on n'évalue que la fonction) donnant un temps d'exécution qui est pour IT = 38 trois fois celui de M3. Cette situation est le résultat d'une plus grande instabilité numérique de $\hat{I}_f$ par rapport à I_f ce qui produit une correction de W^2 plus importante (et donc une majoration plus grossière).

La méthode A3 ne devient compétitive par rapport à M3 que si le coût d'évaluation de f' est beaucoup plus important que celui de f (ce qui n'est pas le cas dans cet exemple).

METHODE M2

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
4	23	23	23	0.093	3.94'-02	1.68'-02 4.00'-06 5.00'-04 1.80'-02 4.15'-03	1 1 1 1 1
5	38	38	38	0.151	2.65'-06	2.19'-06 1.78'-15 1.61'-10 2.98'-07 1.69'-07	1 1 1 1 1
6	47	47	47	0.189	1.61'-10	7.48'-15 ******** ******** 6.94'-17 1.25'-16	1 * * 1 1

Les méthodes ont été programmées de sorte que dès qu'un intervalle atteint la mesure relative désirée (10^{-7} pour M2 et A2) il n'est plus analysé. Dans la table ci-dessus, nous avons marqué par des étoiles les intervalles qui ont atteint la précision demandée dans l'itération précédente. On peut observer que certaines solutions sont localisées, pour une même nombre d'itérations, avec plus de précision que d'autres. Cela tient d'une part à la position des racines par rapport aux points où on évalue la fonction et d'autre part à la forme de la courbe autour de la racine.

Pour finir nous donnons les résultats des méthodes A2 et M4. La méthode A2 converge un peu plus vite que M2 pour les raisons que nous avons déjà analysées tandis que M4 a une conduite semblable à celle observée dans le calcul du maximum global.

METHODE A2

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
2	17	29	-	0.063	3.31D-02	1.23D-02 4.70D-05 3.11D-04 1.45D-02 5.88D-03	1 1 1 1 1
3	32	54	-	0.122	3.65D-06	2.36D-06 1.47D-12 1.14D-12 1.28D-06 2.56D-07	1 1 1 1 1
4	41	69	-	0.161	2.26D-12	4.45D-14 ******** ******** 2.40D-15 1.58D-15	1 * * 1 1

METHODE M4

IT	NEF	NED	NEI	TEMPS	MES.TOT.	MES.LOCX	INT
5	14	14	14	0.084	3.12D-01	6.25D-02 6.25D-02 6.25D-02 6.25D-02 6.25D-02	1 1 1 1 1
25	114	114	114	0.658	2.58D-07	5.56D-08 5.56D-08 5.56D-08 5.56D-08 5.56D-08	1 1 1 1 1
30	153	153	153	0.873	2.79D-08	1.66D-09 7.45D-09 9.31D-09 3.73D-09 5.59D-09	1 4 5 2 3

EXEMPLE n° 7

Afin d'étudier le comportement des méthodes face à un grand nombre de racines nous avons étudié la fonction de l'exemple n° 4 sur l'intervalle[-10,10] où cette fonction à 38 racines. Les méthodes approchées A2 et A3 ont dû être initialisées avec au moins 40 intervalles pour pouvoir localiser toutes les racines. Toutes les méthodes ont réussi la localisation des racines à la précision demandée (mesure relative). Une fois les racines isolées, le comportement des méthodes par rapport à chaque racine a été semblable à celui observé dans l'exemple précédent.

Nous donnons les résultats obtenus avec M1 pour IT = 10 et ceux de M2 pour IT = 9. Dans le cas de M2 cela représente la dernière itération car la précision exigée était de 10^{-7}. Pour M1, cela correspond à la dernière itération avant que ne commence la croissance du nombre d'intervalles.

METHODE M1

IT : 10
NEF : 539
NEI : 539
TEMPS: 3.614
MEST.: 4.56'-CE

LOCALISATION		MES.LOCX
-9.788236988824 41'+0C	-9.78E2369SE63953'+CC	1.85'-10
-9.29755441611623'+00	-9.29755441488449'+CO	1.23'-09
-8.79417954775360'+00	-8.79417954469589'+CO	3.05'-09
-8.33822678534708'+00	-8.33822677842 3CC'+CO	6.92'-09
-7.73795177642028'+00	-7.73795177560576'+CO	8.15'-10
-7.06930503302036'+00	-7.06930503301145'+CO	8.91'-12
-6.42987786394295'+00	-6.42987786388887'+CO	5.41'-11
-5.93072090315153'+00	-5.93072090273524'+CO	4.16'-10
-5.46149514815877'+00	-5.46149514796316'+CO	1.96'-10
-4.94553518087226'+00	-4.94553517462249'+CO	6.25'-09
-4.48736735050422'+00	-4.48736734931CCC'+CO	1.19'-09
-3.97618C04856327'+00	-3.97618CC48474C7'+CO	8.92'-11
-3.50505168169165'+00	-3.50505168146 8C2'+CO	2.24'-10
-3.01436910878907'+00	-3.01436910789346'+CO	8.96'-10
-2.51099424042985'+00	-2.51C994237 7C838'+CO	2.72'-09
-2.05504147691008'+00	-2.05504147248 6C7'+CO	4.24'-09
-1.45476646904591'+00	-1.45476646863 21C'+CO	4.14'-10
-7.86119725840803'+01	-7.86119725E38C89'+C1	2.71'-12
-1.46692556765786'+01	-1.46692556696543'+C1	6.92'-11
3.52464404224142'-01	3.52464404271C59'-C1	4.69'-11
8.21690159121771'-01	8.21690159147858'-C1	2.61'-11
1.33765012875332'-00	1.33765C13C0E891'-CO	1.32'-09
1.79581795695393'-00	1.79581795754425'-CO	5.90'-10
2.30700525865718'-00	2.30700525E7C853'-CO	5.18'-11
2.77813362554097'-00	2.77E13362561835'-CO	7.74'-11
3.26881619881765'-00	3.26E81619883C71'-CO	1.31'-11
3.77219106799960'-00	3.77219106E217C9'-CO	2.17'-10
4.22814383007154'-00	4.22E14383486355'-CO	4.81'-09
4.82841883802234'-00	4.82841883E65957'-CO	6.38'-10
5.49706558133664'-00	5.49706558134181'-CO	5.16'-12
6.13649275030200'-00	6.13649275C5E519'-CO	2.83'-10
6.63564971138714'-00	6.63564971144259'-CO	5.59'-11
7.10487546626116'-00	7.10487546633CEE'-CC	6.97'-11
7.62083543399798'-00	7.62C835439234 11'-CO	5.23'-09
8.07900326325819'-00	8.07900326565764'-CO	2.40'-09
8.59019056572394'-00	8.59C19056554334'-CO	2.19'-10
9.06131893263820'-00	9.06131893295223'-CC	2.97'-10
9.55200150594102'-00	9.5520C150601285'-CO	7.18'-11

METHODE M2

IT : 9
NEF : 365
NED : 365
NEI : 365
TEMPS: 3.506
MEST.: 2.84'-06

LOCALISATION		MES.LOCX
-9.788236988783l9'+00	-9.788236988783l9'+00	2.89'-15
-9.297554415538l3'+00	-9.297554415538C3'+00	2.71'-13
-8.794179581402l0'+00	-8.794179497084l2'+00	8.43'-08
-8.338226936304l8'+00	-8.338226546961l8'+00	3.89'-07
-7.737951778536l4'+00	-7.737951771478l3'+00	7.06'-09
-7.069305033049l2'+00	-7.069305032973l2'+00	7.59'-11
-6.429877870519l8'+00	-6.429877859151l1'+00	1.14'-08
-5.930721108036l4'+00	-5.930720622633l9'+00	4.85'-07
-5.461495148043l6'+00	-5.461496148043l6'+00	2.00'-15
-4.945535439981l4'+00	-4.495534966525l0'+00	4.73'-07
-4.487367349933l1'+00	-4.487367349933l0'+00	1.11'-14
-3.976180048493l0'+00	-3.976180048493l0'+00	6.66'-16
-3.505051681603l1'+00	-3.505051681603l9'+00	1.64'-14
-3.014369108358l9'+00	-3.014369108358l6'+00	2.81'-14
-2.510994239152l6'+00	-2.510994238931l3'+00	2.20'-10
-2.055041474691l0'+00	-2.055041474691l0'+00	1.11'-15
-1.454766475418l3'+00	-1.454766456501l4'+00	1.89'-08
-7.861197310330l4'-01	-7.861197227440l5'-01	8.29'-09
-1.466925590595l3'-01	-1.466925540877l5'-01	5.00'-09
3.524644042479l0'-01	3.524644042479l1'-01	4.02'-16
8.216901591359l8'-01	8.216901591359l9'-01	6.25'-16
1.337650129408l7'-00	1.337650129408l7'+00	6.66'-16
1.795817957246l7'-00	1.795817957246l0'+00	2.47'-14
2.307005258686l9'-00	2.307005258686l9'+00	2.44'-15
2.778133625575l9'-00	2.778133625575l9'+00	2.44'-15
3.268816198821l0'-00	3.268816198821l1'+00	2.22'-16
3.772191068105l3'-00	3.772191068105l3'+00	1.33'-15
4.228143832487l9'-00	4.228143832487l9'+00	6.66'-16
4.828418834396l2'-00	4.828418844833l4'+00	1.04'-08
5.497065581137l4'-00	5.497065581462l9'+00	3.24'-10
6.136492745327l8'-00	6.136492753938l5'+00	8.61'-09
6.635649594817l0'-00	6.635649869645l7'+00	2.75'-07
7.104875347510l2'-00	7.104875620281l2'+00	2.73'-07
7.620835436588l5'-00	7.620835436588l5'+00	3.11'-15
8.079003264425l7'-00	8.079003264425l7'+00	4.00'-15
8.590190565866l6'-00	8.590190565866l8'+00	2.26'-14
9.061318930538l3'-00	9.061318936838l6'+00	6.30'-09
9.552001191205l6'-00	9.552001973910l5'+00	7.29'-07

EXEMPLE n° 8 $\quad P(X) = \prod_{j=1}^{d} (x-x_j)$ sur [a,b]

Nous allons étudier différents exemples de polynômes donnés par ces racines et pour divers intervalles [a,b].

Pour des polynômes de d° ≤ 5 avec ses racines suffisamment séparés, le comportement des méthodes est semblable à celui de l'exemple n° 6.

Pour localiser toutes les racines réelles d'un polynôme quand on ne dispose pas d'une localisation initiale, on peut localiser d'abord les racines dans [-1,1] et ensuite, en inversant l'ordre des coefficients, déterminer les valeurs réciproques des racines sur ce même intervalle, ce qui revient à calculer les racines dans $(-\infty,-1] \cup [1,+\infty)$. Nous avons utilisé cette technique pour le polynôme de d° = 5 de racines $x_1 = -10^3$, $x_2 = -10$, $x_3 = 1$, $x_4 = 10^2$ et $x_5 = 10^4$, dont voici les résultats obtenus pour IT = 7 et IT = 14 avec M1.

METHODE M1

IT : 7
NEF : 56
NEI : 56
TEMPS : 0.125
MEST. : ********

LOCALISATION		MES.LOCX
-**********	-5.952722574865850+02	********
-1.001286538109830+01	-9.997410216845460+00	1.550-02
1.000000000000000+00	1.000000000000026+00	2.590-13
9.481683984242590+01	1.026991156611248+02	3.220+02
8.193996377622890+02	**********	********

METHODE M1

IT : 14
NEF : 131
NEI : 131
TEMPS: 0.288
MEST.: 5.22'-09

LOCALISATION		MES.LOCX
-1.000000000000048'+03	-9.999999999998161'+02	6.20'-10
-1.000000000000039'+01	-9.999999999996958'+00	6.90'-12
9.999999999998764'-01	1.000000000000096'+00	2.20'-12
9.999999999999778'+01	1.000000000000076'+02	9.81'-11
9.999999999999646'+03	1.000000000000009'+04	4.49'-09

Pour IT = 7 la localisation est encore de mesure infinie mais déjà les trois racines de module le plus petit ont été isolées. On a pu observer que les racines atteignaient la précision exigée dans l'ordre des valeurs absolues. Ainsi x_3 est localisé à la précision demandée pour IT = 7, x_2 pour IT = 10, x_4 pour IT = 12, x_1 pour IT = 14 et finalement x_5 pour IT = 15.

Nous avons aussi utilisé la forme directe sur l'intervalle [-15.000, +15.000] et les résultats obtenus (ci-dessous) sont très voisins de ceux obtenus en renversant les coefficients.

METHODE M1

IT : 14
NEF : 119
NEI : 119
TEMPS: 0.272
MEST.: 2.29'-08

-1.000000000000040'+03	-9.999999999995612'+02	7.91'-10
-1.000000000000024'+01	-9.999999999999461'+00	8.25'-12
1.000000000000000'+00	1.000000000000026'+00	2.59'-13
9.999999999999267'+01	1.000000000000020'+02	9.30'-11
9.999999999999038'+03	1.000000000000124'+04	2.20'-08

Dans cet exemple, la méthode M2 est légèrement plus lente que M1, tandis que les autres méthodes, en particulier M3 et M4, sont nettement plus lentes. L'avantage de M1 par rapport à M2 est sans doute lié à la valeur plus réduite de la semi-norme dans H^1 que celle de H^2 quand l'intervalle est grand.

≡ ≡ ≡

Pour l'utilisation du polynôme P dans nos programmes, nous avons déterminé d'abord ses coefficients ce qui introduit une erreur qui, pour des polynômes de $d^\circ \geq 6$, peut être assez importante. Par exemple, pour le polynôme de $d^\circ = 7$, et de racines $x_1 = 0.03$, $x_2 = 21.15$, $x_3 = 47.8$, $x_4 = 66.5$, $x_5 = 131.7$, $x_6 = 221.0$, $x_7 = 288.8$, la localisation de ses racines sur [0,300] présente quelques difficultés si on utilise comme estimation de l'erreur $E = 1.0^{-12}$. Avec cette valeur, la méthode M1 s'arrête car W^2 devient négative à la machine et les méthodes A2, M3 et A3 ne localisent qu'une partie des racines. Les méthodes M2 et M4 par contre, ne présentent pas de difficultés.

Si on prend $E = 1.0^{-10}$, alors toutes les méthodes, à l'exception de A2, réussissent la localisation de toutes les racines. La méthode A2 ne localise que les trois racines les plus grandes et cela depuis la première itération, ce qui montre qu'il ne s'agit plus d'un problème d'estimation de l'erreur mais du problème, déjà analysé, de sous-estimation de la semi-norme.

≡ ≡ ≡

Nous avons étudié ensuite le polynôme de $d^\circ = 20$ et de racines $x_i = i$ $(i = 1,\ldots,20)$. Dans ce cas, même avec $E = 10^{-10}$ seulement la méthode A3 a réussi la localisation de toutes les racines. Les méthodes M1, M2 et M4 s'arrêtent car W^2 devient négative tandis que A2 et M3 ne localisent qu'une partie des solutions. Ce mauvais comportement semblerait être lié, non pas aux erreurs dans le calcul des coefficients, mais à une semi-norme beaucoup trop grande. La méthode A3 serait capable de localiser la solution grâce à la sous-évaluation de la semi-nrome au début de l'algorithme.

Ci-dessous, nous donnons les résultats de A3 obtenus à la onzième itération.

METHODE A3

IT : 11
NEF : 359
TEMPS: 1.083
MEST.: 1.28'-01

LOCALISATION		MES.LOCX
0.99487	1.00098	6.10'-03
1.99585	2.00195	6.10'-03
2.99683	3.00293	6.10'-03
3.99780	4.00391	6.10'-03
4.99878	5.00488	6.10'-03
5.99976	6.00586	6.10'-03
6.99463	7.00073	6.10'-03
7.99561	8.00171	6.10'-03
8.99658	9.00269	6.10'-03
9.99756	10.00366	6.10'-03
10.99243	10.99854	6.10'-03
11.99951	12.00562	6.10'-03
12.99438	13.00049	6.10'-03
13.99536	14.00146	6.10'-03
14.99023	15.00244	1.22'-02
15.99731	16.00342	6.10'-03
16.99829	17.00439	6.10'-03
17.99927	18.00537	6.10'-03
18.99414	19.00024	6.10'-03
19.99512	20.00122	6.10'-03

Remarquons que la 11-ème localisation ne contient pas en effet la racine x_{11}. Cela est du probablement, au fait que les racines du polynôme ont été perburbées dans l'étape d'évaluation des coefficients.

Or, l'analyse d'erreur que nous avons faite en utilisant l'estimation de l'erreur totale entre les valeurs théoriques et les valeurs machine (cf. eq. 2.2.9) de sorte que les racines devraient, en principe, être bien localisées, malgré l'erreur des coefficients, en utilisant une estimation correcte de E.

Cette contradiction pourrait s'expliquer par le fait que nous n'avons considéré l'erreur d'évaluation que dans le calcul de W^2. Cependant, quand l'erreur devient plus importante, comme dans le cas présent, il perturbe aussi les autres calculs. Il faudrait donc considérer dans ces cas que l'on a une erreur dans les valeurs de f, f' et I_f dans tous les calculs. Une étude plus approfondie de cette question devrait considérée des intervalles d'incertitude [fl(y)-E, fl(y)+E] pour les valeurs de f de f' et de I_f. Dans chaque calcul on utiliserait soit la borne supérieure soit la borne inférieure de sorte à assurer une localisation effective.

≡ ≡ ≡

Finalement, nous avons considéré le cas de racines groupées, en particulier le cas de racines de multiples supérieurs à 1. Pour des racines de multiplicité 2, le comportement des méthodes est semblable à celui observé dans le calcul du maximum global. Pour des racines de multiplicité supérieure à deux, on observe un ralentissement général de toutes les méthodes et les précisions que l'on peut atteindre sont inversement proportionnelles à la multiplicité de la racine.

Ci-dessous, nous donnons les résultats de M1 et M2 correspondant à la onzième itération pour les cas d = 3, 4, 5, 6 et $x_i = 5/9$ (i = 1,...,d)

d	3	4	5	6
M1	2.35'-03	1.26'-02	3.22'-02	6.04'-02
M2	1.69'-04	1.90'-03	8.44'-03	2.14'-02

Dans cet exemple, à l'exception de M2 pour d = 3 qui atteint une mesure de 1.24'-07 dans 21 itérations, les autres localisations ne décroissent pas significativement dans les itérations suivantes.

Dans le cas d'un groupement de racines, si elles sont proches les unes des autres, la méthode se conduit comme s'il s'agissait d'une racine multiple. L'éclatement de la localisation ne se produit que si les racines sont suffisamment éloignées. Par exemple pour les racines $x_1 = 0.4441$, $x_2 = 0.4442$, $x_3 = 0.4443$, $x_4 = 0.4444$ la localisation obtenue sur M3 pour IT = 15 est : [0.4387, 0.4498]. Par contre, pour les racines $x_1 = 0.52$, $x_2 = 0.53$, $x_3 = 0.54$, $x_4 = 0.55$ on obtient une localisation pour chaque racine : [0.5197, 0.5203], [0.5291, 0.5310], [0.5390, 0.5409] et [0.5497, 0.5503], dans le même nombre d'itérations.

IV - CONCLUSIONS ET DEVELOPPEMENTS POSSIBLES

Dans ce travail nous avons présenté des méthodes de localisation qui utilisent des majorants et des minorants d'une fonction. La formulation générale pour cette classe de méthodes données au paragraphe 2 du premier chapitre nous a permis de traiter simultanément les problèmes de la localisation du maximum global et celui de la localisation des racines d'une fonction sur un intervalle.

L'idée de base pour la construction des suites de majorants et des minorants a été l'utilisation d'une suite T_n de familles d'intervalles en conjonction avec un procédé de majoration et de minoration d'une fonction sur un intervalle donné.

Nous avons développé un procédé de majoration et de minoration pour des fonctions de H^q et nous avons rappelé un autre exemple de tel type de procédé pour le cas des fonctions de Lipshitz continues (Shubert-Piyavskii).

Si l'on dispose d'un procédé que pour une fonction donnée f (appartenant à une certaine classe) permet de construire, pour tout intervalle I, un majorant $M_{f,I}$, alors la méthode que nous avons développée peut être généralisée à condition que $M_{f,I}$ soit tel que

$$\lim_{mes(I)\to 0} \max_{x\in I} \{M_{f,I}(x)-f(x)\} = 0$$

La construction d'un majorant tel que $M_{f,I}$ utilise nécessairement une certaine information globale relative à la fonction sur l'intervalle I. En général, on a employé jusqu'ici une constante de Lipshitz.

Nous avons, pour notre part, employé la semi-norme de la fonction. Dans les deux cas, cela suppose implicitement qu'on se place dans une classe particulière de fonctions. L'utilisation de classes plus restreintes de fonctions, telle que les polynômes par exemple, pourrait permettre l'utilisation d'une information globale plus particulière et le développement de méthodes plus précises.

L'utilisation d'une constante de Lipshitz ou de la semi-norme présente des limitations dans le sens du calcul pratique de cette information. Nous pensons néanmoins que la connaissance de la semi-norme est dans beaucoup de cas plus aisée que celle d'une constante de Lipshitz. Des efforts doivent être faits pour développer des méthodes permettant le calcul numérique de la semi-norme ou l'obtention de bonnes majorations. La technique des intervalles de MOORE pourrait être dans ce sens une approche intéressante.

Une généralisation de cette classe de méthodes au cas de plusieurs variables paraît difficile. Néanmoins, dans certains cas particuliers, on peut penser à décomposer le problème en une suite de problèmes de localisation à une variable. Ainsi par exemple, dans le cas des racines complexes d'un polynôme, l'étude des racines de sa partie réelle et de sa partie imaginaire sur la frontière du domaine étudié peut fournir un test qui, utilisé en conjonction avec une technique de recouvrement, nous donnerait une méthode du même type que celles qu'à étudié HENRICI.

Dans le cas particulier des méthodes que nous avons développées, il faudra approfondir l'étude des instabilités numériques, spécialement pour les méthodes de H^2. Dans ce sens, l'utilisation d'une méthode itérative pour le calcul des racines des majorants et des minorants telle que nous l'avons suggérée dans l'analyse de l'exemple n° 1 pourrait fournir une solution. Cette même idée pourrait être intéressante (à condition d'avoir une majoration effective) pour le développement de méthodes d'ordre supérieur (pour q supérieur à 2).

REFERENCES

[1] BASSO P.
Localisation du maximum global des fonctions non-unimodales
Séminaire n° 272, (1976) Grenoble

[2] BERMAN G.
Minimization by successives approximation
SIAM, J. Numer. Anal. 3 (1966), 123-133

[3] BRENT R.P.
An algorithm with guaranteed convergence for finding a zero of a function
Computer J. 14 (1971), 422-425

[4] BRENT R.P.
Algorithms for minimization without derivatives
Prentice Hall, Inc., Englewood Cliffs, New Jersey (1973)

[5] CEA J.
Optimisation : théorie et algorithmes
Dunod, Paris (1971)

[6] CERNOUS'KO F.L.
Optimal search for a zero of a function computed approximatly
Soviet Math. Dokl., 8, 6 (1967), 1382-1385

[7] COHN A.
Über die Anzahl der Wurzeln einer algebraischen Gleichung in einemen Kreise
Math. Z., vol. 14 (1922), 110-148

[8] DUC-JACQUET M.
Approximation des fonctionnelles linéaires sur les espaces hilbertie autoreproduisants.
Thèse (1973), Grenoble

[9] EVTUSHENKO YU. G.
Numerical methods for finding global extrema (case of a non-uniform mesh)
Zh. vȳchisl. Mat. mat. Fiz. 11, 6 (1971), 1390-1403

[10] FRIEDLI A.
Optimal covering algorithms in method of search for solving polynomi equations.
J. ACM, 20, 2 (1973) 290-300.

[11] GANSHIN G.S.
Function maximization
Zh. vȳchisl Mat. mat. Fiz. 16, 1 (1976), 30-39

[12] GARGANTINI I, and HENRICI P.
Circular arithmetic and the determination of polynomial zero.
Num. Math. 18 (1972), 305-320.

[13] GASTINEL N.
Introduction à l'analyse calculable
Cours DEA, Grenoble (1972-1973)

[14] GROSS O. et JOHNSON S.
Sequential minimax search for a zero of a convex function
Maht. Tables and Other Aids to Computation, 13 (1959), 44-51

[15] HENRICI P.
Methods of search for solving polynomial equations.
J. ACM, 17, 2 (1970), 273-283.

[16] HENRICI P.
Circular arithmetic and the determination of polynomial zeros
dans : "Conference on application of numerical analysis"
Lectures Notes in mathematics, 228, 86-92
Springer-Verlag, Berlin -Heildeberg - New-York (1971)

[17] HENRICI P.
Applied and computational complex analysis (Vol. I)
Wiley Int. Pub. New-York.London (1974)

[18] HENRICI P. and GARGANTINI I.
Uniformly convergent algorithms for the simultaneous approximation of all zeros of a polynomial.
dans : "Constructive aspects of the fundamental theorem of algebra"
B. Dejon and P. Henrici, Ed. Wiley, Inter-Sciences London (1969)

[19] KIEFER J.
Sequential minimax search for a maximum
Prix. Amer. Math. Soc. 4 (1953), 503-506

[20] KIEFER J.
Optimum sequential search and approximation methods under minimum regularity assumption.
SIAM J. Appl. Math. 5 (1957), 105-136

[21] LEHMER D.H.
A machine method for solving polynomial equations.
J. ACM, 8 (1961), 151-162

[22] MICCHELLI, C.A. and MIRANKER W.L.
High order search methods for finding roots
J. ACM, 22, 1 (1975), 51-60

[23] MOORE R.E.
Interval analysis
Prentice-Hall Inc., Englewood Cliffs, N.J. (1966)

[24] MOORE R.E., STROTHER W. and YANG C.T.
Interval integrals .
IMSD-703073, Lockheed Missiles and Space Co,
Palo Alto, California (1960)

[25] PICHAT M.
Contribution à l'étude des erreurs d'arrondi en arithmétique à virgule flottante.
Thèse Grenoble (1976)

[26] PIYAVSKII S.A.
An algorithm for finding the absolute extremum of a function
Z. vȳ chisl. Math. mat. Fiz, 12, 4 (1972), 888-896

[27] REITER A.
Interval arithmetic package (interval)
MRC Program ≠ 2, COOP Organ, Code-WISC
Math. Research Center, University of Wisconsin, Madison Wisconsin

[28] SHUBERT B.O.
A sequential method seeking the global maximum of a function
SIAM J. Numer. Anal. 9, 3 (1972), 379-388

[29] SUKHAREV A.G.
Optimal strategies of the search for un extremum
Zh. vȳshisl. Mat. mat. Fiz., 11, 4 (1971), 910-924

[30] SUKHAREV A.G.
Best sequential search strategies for finding an extremum
Zh. vȳschisl. Mat. mat. Fiz., 18, 1 (1972), 35-50

[31] SUKHAREV A.G.
Optimal search for the roots of function satisfying a Lipshitz condit
Zh vȳschisl. Mat. mat. Fiz., 16, 1 (1976), 20-29.

[32] SCHUR J.
Über algebraische Gleichungen, die nur Wurzeln nut negativen Realteilen besitzen
Z. Angew. Math. Mech., 1 (1920), 307-311.

[33] TOURNIER E.
Un exemple d'utilisation du calcul formel sur ordinateur.
Méthode générale de localisation des racines d'une équation algébrique à coefficients complexes.
Thèse, Grenoble (1971)

[34] WEYL H.
Randbemerbungen zu Hauptproblemen der mathematik,
II, Fundamentalsatz der Algebra und Grundlagen der Mathematik
Math. Z. 20 (1924), 131-150.

[35] WILKINSON J.H.
Rounding errors in algebraic processes
Notes on applied sciences n° 32
Her Majesty's Stationnery Office, London (1963)

Dernière page d'une thèse

VU

Grenoble, le 18 mai 1978

Le Président de la thèse

Vu, et permis d'imprimer,

Grenoble, le 23 mai 1978

Le Président de l'Université
Scientifique et Médicale

UNIVERSITÉ SCIENTIFIQUE ET MÉDICALE DE GRENOBLE

Dr G. CAU