Application de la programmation linéaire et convexe à l'approximation au sens de Tchebycheff avec contraintes

Michel Terrenoire

N° d'ordre

THESE

présentée à la Faculté des Sciences
de l'Université de Grenoble

pour obtenir
le grade de Docteur de Troisième Cycle
"MATHEMATIQUES APPLIQUEES"

par

Michel TERRENOIRE
Licencié ès Sciences

APPLICATION DE LA PROGRAMMATION LINEAIRE ET CONVEXE A L'APPROXIMATION AU SENS DE TCHEBYCHEFF AVEC CONTRAINTES

Thèse soutenue le 22 juin 1967, devant la Commission d'examen :

Monsieur	J. KUNTZMANN	Président
Messieurs	J.R. BARRA	Examinateur
	P.J. LAURENT	Examinateur

F A C U L T E D E S S C I E N C E S

LISTE DES PROFESSEURS

DOYENS HONORAIRES :

M. MORET
M. WEIL

DOYEN :

M. BONNIER E.

PROFESSEURS TITULAIRES :

MM.	NEEL Louis	Chaire de Physique Expérimentale
	HEILMANN René	Chaire de Chimie
	KRAVTCHENKO Julien	Chaire de Mécanique Rationnelle
	CHABAUTY Claude	Chaire de calcul différentiel et intégral
	BENOIT Jean	Chaire de Radioélectricité
	CHENE Marcel	Chaire de Chimie Papetière
	WEIL Louis	Chaire de Thermodynamique
	FELICI Noël	Chaire d'Electrostatique
	KUNTZMANN Jean	Chaire de Mathématiques Appliquées
	BARBIER Reynold	Chaire de Géologie Appliquée
	SANTON Lucien	Chaire de Mécanique des Fluides
	OZENDA Paul	Chaire de Botanique
	FALLOT Maurice	Chaire de Physique Industrielle
	KOSZUL Jean-Louis	Chaire de Mathématiques M.P.C.
	GALVANI O.	Mathématiques
	MOUSSA André	Chaire de Chimie Nucléaire
	TRAYNARD Philippe	Chaire de Chimie Générale

	SOUTIF Michel	Chaire de Physique Générale
	CRAYA Antoine	Chaire d'Hydrodynamique
	REULOS R.	Théorie des Champs
	BESSON Jean	Chaire de Chimie
	AYANT Yves	Physique Approfondie
	GALLISSOT	Mathématiques
Melle	LUTZ Elisabeth	Mathématiques
MM.	BLAMBERT Maurice	Chaire de Mathématiques
	BOUCHEZ Robert	Physique Nucléaire
	LLIBOUTRY Louis	Géophysique
	MICHEL Robert	Chaire de Minéralogie et Pétrographie
	BONNIER Etienne	Chaire d'Electrochimie et d'Electrométallurgie
	DESSAUX Georges	Chaire de Physiologie Animale
	PILLET E.	Chaire de Physique Industrielle et Electrotechnique
	YOCCOZ Jean	Chaire de Physique Nucléaire Théorique
	DEBELMAS Jacques	Chaire de Géologie Générale
	GERBER R.	Mathématiques
	PAUTHENET R.	Electrotechnique
	VAUQUOIS B.	Chaire de Calcul Electronique
	BARJON R.	Physique Nucléaire
	BARBIER Jean-Claude	Chaire de Physique
	SILBER R.	Mécanique des Fluides
	BUYLE-BODIN Maurice	Chaire d'Electronique
	DREYFUS B.	Thermodynamique
	KLEIN J.	Mathématiques
	VAILLANT F.	Zoologie et Hydrobiologie
	ARNOUD Paul	Chaire de Chimie M.P.C.
	SENGEL P.	Chaire de Zoologie
	BARNOUD F.	Chaire de Biosynthèse de la Cellulose
	BRISSONNEAU P.	Physique
	GAGNAIRE Didier	Chaire de Chimie Physique

Mme	KOFLER L.	Botanique
MM.	DEGRANGE Charles	Zoologie
	PEBAY-PEROULA J.C.	Physique
	RASSAT A.	Chaire de Chimie Systématique

PROFESSEURS SANS CHAIRE :

MM.	GIDON P.	Géologie et Minéralogie
	GIRAUD P.	Géologie
	PERRET R.	Servomécanismes
Mme	BARBIER M.J.	Electrochimie
Mme	SOUTIF J.	Physique
MM.	COHEN J.	Electrotechnique
	DEPASSEL R.	Mécanique des Fluides
	GASTINEL N.	Mathématiques Appliquées
	ANGLES-d'AURIAC P.	Mécanique des Fluides
	DUCROS P.	Minéralogie et Cristallographie
	GLENAT R.	Chimie
	LACAZE A.	Thermodynamique
	BARRA J.	Mathématiques Appliquées
	COUMES A.	Electronique
	PERRIAUX J.	Géologie et Minéralogie
	ROBERT A.	Chimie Papetière
	BIAREZ J.P.	Mécanique Physique
	BONNET G.	Electronique
	CAUQUIS G.	Chimie Générale
	BONNETAIN L.	Chimie Minérale
	DEPOMMIER P.	Etude Nucléaire et Génie Atomique
	HACQUES Gérard	Calcul Numérique
	POLOUJADOFF M.	Electrotechnique

MAITRES DE CONFERENCES :

MM.	DODU J.	Mécanique des Fluides
	LANCIA Roland	Physique Automatique
Mme	KAHANE J.	Physique
MM.	DEPORTES C.	Chimie
Mme	BOUCHE L.	Mathématiques
MM.	SARROT-RAYNAUD J.	Géologie Propédeutique
Mme	BONNIER M.J.	Chimie
MM.	KAHANE A.	Physique Générale
	DOLIQUE J.M.	Electronique
	BRIERE G.	Physique M.P.C.
	DESPRE P.	Chimie S.P.C.N.
	LAJZEROWICZ J.	Physique M.P.C.
	VALENTIN P.	Physique M.P.C.
	BERTRANDIAS J.P.	Mathématiques Appliquées T.M.P.
	LAURENT P.	Mathématiques Appliquées T.M.P.
	CAUBET J.P.	Mathématiques Pures
	PAYAN J.J.	Mathématiques
Mme	BERTRANDIAS F.	Mathématiques Pures M.P.C.
MM.	LONGEQUEUE J.P.	Physique
	NIVAT M.	Mathématiques Appliquées
	SOHM J.C.	Electrochimie
	ZADWORNY F.	Electronique
	DURAND F.	Chimie Physique
	CARLIER G.	Biologie Végétale
	AUBERT G.	Physique M.P.C.
	DELPUECH J.J.	Chimie Organique
	PFISTER J.C.	Physique C.P.E.M.
	CHIBON P.	Biologie Animale
	IDELMAN S.	Physiologie Animale
	BLOCH D.	Electrotechnique
	BRUGEL L.	I.U.T.
	SIBILLE R.	I.U.T.

Je tiens à exprimer ma profonde reconnaissance,
à Monsieur le Professeur KUNTZMANN, Directeur de l'Institut
de Mathématiques Appliquées de l'Université de Grenoble,
qui a bien voulu me faire l'honneur de présider le Jury,

à Monsieur le Professeur BARRA et à Monsieur LAURENT,
Maître de Conférences, qui m'ont donné l'idée de ce
travail et qui, par leurs conseils et encouragements,
m'ont aidé à le mener à bien.

Je tiens aussi à remercier Mademoiselle BICAIS
qui a apporté le plus grand soin à la réalisation pratique
de cette thèse, ainsi que tous les membres du Laboratoire
de Calcul qui m'ont aidé dans l'élaboration de ce travail.

INTRODUCTION

Les travaux de Stiefel [14] ont montré que le problème d'approximation discret au sens de Tchebycheff relève de la programmation linéaire ; nous avons repris ce problème (chapitre II) et par la simple utilisation des techniques de la programmation linéaire, avons établi les propriétés d'existence, d'unicité et de caractérisation du meilleur approximant et l'algorithme permettant de le calculer.

Nous avons ainsi dégagé des idées plus générales, qui nous ont permis de généraliser ces résultats classiques pour des problèmes d'approximation au sens de Tchebycheff avec contraintes (chapitres III et IV).

Dans la deuxième partie de ce travail, nous avons considéré les problèmes d'approximation continus correspondant aux problèmes précédents. Il est facile de voir que ces problèmes appartiennent à une classe particulière de problèmes de programmation convexe, et pour cette classe, nous avons démontré des conditions nécessaires et suffisantes d'optimalité valables sous des hypothèses assez peu restrictives (chapitre V) ; ces conditions d'optimalité ont déjà été énoncées sans démonstration par Gol'Stein [8].

Nous avons pu ainsi généraliser les résultats théoriques et l'algorithme de résolution numérique (algorithme de Remès [12]) relatifs au problème d'approximation continu au sens de Tchebycheff à des problèmes d'approximation continus avec contraintes (chapitre VII et VIII).

CHAPITRE - I

APPROXIMATION DISCRETE AU SENS DE TCHEBYCHEFF ET PROGRAMMATION LINEAIRE

Ce chapitre est destiné à montrer les liens existant entre les problèmes d'approximation discrète au sens de Tchebycheff et la programmation linéaire.

Dans un premier paragraphe, nous définissons trois problèmes d'approximation discrète au sens de Tchebycheff et formulons dans le deuxième paragraphe les trois problèmes de programmation linéaire correspondants.

Dans le troisième paragraphe enfin, nous donnons quelques rappels sur les polyèdres convexes et la programmation linéaire.

§1 - DEFINITION DES PROBLEMES D'APPROXIMATION CONSIDERES

Notations :

* $\mathrm{Inf}(f(x) \mid x \in D)$ désigne la borne inférieure de la fonction $f(x)$ lorsque x parcourt le domaine D ; plus particulièrement,

$\mathrm{Min}(x^j - y^j \mid j = 1...N)$ désigne le minimum de $x^j - y^j$ quand j parcourt l'ensemble d'indices $(1...N)$.

* On dit qu'une matrice A est (N,n) si elle a N lignes et n colonnes, et on désigne par A_i^j l'élément de la $i^{ème}$ ligne et de la $j^{ième}$ colonne.

1 - Approximation d'un vecteur au sens de Tchebycheff

Soient g_i, $i = 1...n$, n vecteurs donnés linéairement indépendants appartenant à R^N On désigne par :

$$g = \sum_{i=1}^{n} x_i \, g_i$$

un élément de la variété linéaire V engendrée par les g_i.

Soit f un vecteur donné de $\mathbb{R}^N$, $f \notin V$, on cherche un élément (meilleur approximant) $g^* \in V$ (s'il existe) tel que :

$$\text{Max}(|f^j - g^{*j}| \,|\, j = 1...N) = \text{Inf}\left[\text{Max}(|f^j - g^j| \,|\, j = 1...N) \,\middle|\, g \in V\right] = \rho$$

Ce problème est un problème classique que nous désignerons par PI.

2 - Approximation discrète au sens de Tchebycheff avec contrainte du type inégalité.

Soient g_i, $i = 1...n$, φ_i, $i = 1...n$, $2 \times n$ vecteurs donnés appartenant respectivement à $\mathbb{R}^N$ et à $\mathbb{R}^K$.

On suppose les g_i linéairement indépendants, et on désigne par g un élément de la variété linéaire V qu'ils engendrent :

$$g = \sum_{i=1}^{n} x_i \, g_i$$

Etant donné un scalaire strictement positif M, et le vecteur ϕ appartenant à $\mathbb{R}^K$, on définit le sous-ensemble X de $\mathbb{R}^n$:

$$X = \{x \in \mathbb{R}^n : \left| \sum_{i=1}^{n} x_i \, \varphi_i^j - \phi^j \right| \leqslant M \quad j = 1...K\}$$

Soit alors $\bar{V}$, le sous-ensemble de V défini par :

$$\bar{V} = \{g = \sum_{i=1}^{n} x_i \, g_i : x \in X\}$$

Etant donné le vecteur f de $\mathbb{R}^N$, $f \notin V$, on cherche un élément (meilleur approximant) $g^* \in \bar{V}$ (s'il existe) tel que :

$$\text{Max}\,(|f^j - g^{*j}| \,\big|\, j = K+1...K+N) = \text{Inf}\left[\text{Max}\,(|f^j - g^j| \,\big|\, j = K+1...K+N) \,\middle|\, g \in \bar{V}\right] = \rho$$

Nous désignerons par PII ce problème.

Remarque :

Dans la suite, nous considérerons également le problème P'II suivant, cas particulier intéressant du problème PII :

Soient g_i, i = 1...n, n vecteurs donnés appartenant à $\mathbb{R}^{N+K}$, on désignera toujours par g un élément quelconque de la variété linéaire V engendrée par les g_i, supposés linéairement indépendants.

Soit ϕ un vecteur donné de $\mathbb{R}^K$, on désigne par $\bar{V}$ le sous-ensemble de V défini par :

$$\bar{V} = \{g \in V : g^j \leqslant \phi^j \quad j = 1...K\}$$

Etant donné le vecteur f de $\mathbb{R}^N$, $f \notin V$, on cherche un élément (meilleur approximant) $g^* \in \bar{V}$ tel que :

$$\text{Max}\left(|f^j - g^{*j}| \;\middle|\; j = K+1...K+N\right) = \text{Inf}\left[\text{Max}\left(|f^j - g^j| \;\middle|\; j = K+1...K+N\right) \;\middle|\; g \in \bar{V}\right]$$

3 - <u>Approximation de deux vecteurs au sens de Tchebycheff.</u>

Soient donnés n vecteurs g_i, i = 1...n, appartenant à $\mathbb{R}^N$; on désigne par g un élément quelconque de la variété linéaire V engendrée par les g_i, supposés linéairement indépendants.

Soient donnés f_1 et f_2 deux vecteurs de $\mathbb{R}^N$, f_1 et $f_2 \notin V$; sans restriction de généralité, étant donné le problème qu'on a en vue, on suppose :

$$f_1^j \geqslant f_2^j \qquad j = 1...N \qquad (1)$$

(Si deux vecteurs f_1 et f_2 ne vérifient pas (1), on se ramènera à ce cas en considérant le vecteurs φ_1 et φ_2 de $\mathbb{R}^N$ définis par :

$$\varphi_1^j = \text{Max}(f_1^j, f_2^j) \quad j = 1...N$$
$$\varphi_2^j = \text{Min}(f_1^j, f_2^j) \quad j = 1...N.\ )$$

Pour un vecteur $g \in V$, on pose :

$$d(g) = \text{Max}\left[\text{Max}\left(|f_1^j - g^j| \;\middle|\; j = 1...N\right), \text{Max}\left(|f_2^j - g^j| \;\middle|\; j = 1...N\right)\right]$$

On cherche un élément (meilleur approximant) $g^* \in V$ (s'il existe) tel que :

$$d(g^*) = \text{Inf}(d(g) \,|\, g \in V) = \rho$$

Nous désignerons par PIII ce problème.

Remarque :

Ce problème est très important dans la pratique ; il peut être considéré comme une version plus réaliste du problème PI.

En effet, supposons que le vecteur f de $\mathbb{R}^N$ du problème PI soit le résultat d'un certain nombre de mesures ; ces mesures sont toujours l'objet d'une certaine incertitude, et il est plus exact d'approcher la "bande de confiance" définie par les vecteur f_1 et f_2 du problème PIII.

§2 - FORMULATION DES PROBLEMES DE PROGRAMMATION LINEAIRE CORRESPONDANTS

Notation :

On désigne par b le vecteur de $\mathbb{R}^{n+1}$ défini par :

$$\begin{cases} b^j = 0 & j = 1...n \\ b^{n+1} = 1 \end{cases}$$

1 - Problème PI

Notations :

On note a_1 le vecteur de $\mathbb{R}^{2N}$ défini par :

$$\begin{cases} a_1^j = f^j & j = 1...N \\ a_1^j = -f^{j-N} & j = N+1...2N \end{cases}$$

On désigne par :

G la matrice (N,n) de terme général $G_j^i = g_i^j$ $(j = 1...N,\ i = 1...n)$

et par :

I le vecteur de $\mathbb{R}^N$ défini par : $I^j = 1$ $j = 1...N$.

On définit la matrice $A_1(2\times N,\ n+1)$:

$$A_1 = \left(\begin{array}{c|c} G & I \\ \hline -G & I \end{array}\right)$$

Proposition 1. 1

"Le problème PI est équivalent au problème de programmation linéaire, sous forme canonique, suivant :

Minimiser la forme linéaire

$$z = bx$$

sur le polyèdre convexe défini dans $\mathbb{R}^{n+1}$ par :

$$A_1 x \geqslant a_1 \text{"}$$

En effet, en introduisant une variable auxiliaire x_{n+1}, le problème PI peut s'énoncer ainsi

Minimiser x_{n+1} sur le polyèdre convexe QI défini dans $\mathbb{R}^{n+1}$ par :

$$QI \begin{cases} \sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} \geqslant f^j & j = 1 \ldots N \\ -\sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} \geqslant -f^j & j = 1 \ldots N \end{cases}$$

En employant les notations vectorielles définies ci-dessus on obtient la proposition annoncée.

Proposition 1.2.

"Le problème PI a pour problème dual le problème P^*I suivant

Maximiser la forme linéaire

$$z' = y \, a_1$$

sur le polyèdre convexe défini dans $\mathbb{R}^{2\times N}$ par :

$$y A_1 = b \qquad y \geqslant 0 \text{"}$$

Cette proposition s'obtient directement par application des théorèmes classiques de dualité pour des programmes linéaires [2].

On remarque que le problème P^*I est un problème de programmation linéaire sous forme standard.

Remarque :

En explicitant les écritures matricielles, le problème P^*I s'énonce comme suit :

Maximiser la forme linéaire

$$z' = \sum_{j=1}^{N} y_j \, f^j - \sum_{j=1}^{N} y_{N+j} \, f^j$$

sur le polyèdre convexe $Q^{*}I$ défini dans $\mathbb{R}^{2N}$ par :

$$Q^{*}I \quad \begin{cases} \sum_{j=1}^{N} (y_j - y_{N+j})\, g_i^j = 0 & i = 1...n \\ \sum_{j=1}^{2N} y_j = 1 & \\ y_j \geqslant 0 & j = 1...2N \end{cases}$$

2 - Problème PII

Notations :

On définit le vecteur a_2 de $\mathbb{R}^{2K+2N}$

$$\begin{cases} a_2^j = \phi^j - M & j = 1...K \\ a_2^j = -\phi^{j-K} - M & j = K+1...2K \\ a_2^j = f^{j-K} & j = 2K+1...2K+N \\ a_2^j = -f^{j-N-K} & j = 2K+N+1...2K+2N \end{cases}$$

On désigne par :

G la matrice (N,n) de terme général $G_j^i = g_i^{j+K}$ $\quad i = 1...n,\ j = 1...N$

Γ la matrice (K,n) de terme général $\Gamma_j^i = \varphi_i^j$ $\quad i = 1...n,\ j = 1...K$

I le vecteur de $\mathbb{R}^N$ défini par : $I^j = 1$ $\quad j = 1...N$

Θ le vecteur nul de $\mathbb{R}^K$: $\Theta^j = 0$ $\quad j = 1...K$

On définit la matrice $A_2(2K+2N,\ n+1)$:

$$A_2 = \left(\begin{array}{c|c} \Gamma & \Theta \\ \hline -\Gamma & \Theta \\ \hline G & I \\ \hline -G & I \end{array}\right)$$

Proposition 1.3.

"Le problème PII est équivalent au problème de programmation linéaire, sous forme canonique, suivant :

Minimiser la forme linéaire

$$z = bx$$

sur le polyèdre convexe défini dans $\mathbb{R}^{n+1}$ par

$$A_2 x \geqslant a_2 \text{"}$$

En effet, en introduisant une variable auxiliaire x_{n+1}, le problème PII peut s'énoncer ainsi :

Minimiser x_{n+1} sur le polyèdre convexe QII défini dans $\mathbb{R}^{n+1}$ par :

$$\text{QII} \begin{cases} \sum_{i=1}^{n} x_i \varphi_i^j \geqslant \phi^j - M & j = 1 \ldots K \\ -\sum_{i=1}^{n} x_i \varphi_i^j \geqslant -\phi^j - M & j = 1 \ldots K \\ \sum_{i=1}^{n} x_i g_i^j + x_{n+1} \geqslant f^j & j = K+1 \ldots K+N \\ -\sum_{i=1}^{n} x_i g_i^j + x_{n+1} \geqslant -f^j & j = K+1 \ldots K+N \end{cases}$$

En employant les notations matricielles définies ci-dessus on obtient la proposition annonc

Programme linéaire dual.

Le problème P^*II, dual de PII nous est donné par la proposition 1.2, et en explicitant, le problème P^*II s'énonce comme suit :

Maximimer la forme linéaire

$$z' = \sum_{j=1}^{K} y_j (\phi^j - M) + \sum_{j=K+1}^{2K} y_j (-\phi^{j-K} - M) + \sum_{j=2K+1}^{2K+N} y_j f^{j-K} - \sum_{j=2K+N+1}^{2K+2N} y_j f^{j-K-N}$$

sur le polyèdre convexe Q^*II défini dans $\mathbb{R}^{2K+2N}$ par :

$$Q^*II \begin{cases} \sum_{j=1}^{K} (y_j - y_{K+j}) \varphi_i^j + \sum_{j=2K+1}^{2K+N} (y_j - y_{N+j}) g_i^{j-K} = 0 & i = 1 \ldots n \\ \sum_{2K+1}^{2K+2N} y_j = 1 & \\ y_j \geqslant 0 & j = 1 \ldots 2K+2N \end{cases}$$

3. Problème PIII

Notations

On définit le vecteur a_3 de $\mathbb{R}^{2N}$:

$$\begin{cases} a_3^j = f_1^j & j = 1...N \\ a_3^j = -f_2^{j-N} & j = N+1...2\times N \end{cases}$$

On désigne par

G la matrice (N,n) de terme général $G_j^i = g_i^j$ $\quad i = 1...n,\ j = 1...N$

I le vecteur de $\mathbb{R}^N$ défini par : $I^j = 1$ $\quad j = 1...N.$

On définit la matrice A_3 $(2\times N,\ n+1)$:

$$A_3 = \left(\begin{array}{c|c} G & I \\ \hline -G & I \end{array}\right)$$

Proposition 1.4.

"Le problème PIII est équivalent au problème de programmation linéaire, sous forme canonique suivant :

Minimiser la forme linéaire

$$z = bx$$

sur le polyèdre convexe défini dans $\mathbb{R}^{n+1}$ par :

$$A_3 x \geqslant a_3"$$

En introduisant une variable auxiliaire x_{n+1}, le problème PIII peut s'énoncer ainsi

Minimiser x_{n+1} sur le polyèdre convexe QIII défini dans $\mathbb{R}^{n+1}$ par :

$$\text{QIII} \begin{cases} \sum_{i=1}^{n} x_i\, g_i^j + x_{n+1} \geqslant f_i^j & j = 1...N \quad (1) \\ -\sum_{i=1}^{n} x_i\, g_i^j + x_{n+1} \geqslant -f_1^j & j = 1...N \quad (2) \\ \sum_{i=1}^{n} x_i\, g_i^j + x_{n+1} \geqslant f_2^j & j = 1...N \quad (3) \\ -\sum_{i=1}^{n} x_i\, g_i^j + x_{n+1} \geqslant -f_2^j & j = 1...N \quad (4) \end{cases}$$

Comme on a supposé $f_1^j \geq f_2^j \quad j = 1...N$, (1) implique (3) et (4) implique (2) ; en supprimant les inégalités redondantes, et en employant les notations vectorielles définies ci-dessus, on obtient la proposition annoncée.

Programme linéaire dual.

Le problème P^*III, dual du problème PIII nous est donné par la proposition 1-2, et en explicitant, le problème P^*III s'énonce ainsi :

Maximiser la forme linéaire

$$z' = \sum_{j=1}^{N} y_j f_1^j - \sum_{j=1}^{N} y_{N+j} f_2^j$$

sur le polyèdre convexe Q^*III défini dans $\mathbb{R}^{2N}$ par :

$$Q^*III \begin{cases} \sum_{j=1}^{N} (y_j - y_{N+j}) g_i^j = 0 & i = 1...n \\ \sum_{j=1}^{2N} y_j = 1 & \\ y_j \geq 0 & j = 1...2N \end{cases}$$

4. Remarque

Les trois problèmes PI, PII et PIII relevant de la programmation linéaire, il est possible de les résoudre directement par l'algorithme géométrique du simplex. Mais nous nous proposons d'exploiter la structure très particulière des polyèdres convexes QI, QII, QIII pour obtenir dans le cas du problème PI des résultats classiques concernant le meilleur approximant et dans le cas des problèmes PII et PIII des propriétés nouvelles généralisant ces résultats.

Les algorithmes qui en découleront, seront dans le cas du problème PI l'algorithme de Remes [12] Stiefel [14] , dont nous donnerons une interprétation géométrique, et dans le cas des problèmes PII et PIII des généralisations de l'algorithme de Remes Stiefel, que nous interpréterons également géométriquement.

§3 - RAPPELS SUR LES POLYEDRES CONVEXES [3].

a - Définition d'un polyèdre convexe.

Soient f_i, $i = 1...p$, p formes linéaires sur $\mathbb{R}^N$, et a_i, $i = 1...p$, p scalaires, l'ensemble des points M de $\mathbb{R}^N$ définis par :

$$f_i(M) + a_i \geqslant 0 \qquad i = 1...p$$

est un polyèdre convexe de $\mathbb{R}^N$ que nous désignerons par Q.

) - Faces d'un ensemble convexe.

Si P est un ensemble convexe de $\mathbb{R}^N$, on appelle face de P tout sous-ensemble convexe F e P qui satisfait à la condition : quels que soient deux points distincts A et B de P els que l'intervalle ouvert]A,B[rencontre F, [A,B] est contenu dans F.

- Point extrémal d'un polyèdre convexe.

Etant donné un ensemble convexe K, un point M est extrémal de K, s'il n'existe s deux points M_1 et M_2 distincts de K, tels que M puisse se mettre sous la forme :

$$M = \alpha M_1 + \beta M_2$$

avec $\alpha > 0$, $\beta > 0$ et $\alpha + \beta = 1$.

point extrémal est une face de dimension 0.
proposition suivante permet de déterminer analytiquement les points extrémaux du polyè- convexe Q défini ci-dessus.

osition 1-5.

"S'il existe un système d'équations extrait du système

$$(S) : f_i(M) + a_i = 0 \qquad i = 1...p$$

et de rang N, le point M solution unique de ce système soit est un point extrémal de Q, soit n'appartient pas à Q. Réciproquement, tout point extrémal de Q est solution d'un tel système".

d - Arête d'un polyèdre convexe.

On verrait de même, avec les mêmes notations, que si un système d'équations S1, extrait du système (S) est de rang N-1, la droite D solution unique du système S1, est telle que $D \cap Q$ est soit vide, soit une face de Q de dimension 1, c'est-à-dire une arête.

e - Polyèdres convexes de la programmation linéaires.

On considère fréquemment en programmation linéaire, des polyèdres convexes Q du type suivant :

$$Q = \{x \in \mathbb{R}^N : Ax = d,\ x \geqslant 0\}$$

où A est une matrice de type (p,N) sur $\mathbb{R}$ et d un vecteur de $\mathbb{R}^P$. (on suppose $p > N$). Soit $\mathbb{R}^N_+$, l'ensemble des vecteurs de $\mathbb{R}^N$ dont toutes les coordonnées par rapport à une base B sont positives ou nulles. Le domaine de $\mathbb{R}^N$ défini par $Ax = d$ est une variété affine V_A de $\mathbb{R}^N$, on peut donc écrire :

$$Q = R^N_+ \cap V_A$$

R^N_+ étant de codimension nulle, si Q est non vide, Q possède des points extrémaux.

En désignant par r le rang du système $Ax = d$ (on a supposé $p > N$, par conséquent : $r \leq N$), on a la proposition suivante facilitant la recherche des points extrémaux de Q :

Proposition 1-6

"A tout point extrémal M de Q, correspond au moins une partie J de N-r indices de{1...N} telle que M soit solution unique du système :

$$Ax = d \qquad x_j = 0 \qquad j \in J$$

Réciproquement, un tel système définit un point extrémal si et seulement si sa solution est unique et satisfait à $x \geqslant 0$".

Il est clair qu'une telle représentation d'un point extrémal par une partie J peut ne pas être unique ; on est cependant assuré de cette unicité si le point extrémal M a exactement N-r coordonnées nulles. Nous dirons alors que le point M est simple, par opposition au cas où M a plus de N-r coordonnées nulles, cas dans lequel nous dirons que M est multiple.

La proposition 1-7, qui suit, est l'analogue de la précédente mais relative aux arêtes de Q.

Proposition 1-7.

"A toute arête de Q dont le support est la droite D correspond au moins une partie J de N-r-1 indices de {1...N} telle que la droite D soit solution unique du système:

$$Ax = d \qquad x_j = 0 \qquad j \in J$$

Réciproquement un tel système définit une arête de Q si et seulement si sa solution est une droite D telle que :

$$D \cap \mathbb{R}^N_+ \neq \emptyset.$$"

CHAPITRE - II

APPROXIMATION D'UN VECTEUR AU SENS DE TCHEBYCHEFF

Ce problème classique a déjà été rappelé au chapitre I (problème PI), et nous avons déjà formulé le problème de programmation linéaire correspondant.

L'application des méthodes de programmation linéaire au problème PI, nous permet de retrouver dans un premier paragraphe des propriétés connues (unicité, caractérisation de la solution), d'où découle l'algorithme de Remes-Stiefel que nous rappelons au deuxième paragraphe.

Au troisième paragraphe, nous donnons une interprétation géométrique de cet algorithme, et en établissons les avantages par rapport à un algorithme du simplex dans un quatrième paragraphe consacré à l'aspect numérique.

§1 - APPLICATION DES METHODES DE PROGRAMMATION LINEAIRE AU PROBLEME PI

1 - Existence d'une solution

Proposition 2-1.

"Le problème PI a toujours au moins une solution".

Le polyèdre convexe Q^*I intervenant dans le programme linéaire dual de PI est borné et non vide ; donc le programme linéaire dual de PI a une solution finie. D'après les théorèmes de dualité [2] établissant une correspondance entre les solutions de deux problèmes duaux, on sait que PI a une solution finie.

2 - Caractérisation des points extrémaux de QI

Notations.

A un ensemble I de p indices pris dans (1...2×N), on fait correspondre la sous-matrice $A_1(I)$ (p, n+1) de A_1 et le vecteur $a_1(I)$ de $\mathbb{R}^p$ définis de la façon suivante :

soient α_i^j le terme général de A_1, et $\alpha_i^j(I)$ le terme général de $A_1(I)$, on pose :

$$\alpha_i^j(I) = \alpha_i^j \qquad j = 1...n+1,\ i \in I$$

et

$$a_1^j(I) = a_1^j \qquad j \in I$$

Nous dirons que I est <u>régulier</u>, s'il n'existe pas d'indice $i \in I$ vérifiant :

$$(E) \quad \begin{cases} i + N \in I & \text{si } i \leqslant N \\ i - N \in I & \text{si } i > N \end{cases}$$

et nous dirons que I est <u>non régulier d'ordre 1</u> s'il existe un indice $i \in I$ et un seul vérifiant (E).

<u>Proposition 2-2</u>

"<u>Avec l'hypothèse</u>

Hl : "<u>toute sous-matrice (n,n) de G est régulière</u>",

<u>à tout point extrémal x de QI, correspond au moins un ensemble I de n+1 indices pris dans (1...2×N), régulier ou non régulier d'ordre 1, tel que x soit solution unique du système</u> :

$$A_1(I)x = a_1(I)"$$

Il est facile de montrer que l'hypothèse Hl entraîne que A_1 est de rang n+1, et que si I est non régulier d'ordre supérieur à 1, la matrice $A_1(I)$ correspondante est singulière, la proposition 2-2 découle alors directement de la proposition 1-5.

<u>Remarque 1</u>

L'hypothèse Hl est équivalente à la proposition suivante :

V vérifie la condition de Haar [9] sur l'ensemble d'indices (1...N).

<u>Remarque 2</u>

Soit M un point extrémal de QI caractérisé par un ensemble d'indices I non régulier, alors il existe $i_o \in (1...N)$ tel que les coordonnées x_i de M vérifient

$$\begin{cases} \sum_{i=1}^{n} x_i \, g_i^{i_o} + x_{n+1} = f^{i_o} \\ \sum_{i=1}^{n} x_i \, g_i^{i_o} - x_{n+1} = f^{i_o} \end{cases}$$

Un ensemble I de n+1 indices pris dans (1...2N) non régulier d'ordre 1 ne peut donc caractériser qu'un point extrémal M dont la $(n+1)^{\text{ième}}$ coordonnée est nulle.

3 - Unicité de la solution

Proposition 2-3

"L'hypothèse H1 entraîne l'unicité de la solution".

Remarquons tout d'abord qu'une solution ne peut être obtenue qu'en un point extrémal caractérisé par un ensemble d'indices I régulier ; en effet, on a supposé que $f \notin V$ et par conséquent on doit avoir $\rho > 0$.

Nous allons montrer que x_{n+1} ne peut rester constant sur une arête quelconque issue d'un point extrémal M caractérisé par un ensemble régulier I de n+1 indices.

Les coordonnées $x_i(M)$ $(i = 1...n+1)$ sont solution unique de :

$$A_1(I)x(M) = a_1(I)$$

Soit $j_o \in I$ tel que la droite D solution unique (en raison de l'hypothèse H1) du système

$$A_1(I-j_o)x = a_1(I-j_o)$$

est telle que $D \cap QI$ soit une arête de QI issue de M.

Soient $x_i(P)$ $(i = 1...n+1)$ les coordonnées d'un point courant P de D ; les $x_i(P)$ $(i = 1...n)$ s'expriment linéairement et d'une façon unique en fonction de $x_{n+1}(P)$. Soit N un point de $D \cap QI$, distinct de M, si on avait $x_{n+1}(M) = x_{n+1}(N)$, on aurait donc :

$$x_i(N) = x_i(M) \qquad i = 1...n.$$

ce qui impliquerait N = M.

Remarque :

Cette propriété est bien connue sous le nom de théorème de Haar [9], qui montre de plus que l'hypothèse H1 est nécessaire pour assurer l'unicité.

4 - Caractérisation de la solution.

Notations

Soit I un ensemble de n+1 indices caractérisant un point M extrémal de QI ; il lui correspond un ensemble J de n+1 indices (pas nécessairement distincts) de (1...N) :

à un indice $i \in I$ on fait correspondre l'indice $j \in J$ tel que :

$$j = i \quad \text{si } i < N$$
$$j = i-N \quad \text{si } i > N$$

Nous désignerons par J_1 et J_2 des sous-ensembles de J définis par :

$$J_1 = \{j \in J : \sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} = f^j\}$$
$$J_2 = \{j \in J : \sum_{i=1}^{n} x_i \, g_i^j - x_{n+1} = f^j\}$$

(les x_i (i = 1...n+1) étant les coordonnées de M).

On a : $J_1 \cup J_2 = J$

De plus si I est régulier, on a : $J_1 \cap J_2 = \emptyset$.

Par conséquent à un point extrémal M de QI correspond au moins un ensemble J de n+1 indices de (1...N), dont n au moins sont distincts, tel que les coordonnées x_i (i=1...n+1) de M soient solution unique (en faisant l'hypothèse H1) du système :

$$\begin{cases} \sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} = f^j & j \in J_1 \\ \sum_{i=1}^{n} x_i \, g_i^j - x_{n+1} = f^j & j \in J_2 \end{cases}$$

De plus, si M^* est solution, l'ensemble J^* correspondant est tel que :

$$J_1^* \cap J_2^* = \emptyset.$$

Réciproquement, soient donnés un point extrémal M de QI et les ensembles d'indices J, J_1 et J_2, on leur fait correspondre l'ensemble d'indices I de (1...2N) en posant :

$$I = J_1 \cup \{j \in (N+1...2N) : j - N \in J_2\}$$

Si J a n+1 indices distincts, l'ensemble I est régulier

Si J n'a que n indices distincts, I est non régulier d'ordre 1.

Proposition 2-4.

"Avec l'hypothèse H1, une condition nécessaire et suffisante pour qu'un point extrémal M^* de QI soit la solution est qu'il existe n+1 coefficients λ_j tels que en désignant par J^* l'ensemble de n+1 indices distincts de (1...N) correspondant à M^*, on ait :

$$1 - \quad \sum_{j \in J^*} \lambda_j \, g_i^j = 0 \qquad i = 1...n$$

$$2 - \quad \sum_{j \in J^*} |\lambda_j| = 1$$

$$3 - \quad \begin{cases} \lambda_j > 0 & j \in J_1^* \\ \lambda_j < 0 & j \in J_2^* \end{cases}$$

$$4 - \quad \rho = \sum_{j \in J^*} \lambda_j \, f^j \quad "$$

Condition nécessaire

Soient I^* l'ensemble d'indices régulier qui caractérise M^*, et $\overline{I}^*$ son complémentaire dans (1...2N) ; les coordonnées x_i^* de M^* sont solution unique du système :

$$A_1(I^*)\, x^* = a_1(I^*).$$

Au point M^* correspond le point P^* extrémal de Q^*I solution du problème dual de PI, et d'après le théorème de "Complementary Slackness" [2] les coordonnées y_i^* (i = 1...2N) de P^* vérifient :

$$\begin{cases} y_j^* = 0 & j \in \overline{I}^* \\ y_j^* \geqslant 0 & j \in I^* \end{cases}$$

Les $y_j^* \geq 0$ étant solution du système :

$$(S) \quad \begin{cases} \sum_{j \in I^* \cap (1...N)} y_j^* \, g_i^j - \sum_{j \in I^* \cap (N+1...2N)} y_j^* \, g_i^{j-N} = 0 \qquad i = 1...n \\ \sum_{j \in I^*} y_j^* = 1 \end{cases}$$

et vérifiant de plus :

$$\sum_{j \in I^* \cap (1\ldots N)} y_j^* f^j - \sum_{j \in I^* \cap (N+1\ldots 2N)} y_j^* f^{j-N} = \rho$$

on remarque qu'en raison de l'hypothèse H1, les y_j^* solution de (S) sont tous non nuls ; en effet si l'on avait $y_\ell^* = 0$ (les autres y_j^* solution de (S) étant non tous nuls), le système homogène

$$\sum_{\substack{j \in I^* \cap (1\ldots N) \\ j \neq \ell}} y_j^* g_i^j - \sum_{j \in I^* \cap (N+1\ldots 2N)} y_j^* g_i^j = 0 \qquad i = 1\ldots n$$

(on a supposé sans restriction de généralité que $\ell \in (1\ldots N)$) admettrait une solution non nulle, ce qui serait contraire à l'hypothèse H1.

On obtient donc la condition nécessaire en posant :

$$\begin{cases} \lambda_j = y_j^* & j \in J_1^* \\ \lambda_j = -y_{j+N}^* & j \in J_2^* \end{cases}$$

Condition suffisante

Soient donnés le point extrémal M^* de QI, J^* l'ensemble d'indices correspondant et les λ_j associés à J^* vérifiant les conditions 1 - 2 - 3 - 4 de la proposition 2-4.

Le point $P^* \in \mathbb{R}^{2N}$ de coordonnées y_j^* :

$$\begin{cases} y_j^* = \lambda_j & j \in J_1^* \\ y_{j+N}^* = -\lambda_j & j \in J_2^* \\ y_j^* = 0 & j \notin J^* \end{cases}$$

est le point extrémal de Q^*I solution du problème dual de PI, il lui correspond le point M'^* extrémal de QI et solution de PI, qui d'après le théorème de "Complementary Slackness" ne peut être que M^*.

§2 - ALGORITHME DE REMES-STIEFEL

Nous rappelons cet algorithme pour deux raisons, tout d'abord pour en donner une interprétation géométrique au §3, et ensuite parce que les algorithmes de résolution des problèmes PII et PIII en seront une généralisation.

Le principe de la méthode est de déterminer la meilleure approximation sur un ensemble J de n+1 indices, puis on fait évoluer J par permutation de deux indices de façon à converger vers la meilleure approximation g^*.

Dans toute la suite nous faisons l'hypothèse H1.

Notations (Selon Stiefel [14]).

On appelle référence un ensemble J de n+1 indices de (1...N). Un vecteur $g \in V$ est appelé vecteur de référence pour la référence J si, en posant e = f-g on a :

$$\text{Signe } (e^j) = \text{Signe } (\lambda_j) \text{ pour tout } j \in J$$

les λ_j étant des coefficients attachés à J définis de la façon suivante à un facteur multiplicatif positif près :

$$\sum_{j \in J} \lambda_j g_i^j = 0 \qquad i = 1...n.$$

$$\sum_{j \in J} \lambda_j f^j > 0$$

1 - Description de l'algorithme (cf. Laurent [10])

(a) Meilleure approximation sur une référence J.

Etant donné une référence J, on définit la meilleure approximation g_J sur J :

$$\text{Max}(|f^j - g_J^j| \,|\, j \in J) = \text{Inf}\left[\text{Max}(|f^j - g^j| \,|\, j \in J) \,\middle|\, g \in V\right]$$

On détermine g_J en utilisant les conclusions de la proposition 2-4 :

Les conditions

$$\begin{cases} \sum_{j \in J} \lambda_j \, g_i^j = 0 & i = 1 \ldots n \\ \sum_{j \in J} \lambda_j \, f^j > 0 \\ \sum_{j \in J} |\lambda_j| = 1 \end{cases}$$

déterminent les λ_j d'une façon unique.

On calcule alors les coefficients $x_i(J)$ $(i = 1 \ldots n)$ de $g_J = \sum_{i=1}^{n} x_i(J) \, g_i$ et l'erreur correspondante $x_{n+1}(J)$ par résolution du système linéaire :

$$\sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} \times \text{Signe}\,(\lambda_j) = f^j \qquad j \in J$$

(l'hypothèse H1 étant vérifiée, ce système linéaire a une solution unique).
On remarque que g_J est un vecteur de référence pour J.

(b) Itération.

La méthode consiste à remplacer la référence précédente J^i par une nouvelle référence J^{i+1} par substitution d'indices :

$$J^{i+1} = J^i + k - \ell$$

$(\ell \in J^i,\ k \notin J^i,\ k \in (1 \ldots N))$.

de telle façon que :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i).$$

Remarque

Si g_{J^i} meilleure approximation sur J^i, n'est pas la meilleure approximation g^*, on a :

$$x_{n+1}(J^i) < \rho \circ$$

α Détermination de k.

Soit : $e_{J^i} = f - g_{J^i}$, on choisit k tel que :

$$|e^k_{J^i}| = \text{Max}(|e^j_{J^i}| \,|\, j = 1...N)$$

si $|e^k_{J^i}| = x_{n+1}(J^i)$ alors on a terminé et $g_{J^i} = g^*$.

β Détermination de ℓ.

Cette détermination se fait en utilisant la démonstration constructive de la proposition suivante :

Proposition 2-5. (Théorème d'échange de Stiefel)

"Si $\bar{g}$ est un vecteur de référence pour J^o, et si $k \in (1...N)$, $k \notin J^o$, alors il existe $\ell \in J^o$ tel que $\bar{g}$ soit encore de référence pour la nouvelle référence J^1 définie par :

$$J^1 = J^o + k - \ell \quad "$$

Nous désignerons par λ^o_j ($j \in J^o$) les coefficients attachés à J^o, et par λ^1_j ($j \in J^1$) les coefficients attachés à J^1.
Pour que $\bar{g}$ soit de référence pour J^1, il faut et il suffit que, en notant $\bar{e} = f - \bar{g}$, on ait :

$$(R) \quad \begin{cases} \text{Signe } (\bar{e}^j) = \text{signe } (\lambda^1_j) = \text{signe } (\lambda^o_j) & j \in J^o - \ell \\ \text{Signe } (\bar{e}^k) = \text{signe } (\lambda^1_k). \end{cases}$$

Soit j' un indice quelconque de J^o, on définit $J' = J^o + k - j'$, et on calcule des coefficients λ'_j par résolution du système :

$$\begin{cases} \sum_{j \in J'} \lambda'_j \, g^j_i = 0 & i = 1...n \\ \lambda'_k = \varepsilon & (\varepsilon \text{ fixé tel que : signe } (\varepsilon) = \text{signe } (\bar{e}^k)) \end{cases}$$

En convenant de poser $\lambda'_{j'} = 0$, les λ'_j définis pour $j \notin J^o+k$, vérifient :

$$\lambda'_k \, g_i^k + \sum_{j \notin J^o} \lambda'_j \, g_i^j = 0 \qquad i = 1...n \qquad (1)$$

par ailleurs, les λ^o_j, $j \notin J^o$ vérifient :

$$\sum_{j \notin J^o} \lambda^o_j \, g_i^j = 0 \qquad i = 1...n \qquad (2)$$

En désignant par ℓ l'indice sortant inconnu, on peut obtenir (à un facteur près) en combinant les relations (1) et (2) les coefficients λ^1_j associés à $J^1 = J^o + k - \ell$: on multiplie la relation (1) par λ^o_ℓ, la relation (2) par λ'_ℓ, et on retranche :

$$\lambda'_k \; \lambda^o_\ell \, g_i^k + \sum_{j \in J^o} (\lambda^o_\ell \; \lambda'_j - \lambda^o_j \; \lambda'_\ell) \, g_i^j = 0 \qquad i = 1...n$$

Soit :

$$\lambda'_k \, g_i^k + \sum_{\substack{j \in J^o \\ j \neq \ell}} \lambda^o_j \left(\frac{\lambda'_j}{\lambda^o_j} - \frac{\lambda'_\ell}{\lambda^o_\ell}\right) g_i^j = 0 \qquad i = 1...n$$

En choisissant alors ℓ tel que :

$$\frac{\lambda'_\ell}{\lambda^o_\ell} = \text{Min} \left(\frac{\lambda'_j}{\lambda^o_j} \mid j \in J^o \right)$$

les relations (R) seront bien vérifiées.

2 - Convergence de l'algorithme.

Proposition 2-6

"La suite $x_{n+1}(J^i)$ définie par l'algorithme est strictement croissante".

On va montrer que :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i).$$

Tout d'abord, il est facile de voir que :

$$x_{n+1}(J^{i+1}) = \sum_{j \in J^{i+1}} \lambda_j^{i+1} \; f^j.$$

L'égalité ci-dessus peut encore s'écrire :

$$x_{n+1}(J^{i+1}) = \sum_{j \in J^{i+1}} \lambda_j^{i+1} \; (f^j - g^j_{J^i})$$

En vertu du choix de l'indice sortant ℓ (cf. proposition 2-5), g_{J^i} est un vecteur de référence pour J^{i+1}, on a donc :

$$x_{n+1}(J^{i+1}) = x_{n+1}(J^i) + \sum_{j \in J^{i+1}} |\lambda_j^{i+1}| (|e^j_{J^i}| - x_{n+1}(J^i))$$

(en notant $e_{J^i} = f - g_{J^i}$)

$$x_{n+1}(J^{i+1}) = x_{n+1}(J^i) + |\lambda_k^{i+1}| \; (|e^k_{J^i}| - x_{n+1}(J^i))$$

Conséquence

Le procédé itératif ainsi décrit nous permet d'obtenir une suite de références $J^1, J^2, \ldots, J^i, \ldots$ telle que la suite $x_{n+1}(J^i)$ correspondante soit strictement croissante on ne peut donc trouver deux foix la même référence.

Le nombre de références distinctes étant fini, on obtient la solution au bout d'un nombre fini d'itérations.

§3 - INTERPRETATION GEOMETRIQUE DE L'ALGORITHME DE REMES-STIEFEL

Nous allons montrer que l'algorithme correspond à un cheminement sur le polyèdre convexe Q^*I de point extrémal simple en point extrémal simple. Le fait que ces points extrémaux soient simples étant dû à l'hypothèse H1.

L'algorithme de Remes Stiefel procède donc d'une façon très analogue à celle que suivrait un algorithme de simplex résolvant le problème linéaire dual du problème PI ; néanmoins ces deux algorithmes diffèrent analytiquement fortement quant à l'échange plus exactement, étant donnés un point extrémal M de Q^*I et une arête D issue de M, l'algorithme de Remes-Stiefel n'utilise pas la représentation analytique de cette arête pour en déterminer l'extrémité N, point extrémal adjacent à M ; il caractérisera directement cette extrémité grâce au théorème d'échange de Stiefel (proposition 2-5), profitant ainsi de la structure très particulière de Q^*I.

Reprenons les différentes phases de l'algorithme de Remes - Stiefel pour les interpréter :

1 - Meilleure approximation sur une référence J^o

Soient λ_j^o ($j \in J^o$) les coefficients attachés à J^o.

A J^o, nous faisons correspondre un ensemble d'indices I^o régulier de (1...2N) auquel nou faisons correspondre un point M_o de $\mathbb{R}^{2N}$ de la façon suivante :

A un indice $j \in J^o$, on fait correspondre l'indice $i \in I^o$ défini par :

$$\begin{cases} i = j & \text{si } \lambda_j^o > 0 \\ i = N+j & \text{si } \lambda_j^o < 0 \end{cases}$$

A I^o, on fait alors correspondre le point M_o de coordonnées y_j^o ($j = 1...2N$) telles que :

$$\begin{cases} y_j^o \geq 0 & j \in I^o \\ y_j^o = 0 & j \notin I^o \end{cases}$$

les $y_j^o \geq 0$ étant solution du système :

$$y^o(I^o)\ A_1(I^o) = b.$$

Le point M_o ainsi obtenu est point extrémal de Q^*I et comme $y_j^o \neq 0 \quad j \in I^o$ (en raison de l'hypothèse H1), M_o est un point extrémal simple.
Par conséquent les coefficients λ_j^o associés à J^o sont (au signe près) les coordonnées non nulles d'un point extrémal simple de Q^*I.

2 - Itération

a) Détermination de k.

Dans l'algorithme de Remes-Stiefel, ayant déterminé l'indice k, on s'impose le signe de λ_1^k, soit :

$$\lambda_1^k = \varepsilon \qquad (\varepsilon \text{ fixé})$$

A ce choix simultané de k et du signe de λ_k^1, on peut faire correspondre un indice $\bar{k} \in (1...2N)$

$$\bar{k} = k \quad \text{si } \varepsilon > 0$$
$$\bar{k} = k+N \quad \text{si } \varepsilon < 0$$

tel que, M^o étant un point simple, l'ensemble d'indices régulier $I^o + k$ caractérise une arête D_k de Q^*I de sommet M_o.

b) Détermination de ℓ.

Le choix de ℓ est tel que :

$$\text{Signe } (\lambda_j^o) = \text{signe } (\lambda_j^1) \qquad j \in J^o - \ell \qquad (1)$$

A l'ensemble $J^1 = J^o + k - \ell$, on fait correspondre l'ensemble régulier I^1 de (1...2N) tel que, en vertu de (1) :

$$I^1 = I^o + \bar{k} - \bar{\ell}$$

(avec : $\bar{\ell} = \ell$ si $\lambda_\ell^o > 0$

$\bar{\ell} = \ell+N$ si $\lambda_\ell^o < 0$)

I^1 caractérise donc un point extrémal M_1 de Q^*I, adjacent de M_o, situé sur D_k. Mais c'est précisément sur la détermination de $\bar{\ell}$ que l'algorithme de Remes-Stiefel diffère fortement d'un algorithme du simplex qui opèrerait schématiquement de la façon suivante :

Soit N un point quelconque de D_k ; tout point M de D_k peut s'écrire de la façon suivante :

$$M(\Theta) = \Theta M_o + (1 - \Theta)N$$

Θ étant un scalaire.

On cherche Θ_{Max} (s'il existe) tel que :

$$M(\Theta) \in Q^*I \qquad \forall\, \Theta \leq \Theta_{Max}$$

et on déduit ℓ et le point extrémal $M_1 = M(\Theta_{Max})$ adjacent à M_o.

L'algorithme de Stiefel n'utilise pas cette représentation analytique de l'arête, et la technique qu'il lui substitue (proposition 2-5) n'a pas d'interprétation géométrique simple. En effet cette technique utilise une référence intermédiaire $J' = J^o + k - j'$ (j' quelconque de J^o) à laquelle correspond un ensemble régulier I' de (1...2N) ; mais à cet ensemble I', correspond un point M' de R^{2N} qui n'appartient pas nécessairement à l'arête D_k issue de M_o, (à moins d'avoir précisément choisi $j' = \ell$) le signe des λ_j', $j \in J'$ n'étant pas nécessairement le même que celui des λ_j^o, $j \in J^o$.

§4 - ASPECT NUMERIQUE : Avantages de l'algorithme de Remes-Stiefel sur un algorithme du simplex.

Il est bien entendu que nous parlerons ici d'algorithme analytique et non d'algorithme géométrique, puisqu'on a vu au $3^{ième}$ paragraphe que l'algorithme de Remes Stiefel donnait lieu à la même interprétation géométrique que celui d'un algorithme du simplex.

Un programme R-S désignera un programme résolvant le problème PI suivant l'algorithme de Remes-Stiefel.

Remarquons tout d'abord que la résolution du problème PI par un programme du simplex donne lieu à un volume d'encombrement des mémoires très nettement supérieur à celui que nécessite un programme R-S ; dans le premier cas, il faut rentrer en machine la matrice $A_1(2N, n+1)$, dans le second cas, il suffit de rentrer la matrice G (N,n+1) ; or dans la pratique N peut prendre d'assez grandes valeurs de l'ordre de 100.

D'autre part, la résolution de PI par un programme du simplex appliqué au problème dual de PI, ne nous fournit pas en général directement les x_i^* (i = 1...n) coefficients de la meilleure approximation $g^* = \sum_{i=1}^{n} x_i^* g_i$, il faudra donc les calculer à postériori.

Par ailleurs tout programme du simplex suppose la positivité des variables, par conséquent pour pouvoir appliquer un programme du simplex au problème PI (problème primal), on est amené à faire une nouvelle transformation, par exemple le dédoublement des variables x_i, en considérant un vecteur $y \in \mathbb{R}^{2(n+1)}$ tel que :

$$\begin{cases} y_i = x_i & \text{si } x_i \geq 0 \\ y_{i+(n+1)} = -x_i & \text{si } x_i \leq 0 \end{cases}$$

Le seul reproche qu'on pourrait être finalement tenté de faire à un programme R-S est de résoudre un système linéaire pour passer d'un point extrémal M_o à un point extrémal adjacent M_1, alors qu'un programme du simplex évite cette résolution en utilisant la représentation analytique d'une arête. Mais, n étant dans les applications pratiques, de l'ordre de 5, la taille du système linéaire (égale à n+1) est faible, et finalement cette résolution est préférable aux calculs assez lourds auxquels donnent lieu la représentation analytique de l'arête.

CHAPITRE - III

APPROXIMATION DISCRETE AU SENS DE TCHEBYCHEFF AVEC CONTRAINTE DU TYPE INEGALITE

Ce chapitre est consacré à l'étude du problème PII, défini au chapitre I, et dont nous avons déjà formulé le problème de programmation linéaire correspondant.

Le plan de ce chapitre est le même que celui du chapitre II ; l'algorithme établi au deuxième paragraphe est une généralisation simple de l'algorithme de Remes-Stiefel, donnant lieu à la même interprétation géométrique. Dans un troisième paragraphe nous donnons deux procédures Algol (l'une relative au problème PII, l'autre au problème P'II cas particulier de PII) et divers exemples numériques.

§1 - APPLICATION DES METHODES DE PROGRAMMATION LINEAIRE AU PROBLEME PII

1 - Existence d'une solution

Proposition 3-1.

"Si $\overline{V}$ n'est pas vide, alors le problème PII a toujours au moins une solution".

Le polyèdre convexe $Q^{*}II$ intervenant dans le programme linéaire dual de PII est non vide ; donc le programme linéaire dual a une solution finie ou infinie. D'après les théorèmes de dualité [2], le cas où cette solution est infinie correspond au cas où le polyèdre QII est vide.

Donc si l'on suppose QII non vide, ce qui est équivalent à la supposition de non vacuité de $\overline{V}$, le problème PII a une solution finie.

Remarque 1.

La non vacuité de $\overline{V}$ est équivalente à la non vacuité de X :

$$X = \{x \in R^n : \left| \sum_{i=1}^{n} x_i \varphi_i^j - \phi^j \right| \leq M \quad j = 1 \ldots K\}$$

Posons :

$$\bar{\phi} = \text{Max}(\phi^j \mid j = 1...K)$$
$$\underline{\phi} = \text{Min}(\phi^j \mid j = 1...K)$$

S'il existe $\tilde{x} \in \mathbb{R}^n$, tel que $\tilde{\varphi} = \sum_{i=1}^{n} \tilde{x}_i \varphi_i$ ait toutes ses coordonnées $\tilde{\varphi}^j$ $(j = 1...K)$ égales à un même scalaire non nul quelconque, et si de plus $M \geqslant \frac{\bar{\phi} - \underline{\phi}}{2}$, alors il est facile de voir que X n'est pas vide. (il existe un scalaire λ tel que $\lambda \tilde{x} \in X$).

Remarque 2.

Considérons le problème P'II ; dans ce cas particulier, la non vacuité de $\bar{V}$ est assurée dès qu'il existe un vecteur $\tilde{g} \in V$ tel que, ε étant égal à 1 ou a -1 (mais fixé) on ait :

$$\text{Signe } (\tilde{g}^j) = \varepsilon \qquad j = 1...K$$

2 - Caractérisation des points extrémaux de QII

Notations

Nous utiliserons les notations définies en 2-1-2.:

I étant un ensemble de p indices de $(1...2\times K + 2\times N)$ on lui fait correspondre la sous-matrice $A_2(I)$, $(p, n+1)$, de A_2 et le vecteur $a_2(I)$ de $\mathbb{R}^p$.

Nous dirons que I est régulier s'il vérifie les deux propriétés suivantes :

(1) Il n'existe pas d'indice $i \in I$ vérifiant :

$$R1 \quad \begin{cases} i + K \notin I & \text{si} \quad i \leqslant K \\ i - K \in I & \text{si} \quad K < i \leqslant 2K \end{cases}$$

$$R2 \quad \begin{cases} i + N \in I & \text{si} \quad 2K < i \leqslant 2K+N \\ i - N \in I & \text{si} \quad 2K < i \leqslant 2K+2N \end{cases}$$

(2) $I \cap \{2K+1...2K+2N\} \neq \emptyset$

Nous dirons que I est non régulier d'ordre 1 s'il n'existe pas d'indice $i \in I$ vérifiant les relations R1, et s'il existe un indice $i \in I$ et un seul vérifiant les relations R2, et si de plus I vérifie la propriété (2).
Nous désignerons par Ω la matrice (K+N, n) définie par :

$$\Omega = \begin{pmatrix} \Gamma \\ \hline G \end{pmatrix}$$

Proposition 3-2.

"Avec les hypothèses

H1 : "le polyèdre QII est non vide"

H2 : "toute sous matrice (n,n) de Ω est régulière",

à tout point extrémal x de QII correspond au moins en ensemble I de n+1 indices pris dans (1...2K+2N), régulier ou non régulier d'ordre 1, tel que x soit solution unique du système :

$$A_2(I)x = a_2(I) \quad "$$

Cette proposition se démontre de la même façon que la proposition 2-2.

Remarque :

Comme en 2-1-2 on voit qu'un ensemble I de n+1 indices non régulier d'ordre 1 ne peut caractériser qu'un point extrémal dont la $(n+1)^{ième}$ coordonnée est nulle.

3 - Unicité de la solution

Proposition 3-3

"Les hypothèses H1 et H2 entraînent l'unicité de la solution".

Cette proposition se démontre de la même façon que la proposition 2-2.

Remarque :

L'hypothèse H2 peut être considérée comme une condition de Haar généralisée.

4 - Caractérisation de la solution.

Notations :

A un ensemble I de n+1 indices caractérisant un point extrémal M de QI, on fait correspondre un ensemble J de n+1 indices de (1...K+N) (dont n indices seulement sont distincts si I est non régulier d'ordre 1) :

à un indice $i \in I$ on fait correspondre l'indice $j \in J$ tel que :

$$\begin{aligned} j &= i && \text{si} && i \leqslant K \\ j &= i-K && \text{si} && K+1 \leqslant i \leqslant 2K+N \\ j &= i-N-K && \text{si} && 2K+N+1 \leqslant i \leqslant 2K+2N \end{aligned}$$

Nous désignerons par J_1, J_2, J_3 et J_4 des sous-ensembles de J définis par :

$$J_1 = \{j \in J = \sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} = f^j\}$$

$$J_2 = \{j \in J : \sum_{i=1}^{n} x_i \, g_i^j - x_{n+1} = f^j\}$$

$$J_3 = \{j \in J : \sum_{i=1}^{n} x_i \, \varphi_i^j = \phi^j - M\}$$

$$J_4 = \{j \in J : \sum_{i=1}^{n} x_i \, \varphi_i^j = \phi^j + M\}$$

on a : $J = J_1 \cup J_2 \cup J_3 \cup J_4$.

Remarque :

Soit M un point extrémal de QII, I un ensemble de n+1 indices qui le caractérise, l'ensemble J correspondant est nécessairement (cf. proposition 3-2) tel que :

$$J_1 \cup J_2 \neq \emptyset.$$

Proposition 3-4

"Avec les hypothèses H1 et H2, une condition nécessaire et suffisante pour qu'un point extrémal M^* de QII soit la solution est qu'il existe n+1 coefficients λ_j tels que, en désignant par J^* l'ensemble de n+1 indices distincts de (1...K+N) correspondant à M^*, on ait :

$$1 - \quad \sum_{j \in J_1^* \cup J_2^*} \lambda_j \, g_i^j + \sum_{j \in J_3^* \cup J_4^*} \lambda_j \, \varphi_i^j = 0 \qquad i = 1...n$$

$$2 - \quad \sum_{j \in J_1^* \cup J_2^*} |\lambda_j| = 1$$

$$3 - \quad \begin{cases} \lambda_j > 0 & j \in J_1^* \cup J_3^* \\ \lambda_j < 0 & j \in J_2^* \cup J_4^* \end{cases}$$

$$4 - \quad \rho = \sum_{j \in J_1^* \cup J_2^*} \lambda_j \, f^j + \sum_{j \in J_3^*} \lambda_j \, (\phi^j - M) + \sum_{j \in J_4^*} \lambda_j \, (\phi^j + M) \text{ "}$$

Cette proposition se démontre de la même façon que la proposition 2-4.

§2 - ALGORITHME DE RESOLUTION

Cet algorithme est une généralisation simple de l'algorithme de Remes-Stiefel ; le principe de la méthode est de déterminer la meilleure approximation sur un ensemble J de n+1 indices de (1...K+N) et de faire évoluer J par permutation de deux indices, de façon à converger vers la meilleure approximation g^*.
Dans toute la suite nous faisons les hypothèses H1 et H2.

Notations. (selon Stiefel [14])

On appelle référence un ensemble J de n+1 indices de (1...K+N), tel que :

$$J \cap \{K+1...K+N\} \neq \emptyset$$

On note :

$$J_a = J \cap \{K+1...K+N\}$$

$$J_c = J \cap \{1...K\}.$$

et on associe à J, n+1 coefficients λ_j tous non nuls (en raison de l'hypothèse H2) définis à un coefficient multiplicatif positif près par les conditions :

$$\begin{cases} \sum_{j \in J_a} \lambda_j g_i^j + \sum_{j \in J_c} \lambda_j \varphi_i^j = 0 & i = 1 \ldots n \\ \sum_{j \in J_a} \lambda_j f^j + \sum_{j \in J_{c_1}} \lambda_j (\phi^j - M) + \sum_{j \in J_{c_2}} \lambda_j (\phi^j + M) > 0 \end{cases}$$

avec :

$$J_{c_1} = \{j \in J_c : \lambda_j > 0\}$$
$$J_{c_2} = \{j \in J_c : \lambda_j < 0\}$$

Un vecteur $x \in \mathbb{R}^n$ est <u>appelé vecteur de référence pour J</u> si en désignant par e(J) le vecteur de $\mathbb{R}^{N+K}$ défini par :

$$\begin{cases} e^j(J) = \phi^j - \sum_{i=1}^{n} x_i \varphi_i^j & j = 1 \ldots K \\ e^j(J) = f^j - \sum_{i=1}^{n} x_i g_i^j & j = K+1 \ldots K+N \end{cases}$$

on a :

$$\text{signe } (e^j(J)) = \text{signe } (\lambda^j) \qquad j \in J.$$

1 - <u>Description de l'algorithme</u>.

a - <u>Meilleure approximation sur une référence J.</u>

Etant donné une référence J, on définit la meilleure approximation g_J sur J : soient :

$$X(J) = \{x \in \mathbb{R}^n : \left| \sum_{i=1}^{n} x_i \varphi_i^j - \phi^j \right| \leqslant M \qquad j \in J \cap \{1 \ldots K\}\}$$

$$\bar{V}(J) = \{g \in V : x \in X(J)\}$$

on a :

$$\text{Max}(|f^j - g_J^i| \mid j \in J \cap \{K+1 \ldots K+N\}) = \text{Inf}\left[\text{Max}(|f^j - g^j| \mid j \in J \cap \{K+1 \ldots K+N\}) \mid g \in \bar{V}(J\right.$$

On détermine g_J en utilisant les conclusions de la proposition 3-4 :
Les conditions :

$$\begin{cases} \sum_{j \in J_a} \lambda_j g_i^j + \sum_{j \in J_c} \lambda_j \varphi_i^j = 0 & i = 1 \ldots n \\ \sum_{j \in J_a} \lambda_j f^j + \sum_{j \in J_{c_1}} \lambda_j (\phi^j - M) + \sum_{j \in J_{c_2}} \lambda_j (\phi^j + M) > 0 & \\ \sum_{j \in J_a} |\lambda_j| = 1 & \end{cases}$$

déterminent les λ_j d'une façon unique.

On calcule alors les coefficients $x_i(J)$ de $g_J = \sum_{i=1}^{n} x_i(J)\, g_i$ et l'erreur correspondante $x_{n+1}(J)$ par résolution du système linéaire :

$$\begin{cases} \sum_{i=1}^{n} x_i g_i^j + x_{n+1} \times \text{signe}\,(\lambda_j) = f_j & j \in J_a \\ \sum_{i=1}^{n} x_i \varphi_i^j = \phi^j - M & j \in J_{c_1} \\ \sum_{i=1}^{n} x_i \varphi_i^j = \phi^j + M & j \in J_{c_2} \end{cases}$$

On remarque que par construction le vecteur $x(J) \in \mathbb{R}^n$ de coordonnées $x_i(J)$ $(i = 1 \ldots n)$ est de référence pour J.

b - Itération.

La méthode consiste à remplacer la référence précédente J^i par une nouvelle référence J^{i+1} par substitution d'indices :

$$J^{i+1} = J^i + k - \ell$$

$$(\ell \in J^i,\ k \notin J^i,\ k \in (1 \ldots K+N))$$

de telle façon que :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i)$$

Remarque :

Si g_{J^i} meilleure approximation sur J^i n'est pas la meilleure approximation g^*, on a :

$$x_{n+1}(J^i) < \rho .$$

α - Détermination de k.

On pose :

$$P(J^i) = |e^p(J^i)| = \text{Max}\ (|e^j(J^i)|\ |\ j = K+1 \ldots K+N)$$

$$Q(J^i) = e^q(J^i) - M = \text{Max}\ (e^j(J^i) - M | j = 1 \ldots K)$$

$$R(J^i) = e^r(J^i) + M = \text{Min}\ (e^j(J^i) + M | j = 1 \ldots K).$$

et :

$$S(J^i) = \text{Max}(P(J^i) - x_{n+1}(J^i), Q(J^i), -R(J^i)).$$

on choisit :

$k = p$ si $P(J^i) - x_{n+1}(J^i) = S(J^i)$

$k = q$ si $Q(J^i) = S(J^i)$

$k = r$ si $-R(J^i) = S(J^i)$

Remarque

Si $Q(J^i) \leqslant 0$ et $R(J^i) \geqslant 0$, alors $g_{J^i} \in \overline{V}$.
Si $Q(J^i) \leqslant 0$, $R(J^i) \geqslant 0$ et $P(J^i) - {}^{J^i}x_{n+1}(J^i) = 0$
alors on a terminé et $g_{J^i} = g^*$.

β - Détermination de ℓ.

Cette détermination se fait en utilisant la démonstration constructive, analogue à celle de la proposition 2-5, de la proposition suivante :

Proposition 3-5

"Si $\bar{x}$ est un vecteur de référence pour J^o, et si $k \in (1 \ldots K+N)$, $k \notin J^o$, alors il existe $\ell \in J^o$ tel que $\bar{x}$ soit encore de référence pour la nouvelle référence J^1 définie par :

$$J^1 = J^o + k - \ell \ "$$

2 - Convergence de l'algorithme.

Proposition 3-6

"La suite $x_{n+1}(J^i)$ définie par l'algorithme est strictement croissante".

On montre de la même façon qu'en 2-2-2, que :

$$x_{n+1}(J^{i+1}) = x_{n+1}(J^i) + |\lambda_k^{i+1}|\ S(J^i)$$

Par conséquent :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i)$$

Conséquence.

On montre comme en 2-2-2, la convergence du procédé itératif ainsi décrit.

Remarque.

Cet algorithme donne lieu à la même interprétation géométrique que celle qui a été donnée en 2-3 pour l'algorithme de Remes-Stiefel.

§3 - PROCEDURES ALGOL ET ASPECTS NUMERIQUES

1 - Notice d'emploi de la procédure TCHEBCØNT résolvant le problème PII.

TCHEBCØNT(M,N,K,F,G,GP,S,EPS,A,BUL,E,IMPØS).
Etant donné les entiers M,N,K, le réel S, le vecteur F[1:K+N] et les matrices G[1:M, 1:K+N] (précisant la variété linéaire V de dimension M) et GP[1:M, 1:K], on définit les ensembles $\bar{X}$ de $\mathbb{R}^M$ et $\bar{V}$ de V (cf. 1-1-2) :

$$\bar{X} = \{X \in \mathbb{R}^M : \left| \sum_{I=1}^{M} X[I]\times GP[I,J] - F[J] \right| \leq S \qquad J = 1...K \}$$

$$\bar{V} = \{g \in V : g^J = \sum_{I=1}^{M} X[I]\times G[I,J] \ , \ J = K+1...K+N, \ X \in \bar{X} \}$$

La procédure calcule les coefficients A[I], I = 1...M, du meilleur approximant de F dans $\bar{V}$ sur les indices K+1...K+N, et l'écart correspondant A[M+1] :

$$A[M+1] = \text{Max}\left(\left| F[J] - \sum_{I=1}^{M} A[I]\times G[I,J] \right| \ \middle| \ J = K+1...K+N \right)$$

Le tableau E[1:K+N] donne la valeur de l'erreur :

$$\begin{cases} E[J] = F[J] - \sum_{I=1}^{M} A[I]\times GP[I,J] & J = 1...K \\ \\ E[J] = F[J] - \sum_{I=1}^{M} A[I]\times G[I,J] & J = K+1...K+N \end{cases}$$

La procédure comporte deux tests d'arrêt :

1) L'algorithme calcule à chaque itération une minoration d_k et une majoration m_k de l'écart A[M+1], et on arrête les calculs lorsque $(m_k - d_k)/d_k < EPS$; le booléen BUL prend alors la valeur VRAI.

2) Pour éviter le cyclage, dans le cas où EPS aurait été choisit trop petit, la procédure vérifie que les itérations nous conduisent bien à des situations distinctes ; si tel n'était pas le cas, le calcul s'arrête et le booléen BUL prend la valeur FAUX. Ce procédé permet de connaître exactement, en essayant plusieurs valeurs de EPS, la précision de la solution.

A chaque itération, la procédure calcule la meilleure approximation sur un ensemble de M+1 indices choisis parmi les K+N indices, s'il y a non-unicité de la meilleure approximation sur l'un de ces ensembles, on sort de la procédure par l'étiquette IMPØS.

2 - Exemples numériques traités par la procédure TCHEBCØNT

1er exemple

Considérons l'approximation de la fonction sin(x) sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ par un polynôme de degré 4 ; on impose de plus au meilleur approximant $P^*(x) = \sum_{i=0}^{4} a_i x^i$ de vérifier :

$$|\cos(x) - P'^*(x)| \leq S \qquad x \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

on prend S = 0,9, et on choisit 42 points de discrétisation t_j, avec K = N = 21 :

$$\begin{cases} t_j = -\frac{\pi}{2} + (j-1) \times \frac{\pi}{2} & j = 1...21 \\ t_j = -\frac{\pi}{2} + (j-22) \times \frac{\pi}{2} & j = 22...42 \end{cases}$$

On choisit EPS = 10^{-5}

On trouve :

a_0 =	- 0,000 000 0
a_1 =	0,985 531 1
a_2 =	0
a_3 =	- 0,142 566 6
a_4 =	0,000 000 0

En désignant par ρ l'écart correspondant et par E_j (j = 1...42) l'erreur en chaque point de discrétisation

$$\begin{cases} E_j = \cos(t_j) - \sum_{i=0}^{4} a_i \times (t_j)^i & j = 1...21 \\ E_j = \sin(t_j) - \sum_{i=0}^{4} a_i \times (t_j)^i & j = 22...42 \end{cases}$$

On trouve :

$\rho = 0{,}0044889$

j	E_j	j	E_j
1	- 0,6337612	*22	- 0,0044889
2	- 0,5441629	23	0,0027589
3	- 0,4513812	*24	0,0044889
4	- 0,3591732	25	0,0031142
5	- 0,2711085	26	0,0004717
6	- 0,1904817	27	- 0,0021421
7	- 0,1202308	28	- 0,0039214
8	- 0,0628652	*29	- 0,0044889
9	- 0,0204038	30	- 0,0038236
10	+ 0,0056748	31	- 0,0021801
11	0,0144688	32	0,0000000
12	0,0056748	33	0,0021801
13	- 0,0204038	34	0,0038237
14	- 0,0628652	*35	0,0044889
15	- 0,1202308	36	0,0039214
16	- 0,1904817	37	0,0021421
17	- 0,2711085	38	- 0,0004717
18	- 0,3591732	39	- 0,0031142
19	- 0,4513812	*40	- 0,0044889
20	- 0,5441629	41	- 0,0027589
21	- 0,6337612	*42	0,0044889

On a mis une * devant les indices appartenant à la référence finale, et on remarque qu'aucun de ces indices n'appartient à l'ensemble d'indices {1...21} ; on a donc calculé parmi les polynômes de degré 4, le meilleur approximant de sin(x) sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ sans contrainte.

BUL a pris la valeur VRAI.

2ème exemple.

Nous considérons le même exemple que précédemment, mais avec S = 0,6
On trouve :

a_o =	- 0,0000000
a_1 =	1,0214783
a_2 =	- 0,0823609
a_3 =	- 0,1708184
a_4 =	0,0333796

ρ = 0,0575208

BUL = VRAI

j	E_j	j	E_j
*1	- 0,5999999	*22	- 0,0575209
2	- 0,5457689	23	0,0050293
3	- 0,4799742	24	0,0404198
4	- 0,4071491	25	0,0558665
5	- 0,3316395	*26	0,0575208
6	- 0,2575163	27	0,0505066
7	- 0,1884937	28	0,0389693
8	- 0,1278572	29	0,0621378
9	- 0,0784020	30	0,0143970
10	- 0,0423829	31	0,0053688
11	- 0,0214782	32	0,0000000
12	- 0,0167673	33	- 0,0013450
13	- 0,0287231	34	0,0012101
14	- 0,0572200	35	0,0071492
15	- 0,1015556	36	0,0156555
16	- 0,1604872	37	0,0256998
17	- 0,2322817	38	0,0361216
18	- 0,3147773	39	0,0457015
19	- 0,4054558	40	0,0532226
20	- 0,5015236	*41	0,0575209
*21	- 0,5999999	*42	0,0575209

On a mis une * devant les indices appartenant à la référence finale.

Le temps de calcul et d'impression des résultats est pour chacun des deux exemples d'environ 3 secondes (sur IBM-7044)

3 - Notice d'emploi de la procédure TCHEBINEG résolvant le problème P'II

TCHEBINEG(M,N,K,G,F,C,A,E,EPS,BUL).

Etant donné les entiers M,N,K, les vecteurs $C[1:K]$, $F[K+1, K+N]$ et la matrice $G[1:M, 1:K+N]$ (précisant la variété linéaire V de dimension M) on définit les ensembles $\bar{X}$ de $\mathbb{R}^M$ et $\bar{V}$ de V (cf. 1-1-2) :

$$\bar{X} = \{X \in \mathbb{R}^M : \sum_{I=1}^{M} X[I]\times G[I,J] \leq C[J] , \quad J = 1...K\}$$

$$\bar{V} = \{g \in V : g^J = \sum_{I=1}^{M} X[I]\times G[I,J] , \quad J = K+1...K+N, \ X \in \bar{X}\}$$

La procédure calcule les coefficients $A[I]$, $I = 1...M$, du meilleur approximant de F dans $\bar{V}$ et l'écart correspondant $A[M+1]$:

$$A[M+1] = \text{Max}\left(\left|F[J] - \sum_{I=1}^{M} A[I]\times G[I,J]\right| \Big| \ J = K+1...K+N\right)$$

Le vecteur $E[1:K+N]$ donne la valeur de l'erreur :

$$\begin{cases} E[J] = C[J] - \sum_{I=1}^{M} A[I]\times G[I,J] & J = 1...K \\ E[J] = F[J] - \sum_{I=1}^{M} A[I]\times G[I,J] & J = K+1...K+N \end{cases}$$

La procédure comporte deux tests d'arrêt tout à fait identiques à ceux de la procédure TCHEBCØNT.

4 - Exemple numérique traité par la procédure TCHEBINEG.

Considérons l'approximation de la fonction f(x) sur $[3 ; 4]$ par un polynôme de degré 5, f(x) étant définie par :

$$\begin{cases} f(x) = 4x - 12 \text{ pour } x \in [3 ; 3,5] \\ f(x) = -4x + 16 \text{ pour } x \in [3,5 ; 4] \end{cases}$$

On impose de plus au meilleur approximant $\sum_{i=o}^{5} a_i x^i$ de vérifier :

$$\sum_{i=o}^{5} a_i x^i \leqslant c(x) \text{ pour } x \in [0 ; 2] \cup [5 ; 7],$$

c(x) étant définie par :

$$\begin{cases} c(x) = -\sqrt{-x^2+2x} & \text{pour } x \in [0 ; 2] \\ c(x) = 2x^2 - 24x + 70 & \text{pour } x \in [5 ; 7] \end{cases}$$

On choisit 51 points de discrétisation de la façon suivante :

$$\begin{aligned} t_j &= j \times 0,1 && j = 1...20 \\ t_j &= (j-1) \times 0,1+3 && j = 21...40 \\ t_j &= (j-1) \times 0,1-1 && j = 41...51 \end{aligned}$$

on a donc : K = 40 et N = 11.
On trouve :

a_o =	5,26873
a_1 = -	61,20687
a_2 =	42,57065
a_3 = -	9,66299
a_4 =	0,68969
a_5 =	0,00005

En désignant par ρ l'écart correspondant et par E_j (j = 1...51) l'erreur en chaque point de discrétisation

$$E_j = c(t_j) - \sum_{i=o}^{5} a_i \times (t_j)^i \qquad j = 1...40$$

$$E_j = f(t_j) - \sum_{i=o}^{5} a_i \times (t_j)^i \qquad j = 41...51$$

on trouve :

$\rho = 0,23710$

j	E_j	j	E_j
*1	- 0,00005	27	19,16255
2	4,74601	28	19,78874
3	8,80313	29	20,21695
4	12,20348	30	20,40983
5	14,99077	31	20,32842
6	17,21125	32	19,93207
7	18,91131	33	19,17853
8	20,13667	34	18,02378
9	20,93200	35	16,42228
10	21,34072	36	14,32666
11	21,40491	37	11,68809
12	21,16532	38	8,45589
13	20,66139	39	4,57778
14	19,93129	*40	0,00000
15	19,01214	*41	0,23709
16	17,94040	42	- 0,07295
17	16,75266	43	- 0,23295
18	15,48803	*44	- 0,23709
19	14,19720	45	- 0,08124
20	13,12908	*46	0,23710
21	13,12828	47	- 0,08124
22	14,25168	*48	- 0,23709
23	15,36614	49	- 0,23295
24	16,44432	50	- 0,07295
25	17,45716	*51	0,23709
26	18,37402		

On a mis une * devant les indices appartenant à la référence finale.

On a choisit EPS = 10^{-5} et BUL a pris la valeur FAUX.
Le temps de calcul et d'impression des résultats a été de 5 secondes (sur IBM-7044).

```
    PROCEDURE  TCHEBCONT(M,N,K,F,G,GP,S,EPS,A,BUL,E,IMPOS) ;
    VALEUR M,N,K,S,EPS ;
    ENTIER M,N,K ; REEL S,EPS ;
TABLEAU F,G,GP,A,E; BOOLEEN BUL ; ETIQUETTE IMPOS ;

DEBUT

PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
            VOIR  ANNEXE  .....


  TABLEAU ERST[1:M+1,1:M+1] , ZWEI[1:M+1] ;
   ENTIER TABLEAU REF[1:M+1] ;
   TABLEAU PREM[1:M,1:M],MU[1:M+2],
      SEC,LAMBAUX[1:M],LAMBDA[1:M+1] ;
      ENTIER I,J,I0,J0,P0,Q0,R0 ; REEL R,AMPLI ;

INIT: POUR I:=1 PAS 1 JUSQUA M FAIRE REF[I]:=K+I ;
    REF[M+1]:=K+N ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE
    POUR J:=1 PAS 1 JUSQUA M FAIRE
       PREM[I,J]:=G[I,REF[J]] ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE
       SEC[I]:=G[I,REF[M+1]] ;
    GRESOLSYSLINE(PREM,SEC,LAMBAUX,M,SORTIE) ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE LAMBDA[I]:=LAMBAUX[I] ;
    LAMBDA[M+1]:=-1.0 ;
      R:=-F[REF[M+1]] ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE
          R:=R +LAMBDA[I]*F[REF[I]] ;
    SI  R < 0 ALORS
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE LAMBDA[I]:=-LAMBDA[I] ;
     BUL:= VRAI ;

CAL: R:=0.0 ;
       POUR J:=1 PAS 1 JUSQUA M+1 FAIRE
       POUR I:=1 PAS 1 JUSQUA M FAIRE
    ERST[J,I]:= SI REF[J] INFEG K ALORS GP[I,REF[J]] SINON
       G[I,REF[J]] ;
       POUR J:=1 PAS 1 JUSQUA M+1 FAIRE
 ERST[J,M+1]:= SI REF[J] > K ALORS
  ( SI LAMBDA[J] > 0 ALORS 1.0 SINON -1.0 )
                    SINON 0.0 ;
       POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
    SI REF[I] INFEG K ALORS
   ZWEI[I]:= SI LAMBDA[I] > 0 ALORS F[REF[I]]-S SINON
       F[REF[I]]+S SINON
 ZWEI[I]:=F[REF[I]] ;
       GRESOLSYSLINE(ERST,ZWEI,A,M+1,SORTIE) ;
    POUR J:=1 PAS 1 JUSQUA K+N FAIRE
    DEBUT R:=0.0 ;
       POUR I:=1 PAS 1 JUSQUA M FAIRE
    R:= SI J INFEG K ALORS R+A[I]*GP[I,J]
       SINON A[I]*G[I,J]+R ;
       E[J]:=F[J]-R ;
```

```
    FIN ;
    P0:=K+1 ;
    POUR I:=K+2 PAS 1 JUSQUA K+N FAIRE
    P0:= SI ABS(E[I]) > ABS(E[P0]) ALORS I SINON P0 ;
    Q0:=1 ;
    POUR J:=2 PAS 1 JUSQUA K FAIRE
    Q0:= SI E[J]-S > E[Q0]-S ALORS J SINON Q0 ;
    R0:=1 ;
    POUR J:=2 PAS 1 JUSQUA K FAIRE
    R0:= SI E[J]+S < E[R0]+S ALORS J SINON R0 ;
    AMPLI:= SI E[Q0]-S > 0 ALORS ABS(E[P0])+E[Q0]-S
                                    SINON ABS(E[P0]) ;
    AMPLI:= SI E[R0]+S < 0 ALORS AMPLI-E[R0]-S
                                    SINON AMPLI ;
    SI 1-A[M+1]/AMPLI INFEG EPS ALORS ALLERA  END ;
    SI ABS(E[P0])-A[M+1] > E[Q0]-S ALORS
    DEBUT I0:=P0 ; R:=ABS(E[P0])-A[M+1] FIN
    SINON DEBUT I0:=Q0 ; R:=E[Q0]-S FIN ;
    I0:= SI R > -E[R0]-S ALORS I0 SINON R0 ;
    I:=0 ;
ENC: I:=I+1 ;
    SI I0=REF[I] ALORS
      DEBUT
      BUL:= FAUX ;
      ALLERA  END ;
      FIN ;
    SI I < M+1 ALORS ALLERA  ENC ;
    SI I0=P0 ALORS MU[M+2]:= SI E[P0] > 0
      ALORS 1.0 SINON -1.0  SINON
    SI I0=Q0 ALORS MU[M+2]:=1.0 SINON MU[M+2]:=-1.0 ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE
    POUR J:=1 PAS 1 JUSQUA M FAIRE
    PREM[I,J]:= SI REF[J] INFEG K ALORS GP[I,REF[J]] SINON
          G[I,REF[J]] ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE
    SEC[I]:= SI I0 INFEG K ALORS -MU[M+2]*GP[I,I0]
                                SINON -MU[M+2]*G[I,I0] ;
    GRESOLSYSLINE(PREM,SEC,LAMBAUX,M,SORTIE) ;
    POUR I:=1 PAS 1 JUSQUA M FAIRE MU[I]:=LAMBAUX[I] ;
    MU[M+1]:=0.0 ;
    J0:=1 ;
    POUR I:=2 PAS 1 JUSQUA M+1 FAIRE
    J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
       ALORS I SINON J0 ;
    POUR I:=1 PAS 1 JUSQUA J0-1,J0+1 PAS 1 JUSQUA M+1 FAIRE
    LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
    LAMBDA[J0]:=MU[M+2] ;
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
       REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
       ALLERA  CAL ;
    SORTIE: ALLERA  IMPOS ;
    END: FIN TCHEBCONT ;
```

```
   PROCEDURE TCHEBINEG(M,N,K,G,F,C,A,E,EPS,BUL) ;
VALEUR M,N,K,EPS ;  BOOLEEN BUL ;
      ENTIER M,N,K ; REEL EPS ; TABLEAU G,F,C,A,E ;

   DEBUT

PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
          VOIR  ANNEXE  .....


TABLEAU PREM1[1:M,1:M],PREM[1:M+1,1:M+1],SEC1,LAM1[1:M],
 SEC,LAM[1:M+1],MU[1:M+2]; ENTIER TABLEAU REF[1:M+1];
REEL R,AMPLI; ENTIER I,J,I0,J0,G0,L0 ;

   INIT: REF[1]:=K+1 ;
      POUR I:=2 PAS 1 JUSQUA M FAIRE REF[I]:=K+1+I+2 ;
      REF[M+1]:=K+N ;
      BUL:= VRAI ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
      POUR J:=1 PAS 1 JUSQUA M FAIRE
PREM1[I,J]:=G[I,REF[J]] ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
SEC1[I]:=G[I,REF[M+1]]  ;
   GRESOLSYSLINE(PREM1,SEC1,LAM1,M,END)  ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
 LAM[I]:=LAM1[I]  ;
  LAM[M+1]:=-1.0  ;
      R:=-F[REF[M+1]] ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
      R:=R+LAM[I]*F[REF[I]] ;
      SI R < 0 ALORS POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
         LAM[I]:=-LAM[I] ;

   CALREF: POUR J:=1 PAS 1 JUSQUA M+1 FAIRE
      POUR I:=1 PAS 1 JUSQUA M FAIRE
  PREM[J,I]:=G[I,REF[J]]  ;
      POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
 SI   REF[I] INFEG K ALORS
      DEBUT SEC[I]:=C[REF[I]] ; PREM[I,M+1]:=0.0 ;
 FIN  SINON
      DEBUT SEC[I]:=F[REF[I]] ;
      PREM[I,M+1]:=SIGNE(LAM[I]) ; FIN ;
      GRESOLSYSLINE(PREM,SEC,A,M+1,END) ;
POUR J:=1 PAS 1 JUSQUA K+N FAIRE
      DEBUT R:=0.0 ; POUR I:=1 PAS 1 JUSQUA M FAIRE
      R:=R+A[I]*G[I,J] ;
      E[J]:= SI J INFEG K ALORS C[J]-R SINON F[J]-R ;
  FIN ;
L0:=K+1 ;
POUR I:=K+2 PAS 1 JUSQUA K+N FAIRE
L0:= SI ABS(E[I]) > ABS(E[L0]) ALORS I SINON L0 ;
G0:=1 ;
POUR I:=2 PAS 1 JUSQUA K FAIRE
 G0:= SI E[I] < E[G0] ALORS I SINON G0 ;
I0:= SI ABS(E[L0])-A[M+1] > -E[G0] ALORS
```

```
 L0 SINON G0 ;
 AMPLI:= SI E[G0] < 0 ALORS
 ABS(E[L0])=E[G0] SINON ABS(E[L0]);
 SI 1=A[M+1]/AMPLI INFEG EPS ALORS ALLERA  END ;
I:=0 ;
ENC: I:=I+1 ;
SI I0=REF[I] ALORS DEBUT  BUL:= FAUX ; ALLERA  END; FIN ;
SI I < M+1 ALORS ALLERA  ENC ;
      MU[M+2]:= SI E[I0] < 0 ALORS -1.0 SINON 1.0 ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
      POUR J:=1 PAS 1 JUSQUA M FAIRE
PREM1[I,J]:=G[I,REF[J+1]]  ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
  SEC1[I]:=-MU[M+2]*G[I,I0]  ;
   GRESOLSYSLINE(PREM1,SEC1,LAM1,M,END)  ;
      POUR I:=1 PAS 1 JUSQUA M FAIRE
  MU[I+1]:=LAM1[I]  ;
      MU[1]:=0.0 ;
      J0:=1 ;
      POUR I:=2 PAS 1 JUSQUA M+1 FAIRE
      J0:= SI MU[I]/LAM[I] < MU[J0]/LAM[J0] ALORS I
      SINON J0 ;
      POUR I:=1 PAS 1 JUSQUA J0-1 FAIRE
      LAM[I]:=MU[I]-LAM[I]*MU[J0]/LAM[J0] ;
      POUR I:=J0+1 PAS 1 JUSQUA M+1 FAIRE
         LAM[I]:=MU[I]-LAM[I]*MU[J0]/LAM[J0] ;
LAM[J0]:=MU[M+2]  ;
      POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
      REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
      ALLERA  CALREF ;
END:
FIN TCHEBINEG  ;
```

CHAPITRE - IV

APPROXIMATION DE DEUX VECTEURS AU SENS DE TCHEBYCHEFF

Ce chapitre est consacré à l'étude du problème PIII défini au chapitre I, et dont nous avons déjà formulé le problème de programmation linéaire correspondant.

Ce chapitre est présenté suivant le même plan que le chapitre II. L'algorithme établi au deuxième paragraphe est une généralisation de l'algorithme de Remes-Stiefel, faisant appel à une méthode de perturbation ; nous en donnons une interprétation géométrique au troisième paragraphe et montrons ainsi qu'il procède d'une façon analogue à l'algorithme de Wolfe [15].

Une procédure Algol, ainsi que divers exemples numériques sont proposés au quatrième paragraphe.

§1 - APPLICATION DES METHODES DE PROGRAMMATION LINEAIRE AU PROBLEME PIII.

1 - Existence d'une solution

Proposition 4-1

"Le problème PIII a toujours au moins une solution".

Le polyèdre convexe Q^*_{III} intervenant dans le programme linéaire dual de PIII est borné et non vide ; donc le programme linéaire dual de PIII a une solution finie. D'après les théorèmes de dualité [2] établissant une correspondance entre les solutions de deux problèmes duaux, on sait que PIII a une solution finie.

2 - Caractérisation des points extrémaux de QIII

Notations

Nous utiliserons les notations définies en 2-1-2.

Proposition 4-2

"Avec l'hypothèse

H1 :"toute sous-matrice (n,n) de G est régulière"

à tout point extrémal x de QIII, correspond au moins un ensemble I de n+1 indices pris dans (1...2N), régulier ou non régulier d'ordre 1, tel que x soit solution unique du système :

$$A_3(I)x = a_3(I) \qquad "$$

Cette proposition se démontre de la même façon que la proposition 2-2.

Remarque :

Soit M un point extrémal de QIII caractérisé par un ensemble d'indices I non régulier ; alors il existe $i_o \in (1...N)$ tel que les coordonnées x_i de M doivent vérifie

$$\begin{cases} \sum_{i=1}^{n} x_i \, g_i^{i_o} + x_{n+1} = f_1^{i_o} \\ \sum_{i=1}^{n} x_i \, g_i^{i_o} - x_{n+1} = f_2^{i_o} \end{cases}$$

On en déduit :

$$x_{n+1} = \frac{1}{2} (f_1^{i_o} - f_2^{i_o})$$

Donc contrairement à ce qui se passait aux chapitres II et III, un ensemble I de n+1 indices de (1...2N) non régulier d'ordre 1 peut caractériser un point extrémal de QIII, solution de PIII.

3 - Unicité de la solution.

Proposition 4-3

"Les hypothèses

H1

H2 :"Tous les points extrémaux de QIII sont caractérisés par des ensembles de n+1 indices I réguliers"

entraînent l'unicité de la solution".

La démonstration est la même que celle de la proposition 2-3.

Remarque :

Il est facile de voir qu'en supprimant l'hypothèse H2, la proposition 4-3 n'est plus vraie.

Soit M un point extrémal solution caractérisé par un ensemble $I = (i_1, i_2, \ldots i_{n+1})$ non régulier d'ordre 1, supposons par exemple qu'on ait :

$$i_2 = i_1 + N.$$

On a :

$$x_{n+1}(M) = \frac{1}{2}\,(f_1^{i_1} - f_2^{i_1})$$

Soit $\ell \in I$, $\ell \neq i_1$, $\ell \neq i_2$, tel que la droite D solution unique du système :

$$A_3(I-\ell)X = a_3(I-\ell)$$

est telle que $D \cap QIII$ soit une arête de QIII (issue de M).

Soit P un point quelconque de $D \cap QIII$, il est facile de voir que :

$$x_{n+1}(P) = \frac{1}{2}\,(f_1^{i_1} - f_2^{i_1}) = x_{n+1}(M)$$

4 - Caractérisation d'une solution

Notations

Nous utiliserons les notations définies en 2-1-4 ; à un ensemble I de n+1 indices caractérisant un point extrémal de QIII, on fait correspondre l'ensemble J d'indices de (1...N) et on pose :

$$J_1 = \{j \in J : \sum_{i=1}^{n} x_i\, g_i^j + x_{n+1} = f_1^j\}$$

$$J_2 = \{j \in J : \sum_{i=1}^{n} x_i\, g_i^j - x_{n+1} = f_2^j\}$$

Si I est régulier, on a : $J_1 \cap J_2 = \emptyset$, si I est non régulier d'ordre 1, alors il existe un indice $k \in J$ tel que : $J_1 \cap J_2 = \{k\}$.

Proposition 4-4

"Avec l'hypothèse H1, une condition nécessaire et suffisante pour qu'un point extrémal M^* de QIII soit solution est qu'il existe n+1 coefficients λ_j, tels que, en désignant par J^* un ensemble de n+1 indices de (1...N) caractérisant M^*, on ait :

$$1 - \quad \sum_{j \in J^*} \lambda_j \, g_i^j = 0 \qquad i = 1...n$$

$$2 - \quad \sum_{j \in J^*} |\lambda_j| = 1$$

$$3 - \quad \begin{cases} \lambda_j \geq 0 & j \in J_1^* \\ \lambda_j \leq 0 & j \in J_2^* \end{cases}$$

$$4 - \quad \rho = \sum_{j \in J_1^*} \lambda_j \, f_1^j + \sum_{j \in J_2^*} \lambda_j \, f_2^j \quad "$$

Cette proposition se démontre de la même façon que la proposition 2-4.

Remarque 1

Supposons que M^* solution, soit caractérisé par un ensemble d'indices I^* régulier, alors l'ensemble J^* correspondant à n+1 indices distincts, et il est facile de montrer (cf. la démonstration de la proposition 2-4) que les λ_j associés à J^* sont tous non nuls.

Remarque 2

Supposons que M^* solution, soit caractérisé par un ensemble d'indices I^* non régulier d'ordre 1, alors il existe un indice $k \in J^*$ et un seul tel que :

$$J_1^* \cap J_2^* = \{k\}.$$

Les conditions 1, 2 et 4 de la proposition 4-4 déterminent les λ_j d'une façon unique, et il est facile de voir que les λ_j solution sont alors tels que :

$$\begin{cases} \lambda_{k_1} = -\lambda_{k_2} = 1/2 \\ \lambda_j = 0 \qquad j \in J, \; j \neq k \end{cases}$$

k_1 et k_2 désignant l'indice k suivant qu'il appartient respectivement à J_1^* ou à J_2^*.

§2 - ALGORITHMES DE RESOLUTION

Nous donnons successivement deux algorithmes ; dans la première partie, nous exposons un algorithme "restreint", valable dans le cadre des hypothèses H1 et H2, et dont le principe est le même que celui de Remes-Stiefel. Dans la deuxième partie, nous établissons un algorithme généralisé valable dans le cadre de l'hypothèse H1 et de l'hypothèse H3 peu restrictive suivante :

H3 : "il n'existe pas de couple (i,j), i et j $\in$ (1...N) tel que : $f_1^j - f_2^j = f_1^i - f_2^i$"

1 - Algorithme restreint

Dans toute la suite nous faisons les hypothèses H1 et H2.

Notations (selon Stiefel [14])

Etant donné une référence J, ensemble de n+1 indices distincts de (1...N), on lui associe n+1 coefficients λ_j tous non nuls définis de la façon suivante à un facteur multiplicatif positif près :

$$\begin{cases} \sum_{j \in J} \lambda_j \, g_i^j = 0 & i = 1...n \\ \sum_{j \in J_1} \lambda_j \, f_1^j + \sum_{j \in J_2} \lambda_j \, f_2^j > 0 \end{cases}$$

avec :

$$J_1 = \{j \in J : \lambda_j > 0\} \quad \text{et} \quad J_2 = \{j \in J : \lambda_j < 0\}$$

On dit que $g \in V$ est vecteur de référence pour J, si ayant défini les vecteurs $e_1(J)$ et $e_2(J)$ de $\mathbb{R}^N$ par :

$$\begin{cases} e_1^j(J) = f_1^j - g^j & j = 1...N \\ e_2^j(J) = f_2^j - g^j & j = 1...N \end{cases}$$

on a :

$$\begin{cases} \text{signe } (e_1^j(J)) = \text{signe } (\lambda_j) & j \in J_1 \\ \text{signe } (e_2^j(J)) = \text{signe } (\lambda_j) & j \in J_2 \end{cases}$$

a - Description de l'algorithme

α - Meilleure approximation sur une référence J.

Etant donné une référence J, on définit la meilleure approximation g_J sur J : on pose :

$$d_J(g) = \mathrm{Max}\left[\mathrm{Max}(|f_1^j - g^j| \,|\, j \in J),\ \mathrm{Max}(|f_2^j - g^j| \,|\, j \in J)\right]$$

et :

$$d_J(g_J) = \mathrm{Inf}(d_J(g) \,|\, g \in V).$$

On détermine g_J en utilisant les conclusions de la proposition 4-4 : on calcule les coefficients λ_j associés à J par résolution du système :

$$\begin{cases} \sum_{j \in J} \lambda_j \, g_i^j = 0 & i = 1 \ldots n \\ \sum_{j \in J} |\lambda_j| = 1 \\ \sum_{j \in J_1} \lambda_j \, f_1^j + \sum_{j \in J_2} \lambda_j \, f_2^j > 0 \end{cases}$$

et on obtient les coefficients $x_i(J)$ $(i = 1 \ldots n)$ de g_J et l'erreur correspondante $x_{n+1}(J)$ par résolution du système linéaire :

$$\sum_{i=1}^{n} x_i \, g_i^j + x_{n+1} \times \mathrm{signe}\,(\lambda_j) = f^j \qquad j \in J$$

Le vecteur g_J ainsi obtenu est un vecteur de référence pour J.

β - Itération

On remplace la référence précédente J^i par une nouvelle référence J^{i+1} par substitution d'indices :

$$J^{i+1} = J^i + k - \ell$$

$$(\ell \in J^i,\ k \notin J^i,\ k \in (1 \ldots N))$$

de telle façon que :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i).$$

Remarque :

Si g_{J_i} meilleure approximation sur J^i, n'est pas la meilleure approximation g^*, on a :

$$x_{n+1}(J^i) < \rho \circ$$

1 - Choix de k.

On pose :

$$M_1(J^i) = \text{Max } (e_1^j(J^i) \mid j = 1...N) = e_1^p(J^i)$$

$$M_2(J^i) = \text{Min } (e_2^j(J^i) \mid j = 1...N) = e_2^q(J^i)$$

et on prend $k = p$ si $M_1(J^i) > -M_2(J^i)$, $k = q$ dans le cas contraire.
Si $M_1(J^i) \leq x_{n+1}(J^i)$ et $-M_2(J^i) \leq x_{n+1}(J^i)$, alors on a terminé et $g_{J^i} = g^*$

2 - Choix de ℓ.

Cette détermination se fait en utilisant la démonstration constructive, analogue à celle de la proposition 2-5, de la proposition suivante :

Proposition 4-5

"Si g est un vecteur de référence pour J^o, et si $k \in (1...N)$, $k \notin J^o$, alors il existe $\ell \in J^o$ tel que g soit encore de référence pour la nouvelle référence J^1 définie par :

$$J^1 = J^o + k - \ell"$$

b - Convergence de l'algorithme restreint.

Proposition 4-6

"La suite $x_{n+1}(J^i)$ définie par l'algorithme est strictement croissante".

On montre de la même façon qu'en 2-2-2, que :

$$x_{n+1}(J^{i+1}) = x_{n+1}(J^i) + |\lambda_k^{i+1}|(\text{Max}(M_1(J^i), -M_2(J^i)) - x_{n+1}(J^i))$$

Par conséquent , en raison de choix de k :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i)$$

Conséquence :

On montre comme en 2-2-2 la convergence du procédé itératif ainsi décrit.

2 - Algorithme généralisé

Dans toute la suite nous faisons les hypothèses H1 et H3.

Notations :

On appelle référence dégénérée un ensemble J de n+1 indices de (1...N) dont n seulement sont distincts ; sans restriction de généralité, nous noterons :

$$J = (j_1, j_2, \ldots, j_{n+1})$$

avec

$$j_1 = j_2,\ j_1 \in J_1,\ j_2 \in J_2.$$

Nous noterons G^j (j= 1...N) le vecteur de $\mathbb{R}^n$ ayant pour coordonnées :

$$g_1^j \ldots g_n^j.$$

(a) - Principe de la méthode

Ayant supprimé l'hypothèse H2, il est possible qu'au cours des itérations, l'algorithme restreint nous conduise à une référence J^i dégénérée. On sait alors (cf. 4-1-3) que la meilleure approximation g_{J^i} n'est pas unique, et dans ce cas l'algorithme restreint ne nous permet pas de conclure.

L'algorithme généralisé va nous permettre de passer d'une référence dégénérée J^i à une référence J^{i+1} telle que :

$$x_{n+1}(J^{i+1}) > x_{n+1}(J^i)$$

(à moins que g_{J^i} ne soit déjà une meilleure approximation g^*)

Etant donné une référence dégénérée J, les coefficients λ_j associés à J et définis par :

$$\begin{cases} \sum_{j \in J} \lambda_j \, g_i^j = 0 & i = 1...n \\ \sum_{j \in J} |\lambda_j| = 1 \\ \sum_{j \in J_1} \lambda_j \, f_1^j + \sum_{j \in J_2} \lambda_j \, f_2^j > 0 \end{cases}$$

sont tels que :

$$\begin{cases} \lambda_{j_1} = -\lambda_{j_2} = 1/2 \\ \lambda_{j_i} = 0 \qquad i = 3...n+1 \end{cases}$$

Nous utilisons une méthode de perturbation qui à la référence dégénérée J substitue la référence J' non dégénérée déduite de J par le remplacement de j_1 en j_1' tel que :

$$(T) \quad \begin{cases} G^{j_1'} = G^{j_1} - \varepsilon \, (c_{j_3} \, G^{j_3} + \ldots + c_{j_{n+1}} \, G^{j_{n+1}}) \\ f_1^{j_1'} = f_1^{j_1} \end{cases}$$

Les c_{j_i} $(i = 3...n+1)$ sont des constantes non nulles, ε un scalaire positif.
On a donc :

$$\begin{cases} J' = J - j_1 + j_1' \\ \text{avec} \\ J_1' \cap J_2' = \emptyset \end{cases}$$

(cette méthode de perturbation a été utilisée par Descloux [5] pour la résolution du problème de Tchebycheff discret, la condition de Haar n'étant pas remplie).

Les λ_j' associés à J' sont par construction même de J', tels que :

$$\begin{cases} \lambda_{j_i}' = \lambda_{j_i} & i = 1...2 \\ \lambda_{j_i}' = \varepsilon \, \lambda_{j_1} \, c_{j_i} & i = 3...n+1 \end{cases}$$

et l'on a :

$$x_{n+1}(J') = \frac{\sum_{j \in J'_1} \lambda'_j f_1^j + \sum_{j \in J'_2} \lambda'_j f_2^j}{\sum_{j \in J'} |\lambda'_j|}$$

$$e_1^j(J') = x_{n+1}(J') \times \text{signe}\ (\lambda'_j) \qquad j \in J'_1$$

$$e_2^j(J') = x_{n+1}(J') \times \text{signe}\ (\lambda'_j) \qquad j \in J'_2$$

Si $\varepsilon \to 0$, on obtient à la limite :

$$e_1^{j_1}(J) = x_{n+1}(J) \times \text{signe}\ (\lambda_{j_1})$$

$$e_2^{j_2}(J) = x_{n+1}(J) \times \text{signe}\ (\lambda_{j_2})$$

$$e_1^j(J) = x_{n+1}(J) \times \text{signe}\ (\lambda_{j_1}\ c_j) \qquad j \in J_1 - j_1$$

$$e_2^j(J) = x_{n+1}(J) \times \text{signe}\ (\lambda_{j_1}\ c_j) \qquad j \in J_2 - j_2$$

En modifiant le signe des c_j, nous pourrons donc obtenir 2^{n-1} meilleurs approximants sur J.

(b) Description de l'algorithme généralisé

Cet algorithme procède par échange de références et sur chaque référence, on détermine une meilleure approximation. Soient $x_{n+1}(J^i)$ et $x_{n+1}(J^{i+1})$ les erreurs relatives à deux références J^i et J^{i+1}, J^{i+1} provenant d'un échange avec J^i, on a :

$$x_{n+1}(J^{i+1}) \geqslant x_{n+1}(J^i)$$

Si $x_{n+1}(J^{i+1}) = x_{n+1}(J^i)$ on a un échange statique. On appelle étape une suite d'échange statiques se terminant par un échange non statique.

Précisons les différentes phases de l'algorithme :

(I) Début d'étape : on dispose d'une référence

$$J = (j_1, j_2, \cdots j_{n+1})$$

Si J est non dégénérée, on effectue l'échange comme il est exposé dans l'algorithme restreint.

Si J est dégénéré, on choisit arbitrairement n-1 coefficients auxiliaires non nuls :

$$c_{j_3} \ldots c_{j_{n+1}}$$

on calcule

$$x_{n+1}(J) = \frac{1}{2}\,(f_1^{j_1} - f_2^{j_2})$$ et on passe à la situation (II).

(II) On résout le système :

$$\begin{cases} \sum_{i=1}^{n} x_i\, g_i^j = -x_{n+1}(J) + f_1^j & j \in J_1 - j_1 \\ \sum_{i=1}^{n} x_i\, g_i^j = +x_{n+1}(J) + f_2^j & j \subset J_2 \end{cases}$$

(J_1 étant l'ensemble des $j \subset J$ pour lesquels λ_j, $c_j > 0$ et J_2 l'ensemble des $J \in J$ pour lesquels λ_j, $c_j < 0$).
La solution $x_i (i = 1 \ldots n)$ de ce système nous donne un meilleur approximant g_J sur J.

Ayant calculé les coordonnées des vecteurs $e_1(J)$ et $e_2(J)$ associés à g_J, on calcule :

$$M_1(J) = |e_1^q(J)| = \mathrm{Max}(|e_1^j(J)| \,|\, j = 1 \ldots N)$$

$$M_2(J) = |e_2^r(J)| = \mathrm{Max}(|e_2^j(J)| \,|\, j = 1 \ldots N)$$

Si $x_{n+1}(J) = \mathrm{Max}(M_1(J), M_2(J))$, on a une solution du problème, on arrête l'algorithme.
Si $x_{n+1}(J) < \mathrm{Max}(M_1(J), M_2(J))$ alors on choisira l'indice entrant k de telle sorte que :

$$k = q \quad \text{si } M_1(J) > M_2(J)$$

$$k = r \quad \text{si } M_1(J) < M_2(J)$$

Il nous reste à déterminer l'indice sortant ℓ, pour connaître la nouvelle référence J' :

$$J' = J + k - \ell.$$

Pour ce faire, on calcule des coefficients auxiliaires μ_j de la façon suivante : on choisit

$$\mu_k = 1 \text{ si } M_1(J) > M_2(J)$$

$$\mu_k = -1 \text{ si } M_1(J) < M_2(J).$$

et on résoud le système :

$$\sum_{j \in J+k-j_1} \mu_j \, G^j = 0$$

on pose $\mu_{j_1} = 0$, et on calcule :

$$M = \text{Min } \left(\frac{\mu_{j_i}}{c_{j_i}} \middle| i = 3 \ldots n+1\right) \; ; \; N = \text{Min } \left(\frac{\mu_{j_i}}{\lambda_{j_i}} \middle| \; i = 1,2\right)$$

Deux cas peuvent alors se présenter :
1^{er} cas : $M < 0$ on passe à la situation (III) (échange statique)
$2^{ème}$ cas : $M > 0$ on passe à la situation (IV) (échange non statique)

(III) Soit $\quad M = \dfrac{\mu_{j_p}}{c_{j_p}}$

on a, en posant $j_p = \ell$, la nouvelle référence J' :

$$J' = J + k - \ell$$

et les coefficients λ'_j et c'_j relatifs à J' sont tels que :

$$\begin{cases} c'_{j_i} = c_{j_i} - c_\ell \dfrac{\mu_{j_i}}{\mu_\ell} & i = 3 \ldots n+1 \\ & i \neq p \\ c'_{j_p} = -\dfrac{c_\ell}{\mu_\ell} \mu_k & \\ \lambda'_{j_1} = \lambda_{j_1} & \\ \lambda'_{j_2} = \lambda_{j_2} & \end{cases}$$

on a donc $x_{n+1}(J') = x_{n+1}(J)$.
Si l'un des nouveaux coefficients c'_j s'annulle, on le remplace par un nombre pris au hasard.

On retourne à la situation (II).

(IV) Soit $N = \frac{\mu_{j_2}}{\lambda_{j_2}}$, on a, en posant $j_2 = \ell$ la nouvelle référence J' :

$$J' = J + k - \ell$$

et on retourne à la situation (I)

(Si $N = \frac{\mu_{j_1}}{\lambda_{j_1}}$, on posera $j_1 = \ell$).

(c) Justification de l'algorithme généralisé.

Nous nous plaçons en début d'étape avec une référence dégénérée J. Remplaçons j_1 par j_1' et désignons par J' la nouvelle référence ainsi obtenue :

$$G^{j_1'} = G^{j_1} - \varepsilon(c_{j_3} G^{j_3} + \ldots + c_{j_{n+1}} G^{j_{n+1}})$$

et

$$f_1^{j_1'} = f_1^{j_1}$$

on a la relation :

$$(\lambda_{j_1} + O(\varepsilon))\, G^{j_1'} + (\lambda_{j_2} + O(\varepsilon))\, G^{j_2} + \varepsilon \sum_{i=3}^{n+1} c_{j_i} G^{j_i} = 0$$

$O(\varepsilon)$ étant une fonction de ε , telle que :

$$\lim_{\varepsilon \to 0} O(\varepsilon) = 0$$

on a :

$$x_{n+1}(J') = \frac{(\lambda_{j_1} + O(\varepsilon))\, f_1^{j_1'} + (\lambda_{j_2} + O(\varepsilon)) f^{j_2} + \varepsilon \sum_{j \in J_1 - j_1} c_j f_1^j + \varepsilon \sum_{j \in J_2 - j_2} c_j f_2^j}{\sum_{i=1}^{2} |\lambda_{j_i} + O(\varepsilon)| + \varepsilon \sum_{i=3}^{n+1} |c_{j_i}|}$$

ayant déterminé l'indice k entrant, J' étant non dégénérée, on détermine l'indice ℓ sortant en utilisant la proposition 4-5. On calcule donc les coefficients auxiliaires μ_j comme il a été dit dans la description de l'algorithme généralisé, et on pose :

$$Q = \text{Min}\left(\text{Min}\left(\frac{\mu_{j_i}}{\lambda_{j_i} + O(\varepsilon)} \;\middle|\; i = 1,2\right),\ \text{Min}\left(\frac{\mu_{j_i}}{\varepsilon c_{j_i}} \;\middle|\; i = 3 \ldots n+1\right)\right)$$

1er cas

$$M = \text{Min}\left(\frac{\mu_{j_i}}{c_{j_i}} \mid i = 3\ldots n+1\right) = \frac{\mu_{j_p}}{c_{j_p}} < 0.$$

alors il existe ε_o tel que :

$$M = Q \text{ pour tout } \varepsilon < \varepsilon_o$$

En échangeant k et ℓ ($\ell = j_p$) on obtient la nouvelle référence J" :

$$J'' = (j_1', j_2, \ldots j_{p-1}, k, j_{p+1}, \ldots j_{n+1})$$

à laquelle correspond la relation linéaire :

$$(\lambda_{j_1} + O(\varepsilon))G^{j_1'} + (\lambda_{j_2} + O(\varepsilon))G^{j_2} + \varepsilon(c_{j_3} - \mu_{j_3}\frac{c_{j_p}}{\mu_{j_p}})G^{j_3} + \varepsilon\frac{c_{j_p}}{\mu_{j_p}}\mu_k G^k + \ldots$$

$$\ldots + \varepsilon((c_{j_{n+1}} - \mu_{j_{n+1}}\frac{c_{j_p}}{\mu_{j_p}})G^{j_{n+1}} = 0$$

grâce à l'hypothèse H3, les coefficients des G^j ne sont pas nuls, et le cas accidentel où l'un de ces coefficients est nul sera envisagé plus loin.

Nous supposerons donc J" non dégénérée et on a :

$$x_{n+1}(J'') > x_{n+1}(J')$$

On peut recommencer les mêmes calculs avec J". Une suite d'échanges correspondant au 1er cas fournit ainsi une suite de références J', J",...$J^{(r)}$ avec :

$$x_{n+1}(J') < x_{n+1}(J'') < \ldots < x_{n+1}(J^{(r)}),$$

toutes les références de la suite sont donc différentes. On voit qu'il existe ε_r tel que pour $\varepsilon < \varepsilon_r$, les échanges ainsi obtenus sont identiques à ceux décrits dans l'algorithme généralisé, et que les différentes approximations trouvées par l'algorithme sont la limite de celles du problème perturbé lorsque $\varepsilon \to 0$. On vérifie qu'on a des échanges statiques, c'est-à-dire que

$$x_{n+1}(J) = \lim_{\varepsilon \to 0} x_{n+1}(J') = \lim_{\varepsilon \to 0} x_{n+1}(J'') = \ldots$$

2ème cas

M > 0. Alors k est échangé avec l'un des deux indices qui réalise le minimum de :

$$\frac{\mu_{j_i}}{\lambda_{j_i} + O(\varepsilon)} \qquad i = 1,2.$$

On voit alors facilement, que $x_{n+1}(J'')$ (J'' étant la nouvelle référence ainsi obtenue) est strictement supérieure à $x_{n+1}(J')$; et l'échange est non statique.

Cas d'accident (cf. Descloux [5])

Il peut arriver qu'au cours d'une série d'échanges statiques l'un des coefficients auxiliaires s'annule. Soient c_{j_i} ($i = 3...n+1$) les coefficients auxiliaires initiaux, et $c_{j_i}^{(q)}$ ($i = 3...n+1$) les coefficients auxiliaires après q échanges statiques. Les coefficients $c_{j_i}^{(q)}$ ($i = 3...n+1$) peuvent être considérés comme des formes homogènes dans les variables c_{j_i} ($i = 3...n+1$). Montrons par induction qu'elles sont indépendantes.

Nous supposons qu'après q-1 échanges, nous ayons les formes indépendantes

$$c_{j_i}^{(q-1)} \qquad (i = 3...n+1)$$

Appliquons les relations données dans la situation (III) :

$$\begin{cases} c_{j_i}^{(q)} = c_{j_i}^{(q-1)} - c_{\ell}^{(q-1)} \dfrac{\mu_{j_i}}{\mu_{\ell}} & i = 3...n+1, \quad i \neq p. \\ c_{j_p}^{(q)} = - \dfrac{c_{\ell}^{(q-1)}}{\mu_{\ell}} \mu_k \end{cases}$$

Considérons la relation linéaire :

$$\sum_{i=3}^{n+1} \alpha_i \, c_{j_i}^{(q)} = 0$$

En exprimant les $c_j^{(q)}$ en fonction des $c_j^{(q-1)}$, on obtient :

$$\sum_{\substack{i=3 \\ i \neq p}}^{n+1} \alpha_i \, c_{j_i}^{(q-1)} - \frac{1}{\mu_{\ell}} \left(\sum_{\substack{i=3 \\ i \neq p}}^{n+1} \alpha_i \, \mu_{j_i} + \alpha_p \, \mu_k \right) c_{\ell}^{(q-1)} = 0$$

Cette relation n'est possible que si tous les α_i sont nuls, d'où nous concluons à l'indépendance des $c_j^{(q)}$.

Supposons que pour des valeurs particulières des coefficients initiaux, on ait :

$$c_{j_i}^{(q)} \neq 0,\ i = 3\ldots,r-1,\ r+1,\ldots n+1,\ \text{et}\ c_{j_r}^{(q)} = 0$$

Le système d'équations :

$$\begin{cases} c'^{(q)}_{j_i} = c^{(q)}_{j_i} & i = 3\ldots n+1 \qquad i \neq r \\ c'^{(q)}_{j_r} = \eta \end{cases}$$

possède une et une seule solution dans les variables c_j ; pour η assez petit, les q échanges ne sont pas modifiés et l'algorithme peut continuer. Nous avons ainsi démontré qu'il existe des valeurs initiales c_{j_i} (i = 3...n+1) qui permettent d'achever une étape, et ces considérations justifient le terme d'accident.

Pratiquement, on remplace le coefficient nul par un terme "quelconque", ce qui peut être interprété comme le commencement d'une nouvelle étape.

§3 - INTERPRETATION GEOMETRIQUE DE L'ALGORITHME GENERALISE

De la même façon qu'en 2-3, on voit que l'algorithme restreint correspond à un cheminement sur le polyèdre Q^*III de point extrémal simple en point extrémal simple adjacent ; en supprimant l'hypothèse H2, ce cheminement peut nous conduire à un point extrémal multiple M représenté par une partie d'indices I. Nous allons montrer que la méthode de perturbation consiste à associer au polyèdre Q^*III un polyèdre perturbé $Q(\varepsilon)$, défini par les relations (T), et à appliquer l'algorithme restreint sur le polyèdre $Q(\varepsilon)$ à partir du point extrémal simple M' représenté par la partie d'indices I, jusqu'à ce que l'on investisse une partie d'indices $\bar{I}$ telle que la dégénérescence soit levée, c'est-à-dire qu'on a trouvé une arête de Q^*III issue de M avec un accroissement positif du critère.

On voit que cet algorithme généralisé est identique à l'algorithme de Wolfe [15], avec néanmoins la simplification suivante :

L'hypothèse H3 nous assure de ne pas investir un ensemble d'indices $\overline{I}$ tel que $\overline{M}$ soit un point extrémal multiple de $Q(\varepsilon)$, sans accroissement positif du critère ; nous avons donc une "Méthode de Wolfe à un étage". Nous pourrions supprimer l'hypothèse H3 et construire alors un algorithme généralisé tout à fait identique à l'algorithme de Wolfe, mais étant donné le peu de restriction qu'apporte, dans la pratique, l'hypothèse H3, nous préférons la conserver et avoir ainsi un algorithme plus facile à présenter et à programmer.
Considérons maintenant les choses dans le détail :

Soit J^o une référence dégénérée, on note r l'élément commun à J_1^o et J_2^o, et λ_j^o les coefficients attachés à J^o ; on a :

$$\begin{cases} \lambda_{r_1}^o = -\lambda_{r_2}^o = 1/2 \\ \lambda_j^o = 0 \qquad\qquad j \in J^o,\ j \neq r_1,\ j \neq r_2. \end{cases}$$

A J^o on peut faire correspondre 2^{n-1} ensembles I^o de n+1 indices de (1...2×N), non réguliers d'ordre 1, définis par :

à un indice $j \in J^o$ on fait correspondre l'indice $i \in I^o$ défini par :

$$\begin{cases} i = j \quad \text{si} \quad \lambda_j^o > 0 \\ i = N+j \quad \text{si} \quad \lambda_j^o < 0 \\ i = j \text{ ou } i = N+j \text{ si } \lambda_j^o = 0 \end{cases}$$

Ces ensembles I^o caractérisent le même point extrémal M_o de Q^*III, de coordonnées y_j^o telles que :

$$\begin{cases} y_r^o = 1/2 \\ y_{N+r}^o = -1/2 \\ y_j = 0 \qquad\qquad j = 1...2\times N,\ j \neq r,\ j \neq N+r \end{cases}$$

Le point M^o est un point extrémal multiple.
La transformation (T) qui définit la méthode de perturbation utilisée, revient à modifier la $r^{\text{ième}}$ ligne de la matrice A, et en désignant par $A(\varepsilon)$ la matrice ainsi obtenue, on note $Q(\varepsilon)$ le polyèdre convexe défini par :

$$Q(\varepsilon) = \{y \in \mathbb{R}^{2N} : y\,A(\varepsilon) = c,\ y \geq 0\}.$$

L'algorithme généralisé associe à J^o dégénéré, la référence J' non dégénérée, à laquell on peut faire correspondre l'ensemble régulier I' défini par :

à un indice $j \in J'$ on fait correspondre l'indice $i \in I'$ tel que :

$$\begin{cases} i = j & \text{si } \lambda'_j > 0 \\ i = N+j & \text{si } \lambda'_j < 0 \end{cases}$$

(les λ'_j étant les coefficients tous non nuls attachés à J')

I' caractérise un point extrémal simple M' de $Q(\varepsilon)$, et on a :

$$\lim_{\varepsilon \to 0} M' = M_o$$

La méthode consiste alors à appliquer l'algorithme restreint sur le polyèdre perturbé $Q(\varepsilon)$; on obtient ainsi une suite de points extrémaux de $Q(\varepsilon)$ adjacents deux à deux, M', M''..., (ces points sont des points extrémaux simples en raison de l'hypothèse H3), et l'on arrête le processus lorsqu'on rencontre un point $M^{(m)}$ tel que, $I^{(m)}$ désignant un ensemble d'indices qui le caractérise, on ait :

$$\text{soit } r \notin I^{(m)} \quad \text{soit } N+r \notin I^{(m)}$$

(le point $M^{(m)}$ n'est pas nécessairement un point extrémal simple de $Q(\varepsilon)$)

En notant :

$$M_1 = \lim_{\varepsilon \to 0} (M^{(m)}),$$

M_1 est un point extrémal de Q^*III adjacent à M_o.

§4 - PROCEDURE ALGOL ET ASPECTS NUMERIQUES

1 - Notice d'emploi de la procédure TCHEBDEUX FØNC résolvant le problème PIII.

TCHEBDEUXFØNC(M,N,F1,F2,G,A,DEV,EPS,REF,BUL,IMPØS).

Etant donné les entiers M,N, les vecteurs F1[1:N] , F2[1:N], et la matrice G[1:M, 1:N], (l'algorithme suppose F1[J] ⩾ F2[J] pour J = 1...N), la procédure calcule les coefficients A[I], (I = 1...M) du meilleur approximant de F1 et de F2, et l'écart correspondant DEV :

$$\begin{cases} DEV = \max\left(\max_{J=1\ldots N} \left|F1[J] - \sum_{I=1}^{M} A[I]\times G[I,J]\right|,\ \max_{J=1\ldots N} \left|F2[J] - \sum_{I=1}^{M} A[I]\times G[I,J]\right|\right) \\ \\ DEV = \inf_{X\in\mathbb{R}^M} \left\{\max\left(\max_{J=1\ldots N} \left|F1[J] - \sum_{I=1}^{M} X[I]\times G[I,J]\right|,\ \max_{J=1\ldots N} \left|F2[J] - \sum_{I=1}^{M} X[I]\times G[I,J]\right|\right)\right\} \end{cases}$$

La procédure comporte deux tests d'arrêt :

1) L'algorithme calcule à chaque itération une minoration d_k et une majoration m_k de l'écart DEV, et on arrête les calculs lorsque $(m_k - d_k)/d_k < EPS$; le booléen BUL prend alors la valeur VRAI.

2) Pour éviter le cyclage, dans le cas où EPS aurait été choisi trop petit, la procédure vérifie que les itérations nous conduisent bien à des situations distinctes, si tel n'était pas le cas, le calcul s'arrête et le booléen BUL prend la valeur FAUX.

Si l'algorithme rencontre au cours des itérations, une référence REF_k telle qu'il existe $\alpha,\beta,\gamma,\delta \in REF_k$ tels que :

$$\alpha = \beta$$
$$\gamma = \delta$$

alors on sort de la procédure par l'étiquette IMPØS. (c'est le cas où l'hypothèse H3 de la partie théorique n'est pas réalisée).
Enfin, la procédure range dans le tableau REF[1:M+1] les indices de la dernière référence.

2 - Exemples numériques.

1er exemple

Considérons l'approximation des fonctions $f_1(x) = \exp(x)$ et $f_2(x) = \sqrt{x}$ sur $[0,2]$ par un polynome de degré 4.

On choisit 21 points de discrétisation t_j :

$$t_j = (j-1)\times 0{,}1 \qquad j = 1\ldots 21.$$

Soient a_i, $i = 0...4$, les coefficients du meilleur approximant, on note :

$$\begin{cases} E_j^1 = \sin(t_j) - \sum_{i=0}^{4} a_i \times (t_j)^i & j = 1...21 \\ E_j^2 = \sqrt{(t_j)} - \sum_{i=0}^{4} a_i \times (t_j)^i & j = 1...21 \end{cases}$$

On trouve, pour une valeur de EPS égale à 10^{-5}

a_0 =	-1,52277
a_1 =	-6,44797
a_2 =	31,02942
a_3 =	-25,59697
a_4 =	6,21740

DEV = 2,98742
BUL = VRAI

I	J = REF[I]	E^1_J	E^2_J
1	21	2,98742	-2,98742
2	21	2,98742	-2,98742
3	3	2,98742	2,21323
4	2	2,98742	2,19848
5	13	-0,76274	-2,98741
6	20	2,98742	-2,32006

On constate que l'indice 21 figure deux fois dans la référence finale, et par conséquent, il n'y a pas unicité de la solution.
Le temps de calcul et d'impression des résultats a été de 3 secondes (sur IBM 7044).

2ème exemple

Considérons l'approximation des fonctions $f_1(x) = \sin(\pi x) + 0,1 \times \cos(\pi x/2)$ et $f_2(x) = \sin(\pi x) - 0,1 \times \cos(\pi x/2)$, sur $[-1,1]$ par un polynôme de degré 4.

On choisit 21 points de discrétisation t_j :

$$t_j = -1 + (j-0,5)/11$$

Soient a_i, i = 0...4, les coefficients du meilleur approximant, on note :

$$E_j^1 = \sin(\pi t_j) + 0,1 \times \cos(\pi t_j/2) \qquad j = 1...21$$

$$E_j^2 = \sin(\pi t_j) - 0,1 \times \cos(\pi t_j/2) \qquad j = 1...21$$

On trouve pour une valeur de EPS égale à 10^{-5} :

a_o =	-0,00668
a_1 =	2,88101
a_2 =	0,14993
a_3 =	-3,28751
a_4 =	-0,30836

DEV = 0,13260
BUL = FAUX

I	J=REF[I]	E_J^1	E_J^2
1	21	0,13261	0,09009
2	1	-0,11833	-0,13260
3	4	0,13260	0,03675
4	9	0,05478	-0,13260
5	13	0,13260	-0,06282
6	18	-0,01275	-0,13260

On constate qu'aucun indice ne figure deux fois dans la référence finale, et par conséquent la solution obtenue est l'unique solution du problème.

Le temps de calcul et d'impression des résultats a été de 3 secondes environ. (sur IBM 7044).

```
PROCEDURE TCHEBDEUXFONC(M,N,F1,F2,G,A,DEV,EPS,REF,BUL,IMPOS) ;
   ENTIER M,N ; TABLEAU F1,F2,G,A ; REEL EPS,DEV ;
   ENTIER TABLEAU REF ;
  BOOLEEN BUL ;
   ETIQUETTE IMPOS ;

   DEBUT

    REEL PROCEDURE HASARD(NO) ; ENTIER NO ;
      COMMENTAIRE  CETTE PROCEDURE DONNE UN NOMRE REEL ALEATOIRE
      DE (0,1) AVEC UNE DENSITE UNIFORME.
      NO SERA INITIALISE A UNE VALEUR ENTIERE IMPAIRE (UNE FOIS
      POUR TOUTE),CHAQUE APPEL DE HASARD(NO) DEFINIT UN REEL DE
      (0,1) ET PROVOQUE UN EFFET DE BORD SUR NO .
      LE CORPS DE CETTE PROCEDURE EST ECRIT EN CODE .


'PROCEDURE'GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
           VOIR  ANNEXE  .....


ENTIER TABLEAU ROC[1:M+1] ;
   TABLEAU PREM[1:M,1:M],MU[1:M+2],E1,E2[1:N],SEC,
   LAMBAUX[1:M],LAMBDA[1:M+1] ;
REEL R,T ; BOOLEEN BOOL,DEG ;
      ENTIER I,J,I0,J0,N0,JDOU ;

POUR I:=1 PAS 1 JUSQUA M FAIRE
REF[I]:=I ;
      REF[M+1]:=N ;
   N0:=3 ;
   BUL:= VRAI ;

ETAPE:
      POUR I:=1 PAS 1 JUSQUA M FAIRE
   POUR J:=1 PAS 1 JUSQUA M FAIRE
   PREM[I,J]:=G[I,REF[J]] ;
POUR I:=1 PAS 1 JUSQUA M FAIRE
 SEC[I]:=G[I,REF[M+1]] ;
   GRESOLSYSLINE(PREM,SEC,LAMBAUX,M,END) ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   LAMBDA[I]:=LAMBAUX[I] ;
   LAMBDA[M+1]:=-1.0 ;
   DEV:=-F2[REF[M+1]] ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*
      F1[REF[I]] SINON DEV+LAMBDA[I]*F2[REF[I]] ;
SI DEV < 0 ALORS   DEBUT
   POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
      LAMBDA[I]:=-LAMBDA[I] ;
DEV:=F1[REF[M+1]] ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*
      F1[REF[I]] SINON DEV+LAMBDA[I]*F2[REF[I]] ;
FIN ;
```

```
R:=0.0;
   POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
 R:=R+ABS(LAMBDA[I]) ;
 DEV:=DEV/R ;
DEG:= FAUX ;

CAL: POUR J:=1 PAS 1 JUSQUA M FAIRE
   POUR I:=1 PAS 1 JUSQUA M FAIRE
      PREM[J,I]:=G[I,REF[J+1]] ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   SEC[I]:= SI LAMBDA[I+1] > 0 ALORS
      F1[REF[I+1]]-DEV SINON F2[REF[I+1]]+DEV ;
   GRESOLSYSLINE(PREM,SEC,A,M,END) ;
   POUR J:=1 PAS 1 JUSQUA N FAIRE
      DEBUT R:=0.0 ;
 POUR I:=1 PAS 1 JUSQUA M FAIRE
      R:=R+A[I]*G[I,J] ;
      E1[J]:=F1[J]-R ; E2[J]:=F2[J]-R ;
   FIN ;
   I0:=1 ;
   POUR I:=2 PAS 1 JUSQUA N FAIRE
 I0:= SI E1[I] > E1[I0] ALORS I SINON I0 ;
   J0:=1 ;
   POUR I:=2 PAS 1 JUSQUA N FAIRE
  J0:= SI E2[I]  < E2[J0] ALORS I SINON J0 ;
   SI ABS(E1[I0]) > ABS(E2[J0]) ALORS BOOL:= VRAI SINON
BOOL:= FAUX ;
SI BOOL ET ABS(1-DEV/ABS(E1[I0])) < EPS ALORS
ALLERA  END;
SI NON BOOL ET ABS(1-DEV/ABS(E2[J0])) < EPS ALORS
ALLERA  END;
I0:= SI BOOL ALORS I0 SINON J0 ;
   I:=0 ;
   SU:  I:=I+1 ;
  SI BOOL ET I0=REF[I] ET ABS(E1[I0]-DEV) < 0.001 ALORS
  DEBUT BUL:= FAUX ; ALLERA  END  FIN ;
 SI NON BOOL ET I0=REF[I] ET
  ABS(-E2[I0]-DEV) < 0.001 ALORS
  DEBUT BUL:= FAUX ; ALLERA  END  FIN ;
  SI I < M+1 ALORS ALLERA  SU ;
   I:=0 ;
   ITER: I:=I+1 ;
   SI I0=REF[I] ET DEG ALORS DEBUT JDOU:=I ;
SI ABS(0.5*(F1[I0]-F2[I0])-DEV) < EPS
ALORS ALLERA  IMPOS SINON ALLERA  SAUT ;  FIN ;
   SI I0=REF[I] ALORS DEBUT JDOU:=I ; DEG:= VRAI ;
   ALLERA  DIF FIN ;
   SI I < M+1 ALORS ALLERA  ITER ;
CRA: MU[M+2]:= SI BOOL ALORS 1.0 SINON -1.0 ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   POUR J:=1 PAS 1 JUSQUA M FAIRE
   PREM[I,J]:=G[I,REF[J+1]] ;
   POUR I:=1 PAS 1 JUSQUA M FAIRE
   SEC[I]:=-MU[M+2]*G[I,I0] ;
   GRESOLSYSLINE(PREM,SEC,LAMBAUX,M,END) ;
```

```
    POUR I:=1 PAS 1 JUSQUA M FAIRE MU[I+1]:=LAMBAUX[I] ;
    MU[1]:=0.0 ;
    SI NON DEG ALORS ALLERA  NORMAL ;
    J0:=3 ;
    POUR I:=4 PAS 1 JUSQUA M+1 FAIRE
    J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
       ALORS I SINON J0 ;
    T:=MU[J0]/LAMBDA[J0] ;
SI T > 0 ALORS  DEBUT
       DEG:= FAUX ;
       J0:= SI MU[2]/LAMBDA[2] < 0 ALORS 2 SINON 1 ;
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
ALLERA  ETAPE;
          FIN ;
    POUR I:=3 PAS 1 JUSQUA J0-1 FAIRE
    DEBUT LAMBDA[I]:=LAMBDA[I]-LAMBDA[J0]*MU[I]/MU[J0] ;
SI ABS(LAMBDA[I]) < EPS ALORS DEBUT
B1: LAMBDA[I]:=HASARD(N0)*(-1)**I ;
SI ABS(LAMBDA[I]) < EPS ALORS ALLERA  B1 ;
FIN ;
    FIN ;
    POUR I:=J0+1 PAS 1 JUSQUA M+1 FAIRE
    DEBUT LAMBDA[I]:=LAMBDA[I]-LAMBDA[J0]*MU[I]/MU[J0] ;
SI ABS(LAMBDA[I]) < EPS ALORS DEBUT
B2: LAMBDA[I]:=HASARD(N0)*(-1)**I ;
SI ABS(LAMBDA[I]) < EPS ALORS ALLERA   B2 ;
FIN ;
    FIN ;
    LAMBDA[J0]:=-LAMBDA[J0]*MU[M+2]/MU[J0] ;
    POUR I:=3 PAS 1 JUSQUA M+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
    ALLERA  CAL ;
    NORMAL: J0:=1 ;
    POUR I:=2 PAS 1 JUSQUA M+1 FAIRE
    J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
    ALORS I SINON J0 ;
RESTAU: POUR I:=1 PAS 1 JUSQUA J0-1 FAIRE
    LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
    POUR I:=J0+1 PAS 1 JUSQUA M+1 FAIRE
    LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
    LAMBDA[J0]:=MU[M+2] ;
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
    DEV:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
    DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*
       F1[REF[I]] SINON
       DEV+LAMBDA[I]*F2[REF[I]] ;
R:=0.0;
    POUR I:=1 PAS 1 JUSQUA M+1 FAIRE
 R:=R+ABS(LAMBDA[I]) ;
 DEV:=DEV/R ;
   ALLERA  CAL ;
```

```
DIF:I:=2 ;
POUR I:=I+1 TANTQUE  I ≤ JDOU+1 ET I ≤ M+1 FAIRE
ROC[I]:=REF[I-2] ;
POUR J:=I PAS 1 JUSQUA M+1 FAIRE
 ROC[J]:=REF[J-1] ;
      ALLERA   OUF ;

   SAUT:
    POUR I:=3 PAS 1 JUSQUA JDOU FAIRE
       ROC[I]:=REF[I-1] ;
     POUR I:=JDOU+1 PAS 1 JUSQUA M+1 FAIRE
       ROC[I]:= REF[I] ;

  OUF:
 POUR I:=3 PAS 1 JUSQUA M+1 FAIRE
REF[I]:=ROC[I] ;
REF[1]:=10 ; REF[2]:=10;
   LAMBDA[1]:=0.5 ; LAMBDA[2]:=-0.5 ;
   POUR I:=3 PAS 1 JUSQUA M+1 FAIRE
DEBUT
   LAMBDA[I]:=HASARD(N0)*(-1)**I ;
SI ABS(LAMBDA[I]) ≤ EPS ALORS DEBUT
E3: LAMBDA[I]:=HASARD(N0)*(-1)**I ;
SI ABS(LAMBDA[I]) ≤ EPS ALORS ALLERA   E3 ;
FIN ;
FIN   ;
   DEV:=0.5*(F1[10]-F2[10]) ;
   ALLERA  CAL ;
END: FIN TCHEBDEUXFONC ;
```

CHAPITRE - V

APPROXIMATION UNIFORME ET PROGRAMMATION CONVEXE

Nous étudions dans ce chapitre les liens existant entre les problèmes d'approximation uniforme et la programmation convexe.

Dans un premier paragraphe, nous donnons la formulation continue des problèmes PI, PII, PIII ; et montrons dans le deuxième paragraphe, que ces problèmes sont des cas particuliers d'un problème P de programmation convexe consistant à minimiser une forme linéaire sur un domaine convexe d'une nature assez spéciale (cf. 5-3-1).

Dans un troisième paragraphe, nous établissons des conditions nécessaires et suffisantes d'optimalité pour le problème P ; des conditions presque identiques ont déjà été énoncées sans démonstration par Gol'stein [8].

Les conditions d'optimalité, que nous démontrons par la technique très générale des "déplacements intérieurs" [1], [11], coïncident avec les conditions de Kühn et Tücker généralisées établies par Russel [13] ; mais étant donné les particularités du problème P, ces conditions sont valables sous des hypothèses moins restrictives et surtout plus faciles à vérifier dans la pratique que celles faites par Russel.

§1 - FORMULATION CONTINUE DES PROBLEMES PI, PII et PIII.

1 - Approximation uniforme d'une fonction continue sur un compact.

Soit S un compact d'un espace métrique, on désigne par $\mathcal{C}(S)$ l'ensemble des fonctions continues à valeur réelle sur S, et on pose

$$||h|| = \text{Max}\ (|h(t)|\ \big|\ t \in S)$$

pour un élément $h \in \mathcal{C}(S)$.

Soient g_i, $i = 1...n$, n éléments donnés linéairement indépendants, appartenant à $\mathcal{C}(S)$, on désigne par

$$g = \sum_{i=1}^{n} x_i \, g_i$$

un élément de la variété linéaire V engendrée par les g_i.
Etant donné un élément $f \in \mathcal{C}(S)$, $f \notin V$, on cherche un élément $g^* \in V$, meilleur approximant, (s'il existe) tel que :

$$\|f - g^*\| = \text{Inf} \; (\|f-g\| \;\big|\; g \in V) = \rho$$

Ce problème est un problème classique que nous désignerons par PIC, c'est la version continue du problème PI.

2 - Approximation uniforme d'une fonction continue sur un compact avec contrainte du type inégalité.

Soient K et H deux compacts quelconques d'un même espace métrique ; pour un élément $h \in \mathcal{C}(K)$ on note :

$$\|h\| = \text{Max} \; (|h(t)| \;\big|\; t \in K).$$

Soient donnés n élément g_i, $i = 1...n$, linéairement indépendants, appartenant à $\mathcal{C}(K)$; on désigne par :

$$g = \sum_{i=1}^{n} x_i \, g_i$$

un élément de la variété linéaire V engendrée par les g_i.

Etant donnés par ailleurs, $_i$, $i = 1...n$, et ϕ, n+1 éléments appartenant à (H), et un scalaire M strictement positif, on définit :

$$X = \{x \in \mathbb{R}^n : \left| \sum_{i=1}^{n} x_i \, \varphi_i(t) - \phi(t) \right| \leq M \qquad \forall \; t \in H\}$$

Soit alors $\overline{V}$ le sous-ensemble de V défini par :

$$\overline{V} = \{g = \sum_{i=1}^{n} x_i \, g_i \; : \; x \in X\}$$

Etant donné $f \in \mathcal{C}(K)$, $f \notin V$, on cherche un élément $g^* \in \overline{V}$, meilleur approximant, (s'il existe) tel que :

$$\|f - g^*\| = \text{Inf } (\|f-g\| \;\big|\; g \in \overline{V}) = \rho$$

Nous désignerons par PIIC ce problème, version continue du problème PII.

Remarque

Dans la suite, nous considérerons également le problème P'IIC, suivant, cas particulier intéressant du problème PIIC, version continue du problème P'II :

Soient K et H deux compacts disjoints d'un espace métrique ; on pose $S = K \cup H$, et pour un élément $h \in \mathcal{C}(K)$ on note :

$$\|h\| = \text{Max } (|h(t)| \;\big|\; t \in K).$$

Soient g_i, $i = 1...n$, n éléments donnés appartenant à $\mathcal{C}(S)$ et linéairement indépendants sur K ; on désignera toujours par g un élément quelconque de la variété linéaire g engendrée par les g_i.

Soit donné un élément $\phi \in \mathcal{C}(H)$, on désigne par $\overline{V}$ le sous-ensemble des V défini par :

$$\overline{V} = \{g \in V \; : \; g(t) \leq \phi(t) \quad \forall \; t \in H\}.$$

Etant donné $f \in \mathcal{C}(K)$, $f \notin V$, on cherche un élément $g^* \in \overline{V}$, meilleur approximant, (s'il existe) tel que :

$$\|f - g^*\| = \text{Inf } (\|f-g\| \;\big|\; g \in \overline{V}) = \rho \; .$$

3 - Approximation uniforme de deux fonctions continues sur un compact.

Soit S un compact d'un espace métrique ; pour un élément $h \in \mathcal{C}(S)$ on note :

$$\|h\| = \text{Max}\ (|h(t)|\ |\ t \in S).$$

Soient g_i, $i = 1...n$, n éléments donnés appartenant à $\mathcal{C}(S)$ et linéairement indépendants, on désigne toujours par g un élément quelconque de la variété linéaire V engendrée par les g_i.

Soient donnés f_1 et $f_2 \in \mathcal{C}(S)$, f_1 et $f_2 \notin V$; sans restriction de généralité, étant donné le problème qu'on a en vue, on suppose :

$$f_1(t) \geq f_2(t) \qquad \forall\ t \in S \qquad (1)$$

(Si deux fonctions f_1 et f_2 ne vérifient pas (1), on se ramènera à ce cas en considérant les fonctions φ_1 et $\varphi_2 \in \mathcal{C}(S)$ définies par :

$$\varphi_1(t) = \text{Max}\ (f_1(t), f_2(t)) \qquad \forall\ t \in S$$

$$\varphi_2(t) = \text{Min}\ (f_1(t), f_2(t)) \qquad \forall\ t \in S. \quad)$$

Pour un élément $g \in V$, on pose :

$$d(g) = \text{Max}\ (\|f_1 - g\|\ ,\ \|f_2 - g\|)$$

On cherche un élément $g^* \in V$, meilleur approximant, (s'il existe) tel que :

$$d(g^*) = \text{Inf}\ (d(g)\,|\,g \in V) = \rho$$

Nous désignerons par PIIIC ce problème, version continue du problème PIII.

§2 - FORMULATION DES PROBLEMES DE PROGRAMMATION CONVEXE CORRESPONDANTS.

1 - Problème PIC.

En introduisant une variable auxiliaire x_{n+1}, le problème PIC peut s'énoncer ainsi :
Minimiser x_{n+1} sur le domaine convexe DI défini dans $\mathbb{R}^{n+1}$ par :

$$\text{DI} \quad \begin{cases} \sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq f(t) & \forall\, t \in S \\ -\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq -f(t) & \forall\, t \in S \end{cases}$$

2 - Problème PIIC

En introduisant une variable auxiliaire x_{n+1}, le problème PIIC peut s'énoncer ainsi :

Minimiser x_{n+1} sur le domaine convexe DII défini dans $\mathbb{R}^{n+1}$ par :

$$\text{DII} \quad \begin{cases} \sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq f(t) & \forall\, t \in K \\ -\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq -f(t) & \forall\, t \in K \\ \sum_{i=1}^{n} x_i\, \varphi_i(t) \geq \phi(t) - M & \forall\, t \in H \\ -\sum_{i=1}^{n} x_i\, \varphi_i(t) \geqslant -\phi(t) - M & \forall\, t \in H \end{cases}$$

3 - Problème PIIIC

En introduisant une variable auxilaire x_{n+1}, le problème PIIIC peut s'énoncer ainsi :

Minimiser x_{n+1} sur le domaine convexe DIII défini dans $\mathbb{R}^{n+1}$ par :

$$
\text{DIII} \quad \begin{cases}
\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq f_1(t) & \forall\, t \in S \qquad (1) \\
-\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq -f_1(t) & \forall\, t \in S \qquad (2) \\
\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq f_2(t) & \forall\, t \in S \qquad (3) \\
-\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq -f_2(t) & \forall\, t \in S \qquad (4)
\end{cases}
$$

Comme on a supposé $f_1(t) \geq f_2(t) \quad \forall\, t \in S$, (1) implique (3) et (4) implique (2) ; en supprimant les inégalités redondantes, DIII est défini par :

$$
\text{DIII} \quad \begin{cases}
\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} > f_1(t) & \forall\, t \in S \\
-\sum_{i=1}^{n} x_i\, g_i(t) + x_{n+1} \geq -f_2(t) & \forall\, t \in S
\end{cases}
$$

§3 - DEFINITION D'UN PROBLEME DE PROGRAMMATION CONVEXE PARTICULIER ; CONDITIONS D'OPTIMALITE

1 - Définition d'un problème particulier P de programmation convexe.

Au vu des problèmes précédents, nous allons définir le problème P de programmation convexe suivant, dont le formalisme inclut celui des problèmes PIC, PIIC et PIIIC :

Soient K_j, $j = 1...p$, p compacts quelconques d'un espace métrique, et $a_i^j(t)$, $i = 1...n$, $j = 1...p$, $b^j(t)$, $j = 1...p$, des fonctions à valeur dans $\mathbb{R}$, définies continues sur K_j.

On désigne par $a^j(t)$ le vecteur de $\mathbb{R}^n$ ayant pour composantes les $a_i^j(t)$ $(i = 1...n)$; et étant donné deux vecteurs x et y de $\mathbb{R}^n$ on désigne par $x.y$ leur produit scalaire.

On considère les ensembles suivants, convexes, fermés de $\mathbb{R}^n$:

$$
D_j = \{x \in \mathbb{R}^n : a^j(t).x \geq b^j(t) \quad \forall\, t \in K_j\} \quad \text{pour } j = 1...p
$$

et

$$
D = \bigcap_{j=1}^{p} D_j
$$

Soit c un vecteur de $\mathbb{R}^n$ donné, non nul, on cherche $x^* \in D$ tel que :

$$c.x^* = \text{Inf } (c.x \,|\, x \in D).$$

2 - Rappels et Notations.

Notations :

Soit $x \in D$, on définit :

$$E_j(x) = \{t \in K_j : a^j(t).x = b^j(t)\}$$

$$E(x) = \bigcup_{j=1}^{p} E_j(x)$$

$$\Omega_j(x) = \{y \in \mathbb{R}^n : a^j(t).y > o \quad \forall\, t \in E_j(x)\}$$

Soit enfin ξ un sous ensemble de E(x), on note :

$$\xi_j = \xi \cap E_j(x).$$

Rappels [1]

Soit un espace vectoriel X muni d'une topologie $\mathcal{T}$, invariante par homothétie et translation, et telle qu'à tout voisinage ouvert u de l'origine, on peut faire correspondre pour tout $x \in X$, un voisinage V de x et un scalaire $\varepsilon > 0$ tel que :

$$0 \leqslant \varepsilon' \leqslant \varepsilon,\ y \in V \Longrightarrow \varepsilon' y \in u.$$

Soient alors x_o un vecteur de X et E une partie de X, on appelle :
déplacement intérieur : "tout vecteur non nul $x \in X$ tel qu'il existe un voisinage V de x et un réel $\varepsilon>0$ tel que :

$$\forall\, \eta \in]0,\varepsilon[\ ,\ \forall\, y \in V \Longrightarrow x_o + \eta y \in E"$$

On peut vérifier que l'ensemble des déplacements intérieurs issus de x_o est un cône ouvert de sommet l'origine que l'on notera $\Gamma(E,x_o)$.

Propriété 1

On montre [11] que : "Si E est une partie convexe dont l'intérieur $\overset{o}{E}$ n'est pas vide, et si $x_o \in \bar{E}$, alors $\Gamma(E,x_o)$ est le cône de sommet l'origine translaté du cône de sommet x_o s'appuyant sur $\overset{o}{E}$ "

3 - Conditions de régularité imposées au domaine D.

Dans la suite, nous ferons les hypothèses H1 et H2 suivantes :

H1 : "Les ensembles $\overset{o}{D}_j$ (j = 1...p) définis par :

$$\overset{o}{D}_j = \{x \in \mathbb{R}^n : a^j(t).x > b^j(t) \quad \forall\, t \in K_j\}$$

sont non vides".

H2 : "L'intersection des ensembles $\overset{o}{D}_j$ (j = 1...p) est non vide".

Propriété 2.

"Si l'hypothèses H1 est vérifiée, alors $a^j(t) \neq 0 \quad \forall\, t \in E_j(x)$".

En effet, soit $\bar{x} \in D_j$, et $t_o \in K_j$ tels que :

$$a^j(t_o).\bar{x} = b^j(t_o)$$

Si l'on avait $a^j(t_o) = 0$, cela impliquerait $b^j(t_o) = 0$, et par conséquent il n'existerait pas de vecteur x de $\mathbb{R}^n$ tel que :

$$a^j(t_o).x > b^j(t_o),$$

donc $\overset{o}{D}_j$ serait vide.

Corollaire 1

"Si l'hypothèse H1 est vérifiée, $\overset{o}{D}_j$ est l'intérieur de D_j".

Remarquons tout d'abord que $\overset{o}{D}_j$ est un ouvert :

La fonction $f(t,x) = a^j(t).x - b^j(t)$ continue par rapport au couple de variables (t,x) est strictement positive pour $x \in \overset{o}{D}_j$ et pour $t \in K_j$; en raison de la compacité

de K_j et de la continuité de f par rapport au couple des variables, il est facile de voir que pour tout $x \in \overset{\circ}{D}_j$, il existe un voisinage V de x tel que :

$$\forall\, y \in V \Longrightarrow f(t,y) > 0 \qquad \forall\, t \in K_j .$$

Par ailleurs, et en raison de la propriété 2, il est facile de voir que tout ouvert inclus dans D_j est inclus dans $\overset{\circ}{D}_j$, donc $\overset{\circ}{D}_j$ est l'intérieur de D_j.

Corollaire 2.

"Si l'hypothèse H1 est vérifiée, et si $x_o \in D$, alors les cônes $\Gamma(D_j, x_o)$ $(j = 1...p)$, sont tous non vides, et si l'hypothèses H2 est vérifiée, leur intersection est non vide".

Cette propriété résulte directement de la propriété 1 et du corollaire 1.

4 - Caractérisation d'une solution x^* du problème P.

Lemme 1

"Etant donné un compact K d'un espace métrique, et deux applications f et g continues de K dans $\mathbb{R}$ telles que :

$$f(t) \geqslant 0 \qquad \forall\, t \in K$$
$$g(t) > 0 \qquad \forall\, t \in E$$

avec : $E = \{t \in K : f(t) = 0\}$

alors il existe $\lambda > 0$ tel que

$$f(t) + \lambda g(t) > 0 \qquad \forall\, t \in K"$$

On définit le sous ensemble $\mathcal{O}$, ouvert de K :

$$\mathcal{O} = \{t \in K : g(t) > 0\}$$

on a : $\mathcal{O} \supset E$

et : $\forall\, t \in \mathcal{O}$ et $\forall\, \lambda > 0$, on a : $f(t) + \lambda g(t) > 0$.

Soit $K' = K - \mathcal{O}$, K' est compact, et on pose :

$$\alpha = \text{Inf } (f(t) \,|\, t \in K'), \text{ on a : } \alpha > 0$$
$$\beta = \text{Sup } (|g(t)| \,|\, t \in K'), \ \beta \text{ est un nombre fini}$$

Il existe donc $\varepsilon > 0$ tel que :

$$\varepsilon\beta \leq \frac{\alpha}{2}$$

par conséquent, pour tout $\lambda \in]0,\varepsilon[$, on a :

$$f(t) + \lambda g(t) > 0 \qquad \forall\, t \in K$$

Lemme 2.

"Soit $\bar{x} \in D_j$, si l'hypothèses H1 est vérifiée, alors les ensembles $\Omega_j(\bar{x})$ et $\Gamma(D_j,\bar{x})$ coïncident".

Remarquons tout d'abord que si $E_j(\bar{x}) = \emptyset$, alors $\bar{x} \in \overset{\circ}{D}_j$, et D_j est un voisinage de $\bar{x}$, par conséquent [1], $\Gamma(D_j,\bar{x})$ est identique à $\mathbb{R}^n$, et il est facile de voir qu'il en est de même pour $\Omega_j(\bar{x})$.

Dans la suite, nous supposerons donc $E_j(\bar{x}) \neq \emptyset$:

- Montrons tout d'abord que $\Gamma(D_j,\bar{x}) \subset \Omega_j(\bar{x})$.

Soit $y \in \Gamma(D_j,\bar{x})$, il existe $\lambda > 0$ et $x \in \overset{\circ}{D}_j$ tel que :

$$y = \lambda(x - \bar{x})$$

Or, on a :

$$a^j(t).x > b^j(t) \qquad \forall\, t \in E_j(\bar{x})$$

et

$$a^j(t).\bar{x} = b^j(t) \qquad \forall\, t \in E_j(\bar{x}).$$

donc :

$$a^j(t).y > 0 \qquad \forall\, t \in E_j(\bar{x})$$

et par conséquent

$$y \in \Omega_j(\bar{x}).$$

- Montrons maintenant que $\Gamma(D_j,\bar{x}) \supset \Omega_j(\bar{x})$

Soit $y \in \Omega_j(\bar{x})$, on a :

$$a^j(t).\bar{x} + a^j(t).y - b^j(t) > 0 \qquad \forall\, t \in E_j(\bar{x})$$

En raison du lemme 1, il existe $\lambda>0$ tel que :

$$a^j(t).\bar{x} + \lambda\, a^j(t).y - b^j(t) > 0 \qquad \forall\, t \in K_j$$

donc : $\bar{x} + \lambda y \in \overset{\circ}{D}_j$ et par suite $y \in \Gamma(D_j,\bar{x})$.

Proposition 5-1

"Avec l'hypothèse H2, une condition nécessaire et suffisante pour que x^* soit solution du problème P est qu'il existe un ensemble de k points, ξ inclus dans $E(x^*)$, $1 \leqslant k \leqslant n$, et k coefficients $\lambda_r > 0$ tels que :

$$\sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r\, a^j(t_r) = c".$$

Condition nécessaire

Notations

Soit un sous-ensemble A de $\mathbb{R}^n$, on désigne par CO(A) l'enveloppe convexe de A, et par CC(A) le cône de sommet l'origine, enveloppe conique convexe de A.

Remarquons tout d'abord que si x^* est solution, il existe au moins un indice $j \in (1...p)$ tel que $E_j(x^*) \neq \emptyset$.
On définit :

$$D_o = \{x \in \mathbb{R}^n : cx < cx^*\}$$

Les cônes $\Gamma(D_j,x^*)$, $j = 0,1,...p$, étant tous non vides, on sait [1] qu'une condition nécessaire pour que x^* soit solution est que :

$$\bigcap_{j=o}^{p} \Gamma(D_j,x^*) = \emptyset \qquad (1)$$

En vertu du lemme 2, la relation (1) est équivalente à :

$$\left(\bigcap_{j=1}^{p} \Omega_j(x^*) \right) \cap \Gamma(D_o, x^*) = \emptyset \qquad (2)$$

En posant :

$$\Omega_o(x^*) = \Gamma(D_o, x^*),$$

la relation (2) s'écrit :

$$\bigcap_{j=o}^{p} \Omega_j(x^*) = \emptyset \qquad (2')$$

et d'après le théorème de Dubovickii et Miljutin [6], une condition nécessaire et suffisante pour que (2') soit vérifié est qu'il existe des formes linéaires continues ω_j, non toutes nulles, positives ou nulles respectivement sur les cônes $\Omega_j(x^*)$ et telles que :

$$\sum_{j=o}^{p} \omega_j = 0 \qquad (3)$$

La relation (3) peut s'écrire :

$$-\omega_o = \sum_{j=1}^{p} \omega_j \qquad (3')$$

Si l'on avait $\omega_o = 0$, on aurait $\sum_{j=1}^{p} \omega_j = 0$, et d'après le théorème de Dubovickii-Miljutin, cela impliquerait

$$\bigcap_{j=1}^{p} \Omega_j(x^*) = \emptyset$$

ce qui est impossible en raison de l'hypothèse H2.
On a donc :

$$\omega_o = - \mu_o \, c \text{ avec } \mu_o > 0.$$

et, en prenant $\mu_o = 1$, la relation (3') s'écrit :

$$c = \sum_{j=1}^{p} \omega_j \qquad (3'').$$

Soit $\bar{\Omega}_j(x^*)$ l'adhérence de $\Omega_j(x^*)$, ω_j est encore positive ou nulle sur $\bar{\Omega}_j(x^*)$, et par conséquent, ω_j appartient au cône $\tilde{\bar{\Omega}}_j(x^*)$ orthogonal du cône $\bar{\Omega}_j(x^*)$.

Soit $t \in E_j(x^*)$, on définit :

$$\Delta(t) = \{y \in \mathbb{R}^n : a^j(t).y \geq 0\}$$

et on a :

$$\bar{\Omega}_j(x^*) = \bigcup_{t \in E_j(x^*)} \Delta(t)$$

Soit $\tilde{\Delta}(t)$ le cône orthogonal du cône $\Delta(t)$, les cônes $\Delta(t)$ de sommet l'origine étant convexes, fermés, on a [4] :

$$\tilde{\bar{\Omega}}_j(x^*) = \overline{CO}\left(\bigcup_{t \in E_j(x^*)} \tilde{\Delta}(t)\right)$$

soit :

$$\tilde{\bar{\Omega}}_j(x^*) = \overline{CO}\left(\{\mu a^j(t),\ t \in E_j(x^*),\ \mu \geq 0\}\right)$$

$$\tilde{\bar{\Omega}}_j(x^*) = \overline{CC}\left(\overline{CO}(\{a^j(t),\ t \in E_j(x^*)\})\right)$$

on définit :

$$A_j = \{a^j(t),\ t \in E_j(x^*)\}$$

A_j est un compact de $\mathbb{R}^n$, donc on sait [7] que :

$$\overline{CO}(A_j) = CO(A_j)$$

Par ailleurs, en raison de la propriété 2, A_j ne contient pas l'origine, il en est de même de $CO(A_j)$, et par conséquent le cône convexe de sommet l'origine, enveloppe conique du compact $CO(A_j)$ est fermé [4] :

$$\overline{CC}(CO(A_j)) = CC(CO(A_j)) = CC(A_j)$$

donc :

$$\tilde{\bar{\Omega}}_j(x^*) = CC(A_j)$$

Soit alors :

$$\bar{\Omega}(x^*) = \bigcap_{j=1}^{p} \bar{\Omega}_j(x^*)$$

on a :

$$\tilde{\tilde{\Omega}}(x^*) = CC(\bigcup_{j=1}^{p} A_j)$$

Or la relation (3") peut s'interpréter comme suit :

$$c \in \tilde{\tilde{\Omega}}(x^*)$$

En vertu d'une conséquence simple du théorème de Carathéodory [7], on sait qu'il existe k points de $\bigcup_{j=1}^{p} A_j$, $k \leq n$, tels que c appartienne encore au cône convexe de sommet l'origine enveloppe de ces k points.
Il existe donc ξ ensemble de k points de $E(x^*)$, $k \leq n$, et k coefficients $\lambda_r > 0$ tels que :

$$c = \sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, a^j(t_r) \qquad (4)$$

Condition suffisante.

Soit $x^* \in D$, tel qu'il existe ξ, ensemble de k points ($1 \leq k \leq n$) de (x^*) et k coefficients $\lambda_r > 0$ vérifiant (4).
Supposons que x^* ne soit pas solution et qu'il existe $\bar{x} \in D$ tel que :

$$c.\bar{x} < c.x^* \qquad (5)$$

En tenant compte de (4), la relation (5) s'écrit :

$$\sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, a^j(t_r).(\bar{x}-x^*) < 0$$

soit :

$$\sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, a^j(t_r).\bar{x} < \sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, b^j(t_r)$$

Les λ_r étant strictement positifs, et comme $\bar{x} \in D$, on aboutirait à la contradiction :

$$\sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, b^j(t_r) < \sum_{j=1}^{p} \sum_{t_r \in \xi_j} \lambda_r \, b^j(t_r)$$

CHAPITRE - VI

APPROXIMATION UNIFORME D'UNE FONCTION CONTINUE SUR UN COMPACT

Ce problème classique a été rappelé au chapitre V (problème PIC), et nous avons déjà formulé le problème de programmation convexe correspondant.

L'application des méthodes de programmation convexe au problème PIC, nous permet de retrouver dans un premier paragraphe des propriétés connues (caractérisation de la solution, unicité), d'où découle l'algorithme de Remès que nous rappelons au deuxième paragraphe.

§1 - APPLICATION DES METHODES DE PROGRAMMATION CONVEXE AU PROBLEME PIC

1 - Existence d'une solution

Proposition 6-1

"Le problème PIC a au moins une solution"

On a vu que le problème PIC était équivalent au problème suivant :

Minimiser x_{n+1} sur le domaine convexe DI défini dans $\mathbb{R}^{n+1}$.

Nous allons montrer qu'une meilleure approximation $x^* \in \mathbb{R}^{n+1}$ ne peut se trouver que dans une partie Ω fermée, bornée de DI.

Soit $x^o \in$ DI ; un tel x^o existe toujours, DI n'étant pas vide ; on suppose $x^o_{n+1} > \rho$ (sinon x^o serait une meilleure approximation).

A x^o correspond le vecteur $g^o \in V$ tel que :

$$||f - g^o|| = x^o_{n+1}$$

Une meilleure approximation, si elle existe, appartient nécessairement au domaine convexe Γ défini dans $\mathbb{R}^{n+1}$ par :

$$\Gamma = \{x \in \mathbb{R}^{n+1} : x \in DI \cap \{x_{n+1} \leq x^o_{n+1}\}\}$$

On définit pour x la norme suivante sur $\mathbb{R}^{n+1}$:

$$||x||_1 = \text{Max}(\|g\|, |x_{n+1}|).$$

Soit $\bar{x} \in$ DI, tel que $\bar{g}$ étant le vecteur de V qui lui correspond, on ait :

$$\|\bar{g}-g_o\| > 2\, x^o_{n+1}$$

on a :

$$\bar{x}_{n+1} = \|f-\bar{g}\| > \left| \|\bar{g}-g_o\| - \|f-g^o\| \right| > 2\, x^o_{n+1} - x^o_{n+1}$$

soit $\bar{x}_{n+1} > x^o_{n+1}$

Donc $\bar{x}$ n'appartient pas à Γ, et pour un $x \in \Gamma$ on a nécessairement :

$$\|x-x^o\|_1 = \text{Max}(\|g-g^o\|, |x_{n+1}-x^o_{n+1}|) \leq 2\, x^o_{n+1}$$

Il suffit donc de chercher une meilleure approximation dans le domaine Ω fermé borné de DI

$$\Omega = \Gamma \cap BF(x^o, 2x^o_{n+1})$$

$BF(x^o, 2x^o_{n+1})$ étant la boule fermée de centre x^o et de rayon $2\, x^o_{n+1}$.
La fonction $h(x) = x_{n+1}$, continue de x, atteint donc sa borne inférieure sur le compact Ω.

2 - Caractérisation d'une solution.

Notations

Etant donné un vecteur $x \in$ DI, on définit les sous-ensembles $\mathcal{E}^+(x)$ et $\mathcal{E}^-(x)$ de S :

$$\mathcal{E}^+(x) = \{t \in S : f(t) - \sum_{i=1}^{n} x_i\, g_i(t) = x_{n+1}\}$$

$$\mathcal{E}^-(x) = \{t \in S : f(t) - \sum_{i=1}^{n} x_i\, g_i(t) = -x_{n+1}\}$$

et la fonction erreur $e \in \mathcal{C}(S)$:

$$e(t) = f(t) - \sum_{i=1}^{n} x_i \, g_i(t) \qquad \forall\, t \in S.$$

On remarque que :

$$\mathcal{E}^+(x) = \{t \in S : e(t) = \|e\| \}$$

$$\mathcal{E}^-(x) = \{t \in S : e(t) = -\|e\| \}$$

L'ensemble $\mathcal{E}^+(x) \cup \mathcal{E}^-(x)$ constitue l'ensemble des points extrémaux de l'erreur.

Proposition 6-2

"Une condition nécessaire et suffisante pour que $g^* \in V$ soit meilleure approximation de f, est qu'il existe k points extrémaux

$$t_1, t_2, \ldots t_k \qquad (k \leqslant n+1)$$

pour l'erreur $e^* = f - g^*$, et k coefficients λ_j tels que :

1- $\displaystyle\sum_{j=1}^{k} \lambda_j \, g_i(t_j) = 0 \qquad i = 1 \ldots n$

2- $\lambda_j > 0$ si $t_j \in \mathcal{E}^+(x^*)$

   $\lambda_j < 0$ si $t_j \in \mathcal{E}^-(x^*)$

3- $\displaystyle\sum_{j=1}^{k} |\lambda_j| = 1$ "

Cette proposition découle directement de la proposition 5-1.

3 - Unicité de la solution.

Proposition 6-3

"L'hypothèse H1,

H1 : "V vérifie la condition de Haar sur S [9]",

entraîne l'unicité de la solution".

En raison de l'hypothèse H1, le système

$$\sum_{j=1}^{k} \lambda_j \, g_i(t_j) = 0 \qquad i = 1...n$$

ne peut avoir de solution non identiquement nulle que si k est au moins égal à n+1. Par conséquent le nombre de points extrémaux (cf. Proposition 6-2) servant à caractériser une meilleure approximation est exactement égal à n+1.

Supposons alors qu'il existe deux solutions x^* et x^{**}, nous noterons g^* et g^{**} les meilleurs approximants correspondants, $\bar{g} = \frac{1}{2}(g^*+g^{**})$ est aussi meilleur approximant, et nous désignerons par :

$$\bar{t}_1, \bar{t}_2, \ldots, \bar{t}_{n+1}$$

les points extrémaux de $\bar{e} = f-\bar{g}$ intervenant dans la caractérisation de $\bar{g}$. On voit facilement que ces points sont aussi extrémaux pour e^* et e^{**} et de même nature, c'est-à-dire :

$$\text{Si } \bar{t}_j \in \mathcal{E}^+(\bar{x}) \text{ alors } \bar{t}_j \in \mathcal{E}^+(x^*) \text{ et } \bar{t}_j \in \mathcal{E}^+(x^{**})$$

$$\text{Si } \bar{t}_j \in \mathcal{E}^-(\bar{x}) \text{ alors } \bar{t}_j \in \mathcal{E}^-(x^*) \text{ et } \bar{t}_j \in \mathcal{E}^-(x^{**})$$

On pose :

$$d = g^* - g^{**}$$

On a :

$$d \in V \text{ et } d(\bar{t}_j) = 0 \qquad j = 1...n+1$$

En posant :

$$d(t) = \sum_{i=1}^{n} \alpha_i \, g_i(t)$$

les α_i (i=1...n) doivent vérifier :

$$\sum_{i=1}^{n} \alpha_i \, g_i(\bar{t}_j) = 0 \qquad j = 1...n+1$$

et en vertu de l'hypothèse H1, le déterminant de la matrice de terme général

$$g_i(\bar{t}_j) \qquad i = 1...n, \; j = 1...n$$

est non nul, ce qui implique

$$\alpha_i = 0 \qquad i = 1...n$$

§2 - ALGORITHME DE REMES [12]

Dans toute la suite, nous ferons l'hypothèse H1.

Cet algorithme découle des propositions 6-2 et 6-3, le principe de la méthode est le suivant :

On détermine la meilleure approximation sur un ensemble de n+1 points de S, et on fait évoluer ces points de façon à converger vers la meilleure approximation g^*.

1 - Description de l'algorithme.

a - Meilleure approximation sur un ensemble M_k de n+1 points de S.

$$M_k = \{t_j^k,\ j = 1...n+1\}.$$

on définit la meilleure approximation g^k sur M_k :

$$\text{Max}(|f(t)-g^k(t)|\ /\ t \in M_k) = \text{Inf}\left[\text{Max}(|f(t)-g(t)|\ /\ t \in M_k)\ /\ g \in V\right] = \rho_k$$

Soient λ_j^k, $j = 1...n+1$, les coefficients associés à M_k, (cf. proposition 6-2)
Les λ_j^k doivent vérifier :

$$(R)\quad \begin{cases} \sum_{j=1}^{n+1} \lambda_j^k\, g_i(t_j^k) = 0 & i = 1...n \\ \sum_{j=1}^{n+1} |\lambda_j^k| = 1 \end{cases}$$

Les relations (R) définissent les λ_j^k au signe près, la condition suivante, correspondant à la condition 2. de la proposition 6-2, permet de les déterminer complètement :

$$\sum_{j=1}^{n+1} \lambda_j^k\, f(t_j^k) > 0$$

(en effet, la condition 2 de la proposition 6-2 peut s'énoncer ainsi :

$$\text{signe}\ (\lambda_j^k) = \text{signe}\ (e_j^k) \qquad j = 1...n+1$$

donc :

$$\sum_{j=1}^{n+1} \lambda_j^k f(t_j^k) = \sum_{j=1}^{n+1} \lambda_j^k (f(t_j^k)-g^k(t_j^k)) = \rho_k \sum_{j=1}^{n+1} |\lambda_j^k| = \rho_k > 0).$$

On obtient alors les coefficients x_i^k ($i = 1...n$) de la meilleure approximation g^k sur M_k et l'écart $x_{n+1}^k = \rho_k$ correspondant par résolution du système :

$$\sum_{i=1}^{n} x_i\, g_i(t_j^k) + x_{n+1} \times \text{signe}(t_j^k) = f(t_j^k) \qquad j = 1...n+1.$$

Remarque :

Si g^k n'est pas la meilleure approximation g^*, on a $\rho_k < \rho$. En effet, soit :

$$d = g^* - g^k = (f-g^k) - (f-g^*)$$

si l'on avait :

$$\rho_k > \rho$$

on aurait :

$$\text{signe}(d(t_j^k)) = \text{signe}(e^k(t_j^k)) \quad j = 1...n+1$$

on aurait donc :

$$\sum_{j=1}^{n+1} \lambda_j^k d(t_j^k) > 0$$

d'où la contradiction puisque $d \in V$.
De plus, si $\rho_k = \rho$, g^k ne peut être que la meilleure approximation g^* d'après la proposition 6-2

b - Itération

Nous allons remplacer M_k par M_{k+1}, ne différant de M_k que par un point, de telle façon que l'écart ρ_{k+1} de la meilleure approximation g^{k+1} sur M_{k+1} soit tel que :

$$\rho_{k+1} > \rho_k$$

(ceci, dans le cas où g^k n'est pas déjà la meilleure approximation g^*)

α On choisit tout d'abord le point $\tau \in S$ qui va être introduit, tel que :

$$|e^k(\tau)| = ||e^k|| .$$

Remarque :

Si $\rho_k = ||e^k||$, on a terminé ; $g^k = g^*$.

β Nous allons maintenant choisir parmi les t_j^k, le point t_μ^k qui est supprimé, le nouvel ensemble M_{k+1} étant alors défini par :

$$\begin{cases} t_j^{k+1} = t_j^k & j = 1...n+1,\ j \neq \mu \\ t_\mu^{k+1} = \tau \end{cases}$$

Nous chosissons μ de telle façon que les coefficients λ_j^{k+1} associés à M_{k+1} vérifient les conditions (C) suivantes :

$$(C) \begin{cases} \text{signe } (\lambda_j^{k+1}) = \text{signe } (\lambda_j^k)\ j = 1...n+1,\ j \neq \mu \\ \text{signe } (\lambda_\mu^{k+1}) = \begin{cases} 1 \text{ si } e^k(\tau) = ||e^k|| \\ -1 \text{ si } e^k(\tau) = -||e^k|| \end{cases} \end{cases}$$

Nous verrons plus loin, lors de la démonstration de la convergence de l'algorithme, qu'avec un tel échange, on a :

$$\rho_{k+1} > \rho_k .$$

Indiquons maintenant comment obtenir en pratique les conditions (C) :

Détermination de μ et des coefficients λ_j^{k+1}

Etant donné les λ_j^k ($j = 1...n+1$) associés à M_k, on définit le vecteur $\Lambda^k \in \mathbb{R}^{n+2}$:

$$\begin{cases} \Lambda_j^k = \lambda_j^k\ j = 1...n+1 \\ \Lambda_{n+2}^k = 0 \end{cases}$$

Soit l'ensemble M de n+1 points, $M = (t_1^k, t_2^k, \ldots, t_n^k, \tau)$, on calcule les coefficients ω_j (j=1...n+1) associés à M par résolution du système :

$$\sum_{j=1}^{n} \omega_j \, g_i(t_j^k) + \omega_{n+1} \, g_i(\tau) = 0 \qquad i = 1 \ldots n$$

On définit le vecteur $\Omega \in \mathbb{R}^{n+2}$:

$$\begin{cases} \Omega_j = \omega_j & j = 1 \ldots n \\ \Omega_{n+1} = 0 \\ \Omega_{n+2} = \omega_{n+1} \end{cases}$$

Aux coefficients λ_j^{k+1} correspond un vecteur $\Lambda^{k+1} \in \mathbb{R}^{n+2}$, qui peut se mettre sous la forme :

$$\Lambda^{k+1} = \alpha \, \Lambda^k + \beta \, \Omega$$

α et β étant tels que : $\Lambda_\mu^{k+1} = 0$

$$\Lambda^{k+1} = \beta(\Omega - \frac{\omega_\mu}{\lambda_\mu^k} \Lambda^k)$$

Soit :

$$\Lambda_j^{k+1} = \beta \, \lambda_j^k \, (\frac{\omega_j}{\lambda_\mu^k} - \frac{\omega_\mu}{\lambda_\mu^k}) \qquad j = 1 \ldots n+1$$

$$\Lambda_{n+2}^{k+1} = \beta \omega_{n+2}$$

On peut choisir Ω de telle sorte que :

ω_{n+2} soit positif si $e^k(\tau) = \|e^k\|$
ω_{n+2} soit négatif si $e^k(\tau) = - \|e^k\|$

Pour que les conditions (C) soit satisfaites, il suffit d'avoir Λ_j^{k+1} du même signe que Λ_j^k, on prendra donc μ vérifiant :

$$\frac{\omega_\mu}{\lambda_\mu^k} = \mathrm{Inf}(\frac{\omega_j}{\lambda_j^k} \mid j = 1 \ldots n+1)$$

β est un coefficient positif que l'on choisit alors de façon à avoir :

$$\sum_{j=1}^{n+2} \left| \Lambda_j^{k+1} \right| = 1$$

Les coefficients λ_j^{k+1} ($j = 1...n+1$) associés à M_{k+1} sont alors définis par :

$$\begin{cases} \lambda_j^{k+1} = \Delta_j^{k+1} & j = 1...n+1,\ j \neq \mu \\ \lambda_\mu^{k+1} = \Delta_{n+2}^{k+1} \end{cases}$$

2 - Convergence de l'algorithme.

Proposition 6-4.

"La suite ρ_k est strictement croissante"

On a :

$$\rho_{k+1} = \sum_{j=1}^{n+1} \lambda_j^{k+1} f(t_j^{k+1})$$

soit :

$$\rho_{k+1} = \sum_{j=1}^{n+1} \lambda_j^{k+1} (f(t_j^{k+1}) - g^k(t_j^{k+1}))$$

En vertu des conditions (C) :

$$\rho_{k+1} = \sum_{j=1}^{n+1} |\lambda_j^{k+1}|\ |e^k(t_j^{k+1})|$$

soit :

$$\rho_{k+1} = \rho_k \sum_{\substack{j=1 \\ j\neq\mu}}^{n+1} |\lambda_j^{k+1}| + |\lambda_\mu^{k+1}|\ |e^k(\tau)|$$

$$\boxed{\rho_{k+1} = \rho_k + |\lambda_\mu^{k+1}|\ (\|e^k\| - \rho_k).} \qquad (1)$$

Comme $|\lambda_\mu^{k+1}|$ ne peut être nul, on a $\rho_{k+1} > \rho_k$, sauf si $\|e^k\| = \rho_k$, ce qui entraînerait : $g^k = g^*$.

Proposition 6-5.

"Il existe une constante $s > 0$ telle que les λ_j^k obtenus par le procédé vérifient :

$|\lambda_j^k| \geq s$ pour $j = 1...n+1$, $k = 1,2...$"

Supposons qu'il existe $j_o \in (1...n+1)$ tel que :

$$s_{j_o} = \text{Inf}\ (\lambda^k_{j_o} \mid k = 1,2,...) = 0.$$

Les vecteurs $\lambda^k \in \mathbb{R}^{n+1}$ parcourant un compact, il existerait alors une sous-suite λ^{k_m}, telle que :

$$\lim_{m\to\infty} (\lambda^{k_m}) = \bar{\lambda} \quad \text{avec } \bar{\lambda}_{j_o} = 0$$

et telle que les sous-suites $t_j^{k_m}$ correspondantes convergent également :

$$\lim_{m\to\infty} (t_j^{k_m}) = \bar{t}_j \qquad j = 1...n+1$$

(les t_j^k parcourant le compact S)
on a :

$$\begin{cases} \sum_{j=1}^{n+1} \lambda_j^{k_m}\, g_i(t_j^{k_m}) = 0 & i = 1...n \\ \sum_{j=1}^{n+1} |\lambda_j^{k_m}| = 1 \end{cases}$$

Par continuité, on aurait donc à la limite :

$$\begin{cases} \sum_{j=1}^{n+1} \bar{\lambda}_j\, g_i(\bar{t}_j) = 0 & i = 1...n \\ \sum_{j=1}^{n+1} |\bar{\lambda}_j| = 1 \\ \text{et} \quad \bar{\lambda}_{j_o} = 0 \end{cases}$$

D'où la contradiction, puisque, en vertu de l'hypothèse H1, le système

$$\sum_{\substack{j=1 \\ j\neq j_o}}^{n+1} \lambda_j\, g_i(\bar{t}_j) = 0 \qquad i = 1...n$$

ne peut avoir qu'une solution identiquement nulle.

Proposition 6-6.

"La suite g^k obtenue par le procédé, converge uniformement vers la meilleure approximation g^*".

En tenant compte de la proposition 6-5, l'inégalité (1) s'écrit :

$$\rho_{k+1} - \rho_k \geq s.(\|e^k\| - \rho_k)$$

$\{\rho_k\}$ est une suite croissante bornée par ρ , elle est donc convergente, et nous désignons par β sa limite :

Supposons $\beta < \rho$:

$$\lim_{k\to\infty} (\|e^k\| - \rho_k) = 0$$

donc

$$\lim_{k\to\infty} \|e^k\| = \beta$$

De la suite g^k, bornée de V, on pourrait alors extraire une sous-suite g^{k_ℓ} convergente de limite $\bar{g} \in V$ telle que :

$$\|f-\bar{g}\| = \beta < \rho$$

Il y a donc contradiction, et la suite $\{\rho_k\}$ a pour limite ρ.

Par conséquent toute sous-suite convergente de la suite $\{g^k\}$ converge nécessairement vers g^* (qui est unique).

D'autre part, $\{g^k\}$ étant une suite d'un compact, si g^k ne convergeait pas vers g^*, on pourrait extraire une sous-suite $\{g^{k_\ell}\}$ dont les éléments resteraient à une distance supérieure à un certain $\varepsilon>0$ de g^*. De la suite $\{g^{k_\ell}\}$ on pourrait à nouveau extraire une sous-suite convergente dont la limite ne pourrait être que g^*, d'où la contradiction.

CHAPITRE - VII

APPROXIMATION UNIFORME D'UNE FONCTION CONTINUE SUR UN COMPACT AVEC CONTRAINTE DU TYPE INEGALITE

Ce chapitre est consacré à l'étude du problème PIIC, défini au chapitre V.

Le plan de ce chapitre est le même que celui du chapitre VI, l'algorithme établi au deuxième paragraphe est une généralisation simple de l'algorithme de Remès. Dans un troisième paragraphe, nous donnons deux procédures Algol (l'une relative au problème PIIC, l'autre au problème P'IIC cas particulier de PIIC) et divers exemples numériques.

§1 - APPLICATION DES METHODES DE PROGRAMMATION CONVEXE AU PROBLEME PIIC.

1 - Existence d'une solution.

Proposition 7-1.

"Si DII n'est pas vide, alors le problème PIIC a au moins une solution".

Cette proposition se démontre de la même manière que la proposition 6-1.

Remarque 1

Posons :

$$\bar{\Phi} = \mathrm{Max}(\Phi(t) \mid t \in H)$$
$$\underline{\Phi} = \mathrm{Min}(\Phi(t) \mid t \in H)$$

S'il existe $\tilde{x} \in \mathbb{R}^n$ tel que $\psi = \sum_{i=1}^{n} x_i \psi_i$ soit constante, non nulle, sur H, et si de plus $M \geq \frac{\bar{\Phi}-\underline{\Phi}}{2}$, alors il est facile de voir que X n'est pas vide (il existe un scalaire λ tel que $\lambda \tilde{x} \in X$), et que par conséquent DII n'est pas vide.

Remarque 2

Considérons le problème P'IIC ; dans ce cas particulier, la non vacuité de $\overline{V}$ est assurée dès qu'il existe une fonction $\tilde{g} \in V$ de signe constant sur H.

2 - Caractérisation d'une solution.

Notations :

Etant donné un vecteur $x \in DII$, on définit les ensembles suivants :

$$\mathcal{E}^+(x) = \{t \in K : f(t) - \sum_{i=1}^{n} x_i\, g_i(t) = x_{n+1}\}$$

$$\mathcal{E}^-(x) = \{t \in K : f(t) - \sum_{i=1}^{n} x_i\, g_i(t) = -x_{n+1}\}$$

$$\mathcal{S}^+(x) = \{t \in H : \sum_{i=1}^{n} x_i\, \varphi_i(t) = \phi(t) - M\}$$

$$\mathcal{S}^-(x) = \{t \in H : \sum_{i=1}^{n} x_i\, \varphi_i(t) = \phi(t) + M\}$$

et les fonctions erreurs $e \in \mathcal{C}(K)$ et $\varepsilon \in \mathcal{C}(H)$:

$$\begin{cases} e(t) = f(t) - \sum_{i=1}^{n} x_i\, g_i(t) & \forall\, t \in K \\ \varepsilon(t) = \phi(t) - \sum_{i=1}^{n} x_i\, \varphi_i(t) & \forall\, t \in H \end{cases}$$

on remarque que :

$$\mathcal{E}^+(x) = \{t \in K : e(t) = \|e\|\}$$
$$\mathcal{E}^-(x) = \{t \in K : e(t) = -\|e\|\}$$
$$\mathcal{S}^+(x) = \{t \in H : \varepsilon(t) = M\}$$
$$\mathcal{S}^-(x) = \{t \in H : \varepsilon(t) = -M\}$$

L'ensemble $\mathcal{E}^+(x) \cup \mathcal{E}^-(x)$ constitue l'ensemble des points extrémaux de l'erreur.
$\mathcal{S}^+(x) \cup \mathcal{S}^-(x)$ sera appelé ensemble des points de saturation.
$C(x) = \mathcal{E}^+(x) \cup \mathcal{E}^-(x) \cup \mathcal{S}^+(x) \cup \mathcal{S}^-(x)$ sera appelé ensemble des points critiques.

Nous désignerons par H1, l'hypothèse suivante :

"il existe un vecteur $\bar{x} \in \mathbb{R}^{n+1}$ tel que :

$$\begin{cases} \bar{x}_{n+1} > \left| f(t) - \sum_{i=1}^{n} \bar{x}_i \, g_i(t) \right| & \forall\, t \in K \\ \left| \sum_{i=1}^{n} \bar{x}_i \, \varphi_i(t) - \phi(t) \right| < M & \forall\, t \in H \end{cases}$$ "

Proposition 7-2.

"Avec l'hypothèse H1, une condition nécessaire et suffisante pour que $x^* \in$ DII soit solution, est qu'il existe k points appartenant à $C(x^*)$,

$$t_1, t_2, \ldots, t_k \qquad (1 \leq k \leq n+1)$$

et k coefficients λ_j tels que :

1 - $$\sum_{j \in E} \lambda_j \, g_i(t_j) + \sum_{j \in S} \lambda_j \, \varphi_i(t_j) = 0 \qquad i = 1 \ldots n$$

(en notant :

$$E = \{j \in (1 \ldots k) : t_j \in \mathcal{E}^+(x^*) \cup \mathcal{E}^-(x^*)\}$$
$$S = \{j \in (1 \ldots k) : t_j \in \mathcal{S}^+(x^*) \cup \mathcal{S}^-(x^*)\})$$

2 - $\lambda_j > 0$ si $t_j \in \mathcal{E}^+(x^*) \cup \mathcal{S}^+(x^*)$

$\lambda_j < 0$ si $t_j \in \mathcal{E}^-(x^*) \cup \mathcal{S}^-(x^*)$

3 - $$\sum_{j \in E} |\lambda_j| = 1$$ "

Cette proposition découle directement de la proposition 5-1.

3 - Unicité de la solution.

Notations :

Soit $R = \{t_1, t_2, \ldots, t_n\}$ un ensemble de n points de $K \cup H$, et G la matrice de terme général G_i^j tel que :

$$G_i^j = \begin{cases} g_i(t_j) & i = 1 \ldots n \\ \text{ou} & \\ \varphi_i(t_j) & i = 1 \ldots n \end{cases}$$

pour $j = 1 \ldots n$.

Nous désignerons par H2 l'hypothèse suivante :

"Les matrices G sont régulières pour tout $R \subset K \cup H$"

Cette hypothèse, qui peut être considérée comme une condition de Haar généralisée, est évidemment rarement vérifiée dans la pratique. En fait, l'hypothèse H2, n'est pas nécessaire pour assurer la validité des résultats qui vont suivre, mais elle permet d'en clarifier et d'en simplifier l'exposé. Nous verrons au deuxième paragraphe que l'on peut se contenter dans la pratique d'une hypothèse moins restrictive.

Proposition 7-3.

"Avec les hypothèses H1 et H2, la solution du problème PIIC est unique".

Cette proposition se démontre de la même façon que la proposition 6-3 ; en raison de l'hypothèse H2, le système

$$\sum_{j=1}^{r} \lambda_j \, g_i(t_j) + \sum_{j=r+1}^{k} \lambda_j \, \varphi_i(t_j) = 0 \qquad i = 1 \ldots n$$

ne peut avoir de solution non identiquement nulle, que si k est au moins égal à n+1. Le nombre de points critiques servant à caractériser une meilleure approximation est donc exactement égal à n+1.

S'il existait alors deux solutions x^* et x^{**}, en notant g^* et g^{**} les meilleurs approximants correspondants, la fonction $\bar{g} = \frac{1}{2}(g^* + g^{**})$ serait aussi meilleur approximant et ses points critiques intervenant dans sa caractérisation seraient aussi points critiques et de même nature pour g^* et g^{**}, ce qui entraînerait $x^* \equiv x^{**}$ en raison de l'hypothèse H2.

§2 - ALGORITHME DE RESOLUTION.

Cet algorithme est une généralisation de l'algorithme de Remes, dans toute la suite nous ferons l'hypothèse H2 et l'hypothèse H3 suivante :

H3 : "Il existe $\tilde{x} \in \mathbb{R}^n$ tel que $\tilde{\varphi} = \sum_{i=1}^{n} \tilde{x}_i \, \varphi_i$ soit constante, non nulle, sur H, et on a de plus (avec les mêmes notations qu'en 7-1-1) :

$$M > \frac{\phi - \underline{\phi}}{2} \quad \text{et} \quad M < \bar{\phi} \qquad ".$$

L'hypothèse H3 entraîne l'hypothèse H1, et elle sera utile pour la démonstration du procédé numérique.

1 - Description de l'algorithme.

a - Meilleure approximation sur un ensemble M_k de n+1 points de $K \cup H$.

Soient donnés $M_k = \{t_j^k : j = 1...n+1\}$, $M_k \subset K \cup H$, et le vecteur $Z^k \in \mathbb{R}^{n+1}$ tel que

$$\begin{cases} Z_j^k = 1 & \text{si } t_j^k \text{ est appelé à être point extrémal} \\ Z_j^k = 0 & \text{si } t_j^k \text{ est appelé à être point de saturation.} \end{cases}$$

(Nous verrons par la suite, comment faire évoluer Z^k à chaque itération).

On désigne par :

$$E^k = \{j \in (1...n+1) : Z_j^k = 1\}$$

$$S^k = \{j \in (1...n+1) : Z_j^k = 0\}$$

et on définit :

$$X_k = \{x \in \mathbb{R}^n : \left| \sum_{i=1}^{n} x_i \varphi_i(t_j^k) - \phi(t_j^k) \right| \leqslant M \quad \forall\, j \in S^k\}$$

$$\bar{V}_k = \{g = \sum_{i=1}^{n} x_i g_i : x \in X_k\}$$

et la meilleure approximation g^k sur M_k par rapport à Z^k :

$$\rho_k = \text{Max}(|f(t_j^k) - g^k(t_j^k)| \,\big|\, j \in E^k) = \text{Inf}\left[\text{Max}(|f(t_j^k) - g(t_j^k)| \,\big|\, j \in E^k) \,\big|\, g \in \bar{V}_k\right]$$

Soient λ_j^k, j = 1...n+1, les coefficients associés à (M_k, Z^k) (cf. proposition 7-2), ils doivent vérifier :

$$(R) \begin{cases} \sum_{j \in E^k} \lambda_j g_i(t_j^k) + \sum_{j \in S^k} \lambda_j \varphi_i(t_j^k) = 0 & i = 1...n \\ \sum_{j \in E^k} |\lambda_j| = 1 \end{cases}$$

Les relations (R) définissent les λ_j^k au signe près, la condition suivante correspondant à la condition 2. de la proposition 7-2, permet de les déterminer complètement :

$$\sum_{j \in E^k} \lambda_j^k f(t_j^k) + \sum_{j \in S^k} \lambda_j^k (\phi(t_j^k) - M \text{ signe } (\lambda_j^k)) > 0$$

(en effet, on montrerait comme en 6-2-1 que

$$\rho_k = \sum_{j \in E^k} \lambda_j^k f(t_j^k) + \sum_{j \in S^k} \lambda_j^k (\phi(t_j^k) - M \text{ signe } (\lambda_j^k)).$$

On obtient alors les coefficients x_i^k $(i = 1...n)$ de la meilleure approximation g^k sur M_k par rapport à Z_k, et l'écart $x_{n+1}^k = \rho_k$ correspondant par résolution du système :

$$\begin{cases} \sum_{i=1}^{n} x_i \, g_i(t_j^k) + x_{n+1} . \text{signe } (\lambda_j^k) = f(t_j^k) & j \in E^k \\ \sum_{i=1}^{n} x_i \, \varphi_i(t_j^k) = \phi(t_j^k) - M . \text{signe } (\lambda_j^k) & j \in S^k \end{cases}$$

Remarque :

Si g^k n'est pas la meilleure approximation g^*, on a : $\rho_k < \rho$.
En effet, soit :

$$d = g^* - g^k = (f - g^k) - (f - g^*) ;$$

Si l'on avait $\rho_k > \rho$, $d(t_j^k)$ serait du signe de $f(t_j^k) - g^k(t_j^k)$ pour $j \in E^k$.
Par ailleurs, soit

$$\delta = \sum_{i=1}^{n} x_i^* \varphi_i - \sum_{i=1}^{n} x_i^k \varphi_i,$$

puisque $g^* \in \bar{V}$, on a :

$$\left.\begin{array}{l} \delta(t_j^k) \geqslant 0 \text{ pour } \lambda_j^k > 0 \\ \delta(t_j^k) \leqslant 0 \text{ pour } \lambda_j^k < 0 \end{array}\right\} j \in S^k$$

on aurait donc, puisque E^k doit être non vide :

$$\sum_{j \in E^k} \lambda_j^k \, d(t_j^k) + \sum_{j \in S^k} \lambda_j^k \, \delta(t_j^k) > 0$$

d'où la contradiction puisque $d \in V$.
De plus si $\rho_k = \rho$, alors g^k ne peut être que la meilleure approximation d'après la proposition 7-2.

b - Itération.

Nous allons remplacer M_k par M_{k+1}, ne différant de M_k que par un point, de telle façon que l'écart correspondant ρ_{k+1} soit supérieur à ρ_k. (dans le cas où g_k n'est pas déjà la meilleure approximation).

α On choisit tout d'abord le point $\tau \in K \cup H$ qui va être introduit.
On définit :

$$\begin{cases} e^k(t) = f(t) - g^k(t) & \forall\, t \in K \\ \varepsilon^k(t) = \phi(t) - \sum_{i=1}^{n} x_i^k \varphi_i(t) & \forall\, t \in H \end{cases}$$

et les points $\tau_1 \in K$, et τ_2 et $\tau_3 \in H$ tels que :

$$|e^k(\tau_1)| = \|e^k\|$$
$$\varepsilon^k(\tau_2) = \text{Min}\,(\varepsilon^k(t) \,|\, t \in H) = m_1^k$$
$$\varepsilon^k(\tau_1) = \text{Max}\,(\varepsilon^k(t) \,|\, t \in H) = m_2^k$$

on choisit :

$$\tau = \tau_1 \text{ si } \quad \|e^k\| - \rho_k = \text{Max}(\|e^k\| - \rho_k,\ m_2^k - M,\ -m_1^k - M)$$
$$\tau = \tau_2 \text{ si } \quad -m_1^k - M = \text{Max}(\|e^k\| - \rho_k,\ m_2^k - M,\ -m_1^k - M)$$
$$\tau = \tau_3 \text{ si } \quad m_2^k - M = \text{Max}(\|e^k\| - \rho_k,\ m_2^k - M,\ -m_1^k - M)$$

Remarque :
Si $m_1^k \geqslant -M$, $m_2^k \leqslant M$, et $\|e^k\| = \rho_k$, alors on a terminé, et $g^k = g^*$.

β On choisit maintenant parmi les t_j^k, le point t_μ^k qui est supprimé, le nouvel ensemble M_{k+1} étant alors défini par :

$$\begin{cases} t_j^{k+1} = t_j^k & j = 1...n+1,\ j \neq \mu \\ t_\mu^{k+1} = \tau \end{cases}$$

Par ailleurs, on définit le vecteur Z^{k+1} à partir de Z^k :

$$Z_j^{k+1} = Z_j^k \qquad j = 1...n+1,\ j \neq \mu$$

$$Z_\mu^{k+1} = \begin{cases} 1 & \text{si } \tau = \tau_1 \\ 0 & \text{si } \tau = \tau_2, \text{ ou si } \tau = \tau_3. \end{cases}$$

Nous choisissons μ de telle façon que les coefficients λ_j^{k+1} associés à (M_{k+1}, Z^{k+1}) vérifient les conditions (C) suivantes :

$$(C) \quad \begin{cases} \text{signe } (\lambda_j^{k+1}) = \text{signe } (\lambda_j^k) & j \neq \mu \\ \text{signe } (\lambda_\mu^{k+1}) = \begin{cases} +1 \text{ si } \tau = \tau_1 \text{ avec } e^k(\tau_1) = \|e^k\|, \text{ ou si } \tau = \tau_3 \\ -1 \text{ si } \tau = \tau_1 \text{ avec } e^k(\tau_1) = -\|e^k\|, \text{ ou si } \tau = \tau_2 \end{cases} \end{cases}$$

Pour l'obtention en pratique des conditions (C), on opère d'une facon analogue à ce qui a été fait en 6-2-1.

2 - Convergence de l'algorithme.

Proposition 7-4

"La suite ρ_k est strictement croissante".

on a :

$$\rho_{k+1} = \sum_{j \in E^{k+1}} \lambda_j^{k+1} f(t_j^{k+1}) + \sum_{j \in S^{k+1}} \lambda_j^{k+1} (\phi(t_j^{k+1}) - M.\text{signe}(\lambda_j^{k+1}))$$

D'une façon analogue à ce qui a été fait en 6-2-2, et en vertu des conditions (C), on montre que :

$$\rho_{k+1} = \rho_k + |\lambda_\mu^{k+1}| \cdot \mathrm{Max}(\| e^k \| - \rho_k, -m_1^k - M, m_2^k - M) \qquad (1)$$

Comme λ_μ^{k+1} ne peut être nul, on a $\rho_{k+1} > \rho_k$, sauf si $\|e^k\| = \rho_k$, $m_1^k \geqslant -M$, $m_2^k \leqslant M$, ce qui entraînerait alors $g^k = g^*$.

Proposition 7-5

"Il existe une constante $s>0$ telle que les λ_j^k obtenus par le procédé vérifient :

$$|\lambda_j^k| > s \qquad \text{pour } j = 1\ldots n+1,\ k = 1, 2\ldots"$$

Pour démontrer cette proposition, il nous suffit de montrer, qu'il existe une constante $\gamma > 0$ telle que :

$$\sum_{j=1}^{n+1} |\lambda_j^k| \leqslant \gamma \qquad \text{pour tout } k = 1,2\ldots,$$

la suite de la démonstration étant alors identique à la démonstration de la proposition 6-5.

Ayant fait l'hypothèse H3, il existe $\tilde{x} \in \mathbb{R}^n$, tel que

$$\sum_{i=1}^{n} \tilde{x}_i \varphi_i(t) = \xi \qquad \forall\, t \in H$$

nous supposerons $\xi > 0$, et on note :

$$\tilde{g} = \sum_{i=1}^{n} \tilde{x}_i g_i$$

on a :

$$\sum_{j \in E^k} \lambda_j^k \tilde{g}(t_j^k) + \sum_{j \in S^k} \lambda_j^k \xi = 0 \qquad \text{pour } k = 1,2\ldots \qquad (1)$$

on définit :

$$S_+^k = \{j \in S^k : \lambda_j^k > 0\}$$

$$S_-^k = \{j \in S^k : \lambda_j^k < 0\}$$

et $A = \text{Max}\ (|\tilde{g}(t)|\ /\ t \in K)$

On a donc :

$$\sum_{j \in S_+^k} |\lambda_j^k| - \sum_{j \in S_-^k} |\lambda_j^k| \leq \frac{A}{\xi} \qquad (2)$$

Par ailleurs :

$$\rho_k = \sum_{j \in E^k} \lambda_j^k f(t_j^k) + \sum_{j \in S_+^k} \lambda_j^k (\phi(t_j^k)-M) + \sum_{j \in S_-^k} \lambda_j^k (\phi(t_j^k)+M)$$

En retranchant (1) de l'égalité ci-dessus et en notant

$$B = \text{Max}(|f(t) - \tilde{g}(t)|\ /\ t \in K)$$

on obtient :

$$(\underline{\phi}+M-\xi) \times \sum_{j \in S_-^k} |\lambda_j^k| \leq B + \sum_{j \in S_+^k} |\lambda_j^k| \times (\bar{\phi}-M-\xi) \qquad (3)$$

En combinant (2) et (3), il vient :

$$(\underline{\phi} - \bar{\phi} + 2M) \times \sum_{j \in S_-^k} |\lambda_j^k| \leq B + \frac{A}{\xi} (\bar{\phi} - M - \xi)$$

Par conséquent, en vertu de l'hypothèse H3, $\sum_{j \in S_-^k} |\lambda_j^k|$ est borné ; en tenant compte de (2), $\sum_{j \in S_+^k} |\lambda_j^k|$ est également borné, ce qui achève la démonstration.

Proposition 7-6

"La suite g^k obtenue par le procédé, converge uniformément vers la meilleure approximation g^*"

En tenant compte de la proposition 7-5, l'inégalité (1) de la proposition 7-4 s'écrit :

$$\rho_{k+1} - \rho_k \geq s.\text{Max}\ (||e^k|| - \rho_k,\ -m_1^k - M,\ m_2^k - M)$$

La suite de la démonstration est identique à celle de la proposition 6-6, on montre tout d'abord que la suite convergente $\{\rho_k\}$ a pour limite ρ, et ensuite que g^k converge uniformément vers g^*.

§3 - PROCEDURES ALGOL ET ASPECTS NUMERIQUES

1 - Notice d'emploi de la procédure TCHEBCØNT 2 résolvant le problème PIIC.

TCHEBCØNT2 (AI,AS,CI,CS,NA,NC,N,F,FP,CAL,CALP,EPS,S,X,REF,E,BUL,IMPØS) ;

Les tableaux AI et AS de dimension NA, et les tableaux CI et CS de dimension NC définissent respectivement les compacts K et H de $\mathbb{R}$:

$$K = \bigcup_{I=1}^{NA} [AI[I], AS[I]]$$

$$H = \bigcup_{I=1}^{NC} [CI[I], CS[I]]$$

La fonction F définie sur le compact K est précisée par la réelle procédure F et la fonction ϕ définie sur le compact H est précisée par la réelle procédure FP.

La variété V de dimension N est précisée par la procédure CAL ; l'appel CAL(T,G) affecte à G[I], I = 1...N, la valeur $g_I(T)$ pour $T \in K$.

De même, l'appel de la procédure CALP (T,PHI) affecte à PHI[I] , I = 1...N, la valeur $\varphi_I(T)$ pour $T \in H$.

On considère les ensembles $\bar{X}$ de $\mathbb{R}^N$ et $\bar{V}$ de V (cf. 5-1-2) :

$$\bar{X} = \{A \in \mathbb{R}^N : \left| \sum_{I=1}^{N} A[I] \times \varphi_I(T) - \phi(T) \right| \leq S \qquad \forall\, T \in H\}$$

$$\bar{V} = \{g \in V : g(T) = \sum_{I=1}^{N} A[I].g_I(T) \qquad \forall\, T \in K,\ A \in \bar{X}\}$$

La procédure calcule les coefficients X[I], I = 1...N, du meilleure approximant de F dans $\bar{V}$, et l'écart correspondant X[N+1] :

$$X[N+1] = \text{Max}(\left|F(T) - \sum_{I=1}^{N} X[I] \times g_I(T)\right| \mid T \in K)$$

Soient les fonctions ε_1 et ε_2 définies par :

$$\begin{cases} \varepsilon_1(T) = F(T) - \sum_{I=1}^{N} X[I] \times g_I(T) & T \in K \\ \varepsilon_2(T) = \phi(T) - \sum_{I=1}^{N} X[I] \times \varphi_I(T) & T \in H \end{cases}$$

Les tableaux REF[1:N+1] et E[1:N+1] donnent respectivement la valeur des abcisses des points critiques de la solution et la valeur de l'erreur correspondante :

$$E[I] = \begin{cases} \varepsilon_1(REF[I]) & \text{si REF[I] est un point extrémal} \\ \varepsilon_2(REF[I]) & \text{si REF[I] est un point de saturation} \end{cases}$$

La procédure comporte deux tests d'arrêt :

1 L'algorithme calcule à chaque itération une minoration d_k et une majoration m_k de l'écart, et on arrête les calculs lorsque $(m_k - d_k)/d_k < EPS$; le booléen BUL prend alors la valeur VRAI.

2 Pour éviter le cyclage, dans le cas où EPS aurait été choisi trop petit, la procédure vérifie que les itérations nous conduisent bien à des situations distinctes ; si tel n'était pas le cas, le calcul s'arrête et le booléen BUL prend la valeur FAUX.

A chaque itération, la procédure calcule la meilleure approximation sur un ensemble de N+1 points de $K \cup H$; s'il y a non-unicité de la meilleure approximation sur l'un de ces ensembles, on sort de la procédure par l'étiquette IMPØS.

2 - Exemples numériques traités par la procédure TCHEBCØNT 2.

1er exemple.

On considère la version continue de l'exemple traité en 3-3-2. On cherche parmi les polynômes de degré 4, le meilleur approximant de la fonction sin(t) sur $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, et on impose de plus au meilleur approximant $P^*(t) = \sum_{i=1}^{5} x_i\, t^{i-1}$ de vérifier :

$$|\cos(t) - P'^*(t)| \leq S \qquad t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$$

On prend S = 0,7 et EPS = 10^{-4} ; on trouve :

x_1 = 0
x_2 = 0,985 529 5
x_3 =-0,000 000 0
x_4 =-0,142 566 7
x_5 = 0,000 000 0

REF[1] = -1,570 795 0	E[1] = -0,004 491 7
REF[2] = -1,265 134 1	E[2] = 0,004 491 7
REF[3] = -0,479 981 9	E[3] = -0,004 491 7
REF[4] = 0,480 350 5	E[4] = 0,004 491 7
REF[5] = 1,265 011 5	E[5] = -0,004 491 6
REF[6] = 1,570 795 0	E[6] = 0,004 491 7

On remarque que tous les points critiques sont extrémaux, on a donc calculé, parmi les polynômes de degré 4, le meilleur approximant de sin(t) sur $[-\frac{\pi}{2}, \frac{\pi}{2}]$ sans contrainte.

L'écart X[6] a pris la valeur : 0,004 491 7 ; et BUL a pris la valeur VRAI.

2ème exemple

On considère le même exemple que précédemment, mais avec S = 0,6 ; on trouve :

x_1 = 0,000 000 0
x_2 = 0,918 647 2
x_3 = 0,007 798 4
x_4 =-0,129 142 5
x_5 =-0,003 160 6

REF[1] = 1,570 795 0	E[1] = 0,057 520 9
*REF[2] = 1,570 795 0	E[2] = -0,599 999 9
REF[3] = -1,570 795 0	E[3] = -0,057 520 9
REF[4] = -1,158 982 7	E[4] = -0,057 520 9
REF[5] = -1,175 427 0	E[5] = -0,057 520 9
*REF[6] = -1,570 795 0	E[6] = -0,599 999 9

On a mis une * devant les points critiques de saturation.
L'écart X[6] a pris la valeur : 0,057 520 9 ; et BUL a pris la valeur VRAI.
Le temps de calcul et d'impression des résultats a été pour l'ensemble des deux exemples de 1 minute et 36 secondes (sur IBM 7044).

3 - Notice d'emploi de la procédure TCHEBINEG2 résolvant le problème P'IIC.

TCHEBINEG2(AI,AS,CI,CS,NA,NC,N,F,EPS,CAL,X,REF,E,BUL).

Les tableaux AI et AS de dimension NA, et les tableaux CI et CS de dimension NC définissent respectivement les compacts K et H de $\mathbb{R}$:

$$K = \bigcup_{I=1}^{NA} [AI[I], AS[I]]$$

$$H = \bigcup_{I=1}^{NC} [CI[I], CS[I]]$$

(et on doit avoir : $K \cap H = \emptyset$. (cf.5-1-2)).

La variété V de dimension N est précisée par la procédure CAL ; l'appel CAL(T,G) affecte à G[I], I = 1...N, la valeur $g_I(T)$ pour $T \in K \cup H$.
(On suppose que V vérifie la condition de Haar).

Les fonctions F définies sur K et $\emptyset$ définie sur H sont précisées toutes les deux par la réelle procédure F.
On définit le sous-ensemble $\overline{V}$ de V (cf. 5-1-2) :

$$\overline{V} = \{g \in V : g(T) \leq \phi(T) \qquad \forall\, t \in H\}$$

La procédure calcule les coefficients X[I], I = 1...N, du meilleur approximant de F dans $\overline{V}$ et l'écart correspondant X[N+1] :

$$X[N+1] = \text{Max}\left(\left|F(T) - \sum_{I=1}^{N} X[I]\times g_I(T)\right| \,\middle|\, t \in K\right).$$

Soient les fonctions ε_1 et ε_2 définies par :

$$\begin{cases} \varepsilon_1(T) = F(T) - \sum_{I=1}^{N} X[I]\times g_I(T) & T \in K \\ \varepsilon_2(T) = \phi(T) - \sum_{I=1}^{N} X[I]\times g_I(T) & T \in H \end{cases}$$

Les tableaux REF[1:N+1] et E[1:N+1] donnent respectivement la valeur des abcisses des points critiques de la solution et la valeur de l'erreur correspondante :

$$E[I] = \begin{cases} \varepsilon_1(REF[I]) \text{ si } REF[I] \text{ est un point extrémal} \\ \\ \varepsilon_2(REF[I]) \text{ si } REF[I] \text{ est un point critique non extrémal.} \end{cases}$$

La procédure comporte deux tests d'arrêt tout à fait identiques à ceux de la procédure TCHEBCØNT2.

4 - Exemple numérique traité par la procédure TCHEBINEG2

On considère la version continue de l'exemple traité en 3-3-4. On cherche parmi les polynômes de degré 5 le meilleur approximant sur $[3 ; 4]$ de la fonction f(t) définie par :

$$\begin{cases} f(t) = 4t - 12 & \text{pour } t \in [3 ; 3,5] \\ f(t) = -4t + 16 & \text{pour } t \in [3,5 ; 4] \end{cases}$$

On impose de plus au meilleur approximant $\sum_{i=1}^{6} x_i\, t^{i-1}$ de vérifier :

$$\sum_{i=1}^{6} x_i\, t^{i-1} \leqslant \phi(t) \quad \text{pour } t \in [0 ; 2] \cup [5 ; 7],$$

$\phi(t)$ étant définie par :

$$\begin{cases} \phi(t) = -\sqrt{-t^2+2t} & \text{pour } t \in [0 ; 2] \\ \phi(t) = 2t^2 - 24t + 70 & \text{pour } t \in [5 ; 7] \end{cases}$$

On a choisi EPS = 10^{-3}, on trouve :

x_1 =	0
x_2 =	-55, 108 2
x_3 =	39, 925 8
x_4 =	- 9, 153 8
x_5 =	0, 652 7
x_6 =	0, 000 1

* REF[1] = 0	E[1] = 0
REF[2] = 3,000 0	E[2] = 0,246 1
REF[3] = 3,252 0	E[3] = -0,246 1
REF[4] = 3,500 0	E[4] = 0,246 1
REF[5] = 3,751 4	E[5] = -0,246 1
REF[6] = 4,000 0	E[6] = 0,246 1
* REF[7] = 7,000 0	E[7] = -0,000 1

On a mis une * devant les points critiques non extrémaux.

L'écart X[7] a pris la valeur : 0,246 1 ; et BUL a pris la valeur FAUX.

Le temps de calcul et d'impression des résultats a été de 1 minute et 16 secondes (sur IBM 7044).

```
   PROCEDURE  TCHEBCONT2(AI,AS,CI,CS,NA,NC,N,F,FP,CAL,CALP,EPS,S,
     X,REF,E,BUL,IMPOS) ;
   VALEUR NA,NC,N,EPS,S ;
   TABLEAU AI,AS,CI,CS,REF,E,X ;
   ETIQUETTE IMPOS ;
   BOOLEEN BUL ;
   REEL PROCEDURE F, FP ;
   ENTIER NA,NC,N ; REEL EPS,S ;
   PROCEDURE CAL,CALP ;

   DEBUT

PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
           VOIR  ANNEXE  .....

PROCEDURE ABS MAX(E,A,B,X0,MAX) ; .....
           VOIR  ANNEXE  .....


   REEL PROCEDURE ER(T) ;
  VALEUR T ; REEL T ;
     DEBUT
   REEL R ; TABLEAU G[1:N] ; ENTIER J ;
            CAL(T,G) ; R:=0.0 ;
            POUR J:=1 PAS 1 JUSQUA N FAIRE
            R:=R+X[J]*G[J] ;
    ER:=F(T) -R ;
      FIN ER ;



   REEL PROCEDURE ERP(T) ;
  VALEUR T ; REEL T ;
     DEBUT
   REEL R ; TABLEAU G[1:N] ; ENTIER J ;
            CALP(T,G) ; R:=0.0 ;
            POUR J:=1 PAS 1 JUSQUA N FAIRE
            R:=R+X[J]*G[J] ;
   ERP:=FP(T) -R ;
      FIN ERP ;



      TABLEAU ERST[1:N+1,1:N+1],PREM[1:N,1:N],
         MU[1:N+2],SEC,G,LAMBAUX[1:N],LAMBDA,ZWEI[1:N+1] ;
      BOOLEEN TABLEAU BOOL[1:N+1] ;
      ENTIER I,J,J0 ;
      REEL T,R,AMPLI,I0,P0,Q0,MP,MQ ;

   INIT: POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
      DEBUT
         REF[I]:=AI[1]+(I-0.5)*(AS[1]-AI[1])/(N+1) ;
         BOOL[I]:= VRAI ;
      FIN ;
      BUL:= VRAI ;
```

```
      POUR J:=1 PAS 1 JUSQUA N FAIRE
      DEBUT
         CAL(REF[J],G) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE PREM[I,J]:=G[I] ;
      FIN ;
   CAL(REF[N+1],G) ;
   POUR I:=1 PAS 1 JUSQUA N   FAIRE
      SEC[I]:=-G[I] ;
   GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,SORTIE) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
      LAMBDA[I]:=LAMBAUX[I] ;
      LAMBDA[N+1]:=1.0 ;
        T:=F(REF[N+1]) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
        T:=T  +LAMBDA[I]*F(REF[I]) ;
   SI T  < 0 ALORS
         POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
         LAMBDA[I]:=-LAMBDA[I] ;

   BOUCLE:
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   SI BOOL[I] ALORS
      DEBUT
         CAL(REF[I],G) ;
         POUR J:=1 PAS 1 JUSQUA N FAIRE
         ERST[I,J]:=G[J] ;
      FIN
         SINON
      DEBUT
         CALP(REF[I],G) ;
         POUR J:=1 PAS 1 JUSQUA N FAIRE
         ERST[I,J]:=G[J] ;
      FIN ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   ERST[I,N+1]:= SI BOOL[I] ALORS
  ( SI LAMBDA[I] > 0 ALORS 1.0 SINON -1.0) SINON 0.0 ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   ZWEI[I]:= SI BOOL[I] ALORS F(REF[I])
         SINON ( SI LAMBDA[I] > 0 ALORS FP(REF[I])-S
            SINON FP(REF[I])+S);
   GRESOLSYSLINE(ERST,ZWEI,X,N+1,SORTIE) ;
   MP:=0.0 ;
   POUR I:=1 PAS 1 JUSQUA NA FAIRE
      DEBUT
  ABSMAX(ER,AI[I],AS[I],T,R) ;
  SI R > MP ALORS
            DEBUT
            MP:=R ;
            P0:=T
            FIN ;
      FIN ;
   MQ:=0.0 ;
   POUR I:=1 PAS 1 JUSQUA NC FAIRE
      DEBUT
     ABSMAX(ERP,CI[I],CS[I],T,R) ;
```

```
    SI R > MQ ALORS
            DEBUT
              MQ:=R ;
              Q0:=T
            FIN ;
       FIN ;
   SI MP-X[N+1] >      MQ- S ALORS I0:=P0 SINON I0:=Q0 ;
AMPLI:= SI MQ-S > 0 ALORS MP+2*(MQ-S) SINON MP ;
   SI 1-X[N+1]/AMPLI INFEG EPS ALORS ALLERA  END ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   DEBUT
     SI I0=P0 ET ABS(I0-REF[I]) INFEG 0.0001 ET
      ABS(ER(I0)-X[N+1]) INFEG 0.001 ALORS
       DEBUT BUL:= FAUX ; ALLERA  END FIN ;
     SI I0=Q0 ET ABS(I0-REF[I]) INFEG 0.0001 ET
      ABS(ERP(I0)-X[N+1]) INFEG 0.001 ALORS
       DEBUT BUL:= FAUX ; ALLERA  END FIN ;
   FIN ;
   SI I0=P0 ALORS
         MU[N+2]:= SI ER(P0) > 0 ALORS 1.0 SINON -1.0
         SINON
         MU[N+2]:= SI ERP(Q0) > 0 ALORS 1.0 SINON -1.0 ;
   POUR J:=1 PAS 1 JUSQUA N FAIRE
   SI BOOL[J] ALORS
      DEBUT
         CAL(REF[J],G) ;
         POUR I:=1 PAS 1 JUSQUA N FAIRE
         PREM[I,J]:=G[I]
      FIN
         SINON
      DEBUT
         CALP(REF[J],G) ;
         POUR I:=1 PAS 1 JUSQUA N FAIRE
         PREM[I,J]:=G[I]
      FIN ;
   SI I0=P0 ALORS
      DEBUT CAL(I0,G) ;
         POUR I:=1 PAS 1 JUSQUA N FAIRE
         SEC[I]:=-MU[N+2]*G[I]
      FIN
         SINON
      DEBUT
         CALP(I0,G) ;
         POUR I:=1 PAS 1 JUSQUA N FAIRE
         SEC[I]:=-MU[N+2]*G[I]
      FIN ;
   GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,SORTIE) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
   MU[I]:=LAMBAUX[I] ;
   MU[N+1]:=0.0 ;
   J0:=1 ;
   POUR I:=2 PAS 1 JUSQUA N+1 FAIRE
   J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
   ALORS I SINON J0 ;
   POUR I:=1 PAS 1 JUSQUA J0-1,J0+1 PAS 1 JUSQUA N+1 FAIRE
```

```
   LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
   LAMBDA[J0]:=MU[N+2] ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
   SI I0=P0 ALORS BOOL[J0]:= VRAI SINON BOOL[J0]:= FAUX ;
   ALLERA  BOUCLE ;
   SORTIE: ALLERA  IMPOS ;
  END:
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
  E[I]:= SI BOOL[I] ALORS ER(REF[I]) SINON ERP(REF[I]) ;
 FIN TCHEBCONT2 ;
```

****

```
 PROCEDURE TCHEBINEG2(AI,AS,CI,CS,NA,NC,N,F,EPS,CAL,X,REF,E,BUL) ;
   VALEUR NA,NC,N,EPS;
   ENTIER NA,NC,N ; REEL EPS;
  BOOLEEN BUL ;
   TABLEAU AI,AS,CI,CS,REF,E,X ;
      REEL PROCEDURE F ;
    PROCEDURE CAL ;

   DEBUT

PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
          VOIR  ANNEXE  .....

PROCEDURE ABS MAX(E,A,B,X0,MAX) ; .....
          VOIR  ANNEXE  .....

PROCEDURE  ALGMIN(E,A,B,X0,MIN) ; .....
            VOIR  ANNEXE .....

   REEL PROCEDURE ER(T) ;
   VALEUR T ; REEL T ;
      DEBUT
   REEL S ;
  ENTIER I; TABLEAU Y[1:N] ;
   CAL(T,Y) ;  S:=0.0 ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE
   S:=S+X[I]*Y[I] ;
    ER:=F(T) - S ;
   FIN   ER ;

  TABLEAU ERST[1:N+1,1:N+1] ,ZWEI[1:N+1] ;
```

```
      TABLEAU PREM[1:N,1:N],MU[1:N+2],SEC,LAMBAUX[1:N],
         LAMBDA[1:N+1],G[1:N] ;
      BOOLEEN TABLEAU BOOL[1:N+1] ;
      ENTIER I,J,J0 ; REEL R,AMPLI,MP,MQ,I0,P0,Q0 ;
      REEL S ;

   INIT: POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
      DEBUT
         REF[I]:=AI[1]+(I-0.5)*(AS[1]-AI[1])/(N+1) ;
         BOOL[I]:= VRAI
      FIN ;
   BUL:= VRAI ;
      POUR J:=1 PAS 1 JUSQUA N FAIRE
      DEBUT
         CAL(REF[J],G) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE PREM[I,J]:=G[I]
      FIN ;
         CAL(REF[N+1],G) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE
         SEC[I]:=-G[I] ;
      GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,END) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE
         LAMBDA[I]:=LAMBAUX[I] ;
         LAMBDA[N+1]:=1.0 ;
           S:=F(REF[N+1]) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE
           S:= S +LAMBDA[I]*F(REF[I]) ;
      SI  S < 0 ALORS
      POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
      LAMBDA[I]:=-LAMBDA[I] ;

    BOUCLE:
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
      DEBUT CAL(REF[I],G) ;
      POUR J:=1 PAS 1 JUSQUA N FAIRE ERST[I,J]:=G[J] ;
      FIN ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   ERST[I,N+1] := SI BOOL[I] ALORS
( SI LAMBDA[I] > 0 ALORS 1.0 SINON -1.0)
              SINON 0.0 ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   ZWEI[I]:=F(REF[I]) ;
    GRESOLSYSLINE(ERST,ZWEI,X,N+1,END) ;
    MP:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA NA FAIRE
      DEBUT
ABSMAX(ER,AI[I],AS[I],R,S) ;
   SI  S > MP ALORS
      DEBUT
     MP:=S ;  P0:=R ;
      FIN ;
      FIN ;
    MQ:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA NC FAIRE
      DEBUT
```

```
   ALGMIN(ER,CI[I],CS[I],R,S) ;
    SI S < MQ ALORS
      DEBUT
   MQ:=S ;Q0:=R ;
      FIN ;
      FIN ;
I0:= SI MP-X[N+1] > -MQ ALORS P0 SINON Q0 ;
      AMPLI:= SI MQ < 0 ALORS MP-MQ SINON MP ;
SI 1-X[N+1]/AMPLI INFEG EPS ALORS ALLERA  END ;
 I:=0 ;
SU : I:=I+1 ;
 SI ABS(I0-REF[I]) INFEG 0.0001 ALORS
DEBUT BUL:= FAUX ; ALLERA  END FIN ;
SI I < N+1 ALORS ALLERA  SU  ;
      SI I0=P0 ALORS
      DEBUT CAL(P0,G) ; R:=0.0 ;
         POUR I:=1 PAS 1 JUSQUA N FAIRE R:=R+X[I]*G[I] ;
      MU[N+2]:= SI F(P0)-R > 0 ALORS 1.0 SINON -1.0 ;
      FIN
         SINON
      MU[N+2]:=-1.0 ;
      POUR J:=1 PAS 1 JUSQUA N FAIRE
      DEBUT
         CAL(REF[J],G) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE PREM[I,J]:=G[I]
      FIN ;
         CAL(I0,G) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE
         SEC[I]:=-MU[N+2]*G[I] ;
      GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,END) ;
      POUR I:=1 PAS 1 JUSQUA N FAIRE MU[I]:=LAMBAUX[I] ;
      MU[N+1]:=0.0 ;
      J0:=1 ;
      POUR I:=2 PAS 1 JUSQUA N+1 FAIRE
      J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
      ALORS I SINON J0 ;
     POUR I:=1 PAS 1 JUSQUA J0-1,J0+1 PAS 1 JUSQUA N+1 FAIRE
      LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
      LAMBDA[J0]:=MU[N+2] ;
      POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
      REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
     SI I0=P0 ALORS BOOL[J0]:= VRAI SINON BOOL[J0]:= FAUX ;
      ALLERA  BOUCLE ;
END:
      POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
 E[I]:=ER(REF[I]) ;
FIN TCHEBINEG2 ;
```

CHAPITRE - VIII

APPROXIMATION UNIFORME DE DEUX FONCTIONS CONTINUES SUR UN COMPACT

Ce chapitre est consacré à l'étude du problème PIIIC défini au chapitre V.

Le plan de ce chapitre est le même que celui du chapitre VI, l'algorithme établi au deuxième paragraphe est une généralisation de l'algorithme de Remès, faisant appel à une méthode de perturbation tout-à-fait semblable à celle utilisée au chapitre IV.

Une procédure Algol et un exemple numérique sont proposés au troisième paragraphe.

§1 - APPLICATION DES METHODES DE PROGRAMMATION CONVEXE AU PROBLEME PIIIC

1 - Existence d'une solution.

Proposition 8-1

"Le problème PIIIC a au moins une solution"

Cette proposition se démontre de la même manière que la proposition 6-1-, à la variante près suivante : à un vecteur $x^o \in$ DIII, correspond le vecteur $g^o \in$ V tel que :

$$\text{Max}(\|f_1 - g^o\| , \|f_2 - g^o\|) = x^o_{n+1}.$$

2 - Caractérisation d'une solution.

Notations

Etant donné un vecteur $x \in$ DIII, on définit les ensembles suivants :

$$\mathcal{E}^+(x) = \{t \in S : f_1(t) - \sum_{i=1}^{n} x_i\, g_i(t) = x_{n+1}\}$$

$$\mathcal{E}^-(x) = \{t \in S : f_2(t) - \sum_{i=1}^{n} x_i\, g_i(t) = -x_{n+1}\}$$

et les fonctions erreurs e_1 et $e_2 \in \mathcal{C}(S)$:

$$e_1(t) = f_1(t) - \sum_{i=1}^{n} x_i\, g_i(t) \qquad \forall\, t \in S$$

$$e_2(t) = f_2(t) - \sum_{i=1}^{n} x_i\, g_i(t) \qquad \forall\, t \in S$$

On remarque :

$$\mathcal{E}^+(x) = \{t \in S : e_1(t) = \|e_1\| \}$$
$$\mathcal{E}^-(x) = \{t \in S : e_2(t) = -\|e_2\| \}$$

L'ensemble $\mathcal{E}^+(x) \cup \mathcal{E}^-(x)$ représente l'ensemble des points extrémaux.

Proposition 8-2

"Une condition nécessaire et suffisante pour que $g^* \in V$ soit meilleure approximation de f_1 et de f_2 est qu'il existe k points appartenant à $\mathcal{E}^+(x^*) \cup \mathcal{E}^-(x^*)$,

$t_1, t_2, \ldots, t_k \qquad (1 \leqslant k \leqslant n+1)$

et k coefficients λ_j tels que :

$$1 - \sum_{j=1}^{k} \lambda_j\, g_i(t_j) = 0 \qquad i = 1 \ldots n$$

$$2 - \lambda_j > 0 \text{ si } t_j \in \mathcal{E}^+(x^*)$$
$$\lambda_j < 0 \text{ si } t_j \in \mathcal{E}^-(x^*)$$

$$3 - \sum_{j=1}^{k} |\lambda_j| = 1 \qquad "$$

Cette proposition découle directement de la proposition 5-1.

3 - Unicité de la solution.

Proposition 8-3

"Les hypothèses

HL : "V vérifie la condition de Haar sur S"

H2 : "$\mathcal{E}^+(x) \cap \mathcal{E}^-(x) = \emptyset \quad \forall\, x \in DIII$ "

entraînent l'unicité de la solution".

Cette proposition se démontre de la même manière que la proposition 6-3.

Remarque 1 :

L'hypothèse H2 est évidemment rarement vérifiée dans la pratique ; en fait cette hypothèses n'est pas nécessaire, mais elle permet de simplifier et de clarifier notre exposé. Nous verrons au deuxième paragraphe que l'on peut se contenter d'une hypothèse moins restrictive.

Remarque 2 :

Soit x^* une solution, et $M^* = (t_1, t_2, \ldots, t_{n+1})$ l'ensemble des points extrémaux servant à caractériser le meilleur approximant g^* correspondant.
Supposons qu'on ait :

$$t_1 = t_2,\ t_1 \in \mathcal{E}^+(x^*),\ t_2 \in \mathcal{E}^-(x^*)$$

Les autres t_j, $j = 3 \ldots n+1$ étant distincts.
Les λ_j associés à M^* (cf. proposition 8-2) sont alors tels que :

$$\begin{cases} \lambda_1 = -\lambda_2 = \frac{1}{2} \\ \text{et} \quad \lambda_j = 0 \quad j = 3 \ldots n+1. \end{cases}$$

on remarque que :

$$\rho = \frac{1}{2} \left(f_1(t_1) - f_2(t_2) \right)$$

et il est clair que dans ce cas, il n'y a pas unicité de la solution.

§2 - ALGORITHMES DE RESOLUTION

Comme au chapitre IV, nous donnons successivement deux algorithmes ; dans une première partie, nous exposons un algorithme "restreint", valable dans le cadre des hypothèses H1 et H2, et qui est une généralisation simple de l'algorithme de Remès. Dans la deuxième partie, l'utilisation d'une méthode de perturbation nous permet d'établir un algorithme généralisé valable dans le cadre de l'hypothèses H1 et de l'hypothèse H3 peu restrictive suivante :

H3 : "il n'existe pas de couple (t,τ), $t \in S$, $\tau \in S$ tel que : $f_1(t) - f_2(t) = f_1(\tau) - f_2(\tau)$.

1 - Algorithme restreint

Dans toute la suite nous faisons les hypothèses H1 et H2.

a) Description de l'algorithme

α - Meilleure approximation sur un ensemble M_k de n+1 points de S.

$$M_k = \{t_j^k,\ j = 1...n+1\},\quad M_k \subset S.$$

On définit la meilleure approximation g^k sur M_k :

$$\text{Max}\left[\text{Max}\left(|f_1(t)-g^k(t)|\ \middle|\ t \in M_k\right),\ \text{Max}\left(|f_2(t)-g^k(t)|\ \middle|\ t \in M_k\right)\right] =$$

$$\underset{g \in V}{\text{Inf}}\left\{\text{Max}\left[\text{Max}\left(|f_1(t)-g(t)|\ \middle|\ t \in M_k\right),\ \text{Max}\left(|f_2(t)-g(t)|\ \middle|\ t \in M_k\right)\right]\right\} = \rho_k$$

Soient λ_j^k, j = 1...n+1, les coefficients associés à M_k, (cf. proposition 8-2), les λ_j^k doivent vérifier :

$$(R)\begin{cases} \sum_{j=1}^{n+1} \lambda_j^k\, g_i(t_j^k) = 0 \qquad i = 1...n \\ \\ \sum_{j=1}^{n+1} |\lambda_j^k| = 1 \end{cases}$$

Les relations (R) définissent les λ_j^k au signe près, la condition suivante, correspondant à la condition 2. de la proposition 8-2 permet de les déterminer complètement :

on note :

$$E_+^k = \{j \in (1...n+1) : \lambda_j^k > 0\}$$

$$E_-^k = \{j \in (1...n+1) : \lambda_j^k < 0\}$$

et on doit avoir :

$$\sum_{j \in E_+^k} \lambda_j^k\, f_1(t_j^k) + \sum_{j \in E_-^k} \lambda_j^k\, f_2(t_j^k) > 0$$

(En effet on montre facilement, de la même façon qu'en 6-2-1, que :

$$\rho_k = \sum_{j \in E_+^k} \lambda_j^k f_1(t_j^k) + \sum_{j \in E_-^k} \lambda_j^k f_2(t_j^k))$$

On obtient alors les coefficients x_i^k (i = 1...n) de la meilleure approximation g^k sur M_k, et l'écart $x_{n+1}^k = \rho_k$ correspondant par résolution du système :

$$\begin{cases} \sum_{i=1}^{n} x_i\, g_i(t_j^k) + x_{n+1} = f_1(t_j^k) & j \in E_+^k \\ \sum_{i=1}^{n} x_i\, g_i(t_j^k) - x_{n+1} = f_2(t_j^k) & j \in E_-^k \end{cases}$$

Remarque

Comme en 6-2-1, on montre facilement que si g^k n'est pas la meilleure approximation g^*, on a $\rho_k < \rho$ et que de plus si $\rho_k = \rho$, g^k ne peut être que la meilleure approximation g^* d'après la proposition 8-2.

β - Itération

On remplace M_k par M_{k+1}, ne différant de M_k que par un point, de telle façon que :

$$\rho_{k+1} > \rho_k.$$

(ceci dans le cas où g^k n'est pas déjà la meilleure approximation g^*).

1 - On choisit tout d'abord le point $\tau \in S$ qui va être introduit :

Soient e_1^k et e_2^k les fonctions erreurs associées à g^k, on détermine, τ_1 et τ_2 tels que :

$$e_1^k(\tau_1) = \mathrm{Max}(e_1^k(t) \,|\, t \in S)$$

$$e_2^k(\tau_2) = \mathrm{Min}(e_2^k(t) \,|\, t \in S)$$

et on choisit $\tau = \tau_1$ si $e_1^k(\tau_1) > -e_2^k(\tau_2)$ et $\tau = \tau_2$ dans le cas contraire.

2 - On choisit ensuite parmi les points t_j^k le point t_μ^k qui est supprimé, le nouvel ensemble M_{k+1} étant alors défini par :

$$\begin{cases} t_j^{k+1} = t_j^k & \forall j = 1...n+1,\ j \neq \mu \\ t_\mu^{k+1} = \tau \end{cases}$$

On choisit μ de telle façon que les coefficients λ_j^{k+1} associés à M_{k+1} vérifient les conditions (C) suivantes :

$$(C) \begin{cases} \text{signe } (\lambda_j^{k+1}) = \text{signe } (\lambda_j^k) \quad j = 1...n+1,\ j \neq \mu \\ \text{signe } (\lambda_\mu^{k+1}) = \begin{cases} 1 \text{ si } \tau = \tau_1 \\ -1 \text{ si } \tau = \tau_2 \end{cases} \end{cases}$$

Pour l'obtention en pratique des conditions (C) on opère d'une façon analogue à ce qui a été fait en 6-2-1.

b - Convergence de l'algorithme.

Proposition 8-4

"La suite ρ_k est strictement croissante"

on a :

$$\rho_{k+1} = \sum_{j \in E_+^{k+1}} \lambda_j^{k+1} f_1(t_j^{k+1}) + \sum_{j \in E_-^{k+1}} \lambda_j^{k+1} f_2(t_j^{k+1})$$

D'une façon analogue à ce qui a été fait en 6-2-2, et en vertu des conditions (C), on montre que :

$$\rho_{k+1} = \rho_k + |\lambda_\mu^{k+1}| \cdot \text{Max}(\|e_1^k\| - \rho_k, \|e_2^k\| - \rho_k)$$

Comme λ_μ^{k+1} ne peut être nul, on a : $\rho_{k+1} > \rho_k$, sauf si $\|e_1^k\| = \|e_2^k\| = \rho_k$, ce qui entraînerait alors $g^k = g^*$.

Proposition 8-5

"La suite g^k converge uniformément vers la meilleure approximation g^*"

La proposition 6-5 s'applique directement, et on montre de la même façon qu'en 6-2-2 que la suite convergente ρ_k a pour limite ρ, et ensuite que g^k converge uniformément vers g^*.

2 - Algorithme généralisé

Dans toute la suite nous faisons les hypothèses H1 et H3.

Notations

On dit qu'un ensemble M_k est dégénéré, s'il y figure deux fois le même point ; sans restriction de généralité, nous supposerons que ces points sont les deux premiers et nous noterons :

$$M_k = (t_1^k, t_2^k, \ldots, t_{n+1}^k)$$

avec $t_1^k = t_2^k$, les t_j^k, $j = 3 \ldots n+1$, étant distincts et nous considérerons t_1^k comme point extrémal positif et t_2^k comme point extrémal négatif ($t_1^k \in \mathcal{E}^+(x^k)$, $t_2^k \in \mathcal{E}^-(x^k)$).
On désigne par $G(t)$ le vecteur de R^n ayant pour composantes :

$$g_i(t) \quad i = 1 \ldots n.$$

(a) Principe de la méthode.

Ayant supprimé l'hypothèses H2, l'algorithme restreint peut nous conduire à un ensemble M_k dégénéré. On sait alors (cf. 8-1-3) que le meilleur approximant g^k sur M_k n'est pas unique, et l'algorithme restreint ne nous permet pas de conclure.

L'algorithme généralisé va nous permettre de passer d'un ensemble M_k dégénéré à un ensemble M_{k+1} tel que :

$$\rho_{k+1} > \rho_k$$

(à moins que g^k ne soit déjà une meilleure approximation g^*)

Nous utilisons une méthode de perturbation, tout à fait semblable à celle utilisée au chapitre IV, qui à un ensemble M_k dégénéré associe l'ensemble M'_k non dégénéré déduit de M_k par le remplacement de t_1^k en t'^k_1 tel que :

$$(T)\begin{cases} G(t'^k_1) = G(t_1^k) - \varepsilon\left[c_3^k\, G(t_3^k) + \ldots + c_{n+1}^k\, G(t_{n+1}^k)\right] \\ f_1(t'^k_1) = f_1(t_1^k) \end{cases}$$

Les c_j^k, $j = 3...N+1$, sont des constantes non nulles, ε un scalaire positif.

On a :

$$M'_k = (t'^k_1, t^k_2, \ldots, t^k_{n+1})$$

En notant :

$$E'^k_+ = \{j : c^k_j > 0\}$$

$$E'^k_- = \{j : c^k_j < 0\}$$

On montre, de la même façon qu'en 4-2-2, que :

$$e^k_1(t^k_1) = \rho_k \cdot \text{signe}(\lambda^k_1)$$
$$e^k_2(t^k_2) = \rho_k \cdot \text{signe}(\lambda^k_2)$$
$$e^k_1(t^k_1) = \rho_k \cdot \text{signe}(\lambda^k_1 c^k_j) \qquad j \in E'^k_+$$
$$e^k_2(t^k_2) = \rho_k \cdot \text{signe}(\lambda^k_1 c^k_j) \qquad j \in E'^k_-$$

En modifiant le signe des c^k_j, on est donc en mesure d'obtenir 2^{n-1} meilleurs approximants sur M_k.

(b) Description de l'algorithme généralisé.

Nous employons la même terminologie qu'en 4-2-2. Précisons les différentes phases de l'algorithme :

(I) Début d'étape : on dispose d'un ensemble M_k.

$$M_k = (t^k_1, \ldots, t^k_{n+1})$$

Si M_k est non dégénéré, on effectue l'échange comme il est exposé dans l'algorithme restreint. Si M_k est dégénéré, on choisit arbitrairement n-1 coefficients auxiliaires non nuls $c^k_3, \ldots, c^k_{n+1}$.

On calcule :

$$\rho_k = \frac{1}{2}(f_1(t^k_1) - f_2(t^k_2))$$

et on passe à la situation (II).

(II) on note :

$$E'^{k}_{+} = \{j : c^{k}_{j} > 0\}$$
$$E'^{k}_{-} = \{j : c^{k}_{j} < 0\}$$

et on résout le système :

$$\begin{cases} \sum_{i=1}^{n} x_i \, g_i(t^k_j) = -\rho_k + f_1(t^k_j) & j \in E'^{k}_{+} \\ \sum_{i=1}^{n} x_i \, g_i(t^k_j) = \rho_k + f_2(t^k_j) & j = \{2\} \cup E'^{k}_{-} \end{cases}$$

La solution x^k_i (i = 1...n) de ce système nous donne un meilleur approximant g^k sur M_k.

On calcule τ_1 et τ_2 comme il a été fait dans l'algorithme restreint, et on détermine le point τ entrant. (Si $\|e^k_1\| = \|e^k_2\| = \rho_k$, g^k est une solution du problème et on arrête l'algorithme). Pour déterminer le point t^k_μ sortant, on opère également de la même façon que dans l'algorithme restreint :

On détermine des coefficients auxiliaires ω_j (j = 1...n+2), tels que :

$$\begin{cases} \omega_1 = 0 \\ \omega_{n+2} = 1 \text{ si } \tau = \tau_1 \\ \omega_{n+2} = -1 \text{ si } \tau = \tau_2 \\ \sum_{j=2}^{n+2} \omega_j \, G(t^k_j) = 0 \end{cases}$$

On considère ensuite les quantités :

$$Q = \text{Min}\left(\frac{\omega_j}{c^k_j} \;\middle|\; j = 3...n+1\right), \quad P = \text{Min}\left(\frac{\omega_j}{\lambda^k_j} \;\middle|\; j = 1,2\right)$$

Deux cas peuvent alors se présenter :

Si Q < 0 on passe à la situation (III) (échange statique).

Si Q > 0 on passe à la situation (IV) (échange non statique).

(III) Soit : $Q = \frac{\omega_\mu}{c^k_\mu}$

On obtient le nouvel ensemble M_{k+1} en échangeant τ et t^k_μ, et on retourne à la situation (II) avec les coefficients c^{k+1}_j tels que :

$$\begin{cases} c^{k+1}_j = c^k_j - \omega_j \cdot \frac{c^k_\mu}{\omega_\mu} & j = 3\ldots n+1,\ j \neq \mu \\ c^{k+1}_\mu = -\frac{c^k_\mu}{\omega_\mu} \end{cases}$$

(λ^{k+1}_1, λ^{k+1}_2 étant restectivement égaux à λ^k_1 et λ^k_2, et ρ_{k+1} étant égal à ρ_k).

Si l'un (ou plusieurs) des c^{k+1}_j s'annulent, nous avons un "accident", et nous le remplaçons par un nombre pris au hasard.

(IV) Soit $P = \frac{\omega_2}{\lambda^k_2}$

on échange t_2 et τ, et l'on retourne à la situation (I) avec l'ensemble M_{k+1} :

$$M_{k+1} = (t_1, \tau, t_3, \ldots, t_{n+1}).$$

(c) Justification de l'algorithme généralisé.

Plaçons nous au début d'une étape avec M_k dégénéré. On a déjà justifié en (a), lors de l'exposé du principe de la méthode, les situations (I) et (II), il nous reste à justifier les situations (III) et (IV).

Soient M'_k l'ensemble déduit de M_k par la transformation (T), et ω_j les coefficients auxiliaires définis en (II) ; l'ensemble M'_k étant non dégénéré, on opère l'échange comme il a été exposé dans l'algorithme restreint, et on calcule :

$$K = \text{Min}\left[\left(\frac{\omega_j}{\lambda^k_j} \mid j = 1,2\right),\ \left(\frac{\omega_j}{\varepsilon\lambda^k_1 c^k_j} \mid j = 3\ldots n+1\right)\right]$$

1er cas.

$$Q = \text{Min}\left(\frac{\omega_j}{c^k_j} \mid j = 3\ldots n+1\right) < 0$$

alors il existe ε_1 tel que $K = \frac{1}{\varepsilon\lambda^k_1} Q$ pour tout $\varepsilon < \varepsilon_1$, et en posant $Q = \frac{\omega_\mu}{c^k_\mu}$, le nouvel ensemble M''_k est tel que :

$$M''_k = (t'^k_1,\ t^k_2,\ldots,t^k_{\mu-1},\tau,\ t^k_{\mu+1},\ldots,t^k_{n+1})$$

En supposant, sans restriction de généralité que $\tau=\tau_1$, c'est-à-dire que $\omega_{n+2}=1$, les ω_j vérifient :

$$\sum_{j=1}^{n+1} \omega_j\ G(t^k_j) + G(\tau) = 0 \qquad (1)$$

Par ailleurs, on a :

$$\lambda^k_1\ G(t'^k_1) + \lambda^k_2\ G(t^k_2) + \varepsilon\lambda^k_1 \sum_{j=3}^{n+1} c^k_j\ G(t^k_j) = 0 \qquad (2)$$

en multipliant (1) par $\varepsilon\lambda^k_1\ \dfrac{c^k_\mu}{\omega_\mu}$ et en retranchant de (2), il vient :

$$\lambda^k_1\ G(t'^k_1) + (\lambda^k_2+O(\varepsilon))\ G(t^k_2) + \varepsilon\lambda^k_1 \sum_{\substack{j=3\\ j\neq\mu}}^{n+1} \left(c^k_j-\omega_j\ \frac{c^k_\mu}{\omega_\mu}\right) G(t^k_j) = \varepsilon\lambda^k_1\ \frac{c^k_\mu}{\omega_\mu}\ G(\tau)$$

La fonction $O(\varepsilon)$ tendant vers 0 quand $\varepsilon\to 0$.

En posant :

$$\begin{cases} c^{k+1}_j = c^k_j - \omega_j\ \dfrac{c^k_\mu}{\omega_\mu} & j = 3\ldots n+1,\ j \neq \mu \\ c^{k+1}_\mu = -\dfrac{c^k_\mu}{\omega_\mu} \end{cases}$$

on a bien :

$$\text{signe}\ (c^{k+1}_j) = \text{signe}\ (c^k_j) \qquad j = 3\ldots n+1,\ j \neq \mu$$
$$\text{et}\ c^{k+1}_\mu > 0$$

Les conditions (C) de l'algorithme restreint sont remplies et par conséquent, l'erreur ρ''_k de la meilleure approximation sur M''_k sera telle que :

$$\rho''_k > \rho'_k \text{ (cf. proposition 8-4).}$$

(ρ'_k étant l'erreur de la meilleure approximation sur M'_k)

Remarque

Ayant fait l'hypothèse H3, les c^{k+1}_j ne sont en général pas nuls, le cas accidentel où l'un de ces coefficients est nul sera examiné plus loin.

Nous pouvons recommencer les mêmes calculs avec M_k'', une suite d'échanges correspondant au premier cas fournit des ensembles M_k', M_k'',...,$M_k^{(r)}$, tels que :

$$\rho_k' < \rho_k'' < \ldots < \rho_k^{(r)}$$

On vérifie aisément qu'il existe ε_r, tel que pour tout $\varepsilon < \varepsilon_r$, les échanges ainsi obtenus sont identiques à ceux que l'on trouve décrit dans l'algorithme généralisé, et que les différentes approximations trouvées par cet algorithme sont la limite de celles du problème perturbé lorsque ε tend vers zéro ; on vérifie également que l'on a des échanges statiques, c'est-à-dire que :

$$\rho_k = \lim_{\varepsilon \to 0} \rho_k' = \lim_{\varepsilon \to 0} \rho_k'' = \text{---}$$

2ème cas

$$Q = \text{Min}\left(\frac{\omega_j}{c_j^k} \mid j = 3\ldots n+1\right) > 0.$$

Soit alors, $P = \text{Min}\left(\frac{\omega_j}{\lambda_j^k} \mid j = 1,2\right) = \frac{\omega_2}{\lambda_2^k}$, le nouvel ensemble M_{k+1} est tel que :

$$M_{k+1} = (t_1^k, \tau, t_3^k, \ldots, t_{n+1}^k)$$

Soit ρ_{k+1} l'erreur d'une meilleure approximation sur M_{k+1}, on a :

$$\rho_{k+1} > \rho_k'$$

il est facile de vérifier que cette stricte inégalité subsiste quand $\varepsilon \to 0$; l'échange est alors non statique.

Cas d'accident.

Il peut arriver qu'au cours d'une série d'échanges statiques, l'un des coefficients auxiliaires s'annule. Soient c_j^k (j = 3...n+1) les coefficients auxiliaires initiaux et c_j^r (j = 3...n+1) les coefficients auxiliaires après q = r-k échanges.
On montre, de la même façon qu'au chapitre IV, que les c_j^r sont des formes homogènes indépendantes dans les variables c_j^k.

Supposons alors que pour des valeurs particulières des coefficients initiaux, on ait :

$$c_j^r \neq 0,\ j = 3...n+1,\ j \neq p,\ \text{et}\ c_p^r = 0$$

Le système d'équations :

$$\begin{cases} c'^r_j = c^r_j & j = 3...n+1,\ j \neq p \\ c'^r_p = \eta \end{cases}$$

possède une et une seule solution dans les variables c_j^k ; pour η assez petit, les q échanges ne sont pas modifiés et l'algorithme peut continuer. Nous avons ainsi démontré qu'il existe des valeurs initiales c_j^k qui permettent d'achever une étape et ces considérations justifient le terme d'accident.

§3 - PROCEDURE ALGOL ET ASPECTS NUMERIQUES

1 - Notice d'emploi de la procédure TCHEBDEUXFØNC2 résolvant le problème PIIIC.

TCHEBDEUXFØNC2(AI,AS,NA,N,F1,F2,CAL,EPS,DEV,X,REF,BUL,IMPØS).

Les fonctions F1 et F2 (précisées par les réelles procédures F1 et F2) sont définies sur le compact S de $\mathbb{R}$:

$$S = \bigcup_{I=1}^{NA} [AI[I],\ AS[I]].$$

La variété V de dimension N est précisée par la procédure CAL ; l'appel CAL(T,G) affecte à G[I], I = 1...N, les valeurs $g_I(T)$, pour $T \in S$.
La procédure calcule les coefficients X[I], I = 1...N, du meilleur approximant de F1 et de F2 sur S, et l'écart correspondant DEV :

$$DEV = Max\left[Max_{T\in S}(|F1(T) - \sum_{I=1}^{N} X[I]\times g_I(T)|),\ Max_{T\in S}(|F2(T) - \sum_{I=1}^{N} X[I]\times g_I(T)|)\right]$$

La procédure comporte deux tests d'arrêt tout-à-fait identiques à ceux de la procédure TCHEBCØNT2 (cf. 7-3-1).
Si l'algorithme rencontre au cours des itérations une référence REF_k, telle qu'il existe $\alpha,\beta,\gamma,\delta \in REF_k$ tels que :

$\alpha = \beta$

$\gamma = \delta$

alors on sort de la procédure par l'étiquette IMPØS. (c'est le cas où l'hypothèse H3 (cf. 8-2) n'est pas réalisée).

Enfin, la procédure range dans le tableau REF[1 : N+1] les abscisses des points de la référence finale.

(On suppose que V vérifie la condition de Haar).

2 - Exemple numérique.

Considérons l'approximation des fonction $f_1(x) = \exp(x)$ et $f_2(x) = \sqrt{x}$ sur $[0 ; 2]$ par un polynôme de degré 4. On a choisi EPS = 10^{-2} ; on trouve :

DEV = 2,98742
BUL = VRAI

X[1]= 2,98742	REF[1]= 2,00000
X[2]= -30, 38956	REF[2]= 2,00000
X[3]= 63,37348	REF[3]= 0
X[4]= -41,98827	REF[4]= 1,38701
X[5]= 9,03785	REF[5]= 1,27322
	REF[6]= 0,37759

On remarque que le point d'abscisse 2 figure deux fois dans la référence finale, et par conséquent, il y a non unicité de la solution.

Il y a eu 9 itérations, le temps de calcul et d'impression des résultats a été de 1 minute et 16 secondes (sur IBM 7044).

```
   PROCEDURE TCHEBDEUXFONC2(AI,AS,NA,N,F1,F2,CAL,EPS,
   DEV,X,REF,BUL,IMPOS) ;
   VALEUR NA,N,EPS ;
   ENTIER NA,N ; REEL EPS,DEV ;
  BOOLEEN BUL ;
   PROCEDURE CAL ; REEL PROCEDURE F1,F2 ;
   TABLEAU X,REF,AI,AS ;
   ETIQUETTE IMPOS ;

DEBUT

    REEL PROCEDURE HASARD(NO) ; ENTIER NO ;
      COMMENTAIRE  CETTE PROCEDURE DONNE UN NOMRE REEL ALEATOIRE
      DE (0,1) AVEC UNE DENSITE UNIFORME.
      NO SERA INITIALISE A UNE VALEUR ENTIERE IMPAIRE (UNE FOIS
      POUR TOUTE),CHAQUE APPEL DE HASARD(NO) DEFINIT UN REEL DE
      (0,1) ET PROVOQUE UN EFFET DE BORD SUR NO .
      LE CORPS DE CETTE PROCEDURE EST ECRIT EN CODE .


'PROCEDURE'GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE); .....
           VOIR  ANNEXE  .....

PROCEDURE ALGMAX(E,A,B,XO,MAX) ;  .....
     VOIR ANNEXE .....

PROCEDURE ALGMIN(E,A,B,XO,MIN) ;  .....
     VOIR ANNEXE .....

   REEL PROCEDURE ER1(T) ; VALEUR T ; REEL T ;
DEBUT REEL R ; TABLEAU G[1:N] ; ENTIER J ;
   CAL(T,G) ; R:=0.0 ;
   POUR J:=1 PAS 1 JUSQUA N FAIRE
   R:=R+X[J]*G[J] ;
   ER1:=F1(T)-R ;
FIN ER1 ;


   REEL PROCEDURE ER2(T) ; VALEUR T ; REEL T ;
DEBUT REEL R ; TABLEAU G[1:N] ; ENTIER J ;
   CAL(T,G) ; R:=0.0 ;
   POUR J:=1 PAS 1 JUSQUA N FAIRE
   R:=R+X[J]*G[J] ;
   ER2:=F2(T)-R ;
FIN ER2 ;


   TABLEAU PREM[1:N,1:N],MU[1:N+2],SEC,LAMBAUX[1:N],
   LAMBDA[1:N+1],G[1:N],ROC[1:N+1] ;
   BOOLEEN BOOL,DEG ;
   REEL S,R,T,I0,I1,I2,M1,M2 ;
   ENTIER I,J,J0,N0,JDOU ;
INIT:
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   REF[I]:=AI[1]+(I-0.5)*(AS[1]-AI[1])/(N+1) ;
```

```
    NO:=3 ;
    BUL:= VRAI ;
ETAPE:
    POUR J:=1 PAS 1 JUSQUA N FAIRE
DEBUT
    CAL(REF[J],G) ;
    POUR I:=1 PAS 1 JUSQUA N FAIRE
    PREM[I,J]:=G[I] ;
FIN ;
    CAL(REF[N+1],G) ;
    POUR I:=1 PAS 1 JUSQUA N FAIRE
    SEC[I]:=G[I] ;
    GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,END) ;
    POUR I:=1 PAS 1 JUSQUA N FAIRE
    LAMBDA[I]:=LAMBAUX[I] ;
    LAMBDA[N+1]:=-1.0 ;
    DEV:=-F2(REF[N+1]) ;
    POUR I:=1 PAS 1 JUSQUA N FAIRE
    DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*F1(REF[I])
    SINON DEV+LAMBDA[I]*F2(REF[I]) ;
    SI DEV < 0 ALORS
DEBUT
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    LAMBDA[I]:=-LAMBDA[I] ;
    DEV:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*F1(REF[I])
    SINON DEV+LAMBDA[I]*F2(REF[I]) ;
FIN ;
    R:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    R:=R+ABS(LAMBDA[I]) ;
    DEV:=DEV/R ;
    DEG:= FAUX ;

BOUCLE:
    POUR I:=1 PAS 1 JUSQUA N FAIRE
DEBUT
    CAL(REF[I+1],G) ;
    POUR J:=1 PAS 1 JUSQUA N FAIRE
    PREM[I,J]:=G[J] ;
FIN ;
    POUR I:=1 PAS 1 JUSQUA N FAIRE
    SEC[I]:= SI LAMBDA[I+1] > 0 ALORS
    F1(REF[I+1])-DEV SINON F2(REF[I+1]) + DEV ;
    GRESOLSYSLINE(PREM,SEC,X,N,END) ;
    M1:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA NA FAIRE
DEBUT
    ALGMAX(ER1,AI[I],AS[I],R,S) ;
    SI S > M1 ALORS
DEBUT
    M1:=S ;
    I1:=R ;
    FIN ;
```

```
FIN ;
   M2:=0.0 ;
   POUR I:=1 PAS 1 JUSQUA NA FAIRE
DEBUT
   ALGMIN(ER2,AI[I],AS[I],R,S) ;
   SI S < M2 ALORS
DEBUT
   M2:=S ;
   I2:=R ;
   FIN ;
FIN ;
   BOOL:= SI M1 > -M2 ALORS VRAI SINON FAUX ;
   IO:= SI BOOL ALORS I1 SINON I2 ;
   SI BOOL ET (1-DEV/M1) < EPS ALORS ALLERA  END ;
   SI NON BOOL ET (1+DEV/M2) < EPS ALORS ALLERA  END ;
   POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
   DEBUT
 SI BOOL ET ABS(IO-REF[I]) < 0.0001 ET
  ABS(ER1(IO)-DEV) < 0.001 ALORS
  DEBUT BUL:= FAUX ; ALLERA  END  FIN ;
   SI NON BOOL ET ABS(IO-REF[I]) < 0.0001 ET
  ABS(-ER2(IO)-DEV) < 0.001 ALORS
  DEBUT BUL:= FAUX ; ALLERA  END  FIN ;
   FIN ;
   ETUDE DE DEGENERESCENCE:
   I:=0 ;
   ITER: I:=I+1 ;
  SI ABS(IO-REF[I]) < 0.0001 ET DEG ALORS
DEBUT
   JDOU:=I ;
   SI ABS(0.5*(F1(IO)-F2(IO))-DEV) < EPS ALORS ALLERA  IM
   SINON ALLERA  SAUT ;
FIN ;
  SI ABS(IO-REF[I]) < 0.0001 ALORS
DEBUT
   JDOU:=I ;
   DEG:= VRAI ;
   ALLERA  DIF ;
FIN ;
   SI I < N+1 ALORS ALLERA  ITER ;
   CRA:
   MU[N+2]:= SI BOOL ALORS 1.0 SINON -1.0 ;
   POUR J:=1 PAS 1 JUSQUA N FAIRE
DEBUT
   CAL(REF[J+1],G) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
   PREM[I,J]:=G[I] ;
FIN ;
   CAL(IO,G) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
   SEC[I]:=-MU[N+2]*G[I] ;
   GRESOLSYSLINE(PREM,SEC,LAMBAUX,N,END) ;
   POUR I:=1 PAS 1 JUSQUA N FAIRE
   MU[I+1]:=LAMBAUX[I] ;
   MU[1]:=0.0 ;
```

```
    SI NON DEG ALORS ALLERA  NORMAL ;
    J0:=3 ;
    POUR I:=4 PAS 1 JUSQUA N+1 FAIRE
    J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
    ALORS I SINON J0 ;
    T:=MU[J0]/LAMBDA[J0] ;
    SI T > 0 ALORS
DEBUT
    DEG:= FAUX ;
    J0:= SI MU[2]/LAMBDA[2] < 0 ALORS 2 SINON 1 ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
    ALLERA  ETAPE ;
FIN ;
    POUR I:=3 PAS 1 JUSQUA J0-1,J0+1 PAS 1 JUSQUA N+1 FAIRE
DEBUT
    LAMBDA[I]:=LAMBDA[I]-LAMBDA[J0]*MU[I]/MU[J0] ;
    SI ABS(LAMBDA[I]) < EPS ALORS
DEBUT
    B1: LAMBDA[I]:=HASARD(N0)*(-1)**I ;
    SI ABS(LAMBDA[I]) < EPS ALORS ALLERA  B1 ;
FIN ;
FIN ;
    LAMBDA[J0]:=-LAMBDA[J0]*MU[N+2]/MU[J0] ;
    POUR I:=3 PAS 1 JUSQUA N+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
    ALLERA  BOUCLE ;
NORMAL:
    J0:=1 ;
    POUR I:=2 PAS 1 JUSQUA N+1 FAIRE
    J0:= SI MU[I]/LAMBDA[I] < MU[J0]/LAMBDA[J0]
    ALORS I SINON J0 ;
    POUR I:=1 PAS 1 JUSQUA J0-1,J0+1 PAS 1 JUSQUA N+1 FAIRE
    LAMBDA[I]:=MU[I]-LAMBDA[I]*MU[J0]/LAMBDA[J0] ;
    LAMBDA[J0]:=MU[N+2] ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    REF[I]:= SI I ≠ J0 ALORS REF[I] SINON I0 ;
    DEV:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    DEV:= SI LAMBDA[I] > 0 ALORS DEV+LAMBDA[I]*F1(REF[I])
    SINON DEV+LAMBDA[I]*F2(REF[I]) ;
    R:=0.0 ;
    POUR I:=1 PAS 1 JUSQUA N+1 FAIRE
    R:=R+ABS(LAMBDA[I]) ;
    DEV:=DEV/R ;
    ALLERA  BOUCLE ;
DIF:
    I:=2 ;
    POUR I:=I+1 TANTQUE  I < JDOU+1 ET I < N+1 FAIRE
    ROC[I]:=REF[I-2] ;
    POUR J:=I PAS 1 JUSQUA N+1 FAIRE
    ROC[J]:=REF[J-1] ;
       ALLERA   OUF ;

    SAUT:
```

```
     POUR I:=3 PAS 1 JUSQUA JDOU FAIRE
        ROC[I]:=REF[I-1] ;
      POUR I:=JDOU+1 PAS 1 JUSQUA N+1 FAIRE
        ROC[I]:= REF[I] ;

   OUF:
    POUR I:=3 PAS 1 JUSQUA N+1 FAIRE
    REF[I]:=ROC[I] ;
    REF[1]:=10 ; REF[2]:=10 ;
    LAMBDA[1]:=0.5 ; LAMBDA[2]:=-0.5 ;
    POUR I:=3 PAS 1 JUSQUA N+1 FAIRE
DEBUT
    LAMBDA[I]:=HASARD(N0)*(-1)**I ;
    SI ABS(LAMBDA[I]) < EPS ALORS
DEBUT
    B2: LAMBDA[I]:=HASARD(N0)*(-1)**I ;
    SI ABS(LAMBDA[I]) < EPS ALORS ALLERA  B2 ;
    FIN ;
FIN ;
    DEV:=0.5*(F1(10)-F2(10)) ;
    ALLERA  BOUCLE ;
END: FIN TCHEBDEUXFONC2 ;
```

```
PROCEDURE GRESOLSYSLINE(A,B,X,N,IMPOSSIBLE);
       TABLEAU A,B,X; ENTIER N; ETIQUETTE IMPOSSIBLE;
DEBUT TRIANGULARISATION:
       DEBUT ENTIER I,J,K; REEL R;
                POUR K:=1 PAS 1 JUSQUA N-1 FAIRE
       DEBUT NORMAL:
     DEBUT SI ABS(A[K,K])=0 ALORS ALLERA  ECHANGE DE LIGNES;
                       POUR I:=K+1 PAS 1 JUSQUA N FAIRE
                           DEBUT R:=A[I,K]/A[K,K];
                                  POUR J:=K+1 PAS 1 JUSQUA N FAIRE
                                     A[I,J]:=A[I,J]-R*A[K,J];
                                     B[I]:=B[I]-R*B[K] FIN
                FIN ;
                ALLERA  RETOUR;
   ECHANGE DE LIGNES:
   DEBUT ENTIER M ;
      M:=K+1 ;
   RET: SI ABS(A[M,K])=0 ALORS M:=M+1
        SINON ALLERA  CONT ;
      SI M=N+1 ALORS ALLERA  IMPOSSIBLE SINON ALLERA  RET ;
   CONT: POUR J:=K PAS 1 JUSQUA N FAIRE
      DEBUT R:=A[K,J] ;
         A[K,J]:=A[M,J] ;
         A[M,J]:=R
      FIN ;
   R:=B[K] ; B[K]:=B[M] ; B[M]:=R ;
   ALLERA  NORMAL
   FIN ;
RETOUR: FIN FIN TRIANGULARISATION ;
   RESSYSTRI:
DEBUT ENTIER I,J; REEL TX;
       POUR I:=N PAS -1 JUSQUA 1 FAIRE
                DEBUT TX:=0;
                    POUR J:=N PAS -1 JUSQUA I+1 FAIRE
                         TX:=TX-X[J]*A[I,J];
                    SI A[I,I]=0 ALORS ALLERA  IMPOSSIBLE;
                    X[I]:=(B[I]+TX)/A[I,I] FIN
   FIN RESSYSTRI
FIN GRESOLSYSLINE;
```

```
PROCEDURE ALGMIN(E,A,B,X0,MIN) ;
   REEL PROCEDURE E ; REEL A,B,X0,MIN ;
DEBUT
      REEL S,RES,H0,H,V1,V2,V3,T0,T ; BOOLEEN BOL ;
   S:=E(A) ; RES:=A ; T:=A ;
   H0:=H:=(B-A)/100 ;
   REP: BOL:= FAUX ;
   DEP: V1:=E(T) ; V2:=E(T+H) ; V3:=E(T+2*H) ;
   ITER: SI V3-V2 > 0 ET V1-V2 > 0
   ALORS ALLERA  DIV ;
   SI T+2,5*H > B ALORS ALLERA  TEST ;
   T:=T+H ; V1:=V2 ; V2:=V3 ; V3:=E(T+2*H) ;
   ALLERA  ITER ;
   TEST: SI E(B) < S ALORS
      DEBUT S:=E(B) ; RES:=B FIN ;
      ALLERA  SORT ;
DIV: SI NON BOL ALORS T0:=T+H ;
   H:=H/2 ; BOL:= VRAI ;
   SI H/H0 > 0,001 ALORS ALLERA  DEP ;
   SI V2 < S ALORS
   DEBUT S:=V2 ; RES:=T+H FIN ;
   T:=T0 ; H:=H0 ; ALLERA  REP ;
   SORT: X0:=RES ; MIN:=S ;
FIN ALGMIN ;
```

****

```
PROCEDURE ALGMAX(E,A,B,X0,MAX) ;
   REEL PROCEDURE E ; REEL A,B,X0,MAX ;
DEBUT
      REEL S,RES,H0,H,V1,V2,V3,T0,T ; BOOLEEN BOL ;
   S:=E(A) ; RES:=A ; T:=A ;
   H0:=H:=(B-A)/100 ;
   REP: BOL:= FAUX ;
   DEP: V1:=E(T) ; V2:=E(T+H) ; V3:=E(T+2*H) ;
   ITER: SI V3-V2 < 0 ET V1-V2 < 0
   ALORS ALLERA  DIV ;
   SI T+2,5*H > B ALORS ALLERA  TEST ;
   T:=T+H ; V1:=V2 ; V2:=V3 ; V3:=E(T+2*H) ;
   ALLERA  ITER ;
   TEST: SI E(B) > S ALORS
      DEBUT S:=E(B) ; RES:=B FIN ;
      ALLERA  SORT ;
DIV: SI NON BOL ALORS T0:=T+H ;
   H:=H/2 ; BOL:= VRAI ;
   SI H/H0 > 0,001 ALORS ALLERA  DEP ;
   SI V2 > S ALORS
   DEBUT S:=V2 ; RES:=T+H FIN ;
   T:=T0 ; H:=H0 ; ALLERA  REP ;
   SORT: X0:=RES ; MAX:=S ;
FIN ALGMAX ;
```

```
PROCEDURE ABS MAX(E,A,B,X0,MAX) ;
   REEL PROCEDURE E ; REEL A,B,X0,MAX ;
DEBUT
      REEL S,RES,H0,H,V1,V2,V3,T0,T ; BOOLEEN BOL ;
   S:=ABS(E(A)) ; RES:=A ; T:=A ;
   H0:=H:=(B-A)/100 ;
   REP: BOL:= FAUX ;
   DEP: V1:=E(T) ; V2:=E(T+H) ; V3:=E(T+2*H) ;
   ITER: SI (V3-V2)*(V2-V1) < 0 ALORS ALLERA  DIV ;
   SI T+2,5*H > B ALORS ALLERA  TEST ;
   T:=T+H ; V1:=V2 ; V2:=V3 ; V3:=E(T+2*H) ;
   ALLERA  ITER ;
   TEST: SI ABS(E(B)) > S ALORS
   DEBUT
      S:=ABS(E(B)) ; RES:=B FIN ;
   ALLERA  SORT ;
DIV: SI NON BOL ALORS T0:=T+H ;
   H:=H/2 ; BOL:= VRAI ;
   SI H/H0 > 0,001 ALORS ALLERA  DEP ;
   SI ABS(V2) > S ALORS
   DEBUT S:=ABS(V2) ; RES:=T+H ; FIN ;
   T:=T0 ; H:=H0 ; ALLERA  REP ;
   SORT: X0:=RES ; MAX:=S ;
FIN ABS MAX ;
```

B I B L I O G R A P H I E

[1] J.R. BARRA — Problèmes d'extréma sous contraintes. Séminaire d'Analyse Fonctionnelle - Université de GRENOBLE -1966-

[2] J.R. BARRA, A. AUSLENDER — Programmation linéaire. Polycopié - Faculté des Sciences de Grenoble -1965-

[3] J.R. BARRA, M. LEMARIE — Polyèdres convexes. Polycopié - Faculté des Sciences de Grenoble -1966-

[4] N. BOURBAKI — Espaces vectoriels topologiques (Chapitres 1 et 2) (Eléments de Mathématiques Fascicule XV) -1966-

[5] J. DESCLOUX — Contribution au calcul des approximations de Tchebycheff (Thèse présentée à l'Ecole Polytechnique de Zurich) -1961-

[6] DUBOVICKII, MILYUTIN — Extremum problems with contraints. Soviet Mathematics -1963- Tome 4 - pp. 452-455

[7] H.G. EGGLESTON — Convexity- pp. 34-38 -1963-

[8] E.G. GOL'STEIN — An infinite dimensional analogue of the linear programming problem and its applications to some problems in the theory of approximation. Dokl - Akad-Navk SSSR 140, 23-26 -1961-

[9] A. HAAR — Die Minkowskische Geometrie und die Annährung stetiger Fonktionen. Math. Annalen 78, pp. 294-311 -1918-

[10] P.J. LAURENT — Théorie de l'approximation (Fascicule 1). Polycopié Faculté des Sciences de Grenoble -1964-

[11] LOBRY — Etude géométrique des problèmes d'optimisation en présence de contraintes. Thèse de 3ème Cycle - Faculté des Sciences de Grenoble -1967- à paraître

[12] E. REMES — Sur le calcul effectif des polynômes d'approximation de Tchebycheff. C.R. Acad. Sci. Paris 199, pp. 337-340 -1934-

[13] D.L. RUSSEL — The Kühn-Tücker conditions in Banach space with an application to Control Theory. Journal of Mathematical Analysis and Applications. Vol 15 N° 2 (August 1966)

[14] E. STIEFEL — Über diskrete und lineare Tchebycheff Approximationen. Num-Math 1, pp. 1-28 -1959-

[15] WOLFE — A technique for resolving degeneracy in linear programming. Rand Corporation -1963-

TABLE DES MATIERES

VU

Grenoble, le

Le Président de la Thèse

VU

Grenoble, le

Le Doyen de la Faculté des Sciences

VU, et permis d'imprimer,

Le Recteur de l'Académie de GRENOBLE