Quelques outils numériques pour la résolution de systèmes algébrodifférentiels de grande dimension : applications au projet CASCADE

Mariano Bona

▶ **To cite this version:**

Mariano Bona. Quelques outils numériques pour la résolution de systèmes algébrodifférentiels de grande dimension : applications au projet CASCADE. Modélisation et simulation. Université Joseph-Fourier - Grenoble I, 1983. Français. NNT : . tel-00308063

HAL Id: tel-00308063

https://theses.hal.science/tel-00308063

Submitted on 29 Jul 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

présentée à

l' Université Scientifique et Médicale de Grenoble

pour obtenir le grade de
DOCTEUR INGENIEUR

par

BONA Mariano

QUELQUES OUTILS NUMERIQUES POUR LA RESOLUTION DE SYSTEMES ALGEBRODIFFERENTIELS DE GRANDE DIMENSION.

APPLICATIONS AU PROJET CASCADE.

Thèse soutenue le 3 novembre 1983 devant la commission d'examen.

G. VEILLON	Président
J. DELLA-DORA	Examinateurs
M. HEYDEMANN	
C. LE FAOU	
J. MERMET	
F. ROBERT	
J.P. UHRY	

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

Année universitaire 1982-1983

Président de l'Université : M. TANCHE

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

(RANG A)

SAUF ENSEIGNANTS EN MEDECINE ET PHARMACIE

PROFESSEURS DE 1ère CLASSE

ARNAUD Paul	Chimie organique
ARVIEU Robert	Physique nucléaire I.S.N.
AUBERT Guy	Physique C.N.R.S.
AYANT Yves	Physique approfondie
BARBIER Marie-Jeanne	Electrochimie
BARBIER Jean-Claude	Physique expérimentale C.N.R.S. (labo de magnétisme)
BARJON Robert	Physique nucléaire I.S.N.
BARNOUD Fernand	Biosynthèse de la cellulose-Biologie
BARRA Jean-René	Statistiques - Mathématiques appliquées
BELORISKY Elie	Physique
BENZAKEN Claude (M.)	Mathématiques pures
BERNARD Alain	Mathématiques pures
BERTRANDIAS Françoise	Mathématiques pures
BERTRANDIAS Jean-Paul	Mathématiques pures
BILLET Jean	Géographie
BONNIER Jean-Marie	Chimie générale
BOUCHEZ Robert	Physique nucléaire I.S.N.
BRAVARD Yves	Géographie
CARLIER Georges	Biologie végétale
CAUQUIS Georges	Chimie organique
CHIBON Pierre	Biologie animale
COLIN DE VERDIERE Yves	Mathématiques pures
CRABBE Pierre (détaché)	C.E.R.M.O.
CYROT Michel	Physique du solide
DAUMAS Max	Géographie
DEBELMAS Jacques	Géologie générale
DEGRANGE Charles	Zoologie
DELOBEL Claude (M.)	M.I.A.G. Mathématiques appliquées
DEPORTES Charles	Chimie minérale
DESRE Pierre	Electrochimie
DOLIQUE Jean-Michel	Physique des plasmas
DUCROS Pierre	Cristallographie
FONTAINE Jean-Marc	Mathématiques pures
GAGNAIRE Didier	Chimie physique

.../...

GASTINEL Noël	Analyse numérique - Mathématiques appliquées
GERBER Robert	Mathématiques pures
GERMAIN Jean-Pierre	Mécanique
GIRAUD Pierre	Géologie
IDELMAN Simon	Physiologie animale
JANIN Bernard	Géographie
JOLY Jean-René	Mathématiques pures
JULLIEN Pierre	Mathématiques appliquées
KAHANE André (détaché DAFCO)	Physique
KAHANE Josette	Physique
KOSZUL Jean-Louis	Mathématiques pures
KRAKOWIAK Sacha	Mathématiques appliquées
KUPTA Yvon	Mathématiques pures
LACAZE Albert	Thermodynamique
LAJZEROWICZ Jeannine	Physique
LAJZEROWICZ Joseph	Physique
LAURENT Pierre	Mathématiques appliquées
DE LEIRIS Joël	Biologie
LLIBOUTRY Louis	Géophysique
LOISEAUX Jean-Marie	Sciences nucléaires I.S.N.
LOUP Jean	Géographie
MACHE Régis	Physiologie végétale
MAYNARD Roger	Physique du solide
MICHEL Robert	Minéralogie et pétrographie (géologie)
MOZIERES Philippe	Spectrométrie - Physique
OMONT Alain	Astrophysique
OZENDA Paul	Botanique (biologie végétale)
PAYAN Jean-Jacques (détaché)	Mathématiques pures
PEBAY PEYROULA Jean-Claude	Physique
PERRIAUX Jacques	Géologie
PERRIER Guy	Géophysique
PIERRARD Jean-Marie	Mécanique
RASSAT André	Chimie systématique
RENARD Michel	Thermodynamique
RICHARD Lucien	Biologie végétale
RINAUDO Marguerite	Chimie CERMAV
SENGEL Philippe	Biologie animale
SERGERAERT Francis	Mathématiques pures
SOUTIF Michel	Physique
VAILLANT François	Zoologie
VALENTIN Jacques	Physique nucléaire I.S.N.
VAN CUTSEN Bernard	Mathématiques appliquées
VAUQUOIS Bernard	Mathématiques appliquées
VIALON Pierre	Géologie

PROFESSEURS DE 2ème CLASSE

ADIBA Michel	Mathématiques pures
ARMAND Gilbert	Géographie

.../...

AURIAULT Jean-Louis	Mécanique
BEGUIN Claude (M.)	Chimie organique
BOEHLER Jean-Paul	Mécanique
BOITET Christian	Mathématiques appliquées
BORNAREL Jean	Physique
BRUN Gilbert	Biologie
CASTAING Bernard	Physique
CHARDON Michel	Géographie
COHENADDAD Jean-Pierre	Physique
DENEUVILLE Alain	Physique
DEPASSEL Roger	Mécanique des fluides
DOUCE Roland	Physiologie végétale
DUFRESNOY Alain	Mathématiques pures
GASPARD François	Physique
GAUTRON René	Chimie
GIDON Maurice	Géologie
GIGNOUX Claude (M.)	Sciences nucléaires I.S.N.
GUITTON Jacques	Chimie
HACQUES Gérard	Mathématiques appliquées
HERBIN Jacky	Géographie
HICTER Pierre	Chimie
JOSELEAU Jean-Paul	Biochimie
KERCKOVE Claude (M.)	Géologie
LE BRETON Alain	Mathématiques appliquées
LONGEQUEUE Nicole	Sciences nucléaires I.S.N.
LUCAS Robert	Physiques
LUNA Domingo	Mathématiques pures
MASCLE Georges	Géologie
NEMOZ Alain	Thermodynamique (CNRS - CRTBT)
OUDET Bruno	Mathématiques appliquées
PELMONT Jean	Biochimie
PERRIN Claude (M.)	Sciences nucléaires I.S.N.
PFISTER Jean-Claude (détaché)	Physique du solide
PIBOULE Michel	Géologie
PIERRE Jean-Louis	Chimie organique
RAYNAUD Hervé	Mathématiques appliquées
ROBERT Gilles	Mathématiques pures
ROBERT Jean-Bernard	Chimie physique
ROSSI André	Physiologie végétale
SAKAROVITCH Michel	Mathématiques appliquées
SARROT REYNAUD Jean	Géologie
SAXOD Raymond	Biologie animale
SOUTIF Jeanne	Physique
SCHOOL Pierre-Claude	Mathématiques appliquées
STUTZ Pierre	Mécanique
SUBRA Robert	Chimie
VIDAL Michel	Chimie organique
VIVIAN Robert	Géographie

A LA PAIX

J'exprime toute ma reconnaissance à Messieurs J. Della-Dora, C. Le Faou, J. Mermet et F. Robert qui ont bien voulu diriger cette thèse, pour la sympathie, les critiques et les encouragements prodigués au cours de mon travail.

Je suis très sensible à l'honneur que me fait Monsieur G. Veillon en acceptant de présider mon jury. Je tiens à lui exprimer ma sincère reconnaissance.

Mes remerciements vont également à Messieurs J. P. Uhry et M. Heydemann qui ont bien voulu participer à mon jury.

Je tiens à remercier l'ensemble de l'équipe d'Analyse Numérique et celui de l'équipe de Communication Graphique et Méthodologie de la C.A.O. pour l'ambiance amicale dans laquelle j'ai été accueilli.

Enfin je tiens à remercier ceux qui ont participé à la réalisation matérielle de ce travail et tout particulièrement Monsieur C. Anguille, Monsieur P. Chenin, Monsieur M. Cosnard, Monsieur D. Iglesias et Madame C. Meyrieux.

SOMMAIRE

INTRODUCTION GENERALE 1

CHAPITRE I 3
INTRODUCTION 5
préliminaires 7
I) notations utilisées 8
II) stabilité d'un système algébrodifférentiel 11
III) cas explicite 16
IV) retour au cas implicite 27
V) stabilité des méthodes de résolution 32
CONCLUSION 49
ANNEXE DU CHAPITRE I 53

CHAPITRE II 93
INTRODUCTION 95
I) préliminaire 97
II) présentation du problème 97
III) nouveaux outils numériques ...103
IV) algorithmes de partition et de renumérotation des équations 108
V) méthode de Newton à plusieur niveaux et algorithmes correspondants 130
VI) découplage d'un système couplé 138
CONCLUSION 145

CHAPITRE III 149
INTRODUCTION 151
I) principe 153
II) calcul de l'erreur de troncature 155
III) choix du pas et de l'ordre ..157
IV) divers problèmes numériques liés au calcul du pas et de l'ordre 159
V) résolution de problèmes-test ..166
VI) élaboration d'un code pour la résolution de systèmes algébrodifférentiels raides ..190
VII) comparaison des différentes méthodes entre elles 197
VIII) étude du coût des algorithmes de partitionnement(procédure TRI) 205
CONCLUSION 211
ANNEXE DU CHAPITRE III 213

CONCLUSION GENERALE 225

BIBLIOGRAPHIE 227

INTRODUCTION GENERALE

Depuis une quinzaine d'années, des efforts considérables sont consentis de par le monde pour le développement de programmes de simulation de circuits intégrés, qui sont aujourd'hui des outils indispensables pour la conception et la vérification des circuits.

Cependant, la complexité toujours croissante des circuits intégrés fait que les outils existants sont vites dépassés, ce qui fait de ce champ de recherches un domaine en continuelle évolution.

La simulation d'un circuit électrique au niveau le plus bas (niveau électrique fin) transforme un modèle électrique en un système algébrodifférentiel qui le plus souvent est "raide", c'est à dire dont les variables effectuent au cours du temps une ou plusieurs transitions exponentielles entre deux états stables. Le graphe de la décharge d'une capacité dans une résistance fournit un exemple assez typique de ce type de fonction; un autre exemple est ce qui constitue la famille des fonctions "créneaux" ou "dents de scie", qui se rencontre souvent lors de la simulation de circuits électriques. Nous pouvons également rencontrer des problèmes "raides" dans la modélisation de réactions chimiques, de phénomènes physiques, etc...

Cette thèse est consacrée à l'étude de problèmes numériques liés à la résolution des systèmes algébrodifférentiels "raides" de grande dimension, et se décompose en trois parties :

Au premier chapitre, nous nous intéressons au concept de stabilité d'un système algébrodifférentiel, et, tout en rappelant les résultats fondamentaux existant en la matière, nous introduisons la notion de forte stabilité et donnons quelques caractérisations des systèmes fortement stables. Nous étudions ensuite la stabilité d'un schéma de résolution (A-stabilité, A(θ)-stabilité, quasi-stabilité).

Le second chapitre développe un certain nombre de techniques numériques, qui en réalisant la partition du système initial en sous-systèmes faiblement couplés, permettent la prise en compte de la latence et l'utilisation de méthodes de surrelaxation pour la résolution de tels systèmes.

Le troisième chapitre est lui consacré aux expérimentations numériques, ainsi qu'à la description d'un logiciel conversationnel utilisant les techniques décrites dans le chapitre II. Ce logiciel est testé sur un certain nombre de problèmes raides, et les graphes des solutions obtenues sont donnés. Nous finissons ce chapitre sur une étude expérimentale des coûts relatifs des différentes méthodes et de celui des algorithmes de partitionnement.

CHAPITRE I

INTRODUCTION

Dans ce chapitre,nous présentons des notions de base sur la théorie de la stabilité des équations différentielles ordinaires.

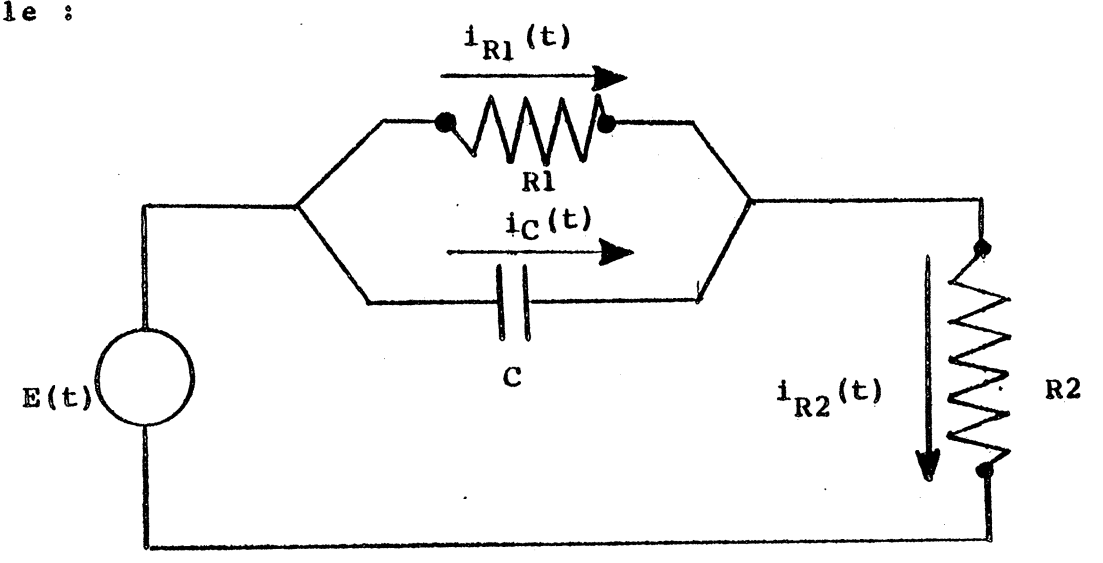
Nous rappelons les notions et résultats,désormais classiques, obtenus en la matière par Liapounov, et nous introduisons la définition d'un nouveau type de stabilité:la forte stabilité. Un certain nombre de propriétés nous permettra de caractériser des familles d'équations différentielles ordinaires fortement stables.

Dans un deuxième temps,nous faisons le point sur la stabilité des méthodes numériques usitées actuellement, en citant notamment les résultats de Henrici, Dahlquist et Gear.

Nous montrons par ailleurs que les schémas numériques issus des travaux de Gear ont des propriétés de stabilité intéressantes lorsqu'ils sont appliqués à une famille d'équations différentielles ordinaires que nous caractérisons.

PRELIMINAIRES

Vouloir connaître l'activité d'un circuit électrique au cours du temps conduit à la résolution de systèmes algébrodifférentiels non-linéaires. Afin d'expliciter notre propos, nous allons étudier un exemple de circuit très simple :



t ($\in$ $[t_0,T]$) désigne le temps.

E(t) désigne une source de tension variable avec le temps.

$i_{R1}(t)$ désigne l'intensité du courant qui traverse la résistance R1 au temps t.

$i_{R2}(t)$ désigne l'intensité du courant qui traverse la résistance R2 au temps t.

$i_C(t)$ désigne l'intensité du courant qui traverse la capacité C au temps t.

R1 a une résistance de r1(t)

R2 a une résistance de r2(t)

C a une capacité de c(t)

Ces diverses grandeurs sont liées par le système

$$\begin{cases} E(t) - R1.i_{R1}(t) - R2.i_{R2}(t) = 0 \\ i_C(t) - c(t) \cdot \left(\frac{d}{dt}(E(t) - R2.i_{R2}(t))\right) = 0 \\ i_{R2}(t) - i_{R1}(t) - i_C(t) = 0 \end{cases}$$

(S) est constitué d'un système algébrique(équations 1) et 3)) et d'un système différentiel(équation 2)) couplés, d'où le nom de système algébrodifférentiel donné à (S).

L'intervention de variables différentielles(i_{R2}) nous amène à compléter (S) avec la valeur de i_{R2} au temps t_0. Ces valeurs initiales sont obtenues en étudiant le circuit précédent en régime continu, c'est à dire en considérant pour notre exemple que la source de tension est constante et égale à $E(t_0)$: cela conduit à résoudre des systèmes purement algébriques [Rey, 1978, chapitre II]. Nous ne nous intéresserons pour notre part qu'au régime transitoire(quand les sources sont variables) en supposant connues les valeurs initiales des variables différentielles.

I) NOTATIONS UTILISEES

D'une manière plus générale, si nous posons $\dot{z}(t)= \frac{d}{dt}z(t)$,notre problème se ramène à la résolution de :

$$(S) \quad \begin{cases} f(y(t),z(t),\dot{z}(t),t)= \Theta_{q+p} \\ t \in [t_0,T] \\ z(t_0) = \mu \end{cases}$$

- $y(t)$ ($\in \mathbb{R}^q$) et $z(t)$ ($\in \mathbb{R}^p$) ont pour composantes des tensions et des intensités.
- f est une fonction de $\mathbb{R}^{q+2p} \times [t_0,T]$ dans $\mathbb{R}^{q+p}$, qui à (y,z,d,t) associe $(f_i(y,z,d,t))_{i=1,\ldots q+p}$.
- nous désignerons par

$\frac{\partial f}{\partial y}(u)$, $\frac{\partial f}{\partial z}(u)$, $\frac{\partial f}{\partial \dot{z}}(u)$, $\frac{\partial f}{\partial t}(u)$ les matrices jacobiennes de f par rapport aux variables y,z,d et t calculées au point u ($u \in E=\mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^p \times [t_0,T]$).

nous appellerons $\frac{\partial f}{\partial y\dot{z}}(u)$ la matrice $\left(\frac{\partial f}{\partial y}(u) , \frac{\partial f}{\partial \dot{z}}(u) \right)$.

Les méthodes de résolution de (S) se prêtant le mieux à l'utilisation des moyens de calcul actuels sont les méthodes de **discrétisation**, qui consistent à obtenir les valeurs de y

et z en un nombre fini de points t_i ($t_i \in [t_0,T]$).

Toutes les les méthodes existantes utilisent pour calculer les valeurs de y et z au point t_i , leurs valeurs aux points t_j ($j < i$). La précision d'un ordinateur n'étant pas infinie, ces valeurs sont entachées d'erreurs : ce sont les **erreurs d'arrondi**. De plus z n'étant pas connu en tout point d'un voisinnage de t_i ,la valeur $\dot{z}(t_i)$ n'est pas accessible. Il est nécessaire de l'approcher par diverses formules. C'est une deuxième source d'erreurs que nous appellerons erreurs de méthode ou **erreurs de troncature**.

Un problème crucial est celui de la propagation de ces erreurs.

Le choix de la méthode utilisée n'est pas du tout indifférent en ce qui concerne ce problème. Considérons l'exemple suivant, donné par Gear [Gear,1971, b), p. 17] :

Soit le système (S_1)

$$\begin{cases} \dot{z}(t) + 1000.(z(t) - t^2) - 2t = 0 \\ t \in [0,1] \\ z(0) = 0 \end{cases}$$

dont la solution est $z(t) = t^2$.
Considérons la méthode numérique de résolution simple qui est la méthode d'Euler **explicite** :

$h=1/k$
(h est appelé pas de discrétisation et est ici supposé constant sur tout l'intervalle [0,1])

$$\dot{z}(t_n) = \frac{z(t_{n+1}) - z(t_n)}{h}$$

Reporté dans le système précédent, cela donne :

$$\begin{cases} z(t_{n+1}) = (1 - 1000h).z(t_n) + (2t_n + 1000\, t_n^2).h \\ z(0) = 0 \end{cases}$$

ou encore,puisque $t_n = n.h$:

$$\begin{cases} z(t_{n+1}) = (1 - 1000.h).z(t_n) + (2n + 1000.n^2.h)^2.h \\ z(0) = 0 \end{cases}$$

Si nous essayons de calculer z(1) par cette méthode avec des pas de taille 10^{-m} (m = 0,1,...,5) et avec une précision de 10 chiffres significatifs, alors nous avons les résultats suivants :

h	nombre de pas	z(1)
1	1	0
10^{-1}	10	$0.90438207503 \times 10^{16}$
10^{-2}	100	dépassement de capacité-machine
10^{-3}	1000	0.99999900001
10^{-4}	10000	0.99999990000
10^{-5}	100000	0.99999998997

Ces résultats se comprennent assez bien puisque pour un pas h donné, une erreur est multipliée à chaque pas par (1 - 1000.h). Ainsi pour $h = 10^{-2}$, au bout du compte elle est multipliée par 9^{100} !

La méthode ne donne de résultats significatifs que pour h inférieur à 10^{-3} . L'erreur est alors sensiblement égale à 10^{-k-3} (m > 3).

Considérons maintenant une méthode d'Euler **implicite** :

$$\dot{z}(t_n) = \frac{z(t_n) - z(t_{n-1})}{h}$$

Nous obtenons alors les relations :

$$\begin{cases} z(t_n) = (1+1000h)^{-1}(z(t_{n-1}) + (2n + 1000.n^2.h).h^2)) \\ z(0) = 0 \end{cases}$$

Nous avons les résultats suivants:

h	nombre de pas	z(1)
1	1	1.000999001
10^{-1}	10	1.000100000
10^{-2}	100	1.000010000
10^{-3}	1000	1.000001000
10^{-4}	10000	1.000000100
10^{-5}	100000	1.000000010

Ici, les résultats sont toujours significatifs, même pour h=1. L'erreur est sensiblement égale à 10^{-m-3} ($m > 0$). Cela nous amène à rechercher une certaine stabilité des résultats et à caractériser cette stabilité.

Nous allons distinguer deux types de stabilité :

-stabilité d'un système algébrodifférentiel

-stabilité d'une méthode numérique

II)STABILITE D'UN SYSTEME ALGEBRODIFFERENTIEL

Nous reprenons pour les systèmes algébrodifférentiels les notions de "stabilité au sens de Liapounov" et "d'attractivité" [Liap, 1907], [Rou, 1973, tome 2, p.6 et 76]

Cependant,ces notions n'ont pas été reprises telles quelles. Il nous a semblé plus simple,en considérant le type de problème auquel nous étions confrontés de ne pas considérer la stabilité d'un système par rapport à l'origine mais par rapport à sa solution(en supposant bien sûr l'existence d'une telle solution). Les cas concrets rencontrés sont du type

$$(S_T) \quad \begin{cases} f(y,z,\dot{z},t) = \theta_{q+p} \\ t \in [t_0,T] \\ z(t_0) = \mu \end{cases}$$

mais il est préférable pour étudier les différentes notions de stabilité,de pouvoir observer les conséquences des perturbations lorsque t tend vers l'infini. Aussi est-il nécessaire de travailler sur des systèmes du type

$$(S) \quad \begin{cases} f(y,z,\dot{z},t)=\theta_{q+p} \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$$

Introduisons maintenant le problème perturbé

$$(S_{\varepsilon_1,t_1}) \quad \begin{cases} f(y_\varepsilon,z_\varepsilon,\dot{z}_\varepsilon,t) = \theta_{q+p} \\ t \in [t_1,+\infty[\\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \end{cases}$$

DEFINITIONS

Nous supposerons dans les définitions suivantes que (S) admet une solution sur $[t_0,+\infty[$.

(D.1) (S) sera dit **stable**(au sens de Liapounov) au point t_1 si et seulement si :

$$\exists \rho > 0 \ / \ \forall \delta \in]0,\rho[, \ \exists \sigma > 0 \ /$$

$$\|\varepsilon_1\| < \sigma \quad \Rightarrow \quad \forall t \in [t_1,+\infty[\begin{cases} \|y_\varepsilon(t) - y(t)\| < \delta \\ \|z_\varepsilon(t) - z(t)\| < \delta \end{cases}$$

(D.2) (S) sera dit **uniformément stable**(au sens de Liapounov) si et seulement si :

$$\exists \rho > 0 \ / \ \forall \delta \in]0,\rho[, \exists \sigma > 0 \ / \ \forall t_1 \in [t_0, +\infty[.$$

$$\|\varepsilon_1\| < \sigma \quad \Rightarrow \quad \forall t \in [t_1,+\infty[\begin{cases} \|y_\varepsilon(t) - y(t)\| < \delta \\ \|z_\varepsilon(t) - z(t)\| < \delta \end{cases}$$

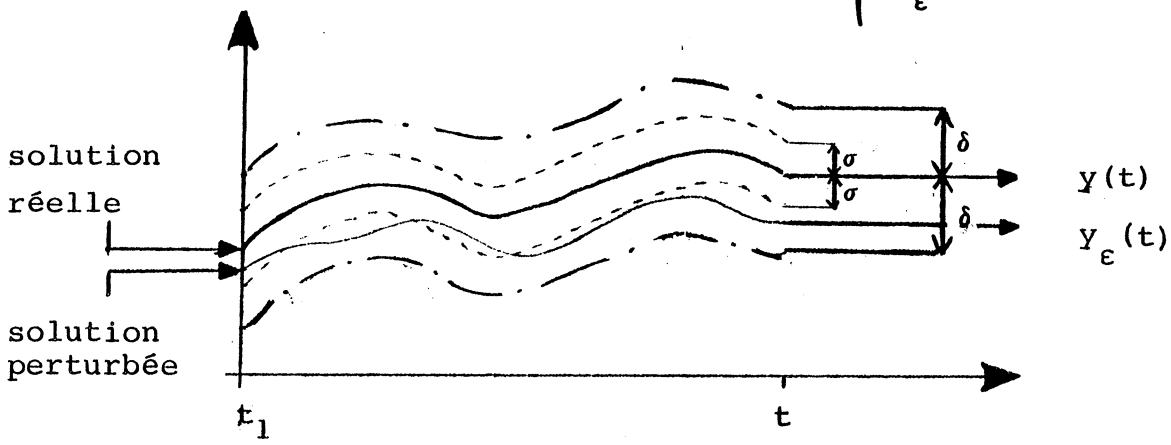


illustration de la stabilité (au sens de Liapounov)

(D.3) (S) sera dit attractif au point t_1 si et seulement si :

$$\exists \sigma > 0 \quad \|\varepsilon_1\| < \sigma \quad \Rightarrow \begin{cases} \lim\limits_{t \to +\infty} \|y_\varepsilon(t) - y(t)\| = 0 \\ \lim\limits_{t \to +\infty} \|z_\varepsilon(t) - z(t)\| = 0 \end{cases}$$

(D.4) (S) sera dit uniformément attractif si et seulement si :

$$\exists \sigma > 0 \quad / \; \forall t_1 \in [t_0, +\infty[\,,$$

$$\|\varepsilon_1\| < \sigma \quad \Rightarrow \begin{cases} \lim\limits_{t \to +\infty} \|y_\varepsilon(t) - y(t)\| = 0 \\ \lim\limits_{t \to +\infty} \|z_\varepsilon(t) - z(t)\| = 0 \end{cases}$$

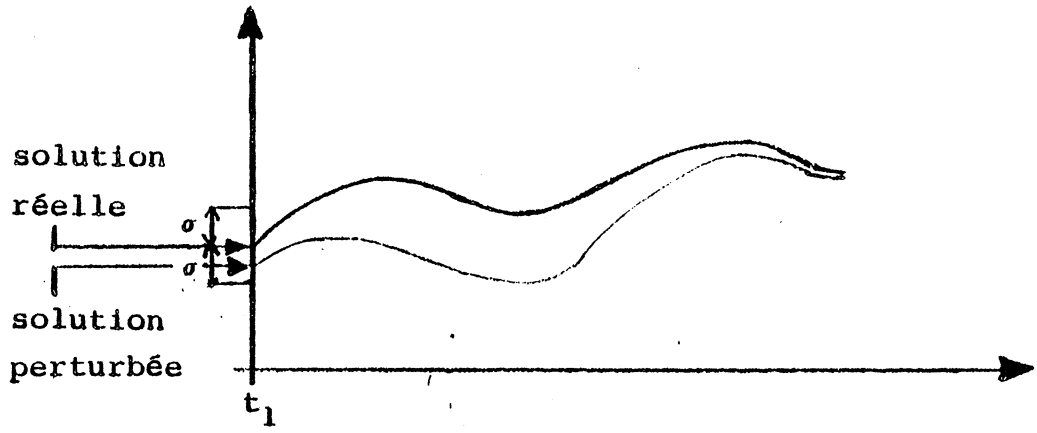


illustration de l'attractivité

(D.5) (S) sera dit asymptotiquement stable si et seulement si (S) est à la fois stable et attractif.

(D.6) (S) sera dit uniformément asymptotiquement stable si et seulement si (S) est à la fois uniformément stable et uniformément attractif.

M. Crouzeix [Crou, 1980, p.7] indique que si nous posons

$$(S)\quad \begin{cases} \dot{z} = f(z,t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

$$(S_{\varepsilon,t_1})\quad \begin{cases} \dot{z}_\varepsilon(t) = f(z_\varepsilon(t),t) + \varepsilon(t) \\ t \in [t_1, +\infty[\\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \end{cases}$$

et s'il existe un produit scalaire ,noté (.,.), tel que:

$$(R_1)\ \forall\, z_1 \in \mathbb{R}^p\ ,\ \forall\, z_2 \in \mathbb{R}^p,$$
$$(f(z_1,t) - f(z_2,t), z_1 - z_2) \leq \ell \cdot \| z_1 - z_2 \|_2^2$$

où ℓ désigne un nombre réel(positif ou négatif) et $\|\cdot\|_2$ la norme associée au produit scalaire, alors :

$$\| z_\varepsilon(t) - z(t) \|_2 \leq e^{\ell \cdot (t - t_0)} \cdot \| \varepsilon_1 \|_2 + \int_{t_1}^{t} e^{\ell \cdot (t - s)} \cdot \| \varepsilon(s) \|_2 ds$$

Nous pouvons remarquer que si ℓ est négatif,alors (S) est uniformément asymptotiquement stable. En effet, si nous ne créons une perturbation qu'au point t_1 $(\varepsilon(t) = 0 \quad \forall\, t > t_1)$, nous avons :

$$\| z_\varepsilon(t) - z(t) \|_2 \leq e^{\ell \cdot (t - t_1)} \cdot \| \varepsilon_1 \|_2$$

et donc (D.2) et (D.4) seront vérifiées.

Bien plus, si nous perturbons constamment le système $(\varepsilon(t) \neq 0)$ nous avons , si nous supposons ℓ **négatif** :

$$\begin{cases} \| \varepsilon_1 \|_2 < \delta \\ \| \varepsilon(t) \|_2 < \delta \Rightarrow \forall\, t \in [t_1, +\infty[,\ \| z_\varepsilon(t) - z(t) \|_2 < (1 - 1/\ell) \cdot \delta \\ \forall\, t \in [t_1, +\infty[\end{cases}$$

puisque nous disposons de l'inégalité suivante :

$$\int_{t_1}^{t} e^{\ell \cdot (t - s)} \cdot \|\varepsilon(s)\|_2 \, ds \leq (-1/\ell) \cdot \sup_{t \geq t_1} \|\varepsilon(s)\|_2$$

Il faut remarquer que nous travaillons dans des espaces vectoriels de dimension finie, donc toutes les normes sont équivalentes entre elles. Les définitions précédentes sont donc indépendantes des normes utilisées.

Considérons le système algébrodifférentiel

$$(S) \quad \begin{cases} f(y,z,\dot{z},t) = \theta_{q+p} \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$$

et perturbons-le de la manière suivante :

$$(S_{\varepsilon,t_1}) \quad \begin{cases} f(y_\varepsilon,z_\varepsilon,\dot{z}_\varepsilon + \varepsilon(t),t) = \theta_{q+p} \\ t \in [t_1,+\infty[\\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \end{cases}$$

Nous proposons qu'un système qui en outre de la stabilité uniformément asymptotique vérifie :

$$\exists \rho > 0 \ , \ \forall \xi \in]0,\rho[\ , \ \exists \sigma > 0 \ / \ \forall \ t_1 \in [t_0,+\infty[$$

$$\begin{cases} \|\varepsilon_1\| < \sigma \\ \sup\limits_{t \in [t_1,+\infty[} \|\varepsilon(t)\| < \sigma \end{cases} \Rightarrow \begin{cases} \|y_\varepsilon(t) - y(t)\| < \xi \\ \|z_\varepsilon(t) - z(t)\| < \xi \\ \forall \ t \in [t_1,+\infty[\end{cases}$$

soit dit **fortement stable**(définition (D.7))

Cette notion correspond aux perturbations qui se rencontrent le plus fréquemment,puisque les méthodes numériques existant aujourd'hui pour résoudre (S) approximent $\dot{z}(t)$, et donc perturbent constamment le système.

Les différentes définitions précédentes s'adaptent de manière naturelle aux restrictions que sont les cas purement différentiels ou explicites.

III) CAS EXPLICITE

Nous allons étudier la forte stabilité des systèmes

$$(S) \quad \begin{cases} \dot{z}(t) = f(z(t),t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

où z appartient à $\mathbb{R}^p$.

proposition 1

$$\text{soit} \quad (S_1) \quad \begin{cases} \dot{z}(t) = A \cdot z(t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

où A est une matrice réelle à coefficients constants p-lignes p-colonnes .
Soient $\{\lambda_i\}_{i=1,\ldots p}$ les valeurs propres de A comptées avec leur multiplicité algébrique.

Si $\max_{i=1,\ldots p} \mathrm{Re}(\lambda_i) < 0$, alors (S_1) est fortement stable.

(démonstration en annexe du chapitre I)

proposition 2

$$\text{Soit} \quad (S_2) \quad \begin{cases} \dot{z}(t) = A \cdot z(t) + f(z(t),t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

Supposons que les hypothèses (H_1) suivantes soient vérifiées :

i) f est continue et lipschitzienne par rapport à la première variable.

ii) $\max_{i=1,\ldots p} \mathrm{Re}(\lambda_i) < 0$ où $\{\lambda_i\}$ est l'ensemble des valeurs propres de A.

iii) $\exists \xi_z > 0 \ / \forall \varepsilon > 0, \ \exists \alpha > 0 \ / \forall t \in [t_0, +\infty[$,
$\forall z_1 \in B_o(z(t), \xi_z)$, $\forall z_2 \in B_o(z(t), \xi_z)$,

$$\|z_1 - z_2\| < \alpha \Rightarrow \|f(z_1,t) - f(z_2,t)\| < \varepsilon \cdot \|z_1 - z_2\|$$

(nous notons z(t) la solution de (S_2))

alors (S_2) est fortement stable.

(démonstration en annexe du chapitre I)

Nous savons grâce aux travaux de Liapunov , que si $\max_{i=1,...p} Re(\lambda_i) < 0$ alors (S_1) (resp. (S_2) si Hl est vérifié) est instable au sens de Liapunov [Liap , 1907, p 291] .Nous n'avons par contre pas de résultats probants en ce qui concerne la stabilité du cas $\max_{i=1,..p} Re(\lambda_i) = 0$. Cependant diverses études intéressantes sur le comportement des solutions ont été faites, notemment par V. Arnold [Arn, 1974, p 160 et suivantes]

Nous allons maintenant aborder les cas plus généraux que sont les systèmes différentiels

$$\begin{cases} \dot{z}(t) = P(t) \cdot z(t) + f(z(t),t) \\ t \in [t_0, +\infty[\end{cases}$$

Jusqu'à présent,les seules transformations utilisées sont les transformations linéaires effectuées par une matrice inversible à coefficients constants.

Si nous voulons aller plus loin dans notre propos, nous devons élargir le champs des transformations utilisables.

soit le système (S)
$$\begin{cases} \dot{z}(t) = f(z,t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

où f est lipschitzienne par rapport à la première variable, continue et définie sur $\mathbb{R}^p \times [t_0, +\infty[$.

En nous inspirant de la définition donnée par Lefschetz [Lef, 1957, p 80] d' une transformation uniformément régulière nous proposons la définition suivante(**D.8**) :

Soit T une application de $\mathbb{R}^p \times [t_0,+\infty[$ vers $\mathbb{R}^p$. Soit z(t) la solution unique de (S).

Nous dirons que T est une **transformation fortement régulière** de (S) si et seulement si elle vérifie :

i) $\forall\ t \in [t_0,+\infty[\ ,\ \exists\ \alpha(t) > 0,\ \exists\ \beta(t) > 0\ /\ \tau_t$ est un homéomorphisme de $B_o(z(t),\alpha(t))$ dans $B_o(y(t),\beta(t))$ si nous posons $\tau_t(z) = T(z,t)$ et $y(t) = \tau_t(z(t))$.

ii) $\exists\ \hat{\alpha} > 0,\ \exists\ \hat{\beta} > 0 / \forall\ t \in [t_0,+\infty[\ ,$ $\begin{cases} \alpha(t) > \hat{\alpha} \\ \beta(t) > \hat{\beta} \end{cases}$

iii) $\forall\ \rho > 0,\ \exists\ \rho_1 > 0\ /\ \forall\ t \in [t_0,+\infty[\ ,$

$\|z_\varepsilon(t) - z(t)\| > \rho \quad \Rightarrow \quad \|\tau_t(z) - \tau_t(z(t))\| > \rho_1$

iv) $\forall\ \rho^* > 0,\ \exists\ \rho_1^* > 0 / \forall\ t \in [t_0,+\infty[\ ,$

$\|y - y(t)\| > \rho^* \Rightarrow \|\tau_t^{-1}(y) - \tau_t^{-1}(t)\| > \rho_1^*$

v) Posons $T^*(y,t) = \tau_t^{-1}(y)$

Supposons alors que les différentielles partielles $\frac{\partial T}{\partial z}$, $\frac{\partial T}{\partial t}$, $\frac{\partial T^*}{\partial y}$, et $\frac{\partial T^*}{\partial t}$ soient définies, continues, lipschitziennes par rapport à la première variable et bornées sur

$E^\alpha = \{ (z,t) \in \mathbb{R}^p \times [t_0,+\infty[\ /\ \|z - z(t)\| < \alpha(t) \}$

et $E^*_\beta = \{ (y,t) \in \mathbb{R}^p \times [t_0,+\infty[\ /\ \|y - y(t)\| < \beta(t) \}$

La norme qui est prise en considération pour T et T* est

$$\| f \| = \sup_{(z,t)\neq(0,0)} \frac{\| f(z,t) \|}{\|(z,t)\|} .$$

définition (D.9) :

Soient les opérateurs L_T et L_{T*} définis comme suit :

$$\forall f \in (E_\beta^*)^{E_\alpha}, \forall t \in [t_0,+\infty[, \forall y \in B_o(y(t),\beta(t)),$$

$$\left[[L_T(f)](y,t) = \frac{\partial T}{\partial z}(T^*(y,t),t) \circ f(T^*(y,t),t) + \frac{\partial T}{\partial t}(T^*(y,t),t) \right.$$

$$\forall f \in (E_\alpha)^{E_\beta^*}, \forall t \in [t_0,+\infty[, \forall z \in B_o(z(t),\alpha(t)),$$

$$\left[[L_{T*}(f)](z,t) = \frac{\partial T^*}{\partial y}(T(z,t),t) \circ f^*(T(z,t),t) + \frac{\partial T^*}{\partial t}(T(z,t),t) \right.$$

où B^A désigne l'ensemble des applications de A vers B.

proposition 3

Soit f un élément de $(E_\beta^*)^{E_\alpha}$ (reps. f* un élément de $(E_\alpha)^{E_{\beta*}}$) alors $L_{T*}(L_T(f)) = f$ (resp. $L_T(L_{T*}(f^*)) = f^*$)

(démonstration en annexe du chapitre I)

dT et dT* sont continues. Il est donc clair que si f $\in$ $(E_{\beta*})^{E_\alpha}$ (resp. f* $\in$ $(E_\alpha)^{E_{\beta*}}$) est continue, alors $L_T(f)$ (resp. $L_{T*}(f^*)$) est continue.

proposition 4

Soit f un élément de $(E_{\beta*})^{E_\alpha}$ (resp. f* un élément de $(E_\alpha)^{E_{\beta*}}$).

Si f(resp. f*) est lipschitzienne par rapport à la première variable, alors $L_T(f)$ (resp. $L_{T*}(f^*)$) est lipschitzienne par rapport à la première variable.

(démonstration en annexe du chapitre I)

T transforme le système (S)
$$\begin{cases} \dot{z}(t) = f(z(t),t) \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$$

en (S*)
$$\begin{cases} \dot{y}(t) = f^*(y,t) \\ t \in [t_0,+\infty[\\ y(t_0) = \mu^* \end{cases}$$

si nous posons $y(t) = T(z(t),t)$, $f^* = L_T(f)$ et $\mu^* = T(\mu,t_0)$.

proposition 5

Une condition nécessaire et suffisante pour que (S) soit stable au sens de Liapunov (resp. attractif, uniformément stable, uniformément attractif, asymptotiquement stable, uniformément asymptotiquement stable, fortement stable) est que (S*) soit stable au sens de Liapunov (resp. idem)

(démonstration en annexe du chapitre I)

Toutes les transformations de la forme
$T(z,t)=A(t)\cdot z + a(z,t)$, où

- A(t) est inversible, continue et d'inverse continue

- $a_i(z,t)$ sont développables en puissance de z_j avec un premier terme de degré au moins égal à 2, les coefficients de la série étant continus et bornés uniformément en t, pour $t \in [t_0,+\infty[$

- la série issue du développement de a_i converge dans une sphère de rayon donné $\|z - z(t)\| < B$

vérifient i),ii),iii) et iv) [Lef, 1957, p 131]

Nous allons dans un premier temps, nous intéresser

aux systèmes de la forme (S_3)
$$\begin{cases} \dot{z}(t) = P(t) \cdot z \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$$
où P est une matrice p-lignes p-colonnes, continue et bornée.

Un résultat important pour l'étude de ces systèmes est le théorème de réduction de Perron [Lef, 1957, p 131] qui indique :

"Il existe pour tout $t > t_0$ une matrice $u(t)$ réelle, continuement différentiable et orthonormale telle que $y(t)$ définie par $y(t) = {}^t u(t) \cdot z(t)$ vérifie :

$$(S_3^*) \quad \begin{cases} \dot{y}(t) = Q(t) \cdot y(t) \\ t \in [t_0, +\infty[\\ y(t_0) = {}^t u(t_0) \cdot \mu \end{cases}$$

où $Q(t)$, déterminée par la relation $Q(t) = {}^t u(t) \cdot [P(t) \cdot u(t) - \dot{u}(t)]$, est bornée, continue et triangulaire supérieure."

Plusieurs remarques s'imposent :

- Le même énoncé avec "triangulaire inférieure " remplaçant "triangulaire supérieure" reste valable. En effet, il suffit d'utiliser la transformation linéaire constante et inversible (donc fortement régulière) qui à y_j associe y_{p-j}

- $u(t)$ étant orthonormale, $u(t)$ et $u^{-1}(t) = {}^t u(t)$ sont bornées.

La transformation $T(z,t) = {}^t u(t) \cdot z$ vérifie, d'après une propriété énoncée précédemment, i), ii), iii), et iv) de la définition (D.8).

Comme nous avons

$\frac{\partial T}{\partial z}(z,t) = {}^t u(t)$, $\frac{\partial T}{\partial t}(z,t) = \frac{d({}^t u)}{dt}(t) \cdot z$, $\frac{\partial T^*}{\partial y}(y,t) = u(t)$,

et $\frac{\partial T^*}{\partial t}(y,t) = \frac{du}{dt}(t) \cdot y$,

et puisque $\frac{du}{dt}(t)$ (resp. $\frac{d({}^t u)}{dt}(t)$) qui vaut $P(t) \cdot u(t) - u(t) \cdot Q(t)$ (resp. ${}^t u(t) \cdot {}^t P(t) - {}^t Q(t) \cdot {}^t u(t)$) bornée, les différentielles partielles de T et de T* sont lipschitziennes par rapport à la première variable et bornées

sur $\mathbb{R}^p \times [t_0, +\infty[$. T est donc une transformation **fortement régulière.**

proposition 6:

Posons $Q(t) = (q_{ij}(t))_{\substack{i=1,\ldots p \\ j=1,\ldots p}}$

Si $\exists\, T \in [t_0, +\infty[\;/\; \max_{i=1,\ldots p} \left(\sup_{t \in [T, +\infty[} (q_{ij}(t)) \right) \leqslant 0$

alors (S_3) est fortement stable.

(démonstration en annexe du chapitre I)

proposition 7

$$\text{Posons } (S_4) \quad \begin{cases} \dot{z}(t) = P(t) \cdot z(t) + f(z(t), t) \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

Supposons que les hypothèses (H_2) suivantes soient vérifiées :

i) f est continue et lipschitzienne par rapport à la première variable

ii) P(t) est continue et bornée sur $\mathbb{R}^p \times [t_0, +\infty[$

iii) si nous posons $Q(t) = (q_{ij}(t))$ la matrice réduite de P(t), alors $\exists\, T \in [t_0, +\infty[\;/$

$$\max_{i=1,\ldots p} \left(\sup_{t \in [T, +\infty[} (q_{ij}(t)) \right) \leqslant 0$$

$\exists\, \xi_z > 0 \;/ \forall\, \varepsilon > 0, \; \exists\, \sigma(\varepsilon) > 0 / \; \forall\, t \in [t_0, +\infty[,$

$\forall\, z_1 \in B_o(z(t), \xi_z)$, $\forall\, z_2 \in B_o(z(t), \xi_z)$,

$\| z_1 - z_2 \| < \sigma(\varepsilon) \Rightarrow \| f(z_1, t) - f(z_2, t) \| < \varepsilon \| z_1 - z_2 \|$

alors (S_4) est fortement stable

(démonstration en annexe du chapitre I)

Dans son mémoire fondamental Liapunov [Liap, 1967, p 223-254] introduit la notion de **nombre caractéristique** d'une fonction f d'une variable réelle t :

soient les deux ensembles

$$\Lambda(f) = \{ \lambda \in \mathbb{R} / \lim_{t \to +\infty} \| f(t) \cdot e^{\lambda t} \| = 0 \}$$

et $$\Lambda^*(f) = \{ \lambda \in \mathbb{R} / \lim_{t \to +\infty} \| f(t) \cdot e^{\lambda t} \| = +\infty \} ,$$

alors $\lambda_0(f) = \sup_{\lambda \in \Lambda(f)} \lambda$, qui est aussi égal à $\inf_{\lambda \in \Lambda^*(f)} \lambda$,

sera appelé nombre caractéristique de f.

Il est facile de montrer que :

$$\lambda_0(f \cdot g) \geq \lambda_0(f) + \lambda_0(g)$$

$$\lambda_0(f + g) \begin{cases} = \min(\lambda_0(f), \lambda_0(g)) \text{ si } \lambda_0(f) \neq \lambda_0(g) \\ = \lambda_0(f) \text{ sinon} \end{cases}$$

Soit $Z = \{ z_i \}_{i=1,\dots p}$ un ensemble de p solutions indépendantes du système

$$(S^P) \begin{cases} \dot{z}(t) = P(t) \cdot z(t) \\ t \in [t_0, +\infty[\end{cases}$$

où P(t) est continue et bornée sur $[t_0, +\infty[$

Posons $\Sigma(Z) = \sum_{i=1}^{p} \lambda_0(z_i)$

Si quelque soit l'ensemble de p solutions indépendantes Z', nous avons $\Sigma(Z') < \Sigma(Z)$, alors Z est dit **normal**

Soit u(t) une matrice réelle orthogonale continue et bornée.

Pour toute application g de $[t_0,+\infty[$ dans $\mathbb{R}^p$, nous avons :

$$\exists M_1 > 0, \quad \exists M_2 > 0 \ / \ \forall t \in [t_0,+\infty[,$$
$$M_1\|u(t).g(t)\| < \|g(t)\| < M_2\|u(t).g(t)\|$$

Il est donc aisé de montrer que $\lambda_0(u\cdot g) = \lambda_0(g)$

Si nous considérons Q(t) la matrice réduite de P(t) (cf.théorème de Perron), (S^{Q*}) le système associé, et si nous appelons $T(z,t) = u^{-1}(t)\cdot z$ la transformation de réduction, alors, de manière claire :

- z_i solution de (S^P) $\Leftrightarrow$ $u^{-1}\cdot z$ solution de (S^{Q*})
- $Z = \{ z_i \}_{i=1,\ldots p}$ normal $\Leftrightarrow$ $Z^* = \{ u^{-1}\cdot z_i \}$ normal
- $\lambda_0(z_i) = \lambda_0(u^{-1}\cdot z_i)$

Posons $P(t) = (p_{ij}(t))_{\substack{i=1,..p \\ j=1,..p}}$

Soient $H(t) = e^{(\int_{t_0}^{t} \sum_{i=1}^{p} p_{ii}(u)\cdot du)}$

et

$$H^*(t) = e^{-(\int_{t_0}^{t} \sum_{i=1}^{p} p_{ii}(u)\cdot du)}$$

Puisque pour tout t le produit $H(t)\cdot H^*(t)$ vaut 1, nous avons la relation $\lambda_0(H) + \lambda_0(H^*) \leq \lambda_0(H\cdot H^*) = 0$

Tout système normal Z vérifie donc $\Sigma(Z) \leq -\lambda_0(H^*)$

(S^P) sera dit régulier si et seulement si $\Sigma(Z) = -\lambda_0(H^*)$

Posons $G_i(P)(t) = e^{\int_{t_0}^{t} p_{ii}(u)\cdot du}$ et $G_i^*(P)(t) = e^{-\int_{t_0}^{t} p_{ii}(u)\cdot du}$ $(i = 1,\ldots,p)$

Si nous supposons P(t) triangulaire, une condition nécessaire et suffisante de régularité est
$\lambda_0(G_i(P)) + \lambda_0(G_i^*(P)) = 0$ $(i = 1,\ldots p)$

D'autre part Liapunov indique que la condition nécessaire et suffisante précédente est équivalente à

$\lim_{t \to +\infty} \int_{t_0}^{t} p_{ii}(u)du$ existe et est finie pour i allant de 1 à p

Donc une condition nécessaire et suffisante de régularité de (S^P) est
$\lambda_0(G_i(Q)) + \lambda_0(G_i^*(Q)) = 0$

ou encore $\lim_{t \to +\infty} \int_{t_0}^{t} q_{ii}(u)\cdot du$ existe et est finie.

Considérons maintenant

$$f_i(z,t) = P_i(t)\cdot z + \Sigma\, P_i^{(m_1,\ldots,m_p)}(t) \cdot z_1^{m_1} \cdots z_p^{m_p}$$

où la sommation s'étend à toutes les valeurs non-négatives des entiers $m_1,\ldots,m_p$ satisfaisant à $m_1 + \ldots + m_p \leq 2$

Nous supposerons de plus qu'il existe deux réels positifs M et A tels que

$$\| P_i^{(m_1,\ldots,m_p)}(t) \| \leqslant \frac{M}{A^{m_1+\ldots m_p}} \quad \forall\, t \in [t_0,+\infty[$$

Considérons (S_5)

$$\begin{cases} \dot{z}(t) = f(z,t) \\ t \in [t_0,+\infty[\end{cases}$$

où $f(z,t) = (f_i(z,t))_{i=1,\ldots p}$ et appelons **première approximation** de (S_5) le système

$$(S_{ap}) \quad \begin{cases} \dot{z}(t) = P(t)\cdot z(t) \\ t \in [t_0,+\infty[\end{cases}$$

où $P(t) = (P_i(t))_{i=1,\ldots p}$

Liapunov démontre alors les deux résultats de stabilité remarquables suivants [Liap, 1907, p.252-253] :

théorème 1

Si le système d'équations différentielles de la première aproximation est régulier,et si tous ses nombres caractéristiques sont positifs,alors (S_5) est stable.

théorème 2

Si le système d'équations de la première approximation est régulier et si un de ses nombres caractéristiques est négatif,alors (S_5) est instable.

La correspondance entre ces résultats et le cas où P(t) est une matrice constante est d'autant plus grande que les nombres caractéristiques des solutions dans le cas où P(t) est une matrice constante sont les **valeurs propres de P(t) changées de signe.**

En ce qui concerne la forte stabilité, nous n'avons pas pour l'instant de résultats comparables liant nombres caractéristiques et forte stabilité.

IV)RETOUR AU CAS IMPLICITE

Nous supposerons que f est une application de classe C^k ($k \geq 1$) allant de $E = \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^p \times [t_0, +\infty[$ vers $\mathbb{R}^{q+p}$.

proposition 8

Supposons réalisées les hypothèses (H_3) suivantes :

- $\forall\, u \in \Omega_T = \{ (c,z,d,t) \in \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^p \times [t_0,T] \;/\; f(c,z,d,t) = \theta_{q+p} \}$
  $\dfrac{\partial f}{\partial \dot{y}\dot{z}}(u)^{-1}$ existe
- $\exists\, \beta_z(T) > 0,\ \exists\, \beta_t(T) > 0 \ /\forall\, u \in \Omega_T,$
  $\left\| \dfrac{\partial f}{\partial \dot{y}\dot{z}}(u)^{-1} \circ \dfrac{\partial f}{\partial z}(u) \right\| < \beta_z(T)$
  $\left\| \dfrac{\partial f}{\partial \dot{y}\dot{z}}(u)^{-1} \circ \dfrac{\partial f}{\partial t}(u) \right\| < \beta_t(T)$
- $\exists!\, (c,d) \in \mathbb{R}^q \times \mathbb{R}^p \;/\; f(c,\mu,d,t_0) = \theta_{q+p}$

alors si nous posons (S_T) $\begin{cases} f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p} \\ t \in [t_0,T] \\ z(t_0) = \mu \end{cases}$

il existe exactement une application y de classe C^0 et une application z de classe C^1 telles que (S_T) soit vérifié. De plus y est de classe C^k et z est de classe C^{k+1} .

Si nous rajoutons à (H_3) les hypothèses (H_4) suivantes :

- $\forall\, u \in \Omega = \{(c,z,d,t) \in \mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^p \times [t_0,+\infty[,\ f(c,z,d,t) = \theta_{q+p} \}$
  $\dfrac{\partial f}{\partial \dot{y}\dot{z}}(u)^{-1}$ existe

• $\forall\, t \in [t_0, +\infty[\ ,\ \exists b_z(t) > 0\ /\ \forall\, u \in \Omega_t$

$$\begin{cases} \left\| \dfrac{\partial f}{\partial \dot{y}z}(u)^{-1} \cdot \dfrac{\partial f}{\partial z}(u) \right\| \leq \beta_z(t) \\ \left\| \dfrac{\partial f}{\partial \dot{y}z}(u)^{-1} \cdot \dfrac{\partial f}{\partial t}(u) \right\| \leq \beta_t(t) \end{cases}$$

alors y et z sont prolongeables de manière unique sur $[t_0, +\infty[$ par deux applications $\bar{y}$ et $\bar{z}$, respectivement de classe C^0 et C^1 de telle sorte que (S_6) soit vérifié, si nous posons

$$(S_6) \quad \begin{cases} f(y(t), z(t), \dot{z}(t), t) = \Theta_{q+p} \\ t \in [t_0, +\infty[\\ z(t_0) = \mu \end{cases}$$

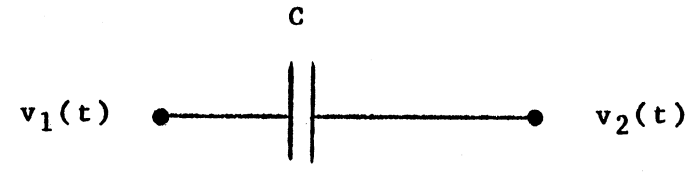
En outre $\bar{y}$ est de classe C^k et $\bar{z}$ est de classe C^{k+1}.

(démonstration en annexe du chapitre I)

Ces conditions d'existence d'une solution ne sont pas entièrement satisfaisantes, surtout à cause de l'exigence d'inversibilité de $\dfrac{\partial f}{\partial \dot{y}z}(u)$. L'idéal serait de pouvoir remplacer cette condition par celle de l'existence d'une permutation σ de $\{1, \ldots, i\}$, d'une permutation μ de $\{1, \ldots, j\}$, d'une permutation ν de $\{1, \ldots, k\}$, avec $i+j+k = p+q$ telle que $\dfrac{\partial f}{\partial \omega}(u)$ soit inversible, si nous posons $\omega = \{y_{\sigma(1)}, \ldots, y_{\sigma(i)}, z_{\mu(1)}, \ldots, z_{\mu(j)}, \dot{z}_{\nu(1)}, \ldots, \dot{z}_{\nu(k)}\}$

L'exemple suivant est instructif à ce sujet :

Le circuit



peut être modélisé par

$$(Cir) \quad \begin{cases} y_1(t) + y_2(t) = 0 \\ z(t) - v_1(t) + v_2(t) = 0 \\ y_1(t) - c.\dot{z}(t) = 0 \end{cases}$$

Les variables algébriques sont y_1, y_2 . z est variable différentielle. (Cir) admet une solution unique qui est :

$$\begin{cases} z(t) = v_2(t) - v_1(t) \\ y_1(t) = c.(\dot{v}_2(t) - \dot{v}_1(t)) \\ y_2(t) = -c.(\dot{v}_2(t) - \dot{v}_1(t)) \end{cases}$$

Or $\dfrac{\partial f}{\partial y \dot{z}}(u) =$ $\begin{array}{c} \begin{matrix} y_1 & y_2 & \dot{z} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & -c \end{bmatrix} \end{array}$ n'est pas inversible

alors que $\dfrac{\partial f}{\partial y z}(u) =$ $\begin{array}{c} \begin{matrix} y_1 & y_2 & z \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$ l'est.

Se passer de l'hypothèse $\dfrac{\partial f}{\partial y \dot{z}}(u)$ inversible, revient à traiter localement des problèmes que nous ne savons pas résoudre pour l'instant(tels les problèmes du type $(z(t) = f(\dot{z}(t),t))$

proposition 9

Supposons (H_3) et (H_4) réalisées.
Supposons en outre (H_5) :

1) $\exists \gamma_1 > 0 \ / \ \forall t \in [t_0, +\infty[$,

$$\left\| \frac{\partial f}{\partial y \dot{z}}(y(t),z(t),\dot{z}(t),t)^{-1} \right\| < \gamma_1$$

ii) $\exists \xi_z > 0, \ \exists \xi_y > 0, \ \exists \xi_{\dot{z}} > 0, \ \exists \gamma_f > 0, \ \exists \gamma_{\partial f} > 0 \ /$

$\cdot \ \forall w \in W, \ \|f(w)\| < \gamma_f$

$$\left\| \frac{\partial f}{\partial y \dot{z}}(w) \right\| < \gamma_{\partial f}$$

• f et $\frac{\partial f}{\partial \dot{y}z}$ sont uniformément continues sur W si nous posons

$$W = \{ (c,z,d,t) \;/\; t \in [t_0,+\infty[\;,\; c \in B_o(y(t),\xi_y) \;,\; z \in B_o(z(t),\xi_z) \;,\; d \in B_o(\dot{z}(t),\xi_{\dot{z}}) \}$$

alors il existe quatre constantes strictement positives η_y , η_z , $\eta_{\dot{z}}$, η_t , telles que :

$$\forall\, t \in [t_0,+\infty[\;,\; \forall\, z \in B_o(z(t),\eta_z) \;,$$
$$\forall\, \tau \in]t-\eta_t, t+\eta_t[\cap [t_0,+\infty[\;,$$

$$\exists!\, c \in B_o(y(t),\eta_y) \;,\; \exists! d \in B_o(\dot{z}(t),\eta_{\dot{z}}) \;/$$
$$f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p}$$

(démonstration en annexe du chapitre I)

Appelons

$$E_{\eta_z} = \{ (z,t) \in \mathbb{R}^p \times [t_0,+\infty[\;/\; \|z - z(t)\| < \eta_z \}$$

$$E_{\eta_y} = \{ (c,t) \in \mathbb{R}^q \times [t_0,+\infty[\;/\; \|c - y(t)\| < \eta_y \}$$

$$E_{\eta_{\dot{z}}} = \{ (d,t) \in \mathbb{R}^p \times [t_0,+\infty[\;/\; \|d - \dot{z}(t)\| < \eta_{\dot{z}} \}$$

La proposition précédente nous permet d'affirmer que $\tilde{\Phi}$, allant de E_{η_z} vers $E_{\eta_y} \times E_{\eta_{\dot{z}}}$, qui au couple (z,t) associe les couples (c,t) et (d,t) tels que $f(c,z,d,t) = \theta_{q+p}$ est une application. Nous appellerons $\Phi_y(z,t)$ et $\Phi_{\dot{z}}(z,t)$ les quantités

c($\in \mathbb{R}^q$) et d($\in \mathbb{R}^p$), $\Phi(z,t)$ le couple (c,d).

L'utilisation en tous points du théorème des fonctions implicites nous apprend que :

$$\begin{cases} \dfrac{\partial \Phi}{\partial z}(z,t) = -\dfrac{\partial f}{\partial y\dot{z}}(\Phi_y(z,t),z,\Phi_{\dot{z}}(z,t),t)^{-1} \cdot \dfrac{\partial f}{\partial z}(\Phi_y(z,t),z,\Phi_{\dot{z}}(z,t),t) \\ \\ \dfrac{\partial \Phi}{\partial t}(z,t) = -\dfrac{\partial f}{\partial y\dot{z}}(\Phi_y(z,t),z,\Phi_{\dot{z}}(z,t),t)^{-1} \cdot \dfrac{\partial f}{\partial t}(\Phi_y(z,t),z,\Phi_{\dot{z}}(z,t),t) \end{cases}$$

Les conditions (H_3), (H_4) et (H_5) nous permettent de sortir du caractère local des solutions d'un système implicite pour aboutir à des comportements globaux.

Cependant, les hypothèses prises sont fortes et il serait souhaitable de pouvoir établir des résultats avec des hypothèses moins contraignantes.

proposition 10

Supposons (H_3), (H_4) et (H_5) réalisées.

Supposons en outre :

i) $\beta_z = \sup_{t \in [t_0,+\infty[} \left(\beta_t(t)\right) < +\infty$

ii) il existe un produit scalaire, noté $(\cdot,\cdot)$, et un réel strictement négatif ℓ tel que, si nous notons $\| \ \|_2$ la norme associée à ce produit scalaire, nous puissions écrire :

$$\forall\ (z_1,t) \in E_{\eta_z}\ ,\ \forall\ (z_2,t) \in E_{\eta_z}\ ,$$

$$(\Phi_{\dot{z}}(z_1,t) - \Phi_{\dot{z}}(z_2,t),\ z_1 - z_2) \leq \ell \cdot \| z_1 - z_2 \|_2^2$$

alors le système (S) $\begin{cases} f(y(t),z(t),\dot{z}(t),t) = \Theta_{q+p} \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$

est fortement stable.

(démonstration en annexe du chapitre I)

Posons

$$q(z,t) = \Phi(z,t) - \Phi(z(t),t) - \frac{\partial \Phi}{\partial z}(z(t),t) \cdot (z - z(t))$$

Si nous supposons (H_6) :

$$\exists \xi_q > 0 \ / \ \forall \varepsilon > 0 \ / \ \forall t \in [t_0,+\infty[\ ,$$

$$\forall z_1 \in B_o(z(t),\xi_q) \ , \ \forall z_2 \in B_o(z(t),\xi_q) \ ,$$

$$\|q(z_1,t) - q(z_2,t)\| < \varepsilon \|z_1 - z_2\|$$

alors nous pourrons, si (H_3), (H_4) et (H_5) sont réalisées, appliquer à (S) les résultats obtenus pour les systèmes explicites.

V)STABILITE DES METHODES DE RESOLUTION

Nous demandons aux méthodes numériques de résolution, en absolu, de fournir une solution "numérique" proche de la solution réelle, si nous maintenons les diverses sources de perturbation(erreurs d'arrondis, erreurs dues aux approximations faites, etc..) en deçà d'un certain seuil. Les méthodes les plus efficaces de ce point de vue sont les **méthodes linéaires multipas** et les **méthodes Runge-Kutta.**

La plus grande souplesse des méthodes linéaires multipas font que nous nous intéresserons exclusivement à celles-ci (pour plus de précision sur la stabilité des méthodes Runge-Kutta, se reporter à l'excellent article de M.Crouzeix sur ce sujet [Crou, 1980])

Les méthodes linéaires multipas sont définies au point t_n par,si nous notons z_n l'approximation numérique de $z(t_n)$:

- un pas $h_n = t_n - t_{n-1}$

- un degré k_n

- des coefficients $(a_{i,n})_{i=0,\ldots,k_n}$ et $(b_{i,n})_{i=0,\ldots k_n}$

- la relation (R_2) $\sum_{i=0}^{k_n} a_{i,n} z_{n-i} + \sum_{i=0}^{k_n} b_{i,n} \dot{z}_{n-i} = \theta_p$

Si b_0 est nul, la méthode est dite **explicite** au point t_n, sinon elle est dite **implicite**.

Ces méthodes permettent de transformer un système algébrodifférentiel

$$(S_T) \quad \begin{cases} f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p} \\ t \in [t_0,T] \\ z(t_0) = \mu \end{cases}$$

en une suite de problèmes

$$(S_n) \quad \begin{cases} F_n(y_n,z_n) = \theta_{q+p} \\ n=1,\ldots \end{cases}$$

où (y_n,z_n) sont une approximation de $(y(t_n),z(t_n))$.

La plupart des résultats disponibles pour l'instant concernent les méthodes linéaires multipas où l'ordre k_n, le pas h_n et les coefficients $(a_{i,n})$ et $(b_{i,n})$ sont constants. Nous appellerons ces méthodes **invariantes**

Cependant des résultats intéressants ont été obtenus très récemment pour les méthodes à pas variables, pour des systèmes algébrodifférentiels vérifiant la condition(R_1). Le lecteur intéressé par ces développements nouveaux pourra consulter avec profit [Dah, 1980] et [Lin, 1982] .

Il est habituel d'écrire pour les méthodes la relation (R_2) sous la forme (R_3)

$$(R_3) : \sum_{i=0}^{k} \alpha_i \cdot z_{n+i} - h \cdot \left(\sum_{i=0}^{k} \beta_i \cdot \dot{z}_{n+i} \right) = \theta_p$$

où k ,h ,$(\alpha_i)_{i=1,..k}$ et $(\beta_i)_{i=1,...k}$ sont des constantes.

Nous dirons qu'un système différentiel est un problème de Cauchy-Lipschitz s'il est de la forme

$$\begin{cases} \dot{z}(t) = f(z(t),t) \\ t \in I \\ z(t_0) = \mu \end{cases}$$

où I est un intervalle de $\mathbb{R}$, f est une fonction continue et lipschitzienne par rapport à la première variable sur I.

(R_3) peut alors être remplacée par

$$(R_4): \quad \sum_{i=0}^{k} \alpha_i \cdot z_{n+i} - h \cdot \left(\sum_{i=0}^{k} \beta_i \cdot f(z_{n+i}, t_{n+i}) \right) = \theta_p$$

Nous pouvons demander légitimement à une méthode de résolution qu'elle donne des résultats d'autant plus précis que le pas choisi est faible.

Nous devons à P.Henrici la formalisation de cette notion et trois théorèmes extrêmement importants sur cette question.

définition (D.10) [Hen, 1962, p 217] :
Une méthode invariante est dite **convergente** si et seulement si, appliquée à un problème de Cauchy-Lipschitz quelconque, en notant z(t) la solution de ce problème et $(z_i(h))_{i=0,..}$ l'approximation numérique obtenue en utilisant la méthode avec un pas h, nous avons :

i) $\forall\, t \in [t_0,T]$, $\lim_{t \to +\infty} z_n(h_n) = z(t)$

en posant $h_n = (t - t_0)/n$

ii) $\lim_{h \to 0} z_i(h) = \mu \quad i=0,...k-1$

l'ordre d'une méthode linéaire multipas invariante est le plus grand entier m tel que

$L_h\phi(t) = 0 \ \forall \ \phi \in \mathbb{R}_m[X]$ où L_h est l'opérateur $\rho(E_h) - hD\sigma(E_h)$

avec :

E_h : $(t \longrightarrow x(t)) \longrightarrow (t \longrightarrow x(t + h))$

(opérateur de "déplacement")

ρ : $\xi \longrightarrow \sum_{i=0}^{k} \alpha_i \cdot \xi^i$

σ : $\xi \longrightarrow \sum_{i=0}^{k} \beta_i \cdot \xi^i$

ρ et σ sont appelés les polynômes associés à la méthode

D : $x \longrightarrow \dot{x}$

Pour une fonction ϕ de classe C^{m+1}, nous avons la relation (R_5) $L_h\phi(t) = -c \cdot h^{m+1} \cdot \phi^{(m+1)}(t) + o(h^{m+1})$, c étant indépendante de ϕ.

théorème 3 [Hen, 1962, p 218]

Une condition **nécessaire** pour la convergence d'une méthode linéaire multipas invariante est que le module des racines du polynôme associé ρ ne dépasse pas 1 et que les racines de module 1 soient simples.

Cette condition est appelée - de manière quelque peu équivoque - **condition de stabilité**

théorème 4 [Hen, 1962, p 224]

Une condition **nécessaire** pour la convergence d'une méthode linéaire multipas invariante est que l'ordre de la

méthode soit au moins égal à un. Une telle méthode sera dite **consistante.**

Une condition **nécessaire et suffisante** de consistance est que les polynômes associés à la méthode ρ et σ vérifient la relation

$$(R_6) \quad \begin{cases} \rho(1) = 0 \\ \sigma(1) = \rho'(1) \end{cases}$$

théorème 5 [Hen, 1962, p 244]

Une condition **nécessaire et suffisante** de convergence pour une méthode linéaire multipas invariante est qu'elle vérifie la condition de stabilité, la condition de consistance, et que toute solution numérique $(z_i(h))_{i=0,\ldots}$ d'un problème de Cauchy-Lipschitz quelconque, calculée en utilisant cette méthode avec un pas h, vérifie

$\lim_{h \to 0} z_i(h) = \mu$ pour i appartenant à $\{ 0,\ldots,k-1 \}$, où μ désigne la condition initiale du problème différentiel.

La condition " $\lim_{h \to 0} z_i(h) = \mu \quad i = 0,\ldots k-1$" provient du fait qu'une méthode multipas(invariante ou non), sauf si elle est de degré 1, ne peut pas démarrer toute seule. Elle a besoin des valeurs $z_i, i=0,\ldots k-1$ qui sont fournies par une autre méthode. Ces valeurs de départ(couramment désignées par "starting values" dans la littérature anglo-saxonne) sont donc astreintes à tendre vers la valeur initiale quand le pas tend vers 0 afin d'assurer la convergence de la méthode.

Dorénavant, nous demanderons à une méthode linéaire multipas invariante de vérifier les conditions du théorème 5.

Nous avons déjà noté que, pour toute application ϕ de classe C^{k+1} ,nous avions la relation :

$$L_h\psi(t) = -c \cdot h^{k+1} \cdot \phi^{(k+1)}(t)$$

Prenons $\phi(t) = e^t$ et posons $e^h = h$. Alors :

$$\rho(\xi) - \sigma(\xi) \cdot \mathrm{Log}(\xi) = -c \cdot (\xi - 1)^{k+1} + o((\xi - 1)^{k+1})$$

Puisque 1 est racine simple de ρ, nous avons $\sigma(1) \neq 0$, et donc σ est non-nul dans un voisinage de 1.

Si nous posons $c^* = c/\sigma(1)$, alors :

$$(R_7) \qquad \mathrm{Log}(\xi) - \rho(\xi)/\sigma(\xi) = c^* \cdot (\xi - 1)^{k+1})$$

c^* est appelée la constante d'erreur de la méthode, et constitue une quantité adéquate pour mesurer la précision de méthodes ayant le même ordre k.

Si ρ et σ ont un P.G.C.D. Δ autre qu'un polynôme constant, nous pouvons utiliser $\rho^* = \rho/\Delta$ et $\sigma^* = \sigma/\Delta$ sans affecter ni la constante d'erreur, ni l'ordre de la méthode. Nous supposerons donc, par souci d'économie, que ρ et σ sont premiers entre eux.

Lorsqu'un système différentiel est instable, il est logique que le pas à utiliser pour suivre de manière satisfaisante la solution réelle soit "petit". En effet, les perturbations dues aux erreurs d'arrondi ou de méthode se répercutent de manière exponentielle sur la suite des calculs et il est nécessaire de limiter de manière radicale ces dernières.

Par contre, il serait intéressant de pouvoir utiliser pour des systèmes stables, des pas de "grande taille", et obtenir néanmoins un résultat proche de la solution réelle.

Pour mener à bien l'étude des méthodes ayant de telles propriétés, Dahlquist [Dahl, 1962] a introduit la notion de système-test

$$(S_\lambda^\mu) \qquad \begin{cases} \dot{z}(t) = \lambda \cdot z(t) \\ t \in [0, +\infty[\\ z(0) = \mu \end{cases}$$

définition (D.11)

Une méthode linéaire multipas invariante est dite A-stable si, appliquée à un (S_λ^μ) quelconque tel que $Re(\lambda) < 0$, toutes les solutions de

$$(R_8) \qquad [\ (\sum_{i=0}^{k} \alpha_i \cdot z_{n+i} - h \cdot (\sum_{i=0}^{k} \beta_i \cdot \dot{z}_{n+i}) = 0\ ]$$

tendent vers 0 lorsque n tends vers l'infini.

Dahlquist a montré qu'une méthode A-stable est forcément implicite($\beta_k \neq 0$) et d'ordre 2 au plus.

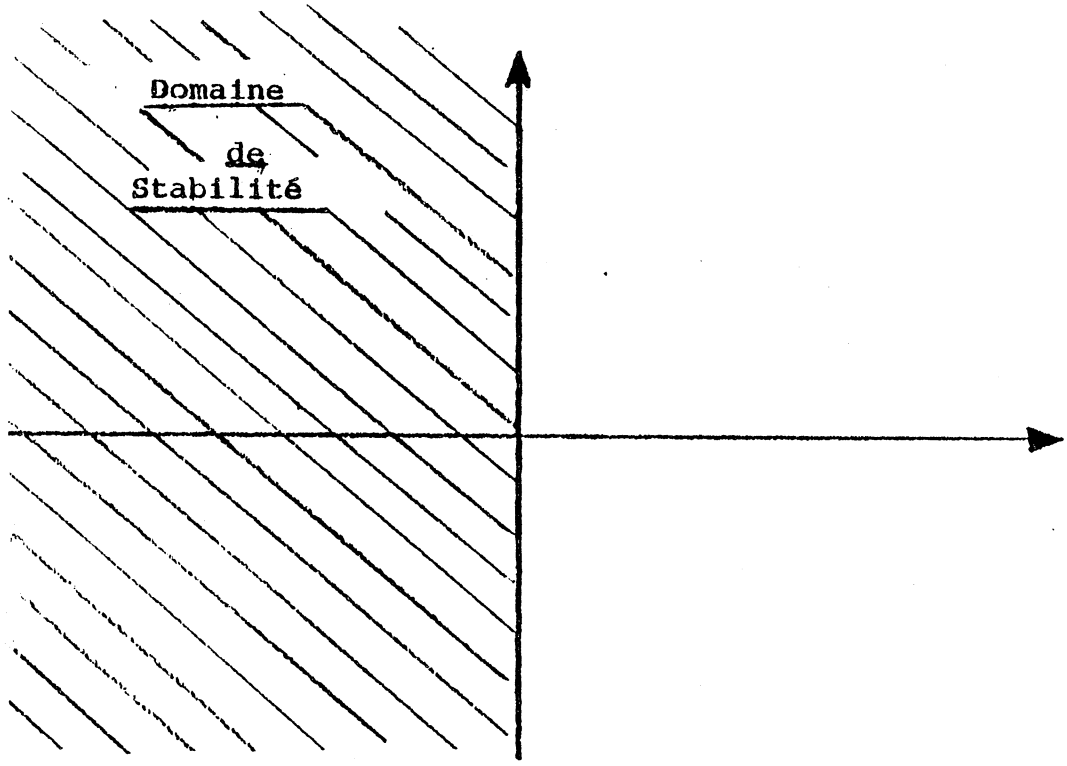
De plus il a montré que la méthode d'ordre 2 la plus efficace (ayant la constante d'erreur c^* la plus petite) était la méthode dite des trapèzes :

$$z_{n+1} = z_n + h/2 \cdot (\ \dot{z}_{n+1} + \dot{z}_n\ )$$

ou encore, en utilisant les polynômes associés à la méthode des trapèzes :

$$\begin{cases} \rho(\xi) = \xi - 1 \\ \sigma(\xi) = (\ \xi + 1\ )/2 \end{cases}$$

Si nous représentons les valeurs $h\lambda$ par les points du plan complexe, alors une méthode A-stable est stable sur tout le demi-plan gauche :



Cependant n'avoir droit qu'à un ordre 2 au plus est un handicap sérieux pour une méthode de résolution. En effet, l'erreur d'une méthode A-stable est au mieux en $o(h^3)$ ce qui est de peu d'importance en face de systèmes stables, mais qui peut se révéler très génant pour des systèmes instables, pour lesquels nous aimerions avoir des ordres aussi élevés que possible. La limite de l'ordre d'une méthode fixe également la limite de sa capacité à suivre les transitions non-linéaires d'une solution. Ainsi, comme nous le verrons dans le chapitre III lors des expérimentations numériques, des problèmes très raides comme celui de l'oscillateur de Van der Pol réclament pour leur résolution des ordres allant jusqu'à 6.

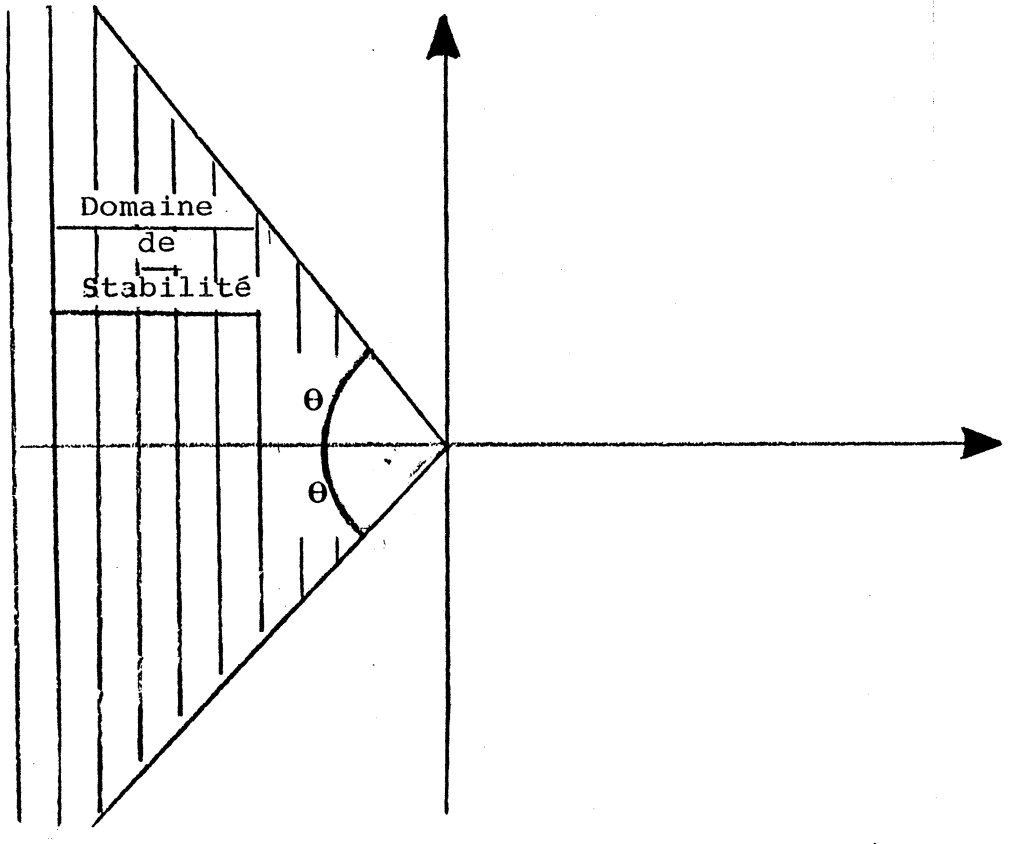
IL est donc nécessaire d'affaiblir la notion d' A-stabilité pour avoir des ordres plus élevés. Cet

affaiblissement peut s'opérer de plusieur manières, qui reviennent toutes à restreindre le domaine sur lequel nous demandons la stabilité.

Les deux principales restrictions introduites à ce jour sont:

• l'**A(θ)-stabilité (D.11)** [Wid, 1967]

Une méthode est A(θ)-stable, $\theta \in [0,\pi/2]$, si, appliquée à un (S_λ^μ) quelconque tel que $\lambda \neq 0$ et $Arg(-\lambda) < \theta$, toutes les solutions de (R_8) tendent vers 0 lorsque n tend vers l'infini.

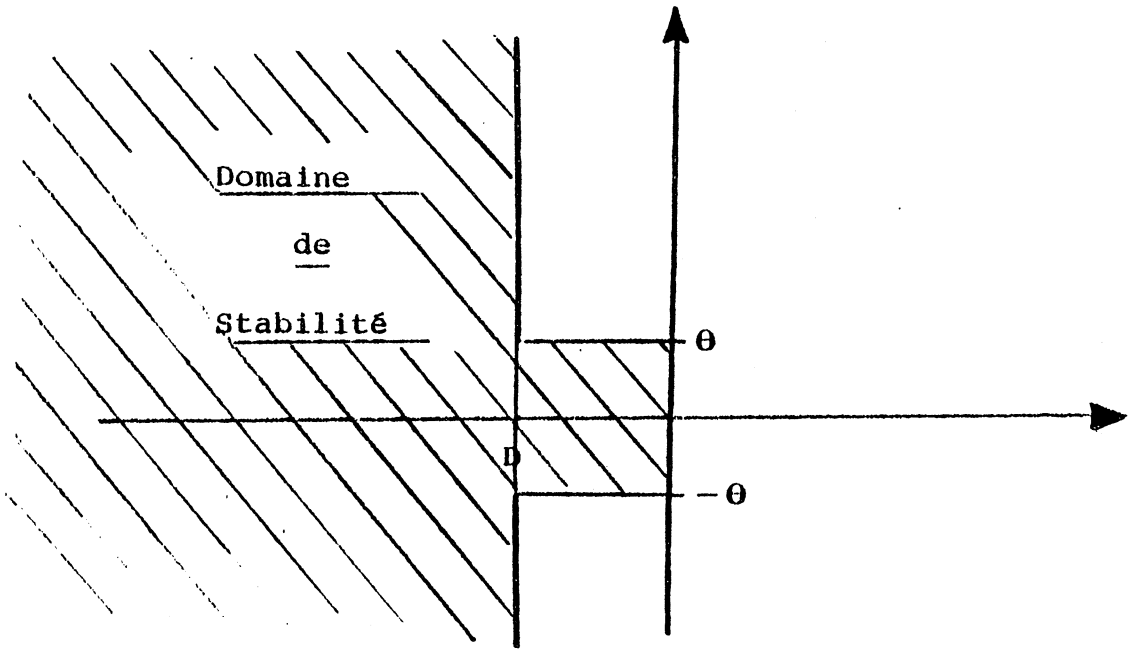


Nous pouvons remarquer que pour $\theta = \pi/2$, nous retrouvons l'A-stabilité. Widlund a montré qu'il est possible d'obtenir des méthodes d'ordre 3 ou 4 pour θ appartenant à $[0,\pi/2[$ (les constantes d'erreurs de ces méthodes tendent vers l'infini lorsque θ tend vers $\pi/2$)

• la **quasi-stabilité** (ou stiff-stabilité) (D.13) [Gear, 1967]

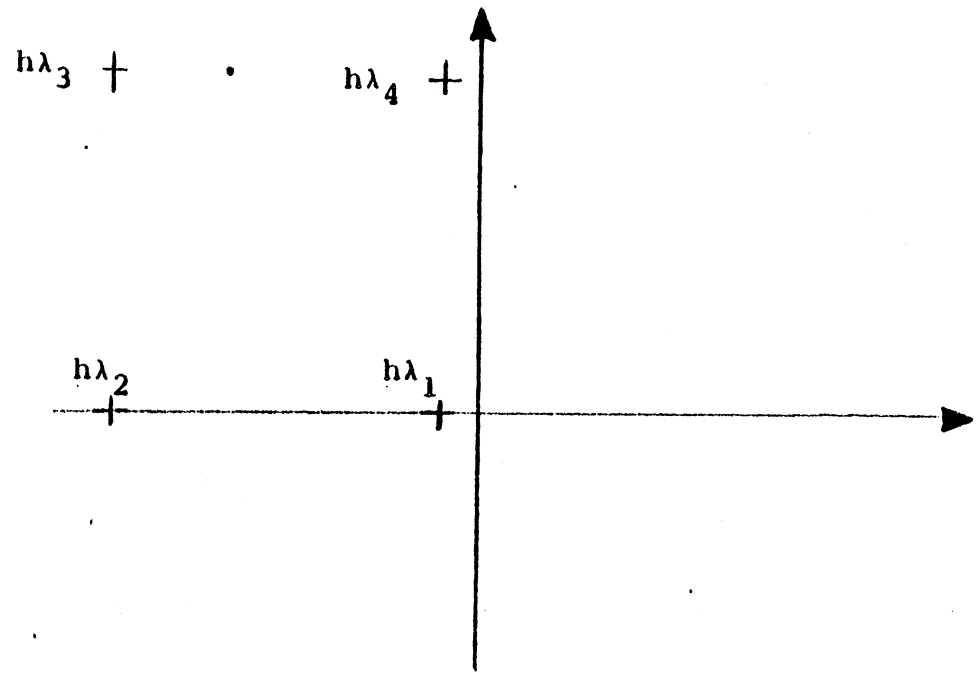
Une méthode est dite quasi-stable s'il existe deux réels $D (\in \mathbb{R}_-)$ et $\theta (\in [0,2\pi[)$ tels que, la méthode étant appliquée

à un (S_λ^μ) quelconque avec un pas h donné, toutes les solutions de (R_8) tendent vers 0 lorsque n tend vers l'infini dès que $Re(h\lambda)$ D ou que $h\lambda \in S_1 = \{ u \in \mathbb{C} / D < Re(u) \leq 0, |Im(u)| < \theta \}$



Cette dernière restriction se justifie par les considérations suivantes :

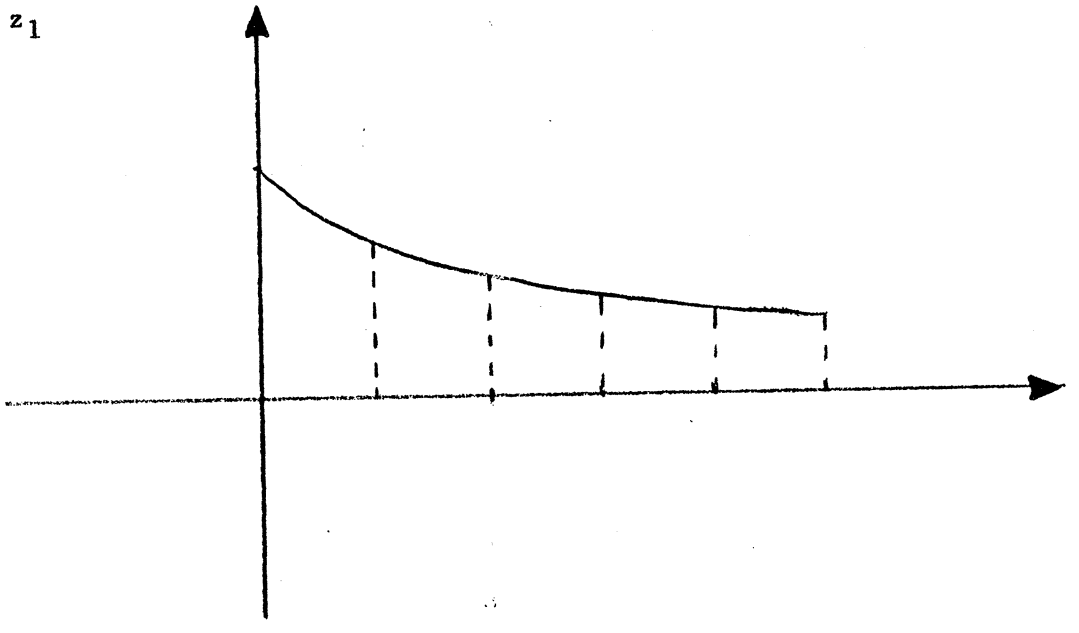
Soient quatre valeurs $h\lambda_1$, $h\lambda_2$, $h\lambda_3$, et $h\lambda_4$ disposées dans le plan complexe comme ci-dessous:



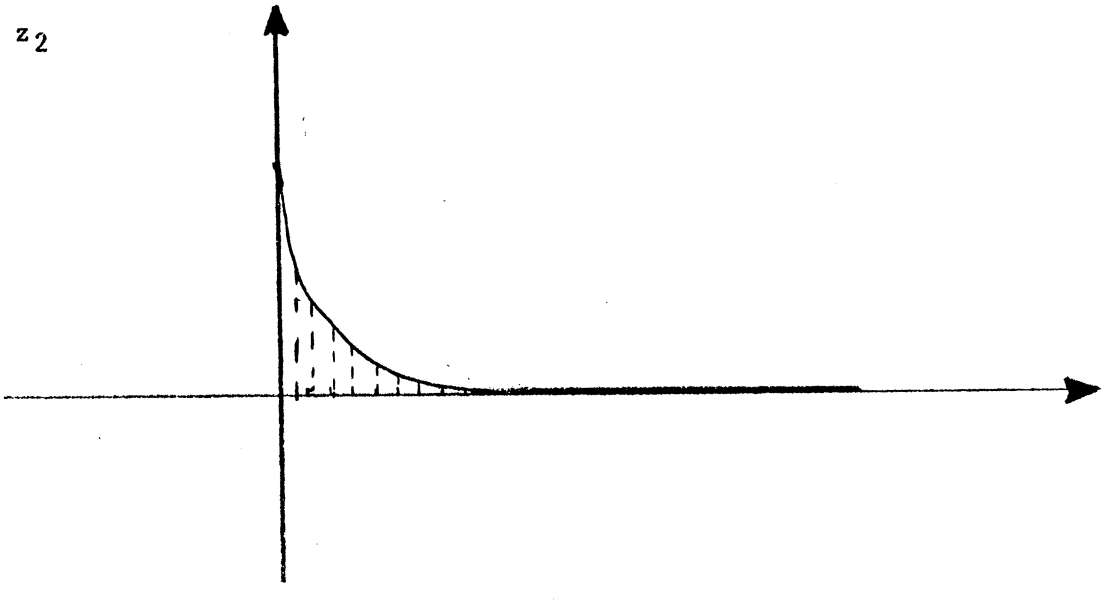
Associons-leur les équations-test

$$\begin{cases} \dot{z}_i(t) = \lambda_i \cdot z_i(t) \\ z_i(0) = 1 \\ (i = 1,2,3,4) \end{cases}$$

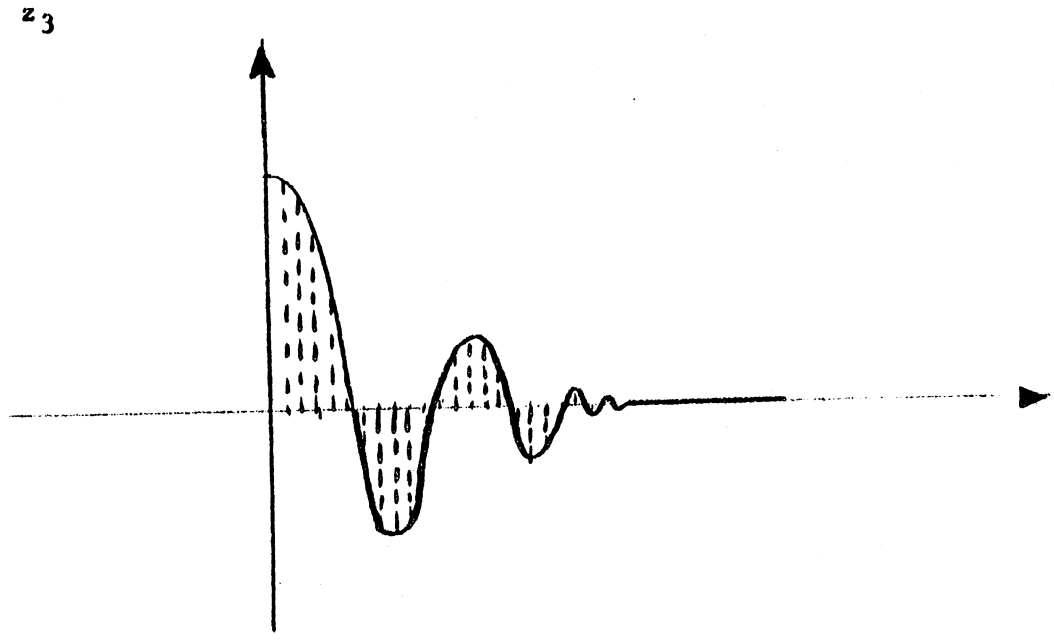
L'allure des solutions va être :



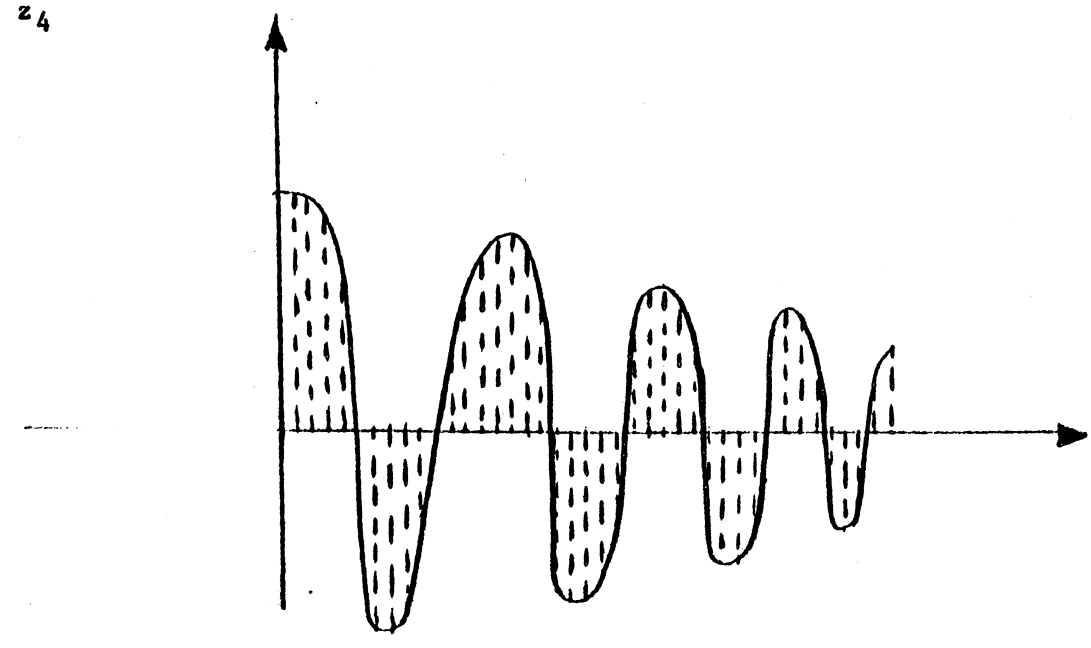
Il faut un temps assez long pour que z_1 devienne négligeable, mais un grand pas est acceptable.



Il faut un petit laps de temps pour que z_2 devienne négligeable.Un petit pas est nécessaire.



Il faut un petit laps de temps pour que z_3 devienne négligeable, mais un très petit pas est nécessaire.



Il faut un grand laps de temps pour que z_4 devienne négligeable, et il est nécessaire d'avoir un très petit pas d'intégration.

Nous voyons clairement qu'en présence d'une valeur de type λ_4, il faudra de toute façon avoir un pas très petit,

ramenant $h\lambda$ dans S_1; Il est donc inutile de réclamer la stabilité pour les valeurs de $h\lambda$ appartenant à S_2.

Gear a montré qu'il existait des méthodes quasi-stables d'ordre allant jusqu'à 6 inclus en prenant comme polynôme associé $\sigma(\xi) = \xi^k$.

Les domaines de stabilité d'une méthode peuvent être considérés comme l'ensemble des $h\lambda$ tels que les racines de $\rho(\xi) - h\lambda \cdot \sigma(\xi)$ soient de module inférieur à un.

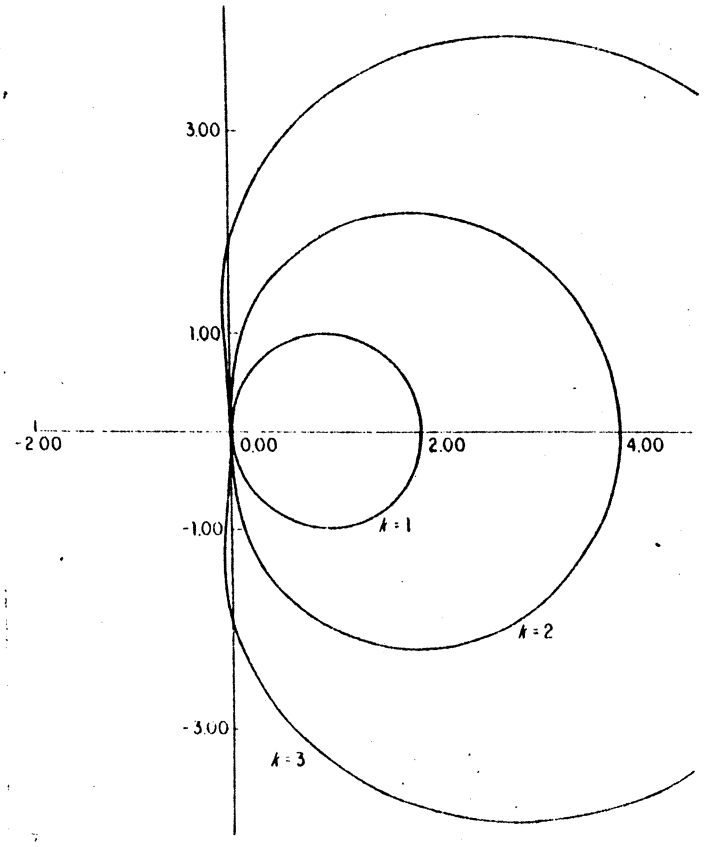
En nous servant de (R_7), nous voyons que pour un σ donné, il n'existe qu'un seul ρ tel que la relation soit vérifiée à l'ordre voulu.

Puis en calculant $\mu = \dfrac{\rho(e^{i\theta})}{\sigma(e^{i\theta})}$,θ C $[0,2\pi]$,nous obtenons,

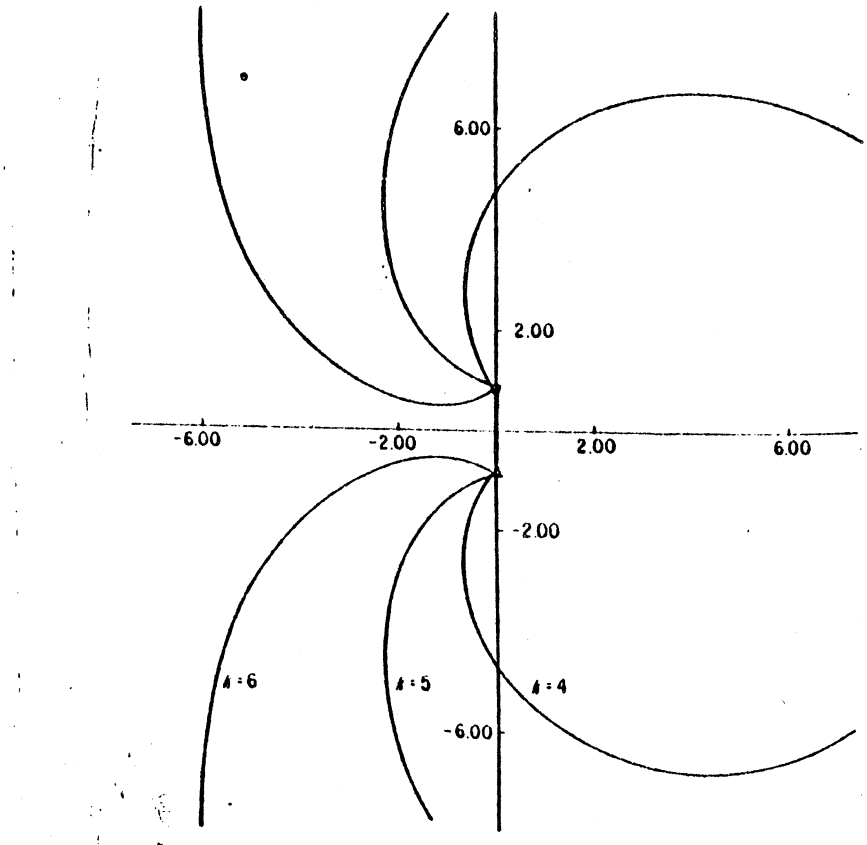
pour k allant de 1 à 6, deux ensembles déconnectés (S_k) et (I_k) ,(S_k) étant le domaine relié à l'infini.

Pour $\mu = \infty$, les racines de $\rho(\xi) + \mu \cdot \sigma(\xi)$ sont celles de $\sigma(\xi)$, donc toutes égales à 0. Puisque les racines sont des fonctions continues de μ , toutes les racines de $\rho(\xi) + \mu \cdot \sigma(\xi)$ sont de module plus petit que un.

domaines de stabilité pour k allant de 1 à 3 :



domaines de stabilité pour k allant de 4 à 6 :



Nous trouvons pour D les valeurs suivantes :

k	D
1	0
2	0
3	- 0.1
4	- 0.7
5	- 2.4
6	- 6.1

Il est difficile de représenter une fonction avec moins de 6 à 10 points par cycle de la fréquence maximale. Il faut donc s'attendre à que ces méthodes ne soient valables que pour θ appartenant à $[\pi/5, 2\pi/5]$.

De plus nous sommes en mesure de montrer la

proposition 11

Soit le système (S)
$$\begin{cases} \dot{z}(t) = f(z(t),t) \\ t \in [t_0,+\infty[\\ z(t_0) = \mu \end{cases}$$

où f est un fonction de $\mathbb{R}^p \times [t_0,+\infty[$ dans $\mathbb{R}^p$.

Supposons qu'il existe $t_1 \in [t_0,+\infty[$, $\mu \in \mathbb{R}$ et un produit scalaire (.,.) tels que :

$\forall\, t \in [t_1,+\infty[$, $\forall\, z_1 \in \mathbb{R}^p$, $\forall\, z_2 \in \mathbb{R}^p$,

$$(H_7) \qquad (f(z_1,t) - f(z_2,t), z_1 - z_2) \leq \mu . \|z_1 - z_2\|_2^2$$

Soient $\alpha_0, \alpha_1, \ldots, \alpha_k$ les coefficients de la méthode d'ordre k quasi-stable.

Soient deux suites $(z_i)_{i=0,..}$ et $(z'_i)_{i=0,..}$ déterminées par :

$$\begin{cases} t_n = n.h + t_1 \\ \alpha_0 . z_n - h.f(z_n,t_n) = -\sum_{i=1}^{k} \alpha_i . z_{n-i} \\ \alpha_0 . z'_n - h.f(z'_n,t_n) = -\sum_{i=1}^{k} \alpha_i . z'_{n-i} \\ (\, n = k, \ldots . \,) \end{cases}$$

et la connaissance de $(z_0, \ldots, z_{k-1})$ et $(z'_0, \ldots, z'_{k-1})$

si $\mu.h < \alpha_0 - \sum_{i=1}^{k} |\alpha_i|$

alors

1) il existe r_k appartenant à $[0,1[$ et λ_k appartenant à $\mathbb{R}^+_*$ tels que :

$\forall\, n \in \mathbb{N}$ $\|z_n - z'_n\|_2 \leq \lambda_k . r_k^{\,n}$

2) si pour tout i allant de 0 à k-1 $\|z_i - z'_i\|_2 < \delta$,
alors
$\forall\, n \in \mathbb{N} \quad \|z_n - z'_n\|_2 < \delta$

(démonstration en annexe du chapitre I)

Ces propositions nous permettent d'avoir le pendant pour les approximations numériques, des propriétés de stabilité uniforme (grâce à 2)) et d'attractivité uniforme(grâce à 1)).

proposition 12

Supposons (H_7) réalisée. Soit $(z_n)_{n=0,..}$ une suite définie de la même façon que dans le corps de la proposition 11.

Posons $e_n = \sum_{i=0}^{k} \alpha_i . z(t_{n-i}) - h.f(z(t_n), t_n)$
alors

$$\begin{cases} \sup_{n>0} \|e_n\|_2 < \delta/(2.\sum_{i=1}^{k} \alpha_i \,)) \\ \|z_i - z(t_i)\| < \delta/2 \\ \quad i = 0, \ldots k-1 \end{cases} \Rightarrow \forall\, n \in \mathbb{N} \quad \|z_n - z(t_n)\| < \delta$$

(démonstration en annexe du chapitre I)

CONCLUSION

Ce chapitre nous a permis de préciser les contours de cette notion quelque peu intuitive qu'est la stabilité, et nous a donné une meilleure maîtrise des critères qui permettent de choisir une méthode de résolution numérique parmi toutes celles existant.De même, nous avons pu poser quelques jalons pour l'étude des systèmes algébrodifférentiels, qui revêtent une grande importance pour l'étude des circuits électriques.

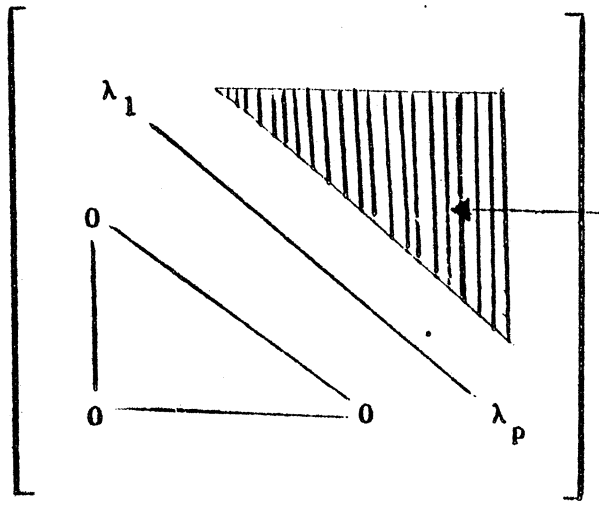
L'étude des propriétés de stabilité des schémas numériques utilisés pour la résolution des équations différentielles ordinaires et des systèmes algébrodifférentiels mérite que des travaux y soient consacrés avec l'ambition de mettre au point, à terme, de nouvelles méthodes numériques plus adaptées aux problèmes raides.

ANNEXE DU CHAPITRE I

démonstration de la proposition 1

D'après un lemme de Arnold [Arn, 1974, p 148], nous pouvons, pour tout ε_0 , choisir une base B_{ε_0} de $\mathbb{C}^p$ telle que, si nous notons P la matrice de passage de la base canonique à cette base B_{ε_0} , $J = P^{-1} \cdot A \cdot P$ est triangulaire supérieure et tous les éléments situés strictement au-dessus de la diagonale sont de module inférieur à ε_0 .

$J =$



Soit la fonction $j(x) = \sum_{k,l=1}^{p} j_{kl} \cdot x_k \cdot \bar{x}_l$

si $x \in \mathbb{C}^p$ et $J = (j_{kl})_{k,l=1,\ldots p}$. J se décompose en $a(x) = \sum_{i=1}^{p} \lambda_i \cdot \| x_i \|^2$ et $b(x) = \sum_{k\ l} j_{kl} \cdot x_k \cdot \bar{x}_l$.

Considérons la fonction de $\mathbb{C}^p$ dans $\mathbb{C}$ qui à z associe $L(z) = Re(Jz \cdot z)$.

Nous pouvons écrire $L(z) = Re(a(z)) + Re(b(z))$

La fonction a est indépendante du choix de la base précédente et ne dépend que des valeurs propres de A.

$$\mathrm{Re}(a(z)) = \sum_{i=1}^{p} \mathrm{Re}(\lambda_i) \cdot \| z_i \|^2 \leq \max_{i=1,..p} \mathrm{Re}(\lambda_i) \cdot \left(\sum_{i=1}^{p} \| z_i \|^2 \right)$$

$$\| \mathrm{Re}(b(z)) \| = \| \mathrm{Re}\left(\sum_{k<l} j_{kl} \cdot z_k \cdot \bar{z}_l \right) \| \leq \sum_{k<l} \| j_{kl} \cdot z_k \cdot \bar{z}_l \|$$

$$\leq \varepsilon_0 \cdot \sum_{k<l} \| \bar{z}_k \cdot z_l \|$$

$$\leq \varepsilon_0 \cdot \left(\sum_{k=1}^{p} \| z_k \| \right)^2$$

$\| z \|_1^2 = \left(\sum_{i=1}^{p} \| z_i \| \right)^2$ et $\| z \|_2^2 = \left(\sum_{i=1}^{p} \| z_i \|^2 \right)$ sont équivalentes. Donc il existe β tel que

$$\| \mathrm{Re}(b(z)) \| \leq \beta^2 \cdot \varepsilon_0 \cdot \| z \|_2^2 \ .$$

Posons $\ell = \max_{i=1,..p} \mathrm{Re}(\lambda_i)$. D'après nos hypothèses ℓ est strictement négatif. Si nous prenons $\varepsilon_0 \leq (-\ell)/(2\beta^2)$, nous avons, puisque $\mathrm{Re}(a(z)) \leq \ell \cdot \| z \|_2^2$:

$$\mathrm{Re}(a(z) + b(z)) \leq \tfrac{1}{2} \ell \cdot \| z \|_2^2$$

ou encore

$$L(z) \leq \tfrac{1}{2} \ell \cdot \| z \|_2^2$$

Posons $B_{\varepsilon_0} = (b_i)_{i=1,..p}$.

Soit la forme hermitienne notée $(.,.)_{\varepsilon_0}$ qui est définie par $(b_i, b_j) = \delta_{ij}$.
Soit z un vecteur de $\mathbb{C}^p$.
Soient $(z_i)_{i=1,...p}$ ses coordonnées dans la base canonique et $(y_i)_{i=1,...p}$ ses coordonnées dans la base B_{ε_0} .

Nous avons $(y_i) = P^{-1} \cdot (z_i)$

Donc $(A\cdot z, z) = (J\cdot(y_i))\cdot(\bar{y}_i)$

Considérons la fonction de $\mathbb{R}^p \times \mathbb{R}^p$ vers $\mathbb{R}$ notée $(\cdot,\cdot)$ et telle que $(x,y) = Re((x,y)_{\varepsilon_0})$.

$(\cdot,\cdot)$ est un produit scalaire. En effet :

- $(x,y) = Re((x,y)_{\varepsilon_0})$
  $= Re((y,x)_{\varepsilon_0})$
  $= (y,x)$

- $\forall (\alpha,\beta) \in \mathbb{R}^2, (\alpha\cdot x + \beta\cdot y, z) = Re((\alpha.x+\beta.y,z)_{\varepsilon_0})$
  $= Re(\alpha.(x,z)_{\varepsilon_0} + \beta.(y,z)_{\varepsilon_0})$

- $(x,x) = Re((x,x)_{\varepsilon_0})$
  $= (x,x)_{\varepsilon_0}$, qui est strictement positif lorsque x est différent de zéro.

Nous avons l'égalité $(Az,z) = L(z)$.

Il existe donc un produit scalaire tel que, si nous le notons $(\cdot,\cdot)$, nous ayons :
$\forall z_1 \in \mathbb{R}^p, \forall z_2 \in \mathbb{R}^p$
$(A\cdot(z_1 - z_2), z_1 - z_2) \leq (\ell/2) \cdot \|z_1 - z_2\|_2^2$
avec $\ell \in \mathbb{R}^*$.

(S_1) est donc fortement stable □

démonstration de la proposition 2

Nous savons qu'il existe un produit scalaire, noté $(\cdot,\cdot)$ tel que :
$\forall z_1 \in \mathbb{R}^p, \forall z_2 \in \mathbb{R}^p,$
$(A\cdot(z_1 - z_2), z_1 - z_2) \leq (\ell/2) \cdot \|z_1 - z_2\|_2^2$
avec $\ell < 0$ et $\ell = \max_{i=1,\ldots,p} Re(\lambda_i)$

D'après nos hypothèses, nous pouvons écrire :

$\exists \sigma > 0 \ / \ \forall t > t_0$

$\forall z_1 \in B_o(z(t),\xi_z)$, $\forall z_2 \in B_o(z(t),\xi_z)$

$(\|z_1 - z_2\| < 2\sigma)$

$\Rightarrow (\|f(z_1,t) - f(z_2,t)\| < (-\ell)/4 \cdot \|z_1 - z_2\|)$

ou encore, puisque les normes $\|\cdot\|$ et $\|\cdot\|_2$ sont équivalentes

$\exists \sigma > 0 \ / \ \forall t \in [t_0,+\infty[$,

$\forall z_1 \in B_o^2(z(t),\sigma)$, $\forall z_2 \in B_o^2(z(t),\sigma)$,

$\|f(z_1,t) - f(z_2,t)\|_2 < (-\ell/4) \cdot \|z_1 - z_2\|_2$

donc :

$\forall t \in [t_0,+\infty[$, $\forall z_1 \in B_o^2(z(t),\sigma)$, $\forall z_2 \in B_o^2(z(t),\sigma)$,

$((A\cdot z_1 + f(z_1,t)) - (A\cdot z_2 + f(z_2,t)), z_1 - z_2) < (-\ell/4)\cdot \|z_1 - z_2\|_2^2$

(S_2) est donc fortement stable □

démonstration de la proposition 3

$\forall t \in [t_0,+\infty[$, $\forall y \in B_o(y(t),\beta(t))$,

$$L_{T*}(L_T(f))(z,t) = \frac{\partial T^*}{\partial y}(T(z,t),t)\cdot(L_T(f))\cdot(T(z,t),t) + \frac{\partial T^*}{\partial t}(T^*(z,t),t)$$

$$\Rightarrow L_{T*}(L_T(f))(z,t) = \frac{\partial T^*}{\partial y}(T(z,t),t)\cdot\left[\frac{\partial T}{\partial z}(T^*(T(z,t),t),t)\cdot f(T^*(T(z,t),t,t) + \frac{\partial T}{\partial t}(T^*(T(z,t),t),t)\right] + \frac{\partial T^*}{\partial t}(T(z,t),t) \qquad (9)$$

Or $\forall t \in [t_0,+\infty[$, $\forall z \in B_o(z(t),\alpha(t))$,

$T^*(T(z,t),t) = z$

Donc

$$(1) \quad \Rightarrow \quad L_{T*}(L_T(f))(z,t) = \frac{\partial T^*}{\partial y}(T(z,t),t)\cdot\left[\frac{\partial T}{\partial z}(z,t)\cdot f(z,t) + \frac{\partial T}{\partial t}(z,t)\right] + \frac{\partial T^*}{\partial t}(T(z,t),t)$$

De plus

$$\frac{\partial T^*}{\partial y}(T(z,t),t)\cdot\left[\frac{\partial T}{\partial z}(z,t)\cdot dz + \frac{\partial T}{\partial t}(z,t)\cdot dt\right] + \frac{\partial T^*}{\partial t}(T(z,t),t)\cdot dt = dz$$

d'où les égalités
$$\begin{cases} \frac{\partial T^*}{\partial y}(T(z,t),t)\cdot\frac{\partial T}{\partial z}(z,t) = I_p \\ \frac{\partial T^*}{\partial y}(T(z,t),t)\cdot\frac{\partial T}{\partial t}(z,t) + \frac{\partial T^*}{\partial t}(T(z,t),t) = 0 \end{cases}$$

en notant I_p l'identité sur $\mathbb{R}^p$.

Nous pouvons alors écrire :
$\forall\, t \in [t_0,+\infty[$, $\forall\, y \in B_o(y(t),\beta(t))$,
$L_{T*}(L_T(f))(z,t) = f(z,t)$, ou encore $L_{T*}(L_T(f)) = f$

T et T* jouant un rôle symétrique,une même démonstration vaut pour la proposition 3 ($L_T(L_{T*}(f^*)) = f^*$)

démonstration de la proposition 4

Soit $f \in E^{E_\alpha}_{\beta*}$ une application lipschitzienne par rapport à la première variable.

$\forall\, t \in [t_0,+\infty[$, $\forall\, y_1 \in B_o(y(t),\beta(t))$, $\forall\, y_2 \in B_o(y(t),\beta(t))$,

$\| [L_T(f)](y_1,t) - [L_T(f)(y_2,t)\| =$

$\| \left(\frac{\partial T}{\partial z}(T^*(y1,t),t)\cdot f(T^*(y_1,t),t) + \frac{\partial T}{\partial t}(T^*(y_1,t),t)\right)$

$$- \quad (\frac{\partial T}{\partial z}(T^*(y_2,t),t)\cdot f(T^*(y_2,t) + \frac{\partial T}{\partial t}(T^*(y_2,t),t)) \|$$

$$\leq \|\frac{\partial T}{\partial z}(T^*(y_1,t),t)\| \cdot \|f(T^*(y_1,t),t) - f(T^*(y_2,t),t)\|$$

$$+ \|\frac{\partial T}{\partial z}(T^*(y_1,t),t) - \frac{\partial T}{\partial z}(T^*(y_2,t),t)\| \cdot \|f(T^*(y_2,t),t)\|$$

$$+ \|\frac{\partial T}{\partial t}(T^*(y_1,t),t) - \frac{\partial T}{\partial t}(T^*(y_2,t),t)\|$$

$\frac{\partial T}{\partial z}$, $\frac{\partial T}{\partial t}$ et f étant lipschitziennes par rapport à la première variable, nous pouvons écrire :

$\exists \gamma > 0 /$

$\forall\ t \in [t_0,+\infty[\ ,\ \forall\ y_1 \in B_o(y(t),\beta(t)),\ \forall\ y_2 \in B_o(y(t),\beta(t)),$

$\|[L_T(f)](y_1,t) - [L_T(f)](y_2,t)\| \leq \gamma \cdot \|T^*(y_1,t) - T^*(y_2,t)\|$

dT* étant bornée, T* est lipschitzienne sur $B_o(y(t),\beta(t))$ qui est connexe.

$\exists \gamma > 0 /$

$\forall\ t \in [t_0,+\infty[,\ \forall\ y_1 \in B_o(y(t),\beta(t)),\ \forall\ y_2 \in B_o(y(t),\beta(t)),$

$\|[L_T(f)](y_1,t) - [L_T(f)](y_2,t)\| \leq \gamma^* \cdot \|y_1 - y_2\|$

T et T* jouant un rôle symétrique, la même démonstration vaut pour montrer que "si $f^* \in E_\alpha^{E_\beta^*}$ est lipschitzienne par rapport à la première variable, alors $L_{T^*}(f^*)$ l'est aussi".

démonstration de la proposition 5

Nous allons d'abord traiter les stabilités autres que fortes. Nous reprendrons pour cela, de manière très large, les démonstrations faites par Lefschetz [Lef, 1957, p.80] en leur

donnant une nouvelle forme, plus conforme aux différentes définitions de la stabilité données dans le présent chapitre.

Considérons les systèmes

$$(S_{\varepsilon,t_1}) \quad \begin{cases} \dot{z}(t) = f(z_\varepsilon(t),t) \\ t \in [t_1,+\infty[\\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \end{cases}$$

et $(S^*_{\varepsilon^*,t_1})$

$$\begin{cases} \dot{y}_{\varepsilon^*}(t) = f^*(y_\varepsilon(t),t) \\ t \in [t_1,+\infty[\\ y_{\varepsilon^*}(t_1) = y(t_1) + \varepsilon^*_1 \end{cases}$$

où nous posons $f^* = L_T(f)$.

Si nous prenons $\varepsilon_1 = T^*(y_\varepsilon(t_1),t_1) - T^*(y(t_1),t_1)$, alors $y_{\varepsilon^*}(t) = T(z_\varepsilon(t),t)$.

D'après le iii) de la définition d'une transformation fortement régulière (définition (D.8)) nous avons :

$\forall\ \theta \in]0,\hat{\beta}/2[,\ \exists\ \theta_1(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$\hat{\beta} > \|y - y(t)\| > \theta \Rightarrow \|T^*(y,t) - T^*(y,(t),t)\| > \theta_1(\theta)$

ou encore

$\forall\ \theta \in]0,\hat{\beta}/2[,\ \exists \theta_1(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$$\|T^*(y,t) - T^*(y(t),t)\| < \theta_1(\theta) \Rightarrow \begin{cases} \|y - y(t)\| < \theta \\ \text{ou} \\ \|y - y(t)\| > \hat{\beta} \end{cases}$$

τ_t^{-1} étant un homéomophisme de V(t) vers U(t), si nous choisissons $\theta_1(\theta)$ inférieur à α, nous pouvons écrire :

$\forall\ \theta \in\ ]0, /2[,\quad \theta_1(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$$\|z - z(t)\| < \theta_1(\theta) \quad \Rightarrow \quad \begin{cases} \|T(z,t) - T(z(t),t)\| < \theta \\ \quad \text{ou} \\ \|T(z,t) - T(z(t),t)\| > \hat{\beta} \end{cases}$$

τ_t est continue de U(t) vers V(t)

$B_o(z(t),\theta_1(\theta))$ est connexe et contenue dans U(t),

donc $\tau_t(B_o(z(t),\theta_1(\theta))$ est contenue dans V(t) et est connexe.

Cela nous permet d'écrire :

$\forall\ \theta \in\ ]0,\hat{\beta}/2[,\ \exists\ \theta_1(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$\|z - z(t)\| < \theta_1(\theta) \Rightarrow \|T(z,t) - T(z(t),t)\| < \theta$

ou encore :

(1) $\forall\ \theta \in\ ]0,\hat{\beta}[,\ \exists\ \theta_1(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$\|z - z(t)\| < \theta_1(\theta) \Rightarrow \|y - y(t)\| < \theta$

T et T* jouant un rôle symétrique nous pouvons écrire :

(2) $\forall\ \theta \in\ ]0,\alpha/2[,\ \exists\ \theta_2(\theta) > 0\ /\ \forall\ t \in [t_1,+\infty[,$

$\|y - y(t)\| < \theta_2(\theta) \Rightarrow \|z - z(t)\| < \theta$

Ecrivons maintenant les définitions de a)la stabilité et de b)la stabilité uniforme(définitions (D.1) et (D.2)) :

a) $\exists\ \rho > 0\ /\ \forall\ \delta \in\ ]0,\rho[,\ \exists\ \sigma(\delta) > 0\ /$

$\varepsilon_1 < \sigma(\delta) \Rightarrow \forall\ t \in\ [t_1,+\infty[\ ,\ \|z(t) - z_\varepsilon(t)\| < \delta$

b) $\exists \rho > 0 \ / \ \forall \ \delta \in]0,\rho[, \ \exists \sigma(\delta) > 0 \ / \ \forall \ t_1 \in [t_1,+\infty[,$

$\|\varepsilon_1\| < \sigma(\delta) \Rightarrow \forall \ t \in [t_1,+\infty[\ , \ \|z(t) - z_\varepsilon(t)\| < \delta$

Pour $\theta_1(\theta)$ donné, nous avons :

d'après a)

(3) $\exists \ \sigma(\theta) > 0 \ /$

$\|\varepsilon_1\| < \sigma(\theta) \Rightarrow \forall \ t \in [t_1,+\infty[\ , \ \|z(t) - z_\varepsilon(t)\| < \theta_1(\theta)$

et d'après b)

(4) $\exists \theta_2(\theta) > 0 \ /$

$(\forall \ t_1 \in [t_0,+\infty[\ , \ \|\varepsilon_1\| < \sigma(\theta)) \Rightarrow$

$(\forall \ t \in [t_1,+\infty[\ , \ \|z(t) - z_\varepsilon(t)\| < \theta_1(\theta))$

(2) nous permet d'écrire :

$\exists \theta_2(\theta) > 0 \ / \ \forall \ t \in [t_1,+\infty[\ , \ \|\varepsilon_1^*\| < \theta_2(\theta) \Rightarrow \|\varepsilon_1\| < \sigma(\theta)$

(3) combiné avec (2) et (1) donne

$\forall \ \theta \in]0,\hat{\beta}/2[, \ \exists \theta_2(\theta) > 0 \ / \ \forall \ t_1 \in [t_0,+\infty[, \forall t \in [t_1,+\infty[$

$\|\varepsilon_1^*\| < \theta_2(\theta) \Rightarrow \|y_{\varepsilon^*}(t) - y(t)\| < \theta$

(4) combiné avec (2) et (1) donne

$\forall \ \theta \in]0,\hat{\beta}/2[, \ \exists \theta_2(\theta) > 0 \ / \ \forall \ t_1 \in [t_0,+\infty[, \forall t \in [t_1,+\infty[$

$\|\varepsilon_1^*\| < \theta_2(\theta) \Rightarrow \|y_{\varepsilon^*}(t) - y(t)\| < \theta$

Donc, si (S) est stable(resp. uniformément stable) alors (S*) est stable(resp. uniformément stable).

De plus, puisque T est continue,

" $\lim_{t \to +\infty} \|z(t) - z_\varepsilon(t)\| = 0$" implique

" $\lim_{t \to +\infty} \|y(t) - y_{\varepsilon*}(t)\| = 0$" . Donc si (S) est attractif, (S*) l'est aussi, de même pour l'attractivité uniforme, la stabilité asymptotique et la stabilité uniformément asymptotique (définitions (D.3),(D.4),(D.5) et (D.6))

Supposons que (S) soit fortement stable.

Considérons

$$(S^*_{\varepsilon*,t_1}) \quad \begin{cases} \dot{y}_{\varepsilon*}(t) = f^*(y_{\varepsilon*}(t),t) + \varepsilon^*(t) \\ t \in [t_1,+\infty[\\ y_{\varepsilon*}(t_1) = y(t_1) + \varepsilon_1^* \end{cases}$$

Si nous prenons $\varepsilon(t) = \frac{\partial T^*}{\partial y}(y_{\varepsilon*}(t),t)\cdot\varepsilon^*(t)$, pour $t \in [t_1,+\infty[$ et $\varepsilon_1 = T^*(y_\varepsilon(t_1),t_1) - T^*(y(t_1),t_1)$ alors

$y_{\varepsilon*}(t) = T(z_\varepsilon(t),t)$.

Le fait que T* soit continue et que $\frac{\partial T^*}{\partial y}$ soit bornée nous permet d'écrire

$\forall\ \sigma \in]0,\hat{\alpha}/2[,\ \exists\ \eta(\sigma) > 0\ /\ \forall\ t_1 \in [t_1,+\infty[$

$$\begin{cases} \|\varepsilon_1^*\| < \eta(\sigma) \\ \left(\sup_{t \in [t_1,+\infty[} \|\varepsilon^*(t)\|\right) < \eta(\sigma) \end{cases} \Rightarrow \begin{cases} \|\varepsilon_1\| < \sigma \\ \left(\sup_{t \in [t_1,+\infty[} \|\varepsilon(t)\|\right) < \sigma \end{cases}$$

Ecrivons maintenant la définition (D.7) de la forte stabilité :

c) $\exists \rho > 0,\ \forall\ \xi \in\ ]0,\rho[,\ \exists \sigma > 0\ /\ \forall\ t_1 \in [t_0,+\infty[,$

$$\begin{cases} \|\varepsilon_1\| < \sigma \\ \left(\sup_{t \in [t_1,+\infty[} \|\varepsilon^*(t)\|\right) < \sigma \end{cases} \Rightarrow \begin{cases} \forall\ t \in [t_1,+\infty[\\ \|z(t) - z_\varepsilon(t)\| < \xi \end{cases}$$

(S) vérifie cette définition. Nous pouvons choisir ρ inférieur à $\alpha/2$.

c) nous permet d'écrire, pour $\theta_1(\theta) \in\ ]0,\alpha/2[$ donné :

d) $\exists \sigma(\theta) > 0\ /\ \forall\ t_1 \in [t_0,+\infty[,$

$$\begin{cases} \|\varepsilon_1\| < \sigma(\theta) \\ \sup_{t \in [t_1,+\infty[} \|\varepsilon(t)\| < \sigma(\theta) \end{cases} \Rightarrow \begin{cases} \forall\ t \in [t_1,+\infty[\\ \|z(t) - z_\varepsilon(t)\| < \theta_1(\theta) \end{cases}$$

(5) nous amène à

(6) $\exists \eta(\sigma(\theta)) > 0\ /\ \forall\ t_1 \in [t_0,+\infty[,$

$$\begin{cases} \|\varepsilon_1\| < \eta(\sigma(\theta)) \\ \left(\sup_{t \in [t_1,+\infty[} \|\varepsilon^*(t)\|\right) < \eta(\sigma(\theta)) \end{cases} \Rightarrow \begin{cases} \|\varepsilon_1\| < \sigma(\theta) \\ \left(\sup_{t \in [t_1,+\infty[} \|\varepsilon(t)\|\right) < \sigma(\theta) \end{cases}$$

d) et 6) combinés à (1) nous donnent :

$$\forall\ \theta \in\ ]0, \hat{\beta}/2[,\ \exists \eta(\sigma(\theta)) > 0\ /\ \forall\ t_1 \in [t_0, +\infty[,$$

$$\begin{cases} \|\varepsilon_1^*\| < \eta(\sigma(\theta)) \\ \\ \left(\sup_{t \in [t_1, +\infty[} \|\varepsilon^*(t)\| \right) < \eta(\sigma(\theta)) \end{cases} \Rightarrow \begin{cases} \forall\ t \in [t_1, +\infty[\\ \\ \|y_{\varepsilon^*}(t) - y(t)\| < \theta \end{cases}$$

et donc (S*) est fortement stable.

Comme T et T* ont les mêmes propriétés, les différents types de stabilité se transportent donc également de (S*) à (S) □

démonstration de la proposition 6

Nous pouvons remarquer qu'une transformation linéaire "x = A•y" ou A est une matrice inversible à coefficients constants est fortement régulière.

Soit $D(\sigma) =$

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \sigma^{-1} & & \\ & & \sigma^{-2} & \\ & & & \sigma^{-(p-2)} \\ 0 & & & 0 \end{bmatrix} \quad \begin{matrix} 0 \\ \\ \\ 0 \\ \sigma^{-(p-1)} \end{matrix}$$

Si nous posons $Q(t) = (q_{ij}(t))_{\substack{i=1,\dots p \\ j=1,\dots p}}$ la transformation "$x = D(\sigma)\cdot y$" transforme (S_3^*) en $(S_3(\sigma))$:

$$\left\{\begin{array}{l} x_1(t) = q_{11}(t)\cdot x_1(t) + \sigma\cdot q_{12}(t)\cdot x_2(t) + \dots + \sigma^{p-1}\cdot q_{1p}(t)\cdot x_p(t) \\ x_2(t) = q_{22}(t)\cdot x_2(t) + \sigma\cdot q_{23}(t)\cdot x_3(t) + \dots + \sigma^{p-2}\cdot q_{2p}(t)\cdot x_p(t) \\ \qquad\qquad \dots\dots\dots\dots\dots\dots \\ x_p(t) = q_{pp}(t)\cdot x_p(t) \\ t \in [t_0, +\infty[\\ x(t_0) = \mu(\sigma) \\ (\text{si nous posons } \mu(\sigma) = D(\sigma)\cdot\mu^* \ ) \end{array}\right.$$

Les coefficients $q_{ij}(t)$ sont bornés pour $t \in [t_0, +\infty[$.

Soit $M(T) = \max_{i,j=1\dots p} \left(\sup_{t \in [T,+\infty[} \|q_{ij}(t)\| \right) + 1$

En prenant $\sigma < \inf(1, \xi/M(T))$, nous pouvons rendre les coefficients non-diagonaux de $Q^{(\sigma)}(t) = \left(\frac{1}{\sigma^{j-i}} q_{ij}(t)\right)_{\substack{i=1,\dots p \\ j=1,\dots p}}$ inférieurs en valeur absolue à ξ, et ce pour tout $t \in [T, +\infty[$, puisque $q_{ij}(t)$ est nul pour $i < j$.

Soit $Q^{\xi}(t)$ la matrice ainsi transformée.

Posons $Q^{\xi}(t) = \left(q^{\xi}_{ij}(t)\right)_{\substack{i=1,\dots p \\ j=1,\dots p}}$ et (S_3^{ξ}) le système associé.

Considérons le produit scalaire $(x,y) = \sum_{i=1}^{p} x_i y_i$

Nous avons la relation suivante :

$$(Q^{\xi}(t)\cdot(x_1 - x_2),(x_1 - x_2)) = \sum_{i=1}^{p} q_{ii}^{\xi}(t)\cdot(x_{1i} - x_{2i})^2 + \sum_{i>j} q_{ij}^{\xi}(t)\cdot(x_{1i} - x_{2i})\cdot(x_{1j} - x_{2j})$$

Or

$$\| \sum_{i>j} q_{ij}^{\xi}(t)\cdot(x_{1i} - x_{2i})\cdot(x_{1j} - x_{2j})\| \leq \xi\cdot(\sum_{i>j} \|x_{1i} - x_{2i}\|\cdot\|x_{1j} - x_{2j}\|$$

ou encore

$$\| \sum_{i>j} q_{ij}^{\xi}(t)\cdot(x_{1i} - x_{2i})\cdot(x_{1j} - x_{2j})\| \leq \xi\cdot(\sum_{i=1}^{p} \|x_{1i} - x_{2i}\|)^2$$

Toutes les normes étant équivalentes, nous pouvons écrire :

$$\exists \gamma > 0 \ / \ \forall\ z \in \mathbb{R}^p \quad \gamma\cdot(\sum_{i=1}^{p} (z_i)^2) \leq (\sum_{i=1}^{p} \|z_i\|)^2$$

donc :

$$\| \sum_{i>j} q_{ij}^{\xi}(t)\cdot(x_{1i} - x_{2i})\cdot(x_{1j} - x_{2j})\| \leq \gamma\cdot\xi\cdot(\sum_{i=1}^{p} (x_{1i} - x_{2i})^2)$$

ce qui amène, puisque $q_{ii}(t) = q_{ii}^{\xi}(t)$:

$\forall\ t \in [T,+\infty[$, $\forall\ x_1 \in \mathbb{R}^p$, $\forall\ x_2 \in \mathbb{R}^p$,

$$(Q^{\xi}(t)\cdot(x_1 - x_2),(x_1 - x_2)) \leq (\ell+\gamma\cdot\xi)\cdot(\sum_{i=1}^{p} (x_{1i} - x_{2i})^2)$$

si nous posons $\ell = \max_{i=1,\ldots p} \left(\sup_{t \in [T,+\infty[} (q_{ii}(t)) \right)$

Il suffit donc de prendre ξ inférieur à $(-\ell)/2\gamma$ pour pouvoir écrire, si nous notons $Q^{**}(t) = Q^{(-\ell)/2\gamma}(t)$ et $(S^{**}) = (S^{(-\ell)/2\gamma})$:

$\forall\ t \in [T,+\infty[$, $\forall\ x_1 \in \mathbb{R}^p$, $\forall\ x_2 \in \mathbb{R}^p$,

$$(Q^{**}(t)\cdot(x_1 - x_2),x_1 - x_2) \leq (\ell/2)\cdot\|x_1 - x_2\|_2^2$$

Comme ℓ est strictement négatif, nous pouvons écrire :

(1) $\forall\ \xi > 0,\ \exists \alpha(\xi) > 0\ /\ \forall\ t \in [T, +\infty[,$

$$\begin{cases} \|\varepsilon(t)\| < \alpha(\xi) \\ t \in [T, +\infty[\\ \|\varepsilon_1\| < \alpha(\xi) \end{cases} \Rightarrow \begin{cases} \|x(t) - x_\varepsilon(t)\| < \xi \\ \\ t \in [t_1, +\infty[\end{cases}$$

De plus, puisque Q(t) est continue et bornée, il existe $k > 0$ tel que :

$\forall\ t \in [t_0, T],\ \forall\ x_1 \in \mathbb{R}^p, \forall\ x_2 \in \mathbb{R}^p,$

$\|Q^{**}(t) \cdot (x_1 - x_2)\| < k \cdot \|x_1 - x_2\|$

ce qui permet d'écrire [Crou, 1980, p.6] :

$\forall\ t_1 \in [t_0, +\infty[,\ \forall\ t \in [t_1, +\infty[,$

$$\|x(t) - x_\varepsilon(t)\| < e^{k\cdot(t-t_1)} \cdot \|\varepsilon_1\| + \int_{t_1}^{t_0} e^{k\cdot(t-s)} \cdot \|\varepsilon(s)\| \cdot ds$$

ou encore :

(2) $\exists \sigma(\xi) > 0\ /\ \forall\ t_1 \in [t_1, +\infty[,$

$$\begin{cases} \|\varepsilon_1\| < \sigma(\xi) \\ \|\varepsilon(t)\| < \sigma(\xi) \\ t \in [t_1, T] \end{cases} \Rightarrow \begin{cases} \|x(t) - x_\varepsilon(t)\| < \xi \\ \\ t \in [t_1, T] \end{cases}$$

(1) et (2) nous permettent d'affirmer la forte stabilité de (S_3^{**}) et donc celle de (S_3) □

démonstration de la proposition 7

Soit (S_4^{**}) le système réduit associé à (S_4) :

$$(S_4^{**}) \begin{cases} \dot{x}(t) = Q^{**}(t)\cdot x(t) + f^{**}(x(t),t) \\ t \in [t_0, +\infty[\\ x(t_0) = \mu^{**} \end{cases}$$

avec si "$T(z,t) = u^{-1}(t)\cdot z$" est la transformation de réduction, $f^{**}(x,t) = u^{-1}(t)\cdot f(u(t)\cdot x,t)$ et $\mu^{**} = u^{-1}(t)\cdot\mu$

Soit $\|\cdot\|_2$ la norme associée au produit scalaire $(x_1,x_2) = \sum_{i=1}^{p} x_{1i}x_{2i}$

Nous allons dans un premier temps montrer que f** vérifie les mêmes propriétés que f.

u(t) étant bornée et continue, f** est donc continue et lipschitzienne par rapport à la première variable. De plus, d'après le point iv) de (D.8), nous pouvons écrire, en posant $M = \sup_{t \in [t_0,+\infty[} \|u(t)\|$:

(1) $\exists \xi_z > 0 \ / \ \forall\ \xi > 0, \quad \exists \sigma(\xi) > 0 \ / \ \forall\ t \in [T,+\infty[$

$\forall\ z_1 \in B_o(z(t),\xi_z)$, $\forall\ z_2 \in B_o(z(t),\xi_z)$,

$\|z_1 - z_2\| < \eta(\xi) \Rightarrow (\|f(z_1,t) - f(z_2,t)\| < (\xi/M^2)\|z_1 - z_2\|)$

Posons $\xi_x = \xi_z/M$, $\eta(\xi) = \sigma(\xi)/M$, et $x(t) = u^{-1}(t)\cdot z(t)$.

Nous avons :

$\forall\ x_1 \in B_o(x(t),\xi_x)$, $\forall\ x_2 \in B_o(x(t),\xi_x)$,

$\|x_1 - x_2\| < \eta(\xi) \Rightarrow \|u(t)\cdot x_1 - u(t)\cdot x_2\| < \sigma(\xi)$

Donc, d'après (1), nous pouvons écrire :

$$\exists \xi_x > 0 \ / \ \forall \ \xi > 0, \ \exists \eta(\xi) > 0 \ / \ \forall \ t \in [T,+\infty[,$$

$$\forall \ x_1 \in B_o(x(t),\xi_x), \ \forall \ x_2 \in B_o(x(t),\xi_x),$$

$$(\ \|x_1 - x_2\| < \eta(\xi) \ ) \Rightarrow \ (\ \|f(u(t)\cdot x_1,t) - f(u(t)\cdot x_2,t)\|$$

$$< \ (\varepsilon/M^2)\cdot\|u(t)\cdot x_1 - u(t)\cdot x_2\| \ )$$

ou encore, puisque $\|u(t)\cdot x\| \leq M\cdot\|x\|$ et que :

$$\|f^{**}(x_1,t) - f(^{**}(x_2,t)\| \leq M\cdot\|f(u(t)\cdot x_1,t) - f(u(t)\cdot x_2,t)\|$$

nous avons :

$$\exists \xi_x > 0 \ / \ \forall \ \xi > 0, \ \exists \eta(\xi) > 0 \ / \ \forall \ t \in [T,+\infty[,$$

$$\forall \ x_1 \in B_o(x(t),\xi_x), \ \forall \ x_2 \in B_o(x(t),\xi_x),$$

$$\|x_1 - x_2\| < \eta(\xi) \Rightarrow \|f^{**}(x_1,t) - f^{**}(x_2,t)\| \leq \xi\cdot\|x_1 - x_2\|$$

et donc f** vérifie iv).

Puisque toutes les normes sont équivalentes, nous pouvons écrire également :

$$\exists \xi_x > 0 \ / \ \forall \ \xi > 0, \ \exists \eta(\xi) > 0 \ / \ \forall \ t \in [T,+\infty[\ ,$$

$$\forall \ x_1 \in B_o^2(x(t),\xi_x), \ \forall \ x_2 \in B_o^2(x(t),\xi_x)$$

$$\|x_1 - x_2\| < 2\eta(\xi) \Rightarrow \|f^{**}(x_1,t) - f^{**}(x_2,t)\| < \xi\cdot\|x_1 - x_2\|$$

en notant B_o^2 les boules ouvertes par rapport à $\|\cdot\|_2$.

Nous supposons que la réduction effectuée est telle que, si $\ell = \max\limits_{i=1,\ldots p} \left(\sup\limits_{t \in [T,+\infty[} (q_{ii}(t)) \right)$, nous ayons :

$$\forall \ x \in \mathbb{R}^p \ (Q^{**}(t)\cdot x, x) \leq (\ell/2)\cdot\|x\|_2^2 \ .$$

Si nous prenons $\xi = (-\ell)/4$, nous avons :

$$\forall\, t \in [T,+\infty[\ ,\ \forall\, x_1 \in B_o^2(x(t),\eta(\xi))\ ,$$

$$\forall\, x_2 \in B_o^2(x(t),\eta(\xi))$$

$$\left((Q^{**}(t)\cdot x_1 + f^{**}(x_1,t)) - (Q^{**}(t)\cdot x_2 + f^{**}(x_2,t)), x_1 - x_2 \right) \leq \ell/4 \cdot \|x_1 - x_2\|_2^2$$

Nous pouvons donc écrire :

$$\forall\, \xi > 0,\ \exists \alpha(\xi) > 0 \ /\ \forall\, t \in [T,+\infty[\ ,$$

$$\begin{cases} \|\varepsilon(t)\| < \alpha(\xi) \\ t \in [T,+\infty[\\ \|\varepsilon_1\| < \alpha(\xi) \end{cases} \Rightarrow \begin{cases} \|x_\varepsilon(t) - x(t)\| < \xi \\ t \in [T,+\infty[\end{cases}$$

De plus, puisque f^{**} est lipschitzienne par rapport à la première variable, continue et que $Q^{**}(t)$ est continue et bornée, nous avons :

$$\exists k > 0 \ /\ \forall\, t \in [t_0,+\infty[\ ,\ \forall\, x_1 \in \mathbb{R}^p,\ \forall\, x_2 \in \mathbb{R}^p,$$

$$\|Q^{**}(t)\cdot(x_1 - x_2) + f^{**}(x_1,t) - f^{**}(x_2,t)\| \leq k \cdot \|x_1 - x_2\|$$

ce qui permet, comme dans la démonstration de la proposition 6, d'écrire :

$$\exists \sigma(\xi) > 0 \ /\ \forall\, t_1 \in [t_0,T],$$

$$\begin{cases} \|\varepsilon(t)\| < \sigma(\xi) \\ t \in [t_1,T] \\ \|\varepsilon_1\| < \sigma(\xi) \end{cases} \Rightarrow \begin{cases} \|x(t) - x_\varepsilon(t)\| < \xi \\ t \in [t_1,T] \end{cases}$$

et de conclure à la forte stabilité de (S_4) □

démonstration de la proposition 8

Les hypothèses prises nous permettent d'appliquer le

théorème des fonctions implicites en tout point de Ω_T[L-F-Arn, 1977, tome 2, p.240]

Soit J l'ensemble des intervalles I incluant t_0 et contenus dans $[t_0,T]$, sur lequel nous pouvons construire une application y continue et une application z de classe C^1 telles que (S_6) soit vérifié sur I.

Posons $t_\mu = \sup_{I \subset J} (\sup_{t \in I} t)$

Montrons que pour tout I, y et z sont uniques.

Soient (y*,z*) et (y**,z**) deux couples de solutions de (S_6) sur I. Supposons ces deux couples distincts.

Soient $H_z = \{ t \in I / z^*(t) = z^{**}(t) \}$

$H_y = \{ t \in I / y^*(t) = y^{**}(t) \}$

$H_{\dot{z}} = \{ t \in I / \dot{z}^*(t) = \dot{z}^{**}(t) \}$

Puisque ce sont des applications continues, H_z et H_y sont des fermés relatifs de I. H_z, H_y, et $H_{\dot{z}}$ sont non-vides car ils contiennent t_0 et le système $f(c,\mu,d,t_0)$ n'admet qu'une solution.

$$\exists t_{1z} > t_0,\ \exists t_{2z} > t_0 \ /$$
$$\begin{cases} \forall\, t \in [t_0,t_{1z}],\ z^*(t) = z^{**}(t) \\ \forall\, t \in]t_{1z},t_{2z}[,\ z^*(t) \neq z^{**}(t) \end{cases}$$

$$\exists t_{1y} > t_0,\ \exists t_{2y} > t_0 \ /$$
$$\begin{cases} \forall\, t \in [t_0,t_{1y}],\ y^*(t) = y^{**}(t) \\ \forall\, t \in]t_{1y},t_{2y}[,\ y^*(t) \neq y^{**}(t) \end{cases}$$

De plus,de manière claire, $\dot{z}^*$ est égal à $\dot{z}^{**}$ sur $[t_0,t_{1z}]$.

Montrons que $t_{1z} = t_\mu$

Pour cela supposons par l'absurde que $t_{1z} < t_\mu$.

Appliquons le théorème des fonctions implicites au point $(y^*(t_{1z}), z^*(t_{1z}), \dot{z}^{**}(t_{1z}), t_{1z})$ qui est aussi le point $(y^{**}(t_{1z}), z^{**}(t_{1z}), \dot{z}^{**}(t_{1z}), t_{1z})$.

$$\exists \varepsilon_t > 0, \quad \exists \varepsilon_y > 0, \quad \exists \varepsilon_z > 0, \quad \exists \varepsilon_{\dot{z}} > 0 \ /$$
$$\forall t \in]t_{1z} - \varepsilon_t, t_{1z} + \varepsilon_t[, \ \forall z \in B_o(z^*(t_{1z}), \varepsilon_z),$$
$$\exists! c \in B_o(y^*(t_{1y}), \varepsilon_y), \ \exists! d \in B_o(\dot{z}^*(t_{1z}), \varepsilon_{\dot{z}}) \ /$$
$$f(c,z,d,t) = \Theta_{q+p}$$

De plus ϕ définie par

$$\begin{cases} \phi_y(z,t) = c \\ \phi_{\dot{z}}(z,t) = d \end{cases}$$

est continue, différentiable, sa différentielle valant

$$\begin{cases} \frac{\partial \phi}{\partial t}(z,t) = - \frac{\partial f}{\partial y_{\dot{z}}}(c,z,d,t)^{-1} \cdot \frac{\partial f}{\partial t}(c,z,d,t) \\ \frac{\partial \phi}{\partial z}(z,t) = - \frac{\partial f}{\partial y_{\dot{z}}}(c,z,d,t)^{-1} \cdot \frac{\partial f}{\partial z}(c,z,d,t) \end{cases}$$

D'après les hypothèses, nous avons $\| \frac{\partial \phi}{\partial z}(z,t) \| < \beta_z(T)$

d'où, puisqu'une boule ouverte est convexe :

$$\forall z_1 \in B_o(z(t_{1z}), \varepsilon_z), \ \forall z_2 \in B_o(z(t_{1z}), \varepsilon_z),$$
$$\forall t \in]t_{1z} - \varepsilon_t, t_{1z} + \varepsilon_t[$$
$$\| \phi(z_1,t) - \phi(z_2,t) \| < \beta_z(T) \cdot \| z_1 - z_2 \|$$

$\phi_{\dot{z}}$ est donc lipschitzienne en z.

Puisque z* et z** sont continues, alors :

$\exists \alpha > 0$ (que nous prendrons inférieur à ε_t et à $t_{2z} - t_{1z}$)/

$$\forall\, t \in [t_1, t_1 + \alpha[, \quad \begin{cases} z^*(t) \in B_o(z^*(t_{1z}), \varepsilon_z) \\ z^{**}(t) \in B_o(z^*(t_{1z}), \varepsilon_z) \end{cases}$$

$$\text{donc le problème} \quad \begin{cases} \dot{z} = \phi_{\dot{z}}(z,t) \\ t \in [t_{1z}, t_{1z} + \alpha[\\ z(t_{1z}) = z^*(t_{1z}) \end{cases}$$

défini sur $B_o(z^*(t_1), \varepsilon_z)$ admet deux solutions distinctes, ce qui est contradictoire avec le théorème de Cauchy-Lipschitz [L-F-Arn, 1977, tome 4, p.99]

D'où le fait que $t_{1z} = t_\mu$, et donc que le couple (y,z) est unique.

De plus, nous pouvons déduire de la démonstration précédente que, si I_1 est inclus dans I_2, alors le couple (y_1, z_1) est solution de (S_6) sur I_2. Donc si nous posons $I^* = \bigcup_{ICJ} I$, il existe un couple unique (y,z) solution de (S_6) sur I*.

Il existe localement deux applications ϕ_y et $\phi_{\dot{z}}$ de classe C_k (données par le théorème des fonctions implicites) telles que

$$\begin{cases} \dot{z}(t) = \phi_{\dot{z}}(z(t),t) \\ y(t) = \phi_y(z(t),t) \end{cases}$$

Un simple raisonnement par récurrence sur la classe de z montre que z est de classe C^{k+1} et donc que y est de classe C^k, et ce sur tout I*, puisque t a été choisi quelconque.

Soit l'application g de I* vers $\mathbb{R}^{q+p}$ qui à t associe $f(y(t),z(t),\dot{z}(t),t)$.

$$\forall\ t \in I^*,\ g(t) = \theta_{q+p}$$

donc $\forall\ t \in I^*, \dot{g}(t) = \theta_{q+p}$ ou encore

$$\frac{\partial f}{\partial y \dot{z}}(y(t),z(t),\dot{z}(t),t)\cdot\begin{pmatrix}\dot{y}(t)\\ \dot{z}(t)\end{pmatrix} + \frac{\partial f}{\partial z}(y(t),z(t),\dot{z}(t),t)\cdot\dot{z}(t)$$

$$+ \frac{\partial f}{\partial t}(y(t),z(t),\dot{z}(t),t) = \theta_{q+p}$$

donc

$$\forall\ t \in I \quad \begin{cases} \|\dot{y}(t)\| \leq \beta_z(T)\cdot\|\dot{z}(t)\| + \beta_t(T) \\ \|\dot{z}^{\cdot}(t)\| \leq \beta_z(T)\cdot\|\dot{z}(t)\| + \beta_t(T) \end{cases}$$

Puisque z est de classe C^2, nous avons :

$$\int_{t_0}^{t} \dot{z}(u).du = \dot{z}(t) - \dot{z}(t_0)$$

ou encore :

$$\int_{t_0}^{t} \|\ddot{z}(u)\|.du \geq \|\dot{z}(t)\| - \|\dot{z}(t_0)\|$$
$$\Rightarrow \|\dot{z}(t)\| \leq \|\dot{z}(t_0)\| + \int_{t_0}^{t} \|\ddot{z}(u)\|.du$$

ce qui entraîne, vu les inégalités précédentes :

$$\|\dot{z}(t)\| \leq \|\dot{z}(t_0)\| + \beta_z(T).\int_{t_0}^{t} \|\dot{z}(u)\|.du + (t - t_0).\beta_z(T)$$

ou encore :

$$\|\dot{z}(t)\| \leq \|\dot{z}(t_0)\| + \beta_z(T).\int_{t_0}^{t} \|\dot{z}(u)\|.du + (T - t_0).\beta_z(T)$$

Posons $\xi(t) = \int_{t_0}^{t} \|\dot{z}(u)\|.du$.

Alors :

$$\dot{\xi}(t) \leq \dot{\xi}(t_0) + \beta_z(T).\xi(t) + (T - t_0).\beta_z(T)$$

Soit $M_\xi^* = \dot{\xi}(t_0) + (T - t_0).\beta_z(T)$

Nous avons, puisque $\beta_z(T)$, M_ξ^* et ξ sont strictement positifs :

$$\frac{\beta_z(T).\dot{\xi}(t)}{\beta_z(T).\xi(t) + M_\xi^*} \leq \beta_z(T)$$

ou encore, en intégrant terme à terme :

$$\text{Log}((\beta_z(T).\xi(t) + M_\xi^*)/M_\xi^*) \leq \beta_z(T).(t - t_0)$$
$$\Rightarrow \beta_z(T).\xi(t) + M_\xi^* \leq M_\xi^*.e^{\beta_z(T).(t - t_0)}$$
$$\Rightarrow \xi(t) \leq (M_\xi^*/\beta_z(T)).(e^{\beta_z(T).(t - t_0)} - 1)$$

Donc il existe $M_\xi > 0$ tel que :
$\forall\, t \in [t_0, t_\mu[,\ \xi(t) \leq M_\xi$
$\Rightarrow \int_{t_0}^{t} \|\dot{z}(u)\|.du \leq M_\xi$
$\Rightarrow \|\dot{z}(t)\| \leq \|\dot{z}(t_0)\| + \beta_z(T).M_\xi + (T - t_0).\beta_z(T)$

D'où le fait qu'il existe M_y et $M_{\dot{z}}$ réels strictement positifs tels que :

$$\forall\, t \in [t_0, t_\mu[, \quad \begin{cases} \|\dot{z}(t)\| \leq M_z \\ \|\dot{y}(t)\| \leq M_y \end{cases}$$

Le théorème des accroissements finis nous indique alors

$$\forall\, t \in [t_0, t_\mu[\quad \begin{cases} \|z(t) - z(t_0)\| \leq M_z (t - t_0) \\ \|y(t) - y(t_0)\| \leq M_y (t - t_0) \end{cases}$$

Donc z et y sont bornées sur $[t_0, t_\mu[$.

$z, y, \dot{z}$ sont alors uniformément continues sur $[t_0, t_\mu[$.

Il existe donc des prolongements continus

$\tilde{z}$, $\tilde{y}$, $\tilde{\dot{z}}$ sur $[t_0, t_\mu]$.

$\tilde{z}$, $\tilde{y}$, $\tilde{\dot{z}}$ vérifient $\begin{cases} \tilde{y}(t) = y(t) \\ \tilde{z}(t) = z(t) \\ \tilde{\dot{z}}(t) = \dot{z}(t) \end{cases} \forall\, t \in [t_0, t_\mu[$

et $\begin{cases} \tilde{y}(t_\mu) = \lim\limits_{t \to t_\mu} y(t) \\ \tilde{z}(t_\mu) = \lim\limits_{t \to t_\mu} z(t) \\ \tilde{\dot{z}}(t_\mu) = \lim\limits_{t \to t_\mu} \dot{z}(t) \end{cases}$

[L-F-Arn, 1977, tome 2, p 86]

Nous savons de plus [L-F-Arn, 1977, tome 2, p 135] que

$\tilde{\dot{z}}(t_\mu) = \dot{\tilde{z}}(t_\mu)$

Puisque f est continue, et que

$\lim_{t \to t_\mu} f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p}$, alors

$f(\tilde{y}(t_\mu),\tilde{z}(t_\mu),\dot{\tilde{z}}(t_\mu),t_\mu) = \theta_{q+p}$

Donc $(\tilde{y},\tilde{z})$ est solution de (S*).

D'où $I^* = [t_0,t_\mu]$

Si t_μ était inférieur à T, alors, grâce au théorème des fonctions implicites appliqué au point $(y(t_\mu),z(t_\mu),\dot{z}(t_\mu),t_\mu)$, nous pourrions prolonger y et z au delà de t_μ, ce qui est contradictoire avec l'hypothèse de maximalité de I*. Donc $I^* = [t_0,T]$.

(H_2) nous permet de construire y_τ et z_τ sur $[t_0,\tau]$, quelque soit la valeur de $\tau > t_0$, telles que
$\forall\, t \in [t_0,t_\mu]$, $f(y_\tau(t),z_\tau(t),\dot{z}_\tau(t),t) = \theta_{q+p}$.

En reprenant la démonstration de l'unicité de y et z sur I*, nous pouvons démontrer que y_τ et z_τ sont uniques, et sont le prolongement de y_{τ^*}et z_{τ^*} , pour $\tau^* \in [t_0,\tau]$

Donc nous pouvons construire y et z solution de

$$\begin{cases} f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p} \\ t \in [t_0,+\infty[\\ z(t_0)=\mu \end{cases}$$ □

démonstration de la proposition 9

Nous allons reprendre la démonstration du théorème des fonctions implicites proposée par G. Chilov [Chilov, 1974, p 64] et essayer de passer de résultats locaux à des résultats globaux grâce aux hypothèses d'uniformité qui ont été faites :

$$\text{Posons } \delta = (\alpha_z,\alpha_t) \in \mathbb{R}_+^{*2},\ \rho = (\alpha_y,\alpha_{\dot z}) \in \mathbb{R}_+^{*2},$$
$$U_\delta^t = B_o(z(t),\alpha_z) \times]t-\alpha_t,t+\alpha_t[\cap [t_0,+\infty[$$
$$V_\rho^t = B_o(y(t),\alpha_y) \times B_o(\dot z(t),\alpha_{\dot z})$$

Notons $V_\rho^t(U_\delta^t)$ l'espace vectoriel normé de toutes les fonctions bornées et continues allant de U_δ^t dans V_ρ^t, muni de la norme $\|g\| = \sup_{(z,t)\in U_\delta^t} \|g(z,\tau)\|$.

La fonction implicite $\phi^t = (\phi_y^t,\phi_{\dot z}^t)$ vérifie :

$$\frac{\partial f}{\partial y\dot z}(y(t),z(t),\dot z(t),t)^{-1}\cdot\Big[\frac{\partial f}{\partial y\dot z}(y(t),z(t),\dot z(t),t)\cdot$$
$$\phi^t(z,t) - f(\phi_y^t(z,\tau),z,\phi_{\dot z}^t(z,\tau),\tau)\Big] = \phi^t(z,\tau)$$

Posons

$$F^t(g) = \frac{\partial f}{\partial y\dot z}(y(t),z(t),\dot z(t),t)^{-1}\cdot\Big[\frac{\partial f}{\partial y\dot z}(y(t),z(t),\dot z(t),t)\cdot g$$
$$- f(g_y,I_z,g_{\dot z},I_\tau)\Big]$$

Nous appelons I_z (resp. I_t) la projection de U_δ^t sur $B_o(z(t),\alpha_z)$ (resp. $]t-\alpha_t,t+\alpha_t[\cap[t_0,+\infty[$

f est bornée et uniformément continue sur W.

$\frac{\partial f}{\partial \dot{yz}}$ est bornée et uniformément continue sur W.

F^t est donc dérivable dans $V^t_{\rho_1}$ ($U^t_{\delta_1}$)

si nous posons $\delta_1 = (\xi_z, \xi_t)$ et $\rho_1 = (\xi_y, \xi_{\dot{z}})$.

alors

$$F^{t\prime}(g) = -\frac{\partial f}{\partial \dot{yz}}(y(t),z(t),\dot{z}(t),t) \cdot$$

$$\left[\frac{\partial f}{\partial \dot{yz}}(y(t),z(t),\dot{z}(t),t) - \frac{\partial f}{\partial \dot{yz}}(g_y, I_z, g_{\dot{z}}, I_t)\right]$$

$$\|F^{t\prime}(g)\| \leq \gamma_1 \sup_{\substack{(z,t) \in U^t_{\delta_1} \\ (c,d) \in V^t_{\rho_1}}} \left[\left\| \frac{\partial f}{\partial \dot{yz}}(y(t),z(t),\dot{z}(t),t) - \frac{\partial f}{\partial \dot{yz}}(c,z,d,t)\right\|\right]$$

$\frac{\partial f}{\partial \dot{yz}}$ est uniformément continue sur W.

$\exists \nu_z > 0$, $\exists \nu_y > 0$, $\exists \nu_{\dot{z}} > 0$, $\exists \nu_t$ 0 (que nous prendrons inférieurs à $\xi_z, \xi_y, \xi_{\dot{z}}$, et ξ_t)/

$$\forall\, t \in [t_0, +\infty[\ ,\ \forall\, z \in B_o(z(t),\nu_z)\ ,\ \forall\, \tau \in]t-\nu_t, t+\nu_t[\cap [t_0, +\infty[\ ,\ \exists!\, c \in B_o(y(t),\nu_y)\ ,\ \exists!\, d \in B_o(\dot{z}(t),\xi_{\dot{z}})/$$
$$\left\| \frac{\partial f}{\partial \dot{yz}}(y(t),z(t),\dot{z}(t),t) - \frac{\partial f}{\partial \dot{yz}}(c,z,d,\tau)\right\| \leq 1/2\gamma_1$$

f est uniformément continue sur W.

$\exists \eta_z > 0$, $\exists \eta_y > 0$, $\exists \eta_{\dot{z}} > 0$, $\exists \eta_t > 0$ (que nous prendrons inférieurs à $\nu_z, \nu_y, \nu_{\dot{z}}$, et ν_t)/

$$\forall\, t \in [t_0, +\infty[\ ,\ \forall\, z \in B_o(z(t),\eta_z)\ ,\ \forall\, \tau \in]t-\eta_t, t+\eta_t[\cap [t_0, +\infty[\ ,\ \forall\, c \in B_o(y(t),\eta_y)\ ,\ \forall\, d \in B_o(\dot{z}(t),\eta_{\dot{z}})\ ,$$
$$\|f(c,z,d,t)\| \leq 1/2\gamma_1$$

Donc, si nous posons $\delta_1 = (\eta_z, \eta_t)$, $\rho_1 = (\eta_y, \eta_{\dot{z}})$, nous pouvons écrire :

$$\forall\, t \in [t_0, +\infty[\ ,\ \forall\, g \in V^t_{\rho_1}(U^t_{\delta_1}),\ \|F^{t\,\prime}(g)\| < 1/2$$

Posons $\tilde{b}^t(z,\tau) = (\tilde{b}^t(z,\tau), \tilde{b}^t(z,\tau))$, avec

$$\begin{cases} \forall\, z \in B_o(z(t), \eta_z) \\ \forall\, \tau \in\,]t - \eta_t, t + \eta_t[\,\cap\, [t_0, +\infty[\end{cases} \qquad \begin{cases} \tilde{b}^t_y(z,\tau) \equiv y(t) \\ \tilde{b}^t_{\dot{z}}(z,\tau) \equiv \dot{z}(t) \end{cases}$$

Alors :

$$\forall\, t \in [t_0, +\infty[\ ,\ \|F^t(\tilde{b}^t) - \tilde{b}^t\| < 1/2$$

Les conditions sont alors réunies pour utiliser le théorème du point fixe [Chilov, 1974, p 54]

Nous pouvons écrire :

$$\exists \theta_1 \in \mathbb{R}^{\sharp 2},\ \exists \delta_1 \in \mathbb{R}^{\sharp 2} /$$
$$\forall\, t \in [t_0, +\infty[\ ,\ \exists!\ \phi^t \in V^t_{\rho_1}(U^t_{\delta_1}) /\ F^t(\phi^t) = \phi^t$$

ou encore :

$$\exists \eta_t > 0,\ \exists \eta_y > 0,\ \exists \eta_{\dot{z}} > 0\ /$$
$$\forall\, t \in [t_0, +\infty[\ ,\ \forall\, z \in B_o(z(t), \eta_z)\ ,\ \forall\, \tau \in$$
$$]t - \eta_t, t + \nu_t[\,\cap\, [t_0, +\infty[\ ,\ ,\ \exists!\ c \in B_o(y(t), \eta_y)\ ,$$
$$\exists! d \in B_o(\dot{z}(t), \eta_{\dot{z}}) /\ \ f(c,z,d,t) = \theta_{q+p}$$

démonstration de la proposition 10

Considérons le système perturbé

$$\begin{cases} f(y_\varepsilon(t), z_\varepsilon(t), \dot{z}_\varepsilon(t) + \varepsilon(t), t) = \theta_{q+p} \\ t \in [t_1, +\infty[\\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \end{cases}$$

que nous pouvons remplacer par

$$\begin{cases} y_\varepsilon(t) = \phi_y(z_\varepsilon(t),t) \\ \dot{z}_\varepsilon(t) = \phi_{\dot{z}}(z_\varepsilon(t),t) - \varepsilon(t) \\ z_\varepsilon(t_1) = z(t_1) + \varepsilon_1 \\ t \in [t_0,+\infty[\end{cases}$$

Nous avons :

$$(\phi_{\dot{z}}(z_1,t) - \phi_{\dot{z}}(z_2,t), z_1 - z_2) \leq \ell \, \|z_1 - z_2\|_2^2$$

avec $\ell < 0$

D'après (R_2) nous pouvons écrire :

$$\|z(t) - z_\varepsilon(t)\|_2 \leq e^{\ell(t - t_1)} \cdot \|\varepsilon_1\| + \int_{t_1}^{t} e^{\ell(t - s)} \cdot \|\varepsilon(s)\|_2 ds$$

Nous avons $y(t) - y_\varepsilon(t) = \phi_y(z(t),t) - \phi_y(z_\varepsilon(t),t)$

Nous pouvons appliquer le théorème des accroissements finis car le segment $[z(t),z_\varepsilon(t)]$ est contenu dans W.

(H_2) nous permet d'écrire :

$$\|y(t) - y_\varepsilon(t)\| \leq \beta_z \cdot \|z(t) - z_\varepsilon(t)\|$$

$\|.\|$ et $\|.\|_2$ sont équivalentes :

$$\exists a > 0, \quad \exists b > 0 \ / \ \forall \ j \in \mathbb{R}^q,$$
$$a\|j\|_2 \leq \|j\| \leq b\|j\|_2$$

Donc

$$\|y(t) - y_\varepsilon(t)\| \leq (b.\beta_z/a) \cdot \|z(t) - z_\varepsilon(t)\|_2$$

Soit $c = (b.\beta_z/a)$

Posons $\alpha = \min(\eta_{\dot{z}}, (-\ell.\xi)/c, \xi/c, -\ell.\xi, \xi)$

Nous pouvons écrire :

$$\forall\ \xi \in\ ]0,\eta_z[,\ \exists \alpha > 0\ /$$

$$\sup_{t \in [t_0,+\infty[} \|\varepsilon(t)\|_2 < \alpha \Rightarrow \begin{cases} \|z(t) - z_\varepsilon(t)\|_2 < \xi \\ \\ \|y(t) - y_\varepsilon(t)\|_2 < \xi \end{cases}$$

démonstration de la proposition 11

Nous allons d'abord considérer la suite $(g_n^k)_{n\geq 0}$ déterminée par :

$$\begin{cases} \|g_i^k = \|z_i - z'_i\|_2 \\ i=1,\ldots k-1 \\ g_n^k = -\frac{1}{(\mu h - \alpha_0)} \left(\sum_{i=1}^{k} |\alpha_i| \cdot g_{n-i}^k \right) \end{cases}$$

Soit $P_{\mu h}^k(x) = x^k + \frac{1}{(\mu h - \alpha_0)} \left(\sum_{i=1}^{k} |\alpha_i| \ x^{k-1} \right)$.

Posons $\{\xi_i^k\}_{i=1,\ldots k}$ les racines de $(P_{\mu h}^k)$.

La théorie des suites récurrentes [Hac, 1971, p 193] nous apprend qu'il existe un ensemble de valeurs $\{\gamma_i^k\}_{i=1,\ldots k}$ telles que :

$$(E)\ \forall\ n > 0 \quad g_n^k = \sum_{i=1}^{k} \gamma_i^k (\xi_i^k)^n$$

De plus, il est clair que, si $(\mu h - \alpha_0)$ est négatif, g_n^k est positif. Les racines de $P_{\mu h}$ évoluent continuement en fonction de μh.

$(P^k_{-\infty})$ admet pour unique racine la racine d'ordre k 0. En résolvant pour $\theta \in [0,2\pi]$, l'équation $\mu h = \alpha_0 - \sum_{i=1}^{k} \alpha_i e^{-i\theta}$ (qui provient de la résolution de $P^k_{\mu h}(e^{i\theta}) = 0$), nous trouverons l'ensemble des valeurs μh telles que $P^k_{\mu h}$ admette des racines de module 1. Nous savons déjà que $P^k_{\mu h}$ admet 1 pour racine si $\mu h = \alpha_0 - \sum_{i=1}^{k} |\alpha_i|$

Les coefficients α_i sont les suivants [Gear, 1967]

k	α_0	α_1	α_2	α_3	α_4	α_5	α_6	$\alpha_0 - \sum_{i=1}^{k} \alpha_i$
1	1	-1						0
2	3/2	-2	1/2					-1
3	11/6	-3	3/2	-1/3				- 3
4	25/12	-4	3	-4/3	1/4			- 13/2
5	137/60	-5	5	-10/3	5/4	-1/5		-25/2
6	147/60	-6	15/2	-20/3	15/4	-6/5	1/6	-691/30

Pour k=1 nous avons :

$$\begin{cases} \mu h = 1 - e^{-i\theta} \\ \mu h \in \mathbb{R} \end{cases}$$

$$\begin{cases} \mu h = (1 - \cos(\theta)) + i\sin(\theta) \\ \mu h \in \mathbb{R} \end{cases}$$

$$\begin{cases} \mu h = 0 \\ \text{ou} \\ \mu h = 2 \end{cases}$$

Donc, d' après la remarque précédente, portant sur la continuité des racines de $(P^k_{\mu h})$ et sur le fait

que $(P^k_{-\infty})$ admette pour unique racine 0, nous pouvons affirmer que :

$$\mu h < 0 \Rightarrow \xi_1^1 < 1$$

Pour k allant de 2 à k, en regardant les graphes des pages suivantes, nous pouvons constater qu'il n'y a pas d'autre valeurs réelles de μh après celle de

$$\mu h = \alpha_0 - \sum_{i=1}^{k} |\alpha_i| .$$

Donc de la même manière :

$$\mu h < \alpha_0 - \sum_{i=1}^{k} |\alpha_i| \Rightarrow \xi_i^k < 1.$$

Nous pouvons remarquer que $\mu h - \alpha_0$ est négatif dans ce cas-là.

Posons $\lambda_k = \sum_{i=1}^{k} \gamma_i^k$, $r_k = \max_{i=1,\ldots k} \xi_i^k$.

Nous avons, d'après ce qui précède, pour μh inférieur à $\alpha_0 - \sum_{i=1}^{k} |\alpha_i|$, r_k strictement inférieur à 1.

Nous avons, d'après (E) :

$\forall\, n \in \mathbb{N}\ g_n^k \leq \lambda_k r_k^n$

Soit $\mathscr{J}(n)$: " $\|z_n - z'_n\| \leq g_n^k$ "

$\mathscr{J}(0), \mathscr{J}(1), \ldots \mathscr{J}(k-1)$ sont vérifiées.

Supposons $\mathscr{J}(p), \mathscr{J}(p-1), \ldots \mathscr{J}(p-k+1)$ vraies. Qu'en est-il de $\mathscr{J}(p+1)$?

Nous avons :

$$\alpha_0(z_{p+1} - z'_{p+1}) + \sum_{i=1}^{k} \alpha_i (z_{p-i+1} - z_{p-i+1})$$
$$= h.(f(z_{p+1}, t_{p+1}) - f(z'_{p+1}, t_{p+1}))$$

$$= \sum_{i=1}^{k} \alpha_i (z_{p+1-i} - z'_{p+1-i}, z_{p+1} - z'_{p+1}) \leq (\mu h - \alpha_0) \|z_{p+1} - z'_{p+1}\|_2^2$$

$$= \|z_{p+1} - z'_{p+1}\|_2 \leq - \sum_{i=1}^{k} \frac{\alpha_i}{(\mu h - \alpha_0)} \|z_{p+1-i} - z'_{p+1-i}\|_2$$

= (hyp. de récurrence)

$$\|z_{p+1} - z'_{p+1}\| \leq - \sum_{i=1}^{k} \frac{\alpha_i}{(\mu h - \alpha_0)} g^k_{p+1-i}$$

$$= \|z_{p+1} - z'_{p+1}\|_2 \leq g^k_{p+1}$$

par définition de $(g^k_n)_{n=0,\ldots}$.

$\mathcal{J}(p+1)$ est donc vraie.

Donc

$$\forall\, n \in \mathbb{N} \quad \|z_n - z'_n\| \leq g^k_n$$

Or $\forall\, n \in \mathbb{N}$, $g^k_n \leq \lambda_k . r^n_k$

D'où le fait que $\forall\, n \in \mathbb{N}$ $\|z_n - z'_n\| \leq \lambda_k . r^n_k$

Le 1) est démontré.

Le 2) se démontre quant à lui beaucoup plus simplement :

Soit $\mathcal{H}(n)$ " $\|z_n - z'_n\|_2 \leq \delta$ "

Par hypothèse, $\mathcal{H}(0), \mathcal{H}(1), \ldots \mathcal{H}(k-1)$ sont vérifiées.

Supposons $\mathcal{H}(p), \mathcal{H}(p-1), \ldots \mathcal{H}(p-k+1)$ vraies.

Qu'en est-il de $\mathcal{H}(p+1)$?

Nous pouvons vérifier très simplement que

$$(I) \Rightarrow \|z_{p+1} - z'_{p+1}\|_2 < \delta\left(1 - \sum_{i=1}^{k} \frac{|\alpha_i|}{(\mu h - \alpha_0)}\right)$$

$$\Rightarrow \|z_{p+1} - z'_{p+1}\|_2 < \delta \text{ puisque } \mu h < \alpha_0 - \sum_{i=1}^{k} |\alpha_i|$$

$\mathcal{H}(p+1)$ est donc vérifiée.

La propriété est donc démontrée pour tout $n \in \mathbb{N}$ □

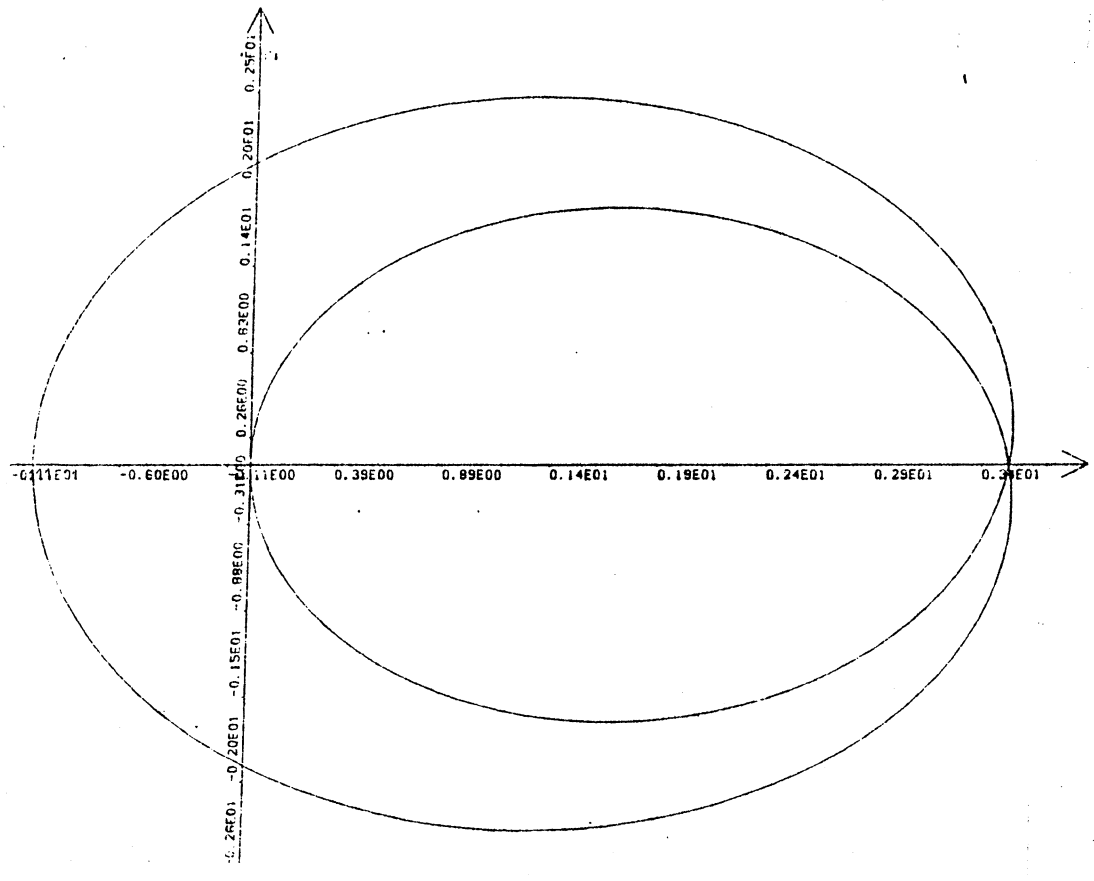
Les graphes qui suivent représentent les graphes de

$$\mu h = \alpha_0 - \sum_{i=1}^{k} \alpha_i e^{-i\theta}, \quad \theta \in [0, 2\pi[$$

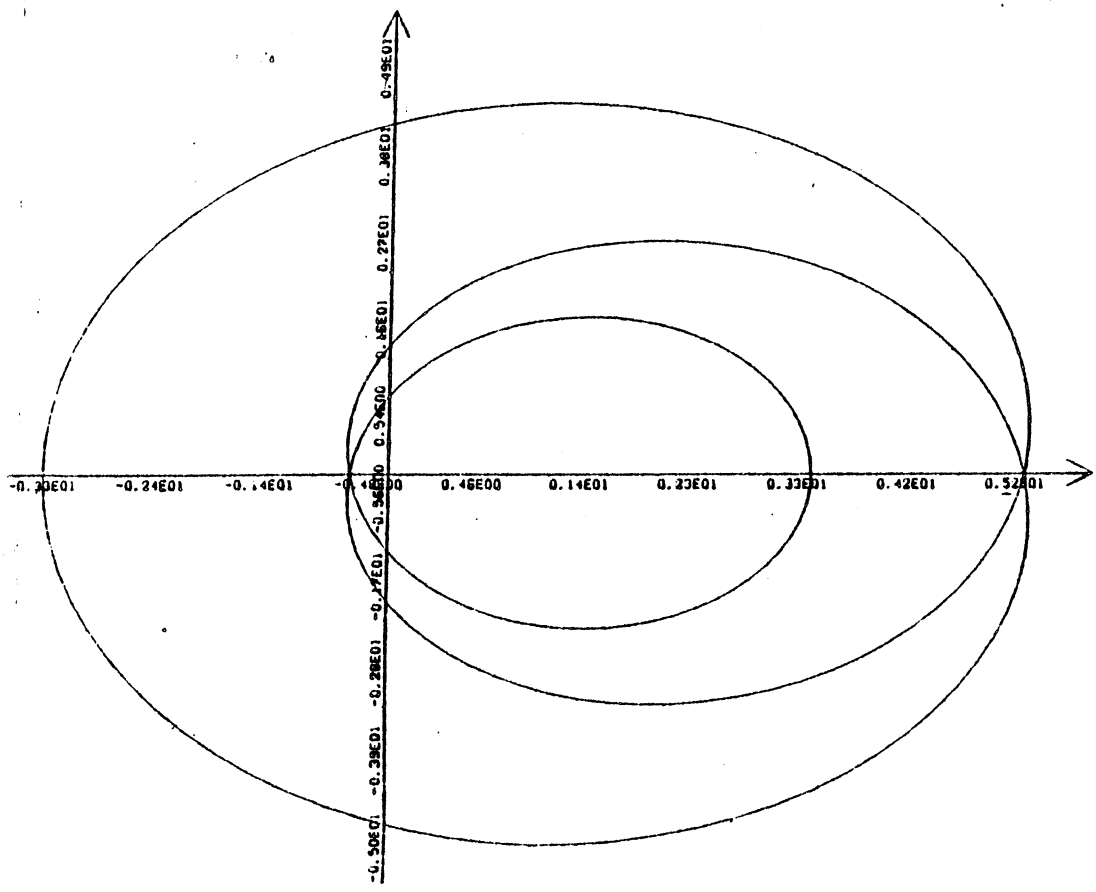
$k = 2, \ldots, 6$

Ils forment l'ensemble des valeurs de μh telles que $P^k_{\mu h}$ admette des racines de module un . Les courbes qui suivent coupent respectivement pour k allant de 2 à 6 l'axe des réels négatifs en -1, -3, -6.5, -12.5, -23.03

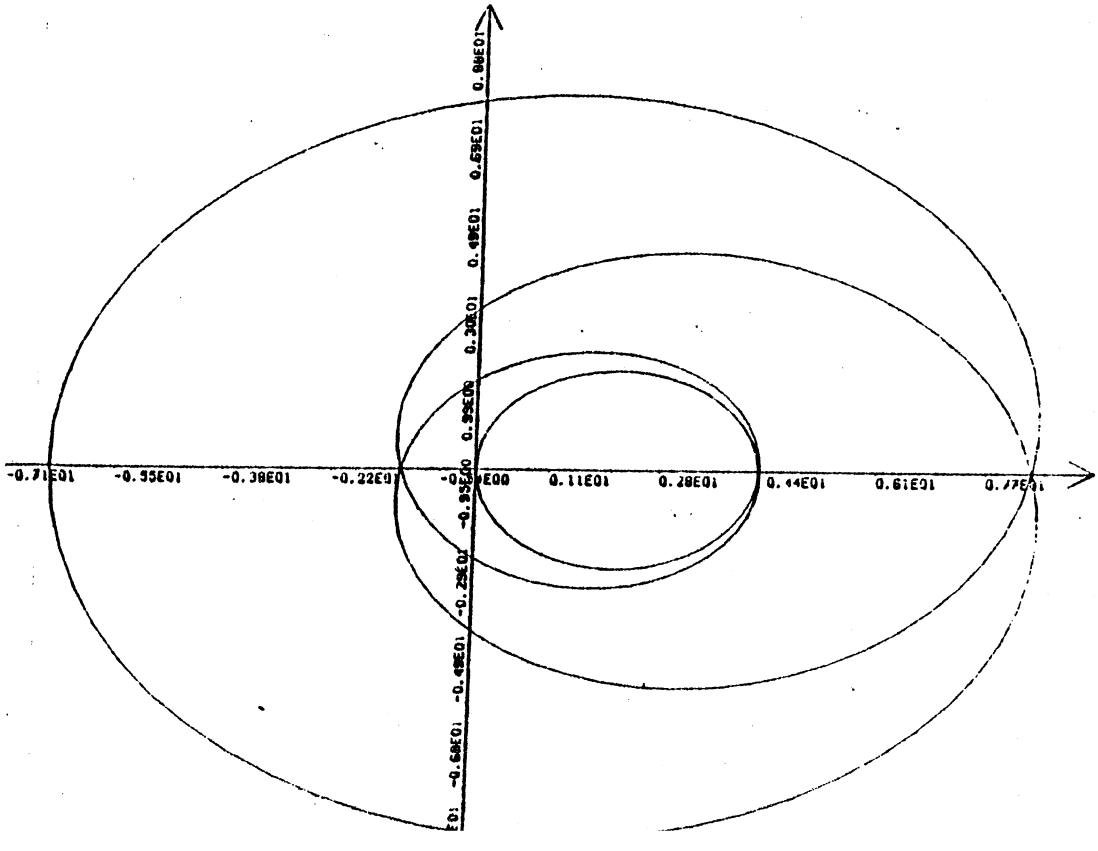
lieux des μh pour k = 2



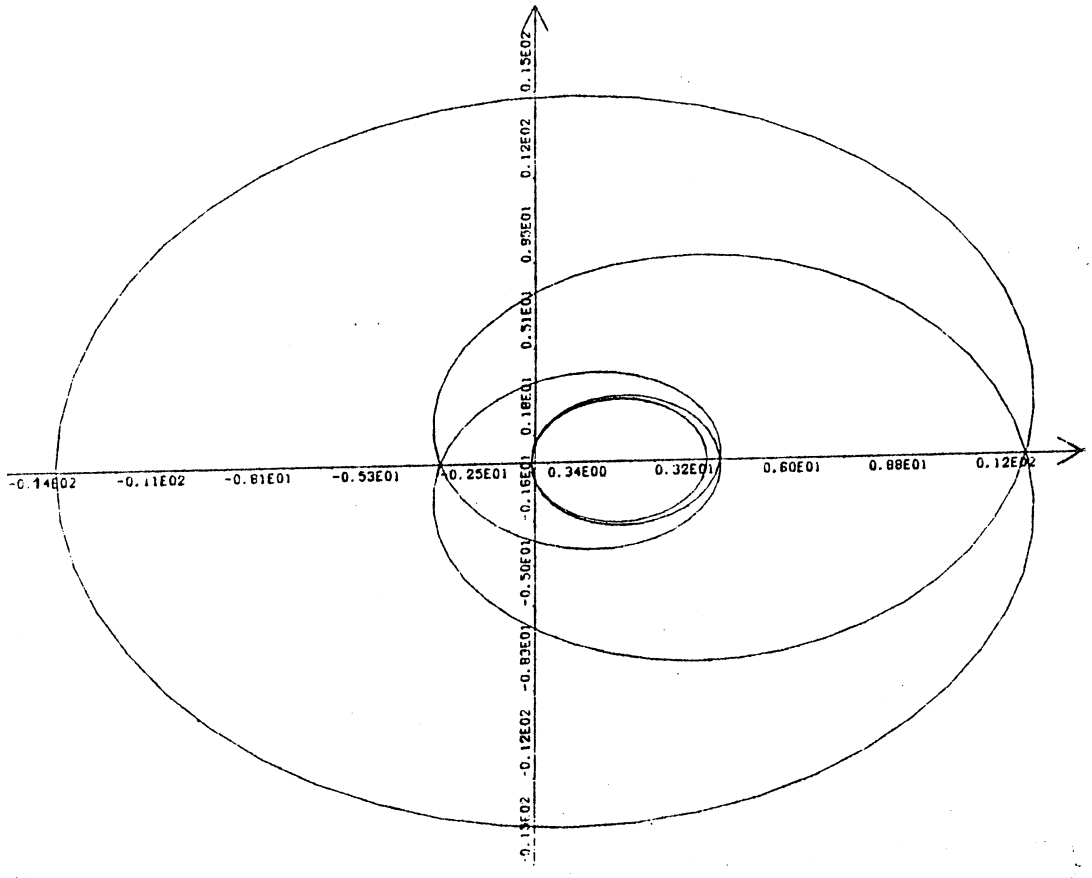
lieux des μh pour k = 3



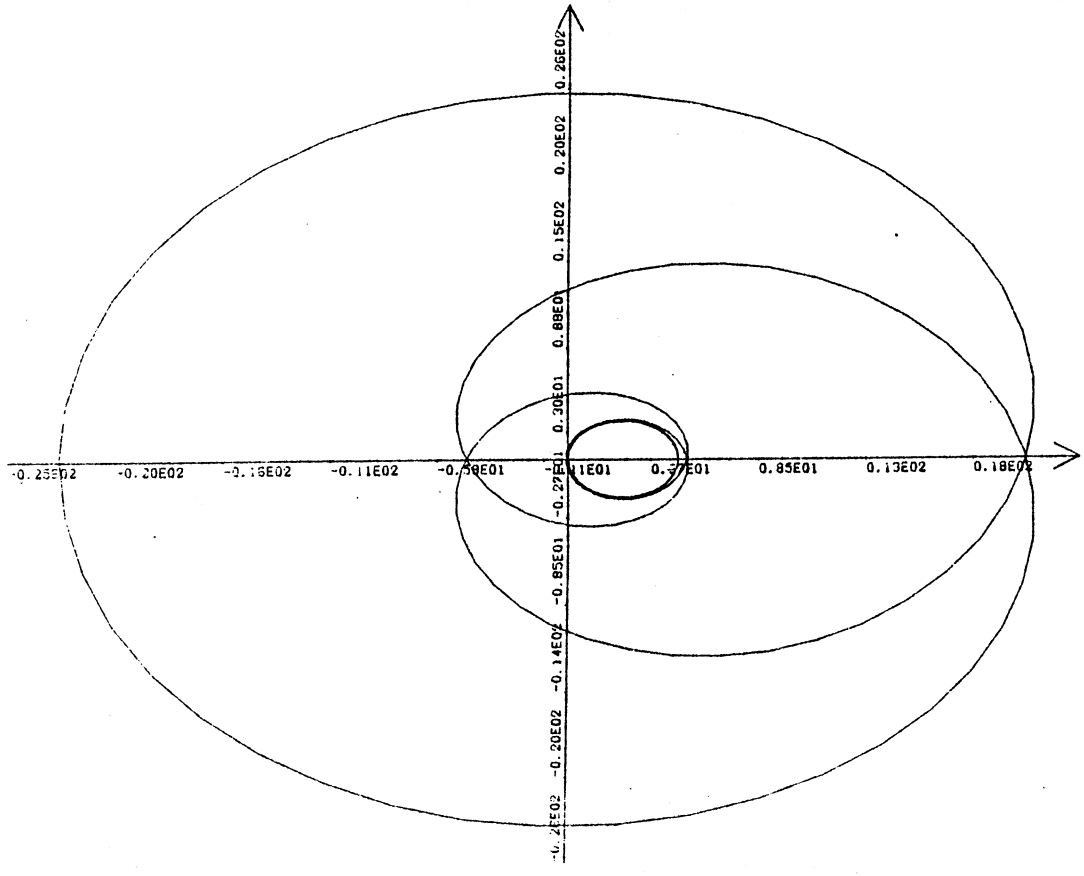
lieux des μh pour k = 4



lieux des μh pour k = 5



lieux des μh pour k = 6



démonstration de la proposition 12

Nous avons pour tout n :

$$\alpha_0 z_n + \sum_{i=1}^{k} \alpha_i z_{n-i} = \mu h . f(z_n, t_n)$$

$$\alpha_0 z(t_n) + \sum_{i=1}^{k} \alpha_i z(t_{n-i}) = \mu h . f(z(t_n, t_n) + e_n$$

Donc :

$$- \sum_{i=1}^{k} |\alpha_i| \, \| z_{n-i} - z(t_{n-i}) \| \leq (\mu h - \alpha_0) \| z_n - z(t_n) \| + \| e_n \|$$

=>(d'après nos hypothèses de départ) $\| z_n - z(t_n) \| \leq \delta$ □

CHAPITRE II

INTRODUCTION

Dans ce chapitre, nous présentons quelques algorithmes de recherche opérationnelle(décomposition d'un graphe en composantes faiblement connectées entre elles, recherche d'un couplage maximum,...) qui nous permettent de construire divers algorithmes dérivés des outils numériques les plus récemment proposés.

Ces algorithmes permettent la prise en compte de la latence et l'utilisation de méthodes de surrelaxation pour la résolution de tels systèmes.

I)PRELIMINAIRES

Dans le chapitre I nous avons pu constater la portée des travaux de P.Henrici [Hen, 1962] ,de Dahlquist [Dahl, 1962],de Gear [Gear, 1967] qui permirent la mise au point de schémas numériques quasi-stables(ou "stiffly-stables") adaptés aux caractéristiques de la simulation electrique.Les efforts se sont ensuite portés sur la mise en œuvre de ces schémas.Dans un premier temps Hachtel, Brayton et Gustavson [Hach, 1971] introduisirent les techniques de matrices creuses, tandis que Van Bokhoven s'attacha à la mise au point d'algorithmes efficaces pour le calcul des predicteurs, du pas et de l'ordre(pour plus de précision cf. le chapitre III), et que . Récemment,ce sont les travaux de Rabbat [Rabbat, 1979] et de Sangiovanni-Vincentelli [Sang, 1980] qui introduisirent les concepts de "macromodèles" et de latence. Cette approche retiendra particulièrement notre attention dans le texte qui suit et nous essaierons de dégager des outils numériques permettant la prise en compte de macromodèles et permettant éventuellement la prise en compte des phénomènes de latence.

II) PRESENTATION DU PROBLEME

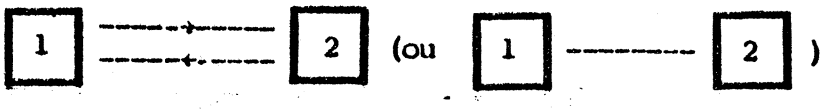
1)présentation sommaire de la méthode de Newton à plusieurs niveaux

A priori,un circuit intégré est décrit de manière modulaire par son concepteur. Les relations existant entre deux modules peuvent être de plusieurs sortes :

_ relation de type directionnel

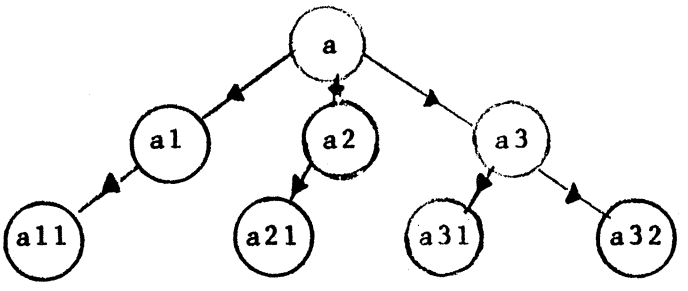
Le module 2 est affecté par le module 1 et non pas l'inverse:il s'agit d'un transfert d'information.

_ relations de type non-directionnel



Les modules 1 et 2 intéragissent ensemble.Il s'agit d'un équilibre fonctionnel.

Un travail préalable consiste à calculer les composantes fortement connexes de l'ensemble des modules ainsi connectés et à les ordonner. Nous obtenons alors un graphe de ce type :



Où chaque sommet du graphe correspond à une composante fortement connexe.

Chacune de ces composantes peut être évaluée indépendemment des autres,du moment que celles en amont sont connues.

Le problème qui demeure est donc celui de l'évaluation du comportement d'une composante connexe.

Cette démarche hiérarchisée est celle adoptée par le programme de C.A.O CASCADE [Bor, 1983], [Le Faou, 1982].

2)Nécessité de nouvelles approches du problème numérique

Chaque module correspond en fait à une portion de circuit électrique. Analyser l'activité d'un circuit au cours du temps conduit à la discrétisation et à la résolution d'un système algébrodifférentiel de la forme :

(si nous posons $\dot{z}(t)= \frac{dz}{dt}(t)$)

$$\begin{cases} f(y(t),z(t),\dot{z}(t),t)=\theta_{q+p} \\ z(t0) = \mu \end{cases}$$

où $\begin{cases} y(t) \in \mathbb{R}^q \\ z(t) \in \mathbb{R}^p \end{cases}$ ont pour composantes des tensions et des intensités.

θ_{q+p} est l'origine de $\mathbb{R}^{q+p}$.

Ces systèmes sont souvent des systèmes de très grande dimension, plusieurs centaines éventuellement, de telle sorte que les méthodes traditionnelles (Newton-Raphson notamment)

deviennent d'un coût prohibitif, et limitent la portée des programmes qui les utilisent.

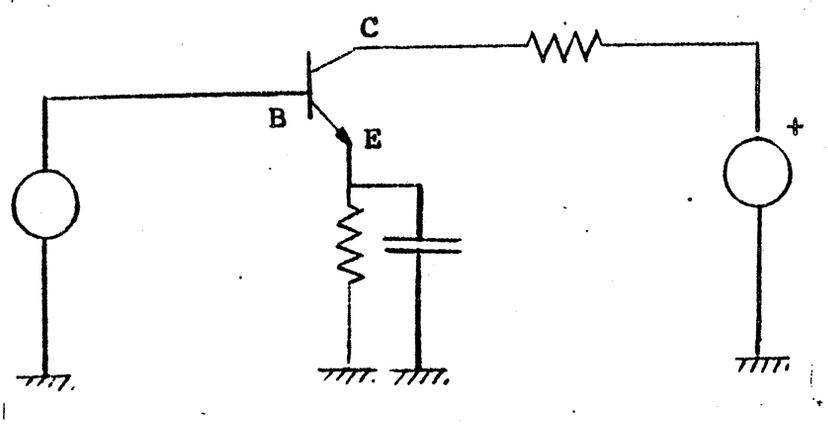
Une première remarque que nous pouvons faire est que les composantes de $y(t)$ et de $z(t)$ sont souvent latentes, c'est-à-dire qu'elles restent plus ou moins constantes sur des intervalles de temps de longueur non négligeable. La latence d'une variable peut occuper jusqu'à 90-95% du temps d'intégration ,lorsque le système est issu de l'étude d'un circuit intégré de grande taille.

Il serait attrayant de pouvoir tenir compte de cela afin de réduire de manière sensible le nombre de calculs.

Le problème est donc de déterminer avec précision quand une variable devient latente et surtout quand elle cesse de l'être.

Nous avons d'abord expérimenté diverses démarches "empiriques" sur un système algébrodifférentiel de quatre équations à quatre inconnues, suffisamment petit pour pouvoir la suivre "à la main", et représentatif des problèmes que nous nous posons, les variables étant latentes 30 % du temps d'intégration.

Schéma du circuit test



Description du système issu de ce circuit

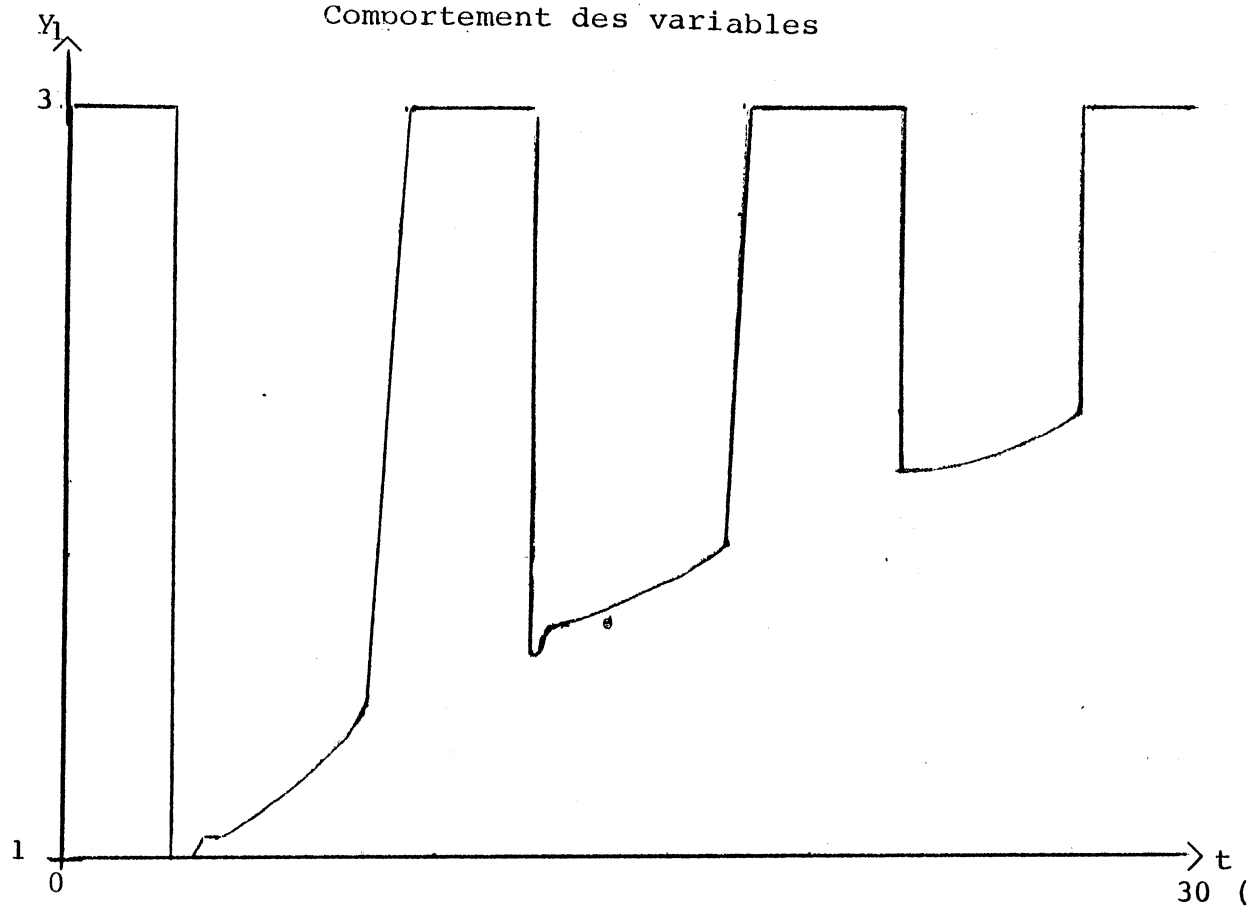
$$a_{11}\cdot y_1+a_{12}\cdot z_1+a_{13}\cdot z_2+a_{14}\cdot z_3 = b_1$$

$$a_{21}\cdot y_1+a_{22}\cdot z_1+a_{23}\cdot z_2+a_{24}\cdot z_3+a_{25}\cdot \dot{z}_1 = b_2(t)$$

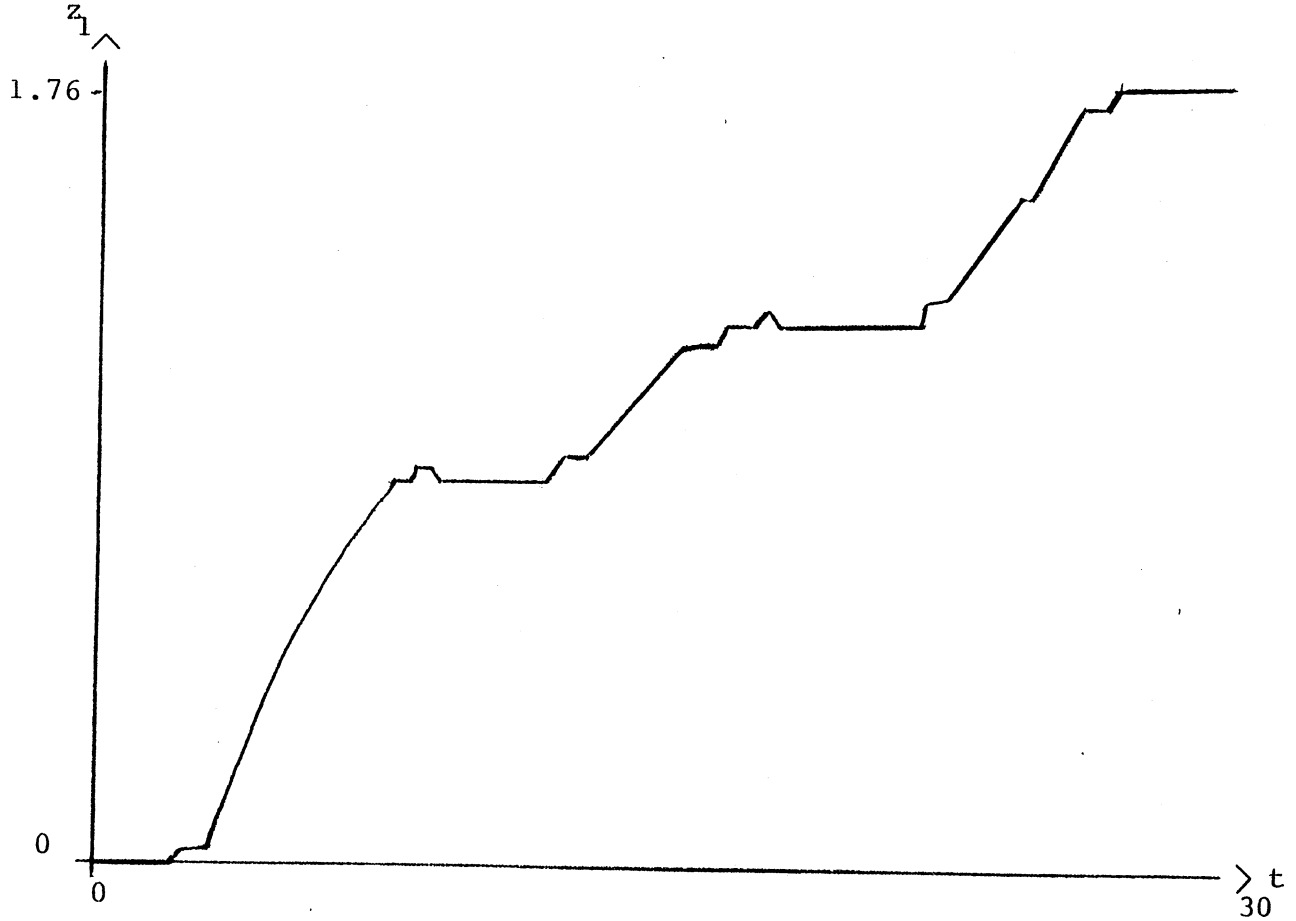
$$a_{31}\cdot y_1+a_{32}\cdot z_1+a_{33}\cdot z_2+a_{34}\cdot z_3+a_{35}(y_2,y_3)-a_{36}\cdot \dot{z}_2 = b_3(t)$$

$$a_{41}\cdot y_1+a_{42}\cdot z_1+a_{43}\cdot z_2+a_{44}\cdot z_3+a_{45}(y_2,y_3)-a_{46}\cdot \dot{z}_3 = b_4(t)$$

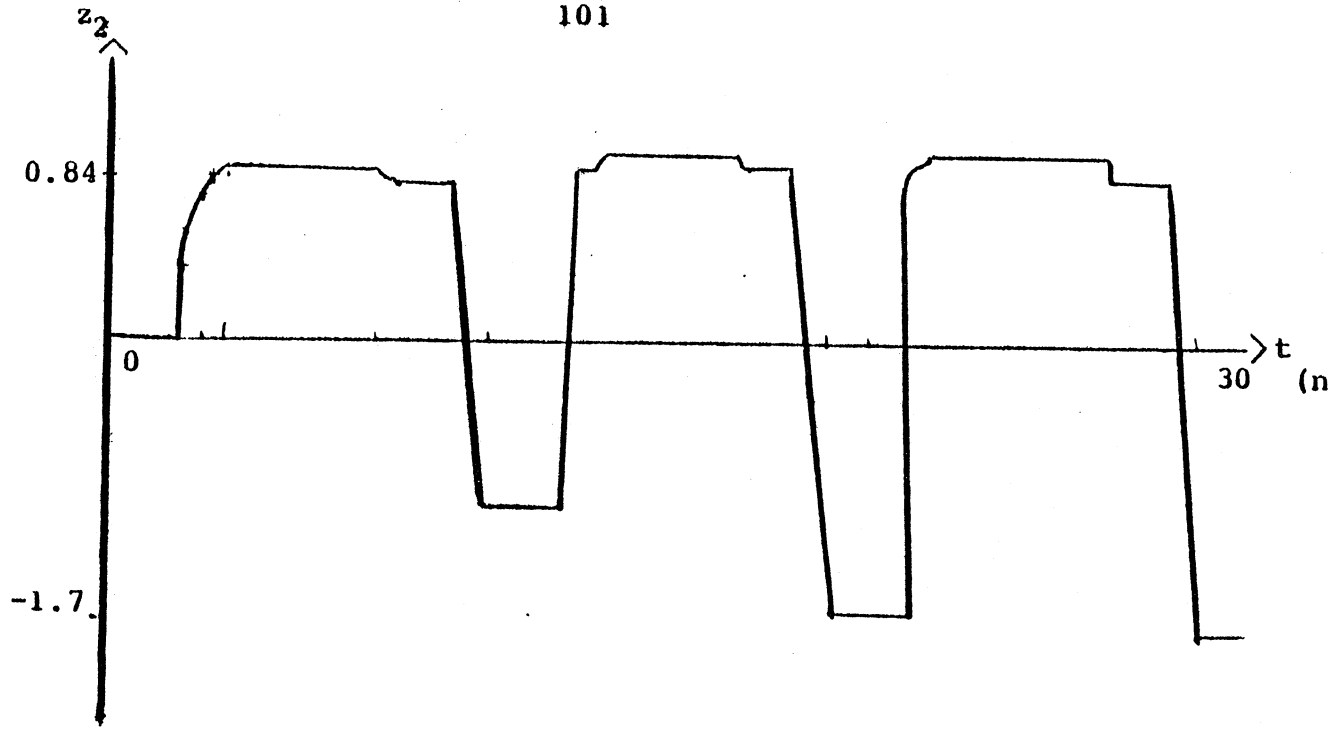
Comportement des variables



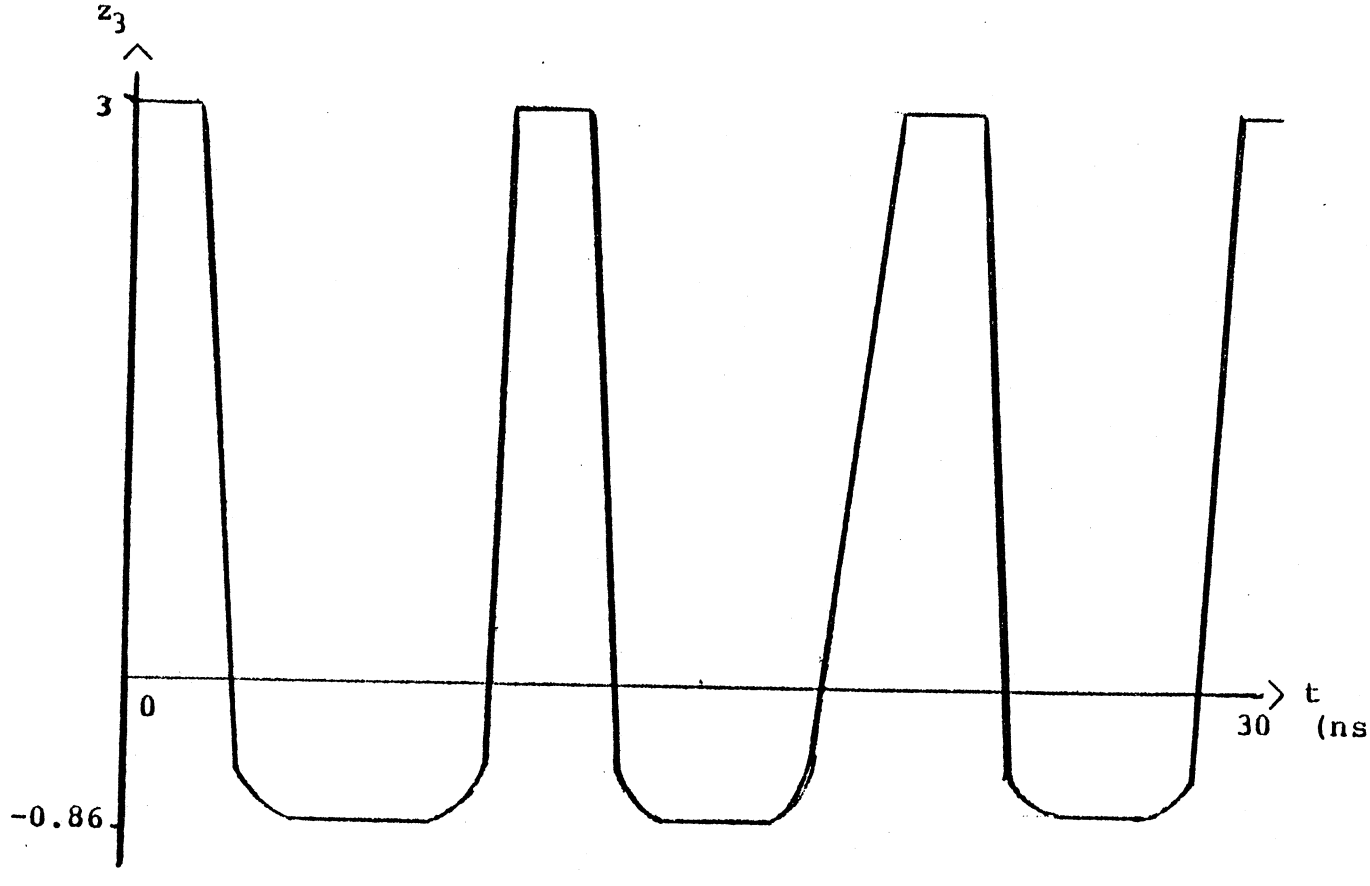
variable n°1



variable n°2



variable n°3



variable n°4

Nous avons essayé de trouver des tests satisfaisants pour la détection des variables "latentes' afin de ne pas avoir à les calculer.

Les questions qui se sont posées ont été de deux types:

*) Trouver des critères de latence satisfaisants

C'est-à-dire trouver k et ε_j tels que, dès que nous trouvons $\frac{\| z^j_{n+1} - z^j_{n-i} \|}{test + \| z^j_{n-k} \|} < \varepsilon_j$, i=0... k nous puissions conclure à la latence de la jième variable (nous posons $z^j_{n+1} = z^j(t_n)$ test est un réel strictement positif donné).

**) Savoir quand la latence prend fin pour commencer à réévaluer zj

Il est nécessaire que ε_j, si nous appelons ainsi le critère de latence attribué à la jième variable, soit inférieur à l'erreur de troncature tolérée, ceci afin de ne pas introduire d'erreurs additionnelles

Ce seul critère se révèle cependant nettement insuffisant pour les variables intervenant dans des expressions non linéaires, car l'impact sur les autres variables d'une erreur sur zj est difficilement maîtrisable notamment dans des expressions où interviennent des exponentielles.

Il serait intéressant de disposer d'un outil de mesure adéquat qui permette de moduler le critère ε_j en fonction de la plus ou moins grande linéarité de l'expression dans laquelle la jième variable intervient.

Détecter la fin de la latence d'une composante n'est pas un problème simple .

En effet contrairement à ce que nous pensions au départ quand une variable est dite de manière erronée "latente", le système ne diverge pas tout de suite, ce qui aurait constitué un moyen simple de constater la fin de la latence, mais dérive vers une solution n'ayant que peu de rap-

ports avec la solution réelle, pour diverger éventuellement bien plus tard. Pour parer à cet inconvénient, une approche adoptée fut de recalculer de temps à autre une composante dite "latente" et de vérifier son état de latence, mais le peu d'efficacité de cette démarche conduit à multiplier les tests, et donc diminue de manière considérable le profit retiré de la latence(dans notre exemple,sur 1100 appels de fonction,seul un gain de 100 appels a pu être réalisé).

3) Conclusion

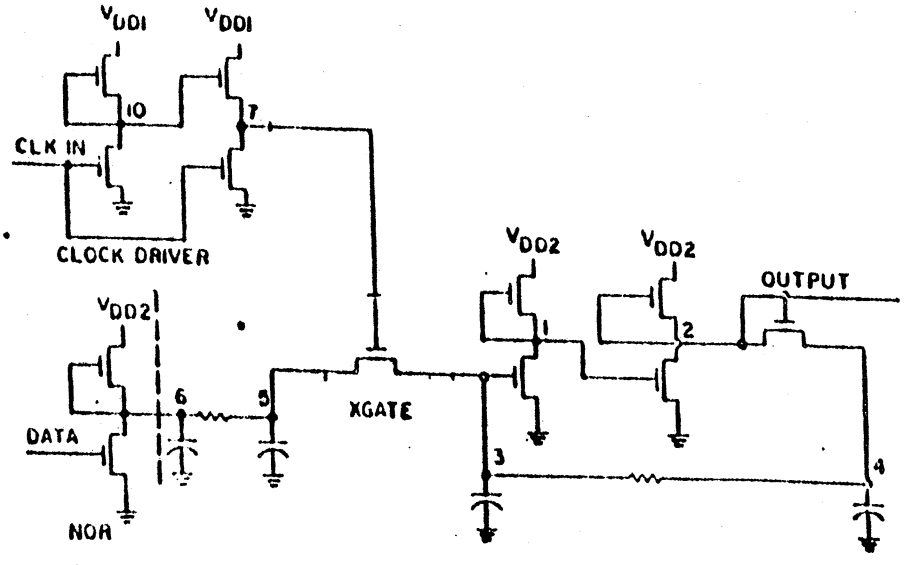
L'échec des méthodes "empiriques" précédemment exposées est celui d'une démarche voulant adapter les méthodes existantes sans les modifier de manière conséquente.Il apparaît donc nécessaire de se doter d'outils spécifiques si nous voulons des méthodes ayant un caractère local.

III) NOUVEAUX OUTILS NUMERIQUES

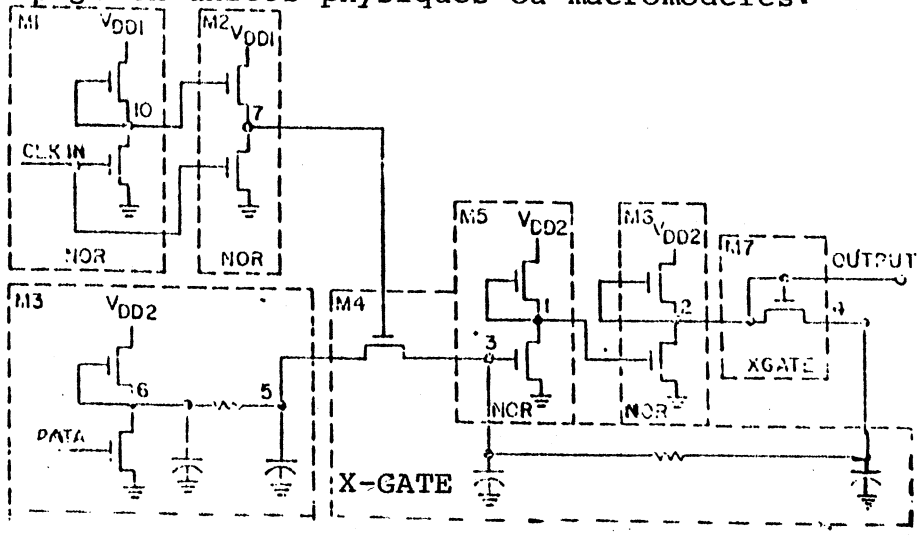
1) Présentation sommaire de la méthode de Newton à plusieurs niveaux

Sangiovanni-Vincentelli [Sang,1979] a construit une méthode de Newton s'appuyant sur la mise en évidence d'unités physiques ('macromodèles') dans un circuit, connectées faiblement entre elles.

Exemple : un circuit à registre



son découpage en unités physiques ou macromodèles:



L'idée fondamentale de Sangiovanni-Vincentelli est que les variables attachées à une unité physique "dépendent" de la valeur des variables attachées aux connexions qui lient les différents modules entre eux Il s'agit donc de construire un outil à deux niveaux · nous faisons varier lentement les variables liées aux connexions liant les modules, et à chaque variation de ces variables, on ajuste les variables dépendants des unités physiques. Le processus s'arrête lorsqu'il y a équilibre fonctionnel L'inconvénient de l'algorithme proposé par Sangiovanni-Vincentelli est qu'il nécessite la connaissance à priori des variables qui servent d'entrée à l'unité physique en question, et de celles qui sont délivrées en sortie ce qui est la plupart du temps ignoré.

D'autre part il n'est pas toujours aisé de discerner la modularité du circuit.

Il nous semble préférable de nous attarder d'abord sur la structure même du système algébrodifférentiel.

2) **Graphe associé à un système algébrodifférentiel, et problème de sa partition:**

La discrétisation du système

$$f(y(t)\ z(t), \dot{z}(t)\ t) = \theta_{q+p}$$

conduit à résoudre le système algébrique $Fn(yn\ zn) = \theta_{q+p}$ au point $t=t_n$·

Soit le graphe $\underline{A} = (A, X)$.

X est un ensemble de q+p sommets numérotés de 1 à

$q+p$ Le sommet i représente à la fois la $i$ème variable et la $i$ème équation si nous numérotons les variables de 1 à $q+p$ en considérant d'abord y_n, puis z_n.

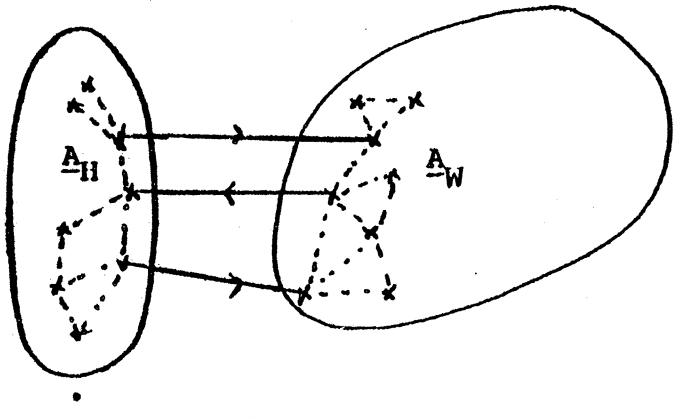
A est un ensemble d'arcs a_{ij} allant de i à j a_{ij} existe si et seulement si la $i^{ième}$ inconnue intervient dans la $j$ième équation.

Nous pouvons remarquer que $\underline{A}$ est similaire au graphe de connexion de F_n puisque, si nous posons

$$\frac{\partial F_n}{\partial yz} = (u_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}} \quad , \qquad u_{ij} \neq 0 \quad \langle = \rangle \quad a_{ji} \text{ existe.}$$

Soit $\underline{A}_H$ un sous graphe de $\underline{A}$ n'étant connecté à son complémentaire que par un nombre restreint d'arcs

$\underline{A}$.



$\underline{A}_H$ correspond à un sous-système carré H de F_n .

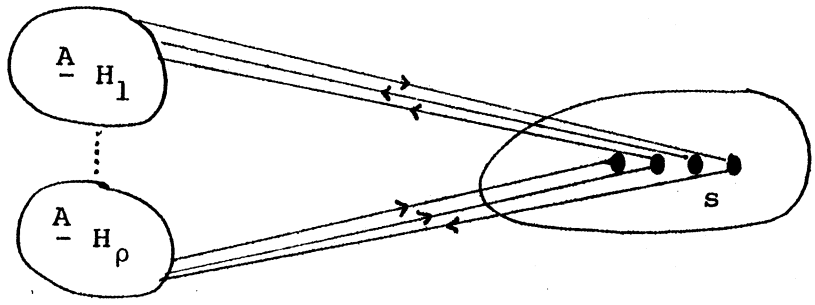
- H est "piloté' par les variables correspondant aux sommets de A_W reliées à A_H par des arcs entrants dans A_H .
  Nous appelons ces variables les "entrées" de H .
- Les arcs sortants de A_H pour aller dans A_W correspondent à des variables de H intervenant dans d autres équations que celles de H Ces variables seront dites "sorties".
- Les autres variables seront appelées "internes".

Notons
- u ($\in \mathbb{R}^{\sigma 1}$) les entrées
- y ($\in \mathbb{R}^{\sigma 2}$) les sorties
- x ($\in \mathbb{R}^{\pi}$) les variables internes.

Le nombre de sommets de A_H est $(\pi+\sigma_2)$. Le sous système H s'écrit alors $H(u,x,y) = \Theta_{\sigma_2+\pi}$, où $\Theta_{\sigma_2+\pi}$ est l'origine de $R^{\sigma 2+\pi}$

H est une application allant de $\mathbb{R}^{\sigma 1} \times \mathbb{R}^{\pi} \times \mathbb{R}^{\sigma 2}$ vers $\mathbb{R}^{\sigma 2+\pi}$.

Dans un premier temps nous nous proposons comme but de déconnecter Fn en ρ sous-systèmes H_i en minimisant le nombre de variables "pilotes" et le nombre d'équations dans lesquelles interviennent les sorties équations que nous nommerons "équations de couplage".



Si nous appelons S l'ensemble des sommets correspondants aux entrées des H_i et aux équations de couplage, il s'agit pour nous de minimiser l'ensemble S qui sera appelé ensemble séparateur.

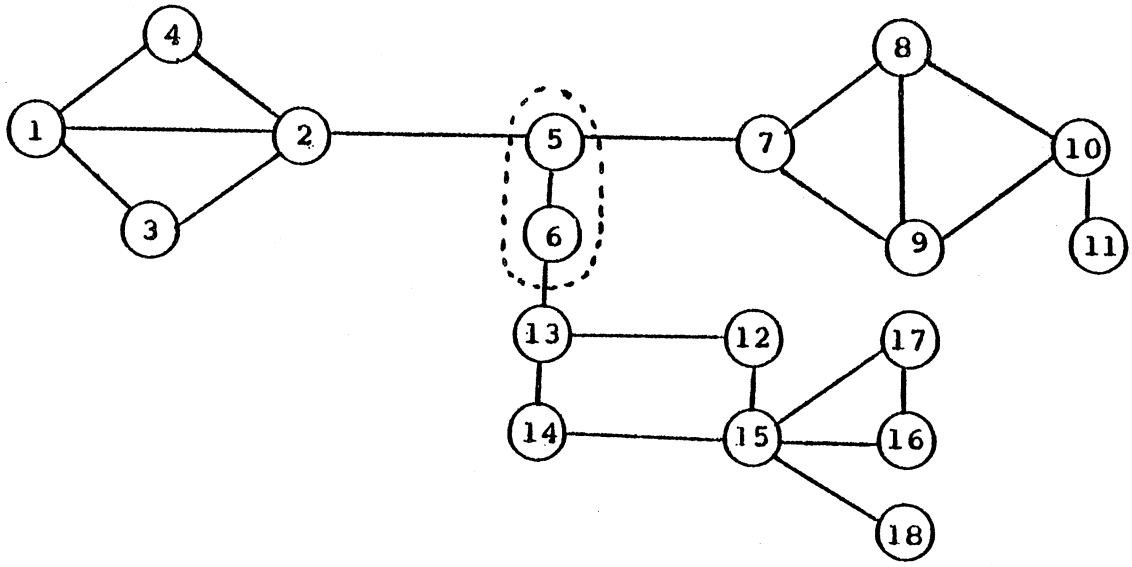
Construisons le graphe non-orienté A^*

$$\underline{A}^* = (A^*, X) \qquad A^* = \{a^*_{ij}\}$$

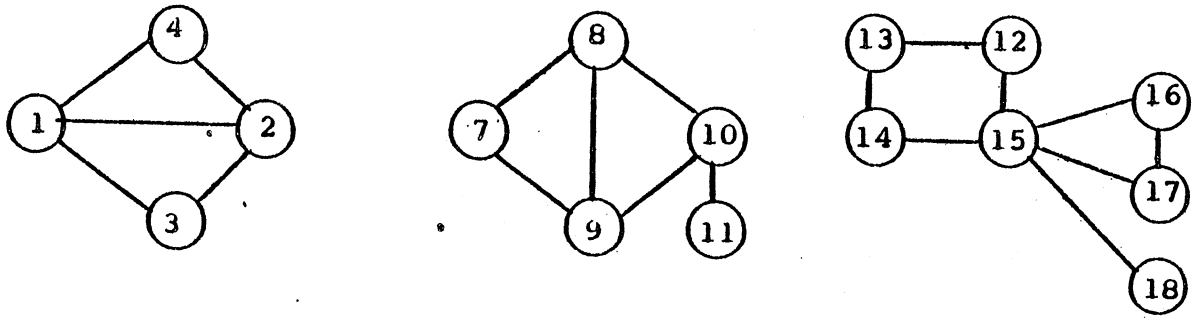
a^*_{ij} relie i à j si aij ou aji existe (A^* est le symétrisé de A).

Le problème précédent revient donc à minimiser le nombre de sommets qui déconnectent $\underline{A}^*$ en ρ sous-graphes $\underline{A}^*_{H_i}$.

Exemple :

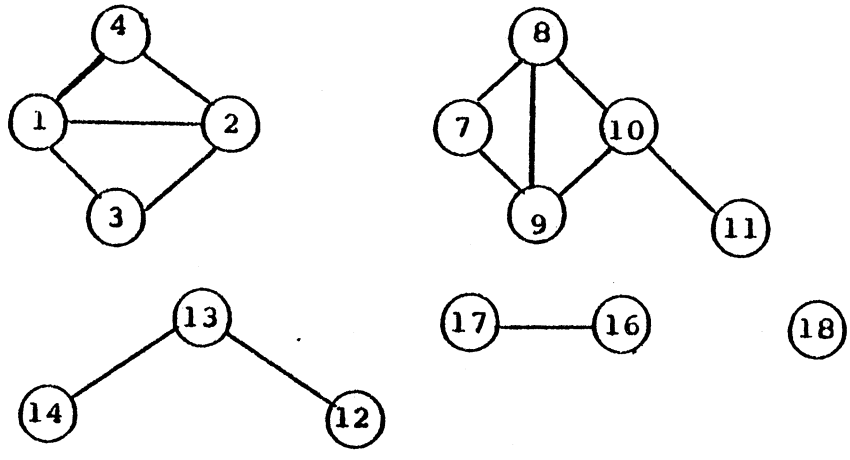


L'ensemble S = {5,6} déconnecte le graphe en



Cependant nous devons spécifier plus le problème si nous voulons le résoudre.

En effet dans l'exemple précédent S' = {5,6,15} pris comme séparateur déconnecte le graphe en



Ce résultat est moins régulier. Si nous voulons pallier à ce type de problème, nous devons imposer un ordre de grandeur pour la taille des sous-graphes que nous désirons obtenir.

Notre recherchons alors un ensemble minimal de sommets déconnectant $\underline{A}^*$ en ρ sous-graphes de taille comprise entre deux nombres nmin et nmax.

Pratiquement, un compromis est à chercher entre une régularité parfaite du graphe qui correspond à imposer nmax = nmin , et nous courons le risque d'avoir un ensemble S assez gros et un minimisation totale de S qui correspond à nmin = 0 , nmax = ∞ , avec le risque d'avoir un découpage irrégulier ne reposant sur rien de concret.

Il s'avère nécessaire de prendre nmin =α*nmax , α appartenant à [0.5;0.8] et d'avoir une idée de l'ordre de grandeur de nmax. La connaissance du problème traité et de la taille probable des unités physiques s'avère donc précieuse.

De plus, il s'agit pour nous, de trouver une bonne solution à notre problème pour un coût raisonnable.

IV) ALGORITHMES DE PARTITION ET DE RENUMEROTATION DES EQUATIONS

1) Algorithme de partition

L'algorithme présenté ci-dessous à un coût qui évolue en $O(\ell\, np)$, où ℓ est le nombre d'arcs du graphe et np le nombre de sommets du graphe ([Sang, 1977]).

```
PARTITION
k=0
S=∅
W=X
tant que(W≠∅) faire
début  H^0={x_0}
          S^0=V(x_0)
          k=k+1
          M_k=∅
          i=0;
```

Commentaire x_0 est un élément de W (de préférence faiblement connecté) Nous désignerons par V(x) le voisinage premier de x.

tant que (H^i < nmin) et ($S^i \neq \emptyset$) faire

début chercher x_{i+1} tel que

$$|V(x_{i+1}) \cup S^i| = \min_{x \in S^i} |V(x) \cup S^i|$$

$$H^{i+1} = H^i \cup \{x_{i+1}\}$$

$$S^{i+1} = S^i \cup V(x_{i+1}) - \{x_{i+1}\}$$

fin tant que

si ($S^i = \emptyset$) alors

début $H_k = H^i$

$W = W - H_k$

fin

sinon

début $I = \emptyset$

tant que (H^i < nmax) et ($S^i \neq \emptyset$) faire

début chercher x_{i+1} tel que

$$|V(x_{i+1}) \cup S^i| = \min_{x \in S^i} |V(x) \cup S^i|$$

$$S^{i+1} = S^i \cup V(x_{i+1}) - \{x_{i+1}\}$$

$$I = I \cup \{i\}$$

$$i = i+1$$

fin tant que

chercher $i^*/|S^{i^*}| = \min_{i \in I} |S^i|$

$S = S^{i^*} \cup S$

$H_k = H^{i^*}$

$W = W - S - H_k$

fin sinon

fin tant que

fin Partition

CASCADE calcule de manière formelle avant tout calcul des solutions le jacobien de F_n Nous allons donc utiliser la structure de la matrice obtenue pour appliquer notre heuristique puisque nous l'avons vu le symétrisé du graphe de connexion de F_n est similaire à celui du graphe de connexion de $\frac{\partial F_n}{\partial yz}$

2) Problème de la numérotation des équations

Cependant la numérotation des équations ne nous est pas indifférente. En effet utilisée telle quelle, cette numérotation peut nous amener à des sous-systèmes H non inversibles, ce qui nous le verrons plus loin, empêche l'utilisation d'une méthode de Newton à plusieurs niveaux.

Nous imposerons, afin de nous assurer au maximum de l'inversibilité des blocs obtenus que la ième inconnue intervienne dans la ième équation.

Il est toujours possible de numéroter ainsi les équations d'un système.

En effet, si le système $F_n(y_n, z_n) = \Theta_{p+q}$ admet une solution unique, alors son jacobien est inversible dans un voisinage de cette solution. Il nous suffit donc de montrer qu'une telle numérotation existe pour $\frac{\partial F_n}{\partial yz}$ pour montrer son existence pour $F_n(y_n\ z_n) = \Theta_{p+q}$.

Soit $\mathcal{M}_n(\mathbb{R})$ l'ensemble des matrices à coefficients réels n lignes - n colonnes.

Soit S_n l'ensemble des permutations de $\{1, \ldots n\}$.

PROPOSITION P(n) :

$$\forall A \in \mathcal{M}_n(\mathbb{R}) \quad \det A \neq 0 \quad \Rightarrow \quad \begin{cases} \exists \sigma \in S_n \\ \forall i \in \{1 \ldots n\} \\ a_{i\sigma(i)} \neq 0 \end{cases}$$

si $A = (a_{ij})$; $1 \leq i \leq n \qquad 1 \leq j \leq n$

P(1) est vraie de toute évidence.

Supposons P(r) vraie. Qu'en est il de P(r+1) ?

Soit A un élément de $\mathcal{M}_{r+1}(\mathbb{R})$ et A_j la sous-matrice de A obtenue en ôtant à A la colonne j et la ligne r+1 .

$$\det A = \sum_{j=1}^{r+1} (-1)^{r+1+i} a_{r+1\ j} \det A_j$$

Comme $\det A \neq 0$ alors il existe j_0 tel que $a_{r+1\ j_0} \det A_{j_0} \neq 0$.

$A_{j_0} = (a'_{ij})$ $1 \leq i \leq r$ $1 \leq j \leq r$ Nous avons $a'_{ij} = a_{ij}$ $j \in \{1,\ldots,j_0-1\}$; $a'_{ij} = a_{ij+1}$ $j \in \{j_0,\ldots,r\}$

P(r) étant vraie alors puisque $\det A_{i_0} \neq 0$ il existe σ^* appartenant à S_r tel que :

$$\forall\ i \in \{1,\ldots,r\} \qquad a'_{i\sigma^*(i)} \neq 0$$

Soient $H(j_0) = \{j \in \{1,\ldots,r\} \quad \sigma^*(j) < j_0\}$

et $\bar{H}(j_0) = \{1,\ldots,r\} - H(j_0)$

Soit σ
$$\begin{cases} \sigma(j) = \sigma^*(j) & j \in H(i_0) \\ \sigma(j) = \sigma^*(j)+1 & j \in \bar{H}(i_0) \\ \sigma(r+1) = j_0 \end{cases}$$

alors

$\forall\ j \in H(j_0)$ $\quad a_{j\ \sigma(j)} = a_{j\sigma^*(j)} = a_{j\ \sigma^*(j)} \neq 0$

$\forall\ j \in \bar{H}(j_0)$ $\quad a_{j\ \sigma(j)} = a_{j\ \sigma^*(j)+1} = a_{j\ \sigma^*(j)} \neq 0$

$a_{r+1\ \sigma(r+1)} = a_{r+1\ j_0} \neq 0$

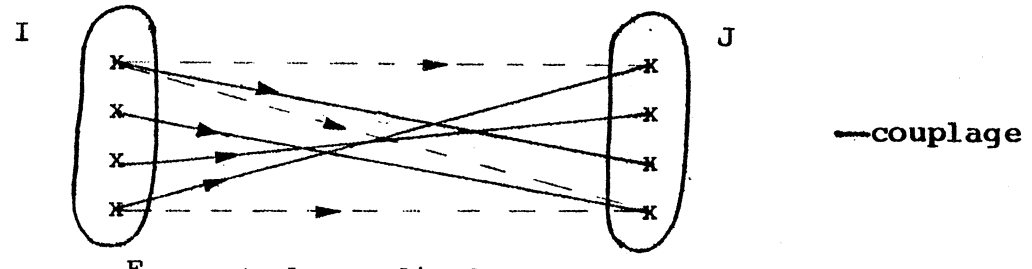
Donc $\forall\ j \in \{1,\ldots,r+1\}$ $a_{j\ \sigma(j)} \neq 0$.

σ^* est une bijection de $\{1,\ldots,r\}$ dans $\{1,\ldots r\}$ donc σ est une bijection de $H(j_0)$ dans $\{1,\ldots j_0-1\}$ et de $H(j_0)$ dans $\{j_0+1,\ldots,r+1\}$ Comme de plus $\sigma(r+1) = j_0$, alors σ est une bijection de $\{1,\ldots,r+1\}$ dans $\{1,\ldots,r+1\}$ et appartient donc à S_{r+1}.
P(r+1) est vérifiée.
Donc P(n) **est vraie quelque soit** n .

Si nous donnons à l'équation n° i le n° $\sigma(i)$ nous obtiendrons la numérotation cherchée.

3)Correspondance avec le problème du couplage parfait :

En réalité, nous venons de montrer l'existence d'un couplage E_0 entre l'ensemble I des lignes et l'ensemble J des colonnes où un arc relie une ligne i à une colonne j si a_{ij} est non-nulle



E_0 est de cardinal p+q

Ce problème peut aussi s'exprimer en termes de "problème de cases admissibles" ([Sak 1979 p. 136]).

En effet soit K un tableau de n lignes et n colonnes . La case K(i,j) sera dite admissible si $a_{ij}\neq 0$. Nous voulons affecter à K(i,j) une valeur x_{ij} ($x_{ij} \in \{0,1\}$) telle que :

$$\begin{cases} \forall\, i\,, & \sum_{j=1}^{n} x_{ij} = 1 \\ \forall\, j\,, & \sum_{i=1}^{n} x_{ij} = 1 \end{cases}$$

et si K(i,j) n'est pas admissible alors $x_{ij} = 0$.

Nous savons que dans les cas qui nous concernent, ce problème admet une solution. Il s'agit maintenant d'être à même d'en exhiber une .

DEFINITION :

Une solution (x_{ij}) , $1 \leq i \leq n$, $1 \leq j \leq n$, $x_{ij} \in \{0,1\}$ sera dite réalisable si et seulement si :

$$\begin{cases} \forall\, i \in \{1,\ldots,n\} \quad , \quad \sum_{j=1}^{n} x_{ij} \leq 1 \\ \forall\, j \in \{1,\ldots,n\} \quad , \quad \sum_{i=1}^{n} x_{ij} \leq 1 \\ x_{ij} = 0 \ , \ \forall\, (i,j) \ / \ K(i,j) \text{ non admissible.} \end{cases}$$

Une première remarque est que si pour une solution réalisable donnée, il y a exactement m lignes de K telles que les sommes sur ces lignes soient nulles, alors il y a exactement m colonnes de K telles que les sommes sur ces colonnes soient nulles.

Les algorithmes existants transforment une solution réalisable quelconque non optimale (n-t lignes dont les sommes sont nulles) en une meilleure solution réalisable (n-t-1 lignes dont les sommes sont nulles).

Une première approche serait de partir de la solution nulle (toujours réalisable) pour en n pas, obtenir une solution optimale. Cependant, ces algorithmes sont coûteux ($O(\ell)$ opérations par pas). Il est préférable de mettre en œuvre une heuristique afin d'obtenir une solution réalisable meilleure que la solution nulle et ensuite d'appliquer un algorithme de cases admissibles.

L'heuristique proposée plus loin est en $O(\ell/n)$ par pas (cf. chapitre III pour une étude expérimentale de la complexité). Elle trouve la plupart du temps une solution optimale en n pas. Dans ce cas, le traitement complet est en $O(\ell)$. Si, par aventure, elle n'arrive pas à trouver une solution optimale, elle fournit une bonne solution réalisable, que nous améliorons par un algorithme de cases admissibles jusqu'à trouver une solutions optimale.

Il est à noter que dans tous les cas, nous faisons n pas, mais que suivant la réussite ou l'échec de l'heuristique, les derniers pas peuvent être plus coûteux.

4) DESCRIPTION DE L'HEURISTIQUE :

$NL(i) = \{j / a_{ij} \neq 0,\ j \in \{1,\ldots,n\}\}$ $\quad i=1 \ldots n$

$NC(j) = \{i / a_{ij} \neq 0 \quad i \in \{1,\ldots,n\}\}$ $\quad j=1 \ldots n$

ECHOUE=Faux;

faire pour i allant de 1 à n ILORD(i)=0;

iter=1.

tant que (iter < n) faire

début si ($\exists$ i $\in \bar{I}$/NL(i)=1) alors

début iiter=i

chercher j $\in \bar{J}$/$a_{iiter\ j} \neq 0$

jiter=j

fin

sinon si ($\exists$j $\in \bar{J}$/N(ij)=1) alors

début jiter=j

chercher i $\in \bar{I}$/$a_{ijiter} \neq 0$

iiter=i

fin

sinon si ($\exists$i $\in \bar{I}$/$a_{ij} \neq 0$) alors

début choisir i et j tels que

- $i \in \bar{I}$
- $j \in \bar{J}$
- $a_{ij} \neq 0$
- $NL(i) = \min_{k \in \bar{I}} NL(k)$
- $NC(j) = \min_{k \in \bar{J}} \{NC(h) / a_{ih}\}$

fin

sinon ECHOUE=Vrai.

si (non ECHOUE) alors

début ILORD(iiter)=jiter

faire pour j allant de 1 à n

si $a_{iiter\ j} \neq 0$ alors NC(j)=NC(j)-1

faire pour i allant de 1 à n

si $a_{i\ jiter} \neq 0$ alors NL(i)=NL(i)-1.

I = I U {iiter}

J = J U {jiter}

$\bar{I}$ = $\bar{I}$-{iiter}

$\bar{J}$ = $\bar{J}$-{jiter}

```
        iter=iter+1
fin
sinon sortir de la boucle tant que et utiliser un algorithme
de cases admissibles
fin tant que
fin HEURISTIQUE
```

5) **Description de l'algorithme de cases admissibles**

```
ERREUR=Faux
tant que iter < n faire
début   FINI=Faux
        I*=Ī
        J*=∅
        pour i ∈ Ī faire ICORD(i)=0
        tant que (non FINI) faire
        début   si ( ∃i ∈ I* et j ∉J*/a_ij≠0) alors
                début J*=J* U {i}
                      ICORDINV(j)=i
                      si ( ∃ τ ∈ I/ILORD(τ)=j)
                      alors FINI=Vrai
                      sinon si ( ∃i ∉ I* et j ∈ J*/ILORD(i)=j)
                            alors
                            début I*=I* U {i}
                                  ICORD(i)=j
                            fin
                            sinon FINI=Vrai
                fin
        fin tant que
        si (non ERREUR) alors
        début   ILORD(i)=j
                I=I U {i}
                J=J U {j}
                tant que (ICORD(i)≠0) faire
                début   j=ICORD(i)
                        i=ICORDINV(j)
                        ILORD(i)=j
                        I=I U {i}
                fin tant que
```

iter=iter+1
$\bar{I}=\{1,\ldots,n\}-I$
$\bar{J}=\{1\ \ldots\ n\}-J$

fin;
sinon iter=n+1
fin tant que
fin ALGORITHME

L'intérêt de cette heuristique est qu'elle nous permet de savoir si un sous-système de $Fn(yn\ zn) = \theta_{q+p}$ est triangulaire. donc resoluble simplement.Cela nous permettra de réduire la taille du système à résoudre.

En effet, si pour iter allant de 1 à k, nous avons NL(iiter)=1 alors la sous-matrice de $\frac{\partial F_n}{\partial yz}(y_n\ z_n)$ constituée par les lignes i ∈ I ,et par les colonnes j ∈ J, où I={iiter,iter=1,... k} et J = {jiter,iter=1 ...k} est une matrice triangulaire inférieure.

EXEMPLE :

Soit la matrice de départ A (* représente un élément non nul de A).

$$A = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} * & & & \\ * & * & * & * \\ * & & & * \\ * & & * & * \end{bmatrix} \end{array}$$

. Mettons en oeuvre l'heuristique précédente :

iter=1

NL(1)= 1 NC(1)= 4 iiter=i_1=1
jiter=j_1=1
NL(2)= 4 NC(2)= 1
NL(3)= 2 NC(3)= 2
NL(4)= 3 NC(4)= 3

iter=2

NL(1)= 0	NC(1)= 0	
NL(2)= 3	NC(2)= 1	
NL(3)= 1	NC(3)= 2	iiter=i_2=3
		jiter=j_2=4
NL(4)= 2	NC(4)= 3	

iter=3

NL(1)= 1	NC(1)=-1	
NL(2)= 2	NC(2)= 1	
NL(3)= 0	NC(3)= 2	
NL(4)= 1	NC(4)= 0	iiter=i_3=4
		jiter=j_3=3

iter=4

NL(1)=-2	NC(1)=-2	
NL(2)= 1	NC(2)= 1	iiter=i_4=2
		jiter=j_4=2
NL(3)=-1	NC(3)= 0	
NL(4)= 0	NC(4)=-1	

Si nous plaçons les équations n°iiter à la place jiter, alors nous avons:

	1	2	3	4
1	*			
2	*	*	*	*
4		*	*	*
3	*			*

Cette matrice est semblable à

	1	4	3	2
	(j_1)	(j_2)	(j_3)	(j_4)
1 (i_1)	*			
3 (i_2)	*	*		
4 (i_3)	*	*	*	
2 (i_4)	*	*	*	*

L'exemple suivant est instructif sur les limites de cette heuristique

soit A la matrice suivante:

	1	2	3	4	5	6
1	*	*				*
2		*			*	*
3	*		*		*	
4			*	*	*	
5			*	*	*	
6			*	*	*	

L' heuristique,qui au départ a le choix entre toutes les lignes puisqu'elles ont toutes 3 elements non-nuls attribuera l'équation n°1 à la variable n°1 ce qui rendra impossible l'attribution d'une équation à la variable n°6 .
Cependant l'heuristique attibuera les lignes 1,2,3,4 5 aux équations n°1 2 3 4,5 et fournira ainsi une solution réalisable à l'algorithme de cases admissibles qui en un pas nous permettra d'avoir la solution optimale suivante:

	1	2	3	4	5	6
3	*		*		*	
2		*			*	*
4			*	*	*	
5			*	*	*	
6			*	*	*	
1	*	*				*

La structure de $\frac{\partial F_n}{\partial yz}$ est souvent celle d' une matrice très creuse.Il sera donc avantageux de la présenter sous forme de liste

Le sous programme TRI qui a été composé prend en entrée une structure en 6 tableaux ·

- IDEBLIG(i) adresse de l'élément non-nul de A qui débute la ligne i
- IDEBCOL(i) adresse de l'élément non-nul de A qui

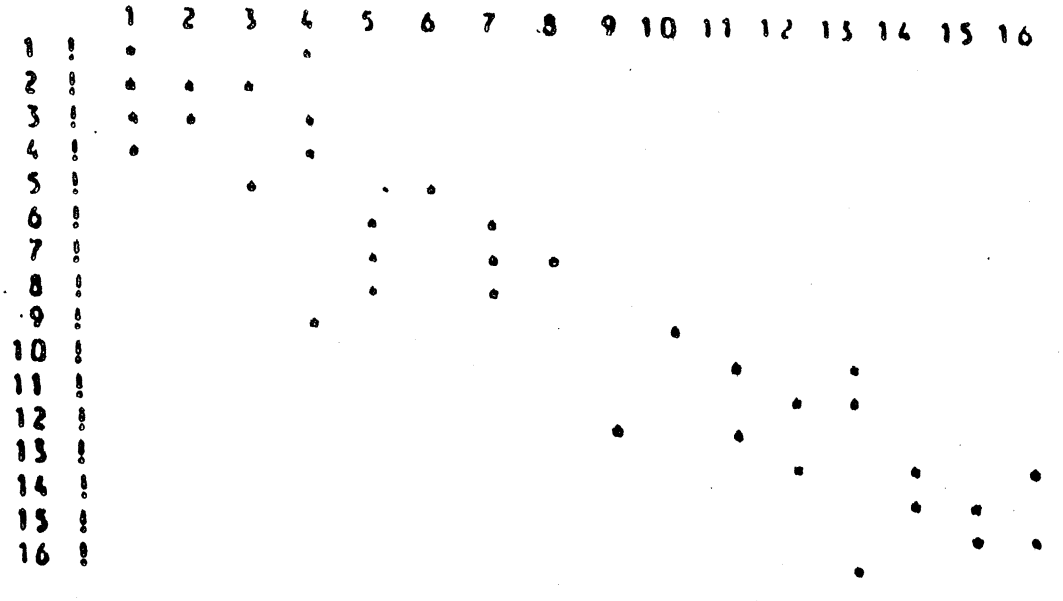
débute la colonne i

- IL(i) indice ligne de l'élément qui a pour adresse i
- IC(i) indice colonne de l'élément qui a pour adresse i
- SVL(i) adresse de l'élément non-nul suivant qui se trouve sur la même ligne s'il existe 0 sinon
- SVC(i) adresse de l'élément non-nul suivant qui se trouve sur la même colonne s'il existe , 0 sinon

Ce sous-programme opère une permutation sur les indices lignes pour obtenir une matrice diagonale (étape 1).

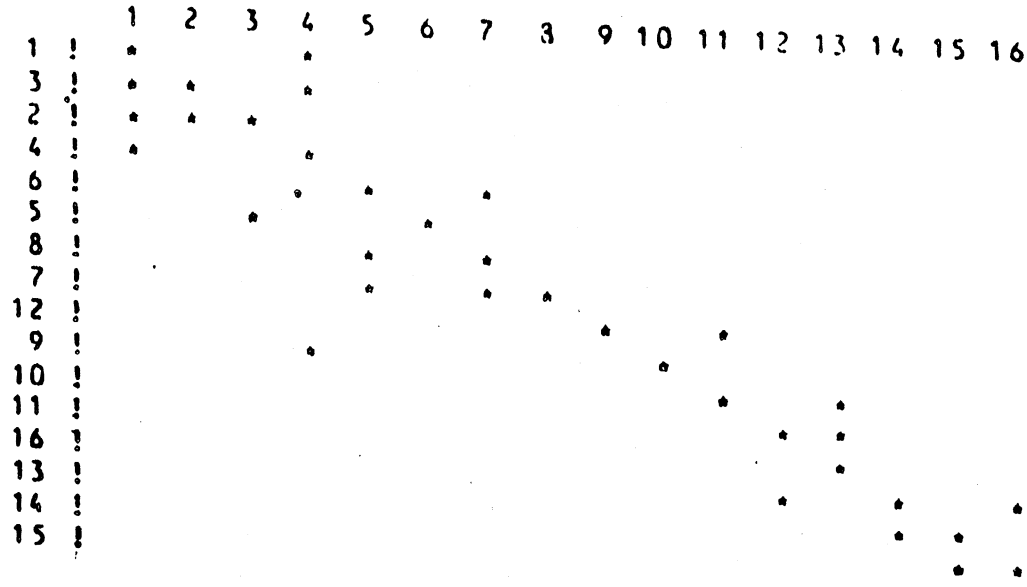
Puis il procède au traitement du graphe associé pour obtenir finalement une matrice par blocs avec bord grâce à une permutation symétrique des lignes et des colonnes (voir les exemples)

exemple n° 1



n=16

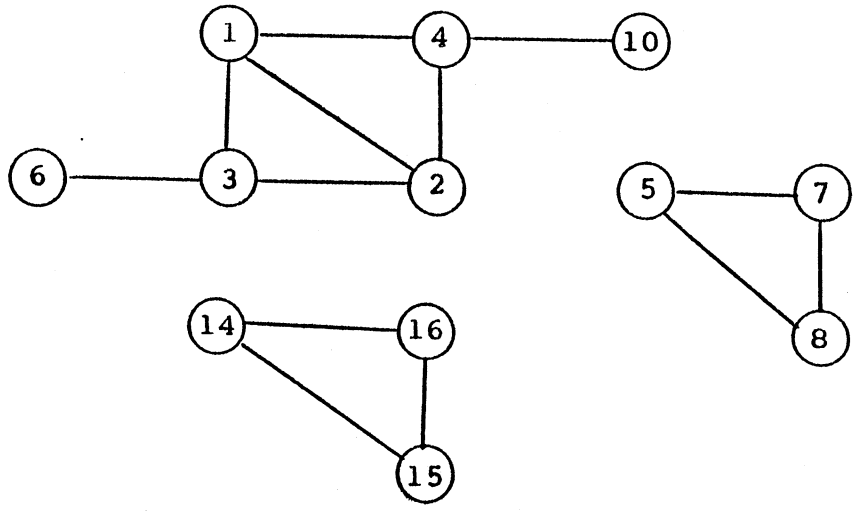
nombre d'éléments non-nuls=35



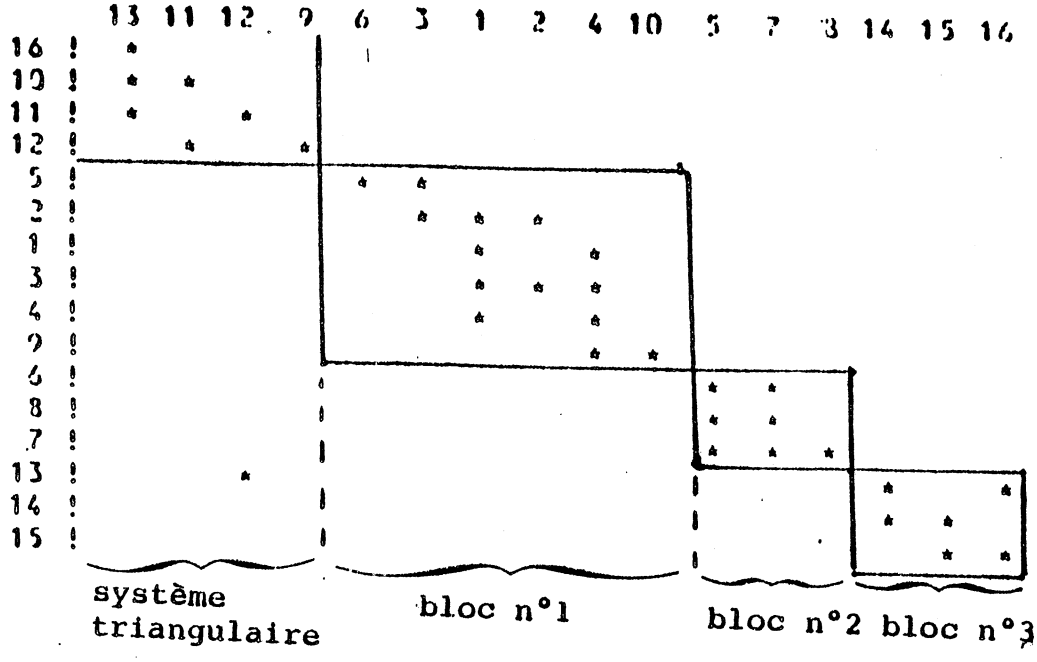
etape n°1

L'étape n 1 nous indique que les équations {16 10,12,11} associées aux inconnues {13,9 11,12} forment un système triangulaire

nous avons le graphe suivant:



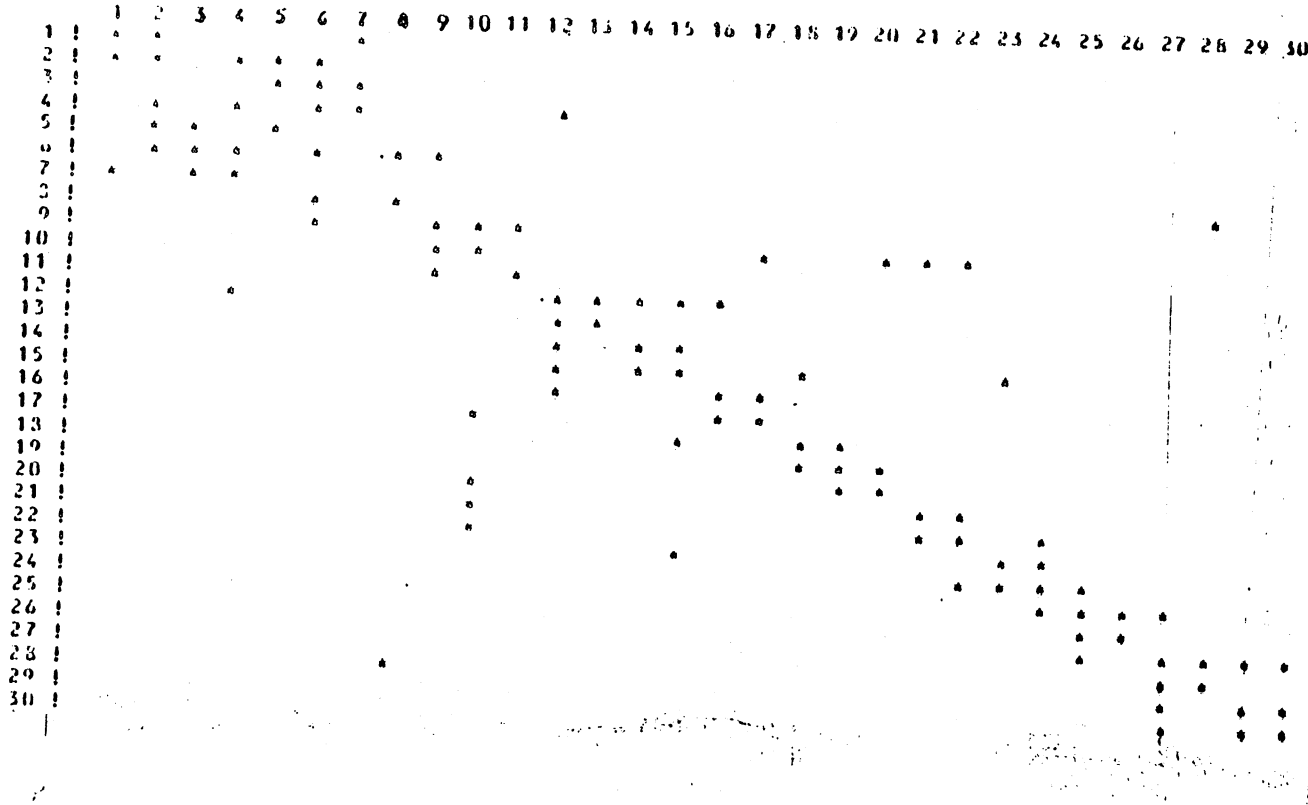
qui est celui d'un graphe déconnecté en 3 grappes.Il est à noter que le fait d'avoir décelé un sous-système triangulaire nous permet de discerner une décomposition totalement découplée du système associé au graphe précédent



étape n°2

exemple n°2

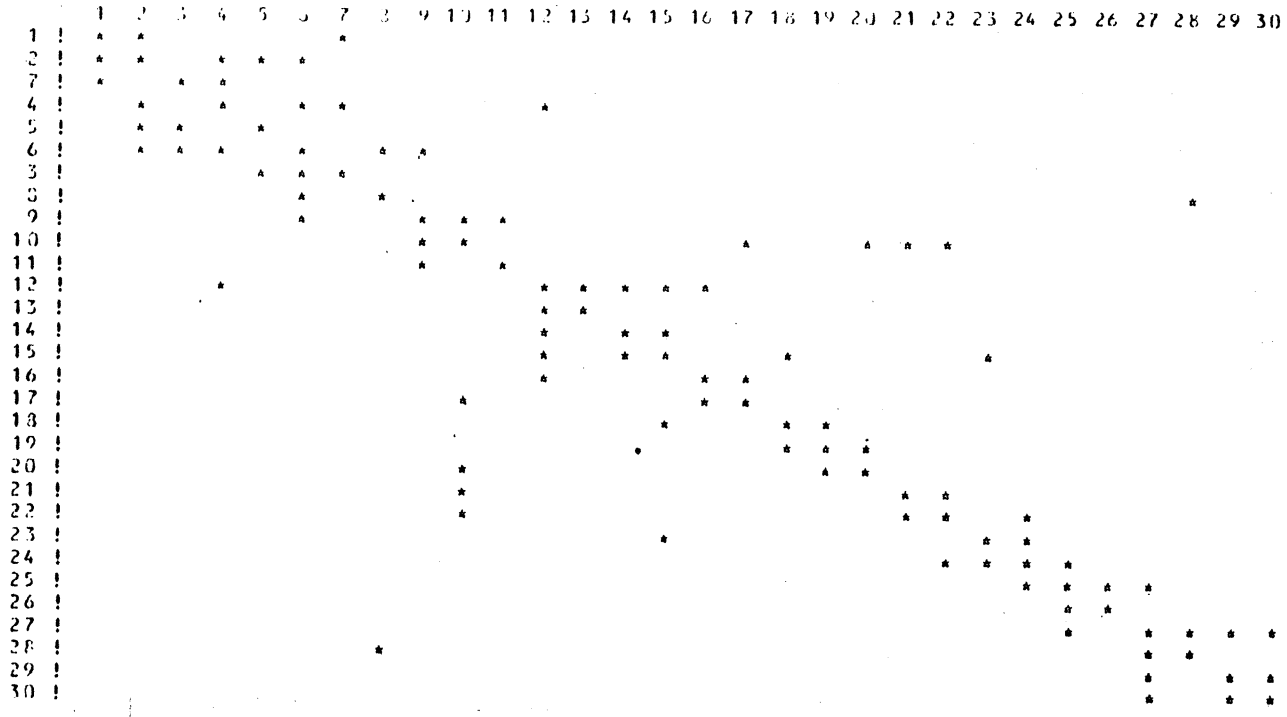
matrice initiale



n=30

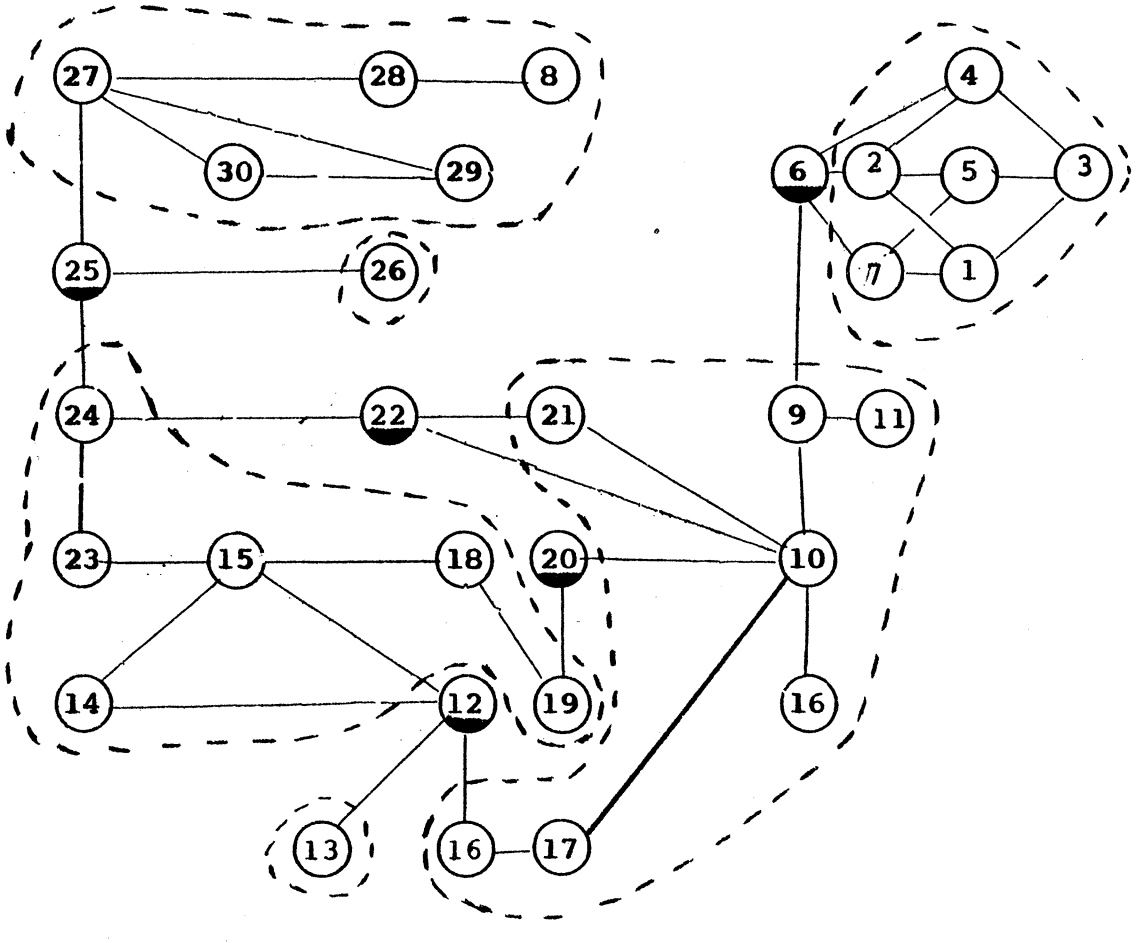
nombre d'éléments non-nuls:108

étape n°1

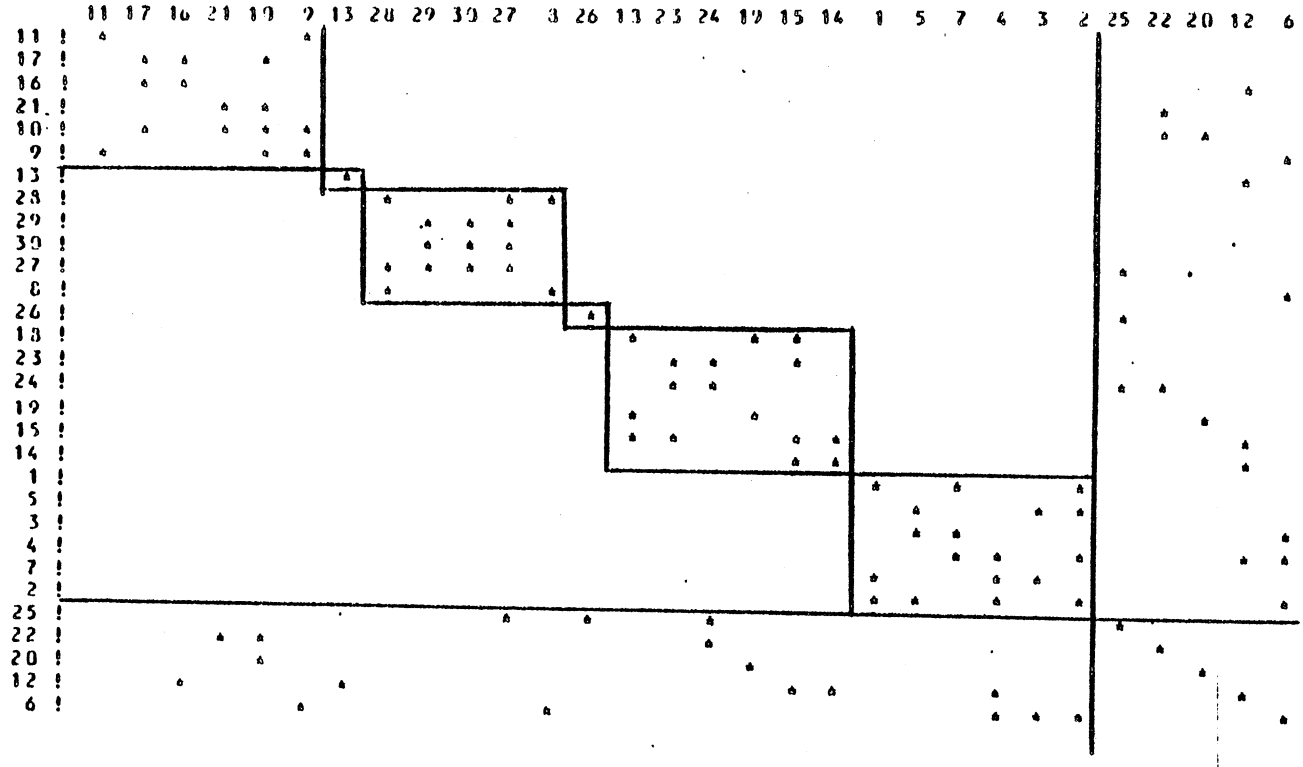


graphe associé

(nmin=10,nmax=14)

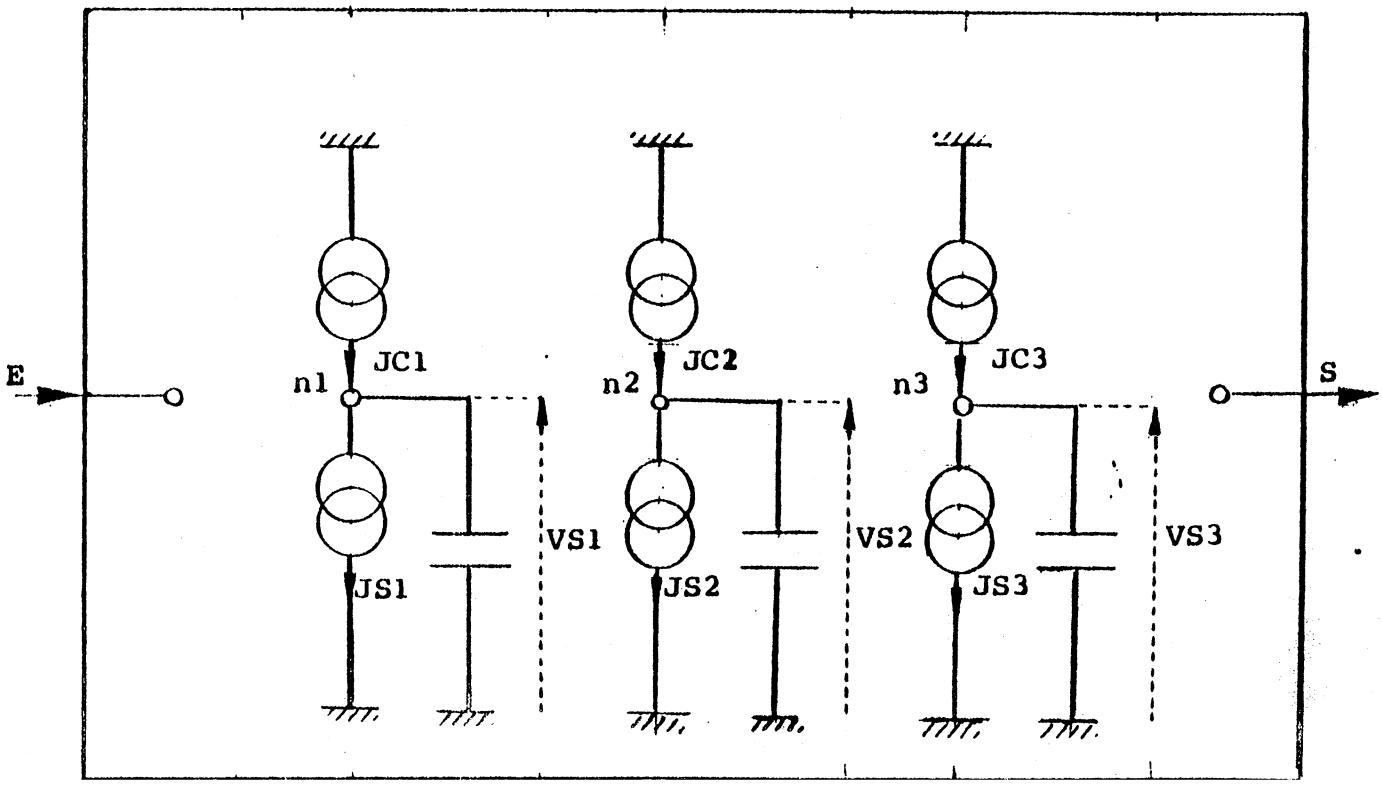


d'où la structure en bloc suivante



exemple n°3

cet exemple est issu de l'étude du circuit ci-dessus·



expressions définies par l'utilisateur

$\beta_1 = 10^{-6}.E$ $\beta_2 = 10^{-6}.VS1$ $\beta_3 = 10^{-6}.VS2$

$\alpha_{\gamma 1} = 10^{-6}.VS1$ $\alpha_2 = 10^{-6}.VS2$ $\alpha_3 = 10^{-6}.VS3$

$1 = \alpha_1$ $\gamma_2 = \alpha_2$ $\gamma_3 = \alpha_3$

VS1 = Vn1 VS2 = Vn2 VS3 = Vn3

$JS1=\alpha_1.VS1+\beta_1.E+VS3$ $JS2=\alpha_2.VS2+\beta_2.VS1$ $JS3=\alpha_3.VS3+\beta_3.VS2$

$JC1 = \gamma_1.VS1$ $JC2 = \gamma_2.VS2$ $JC3 = \gamma_3.VS3$

mise en équations :

1. Vp1 - Vn1 = 0
2. Vp2 = 0
3. Vp3 = 0
4. Vp4 - Vn1 = 0

.....

17. Vp17 - Vn3 = 0
18. Vp18 = 0

---

19. ip1 + ip2 = 0
20. ip3 + ip4 = 0

.....

27. ip17 + ip18 = 0

---

28. ip1 + ip4 + ip13 = 0
29. ip5 + ip8 + ip15 = 0
30. ip9 + ip12 + ip17 = 0

---

31. ip1 + JS1 = 0
32. ip3 + JC1 = 0

.....

36. ip11 + JC3 = 0
37. $ip13 - c_1.\hat{V}p13 + c_1.\hat{V}p14 = 0$
38. $ip15 - c_2.\hat{V}p15 + c_2.\hat{V}p16 = 0$
39. $ip17 - c_3.\hat{V}p17 + c_3.\hat{V}p18 = 0$

Posons $(y_1,\ldots,y_{33},z_1,z_2,\ldots,z_6)$
$=(Vp1,\ldots,Vp12,ip1,\ldots,ip18,Vn1,Vn2,Vn3,Vp13,\ldots,Vp18)$
Alors nous avons à résoudre le système

$$f(y_1,\ldots,y_{12},\ldots,y_{33},z_1,\ldots,z_6,\dot{z}_1,\ldots,\dot{z}_6,t) = \theta_{39}$$

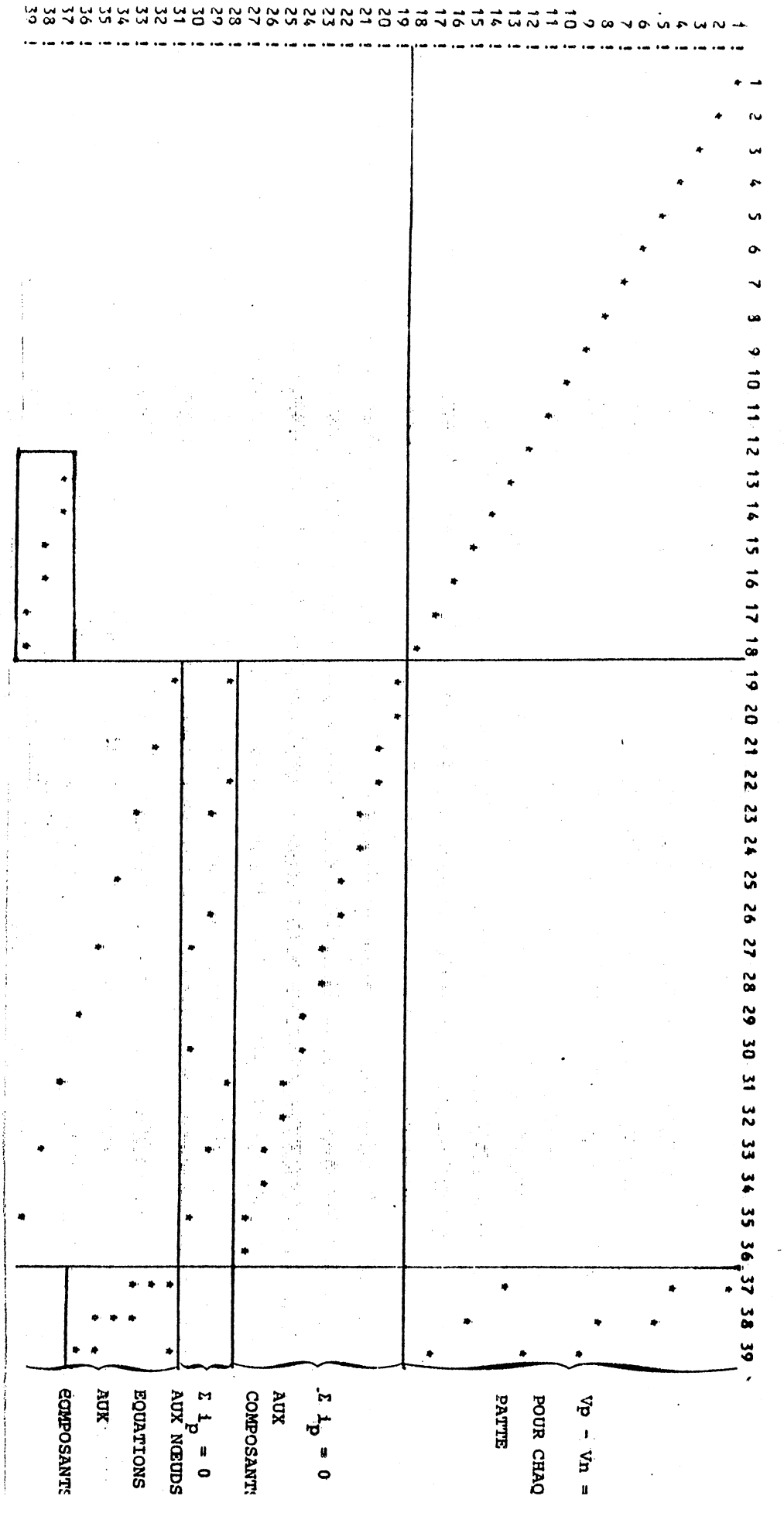
ou encore, après discrétisation du problème au point tn,

$$F_n(y_n, z_n)=\theta_{39}$$

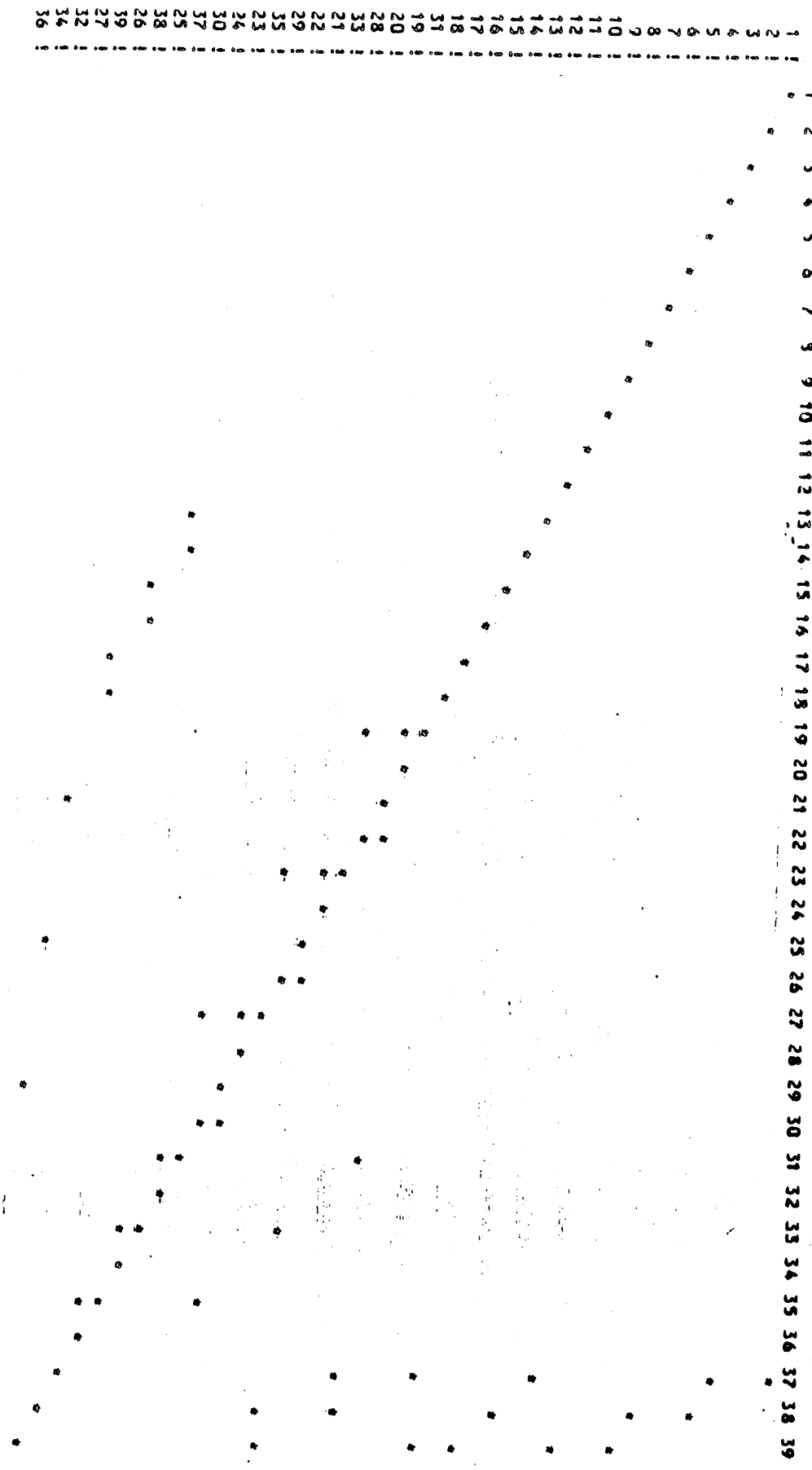
La structure du jacobien $\frac{\partial F_n}{\partial yz}$,son graphe associé et son traitement par TRI sont visibles sur les figures ci-après.

Il est à remarquer que la symétrie ternaire u circuit a été preservée par le découpage en blocs du jacobien, ce qui met ainsi en évidence les liens qui existent entre le graphe de connexion du jacobien et la structure "topologique" du circuit.

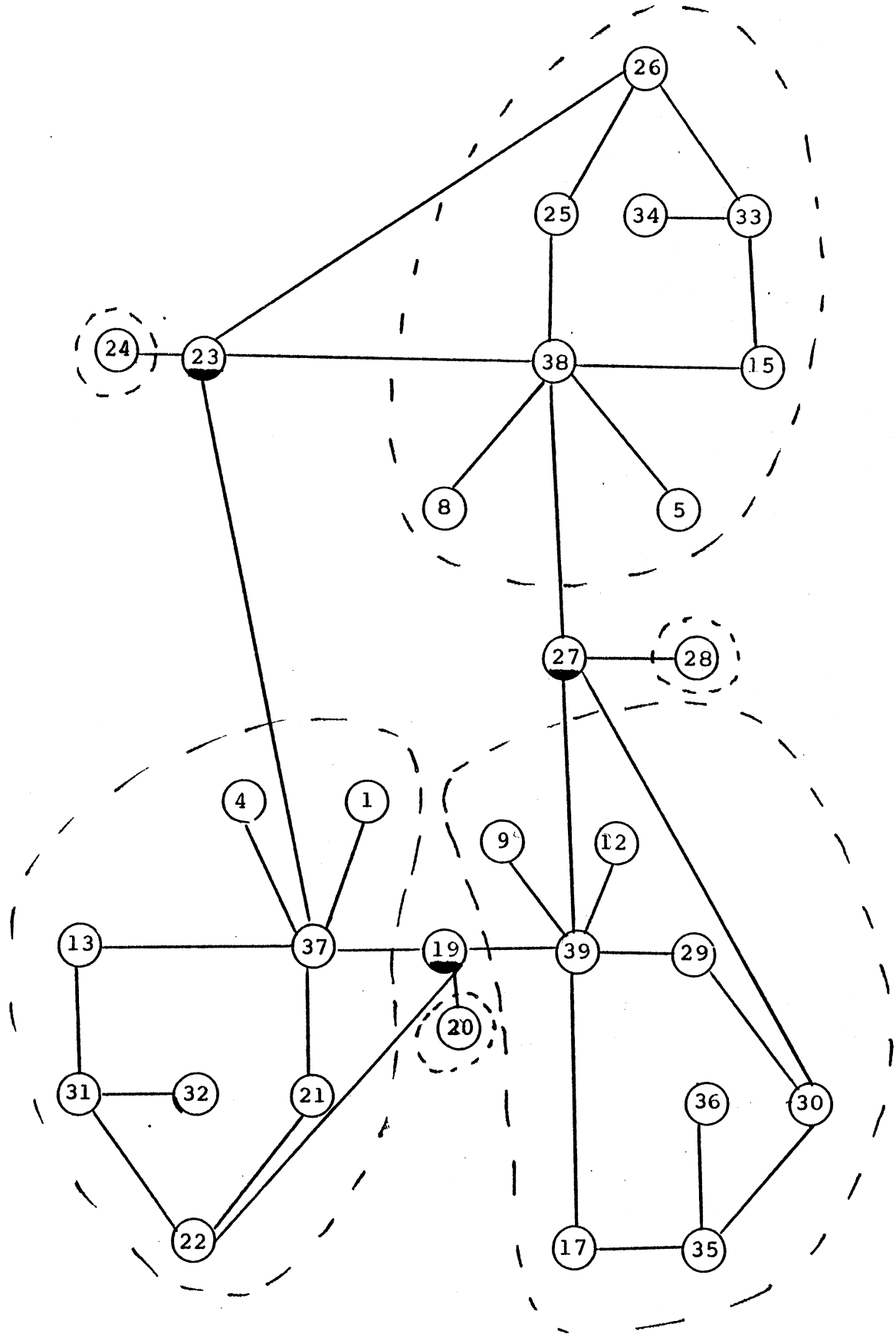
structure du jacobien



étape n° 1

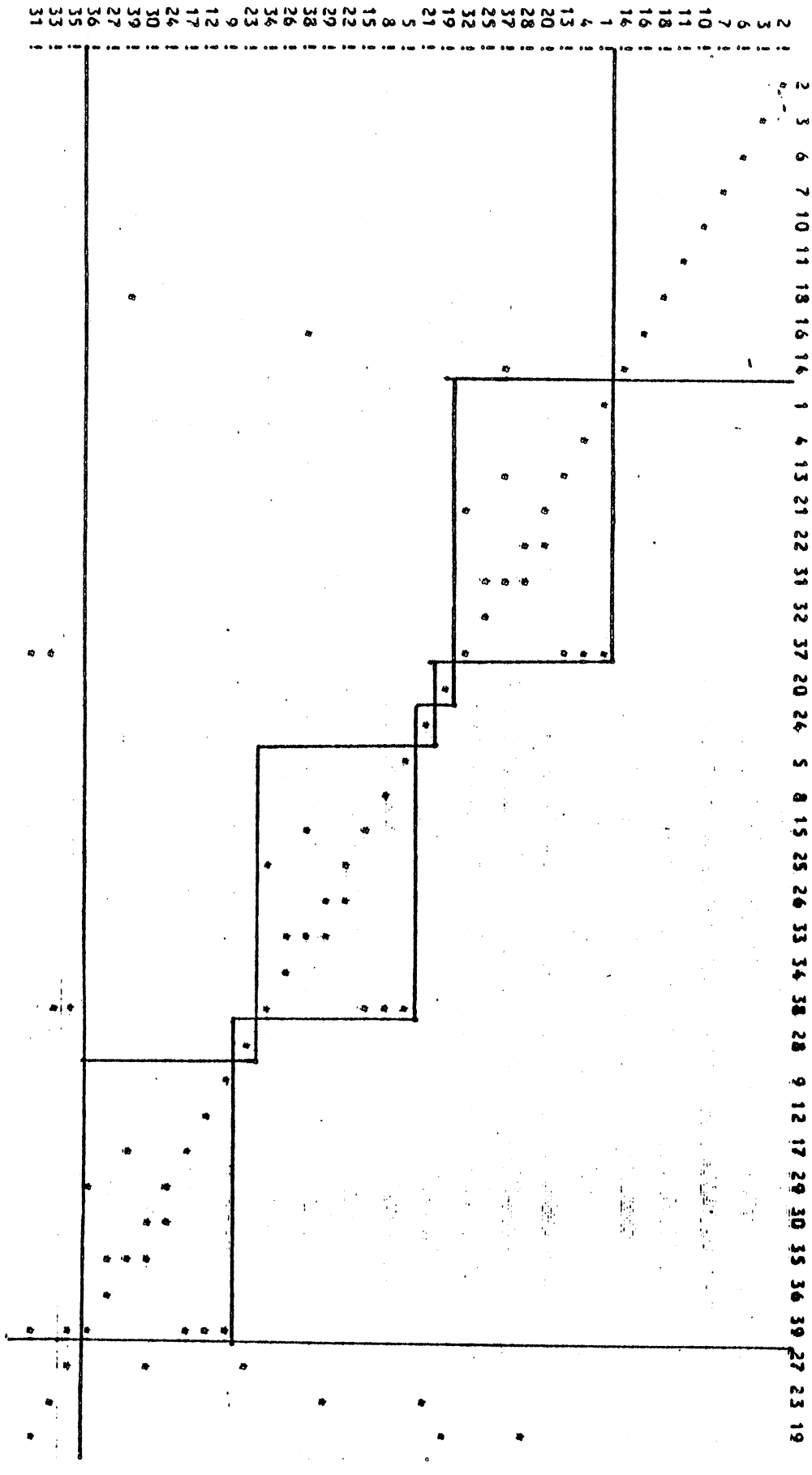


graphe associe au jacobien



S = { 19;23;27 }

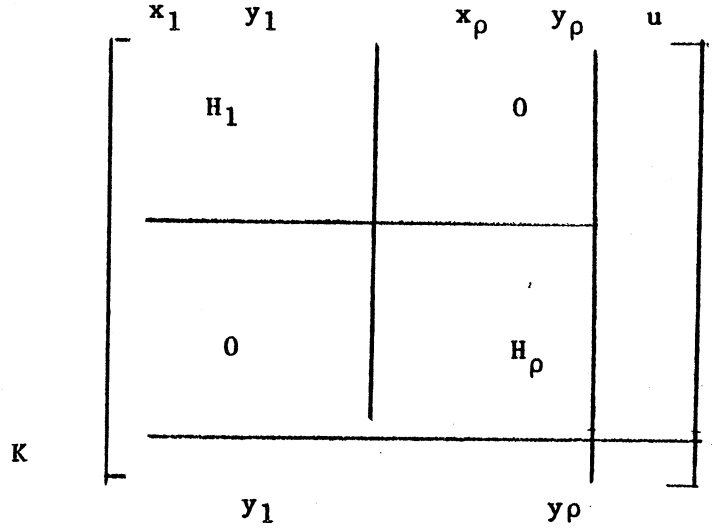
étape n°2



V) METHODE DE NEWTON A PLUSIEURS NIVEAUX ET ALGORITHMES CORRESPONDANTS

Revenons à l'idée exprimée par Rabbat et Sangiovanni-Vincentelli de remplacer une méthode de Newton globale par une méthode hiérarchisée tenant compte de modularités physiques.

Imaginons que nous ayons, grâce aux algorithmes précédents, pu mettre en évidence dans la matrice de $\frac{\partial F_n}{\partial yz}$ la structure suivante :



Le lecteur voudra bien nous excuser de la confusion qui existe entre les variables algébriques notées y_n et ce qui est appelé par la suite "sorties d'un sous-système H_j et qui sont notées y_j, mais nous sommes tributaires de notations largement utilisées et adoptées par ailleurs.

Pour aider à la compréhension du texte qui suit, nous avons pris soin de ne faire intervenir la notation "y_n" concernant les variables algébriques que dans l'expression "$F_n(y_n, z_n)$".

1) Formulation mathématique de la méthode

Les variables de H_j peuvent être décomposées en trois catégories (en reprenant des notations déjà utilisées précédemment)

$$\begin{cases} u_j & \text{les entrées} \quad (\in \mathbb{R}^{\sigma 1 j}) \\ x_j & \text{les variables internes} \quad (\in \mathbb{R}^{\pi j}) \\ y_j & \text{les sorties} \quad (\in \mathbb{R}^{\sigma 2 j}) \end{cases}$$

$F_n(y_n, z_n) = \theta_{q+p}$ s'écrit alors, si nous posons $n_j = \pi_j + \sigma 2_j$ et $n_u = p+q+\sum_{j=1}^{\rho} n_j$:

$$\begin{cases} H_j(x_j, y_j, u_j) = \theta_{nj} \\ \qquad j = 1, \ldots \rho \\ K(y_1, \ldots, y_\rho, u) = \theta_{nu} \end{cases}$$

Soit $\Omega_j = \{(u\ x\ y)/H_j(u\ x\ y) = \theta_{nj}\}$
Soit $U_j = \{u/\ (u,x,y) \in \Omega_j\}$
Soient les hypothèses (H) suivantes :

$$\begin{cases} \bullet\ \exists\, L > 0\ /\ \forall\ (u_1,x_1,y_1) \in \Omega_j, (u_2,x_2,y_2) \in \Omega_j \\ \qquad \|dH(u_1,x_1,y_1) - dH(u_2,x_2,y_2\| \leq L\|(u_1,x_1,y_1) - (u_2,x_2,y_2)\| \\ \bullet\ \exists\, \alpha > 0\ /\ \forall\ (u,x,y) \in \Omega_j \\ \qquad \|dH(u,x,y)\| \leq \alpha \\ \bullet\ \frac{\partial H}{\partial xy}\ (u\ x\ y)^{-1} \text{ existe} \end{cases}$$

alors il existe une fonction $G_j = (G_{xj}, G_{yj})$ continuement differentiable telle que

$$\forall\ u \in U_j \quad H_j(u, G_{xj}(u), G_{yj}(u)) = \theta_{nj}$$

G_j sera appelée **macromodèle fonctionnel** de H_j . De plus

$$\frac{\partial G_j}{\partial u}(u) = -\frac{\partial H_j}{\partial x_j y_j}(u\ G_{xj}(u)\ G_{yj}(u))^{-1} \cdot \frac{\partial H_j}{\partial u}(u, G_{xj}(u), G_{yj}(u))$$

et

$$\frac{\partial K\,(G_{y1}(u) \quad G_{y\rho}(u),u)}{\partial u} =$$

$$\frac{\partial K}{\partial u}(G_{y1}(u),\ldots,G_{y\rho}(u)\ u) + \sum_{j=1}^{\rho} \frac{\partial K}{\partial yj}\,(G_{y1}(u),\ldots,j) \cdot \frac{\partial G_{yj}}{\partial u}\,(u)$$

(pour la démonstration de cette proposition voir [Sang, 1979])

2) Description d'un algorithme de Newton à plusieurs niveaux

paramètres : $\tau^{\circ} \in \mathbb{R}+$, $\eta \in \mathbb{R}+$
données u^{o} , $x_j^{o\,o}$, $y_j^{o\,o}$
tant que ($\|\Delta u\| \geq \eta$) faire
début i=0
pour j allant de 1 à ρ faire
debut tant que ($\|(\Delta x_j, \Delta y_j)\| > \tau^i$) faire
début calculer $H_j(u^i, x_j^{i,k}, y_j^{i,k})$
calculer $\frac{\partial H_j}{\partial x_j y_j}\,(u^i, x_j^{i\,k}, y_j^{i\,k})$
résoudre

$$\frac{\partial H_j}{\partial x_j y_j}\,(u^i\ x_j^{i\,k}, y_j^{i\,k}) \cdot \begin{pmatrix} \Delta x_j \\ \Delta y_j \end{pmatrix} + H_j(u^i, x_j^{i,k}, y_j^{i\,k}) = \theta_{n_i}$$

k=k+1
fin tant que

$$(x_j^{i+1,o}, y_j^{i+1,o}) = (x_j^{i,k+1}, y_j^{i,k+1})$$

$$G_{yj}(u^i) = y_j^{i+1,o}$$

$$\frac{\partial G}{\partial u_j}\,(u^i) = -\frac{\partial H_j}{\partial x_j y_j}\,(u^i\ x_j^{i+1,o}, y_j^{i+1,o})^{-1} \cdot \frac{\partial H_j}{\partial u_j}\,(u^i\ x_j^{i+1,o}, y_j^{i+1,o})$$

fin pour

$$\text{calculer } A = \frac{\partial K}{\partial u}\,(G_{y1}(u^i),\ldots,G_{y\rho}(u^i), u^i)$$

calculer

$$B = \sum_{j=1}^{\rho} \frac{\partial K}{\partial y_y} (G_{y1}(u^i), \ldots, G_{y\rho}(u^i)\ u^i) \cdot \frac{\partial Gy_j}{\partial u} (u^i).$$

calculer $C = K(G_{y1}(u^i) \qquad G_{y\rho}(ui)\ ui)$;

résoudre $(A+B) \cdot \Delta u + C = \theta_{n_u}$

$ui+1 = ui + \Delta u$

$\tau i+1 = \min(\tau^i, \|\Delta u\|^2)$

fin tant que

fin ALGORITHME

Sangiovanni-Vincentelli précise dans son exposé que $\tau^{i+1} = \min(\tau^i, \|\Delta u\|^2)$ permet d'obtenir une convergence quadratique alors qu'elle n'est que linéaire si nous prenons $\tau^{i+1} = \min(\tau^i, \|\Delta u\|)$.

De plus, nous n'avons pas besoin de calculer entièrement

$$\frac{\partial G}{\partial u_j} (u^i) = \begin{pmatrix} \frac{\partial Gx_j}{\partial u_j} (u^i) \\ \\ \frac{\partial Gy_j}{\partial u_j} (u^i) \end{pmatrix} \quad \text{car seul} \quad \frac{\partial Gy_j}{\partial u_j} (u^i)$$

nous intéresse . Il est proposé de remplacer le calcul de $\frac{\partial H_j}{\partial x_j y_j} (u^i, x_j^{i+1,o}, y_j^{i+1,o})^{-1}$ qui représente n_j systèmes linéaires de dimension n_j à résoudre, par la résolution des $\sigma 2_j$ systèmes linéaires suivants

$$\frac{\partial H_j}{\partial x_j y_j} (u^i, x_j^{i+1\ 0}, y_j^{i+1\ 0})^T \cdot M = \begin{bmatrix} 0 & \pi_j \\ I & \sigma_{2j} \end{bmatrix}$$

où $0_{\pi j}$ représente la matrice nulle $\pi j \times \sigma 2_j$ et $I_{\sigma 2j}$ la matrice identité $\sigma_{2j} \times \sigma 2j$,suivi du calcul de:

$$\frac{\partial G_{y_j}}{\partial u} (u^i) = - M^T \cdot \frac{\partial H_j}{\partial u} (u^i\ x_j^{i+1,0}, y_j^{i+1,0})$$

3) **Etude de la latence**

La structure numérique précédente s'adapte bien à une étude de la latence.

Soit $H_j(u\ x_j\ y_j) = \theta_{nj}$ un sous système de $Fn(y_n, z_n) = \theta_{q+p}$

Soit $H_j^*(u, \dot{u}_d, x, \dot{x}_d, y, \dot{y}_d, t) = \theta_{nj}$ le sous-système algébrodifférentiel correspondant à H_j

Nous dirons que H_j^* dépend d'un signal s'il existe une application G^* continuement différentiable de

$$\mathbb{R}^{\sigma 1j} \times \mathbb{R}^{\sigma 1jd} \times \mathbb{R}^{\pi j} \times \mathbb{R}^{\pi jd} \times \mathbb{R}^{\sigma 2j} \times \mathbb{R}^{\sigma 2jd} \times [t_0;T]$$

vers $\mathbb{R}^{\sigma 1j+\pi j+\sigma 2j}$ telle que

$$\forall\ t \in [t_0;T] \qquad \forall\ (u,p_u) \in \mathbb{R}^{\sigma 1j} \times \mathbb{R}^{\sigma 1jd}$$

$$\forall\ (x,p_x) \in \mathbb{R}^{\pi j} \times \mathbb{R}^{\pi jd}\ ,\quad \forall\ (y,p_y) \in \mathbb{R}^{\sigma 2j} \times \mathbb{R}^{\sigma 2jd}$$

$$H_j^*(u,pu,x,px,y,py,t) = G^*(u,p_u,x,px,y,py,g(t))$$

Hypothèses

1) le système admet une solution unique

2) $\forall\ t \in [t_1,t_2]$ $\begin{cases} g(t) = c_t \\ u(t) = c_u \end{cases}$

3) $\exists\ \varepsilon > 0\ /\ \forall\ t \in [t_1, t_1+\varepsilon]\quad x(t) = c_x\ ,\ y(t) = c_y$

alors

$$\forall\ t \in [t_1,t_2]\quad x(t) = c_x \text{ et } y(t) = c_y$$

En effet. c_x c_y et c_u étant solutions sur $[t_1\ t_1+\varepsilon]$ nous avons $G^*(c_u, \theta_{\sigma_{1j}}, c_x, \theta_{\pi_j}, c_y, \theta_{\sigma_{2j}}, c_t) = \theta_{n_j}$

Donc le problème

$$\begin{cases} G^*(u,\dot{u}_d\ x,\dot{x}_d,y,\dot{y}_d,g(t)) = \theta_{n_i} \\ t \in [t_1,t_2] \\ x(t_1) = c_x \\ y(t_1) = c_y \end{cases}$$

admet la solution $\begin{cases} x(t) = c_x \\ y(t) = c_y \\ t \in [t_1, t_2] \end{cases}$

Pratiquement si nous savons que g(t) reste constant pendant un certain laps de temps, par exemple si g est une fonction créneau, et que nous constations que u(t), x(t), y(t) restent constants alors nous pouvons en conclure que tant que u(t) et g(t) restent constants x(t) et y(t) le restent aussi.

Cela se traduit numériquement par (en posant $g_n = g(t_n)$)

Si $\begin{cases} x_{n+1} = x_n \\ y_{n+1} = y_n \text{ alors} \\ u_{n+1} = u_n \\ g_{n+1} = g_n \end{cases}$

début i=0
tant que $((u_{n+i}=u_n)$ et $(g_{n+i}=g_n))$ faire
début i=i+1
$x_{n+i} = x_n$
$y_{n+i} = y_n$
fin tant que
fin si

Malheureusement cette formulation ne tient pas compte des err urs dues à la méthode et à la précision machine.

Elle est inutilisable telle quelle.

Soit $E_x^{t_r}$ l'erreur de troncature tolérée pour la variable x

Si nous avons $\begin{cases} i=-1,\ldots -j \\ \dfrac{\| x_{n+i} - x_n \|}{\| x_n \| + test} \leq E_x^{t_r} \\ \dfrac{\| y_{n+i} - y_n \|}{\| y_n \| + test} \leq E_y^{t_r} \\ g(t_{n+i}) = g(t_n) \end{cases}$

et si nous pouvons déterminer ε_u tel que, tant que

$$\frac{\| u_{n+i} \; u_n \|}{\| u_n \| + test} \leq \varepsilon_u \quad (i=-j,\ldots,0,1\ldots.)$$

$$\frac{\| x_{n+i} - x_n \|}{\| x_n \| + test} \leq E_x^{t_r} \quad \text{et} \quad \frac{\| y_{n+i} \; y_n \|}{\| y_n \| + test} \leq E_y^{t_r} \quad \text{alors} \quad x \text{ et } y$$

seront déclarés latents tant que le signal $g(t)$ restera constant et que les variations relatives de u seront inférieures à ε_u .

ε_u est délicat à déterminer car il dépend des expressions dans lesquelles u intervient

Nous pouvons remarquer que :

$$Hj(u_{n+1},x_{n+1},y_{n+1}) \simeq Hj(u_n,x_n,y_n) + \frac{\partial H_j}{\partial u_j}(u_n,x_n,y_n) \cdot (u_{n+1} - u_n) + \frac{\partial H_j}{\partial xy}(u_n,x_n,y_n) \cdot \begin{pmatrix} x_{n+1}-x_n \\ y_{n+1}-y_n \end{pmatrix}$$

Donc

$$\begin{pmatrix} x_{n+1}-x_n \\ y_{n+1}-y_n \end{pmatrix} \simeq \left(- \frac{\partial H_j}{\partial xy}(u_n \; x_n y_n)^{-1} \cdot \frac{\partial H_j}{\partial uj}(u_n , x_n \; y_n)\right) \cdot (u_{n+1} - u_n)$$

ou encore :

$$\begin{pmatrix} x_{n+1}-x_n \\ y_{n+1}-y_n \end{pmatrix} \simeq \frac{\partial G_j}{\partial u_j}(u_n) \cdot (u_{n+1} - u_n)$$

Les variations de x et de y sont donc fonction des variations de u .

Il n'est pas très utile d'avoir une très grande précision sur x car celui ci n'intervient pas en dehors de son bloc et n'influe pas en retour sur u . Contrôler u suffit à contrôler x .

Par contre y intervient dans les équations de couplage et une erreur sur y a des répercussions sur u ce qui a comme corollaire la possibilité de solutions parasites. Nous nous attacherons à :

$$\Delta y_{n+1} = \frac{\partial G_{y_j}}{\partial u_j}(u_n) \cdot \Delta u_{n+1}$$

$\frac{\partial G_{y_j}}{\partial u_j}$ est calculé au cours de l'algorithme de Newton à plusieurs niveaux.

L'algorithme adopté pour la latence est le suivant:

LATENCE
LATENT=Faux
si $(g(t_{n+1})=g(t_n))$ alors
début $E_u = \frac{\| u_n - u_{n+1}^{p_r} \|}{\| u_n \| + test}$

si $(E_u \leq E_u^{t_r})$ alors
début $E_x = \frac{\| x_n - x_{n+1}^{p_r} \|}{\| x_n \| + test}$

$E_y = \frac{\| y_n - y_{n+1}^{p_r} \|}{\| y_n \| + test}$

si $((E_x \leq E_x^{t_r})$ et $(E_y \leq E_y^{t_r}))$ alors
début $\Delta y = \frac{\partial G_{y_j}}{\partial u_j} \cdot (u_{n+1}^{p_r} - u_n)$

si $(\frac{\| \Delta y \|}{\| y_n \| + test} \leq E_y^{t_r})$

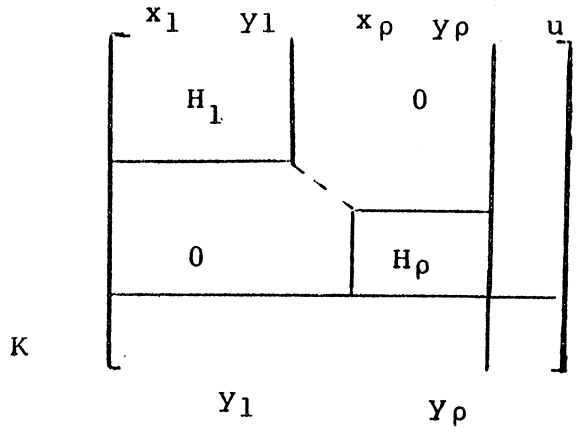
alors LATENT=Vrai
fin si

fin si
fin si
fin LATENCE

Nous avons donc là un outil simple qui répond aux questions que nous nous étions posées lors des expérimentations sur les critères de latence.

VI) DECOUPLAGE D'UN SYSTEME COUPLE

Revenons à la structure en blocs de F_n



Soit U_i l'ensemble des entrées de H_i et $U = \bigcup_{i=1}^{\rho} U_i$

Nous avons vu précédemment que la renumérotation σ des équations était en réalité un couplage de l ensemble des équations sur l'ensemble des inconnues, qui attribue à chaque équation une inconnue. Nous appelons K_{u_i} le sous-système de K constitué par les équations "correspondant" aux inconnues appartenant à U_i

Nous allons maintenant introduire un algorithme de surrelaxation qui nous permettra de traiter séparément chaque bloc.

RELAXATION

paramètre ε_0

pour i=1,..., ρ faire le calcul de u_j^o, x_j^o, y_j^o par des formules de prédiction.

pour i=1,...,ρ faire

début $\mu = 2\varepsilon_0$

$j=0$

tant que $(\mu > \varepsilon_0)$ faire

début calculer y_i^j et $\frac{\partial G_{y_i}}{\partial u}(u_i^j)$ en résolvant

$H_i(x_i^j, y_i^j, u_i^j) = \theta_{n_i}$

former Vect = $(u_1^{(n)},\ldots,u_{i-1}^{(n)},u_i^j\ u_{i+1}^o,\ldots,y_\rho^o$
$x_1^{(n)},\ldots,x_{i-1}^{(n)}\ x_i^j\ x_{i+1}^o,\ldots,x_\rho^o$
$y_1^{(n)},\ldots,y_{i-1}^{(n)},y_i^j\ y_{i+1}^o,\ldots,y_\rho^o)$

calculer

$$u_i^{j+1}=u_i^j-\left[\frac{\partial K_{u_i}}{\partial u_i}(vect)+\frac{\partial K_{u_i}}{\partial y_i}(vect)\circ\frac{\partial G_{y_i}}{\partial u_i}(u_i^j)\right]^{-1}\circ K_{u_i}(vect)$$

$\mu=\|u_i^{j+1}-u_i^j\|$
$j=j+1$
fin tant que
$x_i^{(n)}=x_i^{j-1}$;
$y_i^{(n)}=y_i^{j-1}$
$u_i^{(n)}=u_i^j$;
fin pour.
fin RELAXATION

L'utilisation de cet algorithme, proposé par Sangiovanni-Vincentelli Sang,1980 permet de diminuer de manière considérable la complexité des calculs puisque, lorsque les calculs sont menés pour un bloc M_i toutes les variables autres que celles du bloc sont considérées comme constantes

Cependant cet algorithme n'a pas les propriétés de convergence d'une méthode de Newton à plusieurs niveaux

D'une part sa convergence est linéaire (au lieu d'être quadratique) d'autre part les conditions de convergence sont plus sévères.

Si nous posons

$$B(u) = \begin{bmatrix} \frac{\partial K_{u_1}}{\partial u_1}(u_1) & 0 & \cdots & 0 \\ 0 & 0 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \frac{\partial K_{u_\rho}}{\partial u_\rho}(u_\rho) \end{bmatrix}$$

$$L(u) = \begin{bmatrix} 0 & \cdots & 0 \\ \frac{\partial K_{u_2}}{\partial u_2}(u_2) & 0 & 0 \\ \vdots & & \vdots \\ \frac{\partial K_{u_\rho}}{\partial u_1}(u_\rho) & \cdots & \frac{\partial K_{u_\rho}}{\partial u_{\rho-1}}(u_\rho) \quad 0 \end{bmatrix}$$

$$V(u) = \begin{bmatrix} 0 & \frac{\partial K_{u_1}}{\partial u_2}(u_1) & \cdots & \frac{\partial K_{u_1}}{\partial u_\rho}(u_1) \\ 0 & & & \\ \vdots & & & \frac{\partial K_{u_{\rho-1}}}{\partial u_\rho}(u_{\rho-1}) \\ 0 & \cdots & & 0 \end{bmatrix}$$

Il faut alors en plus des conditions d'existence d'un macromodèle fonctionnel que soient vérifiées les conditions suivantes(nous supposons que $\hat{u}$ est solution):

1) $H(\hat{u}) = -(B(\hat{u})+L(\hat{u}))^{-1} \cdot V(\hat{u})$ existe

2) le rayon spectral de $H(\hat{u})$ est inférieur à 1

3) $B(\hat{u})$ est inversible

alors Relaxation converge pour toute valeur initiale uo prise dans un voisinage de $\hat{u}$. De plus sa convergence est linéaire [Sang, 1980] .

Nous pouvons aller plus loins et découpler totalement Fn en utilisant des pas et des ordres différents pour chaque bloc. Attribuons à chaque bloc H_i un pas hi (calculé en tenant compte des seules variables de H_i par des méthodes similaires à celles de Brayton [Bray, 1972] ou Van Bokhoven [Van Bock , 1975]) et un temps t_i^n $(= t_i^{n-1}+h_i)$ qui est le point où H_i doit être résolu à nouveau.

Les H_i sont classés de telle sorte que les valeurs ti soient par ordre croissant Nous utiliserons pour cela la structure de liste suivante (3 tableaux)

temps	pas	après

temps (i) = t_i^n
pas (i) = h_i

après (i) = j si H_j est le bloc classé après H_i
(n+1 si dernier)

début pointe sur le premier bloc (début = k si H_k est le premier bloc). Nous traiterons toujours Hdébut. Après la résolution, nous restructurons la liste pour que les t_j soient toujours par ordre croissant).

Voici l'algorithme :

```
DECOUPLAGE
        tA=t
        tant que (tA < tmax) faire
        début
            calculer(xdébut,ydébut,udébut)
            Hdébut(xdébut,ydébut,udébut) = θdébut
            Ku_début(ydébut,udébut) = θn_u_début
        si (non convergence) alors
            calculer nouveau pas  h*  et nouvel ordre k*
        début   tA=tA-pas(début)+h*
                pas(début)=h*
                temps(début)=tA
        fin
        sinon
        début   calculer nouveau pas h* et nouvel ordre k*;
                temps(début)=temps(début)+h*
                pas(début)=h*
                chercher j et k tels que
                temps(j)<temps(début)<temps(k)
                après(j)=k
                debl=après(début)
                après(j)=début
                après(début)=k
                début=debl
                tA=temps(début)
            commentaire :
            le cas où j n'existe pas (lorsque par
            exemple début reste en tête) doit être traité à part
        fin si
fin tant que
fin DECOUPLAGE
```

CONCLUSION

Tous les algorithmes présentés dans le chapitre II doivent être testés numériquement pour pouvoir rendre compte de leur efficacité et de leur validité. C'est l'objet du prochain chapitre.

Les outils de recherche opérationnelle utilisés dans ce chapitre sont sujets à améliorations, et des travaux devront être entrepris en ce sens s'il s'avère que leur coût croît de manière trop rapide avec la taille du problème.

CHAPITRE III

INTRODUCTION

Le début de ce chapitre est consacré à l'étude de l'erreur de troncature et des formules liant la taille du pas et l'ordre utilisé à cette dernière dans des équations non-linéaires. Diverses méthodes de résolution sont proposées et des comparaisons de temps effectuées. Nous nous attachons ensuite à la mise au point d'une démarche cohérente qui nous permette de conduire au mieux l'intégration d'un système algébrodifférentiel. Nous testons cette démarche en résolvant divers problèmes raides classiques, dont celui lié à l'oscillateur de Van der Pol.

La suite du chapitre est consacrée à l'élaboration d'un logiciel de résolution des systèmes algébrodifférentiels raides, intégrant la méthode de Newton classique, la méthode de Newton à plusieurs niveaux et la méthode de surrelaxation (ces deux dernières méthodes ont été décrites dans le chapitre II). Nous comparant sur un exemple les trois méthodes tant du point de vue nombre de pas que de celui du temps C.P.U. utilisé.

En fin de chapitre, une étude est faite du coût des méthodes de partitionnement en blocs d'une matrice.

I)PRINCIPE

Soit le système (S) $\begin{cases} f(y(t),z(t),\dot{z}(t),t) = \theta_{q+p} \\ t \in [t_0,T] \\ z(t_0) = \mu \end{cases}$

avec f de classe $C^k (k > 1)$ allant de $\mathbb{R}^q \times \mathbb{R}^p \times \mathbb{R}^p \times [t_0,T]$ vers $\mathbb{R}^{q+p}$.

Nous allons utiliser pour résoudre ce système des méthodes issues des schémas quasi-stables mises en évidence par Gear (cf. chapitre I, partie V)).

Posons $g_{k_n}(z) = \alpha_0 z + \sum_{i=1}^{k_n} \alpha_i z_{n-i}$

en calculant les valeurs réelles α_i de telle sorte que

$$\dot{P}(t_n) = g_{k_n}(P(t_n))$$

pour tout polynôme P^n de degré inférieur ou égal à k_n $(1 \leq k_n \leq 6)$, ce qui assure à la méthode un ordre k_n.

z_{n-i} désigne des approximations numériques de $z(t_{n-i})$.

L'approximation de $\dot{z}(t)$ par $g_{k_n}(z(t))$ nous permet de transformer (S) en une suite de systèmes algébriques

$$(S_n)_{n=1,\ldots} = \{ F_n(y_n,z_n) = \theta_{q+p} \}_{n=1,\ldots}$$

où F_n, allant de $\mathbb{R}^q \times \mathbb{R}^p$ vers $\mathbb{R}^{q+p}$ est de classe supérieure ou égale à 1.

Pour résoudre ce système, on utilise généralement une méthode de Newton [Hac, 1971, p.146] ou des méthodes dérivées (cf chapitre II) qui ont une convergence quadratique pour une valeur initiale (y_n^{pr}, z_n^{pr}) suffisamment proche de la valeur exacte (y_n, z_n).

La démarche la plus efficace à l'heure actuelle consiste à calculer des valeurs réelles $(\gamma_i)_{i=1,\ldots,k_n+1}$

telles que, pour tout polynôme Q de degré inférieur ou égal à k_n, $Q(t_n) = \sum_{i=1}^{k_n+1} \gamma_i Q(t_{n-i})$

Brayton[Bray, 1972] proposa le premier un algorithme efficace pour le calcul des α_i et des valeurs de prédiction y_n^{pr} et z_n^{pr}, basé sur l'utilisation des polynômes de Lagrange.

Cet algorithme, qui utilise une approche matricielle, est cependant lourd à mettre en œuvre et de ce fait, les méthodes à ordre et pas variables n'étaient utilisées que lorsqu'on voulait une précision élevée. Les travaux de Van Bokhoven furent à cet égard décisifs, puisqu'il proposa un algorithme reposant sur l'utilisation des polynômes de Newton, beaucoup plus maniable et demandant moins d'opérations arithmétiques pour sa mise en œuvre.

DESCRIPTION DE L'ALGORITHME

(note : nous supposons $(S_i)_{i=0,\ldots,n-1}$ résolus; nous posons $h_i = t_n - t_{n-i}$, $h'_i = t_{n-1} - t_{n-i-1}$; pour $i > 1$ $\bar{z}_n^{\,i}$ désigne le prédicteur d'ordre i-1)

<u>début</u>

$\beta = 1/h_1$

$d = h_1/h'_1$

$\bar{z}_n^{\,1} = z_n$

$S = \dfrac{\bar{z}_n^{\,1}}{h_1}$

pour i allant de 1 à k_n-1 faire

début

$$\bar{z}_n^{i+1} = \bar{z}_n^{i} + d\cdot(z_{n-1} - \bar{z}_{n-1}^{i})$$

$$d = d\cdot(h_{i+1}/h'_{i+1})$$

$$\beta = \beta + 1/h_{i+1}$$

$$S = S + (\bar{z}_n^{i+1}/h_{i+1})$$

fin pour

utiliser $g_{k_n}(z) = \beta\cdot z - S$ pour résoudre (S_n)

fin algorithme

Cet algorithme permet de changer de pas et d'ordre sans difficultés, la seule contrainte étant $\| k_n - k_{n-1} \| \leq 1$. Or toutes les stratégies adoptées jusqu'à présent pour le choix du pas et de l'ordre ne proposent pas des variations dans l'ordre supérieure à 1, pour ne pas créer des instabilités durant l'intégration du système algébrodifférentiel à résoudre [Gear, 1971] [Nord, 1962].

II)CALCUL DE L'ERREUR DE TRONCATURE

Les démarches adoptées dans les principaux logiciels de résolution qui ont vu le jour jusqu'à présent (une rapide description de ces logiciels est donnée dans [Rueh, 1981]) s'appuient sur une évaluation la plus fine possible de l'erreur faite à chaque pas.

Un premier outil important fut la formule montrée par Brayton [Bray, 1972] :

$$\dot{z}(t_n) - g_{k_n}(z(t_n)) = \frac{z(t_n) - z_n}{h_{k_n+1}} + o(h_1^{k_n+1})$$

Mais une fois de plus, ce sont les travaux de Van-Bokhoven qui donnent l formule usitées aujourd'hui :

$$z_n - z(t_n) = \frac{z_n - z_n^{pr}}{1 + \alpha_0 h_{k_n+1}} + o(\|z(t_n) - z_n\|)$$

qui a été démontrée pour des systèmes différentiels purs.

Nous pouvons étendre cette formule aux systèmes algébrodifférentiels.

proposition 1

Supposons que l'hypothèse (H_3, chapitre I, partie IV) , proposition 8) soit vérifiée.

Supposons en outre (H_1) :

$$\|\frac{\partial f}{\partial \dot{yz}}(y(t_n),z(t_n),\dot{z}(t_n),t_n)^{-1} \cdot \frac{\partial f}{\partial z}(y(t_n),z(t_n),\dot{z}(t_n),t_n)\| \ll \alpha_0$$

alors

i) $$z(t_n) - z_n = \frac{1}{1 + \alpha_0 h_{k_n+1}}(z_n^{pr} - z_n) + o(\|z(t_n) - z_n\|)$$

ii) $$\|y(t_n) - y_n\| = O(\|z(t_n) - z_n\|)$$

(démonstration en annexe du chapitre III)

Nous voyons que le résultat est moins précis pour les variables algébriques que pour les variables différentielles. Nous ne disposons par pour l'instant de meilleurs résultats, ce qui est d'autant plus regrettable que les systèmes traités comportent beaucoup plus de variables algébriques que de variables différentielles.

. Van Bokhoven [Van Bok, 1975] a montré que l'erreur de troncature était proportionnelle à

$$K_{k_n}(h) = \frac{\prod_{i=1}^{k_n} h_i}{\sum_{i=1}^{k_n} \frac{1}{h_i}}$$

ou encore à

$$K_{k_n}(h) = \frac{h \prod_{i=2}^{k_n} (h + h'_{i-1})}{\frac{1}{h} + \sum_{i=2}^{k_n} \frac{1}{(h + h'_{i-1})}}$$

De plus, si nous notons $E^j_{k_n}$ l'erreur de troncature faite sur la jème composante différentielle au temps t_n par une méthode d'ordre k_n, nous avons :

$$E^j_{k_n+1} = \frac{\bar{z}^{k_n+2}_{n,j} - z_{n,j} - E^j_{k_n}}{h_{k_n+2} \left(\sum_{i=1}^{k_n+1} \frac{1}{h_i} \right)}$$

$$E^j_{k_n-1} = \frac{\bar{z}^{k_n}_{n,j} - z_{n,j} - E^j_{k_n}}{h_{k_n} \left(\sum_{i=1}^{k_n-1} \frac{1}{h_i} \right)}$$

III)CHOIX DU PAS ET DE L'ORDRE

La stratégie suivie pour le choix du pas et de l'ordre va être la suivante :

Supposons donnée une valeur "objectif" α et une valeur "limite" β.

Supposons que nous ayons calculé au temps t_n la valeur z_n grâce à une méthode d'ordre k_n, avec un pas h.

début STRATEGIE

pour m allant de k_n-1 à k_n+1

début

pour j allant de 1 à p faire

début

calculer E_m^j

rejeter le pas si $m = k_n$ et $\|E_m^j\| > \beta$

fin pour

poser $E_m^* = \max_{j=1,\ldots p} \|E_m^j\|$

trouver h_m^* tel que $\frac{K_m(h_m^*)}{K_m(h)} = \frac{\alpha}{E_m^*}$

fin pour

trouver m* tel que $h_{m^*}^* = \max_{m=k_n-1,k_n,k_n+1} h^*$

le nouveau pas sera $h_{m^*}^*$, le nouvel ordre sera m*

fin STRATEGIE

Il faut remarquer que la stratégie habituellement cherche le $\max_{m=k_n-1,k_n,k_n+1} (\min_{j=1,\ldots p} h_m^j)$

ou h_m^j est tel que $\frac{K_m(h_m^j)}{K_m(h)} = \frac{\alpha}{E_m^j}$

Mais comme K_m est croissante - voir la démonstration de la proposition 2, notre démarche est équivalente à cette dernière et plus économe en opérations arithmétiques.

Une deuxième remarque est que le calcul de troncature n'est fait que pour les variables différentielles, en vertu de la proposition 1.

IV) DIVERS PROBLEMES NUMERIQUES LIES AU CALCUL DU PAS ET DE L'ORDRE

Le problème qui se pose à nous maintenant est celui de la résolution, h et λ étant donnés, de

$$(E) \qquad \frac{K_m(hx)}{K_m(h)} = \alpha$$

proposition 2

Soit $\alpha > 0$

Soit $m > 2$

Supposons qu'il existe h_x^s solution de (E) dans l'intervalle réel $[h_{min}, h_{max}]$.

Alors :

1) h_x^s est unique

ii) la suite des itérés $h_x^{p+1} = G^p(hx)$ déterminé par h_x^0 choisi quelconque dans $[h_{min}, h_{max}]$ et

$$G(hx) = hx - \frac{K_m(h)h_{max}}{2m\, K_m(h_{max})} \left(\frac{K_m(hx)}{K_m(h)} - \alpha \right)$$

(méthode du point fixe)

est convergente et sa limite hx* est telle que $\frac{K_m(hx^*)}{K_m(h)} = \alpha$

(démonstration en annexe du chapitre III)

Pratiquement, il est préférable d'utiliser une variante de la méthode du point fixe qui est la méthode de Steffesen

$$h_x^{p+1} = \frac{G(G(h_x^p)).h_x^p - [G(h_x^p)]^2}{G(G(hx^p)) - 2G(h_x^p) + h_x^p}$$

Une deuxième méthode utilisable est de résoudre (E) par une méthode de Newton-Raphson

proposition 3

Soit $\alpha \geq 0$

Soit $m > 2$.

Supposons qu'il existe hx^s solution de (E) dans l'intervalle réel $[h_{min}, h_{max}]$

alors la suite $hx^{p+1} = G(hx^p)$ déterminée par hx^0 choisie quelconque dans $[h_{min}, h_{max}]$, et

$$G(hx) = hx - \frac{K_m'(hx)}{K_m(h)\left(\frac{K_m(hx)}{K_m(h)} - \alpha\right)}$$

est convergente et sa limite hx* est telle que

$$\frac{K_m(hx^*)}{K_m(h)} = \alpha$$

(démonstration en annexe du chapitre III))

Nous avons donc à notre disposition deux méthodes quadratiques dont la convergence est assurée, pour résoudre (E) lorsque m est strictement supérieur à 1

Pour m = 1, (E) se réduit à $\frac{hx^2}{h^2} - \alpha = 0$

ce qui nous donne immédiatement $hx^* = \sqrt{\alpha}\, h$

Il est préjudiciable pour la stabilité du schéma numérique de faire varier de manière trop brusque la taille du pas, notamment lorsqu'on augmente le pas. Concrétement, si hx s'avère être supérieur à 2h (respectivement inférieur à h/4), nous prendrons hx = 2h (resp. hx = h/4).

De plus, si α est supérieur(resp. inférieur à 1), hx sera supérieur(resp. inférieur) à h. Cela nous amène à poser $h_{max} = 2h$ et $h_{min} = h$ (resp. $h_{max} = h$ et $h_{min} = h/4$) lorsque α est supérieur(resp. inférieur) à 1.

Puisque nous savons que notre problème admet une solution unique dans l'intervalle $[h_{min}, h_{max}]$, nous pouvons sans problèmes utiliser un algorithme de type "regula-falsi" ou "dichotomique" [Hac, 1971, p 136].

Parmi ces types de méthode, la plus efficace est sans conteste la méthode dite de "section dorée" qui converge en N pas si nous posons N le plus petit entier tel que $\lambda^N.\varepsilon$ soit supérieur à $(h_{max} - h_{min})$, où λ est le "nombre d'or" ($= \frac{-1 + \sqrt{5}}{2}$) et ε la précision demandée sur hx [Kar, 1975].

Nous allons comparer ces trois méthodes entre elles (point fixe, Newton-Raphson, section dorée) du point de vue du temps d'unité centrale (temps C.P.U.) utilisé par chacune d'elles. C'est un critère important, car la recherche de la solution de (E) intervient 3 fois par pas (pour $m = k_n - 1$, k_n et $k_n + 1$), ce qui représente en tout plusieurs centaines de résolutions, voire plusieurs milliers lorsqu'un algorithme découplage de blocs (cf. chapitre II, VI)) est utilisé.

Pour comparer deux méthodes entre elles, nous allons tirer au sort la valeur de α, des h'_1 et de h, puis nous évaluerons le temps mis par les deux méthodes pour résoudre $\frac{K_m(hx)}{K_m(h)} = \alpha$ en faisant varier m de 2 à 6.

Pour ce faire, nous utilisons la procédure X05GAF de la bibliothèque NAG [NAG] qui fournit des nombres aléatoires compris entre 0 et 1

Pour évaluer le temps C.P.U. utilisé par chacune des méthodes, nous utilisons deux procédures MULTICS [Mult, 1980] setchrono et vtime qui donne le temps C.P.U. écoulé depuis le dernier appel à setchrono.

Les résultats sont les suivants :

1ère comparaison

ordre m	2	3	4	5	6
méthodes					
section dorée	2.1133	2.7358	3.4020	3.9268	4.5547
point fixe	1.2640	1.5979	1.6189	1.6961	1.7143

2ème comparaison

ordre m	2	3	4	5	6
méthodes					
point fixe	1.2268	1.5196	1.5432	1.6991	1.7370
Newton Raphson	1.7006	2.0027	2.4407	2.9370	3.2094

tous les temps indiqués sont en milli-secondes C.P.U.

Les méthodes ont été testées à chaque comparaison sur un échantillon de 100 valeurs.

Comme il était prévisible, les méthodes "quadratiques" sont plus rapides que les méthodes de section dorée. Il est remarquable que la méthode de point fixe soit notablement plus efficace que la méthode de Newton-Raphson. Nous retiendrons donc à l'avenir la méthode de point fixe-Steffesen pour le calcul du pas.

Nous avons vu précédemment que la discrétisation du problème

$$\begin{cases} f(y, z, \dot{z}, t) = \theta_{q+p} \\ z(t_0) = \mu \end{cases}$$

générait une suite de problèmes algébriques à résoudre

$(S_n)_{n=0,1,\ldots}$ et qu'ils étaient résolus par une méthode de Newton ou une méthode dérivée.

Une question que nous devons aborder est celui des variables d'arrêt.

Les variables issues de la modélisation de circuits électroniques sont d'ordre de grandeur disparates.

Il est donc souhaitable que la précision demandée pour décider de l'arrêt de la suite d'itérés engendrée par la méthode de Newton soit une **précision relative**, tout en veillant à prendre en compte le cas où la variable testée est nulle.

Aussi, si nous appelons $(y_n^k)_{k=0,\ldots}$ et $(z_n^k)_{k=0,\ldots}$ les deux suites d'itérés engendrées, le test d'arrêt sera de la forme :

$$\frac{\|y_n^{k+1} - y_n^k\|}{test + \|y_{n-1}\|} \leqslant \varepsilon_n$$

$$\frac{\|z_n^{k+1} - z_n^k\|}{test + \|z_{n-1}\|} \leqslant \varepsilon_n$$

où ε_n est la précision demandée au nième pas et test un réel strictement positif.

test doit être de l'ordre de grandeur de ce qui sera compté comme négligeable devant la variable, en quelque sorte du "bruit" toléré. Son choix dépendra donc de la précision demandée et du comportement présumé de la variable. Il est à remarquer que pour $z_{n-1}=0$ (ou $y_{n-1}=0$) la précision demandée est absolue et égale à test.ε_n.

Il est important de faire le test relatif par rapport au pas antérieur, c'est à dire par rapport à une valeur déjà calculée, ceci afin d'éviter que des divergences dans les itérées de Newton soient interprétées comme une réussite au test.

En effet le test $\frac{\|z_n^{k+1} - z_n^k\|}{test + \|z_n\|} \leqslant \varepsilon_n$ est satisfait lorsque z_n^{k+1} tends vers l'infini, alors que la distance entre

deux itérés successifs reste finie.

Au vu de la démonstration de la proposition 1 sur l'erreur de troncature,qui suppose que les valeurs (y_n, z_n) sont **exactement** celles satisfaisant à "$F_n(y_n, z_n) = \theta_{q+p}$", il est souhaitable d'avoir un ε_n faible devant l'erreur de troncature. Pour que les comparaisons soient homogènes, il est donc indiqué de calculer également des errreurs de troncature relatives,et de se fixer de valeurs "objectif" α et "limite" β relatives, dans la stratégie de choix de pas et d'ordre définie précédemment.

Cela se fait sans mal en remarquant que

$$\frac{\alpha}{E_m^j} = \frac{\dfrac{\alpha}{\text{test} + \|z_{n,j}\|}}{\dfrac{E_m^j}{\text{test} + \|z_{n,j}\|}} = \frac{\alpha^*}{E_m^{j*}}$$

si nous appelons α^* la valeur objectif relative choisie, et E_m^{j*} l'erreur de troncature relative faite sur la jème variable différentielle.

De plus,surtout si nous autorisons le pas à varier de manière conséquente (dans les exemples traités plus loin, le rapport entre la plus grande valeur du pas et la plus petite valeur prise au cours de l'intégration peut atteindre plusieurs millions), il est nécessaire que la précision demandée tienne compte de la taille du pas.

Pour toutes ces raisons, nous avons été amenés à utiliser pour les erreurs et tests d'arrêts l'ensemble des règles suivantes :

Soit $[t_{min}, t_{max}]$ l'intervalle d'intégration.

Soit ε la précision relative demandée pour un pas standard ($h_{stand} = \dfrac{t_{max} - t_{min}}{100}$).

Supposons que nous voulions résoudre S_n.

Alors nous prendrons pour

- la valeur objectif la valeur $\alpha^* = \varepsilon.(t_n - t_{n-1})/h_{stand}$
- la valeur limite la valeur $\beta^* = 10\ \alpha^*$
- le test d'arrêt pour les itérations de Newton $\varepsilon_n = .1\ \alpha^*$

Ainsi que nous l'avons indiqué précédemment, pour des raisons de stabilité, nous autorisons entre deux résolutions successives le pas à croître seulement d'un facteur 2 au plus et à décroître d'un facteur 4 au plus. C'est à dire

$$h_{n+1} \in [h_n, 2\ h_n] \text{ lorsqu'il y a croissance}$$
$$h_{n+1} \in [h_n, 2\ h_n] \text{ lorsqu'il ya décroissance}$$

De plus, nous limitons supérieurement la taille du pas par une valeur h_{max} (égale au pas standard en général), ceci afin d'avoir un nombre minimal de points pour la solution et d'assurer la précision des résultats, et inférieurement par une valeur h_{min} (fixée par le rapport donné en entrée h_{min}/h_{max} dans le logiciel de résolution (cf. VI)) ,pour ne pas se trouver en présence de schémas numériques qui n'avancent pas.

Enfin,pratiquement, la majeure partie du temps d'intégration l'erreur de troncature est très faible et la résolution de l'équation

$$\frac{K_m(hx)}{K_m(h)} = \frac{\alpha^*}{E_m^*}$$

conduit à des pas h_{n+1} bien plus grands que $2.h_n$. Ainsi, pour éviter des résolutions inutiles, lorsque E_m^{j*} est trop faible(inférieur à $.1\ \alpha^*$), nous doublons le pas systématiquement. De même, si E_m^{j*} est supérieur à β^* ou si le nombre d'itérations de Newton est trop élevé, nous rejetons le pas et le divisons par 4.

Limiter le nombre d'itérations de Newton est, on le concoit aisément, nécessaire pour éviter d'éventuelles non-convergences qui feraient "boucler" le programme de résolution. Mais de plus, ce nombre limite ne doit pas être très élevé, en pratique de l'ordre de 4 ou 5. En effet, même si au-delà les itérations convergeaient, le résultat final serait éloigné du prédicteur, et par la force des choses, nous serions en présence d'une erreur de troncature élevée.

V)RESOLUTION DE PROBLEMES TESTS

Il existe dans la littérature des problèmes d'équations différentielles ordinaires "raides" dont la résolution s'avère délicate, et qui servent à tester les méthodes d'intégration. Nous devons à [Bju, 1970, section 4, p 1 et suivantes] l'ensemble des problèmes test moins le dernier et des commentaires sur leur origine. Nous allons, grâce à ceux-ci, "tester" l'ensemble de notre démarche, en demandant pour chacun des problèmes suivants une précision relative par pas standard de $\varepsilon=.01$, un pas maximal égal au centième de l'intervalle d'intégration et un rapport h_{min}/h_{max} égal à 10^{-6}

problème n°1

$$\begin{cases} \dot{z}_1 + 2000\ z_1 - 1000\ z_2 - 1 = 0 \\ \dot{z}_2 - z_1 + z_2 = 0 \\ t \in [0,4] \\ z_1(0) = 0 \\ z_2(0) = 0 \end{cases}$$

Ce problème dû à Warten et Fowler, est un problème linéaire à coefficients constants. Les valeurs propres de

$$-\frac{\partial f}{\partial \dot{yz}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t) = \begin{bmatrix} -2000 & 1000 \\ 1 & -1 \end{bmatrix}$$

sont égales à $\lambda_1 \simeq -2000.5$ et $\lambda_2 \simeq -0.5$.

Si nous regardons l'allure des solutions ,nous observons que z_1 a une très forte irrégularité à l'origine puis qu'elle se stabilise.

Dans les pages qui suivent, nous entendons par intervalle d'intégration non pas l'intervalle total, mais un sous-intervalle de ce dernier (égal en taille au dixième de celui-ci). Tous les résultats présentés concernent exclusivement ce sous-intervalle.

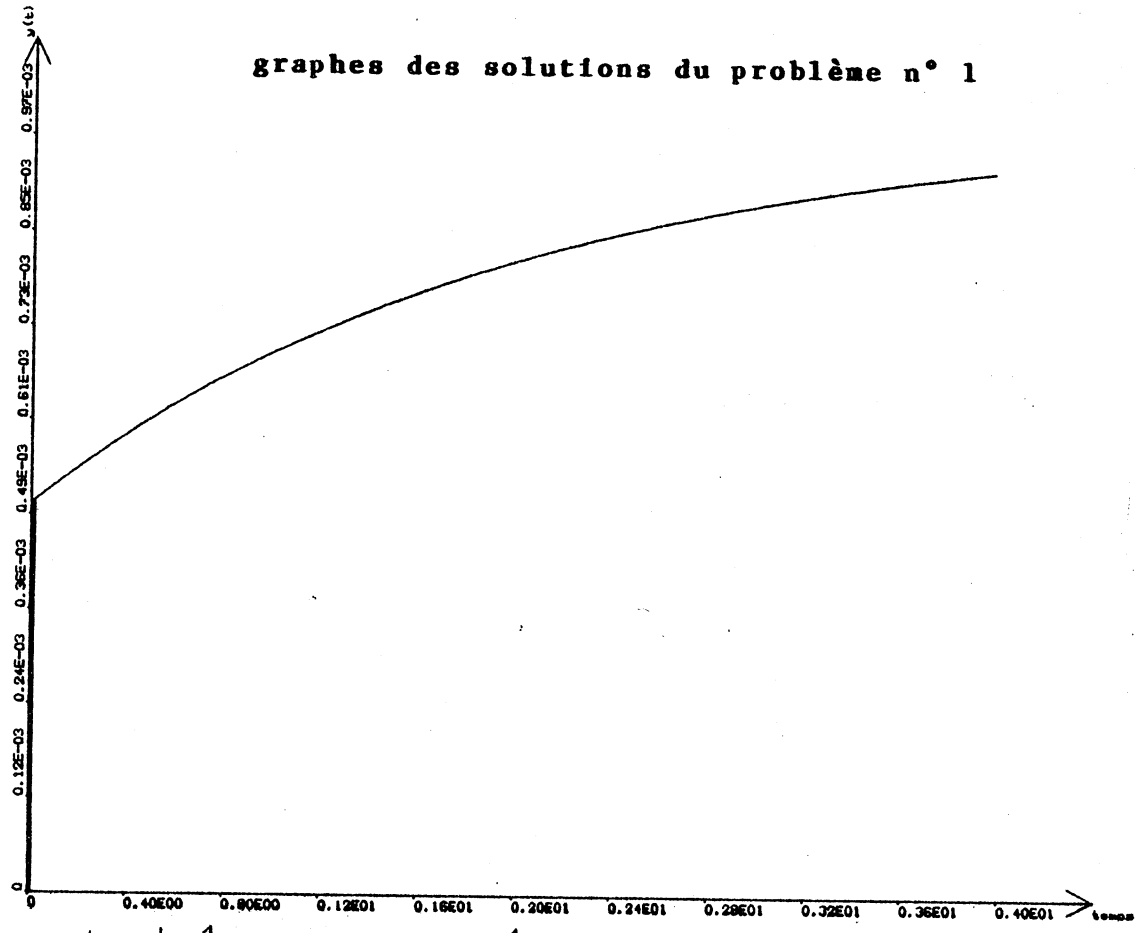
Résultats

intervalle d' intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 0.4]	5	3	35	6
]0.4, 0.8]	1	1	10	0
]0.8, 1.2]	1	1	10	0
]1.2, 1.6]	1	1	10	0
]1.6, 2.0]	1	1	10	0
]2.0, 2.4]	1	1	10	0
]2.4, 2.8]	1	1	10	0
]2.8, 3.2]	1	1	10	0
]3.2, 3.6]	1	1	10	0
]3.6, 4.0]	1	1	10	0

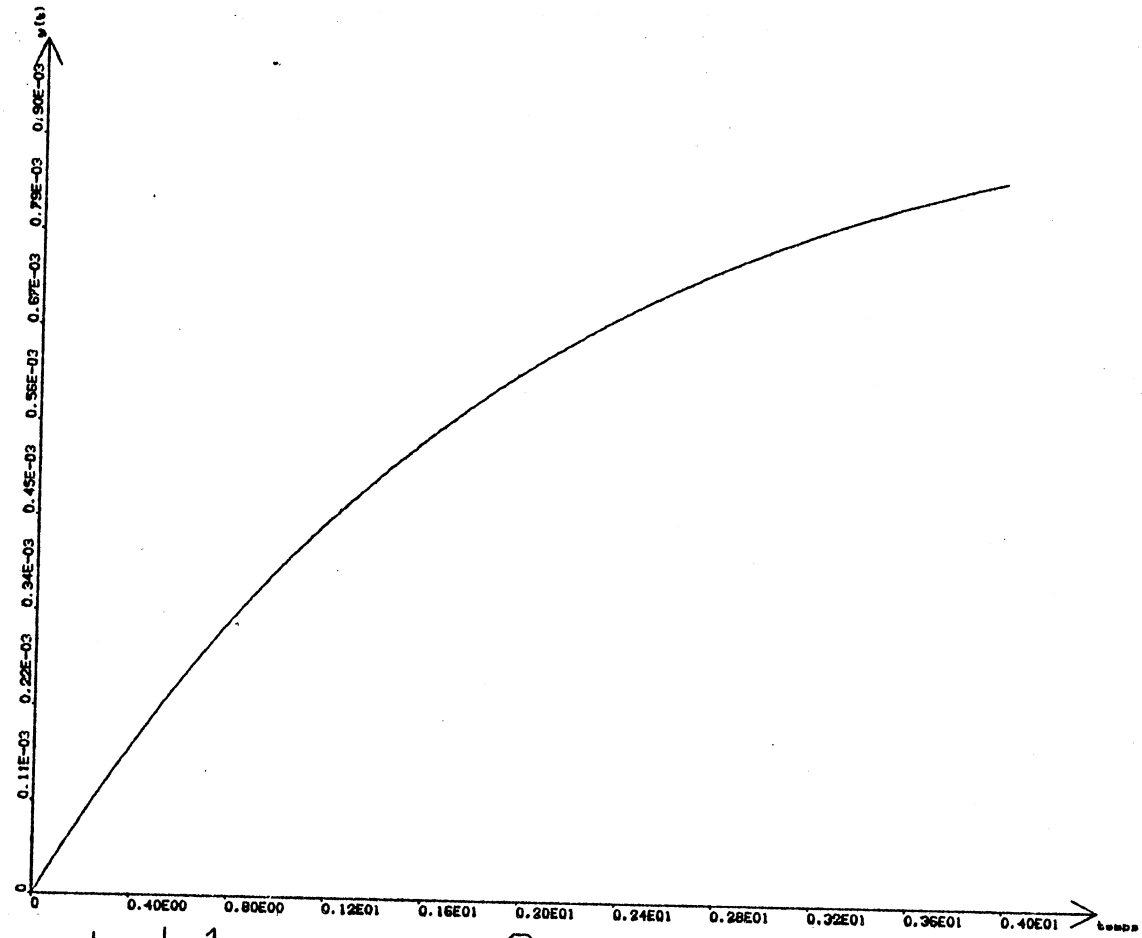
nombre de pas acceptés : 125

nombre de pas refusés : 6

graphes des solutions du problème n° 1



variable no 1



variable no 2

problème n° 2

$$\begin{cases} \dot{z}_1 - 3000\ (1 - z_1^2) + 10\ z_2 = 0 \\ \dot{z}_2 - 0.04\ (1 - z_2) - (1 - z_1)\ z_2 - 10^4\ (1 - z_1)^2 = 0 \\ t \in [0,\ 2.6] \\ z_1(0) = 1 \\ z_2(0) = 0 \end{cases}$$

Ce problème est également dû à Warten et Fowler. Il est non-linéaire, mais les valeurs propres ne ne bougent pas beaucoup si $z_1(0)$ est choisi pres de la solution stationnaire.

Les valeurs propres de

$$-\frac{\partial f}{\partial y\dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t) = \begin{bmatrix} -5000\ z1 & -10 \\ -z_2 - 2\ 10^4\ z_1 & -0.04 - (1 - z_1) \end{bmatrix}$$

sont réelles, négatives et bien séparées.

$t = 0$	$\lambda_1 \simeq -6000$	$\lambda_2 \simeq -0.04$
$t = 2.6$	$\lambda_2 \simeq -6000$	$\lambda_2 \simeq -0.04$

Si nous regardons les solutions, celles-ci apparaissent comme très lisses : z_1 est stationnaire et z_2 est linéaire .

Ce n'est donc pas l'allure de ces dernières qui pose difficulté. Le problème réside dans le fait que les erreurs sont amplifiées au départ de manière notable, vu la valeur du rapport des valeurs propres.

Celles-ci étant négatives et plus ou moins constantes, l'étude faite au chapitre I sur les systèmes fortement stables s'applique , et la solution perturbée doit donc rester a une distance finie de la solution exacte
(si nous prenons comme distance $\max_{t \in [t_0,T]} \|z(t) - z_\varepsilon(t)\|$)

Cette distance tend vers zéro en même temps que la taille des perturbations.

Mais cela n'implique pas, nous l'avons vu au chapitre I, que la solution numérique puisse elle être rendue aussi proche que l'on veut de la solution exacte. Cela dépend fortement de la stabilité des méthodes utilisées.

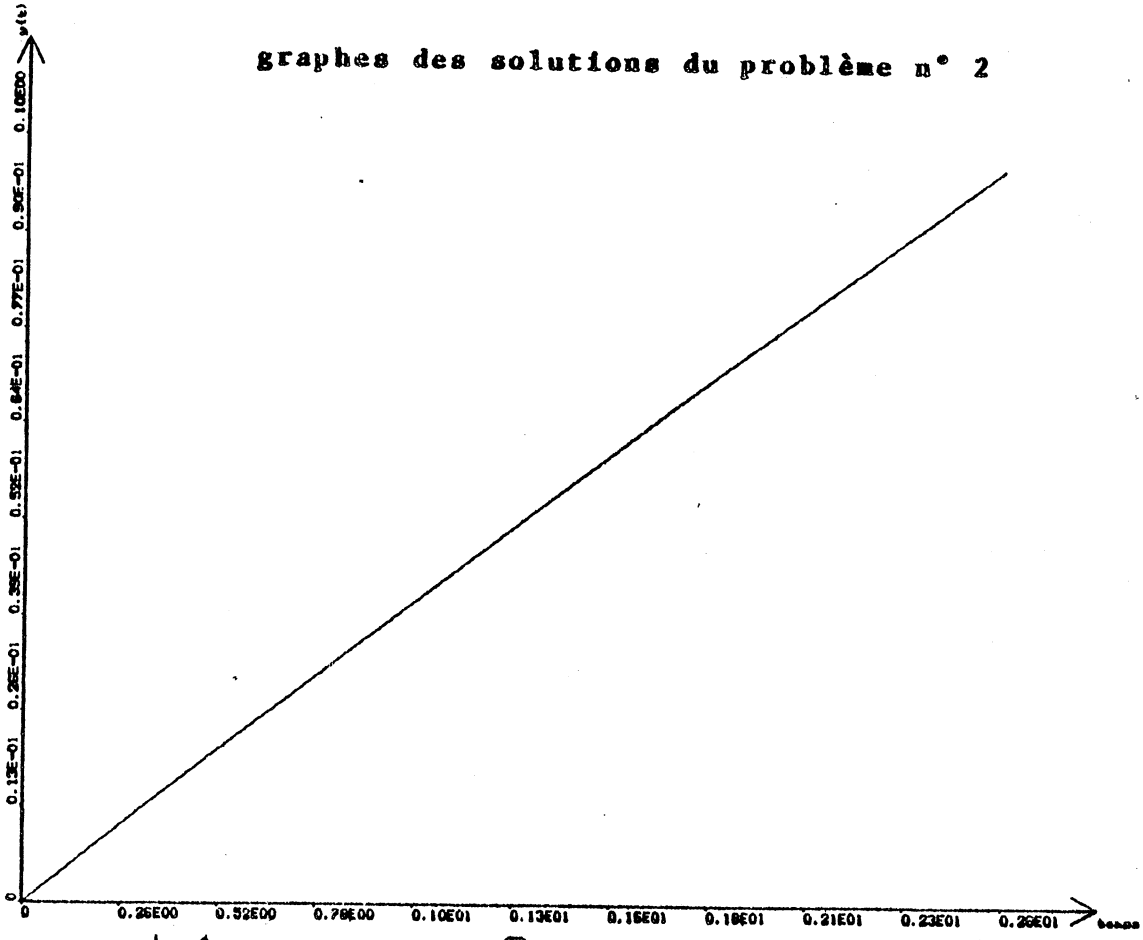
Résultats

intervalle d' integration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 0.26]	1	1	11	0
]0.26, 0.52]	1	1	10	0
]0.52, 0.78]	1	1	10	0
]0.78, 1.04]	1	1	10	0
]1.04, 1.30]	1	1	10	0
]1.30, 1.56]	1	1	10	0
]1.56, 1.82]	1	1	10	0
]1.82, 2.08]	1	1	10	0
]2.08, 2.34]	1	1	10	0
]2.34, 2.60]	1	1	10	0

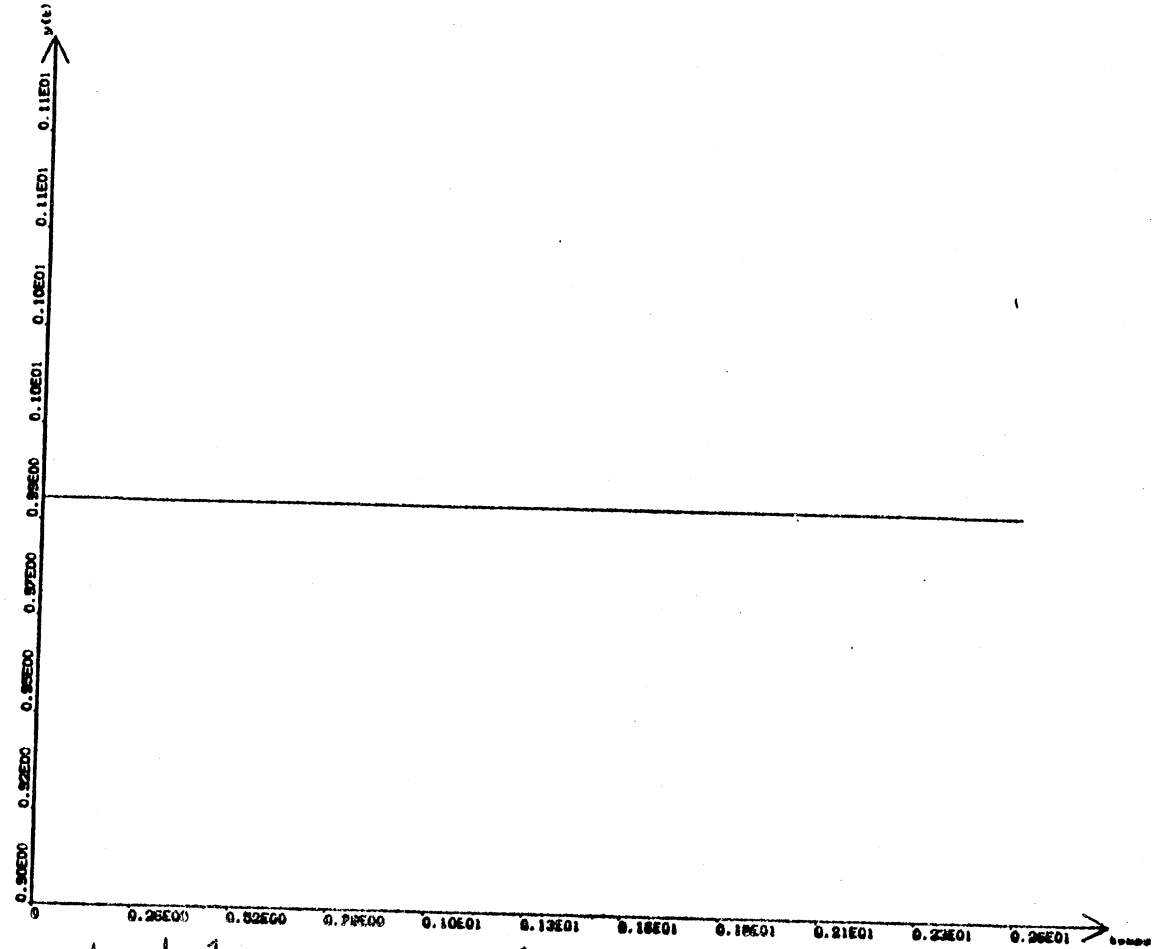
nombre de pas acceptés : 101

nombre de pas refusés : 0

graphes des solutions du problème n° 2



variable no 2



variable no 1

problème n° 3

$$\begin{cases} \dot{z}_1 - 0.2\, z_2 + (60 - 0.125\, t)\, z_1 - 0.125\, t = 0 \\ \dot{z}_2 - 0.2\, z_1 + 0.2\, z_2 = 0 \\ t \in [0,\ 400] \\ z_1(0) = 0 \\ z_2(0) = 0 \end{cases}$$

Ce problème, utilisé par Liniger et Willoughby, est lié au mouvement d'une baguette de contrôle dans un réacteur nucléaire. C'est un problème linéaire à coefficients variables.

Les valeurs propres de $\begin{bmatrix} -(60 - 0.125\, t) & 0.2 \\ 0.2 & -0.2 \end{bmatrix}$

varient avec le temps, et leur rapport est grand :

$t = 0$	$\lambda_1 = -60.0$	$\lambda_2 = -0.17$
$t = 320$	$\lambda_1 = -20.1$	$\lambda_2 = -0.10$
$t = 400$	$\lambda_1 = -10.2$	$\lambda_2 = 0$

les valeurs propres de $-\frac{\partial f}{\partial \dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t)$ sont négatives, réelles et leur rapport est élevé à l'origine. Le fait que leur rapport soit très grand pour $t = 400$ n'est pas important car l'importance des perturbations diminue considérablement avec le temps lorsque les valeurs propres sont très négatives.

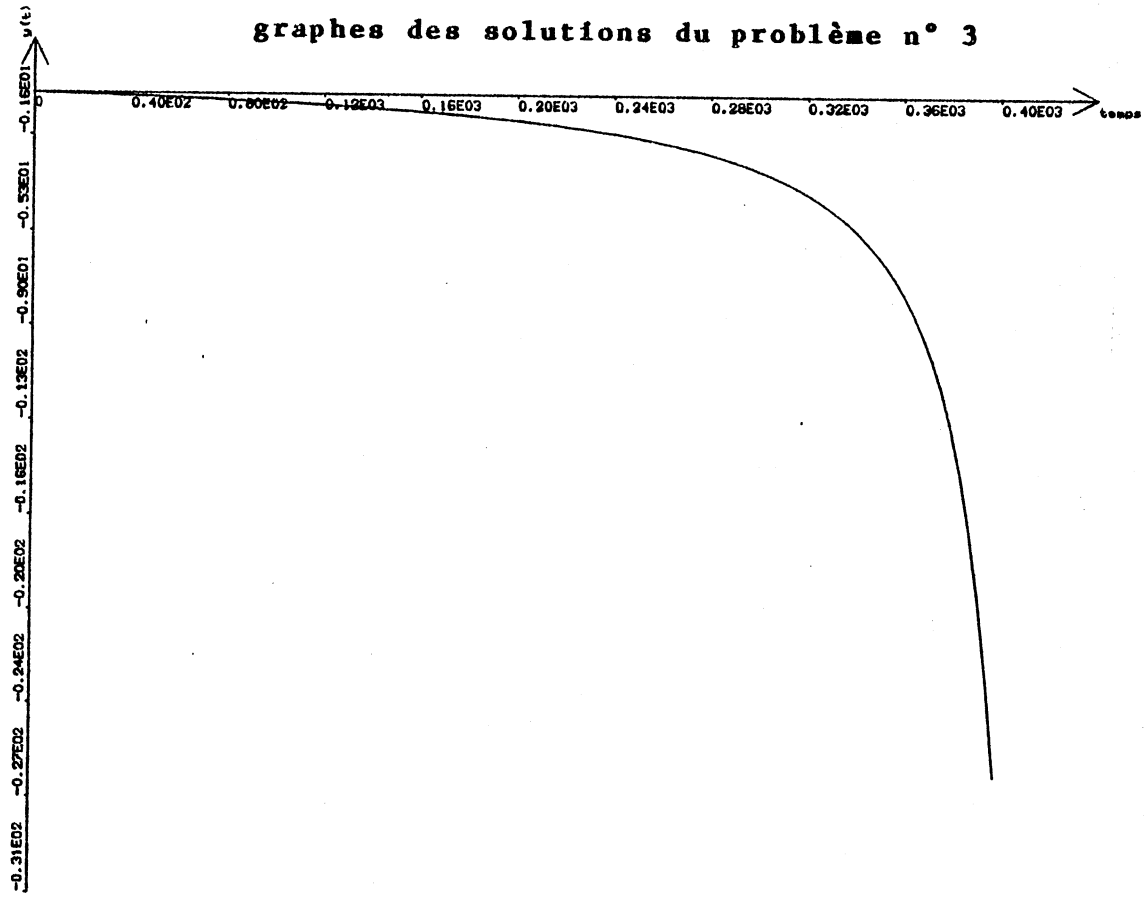
Résultats

intervalle d' intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 40]	6	1	41	8
]40, 80]	2	1	10	0
]80, 120]	1	1	10	0
]120, 160]	1	1	10	0
]160, 200]	1	1	10	0
]200, 240]	1	1	10	0
]240, 280]	1	1	10	0
]280, 320]	1	1	10	0
]320, 360]	1	1	10	0
]360, 400]	1	1	10	0

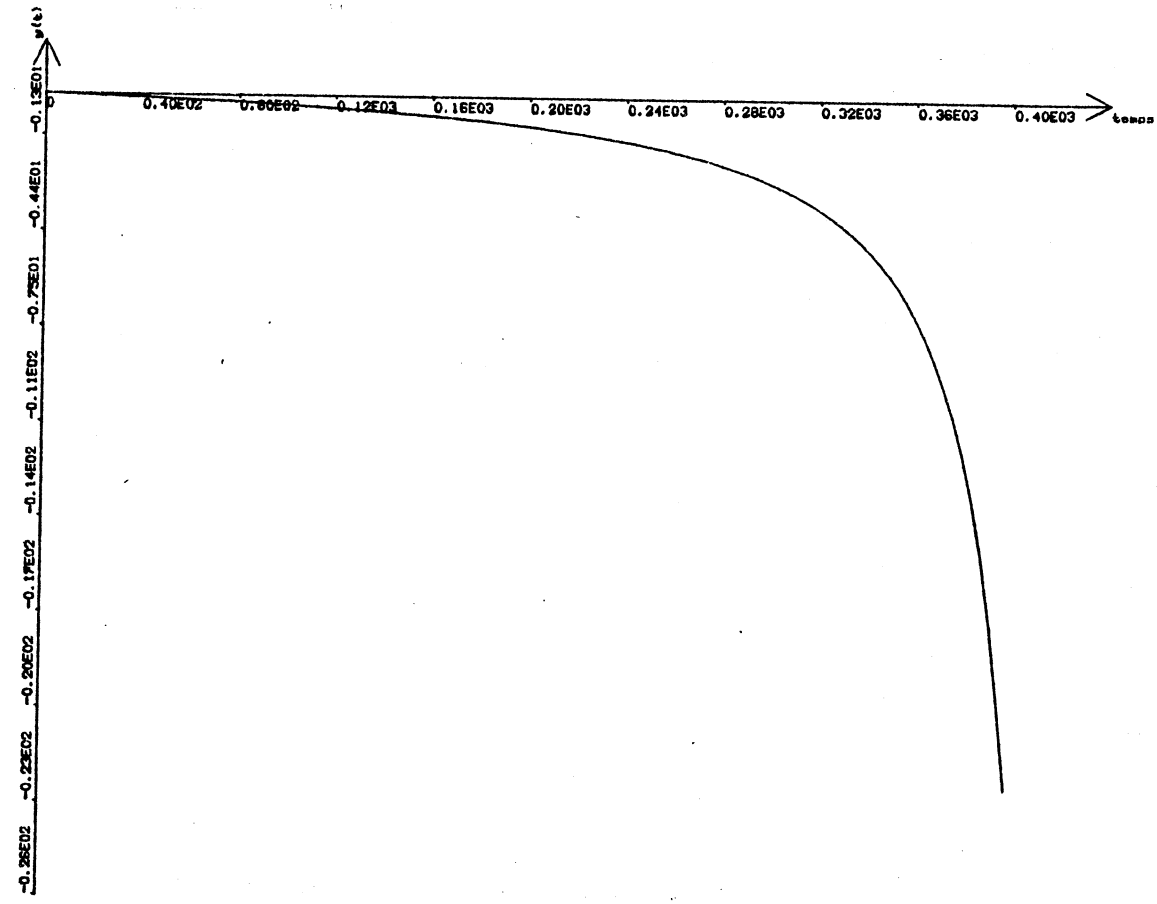
nombre de pas acceptés : 131

nombre de pas refusés : 8

graphes des solutions du problème n° 3



variable no 1



variable no 2

problème n° 4

$$\begin{cases} \dot{z}_1 + [1 + (z_1 + 1000)(z_1 + 1)](0.01 + z_1 + z_2) - 0.01 = 0 \\ \dot{z}_2 + (1 + \dot{z}_2)(0.01 + z_1 + z_2) - 0.01 = 0 \\ t \in [0,100] \\ z_1(0) = 0 \\ z_2(0) = 0 \end{cases}$$

C'est un problème de cinétique chimique décrit par un système différentiel non-linéaire. Comme le précédent, il est dû à Liniger et Willoughby.

La matrice $= \frac{\partial f}{\partial \dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t)$ vaut

$$\begin{bmatrix} -3z_1^2 - 2z_1z_2 - 2002.02z_1 - 1001z_2 - 1011.01 & -z_1^2 - 1001z_1 - 1001 \\ -z_2^2 - 1 & -3z_2^2 - 2z_1z_2 - 0.02z_2 - 1 \end{bmatrix}$$

Ses valeurs propres sont négatives et réelles. Leur rapport est très grand à l'origine ($\lambda_1/\lambda_2 \simeq 10^5$)

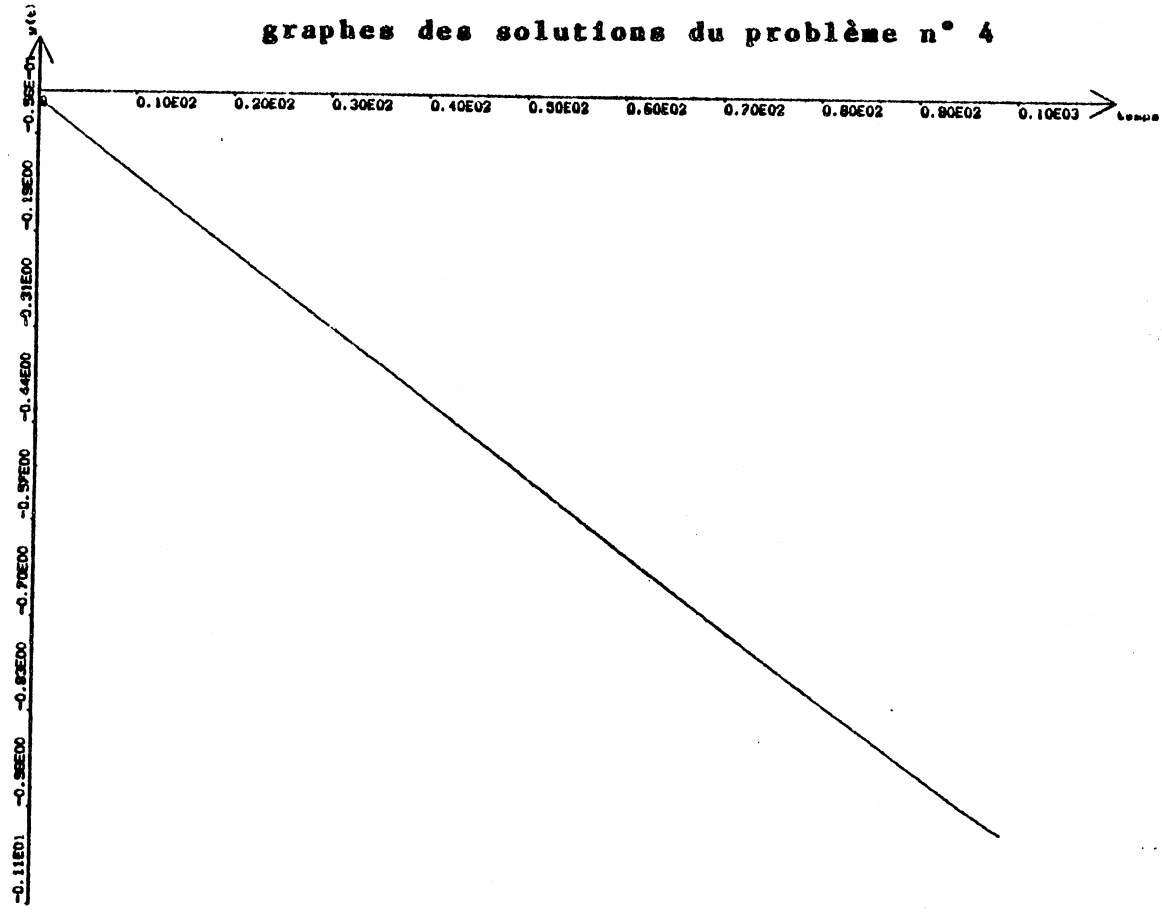
Résultats

intervalle d' intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 10]	6	1	62	12
]10, 20]	1	1	10	0
]20, 30]	1	1	10	0
]30, 40]	1	1	10	0
]40, 50]	1	1	10	0
]50, 60]	1	1	10	0
]60, 70]	1	1	10	0
]70, 80]	1	1	10	0
]80, 90]	1	1	10	0
]90, 100]	1	1	10	0

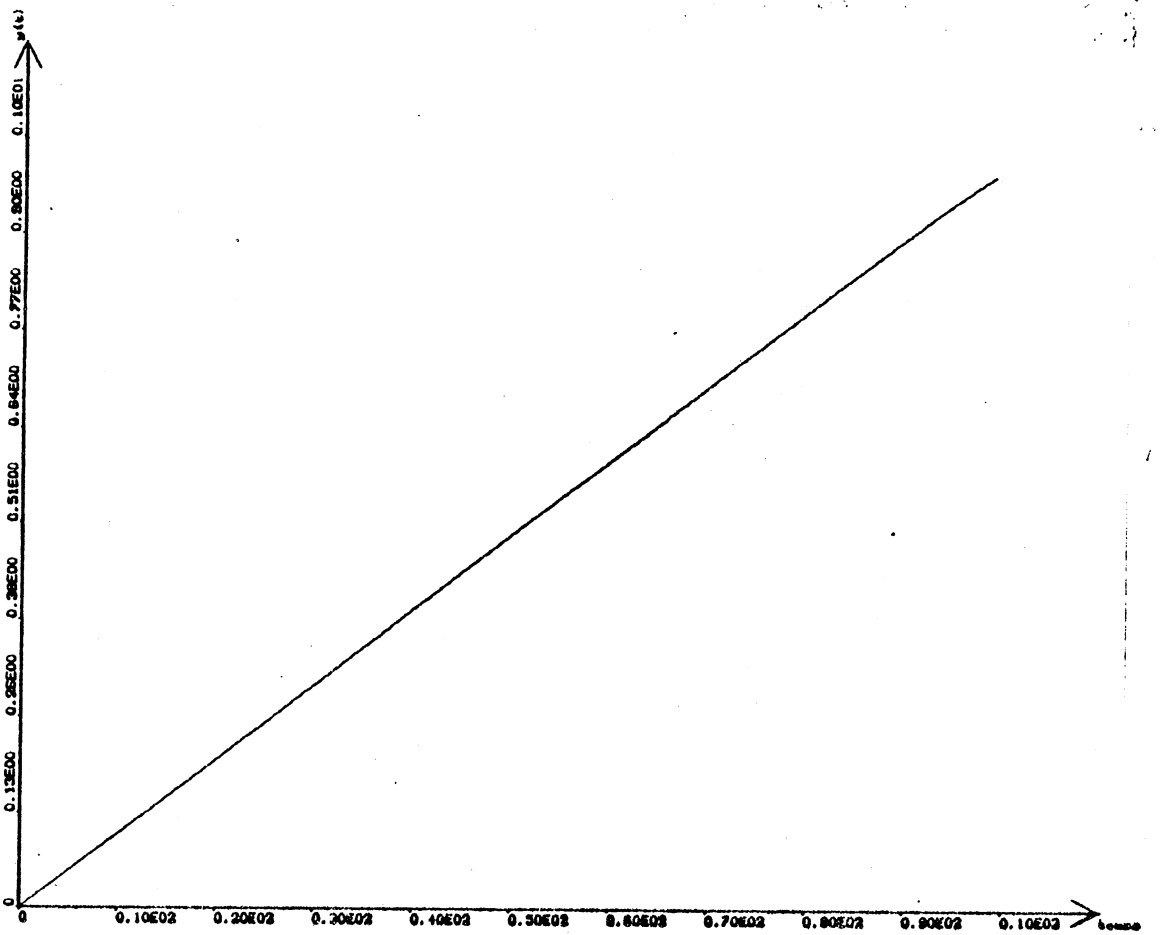
nombre de pas aceptés : 152

nombre de pas refusés: 12

graphes des solutions du problème n° 4



variable no 1



variable no 2

problème n°5 :

Ce problème proposé par Gear est dérivé d'un problème physique

$$\begin{cases} \dot{z}_1 + (55 + z_3)\, z_1 - 65\, z_2 = 0 \\ \dot{z}_2 - 0.0785(z_1 - z_2) = 0 \\ \dot{z}_3 - 0.1\, z_2 = 0 \\ t \in [0,\ 400] \\ z_1(0) = 1 \\ z_2(0) = 1 \\ z_3(0) = 0 \end{cases}$$

La matrice $-\frac{\partial f}{\partial \dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t)$ vaut

$$\begin{bmatrix} -55 + z_3 & 65 & -z_1 \\ 0.785 & -0.0785 & 0 \\ 0 & 0.1 & 0 \end{bmatrix}$$

$t = 0$ $\qquad \lambda_1 \simeq -55 \quad \lambda_{2,3} \simeq 0.0625 \pm 0.2\ i$

Ces valeurs propres correspondent à des oscillations très amorties.

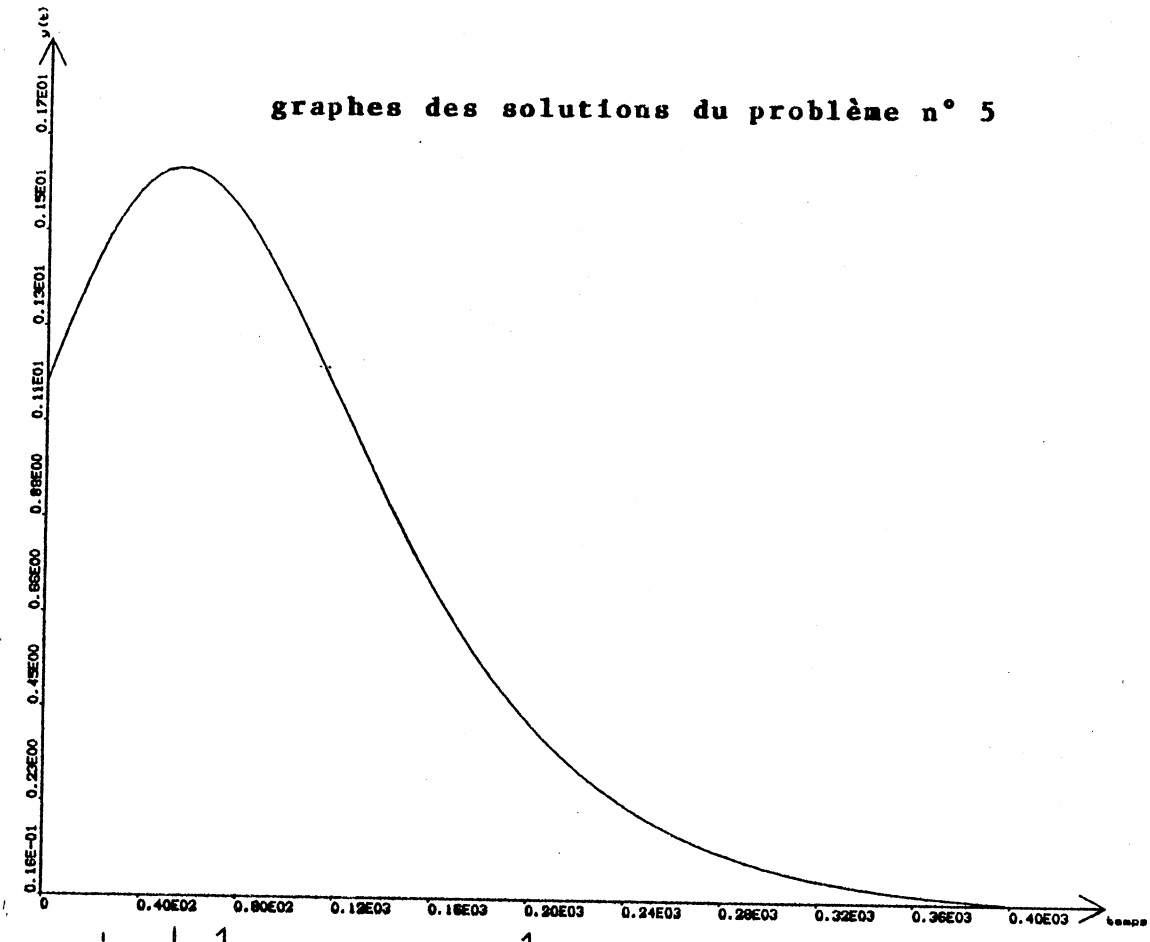
Résultats :

intervalle d° intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 40]	2	1	148	7
]40,80]	1	1	10	0
]80, 120]	1	1	10	0
]120, 160]	1	1	10	0
]160, 200]	1	1	10	0
]200, 240]	1	1	10	0
]240, 280]	1	1	10	0
]280, 320]	1	1	10	0
]320, 360]	1	1	10	0
]360, 400]	1	1	10	0

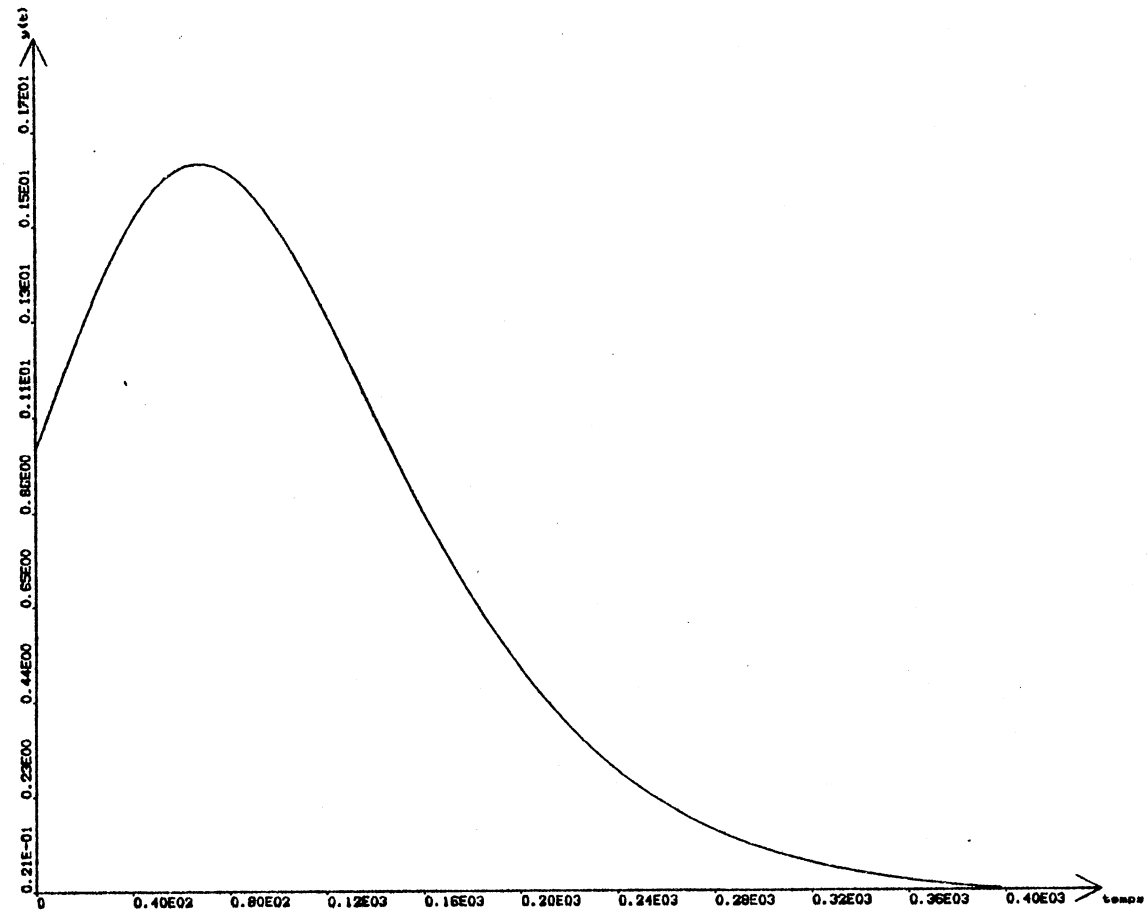
nombre de pas acceptés : 238

nombre de pas refusés : 7

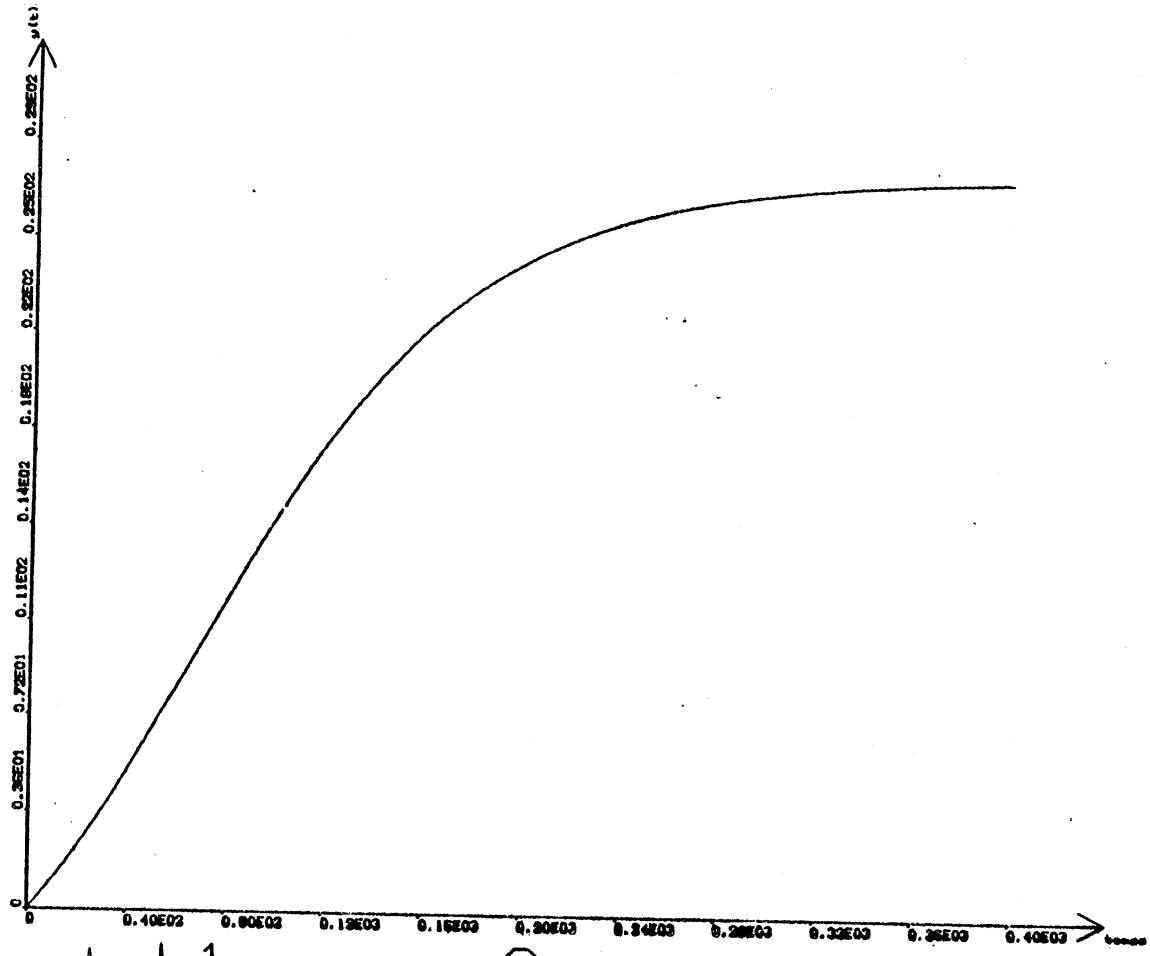
graphes des solutions du problème n° 5



variable no. 1



variable no 2



variable no 3

problème n° 6

Ce problème non-linéaire proposé par Gear est issu d'un problème de chimie et est considéré comme très raide.

$$\begin{cases} \dot{z}_1 + 0.013\ z_2 + 1000\ z_2\ z_1 + 2500\ z_3\ z_1 = 0 \\ \dot{z}_2 + 0.013\ z_2 + 1000\ z_2\ z_1 = 0 \\ \dot{z}_3 + 2500\ z_3\ z_1 = 0 \\ t \in [0,\ 50] \\ z_1(0) = 0 \\ z_2(0) = 1 \\ z_3(0) = 0 \end{cases}$$

La matrice $- \frac{\partial f}{\partial \dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t)$ vaut

$$\begin{bmatrix} -0.013 - 1000\ z_2 - 2000\ z_1 & -1000\ z_1 & -2500\ z_1 \\ -1000\ z_2 & -0.013\ z_1 & 0 \\ -2500\ z_2 & 0 & -2500\ z_1 \end{bmatrix}$$

$t = 0 \quad \lambda_1 \simeq -3500 \quad \lambda_2 \simeq 0.017 \quad \lambda_3 = 0$

La première variable présente une forte irrégularité à l'origine. Les deux autres variables sont stationnaires.

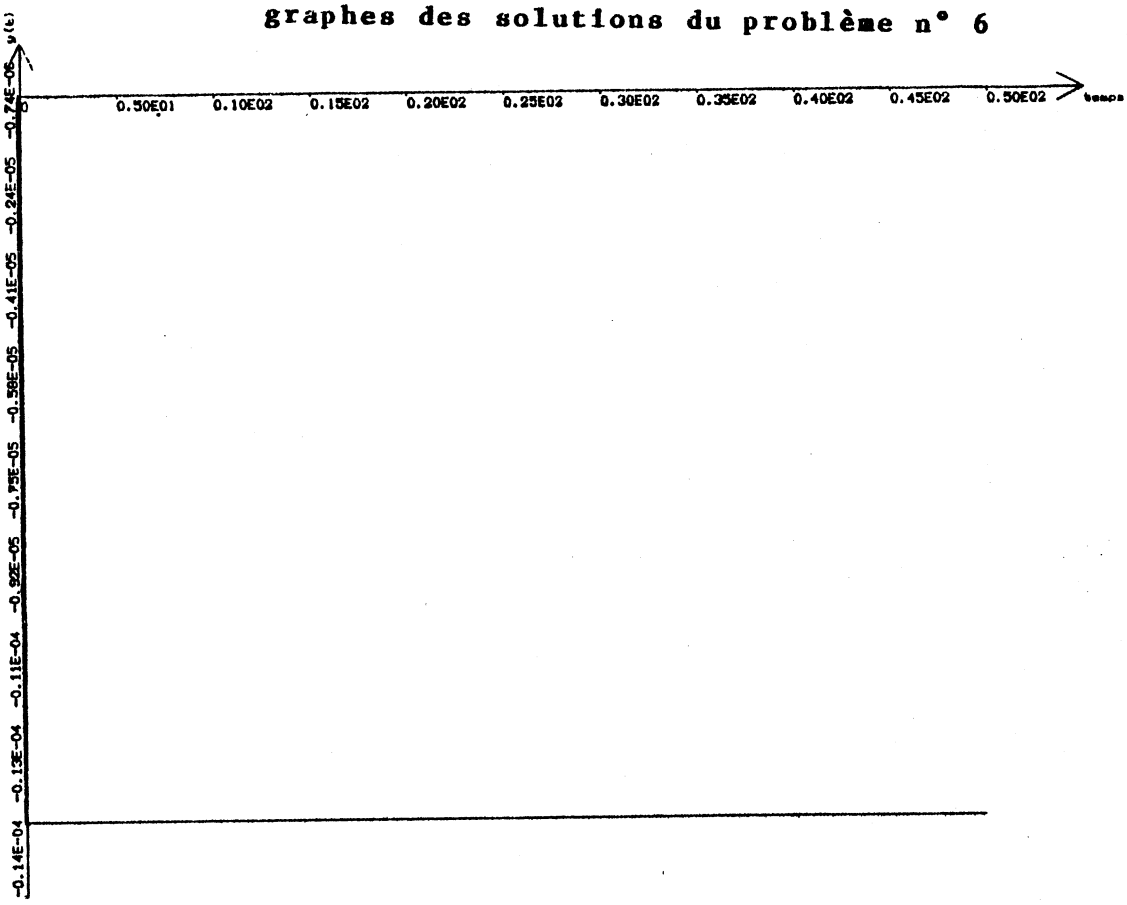
Résultats :

intervalle d' intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 5]	6	1	47	9
]5, 10]	1	1	10	0
]10, 15]	1	1	10	0
]15, 20]	1	1	10	0
]20, 25]	1	1	10	0
]25, 30]	1	1	10	0
]30, 35]	1	1	10	0
]35, 40]	1	1	10	0
]40, 45]	1	1	10	0
]45, 50]	1	1	10	0

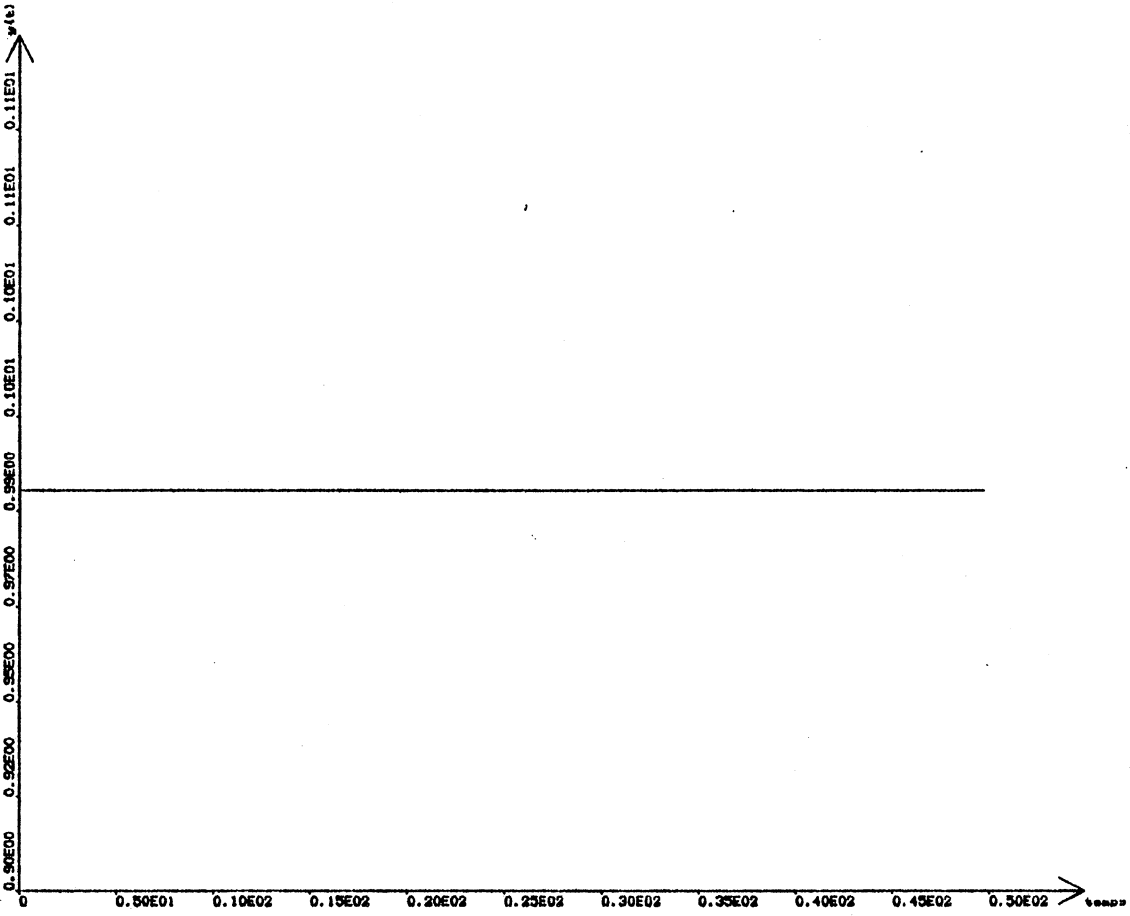
nombre de pas acceptés : 137

nombre de pas refusés : 9

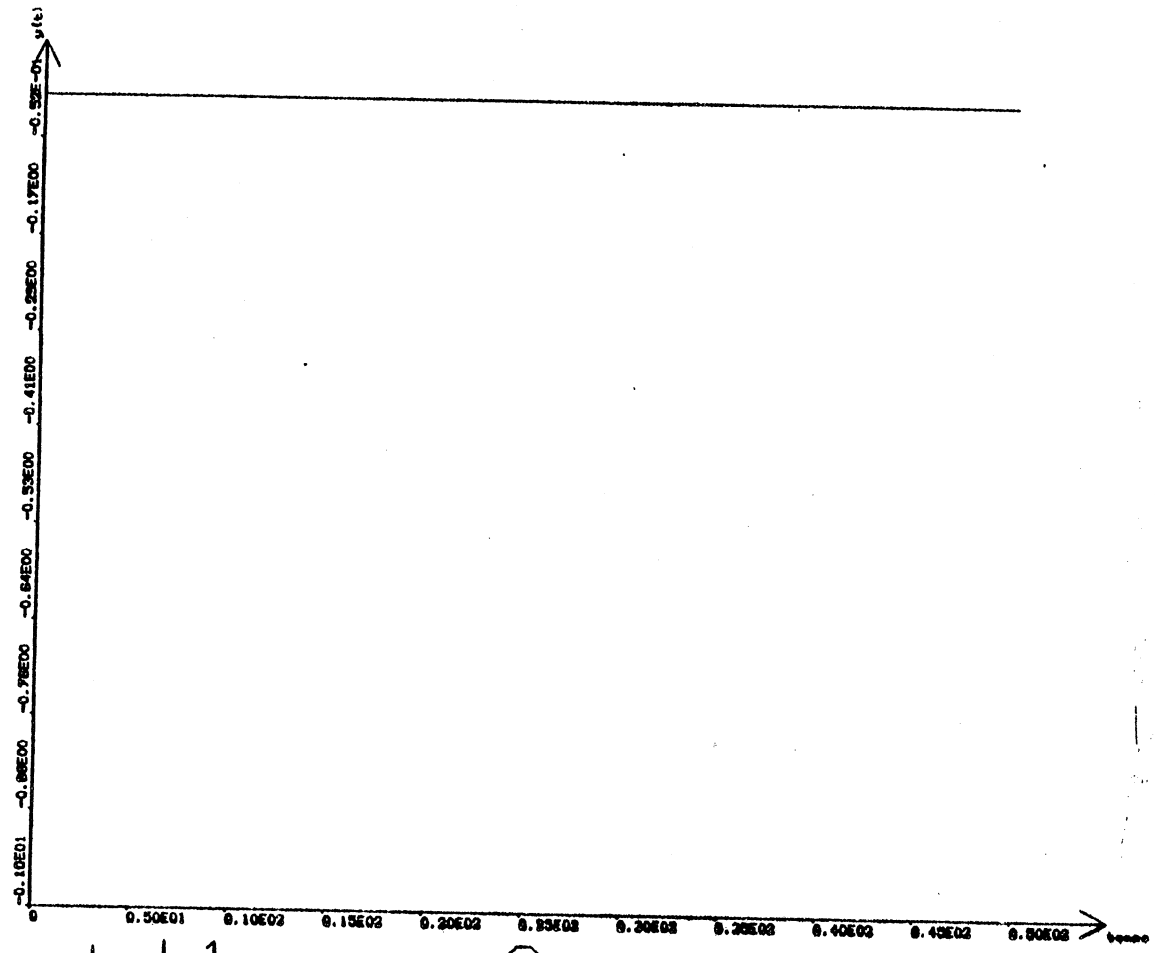
graphes des solutions du problème n° 6



variable no 1



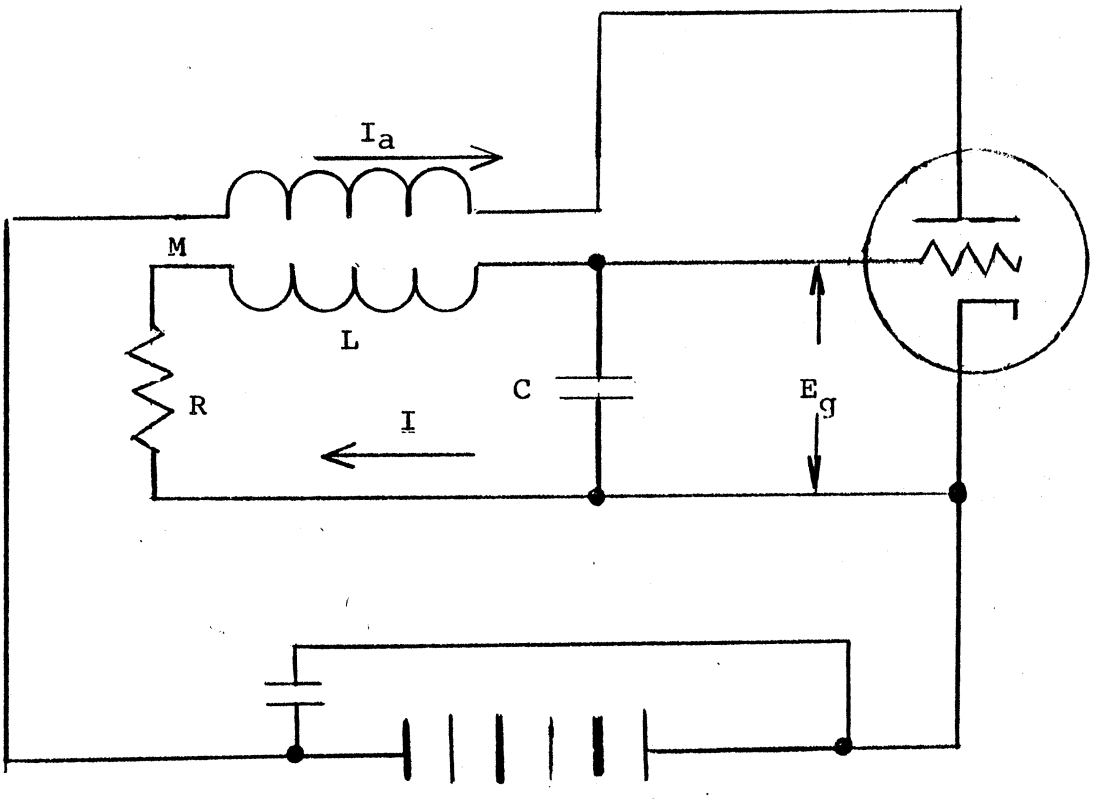
variable no 2



variable no 3

Le dernier problème test retenu est issu de la modélisation d'un circuit électrique nommé oscillateur de Van der Pol. Ce problème est connu pour sa difficulté et est cité dans [Ker, 1981, p 67] et [Wan, 1977].

Schéma de l'oscillateur de Van der Pol :



L'équation correspondante à ce circuit est :

$$(E) \quad L \frac{dI}{d\tau} + R\,I + \frac{Q}{C} = M \frac{dIa}{d\tau}$$

τ est le temps et M est la valeur de l'inductance mutuelle.

Nous avons également :

$$Ia = \sigma \left(Eg - \frac{1}{3} \frac{E_g^3}{E_s^2} \right)$$

où Eg est la tension de grille, Es la tension de saturation et σ la conductance du tube.

$$\text{De plus } I = C \frac{dEg}{d\tau}$$

Reporté dans (E), ces renseignements donnent :

$$\frac{d^2V}{d\bar{t}} + R\,\frac{\sqrt{C}}{\sqrt{L}}\,\frac{dV}{d\bar{t}} + V = \frac{\pi\,\sigma}{\sqrt{L}\,\sqrt{C}}\,(1 - V^2)\,\frac{dV}{d\bar{t}}$$

en posant $\omega = \frac{1}{\sqrt{(LC)}}$ et $\bar{t} = \omega\,\tau$

Si nous posons $y(\bar{t}) = \frac{V(\bar{t})}{A}$ et $A = \sqrt{(1 - \frac{RC}{M\sigma})}$ (E) devient

$$(E') \quad \frac{d^2y}{d\bar{t}^2} - \varepsilon(1 - y^2)\,\frac{dy}{d\bar{t}} + y = 0 \quad \text{avec} \quad \varepsilon = \frac{M\sigma A^2}{\sqrt{(LC)}}$$

Le problème ainsi posé est très irrégulier pour ε grand. Transformons-le en un problème différentiel du premier ordre :

$$\begin{cases} \dot{z}_1 - z_2 = 0 \\ \dot{z}_2 - \varepsilon(1 - z_1^2)z_2 + z_1 = 0 \end{cases}$$

Wanner [Wan, 1977] suggère de l'étudier avec les valeurs $\varepsilon = 100$, $t \in [0, 100]$, $z_1(0) = 2$ et $z_2(0) = 0$

La matrice $-\frac{\partial f}{\partial \dot{z}}(z,t)^{-1} \cdot \frac{\partial f}{\partial z}(z,t)$ vaut:

$$\begin{bmatrix} 0 & 1 \\ -200\,z_1\,z_2 - 1 & 100(1 - z_1^2) \end{bmatrix}$$

$t = 0$	$\lambda_1 \simeq -300$	$\lambda_2 \simeq -3.4\ 10^{-3}$
$t = 80$	$\lambda_{1,2} \simeq \pm 50\ i$	
$t = 100$	$\lambda_1 \simeq -300$	$\lambda_2 \simeq -3.4\ 10^{-3}$

Nous aurons donc, de manière prévisible, un problème à l'origine (valeurs propres négatives de rapport très grand), et un autre à $t = 80$ (valeurs propres imaginaires pures avec un grand module)

Nous avons du prendre un rapport $h_{min}/h_{max} = 10^{-7}$

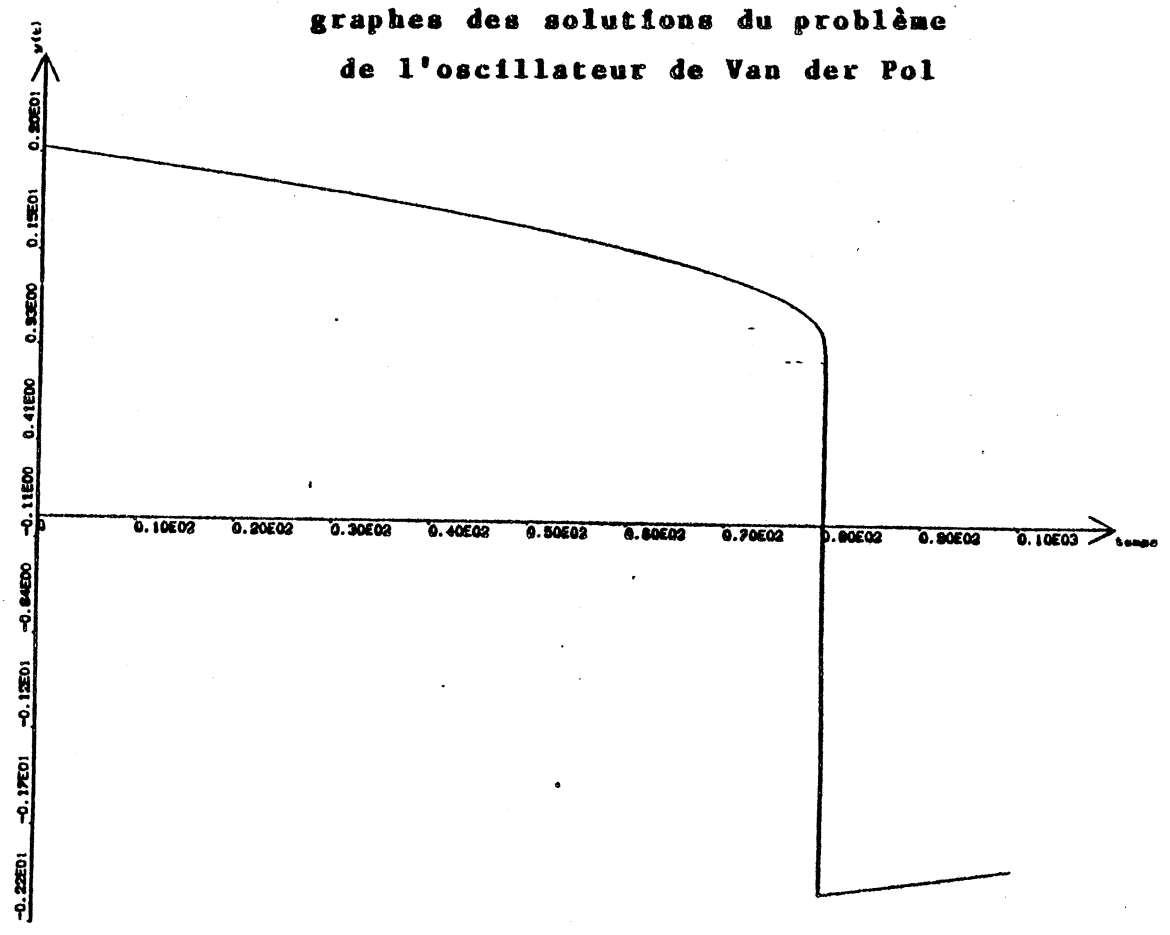
Résultats :

intervalle d' intégration	plus grand ordre utilisé	plus petit ordre utilisé	nombre de pas acceptés	nombre de pas refusés
[0, 10]	2	1	54	10
]10, 20]	1	1	10	0
]20, 30]	1	1	10	0
]30, 40]	1	1	10	0
]40, 50]	1	1	10	0
]50, 60]	1	1	10	0
]60, 70]	1	1	10	0
]70, 80]	3	1	11	3
]80, 90]	6	1	160	12
]90, 100]	1	1	10	0

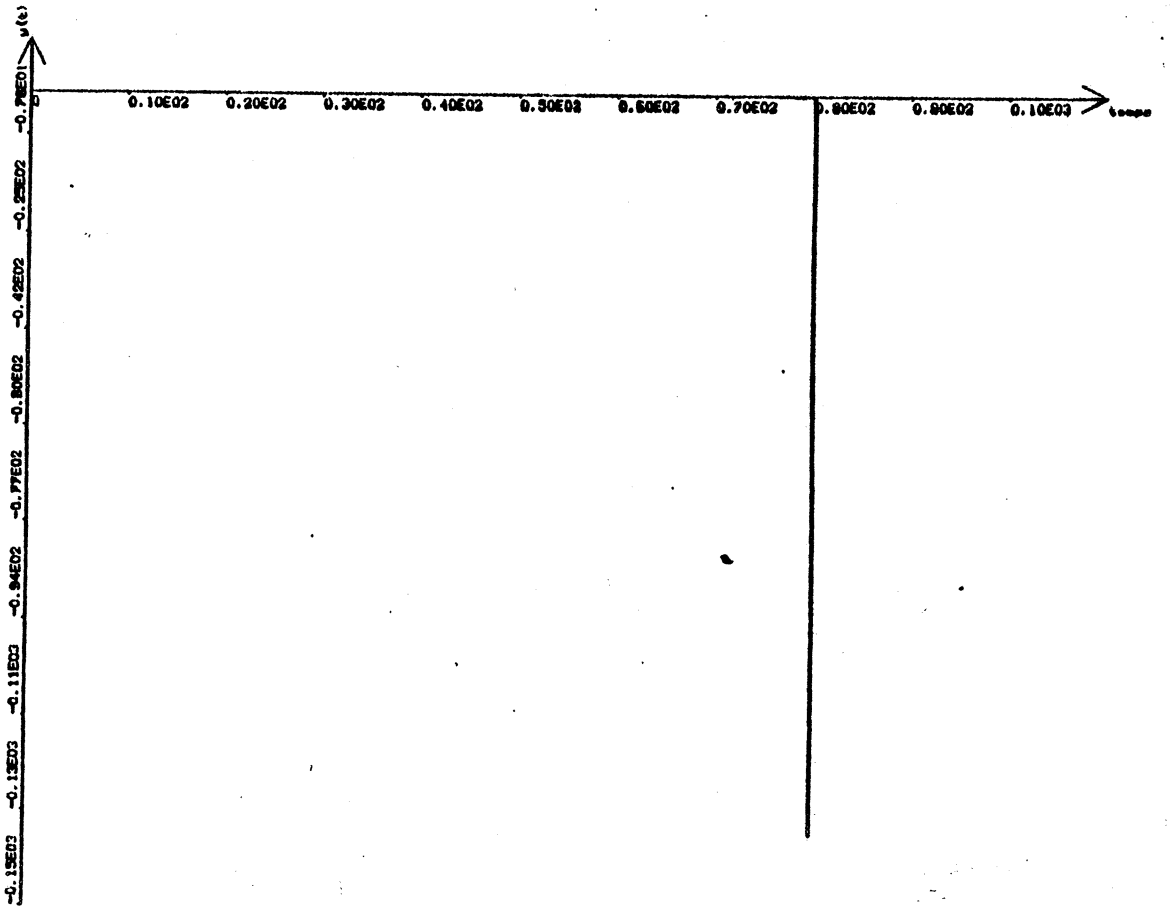
nombre de pas acceptés : 295

nombre de pas refusés : 25

graphes des solutions du problème de l'oscillateur de Van der Pol



variable no 1



Les résultats des expérimentations précédentes indiquent que l'ensemble de notre démarche (méthode quasi-stable, calcul d'erreur, adaptation du pas et tests d'arrêts) est bien adapté aux systèmes raides, c'est à dire aux systèmes algébrodifférentiels dont la matrice

$-\frac{\partial f}{\partial \dot{y}z}(y,z,\dot{z},t)^{-1} \cdot \frac{\partial f}{\partial z}(y,z,\dot{z},t)$ a des valeurs propres d'ordre de grandeur très différents. Ces systèmes ont généralement des solutions qui connaissent des zones de "calme" relatif succédant à des variations très brusques (cf. l'oscillateur de Van der Pol). Les variables issues de la modélisation des circuits électroniques se comportent souvent comme des fonctions "créneaux". Les méthodes exposées précédemment conviennent donc particulièrement pour ce type de variables.

VI)ELABORATION D'UN CODE POUR LA RESOLUTION DESYSTEMES ALGEBRODIFFERENTIELS RAIDES

La question qui demeure est que les systèmes issus de la modélisation des circuits électroniques de grande dimension (plusieur centaines,voire milliers d'équations). Il est donc important de tester la validité numérique des méthodes exposées dans le chapitre II et de comparer entre elles la méthode de Newton classique, la méthode de Newton à plusieurs niveaux(méthode MLNA) et la méthode de surrelaxation basée sur un découplage de blocs (méthode RELAX).

Pour cela nous utiliserons CREDO (Conversationnel pour la Résolution des Equations Différentielles Ordinaires), un ensemble de 19 programmes et sous-programmes formant un tout de trois à quatre mille instructions. Cet ensemble comporte une partie conversationnelle(DISCOURS) qui permet à l'utilisateur de spécifier son problème(intervalle d'intégration,valeurs initiales des variables), la méthode choisie(NEWT, MLNA ou RELAX), le nombre minimal de pas desiré, le rapport h_{min}/h_{max}, la précision voulue, les variables dont on désire des sorties graphiques.

Liste et fonctions des programmes et sous-programmes de CREDO

_ BOKHOVEN regroupe les méthodes à pas calculé

globalement NEWT et MLNA.

_ EDO contient la méthode RELAX.

_ TRI réalise le découpage en blocs de la matrice jacobienne après qu'on lui ait fourni le taux de remplissage voulu pour les blocs.

_ VALDIF calcule les valeurs initiales des dérivées des variables différentielles.

_ TRONCAT et TRONREL calculent grâce à RESOUD(méthode de point fixe-Steffesen) l'erreur de troncature respectivement pour les méthodes à pas calculé globalement et pour les méthodes de surrelaxation.

_ DGY et DGYRELAX calculent le jacobien

$\frac{\partial H}{\partial xy}(u,x,y)^{-1} \cdot \frac{\partial H}{\partial u}(u,x,y)$ (pour plus de détails se reporter à la description dans le chapitre II des algorithmes MLNA et de découplage)

_ RESO résoud le système triangulaire (s'il existe) que TRI a détecté.

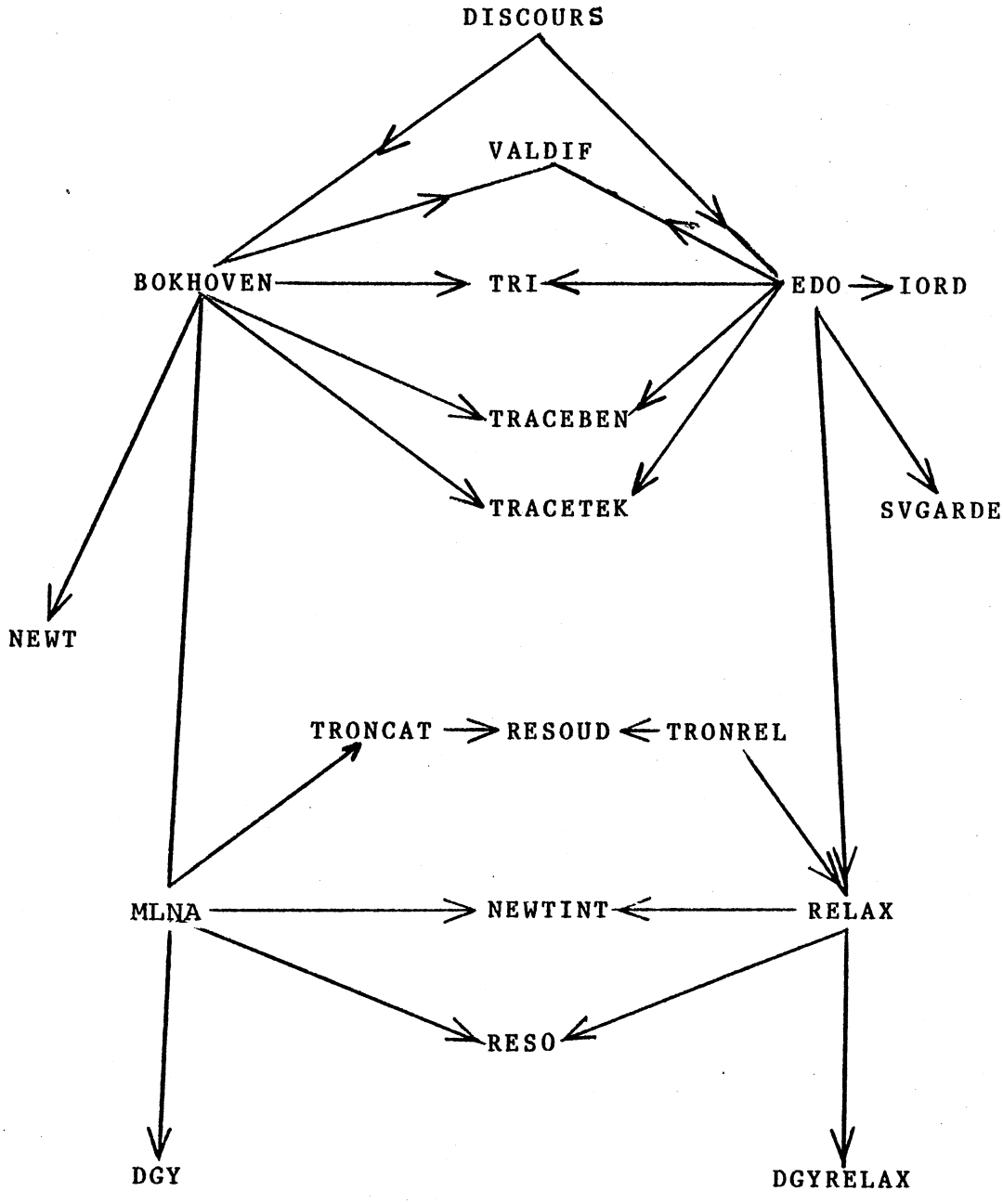
_ NEWTINT est une méthode de Newton utilisée pour la résolution de $H_1(u,x,y) = 0$ (niveau interne de MLNA ou de RELAX)

_ SVGARDE est appelé par EDO. Il sert à garantir la bonne synchronisation des différents blocs. Il est utilisé lorsqu'une désynchronisation est constatée(des précisions concernant ce sous-programme sont données plus loin). En temps normal,la synchronisation est assurée par l'algorithme de découplage tel qu'il est décrit dans le chapitre II)

_ IORD ordonne les blocs et indique le numéro du bloc que RELAX devra traiter.

_ TRACEBEN et TRACETEK sont des procédures graphiques permettant le tracé des variables dont on désire les graphes (BENSON et TEKTRONIX respectivement)

La structure d'appels est la suivante:



Un problème délicat s'est posé à nous lors de la réalisation pratique de EDO est des procédures qui y sont attachées. Il concerne les variables servant "d'entrées" à plusieurs blocs.

Supposons, pour simplifier, que la variable u soit une entrée à la fois pour le bloc H_1 et pour le bloc H_2(dont les variables internes et les "sorties" respectives sont notées x_1, y_1 et x_2, y_2)

H_1 a été évalué au temps t_{N1} et H_2 a été évalué au temps t_{N2} (avec t_{N1} inférieur à t_{N2}), et le code de résolution a accordé à H_1 un pas h_1 et un ordre k_1, à H_2 un pas h_2 et un ordre k_2.

Supposons que le processus d'intégration en soit au temps $t_{N1}+h_1$.

Pour évaluer la valeur des variables de H_1 nous allons, comme nous l'avons vu précédemment, résoudre un système algébrique. Pour cela, et pour estimer l'erreur de troncature, nous calculons des valeurs (prédicteurs) basées-en ce qui concerne les variables internes et les sorties- sur les valeurs x_{N1}, x_{N1-1}, ...,x_{N1-k_1}, et y_{N1}, y_{N1-1},...,y_{N1-k_1}, qui sont les dernières valeurs calculées pour ces variables.

Ce n'est pas le cas pour u qui a été évalué en plus au moment t_{N2}.

Deux choix s'offrent à nous:

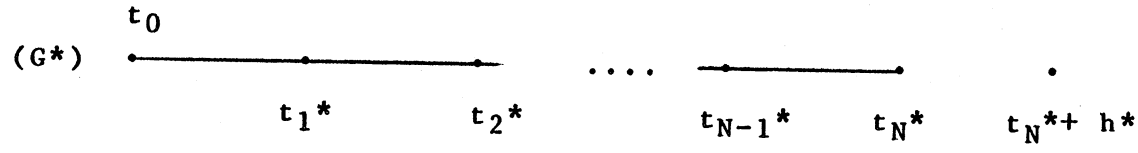
- i) nous ne tenons pas compte de l'évaluation faite au temps t_{N2}. Le calcul des prédicteurs et de l'erreur de troncature est donc identique à ceux de x et y. Les précision demandées sur x, y et u sont les mêmes.
- ii) nous prenons en compte l'évaluation faite au temps t_{N2}. Dans ce cas-là,le polynôme d'interpolation et le calcul de l'erreur de troncature doivent être adaptés. De plus, la précision demandée sur x, y et u qui ne sont pas calculés sur un même pas (h_1 pour x et y, $h_u = t_{N1}+h_1-t_{N2}$ pour u) ne pourront être identiques.

Nous avons opté pour le ii), malgrès sa plus grande difficulté dans la mise en œuvre. Nous pensons qu'il assure une plus grande stabilité au schéma d'ensemble.

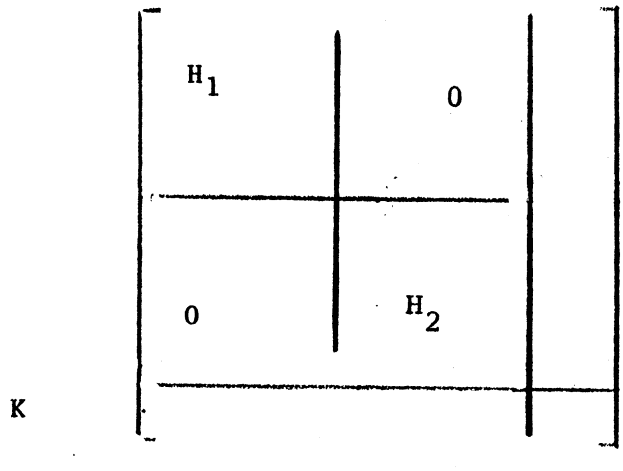
Nous allons maintenant donner plus de détails sur la procédure SVGARDE et les raisons qui ont amené à l'élaborer.

Définissons d'abord ce qu'est le graphe d'avancement d'un bloc H.

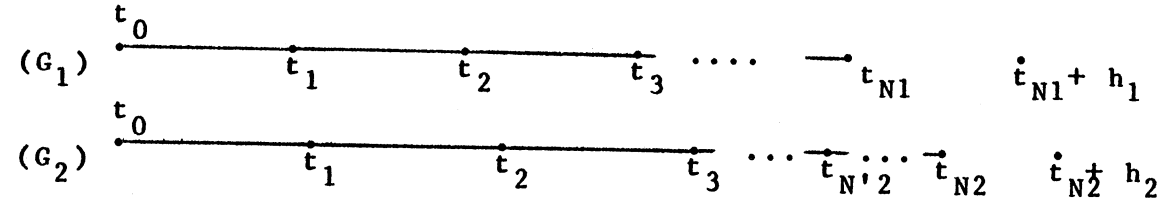
Supposons que les variables propres au bloc H* aient été calculées aux points t_0, t_1*, ..., t_N* et que le calcul du nouveau pas au point t_N* ait donné h*. Le graphe d'avancement (G*) de H* sera constitué des points t_0, t_1*, ..., t_N* reliés entre eux par des traits pleins et du point t_N*+h* :



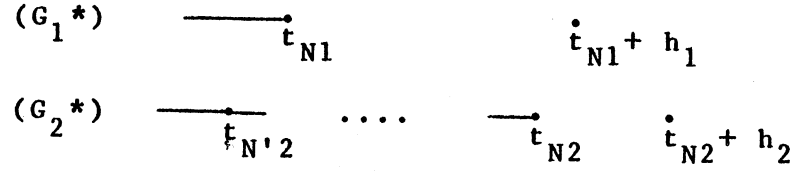
Supposons maintenant, pour simplifier, que nous ayons partitionné la matrice jacobienne du système en deux blocs H_1 et H_2



et que ces deux blocs aient les graphes d'avancement respectifs (G_1) et (G_2) suivants :



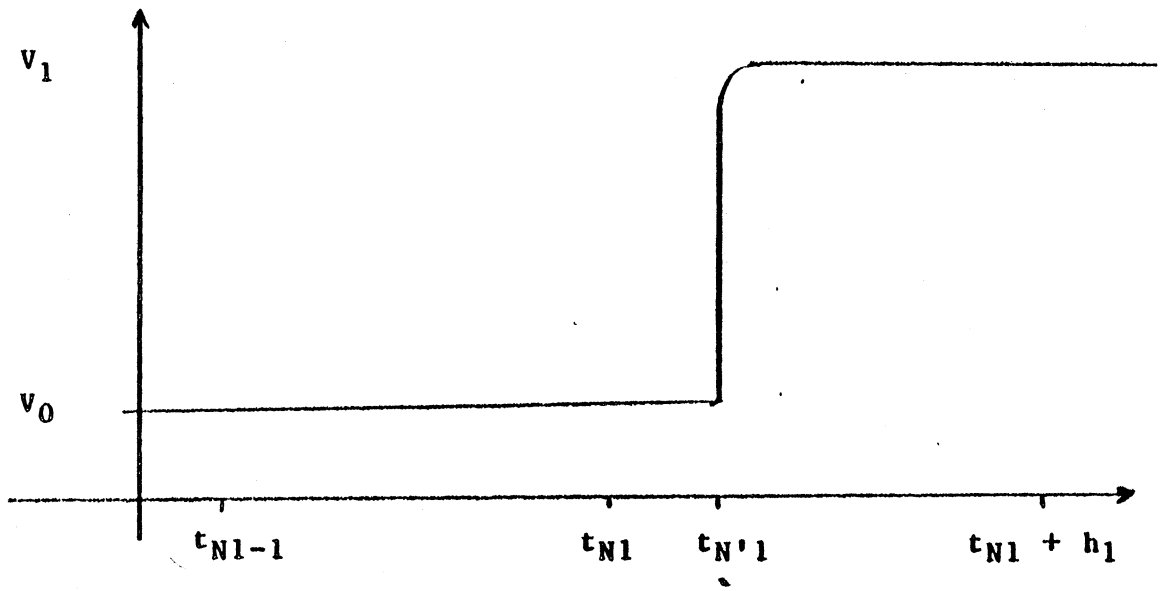
"Grossissons" les derniers tronçons des graphes (G_1) et (G_2) :



Reconstituons maintenant le déroulement des calculs. Au temps t_{N1}, les différents tests d'erreur portant sur les variables du bloc H_1 étaient bien satisfaits, et ce dernier s'est vu "attribuer" un certain pas h_1. H_2, lui, satisfait moins bien aux tests d'erreurs et se voit attribuer au temps t_{N2} un pas nettement plus petit h_2. Les variables de H_2 sont donc évaluées jusqu'au temps t_{N2}, moment où il doit "céder" la place, puisque $t_{N2}+h_2$ est supérieur à $t_{N1}+h_1$.

Pour pouvoir évaluer les variables de H_2, il a été nécessaire, afin de résoudre K, de faire des estimations de la valeur des variables de H_1. Ces estimations sont faites grâce à l'utilisation de polynômes d'interpolation, dont les valeurs sont supposées ne pas trop différer des valeurs réelles prises par les variables de H_1.

Supposons maintenant qu'une des variables de H_1 ait le graphe suivant :



Ainsi, au temps t_{N1}, les contrôles d'erreurs étaient très bien satisfaits car la variable en question (désignons-la par y_i) était une fonction constante ou presque. Le programme n'étant pas en mesure de prévoir le "futur" de la variable, alloue à H_1 un grand pas .Pour les calculs liés à l'évaluation de H_2, l'estimation de y_i qui sera donnée sera peu différente de V_0., alors qu'à partir du temps $t_{N'1}$, il aurait fallu donner la valeur V_1. Les résultats des calculs des variables

de H_2 sont donc fortement entachés d'erreurs. Comment y remédier?

Revenons au temps $t_{N1} + h_1$.
Le point de départ pour la méthode de Newton relatif au bloc H_1 va se révéler très éloigné de la valeur réelle : soit les itérations de Newton vont diverger soit de toutes façons l'erreur de troncature va être considérable. Le pas va donc être rejeté et divisé par quatre jusqu'à ce que les tests d'erreurs soient satisfaits, dans notre cas jusqu'à $t_{N'1}$. Il s'avère donc que l'idéal serait de pouvoir revenir au temps t_{N1}, d'être à même de "régénérer" les valeurs des variables, des ordres des méthodes,... en ce temps là et de faire redémarrer l'algorithme en possession cette fois-ci de l'information que la prochaine évaluation des variables de H_1 se fait au temps $t_{N'1}$.

C'est ce que réalise la procédure SVGARDE. Son mécanisme est simple :
chaque bloc H_i se voit attribuer un fichier "mémoire" $ifmem_i$ dans lequel on range tout ce qui est nécessaire à un redémarrage éventuel de l'algorithme. Dès qu'un problème surgit (signalé par le fait que l'ordre normal d'évaluation des blocs n'a pas été respecté) la procédure SVGARDE est appelée. Cette procédure lit dans $ifmem_i$ les renseignements indiqués précédemment et ramène l'ensemble du système à un temps antérieur (dans notre exemple $t_{N'1}$).

Les opérations d'entrées-sorties sur les fichiers coûtent cher. Il nous faut les optimiser, pour ne les réaliser qu'à bon escient. Diverses remarques nous aident dans ce sens :

1) il est inutile d'associer un fichier aux blocs n'ayant pas de "sorties". Ainsi, dans notre exemple, si les variables de H_1 n'intervenaient pas dans K, la mauvaise estimation de y_i n'aurait aucune influence sur les calculs faits pour les variables de H_2 et il n'était pas alors nécessaire de faire revenir H_2 à $t_{N'1}$.

2) dans un fichier, il est inutile d'écrire les informations relatives aux blocs n'ayant pas d'entrées. En effet, ceux-ci n'ont pas pu être affectés par une mauvaise résolution de K.

3) il n'est nécessaire de mettre à jour les fichiers $ifmem_i$ que lorsque le bloc évalué au temps suivant est un H_j avec j différent de i. Ainsi, dans notre exemple, le fichier $ifmem_2$ sera mis à jour aux temps $t_{N'2}$ et t_{N2}, mais pas aux temps intermédiaires.

4) De plus, pour un bloc H_i donné, il n'est utile de mémoriser dans $ifmem_i$ que les blocs qui seront appelés plus d'une fois avant le "retour" de H_i. Concrétement, les renseignements relatifs à H_j ne seront écrits dans $ifmem_i$ que si, en notant t_{Nj} le moment où H_j est évalué et h_j le pas que lui accorde les procédures de calcul de pas et d'ordre, nous avons $t_{Nj}+h_j$ inférieur à $t_{Ni}+h_i$.

VII)COMPARAISON DES DIFFERENTES METHODES ENTRE ELLES

Nous avons testé les trois méthodes(NEWT, MLNA, et RELAX) sur le problème suivant :

$$
\begin{cases}
\dot{z}_1 - z_2 = 0 \\
\dot{z}_2 - 100(1-z_1^2)z_2 + z_1 = 0 \\
\dot{z}_3 + 0.1\sin(0.1t) = 0 \\
\dot{z}_4 - 0.1z_3 = 0 \\
5\dot{z}_5 + z_5 - 2 + (t+1)z_7 = 0 \\
\dot{z}_6 + 10^5(z_6 - t^3) - (t+1)z_7 + 1 - 3t^2 = 0 \\
z_5 - 1 + e^{-0.2t} + \dot{z}_7 + 10^5\,(z_7 - 1/(t+1)) + 1/(t+1)^2 = 0 \\
t \in [0, 100] \\
z_1(0) = 2 \qquad z_2(0) = 0 \; z_3(0) = 1 \\
z_4(0) = 0 \qquad z_5(0) = 0 \; z_6(0) = 0 \\
z_7(0) = 1
\end{cases}
$$

z_1 et z_2 sont les solutions du problème de Van der Pol

$z_3(t) = \cos(0.1t)$ et $z_4(t) = \sin(0.1t)$

$z_5(t) = 1 - e^{-0.2t}$

$z_6(t) = t^3$

$z_7(t) = 1/(t+1)$

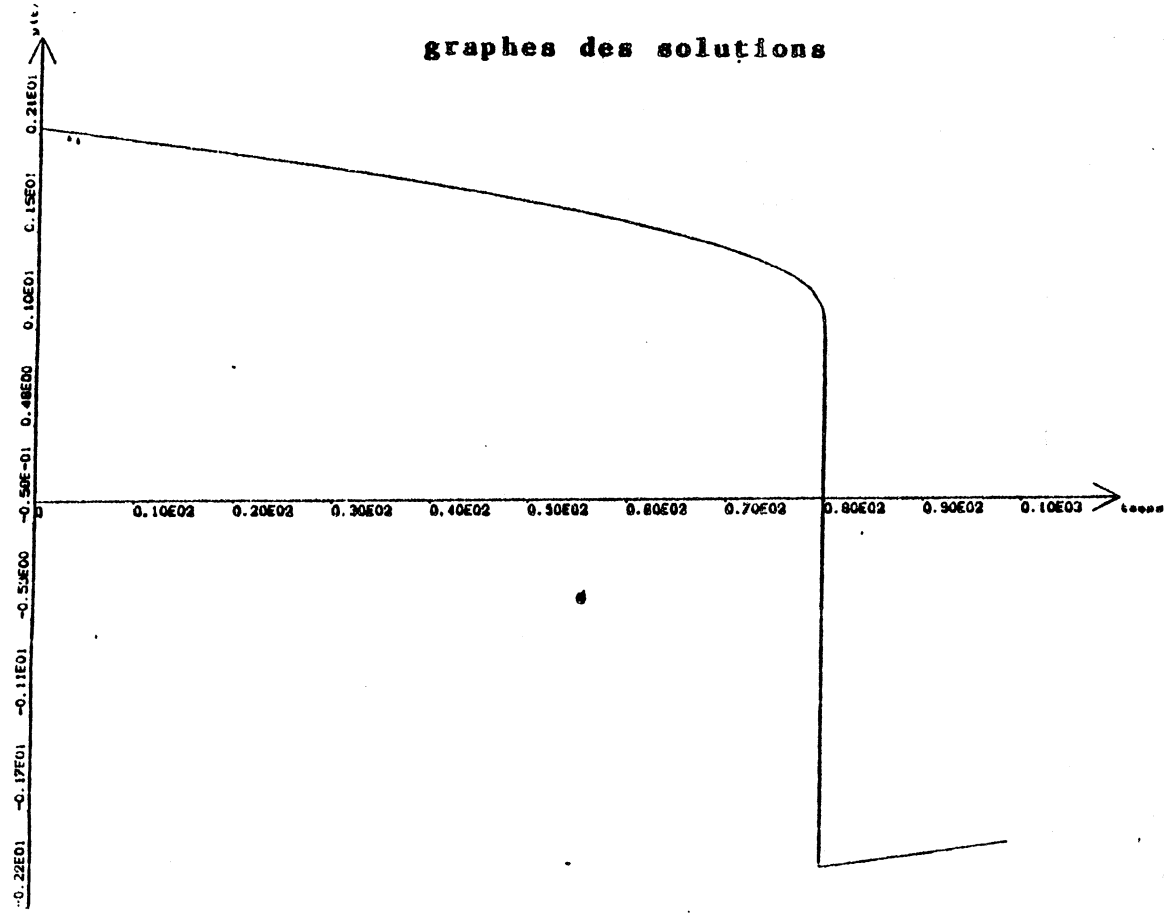
Ce système comporte un certain nombre de difficultés liées d'une part aux solutions, d'autre part au système lui-même :

difficultés liées aux solutions :

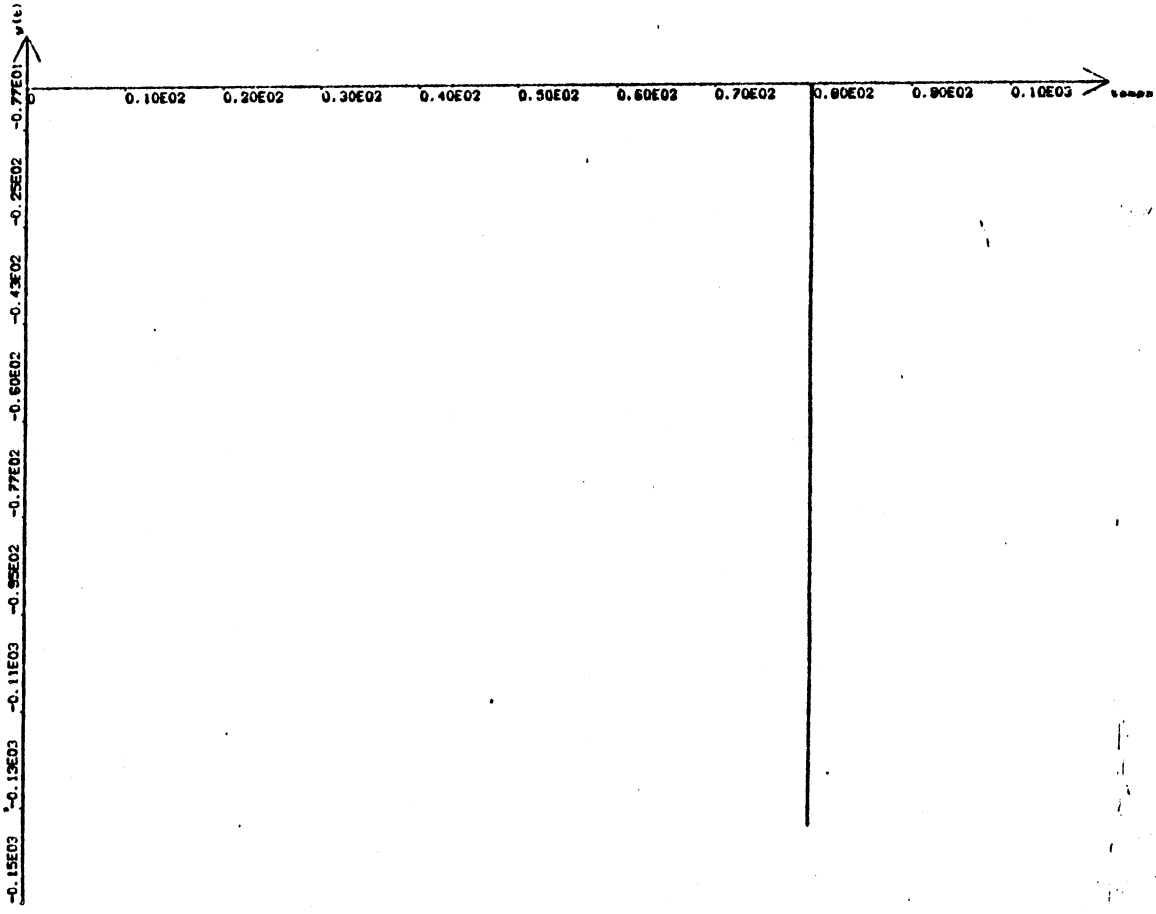
— z_5 et z_7 sont raides à l'origine, mais deviennent par la suite très régulières (voire constante pour z_5). Le pas devra donc pouvoir varier très rapidement

— z_6 a une pente toujours croissante. La solution numérique devra donc être aussi ajustée que possible pour éviter la dérive numérique. Cependant z_6 n'est pas exponentielle (ce qui aurait condamné le pas à devenir de plus en plus petit) mais polynomiale et de degré inférieur à 6, ce qui doit permettre aux méthodes quasi-stables de suivre la solution.

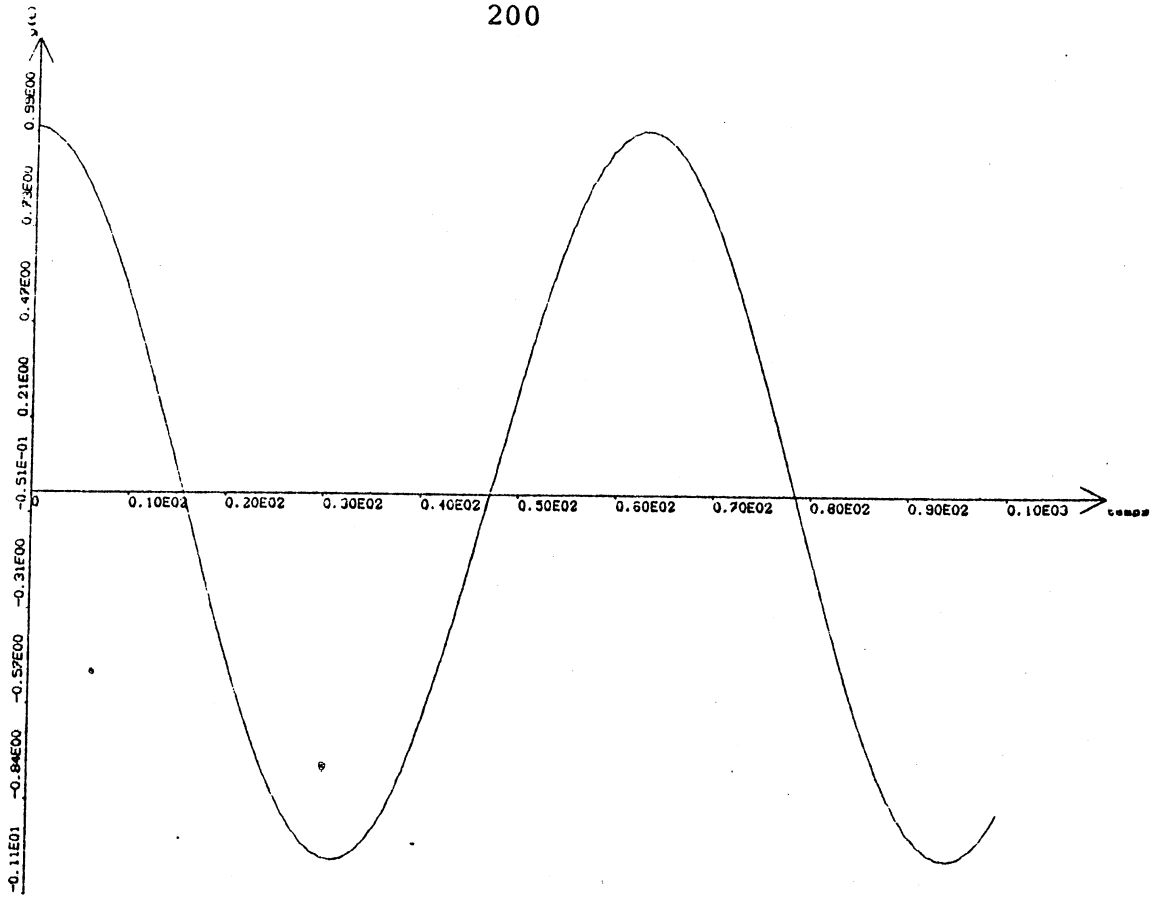
graphes des solutions



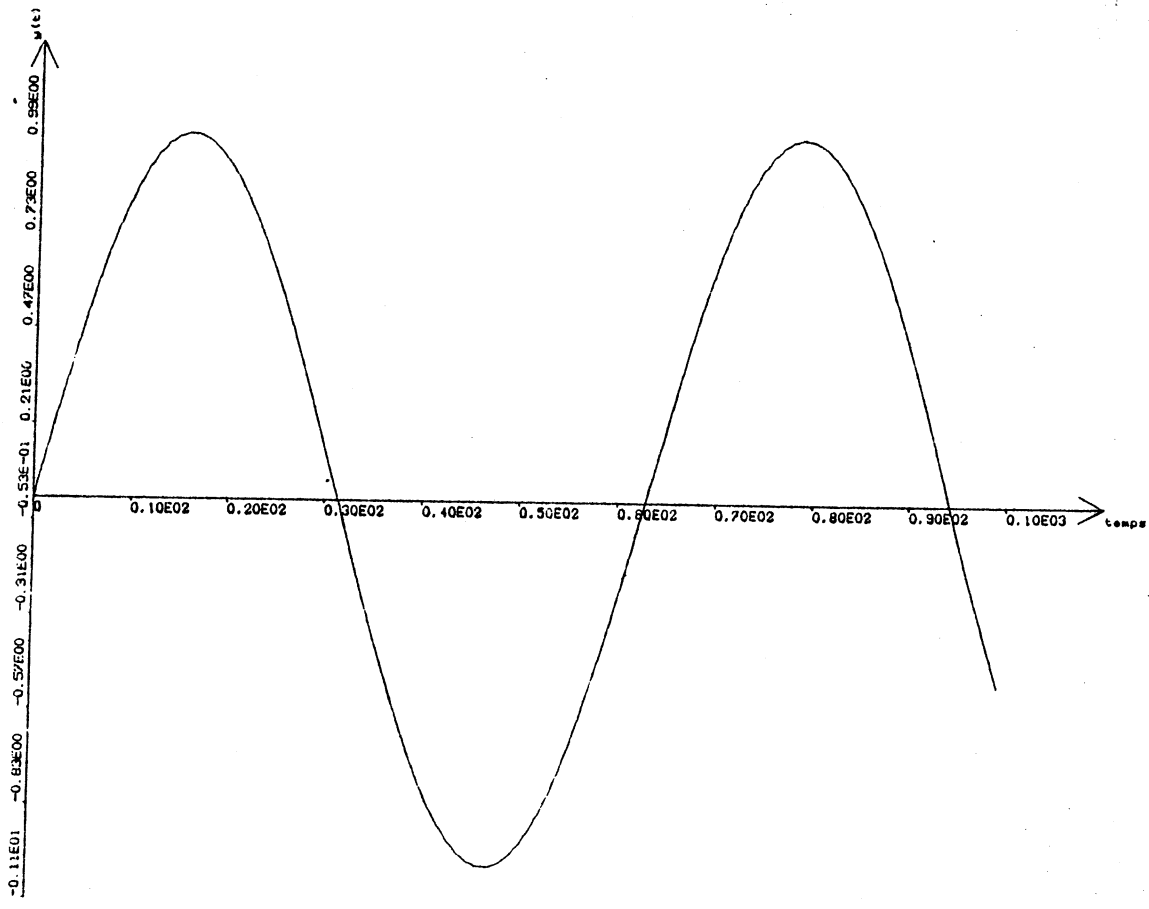
variable no 1



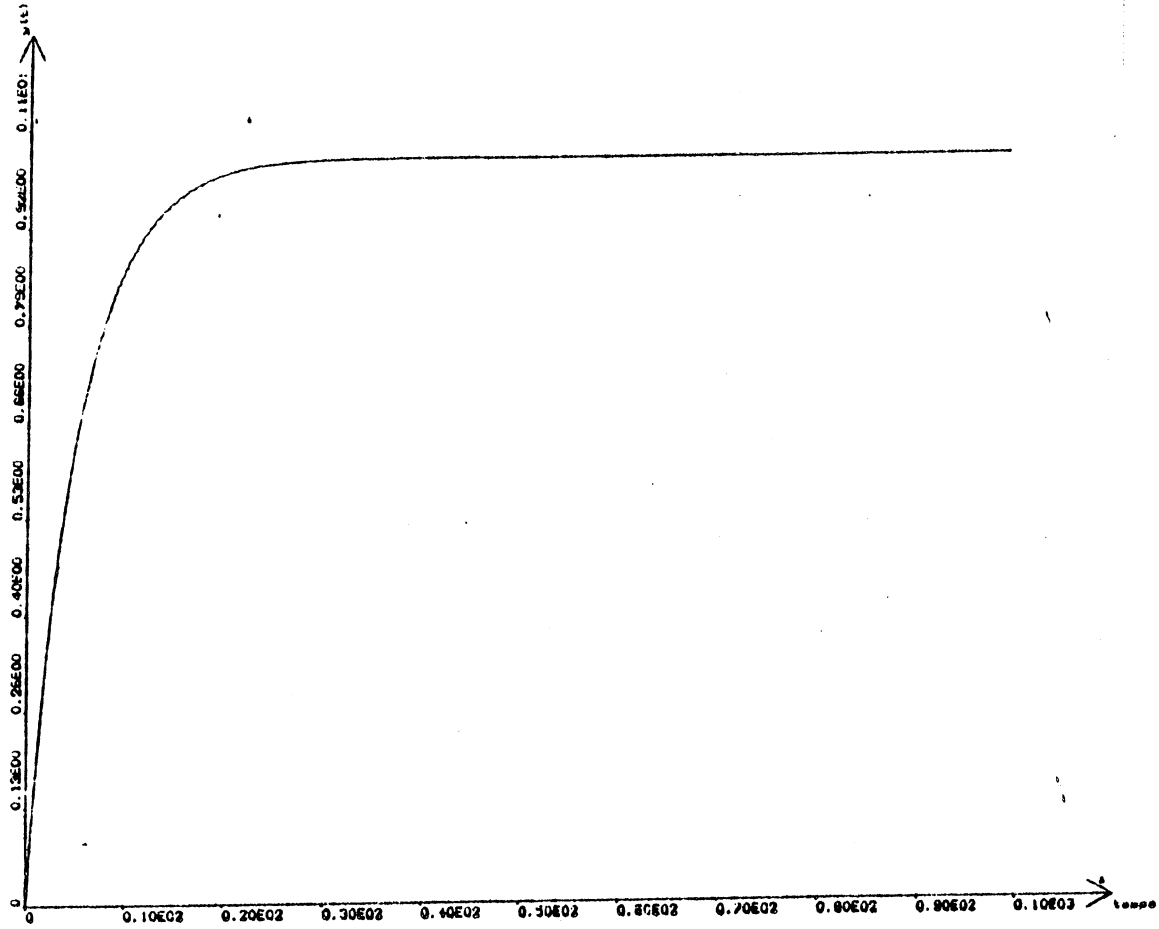
variable no 2



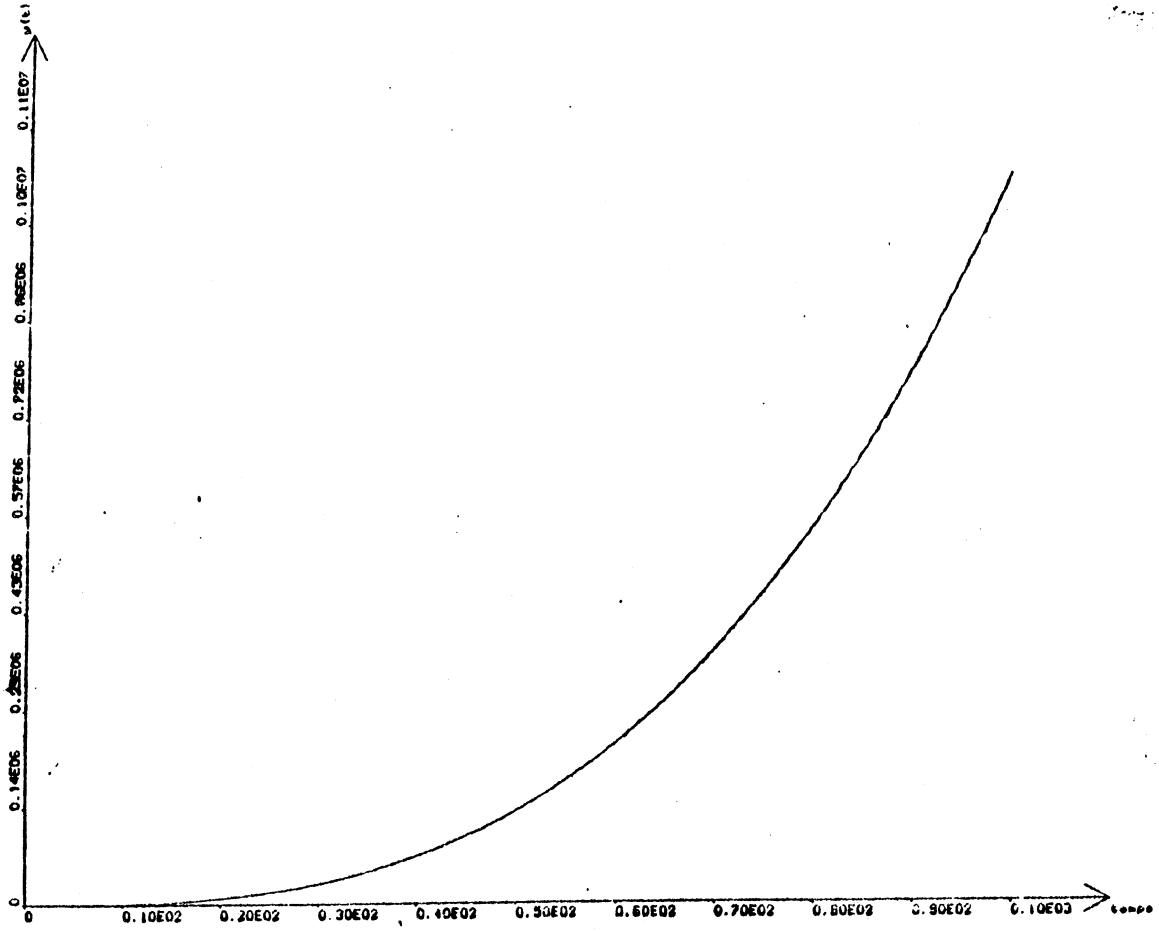
variable no 3



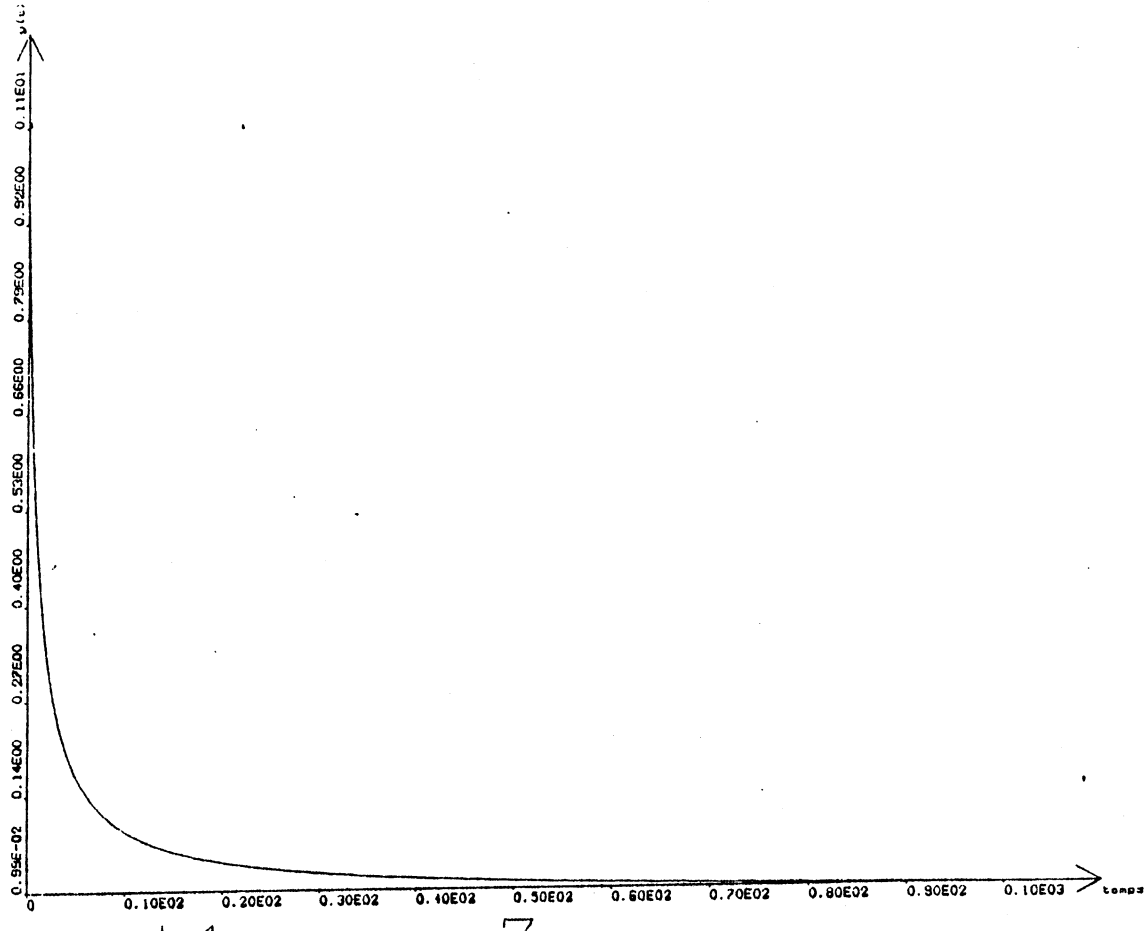
variable no 4



variable no 5



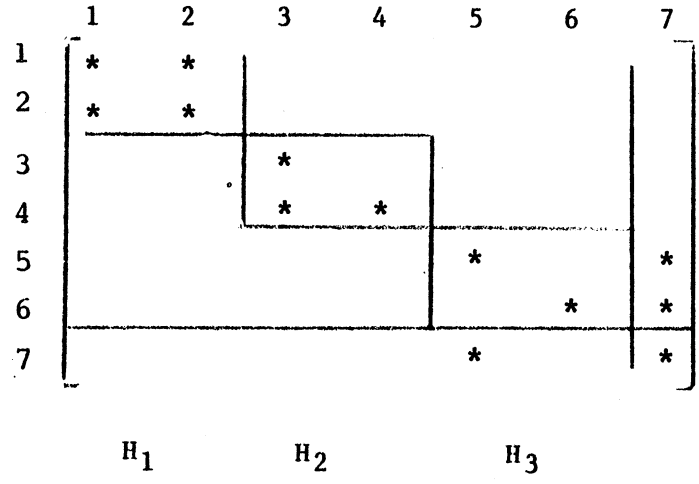
variable no 6



variable no 7

difficultés liées au systèmes :

Nous avons partitionné la matrice jacobienne du système en trois blocs H_1, H_2 et H_3 :



Les difficultés liées à H_1 ont été décrites dans le problème n°6 de la partie précédente .

Celles liées à H_2 viennent de ce que les valeurs propres de $-\frac{\partial f_{3,4}}{\partial \dot{z}_3 \dot{z}_4}(z,\dot{z},t)^{-1} \cdot \frac{\partial f_{3,4}}{\partial z_3 z_4}(z,\dot{z},t)$ sont égales à 0

Les valeurs propres de
$-\frac{\partial f_{5,6}}{\partial \dot{z}_5 \dot{z}_6}(z,\dot{z},t)^{-1} \cdot \frac{\partial f_{5,6}}{\partial z_5 z_6}(z,\dot{z},t)$ sont réelles, négatives bien séparées: $\lambda_1 \simeq -10^5$, $\lambda \simeq -1/5$

Les résultats numériques obtenus sont les suivants (la précision relative demandée est 10^{-2}, et afin de pouvoir comparer NEWT et MLNA avec RELAX nous avons multiplié le nombre de pas par le nombre des variables concernées) :

méthode de Newton

nombre total de pas acceptés×nombre de variables : 885×7= 6195
nombre total de pas refusés×nombre de variables : 331×7= 2317

temps C.P.U. utilisé pour :

- le calcul des jacobiens : 8.2 s
- les évaluations de fonction : 2.4 s
- les résolutions de systèmes : 78.2 s

- le calcul des prédicteurs : 2.1 s
- le calcul de l'erreur de troncature, du pas et de l'ordre : 4.0 s

temps C.P.U. utilisé : 94.9 s

méthode MLNA

nombre total de pas acceptés×nombre de variables : 859×7= 5894
nombre total de pas refusés×nombre de variables : 319×7= 2233

temps C.P.U. utilisé pour :

- le calcul de jacobiens : 4.1 s
- les évaluations de fonctions : 2.7 s
- les résolutions de systèmes : 47.5 s

- le calcul des prédicteurs : 2.1 s
- le calcul de l'erreur de troncature, du pas et de l'ordre : 4.0 s

- temps C.P.U. utilisé : 60.4 s

méthode RELAX

nombre de pas acceptés pour le bloc H_1 : 290×2 = 580
H_2 : 109×2 = 218
H_3 : 121×2 = 242
nombre de pas effectués par les entrées : 121×1 = 121
nombre total de pas acceptés : 1261

nombre de pas refusés pour le bloc H_1 : 41×2 = 82
H_2 : 3×2 = 6
H_3 : 7×2 = 14
nombre de refus dans lesquels étaient impliquées des entrées : 7×1 = 7
nombre total de pas refusés : 169

temps C.P.U. utilisé pour :

- le calcul de jacobiens : 1.0 s
- les évaluations de fonctions : 0.6 s
- les résolutions de systèmes : 9.9 s

_ le calcul des prédicteurs : 1.8 s
_ le calcul de l'erreur de
troncature, du pas et de l'ordre : 2.0 s
_ les sauvegardes sur fichier : 4.0 s

_ temps C.P.U. total : 18.3 s

Nous pouvons donc constater une très nette supériorité de la méthode de surrelaxation par rapport aux deux autres méthodes, et ce à tout point de vue.

La méthode de Newton à plusieurs niveaux(MLNA) est meilleure que la méthode de Newton.

Cependant, il nous faut remarquer que l'ensemble des systèmes que la méthode de Newton ou celle de Newton à plusieurs niveaux peut résoudre est plus vaste que celui de la méthode de relaxation (cf. chapitre II, partie VI)).

VIII) ETUDE DU COUT DES ALGORITHMES DE PARTITIONNEMENT (PROCEDURE TRI)

Nous avons construit une procédure GROSTEST(n) qui génère sous forme de liste des matrices n×n, ayant tous ses éléments diagonaux non-nuls et un nombre d'éléments non-nuls extra-diagonaux valant partie entière de [(2+2x).n], où x est un nombre aléatoire compris entre 0 et 1 (notre choix provient du fait que les matrices jacobiennes des systèmes algébrodifférentiels issus de la modélisation par CASCADE de circuits électriques ont moins de cinq éléments non-nuls par ligne en général). Les éléments extra-diagonaux sont placés aléatoirement dans la matrice.

La procédure TRI se décompose en deux temps:

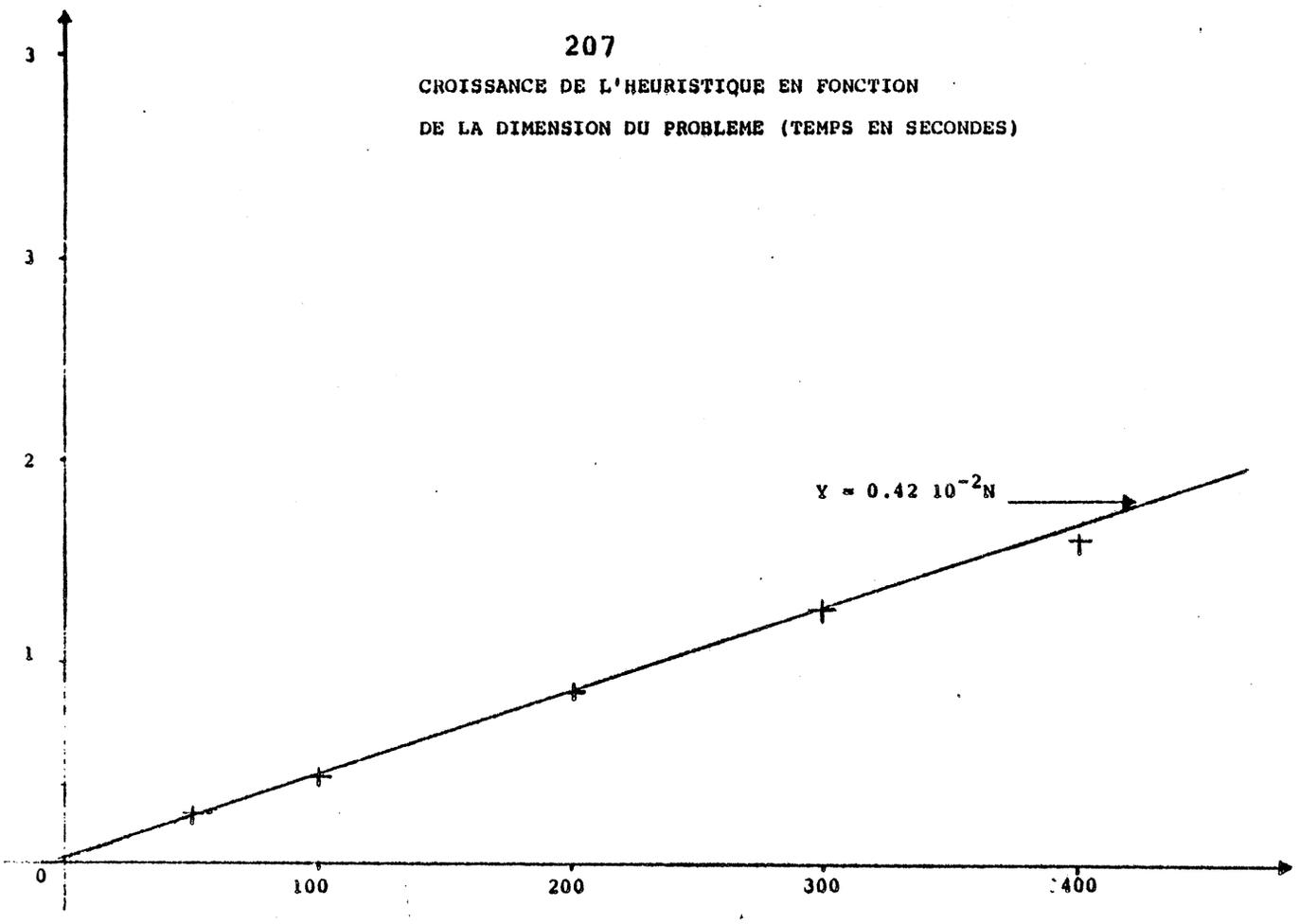
1) recherche d'une permutation des indices lignes tels que la diagonale ne soit composée que d'éléments non-nuls.
2) transformation de la matrice en matrice diagonale par blocs avec bande

Le fait que la diagonale de la matrice à traiter soit déjà non-nulle ne diminue en rien le travail que fournit la première partie de TRI, puisqu'au départ l'heuristique ne travaille que sur le nombre d'éléments existant dans chaque ligne et chaque colonne, et non pas sur une connaissance préalable de la structure de la matrice (cf. chapitre II, partie IV)). Nous nous assurons simplement ainsi de la faisabilité de la tâche confiée à TRI.

Les performances de TRI dépendent énormément de la structure de la matrice. Aussi avons-nous préféré travailler sur des échantillons de 20 à 25 matrices pour chaque n étudié (n= 100, 200, 300, 400) et faire des moyennes sur ces échantillons. Nous avons obtenu les graphes qui suivent.

CROISSANCE DE L'HEURISTIQUE EN FONCTION

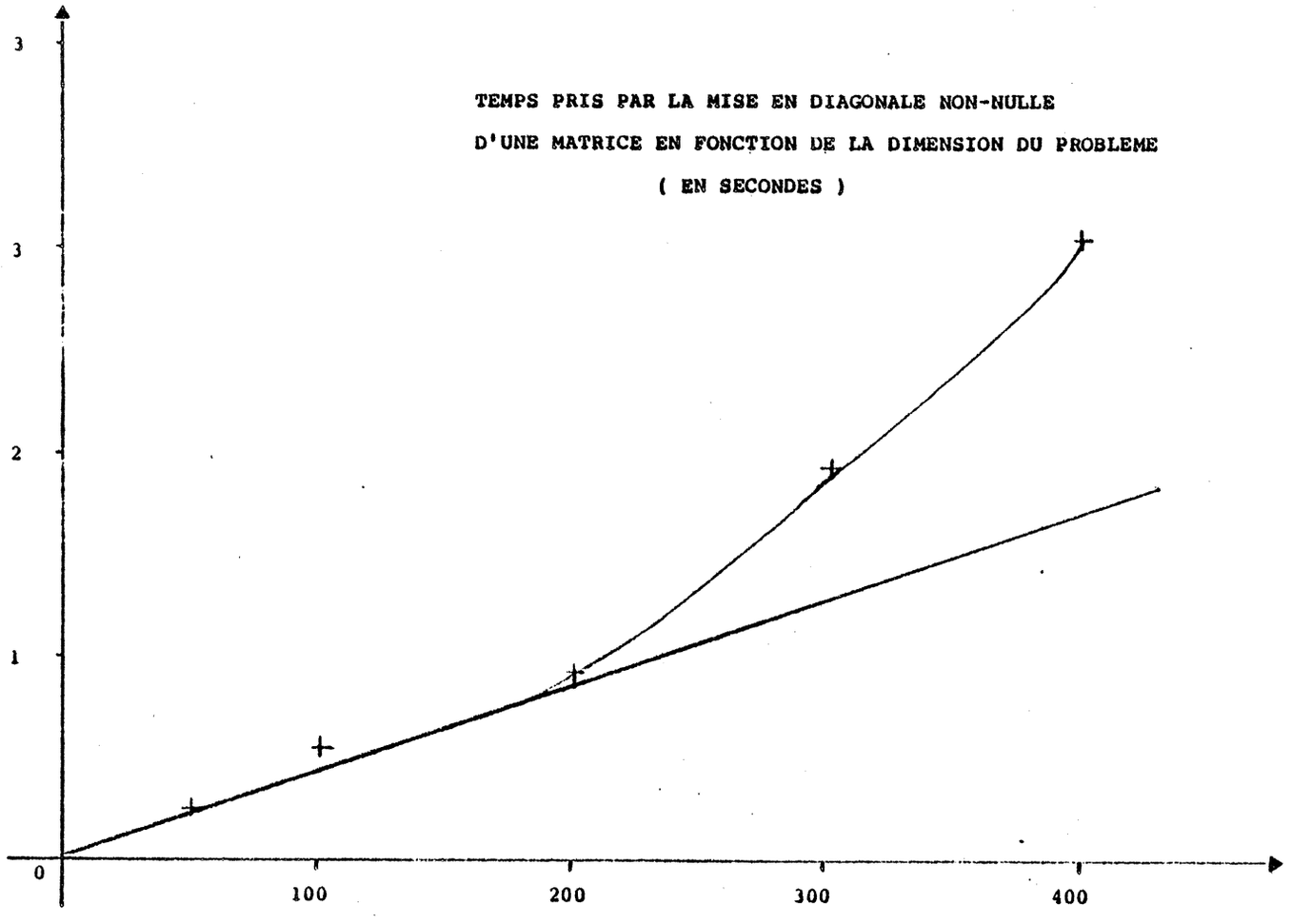
DE LA DIMENSION DU PROBLEME (TEMPS EN SECONDES)



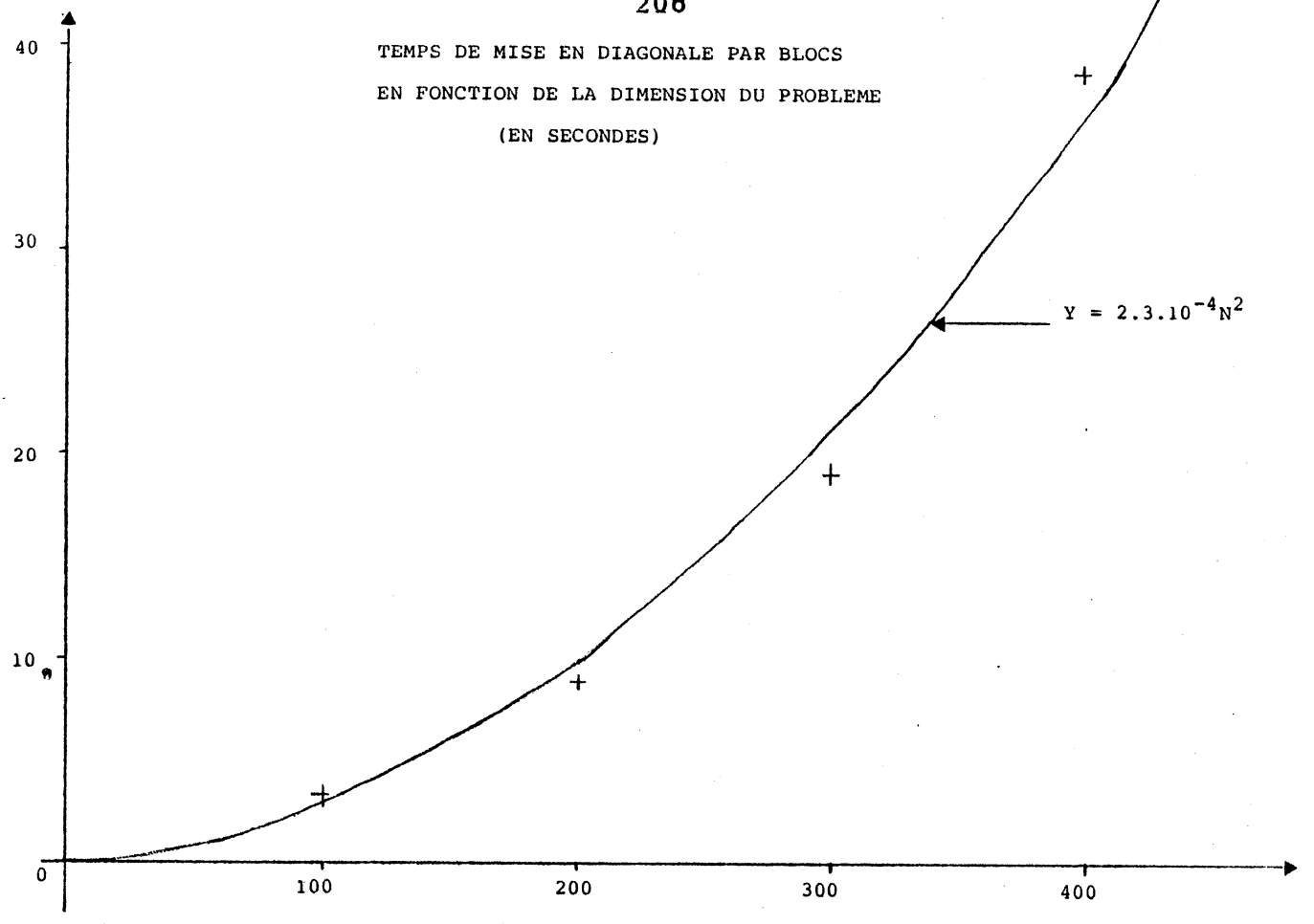
TEMPS PRIS PAR LA MISE EN DIAGONALE NON-NULLE

D'UNE MATRICE EN FONCTION DE LA DIMENSION DU PROBLEME

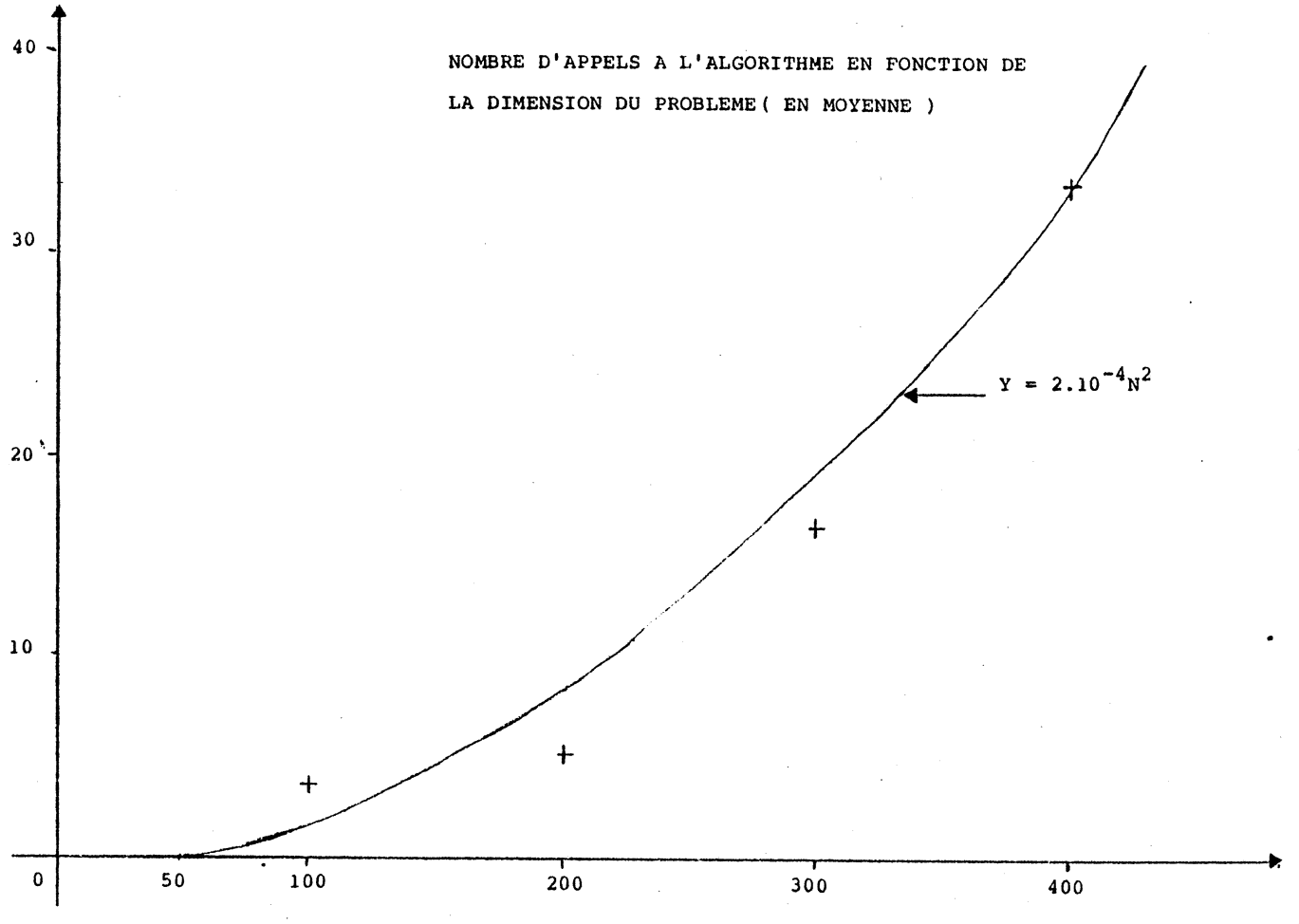
(EN SECONDES)



TEMPS DE MISE EN DIAGONALE PAR BLOCS
EN FONCTION DE LA DIMENSION DU PROBLEME
(EN SECONDES)



NOMBRE D'APPELS A L'ALGORITHME EN FONCTION DE
LA DIMENSION DU PROBLEME (EN MOYENNE)



Les enseignements qui découlent de ces graphes sont les suivants :

1) le temps pris par la transformation de la matrice en matrice diagonale par blocs avec bande est une fonction quadratique en n, ce qui corrobore l'idée que cet algorithme est en $O(\ell n)$ où ℓ est le nombre d'éléments non-nuls et n la dimension de la matrice traitée, vu que ℓ est une fonction linéaire de n

2) comme nous l'affirmions dans le chapitre II partie IV) le temps pris par l'heuristique est une fonction linéaire en ℓ/n, qui est ici plus ou moins constant (compris entre 3 et 5).

3) le nombre d'appels de l'algorithme de "cases admissibles" croît quadratiquement en n, ce qui pénalise le temps de mise en diagonale et fait que la courbe de ce dernier en fonction de la dimension de la matrice s'écarte de plus en plus de la droite initiale.

L'ensemble du traitement ne dépasse pas la minute pour une matrice 400×400 ayant un taux de remplissage compris entre 3/400 et 5/400 (donc aux alentours de 1%) et, en extrapolant, devrait être aux alentours de quatre minutes pour une matrice 1000×1000 ayant un taux de remplissage compris entre 3/1000 et 5/1000 ($\simeq$ 0.5%). Ce sont somme toute des temps raisonnables en égard à la taille du problème et du temps que prendra la résolution effective d'un système algébrodifférentiel de cette taille.

CONCLUSION

Diverses questions sont posées par les études réalisées dans ce chapitre et constituent à notre avis des voies de recherche intéressantes.

D'une part, le calcul de la taille du pas et de l'ordre ne concernent que les variables différentielles alors que dans les problèmes issus de la modélisation de circuits électriques, ce sont les variables algébriques qui sont largement majoritaires. Il est donc important d'arriver à obtenir des résultats plus précis sur l'erreur faite lors des approximations portant sur des variables algébriques.

Ensuite, les essais numériques réalisés portent sur des problèmes de petite dimension, alors que les systèmes qui devront être traités réellement sont tous de dimension importante. Notre étude comparative sur la méthode de Newton, la méthode de Newton à plusieurs niveaux et la méthode de surrelaxation doit donc être complétée.

Pour finir, il est sûrement intéressant d'étudier les méthodes de partionnement de matrices afin d'obtenir des algorithmes linéaires.

ANNEXE DU CHAPITRE III

démonstration de la proposition 1

Nous avons

$$f(y(t_n),z(t_n),\dot{z}(t_n),t_n) - f(y_n,z_n,g_{k_n}(z_n),t_n)$$
$$= \frac{\partial f}{\partial z}(y(t_n),z(t_n),\dot{z}(t_n),t_n) \cdot (z(t_n) - z_n)$$
$$+ \frac{\partial f}{\partial y\dot{z}}(y(t_n),z(t_n),z(t_n),t_n) \cdot \begin{pmatrix} y(t_n) - y_n \\ z(t_n) - z_n \end{pmatrix}$$
$$+ o(\|y(t_n)-y_n\|,\|z(t_n)-z_n\|,\|z(t_n) - g_{k_n}(z_n)\|)$$

Comme $f(y(t_n),z(t_n),\dot{z}(t_n),t_n) = \theta_{q+p}$ et

$f(y_n,z_n,g_{k_n}(z_n),t_n) = \theta_{q+p}$, nous avons :

$$\frac{\partial f}{\partial z}(y(t_n),z(t_n),\dot{z}(t_n),t_n) \cdot (z(t_n) - z_n)$$
$$+ \frac{\partial f}{\partial y\dot{z}}(y(t_n),z(t_n),\dot{z}(t_n),t_n) \cdot \begin{pmatrix} y(t_n) - y_n \\ z(t_n) - g_{k_n}(z_n) \end{pmatrix}$$
$$+o(\|y(t_n) - y_n\|,\|z(t_n)-z_n\|,\|z(t_n)-g_{k_n}(z_n)\|) = \theta_{q+p}$$

ou encore :

$$\begin{pmatrix} y(t_n) - y_n \\ \dot{z}(t_n) - g_{k_n}(z_n) \end{pmatrix}$$
$$= - \frac{\partial f}{\partial y\dot{z}}(y(t_n),z(t_n),z(t_n),t_n)^{-1} \cdot \frac{\partial f}{\partial z}(y(t_n),z(t_n),\dot{z}(t_n),t_n) \cdot$$
$$(z(t_n) - z_n) + o(\|y(t_n) - y_n\|,\|z(t_n) - z_n\|,\|z(t_n) - g_{k_n}(z_n)\|)$$

Or $g_{k_n}(z_n) = g_{k_n}(z_n) + \alpha_0(z_n - z(t_n))$

et $z(t_n) - g_{k_n}(z(t_n)) = \dfrac{z(t_n) - z_n^{pr}}{h_{k_n+1}} + o(h^{k_n+1})$

Donc, si nous posons

$$-\frac{\partial f}{\partial \dot{y}z}(y(t_n),z(t_n),\dot{z}(t_n),t_n)^{-1}\cdot\frac{\partial f}{\partial z}(y(t_n),z(t_n),\dot{z}(t_n),t_n)$$
$$=\begin{pmatrix} M_y^n \\ M_z^n \end{pmatrix}$$

où M_y^n est une matrice q-lignes et p-colonnes, et M_z^n une matrice p-lignes p colonnes, nous avons :

$$(1)\quad \begin{cases} y(t_n) - y_n + M_y^n(z(t_n) - z_n) = o(\|y_n - y(t_n)\|, \|z(t_n) - z_n\|) \\ \alpha_0(z(t_n) - z_n) + \dfrac{z(t_n) - z_n^{pr}}{h_{k_n+1}} + M_z^n(z(t_n) - z_n) \\ = o(\|y_n - y(t_n)\|, \|z_n - z(t_n)\|) \end{cases}$$

(H_1) nous permet de déduire de (1)

$$\begin{cases} z(t_n) - z_n = \dfrac{z_n^{pr} - z_n}{1 + h_{k_n+1}} + o(\|z(t_n) - z_n\|) \\ \|y(t_n) - y_n\| = O(\|z(t_n) - z_n\|) \end{cases}$$

□

démonstration de la proposition 2

Calculons $\frac{K_m'(hx)}{K_m(hx)}$

$$\frac{K_m'(hx)}{K_m(hx)} = \frac{1}{hx} + \sum_{i=2}^{m} \frac{1}{hx + h'_{i-1}} +$$

$$\frac{\frac{1}{hx^2} + \sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})^2}}{\frac{1}{hx} + \sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})}}$$

Clairement nous avons $\frac{K_m'(hx)}{K_m(hx)} > 0$ puisque toutes les quantités sont positives.

K_m est donc croissante sur l'intervalle $[h_{min}, h_{max}]$. D'où le fait que h_x^g soit unique.

D'autre part :

$$\frac{K_m'(hx)}{Km(hx)} < \frac{1}{hx} + \frac{m-1}{h_x} + \frac{\frac{1}{hx^2} + \frac{m-1}{hx^2}}{\frac{1}{hx}}$$

ou encore :

$$\frac{K_m'(hx)}{K_m(hx)} < \frac{2m}{hx}$$

En conclusion, nous avons l'encadrement suivant :

$$0 < \frac{K_m(h_x)}{hx} < K_m'(hx) < \frac{2m}{hx} K_m(hx)$$

Posons $G_m(hx) = \frac{K_m(hx)}{hx}$

$$\frac{G'_m(hx)}{G_m(hx)} = \sum_{i=2}^{m} \frac{1}{hx + h'_{i-1}} + \frac{\sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})^2}}{\sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})}}$$

Donc G'_m est positive.

$G_m(hx)$ est une fonction croissante en hx.

d'où l' encadrement suivant :

$$0 < \frac{h_{max} K_m(h_{min})}{2m h_{min} K_m(h_{max})} < \frac{h_{max}}{2m K_m(h_{max})} K'_m(hx) < 1$$

Posons $c = 1 - \frac{h_{max} K_m(h_{min})}{2m\, h_{min} K_m(hmax)}$

D'après ce qui précède, c appartient à l'intervalle]0,1[.

Si nous prenons $G(hx) = hx - \frac{K_m(hx)}{2mK_m(h_{max})}\left(\frac{K_m(hx)}{K_m(h)} - \alpha\right)$

nous avons la dérivée de G qui est strictement comprise entre 0 et c, et ce pour tout hx $\in [h_{min}, h_{max}]$

De plus

$$G(h_{max}) = h_{max}\left[1 - \frac{1}{2m} + \frac{\alpha K_m(h)}{2mK_m(h_{max})}\right]$$

Or $\frac{K_m(h_{max})}{K_m(h)} > \frac{K_m(hx)}{K_m(h)} = \alpha$ puisque K_m est croissante.

Donc $G(h_{max}) < h_{max}$

$$G(h_{min}) = h_{min}\left[\, 1 - \frac{K_m(h)h_{max}}{2mK_m(h_{max})h_{min}} \left(\frac{K_m(h_{min})}{K_m(h)} - \alpha \right) \right]$$

Donc $G(h_{min}) < h_{min}$

En définitive, $G([h_{min}, h_{max}]) \subset [h_{min}, h_{max}]$

Nous pouvons donc appliquer le théorème du point fixe.

Il existe hx^* tel que $G(hx^*) = hx^*$

De plus $hx^* = \lim_{p \to \infty} G^p(hx^0)$ □

démonstration de la proposition 3

Posons $$s_1(hx) = \frac{1}{hx} + \sum_{i=2}^{m} \frac{1}{hx + h'_{i-1}}$$

$$s_2(hx) = \frac{1}{hx^2} + \sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})^2}$$

$$s_3(hx) = \frac{1}{hx^3} + \sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})^3}$$

$$P(hx) = hx \prod_{i=2}^{m} (hx + h'_{i-1})$$

Nous avons :

$$P'(hx) = \prod_{i=2}^{m} (hx + h'_{i-1}) + hx\left(\sum_{i=2}^{m} \prod_{\substack{i=2 \\ i \neq j}}^{m} (hx + h'_{i-1}) \right)$$

$$= P(hx) \left(\frac{1}{hx} + \sum_{i=2}^{m} \frac{1}{hx + h'_{i-1}} \right)$$

$$= P(hx)\, s_1(hx)$$

$$s'_1(hx) = -\frac{1}{hx^2} - \sum_{i=2}^{m} \frac{1}{(hx + h'_{i-1})^2}$$

$$= -s_2(hx)$$

D'où, puisque $K_m(hx) = \frac{P(hx)}{s_1(hx)}$:

$$K'_m(hx) = \left(\frac{P(hx)}{s_1(hx)} \right)'$$

$$= \frac{P'(hx)\, s_1(hx) - s'_1(hx)\, P(hx)}{[\, s_1(hx)\,]^2}$$

$$= K_m(hx) \quad \frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)}$$

De plus

$$K''_m(hx) = K'_m(hx) \quad \frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)}$$

$$+ \quad K_m(hx) \left(\frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)} \right)'$$

$$= K_m(hx) \left(\frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)} \right)^2$$

$$+ \quad \left(\frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)} \right)'$$

Or :

$$\left(\frac{[s_1(hx)]^2 + s_2(hx)}{s_1(hx)} \right)'$$

$$= \frac{-2\, s_1(hx)(s_1(hx)s_2(hx)+s_3(hx)) + s_2(hx)([s_1(hx)]^2+s_2(hx))}{[s_1(hx)]^2}$$

$$= \frac{[s_2(hx)]^2 - [s_1(hx)]^2\, s_2(hx) - 2\, s_1(hx)\, s_3(hx)}{[s_1(hx)]^2}$$

Donc

$$\frac{K_m''(hx)}{K_m'(hx)} = \frac{[s_1(hx)]^4+[s_1(hx)]^2 s_2(hx)+2[s_2(hx)]^2-2s_1(hx)s_3(hx)}{[s_1(hx)]^2}$$

Or nous avons :

$$[s_1(hx)]^2 > s_2(hx)$$

$$[s_1(hx)]^3 > s_3(hx)$$

$$s_1(hx)\, s_2(hx) > s_3(hx)$$

Donc :

$$[s_1(hx)]^4 > s_1(hx)\, s_3(hx)$$

$$[s_1(hx)]^2 s_2(hx) > s_1(hx)\, s_3(hx)$$

D'où le fait que $K_m''(hx) > 0$

la dérivée de l'opérateur de Newton s'écrit :

$$G'(hx) = \frac{K_m(hx)\, K_m''(hx)}{[K_m'(hx)]^2}$$

G'(hx) est donc strictement positive.

$$G'(hx) = \frac{[s_1(hx)]^4 + [s_1(hx)]^2 s_2(hx) + 2\,[s_2(hx)]^2 - 2\,s_1(hx)\,s_3(hx)}{([s_1(hx)]^2 + s_2(hx))^2}$$

$$= 1 + \frac{[s_2(hx)]^2 - [s_1(hx)]^2 s_2(hx) - 2\,s_1(hx)\,s_3(hx)}{([s_1(hx)]^2 + s_2(hx))^2}$$

$[s_2(hx)]^2 - [s_1(hx)]^2\, s_2(hx)$ est strictement négatif. Donc :

$$G'(hx) < 1 - \frac{2\,s_1(hx)\,s_3(hx)}{([s_1(hx)]^2 - s_2(hx))^2}$$

ou encore $\left(s_2(hx) < s_1(hx)^2\right)$:

$$G'(hx) < 1 - \frac{s_1(hx)\,s_3(hx)}{2\,s_1(hx)^4}$$

$$\Rightarrow G'(hx) < 1 - \frac{s_3(hx)}{2\,s_1(hx)^3}$$

Nous avons

$$s_3(hx) < \frac{1}{(h_{max})^3}$$

$$s_1(hx) < \left(\sum_{i=1}^{m} \frac{1}{i}\right) \frac{1}{h_{min}}$$

Posons $\lambda_m = \sum_{i=1}^{m} \frac{1}{i}$

Nous pouvons alors écrire :

$$G'(hx) < 1 - \frac{(h_{min})^3}{2\,\lambda_m\,(h_{max}^3)}$$

Comme λ_m est supérieur à 1, il existe c appartenant strictement à]0,1[tel que $G'(hx) < c$

Nous avons vu précédemment que $G'(hx) > 0$.

Donc :

$\exists c \in]0,1[\ / \ 0 < G'(hx) < c$

De plus :

$$G(h_{max}) = h_{max} - \frac{K'_m(h_{max})}{K_m(h)\left(\frac{K_m(hmax)}{K_m(h)} - \alpha\right)}$$

Or

$$\frac{K_m(h_{max})}{K_m(h)} \qquad \frac{K_m(h_x^s)}{K_m(h)}$$

Comme de plus $K'_m(h_{max}) > 0$, alors $G(h_{max}) < h_{max}$.

$$G(h_{min}) = hmin - \frac{K'_m(h_{min})}{K_m(h)\left(\frac{K_m(h_{min})}{K_m(h)} - \alpha\right)}$$

$$\frac{K_m(h_{min})}{K_m(h)} < \frac{K_m(h_x^s)}{K_m(h)}$$

Donc $G(h_{min}) > h_{min}$

G étant croissante, nous avons alors

$G([h_{min}, h_{max}]) \subset [h_{min}, h_{max}]$

Le théorème du point fixe s'applique donc ici.

La suite $hx^p = G^p(hx^0)$ converge et sa limite est telle que $G(hx^*) = hx^*$ ou encore $\frac{K_m(hx^*)}{K_m(h)} = \alpha$

CONCLUSION GENERALE

En entamant ce travail, nous avions pour simple point de départ la volonté de mettre au point des outils numériques stables permettant la prise en compte de la réalité physique du problème proposé (notemment les phénomènes de latence, qui ne sont que la traduction mathématique du parcours dans un circuit d'un signal électrique), sans savoir de manière précise ce que recouvrait la notion de stabilité ou la manière dont la latence pouvait être caractérisée. Aujourd'hui, nous sommes à même de proposer des outils numériques répondant au problème posé, du moins en partie. Notre travail est évidemment incomplet et les lacunes qu'il présente ont été soulignées tout au long des pages de cet ouvrage. Nous pensons néanmoins que nous avons contribué à dégager des voies de recherche dans des domaines importants comme ceux concernant la stabilité numérique ou l'évaluation de l'erreur réalisée.

La vitalité des travaux menés dans le domaine de la résolution des systèmes différentiels ou algébrodifférentiels est telle qu'il nous semble difficile de ne pas citer les développements les plus récents dans ce domaine. Nous en citerons deux :

D'abord les travaux de Saad[Saad, 1983] qui en appliquant des méthodes dérivées des méthodes de gradient conjugué a obtenu des résultats remarquables, tant du point de vue de la stabilité que de celui du coût des méthodes utilisées.

Très nouvelle est la méthode de Lelarasmee et Sangiovanni-Vincentelli connue sous le nom de RELAX [Lelar, 1982], et qui est basée sur la méthode de relaxation par vagues.

Il nous semble que certains travaux portant sur la relaxation chaotique [Robert, 1975][Mah, 1978][Abt, 1978] pourraient trouver là une application originale.

BIBLIOGRAPHIE

[Abt, 1977] A. ABTROUN
"Recherche d'une permutation optimale des variables dans la méthode itérative de Gauss-Seidel."
THESE DE 3EME CYCLE, GRENOBLE, Mai 1977.

[Arn, 1974] V. ARNOLD
"equations differentielles ordinaires."
EDITIONS MIR, MOSCOU, 1974.

[Arn, 1978] V. ARNOLD
"chapitres supplémentaires de la théorie des équations différentielles ordinaires."
EDITIONS MIR, MOSCOU, 1978.

[Bju, 1970] G. BJUREL, G. DAHLQUIST, B. LINDBERG, S. LINDE, L. ODEN
"Survey of stiff ordinary differential equations."
rapport NA 70-11, departement of information processing computer science, the royal institute of technology, STOCKHOLM, SUEDE, Novembre 1970.

[Bor, 1983] D. BORRIONE,M. HUMBERT, C. LE FAOU
"projet CASCADE: mécanisme de simulation mixte hierarchisée."
RR n° 354, I.M.A.G., GRENOBLE, Mars 1983.

[Bray, 1972] R.K. BRAYTON, F.G. GUSTAVSON, G. G. HACHTEL
"A new efficient algorithm for solving differential algebraic systems using implicit backward differentiation formulas".
IEEE PROCEEDINGS, vol. 60, p. 98-108, Janvier 1972.

[Bre, 1977] C. BREZINSKI
"Acceleration de la convergence en analyse numerique."
lecture notes in mathematics, 584, SPRINGER-VERLAG BERLIN HEIDELBERG NEW-YORK, 1977.

[Chilov, 1975] G. CHILOV
"Analyse mathématique:fonctions de plusieurs variables réelles (1ere et 2eme partie)"
EDITIONS MIR, MOSCOU, 1975.

[Chu, 1975] CHUNG-WEN HO, A. E. RUEHLI, P. A. BRENNAN
"The modified nodal approach to network analysis."
IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, vol CAS-22, no 6 , p 504-509, Juin 1975.

[Crou, 1980] M. CROUZEIX
"Différentes notions de stabilité utiles pour la résolution numérique des équations différentielles raides."
GROUPEMENT POUR L'AVANCEMENT DES METHODES NUMERIQUES POUR L'INGENIEUR(GAMNI), bulletin n° 15, PARIS, Juillet 1980.

[Dah, 1963] G.G. DAHLQUIST
A special stability problems for linears multistep methods."
BIT 3, p. 27-43, 1963.

[Dah, 1980] G. DAHLQUIST, W. LINIGER, O. NEVANLINNA
"Stability of two-steps methods for variable integration steps"
Report RC 8494, IBM Research center, Yorktown Heights, NEW-YORK 10566, 1980

[Gam, 1977] D. GAMBART
"Implantation et optimisation de la méthode de Brayton dans un programme de simulation de circuits électroniques."
THESE DE DOCTEUR DE SPECIALITE INFORMATIQUE(METHODES), TOULOUSE, Novembre 1977.

[Gam, 1980] D. GAMBART-DUCROS, G. MARAL
"An efficient algorithm reduces computing time in solving a system of stiff ordinary differential equations."
IEEE TRANSACTIONS ON CICUITS AND SYSTEMS, vol CAS-27, n° 9, p. 747-755, Septembre 1980.

[Gear, 1967] C. W. GEAR
"Numerical integration of stiff ordinary differential equations."
report n° 221, department of computer science, university of ILLINOIS, URBANA, ILLINOIS, Janvier 1967

[Gear, 1969] C.W. GEAR
"The automatic integration of stiff ordinary differential equations."
A.J.H. MORREL NORTH-HOLLAND PUBLISHING COMPANY, AMSTERDAM, p. 187-193, 1969.

[Gear, 1971, a)] C. W. GEAR
"The automatic integration of ordinary differential equations."
communication of A.C.M., vol. 14 n° 3, p. 176-179, Mars 1971.

[Gear, 1971, b)] C. W. GEAR
"Numerical initial value problems in ordinary differential equations."
PRENTICE-HALL series in automatic computation, inc. ENGLEWOOD CLIFFS, NEW JERSEY, 1971.

[Hac, 1971] G. HACQUES
"Algorithmique numérique."
collection U, ARMAND-COLLIN, 1971.

[Hach, 1971] G. D. HACHTEL, R. K. BRAYTON, F. G. GUSTAVSON
"The sparse tableau approach to network analysis and design."
IEEE TRANSACTIONS ON CIRCUIT THEORY, vol CT-18, p 101-112, Janvier 1971.

[Hen, 1962] P. HENRICI
"Discrete variables methods in ordinary differential equations."
JOHN WILEY AL SOUS, inc., NEW-YORK-LONDON, 1962.

[Hou, 1979] N. HOUBAK, P. G THOMSEN
"A fortran subroutine for the solution of large systems of stiff ODE's with sparse jacobians."
rapport n° NI-79-02, numerisk institut, DANMARKS TEKNISK HØJSKOLE, 2800 LYNGBY, DANEMARK.

[Kar, 1975] V. KARMANOV
"Programmation mathématique"
éditions MIR, MOSCOU, 1977.

[Ker, 1981] J. KERVORIAN, J. D. COLE
"Perturbation methods in applied mathematics."
applied mathematical sciences 34, SPRINGER-VERLG, NEW-YORK HEIDELBERG BERLIN, 1981

[Jac, 1977] C. JACQUEMARD
"Contribution à l'étude d'algorithmes de relaxation a convergence monotone."
THESE DE DOCTEUR EN MATHEMATIQUES, MENTION MATHEMATIQUES PURES, BESANCON, Mai 1977

[Le Faou, 1982] C. LE FAOU
"CASCADE : système de C.A.O. intégré pour ensembles logiques et électroniques. Introduction du niveau électrique"
R.R. N° 308, IMAG, GRENOBLE, 1982.

[L-F-Arn, 1977] J. LELONG-FERRAND, .M. ARNAUDIES
"cours de mathematiques(tomes 1 à 4)."
DUNOD UNIVERSITE, BORDAS, PARIS, 1977.

[Lelar, 1982] E. LELARASMEE, A.SANGIVANNI-VINCENTELLI
"RELAX: A new circuit simulator for large scale MOS integrated circuits."
IEEE, 19th design automation conference, article 38.2, 1982.

[Lef, 1957] S. LEFSCHETZ
"Differential equations:geometric theory."
pure and applied mathematics, volume VI,INTERSCIENCE PUBLISHERS, INC.,NEW-YORK, 1957.

[Liap, 1907] M.A. LIAPUNOFF
"Problème général de la stabilité du mouvement."
PRINCETON UNIVERSITY PRESS, PRINCETON, 1947.
(cet ouvrage est la traduction du russe de l'original de Liapunov paru dans "Annales de la Faculté des Sciences de Toulouse, deuxième série,volume 9, 1907.)

[Lin, 1982] W. LINIGER
"Accurate multistep methods for smooth stiff problems."
Thomas J. Watson Research Center, Yorktown heights, NEW-YORK, Juillet 1982

[Mah, 1977] Z. MAHJOUB
"Expérimentation de stratégies itératives chaotiques sur des problèmes de point fixe à grand nombre de variables."
THESE DE DOCTEUR-INGENIEUR, GRENOBLE, Mai 1977.

[Mir, 1975] W. L. MIRANKER
"The computationnal theory of stiff differential equations."
publication mathématique d'ORSAY, PARIS XI, UER MATHEMATIQUES, 91407 ORSAY, 1975.

[Mult, 1980] Series 60(level 68) MULTICS
MULTICS Pocket Guide Commands and active Functions, software
CII HONEYWELL BULL, Avril 1980.

[NAG] NAG CENTRAL OFFICE
MAYFIELD HOUSE
256 Banbury road, OXFORD OX2 7DE (UK), GRANDE BRETAGNE.

[Nord, 1962] A. NORDSIECK
"On numerical integration of differential equations."
MATH. COMP., vol 16, p. 22-49, 1962.

[Rabbat, 1979] N.B. RABBAT, A.L. SANGIOVANNI-VINCENTELLI, HSUEJ Y. HSIEH
"A multilevel Newton algorithm with macromodeling and latency for the analysis of large scale non-linear circuits in the time domain."
IEEE CIRCUITS AND SYSTEMS, vol. CAS-26, n° 9, 1979.

[Rey, 1978] J. C. REYNAUD
" IMAG3 : un système de simulation et d'optimisation de circuits électroniques."
THESE DE 3ème CYCLE, IMAG, GRENOBLE, Mai 1978.

[Robert, 1975] F. ROBERT, M. CHARNAY, F. MUSY
"Itérations chaotiques série-parallèle pour des équations non-linéaires de point fixe."
aplikace matematiky, 20, CESKOSLOVENSKA akademia ved, PRAGUE, 1975.

[Rou, 1973] N. ROUCHE, J. MAWHIN
"Equations différentielles ordinaires, tome 2 : stabilité et solutions périodiques."
MASSON ET CIE, PARIS, 1973.

[Rueh, 1981] A. E. RUEHLI
"Survey of analysis, simulation and modeling for large scale logic circuits."
IBM T. J. WATSON RESEARCH CENTER, post office box 218, YORKTOWN HEIGHTS, NY 10598, 1981.

[Saad, 1983] Y. SAAD
"Méthodes numériques pour la résolution de problèmes matriciels de grandes dimensions"
THESE D'ETAT, I.M.A.G. , GRENOBLE, Septembre 1983

[Sak, 1979] M. SAKAROVITCH
"Techniques mathématiques de la recherche opérationnelle : III Optimisation dans les réseaux."
Polycopié E.N.S.I.M.A.G. , Septembre 1979.

[Sang, 1977] A. SANGIOVANNI-VINCENTELLI, LIKUAN CHOU, LEON O. CHUA
"An efficient heuristic cluster algorithm for tearing large scale networks."
IEEE CIRCUITS AND SYSTEMS, vol. 24, n° 12, 1977.

[Sang, 1980] A. SANGIOVANNI-VINCENTELLI
"Techniques for the time domain analysis of LSI circuits."
R.R. IBM THOMAS J. WATSON RESEARCH CENTER, post office box 218, YORKTOWN HEIGHTS, NEW-YORK 10598.

[Van Bok, 1975] W.M.G. Van Bokhoven
"Linear implicit differentiation formulas of variable step and order."
IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS, vol. 22, n° 2, 1975.

[Wan, 1977] G. WANNER
"On the integration of stiff differential equations."
G. WANNER, section de mathématiques, 2-4 rue du Lièvre, case postale 124, 1211 GENEVE, SUISSE, 1977.

[Zla, 1979, a)] Z. ZLATEV, P. G. THOMSEN
"An automatic differential equations solver based on linear multistep methods."
report n° NI-79-01, numerisk institut, DANMARKS tekniske højskole 2800 LYNGBY, DANEMARK, Mars 1979.

[Zla, 1979, b)] Z. ZLATEV, P. G. THOMSEN
Two-parameters families of predictor-corrector methods for the solution of ordinary differential equations."
BIT 19, p. 503-517, institute for numerical analysis, technical university of DANMARK, 2800 LYNGBY, DANEMARK, 1979.

DERNIERE PAGE D'UNE THESE

3È CYCLE, DOCTEUR INGÉNIEUR OU UNIVERSITÉ

Vu les dispositions de l'arrêté du 16 avril 1974,

Vu les rapports de M. ..F. ROBERT .....

M. .....,.....

M. .....BONA Mariano..... ..... est autorisé

à présenter une thèse en vue de l'obtention du grade de DOCTEUR - Ingénieur

.....

Grenoble, le 11 OCT. 1983

Le Président de l'Université Scientifique
et Médicale

M. TANCHE

D. Tanche