

Jacques Louis Lions (Ed.)

Numerical Analysis of Partial Differential Equations

44

Ispra, Italy 1967



Jacques Louis Lions (Ed.)

Numerical Analysis of Partial Differential Equations

Lectures given at a Summer School of the
Centro Internazionale Matematico Estivo (C.I.M.E.),
held in Ispra (Varese), Italy,
July 3-11, 1967

 Springer



FONDAZIONE
CIME
ROBERTO CONTI

C.I.M.E. Foundation
c/o Dipartimento di Matematica “U. Dini”
Viale Morgagni n. 67/a
50134 Firenze
Italy
cime@math.unifi.it

ISBN 978-3-642-11056-6 e-ISBN: 978-3-642-11057-3
DOI:10.1007/978-3-642-11057-3
Springer Heidelberg Dordrecht London New York

©Springer-Verlag Berlin Heidelberg 2010
Reprint of the 1st Ed. C.I.M.E., Ed. Cremonese, Roma 1968
With kind permission of C.I.M.E.

Printed on acid-free paper

Springer.com

CENTRO INTERNAZIONALE MATEMATICO ESTIVO

(C. I. M. E.)

2^o Ciclo - Ispra (Varese) dal 3-11 Luglio 1967

NUMERICAL ANALYSIS OF PARTIAL EQUATIONS

Coordinatore : J. L. LIONS

S. ALBERTONI : Alcuni metodi di calcolo nella teoria della diffusione dei neutroni.	pag. 1
I. BABUSKA : Optimization and numerical stability in computations "	25
J. H. BRAMBLE : Error estimates in elliptic boundary value problems "	69
G. CAPRIZ : The numerical approach to hydrodynamic problems "	109
A. DOU : Energy inequalities in an elastic cylinder	161
DUPONT TODD : On the existence of an iterative method for the solution of elliptic difference equation with an improved work estimate.	pag. 175
J. DOUGLAS and J. R. CANNON: The approximation of harmonic and p-bolic functions on Half-spaces from interior data	pag. 193
B. E. HUBBARD: Error estimates in the fixed Membrane problem "	231
K. JORGENS : Calculation of the spectrum of a Schrodinger operator"	245
A. LASOTA : Contingent equations and boundary value problems "	255
J. L. LIONS : Reduction a des problèmes du type Cauchy Kowaleska"	267
J. L. LIONS : Problèmes aux limites non homogenes a données irregulières; une methode d'approximation	283
J. L. LIONS : Remarques sur l'approximation regularisée de problèmes aux limites.	pag. 293
W. V. PETRYSHYN: Iterative construction of fixed points of contractive type mappings in Banach spaces	" 307
W. V. PETRYSHYN: On the approximation-solvability of nonlinear functional equations in normed linear spaces	pag. 341
P. A. RAVIART: Approximation des equations d'evolution par des methodes variationnelles	pag. 357
M. SIBONY et H. BREZIS : Methodes d'approximation et d'iteration pour les operateurs monotones	pag. 407
V. THOMEE : Some topics in stability theory for partial difference operators.	pag. 441

CENTRO INTERNAZIONALE MATEMATICO ESTIVO

- S. ALBERTONI -

" ALCUNI METODI DI CALCOLO NELLA TEORIA DELLA DIFFUSIONE DEI
NEUTRONI "

Corso tenuto ad Ispra dal 3-11 Luglio 1967

ALCUNI METODI DI CALCOLO NELLA TEORIA A MULTIGRUPPI DELLA DIFFUSIONE DEI NEUTRONI

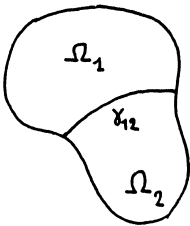
- S. Albertoni -

PARTE 1^a - Teoria stazionaria a multigruppi, (1,2)

§ 1 - Notazioni e Problemi

Sia $\Omega \subset \mathbb{R}^n$ un aperto limitato e $x = (x_1, x_2, \dots, x_n) \in \Omega$. Nel seguito considereremo gli spazi $L^2(\Omega)$, $H^1(\Omega)$, $H_0^1(\Omega)$, e se g è intero > 0 , $(L^2)^g, (H^1)^g, (H_0^1)^g$ saranno i g -prodotti diretti di L^2, H^1, H_0^1 .

Volendo considerare principalmente problemi di "trasmissione" supporremo $\Omega = \bigcup_1^r \Omega_i$, e siano $\Gamma = \partial\Omega$, $\Gamma_i = \partial\Omega_i$ rispettivamente le frontiere di Ω, Ω_i : $\gamma_{rs} = \partial\Omega_r \cap \partial\Omega_s$ sono le "interfacce", e cioè le parti di frontiera comuni a $\bar{\Omega}_r, \bar{\Omega}_s$.



Infine ν_r sia la normale, diretta verso l'esterno, di $\partial\Omega_r$. Premesso ciò supponiamo assegnate le funzioni reali: $D_{pq}^i(x), A^i(x), C^i(x) (c^1 \equiv 0), B^i(x), f^i(x) \cdot (p, q = 1, 2, \dots, n; i = 1, 2, \dots, g)$.

I problemi da risolvere sono due, e cioè:

Problema A - Trovare la soluzione $u(x)$ del sistema:

$$-\sum_{p,q}^n \frac{\partial}{\partial x_p} (D_{pq}^i \frac{\partial u^i}{\partial x_q}) + A^i u^i = c^i u^{i-1} + \frac{1}{\lambda} \omega^i(x) + f^i$$

(1)

$$\omega^i(x) = x_i \sum_{j=1}^g \nu_j B^j(x) u^j; (i=1, 2, \dots, g),$$

soddisfacente a $u|_r = 0$, e alle condizioni di "trasmissione"

$$u^r = u^s$$

(1)'

$$\frac{du^r}{dv_r} = - \frac{du^s}{dv_s}$$

valide per $x \in \gamma_{rs}$, essendo $\frac{d}{dv_r}$ la derivata conormale a $\partial\Omega_r$ e u^r, u^s le restrizioni di u a Ω_r e Ω_s rispettivamente.

Osservazione I - Dal punto di vista fisico le (1) forniscono la distribuzione stazionaria del flusso neutronico entro un reattore Ω (composto da regioni di materiali diversi Ω_i) ripartito in g gruppi di energia decrescente (u^i è il flusso dei neutroni di i^{ma} energia) ove f^i sono le sorgenti neutroniche esterne e:

- D^i un "coefficiente" di diffusione
- A^i un "coefficiente" di assorbimento
- C^i un "coefficiente" di rimozione dal gruppo $i-1$ al gruppo i
(Ecco perchè $C^1=0$)
- B^i un "coefficiente" di fissione.

Osservazione II - E' utile per il seguito dare più esplicitamente la struttura formale del sistema (1):

$$\begin{aligned} & - \sum_{p,q=1}^n \frac{\partial}{\partial x_p} (D_{pq}^1 \frac{\partial u^1}{\partial x_q}) + A^1 u^1 = \frac{1}{\lambda} \omega^1 + f^1 \\ (2) \quad & - \sum_{p,q=1}^n \frac{\partial}{\partial x_p} (D_{pq}^2 \frac{\partial u^2}{\partial x_q}) + A^2 u^2 = C^2 u^1 + \frac{1}{\lambda} \omega^2 + f^2 \\ & \text{-----} \\ & - \sum_{p,q=1}^n \frac{\partial}{\partial x_p} (D_{pq}^g \frac{\partial u^g}{\partial x_q}) + A^g u^g = C^g u^{g-1} + \frac{1}{\lambda} \omega^g + f^g \end{aligned}$$

Ipotesi sul problema A:

- 1) Positività dei coefficienti D^i, A^i, B^i, C^i (supposti misurabili e limitati in Ω) e cioè: $A^i(x) \geq a > 0, B^i(x) \geq b > 0, C^i(x) \geq c > 0$ ($i=2,3,\dots$), $D_{pq}^i(x) \geq d > 0$; χ_i, ν sono coefficienti numerici > 0 .

Inoltre si suppone $C^1(x) \equiv 0$.

- 2) Ellitticità uniforme e cioè: $\forall \{\lambda_p; p=1,2,\dots,n, \lambda_p \in \mathbb{R}_1, x \in \bar{\Omega}\}$ esistono $m, M > 0$ tali che:

$$m \sum_{p=1}^n \lambda_p^2 \leq \sum_{p,q=1}^n D_{pq}^i \lambda_p \lambda_q \leq M \sum_{p=1}^n \lambda_p^2$$

per ogni $i=1,2,\dots,g$. Inoltre Ω sarà abbastanza regolare da garantire che tutte le operazioni di traccia abbiano un senso.

Problema B - Trovare il massimo autovalore $> 0, \lambda_0$, e la corrispondente autosoluzione positiva $u_0(x) > 0$, per il Problema A omogeneo, e cioè per $f^i=0$ e $u|_{\Gamma}=0$; tal caso fisicamente corrisponde ad una ripartizione di neutroni autosostenentesi (reattore critico).

§ 2 - Soluzione del problema A

Cercheremo u in $H_0^1(\Omega) (u|_{\Gamma}=0)$: le condizioni di trasmissione appariranno come "condizioni naturali", automaticamente soddisfatte nella formulazione variazionale, come è ben noto nel caso d'una equazione sola ($g=1$). Per $g>1$ il Pb non è autoaggiunto. Si è allora preferito seguire un ragionamento elementare basato sul fatto che $C^1(x) \equiv 0$. Introdotte le forme lineari:

$$\begin{aligned}
 a(u,v) &= \sum_{i=1}^g \int_{\Omega} \left(\sum_{p,q=1}^n D_{pq}^i \frac{\partial u^i}{\partial x_p} \frac{\partial v^i}{\partial x_q} + A^i u^i v^i \right) d\Omega \\
 (3) \quad b(u,v) &= \sum_{i=1}^g \int_{\Omega} \omega^i(x) v^i d\Omega \quad ; \quad \forall u,v \in (H_0^1)^g \\
 c(u,v) &= \sum_{i=1}^g \int_{\Omega} c^i(x) u^{i-1} v^i d\Omega
 \end{aligned}$$

siano A, B, C i corrispondenti operatori (matriciali) generati (nel senso delle distribuzioni). Per le ipotesi fatte sui coefficienti essi risultano definiti su tutto $(H_0^1)^g$ con codominio $C(H^{-1})^g$. Gli elementi di matrice corrispondenti sono:

$$\begin{aligned}
 A_{ij} &= \delta_{ij} A^i(x) \\
 (4) \quad B_{ij} &= x_i v B^j(x) \quad ; \quad (i,j=1,2,\dots,g) \\
 C_{ij} &= \delta_{i,j+1} C^j(x)
 \end{aligned}$$

ove:

$$(4)' \quad A^i \cdot = - \sum_{p,q=1}^n D_{pq}^i \frac{\partial}{\partial x_p} \left(D_{pq}^i \frac{\partial}{\partial x_q} \cdot \right) + A^i \cdot \quad ; \quad (i=1,2,\dots,g)$$

Introdotti A, B, C il Problema A si riconduce alla risoluzione dell'equazione in u :

$$(5) \quad A_{\mu} u \equiv (A - \mu B - C)u = f \quad ; \quad u \in (H_0^1)^g \quad ; \quad \mu = \frac{1}{\lambda}$$

Il procedimento esistenziale è allora il seguente:

Se $\mu=0$ allora la:

(6) $A_0 u \equiv (A - C)u = f$ è costituita (si osservi la struttura delle (2)) da g equazioni disaccoppiate ($C^1 \equiv 0$) del tipo:

$$(7) \quad A^i u^i = - \sum_{p,q=1}^n D_{pq}^i \frac{\partial}{\partial x_p} \left(D_{pq}^i \frac{\partial u^i}{\partial x_q} \right) + A^i u^i = C^i u^{i-1} + f^i$$

Assumendo $f \in (H^{-1})^g$ (o più semplicemente $(L^2)^g$) per le ipotesi

1) e 2)) esiste allora un operatore G_0^1 di Green^(°) che ci fornisce $u^1 = G_0^1 f^1$ (G_0^1 è un isomorfismo di $H^{-1}(\Omega)$ su $H_0^1(\Omega)$). Trovato u^1 , u^2 è fornita da: $u^2 = G_0^2(C^2 u^1 + f^2) = G_0^2(C^2 G_0^1 f^1 + f^2)$, e in generale si ha: $u^i = G_0^i(C^i G_0^{i-2} f^{i-1} + f^i) \in H_0^i(\Omega)$.

Ne consegue l'esistenza ($\mu=0$) di un operatore matriciale di Green, G_0 , per il nostro problema, che realizza un isomorfismo di $(H^{-1})^{\mathcal{E}}$ su $(H_0^1)^{\mathcal{E}}$, e $u = G_0 f \in (H_0^1)^{\mathcal{E}}$ soddisfa "naturalmente" alle condizioni di trasmissione.

Il caso $\mu \neq 0$ si tratta riducendo la (5) nella forma:

$$(9) \quad u = G_0 f + G_0(Bu) = g + \mu Tu \quad ; \quad (g = G_0 f \quad ; \quad G_0 B = T)$$

Essendo G_0 un operatore limitato da $(L^2)^{\mathcal{E}} \rightarrow (H_0^1)^{\mathcal{E}}$, B_{ij} moltiplicatori in $L^2(\Omega)$, e l'immersione di $(H_0^1)^{\mathcal{E}} \rightarrow (L^2)^{\mathcal{E}}$ compatta, allora T è compatto da $(H_0^1)^{\mathcal{E}} \rightarrow (H_0^1)^{\mathcal{E}}$, e pertanto il Problema A è ricondotto ad un classico Problema di Riesz-Fredholm. Lo spettro puntuale $\{\mu_j\}$ ammette l'unico punto di accumulazione all'infinito, e i $\lambda_j = \frac{1}{\mu_j}$ si accumulano solo verso zero.

§ 3 - Soluzione del Problema B

a) In tal caso si ha:

$$(10) \quad (A-C)u = Lu = \mu Bu \quad ; \quad \lambda Lu = Bu$$

(°) $G_0 = (A-C)^{-1} = A_0^{-1}$

Per risolvere la (10) si è trovata una rappresentazione a nucleo (funzione di Green) dell'operatore L^{-1} . Se $g=1$ la funzione di Green $G^1(x,y)$ è stata trovata, nel caso dei coefficienti discontinui, e per la 1^a volta, da Stampacchia^(°) come soluzione di:

$$(11) \quad A^1 G^1(x,y) = \delta_{x,y} \quad (\delta_{x,y} \text{ misura di Dirac in } (x=y))$$

Allora

$u = \int_{\Omega} G^1(x,y) f(y) dy$ è soluzione di $A^1 u = f$ e si ha

$$\int_{\Omega} G^1(x,y) A^1 u(y) dy = u(x) \quad ; \quad \forall u \in H_0^1(\Omega)$$

$$(10) \quad G^1(x,y) > 0 \quad , \quad G^1(x,y) = G^1(y,x) \quad ;$$

$$G^1(x,y) = \frac{M^1(x,y)}{(x-y)^{n-2}} \quad ; \quad \forall (x,y) \in \Omega' \text{ compatto } \subset \Omega \quad .$$

Osservazione - Questo fatto di $\Omega' \subset \Omega$ non ci permette (benchè plausibile) di concludere che $G^1 \in L^2(\Omega \times \Omega)$ (e i suoi iterati).

Allora sfruttando il fatto che per l'operatore L si possono (vedi (6)) fare dei ragionamenti per singole equazioni (essendo queste disaccoppiabili) si può costruire subito una matrice "formale" di Green di elementi $\mathfrak{L}_{ik}^{-1}$:

$$\mathfrak{L}_{ik}^{-1} = \begin{cases} 0 & ; \quad i < k \\ \int_{\Omega} dy \, G^i(x,y) & ; \quad i = k \\ \mathfrak{L}_{ii}^{-1} C^i(x) \mathfrak{L}_{i-1,k}^{-1} & ; \quad i > k \end{cases}$$

(°) Si osservi che Habetler e Martino (8) già nel 1958 avevano considerato il Problema B assumendo però formalmente l'esistenza della funzione di Green.

Ad esempio:

$$(11) \quad \ell_{21}^{-1}(\cdot) = \int_{\Omega} dy G^2(x,y) C^2(y) \int_{\Omega} dz G^1(x,y)(\cdot) = \int_{\Omega} dy G^{21}(x,y)(\cdot)$$

se i G^i fossero $\in L^2(\Omega \times \Omega)$.

Si verifica poi subito che $L^{-1} = \{\ell_{ik}^{-1}\}$ è tale che $L^{-1}Lu = u, u \in (H_0^1)^E$, e pertanto $L^{-1} = \{\ell_{ik}^{-1}\}$ è la matrice di Green del problema $Lu = \phi$.

Osservazione - Per le ipotesi di positività fatte sulle C^i l'operatore L^{-1} lascia invariato il cono, in $(L^2)^E$, dei vettori >0 , come pure, per le ipotesi sulle B^i , l'operatore $L^{-1}B$. (compatto in $(L^2)^E$).

Ne consegue subito, per noti risultati di Krein-Rutman, (1), che esiste un autovalore massimo (dominante e semplice) λ_0 del problema, $\lambda_0 u_0 = L^{-1}B u_0$, positivo e maggiore del valore assoluto di ogni altro autovalore, al quale corrisponde un'autosoluzione u_0 pure >0 in Ω .

b) - Determinazione iterativa di λ_0 .

Richiamiamo il Teorema di I. Marek. (6)

Se H, K sono due operatori lineari con $D_H, D_K \subset K$, spazio di Hilbert, e se K è limitato e H^{-1} esiste limitato da $X \rightarrow D_H$, e $H^{-1}K$ è limitato con autovalore dominante λ_0 (e autosoluzione x_0) allora il processo iterativo di Rayleigh-Kellogg descritto dalle:

$$(12) \quad \begin{aligned} u^{(0)} &= x^{(0)} && \text{(approssimazione zero)} \\ v^{(m)} &= K u^{(m)} && ; \quad H u^{(m+1)} = v^{(m)} && ; \quad u_{(m)} = \frac{u^{(m)}}{(u^{(m)}, u^{(m)})} \\ \lambda_m &= \frac{(u^{(m+1)}, u^{(m+1)})}{(u^{(m)}, u^{(m)})} && , \end{aligned}$$

converge:

$$\lambda_m \rightarrow \lambda_0, \quad u_{(m)} \rightarrow x_0$$

Per avere una determinazione iterativa di λ_0 basta applicare tale Teorema nel nostro caso $K=B$, $H=L$, $X \equiv (L^2)^g$.

Si osservi ora che il calcolo dell'iterazione $m+1$ si deve risolvere un'equazione del tipo:

$$(13) \quad - \sum_{p,q}^n \frac{\partial}{\partial x_p} (D_{pq}^i \frac{\partial u^{(m+1),i}}{\partial x_q}) + A^i u^{(m+1),i} - C^i u^{(m+1),i-1} = v^{(m)}$$

e le (13), come al solito disaccoppiandosi, permettono di trovare le $u^{(m+1),i}$ una per volta ($i=1,2,\dots,g$) risolvendo problemi del tipo:

$$(14) \quad \begin{aligned} & - \sum_{p,q}^n \frac{\partial}{\partial x_p} (D_{pq}^1 \frac{\partial u^{(m+1),1}}{\partial x_q}) + A^1 u^{(m+1),1} = v^{(m)} = f_1^{(m)} \quad (\text{nota}) \\ & - \sum_{p,q}^n \frac{\partial}{\partial x_p} (D_{pq}^2 \frac{\partial u^{(m+1),2}}{\partial x_q}) + A^2 u^{(m+1),2} = C^2 u^{(m+1),1} + v^{(m)} = f_2^{(m)} \quad (\text{nota}) \end{aligned}$$

- - - - -

A ciascuna iterazione si devono perciò risolvere problemi di tipo ben noto. Questo consente perciò di applicare tecniche svariate studiate per problemi del tipo $A^i u = f$.

Ad esempio la $u^{(m+1),i}$ può essere trovata, (9), minimizzando il funzionale:

$$(15) \quad I^i(\phi) = \frac{1}{2} \int_{\Omega} \left(\sum_{p,q}^n D_{pq}^i \frac{\partial \phi^i}{\partial x_p} \frac{\partial \phi^i}{\partial x_q} + A^i (\phi^i)^2 - 2 \tilde{f}_i^i \phi \right) d\Omega$$

essendo $\tilde{f}_i^i$ nota dalla precedente iterazione.

Ne consegue che adattando questo metodo si ha un ciclo doppiamente iterativo-variazionale per trovare u_0 , λ_0 .

Evidentemente la $u^{(m+1),i}$ può essere trovata anche col metodo delle differenze finite, e appunto si sono fatte delle esperienze numeriche comparative al riguardo.

§ 4 - Esperienze numeriche

a) Risoluzione del Problema (15) attraverso il metodo di Riesz.

Se si assume una "base" finita $F_v(x)$; $v=1,2,\dots N$, e se si rappresentano le ϕ^i come $\phi^i = \sum_1^N \alpha_{i,v} F_v$, le condizioni di minimo per il funzionale (15) diventano:

$$\frac{\partial I}{\partial \alpha_{i,v}} = 0 \quad ; \quad \begin{matrix} i = 1,2,\dots g \\ v = 1,2,\dots N \end{matrix}$$

Queste ci danno un sistema algebrico lineare del tipo (per ogni i):

$$R_{\alpha^{(m+1),i}=\beta^{(m),i}}^i \quad ; \quad \alpha = (\alpha_1^{(m+1),i}, \dots, \alpha_N^{(m+1),i}) \quad ; \quad (i=1,2,\dots g)$$

$$\beta_{\lambda^{(m),i}}^{(m),i} = \frac{x_i}{\lambda^{(m)}} \sum_1^g S_j^i \alpha^{(m),i} + Q^i \alpha^{(m+1),i-1} \quad ,$$

essendo R,S,Q matrici NXN di elementi noti.

Ad esempio:

$$S_{uv}^i = \sum_1^r \int_{\Omega_\ell} B^i F_u F_v d\Omega_\ell$$

L'inversione delle matrici R è stata fatta con l'algoritmo di Gauss.

b) Confronti tra il metodo iterativo variazionale descritto e quello delle differenze finite (assunto come elemento di confronto).

Caso monodimensionale: $g=2$, $r=5$, D,A,B,C funzioni costanti a tratti; $N=20$.

$$\lambda_0 \text{ (con differenze finite)} = \lambda_{df}$$

F_v sono le autosoluzioni dell'operatore di Laplace monodimensionale

$$\lambda_0 \text{ (con il nostro metodo)} = \lambda_N$$

Ecco una tabella riferentesi ai vari casi:

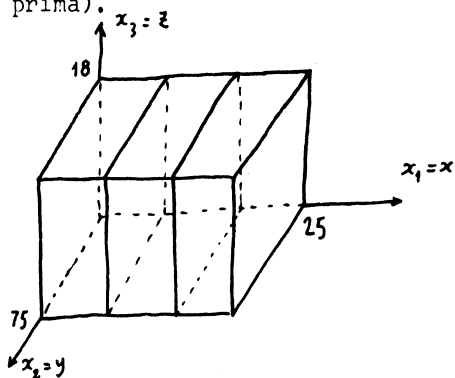
λ_N	λ_{df}	e%
1,082447	1,082678	0,00637
1,226778	1,226704	0,00603
1,000041	1,000011	0,00299
0,439297	0,438788	0,116
1,141368	1,141318	0,00438

L'errore % è al più attorno allo 0,1%. Circa l'andamento delle soluzioni u^1, u^2 nei casi sperimentati si ha un accordo del nostro metodo con quello alle differenze finite sino a 3 cifre significative nelle zone centrali, e uno meno buono (2 cifre) nelle

altre zone.

Caso bidimensionale - Risultati analoghi ai precedenti, perchè l'errore % (N eguale sia per l'asse x_1 che per l'asse x_2) nei nostri esperimenti non ha mai superato lo 0,3%. (F_v sono prodotti di autosoluzione dell'operatore precedente).

Caso tridimensionale: $g=2, r=3$; D,A,B,C costanti a pezzi, simmetria rispetto ai piani $x_1=0, x_2=0$. (F_v prodotti di autosoluzione come prima).

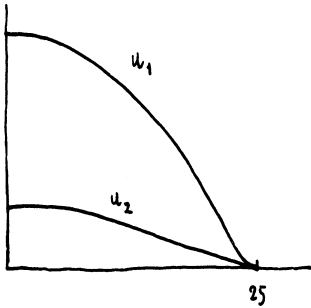


Risultati

- $N_z = N_y = 1, N_x = 3$ e cioè N totale abbastanza piccolo ($N=N_x N_y N_z$): con 6 iterazioni (30" IBM 7090) λ_N approssima λ_{df} entro il 3,3%.
- aumentando N_y da 3 a 10 l'e% si riduce al 3% (1'47" di macchi na).

In generale $\lambda = \lambda_N$ è di tipo monotono (crescente in N), e in 15 iterazioni al più, nei casi considerati, si ha l'autovalore con la approssimazione cercata (1%) mentre è ben noto che con il metodo delle differenze finite il numero delle iterazioni sale, in genere, almeno a circa 50+60.

u_1, u_2 (sulle rette $y=6$, $Z=8$ in figura) sono in buon accordo con i valori ottenuti a differenze finite, tranne nelle interfacce ove lo scarto è $\sim 2,3\%$.



Osservazione - Il metodo iterativo variazionale fornisce l'autovalore massimo in modo assai soddisfacente, sia per precisione che per tempi di calcolatore (rispetto al metodo delle differenze finite), ma invece è inferiore a quest'ultimo metodo per la precisione della tabulazione della soluzione specie per quel che riguarda l'andamento della u_2 che può presentare dei "picchi" nelle zone non centrali (e vicino alle interfacce), picco a volte assai mal descrivibile col metodo variazionale.



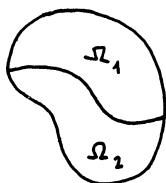
PARTI 2^a - Teoria a multigruppi dipendente dal tempo

§ 1 - E' ben noto che nella teoria della diffusione dei neutroni nell'approssimazione a più gruppi g di velocità, che supporremo due per semplicità, l'evoluzione del tempo dei flussi veloce e lento, rappresentati da u_1, u_2 e della concentrazione dei così detti "neutroni ritardati" rappresentata da C , è retta dal seguente sistema:

$$\begin{aligned} (1) \quad & \frac{1}{v_1} \frac{\partial u_1}{\partial t} - D_1 \Delta u_2 + a_{11} u_1 - a_{12} u_2 - \lambda C = f_1 \\ & \frac{1}{v_2} \frac{\partial u_2}{\partial t} - D_2 \Delta u_2 + a_{22} u_2 - a_{21} u_1 = f_2 \\ & \frac{\partial C}{\partial t} + \lambda C - h u_2 = f_3 \end{aligned}$$

essendo assegnate:

- 1) i coefficienti (funzioni misurabili e limitate essenzialmente >0) e le funzioni di "sorgente" $f_{1,2,3}$,
- 2) le condizioni (Dirichlet) per le u_1, u_2 al contorno $\partial \Omega \times]0, T[$, $0 < t < T$, $\Omega \subset \mathbb{R}_n$,
- 3) le condizioni di "trasmissione" relative ad u_1, u_2 rispetto una interfaccia γ (una sola per semplicità) che rappresenta la frontiera comune a due subregioni Ω_1, Ω_2



- 4) la condizione iniziale di Cauchy per tutte le incognite:

$$u_{1,2}(x, 0) = u_{1,2}^0(x)$$

$$C(x, 0) = C^0(x)$$

Il sistema stazionario associato ad (1) è ellittico secondo

Nirenberg-Douglis, ma non secondo Petrowsky, e pertanto essenzialmente diverso, ad esempio, del sistema ellittico del tipo elastici tà. Infatti, l'operatore matriciale associato è del tipo:

$$(2) \quad A = \begin{vmatrix} -D_1\Delta + a_{11} & -a_{12} & -\lambda \\ -a_{21} & -D_2\Delta + a_{22} & 0 \\ 0 & -h & \lambda \end{vmatrix}$$

da cui si vede subito che sulla diagonale principale l'ordine degli operatori sono rispettivamente 2, 2, 0.

Risulta quindi naturale assegnare le condizioni al contorno (e di trasmissione) solo per u_2, u_2 .

Inoltre essendo il sistema a coefficienti discontinui i risultati di Agmon-Douglis-Nirenberg per i sistemi ellittici (che esigono la continuità dei coefficienti) non si applicano. Appare perciò naturale, pensare ad una impostazione di tipo variazionale. Per questo vedi un lavoro di prossima pubblicazione di Albertoni-Daneri-Geymonat. [5]

Non presenteremo, pertanto, qui risultati teorici circa (1) essendo le ricerche in corso, ma considereremo solo il problema discretizzato che si ottiene da (1) mediante il metodo delle linee (discretizzazione solo sulle variabili spaziali) ottenendo:

$$(2) \quad V^{-1} \frac{d\phi}{dt} + B\phi = 0 \quad ; \quad \frac{d\phi}{dt} = Q\phi \quad , \quad Q = -VB \quad ,$$

essendo ϕ = vettore di R_{2N+N_1} ove N è il numero dei punti del reticolo spaziale (esclusa la frontiera) e N_1 è il numero dei punti ove $h(x) \neq 0$. Inoltre è:

$$V^{-1} = \begin{pmatrix} V_2^{-1} & 0 & 0 \\ 0 & V_2^{-1} & 0 \\ 0 & 0 & E \end{pmatrix} \quad \text{ove:} \quad \begin{cases} V_2^{-1} \text{ matrice diagonale } NXN \\ V_2^{-1} \text{ matrice diagonale } NXN \\ E \text{ identità in } R_{N_1} \end{cases}$$

$$B = \begin{pmatrix} L_1 + A_{11} & -A_{12} & -\lambda E_2 \\ -A_{21} & L_2 + A_{22} & 0 \\ 0 & -H & \lambda E \end{pmatrix}$$

essendo:

- $L_1 \rightarrow D_1 \Delta$ una matrice pentadiagonale $N \times N$,
 $L_2 \rightarrow D_2 \Delta$ una matrice pentadiagonale $N \times N$,
 $A_{11} \rightarrow a_{11}$ una matrice diagonale $N \times N$ a coefficienti > 0 ,
 $A_{12} \rightarrow a_{12}$ una matrice diagonale $N \times N$ a coefficienti > 0 ,
 $A_{21} \rightarrow a_{21}$ una matrice diagonale $N \times N$ a coefficienti > 0 ,
 $A_{22} \rightarrow a_{22}$ una matrice diagonale $N \times N$ a coefficienti > 0 ,
 $H \rightarrow h$ una matrice $N_1 \times N$ a coefficienti > 0 ,
 $E_1 \rightarrow \lambda$ una matrice $N \times N_1$ a coefficienti > 0 .

Osservazione I - Lo studio del problema discreto (2) è egualmente molto importante in analisi numerica, perchè, a differenza di quanto accade nei casi parabolici, concernenti in generale più o meno la diffusione del calore, la matrice può avere autovalori > 0 , il che comporta a volte una crescenza molto rapida di ϕ cosa sempre delicata da controllare dal punto di vista numerico.

Inoltre le $v_{1,2}$ (in generale costanti) sono $\approx 10^6$, mentre gli altri coefficienti sono ≈ 1 , e pertanto dal punto di vista numerico si incontrano difficoltà simili a quelle che si hanno nei problemi di "boundary layer" connessi con equazioni differenziali contenenti piccoli parametri nelle derivate più alte. Usando schemi impliciti (per ragioni di stabilità) si generano da (2) "grossi" sistemi lineari per i quali occorrono metodi iterativi la cui convergenza, che era da indagare, è stata verificata in.(3)

§ 2 - Proprietà del sistema (2).

In (3) sono stati ottenuti i seguenti risultati:

- 1) Q è irriducibile;
- 2) Q è essenzialmente >0 ;
- 3) Q possiede un autovalore $\omega_0 > -\lambda$ cui corrisponde un autovettore $v > 0$, tale che se α_i è un qualsiasi altro autovalore è:
 $R \alpha_i < \omega_0$;
- 4) dalle 1), 2), 3), seguendo Birhoff-Varga (7), si dimostra ($t \rightarrow \infty$):
 $\phi(t) = K e^{\omega_0 t} v + O(e^{\mu t})$ con $R \alpha_i < \mu < \omega_0$
 (K dipende da $\phi(0)$);
- 5) le matrici di Jacobi e di Gauss-Seidel associate alla matrice $\alpha I - Q$ sono convergenti per $\alpha > \omega_0$.

§ 3 - Metodi di risoluzione di (2).

- a) Metodo Esplicito: $\phi(t) = (I + \Delta t Q) \phi(t - \Delta t)$. Tale metodo è stato scartato nei nostri casi perchè ha una soglia di stabilità troppo bassa (Δt troppo piccolo).
- b) Metodo Implicito: $\phi(t) = (I - \Delta t Q)^{-1} \phi(t - \Delta t)$.

Ad ogni passo temporale c'è da risolvere un sistema del tipo:

$$\left(\frac{I}{\Delta t} - Q\right) x = K \quad (\text{noto}).$$

Se $\Delta t \omega_0 < 1$ allora, in base alla proprietà 5 del §2, i metodi di Jacobi e Gauss-Seidel relativi sono convergenti.

Osservazione - Quando si ha un "transiente" molto rapido si è trovato che anche il metodo implicito (e pure quelli di Crank-Nicolson e di Saul'yev (4)) risultano molto imprecisi. Pertanto si pone il problema di trovare qualche metodo meno impreciso. La valutazione (4) ha fornito l'idea base per il seguente metodo che chiameremo metodo ω .

Nel metodo ω s'è pensato di esprimere la soluzione nella forma:

$$(3) \quad \phi(t) = e^{\omega_0 t} \psi(t)$$

Allora la (2) si trasforma in:

$$(4) \quad V^{-1} \frac{d\psi}{dt} + (B + \omega_0 V^{-1})\psi = 0 \quad ; \quad \psi(0) = \psi(0) .$$

In ogni caso però c'è il problema di determinare ω_0 che non è determinabile con procedimento tipo "metodo delle potenze Rayleigh-Kellog" non essendo l'autovalore quello di modulo massimo. Questa questione è abbastanza difficile dal punto di vista numerico, perchè è vero che si può tentare di "traslare" lo spettro al fine di condursi ad un problema di autovalore di massimo modulo, ma così facendo, dovendo poi sottrarre il passo di traslazione, se questo è molto grande si può perdere ogni significato.

Allora posto $B=M-N$:

$$M = \begin{pmatrix} L_{11}+A_{11} & 0 & 0 \\ -A_{21} & L_2+A_{22} & 0 \\ 0 & -H & \lambda E \end{pmatrix} ; \quad N = \begin{pmatrix} 0 & A_{12} & \lambda E_1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

ci si riduce, come equazione agli autovalori per ω_0 (essendo ω_0 autovalore di $-VB$) alla (5): $(M + \omega_0 V^{-1}) x = Nx$.

Introdotta un parametro fittizio μ si dimostra che l'equazione $(M + \omega V^{-1})x = \frac{1}{\mu} Nx$ possiede un autovalore di massimo modulo cui corrisponde un autovettore x_0 ottenibile col metodo iterativo delle potenze. Quello che si dimostra è che $\mu = \mu(\omega)$ è monotona decrescente, e che $\mu=1$ individua ω_0 .

Il sistema (4) è poi risolto con il metodo implicito Crank-Nicolson ed i relativi metodi iterativi risultano convergenti. (3)

§ 4 - Metodo dei passi frazionari

Recentemente, (5) , abbiamo pure sperimentato il metodo di Marchuck ed altri per la (1) ($g=1$) assumendo $C=0$ e come decomposizione dell'operatore A una del tipo (vedi § 5) :

$$\begin{aligned} A &= A_1 + A_2 \\ (7) \quad A_1 &= \frac{\partial}{\partial x} \left(D \frac{\partial}{\partial x} \right) + \frac{1}{2}b \\ A_2 &= \frac{\partial}{\partial y} \left(D \frac{\partial}{\partial y} \right) + \frac{1}{2}b \end{aligned}$$

Lo schema alternato per il passaggio da $t_n \rightarrow t_{n+1}$, $t \in [t_n, t_{n+1}]$ è il seguente:

$$\begin{aligned} (8) \quad \frac{1}{v} \frac{d u^{n+\frac{1}{2}}}{dt} + A_1 u^{n+\frac{1}{2}} &= 0 \\ u^{n+\frac{1}{2}}(t_n) &= u(t_n) \\ \frac{1}{v} \frac{d u^{n+1}}{dt} + A_2 u^{n+1} &= 0 \\ u^{n+1}(t_n) &= u^{n+\frac{1}{2}}(t_{n+1}) \end{aligned}$$

formente $u(t_{n+1}) = u^{n+1}(t_{n+1})$. La discretizzazione della (8) ci dà poi sistemi del tipo:

$$\begin{aligned} (9) \quad V^{-1} \frac{du}{dt} &= \alpha_1 u ; & V^{-1} &\text{ matrice diagonale;} \\ V^{-1} \frac{du}{dt} &= \alpha_2 u ; & \alpha_1, \alpha_2 &\text{ matrici tridiagonali} \end{aligned}$$

Osservazione - Un primo vantaggio del metodo è che abbiamo ora a che fare con matrici tridiagonali (invece di pentadiagonali) invertibili anche con metodi diretti.

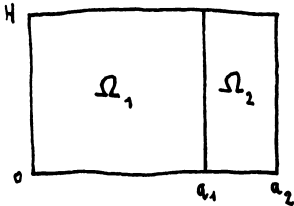
§ 5 - Risultati numerici

Si è sviluppato un caso "test" per il quale si conosce la soluzione esatta e si sono fatti confronti con i metodi a), b), c) (§ 3).

Problema "test": $g = 1$ (e senza equazione in C).

$$\frac{1}{v} \frac{\partial u}{\partial t} = \text{div}(d \text{ grad } u) + bu \equiv Au$$

$u|_{\Gamma} = 0$ e condizioni di trasmissione su γ_{12}



essendo

$$v = \begin{cases} v_1 & ; x \in \Omega_1 \\ v_2 & ; x \in \Omega_2 \end{cases} ; \quad d = \begin{cases} d_1 & ; x \in \Omega_1 \\ d & ; x \in \Omega_2 \end{cases} ; \quad b = \begin{cases} b_1 & ; x \in \Omega_1 \\ b_2 & ; x \in \Omega_2 \end{cases}$$

Se

$$\frac{d_1}{d_2} = \frac{v_1}{v_2} ; \quad b_2 = \{b_1 v_1 + (\frac{\pi}{H})^2 (d_2 v_2 - d_1 v_1)\} / v_2 ; \quad a = a_1 + (a_2 - a_1) \frac{d_1}{d_2}$$

allora:

$$u_j = \sin \frac{\pi}{H} y \cdot \begin{cases} \sin \frac{\pi}{a} j x & ; x \in \Omega_1 \\ \sin \frac{\pi}{a} j (\frac{d_1}{d_2} x + (1 - \frac{d_1}{d_2}) a_1) & ; x \in \Omega_2 \end{cases}$$

sono delle autosoluzioni di A, e $\omega_j = \{a_1 - d \left[(\frac{\pi}{a} j)^2 + (\frac{\pi}{H})^2 \right]\} v_1$ sono i corrispondenti autovalori. Inoltre se $u(x, y, 0) = \sum_{j=1}^N C_j u_j$; si ha la soluzione esatta:

$$(6) \quad u(x, y, t) = \sum_{j=1}^N C_j e^{\omega_j t} u_j$$

L' ω_0 calcolato come indicato alla fine del § 3 risulta $\omega_0 = 63,69$, mentre il valore esatto è 63,21. (Le differenze finite sovrastimano l'autovalore e, a nostra conoscenza, ci sono risultati teorici per questa stima solo nel caso di coefficienti continui in $\bar{\Omega}$).

I risultati numerici sono riportati in tabella dove $\bar{u}$ è il valore medio su Ω della soluzione esatta, $\bar{u}_{M\omega}$ è la soluzione approssimata ottenuta con la trasformazione ω ed il metodo di Marchuck, $\bar{u}_M$ l'analogia soluzione senza trasformazione ω , $\bar{u}_{IM\omega}$ quello ottenuto con il metodo implicito che fa seguito alla trasformazione ω , mentre la $\bar{u}_{IM}$ è quella che si riferisce al metodo implicito diretto.

Come si vede in ogni caso la trasformazione ω , comunque sia associata ad altre tecniche, dà i risultati migliori, ed il metodo dei passi frazionari si è rivelato superiore, nei casi fatti, al metodo implicito.

T A B E L L A ($\Delta t = 10^{-3}$)

t	Soluzione esatta $\bar{u}$	$\bar{u}_{Mw}$	$\bar{u}_M$	$\bar{u}_{IMw}$	$\bar{u}_{IM}$
.025	1.976	2.000	2.039	2.024	2.132
.030	2.708	2.749	2.815	2.780	2.960
.035	3.714	3.779	3.886	3.821	4.111
.040	5.094	5.200	5.366	5.252	5.711
.045	6.986	7.153	7.411	7.220	7.934
.050	9.583	9.838	10.23	9.926	11.02
.055	13.14	13.53	14.14	13.65	15.31
.060	18.03	18.62	19.52	18.76	21.28
.065	24.73	25.61	26.96	25.80	29.57
.070	33.92	35.23	37.24	35.47	41.08
.075	46.54	48.47	51.44	48.77	57.09
.080	63.83	66.69	71.05	67.06	79.32
.085	87.56	91.77	98.14	92.20	110.2
.090	120.1	126.68	135.22	126.8	153.1
.095	164.7	173.68	187.22	174.3	212.9
.100	226.0	238.89	258.59	239.7	295.6

Bibliografia

- (1) S.Albertoni: "Metodi variazionali per certi sistemi di equazioni a derivate parziali" - Ist. Lombardo Scienze e Lettere Vol. 100-1966.
- (2) S.Albertoni - M.Lunelli - G.Maggioni: "Metodi iterativi variazionali per problemi ellittici nella teoria dei reattori nucleari" Atti Seminario Mat. e Fis. Univ.Modena Vol.XIV,1965.
- (3) A.Daneri - A.Daneri - I.Galligani: "A numerical approach to the time dependent neutron diffusion equations" EUR 3742e - 1968.
- (4) I.Galligani: "Numerical solutions of the time dependent diffusion equations using the alternative method of Saul'yev" - Calcolo Vol.2, Suppl. 1, 111 - 1965.
- (5) S.Albertoni - A.Daneri - G.Geymonat: "Existence and approximation theory for general differential equations of the multi-group diffusion reactor theory" (in corso di pubblicazione).
- (6) I.Marek: "Iterations of linear unbounded operator in non self-adjoint eigenvalues problems and Kellog iteration processes" (Cech.Meth.Journal 12 (1962)).
- (7) G.Birkoff - R.S.Varga: "Reactor Criticality and non negative matrices" J.Soc.Indust.Appl.Math. 6, 354-377 (1958).
- (8) G.I. Habetler - M.A. Martino: "The multigroup diffusion equations of reactor physics, KAPL, 1886, July 28, 1958
- (9) G.Stampacchia: "Su un problema relativo alle equazioni di tipo ellittico del 2° ordine" Ricerche di Mat., Vol. V (1956).

CENTRO INTERNAZIONALE MATEMATICO ESTIVO
(C. I. M. E.)

I. BABUŠKA

"PROBLEMS OF OPTIMIZATION AND NUMERICAL STABILITY'
IN COMPUTATIONS"

Corso tenuto ad Ispra dal 3-11 Luglio 1967

PROBLEMS OF OPTIMIZATION AND NUMERICAL STABILITY
IN COMPUTATIONS ¹⁾

by

I. Babuška (Praga)

1. Introduction

Computer Science is a new scientific discipline. An important part of this discipline is the numerical mathematics. The "Art of Computation" is becoming science ; new questions and problems become important.

A typical problem is the problem of the creation of numerical methods, the determination of their "worth" and, in general, the choice of the most suitable method for the given purpose.

For example, the program-library in a computing centre contains mostly many algorithms for solving single mathematical problems. Opinions on the expedience of these algorithms are usually quite different and subjective. This statement is still more apparent when a method of applied mathematics is to be appreciated, especially in the field of scientific-technical computations. These scientific - technical computations are that part of the computer science in which I have some experience.

My paper will deal with questions which are more or less associated with this kind of computations.

I think that these computations may be characterized as a mathematical and constructive way of processing (transformation) of the given information to the required one ²⁾. I am sure that in scientific-technical computations it is necessary to emphasize the knowledge of information which we may collect and the appreciation of its reliability. Further it is

¹⁾ In this paper some results obtained recently in Prague will be given

²⁾ Henrici ²³ defines numerical analysis as the theory of constructive methods in mathematical analysis (with emphasis on the word "constructive").

necessary to formulate clearly the required information on the given problem. The necessity of a mathematical and constructive way of this processing is obvious here.

The "clarity" of the given and required information is an important part for a successful solution of a technical problem. Numerical mathematics are the rudiments of this constructive processing of information.

Numerical method generates (in a constructive manner) a mapping, from the class (space) of the given information to the class of the required one. It is important that this mapping is defined on the entire class of information. This class will be the domain of definition of the given method (mapping).

Numerical process is an exact constructive law (prescription) of creation of the given mapping.

Computation is a concrete realisation of the numerical process in the given case. We shall talk about exact realisation when we compute without round-off errors and about a realization (or disturbed realisation) in a real computation.

It is obvious that there are many different manners of a constructive creation of one given mapping, i.e. many processes exist which transform the given information to the requested one and solve the same mathematical problem. It is evident that the question of choosing a process is very important.

It is clear that the choice and every optimization must necessarily be relative to the given information. This does not mean, however, that some methods might not be advantageous in a certain generality.

The manner in which we appreciate the method is of great importance. My experience is that, from the practical point of view, it is very important to respect an incredulity of the given information. This incredulity can be of different kinds. Some of them will be shown in the next

part of the paper . It is essential that the method -and in general all conclusion - be stable with respect to these incredulities. I think that this stability is one of the most important points when choosing a method in practice.

In the next part I shall point out some aspects of these questions.

2. The problem of quadrature formulas ¹⁾

In this section I shall show some aspects of ideas, which I mentioned previously, in a simple case of quadrature formulas.

Let our task be to determine numerically

$$(2.1) \quad J(f) = \frac{1}{2\pi} \int_0^{2\pi} f(x) \, dx .$$

We shall suppose that we know the following about the integrated function $f(x)$:

1. The function $f(x)$ is a continuous periodic function with the period 2π .

2. We can evaluate only the function $f(x)$ (i.e. compute the values of $f(x)$) .

In this case, the simplest quadrature formula $T_n(f)$ is mostly used in practice, with

$$(2.2) \quad T_n(f) = \frac{1}{n} \sum_{j=1}^n f\left(\frac{2\pi}{n} j\right) .$$

This formula is the well known trapezoid formula .

I will now analyse the question, if there are any reasons for selecting

¹⁾ In this part we are not dealing with the problems of the round-off errors.