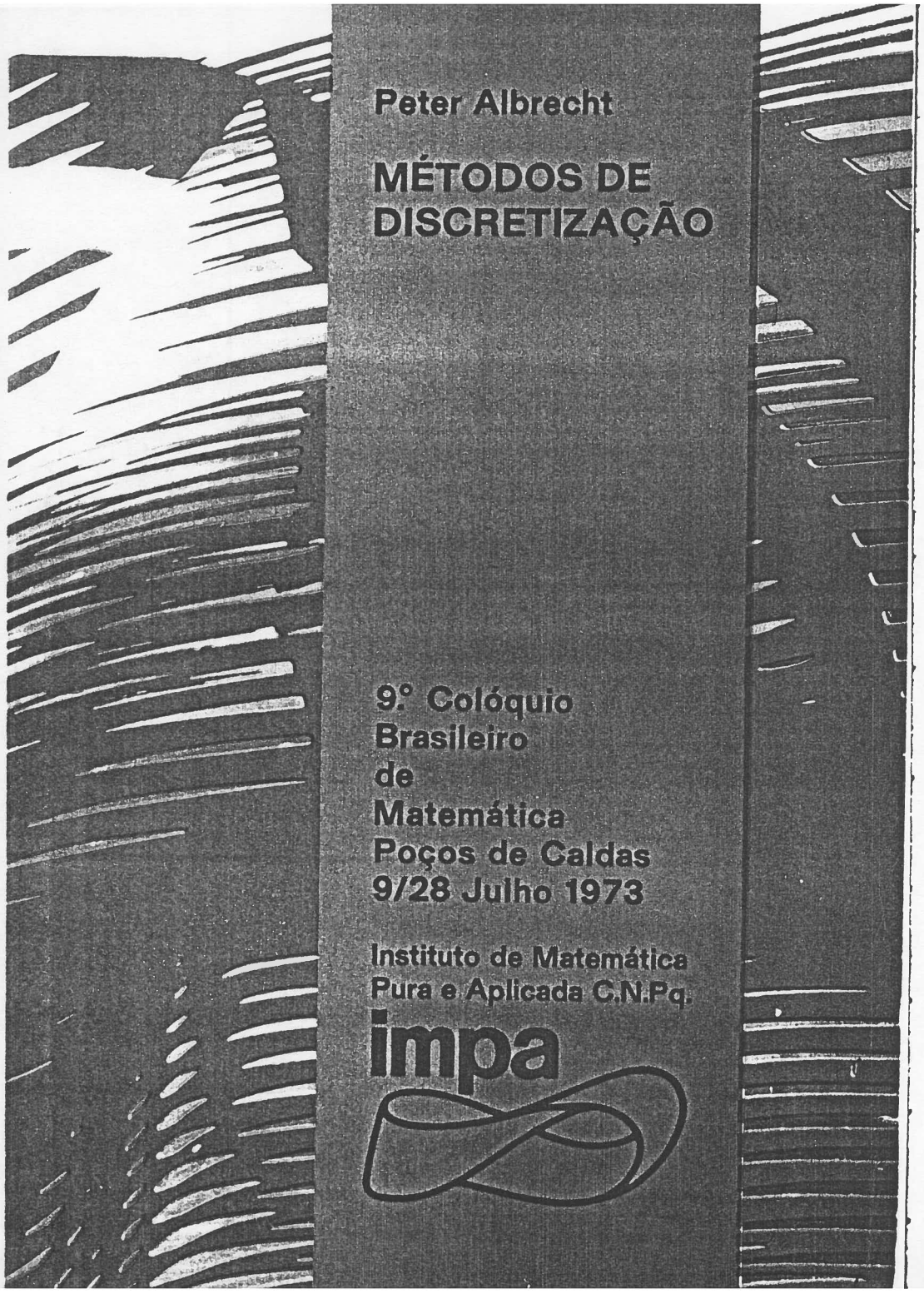


Peter Albrecht

MÉTODOS DE DISCRETIZAÇÃO

9.° Colóquio
Brasileiro
de
Matemática
Poços de Caldas
9/28 Julho 1973

Instituto de Matemática
Pura e Aplicada C.N.Pq.

impa

Peter Albrecht

MÉTODOS DE DISCRETIZAÇÃO

PREFÁCIO

Desde que uma grande parte dos problemas da Matemática Numérica é resolvida por discretização, tentamos fazer, deste curso, uma apresentação unificada (i) dos métodos gerais de discretização e (ii) da aceleração de sua convergência. Procuramos uma abordagem completa mas, ainda assim, elementar, tentando tratar várias aplicações sob pontos de vista comuns e unificados.

Na primeira parte conseguimos uma apresentação unificada introduzindo operadores de discretização, para os quais são definidas a consistência (seção 1.2) e a estabilidade (seção 1.3) de um modo bem geral. Para manter o carater elementar, a teoria geral é apresentada sobre um caso particular, o da discretização de sistemas de equações diferenciais ordinárias. Chegamos, assim ao mesmo tempo, a uma introdução detalhada ao tratamento numérico destes sistemas. Entretanto, esta abordagem se afasta das apresentações usuais pois está fundamentalmente baseada num novo critério geral de estabilidade (teorema 1.3), do qual se poderia, ainda, derivar outros novos resultados.

Na segunda parte, inicialmente, apresentamos a extrapolação ao limite através de um exemplo - a quadratura numérica com o método de Romberg - e em seguida, a teoria geral (seção 2.2). Procuramos dar uma apresentação, em contraposição às usuais, bastante simplificada, a

bordando a extrapolação ao limite de uma maneira nova e generalizada. Exemplificamos o método através da diferenciação numérica (seção 2.3) e da integração de sistemas de equações diferenciais ordinárias (seção 2.4).

A segunda parte foi tirada do capítulo VIII do nosso livro sôbre Análise Numérica[1] ; a primeira é uma revisão de um capítulo da versão alemã, a ser publicada. Do mesmo livro vem a idéia de separar completamente a parte teórica da parte computacional, apresentando os algoritmos apenas como "receitas" de fácil programação. Em cada algoritmo, entretanto, está indicado um teorema que contém detalhes teóricos do método.

Este curso teve de ser preparado, traduzido, revisado e datilografado em seis semanas. Isto não seria possível sem a ajuda dos colegas do Departamento de Informática da PUC-RJ; especialmente agradeço a Therezinha Chaves pelo espírito cooperativo e a ajuda dedicada na preparação do texto português, Marcelo Klein pela revisão rigorosa e indicação de imprecisões da primeira parte e Albrecht von Plehwe por várias sugestões para a primeira parte. Finalmente destaco o trabalho de Antonia Oka, datilografando excelentemente o texto, em tempo mínimo.

Rio de Janeiro, Abril 1973

[1] Um Curso de Análise Numérica, Ao Livro Técnico R.J. 1973.

INDICE

PARTE I : MÉTODOS DE DISCRETIZAÇÃO PARA EQUAÇÕES DIFERENCIAIS

Introdução 1

1.1. Definições e Conceitos 2

1.2. Consistência 6

1.3. Estabilidade para métodos de discretização geral 18

1.3.1. Exemplo 19

1.3.2. Definição de estabilidade 21

1.3.3. Um critério geral de estabilidade 26

1.3.4. Estabilidade e estimativa do erro para métodos de passo simples 32

1.3.5. Estabilidade e estimativa do erro para métodos de passo múltiplo 34

1.3.6. Condições de estabilidade para equações diferenciais parciais 39

1.4. Convergência dos métodos de passo simples ou múltiplo 42

1.5. Dedução de métodos de passo simples 47

1.5.1. Métodos com derivadas 47

1.5.2. Métodos de Runge-Kutta 49

1.5.3. Métodos combinados 61

1.6. Obtenção de métodos de passo múltiplo 62

1.6.1. Os métodos de Adams-Bashforth e Nyström 65

1.6.2. Métodos preditor-corretores clássicos 67

1.7. Estabilidade e ordem de métodos preditor-corretores 73

1.8. Comparação dos métodos 77

1.8.1. Métodos de passo simples 78

1.8.2. Métodos de passo múltiplo 79

1.8.3. Métodos de extrapolação 79

Observações Bibliográficas 81

Bibliografia da parte I 84

PARTE II : EXTRAPOLAÇÃO AO LIMITE

Introdução 86

2.1. Quadratura numérica por extrapolação ao límite 92

2.1.1. O método de Romberg 92

2.1.2. Casos particulares do método de Romberg 96

2.1.3. Considerações sôbre o êrro e a convergência 100

2.2. Discussão geral da extrapolação ao limite 108

2.2.1. Tratamento geral do método 109

2.2.2. Dois algoritmos particulares 113

2.2.3. Considerações sobre o êrro nos algoritmos A e B .. 116

2.3. Diferenciação numérica por extrapolação 122

2.4. Integração numérica de equações diferenciais por extrapolação ao limite 128

Observações Bibliográficas 128

Bibliografia da parte II 130

PARTE I

MÉTODOS DE DISCRETIZAÇÃO PARA EQUAÇÕES DIFERENCIAIS

INTRODUÇÃO

Neste capitulo trataremos de métodos numéricos para resolução de sistemas de equações diferenciais ordinárias de primeira ordem com condições iniciais, da forma

$$T[y](x) \equiv y'(x) - f(x,y)) = 0 \; ; \quad y(a) = \eta_0 \tag{1.1}$$

$$y:[a,b] \Rightarrow \mathbb{R}^n \; ; \; f:[a,b] \times \mathbb{R}^n \Rightarrow \mathbb{R}^n$$

Como se sabe, (1.1) tem uma única solução $y \in \mathbb{C}^1[a,b]$ se $f(x,y)$ é contínua e satisfaz à seguinte condição de Lipschitz em $G: \{a \le x \le b \; ; \; y \text{ qualquer}\}$:

$$\| f(x,y) - f(x,\bar{y}) \| \le L \| y - \bar{y} \| \; ; \quad L \text{ constante.} \tag{1.2}$$

Daqui em diante assumiremos sempre satisfeita esta condição e como norma no $\mathbb{R}^n$ usaremos sempre

$$\| u \| = \max_{1 \le j \le n} |u_j| \; ; \quad u = (u_1, u_2, \ldots, u_n)^T$$

Limitaremos nossas considerações a problemas de condições iniciais da forma (1.1), entretanto as definições e a maioria dos teoremas sobre consistência, estabilidade e convergência podem ser aplicados a problemas de discretização muito mais gerais, especialmente ao tratamento numérico de sistemas de ordem maior e problemas de contôrno, bem como para problemas de valores iniciais e de contôrno de equações diferenciais parciais. Estas aplicações serão apresentadas em vários exemplos mas, renunciamos a uma apresentação mais geral para não complicar desnecessariamente este curso.

§ 1.1 - DEFINIÇÕES E CONCEITOS

Todos os métodos de discretização para obtenção de uma solução numérica de (1.1), consistem em, dada uma rede

$$I_h = \{x_k = a + kh \mid k = 0,1,\ldots,m;\ h = \frac{b-a}{m}\}$$

aproximar os valores $y(x_k)$ da solução y de (1.1) por determinados valores y_k^*.

Isto é feito substituindo-se adequadamente o operador diferencial T em (1.1) por um operador de diferenças T_h e determinando-se as aproximações y_k^* a partir do sistema de diferenças assim obtido.

Exemplo:

Se aproximarmos

$$T[y](x) \equiv y'(x) - f(x,y(x)) \ , \quad y: \mathbb{R} \to \mathbb{R} \ ,$$

por $$T_h[y](x) \equiv \frac{y(x+h)-y(x)}{h} - f(x,y(x)) \ ; \ y: \mathbb{R} \to \mathbb{R}$$

então o problema (1.1) se reduz à equação de diferenças

$$T_h[y^*](x) = 0 \ , \quad y^*(a) = \eta_0$$

e com $x = x_k$ obtemos sucessivamente as aproximações y_k^* de

$$y_0^* = \eta_0 \ ; \ y_{k+1}^* - y_k^* - hf(x_k, y_k^*) = 0 \ ; \ k = 0,1,\ldots,m-1 \quad (1.4$$

(1.4) é conhecido como método de Euler e é o mais simples dos chamados métodos de passo simples.

Definição 1.1:

Um método de diferenças é chamado de passo simples (one-step method) se a aproximação y_{k+1}^* é calculada a partir somente do resultado y_k^* do passo anterior. Se são necessarios os resultados y_k^*, $y_{k-1}^*,\ldots, y_{k-p+1}^*$, $p>1$, para a determina-

ção de y^*_{k+1} , então o procedimento é chamado <u>método de passo múltiplo</u> (multistep method), particularmente <u>método de passo p</u>.

No parágrafo 1.5 vamos apresentar métodos de passo simples da forma

$$hT_h[y^*](x_k) \equiv y^*_{k+1} - y^*_k - h\phi(x_k, y^*_k; h) = 0 \quad ; \quad y^*_0 = \eta_0 \qquad (1.5)$$
$$k=0,1,\ldots,m-1$$

e no parágrafo 1.6 métodos de passo múltiplo da forma

$$hT_h[y^*](x_k) \equiv \alpha_p y^*_{k+p} + \alpha_{p-1} y^*_{k+p-1} + \ldots + \alpha_0 y^*_k - h\{\beta_p f_{k+p} + \ldots + \beta_0 f_k\} = 0 \qquad (1.6)$$

$$\alpha_p = 1 \; ; \; y^*_0 = \eta_0 \; ; \; y^*_1 = \eta_1(h) \; ; \; \ldots \; ; \; y^*_{p-1} = \eta_{p-1}(h); \; |\alpha_0| + |\beta_0| \neq 0$$

$$f_k = f(x_k, y^*_k) \quad ; \quad k = 0,1,\ldots,m-p$$

onde os $\eta_i(h)$, $i = 1,\ldots,p-1$ são determinados por outros métodos (por exemplo de passo simples) num cálculo inicial. Se $\beta_p = 0$ o método é chamado <u>explícito</u> e se $\beta_p \neq 0$, <u>implícito</u> pois, neste caso, y^*_{k+p} ainda aparece em $f_{k+p} = f(x_{k+p}, y^*_{k+p})$. Por isso, como veremos no parágrafo 1.6, os métodos implícitos necessitam, em cada passo k, uma iteração para determinar y^*_{k+p} .

<u>Exemplos</u>:

1.- Substituindo-se em (1.1) $y'(x)$ por $\frac{1}{2h}\left[y(x+h)-y(x-h)\right]$ então, para $x = x_k$, obtem-se o seguinte regra do ponto médio:

$$y^*_{k+1} = y^*_{k-1} + 2hf(x_k, y^*_k) \; ; \; k = 1,2,\ldots,m-1 \qquad (1.7)$$

com os valores iniciais $y^*_0 = \eta_0$; $y^*_1 = y^*_0 + hf(x_0, y^*_0)$, onde y^*_1 foi determinado pelo método de Euler.

2.- Um outro método de passo simples é o <u>método de Euler modificado</u> (vide algoritmo 1.2):

$$y^*_0 = \eta_0 \; ; \; y^*_{k+1} = y^*_k + hf\left(x_k + \frac{h}{2}, \; y^*_k + \frac{h}{2} f(x_k, y^*_k)\right) \qquad (1.8)$$

$$k = 0,1,\ldots,m-1$$

Neste caso temos:

$$T_h[y](x) \equiv \frac{y(x+h)-y(x)}{h} - f\left(x + \frac{h}{2}, \; y(x) + \frac{h}{2} f(x, y(x))\right)$$

3.- (1.8) pode ser usado para calcular o valor inicial y^*_1 para o <u>método de Milne-Simpson</u>:

$$y^*_{k+1} = y^*_{k-1} + \frac{h}{3}\left(f_{k+1} + 4f_k + f_{k-1}\right) \; ; \; k = 1,2,\ldots,m-1 \qquad (1.9)$$

$$y_0^* = \eta_0 \; ; \; y_1^* = y_0^* + hf(x_0 + \frac{h}{2} \, , \, y^* + \frac{h}{2} f(x_0, y_0^*))$$

Este é um método de passo 2 implícito. O operador de diferenças $T_h[y]$ associado é

$$T_h[y](x) \equiv \frac{y(x+h)-y(x-h)}{2h} - \frac{1}{6}\left[f(x+h,y(x+h))+4f(x,y(x))+f(x-h,y(x-h))\right]$$

Detalhes sobre a obtenção e aplicação desses procedimentos serão apresentados nos parágrafos 1.5 e 1.6.

A seguir chamaremos uma função $y(x)$, definida somente nos pontos $x \in I_h$, de <u>função de rede</u> e a denotaremos por $\{y_k\}_h$. Além disso denotaremos por I_h' o conjunto de todos os $x \in I_h$ para os quais o operador $T_h[y](x)$ está definido.

§ 1.2 - CONSISTÊNCIA

O operador T_h pode ser considerado como uma aproximação de T num espaço $\mathbb{F}$ de funções quando

$$\lim_{h\to 0} \max_{x\in I_h'} \| T_h[z](x) - T[z](x) \| = 0$$

para todas as funções $z \in \mathbb{F}$.

Para operadores T_h não lineares (por exemplo no caso dos métodos de passo múltiplo (1.6), quando f não é linear) a condição acima somente se verifica para todo $z \in \mathbb{F}$ em casos particulares. Por isso nos restringiremos apenas a operadores que aproximam T somente para y solução de $T[y](x) = \sigma$. Chamaremos então T_h de <u>consistente</u> com T quando o <u>erro da substituição</u> $T_h[y](x) - T[y](x) \equiv T_h[y](x)$ converge uniformemente para zero, em h:

<u>Definição 1.2</u>:

Seja y solução de $T[y](x) = \sigma$ e I_h' o conjunto de todos os $x \in I_h$ para os quais $T_h[y](x)$ está definido. Então o operador T_h, e também o método de diferenças definido por ele, é chamado de <u>consistente</u> (com o problema $T[y](x) = \sigma$ (1)) se

$$\lim_{h\to 0} \max_{x\in I_h'} \| T_h[y](x) \| = 0 \qquad (1.10)$$

Se, além disso, existem constantes D>0 e q>0 tais que, para $h \le h_0$

$$\max_{x\in I_h'} \| T_h[y](x) \| \le D\, h^q$$

então o método é dito ter <u>ordem</u> (de consistência) q.

---

(1) Esta observação pode ser feita suprimida se não há dúvidas sôbre o problema em consideração

Se também as condições iniciais $\eta_k(h)$ de um método de passo $p \geq 1$ são aproximações dos valores exatos $y(x_k)$, isto é, quando se tem

$$\lim_{h\to 0} \| \eta_k(h) - y(x_k) \| = 0 \ , \ k = 0,1,\ldots,p-1 \qquad (1.11)$$

podemos assumir que, para h "suficientemente pequeno",- possivelmente sob hipóteses suplementares - a solução y_k^* de um método de diferenças consistente aproxima a solução $y(x_k)$ da equação diferencial (1.1) nos pontos $x_k \in I_h$. Como veremos adiante (§ 1.4,teorema 1.5) isto realmente acontece.

Muitas vezes (1.11) é chamado de <u>condição de consistência para os valores iniciais</u> pois, em casos mais gerais, podemos representar os valores iniciais e de contorno de $T[y](x) = \sigma$, pelos valores $U[y](x) = \sigma$ de um operador U e então definir a consistência da discretização U_h de U.

<u>Observação</u>:

Mostraremos no § 1.4 que a solução y_k^* de um método com ordem de consistência q converge sob hipóteses adequadas a $y(x_k)$ com ordem q em h ($\mathcal{O}(h^q)$). Por isso não distinguiremos a ordem de consistência da ordem de convergência e nos referiremos sempre apenas à <u>ordem</u> do método.

<u>Exemplos</u>:

1. A regra do ponto médio(1.7) é consistente e tem ordem 2 se $y \in \mathbb{C}^3[a,b]$, pois

$$\|T_h[y](x)\| \equiv \| \frac{1}{2h} (y(x+h) - y(x-h)) - f(x,y(x)) \|$$

$$= \| \frac{1}{2h} (y(x+h) - y(x-h)) - y'(x) \|$$

$$= \frac{h^2}{3!} \| y'''(\xi) \| \quad ; \quad \xi \in [a,b]$$

Além disso, como é óbvio, vale (1.11), isto é:

$$\lim_{h\to 0} \| y_0^* + hf(x_0, y_0^*) - y(x_1) \| = 0$$

Igualmente se demonstra que T_h tem ordem 1 quando apenas $y \in \mathbb{C}^2[a,b]$. Como, para todo $z \in \mathbb{C}^3[a,b]$, temos

$$\| T_h[z](x) - T[z](x) \| = \frac{h^2}{3!} \| z'''(\xi) \| ,$$

T_h é, para h suficientemente pequeno, uma boa aproximação do operador T até mesmo em <u>todo</u> o espaço $\mathbb{C}^3[a,b]$ (e não só para a solução y de (1.1)).

2. - O método (1.9) de Milne-Simpson é consistente e tem ordem 4 quando $y \in C^5[a,b]$ pois

$$\|T_h[y](x)\| = \left\| \frac{y(x+h)-y(x-h)}{2h} - \frac{1}{6}\left\{f(x+h,y(x+h))+4f(x,y(x))+f(x-h,y(x-h))\right\}\right\|$$

$$= \left\| \frac{y(x+h)-y(x-h)}{2h} - \frac{1}{6}\left(y'(x+h) + 4y'(x) + y'(x-h)\right)\right\|$$

$$= \left\| y'(x)+\frac{h^2}{3!}y'''(x)+\frac{h^4}{5!}y^{(5)}(\xi)-\frac{1}{6}\left[6y'(x)+h^2y'''(x)+\frac{h^4}{12}y^{(5)}(\eta)\right]\right\|$$

$$\leq Ch^4 \text{ com } C = \frac{1}{45}\max_{x\in[a,b]}\|y^{(5)}(x)\| \; ; \; x<\xi,\eta<x+h \qquad (1.12)$$

3. - Seja a equação diferencial parcial ~~hiper~~parabólica

$$T[u](x,t) \equiv u_t - \sigma u_{xx} - f(x,t,u) = 0 \; ; \; \sigma > 0 \qquad (1.13)$$

$$u(x,0) = \eta(x) \quad ; \quad 0 \leq x \leq b$$

$$u(0,t) = u(b,t) = 0 \; ; \; 0 \leq t \leq c$$

aproximada por

a) $$T^{(a)}_{\Delta x,\Delta t}[u^*](x,t) \equiv \frac{u^*(x,t+\Delta t)-u^*(x,t)}{\Delta t} - \sigma\,\frac{u^*(x+\Delta x,t)-2u^*(x,t)+u^*(x-\Delta x,t)}{(\Delta x)^2} -$$

$$- f(x,t,u^*) = 0 \qquad (1.14a)$$

b)

$$T^{(b)}_{\Delta x,\Delta t}[u^*](x,t) \equiv \frac{u^*(x,t+\Delta t)-u^*(x,t-\Delta t)}{2\Delta t} - \sigma\,\frac{u^*(x+\Delta x,t)-2u^*(x,t)+u^*(x-\Delta x,t)}{(\Delta x)^2}$$

$$- f(x,t,u^*) = 0 \qquad (1.14b)$$

c)

$$T^{(c)}_{\Delta x,\Delta t}[u^*](x,t) \equiv \frac{u^*(x,t+\Delta t)-u^*(x,t)}{\Delta t} - \sigma\,\frac{u^*(x+\Delta t,t+\Delta t)-2u^*(x,t+\Delta t)+u^*(x-\Delta x,t+\Delta t)}{(\Delta x)^2}$$

$$- f(x,t,u^*) = 0 \qquad (1.14c)$$

com $\Delta x = \frac{b}{m_1}$; $\Delta t = \frac{c}{m_2}$; $m_1, m_2 \in \mathbb{N}$

então, obtemos com

$$x_j = j\Delta x \; ; \; t_k = k\Delta t \; ; \; u(x_j,t_k) = u_j^{(k)} \; ; \; f(x_j,t_k;u(x_j,t_k)) = f_j^{(k)}$$

e $j = 1,2,\ldots,m_1 - 1$:

a) O método explícito:

$$u_j^{*(k+1)} = \left(1 - \frac{2\sigma\Delta t}{(\Delta x)^2}\right)u_j^{*(k)} + \frac{\sigma\Delta t}{(\Delta x)^2}\left(u_{j+1}^{*(k)} + u_{j-1}^{*(k)}\right) + \Delta t f_j^{(k)} \; ; \qquad (1.15a)$$

$$k = 0,1,\ldots,m_1 - 1$$

b) o método de passo 2:

$$u_j^{*(k+1)} = u_j^{*(k-1)} + \frac{2\sigma\Delta t}{(\Delta x)^2}\left(u_{j+1}^{*(k)} - 2u_j^{*(k)} + u_{j-1}^{*(k)}\right) + 2\Delta t\, f_j^{(k)}; \qquad (1.15b)$$

$$k = 0,1,\ldots,m_1 - 1$$

c) o método implícito (Crank-Nicolson):

$$-\frac{\sigma\Delta t}{(\Delta x)^2} u_{j+1}^{*(k+1)} + \left(1 + \frac{2\sigma\Delta t}{(\Delta x)^2}\right) u_j^{*(k+1)} - \frac{\sigma\Delta t}{(\Delta x)^2} u_{j-1}^{*(k+1)} = u_j^{*} + \Delta t f_j^{(k)}; \qquad (1.15c)$$

$$k = 0,1,\ldots,m_1 - 1$$

Numa generalização natural da definição (1.2) chamamos $T_{\Delta x,\Delta t}$ de consistente quando vale, para a solução u de (1.13)

$$\lim_{\Delta x,\Delta t \to 0} \max_{x,t\in I'} \left\| T_{\Delta x,\Delta t}[u](x,t) \right\| = 0 \qquad (1.16)$$

onde I' indica o conjunto de todos os x e t para os quais $T_{\Delta x,\Delta t}$ é definido. $T_{\Delta x,\Delta t}$ tem ordem de consistência $q = \min(r,s)$ se, para a solução $u(x,t)$ de (1.13):

$$\max_{x,t\in I'} \left\| T_{\Delta x,\Delta t}[u](x,t) \right\| = \mathcal{O}\left((\Delta x)^r\right) + \mathcal{O}\left((\Delta t)^s\right)$$

Obtemos, por desenvolvimento de Taylor,

$$T^{(a)}_{\Delta x,\Delta t}[u](x,t) = \frac{\Delta t}{2!} u_{tt}(x,t) - \frac{(\Delta x)^2}{12} u_{xxxx}(x,t) + \text{termos de ordem maior em } \Delta x, \Delta t$$

$$T^{(b)}_{\Delta x,\Delta t}[u](x,t) = \frac{(\Delta t)^2}{3!} u_{ttt}(x,t) - \frac{\sigma(\Delta x)^2}{12} u_{xxxx}(x,t) + \text{termos de ordem maior em } \Delta x, \Delta t$$

$$T^{(c)}_{\Delta x,\Delta t}[u](x,t) = \frac{\Delta t}{2} \frac{\partial}{\partial t} f(x,t,u(x,t)) + \frac{\sigma}{12} \frac{(\Delta t)^4}{(\Delta x)^2} u_{tttt}(x,t) + \text{termos com } (\Delta x)^r,\ (\Delta t)^r,\ \frac{(\Delta t)^{r+2}}{(\Delta x)^r} \quad r = 2,3,\ldots.$$

Por isso os métodos a) e b) são consistentes e tem ordens, respectivamente, 1 e 2 se u for suficientemente diferenciável.

No caso do procedimento c), (1.16) somente é satisfeita quando existe uma relação entre Δx e Δt tal que $\lim_{\Delta x,\Delta t \to 0} \frac{(\Delta t)^2}{\Delta x} = 0$.

Neste caso o método é chamado condicionalmente consistente.

Então o procedimento c) tem ordem 2 para $\Delta t = c\Delta x$ quando $\frac{\partial}{\partial t} f(x,t,u(x,t)) \equiv 0$; senão tem ordem 1.

Estudando a consistência dos métodos de passo simples ou múltiplo da forma (1.5) ou (1.6), chegamos aos seguintes teoremas de consistência:

Teorema 1.1

O método de passo simples $y^*_{k+1} = y^*_k + h\phi(x_k, y^*_k; h)$; $y^*_0 = \eta_0$ que aproxima a solução $y(x)$ de (1.1) é consistente se $\phi(x,y;0) \equiv f(x,y)$ e $\phi(x,y;h)$ é continua na região G^*, $G^* = \{a \le x \le b;\ y \text{ qualquer};\ 0 \le h \le h_0\}$. Se, além disso, a função f tem todas suas derivadas parciais de ordem q contínuas e se

$$\phi(x,y;h) = f(x,y) + \frac{h}{2!} Df(x,y) + \ldots + \frac{h^{q-1}}{q!} D^{q-1} f(x,y) + \mathcal{O}(h^q) \qquad (1.17)$$

então o método acima tem ordem q.

$Df(x,y)$ é definido por $Df(x,y) = \left(\frac{\partial}{\partial x} + f \frac{\partial}{\partial y}\right) f(x,y)$.

Por exemplo:

$$Df = f_x + f f_y;$$

$$D^2 f = \left(\frac{\partial}{\partial x} + f \frac{\partial}{\partial y}\right)\left(\frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y}\right)$$

$$= f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_x f_y + f f_y^2 \ ; \ \text{etc...}$$

Demonstração:

Para y vale:

$$\max_{x \in I'_h} \| T_h[y](x) \| = \max_{x \in I'_h} \left\| \frac{y(x+h)-y(x)}{h} - \phi(x,y(x);h) \right\| \le$$

$$\le \max_{x \in I'_h} \| y'(x+\vartheta h) - \phi(x,y(x);0) \| + \max_{x \in I'_h} \| \phi(x,y(x);0) - \phi(x,y(x);h) \|$$

$$0 < \vartheta < 1$$

Desde que $\phi(x,y(x);0) = f(x,y(x)) = y'(x)$, $y \in \mathbb{C}^1[a,b]$ e tendo em vista a continuidade uniforme de $\phi(x,y(x);h)$ na região limitada $a \le x \le b$; $0 \le h \le h_0$, esta expressão tende para 0 quando $h \Rightarrow 0$. Logo o método é consistente. Se todas as derivadas parciais de ordem q de f são contínuas, então $y \in \mathbb{C}^{q+1}[a,b]$ e

$$\| T_h[y](x) \| = \left\| \frac{y(x+h)-y(x)}{h} - \phi(x,y(x);h) \right\|$$

$$= \left\| y'(x) + \frac{h}{2!} y'(x) + \ldots + \frac{h^q}{(q+1)!} y^{(q+1)}(\xi) - \phi(x,y(x);h) \right\| ;$$

$$x < \xi < x+h$$

$$= \mathcal{O}(h^q) \ ; \ \text{pois de } y' = f(x,y(x)) \text{ segue}$$

$$y^{(k+1)}(x) = D^k f(x,y(x)); \ k = 1,2,\ldots,q-1$$

Teorema 1.2

O método de passo $p \geq 1$ da forma

$$\alpha_p y^*_{k+p} + \alpha_{p-1} y^*_{k+p-1} + \ldots + \alpha_0 y^*_k - h\{\beta_p f_{k+p} + \beta_{p-1} f_{k+p-1} + \ldots + \beta_0 f_k\} = \emptyset;$$

$$\alpha_p = 1$$

é consistente (com (1.1)) se

$$\sum_{j=0}^{p} \alpha_j = 0 \quad ; \quad \sum_{j=1}^{p} j.\alpha_j = \sum_{j=0}^{p} \beta_j \tag{1.18}$$

e tem ordem q se $y \in \mathbb{C}^{q+2}[a,b]$ e se

$$c_0 = c_1 = \ldots = c_q = 0 \quad ; \quad c_{q+1} \neq 0$$

onde as constantes c_j são definidos por

$$c_0 = \sum_{j=0}^{p} \alpha_j \; ; \; c_1 = \sum_{j=1}^{p} j\alpha_j - \sum_{j=0}^{p} \beta_j \; ; \tag{1.19}$$

$$c_k = \frac{1}{k!} \sum_{j=1}^{p} j^k \alpha_j - \frac{1}{(k-1)!} \sum_{j=1}^{p} j^{k-1} \beta_j \; ; \; k = 2,3,\ldots,q+1$$

Demonstração:

Para $y \in \mathbb{C}^1[a,b]$ temos:

$$\| T_h[y](x) \| = \| \frac{1}{h} \sum_{j=0}^{p} \alpha_j y(x+jh) - \sum_{j=0}^{p} \beta_j f(x+jh, y(x+jh)) \|$$

$$= \| \frac{y(x)}{h} \sum_{j=0}^{p} \alpha_j + \sum_{j=0}^{p} j y'(\xi_j)\alpha_j - \sum_{j=0}^{p} \beta_j f(x+jh, y(x+jh)) \| \; ;$$

$$x \leq \xi_j \leq x+jh$$

Como $\lim_{h \to 0} y'(\xi_j) = y'(x) = f(x,y(x)) \; ; \; j = 1,2,\ldots,p$

$$\lim_{h \to 0} \sum_{j=0}^{p} \beta_j f(x+jh, y(x+jh)) = f(x,y(x)). \sum_{j=0}^{p} \beta_j$$

temos $\lim_{h \to 0} \| T_h[y](x) \| = 0$ se a equação (1.18) é satisfeita.

Se $y \in \mathbb{C}^{q+2}[a,b]$, então obtemos de $f(x,y(x)) = y'(x)$ e do desenvolvimento de Taylor:

$$\| T_h[y](x) \| = \| \frac{1}{h} \sum_{j=0}^{p} \alpha_j y(x+jh) - \sum_{j=0}^{p} \beta_j f(x+jh, y(x+jh)) \|$$

$$= \| \frac{c_0}{h} y(x) + c_1 y'(x) + c_2 h y''(x) + \ldots + c_q h^{q-1} y^{(q)}(x) + c_{q+1} h^q y^{(q+1)}(x) + \emptyset(h^{q+1}) \|$$

$$= \mathcal{O}(h^q) \text{ , se } c_0 = c_1 = \ldots = c_q = 0; \quad c_{q+1} \neq 0$$

o que demonstra a segunda parte do teorema.

Caracterizando o método (1.6) por seus polinômios geradores

$$\rho(z) = z^p + \alpha_{p-1}z^{p-1} + \ldots + \alpha_1 z + \alpha_0$$

$$\sigma(z) = \beta_p z^p + \beta_{p-1}z^{p-1} + \ldots + \beta_1 z + \beta_0 \qquad (1.20)$$

então as condições de consistência (1.18) se simplificam na forma

$$\rho(1) = 0 \quad ; \quad \rho'(1) = \sigma(1) \qquad (1.21)$$

§ 1.3 - ESTABILIDADE PARA MÉTODOS DE DISCRETIZAÇÃO GERAL.

A consistência do operador T_h garante uma "boa aproximação" da equação $T[y](x) = 0$ pela equação de diferenças $T_h[y^*](x) = 0$, o que não implica entretanto, que também as soluções destas equações se aproximam.

Existem por exemplo equações de diferenças $T_h[y^*](x) = 0$ cuja solução, para uma pequena mudança nas condições iniciais ou no pró-

prio operador T_h sofre mudanças "arbitrariamente grandes" (apenas quando h é "suficientemente pequeno"). Tais equações são chamadas **instáveis**.

1.3.1 - Exemplo:

Aproximemos a equação

$$T[y](x) \equiv y'(x) + y(x) = 0 \; ; \quad y(0) = 1, \qquad (1.22)$$

cuja solução é $y(x) = e^{-x}$,

por $T_h[y^*](x) = \frac{-y^*(x+h)+4y^*(x)-3y^*(x-h)}{2h} + y^*(x) = 0; \; y^*(0) = 1;$

$$y^*(h) = \eta_1 \qquad (1.23)$$

É fácil demonstrar que (1.23) é consistente com (1.22).

Para $x_k = kh$ e $y^*(hk) = y_k^*$ obtemos a equação de diferenças:

$$y_{k+1}^* - (4+2h)y_k^* + 3y_{k-1}^* = 0; \; y_0^* = 1 \; ; \; y_1^* = \eta_1 \qquad (1.24)$$

Pela substituição $y_k^* = Cr^k$ temos a equação característica:

$$r^2 - (4 + 2h)r + 3 = 0 \qquad (1.25)$$

A solução geral de (1.24) é

$$y_k^* = C_1 r_1^k + C_2 r_2^k$$

sendo $r_1(h)$ e $r_2(h)$ as raízes de (1.25):

$$r_1(h) = (2+h) + \sqrt{1+4h+h^2} = 3e^h + \mathcal{O}(h^2)$$

$$r_2(h) = (2+h) - \sqrt{1+4h+h^2} = e^{-h} + \mathcal{O}(h^2)$$

Das condições iniciais de (1.24) segue:

$$1 = C_1 + C_2; \qquad \eta_1 = C_1 r_1 + C_2 r_2$$

$$C_1 = \frac{r_1 - \eta_1}{r_2 - r_1}; \qquad C_2 = \frac{r_1 - \eta_1}{r_1 - r_2}$$

Assim obtemos

$$y_k^* = \frac{r_2 - \eta_1}{r_2 - r_1}(3e^h + \mathcal{O}(h^2))^k + \frac{r_1 - \eta_1}{r_1 - r_2}(e^{-h} + \mathcal{O}(h^2))^k$$

e com $hk = x_k$:

$$y_k^* = \frac{r_2 - \eta_1}{r_2 - r_1} \cdot 3^{x_k/h} e^{x_k} + \frac{r_1 - \eta_1}{r_1 - r_2} e^{-x_k} + \mathcal{O}(h^2) \qquad (1.26)$$

Para $\eta_1 = r_2$, (1.26) aproxima a solução $y(x_k) = e^{-x_k}$ da equação diferencial (1.22) (com ordem 2 em h).

Entretanto, se η_1 difere, mesmo que "ligeiramente", de r_2, (o que pode acontecer por arredondamento), então chegamos a soluções (1.26) que, para h "suficientemente pequeno", divergem arbitrariamente de e^{-x_k}. Por isso o procedimento (1.24) é dito instável.

Notamos que a instabilidade é devida ao termo $3^{\frac{x_k}{h}}$, isto é, ao fato que $r_1(0) = 3$. Veremos em 1.3.3 que os métodos de passo múltiplo sempre são instáveis quando vale, para uma raiz r(h) da equação característica: $|r(0)| > 1$.

1.3.2. - Definição de estabilidade.

A existência de um método numérico, mesmo convergente para a solução de um problema matemático, não implica necessariamente que a solução (ou uma sua "boa" aproximação numérica) pode ser calculada por este método. Esta contradição aparente pode ser explicada facilmente pelo exemplo 1.3.1: o procedimento (1.24) converge, quando $h \to 0$ e $\eta_1 = r_2$, para a solução $y(x) = e^{-x}$ da equação diferencial (1.22). Entretanto, ele não pode ser usado na prática para determinar "boas" aproximações de y(x) pois, como mostramos acima, para h "pequeno" estas aproximações são distorcidas por erros de arredondamento inevitáveis. Por isso um

método de discretização somente tem utilidade prática quando a influência de erros é limitada para todos os valores do parâmetro de discretização h[1] . Neste caso chamamos este procedimento de estável. Então a estabilidade é, essencialmente, a propriedade que permite a aplicação prática de um método.

Vamos agora tornar mais preciso o conceito da estabilidade:

Sejam $\{v_j\}_h$ e $\{w_j\}_h$ soluções dos métodos de passo $p \geq 1$:

$$T_h[v](x_j) + S_h[v](x_j) = \mathcal{O} \quad \text{com} \quad v(x_j) = \eta_j + \mu_j$$

$$j = 0,1,\ldots,p-1$$

$$T_h[w](x_j) + R_h[w](x_j) = \mathcal{O} \quad \text{com} \quad w(x_j) = \eta_j + \omega_j$$

onde consideramos S_h, R_h, μ_j e ω_j como perturbações na equação $T_h[y^*](x_j) = \mathcal{O}$, $y^*(x_j) = \eta_j$ $(j = 0,1,\ldots,p-1)$ surgidas de arredondamento ou truncamento. Se S_h e R_h "diferem pouco" para $h \in (0,h_0]$, isto é: para um certo δ_1 "suficientemente pequeno",

$$\| S_h[v](x_j) - R_h[w](x_j) \| < \delta_1 \ , \ j = 0,\ldots,m-p$$

1) Por exemplo, em (1.26) para $\eta_1 \neq r_2$, $(y_k^* - e^{-x_k})$ não é limitado para "pequenos" valores de h.

e os valores iniciais também "diferem pouco", isto é: para um certo δ_2 "suficientemente pequeno" $\| \mu_j - \omega_j \| < \delta$, $j = 0,1,\ldots,p-1$, então chamaremos $T_h[y^*](x_j) = \mathcal{O}$ estável quando v_j e w_j , $j = p,p+1,\ldots,m$ diferem pouco. Definimos assim, de um modo um pouco menos geral:

Definição 1.3

Um método de passo $p \geq 1$

$$T_h[y^*](x_j) = \mathcal{O} \ , \ y^*(x_j) = \eta_j \quad j = 0,1,\ldots,p-1$$

é chamado estável quando existe uma constante M independente de h tal que, para funções de rede arbitrárias $\{v_j\}_h, \{w_j\}_h$ vale:

$$\| v_k - w_k \| \leq M\{ \max_{0 \leq j \leq p-1} \| v_j - w_j \| + \max_{0 \leq j \leq m-p} \| T_h[v](x_j) - T_h[w](x_j) \| \}$$

$$k = 0,1,\ldots,m-1 \quad (1.27)$$

Caso contrário o método é dito instável.

Se T_h é linear então a condição (1.27), para $z_k = v_k - w_k$, se reduz a

$$\| z_k \| \leq M\{ \max_{0 \leq j \leq p-1} \| z_j \| + \max_{0 \leq j \leq m-p} \| T_h[z](x_j) \| \} \quad (1.28)$$

Exemplo:

Consideramos o método de Euler:

$$h\, T_h[y^*](x_k) \equiv y_{k+1}^* - y_k^* - hf(x_k, y_k^*) = 0 \quad (1.29)$$

Para funções de rede arbitrárias $\{v_j\}_h$ e $\{w_j\}_h$ obtemos as identidades

$$v_{k+1} = v_k + hf(x_k,v_k) + hT_h[v](x_k)$$

$$w_{k+1} = w_k + hf(x_k,w_k) + hT_h[w](x_k)$$

e com a condição (1.2) de Lipschitz,

$$\|v_{k+1} - w_{k+1}\| \le (1+hL)\|v_k - w_k\| + h \max_{0\le j\le m-1} \|T_h[v](x_j) - T_h[w](x_j)\| \ ;$$

$$k = 0,1,\ldots,m-1$$

Fazendo $(1+hL) = A$; $\max_{0\le j\le m-1} \|T_h[v](x_j) - T_h[w](x_j)\| = B$ então

$$\|v_1 - w_1\| \le A\|v_0 - w_0\| + hB$$

$$\|v_2 - w_2\| \le A\|v_1 - w_1\| + hB \le A^2\|v_0 - w_0\| + hB(A+1)$$

$$\|v_3 - w_3\| \le A\|v_2 - w_2\| + hB \le A^3\|v_0 - w_0\| + hB(A^2+A+1); \text{ etc.}$$

Por indução obtemos:

$$\|v_k - w_k\| \le A^k\|v_0 - w_0\| + hB(A^{k-1} + A^{k-2} + \ldots + A+1)$$

$$\le A^k\|v_0 - w_0\| + hB\,\frac{A^k - 1}{A - 1}$$

$$\|v_k - w_k\| \le (1+hL)^k\|v_0 - w_0\| + L^{-1}((1+hL)^k - 1) \max_{0\le j\le m-1} \|T_h[v](x_j) - T_h[w](x_j)\| \quad (1.30)$$

Com $(1+hL)^k \le e^{hkL} = e^{L(x_k - a)}$ e $M = e^{(b-a)L}\max(1,L^{-1})$ obtemos

$$\|v_k - w_k\| \le M\{\|v_0 - w_0\| + \max_{0\le j\le m-1} \|T_h[v](x_j) - T_h[w](x_j)\|\}$$

e assim, a estabilidade do método de Euler.

Fazendo $v_k = y(x_k)$ e $w_k = y_k^*$, sendo y_k^* a solução de $T_h[y^*](x_k) = \sigma$; $y_0^* = \eta_0$ e y a solução de (1.1), obtemos de (1.30) a seguinte <u>estimativa do erro</u> do método de Euler.

$$\|y(x_k) - y_k^*\| \le \frac{(1+hL)^k - 1}{L} \max_{0\le j\le m-1} \|T_h[y](x_j)\|$$

De

$$\|T_h[y](x_j)\| = \left\| \frac{y(x_j+h) - y(x_j)}{h} - f(x_j, y(x_j)) \right\| = \frac{h}{2}\|y''(\xi_j)\| \le hK$$

$$x_j < \xi_j < x_j + h$$

segue, se $y \in C^2[a,b]$ e $K = \frac{1}{2}\max_{x\in[a,b]} \|y''(x)\|$:

$$\|y(x_k) - y_k^*\| \le hKL^{-1}\{(1+hL)^k - 1\} \le hKL^{-1}\, e^{L(x_k - a)}$$

Então y_k^* converge para $y(x_k)$ (pelo menos com ordem 1 em h).

<u>Observação</u>:

Se na determinação de y_{k+1}^*, a partir de (1.29), for intro-

duzido, em cada passo, um erro de arredondamento ε_k tal que, para um certo $\varepsilon > 0$, $0 < \|\varepsilon_k\| < \varepsilon$, $k = 1,2,\ldots,m$, isto é, se considerarmos, ao invés do operador T_h em (1.29) o operador $\hat{T}_h$, com

$$h\hat{T}_h[y^*](x_k) = y^*_{k+1} - y^*_k - hf(x_k,y^*_k) - \varepsilon_k$$

então, $\hat{T}_h$ <u>não</u> é consistente com T e o erro de substituição $\|\hat{T}_h[y](x_k)\|$, (y : solução de (1.1)) <u>não é limitado</u> em $(0,h_0]$. Obtemos neste caso $\|T_h[y](x_k)\| \le hK + \frac{\varepsilon}{h}$, e, assim, a estimativa do erro:

$$\|y(x_k) - y^*_k\| \le (hK + \frac{\varepsilon}{h})L^{-1}\{(1+hL)^k - 1\}$$

Observamos que a parte do erro, devido ao erro de arredondamento cresce ilimitadamente quando $h \to 0$; apesar disso o método é estável em termos da nossa definição!

1.3.3. - <u>Um Critério Geral de Estabilidade</u>.

Os resultados da seção anterior podem ser consideravelmente generalizados pelo teorema seguinte que forma a base para considerações de estabilidade e consistência, assim como estimativas de erro para métodos muito gerais de passo simples ou múltiplo.

<u>Teorema 1.3</u>

Seja T_h definido por

$$hT_h[y](x) = y(x+h) - Ay(x) - h\phi(x,y(x),y(x+h);h)$$

com $A \in \mathbb{R}(n,n)$, $y:[a,b] \to \mathbb{R}^n$; $\phi:[a,b]\times\mathbb{R}^n\times\mathbb{R}^n\times[0,h_0] \to \mathbb{R}^n$. Suponhamos que, para $v,w,y,z \in \mathbb{R}^n$, $x \in [a,b]$, $h \in (0,h_0]$

$$\|\phi(x,v,y;h) - \phi(x,w,z;h)\| \le K_1\|v-w\| + K_2\|y-z\| \qquad (1.31)$$

com K_1 e K_2 constantes, onde $h_0K_2 < 1$.

Então o procedimento

$$y^*_{k+1} = Ay^*_k + h\phi(x_k,y^*_k,y^*_{k+1},h) \; ; \; y^*_0 = \eta_0 \qquad (1.32)$$

$$k = 0,1,\ldots,m-1$$

é estável se e somente se existe um numero $D>0$ tal que $\|A^j\| \le D$ para todo $j \in \mathbb{N}$. Neste caso, vale para uma função de rede arbitrária $\{v_k\}_h$:

$$\|v_k - y^*_k\| \le De^{hkDE}\|v_0 - \eta_0\| + \frac{(e^{hkDE}-1)}{E(1-hK_2)} \max_{0\le j\le m-1} \|T_h[v](x_j)\| \qquad (1.33)$$

com $E = \dfrac{K_1 + K_2\|A\|}{1 - hK_2}$

<u>Observações</u>:

1. Desde que $h_0K_2 < 1$, y^*_{k+1} é unicamente determinado por (1.32) e pode ser calculado por iteração;

2. O número D existe se e somente se:

   a. $\rho(A) \leq 1$, $\rho(A)$: raio espectral de A;

   b. para os autovalores de A de módulo 1, (se existem), só estão associados divisores elementares de grau 1.

A demonstração é uma generalização da demonstração de estabilidade do método de Euler.

Para $v,w,y,z \in \mathbb{R}^n$ quaisquer, com $\max_j |v_j - w_j| = |v_J - w_J|$ e $\max_j |y_j - z_j| = |y_L - z_L|$, definimos dois operadores lineares $K^{(1)}(v,w) = (k_{ij}^{(1)})$ e $K_h^{(2)}(y,z) = (k_{ij}^{(2)})$ por

$$k_{ij}^{(1)}(v,w) = \begin{cases} \dfrac{\phi_i(x,v,y;h) - \phi_i(x,w,y;h)}{v_J - w_J} & \text{para } j=J \text{ e } v \neq w \\ 0 & \text{nos outros casos} \end{cases}$$

$$k_{ij}^{(2)}(y,z) = \begin{cases} \dfrac{\phi_i(x,w,y;h) - \phi_i(x,w,z;h)}{y_L - z_L} & \text{para } j=L \text{ e } y \neq z \\ 0 & \text{nos outros casos} \end{cases}$$

Então, $\phi(x,v,y;h) - \phi(x,w,z;h) = K_h^{(1)}(v,w)(v-w) + K_h^{(2)}(y,z)(y-z)$ e usando (1.31)

$$\| K_h^{(1)}(v,w) \| = \max_i |k_{iJ}^{(1)}(v,w)| \leq K_1$$

$$\| K_h^{(2)}(y,z) \| = \max_i |k_{iL}^{(2)}(y,z)| \leq K_2$$

obtem-se, a partir das identidades

$$v_{k+1} = Av_k + h\phi(x_k, v_k, v_{k+1}; h) + hT_h[v](x_k)$$

$$w_{k+1} = Aw_k + h\phi(x_k, w_k, v_{k+1}; h) + hT_h[w](x_k)$$

para arbitrárias funções de rede $\{v_k\}_h$ e $\{w_k\}_h$ a equação:

$$v_{k+1} - w_{k+1} = A(v_k - w_k) + hK_h^{(1)}(v_k, w_k)(v_k - w_k) + hK_h^{(2)}(v_{k+1}, w_{k+1})(v_{k+1} - w_{k+1})$$
$$+ h(T_h[v](x_k) - T_h[w](x_k))$$

Usando as abreviações

$$A_k = \left(I - hK_h^{(2)}(v_{k+1}, w_{k+1})\right)^{-1}\left(A + hK_h^{(1)}(v_k, w_k)\right)$$

$$B_k = \left(I - hK_h^{(2)}(v_{k+1}, w_{k+1})\right)^{-1}\left(T_h[v](x_k) - T_h[w](x_k)\right) \qquad (1.34)$$

(a inversa existe desde que $hK_2 < 1$) obtemos, resolvendo para $(v_{k+1} - w_{k+1})$:

$$v_{k+1} - w_{k+1} = A_k(v_k - w_k) + hB_k \; ; \; k = 0,1,\ldots,m-1$$

Por indução:

$$v_k - w_k = A_{k-1}A_{k-2}\cdots A_0(v_0 - w_0) + h(A_{k-1}\cdots A_1 B_0 + A_{k-1}\cdots A_2 B_1 + \cdots + A_{k-1}B_{k-2} + B_{k-1}) \qquad (1.35a)$$

$$\| v_k - w_k \| \le \| P_{k-1}^{(0)} \| \|v_0 - w_0\| + h(\| P_{k-1}^{(1)}\| + \| P_{k-1}^{(2)}\| + \cdots + \|P_{k-1}^{(k-1)}\| + 1) \max_{0\le j\le k-1} \|B_j\| \qquad (1.35b)$$

onde $P_{k-1}^{(r)} = A_{k-1}A_{k-2}\cdots A_r$; $r = 0,1,\ldots,k-1$.

Escrevendo A na forma

$$A = (I - hK_h^{(2)}(v_{k+1},w_{k+1}))A + hK_h^{(2)}(v_{k+1},w_{k+1})A$$

obtemos de (1.34) que $A_k = A + hE_k$,

com $$E_k = (I - hK_h^{(2)}(v_{k+1},w_{k+1}))^{-1}\left(K_h^{(1)}(v_k,w_k) + K_h^{(2)}(v_{k+1},w_{k+1})A\right)$$

e $$\| E_k \| \le E = \frac{K_1 + K_2\|A\|}{1 - hK_2}, \quad hK_2 < 1 .$$

$P_{k-1}^{(r)}$, $0 \le r \le k-1$ consiste de 2^{k-r} termos, dos quais $\binom{k-r}{s}$ consistem de um produto de s fatores da forma hE_j e $(k-r-s)$ fatores A. Cada um desses $\binom{k-r}{s}$ termos contem no máximo $s+1$ grupos de A consecutivos, pois só existem s operadores da forma hE_j para separar os

A_s'. A norma de cada um destes grupos é limitada por D. Com $\| E_j \| < E$ obtemos então:

$$\| P_{k-1}^{(r)} \| \le \sum_{s=0}^{k-r} \binom{k-r}{s} D^{s+1}(hE)^s = D(1+hED)^{k-r}; \quad 0\le r\le k-1$$

Daí segue:

$$1 + \sum_{r=1}^{k-1} \| P_{k-1}^{(r)} \| \le \frac{D(1+hED)^{k-1}}{hED}$$

Considerando

$$\| B_j \| \le \| (I - hK_h^{(2)}(v_{j+1},w_{j+1}))^{-1} \| \, \| T_h[v](x_j) - T_h[w](x_j) \|$$

$$\le (1-hK_2)^{-1} \| T_h[v](x_j) - T_h[w](x_j) \|$$

obtemos de (1.35b) a relação

$$\|v_k - w_k\| \le D(1+hED)^k \|v_0 - w_0\| + hD \frac{(1+hED)^{k-1}}{(hED(1-hK_2)} \max_{0\le j\le m-1} \| T_h[v](x_j) - T_h[w](x_j) \|$$

Com $w_k = y_k^*$, $T_h[y^*](x_k) = 0$ e $(1+hED) \le e^{hED}$ obtemos o erro (1.33).

Além disso, fazendo $hk \le H$ e $M = \max\left\{De^{HED}, \frac{e^{HED} - 1}{K_1 + K_2\|A\|}\right\}$, é satisfeita a condição de estabilidade (1.27).

Como (1.32) é estável quando $\phi \equiv 0$, obtemos neste caso, para z_0 qualquer (vide (1.28))

$$\| z_{k+1} \| = \| A^k z_0 \| \leq M \| z_0 \| .$$

Segue-se, como sabemos da teoria das matrizes, que $\rho(A) \leq 1$ e que os divisores elementares associados aos autovalores de módulo 1, são de grau 1.

A ampla significação do teorema 1.3 verifica-se nas aplicações seguintes:

1.3.4 - Estabilidade e estimativa do erro para métodos de passo simples.

Se $A=I$ (matriz identidade) e se $\phi(x,v,w,h)$ é independente de w, então (1.32) toma a forma (1.5) dos métodos de passo simples. Neste caso o teorema 1.3 com D=1 resulta nas seguintes condições de estabilidade e de estimativa de erro:

Corolário 1.3a

Métodos de passo simples da forma

$$y^*_{k+1} = y^*_k + h\phi(x_k, y^*_k; h) \quad ; \quad y^*_0 = \eta_0$$

$y^*_k \in \mathbb{R}^n$; $\phi : [a,b] \times \mathbb{R}^n \times [0,h] \Rightarrow \mathbb{R}^n$

são estáveis se ϕ satisfaz à condição de Lipschitz:

$$\| \phi(x,y;h) - \phi(x,\hat{y};h) \| \leq K \| y - \hat{y} \| .$$

Se y é solução de (1.1) então vale a estimativa do erro

$$\| y(x_k) - y^*_k \| \leq K^{-1}(e^{(x_k-a)K} - 1) \max_{x \in [a,b]} \left\| \frac{y(x+h)-y(x)}{h} - \phi(x,y(x);h) \right\| \qquad (1.36)$$

Sob hipótese adicionais podemos obter da demonstração do teorema 1.3 outras informações sobre o erro dos métodos de passo simples:

Se $n = 1$ então $A, K_h^{(1)}$ e $K_h^{(2)}$ são os números: $A=1, K_h^{(2)} = 0$

e $K_h^{(1)}(v,w) = \dfrac{\phi(x,v;h)-\phi(x,w;h)}{v-w}$ para $v \neq w$ ou

$K_h^{(1)}(v,w) = 0$ para $v = w$.

Da condição de Lipschitz, segue a existência de dois números M_1 e M_2 com $M_1 \leq K_h^{(1)}(v,w) \leq M_2$. Se escolhemos h "não pequeno" tal que $(1 + hM_1) \geq 0$, então $0 \leq (1+hM_1) \leq A_j \leq (1 + hM_2)$ e por isso

$$\frac{(1+hM_1)^k - 1}{hM_1} = \sum_{r=1}^{k} (1+hM_1)^{k-r} \leq 1 + \sum_{r=1}^{k-1} P_{k-1}^{(r)} \leq \sum_{r=1}^{k} (1+hM_2)^{k-r} = \frac{(1+hM_2)^k - 1}{hM_2}$$

Para $v_k = y(x_k)$, $w_k = y^*_k$, $y^*_0 = \eta_0$ e $B_j = T_h[y](x_j)$, sendo

y a solução de (1.1), então temos de (1.35a), os limites superior e inferior do erro:

$$\frac{(1+hM_1)^{k}-1}{M_1} \min_{0\le j\le k-1} T_h[y](x_j) \le y(x_k)-y_k^* \le \frac{(1+hM_2)^{k}-1}{M_2} \max_{0\le j\le k-1} T_h[y](x_j) \quad (1.37a)$$

Se especialmente $\frac{\partial \phi}{\partial y} < 0$ para $a\le x\le b$, y arbitrário, $0\le h\le h_o$, então $M_2 < 0$ e obtemos de (1.37a):

$$\| y(x_k)-y_k^* \| \le \frac{1}{|M_2|} \max_{0\le j\le k-1} \| T_h[y](x_j) \| \quad (1.37b)$$

Concluimos que os métodos de passo simples (1.17) produzem , para h "pequeno", resultados consideravelmente melhores, quando $f_y < 0$ do que quando $f_y > 0$.

3.5 - Estabilidade e estimativa do erro para métodos de passo múltiplo.

O método de passo múltiplo (1.6)

$$= -\alpha_{p-1}y^*_{k+p-1} - \alpha_{p-2}y^*_{k+p-2} - \dots - \alpha_0 y_k^* + h\left(\beta_p f(x_{k+p},y^*_{k+p}) + \right.$$

$$\left. \dots + \beta_0 f(x_k,y_k^*)\right)$$

se reduz à forma (1.32)

$$z^*_{k+1} = Az^*_k + hBFz^*_k + hCFz^*_{k+1}$$

se fizemos:

$$z_k^* = \begin{pmatrix} y_k^* \\ y^*_{k+1} \\ \vdots \\ y^*_{k+p-1} \end{pmatrix} \quad ; \quad Fz_k^* = \begin{pmatrix} f(x_k,y_k^*) \\ f(x_k+h,y^*_{k+1}) \\ \vdots \\ f(x_k+(p-1)h,\ y^*_{k+p-1}) \end{pmatrix} \quad ; \quad y_k^* \in \mathbb{R}; \quad f:\mathbb{R}^2 \to \mathbb{R}$$

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \multicolumn{5}{c}{\text{----------}} \\ 0 & 0 & 0 & \dots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \dots & -\alpha_{p-1} \end{bmatrix} \quad ; \quad B = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \multicolumn{3}{c}{\text{----------}} \\ 0 & 0 & \dots & 0 \\ \beta_0 & \beta_2 & \dots & \beta_{p-1} \end{bmatrix} \quad ; \quad C = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \multicolumn{3}{c}{\text{----------}} \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \beta_p \end{bmatrix}$$

Se L é a constante de Lipschitz de $f(x,y)$ (vide (1.2)) tão

$$\|BF\| \le L\sum_{j=1}^{p-1}|\beta_j| = K_1; \quad \|CF\| \le L|\beta_p| = K_2 \quad ; \quad \|A\| = \sum_{j=0}^{p-1}|\alpha_j| \quad (1)$$

(1) $\|A\|$ tem esta forma porque, para um método de passo múltiplo consistente vale $\sum_{j=0}^{p-1}\alpha_j = 1$.

consequentemente

$$E = \frac{K_1 + K_2 \|A\|}{1 - hK_2} = \frac{L\sum_{j=0}^{p-1}|\beta_j| + L|\beta_p|\sum_{j=0}^{p-1}|\alpha_j|}{1 - hL|\beta_p|}$$

Sendo a matriz A de Frobenius, seus autovalores são as raizes do polinômio

$$\rho(z) = z^p + \alpha_{p-1}z^{p-1} + \ldots + \alpha_1 z + \alpha_0$$

e seus divisores elementares tem grau maior que 1 para raizes múltiplas de $\rho(z)$.

É fácil verificar que no caso em que $y_k^* \in \mathbb{R}^n$, A é uma matriz bloco de Frobenius, seus autovalores então, são as raizes de $\sigma(z) = (\rho(z))^n$. Também neste caso, A tem divisores elementares de grau maior do que 1 quando $\rho(z)$ tem raizes múltiplas.

endo $\|z(x_k) - z_k^*\| = \max_{0\le j\le p-1} \|y(x_{k+j}) - y_{k+j}^*\|$, temos

$$\|z(x_k) - z_k^*\| \ge \|y(x_{k+p-1}) - y_{k+p-1}^*\|$$ e segue do teorema 1.3, corolário seguinte:

Corolário 1.3b

Satisfeita a condição (1.2) e $h|\beta_p|L < 1$, então o método de passo múltiplo (1.6) é estável se e somente se:

1. nenhuma raiz de $\rho(z)$ tem módulo maior do que 1;

2. as raizes de módulo 1 são simples.

Podemos neste caso dar a seguinte estimativa do erro:

$$\|y(x_{k+p-1}) - y_{k+p-1}^*\| \le De^{hkDE} \max_{0\le j\le p-1} \|y(x_j) - \eta_j\| +$$

$$\frac{(e^{hkDE}-1)}{E(1-h|\beta_p|L)} \max_{0\le j\le m-p} \|T_h[y](x_j)\| \qquad (1.38)$$

$$k = 1,2,\ldots,m-p+1$$

sendo y a solução de (1.1) e L a constante de Lipschitz definida em (1.2) e D, E e T_h definido por:

$$\|A^j\| \le D; \quad E = (1-h|\beta_p|L)^{-1}L(|\beta_p|\sum_{j=1}^{p-1}|\alpha_j| + \sum_{j=0}^{p-1}|\beta_j|)$$

$$\|T_h[y](x)\| = \left\|\frac{1}{h}\sum_{i=0}^{p}\alpha_i y(x+ih) - \sum_{i=0}^{p}\beta_i f(x+ih, y(x+ih))\right\|; \quad \alpha_p = 1$$

Exemplo:

No caso do método (1.9) de Milne-Simpson temos:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \quad \alpha_0 = 1 \; ; \; \alpha_1 = 0; \; \beta_0 = \beta_2 = \frac{1}{3} \; ; \; \beta_1 = \frac{4}{3}$$

Os autovalores de A são $\lambda_1 = 1$ e $\lambda_2 = -1$, por isso o método é estável para $\frac{Lh}{3} < 1$.

Obtemos de um desenvolvimento de Taylor:

$$\left| y(x_1) - \eta_0 - hf\left(x_0 + \frac{h}{2}, \eta_0 + \frac{h}{2} f(x_0;\eta_0)\right)\right| \leq C^* h^2$$

Como $D=1$; $E = (1 - \frac{hL}{3})^{-1} 2L$; $y_0^* = \eta_0$

e $$|T_h[y](x)| \leq Ch^4 \qquad \text{(vide(1.12))}$$

chegamos ao erro:

$$|y(x_{k+1}) - y^*_{k+1}| < e^{hkE} h^2 \left(C^* + \frac{C}{2L} h^2\right) \; ; \; k = 1,2,\ldots,m-1$$

Na prática é mais indicado usar um método de passo simples de ordem 4 ao invés de 2, para aproximar $y(x_1)$ pois, neste caso, o limite do erro seria da ordem de h^4.

Muitas vezes a fórmula do erro (1.38) se mostra muito grosseira; por isso cabe a pergunta de sôbre quais hipóteses podemos, através de outras definições de estabilidade, chegar a fórmulas mais rigorosas. Estudos

sobre isso encontram-se no trabalho de Spejker [5] no qual, também foi demonstrado que existe uma fórmula de erro da forma

$$\|y(x_k) - y_k^*\| \leq M \left\{ \sum_{j=0}^{p-1} \| y(x_j) - \eta_j \| + \max_{p \leq j \leq m} \| h \sum_{k=0}^{j-p} T_h[y](x_k) \| \right\}$$

se e somente se os módulos das raízes de $\rho(z)$ são todos menores que 1.

1.3.6. - Condições de estabilidade para equações diferenciais parciais

Com $u_k = (u_1^{(k)}, u_2^{(k)}, \ldots, u_{m_1-1}^{(k)})^T$;

$f_k = (f_1^{(k)}, f_2^{(k)}, \ldots, f_{m_1-1}^{(k)})^T$; $a = \frac{\sigma \Delta t}{(\Delta x)^2}$

$$A_1 = \begin{bmatrix} (1-2a) & a & 0 & \ldots & 0 \\ a & (1-2a) & a & \ldots & 0 \\ 0 & a & (1-2a) & \ldots & 0 \\ \multicolumn{5}{c}{\text{---------------------------}} \\ 0 & 0 & \ldots & a & (1-2a) \end{bmatrix} ; \; C = \begin{bmatrix} -2 & 1 & 0 & \ldots & 0 \\ 1 & -2 & 1 & \ldots & 0 \\ 0 & 1 & -2 & \ldots & 0 \\ \multicolumn{5}{c}{\text{---------------}} \\ 0 & 0 & \ldots & 1 & -2 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} (1+2a) & -a & 0 & \ldots & 0 \\ -a & (1+2a) & -a & \ldots & 0 \\ 0 & -a & (1+2a) & \ldots & 0 \\ \multicolumn{5}{c}{\text{---------------------------}} \\ 0 & 0 & \ldots & -a & (1+2a) \end{bmatrix}$$

os métodos (1.15a/c) para solução da equação diferencial parcial (1.13) se reduzem a

a) $u^*_{k+1} = A_1 u^*_k + \Delta t\, f_k \quad ; \quad k = 0,1,\ldots,m_2-1$ (1,39a)

b) $u^*_{k+1} = 2aCu^*_k + u^*_{k-1} + 2\Delta t f_k \quad ; \quad k = 1;2,\ldots,m_2-1$

que se transforma no método de passo simples da forma (1.32):

$$v^*_{k+1} = A_2 v^*_k + 2\Delta t g_k \; ; \; v_0 = \begin{pmatrix} u^*_0 \\ u^*_1 \end{pmatrix} \qquad (1.39b)$$

se fizemos $v^*_k = \begin{pmatrix} u^*_{k-1} \\ u^*_k \end{pmatrix}$; $A_2 = \begin{pmatrix} \sigma & I \\ I & 2aC \end{pmatrix}$

$$g_k = \begin{pmatrix} \sigma \\ f_k \end{pmatrix} \in \mathbb{R}^{2(m_1-1)} \; ; \; I \in \mathbb{R}(m_1-1, m_1-1)$$

c) $A_3 u^*_{k+1} = u^*_k + \Delta t f_k$;

$$u^*_{k+1} = A_3^{-1} u^*_k + \Delta t\, A_3^{-1} f_k \quad ; \quad k = 0,1,\ldots,m_2-1 \qquad (1.39c)$$

com $u^*_0 = \left(\eta(\Delta t), \eta(2\Delta t), \ldots, \eta((m_1-1)\Delta t)\right)^T$.

No caso b, o vetor inicial u^*_1 , deve ser determinado por algum outro método como acontece sempre nos métodos de passo múltiplo.

Assim os três métodos são reduzidos à forma (1.32); sua estabilidade depende, conforme o teorema 1.3, dos autovalores das matrizes

A_1, A_2 e A_3 :

a) Podemos demonstrar que os autovalores de A_1 são:

$$\lambda_j = 1 - 4\,\mathrm{asen}^2 \frac{j\Pi}{2m_1} \quad , \; j = 1,2,\ldots,m_1-1 \; ;$$

por isso vale $\rho(A_1) < 1$ quando $a \leq \frac{1}{2}$. Logo o método explícito (1.15a) é <u>estável para</u> $a = \frac{\sigma \Delta t}{\Delta x^2} \leq \frac{1}{2}$

b) Se $\lambda \neq 0$ é autovalor de A_2 , vem de $\det(A_2 - \lambda I) = 0$, pela multiplicação por $\det \begin{pmatrix} I & \sigma \\ I & \lambda I \end{pmatrix}$,

$$\det(2a\lambda C + (1-\lambda^2) I) = 0$$

$$\det(C + \frac{1-\lambda^2}{2a\lambda} I) = 0$$

Logo $\lambda = a\mu \pm \sqrt{a^2\mu^2+1}$, onde μ é autovalor de C. Como todos os autovalores de C são diferentes de zero existem autovalores λ com $|\lambda_1| > 1$ para qualquer a. Assim o método de passo 2 (1.15b), é <u>instável para todos os Δx e Δt.</u>

Entretanto se substituirmos em (1.15b) o termo $2u_j^{*(k)}$ por $(u_j^{*(k-1)} + u_j^{*(k+1)})$, obtemos o método de Du Fort-Frankel, estável <u>para todo Δx e Δt!</u>

c) Conforme o teorema de Gerschgorin, todos os autovalores de A_3 são, em módulo, maiores que 1 para $a>0$. Logo $\rho(A_3^{-1}) < 1$ para qualquer $a>0$. Assim o método implícito (1.15c) é estável para todo Δx e Δt.

§ 1.4 - CONVERGÊNCIA DOS MÉTODOS DE PASSO SIMPLES OU MÚLTIPLO

É conveniente relembrar que a solução y^* de um método de diferenças é uma função definida apenas na rede I_h e tal que $T_h[y^*](x) = \sigma$ para todo $x_k \in I_h'$.

Definição 1.4

A solução y^* de um método de diferenças converge para a solução y de (1.1) se

$$\lim_{h\to 0} \max_{x\in I_h} \| y^*(x) - y(x) \| = 0 \qquad (1.40)$$

Diz-se que o método de diferenças converge quando a sua solução converge.

A consistência e a estabilidade de um método de discretização implicam na sua convergência, segundo o teorema:

Teorema 1.4

Um método de passo simples ou múltiplo

$$T_h[y^*](x_k) = \sigma \quad ; \quad k = 0,1,\ldots,m-p;\ p\geq 1$$

$$y_k^* = \eta_k = \eta_k(h) \quad ; \quad k = 0,1,\ldots,p-1$$

é convergente para $h\to 0$ quando é consistente e estável e se

$$\lim_{h\to 0} \max_{0\leq k\leq p-1} \| \eta_k - y(x_k) \| = 0 \qquad (1.41)$$

Demonstração:

Da condição de estabilidade obtemos, com $v_k = y_k^*$, $w_k = y(x_k)$ e $T_h[y^*](x_k) = \sigma$:

$$\| y_k^* - y(x_k) \| \leq M\{ \max_{0\leq j\leq p-1} \| \eta_j - y(x_j) \| + \max_{0\leq j\leq m-p} \| T_h[y](x_j) \| \} \qquad (1.42)$$

Então, usando (1.10) e (1.41) fica demonstrado o teorema.

Por razões didáticas formulamos o teorema acima somente para a solução do problema (1.1); entretanto a sua demonstração é independente do problema em consideração o que possibilita a sua aplicação a problemas mais gerais. Por isso podemos, por exemplo, aplicá-lo na verificação da convergência dos métodos consistentes e estáveis (1.15a/c) para solução de equações diferenciais parciais (1.3).

O corolário seguinte, vem diretamente de (1.41):

Corolário 1.4:

Se além das hipóteses do teorema 1.4, temos

$$\| \eta_j - y(x_j) \| = \mathcal{O}(h^r) \quad ; \quad j = 0,1,\ldots,p-1$$

e desde que $T_h[y^*](x_k) = \sigma$ tenha a ordem de consistência q, vale, para $k = 0,1,\ldots,m$:

$$\| y_k^* - y(x_k) \| = \mathcal{O}(h^r) + \mathcal{O}(h^q) \qquad (1.42)$$

Por isto, como já foi mencionado (pag. 8), a ordem de convergência de um método é igual a sua ordem de consistência q quando

$$\| \eta_j - y(x_j) \| = \mathcal{O}(h^q) \quad ; \quad j = 0,1,\ldots,p-1$$

Resumindo agora os resultados obtidos, temos os seguintes teoremas de convergência para métodos de passo simples ou múltiplo da forma (1.5) e (1.6):

Teorema 1.5

Seja y solução do problema de valor inicial (1.1) e

$$y_{k+1}^* = y_k^* + h\phi(x_k, y_k^* ; h) \; ; \; y_0^* = \eta_0 \; ; \; k = 1,2,..,m-1$$

um método de passo simples para a aproximação de y em I_h.

Este método converge se:

1. $\phi(x,y;h)$ for contínua em $a \le x \le b$; y qualquer; $0 \le h \le h_0$ e tal que

$$\| \phi(x,y;h) - \phi(x,\hat{y};h) \| \le K \| y - \hat{y} \| \; ;$$

2. $\phi(x,y;0) \equiv f(x,y)$

Se além disso:

3. Todas as derivadas parciais de $f(x,y)$ de ordem $q \ge 1$, forem contínuas e

4. $\phi(x,y;h) = f(x,y) + \frac{h}{2!} Df(x,y) + \ldots + \frac{h^{q-1}}{q!} D^{q-1} f(x,y) + \mathcal{O}(h^q)$

então o método tem a ordem q e vale a estimativa do erro:

$$\| y(x_k) - y_k^* \| \le K^{-1} (e^{(x_k - a)K} - 1) C h^q \quad ; \qquad (1.43)$$

$$k = 0,1,\ldots,m$$

com

$$\| y(x+h) - y(x) - h\phi(x,y(x);h) \| \le C\, h^q \quad \text{(vide (1.17))}$$

A demonstração é consequência do teorema 1.1 (consistência e ordem), corolário 1.3a (estabilidade e erro) e teorema 1.4 (convergência).

Teorema 1.6

Seja y solução de (1.1), L constante de Lipschitz de (1.2) e

$$\alpha_p y^*_{k+p} + \alpha_{p-1} y^*_{k+p-1} + \ldots + \alpha_0 y^*_k - h(\beta_p f_{k+p} + \beta_{p-1} f_{k+p-1} + \ldots + \beta_0 f_k) = \mathcal{O};$$

$$k = 0,1,\ldots,m-p$$

com $y^*_k = \eta_k$; $k = 0,1,\ldots,p-1$; $\alpha_p = 1$; $f_j = f(x_j, y^*_j)$;

$h|\beta_p|L < 1$

um método de passo múltiplo para a aproximação de $y(x_k)$, $x_k \in I_h$.

Usando os polinômios geradores $\rho(z)$ e $\sigma(z)$ definidos em (1.20), o método converge se

1. $\rho(1) = 0$; $\rho'(1) = \sigma(1)$

2. nenhuma raiz de $\rho(z)$ fôr , em módulo, maior do que 1;

3. Todas as raizes de $\rho(z)$ de módulo 1 forem simples.

Se além disso:

4. $y \in \mathcal{C}^{q+2}[a,b]$, $q \geq 1$;

5. $c_0 = c_1 = \ldots = c_q = 0$; $c_{q+1} \neq 0$ sendo c_k definido por (1.19) e $\| y(x_j) - \eta_j \| = \mathcal{O}(h^q)$, então o método tem ordem q e vale a estimativa de erro:

$$\| y(x_{k+p-1}) - y^*_{k+p-1} \| \leq De^{hkDE} \max_{0 \leq j \leq p-1} \| y(x_j) - \eta_j \|$$

$$+ \frac{(e^{hkDE}-1)h^q}{E(1-hL|\beta_p|)} \max_{x \in [a,b]} \| c_{q+1} y^{(q+1)}(x) \| \quad (1.44)$$

$$k = 1,2,\ldots,m-p+1$$

sendo D e E definidos no corolário 1.3b.

A demonstração é consequência do teorema 1.2 (consistência e ordem), corolário 1.3b (estabilidade e erro) e teorema 1.4 (convergência).

Após essas considerações teóricas, trataremos agora o problema prático de como obter métodos de passo simples ou múltiplo para a solução de (1.1).

§ 1.5 - DEDUÇÃO DE MÉTODOS DE PASSO SIMPLES

1.5.1 - Métodos com derivadas.

Sejam todas as derivadas parciais de $f(x,y)$ de ordem q , contínuas em G: $\{a \leq x \leq b$; y qualquer$\}$. Então, vem diretamente do teorema 1.5 o seguinte método de passo simples de ordem q (consistente, estável e assim convergente):

$$y^*_0 = \eta_0 \; ; \quad y^*_{j+1} = y^*_j + h\hat{\phi}(x_j, y^*_j; h); \quad j = 0,1,\ldots,m-1 \tag{1.45}$$

com $$\hat{\phi}(x_j,y_j;h) = f(x_j,y_j) + \frac{h}{2!} Df(x_j,y_j) + \ldots + \frac{h^{q-1}}{q!} D^{q-1} f(x_j,y_j)$$

Entretanto, este método com as derivadas

$$D^k f(x_j;y_j) = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right)^k f(x_j,y_j)$$

só tem valor prático quando as expressões $D^k f(x,y)$ podem ser facilmente determinadas, o que é raro. Uma vantagem deste método é que pode ter ordem arbitrária quando $f(x,y)$ for "suficientemente" diferenciável.

Exemplo

Para a equação diferencial

$$y' = -xy^2 \quad ; \quad y(1) = 2$$

obtemos de (1.45), com $q=2$, o método de ordem 2

$$y_0^* = 2 \; ; \; y_{j+1}^* = y_j^* + h y_j^{*2}(h x_j^2 y_j^* - x_j - \frac{1}{2}) \; ; \qquad (1.46)$$

$$j = 0,1,\ldots,m-1$$

Para $q=1$ obtem-se, como caso particular de (1.45), o método de Euler:

Algoritmo 1.1 (Euler)

Escolhe-se $m \in \mathbb{N}$ e calcula-se para $h = \frac{b-a}{m}$ e $j = 0,1,\ldots,m-1$.

$$x_0 = a \quad ; \quad x_{j+1} = x_j + h$$

$$y_0^* = \eta_0 \quad ; \quad y_{j+1}^* = y_j^* + h f(x_j, y_j^*)$$

(vide teorema (1.5))

1.5.2. - Métodos de Runge-Kutta

A desvantagem dos métodos da seção anterior, que consiste na determinação das expressões $D^k f$ de cálculo dificil, pode ser evitada substituindo-se $D^k f$ por combinações lineares de valores $f(x,y)$ tomados em r pontos discretos. Para isso fazemos

$$\phi_r(x,y;h) = \sum_{i=1}^{r} \mu_i k_i \qquad (1.47a)$$

com

$$\begin{aligned} k_1 &= f(x,y) \\ k_2 &= f(x + \alpha_2 h, \; y + \beta_{21} k_1 h) \\ k_3 &= f(x + \alpha_3 h, \; y + \beta_{31} k_1 h + \beta_{32} k_2 h) \qquad (1.47b) \\ &\cdots\cdots\cdots\cdots \\ k_r &= f(x + \alpha_r h, \; y + \beta_{r1} k_1 h + \beta_{r2} k_2 h + \ldots + \beta_{r,r-1} k_{r-1} h) \end{aligned}$$

e determinamos os parâmetros α_i, β_{ij} e μ_k tais que, para f suficientemente diferenciável, tenhamos:

$$\phi_r(x,y;h) - \hat{\phi}(x,y;h) = \mathcal{O}(h^q) \qquad (1.48)$$

com $\hat{\phi}(x,y;h)$ definida por (1.45).

Obtemos assim os chamados métodos de Runge-Kutta de nível r da forma

$$y_0^* = \eta_0 \; ; \; y_{j+1}^* = y_j^* + h\phi_r(x_j, y_j^*; h) \quad , \; j = 0,1,\ldots,m-1$$

Se f satisfaz a uma condição de Lipschitz, então ϕ_r também satisfará e assim, estes métodos convergem conforme o teorema 1.5 e tem ordem q (vide item 4 do teorema 1.5 e (1.48)), quando f(x,y) for q vezes continuamente diferenciável.

O sistema para a determinação dos α_j, β_{ij} e μ_k é não-linear e, em geral, não tem uma solução única. Existem na realidade, para r>1, infinitas funções $\phi_r(x,y;h)$ que satisfazem (1.48). É comum caracterizar métodos de Runge-Kutta de nível r com o seguinte esquema de parâmetros:

$$\begin{array}{c|cccc} \alpha_2 & \beta_{21} & & & \\ \alpha_3 & \beta_{31} & \beta_{32} & & \\ \alpha_4 & \beta_{41} & \beta_{42} & \beta_{43} & \\ \vdots & & & & \\ \alpha_r & \beta_{r1} & \beta_{r2} & \beta_{r3} \ \ldots\ldots\ \beta_{r,r-1} & \\ \hline & \mu_1 & \mu_2 & \mu_3 \ \ldots\ldots\ \mu_{r-1} & \mu_r \end{array} \qquad , \quad (1.49)$$

Em seguida consideramos os casos particulares r = 1,2,3 e 4.

Fórmulas de Runge-Kutta de ordens 1,2 e 3

1.- Para r=1 obtemos o método de Euler, já conhecido, que tem ordem 1 se f(x,y) tem derivadas parciais continuas de ordem 1.

2. Para r=2 e f(x,y) duas vezes continuamente diferenciável, obtemos

$$k_1 = f(x,y)$$

$$k_2 = f(x+\alpha_2 h, y+\beta_{21} k_1 h)$$

$$= f(x,y)+\alpha_2 h f_x(x,y)+\beta_{21} h f(x,y) f_y(x,y)+ \mathcal{O}(h^2)$$

Logo:

$$\phi_2 - \bar{\phi} = \mu_1 f + \mu_2 (f+\alpha_2 h f_x+\beta_{21} h f f_y) - f - \frac{h}{2} f_x - \frac{h}{2} f f_y + \mathcal{O}(h^2)$$

Esta diferença tem ordem 2 em h quando

$$\mu_1 + \mu_2 - 1 = 0; \quad \mu_2\alpha_2 - \frac{1}{2} = 0 \ ; \quad \mu_2\beta_{21} - \frac{1}{2} = 0$$

Para $\alpha_2 \neq 0$ obtem-se

$$\mu_1 = 1 - \frac{1}{2\alpha_2} \quad ; \quad \mu_2 = \frac{1}{2\alpha_2} \quad ; \quad \beta_{21} = \alpha_2$$

e assim a forma mais geral de um método de Runge-Kutta de ordem 2 e nível 2:

$$y_0^* = \eta_0; \quad y_{j+1}^* = y_j^* + h\left(1 - \frac{1}{2\alpha_2}\right) f(x_j, y_j^*) + \frac{h}{2\alpha_2} f\left(x_j+\alpha_2 h, y_j^*+\alpha_2 h f(x_j, y_j^*)\right)$$

a) para $\alpha_2 = \frac{1}{2}$ chegamos ao método de Euler modificado:

Algoritmo 1.2 (Euler modificado)

Escolhe-se $m \in \mathbb{N}$ e determina-se para $h = \frac{b-a}{m}$

e $j = 0,1,\ldots,m-1$:

$$x_0 = a \; ; \quad x_{j+1} = x_j + h$$

$$y_0^* = \eta_0 \; ; \quad y_{j+1}^* = y_j^* + hf(x_j + \frac{h}{2}, y_j^* + \frac{h}{2} f(x_j, y_j^*))$$

(vide teorema 1.5)

b) para $\alpha_2 = 1$ chegamos ao método de Heun:

Algoritmo 1.3 (Heun)

Escolhe-se $m \in \mathbb{N}$ e calcula-se para $h = \frac{b-a}{m}$ e

$j = 0,1,\ldots,m-1$:

$$x_0 = a \; ; \quad x_{j+1} = x_j + h;$$

$$y_0^* = \eta_0; \; y_{j+1}^* = y_j^* + \frac{h}{2} f(x_j, y_j^*) + \frac{h}{2} f(x_j + h, y_j^* + hf(x_j, y_j^*))$$

(vide teorema 1.5)

3. Analogamente obtemos, para $r=3$, o caso particular de um método de Runge-Kutta de ordem 3:

$$y_0^* = \eta_0 \; ; \quad y_{j+1}^* = y_j^* + h\phi_3(x_j, y_j^*; h)$$

sendo $\phi_3(x,y;h) = \frac{1}{6}(k_1 + 4k_2 + k_3)$

$$k_1 = f(x,y); \quad k_2 = f(x + \frac{h}{2}, y + \frac{h}{2} k_1);$$

$$k_3 = f(x + h, y - hk_1 + 2hk_2)$$

Na notação (1.49) este método é caracterizado pelo seguinte esquema de coeficientes:

$\frac{1}{2}$	$\frac{1}{2}$		
1	-1	2	
	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{1}{6}$

Exemplo:

Aplicando o método de Heun ao exemplo anterior:

$$y = -xy^2 \; ; \quad y(1) = 2$$

resulta em:

$$y_0^* = \eta_0 \; ; \quad y_{j+1}^* = y_j^* - \frac{h}{2} x_j y_j^* - \frac{h}{2}(x_j + h)(y_j^* - hx_j y_j^{*2})^2$$

Para $h = 0{,}1$ e $x_j = 1+0{,}1j$, $j = 0,1,\ldots,10$, obtemos os erros $e_H = y_j^* - y(x_j)$ que, na tabela abaixo, são confrontados aos erros e do algoritmo (1.46)

x_j	e_H	e
1,0	0,00 00	0,00 00
1,1	0,00 63	0,00 71
1,2	0,00 85	0,00 96
1,3	0,00 89	0,00 99
1,4	0,00 84	0,00 94
1,5	0,00 77	0,00 96
1,6	0,00 69	0,00 76
1,7	0,00 61	0,00 67
1,8	0,00 53	0,00 59
1,9	0,00 47	0,00 52
2,0	0,00 41	0,00 45

O método clássico de Runge - Kutta

Se calculamos os parâmetros α, β e μ em (1.47a/b) para r=4 tais que

$$\phi_4(x,y;h) - \hat{\phi}(x,y;h) = \mathcal{O}(h^4)$$

então obtemos, com $\beta_{21} = \alpha_2$; $\beta_{31} = \alpha_3 - \beta_{32}$; $\beta_{41} = \alpha_4 - \beta_{42} - \beta_{43}$; o sistema não linear:

$$\mu_1 + \mu_2 + \mu_3 + \mu_4 = 1$$

$$\alpha_2\mu_2 + \alpha_3\mu_3 + \alpha_4\mu_4 = \frac{1}{2}$$

$$\alpha_2^2\mu_2 + \alpha_3^2\mu_3 + \alpha_4^2\mu_4 = \frac{1}{3}$$

$$\alpha_2^3\mu_2 + \alpha_3^3\mu_3 + \alpha_4^3\mu_4 = \frac{1}{4}$$

$$\alpha_2\beta_{32}\mu_3 + (\alpha_2\beta_{42} + \alpha_3\beta_{43})\mu_4 = \frac{1}{6}$$

$$\alpha_2\alpha_3\beta_{32}\mu_3 + (\alpha_2\beta_{42} + \alpha_3\beta_{43})\mu_4\alpha_4 = \frac{1}{8}$$

$$\alpha_2^2\beta_{32}\mu_3 + (\alpha_2^2\beta_{42} + \alpha_3^2\beta_{43})\mu_4 = \frac{1}{12}$$

$$\alpha_2\beta_{32}\beta_{43}\mu_4 = \frac{1}{24}$$

Cada solução deste sistema define uma fórmula de ordem 4 de Runge-Kutta; usando a notação (1.49) chegamos aos casos especiais:

$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	0	$\frac{1}{2}$		
1	0	0	1	
	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

método clássico

$\frac{1}{3}$	$\frac{1}{3}$			
$\frac{2}{3}$	$-\frac{1}{3}$	1		
1	1	-1	1	
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

método $\frac{3}{8}$

$\frac{1}{2}$	$\frac{1}{2}$			
$\frac{1}{2}$	$(\sqrt{2}-1)/2$	$(2-\sqrt{2})/2$		
1	0	$-\sqrt{2}/2$	$(1+\sqrt{2})/2$	
	$\frac{1}{6}$	$(2-\sqrt{2})/6$	$(2+\sqrt{2})/6$	$\frac{1}{6}$

método de Gill

O método clássico de Runge-Kutta é o mais usado.

Algoritmo 1.4 (Runge-Kutta clássico de ordem 4)

Escolhe-se $m \in \mathbb{N}$ e calcula-se para $h = \frac{b-a}{m}$, e $j = 0,1,\ldots,m-1$:

$$x_0 = a \; ; \quad x_{j+1} = x_j + h$$

$$y_0^* = \eta_0 \; ; \quad k_1 = f(x_j, y_j^*)$$

$$k_2 = f(x_j + \frac{h}{2}, y_j^* + \frac{h}{2} k_1)$$

$$k_3 = f(x_j + \frac{h}{2}, y_j^* + \frac{h}{2} k_2)$$

$$k_4 = f(x_j + h, y_j^* + h\, k_3)$$

$$y_{j+1}^* = y_j^* + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

(vide teorema 1.5)

Se f não depende de y, então o algoritmo 1.4 recai na fórmula de quadratura, bem conhecida, de Simpson:

$$y_{j+1} = y_j + \int_{x_j}^{x_{j+1}} f(x)dx \approx y_j + \frac{h}{6}\left(f(x_j) + 4f(x_j + \frac{h}{2}) + f(x_j + h)\right)$$

Ilustramos o algoritmo 1.4 no caso do sistema de equações diferenciais

$$(y')^{(1)} = f^{(1)}(x, y^{(1)}, y^{(2)}, y^{(3)})$$

$$(y')^{(2)} = f^{(2)}(x, y^{(1)}, y^{(2)}, y^{(3)}) \; ; \; y(a) = \eta_0 = (\eta_0^{(1)}, \eta_0^{(2)}, \eta_0^{(3)})$$

$$(y')^{(3)} = f^{(3)}(x, y^{(1)}, y^{(2)}, y^{(3)})$$

Neste caso k_1, k_2, k_3 e $k_4 \in \mathbb{R}^3$ e os 4 níveis para a determinação de y_{j+1}^* consistem no cálculo de:

$k_1 = f(x, y^{(1)}, y^{(2)}, y^{(3)})$ para $x = x_j$; $y^{(1)} = y_j^{(1)}$; $y^{(2)} = y_j^{(2)}$; $y^{(3)} = y_j^{(3)}$

$k_2 = f(x, y^{(1)}, y^{(2)}, y^{(3)})$ para $x = x_j + \frac{h}{2}$; $y^{(1)} = y_j^{(1)} + \frac{h}{2} k^{(1)}$;

$$y^{(2)} = y_j^{(2)} + \frac{h}{2} k_1^{(2)}; \; y^{(3)} = y_j^{(3)} + \frac{h}{2} k_1^{(3)}$$

$k_3 = f(x, y^{(1)}, y^{(2)}, y^{(3)})$ para $x = x_j + \frac{h}{2}$; $y^{(1)} = y_j^{(1)} + \frac{h}{2} k_2^{(1)}$;

$$y^{(2)} = y_j^{(2)} + \frac{h}{2} k_2^{(2)}; \quad y^{(3)} = y_j^{(3)} + \frac{h}{2} k_2^{(3)}$$

$k_4 = f(x, y^{(1)}, y^{(2)}, y^{(3)})$ para $x = x_j + h$; $y^{(1)} = y_j^{(1)} + h k_3^{(1)}$;

$$y^{(2)} = y_j^{(2)} + h k_3^{(2)}; \quad y^{(3)} = y_j^{(3)} + h k_3^{(3)}$$

Então o novo vetor y_{j+1}^* resulta de:

$$y_{j+1}^* = y_j^* + \frac{h}{6} s \quad ; \quad s = (k_1 + 2k_2 + 2k_3 + k_4) \in \mathbb{R}^3$$

Este cálculo está resumido no seguinte esquema:

ESQUEMA

x	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$f^{(1)}(x, y^{(1)}, y^{(2)}, y^{(3)})$	$f^{(2)}(x, y^{(1)}, y^{(2)}, y^{(3)})$	$f^{(3)}(x, y^{(1)}, y^{(2)}, y^{(3)})$
x_j	$y_j^{(1)}$	$y_j^{(2)}$	$y_j^{(3)}$	$k_1^{(1)}$	$k_1^{(2)}$	$k_1^{(3)}$
$x_j + \frac{h}{2}$	$y_j^{(1)} + \frac{h}{2} k_1^{(1)}$	$y_j^{(2)} + \frac{h}{2} k^{(2)}$	$y_j^{(3)} + \frac{h}{2} k_1^{(3)}$	$k_2^{(1)} \cdot 2$	$k_2^{(2)} \cdot 2$	$k_2^{(3)} \cdot 2$
$x_j + \frac{h}{2}$	$y_j^{(1)} + \frac{h}{2} k_2^{(1)}$	$y_j^{(2)} + \frac{h}{2} k_2^{(2)}$	$y_j^{(3)} + \frac{h}{2} k_2^{(3)}$	$k_3^{(1)} \cdot 2$	$k_3^{(2)} \cdot 2$	$k_3^{(3)} \cdot 2$
$x_j + h$	$y_j^{(1)} + h\, k_3^{(1)}$	$y_j^{(2)} + h k_3^{(2)}$	$y_j^{(3)} + h k_3^{(3)}$	$k_4^{(1)}$	$k_4^{(2)}$	$k_4^{(3)}$
				$s^{(1)}$	$s^{(2)}$	$s^{(3)}$
x_{j+1}	$y_{j+1}^{(1)} = y_j^{(1)} + \frac{h}{6} s^{(1)}$	$y_{j+1}^{(2)} = y_j^{(2)} + \frac{h}{6} s^{(2)}$	$y_{j+1}^{(3)} = y_j^{(3)} + \frac{h}{6} s^{(3)}$			

Fórmulas de Runge-Kutta de ordem mais elevada.

Nos casos até agora considerados sempre foi possível construir fórmulas de ordem q necessitando de, apenas, r=q níveis, isto é, de q cálculos de f(x,y) para cada passo. Infelizmente isso não vale mais para q>4 . Butcher mostrou em [6] que não existe um método de Runge-Kutta de ordem 5 com 5 níveis. Um método de ordem 5 com 6 níveis foi derivado por Nyström:

$\frac{1}{3}$	$\frac{1}{3}$					
$\frac{2}{5}$	$\frac{4}{25}$	$\frac{6}{25}$				
1	$\frac{1}{4}$	$-\frac{12}{4}$	$\frac{15}{4}$			
$\frac{2}{3}$	$\frac{6}{81}$	$\frac{90}{81}$	$-\frac{50}{81}$	$\frac{8}{81}$		
$\frac{4}{5}$	$\frac{6}{75}$	$\frac{36}{75}$	$\frac{10}{75}$	$\frac{8}{75}$	0	
	$\frac{23}{192}$	0	$\frac{125}{192}$	0	$-\frac{81}{192}$	$\frac{125}{192}$

Métodos de ordem 6 com 7 e 8 níveis são indicados por Butcher em |6| e Huťa em |4|.

Como foi demonstrado também por Butcher em |7| não se pode excluir a existência de métodos com r níveis (r>4) que funcionam para uma equação diferencial de ordem 1 mas falham para sistemas de equações diferenciais! A demonstração que os métodos de Huťa valem para sistemas encontra-se em |7|.

1.5.3. - Métodos combinados.

Os métodos das seções 1.5.1 e 1.5.2 podem ser combinados, o que é vantajoso, quando $Df = \left(\frac{\partial}{\partial x} + f\frac{\partial}{\partial y}\right) f$ é facilmente calculável, mas $D^k f$, $k>1$, não.

Obtemos, por exemplo, uma fórmula de ordem 3, fazendo:

$$\phi(x,y;h) = f(x,y) + \frac{h}{2} Df(x + \alpha h,\ y + \beta h f(x,y))$$

e determinando α e β tais que

$$f(x,y) + \frac{h}{2!} Df(x,y) + \frac{h^2}{3!} D^2 f(x,y) - \phi(x,y;h) = \mathcal{O}(h^3)$$

Com um desenvolvimento de Taylor chegamos a $\alpha = \beta = \frac{1}{3}$ e assim ao algoritmo de ordem 3:

Algoritmo 1.5 (combinado de ordem 3)

Escolhe-se $m \in \mathbb{N}$ e calcula-se para $h = \frac{b-a}{m}$ e $j = 0,1,\ldots,m-1$.

$$x_0 = a \quad ; \quad x_{j+1} = x_j + h$$

$$y_0^* = \eta_0 \; ; \; y_{j+1}^* = y_j^* + hf(x_j, y_j^*) + \frac{h^2}{2} Df(x_j + \frac{h}{3}, \; y_j^* + \frac{h}{3} f(x_j, y_j^*))$$

(vide teorema 1.5)

Considerações de convergência e erro para todos os métodos apresentados neste parágrafo resultam do teorema 1.5.

§ 1.6 - OBTENÇÃO DE MÉTODOS DE PASSO MÚLTIPLO

Devido ao teorema 1.2, os coeficientes α_j e β_j ($j=0,1,\ldots,p$; $\alpha_p = 1$) de um método de passo múltiplo consistente de ordem q da forma (1.6), podem ser determinados a partir do sistema linear

$$c_i = 0; \quad i = 0,1,\ldots,q \tag{1.51}$$

sendo c_i dados por (1.19).

Assim, teoricamente é possível obter métodos implícitos ($\beta_p \neq 0$) com ordem $q = 2p$ e explícitos ($\beta_p = 0$) com ordem $q = 2p-1$, pois es-

tão disponiveis respectivamente (2p+1) e 2p coeficientes para satisfazer as (q+1) equações (1.51). O seguinte teorema (da Dahlquist |9|) mostra entretanto que, para p fixo, todos os métodos, a partir de uma certa ordem são instáveis:

Teorema 1.7

Todos os métodos explícitos de passo p, ($\beta_p = 0$), da forma (1.6) com ordem $q>p$ são instáveis. Todos os métodos implícitos de passo p, ($\beta_p \neq 0$), com ordem $q>p+2$ (para p par) e $q>p+1$ (para p impar) são instáveis.

Exemplos:

1. - O método explícito de passo 3

$$y_{k+1}^* = \frac{1}{3}(y_k^* + y_{k-1}^* + y_{k-2}^*) + \frac{h}{6}(13f_k - 4f_{k-1} + 3f_{k-2})$$

tem ordem 3 e é estável.

2. - O método implícito de passo 2 de Milne-Simpson

$$y_{k+1}^* = y_{k-1}^* + \frac{h}{3}(f_{k+1} + 4f_k + f_{k-1})$$

tem ordem 4 e é estável (como foi mostrado no exemplo 2 da seção 1.2).

Fica claro pelo teorema 1.7, que aumentar a ordem da aproximação do operador diferencial T pelo operador de diferenças T_h não resulta,necessariamente, em métodos mais precisos. Entretanto, fazendo p "suficientemente grande" , podemos construir métodos estáveis com ordem arbitrária, simplesmente determinando os coeficientes α_j em (1.6) tais que as raizes de $\rho(z)$ satisfaçam às condições 1 e 2 do corolário 1.3b e subsequentemente calculando os β_j de (1.51).

Uma alternativa mais simples para a obtenção de métodos de passo múltiplo estáveis com ordem qualquer, consiste na redução do problema $y' = f(x,y)$ ao problema de quadratura

$$y(x_{k+1}) = y(x_{k-r}) + \int_{x_{k-r}}^{x_{k+1}} f(x,y(x))dx \quad ; \; r \in \mathbb{N} \cup \{0\} \qquad (1.52)$$

e na avaliação desta integral, para $k = p-1,p,\ldots,m-1$, com métodos de Newton-Cotes (abertas ou fechadas), usando $p \geq r$ aproximações iniciais y_k^* , $k = 0,1,\ldots,p-1$, determinados anteriormente por métodos de passo simples.

Variando r, temos,assim, vários métodos de passo múltiplo entre os quais, são clássicos e muito usados os métodos explícitos de Adams - Bashforth e Nystrøm e os métodos implícitos de Adams - Moulton e Milne - Simpson.

1.6.1 - Os métodos de Adams - Bashforth e Nystrøm.

Substituindo em (1.52) a função $f(x,y(x))$ por seu polinômio de interpolação nos pontos $(x_j, f(x_j, y_j^*))$, $j = k-p+1, k-p+2,\ldots,k$:

$$P(u) = \sum_{j=0}^{p-1} (-1)^j \binom{-u}{j} \nabla^j f_k \quad ; \quad u = \frac{x - x_k}{h}$$

(vide |14|,(7.27))

e integrando de x_{k-r} até x_{k+1} , obtemos de (1.52) a seguinte classe de métodos explícitos de passo múltiplo

$$y_{k+1}^* = y_{k-r} + h \sum_{j=0}^{p-1} b_j(r) \nabla^j f_k \quad \text{com} \quad b_j(r) = (-1)^j \int_{-1}^{r} \binom{u}{j} du \qquad (1.53)$$

A convergência destes procedimentos, para $p \in \mathbb{N}$, $r = 0$ ou $r \in \mathbb{N}$ quaisquer, é consequência do teorema 1.6. Para $y \in \mathbb{C}^{p+2}[a,b]$ estes métodos tem ordem p, como pode ser verificado com um estudo mais detalhado.

Casos Especiais:

1. - Para $r = 0$ o método (1.53) se reduz ao método de passo p de Adams - Bashforth:

$$y_{k+1}^* = y_k^* + h(f_k + \frac{1}{2}\nabla f_k + \frac{5}{12}\nabla^2 f_k + \frac{3}{8}\nabla^3 f_k + \frac{251}{720}\nabla^4 f_k + \ldots + b_{p-1}(0)\nabla^{p-1} f_k)$$

Algoritmo 1.6 (Adams - Bashforth)

Escolhe-se $m \in \mathbb{N}$ e um natural p igual à ordem desejada. Calculam-se para $h = \frac{b-a}{m}$ e $k = 0,1,\ldots,p-2$ as aproximações y^*_{k+1} com métodos de passo simples de ordem p e os valores:

$$x_0 = a \quad ; \quad f_0 = f(x_0,y^*_0) = \nabla^0 f_0$$

$$x_{k+1} = x_k + h \; ; \; f_{k+1} = f(x_{k+1},y^*_{k+1}) = \nabla^0 f_{k+1}$$

$$\nabla^j f_{k+1} = \nabla^{j-1} f_{k+1} - \nabla^{j-1} f_k \; ; \; j = 1,2,\ldots,k+1$$

A seguir para $k = p-1, p,\ldots,m-1$:

$$x_{k+1} = x_k + h$$

$$y^*_{k+1} = y^*_k + h \sum_{j=0}^{p-1} b_j(0) \nabla^j f_k \qquad (1.54)$$

$$\nabla^j f_{k+1} = \nabla^{j-1} f_{k+1} - \nabla^{j-1} f_k \quad ; \; j = 1,2,\ldots,p-1$$

sendo $b_j(0)$, $j = 0,1,2,\ldots$, dados por (1.53),

ou seja: $1; \frac{1}{2}; \frac{5}{12}; \frac{3}{8}; \frac{251}{720}; \frac{95}{288}; \ldots$

2. - Para $r = 1$, $p \geq 2$ obtemos de (1.53) os métodos de passo p de Nystrøm:

$$y^*_{k+1} = y^*_{k-1} + h\left(2f_k + \frac{1}{3}\nabla^2 f_k + \frac{1}{3}\nabla^3 f_k + \frac{29}{90}\nabla^4 f_k + \ldots + b_{p-1}(1)\nabla^{p-1} f_k\right) \qquad (1.55a)$$

e para $r = p = 1$ a regra (1.7) do ponto médio:

$$y^*_{k+1} = y^*_{k-1} + 2hf(x_k,y^*_k) \; ; \; k = 1,2,\ldots,m-1 \qquad (1.55b)$$

O algoritmo de Nystrøm é igual ao de Adams - Bashforth se substituimos (1.54) por

$$y^*_{k+1} = y^*_{k-1} + h \sum_{j=0}^{p-1} b_j(1) \nabla^j f_k$$

sendo $b_j(1)$, $j = 0,1,2,\ldots$, dados por (1.53),

ou seja: $2; 0; \frac{1}{3}; \frac{1}{3}; \frac{29}{90}; \frac{14}{45}; \frac{1139}{3780}; \ldots\ldots$

§ 1.6.2 - MÉTODOS PREDITOR - CORRETORES CLÁSSICOS

Substituindo em (1.52) a função $f(x,y(x))$ por seu polinômio de interpolação nos pontos $(x_j,f(x_j,y^*_j))$, $j = k-p+1,k-p+2,\ldots,k+1$:

$$P(u) = \sum_{j=0}^{p} (-1)^j \binom{-u}{j} \nabla^j f_{k+1} \; ; \quad u = \frac{x-x_{k+1}}{h}$$

(vide|14| (7.22b))

e integrando de x_{k-r} até x_{k+1} obtemos de (1.52) a seguinte classe de fórmulas <u>implícitas</u> de passo múltiplo:

$$y^*_{k+1} = y^*_{k-r} + h \sum_{j=0}^{p} a_j(r+1) \nabla^j f_{k+1}; \quad a_j(r) = (-1)^j \int_0^r \binom{u}{j} du \qquad (1.56a)$$

Sendo g_k a soma de todas as parcelas que não dependem de y^*_{k+1} e considerando-se a relação

$$\sum_{j=0}^{p} a_j(r+1) = b_p(r)$$

obtemos de (1.56a)

$$y^*_{k+1} = hb_p(r)\, f(x_{k+1}, y^*_{k+1}) + g_k \qquad (1.56b)$$

Esta equação em y^*_{k+1}, $k = p-1, p, \ldots, m-1$, é resolvida por iteração:

$$y^{*(i+1)}_{k+1} = hb_p(r)\, f(x_{k+1}, y^{*(i)}_{k+1}) + g_k \; ; \; i = 0,1,\ldots,I.$$

onde $y^{*(0)}_{k+1}$ é determinado com uma fórmula explícita, por exemplo (1.53). Esta fórmula chama-se <u>Preditor</u>.

Com $f^{(i)}_{k+1} = f(x_{k+1}, y^{*(i)}_{k+1})$ obtemos

$$y^{*(i+1)}_{k+1} = y^{*(i)}_{k+1} + h\, b_p(r) \left(f^{(i)}_{k+1} - f^{(i-1)}_{k+1} \right). \qquad (1.56c)$$

Então, a iteração acima converge para $|hb_p(r)L| < 1$ sendo L a constante de Lipschitz de $f(x,y)$, vide (1.2). Esta condição já surgiu na demonstração do corolário 1.3b[1] e sempre pode ser satisfeita fazendo-se h suficientemente pequeno.

Resumimos a seguir, os passos de tal <u>método preditor-corretor</u>:

1. Como no caso dos métodos explícitos da seção anterior, os y^*_k são determinados para $k = 1,2,\ldots,p-1$ com métodos de passo simples;

2. A seguir calculamos, para cada $k \geq p-1$, um valor inicial $y^{*(0)}_{k+1}$ usando um <u>preditor</u> da forma (1.53);

3. O primeiro passo de iteração é feito com o <u>corretor</u> (1.56a), onde as diferenças ∇^j são calculadas com $y^{*(0)}_{k+1}$;

4. Os demais passos de iteração são feitos com (1.56c).

Como o valor inicial $y^{*(0)}_{k+1}$ já é "razoavelmente exato", dois passos de iteração são suficientes para a determinação de y^*_{k+1} (vide |10|).

---

1) (1.56b) mostra que $\beta_p = b_p(r)$

Casos especiais:

1. - Para $r = 0$, obtemos de (1.56a) os métodos de passo p de Adams - Moulton:

$$y^*_{k+1} = y^*_k + h(f_{k+1} - \frac{1}{2}\nabla f_{k+1} - \frac{1}{12}\nabla^2 f_{k+1} - \frac{1}{24}\nabla^3 f_{k+1} - \frac{19}{720}\nabla^4 f_{k+1} + \ldots$$

$$\ldots + a_p(1)\nabla^p f_{k+1})$$

Usando como preditor o método de Adams-Bashforth obtemos o seguinte algoritmo:

Algoritmo 1.7 (Adams - Moulton)

Escolhe-se $p=q-1$, sendo q a ordem desejada do método e $m \in \mathbb{N}$ tal que $\left|\frac{b-a}{m} Lb_p(0)\right| < 1$. Para $k = 0,1,\ldots,p-2$ calculam-se, com $h = \frac{b-a}{m}$, as aproximações y^*_{k+1} com um método de passo simples de ordem p e em seguida os valores:

$$x_0 = a \quad ; \quad f_0 = f(x_0, y^*_0) = \nabla^0 f_0$$

$$x_{k+1} = x_k + h \; ; \; f_{k+1} = f(x_{k+1}, y^*_{k+1}) = \nabla^0 f_{k+1}$$

$$\nabla^j f_{k+1} = \nabla^{j-1} f_{k+1} - \nabla^{j-1} f_k \; ; \; j = 1,2,\ldots,k+1$$

A seguir, para $k = p-1, p, \ldots, m-1$:

$$x_{k+1} = x_k + h; \; y^{*(0)}_{k+1} = y^*_k + h \sum_{j=0}^{p-1} b_k(0)\nabla^j f_k \qquad \text{(Preditor)}$$

$$f^{(0)}_{k+1} = f(x_{k+1}, y^{*(0)}_{k+1}) = \nabla^0 f^{(0)}_{k+1}$$

$$\nabla^j f^{(0)}_{k+1} = \nabla^{j-1} f^{(0)}_{k+1} - \nabla^{j-1} f^{(0)}_k \quad ; \quad j = 1,2,\ldots,p$$

$$y^{*(1)}_{k+1} = y^*_k + h \sum_{j=0}^{p} a_j(1)\nabla^j f^{(0)}_{k+1} \qquad \text{(1. passo de iteração)}$$

$$f^{(1)}_{k+1} = f(x_{k+1}, y^{*(1)}_{k+1})$$

$$y^{*(2)}_{k+1} = y^{*(1)}_{k+1} + b_p(0)\, h(f^{(1)}_{k+1} - f^{(0)}_{k+1}) \qquad \text{(2. passo de iteração)}$$

$$y^*_{k+1} = y^{*(2)}_{k+1}; \; \nabla^0 f_{k+1} = f(x_{k+1}, y^*_{k+1})$$

$$\nabla^j f_{k+1} = \nabla^{j-1} f_{k+1} - \nabla^{j-1} f_k \; ; \; j = 1,2,\ldots,p$$

Os $a_j(1)$, $j = 0,1,2,\ldots$, são definidos em (1.56a) e dados por: $1; -\frac{1}{2}; -\frac{1}{12}; -\frac{1}{24}; -\frac{19}{720}; \ldots$

2. - Para $r = 1$ obtemos de (1.56a) os métodos de passo p de Milne - Simpson, $p \geq 2$:

$$y^*_{k+1} = y^*_{k-1} + h(2f_{k+1} - 2\nabla f_{k+1} + \frac{1}{3}\nabla^2 f_{k+1} - \frac{1}{90}\nabla^4 f_{k+1} - \frac{1}{90}\nabla^5 f_{k+1} + \ldots$$

$$\ldots + a_p(2)\nabla^p f_{k+1})$$

e para $r=p=1$, novamente a regra do ponto médio (1.7)

O algoritmo de Milne-Simpson se transforma no de Adams-Moulton substituindo-se $b_k(0)$ por $b_k(1)$, $b_p(0)$ por $b_p(1)$ e o 1º passo de iteração por

$$y^{*(1)}_{k+1} = y^*_{k-1} + h \sum_{j=0}^{p} a_j(2)\, \nabla^j f^{(a)}_{k+1}$$

A convergência dos métodos de Adams-Moulton e Milne-Simpson é consequência do teorema 1.6. Para $y(x)$ suficientemente diferenciável eles têm ordem $p+1$, com exceção do método de passo 2 de Milne-Simpson que tem ordem 4, como já sabemos (vide § 1.2). Estas afirmações só valem no caso (teórico) em que os y^*_{k+1} são determinados exatamente, isto é, com um número infinito de passos de iteração. Quando usamos apenas um ou um número finito de passos de iteração, então a convergência e a ordem de um método dependem, naturalmente, do preditor, como será mostrado na seção seguinte.

§ 1.7. - ESTABILIDADE E ORDEM DE MÉTODOS PREDITOR-CORRETORES

Seja[1]
$$z^*_{k+1} = A_1 z^*_k + h\, B_1 F\, z^*_k \qquad (1.57a)$$

um preditor de ordem q_1 e

$$z^*_{k+1} = A_2 z^*_k + h\, B_2 F\, z^*_k + h\, C\, F\, z^*_{k+1} \qquad (1.57b)$$

um corretor de ordem q_2 com $|\beta_p| Lh < 1$. Se executamos apenas uma iteração com o corretor, o método associado pode ser representado por

$$z^*_{k+1} = A_2 z^*_k + h\, B_2 F\, z^*_k + h\, C\, F\, \hat{z}_{k+1} \qquad (1.58)$$

com
$$\hat{z}_{k+1} = A_1 z^*_k + h\, B_1 F\, z^*_k$$

Como (1.58) tem a forma (1.32), obtemos imediatamente, do teorema 1.3:

Teorema 1.8

Um método preditor-corretor é estável se e sòmente se o corretor é estável.

Concluimos que o preditor pode, até, ser instável. Isso poderia nos levar a usar preditores de ordem mais alta possível, isto é, de

(1) Usaremos, a seguir, a notação introduzida em 1.3.5

ordem 2p-1 que, pelo teorema 1.7, são instáveis. Entretanto, as considerações abaixo nos mostram que assim não obteríamos uma melhoria na precisão, pois a ordem (e portanto, em geral, a precisão) de um método preditor-corretor depende somente do corretor.

Se denominamos de $T_h^{(1)}[z](x_k)$ e $T_h^{(2)}[z](x_k)$ os erros de substituição de (1.57a) e (1.57b), temos:

$$z(x_{k+1}) = A_1 z(x_k) + hB_1 Fz(x_k) + h\, T_h^{(1)}[z](x_k)$$

$$z(x_{k+1}) = A_2 z(x_k) + hB_2 Fz(x_k) + hCFz(x_{k+1}) + hT_h^{(2)}[z](x_k) \qquad (1.59)$$

com $$\max_{x\in I_h'} \| T_h^{(i)}[z](x)\| \le D_i h^{q_i} \; ; \quad i = 1,2. \qquad (1.60)$$

Assim, o erro de substituição $T_h^{(3)}[z](x_k)$ do método (1.58) é dado por

$$z(x_{k+1}) = A_2 z(x_k) + hB_2 Fz(x_k) + hCF\hat{z}(x_{k+1}) + hT_h^{(3)}[z](x_k) \qquad (1.61a)$$

com $$\hat{z}(x_{k+1}) = z(x_{k+1}) - hT_h^{(1)}[z](x_k) \qquad (1.61b)$$

Subtraindo (1.61a) de (1.59), vem:

$$T_h^{(3)}[z](x_k) = T_h^{(2)}[z](x_k) + C\left(Fz(x_{k+1}) - F\hat{z}(x_{k+1})\right) \qquad (1.62)$$

Com $\| C(Fz-F\hat{z})\| \le |\beta_p| L \, \| z-\hat{z}\|$ obtemos de (1.61b) e (1.62):

$$\| T_h^{(3)}[z](x_k)\| \le \| T_h^{(2)}[z](x_k)\| + |\beta_p| Lh \| T_h^{(1)}[z](x_k)\|$$

e, tomando em consideração (1.60), obtemos:

$$\max_{x\in I_h'} \| T_h^{(3)}[z](x)\| \le D_2 h^{q_2} + D_1^* h^{q_1+1} \; ; \; D_1^* = |\beta_p| L \cdot D_1$$

Assim fica demonstrado o seguinte teorema:

Teorema 1.9

Seja q_1 a ordem do preditor e q_2 a do corretor, onde $q_1 \ge q_2 - 1$. Então o método preditor-corretor associado, com apenas um passo de iteração do corretor, tem ordem q_2.

Observamos, então, que a ordem de um método preditor-corretor não se altera, resolvendo-se exatamente (1.57b) (o que acarretaria um número infinito de iterações) ou executando um único passo de iteração.

Recomendamos 2 passos de iteração com o corretor (conforme |10|), com a vantagem adicional que o incremento h pode ser controlado pelo valor:

$$\Delta_{k+1} = \frac{1}{|\beta_p|} \frac{\left\| y_{k+1}^{*(2)} - y_{k+1}^{*(1)} \right\|}{\left\| y_{k+1}^{*(1)} - y_{k+1}^{*(0)} \right\|} = h \left\| \frac{\partial f}{\partial y} (x_{k+1}, \eta_{k+1}) \right\| ;$$

$$\eta \in \left[y_{k+1}^{*(1)}, y_{k+1}^{*(0)} \right]$$

pois, de (1.56c) segue, levando-se em consideração que $b_p(r) = \beta_p$:

$$y_{k+1}^{*(2)} - y_{k+1}^{*(1)} = h\beta_p \left(f(x_{k+1}, y_{k+1}^{*(1)}) - f(x_{k+1}, y_{k+1}^{*(0)}) \right)$$

$$= h\beta_p \frac{\partial f}{\partial y} (x_{k+1}, \eta_{k+1}) (y_{k+1}^{*(1)} - y_{k+1}^{*(0)})$$

Como regra prática recomendamos escolher h tal que, para todo k, $\Delta_k \simeq 10^{-1}$. O incremento h deve ser mudado se essa quantidade Δ_k não obtiver o valor aproximado desejado . Isso pode ser feito retornando-se à subrotina que calcula, através de métodos de passo simples, novos valores para reiniciar o processo no ponto interrompido com um novo incremento h.

Finalmente surge o problema de como escolher, para determinado corretor, um preditor "otimo". Isto tem sido pouco estudado e ainda não existe uma resposta definitiva. Seria possível chegar a resultados aplicando-se a fórmula de erro (1.38) à (1.58), obtendo-se assim , um limite do erro para o método preditor-corretor completo (ao invés de,

como no § 1.3.5, limites distintos para o preditor e o corretor), e determinando-se os α_k e β_k do preditor, tais que este limite seja mínimo. Bons resultados práticos foram conseguidos com preditores de passo p tomando-se

$$\alpha_k = -\frac{1}{p} \quad , \; k = 0,1,\ldots,p-1 \; ; \; \alpha_p = 1 \quad \text{(vide } |11|\text{)}.$$

§ 1.8 - COMPARAÇÃO DOS MÉTODOS

No caso de tipos particulares de equações diferenciais é aconselhável o uso de métodos especiais ao invés dos anteriormente apresentados. Nos últimos anos vários destes métodos especiais foram desenvolvidos e toda a teoria do tratamento numérico de equações diferenciais parciais mostra uma tendência crescente nesta direção.

A eficiência de um método depende fundamentalmente do problema em consideração; por isso observações gerais sobre "melhores métodos" não são possíveis e temos que nos limitar à comparação dos processos usando critérios bastante gerais. A seguir comparamos métodos de passo simples, múltiplo e de extrapolação (apresentados na parte II), baseando-nos nos seguintes critérios:

1. O método é auto-iniciável ?

2. É possível variar o incremento h sem "grandes esforços computacionais"?

3. A ordem do método pode ser variada?

4. Quantas avaliações de f(x,y) são necessárias por passo de computação?

5. O método pode ser usado sem a ajuda de computadores?

1.8.1 - Métodos de passo simples:

Vantagens:

1. São auto-iniciáveis.

2. A variação de h é facilmente executada; assim h pode ser ajustado às exigências do problema facilitando um compromisso entre a precisão desejada e um mínimo de esfôrço computacional.

Desvantagens:

1. A ordem é baixa e invariável, por isso h deve ser "pequeno" para limitar o erro de discretização; como consequência disto, o erro de arredondamento e desnecessariamente "grande".

2. O número de avaliações de f(x,y) por passo é "grande" em comparação com outros métodos (temos, por exemplo, 4 avaliações no método clássico de Runge-Kutta).

3. Os métodos são complicados em relação aos de passo múltiplo.

1.8.2 - Métodos de passo múltiplo

Vantagens:

1. São necessárias apenas (I+1) avaliações de f(x,y) por passo de computação, com I=0 no caso dos métodos explícitos e I=1 ou I=2 no dos implícitos.

2. O esforço computacional é quase independente da ordem do método. Isto torna fácil o uso de método de ordem elevada.

3. Os métodos são muito simples.

Desvantagens:

1. Não são auto-iniciáveis.

2. Uma variação de h é "relativamente" complicada.

3. A ordem, uma vez escolhida, é fixa.

1.8.3 - Métodos de extrapolação (a ser apresentados na seção 2.4 da parte II).

Vantagens:

1. São auto-iniciáveis.

2. A variação de h e o controle contínuo da precisão são facilmente executados.

3. A ordem é variável e pode ser adaptada ao problema.

4. O número de avaliações de f(x,y) é "aproximadamente" igual ao dos métodos de passo múltiplo, às vêzes até menor.

Desvantagens:

1. Os métodos são complicados e não indicados para cálculo manual.

Uma boa biblioteca de programas deveria conter todos os três algoritmos: para funções f relativamente simples (menos de 25 operações algébricas por componente) um método de extrapolação com ordem entre 6 e 12; para funções complicadas, um método preditor-corretor com ordem variável (até ordem 12) e, para funções simples e precisão restrita a mais ou menos 10^{-4}, o método de Runge-Kutta de ordem 4.

Uma comparação rigorosa dos métodos numéricos para equações diferenciais encontra-se em |15|.

OBSERVAÇÕES BIBLIOGRÁFICAS

1. Os primeiros resultados básicos sobre estabilidade e convergência de métodos de passo múltiplo foram publicados por Dahlquist em 1956, |9|. Nesta publicação, entre outras coisas, encontra-se a seguinte definição de um operador estável que tornou-se classica, sendo usada na maioria dos livros sobre o assunto: Um operador de diferenças da forma (1.6) é estável se todas as raízes de $\rho(z)$ (vide (1.20)) encontram-se sobre ou dentro do círculo unitário do plano z e se as raízes sobre o círculo são simples.

   Esta definição nos pareceu insatisfatória por duas razões: (i) não permite uma generalização aos demais problemas de discretização, por exemplo para equação diferenciais parciais; (ii) sendo tão especializada fica dificil perceber a sua relação com os demais conceitos de estabilidade, tendo assim mais o aspecto de um teorema do que de uma definição mesma.

   Por isso preferimos usar a definição 1.3 da estabilidade, que é muito mais geral, e deduzimos daí a definição de Dahlquist como teorema (vide teorema 1.6).

2. A definição 1.3 foi tirada do livro |13| de Keller (e pode ser encontrada também em outras publicações recentes). Neste livro são

apresentados também exemplos para considerações de estabilidade para problemas de contorno. Estudos gerais sobre a "melhor" definição de estabilidade em relação à estimativa mais vantajosa do erro encontram-se numa publicação muito interessante de Spijker |5|.

3. O critério geral de estabilidade (teorema 1.3) é novo; está baseado em trabalhos próprios do autor, e permite a obtenção de vários novos resultados entre os quais destacamos o teorema 1.8 (também ainda não publicado[1]). O aspecto talvez mais importante do teorema 1.3 é a possibilidade de considerar métodos preditor-corretores como métodos integrados ao invés de, como é feito em quase toda a literatura, tratar cada fórmula separadamente. Assim obtemos a possibilidade de estudar a influência do preditor no comportamento do método completo (o que é feito, por exemplo, nos teoremas 1.8 e 1.9), resolvendo, assim, um problema, até então em aberto[2], da Análise Numérica.

---

1) Encontramos, entretanto, no livro de Henrici |1|, pág. 283/284 a observação de que existem métodos preditor-corretores convergentes com preditores instáveis. Mas, esta afirmação está baseada apenas em experiências numéricas.

2) Vide Gear |12| pag. 142

4. Estarão disponíveis no Departamento de Informática da PUC-RJ, a partir do fim do ano corrente, programas modernos e eficazes, em FORTRAN, para resolução numérica de sistemas de equações diferenciais ordinárias: (i) por Runge-Kutta de ordem 4; (ii) preditor-corretores, (iii) método de extrapolação ao limite. Estes programas admitem um controle automático do erro, mudança, também automática, do incremento, além de outras vantagens.

BIBLIOGRAFIA DA PARTE I

1. Henrici, P. : Discrete Variable Methods in Ordinary Differential Equations, Wiley 1965.

2. Henrici, P. : Error Propagation for Difference Methods, Wiley 1963

3. Collatz, L. : The Numerical Treatment of Differential Equations, Springer 1966

4. Ceschino F. ; Kuntzmann, J. : Numerical Solution of Initial Value Problems, Prentice Hall, 1966

5. Spijker, M. N. : On the Structure of Error Estimates for Finite - Difference Methods, Numer. Math.18, pg.73-100,1971

6. Butcher, J. C.: On Runge-Kutta-Processes of Higher Order, J. Austral. Math. Soc. 4, 1964.

7. Butcher, J. C.: On the Integration Processes of A. Huťa, J.Austral. Math. Soc. 3, 1963

8. Nyström, E.J. : Über die num. Integration von Differentialglei = chungen, Acta Soc. Sci. Fennicae 50, No. 13, 1925.

9. Dahlquist, G. : Convergence and Stability in the Numerical Integration of Ordinary Differential Equations, Math. Scand. 4, 1956.

10. Hull, T. E., ; Creemer, A.L.: Efficiency of Predictor-Corrector Procedures, J ACM 10, 1963

11. Chaves, T. : Preditores: Experiências para a obtenção de me - lhores formulas, Tese de Mestrado, PUC-RJ, 1971.

12. Gear, C. W. : Numerical Initial Value Problems in Ordinary Differential Equations, Prentice Hall, 1971.

13. Keller, H.B. : Numerical Methods for Two-Point Boundery - Value Problems, Blaisdell, 1968.

14. Albrecht, P. : Um Curso de Análise Numérica, Ao Livro Técnico , R.J. 1973.

15. Hull, T.E. e outros : Comparing Numerical Methods for Ordinary Differential Equations, SIAM J. Numer. Anal., Vol. 9, No. 4, 1972.

PARTE II

EXTRAPOLAÇÃO AO LIMITE

INTRODUÇÃO

Na parte anterior foi apresentada uma teoria geral de discretização e abordada a maneira de se obter, no caso de equações diferenciais ordinárias, os algoritmos de discretização.

Nesta segunda parte do curso estudaremos o problema de como a partir de algoritmos de determinada ordem, obter outros que convergem com ordem mais elevada.

Não nos limitaremos apenas a problemas de equações diferenciais e consideraremos também métodos para quadratura e diferenciação numéricas.

Observamos que estes métodos podem ser considerados como casos particulares dos métodos gerais da Parte I, de modo que se enquadram na teoria lá apresentada.

Por exemplo, o problema de quadratura

$$y = \int_a^b f(u)du \qquad (2.1)$$

pode ser escrita na forma[1)]

$$T[y] = 0 \qquad \text{com} \qquad T[y] = y - \int_a^b f(u)du \ .$$

e reduzido, discretizando a integral, ao problema

$$T_h[y^*] = 0 \qquad \text{com} \qquad T[y^*] \equiv y^* - \frac{h}{2}\left[f(a)+f(b)+2\sum_{j=1}^{N-1} f(a+jh)\right] \ ;$$

$$h = \frac{b-a}{N}$$

cuja solução é a constante (dependente de h):

$$y^* = \frac{h}{2}\left[f(a)+f(b) + \sum_{j=1}^{N-1} f(a+jh)\right] \qquad (2.2)$$

Esta é a chamada regra dos trapézios.

Do mesmo modo podemos representar o problema da determinação da derivada de uma função $f: \mathbb{R} \to \mathbb{R}$ no ponto x, na forma

$$T[y](x) \equiv y - f'(x) = 0$$

e resolvê-lo pela discretização

$$T_h[y^*](x) \equiv y^*(x) - \frac{f(x+h)-f(x-h)}{2h} = 0$$

---

1) Neste caso a função $y : \mathbb{R} \to \mathbb{R}$ é constante.

cuja solução é $y^*(x) = \frac{f(x+h) - f(x-h)}{2h}$ (2.3)

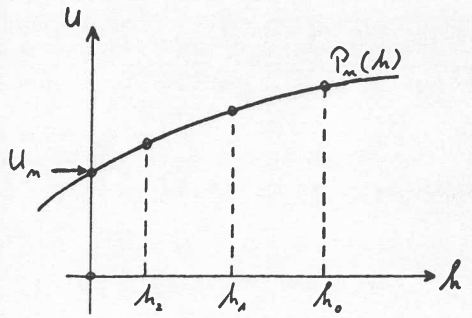
Consideraremos, nas discussões seguintes $y^*(x)$, para x fixo, função do incremento h. Para acentuar este fato e simplificar a notação, introduziremos uma nova variável U tal que:

$$U(h) \equiv y^*(x,h) \quad ; \quad x \text{ fixo} \tag{2.4}$$

Então, U(h) é uma aproximação da solução y(x) de $T[y](x)=0$ para x fixo, obtida por algum método de discretização com o incremento h.

A aceleração da convergência por extrapolação ao limite se baseia na seguinte idéia:
U(h) será calculado com um algoritmo de discretização não só para um determinado valor h (como na Parte I) mas sim para uma sequência de incrementos h_k, $k = 0,1,\ldots,n$.

Em geral podemos esperar que, à medida que h_k diminui, as aproximações $U(h_k)$ melhoram. Obviamente U(h) não pode ser calculado no limite h=0, entretanto é evidente a idéia de que o valor em h=0, do polinômio de interpolação de U(h) pode fornecer uma nova aproximação U_n mais exata. Esta extrapolação ao limite (ou extrapolação de Richardson) é explicado pela seguinte figura:



$P_n(h)$: polinômio de interpolação de U(h) ; $U_n = P_n(0)$.

Será demonstrado, mais tarde, que U_n converge para y(x), x fixo, quando $n \to \infty$, com ordem maior do que a sequência $U(h_k)$, $k \to \infty$, se existe num determinado intervalo $[0,h_0]$ um desenvolvimento da forma

$$U(h) = y(x) + \tau_1 h + \tau_2 h^2 + \ldots + \tau_n h^n + h^{n+1} K_{n+1}(h) \tag{2.5}$$

sendo τ_j independentes de h e $|K_{n+1}(h)| \leq M_{n+1} = \text{const.}$

Exemplos:

1. - Seja a derivada f'(x) aproximada por (2.3), isto é

$$U(h) = \frac{f(x+h) - f(x-h)}{2h}$$

Então, se $f \in \mathbb{C}^{2n+1}$ numa determinada vizinhança de x, obtemos, pelo desenvolvimento de Taylor:

$$U(h) = f'(x) + \frac{f'''(x)}{3!} h^2 + \ldots + \frac{f^{(2n-1)}(x)}{(2n-1)!} h^{2n-2} + h^{2n} K_{n+1}(h) \tag{2.6}$$

com $K_{n+1}(h) = \frac{1}{(2n+1)!} f^{(2n+1)}(\xi)$; $x-h < \xi < x+h$

A diferenciação numérica por extrapolação ao limite será apresentada no parágrafo 2.3.

2. - Seja $y = \int_a^b f(u)du$

aproximada pela regra de trapézios (2.2), isto é:

$$U(h) = \frac{h}{2}\left[f(a)+f(b) + 2\sum_{j=1}^{N-1} f(a+jh)\right] \quad ; \quad h = \frac{b-a}{N}$$

Então, se $f \in \mathbb{C}^{2n+2}[a,b]$, vale a fórmula de Euler-Maclaurin (vide e. g. [2], pg. 290 ou [3], pg. 127):

$$U(h) = y + \tau_1 h^2 + \tau_2 h^4 + \ldots + \tau_n h^{2n} + h^{2n+2} K_{n+1}(h) \qquad (2.7)$$

onde $y = \int_a^b f(u)du$; $\tau_j = \frac{B_{2j}}{(2j)!}\left\{f^{(2j-1)}(a) - f^{(2j-1)}(b)\right\}$; $j=1,2,\ldots,n$

$$K_{n+1}(h) = (-1)^n \int_a^b f^{(2n+2)}(u)\left\{\sum_{j=1}^{\infty} \frac{2-2\cos(2\pi j(u-a)h^{-1})}{(2\pi j)^{2n+2}}\right\} du \qquad (2.8)$$

B_{2j}: numeros de Bernoulli.

A aplicação da extrapolação ao limite à quadratura numérica resulta no método de Romberg que será abordado no parágrafo 2.1.

3. - Seja $y^*(x_k)$ a aproximação da solução $y(x)$ de

$$y' = f(x,y) \quad ; \quad y(a) = \eta_0$$

obtida pela regra do ponto médio (1.7):

$$y_1^* = \eta_0 + hf(x_0,\eta_0) \quad ; \quad y_{k+1}^* = y_{k-1}^* + 2hf(x_k,y_k^*) \quad ; \quad k = 1,2,\ldots$$

Se $f(x,y)$ tem derivadas parciais de ordem $2n+2$ então pode-se demonstrar (vide Gragg |4;5|) a existência de um desenvolvimento da forma:

$$U(h) = y(x) + \tau_1(x)h^2 + \tau_2(x)h^4 + \ldots + \tau_n(x)h^{2n} + K_{n+1}(h,x)h^{2n+2} \qquad (2.9)$$

onde $U(h) = y^*(x)$ e os $\tau_j(x)$, $j = 1,\ldots,n$, não dependem de h e $|K_{n+1}(h,x)| < M$; $x \in [a,b]$; $h \in [0,h_0]$ sendo M e h_0 constantes.

A integração numérica das equações diferenciais por extrapolação ao limite será apresentada no parágrafo 2.4.

Como a extrapolação ao limite será generalizada em 2.2.1. a algoritmos para os quais existe um desenvolvimento da forma

$$U(h) = y + \tau_1 h^{\gamma_1} + \tau_2 h^{\gamma_2} + \ldots + \tau_n h^{\gamma_n} + K_{n+1}(h) h^{\gamma_{n+1}} \qquad (2.10)$$

com $0 < \gamma_1 < \gamma_2 < \ldots < \gamma_{n+1}$, apresentamos como último exemplo um algoritmo que tem tal desenvolvimento:

4. - Integrais impróprias da forma $y = \int_0^a \frac{g(x)}{\sqrt{x}}\, dx$, $g(x)$ analítica, podem ser avaliadas pela seguinte modificação da regra dos trapézios:

$$U(h) = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{N-1} f(jh)\right] \; ; \quad h = \frac{a}{N}$$

Pode-se provar que (vide Rutishauser |6|):

$$U(h) = y + \tau_1 h^{\frac{1}{2}} + \tau_2 h^{\frac{3}{2}} + \tau_3 h^{2} + \tau_4 h^{\frac{5}{2}} + \tau_5 h^{\frac{7}{2}} + \ldots \qquad (2.11)$$

§ 2.1. - QUADRATURA NUMÉRICA POR EXTRAPOLAÇÃO AO LIMITE

2.1.1. - O método de Romberg

Seja $$U(h_k) = \frac{h_k}{2}\{f(a)+f(b) + 2\sum_{j=1}^{N_k-1} f(a+jh_k)\} \qquad (2.12)$$

com $$N_k = \frac{b-a}{h_k} \in \mathbb{N} \; ; \quad h_{k+1} < h_k \; ; \quad k = 0,1,2,\ldots$$

uma sequência de aproximações da integral $y = \int_a^b f(u)du$ obtida pela regra dos trapézios (2.2). Se f é Riemann - integravel e $\lim_{k\to\infty} h_k = 0$, então $\lim_{k\to\infty} U(h_k) = y$. Se $f \in \mathbb{C}^{[a,b]}$, a convergência de $U(h)$ é de ordem 2 em h como se vê na fórmula (2.7) de Maclaurin.

Interpolando nos pontos $(h_k, U(h_k))$ por um polinômio em h^2 obtemos:

$$P_{2n}(h) = \sum_{k=0}^{n} U(h_k) \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{h^2 - h_j^2}{h_k^2 - h_j^2} \qquad (2.13)$$

Definimos U_n por $U_n = P_{2n}(0)$ e o consideraremos como nova aproximação de $y = \int_a^b f(u)du$, isto é,

$$U_n = \sum_{k=0}^{n} c_k(n)\, U(h_k) \qquad (2.14)$$

com $$c_k(n) = \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{h_j^2}{h_j^2 - h_k^2} \qquad (2.15)$$

Veremos na seção 2.1.3. (Corolário 2.2) que se $f \in \mathbb{C}^{2n+2}[a,b]$ o algoritmo assim obtido é de ordem mais elevada do que o algoritmo inicial (ou seja, a regra dos trapézios).

Na prática não será necessario determinar o polinômio $P_{2n}(h)$ explícitamente, sendo U_n facilmente calculado com o algoritmo de Neville (vide e.g., |17|, sec. 6.5.2). Chegamos, assim, ao algoritmo de Romberg:

Algoritmo 2.1 (Romberg)

Escolhe-se $\varepsilon > 0$, um natural N (em geral $N \leq 8$) e uma sequência $h_k = \frac{b-a}{N_k}$, N_k naturais, $N_{k+1} > N_k$, $k=0,1,2,\ldots$

Calcula-se $U_0^{(0)} = U(h_0)$ e determina-se para $k=0,1,2,\ldots$:

$$U_0^{(k+1)} = U(h_{k+1})$$

$$U_{d+1}^{(k-d)} = U_d^{(k-d+1)} + \frac{U_d^{(k-d+1)} - U_d^{(k-d)}}{\left(\frac{h_{k-d}}{h_{k+1}}\right)^2 - 1} \quad ; \; d=0,1,2,\ldots,m \qquad (2.16)$$

$m=\min(k,n-1,N)$

m: vide (2.9)

(vide algoritmo de Neville)

Termina-se o processo, se $\left|U_m^{(k-m)} - U_m^{(k-m+1)}\right| < \varepsilon$ e usa-se $U_{m+1}^{(k-m)}$ como aproximação de $y = \int_a^b f(u)\,du$

(vide teorema 2.2.)

Esquema:

h_0	$U(h_0)$	$U_1^{(0)}$	$U_2^{(0)}$	. . . . .	$U_{n-1}^{(0)}$	$U_n^{(0)}$
h_1	$U(h_1)$	$U_1^{(1)}$	$U_2^{(1)}$	. . . . .	$U_{n-1}^{(1)}$	
h_2	$U(h_2)$	$U_1^{(2)}$	$U_2^{(2)}$	. . . . .		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
h_{n-1}	$U(h_{n-1})$	$U_1^{(n-1)}$				
h_n	$U(h_n)$					

Se $n > N+1$, então o natural N limita o número de colunas no esquema dos $U_k^{(j)}$ com o objetivo de se limitar a acumulação do êrro.

A sequência $\{h_k\}$ pode ser escolhida de várias maneiras. Em geral, usa-se uma das seguintes:

a) $h_k = \frac{h_0}{k+1}$

b) $h_k = h_0 p^{-k}$; $p>1$, inteiro

(2.17)

c) $h_k = h_0 p_k$ com $p_0 = 1$; $p_k = \begin{cases} \frac{2}{3}\, 2^{-\frac{k}{2}} & ; \; k: \text{par} \\ 2^{-\frac{(k+1)}{2}} & ; \; k: \text{impar} \end{cases}$

isto é: $p_k = 1; \frac{1}{2}; \frac{1}{3}; \frac{1}{4}; \frac{1}{6}; \frac{1}{8}; \frac{1}{12}; \ldots$

onde $h_0 = \frac{b-a}{N_0}$, $N_0 \in \mathbb{N}$; $k = 0,1,2,\ldots$

A escolha b) com $p = 2$ nos dá o algoritmo original de Romberg |7|. A seguir, entretanto, chamaremos o algoritmo 2.1 de método de Romberg, qualquer que seja a escolha da sequência $\{h_k\}$.

Exemplo:

$$\int_0^{0,64} \frac{\operatorname{sen} x}{x}\, dx \approx 0,625\ 614$$

h_k	$U(h_k)$	$U_1^{(k)}$	$U_2^{(k)}$
0,64	0,618 60	0,625 62	0,625 61
0,32	0,623 87	0,625 61	
0,16	0,625 18		
0,08	0,625 50		
0,04	0,625 59		

Vemos que se considerarmos 3 avaliações do integrando (isto é, $h_0 = 0,64$ e $h_1 = 0,32$) já obtemos a aproximação $U_1^{(0)} = 0,625\ 62$ com quatro algarismos significativos. Usando-se a regra dos trapézios só obteremos esta precisão para $h = 0,04$, ou seja após 17 avaliações do integrando.

2.1.2 - Casos particulares do método de Romberg

1. Aplicamos o algoritmo de Romberg para $n=1$ com $h_0 = b-a$; $h_1 = \frac{b-a}{N_1}$ e calculamos $U_1^{(0)}$ nos casos $N_1 = 2,3,4$. Obtemos:

a) $h_0 = b - a$; $h_1 = \frac{b-a}{2}$

$$U_1^{(0)} = U(h_1) + \frac{1}{3}(U(h_1) - U(h_0))$$

$$= \frac{4}{3} U(h_1) - \frac{1}{3} U(h_0)$$

$$= \frac{4}{3} \cdot \frac{h_1}{2} (f(a) + 2f(a+h_1) + f(b)) - \frac{1}{3}\frac{h_0}{2}(f(a) + f(b))$$

$$= \frac{h_1}{3} (f(a) + 4f(a + h_1) + f(b)) \quad \text{já que } h_0 = 2h_1$$

Chegamos, assim, à fórmula

$$\int_a^b f(u)du \approx \frac{h}{3}(f(a)+4f(a+h)+f(b)) \quad ; \quad h = \frac{b-a}{2} \qquad (2.18)$$

que é a regra de Simpson com 3 pontos.

b) $h_0 = b - a$; $h_1 = \frac{b-a}{3}$

$$U_1^{(0)} = U(h_1) + \frac{1}{8}(U(h_1) - U(h_0))$$

como $h_0 = 3h_1$, obtemos

$$U_1^{(0)} = \frac{3}{8} h_1 (f(a) + 3f(a + h_1) + 3f(a + 2h_1) + f(b))$$

Isto é: $$\int_a^b f(u)du \approx \frac{3}{8} h\left(f(a)+3f(a+h)+3f(a+2h)+f(b)\right) \qquad (2.19)$$

$$h = \frac{b-a}{3}$$

que é a "pulcherrima" ou regra 3/8 de Newton.

c) $h_0 = b - a$; $h_1 = \frac{b-a}{4}$

De maneira análoga, obtemos:

$$U_1^{(0)} = \frac{2}{15} h_1(3f_0 + 8f_1 + 8f_2 + 8f_3 + 3f_4) \qquad (2.20)$$

As fórmulas (2.18) e (2.19) são chamadas fórmulas fechadas de Newton-Cotes; elas podem ser também obtidas (vide |17| cap. VII) por interpolação de $f(u)$ e posterior integração do polinômio de interpolação. Entretanto, (2.20) não é uma destas fórmulas, o que mostra que o método de Romberg fornece, também, novos algoritmos. A grande vantagem das fórmulas de Romberg de ordem elevada é que convergem qualquer que seja f, ao contrário das de Newton-Cotes (vide teorema 2.3).

2. - Agora fazemos $h_k = (b - a)2^{-k}$, $k = 0,1,2,\ldots$ e computamos $U_1^{(k)}$:

$$U_1^{(k)} = U(h_{k+1}) + \frac{1}{3}\left(U(h_{k+1}) - U(h_k)\right)$$

$$= \frac{1}{3}\left\{4.\frac{1}{2}h_{k+1}\left[f(a)+f(b)+2\sum_{j=1}^{2m-1} f(a+jh_{k+1})\right]\right.$$

$$\left. - \frac{h_k}{2}\left[f(a) + f(b) + 2\sum_{j=1}^{m-1} f(a+jh_k)\right]\right\}, \quad m = 2^k$$

Já que $h_k = 2h_{k+1}$, obtemos com $h_{k+1} = h$

$$U_1^{(k)} = \frac{h}{3}\{f(a) + f(b) + 4\sum_{j=1}^{2m-1} f(a+jh) - 2\sum_{j=1}^{m-1} f(a+j2h)\}$$

$$= \frac{h}{3}\{f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \ldots$$

$$\ldots + 2f(a + (2m-2)h) + 4f(a+ (2m-1)h + f(b)\} \qquad (2.21)$$

Essa é a fórmula de Simpson com (2m+1) pontos. Assim está demonstrado o teorema seguinte:

Teorema 2.1

Se $h_k = (b - a)2^{-k}$ então a segunda coluna $U_1^{(k)}$ do esquema de Romberg é idêntica aos valores obtidos pela regra (2.21) de Simpson.

Chamamos a atenção para o fato de que a representação de $U_d^{(k)}$ pelos valores f_j foi vista apenas para demonstrar que o método de Romberg contém algumas fórmulas de Newton-Cotes como casos particulares. Na práti-

ca essa representação nunca é feita!

2.1.3 - Considerações sôbre o êrro e a convergência

Sabemos de 2.1.1 que a primeira coluna do esquema de Romberg tende para y. Estudaremos agora a convergência das demais colunas do esquema e determinaremos sob que hipóteses $U_n^{(k)}$ tende para y mais rãpidamente que $U(h_k)$ quando $k \to \infty$.

Observamos que, se $U(h)$ fôr um polinômio de grau $\leq n$ em h^2, então, a interpolação (2.13) é exata, isto é $P_{2n}(h) \equiv U(h)$ (pois, a interpolação em n pontos de um polinômio de grau $\leq n$ resulta, como se sabe, no próprio polinômio). Por outro lado, os coeficientes $c_k(n)$ em (2.14) são independentes da função $U(h)$; por isso podemos usar qualquer função $U(h)$ para chegar a resultados sôbre eles.

Fazendo assim:

a) $U(h) \equiv 1$;

b) $U(h) \equiv h^{2j}$, $j = 1,2,\ldots,n$;

c) $U(h) \equiv h^{2(n+1)} - (h^2 - h_0^2)(h^2 - h_1^2)\ldots(h^2 - h_n^2)$

obtemos de $P_{2n}(0) = U(0)$ as seguintes relações para os $c_k(n)$:

$$\text{a)} \quad \sum_{k=0}^{n} c_k(n) = 1 \quad ;$$

$$\text{b)} \quad \sum_{k=0}^{n} c_k(n) h_k^{2j} = 0 \quad ; \quad j = 1,2,\ldots,n \qquad (2.22)$$

$$\text{c)} \quad \sum_{k=0}^{n} c_k(n) h_k^{2n+2} = (-1)^n h_0^2 . h_1^2 \ldots h_n^2$$

Podemos agora demonstrar:

Teorema 2.2

Seja $f \in C^{2n+2}[a,b]$ e seja h_k, $k = 0,1,\ldots$, uma das seqüências (8.17). Então vale para os valores $U_n^{(j)}$ do algoritmo 2.1:

$$U_n^{(j)} = \int_a^b f(u)du + R_n^{(j)}$$

$$\text{com} \quad R_n^{(j)} = (b-a) h_j^2 . h_{j+1}^2 \ldots h_{j+n}^2 \frac{2S_{2n+2}}{(2\pi)^{2n+2}} f^{(2n+2)}(\eta) \qquad (2.23a)$$

$$\text{onde} \quad a < \eta < b \quad \text{e} \quad S_m = \sum_{k=1}^{\infty} k^{-m}$$

Como (vide |2|, pg. 251)

$$S_{2m} = (-1)^m \frac{(2\pi)^{2m}}{2(2m)!} B_{2m} \quad , \quad B_{2m}: \text{números de Bernoulli}$$

obtemos uma outra representação de $R_n^{(j)}$:

$$R_n^{(j)} = (b - a) \prod_{k=j}^{j+n} h_k^2 \frac{|B_{2n+2}|}{(2n+2)!} f^{(2n+2)}(\eta) \qquad (2.23b)$$

Como $S_{2n+2} \simeq 1$ para $n>1$, preferimos (2.23a).

Demonstração:

A fórmula de Euler - Maclaurin (2.7) juntamente com (2.14) nos dá:

$$U_n^{(0)} = \sum_{k=0}^{n} c_k(n)(y + \tau_1 h_k^2 + \ldots + \tau_n h_k^{2n} + h_k^{2n+2} K_{n+1}(h_k))$$

onde K_{n+1} é dado por (2.8).

De (2.22) obtemos com $y = \int_a^b f(u)du$:

$$U_n^{(0)} - \int_a^b f(u)du = \sum_{k=0}^{n} c_k(n) h_k^{2n+2} K_{n+1}(h_k) \qquad (2.24)$$

$$U_n^{(0)} - \int_a^b f(u)du = (-1)^n \int_a^b f^{(2n+2)}(u)G(u)du$$

com $$G(u) = \sum_{k=0}^{n} h_k^{2n+2} c_k(n) \sum_{j=1}^{\infty} \frac{2 - \cos 2\pi j(u - a)h_k^{-1}}{(2\pi j)^{2n+2}}$$

Pode-se demonstrar (vide |9|), que $G(u)$ não muda de sinal em $[a,b]$, se $\{h_k\}$ é uma das sequências (2.17). Por isso:

$$U_n^{(0)} - \int_a^b f(u)du = (-1)^n f^{(2n+2)}(\eta) \int_a^b G(u)du \quad ; \quad a<\eta<b \ .$$

Como $\frac{b - a}{h_k}$ são números inteiros, as integrais dos têrmos em coseno de $G(u)$ se anulam e, assim, obtemos:

$$\int_a^b G(u)du = \sum_{k=0}^{n} h_k^{2n+2} c_k(n) \sum_{j=1}^{\infty} \frac{2(b - a)}{(2\pi j)^{2n+2}}$$

Usando (2.22b/c) vem:

$$\int_a^b G(u)du = (-1)^n h_0^2 . h_1^2 \ldots h_n^2 \cdot \frac{2(b - a)}{(2\pi)^{2n+2}} S_{2n+2}$$

Com a mudança dos índices obtemos (2.23a)

c.q.d.

Definindo a ordem do algoritmo 2.1 como a ordem em h_j de $R_n^{(j)}$, segue:

Corolário 2.2

Os elementos $U_n^{(j)}$ da (n+1)-ésima coluna do esquema de Romberg convergem para $\int_a^b f(u)du$ com ordem (2n+2). Então a adição de outra coluna no esquema aumenta, em 2, a ordem do algoritmo 2.1.

Demonstração:

A convergência resulta de que $\lim_{j\to\infty} h_j = 0$.

Como $h_{k+1}^2 = q_k h_k^2$; $q_k < 1$, temos: $h_{k+1}^2 = h_0^2 \prod_{j=0}^{k} q_j$.

Então $\prod_{k=0}^{n} h_k^2 = C h_0^{2n+2}$, com $C = q_0^{n} . q_1^{n-1} \ldots q_{n-1} < 1$

c.q.d.

Entretanto, a ordem de uma fórmula é apenas uma estimativa grosseira de sua eficiência. Por isso pode acontecer que $U_0^{(n)}$ forneça uma melhor aproximação de y do que $U_n^{(0)}$, dependendo das derivadas de f(u).

No caso da sequência $h_k = \frac{h_0}{k+1}$, k = 0,1,2,..., o algoritmo de Romberg pode divergir, isto é, $\lim_{n\to\infty} R_n^{(j)} \neq 0$ para j fixo. No exemplo abaixo ilustramos êste caso.

Exemplo:

$$\int_{-1}^{1} \frac{x^2\,dx}{1+25x^2} = \frac{2}{125}(5 - \text{arc tg } 5) = 0{,}058\ 026$$

h_k	$U(h_k)$	$U_1^{(j)}$	$U_2^{(j)}$	$U_3^{(j)}$	$U_4^{(j)}$	$U_5^{(j)}$	$U_6^{(j)}$
$2/_{10}$	0,057 95	0,058 76	0,054 94	0,067 31	0,035 28	0,106 50	-0,035 61
$2/_{11}$	0,058 08	0,057 60	0,059 99	0,051 62	0,074 85	0,019 90	
$2/_{12}$	0,058 01	0,058 28	0,056 79	0,062 36	0,045 87		
$2/_{13}$	0,058 05	0,057 88	0,058 79	0,055 15			
$2/_{14}$	0,058 03	0,058 11	0,057 55				
$2/_{15}$	0,058 04	0,057 98					
$2/_{16}$	0,058 03						

Temos, entretanto, o seguinte teorema:

Teorema 2.3

Para $h_k = h_0 p^{-k}$, $p>1$, $k = 0,1,\ldots$ o algoritmo 2.1 sempre converge quando $n \to \infty$, ou seja:

$$\lim_{n\to\infty} U_n^{(0)} = y$$

qualquer que seja f.

O teorema vale também no caso mais geral quando

$$\lim_{k\to\infty} h_k = 0 \quad e \quad \frac{h_{k+1}}{h_k} \leq p < 1$$

e não implica, necessariamente, que $U_n^{(0)}$ converge mais rápido para y do que $U(h_k)$, pois êle não exige a existência do desenvolvimento (2.7)!

A demonstração é consequência do:

Teorema de Toeplitz:

Se a) $\lim_{k\to\infty} U(h_k) = y$

b) $\lim_{n\to\infty} c_k(n) = 0$, para cada k

c) $\sum_{k=0}^{n} c_k(n) = 1$, para todo n

d) $\sum_{k=0}^{n} |c_k(n)| \leq C$, para todo n; C: const.

então

$$\lim_{n\to\infty} \sum_{k=0}^{n} c_k(n).U(h_k) = y \text{ , isto é, } \lim_{n\to\infty} U_n^{(0)} = y$$

a) é satisfeita com $y = \int_a^b f(u)du$ já que $f(u)$ é integrável

b) é consequência de (2.15).

c) foi provado em (2.22a)

d) Com $h_k = h_0 p^{-k}$ obtemos de (2.22b):

$$\sum_{k=0}^{n} c_k(n)(p^{-2j})^k = 0 \; ; \quad j = 1,2,\ldots,n.$$

Então o polinômio:

$$Q_n(x) = \sum_{k=0}^{n} c_k(n)x^k \quad \text{tem os zeros} \quad x_j = p^{-2j} \text{ ,}$$

$j = 1,2,\ldots,n$, e de (2.22a) vemos, que $Q_n(1)=1$. Obtemos, então:

$$Q_n(x) = \sum_{k=0}^{n} c_k(n)x^k = \prod_{j=1}^{n} \frac{x - p^{-2j}}{1 - p^{-2j}} \quad ;$$

Por isso os $c_k(n)$, $k = 0,1,\ldots,n$ têm sinais alternados e assim:

$$(-1)^n Q_n(-1) = \sum_{k=0}^{n} |c_k(n)| = \prod_{j=1}^{n} \frac{1+p^{-2j}}{1-p^{-2j}} < \prod_{j=1}^{\infty} \frac{1+p^{-2j}}{1-p^{-2j}} =$$

$$= \prod_{j=1}^{\infty} \left(1 + \frac{2p^{-2j}}{1 - p^{-2j}}\right)$$

O produto converge se $p>1$, pois $\sum_{j=1}^{\infty} \frac{p^{-2j}}{1 - p^{-2j}}$ converge. Então existe uma constante C tal que

$$\sum_{k=0}^{n} |c_k(n)| < C$$

c.q.d.

§ 2.2 - DISCUSSÃO GERAL DA EXTRAPOLAÇÃO AO LIMITE

Neste parágrafo, o método usado anteriormente é generalizado de maneira que se torna aplicável também a problemas diferentes da quadratura. Além disso apresentaremos uma abordagem distinta e mais fácil, fornecendo, assim,também uma nova interpretação do método de Romberg.

2.2.1 - Tratamento geral do método

O método de Romberg foi baseado no desenvolvimento (2.7). Generalizando, assumimos agora que $U(h)$ é a solução de um algoritmo finito dependendo de $h>0$, que converge, quando $h\to 0$, para a solução y de um problema tal que

$$U(h) = y + \tau_1 h^{\gamma_1} + \tau_2 h^{\gamma_2} + \ldots + \tau_n h^{\gamma_n} + K_{n+1}(h)h^{\gamma_{n+1}} \qquad (2.25)$$

onde $0 < \gamma_1 < \ldots < \gamma_{n+1}$; $|K_{n+1}(h)| < M_{n+1}$, $h \in [0,h_0]$

τ_j , $j = 1,2,\ldots,n$, independentes de h.

Nosso objetivo será o de construir, a partir dos valores $U(h_k)$, $k = 0,1,\ldots,n$, novos algoritmos que convergem com ordem $\gamma_n > \gamma_1$.

De (2.25) obtemos o sistema:

$$U(h_k) - K_{n+1}(h_k)h_k^{\gamma_{n+1}} = y + \tau_1 h_k^{\gamma_1} + \ldots + \tau_n h_k^{\gamma_n}$$

ou seja $t - s = H\tau$; $t,s,\tau \in \mathbb{R}^{n+1}$ (2.26)

com $t_k = U(h_k)$; $s_k = K_{n+1}(h_k)h_k^{\gamma_{n+1}}$; $\tau_0 = y$

$$H = \begin{bmatrix} 1 & h_0^{\gamma_1} & h_0^{\gamma_2} & \cdots & h_0^{\gamma_n} \\ 1 & h_1^{\gamma_1} & h_1^{\gamma_2} & \cdots & h_1^{\gamma_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & h_n^{\gamma_1} & h_n^{\gamma_2} & & h_n^{\gamma_n} \end{bmatrix}$$

tão, aplicando a regra de Cramer:

$$y = \frac{D}{\det H} = \frac{1}{\det H} \sum_{k=0}^{n} (-1)^k D_k \{U(h_k) - K_{n+1}(h_k) h_k^{\gamma_{n+1}}\} \qquad (2.27)$$

ndo D_k os menores dos elementos da 1º coluna de H.

r outro lado obtemos de (2.26): $\tau = H^{-1}(t-s)$, ou seja

$$y = \sum_{k=0}^{n} c_k(n) \left(U(h_k) - K_{n+1}(h_k) h_k^{\gamma_{n+1}}\right) \qquad (2.28)$$

ndo $c_k(n)$ os elementos da 1ª linha de H^{-1}.

terpretando $U_n^{(0)} = \sum_{k=0}^{n} c_k(n) U(h_k)$ como nova aproximação de y, obte-
s de (2.28) o erro:

$$R_n^{(0)} = U_n^{(0)} - y = \sum_{k=0}^{n} c_k(n) K_{n+1}(h_k) h_k^{\gamma_{n+1}} \qquad (2.29)$$

Obviamente, (2.29) é uma generalização de (2.24). Nota-se, que $R_n^{(0)}$ é de ordem γ_{n+1} em h enquanto que o erro de U(h) somente é de ordem $\gamma_1 < \gamma_{n+1}$.

De $H^{-1}H = I$ segue a seguinte generalização das relações (2.22a/b):

$$\sum_{k=0}^{n} c_k(n) = 1 \; ; \; \sum_{k=0}^{n} c_k(n) h_k^{\gamma_j} = 0 \; ; \; j = 1,2,\ldots,n. \qquad (2.30)$$

Comparando (2.27) e (2.28) obtemos

$$c_k(n) = (-1)^k \frac{D_k}{\det H} \quad ; \quad k = 0,1,\ldots,n \qquad (2.31)$$

e, assim

$$U_n^{(0)} = \sum_{k=0}^{n} \frac{(-1)^k D_k}{\det H} U(h_k) \qquad (2.32)$$

Se $h_{k+1} < h_k$, então [1] ~~a matriz H é totalmente negativa (vide, e.g., |12|, pg. 98), isto é: $\det H < 0$, $D_k < 0$, $k = 0,1,\ldots,n$~~. Por isso, os sinais dos $c_k(n)$ são alternados e de (2.29) e (2.31) segue com $M_{n+1} = \sup_{h \in [0,h_0]} |K_{n+1}(h)|$:

$$|R_n^{(0)}| \leq M_{n+1} \sum_{k=0}^{n} |c_k(n)| h_k^{\gamma_{n+1}} = M_{n+1} \left| \sum_{k=0}^{n} \frac{D_k}{\det H} h_k^{\gamma_{n+1}} \right| ; \qquad (2.33)$$

[1] todos D_k, $k = 0,1,\ldots,n$, têm o mesmo sinal, pois por $\leq (n-1)$ trocas de linha podem ser reduzidas a matrizes totalmente positivas (vide |12|, pg. 98).

finindo

$$H_1 = \begin{bmatrix} h_0^{\gamma_{n+1}} & h_0^{\gamma_1} & h_0^{\gamma_2} & \cdots & h_0^{\gamma_n} \\ h_1^{\gamma_{n+1}} & h_1^{\gamma_1} & h_1^{\gamma_2} & \cdots & h_1^{\gamma_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (-1)^n h_n^{\gamma_{n+1}} & h_n^{\gamma_1} & h_n^{\gamma_2} & \cdots & h_n^{\gamma_n} \end{bmatrix}$$

temos finalmente

$$|R_n^{(0)}| \leq M_{n+1} \left| \frac{\det H_1}{\det H} \right|$$

stá, assim, demonstrado

eorema 2.4

Seja válido o desenvolvimento (2.25) e seja

$U_n^{(0)} = \sum_{k=0}^{n} c_k(n) U(h_k)$ uma nova aproximação de y, sendo

$c_k(n)$ dados por (2.31) e $h_{k+1} < h_k$.

Então vale a estimativa do êrro:

$$|R_n^{(0)}| \leq M_{n+1} \left| \frac{\det H_1}{\det H} \right| ; \quad M_{n+1} = \sup_{h \in [0,h_0]} |K_{n+1}(h)| \qquad (2.34)$$

Então, a extrapolação ao limite, na sua forma mais geral, pode ser caracterizada da seguinte maneira: Os resultados $U(h_k)$, obtidos por um certo algoritmo com ordem γ_1, são combinados (não necessariamente por extrapolação!) de maneira a formarem novos algoritmos com ordem mais elevada.

2.2.2 - Dois algoritmos particulares

Em seguida consideramos casos especiais que permitem a determinação rápida de $U_n^{(0)}$ e $R_n^{(0)}$. Observamos que em dois casos o det H é de Vandermonde e, assim, facilmente calculado.

Caso A:

$\gamma_k = k\gamma$; h_k : arbitrários; $h_{k+1} < h_k$. Neste caso $U_n^{(0)}$ pode ser determinado recursivamente de (2.32) pelo algoritmo seguinte:

Algoritmo A (Extrapolação ao limite)

Escolhe-se $h_0 > 0$, $\varepsilon > 0$ e um natural N (em geral $N \leq 8$).
Calcula-se $U_0^{(0)} = U(h_0)$ e determina-se para $k = 0,1,2,..$

$$h_{k+1} < h_k \quad ; \quad U_0^{(k+1)} = U(h_{k+1})$$

(2.35)

$$U_{d+1}^{(k-d)} = U_d^{(k-d+1)} + \frac{U_d^{(k-d+1)} - U_d^{(k-d)}}{\left(\frac{h_{k-d}}{h_{k+1}}\right)^{\gamma} - 1} \quad ; \; d=0,1,2,\ldots,m \qquad m=\min(k,n-1,N)$$

(n definido em (2.25))

Termina-se o processo se $|U_m^{(k-m)} - U_m^{(k-m+1)}| < \varepsilon$ e usa-se $U_{m+1}^{(k-m)}$ como aproximação de y.

(vide teorema 2.5)

O natural N limita o número de colunas no esquema dos $U_k^{(j)}$ com o objetivo de se limitar a acumulação do êrro.

Para os $U_k^{(j)}$ usa-se o mesmo esquema do algoritmo 2.1. Aplicações do algoritmo A encontram-se nos parágrafos 2.3 e 2.4.

O algoritmo A é idêntico à extrapolação de Neville e representa o caso clássico da extrapolação de Richardson. Nota-se, entretanto, que foi obtido sem a necessidade da aplicação de uma extrapolação, de uma maneira bem elementar.

O fato de que o procedimento aqui desenvolvido é independente da extrapolação fica mais óbvio no seguinte:

Caso B:

$$h_k = h_0 p^{-k} \quad ; \; p > 1$$

Calculando (2.32) sucessivamente, obtemos o seguinte algoritmo:

Algoritmo B (Extrapolação ao limite)

Escolhe-se $h_0 > 0$, $\varepsilon > 0$ e os naturais $p>1$ e N (em geral $p=2$, $N \leq 8$). Calcula-se $U_0^{(0)} = U(h_0)$ e determina-se para $k = 0,1,2,\ldots$:

$$h_{k+1} = p^{-1} h_k \quad ; \quad U_0^{(k+1)} = U(h_{k+1})$$

$$U_{d+1}^{(k-d)} = U_d^{(k-d+1)} + \frac{U_d^{(k-d+1)} - U_d^{(k-d)}}{p^{\gamma_{d+1}} - 1} \quad ; \; d=0,1,\ldots,m \qquad m=\min(k,n-1,N) \qquad (2.36)$$

onde n e γ_d são definidos por (2.25). Termina-se o processo, quando

$$|U_m^{(k-m)} - U_m^{(k-m+1)}| < \varepsilon$$

e usa-se $U_{m+1}^{(k-m)}$ como aproximação de y.

(vide teorema 2.6)

É um fato notável que o algoritmo B, apesar de obtido com as mesmas considerações que resultam no algoritmo A, não pode ser derivado de uma extrapolação.

2.2.3 - Considerações sobre o erro nos algoritmos A e B.

Baseado nos resultados da seção anterior, podemos, agora, demonstrar o seguinte teorema:

Teorema 2.5

Seja válido o desenvolvimento (2.25) com $\gamma_j = j\gamma$; $j=1,\ldots,n$ e seja $U_n^{(0)}$ obtido pelo algoritmo A. Então vale para o erro:

$$|R_n^{(0)}| = |U_n^{(0)} - y| \leq M_{n+1} N_n^{(0)} \prod_{i=0}^{n} h_i^{\gamma} \qquad (2.37)$$

sendo M_{n+1} definido em (2.25) e

$$N_n^{(0)} = \sum_{\ell=0}^{n} h_\ell^{n\gamma} \prod_{\substack{i=0 \\ i\neq\ell}}^{n} |h_i^{\gamma} - h_\ell^{\gamma}|^{-1}$$

Demonstração:

Se $\gamma_j = j\gamma$, obtemos de (2.31): $|c_k(n)| = \prod_{\substack{i=0 \\ i\neq k}}^{n} h_i^{\gamma} |h_i^{\gamma} - h_k^{\gamma}|^{-1}$ e de (2.33):

$$|R_n^{(0)}| \leq M_{n+1} \sum_{k=0}^{n} h_k^{(n+1)\gamma} \prod_{\substack{i=0 \\ i\neq k}}^{n} h_i^{\gamma} |h_i^{\gamma} - h_k^{\gamma}|^{-1}$$

c.q.d.

Obviamente obtemos limites para o erro $R_\ell^{(j)}$ de cada valor $U_\ell^{(j)}$ se substituimos em (2.37) n por ℓ e se mudamos os índices de h (isto é equivalente à escolha de h_j em vez de h_0 como parâmetro inicial). Assim chegamos a:

Corolário 2.5a

Para o erro $R_\ell^{(j)}$ de cada valor $U_\ell^{(j)}$, $\ell \leq n$, temos

$$|R_\ell^{(j)}| \leq M_{\ell+1} N_\ell^{(j)} \prod_{i=0}^{\ell} h_{j+i}^{\gamma} \qquad (2.38)$$

sendo $N_\ell^{(j)} = \sum_{m=0}^{\ell} h_{j+m}^{\ell\gamma} \prod_{\substack{i=0 \\ i\neq m}}^{\ell} |h_{j+i}^{\gamma} - h_{j+m}^{\gamma}|^{-1}$

Se $\lim_{j\to\infty} \frac{h_{j+1}}{h_j} < 1$, então $\lim_{j\to\infty} N_\ell^{(j)} < \infty$ para cada ℓ fixo. Obtemos, então:

Corolário 2.5b

Se, além de $h_{j+1} < h_j$, $\lim_{j\to\infty} \frac{h_{j+1}}{h_j} < 1$, então para $\ell \leq n$

$$\lim_{j\to\infty} U_\ell^{(j)} = y$$

No caso especial do método de Romberg, temos (vide (2.23a)):

$$M_{n+1} = \frac{2(b-a)S_{2n+2}}{(2\pi)^{2n+2}} \max_{u\in[a,b]} |f^{(2n+2)}(u)|$$

Por isso os números $N_n^{(0)}$ em (2.37) representam as razões entre os limites (2.37) e (2.23a) e facilitam, assim, uma comparação de ambos os erros. Para $\gamma = 2$ e a sequência $h_k = h_0 2^{-k}$, $k = 0,1,\ldots$, obtemos para todo n: $N_n^{(0)} < 2$. Este é um resultado muito satisfatório se levarmos em consideração o fato de que a fórmula (2.37) foi obtida com pouco esforço. Notamos, além disso, que a fórmula (2.37) vale também para a sequência $\{h_k\}$ dada por (2.17c), caso em que a validez de (2.23a) não é garantida.

No caso B obtemos de (2.33):

$$|R_n^{(0)}| \le M_{n+1} h_0^{\gamma_{n+1}} \sum_{k=0}^{n} |c_k(n)| p^{-k\gamma_{n+1}} \qquad (2.39)$$

Com isso podemos demonstrar o seguinte teorema:

Teorema 2.6

Seja válido o desenvolvimento (2.25) e seja $U_n^{(0)}$ obtido pelo algoritmo B. Então vale, para o erro:

$$|R_n^{(0)}| = |U_n^{(0)} - y| \le M_{n+1} N_n^* h_0^{\gamma_{n+1}} \qquad (2.40)$$

com $$N_n^* = \prod_{j=1}^{n} \frac{p^{-\gamma_{n+1}} + p^{-\gamma_j}}{1 - p^{-\gamma_j}}$$

Demonstração:

Substituindo h_k por $h_0 p^{-k}$ em (2.30) obtemos

$$\sum_{k=0}^{n} c_k(n) (p^{-\gamma_j})^k = 0$$

Então o polinômio $Q_n(x) = \sum_{k=0}^{n} c_k(n) x^k$ tem as n raizes positivas $p^{-\gamma_j}$, $j = 1,2,\ldots,n$. Além disso, como consequência de (2.30): $Q_n(1) = 1$. Por isso

$$Q_n(x) = \prod_{j=1}^{n} \frac{(x - p^{-\gamma_j})}{(1 - p^{-\gamma_j})}$$

Como os sinais dos $c_k(n)$ são alternados, o somatório em (2.39) é igual a

$$|Q_n(-p^{-\gamma_{n+1}})| = \prod_{j=1}^{n} \frac{p^{-\gamma_{n+1}} + p^{-\gamma_j}}{1 - p^{-\gamma_j}} = N_n^* .$$

c.q.d.

Poderiamos ter chegado a (2.40) calculando (2.34) diretamente, levando

em consideração, que, no caso B, H_1 também é de Vandermonde.

Como no corolário (2.5a) obtemos limites para o erro $R_\ell^{(j)}$ de cada valor $U_\ell^{(j)}$, $\ell \le n$, se substituimos em (2.40) n por ℓ e h_0 por h_j:

Corolário 2.6

Para o erro $R_\ell^{(j)}$ de cada valor $U_\ell^{(j)}$, $\ell \le n$, temos

$$|R_\ell^{(j)}| \le M_{\ell+1} N_\ell^* h_j^{\gamma_{\ell+1}}$$

Isto é: A sequência $U_\ell^{(j)}$, $j = 0,1,\ldots$ e ℓ fixo, $\ell \le n$ converge a y com ordem $\gamma_{\ell+1}$

Exemplo:

A integral $y = \int_0^{\frac{\pi^2}{4}} \frac{2\cos\sqrt{u}}{\sqrt{u}}\,du = 4$ não pode ser calculada pelo método de Romberg pois não valem as hipóteses da fórmula de Euler-Maclaurin. Entretanto, o desenvolvimento (2.11) nos dá $\gamma_1 = \frac{1}{2}$, $\gamma_2 = \frac{3}{2}$, $\gamma_3 = 2$, $\gamma_4 = \frac{5}{2}$, $\gamma_5 = \frac{7}{2}$, e com êsses valores aplicamos o algoritmo B com $h_k = \pi^2 . 2^{-(k+4)}$ obtendo assim:

h_k	$U(h_k)$	$U_1^{(j)}$	$U_2^{(j)}$	$U_3^{(j)}$
$\pi^2/16$	2,572 56	3,972 57	3,999 46	4,000 00
$\pi^2/32$	2,982 61	3,989 95	3,999 87	
$\pi^2/64$	3,277 65	3,996 36		
$\pi^2/128$	3,348 82			

Sabemos do corolário 2.6, que:

$$y - U_1^{(j)} = \mathcal{O}(h_j^{\frac{3}{2}});\quad y - U_2^{(j)} = \mathcal{O}(h_j^2);\quad y - U_3^{(j)} = \mathcal{O}(h_j^{\frac{5}{2}})$$

Comparamos estes resultados com os obtidos do algoritmo A com $\gamma = 0,5$:

$U_1^{(j)}$	$U_2^{(j)}$	$U_3^{(j)}$
3,972 57	4,007 33	4,000 27
3,989 95	4,002 77	
3,996 36		

O passo de $U_1^{(j)}$ para $U_2^{(j)}$ é inútil, pois serve para "remover" um têrmo em h não existente no desenvolvimento (2.25).

O algoritmo de Romberg ($\gamma = 2$) fornece valores falsos:

$U_1^{(j)}$	$U_2^{(j)}$	$U_3^{(j)}$
3,119 30	3,393 11	3,573 30
3,376 00	3,570 48	
3,558 33		

pois o têrmo em $h^{\frac{1}{2}}$ não foi "removido".

§ 2.3. - DIFERENCIAÇÃO NUMÉRICA POR EXTRAPOLAÇÃO

O uso da extrapolação ao limite nos permite obter fórmulas de qualquer ordem para a diferenciação numérica, desde que a função a ser derivada seja "suficientemente" diferenciável.

Como exemplo, seja:

$$U(h) = h^{-2}\left(f(x_0 + h) - 2f(x_0) + f(x_0 - h)\right) \qquad (2.41)$$

uma aproximação de $f''(x_0)$. O desenvolvimento de Taylor nos dá:

$$U(h) = f''(x_0) + \frac{1}{12} f^{(4)}(x_0)h^2 + \frac{1}{360} f^{(6)}(x_0)h^4 + \ldots$$

mostrando que o método (2.41) tem ordem 2. Aplicando o algoritmo A com $\gamma = 2$ e $h_0 = 2h$; $h_1 = h$ obtemos:

$$U_1^{(0)} = U_0^{(1)} + \frac{1}{3}\left(U_0^{(1)} - U_0^{(0)}\right)$$

$$= \frac{-h^{-2}}{12}\left\{f(x_0 - 2h) - 16f(x_0 - h) + 30f(x_0) - 16f(x_0 + h) + f(x_0 + 2h)\right\}$$

$U_1^{(0)}$ tem erro $R_5 = \mathcal{O}(h^4)$.

Deve ficar claro que, na prática, os $U_4^{(j)}$ nunca são expressos pelos valores funcionais mas sim obtidos diretamente do algoritmo A. Assim o procedimento acima serve apenas como ilustração.

Exemplo:

Determinando $f''(1)$ com $f(x) = \frac{1}{\text{sen } x}$, o algoritmo A nos dá a seguinte tabela:

h_k	$U(h_k)$	$U_1^{(j)}$	$U_2^{(j)}$	$U_3^{(j)}$
0,64	3,5750	2,0096	2,1725	2,1683
0,32	2,4009	2,1622	2,1683	
0,16	2,2219	2,1680		
0,08	2,1815			

; Valor exato: $f''(1) = \frac{1 + \cos^2 1}{\text{sen}^3 1} = 2,168\ 30\ldots$

§ 2.4. - INTEGRAÇÃO NUMÉRICA DE EQUAÇÕES DIFERENCIAIS POR EXTRAPOLAÇÃO AO LIMITE

Tratamos agora da aplicação da extrapolação ao limite à resolução numérica de equações diferenciais, como já foi indicado no exemplo 3 da introdução.

Chamemos de $U(h)$ a aproximação da solução $y(x)$ de $y'(x) = f(x,y)$; $y(x_0) = \eta_0$ obtida no ponto $x \in [a,b]$, por algum método de ordem q com passo h. Para se obter aproximações de ordem superior a q com uma extrapolação ao limite é necessário assegurar a existência de um desenvolvimento de $U(h)$ da forma

$$U(h) = y(x)+\tau_q(x)h^q+\tau_{q+1}(x)h^{q+1}+\ldots+\tau_n(x)h^n+K_{n+1}(h,x).h^{n+1} \qquad (2.42)$$

onde os $\tau_i(x)$, $i = q, q+1,\ldots,n$ não dependem de h e $|K_{n+1}(h,x)|<M$ para $h \in [0,h_0]$ e $x \in [a,b]$.

Pode-se demonstrar, vide |4;5| , que para qualquer método de passo simples ou múltiplo êsse desenvolvimento existe se $f(x,y)$ tem derivadas parciais de ordem $(n+1)$ contínuas.

No caso da regra do ponto médio (vide (1.7), se $f \in \mathbb{C}^{2n+2}[a,b]$ temos:

$$U(h) = y(x) + \tau_1(x)h^2 + \tau_2(x)h^4 + \ldots + \tau_n(x)h^{2n} + K_{n+1}(h,x)h^{2n+2}$$

O mesmo desenvolvimento vale para a seguinte modificação da regra de ponto médio, devida a Gragg:

$$x_k = x_0 + kh \qquad k = 0,1,\ldots,N$$

$$z_0^* = \eta_0 \; ; \; z_1^* = \eta_0 + hf(x_0,\eta_0)$$

$$z^*_{k+1} = z^*_{k-1}+2hf(x_k,z_k^*) \qquad k = 1,2,\ldots,N-1$$

$$y_N^* = \frac{1}{2}\,(z^*_{N-1} + z_N^* + hf(x_N, z_N^*))$$

A extrapolação ao limite é, portanto, aplicável e nos dá:

Algoritmo 2.2

Escolhe-se $\varepsilon > 0$, um natural N (recomenda-se $N=6$) e uma sequência h_j com $h_j = \dfrac{(x_1 - x_0)}{N_j}$, $N_j < N_{j+1}$ naturais e pares.

Calcula-se

$$t_1 = x_0 + h_0 \; ; \; z_0^* = \eta_0; \; z_1^* = \eta_0 + h_0\, f(x_0,\eta_0)$$

$$t_{k+1} = t_k + h \; ; \; z^*_{k+1} = z^*_{k-1} + 2h_0 f(t_k, z^*_k) \; ; \quad k = 1,2,\ldots,N_0 - 1$$

$$U_0^{(0)} = \frac{1}{2}\left(z^*_{N_0 - 1} + z^*_{N_0} + h_0 f(t_{N_0}, z^*_{N_0})\right)$$

e determina-se para $j = 1,2,\ldots$

$$t_1 = x_0 + h_j \; ; \quad z^*_0 = \eta_0 ; \; z^*_1 = \eta_0 + h_j \, f(x_0, \eta_0)$$

$$t_{k+1} = t_k + h_j \; ; \; z^*_{k+1} = z^*_{k-1} + 2h_j \, f(t_k, z^*_k) \; ;$$

$$k = 1,2,\ldots,N_j - 1$$

$$U_0^{(j)} = \frac{1}{2}\left(z^*_{N_j - 1} + z^*_{N_j} + h_j f(t_{N_j}, z^*_{N_j})\right)$$

$$U_{d+1}^{(j-d)} = U_d^{(j-d+1)} + \frac{U_d^{(j-d+1)} - U_d^{(j-d)}}{\left(\frac{h_{j-d}}{h_{j+1}}\right)^2 - 1} \; ; \quad d = 0,1,\ldots,m$$

$m = \min(j, n-1, N)$ (n definido em (2.42))

Termina-se o processo, se $|U_m^{(j-m)} - U_m^{(j-m+1)}| < \varepsilon$ e usa-se $U_{m+1}^{(j-m)}$ como aproximação de $y(x_1)$.

Substitui-se x_0 por x_1 e η_0 por $U_{m+1}^{(j-m)}$ e repete-se o algoritmo para a determinação de $y(x_2)$, etc.

Recomendamos a escolha $N_j = 2,4,6,8,12,16,24,\ldots$. O natural N limita o número de colunas no esquema dos $U_k^{(j)}$ com o objetivo de se limitar a acumulação do êrro.

OBSERVAÇÕES BIBLIOGRÁFICAS

1. - Já Arquimedes, em 250 A.C., usou, para determinação do número π, um método que pode ser considerado como caso particular da extrapolação ao limite. Em 1910 Richardson introduziu a denominação "deferred approach to the limit" e aplicou este método em vários problemas diferentes. Uma sobrevista minuciosa do desenvolvimento do método e de suas várias aplicações encontra-se na publicação de Joyce |1|.

2. - Os algoritmos A e B e considerações de erro e convergência, são tratados detalhadamente no trabalho clássico de Bulirsch e Stoer |9|. Essas considerações sendo muito complicadas se simplificam bastante com a nova abordagem do método apresentado em 2.2.1, que é baseado em trabalhos próprios do autor |16|.

3. - O teorema 2.4 também é novo e generaliza teoremas correspondentes de Bulirsch e Stoer, simplificando muito suas demonstrações, além disso, fornece uma estimativa do erro para um algoritmo apresentado por Pereyra em |13| (fórmulas (1.3) - (1.5)), que representa o caso mais geral da extrapolação ao limite e, assim, contém os algoritmos A e B como casos especiais.

4. - Uma teoria geral sobre o desenvolvimento de U(h) em uma série da forma (2.25), no caso de métodos de discretizações muito gerais, encontra-se na publicação de Stetter |14|, que é uma continuação da tese de doutorado |4| de Gragg.

5. - Calculando-se $Q_n(-1)$, sendo $Q_n(x)$ definido na demonstração do teorema 2.6., chega-se facilmente à conclusão que $\sum_{k=0}^{n} |c_k(n)|$ é limitado no caso B. Isto simplifica bastante a demonstração do teorema 1 em |8|.

BIBLIOGRAFIA DA PARTE II

1. Joyce, D.C. : Survey of Extrapolation Processes in Numerical Analysis, SIAM Review, Vol. 13, No. 4, 1971.

2. Smirnov,V.I. : A Course of Higher Mathematics, III, 2 Addison - Wesley, 1964.

3. Whittaker, E.T. ; Watson, G.N. : A Course of Modern Analysis, Cambridge Univ. Press, 4ª ed., 1969.

4. Gragg, W.B. : Repeated Extrapolation to the Limit in the Numerical Solution of Ordinary Differential Equations Doctoral Thesis, Univ. of California, 1964.

5. —— : On Extrapolation Algorithms for Ordinary Initial Value Problems, SIAM J. Numer. Anal. 2, 1965.

6. Rutishauser, H. : Ausdehnung des Rombergschen Prinzips, Num. Math. 5, 1963.

7. Romberg, W. : Vereinfachte numerische Integration, Kgl. Norske Vid. Selsk. Forh. 28, 1955.

8. Bulirsch, R. : Bemerkungen zur Romberg-Integration, Num. Math. 6, 1964.

9. Bulirsch. R.; Stoer, J. : Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-

Typus, Num. Math. 6, 1964.

10. —— : Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods, Num. Math. 8, 1966.

11. —— : Numerical Quadrature by Extrapolation, Num. Math. 9, 1967.

12. Gantmacher, F. R. : The Theory of Matrices, Vol. II, Chelsea Publ. Comp. 1971.

13. Pereyra, V.: Accelerating the Convergence of Discretisation Algorithms, SIAM J. Num. Anal., Vol. 4, No.4, 1967.

14. Stetter, H.: Asymptotic Expansions for the Error of Discretisation Algorithms for Non-linear Functional Equations Num. Math. 7, 1965.

15. Lindberg, B. : A Simple Interpolation Algorithm for Improvement of the Numerical Solution of a Differential Equation, SIAM J. Numer. Anal., Vol. 9, No.4, 1972.

16. Albrecht, P.: Considerações Sôbre o Erro na Extrapolação de Richardson, Proceedings do Primer Congreso Iberoamericano de Informática, Buenos Aires, 1972.

17. —— : Um Curso de Análise Numérica Ao Livro Técnico, R.J. 1973.