UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

INSTITUTO DE GEOCIÊNCIAS

DEPARTAMENTO DE ASTRONOMIA

OBSERVATÓRIO DO VALONGO

DETERMINAÇÃO DE ÓRBITAS PRELIMINARES PERTURBADAS A PARTIR DE ÓRBITAS DE DOIS CORPOS OBTIDAS PELO MÉTODO DE GAUSS

ALUNAS: DANIELA LAZZARO E MERY PASSOS PINHEIRO

ORIENTADOR: JOSÉ AUGUSTO BUARQUE DE NAZARETH

RIO DE JANEIRO, MARÇO, 1979

ERRATA

pag.ii, 9ª linha - onde se lê "...cálculo das perturbações."
leia-se "...cálculo das perturbações em termos mais rigorosos."

pag.v, 7ª linha - onde se lê "VI.6 - Cometa 1977 HB..."
leia-se "VI.6 - 1977 HB..."; também ao longo de todo o trabalho, o corpo celeste 1977 HB foi erroneamente denominado cometa, já que os Telegramas do I.A.U. nada especificam sobre a natureza de dito corpo.

pag.7, 5ª linha - onde se lê "...energia potencial dada pelo gradiente da força."
leia-se "...energia potencial cujo gradiente determina a força."

pag.17, 8ª linha - onde se lê "...da mesma corrigir a paralaxe diurna."
leia-se "...da mesma tornar desnecessária a correção da paralaxe diurna."

pag.24, 2ª linha - onde se lê "...utiliza-se um método de aproximação."
leia-se "...utiliza-se um método de aproximações sucessivas."

pag.24, OBS. 1 - no final do parágrafo, foi omitida a frase: "Este problema será discutido, com maiores detalhes, nas Conclusões, item VII.3 ."

pag.31, 3ª linha - onde se lê "...correções diferenciais de Gauss..."
leia-se "...correções diferenciais gaussianas..."
A mesma correção deve ser feita nas seguintes páginas:
pag.31, 7ª linha,
pag.34, 13ª linha,
pag.37, 10ª linha.

pag.43,22ª linha - onde se lê "...as perturbações com elevados movimentos médios..."
leia-se "...as perturbações dos planetas com elevados movimentos médios..."

pag.52,14ª linha - onde se lê "...movimento médio em radiano por dia solar médio..."
leia-se "...movimento médio em graus..."; também em todas as tabelas da seção Análise de Resultados, as unidades do movimento médio são dadas erroneamente como rad/dsm, quando, na realidade, são graus.

pag.61, 4ª linha - onde se lê "...elementos orbitais para Ceres:"
leia-se "...elementos orbitais para Ceres, observando-se que a anomalia média está referenciada a 1º de Janeiro de 1978:"

pag.73,21ª linha - onde se lê "...é aplicável se uma quarta for utilizada."
leia-se "...é aplicável se uma quarta observação é utilizada."

pag.74, item VII.6 - foi omitido o parágrafo:
"Para os casos em que D é nulo ou muito próximo a zero, o sistema (9) não pode ser obtido. Neste caso outras formas de solução para o sistema (8) devem ser utilizadas, por exemplo, o processo de eliminação de Gauss."
O parágrafo acima antecede imediatamente o item VII.7 das Conclusões.

pag.73, item VII.5 - A redação correta deste parágrafo é:

"Para exemplificar, quando utilizamos as datas 31 de outubro, 7 e 14 de novembro de 1978, o processamento da órbita do cometa Kohler foi interrompido devido a uma raiz inválida ($\sqrt{1 - e^2}$ para $e > 1$). Contudo, com as datas 1, 11 e 21 de junho de 1978, os resultados foram satisfatórios (ver secção VI). A primeira idéia é que o esquema geral utilizado seja, de alguma forma, sensível ao espaçamento entre as datas de observação. A solução deste problema envolverá, certamente, processamento de um maior número de órbitas, juntamente com um reexame das fórmulas utilizadas."

Aos nossos pais

Agradecemos ...

... ao professor e amigo José Augusto Buarque de Nazareth, pela formação que nos proporcionou durante o último ano de curso, no qual lançou bases fundamentais para o desenvolvimento deste projeto; bem como pelo constante apoio demonstrado ao longo da realização deste.

... ao professor Getúlio de Jesus Villar e ao amigo Valdomiro de Oliveira Junior pelo auxílio que nos forneceram na parte computacional.

... a Maria Grazia pelo excelente trabalho de datilografia e pela paciência demonstrada na realização deste.

... finalmente, aos professores e colegas do Observatório do Valongo, que pela insistente pergunta " Saiu o programa?" forçaram a nos e ao computador a rápida conclusão do projeto.

INTRODUÇÃO

O objetivo do projeto é a formulação de um roteiro para cálculo de órbitas preliminares, que possa servir como ponto de partida para determinação de órbitas perturbadas mais rigorosas.

Devido ao seu interesse intrínseco em problemas astronômicos, optamos pela determinação de órbitas de dois corpos (elíticas) a partir do método de Gauss. A partir desta fase, o programa incorpora a integração das equações de movimento perturbadas em coordenadas retangulares, e produz, em suas partes finais, elementos orbitais e efemérides já corrigidas de perturbações. Os exemplos apresentados se referem a cometas e asteróides (ver secção VI) e como presença perturbadora considerando-se apenas a do planeta Júpiter, porém nenhuma limitação, no programa, foi imposta quanto ao número de corpos perturbadores.

Conforme observado nas Conclusões (secção VII), alguns problemas foram detectados em exemplos específicos. Dependendo de circumstâncias específicas, das quais algumas já foram detectadas, o programa na sua forma atual (ver anexo II) pode produzir distâncias geocêntricas negativas e, ou, não convergência das correções diferenciais. Para o primeiro deles, possíveis alternativas foram propostas, mas não incorporadas ao programa. Quanto ao segundo, as causas que dão origem às divergências observadas ainda não foram fixadas. Certamente esta questão e outras que porventura vierem a ser levantadas, dependem de aplicações sistemáticas e posteriores estudos, a

fim de que possam ser sanadas.

Ainda na secção VI, notamos que os resultados perturbados não foram submetidos a comparações com os dados de outros autores, tal como o fizemos para os elementos orbitais obtidos por órbitas dois corpos. A razão é evidente e se deve apenas ao fato de que antes de eliminarmos as fontes de erro das órbitas elíticas, nenhuma garantia se pode ter quanto aos resultados perturbados, que constituem o ponto de partida para o cálculo das perturbações.

Se as equações rigorosas a serem integradas numericamente forem as equações planetárias de Lagrange, então este programa produzirá tabelas numéricas que são eficazes na inicialização daquelas integrações. Apesar da mesma observação se aplicar para o caso de adotarmos o método de Cowell, observamos que a subrotina DREBS (ver subsecção V.2) não poderá mais ser utilizada, pois que neste caso a integração é diretamente sobre equações diferenciais de 2ª ordem; aqui a sugestão é de uma integração numérica pelo denominado procedimento Σ^2, ou de Gauss-Jackson. Quanto ao método de Enck, o processo Σ^2 ainda será útil, porém algumas modificações de conteúdo deverão ser processadas (Herrick, 9, cap. XIV). Ainda sobre a DREBS, sabemos que ela opera sobre sistemas de equações diferenciais da forma

$$\frac{dy}{dt} = F(y,t) ,$$

mas porque não obtivemos as expansões das funções de perturbação segundo os elementos orbitais dos corpos perturbadores,

convertemos o sistema acima em um autônomo, isto é, da forma

$$\frac{dy}{dt} = F(y),$$

admitindo que o corpo perturbador não varia sua posição no tempo apreciavelmente. Este tema será discutido na subseção V.1. Entretanto voltamos a frisar que o resultado deste programa pretende apenas fornecer elementos de inicialização para integrações mais rigorosas. Além das alternativas sugeridas na citada subsecção, observamos que integrações em coordenadas retangulares podem também ser feitas a partir do método de Hamsen (Brouwer e Clemence, 1), porém aqui também estas questões só poderão ser decididas a partir de um exame mais profundo das diferentes possibilidades.

Finalizando, o programa tal como foi formulado ao longo do segundo semestre de 1978 não deve ser utilizado em trabalhos sistemáticos antes que as correções que são referidas ao longo do texto sejam incorporadas.

SUMÁRIO

Pag.

Pag.

I - CONCEITOS FUNDAMENTAIS DO MOVIMENTO ELÍTICO

I.1 - Introdução

A presente secção visa fornecer informações mínimas necessárias ao desenvolvimento do projeto. As demonstrações não serão apresentadas, dado que fazem parte de disciplinas obrigatórias do curso de Astronomia (Astronomia V e Mecânica Teórica). Os detalhes matemáticos que foram omitidos podem ser encontrados em, por exemplo, Fitzpatrick (6 , cap. 2 e 3) e Danjon (5 , cap. 10).

I.2 - O Problema dos dois corpos

O problema do movimento de dois corpos que se atraem com uma força que depende somente da distância entre eles, é fundamental em Mecânica Celeste.

O movimento de um planeta do sistema solar é devido, em primeira aproximação, somente ao campo gravitacional do Sol, não considerando forças externas. A mesma aproximação é usada em estudos preliminares do movimento de asteróides e cometas em torno do Sol, e de satélites em torno de seus primários. Assim, o movimento de planetas, asteróides e satélites é descrito com razóavel precisão por trajetórias elíticas.

O movimento de um corpo celeste cuja trajetória é uma elipse, é completamente caracterizado por seis constantes , mutuamente independentes, denominadas elementos orbitais, geometricamente representadas nas figuras que se seguem.

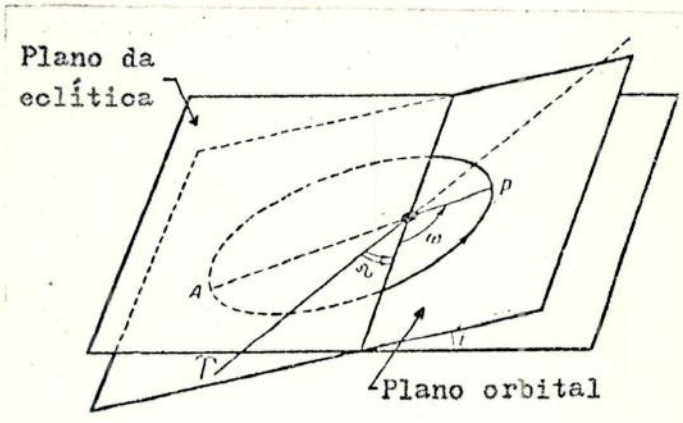


Fig.1 - Elementos orbitais

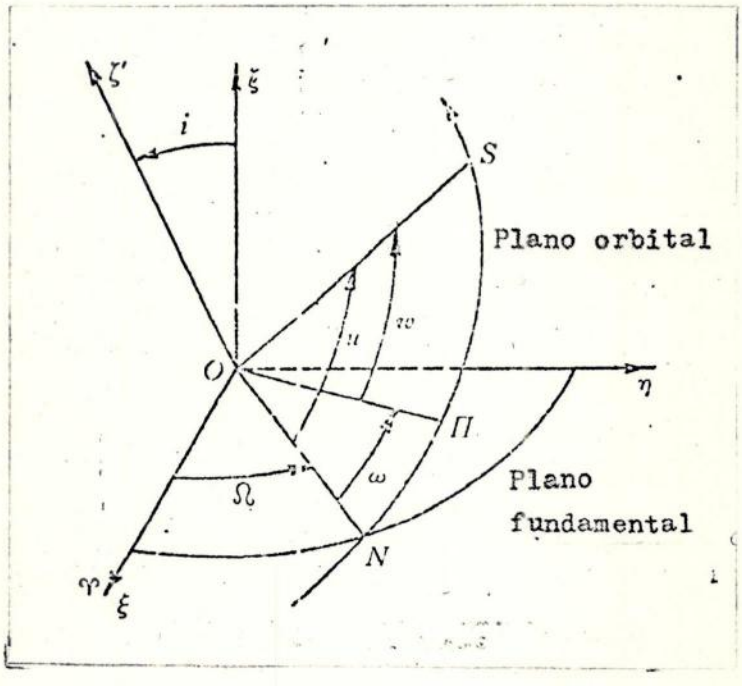


Fig.2 - Elementos orbitais

Dois elementos - a inclinação e a longitude do nodo ascendente - definem a posição no espaço do plano que contém a órbita:

- a inclinação (i) é o ângulo entre o plano da órbita e o plano de referência; sua variação está contida no intervalo $[0º,180º]$. O movimento é dito retrógrado se $i > 90º$ e direto se $i \leq 90º$;
- a longitude do nodo ascendente (Ω) é o ângulo entre a direção do ponto vernal e a direção do nodo ascendente, medido no plano de referência.

O terceiro elemento - argumento do pericentro - define a orientação da órbita:

- o argumento do pericentro (ω) é o ângulo entre a direção do nodo ascendente e a direção do pericentro . Está contido no plano da órbita e é medido de 0º a 360º na direção do movimento.

O quarto e quinto elementos - semi-eixo maior e excentricidade - definem o tamanho e a forma da órbita. O intervalo de variação da excentricidade é (0,1) para a elipse , 1 para a parábola, e maior que 1 para a hipérbole.

O sexto elemento define a posição do corpo na órbita num determinado instante de tempo, geralmente o instante da passagem pelo pericentro (T).

O conjunto de elementos descrito acima, é geralmente denominado de conjunto "clássico". A partir dele, e segundo as necessidades do problema, outras constantes podem ser deduzidas :

$$n = \sqrt{\frac{\mu}{a^3}} \quad , \quad P = \frac{2\pi}{n} \quad , \quad M_o = n(T_o - T) \quad , \quad q = a(1 - e),$$

$$p = a(1 - e^2) \ , \ \tilde{\omega} = \omega + \Omega \ , \ \tilde{\omega}_r = \omega - \Omega \ ,$$

e as componentes de $\vec{P}$ e $\vec{Q}$ (fig. 3).

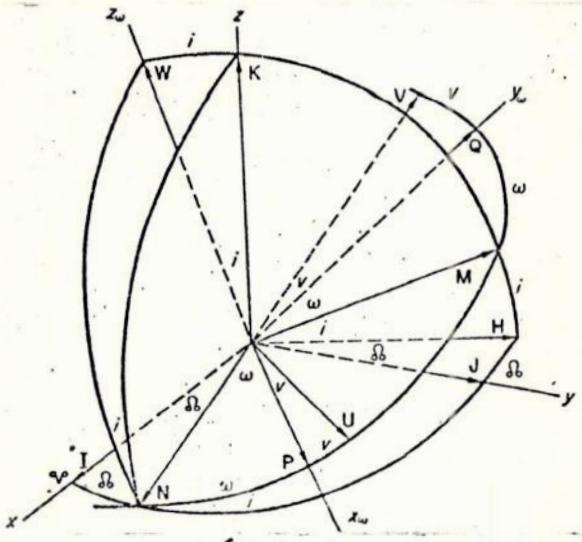


Fig.3

No entanto, para a parábola tem-se $a = \infty$ e $e = 1$, de tal maneira que não são suficientes para diferenciar uma parábola de outra, e portanto, usa-se q no lugar de a . No caso da inclinação ser pequena, tanto ω quanto Ω são determinados de forma precária, e, então, um deles pode ser substituído por $\tilde{\omega}$, melhor determinado. Se a inclinação é próxima a 180º, $\tilde{\omega}$, ω e Ω são também mal determinados, e $\tilde{\omega}_r$ é bem determinado. Para finalizar, as componentes $\vec{P}$ e $\vec{Q}$ são, às vezes, preferidas, pois elas permitem calcular as efemérides mesmo sem ter calculado i, ω, Ω ; é bom, no entanto, ressaltar que as relações entre eles devem ser verificadas:

$$\vec{P} \cdot \vec{P} = 1 \ ,$$

$$\vec{Q} \cdot \vec{Q} = 1 \ ,$$

$$\vec{P} \cdot \vec{Q} = 0 \ .$$

Obviamente, para uma maior generalização, pode-se usar, como

elementos orbitais, o conjunto:

$$\vec{r}_o = (x_o, y_o, z_o)$$

$$\dot{\vec{r}}_o = (\dot{x}_o, \dot{y}_o, \dot{z}_o)$$

Para maiores detalhes, vide Herrick (8 , cap. 3).

I.3 - Equações Fundamentais

A equação do movimento relativo de duas partículas de massa m_o e m, sujeitas a uma força que depende somente da distância entre elas, e considerando-se constante a intensidade de possíveis campos externos, é

$$\ddot{\vec{r}} + k^2 M \frac{F(r)}{r} \vec{r} = \vec{0} , \qquad (1)$$

onde

- $M = m_o + m$
- $\vec{r}$ é o vetor posição da massa m em relação à massa m_o, considerada como origem,
- k^2 é a constante gravitacional, k = 0,01720209895 ("The American Ephemerides and Nautical Almanac of 1978")

A força $\vec{F}$ que age sobre a partícula de massa m pode ser

expressa por

$$\vec{F} = - k^2 m_o m \, F(r) \vec{\mu}_{o,m} ,$$

onde $\vec{\mu}_{o,m}$ é o vetor unitário dirigido de m_o a m.

O problema de dois corpos descrito pela equação (1) é equivalente ao problema de um corpo sobre o qual age uma força da forma

$$\vec{F} = - k^2 m M \left[\frac{F(r)}{r}\right] \vec{r} ,$$

dirigida para o centro de força. Portanto, o problema se resume ao de uma partícula de massa m, movendo-se em torno de um centro de força fixo, tomado como origem, sob a influência de uma força central da forma

$$F = - \mu m \frac{F(r)}{r} \vec{r} , \qquad (2)$$

onde

$$\mu = k^2 (m_o + m) .$$

No caso em que m_o é a massa do Sol, considerada unitária, m é a massa de um cometa ou de um asteróide, e apenas dois corpos são considerados, podemos tomar

$$m \approx 0 \qquad e \qquad \mu = k^2 .$$

Considerando o caso em que $F(r) = r^{-2}$, temos, a partir da equação (2),

$$\vec{F} = - \mu m r^{-3} \vec{r} .$$

Sabendo-se que o campo é conservativo, podemos associar a ele uma energia potencial dada pelo gradiente da força. Isto nos permite obter uma equação que exprime a lei de conservação de energia e, em consequência, a equação da trajetória, expressa em coordenadas polares (r, θ) por

$$|\vec{r}| = \frac{h^2 \mu^{-1}}{1 + \left[1 + (2h^2 E / m\mu^2)\right]^{1/2} \cos w} , \qquad (3)$$

onde

- $h = |\vec{h}|$ é uma constante de integração denominada constante das áreas,
- E: a energia total do sistema,
- $w = \theta + \omega^*$,
- ω^*: uma constante de integração determinada pelas condições iniciais.

A equação (3) é a equação de uma cônica. Comparando-se a equação (3) com

$$r = \frac{2 e p}{1 + e \cos \theta} ,$$

temos

$$e = \left(1 + \frac{2h^2 E}{m\mu^2}\right)^{1/2}$$

e

$$2ep = \frac{h^2}{\mu}$$

onde e é a excentricidade da cônica e p é um parâmetro associado ao semi-latus retum através da expressão 2ep.

Para a elipse, as seguintes equações são verdadeiras:

$$2ep = a\,(1 - e^2)$$

$$h^2 = \mu a\,(1 - e^2)$$

$$v^2 = \mu\left(\frac{2}{r} - \frac{1}{a}\right)$$

onde

- v é a velocidade
- a o semieixo maior da elipse.

Da figura 4 deduzem-se as seguintes relações:

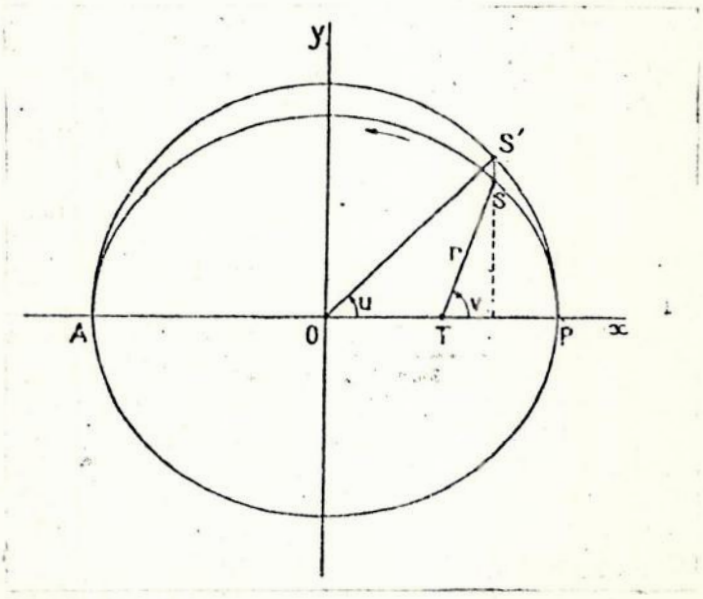


Fig.4 - Relação entre as anomalias verdadeira, v, e excêntrica, u.

$$\begin{cases} x = a \cos U, \\ y = b \operatorname{sen} U = a \sqrt{1 - e^2} \operatorname{sen} U, \end{cases}$$

$$\begin{cases} r \cos V = x - ae = a (\cos U - e), \\ r \operatorname{sen} V = y = a \sqrt{1 - e^2} \operatorname{sen} U, \end{cases}$$

$$\begin{cases} r = a (1 - e \cos U), \\ \cos V = \dfrac{\cos U - e}{1 - e \cos U}, \\ \operatorname{sen} V = \dfrac{\sqrt{1 - e^2} \operatorname{sen} U}{1 - e \cos U}, \end{cases}$$

onde

- U = anomalia excêntrica
- V = anomalia verdadeira.

Tendo em vista que as coordenadas foram expressas somente em função de U, resta calcular a anomalia excêntrica em função do tempo a fim de que se possa determinar a posição do planeta em qualquer instante t. Para tanto, consideremos as relações

$$\begin{cases} r^2 \dfrac{dV}{dt} = h \text{ , ou, } r^2 \, dV = na^2 \sqrt{1 - e^2} \, dt \text{ ,} \\ \\ \operatorname{tg} \dfrac{V}{2} = \sqrt{\dfrac{1 + e}{1 - e}} \operatorname{tg} \dfrac{U}{2} \text{ .} \end{cases} \tag{4}$$

Diferenciando-se a segunda das expressões do sistema acima ,

$$\frac{dV}{\cos^2 \frac{V}{2}} = \sqrt{\frac{1+e}{1-e}} \frac{dU}{\cos^2 \frac{U}{2}} ,$$

ou

$$\frac{dV}{dU} = \sqrt{1-e^2} \frac{a}{r} = \frac{\text{sen } V}{\text{sen } U} . \qquad (5)$$

Eliminando-se V entre as relações (4) e (5), obtemos a importante relação

$$\frac{dU}{dt} = \frac{na}{r} = \frac{n}{1 - e \cos U} ,$$

onde n é o movimento médio, dado por $(\mu/a^3)^{1/2}$. Esta é uma equação diferencial com variáveis separadas, que, quando integrada, fornece a seguinte expressão:

$$U - e \text{ sen } U = n(t - t_o) ,$$

denominada equação de Kepler e mais comumente representada na forma

$$M = U - e \text{ sen } U,$$

onde

- M é anomalia média dada por $M = n(t - t_o)$,

- t_o a época da passagem pelo pericentro,
- t o instante para o qual se deseja calcular a anomalia média.

I.4 - Sistemas de Coordenadas Astronômicas

A escolha do sistema de referência é circunstancial, isto é, é ditada pelo problema em questão.

As observações de qualquer corpo celeste são geralmente dadas num sistema de coordenadas equatoriais topocêntricas. A origem é colocada no ponto de observação (M), sobre a superfície da Terra (vide fig. 5).

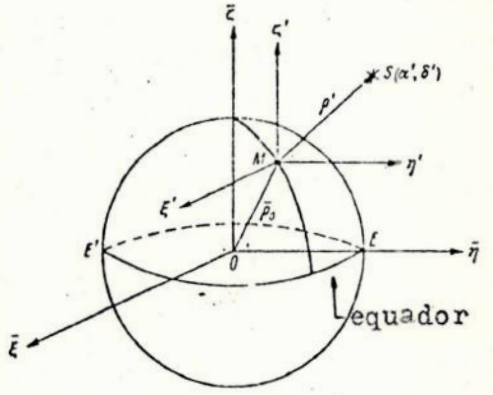


Fig.5 - Coordenadas topocêntricas e geocêntricas.

O plano fundamental $\xi'\eta'$ é paralelo ao plano do equador terrestre, o eixo ξ' é perpendicular a este plano e passa pelo ponto vernal. As coordenadas esféricas deste sistema são

ρ', α' e δ', relacionadas com as coordenadas retangulares pelas fórmulas

$$\xi' = \rho' \cos\delta' \cos\alpha' ,$$

$$\eta' = \rho' \cos\delta' \operatorname{sen}\alpha' ,$$

$$\zeta' = \rho' \operatorname{sen}\delta' ,$$

onde:

- ρ' é o raio vetor topocêntrico,
- α' e δ' a ascensão reta e declinação topocêntricas, respectivamente.

Considere-se agora o sistema de coordenadas equatoriais geocêntricas $O\bar{\xi}\bar{\eta}\bar{\zeta}$ (fig. 5), que é obtido a partir do sistema topocêntrico através de uma translação ao longo do vetor $\vec{\rho}_o$. Este é o vetor que determina o ponto de observação M em relação ao centro de inércia da Terra. Centro de massa ou centro de inércia de um corpo sólido é definido como o ponto geométrico C, cujas coordenadas são dadas por

$$X = \frac{1}{M} \iiint \rho x dV ,$$

$$Y = \frac{1}{M} \iiint \rho y dV ,$$

$$Z = \frac{1}{M} \iiint \rho z dV ,$$

onde ρ é a densidade de massa por unidade de volume, e a integração pode ser estendida sobre todo o volume ou sobre to-

do o espaço, já que $\rho = 0$ para os pontos externos ao corpo .

Para passar do sistema de coordenadas topocêntricas para geocêntricas torna-se necessário conhecer as coordenadas do ponto de observação em relação ao centro de inércia da Terra. Entretanto, como bem diz Chebotarev (4), "a figura da Terra (geóide) e a posição do centro de inércia dentro da Terra são, falando rigorosamente, desconhecidas." O cálculo destas coordenadas na prática é feito associando-se à forma da Terra um elipsóide de revolução, onde o centro deste coincide com o centro de inércia da Terra (vide fig. 6).

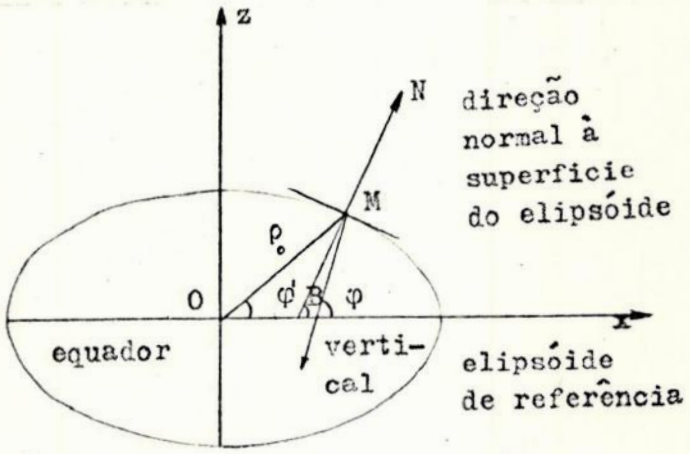


Fig.6 – Latitudes geocêntrica (φ'), astronômica (φ), e geodésica (B) de um ponto de observação M.

Numericamente, ρ_o é obtido a partir da latitude astronômica (ou geográfica), por meio da fórmula

$$\rho_o = 0{,}9983200 + 0{,}0016835 \cos 2\varphi - 0{,}0000035 \cos 4\varphi \ .$$

Para maiores detalhes, vide Danjon (5) ou Herrick (8).

A posição do ponto de observação é, portanto, obtida pelas equações

$$\xi_1 = \rho_o \cos\varphi' \cos TSL \ ,$$

$$\eta_1 = \rho_o \cos \varphi' \operatorname{sen} TSL ,$$

$$\zeta_1 = \rho_o \operatorname{sen} \varphi' ,$$

onde

- TSL é tempo sideral local da observação.

Em consequência, as equações que passam do sistema geocêntrico para o topocêntrico serão

$$\bar{\xi} = \xi' + \xi_1 ,$$

$$\bar{\eta} = \eta' + \eta_1 ,$$

$$\zeta = \zeta' + \zeta_1 .$$

Para transformar coordenadas geocêntricas em coordenadas heliocêntricas, utilizam-se as relações

$$\bar{x} = \rho \cos \delta \cos \alpha - X ,$$

$$\bar{y} = \rho \cos \delta \operatorname{sen} \alpha - Y ,$$

$$\bar{z} = \rho \operatorname{sen} \delta - Z ,$$

onde

- $\bar{x}$, $\bar{y}$, $\bar{z}$ são as coordenadas retangulares equatoriais heliocêntricas,
- X, Y, Z são as coordenadas geocêntricas equatoriais do Sol.

A transformação das coordenadas equatoriais heliocêntricas em eclíticas heliocêntricas é feita através das seguintes equações:

$$x' = \bar{x} ,$$

$$y' = \bar{y} \cos \varepsilon - \bar{z} \operatorname{sen} \varepsilon ,$$

$$z' = \bar{y} \operatorname{sen} \varepsilon - \bar{z} \cos \varepsilon ,$$

onde ε é o ângulo entre os planos da eclítica e do equador para uma época t_o.

Sendo que as perturbações causadas pelo Sol, Lua e planetas mudam constantemente a posição do plano do equador no espaço, torna-se necessário especificar a época na qual se referencia um determinado sistema de coordenadas. É comum em trabalhos astronômicos adotar-se como épocas de referência 1900,0 , 1925,0 , 1950,0 ou 1975,0. No presente trabalho adotou-se, como época de referência, 1950,0.

II - MÉTODO DE GAUSS

II.1 - Aspectos Fundamentais

Conhecidas três posições sucessivas de um corpo celeste, a pequenos intervalos de tempo, o método de Gauss, utilizando aproximações sucessivas, fornece os elementos que definem a órbita daquele corpo. Teoricamente, nada impede que os intervalos de tempo utilizados sejam arbitrariamente grandes ou pequenos. Mas, a experiência astronômica tem demonstrado que melhores resultados são obtidos se forem considerados intervalos de alguns dias, no caso de cometas, e de três a quatro semanas, no caso de asteróides suficientemente longe da esfera de ação de um planeta perturbador.

Sejam XT, YT, ZT as coordenadas topocêntricas do Sol; x, y, z as coordenadas equatoriais heliocêntricas do corpo celeste considerado; λ, μ, ν os co-senos diretores das posições observadas. Tem-se, então:

$$\begin{cases} x_i = \lambda_i \Delta_i - XT_i , & i = 1,2,3 \\ y_i = \mu_i \Delta_i - YT_i , \\ z_i = \nu_i \Delta_i - ZT_i , \end{cases} \qquad (6)$$

onde

Δ = distância geocêntrica do corpo celeste,

$\lambda = \cos\delta \cos\alpha$,

$\mu = \cos\delta \operatorname{sen}\alpha$,

ν = sen δ ,

δ, α = declinação e ascensão reta do corpo celeste obtidos pela observação.

OBS.: A) O índice i corresponde a observações feitas nos instantes t_1, t_2 e t_3.

B) A razão de considerar-se as coordenadas topocêntricas do Sol em vez de coordenadas geocêntricas deve-se ao fato da mesma corrigir a paralaxe diurna.

As coordenadas heliocêntricas do corpo celeste estão relacionadas com as razões das áreas S_1, S_2, S_3 (fig.7) como se segue:

$$\begin{cases} x_1 \dfrac{S_2}{S_3} - x_2 + x_3 \dfrac{S_1}{S_3} = 0 , \\ y_1 \dfrac{S_2}{S_3} - y_2 + y_3 \dfrac{S_1}{S_3} = 0 , \\ z_1 \dfrac{S_2}{S_3} - z_2 + z_3 \dfrac{S_1}{S_3} = 0 . \end{cases} \qquad (7)$$

Para uma demonstração deste sistema veja Apêndice 1.

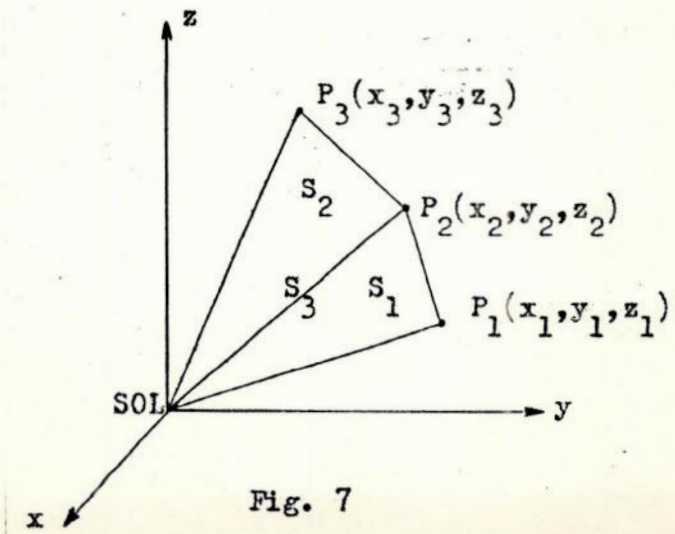


Fig. 7

Substituindo-se as equações do sistema (6) no sistema (7), tem-se

$$\begin{cases} (\lambda_1\Delta_1 - XT_1)\dfrac{S_2}{S_3} - \lambda_2\Delta_2 + XT_2 + (\lambda_3\Delta_3 - XT_3)\dfrac{S_1}{S_3} = 0\,, \\ (\mu_1\Delta_1 - YT_1)\dfrac{S_2}{S_3} - \mu_2\Delta_2 + YT_2 + (\mu_3\Delta_3 - YT_3)\dfrac{S_1}{S_3} = 0\,, \\ (\nu_1\Delta_1 - ZT_1)\dfrac{S_2}{S_3} - \nu_2\Delta_2 + ZT_2 + (\nu_3\Delta_3 - ZT_3)\dfrac{S_1}{S_3} = 0\,, \end{cases}$$

ou,

$$\begin{cases} \lambda_1\Delta_1\dfrac{S_2}{S_3} - \lambda_2\Delta_2 + \lambda_3\Delta_3\dfrac{S_1}{S_3} = \dfrac{S_2}{S_3}XT_1 - XT_2 + \dfrac{S_1}{S_3}XT_3 = L \\ \mu_1\Delta_1\dfrac{S_2}{S_3} - \mu_2\Delta_2 + \mu_3\Delta_3\dfrac{S_1}{S_3} = \dfrac{S_2}{S_3}YT_1 - YT_2 + \dfrac{S_1}{S_3}YT_3 = M \\ \nu_1\Delta_1\dfrac{S_2}{S_3} - \nu_2\Delta_2 + \nu_3\Delta_3\dfrac{S_1}{S_3} = \dfrac{S_2}{S_3}ZT_1 - ZT_2 + \dfrac{S_1}{S_3}ZT_3 = N \end{cases} \qquad (8)$$

A solução do sistema acima é da forma

$$\begin{cases} \dfrac{S_2}{S_3}\Delta_1 = A_1L + B_1M + C_1N \\ \Delta_2 = A_2L + B_2M + C_2N \\ \dfrac{S_1}{S_3}\Delta_3 = A_3L + B_3M + C_3N \end{cases} \qquad (9)$$

onde

$$A_1 = \frac{\mu_2\nu_3 - \mu_3\nu_2}{D}, \quad A_2 = \frac{\mu_1\nu_3 - \mu_3\nu_1}{D}, \quad A_3 = \frac{\mu_1\nu_2 - \mu_2\nu_1}{D},$$

$$B_1 = \frac{\nu_2\lambda_3 - \nu_3\lambda_2}{D}, \quad B_2 = \frac{\nu_1\lambda_3 - \nu_3\lambda_1}{D}, \quad B_3 = \frac{\nu_1\lambda_2 - \nu_2\lambda_1}{D},$$

$$C_1 = \frac{\lambda_2\mu_3 - \lambda_3\mu_2}{D}, \quad C_2 = \frac{\lambda_1\mu_3 - \lambda_3\mu_1}{D}, \quad C_3 = \frac{\lambda_1\mu_2 - \lambda_2\mu_1}{D},$$

$$D = \begin{vmatrix} \lambda_1 & \lambda_2 & \lambda_3 \\ \mu_1 & \mu_2 & \mu_3 \\ \nu_1 & \nu_2 & \nu_3 \end{vmatrix}$$

Os valores das razões das áreas não são conhecidos, portanto as quantidades L, M, N também não o são. Calcular-se-á as razões das áreas por aproximações sucessivas como será feito a seguir.

II.2 - Aproximações Sucessivas

II.2.1 - Primeira aproximação

As áreas são dadas por

$$2S_1 = \theta_1 h \left[1 - \frac{k}{6} \frac{1}{R_2^3} \theta_1^2 + \frac{k}{4} \frac{1}{R_2^4} \theta_1^3 \left(\frac{dR}{dt}\right)_2 \cdots \right], \qquad (10)$$

$$2S_2 = \theta_2 h \left[1 - \frac{k}{6} \frac{1}{R_2^3} \theta_2^2 + \frac{k}{4} \frac{1}{R_2^4} \theta_2^3 \left(\frac{dR}{dt}\right)_2 \cdots \right], \qquad (11)$$

$$2S_3 = \theta_3 h \left[1 - \frac{k}{6} \frac{1}{R_1^3} \theta_3^2 + \frac{k}{4} \frac{1}{R_1^4} \theta_3^2 \left(\frac{dR}{dt}\right)_1 \cdots \right], \qquad (12)$$

onde, $\theta_1 = t_2 - t_1$, $\theta_2 = t_3 - t_2$, $\theta_3 = \theta_1 + \theta_2$.

Para uma demonstração destas fórmulas veja Apêndice 2.

Nas área S_1 e S_2 será considerada como posição de origem a posição média, e essa posição média é o raio vetor $\vec{R}_2$, enquanto que a área S_3 é expressa em relação à posição um.

O desenvolvimento em série de Taylor do raio vetor da posição um em torno da posição dois será

$$\frac{1}{R_1^3} = \frac{1}{R_2^3} + \frac{3}{R_2^4} \left(\frac{dR}{dt}\right)_2 \theta_1 .$$

Portanto,

$$2S_3 = h\,\theta_3 \left[1 - \frac{k}{6} \frac{1}{R_2^3} \theta_3^2 + \frac{k}{4} \frac{1}{R_2^4} \left(\frac{dR}{dt}\right)_2 \theta_3^2 (\theta_2 - \theta_1) + \cdots \right]. \qquad (13)$$

Utilizando-se as relações (10), (11) e (13) acham-se as razões das áreas

$$\frac{S_1}{S_3} = \frac{\theta_1}{\theta_3} \left[1 + \frac{k}{6} \frac{1}{R_2^3} (\theta_3^2 - \theta_1^2) + \frac{k}{4} \frac{1}{R_2^4} \left(\frac{dR}{dt}\right)_2 \theta_2 (\theta_2 \theta_3 - \theta_1^2) + \cdots \right],$$

$$\frac{S_2}{S_3} = \frac{\theta_2}{\theta_3}\left[1 + \frac{k}{6}\ \frac{1}{R_2^3}\,(\theta_3^2 - \theta_2^2) + \frac{k}{4}\ \frac{1}{R_2^4}\,\left(\frac{dR}{dt}\right)_o \theta_1(\theta_1\theta_3 - \theta_2^2) + \ldots\right].$$

Desprezando-se os termos de ordem superior a dois nos intervalos de tempo, chega-se finalmente a

$$\begin{cases} \dfrac{S_1}{S_3} = \dfrac{\theta_1}{\theta_3}\left[1 + \dfrac{k}{6R_2^3}\,(\theta_3^2 - \theta_1^2)\right] \\[2ex] \dfrac{S_2}{S_3} = \dfrac{\theta_2}{\theta_3}\left[1 + \dfrac{k}{6R_2^3}\,(\theta_3^2 - \theta_2^2)\right] \end{cases} \qquad (14)$$

Sejam

$$AO = \frac{\theta_1}{\theta_3},$$

$$BO = \frac{\theta_1}{\theta_3} \cdot \frac{k}{6} \cdot (\theta_3^2 - \theta_1^2),$$

$$A1 = \frac{\theta_2}{\theta_3},$$

$$B1 = \frac{\theta_2}{\theta_3} \cdot \frac{k}{6} \cdot (\theta_3^2 - \theta_2^2).$$

Com estas expressões, o sistema (14) toma a forma

$$\begin{cases} \dfrac{S_1}{S_3} = AO + \dfrac{BO}{R_2^3} \\[2ex] \dfrac{S_2}{S_3} = A1 + \dfrac{B1}{R_2^3} \end{cases} \qquad (15)$$

Substituidos esses valores em (8),

$$L = A1 \cdot XT_1 - XT_2 + A0 \cdot XT_3 + \frac{(B1 \cdot XT_1 + B0 \cdot XT_3)}{R_2^3} .$$

Seja

$$A2 = A1 \cdot XT_1 - XT_2 + A0 \cdot XT_3 ,$$

$$B2 = B1 \cdot XT_1 + B0 \cdot XT_3 .$$

Portanto,

$$L = A2 + \frac{B2}{R_2^3} .$$

De maneira análoga, define-se

$$A3 = A1 \cdot YT_1 - YT_2 + A0 \cdot YT_3 ,$$

$$A4 = A1 \cdot ZT_1 - ZT_2 + A0 \cdot ZT_3 ,$$

$$B3 = B1 \cdot YT_1 + B0 \cdot YT_3 ,$$

$$B4 = B1 \cdot ZT_1 + B0 \cdot ZT_3 ,$$

donde

$$M = A3 + \frac{B3}{R_2^3} ,$$

$$N = A4 + \frac{B4}{R_2^3} .$$

Substituindo-se esses valores na segunda equação do sistema (9), tem-se

$$\Delta_2 = A5 + \frac{B5}{R_2^3} ,$$

onde

$$A5 = A_2 \cdot A2 + B_2 \cdot A3 + C_2 \cdot A4 ,$$

$$B5 = A_2 \cdot B2 + B_2 \cdot B3 + C_2 \cdot B4 .$$

Uma segunda equação é obtida escrevendo-se

$$R_2^2 = x_2^2 + y_2^2 + z_2^2 .$$

Substituindo-se as equações do sistema (6), a equação acima transforma-se em

$$R_2^2 = \Delta_2^2 - 2\Delta_2 \cdot (\lambda_2 \cdot XT_2 + \mu_2 \cdot YT_2 + \nu_2 \cdot ZT_2) + XT_2^2 + YT_2^2 + ZT_2^2 .$$

Foram obtidas assim as duas equações fundamentais do problema

$$\begin{cases} \Delta_2 = A5 + \dfrac{B5}{R_2^3} \\ \\ R_2^2 = \Delta_2^2 - 2\Delta_2(\lambda_2 \cdot XT_2 + \mu_2 \cdot YT_2 + \nu_2 \cdot ZT_2) + XT_2^2 + YT_2^2 + ZT_2^2 . \end{cases} \tag{16}$$

A solução algébrica deste sistema não é praticável, e portanto, utiliza-se um método de aproximação. Atribui-se um valor inicial, R_E, para o raio vetor da observação média (R_2), obtendo-se assim um valor para Δ_2 (vide observação 1). Com este valor de Δ_2, obtem-se um novo valor R_E do raio R_2. Testa-se, então, o valor inicial do raio, R_E, com o valor calculado R_C. Se a diferença $R_C - R_E$ for menor que o erro desejado, então adota-se o valor $R_2 = R_C$ (vide observação 2). Se a diferença for maior que o erro estipulado, calcula-se um novo valor de Δ_2, adotando $R_2 = R_C$, e obtem-se, assim, um outro valor de R_2. Testa-se esse valor com o anterior (R_C), iteragindo esse procedimento até alcançar o erro desejado.

OBS.1 : Esse valor depende da natureza da órbita; se geocêntrica, R_2 = 1.1 gr , e se heliocêntrica, R_2 =2.5U.A. para planetas menores e R_2 = 1 U.A. para planetas maiores de movimento rápido, como Betúlia, Toro, Icarus, Geographus.

OBS.2 : O erro é estipulado em função da eficiência do instrumental utilizado para observações. É de praxe, em Astronomia, utilizar-se um erro igual a 10^{-6}, que equivale a 1''.

II.2.2 - Segunda aproximação

Com os valores das distâncias geocêntricas e dos raios vetores calculados na primeira aproximação, é possível obter novos valores das razões das áreas por meio das fórmulas de Gibbs, a saber:

$$\frac{S_1}{S_3} = \frac{\theta_1}{\theta_3} \frac{1 + \frac{\psi_3}{r_3^3}}{1 - \frac{\psi_2}{r_2^3}},$$

$$\frac{S_2}{S_3} = \frac{\theta_2}{\theta_3} \frac{1 + \frac{\psi_1}{r_1^3}}{1 - \frac{\psi_2}{r_2^3}},$$

onde

$$\psi_1 = \frac{k}{12} (\theta_1\theta_3 - \theta_2^2),$$

$$\psi_2 = \frac{k}{12} (\theta_1\theta_2 - \theta_3^2),$$

$$\psi_3 = \frac{k}{12} (\theta_2\theta_3 - \theta_1^2).$$

Para uma demonstração das fórmulas de Gibbs veja Apêndice 3 .

Tendo sido obtidos novos valores para as razões das áreas, por meio do sistema (9), calculamos valores mais aproximados das distâncias geocêntricas Δ_1, Δ_2, Δ_3 e, portanto, dos raios vetores R_1, R_2, R_3.

II.2.3 - Terceira aproximação

A terceira e última aproximação do método de Gauss utiliza uma relação mais aproximada das áreas triangulares, em particular

$$\gamma = \frac{S_S}{S_T},$$

onde S_S é a área do setor curvilíneo, e S_T a área do triângulo correspondente (fig.8). A área do setor curvilíneo pode ser escrita, utilizando-se as equações do movimento elítico , da seguinte maneira,

$$2S_S = h\theta_3 = na^2\sqrt{1-e^2}\cdot\theta_3 = a^2\sqrt{1-e^2}\,(M_3 - M_1)\ ,$$

onde:

$$\theta_3 = t_3 - t_1\ ,$$

e

M_3, M_1 são as anomalias médias nos instantes t_3 e t_1.

Pela equação de Kepler podemos escrever:

$$2S_S = a^2\sqrt{1-e^2}\left[2g - e(\operatorname{sen} U_3 - \operatorname{sen} U_1)\right] \qquad (17)$$

onde

$$2g = U_3 - U_1\ .$$

U_3, U_1 são as anomalias excêntricas para os instantes t_3 e t_1.

A área do triângulo será dada por

$$2S_T = |\vec{r}_1|\cdot|\vec{r}_2|\ \operatorname{sen}(V_3 - V_1),$$

mas também pode ser dada através

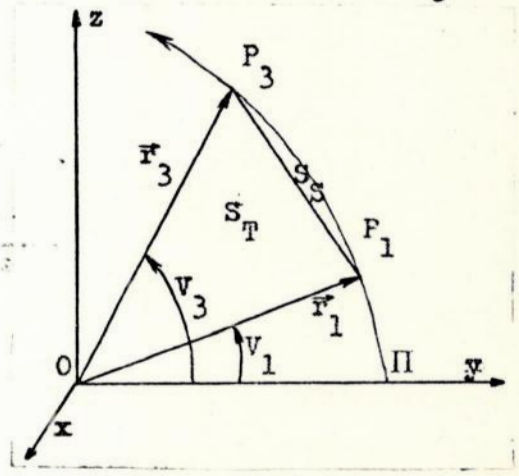


Fig.8

de manipulações puramente algébricas das equações do movimento elítico:

$$2S_T = a^2 \sqrt{1 - e^2} \left[\text{sen } 2g - e(\text{sen } U_3 - \text{sen } U_1) \right] \quad . \qquad (18)$$

Subtraindo-se a equação (18) da equação (17), temos

$$S_S - S_T = \frac{a^2}{2} \sqrt{1 - e^2} (2g - \text{sen } 2g) \quad . \qquad (19)$$

As quantidades a, g, e, são desconhecidas. Mas sabe-se que por meio de combinações convenientes das equações do movimento elítico, as igualdades seguintes são válidas (Danjon, 5 , pag.207):

$$\begin{cases} \sqrt{r_1 \cdot r_3} \ \text{sen} \left(\frac{V_3 - V_1}{2}\right) = a \sqrt{1 - e^2} \ \text{sen } g \ , \\ \sqrt{r_1 \cdot r_3} \ \cos \left(\frac{V_3 - V_1}{2}\right) = \frac{\chi_3}{\sqrt{2}} = a \left[\cos g - e \cos \left(\frac{U_3 - U_1}{2}\right) \right] \\ \frac{1}{2}(r_1 + r_3) = a \left[1 - e \cos g \cos \left(\frac{U_3 + U_1}{2}\right), \right] \end{cases} \qquad (20)$$

onde

$$\chi_3 = \sqrt{2r_1 \cdot r_3} \ \cos \left(\frac{V_3 - V_1}{2}\right) \quad .$$

Multiplicando membro a membro as duas primeiras igualdades ,

obtemos uma nova expressão para a área do triângulo:

$$S_T = \frac{1}{\sqrt{2'}} (\chi_3 \cdot a \sqrt{1 - e^{2'}}) \cdot \text{seng} \, , \qquad (21)$$

Dividindo-se a equação (19) pela equação (21),

$$\gamma - 1 = \frac{a}{\chi^3 \sqrt{2'}} \cdot \frac{(2g - \text{sen } 2g)}{\text{sen } g} \, , \qquad (22)$$

e observando que

$$\rho = a (1 - e^2) = \frac{h^2}{k} = \frac{4S_S^2}{k\theta_3^2} \, .$$

Portanto, a equação (21) toma a forma

$$S_T^2 = \frac{1}{2} \chi_3^2 \, a \, \rho \, \text{sen}^2 g = \frac{2\chi_3^2 \, a \, S_S^2}{k\theta_3^2} \cdot \text{sen}^2 g \, ,$$

e

$$\gamma^2 = \frac{k\theta_3^2}{2\chi_3^2 \, a \, \text{sen}^2 g} \, . \qquad (23)$$

Seja

$$m = \frac{k\theta_3^2}{2\sqrt{2}\chi_3^3} \, .$$

Multiplicando membro a membro as equações (22) e (23), elimi-

namos a :

$$\gamma^3 - \gamma^2 = m\left(\frac{2g - \operatorname{sen} 2g}{\operatorname{sen}^3 g}\right). \qquad (24)$$

Feito isso, permanecem desconhecidas duas quantidades, γ e g. Para estabelecer uma segunda relação entre essas mesmas incógnitas, elimina-se a excentricidade através das duas últimas relações do sistema (20):

$$a \operatorname{sen}^2 g = \chi_3 \sqrt{2} \cdot \left(1 + \operatorname{sen}^2 \frac{g}{2}\right) .$$

Seja

$$\ell = \frac{r_1 + r_3}{2\sqrt{2}\,\chi_3} - \frac{1}{2} .$$

Portanto, a equação (23) pode ser escrita como

$$\gamma^2 = \frac{m}{\ell + \operatorname{sen}^2 \frac{g}{2}} . \qquad (25)$$

Observa-se então, que as equações (24) e (25) formam um sistema de duas equações e duas incógnitas; que são algébricas em relação a γ , mas não o são em relação a g . Essas equações são rigorosas, mas o valor de γ só pode ser obtido por meio de aproximações, as quais levam à equação de Gauss

$$\gamma^3 - \gamma^2 - H\gamma - \frac{H}{9} = 0 , \qquad (26)$$

onde

$$H = \frac{k\theta_3^2}{\chi_3^2\left[r_1 + r_3 + \frac{2\sqrt{2}}{3}\chi_3(1+3\xi)\right]},$$

$$\xi = \frac{2}{35}\rho^2 + \frac{52}{1575}\rho^3 + \ldots .$$

Para uma demonstração veja Apêndice 4.

A expressão H contém ainda uma incógnita, ξ, pois ela depende de ρ, que, por sua vez, depende de $\underline{g}$ através da equação

$$\rho = \operatorname{sen}^2 \frac{g}{2} .$$

Sendo que $\rho \in [0,1]$, então ξ será uma quantidade pequena e, portanto, desprezível numa primeira aproximação para o cálculo de γ. Considerando

$$\rho = \frac{m}{\gamma^2} - 1 ,$$

pode-se assim determinar o valor aproximado de ξ. Desta maneira obtém-se um novo valor de H e, consequentemente, um novo valor para γ. Isso feito, obtém-se um valor mais aproximado para ξ. Essa iteração é feita até obter-se um ξ_n que difira de ξ_{n-1} por uma quantidade arbitrariamente pequena, previamente determinada (vide obs.2, pag.24).

III - CORREÇÕES DIFERENCIAIS

III.1 - Correções diferenciais gaussianas

As correções diferenciais de Gauss são análogas às correções diferenciais de Leuschner, observando-se que estas últimas originariamente foram desenvolvidas especificamente para o método de determinação de órbita de Laplace.

As correções diferenciais de Gauss são baseadas na equação:

$$\rho_2 = Q_1 c_1 - Q_2 + Q_3 c_3 , \tag{27}$$

onde

$$\rho_2 = \Delta_2 ,$$

$$c_1 = \frac{S_2}{S_3} ,$$

$$c_3 = \frac{S_1}{S_3} ,$$

$$Q_1 = A_2 \cdot XT_1 + B_2 \cdot YT_1 + C_2 \cdot ZT_1 ,$$

$$Q_2 = A_2 \cdot XT_2 + B_2 \cdot YT_2 + C_2 \cdot ZT_2 ,$$

$$Q_3 = A_2 \cdot XT_3 + B_2 \cdot YT_3 + C_2 \cdot ZT_3 .$$

Deve-se observar que existem três conjuntos, c_1, c_3 e ρ_2:

1º) Um conjunto "preliminar" (P), determinado pela equação (15) da secção anterior

$$c_1 = c_{1P} = A0 + \frac{B0}{R_2^3} ,$$

$$c_3 = c_{3P} = A1 + \frac{B1}{R_2^3} .$$

2º) Um conjunto "computado" (C), determinado pelas expressões de Gibbs:

$$c_1 = c_{1C} = \frac{\theta_2}{\theta_3} \frac{1 + \frac{\psi_1}{R_1^3}}{1 - \frac{\psi_2}{R_2^3}} ,$$

$$c_3 = c_{3C} = \frac{\theta_1}{\theta_3} \frac{1 + \frac{\psi_3}{R_3^3}}{1 - \frac{\psi_2}{R_2^3}}$$

3º) Um conjunto "objetivo" (O), c_1, c_3, que resultará das correções diferenciais e que estará de acordo, nos limites dos erros permissíveis, com os valores observados de α e δ.

Tem-se, consequentemente, três conjuntos de diferenças ou resíduos:

$$\Delta c_j = c_j - c_{jC} ,$$

$$\Delta' c_j = c_{jP} - c_{jC} , \qquad j = 1,3$$

$$\Delta'' c_j = c_j - c_{jP} ,$$

ou

$$\Delta''c_j = \Delta c_j - \Delta' c_j \,, \tag{28}$$

e expressões análogas em ρ_2 com as condições especiais

$$\rho_2 = \rho_{2C} \,, \tag{29}$$

$$\Delta'\rho_2 = \rho_{2P} - \rho_{2C} = 0 \,,$$

$$\Delta''\rho_2 = \rho_2 - \Delta'\rho_2 = \Delta\rho_2 \,.$$

As diferenças Δc_1 e Δc_3 são as correções desejadas para os valores (C), e estão relacionadas a $\Delta\rho_2$ através das equações

$$\Delta c_j = -B_j''' \Delta\rho_2 \,, \qquad j = 1,3 \,, \tag{30}$$

onde

$$B_1'' = 3(c_1 - A0)(\rho_2 - F)/R_2^2 \,,$$

$$B_3'' = 3(c_3 - A1)(\rho_2 - F)/R_2^2 \,,$$

$$F = XT_2 \cdot \lambda_2 + YT_2 \cdot \mu_2 + ZT_2 \cdot \nu_2 \,.$$

As diferenças $\Delta' c_1$ e $\Delta' c_3$ entram no desenvolvimento das correções diferenciais, pois ambos os valores (P) e (O) de c_1, c_3 e ρ_2 devem satisfazer a equação (27), isto é,

$$\rho_2 = Q_1 c_1 - Q_2 + Q_3 c_3 \tag{31}$$

$$\rho_{2P} = Q_1 c_{1P} - Q_2 + Q_3 c_{3P} \tag{32}$$

Subtraindo (32) de (31),

$$\Delta''\rho_2 = Q_1 \cdot \Delta'' c_1 + Q_3 \Delta'' c_3 .$$

Substituindo a equação (28) temos

$$\Delta''\rho_2 = Q_1(\Delta c_1 - \Delta' c_1) + Q_3(\Delta c_3 - \Delta' c_3) .$$

Pela equação (29),

$$\Delta\rho_2 = Q_1(\Delta c_1 - \Delta' c_1) + Q_3(\Delta c_3 - \Delta' c_3);$$

utilizando-se a equação (30)

$$\Delta\rho_2 = Q_1(- B_1''\Delta\rho_2 - \Delta' c_1) + Q_3(- B_3''\Delta\rho_2 - \Delta' c_3) ,$$

$$\Delta\rho_2(1 + Q_1 B_1'' + Q_3 B_3'') = - Q_1\Delta' c_1 - Q_3\Delta' c_3 . \tag{33}$$

A equação (33) é uma das equações básicas das correções diferenciais de Gauss, e pode ser resolvida diretamente para a incógnita $\Delta\rho_2$. Com $\Delta\rho_2$ obtido deste modo, Δc_1 e Δc_3 são computados a partir da equação (30), obtendo, portanto, os

valores "objetivos" de c_1 e c_3, que serão usados para calcular um novo ρ_2 que, testado com $\Delta\rho_2$, indicará a necessidade ou não de uma nova iteração.

III.2 - Resíduos Gaussianos muito lineares

Consideremos as expressões

$$c_1 = \frac{\theta_2}{\theta_3} \frac{1 + \frac{\psi_1}{R_1^3}}{1 - \frac{\psi_2}{R_2^3}},$$

$$c_3 = \frac{\theta_1}{\theta_3} \frac{1 + \frac{\psi_3}{R_3^3}}{1 - \frac{\psi_2}{R_2^3}}.$$

Todos os termos dessas expressões podem ser obtidos com precisão satisfatória, mas R_2 é necessariamente um valor preliminar, cujo valor computado seria dado pela expressão:

$$\vec{R}_2 = c_1\vec{R}_1 + c_3\vec{R}_3 , \qquad (34)$$

onde

$$\vec{R}_i = (x_i, y_i, z_i) , \qquad i = 1,2,3.$$

A equação (34) pode ser substituída por

$$\vec{R}_{2C} = (c_{1C} + \partial c_1)\vec{R}_1 + (c_{3C} + \partial c_3)\vec{R}_3 , \qquad (35)$$

onde

$$\partial c_j = B_j'''\Delta' R_2 ,$$

$$B_j''' = \frac{3B_2 c_j}{R_2^4(1 - B_2/R_2^3)} , \qquad j = 1,3 .$$

$$B_2 = (\theta_1^2 - 3\theta_1\theta_2 - \theta_2^2)/12 ,$$

e $\Delta' R_2$ é o negativo da correção no valor preliminar de R_2.

Sendo que

$$\vec{R}_{2P} = c_{1P}\vec{R}_1 + c_{3P}\vec{R}_3$$

contém os mesmos valores R_1 e R_3 da expressão (35), podemos escrever

$$\Delta'\vec{R}_2 = \vec{R}_1(\Delta' c_1 - \partial c_1) + \vec{R}_3(\Delta' c_3 - \partial c_3) ,$$

ou, se tomarmos

$$\delta\vec{R}_2 = \vec{R}_1\Delta' c_1 + \vec{R}_3\Delta' c_3 ,$$

e

$$\vec{B}_2 = \vec{R}_1 \cdot B_1''' + \vec{R}_3 \cdot B_3''' ,$$

então,

$$\Delta' \vec{R}_2 = \delta \vec{R}_2 - \vec{B}_2 \cdot \Delta' R_2 . \qquad (36)$$

Fazendo-se o produto escalar de $\vec{R}_2$ com a equação (36), obtemos

$$\Delta' R_2 = (\vec{R}_2 \cdot \delta \vec{R}_2)/(R_2 + \vec{R}_2 \cdot \vec{B}_2) .$$

Alguns autores, como, por exemplo, Herrick (8), sugerem que este método seja adotado como critério de verificação da necessidade de correção nas razões das áreas. Se necessário, introduzem-se as correções diferenciais de Gauss . Já outros autores consideram os dois tipos de correções independentes, e esta foi a orientação seguida neste trabalho.

IV - PERTURBAÇÃO

A equação do movimento de uma partícula de massa m_i, sob a atração predominante da massa m_o, mas também influenciada por outras massas m_j, $j = 1,2,\ldots n-1$, $j \neq i$, e por um campo externo $\vec{E}$, é

$$\ddot{\vec{r}}_i - K(m_o + m_i)F(r_i)\vec{\mu}_{oi} = \nabla_i R_i + \vec{E}_i - \vec{E}_o \,, \qquad (37)$$

onde

$$\nabla_i = \frac{\partial}{\partial \xi_i}\vec{i} + \frac{\partial}{\partial \eta_i}\vec{j} + \frac{\partial}{\partial \zeta_i}\vec{k} \,,$$

$$R_i = -K \sum_{\substack{j=1 \\ j\neq i}}^{n-1} m_j \left[P_{ij} + \frac{F(r_j)}{r_j}(\xi_i \xi_j + \eta_i \eta_j + \zeta_i \zeta_j) \right] \,,$$

considerando o sistema retangular $O\xi\eta\zeta$ com origem em m_o, e P_{ij} satisfazendo a equação

$$\frac{dP_{ij}}{d\rho_{ij}} = F(\rho_{ij}) \,, \qquad \text{para cada } j.$$

Se nos restringirmos somente ao caso em que, para cada j

$$F(\rho_{ij}) = \rho_{ij}^{-2} \,,$$

segue que

$$F(r_j) = r^{-2} ,$$

e, portanto,

$$P_{ij} = \frac{1}{\rho_{ij}} .$$

Se abandonarmos o índice i, podemos escrever

$$R = K \sum_{j=1}^{n-2} m_j \left(\frac{1}{\rho_j} - \frac{\xi\xi_j + \eta\eta_j + \zeta\zeta_j}{r_j^3}\right).$$

Considerando-se que exista uma só partícula perturbadora ($j = 1$), $m_1 = m'$ (ver fig. 9), tem-se

$$R = Km'\left(\frac{1}{\rho} - \frac{\xi\xi' + \eta\eta' + \zeta\zeta'}{r'^3}\right) ,$$

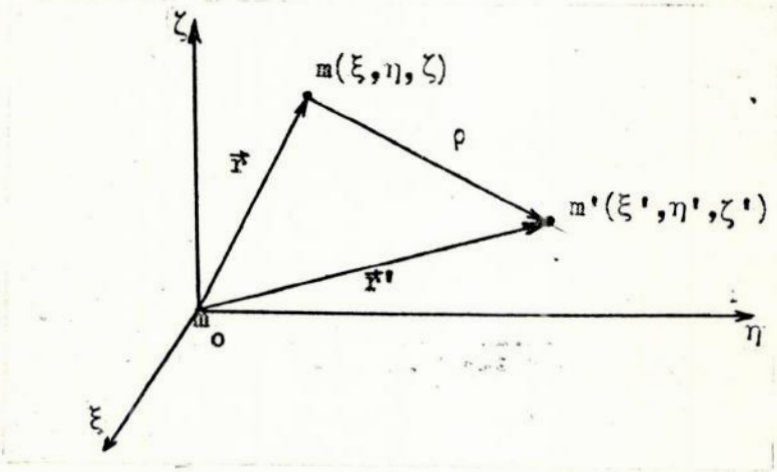


Fig.9

onde $\xi, \eta, \zeta, \xi', \eta', \zeta'$ são coordenadas retangulares referenciadas a qualquer sistema retangular com eixos fixos no espaço, e origem no primário; r e r' são as distâncias ao primário, e

$$\rho^2 = (\xi - \xi')^2 + (\eta - \eta')^2 + (\zeta - \zeta')^2$$

é a distância entre m e m'.

O termo $\frac{1}{\rho}$ é conhecido como a parte principal da função de perturbação. O outro termo é chamado a parte indireta; ela expressa a ação do planeta perturbador sobre o Sol. Surge devido ao fato de que se utilizarmos coordenadas heliocêntricas este termo se anularia se tomassemos como origem o centro de massa do sistema Sol - planeta perturbador.

Da equação (37) obtem-se, não considerando os campos externos,

$$\ddot{\vec{r}} + \frac{\mu \vec{r}}{r^3} = \nabla R.$$

Fazendo-se $\nabla R = \vec{F}$ temos, em função das componentes de $\vec{F}$ ao longo dos eixos ξ, η, ζ,

$$\begin{cases} \ddot{\xi} + \dfrac{\mu\xi}{r^3} = F_\xi , \\ \ddot{\eta} + \dfrac{\mu\eta}{r^3} = F_\eta , \\ \ddot{\zeta} + \dfrac{\mu\zeta}{r^3} = F_\zeta . \end{cases} \qquad (38)$$

As substituições $\dot{\xi} = v_\xi$, $\dot{\eta} = v_\eta$ e $\dot{\zeta} = v_\zeta$ transformarão as três equações diferenciais de segunda ordem em seis equações diferenciais de primeira ordem:

$$\dot{\xi} = v_\xi \,,$$

$$\dot{\eta} = v_\eta \,,$$

$$\dot{\zeta} = v_\zeta \,,$$

$$\dot{v}_\xi + \frac{\mu\xi}{r^3} = F_\xi \,,$$

$$\dot{v}_\eta + \frac{\mu\eta}{r^3} = F_\eta \,,$$

$$\dot{v}_\zeta + \frac{\mu\zeta}{r^3} = F_\zeta \,.$$

Integrando essas equações, para um pequeno intervalo de tempo, obteremos novos valores para as velocidades e para as coordenadas. Nesta integração consideraremos constantes os elementos orbitais, o que pode ser justificado teoricamente pois o intervalo é pequeno (Herrick, 8).

Esses novos valores, são da órbita verdadeira, isto é, perturbada, em um instante t_1. Considera-se agora, uma órbita de dois corpos e calcula-se novos elementos orbitais. Esses elementos caracterizam uma órbita chamada "órbita osculadora" no ponto em questão. O corpo celeste em sua órbita perturbada tem neste instante, t_1, as mesmas coordenadas e a mesma velocidade que teria se estivesse se movendo na órbita osculadora neste mesmo instante.

V - MÉTODO DE CÁICULO

V.1 - Esquema Computacional

Com base na teoria formulada nos itens anteriores, desenvolvemos um programa em FORTRAN IV cujo objetivo é calcular, a partir de três observações astronômicas, uma órbita preliminar tal que seus elementos possam ser usados como inicializador ("starter") numa integração numérica, por exemplo, das equações planetárias de Lagrange. O programa foi processado no Burroughs B-6700 do Núcleo de Computação Eletrônica da Universidade Federal do Rio de Janeiro.

O programa seguiu, basicamente, o seguinte esquema

A partir de três conjuntos de observações obtem-se, pelo método de Gauss, uma ajustante a estes valores. Essa ajustante é uma elipse (em geral, uma cônica) pois, inicialmente,

o problema foi tratado como o de dois corpos. A ajustante, se necessário, é otimizada através de correções diferenciais.

Introduz-se a seguir as forças perturbadoras em questão e obtem-se equações diferenciais em coordenadas retangulares, que são integradas para intervalos pequenos, especificamente, um dia antes e um dia após cada data de observação. A escolha deste intervalo decorre de razões teóricas e experimentais: sob o aspecto teórico, a posição dos planetas perturbadores é considerada como invariável ao longo do intervalo de integração e coincidente com os instantes das observações ; sob o aspecto experimental, a escolha de um dia foi motivada pela apresentação das efemérides planetárias que listam posições em intervalos de um dia. Certamente, a hipótese da invariância é tanto mais rigorosa quanto menor for o movimento médio do planeta, o que não é crítico no caso de Jupiter e Saturno, cujas perturbações são as mais importantes no problema de asteróides e cometas. Finalmente, esta invariância é uma exigência do método adotado para as integrações das equações do movimento. Pesquisas posteriores deverão examinar com maiores detalhes os erros introduzidos em tais aproximações, bem como alterações que se fizerem necessárias a fim de que as perturbações com elevados movimentos médios possam ser incorporadas.

A partir dessas integrações calcula-se uma osculadora para cada data de observação, obtendo-se assim novos valores para α e δ nas datas de observação. Com esses novos valores de α e δ, que incorporam perturbações, retorna-se ao método de Gauss, obtendo-se uma nova ajustante para as datas de observação . É bom frisar que esses α e δ desempenham, no método de Gauss, a mesma função que os α e δ obtidos da observação, isto é, seriam novos valores observados. Em princípio, a escolha da curva que se ajustaria aos conjuntos de AR

e DEC é arbitrária. Escolhemos cônicas apenas porque isto nos permite a realimentação do programa a partir de suas fases iniciais, porém técnicas como as dos mínimos quadrados em aproximações polinomiais poderiam fornecer arcos da órbita real, possivelmente com precisões superiores às elipses. Neste projeto admiteremos, por hipótese, que as cônicas, fornecerão resultados aceitáveis, pelo menos em pequenos intervalos de tempo. A validade da hipótese será testada por comparação com outros resultados (ver secção VI).

V.1.1 - Definição das variáveis do programa

Nas variáveis abaixo, tem-se que I = 1,2,3.

TT(I)	- instantes de observação
ALFA(I), DELTA(I)	- ascensão reta e declinação das observações 1, 2 e 3
XS(I), YS(I), ZS(I)	- coordenadas geocêntricas do Sol
TS(I)	- tempo sideral local das observações
K	- constante gaussiana da gravitação
LAT, LOM	- latitude geocêntrica e longitude geocêntrica do local das observações
E	- obliquidade da eclítica para a época das observações
ORO	- distância geocêntrica do lugar das observações
XT(I), YT(I), ZT(I)	- coordenadas topocêtricas do Sol
LA(I), MI(I), NI(I)	- cossenos diretores das três observações
A(I), B(I), C(I)	- quantidades auxiliares do método de Gauss definidas no item II, pag. 19

ED	- determinante da matriz definida no método de Gauss, pag. 19
TE(I)	- intervalos de tempo entre as observações
D(I)	- distâncias geocêntricas do corpo considerado
SI	- razão entre as áreas S_1 e S_3
SII	- razão entre as áreas S_2 e S_3
SIII	- área S_3
L, M, N	- quantidades auxiliares do método de Gauss, definidas na pag.18
R(I)	- distâncias do Sol ao corpo nos dias das observações
F1, F2, F3	- quantidades utilizadas nas fórmulas de Gibbs
XA(I), YA(I), ZA(I)	- coordenadas heliocêntricas do corpo
QU(I)	- quantidades auxiliares do método de Gauss, (I), pag. 27
H(I)	- quantidades auxiliares do método de Gauss, pag. 30
GA(I)	- razões entre os setores curvilíneos e as áreas triangulares
LE(I)	- quantidades auxiliares do método de Gauss, (I), pag. 29
MA(I)	- quantidades auxiliares do método de Gauss, (I), pag. 28
RO(I)	- quantidades auxiliares do método de Gauss, (I), pag. 30
QS(I)	- quantidades auxiliares do método de Gauss, (I), pag. 30
XH(I), YH(I), XH(I)	- coordenadas eclíticas heliocêntricas do corpo
P	- parâmetro definido a partir da constante das áreas
CI, SE	- quantidades usadas para definir os cossenos e senos, respectivamente, dos ângulos cujos quadrantes desejamos calcular

II, IIG	- ângulo que define a inclinação da órbita
OMEGA, OMEGAG	- longitude do nodo ascendente
V(1), V(3)	- anomalias verdadeiras dos instantes t_1 e t_3
EX	- excentricidade
AE	- semi-eixo maior
NM	- movimento médio
WP, WPG	- argumento do periçentro
AM, AMG	- anomalia média
U(I)	- anomalias excêntricas das três posições
TZ	- tempo tomado como referência
XF(I), YF(I), ZF(I)	- coordenadas retangulares heliocêntricas do corpo
MF(I)	- anomalias médias finais das três observações
UF(I)	- anomalias excêntricas finais das observações
QSI(I), ETA(I), ZET(I)	- coordenadas retangulares geocêntricas do corpo
AF(I), DF(I)	- ascensões retas e declinações calculadas
ERRA(I), ERRD(I)	- erros relativos entre as ascensões retas e declinações, respectivamente, observadas e calculadas
DIX	- data para a qual deseja-se interpolar
DIO, TA(I)	- data imediatamente anterior a DIX
DII, TD(I)	- data imediatamente posterior a DIX
ALZEP(I), ALPI(I)	- ascensões retas correspondentes às datas DIO e DII, respectivamente do planeta perturbador

DEPZE(I), DEPI(I) - declinações correspondentes às datas DIO e DIX, respectivamente, do planeta perturbador

ALPMI(I), ALPII(I) - ascensões retas imediatamente anteriores e posteriores a ALZEP(I) e ALPI(I), respectivamente

DEPMI(I), DEPII(I) - declinações imediatamente anteriores e posteriores a DEPZE(I) e DEPI(I), respectivamente

XP(I), YP(I), ZP(I) - coordenadas retangulares geocêntricas do planeta perturbador

ALFAP(I), DELTAP(I) - ascensões retas e declinações do planeta perturbador para as datas de observação

ROP(I) - distâncias geocêntricas do planeta perturbador

FX(I), FY(I), FZ(I) - coordenadas das forças de perturbação para as datas das observações

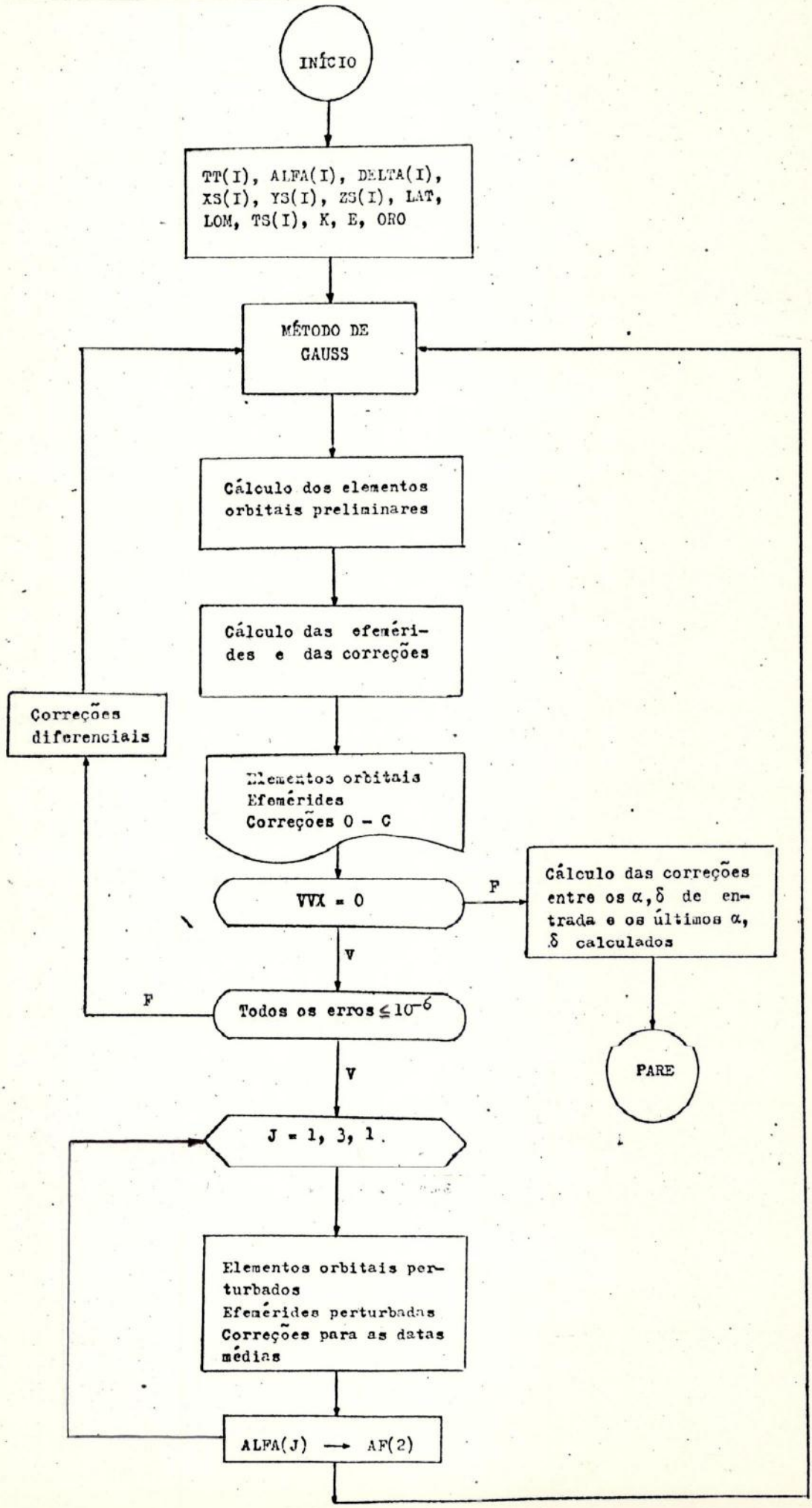
VX(I), VY(I), VZ(I) - componentes das velocidades do corpo

Y(1), Y(2), Y(3) - coordenadas equatoriais heliocêntricas perturbadas para um dia antes e um dia depois das datas de observação

Y(4), Y(5), Y(6) - componentes das velocidades perturbadas para um dia antes e um dia depois das datas de observação

ERROA, ERROD - diferenças entre os α e δ observados e os α e δ calculados pelo método de Gauss depois de ter sido efetuada a perturbação

V.1.2 - Fluxograma



V.2 - Método de Integração Numérica de Bulirsch - Stoer (B-S)

O método B-S é utilizado para resolver um sistema de equações diferenciais ordinárias da forma

$$y'_i = f_i(x, y_1, y_2, \ldots, y_n) , \quad i = 1, \ldots, n$$

sendo dados valores iniciais.

Essencialmente é baseado em extrapolações, pois que a extrapolação é um meio poderoso na aceleração de convergência de soluções. Tem sido provado que extrapolações baseadas em polinomiais interpoladoras ou funções racionais, fornecem soluções bastante precisas; entretanto, a experiência tem mostrado a superioridade da extrapolação por funções racionais sobre a extrapolação polinomial.

Uma comparação entre o método B-S e os métodos de Runge Kutta, Adams - Moulton - Bashforth, e o de extrapolação com polinomiais baseadas na regra do ponto médio modificada, mostrou que os resultados são muito mais precisos e o número de operações necessárias para obte-los, é muito menor quando se faz uso do método B-S. Além do mais, este é mais fácil de ser programado, pois não é necessário computar valores iniciais especiais e a ordem de aproximação não é prefixada, podendo ser modificada de acordo com o problema em questão. E para finalizar, não é preciso nenhuma preparação especial das equações diferenciais a serem integradas.

Para uma melhor visualização das vantagens deste método sobre outros citados acima, sugerimos um exame dos exemplos apresentados no artigo "Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods" do anexo 1, em

especial do exemplo 3, pois refere-se a um problema de Mecânica Celeste. Remetemos também ao Anexo 1 o leitor interessado em pesquisar os aspectos teóricos e práticos do método em questão.

VI - ANÁLISE DOS RESULTADOS

VI.1 - Introdução

O programa desenvolvido foi testado para três asteróides, 683 Lanzia (1909 HC), 1342 Brabantia (1935 CV) e Ceres. Em seguida, foi testado para dois cometas, 1977 HB e Kohler, com dados obtidos dos telegramas do I.A.U. (União Astronômica Internacional).

Os resultados de cada um serão analisados separadamente, fazendo parte da análise comparações com resultados obtidos por outros autores.

As coordenadas retangulares geocêntricas do Sol, as coordenadas esféricas geocêntricas de Jupiter, as coordenadas dos observatórios (quando necessárias), a obliquidade média da eclítica, a redução para tempo sideral e, finalmente, as correções para a precessão, foram obtidas de "Nautical Almanac". A constante gravitacional adotada, k, também foi obtida da mesma publicação, edição de 1978, e o seu valor é

$$k = 0{,}0002959122$$

As ascensões retas e declinações topocêntricas de entrada foram reduzidas para radianos e, quando necessário, para a data de observação.

As anomalias médias, os argumentos do pericentro e as inclinações são referenciadas à época de observação.

Alguns problemas que surgiram quando da determinação das referidas órbitas serão examinados na secção VII.

VI.2 - Asteróide 683 Lanzia

Os dados de entrada deste asteróide foram obtidos de Herrick (8, pag.384):

DATAS	7,8205	26,7480	48,626
α_{1910}	$3^h\ 50^m\ 24^s,3$	$3^h\ 13^m\ 3^s,0$	$4^h\ 54^m\ 19^s,5$
δ_{1910}	25° 11' 10",5	22° 29' 31",3	20° 14' 51",9

As datas t_1, t_2 e t_3 são dadas em tempo médio de Greenwich , referenciadas a 1º de Novembro de 1910.

Os resultados por nós obtidos, utilizando correções diferenciais gaussianas, estão nas tabelas nº 1 e 2, onde a anomalia média, o argumento do pericentro, a inclinação, o nodo ascendente, DELTAF e ERRD são dados em graus e frações; o semi-eixo maior XF, YF, ZF, QSI, ETA, ZET, em unidades astronômicas; o movimento médio em radiano por dia solar médio e , finalmente, ALFA e ERRA em horas e frações.

TABELA 1 - ELEMENTOS ORBITAIS

ANOMALIA MEDIA=221°.181650
SEMI EIXO MAIOR= 3.121299 U.A.
MOVIMENTO MEDIO= 0.178731 rad/dsm
EXCENTRICIDADE= 0.048770
ARGUMENTO DO PERICENTRO= 266°.870941
INCLINACAO= 10°.493860
NODO ASCENDENTE=260°.681719

TABELA 2 – EFEMÉRIDES E CORREÇÕES

EFEMERIDES

XF	YF	ZF
2.8660845556	0.7833286532	1.2998278071
2.7954234713	0.9472833256	1.3339391821
2.7029046961	1.1272947316	1.3681763459

QSI	ETA	ZET
2.1660158556	0.1453387532	1.0209067071
2.3647327713	0.1329337256	0.9806646821
2.6400675961	0.2265849316	0.9774346459

CORRECOES

ALFA	ERRA	DELTAF	ERRD
0.2560056753	-0.0000001492	25.1862449762	-0.0000000593
0.2145001637	-0.0000001586	22.4920258062	-0.0000003294
0.3270275555	0.0000001743	20.2477489530	0.0000008401

$\cos \delta_2 \Delta\alpha_2 = -0",0085618$

$\Delta\delta_2 = -0",6705766$

Os resultados obtidos por Herrick (8, pag.399-400) , através do método de Laplace, foram:

SEMI EIXO MAIOR = 3,1212221 U.A.

MOVIMENTO MÉDIO = 0,1813481 rad/dsm

EXCENTRICIDADE = 0,04888372

ARGUMENTO DO PERICENTRO = $267^{\circ},05142$

INCLINAÇÃO = $18^{\circ},49806$

NODO ASCENDENTE = $260^{\circ},65696$

$\cos \delta_2 \Delta\alpha_2 = -0",35$

$\Delta\delta_2 = -0",09$

Usando-se os resíduos gaussianos muito lineares obtivemos

TABELA 3 - ELEMENTOS ORBITAIS, EFEMÉRIDES E CORREÇÕES

ANOMALIA MEDIA=221°.181650

SEMI EIXO MAIOR= 3.121299 U.A.

MOVIMENTO MEDIO= 0.178731 rad/dsm

EXCENTRICIDADE= 0.048770

ARGUMENTO DO PERICENTRO= 266°.876941

INCLINACAO= 18°.493880

NODO ASCENDENTE=260°.681719

EFEMERIDES

	XF	YF	ZF
1	2.8660845556	0.7883286532	1.2998278071
2	2.7954234713	0.9472833256	1.3339391621
3	2.7029046961	1.1272947316	1.3681763459

QSI	ETA	ZET
2.1660158556	0.1453387532	1.0209067071
2.3647327713	0.1329337256	0.9806646821
2.6400675961	0.2265849316	0.9774346459

CORRECOES

ALFA	ERRA	DELTAF	ERRD
0.2560056753	-0.0000001492	25.1862449762	-0.0000000593
0.2145001637	-0.0000001586	22.4920258062	-0.0000003294
0.3270275559	0.0000001743	20.2477469530	0.0000008401

Da tabela nº 3, tem-se

$\cos \delta_2 \Delta\alpha_2 = - 0",0085641,$

enquanto que Herrick (8, pag.399-400) fornece

$\cos \delta_2 \Delta\alpha_2 = - 0",35$

Neste exemplo não foi considerada a perturbação planetária; ele serviu para testar o método de Gauss, que foi a base do programa.

VI.3 - Asteróide 1342 Brabantia

Os dados de entrada obtidos de Danjon (5, pag.97) foram

DATAS	13,05561	28,90839	37,96207
$\alpha_{1935,0}$	$10^h\ 25^m\ 28^s,39$	$10^h\ 1^m\ 58^s,85$	$9^h\ 49^m\ 14^s,92$
$\delta_{1935,0}$	$- 5^o\ 27'\ 18",7$	$- 8^o\ 39'\ 01",1$	$- 9^o\ 57'\ 48",0$

As datas são apresentadas em TU (Tempo Universal) e referenciadas a 1º de Fevereiro de 1935.

A fim de possibilitar uma comparação com os resultados dados por Danjon (5, pag.102), a tabela nº 4 foi obtida sem nenhuma correção diferencial e sem perturbação.

TABELA 4 - ELEMENTOS ORBITAIS. EFEMÉRIDES E CORREÇÕES

ANOMALIA MEDIA=345°.076405
SEMI EIXO MAIOR= 2.293485 U.A.
MOVIMENTO MEDIO= 0.283766 rad/dsm
EXCENTRICIDADE= 0.202014
ARGUMENTO DO PERICENTRO= 227°.085925
INCLINACAO= 21°.063110
NODO ASCENDENTE=312°.988160

EFEMERIDES

	XF	YF	ZF
1	-1.6271846326	0.9048992217	0.1472008677
2	-1.6921141171	0.7527553936	0.0051719723
3	-1.7203854151	0.6618799061	-0.0760655003

	QSI	ETA	ZET
	-0.8342996491	0.3650462331	-0.0669652597
	-0.7646748335	0.4327161425	-0.1336687591
	-0.7472916284	0.4795267608	-0.1551765277

CORRECOES

	ALFA	ERRA	DELTAF	ERRD
1	10.4245539162	-0.0000011339	-5.4552369166	0.0000424944
2	10.0330157306	-0.0000018318	-8.6503687812	0.0000632396
3	9.8208179181	-0.0000066035	-9.9133242451	-0.0000091163

	1	2	3
ERRA	- 0^s,00408204	- 0^s,00659448	- 0^s,0244926
ERRD	0",15297984	0",22766256	- 0",03281868

Danjon (5, pag.225-228) apresenta os seguintes elementos orbitais:

ANOMALIA MÉDIA = 345°,823

SEMI EIXO MAIOR = 2,29265 U.A.

MOVIMENTO MÉDIO = 0,283921 rad/smd

EXCENTRICIDADE = 0,20202

ARGUMENTO DO PERICENTRO = 227°,976

INCLINAÇÃO = 21°,048 1935,0

NODO ASCENDENTE = 312°,974

ERRA	$-0^s,01$	$0^s,01$	$-0^s,01$
ERRD	0",1	- 0",1	- 0",2

Neste caso particular o processo computacional não apresentou convergência em ERRA e ERRD por nenhuma das técnicas de correções diferenciais. Tentamos então obter a convergência calculando, primeiro, a perturbação devida a Jupiter, e, em seguida, as correções diferenciais. Mas a convergência, para 10^{-6}, não foi obtida, e os erros se mantiveram na ordem de 10^{-4}.

Tendo em vista o exposto acima, diminuimos o critério de convergência, neste caso específico, para 10^{-4} e, considerando a perturbação devida a Júpiter, obtivemos a seguinte tabela:

TABELA 5 – ELEMENTOS ORBITAIS, EFEMÉRIDES, CORREÇÕES E CORREÇÕES FINAIS

ANOMALIA MEDIA=345°.985206

SEMI EIXO MAIOR= 2.295557U.A.

MOVIMENTO MEDIO= 0.283382 rad/dsm

EXCENTRICIDADE= 0.202108

ARGUMENTO DO PERICENTRO= 227°.712931

INCLINACAO= 21°.096537

NODO ASCENDENTE=315°.011960

EFEMERIDES

	XF	YF	ZF
1	-1.6281756503	0.9053369205	0.1470663950
2	-1.6931266246	0.7533629767	0.0049633572
3	-1.7214192604	0.6625306746	-0.0763149240

	QSI	ETA	ZET
	-0.6352906667	0.3654839319	-0.0870660776
	-0.7656873409	0.4332837196	-0.1338397154
	-0.7463254737	0.4801777294	-0.1553862966

CORRECOES

	ALFA	ERRA	DELTAF	ERRD
1	10.4245385892	-0.0000011291	-5.4546220660	0.0000421241
2	10.0330360717	-0.0000018134	-8.6500105832	0.0000630154
3	9.8208628809	-0.0000067504	-9.9129653231	-0.0000087612

CORRECOES FINAIS

	ERROA	ERROD
1	0.0000141915	-0.0003723562
2	-0.0000221744	-0.0002949584
3	-0.0000517679	-0.0003680382

VI.4 - Ceres

Os dados de entrada, bem como as coordenadas retangulares geocêntricas do Sol e as coordenadas esféricas geocêntricas de Júpiter, foram tiradas do "Nautical Almanac" para 1978

DATAS	1,0	16,0	31,0
$\alpha_{1950,0}$	$18^h\ 57^m\ 16^s,91$	$18^h\ 10^m\ 47^s,03$	$19^h\ 27^m\ 41^s,32$
$\delta_{1950,0}$	$-\ 30^o\ 42'\ 46",0$	$-\ 30^o\ 12'\ 08",7$	$-\ 29^o\ 31'\ 50",2$

As datas estão em TU (Tempo Universal), e são referenciadas a 1º de Outubro de 1978.

Por meio de correções diferenciais gaussianas, e sem considerar a perturbação planetária, obtivemos os resultados das tabelas nº 6 e 7.

TABELA 6 - ELEMENTOS ORBITAIS

ANOMALIA MEDIA= 6.9448270
SEMI EIXO MAIOR= 2.773668 U.A.
MOVIMENTO MEDIO= 0.213365 rad/dsm
EXCENTRICIDADE= 0.077434
ARGUMENTO DO PERICENTRO= 74º623442
INCLINACAO= 10º597362
NODO ASCENDENTE= 80º102763

TABELA 7 - EFEMÉRIDES E CORREÇÕES

EFEMERIDES

XF	YF	ZF
-1.7699683772	-1.9567755166	-0.5636979581
-1.6582810851	-2.0484072221	-0.6297357823
-1.5410425763	-2.1331506434	-0.6936657279

QSI	ETA	ZET
-2.7629224772	-2.0737946166	-0.6144344581
-2.5816836851	-2.3930029221	-0.7791482323
-2.3331402763	-2.6821976434	-0.9317232279

CORRECCES

ALFA	ERRA	DELTAF	ERRD
14.4594108945	0.0000002177	-10.0851104994	-0.0000006109
14.8551944694	-0.0000000252	-12.4805555362	-0.0000000184
15.2654166885	-0.0000000208	-14.6863666406	-0.0000002464

Efetuando-se a perturbação, devido a Júpiter, obtivemos como resultados finais os apresentados nas tabelas nº 8 e 9.

TABELA 8 - ELEMENTOS ORBITAIS

ANOMALIA MEDIA= 68°612206
SEMI EIXO MAIOR= 2.772971 U.A.
MOVIMENTO MEDIO= 0.213445 rad/dsm
EXCENTRICIDADE= 0.077529
ARGUMENTO DO PERICENTRO= 74°457578
INCLINACAO= 10°598334
NODO ASCENDENTE= 80°101090

TABELA 9 - EFEMÉRIDES, CORREÇÕES E CORREÇÕES FINAIS

EFEMERIDES

XF	YF	ZF
-1.7697955988	-1.9566503097	-0.5636607491
-1.6581425736	-2.0482826653	-0.6296949444
-1.5409351596	-2.1330697690	-0.6936273743

QSI	ETA	ZET
-2.7627496988	-2.0736694097	-0.6143972491
-2.5815451736	-2.3928783653	-0.7791074444
-2.3330328596	-2.6820767690	-0.9316808743

CORRECCES

ALFA	ERRA	DELTAF	ERRD
14.4594148539	0.0000002196	-10.0851223390	-0.0000006184
14.6551975189	-0.0000000266	-12.4305615599	-0.0000000059
15.2654185311	-0.0000000222	-14.6863890299	-0.0000002368

CORRECCES FINAIS

ERRDA	ERRDD
-0.0000037417	0.0000112286
-0.0000030747	0.0000060053
-0.0000018634	0.0000001429

As efemérides planetárias ("Ephemerides of Minor Planets for 1978") forneceram os seguintes elementos orbitais para Ceres:

ANOMALIA MÉDIA = 348°,581

SEMI EIXO MAIOR = 2,7674 U.A.

ARGUMENTO DO PERICENTRO = 71°,335

INCLINAÇÃO = 10°,610 1950,0

NODO ASCENDENTE = 80°,486

Sendo que os dados de entrada foram tirados do "Nautical Almanac", eles já estão corrigidos da perturbação de Júpiter, e dos outros oito planetas. Portanto, o que fizemos foi calcular uma nova perturbação devido a Júpiter. Isto implica em dizer que calculamos a perturbação sobre Ceres devido a um planeta de massa equivalente a duas vezes a massa de Júpiter, o que foi feito a fim de testar o grau de influência de Júpiter sobre Ceres. Como pode ser observado, esta influência é pequena, pelo menos na configuração estudada.

VI.5 - Cometa Kohler (1977 m)

Os dados de entrada foram tirados das efemérides apresentadas nos Telegramas da I.U.A., circular nº3205. As coordenadas retangulares geocêntricas do Sol e as coordenadas esféricas geocêntricas de Júpiter foram obtidas do "Nautical Almanac" para 1978.

DATAS	1,0	11,0	21,0
$\alpha_{1950,0}$	$5^h\ 57^m,95$	$6^h\ 13^m,23$	$6^h\ 27^m,83$
$\delta_{1950.0}$	$-\ 14^\circ\ 28',4$	$14^\circ\ 21',9$	$-\ 14^\circ\ 24',7$

As datas estão em TU (Tempo Universal) e são referenciadas a 1º de Junho de 1978.

Não considerando correções diferenciais e a perturbação planetária, obtivemos os dados da tabela nº 10.

TABELA 10 - ELEMENTOS ORBITAIS, EFEMÉRIDES E CORREÇÕES

ANOMALIA MEDIA= 0°.205978
SEMI EIXO MAIOR=100.685288U.A.
MOVIMENTO MEDIO= 0.000976 rad/dsm
EXCENTRICIDADE= 0.990523
ARGUMENTO DO PERICENTRO= 161°.972297
INCLINACAO= 48°.782448
NODO ASCENDENTE=181°.728280

EFEMERIDES

	XF	YF	ZF
1	-0.3176849637	2.7688132334	-1.3184068973
2	-0.4057148376	2.8618820052	-1.3646647639
3	-0.4933710433	2.9523493992	-1.4096820625

QSI	ETA	ZET
0.0330771363	3.6417086334	-0.9399128973
-0.2178394376	3.7773398052	-0.9677159639
-0.4736072433	3.8844844992	-1.0054956625

CORRECOES

ALFA	ERRA	DELTAF	ERRD
5.9653069777	0.0005263009	-14.4714498624	-0.0018823023
6.2200397045	0.0004603269	-14.3466911892	-0.0016445029
6.4634235497	0.0004097475	-14.4102176852	-0.0014470319

Incluindo correções diferenciais gaussianas, as tabelas a seguir foram obtidas.

TABELA 11 - ELEMENTOS ORBITAIS

ANOMALIA MEDIA= 0.062938
SEMI EIXO MAIOR=222.133756
MOVIMENTO MEDIO= 0.000298
EXCENTRICIDADE= 0.995676
ARGUMENTO DO PERICENTRO= 162.441167
INCLINACAO= 48.778255
NODO ASCENDENTE=181.706492

TABELA 12 - EFEMÉRIDES E CORREÇÕES

		EFEMERIDES
XF	YF	ZF
-0.3165092562	2.7691066232	-1.3181864019
-0.4046411366	2.8623353023	-1.3644866949
-0.4924263852	2.9530020475	-1.4095660890
QSI	ETA	ZET
0.0342528438	3.6420020232	-0.9396924019
-0.2167657366	3.7777931023	-0.9675378949
-0.4726625852	3.8851371475	-1.0053796890

CORRECOES

ALFA	ERRA	DELTAF	ERRD
5.9640767998	0.0017564787	-14.4670396217	-0.0062925430
6.2189313237	0.0015687078	-14.3427395889	-0.0055961032
6.4624312833	0.0014020139	-14.4067419345	-0.0049227826

O telegrama apresenta os seguintes elementos orbitais:

EXCENTRICIDADE = 0,999502

ARGUMENTO DO PERICENTRO = 163°,4880

NODO ASCENDENTE = 181°,8240 1950,0

INCLINAÇÃO = 48°,7181

Como pode ser notado examinando-se as tabelas acima, as correções diferenciais, apesar de melhorarem os elementos orbitais, pioraram os erros em alfa e delta, não levando à convergência desejada no nível 10^{-6}.

VI.6 - Cometa 1977 HB

Os dados de entrada foram obtidos a partir das efemérides dos Telegramas do I.A.U., circular nº3159.

As coordenadas retangulares geocêntricas do Sol e as coordenadas esféricas geocêntricas de Júpiter foram obtidas do "Nautical Almanac" para 1978.

DATAS	1,0	11,0	21,0
$\alpha_{1950,0}$	$23^h\ 17^m,16$	$23^h\ 55^m,0$	$0^h\ 34^m,76$
$\delta_{1950,0}$	$-\ 3^\circ\ 50',0$	$2^\circ\ 5',5$	$8^\circ\ 22',3$

As datas estão em TU (Tempo Universal) e são referenciadas a 1º de Fevereiro de 1978. Obtivemos como resultado:

INCLINAÇÃO = $3^\circ,64872983$

NODO ASCENDENTE = $320^\circ,59351595$

EXCENTRICIDADE > 1

enquanto que o telegrama fornece

INCLINAÇÃO = $9^\circ,4231$

NODO ASCENDENTE = $32^\circ,7856$

EXCENTRICIDADE = 0,3449508

A causa da excentricidade ser maior que um foi devido ao fato de obtermos com o programa distância geocêntrica negativa na primeira aproximação calculada pela subrotina RAIO. Este problema será retomado na próxima seção.

VII - CONCLUSÕES E PERSPECTIVAS FUTURAS

VII.1 - Preparação de dados

A medida que as diferentes órbitas foram sendo computadas, verificamos que para aplicações sistemáticas do esquema proposto (ver pag.48), se faria necessário a montagem de uma subrotina que processasse as seguintes conversões:

a) tempo universal em tempo das efemérides - esta conversão é necessária a fim de compatibilizar as unidades de tempo empregadas nas observações (TU) e nas efemérides do Sol e dos planetas (TE);

b) coordenadas aparentes em médias - isto é necessário, uma vez que para fins computacionais e na hipotése de publicações, devemos referir as coordenadas observadas (aparentes) a algum equinócio fixo, usualmente 1950,0 , bem como na transformação das coordenadas dos planetas perturbadores para o mesmo equinócio, se estas coordenadas são retiradas do "Nautical Almanac". Somos de opinião que a conversão de coordenadas aparentes em médias, bem como a passagem de um equinócio a outro, deva ser feita a partir de fórmulas rigorosas e não das indicadas no "Nautical Almanac", que são aproximadas (1978, pag. 9), dado que em grandes programas que, como no presente caso, operam com base em iterações, pequenos erros nos valores de entrada podem se converter em erros apreciáveis ao final do processamento;

c) subrotina SOL - esta subrotina que realiza interpolações besselianas nas coordenadas do Sol e dos planetas perturbadores, a fim de obter valores correspondentes às datas de observação, necessita ser modificada com o objetivo de in-

cluir diferenças de ordens superiores às segundas, de tal maneira que precisões da ordem de 10^{-7} nestas coordenadas sejam asseguradas. Idealisticamente, o programa deveria conter subrotinas que computassem, a partir das teorias planetárias, as coordenadas do Sol e dos planetas, quando então as coordenadas destes corpos deixariam de ser dados de entrada. Do ponto de vista prático, contudo, isto implicaria em um tal acréscimo do tempo de processamento (significaria, por exemplo, resolver um conjunto de 81 equações planetárias de Lagrange - equações diferenciais ordinárias de 1ª ordem não lineares), que só se justificaria caso fosse economicamente viável a montagem de uma central de cálculo de órbitas;

d) correções de paralaxe e aberração planetária - a fim de tornar possível a utilização de observações feitas em diferentes observatórios.

VII.2 - Subrotina RAIO

Esta subrotina resolve o sistema de equações

$$\Delta = A - \frac{B}{r^3} ,$$

$$r^2 = \Delta^2 - (2R \cos \psi) + R^2 ,$$

onde A e B são constantes obtidas a partir dos dados de entrada e

$$R \cos \psi = \lambda XT + \mu YT + \nu ZT ,$$

$$R^2 = XT^2 + YT^2 + ZT^2 ,$$

a partir de uma estimativa inicial da distância heliocêntrica do astro, r, e de acordo com o esquema abaixo:

$$\Delta^{(i)} = A - \frac{B}{\left[r^{(i-1)}\right]^3},$$

$$\left[r^{(i)}\right]^2 = \left[\Delta^{(i)}\right]^2 - (2R\cos\psi)\Delta^{(i)} + R^2,$$

sendo $i = 1,2,\ldots,$ e $r^{(o)}$ a estimativa inicial. O ciclo de iterações é processado até que

$$\left|r^{(i)} - r^{(i-1)}\right| \leq 10^{-6}.$$

Este esquema, entretanto, apresentou uma circunstância de não solução: se $A<0$, obtivemos $\Delta<0$, o que é impossível. Testes realizados em calculadoras demonstraram que nestas circunstâncias outros esquemas, como, por exemplo, aproximações de Newton - Raphsow (Herrick, 8, pags. 386-387, 390-391) podem ser usados. Ilustrativamente, tendo-se $r^{(o)}$, podemos obter

$$\Delta^{(i-1)} = R\cos\psi \pm \sqrt{\left[r^{(i-1)}\right]^2 - R^2 + (R\cos\psi)^2}, \quad (39)$$

$$f(r^{(i-1)}) = \frac{A}{B} - \frac{\Delta^{(i-1)}}{B} - \frac{1}{\left[r^{(i-1)}\right]^3}, \quad (40)$$

$$-f'(r^{(i-1)}) = \frac{r^{(i-1)}}{B(\Delta^{(i-1)} - R\cos\psi)} - \frac{3}{\left[r^{(i-1)}\right]^4}, \quad (41)$$

$$\delta\Delta^{(i-1)} = f(r^{(i-1)})/-f'(r^{(i-1)}), \quad (42)$$

$$\Delta^{(i)} = \Delta^{(i-1)} + \delta\Delta^{(i-1)}, \qquad (43)$$

$$r^{(i)} = \left[\Delta^{(i)}\right]^2 - (2R\cos\psi)\Delta^{(i)} + R^2 \quad 1/2 \qquad (44)$$

onde $i = 1,2,\ldots$ O ciclo de operações é fechado entre as fórmulas (44) e (40), a fórmula (39) sendo utilizada apenas se $i = 1$. A iteração será interrompida se

$$\left|\delta\Delta^{(i-1)} - \delta\Delta^{(i)}\right| < 10^{-7}.$$

Como é fácil de ser observado, o tempo de processamento é bem menor no primeiro esquema do que no segundo, uma vez que aquele opera sobre duas fórmulas e este sobre cinco. Assim sendo, a subrotina RAIO deveria conter um teste sobre o sinal de A e sobre $|\Delta - R\cos\psi|$. A necessidade de testar o módulo é devida ao fato de que a diferença indicada aparece, na equação (41), em denominador: se esta diferença se aproxima de zero, seu inverso crescerá para além de qualquer limite, induzindo indeterminações programáticas (Herrick, 8, pag 176).

Dado que, em programas muito longos, o ganho em tempo de processamento é fator crítico, os testes sobre A e $|\Delta - R\cos\psi|$ na subrotina RAIO, deverão, por exemplo, obedecer a uma hierarquia do tipo:

Se $A > 0$ - processe RAIO

Se $A < 0$ - teste se $|\Delta - R\cos\psi| < \varepsilon$,
caso negativo processe RAIO 2 (fórmulas (39) a (44)),
caso positivo processe RAIO 3 (ver fórmulas (45) a (48)).

Como proposta para uma RAIO 3, temos

$$\Delta^{(i-1)} = A - B/\left[r^{(i-1)}\right]3 , \tag{45}$$

$$f(r^{(i-1)}) = \left[\Delta^{(i-1)}\right]^2 - (2R\cos\psi)\Delta^{(i-1)} + R^2 - \left[r^{(i-1)}\right]^2, \tag{46}$$

$$-f'(r^{(i-1)}) = 2r^{(i-1)}\left[1 - \frac{3B}{\left[r^{(i-1)}\right]5}(\Delta^{(i-1)} - R\cos\psi)\right] \tag{47}$$

$\delta\Delta^{(i-1)}$, $\Delta^{(i)}$ e $r^{(i)}$ obtidos pelas fórmulas (42), (43), (44). (48)

A não utilização da RAIO 3, que não contém denominadores capazes de se anularem, em lugar da RAIO 2, decorre do fato de que a presença dos denominadores acelera a convergência, desde que, é claro, estejamos operando com diferenças $|\Delta - R\cos\psi| > \varepsilon$. Pesquisas adicionais terão que ser realizadas a fim de que possamos determinar um valor para ε, caso exista apenas um que possa ser satisfatório em todas as circunstâncias orbitais. Não encontramos, nos textos consultados, uma solução para este problema.

VII.3 - Estimativa inicial da distância heliocêntrica

Porque a solução de todo o problema do cálculo da órbita depende da possibilidade de serem realizadas estimativas

mais ou menos confiáveis para $r^{(o)}$, vários esquemas têm sido propostos para solução desta questão, e são baseadas em gráficos ou tabelas (Herrick, 8, pag.386, 389). Este tipo de solução pressupõe ou a interrupção do processamento a fim de que os gráficos e tabelas sejam consultados, ou a utilização de arquivos no computador. A primeira hipótese não deve ser considerada por razões evidentes. Quanto à segunda, somos de opinião que soluções alternativas possam, talvez, serem testadas. Como exemplo, dado o seu interesse para a determinação de efemérides de procura, que são provisórias e servem apenas como orientação aos observadores, acreditamos que o método de Olbers para a determinação de órbitas parabólicas, possa ser estabelecido como uma subrotina capaz de:

a) fornecer estimativas mais precisas para Δ_2 ou r_2 (o índice 2 referindo-se à data intermediária);

b) fornecer efemérides de procura no caso de cometas ou asteróides recém descobertos.

Certamente, este método pressupõe a possibilidade de estimativas de Δ na posição 1 estarem disponíveis, o que parece retornar o problema ao seu ponto de partida. Entretanto, o que acreditamos ser possível realizar é a combinação dos métodos de Olbers e de Gauss na solução do problema, uma vez que neste devemos estimar Δ_2, ou r_2, e naquele Δ_1. A forma pela qual esta combinação poderá, ou não, ser processada, depende de estudos posteriores.

VII.4 - Correções diferenciais

Que as correções diferenciais, tanto gaussianas como pe-

los resíduos muito lineares, tenham apresentado circunstâncias de não convergência, é questão ainda a ser justificada, pois que na bibliografia consultada não encontramos indicações dos motivos pelos quais a não convergência ocorreu em alguns casos (Kohler e Brabantia), mas não em outros (Lanzia e Ceres).

VII.5 - Espaçamento entre as datas das observações

Para exemplificar, quando utilizamos as datas 1, 11 e 21 de fevereiro de 1978, o processamento da órbita de 1977HB foi interrompido devido a uma raiz inválida ($\sqrt{1 - e^2}$ para $e > 1$). Contudo, com as datas 11 e 21 de fevereiro e 8 de março, os resultados foram satisfatórios (ver secção VI). A primeira idéia é que o esquema geral utilizado seja, de alguma forma, sensível ao espaçamento entre as datas de observação. A solução deste problema envolverá, certamente, processamento de um maior número de órbitas, juntamente com um reexame das fórmulas utilizadas.

VII.6 - Algumas circunstâncias de não solução não previstas no programa

1) Se a latitude do observador é muito próxima a zero, o método de Gauss só é aplicável se uma quarta for utilizada. Em caso contrário, apenas com três observações, pode-se demonstrar que D tende a zero (Danjon, 5, pag.218).

2) Se as três posições observadas do objeto estiverem

sobre um grande círculo, D também se anula.

3) Tomando a representação

$$\vec{r} = \vec{L}\Delta - \vec{R} ,$$

com significados óbvios para $\vec{L}$ e $\vec{R}$, temos que se $\vec{R}$ e $\ddot{\vec{R}}$ estiverem contidos no plano determinado por $\vec{L}$ e $\dot{\vec{L}}$, nenhuma solução é possível. Apesar do método de Gauss não operar com $\ddot{\vec{R}}$ e $\dot{\vec{L}}$, esta condição pode ser usada para identificação de não soluções.

4) "As circunstâncias de não soluções ocorrem nas órbitas heliocêntricas quando o objeto está se movendo próximo ao plano da eclítica; para órbitas geocêntricas quando o objeto e o ponto de observação estão se movendo próximos ao plano do equador."

As observações 2, 3, 4 foram tiradas de Herrick (8, pag 379).

Evidentemente estas circunstâncias de não solução terão que ser incorporadas ao programa.

VII.7 - Outros tipos de órbitas

Apesar do projeto pretender a determinação de órbitas elíticas, é evidentemente útil que outras possibilidades sejam acessíveis. Explicitamente, a sugestão é a incorporação de um comando capaz de testar o valor da excentricidade, antes que os elementos orbitais sejam computados, de tal maneira que subrotinas para cálculo de trajetórias hiperbólicas e circulares possam ser acionadas.

VII.8 - Perturbações

O movimento de dois corpos foi todo formulado a partir da utilização das coordenadas esféricas (r, α, δ). Entretanto a integração das equações de movimento para o problema perturbado foi feita a partir de fórmulas em coordenadas retangulares (x, y, z). A utilização de ambos os sistemas em um mesmo problema depende de um exame mais cuidadoso a fim de determinarmos se a transformação entre as coordenadas é capaz de produzir erros observáveis. Além disso, apesar dos resultados não perturbados terem sido bastante testados, o mesmo não ocorreu com os perturbados. Dessa forma, as previsões feitas com base na técnica adotada para o cálculo das perturbações não devem ser, ainda, tidas como confiáveis.

VII.9 - Observações redundantes

É fato estabelecido pela experiência que a utilização de um número de observações precisas superiores a três é capaz, mesmo que perturbações não sejam incorporadas, de melhorar consideravelmente a confiabilidade dos resultados (Danjon, 5, pag.251).

VII.10 - Análise de erros

Com este título estamos indicando explicitamente uma análise que nos permita fixar com segurança os efeitos que os erros nos alfa e delta observados possa ter sobre os ele-

mentos orbitais, após todo o processamento ser executado. A alternativa a um estudo teórico de propagação de erro, em geral extremamente difícil em problemas longos, é o método das tentativas a partir de ensaios computacionais. Em outras palavras, são dados pequenas alterações nos valores de entrada para muitas circunstâncias astronomicamente possíveis, e os resultados comparados. Esta técnica não produz nenhuma demonstração, porém com a experiência acumulada ao longo de muitas computações, é capaz de fornecer indicadores razoavelmente seguros sobre esta questão.

APÊNDICE 1

Considere-se a figura 7 , onde os P_i, i = 1,2,3, são as sucessivas posições do corpo celeste; x_i, y_i, z_i, i = 1,2,3, são as coordenadas retangulares heliocêntricas destas posições; e S_i, i = 1,2,3, são as áreas dos triângulos formados por essas posições tendo o Sol como vértice comum. Através da Geometria Analítica, sabe-se que o plano gerado por P_1,P_2, P_3, e que passa pela origem, tem o determinante das suas coordenadas nulo, isto é:

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} = 0$$

Desenvolvendo-se o determinante em relação à primeira coluna, tem-se

$$x_1(y_2z_3 - y_3z_2) - x_2(y_1z_3 - y_3z_1) + x_3(y_1z_2 - y_2z_1) = 0 \quad (1)$$

As quantidades entre parênteses são o dobro das projeções das áreas S_2, S_3 e S_1 sobre o plano yz. Segue-se a demonstração para apenas uma das áreas, no caso, S_2; as outras são obtidas de maneira análoga.

Seja a figura A1, onde S_R é a área do retângulo cujos lados são z_3 e y_2.

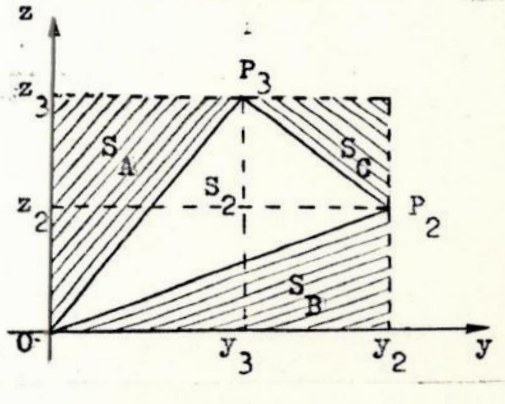


Fig.A1

A área S_2 será dada por

$$S_2 = S_R - S_A - S_B - S_C ,$$

onde

$$S_R = y_2 z_3 ,$$

$$S_A = \frac{y_3 z_3}{2} ,$$

$$S_B = \frac{y_2 z_2}{2} ,$$

$$S_C = \frac{(y_2 - y_3)(z_3 - z_2)}{2} .$$

Portanto,

$$S_2 = y_2 z_3 - \frac{y_3 z_3}{2} - \frac{y_2 z_2}{2} - \frac{(y_2 - y_3)(z_3 - z_2)}{2} .$$

Multiplicando ambos os membros da equação por 2 tem-se

$$2S_2 = 2y_2 z_3 - y_3 z_3 - y_2 z_2 - y_2 z_3 + y_2 z_2 + y_3 z_3 - y_3 z_2 ,$$

donde

$$2S_2 = y_2 z_3 - y_3 z_2 ,$$

como queria-se demonstrar.

Dividindo a equação (1) por $(y_1z_3 - y_3z_1)$, tem-se

$$x_1\frac{(y_2z_3 - y_3z_2)}{(y_1z_3 - y_3z_1)} - x_2 + x_3\frac{(y_1z_2 - y_2z_1)}{(y_1z_3 - y_3z_1)} = 0$$

Como

$$y_2z_3 - y_3z_2 = 2S_2 ,$$

$$y_1z_3 - y_3z_1 = 2S_3 ,$$

$$y_1z_2 - y_2z_1 = 2S_1 ,$$

temos a equação

$$x_1\frac{S_2}{S_3} - x_2 + x_3\frac{S_1}{S_3} = 0$$

De forma similar resolve-se o determinante em relação às outras duas colunas e obtem-se o sistema de equações fundamentais do método de Gauss e d'Olbers

$$\begin{cases} x_1\dfrac{S_1}{S_3} - x_2 + x_3\dfrac{S_1}{S_3} = 0 \\ y_1\dfrac{S_1}{S_3} - y_2 + y_3\dfrac{S_1}{S_3} = 0 \\ z_1\dfrac{S_1}{S_3} - z_2 + z_3\dfrac{S_1}{S_3} = 0 \end{cases}$$

APÊNDICE 2

Considere-se a figura A2, onde

⊙ : Sol

P_o e P: posições do corpo celeste nos instantes t_o e $t=t_o+0$,

S_T: área do triângulo $P_o \odot P$,

S_S: área do setor curvilíneo $P_o \odot P$,

X⊙Y: sistema de referência retangular heliocêntrico situado no plano da órbita.

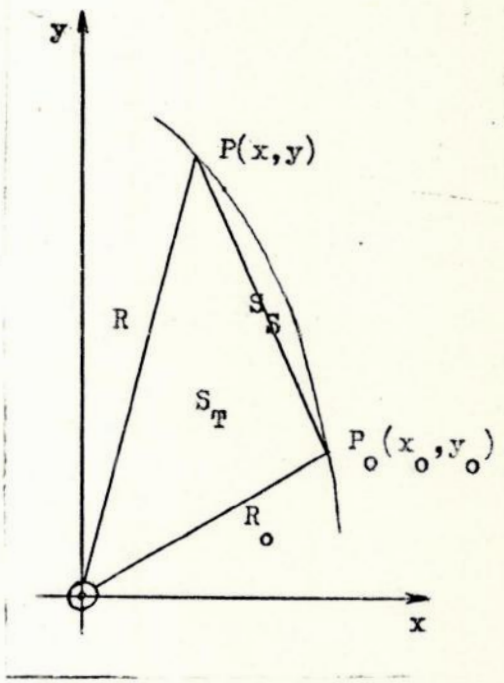


Fig.A2

A área S_T é dada pela expressão

$$2S_T = x_o y - y_o x \qquad (1)$$

Pela segunda lei de Kepler, segundo a qual o raio vetor varre áreas iguais em tempos iguais, tem-se a expressão para a área do setor curvilíneo

$$2S_S = h\,\theta \ , \qquad (2)$$

onde h é a constante das áreas por unidade de massa.

Fazendo-se o desenvolvimento em série de Taylor de x e y em função do tempo em torno de t_o , tem-se

$$x = x_o + \theta\left(\frac{dx}{dt}\right)_o + \frac{\theta^2}{2!}\left(\frac{d^2x}{dt^2}\right)_o + \frac{\theta^3}{3!}\left(\frac{d^3x}{dt^3}\right)_o + \cdots,$$

$$y = y_o + \theta\left(\frac{dy}{dt}\right)_o + \frac{\theta^2}{2!}\left(\frac{d^2y}{dt^2}\right)_o + \frac{\theta^3}{3!}\left(\frac{d^3y}{dt^3}\right)_o + \cdots,$$

Substituindo-se em (1),

$$2S_T = x_o y_o + x_o\,\theta\left(\frac{dy}{dt}\right)_o + x_o\,\frac{\theta^2}{2!}\left(\frac{d^2y}{dt^2}\right)_o + x_o\,\frac{\theta^3}{3!}\left(\frac{d^3y}{dt^3}\right)_o +$$

$$- x_o y_o - y_o\,\theta\left(\frac{dx}{dt}\right)_o - y_o\,\frac{\theta^2}{2!}\left(\frac{d^2x}{dt^2}\right)_o - y_o\,\frac{\theta^3}{3!}\left(\frac{d^3x}{dt^3}\right)_o + \cdots,$$

$$2S_T = \theta\left[x\,\frac{dy}{dt} - y\,\frac{dx}{dt}\right]_o + \frac{\theta^2}{2!}\left[x\,\frac{d^2y}{dt^2} - y\,\frac{d^2x}{dt^2}\right]_o +$$

$$+ \frac{\theta^3}{3!}\left[x\,\frac{d^3y}{dt^3} - y\,\frac{d^3x}{dt^3}\right]_o + \cdots \qquad (3)$$

Sendo que h é o momento angular por unidade de massa , pode-se escrever

$$h = \vec{r} \times \dot{\vec{r}},$$

e portanto,

$$\vec{r} \times \dot{\vec{r}} = \begin{vmatrix} \vec{i} & \vec{j} & \vec{k} \\ x & y & 0 \\ \dot{x} & \dot{y} & 0 \end{vmatrix},$$

sendo que $z = 0$ e $\dot{z} = 0$, devido ao fato de os eixos xy estarem no plano da órbita. Portanto,

$$h = x\dot{y} - y\dot{x} \tag{4}$$

Derivando-se a expressão acima sucessivamente

$$x \frac{d^2y}{dt^2} - y \frac{d^2x}{dt^2} = 0 , \tag{5}$$

$$x \dddot{y} - y \dddot{x} + \ddot{y} \dot{x} - \dot{y} \ddot{x} = 0 ,$$

$$x \dddot{y} - y \dddot{x} = \dot{y} \ddot{x} - \dot{x} \ddot{y} , \tag{6}$$

$$x\, y^{iv} - y\, x^{iv} - 2 (\dot{y} \dddot{x} - \dot{x} \dddot{y}) = 0 ,$$

$$x\, y^{iv} - y\, x^{iv} = 2 (\dot{y} \dddot{x} - \dot{x} \dddot{y}) , \tag{7}$$

Observa-se que os termos $\ddot{x}$ e $\ddot{y}$ são as projeções da aceleração newtoniana do corpo. Tem-se portanto,

$$\ddot{x} = - \frac{kx}{R^3} , \tag{8}$$

$$\ddot{y} = - \frac{ky}{R^3} , \tag{9}$$

onde k é o quadrado da constante gravitacional. Derivando-se as expressões acima, obtém-se

$$\ddot{x} = -\frac{k}{R^3}\dot{x} + \frac{3k}{R^4}x\frac{dR}{dt}, \qquad (10)$$

$$\ddot{y} = -\frac{k}{R^3}\dot{y} + \frac{3k}{R^4}y\frac{dR}{dt}. \qquad (11)$$

Substituindo-se as expressões (8) e (9) em (6),

$$x\ddot{y} - y\ddot{x} = -\dot{y}\frac{kx}{R^3} + \dot{x}\frac{ky}{R^3},$$

$$x\dddot{y} - y\dddot{x} = -\frac{k}{R^3}(x\dot{y} - y\dot{x}).$$

Obtemos através de (4):

$$x\dddot{y} - y\dddot{x} = -\frac{k}{R^3}h.$$

Substituindo (10) e (11) em (7) e utilizando (4) chega-se a

$$x\dddot{y} - y\dddot{x} = 2\left[\dot{y}\left(-\frac{k\dot{x}}{R^3} + \frac{3k}{R^4}x\frac{dR}{dt}\right) - \dot{x}\left(-\frac{k\dot{y}}{R^3} + \frac{3k}{R^4}y\frac{dR}{dt}\right)\right]$$

$$= \frac{2k}{R^3}\left(-\dot{x}\dot{y} + \frac{3}{R}x\dot{y}\frac{dR}{dt} + \dot{x}\dot{y} - \frac{3}{R}y\dot{x}\frac{dR}{dt}\right)$$

$$= \frac{6k}{R^4}\frac{dR}{dt}(x\dot{y} - y\dot{x})$$

$$= \frac{6k}{R^4}h\frac{dR}{dt}.$$

A área do triângulo será finalmente dada substituindo-se

esses resultados na equação (3):

$$2S_T = 0h + \frac{\theta^2}{2!}(0) + \frac{\theta^3}{3!}\left(-\frac{k}{R^3}h\right) + \frac{\theta^4}{4!}\left(\frac{6kh}{R^4}\frac{dR}{dt}\right) + \cdots$$

$$2S_T = 0h\left(1 - \frac{k}{6}\frac{1}{R_o^3}\theta^2 + \frac{k}{4}\frac{1}{R_o^4}\theta^3\frac{dR}{dt} + \cdots\right) .$$

APÊNDICE 3

Sejam $x_1y_1z_1$, $x_2y_2z_2$ e $x_3y_3z_3$ as coordenadas heliocêntricas correspondentes às posições P_1, P_2 e P_3 do corpo em estudo. Pode-se desenvolver estas coordenadas em função do tempo e da posição média como segue,

$$x = x_2 + a(t - t_2) + b(t - t_2)^2 + c(t - t_2)^3 + d(t - t_2)^4 + \ldots, \qquad (1)$$

Se

$$\theta_1 = t_2 - t_1 ,$$

$$\theta_2 = t_3 - t_2 ,$$

$$\theta_3 = \theta_1 + \theta_2 ,$$

então,

$$x_1 = x_2 - a\,\theta_1 + b\,\theta_1^2 - c\,\theta_1^3 + d\,\theta_1^4 ,$$

$$x_2 = x_2 ,$$

$$x_3 = x_2 + a\,\theta_2 + b\,\theta_1^2 + c\,\theta_2^3 + d\,\theta_2^4 .$$

Derivando-se duas vezes a expressão (1) em relação ao tempo,

$$\frac{d^2x}{dt^2} = 2b + 6c(t - t_2) + 12d(t - t_2)^2 ,$$

donde se obtém, para as três posições,

$$\frac{d^2x_1}{dt_1^2} = 2b - 6c\,\theta_1 + 12d\,\theta_1^2 ,$$

$$\frac{d^2x_2}{dt_2^2} = 2b ,$$

$$\frac{d^2x_3}{dt_3^2} = 2b + 6c\,\theta_2 + 12d\,\theta_2^2 .$$

Lembrando que num campo de forças newtonianas temos

$$\frac{d^2x}{dt^2} = -\frac{kx}{r^3} ,$$

se obtém o sistema

$$x_1 = x_2 - a\,\theta_1 + b\,\theta_1^2 - c\,\theta_1^3 + d\,\theta_1^4 , \qquad (2)$$

$$x_2 = x_2 , \qquad (3)$$

$$x_3 = x_2 + a\,\theta_2 + b\,\theta_2^2 + c\,\theta_2^3 + d\,\theta_2^4 , \qquad (4)$$

$$-\frac{kx_1}{r_1^3} = 2b - 6c\theta_1 + 12d\theta_1^2 , \qquad (5)$$

$$-\frac{kx_2}{r_2^3} = 2b , \qquad (6)$$

$$-\frac{kx_3}{r_3^3} = 2b + 6c\theta_2 + 12d\theta_2^2 . \qquad (7)$$

Multiplicando a expressão (2) por θ_2, a expressão (3) por $-\theta_3$ e a expressão (4) por θ_1, e somando,

$$\theta_2 x_1 - \theta_3 x_2 + \theta_1 x_3 = (\theta_2 - \theta_1 - \theta_2 + \theta_1)x_2 + b\theta_1\theta_2(\theta_1 + \theta_2) +$$

$$+ c\theta_1\theta_2(\theta_2^2 - \theta_1^2) + d\theta_1\theta_2(\theta_1^3 + \theta_2^3) ,$$

$$\theta_2 x_1 - \theta_3 x_2 + \theta_1 x_3 = b\theta_1\theta_2\theta_3 + c\theta_1\theta_2(\theta_1 + \theta_2)(\theta_1 - \theta_2) +$$

$$+ d\theta_1\theta_2(\theta_1 + \theta_2)(\theta_1^2 - \theta_1\theta_2 + \theta_2^2) ,$$

$$\theta_2 x_1 - \theta_3 x_2 + \theta_1 x_3 = \theta_1\theta_2\theta_3\left[b + c(\theta_1 - \theta_2) + d(\theta_1^2 - \theta_1\theta_2 + \theta_2^2)\right] . \qquad (8)$$

Repetindo o processo acima com as expressões (5), (6) e (7),

$$-\frac{k}{12}\left[\left(\frac{\theta_2 x_1}{r_1^3} - \frac{\theta_2 x_2}{r_2^3} + \frac{\theta_1 x_3}{r_3^3}\right)\right] = d\,\theta_1\theta_2\theta_3 . \qquad (9)$$

Somando as expressões (5), (6) e (7), obtemos a terceira equação do sistema:

$$-\frac{k}{6}\left[\left(\frac{x_1}{r_1^3}+\frac{x_2}{r_2^3}+\frac{x_3}{r_3^3}\right)\right] = b + c(\theta_2 - \theta_1) + 2d(\theta_1^2 + \theta_2^2) \quad .(10)$$

A fim de eliminar as incógnitas b, c, d do sistema, multiplicamos as expressões (8), (9), (10), respectivamente, por 1, $\theta_1^2 + \theta_1\theta_2 + \theta_2^2$, $-\theta_1\theta_2\theta_3$ e em seguida somamos as novas expressões:

$$-\frac{k}{12}(\theta_1^2 + \theta_1\theta_2 + \theta_2^2)\left(\frac{\theta_2 x_1}{r_1^3} - \frac{\theta_3 x_2}{r_2^3} + \frac{\theta_1 x_3}{r_3^3}\right) + \frac{k}{6}\theta_1\theta_2\theta_3\left(\frac{x_1}{r_1^3}+\frac{x_2}{r_2^3}+\frac{x_3}{r_3^3}\right) +$$

$$+\theta_2 x_1 - \theta_3 x_2 + \theta_1 x_3 = 0 ,$$

$$\frac{k}{12}\frac{\theta_2 x_1}{r_1^3}(-\theta_1^2 - \theta_1\theta_2 - \theta_2^2 + 2\theta_1\theta_3) + \frac{k}{12}\frac{\theta_3 x_2}{r_2^3}(\theta_1^2 + \theta_1\theta_2 + \theta_2^2 + 2\theta_1\theta_2) +$$

$$+\frac{k}{12}\frac{\theta_1 x_3}{r_3^3}(-\theta_1^2 - \theta_1\theta_2 - \theta_2^2 + 2\theta_2\theta_3) + \theta_2 x_1 - \theta_3 x_2 + \theta_1 x_3 = 0 ,$$

$$\theta_2 x_1\left\{1 + \frac{k}{12r_1^3}\left[-\theta_2^2 - \theta_1(\theta_1 + \theta_2) + 2\theta_1(\theta_1 + \theta_2)\right]\right\} +$$

$$-\theta_3 x_2\left\{1 - \frac{k}{12r_2^3}\left[\theta_1(\theta_1 + \theta_2) + \theta_2(\theta_2 + \theta_1) + \theta_1\theta_2\right]\right\} +$$

$$+\theta_1 x\left\{1 + \frac{k}{12r_3^3}\left[-\theta_1^2 - \theta_2(\theta_1 + \theta_2) + 2\theta_2\theta_3\right]\right\} = 0 ,$$

$$\theta_2 x_1 \left[1 + \frac{k}{12r_1^3}(\theta_1\theta_3 - \theta_2^2)\right] - \theta_3 x_2 \left\{1 - \left[\frac{k}{12r_2^3}\ \theta_3(\theta_1 + \theta_2) + \right.\right.$$

$$\left.\left. + \theta_1\theta_2\right]\right\} + \theta_1 x_3 \left[1 + \frac{k}{12r_3^3}(\theta_2\theta_3 - \theta_1^2)\right] = 0 ,$$

$$\theta_2 x_1 \left[1 + \frac{k}{12r_1^3}(\theta_1\theta_3 - \theta_2^2)\right] - \theta_3 x_2 \left[1 - \frac{k}{12r_2^3}(\theta_1\theta_2 + \theta_3^2)\right] +$$

$$+ \theta_1 x_3 \left[1 + \frac{k}{12r_3^3}(\theta_2\theta_3 - \theta_1^2)\right] = 0 .$$

Ou, fazendo

$$\psi_1 = \frac{k}{12}(\theta_1\theta_3 - \theta_2^2) ,$$

$$\psi_2 = \frac{k}{12}(\theta_1\theta_2 - \theta_3^2) ,$$

$$\psi_3 = \frac{k}{12}(\theta_2\theta_3 - \theta_1^2) ,$$

vem

$$\theta_2 x_1 \left(1 + \frac{\psi_1}{r_1^3}\right) - \theta_3 x_2 \left(1 - \frac{\psi_2}{r_2^3}\right) + \theta_1 x_3 \left(1 + \frac{\psi_3}{r_3^3}\right) = 0 . \quad (11)$$

De maneira análoga obtemos o sistema

$$\begin{cases} \theta_2 x_1 \left(1 + \dfrac{\psi_1}{r_1^3}\right) - \theta_3 x_2 \left(1 - \dfrac{\psi_2}{r_2^3}\right) + \theta_1 x_3 \left(1 + \dfrac{\psi_3}{r_3^3}\right) = 0 \\ \theta_2 y_1 \left(1 + \dfrac{\psi_1}{r_1^3}\right) - \theta_3 y_2 \left(1 - \dfrac{\psi_2}{r_2^3}\right) + \theta_1 y_3 \left(1 + \dfrac{\psi_3}{r_3^3}\right) = 0 \\ \theta_2 z_1 \left(1 + \dfrac{\psi_1}{r_1^3}\right) - \theta_3 z_2 \left(1 - \dfrac{\psi_2}{r_2^3}\right) + \theta_1 z_3 \left(1 + \dfrac{\psi_3}{r_3^3}\right) = 0 \end{cases} \quad (12)$$

Seja agora o sistema de equações fundamentais de Gauss:

$$x_1 \frac{S_2}{S_3} - x_2 + x_3 \frac{S_1}{S_3} = 0 ,$$

$$y_1 \frac{S_2}{S_3} - y_2 + y_3 \frac{S_1}{S_3} = 0 ,$$

$$z_1 \frac{S_2}{S_3} - z_2 + z_3 \frac{S_1}{S_3} = 0 .$$

Comparando com o sistema (12), chega-se às fórmulas de Gibbs

$$\frac{S_2}{S_3} = \frac{\theta_2}{\theta_3} \frac{1 + \dfrac{\psi_1}{r_1^3}}{1 - \dfrac{\psi_2}{r_2^3}} ,$$

e

$$\frac{S_1}{S_3} = \frac{\theta_2}{\theta_3} \frac{1 + \dfrac{\psi_3}{r_3^3}}{1 - \dfrac{\psi_2}{r_2^3}} .$$

APÊNDICE 4

Seja

$$\operatorname{sen}^2 \frac{g}{2} = \rho \ , \tag{1}$$

onde

$$g = \frac{v_3 - v_1}{2} \ .$$

Assim,

$$\cos g = 1 - 2\rho \ , \tag{2}$$

e

$$\operatorname{sen}^2 g = 4\rho(1 - \rho) \ . \tag{3}$$

Expandindo

$$\frac{1}{\sqrt{1 - x^2}}$$

e integrando, vem

$$\text{arc sen } x = x + \frac{x^3}{2 \cdot 3} + \frac{1 \cdot 3}{2 \cdot 4 \cdot 5} x^5 + \frac{1 \cdot 3 \cdot 5}{2 \cdot 4 \cdot 6 \cdot 7} x^7 + \cdots$$

Fazendo-se $x = \text{sen}2g$ na série acima,

$$2g = \text{sen}2g + \frac{\text{sen}^3 2g}{2\cdot 3} + \frac{1\cdot 3\ \text{sen}^5 2g}{2\cdot 4\cdot 5} + \frac{1\cdot 3\cdot 5\ \text{sen}^7 2g}{2\cdot 4\cdot 6\cdot 7} + \ldots ,$$

$$2g - \text{sen}2g = \frac{1}{6}\,\text{sen}^3 2g + \frac{3}{40}\,\text{sen}^5 2g + \frac{5}{112}\,\text{sen}^7 2g + \ldots .$$

Sabendo-se que $\text{sen}2g = 2\ \text{sen}\ g\ \cos g$,

$$2g - \text{sen}2g = \frac{2^3}{6}\,\text{sen}^3 g\,\cos^3 g + \frac{3\cdot 2^5}{40}\,\text{sen}^5 g\,\cos^5 g + \frac{5\cdot 2^7}{112}\,\text{sen}^7 g\,\cos^7 g + \ldots$$

Dividindo-se ambos os membros por $\text{sen}^3 g$,

$$\frac{2g - \text{sen}2g}{\text{sen}^3 g} = \frac{4}{3}\cos^3 g + \frac{12}{5}\,\text{sen}^2 g\,\cos^5 g + \frac{40}{7}\,\text{sen}^4 g\,\cos^7 g + \ldots$$

Fazendo uso de (1), (2) e (3),

$$\frac{2g - \text{sen}2g}{\text{sen}^3 g} = \frac{4}{3}(1 - 2\rho)^3 + \frac{48}{5}\rho(1 - \rho)(1 - 2\rho)^5 +$$

$$+ \frac{640}{7}\rho^2(1 - \rho)^2(1 - 2\rho)^7 + \ldots ,$$

$$\frac{2g - \text{sen}2g}{\text{sen}^3 g} = \frac{4}{3}\left(1 + \frac{6}{5}\rho + \frac{48}{35}\rho^2 + \frac{32}{21}\rho^3 + \ldots\right) .$$

O termo de ordem k é dado pela expressão:

$$\frac{4\cdot 6 \ldots (2k + 2)}{3\cdot 5 \ldots (2k + 1)}\,\rho^{k-1} .$$

Gauss, no seu trabalho original, destacou que o desenvolvimento da quantidade inversa converge mais rapidamente, o que pode ser verificado observando a série

$$\frac{\text{sen}^3 g}{2g - \text{sen}^3 g} = \frac{3}{4} - \frac{9}{10}\rho + \frac{9}{175}\rho^2 + \frac{26}{875}\rho^3 + \dots .$$

A inversão pode ser obtida por divisão ou pelo método dos coeficientes indeterminados. Para tanto, tomamos.

$$\frac{\text{sen}^3 g}{2g - \text{sen}^3 g} = \frac{3}{4} - \frac{9}{10}(\rho - \xi) ,$$

onde

$$\xi = \frac{2}{35}\rho^2 + \frac{52}{1575}\rho^3 + \dots$$

Sendo

$$\gamma^3 - \gamma^2 = m\,\frac{2g - \text{sen}2g}{\text{sen}^3 g} ,$$

e

$$\gamma^2 = \frac{m}{1 + \text{sen}^2\left(\frac{\xi}{2}\right)} ,$$

temos

$$\gamma^3 - \gamma^2 = \frac{m}{\frac{3}{4} - \frac{9}{10}(\rho - \xi)} ,$$

e

$$\gamma^2 = \frac{m}{1+\rho}$$

Utilizando o sistema acima para eliminar ρ , chegamos a equação, denominada de Gauss,

$$\gamma^3 - \gamma^2 - H\gamma - \frac{H}{9} = 0 ,$$

onde

$$H = \frac{m}{\frac{5}{6} + 1 + \xi} .$$

A N E X O - 1

Artigos sobre os aspectos teóricos e práticos do método de integração numérica de Bulirsch - Stoer

IMSL ROUTINE NAME - DREBS

PURPOSE - DIFFERENTIAL EQUATION SOLVER - EXTRAPOLATION METHOD

USAGE - CALL DREBS (FCN,Y,X,N,JM,IND,JSTART,H,HMIN, TOL,R,S,WK,IER)

ARGUMENTS

FCN - NAME OF SUBROUTINE FOR EVALUATING FUNCTIONS. (INPUT)
THE SUBROUTINE ITSELF MUST ALSO BE PROVIDED BY THE USER AND IT SHOULD BE OF THE FOLLOWING FORM

```
SUBROUTINE FCN(N,X,Y,YPRIME)
REAL Y(N),YPRIME(N)
      .
      .
      .
```

FCN SHOULD EVALUATE YPRIME(1),...,YPRIME(N) GIVEN N,X, AND Y(1),...,Y(N). YPRIME(I) IS THE FIRST DERIVATIVE OF Y(I) WITH RESPECT TO X.
FCN MUST APPEAR IN AN EXTERNAL STATEMENT IN THE CALLING PROGRAM AND N,X,Y(1),...,Y(N) MUST NOT BE ALTERED BY FCN.

Y - DEPENDENT VARIABLES, VECTOR OF LENGTH N. (INPUT AND OUTPUT)
ON INPUT, Y(1),...,Y(N) SUPPLY INITIAL VALUES.
ON OUTPUT, Y(1),...,Y(N) ARE REPLACED WITH AN APPROXIMATE SOLUTION AT X (AS SET ON OUTPUT).

X - INDEPENDENT VARIABLE. (INPUT AND OUTPUT)
ON INPUT, X SUPPLIES THE INITIAL VALUE.
ON OUTPUT, X IS REPLACED WITH THE UPDATED VALUE OF THE INDEPENDENT VARIABLE.

N - THE NUMBER OF EQUATIONS. (INPUT)

JM - THE MAXIMUM ORDER OF THE RATIONAL APPROXIMATION. (INPUT) JM MUST BE LESS THAN 7. A SUGGESTED VALUE IS JM=6. SEE REMARKS.

IND - CONVERGENCE TYPE INDICATOR. (INPUT)
IND = 1 SPECIFIES THE STANDARD ERROR TEST
IND = 2 SPECIFIES THE RELATIVE ERROR TEST
IND = 3 SPECIFIES THE ABSOLUTE ERROR TEST
SEE REMARKS FOR FURTHER DETAILS.

JSTART - INDICATOR. (INPUT)
THE USER MUST SET JSTART TO 0 OR -1
JSTART = 0 IMPLIES PERFORM A STEP.
THE FIRST STEP MUST BE DONE WITH THIS VALUE OF JSTART SO THAT THE SUBROUTINE CAN INITIALIZE ITSELF.
JSTART = -1 IMPLIES REPEAT THE LAST STEP WITH A NEW VALUE OF H OR JM.
THE INITIAL VALUES OF Y, S, AND X ARE SET TO THE INITIAL VALUES OF Y, S, AND X FROM THE MOST RECENT CALL TO DREBS WITH JSTART = 0.

3. AT EACH STEP OF THE INTEGRATION, THE EXTRAPOLATION PROCESS IS CONSIDERED TO HAVE CONVERGED WHEN EACH Y(I), I=1,...,N, HAS SATISFIED A CONVERGENCE CRITERION SPECIFIED BY THE USER. THE USER MAY CHOOSE ONE OF THREE CONVERGENCE CRITERIA. IN TESTING FOR CONVERGENCE, TWO SUCCESSIVE EXTRAPOLATED VALUES (FOR EACH COMPONENT) AT THE POINT IN QUESTION ARE COMPARED. LET THE DIFFERENCE BETWEEN THE TWO FOR THE J-TH COMPONENT BE CALLED D(J). THE THREE CONVERGENCE CRITERIA CAN BE STATED IN THE FOLLOWING MANNER;

   A. STANDARD ERROR
      LET YMAX(J) BE THE LARGEST ABSOLUTE VALUE ATTAINED SO FAR IN THE INTEGRATION BY THE DEPENDENT VARIABLE Y(J). THE CONVERGENCE REQUIREMENT IS
        ABS(D(J)/YMAX(J)) .LE. TOL, FOR J=1,...,N
      IF YMAX(J) IS LESS THAN TOL, IT IS REPLACED IN THE TEST BY TOL.

   B. RELATIVE ERROR
      LET Y(J) BE THE CURRENT APPROXIMATION TO THE RESPECTIVE DEPENDENT VARIABLE. THE CONVERGENCE REQUIREMENT IS
        ABS(D(J)/Y(J)) .LE. TOL, FOR J=1,...,N
      IF ABS(Y(J)) IS LESS THAN TOL, IT IS REPLACED IN THE TEST BY TOL.

   C. ABSOLUTE ERROR
      THE CONVERGENCE REQUIREMENT IS
        ABS(D(J)) .LE. TOL, J=1,...,N

4. JM, THE ORDER OF THE RATIONAL APPROXIMATION, DOES NOT HAVE TO EQUAL 1, THE ORDER OF THE DIFFERENTIAL EQUATIONS. AT EACH INTEGRATION STEP, AS MANY AS JM APPLICATIONS OF THE MIDPOINT RULE ARE COMPUTED FOR SUCCESSIVELY SMALLER VALUES OF H AND EXTRAPOLATED TO H=0 IN ATTEMPTING TO ACHIEVE CONVERGENCE. TYPICAL USAGE OF DREBS WOULD BE WITH JM SET TO 6.

Algorithm

DREBS performs one step in the integration of Y'=f(Y,X) with $Y(X_{input})$ given. The value of $Y(X_{output})$ is returned from DREBS.

DREBS is a modification of the Bulirsch-Stoer ALGOL procedure DESUB.

See references:

1. Bulirsch, R., and Stoer, J., "Numerical treatment of ordinary differential equations by extrapolation methods," Numerische Mathematik, 8(1)1966, 1-13.

2. Gragg, W.B., "On extrapolation algorithms for ordinary initial-value problems", J. SIAM Numerical Analysis, Series B, 2(1965), 384-403.

H - STEP SIZE. (INPUT AND OUTPUT)
ON INPUT, H IS AN INITIAL GUESS FOR THE STEP SIZE.
ON OUTPUT, H IS REPLACED BY A SUGGESTED STEP SIZE FOR THE NEXT STEP. THE SUGGESTED VALUE MAY BE LARGER OR SMALLER THAN THE ORIGINAL STEP SIZE.

HMIN - THE SMALLEST PERMISSIBLE STEP SIZE. (INPUT)
DREBS WILL DECREASE THE STEP SIZE UNTIL CONVERGENCE CAN BE OBTAINED.

TOL - TOLERANCE FOR ERROR CONTROL. (INPUT)

R - VECTOR OF LENGTH N. (OUTPUT)
ON OUTPUT, R CONTAINS THE ABSOLUTE ERRORS IN EACH COMPONENT FOR THE CURRENT STEP.

S - VECTOR OF LENGTH N. (INPUT AND OUTPUT)
IF IND = 1,
BEFORE THE FIRST CALL TO THE ROUTINE, S(I) SHOULD BE SET TO Y(I), I=1,...,N.
ON OUTPUT, S CONTAINS THE LARGEST VALUE OF EACH Y COMPUTED SINCE THE START OF THE INTEGRATION.
IF IND = 2,
BEFORE THE FIRST CALL TO THE ROUTINE, S(I) SHOULD BE SET TO Y(I), I=1,...,N.
ON OUTPUT, S CONTAINS THE LARGEST VALUE OF EACH Y COMPUTED DURING THE CURRENT STEP.
IF IND = 3,
BEFORE THE FIRST CALL TO THE ROUTINE, S(I) SHOULD BE SET TO 1.0, I=1,...,N.
ON OUTPUT, S IS UNCHANGED.

WK - WORK VECTOR OF LENGTH 29*N.
WK MUST REMAIN UNCHANGED BETWEEN SUCCESSIVE CALLS DURING INTEGRATION.

IER - ERROR PARAMETER. (OUTPUT)
TERMINAL ERROR
IER = 129 INDICATES CONVERGENCE COULD NOT BE OBTAINED WITH CURRENT VALUES OF H AND HMIN. Y,X, AND H HAVE BEEN UPDATED.
WARNING ERROR (WITH FIX)
IER = 66 INDICATES JM IS LESS THAN 1 OR GREATER THAN 6. JM IS RESET TO 6.

PRECISION/HARDWARE - SINGLE AND DOUBLE/H32
- SINGLE/H36,H48,H60

REQD. IMSL ROUTINES - UERTST,UGETIO

NOTATION - INFORMATION ON SPECIAL NOTATION AND CONVENTIONS IS AVAILABLE IN THE MANUAL INTRODUCTION OR THROUGH IMSL ROUTINE UHELP

REMARKS 1. THE SOLUTION Y, THE INDEPENDENT VARIABLE X, AND THE SUGGESTED STEP SIZE H ARE ALWAYS UPDATED EVEN IF CONVERGENCE IS NOT OBTAINED.

2. IN GENERAL, HMIN SHOULD BE MUCH SMALLER THAN H TO ALLOW THE PROGRAM TO ADJUST FOR RAPIDLY CHANGING SOLUTIONS (WITH RESPECT TO H).

3. Fox, P.A., "DESUB: Integration of a first-order system of ordinary differential equations", Mathematical Software (John R. Rice, Editor), Academic Press, New York, 1971, Chapter 9.

Example

This example illustrates the usage of DREBS. The system of differential equations

$$y_1' = y_2 \qquad y_1 = 1 \qquad \text{at } X=0$$

$$y_2' = y_1 \qquad y_2 = -1$$

is to be solved at X=4. We can proceed as follows:

Input:

```
      INTEGER   N,JM,IND,JSTART,IER
      REAL      Y(2),X,H,HMIN,TOL,R(2),S(2),WK(58)
      EXTERNAL  FCN
      N      =  2
      JM     =  6
      IND    =  2
      JSTART=   0
      Y(1)   =  1.0
      Y(2)   = -1.0
      X      =  0.0
      H      = .1
      HMIN   = .005
      TOL    =  1.0E-3
      S(1)   =  Y(1)
      S(2)   =  Y(2)
C                                  H = AMIN1(H,4.0-X) NECESSARY TO HIT
C                                  4.0 EXACTLY AT THE END.
    5 H = AMIN1(H,4.0-X)
C                                  CALL DREBS TO TAKE A STEP
      CALL DREBS (FCN,Y,X,N,JM,IND,JSTART,H,HMIN,TOL,R,S,WK,IER)
      IF (IER.NE.0) GO TO 15
      IF (X.LT.4.0-HMIN) GO TO 5
C                                  Insert statements to write solution here
      GO TO 20
   15 CONTINUE
C                                  Handle IER .GT. 0 here.  Items that
C                                  may help to diagnose the problem should
C                                  be output here, e.g. Y,X,R,S,H.
   20 CONTINUE
        .
        .
        .
      STOP
      END

      SUBROUTINE FCN(N,X,Y,YPRIME)
      REAL Y(N),YPRIME(N)
      YPRIME(1)=Y(2)
      YPRIME(2)=Y(1)
      RETURN
      END
```

OUT PUT:

IER = 0

X = 4.0

Y(1) = .0183

Y(2) = -.0183

Numerische Mathematik 8, 1–13 (1966)

Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods

ROLAND BULIRSCH and JOSEF STOER*

Received June 3, 1965

1. Introduction

Extrapolation constitutes a powerful means of numerical analysis for accelerating the convergence of solutions arising from discretization methods: If the underlying discretization method gives a result $T(h)$ for a finite stepsize $h \neq 0$, then the exact result $T(0)$ usually is very accurately approximated by the extrapolated value $\hat{T}_m(0)$ of an interpolating polynomial or rational function $\hat{T}_m(h)$ satisfying

$$\hat{T}_m(h_j) = T(h_j); \qquad j = 0, \ldots, m$$

for a sequence h_j of stepsizes tending to zero. In [7], the convergence behaviour of such extrapolation methods has been studied extensively provided $T(h)$ has an asymptotic expansion of the form

$$T(h) = \tau_0 + \tau_1 h^{\gamma_1} + \cdots + \tau_k h^{\gamma_k} + R_{k+1}(h)\, h^{\gamma_{k+1}}, \tag{1}$$

τ_i independent of h. For the existence of such expansions see GRAGG [2] and STETTER [8].

If $\gamma_i = i\gamma$ for all i, then it has been shown in [7] that there is a bound for the error $\hat{T}_m(0) - T(0)$ of the type

$$|\hat{T}_m(0) - T(0)| \leqq C h_0^{\gamma} \ldots h_m^{\gamma}.$$

Thus the error decreases with increasing γ. Moreover, experience has shown that extrapolation based on rational functions is normally better than polynomial extrapolation.

The present paper can be considered as a continuation of the more theoretical papers [7, 8]. Here, the practical aspects of extrapolation methods for solving the initial value problem for systems of ordinary differential equations of the form

$$y_i' = f_i(x, y_1, \ldots, y_n), \qquad i = 1, \ldots, n \tag{2}$$

are described.

The algorithm to be given uses rational function for extrapolation and is based on the midpoint-rule in a slightly modified form due to GRAGG** as the

* The research reported in this paper has been sponsored by the Air Force Office of Scientific Research under Grant AF EOAR 63-77 through the European Office of Aerospace Research (OAR), USAF.

** Personal communication to H. J. STETTER.

underlying discretization method. It can be shown that for the midpoint-rule the expansion (1) proceeds with even powers of h.

The proposed method is compared with the following alternatives

1. Runge-Kutta method.
2. The linear multistep method of Adams-Moulton-Bashforth (of order 6).
3. Extrapolation with polynomials based on the modified midpoint rule.

The comparison shows clearly that rational extrapolation yields

1. more accurate results;
2. needs much less operations in order to obtain these results;
3. is much more easy to program than the alternative methods, due to the following reasons:

There is no need to compute extra starting values as is the case for linear multistep methods. The order of approximation is not fixed and is automatically adapted to the special problem to be treated. Moreover, no special preparation of the differential equation is necessary, such as building up total derivatives and so on.

Historical remarks. COREY (1906) was perhaps the first, who tried to accelerate the convergence of trapezoidal sums $T(h)$ towards $T(0)$ by forming suitable linear combinations of $T(h_i)$, $i=0, 1, \ldots$. RICHARDSON and GAUNT (1927) applied this scheme for solving ordinary differential equations. KOMMERELL (1936) used it for the calculation of π. ROMBERG (1955) presented an extrapolation algorithm (for the stepsize sequence $h_i=h_0/2^i$) for the numerical quadrature. BOLTON and SCOINS (1956) described an extrapolation method for general sequences h_i and used it for solving the eigenvalue problem for ordinary and partial differential equations. BAUER and RUTISHAUSER-STIEFEL (1961) investigated the convergence behaviour of Romberg's method. RUTISHAUSER (1963) also used general sequences for the quadrature of functions and improved Euler's method for solving ordinary differential equations by extrapolation; he also applied extrapolation for numerical differentiation. LAURENT (1963) investigated the convergence of (polynomial) extrapolation schemes in the general case and applied extrapolation to various problems. Further investigations were made by LYNESS and MCHUGH (1963), MEIR and SHARMA (1965) and FILIPPI (1964); see also [2, 7, 8] and the references of these papers.

The authors wish to thank CH. REINSCH for useful discussions and for improving and testing the ALGOL program of section 5. They are also grateful to D. GRIES for his careful reading of the manuscript.

2. The Method of Computation

In the sequel the system (2) of differential equations is written in vector form

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

the initial values being

$$x_0,\ y_0=y(x_0).$$

We denote by $\eta(x, h)$ a suitable defined approximation to the exact value $y(x)$ obtained by a discretization method, e.g. the midpoint-rule, for stepsize $h\neq 0$.

The modified midpoint-rule proceeds as follows: If

$$x=x_0+l\,h, \qquad l \text{ an integer},$$

then

$$T(h)=T(h, x)$$

is defined recursively by

$$\begin{aligned}
&x_{i+1}=x_i+h, && i=0, 1, \ldots, l-1,\\
&\eta(x_1, h)=y_0+h f(x_0, y_0),\\
(3)\qquad &\eta(x_{i+1}, h)=\eta(x_{i-1}, h)+2h f\big(x_i, \eta(x_i, h)\big), && i=1, 2, \ldots, l-1,\\
&S(h, x)=\tfrac{1}{2}\big[\eta(x_l, h)+\eta(x_{l-1}, h)+h f\big(x_l, \eta(x_l, h)\big)\big],\\
&T(h, x)=S\Big(\frac{h}{2}, x\Big).
\end{aligned}$$

GRAGG has shown that under suitable differentiability assumptions, the asymptotic expansion of $T(h, x)$ proceeds with even powers of h

$$T(h, x)=y(x)+\tau_1(x)h^2+\tau_2(x)h^4+\cdots.$$

It can be easily proved that $\eta(x, h)=\eta(x, -h)$ holds. This shows that if $\eta(x, h)$ has an asymptotic expansion of the form

$$\eta(x, h)=\sigma_0(x)+\sigma_1(x)\,h+\sigma_2(x)\,h^2+\cdots,$$

then the odd terms $\sigma_{2i+1}(x)$ vanish. Indeed, the recursion formulae (3) are equivalent to

(a) $$\eta(x_0-h, h)=y_0-h f\big(x_0, \eta(x_0, h)\big),$$

(b) $$\eta(x_0, h)=y_0,$$

(c) $$\eta(x_i+h, h)=\eta(x_i-h, h)+2h f\big(x_i, \eta(x_i, h)\big), \qquad i=0, 1, 2, \ldots.$$

It follows from (b)

$$\eta(x_0, h)=\eta(x_0, -h)=y_0,$$

and by (a)

$$\begin{aligned}
\eta(x_0+h, -h)&=y_0+h f\big(x_0, \eta(x_0, -h)\big)\\
&=y_0+h f(x_0, y_0).
\end{aligned}$$

This proves

$$\eta(x_1, -h)=\eta(x_1, h).$$

This in turn implies by (c)

$$\begin{aligned}
\eta(x_2, -h)&=\eta(x_0, -h)+2h f\big(x_1, \eta(x_1, -h)\big)\\
&=\eta(x_0, h)+2h f\big(x_1, \eta(x_1, h)\big)\\
&=\eta(x_2, h).
\end{aligned}$$

In the same way follows

(d) $$\eta(x_i, h)=\eta(x_i, -h), \qquad i=3, 4, \ldots.$$

Now if x is a fixed number and

$$h_k:=\frac{x-x_0}{k}, \qquad k=1, 2, \ldots$$

we obtain by (d)

$$\eta(x, h_k)=\eta(x, -h_k), \qquad k=1, 2, \ldots,$$

proving

$$\sigma_{2i+1}(x)\equiv 0.$$

In the same way it can be shown that the asymptotic expansion of $T(h, x)$ contains only even powers of h. (For a generalization of this result see a forthcoming paper of STETTER.)

Therefore, it can be expected that extrapolation based on $T(h, x)$ thus defined will yield especially good results (see [7]). We propose to use rational extrapolation in order to approximate $T(0, x)=y(x)$. For the computation of the extrapolated values, the algorithm of [7] in a slightly modified form is used. Since this algorithm is applied to each component of $T(h, x)\in R^n$ separately, an index denoting the individual component is suppressed in the sequel in order to simplify the notation.

If $\hat{T}_m^{(i)}(h)$ denotes the rational function

$$\hat{T}_m^{(i)}(h)=\frac{p_0^{(i)}+p_1^{(i)}h^2+\cdots+p_\mu^{(i)}h^{2\mu}}{q_0^{(i)}+q_1^{(i)}h^2+\cdots+q_\nu^{(i)}h^{2\nu}}, \qquad \mu=\left[\frac{m}{2}\right], \qquad \nu=m-\mu$$

defined by the requirement

$$\hat{T}_m^{(i)}(h_k)=T(h_k, x), \qquad k=i, i+1, \ldots, i+m,$$

where $\{h_k\}$ is a strictly decreasing sequence of stepsizes tending to zero, the extrapolated values

$$T_m^{(i)} := \hat{T}_m^{(i)}(0)\approx T(0, x)$$

can be computed from $T(h_i, x)$ by the following set of formulae (see [7])

$$\begin{aligned} &T_{-1}^{(i)}=0,\\ &T_0^{(i)}=T(h_i, x),\\ &T_k^{(i)}=T_{k-1}^{(i+1)}+\frac{T_{k-1}^{(i+1)}-T_{k-1}^{(i)}}{\left(\frac{h_i}{h_{i+k}}\right)^2\left[1-\frac{T_{k-1}^{(i+1)}-T_{k-1}^{(i)}}{T_{k-1}^{(i+1)}-T_{k-2}^{(i+1)}}\right]-1}, \qquad k\geqq 1 \end{aligned} \tag{4}$$

connecting the elements $T_{k-2}^{(i+1)}, T_{k-1}^{(i)}, T_{k-1}^{(i+1)}, T_k^{(i)}$ of the tableau

$$\begin{array}{ccccccc} T_0^{(0)} & & & & & & \\ T_{-1}^{(1)} & T_1^{(0)} & & & & & \\ & T_0^{(1)} & \ddots & & & & \\ T_{-1}^{(2)} & \vdots & & T_{m-1}^{(0)} & & & \\ \vdots & \vdots & & & T_m^{(0)} & & \\ \vdots & \vdots & & T_{m-1}^{(1)} & & & \\ T_{-1}^{(m)} & \vdots & \iddots & & & & \\ & T_0^{(m)} & & & & & \end{array} \tag{5}$$

by a rhombus-rule. In order to avoid repeated formation of differences in (4) as far as possible, we propose recursive calculation of the differences

$$\begin{aligned} \Delta T_k^{(i)} &:= T_k^{(i)}-T_{k-1}^{(i+1)},\\ C_k^{(i)} &:= T_k^{(i)}-T_{k-1}^{(i)}. \end{aligned}$$

We easily obtain from (4) the formulae (which are equivalent to (4))

$$\begin{aligned} \Delta T_0^{(m)} &= T(h_m, x),\\ C_0^{(m)} &= T(h_m, x),\\ \Delta T_k^{(m-k)} &= \frac{C_{k-1}^{(m-k+1)}\,W_{k-1}^{(m-k+1)}}{\left(\frac{h_{m-k}}{h_m}\right)^2\Delta T_{k-1}^{(m-k)}-C_{k-1}^{(m-k+1)}}, \qquad k=1, 2, \ldots, m,\\ C_k^{(m-k)} &= \frac{\left(\frac{h_{m-k}}{h_m}\right)^2\Delta T_{k-1}^{(m-k)}\,W_{k-1}^{(m-k+1)}}{\left(\frac{h_{m-k}}{h_m}\right)^2\Delta T_{k-1}^{(m-k)}-C_{k-1}^{(m-k+1)}}, \qquad k=1, 2, \ldots, m\\ &= W_{k-1}^{(m-k+1)}+\Delta T_k^{(m-k)},\\ T_m^{(0)} &= \sum_{k=0}^{m}\Delta T_k^{(m-k)}, \end{aligned} \tag{6}$$

with the abbreviation

$$W_k^{(i)} := C_k^{(i)}-\Delta T_k^{(i-1)}\,(\equiv T_k^{(i)}-T_k^{(i-1)}).$$

These formulae are evaluated successively for $m=0, 1, 2, \ldots$. The indexing in (6) is chosen so as to indicate the actual sequence of calculations. Note further, that for programming (6) only one linear array for storing the differences $\Delta T_k^{(m-k)}$, $k=0, \ldots, m$ is needed.

As to the sequence of h_i to be employed for extrapolation, the sequence (see [7])

$$\mathfrak{H} := \left\{h_0, \frac{h_0}{2}, \frac{h_0}{3}, \frac{h_0}{4}, \frac{h_0}{6}, \frac{h_0}{8}, \ldots\right\} \tag{7}$$

can be used without increasing the sensitivity to round-off, since the expansion (1) contains only even powers of h. The sequence

$$\left\{h_0, \frac{h_0}{2}, \frac{h_0}{4}, \frac{h_0}{8}, \ldots\right\}$$

yields equal accuracy at the expense of doubling the number of operations.

Section 5 contains an ALGOL-description of this rational extrapolation method.

3. Examples

The following examples demonstrate the effectivity of the method described above. Rational extrapolation has been compared with the following most commonly used methods for solving the initial value problem for ordinary differential equations:

1. The well known method of Runge-Kutta, which gives rise to an asymptotic expansion (1) of the form

$$T(h, x)=y(x)+\tau_4(x)h^4+\tau_5(x)h^5+\cdots.$$

2. The linear multistep method based on the predictor formula of Adams-Bashforth

$$y_{i+1} = y_i + \frac{h}{1440}(4227 f_i - 7673 f_{i-1} + 9482 f_{i-2} - 6798 f_{i-3} + 2627 f_{i-4} - 425 f_{i-5}), \qquad f_i = f(x_i, y_i),$$

and the corrector formula of Adams-Moulton

$$y_{i+1}^{(j+1)} = y_i + \frac{h}{1440}(475 f(x_{i+1}, y_{i+1}^{(j)}) + 1427 f_i - 798 f_{i-1} + 482 f_{i-2} - 173 f_{i-3} + 27 f_{i-4})$$

(see, e.g. HENRICI [1], p. 194–199).

For starting this algorithm the values $f_0, \ldots, f_5$ are needed. For the sake of simplicity we have taken the values of the exact solution for this purpose. Then this method leads to an asymptotic expansion (1) of order 6

$$T(h, x) = y(x) + \tau_6(x) h^6 + \tau_7(x) h^7 + \cdots.$$

3. Finally, the extrapolation with polynomials

$$\hat{T}_m^i(h) = a_0^{(i)} + a_1^{(i)} h^2 + \cdots + a_m^{(i)} h^{2m},$$

$$\hat{T}_m^i(h_j) = T(h_j, x), \qquad j = i, i+1, \ldots, i+m,$$

where $T(h, x)$ again denotes the values yielded by the modified midpoint-rule (4), was considered (for the recursion formulae see [7]).

For measuring the amount of work involved in these methods we have taken the number A of evaluations of the righthand side $f(x, y)$ of the differential equation. The stepsize h for the methods 1. and 2. has been chosen maximal so as to keep the (relative) error below 10^{-11} after a simple step with this stepsize. Further note, that the accuracy obtained with these methods (1. and 2.) has not been controlled by such devices as comparing the results for the stepsizes h and $h/2$ (e.g. Runge's method). This control would have doubled the number A. On the other hand, an accuracy control has been introduced into the extrapolation methods into a natural way: One compares two successive values $T_{m-1}^{(0)}$ and $T_m^{(0)}$ and increases the index m, till this difference is small enough. Clearly this procedure will also increase the number A by some factor about $\sqrt{2}$.

The computations have been performed on the PERM computer at the Technische Hochschule München (word length: 40 bit in the mantissa). The following examples have been used for comparison

a) $y' = -y, \qquad x_0 = 0, \qquad y(x_0) = 1,$

$y(x) = e^{-x}.$

This example has been taken in order to test the stability of the extrapolation procedure, since the midpoint-rule itself is a weakly unstable method.

The results are shown in Table 1. There, the *relative* error ε_{rel} and the number $A(x)$ needed for the calculation of $T(h, x)$ from 0 up to x for the various methods are tabulated.

Table 1. *Relative errors by integrating* $y' = -y$

	Runge-Kutta $h = 10^{-3}$ *		Adams-Moulton-Bashforth $h = 2_{10}-2$ *		Extrapolation with			
					Polynomials $h_0 = 0.5$		Rational functions $h_0 = 0.5$	
x	A	ε_{rel}	A	ε_{rel}	A	ε_{rel}	A	ε_{rel}
0	0	0	0		0	0	0	0
5	2000	$2.6_{10}-10$	500	$2.1_{10}-11$	480	$2.2_{10}-11$	330	$4.0_{10}-11$
10	4000	$5.2_{10}-8$	1000	$1.7_{10}-10$	980	$9.7_{10}-11$	660	$8.2_{10}-11$
15	6000	$2.1_{10}-5$	1500	$3.0_{10}-10$	1470	$1.5_{10}-10$	990	$1.2_{10}-10$
20	8000	$1.6_{10}-3$	2000	$3.4_{10}-9$	1960	$2.1_{10}-10$	1320	$1.6_{10}-10$

* It is well-known that for $y' = -y$ it is the best strategy to retain a constant stepsize throughout the computation (see e.g. [1], p. 104–105).

b) Euler's equation of motion for a rigid body without external forces

$$y' = \begin{pmatrix} y_1' \\ y_2' \\ y_3' \end{pmatrix} = \begin{pmatrix} y_2\, y_3 \\ -y_1\, y_3 \\ -k^2\, y_1\, y_2 \end{pmatrix}, \qquad x_0 = 0, \qquad y(x_0) = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \qquad k^2 = 0.51,$$

Solution:

$$y(x) = \begin{pmatrix} \operatorname{sn}(x; k) \\ \operatorname{cn}(x; k) \\ \operatorname{dn}(x; k) \end{pmatrix}.$$

Results are shown in Table 2. There, the maximal *absolute* error $\varepsilon_{\max}$ of the three components is tabulated.

Table 2. *Absolute errors by integrating Euler's equation*

	Runge-Kutta $h = 5_{10}-3$		Adams-Moulton-Bashforth $h = 10^{-4}$		Extrapolation with			
					Polynomials $h_0 = 1$		Rational functions $h_0 = 1$	
x	A	$\varepsilon_{\max}$	A	$\varepsilon_{\max}$	A	$\varepsilon_{\max}$	A	$\varepsilon_{\max}$
0	0	0	0	0	0	0	0	0
10	8000	$1.6_{10}-9$	2000	$3.5_{10}-10$	858	$1.0_{10}-10$	794	$8.9_{10}-11$
20	16000	$2.3_{10}-9$	4000	$3.0_{10}-9$	1780	$2.5_{10}-10$	1620	$2.5_{10}-10$
30	24000	$1.6_{10}-9$	6000	$5.5_{10}-9$	2638	$1.2_{10}-10$	2414	$4.2_{10}-10$
40	32000	$4.8_{10}-8$	8000	$2.5_{10}-9$	3496	$4.7_{10}-10$	3176	$3.1_{10}-10$
50	40000	$1.2_{10}-7$	10000	$1.4_{10}-9$	4386	$6.1_{10}-10$	4020	$1.2_{10}-10$
60	48000	$2.4_{10}-7$	12000	$7.1_{10}-9$	5276	$1.2_{10}-9$	4860	$2.6_{10}-10$

Examples a) and b) clearly show the superiority of rational extrapolation.

The next example is more difficult in nature. It arises from the astronomical three body-problem of the system Sun-Jupiter-8th moon of Jupiter. This example has been chosen for the following reason: it has been taken as test problem for other high accuracy methods for solving differential equations, such as the Runge-Kutta-Fehlberg method [4] and the method of LIE series due to GRÖBNER (see FILIPPI [5] and REUTTER, KNAPP [9]).

The differential equations for the three body-problem are

$$\dot{x}_m = u_m,$$

$$\dot{u}_m = -\frac{m_2}{|x_m|^3} x_m + m_3 \left\{ \frac{x_s - x_m}{|x_s - x_m|^3} - \frac{x_s}{|x_s|^3} \right\},$$

with

$$x_m = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \qquad u_m = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix};$$

the constants m_2 and m_3, the complicated vector function $x_s(t)$ and the starting values being given in [5, 9].

Table 3. *Three body problem*

	t	A	x_1	x_2	u_1	u_2	u_3
Starting values	0	0	−0.185921 387400	$0.712376\,370000_{10}-2$	$0.775628\,307000_{10}-1$	$0.206230\,159000_{10}-3$	$0.894287\,280000_{10}-3$	$-0.335610\,452000_{10}-3$
results:								
$h_0=10$	100	210	−0.128523 007306	$0.859964\,263800_{10}-1$	$0.301406\,208017_{10}-1$	$0.996344\,855954_{10}-3$	$0.596281\,897038_{10}-3$	$-0.604248\,626094_{10}-3$
$h_0=33.33\ldots$	100	99	...08	$\ldots 3772_{10}-1$	$\ldots 008_{10}-1$	$\ldots 947_{10}-3$	$\ldots 017_{10}-3$	$\ldots 098_{10}-3$
$h_0=50$	100	98	...7299	$\ldots 3781_{10}-1$	$\ldots 07979_{10}-1$	$\ldots 976_{10}-3$	$\ldots 002_{10}-3$	$\ldots 090_{10}-3$
$h_0=100$	100	105	...09	$\ldots 3785_{10}-1$	$\ldots 08079_{10}-1$	$\ldots 970_{10}-3$	$\ldots 024_{10}-3$	$\ldots 087_{10}-3$

Table 3 contains the results obtained by rational extrapolation for different basic stepsizes h_0. Again, a comparison with the results of [5] and [9] shows that rational extrapolation gives comparable accuracy, whereas the labor involved has been reduced.

4. Automatic Stepsize Correction

Since the sensitivity to round-off of the extrapolation process increases with the order of extrapolation, it is useful to limit the number of columns of (5) to be evaluated. Thus, the program of section 5 computes only $T_k^{(i)}$ for $k \leqq 6$ which has proved reasonable for a machine accuracy of 40 bits in the mantissa and for $m > 6$ the elements $T_6^{(m-6)}$ of (5) are taken as successive approximations.

Since by the results of [7], the errors $T_6^{(m-6)} - T(0, x)$ are of the type

$$T_6^{(m-6)} - T(0, x) = O(h_{m-6}^2 \ldots h_m^2),$$

a new stepsize $\bar{h}_0$ can be evaluated by the equation

$$\bar{h}_0^2 \ldots \bar{h}_6^2 = h_{m-6}^2 \ldots h_m^2. \tag{8}$$

If $\bar{h}_0$ is taken as the basic stepsize for the next step, then it will be probable that in this step $T_6^{(0)}$ will be a sufficient approximation. Thus an automatic stepsize correction can be easily built in into the extrapolation method. However, it should be noted that a correction of h_0 is not so important for the accuracy of calculations as it is the case with methods of fixed order: It is typical

for extrapolation methods that the actual stepsize h_j (not the basic stepsize h_0) is automatically reduced to h_{j+1} if the approximation is not good enough. This argument shows that apart from extreme cases it is not necessary in principle to reduce the basic step h_0. Nevertheless, it may be profitable to increase h_0, if it is too small.

If a correction of h_0 is wanted then it is not advisable to determine $\bar{h}_0$ by the complicated equation (8); for the sequence $\mathfrak{F}$ the simple rule

$$\bar{h}_0 = h_0 \cdot 0.9 \cdot (0.6)^{m-7}, \qquad m \geqq 7$$

is sufficient. This rule has been chosen, since the quotient h_{j+1}/h_j is approximately 0.6 and the computation of $T_6^{(1)}$ is needed in order to control the accuracy of $T_6^{(0)}$. If $m < 7$, i.e., if presumably h_0 is too small and should be magnified, the above argumentation via (8) breaks down. As a rule of thumb, the choice

$$\bar{h}_0 = h_0 \cdot 1.5, \qquad m < 7$$

is recommended.

As an example* take the following system of differential equations arising from the restricted problem of three bodies (earth-moon-spaceship, see FEHLBERG [3], FILIPPI [6])

$$\ddot{x} = x + 2\dot{y} - \mu' \frac{x+\mu}{[(x+\mu)^2+y^2]^{\frac{3}{2}}} - \mu \frac{x-\mu'}{[(x-\mu')^2+y^2]^{\frac{3}{2}}}, \qquad \mu = \frac{1}{82.45}$$

$$\ddot{y} = y - 2\dot{x} - \mu' \frac{y}{[(x+\mu)^2+y^2]^{\frac{3}{2}}} - \mu \frac{y}{[(x-\mu')^2+y^2]^{\frac{3}{2}}}, \qquad \mu' = 1 - \mu \tag{9}$$

initial values:

$$t_0 = 0, \quad x_0 = 1.2, \quad \dot{x}_0 = 0,$$

$$y_0 = 0, \quad \dot{y}_0 = -1.04935\,750983.$$

The solution $x(t)$, $y(t)$ is a closed orbit with period $T = 6.19216\,9331396\ldots$. As initial stepsize $h_0 = 0.2$ has been chosen. The next stepsizes were calculated according to the above rules. Fig. 1 shows the points of the orbit determined by the machine and illustrates the effect of automatic step size correction. After one

* The authors wish to thank S. FILIPPI who communicated this example to them.

period the values found by the computer were

$$x = 1.1999\,9999\,672, \qquad \dot{x} = -0.0000\,0000\,326,$$
$$y = 0.0000\,0000\,033, \qquad \dot{y} = -1.0493\,5750\,691.$$

In order to obtain these results the machine needed 48 steps with 5218 evaluation of the righthand side of the differential equation (9). The example shows that about three digits were lost during the computation.

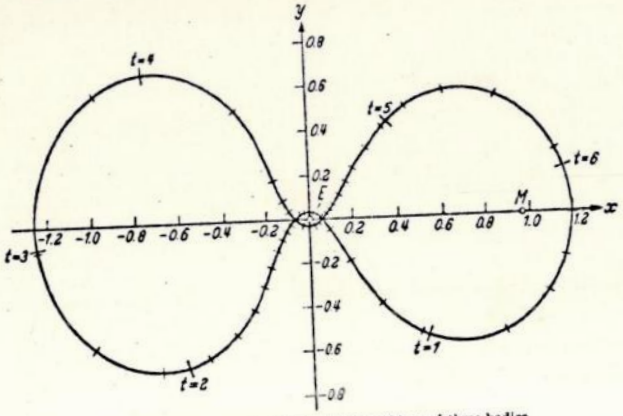


Fig. 1. Periodic orbit for the restricted problem of three bodies

5. ALGOL procedure

The previous examples have been computed with the following ALGOL program.

```
procedure diffsys          (n)
          initial values:  (x, y)
          basic stepsize:  (h0)
          error bounds:    (eps, s)
          procedure:       (f)
          exit:            (exit);
value n, eps; real x, h0, eps; integer n; procedure f; array y, s; label exit;
comment diffsys performs one integration step with a stepsize h≦h0 for a
        system of n first order ordinary differential equations of the form
        dz/dx = f(x, z), the righthand side of which must be given as a pro-
        cedure with the heading
                procedure f(x, z) result: (dz)
                value x    real x    array z, dz
        where the arrays z, dz are supposed to be of the format z, dz[1: n].
        The program takes the first of the numbers h0, h0/2, h0/4, ..., as step
        size h for which no more than 9 extrapolation steps are needed to
        obtain a sufficiently accurate result. If h ≠ h0 the program is left by
        the exit exit.
```



```
        x and y[1: n] are the initial values. After leaving the procedure,
        the original values of the parameters x and y are replaced by x+h
        and y(x+h), respectively. Also h0 may be changed (automatic step
        size correction). The output value of h0 is chosen so as to be the
        (presumably) optimal step size for the next integration step. The
        array s[1: n] and the constant eps are used to control the accuracy
        of the computed values: The procedure is left, if for all i = 1, 2, ..., n
        two successive values for y[i] differ at most by an amount of eps × s[i].
        eps should be not smaller than 10^(-D+3), where D is the number
        of digits of the machine number representation. For the first integra-
        tion step it is advisable to set s[i] = 0. After leaving, the array s[1: n]
        is changed to s_i := max_{ξ∈[x, x+h]} {s_i, |y_i(ξ)|};
begin real a, b, b1, c, g, u, v, ta, fc; integer i, j, k, kk, jj, l, m, r, sr;
    array ya, yl, ym, dy, dz[1: n], dl[1: n, 0: 6], d[0: 6], yg, yh[0: 7, 1: n];
    Boolean konv, bo, bh, fin;
    f(x, y, dz); bh := fin := false; for i := 1 step 1 until n do ya[i] := y[i];
an: a := h0 + x; fc := 1.5; bo := false; m := 1; r := 2; sr := 3; jj := -1;
    for j := 0 step 1 until 9 do
    begin
        if bo then begin d[1] := 16/9; d[3] := 64/9; d[5] := 256/9 end
            else begin d[1] := 9/4; d[3] := 9; d[5] := 36 end;
        if j > 2 then konv := true else konv := false;
        if j > 6 then begin l := 6; d[6] := 64; fc := .6 × fc end
            else begin l := j; d[l] := m × m end;
        m := m × 2; g := h0/m; b := g × 2;
      if bh ∧ j < 8 then
        begin
            for i := 1 step 1 until n do
            begin ym[i] := yh[j, i]; yl[i] := yg[j, i] end
        end
      else
        begin
            kk := (m - 2)/2; m := m - 1;
            for i := 1 step 1 until n do
            begin yl[i] := ya[i]; ym[i] := ya[i] + g × dz[i] end;
            for k := 1 step 1 until m do
            begin
                f(x + k × g, ym, dy);
                for i := 1 step 1 until n do
                begin
                    u := yl[i] + b × dy[i]; yl[i] := ym[i]; ym[i] := u;
                    u := abs(u); if u > s[i] then s[i] := u
                end;
                if k = kk ∧ k ≠ 2 then
                begin
                    jj := 1 + jj; for i := 1 step 1 until n do
```

```
                begin yh[jj, i] := ym[i]; yg[jj, i] := yl[i] end
            end
        end
    end;
        f(a, ym, dy);
        for i := 1 step 1 until n do
        begin
            v := dt[i, 0]; ta := c := dt[i, 0] := (ym[i] + yl[i] + g × dy[i])/2;
            for k := 1 step 1 until l do
            begin
                b1 := d[k] × v; b := b1 − c; u := v;
                if b ≠ 0 then
                begin b := (c − v)/b; u := c × b; c := b1 × b end;
                v := dt[i, k]; dt[i, k] := u; ta := u + ta
            end;
            if abs(y[i] − ta) > eps × s[i] then konv := false; y[i] := ta
        end;
        if konv then goto end;
        d[2] := 4; d[4] := 16; bo := ¬ bo; m := r; r := sr; sr := m × 2
    end;
    bh := ¬ bh; fin := true; h0 := h0/2; goto anf;
end: h0 := fc × h0; x := a; if fin then goto exit;
end diffsys
```

The following program is to illustrate the typical use of *diffsys*. It refers to the integration of $y' = y$, $x_0 = 0$, $y_0 = 1$, in the interval [0, 10].

```
begin
    integer n;
    n := 1;
    begin real h0, eps, x; array s, y[1:n];
    procedure f(x, z) result:(dz);
    value x; real x; array z, dz;
    begin
        dz[1] := z[1];
    end;
    procedure diffsys ...;
        h0 := 0.5; eps := 10−8;
        x := 0; y[1] := 1; s[1] := 0;
    marke: diffsys(n, x, y, h0, eps, s, f, exit);
    exit: print(x, y[1]);
        if x < 10 then goto marke;
    end
end
```


References

[1] HENRICI, P.: Discrete variable methods. New York: Wiley 1962.

[2] GRAGG, W.: Repeated extrapolation to the limit in the numerical solution of ordinary differential equations. Thesis UCLA (1963).

[3] FEHLBERG, E.: Runge-Kutta type formulas of high-order accuracy and their application to the numerical integration of the restricted problem of three bodies. Colloque international des Techniques de Calcul Analogique et Numérique en Aéronautique à Liège, 1963.

[4] — New high-order Runge-Kutta formulas with an arbitrarily small truncation error. To appear in ZAMM 45, (1965).

[5] FILIPPI, S.: Angenäherte Lösung eines astronomischen Drei-Körperproblems. Teil II. Elektronische Datenverarbeitung, Heft 6, 264–268 (1963).

[6] — Zum Verfahren von Runge-Kutta-Fehlberg. MTW 11, 147–153 (1964).

[7] BULIRSCH, R., u. J. STOER: Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus. Num. Math. 6, 413–427 (1964).

[8] STETTER, H. J.: Asymptotic expansions for the error of discretization algorithms for non-linear functional equations. Num. Math. 7, 18–31 (1965).

[9] REUTTER, F., u. J. KNAPP: Untersuchungen über die numerische Behandlung von Anfangswertproblemen gewöhnlicher Differentialgleichungssysteme mit Hilfe von Lie-Reihen und Anwendungen auf die Berechnung von Mehrkörperproblemen. Köln u. Opladen: Westdeutscher Verlag 1964.

[10] COREY: The American Math. Monthly 13 (1906).

[11] RICHARDSON, C., and J. GAUNT: The deferred approach to the limit. Trans. Roy. Soc. London 226, 300–361 (1927).

[12] BOLTON, H. C., and H. I. SCOINS: Eigenvalues of differential equations by finite-difference methods. Proc. of the Cambridge Phil. Soc. 52, 215–229 (1956).

[13] BAUER, F. L., H. RUTISHAUSER, and E. STIEFEL: New aspects in numerical quadrature. Proc. of Symposia in Applied Mathematics 15, 199–218, Am. Math. Soc. (1963).

[14] RUTISHAUSER, H.: Ausdehnung des Rombergschen Prinzips. Num. Math. 5, 48–54 (1963).

[15] LAURENT, P. J.: Étude de procédés d'extrapolation en analyse numérique. Grenoble: Thèse presentée à la Faculté des Sciences de l'Université de Grenoble 1964.

[16] LYNESS, J. N., and B. J. J. MCHUGH: Integration over multidimensional hypercubes, I. A progressive procedure. The Computer J. 6, 264–270 (1963).

[17] MEIR, A., and A. SHARMA: On the method of Romberg quadrature. J. SIAM Numer. Anal. Ser. B, 2, 250–258 (1965).

[18] FILIPPI, S.: Das Verfahren von Romberg-Stiefel-Bauer als Spezialfall des allgemeinen Prinzips von Richardson, II. Teil. MTW 11, 98–100 (1964).

[19] KOMMERELL, K.: Das Grenzgebiet der elementaren und höheren Mathematik. Leipzig: K. F. Koehler Verlag 1936.

Mathematisches Institut
der Technischen Hochschule
8 München 2, Arcisstr. 21

Numerische Mathematik 7, 18–31 (1965)

Asymptotic Expansions for the Error of Discretization Algorithms for Non-linear Functional Equations

By
HANS J. STETTER*

§ 1. Introduction

Assume that the solution $\eta(h)$ of a finite algorithm depending upon a parameter $h>0$ converges for $h\to 0$ to the solution y of a certain infinitesimal problem. We consider *asymptotic expansions of the discretization error* $\varepsilon(h):=\eta(h)-y$:

$$\varepsilon(h)=\varepsilon_1 h^{p_1}+\cdots+\varepsilon_N h^{p_N}+\bar{\varepsilon}^N(h) \quad \text{with} \quad \|\bar{\varepsilon}^N(h)\|=o(h^{p_N}) \tag{1.1}$$

where $0<p_1<\cdots<p_N$ and the ε_ν do not depend on h.

These asymptotic expansions form the basis for the so-called *Richardson-extrapolation*: The desired value $\eta(0):=\lim_{h\to 0}\eta(h)=y$ is approximated by extrapolation from several values $\eta(h_\mu)$, $h_\mu>0$. Except in the case of the Euler-Maclaurin sum formula representing the expansion (1.1) for the approximation of definite integrals by trapezoidal sums, the existence of an asymptotic expansion and its sequence of exponents $\{p_\nu\}$ had only been conjectured in applications of Richardson-extrapolation to functional equations. Quite recently, GRAGG treated the case of initial value problems for first order differential equations (see [7]).

In § 2 of this paper, we will — under suitable conditions — prove the existence of such expansions (usually with $p_\nu=p+\nu-1$) for a very general class of discretization algorithms for non-linear functional equations in Banach-spaces. In the proof, the sequence $\{\varepsilon_\nu\}$ will be recursively constructed. If the expansion of the local discretization error (see (2.2)) contains only even powers of h, this fact is preserved in the expansion of $\varepsilon(h)$. In these cases, Richardson-extrapolation is particularly effective in improving the numerical results.

In § 3, we will apply our abstract theorem to several important functional equations and their discretizations: Initial and boundary value problems for both ordinary and partial differential equations, integral equations and integrodifferential equations. For all these infinitesimal functional equations our theorem will provide hypotheses under which the application of Richardson-extrapolation is justified for a given discretization algorithm.

In § 4, we will actually compute the first terms of the expansion (1.1) for a non-linear boundary value problem of the third kind by the methods displayed

* The research reported in this paper has been sponsored in part by the United States Air Force under Grant AF EOAR 63-77 and monitored by the European Office, OAR.

in § 2 and see that the actual error of the numerical solution of the problem is well represented.

Historical remarks on Richardson-extrapolation: While the original suggestion of RICHARDSON and GAUNT [1] is almost 40 years old, systematic investigations have begun quite recently. They started with ROMBERG's well-known extrapolation for trapezoidal sums, with $h_{\mu+1}=h_\mu/2$ (see [2], [3], et al.). In the meantime, LAURENT [4] and BULIRSCH [5] have considered more general sequences $\{h_\mu\}$, RUTISHAUSER [6] has applied the general principle to other approximation processes, and GRAGG [7] has carefully investigated the basis for Richardson-extrapolation in the numerical solution of ordinary differential equations. While polynomial extrapolation has been used in all the above instances, recent investigations by BULIRSCH and STOER [8] have shown that rational extrapolation will normally give better results.

§ 2. The asymptotic expansion

2.1. Preparations

We consider functional equations

$$F(y)=0 \tag{2.1a}$$

with side conditions (e.g. initial or boundary conditions)

$$R(y)=0, \tag{2.1b}$$

where $F\colon D^1\to E^1$ and $R\colon D^2\to E^2$ are two generally nonlinear operators from subspaces D^1 and D^2 of a Banach-space E into Banach-spaces E^1 and E^2. We will always assume that (2.1) has a unique solution $y\in D<D^1\cap D^2$.*

For the purpose of numerical solution the problem (2.1) is *discretized* in the following sense:

We define families — depending on a real parameter $h\in H:=(0,h_0]$, with $h_0>0$ fixed — of B-spaces E_h, E_h^1, E_h^2 and of linear transformations Δ_h, Δ_h^1, Δ_h^2 which map E, E^1, E^2 into E_h, E_h^1, E_h^2 resp.

Then we choose two families of (nonlinear) operators $\Phi_h\colon E_h\to E_h^1$ and $\mathsf{P}_h\colon E_h\to E_h^2$ such that for $z\in D$ and $h\in H$

$$\Phi_h(\Delta_h z)=h^{n_1}\cdot\Delta_h^1\Big\{F(z)+\sum_{\nu=p}^{N}h^\nu\cdot f_\nu(z)\Big\}+O(h^{n_1+N+1}), \tag{2.2a}$$

$$\mathsf{P}_h(\Delta_h z)=h^{n_2}\cdot\Delta_h^2\Big\{R(z)+\sum_{\nu=p}^{N}h^\nu\cdot r_\nu(z)\Big\}+O(h^{n_2+N+1}), \tag{2.2b}$$

where $f_\nu\colon D\to E^1$ and $r_\nu\colon D\to E^2$ do not depend upon h.

The expressions $\Phi_h(\Delta_h y)$ and $\mathsf{P}_h(\Delta_h y)$ formed with the solution y of (2.1) are often called the *local discretization errors* of (Φ_h,P_h). n_1, n_2, and $p\geqq 1$ are suitably chosen integers (comp. the applications in § 3).

The original problem (2.1) is now replaced by the *"algorithm"*

$$\Phi_h(\eta)=0, \qquad \mathsf{P}_h(\eta)=0, \tag{2.3}$$

* Naturally we could regard $(F,R)\colon D\to E^1\times E^2$. However, the discretizations of F and R have, in general, different stability properties (see below).

which is supposed to have a unique solution $\eta(h)\in E_h$ for $h\in H$. The *global discretization error* of (2.3) is defined as

$$\varepsilon(h) := \eta(h) - \Delta_h y \in E_h$$

where y is again the solution of (2.1).

(2.3) is *convergent of order* p ($p \geqq 1$) if

$$\|\varepsilon(h)\| \leqq C h^p \quad \text{for} \quad h\in H. \tag{2.4}$$

$\|..\|$ will always denote the norm of E_h.

The global discretization error $\varepsilon(h)$ admits an *asymptotic expansion to the order* N ($N \geqq p$) if there are $e_\nu \in E$, $\nu = p(1)N$, e_ν independent of h, such that

$$\left\| \varepsilon(h) - \Delta_h \sum_{\nu=p}^{N} h^\nu \cdot e_\nu \right\| \leqq C_N h^{N+1} \quad \text{for} \quad h\in H. \tag{2.5}$$

Φ_h and P_h will always be assumed to possess at least one Frechet-derivative. We will call (2.3) *m_1, m_2-stable for $z\in E$*, z fixed, if there is a constant S independent of h such that each solution $\varepsilon\in E_h$ of

$$\Phi_h'(\Delta_h z)\,\varepsilon = \varphi, \qquad P_h'(\Delta_h z)\,\varepsilon = \varrho \tag{2.6a}$$

satisfies ($\|.\|_i$ is the norm of E_h^i, $i=1,2$)

$$\|\varepsilon\| \leqq S[h^{-m_1}\|\varphi\|_1 + h^{-m_2}\|\varrho\|_2] \quad \text{for} \quad h\in H. \tag{2.6b}$$

(2.6) is a natural extension of many of the usual concepts of stability as we will see in § 3.

Remarks. 1. Compared to the formulation of (1.1) we have now restricted ourselves to the case of integer p_ν. More general sequences $\{p_\nu\}$ can be treated similarly (the p_ν are always rational).

2. In some applications it is necessary to subdivide F, R, Φ_h, P_h, and the corresponding B-spaces because the stability properties of various parts of Φ_h and P_h differ from each other. In these cases, m_1 and m_2 are vectors and (2.6b) is modified in an obvious manner.

3. A stability concept similar to (2.6) is found in [9], but with an important restriction: CHÜN calls (2.3) stable only if (2.6) holds *for all* $z\in E$, this is rarely the case in applications to non-linear problems.

4. Naturally the existence of a "local" expansion (2.2) and the size of N depend on the differentiability properties of (F, R), the solution y, and of (Φ_h, P_h). Actually one has a sequence of subspaces $D_\nu \subset E$ (which contain elements with certain differentiability properties) and corresponding subspaces E_ν^1 and E_ν^2, in particular with respect to assumption c of Theorem 1. This situation is displayed with the example of sect. 3.1.

2.2. *The existence of an asymptotic expansion*

Theorem 1. Let

a) an expansion (2.2), with $N \geqq p$, hold for (Φ_h, P_h);

b) (2.3) be n_1, n_2-stable for the solution y of (2.1), with n_1 and n_2 from (2.2);

c) (2.3) be convergent of order $p \geqq 1$, with p from (2.2);

d) the operators in (2.2) be M-times Frechet-differentiable at y, with $M \geqq (N+1)/p$;

e) $F'(y)e = b \in E^1$, $R'(y)e = c \in E^2$ have a unique solution $e\in D$.

Then the global discretization error of (Φ_h, P_h) possesses an asymptotic expansion (2.5) to the order N.

*Proof**. We will determine $b_\nu \in E^1$ and $c_\nu \in E^2$, $\nu = p(1)N$, such that the solutions e_ν of

$$F'(y)\,e_\nu = b_\nu, \qquad R'(y)\,e_\nu = c_\nu \tag{2.7}$$

satisfy (2.5). By assumption e, the e_ν are uniquely determined and $e_\nu \in D$. At first, b_ν and c_ν remain arbitrary elements from E^1 and E^2 resp.

We consider

$$s^N(h) := \sum_{\nu=p}^{N} h^\nu \cdot e_\nu,$$
$$\bar{\varepsilon}^N(h) := \varepsilon(h) - \Delta_h s^N(h). \tag{2.8}$$

Naturally $\|\Delta_h s^N(h)\| = O(h^p)$, hence by assumption c and (2.8)

$$\|\bar{\varepsilon}^N(h)\| = O(h^p). \tag{2.9}$$

We now form $\Phi_h'(\Delta_h y)\,\bar{\varepsilon}^N(h)$. (The argument $\Delta_h y$ of the multilinear operators $\Phi_h^{(\mu)}(\Delta_h y): E_h \times \cdots \times E_h \to E_h^1$ as well as the parameter h with ε, $\bar{\varepsilon}^N$ and s^N will from now on be omitted.)

$$\begin{aligned}\Phi_h'\bar{\varepsilon}^N &= \Phi_h'(\Delta_h s^N + \bar{\varepsilon}^N) - \Phi_h'\Delta_h s^N \\ &= -[\Phi_h(\Delta_h y + \Delta_h s^N + \bar{\varepsilon}^N) - \Phi_h(\Delta_h y) - \Phi_h'(\Delta_h s^N + \bar{\varepsilon}^N)] \\ &\quad - \Phi_h(\Delta_h y) - \Phi_h'\Delta_h s^N,\end{aligned}$$

since $\Phi_h(\Delta_h y + \Delta_h s^N + \bar{\varepsilon}^N) = \Phi_h(\eta(h)) = 0$.

By assumption d and (2.2) we have through the linearity of Δ_h and Δ_h^1 for $e\in D$

$$\Phi_h^{(\mu)}(\Delta_h e)^\mu = h^{n_1}\Delta_h^1\left\{F^{(\mu)}(y)\,e^\mu + \sum_{\nu=p}^{N} h^\nu f_\nu^{(\mu)}(y)\,e^\mu\right\} + O(h^{n_1+N+1}), \quad \mu = 1(1)M; \tag{2.10}$$

it is clear that under our differentiability assumptions the order in h of the remainder term is not affected by the differentiation. (2.10) implies

$$\|\Phi_h^{(\mu)}\| = O(h^{n_1}), \qquad \mu = 1(1)M. \tag{2.11}$$

* The structure of this proof was suggested by some proofs in the doctoral dissertation [7] of W. GRAGG whom I wish to thank for interesting discussions on the subject.

Assumption d permits the use of the generalized Taylor-expansion

$$\begin{aligned}
&\Phi_h(\eta(h)) - \Phi_h(\Delta_h y) - \Phi_h'(\Delta_h s^N + \bar\varepsilon^N)\\
&= \sum_{\mu=2}^{M-1} \frac{1}{\mu!}\, \Phi_h^{(\mu)}(\Delta_h s^N + \bar\varepsilon^N)^\mu + O(h^{n_1+Mp}) \quad \text{by (2.11) and assumption } c\\
&= \sum_{\mu=2}^{M-1} \frac{1}{\mu!}\, \Phi_h^{(\mu)}(\Delta_h s^N)^\mu + O\left(h^{n_1+p}\|\bar\varepsilon^N\|\right) + O(h^{n_1+N+1}) \quad \text{by (2.9) and } Mp \geqq N+1\\
&= h^{n_1}\left\{\Delta_h^1 \sum_{\nu=2p}^{N} h^\nu g_\nu(y, e_p, \ldots, e_{\nu-p}) + O\left(h^p\|\bar\varepsilon^N\|\right) + O(h^{N+1})\right\},
\end{aligned}$$

where we have defined the nonlinear operators g_ν via (2.10) by

$$\sum_{\mu=2}^{M-1} \frac{1}{\mu!}\left[F^{(\mu)}(y) + \sum_{\lambda=p}^{N} h^\lambda f_\lambda^{(\mu)}(y)\right]\left(\sum_{\nu=p}^{N} h^\nu e_\nu\right)^\mu =: \sum_{\nu=2p}^{N} h^\nu g_\nu + O(h^{N+1}). \tag{2.12}$$

It is easily seen from (2.12) that the g_ν depend only on e_μ's with $\mu \leqq \nu - p$.

For the remaining parts of $\Phi_h'\bar\varepsilon^N$ we obtain from (2.2) and (2.1)

$$\Phi_h(\Delta_h y) = h^{n_1}\left\{\Delta_h^1 \sum_{\nu=p}^{N} h^\nu f_\nu(y) + O(h^{N+1})\right\} \tag{2.13}$$

and from (2.10) and (2.7)

$$\begin{aligned}
\Phi_h'\Delta_h s^N &= h^{n_1}\left\{\Delta_h^1 \sum_{\nu=p}^{N} h^\nu\left[b_\nu + \sum_{\lambda=p}^{N} h^\lambda f_\lambda'(y)\, e_\nu\right] + O(h^{N+1})\right\}\\
&= h^{n_1}\left\{\Delta_h^1 \sum_{\nu=p}^{N} h^\nu\left[b_\nu + \sum_{\lambda=p}^{\nu-1} f_\lambda'(y)\, e_{\nu-\lambda}\right] + O(h^{N+1})\right\}.
\end{aligned} \tag{2.14}$$

Collecting the various expressions we have — with $g_\nu = 0$ for $\nu < 2p$ —

$$\begin{aligned}
\Phi_h'\bar\varepsilon^N = -\,h^{n_1}\Bigg\{&\Delta_h^1 \sum_{\nu=p}^{N} h^\nu\left[g_\nu(y, e_p, \ldots, e_{\nu-p}) + f_\nu(y) + \sum_{\lambda=p}^{\nu-1} f_\lambda'(y)\, e_{\nu-\lambda} + b_\nu\right]\\
&+ O\left(h^p\|\varepsilon^N\|\right) + O(h^{N+1})\Bigg\}.
\end{aligned} \tag{2.15a}$$

Completely analogously we obtain

$$\begin{aligned}
\mathrm{P}_h'\bar\varepsilon^N = -\,h^{n_2}\Bigg\{&\Delta_h^2 \sum_{\nu=p}^{N} h^\nu\left[t_\nu(y, e_p, \ldots, e_{\nu-p}) + r_\nu(y) + \sum_{\lambda=p}^{\nu-1} r_\lambda'(y)\, e_{\nu-\lambda} + c_\nu\right]\\
&+ O\left(h^p\|\varepsilon^N\|\right) + O(h^{N+1})\Bigg\},
\end{aligned} \tag{2.15b}$$

where

$$\sum_{\mu=2}^{M-1} \frac{1}{\mu!}\left[R^{(\mu)}(y) + \sum_{\lambda=p}^{N} h^\lambda r_\lambda^{(\mu)}(y)\right]\left(\sum_{\nu=p}^{N} h^\nu e_\nu\right)^\mu =: \sum_{\nu=2p}^{N} h^\nu t_\nu + O(h^{N+1})$$

and $t_\nu = 0$ for $\nu < 2p$.

For $\nu = p(1)N$ we can now recursively choose b_ν and c_ν which annihilate the brackets in (2.15) since the corresponding conditions for the b_ν and c_ν contain only e_μ's with $\mu < \nu$ while the ones for b_p and c_p do not contain an e_μ at all. Thus through (2.7), all the b_ν, c_ν, and e_ν are uniquely defined for $\nu = p(1)N$.

With this choice of the e_ν, (2.15) is reduced to

$$\begin{aligned}
\Phi_h'\bar\varepsilon^N &= h^{n_1}\left\{O\left(h^p\|\bar\varepsilon^N\|\right) + O(h^{N+1})\right\},\\
\mathrm{P}_h'\bar\varepsilon^N &= h^{n_2}\left\{O\left(h^p\|\bar\varepsilon^N\|\right) + O(h^{N+1})\right\}.
\end{aligned}$$

By assumption b and (2.6) we conclude from (2.9) inductively $\|\bar\varepsilon^N\| = O(h^{jp})$, $j = 2, 3, \ldots$ until jp would surpass $N+1$ and the final estimate $\|\bar\varepsilon^N\| = O(h^{N+1})$ is reached.

Remarks. 1. It is clear that the constant C_N of (2.5) may actually be determined (in terms of certain bounds on derivatives of the various operators) for each particular application of Theorem 1, although this may present a formidable task. See e.g. [5].

2. Since the unique solvability of the non-linear problem (2.1) is required to have the discretization error $\varepsilon(h)$ well defined, condition c on the analogous linear problem usually presents no difficulties. See also remark 4 of sect. 2.1.

2.3. *Expansions in h^2*

For particular algorithms (Φ_h, P_h) it may happen that the expansions (2.2) proceed by even powers of h only, p also being even. This property is inherited by the asymptotic expansion of the global discretization error.

Theorem 2. Let all assumptions of Theorem 1 hold and let f_ν and r_ν be zero operators for odd ν in (2.2).

Then the asymptotic expansion of the global discretization error of (Φ_h, P_h) contains only even powers of h.

Proof. We try, for even N,

$$s^N(h) = \sum_{\nu=p/2}^{N/2} h^{2\nu} e_{2\nu}, \tag{2.16}$$

i.e. $e_\nu = 0$ for odd ν, and check the proof of Theorem 1 for possible inconsistencies:

It is easily checked, however, that now the brackets in (2.15) vanish for odd ν, if $b_\nu = c_\nu = 0$ for ν odd; this is consistent with assumption (2.16).

Remark. To make the odd power terms vanish in (2.2) it is often necessary to choose Δ_h^1 and Δ_h^2 judiciously, see e.g. sect. 3.1 and 3.3.

§ 3. Applications

3.1. *Ordinary differential equations, initial value problems**

We consider a system of l first order differential equations

$$\begin{aligned}
F(y) &= y'(x) - G\left(y(x)\right) = 0,\\
R(y) &= y(a) - y_0 = 0,
\end{aligned} \qquad x \in [a, b]. \tag{3.1}$$

We have $E = E^1 = C_l[a, b]$ (= space of continuous functions $y: [a, b] \to R_l$ into the l-dimensional real space), $E^2 = R_l$. If the independent variable should occur

* This case has been thoroughly investigated in [7] with special methods. We use it as an introductory example and to display the influence of differentiability assumptions (comp. remark 4 of sect. 2.1).

explicitly it can be disguised as a dependent variable by adding the differential equation $x'=1$ with $x(a)=a$. $D=D^1=C_l^{(1)}[a,b]$, $D^2=C_l[a,b]$.

Let $[a,b]_h:=\left\{x: x=a+ih=:x_i,\ i=0(1)\left\lfloor\frac{b-a}{h}\right\rfloor\right\}$. We consider the discretizations Δ_h and Δ_h^1 which restrict functions from $C_l[a,b]$ to functions on $[a,b]_h$, while Δ_h^2 is the identity. Hence we have $E_h=E_h^1=\{\eta: [a,b]_h\to R_l\}$, $E_h^2=R_l$.

For simplicity we choose at first *Euler's method* as our numerical integration algorithm:

$$\begin{aligned}&\Phi_h(\eta)=(T_h-I)\eta-hG(\eta)=0, && \eta\in E_h,\\ &\mathrm{P}_h(\eta)=\eta(x_0)-y_0=0,\end{aligned}\tag{3.2}$$

where $T_h\eta(x_i):=\eta(x_{i+1})$, I the identity.

Assume that G is M-times continuously differentiable ($M\geqq 2$) in a closed region of the R_l containing the solution of (3.1) and a sufficiently large neighborhood, y is then in $C_l^{(M+1)}[a,b]$. For $z\in C_l^{(m)}[a,b]$ we have an expansion (2.2a) with $1\leqq N\leqq m-2$:

$$\begin{aligned}&(T_h-I)z(x_i)-hG(z(x_i))\\ &\quad=h\left\{[z'(x)-G(z(x))]+\sum_{\nu=1}^{N}\frac{h^\nu}{(\nu+1)!}\frac{d^\nu}{dx^\nu}z(x)\right\}_{x=x_i}+O(h^{N+2}),\end{aligned}$$

i.e.

$$l_\nu:=\frac{1}{(\nu+1)!}\frac{d^\nu}{dx^\nu}:\ C_l^{(m)}[a,b]\to C_l^{(m-\nu-1)}[a,b].$$

(2.2b) is trivial, all the r_ν vanish:

$$\eta(x_0)-y_0=h^0\{[y(a)-y_0]\}.$$

Hence we have $n_1=1$, $n_2=0$, $p=1$, and the order N of the asymptotic expansion is limited to $N\leqq M-1$.

It is well-known that the Euler-algorithm (3.2) is 1,0-stable in our sense and convergent of order 1 (see e.g. [10]). Assumption d of Theorem 1 applies only to G since all other operators in (2.2) are linear, again we obtain the restriction $N\leqq M-1$. Assumption e requires the unique solvability of

$$e'-\frac{\partial G}{\partial y}(y(x))\,e=b\in E^1,\qquad e(a)=c\in E^2,$$

which is trivial ($M\geqq 2$!).

The analysis of the recursive definition of the b_ν and e_ν (all the c_ν vanish) shows that (see (2.7) and (2.15a))

$$\begin{aligned}&b_1\in C_l^{[M-1]}[a,b] \quad\text{implies}\quad e_1\in C_l^{[M]}[a,b]\\ &\text{implies}\quad b_2\in C_l^{[M-2]}[a,b] \quad\text{implies}\quad e_2\in C_l^{[M-1]}[a,b]\\ &\qquad\cdots\qquad\qquad\cdots\\ &\text{implies}\quad b_n\in C_l^{[M-N]}[a,b] \quad\text{implies}\quad e_N\in C_l^{[M-N+1]}[a,b]\end{aligned}$$

and hence $e_\nu\in D$ for each $\nu=1(1)N$.

By Theorem 1 we conclude, for $N\leqq M-1$

$$\eta(x_i,h)=y(x_i)+\sum_{\nu=1}^{N}h^\nu e_\nu(x_i)+O(h^{N+1})\quad\text{for}\quad x_i\in[a,b]_h.$$

For *general one-step methods* (e.g. Runge-Kutta) it is often cumber-some to explicitly derive the expansion (2.2a); but only the existence of such an expansion and the consistency of the differentiability properties are needed. These are verified quite easily (see [7]).

Linear *multistep methods* (see e.g. [10], Chapter 5)

$$\sum_{\kappa=0}^{k}\alpha_\kappa\eta(x_{i+\kappa})-h\cdot\sum_{\kappa=0}^{k}\beta_\kappa G\left(\eta(x_{i+\kappa})\right)=0,\qquad \alpha_k\neq 0,$$

may be treated in the same manner after they have been formally reduced to one-step methods by regarding the vectors $\bar\eta(x_i)^T:=(\eta(x_i),\eta(x_{i+1}),\ldots,\eta(x_{i+k-1}))$.

As an example of a "symmetric" method we choose the *trapezoidal rule*

$$\Phi_h(\eta)=(T_h-I)\eta-\frac{h}{2}(T_h+I)G(\eta)=0.\tag{3.3}$$

In order to obtain an expansion (2.2a) in even powers of h we have to use the following discretization operator Δ_h^1 (see remark of sect. 2.3):

$$\Delta_h^1 z:=\frac{1}{2}(T_h+I)\Delta_h z$$

where Δ_h is the trivial discretization.

Then we have for sufficiently differentiable z and even N

$$\Phi_h(\Delta_h z)=h\Delta_h^1\left\{F(z)+\sum_{\nu=1}^{N/2}h^{2\nu}\frac{1}{(2\nu+1)!}\frac{d^{2\nu+1}}{dx^{2\nu+1}}z\right\}+O(h^{N+3})$$

and, by Theorem 2, the asymptotic expansion of the global discretization error contains only even powers of h ($p=2$). For the special case of Romberg-integration this asymptotic expansion is explicitly given by the Euler-Maclaurin sum formula (B_ν are the Bernoulli numbers):

$$\begin{aligned}&\frac{h}{2}\left[f(a)+2\sum_{i=1}^{n-1}f(x_i)+f(x_n)\right]-\int_a^{x_n}f(t)\,dt\\ &\quad=\sum_{\nu=1}^{N/2}\frac{h^{2\nu}}{2\nu!}B_{2\nu}\left[f^{(2\nu-1)}(x_n)-f^{(2\nu-1)}(a)\right]+O(h^{N+2}).\end{aligned}$$

3.2. Ordinary differential equations: boundary value problems

As an example for the application of Theorems 1 and 2 to nonlinear boundary value problems for ordinary differential equations we consider the second order equation for one function $y(x)$*

$$F(y)=-y''(x)+G(x,y(x))=0,\qquad x\in[a,b];\tag{3.4}$$

with $G_y(x,y)\geqq 0$ in a suitable region. We assume sufficient differentiability for G to justify all expansions, we will not analyse the precise requirements here.

* Systems are treated analogously.

For Δ_h and Δ_h^1 we take the trivial discretization $y(x) \to \{y(x_i)\}$, with $h = (b-a)/n$, n integer. The following symmetric algorithm for (3.4) is widely used (see e.g. [11]):

$$\Phi_h(\eta) = (T_h^{-1} + 2I - T_h)\,\eta + \frac{h^2}{2}\,(2\beta I + (1-\beta)\,(T_h^{-1} + T_h))\,G(\xi,\eta) = 0, \tag{3.5}$$

where $\xi = \Delta_h x$.

The expansion (2.2a) for (3.4)/(3.5) contains even powers of h only, for sufficiently differentiable z we have

$$\Phi_h(\Delta_h z) = h^2 \Delta_h^1 \left\{F(z) + \sum_{\nu=p/2}^{N/2} h^{2\nu} r_{2\nu}(z)\right\} + O(h^{N+4}), \tag{3.5a}$$

with $p=4$ for $\beta = \frac{5}{6}$, $p=2$ otherwise (e.g. for $\beta = 1$).

The case of *boundary conditions of the first kind:*

$$R(y) = \begin{Bmatrix} y(a) - A \\ y(b) - B \end{Bmatrix} = 0, \tag{3.6}$$

$$P_h(\eta) = \begin{Bmatrix} \eta(x_0) - A \\ \eta(x_n) - B \end{Bmatrix} = 0, \tag{3.7}$$

presents no difficulties. To establish the 2, 0-stability of (3.5)/(3.7) we have to show that the solution $\varepsilon \in R_{n+1}$ of the linear system (with y from (3.4)/(3.6))

$$\begin{aligned} \Phi_h'(\Delta_h y)\,\varepsilon &= \varphi \in R_{n-1} \quad \text{(equ. at } x_1, x_2, \ldots, x_{n-1}) \\ P\varepsilon &= \varrho \in R_2 \quad \text{(defining } \varepsilon(x_0) \text{ and } \varepsilon(x_n)) \end{aligned}$$

satisfies $\|\varepsilon\| \leqq S[h^{-2}\|\varphi\|_1 + \|\varrho\|_2]$ ($\|\cdot\|, \|\cdot\|_1, \|\cdot\|_2$ are norms in R_{n+1}, R_{n-1}, R_2 resp.). But this fact is well-known (in different terminology), see e.g. [10], Theorem 7.8.

Since the other hypotheses of Theorem 1 are satisfied there exists an asymptotic expansion in even powers of h for the difference $\varepsilon(h)$ between the exact solution η of the algorithm (3.5)/(3.7) and the solution y of (3.4)/(3.6).

To be able to replace *nonlinear boundary conditions of the third kind*

$$R(y) = \begin{pmatrix} y'(a) - C(y(a)) \\ y'(b) + D(y(b)) \end{pmatrix} = 0, \quad \begin{matrix} C' \geqq 0, \\ D' \geqq 0, \end{matrix} \quad C' + D' > 0, \tag{3.8}$$

by a symmetric boundary condition, we extend our elements $\eta \in E_h$ by values $\eta(x_{-1})$ and $\eta(x_{n+1})$* and choose

$$P_h(\eta) = \begin{pmatrix} \frac{1}{2}(T_h - T_h^{-1})\,\eta(x_0) - h \cdot C(\eta(x_0)) \\ \frac{1}{2}(T_h - T_h^{-1})\,\eta(x_n) + h \cdot D(\eta(x_n)) \end{pmatrix} = 0. \tag{3.9}$$

Assuming $z \in E$ to be extended beyond $[a,b]$ by Taylor-expansion we have for sufficiently differentiable z

$$P_h(\Delta_h z) = h\,\Delta_h^2 \left\{R(z) + \sum_{\nu=p/2}^{N/2} h^{2\nu} r_{2\nu}(z)\right\} + O(h^{N+3}). \tag{3.9a}$$

A discretization of (3.8) with $p=4$ may be obtained by using the differential equation (3.4) to knock out the h^2-term in (3.9a).

* This includes a modification of E_h and E_h^1, of course.

We will now establish the 2, 1-stability of (3.5)/(3.9). From this stability the convergence of order p of the algorithm (3.5)/(3.9) for $h \to 0$ may be deduced as usual (see e.g. [10], sect. 7.3) and by Theorems 1 and 2 we have the existence of an asymptotic expansion in terms of h^2 for the global discretization error.

We have to show that the solution $\varepsilon = (\varepsilon_{-1}, \ldots, \varepsilon_{n+1}) \in R_{n+3}$ of the linear system (y is the solution of (3.4)/(3.8))

$$\begin{aligned} \Phi_h'(\Delta_h y)\,\varepsilon &= \varphi \in R_{n+1} \quad &&\text{(equ. at } x_0, \ldots, x_n), \\ P_h'(\Delta_h y)\,\varepsilon &= \varrho \in R_2 \quad &&\text{(equ. no. } -1 \text{ and } n+1, \text{ at } x_0 \text{ and } x_n) \end{aligned} \tag{3.10}$$

satisfies $\|\varepsilon\| \leqq S\,(h^{-2}\|\varphi\|_1 + h^{-1}\|\varrho\|_2)$.

Let $\Psi = (\psi_{ik}) = \Psi^{(1)} + h^2 \cdot \Psi^{(2)}$ be the matrix of (3.10), where $h^2\Psi^{(2)}$ contains the h^2-parts of Φ_h' (see (3.5)). Then Ψ as well as $\Psi^{(1)}$ have a positive inverse (for sufficiently small h) by virtue of the row sum criterion and $\Psi - \Psi^{(1)}$ is also positive. Hence

$$\Psi^{(1)-1} - \Psi^{-1} = \Psi^{(1)-1}(\Psi - \Psi^{(1)})\,\Psi^{-1} \geqq 0$$

i.e. the elements $\bar{\psi}_{ik}$ of Ψ^{-1} are smaller than those of $\Psi^{(1)-1}$. The fact that the latter ones are all of order $O(1/h)$ in h is easily verified by explicit computation. From

$$\varepsilon_\mu = \bar{\psi}_{\mu,-1}\,\varrho_{-1} + \sum_{\nu=0}^{n} \bar{\psi}_{\mu\nu}\,\varphi_\nu + \bar{\psi}_{\mu,n+1}\,\varrho_{n+1} \quad \text{and} \quad n = O\left(\frac{1}{h}\right)$$

we have the desired result.

(This also shows that (3.9) is a sensible discrete boundary condition to be used with (3.5).)

3.3. *Partial differential equations: initial value problems*

From the variety of problems we choose as an example the Cauchy-problem for a system of quasilinear hyperbolic equations in two independent variables with two characteristic directions. Such a system may be reduced to the following normal form ($y^k = y^k(\lambda,\mu)$, see e.g. [12]):

$$\begin{cases} \sum_{k=1}^{K} a^{ik}(y)\,\dfrac{\partial}{\partial\lambda}\,y^k = 0, & i = 1(1)\,K', \\ \sum_{k=1}^{K} a^{ik}(y)\,\dfrac{\partial}{\partial\mu}\,y^k = 0, & i = (K'+1)\,(1)\,K, \end{cases} \qquad 1 \leqq K' < K \tag{3.11a}$$

in $D := \{(\lambda,\mu) : \lambda + \mu \geqq 0,\ \lambda \leqq 1,\ \mu \leqq 1\}$, with initial conditions

$$y^k(\lambda, -\lambda) - \bar{y}^k(\lambda) = 0, \quad k = 1(1)\,K, \quad \text{in } |\lambda| \leqq 1. \tag{3.11b}$$

The a^{ik} depend on the y^k but not on their derivatives, $\bar{y}^k(\lambda)$ is given.

A theory of finite-difference methods for problems of this type has been presented in [13]. Here we consider the "mean-value method":

$$\begin{cases} \sum_{k=1}^{K} a^{ik}\left(\dfrac{\eta_{l+1,m} + \eta_{l,m}}{2}\right)[\eta^k_{l+1,m} - \eta^k_{l,m}] = 0, & i \leqq K', \\ \sum_{k=1}^{K} a^{ik}\left(\dfrac{\eta_{l,m+1} + \eta_{l,m}}{2}\right)[\eta^k_{l,m+1} - \eta^k_{l,m}] = 0, & i > K', \end{cases} \tag{3.12a}$$

with

$$\eta^k_{l,-1} = \bar{y}^k(lh). \tag{3.12b}$$

Obviously, E, E^1, and E^2 are B-spaces of functions from D and $[-1,+1]$ resp. to the R_K while the functions of E_h, E^1_h, and E^2_h are from the nodes of a square mesh of mesh-size $h=1/n$, n integer; $\eta_{l,m} := \eta(lh, mh)$, l, m integer. Again we will not regard differentiability properties, in any case we have to restrict ourselves to a closed subregion $D^* \subset D$ in which the solution y of (3.11) and its derivatives are bounded. It is well-known that D^* may be much smaller than D even for highly differentiable a^{ik}.

In order to carry the symmetry of (3.12a) into the expansion (2.2a) we choose a $\Delta^1_h: E^1 \to E^1_h$ of a special structure

$$\Delta^1_h \begin{pmatrix} z^1 \\ \cdot \\ \cdot \\ \cdot \\ z^K \end{pmatrix} := \frac{1}{2} \begin{pmatrix} z^1((l+1)h, mh) + z^1(lh, mh) \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ z^{K'}((l+1)h, mh) + z^{K'}(lh, mh) \\ z^{K'+1}(lh, (m+1)h) + z^{K'+1}(lh, mh) \\ \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \cdot \\ z^K(lh, (m+1)h) + z^K(lh, mh) \end{pmatrix}.$$

It is easily verified that this choice leads to an expansion (2.2a) in even powers of h, with $n_1=1$ and $p=2$.

From [13] we know that the algorithm (3.12) converges of order 2 and that it is 1,0-stable. By Theorems 1 and 2 (the remaining hypotheses present no difficulties) it follows that there exists an asymptotic expansion in even powers of h for the global discretization error. The order of this expansion is naturally limited by the differentiability properties of the solution y.

Further applications to initial value problems for partial differential equations will be treated in a separate publication, in particular the question of m_1, m_2-stability under various norms.*

3.4. Partial differential equations: boundary value problems

The application of the theory of § 2 to boundary value problems for partial differential equations of elliptic type follows the ideas presented in sect. 3.2. A detailed investigation of some aspects of Richardson-extrapolation for problems of this kind will be presented by P. HOFMANN in his doctoral thesis.

3.5. Integral equations

Consider the nonlinear integral equation

$$F(y) = y(s) - \int_a^b K(s, t, y(t))\, dt = 0 \tag{3.13}$$

which is assumed to have a unique solution.

As a discretization algorithm for the numerical solution of (3.13) we may use the following nonlinear system for the $\eta(s_i)$:

$$\Phi_h(\eta_i) = \left(\eta(s_i) - h \sum_{j=0}^{n} \beta_j K(s_i, t_j, \eta(t_j)),\ i = 0(1)\, n\right) = 0. \tag{3.14}$$

* For partial difference equations, stability may depend upon the norm which is used, see e.g. [14].

Natural choices for the β_j are $\beta_0=\beta_n=\frac{1}{2}$, $\beta_j=1$ else (trapezoidal rule) and $\beta_0=\beta_n=\frac{1}{3}$, $\beta_{2j}=\frac{2}{3}$, $\beta_{2j-1}=\frac{4}{3}$ (Simpson-rule). Both choices yield an expansion (2.2a) in even powers of h, with $n_1=0$ and $p=2$ or 4 resp. Naturally we assume that (3.14) possesses a unique solution for sufficiently small h.

There are no boundary conditions. The definition of the B-spaces E, E^1, etc. is obvious, the discretization is the usual one.

In order to apply Theorems 1 and 2 we have to assume that the linear integral equation of the second kind

$$F'(y)\, e = e(s) - \int_a^b \frac{\partial K}{\partial y}(s, t, y(t))\, e(t)\, dt = b(t) \in C[a, b]$$

has a unique solution, i.e. that 1 is not an eigenvalue of the kernel

$$K(s,t) := \frac{\partial K}{\partial y}(s, t, y(t)).$$

From this assumption, however, it follows that for sufficiently small h the matrix of the linear system

$$\Phi'_h(\Delta_h y)\, \varepsilon = \left(\varepsilon_i - h \sum_{j=0}^{n} \beta_j K(s_i, t_j)\, \varepsilon_i = \varphi_i,\ i = 0(1)\, n\right)$$

has an inverse which is $O(1)$ in h (see [15], Theorem 1, p. 35). This is equivalent to the 0-stability of (3.14).

The convergence of (3.14) of order p can be proved under certain conditions on (3.13) (see e.g. [15]), thus we have again the result that the discretization error of the solution of (3.14) possesses an asymptotic expansion in even powers of h.

Integrodifferential equations are treated in a similar fashion.

§ 4. Examples

4.1. Euler's Method

The solution of $y'=y$, $y(0)=1$ at $x=1$ by Euler's Method is

$$\eta(1; h) = (1+h)^{1/h} = e^{\frac{1}{h}\cdot\log(1+h)} = \sum_{\nu=0}^{\infty} \frac{1}{\nu!}\left(1 - \frac{h}{2} + \frac{h^2}{3} - \frac{h^3}{4} + \cdots\right)^\nu$$

which yields after some manipulation

$$\eta(1; h) = e\left(1 - \tfrac{1}{2}h + \tfrac{11}{24}h^2 - \tfrac{7}{16}h^3 + \cdots\right).$$

By the method of sect. 2.2 we obtain for the first terms of the asymptotic expansion:

$$e_1(x) = -\frac{x}{2}e^x, \qquad e_1(1) = -\frac{1}{2}e,$$

$$e_2(x) = +\frac{x}{24}(3x+8)e^x; \qquad e_2(1) = +\frac{11}{24}e,$$

$$e_3(x) = -\frac{x}{48}(x^2+8x+12)e^x, \qquad e_3(1) = -\frac{7}{16}e.$$

4.2. A boundary value problem of the third kind

Consider the non-linear boundary value problem

$$
\begin{aligned}
&-y'' + 6\sqrt{y-x} = 0 \quad \text{in} \quad [1,2],\\
&y'(1) - y^2(1) - \frac{7}{16} = 0,\\
&y'(2) - 54/y(2) = 0,
\end{aligned}
\tag{4.1}
$$

with the solution $y(x) = x^4/4 + x$, and its discretization according to sect. 3.2

$$
\begin{aligned}
&-\eta_{i-1} + 2\eta_i - \eta_{i+1} + 6h^2\sqrt{\eta_i - x_i} = 0, \qquad i = 0(1)n,\\
&(-\eta_{-1} + \eta_1)/2 - h\left(\eta_0^2 + \frac{7}{16}\right) = 0,\\
&(-\eta_{n-1} + \eta_{n+1})/2 - 54h/\eta_n = 0,
\end{aligned}
\tag{4.2}
$$

where $n := 1/h$ integer.

The $f_{2\nu}$ and $r_{2\nu}$ of the expansions (3.5a) and (3.9a) are found to be

$$
f_{2\nu}(z) = \frac{2}{(2\nu+2)!}\, z^{(2\nu+2)}(x), \qquad r_{2\nu}(z) = \frac{1}{(2\nu+1)!}\, z^{(2\nu+2)}\Big|_{x=\xi}.
$$

The procedure of sect. 2.2 yields the following linear boundary value problems of the third kind for the first two terms e_2 and e_4 of the asymptotic expansion:

$$
\begin{aligned}
&-e_2'' + \frac{6}{x^2}\, e_2 = \frac{1}{2},\\
&e_2'(1) - \frac{5}{2}\, e_2(1) = -1, \qquad e_2'(2) + \frac{3}{2}\, e_2(2) = -2;
\end{aligned}
\tag{4.3}
$$

$$
\begin{aligned}
&-e_4'' + \frac{6}{x^2}\, e_4 = \frac{6}{x^4}\, e_2^2 + \frac{6}{x^4}\, e_2 - \frac{2}{x^3}\, e_2' - \frac{1}{4x^2},\\
&e_4' - \frac{5}{2}\, e_4 = e_2^2 - \frac{1}{2}\, e_2 + 1 \quad \text{at} \quad x = 1,\\
&e_4' + \frac{3}{2}\, e_4 = \frac{1}{4}\, e_2^2 + \frac{5}{8}\, e_2 + \frac{1}{2} \quad \text{at} \quad x = 2.
\end{aligned}
\tag{4.4}
$$

The algorithm (4.2) was solved by Newton's method for $h_\mu = \frac{1}{10}\, 2^{-\mu}$, $\mu = 0(1)3$, with the poor initial approximation $\eta(x) = 2.5\,x + 6$ (which is not too far from satisfying the boundary conditions). Then (4.3) and (4.4) were also solved numerically with sufficient accuracy, and the beginning of the asymptotic expansion was compared to the errors of the computed values η.

Except for h_0 — where the $h^6 e_6$ term is still non-negligible — the values of $\eta(x, h)$ and of $y(x) + h^2 e_2(x) + h^4 e_4(x)$ coincided within the accuracy which had been obtained for η over the whole interval $[1, 2]$ including the boundary values. Sample values obtained for $h_1 = \frac{1}{20}$ are shown in the Table below:

x	y	$y + h^2 e_2$	$y + h^2 e_2 + h^4 e_4$	$\eta(h)$
1.0	1.2500	1.2504 54381	... 53545	... 53547
1.2	1.7184	1.7185 96027	... 95699	... 95699
1.4	2.3604	2.3603 23037	... 22070	... 22969
1.6	3.2384	3.2379 91678	... 91793	... 91791
1.8	4.4244	4.4235 72511	... 72784	... 72782
2.0	6.0000	5.9986 42278	... 42710	... 42709

As to be expected, Richardson-extrapolation gives an excellent improvement on the values of η: Although the error of η for $h_2 = \frac{1}{40}$ was $1 - 3 \times 10^{-4}$, the values extrapolated from η for h_0, h_1, h_2 were correct within 2×10^{-9}!

References

[1] RICHARDSON, C., and J. GAUNT: The deferred approach to the limit. Trans. Roy. Soc. Lond. 226, 300–361 (1927).

[2] ROMBERG, W.: Vereinfachte numerische Integration. Det. Kong. Norske Videnskabers Selskab Forhandl. 28, 7, Trondheim 1955.

[3] BAUER, F. L., H. RUTISHAUSER, and E. STIEFEL: New aspects in numerical quadrature. Proceed. Symposia Appl. Math. 15, 199–218 (1963).

[4] LAURENT, P.-J.: Un théorème de convergence pour le procédé d'extrapolation de Richardson. Compt. Rend. Acad. Sc. 256, 1435–1437 (1963).

[5] BULIRSCH, R.: Bemerkungen zur Romberg-Integration. Num. Math. 6, 6–16 (1964).

[6] RUTISHAUSER, H.: Ausdehnung des Rombergschen Prinzips. Num. Math. 5, 48–54 (1963).

[7] GRAGG, W.: Repeated extrapolation to the limit in the numerical solution of ordinary differential equations, dissertation, UCLA (1963).

[8] BULIRSCH, R., and J. STOER: Über Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus. Num. Math. 6, 413 (1964).

[9] LIN CHÜN: A discussion on the difference method for the solution of nonlinear differential equations. Scient. Sin. 10, 414–419 (1961).

[10] HENRICI, P.: Discrete variable methods in differential equations. New York: Wiley 1962.

[11] FOX, L.: The numerical solution of two-point boundary problems in differential equations. Oxford: Clarendon Press 1957.

[12] SAUER, R.: Anfangswertprobleme bei partiellen Differentialgleichungen, 2. Aufl. Berlin-Göttingen-Heidelberg: Springer 1958.

[13] STETTER, H. J.: On the convergence of characteristic finite-difference methods of high accuracy for quasi-linear hyperbolic equations. Num. Math. 3, 321–344 (1961).

[14] — Maximum bounds for the solutions of initial value problems for partial difference equations. Num. Math. 5, 399–424 (1963).

[15] KANTOROVICH, L. V.: Functional analysis and applied mathematics, NBS-translation, 1952, edited by G. E. FORSYTHE.

Mathematisches Institut der Technischen Hochschule
8 München 2, Arcisstraße 21

(Received May 28, 1964)

SIAM J. NUMER. ANAL.
Ser. B, Vol 2, No. 3
1965

ON EXTRAPOLATION ALGORITHMS FOR ORDINARY INITIAL VALUE PROBLEMS*

WILLIAM B. GRAGG†

1. Introduction. The algorithm of Romberg [20], [3] and its generalizations [1], [5] for the numerical evaluation of definite integrals are based on the fact that, under suitable regularity assumptions on the integrand, the trapezoidal approximation with step h has an asymptotic expansion in powers of h^2. It is proposed in [3], [5] to apply similar ideas to the solution of first order ordinary initial value problems using Euler's method as the basic discretization. The corresponding asymptotic expansion then contains also odd powers of h. The main purpose of this paper is to establish the existence of simple discretizations of both first and special second order systems which have asymptotic expansions in powers of h^2. These schemes, coupled with a slower mesh refinement [4] and the use of rational function extrapolation [5] should lead to effective algorithms of this type for ordinary initial value problems. Numerical results are given for the restricted two body problem, including comparison with some classical techniques.

2. Extrapolation schemes. Let $D(h)$ be a complex valued discrete approximation defined for *steps* $h \in H = (0, h_0]$ to the solution $D(0)$ of an infinitesimal problem. Under the assumption that $D(h)$ has an *asymptotic expansion*,

$$D(h) \sim e_0 + e_1 h^2 + e_2 h^4 + \cdots, \qquad h \in H, \tag{2.1}$$

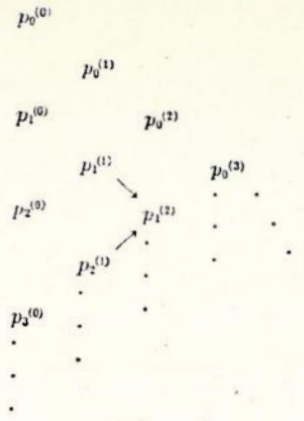
Richardson [18], [19] proposed to obtain improved approximations from two or more values of $D(h)$, say at $h_0 > h_1 > \cdots > h_n$, by requiring that the linear combinations

$$p_0^{(n)} \equiv \sum_{m=0}^{n} c_m^{(n)} D(h_m)$$

satisfy

$$p_0^{(n)} = D(0) + O(h_0^{2n+2}), \qquad h_0 \to 0^+.$$

It is important that the constants $c_m^{(n)}$ need not be calculated. The $p_0^{(n)}$ can be found indirectly with the Neville algorithm for the recursive construction of $p_0^{(n)} = p_n^{(n)}(0)$, where $p_n^{(m)}(h^2)$ is the polynomial of degree m in h^2 which interpolates $D(h)$ at $h = h_k$, $k = n, \cdots, n+m$. One forms the triangular array



from the linear recursion

$$\begin{aligned} p_n^{(0)} &= D(h_n), \\ p_n^{(m)} &= p_{n+1}^{(m-1)} + \frac{p_{n+1}^{(m-1)} - p_n^{(m-1)}}{(h_n/h_{n+m})^2 - 1}. \end{aligned} \tag{2.2}$$

For the applications discussed in this paper the main computational effort occurs in the evaluation of the first column. The scheme is built up by generating, at the nth stage, the upward sloping diagonal beginning with $p_n^{(0)}$. See, for example, the algorithms in [2], [5].

The following theorem states that, under mild assumptions on $D(h)$ and the rate of refinement of the mesh, the linear sequence to sequence transformation $p_n^{(0)} \to p_0^{(n)}$ of the first column into the diagonal is convergence and limit preserving. The sufficiency was stated, in a special case, by Stiefel and Rutishauser [23]. A more general theorem is that of Laurent [14].

THEOREM 2.1. *A necessary and sufficient condition that* $\lim_{n\to\infty} p_0^{(n)} = D(0)$ *for all functions* $D(h)$ *continuous from the right at* $h = 0$ *is that*

$$\alpha = \sup_{n\geq 0} \frac{h_{n+1}}{h_n} < 1. \tag{2.3}$$

In particular, (2.3) implies the Toeplitz condition

$$C = \sup_{n\geq 0} \sum_{m=0}^{n} |c_m^{(n)}| < \infty. \tag{2.4}$$

The constant C is a measure of the numerical stability of the scheme. The sequences $h(\alpha)$, $0 < \alpha \leqq 1$, defined by

$$h_n(\alpha) = \frac{h_0}{k_n}, \qquad \begin{cases} k_0 = 1, \\ k_{n+1} = \text{entier}(k_n/\alpha) + 1, \end{cases} \tag{2.5}$$

* Received by the editors May 21, 1965.

† Oak Ridge National Laboratory, Oak Ridge, Tennessee. This research was sponsored by the United States Atomic Energy Commission under contract with the Union Carbide Corporation.

we the following values for $C(\alpha)$:

(2.6)

α	$\frac{1}{2}^+$	$\frac{1}{4}$	$\frac{2}{5}$	$\frac{5}{11}$	$\frac{4}{9}$	$\frac{5}{9}$	1
$C(\alpha)$	1.97	2.71	5.4	11	48	4850	$+\infty$

The next theorem provides statements about the rates of convergence of the columns and principal diagonal of the p-scheme. It follows from results in [11], [5].

THEOREM 2.2. *Let $D(h)$ have the asymptotic expansion* (2.1) *and let* $\sup_{n\geq 0} h_{n+1}/h_n \leq \alpha < 1$. *Then, as* $n \to \infty$,

$$p_n^{(m)} - D(0) = (-1)^m c_{m+1}(h_n \cdots h_{n+m})^2 + o((h_n \cdots h_{n+m})^2). \tag{2.7}$$

If, in addition, $0 < \gamma \leq \inf_{n\geq 0} h_{n+1}/h_n$ *then there exist constants* E_m *such that, for each* $m \geq 0$,

$$|p_0^{(n)} - D(0)| \leq E_{m+1}(h_n \cdots h_{n+m})^2, \qquad n \geq 0. \tag{2.8}$$

In the normal case where $e_m \neq 0$, $m = 1, 2, \cdots$, (2.7) states that each column of the p-scheme converges to $D(0)$ faster than the preceding one, and (2.8) shows that the principal diagonal converges faster than any column. Under mild restrictions on the rate of growth of the order constants implied by (2.1), it can be shown that $p_0^{(n)}$ converges *superlinearly* to $D(0)$ in the sense that $|p_0^{(n)} - D(0)| \leq K_n$ and $\lim_{n\to\infty} K_{n+1}/K_n = 0$. Such is the case if $D(h)$ can be extended to a function which is analytic at $h = 0$.

An important generalization of (2.2) has recently been proposed in [5]. It uses the algorithm of Stoer [24] to construct $r_n^{(m)} = r_n^{(\mu,\nu)}(0)$, where $r_n^{(\mu,\nu)}(h^2)$ is "the" rational function with numerator degree μ and denominator degree ν ($\mu + \nu = m$) which interpolates $D(h)$ at $h = h_k$, $k = n, \cdots, n + m$. Choosing the sequence $(\mu, \nu) = (0, 0), (0, 1), (1, 1), (1, 2), \cdots$ gives the nonlinear recursion

$$r_n^{(-1)} = 0, \qquad r_n^{(0)} = D(h_n),$$

$$r_n^{(m)} = r_{n+1}^{(m-1)} + \frac{r_{n+1}^{(m-1)} - r_n^{(m-1)}}{\left(\frac{h_n}{h_{n+m}}\right)^2 \left[1 - \frac{r_{n+1}^{(m-1)} - r_n^{(m-1)}}{r_{n+1}^{(m-1)} - r_{n+1}^{(m-2)}}\right] - 1} \tag{2.9}$$

with the diagram

$$\begin{array}{ccc} & & r_n^{(m-1)} \\ & \nearrow & \\ r_{n+1}^{(m-2)} & \longrightarrow & r_n^{(m)} \\ & \searrow & \\ & & r_{n+1}^{(m-1)} \end{array}$$

A statement analogous to that of Theorem 2.2 on the rate of convergence of the columns of the Stoer scheme involves the Hankel determinants

$$H_p^{(q)} = \begin{vmatrix} e_p & e_{p+1} & \cdots & e_{p+q-1} \\ e_{p+1} & e_{p+2} & \cdots & e_{p+q} \\ \vdots & \vdots & & \vdots \\ e_{p+q-1} & e_{p+q} & \cdots & e_{p+2q-2} \end{vmatrix}.$$

THEOREM 2.3. *In addition to the hypotheses of Theorem* 2.2 *let* $H_p^{(q)} \neq 0$, $p = 0, 1, q = 1, 2, \cdots$. *If h_0 is sufficiently small the nth column of the Stoer scheme exists and*

$$r_n^{(m)} - D(0) \sim (-1)^m \tilde{e}_{m+1}(h_n \cdots h_{n+m})^2, \qquad n \to \infty,$$

where

$$\tilde{e}_{2q} = \frac{H_0^{(q+1)}}{H_0^{(q)}}, \qquad \tilde{e}_{2q+1} = \frac{H_1^{(q+1)}}{H_1^{(q)}}.$$

The algorithm of Romberg [20] for the evaluation of definite integrals,

$$T(0) = \int_a^b f(t)\,dt, \qquad \begin{cases} I = [a, b] \text{ finite}, \\ f \in C^\infty(I), \end{cases}$$

has been studied in the interesting papers [1], [23], [22], [3] by Bauer, Rutishauser, and Stiefel and, for more general h-sequences, by Bulirsch [4]. The discretization is the *trapezoidal rule*

$$T(h) = h[\tfrac{1}{2}f(a) + f(a + h) + \cdots + f(b - h) + \tfrac{1}{2}f(b)] \tag{2.10}$$

which, according to the Euler-Maclaurin formula, has the asymptotic expansion

$$T(h) \sim T(0) + \sum_{m=1}^{\infty} \frac{B_{2m}}{(2m)!}[f^{(2m-1)}(b) - f^{(2m-1)}(a)]h^{2m}. \tag{2.11}$$

The B_{2m} are the Bernoulli numbers

$$\frac{B_{2m}}{(2m)!} = \frac{2(-1)^{m-1}\zeta(2m)}{(2\pi)^{2m}},$$

and $\zeta(z)$ is the Riemann zeta function. If f is analytic on I then (2.3) implies that $p_0^{(n)} \to T(0)$ superlinearly as $n \to \infty$.

One can also base similar schemes on the *midpoint rule*

$$M(h) = h\left[f\left(a + \frac{h}{2}\right) + f\left(a + \frac{3h}{2}\right) + \cdots + f\left(b - \frac{h}{2}\right)\right].$$

Since

$$T\left(\frac{h}{2}\right) = \frac{1}{2}[T(h) + M(h)], \tag{2.12}$$

the Euler-Maclaurin formula shows that (2.1) holds for $M(h)$ with

$$c_m = -\left(1 - \frac{2}{4^m}\right)\frac{B_{2m}}{(2m)!}[f^{(2m-1)}(b) - f^{(2m-1)}(a)].$$

The relation (2.12) was used by Romberg, with the sequence $h(\frac{1}{2}^+)$, to construct the first column of his T-scheme.

3. Two one-step methods for first order systems. Let f be continuous and uniformly Lipschitzian with respect to its second argument on the set $D = I \times C_l$, where $I = [a, b]$ is a finite t-interval and C_l is the complex normed linear space of l-tuples $x = (x^{(1)}, \cdots, x^{(l)})$. Let it be required to find $\phi(t)$ at a fixed point $t = a + h_0 \in I$, where ϕ is the unique solution of the initial value problem

$$\begin{aligned} x(a) &= s, \\ x' &= f(t, x), \qquad t \in I. \end{aligned} \tag{3.1}$$

If $\phi(t)$ is wanted at a number of points $t \in I$ the algorithms described below, coupled with the extrapolation schemes (2.2) or (2.9), can be applied over the subintervals between successive points. When $l > 1$ the extrapolation schemes are applied to the individual components of the numerical solution. Two familiar one-step methods are considered in this section: Euler's method and the usual generalization to differential equations of the trapezoidal rule. For the special case where f is independent of x, their asymptotic expansions reduce to the Euler-Maclaurin formula (2.11). The proofs, which are easier than the proof of Theorem 4.2, appear in [11], [21].

It is assumed further that $f \in C^\infty(D)$. Denote by J the Jacobian matrix of f, evaluated at the solution ϕ,

$$J(t) = \frac{\partial f}{\partial x}(t, \phi(t)), \quad t \in I,$$

and define the symmetric k-linear operators $f^{(k)}(t, \phi(t))$, $t \in I$, from C_l to C_l by

$$f^{(k)}(t, \phi(t))x_1 \cdots x_k = \sum_{i_1=1}^{l} \cdots \sum_{i_k=1}^{l} \frac{\partial^k f(t, \phi(t))}{\partial x_1^{(i_1)} \cdots \partial x_k^{(i_k)}} x_1^{(i_1)} \cdots x_k^{(i_k)}.$$

The properties of such operators are discussed in [15]. This device reduces the formal study of systems to that of a single differential equation.

The coefficients of several asymptotic expansions to be given below can be defined as solutions of certain recursive systems of linear differential equations. Put

$$e_0(t) \equiv \phi(t), \tag{3.2a}$$

and, for $m = 1, 2, \cdots$, let $e_m(t)$ satisfy

$$\begin{aligned} e_m(a) &= 0, \\ e_m' &= J(t)e_m + a_m(t) + b_m(t), \qquad t \in I, \end{aligned} \tag{3.2b}$$

where

$$a_m(t) = -\sum_{k=1}^{m} \alpha_k e_{m-k}^{(qk+1)}(t) \tag{3.2c}$$

and

$$\sum_{m=1}^{\infty} b_m(t)z^m \equiv \sum_{k=2}^{\infty} \frac{1}{k!} f^{(k)}(t, \phi(t)) \left(\sum_{m=1}^{\infty} e_m(t)z^m\right)^k. \tag{3.2d}$$

The integer q and constants α_k will be specified in each particular case by the generating function

$$A(z) = \sum_{k=0}^{\infty} \alpha_k z^{qk}. \tag{3.2e}$$

It was proposed in [3], [5] to use Euler's method as a simple discretization of (3.1). Thus put

$$E(t; h) = x_N(h), \qquad Nh = t - a,$$

where the sequence $x_n(h)$, $n = 0, \cdots, N$, satisfies the difference equation

$$\begin{aligned} x_0 &= s, \\ x_{n+1} &= x_n + hf(t_n, x_n), \end{aligned}$$

with $t_n = a + nh$.

THEOREM 3.1. *Let the functions $e_m(t)$ be defined by (3.2) with*

$$A(z) = \frac{e^z - 1}{z} = \sum_{k=0}^{\infty} \frac{1}{(k+1)!} z^k.$$

Then

$$E(t; h) \sim e_0(t) + e_1(t)h + e_2(t)h^2 + \cdots \tag{3.3}$$

uniformly for $t \in I$ and steps $h \in H$.

Since (3.3) contains odd powers of h the extrapolation schemes must be modified in an obvious way. For example, the Neville scheme becomes

$$p_n^{(m)} = p_{n+1}^{(m-1)} + \frac{p_{n+1}^{(m-1)} - p_n^{(m-1)}}{\dfrac{h_n}{h_{n+m}} - 1}.$$

This results in a loss of numerical stability. Some values of the corresponding constants $C(\alpha)$ (see (2.4)-(2.6)) are:

α	$\frac{1}{2}^+$	$\frac{4}{7}$	$\frac{2}{3}$	$\frac{8}{11}$	$\frac{4}{5}$	$\frac{8}{9}$	1
$C(\alpha)$	8.3	17.4	79	370	8500	$>10^8$	$+\infty$

Note that, for the differential equation $x(0) = s$, $x' = ax$, $a = \text{const.}$,

$$E(t; h) = (1 + ah)^N s = \exp\left[at\,\frac{\log(1 + ah)}{ah}\right]s$$

$$= e^{at}[1 + p_1(at)ah + p_2(at)(ah)^2 + \cdots]s$$

is analytic for $|h| < 1/|a|$. The $p_m(t)$ are polynomials of degree m. It follows from this and a previous remark that if $\alpha < 1$ then $p_0^{(n)} \to e^{at}s$ superlinearly for the Neville scheme. However, this superlinear convergence is slower for larger values of $|a|$. This behavior generalizes to the other methods studied below.

An obvious choice for a discretization of (3.1) with an h^2-expansion is the usual generalization of the trapezoidal rule:

$$T(t; h) = x_N(h), \qquad Nh = t - a,$$

with

$$x_0 = s,$$
$$x_{n+1} = x_n + \frac{h}{2}[f(t_{n+1}, x_{n+1}) + f(t_n, x_n)]. \tag{3.4}$$

THEOREM 3.2. *Let the functions $e_m(t)$ be defined by* (3.2) *with*

$$A(z) = \frac{2}{z}\tanh\left(\frac{z}{2}\right) = \left(\sum_{k=0}^{\infty}\frac{B_{2k}}{(2k)!}z^{2k}\right)^{-1}.$$

If h_0 is sufficiently small the difference equation (3.4) *has a unique solution $x_n(h)$, $n = 0, \cdots, N$, and*

$$T(t; h) \sim e_0(t) + e_1(t)h^2 + e_2(t)h^4 + \cdots$$

uniformly for $t \in I$ and steps $h \in II$.

This generalization of the trapezoidal rule (2.10) has an important stability property. It has been shown by Dahlquist [7], [8] that any linear multistep method which preserves the asymptotic stability of solutions of $x' = \lambda x$, $\operatorname{Re}\lambda(\lambda) < 0$, for all $h > 0$ necessarily is of order $\leqq 2$ and that, among the second order methods with this property, the trapezoidal rule has the smallest error constant. This is of interest when λ has some eigenvalues with large negative real parts so that the general solution contains rapidly decaying transients. The trapezoidal rule prevents these components from reentering the numerical solution once they have decayed. Dahlquist then proposes using *global* extrapolation to increase the order of the approximation.

It is not possible to base a general purpose procedure on extrapolation of the trapezoidal rule since the h^2-expansion does not hold unless the system (3.4) is solved exactly at each step. The classical predictor-corrector technique requires in general infinitely many evaluations of f to obtain the h^2-expansion. On the other hand, if it is relatively easy to solve (3.4) exactly the use of extrapolation gives very good results.

4. A composite rule. The starting point for the main result on first order systems is Nyström's second order method, commonly called the midpoint rule:

$$\mathfrak{N}(t; h) = x_N(h), \qquad Nh = t - a,$$
$$x_0 = s, \qquad x_1 = s_1(h), \tag{4.1}$$
$$x_{n+1} = x_{n-1} + 2hf(t_n, x_n).$$

This is a two-step method and thus requires an additional starting value $s_1(h)$. It is the simplest linear k-step method [6],

$$\rho(E)x_n = h\sigma(E)f(t_n, x_n),$$

which is symmetric in the sense that

$$\rho(z) + z^k\rho(z^{-1}) = \sigma(z) - z^k\sigma(z^{-1}) = 0.$$

The requirement of stability implies that all zeros of $\rho(z)$ are of unit modulus for a symmetric method. If $k > 1$ and negative growth parameters exist, weak instability can occur. This is less important in the step-by-step use of symmetric methods with extrapolation schemes. It does require a moderate control of the step h_0, however.

For symmetric methods it is theoretically possible, by a suitable choice of starting values, to obtain asymptotic expansions in powers of h^2. The following theorem was given, in part, by de Vogelaere [9] who extended a result of Gaunt [10]. It generalizes easily to the class of symmetric multistep methods.

THEOREM 4.1. *Let the functions $e_m(t)$ be defined by* (3.2) *with*

$$A(z) = \frac{\sinh z}{z} = \sum_{k=0}^{\infty}\frac{1}{(2k+1)!}z^{2k}.$$

If the starting function $s_1(h)$ satisfies

$$s_1(h) \sim e_0(a + h) + e_1(a + h)h^2 + e_2(a + h)h^4 + \cdots$$
$$\sim \phi(a) + \phi'(a)h + \tfrac{1}{2}\phi''(a)h^2 - \tfrac{1}{12}[J(h)\phi'''(a) + \tfrac{1}{2}\phi^{(4)}(a)]h^4 + \cdots \tag{4.2}$$

for $h \in H$, *then*

$$\mathfrak{N}(t; h) \sim e_0(t) + e_1(t)h^2 + e_2(t)h^4 + \cdots$$

uniformly for $t \in I$ *and steps* $h \in H$.

Note that (4.2) does not require $s_1(h) \equiv \phi(a + h)$. It is difficult to obtain since it requires a knowledge of J and high order derivatives of the solution ϕ. De Vogelaere proposes the use of methods of Runge-Kutta type to satisfy (4.2) approximately. This appears cumbersome and in general does not lead to an infinite h^2-expansion.

The most natural choice for the starting function $s_1(h)$, in terms of the data of the problem (3.1), is

$$s_1(h) = s + f(a, s)h. \tag{4.3}$$

It is a remarkable fact that this choice leads to a certain type of *infinite* h^2-expansion. The statement of this result requires the recursive definition, similar to (3.2), of *two* sequences of functions $e_m(t)$, $f_m(t)$. Put

$$e_0(t) = f_0(t) \equiv \phi(t), \tag{4.4a}$$

and, for $m = 1, 2, \cdots$, let $e_m(t), f_m(t)$ satisfy

$$\begin{gathered} e_m(a) = 0, \qquad f_m(a) = -\sum_{k=1}^{m} \frac{1}{(2k)!} f_{m-k}^{(2k)}(a), \\ e_m' = J(t)f_m + a_m(t) + b_m(t), \qquad f_m' = J(t)e_m + c_m(t) + d_m(t), \\ t \in I, \end{gathered} \tag{4.4b}$$

where

$$\begin{aligned} a_m(t) &= -\sum_{k=1}^{m} \frac{1}{(2k+1)!} e_{m-k}^{(2k+1)}(t), \\ c_m(t) &= -\sum_{k=1}^{m} \frac{1}{(2k+1)!} f_{m-k}^{(2k+1)}(t), \end{aligned} \tag{4.4c}$$

and

$$\begin{aligned} \sum_{m=1}^{\infty} b_m(t)z^m &\equiv \sum_{k=2}^{\infty} \frac{1}{k!} f^{(k)}(t, \phi(t)) \left(\sum_{m=1}^{\infty} f_m(t)z^m\right)^k, \\ \sum_{m=1}^{\infty} d_m(t)z^m &\equiv \sum_{k=2}^{\infty} \frac{1}{k!} f^{(k)}(t, \phi(t)) \left(\sum_{m=1}^{\infty} e_m(t)z^m\right)^k. \end{aligned} \tag{4.4d}$$

THEOREM 4.2. *Let* $\mathfrak{N}(t; h)$ *be constructed from the algorithm* (4.1) *with* $s_1(h) = s + f(a, s)h$. *Then*

$$\mathfrak{N}(t; h) \sim \sum_{m=0}^{\infty} \begin{Bmatrix} e_m(t) \\ f_m(t) \end{Bmatrix} h^{2m}, \qquad t \in I, h \in H. \tag{4.5}$$

The upper (lower) expression is taken when N is even (odd).

This result shows that there exist two distinct h^2-expansions arising from Nyström's method with the starting function (4.3). Extrapolation is therefore possible with a sequence of even N's *or* with a sequence of odd N's. Since $e_m(a) = 0$ but in general $f_m(a) \neq 0$ for $m \geqq 1$, the former procedure is perhaps preferred. To better understand Theorem 4.2, put

$$u_m(t) = \tfrac{1}{2}[e_m(t) + f_m(t)], \qquad v_m(t) = \tfrac{1}{2}[e_m(t) - f_m(t)],$$

and compare with Theorem 4.2 of Henrici [13]. The functions u_m and v_m satisfy differential equations of the form

$$\begin{aligned} u_m' &= J(t)u_m + \text{inhomogeneous terms}, \\ v_m' &= -J(t)v_m + \text{inhomogeneous terms}; \end{aligned}$$

the expansion (4.5) becomes

$$\mathfrak{N}(t; h) \sim \sum_{m=0}^{\infty} [u_m(t) + (-1)^N v_m(t)]h^{2m}. \tag{4.6}$$

The functions v_m are the "weakly unstable" components of the discretization error. Note that $v_0(t) \equiv 0$. By choosing a more *accurate* starting value $s_1(h)$, it is possible to obtain an expansion of the form (4.6) where, in addition, $v_1(t) \equiv 0$. Such is the case if

$$s_1(h) = \phi(a) + \phi'(a)h + \tfrac{1}{2}\phi''(a)h^2,$$

but this requires the knowledge of $f_t(a, s)$ and $J(a)$. Even then $v_2(t) \not\equiv 0$ in general. It will be seen later how to annihilate $v_1(t)$ which is the leading unstable component.

Proof of Theorem 4.2. For $p \geqq 1$ and $t = t_n = a + nh$ let

$$\begin{aligned} \epsilon_n(h) &\equiv x_n(h) - \phi(t) - \delta_n(h), \\ \delta_n(h) &\equiv \sum_{m=1}^{p-1} \begin{Bmatrix} e_m(t) \\ f_m(t) \end{Bmatrix} h^{2m}, \qquad n \begin{Bmatrix} \text{even} \\ \text{odd} \end{Bmatrix}. \end{aligned}$$

It will be shown that $\epsilon_n(h) = O(h^{2p})$ uniformly for $t \in I$ and steps $h \in H$. This is known for $p = 1$ [13, Theorem 4.1]; thus

$$\epsilon_n(h) = O(h^2), \qquad t \in I, h \in H. \tag{4.7}$$

Define the linear operator $\mathfrak{L}$ by

$$\mathfrak{L}\epsilon_n = \epsilon_{n+1} - \epsilon_{n-1} - 2hJ(t)\epsilon_n.$$

For $p > 1$ the result will follow from

$$\epsilon_0(h) = 0, \qquad \epsilon_1(h) = O(h^{2p}), \tag{4.8a}$$

$$\mathfrak{L}\epsilon_n(h) = O(h^3 \| \epsilon_n(h) \|) + O(h^{2p+1}), \qquad t \in I, h \in H, \tag{4.8b}$$

by (4.7) and $p - 1$ applications of [13, Lemma 3.2].

The first equation of (4.8a) holds since $x_0(h) = s = \phi(a)$ and $e_m(a) = 0$, $m = 1, 2, \cdots$. Similarly

$$\epsilon_1(h) = \phi(a) + \phi'(a)h - \sum_{m=0}^{p-1} f_m(a+h)h^{2m}.$$

Expanding $f_m(a+h)$ in finite Taylor series about $h = 0$, rearranging into powers of h, and estimating remainders gives

$$-\epsilon_1(h) = \sum_{m=0}^{p-1}\left[f_m(a) + \sum_{k=1}^{m}\frac{1}{(2k)!}f_{m-k}^{(2k)}(a)\right]h^{2m}$$

$$+ \sum_{m=0}^{p-1}\left[f_m'(a) + \sum_{k=1}^{m}\frac{1}{(2k+1)!}f_{m-k}^{(2k+1)}(a)\right]h^{2m+1} + O(h^{2p}), \qquad h \in H.$$

The sums vanish because of the initial value problems defining the f_m. By (4.4d), $d_m(a) = 0$ since $e_m(a) = 0$. This completes the proof of (4.8a).

To show (4.8b), write

$$\mathfrak{L}\epsilon_n = \mathfrak{L}x_n - \sum_{m=0}^{p-1}\mathfrak{L}\begin{Bmatrix} e_m(t) \\ f_m(t)\end{Bmatrix}h^{2m} \tag{4.9}$$

and consider each term on the right separately. From the difference equation (4.1),

$$\mathfrak{L}x_n = 2h[f(t, \phi(t) + \delta_n + \epsilon_n) - J(t)(\phi(t) + \delta_n + \epsilon_n)].$$

Expanding $f(t, \phi(t) + \delta_n + \epsilon_n)$ in a finite (Fréchet) Taylor series about $\phi(t)$ and estimating remainders, using the fact that both $\delta_n(h)$ and $\epsilon_n(h)$ are uniformly $O(h^2)$, gives

$$f(t, \phi(t) + \delta_n + \epsilon_n) - f(t, \phi(t)) - J(t)(\delta_n + \epsilon_n)$$

$$= \sum_{k=2}^{p-1}\frac{1}{k!}f^{(k)}(t, \phi(t))(\delta_n + \epsilon_n)^k + O(h^{2p})$$

$$= \sum_{k=2}^{p-1}\frac{1}{k!}f^{(k)}(t, \phi(t))\delta_n{}^k + O(h^2\|\epsilon_n\|) + O(h^{2p}), \qquad t \in I, h \in H.$$

Combining the last two expressions with (4.4d) yields

$$\mathfrak{L}x_n = 2h\left[f(t, \phi(t)) - J(t)\phi(t) + \sum_{m=1}^{p-1}\begin{Bmatrix} d_m(t) \\ b_m(t)\end{Bmatrix}h^{2m}\right] \tag{4.10}$$

$$+ O(h^3\|\epsilon_n\|) + O(h^{2p+1}), \qquad t \in I, h \in H.$$

The mth term of the sum (4.9) can be expanded similarly:

$$\mathfrak{L}\begin{Bmatrix} e_m(t) \\ f_m(t)\end{Bmatrix} = \begin{Bmatrix} f_m(t+h) - f_m(t-h) - 2hJ(t)e_m(t) \\ e_m(t+h) - e_m(t-h) - 2hJ(t)f_m(t)\end{Bmatrix}$$

$$= 2h\left[\begin{Bmatrix} f_m'(t) - J(t)e_m(t) \\ e_m'(t) - J(t)f_m(t)\end{Bmatrix} + \sum_{k=1}^{p-m-1}\frac{1}{(2k+1)!}\begin{Bmatrix} f_m^{(2k+1)}(t) \\ e_m^{(2k+1)}(t)\end{Bmatrix}h^{2k}\right]$$

$$+ O(h^{2(p-m)+1}), \qquad t \in I, h \in H.$$

Rearranging into powers of h^2 and applying (4.4c) then gives

$$\mathfrak{L}\sum_{m=0}^{p-1}\begin{Bmatrix} e_m(t) \\ f_m(t)\end{Bmatrix}h^{2m} = 2h\sum_{m=0}^{p-1}\begin{Bmatrix} f_m'(t) - J(t)e_m(t) - c_m(t) \\ e_m'(t) - J(t)f_m(t) - a_m(t)\end{Bmatrix}h^{2m}$$

$$+ O(h^{2p+1}), \qquad t \in I, h \in H.$$

The verification of (4.8) is completed by combining this with (4.9–10) and recalling the differential equations (3.1), (4.4b).

Because of Theorem 4.2 it seems natural to separate the "even" and "odd" parts of $\mathfrak{N}(t; h)$. Also, noting the special case of ordinary integration when $f(t, x)$ is independent of x, one is led to define generalizations of $M(h)$ and $T(h)$ by

$$M(t; h) = x_N(h), \qquad Nh = t - a,$$

$$T(t; h) = y_N(h) - \frac{h}{2}f(t, x_N(h)),$$

where

$$x_0 = s, \qquad y_0 = s + \frac{h}{2}f(a, s),$$

$$x_{n+1} = x_n + hf(t_{n+\frac{1}{2}}, y_n), \tag{4.11}$$

$$y_{n+1} = y_n + hf(t_{n+1}, x_{n+1}).$$

These rules are related to $\mathfrak{N}(t; h)$ by

$$M(t; h) = \mathfrak{N}\left(t; \frac{h}{2}\right), \tag{4.12}$$

$$T(t; h) = \mathfrak{N}\left(t + \frac{h}{2}; \frac{h}{2}\right) - \frac{h}{2}f\left(t, \mathfrak{N}\left(t; \frac{h}{2}\right)\right). \tag{4.13}$$

It follows directly from (4.12) that

$$M(t; h) \sim \sum_{m=0}^{\infty} e_m(t)\left(\frac{h}{2}\right)^{2m}, \qquad t \in I, h \in H,$$

and, by expanding the right side of (4.13) using (4.4-5), that

$$T(t; h) \sim \sum_{m=0}^{\infty} g_m(t) \left(\frac{h}{2}\right)^{2m}, \qquad t \in I, h \in H,$$

where

$$g_m(t) = \sum_{k=0}^{m} \frac{1}{(2k)!} f_{m-k}^{(2k)}(t).$$

Notice that now, by the initial conditions for the functions f_m, $g_m(a) = 0$, $m = 1, 2, \cdots$.

The rules $M(t; h)$ and $T(t; h)$ again provide two distinct h^2-expansions, and extrapolation for $\phi(t)$ is possible in either with an *arbitrary* sequence of N's. More generally one could consider the linear combination $\gamma M(t; h) + (1 - \gamma)T(t; h)$. Noting (2.12) it is natural to take $\gamma = \frac{1}{2}$ and thus to put

$$A(t; h) = \tfrac{1}{2}[M(t; h) + T(t; h)].$$

The rule $A(t; h)$ has the asymptotic expansion

$$A(t; h) \sim \sum_{m=0}^{\infty} [e_m(t) + g_m(t)] \left(\frac{h}{2}\right)^{2m}, \qquad t \in I, h \in H.$$

In particular

$$e_1(t) + g_1(t) = u_1(t) + \tfrac{1}{4}x''(t)$$

does not contain $v_1(t)$. This is reminiscent of the averaging procedure of Milne and Reynolds [17] for annihilating the leading "unstable" component of the discretization error.

Finally, the following observation guarantees the numerical stability of the (practical) step-by-step algorithms. If either of the rules M, T, or A is coupled with the Neville (Stoer) scheme using a fixed number of extrapolations per step h_0, then the entire process is a Runge-Kutta (one-step) method. The existence of the Stoer schemes at each step must be assumed in the latter case.

5. Special second order systems. Let f again satisfy the hypotheses of §3 and now let it be required to find $\phi(t)$ at a fixed point $t \in I$, where ϕ is the unique solution of the special second order system

$$x(a) = s, \qquad x'(a) = s', \qquad x'' = f(t, x), \qquad t \in I. \tag{5.1}$$

The simplest linear k-step method of the form

$$\rho(E)x_n = h^2\sigma(E)f(t_n, x_n)$$

for the solution of (5.1) is the Störmer second order scheme:

$$x_{n+1} - 2x_n + x_{n-1} = h^2 f(t_n, x_n).$$

The simplest choice of starting values compatible with this method is

$$x_0 = s, \qquad x_1 = s + hs' + \frac{h}{2}f(a, s).$$

In its summed form, which reduces the accumulation of rounding errors [12, §6.4], the scheme takes a form similar to (4.11) but with one function evaluation per step:

$$x_0 = s, \qquad y_0 = s' + \frac{h}{2}f(a, s), \tag{5.2a}$$

$$x_{n+1} = x_n + hy_n,$$

$$y_{n+1} = y_n + hf(t_{n+1}, x_{n+1}).$$

The rules $S(t; h)$ and $S^*(t; h)$ are now defined by

$$S(t; h) = x_N(h), \qquad Nh = t - a, \tag{5.2b}$$

$$S^*(t; h) = y_N(h) - \frac{h}{2}f(t, x_N(h)).$$

In order to state results similar to those of §4 let the functions $e_m(t)$, $t \in I$, be defined recursively by

$$e_0(t) \equiv \phi(t), \tag{5.3a}$$

and for $m = 1, 2, \cdots$ by

$$e_m(a) = 0, \qquad e_m'(a) = -\sum_{k=1}^{m} \frac{1}{(2k+1)!} e_{m-k}^{(2k+1)}(a), \tag{5.3b}$$

$$e_m'' = J(t)e_m + a_m(t) + b_m(t), \qquad t \in I,$$

where

$$a_m(t) = -2\sum_{k=1}^{m} \frac{1}{(2k+2)!} e_{m-k}^{(2k+2)}(t) \tag{5.3c}$$

and

$$\sum_{m=1}^{\infty} b_m(t)z^m \equiv \sum_{k=2}^{\infty} \frac{1}{k!} f^{(k)}(t, \phi(t)) \left(\sum_{m=1}^{\infty} e_m(t)z^m\right)^k. \tag{5.3d}$$

In addition put, for $m = 0, 1, \cdots$,

$$e_m^*(t) = \sum_{k=0}^{m} \frac{1}{(2k+1)!} e_{m-k}^{(2k+1)}(t), \tag{5.4}$$

and note that

$$e_0^*(t) \equiv \phi'(t).$$

THEOREM 5.1. *The rules $S(t; h)$ and $S^*(t; h)$ have the asymptotic expansions*

$$S(t; h) \sim \sum_{m=0}^{\infty} e_m(t)h^{2m}, \tag{5.5}$$

$$S^*(t; h) \sim \sum_{m=0}^{\infty} e_m^*(t)h^{2m}, \tag{5.6}$$

uniformly for $t \in I$ and steps $h \in H$.

The result (5.5) was recently stated by Mayers [16], but without explicit expressions for the functions $e_m(t)$. Theorem 5.1 is actually more satisfying than the corresponding results for first order systems. There is no fear of instability brought on by the numerical method. If instability exists it is normally inherent in the differential equation (5.1). The methods for special second order equations are usually justified by the fact that a saving can be achieved if one avoids computation of the derivative $\phi'(t)$. It is necessary to know $\phi'(t)$ for the step-by-step use of (5.2) coupled with extrapolation schemes. It is therefore noteworthy that an h^2-expansion can be obtained for its calculation with *no increase* in the number of evaluations of f.

Proof of Theorem 5.1. For $p \geqq 1$ and $t = t_n = a + nh$ let

$$\epsilon_n(h) \equiv x_n(h) - \phi(t) - \delta_n(h),$$

$$\delta_n(h) \equiv \sum_{m=1}^{p-1} e_m(t)h^{2m}.$$

To prove (5.5) it must be shown that $\epsilon_n(h) = O(h^{2p})$ uniformly for $t \in I$ and steps $h \in H$. This is known for $p = 1$ [12, Theorem 6.7]. Define the linear operator $\mathfrak{L}$ by

$$\mathfrak{L}\epsilon_n = \epsilon_{n+1} - 2\epsilon_n + \epsilon_{n-1} - h^2J(t)\epsilon_n.$$

For $p > 1$ the required result will follow from

$$\epsilon_0(h) = 0, \qquad \epsilon_1(h) = O(h^{2p+1}), \tag{5.7a}$$

$$\mathfrak{L}\epsilon_n(h) = O(h^4 \| \epsilon_n(h) \|) + O(h^{2p+2}), \qquad t \in I,\ h \in H, \tag{5.7b}$$

by the result for $p = 1$ and $p - 1$ applications of [12, Lemma 6.3].

The verification of (5.7) is accomplished by showing, similar to the proof of Theorem 4.2, that

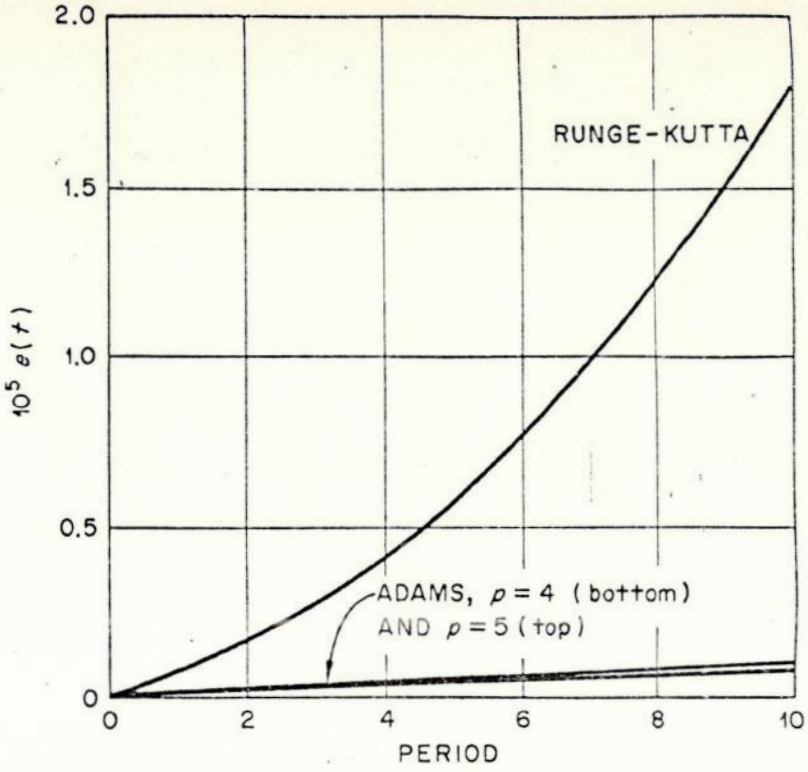


FIG. 1

$$-\epsilon_1(h) = \sum_{m=1}^{p-1}\left[e_m'(a) + \sum_{k=1}^{m} \frac{1}{(2k+1)!} e_{m-k}^{(2k+1)}(a)\right]h^{2m+1}$$

$$+ \sum_{m=1}^{p-1}\left[\frac{1}{2}e_m''(a) + \sum_{k=1}^{m} \frac{1}{(2k+2)!} e_{m-k}^{(2k+2)}(a)\right]h^{2m+2} + O(h^{2p+1}), \qquad h \in H,$$

and

$$-\mathfrak{L}\epsilon_n(h) = h^2\sum_{m=1}^{p-1}[e_m''(t) - J(t)e_m(t) - a_m(t) - b_m(t)]h^{2m}$$

$$+ O(h^4 \| \epsilon_n(h) \|) + O(h^{2p+2}), \quad t \in I, h \in H,$$

and then applying the definitions (5.3).

To prove (5.6) note that

$$S^*(t; h) = h^{-1}[S(t + h; h) - S(t; h)] - \frac{h}{2}f(t, S(t; h)),$$

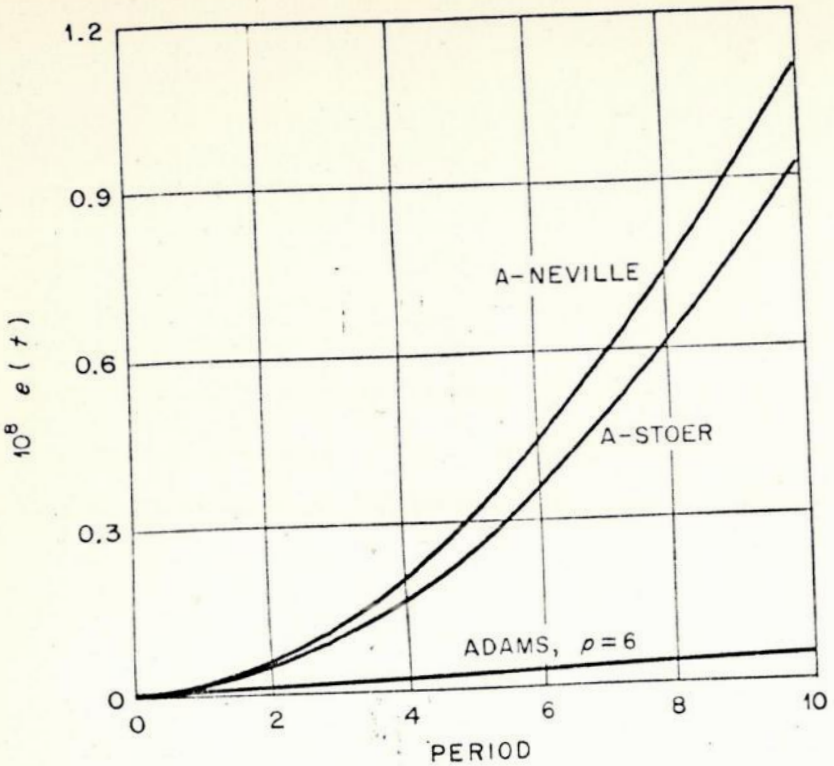


FIG. 2

and expand the right side using (5.3-5).

6. Numerical results. The restricted two-body problem

$$x(0) = [1, 0]^T, \qquad x'(0) = [0, 1]^T,$$

$$x'' = -\frac{x}{\| x \|^3}, \qquad 0 \leqq t \leqq 20\pi, \tag{6.1}$$

$$\| x \| = \text{sqrt}\ (x_1^2 + x_2^2),$$

with exact solution

$$\phi(t) = [\cos t, \sin t]^T,$$

was solved numerically with the rules A and S, S^* coupled with the Neville and Stoer algorithms. To compare the A-schemes with some classical methods the formulation of (6.1) as a first order system was also solved by the Runge-Kutta method and Adams predictor-corrector pairs of order $p = 4, 5, 6$ with two corrections per step. The number of evaluations of f varied approximately constant in this comparison.

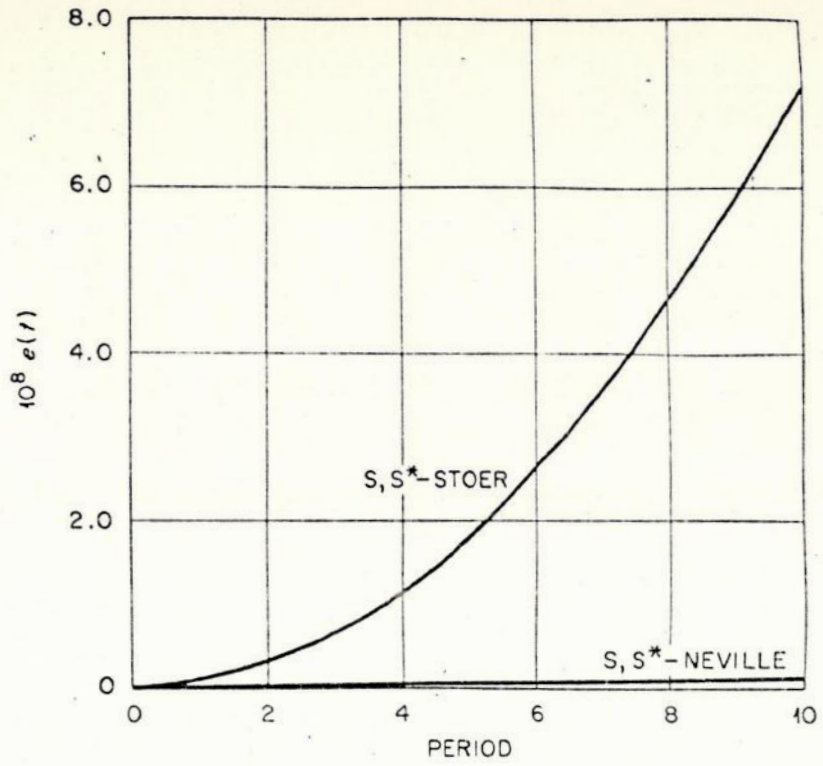


FIG. 3

The evaluations of f were accurate to 39 binary places; high precision was used in the remaining computations. The extrapolation schemes used the sequence (2.5) with $h_0 = \pi/3$, $\alpha = 1/\text{sqrt}\ (2)$, and six extrapolations per "global" step h_0. Figs. 1-3 show the results of these experiments. The error $e(t) = \| \tilde{x}(t) - \phi(t) \|$, where $\tilde{x}(t)$ is the numerical solution, is plotted as a function of the number of periods.

The error curves had roughly the same shape in each case; some appear linear due to the scale. The efficiency of the extrapolation algorithms is somewhat lower than the Adams sixth order method in this example. This standing can be improved by using higher precision and, perhaps, a slightly larger value of α. A similar comparison with seven extrapolations per step gave maximum errors of $\sim 2.10^{-11}$ for both the A-Neville scheme and the Adams sixth order method. The error was pure rounding error in these cases. It is interesting that the Neville algorithm gave better results, when coupled with S, S^*, than the Stoer algorithm. This does not appear to be a typical example however.

To test the effect of weakest stability [illegible]

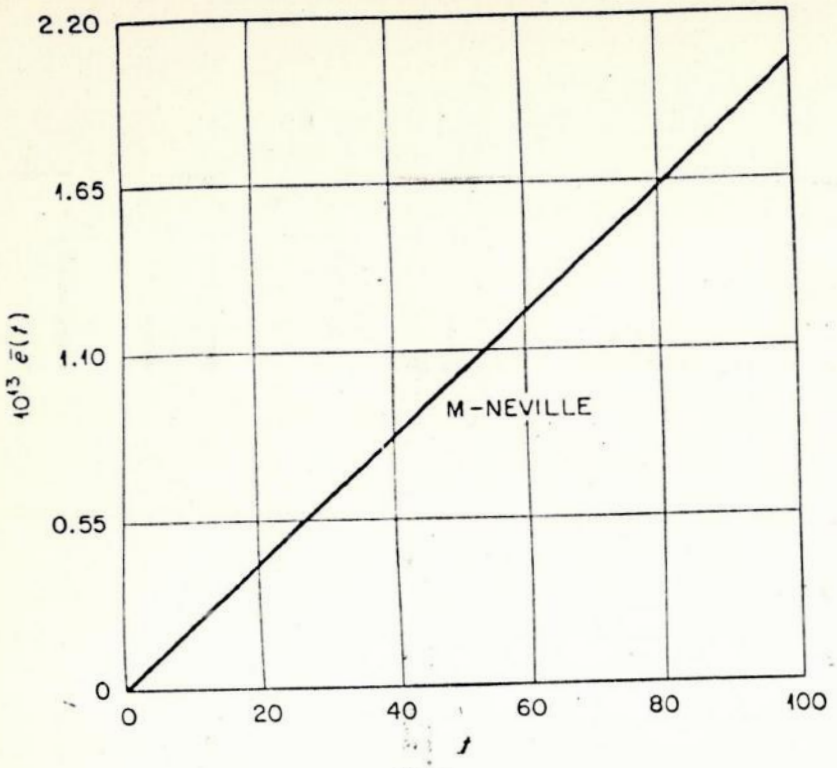


FIG. 4

$$x(0) = [0, 1]^T,$$

$$x' = Ax, \qquad 0 \leqq t \leqq 100, \qquad A = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix},$$

with exact solution

$$\phi(t) = e^{-t}[\sin t, \cos t]^T,$$

was solved (in high precision) with the sequence $h(\frac{1}{2}^+)$, $h_0 = 1$, and six extrapolations per step. Fig. 4 shows the *relative* error $\bar{e}(t) = e^t \| \bar{x}(t) - \phi(t) \|$ as a function of t.

In conclusion, it should be noted that the extrapolation algorithms provide good estimates of the "local error" and are extremely flexible with regard to variation of the step h_0.

REFERENCES

[1] F. L. BAUER, *La méthode d'integration numérique de Romberg*, Colloque sur l'analyse numérique (Mons 1961), Gauthier-Villars, Paris, 1961, pp. 119–129.

[2] ———, *Algorithm 60, Romberg integration*, Comm. ACM, 4(1961), p. 255.
[3] F. L. BAUER, H. RUTISHAUSER, AND E. STIEFEL, *New aspects in numerical quadrature*, Proceedings of Symposia in Applied Mathematics, vol. 15, American Mathematical Society, 1963, pp. 199–218.
[4] R. BULIRSCH, *Bemerkungen zur Romberg-Integration*, Numer. Math., 6(1964), pp. 6–16.
[5] R. BULIRSCH AND J. STOER, *Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus*, Ibid., 6(1964), pp. 413–427.
[6] G. DAHLQUIST, *Convergence and stability in the numerical integration of ordinary differential equations*, Math. Scand., 4(1956), pp. 33–53.
[7] ———, *Stability questions for some numerical methods for ordinary differential equations*, Proceedings of Symposia in Applied Mathematics, vol. 15, American Mathematical Society, 1963, pp. 147–158.
[8] ———, *A special stability problem for linear multistep methods*, Nordisk Tidskr. Informations-Behandling, 3(1963), pp. 27–43.
[9] R. DE VOGELEARE, *On a paper of Gaunt concerned with the start of numerical solutions of differential equations*, Z. Angew. Math. Phys., 8(1957), pp. 151–156.
[10] J. A. GAUNT, *The deferred approach to the limit, II-Interpenetrating lattices*, Philos. Trans. Roy. Soc. London Ser. A., 226 (1927), pp. 350–361.
[11] W. B. GRAGG, *Repeated extrapolation to the limit in the numerical solution of ordinary differential equations*, Doctoral dissertation, University of California, Los Angeles, 1964.
[12] P. HENRICI, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley, New York, 1962.
[13] ———, *Error Propagation for Difference Methods*, John Wiley, New York, 1963.
[14] P. LAURENT, *Un théorème de convergence pour le procédé d'extrapolation de Richardson*, C. R. Acad. Sci. Paris, 256 (1963), pp. 1435–1437.
[15] L. A. LIUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Frederick Ungar, New York, 1961.
[16] D. F. MAYERS, *The deferred approach to the limit in ordinary differential equations*, Comput. J., 7(1964), pp. 54–57.
[17] W. E. MILNE AND R. R. REYNOLDS, *Stability of a numerical solution of differential equations*, J. Assoc. Comput. Mach., 6(1959), pp. 196–203; 7(1960), pp. 46–56.
[18] L. F. RICHARDSON, *The approximate arithmetical solution by finite differences of physical problems involving differential equations with an application to the stresses in a masonry dam*, Philos. Trans. Roy. Soc. London Ser. A, 210 (1910), pp. 307–357.
[19] ———, *The deferred approach to the limit, I-Single lattice*, Ibid., 226(1927), pp. 299–349.
[20] W. ROMBERG, *Vereinfachte numerische Integration*, Norske Vid. Selsk. Forh. (Trondheim), 28(1955), pp. 30–36.
[21] H. J. STETTER, *Asymptotic expansions for the error of discretization algorithms for non-linear functional equations*, Numer. Math., 7(1964), pp. 18–31.
[22] E. STIEFEL, *Altes und Neues über numerische Quadratur*, Z. Angew. Math. Mech., 41(1961), pp. 408–413.
[23] E. STIEFEL AND H. RUTISHAUSER, *Remarques concernant l'integration numérique*, C. R. Acad. Sci. Paris, 252(1961), pp. 1899–1900.
[24] J. STOER, *Über zwei Algorithmen zur Interpolation mit rationalen Funktionen*, Numer. Math., 3(1961), pp. 285–304.

$s \leqq \nu \leqq t$ bezeichnet,

$$b_{i,i} - b_{k,k} \leqq \frac{1}{\Delta - 3\gamma} \sum_{\nu \notin T} (a_{\nu,k}^2 + a_{\nu,i}^2) \leqq \frac{|a_{i,k}|}{\Delta - 3\gamma} \sum_{\nu \notin T} (|a_{\nu,k}| + |a_{\nu,i}|)$$
$$\leqq \frac{|a_{i,k}|}{\Delta - 3\gamma} \sqrt{2n-4} \cdot \frac{\gamma}{\sqrt{2}}.$$

In Verbindung mit (7) ergibt sich $2(\Delta - 3\gamma) \leqq \sqrt{n-2} \cdot \gamma$ im Widerspruch zu (6).

Sobald (6) erfüllt ist, wird also das maximale $a_{i,k}$ nur noch außerhalb der Teilmatrix $(a_{\nu,\mu})$ $(s \leqq \nu, \mu \leqq t)$ gefunden. In [1] ergab sich nach $\frac{n}{2}(n-1)$ Rotationen ein $\gamma'' \leqq \sqrt{\frac{n}{2} - 1}\, \frac{\gamma^2}{\Delta - 2\gamma}$; daraus folgt $\gamma'' < \gamma$, sobald $\Delta - 2\gamma > \gamma \sqrt{\frac{n}{2} - 1}$ erreicht ist. Diese Bedingung ist von ähnlicher Größenordnung wie (6).

Hat die Matrix m Eigenwerte mit Vielfachheiten $v_1, v_2, \ldots, v_m$, dann erhält man nach dem hier bewiesenen Satze schon nach höchstens $\frac{n}{2}(n-1) - \sum_{\mu=1}^{m} \frac{v_\mu}{2}(v_\mu - 1)$ Rotationen ein $\gamma'' \leqq \frac{n}{2} \cdot \frac{\gamma^2}{\Delta - 2\gamma}$ (man vgl. (16) in [1]). In diesem Sinne wird die Konvergenz durch die Existenz mehrfacher Eigenwerte beschleunigt.

Literatur

[1] SCHÖNHAGE, A.: Zur Konvergenz des Jacobi-Verfahrens. Num. Math. 3, 374–380 (1961).

[2] WILKINSON, J. H.: Note on the Quadratic Convergence of the Cyclic Jacobi Process. Num. Math. 4, 296–300 (1962).

Institut für Angewandte Mathematik der Universität
Köln-Lindenthal
Weyertal 88

(Eingegangen am 29. April 1964)

Numerische Mathematik 6, 413–427 (1964)

Fehlerabschätzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson-Typus

Von

ROLAND BULIRSCH und JOSEF STOER

1.

Sei $T(h)$ die zu einem Diskretisierungsparameter h gewonnene numerische Näherung eines exakten Problems, definiert durch $\lim_{h \to 0} T(h) = T(0)$. Von RICHARDSON [8] stammt folgende Idee einer Verbesserung der $T(h)$: Man berechne $T(h_i)$ für verschiedene h_i, $i = 0, \ldots, m$, lege durch die Stützpunkte $(h_i, T(h_i))$ ein interpolierendes Polynom $\hat{T}_m(h)$ und nehme $\hat{T}_m(0)$ als Näherungswert für $T(0)$. LAURENT [7] hat kürzlich in einem allgemeinen Rahmen gezeigt, daß unter geeigneten Voraussetzungen $\hat{T}_m(0)$ mit wachsendem m gegen $T(0)$ konvergiert.

RICHARDSONs Methode, verbunden mit einem von NEVILLE bzw. AITKEN stammenden Interpolationsalgorithmus zur Ermittlung von $\hat{T}_m(0)$, liefert im allgemeinen mit geringem Aufwand sehr gute numerische Resultate. Beispiel dafür ist das bekannte, nach ROMBERG [9] benannte Quadraturverfahren; vgl. dazu die Arbeit [1] von BAUER, RUTISHAUSER und STIEFEL und die Arbeit [3]. Weitere Anwendungen finden sich in Arbeiten von RUTISHAUSER [10], BOLTON and SCOINS [2] u.a. Siehe dazu auch die Arbeiten [13, 14, 15, 16][1].

Voraussetzung für die numerische Brauchbarkeit der Extrapolationsmethode ist allerdings die Existenz einer Entwicklung der Form

$$T(h) = \tau_0 + \tau_1 h^{\gamma_1} + \cdots + \tau_k h^{\gamma_k} + R_{k+1}(h) h^{\gamma_{k+1}} \tag{1}$$

mit $|R_{k+1}(h)| \leqq M_{k+1}$ für alle $h > 0$, $\tau_0, \ldots, \tau_k$ unabhängig von h und $0 < \gamma_1 < \cdots < \gamma_{k+1}$. STETTER [11] hat im Anschluß an GRAGG [4] gezeigt, daß solche Entwicklungen bei großen Klassen praktisch wichtiger Diskretisierungsverfahren (Differenzenmethoden) existieren. Spezielle Probleme hat NAVOT [17] untersucht[1].

Die vorliegende Arbeit gliedert sich in zwei Teile. Im ersten Teil werden unter der Voraussetzung (1) Abschätzungen für den Extrapolationsfehler $|\hat{T}_m(0) - T(0)|$ hergeleitet. Der zweite Teil gibt eine neue Version des Extrapolationsgedankens: Legt man durch die Stützstellen $(h_i, T(h_i))$ eine interpolierende rationale Funktion $\hat{T}_{\mu,\nu}(h)$ (Zählergrad μ, Nennergrad ν, $\mu + \nu = m$), so kann auch $\hat{T}_{\mu,\nu}(0)$ als Näherungswert für $T(0)$ genommen werden. Die Extrapolation auf $\hat{T}_{\mu,\nu}(0)$ wird hier mit einem in [12] beschriebenen Algorithmus durchgeführt. Für den Fall der Trapezsummen-Extrapolation kann der Extrapolationsfehler wegen der speziellen Gestalt des Restgliedes der Euler-McLaurinschen

[1] Erst nach Abschluß dieser Arbeit zur Kenntnis gelangt.

Formel angegeben werden. Beispiele zeigen die Überlegenheit dieses Verfahrens, das, obwohl nur geringfügig komplizierter, in allen untersuchten Fällen nicht schlechter, meistens sogar erheblich besser konvergierte als entsprechende Polynom-Verfahren.

2.

Zunächst sollen lineare Extrapolationsoperatoren Λ_m^i mit den Eigenschaften

$$\text{(2)}\quad \begin{aligned} &\text{a)}\quad \Lambda_m^i T = T_m^{(i)} = \sum_{j=i}^{i+m} c_{mj}^{(i)} T(h_j) \\ &\text{b)}\quad \Lambda_m^i 1 = 1 \\ &\text{c)}\quad \Lambda_m^i h^{\gamma_j} = 0, \qquad j = 1, \ldots, m \end{aligned}$$

rekursiv konstruiert werden. Dabei gelte $h_0 > h_1 > \cdots > 0$. Definiert man den Ausdruck (in Spezialfällen das Polynom)

$$\hat{T}_m(h) = a_0 + a_1 h^{\gamma_1} + \cdots + a_m h^{\gamma_m}$$

durch die Forderung

$$\hat{T}_m(h_j) = T(h_j), \qquad j = i, i+1, \ldots, i+m,$$

so ist die Bestimmung von $\Lambda_m^i T$ gleichwertig der Berechnung von $\hat{T}_m(0)$. Ist nämlich $\hat{T}_m(h)$ ein solcher Ausdruck, so gilt wegen (2)

$$\Lambda_m^i \hat{T}_m = \Lambda_m^i T = a_0 = \hat{T}_m(0).$$

Zur Konstruktion der Λ_m^i läßt sich also die Theorie der Interpolation heranziehen. Für den Fall $\gamma_j = j\gamma$ liefern die Nevilleschen Interpolationsformeln, angewandt auf ein Polynom in h^γ (s. [5], [6]), die Lösung

$$\text{(3)}\quad \begin{aligned} &\Lambda_0^i T = T_0^{(i)} = T(h_i), \\ &\Lambda_m^i T = T_m^{(i)} = \frac{h_i^\gamma T_{m-1}^{(i+1)} - h_{i+m}^\gamma T_{m-1}^{(i)}}{h_i^\gamma - h_{i+m}^\gamma} \\ &\qquad = T_{m-1}^{(i+1)} + \frac{T_{m-1}^{(i+1)} - T_{m-1}^{(i)}}{\left(\frac{h_i}{h_{i+m}}\right)^\gamma - 1}, \qquad m \geqq 1. \end{aligned}$$

Ordnet man die $T_m^{(i)}$ in dem Schema an

$$\text{(4)}\quad \begin{array}{lllll} T(h_0) = T_0^{(0)} & & & & \\ & T_1^{(0)} & & & \\ T(h_1) = T_0^{(1)} & & T_2^{(0)} & & \\ & T_1^{(1)} & & T_3^{(0)} & \\ T(h_2) = T_0^{(2)} & & T_2^{(1)} & & \\ & T_1^{(2)} & & & \\ T(h_3) = T_0^{(3)} & & & & \\ \cdot & & \cdot & & \\ \cdot & & & & \end{array}$$

so lassen sich die $T_m^{(i)}$ ausgehend von der ersten Spalte rekursiv berechnen.

Für die $c_{mj}^{(i)}$ aus Gl. (2) liefert die Interpolationsformel von LAGRANGE

$$\text{(5)}\quad c_{mj}^{(i)} = \prod_{\varrho=i}^{i+m}{}' \frac{h_\varrho^\gamma}{h_\varrho^\gamma - h_j^\gamma}.$$

Die Formeln (3) und (5) gelten für beliebige h_i, soferne nur $\gamma_j = j\gamma$ ist. Für allgemeinere Sequenzen γ_j ist eine dem Nevilleschen Algorithmus entsprechende, einfache Interpolationsformel nicht bekannt.

Im Falle allgemeiner Sequenzen γ_j lassen sich jedoch die Operatoren Λ_m^i für die speziellen Schrittweiten $h_j = h_0 b^j$, $0 < b < 1$, leicht rekursiv konstruieren (vgl. hierzu auch [10], § 4):

Aus 2a) und 2c) ergibt sich nämlich

$$\Lambda_m^i h^{\gamma_\varrho} = \sum_{j=i}^{i+m} c_{mj}^{(i)} h_j^{\gamma_\varrho} = h_0^{\gamma_\varrho} b^{i\gamma_\varrho} \sum_{j=0}^{m} c_{m,j+i}^{(i)} (b^{\gamma_\varrho})^j = 0, \qquad \varrho = 1, \ldots, m;$$

daher sind die b^{γ_ϱ} gerade die Nullstellen des Polynoms

$$P_m^{(i)}(x) \equiv \sum_{j=0}^{m} c_{m,j+i}^{(i)} x^j$$

und wegen 2b) hat $P_m^{(i)}(x)$ die Gestalt

$$\text{(6)}\quad P_m^{(0)}(x) = P_m^{(i)}(x) = \prod_{\varrho=1}^{m} \frac{x - b^{\gamma_\varrho}}{1 - b^{\gamma_\varrho}}.$$

Damit gilt $c_{m,j+i}^{(i)} = c_{mj}^{(0)}$ für alle i, und die $c_{mj}^{(0)}$ sind bis auf einen gemeinsamen Faktor die elementarsymmetrischen Funktionen der b^{γ_ϱ}, $\varrho = 1, \ldots, m$.

Wegen

$$P_m^{(0)}(x) = \frac{x - b^{\gamma_m}}{1 - b^{\gamma_m}} P_{m-1}^{(0)}(x)$$

erhält man für $\Lambda_m^i T$ die einfachen Rekursionsformeln

$$\text{(7)}\quad \begin{aligned} &\Lambda_0^i T = T_0^{(i)} = T(h_i), \\ &\Lambda_m^i T = T_m^{(i)} = \frac{T_{m-1}^{(i+1)} - b^{\gamma_m} T_{m-1}^{(i)}}{1 - b^{\gamma_m}}, \qquad m \geqq 1. \end{aligned}$$

Für $\gamma_m = m\gamma$ geht (7) in (3) über.

3.

Existiert für $T(h)$ eine Entwicklung der Form (1) und wendet man auf T den Operator Λ_m^0 an, so läßt sich der Fehler $|T_m^{(0)} - T(0)|$ wegen (2) leicht abschätzen. Man erhält

$$\text{(8)}\quad |T_m^{(0)} - T(0)| \leqq M_{k+1} \sum_{j=0}^{m} |c_{mj}^{(0)}| h_j^{\gamma_{k+1}},$$

wobei

$$M_{k+1} = \sup_{h_j} \{|R_{k+1}(h_j)|\}.$$

Ist $T(h)$ Element eines Banach-Raums, so können die Abschätzungen im Sinne der Norm des Banach-Raums interpretiert werden. Es gilt nun der

Satz 1. *Für $T(h)$ existiere eine Entwicklung der Form (1). $T_m^{(i)} = \Lambda_m^i T$ sei der aus den $T(h_j)$, $j = i, \ldots, i+m$ extrapolierte Näherungswert für $T(0)$. Dann gilt für die Extrapolationsverfahren (3) und (7), falls $\frac{h_{j+1}}{h_j} \leqq b < 1$, $m \geqq k$ und $\gamma_{j+1} - \gamma_j \geqq \gamma > 0$, $\gamma_0 = 0$, ist, die Abschätzung*

$$|T_m^{(i)} - T(0)| \leqq M_{k+1} C(b^\gamma) h_i^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1} + \sum_{j=1}^{k} \gamma_j}$$

mit Konstanten $C(b^\gamma)$ und M_{k+1}.

Satz 1 gibt über das Konvergenzverhalten dieser Extrapolationsverfahren Auskunft. Besitzt $T(h)$ eine asymptotische Entwicklung (1) von höchstens k Gliedern ($k=\infty$ ist möglich), so konvergiert der Fehler in der n-ten Spalte von (4) für $n \leqq k+1$ wie $h_i^{\gamma_n}$ gegen 0, für $n > k+1$ wie $h_i^{\gamma_{k+1}}$. Außerdem sieht man, daß die Abschätzungen für den Extrapolationsfehler $|T_j^{(m-j)} - T(0)|$, $j=k, k+1, \ldots, m$, dieselbe obere Schranke liefern. Das legt die Vermutung nahe, daß nur die Berechnung der ersten $k+1$ Spalten von (4) sinnvoll ist. Dies wird durch die Erfahrung bestätigt. Existiert ferner für jedes k das Restglied $R_{k+1}(h)$ in (1) und bleiben die $R_{k+1}(h)$ gleichmäßig beschränkt, so zeigt Satz 1, daß $|T_m^{(i)} - T(0)|$ für $m \to \infty$ superlinear, nämlich wie $h_i^{\gamma_{m+1} b_1^{\sum_1^i \gamma_\nu}}$ gegen 0 strebt.

Zum Beweise von Satz 1 werden zwei Hilfssätze vorausgeschickt. Zunächst folgt für das Extrapolationsverfahren (3), das auf dem Neville-Algorithmus beruht, das

Lemma 1. *Falls für alle j $\gamma_j = j\gamma$, $\gamma > 0$, und $\frac{h_{j+1}}{h_j} \leqq b < 1$, dann ist für $k \leqq m$*

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} \leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma}\, \overline{C}_1(b^\gamma, k)$$

mit Konstanten $\overline{C}_1(b^\gamma, k) \leqq C_1(b^\gamma)$.

Für $h_j = \frac{h_0}{1+j}$, $\gamma = 2$ und $k = m$ gilt noch

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{(m+1)\gamma} \leqq h_0^{\gamma} \ldots h_m^{\gamma} \cdot 2(m+1) = \frac{2\, h_0^{2m+2}}{m!\,(m+1)!}.$$

Beweis. Man hat

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} = \sum_{j=0}^{m-k-1} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} + \sum_{j=m-k}^{m} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma}.$$

Nun läßt sich abschätzen

$$\begin{aligned}
\sum_{j=0}^{m-k-1} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} &= h_{m-k}^{\gamma} \ldots h_m^{\gamma} \sum_{j=0}^{m-k-1} {\prod_{\mu=0}^{m-k-1}}' \frac{h_\mu^\gamma}{|h_\mu^\gamma - h_j^\gamma|} \times \prod_{\mu=m-k}^{m} \frac{h_j^\gamma}{h_j^\gamma - h_\mu^\gamma} \\
&= h_{m-k}^{\gamma} \ldots h_m^{\gamma} \sum_{j=0}^{m-k-1} \frac{1}{\frac{h_{m-k-j-1}^\gamma}{h_{m-k-j}^\gamma} - 1} \cdots \frac{1}{\frac{h_{m-k-j-1}^\gamma}{h_{m-k-1}^\gamma} - 1} \times \\
&\quad \times \prod_{\mu=0}^{m-k-j-2} \frac{1}{1 - \frac{h_{m-k-j-1}^\gamma}{h_\mu^\gamma}} \times \prod_{\mu=m-k}^{m} \frac{1}{1 - \frac{h_\mu^\gamma}{h_{m-k-j-1}^\gamma}} \\
&\leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma} \sum_{j=0}^{m-k-1} \frac{1}{b^{-\gamma} - 1} \cdots \frac{1}{b^{-j\gamma} - 1} \times \\
&\quad \times \prod_{\mu=0}^{m-k-j-2} \frac{1}{1 - b^{(m-k-j-1-\mu)\gamma}} \times \prod_{\mu=m-k}^{m} \frac{1}{1 - b^{(\mu+k+j+1-m)\gamma}} \\
&\leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma}\, \overline{C}_1'(b^\gamma, k) \leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma}\, C_1'(b^\gamma),
\end{aligned}$$

weil der Faktor von $h_{m-k}^{\gamma} \ldots h_m^{\gamma}$ für alle m und k beschränkt bleibt.

Analog erhält man

$$\begin{aligned}
\sum_{j=m-k}^{m} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} &= h_{m-k}^{\gamma} \ldots h_m^{\gamma} \sum_{j=0}^{k} \frac{1}{\frac{h_{m-k+j-1}^\gamma}{h_{m-k+j}^\gamma} - 1} \cdots \frac{1}{\frac{h_{m-k}^\gamma}{h_{m-k+j}^\gamma} - 1} \times \\
&\quad \times \prod_{\mu=m-k+j+1}^{m} \frac{1}{1 - \frac{h_\mu^\gamma}{h_{m-k+j}^\gamma}} \times \prod_{\mu=0}^{m-k-1} \frac{1}{1 - \frac{h_{m-k+j}^\gamma}{h_\mu^\gamma}} \\
&\leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma}\, \overline{C}_1''(b^\gamma, k) \leqq h_{m-k}^{\gamma} \ldots h_m^{\gamma}\, C_1''(b^\gamma).
\end{aligned}$$

Die erste Behauptung ergibt sich jetzt mit

$$\overline{C}_1(b^\gamma, k) = \overline{C}_1'(b^\gamma, k) + \overline{C}_1''(b^\gamma, k) \leqq C_1(b^\gamma).$$

Weiter gilt mit Hilfe der Produktdarstellung von $\sin z$

$$\begin{aligned}
&\left|1 - \left(\frac{j+1}{1}\right)^2\right| \cdots \left|1 - \left(\frac{j+1}{j}\right)^2\right| \times \\
&\quad \times \left|1 - \left(\frac{j+1}{j+2}\right)^2\right| \cdots \left|1 - \left(\frac{j+1}{m+1}\right)^2\right| \cdot \left[\left|1 - \left(\frac{j+1}{m+2}\right)^2\right| \cdots\right] = \frac{1}{2},
\end{aligned}$$

woraus

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{2m+2} = h_0^2 \ldots h_m^2 \sum_{j=0}^{m} {\prod_{\mu=0}^{m}}' \frac{1}{\left|1 - \left(\frac{j+1}{\mu+1}\right)^2\right|} \leqq h_0^2 \ldots h_m^2 \sum_{j=0}^{m} 2 = \frac{2\, h_0^{2m+2}}{m!\,(m+1)!}.$$

Auf Verallgemeinerungen wird verzichtet.

Für die später häufig benutzte Folge $\mathfrak{F} = \{h_i\} = \left\{h_0, \frac{h_0}{2}, \frac{h_0}{3}, \frac{h_0}{4}, \frac{h_0}{6}, \frac{h_0}{8}, \frac{h_0}{12}, \cdots\right\}$ ergibt sich mit $\gamma = 2$

$$C_1 < 5.$$

Für den Algorithmus (7), wo $h_j = h_0 b^j$, gilt das schärfere

Lemma 2. *Ist $h_j = h_0 b^j$ und die Sequenz der γ_i in (1) so beschaffen, daß für festes $q^2 < 1$*

$$b^{\gamma_{k+1} - \gamma_l} \leqq q^2 < 1, \qquad b^{\gamma_l} \leqq q^{2(l+1)}, \qquad l \geqq 0 \text{ ganz},$$

so gilt für $k \leqq m$

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{\gamma_{k+1}} \leqq \overline{C}_2(q^2, k, l)\, h_0^{\gamma_{k+1}}\, b^{(m-k)\gamma_{k+1} + \sum_{i=1}^{k} \gamma_i},$$

wo

$$\begin{aligned}
\overline{C}_2(q^2, k, l) &= 1 \cdot (1 - q^2) \ldots (1 - q^{2l})\, [\vartheta_3(0, q)\, \vartheta_4(0, q)]^{-\frac{1}{2}} \times \\
&\quad \times \begin{cases} 1, & \text{falls } k = m \\ 2(1 + q^2) \ldots (1 + q^{2(m-k-1)}), & \text{falls } m - k - 1 \text{ fest} \\ 2(1 + q^2) \ldots (1 + q^{2k}), & \text{falls } k \text{ fest} \end{cases}
\end{aligned}$$

und allgemein für beliebiges $k \leqq m$

$$\overline{C}_2(q^2, k, l) \leqq 1 \cdot (1 - q^2) \ldots (1 - q^{2l}) \left(\frac{2^8}{q}\right)^{\frac{1}{4}} \left[\frac{\vartheta_2(0, q)}{\vartheta_2^2(0, q)\, \vartheta_4^2(0, q)}\right]^{\frac{1}{4}} = C_2(q^2, l),$$

$\vartheta_i(z, q)$ Jacobische Thetafunktionen.

Für $q \geqq \frac{1}{3}$ gelten die für diese Zwecke ausreichenden Näherungsformeln

$$[\vartheta_3(0,q)\,\vartheta_4(0,q)]^{-\frac{1}{2}} \approx \sqrt{\frac{-\ln q}{2\pi}} \exp\left(-\frac{\pi^2}{8\ln q}\right)$$

$$\left(\frac{2^8}{q}\right)^{\frac{1}{24}} \left[\frac{\vartheta_2(0,q)}{\vartheta_3^2(0,q)\,\vartheta_4^2(0,q)}\right]^{\frac{1}{2}} \approx q^{-\frac{1}{24}} \sqrt{\frac{-\ln q}{\pi}} \exp\left(-\frac{\pi^2}{6\ln q}\right).$$

Beweis. Man hat

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{\gamma_{k+1}} = h_0^{\gamma_{k+1}} \sum_{j=0}^{m} |c_{mj}^{(0)}|\, (b^{\gamma_{k+1}})^j = \frac{h_0^{\gamma_{k+1}}}{(1-b^{\gamma_1})\dots(1-b^{\gamma_m})} (b^{\gamma_{k+1}}+b^{\gamma_1})\dots(b^{\gamma_{k+1}}+b^{\gamma_k})$$

$$= \frac{(1+b^{\gamma_{k+1}-\gamma_{k+1}})\dots(1+b^{\gamma_m-\gamma_{k+1}})\cdot(1+b^{\gamma_{k+1}-\gamma_k})\dots(1+b^{\gamma_{k+1}-\gamma_1})}{(1-b^{\gamma_1})\dots(1-b^{\gamma_m})} \times$$

$$\times\, h_0^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1}+\sum_{j=0}^{k}\gamma_j}$$

$$\leqq (1-q^2)\dots(1-q^{2l}) \frac{2(1+q^2)\dots(1+q^{2(m-k-1)})\,(1+q^2)\dots(1+q^{2k})}{(1-q^2)\dots(1-q^{2(m+l)})} \times$$

$$\times\, h_0^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1}+\sum_{j=0}^{k}\gamma_j}.$$

Das Weitere folgt jetzt aus den Beziehungen

$$\vartheta_1'(0,q) = 2q^{\frac{1}{4}} \prod_{n=1}^{\infty} (1-q^{2n})^3, \qquad \vartheta_2(0,q) = 2q^{\frac{1}{4}} \prod_{n=1}^{\infty} (1+q^{2n})^2 (1-q^{2n})$$

und

$$\vartheta_1'(0,q) = \vartheta_2(0,q)\,\vartheta_3(0,q)\,\vartheta_4(0,q).$$

Am Rande sei vermerkt, daß für $\gamma_j = (l+j)\gamma$, $(q = b^{\gamma/2})$

$$\lim_{m\to\infty} \sum_{j=0}^{m} |c_{mj}^{(0)}| = \frac{1\cdot(1-q^2)\cdots(1-q^{2l})}{1\cdot(1+q^2)\cdots(1+q^{2l})} [\vartheta_3(0,q)\,\vartheta_4(0,q)]^{-\frac{1}{2}}.$$

In diesem Fall sind die Abschätzungen in Lemma 2 scharf. Speziell für $q = \frac{1}{3}$ (s. z.B. Romberg-Verfahren) ist

$$[\vartheta_3(0,\tfrac{1}{3})\,\vartheta_4(0,\tfrac{1}{3})]^{-\frac{1}{2}} = 1{,}969\dots$$

und

$$C_2\left(\frac{1}{4},0\right) = 2^{\frac{1}{2}} \left[\frac{\vartheta_2(0,\frac{1}{3})}{\vartheta_3^2(0,\frac{1}{3})\,\vartheta_4^2(0,q)}\right]^{\frac{1}{2}} = 5{,}3\dots;$$

letztere Abschätzung für $C_2(\frac{1}{4}, k, 0)$ ist scharf, falls mit m auch $m-k$ und k nach ∞ strebt.

Der Beweis von Satz 1 ist jetzt einfach.

Für das Extrapolationsverfahren (3) ergibt sich aus Lemma 1 wegen $h_j \leqq b^j h_0$

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{(k+1)\gamma} \leqq h_{m-k}^{\gamma} \dots h_m^{\gamma} C_1(b^\gamma, k) \leqq h_0^{(k+1)\gamma} b^{(m-k+m-k+1+\dots+m)\gamma} C_1(b^\gamma, k)$$

$$= C_1(b^\gamma, k)\, h_0^{(k+1)\gamma} b^{(m-k)(k+1)\gamma + \sum_{j=1}^{k} j\gamma} \leqq C_1(b^\gamma)\, h_0^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1}+\sum_{j=1}^{k}\gamma_j}$$

und für das Extrapolationsverfahren (7) mit $q^2 = b^\gamma$ und $l = 0$ unmittelbar aus Lemma 2

$$\sum_{j=0}^{m} |c_{mj}^{(0)}|\, h_j^{\gamma_{k+1}} \leqq C_2(b^\gamma, 0)\, h_0^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1}+\sum_{j=1}^{k}\gamma_j}.$$

(8) liefert jetzt

$$|T_m^{(0)} - T(0)| \leqq M_{k+1} C(b^\gamma)\, h_0^{\gamma_{k+1}} b^{(m-k)\gamma_{k+1}+\sum_{j=1}^{k}\gamma_j},$$

womit der Beweis von Satz 1 erbracht ist, denn der Übergang von $(h_0, T_m^{(0)})$ nach $(h_i, T_m^{(i)})$ kann als „Umnumerierung" der h_j interpretiert werden.

Als letztes läßt Satz 1 die Frage nach dem Konvergenzverhalten von $T_m^{(i)}$ offen, wenn von $T(h)$ nur bekannt ist, daß $\lim_{j\to\infty} T(h_j) = T(0)$. Es gilt hier

Satz 2. *Es sei* $\lim_{j\to\infty} T(h_j) = T(0)$. *Ist für alle* j *entweder* $\gamma_j = j\gamma$ *und* $\frac{h_{j+1}}{h_j} \leqq b < 1$ *(Rekursionsformel (3)) oder* $h_j = h_0 b^j$, $0 < b < 1$ *und* $\sum_{j=1}^{\infty} b^{\gamma_j}$ *konvergent (Rekursionsformel (7)), so gilt* $\lim_{m\to\infty} T_m^{(i)} = T(0)$, $i = 1, 2, \dots$.

Beweis. Für den Fall $\gamma_j = j\gamma$ ist dieser Satz bekannt ([1], [7], [3]). Für den Rest[1] folgt aus Gl. (2a) nach TOEPLITZ (s. [3], Satz 1) $\lim_{m\to\infty} T_m^{(i)} = \lim_{j\to\infty} T(h_j)$, falls

1. $\sum_{j=i}^{i+m} c_{mj}^{(i)} = 1$,
2. $\sum_{j=i}^{i+m} |c_{mj}^{(i)}| \leqq \text{const}$, für alle m
3. $\lim_{m\to\infty} c_{mj}^{(i)} = 0$, $i \leqq j$, j fest.

Diese Bedingungen sind erfüllt, denn 1. folgt aus Gl. (2b), 2. ergibt sich aus (6) zu

$$\sum_{j=i}^{i+m} |c_{mj}^{(i)}| = \prod_{j=1}^{m} \frac{1+b^{\gamma_j}}{1-b^{\gamma_j}} \leqq \prod_{j=1}^{\infty} \frac{1+b^{\gamma_j}}{1-b^{\gamma_j}} = \text{const}, \quad \text{da } \sum_{j=1}^{\infty} b^{\gamma_j} \text{ konvergiert},$$

und 3. folgt schließlich daraus, daß die $c_{m,j+i}^{(i)} = c_{mj}^{(0)}$ im wesentlichen die elementarsymmetrischen Funktionen der b^{γ_ϱ} sind, $\varrho = 1, \dots, m$, was zu der Abschätzung führt

$$|c_{mj}^{(i)}| \leqq m^{j-i} b^{\gamma_1(m+i-j)} \prod_{n=1}^{\infty} (1-b^{\gamma_n})^{-1}.$$

Die folgenden Beispiele illustrieren das Konvergenzverhalten bei der Extrapolation. Im ersten Beispiel einer Quadratur wurde $T_m^{(0)}$ aus der Trapezsumme (s. [3], Gl. (14))

$$T(h) = \int_0^1 \sqrt{x}\,dx + \tau_1 h^{\frac{1}{2}} + \tau_2 h^2 + \tau_3 h^4 + \dots + R(h)$$

für $h_m = 2^{-m}$ einmal nach Gl. (3) mit $\gamma = 2$ und einmal nach (7) mit $b = \frac{1}{2}$, $\gamma_1 = \frac{3}{2}$, $\gamma_j = 2(j-1)$, $j > 1$, berechnet (vgl. auch [10]).

Tabelle 1

m	$T_0^{(m)} = T(h_m)$	$T_m^{(0)}, \gamma = 2$	$T_m^{(0)}, \gamma_1 = \frac{3}{2}, \dots$
3	0,6581 3022 1626	0,6636 0756 9117	0,6666 6614 7251
4	0,6635 8119 6876	0,6655 9286 5132	0,6666 6666 3132
5	0,6655 5893 6282	0,6662 8769 9043	0,6666 6666 6667

[1] Für die Folge $(\gamma_j) = (\frac{1}{2}, \frac{3}{2}, 2, \frac{5}{2}, \frac{7}{2}, 4, \dots)$ findet sich allerdings eine entsprechende Bemerkung in [10], § 4.

Das nächste Beispiel betrifft die Integration der Differentialgleichung $y'=y$, $y(0)=1$, mit Hilfe der Runge-Kutta-Formeln. In diesem Fall lautet die Entwicklung (vgl. [4])

$$T(h, x)=y(x)+h^4\tau_1(x)+h^5\tau_2(x)+\cdots+R(x, h).$$

Zur Extrapolation wurde für $h_m=2^{-m}$ Gl. (7) mit $b=\frac{1}{2}$, $\gamma_j=3+j$ verwendet.

Tabelle 2

m	$T_0^{(m)}=T(h_m, 1)$	$T_m^{(0)}$
0	2,70833333331	
1	2,71734619139	2,71794704860
2	2,71820993919	2,71827786025
3	2,71827684440	2,71828181110
4	2,71828150036	2,71828182844
.	.	
.	.	
8	2,71828182844	
	(e=2,71828182846)	

4.

Die bisher untersuchten Verfahren beruhten auf der Interpolation durch polynomartige Ausdrücke. Es liegt nun nahe, im Fall $\gamma_j=j\gamma$ statt dessen rationale Funktionen zu benutzen und als Näherungswert für $T(0)$ den Wert $T_{\mu,\nu}^{(i)}=\hat{T}_{\mu,\nu}^{(i)}(0)$ derjenigen rationalen Funktion in h^γ

$$\hat{T}_{\mu,\nu}^{(i)}(h)\equiv\frac{P_{\mu,\nu}^{(i)}(h)}{Q_{\mu,\nu}^{(i)}(h)}=\frac{p_0^{(i)}+p_1^{(i)}h^\gamma+\cdots+p_\mu^{(i)}h^{\mu\gamma}}{q_0^{(i)}+q_1^{(i)}h^\gamma+\cdots+q_\nu^{(i)}h^{\nu\gamma}} \tag{9}$$

zu nehmen, für welche gilt

$$\hat{T}_{\mu,\nu}^{(i)}(h_j)=T(h_j), \qquad j=i, i+1, \ldots, i+\mu+\nu=i+m. \tag{10}$$

In [12] wurde gezeigt, wie man die Werte $\hat{T}_{\mu,\nu}^{(i)}(\xi)$ an einer festen Stelle ξ aus den gegebenen Werten $T(h_j)$ rekursiv berechnen kann, ohne die rationale Funktion $\hat{T}_{\mu,\nu}^{(i)}(h)$ selbst, d.h. ihre Koeffizienten, zu bestimmen. Setzt man in die in [12] angegebenen Formeln $\xi=0$ ein, so erhält man

$$T_{0,0}^{(i)}=T(h_i)$$

$$T_{\mu,0}^{(i)}=\frac{h_i^\gamma T_{\mu-1,0}^{(i+1)}-h_{i+\mu}^\gamma T_{\mu-1,0}^{(i)}}{h_i^\gamma-h_{i+\mu}^\gamma}$$

$$T_{0,\nu}^{(i)}=\frac{h_i^\gamma-h_{i+\nu}^\gamma}{\dfrac{h_i^\gamma}{T_{0,\nu-1}^{(i+1)}}-\dfrac{h_{i+\nu}^\gamma}{T_{0,\nu-1}^{(i)}}}$$

und für $\mu, \nu\geqq 1$

$$T_{\mu,\nu}^{(i)}=\frac{h_i^\gamma T_{\mu-1,\nu}^{(i+1)}(T_{\mu-1,\nu}^{(i)}-T_{\mu-1,\nu-1}^{(i+1)})-h_{i+\mu+\nu}^\gamma T_{\mu-1,\nu}^{(i)}(T_{\mu-1,\nu}^{(i+1)}-T_{\mu-1,\nu-1}^{(i+1)})}{h_i^\gamma(T_{\mu-1,\nu}^{(i)}-T_{\mu-1,\nu-1}^{(i+1)})-h_{i+\mu+\nu}^\gamma(T_{\mu-1,\nu}^{(i+1)}-T_{\mu-1,\nu-1}^{(i+1)})}$$

oder

$$T_{\mu,\nu}^{(i)}=\frac{h_i^\gamma T_{\mu,\nu-1}^{(i+1)}(T_{\mu,\nu-1}^{(i)}-T_{\mu-1,\nu-1}^{(i+1)})-h_{i+\mu+\nu}^\gamma T_{\mu,\nu-1}^{(i)}(T_{\mu,\nu-1}^{(i+1)}-T_{\mu-1,\nu-1}^{(i+1)})}{h_i^\gamma(T_{\mu,\nu-1}^{(i)}-T_{\mu-1,\nu-1}^{(i+1)})-h_{i+\mu+\nu}^\gamma(T_{\mu,\nu-1}^{(i+1)}-T_{\mu-1,\nu-1}^{(i+1)})}.$$

Wählt man für die Extrapolation folgende Sequenz der (μ, ν): $(0,0)\to(0,1)\to(1,1)\to(1,2)\to\cdots\to(i,i)\to(i,i+1)\to(i+1,i+1)\to\cdots$ (Zählergrad = Nennergrad (-1)) und schreibt man $T_{\mu+\nu}^{(i)}$ für $T_{\mu,\nu}^{(i)}$, $(\mu+\nu=m)$, so reduzieren sich die obigen Formeln auf

$$T_{-1}^{(i)}=0$$

$$T_0^{(i)}=T(h_i)$$

$$T_m^{(i)}=T_{m-1}^{(i+1)}+\frac{T_{m-1}^{(i+1)}-T_{m-1}^{(i)}}{\left(\dfrac{h_i}{h_{i+m}}\right)^\gamma\left[1-\dfrac{T_{m-1}^{(i+1)}-T_{m-1}^{(i)}}{T_{m-1}^{(i+1)}-T_{m-2}^{(i+1)}}\right]-1}, \qquad m\geqq 1. \tag{11}$$

Man erkennt die formale Ähnlichkeit mit dem Neville-Algorithmus (3). Verwendet man die Formeln (11), so muß man allerdings prüfen, ob die auftretenden Nenner nicht zu klein sind. Insbesondere besteht diese Gefahr bei größeren m, da $|T_{m-1}^{(i+1)}-T_{m-2}^{(i+1)}|$ bald sehr klein wird. Dieses Verhalten ist jedoch ungefährlich, weil es nur die schnelle Konvergenz des Verfahrens bestätigt und in der Regel nicht auf algebraischen Ausnahmesituationen beruht, wie sie in [12] ausführlich beschrieben sind.

Bei der Extrapolation durch rationale Funktionen läßt sich nun ebenfalls ein Ausdruck für den Fehler $T_{\mu,\nu}^{(i)}-T(0)$ angeben. Ist nämlich $k\leqq\mu$, so gilt wegen (9) und (10)

$$P_{\mu,\nu}^{(i)}(h_j)=T(h_j)\,Q_{\mu,\nu}^{(i)}(h_j), \qquad j=i, i+1, \ldots, i+\mu+\nu=i+m.$$

Multipliziert man diese Gleichungen mit den Lagrange-Koeffizienten $c_{mj}^{(i)}$ (s. (5)) und summiert über j, so erhält man wegen (1) und (2)

$$P_{\mu,\nu}^{(i)}(0)=T(0)\,Q_{\mu,\nu}^{(i)}(0)+\sum_{j=i}^{i+m}c_{mj}^{(i)}h_j^{(k+1)\gamma}R_{k+1}(h_j)\,Q_{\mu,\nu}^{(i)}(h_j)$$

oder falls $Q_{\mu,\nu}^{(i)}(0)=q_0^{(i)}\neq 0$

$$T_{\mu,\nu}^{(i)}-T(0)=\frac{1}{q_0^{(i)}}\sum_{j=i}^{i+m}c_{mj}^{(i)}h_j^{(k+1)\gamma}R_{k+1}(h_j)\,Q_{\mu,\nu}^{(i)}(h_j). \tag{12}$$

Leider enthält dieser Ausdruck für den Extrapolationsfehler noch das Nennerpolynom $Q_{\mu,\nu}^{(i)}(h_j)$, so daß er zunächst nur bedingten Wert hat. Wäre allerdings $R_{k+1}(h_j)=\text{const}$, $j=\ldots, i+m$, so ergäbe sich für $k=\mu$

$$T_{\mu,\nu}^{(i)}-T(0)=(-1)^m h_i^\gamma\ldots h_{i+m}^\gamma\frac{q_\nu^{(i)}}{q_0^{(i)}}R_{k+1}, \qquad \mu+\nu=m.$$

Gilt lediglich $\lim_{j\to\infty}T(h_j)=T(0)$, so konvergiert $T_{\mu,\nu}^{(i)}$ für $\mu+\nu\to\infty$ gegen $T(0)$ (vgl. Satz 2), falls $\frac{h_{j+1}}{h_j}\leqq b<1$ und

$$\left|\frac{Q_{\mu,\nu}^{(i)}(h_j)}{Q_{\mu,\nu}^{(i)}(0)}\right|\leqq C_i, \quad \text{für alle } \mu, \nu, j. \tag{13}$$

Denn wie in (12) folgt

$$T_{\mu,\nu}^{(i)}=\frac{1}{Q_{\mu,\nu}^{(i)}(0)}\sum_{j=i}^{i+m}c_{mj}^{(i)}\,Q_{\mu,\nu}^{(i)}(h_j)\,T(h_j)=\sum_{j=i}^{i+m}\bar{c}_{mj}^{(i)}\,T(h_j), \qquad \bar{c}_{mj}^{(i)}=c_{mj}^{(i)}\frac{Q_{\mu,\nu}^{(i)}(h_j)}{Q_{\mu,\nu}^{(i)}(0)}$$

und mit den $c_{mj}^{(i)}$ (s. [3], Satz 1) erfüllen wegen (13) auch die $\bar{c}_{mj}^{(i)}$ die Bedingungen von TOEPLITZ

1. $\sum_{j=i}^{i+m} \bar{c}_{mj}^{(i)} = 1$

2. $\sum_{j=i}^{i+m} |\bar{c}_{mj}^{(i)}| \leqq \text{const}$, für alle m

3. $\lim_{m\to\infty} \bar{c}_{mj}^{(i)} = 0$, $i \leqq j$, j fest.

5.

Überschaubare Verhältnisse bei der Extrapolation mit rationalen Funktionen erhält man bei der *Quadratur durch Trapezsummenextrapolation*. Hier ist mit $\gamma=2$ und $\mu=k$

$$T(h) = \int_0^1 f(x)\,dx + \tau_1 h^2 + \cdots + \tau_k h^{2k} + \\ + (-1)^k h^{2k+2} \int_0^1 f^{(2k+2)}(x) \sum_{n=1}^{\infty} \frac{2\left(1-\cos 2n\pi\frac{x}{h}\right)}{(2n\pi)^{2k+2}}\,dx,$$

woraus

$$T_{k,m-k}^{(i)} - \int_0^1 f(x)\,dx = \frac{(-1)^k}{q_0^{(i)}} \int_0^1 f^{(2k+2)}(x) \times \\ \times \left\{ \sum_{j=i}^{i+m} c_{mj}^{(i)} h_j^{2k+2} Q_{k,m-k}^{(i)}(h_j) \sum_{n=1}^{\infty} \frac{2\left(1-\cos 2n\pi\frac{x}{h_j}\right)}{(2n\pi)^{2k+2}} \right\} dx.$$

Wenn die Funktion $S(x)$ in der geschweiften Klammer ihr Vorzeichen nicht ändert, liefern Mittelwertsatz und Formel (2)

$$(14)\quad T_{k,m-k}^{(i)} - \int_0^1 f(x)\,dx = (-1)^m h_i^2 \cdots h_{i+m}^2 \frac{q_{m-k}^{(i)}}{q_0^{(i)}} \frac{B_{2k+2}}{(2k+2)!} f^{(2k+2)}(\xi), \quad 0<\xi<1,$$

B_{2k+2} Bernoulli-Zahlen.

Für $k=m$ (Nennergrad 0) erhält man die bereits in [3] abgeleitete Beziehung

$$T_{m,0}^{(i)} - \int_0^1 f(x)\,dx = (-1)^m h_i^2 \cdots h_{i+m}^2 \frac{B_{2m+2}}{(2m+2)!} f^{(2m+2)}(\xi)$$

zurück.

Im Gegensatz zum Fall $k=m$ ist für $k<m$ eine a priori Entscheidung über Vorzeichenwechsel von $S(x)$ nicht möglich, weil $S(x)$ von dem zu berechnenden Nennerpolynom $Q_{k,m-k}^{(i)}$ abhängt. Trotzdem erlaubt die Beziehung (14) wenigstens einen qualitativen Überblick über den zu erwartenden Fehler: Da nämlich im allgemeinen $\left|\frac{B_{2k+2}}{(2k+2)!} f^{(2k+2)}(x)\right|$ mit k über alle Grenzen wächst, scheint (14)

zunächst den Schluß nahezulegen, daß $k=0$ den kleinsten Fehler bei der Extrapolation liefert. Andererseits gehen jedoch in $\frac{q_{m-k}^{(i)}}{q_0^{(i)}}$, wie sich zeigen läßt, Ableitungen bis zur Ordnung $f^{(2r+2)}(x)$, $r=\max\{k, m-k\}$, in komplizierter Weise ein. Man wird also in der Regel am günstigsten mit $k=\left[\frac{m}{2}\right]$ (Zählergrad $\approx$ Nennergrad) arbeiten. Die numerische Erfahrung auf der Rechenanlage PERM (TH München) hat diesen Schluß bestätigt.

In den folgenden Beispielen zur Trapezsummen-Extrapolation sind die Ergebnisse bei unterschiedlicher Extrapolation gegenübergestellt. Als Schrittweitenfolge wurde die bereits in [3] erwähnte Folge $\mathfrak{F} = \left\{L, \frac{L}{2}, \frac{L}{3}, \frac{L}{4}, \frac{L}{6}, \frac{L}{8}, \frac{L}{12}, \ldots\right\}$ gewählt (L Länge des Integrationsintervalls), die auch hier in bezug auf Arbeitsaufwand und Genauigkeit die besten Resultate liefert. In der folgenden Tabelle gibt m die Anzahl der Extrapolationsschritte, A_m die Anzahl der bis zur Schrittweite h_m zu berechnenden Funktionswerte und $T(h_m)$ die berechneten Trapezsummen an. Die weiteren Spalten enthalten in der Reihenfolge die extrapolierten Werte bei der Interpolation durch

a) Polynome,

b) rationale Funktionen (Zählergrad = Nennergrad (-1)),

c) reziproke Polynome.

(Es wurden zwei „schlechte" und ein „gutes" Beispiel ausgewählt!)

Tabelle 3

	m	A_m	$T(h_m)$	$T_{m,0}^{(0)}$	$T_{[m/2],m-[m/2]}^{(0)}$	$\bar{T}_{0,m}^{(0)}$
$\int_1^4 \frac{dx}{x}$	4	9	1,0058 5652 083	1,0000 0281 498	1,0000 0100 964	1,0000 0356 225
	5	13	1,0033 0707 022	1,0000 0014 667	1,0000 0003 995	1,0000 0017 065
	6	17	1,0014 7397 628	1,0000 0000 488	1,0000 0000 080	1,0000 0000 541
	7	25	1,0008 2994 518	1,0000 0000 010	1,0000 0000 000	1,0000 0000 010
	8	33	1,0003 6913 108	1,0000 0000 000	—	1,0000 0000 000
$\int_0^{\frac{\pi}{2}} \sin x\,dx$	2	5	0,9770 4861 6664	$0{,}9999\,9849\,5995_{10^{-1}}$	$0{,}9999\,9957\,0210_{10^{-1}}$	$0{,}9994\,8556\,0817$
	3	7	0,9871 1580 0974	1,0000 0005 5625	1,0000 0000 656	1,0000 0816 914
	4	9	0,9942 8188 8297	$0{,}9999\,9999\,9992_{10^{-1}}$	1,0000 0000 000	0,9999 9942 886
	5	13	0,9967 8517 1887	1,0000 0000 000	—	1,0000 0000 021
	6	17	0,9985 7169 7907	—	—	1,0000 0000 000
$-\int_{-\frac{\pi}{6}}^{\pi-\frac{\pi}{6}} x\sin 3x\,dx$	4	9	0,2741 5567 7808	0,1882 3479 9393	0,2226 5227 5416	0,2223 4864 0655
	5	13	0,2498 1116 1476	0,2237 4428 3778	0,2222 2326 8686	0,2222 2334 6708
	6	17	0,2340 0650 8335	0,2221 9366 2180	0,2222 2221 7815	0,2222 2222 5452
	7	25	0,2287 6077 5525	0,2222 2251 4134	0,2222 2222 2178	0,2222 2222 2186
	8	33	0,2251 0016 5925	0,2222 2222 0926	0,2222 2222 2223	0,2222 2222 2223
	9	49	0,2238 3559 4426	0,2222 2222 2186	—	
	10	65	0,2229 3754 7635	0,2222 2222 2213	—	

Man sieht, daß die Werte $T_{m/2,m-[m/2]}^{(0)}$ jeweils am schnellsten konvergieren.

Ein weiteres interessantes Beispiel liefert die Integration der Differentialgleichung $y' = f(x, y)$ nach dem Euler-Verfahren $y_{n+1} = y_n + h f(x_n, y_n)$. Das

Fehlerverhalten ist hier gegeben durch

$$T(h, x) = y(x) + h\tau_1(x) + h^2\tau_2(x) + h^3\tau_3(x) + \cdots. \tag{15}$$

Als spezielles Beispiel wurde wieder $y' = y$, $y(0) = 1$ gewählt, vgl. dazu auch [1], § 10. $\left(\text{In diesem Fall ist } T(h, x) = e^x - h\frac{x}{2}e^x + h^2\left(\frac{x}{3} + \frac{x^2}{8}\right)e^x + \cdots\right)$.

Mit der Schrittweite $h_m = 2^{-m}$ ergibt sich

Tabelle 4

m	$T(h_m, 1)$	$T^{(0)}_{m,0}$	$T^{(0)}_{[m/2], m-[m/2]}$
3	2,56578451393	2,71387899486	2,71815516173
4	2,63792849736	2,71802983460	2,71828078589
5	2,67699012937	2,71827434380	2,71828181791
6	2,69734495258	2,71828171514	2,71828182855
·			
·			
9	2,71563200012	2,71828182855	—

Man erhält also mit der einfachen Eulerschen Methode für $m = 6$ durch Bildung von $T^{(0)}_{3,3}$ praktisch dieselbe Genauigkeit wie mit den Runge-Kutta-Formeln mit anschließender Extrapolation (vgl. das Beispiel am Ende von § 3). Die Anzahl der zu berechnenden Funktionswerte $f(x_n, y_n)$ beträgt in diesen beiden Fällen 121 bzw. 120. Das Verhältnis verschiebt sich sogar noch zugunsten der Euler-Integration, falls die Schrittweitenfolge $\mathfrak{H}$ benutzt wird. In diesem Fall erhält man durch Berechnung von nur 30 Funktionswerten das Ergebnis $T^{(0)}_{3,3} = 2{,}71828182652$. Höhere Genauigkeiten können hier mit $\mathfrak{H}$ im Gegensatz zur Trapezsummen-Extrapolation leider nicht erzielt werden, da wegen der lediglich nach Potenzen von h fortschreitenden Entwicklung (15) durch die nahe bei 1 liegenden Quotienten $\frac{h_{j+1}}{h_j}$ die Rundungsfehler stark ins Gewicht fallen.

Das folgende ALGOL-Programm zur Berechnung von $\int_{UG}^{OG} f(x)\,dx$ benutzt zur Trapezsummen-Extrapolation rationale Funktionen (Zählergrad $\approx$ Nennergrad, s. Formel (11)). Verwendet wird die Folge $\mathfrak{H} = \left\{L, \frac{L}{2}, \frac{L}{3}, \frac{L}{4}, \frac{L}{6}, \frac{L}{8}, \ldots\right\}$, $L = OG - UG$. Das Programm ist optimal konstruiert, bereits berechnete $f(n\,h_j)$ werden zur Ermittlung neuer $T^{(m)}_0$ wieder verwendet. Bezeichnet A_m die Anzahl der $f(n\,h_j)$ zur Berechnung von $T^{(0)}_0, \ldots, T^{(m)}_0$, so ist

$$A_m = 1 + \begin{cases} 2^{\frac{m}{2}+1}, & m \text{ gerade} \\ 2^{\frac{m+1}{2}} + 2^{\frac{m-1}{2}}, & m \text{ ungerade} \end{cases}, \quad m \geqq 2$$

(vgl. Tabelle 1 in [3]).

Das Programm ermittelt zu vorgegebenem ord die $T^{(i)}_{m-i}$ für

$$m = 0, 1, \ldots, \text{ord}$$

$$i = m, m-1, \ldots, r(m) \quad \text{mit} \quad r(m) = \max\{0, m-7\}$$

und liefert als Näherung für $\int_{UG}^{OG} f(x)\,dx$ den Wert $T^{(r(\text{ord}))}_{\text{ord}-r(\text{ord})}$.

(Die Erfahrung zeigt, daß es keinen Sinn hat, über mehr als 8 Stützpunkte $(h_j, T^{(j)}_0)$ zu extrapolieren.)

Voraussetzung für die schnelle Konvergenz der $T^{(i)}_{m-i}$ ist allerdings die Existenz einer Entwicklung (1) für $T(h)$ mit $\gamma_j = 2j$ und $k \geqq 7$ (trifft zu, falls $\int_{UG}^{OG} |f^{(16)}(x)|\,dx$ existiert).

```
real procedure Trapezsummenextrapolation (ug, og, ord, eps) procedure: (f);
comment Die Prozedur Trapezsummenextrapolation liefert den Näherungswert
        für das Integral der Funktion f(x) zwischen der unteren Grenze ug
        und der oberen Grenze og, den man durch ein Extrapolationsver-
        fahren mit rationalen Funktionen von der Ordnung ord (≧2) erhält.
        Die Extrapolation wird vor ihrem natürlichen Ende abgebrochen,
        wenn sich zwei aufeinanderfolgende Näherungswerte t[i] und t[i+1]
        für das Integral um weniger als eps×abs(t[i+1]) unterscheiden;
value ug, og, ord, eps;
real ug, og, eps;
integer ord;
real procedure f;
begin real h, e, t0, t2a, t2, tn, t3, ha, hg;
      integer m, nn, i;
      integer array n[0:ord];
      array t[0:7];
      boolean bo;

      procedure extr(m);
      value m;
      integer m;
      begin real v, d, u, hv, hu;
            integer i, mr;
            v := 0; u := t[0]; h := t[0] := tn;
            if m > 7 then mr := 7 else mr := m;
            for i := 1 step 1 until mr do
            begin d := n[m]/n[m-i]; d := d×d;
                  hv := h - v; hu := h - u;
                  if hv ≠ 0 then
                    begin h := h + hu/(d×(1 - hu/hv) - 1);
                          v := u; u := t[i]; t[i] := h;
                    end
                  else go to ende;
            end;
      end extr;
```

```
    n[0]:=1; n[1]:=2; n[2]:=3;
    for i:=3 step 1 until ord do n[i]:=n[i-2]×2;

    e:=og-ug; bo:=true; t0:=(f(og)+f(ug))/2;
    t[0]:=t0×e; t2a:=t2:=f(ug+e/2); tn:=(t0+t2)×e/2;
    extr(1);
    t3:=f(ug+e/3)+f(og-e/3); tn:=(t0+t3)×e/3;
    extr(2); ha:=h;
  for m:=3 step 1 until ord do
  begin nn:=n[m]; hg:=e/nn;
        if bo then
          begin for i:=1 step 2 until nn do
                      t2:=t2+f(ug+i×hg);
                tn:=(t2+t0)×hg;
          end
        else
          begin for i:=1 step 6 until nn, i:=5 step 6 until nn do
                      t3:=t3+f(ug+i×hg);
                tn:=(t3+t2a+t0)×hg; t2a:=t2;
          end;
        extr(m); bo:=¬bo;
        if abs(ha-h)<abs(h)×eps then go to ende;
        ha:=h;
  end;
ende: trapezsummenextrapolation:=h;
end trapezsummenextrapolation
```

Literatur

[1] BAUER, F. L., H. RUTISHAUSER, and E. STIEFEL: New Aspects in Numerical Quadrature. Proc. of Symposia in Applied Mathematics **15**, Am. Math. Soc. (1963).

[2] BOLTON, H. C., and H. I. SCOINS: Eigenvalues of Differential Equations by Finite-Difference Methods. Proc. of the Cambridge Phil. Soc. **52**, 215–229 (1956).

[3] BULIRSCH, R.: Bemerkungen zur Romberg-Integration. Num. Math. **6**, 6–16 (1964).

[4] GRAGG, W.: Repeated Extrapolation to the Limit in the Numerical Solution of Ordinary Differential Equations. Thesis UCLA (1963).

[5] HARTREE, D. R.: Numerical Analysis. Oxford: At the Clarendon Press 1958.

[6] KUNTZMANN, J.: Méthodes numériques, interpolation-dérivées. Paris: Dunod 1959.

[7] LAURENT, P.-J.: Un théorème de convergence pour le procédé d'extrapolation de RICHARDSON. Comptes Rendus de l'Académie des Sciences (Paris) **256**, 1435–1437 (1963).

[8] RICHARDSON, C., and J. GAUNT: The Deferred Approach to the Limit. Trans. Roy. Soc. Lond. **226**, 300–361 (1927).

[9] ROMBERG, W.: Vereinfachte numerische Integration. Det. Kong. Norske Videnskabers Selskab Forhandlinger **28**, Nr. 7, Trondheim 1955.

[10] RUTISHAUSER, H.: Ausdehnung des Rombergschen Prinzips. Num. Math. **5**, 48–54 (1963).

[11] STETTER, H.-J.: Asymptotic Expansions for the Error of Discretisation Algorithms for Non-linear Functional Equations. Erscheint in Num. Math.

[12] STOER, J.: Über zwei Algorithmen zur Interpolation mit rationalen Funktionen. Num. Math. **3**, 285–304 (1961).

[13] COREY: The American math. Monthly **13** (1906).

[14] SALVADORI, M. G., and M. L. BARON: Numerical Methods in Engineering. Englewood Cliffs, N. J.: Prentice-Hall 1952.

[15] LYNESS, J. N., and B. J. J. MCHUGH: Integration over multidimensional Hypercubes, I. A progressive Procedure. The Computer J. **6**, 264–270 (1963).

[16] LAURENT, P.-J.: Formules de quadrature approchée sur domaines rectangulaires convergentes pour toute fonction intégrable Riemann. Comptes Rendus de l'Académie des Sciences (Paris) **258**, 793–801 (1964).

[17] NAVOT, I.: The Euler-Maclaurin Functional for Functions with a Quasi-Step Discontinuity. Math. Comp. **17**, 337–345 (1963).

Mathematisches Institut der
Technischen Hochschule
8 München 2, Arcisstr. 21

(Eingegangen am 26. März 1964)

A N E X O - 2

Programa em FORTRAN IV

PROJ
= = = =

```
C       R(2)=1.
C       COMETA  KOHLER
        IMPLICIT DOUBLE PRECISION  (A-Z)
        INTEGER I,I2,I3,I1,J
        REAL LAT,LOM,E,ALFA,DELTA,XS,YS,ZS,TS,ORO,TT,TE,TZ,ALFAP,DELTAP,RO
       *P,MP,DIX,DIO,DII,TA,TD,ALPMI,ALZEP,ALPI,ALPII,DEPMI,DEPZE,DEPI,DEP
       *II,EMI,EZE,EI,EII,II,XXF,YYF,ZZF,VVX,VVY,VVZ,FFX,FFY,FFZ,T,Y,KK,RR
       *A,PALFA,PDELTA,PALF,PDEF,TJ,ALFAO,DELTAO,XSA,XSZ,XSI,XSII,YSA,YSZ,
       *YSI,YSII,ZSA,ZSZ,ZSI,ZSII
        DIMENSION TT(3),ALFA(3),DELTA(3),XS(3),YS(3),ZS(3),TS(3),TE(3),D(3
       *),M(3),GA(3),LE(3),MA(3),RO(3),QS(3),XA(3),YA(3),ZA(3),R(3),XT(3),
       *TAF(3),SDF(3),ERRA(3),ERRD(3),MF(3),UF(3),VG(3),MFG(3),UFG(3),RV(3
       *),DS(3),DX(6),DY(6),DZ(3),ALFAP(3),DELTAP(3),ROP(3),XP(3),YP(3),ZP
       *(3),FX(3),FY(3),FZ(3),TA(3),TD(3),ALPMI(3),ALZEP(3),ALPI(3),ALPII(
       *3),DEPMI(3),DEPZE(3),DEPI(3),DEPII(3),VX(3),VY(3),VZ(3),Y(6),YT(3)
       *,YM(3),ZM(3),V(3),U(3),XF(3),YF(3),ZF(3),QSI(3),ETA(3),ZET(3),AF(3
       *),DF(3),ZT(3),LA(3),MI(3),NI(3),A(3),B(3),C(3),OU(3),RA(3),XM(3),P
       *R(3),ROPMI(3),ROPZE(3),ROPI(3),ROPII(3),XAA(3),YAA(3),ZAA(3),XAAA(
       *3),YAAA(3),ZAAA(3),XFA(3),YFA(3),ZFA(3),ALFAF(3),DELTAF(3),TJ(3),A
       *LFAO(3),DELTAO(3),ERROA(3),ERROD(3),VAX(3),VAY(3),VAZ(3),XSA(3),XS
       *Z(3),XSI(3),XSII(3),YSA(3),YSZ(3),YSI(3),YSII(3),ZSA(3),ZSZ(3),ZSI
       *(3),ZSII(3)
        READ(5,1) (TT(I),I=1,3),(ALFA(I),I=1,3),(DELTA(I),I=1,3),(XS(I),I=
       *1,3),(YS(I),I=1,3),(ZS(I),I=1,3),(TS(I),I=1,3)
        READ(5,2)K,LAT,LOM,E,ORO
        READ(5,25) (TA(I),I=1,3),(TD(I),I=1,3),(ALPMI(I),I=1,3),(ALZEP(I),
       *I=1,3),(ALPI(I),I=1,3),(ALPII(I),I=1,3),(DEPMI(I),I=1,3),(DEPZE(I)
       *,I=1,3),(DEPI(I),I=1,3),(DEPII(I),I=1,3),MP,(ROPMI(I),I=1,3),(ROPZ
       *E(I),I=1,3),(ROPI(I),I=1,3),(ROPII(I),I=1,3)
        READ(5,70) (XSA(I),I=1,3),(XSZ(I),I=1,3),(XSI(I),I=1,3),(XSII(I),I
       *=1,3),(YSA(I),I=1,3),(YSZ(I),I=1,3),(YSI(I),I=1,3),(YSII(I),I=1,3)
       *,(ZSA(I),I=1,3),(ZSZ(I),I=1,3),(ZSI(I),I=1,3),(ZSII(I),I=1,3)
        NN=1
        R(2)=1.
        CALL SOL(XSA,XSZ,XSI,XSII,YSA,YSZ,YSI,YSII,ZSA,ZSZ,ZSI,ZSII,XS,YS,
       *ZS,TT,TA,TD)
        DO 19 I=1,3
        ALFAO(I)=ALFA(I)*180/(3.1415926535*15)
   19   DELTAO(I)=DELTA(I)*180/3.1415926535
        CALL TOPOC (XS,YS,ZS,TS,LAT,ORO,XT,YT,ZT)
   36   CALL CODIR(ALFA,DELTA,LA,MI,NI,A,B,C,ED)
        TE(1)=TT(2)-TT(1)
        TE(2)=TT(3)-TT(2)
        TE(3)=TE(1)+TE(2)
        CALL RAIO (A,B,C,XT,YT,ZT,LA,MI,NI,TE,K,R,D)
        SI=(TE(1)/TE(3))*(1+(K/6)*(TE(3)**2-TE(1)**2)*(1/(R(2)**3)))
        SII=(TE(2)/TE(3))*(1+(K/6)*(TE(3)**2-TE(2)**2)*(1/(R(2)**3)))
C       SI,A RAZAO ENTRE AS AREAS S1 E S3;SII, A RAZAO ENTRE S2 E S3
   31   L=SII*XT(1)-XT(2)+SI*XT(3)
        M=SII*YT(1)-YT(2)+SI*YT(3)
        N=SII*ZT(1)-ZT(2)+SI*ZT(3)
        D(1)=(1/SII)*(A(1)*L+B(1)*M+C(1)*N)
        D(2)=A(2)*L+B(2)*M+C(2)*N
        D(3)=(1/SI)*(A(3)*L+B(3)*M+C(3)*N)
        R1=D(1)**2-2*D(1)*(LA(1)*XT(1)+MI(1)*YT(1)+NI(1)*ZT(1))+XT(1)**2+Y
       *T(1)**2+ZT(1)**2
        R(1)=DSQRT(R1)
        R3=D(3)**2-2*D(3)*(LA(3)*XT(3)+MI(3)*YT(3)+NI(3)*ZT(3))+XT(3)**2+Y
       *T(3)**2+ZT(3)**2
```

```
      R(3)=DSQRT(R3)
      CALL TEMPO (TT,TE,D)
      F1=(K/12)*(TE(1)**2+TE(1)*TE(2)-TE(2)**2)
      F2=(K/12)*(TE(1)**2+3*TE(1)*TE(2)+TE(2)**2)
      F3=(K/12)*(-TE(1)**2+TE(1)*TE(2)+TE(2)**2)
      SI=(TE(1)/TE(3))*((1+F3/R(3)**3)/(1-F2/R(2)**3))
      SII=(TE(2)/TE(3))*((1+F1/R(1)**3)/(1-F2/R(2)**3))
      CALL RAIOS (SI,SII,A,B,C,D,RA,XT,YT,ZT,LA,MI,NI,R)
      DO 7 I=1,3
    7 R(I)=RA(I)
      CALL TEMPO (TT,TE,D)
      CALL HELIO (LA,MI,NI,D,XT,YT,ZT,XA,YA,ZA)
      CALL QUI (R,XA,YA,ZA,QU)
      H(1)=(K*TE(1)**2)/(QU(1)**2*(R(1)+R(2)+(2*SQRT(2)/3)*QU(1)))
      H(2)=(K*TE(2)**2)/(QU(2)**2*(R(2)+R(3)+(2*SQRT(2)/3)*QU(2)))
      H(3)=(K*TE(3)**2)/(QU(3)**2*(R(3)+R(1)+(2*SQRT(2)/3)*QU(3)))
      DO 3 I=1,3
      HH=H(I)
      CALL CUBIC(HH,GANEXT)
    3 GA(I)=GANEXT
      SI=(TE(1)/TE(3))*(GA(3)/GA(1))
      SII=(TE(2)/TE(3))*(GA(3)/GA(2))
      CALL RAIOS (SI,SII,A,B,C,D,RA,XT,YT,ZT,LA,MI,NI,R)
      DO 11 I=1,3
   11 R(I)=RA(I)
      CALL HELIO (LA,MI,NI,D,XT,YT,ZT,XA,YA,ZA)
      CALL QUI (R,XA,YA,ZA,QU)
   20 LE(1)=(R(1)+R(2)-SQRT(2)*QU(1))/(2*SQRT(2)*QU(1))
      LE(2)=(R(2)+R(3)-SQRT(2)*QU(2))/(2*SQRT(2)*QU(2))
      LE(3)=(R(1)+R(3)-SQRT(2)*QU(3))/(2*SQRT(2)*QU(3))
      DO 13 I=1,3
      MA(I)=(K*TE(I)**2)/(2*SQRT(2)*QU(I)**3)
      RO(I)=(MA(I)/GA(I)**2)-LE(I)
   13 QS(I)=((2./35.)*RO(I)**2)+((52./1575.)*RO(I)**3)
      CALL TEMPO (TT,TE,D)
      H(1)=(K*TE(1)**2)/(QU(1)**2*(R(1)+R(2)+2*(SQRT(2)/3)*QU(1)*(1+3*QS
     *(1))))
      H(2)=(K*TE(2)**2)/(QU(2)**2*(R(2)+R(3)+2*(SQRT(2)/3)*QU(2)*(1+3*QS
     *(2))))
      H(3)=(K*TE(3)**2)/(QU(3)**2*(R(1)+R(3)+2*(SQRT(2)/3)*QU(3)*(1+3*QS
     *(3))))
      DO 15 I=1,3
      HH=H(I)
      CALL CUBIC(HH,GANEXT)
   15 GA(I)=GANEXT
      SI=(TE(1)/TE(3))*(GA(3)/GA(1))
      SII=(TE(2)/TE(3))*(GA(3)/GA(2))
      CALL RAIOS (SI,SII,A,B,C,D,RA,XT,YT,ZT,LA,MI,NI,R)
C     COMECA O CALCULO DOS ELEMENTOS ORBITAIS
      CALL COROIF(FD,TE,K,RA,XT,YT,ZT,LA,MI,NI,D,SI,SII,DS)
C     PRIMEIRO: CALCULO DAS COORDENADAS ECLITICAS HELIOCENTRICAS
   30 CALL HELIO (LA,MI,NI,D,XT,YT,ZT,XA,YA,ZA)
      DO 21 I=1,3
      XH(I)=XA(I)
      YH(I)=YA(I)*COS(E)+ZA(I)*SIN(E)
   21 ZH(I)=-YA(I)*SIN(E)+ZA(I)*COS(E)
C     SEGUNDO:CALCULO DE  P*, SII: A RAZAO ENTRE AS AREAS 1 E 3
      SIII=DSQRT((XH(1)*YH(3)-XH(3)*YH(1))**2+(YH(1)*ZH(3)-YH(3)*ZH(1))*
     **2+(XH(1)*ZH(3)-XH(3)*ZH(1))**2)/2
      P=(4*SIII**2*GA(3)**2)/(K*TE(3)**2)
C     TERCEIRO: CALCULO DA INCLINACAO (I) E DA LONGITUDE DO NODO, OMEGA
      CI=(XH(1)*YH(3)-XH(3)*YH(1))/(2*SIII)
      SE=DSQRT(1-CI**2)
      CALL ANGULO (CI,SE,II)
      IIG=(II*180)/3.1415926535
      CO=(XH(1)*ZH(3)-XH(3)*ZH(1))/(2*SIII*SE)
      SO=(YH(1)*ZH(3)-YH(3)*ZH(1))/(2*SIII*SE)
      CI=CO
      SE=SO
```

```
      CALL ANGULO (CI,SE,II)
      OMEGA=II
      OMEGAG=(OMEGA*180)/3.1415926535
C     QUARTO: CALCULO DAS ANOMALIAS VERDADEIRAS V(1), V(3)
      CALL ANOVER (RA,SIII,P,V,QU,D1,D3)
      VG(1)=(V(1)*180)/3.1415926535
      VG(3)=(V(3)*180)/3.1415926535
C     QUINTO: CALCULO DA EXCENTRICIDADE , EX
      EX1=D1*(1/DCOS(V(1)))
      EX2=D3*(1/DCOS(V(3)))
      EX=(EX1+EX2)/2
C     SEXTO: CALCULO DO SEMI EIXO MAIOR, AE, E DO MOVIMENTO MEDIO, NM
      AE=P/(1-EX**2)
      NM=(DSQRT(K))/(DSQRT(DABS(AE**3)))
      NMG=NM*180/3.1415926535
C     SETIMO: CALCULO DO ARGUMENTO DO PERICENTRO, WP
      II=(IIG*3.1415926535)/180
      SA=ZH(1)/(RA(1)*DSIN(II))
      CA=((XH(1)/RA(1))*DCOS(OMEGA))+((YH(1)/RA(1))*DSIN(OMEGA))
      CI=CA
      SE=SA
      CALL ANGULO (CI,SE,II)
      W1=II
      WP=W1-V(1)
      IF(WP.GT.0) GO TO 40
      WP=WP+2*3.1415926535
   40 WPG=(WP*180)/3.1415926535
C     OITAVO: CALCULO DA ANOMALIA MEDIA AM
      II=(IIG*3.1415926535)/180
      CALL ANOMED (V,EX,NM,TT,AM,U,TZ)
      AMG=(AM*180)/3.1415926535
C     VERIFICACAO
      DO  120 I=1,3,2
  120 RV(I)=AE*(1-EX*DCOS(U(I)))
      CALL EQUAT1(OMEGA,WP,II,AE,E,EX,AA,AC,BA,BC,AB,AX,AY,AZ,BX,BY,BZ)
      CALL EQUAT2(AX,AY,AZ,BX,BY,BZ,U,TT,TZ,NM,AM,EX,XF,YF,ZF,MF,UF)
      DO 131 I=1,3
      MFG(I)=(MF(I)*180)/3.1415926535
  131 UFG(I)=(UF(I)*180)/3.1415926535
      CALL EQUAT3(XF,YF,ZF,XT,YT,ZT,ALFA,DELTA,QSI,ETA,ZET,AF,DF,TAF,SDF
     *,ERRA,ERRD)
      WRITE(6,9)
      WRITE(6,8) AMG,AE,NMG,EX,WPG,IIG,OMEGAG
      WRITE(6,10)
      WRITE(6,24)
      DO 500 I=1,3
  500 WRITE(6,14) I,XF(I),YF(I),ZF(I),QSI(I),ETA(I),ZET(I)
      WRITE(6,12)
      DO 600 I=1,3
  600 WRITE(6,23) I,AF(I),ERRA(I),DF(I),ERRD(I)
      IF(VVX.EQ.0) GO TO 48
      CONTINUE
      DO 46 I=1,3
      ERRDA(I)=ALFAO(I)-AF(I)
   46 ERRDD(I)=DELTAO(I)-DF(I)
      WRITE(6,49)
      DO 700 I=1,3
  700 WRITE(6,59) I,ERRDA(I),ERRDD(I)
      GO TO 800
   48 IF(DABS(ERRA(1)).LE.1E-6) GO TO 32
      GO TO 31
   32 IF(DABS(ERRA(3)).LE.1E-6) GO TO 33
      GO TO 31
   33 IF(DABS(ERRD(1)).LE.1E-4) GO TO 34
      GO TO 31
   34 IF(DABS(ERRD(3)).LE.1E-6) GO TO 35
      GO TO 31
```


NUCLEO DE COMPUTAÇÃO ELETRONICA

```
C  INICIO DO CALCULO DOS ELEMENTOS ORBITAIS PERTURBADOS PARA DATAS
C  PROXIMAS DAS DATAS DE OBSERVACAO.  ESSES ELEMENTOS SERAO USADOS
C  COMO STARTER NA INTEGRACAO NUMERICA DAS EQUACOES PLANETARIAS
   35 CONTINUE
      DO 100 I=1,3
      DIX=TT(I)
      DIO=TA(I)
      DII=TD(I)
      FMI=ALPMI(I)
      FZE=ALZEP(I)
      FI=ALPI(I)
      FII=ALPII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
      ALEAP(I)=FXX
      FMI=DEPMI(I)
      FZE=DLPZE(I)
      FI=DEPI(I)
      FII=DEPII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
      DELTAP(I)=FXX
      FMI=ROPMI(I)
      FZE=ROPZE(I)
      FI=ROPI(I)
      FII=ROPII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
      ROP(I)=FXX
  100 CONTINUE
      CALLTRANSF(ALEAP,DELTAP,ROP,XP,YP,ZP)
      CALL PERTUR(ZP,YP,XP,XS,YS,ZS,K,MP,FX,FY,FZ,XF,YF,ZF)
      CALL VELOCI(RA,K,AE,UF,EX,MP,OMEGA,II,VX,VY,VZ)
      DO 141 I=1,3
      YFA(I)=YF(I)
      ZFA(I)=ZF(I)
  141 XFA(I)=XF(I)
      DO 143 I=1,3
  143 TJ(I)=TT(I)
      DO 140 J=1,3
      TI=TJ(J)
      XXXF=XFA(J)
      YYYF=YFA(J)
      ZZZF=ZFA(J)
      VVVX=VX(J)
      VVVY=VY(J)
      VVVZ=VZ(J)
      RRRA=RA(J)
      FFFX=FX(J)
      FFFY=FY(J)
      FFFZ=FZ(J)
      XXF=SNGL(XXXF)
      YYF=SNGL(YYYF)
      ZZF=SNGL(ZZZF)
      VVX=SNGL(VVVX)
      VVY=SNGL(VVVY)
      VVZ=SNGL(VVVZ)
      RRA=SNGL(RRRA)
      FFX=SNGL(FFFX)
      FFY=SNGL(FFFY)
      FFZ=SNGL(FFFZ)
      KK=SNGL(K)
      CALL INTEGR(TI,XXF,YYF,ZZF,VVX,VVY,VVZ,FFX,FFY,FFZ,KK,T,Y,RRA)
      XA(3)=Y(1)
      YA(3)=Y(2)
      ZA(3)=Y(3)
      VAX(3)=Y(4)
      VAY(3)=Y(5)
      VAZ(3)=Y(6)
```

```
      CALL INTEGN(TI,XXF,YYF,ZZF,VVX,VVY,VVZ,FFX,FFY,FFZ,KK,I,Y,RRA)
      XA(1)=Y(1)
      YA(1)=Y(2)
      ZA(1)=Y(3)
      VAX(1)=Y(4)
      VAY(1)=Y(5)
      VAZ(1)=Y(6)
      XA(2)=XFA(J)
      YA(2)=YFA(J)
      ZA(2)=ZFA(J)
      VAX(2)=VX(J)
      VAY(2)=VY(J)
      VAZ(2)=VZ(J)
      TE(1)=1
      TE(2)=1
      TE(3)=2
      TT(1)=TJ(J)-1
      TT(2)=TJ(J)
      TT(3)=TJ(J)+1
      DO 150 I=1,3
      RR(I)=XA(I)*XA(I)+YA(I)*YA(I)+ZA(I)*ZA(I)
  150 R(I)=DSQRT(RR(I))
      CALL QUI (R,XA,YA,ZA,QU)
      DO 161 I=1,3
  161 RA(I)=R(I)
      M(3)=(K*TE(3)**2)/(QU(3)**2*(R(3)+R(1)+(2*SQRT(2)/3)*QU(3)))
      MM=M(3)
      CALL CUBIC(MM,GANEXT)
      GA(3)=GANEXT
      LE(3)=(R(1)+R(3)-SQRT(2)*QU(3))/(2*SQRT(2)*QU(3))
      MA(3)=(K*TE(3)**2)/(2*SQRT(2)*QU(3)**3)
      RO(3)=(MA(3)/GA(3)**2)-LE(3)
      QS(3)=((2./35.)*RO(3)**2)+((52./1575.)*RO(3)**3)
      M(3)=(K*TE(3)**2)/(QU(3)**2*(R(1)+R(3)+2*(SQRT(2)/3)*QU(3)*(1+3*QS
     *(3))))
      MM=M(3)
      CALL CUBIC (MM,GANEXT)
      GA(3)=GANEXT
C     COMECA O CALCULO DOS ELEMENTOS ORBITAIS PERTURBADOS
C     PRIMEIRO: CALCULO DAS COORDENADAS ECLITICAS HELIOCENTRICAS
      DO 162 I=1,3
      XH(I)=XA(I)
      YH(I)=YA(I)*COS(E)+ZA(I)*SIN(E)
  162 ZH(I)=-YA(I)*SIN(E)+ZA(I)*COS(E)
C     SEGUNDO:CALCULO DE  P*, SII: A RAZAO ENTRE AS AREAS 1 E 3
      SIII=DSQRT((XH(1)*YH(3)-XH(3)*YH(1))**2+(YH(1)*ZH(3)-YH(3)*ZH(1))*
     **2+(XH(1)*ZH(3)-XH(3)*ZH(1))**2)/2
      P=(4*SIII**2*GA(3)**2)/(K*TE(3)**2)
C     TERCEIRO: CALCULO DA INCLINACAO (I) E DA LONGITUDE DO NODO, OMEGA
      CI=(XH(1)*YH(3)-XH(3)*YH(1))/(2*SIII)
      SE=DSQRT(1-CI**2)
      CALL ANGULO (CI,SE,II)
      IIG=(II*180)/3.1415926535
      CO=(XH(1)*ZH(3)-XH(3)*ZH(1))/(2*SIII*SE)
      SO=(YH(1)*ZH(3)-YH(3)*ZH(1))/(2*SIII*SE)
      CI=CO
      SE=SO
      CALL ANGULO (CI,SE,II)
      OMEGA=II
      OMEGAG=(OMEGA*180)/3.1415926535
C     QUARTO: CALCULO DAS ANOMALIAS VERDADEIRAS V(1), V(3)
      CALL ANOVER (RA,SIII,P,V,QU,D1,D3)
      VG(3)=(V(3)*180)/3.1415926535
      VG(1)=(V(1)*180)/3.1415926535
C     QUINTO: CALCULO DA EXCENTRICIDADE , EX
      EX1=D1*(1/DCOS(V(1)))
      EX2=D3*(1/DCOS(V(3)))
      EX=(EX1+EX2)/2
```

```
C       SEXTO: CALCULO DO SEMI EIXO MAIOR, AE, E DO MOVIMENTO MEDIO, NM
        AE=P/(1-EX**2)
        NM=(DSQRT(K))/(DSQRT(AE**3))
        NMG=NM*180/3.1415926535
C       SETIMO: CALCULO DO ARGUMENTO DO PERICENTRO, WP
        II=(IIG*3.1415926535)/180
        SA=ZH(1)/(RA(1)*DSIN(II))
        CA=((XH(1)/RA(1))*DCOS(OMEGA))+((YH(1)/RA(1))*DSIN(OMEGA))
        CI=CA
        SE=SA
        CALL ANGULO (CI,SE,II)
        W1=II
        WP=W1-V(1)
        IF(WP.GT.0) GO TO 163
        WP=WP+2*3.1415926535
  163   WPG=(WP*180)/3.1415926535
C       OITAVO: CALCULO DA ANOMALIA MEDIA AM
        II=(IIG*3.1415926535)/180
        CALL ANOMED (V,EX,WM,TT,AM,U,TZ)
        AMG=(AM*180)/3.1415926535
        CALL EQUAT1(OMEGA,WP,II,AE,E,EX,AA,AC,BA,BC,AB,AX,AY,AZ,BX,BY,BZ)
        CALL EQUAT2(AX,AY,AZ,BX,BY,BZ,U,TT,TZ,WM,AM,EX,XF,YF,ZF,MF,UF)
        PALFA=ALFA(J)
        PDELTA=DELTA(J)
        PXT=XT(J)
        PYT=YT(J)
        PZT=ZT(J)
        CALL EQUAT4(XF,YF,ZF,PXT,PYT,PZT,PALFA,PDELTA,WSI,ETA,ZET,AF,DF,ER
       *RA,ERRD)
        WRITE(6,45)
        WRITE(6,8) AMG,AE,NMG,EX,WPG,IIG,OMEGAG
        WRITE(6,41)
        WRITE( 6,24)
        DO 42 I=1,3
   42   WRITE(6,14) I,XF(I),YF(I),ZF(I),WSI(I),ETA(I),ZET(I)
        WRITE(6,12)
        I=2
        WRITE(6,23) I,AF(2),ERRA(2),DF(2),ERRD(2)
        ALFA(J)=(AF(2)*3.1415926535*15)/180
        DELTA(J)=DF(2)*3.1415926535*1/180
  140   CONTINUE
        DO 144 I=1,3
  144   TT(I)=TJ(I)
        GO TO 36
    1   FORMAT (4F18.4)
    2   FORMAT (4F18.4)
    9   FORMAT(25X,'TABELA DOS ELEMENTOS ORBITAIS')
    8   FORMAT(//,2X,'ANOMALIA MEDIA=',F10.6,//,2X,'SEMI EIXO MAIOR=',F10.6,/
       */,2X,'MOVIMENTO MEDIO=',F10.6,//,2X,'EXCENTRICIDADE=',F10.6,//,2X,'A
       *RGUMENTO DO PERICENTRO=',F11.6,//,2X,'INCLINACAO=',F10.6,//,2X,'NODO
       * ASCENDENTE=',F10.6)
   10   FORMAT(///,35X,'EFEMERIDES',//)
   24   FORMAT(9X,'XF',13X,'YF',13X,'ZF',13X,'WSI',12X,'ETA',12X,'ZET')
   14   FORMAT(3(2X,I1,6F15.10,/))
   12   FORMAT(//,35X,'CORRECOES',//7X,'ALFA',12X,'ERRA',9X,'DELTAF',11X,'
       *ERRD')
   23   FORMAT(3(2X,I1,4F15.10,/))
   25   FORMAT(4F18.4)
   45   FORMAT(25X,'TABELA DOS ELEMENTOS ORBITAIS PERTURBADOS')
   41   FORMAT(///,35X,'EFEMERIDES PERTURBADAS',//)
   49   FORMAT(//,35X,'CORRECOES FINAIS',//,7X,'ERRA',9X,'ERRD')
   59   FORMAT(3(2X,I1,2F15.10,/))
   70   FORMAT(4F18.4)
  800   STOP
        END
```

```
      SUBROUTINE SOL(XSA,XSZ,XSI,XSII,YSA,YSZ,YSI,YSII,ZSA,ZSZ,ZSI,ZSII,
     *XS,YS,ZS,TT,TA,TD)
      DOUBLE PRECISION FXX
      DIMENSION XSA(3),XSZ(3),XSI(3),XSII(3),YSA(3),YSZ(3),YSI(3),YSII(3
     *),ZSA(3),ZSZ(3),ZSI(3),ZSII(3),XS(3),YS(3),ZS(3),TT(3),TA(3),TD(3)
      DO 1 I=1,3
      DIO=TA(I)
      DIX=TT(I)
      DII=TD(I)
      FMI=XSA(I)
      FZE=XSZ(I)
      FI=XSI(I)
      FII=XSII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
      XS(I)=FXX
      FMI=YSA(I)
      FZE=YSZ(I)
      FI=YSI(I)
      FII=YSII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
      YS(I)=FXX
      FMI=ZSA(I)
      FZE=ZSZ(I)
      FI=ZSI(I)
      FII=ZSII(I)
      CALL BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
    1 ZS(I)=FXX
      RETURN
      END
```

```
      SUBROUTINE TOPOC(XS,YS,ZS,TS,LAT,ORO,XT,YT,ZT)
C     TRANFORMA COORDENADAS GEOCENTRICAS DO SOL
C     EM COORDENADAS TOPOCENTRICAS
      DIMENSION XS(3),YS(3),ZS(3),TS(3),DX(3),DY(3),DZ(3),XT(3),YT(3),ZT
     *(3)
      DOUBLE PRECISION XT,YT,ZT,DX,DY,DZ
      W=.0000427
      DO 30 I=1,3
      DX(I)=-ORO*COS(LAT)*COS(TS(I))*W
      DY(I)=-ORO*COS(LAT)*SIN(TS(I))*W
      DZ(I)=-ORO*SIN(LAT)*W
      XT(I)=XS(I)+DX(I)
      YT(I)=YS(I)+DY(I)
   30 ZT(I)=ZS(I)+DZ(I)
      RETURN
      END
```


NÚCLEO DE COMPUTAÇÃO ELETRÔNICA UFRJ

```
      SUBROUTINE CODIR (ALFA,DELTA,LA,MI,NI,A,B,C,ED)
C     DETERMINA OS COS DIRETORES E AS  CONSTANTES DO SISTEMA
C     QUE DEFINE LMN
       DOUBLE PRECISION LA,MI,NI,A,B,C,F,G,Q,ED
      DIMENSION ALFA(3),DELTA(3),LA(3),MI(3),NI(3),A(3),B(3),C(3)
      DO 40 I=1,3
      LA(I)=COS(DELTA(I))*COS(ALFA(I))
      MI(I)=COS(DELTA(I))*SIN(ALFA(I))
   40 NI(I)=SIN(DELTA(I))
      F=MI(2)*NI(3)-MI(3)*NI(2)
      G=-LA(2)*NI(3)+LA(3)*NI(2)
      Q=LA(2)*MI(3)-LA(3)*MI(2)
      D=LA(1)*F+MI(1)*G+NI(1)*Q
      A(1)=F/D
      B(1)=G/D
      C(1)=Q/D
      A(2)=(NI(1)*NI(3)-MI(3)*NI(1))/D
      B(2)=(NI(1)*LA(3)-NI(3)*LA(1))/D
      C(2)=(LA(1)*MI(3)-LA(3)*MI(1))/D
      A(3)=(MI(1)*NI(2)-NI(2)*NI(1))/D
      B(3)=(NI(1)*LA(2)-NI(2)*LA(1))/D
      C(3)=(LA(1)*MI(2)-LA(2)*MI(1))/D
      ED=D
      RETURN
      END
```

```
      SUBROUTINE RAIO (A,B,C,XT,YT,ZT,LA,MI,NI,TE,K,R,D)
      DOUBLE PRECISION LA,MI,NI,A,B,C,XT,YT,ZT,R,D,RA,A0,B0,A1,B1,A2,B2,
     *A3,B3,A4,B4,A5,B5,A6,B6,K
      DIMENSION A(3),B(3),C(3),XT(3),YT(3),ZT(3),LA(3),MI(3),NI(3),TE(3)
     *,R(3),RA(3),D(3)
      A0=TE(1)/TE(3)
      B0=(TE(1)/TE(3))*(K/6)*(TE(3)**2-TE(1)**2)
      A1=TE(2)/TE(3)
      B1=(TE(2)/TE(3))*(K/6)*(TE(3)**2-TE(2)**2)
      A2=A1*XT(1)-XT(2)+A0*XT(3)
      B2=B1*XT(1)+B0*XT(3)
      A3=A1*YT(1)-YT(2)+A0*YT(3)
      B3=B1*YT(1)+B0*YT(3)
      A4=A1*ZT(1)-ZT(2)+A0*ZT(3)
      B4=B1*ZT(1)+B0*ZT(3)
      A5=A(2)*A2+B(2)*A3+C(2)*A4
      B5=A(2)*B2+B(2)*B3+C(2)*B4
      A6=LA(2)*XT(2)+MI(2)*YT(2)+NI(2)*ZT(2)
      B6=XT(2)**2+YT(2)**2+ZT(2)**2
   32 D(2)=A5+(B5/(R(2)**3))
      RA(2)=DSQRT(D(2)**2-2*D(2)*A6+B6)
      IF(RA(2)-R(2),LT,1E-6) GO TO 31
      R(2)=RA(2)
      GO TO 32
   31 R(2)=RA(2)
      RETURN
      END
```

```
      SUBROUTINE TEMPO (TT,TE,D)
C     CALCULA OS TEMPOS E INTERVALOS CORRIGIDOS DA ABERRACAO
      DOUBLE PRECISION D
      DIMENSION TT(3),TE(3),D(3),TA(3)
      DO 50 I=1,3
      TA(I)=.00577*D(I)
   50 TT(I)=TT(I)-TA(I)
      TE(1)=TT(2)-TT(1)
      TE(2)=TT(3)-TT(2)
      TE(3)=TE(1)+TE(2)
      RETURN
      END
```

```
      SUBROUTINE RAIOS (SI,SII,A,B,C,D,RA,XT,YT,ZT,LA,MI,NI,R)
C  CALCULA OS DELTAS E OS RAIOS VETORES
      DOUBLE PRECISION A,B,C,D,R,XT,YT,ZT,LA,MI,NI,RA,RO,L,M,N,SI,SII
      DIMENSION A(3),B(3),C(3),D(3),R(3),XT(3),YT(3),ZT(3),LA(3),MI(3),N
     *I(3),RA(3),RO(3)
      L=SII*XT(1)-XT(2)+SI*XT(3)
      M=SII*YT(1)-YT(2)+SI*YT(3)
      N=SII*ZT(1)-ZT(2)+SI*ZT(3)
      D(1)=(1/SII)*(A(1)*L+B(1)*M+C(1)*N)
      D(2)=A(2)*L+B(2)*M+C(2)*N
      D(3)=(1/SI)*(A(3)*L+B(3)*M+C(3)*N)
      DO 60 I=1,3
      RO(I)=D(I)**2-2*D(I)*(LA(I)*XT(I)+MI(I)*YT(I)+NI(I)*ZT(I))+XT(I)**
     *2+YT(I)**2+ZT(I)**2
   60 RA(I)=DSQRT(RO(I))
      RETURN
      END
```

```
      SUBROUTINE HELIO (LA,MI,NI,D,XT,YT,ZT,XA,YA,ZA)
C     CALCULA AS COORDENADAS HELIOCENTRICAS DO ASTRO
      DOUBLE PRECISION LA,MI,NI,D,XT,YT,ZT,XA,YA,ZA
      DIMENSION LA(3),MI(3),NI(3),D(3),XT(3),YT(3),ZT(3),XA(3),
     *YA(3),ZA(3)
      DO 70 I=1,3
      XA(I)=LA(I)*D(I)-XT(I)
      YA(I)=MI(I)*D(I)-YT(I)
   70 ZA(I)=NI(I)*D(I)-ZT(I)
      RETURN
      END
```

```
      SUBROUTINE QUI (R,XA,YA,ZA,QU)
C     CALCULA AS QUANT AUXILIARES QU(1),QU(2),QU(3)
      DOUBLE PRECISION R,XA,YA,ZA,QU,QU1,QU2,QU3
      DIMENSION R(3),XA(3),YA(3),ZA(3),QU(3)
      QU1=R(1)*R(2)+XA(1)*XA(2)+YA(1)*YA(2)+ZA(1)*ZA(2)
      QU(1)=DSQRT(QU1)
      QU2=R(2)*R(3)+XA(2)*XA(3)+YA(2)*YA(3)+ZA(2)*ZA(3)
      QU(2)=DSQRT(QU2)
      QU3=R(1)*R(3)+XA(1)*XA(3)+YA(1)*YA(3)+ZA(1)*ZA(3)
      QU(3)=DSQRT(QU3)
      RETURN
      END
```

```
      SUBROUTINE CUBIC(HH,GANEXT)
      DOUBLE PRECISION HH,GAA,GANEXT,EPSLO
      POL(GAA)=GAA**3-GAA**2-HH*GAA-HH/9
      PRIME(GAA)=3*GAA**2-2*GAA-HH
      GANEXT=1
      EPSLO=1E-8
      LIMIT=20
      N=0
   30 N=N+1
      GAA=GANEXT
      IF(PRIME(GAA)) 10,1,10
    1 CONTINUE
      CALL EXIT
   10 GANEXT=GAA-POL(GAA)/PRIME(GAA)
      IF(DABS(GAA-GANEXT)-EPSLO) 2,2,20
    2 CONTINUE
      GO TO 5
   20 IF(N-LIMIT) 30,30,3
    3 GO TO 5
    5 RETURN
      END
```

```
      SUBROUTINE ANGULO (CI,SE,II)
C     CALCULA EM QUE QUADRANTE ESTA O ANGULO DESEJADO
      DOUBLE PRECISION CI,SE,II,SEI
      IF(CI.EQ.0) STOP
      SEI=SE/CI
      IF(SE+0) 80,81,82
   80 IF(SEI+0) 83,83,84
   83 II=DATAN(SEI)+6.2831853071
      GO TO 91
   84 II=3.1415926535+DATAN(SEI)
      GO TO 91
   81 II=0
      GO TO 91
   82 IF(SEI+0) 85,85,86
   85 II=DATAN(SEI)+3.1415926535
      GO TO 91
   86 II=DATAN(SEI)
   91 RETURN
      END
```

```
      SUBROUTINE ANOVER(RA,SIII,P,V,QU,D1,D3)
      DIMENSION V(3),RA(3), QU(3)
      DOUBLE PRECISION RA,SIII,P,V,D1,D3,TV2,V2,QU
      SV1=(2*SIII)/(RA(1)*RA(3))
      V1=ARSIN(SV1)
      D1=(P/RA(1))-1
      D3=(P/RA(3))-1
      TV2=((D1-D3)/(D1+D3))*(QU(3)**2)/(2*SIII)
      IF(TV2)1,1,2
    1 IF(D1-D3)3,4,5
    3 V2=6.2831853071+DATAN(TV2)
      GO TO 6
    4 V2=0
      GO TO 6
    5 V2=3.1415926535+DATAN(TV2)
      GO TO 6
    2 IF(D1-D3)5,4,9
    9 V2=DATAN(TV2)
    6 V(3)=V2+(V1/2)
      V(1)=V2-(V1/2)
      RETURN
      END
```

```
      SUBROUTINE ANOMED (V,EX,NM,TT,AM,U,TZ)
      DIMENSION V(3),U(3),M(3),C(3),MZ(3),TT(3)
      DOUBLE PRECISION V,EX,NM,AM,U,TU,UI,M,C,MZ
      DO 27 I=1,3,2
      TU=DSQRT((1-EX)/(1+EX))*DTAN(V(I)/2)
      UI=2*DATAN(TU)
      IF(DTAN(UI)) 1,1,2
    1 IF(V(I).LT.3.1415926535) GO TO 3
      U(I)=DATAN(DTAN(UI))+2*3.1415926535
      GO TO 5
    3 U(I)=DATAN(DTAN(UI))+3.1415926535
      GO TO 5
    2 IF(V(I).GT.3.1415926535) GO TO 3
      U(I)=DATAN(DTAN(UI))
    5 M(I)=U(I)-EX*DSIN(U(I))
      TZ=HFIX(TT(2))+1
      C(I)=-NM*(TT(I)-TZ)
   27 MZ(I)=M(I)+C(I)
      AM=(MZ(1)+MZ(3))/2
      RETURN
      END
```

```
      SUBROUTINE EQUAT1(OMEGA,WP,II,AE,E,EX,AA,AC,BA,BC,AB,AX,AY,AZ,BX,B
     *Y,BZ)
      DOUBLE PRECISION AX,AY,AZ,BX1,BX,BY,BZ,AA,AC,BA,BC,AB,OMEGA,WP,II,
     *EX,AE
      AX=(DCOS(OMEGA)*DCOS(WP)-DCOS(II)*DSIN(OMEGA)*DSIN(WP))*AE
      AY=((DSIN(OMEGA)*DCOS(WP)+DCOS(II)*DCOS(OMEGA)*DSIN(WP))*COS(E)-DS
     *IN(II)*DSIN(WP)*SIN(E))*AE
      AZ=((DSIN(OMEGA)*DCOS(WP)+DCOS(II)*DCOS(OMEGA)*DSIN(WP))*SIN(E)+DS
     *IN(II)*DSIN(WP)*COS(E))*AE
      BX1=AE*DSQRT(1-EX**2)
      BX=(-DCOS(OMEGA)*DSIN(WP)-DCOS(II)*DSIN(OMEGA)*DCOS(WP))*BX1
      BY=((-DSIN(OMEGA)*DSIN(WP)+DCOS(II)*DCOS(OMEGA)*DCOS(WP))*COS(E)-D
     *SIN(II)*DCOS(WP)*SIN(E))*BX1
      BZ=((-DSIN(OMEGA)*DSIN(WP)+DCOS(II)*DCOS(OMEGA)*DCOS(WP))*SIN(E)+D
     *SIN(II)*DCOS(WP)*COS(E))*BX1
      AA=AX**2+AY**2+AZ**2
      AC=AE**2
      BA=BX**2+BY**2+BZ**2+0.0
      BC=AE**2*(1-EX**2)
      AB=AX*BX+ AY*BY+AZ*BZ
      RETURN
      END
```

```
      SUBROUTINE EQUAT2(AX,AY,AZ,BX,BY,BZ,U,TT,TZ,NM,AM,EX,XF,YF,ZF,MF,U
     *F)
      IMPLICIT DOUBLE PRECISION (A-Z)
      REAL TT,TZ
      INTEGER I
      DIMENSION U(3),C(3),A(3),B(3),TT(3),XT(3),YT(3),ZT(3),XF(3),YF(3),
     *ZF(3),MF(3),UF(3)
      C(1)=-NM*(TT(1)-TZ)
      C(2)=-NM*(TT(2)-TZ)
      C(3)=-NM*(TT(3)-TZ)
      DO 1 I=1,3
    1 MF(I)=AM-C(I)
      MFF=MF(1)
      CALL UFIN(EX,MFF,UFF)
      UF(1)=UFF
      MFF=MF(2)
      CALL UFIN(EX,MFF,UFF)
      UF(2)=UFF
      MFF=MF(3)
      CALL UFIN(EX,MFF,UFF)
      UF(3)=UFF
      DO 2 I=1,3
      A(I)=DCOS(UF(I))-EX
      B(I)=DSIN(UF(I))
      XF(I)=AX*A(I)+BX*B(I)
      YF(I)=AY*A(I)+BY*B(I)
    2 ZF(I)=AZ*A(I)+BZ*B(I)
      RETURN
      END
```

```
      SUBROUTINE EQUAT3(XF,YF,ZF,XT,YT,ZT,ALFA,DELTA,QSI,ETA,ZET,AF,DF,T
     *AF,SDF,ERRA,ERRD)
      IMPLICIT DOUBLE PRECISION (A-Z)
      INTEGER I
      REAL ALFA,DELTA,ALFAF,DELTAF,ALFAD,DELTAD
      DIMENSION XF(3),YF(3),ZF(3),QSI(3),ETA(3),ZET(3),ALFA(3),DELTA(3),
     *TAF(3),SDF(3),AF(3),DF(3),AC(3),ERRA(3),ERRD(3),XT(3),YT(3),ZT(3),
     *CDF(3),TDF(3),ALFAF(3),DELTAF(3),ALFAD(3),DELTAD(3)
      DO 1 I=1,3
      QSI(I)=XF(I)+XT(I)
      ETA(I)=YF(I)+YT(I)
      ZET(I)=ZF(I)+ZT(I)
      TAF(I)=(ETA(I)/QSI(I))
      SDF(I)=(ZET(I)/(DSQRT(QSI(I)**2+ETA(I)**2+ZET(I)**2)))
      AC(I)=DATAN(TAF(I))
      IF(ALFA(I).LT.3.1415926535/2) GO TO 2
      IF(ALFA(I).LT.3*3.1415926535/2) GO TO 3
      AF(I)=AC(I)+2*3.1415926535
      GO TO 4
    2 AF(I)=AC(I)
      GO TO 4
    3 AF(I)=AC(I)+3.1415926535
      GO TO 4
    4 AF(I)=(AF(I)*180)/(3.1415926535*15)
      CDF(I)=DSQRT(1-SDF(I)**2)
      TDF(I)=SDF(I)/CDF(I)
      DF(I)=DATAN(TDF(I))
      DF(I)=(DF(I)*180)/3.1415926535
      ALFAF(I)=ALFA(I)*180/(3.1415926535*15)
      ERRA(I)=ALFAF(I)-AF(I)
      DELTAF(I)=DELTA(I)*180/3.1415926535
    1 ERRD(I)=DELTAF(I)-DF(I)
      RETURN
      END
```

```
      SUBROUTINE UFIN(EX,MFF,UFF)
      IMPLICIT DOUBLE PRECISION (A-Z)
      U=0
    3 ARG=MFF+U
      UF=EX*DSIN(ARG)
      IF(DABS(DABS(UF)-DABS(U)).LE.1E-6) GO TO 2
      U=UF
      GO TO 3
    2 UFF=MFF+UF
      RETURN
      END
```

```
      SUBROUTINE CORDIF(ED,TE,K,RA,XT,YT,ZT,LA,MI,NI,D,SI,SII,DS)
C     CALCULA AS CORRECOES DIFERENCIAIS DAS RAZOES DAS AREAS - BASEADO NAS
C     CORRECOES DIFERENCIAIS DE LEUSCHNER.
      IMPLICIT DOUBLE PRECISION (A-Z)
      REAL TE
      INTEGER I
      DIMENSION XT(3),YT(3),ZT(3),D(3),TE(3),RA(3),LA(3),MI(3),NI(3),A(3
     *),SC(3),SP(3),F(3),FA(3),FB(3),FC(3),PP(3),B(3),DSL(3),DS(3)
      A(1)=TE(1)/TE(3)
      A(3)=TE(2)/TE(3)
      SP(1)=A(1)*(1+(K/6)*(TE(3)**2-TE(1)**2)*(1/RA(2)**3))
      SP(3)=A(3)*(1+(K/6)*(TE(3)**2-TE(2)**2)*(1/RA(2)**3))
      F3=(K/12)*(-TE(1)**2+TE(1)*TE(2)+TE(2)**2)
      F2=(K/12)*(TE(1)**2+3*TE(1)*TE(2)+TE(2)**2)
      F1=(K/12)*(TE(1)**2+TE(1)*TE(2)-TE(2)**2)
      SC(1)=A(1)*((1+F3/RA(3)**3)/(1-F2/RA(2)**3))
      SC(3)=A(3)*((1+F1/RA(1)**3)/(1-F2/RA(2)**3))
      RCO=XT(2)*LA(2)+YT(2)*MI(2)+ZT(2)*NI(2)
      DO 1 I=1,3
      FA(I)=LA(1)*(YT(I)*NI(3)-ZT(I)*MI(3))
      FB(I)=MI(1)*(XT(I)*NI(3)-ZT(I)*LA(3))
      FC(I)=NI(1)*(XT(I)*MI(3)-YT(I)*LA(3))
    1 F(I)=FA(I)-FB(I)+FC(I)
      P=D(2)-RCO
      PA=3*P/RA(2)**2
      DO 2 I=1,3,2
      PP(I)=SC(I)-A(I)
    2 B(I)=PP(I)*PA
      DO 3 I=1,3,2
    3 DSL(I)=SP(I)-SC(I)
      Q=ED+F(1)*B(1)+F(3)*B(3)
      DD=(-F(1)*DSL(1)-F(3)*DSL(3))/Q
      DO 4 I=1,3,2
    4 DS(I)=-B(I)*DD
      SI=SI+DS(1)
      SII=SII+DS(3)
      RETURN
      END
```

```
      SUBROUTINE BESSEL(DIO,DIX,DII,LOM,FMI,FZE,FI,FII,FXX)
C     ATE AS DIFERENCAS SEGUNDAS
      IMPLICIT DOUBLE PRECISION (A-Z)
      REAL LOM,DIO,DIX,DII,FMI,FZE,FI,FII
C     DIX=DATA PARA A QUAL DESEJA-SE INTERPOLAR
C     NB= FATOR DE INTERPOLACAO
C     DIO=DATA IMEDIATA/ ANTERIOR A DIX; DII=IMEDIATA/ POSTERIOR A DIX
C     FZE,FI DADOS CORRESPONDENTES A DIO E DIx RESPECTIVA/
C     FMI,FII DADOS IMEDIATA/ ANTERIORES E POSTERIORES AFZE E FI
      LOM=LOM*180/(3.1415926535*15*24)
      NB=(1/(DII-DIO))*(DIX-DIO+LOM)
      FD1=FZE-FMI
      FD2=FI-FZE
      FD3=FII-FI
      FDO=FD2-FD1
      FDL=FD3-FD2
      FXX=FZE+NB*FD2+(NB*(NB-1)/4)*(FDO+FDL)
      RETURN
      END
```

```
      SUBROUTINE TRANSF(ALFAP,DELTAP,ROP,XP,YP,ZP)
C     TRANSFORMA COORDENADAS ESFERICAS GEOCENTRICAS DOS PLANETAS
C    PERTURBADORES (ALFA,DELTA,RO) EM COORDENADAS RETANGULARES GEOCENTRICA
C    (XP,YP,ZP)
      DOUBLE PRECISION XP,YP,ZP
      DIMENSION ALFAP(3),DELTAP(3),ROP(3),XP(3),YP(3),ZP(3)
      DO 1 I=1,3
      XP(I)=ROP(I)*COS(DELTAP(I))*COS(ALFAP(I))
      YP(I)=ROP(I)*COS(DELTAP(I))*SIN(ALFAP(I))
    1 ZP(I)=ROP(I)*SIN(DELTAP(I))
      RETURN
      END
```

```
      SUBROUTINE PERTUR(XP,YP,ZP,XS,YS,ZS,K,MP,FX,FY,FZ,XF,YF,ZF)
C   CALCULA AS COMPONENTES DA FORCA DE PERTURBACAO DE CADA PLANETA
      IMPLICIT DOUBLE PRECISION (A-Z)
      INTEGER I
      REAL MP,XS,YS,ZS
      DIMENSION XP(3),YP(3),ZP(3),XS(3),YS(3),ZS(3),FX(3),FY(3),FZ(3),XF
     *(3),YF(3),ZF(3),RP(3),ROP(3),A(3),B(3),C(3),R1(3),R(3),R2(3),QSI(3
     *),ETA(3),ZET(3),X(3),Y(3),Z(3)
      DO 1 I=1,3
      QSI(I)=XF(I)+XS(I)
      ETA(I)=YF(I)+YS(I)
      ZET(I)=ZF(I)+ZS(I)
      X(I)=QSI(I)-XP(I)
      Y(I)=ETA(I)-YP(I)
      Z(I)=ZET(I)-ZP(I)
      RP(I)=X(I)*X(I)+Y(I)*Y(I)+Z(I)*Z(I)
      ROP(I)=RP(I)*DSQRT(RP(I))
      A(I)=XS(I)-XP(I)
      C(I)=ZS(I)-ZP(I)
      B(I)=YS(I)-YP(I)
      R1(I)=A(I)*A(I)+B(I)*B(I)+C(I)*C(I)
      R(I)=R1(I)*DSQRT(R1(I))
      FX(I)=K*MP*(X(I)/ROP(I)-A(I)/R(I))
      FY(I)=K*MP*(Y(I)/ROP(I)-B(I)/R(I))
    1 FZ(I)=K*MP*(Z(I)/ROP(I)-C(I)/R(I))
      WRITE(6,/) FX,FY,FZ
      RETURN
      END
```

```
      SUBROUTINE VELOCI(RA,K,AE,UF,EX,WP,OMEGA,II,VX,VY,VZ)
C  CALCULA AS VELOCIDADES NO SISTEMA EQUATORIAL HELIOCENTRICO
      IMPLICIT DOUBLE PRECISION (A-Z)
      INTEGER I
      DIMENSION RA(3),UF(3),QSL(3),ETL(3),VX(3),VY(3),VZ(3)
      DO 1 I=1,3
      QSL(I)=-(1/RA(I))*DSQRT(K*AE)*DSIN(UF(I))
      A=K*AE*(1-EX*EX)
    1 ETL(I)=(1/RA(I))*DSQRT(A)*DCOS(UF(I))
      B=DCOS(WP)
      C=DSIN(WP)
      D=DCOS(II)
      E=DSIN(II)
      F=DCOS(OMEGA)
      G=DSIN(OMEGA)
      D11=B*F-C*G*D
      D12=B*G+C*F*D
      D13=C*E
      D21=-C*F-B*G*D
      D22=-C*G+B*F*D
      D23=B*E
      DO 2 I=1,3
      VX(I)=D11*QSL(I)+D21*ETL(I)
      VY(I)=D12*QSL(I)+D22*ETL(I)
    2 VZ(I)=D13*QSL(I)+D23*ETL(I)
      RETURN
      END
```

```
      SUBROUTINE INTEGR(TI,XXF,YYF,ZZF,VVX,VVY,VVZ,FFX,FFY,FFZ,KK,T,Y,RR
     *A)
C     PROCESSA A INTEGRAÇAO NUMERICA PARA DATAS PROXIMAS AS DE OBSERVACA
      EXTERNAL DFN
      REAL KK
      DIMENSION Y(6),R(6),S(6),WK(174),DY(6)
      Y(1)=XXF
      Y(2)=YYF
      Y(3)=ZZF
      Y(4)=VVX
      Y(5)=VVY
      Y(6)=VVZ
      S(1)=XXF
      S(2)=YYF
      S(3)=ZZF
      S(4)=VVX
      S(5)=VVY
      S(6)=VVZ
      T=0
      H=.15
      HMIN=0.01
      EPS=0.0000001
      N=6
      JM=0
      IND=2
      JSTART=0
      DO 2 I=1,5
      B=0.2*FLOAT(I)
    1 IF(H.GT.B-T) H=B-T
      CALL DYA(RRA,FFX,FFY,FFZ,KK,Y,DY)
      CALL DREBS (DFN,Y,T,N,JM,IND,JSTART,H,HMIN,EPS,R,S,WK,IER,DY)
      IF(IER.NE.0) STOP
      IF(T.LT.B-HMIN) GO TO 1
    2 WRITE(6,/) T,Y
      RETURN
      END
```

```
      SUBROUTINE INTEGN(TI,XXF,YYF,ZZF,VVX,VVY,VVZ,FFX,FFY,FFZ,KK,T,Y,RR
     *A)
C     PROCESSA A INTEGRACAO NUMERICA PARA DATAS PROXIMAS AS DE OBSERVACA
      EXTERNAL DFN
      REAL KK
      DIMENSION Y(6),R(6),S(6),WK(174),DY(6)
      Y(1)=XXF
      Y(2)=YYF
      Y(3)=ZZF
      Y(4)=VVX
      Y(5)=VVY
      Y(6)=VVZ
      S(1)=XXF
      S(2)=YYF
      S(3)=ZZF
      S(4)=VVX
      S(5)=VVY
      S(6)=VVZ
      T=0
      H=-0.15
      HMIN=-0.01
      EPS=0.0000001
      N=6
      JM=6
      IND=2
      JSTART=0
      DO 2 I=1,5
      B=0.2*FLOAT(I)
    1 IF(ABS(H).GT.(B-ABS(T))) H=-B-T
      CALL DYA(RRA,FFX,FFY,FFZ,KK,Y,DY)
      CALL DREBS (DFN,Y,T,N,JM,IND,JSTART,H,HMIN,EPS,R,S,WK,IER,DY)
      IF(IER.NE.0) STOP
      IF(ABS(T).LT.(B-ABS(HMIN))) GO TO 1
    2 WRITE(6,/) T,Y
      RETURN
      END
```

```
      SUBROUTINE DYA(RRA,FFX,FFY,FFZ,KK,Y,DY)
      REAL KK
      DIMENSION Y(6),DY(6)
      DY(4)=-KK*Y(1)/RRA**3+FFX
      DY(5)=-KK*Y(2)/RRA**3+FFY
      DY(6)=-KK*Y(3)/RRA**3+FFZ
      RETURN
      END
```

```
      SUBROUTINE DFN(Y,T,N,DY)
      REAL KK
      DIMENSION Y(6),DY(6)
      DY(1)=Y(4)
      DY(2)=Y(5)
      DY(3)=Y(6)
      DY(4)=DY(4)
      DY(5)=DY(5)
      DY(6)=DY(6)
       RETURN
       END
```

```
      SUBROUTINE EQUAT4(XF,YF,ZF,PXT,PYT,PZT,PALFA,PDELTA,QSI,ETA,ZET,AF
     *,DF,ERRA,ERRD)
      IMPLICIT DOUBLE PRECISION (A-Z)
      REAL PALFA,PDELTA,PALF,PDEF
      INTEGER I
      DIMENSION QSI(3),ETA(3),ZET(3),XF(3),YF(3),ZF(3),TAF(3),SDF(3),AC(
     *3),AF(3),CDF(3),TDF(3),DF(3),ERRA(3),ERRD(3)
      QSI(2)=XF(2)+PXT
      ETA(2)=YF(2)+PYT
      ZET(2)=ZF(2)+PZT
      TAF(2)=(ETA(2)/QSI(2))
      SDF(2)=(ZET(2)/DSQRT(QSI(2)**2+ETA(2)**2+ZET(2)**2))
      AC(2)=DATAN(TAF(2))
      IF(PALFA.LT.3.1415926535/2) GO TO 2
      IF(PALFA.LT.3*3.1415926535/2) GO TO 3
      AF(2)=AC(2)+2*3.1415926535
      GO TO 4
    2 AF(2)=AC(2)
      GO TO 4
    3 AF(2)=AC(2)+3.1415926535
      GO TO 4
    4 AF(2)=(AF(2)*180)/(3.1415926535*15)
      CDF(2)=DSQRT(1-SDF(2)**2)
      TDF(2)=SDF(2)/CDF(2)
      DF(2)=DATAN(TDF(2))
      DF(2)=(DF(2)*180)/3.1415926535
      PALF=PALFA*180/(3.1415926535*15)
      ERRA(2)=PALF-AF(2)
      PDEF=PDELTA*180/3.1415926535
    1 ERRD(2)=PDEF-DF(2)
      RETURN
      END
```

BIBLIOGRAFIA

1. BROUWER D., CLEMENCE G.M. Methods of Celestial Mechanics Academic Press, New York, 1966.

2. BULIRSCH R., STOER J. "Fehlerabschatzungen und Extrapolation mit rationalen Funktionen bei Verfahren vom Richardson - Typus". Numerische Mathematik. 6, 1964 , pags. 413-427.

3. BULIRSCH R., STOER J. "Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods" . Numerische Mathematik. 8, 1966, pags. 1-13.

4. CHEBOTAREV G.A. "Analitical and Numerical Methods of Celestial Mechanics. American Elvesier, 1967.

5. DANJON A. Astronomie Générale. J. R. Sennac Editeurs, Paris, 1959.

6. FITZPATRICK P.M. Principles of Celestial Mechanics . Academic Press, New York, 1970.

7. GRAGG W.B. "On Extrapolation Algorithms for Ordinary Initial Value Problems". Siam Numer. Anal., Ser.B, Vol.2, No.3, 1965, pags. 384-402.

8. HERRICK S. Astrodynamics - Orbit Determination, Space Navegation, Celestial Mechanics. Vol.1. Van Nostrand Reinhold Company, London, 1971.

9. HERRICK S. Astrodynamics - Orbit Correction, Perturbation Theory, Integration. Vol.2. Van Nostrand Reinhold Company, London, 1972.

10. RYABOV Y. An Elementary Survey of Celestial Mechanics . Dover Publ. Catlons Inc., New York, 1961.

11. STETTER H.J. "Asymptotic Expansions for the Error of Discretization Algorithms for Non - Linear Functional Equations". Numerische Mathematik. 7, 1965, pag. 18-31.