STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

AFDELING TOEGEPASTE WISKUNDE

COLLEGE:

NUMERIEKE METHODEN VOOR

PARTIËLE DIFFERENTIAAL VERGELIJKINGEN

DR. R.P. van de RIET

2e SEMESTER 1968-1969.

INHOUD:

College:

Numerieke methoden voor partiële differentiaal vergelijkingen

Dr. R.P.van de Riet

2e semester 1968-1969.

Inleiding

Het doel van dit college is om processen op te stellen waarmee men m.b.v. een rekenautomaat partiële differentiaalvergelijkingen (p.d.v.'s) met rand-en begin voorwaarden tot (benaderde) oplossing kan brengen.
In het bijzonder zal de aandacht gericht zijn op parabolische en hyperbolische p.d.v.'s; d.w.z. begin-randwaardeproblemen.
Het niet, of nauwelijks, behandelen van zuivere randwaarde problemen (elliptische p.d.v.'s) vindt zijn oorzaak in het feit dat de processen voor deze problemen iteratief van aard zijn, terwijl de processen die wij zullen behandelen meer "recht-toe-recht-aan" van aard zijn.

Zowel bij het vertalen van de differentiaal operator werkend op functies van continue variabelen naar een differentie operator werkend op functies van discrete variabelen,als bij het rekenen op de rekenautomaat worden fouten gemaakt (afbreekfouten,respectievelijk rekenfouten). Beide fouten zijn de oorzaak dat de berekende oplossing niet gelijk is aan de echte analytische oplossing, waarvan we de existentie altijd impliciet aannemen. Bij de vraag hoeveel beide oplossingen dan wel verschillen spelen de begrippen consistentie, stabiliteit en convergentie een belangrijke rol. Deze begrippen en de problemen die zich voordoen zullen aan de hand van een eenvoudige gewone differentiaal vergelijking geadstrueerd worden.
In een volgend hoofdstuk worden deze begrippen op strenge wijze ingevoerd.
Vervolgens worden enkele processen voor de diffusie vergelijking en de golf vergelijking bestudeerd in verband met bovenstaande begrippen.

Literatuur:

Ter voorbereiding van dit college is vooral gebruikt:

Godunov, J.K. en V.S. Rjabenski: Theory of difference schemes - an introduction, Noord Hollandse Uitgevers Maatschappij, Amsterdam (1964).

en

Förch, G.J.R., P.J. van der Houwen en R.P. van de Riet: Colloquium Stabiliteit van differentieschema's, deel 1, M.C. Syllabus 2.1, Mathematisch Centrum, Amsterdam (1967).

In het tweede boekje treft men een uitgebreide literatuurlijst aan. Verder vermelden we nog:

Dekker, L., T.J. Dekker, P.J. van der Houwen en M.N. Spijker: Colloquium Stabiliteit van differentieschema's, deel 2, M.C. Syllabus 2.2, Mathematisch Centrum, Amsterdam (1968).

Van der Houwen, P.J.: Finite difference methods for solving partial differential equations, Mathematical Centre Tracts 20, Mathematisch Centrum, Amsterdam (1968).

1. Een "gewone" differentiaalvergelijking; "gewone" differentieschema's.

In dit hoofdstuk worden differentieschema's voor een gewone differentiaalvergelijking opgesteld. Naar aanleiding hiervan en ter fundering voor de later aan de orde komende "partiële" differentieschema's wordt een summiere theorie van gewone differentieschema's gegeven.

1.0. Een "gewoon" differentieschema voor een "gewone" differentiaalvergelijking.

Opmerking vooraf: In deze paragraaf worden niet de standaard methoden voor een gewone differentiaalvergelijking (Runge-Kutta, Romberg) behandeld.
De inhoud van deze paragraaf dient slechts ter illustratie.

In het vervolg zullen functies van een continue variabele, die optreden in de oorspronkelijke probleemstelling met een hoofdletter aangeduid worden. De functies van een discrete variabele, die in het rekenproces een rol spelen zullen met een kleine letter worden aangeduid.

Beschouw het probleem:

$$\frac{dU}{dt} + U(t) = F(t),\ U(0) = 1 \qquad (1.1)$$

dat, zoals bekend, de oplossing:

$$U(t) = e^{-t}\ (1 + \int_0^t e^{t_1}\ F(t_1)\ dt_1) \qquad (1.2)$$

bezit.
Kies op het t interval $[0,\infty)$ de punten:
$t_i = i\tau$, $i = 0, 1, 2, \ldots$, waarin τ een reëel positief getal is.
Gevraagd te berekenen de <u>roosterfunctie</u> u(i), gedefinieerd op de punten t_i, die voldoet aan:

$$\left.\begin{aligned} &\frac{u(i+1) - u(i)}{\tau} + u(i) = f(i),\ i = 0, 1, 2, \ldots \\ &\text{en } u(0) = 1 \end{aligned}\right\} \qquad (1.3)$$

We duiden de roosterfunctie ook wel aan met u of met u_i.
Het probleem (1.3) heeft veel overeenkomst met probleem (1.1); immers:

$$\frac{U(t_{i+1}) - U(t_i)}{\tau} = U_t(t_i) + \frac{\tau}{2} U_{tt}(t_i + \theta_i\tau) \qquad (1.4)$$

waarin U_t en U_{tt} voorstellen $\frac{dU}{dt}$ en $\frac{d^2U}{dt^2}$ en waarin $0 < \theta_i < 1$,

zodat $U(t_i)$ voldoet aan (1.3) met

$$f(i) = F(t_i) + \frac{\tau}{2} U_{tt}(t_i + \theta_i\tau) \qquad (1.5)$$

Een oplossing van (1.3) is op de volgende wijze eenvoudig te berekenen:

$$u_0 = 1, u_{i+1} = (1 - \tau)u_i + \tau f_i,\ i = 0, 1, 2, \ldots \qquad (1.6)$$

M.b.v. (1.5) zouden we dus makkelijk $U(t_i)$ kunnen berekenen voor $i = 0, 1, 2, \ldots$, ware het niet dat $\frac{\tau}{2} U_{tt}(t_i + \theta_i\tau)$ als onbekende in (1.5) optrad.

We maken nu onze eerste fout, door de <u>afbreekfout</u>:

$$\frac{\tau}{2} U_{tt}(t_i + \theta_i\tau) \qquad (1.7)$$

te verwaarlozen.
M.a.w. we kiezen $f_i = F(t_i)$ en lossen (1.6) op, waarbij we opmerken dat de fout waarschijnlijk niet zo erg zal zijn aangezien de afbreekfout de factor τ bevat die we zelf mogen kiezen en die we dus klein kiezen.

1.1. <u>Practische rekenresultaten</u>

Het oplossen van (1.6) gebeurt met de rekenautomaat die we een ALGOL 60 programma: RPR 160469/01 aanbieden waarin $F(t) = \sin(t)$, zodat $U(t) = (3e^{-t} + \sin(t) - \cos(t))/2$.

```
begin comment Enkele print en pons procedures bij T 8182;
  procedure PR nlcr; PR string(‡
  †);
  procedure PR string(s); string s;
  begin PRINTTEXT(s); PUTEXT(s) end;
  procedure PR num(x); value x; real x;
  begin PRINT(x); PUNCH(x) end;

  begin comment Programma 1 bij college numerieke methoden
        voor p.d.v.'s. T8182, RPR 160469/01. R.P. van de Riet;
    real u,U,t,tau,e; integer i;
    procedure volgende u;
    begin i:= i + 1;
      u:= e × u + tau × sin(t); t:= t + tau;
      if i = i : 5 × 5 then
      begin PR nlcr; PR num(t); U:= exp(-t) × 1.5 + (sin(t) - cos(t))/2;
        PR string(‡abs. fout†); PR num(abs(U - u));
        PR string(‡rel. fout†); PR num(abs((U - u)/U))
      end
    end;

    PR nlcr; PR string(‡resultaten RPR 160469/01 tau =†);
    i:= 0; t:= 0; tau:= 10-1; u:= 1; PR num(tau); e:= 1 - tau;
    for i:= i while i < 10 do volgende u;
    PR nlcr; PR string(‡tau =†);
    i:= 0; t:= 0; tau:= 10-2; u:= 1; PR num(tau); e:= 1 - tau;
    for i:= i while i < 100 do volgende u;
    PR nlcr; PR string(‡100 × 10-2 =†); PR num(i × tau);
    PR nlcr; PR string(‡de exacte oplossing van het differentieschema:†);
    PR num(e ↑ 100 + (sin(.99) - e × sin(1) +
                        e ↑ 100 × sin(tau))/ (1 - 2 × e × cos(tau) + e ↑ 2)
                        × tau);
    PR nlcr; PR string(‡de boven berekende u geeft†);
    PR num(u);
    comment De bovenstaande exacte oplossing is niet zo erg exact daar er,
      bij voorbeeld in de berekening van 1 - 2 × e × cos(tau) + e↑2, cijfer-
      verlies gaat optreden. Een nauwgezette analyse levert het volgende
      programma op waarin geen cijferverlies optreedt.;
    PR nlcr; PR string(‡de werkelijk exacte oplossing is:†);
    U:= exp(-1) × exp(-.50335 85350 14₁₀-2);
```

```
    PR num(U + (-2 × cos(.995) × sin(.005) + .01 × sin(1) + U × sin(.01)) ×
          .01 /(.0001 + 4 × .99 × sin(.005)↑2));
    PR nlcr; PR string(«(1 - tau) ↑ 100 =»);
    u:= 1; for i:= 1 step 1 until 100 do u:= u × e; PR num(u);
    PR string(« en in 12 decimalen nauwkeurig:»); PR num(U)
  end
end
```

resultaten RPR 160469/01 tau =+.1000000000000_{10}- 0

+.5000000000000_{10}- 0 abs. fout+.3131875323015_{10}- 1 rel. fout+.4406638953287_{10}- 1

+.1000000000002_{10}+ 1 abs. fout+.3353143057302_{10}- 1 rel. fout+.4773813130834_{10}- 1

tau =+.1000000000001_{10}- 1

+.5000000000001_{10}- 1 abs. fout+.4786572781086_{10}- 3 rel. fout+.5025491738300_{10}- 3

+.1000000000000_{10}- 0 abs. fout+.9099290182348_{10}- 3 rel. fout+.1000283910960_{10}- 2

+.1500000000001_{10}- 0 abs. fout+.1296708737755_{10}- 2 rel. fout+.1488082908162_{10}- 2

+.2000000000000_{10}- 0 abs. fout+.1641723238208_{10}- 2 rel. fout+.1960506481407_{10}- 2

+.2500000000000_{10}- 0 abs. fout+.1947543467395_{10}- 2 rel. fout+.2411977013928_{10}- 2

+.3000000000011_{10}- 0 abs. fout+.2216594706624_{10}- 2 rel. fout+.2836989972334_{10}- 2

+.3500000000022_{10}- 0 abs. fout+.2451166131323_{10}- 2 rel. fout+.3230341737833_{10}- 2

+.4000000000033_{10}- 0 abs. fout+.2653419753187_{10}- 2 rel. fout+.3587356706820_{10}- 2

+.4500000000044_{10}- 0 abs. fout+.2825398810273_{10}- 2 rel. fout+.3904094480379_{10}- 2

+.5000000000055_{10}- 0 abs. fout+.2969035583192_{10}- 2 rel. fout+.4177518740441_{10}- 2

+.5500000000065_{10}- 0 abs. fout+.3086158741098_{10}- 2 rel. fout+.4405613138339_{10}- 2

+.6000000000076_{10}- 0 abs. fout+.3178500176546_{10}- 2 rel. fout+.4587435051420_{10}- 2

+.6500000000087_{10}- 0 abs. fout+.3247701379223_{10}- 2 rel. fout+.4723105097135_{10}- 2

+.7000000000098_{10}- 0 abs. fout+.3295319410427_{10}- 2 rel. fout+.4813737210576_{10}- 2

+.7500000000109_{10}- 0 abs. fout+.3322832435515_{10}- 2 rel. fout+.4861319674987_{10}- 2

+.8000000000120_{10}- 0 abs. fout+.3331644877107_{10}- 2 rel. fout+.4868561414945_{10}- 2

+.8500000000131_{10}- 0 abs. fout+.3323092225401_{10}- 2 rel. fout+.4838719299570_{10}- 2

+.9000000000142_{10}- 0 abs. fout+.3298445479231_{10}- 2 rel. fout+.4775421439675_{10}- 2

+.9500000000153_{10}- 0 abs. fout+.3258915277911_{10}- 2 rel. fout+.4682499277322_{10}- 2

+.1000000000016_{10}+ 1 abs. fout+.3205655725651_{10}- 2 rel. fout+.4563837908044_{10}- 2

$100 \times {}_{10}-2 =$ +1

de exacte oplossing van het differentieschema:+.6991978452934_{10}- 0

de boven berekende u geeft+.6991978455035_{10}- 0

de werkelijk exacte oplossing is:+.6991978455153_{10}- 0

(1 - tau) ↑ 100 =+.3660323412655_{10}- 0

en in 12 decimalen nauwkeurig:+.3660323412732_{10}- 0

Voor deze keuze van f_i (= sin(iτ)) is (1.6) exact op te lossen.

$$u_i = (1-\tau)^i + \frac{\sin(i-1)\tau - (1-\tau)\sin(i\tau) + (i-\tau)^i \sin(\tau)}{1 - 2(1-\tau)\cos\tau + (1-\tau)^2}\,\tau \qquad (1.8)$$

Zowel u_{100}, volgens (1.8), als u_{100}, volgens (1.6), zijn berekend. De grote discrepantie: $.2101_{10}{}^{-9}$ is te wijten aan de slechtheid van de formule (1.8) waarin verlies van cijfers optreedt door het aftrekken van vrijwel even grote getallen.
Bovendien wordt $(1-\tau)^{100}$ niet zo precies uitgerekend daar de relatieve fout in 1-τ met 100 vermenigvuldigd wordt in het resultaat; vandaar dat $(1-\tau)^{100}$ is berekend volgens:

$$(1-\tau)^{100} = e^{100\,\ln(1-\tau)}$$

en uit een tabel halen we:

$$\ln(1-\tau) = \ln 0.99 \approx -.010050335\ 85350\ 14.$$

Met behulp van een formule waarin geen cijferverlies optreedt en een betere berekening van $(1-\tau)^{100}$ is u_{100} nogmaals berekend.

We concluderen dat er een fout: $.118_{10}{}^{-10}$ in de berekening van u_{100} volgens (1.6) is gemaakt. De oorzaak is dat in elke stap rekenfouten worden gemaakt.

De berekende u_i's voldoen dan ook niet aan schema (1.6) maar aan het volgende schema:

$$\tilde{u}_0 = 1, \tilde{u}_{i+1} = (1-\tau)\tilde{u}_i + \tau f_i + \delta_i,\ i = 0, 1, 2, \ldots, \qquad (1.9)$$

waarin δ_i de fout is die per stap gemaakt wordt.

De resultaten van de bovengedane berekening doen vermoeden dat, als we τ maar kleiner maken, dat dan het antwoord beter wordt. Dit vermoeden wordt ten dele bevestigd in de theoretische beschouwing van paragraaf 1.3.

1.2. Theorie van eerste en hogere orde differentieschema's

Gegeven het eerste orde differentieschema:

$$a\,u_{k+1} - b\,u_k = d_k,\ k \geq 0. \tag{1.10}$$

De term "differentieschema" is gekozen omdat (1.10) ook in de vorm van differenties is te schrijven; namelijk:

$$a\Delta u_k - (b-a)u_k = d_k,\ k \geq 0, \tag{1.11}$$

waarin de operator Δ is gedefinieerd door

$$\Delta u_k = u_{k+1} - u_k. \tag{1.12}$$

Het bij (1.10) behorende differentieschema is:

$$a\,u_{k+1} - b\,u_k = 0,\ k \geq 0, \tag{1.13}$$

waarvan de algemene oplossing gegeven is door

$$u_k = (b/a)^k\,u_0,\ k \geq 1, \tag{1.14}$$

waarin u_0 willekeurig is.

De algemene oplossing van (1.10) is de som van een particuliere oplossing van (1.10) en de algemene oplossing van (1.13).
Een particuliere oplossing van (1.10) is:

$$u_{pk} = \frac{1}{a}\left(\frac{b}{a}\right)^{k-1} \sum_{i=0}^{k-1} \left(\frac{a}{b}\right)^i d_i. \tag{1.15}$$

Opmerking: (1.10) is een differentieschema met constante coëfficienten a en b. Interessante theorieën zijn opgesteld voor differentieschema's met niet constante coëfficienten.

Zie bijvoorbeeld:

Nörlund, N.E., Vorlesungen über Differenzen gleichungen, Springer, Berlin, 1924.

of Milne-Thomson, L.M., The calculus of finite differences, The Macmillan Co., New York, 1933.

of Miller, K.S., An introduction to the calculus of finite differences and difference equations, Dover, New York, 1966.

Beschouw nu het tweede-orde differentieschema:

$$a\,u_{k+2} - 2b\,u_{k+1} + c\,u_k = d_k,\ k \geq 0 \text{ en } a \neq 0, \qquad (1.16)$$

met als variant in differenties geschreven:

$$\{a\Delta^2 + 2(a - b)\Delta + (a - 2b + c)\}u_k = d_k,\ k \geq 0. \qquad (1.17)$$

Het homogene schema ($d_k = 0$) heeft als algemene oplossing:

$$u_k = C_1\alpha_1^{\,k} + C_2\alpha_2^{\,k}, \qquad (1.18)$$

waarin C_1 en C_2 willekeurig zijn en waarin:

$$\alpha_1 = (b + \sqrt{b^2 - ac})/a \text{ en } \alpha_2 = (b - \sqrt{b^2 - ac})/a, \qquad (1.19)$$

uiteraard onder de voorwaarde dat $b^2 \neq ac$.
Als $b^2 = ac$ dan vinden we de algemene oplossing als volgt:
Stel $\bar{c} = c - \delta$, $\delta > 0$, dan is $b^2 \neq a\bar{c}$,
De algemene oplossing voor het differentieschema:

$$a\,\bar{u}_{k+2} - 2b\,\bar{u}_{k+1} + \bar{c}\,u_k = 0,\ k \geq 0, \qquad (1.20)$$

is

$u_k = C_1 \bar{\alpha}_1^k + C_2 \bar{\alpha}_2^k$ met $\bar{\alpha}_1 = (b + \sqrt{a\delta})/a$ en $\alpha_2 = (b - \sqrt{a\delta})/a$.

kies $C_2 = - C_1 + D_1$, met D_1 willekeurig.
Dan geldt:

$$\bar{u}_k = C_1(\bar{\alpha}_1^k - \bar{\alpha}_2^k) + D_1 \bar{\alpha}_2^k =$$

$$= C_1 ((b + \sqrt{a\delta})^k - (b - \sqrt{a\delta})^k)/a^k + D_1 \bar{\alpha}_2^k$$

$$= C_1 (b^k + kb^{k-1} \sqrt{a\delta} + \frac{k(k-1)}{2} b^{k-2} a\delta + \ldots$$

$$- b^k + kb^{k-1} \sqrt{a\delta} - \frac{k(k-1)}{2} b^{k-2} a\delta + \ldots)/a^k$$

$$+ D_1 \bar{\alpha}_2^k$$

$$= C_1 \sqrt{a\delta} \frac{2}{b} k (b/a)^k + D_1 \bar{\alpha}_2^k + C_1 \delta R_k,$$

kies: $\bar{C}_1 = C_1 \sqrt{a\delta} \frac{2}{b}$; dan is

$$\bar{u}_k = \bar{C}_1 k (b/a)^k + D_1 \bar{\alpha}_2^k + \bar{C}_1 \sqrt{\delta} \bar{R}_k,$$

waarin $\bar{C}_1$ weer volmaakt willekeurig is.

Neem tenslotte de limiet voor $\delta \to 0$

$$u_k = \lim_{\delta \to 0} \bar{U}_k = \bar{C}_1 k(b/a)^k + D_1 (b/a)^k$$

We zullen het bewijs dat $k(b/a)^k$ een oplossing is nog even uitstellen.

Het geval $b^2 < ac$ behoeft nog enig commentaar.
We stellen $\alpha_1 = \rho e^{i\phi}$, zodat $\alpha_2 = \rho e^{-i\phi}$.

Dan is

$$u_k = C_1 \rho^k e^{-ik\phi} + C_2 \rho^k e^{-ik\phi}$$

$$= \rho^k \{(C_1 + C_2) \cos k\phi + (C_1 - C_2)\, i \sin k\phi\}.$$

Kies nu $D_1 = C_1 + C_2$ en $D_2 = (C_1 - C_2)\, i$, welke ook willekeurig zijn, omdat C_1 en C_2 willekeurig zijn; dan volgt als reële algemene oplossing:

$$u_k = \rho^k (D_1 \cos k\phi + D_2 \sin k\phi), \tag{1.21}$$

of

$$u_k = \rho^k A \sin (k\Phi + \psi), \tag{1.22}$$

met $A = (D_1^2 + D_2^2)^{1/2}$ en $\sin \psi = D_1/A$,
waarin ook A en ψ willekeurig zijn.

De algemene oplossing van het inhomogene probleem ($d_k \neq 0$ in (1.16)) is de som van een particuliere oplossing van (1.16) met de algemene oplossing van het homogene probleem.

Een particuliere oplossing vinden we m.b.v. "variatie van constanten" zoals in de theorie van gewone differentiaalvergelijkingen gebruikelijk is.
We nemen in plaats van (1.16), (1.17) als uitgangspunt:

$$(\alpha\Delta^2 + \beta\Delta + \gamma)\, u_k = d_k,\ k \geq 0, \tag{1.23}$$

met $\alpha = a$, $\beta = 2(a - b)$ en $\gamma = a - 2b + c$.

Zij $u_k = C_1 u_{1,k} + C_2 u_{2,k}$ de algemene oplossing van het homogene probleem.

We stellen:

$$u_{p,k} = C_{1,k}\, u_{1,k} + C_{2,k}\, u_{2,k}, \tag{1.24}$$

waarin $C_{1,k}$ en $C_{2,k}$ te bepalen zijn zó dat aan (1.23) voldaan is door $u_{p,k}$.

We merken op dat

$$\Delta(f_k.g_k) = f_{k+1}\, \Delta g_k + g_k\, \Delta f_k. \tag{1.25}$$

en

$$\Delta(f_k.g_k) = f_k\, \Delta g_k + g_{k+1}\, \Delta f_k. \tag{1.26}$$

Dus

$$\Delta u_{p,k} = C_{1,k+1}\, \Delta u_{1,k} + u_{1,k}\, \Delta C_{1,k} + \ldots, \tag{1.27}$$

en

$$\Delta u_{p,k} = C_{1,k}\, \Delta u_{1,k} + u_{1,k+1}\, \Delta C_{1,k} + \ldots. \tag{1.28}$$

We willen ervoor zorgen dat (1.23) overgaat in:

$$C_{1,k}\, (\alpha\Delta^2 + \beta\Delta + \gamma)\, u_{1,k} + C_{2,k}\, (\alpha\Delta^2 + \beta\Delta + \gamma)\, u_{2,k} + \ldots = d_k,$$

zodat we (1.28) kiezen.
In navolging van de theorie van de gewone differentiaalvergelijkingen stellen we

$$u_{1,k+1}\, \Delta C_{1,k} + u_{2,k+1}\, \Delta C_{2,k} = 0, \tag{1.29}$$

zodat

$$\Delta^2 u_{p,k} = C_{1,k}\, \Delta^2 u_{1,k} + \Delta u_{1,k+1}\, \Delta C_{1,k}$$

$$+ C_{2,k}\, \Delta^2 u_{2,k} + \Delta u_{2,k+1}\, \Delta C_{2,k}. \tag{1.30}$$

Substitutie van (1.28) en (1.30) in (1.23) geeft:

$$\alpha(\Delta u_{1,k+1}\, \Delta C_{1,k} + \Delta u_{2,k+1}\, \Delta C_{2,k}) = d_k. \tag{1.31}$$

Uit (1.29) en (1.31) vinden we

$$\Delta C_{1,k} = - u_{2,k+1} \frac{d_k}{\alpha} / (u_{1,k+1} \Delta u_{2,k+1} - u_{2,k+1} \Delta u_{1,k+1}), \qquad (1.32)$$

$$\Delta C_{2,k} = u_{1,k+1} \frac{d_k}{\alpha} / (u_{1,k+1} \Delta u_{2,k+1} - u_{2,k+1} \Delta u_{1,k+1}).$$

Waaruit onmiddellijk volgen:

$$C_{1,k} = C_{1,k} - C_{1,k-1} + C_{1,k-1} - C_{1,k-2} + \ldots$$

$$+ C_{1,1} - C_{1,0} + C_{1,0}$$

$$= \sum_{i=0}^{k-1} \Delta C_{1,i} + C_{1,0},$$

en (1.33)

$$C_{2,k} = \sum_{i=0}^{k-1} \Delta C_{2,i} + C_{2,0}.$$

Substitutie van (1.32) in (1.33) geeft de gevraagde $C_{1,k}$'s en $C_{2,k}$'s, met $C_{1,0}$ en $C_{2,0}$ willekeurig.

Het n- de orde differentieschema:

$$a_n u_{k+n} + a_{n-1} u_{k+n-1} + \ldots + a_0 u_k = d_k, \; k \geq 0, \qquad (1.34)$$

heeft voor het homogene geval, $d_k = 0$, de algemene oplossing:

$$u_k = \sum_{\nu=1}^{p} \sum_{i=0}^{\mu_\nu - 1} C_{\nu,i} \, k^i \, \alpha_\nu^k, \qquad (1.35)$$

waarin α_ν het ν-de nulpunt, met multipliciteit μ_ν, van het polynoom

$$P(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_0 \tag{1.36}$$

is. Kennelijk geldt: $\sum_{\nu=1}^{p} \mu_\nu = n$.

Het is duidelijk dat, als $\mu_\nu = 1$, dat $\alpha_\nu{}^k$ een oplossing van (1.34) is. Beschouwen we nu het geval $\mu_\nu > 1$ dan geldt:

$$\frac{d^i P(x)}{d x^i} = 0, \text{ voor } x = \alpha_\nu \text{ en } i = 0, \ldots, \mu_\nu - 1. \tag{1.37}$$

of

$$\sum_{j=0}^{n} j\,(j-1)\ldots(j-i+1)\, a_j\, x^{j-1} = 0, \; i = 0, \ldots, \mu_\nu - 1. \tag{1.38}$$

Voeren we nu de notatie in:

$$y^{(k)} = y\,(y-1)\ldots(y-k+1). \tag{1.39}$$

dan wordt (1.38):

$$\sum_{j=0}^{n} j^{(i)} a_j\, x^{j-i} = 0, \; i = 0, \ldots, \mu_\nu - 1. \tag{1.40}$$

Nu geldt:

$$y^k = \sum_{j=0}^{k} A_{k,j}\, y^{(j)}, \tag{1.41}$$

met, uiteraard, $A_{k,k} = 1$.

Bedenken we nu dat $y^{(i)} = 0$ voor $y = 0, 1, \ldots, j-1$,
dan zien we dat

$$y = 0 \to 0^k = A_{k,0}\, y^{(0)}, \text{ zodat } A_{k,0} = 0$$

$$y = 1 \to 1^k = A_{k,1}\, 1^{(1)}, \text{ zodat } A_{k,1} = 1,$$

$$y = i \to i^k = \sum_{j=1}^{i-1} A_{k,j}\, i^{(j)} + A_{k,i}\, i^{(i)},$$

zodat

$$A_{k,i} = (i^k - \sum_{j=1}^{i-1} A_{k,j}\, i^{(j)}) \,/\, i! \,. \qquad (1.42)$$

Nu nemen we de oplossing:

$$u_k = k^i\, \alpha_\nu{}^k, \qquad (1.43)$$

en substitueren deze in (1.34); het resultaat is:

$$\sum_{j=0}^{n} a_j\, (k+j)^i\, \alpha_\nu{}^{k+j} =$$

$$= \alpha_\nu{}^k \sum_{j=0}^{n} a_j \sum_{m=0}^{i} \binom{i}{m} k^{i-m}\, j^i\, \alpha_\nu{}^j =$$

$$= \alpha_\nu{}^k \sum_{m=0}^{i} \binom{i}{m} k^{i-m} \sum_{j=0}^{n} j^i\, a_j \alpha_\nu{}^j =$$

$$= \alpha_\nu{}^k \sum_{m=0}^{i} \binom{i}{m} k^{i-m} \sum_{j=0}^{n} \sum_{l=0}^{i} A_{i,l}{}^{(l)}\, a_j\, \alpha_\nu{}^j =$$

$$= \alpha_\nu{}^k \sum_{m=0}^{i} \binom{i}{m} k^{i-m} \sum_{l=0}^{i} A_{i,l} \sum_{j=0}^{n} j^{(l)}\, a_j\, \alpha_\nu{}^j \qquad (1.44)$$

mits l, en dus i, kleiner zijn dan μ_ν, kunnen we gebruik maken van (1.40), zodat (1.44) nul oplevert, zodat inderdaad (1.43) een oplossing van het homogene differentieschema (1.34) is.

(Opmerking de functie $k^{(a)}$ heeft de eigenschap dat $\Delta k^{(a)} = a\, k^{(a-1)}$)

Stilzwijgend zijn we er boven van uitgegaan dat $\alpha_\nu \neq 0$.
Echter, als $\alpha_\nu = 0$ dan zijn $a_0, \ldots, a_{\mu_{\nu-1}}$ alle nul, zodat (1.34) reduceert tot:

$$a_n\, u_{k+n} + \ldots + a_{\beta_\nu}\, u_{k+\beta_\nu} = d_k,\ k \geq 0,$$

hetgeen door een hernummering: $k' = k+\mu_\nu$ overgaat in een $(n-\mu_\nu)$ orde differentieschema.

Nu geldt

$$\int_0^t e^x \sin x \, dx = \sum_{j=0}^{(k-1)} \int_{j\tau}^{(j+1)\tau} e^x \sin x \, dx$$

$$= \sum_{j=0}^{k-1} \int_{j\tau}^{(j+1)\tau} \{e^{j\tau} \sin(j\tau) + (x-j\tau) \frac{d}{dx} (e^x \sin x)_{x = j\tau + \theta\tau}\} \, dx$$

$$= S_1 + O(\tau).$$

Ook $S_2 = O(\tau)$, zodat we vinden:

$$\tilde{u}_k = e^{-t} + e^{-t} \int_0^t e^x \sin x \, dx + O(\tau)$$

$$+ e^{-t} \sum_{j=0}^{k-1} e^{j\tau} \delta_j.$$

Dus

$$\tilde{u}_k - U(t) = O(\tau) + e^{-t} \sum_{j=0}^{k-1} e^{j\tau} \delta_j. \qquad (1.48)$$

We concluderen dat het verschil tussen $\tilde{u}_k$ en $U(t)$ enerzijds veroorzaakt wordt door de afbreekfout (1.7) die zich manifesteert in de $O(\tau)$ term in (1.48) en anderzijds veroorzaakt wordt door de rekenfouten δ_j.

Nemen we het slechtste geval waarin $\delta_j = d$, met d bijvoorbeeld 10^{-10} dan is de totale rekenfout:

$$\nu = \frac{1-e^{-t}}{e^{\tau}-1} d.$$

Verbetering van de berekende $\tilde{u}_k$ betekent derhalve dat τ niet ongelimiteerd klein gekozen kan worden; immers $e^{\tau}-1 = \tau + \tau^2/2 + \ldots$ zodat ν voor grote t gelijk is aan

$$\nu = \frac{d}{\tau}\,(1 + O(\tau)). \qquad (1.49)$$

De situatie is nu wel wat al te gechargeerd voorgesteld; immers de δ_j's zullen in de eerste plaats zowel negatief als positief zijn.
Bovendien zijn de δ_j's relatieve fouten.
Beschouwen we de formule $u_{i+1} = (1-\tau)u_i + \tau\, f_i$ en nemen we aan dat er relatieve fouten worden gemaakt dan zal de rekenautomaat rekenen:

$$\begin{aligned}\tilde{u}_{i+1} &= \{((1-\tau)(1+e_1) \times \tilde{u}_i)(1+e_2) + \tau \times f_i(1+e_3)\}\,(1+e_4)\\ &= (1-\tau)\tilde{u}_i + \tau\, f_i + (1-\tau)\tilde{u}_i\, e_1 + (1-\tau)\tilde{u}_i\, e_2 + \tau\, f_i\, e_3\\ &\quad + ((1-\tau)\tilde{u}_i + \tau\, f_i)\, e_4 + \ldots .\end{aligned}$$

Dus $\delta_i \approx \max(\tau\, f_i, \tilde{u}_i) \times d$.
Voor grote t is echter $\tilde{u}_i \approx (\sin(t) - \cos(t))/2$, zodat er toch geen factor τ in δ_i zal optreden.
We mogen dus <u>niet</u> de limiet voor $\tau \to 0$ nemen.

<u>Opmerking</u>: stellen we $e_i = \frac{1}{2}\,10^{-12}$ en $\delta_i = \frac{1}{2}\,10^{-12}$ dan is voor t = 1 is $\nu \approx \frac{1-e^{-1}}{\tau}\ \frac{1}{2}\,10^{-12}\,\bar{u} \approx .28_{10}{}^{-10}$ (met $\bar{u}$ de gemiddelde u) voor $\tau = 10^{-2}$.
Het programma RPR 160469/01 gaf $\nu = .12_{10}{}^{-10}$ dus ongeveer twee keer zo klein.

Bovenstaande beschouwing is enerzijds wat pessimistisch, anderzijds wat onrealistisch.
Voordat we namelijk last krijgen van de rekenfout moet τ wel verschrikkelijk klein gekozen worden; zó klein, zeg 10^{-5}, dat we 100 000 stappen nodig hebben om een benaderde waarde van U(1) te krijgen, in ruwweg 5 decimalen nauwkeurig.

Gelukkig bestaan er betere methoden om U(1) zo nauwkeurig uit te rekenen (afgezien dan m.b.v. de analytische formule).
Daartoe is het nodig om de afbreekfout aan een nadere analyse te onderwerpen, aangezien dit de boosdoener is; immers als $\tau = 10^{-2}$ en $d = \frac{1}{2}\ 10^{-12}$ dan is $0(10^{-2})$ vele ordes groter dan $\nu \approx 10^{-10}$.

1.4. Hogere orde afbreekfouten

We beschouwen het n-de orde differentieschema

$$Ru_k \overset{d}{=} \sum_{j=0}^{n} a_j\, u_{j+k} = d_k,\ k \geq 0. \tag{1.50}$$

Substitutie van $u_k = U(k\tau)$ in het linkerlid van (1.50) geeft

$$RU(k\tau) = \sum_{j=0}^{n} a_j\, U((j+k)\tau).$$

We ontwikkelen $U((j+k)\tau)$ in het punt $t_{k+m} = (k+m)\tau$,
waarin m nog vrij te kiezen is; dit geeft:

$$RU(k\tau) = \sum_{i=0}^{\infty} \frac{\tau^i}{i!} \left.\frac{d^i U}{dt^i}\right|_{t=t_{k+m}} \sum_{j=0}^{n} (j-m)^i\, a_j. \tag{1.51}$$

Deze uitdrukking willen we zo goed mogelijk laten overeenstemmen met het linkerlid van (1.1) d.w.z. met dU/dt + U genomen in het punt t_{k+m}.

Aangezien we n+1 a_j's hebben kunnen we n+1 condities stellen:

$$\left.\begin{aligned} i = 0 &: \quad \sum_{j=0}^{n} a_j = 1, \\ i = 1 &: \tau \sum_{j=0}^{n} (j-m)\, a_j = 1, \\ 1 < i \leq &: \tau^i \sum_{j=0}^{n} (j-m)^i\, a_j = 0. \end{aligned}\right\} \qquad (1.52)$$

De laatste n vergelijkingen zijn:

$$-m\, a_0 + (-m+1)\, a_1 + \ldots + (-1) a_{m-1} + (1) a_{m+1} + \ldots + (n-m) a_n = 1/\tau$$

$$(-m)^2\, a_0 + (-m+1)^2\, a_1 + \ldots + (-1)^2\, a_{m-1} + (1)^2\, a_{m+1} + \ldots + (n-m)^2\, a_n = 0$$

$$(-m)^n\, a_0 + (-m+1)^n\, a_1 + \ldots + (-1)^n\, a_{m-1} + (1)^n\, a_{m+1} + \ldots + (n-m)^n\, a_n = 0$$

Met als oplossing, voor $i \neq m$:

$$a_i = \frac{\begin{vmatrix} (-m) & \ldots & (-m+i-1) & \frac{1}{\tau} & (-m+i+1) & \ldots & (n-m) \\ (-m)^2 & \ldots & (-m+i-1)^2 & 0 & (-m+i+1)^2 & \ldots & (n-m)^2 \\ & & & - - - - - & & & \\ (-m)^n & \ldots & (-m+i-1)^n & 0 & (-m+i+1)^n & \ldots & (n-m)^n \end{vmatrix}}{\begin{vmatrix} (-m) & \ldots & (n-m) \\ (-m)^2 & \ldots & (n-m)^2 \\ & - - - - - & \\ (-m)^n & \ldots & (n-m)^n \end{vmatrix}}$$

$$= \frac{(-1)^i}{\tau} \prod_{\substack{j\neq m \\ j\neq i}} (-m+j)^2 \prod_{\substack{j\neq m \\ k\neq m \\ j\neq i \\ k\neq i \\ j\neq k \\ (j,k)\ \text{cyclisch}}} (j-k) \Big/ \Big\{ \prod_{j\neq m} (-m+j) \prod_{\substack{j\neq m \\ k\neq m \\ j\neq k \\ (j,k)\ \text{cyclisch}}} (j-k) \Big\}, \qquad (1.53)$$

eb $a_m = 1 - \sum_{i\neq m} a_1$.

Uit

$$RU(t_k) = \left(\frac{dU}{dt} + U\right)\Big|_{t=t_{k+m}} + \frac{\tau^{n+1}}{(n+1)!} \sum_{j=0}^{n} (j-m)^{n+1} a_j \frac{d^{n+1}U}{d\, t^{n+1}} \Big|_{t=\bar{t}_k},$$

met $t_k < \bar{t}_k < t_{k+n}$,

zien we dat met de keus

$$d_k = \left(\frac{dU}{dt} + U\right) t + t_{k+m} = F(t_{k+m}),$$

de afbreekfout $0\,(\tau^n)$ is, immers voor $j\neq m$ geldt:

$$a_j = 0\,(\tfrac{1}{\tau}).$$

De algemene oplossing van het homogene deel van (1.50) is gegeven in (1.35):

$$u_k = \sum_{\nu=1}^{p} \sum_{i=0}^{\mu_\nu - 1} C_{\nu,i}\, k^i\, \alpha_\nu^{\,k}, \qquad (1.35)$$

waarin α_ν, $\nu = 1, \ldots, p$ de nulpunten zijn van $\sum_{j=0}^{n} a_j x^j = 0$ met multipliciteit μ_ν.

Een oplossing van (1.50) vinden we door n begin condities op te leggen:

$$u_i = b_i,\ i = 0, \ldots, n-1, \qquad (1.54)$$

en vervolgens u_n, u_{n+1}, u_{n+2}, etc. te berekenen.

Veronderstel nu dat $F(t) = 0$ en dus $d_i = 0$.
Ongetwijfeld is er een nulpunt, zeg, $\alpha_1 = 1 - \tau + 0(\tau^2)$
zodat de bijbehorende oplossing $u_k \approx (1-\tau)^k \approx \exp(-t_k)$.
De overige nulpunten, die dus niet gelijk zijn aan $1 - \tau + 0(\tau^2)$
zorgen echter voor oplossingen die niet naar $\exp(-t_k)$ convergeren.
Bovendien zorgen nulpunten met multipliciteit groter dan 1 voor oplossingen van de vorm $u_k = k^i \alpha_\nu^k$, die, als $\alpha_\nu \neq 0$, naar oneindig convergeren. Het is ook mogelijk dat de overige nulpunten $0(\tau)$ zijn; immers, als $\alpha_\nu = \varepsilon_\nu \tau + 0(\tau^2)$ dan geldt

$$\sum_{j=0}^{n} a_j x^j = \{x - (1-\tau+0(\tau^2))\} \prod_{\nu=2}^{n} (x-\varepsilon_\nu \tau+0(\tau^2)),$$

zodat
$$\frac{a_0}{a_n} = (-1)^{n+1} (1-\tau) \tau^{n-1} \prod_{\nu=2}^{n} (\varepsilon_\nu) + 0(\tau^2) ;$$

maar
$$\frac{a_0}{a_n} \text{ is hoogstens } 0(\tau), \text{ zodat n hoogstens 2 kan zijn.}^{*})$$

De conclusie is dat de nog te kiezen b_i's zeer bijzonder gekozen moeten worden, n.l. zó dat de coëfficienten $C_{\nu,i}$ alle nul zijn op $C_{1,0}$ na.
Stel dat we deze b_i's zo kiezen, dan mogen we uiteraard geen rekenfout maken, want elke rekenfout zou een oplossing behorende bij een ander nulpunt introduceren.

Voor $m = 0$, $n = 1$, vinden we (1.3) terug met $\alpha_1 = 1-\tau$.

Voor $m = 0$, $n = 2$, vinden we

$$\frac{1}{\tau} (-\tfrac{1}{2}u_{k+2} + 2u_{k+1} - \tfrac{3}{2}u_k) + u_k = d_k, \tag{1.55}$$

met
$$\alpha_1 = 2 - (1+2\tau)^{\frac{1}{2}} \text{ en } \alpha_2 = 2 + (1+2\tau)^{\frac{1}{2}},$$

zodat
$$u_k \approx C_1 e^{-t_k} + C_2 3^k e^{t_k/3}. \tag{1.56}$$

*) Voor n = 2 geldt zoals we zullen zien, $\alpha_2 \neq 0(\tau)$.

Voor m = 1, n = 2 vinden we

$$\frac{1}{2\tau}(u_{k+2} - u_k) + u_{k+1} = d_k, \tag{1.57}$$

met $\alpha_1 = -\tau + (1+\tau^2)^{\frac{1}{2}}$ en $\alpha_2 = -\tau - (1+\tau^2)^{\frac{1}{2}}$,

zodat $$u_k = C_1\, e^{-t_k} + C_2\, (-1)^k\, e^{t_k}. \tag{1.58}$$

Inderdaad zien we in (1.56) en in (1.58) twee oplossingen optreden die voor grote k en grote t de gevraagde oplossing aanzienlijk kunnen verstoren. De beide oplossingen $(\alpha_2^{\,k})$ worden wel <u>parasitaire oplossingen</u> genoemd.

Wel hebben we bereikt dat de afbreekfout een orde kleiner is geworden; de rekenfout is echter in aanzienlijke mate in betekenis gegroeid. Dit is voor het schema (1.55) nog veel belangrijker dan voor schema (1.57); immers als t vast is en $\tau = t/k$, met $k \to \infty$ dan is de invloed van 3^k funest.

De drie differentieschema's die we bestudeerd hebben: (1.3), (1.55) en (1.57) kunnen we voorstellen door:

$$R\, u_k = d_k,\ k \geq 0.$$

We hebben gezien dat

$$R\, U(t_k) - L\, U(t_{k+m}) = O(\tau^i)$$

met

$$L\, U = \frac{dU(t)}{dt} + U(t) \text{ en } i > 0.$$

Er geldt dus $\lim_{\tau \to 0} R\, U(t_k) - L\, U(t_{k+m}) = 0$;

We zeggen daarom dat de differentieschema's <u>consistent</u> zijn.

Voor differentieschema (1.3) zagen we dat een rekenfout niet exponentieel als functie van t en τ aangroeit; het wordt stabiel genoemd. Bij (1.55) groeit de rekenfout exponentieel als functie vast aan en wordt instabiel genoemd. Bij (1.57) groeit de rekenfout exponentieel als functie van t aan.

Als er geen rekenfouten gemaakt worden dan zou de oplossing van het differentieschema (1.3) convergeren naar de echte oplossing; de oplossing van (1.55) convergeert slechts in het homogene geval onder zeer geschikte keuze van de b_i's in (1.54).
Het valt daarom te verwachten dat consistentie en stabiliteit voldoende voorwaarden zijn voor de convergentie.
Hierop komen we in het volgende hoofdstuk terug.

Opmerking: Het heeft de schijn dat het schema (1.50) met de condities (1.52) het meest algemene schema is voor een afbreekfout van de orde τ^n.
Echter het schema:

$$\frac{u_{k+1} - u_k}{\tau} + \frac{u_{k+1} + u_k}{2} = F(t_{k+\frac{1}{2}}), \qquad (1.59)$$

heeft een afbreekfout van de orde τ^2 en valt niet onder het vermelde "algemene" schema.

Opgaven:

1.1. Verklaar het merkwaardige fenomeen in bovenstaande opmerking.

1.2. Ga de berekening van (1.56 - 1.58) na.

1.3. Maak tweede orde differentieschema's voor dU/dt - U(t) = F(t) en ga het effect van de parasitaire oplossing na.

1.4. Kies het schema (1.57) met $d_k = \delta$ en bereken m.b.v. (1.33) de oplossing van (1.57) met begin condities $u_0 = 0$, $u_1 = 0$.

1.4. Verbetering door extrapolatie

In het voorgaande hebben we getracht een betere benadering van de exacte oplossing te vinden door de orde van het differentieschema hoger te kiezen.
Dit had tot resultaat een hogere orde afbreekfout en meer of minder instabiele schema's met grotere rekenfouten.
Door echter m.b.v. schema (1.3) eerst voor zekere τ_0, vervolgens voor $\tau_1 = \tau_0/2, \ldots,$ tot $\tau_n = \tau_{n-1}/2$ het differentieschema door te rekenen tot $k_i\tau_i = t$, kunnen we door extrapolatie een nog nauwkeuriger antwoord vinden met gebruik making van (1.48), waarin we $\delta_j = 0$ stellen.

Bij τ_i vinden we:

$$u_{k_i} - U(t) = O(\tau_i) = C_1\tau_i + C_2\tau_i^2 + \ldots, \tag{1.60a}$$

$$u_{k_{i+1}} - U(t) = \mathcal{O}(\tau_i/2) = \tfrac{1}{2} C_1\tau_i + \tfrac{1}{4} C_2\tau_i^2 + \ldots. \tag{1.60b}$$

Onbekend zijn: $U(t)$, C_1, C_2, ...
2* (1.60b) - (1.60a) geeft:

$$u(t) = 2\, u_{k_{i+1}} - u_{k_i} + \tfrac{1}{2} C_2\tau_i^2 + \ldots \tag{1.61}$$

Gebruiken we bovendien de (i+2)de stap dan vinden we

$$U(t) = 2\, u_{k_{i+2}} - u_{k_{i+1}} + \frac{1}{8} C_2\tau_i^2 + \ldots;$$

en een nog betere benadering van $U(t)$ is:

$$U(t) = \frac{1}{3}\left(8\, u_{k_{i+2}} - 6\, u_{k_{i+1}} + u_{k_i}\right) + O(\tau_i^3). \tag{1.62}$$

Voor $F(t) = t$ en dus $U(t) = 2e^{-t} + t - 1$, is de berekening voor $t = 1$ uitgevoerd, te beginnen met $\tau_0 = 1/40$.
De resultaten zijn de volgende:

	$U(1) = 2e^{-1}$	= .73575 88823
Voor $\tau = \tau_0$ en m.b.v. (1.3):	u_{40}	= .74238 40688
Voor $\tau = \tau_1$ " " "	u_{80}	= .73906 66831
Voor $\tau = \tau_2$ " " "	u_{160}	= .73741 15868
Voor $\tau = \tau_0$ en m.b.v. (1.61):		.73574 93038
Voor $\tau = \tau_1$ en m.b.v. (1.61):		.73575 64867
Voor $\tau = \tau_0$ en m.b.v. (1.62):		.73575 88818

We bereiken dus een zeer goede precisie van 8 decimalen.
Als we inplaats van met absolute fouten, met relatieve fouten rekenen dan moeten we de volgende fout afschattings formule gebruiken.

$$U(t) = u_k \, (1 + \bar{C}_1\tau + \bar{C}_2\tau^2 + \ldots),$$

en er ontstaan de volgende extrapolatie formules:

$$U(t) = u_{k_{i+1}} \, u_{k_i} \, /(2 \, u_{k_i} - u_{k_{i+1}}) + O(\tau^2),$$

en

$$U(t) = 3 \, (u_{k_i}^{-1} - 6 \, u_{k_{i+1}}^{-1} + 8 \, u_{k_{i+2}}^{-1})^{-1} + O(\tau^3).$$

De laatste formule geeft voor bovenstaand voorbeeld:
het getal .73575 8908, dus iets slechter dan het getal dat m.b.v. (1.62) verkregen is.

Conclusie: door in 160 + 80 + 40 punten de u_k's te berekenen vinden we m.b.v. (1.62) $U(1)$ in 8 decimalen nauwkeurig, bovendien vinden we ook $U(i/40)$, $i = 1, \ldots, 39$, in 8 decimalen nauwkeurig. Echter, in de overige punten vinden we de U niet zo nauwkeurig.

2. Consistentie, stabiliteit en convergentie.

Gezocht wordt een functie U(x), met x een punt in een d-dimensionale ruimte E_d, die voldoet aan

$$L\,U = F. \tag{2.1}$$

Met L een operator en F een functie van x. Het punt x wordt geacht te liggen in een gebied G met eventuele rand Γ.
De operator is i.h.a. een differentiaal operator.
De functie U mag een vectorfunctie zijn.
De oplossing

$$U = L^{-1}\,F, \tag{2.2}$$

wordt geacht te bestaan en continu van F af te hangen.

Opmerking: Eventuele rand en begin condities worden alle in (2.1) opgenomen. Het probleem (1.1) luidt:

$$U(0) = 1 \quad t = 0$$

$$\left(\frac{d}{dt} + 1\right) U = F, \quad t > 0$$

zodat in dit geval:

$$LU = \begin{cases} U(0), \text{ als } t = 0 \\ \left(\frac{d}{dt} + 1\right) U, \text{ als } t > 0 \end{cases} \quad \text{en} \quad F \begin{cases} 1, \text{ als } t = 0 \\ F^{*}(t), \text{ als } t > 0. \end{cases}$$

met $F^{*}(t)$ de oude F(t).

Aangenomen wordt dat U in een lineaire ruimte E_U ligt, en dat F in een lineaire ruimte E_F ligt.
In de E_d kiezen we een rooster, d.w.z. een verzameling punten x_n, met $x_n = (x_{1,n_1}, \ldots, x_{d,n_d})$, $n_i = \ldots -2, -1, 0, 1, 2, \ldots$, voor $i = 1, \ldots, d$.

De <u>maaswijdte</u> h van het rooster geeft de dichtheid van het rooster aan. h zou bijvoorbeeld gedefinieerd kunnen worden als

$$h = \sup_n \left(\min_{\substack{m \\ m \neq n}} | x_n - x_m | \right).$$

Zij g een deelverzameling van alle roosterpunten.
Meestal zal gelden dat $g = \{x_n : x_n \in G\}$.
We beschouwen roosterfuncties u_n en f_n gedefinieerd op de punten x_n van g. Deze roosterfuncties liggen in de <u>genormeerde</u> ruimten E_u, respectievelijk E_f.

Zij R een <u>lineaire</u> operator die een u op een f afbeeldt:

$$R_u = f. \tag{2.3}$$

We eisen dat er een eenduidige inverse R^{-1} van R bestaat (d.w.z. er moet een algorithme te vinden zijn zó dat bij gegeven f een u te berekenen is).

Om (2.1) en (2.3) met elkaar in verband te brengen is het nodig om aan U(x) een u_n toe te voegen. We voeren daartoe de <u>discretisatie operator</u> $[\ ]$ in zó dat

$$u = [U] \tag{2.4}$$

een roosterfunctie is en zó dat $u \in E_u$, voor alle $U \in E_U$.
Evenzo moet gelden:

$$f = [F] \in E_f, \text{ voor alle } F \in E_F. \tag{2.5}$$

Aangezien de U functies en de F functies van totaal verschillend karakter zijn kan de keuze van $[U]$ en van $[F]$ wezenlijk verschillend zijn.
Een voorbeeld van $[U]$ is:

$$[U]_n = U(x_n).$$

We eisen dat:

a. De operator $[\ ]$ is lineair.
b. De operator $[\ ]$ is uniform continu.

Voor een rij van roosters met afnemende maaswijdte h lossen we (2.3) op, met steeds een bijbehorende operator R_h en rechterlid f_h.
Dus

$$u_h = R^{-1}{}_h\, f_h; \tag{2.6}$$

en we kiezen de roosters zó dat $h \to 0$.
Uiteraard zijn we geïnteresseerd in:

$$[U]_h - u_h,$$

waarin $[\ ]_h$ de $[\ ]$ operator is van het rooster met maaswijdte h.
Nu geldt

$$\begin{aligned}
[U]_h - u_h &= R_h^{-1}\, R_h\, ([U]_h - u_h) \\
&= R_h^{-1}\, (R_h\, [U]_h - f_h) = \\
&= R_h^{-1}\, (R_h\, [LU]_h - [LU]_h + [LU]_h - f_h) \\
&= R_h^{-1}\, \{(R_h\, [U]_h - [LU]_h + ([F]_h - f_h)\}
\end{aligned}$$

In aanmerking nemende dat de ruimten E_u en E_f genormeerd zijn, kunnen we nu stellen:

$$\| [U]_h - u_h \| \leq \| R_h^{-1} \|\ \| (R_h\, [U]_h - [LU]_h + ([F]_h - f_h) \|, \tag{2.7}$$

waarin

$$\| R_h^{-1} \| = \sup_{f \in E_f} \frac{\| R_h^{-1} f \|}{\| f \|}. \tag{2.8}$$

De ongelijkheid (2.7) is het uitgangspunt voor de volgende definities.

Definitie: De discretiseringsfout Δ_h van het differentieschema $R_h u_h = f_h$ ten opzichte van de vergelijking LU = F is gegeven door

$$\Delta_h = (R_h [U]_h - [LU]_h) + ([F]_h - f_h). \tag{2.9}$$

Definitie: Het differentieschema $R_h u_h = f_h$ heet consistent t.o.v. de vergelijking LU = F als geldt:

$$\lim_{h \to 0} || \Delta_h || = 0. \tag{2.10}$$

Definitie: De differentie operator R_h heet stabiel als de operator R_h^{-1} begrensd is; dat wil zeggen:

$$\exists M > 0,\ h_0 > 0,\ || R_h^{-1} || < M, \text{ voor } h \leq h_0. \tag{2.11}$$

Definitie: Het differentieschema $R_h u_h = f_h$ heet convergent met betrekking tot de vergelijking LU = F, als de oplossing u_h "convergeert" naar de oplossing U, d.w.z.

$$\lim_{h \to 0} || [U]_h - u_h || = 0. \tag{2.12}$$

Stelling: Een consistent en stabiel differentieschema convergeert.

Bewijs: Volgt onmiddellijk uit (2.7).

Aangezien we nog in het geheel de operator $[U]$ niet vast gelegd hebben zegt de convergentie van R nog niet veel.
Daarom spreken we nu af dat:

$$[U]_n \overset{d}{=} U(x_n), \tag{2.13}$$

zodat (2.12) zinvol wordt; immers als we nu een punt x kiezen dat in alle roosters ligt en zo dat

$$\lim_{h\to 0} x_{n_h} = x$$

dan zegt (2.12) dat $\lim_{h\to 0} |U(x_{n_h}) - u_{n_h}| = 0$, als we als norm de maximum norm kiezen.

Door (2.13) te stellen hebben we schijnbaar een vrijheid verloren; dit is echter niet het geval aangezien we $[\bar{F}]$ nog willekeurig kunnen kiezen.

Een belangrijke eigenschap van een stabiel schema is dat rekenfouten begrensd blijven.
Immers, stel dat $\bar{u}$ een oplossing is van

$$R\,\bar{u} = \bar{f}$$

met $\bar{f} = f + \delta$, waarin δ de verstoring is.

Dan geldt:

$$||u - \bar{u}|| = ||R^{-1} f - R^{-1} \bar{f}|| \leq ||R^{-1}||\ ||f - \bar{f}|| \leq ||R^{-1}||\ ||\delta||, \tag{2.14}$$

mits dus $||R^{-1}||$ maar niet al te groot is, blijkt dat $||u - \bar{u}||$ ook niet te groot wordt en van de orde $||\delta||$ is.

Opmerking: Sommige auteurs, zoals:

Richtmeyer, D.R., K.W. Morton, Difference Methods for initial-value problems, Interscience, New York (1967).

laten R als een analytische operator werken op functies U(x) en beschouwen dan $R^{-1} U - L^{-1} U$.

Het gevolg is een prachtige maar moeilijke theorie die min of meer voorbijgaat aan het gegeven feit dat de rekenautomaat de oplossing slechts in discrete punten uitrekent ; zodat in feite slechts $[U] - u$ van belang is.

2.1. Toepassing van de theorie voor "gewone" differentieschema's.

Het differentieschema (1.3) laat zich als (2.3) schrijven met

$$(Ru)_k = \begin{cases} u_0 \text{ als } k = 0 \\ \\ \dfrac{u_k - u_{k-1}}{\tau} + u_{k-1} \text{ als } k > 0, \end{cases}$$

en

$$f_k = \begin{cases} C \text{ als } k = 0 \\ \\ \bar{f}_{k-1} \text{ als } k > 0, \end{cases}$$

waarin, voor (1.3), $C = 1$ en $\bar{f}_{k-1} = F(x_{t_{k-1}})$.

R^{-1} is nu gemakkelijk te bepalen:

$$u_k = (R^{-1}f)_k = (1-\tau)^k f_0 + \tau \sum_{j=1}^{k} (1-\tau)^{k-j} f_j$$

Kiezen we als norm in E_f:

$$||f|| = \max_{j \geq 0} |f_j|$$

dan geldt:

$$|(R^{-1} f)_k| \leq (1-\tau)^k ||f|| + \tau \frac{1-(1-\tau)^k}{\tau} ||f|| \leq ||f||.$$

Kiezen we ook in E_u de maximum norm dan geldt:

$$||R^{-1} f|| \leq ||f||, \text{ dus } ||R^{-1}|| \leq 1,$$

zodat bij deze keuze van normen het schema stabiel is.

Om de consistentie na te gaan moet de discretiseringsfout berekend worden:

$$\Delta = R[U] - [LU] + [F] - f.$$

We merken op dat $\Delta \in E_f$, zodat Δ_0 en Δ_k, $k > 0$, aparte behandelingen vereisen:

$$\Delta_0 = R[U]_0 - [LU]_0 + [F]_0 - f_0,$$

$$= U(0) - [LU]_0 + [F]_0 - f_0,$$

$$\Delta_k = \frac{U(t_k) - U(t_{k-1})}{\tau} + U(t_{k-1}) - [LU]_k + [F]_k - f_k, \quad k > 0$$

De keuze van $|F|_k$, voor $k \geq 0$, is nog niet gemaakt, zodat deze geschikt gekozen kan worden.
Bovendien mogen we f_k nog geschikt kiezen.
Voor $k = 0$ kiezen we:

$$[F]_0 = F(t_0) = F(0),$$

zodat $$[LU]_0 = LU(0) = U(0).$$

Kiezen we nu $f_0 = F(0) = 1$ dan is $\Delta_0 = 0$.

(N.B. $F(0)$ is <u>niet</u> $\sin(0)$, maar <u>wel</u> het rechterlid van de begin conditie.)

Voor $k > 0$, kiezen we

$$[F]_k = F(t_{k-1}),$$

zodat $[L\bar{U}]_k = (\frac{dU}{dt} + U)\Big|_{t=t_{k-1}}$;

dan geldt (zie (1.4))

$$\Delta_k = \frac{\tau}{2} \frac{d^2U}{dt^2} (t_{k-1} + \theta_{k-1}\tau) + F(t_{k-1}) - f_k.$$

De keus $f_k = F(t_{k-1})$ ligt nu voor de hand zodat

$$\Delta_k = 0(\tau) \text{ voor } k > 0.$$

Uit $\lim_{\tau\to 0} ||\Delta||_\tau = \lim_{\tau\to 0} \max_{k\geq 0} |\Delta_{k,\tau}| = 0,$

volgt dan de consistentie van het differentieschema, waaruit weer zijn convergentie volgt.

N.B. Impliciet is boven aangenomen dat $\frac{d^2U}{dt^2}$ begrensd is; in het vervolg zullen we steeds aannemen dat de gezochte functie begrensde afgeleiden bezit.

<u>Opmerking</u>: Om niet steeds indexen k-1 te krijgen zullen we i.h.v. $(Ru)_{k+1}$ of $(Ru)_{k+1}$, etc. definieren.

De differentie operator R voor (1.55) is als volgt gedefinieerd:

$$(Ru)_0 = U_0$$
$$(Ru)_1 = U_1$$
$$(Ru)_{k+2} = \frac{1}{\tau}(-\tfrac{1}{2} u_{k+2} + 2 u_{k+1} - \frac{3}{2} u_k) + u_k, \ (k \geq 0),$$

en $f_0 = a$, $f_1 = b$, $f_{k+2} = d_k$ $(k \geq 0)$,

waarin a en b geschikt gekozen moeten worden, zeg $a = U(0)$ en $b = U(\tau)$ in welk geval we $u_1 = U(\tau)$ kiezen.

Nu geldt:

$$u_k = (R^{-1}u)_k = C_{1,k}u_{1,k} + C_{2,k}u_{2,k} \quad \text{(zie (1.24))}$$

met $u_{1,k} = (2-(1+2\tau)^{\frac{1}{2}})^k$ en $u_{2,k} = (2+(1+2\tau)^{\frac{1}{2}})^k$.

Uit (1.33) vinden we:

$$C_{1,k} = -\sum_{i=0}^{k-1} \frac{u_{2,i+1}\, d_i}{-(\frac{1}{2\tau})} \Big/ (u_{1,i+1}\, \Delta\, u_{2,i+1} - u_{2,i+1}\, \Delta_{1,i+1}) + C_{1,0}.$$

en een zelfde uitdrukking voor $C_{2,k}$.

Na enige analyse vinden we:

$$u_k = \alpha_1^{\,k} \{C_1 + \frac{2\tau}{\alpha_2-\alpha_1} \sum_{i=0}^{k-1} \alpha_1^{\,-(i+1)} f_{i+2}\}$$

$$+ \alpha_2^{\,k} \{C_2 + \frac{2\tau}{\alpha_1-\alpha_2} \sum_{i=0}^{k-1} \alpha_2^{\,-(i+1)} f_{i+2}\},$$

met $\alpha_1 = 2 - (1+2\tau)^{\frac{1}{2}}$ en $\alpha_2 = 2 + (1+2\tau)^{\frac{1}{2}}$.

Uit $u_0 = f_0$, $u_1 = f_1$ volgen

$$C_1 = \frac{f_1-\alpha_1 f_0}{\alpha_2-\alpha_1},$$

$$C_2 = \frac{f_1-\alpha_2 f_0}{\alpha_1-\alpha_2}.$$

Als we $f_0 = 0$, $f_1 = \delta$, $f_k = 0$, $k \geq 0$, $k \geq 2$, kiezen dan is

$$u_k = \frac{\delta}{\alpha_2-\alpha_1} (\alpha_1^{\,k} - \alpha_2^{\,k}).$$

Voor kleine τ geldt: $\alpha_2 - \alpha_1 \approx 2$, $\alpha_1^k \approx e^{-k\tau}$
en $\alpha_2^k \approx 3^k\, e^{k\tau/3}$.

Kiezen we nu een vaste T en k en τ zó dat $k_\tau = T$,
dan geldt

$$u_k \approx \frac{\delta}{2}\,(e^{-T} - 3^{T/\tau}\, e^{T/3}),$$

zodat voor deze f en de maximum norm

$$\lim_{\tau\to 0} \frac{||R^{-1} f||}{||f||} = \lim_{\tau\to 0} \tfrac{1}{2}(e^{-T} - 3^{T/\tau}\, e^{T/3}) = \infty$$

De conclusie is dat het schema instabiel is.

Opgave 2.1: Ga een en ander na voor het differentieschema (1.57) en constateer dat voor kleine τ:

$$||R^{-1}|| \approx 2e^{T},$$

zodat (1.57) stabiel genoemd mag worden.

Opgave 2.2: Ga de consistentie na van schema (1.57).
Opgave 2.3: Ga de stabiliteit en consistentie na van schema (1.59).

3. Differentieschema's voor partiële differentiaalvergelijkingen.

Drie belangrijke typen zijn:

a) Diffusievergelijking: $LU = F$ (3.1)

met

$$LU = \begin{cases} \frac{\partial U}{\partial t} - A(x,t)\,\frac{\partial}{\partial x}\left(B(x,t)\,\frac{\partial U}{\partial x}\right), & t > 0,\ 0 < x < 1, \\ U(x,0), & t = 0,\ 0 \leq x \leq 1, \\ \alpha_1(t)\,U(0,t) + \beta_1(t)\,\frac{\partial U}{\partial x}(0,t), & t > 0,\ x = 0, \\ \alpha_2(t)\,U(1,t) + \beta_2(t)\,\frac{\partial U}{\partial x}(1,t), & t > 0,\ x = 1. \end{cases} \tag{3.2}$$

$$F = \begin{cases} H(x,t), & t > 0,\ 0 < x < 1, \\ \Phi(x), & t = 0,\ 0 \leq x \leq 1, \\ \Psi_1(t), & t > 0,\ x = 0, \\ \Psi_2(t), & t > 0,\ x = 1. \end{cases} \tag{3.3}$$

b) De golfvergelijking, in eenvoudige gedaante geschreven,

$$\left.\begin{aligned} &\frac{\partial^2 U}{\partial t^2} - \frac{\partial^2 U}{\partial x^2} = H(x,t),\ t > 0,\ 0 < x < 1, \\ &U(x,0) = \Phi_1(x),\ U_t(x,0) = \Phi_2(x),\ 0 \leq x \leq 1, \\ &U(0,t) = \Psi_1(t),\ U(1,t) = \Psi_2(t),\ t > 0. \end{aligned}\right\} \tag{3.4}$$

c) De Laplace vergelijking:

$$\left.\begin{aligned} &\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2} = H(x,y),\ (x,y)\in \text{ gebied } G, \\ &\alpha(x,y)U + \beta(x,y)\frac{\partial U}{\partial n} = \Psi(x,y),\ (x,y)\in \Gamma = \text{rand van } G \end{aligned}\right\} \quad (3.5)$$

Een belangrijke generalisatie wordt verkregen als men punten x kiest in een gebied G van een n-dimensionale ruimte. Op de rand Γ van G worden dan randcondities voorgeschreven analoog aan (3.5).
De Laplace vergelijking zou men kunnen opvatten als een bijzonder geval van de diffusievergelijking:

$$\frac{\partial U}{\partial t} - \left(\frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 U}{\partial y^2}\right) = H(x,y) \quad (3.6)$$

immers de oplossing van (3.5) is de limiet voor $t\to\infty$, van de oplossing van (3.6).

In de x (of x,y) ruimte kiezen we een rooster met punten x_n, n = 0,,..., N, en maaswijdte h.
Op de eventuele t-as kiezen we de punten t_k met $t_k = k\tau$, $\tau > 0$.
Op het nu verkregen rooster, met punten (x_n,t_k) zoeken we de roosterfunctie u(n,k), of u_n^k zodanig dat u_n^k een benadering is van $U(x_n,t_k)$.
Om deze roosterfuncties te vinden stellen we weer een differentieschema Ru = f op.

3.1. Een expliciet differentieschema voor de diffusievergelijking.

Voor (3.1) kiezen we R als volgt:

$$
\left.
\begin{aligned}
(Ru)_n^0 &= u_n^0 \ , \ 0 \le n \le N, \\
(Ru)_n^{k+1} &= \frac{u_n^{k+1} - u_n^k}{\tau} - \frac{A_n^k}{h}\left(B_{n+\frac{1}{2}}^k \ \frac{u_{n+1}^k - u_n^k}{h} \right. \\
&\quad \left. - B_{n-\frac{1}{2}}^k \ \frac{u_n^k - u_{n-1}^k}{h} \right), \ k \ge 0, \ 0 < n < N, \\
(Ru)_0^{k+1} &= \alpha_1^{k+1} \ u_0^{k+1} + \beta_1^{k+1} \ \frac{u_1^{k+1} - u_0^{k+1}}{h} \ , \ k \ge 0 \\
(Ru)_N^{k+1} &= \alpha_2^{k+1} \ u_N^{k+1} + \beta_2^{k+1} \ \frac{u_N^{k+1} - u_{N-1}^{k+1}}{h} \ , \ k \ge 0,
\end{aligned}
\right\} \qquad (3.7)
$$

en voor f kiezen we

$$
\left.
\begin{aligned}
f_n^0 &= \Phi_n, \ 0 \le n \le N, \\
f_n^{k+1} &= H_n^k, \ k \ge 0, \ 0 < n < N, \\
f_0^{k+1} &= \Psi_1^{k+1}, \ f_N^{k+1} = \Psi_2^{k+1}, \ k \ge 0.
\end{aligned}
\right\} \qquad (3.8)
$$

Hierin zij $A_n^k, \ldots, \Psi_2^{k+1}$ afkortingen voor $A(t_k, x_n), \ldots, \Psi_2(t_{k+1})$.

Met behulp van (3.7) en (3.8) is een algorithme voor de berekening van u_n^k gemakkelijk opgesteld.
Immers u_n^0 is bekend en u_n^{k+1} kan berekend worden m.b.v. u_n^k, u_{n-1}^k en u_{n+1}^k, dus voor $n = 1, \ldots, N-1$; terwijl u_0^{k+1} en u_n^{k+1} uit de randvoorwaarden zijn te berekenen.

Het doet er in dit geval niet toe hoe ingewikkeld de functies $A, \ldots, \Psi_2$ zijn.
Dit staat in tegenstelling tot analytische methoden om U te bepalen; daarin mogen we voor $A, \ldots, \Psi_2$ slechts zeer eenvoudige functies kiezen.

Opgave 3.1: Stel een ALGOL 60 programma op dat bij gegeven functies A, ..., Ψ_2, de u_n^k berekent voor $0 \leq k \leq K$, $0 \leq n \leq N$.

Bij de vraag of het schema consistent is, nemen we de willekeurige functies A, ..., Ψ_2 mee; dit is helaas niet meer mogelijk bij de bestudering van de stabiliteit.

De discretiseringsfout Δ vinden we uit

$$\Delta_n^0 = (Ru)_n^0 - [LU]_n^0 + [F]_n^0 - f_n^0$$

$$= u_n^0 - U(x_n,0) + \Phi(x_n) - \Phi_n = 0,$$

(waar we $[F]_n^0 = F(x_n,0)$ gekozen hebben),

$$\Delta_n^{k+1} = U_t(x_n,t_k) + \frac{\tau}{2} U_{tt}(x_n,t_k + \Theta_{k,n}\tau)$$

$$- A(x_n,t_k) \{ \frac{\partial}{\partial x} (B \frac{\partial U}{\partial x})_{x_n,t_k}$$

$$+ \frac{h^2}{24} \left[\frac{\partial^3}{\partial x^3} (B \frac{\partial U}{\partial x}) \Big|_{\bar{x}_n,t_k} + B^k_{n+\frac{1}{2}} \frac{\partial^3 U}{\partial x^3} \Big|_{\bar{\bar{x}}_n,t_k} - \right.$$

$$\left. - B^k_{n-\frac{1}{2}} \frac{\partial^3 U}{\partial x^3} \Big|_{\bar{\bar{x}}_n,t_k} \right] \}$$

$$- \{U_t - A \frac{\partial}{\partial x} (B \frac{\partial U}{\partial x})\}_{x_n,t_k}$$

$$+ H(x_n,t_k) - H_n^k$$

$$= O(\tau) + O(h^2), \quad k \geq 0, \quad 0 < n < N,$$

$$\Delta_0^{k+1} = \alpha_1^{k+1}\, U(0,t_{k+1}) + \beta_1^{k+1} \{ \frac{\partial U}{\partial x}\Big|_{0,t_{k+1}} + \frac{h}{2} \frac{\partial^2 U}{\partial x^2} (\Theta h, t_{k+1})\}$$

$$- \{\alpha_1^{k+1}\, U(0,t_{k+1}) + \beta_1^{k+1} \frac{\partial U}{\partial x} (0,t_{k+1})\} + \psi_1^{k+1} - \psi_1^{k+1}$$

$$= o(h),$$

$$\Delta_N^{k+1} = o(h).$$

Opmerking: $[F]_n^k = F(x_n,t_k)$ gekozen.

Conclusie: Het differentieschema is consistent, want er geldt:

$$\Delta_n^k = O(\tau) + O(h), \tag{3.9}$$

Als de maaswijdte = max (τ,h) naar nul gaat, gaat $||\Delta||$ naar nul.

Het is jammer dat de discretisatiefout niet $O(h^2) + O(\tau)$ is hetgeen veroorzaakt wordt door de discretisatie van de randvoorwaarden.
We zien dat
de term $\beta_1^{k+1} \frac{h}{2} \frac{\partial^2 U}{\partial x^2} (\Theta h, t_{k+1})$ dan weggewerkt moet worden.

Nu geldt $\frac{\partial U}{\partial t} - AB \frac{\partial^2 U}{\partial x^2} - A \frac{\partial U}{\partial x} \frac{\partial B}{\partial x} = H$, zodat

$$\frac{\partial^2 U}{\partial x^2} = \frac{1}{AB} \{ \frac{\partial U}{\partial t} - A \frac{\partial U}{\partial x} \frac{\partial B}{\partial x} - H\}.$$

Vervangen we nu $(Ru)_0^{k+1}$ door:

$$(R_1 u)_0^{k+1} = \alpha_1^{k+1}\, u_0^{k+1} + \beta_1^{k+1} \frac{u_1^{k+1} - u_0^{k+1}}{h}$$

$$- \frac{h}{2} \beta_1^k \frac{1}{A_0^k B_0^k} \{ \frac{u_0^{k+1} - u_0^k}{\tau} - A_0^k \left(\frac{\partial B}{\partial x} \right)_0^k \frac{u_1^k - u_0^k}{h} \}, \tag{3.10}$$

dan is

$$\Delta_0^{k+1} = O(h^2) + O(h\tau) - \frac{H_0^k}{A^k{}_0\, B^k{}_0}\ \frac{h}{2}\,\beta_1^k + \psi_1^{k+1} - f_0^{k+1}.$$

Kiezen we $f_0^{k+1} =$

$$f_0^{k+1} = \psi_1^{k+1} - \frac{H_0^k}{A_0{}^k\, B_0{}^k}\ \frac{h}{2}\,\beta_1^k, \tag{3.11}$$

en behandelen we Δ_N^{k+1} analoog dan ontstaat

$$\Delta_n^k = O(\tau) + O(h^2). \tag{3.12}$$

3.1.1. Stabiliteit van het expliciete differentieschema voor de diffusievergelijking met een randvoorwaarde van de 1e soort.

Voor de bestudering van de stabiliteit laten we voorlopig alle algemeenheid varen en stellen $A = B = 1$, $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 0$. In dit geval kunnen we de volgende algorithme opstellen:

$$u^0{}_n = f^0{}_n$$

$$u_n^{k+1} = \frac{\tau}{h^2}\ u_{n-1}^k + (1-2\frac{\tau}{h^2})\ u_n^k + \frac{\tau}{h^2}\ u_{n+1}^k + f_n^{k+1},\ 0 < n < N,$$

$$u_0^{k+1} = f_0^{k+1};\ u_N^{k+1} = f_N^{k+1}. \tag{3.13}$$

Vatten we u_n^k als de componenten van de vector u^k op, dan is (3.13) te schrijven als:

$$u^{k+1} = A\,u^k + \tau f^{k+1},\ k \geq 0, \tag{3.14}$$

met, als $r = \tau/h^2$:

$$A = \begin{pmatrix} 0 & 0 & 0 & & & \\ r & 1-2r & r & & 0 & \\ & r & 1-2r & r & & \\ & & - \; - & - & & \\ & & & r & 1-2r & r \\ 0 & & & 0 & 0 & 0 \end{pmatrix},$$

en $f^k = (f_0^k/\tau,\ f_1^k,\ \ldots,\ f_{N-1}^k,\ f_N^k/\tau)^{Tr}$.

De oplossing van (3.14) is eenvoudig aan te geven:

$$u^k = A^k\, f^0 + \tau \sum_{j=1}^{k} A^{k-1}\, \rho^j. \tag{3.15}$$

Om nu $||R^{-1}||$ te vinden moeten we $||u||$ bepalen.

We beschouwen een eindig t interval $0 \leq t \leq T$ en we kiezen $\tau = T/K$, met een geheel getal.
We spreken af dat

$$\begin{aligned} ||u|| &= \max_{0 \leq k \leq K} ||u^k||, \\ \text{en} \qquad ||f|| &= \max_{0 \leq k \leq K} ||f^k||. \end{aligned} \tag{3.16}$$

Zij $||A||$ de norm van de matrix A dan volgt uit (3.15)

$$||U|| = ||R^{-1}\, f|| \leq \max_{0 \leq k \leq K} \{ ||A||^k\, ||f^0|| + \tau \sum_{j=1}^{k} ||A||^{k-j}\, ||\rho^j|| \}$$

Nu is $||\rho^j|| = ||(\frac{f_0^j}{\tau},\ f_1^j,\ \ldots,\ f_{N-1}^j,\ \frac{f_N^j}{\tau})^{Tr}||$, zodat met

$$||f^k|| = \max\left(\frac{|f_0^k|}{\tau},\ \max_{0 < n < N} |f_n^k|,\ \frac{|f_N^k|}{\tau}\right),\ k > 0,$$

en $||f^0|| = \max_{0 \leq n \leq N} |f_n^0|$, geldt:

$$\|u\| \leq \max_{0 \leq k \leq K} \{\|A\|^k + \tau \frac{\|A\|^k - 1}{\|A\| - 1}\} \|f\| .$$

Hieruit volgt de belangrijke

Stelling 3.1: Voor stabiliteit van schema (3.15) is het voldoende als geldt:

$$\|A\| \leq 1 + O(\tau). \tag{3.17}$$

Immers, dan geldt, als

$$\|A\| = 1 + C_1 \tau + O(\tau^2),$$
$$\|u\| \leq (1+\tau)e^{C_1 T} \|f\| .$$

Bij de bovengedane keuze van $\|f\|$ hoort die van $\|\Delta\|$.

Aangezien $\Delta_0^k = \Delta_N^k = 0$ volgt $\|\Delta\| = O(\tau) + O(h^2)$.
Bovendien kunnen we $\|A\|$ nu berekenen aangezien we kennelijk met de maximum norm te doen hebben.

Als $v = Aw$ dan geldt: $v_j = \sum_{n=0}^{N} A_{jn} w_n$

$$\text{dus} \quad \max_{0 \leq j \leq N} |v_j| \leq \max_{0 \leq j \leq N} \sum_{n=0}^{N} |A_{jn}| \max_{0 \leq n \leq N} |w_n|$$

$$\text{of} \quad \|v\| \leq \max_{0 \leq j \leq N} \sum_{n=0}^{N} |A_{jn}| \, \|w\|$$

$$\text{zodat} \quad \|A\| = \max_{0 \leq j \leq N} \sum_{n=0}^{N} |A_{jn}| .$$

Als nu $r = \tau/h^2 \leq \frac{1}{2}$ dan zijn alle elementen van de matrix niet-negatief en $\|A\| = (r + 1-2r + r) = 1$.
Er is dus stabiliteit als $\tau/h^2 \leq \frac{1}{2}$. In het vervolg kiezen we $r = \tau/h^2$ vast, zodat de maaswijdte $\tau = rh^2$.

Opgave 3.2: Ga m.b.v. de normkeuze van $||f||$ de consistentie na als $r \leq \frac{1}{2}$ en bewijs de convergentie.

De vraag rijst nu of $r \leq \frac{1}{2}$ een noodzakelijke voorwaarde is. Met twee methoden: een "ad hoc" methode en een algemene methode zullen we de vraag bevestigend beantwoorden.
Eerst de "ad hoc" methode.
We verwaarlozen de randcondities en doen alsof een beginfunctie $U(x,0)$ gegeven is voor $-\infty \leq x \leq +\infty$.
Voor de roosterfunctie u betekent dit dat we het volgende probleem willen oplossen:

$$\begin{aligned} u_n^0 &= f_n^0, \; n = \ldots -2, -1, 0, 1, 2, \ldots \\ u_n^{k+1} &= r\, u_{n-1}^k + (1-2r)u_n^k + r\, u_{n+1}^k, \; k \geq 0 \end{aligned} \tag{3.18}$$

We merken op dat als $r \leq \frac{1}{2}$ dan geldt:

$$|u_n^{k+1}| \leq \{r + (1-2r) + r\} \max \{|u_{n-1}^k|, |u_n^k| \; u_{n+1}^k|\}$$

dus $$\max_n |u_n^{k+1}| \leq \max_n |u_n^k|.$$

Stel nu $r > \frac{1}{2}$, dus $1-2r < 0$.

Neem $f_n^0 = 0$ voor $n \neq 0$ en $f_0^0 = \delta > 0$; we kunnen ons voorstellen dat er éénmalig een rekenfout δ wordt gemaakt.
De eerste bewering is dat de tekens van u_n^k alterneren zowel in de boven index, als in de beneden index (waarbij we afspreken dat 0 zowel een positief als een negatief teken heeft, zodat (0,0) alterneert).
Voor $k = 0$ is dit inderdaad het geval: immers $\ldots, +0, -0, +\delta, -0, +\delta, \ldots$ alterneert; bovendien geldt:

$$u_{-1}^1 = r\,\delta > 0, \; u_0^1 = (1-2r)\,\delta < 0 \text{ en } u_1^1 = r\,\delta > 0$$

Als de tekens van $u^k_{-k}, u^k_{-k+1}, \ldots, u^k_0, \ldots, u^k_{k-1}, u^k_k$ alterneren dan volgt:
als u^k_n negatief is, en dus u^k_{n-1} en u^k_{n+1} niet negatief, dat u^{k+1}_n positief is; zodat het alterneren zich inderdaad herhaalt.

De tweede bewering is:

$$\max_n |u^{k+1}_n| \geq (4r-1)^k \frac{\delta}{2k+1},$$

welke meteen volgt als we opschrijven

$$\sum_{j=-k-1}^{k+1} |u^{k+1}_j| = u^{k+1}_{-k-1} - u^{k+1}_{-k} + \ldots - u^{k+1}_k + u^{k+1}_{k+1}$$

$$= r\,u^k_{-k} - (1-2r)\,u^k_{-k} - r\,u^k_{-k+1}$$

$$+ r\,u^k_{-k} + (1-2r)\,u^k_{-k+1} + r\,u^k_{-k+2}$$

$$- r\,u^k_{-k+1} - (1-2r)\,u^k_{-k+1} - r\,u^k_{-k+3}$$

$$+ \ldots$$

$$= (4r-1)\{u^k_{-k} - u^k_{-k+1} + \ldots - u^k_{k-1} + u^k_{-k}\}$$

$$= (4r-1) \sum_{j=-k}^{k} |u^k_j|$$

Dus $$\max_n |u^{k+1}_n| \geq \frac{1}{2k+1} \sum_{j=-k-1}^{k+1} |u^{k+1}_j| = (4r-1)^k\,\delta.$$

Dit betekent dat, hoe klein δ ook is, dus hoe klein $||f||$ ook is, er altijd wel een k te vinden is waarvoor $||u||$ groter is dan een willekeurig groot gekozen C.
M.a.w. het schema is instabiel.

De tweede methode is algemener, maar bewerkelijker.
We gaan de <u>eigenwaarden</u> en <u>eigenvectoren</u> van de matrix A bestuderen.
We willen aantonen dat het differentieschema instabiel is voor $r > \frac{1}{2}$; het is dus voldoende als we aantonen dat bij willekeurig groot gekozen C er bij elke willekeurig kleine τ_0 een $\tau < \tau_0$ te vinden is zó dat daar weer een f bij te vinden is zódat

$$||R_\tau^{-1} f|| > C ||f|| . \tag{3.19}$$

Of $\quad \forall_C \forall_{\tau_0} \exists_\tau \exists_f, \tau < \tau_0 \Rightarrow (3.19).$

Het is dus geen beperking van de algemeenheid als we ons beperken tot roosterfuncties f, waarvoor geldt $f_n^k = 0$, voor $k > 0$, $0 \leq n \leq N$.
Dit betekent dat we slechts beginfuncties f^0 toelaten.
In het probleem (3.14) mogen we dan $f^k = 0$ stellen,
zodat

$$u^k = A^k f^0. \tag{3.20}$$

Als A de eigenvectoren e_i, met eigenwaarden λ_i, $i=0, \ldots, N$, heeft en er bestaat een λ_m zó dat $|\lambda_m| > 1 + 0(\tau)$, (d.w.z. als

$$\lambda_m = \sum_{i=-l}^{\infty} C_i \tau^i, \text{ en } C_{-l} \neq 0,$$

dan geldt óf $l < 0$ of $l=0$ èn $|C_0| > 1$),
dan kiezen we $f^0 = e_m$, zodat

$$u^k = \lambda_m{}^k e_m.$$

Dan geldt $||u^k|| = |\lambda_m|^k ||e_m||$, zodat
bij gegeven T en $\tau = T/K$, K geheel,

$$||u|| = \max_{0 \leq k \leq K} ||u^k|| = |\lambda_m|^K ||f|| .$$

Onder deze aannamen is het niet moeilijk in te zien dat we bij een willekeurige C een τ en een f kunnen vinden zó dat (3.19) geldt, daar immers $|\lambda_m|^k$ willekeurig groot te krijgen is.

Het probleem de instabiliteit te bewijzen is nu verschoven naar het probleem de eigenwaarden van A te bestuderen.

Het resultaat van bovengedane beschouwingen geven we nu in de volgende

Stelling 3.2: Zij het differentieschema $R_u = f$ te schrijven in de vorm

$$\left.\begin{aligned} u^0 &= f^0, \\ u^{k+1} &= A\,u^k + \tau\rho^k,\ k \geq 0, \end{aligned}\right\} \qquad (3.21)$$

met A een (n+1) bij (n+1) matrix en $\rho = Bf$, dan is R instabiel als voor tenminste één van de eigenwaarden λ van A geldt:

$$|\lambda| > 1 + O(\tau).$$

Of:

Stelling 3.3: Een noodzakelijke voorwaarde voor stabiliteit is dat alle eigenwaarden λ_i van A voldoen aan $|\lambda_i| \leq 1 + O(\tau)$.

Het criterium in bovengenoemde stelling wordt het Von Neumann criterium genoemd.

De bepaling van de eigenwaarden gaat als volgt:
Een eigenwaarde λ en een eigen functie e voldoen aan

$$Ae = \lambda e, \qquad (3.22)$$

of uitgeschreven:

$$\sum_{j=0}^{N} (A_{n_j} - \delta_{n_j}\,\lambda)\, e_j = 0,\ n = 0, \ldots, N. \qquad (3.23)$$

Aangezien A een band matrix is, waarvoor geldt dat alleen op de hoofddiagonaal en op twee aanliggende diagonalen elementen ongelijk nul voorkomen, kunnen we (3.23) als volgt schrijven:

$$\left.\begin{aligned} (A_{00} - \lambda)\, e_0 + A_{11}\, e_1 &= 0 \\ A_{nn-1}\, e_{n-1} + (A_{nn} - \lambda)\, e_n + A_{nn+1}\, e_{n+1} &= 0,\ n = 1, \ldots, N-1 \\ A_{NN-1}\, e_{N-1} + (A_{NN} - \lambda)\, e_N &= 0 \end{aligned}\right\} \quad (3.24)$$

Voor deze matrix geldt bovendien dat

$$\begin{aligned} A_{n,n-1} &= a \quad \text{(onafhankelijk van n)} \\ A_{nn} - \lambda &= -2b \quad \text{(onafhankelijk van n)} \\ A_{n,n+1} &= c \quad \text{(onafhankelijk van n)} \end{aligned} \quad (3.25)$$

Zodat het volgende differentieschema ontstaat:

$$a\, e_{n-1} - 2\, b\, e_n + c\, e_{n+1} = 0, \quad (3.26)$$

waarvan we de algemene oplossing kennen:

$$\begin{aligned} &e_n = \alpha^n \sin(n\phi + \psi), \text{ met} \\ &\alpha = (a/c)^{\frac{1}{2}} \text{ en } \phi = \operatorname{arctg}\{(ac - b^2)^{\frac{1}{2}}/b\}. \end{aligned} \quad (3.27)$$

Om ϕ te vinden, maken we gebruik van:

$$(A_{00} - \lambda) \sin\psi + A_{11}\, \alpha \sin(\phi + \psi) = 0,$$

en (3.28)

$$A_{N,N-1} \sin((N-1)\phi + \psi) + (A_{N,N} - \lambda)\, \alpha \sin(N\phi + \psi) = 0.$$

In ons speciale geval geldt:

$a = r$, $b = (\lambda - 1 + 2r)/2$, $c = r$, $A_{00} = A_{11} = A_{N,N-1} = A_{N,N} = 0$,

zodat

$$\alpha = 1 \text{ en } \phi = \operatorname{arctg} \frac{(4(1-\lambda)r - (1-\lambda)^2)^{\frac{1}{2}}}{\lambda - 1 + 2r},$$

en

$$-\lambda \sin \psi = 0$$
$$-\lambda \sin (N\phi + \psi) = 0.$$

Hieruit volgt, afgezien van $\lambda = 0$, dat $\psi = 0$ en $\phi_i = \frac{i\,\pi}{N}$, $i = 0, 1, 2, \ldots$.
Dus

$$\frac{(2(1-\lambda)r - (1-\lambda)^2)^{\frac{1}{2}}}{\lambda - 1 + 2r} = \operatorname{tg} \frac{i\,\pi}{N},$$

of

$$\lambda_{1i} = 1 - 4r \sin^2 \frac{i\,\pi/2}{N},$$

en

$$\lambda_{2i} = 1 - 4r \cos^2 \frac{i\,\pi/2}{N}.$$

Aangezien echter $\sin^2 \frac{i\,\pi/2}{N} = \cos^2 \frac{(N-i)\,\pi/2}{N}$ volgt dat de eigenwaarden zijn

$$\lambda_i = 1 - 4r \sin^2 \frac{i\,\pi/2}{N}, \; i = 0, 1, 2, \ldots, N. \quad (3.29)$$

en de bijbehorende eigenfuncties zijn:

$$e_n^i = \sin(n \frac{i\,\pi}{N}), \; i = 0, \ldots, N. \quad (3.30)$$

We merken op dat twee eigenfuncties e^0 en e^N degenereren, hetgeen niet zo verwonderlijk is.

Als nu $r > \frac{1}{2}$, dan kiezen we e^{N-1} met λ_{N-1}.

Als N maar groot genoeg is volgt:

$$|\lambda_{N-1}| = 4r \sin^2 \frac{(N-1)\,\pi/2}{N} - 1$$

$$= 4r \cos^2 \frac{\pi/2}{N} - 1 = 4r - 1 - 4r \sin^2 \frac{\pi/2}{N}$$

Nu is $Nh = 1$ en $h = \sqrt{\tau/r}$, dus

$$|\lambda_{N-1}| = 4r - 1 - r\,\pi^2 \frac{\tau}{r} + 0(\tau^2).$$

Deze λ voldoet precies aan $|\lambda| > 1 + 0(\tau)$, zodat het schema R met $r > \frac{1}{2}$ instabiel is.
Opmerking: de bijbehorende eigen functie e^{N-1} is ongelijk aan nul.

3.1.2. Stabiliteit van het expliciete differentieschema voor de differsievergelijking met een randvoorwaarde van de tweede soort.

We beschouwen weer het schema (3.7), maar nu met $A = B = 1$, $\alpha_1 = 0$, $\beta_1 = 1$, $\alpha_2 = 1$, $\beta_2 = 0$.
De randvoorwaarde bij $x = 0$ kunnen we nu als volgt discretiseren:

$$\frac{1}{h}(u_1^{k+1} - u_0^{k+1}) = f_0^{k+1}, \tag{3.31}$$

Om er een schema van de vorm (3.14) van te krijgen schrijven we (3.31) als:

$$u_0^{k+1} = u_1^{k+1} - h\, f_0^{k+1}$$

$$= r\, u_0^k + (1-2r)\, u_1^k + r\, u_2^k + \tau\, f_1^{k+1} - h\, f_0^{k+1}.$$

De matrix A krijgt nu de gedaante:

$$A = \begin{pmatrix} r & 1-2r & r & 0 & & \\ r & 1-2r & r & 0 & & \\ 0 & r & 1-2r & r & & \\ & & \cdot & \cdot & \cdot & \\ & & & r & 1-2r & r \\ & & & 0 & 0 & 0 \end{pmatrix}, \tag{3.32}$$

en ρ^k is nu:

$$\rho^k = (f_1^{k+1} - \frac{h}{\tau} f_0^{k+1}, f_1^{k+1}, \ldots, f_{N-1}^{k+1}, f_N^{k+1}/\tau)^{Tr}. \quad (3.33)$$

Weer geldt, als $r \leq \frac{1}{2}$, dat

$$||A|| = \max_{0 \leq i \leq N} \sum_{j=0}^{N} |A_{ij}| = 1,$$

zodat het schema stabiel is voor $r \leq \frac{1}{2}$, waarbij $||f||$, is gegeven door:

$$||f|| = \max_k \{\max \{|f_1^k - \frac{h}{\tau} f_0^k|, \max_{0 < n < N} |f_n^k|, |f_N^k/\tau|\}\}.$$

Opgave 3.3: Toon aan dat het schema niet consistent is, bij bovenvermelde normkeuze van $||f||$.

(Moeilijke) Opgave 3.4: Toon aan dat het schema met $A = B = 1$, $\alpha_1 = \beta_2 = 0$, $\beta_1 = \alpha_2 = 1$, onder gebruikmaking van de randoperator gedefinieerd in (3.10), stabiel en consistent is als $r \leq \frac{1}{2}$, onder de bovengebruikte maximum normen, voor A en ρ.

Om voor $r > \frac{1}{2}$ de instabiliteit aan te tonen kunnen we de eigenwaarden van A bestuderen. Dit is echter niet meer zo eenvoudig als voorheen omdat we inplaats van (3.28) de volgende vergelijkingen hebben:

$$(r - \lambda) \sin \psi + (1 - 2r) \sin (\phi + \psi) + r \sin (2\phi + \psi) = 0 \quad (3.34)$$

en

$$- \lambda \sin (N\phi + \psi) = 0. \quad (3.35)$$

Geschreven als vergelijkingen in sin ψ en cos ψ, moet er een niet-triviale oplossing voor sin ψ en cos ψ zijn, d.w.z.

$$(r - \lambda + (1 - 2r))\cos \phi + r \cos 2\psi \sin N\phi =$$

$$((1 - 2r) \sin \phi + r \sin 2\psi)\cos N\phi;$$

Met $\phi = \text{arctg}\ \dfrac{\{4(1-\lambda)r - (1-(1-\lambda)^2\}^{\frac{1}{2}}}{\lambda - 1 + 2r}$ (3.36)

geeft dit een onhanteerbare vergelijking in λ.

We memoreren dat het aantonen van instabiliteit betekent dat we bij elke τ, kleiner dan een zekere τ_0, een f_τ aan kunnen wijzen waarvoor geldt dat $||R_\tau^{-1} f_\tau|| \ / \ ||f_\tau||$ onbegrensd groot wordt voor $\tau \to 0$.
We kunnen ons daarom beperken tot een deelklasse van alle roosterfuncties f. Als we in deze deelklasse bovengevraagde f_τ's aan kunnen geven dan is het doel immers bereikt.
We zullen nu de deelklasse beschouwen die bestaat uit die roosterfuncties f die voldoen aan:

a) $f_0^{k+1} = 0$, dit betekent $u_0^{k+1} = u_1^{k+1}$

b) $\rho^{k+1} = 0$, of $f_n^{k+1} = 0$, $n = 0, \ldots, N$, $k \geq 0$.

zodat $u_N^{k+1} = 0$.

Het eigenwaardeprobleem stellen we nu als volgt:

Gevraagd een roosterfunctie e met $e_0 = e_1$ en $e_N = 0$
zó dat voor zekere $\lambda \neq 0$:

$$r\, e_{n-1} + (1-2r)e_n + r\, e_{nn} = \lambda\, e_n,\ 0 < n < N. \qquad (3.37)$$

Opgave 3.5: Formuleer het bovenstaande eigenwaardeprobleem als een eigenwaarde-eigenvectorprobleem van een zekere matrix A^*.

M.b.v. (3.27) weten we dat $e_n = \sin(n\phi + \psi)$
met ϕ als in (3.36)

Uit $e_0 = e_1$ volgt: $\sin\psi = \sin(\phi + \psi)$; (3.38)

Uit $e_N = 0$ volgt: $\sin(N\phi + \psi) = 0$. (3.39)

De conditie (3.38) betekent tg ψ = tg $(\frac{\pi}{2} - \frac{\phi}{2})$;

De conditie (3.39) betekent tg ψ = tg $(-N\phi)$.

Voor ϕ vinden we de volgende waarden die hieraan voldoen:

$$\phi\, i = \frac{2i - 1}{2N - 1}\,\pi, \; i = 0, 1, 2, \ldots. \tag{3.40}$$

We kunnen $\psi = -N\phi$ stellen aangezien voor de oplossing geldt:

$$e_n^{(i)} = \sin((n - N)\,\phi_i + k\pi) = \sin(n - N)\phi_i.$$

We zien bovendien dat

$$e_n^{(i)} = e_n^{(i + 2N - 1)}$$

en

$$e_n^{(i)} = e_n^{(2N - i)},$$

zodat alleen de ϕ_i's van belang zijn voor i = 1, ..., N - 1.

Voor deze ϕ_i's zijn de eigenwaarden gegeven door

$$\lambda_i = 1 - 4r \sin^2 \frac{\phi_i}{2}.$$

Aangezien voor N voldoende groot, d.w.z. h voldoende klein, d.w.z. $\tau = r\,h^2$ voldoende klein, $\phi_{N-1} \approx \pi$ volgt dat $\lambda_{N-1} < -1$, mits $r > \frac{1}{2}$.

Er is dus een roosterfunctie $e_n = \sin(n - N)\,\frac{2N - 3}{2N - 1}\,\pi$
waarvoor geldt dat na toepassing van het differentieschema het resultaat λ e is, welke weer voldoet aan de condities $u_0 = u_1$ en $u_N = 0$.
Kiezen we bij deze e een passende f,
d.w.z.

$$f_n^0 = e_n, \; f_n^{k+1} = 0, \; k \geq 0;$$

zij nu T de bovengrens van het t interval en $K = T/\tau$, dan is

$$\frac{\|R^{-1} f\|}{\|f\|} = |\lambda_{N-1}|^K$$

hetgeen voor $K \to \infty$ onbegrensd toeneemt, waaruit de instabiliteit volgt.

Opgave 3.6: Onderzoek de stabiliteit van het schema (3.7)
met $A = B = 1$, $\alpha_1 = \alpha_2 = 0$, $\beta_1 = \beta_2 = 1$, en
met $A = B = 1$, $\alpha_1 = 1$, $\beta_1 = 0$, $\alpha_2 = 0$, $\beta_2 = 1$.

3.2. Een impliciet schema voor de diffusievergelijking.

Voor het probleem (3.1) met $A = B = 1$, $\alpha_1 = \alpha_2 = 1$, $\beta_1 = \beta_2 = 0$, kunnen we ook het volgende differentieschema opstellen:

$$\left.\begin{aligned}
&(Ru)_n^0 = u_n^0, \quad 0 \le n \le N. \\
&(Ru)_n^{k+1} = \frac{u_n^{k+1} - u_n^k}{\tau} - \frac{\theta}{h^2}\left(u_{n-1}^{k+1} - 2u_n^{k+1} + u_{n+1}^{k+1}\right) \\
&\qquad - \frac{1-\theta}{h^2}\left(u_{n-1}^{k} - 2u_n^{k+1} + u_{n+1}^{k}\right), \quad k \ge 0, \ 0 < n < N, \\
&(Ru)_0^{k+1} = u_0^{k+1}, \quad k \ge 0, \\
&(Ru)_N^{k+1} = u_N^{k+1}, \quad k \ge 0.
\end{aligned}\right\} \qquad (3.41)$$

Voor $\theta \neq 0$ blijkt het niet meteen duidelijk te zijn welk algorithme gekozen moet worden om $u = R^{-1} f$ te berekenen; immers u_n^{k+1} kan niet meer eenvoudig in de u_i^k's uitgedrukt worden. We zullen echter zien dat voor bovenstaand schema toch een betrekkelijk eenvoudig algorithme is op te stellen.
Het belangrijke voordeel van schema (3.41) t.o.v. schema (3.7) is dat het veel stabieler is.
In feite is de stabiliteitsvoorwaarde $r \le \frac{1}{2}$ voor schema (3.7) vaak onrealistisch stringent (bij halvering van de tijdstap moet de plaatsstap tweemaal gehalveerd worden zodat we acht maal zoveel rekenwerk moeten verrichten).

De stabiliteitsvoorwaarde voor (3.41) blijkt te zijn:

$$\text{als } \theta \leq \tfrac{1}{2} : r \leq 1/(2 - 4\theta),$$
$$\text{als } \theta > \tfrac{1}{2} : r \text{ onbeperkt (het schema heet onvoorwaardelijk stabiel)} \qquad (3.42)$$

Voor $\theta = \frac{1}{2}$ is er bovendien nog een voordeel met betrekking tot de discretisatiefout die we nu $0(\tau^2 + h^2)$ kunnen krijgen door f geschikt te kiezen. (Dit levert het schema van Crank-Nicholson).

Opgave 3.7: Bereken de discretisatiefout van (3.41) bij geschikte keuze van f.

Het schema (3.41) is te schrijven als

$$B\, u^{k+1} = A\, u^k + \tau f^{k+1}, \qquad (3.43)$$

met

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & & \\ -\theta r & 1+2\theta r & -\theta r & 0 & & \\ 0 & -\theta r & 1+2\theta r & \theta r & & \\ - & -- & --- & - & & \\ & & -\theta r & 1+2\theta r & -\theta r & \\ & & 0 & 0 & 1 & \end{pmatrix},$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & & \\ (1-\theta)r & 1-2(1-\theta)r & (1-\theta)r & 0 & & \\ 0 & (1-\theta)r & 1-2(1-\theta)r & (1-\theta)r & & \\ - & -- & --- & - & & \\ & & (1-\theta)r & 1-2(1-\theta)r & (1-\theta)r & \\ & & 0 & 0 & 0 & \end{pmatrix},$$

en

$$\rho^{k+1} = (f_0^{k+1}/\tau,\ f_1^{k+1},\ \ldots,\ f_{N-1}^{k+1},\ f_N^{k+1}/\tau)^{Tr}$$

Aannemende dat het rechterlid van (3.43) bekend is, hetgeen voor $k = 0$ het geval is, kunnen we u^{k+1} berekenen als B^{-1} bekend is.

3.2.1. De "double-sweep" methode.

Zij gevraagd te berekenen de roosterfunctie u_i die voldoet aan:

$$\left.\begin{aligned} & u_0 = \phi \\ & a_i\, u_{i-1} - 2\, b_i u_i + c_i u_{i+1} = g_i,\ i = 1, \ldots, N-1, \\ & u_N = \psi. \end{aligned}\right\} \quad (3.44)$$

We stellen $u_i = k_i + l_i\, u_{i+1}$, en trachten recurrente relaties voor k_i en l_i te vinden.
Uit $u_0 = \phi$ volgt $k_0 = \phi$, $l_0 = 0$.
Substitutie van u_{i-1} in (3.44) geeft:

$$a_i(k_{i-1} + l_{i-1}\, u_i) - 2\, b_i u_i + c_i u_{i+1} = g_i,\ i = 1, \ldots, N-1,$$

dus

$$u_i = \frac{a_i k_{1-i} - g_i}{2\, b_i - a_i l_{i-1}} + \frac{c_i}{2\, b_i - a_i l_{i-1}}\, u_{i+1}$$

zodat $\quad k_i = (a_i\, k_{i-1} - g_i)/(2\, b_i - a_i l_{i-1})$

en $\quad l_i = c_i/(2\, b_i - a_i l_{i-1})$

We berekenen nu eerst de k_i en l_i's (1e sweep),
vervolgens berekenen we de u_i's (2e sweep).

Opgave 3.8: Construeer een ALGOL 60 procedure dat m.b.v. meegegeven arrays: a, b, c en g en waarden ϕ en ψ de u berekent.

Stelling 3.4: Als in (3.44) $a_i \geq 0$, $c_i \geq 0$ en $b_i \geq \frac{1}{2}(a_i + c_i) + \delta$ en $\delta > 0$ dan geldt:

$$|u_i| \leq \max\{|\phi|, |\psi|, \frac{1}{2\delta} \max_{0 < i < n} |g_i|\} \quad (3.45)$$

Bewijs: Zij $u_k = \max_{0 \le i \le N} u_i$, en $u_k \ge 0$.

Als $k = 0$ of $k = N$ dan is er niets te bewijzen.
Stel $0 < k < N$.
Dan geldt:

$$g_k = a_k u_{k-1} - 2 b_k u_k + c_k u_{k+1}$$

$$\le (a_k - 2 b_k + c_k) u_k \le -2\delta u_k$$

dus $g_k < 0$.
Bovendien geldt:

$$u_i \le u_k \le \frac{-g_k}{2\delta} \le \frac{1}{2\delta} \max_{0 < i < N} |g_i|.$$

Dus als $u_k \ge 0$ dan is de stelling bewezen. Voor het geval dat $u_k < 0$, beschouwen we hetzelfde stelsel vergelijkingen (3.44), waarin ϕ, ψ en g_i vervangen zijn door $-\phi$, $-\psi$ en $-g_i$.
De oplossing van dit nieuwe stelsel zij v_i, $i = 0, \ldots, N$. Dan geldt $v_i = -u_i$.
Voor v_k (≥ 0) geldt (3.45) met $|u_i|$ vervangen door v_i.
Tenslotte geldt

$$\max_{0 \le i \le N} |u_i| = \max \{ \max_{0 \le i \le N} u_i, \max_{0 \le i \le N} (-u_i) \} =$$

$$= \max \{u_k, \max_{0 \le i \le N} v_i\}$$

waaruit het gestelde onmiddellijk volgt.

Stelling 3.5: Zij u_i de oplossing van (3.44) en $\bar{u}_i$ de oplossing van (3.44), waarin ϕ, ψ, a_i, b_i, c_i en g_i vervangen zijn door $\bar{\phi}$, $\bar{\psi}$, $\bar{a}_i$, $\bar{b}_i$, $\bar{c}_i$ en $\bar{g}_i$.
Als geldt:

1. $\exists\, \delta$, $\delta > 0$, zó dat $a_i \ge \delta$, $c_i \ge \delta$, $b_i \ge \frac{1}{2}(a_i + c_i) + \delta$

2. $|\phi - \bar{\phi}| < \varepsilon$, $|\psi - \bar{\psi}| < \varepsilon$,

$|a_i - \bar{a}_i| < \varepsilon$, $|b_i - \bar{b}_i| < \varepsilon$, $|c_i - \bar{c}_i| < \varepsilon$,

$|d_i - \bar{d}_i| < \varepsilon$, $i = 1, \ldots, N-1$ met $\varepsilon < \delta/3$

dan geldt: $|u_i - \bar{u}_i| < C(\delta,M)\,\varepsilon$,

met C een functie van δ en M, waarin

$$M = \max\{|\Phi|,\ |\psi|,\ \frac{1}{2\delta} \max_{0 < i < N} |g_i|\}.$$

Bewijs: Zie Godunov en Ryabenki pp 146-154.
Opmerking: Bovenstaande stelling betekent dat mits aan de 1e voorwaarden voldaan wordt en mits de rekenfouten niet te groot zijn ($< \delta/3$) dat dan de berekende oplossing niet veel afwijkt van de exacte oplossing.

3.2.2. De algorithme voor het impliciete schema.

De double sweep algorithme toegepast op (3.42) geeft:

$$a_i = -\Theta r,\ b_i = -\tfrac{1}{2}(1 + 2\Theta r),\ c_i = -\Theta r$$

Overgang op $-a_1$, $-b_i$, en $-c_i$ geeft de conditie (stelling 3.4)

$$\exists\delta:\ \tfrac{1}{2}(1 + 2\Theta r) \geq \tfrac{1}{2}(2\Theta r) + \delta,$$
$$\text{dus } \delta = \tfrac{1}{2}.$$

De algorithme zal dus een goede rekenmethode zijn om $R^{-1}f$ te bepalen.

3.2.3. De stabiliteit van het impliciete schema.

We gaan uit van (3.43).
Zij

$$v^k = A\,u^k + \tau\,\rho^{k+1}$$

dan geldt:

$$\max_{0 < n < N} |v_n^k| \leq$$

$$\max_{0 < n < N} |(1 - \Theta)r\,u_{n-1}^k + (1-2(1-\Theta)r)u_n^k + (1 - \Theta r)u_{n+1}^k + \tau\,\rho_n^{k+1}|$$

$$\leq \max_{0 \leq n \leq N} |u_n^k| + \tau \max_{0 \leq n \leq N} |\rho_n^{k+1}|$$

mits

$$\text{1.} \qquad \theta \leq 1,$$

$$\text{2.} \qquad 1-2(1-\theta)r \geq 0 \text{ of } r \leq \frac{1}{2(1-\theta)}. \tag{3.46}$$

Voor de <u>maximum norm</u> geldt dus

$$||v^k|| \leq ||u^k|| + \tau \, ||\rho^{k+1}||. \tag{3.47}$$

Bij gegeven v^k is het niet moeilijk $||u^{k+1}||$ te schatten.
Uit stelling 3.4 volgt:

$$||u^{k+1}|| = \max_{0 \leq n \leq N} |u_n^{k+1}| \leq \max \{|\tau \, \rho_0^{k+1}|, |\tau \, \rho_N^{k+1}|, \max_{0 < n < N} |v_n^k|\},$$

Dus

$$||u^{k+1}|| \leq ||u^k|| + \tau \, ||\rho^{k+1}||$$

$$\leq ||u^0|| + \tau \sum_{j=1}^{k+1} ||\rho^j||.$$

Onder de condities (3.46) is het schema dan stabiel mits voor f een passende norm gekozen wordt:

$$||f|| = \max_k ||f^k|| = \max_k ||\rho^k||$$

$$= \max_k (\max \{|f_0^k|/\tau, \max_{0 < n < N} |f_n^k|, |f_N^k|/\tau\}).$$

Een voldoende criterium voor de stabiliteit is nu gevonden.
De vraag is of dit criterium ook noodzakelijk is.
We beschouwen daartoe (3.43) met $\rho^{k+1} = 0$, dus $f^{k+1} = 0$, $k \geq 0$.
Vervolgens moeten de eigenwaarden van de matrix $B^{-1}A$ berekend worden.
De eigenwaarden λ_{iA} en eigenfuncties e_A^i van A zijn:

$$\left.\begin{aligned} \lambda_{iA} &= 1-4(1-\Phi)r \sin^2 \frac{i\pi/2}{N}, \\ e_A^i &= \sin(n \tfrac{i\pi}{N}) \end{aligned}\right\} \quad i = 1, \ldots, N-1 \qquad (3.48)$$

en twee eigenwaarden $\lambda_{0A} = \lambda_{NA} = 0$ waarbij:

$$e_n^0 = e_n^N = 0.$$

(N.B. Voor speciale waarden van $(1-\Theta)r$, n.l. $(1-\Theta)r = \frac{1}{4}$ en

$$(1-\Theta)r = \frac{t_k + 1 \pm \sqrt{t_k + 1}}{2\, t_k}, \quad t_k = tg \frac{kn}{N},$$

zijn er niet-triviale eigenvectoren bij de twee eigenwaarden

$$\lambda_{0A} = \lambda_{NA} = 0.)$$

De eigenwaarden λ_{iB} en eigenfuncties e_B^i van B zijn:

$$\left.\begin{aligned} \lambda_{iB} &= 1 + 4\Theta r \sin^2 \frac{i\pi/2}{N}, \\ e_B^i &= \sin(n \tfrac{i\pi}{N}). \end{aligned}\right\} \quad i = 1, \ldots, N-1 \qquad (3.49)$$

en twee eigenwaarden $\lambda_{0B} = \lambda_{NB} = 1$, met $e_{Bn}^0 = 1$ en $e_{Bn}^N = n$.

Aangezien A en B dezelfde eigenvectoren bezitten (afgezien van e^0 en e^N) volgt dat de eigenwaarden λ_i van $B^{-1}A$ gegeven zijn door:

$$\lambda_i = \frac{1 - 4(1-\Theta)r \sin^2 (i\pi/2N)}{1 + 4\Theta r \sin^2 (i\pi/2N)}, \quad 0 < i < N. \qquad (3.50)$$

De instabiliteits eis: $|\lambda_i| > 1+0(\tau)$ betekent:

a) $$1-4(1-\Theta)r \sin^2 (i\pi/2N) \geq 1 + 4\Theta r \sin^2 (i\pi/2N)$$

dus

$$-4r \sin^2 (i\pi/2N) > 0$$

hetgeen nooit vervuld is.

b) $$-1 + 4(1-\theta)r \sin^2 (i\pi/2N) > 1 + 4\theta r \sin^2 (i\pi/2N)$$

dus

$$r > 1/\{(2 - 4\theta) \sin^2 (i\pi/2N)\} \geq \frac{1}{2 - 4\theta}.$$

Conclusie: als

$$r > 1(2 - 4\theta) \tag{3.51}$$

dan is het schema instabiel t.o.v. de maximum norm.
Voor

$$1/(2 - 4\theta) \geq r > 1/(2 - 2\theta),$$

is er nog geen uitspraak over de stabiliteit.

We zullen in de volgende sectie aantonen dat, als we inplaats van de maximum norm de twee-norm:

$$||v|| = (\frac{1}{N+1} \sum_{n=0}^{N} v_n^2)^{\frac{1}{2}}, \tag{3.52}$$

nemen, dan blijkt $r \leq 1/(2 - 4\theta)$ een noodzakelijk èn voldoende criterium te zijn voor de stabiliteit, mits we ons beperken tot die roosterfuncties f waarvoor $f_0^{k+1} = f_N^{k+1} = 0$, $k \geq 0$.

3.2.4. Orthogonaliteit en volledigheid van eigenvectoren.

Beschouw een matrix A met eigenvectoren e^i, $i = 1, \ldots, N$ en eigenwaarden λ_i.
De eigenvectoren vormen een volledige basis indien elke vector $v = (v_1, \ldots, v_n)^{Tr}$ als lineaire combinatie van de e^i's is te schrijven:

$$v = \sum_{i=1}^{N} \alpha_i e^i. \tag{3.53}$$

We definieren een inwendig product van twee vectoren v en w als volgt:

$$(v,w) = \frac{1}{N} \sum_{n=1}^{N} v_n \bar{w}_n, \text{ met } \bar{w}_n \text{ de gecomplex toegevoegde van } w_n, \tag{3.54}$$

en als norm van de vector v definieren we de zogenaamde tweenorm:

$$||v|| = (v,v)^{\frac{1}{2}}. \tag{3.55}$$

Zij nu $v = \sum_{i=1}^{N} \alpha_i e^i$ en $w = \sum_{i=1}^{N} \beta_i e^i$ dan geldt:

$$(v,w) = \frac{1}{N} \left(\sum_{n=1}^{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \bar{\beta}_j e_n^i \bar{e}_n^j \right)$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \bar{\beta}_j (e^i, e^j).$$

De eigenvectoren worden een <u>orthonormale</u> basis genoemd indien geldt:

$$(e^i, e^j) = \delta_{i,j} \tag{3.56}$$

met $\delta_{i,j}$ de bekende Kronecker delta.

Als A een orthonormale en volledige basis eigenvectoren bezit dan heet A <u>normaal</u> en er geldt:

$$(v,w) = \sum_{i=1}^{N} \alpha_i \bar{\beta}_i. \tag{3.57}$$

Substitueren we $w = e^j$ dan vinden we:

$$\alpha_j = (v, e^j) \tag{3.58}$$

Zij nu gegeven het differentieschema:

$$u^{k+1} = A u^k + \tau \rho^{k+1}, \tag{3.59}$$

met een normale matrix A.
Uit

$$u^k = A^k u^0 + \tau \sum_{j=1}^{k} A^{k-j} \rho^j,$$

en

$$u^0 = \sum_{i=1}^{N} \alpha_i e^i, \; \rho^k = \sum_{i=1}^{N} \beta_i^k e^i, \text{ volgt:}$$

$$||u^k|| \leq ||A^k u^0|| + \tau \sum_{j=1}^{k} ||A^{k-j} \rho^j||$$

$$= (\sum_{i=1}^{N} |\lambda_i^k \alpha_i|^2)^{\frac{1}{2}} + \tau \sum_{j=1}^{k} (\sum_{i=1}^{N} |\lambda_i^{k-j} \beta_i^j|^2)^{\frac{1}{2}}$$

$$\leq |\lambda_m^k| \, ||u^0|| + \tau \sum_{j=1}^{k} |\lambda_m|^{k-j} \, ||\rho^j|| , \qquad (3.60)$$

met $|\lambda_m| = \max_{0 \leq i \leq N} (|\lambda_i|)$.

Als we de norm van f zo kiezen dat $||u^0|| \leq ||f||$ en $||\rho^j|| \leq ||f||$ dan geldt:

$$||u^k|| \leq \{|\lambda_m|^k + \tau \, \frac{|\lambda_m|^k - 1}{|\lambda_m| - 1}\} \, ||f|| .$$

Een voldoende conditie voor stabiliteit t.o.v. de twee-norm is dus:

$$|\lambda_m| \leq 1 + 0(\tau), \qquad (3.61)$$

mits de matrix A normaal is en mits voor $||f||$ een geschikte norm gekozen wordt.

Stelling 3.2 zegt dat criterium (3.62) een noodzakelijk criterium voor de stabiliteit t.o.v. de maximum norm is. Dat stelling 3.2 ook waar is als we de twee-norm beschouwen is eenvoudig te zien.
(Neem $\rho^k = 0$ en e^i zó dat $\lambda_i > 1 + 0(\tau)$ dan is $u^k = \lambda_i^k u^0$ met $u^0 = e^i$ en $||u^k|| \to \infty$ als $\tau \to 0$).

Stelling 3.6: Een matrix A is dan en slechts dan normaal indien $AA^T = A^TA$.

Bewijs: Zie G.F. Simmons: Introduction to Topology and Modern Analysis, Mc. Graw Hill, 1963 (International Student Edition) pp. 278-297.

Gevolg: Een symmetrische matrix is normaal.

Voor een symmetrische reële matrix A met onderling verschillende eigenwaarden zullen we op elementaire wijze aantonen dat de matrix normaal is:

Bewijs:

1. De eigenvectoren zijn orthogonaal:

$$(A\, e^i,\, e^j) = (\lambda_i\, e^i,\, e^j) = \lambda_i (e^i,\, e^j)$$

$$(A\, e^i,\, e^j) = \frac{1}{N} \sum_{n=1}^{N} \Big(\sum_{m=1}^{N} A_{nm}\, e^i_m \Big)\, \bar{e}^j_n$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} A_{mn}\, e^i_m\, \bar{e}^j_n$$

$$= \frac{1}{N} \sum_{m=1}^{N} \Big(\sum_{n=1}^{N} A_{mn}\, e^j_n \Big)\, \bar{e}^i_m = (e^i,\, Ae^j)$$

$$= (e^i,\, \lambda_j\, e^j) = \lambda_j\, (e^i,\, e^j)$$

(N.B. de eigenwaarden zijn reëel).

dus $(\lambda_i - \lambda_j)\, (e^i,\, e^j) = 0$

en aangezien $\lambda_i \neq \lambda_j$ volgt $(e^i,\, e^j) = 0$ voor $i \neq j$.

Als we nu e_i vervangen door $\tilde{e}_i = e_i / (e_i,\, e_i)^{\frac{1}{2}}$, dan geldt

$$(\tilde{e}^i,\, \tilde{e}^j) = \frac{(e^i,\, e^i)}{(e^i,\, e^i)} = 1.$$

M.a.w. de eigenvectoren $\tilde{e}^i$ zijn orthonormaal.

In het vervolg laten we de slangetjes weer weg en nemen aan dat de eigenvectoren reeds zó genormeerd zijn dat $(e^i,\, e^i) = 1$.

2. De eigenvectoren zijn volledig, d.w.z. ze spannen de gehele N-dimensionale vectorruimte op.
Stel dat dit niet het geval was; d.w.z. er zijn a_i's te vinden zo dat

$$\sum_{i=1}^{N} a_i \, e^i = 0 \text{ en niet alle } a_i\text{'s zijn nul.}$$

Stel $a_1 \neq 0$ dan is

$$e^1 = -\frac{1}{a_1} \sum_{i=2}^{N} a_i \, e^i$$

dus

$$(e^1, e^1) = -\frac{1}{a_1} \sum_{i=2}^{N} a_i \, (e^i, e^1) = 0$$

dus $e^1 = 0$, welke niet tot de eigenvectoren behoort.
q.e.d.

Uit het voorgaande blijkt dat normale matrices abnormaal prettige eigenschappen bezitten.
Dat er voor niet-normale matrices interessante en practische theorieën worden opgesteld behoeft geen betoog (zie bijv. Richtmeyer-Morton).

De resultaten van deze sectie passen we toe op het differentieschema (3.43), dat we schrijven als:

$$u^{k+1} = \tilde{A} \, u^k + \tau \, \tilde{\rho}^{k+1},$$

met

$$\tilde{A} = B^{-1} A \text{ en } \tilde{\rho}^{k+1} = B^{-1} \rho^{k+1}.$$

We nemen nu aan dat

1e. $u_0^0 = u_N^0 = 0$

2e. $\rho_0^{k+1} = \rho_N^{k+1} = 0$, $k \geq 0$; dus $\tilde{\rho}_0^{k+1} = \tilde{\rho}_N^{k+1} = 0$.

Dus voor alle k geldt: $u_0^k = u_n^k = 0$.
In de ruimte E_{N+1} met vectoren v zó dat $v_0 = v_M = 0$ geldt dat A en B dezelfde eigenvectoren e^i bezitten met

$$e_n^i = \sin \frac{i\pi n}{N},\ i = 1, \ldots, N-1$$

en dat de bijbehorende eigenwaarden van de nu symmetrische matrices A en B onderling verschillend zijn zo dat de e^i's een, in deze ruimte, orthonormale en volledige basis vormen.
Deze zijn bovendien eigenvectoren van $\tilde{A}$ waarvan de eigenwaarden gegeven zijn in (3.50).
Voor stabiliteit is het dan nodig en voldoende dat $r \leq 1(2 - 4\theta)$.
D.w.z. t.o.v. de twee-norm, en bij geschikte keus van $\|f\|$.

Er moet gelden $\|u^0\| \leq \|f\|$ en $\|\tilde{\rho}^j\| \leq \|f\|$

dus $$\|B^{-1} \rho^j\| \leq \|f\|$$

Nu geldt (stelling 3.4.) als $\tilde{\rho}^j = B^{-1} \rho^j$

$$|\tilde{\rho}_n^j| \leq \frac{1}{2\delta} \max_{0 < n < N} |\rho_n^j|$$

met $\delta = b_i - \frac{1}{2}(a_i + c_i) = \frac{1}{2}$, dus

$$|\tilde{\rho}_n^j| \leq \max_{0 < n < N} |\rho_n^j|$$

en

$$\|B^{-1} \rho^j\| = \left(\frac{1}{N+1} \sum_{n=0}^{N} |v_n^j|^2\right)^{\frac{1}{2}} \leq \max_{0 < n < N} |\rho_n^j|$$

$$= \max_{0 < n < N} |f_n^j|$$

<u>Conclusie</u>: met de gebruikelijke norm:

$$||f|| = \max_k ||f^k||, \quad ||f^0|| = \max_{0 < n < N} |f_n^0|$$

en $$||f^k|| = \max_{0 < n < N} |f_n^k|$$

en $f_0^{k+1} = f_N^{k+1} = 0$ volgt de beperkte stabiliteit.

Slotconclusie: Om de mooie theorie van inwendig product, orthogonaliteit en volledigheid te kunnen gebruiken moet men genoegen nemen met een beperkt soort stabiliteit; namelijk, die stabiliteit die men verkrijgt als men aanneemt dat de randvoorwaarden exact worden berekend, en dat bovendien geen discretisatiefout aan de rand optreedt.

3.3. Stabiliteit van zuivere homogene beginwaarde problemen.

Beschouw een p.d.v. van de vorm:

$$\frac{\partial U}{\partial t} + \sum_{j=0}^{M} \alpha_j \frac{\partial^j U}{\partial x^j} = 0, \; t > 0, \; -\infty < x < +\infty, \tag{3.62}$$

$$U(x,0) = \Phi(x), \; -\infty < x < +\infty,$$

waarin de α_j's constant zijn.

Een consistent differentieschema Ru = f is gegeven als volgt:

$$\frac{1}{\tau}(u_u^{k+1} - u_n^k) + \sum_{j=0}^{M} \alpha_j \frac{D^j}{h^j} u_n^k = 0, \; k \geq 0, \tag{3.63}$$

$$u_n^0 = f_N^0.$$

Met $D^0 v_n = v_n$,

$$D^1 v_n = \tfrac{1}{2}(v_{n+1} - v_{n-1}),$$

$$D^{j+1} v_n = \tfrac{1}{2}(D^j v_{n+1} - D^j v_{n-1}).$$

Dat het schema consistent is volgt onmiddellijk door volledige inductie:

$$\frac{D'}{h}\,[U]_n^k = \Big[\frac{\partial U}{\partial x}\Big]_n^k + O(h^2),$$

$$\frac{D^{j+1}}{h^{j+1}}\,[U]_n^k = \frac{D^1}{h}\Big(\Big[\frac{\partial^j U}{\partial x^j}\Big]_{n+1}^k - \Big[\frac{\partial^j U}{\partial x^j}\Big]_{n-1}^k + O(h^2)\Big)$$

$$= \Big|\frac{\partial^{j+1} U}{\partial x^j}\Big|_n^k + O(h^2).$$

Dus $\Delta_n^k = O(h^2 + \tau)$.

We nemen nu $f_n^0 = e^{in\phi}$, voor zekere ϕ.
Dan geldt:

$$u_n^i = u_n^0 - \tau \sum_{j=0}^{M} \alpha_j \frac{D^j}{h^j} u_n^0 = e^{in\phi} - \tau \sum_{j=0}^{M} \alpha_j \frac{D^j}{h^j} e^{in\phi}.$$

Nu geldt:

$$D^1 e^{in\phi} = e^{in\phi}\, i \sin\phi,$$

dus

$$D^j e^{in\phi} = (i \sin\phi)^j\, e^{in\phi},$$

zodat

$$u_n^1 = \Big[1 - \tau \sum_{j=0}^{M} \alpha_j \frac{(i \sin\phi)^j}{h^j}\Big]\, e^{in\phi},$$

dus

$$u_n^k = \Big[1 - \tau \sum_{j=0}^{M} \alpha_j \frac{(i \sin\phi)^j}{h^j}\Big]^k e^{in\phi}. \qquad (3.65)$$

Noodzakelijk voor stabiliteit is dus:

$$\Big|1 - \tau \sum_{j=0}^{M} \alpha_j \frac{(i \sin\phi)^j}{h^j}\Big| < 1 + O(\tau). \qquad (3.66)$$

Voor $\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}$ volgt:

$$\left|1 + \tau \frac{(i \sin\phi)^2}{h^2}\right| = \left|1 - \frac{\tau}{h^2} \sin^2\phi\right| < 1 + O(\tau)$$

bedenken we dat $D^2 v_n = v_{n+2} - 2 v_n + v_{n-2}$,

dus de werkelijke stapgrootte is 2 h, dan volgt het oude criterium:

$$r = \tau/(2h)^2 \leq \tfrac{1}{2}.$$

Evenzo gaat:

$$\frac{1}{\tau}(u_n^{k+1} - u_n^k) - \theta \frac{D^2}{h^2} u_n^{k+1} - (1-\theta) \frac{D^2}{h^2} u_n^k = 0.$$

Stellen we $u_n^k = \lambda^k e^{in\phi}$ dan volgt:

$$\frac{1}{\tau}(\lambda - 1) - \theta \lambda \frac{(i \sin\phi)^2}{h^2} - (1 - \theta) \frac{(i \sin\phi)^2}{h^2} = 0,$$

dus $$\lambda = \frac{1 - (1 - \theta)\tau \sin^2\phi / h^2}{1 + \theta \tau \sin^2\phi / h^2}.$$

Hetzelfde procédé kunnen we gebruiken voor het volgende differentieschema (van Richardson):

$$\frac{1}{2\tau}(u_n^{k+1} - u_n^{k-1}) - \frac{D^2}{h^2} u_n^k = 0, \qquad (3.67)$$

waarvan de discretisatiefout $O(\tau^2 + h^2)$ is.

Stellen we $u_n^k = \lambda^k e^{in\phi}$ dan volgt:

$$\frac{1}{2\tau}(\lambda - \frac{1}{\lambda}) - \frac{(i \sin\phi)^2}{h^2} = 0,$$

of

$$\lambda^2 + \frac{2\tau}{h^2} \sin^2\phi - 1 = 0,$$

zodat $$\lambda = \frac{-\tau}{h^2} \sin^2\phi \pm \sqrt{(\frac{\tau}{h^2} \sin^2\phi)^2 + 1}.$$

Als $r = \tau/h^2$ = constant dan volgt:

$$|r \sin^2\phi + \sqrt{(r \sin^2\phi)^2 + 1}\ | > 1;$$

dus het schema van Richardson is altijd instabiel.

Het volgende differentieschema is van Du Fort - Frankel:

$$\frac{1}{2\tau}(u_n^{k+1} - u_n^{k-1}) - \frac{u_{n-1}^k - (u_n^{k+1} + u_n^{k-1}) + u_{n+1}^k}{h^2} = 0. \qquad (3.68)$$

Vullen we in:

$u_n^k = \lambda^k e^{in\phi}$ dan volgt:

$$\frac{1}{2\tau}(\lambda - \frac{1}{\lambda}) - \frac{e^{-i\phi} - (\lambda + \frac{1}{\lambda}) + e^{i\phi}}{h^2} = 0,$$

dus $$(1 + 2r)\lambda^2 - 4r \cos\phi\ \lambda + 2r - 1 = 0,$$

en $$\lambda_{1,2} = \frac{2r \cos\phi \pm \sqrt{1 - 4r^2 \sin^2\phi}}{1 + 2r}.$$

Nu geldt als $1 \geq 4r^2 \sin^2\phi$,

$$|\lambda_{1,2}| \leq \frac{|2r \cos\phi| + |\sqrt{1 - 4r^2 \sin^2\phi}|}{1 + 2r}$$

$$\leq \frac{2r\ |\cos\phi| + 1}{2r + 1} \leq 1.$$

Als $1 < 4r^2 \sin^2\phi$, dan geldt:

$$|\lambda_{1,2}|^2 = \frac{1}{(1+2r)^2} \{4r^2 \cos^2\phi + 4r^2 \sin^2\phi - 1\}$$

$$= \frac{1}{(1+2r)^2} (4r^2 - 1) = \frac{2r - 1}{2r + 1} \leq 1.$$

Conclusie: het schema van Du Fort-Frankel voldoet altijd aan de noodzakelijke eis voor stabiliteit.

De discretiseringsfout blijkt $O(\tau^2 + h^2 + \frac{\tau^2}{h^2})$ te zijn.
Dit betekent dat, hoewel aan τ/h^2, in principe, geen restrictie is opgelegd, τ en h niet naar nul mogen gaan zó dat τ/h = constant. τ moet sneller naar nul gaan dan h (bijvoorbeeld zó dat τ/h^2 = constant!)

N.B. als $\tau/h = C$, dan wordt niet de difussievergelijking benaderd, maar de golfvergelijking:

$$- C \frac{\partial^2 U}{\partial t^2} + \frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0.$$

Een probleem bij Du Fort-Frankel is de start; immers, behalve u_n^0, is ook u_n^1 nodig.

De golfvergelijking kan als volgt worden gediscretiseerd:

$$\frac{1}{\tau^2} (u_n^{k+1} - 2 u_n^k + u_n^{k-1}) - \frac{D^2}{h^2} u_n^k = 0;$$

stellen we $u_n^k = \lambda e^{in\phi}$ dan volgt:

$$\frac{1}{\tau^2} (\lambda - 2 + \frac{1}{\lambda}) + \frac{\sin^2\phi}{h^2} = 0,$$

zodat

$$\lambda_{1,2} = 1 - \frac{\tau^2}{2h^2}\sin^2\phi \pm \sqrt{(1 - \frac{\tau^2}{2h^2}\sin^2\phi)^2 - 1}.$$

Zij $$\gamma = \frac{\tau^2}{2h^2}\sin^2\phi.$$

Als $(1 - \gamma)^2 \leq 1$ dan zijn er twee complexe wortels, zodat

$$|\lambda_{1,2}|^2 = (1 - \gamma)^2 + (1 - (1 - \gamma)^2) = 1.$$

De eis $(1 - \gamma)^2 \leq 1$ betekent: $0 \leq \gamma \leq 2$

dus als $\tau < 2h$ dan is aan de noodzakelijke eis voor stabiliteit voldaan.
Bovendien geldt, als $(1 - \gamma)^2 > 1$, dat

$$\lambda_{1,2} = 1 - \gamma \pm \sqrt{(1 - \gamma)^2 - 1}.$$

zodat, voor het + teken, $|\lambda| > 1$; er is dus instabiliteit.
Bedenken we dat de werkelijke stapgrootte h^* in de x richting 2h is, dan luidt de noodzakelijke eis voor stabiliteit:

$$\tau < h^*, \tag{3.69}$$

hetgeen ook als <u>karakteristieken criterium</u> bekend staat.
We memoreren nog eens de portée van deze paragraaf:
Voor zuivere begonwaarde problemen met een inhomogene term gelijk aan nul en met constante coëfficiënten, kunnen we een noodzakelijke stabiliteitvoorwaarde aangeven (3.66). Deze voorwaarde kan "voldoende-voorwaarde" worden, voor deze speciale problemen, als de beginfunctie f_n^0 te ontwikkelen is in een set exponentiele roosterfuncties $e^{in\phi_j}$:

$$f_n^0 = \sum_{j=0}^{J} \beta_j e^{in\phi_j}$$

zó dat $\quad \|f^0\| = \sum_{j=0}^{J} |\beta_j| \, \|\varepsilon_j\|, \; \varepsilon_{j_n} = e^{in\phi_j};$

d.w.z. t.o.v. een zeker inwendig product. zó dat

$\|f^0\| = (f^0, f^0)^{\frac{1}{2}}$, moeten de ε_j's orthogonaal zijn.

Als we bijvoorbeeld slechts periodieke roosterfuncties f_n^0, met periode N bekijken ($f_n^0 = f_{n+N}^0$) dan kunnen we nemen

$$\varepsilon_{jn} = e^{in\frac{2\pi j}{N}},$$

met als inwendig product:

$(v, w) = \frac{1}{N} \sum_{n=1}^{N} v_n \bar{w}_n$. Dan geldt:

$$(\varepsilon_p, \varepsilon_q) = \frac{1}{N} \sum_{n=1}^{N} e^{in\frac{2\pi p}{N}} e^{-in\frac{2\pi q}{N}} =$$

$$= \frac{1}{N} \sum_{n=1}^{N} e^{in\frac{2\pi(p-q)}{N}} = \begin{cases} 0 \text{ als } p \neq q \\ 1 \text{ als } p = q \end{cases}$$

We hebben gezien dat bovenomschreven problemen zich erg eenvoudig laten analyseren, dit in tegenstelling tot de vroeger behandelde problemen waarin zowel de randcondities als de inhomogene termen een rol speelden. In de praktijk wordt een bepaald probleem met randvoorwaarden en inhomogene term meestal eerst bekeken zonder de randvoorwaarden en inhomogene term; is dan door een zeker differentieschema aan de noodzakelijke stabiliteitsvoorwaarde, voor dat geidealiseerde probleem, voldaan, dan heeft men meestal voldoende vertrouwen om dit schema ook voor het moeilijke, theoretisch niet analyseerbare probleem te gebruiken. Dat een en ander voorzichtig gehanteerd moet worden leert de diffusievergelijking met een scheve randvoorwaarde (sectie 3.1.2.), waar wel stabiliteit optrad voor $r \leq \frac{1}{2}$ maar geen consistentie, ten opzichte van de gekozen normen.