TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

NUMERIEKE METHODEN

Prof. Dr. G.W. Veltkamp

Najaarssemester 1972

Technische Hogeschool Eindhoven

Onderafdeling der Wiskunde

Numerieke Methoden

Dictaatnr. 2.211 Prijs ƒ 4,50

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

Numerieke Methoden

Najaarssemester 1974

Inhoudsbeschrijving

NUMERIEKE METHODEN

Najaarssemester 1972

N U M E R I E K E M E T H O D E N

Prof.dr. G.W. Veltkamp

Najaarssemester 1972

0. Inleiding

0.0. Het doel van dit college is de hoorder te doen kennismaken met een aantal voor de praktijk belangrijke technieken uit de numerieke wiskunde, zoals algoritmiseren, itereren, extrapoleren, discretiseren, locaal lineariseren. We zullen dit doen door voor een aantal toepassingsgebieden, zoals het oplossen van vergelijkingen, approximatie van functies, integratie, oplossen van gewone en partiële differentiaalvergelijkingen, optimalisatie, enkele praktisch bruikbare methoden te bespreken die van de genoemde technieken gebruik maken.

0.1. De noodzaak tot het gebruik van numerieke methoden kan verschillende achtergronden hebben.

a) Het kan zijn dat men de oplossing van een wiskundig geformuleerd probleem kan schrijven in de vorm van een formule die echter nog sommen van oneindige reeksen, integralen, elementaire transcendente functies zoals sin x, e^x, hogere transcendente functies zoals Bessel-functies, e.d. bevat. Is men geïnteresseerd in de getalwaarde van de uitkomst, dan moet deze met numerieke methoden benaderd worden.

b) Er is voor het gestelde probleem geen analytische oplossing bekend.

c) In geval a) is het, als er numerieke resultaten geproduceerd moeten worden, niet altijd zo dat het eindresultaat van de analytische behandeling het meest geschikte startpunt is voor de numerieke behandeling.

Voorbeelden

1) De analytische oplossing van de lineaire differentiaalvergelijking van de eerste orde

$$\frac{dy}{dx} = e^{x^2}\, y + e^x \qquad (1)$$

met de beginvoorwaarde $y = 0$ voor $x = 0$, luidt

$$y(x) = \int_0^x \exp\left(\xi + \int_\xi^x e^{t^2}\, dt\right) d\xi \ .$$

Het is duidelijk dat hieruit een benadering voor het getal y(1) niet zonder nogal wat numeriek werk gevonden kan worden. Het blijkt eenvoudiger te zijn om rechtstreeks een numerieke benadering voor de oplossing van de differentiaalvergelijking te bepalen.

2) De analytische oplossing van het beginwaardeprobleem

$$\frac{dy}{dx} = \frac{1}{e^{y^2}x + e^{y}}, \quad y = 0 \text{ voor } x = 0 \tag{2}$$

wordt gegeven door de relatie

$$x = \int_0^y \exp\left(\eta + \int_\eta^y e^{t^2}\,dt\right) d\eta .$$

Dit is een zg. impliciete formule, bij gegeven x moet y uit een vergelijking opgelost worden. Dit is numeriek wel mogelijk maar uiteraard gecompliceerder dan in het eerste voorbeeld. Anderzijds is voor diverse numerieke methoden voor het oplossen van differentiaalvergelijkingen de vergelijking (2) niet lastiger dan (1).

0.2. Met het voorgaande propageren we niet om bij een gegeven probleem altijd meteen naar een numerieke methode te grijpen die rechtstreeks het gezochte getal aflevert.
Richard Hamming geeft zijn boek over numerieke methoden als motto mee: the purpose of numerical computation is insight, not numbers. En aan ieder die bij de computer een programma inlevert zou hij willen vragen: what are you going to do with the numbers? Zijn bedoeling is er op te wijzen dat men uit een enkel getal als uitkomst weinig inzicht in het onderzochte probleem, noch in de nauwkeurigheid van de gebruikte methode verkrijgt en dat men anderzijds een pak van vele bladzijden met tussenresultaten vaak na enige tijd moedeloos in de prullenmand gooit omdat de informatie die men zoekt te zeer verborgen ligt tussen een veelheid van niet relevant materiaal. Alleen een goede mathematische analyse, begeleid door numerieke berekeningen met uitvoer van een doordachte hoeveelheid tussenresultaten, leert ons de eigenschappen van een probleem kennen en geeft ons vertrouwen in de gevonden oplossing. Want let wel: ook als het werkelijk slechts om één getal gaat dan nog blijft het probleem: hoe overtuig ik me dat het verkregen getal een acceptabele benadering is voor de oplossing van het gestelde probleem (en trivialer: hoe overtuig ik me dat mijn programma precies de algoritme uitvoert die ik wilde uitvoeren).

0.3. Foutenbronnen en fouten

Men kan 6 soorten van fouten naar hun bronnen onderscheiden.

a) Modelfouten. Deze ontstaan doordat het wiskundige probleem een vereenvoudigde beschrijving van het fysische probleem is. Deze zijn voor de numericus slechts in zoverre van belang dat bij een grof model erg nauwkeurige berekening weinig zin heeft.

b) Beginfouten (initial errors). Vele problemen bevatten parameters die door metingen bepaald moeten worden en dus slechts met beperkte nauwkeurigheid bekend zijn. De consequenties voor de numericus zijn gedeeltelijk dezelfde als ad a), gedeeltelijk als ad d).

c) Afbreekfouten (truncation errors). Deze ontstaan als men een infiniet proces door een finiet proces vervangt. Vervangt men $\log(1 + x)$ door $x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4}$, dan maakt men een afbreekfout (die bv. voor $0 \leq x \leq 0.1$ kleiner is dan 0.5×10^{-5}).

Ook als men $\int_{-h}^{h} f(x)dx$ vervangt door $\frac{h}{3}[f(-h) + 4f(0) + f(h)]$ (regel van Simpson) maakt men een afbreekfout. Tenslotte ook bij het afbreken van iteratieprocessen na eindig veel stappen: zij $y > 0$, $\alpha > 0$ en zij de rij getallen $x_0, x_1, \ldots$ bepaald door

$$x_0 = y \ , \quad x_{n+1} = \frac{y + \alpha x_n^2}{1 + 2\alpha x_n} \ , \quad n = 0,1,\ldots \ . \tag{1}$$

Dan bestaat $x = \lim_{n\to\infty} x_n$ en x is de positieve oplossing van $x + \alpha x^2 = y$. In de praktijk is men gedwongen te stoppen na eindig veel stappen en een x_n als benadering voor de oplossing x te beschouwen. Het verschil $x - x_n$ is een afbreekfout.

d) Afrondingsfouten. Deze ontstaan doordat men meestal met een vast aantal decimale of binaire cijfers werkt. Het resultaat van een rekenkundige bewerking wordt dan als regel afgerond, is dus niet exact. Voorts kunnen de meeste reële getallen niet met eindig veel cijfers gerepresenteerd worden.

e) Rekenfouten (vergissingen, machinefouten e.d.). Om deze te onderdrukken zijn controleberekeningen zeer gewenst. Bij veel iteratieve processen (bv. het proces (1)) betekent het maken van een niet te ernstige rekenfout alleen dat het langer duurt voordat de gewenste nauwkeurigheid bereikt wordt. Uit dit oogpunt zijn deze processen dus aanbevelenswaardig.

Bij het werken met rekenautomaten komen rekenfouten aanzienlijk minder voor dan bij handrekenen.

f) Programmeerfouten. Deze geven aanleiding tot onjuiste uitvoering van een op zichzelf correct algoritme, waardoor in feite een onjuiste algoritme uitgevoerd wordt. Het voorkomen of/en opsporen van deze fouten is een van de grootste zorgen bij het werken met rekenautomaten.

Er is geen vaste conventie voor het teken van de fout. Als regel definieert men de fout door

berekende waarde = exacte waarde + fout.

Deze keuze past goed bij afrondingsfouten e.d. Bij afbreekfouten past beter

exacte waarde = waarde van benadering + afbreekfout.

Van belang is ook het begrip relatieve fout:

relatieve fout = fout/exacte waarde.

Naar analogie hiervan spreekt men ook wel van absolute fout i.p.v. fout.

0.4. Floating point representatie

In moderne rekenautomaten gebruikt men voor niet-gehele getallen vrijwel steeds een zg. floating point (drijvende komma) representatie. Bij een floating point representatie op basis van het tientallig stelsel met t cijfers voor de mantisse en q cijfers voor de exponent zien de representeerbare getallen (de zg. machinegetallen) er uit als

$$a = m \times 10^e, \tag{1}$$

waarin m (de zg. mantisse) een tiendelige breuk is die voldoet aan

$$0.1 \leq |m| < 1 \tag{2}$$

en t cijfers achter de punt heeft (zodat $10^t \times m$ geheel is) en e (de zg. exponent) een geheel getal is dat voldoet aan

$$|e| \leq 10^q - 1. \tag{3}$$

Aan deze getallen wordt toegevoegd het getal 0, bv. te representeren met $m = 0$ (er is dan niet aan (2) voldaan) en $e = 0$.

Het is duidelijk dat ieder machinegetal nu beschreven wordt door 2 tekens (van m en van e) en t+q <u>decimale cijfers</u> (die de waarde 0,1,...,9 kunnen hebben). Maar daaruit volgt ook dat er slechts eindig veel machinegetallen bestaan. Zo is het grootste machinegetal

$$a\ max = \underbrace{0.99\ \ldots\ 9}_{t\ cijfers} \times 10^{10^q-1} = (1 - 10^{-t}) \times 10^{10^q-1} .$$

En het kleinste positieve machinegetal is

$$a\ min = \underbrace{0.10\ \ldots\ 0}_{t\ cijfers} \times 10^{-(10^q-1)} = 10^{-10^{-q}} .$$

Als een getal $x \neq 0$ buiten de range ligt dan spreekt men van <u>overflow</u> als $|x| > a\ max$ en van <u>underflow</u> als $0 < |x| < a\ min$.

Het grootste machinegetal dat kleiner is dan 1 is

$$0.99\ \ldots\ 9 \times 10^0 = 1 - 10^{-t}$$

en het kleinste machinegetal dat groter is dan 1 is

$$0.10\ \ldots\ 01 \times 10^1 = 1 + 10^{1-t} .$$

In het algemeen is de afstand van twee opvolgende machinegetallen die tussen 10^{p-1} en 10^p liggen 10^{p-t} (ga na).

Exacte optelling, vermenigvuldiging, etc. van twee machinegetallen levert als regel een uitkomst, die geen machinegetal is. Deze uitkomst moet door de machine afgerond worden tot het meest nabij gelegen getal. Als x in de zg. <u>range</u> ligt:

$$amin \leq |x| \leq amax$$

en p zo is dat

$$10^{p-1} \leq |x| \leq 10^p$$

dan is de bij x behorende afronding van de vorm

$$a = m \times 10^p$$

met

$$0.1 \leq |m| \leq 1 \quad ^{*)} .$$

---

*) Als m = 1 dan wordt a genoteerd als $0.1 \times 10^{p+1}$.

Hoe groot kan het verschil tussen x en a zijn?
Als p = 0 dan geldt (ga na)

$$|x - a| \leq \tfrac{1}{2} \cdot 10^{-t} .$$

En in het algemeen geldt (ga na)

$$|x - a| \leq \tfrac{1}{2} \cdot 10^{p-t} .$$

Hieruit volgt voor de maximale relatieve fout

$$\frac{|x - a|}{|x|} \leq \frac{\tfrac{1}{2} \cdot 10^{p-t}}{10^{p-1}} = 5 \times 10^{-t}.$$

In de meeste rekenautomaten wordt niet in het 10-tallig stelsel gewerkt maar in een β-tallig stelsel met $\beta = 2$, 8 of 16. Men spreekt dan van binaire, octale, hexadecimale representatie. Bij binaire representatie noemt men de cijfers (die slechts de waarden 0 of 1 hebben) bits.
In een β-tallig stelsel hebben we de machinegetallen

$$a = m \times \beta^{e}$$

waarin m een β-tallige breuk is die voldoet aan

$$\beta^{-1} \leq |m| < 1$$

en t cijfers (die de waarden $0,1,\ldots,\beta-1$ kunnen hebben) achter de punt heeft (dus $\beta^{t} \times m$ is geheel), terwijl

$$|e| \leq \beta^{q} - 1 .$$

Ga zelf na hoe de hierboven besproken zaken aangepast moeten worden.

Er bestaan nog vele varianten op de hierboven aangegeven floating point representaties. Voorts geldt niet voor alle machines dat tussenresultaten die geen machinegetal zijn, steeds correct afgerond worden. Ook de handelwijze in het geval van over- of underflow is per machine verschillend.

Tenslotte is er bij niet-decimale machines nog het probleem van conversie van de invoerrepresentatie (meestal 10-tallig) naar de interne representatie en omgekeerd. Ook hierbij zijn als regel afrondingen noodzakelijk.

0.5. Foutenvoortplanting, stabiliteit, conditie

Het is duidelijk dat fouten die in een bepaald stadium van een berekening gemaakt worden, in het algemeen aanleiding zullen geven tot fouten in de resultaten van de verdere berekening. Men spreekt van voortplanting van fouten. Een belangrijk deel van de numerieke analyse is gewijd aan de invloed hiervan op de eindresultaten van een berekening.

Men spreekt van een numeriek stabiele algoritme als de fout in het eindantwoord, veroorzaakt door een ergens in de berekening gemaakte fout, niet essentieel groter is (meestal in relatieve zin) dan de fout in het tussenresultaat.

Men spreekt van een goed geconditioneerd probleem als het eindantwoord "stabiel" van de parameters van het probleem afhangt (d.w.z. bij kleine variaties in de parameters ook slechts weinig varieert).

Het is redelijk te verwachten, dat er bij een goed geconditioneerd probleem een stabiele algoritme bestaat (hoewel niet elke voor de hand liggende algoritme stabiel behoeft te zijn). Anderzijds zal bij een slecht geconditioneerd probleem vaak (maar niet altijd) ieder algoritme weinig stabiel zijn.

Fysische, economische, etc. modellen die tot slecht geconditioneerde problemen aanleiding geven, dienen met wantrouwen bezien te worden.

Voorbeeld van een numeriek instabiele algoritme.
Een fysisch interessante grootheid x hangt met een meetbare grootheid y samen volgens

$$y = x + \alpha x^2 \ . \qquad (1)$$

De relevante waarden van x en y zijn in de buurt van 1, α is van de orde van 0.01.
Beschouwt men (1) als vierkantsvergelijking in x, waarvan we de positieve wortel moeten hebben, dan volgt met de bekende formule

$$x = \frac{\sqrt{1 + 4\alpha y} - 1}{2\alpha} \ . \qquad (2)$$

Deze formule is echter numeriek instabiel. Rekenen we bv. consequent in drie cijfers achter de komma, dan zal de gevonden waarde van $\sqrt{1 + 4\alpha y}$ een fout kunnen hebben van $\frac{1}{2} \times 10^{-3}$. Maar dat betekent in x een mogelijke fout van $(1/4\alpha) \times 10^{-3} \sim \frac{1}{4} \times 10^{-1}$, zodat de gevonden waarde van x niet eens twee goede cijfers hoeft te hebben. Of anders gezegd: om voor x een fout kleiner dan $\frac{1}{2} \times 10^{-3}$ te kunnen garanderen moeten we het tussenresultaat $\sqrt{1 + 4\alpha y}$ uitre-

kenen met een fout van hoogstens $\alpha \times 10^{-3}$. Het feit dat de relatieve fout in x veel groter is dan die in $\sqrt{1 + 4\alpha y}$ wordt veroorzaakt door zg. cijferverlies: als men van $\sqrt{1 + 4\alpha y}$ vijf goede cijfers achter de komma bepaald heeft, dan blijven er na aftrekking van het getal 1 maar drie goede cijfers over omdat vooraan nullen ontstaan.

Kunnen we aan deze numerieke instabiliteit iets doen? We moeten het aftrekken van bijna gelijke getallen zo veel mogelijk vermijden. Dat kan in het geval van formule (2) met de zg. worteltruc: we kunnen schrijven

$$x = \frac{2y}{\sqrt{1 + 4\alpha y} + 1} . \qquad (3)$$

Ga na dat nu de eindnauwkeurigheid in x ca. net zo groot is als die in $\sqrt{1 + 4\alpha y}$, zodat formule (3) wel numeriek stabiel is.

Opmerking. Formule (3) is exact en numeriek stabiel. Maar wel wat bewerkelijk. Eisen we niet te grote nauwkeurigheid dan kunnen we reeksontwikkelen:

$$x = \frac{2y}{2 + 2\alpha y + \ldots} = y(1 + \alpha y + \ldots)^{-1} = y(1 - \alpha y + \ldots) .$$

De hieruit volgende benaderingsformule

$$x \sim y - \alpha y^2$$

kunnen we ook verkrijgen door uit (1) als nulde benadering te halen

$$x_0 = y$$

en als eerste benadering

$$x_0 = y - \alpha x_0^2 .$$

Met dit proces kunnen we desgewenst doorgaan (successieve substitutie).

Opgave. De wortels van de vierkantsvergelijking $ax^2 + bx + c = 0$ worden gegeven door

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{2c}{-b - \sqrt{b^2 - 4ac}}$$

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{2c}{-b + \sqrt{b^2 - 4ac}} .$$

Ga (eventueel aan de hand van voorbeelden) na welke formules in welke gevallen het meest stabiel zijn.
N.b. Als een der wortels op stabiele wijze berekend is dan kan men de ander altijd stabiel berekenen uit de relatie $x_1 x_2 = c/a$.

De regels voor de foutenvoortplanting bij de elementaire bewerkingen zijn de volgende.

A. Zij de te berekenen grootheid

$$c = a + b.$$

Zij $\bar{a}$ de ter beschikking staande benadering voor a, $\bar{b}$ die voor b. Dan is

$$\bar{c} = \bar{a} + \bar{b}$$

de bijbehorende benadering voor c.
Schrijven we de fouten in $\bar{a}$, $\bar{b}$ en $\bar{c}$ als

$$\delta a = \bar{a} - a, \quad \delta b = \bar{b} - b, \quad \delta c = \bar{c} - c,$$

dan is

$$\delta c = \delta a + \delta b.$$

Voor de relatieve fouten geldt

$$\frac{\delta c}{c} = \frac{a}{a+b}\,\frac{\delta a}{a} + \frac{b}{a+b}\,\frac{\delta b}{b}.$$

Dus:

a) De absolute fout in $\bar{c}$ is de som van de absolute fouten in $\bar{a}$ en $\bar{b}$.

b) De relatieve fout in $\bar{c}$ is een lineaire combinatie van de relatieve fouten in $\bar{a}$ en $\bar{b}$ met gewichten a/(a+b) resp. b/(a+b).
Als a en b hetzelfde teken hebben dan zijn de gewichten positief en hun som is 1. In dit geval geldt o.a.

$$\left|\frac{\delta c}{c}\right| \leq \max\left(\left|\frac{\delta a}{a}\right|, \left|\frac{\delta b}{b}\right|\right).$$

Als a en b verschillend teken hebben dan hebben de gewichten verschillend teken en de som van hun absolute waarden is groter dan 1. Er geldt nu o.a.

$$\left|\frac{\delta c}{c}\right| \leq \left|\frac{a}{a+b}\right| \left|\frac{\delta a}{a}\right| + \left|\frac{b}{a+b}\right| \left|\frac{\delta b}{b}\right|.$$

Als $|a+b|$ klein is ten opzichte van $|a|$ en $|b|$, dan zijn beide gewichten groot en dan is de relatieve fout in $\bar{c}$ als regel veel groter dan die in $\bar{a}$ en $\bar{b}$.
Men vermijdt daarom, indien mogelijk, optelling van getallen met verschillend teken en bijna gelijke absolute waarde (zie het bovengenoemde voorbeeld van een onstabiele algoritme).

B. Voor de aftrekking geldt mutatis mutandis hetzelfde. Hier zal bij aftrekken van getallen met gelijk teken (vooral als ze bijna gelijk zijn) de relatieve fout als regel sterk toenemen.

C. Voor de vermenigvuldiging

$$c = a \times b,\ \bar{c} = \bar{a} \times \bar{b}$$

geldt (als we afzien van een tweede orde term)

$$\delta c = \delta a \times b + a \times \delta b,$$

$$\frac{\delta c}{c} = \frac{\delta a}{a} + \frac{\delta b}{b}\ .$$

Hier geldt dus met name: de relatieve fout in $\bar{c}$ is de som van de relatieve fouten in $\bar{a}$ en $\bar{b}$.

D. Voor de deling $c = a/b$ geldt met name: de relatieve fout in $\bar{c} = \bar{a}/\bar{b}$ is het verschil van de relatieve fouten in $\bar{a}$ en $\bar{b}$. Ga dit na.

0.6. Literatuur

1. Carnahan, B., H.A. Luther and J.O. Wilkes, Applied numerical methods, Wiley, New York etc., 1969.

2. Fröberg, C.E., Introduction to numerical analysis, Addison-Wesley, Reading (Mass.), 1965.

3. Hamming, R.W., Numerical methods for scientists and engineers, McGraw-Hill, New York etc., 1962.

4. Henrici, P., Elements of numerical analysis, Wiley, New York etc., 1964.

5. Lapidus, L., Digital computation for chemical engineers, McGraw-Hill, New York etc., 1962.

6. Modern Computing Methods, National Physical Laboratory, HMSO, London, 1961.

7. Moursand, D.G. and C.S. Duris, Elementary theory and application of numerical analysis, McGraw-Hill, New York etc., 1967.

8. Noble, B., Numerical methods, 2 dln., Oliver and Boyd, Edinburgh, 1964.

9. Ralston, A., A first course in numerical analysis, McGraw-Hill, New York etc., 1965.

10. Stoer, J., Einführung in die Numerische Mathematik I,
Stoer, J., - R. Bulirsch, Einführung in die Numerische Mathematik II, Heidelberger Taschenbücher, Springer-Verlag, Berlin etc., 1972.

1. Het oplossen van vergelijkingen

We bespreken in dit hoofdstuk methoden voor het oplossen van een vergelijking

$$F(x) = 0 \tag{1}$$

en van stelsels vergelijkingen

$$F_1(x_1,x_2,\ldots,x_n) = 0$$

$$F_2(x_1,x_2,\ldots,x_n) = 0$$

$$\ldots\ldots\ldots\ldots$$

$$F_n(x_1,x_2,\ldots,x_n) = 0 \ .$$

1.1. Successieve substitutie

Vele methoden voor het oplossen van (1) komen er op neer dat men de vergelijking herschrijft in een vorm

$$x = f(x) \tag{2}$$

die equivalent is met (1), d.w.z., dat een oplossing van (2) ook oplossing van (1) is. Bovendien moet (2) zo zijn dat f(x) "in de buurt van" een oplossing "niet sterk" van x afhangt. Op (2) kan dan successieve substitutie toegepast worden: men "kiest" (op grond van reeds verworven kennis of intuïtie) een nulde benadering x_0 voor de oplossing en bepaalt vervolgens

$$x_1 = f(x_0)$$

$$\ldots\ldots$$

$$x_n = f(x_{n-1}) \tag{3}$$

$$\ldots\ldots$$

Voorbeeld

Als in de vergelijking

$$x + \gamma x^3 = y \tag{4}$$

(y en γ bekend, x onbekend) de term γx^3 niet erg belangrijk is, dan schrijven we hem als

$$x = y - \gamma x^3 \ ,$$

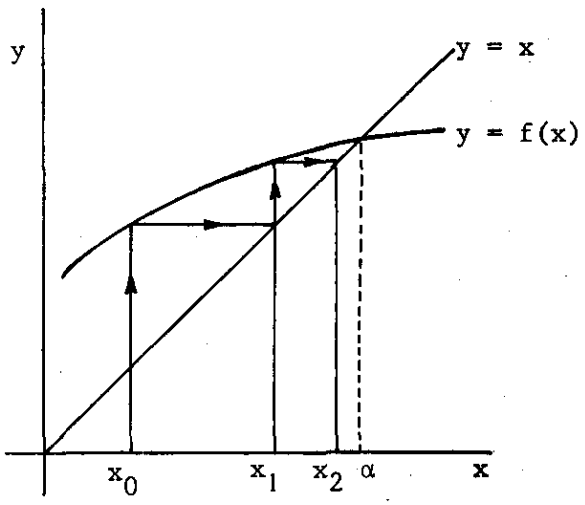
kiezen $x_0 = y$ (of $x_0 = 0$, dan wordt $x_1 = y$!), en bepalen vervolgens

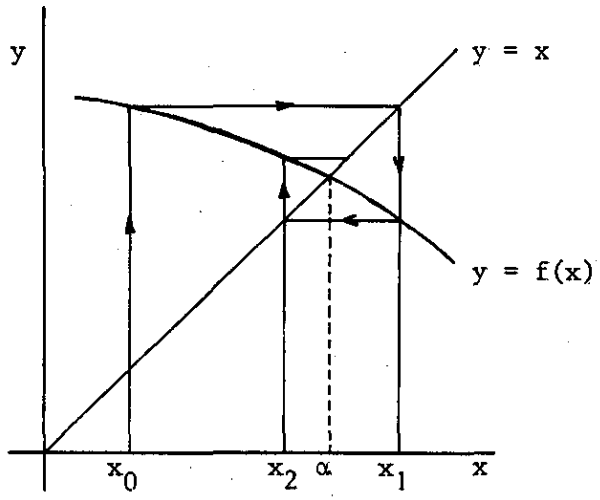
$$x_1 = y - \gamma x_0^3 \ , \ x_2 = y - \gamma x_1^3 \ , \ \ldots \ .$$

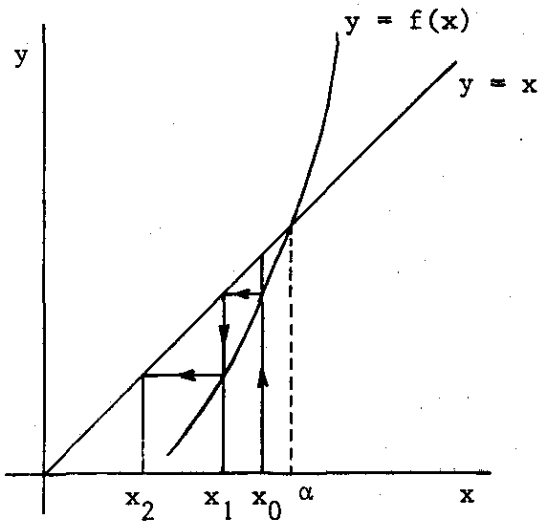
We hopen dat de rij $x_1, x_2, \ldots$ snel naar een limiet nadert; deze limiet is dan oplossing van (4) (waarom?).

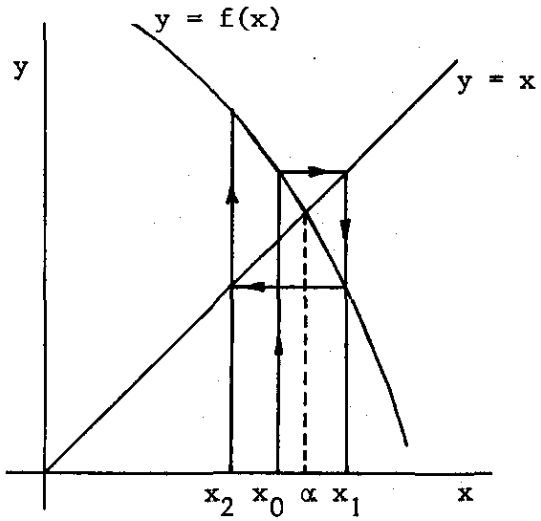
1.1.1. Locale convergentie

De gang van zaken bij de successieve substitutie $x_n = f(x_{n-1})$ wordt duidelijk met de volgende plaatjes (merk op hoe x_1 door een eenvoudige constructie uit x_0 verkregen wordt).









De plaatjes suggereren:

convergentie als $|f'(x)| < 1$,

divergentie als $|f'(x)| > 1$,

monotoon gedrag als $f'(x) > 0$,

oscillerend gedrag als $f'(x) < 0$.

Zij nu α een oplossing van (2) en zij $|f'(\alpha)| < 1$. We bewijzen dat dan het proces convergeert mits x_0 dicht genoeg bij α ligt.

Locale convergentie stelling.

Zij α een oplossing van $x = f(x)$.

Zij $f'(x)$ continu in een omgeving van α.

Zij $f'(\alpha) = A$, met $|A| < 1$.

Dan is er een $\delta > 0$ zodanig dat voor iedere x_0 met $|x_0 - \alpha| \leq \delta$ geldt

i) $\lim_{n\to\infty} x_n = \alpha$,

ii) $\lim_{n\to\infty} \frac{\alpha - x_n}{\alpha - x_{n-1}} = A$,

iii) $\lim_{n\to\infty} \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} = A$,

iv) $\lim_{n\to\infty} \frac{\alpha - x_n}{x_n - x_{n-1}} = \frac{A}{1 - A}$.

Bewijs. Als $|f'(\alpha)| < 1$ en $f'(x)$ continu is dan is er een $\delta > 0$ en een L met $0 < L < 1$ zodanig dat $|f'(x)| \leq L$ voor $|x - \alpha| \leq \delta$.
Zij nu $|x_0 - \alpha| \leq \delta$. Dan is volgens de middelwaardestelling

$$x_1 - \alpha = f(x_0) - f(\alpha) =$$

$$= f'(\xi_0)(x_0 - \alpha) ,$$

met ξ_0 tussen x_0 en α. Dus zeker $|\xi_0 - \alpha| \leq \delta$, dus

$$|x_1 - \alpha| \leq L|x_0 - \alpha|$$

en met name

$$|x_1 - \alpha| \leq \delta .$$

Analoog $|x_n - \alpha| \leq \delta$ en

$$|x_n - \alpha| \leq L|x_{n-1} - \alpha| \leq L^n|x_0 - \alpha| \ . \qquad (5)$$

Daar $L < 1$ volgt hieruit i).

Uit

$$x_n - \alpha = f'(\xi_{n-1})(x_{n-1} - \alpha)$$

met ξ_{n-1} tussen x_{n-1} en α volgt ii), omdat $x_{n-1} \to \alpha$ en dus $\xi_{n-1} \to \alpha$ voor $n \to \infty$ en $f'(x)$ continu is.

Uit

$$x_n - x_{n-1} = f'(\eta_{n-1})(x_{n-1} - x_{n-2})$$

met η_{n-1} tussen x_{n-1} en x_{n-2} volgt op analoge wijze iii).

Uit

$$\frac{\alpha - x_n}{x_n - x_{n-1}} = \frac{\dfrac{x_n - \alpha}{x_{n-1} - \alpha}}{1 - \dfrac{x_n - \alpha}{x_{n-1} - \alpha}}$$

volgt met ii) tenslotte iv). □

Deze stelling is een typische locale convergentie stelling. Uit het gegeven omtrent $f'(x)$ in het punt α volgt convergentie mits x_0 "dicht genoeg" bij α gekozen wordt.
De limietrelaties vertellen hoe de rij $\{x_n\}$ zich "op de duur" gedraagt.
ii) zegt dat de verhouding van de opvolgende fouten nadert tot A (= $f'(\alpha)$).

iii) zegt dat ook de verhouding van de opvolgende correcties nadert tot A. Dit is van groot belang want de getallen

$$A_n := \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} \qquad (6)$$

kunnen we tijdens het proces uitrekenen. En daarmee hebben we een benadering voor A.
iv) geeft aan hoe de fout in x_n "op de duur" samenhangt met de laatste correctie $x_n - x_{n-1}$. We zien hieruit dat, als A dicht bij +1 is, $\alpha - x_n$ aan-

zienlijk groter kan zijn dan $x_n - x_{n-1}$; het is in dit geval dus gevaarlijk om

$$|x_n - x_{n-1}| \leq \epsilon$$

als stopcriterium te gebruiken (als ϵ de opgegeven tolerantie voor $|\alpha - x_n|$ is). Als A dicht bij -1 is dan is de convergentie wel langzaam, maar oscillerend. Uit ii) volgt dat op de duur $|\alpha - x_n| < \frac{1}{2}|x_n - x_{n-1}|$.

1.1.2. Convergentiefactor en convergentie orde

Uit de locale convergentiestelling volgt dat voor een rij $\{x_n\}$, verkregen met het successieve substitutieproces (3), geldt

$$\lim_{n\to\infty} \frac{\alpha - x_n}{\alpha - x_{n-1}} = A \ .$$

De grootheid $A = f'(\alpha)$ wordt de asymptotische convergentiefactor genoemd; de grootheid

$$R := - {}^{10}\log|A| \tag{7}$$

heet de asymptotische convergentiesnelheid van het proces.

Opgave. Ga na dat het aantal iteraties dat nodig is om de fout in x_n met een factor 10 te verminderen asymptotisch gelijk is aan 1/R.

De convergentiesnelheid R is alleen gedefinieerd als $A \neq 0$. Uit (7) kan men concluderen dat als $A = 0$ de convergentie zeer snel zal zijn. In dat geval wordt de convergentiesnelheid met behulp van een andere grootheid aangeduid: namelijk de convergentie orde.

Definitie. Een iteratieproces dat een rij $\{x_n\}$ oplevert, die convergeert naar de limiet α heeft tenminste de convergentie orde p als geldt

$$\lim_{n\to\infty} \frac{|x_n - \alpha|}{|x_{n-1} - \alpha|^p} = B \ .$$

Als $B \neq 0$ dan heet p de convergentie orde van het proces.

Voor de convergentie orde p geldt in ieder geval $p \geq 1$.
Als $p = 1$ dan spreken we van lineaire convergentie. In dat geval geldt $B \leq 1$.
Als $p = 2$, resp. $p = 3$, dan spreken we van kwadratische, resp. kubische convergentie.

Opmerking. Als $p > 1$ dan is de convergentiefactor A van het proces gelijk aan nul. Uit de locale convergentiestelling volgt dat het proces locaal convergeert voor iedere waarde van B. Als $p = 1$ en $B = 1$, dus $A = \pm 1$, dan is het mogelijk dat het proces niet locaal convergent is. Als het proces echter locaal convergent is dan zeggen we ook in dit geval dat het proces lineair convergeert.

Stel dat de asymptotische convergentiefactor $A = f'(\alpha)$ van het successieve substitutieproces (3) nul is. Dan geldt volgens de Taylorreeks met restterm

$$\begin{aligned} x_n - \alpha &= f(x_{n-1}) - f(\alpha) \\ &= f'(\alpha)(x_{n-1} - \alpha) + \tfrac{1}{2}f''(\xi_{n-1})(x_{n-1} - \alpha)^2 \\ &= \tfrac{1}{2}f''(\xi_{n-1})(x_{n-1} - \alpha)^2 \ . \end{aligned}$$

Hieruit volgt

$$\lim_{n\to\infty} \frac{x_n - \alpha}{(x_{n-1} - \alpha)^2} = \tfrac{1}{2}f''(\alpha) \ .$$

Dus als $f'(\alpha) = 0$ dan is het proces kwadratisch convergent als $f''(\alpha) \neq 0$. Geldt ook $f''(\alpha) = 0$ dan is de orde van het proces tenminste 3, aangenomen dat f tenminste driemaal continu differentieerbaar is.

Opmerkingen

1. Als in het geval van kwadratische convergentie

$$|\tfrac{1}{2}f''(x)| \leq M$$

in een omgeving van α, dan geldt zeker (ga na)

$$|M(x_n - \alpha)| \leq |M(x_{n-1} - \alpha)|^2 \ , \qquad (8)$$

dus als bijv. $|M(x_{n-1} - \alpha)| \leq 10^{-p}$, dan is $|M(x_n - \alpha)| \leq 10^{-2p}$. Men drukt dit wel slordig uit door te zeggen dat x_n tweemaal zoveel goede cijfers heeft als x_{n-1}.

2. Men kan bewijzen dat

$$\lim \frac{x_n - x_{n-1}}{(x_{n-1} - x_{n-2})^2} = -\tfrac{1}{2}f''(\alpha) \ .$$

3. Ook kwadratische convergentie is een typisch locale zaak. Dit blijkt al uit (8): het kwadratische karakter wordt pas interessant als bijv. $|x_{n-1} - \alpha| \leq 1/2M$.

1.1.3. Extrapolatie volgens Aitken

In 1.1.2 hebben we geconstateerd dat een proces dat meer dan lineair convergent is zeer snel convergeert. Daarentegen convergeert een lineair convergent proces zeer langzaam als A niet dicht bij nul ligt.

Veronderstel nu dat het successieve substitutieproces lineair convergeert. Dan geldt bij benadering

$$\frac{\alpha - x_n}{x_n - x_{n-1}} \approx \frac{A}{1 - A} \ .$$

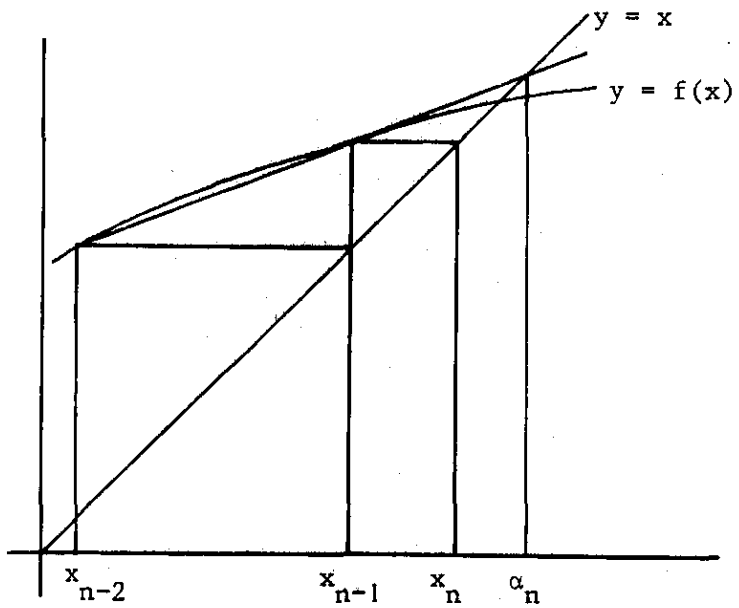
Vervangen we in deze formule A door A_n uit formule (6) dan vinden we als benadering voor de fout in x_n

$$\alpha - x_n \approx \frac{A_n}{1 - A_n} (x_n - x_{n-1}) \ .$$

We kunnen daarom verwachten dat

$$\alpha_n := x_n + \frac{A_n}{1 - A_n} (x_n - x_{n-1}) \qquad (9)$$

een betere benadering voor α zal zijn dan x_n. Ook het hiernavolgende plaatje suggereert dit.



α_n is de x-coördinaat van het snijpunt van de rechte door $(x_{n-2}, f(x_{n-2}))$ en $(x_{n-1}, f(x_{n-1}))$ met de rechte $y = x$. We hebben dus als het ware de kromme $y = f(x)$ vervangen door de rechte door twee punten van deze kromme en hiermee door extrapolatie de benadering α_n gevonden.

Formule (9) kan met behulp van differenties ook op een andere manier geschreven worden.
Bijvoorbeeld met behulp van achterwaartse differenties

$$\nabla x_n := x_n - x_{n-1}, \quad \nabla^2 x_n = \nabla x_n - \nabla x_{n-1},$$

$$\alpha_n = x_n - \frac{(\nabla x_n)^2}{\nabla^2 x_n}$$

of met behulp van voorwaartse differenties

$$\Delta x_{n-2} = x_{n-1} - x_{n-2}, \quad \Delta^2 x_{n-2} = \Delta x_{n-1} - \Delta x_{n-2},$$

$$\alpha_n = x_{n-2} - \frac{(\Delta x_{n-2})^2}{\Delta^2 x_{n-2}}.$$

Deze laatste formule verklaart waarom deze methode Δ^2-extrapolatie van Aitken wordt genoemd.

Als de asymptotische convergentiefactor A dicht bij 0 ligt dan moet x_n dicht bij α zijn voordat de kromme $y = f(x)$ redelijk benaderd wordt door een koorde. De methode werkt in dit geval dan ook vaak averechts.
Als A dicht bij 1 ligt dat is (9) een formule die erg gevoelig is voor afrondingsfouten, dus Aitken extrapolatie is dan een instabiele algorithme. Het oplossen van (2) is in dit geval ook een slecht geconditioneerd probleem (ga na).
Als A negatief is dan hebben we interpolatie (ga na met een plaatje) en werkt de methode vaak voortreffelijk.

1.1.4. Numerieke stabiliteit

We onderzoeken nu het gedrag van het successieve substitutie proces in het geval dat f(x) niet exact uitgerekend kan worden t.g.v. afrond- of afbreekfouten.
Veronderstel dat de getallen $\tilde{x}_1, \tilde{x}_2, \ldots$ voldoen aan

$$\tilde{x}_n = f(\tilde{x}_{n-1}) + \delta_n , \quad n = 1,2,\ldots ,$$

waarbij van δ_n slechts bekend is dat

$$|\delta_n| \leq \delta , \quad \text{alle } n.$$

Zij $x_0, x_1, x_2, \ldots$ de rij die bij exact rekenen verkregen zou zijn.
Veronderstel dat L, $0 < L < 1$, zo is dat

$$|f(x') - f(x'')| \leq L|x' - x''|$$

voor alle relevante x' en x''. Als $f'(x)$ continu is en $|f'(\alpha)| < 1$, dan is zo'n L er zeker voor x' en x'' in een omgeving van α.
We hebben dan

$$\begin{aligned} |\tilde{x}_n - x_n| &\leq |f(\tilde{x}_{n-1}) - f(x_{n-1})| + |\delta_n| \\ &\leq L|\tilde{x}_{n-1} - x_{n-1}| + \delta . \end{aligned}$$

Hieruit volgt door volledige inductie (indien $x_0 = \tilde{x}_0$)

$$|\tilde{x}_n - x_n| \leq \frac{1 - L^n}{1 - L}\,\delta \,. \tag{10}$$

$\tilde{x}_n$ en x_n kunnen dus niet willekeurig ver uit elkaar raken, hoe groot n ook wordt (omdat het effect van vorige storingen weggedempt wordt met een factor L !). Uit (10) en (5) volgt

$$|\tilde{x}_n - \alpha| \leq |\tilde{x}_n - x_n| + |x_n - \alpha|$$

$$\leq \frac{1 - L^n}{1 - L}\,\delta + L^n|x_0 - \alpha| \,.$$

Hieruit zien we dat de fout $|\tilde{x}_n - \alpha|$, die als regel niet naar nul zal gaan, op de duur slechts weinig groter dan $\delta/(1 - L)$ kan zijn. Als de factor $1/(1 - L)$ niet erg groot is, dan kunnen we het proces stabiel noemen tegen storingen in de berekening van f(x). Wel moeten we bij het bedenken van een stopcriterium rekening houden met deze storingen: $|\tilde{x}_n - \tilde{x}_{n-1}|$ hoeft niet kleiner dan $2\delta/(1 - L)$ te worden.

1.1.5. Globale convergentie

We hebben tot dusver alleen gekeken hoe het successieve substitutieproces zich gedraagt vlak bij het limiet punt. We geven nu een uitspraak met een globaal karakter.

Globale convergentie stelling

Zij

1. f(x) continu voor $|x - a| \leq R$,
2. $|f(x) - a| \leq R$ voor alle x met $|x - a| \leq R$,
3. er is een L met $0 < L < 1$ zo dat $|f(x') - f(x'')| \leq L|x' - x''|$ voor alle x' en x'' met $|x' - a| \leq R$, $|x'' - a| \leq R$.

Dan heeft de vergelijking $x = f(x)$ precies één oplossing α in het gebied $|x - a| \leq R$. Voor iedere x_0 met $|x_0 - a| \leq R$ convergeert het successieve substitutie proces en er geldt

$$\left|\frac{x_n - \alpha}{x_{n-1} - \alpha}\right| \leq L \,, \quad \left|\frac{\alpha - x_n}{x_n - x_{n-1}}\right| \leq \frac{L}{1 - L} \,.$$

Dit is een zg. fixed point stelling. De voorwaarde 2. zegt dat bij de afbeelding $x \to f(x)$ van alle punten uit $|x - a| \leq R$ het beeld ook in dat gebied

ligt. De voorwaarde 3. geeft aan dat de afbeelding een zg. contraherende afbeelding is. En de stelling zegt dat er in $|x - a| \leq R$ precies één vast punt α is dat op zichzelf wordt afgebeeld.
Bovendien convergeert bij ieder beginpunt x_0 de rij $x_1, x_2, \ldots$ naar α en wel minstens met een convergentiefactor L. Is L bekend dan kan een bovengrens voor de fout $|\alpha - x_n|$ afgeleid worden uit de grootte van de laatste correctie $x_n - x_{n-1}$ (vergelijk § 1.1.3).

Deze stelling die, bij geschikte interpretatie, ook in meer dimensies geldt, is een van de hoekstenen van de numerieke en de constructieve analyse.

Opmerkingen

1. Aan de voorwaarde 3. is zeker voldaan als in $|x - a| < R$ $f'(x)$ bestaat en $|f'(x)| \leq L$.
2. Aan de voorwaarde 2. is voldaan als aan 3. is voldaan en bovendien $|f(a) - a| \leq (1 - L)R$.
3. Het bewijs van de stelling in één dimensie is niet moeilijk. Uit 2. volgt (ga na) dat $f(x) - x \geq 0$ voor $x = a - R$ en $f(x) - x \leq 0$ voor $x = a + R$. $f(x) - x$ moet dus minstens één nulpunt hebben in het interval $[a-R, a+R]$. Uit 3. volgt echter dat er ook hoogstens één oplossing is: als α en α' beide oplossing waren dan was $|\alpha - \alpha'| = |f(\alpha) - f(\alpha')| \leq L|\alpha - \alpha'|$, dus $|\alpha - \alpha'| = 0$. De rest van het bewijs gaat als bij stelling 1.

1.2. Het herleiden van een vergelijking $F(x) = 0$ tot $x = f(x)$

We willen de vergelijking $F(x) = 0$ omvormen tot een vergelijking $x = f(x)$, zo, dat in een omgeving van een wortel α $|f'(x)| < 1$ is (en liefst zo klein mogelijk).
Een mogelijkheid is te nemen

$$f(x) = x - \varphi(x)F(x) ,$$

waarbij $\varphi(x) \neq 0$ in een omgeving van α. Dan impliceert $\alpha = f(\alpha)$ dat $F(\alpha) = 0$. En $f'(x) = 1 - \varphi'(x)F(x) - \varphi(x)F'(x)$, dus $f'(\alpha) = 1 - \varphi(\alpha)F'(\alpha)$, omdat $F(\alpha) = 0$. De convergentie van successieve substitutie in $x = f(x)$ is dus verzekerd als $|1 - \varphi(\alpha)F'(\alpha)| < 1$ en x_0 dicht genoeg bij α ligt. Het proces convergeert kwadratisch als $\varphi(\alpha)F'(\alpha) = 1$.

Hoe vinden we bij gegeven $F(x)$ een verstandige $\varphi(x)$?

1.2.1. Een "gezond-verstand-methode"

Het komt nogal eens voor dat we F(x) kunnen schrijven als

$$F(x) = (x - \beta)g(x) + \varepsilon h(x) ,$$

met $g(x) \neq 0$ in een omgeving van $x = \beta$ en $|\varepsilon|$ "klein". D.w.z., F(x) is te splitsen als som van een functie die in een bekend punt β een eerste orde nulpunt heeft en een kleine functie.
Het ligt dan voor de hand de vergelijking $F(x) = 0$ te herschrijven als

$$x = \beta - \varepsilon \frac{h(x)}{g(x)} ,$$

d.w.z. te nemen

$$f(x) = \beta - \varepsilon \frac{h(x)}{g(x)} = x - \frac{F(x)}{g(x)} , \quad \text{dus} \quad \varphi(x) = \frac{1}{g(x)} .$$

Er geldt dan

$$f'(x) = \varepsilon\left(\frac{h(x)}{g(x)}\right)' ,$$

en als $|\varepsilon|$ maar klein genoeg is, dan is er zeker goede convergentie.
Als beginschatting nemen we natuurlijk $x_0 = \beta$.

Voorbeeld

$$F(x) = x^3 - 3x^2 + 0.9x + 1.95.$$

We merken op dat

$$x^3 - 3x^2 + x + 2$$

een nulpunt heeft in $x = 2$ en dat

$$x^3 - 3x^2 + x + 2 = (x - 2)(x^2 - x - 1) .$$

Dus $F(x) = (x-2)(x^2-x-1) - 0.1x - 0.05$ en $F(x) = 0$ is equivalent met

$$x = 2 + \frac{0.1x + 0.05}{x^2 - x - 1} .$$

1.2.2. De koorde-methode

Kies $\varphi(x) = \frac{1}{m}$ met m zo dat

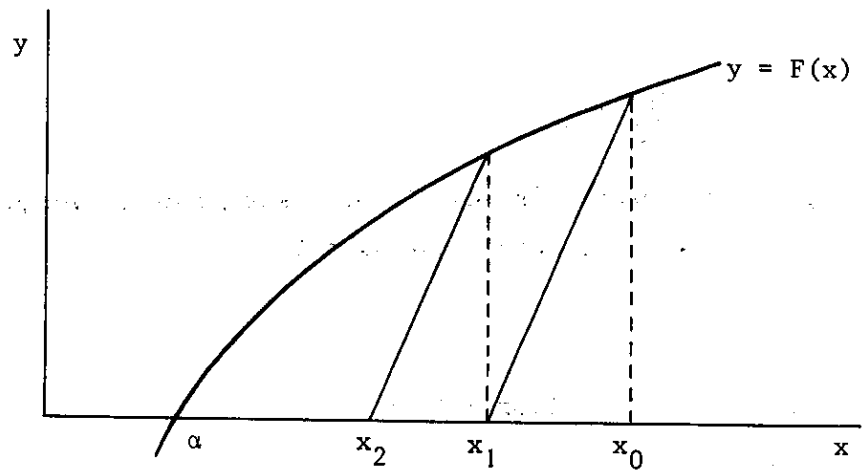
$$\left|1 - \frac{F'(\alpha)}{m}\right| < 1 \quad . \tag{3}$$

Is $F'(\alpha) > 0$ dan betekent dit dat $\frac{1}{2}F'(\alpha) < m < \infty$.
Is $F'(\alpha) < 0$ dan moet $-\infty < m < -\frac{1}{2}|F'(\alpha)|$.

Het proces

$$x_n = x_{n-1} - \frac{1}{m} F(x_{n-1}) \tag{4}$$

convergeert lineair (tenzij $m = F'(\alpha)$ - maar $F'(\alpha)$ is meestel nog onbekend!) en de asymptotische convergentiefactor is $1 - \frac{1}{m} F'(\alpha)$.

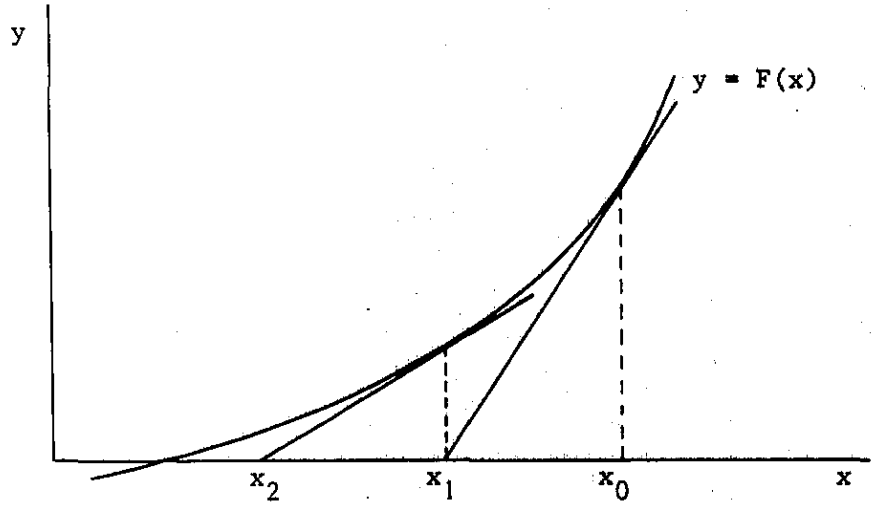


Meetkundig betekent de formule (4) dat men door het punt $(x_{n-1}, F(x_{n-1}))$ een rechte met richtingscoëfficiënt m trekt (vergelijking $y = F(x_{n-1}) + m(x - x_{n-1})$) en het snijpunt hiervan met de x-as als x_n neemt.
De conditie (3) zegt dat de helling van deze rechte meer dan half zo groot moet zijn als die van de raaklijn in $x = \alpha$ aan $y = F(x)$.

1.2.3. De iteratiemethode van Newton-Raphson

Kies $\varphi(x) = \frac{1}{F'(x)}$. De iteratieformule wordt dan

$$x_n = x_{n-1} - \frac{F(x_{n-1})}{F'(x_{n-1})} \quad . \tag{5}$$



Meetkundig betekent (5) dat men in het punt $(x_{n-1},F(x_{n-1}))$ de raaklijn aan de kromme $y = F(x)$ trekt (vergelijking: $y = F(x_{n-1}) + (x - x_{n-1})F'(x_{n-1})$) en het snijpunt hiervan met de x-as als x_n neemt.
Of nog anders gezegd: De op te lossen vergelijking is

$$F(x) = 0 \ .$$

Lineariseer deze vergelijking rond x_{n-1}, d.w.z. ontwikkel F(x) in een Taylor reeks rond x_{n-1} en laat alles behalve nulde- en eerstegraads termen weg:

$$F(x_{n-1}) + (x - x_{n-1})F'(x_{n-1}) = 0 \ .$$

De oplossing van deze gelineariseerde vergelijking nemen we als x_n.
Het is uit 1.2 duidelijk dat dit proces in het algemeen kwadratisch is, althans als $F'(\alpha) \neq 0$, want $\varphi(x)F'(x) = 1$ voor alle x.
Ook blijkt dit uit 1.1.2. Want het proces is van de vorm $x_n = f(x_{n-1})$, met

$$f(x) = x - \frac{F(x)}{F'(x)} \ . \qquad (6)$$

Hieruit volgt (ga na): als $F(\alpha) = 0$, $F'(\alpha) \neq 0$, dan is

$$f(\alpha) = \alpha \ , \quad f'(\alpha) = 0 \ , \quad f''(\alpha) = \frac{F''(\alpha)}{F'(\alpha)} \ .$$

Uit 1.1.2 volgt dan:

$$\lim_{n\to\infty} \frac{x_n - \alpha}{(x_{n-1} - \alpha)^2} = \frac{F''(\alpha)}{2F'(\alpha)} \ . \qquad (7)$$

Als $F''(\alpha) \neq 0$, dan is de convergentie kwadratisch.
Als $F''(\alpha) = 0$, dan is de convergentie orde hoger dan 2.

Opmerkingen

1) We kunnen dit eenvoudig rechtstreeks narekenen. Met de Taylorreeks volgt

$$0 = F(\alpha) = F(x) + F'(x)(\alpha - x) + \tfrac{1}{2}F''(\xi)(\alpha - x)^2 ,$$

dus $F(x) = F'(x)(x - \alpha) - \tfrac{1}{2}F''(\xi)(x - \alpha)^2$, met ξ tussen x en α.
Met (6) volgt hieruit

$$f(x) = \alpha + \frac{F''(\xi)}{2F'(x)} (x - \alpha)^2 .$$

Hieruit volgt direct (7).

2) De globale convergentie van een Newton proces is meestal moeilijk te onderzoeken. Vaak convergeert het proces alleen als x_0 dicht genoeg bij een nulpunt ligt (ga na wat er gebeurt als x_0 dicht bij een nulpunt van $F'(x)$ ligt!). Een situatie waarbij de globale convergentie verzekerd is, is de volgende: $F(a) \leq 0$, $F'(x) > 0$ en $F''(x) \geq 0$ voor $a \leq x < \infty$. Dan heeft $F(x)$ precies één nulpunt in $a \leq x < \infty$, het Newton proces convergeert voor iedere $x_0 \geq a$ en de rij $x_1, x_2, \ldots$ daalt monotoon (ga na met een plaatje). Een Newton proces kan heel langzaam convergeren als x_n nog ver van de limiet α af is.

Voorbeeld: Als $F(x) = x^2 - a$ dan convergeert het proces op grond van het bovenstaande voor iedere $x_0 > 0$ naar $\alpha = \sqrt{a}$. De formule wordt

$$x_n = x_{n-1} - \frac{x_{n-1}^2 - a}{2x_{n-1}} = \tfrac{1}{2}(x_{n-1} + a/x_{n-1}) .$$

Hieruit volgt (met $a = \alpha^2$) $x_n - \alpha = (x_{n-1} - \alpha)^2/(2x_{n-1})$. Dus - uiteraard - kwadratische convergentie. Maar zolang $x_{n-1} >> \alpha$ is $x_n - \alpha \sim \tfrac{1}{2}(x_{n-1} - \alpha)$ zodat we dan slechts lineaire convergentie hebben met factor ongeveer $\tfrac{1}{2}$. Pas als ongeveer $x_{n-1} < 3\alpha$ begint de kwadratische convergentie zichtbaar te worden.

Voorbeeld

$F(x) = x^k - a$, k geheel $\neq 0$. Dan wordt $\alpha = a^{1/k}$ en

$$x_n = x_{n-1} - \frac{x_{n-1}^k - a}{kx_{n-1}^{k-1}} = \frac{1}{k}\left[(k-1)x_{n-1} + \frac{a}{x_{n-1}^{k-1}}\right].$$

Dit is een zeer gebruikelijke methode om $\sqrt[k]{a}$ uit te rekenen.

Voor k = 2 wordt de formule

$$x_n = \tfrac{1}{2}\left[x_{n-1} + \frac{a}{x_{n-1}}\right].$$

Deze formule was al 100 jaar v. Chr. bekend (Heron).

Voor k = -1 krijgen we $x_n = x_{n-1}(2 - ax_{n-1})$. Met deze algoritme kan men dus "delen zonder te delen". Dit proces werd wel gebruikt bij automatische rekenmachines die geen ingebouwde deling hadden.

1.3. Andere iteratieve methoden

Er bestaan natuurlijk ook andere methoden voor het oplossen van een vergelijking F(x) = 0 dan successieve substitutie. We noemen er drie.

1.3.1. Interval halvering

Zij F(x) continu voor $a_0 \leq x \leq b_0$ en zij $F(a_0) < 0 < F(b_0)$ of omgekeerd. Dan heeft F(a) minstens één nulpunt in $a_0 \leq x \leq b_0$. Bepaal nu $c_0 := (a_0 + b_0)/2$ en $F(c_0)$. Als $F(c_0) = 0$ dan hebben we een nulpunt. Als $\text{sign}(F(c_0)) = \text{sign}(F(a_0))$ dan stellen we $a_1 = c_0$, $b_1 = b_0$ en anders $b_1 = c_0$, $a_1 = a_0$. In beide gevallen geldt dan weer $F(a_1) < 0 < F(b_1)$ of omgekeerd. We hebben dus een interval $[a_1, b_1]$ gevonden dat beslist een nulpunt bevat en $b_1 - a_1 = \frac{1}{2}(b_0 - a_0)$. Zo gaan we door tot $b_n - a_n \leq \varepsilon$, waarin ε de geëiste nauwkeurigheid is. Het is duidelijk dat dit proces steeds convergeert, echter slechts met een factor $\frac{1}{2}$ (als we de lengte van het interval (a_n, b_n) waarin we een nulpunt garanderen als maat voor de convergentie nemen).

1.3.2. Successieve interpolatie

Zij weer $a_0 < b_0$ en $F(a_0) < 0 < F(b_0)$ of omgekeerd. Neem nu als punt c_0 het snijpunt van de rechte

$$y = \frac{x - a_0}{b - a_0} F(b_0) + \frac{b_0 - x}{b_0 - a_0} F(a_0)$$

(dat is de rechte die door de punten $(a_0,F(a_0))$ en $(b_0,F(b_0))$ van de grafiek van $y = F(x)$ gaat) met de x-as:

$$c_0 = \frac{a_0 F(b_0) - b_0 F(a_0)}{F(b_0) - F(a_0)}$$

$$= b_0 - F(b_0) \frac{b_0 - a_0}{F(b_0) - F(a_0)} . \qquad (1)$$

En verder handelen we net als bij de interval halvering. Dan hebben we ook steeds convergentie. De convergentiefactor kan van alles zijn, dicht bij 1 of dicht bij 0. Goede convergentie hebben we als zowel a_n als b_n beide tot α naderen, want dan nadert $(b_n - a_n)/(F(b_n) - F(a_n))$ tot $F'(\alpha)$ en dan lijkt formule (1) op die van Newton. Als regel blijft echter op de duur of a_n of b_n vast. Het is dan voordelig het proces te combineren met interval halvering.

1.3.3. Regula Falsi

Als men twee benaderingen x_n en x_{n-1} voor een nulpunt α van $F(x)$ heeft dan kan men als volgende benadering weer nemen het snijpunt

$$x_{n+1} = x_n - F(x_n) \frac{x_n - x_{n-1}}{F(x_n) - F(x_{n-1})}$$

van de rechte door de punten $(x_n,F(x_n))$ en $(x_{n-1},F(x_{n-1}))$ met de x-as. En zo voort. We hebben dan interpolatie als $F(x_n)$ en $F(x_{n-1})$ verschillend teken hebben en extrapolatie in het andere geval. Omdat, in tegenstelling tot de successieve interpolatie, ook extrapolatie kan voorkomen, kunnen we weinig over de globale convergentie zeggen (ga na wat er gebeurt als vrijwel $F(x_n) = F(x_{n-1})$!). Omdat we echter steeds inter- of extrapoleren op basis van de twee laatst gevonden benaderingen, convergeert het proces hard. Men kan bewijzen dat (als $F'(\alpha) \neq 0$)

$$\lim \frac{x_{n+1} - \alpha}{(x_n - \alpha)(x_{n-1} - \alpha)} = \frac{F''(\alpha)}{2F'(\alpha)}$$

(vergelijk dit met de formule die voor het proces van Newton geldt) en dat hieruit volgt dat

$$\lim \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = \left|\frac{F''(\alpha)}{2F'(\alpha)}\right|^{p-1} ,$$

waarin $p = \frac{1}{2}(1 + \sqrt{5}) = 1.62$. De convergentie orde is dus 1.62. Dat is dus minder snel dan bij Newton (waar we werkten met de raaklijn in het punt $(x_n, F(x_n))$), maar wel essentieel meer dan lineair. En men hoeft $F'(x_n)$ niet te berekenen.
Om de globale convergentie te verzekeren kan men het proces combineren met successieve interpolatie en/of interval halvering.

1.4. Stelsels vergelijkingen

Naast één vergelijking met één onbekende ontmoet men ook stelsels van k vergelijkingen met k onbekenden, bv.

$$F_1(x_1, x_2, \ldots, x_k) = 0$$
$$F_2(x_1, x_2, \ldots, x_k) = 0$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$F_k(x_1, x_2, \ldots, x_k) = 0 .$$

We spreken van lineaire vergelijkingen indien de functies F_i (als regel inhomogeen) lineair zijn in $x_1, \ldots, x_k$, dus als $F_i(x_1, \ldots, x_k) = \sum_{j=1}^{k} A_{ij} x_j - b_i$.

Dit type vergelijkingen wordt in hoofdstuk 2 uitvoerig besproken.

We beperken ons hier verder tot $k = 2$ en schrijven de vergelijkingen als

$$F(x,y) = 0 , \quad G(x,y) = 0 . \qquad (1)$$

Merk op dat $z = F(x,y)$ een oppervlak in R_3 voorstelt en dat $F(x,y) = 0$ de snijkromme van dit oppervlak met het vlak $z = 0$ is. Analoog $G(x,y) = 0$.
Zij $x = \alpha$, $y = \beta$ een oplossing van (1). Dan worden de raaklijnen door dit punt aan $F(x,y) = 0$, resp. $G(x,y) = 0$ gegeven door resp.
$(x - \alpha)F_x(\alpha,\beta) + (y - \beta)F_y(\alpha,\beta) = 0$, $(x - \alpha)G_x(\alpha,\beta) + (y - \beta)G_y(\alpha,\beta) = 0$

(met $F_x = \frac{\partial F}{\partial x}$, etc.). Als

$$\begin{vmatrix} F_x(\alpha,\beta) & F_y(\alpha,\beta) \\ G_x(\alpha,\beta) & G_y(\alpha,\beta) \end{vmatrix} = \frac{\partial(F,G)}{\partial(x,y)}\Big|_{(\alpha,\beta)} = 0 \ , \qquad (2)$$

dan vallen de raaklijnen samen. Dit geval, dat moeilijker te behandelen is, sluiten we verder uit.
Vaak is het prettig om (x,y) als vector $\underline{x}$ op te vatten en (F(x,y),G(x,y)) als vector functie $\underline{F}(\underline{x})$. Analoog schrijven we $\underline{\alpha} = (\alpha,\beta)$.

1.4.1. Successieve substitutie

Stel de vergelijkingen (1) herschreven als

$$x = f(x,y) \ , \quad y = g(x,y) \ , \quad \text{of} \quad \underline{x} = \underline{f}(\underline{x}) \ . \qquad (3)$$

Men kan dit, uitgaande van (1), bv. verkrijgen door te nemen

$$\begin{aligned} f(x,y) &= x - A(x,y)F(x,y) - B(x,y)G(x,y) \\ g(x,y) &= y - C(x,y)F(x,y) - D(x,y)G(x,y) \end{aligned} \qquad (4)$$

waarin de functies A, B, C en D zo moeten zijn dat in een omgeving van (α,β)

$$\begin{vmatrix} A(x,y) & B(x,y) \\ C(x,y) & D(x,y) \end{vmatrix} \neq 0$$

(zodat uit $\alpha = f(\alpha,\beta)$, $\beta = g(\alpha,\beta)$ noodzakelijk volgt dat $F(\alpha,\beta) = G(\alpha,\beta) = 0$).
In vectorvorm: $\underline{f}(\underline{x}) = \underline{x} - M(\underline{x})F(\underline{x})$ waarin $M(\underline{x})$ de matrix $\begin{pmatrix} A(x,y) & B(x,y) \\ C(x,y) & D(x,y) \end{pmatrix}$ is, die in de buurt van $\underline{\alpha}$ niet singulier mag zijn.
Kies nu een beginschatting (x_0,y_0) en bepaal volgende benaderingen (x_n,y_n) (n = 1,2,...) door

$$x_{n+1} = f(x_n,y_n) \ , \quad y_{n+1} = g(x_n,y_n) \ , \quad \text{dus} \quad \underline{x}_{n+1} = \underline{f}(\underline{x}_n) \ . \qquad (5)$$

Het is duidelijk dat als de rijen $\{x_n\}$ en $\{y_n\}$ limieten α, resp. β hebben (α,β) een oplossing van (3) is (als f en g continu zijn).
Men kan voorwaarden afleiden die de convergentie van het proces (5) verzekeren. Bijvoorbeeld geldt:

Locale convergentie stelling

Als (α,β) een oplossing van (3) is en voor ieder tweetal punten (x',y') en (x'',y'') uit een omgeving van (α,β) geldt

$$\begin{aligned} |f(x',y') - f(x'',y'')| &\leq L \max\{|x' - x''|, |y' - y''|\} \\ |g(x',y') - g(x'',y'')| &\leq L \max\{|x' - x''|, |y' - y''|\} \end{aligned} \tag{6}$$

met $L < 1$, dan convergeert het proces (4) voor iedere beginschatting (x_0,y_0) uit deze omgeving.

Het bewijs van deze stelling is analoog aan dat van de stelling uit 1.1.1. Deze analogie kan men zeer fraai maken door het volgende begrip afstand van twee "punten" $\underline{x}' = (x',y')$ en $\underline{x}'' = (x'',y'')$ in te voeren:

$$\|\underline{x}' - \underline{x}''\| = \max\{|x' - x''|, |y' - y''|\} .$$

Deze afstand is positief, tenzij $\underline{x}'$ en $\underline{x}''$ samenvallen, en voldoet aan de zg. driehoeksongelijkheid

$$\|\underline{x}' - \underline{x}'''\| \leq \|\underline{x}' - \underline{x}''\| + \|\underline{x}'' - \underline{x}'''\| .$$

Hiermee luidt (6)

$$\|\underline{f}(\underline{x}') - \underline{f}(\underline{x}'')\| \leq L\|\underline{x}' - \underline{x}''\| .$$

Hiermee vinden we nu

$$\|\underline{x}_n - \underline{\alpha}\| = \|\underline{f}(\underline{x}_{n-1}) - \underline{f}(\underline{\alpha})\| \leq L\|\underline{x}_{n-1} - \underline{\alpha}\| \leq L^n\|\underline{x}_0 - \underline{\alpha}\| .$$

En ook

$$\|x_n - \alpha\| \leq L\|x_{n-1} - \alpha\| \leq L(\|x_{n-1} - x_n\| + \|x_n - \alpha\|) ,$$

dus

$$\|x_n - \alpha\| \leq \frac{L}{1 - L}\|x_{n-1} - x_n\| . \tag{7}$$

Het proces is dus in het algemeen van de eerste orde met convergentiefactor $\leq L$. En uit (7) vinden we weer een relatie tussen de onbekende afstand $\|x_n - \alpha\|$ en de bekende afstand $\|x_{n-1} - x_n\|$. Vergelijk deze relatie ook met die van de stelling uit 1.1.5.

1.4.2. De methode van Newton

We gaan uit van de vergelijkingen (1). Zij (α,β) een oplossing en (x_0,y_0) een naburig punt. Dan kan men schrijven (Taylorreeks)

$$F(x,y) = F(x_0,y_0) + (x - x_0)F_x(x_0,y_0) + (y - y_0)F_y(x_0,y_0) + \ldots$$

$$G(x,y) = G(x_0,y_0) + (x - x_0)G_x(x_0,y_0) + (y - y_0)G_y(x_0,y_0) + \ldots ,$$

waarin $F_x = \frac{\partial F}{\partial x}$, etc.

We vervangen nu de vergelijkingen (1) door de gelineariseerde vergelijkingen

$$\begin{aligned} F(x_0,y_0) + (x - x_0)F_x(x_0,y_0) + (y - y_0)F_y(x_0,y_0) &= 0 \\ G(x_0,y_0) + (x - x_0)G_x(x_0,y_0) + (y - y_0)G_y(x_0,y_0) &= 0 . \end{aligned} \tag{8}$$

Of, in vectorvorm,

$$\underline{F}(\underline{x}_0) + \underline{F}'(\underline{x}_0)(\underline{x} - \underline{x}_0) = \underline{0} .$$

waarin $\underline{F}'(\underline{x})$ de matrix

$$\begin{pmatrix} F_x(x,y) & F_y(x,y) \\ G_x(x,y) & G_y(x,y) \end{pmatrix}$$

is. De oplossing van (8):

$$\underline{x} = \underline{x}_0 - (\underline{F}'(\underline{x}_0))^{-1}\ \underline{F}(\underline{x}_0) \tag{9}$$

noemen we $\underline{x}_1$. Etc.

Men kan weer bewijzen dat dit proces in het algemeen de orde twee heeft.

Opmerkingen

1) Het is duidelijk dat moeilijkheden ontstaan indien $F_xG_y - G_xF_y = \frac{\partial(F,G)}{\partial(x,y)} = 0$ in een der punten (x_n,y_n), want dan is de matrix $\underline{F}'(\underline{x})$ hier singulier. Vergelijking met formule (2) uit 1.4 leert dat dit (in de buurt van (α,β) en als F en G continu differentieerbaar zijn) niet kan gebeuren als de kromme $F(x,y) = 0$ en $G(x,y) = 0$ elkaar in (α,β) niet raken.
2) De formule (9) correspondeert met 1.4.1 indien men daar voor de matrix M neemt de inverse van de matrix $\underline{F}'(\underline{x})$.

Zij nl. het aantal vermenigvuldigingen *) nodig om met behulp van deze regels een $n \times n$-determinant uit te rekenen $f(n)$. Dan is kennelijk

$$f(n) = n + nf(n-1) , \quad (n \geq 2) \quad \text{en} \quad f(1) = 0 .$$

Om uit deze recursiebetrekking $f(n)$ te bepalen stellen we $f(n) = n!\, g(n)$. Dan moet

$$g(n) - g(n-1) = \frac{1}{(n-1)!} \quad (n \geq 2) \quad \text{en} \quad g(1) = 0 ,$$

waaruit volgt dat voor $n \geq 2$

$$1 \leq g(n) = \frac{1}{1!} + \frac{1}{2!} + \ldots + \frac{1}{(n-1)!} < e - 1$$

en dus

$$n! \leq f(n) < (e-1) \cdot n! .$$

Voor het uitrekenen van een 20×20-determinant zouden dus ca $20! \sim 2.4 \times 10^{17}$ vermenigvuldigingen nodig zijn. Met een snelle automatische machine met een vermenigvuldigtijd van 10^{-6} sec. zou men dus ca 2.4×10^{11} sec $\sim 10^4$ jaar nodig hebben!

Later zullen we een methode aangeven, waarbij voor de berekening van een $n \times n$-determinant slechts ca $\frac{1}{3} n^3$ vermenigvuldigingen nodig zijn. Ook dan blijft de regel van Cramer niet aanbevelenswaardig aangezien een $n \times n$-stelsel vergelijkingen ook met ca $\frac{1}{3} n^3$ vermenigvuldigingen opgelost blijkt te kunnen worden.

Behalve naar de oplossing van het stelsel (1) dat verkort geschreven kan worden als

$$\underline{Ax} = \underline{b} \tag{1a}$$

kan men ook vragen naar de inverse van de matrix A, dat is een matrix A^{-1} zodanig dat

$$AA^{-1} = A^{-1} A = I \tag{2}$$

---

*) De tijd nodig voor een vermenigvuldiging is - zowel bij het rekenen uit het hoofd, met een tafelmachine of met een automatische rekenmachine - essentieel langer dan die nodig voor een optelling of aftrekking. Daarom telt men meestal alleen het aantal nodige vermenigvuldigingen (en delingen, die meestal met vermenigvuldigingen over één kam geschoren worden).

2. Lineaire vergelijkingen

2.1. Inleiding

In dit hoofdstuk behandelen wij het oplossen van stelsels lineaire vergelijkingen van de vorm

$$\begin{aligned} a_{11}\,x_1 + a_{12}\,x_2 + \ldots + a_{1n}\,x_n &= b_1 \\ a_{21}\,x_1 + a_{22}\,x_2 + \ldots + a_{2n}\,x_n &= b_2 \\ \text{----------} & \\ a_{n1}\,x_1 + a_{n2}\,x_2 + \ldots + a_{nn}\,x_n &= b_n\ . \end{aligned} \qquad (1)$$

Voor de zuiver wiskundige is hier geen probleem aangezien volgens de regel van Cramer (1750!) de oplossing gegeven wordt door

$$x_j = \frac{\begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & b_1 & a_{1,j+1} & \cdots & a_{1n} \\ \text{---} & & & & & & \\ a_{1n} & \cdots & a_{n,j-1} & b_n & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix}}{\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \text{---} & & \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}}\ , \qquad j = 1,\ldots,n$$

en de berekening van determinanten volkomen bepaald is door de regels dat

$$\begin{vmatrix} c_{11} & \cdots & c_{1n} \\ \text{---} & & \\ c_{n1} & \cdots & c_{nn} \end{vmatrix} = \sum_{j=1}^{n} (-1)^{j-1}\, c_{1j} \cdot \begin{vmatrix} c_{21} & \cdots & c_{2,j-1} & c_{2,j+1} & \cdots & c_{2n} \\ \text{---} & & & & & \\ c_{n1} & \cdots & c_{n,j-1} & c_{n,j+1} & \cdots & c_{nn} \end{vmatrix}$$

(ontwikkeling naar de eerste rij waardoor de berekening van n × n-determinanten teruggebracht is tot de berekening van (n-1) × (n-1)-determinanten) en $|c_{11}| = c_{11}$ (berekening van een 1 × 1-determinant).

De numericus is echter met deze algoritme, waarin is aangegeven hoe de oplossing door eindig veel bewerkingen op de coëfficiënten van (1) gevonden kan worden, niet volledig gelukkig. Want toepassing van deze regels leidt, als n enigszins groot is, tot een astronomisch groot aantal bewerkingen.

(waarin I de n × n-eenheidsmatrix is). In componenten geschreven luidt dit

$$\sum_{j=1}^{n} A_{ij}(A^{-1})_{jk} = \sum_{j=1}^{n} (A^{-1})_{ij} A_{jk} = \delta_{ik} , \qquad (2a)$$

waarin

$$\delta_{ik} = \begin{cases} 1 & \text{als } i = k \\ 0 & \text{als } i \neq k \end{cases} \qquad \text{(Kronecker-symbool).}$$

Kent men de matrix A^{-1} dan is de oplossing van het stelsel (1a) ook direct uit te rekenen: namelijk $\underline{x} = A^{-1}\underline{b}$. In de praktijk zal men echter de oplossing van (1a) nooit op deze manier berekenen.

Omgekeerd kan men A^{-1} berekenen door n stelsels van het type (1a) op te lossen. Zijn nl. $\underline{x}_1,\ldots,\underline{x}_n$ de oplossingen van de stelsels

$$A\underline{x}_k = \underline{e}_k , \qquad k = 1,\ldots,n ,$$

waarin $\underline{e}_k$ de k-de eenheidsvector is ($(\underline{e}_k)_i = \delta_{ik}$), dan zijn $\underline{x}_1,\ldots,\underline{x}_n$ de kolommen van de matrix A^{-1}.

In het voorgaande is steeds verondersteld dat de matrix A niet singulier is. Zoals bekend is deze voorwaarde gelijkwaardig met de voorwaarde dat de homogene vergelijking $A\underline{x} = \underline{o}$ uitsluitend $\underline{x} = \underline{o}$ als oplossing heeft en ook met de voorwaarde dat de determinant van A niet nul is.
Een goede standaard-algoritme voor het oplossen van lineaire vergelijkingen zal moeten onderzoeken of een aangeboden matrix singulier of vrijwel singulier is en dan een waarschuwing moeten geven.

2.2. Directe methoden

2.2.1. Triangulaire stelsels

Beschouw een stelsels vergelijkingen $U\underline{x} = \underline{c}$, waarin U een zg. boven-driehoeksmatrix is, d.w.z. $U_{ij} = 0$ voor $j < i$. Het stelsel ziet er dan uit als

$$\left.\begin{aligned} U_{11}x_1 + U_{12}x_2 + \ldots + U_{1n}x_n &= c_1 \\ U_{22}x_2 + \ldots + U_{2n}x_n &= c_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ U_{nn}x_n &= c_n \end{aligned}\right\} . \qquad (1)$$

We nemen aan dat geen der diagonaalelementen U_{jj} $(1 \leq j \leq n)$ nul is (anders was det(U) = 0 !). Dan is het stelsel onmiddellijk op te lossen:

$$x_n = c_n/U_{nn}$$

$$x_{n-1} = (c_{n-1} - U_{n-1,n}\, x_n)/U_{n-1,n-1}$$

$$x_k = (c_k - \sum_{j=k+1}^{n} U_{kj}\, x_j)/U_{kk} \ .$$

In pseudo-ALGOL:

```
for k := n step -1 until 1 do
begin s := c_k;
      for j := k + 1 step 1 until n do s := s - U_kj × x_j;
      x_k := s/U_kk
end
```

Men noemt dit proces wel terugsubstitutie (eerst x_n bepalen uit de laatste vergelijking, deze waarde substitueren in de voorlaatste vergelijking, etc.).

Opgaven

1) Laat zien dat de uitvoering van deze algoritme $\frac{1}{2}n(n + 1)$ vermenigvuldigingen en delingen vraagt.

2) Laat zien dat men (als men na afloop van het proces niet meer in de rechterleden c_k geïnteresseerd is) de getallen x_k zonder bezwaar op de plaatsen van de getallen c_k kan schrijven.
   D.w.z., na uitvoering van

```
for k := n step -1 until 1 do
begin s := c_k;
      for j := k + 1 step 1 until n do s := s - U_kj × c_j;
      c_k := s/U_kk
end
```

   bevatten de elementen c_1 t/m c_n de oplossing!

3) Behandel de oplossing van een stelsel $L\underline{x} = \underline{b}$, waarin L een onder-driehoeksmatrix is ($L_{ij} = 0$ voor $j > i$).

2.2.2. De eliminatiemethode van Gauss

Beschouw nu een algemeen stelsel $\underline{A}\underline{x} = \underline{b}$, of

$$\left.\begin{array}{l} A_{11}\,x_1 + \ldots + A_{1n}\,x_n = b_1 \\ \hline A_{n1}\,x_1 + \ldots + A_{nn}\,x_n = b_n \end{array}\right\}. \qquad (1)$$

Kunnen we dit stelsel tot een triangulaire vorm brengen? Dit kan met de eliminatiemethode van Gauss (ook wel vegen genoemd).

Stel dat $A_{11} \neq 0$. Dan nemen we de eerste vergelijking van (1) als eerste vergelijking van het triangulaire stelsel. Op didactische gronden stellen we $U_{1j} := A_{1j}$ $(j = 1,\ldots,n)$ en $c_1 := b_1$. Definieer nu voor $i = 2,\ldots,n$ $L_{i1} := A_{i1}/U_{11}$, vermenigvuldig de eerste vergelijking met L_{i1} en trek dit af van de i-de vergelijking.
Dan ontstaat het volgende stelsel vergelijkingen

$$\left.\begin{array}{l} U_{11}\,x_1 + U_{12}\,x_2 + \ldots + U_{1n}\,x_n = c_1 \\ A_{22}^{(1)}x_2 + \ldots + A_{2n}^{(1)}x_n = b_2^{(1)} \\ \hline A_{n2}^{(1)}x_2 + \ldots + A_{nn}^{(1)}x_n = b_n^{(1)} \end{array}\right\}. \qquad (2)$$

Hierin is

$$A_{ij}^{(1)} := A_{ij} - L_{i1} \times U_{1j}\,, \quad i,j \geq 2\,,$$

$$b_i^{(1)} := b_i - L_{i1} \times c_1\,, \quad i \geq 2\,.$$

Dit stelsel, waarin x_1 alleen nog in de eerste vergelijking voorkomt, is equivalent met (1), d.w.z. (1) en (2) hebben dezelfde oplossing.

Behandel nu de laatste n-1 vergelijkingen van (2) op dezelfde manier als het stelsel (1). Stel $A_{22}^{(1)} \neq 0$. Dan schrijven we $U_{2j} := A_{2j}^{(1)}$ $(j = 2,\ldots,n)$ en $c_2 := b_2^{(1)}$.
Verder definiëren we voor $i = 3,\ldots,n$ $L_{i2} := A_{i2}^{(1)}/U_{22}$, vermenigvuldigen de tweede vergelijking met L_{i2} en trekken dit af van de i-de vergelijking. Dan ontstaat een stelsel

$$\left.\begin{aligned} U_{11}\,x_1 + U_{12}\,x_2 + U_{13}\,x_3 + \ldots + U_{1n}\,x_n &= c_1 \\ U_{22}\,x_2 + U_{23}\,x_3 + \ldots + U_{2n}\,x_n &= c_2 \\ A^{(2)}_{33}x_3 + \ldots + A^{(2)}_{3n}x_n &= b^{(2)}_3 \\ \text{----------} & \\ A^{(2)}_{n3}x_3 + \ldots + A^{(2)}_{nn}x_n &= b^{(2)}_n \end{aligned}\right\} .$$

Zo gaan we door, aannemende dat geen der kop-elementen (pivots genaamd) A_{11}, $A^{(1)}_{22}$, $A^{(2)}_{33}$,... nul is. Tenslotte ontstaat dan een triangulair stelsel dat equivalent is met (1):

$$\begin{aligned} U_{11}\,x_1 + U_{12}\,x_2 + \ldots + U_{1n}\,x_n &= c_1 \\ U_{22}\,x_2 + \ldots + U_{2n}\,x_n &= c_2 \\ \text{----------} & \\ U_{nn}\,x_n &= c_n \quad . \end{aligned}$$

Uit dit stelsel kunnen nu door terugsubstitutie $x_n, x_{n-1}, \ldots, x_1$ bepaald worden.

Bovendien geldt (ga dit na met de regels voor bewerkingen op determinanten)

$$\det A = \det U = U_{11}\,U_{22} \ldots U_{nn} \quad .$$

In de praktische uitvoering van de algoritme schrijft men natuurlijk de opvolgende stelsels coëfficiënten A_{ij}, $A^{(1)}_{ij}$, $A^{(2)}_{ij}$,... over elkaar heen.
De algoritme kan dan luiden

```
for k := 1 step 1 until n do
begin for j := k step 1 until n do U[k,j] := A[k,j];
      c[k] := b[k];
      for i := k + 1 step 1 until n do
      begin L[i,k] := A[i,k]/U[k,k];
            for j := k + 1 step 1 until n do A[i,j] := A[i,j] - L[i,k] × U[k,j];
            b[i] := b[i] - L[i,k] × c[k]
      end
end
```

Opgaven

1) Laat zien dat de uitvoering van deze algoritme $\frac{1}{3}n(n^2 - 1)$ vermenigvuldigingen en delingen eist voor bewerking van de coëfficiënten matrix en $\frac{1}{2}n(n - 1)$ vermenigvuldigingen voor bewerking van de rechterleden. Samen met het oplossen van het triangulaire stelsel zijn er dus $\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$, dus ca $\frac{1}{3}n^3$ vermenigvuldigingen en delingen nodig voor het oplossen van een $n \times n$ stelsel.

2) Ga na hoe men de algoritme kan wijzigen indien men meerdere stelsels heeft op te lossen die alle dezelfde matrix A, doch verschillende rechterleden hebben.

3) Laat zien dat men zonder bezwaar de elementen L_{ij} $(j < i)$ en U_{ij} $(j \geq i)$ op de plaatsen van de elementen A_{ij} kan schrijven en de elementen c_i op de plaatsen van de elementen b_i, zodat de algoritme wordt:

```
for k := 1 step 1 until n do
for i := k + 1 step 1 until n do
begin A[i,k] := A[i,k]/A[k,k];
      for j := k + 1 step 1 until n do A[i,j] := A[i,j] - A[i,k] × A[k,j];
      b[i] := b[i] - A[i,k] × b[k]
end
```

4) Laat zien dat dankzij het feit dat de vermenigvuldigers L_{ik} bewaard worden, men de bewerking van de rechterleden ook kan doen na voltooiing van de bewerking van de matrix:

```
for k := 1 step 1 until n do
begin c[k] := b[k];
      for i := k + 1 step 1 until n do b[i] := b[i] - L[i,k] × c[k]
end
```

2.2.3. Pivot strategieën

In 2.2.2 is aangenomen dat de hoekelementen $a_{11}, a_{22}^{(1)}, \ldots$ van de gereduceerde stelsels (de zg. "pivots", dat zijn de "spillen", waar de eliminatie om draait) alle $\neq 0$ zijn. Dit is natuurlijk geenszins steeds het geval.

Bij het stelsel

$$\begin{aligned} 2x_1 - 2x_2 + x_3 &= 1 \\ x_1 - x_2 + x_3 &= 1 \\ x_1 + x_2 &= 4 \end{aligned}$$

krijgen we na de eerste slag van de eliminatie

$$\begin{aligned} 2x_1 - 2x_2 + x_3 &= 1 \\ .5x_3 &= .5 \\ 2x_2 - .5x_3 &= 3.5 \ , \end{aligned}$$

zodat $a_{22}^{(1)} = 0$. De remedie ligt voor de hand: we moeten de tweede en de derde vergelijking verwisselen.

In het algemeen: als $a_{kk}^{(k-1)} = 0$, dan zoeken we een element $a_{pk}^{(k-1)}$ $(p > k)$ dat niet nul is en we verwisselen de p-de en de k-de vergelijking.

N.b.: Als $a_{pk}^{(k-1)} = 0$ voor alle $p \geq k$, dan was het stelsel vergelijkingen afhankelijk (ga na!).

Ook als een pivot niet exact nul, doch "klein" is, ontstaan moeilijkheden. Beschouw het stelsel

$$\left.\begin{aligned} .0102x_1 + .9617x_2 &= .8754 \\ -.8813x_1 + .9753x_2 &= .0674 \end{aligned}\right\} . \qquad (1)$$

Nemen we het element $a_{11} = .0102$ als pivot bij de eliminatie van x_1 dan vinden we als gereduceerd stelsel

$$\left.\begin{aligned} .0102x_1 + .9617x_2 &= .8754 \\ 84.0681x_2 &= 75.7037 \end{aligned}\right\} . \qquad (2)$$

Uit de laatste vergelijking volgt

$$x_2 = .900504 \qquad (3a)$$

en terugsubstitutie in de eerste vergelijking van (2) levert

$$.0102x_1 = .8754 - .866015 = .009385 ,$$

waaruit volgt

$$x_1 = .9201 . \tag{3b}$$

We hebben hier voortdurend met "voldoende veel" cijfers gerekend en daarom is het antwoord in 4 cijfers goed. Zouden we echter consequent in 4 cijfers gerekend hebben, dan vinden we in plaats van (2) als gereduceerd stelsel

$$\left.\begin{aligned} .0102x_1 + .9617x_2 &= .8754 \\ 84.07\ x_2 &= 75.70 \end{aligned}\right\} . \tag{4}$$

Bepalen we hiervan - alle tussenresultaten in 4 cijfers afrondend - de oplossing dan vinden we

$$x_1 = .9314 , \quad x_2 = .9004 . \tag{5}$$

Dit wijkt aanzienlijk af van de oplossing (3). De exacte oplossing van (4), afgerond in 4 cijfers, is

$$x_1 = .9262 , \quad x_2 = .9004 . \tag{6}$$

Dit wijkt zowel van (3) als van (5) af.

De oorzaak van de verschillen tussen (3), (5) en (6) ligt in het feit dat de stelsels (2) en (4) zeer slecht geconditioneerd zijn (althans t.a.v. de bepaling van x_1). Dit verklaart

a) dat de exacte oplossingen van (2) en (4) al in het derde cijfer verschillen, hoewel de coëfficiënten van (2) en (4) in 4 cijfers gelijk zijn.

b) dat men van de oplossing van (4) alleen 4 goede cijfers kan vinden indien x_2 in 6 cijfers berekend en teruggesubstitueerd wordt; werkt men consequent in 4 cijfers (dus met een relatieve nauwkeurigheid van 1 op 10^4) dan vindt men een resultaat dat al in het derde cijfer van de exacte oplossing afwijkt (een relatieve fout van ca 50 op 10^4).

Dat de toch zo naburige stelsels (2) en (4), die beide afgeleid zijn van het stelsel (1), zo verschillende oplossingen hebben, kan men ook als volgt begrijpen. De tweede vergelijking van (4) is verkregen door 86.40 (dat is, in 4 cijfers, .8813/.0102) maal de eerste vergelijking van (1) van de tweede

vergelijking van (1) af te trekken en het resultaat op 4 cijfers af te ronden. Tellen we echter bij de tweede vergelijking van (4) weer 86.40 maal de eerste vergelijking van (4) op, dan vinden we dat het stelsel (4) exact equivalent is, niet met (1), maar met

$$\left.\begin{aligned} .0102\,x_1 + .9617\,x_2 &= .8754 \\ -.88128x_1 + .97912x_2 &= .06544 \end{aligned}\right\} \quad . \tag{7}$$

Sommige coëfficiënten hiervan wijken in het derde cijfer af van die van (1)! Dit is een gevolg van het feit dat de coëfficiënten van de tweede vergelijking van (4) ca 100 maal groter zijn dan die van de tweede vergelijking van (1). Hierdoor verliezen we bij de afronding op 4 significante cijfers het derde en vierde cijfer achter de punt, zodat we bij terugrekenen als tweede vergelijking in (7) iets vinden dat in het derde cijfer afwijkt van de tweede vergelijking in (1). Het verschil tussen (7) en (1) (deze stelsels zijn beide goed geconditioneerd) maakt het aannemelijk dat de oplossingen van deze stelsels (nl. (6), resp. (3)) ook in het derde cijfer afwijken.

Indien we in het stelsel (1) de vergelijkingen verwisselen dan verloopt alles zonder moeilijkheden. Als gereduceerd stelsel krijgen we

$$\begin{aligned} -.8813x_1 + .9753x_2 &= .0674 \\ .9730x_2 &= .8762 \ , \end{aligned}$$

waaruit volgt

$$x_1 = .9201 \ , \quad x_2 = .9005 \ .$$

Dit laat zien dat we, rekenend in 4 cijfers, de oplossing van het goed geconditioneerde stelsel (1) in 4 cijfers correct kunnen vinden, mits de goede pivot gekozen wordt.

In het algemeen blijkt dat de volgende pivotstrategie verstandig is.

1) Bepaal de grootste van de getallen

$$|a_{11}|, |a_{21}|, \ldots, |a_{n1}| \ .$$

Als dit $|a_{p,1}|$ is, verwissel dan de eerste en de p-de vergelijking.

2) Voer de eerste stap van de Gauss-algoritme uit (met de nieuwe a_{11} - d.w.z. de oude $a_{p,1}$ - als pivot).

3) Bepaal de grootste van de getallen

$$|a_{22}^{(1)}|, |a_{32}^{(1)}|, \ldots, |a_{n2}^{(1)}| \ .$$

Als dit $a_{p,2}^{(1)}$ is (met als regel een andere waarde voor p dan bij de eerste slag), verwissel dan de tweede en de p-de rij.

4) Voer de tweede stap van de Gauss-algoritme uit. Etc.

In ALGOL kunnen we dit als volgt opschrijven (we nemen aan dat de index p van de rij die, voorafgaande aan de k-de stap van de Gauss-algoritme, met de k-de rij verwisseld wordt, genoteerd wordt in het array element p[k]).

```
for k := 1 step 1 until n do
begin pk := k; max := abs(A[k,k]);
      for i := k + 1 step 1 until n do
      if abs(A[i,k]) > max then begin pk := i; max := abs(A[i,k]) end;
      p[k] := pk;
      for j := k step 1 until n do
      begin U[k,j] := A[pk,j]; A[pk,j] := A[k,j] end;
      for i := k + 1 step 1 until n do
      begin L[i,k] := A[i,k]/U[k,k];
            for j := k + 1 step 1 until n do A[i,j] := A[i,j] - L[i,k] × U[k,j]
      end
end
```

Merk op dat reeds berekende elementen L_{ij} niet meeverwisseld worden. Dit is prettig voor de behandeling achteraf van een rechterlid, die kan luiden:

```
for k := 1 step 1 until n do
begin c[k] := b[p[k]]; b[p[k]] := b[k];
      for i := k + 1 step 1 until n do b[i] := b[i] - L[i,k] × c[k]
end;
for k := n step -1 until 1 do
begin x[k] := c[k]/U[k,k];
      for i := k - 1 step -1 until 1 do c[i] := c[i] - U[i,k] × x[k]
end
```

Ga na dat dit correct is (we hebben het oplossen van het driehoeksstelsel $U\underline{x} = \underline{c}$ anders geschreven dan in 2.2.1, welke manier zou beter zijn?).

2.2.4. Schaling

De oplossing van een stelsel $A\underline{x} = \underline{b}$ verandert niet als men de vergelijkingen met factoren resp. $p_1, p_2, \ldots, p_n$ vermenigvuldigt. Dit komt neer (ga na) op vervanging van het stelsel door

$$PA\underline{x} = P\underline{b} ,$$

waarin $P = \text{diag}(p_1, \ldots, p_n)$.

In dit geval worden de rijen van A met resp. $p_1, p_2, \ldots, p_n$ vermenigvuldigd. Men spreekt van rijschaling van de matrix A.
Ook kan men $A\underline{x} = \underline{b}$ vervangen door

$$(AQ)\underline{y} = \underline{b} , \quad \underline{x} = Q\underline{y} ,$$

met $Q = \text{diag}(q_1, \ldots, q_n)$. Dit komt neer op vermenigvuldiging van de kolommen van A met resp. $q_1, \ldots, q_n$. Na oplossing van $(AQ)\underline{y} = \underline{b}$ is dan $\underline{x} = Q\underline{y}$, of $x_i = q_i y_i$. Dit heet kolomschaling.

Algemeen kunnen we dus $A\underline{x} = \underline{b}$ vervangen door

$$(PAQ)\underline{y} = P\underline{b} , \quad \underline{x} = Q\underline{y}$$

(rij- èn kolomschaling).

Men kan de vervanging van A door $A' := PAQ$ gebruiken om te zorgen dat alle rijen en kolommen van A' ongeveer dezelfde grootte-orde hebben, bv. zo: dat voor alle i en j $|A'_{ij}| \leq 1$, terwijl er in iedere rij en in iedere kolom een element is met absolute waarde ≥ 0.5 (c.q. 0.1). Doet men dit dan heet A' geëquilibreerd. Bij voorkeur neemt men voor de p_i en q_i machten van 2 (c.q. 10), dan treden bij de vermenigvuldigingen geen afrondingsfouten op.

Het is duidelijk dat de in 2.2.3 besproken pivotstrategie niet invariant is tegen schaling. Stel dat A_{11} het grootste element van A is. Neem $p_1 = q_1 = \varepsilon$, $p_2 = \ldots = p_n = q_2 = \ldots = q_n = 1$.
Dan is, als ε voldoende klein is, A'_{11} beslist niet het grootste element van A'.

Een gevolg is dat een erg weinig geëquilibreerde matrix aanleiding kan geven tot een pivot keuze die tot een gereduceerd stelsel voert dat aanzienlijk slechter geconditioneerd is dan het oorspronkelijke stelsel. In de praktijk

komt dit in hoofdzaak voor indien zowel de verschillende onbekenden als de rechterleden fysische grootheden zijn met zeer verschillende grootte-orde (bv. omdat ze verschillende dimensies hebben of in verschillende eenheden gemeten zijn). Als men alle vergelijkingen dimensieloos maakt met behulp van redelijk bij het probleem passende referentie grootheden, dan blijkt als regel de verkregen matrix ook redelijk geëquilibreerd te zijn! Dit is een sterk argument vóór dimensieloos maken.

2.2.5. Foutenanalyse

Het is mogelijk om kwantitatieve uitspraken te doen over de maximale fout in de oplossing als gevolg van afrondingsfouten in de afzonderlijke bewerkingen. Om deze uitspraken te kunnen formuleren gebruiken we als norm voor een vector $\underline{x}$

$$\|\underline{x}\| = \max_{1\le i\le n} |x_i|$$

en als norm voor een matrix A

$$\|A\| = \max_{\|\underline{x}\|=1} \|A\underline{x}\| = \max_{\underline{x}\neq 0} \frac{\|A\underline{x}\|}{\|\underline{x}\|} \qquad (1)$$

Het blijkt dat de aldus gedefinieerde $\|A\|$ vrij eenvoudig uit te drukken is in de matrixelementen A_{ij}:

$$\|A\| = \max_{1\le i\le n} \left(\sum_{j=1}^{n} |A_{ij}|\right) .$$

Uit (1) volgt direct de volgende ongelijkheid

$$\|A\underline{x}\| \le \|A\|\,\|\underline{x}\| . \qquad (2)$$

We zullen nu eerst nagaan hoeveel de oplossing van een stelsel $A\underline{x} = \underline{b}$ kan veranderen, indien de matrix A van het stelsel verandert. Zij E een willekeurige (kleine) matrix en zij $\bar{\underline{x}}$ de oplossing van het (naburige) stelsel

$$(A+E)\bar{\underline{x}} = \underline{b} .$$

Dan geldt

$$\frac{\|\bar{\underline{x}} - \underline{x}\|}{\|\bar{\underline{x}}\|} \le \mathrm{cond}(A)\,\frac{\|E\|}{\|A\|} \qquad (3)$$

waarin

$$\mathrm{cond}(A) = \frac{\max_{\|\underline{x}\|=1} \|A\underline{x}\|}{\min_{\|\underline{x}\|=1} \|A\underline{x}\|} \qquad (4)$$

het zg. conditiegetal van A is. Natuurlijk is cond(A) ≥ 1 en cond(A) zal groot zijn indien de kolommen van A bijna afhankelijk zijn (dan is er een $\underline{x}$ zodanig dat $\|A\underline{x}\|$ veel kleiner is dan $\|A\|\|\underline{x}\|$).
Het conditiegetal is eenvoudig te berekenen indien men de inverse matrix A^{-1} kent: er geldt namelijk

$$\text{cond}(A) = \|A\|\|A^{-1}\| \ .$$

Zij nu $\bar{\underline{x}}$ de "oplossing" van $A\underline{x} = \underline{b}$ die verkregen wordt indien eliminatieproces en terugsubstitutie uitgevoerd worden met een arithmetiek met een relatieve nauwkeurigheid $1 : 2^t$. Dan kan men bewijzen dat er een matrix E is, met

$$\|E\| \leq 3n^2 \cdot 2^{-t} \|U\| \ , \tag{5}$$

zodanig dat $\bar{\underline{x}}$ de exacte oplossing is van het naburige stelsel

$$(A + E)\bar{\underline{x}} = \underline{b} \ . \tag{6}$$

In de praktijk blijkt dat (bij partial pivoting) vrijwel steeds $\|U\| / \|A\|$ niet veel groter is dan 1. Het resultaat (1) is dan zeer bevredigend: het effect van de afrondingsfouten is equivalent met een storing in de matrix waarvan de relatieve grootte $\|E\| / \|A\|$ niet groter is dan enkele malen de relatieve fout in één arithmetische bewerking maal het aantal elementen van de matrix.
Uit (3) en (5) volgt nu

$$\frac{\|\bar{\underline{x}} - \underline{x}\|}{\|\bar{\underline{x}}\|} \leq 3n^2 \cdot 2^{-t} \frac{\|U\|}{\|A\|} \text{cond}(A) \ .$$

Hieruit volgt: De invloed van afrondingsfouten kan alleen aanzienlijk zijn (aannemende dat $3n^2 \cdot 2^{-t}$ klein genoemd mag worden en dat $\|U\| \sim \|A\|$) indien het conditiegetal van A groot is. Maar dit is meer een eigenschap van het sommetje dan van de afrondingsfouten. Bij veranderingen in de matrix tengevolge van meetfouten e.d. treedt volgens (3) ook de factor cond(A) als mogelijke versterkingsfactor op.

Opmerking. De analyse die tot (5) leidt is een "worst case" analyse. In de praktijk kan, in verband met statistische effecten van elkaar compenserende afrondingsfouten, de factor $3n^2$ wel door n vervangen worden.

2.2.6. De algoritme van Crout

Uit de algoritme van Gauss zonder verwisselen, geschreven in de vorm (waarin $A_{ij}^{(0)} = A_{ij}$)

```
for k := 1 step 1 until n do
begin for j = k step 1 until n do U_kj := A_kj^(k-1);
      for i := k + 1 step 1 until n do L_ik := A_ik^(k-1)/U_kk;
      for i := k + 1 step 1 until n do for j := k + 1 step 1 until n do
                                  A_ij^(k) := A_ij^(k-1) - L_ik × U_kj
end
```

volgt door inductie dat

$$A_{ij}^{(k)} = A_{ij} - \sum_{\ell=1}^{k} L_{i\ell} \times U_{\ell j} \ , \quad i > k, \ j > k \ ,$$

en dus ook dat

$$\left.\begin{aligned} U_{kj} &= A_{kj} - \sum_{\ell=1}^{k-1} L_{k\ell} \times U_{\ell j} \ , & j \geq k \\ L_{ik} &= \left(A_{ik} - \sum_{\ell=1}^{k-1} L_{i\ell} \times U_{\ell k}\right)/U_{kk} \ , & i > k \end{aligned}\right\} \ . \qquad (1)$$

We kunnen de bewerkingen die nodig zijn om U_{kj} en L_{ik} te bepalen in een volgorde doen, die meer bij (1) aansluit:

```
for k := 1 step 1 until n do
begin for j := k step 1 until n do
      begin s := A_kj;
            for ℓ := 1 step 1 until k - 1 do s := s - L_kℓ × U_ℓj;
            U_kj := s
      end;
      for i := k + 1 step 1 until n do
      begin s := A_ik;
            for ℓ := 1 step 1 until k - 1 do s := s - L_iℓ × U_ℓk;
            L_ik := s/U_kk
      end
end
```

Dit is de zg. algoritme van Crout. Merk op dat de elementen U en L die bij de k-de slag in rechterleden voorkomen, alle reeds in vorige slagen bepaald zijn. Voeren we de algoritme uit als boven dan gebeurt er arithmetisch precies hetzelfde als bij de Gauss-algoritme (inclusief afrondingen), alleen in een andere volgorde en zonder dat de tussenresultaten $A_{kj}^{(\ell)}$ en $A_{ik}^{(\ell)}$ expliciet in het array A genoteerd worden.

Dit heeft grote voordelen:

a) Bij het werken met tafelmachines kunnen we de partiële sommen s in het optelregister laten staan. Dit scheelt werk (en overschrijffouten). Bovendien hoeven we s niet af te ronden, waardoor de berekening nauwkeuriger wordt.

b) Bij werken met een computer is het iedere keer opzoeken van de arrayplaatsen A_{ij} enigszins tijdrovend. Bovendien is het bij sommige computers mogelijk om de partiële sommen s in zg. dubbele lengte te bewaren (bedenk dat de summanden in s producten zijn die in eerste instantie ook aanleiding geven tot een dubbel lang getal). Pas bij de toekenning van de laatste waarde van s aan U_{kj}, resp. de deling hiervan door U_{kk} wordt weer op de normale (zg. enkele) lengte afgerond.

We halen nog een andere conclusie uit de formules (1). Definieer

$$L_{11} = L_{22} = \ldots = L_{nn} = 1 \ .$$

Dan kunnen we (1) ook schrijven als

$$A_{kj} = \sum_{\ell=1}^{k} L_{k\ell} \times U_{\ell j} \ , \qquad j \geq k$$

$$A_{ik} = \sum_{\ell=1}^{k} L_{i\ell} \times U_{\ell k} \ , \qquad i > k \ .$$

Dus ook

$$A_{ij} = \sum_{\ell=1}^{\min(i,j)} L_{i\ell} \times U_{\ell j} \ , \quad \text{alle } i \text{ en } j. \tag{2}$$

Definieer nu de triangulaire matrices L en U door

$$L = \begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ \vdots & \vdots & \ddots & \\ L_{n1} & L_{n2} & \cdots & L_{nn} \end{pmatrix}, \quad U = \begin{pmatrix} U_{11} & U_{12} & \cdots & U_{1n} \\ & U_{22} & \cdots & U_{2n} \\ & & \ddots & \vdots \\ & & & U_{nn} \end{pmatrix}.$$

Dus

$$L_{i\ell} = 0 \quad \text{voor} \quad \ell > i, \; L_{ii} = 1$$
$$U_{\ell j} = 0 \quad \text{voor} \quad \ell > j \; . \tag{3}$$

Dan is (2) equivalent met

$$L \times U = A \; . \tag{4}$$

Omgekeerd volgt uit (4), met matrices L en U die voldoen aan (3), natuurlijk ook (2) en dus (1).
Derhalve geldt:

Als voor de matrix A de Gauss-algoritme zonder verwisslen werkt (d.w.z., als geen der pivots nul wordt), dan zijn er eenduidig bepaalde matrices L en U die aan (3) en (4) voldoen. Dit is de zg. LU decompositie van A.

Opmerkingen

1) De bewerkingen die bij de Gauss-algoritme op de rechterleden worden uitgevoerd, kunnen we ook in andere volgorde uitvoeren:

```
for k := 1 step 1 until n do
begin s := b_k;
      for ℓ := 1 step 1 until k - 1 do s := s - L_kℓ × c_ℓ;
      c_k := s
end
```

Dit komt neer (ga na) op oplossing van het triangulaire stelsel

$$L \times \underline{c} = \underline{b} \; .$$

Dit klopt met het feit dat we na de Gauss reductie overhouden

$$U \times \underline{x} = \underline{c} \; .$$

2) Ook bij uitvoering van de Crout algoritme kan men rijverwisselingen uitvoeren (en het is gewenst dit te doen, opdat $|L_{ik}| \leq 1$ wordt). Bij het begin van de k-de slag bepaalt men dan eerst de getallen

$$s_i = A_{ik} - \sum_{\ell=1}^{k-1} L_{i\ell} \times U_{\ell k} , \qquad i = k,\ldots,n$$

en het absoluut grootste hiervan kiest men als pivot U_{kk} (na eventueel verwisselen). Werk dit zelf verder uit.
In plaats van (4) krijgen we dan: er is een matrix A', verkregen uit A door rijverwisselingen, zodanig dat A' = L × U.

2.3. Iteratieve methoden

2.3.1. De methoden van Jacobi en van Gauss-Seidel

We beschouwen de volgende iteratiemethode voor de oplossing van $\underline{A}\underline{x} = \underline{b}$. Neem aan dat de diagonaalelementen A_{ii} alle $\neq 0$ zijn. Schrijf dan de vergelijkingen als *)

$$x_i = (b_i - \sum_{j \neq i} A_{ij} x_j)/A_{ii} , \qquad 1 \leq i \leq n .$$

Zij $\underline{x}^{(k)}$ een (k-de) benadering voor de oplossing. Bepaal dan $\underline{x}^{(k+1)}$ uit

$$x_i^{(k+1)} := (b_i - \sum_{j \neq i} A_{ij} x_j^{(k)})/A_{ii}$$

$$= x_i^{(k)} + (b_i - \sum_{j} A_{ij} x_j^{(k)})/A_{ii} , \qquad i = 1,\ldots,n .$$

Dit is het iteratieproces van Jacobi.
Een voor de hand liggende variant is

$$x_i^{(k+1)} = (b_i - \sum_{j<i} A_{ij} x_j^{(k+1)} - \sum_{j>i} A_{ij} x_j^{(k)})/A_{ii}$$

$$= x_i^{(k)} + (b_i - \sum_{j<i} A_{ij} x_j^{(k+1)} - \sum_{j\geq i} A_{ij} x_j^{(k)})/A_{ii} , \qquad i = 1,\ldots,n .$$

*) Merk op dat dit van de vorm $\underline{x} = \underline{f}(\underline{x})$ is. Vergelijk het nu volgende met 1.4.1.

Dit is het proces van Gauss-Seidel. Hier gebruiken we nieuw berekende componenten $x_j^{(k+1)}$ zodra we ze hebben. Bij Jacobi blijven we de oude $x_j^{(k)}$ gebruiken tot we <u>alle</u> $x_j^{(k+1)}$ bepaald hebben. Men spreekt van simultaneous displacement of Gesamtschrittverfahren bij Jacobi en van successive displacement of Einzelschrittverfahren bij Gauss-Seidel.

Wanneer convergeren deze processen? O.a. als de matrix A zg. diagonaal overwegend is, d.w.z., als

$$|A_{ii}| > \sum_{j\neq i} |A_{ij}| \quad \text{voor alle } i.$$

<u>Stelling</u>

Zij voor alle i $A_{ii} \neq 0$ en $\sum_{j\neq i} |A_{ij}| \leq L|A_{ii}|$, met $L < 1$. Dan is de matrix A niet singulier. De processen van Jacobi en Gauss-Seidel convergeren voor iedere beginvector en er geldt (als $\underline{x}$ de oplossing is)

$$\|\underline{x}^{(k)} - \underline{x}\| \leq L^k \|\underline{x}^{(0)} - \underline{x}\| .$$

Er is dus minstens lineaire convergentie met factor L.

<u>Bewijs</u>. Veronderstel dat $A\underline{x} = \underline{o}$. Dan is voor alle i

$$x_i = -\left(\sum_{j\neq i} A_{ij}\, x_j\right)/A_{ii} \quad \text{en dus} \quad |x_i| \leq L \max(|x_j|) = L\|\underline{x}\| .$$

Hieruit volgt dat $\|\underline{x}\| \leq L\|\underline{x}\|$. Daar $L < 1$ volgt hieruit $\|\underline{x}\| = 0$, dus $\underline{x} = \underline{o}$. Dus $A\underline{x} = \underline{o} \Rightarrow \underline{x} = \underline{o}$, dus A niet singulier.

Stel $\underline{x}$ de oplossing van $A\underline{x} = \underline{b}$. Stel $\underline{x}^{(k)} - \underline{x} = \underline{y}^{(k)}$. Dan geldt voor alle i (ga na)

$$y_i^{(k+1)} = -\left(\sum_{j\neq i} A_{ij}\, y_j^{(k)}\right)/A_{ii} , \qquad \text{(Jacobi)}$$

$$y_i^{(k+1)} = -\left(\sum_{j<i} A_{ij}\, y_j^{(k+1)} + \sum_{j>i} A_{ij}\, y_j^{(k)}\right)/A_{ii} , \qquad \text{(Gauss-Seidel).}$$

Voor Jacobi volgt hieruit

$$|y_i^{(k+1)}| \leq L\|\underline{y}^{(k)}\| \quad \text{en dus} \quad \|\underline{y}^{(k+1)}\| \leq L\|\underline{y}^{(k)}\| .$$

Bij Gauss-Seidel geldt

$$|y_i^{(k+1)}| \leq L \max(\|\underline{y}^{(k)}\|, \|\underline{y}^{(k+1)}\|) ,$$

en dus

$$\|\underline{y}^{(k+1)}\| \leq L \max(\|\underline{y}^{(k)}\|, \|\underline{y}^{(k+1)}\|) .$$

De veronderstelling $\|\underline{y}^{(k+1)}\| > \|\underline{y}^{(k)}\|$ voert tot tegenspraak (ga na!) en dus is ook hier

$$\|\underline{y}^{(k+1)}\| \leq L\|\underline{y}^{(k)}\| .$$

Maak het bewijs zelf af.

In de praktijk zijn deze iteratiemethoden langzamer dan directe methoden behalve in de volgende gevallen:

a) L is zo klein en/of de beginschatting $\underline{x}^{(0)}$ is zo goed en/of de te bereiken nauwkeurigheid is zo gering dat niet meer dan $\frac{1}{3}$ n iteratieslagen gedaan hoeven te worden.

b) A heeft zeer weinig elementen $\neq 0$ en deze liggen bovendien onregelmatig verspreid. Bij eliminatie "lopen L en U dan vol", terwijl de sommen $\sum A_{ij}\, x_j^{(k)}$ juist eenvoudig te berekenen zijn.

c) In sommige toepassingen (randwaardeproblemen bij gewone en partiële differentiaalvergelijkingen) treden zeer grote stelsels op (n ~ 1000 à 10000), met matrices met een zeer klein aantal (bv. 5n) elementen $\neq 0$). Hier zijn iteratieve methoden vrijwel onmisbaar, ook al convergeren ze langzaam. Er is een heel arsenaal van iteratiemethoden die voor speciale problemen snellere convergentie geven.

2.3.2. Naïtereren

Stel dat we met een directe methode het stelsel $A\underline{x} = \underline{b}$ behandeld hebben. Het gevonden resultaat $\underline{x}^{(0)}$ zal dan, als gevolg van afrondingsfouten, niet exact gelijk zijn aan $A^{-1}\underline{b}$. Als het stelsel slecht geconditioneerd is kan de afwijking vrij groot zijn.
Bepaal nu zeer nauwkeurig

$$\underline{r}^{(0)} := \underline{b} - A\underline{x}^{(0)}$$

(het zg. residu). Los vervolgens op (met dezelfde LU-decompositie van A) het stelsel $A\underline{y} = \underline{r}^{(0)}$. Zij het resultaat $\underline{y}^{(0)}$ en bepaal

$$\underline{x}^{(1)} := \underline{x}^{(0)} + \underline{y}^{(0)} .$$

Het is te verwachten dat $\underline{x}^{(1)}$ een betere benadering van $A^{-1}\underline{b}$ is dan $\underline{x}^{(0)}$.

Dit proces, dat men kan herhalen, heet naïtereren.

Als de rekennauwkeurigheid en de conditie van A zo zijn dat de directe methode resultaten aflevert waarvan minstens één cijfer goed is, dan convergeert dit proces, mits de residuen nauwkeurig genoeg berekend worden (als regel moet dit in zg. dubbele lengte gebeuren). Men moet stoppen als $\underline{y}^{(k)}$ voldoende klein is (t.o.v. $\underline{x}^{(k)}$) (en niet als $\underline{r}^{(k)}$ klein is, want dat kan zeer misleidend zijn).

Bij niet extreem slechte conditie zijn een à twee slagen voldoende om alle cijfers van $\underline{x}$ goed te krijgen.

3. Numerieke differentiatie en integratie

Zij $f(x)$ gegeven voor $a < x < b$. Onder "gegeven" verstaan we: er is mathematisch een functie $f(x)$ gedefinieerd en er is beschikbaar een procedure die bij een aangeboden getal x een benadering voor het getal $f(x)$ aflevert. Soms bestaat de procedure uit substitutie in een formule, soms is hij veel ingewikkelder (bv. als $f(x)$ de waarde is van de oplossing van een vergelijking waarin x als parameter voorkomt).
Van een dergelijke functie (die voor iedere x uitrekenbaar is, in eindige precisie en eventueel niet zonder moeite) willen we nu door de analyse gedefinieerde getallen bepalen zoals de waarde van een afgeleide, van een integraal, e.d. Deze getallen zijn in de analyse gedefinieerd door limietprocessen die niet in eindig veel stappen uit te voeren zijn. We zullen ze dus moeten vervangen door benaderingen die wel in eindig veel stappen te berekenen zijn en een controleerbare precisie bereiken.

3.1. We lichten de algemene gang van zaken toe aan de hand van een erg eenvoudige methode voor numerieke differentiatie.
Zij $f(x)$ gegeven voor $-a < x < a$ en stel dat $f'(0)$ bestaat. Wiskundig betekent dit dat er bij iedere $\varepsilon > 0$ een $\delta > 0$ is zodat

$$\left|\frac{f(h) - f(0)}{h} - f'(0)\right| < \varepsilon$$

voor $0 < |h| < \delta$. Wat kunnen we hier numeriek mee doen? Als regel kennen we uiteraard de functie $\delta = \delta(\varepsilon)$ niet!
Weten we van $f(x)$ niet meer dan dat $f'(0)$ bestaat, dan kunnen we niet veel meer doen dan de grootheid

$$Df(h) := \frac{f(h) - f(0)}{h} \qquad (1)$$

voor een aantal naar nul gaande waarden van h berekenen en "kijken" of deze rij getallen nadert tot een limietwaarde.

Weten we niet alleen dat $f'(0)$ bestaat, maar ook dat $f''(x)$ bestaat in een omgeving van $x = 0$, dan leert de differentiaalrekening (formule van Taylor) dat er een ξ tussen 0 en h bestaat zodat

$$f(h) = f(0) + hf'(0) + \tfrac{1}{2}h^2 f''(\xi) \ .$$

Hieruit volgt dat we kunnen schrijven

$$f'(0) = Df(h) + R(h) \ , \qquad (2)$$

waarbij

$$R(h) = -\tfrac{1}{2}hf''(\xi) \tag{3}$$

de zg. <u>afbreekfout</u> in de <u>differentiatieformule</u> (2) is.

Uit (3) volgt dat, als $f''(\xi)$ weinig varieert in het interval $(0,h)$, de afbreekfout $R(h)$ ca tweemaal zo groot is als $R(\frac{1}{2}h)$. Daaruit volgt dat dan

$$R(\tfrac{1}{2}h) \sim R(h) - R(\tfrac{1}{2}h) = Df(\tfrac{1}{2}h) - Df(h) \ . \tag{4}$$

Het rechterlid van (4) kunnen we berekenen door niet alleen $Df(h)$ maar ook $Df(\frac{1}{2}h)$ te bepalen. Dat is dubbel werk, maar we krijgen daarmee een schatting voor de afbreekfout $R(\frac{1}{2}h)$ (en ook voor $R(h)$, maar dat is minder interessant).

In zeer veel gevallen weten we dat voor de afbreekfout $R(h)$ in (2) niet alleen zoiets als (3), maar zelfs een formule van de vorm

$$R(h) = c_1 h + c_2 h^2 + \ldots \tag{5}$$

geldt. Voor de differentiatieformule (1) is dit bv. het geval als $f(x)$ een voor $|x| \leq h$ convergente Taylorreeks heeft (ga na). De waarden van c_1, c_2 zijn onbekend (ook als het theoretisch lukt om ze uit te drukken in $f''(0)$, $f'''(0)$, etc.!).
Uit (5) volgt nu

$$Df(\tfrac{1}{2}h) - Df(h) = R(h) - R(\tfrac{1}{2}h) =$$

$$= \tfrac{1}{2}c_1 h + \tfrac{3}{4}c_2 h^2 + \ldots \ .$$

Anderzijds is (6)

$$R(\tfrac{1}{2}h) = \tfrac{1}{2}c_1 h + \tfrac{1}{4}c_2 h^2 + \ldots \ .$$

Hieruit zien we weer dat $Df(\frac{1}{2}h) - Df(h)$ een redelijke indruk geeft van de grootte van $R(\frac{1}{2}h)$, ongeacht of $c_1 = 0$ of niet!
Als echter $|c_1| >> |c_2 h|$, dan geeft $Df(\frac{1}{2}h) - Df(h)$ zelfs een vrij goede benadering voor $R(\frac{1}{2}h)$. Het is dan zinvol te verwachten dat

$$D_1 f(\tfrac{1}{2}h) := Df(\tfrac{1}{2}h) + (Df(\tfrac{1}{2}h) - Df(h)) \tag{7}$$

(altijd op deze manier berekenen!) een betere benadering voor $f'(0)$ levert.

Uit (6) volgt dat we, naar analogie van (2), kunnen schrijven

$$f'(0) = D_1 f(\tfrac{1}{2}h) + R_1(\tfrac{1}{2}h) , \tag{8}$$

met

$$R_1(\tfrac{1}{2}h) = 2R(\tfrac{1}{2}h) - R(h) = -\tfrac{1}{2}c_2 h^2 + \ldots \tag{9}$$

Als $|c_1| >> |c_2\ h|$ dan is $D_1 f(\frac{1}{2}h)$ dus inderdaad een betere benadering dan $D(\frac{1}{2}h)$; als niet $|c_1| >> |c_2\ h|$, dan is $R_1(\frac{1}{2}h)$ van dezelfde orde van grootte als $R(\frac{1}{2}h)$, het gebruik van $D_1 f(\frac{1}{2}h)$ in plaats van $Df(\frac{1}{2}h)$ schaadt dan niet!

Uit (6) volgt ook dat, als we drie waarden van Df hebben, we kunnen kijken naar

$$\frac{Df(\frac{1}{2}h) - Df(h)}{Df(\frac{1}{4}h) - Df(\frac{1}{2}h)} = 2\,\frac{c_1 + \frac{3}{4}c_2\ h + \ldots}{c_1 + \frac{3}{2}c_2\ h + \ldots} . \tag{10}$$

Als $c_1 \neq 0$ dan nadert deze verhouding tot 2 als $h \to 0$. Als de waarde van de breuk dicht bij 2 ligt, dan is dat dus een indicatie dat $|c_1| >> |c_2\ h|$ en dat beschouwing van $D_1 f$ als betere benadering voor $f'(0)$ winst moet leveren.

We kunnen nu het proces herhalen, waarbij (8) en (9) de rol spelen die eerst (2) en (5) speelden. We merken dan eerst op dat

$$D_1 f(\tfrac{1}{4}h) - D_1 f(\tfrac{1}{2}h) = R_1(\tfrac{1}{2}h) - R_1(\tfrac{1}{4}h) = -\frac{3}{8}c_2\ h^2 + \ldots \tag{11}$$

en

$$R_1(\tfrac{1}{4}h) = -\frac{1}{8}c_2\ h^2 + \ldots . \tag{12}$$

Dus $\frac{1}{3}(D_1 f(\frac{1}{4}h) - D_1 f(\frac{1}{2}h))$ geeft weer een schatting voor $R_1(\frac{1}{4}h)$ (ongeacht de waarde van c_2). En als de termen met $c_2\ h^2$ in (11) en (12) overheersend zijn, dan zal

$$D_2 f(\tfrac{1}{4}h) := D_1 f(\tfrac{1}{4}h) + \tfrac{1}{3}\,(D_1 f(\tfrac{1}{4}h) - D_1 f(\tfrac{1}{2}h)) \tag{13}$$

weer een betere benadering voor $f'(0)$ zijn dan $D_1 f(\frac{1}{4}h)$.

Men noemt het idee om, uitgaande van het theoretische gedrag (5), uit twee waarden $Df(h)$ en $Df(\frac{1}{2}h)$ door extrapolatie de betere waarde $D_1 f(\frac{1}{2}h)$ te bepalen, h-extrapolatie volgens Richardson. In de reeksontwikkeling (9) voor $R_1(\frac{1}{2}h)$ komt geen term met h meer voor, wel één met h^2. De formule (13), waarbij uit twee waarden van $D_1 f$ een betere waarde $D_2 f$ bepaald wordt, heet daarom h^2-extrapolatie.

Opmerkingen

1) We kunnen de factor $\frac{1}{3}$ in (13) ook iets rechtstreekser beredeneren. Omdat de formule voor $R_1(\frac{1}{2}h)$ met een term met h^2 begint, is $R_1(\frac{1}{2}h)$ ca vier maal zo groot als $R_1(\frac{1}{4}h)$. De waarde van $R_1(\frac{1}{4}h)$ is dus ca

$$\frac{1}{3}\,(R_1(\tfrac{1}{2}h) - R_1(\tfrac{1}{4}h)) = \frac{1}{3}\,(D_1f(\tfrac{1}{4}h) - D_1f(\tfrac{1}{2}h))$$

en dit bedrag kan dus eventueel als correctie bij $D_1f(\frac{1}{4}h)$ opgeteld worden.

2) Ga na wat de limietwaarde van

$$\frac{D_1f(\frac{1}{4}h) - D_1f(\frac{1}{2}h)}{D_1f(\frac{1}{8}h) - D_1f(\frac{1}{4}h)}$$

is.

Voorbeeld

In onderstaande tabel zijn de bij een gegeven functie horende waarden van Df(h) voor een aantal waarden gegeven. Daarnaast staan de differenties $\nabla Df(h) := Df(h) - Df(2h)$. De laatste waarden in deze kolom suggereren dat de fout in Df(0.0125) van de orde van enkele malen 0.01 is. Verder constateren we dat de waarden in deze kolom inderdaad met ca een factor $\frac{1}{2}$ afnemen. In de volgende kolom staan de volgens (7) berekende waarden van $D_1f(h)$. Deze lijken al veel beter te convergeren. De bijbehorende differentiekolom suggereert dat $D_1f(0.0125)$ een fout van de orde 0.0001 heeft. Aangezien we in de differentiekolom nu - conform de theorie - afname met een factor $\frac{1}{4}$ constateren, doen we ook nog de h^2-extrapolatie, die $D_2f(h)$ levert. Het resultaat is fraai te noemen (de exacte waarde van f'(0) = 0.874326). Dat het laatste cijfer in $D_2f(h)$ onregelmatig verloopt is een gevolg van afrondingsfouten: de functiewaarden f(0) en f(h) zijn alle in 6 cijfers bepaald, maar bij het uitrekenen van Df(h) verliest men nauwkeurigheid en meer naarmate h kleiner is.

h	Df(h)	∇Df(h)	$D_1f(h)$	$\nabla D_1f(h)$	$D_2f(h)$
.2	1.53967				
.1	1.19690	- .34277	0.85413		
.05	1.03308	- .16382	0.86926	0.01515	0.87431
.025	0.95308	- .08000	0.87308	0.00382	0.87435
.0125	0.91352	- .03956	0.87396	0.00088	0.87425

3.2. De in 3.1 geanalyseerde formule voor numerieke differentiatie is erg eenvoudig maar weinig nauwkeurig. Beter is de zg. centrale formule

$$Df(h) := \frac{f(h) - f(-h)}{2h} . \tag{1}$$

Hiervoor geldt (als f een convergente Taylorreeks heeft)

$$Df(h) = f'(0) + c_2 h^2 + c_4 h^4 + \ldots .$$

Bij extrapolatie begint men dus met h^2-extrapolatie:

$$D_1 f(\tfrac{1}{2}h) := Df(\tfrac{1}{2}h) + \frac{1}{3}(Df(\tfrac{1}{2}h) - Df(h)) . \tag{2}$$

Daarna eventueel h^4-extrapolatie, etc.
Voor tweemaal continu differentieerbare functies f geldt weer een uitspraak van de vorm

$$f'(0) = Df(h) + R(h) , \tag{3}$$

met

$$R(h) = -\frac{1}{6} h^2 f'''(\xi) .$$

Een nog nauwkeuriger formule dan (1) is *)

$$Df(h) = \frac{-f(2h) + 8f(h) - 8f(-h) + f(-2h)}{12h} . \tag{4}$$

Hier geldt voor de restterm

$$R(h) = \frac{1}{30} h^4 f^{(5)}(\xi) .$$

Soms wil men f'(0) benaderen met behulp van de functiewaarden f(0), f(-h), f(-2h), ... (bv. bij het oplossen van differentiaalvergelijkingen). Formules van dat type zijn

$$f'(0) = \frac{f(0) - f(-h)}{h} + \tfrac{1}{2}hf''(\xi) ,$$

$$f'(0) = \frac{3f(0) - 4f(-h) + f(-2h)}{2h} + \frac{1}{3} h^2 f'''(\xi) . \tag{5}$$

Ook bestaan er natuurlijk formules voor de tweede en hogere afgeleiden, bv.

---

*) Laat zien dat dit de uitwerking van formule (2) is, met h i.p.v. $\frac{1}{2}h$.

$$f''(0) = \frac{f(h) - 2f(0) + f(-h)}{h^2} + \frac{h^2}{12} f^{(4)}(\xi)$$

$$f''(0) = \frac{-f(2h) + 16f(h) - 30f(0) + 16f(-h) - f(-2h)}{12h^2} + \frac{h^4}{90} f^{(6)}(\xi)$$

$$f''(0) = \frac{2f(0) - 5f(-h) + 4f(-2h) - f(-3h)}{h^2} + \frac{11}{12} h^2 f^{(4)}(\xi) .$$

3.3. We bespreken nu hoe men in het algemeen formules als boven kan afleiden. Als voorbeeld nemen we formule (5) uit 3.2.
We zoeken dus een formule van het type

$$f'(0) = af(0) + bf(-h) + cf(-2h) + R(h) .$$

a) Eerst bepalen we a, b en c zo dat R nul is voor alle polynomen met graad 0,1,2,... (zo hoog mogelijk). Dit levert als vergelijkingen voor a, b, c:

$f(x) = 1$: $a + b + c = 0$

$f(x) = x$: $h(-b - 2c) = 1$

$f(x) = x^2$: $h^2(b + 4c) = 0$.

Door deze drie vergelijkingen zijn a, b en c geheel bepaald:

$$a = \frac{3}{2h} , \quad b = -\frac{4}{2h} , \quad c = \frac{1}{2h} .$$

En met deze waarden blijkt dat voor $f(x) = x^3$ $R(h) \neq 0$, nl.

$$R(h) = -h^3(-b - 8c) = 2h^2 . \qquad (1)$$

b) <u>Veronderstel</u> nu dat R(h) geschreven kan worden als

$$R(h) = Ch^p f^{(q)}(\xi) . \qquad (2)$$

Wat zijn dan de waarden van C, p en q ?
Zeker geldt $q \geq 3$, want $R(h) = 0$ voor <u>alle</u> polynomen met graad ≤ 2 (waarom?). Anderzijds kan niet $q > 3$ zijn, want voor $f(x) = x^3$ is $R(h) \neq 0$. Dus $q = 3$. Nemen we nu weer $f(x) = x^3$ dan volgt door vergelijking van (1) en (2) $p = 2$ en $C = \frac{1}{3}$ (merk op dat we hiervoor de waarde van ξ niet hoeven te kennen).

Met deze methode kan men de meeste formules voor numerieke differentiatie, integratie, etc., afleiden (niet bewijzen, omdat men moet aannemen dat de restterm in de vorm (1) geschreven kan worden - in veel gevallen is dit zo).

Opgave. Leid ook andere formules uit 3.2 op bovenstaande wijze af.

3.4. We kijken nog even naar de invloed van afrondingsfouten in de functiewaarden van f. Neem bv. de laatste formule uit 3.2 en veronderstel dat we in plaats van met $f(0), f(-h), \ldots$, werken met $f(0) + \varepsilon_0$, $f(-h) + \varepsilon_1, \ldots$. Dit geeft in de uitgerekende waarde voor $Df(h)$ een fout

$$\Delta := \frac{2\varepsilon_0 - 5\varepsilon_1 + 4\varepsilon_2 - \varepsilon_3}{h^2} .$$

Veronderstel nu dat we weten dat $|\varepsilon_j| \leq \varepsilon$ $(j = 0,1,2,3)$. Dan geldt

$$|\Delta| \leq \frac{12\varepsilon}{h^2} .$$

We zien dat deze bovengrens voor de fout in $Df(h)$ de factor h^{-2} bevat. Dat betekent dat als we h kleiner nemen, de afbreekfout $R(h)$ kleiner wordt, maar dat de afrondfout Δ als regel groter wordt! Dit stelt een grens aan de nauwkeurigheid waarmee men een afgeleide numeriek kan bepalen! (tenzij men in staat is, bij afnemende h de functiewaarden nauwkeuriger te gaan bepalen).

3.5. Numerieke integratie

Vrijwel alle numerieke integratie methoden benaderen een integraal als volgt

$$\int_a^b f(x)dx = c_0 f(x_0) + \ldots + c_N f(x_N) + R . \qquad (1)$$

Hierin is (a,b) een gegeven integratie interval, $x_0, \ldots, x_N$ en $c_0, \ldots, c_N$ zijn bij het interval en de methode behorende punten en gewichten (onafhankelijk van f(x) !). Vaak kiest men de punten $x_0, \ldots, x_N$ equidistant:

$$x_j = a + jh , \quad \text{met} \quad h = (b - a)/N . \qquad (2)$$

Bij equidistante punten verkrijgt men een integratieformule (1) vaak door samenstelling van een aantal elementaire integratieformules. Bijvoorbeeld

$$\int_0^h f(x)dx = \frac{1}{2} h(f_0 + f_1) - \frac{1}{12} h^3 f''(\xi) \qquad (3)$$

(trapeziumregel)

met

$$f_0 = f(0) \ , \quad f_1 = f(h) \ ;$$

$$\int_a^b f(x)dx = h(\tfrac{1}{2}f_0 + f_1 + \ldots + f_{N-1} + \tfrac{1}{2}f_N)$$

$$- (b - a) \frac{1}{12} h^2 f''(\xi) \qquad (4)$$

(samengestelde trapeziumregel)

met

$$f_j = f(a + jh) \ , \quad h = (b - a)/N \ .$$

Een formule als (3) kan men weer met de methode van 3.3 afleiden. De overgang van (3) naar (4) berust op de splitsing

$$\int_a^b = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} .$$

Voor de restterm in (4) krijgt men dan in eerste instantie met behulp van (3)

$$R(h) = - \frac{1}{12} h^3 \sum_{j=0}^{N-1} f''(\xi_j) \ , \qquad (5)$$

met $x_j < \xi_j < x_{j+1}$; men kan echter bewijzen dat (bij continue f") er een ξ met $a < \xi < b$ is, zodat

$$\sum_{j=0}^{N-1} f''(\xi_j) = Nf''(\xi) \ .$$

Formules analoog aan (3) en (4) zijn

$$\int_0^h f(x)dx = hf_{\frac{1}{2}} + \frac{1}{24} h^3 f''(\xi)$$

(midpointregel);

$$\int_a^b f(x)dx = h(f_{1/2} + f_{3/2} + \ldots + f_{N-\frac{1}{2}}) + (b - a)\,\frac{1}{24}\,h^2\,f''(\xi)$$

(samengestelde midpointregel).

En

$$\int_0^{2h} f(x)dx = \frac{1}{3}\,h(f_0 + 4f_1 + f_2) - \frac{1}{90}\,h^5\,f^{(4)}(\xi)\ ,$$

(regel van Simpson);

$$\int_a^b f(x)dx = \frac{1}{3}\,h(f_0 + 4f_1 + 2f_2 + \ldots + 4f_{N-1} + f_N) - (b - a)\,\frac{1}{180}\,h^4\,f^{(4)}(\xi)$$

(samengestelde regel van Simpson, hierin moet N even zijn).

Uit (5) volgt dat, als $f''(x)$ weinig varieert binnen de intervallen (x_j, x_{j+1}),

$$R(\tfrac{1}{2}h) \sim \tfrac{1}{4}R(h)\ . \qquad (6)$$

Immers, uit (5) volgt dat

$$R(\tfrac{1}{2}h) = -\frac{1}{12}\,(\tfrac{1}{2}h)^3 \sum_{j=0}^{N-1} (f''(\xi_j) + f''(\xi_{j+\frac{1}{2}}))$$

met $x_j < \xi_j < x_{j+\frac{1}{2}} < \xi_{j+\frac{1}{2}} < x_{j+1}$.

Duiden we de trapeziumregel aan met

$$If(h) := h(\tfrac{1}{2}f_0 + \sum_1^{N-1} f_j + \tfrac{1}{2}f_N)\ ,$$

dan is

$$\int_a^b f(x)dx = If(h) + R(h) = If(\tfrac{1}{2}h) + R(\tfrac{1}{2}h)$$

en uit (6) volgt dan dat

$$R(\tfrac{1}{2}h) \sim \frac{1}{3}\,(R(h) - R(\tfrac{1}{2}h)) = \frac{1}{3}\,(If(\tfrac{1}{2}h) - If(h))\ .$$

Berekenen we dus zowel If(h) als If($\frac{1}{2}$h), dan hebben we een schatting voor de fout R($\frac{1}{2}$h) (en ook voor R(h), maar dat is minder interessant) verkregen. Een analoge redenering geldt voor de midpointregel en voor de regel van Simpson (bij de laatste geldt R($\frac{1}{2}$h) ~ $\frac{1}{16}$ R(h)).

Men kan ook bewijzen dat, als f voldoende vaak differentieerbaar is, voor de resttermen R(h) in de samengestelde regels reeksontwikkelingen gelden van de vorm

$$R(h) = c_2h^2 + c_4h^4 + \ldots \qquad (7)$$

(bij de regel van Simpson is $c_2 = 0$). Hiermee vinden we dat

$$\frac{1}{3}\,(If(\tfrac{1}{2}h) - If(h)) = \frac{1}{3}\,(R(h) - R(\tfrac{1}{2}h)) = \frac{1}{4}\,c_2h^2 + \frac{5}{16}\,c_4h^4 + \ldots\ .$$

Hieruit volgt opnieuw dat de grootheid $\frac{1}{3}$ (If($\frac{1}{2}$h) - If(h)) een indruk geeft van de orde van grootte van de fout en dat, als in (7) de term met c_2 overwegend is, deze grootheid een goede benadering is voor R($\frac{1}{2}$h), zodat

$$I_1f(\tfrac{1}{2}h) := If(\tfrac{1}{2}h) + \frac{1}{3}\,(If(\tfrac{1}{2}h) - If(h)) \qquad (8)$$

een betere benadering voor de integraal zal zijn dan If($\frac{1}{2}$h). Dit is h^2-extrapolatie. Overweegt de term met c_2 niet, dan zal $I_1f(\frac{1}{2}h)$ als regel niet essentieel slechter zijn dan If($\frac{1}{2}$h).
Analoog bij de midpointregel en bij de regel van Simpson (bij de laatste moet men h^4-extrapolatie toepassen).

Past men na h^2-extrapolatie ook h^4-, h^6-, etc., extrapolatie toe dan verkrijgt men het zg. <u>schema van Romberg</u>. Men kan bewijzen dat dit schema ook voor minder gladde functies convergeert, zij het niet zo snel als voor gladde functies.

<u>Opmerking</u>. Bij de trapeziumregel geldt:

$$If(\tfrac{1}{2}h) = \tfrac{1}{2}(If(h) + Jf(h))\ ,$$

waarin

$$Jf(h) = h \sum_{j=0}^{N-1} f_{j+\frac{1}{2}}$$

de midpointsom met stap h is. Om na $If(h)$ $If(\frac{1}{2}h)$ te berekenen hebben we dus alleen de functiewaarden in de nieuw toegevoegde punten nodig.
Verder blijkt dat $I_1 f(\frac{1}{2}h)$, zoals gedefinieerd in (8), juist de regel van Simpson met stap $\frac{1}{2}h$ is (ga na).

3.6. Praktische numerieke integratie

De minimale controle die men bij praktische numerieke integratie moet uitvoeren is, de integraal met de gekozen integratieformule voor tenminste 2 waarden van h te berekenen om door vergelijking van de resultaten een indruk van de nauwkeurigheid te krijgen.

Vaak doet zich de omstandigheid voor dat de te integreren functie $f(x)$ in een deel van het interval (a,b) veel minder "glad" is dan in de rest van het interval. Men wenst dan in het minder gladde stuk een kleinere waarde van h te gebruiken (maar ook alleen daar). Voor functies met enigszins gelijkmatig (maar wel aanzienlijk) variërende gladheid kan men dit idee als volgt generaliseren en automatiseren.
Definieer (voor het geval van de trapeziumregel als basisformule)

$$If(x,h) = \frac{1}{2} h(f(x) + f(x+h))$$

$$Rf(x,h) = \frac{1}{3} (If(x,\tfrac{1}{2}h) + If(x+\tfrac{1}{2}h,\tfrac{1}{2}h) - If(x,h))$$

$$I_1 f(x,h) = If(x,\tfrac{1}{2}h) + If(x+\tfrac{1}{2}h,\tfrac{1}{2}h) + Rf(x,h) \ .$$

Volgens de theorie geeft dan $Rf(x,h)$ een indruk van de orde van grootte van de fout in $If(x,\frac{1}{2}h) + If(x+\frac{1}{2}h,\frac{1}{2}h)$ als benadering voor de integraal van x tot x+h en $I_1 f(x,h)$ is een als regel betere en althans niet essentieel slechtere benadering.
Zij nu een positief getal ε gegeven. We wensen dan het interval (a,b) in deelintervallen (x_0,x_1), (x_1,x_2), ..., (x_{N-1},x_N) te splitsen zodat, als $h_j = x_{j+1} - x_j$

$$|Rf(x_j,h_j)| \le tol(h_j) := \frac{h_j}{b-a} \times \varepsilon \ . \qquad (1)$$

Er geldt dan stellig

$$\left| \sum_{j=0}^{N-1} Rf(x_j,h_j) \right| \le \varepsilon \ .$$

Voor "nette" functies f is $Rf(x_j,h_j)$ ongeveer evenredig met h_j^3. Daar het rechterlid van (1) evenredig is met h_j, betekent dit dat we na de berekening van $Rf(x_j,h_j)$ voor zekere waarde van h_j de "ideale" stap h_j^* vanuit x_j kunnen definiëren als

$$h_j^* = \left(\frac{tol(h_j)}{|Rf(x_j,h_j)|}\right)^{\frac{1}{2}} \times h_j \ . \qquad (2)$$

Immers, bij benadering geldt dan

$$|Rf(x_j,h_j^*)| = \left(\frac{h_j^*}{h_j}\right)^3 \times |Rf(x_j,h_j)| = \frac{h_j^*}{h_j} \times tol(h_j) = tol(h_j^*) \ .$$

We handelen daarom, als we gevorderd zijn tot een punt x_j en een suggestie h_j voor de volgende stap hebben, als volgt.

a) Bepaal $Rf(x_j,h_j)$ en $tol(h_j)$.

b) Als $|Rf(x_j,h_j)| > tol(h_j)$, dan bepalen we h_j^* uit (2), stellen $h_j := 0.95 \times h_j^*$ en beginnen opnieuw.

c) Als $|Rf(x_j,h_j)| \leq tol(h_j)$, dan accepteren we de stap h_j en stellen $I := I + I_1 f(x_j,h_j)$, $x_{j+1} := x_j + h_j$.

d) Als $x_{j+1} < b$, dan bepalen we h_j^* uit (2) en nemen $h_{j+1} := \min(0.95 \times h_j^*, b - x_{j+1})$ als suggestie voor de stap vanuit x_{j+1}.

We moeten dit proces starten met $I := 0$, $x_0 := a$. Voor h_0 nemen we hetzij een meegegeven suggestie, hetzij $b - a$ of $(b - a)/5$ of iets dergelijks.

Men noemt een dergelijk proces integratie met zelfzoekende stap. Hoewel het enige administratie vergt, is de winst, doordat we met een min of meer optimale stapgrootte werken, meestal zeer belangrijk, althans bij functies met variërende "gladheid" en waarvan het uitrekenen van een functiewaarde "duur" is.

4. Numerieke integratie van differentiaalvergelijkingen

Beschouw een eerste orde differentiaalvergelijking

$$\frac{dy}{dx} = f(x,y) \ , \tag{1}$$

met beginvoorwaarde

$$y = y_0 \quad \text{voor} \quad x = x_0 \ . \tag{2}$$

Zoals bekend is er (als f een "nette" functie is, bv. als f en $\frac{\partial f}{\partial y}$ continu zijn) bij iedere x_0 en y_0 precies één functie $y = y(x)$ die voldoet aan (1) en (2):

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

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

Om aan te duiden dat de oplossing ook van de beginvoorwaarde afhangt, schrijven we voor de oplossing door een beginpunt (x_0,y_0) ook vaak

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

We vragen nu de functie $\varphi(x,x_0,y_0)$ numeriek te benaderen. Ter onderscheiding van de exacte oplossing zullen we benaderingen steeds met z aanduiden.
Als regel zullen we tevreden zijn met benaderingen $z_1,z_2,\ldots$, voor de waarden $y_1,y_2,\ldots$ van $\varphi(x,x_0,y_0)$ voor een aantal discrete abscissen $x = x_1,x_2,\ldots$.
Veelal zullen deze punten equidistant zijn: $x_n = x_0 + nh$. Natuurlijk nemen we $z_0 = y_0$.

4.1. Enkele eenvoudige methoden

We illustreren een aantal methoden met de bijbehorende nomenclatuur en eigenschappen aan de hand van de volgende eenvoudige methoden.
Uit

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

volgt

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x,y(x))dx$$

en ook

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x,y(x))dx \ .$$

Door toepassing van integratieformules volgt hieruit

$$y(x_{n+1}) = y(x_n) + hf(x_n,y(x_n)) + R_1(h) \tag{1}$$

$$y(x_{n+1}) = y(x_{n-1}) + 2hf(x_n,y(x_n)) + R_2(h) \tag{2}$$

$$y(x_{n+1}) = y(x_n) + \tfrac{1}{2}h(f(x_n,y(x_n)) + f(x_{n+1},y(x_{n+1}))) + R_3(h) \tag{3}$$

met (indien $f(x,y)$ voldoende vaak differentieerbaar is - de oplossing $y(x)$ is het dan ook en de functie $f(x,y(x))$ dus ook)

$$R_1(h) = \tfrac{1}{2}h^2y''(\xi_1) \ , \tag{4}$$

$$R_2(h) = \tfrac{1}{3}\ h^3y'''(\xi_2) \ , \tag{5}$$

$$R_3(h) = -\ \tfrac{1}{12}\ h^3y'''(\xi_3) \ . \tag{6}$$

Stel nu dat we benaderingen $z_1,z_2,\ldots,z_n$ voor $y(x_1),y(x_2),\ldots,y(x_n)$ gevonden hebben. Dan kunnen we op basis van de (exact geldende) formules (1), (2) en (3) de volgende methoden ter bepaling van z_{n+1} als benadering voor $y(x_{n+1})$ opschrijven:

$$z_{n+1} = z_n + hf(x_n,z_n) \quad \text{(Euler)} \tag{7}$$

$$z_{n+1} = z_{n-1} + 2hf(x_n,z_n) \quad \text{(midpoint-regel)} \tag{8}$$

$$z_{n+1} = z_n + \tfrac{1}{2}h(f(x_n,z_n) + f(x_{n+1},z_{n+1})) \quad \text{(trapeziumregel).} \tag{9}$$

We merken enkele overeenkomsten en verschillen op.

a. Euler en trapeziumregel zijn zg. eenstapsmethoden: we hoeven uit het verleden alleen de waarde van z_n te kennen. De midpoint-regel is een tweestapsmethode: zowel z_{n-1} als z_n spelen mee bij de bepaling van z_{n+1}. Dit laatste heeft als consequenties: de programmering is iets ingewikkelder; er is een startmethode nodig: behalve $z_0 = y_0$ is er ook een waarde van z_1 nodig om vervolgens $z_2,z_3,\ldots$ met de midpoint-regel te kunnen bepalen, z_1 kan bv. met de regel van Euler bepaald worden; de theorie van meerstapsmethoden is ingewikkelder dan die van eenstapsmethoden.

b. Euler en midpoint-regel zijn zg. expliciete methoden: de bepaling van z_{n+1} geschiedt door invulling in een formule. De trapeziumregel is een impliciete methode: formule (9) is een vergelijking waaruit z_{n+1} moet worden opgelost. Dit oplossen kan als regel gebeuren met behulp van successieve substitutie: de asymptotische convergentiefactor is daarbij $\frac{1}{2}h \frac{\partial f}{\partial y}(x_{n+1},z_{n+1})$, voor kleine waarden van h is deze dus klein ten opzichte van 1, zodat dan snelle convergentie verzekerd is. Een beginschatting voor de successieve substitutie wordt meestal verkregen door eerst een expliciete formule (met geringere nauwkeurigheid) toe te passen en het resultaat als beginschatting $z_{n+1}^{(0)}$ te gebruiken. Deze expliciete formule (bv. (7) of (8)) noemt men dan de predictorformule, de impliciete formule, waar men mee itereert heet de correctorformule.

Soms substitueert men alleen de uitkomst $z_{n+1}^{(0)}$ van de predictorformule in de correctorformule en beschouwt men het resultaat $z_{n+1}^{(1)}$ als de definitieve waarde z_{n+1} (correcting only once). Schrijven we dit op voor Euler met trapeziumregel, dan kunnen we het proces ook schrijven als

$$\left.\begin{aligned} k_1 &= hf(x_n,z_n) \\ k_2 &= hf(x_n+h,z_n+k_1) \\ z_{n+1} &= z_n + \tfrac{1}{2}(k_1 + k_2) \end{aligned}\right\}. \qquad (10)$$

In deze vorm geschreven hebben we een eenvoudig voorbeeld van een methode van het zg. Runge-Kutta type.

4.1.1. Locale en globale afbreekfout

De afbreekfout van een methode is de fout die gemaakt wordt doordat de differentiaalvergelijking vervangen wordt door een recursieformule.
Eerst definiëren we het begrip locale afbreekfout. Zij

$$z_{n+1} = \Phi(h,x_n,z_{n+1},z_n,\ldots,z_{n-k+1}) \qquad (11)$$

de algemene formule voor een k-staps methode (expliciet als z_{n+1} niet, impliciet als z_{n+1} wel als argument in Φ voorkomt) voor het oplossen van de differentiaalvergelijking (1), pag.66.

Zij

$$y = \varphi(x,x_n,y_n)$$

de oplossing van (1) die door het gegeven punt (x_n,y_n) gaat.
Definieer nu

$$R(h,x_n,y_n) := \frac{1}{h}[y_{n+1} - \Phi(h,x_n,y_{n+1},y_n,\ldots,y_{n-k+1})] \qquad (12)$$

waarin

$$y_j = \varphi(x_j,x_n,y_n), \qquad n-k+1 \le j \le n+1 .$$

Dan heet R de <u>locale afbreekfout</u> van de methode (11) in het punt (x_n,y_n), bij stapgrootte h.
Hoe kunnen we formule (12) begrijpen?
Zij bijvoorbeeld

$$\Phi(h,x_n,z_{n+1},z_n,\ldots,z_{n-k+1}) = z_n + hF(h,x_n,z_{n+1},z_n,\ldots,z_{n-k+1})$$

dan is

$$R(h,x_n,y_n) = \frac{y_{n+1} - y_n}{h} - F(h,x_n,y_{n+1},y_n,\ldots,y_{n-k+1}) .$$

De eerste term is een benadering voor $y'(x_n)$, de tweede term is een benadering voor $f(x_n,y_n)$ en dus is $R(h,x_n,y_n)$ de fout in de benadering van (1) door (11).
Voor de meeste methoden bestaat er (als f(x,y) voldoende vaak differentieerbaar is) een schatting van de vorm

$$|R(h,x_n,y_n)| \le C|h|^m$$

waarin C wel een functie van f is, maar niet van h, x_n en y_n afhangt (althans voor x_n en y_n binnen een zeker gebied van het (x,y)-vlak).
De exponent m heet de <u>orde</u> van de methode.

<u>Voorbeelden</u>. Bij de methode van Euler is

$$\phi(h,x_n,z_n) = z_n + hf(x_n,z_n) .$$

Uit de formules (1) en (4) van pag. 67 volgt dus dat

$$R(h,x_n,y_n) = \frac{1}{h}[y_{n+1}-y_n-hf(x_n,y_n)] = \tfrac{1}{2}hy''(\xi_n) .$$

En dus geldt

$$|R(h,x_n,y_n)| \leq Cf ,$$

als C een bovengrens is voor $|\frac{1}{2}y''(x)|$ in een relevant gebied *) .
Deze methode heeft dus orde 1.
Analoog geldt voor de midpoint-regel

$$R(h,x_n,y_n) = \frac{1}{h}[y_{n+1} - y_{n-1} - 2hy'_n] = O(h^2)$$

en voor de trapeziumregel

$$R(h,x_n,y_n) = \frac{1}{h}[y_{n+1} - y_n - \tfrac{1}{2}h(y'_n + y'_{n+1})] = O(h^2) .$$

Deze methoden hebben dus de orde 2.
Ook de door (10) gegeven methode heeft orde 2.

Zij nu een beginpunt x_0,y_0 gegeven. Zij $z_0 = y_0,z_1,z_2,\ldots$ verkregen met een integratiemethode met stap h en zij

$$y = \varphi(x,x_0,y_0)$$

de oplossing van de differentiaalvergelijking met deze beginvoorwaarden. We willen de globale afbreekfout

$$y_n - z_n = \varphi(x_n,x_0,y_0) - z_n \qquad (13)$$

schatten. En speciaal het gedrag hiervan als h naar 0 gaat maar tegelijk n naar oneindig, zodat het punt x_n op eindige afstand van x_0 blijft.
We voeren de schatting uit voor het geval van de methode van Euler.

---

*) Uit $y'(x) = f(x,y(x))$ volgt

$$y''(x) = f_x(x,y(x)) + f_y(x,y(x))\cdot y'(x) =$$

$$= f_x(x,y(x)) + f_y(x,y(x))\cdot f(x) .$$

Uit bovengrenzen voor f, f_x en f_y volgt dus een waarde voor C.

Hier geldt voor iedere x en y uit een zeker gebied

$$|R(h,x,y)| \leq Ch \ ,$$

waarin

$$R(h,x,y) = \frac{1}{h}[\varphi(x+h,x,y) - y - hf(x,y)] \ .$$

Met name dus

$$y_{n+1} = y_n + hf(x_n,y_n) + hR(h,x_n,y_n) \ . \tag{14}$$

Anderzijds is

$$z_{n+1} = z_n + hf(x_n,z_n) \ . \tag{15}$$

Zij nu

$$|\frac{\partial f}{\partial y}| \leq L$$

in een relevant gebied. Dan volgt uit (14) en (15)

$$|y_{n+1} - z_{n+1}| \leq (1 + L|h|)|y_n - z_n| + Ch^2 \ .$$

Door volledige inductie vinden we hieruit

$$|y_n - z_n| \leq \frac{(1 + L|h|)^n - 1}{L|h|} Ch^2 \ .$$

Of, daar voor alle t $1+t \leq e^t$,

$$|y(x_n) - z_n| \leq \frac{e^{nL|h|} - 1}{L} C|h| = \frac{e^{L|x_n - x_0|} - 1}{L} C|h| \ . \tag{16}$$

Deze formule laat goed de afhankelijkheid van h zien. Als nl. $|h|$ verkleind wordt dan moeten we, om in eenzelfde punt te komen, n in dezelfde verhouding vergroten. De eerste factor in het rechterlid van (16) blijft daarbij gelijk. Hieruit volgt dat de globale afbreekfout voor alle x_n uit een van h onafhankelijk interval kleiner is dan een factor maal $|h|$. De orde van de globale afbreekfout is dus 1, dezelfde als de orde van de methode van Euler.

Voor andere eenstapsmethoden kan analoog bewezen worden dat de orde van de globale afbreekfout gelijk is aan de orde van de methode. Voor meerstapsmethoden is de theorie ingewikkelder, o.a. omdat de startmethode ook een rol speelt. Voor zg. stabiele meerstapsmethoden geldt weer een uitspraak als boven, mits de orde van de startmethode niet meer dan 1 lager is dan die van de meerstapsmethode.

Schattingen van het type (16) zijn de eenvoudigste in hun soort.
Als $f(x,y)$ voldoende vaak differentieerbaar is, dan geldt ook:
er zijn functies $c_1(x), c_2(x), \ldots$ zodanig dat

$$z_n = y(x_n) + hc_1(x_n) + h^2c_2(x_n) + \ldots$$

(bij een methode van de m-de orde is natuurlijk $c_1 = \ldots = c_{m-1} = 0$).
Op basis van deze formule kan men weer fouten schatten (stap h en stap $\frac{1}{2}h$ vergelijken) en extrapolaties uitvoeren (tussenresultaten verkregen met stap $h, \frac{1}{2}h, \frac{1}{4}h, \ldots$).

4.2. Methoden van hogere orde

4.2.1. Expliciete k-staps methoden van de vorm

$$z_{n+1} = z_n + h \sum_{j=0}^{k-1} c_j \, f(x_{n-j}, z_{n-j}) \ .$$

Als we $c_0, \ldots, c_{k-1}$ zo kiezen dat de orde maximaal is, dan blijkt deze k te zijn.

Voorbeelden (we schrijven f_{n-j} in plaats van $f(x_{n-j}, z_{n-j})$):

$$z_{n+1} = z_n + \frac{1}{2} h(3f_n - f_{n-1}) \qquad \text{orde 2,}$$

$$z_{n+1} = z_n + \frac{1}{12} h(23f_n - 16f_{n-1} + 5f_{n-2}) \qquad \text{orde 3,}$$

$$z_{n+1} = z_n + \frac{1}{24} h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \qquad \text{orde 4.}$$

Dit zijn de zg. Adams-Bashforth formules.

4.2.2. Impliciete k-staps methoden van de vorm

$$z_{n+1} = z_n + h \sum_{j=-1}^{k-1} c_j \, f(x_{n-j}, z_{n-j}) \ .$$

Kiezen we $c_{-1}, \ldots, c_{k-1}$ zo dat de orde maximaal is, dan is deze k+1.

Voorbeelden:

$$z_{n+1} = z_n + \frac{1}{12} h(5f_{n+1} + 8f_n - f_{n-1}) \qquad \text{orde 3,}$$

$$z_{n+1} = z_n + \frac{1}{24} h(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \qquad \text{orde 4.}$$

Dit zijn de zg. Adams-Moulton formules. Bij hetzelfde aantal oude punten is de orde één hoger dan bij de Adams-Bashforth formules. Bij dezelfde orde is de constante in de afbreekfout kleiner.

Het ligt voor de hand om een Adams-Moulton formule als corrector formule te gebruiken met daarbij een Adams-Bashforth formule als predictor.

De afbreekfout bij de Adams methoden is wat groter dan bij sommige andere methoden met dezelfde orde en hetzelfde aantal te berekenen functiewaarden per stap. Ze hebben echter een goede stabiliteit (geringe doorwerking van eens gemaakte fouten). Daarom worden ze veel gebruikt voor nauwkeurige integratie over grote trajecten. Wel is steeds een startmethode nodig.

4.2.3. Expliciete k-staps methoden van de vorm

$$z_{n+1} = z_{n-k+1} + h \sum_{j=0}^{k-1} c_j \, f(x_{n-j}, z_{n-j}) \ .$$

De maximaal haalbare orde is weer k.

Voorbeelden:

$$z_{n+1} = z_{n-1} + 2hf_n \qquad \text{orde 2,}$$

$$z_{n+1} = z_{n-2} + \frac{1}{4} h(9f_n + 3f_{n-2}) \qquad \text{orde 3,}$$

$$(*) \qquad z_{n+1} = z_{n-3} + \frac{1}{3} h(8f_n - 4f_{n-1} + 8f_{n-2}) \qquad \text{orde 4.}$$

4.2.4. Impliciete k-staps methoden van de vorm

$$z_{n+1} = z_{n-k+1} + \sum_{j=-1}^{k-1} c_j \, f(x_{n-j}, z_{n-j}) \ .$$

De maximaal haalbare orde is k+1 als k oneven is en k+2 als k even is.

Voorbeelden:

$$z_{n+1} = z_n + \frac{1}{2} h(f_{n+1} + f_n) \qquad \text{orde 2,}$$

$$(**) \qquad z_{n+1} = z_{n-1} + \frac{1}{3} h(f_{n+1} + 4f_n + f_{n-1}) \qquad \text{orde 4,}$$

$$z_{n+1} = z_{n-2} + \frac{3}{8} h(f_{n+1} + 3f_n + 3f_{n-1} + f_{n-2}) \qquad \text{orde 4.}$$

De methoden uit 4.2.3 en 4.2.4 hebben bij gelijke orde een geringere afbreekfout dan de Adams methoden. De stabiliteit is echter veel slechter, hetgeen zich bij integratie over grote trajecten doet gevoelen.
De predictor-corrector methode gebaseerd op (*) als predictor en (**) als corrector heet methode van Milne, hij was zeer populair als handrekenmethode.

4.2.5. Algemene k-staps methoden van de vorm

$$z_{n+1} = \sum_{j=0}^{k-1} b_j \, z_{n-j} + \sum_{j=-1}^{k-1} c_j \, f(x_{n-j}, z_{n-j}) \ . \qquad (1)$$

De haalbare orde is 2k bij impliciete en 2k-1 bij expliciete (we eisen dan $c_{-1} = 0$) methoden van dit type. De methoden zijn echter in het algemeen instabiel. Volgens een beroemde stelling van Dahlquist zijn er geen stabiele

methoden met een orde groter dan k+2 als k even is, resp. k+1 als k oneven is.

Hierbij heet een methode instabiel als de invloed van een verandering in de beginwaarden (of van een tussentijdse fout) op de gevonden waarde in een vast gekozen eindpunt groter wordt naarmate h kleiner wordt. Een nodige en voldoende voorwaarde voor dit soort stabiliteit blijkt te zijn:

Het bij (1) behorende k-de graads polynoom

$$\lambda^k - \sum_{j=0}^{k-1} b_j \lambda^{k-j-1} \qquad (2)$$

heeft geen nulpunten buiten en geen meervoudige nulpunten op de eenheidscirkel van het complexe λ-vlak. Dat nulpunten buiten de eenheidscirkel moeilijkheden geven wordt duidelijk als we opmerken dat

$$z_n = c\lambda^n$$

(met λ een nulpunt van (2)) voldoet aan de bij (1) behorende homogene differentievergelijking

$$z_{n+1} = \sum_{j=0}^{k-1} b_j z_{n-j} .$$

Voorbeelden:

$$z_{n+1} = -4z_n + 5z_{n-1} + h(4f_n + 2f_{n-1})$$

2 staps expliciet, orde 3, instabiel,

$$z_{n+1} = \frac{9}{8} z_n - \frac{1}{8} z_{n-2} + \frac{3}{8} h(f_{n+1} + 2f_n - f_{n-1})$$

3 staps impliciet, orde 4, stabiel.

4.2.6. Runge Kutta methoden.

De zg. Runge Kutta methoden vormen een belangrijke klasse van eenstapsmethoden. Ze kunnen gemotiveerd worden door het feit dat de oplossing y(x) voldoet aan

$$y_{n+1} = y_n + hf(\xi_n, y(\xi_n))$$

met een ξ_n uit (x_n, x_{n+1}). Helaas zijn echter zowel ξ_n als $y(\xi_n)$ onbekend. Daarom doen we een aantal "metingen" van de van de incrementfunctie $k = hf(x,y)$ in listig gekozen punten in de buurt van de oplossingskromme en vervolgens nemen we voor $z_{n+1} - z_n$ een gewogen gemiddelde van de meetwaarden. De eenvoudigste Runge Kutta methode is al in 4.1.b ter sprake gekomen. Deze methode heeft de orde 2.

In het algemeen zien de Runge Kutta methoden er als volgt uit.

Is men gevorderd tot (x_n, z_n) dan wordt z_{n+1} berekend uit

$$k_1 = hf(x_n, z_n)$$

$$k_2 = hf(x_n + \alpha_1 h,\ z_n + \beta_{11} k_1)$$

$$k_3 = hf(x_n + \alpha_2 h,\ z_n + \beta_{21} k_1 + \beta_{22} k_2)$$

$$k_m = hf(x_n + \alpha_m h,\ z_n + \sum_{j=1}^{m-1} \beta_{mj} k_j)$$

$$z_{n+1} = z_n + \sum_{i=1}^{m} \gamma_i k_i$$

De constanten α_i, β_{ij}, γ_i behoren bij de methode (en zijn onafhankelijk van de differentiaalvergelijking). Ze zijn zo gekozen dat de orde van de methode zo hoog mogelijk is (deze eis bepaalt de constanten niet geheel, er zijn meerdere methoden met dezelfde orde).

Voorbeelden:

$$k_1 = hf(x_n, z_n)$$

$$k_2 = hf(x_n + \tfrac{1}{2}h,\ z_n + \tfrac{1}{2}k_1)$$

$$k_3 = hf(x_n + \tfrac{3}{4}h,\ z_n + \tfrac{3}{4}k_2)$$

$$z_{n+1} = z_n + \frac{1}{9}(2k_1 + 3k_2 + 4k_3)$$

orde 3

$$k_1 = hf(x_n, z_n)$$

$$k_2 = hf(x_n + \tfrac{1}{2}h,\ z_n + \tfrac{1}{2}k_1)$$

$$k_3 = hf(x_n + h,\ z_n - k_1 + 2k_2)$$

$$z_{n+1} = z_n + \frac{1}{6}(k_1 + 4k_2 + k_3)$$

orde 3

$$k_1 = hf(x_n, z_n)$$

$$k_2 = hf(x_n + \tfrac{1}{2}h,\ z_n + \tfrac{1}{2}k_1)$$

$$k_3 = hf(x_n + \tfrac{1}{2}h,\ z_n + \tfrac{1}{2}k_2)$$

$$k_4 = hf(x_n + h,\ z_n + k_3)$$

$$z_{n+1} = z_n = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

orde 4.

De Runge Kutta methoden "verbruiken" meer functiewaarden dan een goede meerstaps methode met dezelfde nauwkeurigheid. Maar ze hebben een redelijke stabiliteit en verder alle voordelen van eenstaps methoden (bv. is gemakkelijk de stapgrootte te wijzigen). Daarom worden ze als general purpose methode zeer veel gebruikt. En ook als startmethode bij meerstaps methoden. Er zijn ook varianten waarbij, eventueel ten koste van een nog extra te berekenen functiewaarde, een hulpgrootheid bepaald kan worden die een kwantitatieve indruk van de locale afbreekfout geeft. Met behulp hiervan kan de integratie met zelf-zoekende stap uitgevoerd worden.

4.2.7. Taylor methode.

Tot slot noemen we een methode die slechts in speciale gevallen bruikbaar is. Voor de oplossing $y = y(x)$ van de differentiaalvergelijking geldt

$$y(x + h) = y(x) + hy'(x) + \frac{1}{2}h^2y''(x) + \frac{1}{6}h^3y'''(x) + \ldots . \qquad (1)$$

Uit

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

volgt

$$y''(x) = f_x(x, y(x)) + f_y(x, y(x)) \cdot y'(x)$$

$$y'''(x) = f_{xx}(x, y(x)) + 2f_{xy}(x, y(x)) \cdot y'(x)$$

$$+ f_{yy}(x, y(x))(y'(x)) + f_y(x, y(x)) \cdot y''(x) ,$$

etc.

Definieer de functies

$$c_1(x,y) = f(x,y)$$

$$c_2(x,y) = f_x(x,y) + f_y(x,y)c_1(x,y) \qquad (2)$$

$$c_3(x,y) = f_{xx}(x,y) + 2f_{xy}(x,y)c_1(x,y)$$
$$+ f_{yy}(x,y)c_1^2(x,y) + f_y(x,y)c_2(x,y) \ ,$$

etc.

Dan kunnen we bv. als derde orde Taylor methode definiëren

$$z_{n+1} = z_n + hc_1(x_n,z_n) + \frac{1}{2}h^2c_2(x_n,z_n) + \frac{1}{6}h^3c_3(x_n,z_n) \ .$$

Door vergelijking met (1) volgt dat deze methode de orde 3 heeft.

De Taylor methoden zijn eenstaps methoden met de voordelen van dien. Ze zijn alleen bruikbaar als waarden van de functies $c_2(x,y), c_3(x,y), \ldots$ zonder excessieve moeite uitgerekend kunnen worden.

Als $f(x,y)$ als min of meer eenvoudige formule in x en y gegeven is, dan kunnen formules voor $c_2(x,y), c_3(x,y), \ldots$ met behulp van computer programma's voor zg. "formula manipulation" bepaald worden.

Een soms wat eenvoudiger variant op het werken met de formules (2) is:

Neem voor de door (ξ,η) gaande oplossing $y = \varphi(x,\xi,\eta)$ een reeksontwikkeling

$$\varphi(x,\xi,\eta) = \eta + \sum_{j=1}^{\infty} c_j(\xi,\eta)(x - \xi)^j/j! \qquad (2)$$

aan; substitueer deze in de differentiaalvergelijking, ontwikkel het rechterlid naar machten van $x - \xi$ en stel overeekomstige coëfficiënten links en rechts gelijk.

Voorbeeld:

$$\frac{dy}{dx} = x^2 + y^2 \ . \qquad (3)$$

Uit

$$y = \eta + \sum_{j=1}^{\infty} c_j(x - \xi)^j/j!$$

volgt (ga na), als we ter afkorting $x - \xi = t$ stellen,

$$\frac{dy}{dx} = \sum_{j=0}^{\infty} c_{j+1} t^j/j! \ ,$$

$$y^2 = \eta^2 + 2\eta \sum_{j=1}^{\infty} c_j t^j/j! + \sum_{j=2}^{\infty} t^j/j! \sum_{i=1}^{j-1} \binom{j}{i} c_i c_{j-i} \ .$$

Substitutie in (3) levert nu, door naar de coëfficiënten van $t^0, t^1, t^2, \ldots$ te kijken, de relaties

$$c_1 = \xi^2 + \eta^2 \ ,$$

$$c_2 = 2\xi + 2\eta c_1 \ ,$$

$$c_3 = 2 + 2\eta c_2 + 2c_1^2 \ ,$$

$$c_{j+1} = 2\eta c_j + \sum_{i=1}^{j-1} \binom{j}{i} c_i c_{j-i} \qquad (j \geq 3) \ .$$

Bij gegeven ξ en η kunnen $c_1, c_2, \ldots$ hieruit successief bepaald worden.

4.3. Stelsels differentiaalvergelijkingen
Vergelijkingen van hogere orde

Alle behandelde methoden voor de scalaire differentiaalvergelijking $\frac{dy}{dx} = f(x,y)$ zijn ook bruikbaar voor de vector-differentiaalvergelijking

$$\frac{d\underline{y}}{dx} = \underline{f}(x,\underline{y}) \ , \tag{1}$$

d.w.z., voor stelsels

$$\frac{dy^{(1)}}{dx} = f^{(1)}(x,y^{(1)},\ldots,y^{(n)})$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\frac{dy^{(n)}}{dx} = f^{(n)}(x,y^{(1)},\ldots,y^{(n)}) \ .$$

En hiermee kunnen we ook differentiaalvergelijkingen van hogere orde behandelen, bv.

$$\frac{d^n y}{dx^n} = f\left(x,y,\frac{dy}{dx},\ldots,\frac{d^{n-1}y}{dx^{n-1}}\right) \ . \tag{2}$$

Stel nl.

$$w^{(1)} = y, \ w^{(2)} = \frac{dy}{dx} \ , \ldots, \ w^{(n)} = \frac{d^{n-1}y}{dx^{n-1}} \ .$$

Dan krijgen we het stelsel

$$\frac{dw^{(1)}}{dx} = w^{(2)}$$
$$\ldots\ldots\ldots$$
$$\frac{dw^{(n-1)}}{dx} = w^{(n)}$$
$$\frac{dw^{(n)}}{dx} = f(x,w^{(1)},\ldots,w^{(n)}),$$

of

$$\frac{d\underline{w}}{dx} = \underline{f}(x,\underline{w}) \ , \tag{3}$$

met

$$\underline{f}(x,\underline{w}) = \begin{pmatrix} w^{(2)} \\ \cdot\ \cdot\ \cdot \\ w^{(n)} \\ f(x,w^{(1)},\ldots,w^{(n)}) \end{pmatrix} .$$

Ook met de beginvoorwaarde past het mooi. Bij de vectorvergelijking (1) hoort als passende beginvoorwaarde $\underline{y}(x_0) = \underline{y}_0$. En bij de n-de orde vergelijking (2) horen als passende beginvoorwaarden $y(x_0) = y_0, y'(x_0) = y_0', \ldots, y^{(n-1)}(x_0) = y_0^{(n-1)}$. Stellen we $w_0^{(1)} = y_0, \ldots, w_0^{(n)} = y_0^{(n-1)}$, dan behoort bij (3) dus de beginvoorwaarde $\underline{w}(x_0) = \underline{w}_0$.

Voorbeeld. De tweede orde Runge Kutta methode (zie § 4.1.b) ziet er voor het stelsel differentiaalvergelijkingen (1) als volgt uit.

$$\underline{k}_1 = h\underline{f}(x_n,\underline{z}_n) ,$$

$$\underline{k}_2 = h\underline{f}(x_n+h,\underline{z}_n+\underline{k}_1) ,$$

$$\underline{z}_{n+1} = \underline{z}_n + \tfrac{1}{2}(\underline{k}_1+\underline{k}_2) .$$

In coördinaten uitgeschreven voor het 2×2-stelsel

$$\frac{dy_1}{dx} = f_1(x,y_1,y_2) ,$$

$$\frac{dy_2}{dx} = f_2(x,y_1,y_2)$$

krijgen we de volgende formules:

$$k_{11} = hf_1(x_n,z_{1n},z_{2n}) ,$$

$$k_{21} = hf_2(x_n,z_{1n},z_{2n}) ,$$

$$k_{12} = hf_1(x_n+h,z_{1n}+k_{11},z_{2n}+k_{21}) ,$$

$$k_{22} = hf_2(x_n+h,z_{1n}+k_{11},z_{2n}+k_{21}) ,$$

$$z_{1,n+1} = z_{1n} + \tfrac{1}{2}(k_{11}+k_{12}) ,$$

$$z_{2,n+1} = z_{2n} + \tfrac{1}{2}(k_{21}+k_{22}) .$$

4.3.1. Speciale methoden voor tweede orde vergelijkingen

Er bestaan speciale methoden voor tweede orde vergelijkingen

$$\frac{d^2y}{dx^2} = f(x,y,\frac{dy}{dx}) \ .$$

Met name als $\frac{dy}{dx}$ niet voorkomt, dus bij vergelijkingen van de vorm

$$\frac{d^2y}{dx^2} = f(x,y) \ , \tag{1}$$

zijn deze wat eenvoudiger dan toepassing van de algemene methode voor 2 × 2 stelsels op het stelsel

$$\frac{dy}{dx} = u \ , \quad \frac{du}{dx} = f(x,y) \ . \tag{2}$$

Een voor de hand liggende methode voor vergelijkingen van het type (1) is gebaseerd op de volgende discretisatie van de tweede afgeleide:

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + O(h^2) \ . \tag{3}$$

Deze formule in (1) ingevuld geeft

$$y_{n+1} - 2y_n + y_{n-1} = h^2 f(x_n,y_n) + O(h^4)$$

en dus de methode

$$z_{n+1} = 2z_n - z_{n-1} + h^2 f(x_n,z_n) \ . \tag{4}$$

Deze methode moeten we starten met beginwaarden z_0 en z_1, bijv. te verkrijgen uit

$$z_0 = y_0, \ z_1 = y_0 + hy_0' + \tfrac{1}{2}h^2 f(x_0,y_0) \ .$$

Wensen we ook een benadering voor $y'(x_n)$ te kennen, dan kunnen we hiervoor nemen

$$v_n := \frac{z_{n+1} - z_{n-1}}{2h} \ . \tag{5}$$

De locale afbreekfout is in dit geval gedefinieerd door

$$R(h,x_n,y_n) := \frac{1}{h^2}(y_{n+1} - 2y_n + y_{n-1} - h^2 f(x_n,y_n)) ,$$

waaruit volgt dat de orde van deze methode 2 is. Ook de globale afbreekfout, zowel voor z_n als benadering voor y_n als voor v_n als benadering voor y'_n, heeft orde 2.

We kunnen de invloed van afrondingsfouten nog verminderen door te stellen $z_{n+1} - z_n = hv_{n+\frac{1}{2}}$ en in plaats van (4) te schrijven

$$\left.\begin{aligned} v_{n+\frac{1}{2}} &= v_{n-\frac{1}{2}} + hf(x_n,z_n) \\ z_{n+1} &= z_n + hv_{n+\frac{1}{2}} \end{aligned}\right\} \qquad (6)$$

met als start

$$z_0 = y_0,\ v_{\frac{1}{2}} = y'_0 + \tfrac{1}{2}hf(x_0,y_0) .$$

We hebben hier in feite een discretisatie van (2), waarbij we $y(x)$ in de punten x_n en $u(x)$ in de punten $x_{n+\frac{1}{2}}$ discretiseren. Dit kan omdat in (2) f de variabele u niet bevat.

Uit (5) en (6) volgt nu

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

Berekening van v_n op deze wijze leidt tot een veel kleinere afrondingsfout dan berekening via (5).

Opmerking. Daar

$$y_{n+1} - 2y_n + y_{n-1} = h^2y''_n + \frac{1}{12}h^4y^{(4)}_n + O(h^6) =$$

$$= h^2y''_n + \frac{1}{12}h^2[y''_{n+1} - 2y''_n + y''_{n-1}] + O(h^6) ,$$

kunnen we (1) ook discretiseren door

$$z_{n+1} - 2z_n + z_{n-1} = \frac{1}{12}h^2[f(x_{n+1},z_{n+1}) + 10f(x_n,z_n) + f(x_{n-1},z_{n-1})] . \qquad (7)$$

Dit is een impliciete formule waar z_{n+1} uit opgelost moet worden, hetgeen eenvoudig is als $f(x,y)$ lineair is in y (dus $f(x,y) = g(x)y + h(x)$). Om de startwaarde z_1 bijpassend nauwkeurig te berekenen merken we op dat

$$y_1 - y_{-1} = 2hy'_0 + \frac{1}{3}h^3y'''_0 + O(h^5) =$$

$$= 2hy'_0 + \frac{1}{6}h^2[y''_1 - y''_{-1}] + O(h^5) ,$$

hetgeen aanleiding geeft tot de formule

$$z_1 - z_{-1} = 2hy'_0 + \frac{1}{6} h^2[f(x_1,z_1) - f(x_{-1},z_{-1})] . \qquad (8)$$

Uit (7) (genomen voor n = 0) en (8) kunnen nu z_1 en z_{-1} berekend worden. Deze zg. methode van Numerov heeft orde 4 en is vooral populair in de quantummechanica (Schrödinger vergelijking, deze is lineair).

4.4. Randwaardeproblemen

Behalve beginwaardeproblemen (bij een n-de orde vergelijking n beginwaarden voor $y,y',\ldots,y^{(n-1)}$ in één punt x_0) komen in de practijk ook randwaardeproblemen voor. Hier worden bij een n-de orde vergelijking die in een interval (a,b) moet gelden, n randvoorwaarden gegeven, d.w.z. n relaties tussen de waarden $y(a),y'(a),\ldots,y(b),y'(b),\ldots$.
Bij een tweede orde differentiaalvergelijking

$$\frac{d^2y}{dx^2} = f(x,y,\frac{dy}{dx}), \qquad a < x < b \qquad (1)$$

kunnen we als lineaire randvoorwaarden bv. hebben

$$\left.\begin{aligned} \alpha_1 y(a) + \alpha_2 y'(a) &= \gamma \\ \beta_1 y(b) + \beta_2 y'(b) &= \delta. \end{aligned}\right\} \qquad (2)$$

Een ander voorkomend type randvoorwaarden bij (1) is

$$y(a) = y(b),\ y'(a) = y'(b). \qquad (3)$$

Dit zijn zg. periodiciteitsvoorwaarden.
Als $f(x,y,\frac{dy}{dx})$ als functie van x periodiek is met periode b-a dan impliceren de randvoorwaarden (3) dat we de oplossing y(x) van (1) zoeken die voortgezet kan worden tot een periodieke oplossing met periode b-a.

Men kan een randwaardeprobleem op twee manieren numeriek aanpakken, nl.

a) herleiden tot beginwaardeprobleem
b) rechtstreekse discretisatie, hetgeen tot een groot stelsel, in het algemeen niet lineaire, vergelijkingen leidt.

In veel gevallen verdient de tweede methode de voorkeur.

4.4.1. Om het randwaardeprobleem (1) met de randvoorwaarden (2) te herleiden tot een beginwaardeprobleem kiezen we waarden y(a) en y'(a) die voldoen aan de eerste randvoorwaarde (2) en van een nog vrije parameter s afhangen. Bijv. stellen we

$$-\alpha_2 y(a) + \alpha_1 y'(a) = s$$

waar met (2) uit volgt

$$y(a,s) = \frac{\alpha_1\gamma - \alpha_2 s}{\alpha_1^2 + \alpha_2^2}, \quad y'(a,s) = \frac{\alpha_2\gamma + \alpha_1 s}{\alpha_1^2 + \alpha_2^2}. \tag{4}$$

Bij deze beginwaarden kunnen we nu (voor iedere gekozen getalwaarde van s) de oplossing

$$y = y(x,s)$$

van (1) numeriek bepalen. Hierbij verkrijgen we eindwaarden y(b,s) en y'(b,s) die moeten voldoen aan de tweede randvoorwaarde

$$\beta_1 y(b,s) + \beta_2 y'(b,s) = \delta. \tag{5}$$

Uit deze zg. sluitvergelijking moet de nog onbekende parameter s opgelost worden.

Als de differentiaalvergelijking (1) niet lineair is, dan hangt het linker lid van (5) niet lineair van s af; (5) moet dan met één of ander iteratieve methode, bv. regula falsi, opgelost worden.

Als de differentiaalvergelijking (1) lineair is, d.w.z. van de vorm

$$\frac{d^2y}{dx^2} + a(x)\frac{dy}{dx} + b(x)y = c(x) \tag{6}$$

dan zal y(x,s) lineair van s afhangen. Immers, de functie y(x,s) - y(x,0) voldoet aan de bij (6) behorende <u>homogene</u> differentiaalvergelijking

$$\frac{d^2y}{dx^2} + a(x)\frac{dy}{dx} + b(x)y = 0 \tag{7}$$

en aan de randvoorwaarden

$$y(a) = -\alpha_2 s/(\alpha_1^2 + \alpha_2^2), \quad y'(a) = \alpha_1 s/(\alpha_1^2 + \alpha_2^2). \tag{8}$$

Hieruit volgt dat

$$y(x,s) - y(x,0) = s[y(x,1) - y(x,0)]. \tag{9}$$

We kunnen nu dus volstaan met de bepaling van $y_0(x) := y(x,0)$ als oplossing van (6) met de beginvoorwaarden (4) met s = 0 en die van

$$y_1(x) := y(x,1) - y(x,0)$$

als oplossing van (7) met de beginvoorwaarden (8) met s = 1.
Substitutie van

$$y(x,s) = y_0(x) + s\, y_1(x)$$

in de sluitvergelijking (5) levert nu

$$(\beta_1 y_1(b) + \beta_2 y_1'(b))s = \delta - \beta_1 y_0(b) - \beta_2 y_0'(b). \qquad (10)$$

Hierdoor is s bepaald.

Natuurlijk loopt de bepaling van s uit (10) spaak indien

$$\beta_1 y_1(b) + \beta_2 y_1'(b) = 0. \qquad (11a)$$

Is dit het geval dan is, daar uit (8) volgt dat

$$\alpha_1 y_1(a) + \alpha_2 y_1'(a) = 0, \qquad (11b)$$

de functie $y_1(x)$ een niet triviale oplossing van de homogene differentiaalvergelijking (7) met de bij (2) behorende homogene randvoorwaarden (11). We hebben hier weer de situatie die als regel optreedt bij lineaire problemen: een inhomogeen probleem heeft dan en slechts dan voor ieder rechterlid een eenduidige oplossing indien het bijbehorende homogene probleem uitsluitend de nuloplossing heeft.

Ook bij niet-lineaire randwaardeproblemen kan het voorkomen dat er geen oplossing bestaat.

4.4.2. Als voorbeeld van discretisatie van een randwaardeprobleem bekijken we weer het probleem (1) en (2) uit 4.4.
We verdelen het interval (a,b) in N gelijke delen met lengte h = (b-a)/N en deelpunten $x_j = a + jh$, $j = 0,1,\ldots,N$.
Voor de differentiaalvergelijking (1) ligt de discretisatie

$$z_{j-1} - 2z_j + z_{j+1} - h^2 f(x_j, z_j, \frac{z_{j+1} - z_{j-1}}{2h}) = 0 \qquad (12)$$

voor de hand.

Voor de randvoorwaarden moeten we enkele gevallen onderscheiden:

a) $\alpha_2 = \beta_2 = 0$, $\alpha_1 = \beta_1 = 1$. Dit betekent dat y(a) en y(b) gegeven zijn. We kunnen nu de differentievergelijking (12) laten gelden voor $j = 1,2,\ldots,N-1$, dan komen voor: de N-1 onbekende waarden $z_1, z_2, \ldots, z_{N-1}$ en verder de waarden z_0 en z_N waarvoor we de bekende waarden γ resp. δ van y(a) resp. y(b) nemen. We hebben dus N-1 vergelijkingen met N-1 onbekenden.

b) $\alpha_2 = 1$, $\beta_2 = 0$, $\beta_1 = 1$. De randvoorwaarden in x = a discretiseren we door

$$\alpha_1 z_0 + \frac{z_1 - z_{-1}}{2h} = \gamma$$

(deze formule is centraal t.o.v. het punt x_0 en heeft daarom een afbreekfout van hogere orde dan de voorwaartse formule $\alpha_1 z_0 + (z_1 - z_0)/h = \gamma$). Uit deze vergelijking en de discretisatie (12) van de differentiaalvergelijking in x = a (dus j = 0) elimineren we z_{-1}, hetgeen levert

$$-z_0 + z_1 - h(\gamma - \alpha_1 z_0) + \tfrac{1}{2}h^2 f(x_0, z_0, \gamma - \alpha_1 z_0) = 0.$$

Deze vergelijking, gecombineerd met (12) voor $j = 1,2,\ldots,N-1$ en de eindwaarde $z_N = \delta$ leveren nu N vergelijkingen voor de N onbekenden $z_0, z_1, \ldots, z_{N-1}$.

c) Overige gevallen kunnen op soortgelijke wijze afgedaan worden.

Men kan bewijzen dat, als f voldoende glad is, deze discretisatie een globale afbreekfout van de orde van h^2 heeft, d.w.z. dat voor $0 \le j \le N$

$$|z_j - y(x_j)| \le Ch^2.$$

En bij nadere beschouwing (en voldoende differentieerbaarheid van f) blijkt dat zelfs geldt

$$z_j = y(x_j) + c_1(x_j)h^2 + c_2(x_j)h^4 + \ldots .$$

Hierop kan weer h^2-extrapolatie gebaseerd worden.

Als de functie $f(x,y,\frac{dy}{dx})$ lineair is in y en $\frac{dy}{dx}$ dan zijn de verkregen stelsels lineair.

Als de differentiaalvergelijking de vorm (6) heeft en de randvoorwaarde luiden $y(a) = \gamma$, $y(b) = \delta$, dan gaat (12) over in

$$(1 - \tfrac{1}{2}ha_j)z_{j-1} - (2 - h^2b_j)z_j + (1 + \tfrac{1}{2}ha_j)z_{j+1} = h^2c_j, \qquad (13)$$

voor $j = 1,2,\ldots,N-1$ waarbij in de vergelijking voor $j = 1$ de term $(1 - \tfrac{1}{2}ha_1)z_0 = (1 - \tfrac{1}{2}ha_1)\gamma$ naar rechts gebracht moet worden. Analoog in de vergelijking voor $j = N-1$.

We hebben nu een groot (N is bv. 20 of 100) stelsel lineaire vergelijkingen gekregen. De bijbehorende matrix A is een zg. tridiagonaal matrix, d.w.z. $A_{jk} = 0$ voor $|k-j| > 1$.

Het oplossen van een stelsel vergelijkingen met een tridiagonaal matrix door middel van Gauss-eliminatie is eenvoudig.
We schrijven het stelsel kortheidshalve als

$$A_jz_{j-1} + B_jz_j + C_jz_{j+1} = D_j, \qquad 1 \le j \le N-1$$

(in de vergelijking met $j = 1$ is de term A_1z_0 afwezig, in die met $j = N-1$ de term $C_{N-1}Z_N$).
Bij vegen zonder rijverwisselingen wordt het verkregen driehoeksstelsel van de vorm

$$P_jz_j + C_jz_{j+1} = R_j \qquad 1 \le j \le N-1 \qquad (14)$$

(met $C_{N-1}Z_N$ afwezig).
Daarbij is $P_1 = B_1$, $R_1 = D_1$. En voor $j > 1$ moet (14) bepaald worden door eliminatie van z_{j-1} uit de geveegde j-1-de vergelijking en de j-de originele vergelijking:

$$P_{j-1}z_{j-1} + C_{j-1}z_j = R_{j-1}$$

$$A_jz_{j-1} + B_jz_j + C_jz_{j+1} = D_j,$$

dus

$$P_j = B_j - (A_j/P_{j-1})C_{j-1},$$
$$R_j = D_j - (A_j/P_{j-1})R_{j-1}.$$

Merk op dat het aantal hiervoor nodige bewerkingen van de orde N is (en niet N^3, zoals bij een volle matrix). Het zelfde geldt voor de terug substitutie in (14). Het oplossen van een tridiagonaalstelsel is dus goedkoop.
Men kan inzien, dat rijverwisseling niet nodig is indien voor de coëfficienten uit de differentiaalvergelijking geldt

$$b(x) - \frac{1}{4}a^2(x) - \frac{1}{2}a'(x) < 0.$$

Is hieraan niet voldaan, dan is eliminatie met zo nodig rijverwisseling noodzakelijk, dit is iets, maar omdat er nooit keuze is uit meer dan twee pivots, niet veel ingewikkelder.
Als de functie f niet lineair is, dan moet het stelsel (12) met de bijbehorende randvergelijkingen op de een of andere manier iteratief opgelost worden. Zo mogelijk gebruikt men hiervoor de methode van Newton (in het hier beschouwde verband ook wel quasi-linearisatie genoemd).
We beschouwen alleen het geval dat f niet van $\frac{dy}{dx}$ afhangt en dat de randvoorwaarden zijn $y(a) = y(b) = 0$. Op te lossen is dan het stelsel

$$z_{j-1} - 2z_j + z_{j+1} - h^2 f(x_j,z_j) = 0, \qquad (15)$$

$j = 1,\ldots,N-1$, met $z_0 = z_N = 0$.
Veronderstel nu dat de functie

$$g(x,y) := \frac{\partial f}{\partial y}(x,y)$$

bekend is (eventueel slechts bij benadering).
Veronderstel dat een benadering $z_1^{(\nu)},\ldots,z_{N-1}^{(\nu)}$ voor de oplossing bekend is.
Daar

$$f(x_j,z_j) = f(x_j,z_j^{(\nu)}) + g(x_j,z_j^{(\nu)})(z_j - z_j^{(\nu)}) + \ldots ,$$

kunnen we een (hopelijk) betere benadering $z_1^{(\nu+1)},\ldots,z_{N-1}^{(\nu+1)}$ bepalen uit

$$z_{j-1}^{(\nu+1)} - 2z_j^{(\nu+1)} + z_{j+1}^{(\nu+1)} - h^2[f(x_j,z_j^{(\nu)}) + g(x_j,z_j^{(\nu)}) \cdot (z_j^{(\nu+1)} - z_j^{(\nu)})] = 0,$$

dus uit

$$z_{j-1}^{(\nu+1)} - (2 + h^2 g(x_j,z_j^{(\nu)}))z_j^{(\nu+1)} + z_{j+1}^{(\nu+1)} = h^2[f(x_j,z_j^{(\nu)}) - g(x_j,z_j^{(\nu)})z_j^{(\nu)}].$$

Dit is een lineair stelsel met een tridiagonaal matrix. (We moeten $z_0^{(\nu+1)} = z_N^{(\nu+1)} = 0$ stellen). Het is duidelijk dat als dit iteratieproces convergeert voor $\nu \to \infty$, de limietrij de oplossing van (15) is. In het algemeen zal de convergentie kwadratisch zijn, d.w.z.

$$\max_j |z_j^{(\nu+1)} - z_j| \leq C(\max_j |z_j^{(\nu)} - z_j|)^2.$$

5. Partiële differentiaalvergelijkingen

De in de mathematische fysica voorkomende partiële differentiaalvergelijkingen zijn in veel gevallen van de tweede orde. Als er twee onafhankelijke variabelen zijn, dan zien ze er uit als

$$A \frac{\partial^2 u}{\partial x^2} + 2B \frac{\partial^2 u}{\partial x \partial y} + C \frac{\partial^2 u}{\partial y^2} = D. \qquad (1)$$

Hierin zijn de coëfficienten A, B, C en D in het algemeen nog van x, y, u, $\frac{\partial u}{\partial x}$ en $\frac{\partial u}{\partial y}$ afhankelijk. Als A, B en C alleen van x en y afhangen en D lineair afhangt van u, $\frac{\partial u}{\partial x}$ en $\frac{\partial u}{\partial y}$ dan is de vergelijking lineair.

Het blijkt dat we de vergelijkingen (1) in drie groepen moeten splitsen, nl.

a) Elliptische vergelijkingen. Hier is $AC > B^2$. Prototype voor dit geval is de potentiaalvergelijking

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y). \qquad (2)$$

b) Hyperbolische vergelijkingen. Hier is $AC < B^2$. Prototype is de golfvergelijking

$$\frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial x^2} + f(x,y). \qquad (3)$$

c) Parabolische vergelijkingen. Hier is $AC = B^2$. Prototype is de warmtegeleidingsvergelijking

$$\frac{\partial u}{\partial y} = \frac{\partial^2 u}{\partial x^2} + f(x,y). \qquad (4)$$

Naast deze tweede orde vergelijkingen noemen we nog stelsels van de eerste orde, die, als er twee onafhankelijke variabelen zijn, er uit zien als

$$\frac{\partial \underline{u}}{\partial y} = A \frac{\partial \underline{u}}{\partial x} + \underline{b}. \qquad (5)$$

Hierin zijn $\underline{u}$ en $\underline{b}$ vectoren met n componenten, A een n×n matrix; A en $\underline{b}$ hangen in het algemeen van x, y en $\underline{u}$ af. Als A niet en $\underline{b}$ lineair van $\underline{u}$ afhangt (dus $\underline{b}(x,y,\underline{u}) = B(x,y)\underline{u} + \underline{c}(x,y)$) dan is het stelsel lineair. Als de matrix A uitsluitend reële, onderling verschillende eigenwaarden heeft, dan heet het stelsel hyperbolisch.

Opmerking. Het hyperbolische stelsel

$$\frac{\partial u_1}{\partial y} = \frac{\partial u_2}{\partial x}\ ,\ \frac{\partial u_2}{\partial y} = \frac{\partial u_1}{\partial x} + g(x,y)$$

is equivalent met $\frac{\partial^2 u_1}{\partial y^2} = \frac{\partial^2 u_1}{\partial x^2} + \frac{\partial g}{\partial x}$, dus met (3).

Een partiële differentiaalvergelijking geldt meestal maar in een bepaald gebied van het x,y-vlak en aan de rand van dit gebied moeten randvoorwaarden gegeven zijn. Veel voorkomende gebieden en randvoorwaarden zijn:

a) elliptische vergelijkingen:
gebied: een begrensd of onbegrensd gebied G in het x,y-vlak.
randvoorwaarden: langs de hele rand van G één randvoorwaarde, bv. u of $\frac{\partial u}{\partial n}$ (normale afgeleide) gegeven.

b) hyperbolische vergelijkingen:
gebied: in de positieve y-richting (meestal is de y-variabele de tijd) onbegrensd; in de x-richting bv. $a < x < y$ (met eventueel $a = -\infty$ en/of $b = \infty$).
randvoorwaarden: langs een beginkromme (meestal $y = 0$) zijn u en $\frac{\partial u}{\partial y}$ gegeven (als regel is y de tijd, dan zijn dit zg. beginvoorwaarden), langs randen in de x-richting u of $\frac{\partial u}{\partial x}$ gegeven.

c) parabolische vergelijkingen:
als bij hyperbolische vergelijkingen, langs de beginkromme echter alleen u gegeven.

5.1. De warmtegeleidingsvergelijking

We behandelen de lineaire vergelijking

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a \frac{\partial u}{\partial x}\right) + bu + c, \qquad (1)$$

waarin a, b en c functies van x en t mogen zijn. $a(x,t)$ is overal positief. We gaan de vergelijking (1) discretiseren. Daartoe voeren we roosterpunten $x_j = x_0 + jh$, $t_n = t_0 + nk$ (met h de gekozen stapgrootte in de x-richting en k die in de t-richting en (x_0,t_0) een geschikt gekozen startpunt) in. We noemen

$$u_{j,n} := u(x_j,t_n).$$

5.1.1. De methode van Euler

Als u(x,t) aan (1) voldoet dan geldt (ga na)

$$\frac{u_{j,n+1} - u_{j,n}}{k} = \frac{1}{h}\left(a_{j+\frac{1}{2},n}\,\frac{u_{j+1,n} - u_{j,n}}{h} - a_{j-\frac{1}{2},n}\,\frac{u_{j,n} - u_{j-1,n}}{h}\right)$$

$$+ b_{j,n}\,u_{j,n} + c_{j,n} + R_{j,n}\,, \qquad (2)$$

waarin

$$|R_{j,n}| \leq C_1\,h^2 + C_2 k. \qquad (3)$$

Uit (2) volgt als methode (de door de methode verkregen benadering voor $u_{j,n}$ noemen we $v_{j,n}$)

$$v_{j,n+1} = v_{j,n} + \frac{k}{h^2}\left[a_{j+\frac{1}{2},n}(v_{j+1,n} - v_{j,n}) - a_{j-\frac{1}{2},n}(v_{j,n} - v_{j-1,n})\right]$$

$$+ k(b_{j,n}\,v_{j,n} + c_{j,n}). \qquad (4)$$

Dit is een volledig expliciete methode die we kunnen starten met

$$v_{j,0} = u_{j,0} \quad (= \text{de gegeven beginwaarden}).$$

Ook eventuele randvoorwaarden leveren geen bijzondere moeilijkheden.

Het blijkt echter dat deze methode zich instabiel gaat gedragen als

$$\frac{k}{h^2}\,a > \tfrac{1}{2},$$

in de zin dat in dit geval de invloed van een verandering in de beginwaarden (of van een tussentijdse fout) op de gevonden waarden voor een vast gekozen tijd t groter wordt naarmate h en k kleiner worden. We lichten dit toe aan het geval a = 1, b = c = 0. Het differentieschema luidt dan

$$v_{j,n+1} = v_{j,n} + \alpha(v_{j+1,n} - 2v_{j,n} + v_{j-1,n})$$

$$= (1 - 2\alpha)v_{j,n} + \alpha(v_{j+1,n} + v_{j-1,n}), \qquad (5)$$

waarin

$$\alpha = k/h^2.$$

Voor $\alpha \leq \frac{1}{2}$ is $1 - 2\alpha \geq 0$ en daaruit volgt direct (ga na)

$$\max_j |v_{j,n+1}| \leq \max_j |v_{j,n}| \leq \ldots \leq \max_j |v_{j,0}|. \tag{6}$$

In dit geval kan dus geen versterking van storingen optreden.
Om in te zien dat het mis kan gaan als $\alpha > \frac{1}{2}$ kiezen we als beginvoorwaarde

$$v_{j,0} = \delta_{j,0} := \begin{cases} 1 \text{ als } j = 0 \\ \\ 0 \text{ als } j \neq 0 \end{cases}. \tag{7}$$

Stellen we

$$w_{j,n} = (-1)^{j+n} v_{j,n},$$

dan volgt uit (5)

$$w_{j,n+1} = (2\alpha - 1)w_{j,n} + \alpha(w_{j+1,n} + w_{j-1,n}),\ w_{j,0} = \delta_{j,0}.$$

Als $\alpha \geq \frac{1}{2}$ dan volgt hieruit dat alle $w_{j,n} \geq 0$ zullen zijn, dat $w_{j,n} = 0$ voor $|j| > n$ en dat

$$\sum_{j=-\infty}^{\infty} |v_{j,n+1}| = \sum_j w_{j,n+1} = (4\alpha - 1) \sum_j w_{j,n}$$
$$= (4\alpha - 1)^{n+1} \sum_j w_{j,0} = (4\alpha - 1)^{n+1}.$$

Hieruit volgt

$$\max_j |v_{j,n}| \geq \frac{(4\alpha - 1)^n}{2n+1}.$$

Laten we nu h en k naar 0 gaan zo, dat $\alpha = k/h^2 \geq \alpha_0 > \frac{1}{2}$ blijft, en nemen we $n = T/k$, dan geldt

$$\max_j |v_{j,n}| \geq \frac{(4\alpha_0 - 1)^{T/k}}{2(T/k) + 1} \to \infty \text{ als } k \to 0.$$

Het effect van een storing van de vorm (7) neemt dus exponentieel toe als $k \to 0$ (de veronderstelling dat de storingsbron $v_{0,0}$ een factor k^p bevat zou niet helpen!). Daar de differentievergelijking (5) lineair is, geldt deze bewering ook voor storingen van een algemener karakter.

De methode (4) is derhalve alleen bruikbaar als $k \leq \frac{1}{2}h^2/a$. Kiezen we voor k/h^2 een vaste factor dan blijkt uit (3) dat de locale afbreekfout van de orde k is, dit blijkt ook voor de globale afbreekfout het geval te zijn.

5.1.2. De methode van Crank-Nicolson (trapeziumregel)

Ter wille van het schrijfwerk bespreken we deze methode alleen voor de vergelijking

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + bu + c. \tag{1}$$

Als u(x,t) aan (1) voldoet, dan geldt (reeksontwikkeling t.o.v. het punt $x_j, t_{n+\frac{1}{2}}$)

$$\frac{u_{j,n+1} - u_{j,n}}{k} = \frac{1}{2}\,\frac{u_{j+1,n+1} - 2u_{j,n+1} + u_{j-1,n+1}}{h^2}$$

$$+ \frac{1}{2}\,\frac{u_{j+1,n} - 2u_{j,n} + u_{j-1,n}}{h^2}$$

$$+ \tfrac{1}{2}\,(b_{j,n+1}\,u_{j,n+1} + c_{j,n+1}) + \tfrac{1}{2}(b_{j,n}\,u_{j,n} + c_{j,n})$$

$$+ R_{j,n+\frac{1}{2}}\,, \tag{2}$$

met $|R_{j,n+\frac{1}{2}}| \leq C_1 h^2 + C_2 k^2$. (3)

Hieruit volgt als methode (met $\alpha = k/h^2$)

$$(1 + \alpha - \tfrac{1}{2}\,kb_{j,n+1})v_{j,n+1} - \tfrac{1}{2}\,\alpha(v_{j+1,n+1} + v_{j-1,n+1})$$

$$= (1 - \alpha + \tfrac{1}{2}\,kb_{j,n})v_{j,n} + \tfrac{1}{2}\,\alpha(v_{j+1,n} + v_{j-1,n}) + \frac{k}{2}(c_{j,n+1} + c_{j,n}). \tag{4}$$

We krijgen dus voor iedere j een vergelijking tussen $v_{j,n+1}$, $v_{j+1,n+1}$ en $v_{j-1,n+1}$.

Stel nu dat we als gebied hebben $a < x < b$, en dat de randvoorwaarden luiden

$$u(a,t) = \gamma(t),\ u(b,t) = \delta(t).$$

Nemen we $x_0 = a$, $h = (b - a)/N$, dan levert dit

$$v_{0,n} = \gamma_n,\ v_{N,n} = \delta_n,\ v_{0,n+1} = \gamma_{n+1},\ v_{N,n+1} = \delta_{n+1} \qquad (5)$$

De vergelijkingen (4) moeten nu gelden voor $1 \le j \le N-1$. Daar $v_{0,n+1}$ en $v_{N,n+1}$ bekend zijn, levert dit N-1 vergelijkingen voor $v_{1,n+1},\ldots,v_{N-1,n+1}$. De matrix van dit stelsel is tridiagonaal. Daarom is de bepaling van de $v_{j,n+1}$ wel iets lastiger maar niet essentieel tijdrovender dan bij een expliciete methode(de hoeveelheid werk per tijdstap is in beide gevallen evenredig met N).

We beperken ons nu tot het geval $b = c = \gamma = \delta = 0$. Dan is vrij eenvoudig in te zien dat de methode voor alle α stabiel is. We hebben nu te maken met het stelsel

$$\begin{aligned} &v_{j,n+1} - v_{j,n} = \tfrac{1}{2}\alpha(\delta^2 v_{j,n+1} + \delta^2 v_{j,n}),\ 1 \le j \le N-1 \\ &v_{0,n+1} = v_{N,n+1} = 0 \end{aligned} \qquad (5)$$

(waarin $\delta^2 v_{j,n} := v_{j+1,n} - 2v_{j,n} + v_{j-1,n}$, etc.),

Zij de beginvoorwaarde $v_{j,0} = f_j$, $0 \le j \le N$. (met $f_0 = f_N = 0$).

Uit (5) volgt na vermenigvuldiging met $w_j := v_{j,n+1} + v_{j,n}$ en sommatie over j

$$\sum_{j=1}^{N-1} v_{j,n+1}^2 - \sum_{j=1}^{N-1} v_{j,n}^2 = \tfrac{1}{2}\alpha \sum_{1}^{N-1} w_j\, \delta^2 w_j$$

$$= \tfrac{1}{2}\alpha \sum_{1}^{N-1} (w_j\, w_{j+1} + w_j\, w_{j-1} - 2w_j^2)$$

$$= \tfrac{1}{2}\alpha \sum_{0}^{N-1} (2w_j\, w_{j+1} - w_j^2 - w_{j+1}^2)$$

$$= -\tfrac{1}{2}\alpha \sum_{0}^{N-1} (w_{j+1} - w_j)^2 \le 0.$$

(we hebben hier gebruikt dat $w_0 = w_N = 0$).

Hieruit volgt

$$\sum_{j=1}^{N-1} v_{j,n+1}^2 \leq \sum_{j=1}^{N-1} v_{j,n}^2 \leq \ldots \leq \sum_{j=1}^{N-1} f_j^2.$$

Het effect van een storing in de beginwaarden blijft dus begrensd, hetgeen stabiliteit betekent.

Men kan bewijzen dat bij deze methode ook de globale afbreekfout van de orde $k^2 + h^2$ is.

Impliciete methoden, zoals de hier beschreven methode van Crank-Nicolson verdienen in zo sterke mate de voorkeur boven expliciete methoden (waarbij steeds een ernstig beperkende stabiliteitseis aanwezig blijkt te zijn), dat zij als regel ook gebruikt worden bij niet-lineaire problemen waarbij de $u_{j,n+1}$ uit een niet lineair stelsel iteratief opgelost moeten worden (zo mogelijk gebeurt dit oplossen weer met een variant van het Newton proces, zoals in 4.4.2. aangeduid is).

Opmerking.
Een fraai symmetrische discretisatie voor (1) is het twee-stapsschema

$$v_{j,n+1} = v_{j,n-1} + 2\alpha(v_{j+1,n} - 2v_{j,n} + v_{j-1,n}) + 2k(b_{j,n}\, v_{j,n} + c_{j,n}).$$

Deze methode blijkt echter voor alle waarden van α instabiel te zijn!

5.2. De golfvergelijking

Het ligt voor de hand, de golfvergelijking

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t) \qquad (1)$$

te discretiseren door

$$\frac{1}{k^2}(u_{j,n+1} - 2u_{j,n} + u_{j,n-1}) = \frac{c^2}{h^2}(u_{j+1,n} - 2u_{j,n} + u_{j-1,n}) + f_{jn} + R_{jn}$$

met $|R_{jn}| \leq C_1 h^2 + C_2 k^2$.

Hieruit volgt als methode

$$v_{j,n+1} - 2v_{j,n} + v_{j,n-1} = \alpha^2(v_{j+1,n} - 2v_{j,n} + v_{j-1,n}) + k^2 f_{j,n} \quad (2)$$

Hierin is $\alpha = kc/h$.
Als de beginvoorwaarden zijn

$$u(x,0) = \varphi(x), \quad \frac{\partial u}{\partial t}(x,0) = \psi(x), \quad (3)$$

dan zijn passende beginvoorwaarden voor de differentievergelijking

$$\begin{aligned} v_{j,0} &= \varphi_j; \\ v_{j,1} - v_{j,-1} &= 2k\psi_j. \end{aligned} \quad (4)$$

Samen met (2) (met $j = 0$) levert de tweede voorwaarde

$$v_{j,1} = \varphi_j + k\psi_j + \tfrac{1}{2}\alpha^2(\varphi_{j+1} - 2\varphi_j + \varphi_{j-1}) + \tfrac{1}{2}k^2 f_{j,0}. \quad (5)$$

Het blijkt dat dit differentieschema stabiel is voor $|\alpha| \leq 1$ en instabiel voor $|\alpha| > 1$. Voorts is er ook convergentie als $|\alpha| \leq 1$: bij iedere $T > 0$ is er een constante $C(T)$ zodanig dat voor alle n met $t_n \leq T$ geldt

$$|v_{j,n} - u(x_j,t_n)| \leq C(T).h^2.$$

De globale convergentie orde is dus 2.

Dat de voorwaarde $|\alpha| \leq 1$ nodig is voor convergentie wordt ook duidelijk als men de exacte oplossing van de differentiaalvergelijking (1) beschouwt. Voor het geval dat $f(x,t) = 0$, $\varphi(x) = 0$ luidt deze

$$u(x,t) = \frac{1}{2c} \int_{x-ct}^{x+ct} \psi(\xi)d\xi.$$

De waarde van $u(x,t)$ is dus bepaald door de waarden $\psi(\xi)$ op het interval $|\xi-x| \leq ct$. Anderzijds volgt uit (2) met (4) en (5) dat (als $f = 0$, $\varphi = 0$) de waarde $v_{j,n}$ bepaald is door de waarde ψ_i met $|i-j| < n$, dus door de waarden $\psi(\xi_i)$ met $|\xi_i - x_j| < nh = t_n h/k = ct_n/\alpha$.

Laten we nu bij vaste $\alpha > 1$ h en k naar 0 gaan en nemen we j en n zo, dat $x_j \to x$, $t_n \to t$, dan zou, als $v_{j,n}$ een limiet had, deze limiet bepaald zijn door de waarden van $\psi(\xi)$ met $|\xi - x| < ct/\alpha$, dus uit een kleiner interval dan bij de differentiaalvergelijking. Maar dan kan nooit voor alle ψ deze limiet gelijk zijn aan $u(x,t)$. Neem bijv. $\psi(\xi) = 0$ voor $|\xi - x| \leq ct/\alpha$, $\psi(\xi) > 0$ voor $|\xi - x| > ct/\alpha$.

In het algemeen heeft men bij de vervanging van een hyperbolische differentiaalvergelijking door een expliciete differentievergelijking steeds te maken met een stabiliteitsvoorwaarde.
Voor het stelsel

$$\frac{\partial \underline{u}}{\partial t} = A(x,t)\,\frac{\partial \underline{u}}{\partial x} + \underline{b}(x,t)$$

is een mogelijk differentieschema bv.

$$\underline{v}_{j,n+1} - \underline{v}_{j,n-1} = \frac{k}{h} A_{jn} (\underline{v}_{j+1,n} - \underline{v}_{j-1,n}) + 2k\underline{b}_{j,n}.$$

Dit is stabiel indien voor iedere x en t geldt dat $(k/h) \times \max_{x,t} \|A(x,t)\| \leq 1$ is (met $\| \ \|$ een matrixnorm).

5.3. De potentiaalvergelijking

Een voor de hand liggende discretisatie voor de potentiaalvergelijking

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y) \qquad (1)$$

is

$$\frac{1}{h^2}(4u_{jn} - u_{j,n+1} - u_{j,n-1} - u_{j+1,n} - u_{j-1,n}) = f_{jn} + R_{jn} \qquad (2)$$

met $|R_{jn}| \leq C_1 h^2$.

(Het ligt uiteraard voor de hand om de stapgrootten in x- en y-richting gelijk te maken).
Uit (2) volgt als methode

$$4v_{j,n} - v_{j,n+1} - v_{j,n-1} - v_{j+1,n} - v_{j-1,n} = h^2 f_{j,n}. \qquad (3)$$

We beschouwen alleen het eenvoudigste geval dat het gebied waar (1) geldt de rechthoek $0 < x < Mh$, $0 < y < Nh$ is en dat langs de rand de waarden van u gegeven zijn.

Dan geldt (3) voor $1 \leq j \leq M-1$, $1 \leq n \leq N-1$ en de waarden $v_{j,0}$, $v_{j,N}$, $v_{0,n}$, $v_{M,n}$ zijn bekend. Het stelsel (3) is dan een stelsel van $(M-1) \times (N-1)$ lineaire vergelijkingen voor de $(M-1) \times (N-1)$ onbekenden $v_{j,n}$, $1 \leq j \leq M-1$, $1 \leq n \leq N-1$.

Men kan bewijzen dat dit stelsel altijd een eenduidige oplossing heeft en dat de globale afbreekfout van het resultaat van de orde h^2 is.

Als M en N groot zijn, dan is het stelsel (3) te groot om met een eliminatie methode oplosbaar te zijn. Anderzijds ligt het voor de hand om de vergelijkingen te schrijven als

$$v_{j,n} = \tfrac{1}{4}(v_{j,n+1} + v_{j,n-1} + v_{j+1,n} + v_{j-1,n} + h^2 f_{j,n}) \qquad (4)$$

en successieve substitutie toe te passen, het zij volgens Jacobi, het zij volgens Gauss-Seidel. (zie 2.3.1). Convergeren deze processen? Het stelsel (4) is slechts zg. zwak diagonaal overwegend: als er voor zekere j en n rechts in (4) randpunten voorkomen is de som van de coëfficiënten van de onbekende v's rechts in (4) kleiner dan 1, anders gelijk aan 1. Men kan echter bewijzen dat deze eigenschap voldoende is voor de convergentie, zowel van het Jacobi- als van het Gauss-Seidel proces. Wel is voor grote M en N de convergentie van beide processen erg langzaam. Voor de rechthoek kan men de convergentiefaktor uitrekenen. Voor Jacobi bedraagt deze

$$\lambda_J := 1 - \sin^2 \frac{\pi}{2M} - \sin^2 \frac{\pi}{2N} \sim 1 - \frac{\pi^2}{4} \cdot \frac{M^2 + N^2}{M^2 N^2} \sim \exp\left(-\frac{\pi^2}{2} \cdot \frac{M^2 + N^2}{2M^2 N^2}\right) .$$

Dit betekent dat, als $M \sim N$, circa $2N^2/\pi^2$ slagen gedaan moeten worden om de fout met een factor e^{-1} te doen afnemen.

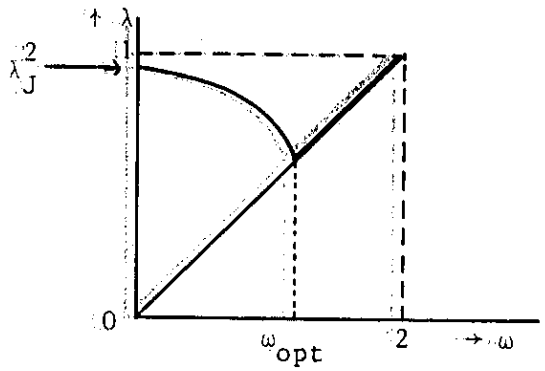
Voor Gauss-Seidel is de convergentiefaktor λ_J^2, de convergentie is dan ca 2 maal zo snel.

Essentieel sneller convergeert de variant van (4) die we krijgen door (4) te schrijven als

$$v_{j,n} = v_{j,n} + \frac{\omega}{4} (v_{j,n+1} + v_{j,n-1} + v_{j+1,n} + v_{j-1,n} - 4v_{jn} + h^2 f_{jn})$$

en daarop Gauss-Seidel toe te passen, na een geschikte keuze van ω.

Dit proces heet systematische overrelaxatie (S.O.R.), ω heet overrelaxatiefactor. Het blijkt dat de convergentiefaktor λ van dit proces zich voor $1 \leq \omega \leq 2$ gedraagt als in onderstaande grafiek aangegeven.



De gunstigste waarde ω_{opt} is

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \lambda_J^2}}$$

en hier is

$$\lambda = \lambda_{opt} = \frac{1 - \sqrt{1 - \lambda_J^2}}{1 + \sqrt{1 - \lambda_J^2}}.$$

Voor de rechthoek geldt voor grote M en N

$$\lambda_{opt} \sim 1 - 2\pi\sqrt{\frac{M^2 + N^2}{2M^2N^2}} \sim \exp\left(-2\pi\sqrt{\frac{M^2 + N^2}{2M^2N^2}}\right)$$

zodat voor $M \sim N$ nu ca $N/(2\pi)$ slagen gedaan moeten worden om een factor e^{-1} te winnen. Dit betekent een factor $\frac{\pi N}{4}$ winst vergeleken bij het Gauss-Seidel proces.

Het hierboven geschetste geldt in grote trekken ook voor meer algemene elliptische vergelijkingen met algemenere gebieden en algemenere randvoorwaarden.

Ook hier is S.O.R. veelal een redelijk convergerend proces. Er bestaan technieken om de waarde van ω_{opt} experimenteel te bepalen.

Naast S.O.R. bestaan er nog verscheidene andere iteratieve methoden om stelsels als (3) op te lossen.

6. Interpolatie en approximatie

6.1. Polynoom interpolatie

Zij van een functie $f(x)$ functiewaarden $f(x_0),\ldots,f(x_n)$ in $n+1$ verschillende punten gegeven. Dan is er precies één polynoom $p(x)$ met de eigenschappen

i) graad $p \leq n$

ii) $p(x_j) = f(x_j)$, $j = 0,\ldots,n$.

Immers, voldeden zowel $p(x)$ als $q(x)$, dan was $p(x) - q(x)$ een polynoom met graad $\leq n$ en $n+1$ verschillende nulpunten $x_0,\ldots,x_n$. Er is dus hoogstens één polynoom dat voldoet. Dit polynoom bestaat, want we kunnen het als volgt construeren. Definieer de $n+1$ zg. polynomen van Lagrange door

$$L_k(x) = \prod_{\substack{i=0 \\ i\neq k}}^{n} \frac{x - x_i}{x_k - x_i} = \frac{(x - x_0)(x - x_1) \overset{(k)}{\cdots} (x - x_n)}{(x_k - x_0)(x_k - x_1) \overset{(k)}{\cdots} (x_k - x_n)} , \quad k = 0,\ldots,n .$$

(Met het symbool $\overset{(k)}{\cdots}$ is aangegeven dat in de teller de factor $x - x_k$ en in de noemer de factor $x_k - x_k$ ontbreekt.)
De $L_k(x)$ zijn polynomen van de graad n en

$$L_k(x_j) = \begin{cases} 0 & \text{als } j \neq k \\ 1 & \text{als } j = k . \end{cases}$$

Hieruit volgt direct dat

$$p(x) = \sum_{k=0}^{n} f(x_k)L_k(x) \qquad (1)$$

het polynoom is dat aan i) en ii) voldoet.

We noemen (1) het interpolatiepolynoom volgens Lagrange behorende bij de punten $x_0,\ldots,x_n$ en de functiewaarden $f(x_0),\ldots,f(x_n)$ (let op: we hebben bewezen dat er bij deze punten en functiewaarden maar één interpolatiepolynoom met graad $\leq n$ bestaat; er zijn echter verschillende schrijfwijzen voor dit polynoom, vandaar de toevoeging: volgens Lagrange).

Levert $p(x)$ nu voor tussenliggende punten een goede benadering voor $f(x)$? Dat hangt geheel af van de "gladheid" van $f(x)$. De volgende stelling geldt:

Als $f(x)$ $n+1$ maal differentieerbaar is dan is er bij iedere x een ξ_x waarvoor geldt

$$\min(x,x_0,\ldots,x_n) < \xi_x < \max(x,x_0,\ldots,x_n)$$

en

$$f(x) = \sum_{k=0}^{n} f(x_k)L_k(x) + R(x) ,$$

met

$$R(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \frac{f^{(n+1)}(\xi_x)}{(n + 1)!} .$$

We bewijzen deze stelling niet. De vorm van de restterm $R(x)$ is gemakkelijk te onthouden door voor $f(x)$ een polynoom met graad $n+1$ te nemen. $R(x)$ moet dan een polynoom met graad $n+1$ zijn dat nul is in $x = x_0,x_1,\ldots,x_n$, dus $R(x) = C(x - x_0) \ldots (x - x_n)$. Daar $R^{(n+1)}(x) = f^{(n+1)}(x)$ (waarom?) is de waarde van de constante C direct te vinden.

Veronderstel nu dat $x_0 < x_1 < \ldots < x_n$, dat $x_{j+1} - x_j \leq h$ en dat

$$|f^{(n+1)}(x)| \leq M_{n+1}$$

voor $x_0 \leq x \leq x_n$.
Dan geldt, als $x_{k-1} \leq x \leq x_k$ $(1 \leq k \leq n)$,

$$|R(x)| \leq \frac{k(k - 1) \ldots \cdot 1 \cdot 1 \ldots \cdot (n - k)(n + 1 - k)}{(n + 1)!} h^{n+1} M_{n+1} =$$

$$= \frac{k!(n + 1 - k)!}{(n + 1)!} h^{n+1} M_{n+1} .$$

Hieruit zien we: deze bovengrens voor de afbreekfout $|R(x)|$ wordt kleiner als h kleiner wordt (punten dichter bij elkaar) en wel evenredig met h^{n+1}. Voorts is de bovengrens kleiner (bij vaste h en n) naarmate k dichter bij $(n+1)/2$ ligt (dus x dichter bij de middelste van de getallen $x_0,\ldots,x_n$). Of vergroting van n de afbreekfout kleiner maakt hangt geheel af van het gedrag van de afgeleiden van $f(x)$.

Voor praktisch gebruik is de schrijfwijze van het interpolatiepolynoom volgens Lagrange niet erg geschikt. Er bestaan diverse andere schrijfwijzen (o.a. volgens Newton). Een prettige algoritme om een waarde $p(x)$ uit te rekenen is die van Aitken en Neville. Deze is gebaseerd op de volgende

Stelling. Zij voor $0 \le \ell \le m \le n$ $p_{\ell m}(x)$ het interpolatiepolynoom met graad $\le m-\ell$ en steunpunten $x_\ell,\ldots,x_m$ (dus $p_{\ell\ell}(x) = f(x_\ell)$). Dan geldt, als $\ell < m$,

$$p_{\ell m}(x) = \frac{(x_m - x)p_{\ell,m-1}(x) + (x - x_\ell)p_{\ell+1,m}(x)}{x_m - x_\ell} . \qquad (2)$$

Immers, $p_{\ell,m-1}$ en $p_{\ell+1,m}$ hebben graad $\le m-\ell-1$, dus $p_{\ell m}$ heeft graad $\le m-\ell$. Uit

$$p_{\ell,m-1}(x_j) = f(x_j) \quad \text{voor } \ell \le j \le m-1$$

en

$$p_{\ell+1,m}(x_j) = f(x_j) \quad \text{voor } \ell+1 \le j \le m$$

volgt (ga na) dat $p_{\ell,m}(x_j) = f(x_j)$ voor $\ell \le j \le m$.

Stel nu dat $x_0 < x_1 < \ldots < x_n$. Stel dat bv. $x_4 < x < x_5$ en dat we $f(x)$ willen benaderen met de waarde van het derdegraads interpolatiepolynoom $p_{3456}(x)$. Dan moeten we de volgende getallen bepalen

$$\begin{array}{lllll} p_3(x) = f_3 & & & & \\ & p_{34}(x) & & & \\ p_4(x) = f_4 & & p_{345}(x) & & \\ & p_{45}(x) & & p_{3456}(x) & \\ p_5(x) = f_5 & & p_{456}(x) & & \\ & p_{56}(x) & & & \\ p_6(x) = f_6 & & & & \end{array}$$

Het is duidelijk dat $p_{45}(x)$ een benadering voor $f(x)$ is met behulp van lineaire interpolatie, $p_{345}(x)$ en $p_{456}(x)$ zijn benaderingen met kwadratische interpolatie. Het verschil tussen deze waarden en $p_{3456}(x)$ geeft een indicatie van de verkregen nauwkeurigheid.

Opmerking. In het geval $n = 1$ (lineaire interpolatie) is de interpolatieformule met basispunten x_0 en x_1

$$p(x) = \frac{(x_1 - x)f(x_0) + (x - x_0)f(x_1)}{x_1 - x_0} =$$

$$= f(x_0) + (x - x_0)\frac{f(x_1) - f(x_0)}{x_1 - x_0} =$$

$$= f(x_1) + (x - x_1)\frac{f(x_1) - f(x_0)}{x_1 - x_0} .$$

Voor tweemaal differentieerbare functies geldt in dit geval

$$f(x) = p(x) + \tfrac{1}{2}(x - x_0)(x - x_1)f''(\xi_x) ,$$

met

$$\min(x,x_0,x_1) < \xi_x < \max(x,x_0,x_1) .$$

6.2. Polynoom interpolatie bij equidistante abscissen

Voor het geval dat de opvolgende basispunten onderling gelijke afstanden

$$x_{j+1} - x_j = h$$

hebben bestaat een heel arsenaal van klassieke interpolatieformules (Gregory-Newton, Gauss, Bessel, Stirling, Everett, etc.). Deze hebben thans, nu niet veel meer met tabellen gewerkt wordt, hun betekenis gedeeltelijk verloren.

We noemen slechts een driepunts en een vierpunts formule, die voor de meeste praktische gevallen voldoende zijn.

6.2.1. Driepunts formule van Stirling

Als $f(x_{-1})$, $f(x_0)$, $f(x_1)$ bekend zijn (we noemen ze f_{-1}, f_0, f_1) en

$$f_s := f(x_0 + sh) ,$$

dan geldt

$$f_s = f_0 + s\frac{f_1 - f_{-1}}{2} + s^2\frac{f_1 - 2f_0 + f_{-1}}{2} + \frac{1}{6}s(s^2 - 1)h^3 f^{(3)}(\xi_s) ,$$

met (als $|s| \leq 1$) $x_{-1} < \xi_s < x_1$.

Men kan deze formule eenvoudig afleiden uit de driepunts formule van Lagrange!

Het bijbehorende tweedegraads interpolatiepolynoom is dus

$$p_s = f_0 + s\frac{f_1 - f_{-1}}{2} + s^2\frac{f_1 - 2f_0 + f_{-1}}{2} .$$

Men zal dit polynoom natuurlijk vooral gebruiken voor $-\frac{1}{2} \leq s \leq \frac{1}{2}$.

Bij praktisch gebruik kan men de grootheid $h^3 f^{(3)}(\xi)$ uit de restterm schatten met behulp van een der differentiatieformules

$$\delta^2 f_1 - \delta^2 f_0 = f_2 - 3f_1 + 3f_0 - f_{-1} = h^3 f^{(3)}(x_{\frac{1}{2}}) + O(h^5)$$

$$\delta^2 f_0 - \delta^2 f_{-1} = f_1 - 3f_0 + 3f_{-1} - f_{-2} = h^3 f^{(3)}(x_{-\frac{1}{2}}) + O(h^5)$$

$$\tfrac{1}{2}(\delta^2 f_1 - \delta^2 f_{-1}) = \tfrac{1}{2}(f_2 - 2f_1 + 2f_{-1} - f_{-2}) = h^3 f^{(3)}(x_0) + O(h^5) \ .$$

Hierin is $\delta^2 f_j := f_{j+1} - 2f_j + f_{j-1}$.

6.2.2. Vierpunts formule van Everett

Als $f_j = f(x_j)$ bekend zijn voor $j = -1,0,1,2$, en

$$f_s = f(x_0 + sh) \ , \quad t = 1 - s \ ,$$

dan geldt

$$f_s = tf_0 + sf_1 + \frac{1}{6}\{t(t^2 - 1)\delta^2 f_0 + s(s^2 - 1)\delta^2 f_1\} +$$

$$+ \frac{1}{24} st(1 + s)(1 + t)h^4 f^{(4)}(\xi_s) \ ,$$

met (als $-1 \leq s \leq 2$) $x_{-1} < \xi_s < x_2$.

Deze formule kan uit de vierpunts Lagrange formule afgeleid worden.
Men zal het bijbehorende interpolatiepolynoom natuurlijk vooral gebruiken voor $0 < s < 1$.
Merk op dat de eerste twee termen corresponderen met lineaire interpolatie. De derde term geeft daar een correctie op.

De grootte van de restterm kan weer geschat worden met een differentieformule, bv.

$$\delta^4 f_j := \delta^2 f_{j-1} - 2\delta^2 f_j + \delta^2 f_{j-1} = h^4 f^{(4)}(x_j) + O(h^6) \ .$$

6.3. Interpolatie met zg. spline functies

Veronderstel dat een functie f bekend is in n+1 punten $x_0,\ldots,x_n$ (al dan niet equidistant). Om f te benaderen in de tussenliggende punten kunnen we een n-de graads interpolatiepolynoom gebruiken. In het algemeen is dit voor $n \geq 5$ echter niet aan te bevelen, onder meer omdat de waarden van dit polynoom erg gevoelig zijn voor afrondfouten in de functiewaarden $f_0,f_1,\ldots,f_n$. Liever benaderen we f(x) stuksgewijs door lagere graads polynomen. Bijvoorbeeld: we benaderen f(x) tussen $(x_{j-1} + x_j)/2$ en $(x_j + x_{j+1})/2$ door een derdegraads interpolatiepolynoom met steunpunten x_{j-1}, x_j en x_{j+1}. Een bezwaar van deze stuksgewijze benadering is dat de benaderingen in de overgangspunten $(x_j + x_{j+1})/2$ niet continu aansluiten.

Een goed continue (zelfs tweemaal continu differentieerbare) benadering krijgen we als volgt.
Zij s(x) voor $x_0 \leq x \leq x_n$ gedefinieerd door

(i) s, s', s'' zijn continu in $[x_0,x_n]$.

(ii) s''' bestaat in ieder der open intervallen (x_j,x_{j+1}) en is daar constant.

(iii) $s''(x_0) = 0$, $s''(x_n) = 0$.

(iv) $s(x_j) = f_j$, $j = 0,\ldots,n$.

Functies die de eigenschappen (i), (ii) en (iii) hebben heten spline functies (van de orde 3) op het interval $[x_0,x_n]$. In ieder der intervallen (x_j,x_{j+1}) is s een derdegraads polynoom. In de punten x_j sluiten deze polynomen aan met continue nulde, eerste en tweede afgeleiden, maar in het algemeen met discontinue derde afgeleiden.
In 6.3.1 bewijzen we dat er steeds precies één spline is die aan de interpolatie eisen (iv) voldoet.
Men kan ook bewijzen dat de gevonden interpolerende spline functie onder alle interpolerende tweemaal continu differentieerbare functies z diegene is, waarvoor

$$\int_{x_0}^{x_n} [z''(x)]^2 dx \qquad (1)$$

minimaal is.

Het blijkt dat de spline interpolatie zeer aantrekkelijke eigenschappen heeft, ook voor numerieke integratie en differentiatie. Natuurlijk zijn ook spline functies van andere orden bruikbaar.

De naam spline functie heeft de volgende achtergrond. Buigt men een dunne elastische staaf zo dat hij gaat door de punten $(x_0,f_0),\ldots,(x_n,f_n)$, dan wordt de positie van de elastische lijn van de staaf juist gegeven door de interpolerende spline functie s(x) (althans als alle hellingen klein zijn, zodat de gelineariseerde differentiaalvergelijking $z''''(x) = 0$ voor de elastische lijn gebruikt mag worden). De randvoorwaarden (iii) corresponderen met het feit dat de staaf buiten het interval (x_0,x_n) recht blijft. Het minimaal zijn van de integraal (1) correspondeert met het principe van minimale potentiële energie. De gunstige interpolatie eigenschappen van dunne houten latten (strooklatten of splines) werd in de scheepsbouw gebruikt voor het tekenen van vloeiende lijnen door een aantal gegeven punten.

6.3.1. Bepaling van een interpolerende spline functie

Stel $s'(x_j) = A_j$, $s''(x_j) = B_j$ $(j = 0,1,\ldots,n)$. Dan geldt (waarom) voor $x_j \le x \le x_{j+1}$

$$s''(x) = B_j \frac{x_{j+1} - x}{h_{j+\frac{1}{2}}} + B_{j+1} \frac{x - x_j}{h_{j+\frac{1}{2}}} ,$$

met $h_{j+\frac{1}{2}} = x_{j+1} - x_j$.

Hieruit volgt voor $x_j \le x \le x_{j+1}$ (ga na)

$$s(x) = \frac{x_{j+1} - x}{h_{j+\frac{1}{2}}} [f_j - \frac{1}{6} B_j (h_{j+\frac{1}{2}}^2 - (x_{j+1} - x)^2)]$$

$$+ \frac{x - x_j}{h_{j+\frac{1}{2}}} [f_{j+1} - \frac{1}{6} B_{j+1} (h_{j+\frac{1}{2}}^2 - (x - x_j)^2)] .$$

Hieruit volgen de relaties

$$A_j = \frac{f_{j+1} - f_j}{h_{j+\frac{1}{2}}} - \frac{1}{6} h_{j+\frac{1}{2}} (2B_j + B_{j+1}) ,$$

$$A_{j+1} = \frac{f_{j+1} - f_j}{h_{j+\frac{1}{2}}} + \frac{1}{6} h_{j+\frac{1}{2}} (B_j + 2B_{j+1}) .$$

Vervangen we in de tweede relatie j+1 door j dan vinden we dat voor $1 \le j \le n-1$ moet gelden

$$\frac{1}{6} h_{j-\frac{1}{2}} B_{j-1} + \frac{1}{3} (h_{j-\frac{1}{2}} + h_{j+\frac{1}{2}}) B_j + \frac{1}{6} h_{j+\frac{1}{2}} B_{j+1} = \frac{f_{j+1} - f_j}{h_{j+\frac{1}{2}}} - \frac{f_j - f_{j-1}}{h_{j-\frac{1}{2}}} . \quad (1)$$

Daar $B_0 = 0$, $B_n = 0$, zijn dit n-1 lineaire vergelijkingen voor de n-1 onbekenden $B_1, \ldots, B_{n-1}$.
De matrix van dit stelsel is tridiagonaal en diagonaal-overwegend (vgl. pag.51). Het stelsel heeft dus altijd een oplossing en deze is eenvoudig te berekenen (vgl. 4.4.2).

Opmerking. In plaats van de randvoorwaarden $s''(x_0) = s''(x_n) = 0$ zijn ook andere randvoorwaarden bruikbaar, bv.

$$s'(x_0) = f'(x_0) , \qquad s'(x_n) = f'(x_n)$$

of

$$s''(x_0) = f''(x_0) , \qquad s''(x_n) = f''(x_n)$$

of

$$\frac{s''(x_2) - s''(x_1)}{h_{3/2}} = \frac{s''(x_1) - s''(x_0)}{h_{\frac{1}{2}}} ,$$

$$\frac{s''(x_n) - s''(x_{n-1})}{h_{n-\frac{1}{2}}} = \frac{s''(x_{n-1}) - s''(x_{n-2})}{h_{n-3/2}} .$$

Men verkrijgt dan soortgelijke stelsels als (1).
Bij deze paren randvoorwaarden geldt de volgende uitspraak over de afbreekfout:
Zij $hmin = \min(h_{j+\frac{1}{2}})$, $hmax = \max(h_{j+\frac{1}{2}})$. Zij $\gamma > 0$. Als f(x) viermaal continu differentieerbaar is in [a,b], dan zijn er slechts van f en van γ afhankelijke constanten C_0, C_1, C_2, C_3, zodanig dat voor alle keuzen van de steunpunten waarvoor $hmin/hmax \ge \gamma$ is, in het interval $[x_0, x_n]$ geldt

$$|s^{(j)}(x) - f^{(j)}(x)| \le C_j h^{4-j} , \qquad j = 0,1,2,3 ,$$

(met $s^{(j)}$ is de j-de afgeleide bedoeld). We krijgen op deze manier dus een fraaie simultane approximatie voor functie en afgeleiden.

6.4. Approximatie

Men spreekt van approximatie indien voor een in principe geheel bekende, maar moeilijk uitrekenbare functie f een benadering gegeven wordt met behulp van een eenvoudiger uitrekenbare functie g.
We beperken ons tot functies f(x) die in het interval $-1 \leq x \leq 1$ geapproximeerd moeten worden.

We meten de kwaliteit van de approximatie met behulp van een norm waarmee de "afstand" van de functies f en g gemeten wordt. Bijvoorbeeld:

a) $$\|f - g\| = \max_{-1 \leq x \leq 1} |f(x) - g(x)| \; ; \qquad (1)$$

dit is de zg. uniforme of Chebyshev-norm.

b) $$\|f - g\| = \left(\int_{-1}^{1} (f(x) - g(x))^2 dx \right)^{\frac{1}{2}} \; ; \qquad (2)$$

dit is de zg. kwadratische of L^2-norm.

c) $$\|f - g\| = \max_{-1 \leq x \leq 1} r(x)|f(x) - g(x)| \; , \qquad (3)$$

met r(x) een gegeven positieve functie; dit is de zg. Chebyshev-norm met gewichtsfunctie.

Meestal approximeren we met een functie g(x) uit een klasse van functies die van een aantal parameters afhangen. Bijvoorbeeld:

a) $$g(x) = a_0 + a_1 x + \ldots + a_{N-1} x^{N-1} \; ; \qquad (4)$$

polynomen met al dan niet gegeven maximale graad.

b) $$g(x) = \frac{a_0 + a_1 x + \ldots + a_{k-1} x^{k-1}}{1 + b_1 x + \ldots + b_{N-k} x^{N-k}} \; , \qquad (5)$$

gebroken rationale functies met al dan niet gegeven maximale teller- en noemer-graad.

c) $$g(x) = \tfrac{1}{2} a_0 + \sum_{1}^{N-1} (a_k \cos(\pi k x) + b_k \sin(\pi k x)) \; . \qquad (6)$$

trigonometrische polynomen met al dan niet gegeven maximale graad.

d) $$g(x) = a_1 e^{\alpha_1 x} + \ldots + a_N e^{\alpha_N x} ;$$

exponentiële approximatie, N en de α's kunnen al dan niet gegeven zijn.

We kunnen nu vragen naar die functie uit een gekozen klasse die f(x) in een gekozen norm het beste benadert. Als het aantal parameters eindig is (dus in de gegeven voorbeelden: als N eindig is) dan is er meestal precies één functie g(x) die de beste benadering is.

Voorbeelden

a) Bij approximatie in de kwadratische norm met trigonometrische polynomen met graad < N krijgen we de beste benadering indien

$$a_k = \int_{-1}^{1} f(x)\cos(\pi kx)dx \qquad k = 0,1,\ldots,N-1$$

$$b_k = \int_{-1}^{1} f(x)\sin(\pi kx)dx \qquad k = 1,\ldots,N-1 .$$

De gevonden a_k en b_k hangen hier niet van de gekozen N af.

b) Bij approximatie in de kwadratische norm met polynomen van graad < N krijgen we de beste benadering als de coëfficiënten $a_0,\ldots,a_{N-1}$ voldoen aan het stelsel

$$\sum_{j=1}^{N} A_{kj} a_{j-1} = \int_{-1}^{1} x^{k-1} f(x)dx , \qquad k = 1,\ldots,N , \qquad (7)$$

waarin

$$A_{kj} = \begin{cases} 0 & \text{als } k + j \text{ oneven} \\ \dfrac{2}{k + j - 1} & \text{als } k + j \text{ even.} \end{cases}$$

Hier hangen de coëfficiënten $a_0,\ldots,a_{N-1}$ wel af van de gekozen N. Het stelsel (7) is voor grotere waarden van N vrij slecht geconditioneerd. We kunnen beter werken met polynomen van de graad N-1 in de vorm

$$g(x) = \sum_{0}^{N-1} b_j P_j(x) ,$$

waarin $P_j(x)$ het j-de Legendre polynoom is, gedefinieerd door

$$P_0(x) = 1 \ , \quad P_1(x) = x \ ,$$

$$P_{j+1}(x) = \frac{2j+1}{j+1} xP_j(x) - \frac{j}{j+1} P_{j-1}(x) \ .$$

In plaats van (7) krijgen we dan (omdat, net als in voorbeeld a), de functies, waaruit we g combineren, onderling orthogonaal zijn):

$$b_k = (k + \tfrac{1}{2}) \int_{-1}^{1} P_k(x) f(x) dx \ .$$

6.4.1. Chebyshev approximatie met polynomen

In veel gevallen is de Chebyshev-norm de voor approximatie meest geschikte norm. De beste approximatie minimaliseert dan de maximale afwijking $|f(x) - g(x)|$ in het interval.
We beschouwen nu approximatie met polynomen met graad < N:

$$g(x) = a_0 + a_1 x + \ldots + a_{N-1} x^{N-1} \ .$$

Liever schrijven we echter g(x) in de vorm

$$g(x) = b_0 T_0(x) + b_1 T_1(x) + \ldots + b_{N-1} T_{N-1}(x) \ , \qquad (1)$$

waarin $T_k(x)$ het k-de Chebyshev polynoom is, gedefinieerd door

$$T_0(x) = 1 \ , \quad T_1(x) = x \ , \qquad (2)$$

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x) \ . \qquad (3)$$

6.4.1.1. We noemen een paar eigenschappen van de Chebyshev polynomen die van belang zijn.

a) $T_k(x)$ is een polynoom met graad k en kopterm $2^{k-1}x^k$ (als $k \geq 1$); dit volgt direct uit (2) en (3).

b) $$T_k(\cos\theta) = \cos(k\theta) \ , \qquad 0 \leq \theta \leq \pi \ . \qquad (4)$$

Uit (2) volgt dat dit geldt voor k = 0 en k = 1. En uit de overeenkomst tussen (3) en de goniometrische formule

$$\cos(k+1)\theta + \cos(k-1)\theta = 2\cos\theta \cos k\theta$$

volgt met inductie dat (4) voor alle k geldt.

We merken op dat het interval $0 \le \theta \le \pi$ door $x = \cos\theta$ 1-1 afgebeeld wordt op $-1 \le x \le 1$.

c) $T_k(x)$ heeft uitsluitend enkelvoudige nulpunten, gelegen in de punten

$$\hat{x}_j := \cos\frac{(2j-1)\pi}{2k}, \qquad j = 1,\ldots,k .$$

Dit zijn nl. k verschillende punten waar volgens (4) $T_k(x) = 0$; als k-de graads polynoom kan T_k geen andere nulpunten meer hebben.

d) $$\|T_k\| = \max_{-1\le x\le 1} |T_k(x)| = 1 . \qquad (5)$$

en de extrema van $T_k(x)$ worden met wisselend teken aangenomen in de k+1 punten $x_j = \cos(j\pi/k)$, $j = 0,\ldots,k$; hier is $T_k(x_j) = (-1)^j$. Ook deze eigenschap volgt direct uit (4). We merken op dat

$$1 = x_0 > x_1 > \ldots > x_k = -1 .$$

Uit de laatste eigenschap volgt

Stelling 1.

Voor ieder polynoom $p(x)$ met graad $k \ge 1$ en kopterm x^k geldt

$$\|p\| = \max_{-1\le x\le 1} |p(x)| \ge 2^{1-k} .$$

Immers, stel $\|p\| < 2^{1-k}$. Zij $q(x) = p(x) - 2^{1-k}T_k(x)$. Dan is in $x_j = \cos(j\pi/k)$

$$(-1)^j q(x_j) = (-1)^j p(x_j) - 2^{1-k} < 0 ,$$

dus $q(x)$ heeft in de punten $x_0, x_1, \ldots, x_k$ afwisselende tekens, moet dus minstens k tussenliggende nulpunten hebben; dit kan echter niet, want $q(x)$ heeft graad $\le k-1$.

Uit deze stelling volgt al het volgende approximatieresultaat.
Zij $f(x)$ een polynoom met graad N:

$$f(x) = a_0 + a_1x + \ldots + a_Nx^N . \qquad (5)$$

Dan is het polynoom $p^*(x)$ van graad $\le N-1$ dat f het beste benadert in de Chebyshev zin (d.w.z. $\|f - p\| \ge \|f - p^*\|$ voor iedere p met graad $\le N-1$)

$$p^*(x) = a_0 + \ldots + a_{N-1}x^{N-1} + a_N(x^N - 2^{1-N}T_N(x)) .$$

Immers, deze p^* heeft graad $\leq N-1$ en

$$\|f - p^*\| = \|2^{1-N}a_N T_N\| = 2^{1-N}|a_N| ,$$

terwijl voor iedere p met graad $\leq N-1$ $f - p$ een polynoom met graad N en kopterm $a_N x^N$ is, zodat volgens stelling 1

$$\|f - p\| \geq 2^{1-N}|a_N| .$$

Opmerking. Als het polynoom f in plaats van door (5) gegeven is door

$$f(x) = b_0 T_0(x) + b_1 T_1(x) + \ldots + b_N T_N(x) ,$$

dan is de beste benadering p^* met graad $\leq N-1$

$$p^*(x) = b_0 T_0(x) + \ldots + b_{N-1} T_{N-1}(x) .$$

6.4.1.2. We vermelden nu enige algemene stellingen over Chebyshev approximatie. Op dezelfde manier als stelling 1 bewijst men

Stelling 2

Zij f(x) gegeven. Zij q(x) een polynoom met graad $\leq N-1$. Zij $x_0,\ldots,x_N$ met

$$1 \geq x_0 > x_1 > \ldots > x_N \geq -1 \qquad (6)$$

en het getal σ zo, dat voor $0 \leq j \leq N$

$$\text{sign}(f(x_j) - q(x_j)) = (-1)^j\sigma . \qquad (7)$$

Dan geldt voor ieder polynoom p met graad $\leq N-1$ dat

$$\|f - p\| \geq \min_{0\leq j\leq N} |f(x_j) - q(x_j)| . \qquad (8)$$

Hieruit volgt onmiddellijk

Stelling 3

Zij f(x) gegeven. Zij $p^*(x)$ een polynoom met graad $\leq N-1$. Zij $x_0,\ldots,x_N$ zo dat (6) geldt en het getal σ zo, dat $|\sigma| = 1$ en dat

$$f(x_j) - p^*(x_j) = (-1)^j\sigma\|f - p^*\| . \qquad (9)$$

Dan is p^* een beste benadering voor f met graad $\leq N-1$.

Immers, nemen we in stelling 2 voor q het polynoom p^*, dan volgt uit (9) dat het rechterlid in (8) gelijk is aan $\|f - p^*\|$.

Verder kan men bewijzen dat de omkering van stelling 3 ook geldt:

Stelling 4

Als p^* een beste benadering voor f met graad $\leq N-1$ is, dan zijn er $x_0,\ldots,x_N$ die aan (6) voldoen en een σ met $|\sigma| = 1$ zo dat (9) geldt.

Uit stelling 4 kan men tenslotte afleiden

Stelling 5

De beste benadering p^* voor f met graad $\leq N-1$ is eenduidig bepaald.

De stellingen 3 en 4 karakteriseren het eenduidige beste polynoom p^*. Stelling 2 is van groot belang omdat voor iedere q(x) met graad $\leq N-1$ waarvoor er $x_0,\ldots,x_N$ en σ bestaan zo, dat (6) en (7) gelden, geldt

$$\min_{0\leq j\leq n} |f(x_j) - q(x_j)| \leq \|f - p^*\| \leq \|f - q\| .$$

We vinden dus een onder- en een bovengrens voor $\|f - p^*\|$ en kunnen op grond hiervan beslissen of q(x) in voldoende mate bijna-optimaal is.

6.4.1.3. We beschrijven nu twee methoden om bij een gegeven functie f een polynoom p met graad $\leq N-1$ te vinden dat bijna-optimaal is.

Zij $x_0,\ldots,x_N$ de extrema van $T_N(x)$, dus

$$x_j = \cos(j\pi/N) . \qquad (1)$$

Bepaal het bij f behorende interpolatiepolynoom q(x) van graad $\leq N$ zo, dat $q(x_j) = f(x_j)$. Schrijf q(x) in de vorm

$$q(x) = \sum_{k=0}^{N} b_k T_k(x) .$$

Neem nu als approximatiepolynoom met graad $\leq N-1$

$$p(x) = \sum_{k=0}^{N-1} b_k T_k(x) . \qquad (2)$$

In de punten x_j geldt dan

$$f(x_j) - p(x_j) = b_N T_N(x_j) = (-1)^j b_N .$$

Met stelling 2 volgt hieruit dat dan voor het optimale polynoom p^* met graad $\leq N-1$ geldt

$$|b_N| \leq \|f - p^*\| \leq \|f - p\| .$$

p is dan bijna-optimaal indien $\|f - p\|$ slechts weinig groter is dan b_N. In veel gevallen blijkt dit het geval te zijn.

Het bepalen van de coëfficiënten b_k is eenvoudig daar

$$\sum_{j=0}^{N''} T_k(x_j)T_\ell(x_j) = \sum_{j=0}^{N''} \cos \frac{jk\pi}{N} \cos \frac{j\ell\pi}{N} =$$

$$= \begin{cases} 0 & \text{als } 0 \leq k < \ell \leq N \\ \frac{1}{2}N & \text{als } 0 < k = \ell < N \\ N & \text{als } k = \ell = 0 \text{ of } k = \ell = N . \end{cases}$$

(Hierin betekent $\sum_{j=0}^{N''}$ dat van de nulde en van de N-de term slechts de helft genomen moet worden.)
Hieruit volgt dat

$$b_k = \frac{2}{N} \sum_{j=0}^{N''} T_k(x_j)f(x_j) \qquad 1 \leq j \leq N-1$$

$$b_0 = \frac{1}{N} \sum_{j=0}^{N''} f(x_j)$$

$$b_N = \frac{1}{N} \sum_{j=0}^{N''} (-1)^j f(x_j) .$$

De tweede methode is als volgt.
Zij voor f(x) in het interval $|x| \leq 1$ een Taylorreeks met restterm bekend:

$$f(x) = \sum_{k=0}^{N} a_k x^k + R_{N+1}(x) .$$

Schrijf het Taylor-polynoom $\sum_{k=0}^{N} a_j x^j$ als lineaire combinatie van Chebyshev-polynomen:

$$f(x) = \sum_{k=0}^{N} b_k T_k(x) + R_{N+1}(x) .$$

Neem nu als approximatiepolynoom met graad $\leq N-1$

$$p(x) = \sum_{k=0}^{N-1} b_k T_k(x) .$$

Stel dat we een getal r_{N+1} kennen waarvoor geldt

$$|R_{N+1}(x)| \leq r_{N+1} \quad \text{voor } |x| \leq 1 \quad \text{en} \quad r_{N+1} < |b_N| .$$

Dan geldt zeker

$$\|f - p\| \leq |b_N| + r_{N+1} .$$

En in de N+1 extrema x_j van $T_N(x)$ is

$$f(x_j) - p(x_j) = (-1)^j [b_N + (-1)^j R_N(x_j)] .$$

Dus dan heeft $f(x) - p(x)$ in de punten $x_0, \ldots, x_N$ wisselende tekens en

$$|f(x_j) - p(x_j)| \geq |b_N| - r_{N+1} .$$

Volgens stelling 2 geldt dan voor het optimale polynoom p^* met graad $\leq N-1$ dat

$$|b_N| - r_{N+1} \leq \|f - p^*\| \leq \|f - p\| \leq |b_N| + r_{N+1} .$$

Hieruit volgt: als $r_{N+1} << |b_N|$, dan is p bijna-optimaal. In veel "mooie" gevallen is hieraan voldaan.

Voorbeeld. Zij $f(x) = \cos(0.4\,x)$. Dan is

$$f(x) = 1 - 0.08\,x^2 + \frac{0.0032}{3}\,x^4 + R_6(x) \ ,$$

met

$$|R_6(x)| \leq \frac{(0.4)^6}{6!} < 0.000006 \ .$$

Daar

$$T_0(x) = 1 \ , \quad T_2(x) = 2x^2 - 1 \ , \quad T_4(x) = 8x^4 - 8x^2 + 1 \ ,$$

is ook

$$f(x) = 0.9604\,T_0(x) - \frac{0.1134}{3}\,T_2(x) + \frac{0.0004}{3}\,T_4(x) + R_6(x) \ .$$

Neem nu

$$p(x) = 0.9604\,T_0(x) - \frac{0.1184}{3}\,T_2(x) = 0.99987 - 0.07893\,x^2 \ .$$

Dan is

$$\|f - p\| \leq \frac{0.0004}{3} + 0.000006 < 0.000140 \ .$$

En uit stelling 2 volgt

$$\|f - p^*\| \geq \frac{0.0004}{3} - 0.000006 > 0.000127 \ .$$

p is dus bijna-optimaal.
Merk op dat het tweedegraads Taylor-polynoom

$$1 - 0.08\,x^2$$

een maximale afwijking heeft van ca 0.0011, dat is dus ca 8 keer zo veel als die van p(x) !

6.5. Aanpassing

Men spreekt van aanpassing indien van een onbekende functie $f(x)$ een aantal waarden $f(x_1),\ldots,f(x_N)$ (al dan niet onnauwkeurig) bekend zijn en men in een bekende, van een aantal parameters afhankelijke functie $g(x;a_1,\ldots,a_n)$ de parameters zo bepaalt dat in de punten $x_1,\ldots,x_N$ f en g zo goed mogelijk overeenstemmen.

Als maat voor de afwijking kiest men weer een norm, nu bv.

$$\|f - g\| = \max_{1\le j\le N} |f(x_j) - g(x_j)| \tag{1}$$

of

$$\|f - g\| = \left(\sum_{j=1}^{N} (f(x_j) - g(x_j))^2 \right)^{\frac{1}{2}} \tag{2}$$

of

$$\|f - g\| = \max_{1\le j\le N} r_j |f(x_j) - g(x_j)| \ . \tag{3}$$

Men spreekt van Chebyshev aanpassing, kleinste kwadraten aanpassing, etc.

Indien bekend is, dat de gemeten waarden $f(x_j)$ behoudens meetfouten gelijk zijn aan de corresponderende waarden $g(x_j;a_1,\ldots,a_n)$ voor een geschikt stel a's, en als deze meetfouten een "stochastisch" karakter hebben, dan is het op statistische gronden verstandig om de kleinste kwadraten norm te gebruiken (in andere gevallen meestal niet).

Als $N \le n$ dan is er als regel een stel parameters zodanig dat $\|f - g\| = 0$. Om de invloed van meetfouten of andere storingen te verminderen kiezen we echter meestal N flink wat groter dan n. Het is duidelijk dat aanpassing dan veel op approximatie gaat lijken. Op technieken om beste aanpassingen te vinden gaan we niet in.