TECHNISCHE HOGESCHOOL EINDHOVEN
Afdeling Algemene Wetenschappen
Onderafdeling der Wiskunde

NUMERIEKE METHODEN I en II

Prof. Dr. G.W. Veltkamp

met medewerking van

Drs. A.J. Geurts

1977/1978

Technische Hogeschool Eindhoven

Onderafdeling der Wiskunde

Numerieke Methoden I en II

Dictaatnr. 2.211 Prijs f 10,-

TECHNISCHE HOGESCHOOL EINDHOVEN

Onderafdeling der Wiskunde

Numerieke Methoden I en II

1977/1978

Prof.dr. G.W. Veltkamp

Drs. A.J. Geurts

Inhoudsopgave

0. Inleiding ([2], ch. 1,2; [15], ch. 1) *)

0.0. Doelstelling

Het doel van de colleges Numerieke Methoden I en II is tweeledig:

- kennismaking met een aantal voor de praktijk belangrijke technieken uit de numerieke wiskunde, zoals algorithmiseren, 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;
- kennismaking met enkele methoden met behulp waarvan het mogelijk is een schatting van de betrouwbaarheid van de berekende resultaten te geven.

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) Het is ook mogelijk dat de oplossing van het probleem wel expliciet in een formule te geven is, maar dat deze formule minder geschikt is voor het verkrijgen van numerieke waarden dan het rechtstreeks numeriek oplossen van het oorspronkelijke probleem.

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

Voorbeelden

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

$$\frac{dy}{dx} = e^{x^2}\, y + e^x$$

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

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

---

*) Een getal tussen vierkante haken verwijst naar een boek uit de literatuurlijst, die aan het einde van dit hoofdstuk is opgenomen.

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 oplossing van het stelsel lineaire vergelijkingen $\underline{A}\underline{x} = \underline{b}$ is met behulp van de regel van Cramer expliciet te formuleren als het berekenen van een aantal determinanten. We zullen later zien dat deze formulering van de oplossing reeds voor betrekkelijk kleine waarden van n (de dimensie van A en $\underline{b}$) voor de numerieke berekening absoluut ongeschikt is.

Met deze voorbeelden willen we niet propageren dat men bij een gegeven probleem altijd meteen moet grijpen naar een numerieke methode 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.

0.1. Fouten en foutenbronnen ([2], p. 22-24)

Het oplossen van een praktisch probleem (dit kan zijn een fysisch, maar ook bijvoorbeeld een economisch probleem) gebeurt in verschillende stappen.

a) Eerst wordt van het practische probleem een wiskundig model gemaakt. De gevolgen van de afwijking tussen de werkelijkheid en dit model noemen we modelfouten. Deze zijn voor de numericus in zoverre van belang dat bij een grof model een erg nauwkeurige berekening weinig zin heeft.

b) Van het wiskundige model wordt zonodig een numeriek model gemaakt, bijvoorbeeld door het probleem te discretiseren of te lineariseren. De hierbij op-

tredende fouten worden in het algemeen afbreekfouten genoemd.

Voorbeeld. De regel van Simpson

$$\int_{-h}^{h} f(x)\,dx = \frac{h}{3}[f(-h) + 4f(0) + f(h)] + R.$$

Hierin is R de afbreekfout.

c) Voor de oplossing van het probleem wordt een algorithme opgesteld. Als deze algorithme in principe niet eindig is, bijvoorbeeld omdat er een iteratieproces in voorkomt, dan wordt een beëindigingscriterium gebruikt. Het resultaat van de algorithme is dan een benadering van de oplossing van het numerieke model. De fout in deze benadering (die in wezen ook een afbreekfout is) zullen we de convergentiefout noemen.

d) Van de algorithme wordt een programma gemaakt. Dit programma wordt uitgevoerd op een rekenautomaat. Deze heeft een arithmetiek met een beperkt aantal cijfers (binair, octaal, decimaal). Daardoor worden zowel bij de invoering van de gegevens als bij het rekenen afrondfouten gemaakt.

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

berekende waarde = exacte waarde + fout.

Bij afbreekfouten is meer gebruikelijk

exacte waarde = waarde van benadering + fout.

In beide gevallen spreekt men ook wel van een absolute fout. Dit in tegenstelling tot de relatieve fout die gedefinieerd is door

relatieve fout = absolute fout / exacte waarde.

Opmerking. Het begrip absolute fout wordt nogal eens verward met het begrip bovengrens (schatting) voor de absolute waarde van de fout.
Zij bijvoorbeeld a een reëel getal en zij $\bar{a}$ een benadering van a met een absolute nauwkeurigheid van drie decimalen. Daarmee bedoelen we dat

$$\bar{a} = a + \delta a$$

met

$$|\delta a| \leq 0.5_{10} - 3.$$

Dan is δa de absolute fout en het getal $0.5_{10} - 3$ is een bovengrens voor de absolute waarde van δa.

0.2. Floating point representatie ([2], p. 42-49; [13], p. 1-8)

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 twee tekens (van m en van e) en t+q decimale cijfers (die de waarde 0,1,...,9 kunnen hebben). Maar daaruit volgt ook dat er slechts eindig veel machinegetallen bestaan. Zo is het grootste machinegetal

$$a\ max = 0.\underbrace{99\ \ldots\ 9}_{t\ cijfers} \times 10^{10^q-1} = (1 - 10^{-t}) \times 10^{10^q-1}.$$

En het kleinste positieve machinegetal is

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

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 machinegetal. Noem de uitkomst x. Als x in de zogenaamde range ligt, d.w.z.

$$a\,min \leq |x| \leq a\,max$$

en p zo is dat

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

dan is de bij x behorende afronding, genoteerd als fl(x), van de vorm

$$fl(x) = m \times 10^{p}$$

met

$$0.1 \leq |m| \leq 1 \quad *) .$$

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

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

En in het algemeen geldt (ga na)

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

Hieruit volgt voor de maximale relatieve fout

$$\frac{|x - fl(x)|}{|x|} \leq \frac{\tfrac{1}{2} \cdot 10^{p-t}}{10^{p-1}} = \tfrac{1}{2} \cdot 10^{1-t} .$$

Het getal $\eta := \tfrac{1}{2}.10^{1-t}$ wordt de <u>machine-nauwkeurigheid</u> (macheps) genoemd.

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 $0 < |x| < a\,min$.

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) <u>bits</u>.
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

---

*) Als m = 1, dan wordt fl(x) genoteerd als $0.\underbrace{10\ldots0}_{t\ \text{cijfers}} \times 10^{p+1}$.

(dus $\beta^t \times m$ is geheel), terwijl

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

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

Het effect van een arithmetische operatie op een rekenmachine kan men dus als volgt beschrijven.
Bij uitvoering van een arithmetische operatie $\oplus$ (waarbij $\oplus$ kan zijn: $+$, $-$, $\times$ of $/$) op twee machinegetallen a en b levert de machine een machinegetal $\bar{c} := fl(a \oplus b)$ af waarvoor geldt *)

$$\frac{|\bar{c} - a \oplus b|}{|a \oplus b|} \leq \eta . \qquad (4)$$

Hierin is $\eta := \frac{1}{2}\beta^{1-t}$ de machine-nauwkeurigheid.
Uit (4) volgt dat

$$\bar{c} = (a \oplus b)(1 + \varepsilon) \quad \text{met } |\varepsilon| \leq \eta . \qquad (5)$$

Als het zo uitkomt, dan kunnen we in plaats van (5) ook schrijven

$$a \oplus b = \bar{c}(1 + \varepsilon') \quad \text{met } |\varepsilon'| \leq \eta . \qquad (6)$$

Opmerking. Voor de B7700 geldt $\beta = 8$, $t = 13$, terwijl van een machinegetal $a \neq 0$ de mantisse m in de representatie (1) een geheel getal is. In plaats van (2) geldt voor m de ongelijkheid

$$1 \leq |m| \leq 8^{13} - 1 .$$

Verder is $q = 2$, zodat voor de exponent e geldt

$$|e| \leq 63 .$$

Ga na dat voor de B7700 geldt $a_{min} \doteq 1.27_{10}{-57}$, $a_{max} \doteq 4.31_{10}68$, en de machine-nauwkeurigheid $\eta = 2^{-37} \doteq 1.1_{10}{-11}$.

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.

---

*) De formules (4), (5) en (6) gelden onder de voorwaarde dat $a \oplus b$ in de range ligt. In het vervolg wordt bij de foutenanalyse steeds verondersteld, dat berekende resultaten in de range liggen.

0.3. Conditie, numerieke stabiliteit, foutenvoortplanting ([2], p. 51-59; [13], 8-19)

Het is duidelijk dat veranderingen in de gegevens van een probleem in het algemeen de oplossing wijzigen. Het is ook duidelijk dat fouten die in een bepaald stadium van een berekening gemaakt worden in het algemeen aanleiding zullen geven tot fouten in de resultaten bij de verdere uitvoering van de berekening, zodat de berekende oplossing afwijkt van het exacte resultaat van de algorithme.
Essentiële begrippen bij de beoordeling van een algorithme voor een probleem zijn de conditie van het probleem en de numerieke stabiliteit van de algorithme.

Een probleem heet goed geconditioneerd als een kleine verandering in de gegevens een kleine verandering in het resultaat geeft; zo niet, dan heet het probleem slecht geconditioneerd.
Zij te berekenen

$$y = f(x) \tag{1}$$

bij een gegeven waarde van x.
Een kleine verandering δx in de gegeven x heeft een verandering δy in het resultaat y tot gevolg.
Hiervoor geldt

$$y + \delta y = f(x + \delta x) \ ,$$

waaruit volgt, als f differentieerbaar is,

$$\delta y = f'(x)\cdot\delta x + \ldots \ .$$

Voor de relatieve veranderingen geldt dan (als $x \neq 0$, $y \neq 0$)

$$\frac{|\delta y|}{|y|} \simeq \frac{|f'(x)|\cdot|x|}{|f(x)|} \cdot \frac{|\delta x|}{|x|} \ . \tag{2}$$

Het getal $c(x) := |f'(x)|\cdot|x|/|f(x)|$ noemen we het conditiegetal van het probleem.

Als een probleem wordt opgelost met behulp van een rekenmachine met machinenauwkeurigheid η, dan bevat de oplossing een fout, onafhankelijk van de algorithme, die ontstaat doordat de gegevens en het eindresultaat slechts in eindige nauwkeurigheid gerepresenteerd kunnen worden. Deze fout in de oplossing heet de onvermijdbare fout.

Omdat de relatieve fout in de representatie van de gegevens en het eindresultaat ten hoogste η bedraagt, volgt uit (2) voor de onvermijdbare fout, zeg $\Delta^{\circ}y$, van probleem (1) de volgende schatting:

$$\frac{|\Delta^{\circ}y|}{|y|} \leq (c(x) + 1)\eta \ . \tag{3}$$

Tijdens de uitvoering van een algorithme worden afrondfouten gemaakt. Dit heeft ook tot gevolg dat de oplossing een fout bevat. Deze fout zullen we de totale rekenfout noemen. Het is het verschil tussen de berekende en de exacte oplossing van het probleem, waarbij de gegevens exact verondersteld zijn.

Een algorithme heet numeriek stabiel als het totale effect van de afrondfouten niet essentieel groter is dan de onvermijdbare fout.

We beschouwen nogmaals het probleem (1). Veronderstel dat we bij een gegeven algorithme voor de totale rekenfout, zeg δy, de volgende schatting hebben afgeleid

$$\frac{|\delta y|}{|y|} \leq A(c(x) + 1)\eta \ . \tag{4}$$

Dan is deze algorithme numeriek stabiel als A maximaal van de orde van grootte van n is, waarbij n het aantal bewerkingen is voor de berekening van f(x).

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

$$y = x + \gamma x^2 \ . \tag{5}$$

De relevante waarden van x en y zijn in de buurt van 1, γ is van de orde van 0.01.
Beschouwt men (5) als vierkantsvergelijking in x, waarvan we de positieve wortel moeten hebben, dan volgt met de bekende formule

$$x = \frac{\sqrt{1 + 4\gamma y} - 1}{2\gamma} \ . \tag{6}$$

Deze formule is echter numeriek instabiel. Rekenen we bv. consequent in drie cijfers achter de komma, dan zal de gevonden waarde van $\sqrt{1 + 4\gamma y}$ een fout kunnen hebben van $\frac{1}{2} \times 10^{-3}$. Maar dat betekent in x een mogelijke fout van $(1/4\gamma) \times 10^{-3} \sim \frac{1}{4} \times 10^{-1}$, 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 $\sqrt{1 + 4\gamma y}$ uitrekenen met een fout van hoogstens $\gamma \times 10^{-3}$. Het feit dat de relatieve fout in x veel groter is dan die in $\sqrt{1 + 4\gamma y}$ wordt veroorzaakt door zg. cijferverlies: als men van $\sqrt{1 + 4\gamma y}$ 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 (6) met de zg. worteltruc: we kunnen schrijven

$$x = \frac{2y}{\sqrt{1 + 4\gamma y} + 1} \quad . \tag{7}$$

Ga na dat nu de eindnauwkeurigheid in x ongeveer net zo groot is als die in $\sqrt{1 + 4\gamma y}$, zodat formule (7) wel numeriek stabiel is.

Opmerking. Formule (7) is exact en numeriek stabiel, maar wel wat bewerkelijk. Eisen we niet te grote nauwkeurigheid, dan kunnen we reeksontwikkelen:

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

De hieruit volgende benaderingsformule

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

kunnen we ook verkrijgen door uit (5) als nulde benadering te halen

$$x_0 = y$$

en als eerste benadering

$$x_1 = y - \gamma x_0^2 \quad ,$$

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

Om te bewijzen dat een algorithme numeriek stabiel is, moeten we een bovengrens vinden voor het totale effect van de tijdens de uitvoering van de algorithme gemaakte afrondfouten. Dit kan op verschillende manieren worden gedaan.

Zij y de exacte waarde van de functie f bij gegeven x, dus

$$y = f(x) \ . \tag{1}$$

Zij $\bar{y}$ het resultaat als f(x) wordt berekend met mogelijke afrondfouten.

a) Bij een <u>voorwaartse foutenanalyse</u> schrijven we

$$\bar{y} = y + \delta y \tag{8a}$$

en we bepalen met behulp van foutenvoortplanting een schatting *) voor δy.

b) Bij een <u>achterwaartse foutenanalyse</u> beschouwen we de berekende $\bar{y}$ als de exacte waarde van f in een naburig punt $x + \delta x$; dat wil zeggen, we schrijven

$$\bar{y} = f(x + \delta x) \tag{8b}$$

en we bepalen een schatting voor de hypothetische storing δx.
Als δx klein is, dan is $\bar{y}$ dus de exacte oplossing van een naburig probleem.

c) Tenslotte bestaat er ook nog een <u>gemengde foutenanalyse</u>. Daarbij schrijven we (deze δx en δy zijn in het algemeen niet dezelfde als de overeenkomstige grootheden uit a) en b)),

$$\bar{y} = f(x + \delta x) + \delta y \ . \tag{9}$$

Zeker in dit geval zijn δx en δy niet eenduidig bepaald. We proberen daarom de fout zo gunstig mogelijk over x en y te verdelen, d.w.z. we proberen te bereiken dat de schattingen voor δx en δy beide klein zijn.

Uit (9) volgt met (2)

$$\left|\frac{\bar{y} - y}{y}\right| \leq c(x)\left|\frac{\delta x}{x}\right| + \left|\frac{\delta y}{y}\right| \ ,$$

zodat de algorithme voor het berekenen van (1) numeriek stabiel is als voor de grootheden uit (9) geldt dat $|\delta x|/|x|$ en $|\delta y|/|y|$ beide van de orde van grootte van afrondfouten in de algorithme (ca. aantal bewerkingen maal machine-nauwkeurigheid) zijn. Immers, in formule (4) kunnen we dan $A := \max(\left|\frac{\delta x}{x}\right|, \left|\frac{\delta y}{y}\right|)/\eta$ nemen.
Analoog bij voorwaartse of achterwaartse foutenanalyse (8a) of (8b).
Als de uitvoer een m-vector $\underline{y}$ is die afhangt van een η-vector $\underline{x}$ van invoergegevens dan kunnen we analoog handelen. Zie 1.4.1 en 5.2.5.

---

*) In de wiskunde (behalve in de statistiek) verstaat men onder een schatting (Eng. estimate) voor een grootheid een (eventueel grove) bovengrens voor de absolute waarde van die grootheid. Dat is dus iets anders dan een benadering (Eng. approximation) voor die grootheid.

In 0.2 hebben we nagegaan op welke wijze afrondfouten gegenereerd worden en wat hun orde van grootte is. Voor een foutenanalyse moeten we ook weten hoe geïntroduceerde fouten zich voortplanten tijdens een berekening.

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

A. Zij de te berekenen grootheid

$$c = a + b.$$

Zij $\overline{a}$ de ter beschikking staande benadering voor a, $\overline{b}$ die voor b. Dan is, als bij de optelling geen nieuwe fout gemaakt wordt,

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

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

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

dan is

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

Voor de relatieve fouten $\varepsilon a := \frac{\delta a}{a}$, $\varepsilon b := \frac{\delta b}{b}$ en $\varepsilon c := \frac{\delta c}{c}$ geldt

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

Dus:

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

b) De relatieve fout in $\overline{c}$ is een lineaire combinatie van de relatieve fouten in $\overline{a}$ en $\overline{b}$ 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.

$$|\varepsilon c| \leq \max(|\varepsilon a|, |\varepsilon b|).$$

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.

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

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 numeriek instabiele algorithme).

Ga na dat $\left|\frac{a}{a + b}\right|$, resp. $\left|\frac{b}{a + b}\right|$, het conditiegetal is bij variatie van a, resp. b.

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. Vermenigvuldiging:
Zij $c = a \times b$ en $\bar{c} = \bar{a} \times \bar{b}$.
Zij $\bar{a} = a(1 + \varepsilon a)$, $\bar{b} = b(1 + \varepsilon b)$, $\bar{c} = c(1 + \varepsilon c)$ dan zijn εa, εb, εc de relatieve fouten in $\bar{a}$, $\bar{b}$, $\bar{c}$, en er geldt

$$c(1 + \varepsilon c) = a(1 + \varepsilon a)\, b(1 + \varepsilon b) ,$$

$$1 + \varepsilon c = (1 + \varepsilon a)(1 + \varepsilon b) .$$

Dus, als we afzien van hogere orde termen, dan geldt

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

Dus 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 dat de relatieve fout in $\bar{c} = \bar{a}/\bar{b}$ het verschil is van de relatieve fouten in $\bar{a}$ en $\bar{b}$. Ga dit na.

0.4. Enkele voorbeelden van een foutenanalyse.

Bij een analyse van de afrondfouten maken we gebruik van formule (5) of (6) van paragraaf 0.2, en van de regels voor de foutenvoortplanting. Daarbij verwaarlozen we tweede en hogere orde termen. Verder veronderstellen we dat de in de voorbeelden optredende getallen machinegetallen zijn.
Als we bijvoorbeeld $c := a_1 \times b_1 + a_2 \times b_2$ willen bepalen, dan geldt voor de verkregen $\bar{c}$

$$\bar{c} = ((a_1 \times b_1)(1 + \varepsilon_1) + (a_2 \times b_2)(1 + \varepsilon_2))(1 + \varepsilon_3)$$

$$= (a_1 \times b_1)(1 + \varepsilon_4) + (a_2 \times b_2)(1 + \varepsilon_5)$$

met

$$|\varepsilon_i| \leq \eta \text{ voor } i = 1,2,3, \quad |\varepsilon_i| \leq 2\eta \text{ voor } i = 4,5$$

(waarbij we termen van de orde η^2 verwaarloosd hebben).

Als $a_1 \times b_1$ en $a_2 \times b_2$ hetzelfde teken hebben, dan geldt ook

$$\bar{c} = (a_1 \times b_1 + a_2 \times b_2)(1 + \varepsilon_6)$$

met $|\varepsilon_6| \leq 2\eta$.

Voorbeeld van een voorwaartse en een achterwaartse foutenanalyse

Zij gevraagd te berekenen de som

$$s = \sum_{i=1}^{n} a_i$$

op een rekenmachine met machine-nauwkeurigheid η.
Als we de berekening uitvoeren met de volgende algorithme

```
s := 0;
for i := 1 step 1 until n do s := s + a_i
```
(10)

dan geldt voor de berekende waarde $\bar{s}$

$$\bar{s} = a_1(1 + \varepsilon_2)\ldots(1 + \varepsilon_n) + a_2(1 + \varepsilon_2)\ldots(1 + \varepsilon_n) + a_3(1 + \varepsilon_3)\ldots(1 + \varepsilon_n) + \ldots + a_n(1 + \varepsilon_n)$$

met $|\varepsilon_i| \leq \eta$.
Hieruit volgt

$$\bar{s} = a_1(1 + \delta_1) + a_2(1 + \delta_2) + \ldots + a_n(1 + \delta_n) \qquad (11)$$

met, onder verwaarlozing van termen van de orde η^2 en hoger,

$$|\delta_1| \leq (n-1)\eta, \quad |\delta_i| \leq (n-i+1)\eta \ (i \geq 2) \ . \qquad (12)$$

Bij een achterwaartse foutenanalyse concluderen we uit (11) dat $\bar{s}$ de exacte som is van de getallen $\tilde{a}_i := a_i(1 + \delta_i)$. Dan volgt uit (12) dat de algorithme (10) numeriek stabiel is.

Bij een voorwaartse foutenanalyse concluderen we uit (11) dat de absolute fout in $\bar{s}$ gelijk is aan

$$\delta s = a_1\delta_1 + a_2\delta_2 + \ldots + a_n\delta_n \ .$$

Hieruit volgt voor de absolute fout in $\bar{s}$ de schatting

$$|\delta s| \leq (n-1)\eta \sum_{i=1}^{n} |a_i| \ .$$

Voor een som van allemaal positieve (negatieve) termen geldt dan

$$\frac{|\delta s|}{|s|} \leq (n-1)\eta \ ,$$

terwijl bij een som van zowel positieve als negatieve termen de relatieve fout essentieel groter kan zijn. Dit laatste is een gevolg van de slechte conditie, want de algorithme (10) is numeriek stabiel (zie pag. 0.13).

Voorbeeld van een gemengde foutenanalyse.

De vierkantsvergelijking

$$y^2 - 2py + q = 0 \tag{13}$$

heeft, als $p^2 - q \geq 0$, de reële oplossing

$$y_1 := p + \sqrt{p^2 - q} \ . \tag{14}$$

Als we y_1 met deze formule berekenen op een rekenmachine met machine-nauwkeurigheid η, dan krijgen we

$$\bar{y}_1 = (p + \sqrt{(p^2(1+\varepsilon_1) - q)(1+\varepsilon_2)}(1+\varepsilon_3))(1+\varepsilon_4) \tag{15}$$

met

$$|\varepsilon_i| \leq \eta, \ i = 1,2,3,4 \ .$$

We vervangen in deze formule p en q door geschikte naburige waarden $\tilde{p}$ en $\tilde{q}$. We kiezen daarvoor $\tilde{p}$ zodanig dat $\tilde{p}^2 = p^2(1+\varepsilon_1)(1+\varepsilon_2)$ en $\tilde{q} = q(1+\varepsilon_2)$.

Zij ε_5 zodanig dat $p = \tilde{p}(1+\varepsilon_5)$, dan geldt $(1+\varepsilon_5)^2(1+\varepsilon_1)(1+\varepsilon_2) = 1$ en hieruit volgt, bij verwaarlozing van tweede en hogere orde termen,

$$|\varepsilon_5| \leq \eta \ .$$

Door deze vervanging gaat (15) over in

$$\begin{aligned}\bar{y}_1 &= (\tilde{p}(1+\varepsilon_5) + \sqrt{\tilde{p}^2 - \tilde{q}}(1+\varepsilon_3))(1+\varepsilon_4) \\ &= \tilde{y}_1\left(1 + \frac{\varepsilon_5\tilde{p} + \varepsilon_3\sqrt{\tilde{p}^2 - \tilde{q}}}{\tilde{y}_1}\right)(1+\varepsilon_4) \ ,\end{aligned}$$

waarin $\tilde{y}_1 = \tilde{p} + \sqrt{\tilde{p}^2 - \tilde{q}}$ de exacte oplossing is van (13) waarin p en q zijn vervangen door $\tilde{p}$ en $\tilde{q}$.

We kunnen nu het volgende concluderen.

De berekende $\bar{y}_1$ is een benadering van $\tilde{y}_1$ met een relatieve fout

$$\varepsilon\tilde{y}_1 = \frac{\varepsilon_5\tilde{p} + \varepsilon_3\sqrt{\tilde{p}^2 - \tilde{q}}}{\tilde{y}_1} + \varepsilon_4 .$$

Als $\tilde{p} > 0$ (dus als $p > 0$), dan geldt

$$|\varepsilon\tilde{y}_1| \leq 2\eta ,$$

dus dan is (14) een stabiele formule.

Opmerking. Als $p < 0$, dan is $y_2 := p - \sqrt{p^2 - q}$ een stabiele formule. En als een der wortels op stabiele wijze berekend is, dan kan men de andere altijd stabiel berekenen uit de relatie $y_1 y_2 = q$.

Opgave. Bepaal de conditiegetallen

$$\left|\frac{p}{y_1}\frac{\partial y_1}{\partial p}\right| \text{ en } \left|\frac{q}{y_1}\frac{\partial y_1}{\partial q}\right| .$$

Ga na dat, als $p \geq 0$, de algorithme (14) voor y_1 goed geconditioneerd is als $q \leq 0$ en slecht geconditioneerd is als q dicht bij p^2 ligt.
Illustreer dit ook met een plaatje waarin de grafieken $z = y^2 + q$ en $z = 2py$ zijn getekend. Welk meetkundig beeld komt dan overeen met slechte conditie?

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 na welke formules in welke gevallen stabiel zijn.

Tenslotte merken we nog op dat de totale fout in de berekende oplossing van een probleem, behalve de onvermijdbare fout en de totale rekenfout, ook nog een fout bevat die veroorzaakt wordt door fouten in de gegevens (ingangsfouten). Dit betekent bijvoorbeeld voor probleem (1), met een ingangsfout δx, waarvoor geldt $\left|\frac{\delta x}{x}\right| \leq \varepsilon$, dat voor de totale fout, zeg Δy, de schatting

$$\left|\frac{\Delta y}{y}\right| \leq c(x).\varepsilon + (A + 1)(c(x) + 1)\eta$$

kan worden gegeven.

Opgave

Zij exp(x) een numeriek stabiele algorithme voor de berekening van e^x.
Dan zijn de algorithmen

$$\sinh(x) := (\exp(x) - \exp(-x))/2$$

en

$$\sinh(x) := (\exp(x) - 1/\exp(x))/2$$

niet numeriek stabiel in een omgeving van $x = 0$.
Als echter $\varphi(x)$ een numeriek stabiele algorithme is voor de berekening van $(e^x - 1)/x$ dan is

$$\sinh(x) := (x/2) * (\varphi(x) + \varphi(-x))$$

wel numeriek stabiel. Hoe zoudt U de algorithme φ maken?

Literatuur

Deze literatuurlijst bevat een aantal studieboeken, die passen bij de onderwerpen van het college. Een verwijzing bij een bepaald onderwerp naar een of meerdere boeken uit deze literatuurlijst betekent dat dezelfde materie daar eveneens wordt behandeld. Voor de bestudering van de collegestof zijn deze verwijzingen niet noodzakelijk. De tentamens zijn uitsluitend gebaseerd op de inhoud van de syllabus.

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

[2] Dahlquist, G. and A. Björck, Numerical methods.
Englewood Cliffs (New Jersey): Prentice-Hall, 1974.

[3] Davis, P.J. and P. Rabinowitz, Methods of numerical integration.
New York etc.: Academic Press, 1975.

[4] Fröberg, C.E., Introduction to numerical analysis.
Reading (Mass.): Addison Wesley, 1972.

[5] Gear, C.W., Numerical initial value problems in ordinary differential equations.
Englewood Cliffs (New Jersey): Prentice-Hall, 1971.

[6] Hamming, R.W., Introduction to applied numerical analysis.
New York etc.: McGraw-Hill, 1971.

[7] Lambert, J.D., Computational methods in ordinary differential equations.
London etc.: Wiley, 1973.

[8] Lapidus, L., Digital computation for chemical engineers.
New York etc.: McGraw-Hill, 1962.

[9] Lapidus, L. and J.H. Seinfeld, Numerical solution of ordinary differential equations.
London etc.: Academic Press, 1971.

[10] Moursund, D.G. and C.S. Duris, Elementary theory and application of numerical analysis.
New York etc.: McGraw-Hill, 1967.

[11] Noble, B., Numerical methods, I,II.
Edinburgh etc.: Oliver and Boyd, 1964.

[12] Phillips, G.M. and P.J. Taylor, Theory and applications of numerical analysis.
London etc.: Academic Press, 1973.

[13] Stoer, J., Einführung in die numerische Mathematik, I (2. neubearb. und erw. Aufl.).
Berlin etc.: Springer-Verlag, 1976. (Heidelberger Taschenbücher; Bd. 105).

[14] Stoer, J. und R. Bulirsch, Einführung in die numerische Mathematik, II.
Berlin etc.: Springer-Verlag, 1973. (Heidelberger Taschenbücher; Bd. 114).

[15] Numerical analysis: an introduction; ed. by J. Walsh.
London etc.: Academic Press, 1966.

[16] Young, D.M. and R.T. Gregory, A survey of numerical mathematics, I,II.
Reading (Mass.): Addison-Wesley, 1973.

1. Het oplossen van vergelijkingen ([2], ch. 6)

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

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

en van stelsels vergelijkingen

$$\begin{aligned} F_1(x_1,x_2,\ldots,x_k) &= 0 \\ F_2(x_1,x_2,\ldots,x_k) &= 0 \\ &\cdots\cdots\cdots \\ F_k(x_1,x_2,\ldots,x_k) &= 0 \ . \end{aligned}$$

Deze laatste kunnen we in vectornotatie schrijven als

$$\underline{F}(\underline{x}) = \underline{0} \ . \tag{2}$$

We zullen zien dat sommige methoden voor het oplossen van (2) kunnen worden opgevat als een generalisatie van een methode voor het oplossen van (1).

Voor het oplossen van (1) beginnen we met een "gezond-verstand methode". Het komt nogal eens voor dat we $F(x)$ kunnen schrijven als

$$F(x) = (x-\beta)g(x) + \varepsilon h(x) \ , \tag{3}$$

met $g(x) \neq 0$ in een omgeving van $x = \beta$, en $|\varepsilon|$ klein. Dat wil zeggen, $F(x)$ is op een kleine storing na een functie met een bekend enkelvoudig nulpunt, nl. β. We mogen daarom verwachten dat $F(x)$ een nulpunt, zeg α, heeft in de buurt van β.
Het ligt dan voor de hand de vergelijking $F(x) = 0$ te herschrijven als

$$x = \beta - \varepsilon \frac{h(x)}{g(x)} \ . \tag{4}$$

Omdat β een goede schatting is voor het nulpunt α en $\varepsilon h(x)/g(x)$ slechts weinig van x afhangt, mogen we verwachten dat

$$x_1 := \beta - \varepsilon \frac{h(\beta)}{g(\beta)}$$

een betere schatting is voor α.
Door herhaald toepassen van dit idee vinden we een rij getallen x_n, gedefinieerd door

$$\begin{aligned} x_0 &:= \beta \ , \\ x_n &:= \beta - \varepsilon \frac{h(x_{n-1})}{g(x_{n-1})} \ , \quad n = 1,2,\ldots \end{aligned}$$

en we verwachten dat x_n een betere benadering is voor α dan x_{n-1} en dat geldt

$$\lim_{n\to\infty} x_n = \alpha .$$

Voorbeeld.

$$F(x) = x^3 - x^2 - 1.1x - 1.95 .$$

We merken op dat

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

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

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

Hieruit volgt voor de rij getallen x_n:

$$\begin{aligned} x_0 &= 2 \\ x_1 &= 2.021429 \\ x_2 &= 2.021406 \\ x_3 &= 2.021406 . \end{aligned}$$

1.1. Successieve substitutie

Veel methoden voor het oplossen van (1) komen neer op het volgende.

a) Men schrijft de vergelijking (1) in de vorm

$$x = f(x) \tag{5}$$

die equivalent is met (1), d.w.z. dat een oplossing α van (5) ook oplossing van (1) is. Bovendien moet (5) zodanig zijn dat $f(x)$ in de buurt van α niet sterk van x afhangt.

b) Men kiest (op grond van reeds verworven kennis of intuïtie) een nulde benadering x_0 voor de oplossing α en bepaalt vervolgens de rij $x_1, x_2, \ldots$ met de formule

$$x_n := f(x_{n-1}) , \quad n = 1,2,\ldots \tag{6}$$

en men hoopt dat

$$\lim_{n\to\infty} x_n = \alpha . \tag{7}$$

Dit is de methode van de successieve substitutie.

Ga na dat als (7) geldt, en f een continue functie is, α oplossing is van (1).

Voorbeeld. Als in de vergelijking

$$x + \gamma x^3 = y$$

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

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

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

$$x_1 := y - \gamma x_0^3$$

$$x_2 := y - \gamma x_1^3$$

. . . . . .

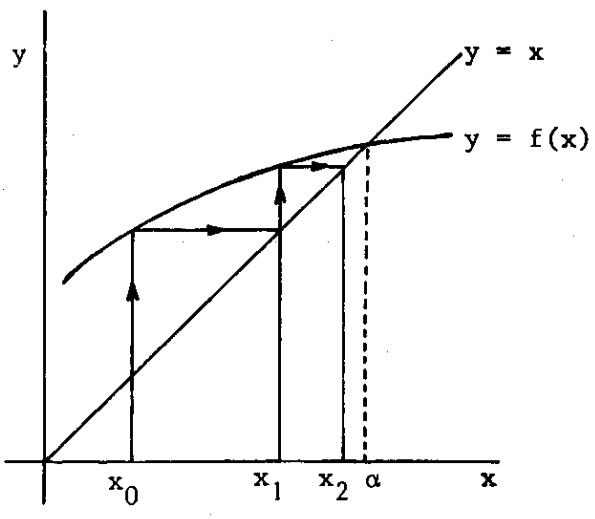
Opmerking. De "gezond-verstand methode" op pag. 1.1 is ook een voorbeeld van successieve substitutie. In dit geval hebben we

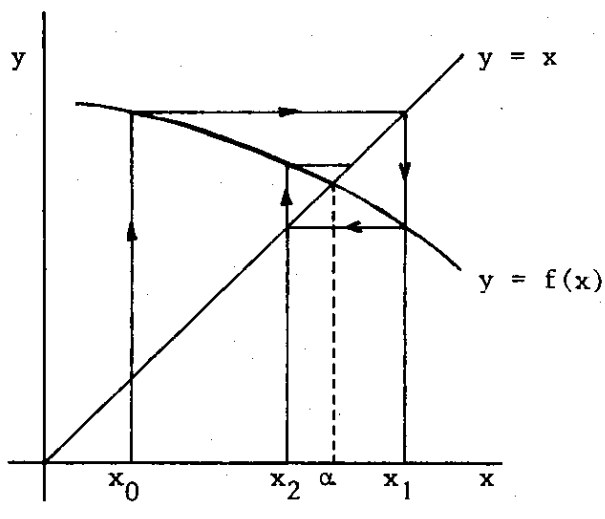
$$f(x) = \beta - \varepsilon \frac{h(x)}{g(x)} = x - \frac{F(x)}{g(x)}$$

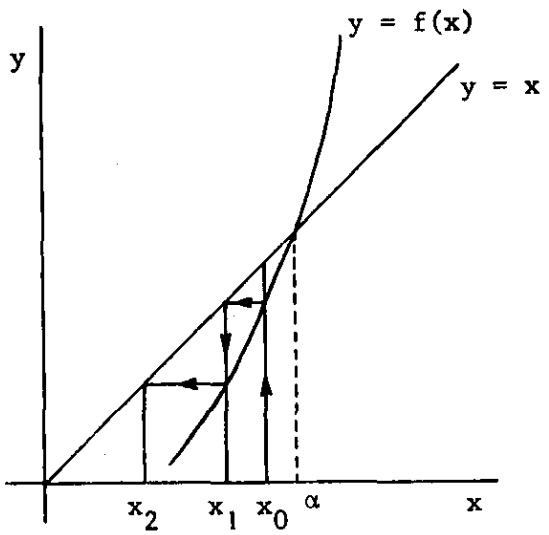
genomen.

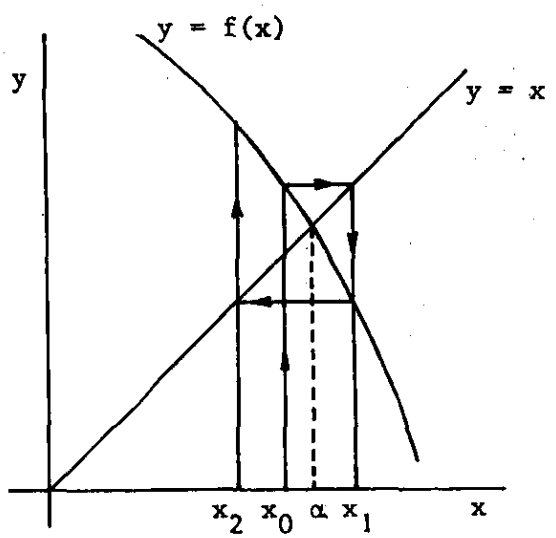
1.1.1. Locale convergentie

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 (5) 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 $x = f(x)$, en zij $x_n = f(x_{n-1})$, $n = 1,2,\ldots$ bij gegeven x_0

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

Als voor zekere k geldt $x_k = \alpha$, dan geldt $x_n = \alpha$, $n \geq k$.

Als $x_n \neq \alpha$ voor alle n, dan geldt

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

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

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

Bewijs. Als $|f'(\alpha)| < 1$ en $f'(x)$ continu is, dan is er een $\delta > 0$ en een L met $0 \leq L < 1$ 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 $|\xi_0 - \alpha| \leq \delta$. Derhalve geldt

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

en met name

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

Analoog $|x_{n-1} - \alpha| \leq \delta$ en

$$|x_n - \alpha| \leq L|x_{n-1} - \alpha| \leq L^n|x_0 - \alpha| . \qquad (8)$$

Daar $0 \leq L < 1$ 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

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

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 den 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

$$A_n := \frac{x_n - x_{n-1}}{x_{n-1} - x_{n-2}} \tag{9}$$

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

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

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 iv) volgt dat dan op den duur $|\alpha - x_n| < \frac{1}{2}|x_n - x_{n-1}|$.

1.1.2. Convergentie orde en convergentiefactor

Uit de locale convergentiestelling volgt dat voor een rij $\{x_n\}$, verkregen met het successieve substitutieproces (6), 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

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

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 asymptotische convergentiesnelheid R is alleen gedefinieerd als $A \neq 0$. Uit (10) 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 en B de asymptotische foutconstante van het proces.
Voor de convergentie orde p geldt in ieder geval $p \geq 1$ (waarom?).
Als $p = 1$, dan spreken we van lineaire convergentie. In dat geval geldt $B \leq 1$ (ga na).
Als $p = 2$, resp. $p = 3$, dan spreken we van kwadratische, resp. kubische, convergentie.

Opmerking. Als $p > 1$, dan is de asymptotische convergentiefactor A van het proces gelijk aan nul. Uit de locale convergentiestelling volgt dat het proces locaal convergeert, ongeacht de 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 (6) 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 $f'(\alpha) = 0$ en $f''(\alpha) \neq 0$, dan is het proces kwadratisch convergent. 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)

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

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 als $f'(\alpha) = 0$

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

3. Ook kwadratische convergentie is een typisch locale zaak. Dit blijkt al uit opmerking 1: 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

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

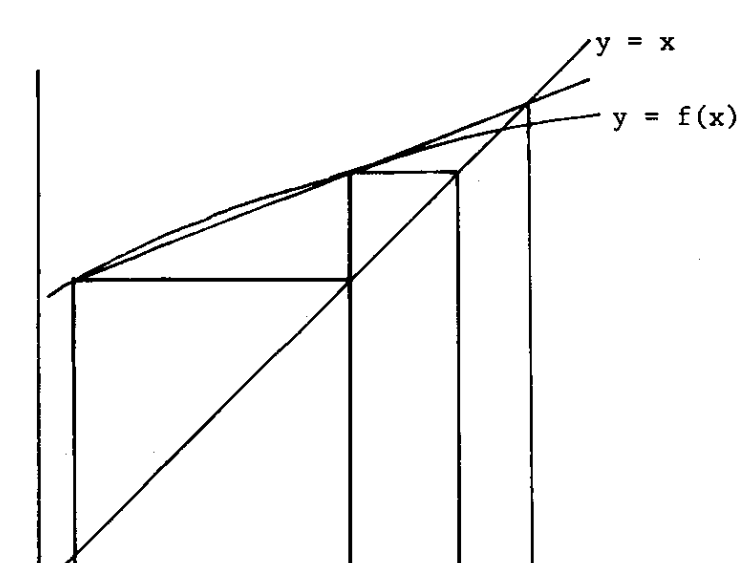
Vervangen we in deze formule A door A_n uit formule (9), dan vinden we als benadering voor de fout in x_n

$$\alpha - x_n \approx \frac{A_n}{1 - A_n} (x_n - x_{n-1}) . \tag{11}$$

We kunnen daarom verwachten dat

$$\alpha_n := x_n + \frac{A_n}{1 - A_n} (x_n - x_{n-1}) \tag{12}$$

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

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

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

of met behulp van voorwaartse differenties

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

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

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

Men kan bewijzen dat voor een lineair convergent proces ($A \neq 0$ en $A \neq 1$) geldt

$$\lim_{n\to\infty} \frac{\alpha_n - \alpha}{x_n - \alpha} = 0 \ .$$

Hieruit volgt dat (11) een goede schatting voor de fout in x_n geeft mits A niet dicht bij 0 ligt.
Als daarentegen A = 0 (maar ook als A bijna nul is), dan heeft de schatting in (11) geen betekenis. Aitken-extrapolatie werkt in dit geval dan ook meestal averechts.
Dit betekent dat Aitken-extrapolatie niet zomaar zonder meer toegepast kan worden. Men moet zich er van tevoren terdege van overtuigd hebben dat het iteratieproces lineair convergeert.

1.1.4. Conditie en numerieke stabiliteit

Het probleem waarvan we de conditie willen onderzoeken is de bepaling van een wortel van de vergelijking (5).
Om de conditie te bepalen gaan we na met welk bedrag $\delta\alpha$ de wortel α verandert als f veranderd wordt in een naburige functie $\tilde{f}$. Als we aannemen dat $\tilde{f}(x) = f(x) + \delta f(x)$, dan geldt

$$\alpha + \delta\alpha = \tilde{f}(\alpha + \delta\alpha) = f(\alpha + \delta\alpha) + \delta f(\alpha) + \ldots$$

$$= f(\alpha) + f'(\alpha)\delta\alpha + \delta f(\alpha) + \ldots \ .$$

Hieruit volgt, als we hogere orde termen verwaarlozen,

$$\frac{\delta\alpha}{\alpha} = \frac{1}{1 - f'(\alpha)} \ \frac{\delta f(\alpha)}{f(\alpha)} \ . \qquad (13)$$

De conditie wordt dus bepaald door de factor $|1/(1 - f'(\alpha))|$ die we het conditiegetal zullen noemen. We zien dus dat het probleem slecht geconditioneerd is als $f'(\alpha) \approx 1$. In alle andere gevallen is het probleem goed geconditioneerd (met name ook als $f'(\alpha) \approx -1$).

De numerieke stabiliteit van het successieve substitutieproces (6) onderzoeken we door na te gaan wat het effect is op de oplossing van het feit dat f(x), ten gevolge van afrondfouten, niet exact berekend wordt.

Veronderstel dat de getallen $\bar{x}_1, \bar{x}_2, \ldots$ voldoen aan

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

waarbij van δ_n slechts bekend is dat

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

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

$$|\bar{x}_n - x_n| \leq |f(\bar{x}_{n-1}) - f(x_{n-1})| + |\delta_n| \leq L|\bar{x}_{n-1} - x_{n-1}| + \delta \ .$$

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

$$|\bar{x}_n - x_n| \leq \frac{1 - L^n}{1 - L}\,\delta \ . \tag{14}$$

$\bar{x}_n$ 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 (8) en (14) volgt

$$|\bar{x}_n - \alpha| \leq |\bar{x}_n - x_n| + |x_n - \alpha| \leq \frac{1 - L^n}{1 - L}\,\delta + L^n|x_0 - \alpha| \ ,$$

zodat voor iedere $\varepsilon > 0$ op den duur geldt

$$|\bar{x}_n - \alpha| \leq \frac{1}{1 - L}\,\delta + \varepsilon \ . \tag{15}$$

We vergelijken deze formule met (13) waarbij we opmerken dat $L \approx |f'(\alpha)|$ genomen mag worden. We zien dan dat, als $f'(\alpha) > 0$, de afwijking in de berekende oplossing op den duur niet essentieel groter is dan de afwijking ten gevolge van de onvermijdbare storing in de gegeven functie. Ook als $f'(\alpha) < 0$, maar niet te dicht bij -1 ligt, zijn de gevonden afwijkingen in (13) en (15) vergelijkbaar in grootte.
Hieruit volgt dat successieve substitutie een numeriek stabiel proces is, als $f'(\alpha)$ niet te dicht bij -1 ligt.

Een andere vraag is nog hoe de fout in de berekende waarde $\bar{x}_n$ samenhangt met de berekende differentie $\bar{x}_n - \bar{x}_{n-1}$.

Opgave. Bewijs dat hiervoor geldt

$$|\bar{x}_n - \alpha| \leq \frac{L}{1 - L} |\bar{x}_n - \bar{x}_{n-1}| + \frac{\delta}{1 - L} .$$

Opmerking. Bij het bedenken van een stopcriterium voor het iteratieproces moeten we er rekening mee houden dat het mogelijk is dat $|\bar{x}_n - \bar{x}_{n-1}|$ niet kleiner wordt dan $2\delta/(1 - L)$.

1.1.5. Globale convergentie

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

Globale convergentiestelling

Zij I het gesloten interval $a \leq x \leq b$.
Zij $f(x)$ een functie die gedefinieerd is voor iedere $x \in I$, met de volgende eigenschappen.
1. f beeldt I in zichzelf af (dus $a \leq f(x) \leq b$ voor alle $x \in I$);
2. er is een L met $0 \leq L < 1$, zo dat voor alle $x' \in I$ en $x'' \in I$ geldt

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

Dan heeft de vergelijking $x = f(x)$ precies één oplossing α in I. Voor iedere $x_0 \in I$ convergeert het successieve substitutieproces en er geldt

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

Dit is een z.g. "fixed point" stelling. De voorwaarde 1. zegt dat bij de afbeelding $x \to f(x)$ van alle punten uit I het beeld ook in I ligt. De voorwaarde 2. geeft aan dat de afbeelding een z.g. contraherende afbeelding is. En de stelling zegt dat er in I precies één vast punt α is dat op zichzelf wordt afgebeeld.
Bovendien convergeert bij ieder beginpunt $x_0 \in I$ 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 2. is zeker voldaan als in I de afgeleide $f'(x)$ bestaat en $|f'(x)| \leq L$.
2. Aan de voorwaarde 1. is voldaan als aan 2. is voldaan en er een $c \in I$ is, zodat het interval bepaald door $|x - c| \leq |f(c) - c|/(1 - L)$ geheel in I ligt.
3. Het bewijs van de stelling in één dimensie is niet moeilijk. Uit 1. volgt (ga na) dat $f(x) - x \geq 0$ voor $x = a$ en $f(x) - x \leq 0$ voor $x = b$
   $f(x) - x$ moet dus minstens één nulpunt hebben in het interval I.
   Uit 2. volgt echter dat er ook hoogstens één oplossing is: als α en α' beide oplossing zijn, dan is $|\alpha - \alpha'| = |f(\alpha) - f(\alpha')| \leq L|\alpha - \alpha'|$, dus $|\alpha - \alpha'| = 0$. De rest van het bewijs gaat als bij de locale convergentiestelling, zie pag. 1.4.

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 van $F(x) = 0$ geldt dat $|f'(x)| < 1$ is (en liefst zo klein mogelijk).
Een mogelijkheid is om voor $f(x)$ 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 **locaal tenminste kwadratisch** als $\varphi(\alpha)F'(\alpha) = 1$.

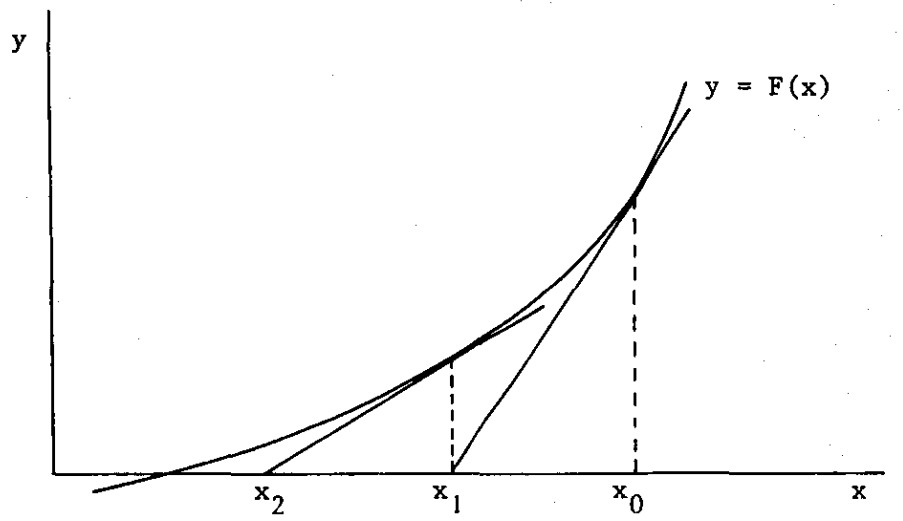
Hoe vinden we bij gegeven $F(x)$ een geschikte $\varphi(x)$?

1.2.1. De iteratiemethode van Newton

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

$$x_n := x_{n-1} - \frac{F(x_{n-1})}{F'(x_{n-1})} \ . \qquad (1)$$

Dit is de formule van Newton.



Meetkundig betekent (1) 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

$$F(x) = 0 ,$$

<u>lineariseer</u> 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

$$f(x) = x - \frac{F(x)}{F'(x)} . \qquad (2)$$

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:

$$\lim_{n\to\infty} \frac{x_n - \alpha}{(x_{n-1} - \alpha)^2} = \frac{F''(\alpha)}{2F'(\alpha)} . \qquad (3)$$

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 (2) volgt hieruit

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

Hieruit volgt direct (3).

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).

3. Een Newton proces kan heel langzaam convergeren als x_n nog ver van de limiet α af is.
Voorbeeld: Als $F(x) = x^2 - a$ met $a > 0$, dan convergeert het proces op grond van opmerking 2 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.

4. De veronderstelling $F'(\alpha) \neq 0$ betekent dat α een enkelvoudig nulpunt is. Als $F(\alpha) = F'(\alpha) = \ldots = F^{(m-1)}(\alpha) = 0$, $F^{(m)}(\alpha) \neq 0$, dan is α een m-voudig nulpunt. Als $m > 1$, dan is het Newton proces locaal slechts lineair convergent met asymptotische convergentiefactor $1 - 1/m$ (ga na).

Voorbeeld

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

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

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

1.2.2. De vaste-richting methode

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

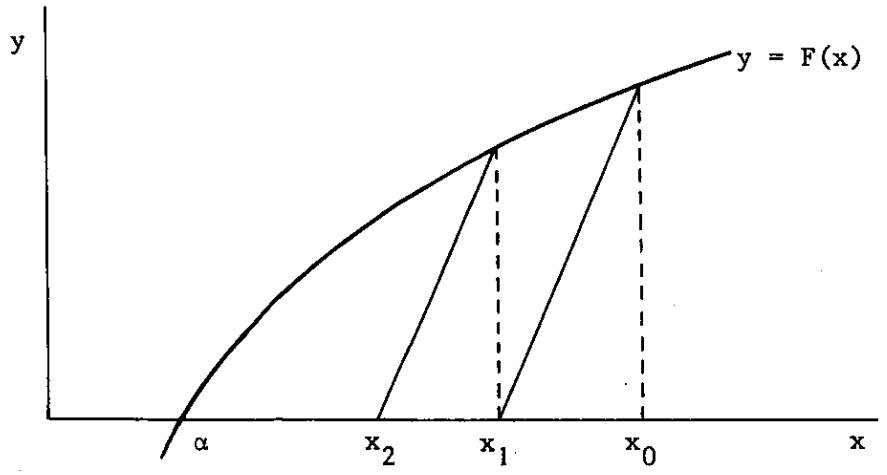
$$\left|1 - \frac{F'(\alpha)}{m}\right| < 1 . \qquad (4)$$

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

$$x_n = x_{n-1} - \frac{1}{m} F(x_{n-1}) \qquad (5)$$

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



Meetkundig betekent de formule (5) 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 (4) zegt dat de helling van deze rechte meer dan half zo groot moet zijn als die van de raaklijn in $x = \alpha$ aan $y = F(x)$.

Opmerking. De vaste-richting methode komt meestal voor, doordat men bij de methode van Newton de waarde van de afgeleide slechts één keer berekent, in het punt x_0. In plaats van (1) gebruikt men dan

$$x_n := x_{n-1} - \frac{F(x_{n-1})}{F'(x_0)} , \qquad (6)$$

(dus $m = F'(x_0)$). Dit wordt ook wel "simplified Newton" genoemd.

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 $a_0 \le x \le b_0$ en zij $F(a_0) < 0 < F(b_0)$ of omgekeerd. Dan heeft F(x) minstens één nulpunt in $a_0 \le x \le b_0$. 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 $b_n - a_n \le \varepsilon$. Dan is $c_n := (a_n + b_n)/2$ een benadering voor een nulpunt van F(x) met een absolute fout die kleiner is dan $\frac{1}{2}\varepsilon$.
In pseudo-ALGOL ziet deze algorithme er als volgt uit:

```
while b - a > eps
do begin c := (a + b)/2;
         if F(c) = 0
         then a := b := c
         else if sign(F(c)) = sign(F(a))
              then a := c else b := c
   end;
c := (a + b)/2
```

Het is duidelijk dat dit proces steeds convergeert, echter slechts met een factor ½ (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 (Regula Falsi)

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:

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

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

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 $1/F'(\alpha)$ en dan lijkt formule (1) op die van Newton. Als regel blijft echter op den duur of a_n of b_n vast. Het is dan voordelig het proces te combineren met interval halvering.

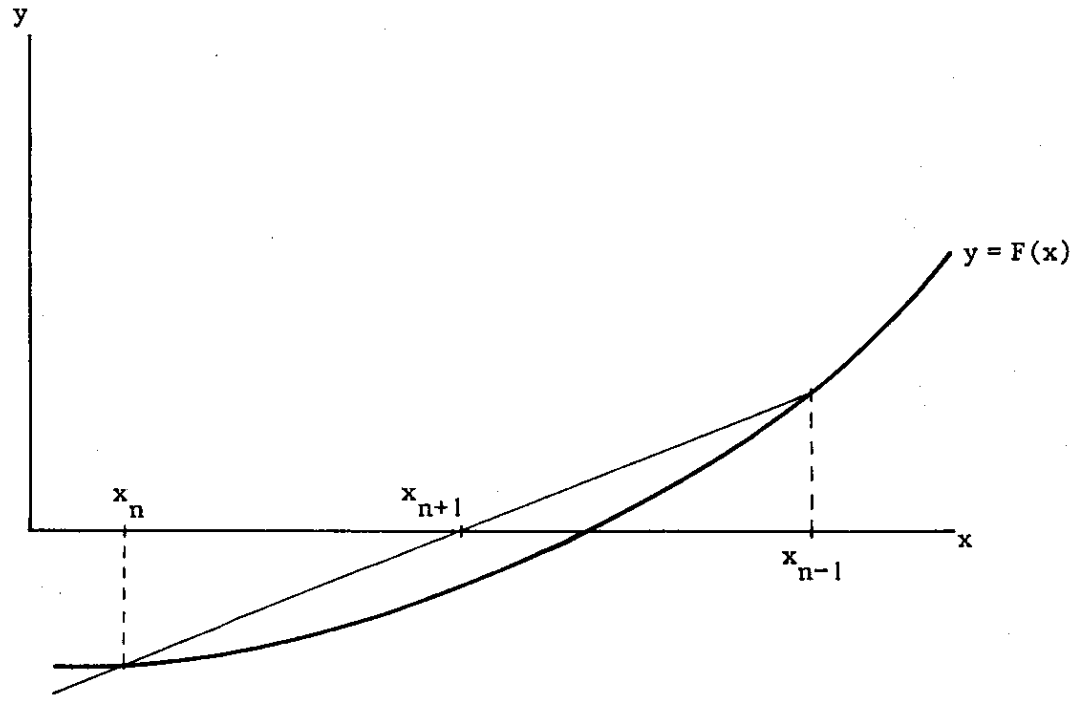
1.3.3. Secant (koorde) methode

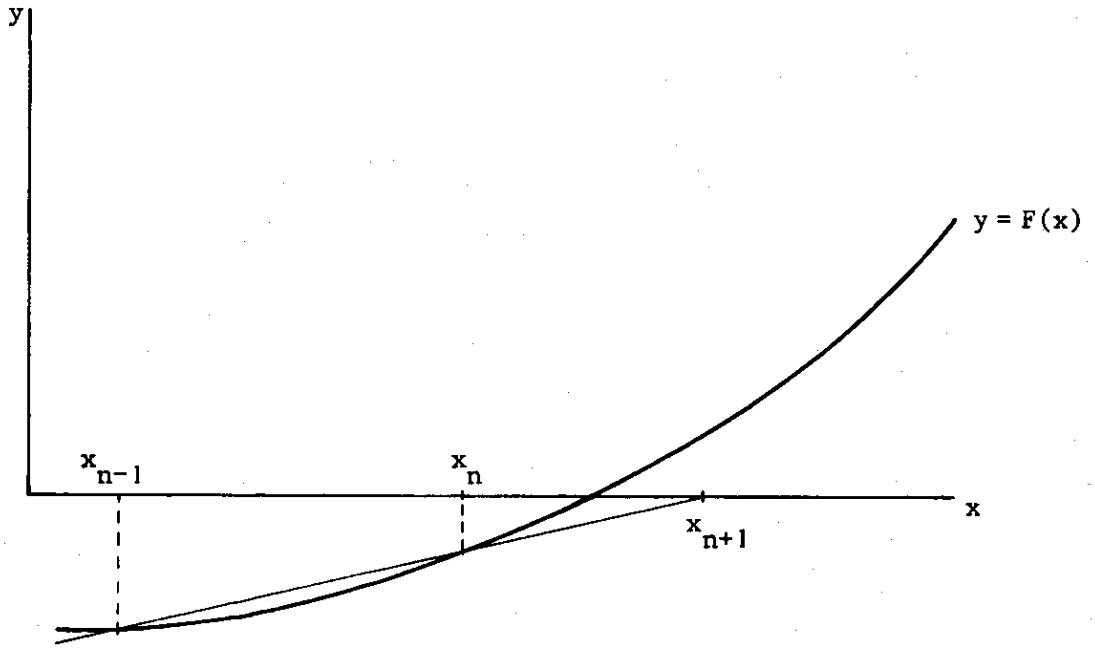
Als men twee benaderingen x_n en x_{n-1} voor een nulpunt α van $F(x)$ heeft, dan kan men als volgende benadering nemen het snijpunt

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

van de rechte door de punten $(x_n,F(x_n))$ en $(x_{n-1},F(x_{n-1}))$ met de x-as.

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 interpoleren of extrapoleren op basis van de twee laatst gevonden benaderingen, nadert, als het proces convergeert,

$\frac{x_n - x_{n-1}}{F(x_n) - F(x_{n-1})}$ tot $(F'(\alpha))^{-1}$, en daarom convergeert het proces snel.

Men kan bewijzen dat (als $F'(\alpha) \neq 0$)

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

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

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

waarin p de positieve wortel is van de vergelijking $p^2 - p - 1 = 0$, dus $p = \frac{1}{2}(1 + \sqrt{5}) \doteq 1.62$. 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

We zullen ons nu bezighouden met het oplossen van stelsels van k vergelijkingen met k onbekenden

$$\begin{aligned} F_1(x_1, x_2, \ldots, x_k) &= 0 \\ F_2(x_1, x_2, \ldots, x_k) &= 0 \\ &\vdots \\ F_k(x_1, x_2, \ldots, x_k) &= 0 \end{aligned} \qquad (1)$$

ofwel in vectornotatie

$$\underline{F}(\underline{x}) = \underline{0} . \qquad (2)$$

We spreken van lineaire vergelijkingen indien de functies F_i (als regel inhomogeen) lineair zijn in $x_1, x_2, \ldots, x_k$, dus als

$$F_i(x_1, x_2, \ldots, x_k) := \sum_{j=1}^{k} A_{ij} x_j - b_i .$$

Dit type vergelijkingen wordt in hoofdstuk 5 uitvoerig besproken. We concentreren ons nu op het geval dat de functies F_i niet lineair zijn.

1.4.1. Normen van vectoren en matrices ([12], p. 229-236)

Het zal blijken dat de methoden die we zullen behandelen k-dimensionale generalisaties zijn van methoden voor het geval van één vergelijking met één onbekende, zoals successieve substitutie, Newton methode, secant methode. Daarvoor hebben we een generalisatie nodig van het begrip absolute waarde. Dit is het begrip norm.

Definitie. Een norm van een vector $\underline{x} \in \mathbb{R}^k$, genoteerd als $\|\underline{x}\|$, is een reëel getal met de volgende eigenschappen

i) $\|\underline{x}\| \geq 0$, $\|\underline{x}\| = 0$ alleen als $\underline{x} = \underline{0}$,

ii) $\|\alpha\underline{x}\| = |\alpha| \, \|\underline{x}\|$, voor ieder reëel getal α,

iii) $\|\underline{x} + \underline{y}\| \leq \|\underline{x}\| + \|\underline{y}\|$.

Eigenschap iii) wordt de driehoeksongelijkheid genoemd.

Men kan op verscheidene manieren een norm voor vectoren definiëren. In dit college zullen we uitsluitend gebruik maken van de volgende twee normen:

a) de maximum norm

$$\|\underline{x}\|_\infty := \max_{1 \leq j \leq k} \{|x_j|\}, \tag{1}$$

b) de euclidische norm

$$\|\underline{x}\|_2 := \left(\sum_{j=1}^{k} x_j^2 \right)^{\frac{1}{2}} . \tag{2}$$

Men kan eenvoudig nagaan dat beide normen voldoen aan de eigenschappen i) tot en met iii).

Het is eveneens eenvoudig te bewijzen dat voor iedere $\underline{x}$ geldt

$$\|\underline{x}\|_\infty \leq \|\underline{x}\|_2 \leq k^{\frac{1}{2}} \|\underline{x}\|_\infty . \tag{3}$$

Met behulp van het begrip norm kunnen we de begrippen afstand en convergentie invoeren.

1) De afstand d tussen de vectoren $\underline{x}$ en $\underline{y}$ is de norm van de verschilvector, dus

$$d(\underline{x},\underline{y}) := \|\underline{x} - \underline{y}\| .$$

In plaats van de norm van $\underline{x}$ spreken we ook wel van de lengte van $\underline{x}$, dit is tevens de afstand van $\underline{x}$ tot de oorsprong.

2) De rij vectoren $\{\underline{x}_n\}$ convergeert naar de vector $\underline{\alpha}$, dus $\lim_{n\to\infty} \underline{x}_n = \underline{\alpha}$, als

$$\lim_{n\to\infty} \|\underline{x}_n - \underline{\alpha}\| = 0 \ .$$

Uit (3) volgt dat een rij $\{\underline{x}_n\}$ die convergeert met betrekking tot de maximum norm, tevens convergeert met betrekking tot de euclidische norm, en omgekeerd.

Met behulp van een norm voor vectoren kunnen we een norm voor matrices (afbeeldingen) invoeren.

Definitie. De norm van de matrix A, genoteerd als $\|A\|$, die afgeleid is van de norm $\|\underline{x}\|$, is

$$\|A\| := \max_{\underline{x}\neq\underline{0}} \frac{\|A\underline{x}\|}{\|\underline{x}\|} \ . \tag{4}$$

Opgave. Bewijs dat $\|A\|$ voldoet aan de eigenschappen i) t/m iii) van de definitie op pag. 1.21, zo dat het inderdaad zinvol is om te spreken van de norm van een matrix.

Voor de normen van matrices en vectoren gelden ook nog de volgende eigenschappen.

$$\|A\underline{x}\| \leq \|A\| \ \|\underline{x}\| \ . \tag{5}$$

Deze eigenschap volgt direct uit de definitie.

$$\|A\ B\| \leq \|A\| \ \|B\| \ . \tag{6}$$

Opgave. Bewijs deze eigenschap.

Het blijkt dat voor de matrixnorm, die behoort bij de maximum norm voor vectoren, geldt

$$\|A\|_\infty = \max_{1\leq i\leq k} \left(\sum_{j=1}^{k} |A_{ij}| \right) \ . \tag{7}$$

Voor de matrixnorm, die behoort bij de euclidische norm voor vectoren geldt

$$\|A\|_2^2 = \rho(A^T A) \ , \tag{8}$$

dit is de spectraalstraal (de grootste eigenwaarde) van de matrix $A^T A$.
Hieruit zien we dat de euclidische norm lastiger te berekenen is dan de maximumnorm. We zullen daarom in het vervolg hoofdzakelijk gebruik maken van de maximum norm (ook wel ∞-norm genoemd).
Ook voor matrices geldt tussen beide normen een eenvoudig verband. Er geldt namelijk voor iedere A

$$k^{-\frac{1}{2}}\|A\|_\infty \le \|A\|_2 \le k^{\frac{1}{2}}\|A\|_\infty \ . \tag{9}$$

Voor het vervolg hebben we ook een generalisatie nodig van de afgeleide van een functie, de Taylorreeksontwikkeling en de middelwaardestelling.

Een geschikte generalisatie van de <u>afgeleide</u> blijkt de functionaalmatrix (matrix van Jacobi) te zijn.

$$F'(\underline{x}) := \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_k} \\ \vdots & & & \vdots \\ \frac{\partial F_k}{\partial x_1} & \frac{\partial F_k}{\partial x_2} & \cdots & \frac{\partial F_k}{\partial x_k} \end{pmatrix} . \tag{10}$$

Met behulp van deze gegeneraliseerde afgeleide kunnen we de eerste twee termen van de <u>Taylorreeksontwikkeling</u> van de vectorfunctie $\underline{F}$ opschrijven.

$$\underline{F}(\underline{x}+\underline{h}) = \underline{F}(\underline{x}) + F'(\underline{x})\underline{h} + \mathcal{O}(\|\underline{h}\|^2), \quad \underline{h} \to \underline{0} \ . \tag{11}$$

Deze formule geldt bijvoorbeeld als alle $\frac{\partial^2 F}{\partial x_i \partial x_j}$ continu zijn in een omgeving van $\underline{x}$.

Als $F'(\underline{x})$ continu is, dan geldt de volgende ongelijkheid

$$\|\underline{F}(\underline{y}) - \underline{F}(\underline{x})\| \le \max_{0\le t\le 1} \|F'(\underline{x} + t(\underline{y}-\underline{x}))\| \, \|\underline{y} - \underline{x}\| \ . \tag{12}$$

Dit kunnen we opvatten als een generalisatie van de <u>middelwaardestelling</u>.

.4.2. <u>Successieve substitutie</u>

Voor het toepassen van de methode van de successieve substitutie wordt de vergelijking

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

herschreven in de vorm

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

Dit kan op diverse manieren, en ook hier kan men soms met "gezond verstand" een geschikte formule vinden. Zij bijvoorbeeld

$$\underline{F}(\underline{x}) = A\underline{x} - \underline{b} - \underline{G}(\underline{x}) ,$$

waarbij A een k × k matrix is, $\underline{b}$ een k-vector is en $\underline{G}$ een niet-lineaire vectorfunctie. Dan is

$$\underline{x} = A^{-1}\underline{b} + A^{-1}\underline{G}(\underline{x}) = \underline{x} - A^{-1}\underline{F}(\underline{x}) \qquad (2a)$$

een goede formule van de gedaante (2) als $A^{-1}\underline{G}(\underline{x})$ klein is ten opzichte van $A^{-1}\underline{b}$.

Voorbeeld.

$$\begin{aligned} 2x - 4y - x^2 - y^2 &= -0.4 \\ 2x + y - xy &= 1.0 . \end{aligned}$$

Dan is

$$A = \begin{pmatrix} 2 & -4 \\ 2 & 1 \end{pmatrix} \qquad \underline{G}(\underline{x}) = \begin{pmatrix} x^2 + y^2 \\ xy \end{pmatrix}, \quad \underline{b} = \begin{pmatrix} -0.4 \\ 1.0 \end{pmatrix} .$$

Bovenstaande formule (2a) wordt dan het stelsel

$$x = 0.36 + 0.1\ (x^2 + y^2 + 4xy)$$

$$y = 0.28 - 0.2\ (x^2 + y^2 - xy) .$$

We zien dat deze formulering geschikt is om een oplossing te bepalen omdat het niet-constante gedeelte relatief klein is t.o.v. de constanten. □

In het algemeen kan men (1) omvormen tot (2) door te stellen

$$\underline{x} = \underline{x} - M(\underline{x})\underline{F}(\underline{x}) , \qquad (3)$$

waarin $M(\underline{x})$ een matrix is die niet-singulier is in de buurt van een oplossing $\underline{\alpha}$ van (1).

De methode van de successieve substitutie voor het bepalen van een oplossing $\underline{\alpha}$ van (2) gaat als volgt:
Kies een nulde benadering $\underline{x}_0$ voor $\underline{\alpha}$ en bereken vervolgens de rij $\{\underline{x}_n\}$ met de formule

$$\underline{x}_n := \underline{f}(\underline{x}_{n-1}) , \quad n = 1,2,\ldots . \qquad (4)$$

Analoog aan het eendimensionale geval gelden voor de convergentie van de methode de volgende stellingen.

Locale convergentiestelling

Zij $\underline{\alpha}$ een oplossing van $\underline{x} = \underline{f}(\underline{x})$, en zij $\underline{x}_n = \underline{f}(\underline{x}_{n-1})$, $n = 1,2,\ldots$ bij gegeven $\underline{x}_0$.

Zij $f'(\underline{x})$, de functionaalmatrix, continu in een omgeving van $\underline{\alpha}$.

Zij $\| f'(\underline{\alpha}) \| = A$ met $0 \leq A < 1$.

Dan is er een $\delta > 0$, zodanig dat voor iedere $\underline{x}_0$ met $\| \underline{x}_0 - \underline{\alpha} \| \leq \delta$ geldt $\lim_{n\to\infty} \underline{x}_n = \underline{\alpha}$.

Bewijs. Uit $\| f'(\underline{\alpha}) \| < 1$ en $f'(\underline{x})$ continu volgt dat er een $\delta > 0$ is en een L met $0 \leq L < 1$, zodanig dat $\| f'(\underline{x}) \| \leq L$ als $\| \underline{x} - \underline{\alpha} \| \leq \delta$.
Met behulp van de gegeneraliseerde middelwaardestelling (zie 1.4.1, formule (12)) volgt dan op dezelfde manier als in het eendimensionale geval, zie pag. 1.5,

$$\| \underline{x}_n - \underline{\alpha} \| \leq L \| \underline{x}_{n-1} - \underline{\alpha} \| \leq L^n \| \underline{x}_0 - \underline{\alpha} \|$$

als $\| \underline{x}_0 - \underline{\alpha} \| \leq \delta$. Daar $0 \leq L < 1$ volgt hieruit dat $\| \underline{x}_n - \underline{\alpha} \| \leq \delta$ voor alle n en dat

$$\lim_{n\to\infty} \| \underline{x}_n - \underline{\alpha} \| = 0, \text{ ofwel } \lim_{n\to\infty} \underline{x}_n = \underline{\alpha} \ . \qquad \square$$

Het proces is dus in het algemeen lineair convergent met een asymptotische convergentiefactor kleiner dan of gelijk aan L.

Globale convergentiestelling (= contractiestelling van Banach)

Zij D een gesloten en begrensd gebied in $\mathbb{R}^k$.

Zij $\underline{f}(\underline{x})$ een vectorfunctie die gedefinieerd is voor iedere $\underline{x} \in D$, met de volgende eigenschappen.

1. $\underline{f}$ beeldt D in zichzelf af, d.w.z. voor iedere $\underline{x} \in D$ geldt $\underline{f}(\underline{x}) \in D$.
2. $\underline{f}$ is een contraherende afbeelding, d.w.z. er is een L met $0 \leq L < 1$, zodat voor $\underline{x}' \in D$ en $\underline{x}'' \in D$ geldt

$$\| \underline{f}(\underline{x}') - \underline{f}(\underline{x}'') \| \leq L \| \underline{x}' - \underline{x}'' \| \ . \qquad (5)$$

Dan heeft de vergelijking $\underline{x} = \underline{f}(\underline{x})$ precies één oplossing $\underline{\alpha}$ in D. Voor iedere $\underline{x}_0 \in D$ geldt dat het successieve substitutie proces convergeert naar $\underline{\alpha}$.

Het bewijs dat de vergelijking $\underline{x} = \underline{f}(\underline{x})$ een oplossing heeft in D wordt niet gegeven. Het bewijs dat de oplossing eenduidig is, volgt rechtstreeks uit (5). Het bewijs dat $\lim_{n\to\infty} \underline{x}_n = \underline{\alpha}$ als $\underline{x}_0 \in D$ is hetzelfde als bij de locale convergentiestelling.

Opmerkingen

1. Aan de voorwaarde 2. is zeker voldaan als het gebied D convex is (d.w.z. als $\underline{x} \in D$ en $\underline{y} \in D$, dan geldt $\underline{x} + t(\underline{y} - \underline{x}) \in D$ voor $0 \leq t \leq 1$), en als in D de functionaalmatrix $f'(\underline{x})$ bestaat en $\| f'(\underline{x}) \| \leq L$.
2. Als aan 2. is voldaan en er een $\underline{a} \in D$ is, zodat de bol B, gedefinieerd door $\| \underline{x} - \underline{a} \| \leq \| \underline{f}(\underline{a}) - \underline{a} \| / (1 - L)$ geheel in D ligt, dan is er precies één oplossing $\underline{\alpha}$ in D, $\underline{\alpha}$ ligt in B en voor ieder $\underline{x}_0 \in B$ convergeert het successieve substitutie proces naar $\underline{\alpha}$.

Met behulp van (5) kan weer een bovengrens voor de fout $\|\underline{x}_n - \underline{\alpha}\|$ worden gegeven, uitgedrukt in de laatste differentie $\|\underline{x}_n - \underline{x}_{n-1}\|$. Er geldt namelijk

$$\|\underline{x}_n - \underline{\alpha}\| \leq L \|\underline{x}_{n-1} - \underline{\alpha}\| \leq L \{\|\underline{x}_{n-1} - \underline{x}_n\| + \|\underline{x}_n - \underline{\alpha}\|\} ,$$

dus

$$\|\underline{x}_n - \underline{\alpha}\| \leq \frac{L}{1-L} \|\underline{x}_n - \underline{x}_{n-1}\| . \qquad (6)$$

Voorbeeld. Beschouw een stelsel van twee vergelijkingen met twee onbekenden:

$$\begin{aligned} x &= f(x,y) \\ y &= g(x,y) \end{aligned} \qquad (7)$$

in het gebied $D := \{|x| \leq R,\ |y| \leq R\}$.
Veronderstel dat voor iedere $(x,y) \in D$ geldt

$$|f(x,y)| \leq R,\ |g(x,y)| \leq R$$

en

$$|\frac{\partial f}{\partial x}| + |\frac{\partial f}{\partial y}| \leq L ,\ |\frac{\partial g}{\partial x}| + |\frac{\partial g}{\partial y}| \leq L ,$$

waarbij $0 \leq L < 1$.
Dan heeft het stelsel (7) precies één oplossing (α,β) met $|\alpha| \leq R$ en $|\beta| \leq R$ en als $(x_0,y_0) \in D$, dan geldt voor de rij $\{(x_n,y_n)\}$, verkregen met successieve substitutie, dus

$$\begin{aligned} x_n &= f(x_{n-1},y_{n-1}) \\ y_n &= g(x_{n-1},y_{n-1}) \end{aligned} , \quad n = 1,2,...$$

dat

$$\lim x_n = \alpha ,\ \lim y_n = \beta$$

en

$$\max(|x_n - \alpha|, |y_n - \beta|) \leq \frac{L}{1-L} \max (|x_n - x_{n-1}|, |y_n - y_{n-1}|) .$$

In het concrete voorbeeld op pag. 1.24 is

$$f(x,y) = 0.36 + 0.1\ (x^2 + y^2 + 4xy)$$

$$g(x,y) = 0.28 - 0.2\ (x^2 + y^2 - xy) .$$

Dan wordt het gebied

$$D := \{0 \leq x \leq 0.5,\ 0 \leq y \leq 0.3\}$$

in zichzelf afgebeeld. De matrix van Jacobi is

$$J := \begin{pmatrix} 0.2x + 0.4y & 0.4x + 0.2y \\ -0.4x + 0.2y & 0.2x - 0.4y \end{pmatrix} .$$

In D geldt $\|J\| \leq 0.48$. Dus volgens opmerking 1. geldt (5) met $L = 0.48$.
Volgens de globale convergentiestelling heeft dit stelsel precies één oplossing in D.
De methode van de successieve substitutie levert bij de startvector $\underline{x}_0 = (0,0)$

n	x_n	y_n
0	0	0
1	0.36	0.28
2	0.42112	0.25856
3	0.42797	0.25294
4	0.42801	0.25222
5	0.42786	0.25223 .

De foutschatting (6) levert bijvoorbeeld dat de fout in $\underline{x}_4$ kleiner is dan $0.7_{10}-3$. Deze schatting is erg grof, onder andere omdat de waarde $L = 0.48$ erg grof is. □

1.4.3. De methode van Newton

De methode van Newton voor het oplossen van het stelsel vergelijkingen

$$\underline{F}(\underline{x}) = \underline{0} \tag{8}$$

kunnen we op de volgende manier krijgen.
We veronderstellen dat $\underline{F}(\underline{x})$ een Taylorreeksontwikkeling (zie 1.4.1, formule (7)) heeft. Dan kunnen we (8) rond $\underline{x}_{n-1}$ lineariseren. Dit levert de vergelijking

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

Als $F'(\underline{x}_{n-1})$ inverteerbaar is, dan kunnen we de oplossing van deze gelineariseerde vergelijking als volgende benadering $\underline{x}_n$ nemen, dus (vergelijk 1.2.1, formule (1)).

$$\underline{x}_n = \underline{x}_{n-1} - (F'(\underline{x}_{n-1}))^{-1}\underline{F}(\underline{x}_{n-1}) . \tag{9}$$

Opmerking. In de praktijk zullen we niet de inverse matrix $(F'(\underline{x}_{n-1}))^{-1}$ berekenen en vervolgens $\underline{x}_n$ met behulp van (9), maar lossen we eerst het stelsel lineaire vergelijkingen

$$F'(\underline{x}_{n-1})\underline{d}_{n-1} = -\underline{F}(\underline{x}_{n-1}) \qquad (10)$$

op (zie hiervoor hoofdstuk 5) en berekenen daarmee

$$\underline{x}_n := \underline{x}_{n-1} + \underline{d}_{n-1} .$$

Voorbeeld. Beschouw het stelsel van twee vergelijkingen met twee onbekenden

$$\begin{aligned} F(x,y) &= 0 \\ G(x,y) &= 0 . \end{aligned} \qquad (11)$$

Zij (α,β) een oplossing van (11). Zij (x_0,y_0) een punt in de buurt van (α,β), dan geldt (Taylorreeksontwikkeling)

$$F(x,y) = F(x_0,y_0) + \left(\frac{\partial F}{\partial x}\right)_0 (x-x_0) + \left(\frac{\partial F}{\partial y}\right)_0 (y-y_0) + \ldots$$

$$G(x,y) = G(x_0,y_0) + \left(\frac{\partial G}{\partial x}\right)_0 (x-x_0) + \left(\frac{\partial G}{\partial y}\right)_0 (y-y_0) + \ldots$$

waarin $\left(\frac{\partial F}{\partial x}\right)_0 = \frac{\partial F}{\partial x}(x_0,y_0)$, etc.

De gelineariseerde vergelijkingen (10) zijn dan

$$\left(\frac{\partial F}{\partial x}\right)_0 d_0 + \left(\frac{\partial F}{\partial y}\right)_0 e_0 = -F(x_0,y_0)$$

$$\left(\frac{\partial G}{\partial x}\right)_0 d_0 + \left(\frac{\partial G}{\partial y}\right)_0 e_0 = -G(x_0,y_0) .$$

Dit stelsel lineaire vergelijkingen heeft een oplossing als de functionaalmatrix

$$F'(\underline{x}_0) := \begin{pmatrix} \left(\frac{\partial F}{\partial x}\right)_0 & \left(\frac{\partial F}{\partial y}\right)_0 \\ \left(\frac{\partial G}{\partial x}\right)_0 & \left(\frac{\partial G}{\partial y}\right)_0 \end{pmatrix}$$

niet-singulier is. Als nieuwe benadering vinden we dan

$$x_1 = x_0 + d_0 .$$

$$y_1 = y_0 + e_0 .$$

In het concrete voorbeeld op pag. 1.24 is

$$F(x,y) = 2x - 4y - x^2 - y^2 + 0.4$$
$$G(x,y) = 2x + y - xy - 1 .$$

De bijbehorende functionaalmatrix is

$$\begin{pmatrix} 2 - 2x & -4 - 2y \\ 2 - y & 1 - x \end{pmatrix} ,$$

deze matrix is niet-singulier in $D := \{0 \le x \le 0.5 , 0 \le y \le 0.3\}$.
Toepassing van de methode van Newton levert bij de startvector $\underline{x}_0 = (0,0)$

n	x_n	y_n
0	0	0
1	0.36	0.28
2	0.4284 30	0.2535 94
3	0.4278 35	0.2522 50
4	0.4278 36	0.2522 50

□

Men kan bewijzen dat, als $\underline{F}$ een voldoende gladde vectorfunctie is, en $F'(\underline{\alpha})$ niet singulier is, het Newtonproces locaal tenminste kwadratisch convergeert. D.w.z. men kan bewijzen dat er een constante $C > 0$ is zodanig dat, als $\underline{x}_0$ voldoende dicht bij $\underline{\alpha}$ ligt, voor alle $n \ge n_0$ geldt

$$\| \underline{x}_n - \underline{\alpha} \| \le C \| \underline{x}_{n-1} - \underline{\alpha} \|^2 .$$

In de praktijk zal men vaak de matrix $F'(\underline{x}_{n-1})$ in (10) na een aantal iteraties niet meer opnieuw berekenen. D.w.z. men vervangt $F'(\underline{x}_{n-1})$ door $F'(\underline{x}_m)$ voor $n-1 > m$. Als $\underline{x}_m$ voldoende dicht bij $\underline{\alpha}$ ligt, dan wordt de convergentie daardoor niet veel langzamer terwijl de berekening van $\underline{x}_n$ veel goedkoper is. Men noemt deze modificatie ook wel "simplified Newton."

1.4.4. Secant methoden

Een andere modificatie van de methode van Newton is een generalisatie van de secant methode. Deze krijgen we door in de n-de iteratiestap de functionaalmatrix $F'(\underline{x}_{n-1})$ te vervangen door een matrix J_{n-1}, die een benadering is van $F'(\underline{x}_{n-1})$.

Een voor de hand liggende manier om een benadering J_{n-1} te vinden, is het vervangen van de afgeleiden door differentiequotiënten, bijvoorbeeld

$$\frac{\partial F_i}{\partial x_j}(\underline{x}) \approx \frac{F_i(\underline{x} + h_{ij}\underline{e}_j) - F_i(\underline{x})}{h_{ij}} , \quad i,j = 1,2,\ldots,k , \tag{12}$$

waarin h_{ij} discretisatieparameters zijn (niet noodzakelijk bij iedere stap dezelfde) en $\underline{e}_j$ de j-de eenheidsvector is. De convergentie van deze methode is lineair als h_{ij} constant, d.w.z. onafhankelijk van n, zijn, maar is in het algemeen meer dan lineair convergent als

$$\lim_{n\to\infty} \max_{ij} |h_{ij}^n| = 0 .$$

Een bezwaar van deze methode is het grote aantal functiewaarden, namelijk k(k+1), dat per iteratiestap moet worden berekend.

We kunnen ook een benadering J_{n-1} van $F'(\underline{x}_{n-1})$ met behulp van reeds bekende functiewaarden bepalen, bijvoorbeeld op de volgende manier.
We vervangen $\underline{F}(\underline{x})$ door een lineaire vectorfunctie, $\tilde{\underline{F}}(\underline{x})$ die in de laatste (k+1) berekende punten $\underline{x}_m$, m = n-k-1 t/m n-1, met $\underline{F}(\underline{x})$ overeenkomt, dus

$$\tilde{\underline{F}}(\underline{x}) = \underline{F}(\underline{x}_{n-1}) + J_{n-1}(\underline{x} - \underline{x}_{n-1}) , \tag{13}$$

waarin de k×k matrix J_{n-1} bepaald is door de interpolatievoorwaarde

$$\underline{F}(\underline{x}_{n-1-i}) = \underline{F}(\underline{x}_{n-1}) + J_{n-1}(\underline{x}_{n-1-i} - \underline{x}_{n-1}) , \quad i = 1,\ldots,k ,$$

ofwel, (door de i-de vergelijking van de $(i-1)$-ste vergelijking af te trekken)

$$\underline{F}(\underline{x}_{n-i}) - \underline{F}(\underline{x}_{n-1-i}) = J_{n-1}(\underline{x}_{n-i} - \underline{x}_{n-i-1}) , \quad i = 1,\ldots,k . \tag{14}$$

De matrix J_{n-1} is hierdoor eenduidig bepaald als de k vectoren $\underline{x}_{n-i} - \underline{x}_{n-i-1}$, i = 1,...,k, lineair onafhankelijk zijn en J_{n-1} is niet-singulier als bovendien de k vectoren $\underline{F}(\underline{x}_{n-i} - \underline{F}(\underline{x}_{n-i-1})$ lineair onafhankelijk zijn.
Als volgende benadering van de oplossing $\underline{\alpha}$ nemen we de oplossing van $\tilde{\underline{F}}(\underline{x}) = \underline{0}$, dus

$$\underline{x}_n = \underline{x}_{n-1} - J_{n-1}^{-1}\underline{F}(\underline{x}_{n-1}) . \tag{15}$$

Men kan bewijzen dat $\lim_{n\to\infty} J_{n-1} = F'(\underline{\alpha})$ als $\lim_{n\to\infty} \underline{x}_n = \underline{\alpha}$, en dat de methode locaal meer dan lineair convergeert als $\underline{F}$ een voldoende gladde functie is.
Er is wel een aparte startprocedure nodig voor de berekening van $\underline{x}_1$ t/m $\underline{x}_k$.
Er bestaan efficiënte algorithmen voor het berekenen van $\underline{x}_n$. Deze maken gebruik van het feit dat het verschil tussen de matrices J_{n-1}^{-1} en J_{n-2}^{-1}, de matrix uit de vorige iteratiestap, een matrix van de rang 1 is, die vrij goedkoop te berekenen is.
Echter, het is mogelijk dat de berekende $\underline{x}_n$ zodanig is dat de k vectoren $\underline{x}_{m+1} - \underline{x}_m$, m = n-k t/m n-1, lineair afhankelijk, resp. bijna afhankelijk, zijn. In dat geval is de matrix J_n niet, resp. slecht bepaald. Deze mogelijke ontsporing, die in praktische gevallen nogal eens voorkomt en die een vorm van instabiliteit is, maakt dat de methode niet erg in trek is.
Men kan de methode verbeteren door niet persé het "oudste" punt, d.i. $\underline{x}_{n-k-1}$, weg te laten, maar het weg te laten punt $\underline{x}_j$ zodanig te kiezen dat de vectoren $\underline{x}_{m+1} - \underline{x}_m$ zo goed mogelijk onafhankelijk zijn.

In de praktijk kiest men bij voorkeur een methode waarbij J_{n-1} door numerieke differentiatie wordt verkregen, als het berekenen van de functiewaarden niet al te duur is.

1.4.5. Minimaliseren van functies ([2], p. 438-442)

Een probleem dat nauw verwant is met het oplossen van een stelsel niet-lineaire vergelijkingen $\underline{F}(\underline{x}) = \underline{0}$ is het minimaliseren van een functie $\varphi(\underline{x})$.
Zij $\underline{F} :=$ grad φ, ofwel

$$F_i(\underline{x}) := \frac{\partial \varphi}{\partial x_i}(\underline{x}),\ i = 1,2,\ldots,k\ . \tag{16}$$

Als $\underline{\alpha}$ een punt is waarin $\varphi(\underline{x})$ minimaal is, dan is in $\underline{x} = \underline{\alpha}$ de gradiënt nul, dus dan is $\underline{\alpha}$ oplossing van het stelsel vergelijkingen $\underline{F}(\underline{x}) = \underline{0}$.
De voorwaarde $\underline{F}(\underline{\alpha}) = \underline{0}$ is nodig, maar niet voldoende.

Beschouw de Taylorreeksontwikkeling van $\varphi(\underline{x})$ rond $\underline{x} = \underline{\alpha}$.

$$\varphi(\underline{\alpha} + \underline{h}) = \varphi(\underline{\alpha}) + \underline{F}(\underline{\alpha})^T\underline{h} + \tfrac{1}{2}\underline{h}^T G(\underline{\alpha})\underline{h} + \mathcal{O}(\|\underline{h}\|^3)\ . \tag{17}$$

Hierin is G de k×k-matrix, gedefinieerd door

$$G_{ij}(\underline{x}) := \frac{\partial^2 \varphi}{\partial x_i \partial x_j}(\underline{x})\ ,\ i,j = 1,2,\ldots,k\ . \tag{18}$$

De matrix G, die ook wel de Hessiaan van φ wordt genoemd, is symmetrisch.

Voor het formuleren van voldoende voorwaarden voor het minimaliseringsprobleem hebben we het begrip positief definiete matrix nodig.

Definitie. Een symmetrische matrix A is positief definiet als voor iedere $\underline{x} \neq \underline{0}$ geldt

$$\underline{x}^T A \underline{x} > 0 \ . \tag{19}$$

Voldoende voorwaarden: Als $\underline{F}(\underline{\alpha}) = \underline{0}$ en $G(\underline{\alpha})$ is een positief definiete matrix dan heeft $\varphi(\underline{x})$ in $\underline{x} = \underline{\alpha}$ een (relatief) minimum, d.w.z. er is een $\delta > 0$ zo dat $\varphi(\underline{\alpha} + \underline{h}) > \varphi(\underline{\alpha})$ voor alle $\underline{h}$ met $0 < \|\underline{h}\| < \delta$.

We zouden voor het oplossing van het minimaliseringsprobleem kunnen volstaan met een verwijzing naar de voorgaande paragrafen over het oplossen van $\underline{F}(\underline{x}) = \underline{0}$. Echter, voor het oplossen van een minimaliseringsprobleem zijn er speciale methoden die, vooral m.b.t. de globale convergentie, beter zijn dan algemene methoden voor het oplossen van stelsels vergelijkingen.

Methoden voor het bepalen van een minimum van de functie $\varphi(\underline{x})$ zien er in grote trekken als volgt uit.

Zij $\underline{x}_{n-1}$ de laatstberekende benadering van het minimum punt $\underline{\alpha}$.

1) Bepaal een zoekrichting $\underline{d}_n$.
2) Bepaal de waarde λ_n, die de lengte van de stap in de richting $\underline{d}_n$ bepaalt.
3) Neem $\underline{x}_n := \underline{x}_{n-1} + \lambda_n \underline{d}_n$ als volgende benadering van $\underline{\alpha}$.

Het verschil tussen de diverse methoden betreft de keuze van de zoekrichting $\underline{d}_n$ en de wijze waarop λ_n wordt bepaald.

In het algemeen zullen we trachten te bereiken dat

$$\varphi(\underline{x}_n) < \varphi(\underline{x}_{n-1}) \ , \ n = 1,2,\ldots \tag{20}$$

waardoor convergentie verzekerd is.

Methoden waarbij $\underline{d}_n$ zodanig wordt bepaald dat voor voldoende kleine $\lambda > 0$ geldt $\varphi(\underline{x}_{n-1} + \lambda \underline{d}_n) < \varphi(\underline{x}_{n-1})$ noemen we descent methoden. Een voldoende voorwaarde hiervoor is dat $\underline{F}(\underline{x}_{n-1})^T \underline{d}_n < 0$.

Een van de meest gebruikte methoden om de staplengte λ_n te bepalen is lijnminimalisering. Hierbij wordt λ_n bepaald, zodanig dat $\underline{x}_{n-1} + \lambda_n \underline{d}_n$ op de lijn $\ell : \underline{x} = \underline{x}_{n-1} + \lambda \underline{d}_n$ het punt is waarin $\varphi(\underline{x})$ minimaal is.

Een voor de hand liggende keuze van $\underline{d}_n$ is de richting van de (negatieve) gradiënt in het punt $\underline{x}_{n-1}$, dus $\underline{d}_n = -\text{grad } \varphi(\underline{x}_{n-1})$, omdat in die richting $\varphi(\underline{x})$ lokaal het sterkst afneemt. Men spreekt dan van een gradiëntmethode. Bepaalt men daarbij λ_n met lijnminimalisering, dan heeft men de methode van de steilste helling (steepest descent).
Een gradiëntmethode voldoet dus aan (20) waardoor convergentie, althans theoretisch, onder zeer ruimte voorwaarden verzekerd is, maar de convergentie is in veel gevallen zeer langzaam.

Als we voor het oplossen van het stelsel vergelijkingen

$$\text{grad } \varphi(\underline{x}) = \underline{0} \tag{21}$$

de in 1.4.3 besproken methode van Newton gebruiken, dan is de bijbehorende functionaalmatrix de Hessiaan en dan zijn de zoekrichting en de staplengte respectievelijk

$$\underline{d}_n := -[G(\underline{x}_{n-1})]^{-1} \text{grad } \varphi(\underline{x}_{n-1}) \tag{22}$$

en

$$\lambda_n := 1 \ .$$

Als $G(\underline{\alpha})$ positief definiet is en $\underline{x}_{n-1}$ voldoende dicht bij $\underline{\alpha}$, dan is $G(\underline{x}_{n-1})$ ook positief definiet, dus niet-singulier en dan bestaat $\underline{d}_n$ zeker.
Tevens volgt dan uit de Taylorreeksontwikkeling van $\varphi(\underline{x})$ rond $\underline{x} = \underline{x}_{n-1}$ (vgl. (17)), met $\underline{F}_{n-1} := \text{grad } \varphi(\underline{x}_{n-1})$,

$$\varphi(\underline{x}_{n-1} + \lambda \underline{d}_n) = \varphi(\underline{x}_{n-1}) + \lambda \underline{F}_{n-1}^T \underline{d}_n + \mathcal{O}(\lambda^2), \ \lambda \to 0$$

ofwel

$$\varphi(\underline{x}_{n-1} + \lambda \underline{d}_n) = \varphi(\underline{x}_{n-1}) - \lambda \underline{d}_n^T G(\underline{x}_{n-1}) \underline{d}_n + \mathcal{O}(\lambda^2) \ ,$$

dat voor voldoende kleine $\lambda > 0$

$$\varphi(\underline{x}_{n-1} + \lambda \underline{d}_n) < \varphi(\underline{x}_{n-1}) \ . \tag{23}$$

Dus dan is de methode van Newton een descent methode. Maar de waarde $\lambda_n = 1$ correspondeert in het algemeen niet met lijnminimalisering en het kan zelfs zijn dat $\varphi(\underline{x}_{n-1} + \underline{d}_n) \geq \varphi(\underline{x}_{n-1})$. In een algorithme, gebaseerd op de methode van Newton, kan men daarom in plaats van $\lambda_n = 1$ te nemen, de waarde van λ_n beter bepalen met lijnminimalisering. Daardoor wordt globaal de convergentie verbeterd, terwijl de locale kwadratische convergentie gehandhaafd blijft. De kosten van een nauwkeurige lijnminimalisering (aantal berekende functiewaarden) zijn echter in veel gevallen zo groot, dat de algorithme daardoor niet efficiënter wordt.

Het grote voordeel van de methode van Newton is de snelle, namelijk locaal kwadratische, convergentie als $\varphi(\underline{x})$ voldoende vaak differentieerbaar is. Daartegenover staat dat in iedere stap de k componenten van de gradiënt (eerste afgeleiden) en $\frac{1}{2}k(k+1)$ elementen van de Hessiaan (tweede afgeleiden) moeten worden bepaald. Een ander nadeel is dat $G(\underline{x}_{n-1})$ niet positief definiet behoeft te zijn (als $\underline{x}_{n-1}$ ver af ligt van $\underline{\alpha}$), zodat het mogelijk is dat de zoekrichting $\underline{d}_n$ niet bestaat (namelijk als $G(\underline{x}_{n-1})$ singulier is) of dat niet voldaan is aan (23).

Het bezwaar van het moeten berekenen van veel afgeleiden kan men ondervangen door de matrix $G(\underline{x}_{n-1})$ te vervangen door een benadering gebaseerd op numerieke differentiatie of, zoals bij de secant methode, door lineaire interpolatie. Men spreekt in deze gevallen van quasi-Newton methoden, bijvoorbeeld de methode van Davidon-Fletcher-Powell.

Een modificatie van de methode van Newton, waarbij de zoekrichting steeds bestaat, wordt verkregen door de matrix $G(\underline{x}_{n-1})$ te vervangen door $G(\underline{x}_{n-1}) + \mu_n I$, met geschikt gekozen positieve μ_n, zodanig dat $G(\underline{x}_{n-1}) + \mu_n I$ positief definiet is.

In de praktijk blijkt dat de meeste minimaliseringsproblemen afkomstig zijn van problemen die bestaan uit curve-fitting, parameterschatting of overbepaalde stelsels niet-lineaire vergelijkingen en die worden opgelost in de zin van de kleinste kwadraten.
In al deze gevallen is het probleem als volgt te formuleren. Er zijn m functies $f_j(\underline{x})$, $j = 1,2,\ldots,m$, $\underline{x} \in \mathbb{R}^k$ en de te minimaliseren functie φ is van de vorm

$$\varphi(\underline{x}) = \sum_{j=1}^{m} [f_j(\underline{x})]^2 . \qquad (24)$$

Voorbeeld. Van een fysische grootheid c(t) wordt aangenomen dat bij benadering geldt

$$c(t) = \alpha_1 e^{-\beta_1 t} + \alpha_2 e^{-\beta_2 t} . \qquad (25)$$

c(t) wordt gemeten in de punten $t_1, t_2, \ldots, t_m$.
De parameters $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ waarmee (25) de beste benadering is in de zin van de kleinste kwadraten zijn dan die waarden waarvoor

$$\sum_{j=1}^{m} \left(c(t_j) - \alpha_1 e^{-\beta_1 t_j} - \alpha_2 e^{-\beta_2 t_j} \right)^2$$

minimaal is.

Dus in bovenstaande formulering is $\underline{x} = (\alpha_1, \alpha_2, \beta_1, \beta_2)$ en

$f_j(\underline{x}) = c(t_j) - \alpha_1 e^{-\beta_1 t_j} - \alpha_2 e^{-\beta_2 t_j}$.

Er bestaan efficiënte algorithmen voor het minimaliseren van een functie φ van de vorm (24), bijvoorbeeld de methode van Gauss-Newton en de methode van Marquardt. Deze methoden zullen in hoofdstuk 7 (d.i. in het college Numerieke Methoden II) worden besproken.

2. Numerieke differentiatie en integratie

In dit hoofdstuk, en in de volgende hoofdstukken, zullen we problemen bespreken uit de numerieke analyse: differentiatie, integratie, oplossen van gewone en partiële differentiaalvergelijkingen. Bij al deze problemen hebben we te maken met de afgeleide(n) of de integraal van een bekende of onbekende functie gedefinieerd op een interval, c.q. een meerdimensionaal gebied. Deze grootheden zijn in de analyse gedefinieerd door een limietproces en zijn daarom in het algemeen niet numeriek, dat wil zeggen in eindig veel bewerkingen, te bepalen.

Voor het oplossen van deze problemen worden afgeleide(n) of integraal vervangen door berekenbare benaderingen, bijvoorbeeld gebaseerd op functiewaarden in een eindig aantal punten. In dit geval spreekt men van discretisatie.
Als de wijze waarop de fout in deze benaderingen afhangt van de afstand tussen de gekozen punten (de zogenaamde stapgrootte) bekend is, dan is het in veel gevallen mogelijk van de fout in de berekende oplossing van het probleem een concrete schatting te geven met behulp van de resultaten van de berekeningen.

Opmerking. Onder een bekende (gegeven) functie verstaan we het volgende: er is mathematisch een functie f gedefinieerd en er is een algorithme gegeven die in een aangeboden punt x van het definitiegebied van f de waarde f(x) berekent. Soms zal de algorithme bestaan uit een formule, soms is hij veel ingewikkelder, bijvoorbeeld als f(x) de waarde is van de oplossing van een vergelijking waarin x als parameter voorkomt.
(In werkelijkheid zal de algorithme niet f(x) maar een benadering voor f(x) afleveren.)

2.1. Foutenanalyse en extrapolatie bij numerieke differentiatie

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 de functie $\delta = \delta(\varepsilon)$ niet.

a) 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 de stapgrootte h berekenen en "kijken" of deze rij getallen nadert tot een limietwaarde.

b) 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

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

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

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

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

dat bij benadering geldt

$$R(\tfrac{1}{2}h) = Df(\tfrac{1}{2}h) - Df(h). \qquad (5)$$

Het rechterlid van (5) 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 goede benadering voor de afbreekfout $R(\frac{1}{2}h)$ (en ook voor R(h), maar dat is minder interessant).

c) Als f(x) voor $|x| \leq h$ een convergente Taylorreeks heeft dan kunnen we de afbreekfout R(h) nader specificeren, want dan geldt

$$R(h) = c_1 h + c_2 h^2 + c_3 h^3 + \ldots \qquad (6)$$

De waarden van c_1, c_2, c_3, ... zijn in het algemeen onbekend (ook al zijn ze theoretisch uit te drukken in f"(0), f'"(0), etc.).
Uit (4) en (6) volgt nu

$$Df(\tfrac{1}{2}h) - Df(h) = \tfrac{1}{2}c_1 h + \frac{3}{4}c_2 h^2 + \frac{7}{8}c_3 h^3 + \ldots \qquad (7)$$

Anderzijds is

$$R(\tfrac{1}{2}h) = \tfrac{1}{2}c_1 h + \tfrac{1}{4}c_2 h^2 + \frac{1}{8}c_3 h^3 + \ldots \qquad (8)$$

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

$$D_1f(\tfrac{1}{2}h) := Df(\tfrac{1}{2}h) + (Df(\tfrac{1}{2}h) - Df(h)) \qquad (9)$$

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

Uit (7), (8) en (9) volgt dat we, naar analogie van (2), kunnen schrijven

$$f'(0) = D_1f(\tfrac{1}{2}h) + R_1(\tfrac{1}{2}h)\ , \qquad (10)$$

met

$$R_1(\tfrac{1}{2}h) = R(\tfrac{1}{2}h) - (Df(\tfrac{1}{2}h) - Df(h)) = -\tfrac{1}{2}c_2h^2 + \ldots \qquad (11)$$

Als $|c_1| >> |c_2\ h|$, dan is $D_1f(\frac{1}{2}h)$ dus inderdaad een betere benadering dan $Df(\frac{1}{2}h)$; 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 $D_1f(\frac{1}{2}h)$ in plaats van $Df(\frac{1}{2}h)$ schaadt dan niet!

In een aantal gevallen weten we uit bekende eigenschappen van f dat $c_1 = 0$.
Uit (7) en (8) blijkt dat we dan beter $\frac{1}{3}(Df(\frac{1}{2}h) - Df(h))$ kunnen nemen als benadering voor $R(\frac{1}{2}h)$, onafhankelijk van wat c_2 is.
Als daarenboven $|c_2| >> |c_3\ h|$, dan is $\frac{1}{3}(Df(\frac{1}{2}h) - Df(h))$ een goede benadering voor $R(\frac{1}{2}h)$, en dan levert

$$D_1f(\tfrac{1}{2}h) := Df(\tfrac{1}{2}h) + \tfrac{1}{3}(Df(\tfrac{1}{2}h) - Df(h)) \qquad (12)$$

een betere benadering voor $f'(0)$.
In dit geval geldt namelijk

$$f'(0) = D_1f(\tfrac{1}{2}h) + R_1(\tfrac{1}{2}h) \qquad (13)$$

met

$$R_1(\tfrac{1}{2}h) = -\tfrac{1}{6}c_3h^3 + \ldots \qquad (14)$$

Hieruit blijkt ook weer dat het gebruik van $D_1f(\frac{1}{2}h)$ geen kwaad kan als c_2 klein is, dwz. als niet $|c_2| >> |c_3\ h|$.

Men noemt de methode om, uitgaande van formule (6) voor de restterm $R(h)$, uit twee waarden $Df(h)$ en $Df(\frac{1}{2}h)$ een betere benadering $D_1f(\frac{1}{2}h)$ met formule (9) te bepalen h-extrapolatie volgens Richardson. Als we formule (12) gebruiken omdat we weten dat $c_1 = 0$, dan spreken we van h^2-extrapolatie. (We gebruiken dan immers het feit dat de restterm begint met h^2.)

2.1.1. Extrapolatie in het algemeen ([2],p. 269-273; [13], p. 115-118)

De hierboven beschreven extrapolatie methode is algemeen toepasbaar op problemen waarbij van een te bepalen grootheid, die we F zullen noemen, een berekenbare benadering G(h), bijvoorbeeld verkregen door discretisatie met stapgrootte h, bestaat die een afbreekfout R(h) heeft waarvoor geldt

$$F = G(h) + R(h) \tag{15}$$

$$R(h) = c_1 h + c_2 h^2 + c_3 h^3 + \ldots \tag{16}$$

Uit (16) volgt dat bij benadering geldt

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

Als we, uitgaande van een stapgrootte h, de rij stapgrootten

$$h_j := h/2^j,\ j = 0,1,2,\ldots$$

nemen, dan krijgen we door h-extrapolatie, zie (9),

$$G_1(h_j) := G(h_j) + (G(h_j) - G(h_{j-1})).$$

Er geldt dan

$$F = G_1(h) + R_1(h)$$

$$R_1(h) = c_2' h^2 + c_3' h^3 + \ldots$$

Omdat we nu het geval hebben dat in de formule van de restterm $c_1' = 0$, geldt bij benadering

$$R_1(\tfrac{1}{2}h) = \tfrac{1}{3}(G_1(\tfrac{1}{2}h) - G_1(h)).$$

Op de rij benaderingen $G_1(h_j)$ kunnen we dus vervolgens h^2-extrapolatie toepassen, zie (12),

$$G_2(h_j) := G_1(h_j) + \tfrac{1}{3}(G_1(h_j) - G_1(h_{j-1})),$$

en dan geldt

$$F = G_2(h) + R_2(h)$$

$$R_2(h) = c_3'' h^3 + c_4'' h^4 + \ldots$$

Nu is $c_1'' = 0$ en $c_2'' = 0$ en dus geldt bij benadering (vergelijk (7) en (8)

$$R_2(\tfrac{1}{2}h) = \tfrac{1}{7}(G_2(\tfrac{1}{2}h) - G_2(h)).$$

Op de rij benaderingen $G_2(h_j)$ kunnen we vervolgens h^3-extrapolatie toepassen, etc.

Samenvattend kunnen we de extrapolatie methode als volgt beschrijven.
Als gegeven is dat

$$F = G(h) + R(h),$$

waarbij bekend is dat de afbreekfout R(h) de volgende reeksontwikkeling heeft (met een bekend natuurlijk getal p en onbekende coefficienten $c_p \neq 0$, $c_{p+1}, \ldots$)

$$R(h) = c_p\, h^p + c_{p+1}\, h^{p+1} + \ldots, \tag{17}$$

dan is, als $|c_p| >> |c_{p+1}\, h|$, bij benadering

$$R(h) = 2^p\, R(\tfrac{1}{2}h). \tag{18}$$

Substitutie hiervan in (4) levert dat dan bij benadering

$$R(\tfrac{1}{2}h) = \frac{1}{2^p - 1}\,(G(\tfrac{1}{2}h) - G(h)) \tag{19}$$

en hieruit volgt dat de extrapolatie formule

$$G_1(\tfrac{1}{2}h) := G(\tfrac{1}{2}h) + \frac{1}{2^p - 1}\,(G(\tfrac{1}{2}h) - G(h)) \tag{20}$$

een betere benadering geeft voor F.

Opgave. Ga na dat geldt

$$F = G_1(h) + R_1(h)$$

met

$$R_1(h) = c'_{p+1}\, h^{p+1} + \ldots,$$

waarbij c'_{p+1} in orde van grootte gelijk is aan c_{p+1}.
Hieruit volgt dan weer dat $G_1(h)$, verkregen door h^p-extrapolatie, in het geval dat $c_p = 0$ geen slechtere benadering is dan G(h).

De formules (19) en (20) zijn gebaseerd op de benaderingsformule (18). Deze formule gaat ervan uit dat

$$|c_p| >> |c_{p+1}\, h|. \tag{21}$$

Om in een practische situatie een indruk te krijgen in hoeverre (21) geldt, beschouwen we het quotient van opvolgende differenties:

$$\frac{G(\frac{1}{2}h) - G(h)}{G(\frac{1}{4}h) - G(\frac{1}{2}h)} = \frac{R(h) - R(\frac{1}{2}h)}{R(\frac{1}{2}h) - R(\frac{1}{4}h)} = 2^p \frac{c_p + c^*_{p+1} h + \ldots}{c_p + \frac{1}{2}c^*_{p+1} h + \ldots} . \qquad (22)$$

Hierin is $c^*_{p+1} = \frac{2^{p+1}-1}{2(2^p-1)} c_{p+1}$, dus c^*_{p+1} is van dezelfde grootte orde als c_{p+1}.

Als de waarde van het linkerlid dicht bij 2^p ligt, dan nemen we aan dat (21) geldt en dat derhalve (19) een goede schatting voor de fout geeft en de extrapolatie formule (20) inderdaad een betere benadering levert voor F.

Opmerking. Als p niet a priori bekend is, dan kunnen we met (22) de waarde van p experimenteel bepalen.

Voorbeeld.

In onderstaande tabel zijn de bij een gegeven functie horende waarden van $Df(h_j)$ voor een aantal waarden van j gegeven. Daarnaast staan de differenties $VDf(h_j) := Df(h_j) - Df(h_{j-1})$. De laatste waarden in deze kolom suggereren dat de fout in $Df(h_4)$ van de orde van 0.04 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 (9) berekende waarden van $D_1f(h_j)$. Deze lijken al veel beter te convergeren. De bijbehorende differentiekolom suggereert dat $D_1f(h_4)$ een fout van de orde 0.0003 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 $D_2f(h_j)$ levert. Het resultaat is fraai te noemen (de exacte waarde van $f'(0) = 0.874326$). Dat het laatste cijfer in $D_2f(h_j)$ onregelmatig verloopt is een gevolg van afrondingsfouten: de functiewaarden $f(0)$ en $f(h_j)$ zijn alle *in 6 decimalen bepaald, maar bij het uitrekenen van* $Df(h_j)$ *verliest men nauw*keurigheid en meer naarmate h kleiner is.

j	h_j	$Df(h_j)$	$VDf(h_j)$	$D_1f(h_j)$	$VD_1f(h_j)$	$D_2f(h_j)$
0	.2	1.53967				
1	.1	1.19690	- .34277	0.85413		
2	.05	1.03308	- .16382	0.86926	0.01513	0.87431
3	.025	0.95308	- .08000	0.87308	0.00382	0.87435
4	.0125	0.91352	- .03956	0.87396	0.00088	0.87425

De resultaten van achtereenvolgende extrapolaties kunnen we in een zogenaamd Romberg schema noteren.

We veronderstellen weer dat voor de afbreekfout R(h) de reeksontwikkeling (16) geldt. We voeren nog de volgende notatie in

$$G_j^k := G_k(h_j)$$

waarbij $G_j^0 := G(h_j)$ genomen wordt. Dan is de algemene extrapolatie formule

$$G_j^{k+1} := G_j^k + \frac{1}{2^{k+1}-1} (G_j^k - G_{j-1}^k)$$

en het Romberg schema

$$\begin{array}{llll} h_0 & G_0^0 & & \\ h_1 & G_1^0 & G_1^1 & \\ h_2 & G_2^0 & G_2^1 & G_2^2 \\ h_3 & G_3^0 & G_3^1 & G_3^2 \quad G_3^3 \\ & \vdots & \vdots & \vdots \quad \vdots \end{array}$$

Opmerking. In een aantal gevallen heeft de afbreekfout R(h) een reeksontwikkeling in h^2, dus

$$R(h) = c_1 h^2 + c_2 h^4 + \ldots$$

Dan wordt achtereenvolgens h^2-extrapolatie, h^4-extrapolatie, etc. toegepast. De algemene extrapolatie formule is dan

$$G_j^{k+1} := G_j^k + \frac{1}{4^{k+1}-1} (G_j^k - G_{j-1}^k) .$$

(Ga dit na.)

2.1.2. Enkele formules voor numerieke differentiatie

De in 2.1. geanalyseerde formule voor numerieke differentiatie is erg eenvoudig maar weinig nauwkeurig. Beter is de zg. centrale formule

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

Voor driemaal continu differentieerbare functies f geldt weer een uitspraak van de vorm

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

met

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

terwijl 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:

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

Daarna eventueel h^4-extrapolatie, etc.

Een nog nauwkeuriger formule dan (1) is *)

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

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

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

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

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

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

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

$$f''(0) = \frac{2f(0) - 5f(-h) + 4f(-2h) - f(-3h)}{h^2} + \frac{11}{12} h^2 f^{(4)}(\xi) . \qquad (6)$$

---

*) Laat zien dat dit de uitwerking van formule (3) is, met h i.p.v. $\frac{1}{2}h$.

2.1.3. De methode van de onbepaalde coefficienten

We bespreken nu hoe men in het algemeen formules als boven kan reconstrueren. Als voorbeeld nemen we formule (5) uit 2.1.2.
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 $R(h) \neq 0$ voor $f(x) = x^3$, nl.

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

b) Veronderstel 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 alle 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) dat $p = 2$ 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 (2) geschreven kan worden - in veel gevallen is dit zo).

Opgave. Leid ook andere formules uit 2.1.2. op bovenstaande wijze af.

2.1.4. De invloed van afrondfouten

We kijken tenslotte naar de invloed van afrondfouten in de functiewaarden van f. Neem bv. formule (6) uit 2.1.2. en veronderstel dat we in plaats van met f(0),f(-h),..., werken met $f(0) + \varepsilon_0$, $f(-h) + \varepsilon_1$,... . Dit geeft in

de berekende 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).

2.2. Numerieke integratie ([3] ch.2,6)

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

$$\int_{x_j}^{x_{j+1}} f(x)dx = \frac{1}{2} h(f_j + f_{j+1}) - \frac{1}{12} h^2 f''(\xi_j)(x_{j+1} - x_j) \qquad (3)$$

(trapeziumregel)

met

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

levert

$$\int_a^b f(x)dx = h(\tfrac{1}{2}f_0 + f_1 + \ldots + f_{N-1} + \tfrac{1}{2}f_N) - \frac{(b - a)}{12} h^2 f''(\xi) \qquad (4)$$

(samengestelde trapeziumregel)

Een formule als (3) kan men weer met de methode van 2.1.3. 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)

$$R(h) = -\frac{1}{12} h^2 \sum_{j=0}^{N-1} f''(\xi_j)(x_{j+1}-x_j), \qquad (5)$$

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

$$\sum_{j=0}^{N-1} f''(\xi_j)\,(x_{j+1}-x_j) = (b-a)f''(\xi).$$

Hieruit volgt voor de afbreekfout

$$R(h) = -\frac{b-a}{12} h^2 f''(\xi).$$

Formules analoog aan (3) en (4) zijn

$$\int_{x_j}^{x_{j+1}} f(x)dx = hf_{j+\frac{1}{2}} + \frac{1}{24} h^2 f''(\xi_j)\,(x_{j+1}-x_j)$$

(midpointregel);

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

(samengestelde midpointregel).

En

$$\int_{x_{2j}}^{x_{2j+2}} f(x)dx = \frac{1}{3} h(f_{2j} + 4f_{2j+1} + f_{2j+2}) - \frac{1}{180} h^4 f^{(4)}(\xi_j)\,(x_{2j+2}-x_{2j})$$

(regel van Simpson);

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

(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

$$R(\tfrac{1}{2}h) = -\frac{1}{12}(\tfrac{1}{2}h)^2 \sum_{j=0}^{N-1} (f''(\eta_j)(x_{j+\frac{1}{2}}-x_j) + f''(\eta_{j+\frac{1}{2}})(x_{j+1}-x_{j+\frac{1}{2}}))$$

met $x_j < \eta_j < x_{j+\frac{1}{2}} < \eta_{j+\frac{1}{2}} < x_{j+1}$.

Duiden we de trapeziumregel aan met

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

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 benadering 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 $R(\frac{1}{2}h) \sim \frac{1}{16}R(h)$).

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 $c_2 = 0$).

Overeenkomstig 2.1.1. volgt hieruit dat voor de trapeziumregel 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 schema van Romberg. Men kan bewijzen dat de kolommen in schema ook voor minder gladde functies convergeren naar de integraal, zij het niet zo snel als voor gladde functies.

Opmerking. 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).

2.2.1. Praktische numerieke integratie

De minimale controle die men bij praktische numerieke integratie moet uitvoeren is, de integraal met de gekozen integratieformule voor tenminste twee 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. Bij integratie met uniforme stap wordt de totale fout dan hoofdzakelijk bepaald door de bijdrage van het gedeelte van (a,b) waar de functie minder glad is. Men kan ook zeggen dat de integraal in een gedeelte van (a,b) veel te nauwkeurig wordt berekend. Men wenst daarom de waarde van de stapgrootte aan te passen aan de gladheid van de functie en wel zodanig dat de totale fout min of meer uniform verdeeld wordt over (a,b).
Zij gevraagd te berekenen $\int_a^b f(x)dx$ met een absolute fout kleiner dan ε. Dan is de toegelaten fout per lengte-eenheid $\varepsilon/(b - a)$. We willen daarom het interval (a,b) verdelen in deelintervallen $(x_0,x_1),(x_1,x_2),\ldots,(x_{N-1},x_N)$ zodat de fout in het interval (x_j,x_{j+1}) kleiner is (maar liefst niet veel) dan $(x_{j+1} - x_j)\varepsilon/(b - a)$.
Voor functies met een enigszins gelijkmatig (maar wel aanzienlijk) variërende gladheid kan de keuze van de stapgrootte als volgt geautomatiseerd worden.

We kiezen als basisformule de trapeziumregel en definiëren voor het interval $(x,x+h)$ de enkelvoudige trapeziumsom

$$T_1(x,h) := \tfrac{1}{2}h(f(x) + f(x+h)) \ .$$

Zij nu $T_2(x,h)$ de benadering van de integraal over $(x,x+h)$ met de trapeziumsom bij verdeling in twee deelintervallen, dus

$$T_2(x,h) := T_1(x,\tfrac{1}{2}h) + T_1(x+\tfrac{1}{2}h,\tfrac{1}{2}h) \ ,$$

dan wordt een goede benadering van de fout in $T_2(x,h)$ gegeven door

$$Rf(x,h) := \frac{1}{3}(T_2(x,h) - T_1(x,h)) \ . \qquad (1)$$

We wensen nu dat het interval (a,b) in deelintervallen (x_j,x_{j+1}), $j = 0,1,\ldots,N-1$, wordt gesplitst, zodat, met $h_j := x_{j+1} - x_j$,

$$|Rf(x_j,h_j)| \le tol(h_j) := \frac{h_j\varepsilon}{b-a} \qquad (2)$$

maar niet essentieel kleiner. Er geldt dan stellig

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

en dus is bij benadering de fout in de berekende waarde

$$\sum_{j=0}^{N-1} T_2(x_j,h_j)$$

kleiner dan ε.

Voor "nette" functies f is $Rf(x_j,h_j)$ ongeveer evenredig met h_j^3. Daar het rechterlid van (2) 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

$$h_j^* = \left(\frac{tol(h_j)}{|Rf(x_j,h_j)|}\right)^{\frac{1}{2}} \times h_j \ . \qquad (3)$$

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 (3), 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 $I := I + T_2(x_j,h_j)$, $x_{j+1} := x_j + h_j$.

d) Als $x_{j+1} < b$, dan bepalen we h_j^* uit (3) 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.

Opmerkingen.

1. In de practijk zal men in plaats van $T_2(x,h)$ de geëxtrapoleerde waarde $T_2(x,h) + Rf(x,h)$ als benadering voor de integraal van x tot x + h nemen.
2. In de hierboven behandelde methode kan men in plaats van de trapeziumregel een willekeurige andere integratieformule als basisformule nemen en de formules (1) en (3) daarbij aanpassen.

Men noemt een methode waarbij de stapgrootte aangepast wordt aan de gladheid van de integrand een adaptieve methode. Naast methoden met zelfzoekende stap, zoals hierboven beschreven, bestaan er methoden die gebaseerd zijn op recursief gebruik van een elementair integratieproces.*)

*) Zie bijvoorbeeld RC Informatie PP 3.1.1. Integratie van enkelvoudige integralen met algemene integrand. Technische Hogeschool Eindhoven.

2.2.2. Enkele andere integratiemethoden

Bij de tot nu toe behandelde elementaire integratieformules (trapeziumregel, midpointregel, regel van Simpson) zijn de punten x_j vooraf gegeven en zijn de coefficiënten c_j bepaald door de eis dat de integratieformule exact is, dwz. dat $R = 0$, voor polynomen van zo hoog mogelijke graad.

Men kan echter ook vragen om zowel de coefficiënten c_j als de punten x_j in de integratieformule zodanig te bepalen dat de integratieformule exact is voor polynomen van zo hoog mogelijke graad. Een dergelijke integratieformule wordt een integratieformule van Gauss genoemd.
Met hetzelfde aantal steunpunten x_j zal bij een integratieformule van Gauss de bereikbare graad, en daarmee ook de orde van de methode, ongeveer twee maal zo hoog zijn dan bij een integratieformule met vooraf gegeven (equidistante) punten x_j.

Voorbeelden van integratieformules van Gauss zijn:

1) De tweepuntsformule

$$\int_{-h}^{h} f(x)dx = h\{f(-\frac{h}{\sqrt{3}}) + f(\frac{h}{\sqrt{3}})\} + R_2$$

Als $f(x)$ een viermaal continu differentieerbare functie is, dan is

$$R_2 = \frac{h^5}{135} f^{(4)}(\xi)$$

2) De driepuntsformule

$$\int_{-h}^{h} f(x)dx = \frac{h}{9}\{5f(-h\sqrt{\frac{3}{5}}) + 8f(0) + 5f(h\sqrt{\frac{3}{5}})\} + R_3$$

Als $f(x)$ een zesmaal continu differentieerbare functie is, dan is

$$R_3 = \frac{h^7}{15750} f^{(6)}(\xi)$$

Voor de k-puntsformule van Gauss

$$\int_a^b f(x)dx = \sum_{j=1}^{k} c_j f(x_j) + R_k \qquad (1)$$

vermelden we de volgende resultaten.

1. Voor iedere k bestaan er getallen c_j en x_j, $j = 1,2,\ldots,k$, zodanig dat (1) exact is, dwz. dat $R_k = 0$, voor alle polynomen van de graad $\leq 2k - 1$.
2. Voor iedere j geldt $x_j \in (a,b)$ en $c_j > 0$.
3. Als $E_k(x) := \prod_{j=1}^{k} (x - x_j)$, dan zijn de punten x_j geheel bepaald door de voorwaarde dat $\int_a^b p(x)\, E_k(x)dx = 0$ voor ieder polynoom p(x) van de graad $< k$.
4. Als f(x) 2k maal continu differentieerbaar is, dan geldt voor de afbreekfout de formule

$$R_k = \frac{(b-a)^{2k+1}(k!)^4}{(2k+1)((2k)!)^3} f^{(2k)}(\xi) \ .$$

Hieruit volgt dat als de lengte van het integratie interval een vast veelvoud is van een stapgrootte h, dus $b - a = mh$

$$R_k = C\, h^{2k+1}\, f^{(2k)}(\xi) \ . \qquad (2)$$

Als de integrand minder dan 2k maal differentieerbaar is, dan is de orde van de afbreekfout kleiner dan $2k + 1$.
Bijvoorbeeld voor de tweepuntsformule kunnen we het volgende bewijzen.
Als f(x) tweemaal continu differentieerbaar is, dan voldoet R_2 aan de volgende ongelijkheid

$$|R_2| \leq C\, h^3 \max_{|x|\leq h} |f''(x)| \ , \qquad (3)$$

waarbij $C \doteq 0.1$.
Vergelijken we deze formule met de restterm van de trapeziumregel $-\frac{1}{12}h^3 f''(\xi)$, dan zien we dat voor een tweemaal continu differentieerbare integrand de tweepuntsformule van Gauss niet slechter is dan de trapeziumregel. Iets dergelijks geldt ook voor de driepuntsformule en de regel van Simpson.

Met behulp van een k-puntsformule van Gauss als elementaire formule kan weer een samengestelde formule worden gemaakt.

Gebruiken we bijvoorbeeld de tweepuntsformule, dan krijgen we

$$\int_a^b f(x)dx = \sum_{i=1}^{N} \int_{x_{2i-2}}^{x_{2i}} f(x)dx =$$

$$= h \sum_{i=1}^{N} \{f(x_{2i-1} - \frac{h}{\sqrt{3}}) + f(x_{2i-1} + \frac{h}{\sqrt{3}})\} + R$$

Hierin is $2h \cdot N = b - a$ en $x_i = a + ih$.
Voor de afbreekfout R geldt de formule

$$R = \sum_{i=1}^{N} \frac{h^5}{135} f^{(4)}(\xi_i),$$

met $x_{2i-2} < \xi_i < x_{2i}$. Hieruit volgt

$$R = \frac{b-a}{270} h^4 f^{(4)}(\xi) \quad , \quad a < \xi < b \ .$$

Voor de berekening van bepaalde typen van oneigenlijke integralen bestaan speciale integratieformules van Gauss. Hierbij wordt de integrand beschouwd als het product van een vaste functie w(x) die continu en positief is, en een willekeurige functie f(x). De functie w(x), ook wel gewichtsfunctie genoemd, beschrijft een speciaal gedrag van de integrand, bijvoorbeeld exponentieel gedrag voor grote x : $w(x) = e^{-x}$ of e^{-x^2} ; singulier gedrag in de eindpunten: $w(x) = (x - a)^{\alpha}(b - x)^{\beta}$ met $\alpha > -1$ en $\beta > -1$.
De k-punts integratieformule heeft dan de gedaante

$$\int_a^b w(x)\, f(x)dx = \sum_{j=1}^{k} c_j\, f(x_j) + R_k, \qquad (4)$$

waarbij mogelijk $a = -\infty$ en/of $b = \infty$.
De coefficiënten c_j en de punten x_j worden nu zodanig bepaald dat de formule exact is, voor het geval dat f(x) een willekeurige polynoom is van zo hoog mogelijke graad.

Voor (4) gelden analoge resultaten als op pag. 2.16 geformuleerd voor (1). Met name geldt voor de afbreekfout, in het geval dat f(x) 2k maal continu differentieerbaar is, een formule van de vorm

$$R_k = d_k f^{(2k)}(\xi)$$

waarbij $d_k \to 0$ als $k \to \infty$.

Voorbeelden

	$[a,b]$	$w(x)$	d_k
Gauss - Laguerre	$[0,\infty)$	e^{-x}	$\frac{(k!)^2}{(2k)!}$
Gauss - Hermite	$(-\infty,\infty)$	e^{-x^2}	$\frac{k!\sqrt{\pi}}{2^k(2k)!}$
Gauss - Chebyshev	$[-1,1]$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{2\pi}{2^{2k}(2k)!}$

2.2.3. Conditie en numerieke stabiliteit

Om na te gaan wat de conditie is van

$$I := \int_a^b f(x)dx, \tag{1}$$

vervangen we $f(x)$ door $f(x) + \delta f(x)$. Daardoor verandert I met een bedrag δI en er geldt

$$I + \delta I = \int_a^b (f(x) + \delta f(x))dx.$$

Dus

$$\delta I = \int_a^b \delta f(x)dx.$$

Bij een relatieve verandering van ten hoogste α van de integrand, dus $|\delta f(x)| \leq \alpha|f(x)|$, geldt voor de relatieve verandering van de integraal

$$\frac{|\delta I|}{|I|} \leq \alpha \frac{\int_a^b |f(x)|dx}{\left|\int_a^b f(x)dx\right|}. \tag{2}$$

Dit betekent dat

$$c(I) := \frac{\int_a^b |f(x)|dx}{\left|\int_a^b f(x)dx\right|}$$

het conditiegetal is van (1), en dat het probleem dus goed geconditioneerd is, als $\left|\int_a^b f(x)\right|dx$ niet essentieel kleiner is dan $\int_a^b |f(x)|dx$.

Uit bovenstaande volgt voor de onvermijdbare fout $\Delta^o I$ de schatting

$$\frac{|\Delta^o I|}{|I|} \leq (c(I)+1)\eta \tag{3}$$

We beschouwen nu de numerieke berekening van (1) met behulp van de integratieformule

$$\int_a^b f(x)dx = \sum_{j=0}^{N} c_j f(x_j) + R_N.$$

Voor voldoend grote N wordt de fout in de benadering voornamelijk bepaald door de fout in de som

$$I_N := \sum_{j=0}^{N} c_j f(x_j).$$

Voor de berekende waarde $\bar{I}_N$ geldt dan (zie voorbeeld pag. 0.13)

$$\bar{I}_N = \sum_{j=0}^{N} c_j f(x_j)(1+\delta_j), \tag{4}$$

waarin

$$|\delta_j| \leq (N+1)\eta, \quad j = 0,1,2,\ldots,N . \tag{5}$$

Voor de totale rekenfout δI_n geldt dus

$$|\delta I_n| \leq \sum_{j=0}^{N} |c_j||f_j| . (N+1)\eta. \tag{6}$$

Als $c_j > 0$ voor alle j, dan is $\sum_{j=0}^{N} |c_j||f_j| = \sum_{j=0}^{N} c_j|f_j|$ een goede benadering voor $\int_a^b |f(x)|dx$ en dan vinden we door vergelijking van (3) en (6) dat de integratieformule <u>numeriek stabiel</u> is.

Als niet geldt $c_j > 0$ voor alle j, dan kan $\sum_{j=0}^{N}|c_j||f_j|$ essentieel groter zijn dan $\int_a^b |f(x)|dx$. In dat geval is de integratieformule <u>numeriek instabiel</u>.

Opmerking De schatting (5) levert in het algemeen een grove bovengrens. In de praktijk zal δ_j niet meer dan enkele malen η bedragen.

Zoals in 2.2.2. is vermeld hebben de elementaire integratieformules van Gauss, en dus ook de daarop gebaseerde samengestelde formules, positieve coefficiënten. Hieruit kunnen we nu concluderen dat elke integratieformule van Gauss numeriek stabiel is.
Ook de integratieformules gebaseerd op de trapeziumregel, de midpointregel, of de regel van Simpson zijn numeriek stabiel. Er zijn echter elementaire integratieformules op equidistante punten waarvan de coefficiënten niet allemaal positief zijn, bijvoorbeeld van een 9-punts formule. De hierop gebaseerde samengestelde formules zijn mogelijk numeriek instabiel.

Bij het schatten van de totale fout in de berekende waarde van de integraal moeten we ook rekening houden met de nauwkeurigheid waarmee de integrand berekend wordt.
Veronderstel dat ε een bovengrens is voor de relatieve nauwkeurigheid waarmee $f(x)$ berekend wordt, dan geeft dat een bijdrage in de fout van $\bar{I}$, die niet groter is dan $\varepsilon \int_a^b |f(x)|\,dx$.

3. Numerieke integratie van differentiaalvergelijkingen([2],ch.8) *)

Beschouw een eerste orde differentiaalvergelijking

$$\frac{dy}{dx} = f(x,y) \ , \ x > x_0, \tag{1}$$

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 in een interval (x_0,x_0+d) 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

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

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 $z_0 = y_0$.

3.1. Enkele eenvoudige methoden

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

*) Zie ook RC-Informatie PP 3.4.1. Integratie van een beginwaardeprobleem voor een gewone differentiaalvergelijking, etc. Technische Hogeschool Eindhoven.

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

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),

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

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

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

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) <u>formules</u> (1), (2) en (3) de volgende <u>methoden</u> 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 tussen deze drie methoden op.

a. Euler en trapeziumregel zijn zogenaamde <u>eenstapsmethoden</u>: we behoeven uit het verleden alleen de waarde van z_n te kennen. Dit houdt o.a. in dat bij iedere volgende stap de stapgrootte opnieuw gekozen kan worden.
De midpoint-regel is een <u>tweestapsmethode</u>: voor de bepaling van z_{n+1} hebben we zowel z_n als z_{n-1} nodig. Dit heeft de volgende consequenties:
- er is een <u>startprocedure</u> nodig, omdat behalve $z_0 = y_0$ ook de waarde van

z_1 nodig is om $z_2, z_3, \ldots$ met de midpoint-regel te kunnen bepalen; z_1 kan bijvoorbeeld met de regel van Euler bepaald worden;

- de stapgrootte h ligt vast, omdat z_{n-1}, d.i. de benadering voor de oplossing in $x_{n-1} = x_n - h$ wordt gebruikt. Verandering van de stapgrootte impliceert opnieuw starten met een startprocedure;
- de theorie van meerstapsmethoden is ingewikkelder dan die van eenstapsmethoden.

b. Euler en midpoint-regel zijn zg. <u>expliciete methoden</u>: de bepaling van z_{n+1} geschiedt door invulling in een formule. De trapeziumregel is een <u>impliciete methode</u>: formule (9) is een vergelijking waaruit z_{n+1} moet worden opgelost. Dit oplossen kan als regel gebeuren met behulp van successieve substitutie:

$$z_{n+1}^{(\ell+1)} := z_n + \tfrac{1}{2}h(f(x_n,z_n) + f(x_{n+1},z_{n+1}^{(\ell)})), \ell = 0,1,2,\ldots ;$$

de asymptotische convergentiefactor is daarbij $\tfrac{1}{2}h\frac{\partial f}{\partial y}(x_{n+1},z_{n+1})$, 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 <u>predictorformule</u>, de impliciete formule, waarmee men itereert heet de <u>correctorformule</u>.

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} (<u>correcting only once</u>). 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. <u>Runge-Kutta</u> type.

3.1.1. Locale en globale afbreekfout

De afbreekfout van een methode is de fout die gemaakt wordt doordat de differentiaalvergelijking door discretisatie wordt vervangen door een recursieformule.
Eerst definiëren we het begrip locale afbreekfout in een punt (x_n,y_n). Zij

$$z_{n+1} := \sum_{j=0}^{k-1} \alpha_j z_{n-j} + hF(x_n,z_{n+1},z_n,\ldots,z_{n-k+1},h) \qquad (11)$$

de algemene formule voor een k-staps methode (expliciet als z_{n+1} niet, impliciet als z_{n+1} wel als argument in F voorkomt) voor het oplossen van de differentiaalvergelijking (1), pag. 3.1.

Zij

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

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

$$R(h,x_n,y_n) := \frac{1}{h}[y_{n+1} - \sum_{j=0}^{k-1} \alpha_j y_{n-j} - hF(x_n,y_{n+1},y_n,\ldots,y_{n-k+1},h)] \ . \qquad (12)$$

Dan heet $R(h,x_n,y_n)$ de locale afbreekfout van de methode (11) in het punt (x_n,y_n), bij stapgrootte h.
Hoe kunnen we formule (12) begrijpen?
Zij bijvoorbeeld $\alpha_0 = 1$, $\alpha_j = 0$ voor $j = 1,2,..,k-1$, dan is

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

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 die we maken door discretisatie van (1).
Voor de meeste methoden bestaat er (als f(x,y) voldoende vaak differentieerbaar is) een schatting van de vorm

$$|R(h,x_n,y_n)| \le Ch^m \qquad (13)$$

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, als m de grootste waarde is waarvoor (13) geldt.

<u>Voorbeelden</u>.
Bij de methode van Euler (7) volgt uit de formules (1) en (4) van pag. 3.2 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

$$|R(h,x_n,y_n)| \leq Ch \ , \tag{14}$$

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

$$R(h,x_n,y_n) = \frac{1}{h}[y_{n+1} - y_{n-1} - 2hy'_n] = \frac{h^2}{3}\, y'''(\xi_n)$$

en voor de trapeziumregel

$$R(h,x_n,y_n) = \frac{1}{h}[y_{n+1} - y_n - \tfrac{1}{2}h(y'_n + y'_{n+1})] = -\frac{1}{12}\, h^2y'''(\xi_n) \ .$$

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

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

de oplossing van de differentiaalvergelijking met deze beginvoorwaarden, en zij $z_0 = y_0$, z_1, z_2, ... verkregen met een integratiemethode met stap h. We willen de <u>globale afbreekfout</u>

$$y_n - z_n = \varphi(x_n,x_0,y_0) - z_n \tag{15}$$

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

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

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

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

Voor de methode van Euler geldt

$$z_{n+1} = z_n + hf(x_n,z_n) \tag{16}$$

terwijl voor de oplossing van het beginwaardeprobleem geldt

$$y_{n+1} = y_n + hf(x_n,y_n) + hR(h,x_n,y_n) , \tag{17}$$

waarin $y_j = \varphi(x_j,x_0,y_0)$ voor $j = n, n+1$.

Zij nu

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

in een relevant gebied. Dan volgt uit (14), (16) en (17)

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

Door volledige inductie vinden we hieruit (daar $y_0 = z_0$)

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

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

$$|y(x_n) - z_n| \leq \frac{e^{nLh} - 1}{L} Ch = \frac{e^{L(x_n - x_0)} - 1}{L} Ch . \tag{18}$$

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 (18) 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 (18) zijn de eenvoudigste in hun soort.

Als f(x,y) voldoende vaak differentieerbaar is, dan kan de globale afbreekfout nader gespecificeerd worden. Dan geldt namelijk het volgende: er zijn van h onafhankelijke functies $c_1(x),c_2(x),\ldots$ zodanig dat

$$y(x_n) - z_n = c_1(x_n)h + c_2(x_n)h^2 + \ldots \quad (19)$$

(bij een methode van de m-de orde is natuurlijk
$c_1(x) \equiv c_2(x) \equiv \ldots \equiv c_{m-1}(x) \equiv 0$ en $c_m(x) \not\equiv 0$).

Op basis van formule (19) kan men weer volgens het principe uit 2.1.1. een schatting van de globale fout krijgen (uit de resultaten met stapgrootten h en $\frac{1}{2}h$) en extrapolaties uitvoeren (met resultaten met stapgrootten h, $\frac{1}{2}h$, $\frac{1}{4}h$, ...).

Voorbeeld.
Beschouw het beginwaardeprobleem

$$y' = xy/(1 + y^2) \ , \ y(0) = 1 \ . \quad (20)$$

De oplossing kan geschreven worden in de vorm van een impliciete vergelijking

$$y^2 = \exp(x^2 - y^2 + 1) \ .$$

Als we (20) oplossen met de methode van Euler voor verschillende waarden van h en daarna extrapoleren, dan krijgen we voor het punt x = 1 de volgende tabel.

$n = \frac{1}{h}$	z_n	∇z_n	$z_n^{(1)}$	$\nabla z_n^{(1)}$	$z_n^{(2)}$
1	1.000000				
2	1.125000	0.125	1.250000		
4	1.187095	$0.621_{10}{}^{-1}$	1.249191	$-0.809_{10}{}^{-3}$	1.248921
8	1.217693	$0.306_{10}{}^{-1}$	1.248290	$-0.901_{10}{}^{-3}$	1.247990
16	1.232834	$0.151_{10}{}^{-1}$	1.247976	$-0.314_{10}{}^{-3}$	1.247872
32	1.240361	$0.753_{10}{}^{-2}$	1.247888	$-0.885_{10}{}^{-4}$	1.247858
64	1.244113	$0.375_{10}{}^{-2}$	1.247864	$-0.233_{10}{}^{-4}$	1.247857

Uit de kolommen voor ∇z_n en $\nabla z_n^{(1)}$ volgt dat de globale afbreekfout inderdaad van de vorm (19) is.

Uit de tabel kunnen we concluderen dat y(1) = 1.247857 met een fout die wezenlijk kleiner is dan $\frac{1}{3} \times 0.233_{10}{}^{-4} \simeq 0.8_{10}{}^{-5}$.

3.2. Methoden van hogere orde.

In deze paragraaf zullen we twee klassen van methoden voor het oplossen van het beginwaardeprobleem

$$y' = f(x,y), \ x > x_0$$
$$y(x_0) = y_0 \quad (1)$$

bespreken, namelijk de lineaire meerstapsmethoden en de expliciete eenstapsmethoden van het Runge-Kutta type. Beide klassen kunnen worden beschouwd als een bijzonder geval van de algemene k-stapsmethode, formule (11) van 3.1.1.

3.2.1. Lineaire meerstapsmethoden.

De algemene lineaire k-stapsmethode is van de vorm

$$z_{n+1} := \sum_{j=0}^{k-1} \alpha_j z_{n-j} + h \sum_{j=-1}^{k-1} \beta_j f(x_{n-j}, z_{n-j}) \quad . \tag{2}$$

We zien hieruit dat z_{n+1} afhangt van de berekende waarden van de oplossing in de voorgaande punten x_{n-k+1} t/m x_n. Als $\beta_{-1} = 0$, dan is het rechterlid van (2) alleen afhankelijk van de voorgaande punten en dan noemen we de methode expliciet. Als $\beta_{-1} \neq 0$, dan komt in het rechterlid de term $\beta_{-1} f(x_{n+1}, z_{n+1})$ voor en dan is de methode impliciet.

De bepaling van de coefficienten α_j en β_j kan weer geschieden met de methode van de onbepaalde coefficienten (zie 2.1.3). Door substitutie van $y(x)$, d.i. de oplossing van (1), in (2) krijgen we

$$y(x_{n+1}) = \sum_{j=0}^{k-1} \alpha_j y(x_{n-j}) + h \sum_{j=-1}^{k-1} \beta_j y'(x_{n-j}) + hR \quad . \tag{2a}$$

We kunnen dan eisen dat de methode exact is, d.w.z. dat $R = 0$, als $y(x)$ een polynoom is van de graad p. Deze eis levert $p + 1$ lineaire vergelijkingen voor de coefficienten $\alpha_0, \ldots, \alpha_{k-1}$, $\beta_{-1}, \beta_0, \ldots, \beta_{k-1}$.
De maximaal haalbare graad, die dan tevens gelijk is aan de orde van de methode, is 2k bij de impliciete methode en 2k-1 bij de expliciete methode (we eisen dan immers dat $\beta_{-1} = 0$).
Methoden met maximaal haalbare graad zijn echter in het algemeen asymptotisch instabiel, hetgeen betekent dat de invloed van een fout in de beginwaarden, of van een tussentijdse afrondfout, op de berekende waarde in een vast gekozen eindpunt groter wordt naarmate de stapgrootte h kleiner wordt. (Zie verder 3.3.1) Volgens een beroemde stelling van Dahlquist zijn er geen asymptotisch stabiele lineaire k-stapsmethoden met een orde groter dan $k + 2$ als k even is, resp. k+1 als k oneven is.

We kunnen in de algemene formule (2) ook een aantal coefficienten voorschrijven en de overige coefficienten berekenen. De orde van de methode zal dan

kleiner zijn dan 2k (resp. 2k-1), maar de methode is dan mogelijk wel asymptotisch stabiel.
We krijgen bijvoorbeeld een asymptotisch stabiele methode als we $\alpha_1 = \alpha_2 = \ldots = \alpha_{k-1} = 0$ nemen. Een dergelijke methode wordt een Adams methode genoemd.

Voorbeelden.

1) Expliciete k-staps Adams methoden, dus $\beta_{-1} = 0$, (we schrijven f_{n-j} in plaats van $f(x_{n-j}, z_{n-j})$):

$k = 1 \quad z_{n+1} := z_n + hf_n$ orde 1,

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

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

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

Dit zijn de zg. Adams-Bashforth formules.

2) Impliciete k-staps Adams methoden, dus $\beta_{-1} \neq 0$:

$k = 0 \quad z_{n+1} := z_n + hf_{n+1}$ orde 1,

$k = 1 \quad z_{n+1} := z_n + \frac{1}{2}h(f_{n+1} + f_n)$ orde 2,

$k = 2 \quad z_{n+1} := z_n + \frac{1}{12}h(5f_{n+1} + 8f_n - f_{n-1})$ orde 3,

$k = 3 \quad z_{n+1} := z_n + \frac{1}{24}h(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})$ 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 voorwaardelijke stabiliteit (zie 3.3.2). Daarom worden ze veel gebruikt voor nauwkeurige integratie met

vaste stap over grote trajecten. Wel is steeds een startmethode nodig (als $k > 1$).

3) Voorbeeld van een predictor corrector methode.

Predictor, derde-orde Adams-Bashforth formule

$$z_{n+1}^{(p)} := z_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) .$$

Corrector, derde-orde Adams-Moulton formule

$$z_{n+1}^{(c)} := z_n + \frac{h}{12}(5f(x_{n+1}, z_{n+1}^{(p)}) + 8f_n - f_{n-1})$$

(correcting only once).

Een mogelijke startprocedure is:

z_1 met Euler-predictor en trapeziumregel-corrector

z_2 met tweede-orde Adams-Bashforth-predictor en derde-orde Adams-Moulton-corrector.

De orde van deze methode is drie, d.w.z. dat in een vast punt $x = x_n = x_0 + nh$ geldt

$$y(x_n) = z_n + O(h^3), \quad h \to 0 .$$

4) Expliciete k-staps methoden van de vorm

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

De maximale haalbare orde is weer k.

$k = 2 \quad z_{n+1} := z_{n-1} + 2hf_n$ orde 2,

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

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

5) Impliciete k-staps methoden van de vorm

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

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

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

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

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

De methoden uit 4) en 5) hebben bij gelijke orde een geringere afbreekfout dan de Adams methoden. De voorwaardelijke 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.

3.2.2. 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 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 op pag. 3.3 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

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

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

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

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

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

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 voorwaardelijke stabiliteit (zie 3.3.2) en verder alle voordelen van eenstaps methoden (zo is bv. de stapgrootte gemakkelijk 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 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.

3.3. Conditie en numerieke stabiliteit.

De conditie van het beginwaardeprobleem

$$y' = f(x,y),\ x > x_0\ ,$$
$$y(x_0) = y_0\ , \qquad (1)$$

als functie van de beginwaarde y_0, hangt samen met het richtingsveld behorende bij de differentiaalvergelijking. Als het richtingsveld in de richting van de positieve x-as divergent is (dit is het geval als $\frac{\partial f}{\partial y} > 0$), dan kan een kleine verandering van de beginwaarde een aanzienlijk andere oplossingskromme opleveren. Is daarentegen het richtingsveld convergent ($\frac{\partial f}{\partial y} < 0$), dan liggen de oplossingskrommen bij verschillende beginwaarden dicht bij elkaar.
Hieruit volgt dat de conditie in ieder geval goed is als $\frac{\partial f}{\partial y} < 0$, en beter naarmate $\frac{\partial f}{\partial y}$ meer negatief is. Als $\frac{\partial f}{\partial y} > 0$, dan hangt het van de gevraagde oplossing van de differentiaalvergelijking (dus van de beginwaarde) af of het probleem goed of slecht geconditioneerd is.

Voorbeeld.

Het beginwaardeprobleem

$$y' = \lambda(y - e^{-x}),\ x > 0$$
$$y(0) = a$$

heeft als oplossing

$$y(x) = \frac{\lambda}{\lambda+1}\, e^{-x} + (a - \frac{\lambda}{\lambda+1})e^{\lambda x}\ ,\ \lambda \neq -1$$

$$y(x) = xe^{-x} + ae^{-x}\ ,\ \lambda = -1\ .$$

Als $\lambda(= \frac{\partial f}{\partial y}) < 0$, dan geeft een kleine verandering van a slechts een geringe wijziging van de oplossing, dus dan is het probleem goed geconditioneerd.
Als $\lambda << -1$, dan is de oplossing practisch onafhankelijk van a, want dan geldt bij benadering $y(x) = e^{-x}$.

Opgave. Het conditiegetal van $y(x)$ bij variatie van a kunnen we als volgt definiëren.

$$c(x) := \left| \frac{dy(x)}{da} \cdot \frac{a}{y(x)} \right|$$

Dan geldt $c(x) \leq 1$ voor $a > 0$ en $\lambda < -1$. Ga dit na.

Als $\lambda > 0$, dan bestaat de oplossing uit een term die begrensd is voor alle x (en voor grote x naar nul gaat) en een term die exponentieel stijgt. Voor $a = \frac{\lambda}{\lambda+1}$ ontbreekt de tweede term, dus dan is de oplossing begrensd ($|y(x)| \leq 1$) voor alle x, terwijl als $a \neq \frac{\lambda}{\lambda+1}$ de exponentieel stijgende term reeds voor niet al te grote waarden van λx sterk domineert. Hieruit zien we dat het probleem slecht geconditioneerd is als a in de buurt van $\frac{\lambda}{\lambda+1}$ ligt. Is dit niet het geval, dan geldt bij benadering voor alle x buiten een kleine omgeving van de oorsprong

$$y(x) = (a - \frac{\lambda}{\lambda+1})e^{\lambda x} ,$$

waaruit volgt dat een kleine relatieve verandering van a een kleine relatieve verandering van $y(x)$ oplevert, dus dan is het probleem goed geconditioneerd.

3.3.1. Asymptotische stabiliteit. ([2], 369-378)

Zij $\{z_n\}$ de oplossing van de k-staps methode

$$z_{n+1} := \sum_{j=0}^{k-1} \alpha_j z_{n-j} + hF(x_n, z_{n+1}, z_n, \ldots, z_{n-k+1}, h) \qquad (2)$$

$$z_0, z_1, \ldots, z_{k-1}, \text{ gegeven.}$$

Zij $\{z_n^*\}$ de oplossing van de "gestoorde" k-staps methode

$$z_{n+1}^* := \sum_{j=0}^{k-1} \alpha_j z_{n-j}^* + hF(x_n, z_{n+1}^*, z_n^*, \ldots, z_{n-k+1}^*, h) + hr_n$$

$$z_i^* = z_i + \rho_i, \quad i = 0, 1, \ldots, k-1 .$$

Veronderstel nu dat r_n voor alle n en ρ_i voor alle i "klein" zijn. Dan noemen we de k-staps methode asymptotisch stabiel als voor iedere vaste $x = x_n$ het verschil $|z_n - z_n^*|$ "klein" is en blijft voor $h \to 0$. Geldt daarentegen dat $|z_n - z_n^*|$ groter wordt naarmate h kleiner wordt, dan is de methode asymptotisch instabiel.

Voorbeeld.

De twee-staps expliciete methode

$$z_{n+1} := -4z_n + 5z_{n-1} + h(4f_n + 2f_{n-1}) \quad \text{orde } 3 \tag{3}$$

is asymptotisch instabiel. We zullen dit aantonen door het gedrag van de oplossing in een vast punt $x = x_n$ te onderzoeken.
We verwaarlozen in eerste instantie de term $h(4f_n + 2f_{n-1})$, hetgeen geoorloofd is omdat we het gedrag van de oplossing willen weten als $h \to 0$.
De algemene oplossing van

$$z_{n+1} = -4z_n + 5z_{n-1} \tag{4}$$

krijgen we door de substitutie $z_n = C\lambda^n$. Dit levert voor λ de karakteristieke vergelijking

$$\lambda^2 + 4\lambda - 5 = 0$$

met als oplossingen $\lambda_1 = 1$ en $\lambda_2 = -5$. De algemene oplossing van (4) is dan

$$z_n = A(1)^n + B(-5)^n , \tag{5}$$

waarbij de waarden van A en B worden bepaald door de beginwaarden z_0 en z_1. Bij de "gestoorde" beginwaarden z_0^* en z_1^* behoren dan "gestoorde" waarden $A + \delta A$ en $B + \delta B$ waaruit volgt dat

$$z_n - z_n^* = \delta A + \delta B(-5)^n .$$

Hieruit zien we dat $|z_n - z_n^*|$ in een vast punt $x = x_n$ groter wordt naarmate n groter, dus h kleiner, wordt. Dit geldt voor de vereenvoudigde vergelijking (4), maar omdat de vereenvoudiging relatief klein is, geldt hetzelfde voor de oorspronkelijke methode (3).

□

De instabiliteit komt voort uit de aanwezigheid van de waarde $\lambda_2 = -5$ in de oplossing (5). Dit blijkt ook algemeen te gelden.
Er geldt namelijk de volgende stelling.

Stelling. Een nodige en voldoende voorwaarde voor de asymptotische stabiliteit van de algemene k-staps methode (2) is dat de bij (2) behorende karakteristieke vergelijking

$$\lambda^k - \sum_{j=0}^{k-1} \alpha_j \lambda^{k-j-1} = 0$$

geen wortels heeft buiten en geen meervoudige wortels heeft op de eenheidscirkel in het complexe λ-vlak.

Met behulp van deze eigenschap kunnen we eenvoudig nagaan dat alle methoden uit 3.2 asymptotisch stabiel zijn.

Opgave. Bewijs dat de drie-staps impliciete methode

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

orde 4 heeft en asymptotisch stabiel is.

Voor alle in het voorafgaande genoemde methoden geldt ook dat de locale afbreekfout $R(h,x_n,y_n) \to 0$ als $h \to 0$. Een methode waarvoor dit geldt noemen we consistent, hetgeen betekent dat de benaderingsformule overgaat in de differentiaalvergelijking als $h \to 0$. Verder noemen we een methode convergent als $\lim_{h\to 0} z_n = y(x_n)$ voor iedere vaste $x = x_n$.

Dan geldt de volgende stelling die aangeeft of een methode in principe bruikbaar is.

Stelling. Voor de convergentie van de methode (2) is nodig en voldoende dat de methode consistent en asymptotisch stabiel is.

3.3.2. Voorwaardelijke stabiliteit.([9], Ch. 3)

De asymptotische stabiliteit is de minimale eis waaraan een methode moet voldoen om bruikbaar te zijn. Het zegt echter alleen iets van het gedrag van de methode als $h \to 0$.

Om iets te kunnen zeggen over het gedrag van de methode voor gegeven waarde van h wordt het begrip voorwaardelijke stabiliteit gehanteerd.

Voor het bepalen van de voorwaardelijke stabiliteit van een methode gebruiken we het beginwaardeprobleem

$$\begin{aligned} y' &= \lambda y, \quad x > 0 \\ y(0) &= a \end{aligned} \tag{6}$$

waarvan de oplossing is $y(x) = ae^{\lambda x}$. Als $\lambda < 0$, dan blijft deze oplossing begrensd voor alle $x > 0$ en we wensen dat dan de numerieke oplossing hetzelfde gedrag vertoont.

Voorbeelden.

1) De methode van Euler

$$z_{n+1} := z_n + hf_n$$

$$= (1 + h\lambda)z_n ,$$

$$z_0 := a$$

heeft als oplossing

$$z_n = a(1 + h\lambda)^n .$$

We eisen nu dat voor $\lambda < 0$ z_n begrensd is voor alle n, bij vaste h. Dit is het geval indien

$$|1 + h\lambda| < 1$$

ofwel

$$-2 < h\lambda < 0 .$$

Het interval (-2,0) waarin $h\lambda$ moet liggen noemen we het stabiliteitsgebied van de methode.

2) De trapeziumregel

$$z_{n+1} := z_n + \tfrac{1}{2}h(f_{n+1} + f_n) ,$$

$$z_0 := a$$

heeft als oplossing

$$z_n = a\left(\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right)^n .$$

Nu moet voor $\lambda < 0$ gelden

$$\left|\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right| < 1 .$$

Hieraan is voldaan voor iedere negatieve waarde van $h\lambda$. Het stabiliteitsgebied is dus het interval $(-\infty,0)$. We zeggen daarom dat de trapeziumregel onvoorwaardelijk stabiel is en de methode van Euler voorwaardelijk stabiel is.

Ter vergelijking geven we hieronder de stabiliteitsgebieden van enkele vierde-orde methoden.

methode	stabiliteitsgebied
Runge-Kutta (orde 4)	(-2.785 , 0)
Adams-Bashforth (orde 4)	(-0.3 , 0)
Adams-Moulton (orde 4)	(-3.0 , 0)
Adams predictor-corrector met beide bovenstaande formules.	
(correcting only once)	(-1.3 , 0)
Milne-corrector	geen

Opmerking. De voorwaardelijke stabiliteit is afgeleid met behulp van de lineaire differentiaalvergelijking $y' = \lambda y$. Omdat de algemene differentiaalvergelijking $y' = f(x,y)$ in een omgeving van een punt (x_n,y_n) bij benadering lineair is

$$y' = f(x,y_n) + (\frac{\partial f}{\partial y})_n (y - y_n) + \ldots ,$$

verwachten we dat de resultaten ook bruikbaar zijn in het algemene geval waarbij we λ vervangen door $\frac{\partial f}{\partial y}$ of een schatting daarvan.

3.3.3. Stijve differentiaalvergelijkingen. ([5], p. 209-222; [9], Ch. 6)

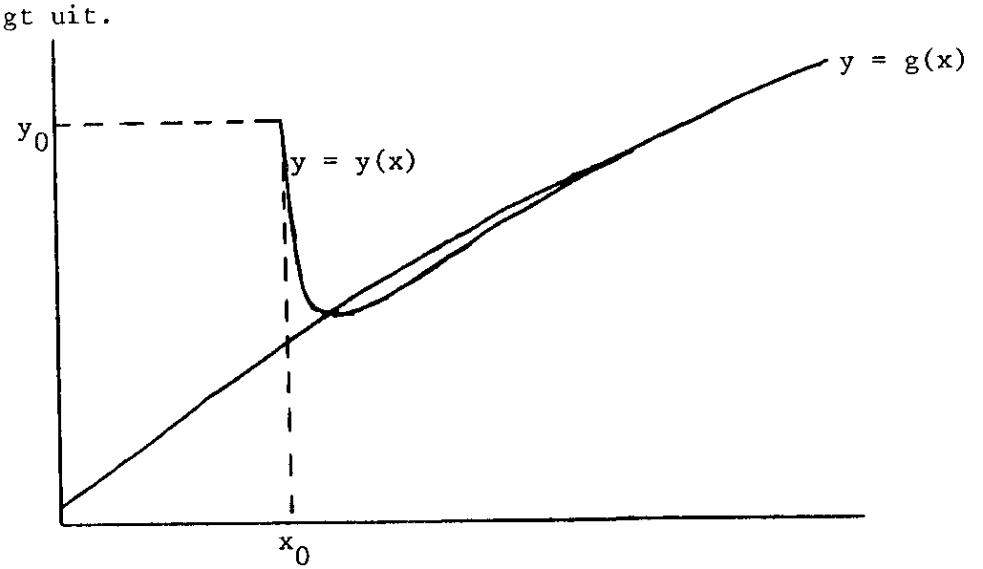
We beschouwen het volgende beginwaardeprobleem

$$\begin{aligned} y' &= \lambda(y - g(x)) \quad x > x_0 \\ y(x_0) &= y_0 . \end{aligned} \tag{7}$$

Als λ sterk negatief is, dan is $y(x) = g(x)$ een goede benadering voor een oplossing van de differentiaalvergelijking, want dan geldt bij benadering

$$y(x) = g(x) + Ae^{\lambda(x - x_0)} + \frac{1}{\lambda} g'(x)$$

met $A := y_0 - g(x_0) - \frac{1}{\lambda} g'(x_0)$. De grafiek van de oplossing ziet er ongeveer als volgt uit.



Dit is een voorbeeld van een zogenaamde stijve differentiaalvergelijking. In het algemeen zullen we een differentiaalvergelijking stijf noemen als $\frac{\partial f}{\partial y}$ sterk negatief is (in een of andere relatieve zin). Het bijbehorende beginwaardeprobleem heeft een zeer goede conditie; de oplossing is practisch onafhankelijk van de beginvoorwaarde met uitzondering van een zeer klein interval in het begin (inschakel-verschijnsel).

Kennelijk spelen bij een probleem met een stijve differentiaalvergelijking twee "tijdschalen" een rol. De ene "tijdschaal" hangt samen met het verloop van $g(x)$, waarvan $|g(x)/g'(x)|$ de karakteristieke tijd genoemd kan worden, de andere met het verloop van het inschakelverschijnsel, waarvan de karakteristieke tijd $1/|\frac{\partial f}{\partial y}| = |\lambda|^{-1}$ is.
Voor een beschrijving van de oplossing in de beginfase is een stapgrootte nodig die een fractie van $|\lambda|^{-1}$ (dus zeer klein) is. Na de zeer korte beginfase is een veel grotere stapgrootte, nl. een fractie van $|g(x)/g'(x)|$ voldoende voor de beschrijving van de oplossing.
Als $g(x)$ een gladde functie is dan verwachten we hetzelfde van de oplossing $y(x)$, behalve vlak bij $x = x_0$, op grond hiervan zouden we kunnen denken dat bijvoorbeeld een vierde orde Runge-Kutta methode met "grote" stap een redelijk nauwkeurig resultaat geeft. Dit is echter niet het geval.
Omdat deze methode een beperkt stabiliteitsgebied heeft, $(-2.785 < h\lambda < 0)$ impliceert de eis van voorwaardelijke stabiliteit dat bij grote negatieve λ de waarde van h zeer klein moet zijn (i.c. $h < -\frac{2.785}{\lambda}$), ook in het gebied waar de oplossing glad verloopt. Daarom zijn methoden met een beperkt stabiliteitsgebied (Runge-Kutta, Adams) ongeschikt voor problemen met een stijve differentiaalvergelijking. Wel geschikt zijn methoden die onvoorwaardelijk stabiel zijn, zoals de trapeziumregel (deze zijn echter ook impliciet).
Er bestaan speciale algorithmen voor het oplossen van stijve differentiaalvergelijkingen, bijvoorbeeld de methode van Gear (zie bijvoorbeeld RC-Informatie PP-3.4.1.).

3.4. Stelsels eerste orde differentiaalvergelijkingen en differentiaalvergelijkingen van hogere orde.

Alle behandelde methoden voor de scalaire differentiaalvergelijking $y' = f(x,y)$ laten zich eenvoudig generaliseren voor het stelsel eerste orde differentiaalvergelijkingen

$$\frac{dy_1}{dx} = f_1(x,y_1,y_2,\ldots,y_k)$$

$$\frac{dy_2}{dx} = f_2(x,y_1,y_2,\ldots,y_k)$$

. . . . . .

$$\frac{dy_k}{dx} = f_k(x,y_1,y_2,\ldots,y_k) \ ,$$

dat in vector-notatie als volgt kan worden beschreven

$$\frac{d\underline{y}}{dx} = \underline{f}(x,\underline{y}) \ . \qquad (1)$$

Voorbeeld. De tweede orde Runge-Kutta methode (zie pag. 3.3) 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:

$$k_{11} := hf_1(x_n,z_{1n},z_{2n}) \ ,$$

$$k_{21} := hf_2(x_n,z_{1n},z_{2n}) \ ,$$

$$k_{12} := hf_1(x_n + h,z_{1n} + k_{11},z_{2n} + k_{21}) \ ,$$

$$k_{22} := hf_2(x_n + h,z_{1n} + k_{11},z_{2n} + k_{21}) \ ,$$

$$z_{1,n+1} := z_{1n} + \tfrac{1}{2}(k_{11} + k_{12}) ,$$

$$z_{2,n+1} := z_{2n} + \tfrac{1}{2}(k_{21} + k_{22}) .$$

Hiermee kunnen we ook differentiaalvergelijkingen van hogere orde behandelen, bijv.

$$\frac{d^k y}{dx^k} = f\left(x, y, \frac{dy}{dx}, \ldots, \frac{d^{k-1} y}{dx^{k-1}}\right) . \qquad (2)$$

Met behulp van de substitutie

$$w_1 = y,\ w_2 = \frac{dy}{dx}\ ,\ \ldots,\ w_k = \frac{d^{k-1} y}{dx^{k-1}} ,$$

gaat (2) over in het stelsel eerste orde vergelijkingen

$$\frac{dw_1}{dx} = w_2$$

. . . . . .

$$\frac{dw_{k-1}}{dx} = w_k$$

$$\frac{dw_k}{dx} = f(x, w_1, \ldots, w_k) ,$$

ofwel in vectornotatie

$$\frac{d\underline{w}}{dx} = \underline{f}(x, \underline{w}) , \qquad (3)$$

met

$$\underline{f}(x, \underline{w}) = \begin{pmatrix} w_2 \\ \ldots \\ w_k \\ f(x, w_1, \ldots, w_k) \end{pmatrix} .$$

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 k-de orde vergelijking (2) horen als passende beginvoorwaarden
$y(x_0) = y_0, y'(x_0) = y'_0, \ldots, y^{(k-1)}(x_0) = y_0^{(k-1)}$.
Stellen we $w_{10} = y_0, \ldots, w_{k0} = y_0^{(k-1)}$, dan behoort bij (3) dus de beginvoorwaarde $\underline{w}(x_0) = \underline{w}_0$.

3.4.1. 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 afrondfouten nog verminderen door te stellen $hv_{n+\frac{1}{2}} := z_{n+1} - z_n$ 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 afrondfout in v_n dan berekening via (5).

Opmerking. Daar

$$y_{n+1} - 2y_n + y_{n-1} = h^2y''_n + \frac{1}{12}h^4y_n^{(4)} + O(h^6) =$$

$$= h^2y''_n + \frac{1}{12}h^2[y''_{n+1} - 2y''_n + y''_{n-1}] + O(h^6) ,$$

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).

3.5. Randwaardeproblemen ([2], p. 359-365; [14], p. 154-158)

Een beginwaardeprobleem bestaat uit een differentiaalvergelijking (stelsel differentiaalvergelijkingen) gedefinieerd voor $x > x_0$ en een aantal beginvoorwaarden, dit zijn voorwaarden voor de oplossing in het punt x_0. Gevraagd wordt dan de oplossing voor $x \geq x_0$, die bij een "nette" differentiaalvergelijking en een "voldoende" aantal beginvoorwaarden bestaat en eenduidig bepaald is (althans in de buurt van x_0).

Naast beginwaardeproblemen komen in de praktijk ook randwaardeproblemen voor. Een randwaardeprobleem bestaat uit een differentiaalvergelijking (stelsel differentiaalvergelijkingen) gedefinieerd op een interval (a,b) en een aantal randvoorwaarden, dit zijn voorwaarden voor de oplossing in de punten a en b. Gevraagd wordt dan de oplossing in het interval (a,b). Het is echter mogelijk dat deze oplossing bij een "nette" differentiaalvergelijking en een "voldoende" aantal randvoorwaarden niet bestaat, of niet eenduidig bepaald is.

Voorbeeld.

Het randwaardeprobleem

$$y'' + k^2y = 0, \quad 0 < x < 1 \ ,$$

$$y(0) = 0, \quad y(1) = \beta \ ,$$

heeft, als $k \neq n\pi$ met n geheel, een eenduidig bepaalde oplossing voor iedere β, nl. $y(x) = \beta\frac{\sin kx}{\sin k}$; heeft, als $k = n\pi$, geen oplossing als $\beta \neq 0$ en oneindig veel oplossingen als $\beta = 0$, nl. $y(x) = A \sin n\pi x$. □

We behandelen in deze paragraaf uitsluitend randwaardeproblemen bestaande uit een tweede orde differentiaalvergelijking

$$\frac{d^2y}{dx^2} = f(x,y,\frac{dy}{dx}), \quad a < x < b \tag{1}$$

met lineaire randvoorwaarden, bijvoorbeeld

$$\begin{aligned} \alpha_1 y(a) + \alpha_2 y'(a) &= \gamma \ , \\ \beta_1 y(b) + \beta_2 y'(b) &= \delta \ . \end{aligned} \tag{2}$$

Een ander mogelijk type randvoorwaarden bij (1) is bijvoorbeeld

$$y(a) = y(b), \ y'(a) = y'(b) \ . \tag{3}$$

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 een beginwaardeprobleem
b) rechtstreekse discretisatie, hetgeen tot een groot stelsel, in het algemeen niet lineaire, vergelijkingen leidt.

Als men een niet te hoge nauwkeurigheid eist, dan verdient de tweede methode, vanwege zijn eenvoud, de voorkeur. Wenst men een hoge nauwkeurigheid, dan is de eerste methode, waarbij men gebruik maakt van een hoge orde integratie methode, vaak beter. Men moet dan wel bedacht zijn op mogelijke numerieke instabiliteit.

3.5.1. Herleiden tot een beginwaardeprobleem (schieten)

We lichten deze methode toe aan de hand van een voorbeeld. Zij gegeven het randwaardeprobleem

$$\begin{aligned} y'' &= f(x,y,y'), \quad a < x < b \ , \\ y(a) &= \gamma, \quad y(b) = \delta \ . \end{aligned} \tag{4}$$

Het bijbehorende beginwaardeprobleem bestaat uit dezelfde differentiaalvergelijking en de beginvoorwaarden

$$y(a) = \gamma, \quad y'(a) = s \tag{5}$$

Voor iedere waarde van s heeft dit beginwaardeprobleem een eenduidig bepaalde oplossing, die we y(x,s) zullen noemen. Dit is een oplossing van het randwaardeprobleem als hij voldoet aan de randvoorwaarde in b, dus als

$$y(b,s) = \delta \ . \tag{6}$$

Uit deze zogenaamde sluitvergelijking moet de waarde van s bepaald worden.

Het op bovenstaande wijze oplossen van een randwaardeprobleem blijkt dus neer te komen op het oplossen van vergelijking (6), c.q. het bepalen van een nulpunt van de functie

$$F(s) := y(b,s) - \delta \ .$$

Dit kan numeriek worden gedaan met een van de methoden uit hoofdstuk 1. Het berekenen van de functiewaarde F(s) impliceert dan het oplossen van een beginwaardeprobleem. Dit laatste kan met een van de methoden uit de voorgaande paragrafen van dit hoofdstuk.

Als de differentiaalvergelijking (4) lineair is, d.w.z. van de vorm

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = r(x) \ , \qquad (7)$$

dan zal y(x,s) lineair van s afhangen. Immers, de functie y(x,s) - y(x,0) voldoet aan de bij (7) behorende <u>homogene</u> differentiaalvergelijking

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = 0 \qquad (8)$$

en aan de randvoorwaarden

$$y(a) = 0, \quad y'(a) = s \ . \qquad (9)$$

Hieruit volgt dat

$$y(x,s) - y(x,0) = s[y(x,1) - y(x,0)] \ .$$

We kunnen nu dus volstaan met de bepaling van $y_0(x) := y(x,0)$ als oplossing van (7) met de beginvoorwaarden (5) met s = 0 en de bepaling van

$$y_1(x) := y(x,1) - y(x,0)$$

als oplossing van (8) met de beginvoorwaarden (9) met s = 1.
Substitutie van

$$y(x,s) = y_0(x) + s\, y_1(x) \qquad (10)$$

in de sluitvergelijking (6) levert nu

$$y_0(b) + s\, y_1(b) = \delta \ . \qquad (11)$$

Hierdoor is s bepaald.

De bepaling van s uit (11) loopt spaak indien

$$y_1(b) = 0 \ . \tag{12a}$$

Is dit het geval,dan is, daar eveneens geldt (zie (9))

$$y_1(a) = 0 \ , \tag{12b}$$

de functie $y_1(x)$ een niet triviale oplossing van de homogene differentiaalvergelijking (8) met de homogene randvoorwaarden (12). 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. (Zie ook het voorbeeld op pag. 3.24.)

Uit het bovenstaande kunnen we concluderen dat de oplossing van het inhomogene randwaarde probleem slecht bepaald zal zijn als $y_1(b)$ "klein" is.
Dit zien we ook aan vergelijking (11), waardoor s bij "kleine" $y_1(b)$ slecht bepaald is.
Dat in dit geval bij de numerieke berekening van s cijferverlies optreedt, omdat $|\delta - y_0(b)| \ll |\delta|$, is derhalve het gevolg van de slechte conditie van het probleem, en niet van numerieke instabiliteit van de schietmethode.

Numerieke instabiliteit kan zich wel voordoen als de oplossing y(x,s) veel kleiner is dan zijn componenten $y_0(x)$ en $sy_1(x)$ in formule (10).
Een <u>voorbeeld</u> hiervan is het volgende randwaardeprobleem

$$\begin{aligned} &y'' - \alpha^2 y = 0 \qquad 0 < x < 1 \ , \\ &y(0) = \gamma \ , \quad y(1) = \delta \ . \end{aligned} \tag{13}$$

De oplossing hiervan is

$$y(x) = \gamma \frac{\sinh \alpha(1 - x)}{\sinh \alpha} + \delta \frac{\sinh \alpha x}{\sinh \alpha} \ .$$

Als γ en δ hetzelfde teken hebben, dan is deze oplossing goed geconditioneerd als functie van γ en δ.
We veronderstellen in het vervolg dat $\gamma > 0$ en $\delta > 0$ en merken dan op dat, als α "groot" is, voor x niet "dicht" bij 0 of 1 geldt $0 < y(x) \ll \max(\gamma,\delta)$.

Met de voorgestelde schietmethode zouden we vinden $y_0(x) = \gamma \cosh \alpha x$ en $y_1(x) = \alpha^{-1} \sinh \alpha x$.

Voor grote waarden van α geldt derhalve $|y_0(x)| >> |y(x,s)|$ en dus treedt er bij de bepaling van $y(x,s)$ via (10) aanzienlijk cijferverlies op.
Een (in dit geval) betere algorithme krijgen we door de oplossing van (13) te schrijven als

$$y(x) = A\varphi_1(x) + B\varphi_2(x) ,$$

met φ_1 en φ_2 oplossingen van $y'' - \alpha^2 y = 0$, bepaald door de beginvoorwaarden

$$\varphi_1(0) = 0 , \varphi_1'(0) = 1 ,$$

resp.

$$\varphi_2(1) = 0 , \varphi_1'(0) = -1 .$$

Dan krijgen we als sluitvergelijkingen

$$B\varphi_2(0) = \gamma, \quad A\varphi_1(1) = \delta$$

en dus

$$y(x) = \gamma \frac{\varphi_2(x)}{\varphi_2(0)} + \delta \frac{\varphi_1(x)}{\varphi_1(1)} .$$

Er treedt nu geen cijferverlies op omdat beide termen in deze laatste formule positief zijn. □

Ook bij niet lineaire problemen kan het voorkomen dat de sluitvergelijking (6), en dus het randwaardeprobleem, geen oplossing heeft.

3.5.2. Herleiden tot een stelsel vergelijkingen (discretisatie)

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

$$z_{j-1} - 2z_j + z_{j+1} - h^2 f(x_j, z_j, \frac{z_{j+1} - z_{j-1}}{2h}) = 0 \qquad (14)$$

voor de hand.

Van de algemene randvoorwaarden (2) bekijken we de volgende twee speciale gevallen

a) $y(a) = \gamma$, $y(b) = \delta$.
We nemen dan natuurlijk $z_0 = \gamma$ en $z_N = \delta$. Daarmee levert de differentievergelijking (14) voor $j = 1,2,\ldots,N-1$ een stelsel van N-1 niet lineaire vergelijkingen voor de N - 1 onbekende waarden $z_1, z_2, \ldots, z_{N-1}$.

b) $y'(a) = \gamma$, $y(b) = \delta$.
De randvoorwaarde in $x = a$ discretiseren we nu door de centrale differentieformule $z_1 - z_{-1}/2h = \gamma$.

Uit deze vergelijking en de differentievergelijking (14) in $x = a$ (dus $j = 0$) elimineren we z_{-1} hetgeen de volgende vergelijking levert.

$$-z_0 + z_1 - \tfrac{1}{2}h^2 f(x_0, z_0, \gamma) = h\gamma \ .$$

Deze vergelijking samen met differentievergelijking (14) voor $j = 1,2,\ldots,N-1$ met in de laatste vergelijking $z_N = \delta$ leveren nu N vergelijkingen voor de N onbekende waarden $z_0, z_1, \ldots, z_{N-1}$.

Men kan bewijzen dat, als f voldoende glad is, bij deze discretisatie de oplossing een globale afbreekfout van de orde 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

$$y(x_j) = z_j + c_1(x_j)h^2 + c_2(x_j)h^4 + \ldots \ .$$

Hierop kan weer h^2-extrapolatie gebaseerd worden.

Als de functie $f(x,y,y')$ lineair is in y en y', dan is het stelsel vergelijkingen lineair. Bijvoorbeeld, voor het randwaardeprobleem

$$y'' + p(x)y' + q(x)y = r(x), \quad 0 < x < 1 \ ,$$

$$y(0) = \gamma, \ y(1) = \delta \ ,$$

gaat (12) over in

$$(1 - \tfrac{1}{2}hp_j)z_{j-1} - (2 - h^2q_j)z_j + (1 + \tfrac{1}{2}hp_j)z_{j+1} = h^2r_j \ . \qquad (15)$$

Dit geldt voor $j = 1,2,\ldots,N-1$, waarbij in de vergelijking voor $j = 1$ de term $(1 - \tfrac{1}{2}hp_1)\, z_0 = (1 - \tfrac{1}{2}hp_1)\gamma$ naar het rechterlid gebracht moet worden. Analoog voor $(1 + \tfrac{1}{2}hp_{N-1})z_N = (1 + \tfrac{1}{2}hp_{N-1})\delta$ in de vergelijking voor $j = N-1$.

We hebben nu een groot (bv. $N = 100$) stelsel lineaire vergelijkingen $A\,\underline{z} = \underline{d}$ gekregen, waarvan de matrix een zogenaamde tridiagonale matrix is.

$$A = \begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & \bigcirc & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & \bigcirc & & b_{N-2} & a_{N-2} & c_{N-2} \\ & & & & b_{N-1} & a_{N-1} \end{pmatrix}, \quad \underline{d} = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix}$$

met $a_j = -(2 - h^2q_j)$, $b_j = 1 - \frac{1}{2}hp_j$, $c_j = 1 + \frac{1}{2}hp_j$, $d_1 = h^2r_1 - (1 - \frac{1}{2}hp_1)\gamma$, $d_j = h^2r_j$ voor j = 2 t/m N-2, $d_{N-1} = h^2r_{N-1} - (1 + \frac{1}{2}hp_{N-1})\delta$.
Door eliminatie van b_j (vegen) gaat dit stelsel over in een driehoeksstelsel.
$U\,\underline{z} = \underline{e}$ met

$$U = \begin{pmatrix} u_1 & c_1 & & & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & u_{N-2} & c_{N-2} \\ & & & & u_{N-1} \end{pmatrix}$$

Dit eliminatieproces geeft, in pseudo-ALGOL, de volgende algorithme

```
u[1] := a[1]; e[1] := d[1];
for j := 2 step 1 until N-1
do begin ℓ := b[j]/u[j-1];
         u[j] := a[j] - ℓ × c[j-1]; e[j] := d[j] - ℓ × e[j-1]
   end
```

De oplossing wordt vervolgens verkregen door terugsubstitutie.

```
z[N-1] := e[N-1]/u[N-1];
for j := N-2 step -1 until 1
do z[j] := (e[j] - c[j] × z[j+1])/u[j]
```

Deze algorithme, die een bijzonder geval is van Gauss-eliminatie (zie hoofdstuk 5) staat ook bekend onder de naam Thomas-algorithme.

Voor de uitvoering van de algorithme is het nodig en voldoende dat $u_j \neq 0$ voor alle j. Men kan voorwaarden afleiden waarvoor dit het geval is. Bijvoorbeeld geldt $u_j \neq 0$ voor alle j als de functies p(x) en q(x) voldoen aan de ongelijkheid

$$q(x) < 0 \ .$$

In dit geval is de algorithme ook numeriek stabiel (zie verder hoofdstuk 5).

Als de functie f(x,y,y') niet lineair is, dan is het stelsel (14) met de bijbehorende randvoorwaarden eveneens niet lineair en moet het met een iteratieve methode opgelost worden.

Voorbeeld.

Beschouw het volgende randwaardeprobleem

$$y'' = f(x,y), \quad a < x < b ,$$

$$y(a) = y(b) = 0 .$$

(In dit geval hangt f dus niet van y' af.)
Het stelsel vergelijkingen (12) wordt dan

$$- z_{j-1} + 2z_j - z_{j+1} + h^2 f(x_j,z_j) = 0, \; j = 1,2,\ldots,N-1$$

met $z_0 = z_N = 0$.
Met behulp van de definities

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad \underline{f}(\underline{z}) = \begin{pmatrix} f(x_1,z_1) \\ f(x_2,z_2) \\ \vdots \\ f(x_{N-1},z_{N-1}) \end{pmatrix}, \quad \underline{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{N-1} \end{pmatrix}$$

luidt (14) in vectornotatie

$$\underline{F}(\underline{z}) := A\,\underline{z} + h^2 \underline{f}(\underline{z}) = \underline{0} \qquad (16)$$

Als de partiele afgeleide van f(x,y) naar y

$$g(x,y) := \frac{\partial f}{\partial y}(x,y)$$

bekend is, dan kunnen we (16) oplossen met de methode van Newton, want de functionaalmatrix van $\underline{F}(\underline{z})$ is gelijk aan

$$F'(\underline{z}) = A + h^2 D(\underline{z}) \qquad (17)$$

waarbij $D(\underline{z})$ de diagonaalmatrix is met

$$D_{ii} = g(x_i,z_i) .$$

In de ν-de iteratieslag moet dan het volgende stelsel vergelijkingen worden opgelost.

$$\underline{F}(\underline{z}^{(\nu)}) + F'(\underline{z}^{(\nu)})(\underline{z} - \underline{z}^{(\nu)}) = \underline{0}$$

ofwel met behulp van (16) en (17)

$$F'(\underline{z}^{(\nu)})\ \underline{z} = h^2(D(\underline{z}^{(\nu)})\ \underline{z}^{(\nu)} - \underline{f}(\underline{z}^{(\nu)}))\ .$$

In coordinaten uitgeschreven is dit

$$-z_{j-1} + (2 + h^2 g(x_j, z_j^{(\nu)}))z_j - z_{j+1} =$$

$$= h^2[g(x_j, z_j^{(\nu)})\ z_j^{(\nu)} - f(x_j, z_j^{(\nu)})],\ j = 1,2,\ldots,N-1$$

met $z_0 = z_N = 0$.
Dit is weer een stelsel met een tridiagonale matrix. De oplossing noemen we $\underline{z}^{(\nu+1)}$.

In het algemeen zal de convergentie van dit proces kwadratisch zijn, d.w.z. dat zal gelden

$$\|\underline{z}^{(\nu+1)} - \underline{z}\| \leq C\,\|\underline{z}^{(\nu)} - \underline{z}\|^2.$$

4. Partiële differentiaalvergelijkingen ([2], p. 383-392; [15], ch. 5,6)

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 van het type (1) in drie groepen kunnen 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 $n \times n$ 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. $a < x < b$ (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 y-richting u of $\frac{\partial u}{\partial x}$ gegeven.

c) parabolische vergelijkingen:
als bij hyperbolische vergelijkingen, langs de beginkromme echter alleen u gegeven.

4.1. De warmtegeleidingsvergelijking

We behandelen de lineaire vergelijking

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right) + bu + c\ ,\quad 0 < x < 1\ ,\ t > 0 \qquad (1)$$

waarin a, b en c functies van x en t mogen zijn, $a(x,t) > 0$ voor alle x en t. Als randvoorwaarden nemen we

$$u(0,t) = p_0(t),\ u(1,t) = p_1(t),\quad t > 0$$

en als beginvoorwaarde

$$u(x,0) = f(x), \quad 0 \le x \le 1 .$$

We gaan de vergelijking (1) discretiseren. Daartoe verdelen we (0,1) in J intervallen ter lengte h = 1/J en nemen we in de t-richting een stapgrootte k. We krijgen dan de roosterpunten $x_j := jh$, $t_n := nk$.
We noemen

$$u_{j,n} := u(x_j,t_n) .$$

4.1.1. 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

$$|R_{j,n}| \le C_1 h^2 + C_2 k. \qquad (3)$$

Uit (2) volgt als methode (de door de methode verkregen benadering voor $u_{j,n}$ noemen we $v_{j,n}$)

$$v_{j,n+1} = v_{j,n} + \frac{k}{h^2}\left[a_{j+\frac{1}{2},n}(v_{j+1,n} - v_{j,n}) - a_{j-\frac{1}{2},n}(v_{j,n} - v_{j-1,n})\right]$$

$$+ k(b_{j,n} v_{j,n} + c_{j,n}) \qquad (4)$$

voor j = 1,2,...,J-1. Hiermee kunnen wij bij gegeven waarden van $v_{j,n}$ voor j = 0,1,...,J, de waarden van $v_{j,n+1}$ voor j = 1,2,...,J-1 berekenen. Voor de randpunten nemen we

$$v_{0,n+1} = p_0(t_{n+1}), \; v_{J,n+1} = p_1(t_{n+1}) . \qquad (5)$$

We starten natuurlijk met

$$v_{j,0} = f(x_j), \quad j = 0,1,...,J \qquad (6)$$

en kunnen dan achtereenvolgens voor n = 0,1,2,... de oplossing $v_{j,n+1}$ expliciet met (4) berekenen. De aldus verkregen expliciete methode wordt de methode van Euler genoemd.

We onderzoeken de asymptotische stabiliteit van deze methode, waaronder we verstaan dat bij een "kleine" verandering van de begin- of randwaarden (of een "kleine" tussentijdse fout) de verandering van de oplossing klein is en blijft als $h,k \to 0$ (vergelijk 3.3.1).
Het probleem is lineair. Dit houdt in dat de verandering in de oplossing t.g.v. storing in de begin- en/of randwaarden voldoet aan de differentievergelijking (4) met als begin- en randvoorwaarden de betreffende storingen.

We nemen nu het speciale geval:

$$a(x,t) \text{ constant}, \; b(x,t) = c(x,t) = 0 \; ,$$
$$p_0(t) = p_1(t) = 0$$

(dus alleen een storing in de beginwaarden).
De differentievergelijking (4) wordt dan

$$v_{j,n+1} = v_{j,n} + \alpha(v_{j+1,n} - 2v_{j,n} + v_{j-1,n})$$
$$= (1 - 2\alpha)v_{j,n} + \alpha(v_{j+1,n} + v_{j-1,n}), \qquad (7)$$

waarin

$$\alpha = ak/h^2 \; .$$

Voor $\alpha \leq \frac{1}{2}$ is $1 - 2\alpha \geq 0$ en daaruit volgt direct (ga na)

$$\max_j |v_{j,n+1}| \leq \max_j |v_{j,n}| \leq \ldots \leq \max_j |v_{j,0}| \; . \qquad (8)$$

In dit geval kan dus geen versterking van storingen optreden.

Om nu voor willekeurige α na te gaan hoe een storing zich voortplant, nemen we als beginwaarden

$$f(x) = \sin m\pi x, \quad m \text{ geheel} \; ^{*)} \qquad (9)$$

Dan kan de oplossing van het gediscretiseerde probleem (7) in gesloten vorm worden gegeven, nl.

---

*) Dit is voldoende algemeen omdat een willekeurige storing, die nul is aan de randen, te schrijven is als een Fourier-sinus-reeks.

$$v_{j,n} = (1 - 4\alpha \sin^2 \tfrac{1}{2}m\pi h)^n \sin m\pi x_j \ .$$

De methode heet asymptotisch stabiel als $v_{j,n}$ begrensd is voor $h,k \to 0$. Hieruit volgt de stabiliteitsvoorwaarde

$$|1 - 4\alpha \sin^2 \tfrac{1}{2}m\pi h| \leq 1 \ , \tag{10}$$

waaraan voldaan is als h en k zodanig naar nul gaan dat

$$\alpha \leq \tfrac{1}{2} \ .$$

Laten we daarentegen h en k zodanig naar nul gaan dat $\alpha \geq \alpha_0 > \tfrac{1}{2}$ dan zijn er waarden van m waarvoor de bij de beginwaarde (9) behorende oplossing exponentieel groeit. De methode is dan instabiel.
Conclusie: De methode van Euler is voorwaardelijk asymptotisch stabiel en de voorwaarde is:

$$\frac{ak}{h^2} \leq \tfrac{1}{2} \text{ voor } h,k \to 0 \ . \tag{11}$$

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.

4.1.2. 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 $x_j, t_{n+\frac{1}{2}}$)

$$\frac{u_{j,n+1} - u_{j,n}}{k} = \tfrac{1}{2} \frac{u_{j+1,n+1} - 2u_{j,n+1} + u_{j-1,n+1}}{h^2}$$

$$+ \tfrac{1}{2} \frac{u_{j+1,n} - 2u_{j,n} + u_{j-1,n}}{h^2}$$

$$+ \tfrac{1}{2} (b_{j,n+1} u_{j,n+1} + c_{j,n+1}) + \tfrac{1}{2}(b_{j,n} u_{j,n} + c_{j,n})$$

$$+ R_{j,n+\frac{1}{2}} \ , \tag{2}$$

met $|R_{j,n+\frac{1}{2}}| \leq C_1 h^2 + C_2 k^2$. (3)

Hieruit volgt als methode (met $\alpha := k/h^2$)

$$(1 + \alpha - \tfrac{1}{2} kb_{j,n+1})v_{j,n+1} - \tfrac{1}{2}\alpha(v_{j+1,n+1} + v_{j-1,n+1})$$

$$= (1 - \alpha + \tfrac{1}{2} kb_{j,n})v_{j,n} + \tfrac{1}{2}\alpha(v_{j+1,n} + v_{j-1,n}) + \frac{k}{2}(c_{j,n+1}+c_{j,n}) . \quad (4)$$

We krijgen dus voor iedere j een vergelijking tussen $v_{j,n+1}$, $v_{j+1,n+1}$ en $v_{j-1,n+1}$.

We nemen weer als gebied $0 \leq x \leq 1$, $t \geq 0$ en als randvoorwaarden

$$u(0,t) = p_0(t), \; u(1,t) = p_1(t) .$$

Dan worden de randpunten

$$v_{0,n} = p_0(t_n), \; v_{J,n} = p_1(t_n) ,$$

$$v_{0,n+1} = p_0(t_{n+1}), \; v_{J,n+1} = p_1(t_{n+1}) . \quad (5)$$

De vergelijkingen (4) moeten nu gelden voor $1 \leq j \leq J-1$. Daar $v_{0,n+1}$ en $v_{J,n+1}$ bekend zijn, levert dit J-1 vergelijkingen voor $v_{1,n+1},\ldots,v_{J-1,n+1}$. 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 J).

We beperken ons nu tot het geval $b = c = p_0 = p_1 = 0$. Dan is vrij eenvoudig in te zien dat de methode voor alle α asymptotisch stabiel is. We hebben nu te maken met het stelsel

$$v_{j,n+1} - v_{j,n} = \tfrac{1}{2}\alpha(\delta^2 v_{j,n+1} + \delta^2 v_{j,n}), \; 1 \leq j \leq J-1$$

$$v_{0,n+1} = v_{J,n+1} = 0 \quad (6)$$

(waarin $\delta^2 v_{j,n} := v_{j+1,n} - 2v_{j,n} + v_{j-1,n}$, etc.) .

Zij de beginvoorwaarde $v_{j,0} = f_j$, $0 \leq j \leq J$ (met $f_0 = f_J = 0$).

Uit (6) volgt na vermenigvuldiging met $w_j := v_{j,n+1} + v_{j,n}$ en sommatie over j

$$\sum_{j=1}^{J-1} v_{j,n+1}^2 - \sum_{j=1}^{J-1} v_{j,n}^2 = \tfrac{1}{2}\alpha \sum_{j=1}^{J-1} w_j \, \delta^2 w_j$$

$$= \tfrac{1}{2}\alpha \sum_{j=1}^{J-1} (w_j \, w_{j+1} + w_j \, w_{j-1} - 2w_j^2)$$

$$= \tfrac{1}{2}\alpha \sum_{j=0}^{J-1} (2w_j \, w_{j+1} - w_j^2 - w_{j+1}^2)$$

$$= -\tfrac{1}{2}\alpha \sum_{j=0}^{J-1} (w_{j+1} - w_j)^2 \leq 0.$$

(we hebben hier gebruikt dat $w_0 = w_J = 0$).

Hieruit volgt

$$\sum_{j=1}^{J-1} v_{j,n+1}^2 \leq \sum_{j=1}^{J-1} v_{j,n}^2 \leq \ldots \leq \sum_{j=1}^{J-1} f_j^2 \, .$$

Het effect van een storing in de beginwaarden blijft dus begrensd, hetgeen asymptotische stabiliteit betekent.

Opgave. Ga na dat het stabiliteitsonderzoek op de manier van de vorige paragraaf leidt tot de stabiliteitsvoorwaarde

$$\left|\frac{1 - 2\alpha \sin^2 \tfrac{1}{2}m\pi h}{1 + 2\alpha \sin^2 \tfrac{1}{2}m\pi h}\right| \leq 1 \, ,$$

waaraan voldaan is voor iedere waarde van h en k. Dus ook op deze manier zien we dat de methode van Crank-Nicolson onvoorwaardelijk asymptotisch stabiel is.□

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 $v_{j,n+1}$ uit een niet lineair stelsel iteratief opgelost moeten worden (zo mogelijk gebeurt dit oplossen weer met een variant van het Newton proces, zoals in 3.5.2. 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 asymptotisch instabiel te zijn.

4.2. De golfvergelijking

Het ligt voor de hand de golfvergelijking

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} + f(x,t) \tag{1}$$

te discretiseren door

$$\frac{1}{k^2}(u_{j,n+1} - 2u_{j,n} + u_{j,n-1}) = \frac{c^2}{h^2}(u_{j+1,n} - 2u_{j,n} + u_{j-1,n})$$

$$+ f_{j,n} + R_{j,n}$$

met $|R_{j,n}| \le C_1 h^2 + C_2 k^2$.

Hieruit volgt als methode

$$v_{j,n+1} - 2v_{j,n} + v_{j,n-1} = \alpha^2(v_{j+1,n} - 2v_{j,n} + v_{j-1,n}) + k^2 f_{j,n} \tag{2}$$

Hierin is $\alpha = kc/h$.

Als de beginvoorwaarden zijn

$$u(x,0) = \varphi(x),\ \frac{\partial u}{\partial t}(x,0) = \psi(x), \tag{3}$$

dan zijn passende beginvoorwaarden voor de differentievergelijking

$$v_{j,0} = \varphi_j,$$

$$v_{j,1} - v_{j,-1} = 2k\psi_j. \tag{4}$$

Samen met (2) (met $n = 0$) levert de tweede voorwaarde

$$v_{j,1} = \varphi_j + k\psi_j + \tfrac{1}{2}\alpha^2(\varphi_{j+1} - 2\varphi_j + \varphi_{j-1}) + \tfrac{1}{2}k^2 f_{j,0}. \tag{5}$$

Voor het stabiliteitsonderzoek gaan we op dezelfde manier te werk als in 4.1.1.

We nemen ter vereenvoudiging aan dat $f(x,t) = 0$. Een elementaire oplossing van (2) met $f_{j,n} = 0$ is

$$v_{j,n} = e^{\beta_m t_n} \sin m\pi x_j \ , \tag{6}$$

waarbij β_m bepaald is door de volgende voorwaarde

$$\tfrac{1}{2}(e^{\beta_m k} + e^{-\beta_m k}) = 1 - 2\alpha^2 \sin^2 \tfrac{1}{2} m\pi h \tag{7}$$

(β_m mag ook complex zijn).

De oplossing van het gediscretiseerde probleem is dan te schrijven als

$$v_{j,n} = \sum_{m=1}^{\infty} (C_m e^{\beta_m t_n} + D_m e^{-\beta_m t_n}) \sin m\pi x_j \ , \tag{8}$$

waarbij de coefficienten C_m en D_m bepaald zijn door de voorwaarden (4) en (5). Voor stabiliteit is vereist dat de termen van de reeks in (8) begrensd zijn voor h en k naar nul. Daarvoor is nodig en voldoende dat $|e^{\pm\beta_m k}| \leq 1$, (dus β_m moet zuiver imaginair zijn) en dit is het geval als

$$|1 - 2\alpha^2 \sin^2 \tfrac{1}{2} m\pi h| \leq 1 \ . \tag{9}$$

Aan deze stabiliteitsvoorwaarde is voldaan als

$$\alpha = \frac{ck}{h} \leq 1 \ . \tag{10}$$

Dus de methode is voorwaardelijk asymptotisch stabiel.

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 $t_n \leq T$ en voor alle j 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 we naar de exacte oplossing kijken van het beginwaardeprobleem

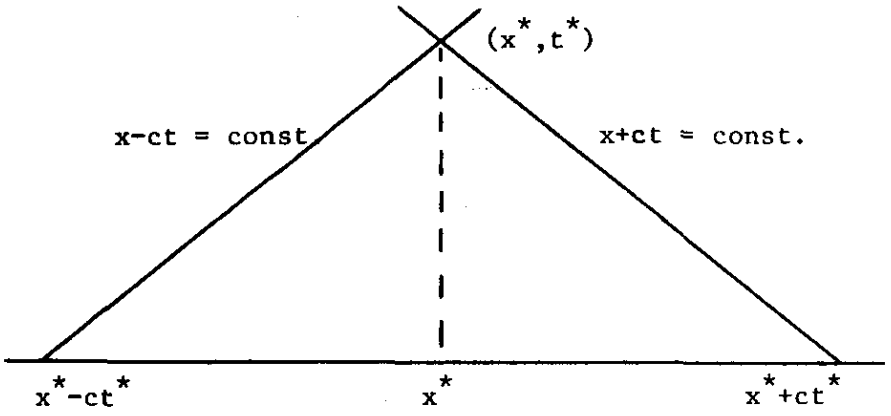
$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \ , \ -\infty < x < \infty$$

$$u(x,0) = \varphi(x), \ \frac{\partial u}{\partial t}(x,0) = \psi(x) \ .$$

Deze oplossing kunnen we als volgt schrijven

$$u(x,t) = \tfrac{1}{2}\{\varphi(x + ct) + \varphi(x - ct)\} + \frac{1}{2c}\int_{x-ct}^{x+ct} \psi(\xi)d\xi \ . \qquad (11)$$

De waarde van $u(x^*,t^*)$ is dus bepaald door de waarden van de beginfuncties $\varphi(x)$ en $\psi(x)$ op het interval $|x - x^*| \le ct^*$.



Dit interval noemen we het afhankelijkheids-gebied van (x^*,t^*). Anderzijds volgt uit (2), (4) en (5) dat de waarde $v_{j,n}$ bepaald is door de waarden φ_i en ψ_i met $|i-j| \le n$, dus door de waarden $\varphi(\xi_i)$ en $\psi(\xi_i)$ met $|\xi_i - x_j| \le nh = ct_n/\alpha$.

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 $\varphi(x)$ en $\psi(x)$ met $|x-x^*| < ct^*/\alpha$, dus uit een kleiner interval dan bij de differentiaalvergelijking. Maar dan kan nooit voor alle φ en ψ deze limiet gelijk zijn aan $u(x^*,t^*)$. Neem bijv. $\varphi(x) = 0$ voor alle x en $\psi(x) = 0$ voor $|x-x^*| \le ct^*/\alpha$, $\psi(x) > 0$ voor $|x-x^*| > ct^*/\alpha$.

Opmerking. Als we $\alpha = 1$ nemen, dan blijkt uit (2), (4) en (5) dat $v_{j,n}$ alleen afhangt van φ_{j-n} en φ_{j+n} en van ψ_i met $|i-j| < n$, geheel in overeenstemming met (11).

Als we als discretisatie van (1) een impliciete formule nemen, bijvoorbeeld

$$\frac{1}{k^2}(v_{j,n+1} - 2v_{j,n} + v_{j,n-1}) = \frac{c^2}{2h^2}(v_{j+1,n+1} - 2v_{j,n+1} + v_{j-1,n+1} + v_{j+1,n-1} - 2v_{j,n-1} + v_{j-1,n-1}) + f_{j,n} \ , \qquad (12)$$

dan moeten we per tijdstap weer een stelsel lineaire vergelijkingen met een tridiagonale matrix oplossen.
Daar staat tegenover dat de met (7) overeenkomende voorwaarde nu is

$$\tfrac{1}{2}(e^{\beta_m k} + e^{-\beta_m k}) = \frac{1}{1 + 2\alpha^2 \sin^2 \tfrac{1}{2} m\pi h} , \qquad (13)$$

zodat de daarbij behorende stabiliteitsvoorwaarde nu luidt

$$\left| \frac{1}{1 + 2\alpha^2 \sin^2 \tfrac{1}{2} m\pi h} \right| \leq 1 .$$

Hieraan is voldaan voor iedere waarde van h en k. Dus de methode (12) is onvoorwaardelijk stabiel.

4.3. 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

$$\frac{1}{h^2} (4u_{j,n} - u_{j,n+1} - u_{j,n-1} - u_{j+1,n} - u_{j-1,n}) = f_{j,n} + R_{j,n} \qquad (2)$$

met $|R_{j,n}| \leq C_1 h^2$.
(Het ligt in dit geval 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.

In het algemeen is het stelsel zeer groot, maar de matrix van het stelsel heeft slechts weinig elementen die niet nul zijn. Een dergelijke matrix wordt een "sparse matrix" genoemd.
Het oplossen van dit stelsel kan gebeuren met speciale eliminatiemethoden voor "sparse matrices". (Zie hoofdstuk 5)
Een andere mogelijkheid is een iteratieve methode. Daartoe schrijven we de vergelijkingen (3) in de vorm

$$v_{j,n} = \tfrac{1}{4}(v_{j,n+1} + v_{j,n-1} + v_{j+1,n} + v_{j-1,n} + h^2 f_{j,n}) \qquad (4)$$

en passen successieve substitutie toe. Daarbij kunnen we op verschillende manieren te werk gaan.

a) In de ν-de iteratieslag worden bij gegeven $v_{i,m}^{(\nu-1)}$ met behulp van (4) de waarden $v_{j,n}^{(\nu)}$ berekend. Dit is de zg. <u>methode van Jacobi</u>. Men kan bewijzen dat deze methode lineair convergeert met asymptotische convergentiefactor

$$\lambda_J := 1 - \sin^2 \frac{\pi}{2M} - \sin^2 \frac{\pi}{2N} \sim 1 - \frac{\pi^2}{4} \cdot \frac{M^2 + N^2}{M^2N^2} \sim \exp(-\frac{\pi^2}{2} \cdot \frac{M^2 + N^2}{2M^2N^2}) .$$

Dit betekent dat, als $M \sim N$, circa $2N^2/\pi^2$ slagen gedaan moeten worden om de fout met een factor e te doen afnemen.

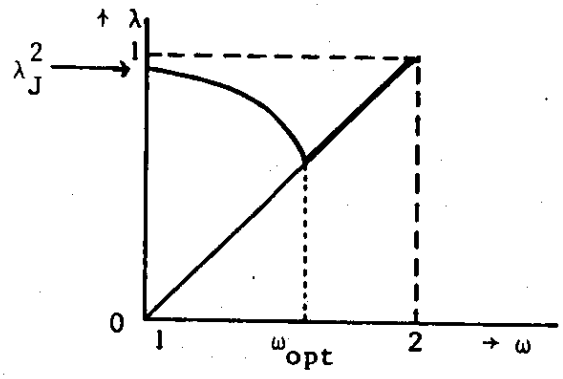
b) In de ν-de slag wordt $v_{j,n}^{(\nu)}$ met behulp van (4) berekend,waarbij in het rechterlid voor iedere term de laatst berekende waarde wordt genomen, dus met name $v_{i,m}^{(\nu)}$ als het punt (i,m) in de ν-de slag al aan de beurt geweest is en $v_{i,m}^{(\nu-1)}$ als het punt (i,m) nog niet aan de beurt geweest is. Dit is de <u>methode van Gauss-Seidel</u>. Men kan bewijzen dat deze methode eveneens lineair convergeert en dat de asymptotische convergentiefactor λ_J^2 is. De convergentie is dus ongeveer twee maal zo snel als bij Jacobi.

c) Essentieel sneller convergeert de variant van (4) die we krijgen door (4) te schrijven als

$$v_{j,n} = v_{j,n} + \frac{\omega}{4}(v_{j,n+1} + v_{j,n-1} + v_{j+1,n} + v_{j-1,n} - 4v_{j,n} + h^2 f_{j,n})$$

en daarop Gauss-Seidel toe te passen, na een geschikte keuze van ω.

Dit proces heet systematische overrelaxatie (S.O.R.), ω heet de overrelaxatiefacto
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 de hierbij behorende convergentiefactor 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

$$\lambda_{opt} \sim 1 - 2\pi\sqrt{\frac{M^2 + N^2}{2M^2N^2}} \sim \exp(-2\pi\sqrt{\frac{M^2 + N^2}{2M^2N^2}})$$

zodat voor $M \sim N$ nu ca.$N/(2\pi)$ slagen gedaan moeten worden om een factor e 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.

5. Lineaire vergelijkingen ([2], 5.1-5.6; [7]; [16], ch. 12)

5.1. Inleiding

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

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

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

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

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

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

(ontwikkeling naar de eerste rij waardoor de berekening van $n \times n$-determinanten teruggebracht is tot de berekening van $(n-1) \times (n-1)$-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.

Zij nl. het aantal vermenigvuldigingen *) nodig om met behulp van deze regels een n × 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 geldt

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

waaruit volgt dat voor n ≥ 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 $20! \sim 2.4_{10}18$ vermenigvuldigingen nodig zijn. Met een snelle automatische machine met een vermenigvuldigtijd van 10^{-6} sec. zou men dus ca $2.4_{10}12$ sec $\sim 7.7_{10}4$ jaar nodig hebben!

Later zullen we een methode aangeven, waarbij voor de berekening van een n × n-determinant slechts ca $\frac{1}{3} n^3$ vermenigvuldigingen nodig zijn. Ook dan blijft de regel van Cramer niet aanbevelenswaardig, aangezien een n × 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 de 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).

(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, omdat bepaling van A^{-1} ca. n^3 vermenigvuldigingen kost.

Men kan 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 $A\underline{x} = \underline{0}$ uitsluitend $\underline{x} = \underline{0}$ 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.

5.2. Directe methoden ([2], 5.3)

5.2.1. Triangulaire stelsels

Beschouw een stelsel 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 $\underline{Lx} = \underline{b}$, waarin L een onder-driehoeksmatrix is ($L_{ij} = 0$ voor $j > i$).

5.2.2. 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

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

waarin

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

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

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):

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

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)

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

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 algorithme 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 coëfficiëntenmatrix 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
```

5.2.3. Pivot strategieën

In 5.2.2 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

$$\begin{bmatrix} 2 & -2 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix}$$

krijgen we na de eerste slag van de eliminatie

$$\begin{bmatrix} 2 & -2 & 1 \\ 0 & 0 & .5 \\ 0 & 2 & -.5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ .5 \\ 3.5 \end{bmatrix}$$

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.

Deze strategie kan alleen mislukken als $A_{pk}^{(k-1)} = 0$ voor alle $p \geq k$. In dit geval bevatten de vergelijkingen k t/m n, zoals ze na de eerste k-1 slagen overgebleven zijn, de onbekende x_k niet. Deze behoeft dus ook niet geëlimineerd te worden, zodat de k-de slag nu uit niets anders bestaat dan uit het opnemen van de k-de vergelijking in het triangulaire stelsel $U\underline{x} = \underline{c}$. Daar nu $U_{kk} = 0$, geldt $\det(U) = 0$. Het stelsel $U\underline{x} = \underline{c}$ is dus singulier en in het algemeen niet oplosbaar, hetzelfde geldt voor het originele stelsel $A\underline{x} = \underline{b}$.

Opgave

Bewijs dat na uitvoering van de gemodificeerde algorithme geldt

$$\det(A) = (-1)^v \det(U) = (-1)^v U_{11} \ldots U_{nn} ,$$

waarin v het aantal uitgevoerde verwisselingen is.

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

$$\begin{pmatrix} 0.001 & 1 \\ -1 & 0.5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} . \qquad (1)$$

Met de eliminatiemethode van Gauss vinden we als gereduceerd stelsel

$$\begin{pmatrix} 0.001 & 1 \\ 0 & 1000.5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1000 \end{pmatrix} , \qquad (2)$$

waaruit bij exact rekenen volgt

$$x_2 = \frac{1000}{1000.5} \sim 0.99950 ,$$

$$x_1 = (1 - \frac{1000}{1000.5})/0.001 = \frac{500}{1000.5} \sim 0.49975 .$$

Stel nu echter dat we werken met een decimale machine met een mantisselengte van 4 decimalen. Dan is het getal $A_{22}^{(1)} = 1000.5$ uit (2) geen machinegetal en het wordt afgerond tot $\bar{A}_{22}^{(1)} = 1000$ of tot $\bar{\bar{A}}_{22}^{(1)} = 1001$, afhankelijk van het afrondmechanisme. In het eerste geval vinden we als gereduceerd stelsel

$$\begin{pmatrix} 0.001 & 1 \\ 0 & 1000 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1000 \end{pmatrix} \qquad (\bar{2})$$

met als oplossing

$$\bar{x}_2 = \frac{1000}{1000} = 1 , \quad \bar{x}_1 = \frac{1 - 1}{0.001} = 0 ,$$

in het tweede geval levert de machine als oplossing

$$\bar{\bar{x}}_2 = \frac{1000}{1001} = 0.9990, \quad \bar{\bar{x}}_1 = \frac{1 - 0.9990}{0.001} = 1 .$$

(bij exact oplossen van het stelsel $(\bar{\bar{2}})$ zouden we vinden $\bar{\bar{x}}_1 = \bar{\bar{x}}_2 = 1000/1001$, dat scheelt dus niet veel).
In beide gevallen vinden we dus in x_2 een relatieve fout van 0.5 ‰, dus even groot als de gemaakte relatieve fout in $A_{22}^{(1)}$, in x_1 is de relatieve fout in beide gevallen 100%. Of, als we liever niet naar individuele componenten doch naar $\|\underline{x} - \bar{\underline{x}}\|_\infty$ als maat voor de fout kijken, dan vinden we

$$\frac{\|\underline{x} - \bar{\underline{x}}\|}{\|\underline{x}\|} \sim \frac{\|\underline{x} - \bar{\bar{\underline{x}}}\|}{\|\underline{x}\|} \sim 0.5 \ .$$

Hoe is deze grote relatieve fout te verklaren? We geven twee beschouwingen, waarvan de eerste meer in detail beschrijft wat er gebeurt, terwijl de tweede globaler is doch meer inzicht geeft in het ontstaan van de ellende en daarmee tevens een remedie suggereert.

a. Tijdens de eliminatie wordt uitsluitend in de coëfficiënt $A_{22}^{(1)}$ een afrondfout gemaakt en deze bedraagt 0.5 ‰ . Daar de bepaling van $\bar{x}_2$ verder exact verloopt is de fout in $\bar{x}_2$ ook 0.5 ‰ . De formule waarmee x_1 bepaald wordt luidt

$$x_1 = \frac{1 - x_2}{0.001} \ .$$

Als geen nieuwe afrondfout gemaakt wordt dan wordt de fout δx_1 in x_1 geheel veroorzaakt door de fout δx_2 in x_2, voor het verband tussen de relatieve fouten geldt

$$\frac{\delta x_1}{x_1} = - \frac{x_2}{1 - x_2} \cdot \frac{\delta x_2}{x_2} \ .$$

En daar $x_2/(1 - x_2) \sim 2000$ "verklaart" dit de relatieve fout van 100% in x_1. Vaak zegt men: bij de bepaling van x_1 treedt "cijferverlies" op. Willen we x_1 en dus ook $1 - x_2$ tot op k cijfers nauwkeurig bepalen dan moet (als $x_2 \sim 0.9995$) x_2 tot op minstens $k + 3$ cijfers nauwkeurig bepaald zijn (inderdaad, nemen we $x_2 = 0.99950025$ dan wordt $x_1 = 0.49975$).

b. In het voorbeeld is $A_{22} = 0.5$, $U_{12} = 1$, $L_{21} = -1000$, zodat we voor $A_{22}^{(1)} = A_{22} - L_{21}U_{12}$ vinden na afronding $\bar{A}_{22}^{(1)} = 1000$ of $\bar{\bar{A}}_{22}^{(1)} = 1001$. Het eerste resultaat zouden we bij exact rekenen gevonden hebben indien $A_{22} = 0$ geweest was, het tweede indien $A_{22} = 1$ geweest was. Dat betekent dat het stelsel $(\bar{2})$ exact hoort bij een stelsel $(\bar{1})$ waarin $\bar{A}_{22} = 0$ en $(\bar{2})$ hoort bij $(\bar{\bar{1}})$ met $\bar{\bar{A}}_{22} = 1$. Daar de oplossing van het stelsel

$$\begin{pmatrix} 0.001 & 1 \\ -1 & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

is

$$x_1 = \frac{A_{22}}{1 + 0.001\ A_{22}} \ , \quad x_2 = \frac{1}{1 + 0.001\ A_{22}} \ ,$$

leidt een verandering van 100% in A_{22} tot een verandering van 100% in x_1 (en -toevallig- slechts tot een geringe verandering in x_2).

Ook hier kunnen we spreken van nauwkeurigheids-, of, liever, informatieverlies. De nauwkeurige waarde van A_{22} speelt bij de bepaling van $A_{22}^{(1)}$ geen rol, voor iedere $\bar{A}_{22}$ uit $-0.05 < \bar{A}_{22} < 0.5$ wordt $\bar{A}_{22}^{(1)} = 1000$, voor ieder $\bar{\bar{A}}_{22}$ uit $0.5 < \bar{\bar{A}}_{22} < 1.5$ wordt $\bar{\bar{A}}_{22}^{(1)} = 1001$ (ga na).
Losjes: de vier decimalen waarmee A_{22} gegeven is gaan vrijwel verloren in de "afrondruis". En omdat een verandering van A_{22} een vrijwel gelijke verandering van x_1 ten gevolge heeft, is hiermee ook de kans op een nauwkeurige bepaling van x_1 verkeken.

De kern van deze beschouwing ligt echter in de opmerking dat het stelsel $(\bar{2})$ en de coëfficiënt L_{21} beschouwd kunnen worden als exact resultaat behorend bij een "naburig" stelsel $(\bar{1})$. We definiëren de coëfficiënt $\bar{A}_{22}$ daarin door

$$\bar{A}_{22}^{(1)} = \bar{A}_{22} - L_{21}U_{12} .$$

Daar het product $L_{21}U_{12}$ toevallig exact berekend wordt geldt

$$\bar{A}_{22}^{(1)} = (A_{22} - L_{21}U_{12})(1 + \varepsilon) = A_{22}^{(1)}(1 + \varepsilon)$$

met $\varepsilon = -0.0005$ (vergelijk dit met 0.3, daaruit zou volgen $|\varepsilon| \leq 0.0005$) en dus is

$$\bar{A}_{22} = A_{22} + \varepsilon\,(A_{22} - L_{21}U_{12}) = A_{22} + \varepsilon A_{22}^{(1)} .$$

Het verschil tussen $\bar{A}_{22}$ en A_{22} is groot omdat $A_{22}^{(1)}$ groot is en dat is weer een gevolg van het feit dat L_{21} groot is. Dit levert meteen de suggestie voor een remedie: zorg dat $A_{22}^{(1)}$ (of, als $n > 2$, de elementen van $A^{(1)},A^{(2)},\ldots,A^{(n-1)}$) niet veel groter worden dan die van A, opdat het na eliminatie verkregen stelsel exact hoort bij een zeer naburig stelsel $\bar{A}x = \bar{b}$ (dit zal in 5.2.5.1 verder uitgewerkt worden). Meestal kunnen we dit bereiken door de pivots zo te kiezen dat de elementen L_{ij} kleiner dan 1 blijven.

In ons voorbeeld kunnen we $A_{22}^{(1)}$ beperkt houden door de eerste en de tweede vergelijking te verwisselen:

$$\begin{pmatrix} -1 & 0.5 \\ 0.001 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \tag{3}$$

hetgeen na eliminatie levert (met $L_{21} = -0.001$)

$$\begin{pmatrix} -1 & 0.5 \\ 0 & 1.0005 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \tag{4}$$

of met afronding $\bar{A}_{22}^{(1)} = 1$, resp. $\bar{\bar{A}}_{22}^{(1)} = 1.001$. Uit de stelsels $(\bar{4})$ of $(\bar{\bar{4}})$ volgen x_2 en x_1, met volledig acceptabele nauwkeurigheid.

In het algemeen blijkt dat de volgende <u>pivotstrategie</u> 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). Er geldt dan $|L_{i1}| \leq 1$ $(i = 2, \ldots, n)$. Dit blijkt van groot belang voor de foutenanalyse te zijn (zie 5.2.5).

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] := if max > 0 then A[i,j]/U[k,k] else 0;
            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 - als geen der $U_{kk} = 0$ - 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 $\underline{U}\underline{x} = \underline{c}$ anders geschreven dan in 5.2.1, namelijk analoog aan de wijze waarop de vector $\underline{c}$ uit de vector $\underline{b}$ verkregen wordt; ga na dat hierdoor in feite niets verandert, precies dezelfde bewerkingen of dezelfde getallen worden uitgevoerd - alleen de volgorde in de tijd is anders).

5.2.4. LU-decompositie. De algoritme van Crout ([2], 5.3.4 en 5.3.5)

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

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

$$L_{ik} = A_{ik}^{(k-1)}/U_{kk} = \left(A_{ik} - \sum_{\ell=1}^{k-1} L_{i\ell} \times U_{\ell k} \right) / U_{kk} , \quad i > k, j = k . \tag{1}$$

Definieer nu

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

Dan kunnen we (1) ook schrijven als

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

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

Dus ook als

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

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

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

Dan is (2) equivalent met

$$A = LU \; . \tag{4}$$

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.

Analoog aan het bovenstaande geldt voor de bewerkingen die bij de Gauss-algoritme op de rechterleden uitgevoerd moeten worden dat

$$b_i^{(k)} = b_i - \sum_{\ell=1}^{k} L_{i\ell} \times c_\ell \; , \qquad 0 \le k < i \le n \; ,$$

en dus

$$c_k = b_k^{(k-1)} = b_k - \sum_{\ell=1}^{k-1} L_{k\ell} \times c_\ell \; , \quad 1 \le k \le n \; , \tag{5}$$

hetgeen we (daar $L_{kk} = 1$) kunnen schrijven als

$$b_k = \sum_{\ell=1}^{k} L_{k\ell} \times c_\ell \; .$$

Dat wil zeggen: de in de Gauss-algoritme verkregen vector $\underline{c}$ is de oplossing van het driehoeksstelsel

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

Dit correspondeert met het feit dat als A = LU met reguliere L, het stelsel $A\underline{x} = \underline{b}$ equivalent is met de stelsels $L\underline{c} = \underline{b}$ en $U\underline{x} = \underline{c}$.

We merken nu op dat we de bewerkingen die nodig zijn om U_{kj} en L_{ik} te bepalen, even goed kunnen uitvoeren in een volgorde 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[k,j];
            for ℓ := 1 step 1 until k - 1 do s := s - L[k,ℓ] × U[ℓ,j];
            U[k,j] := s
      end;
      for i := k + 1 step 1 until n do
      begin s := A[i,k];
            for ℓ := 1 step 1 until k - 1 do s := s - L[i,ℓ] × U[ℓ,k];
            L[i,k] := s/U[k,k]
      end
end
```

En de bewerkingen die bij de Gauss-algoritme op de rechterleden worden uitgevoerd, kunnen we uitvoeren in een volgorde die bij (5) aansluit:

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

(vergelijk dit met 5.2.1).

We hebben hiermee de zg. algoritme van Crout verkregen. Merk op dat de elementen L, U en c 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 $A_{kj}^{(\ell)}$, $A_{ik}^{(\ell)}$ en $b_k^{(\ell)}$ expliciet in de arrays A en b genoteerd worden.

Dit heeft enkele 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 (bij machines met een zg. dubbellengte optelregister).
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. Analoog voor de rechterleden.

Ook bij de uitvoering van de algoritme van Crout kan men rijverwisselingen toepassen (en het is, ook als geen der pivots nul wordt, gewenst dit te doen opdat $|L_{ik}| \leq 1$ wordt). Hiertoe merken we op dat bij de Gauss-algoritme bij het begin van de k-de slag gekeken werd welk van de getallen $|A_{ik}^{(k-1)}|$, $i = k,\ldots,n$, het grootste is. Daar

$$A_{ik}^{(k-1)} = A_{ik} - \sum_{\ell=1}^{k-1} L_{ik} \times U_{kj}$$

moeten we bij de k-de stap van de Crout-algoritme dus beginnen met deze getallen te bepalen.

De algoritme wordt dan (geschreven in een nog wat compactere vorm van pseudo-ALGOL en met een procedure wissel die de waarden van de meegegeven actuele parameters verwisselt):

```
for k := 1 step 1 until n do
begin max := 0;
      for i = k step 1 until n do
      begin s_i := A_ik - Σ(ℓ=1..k-1) L_iℓ × U_ℓk;
            if abs(s_i) > max then begin pk := i; max := abs(s_i) end;
      end;
```

```
    p[k] := pk;
    if pk > k then
    begin for j := 1 step 1 until k - 1 do wissel (L_kj, L_pk,j);
          wissel (s_k, s_pk);
          for j := k + 1 step 1 until n do wissel (A_kj, A_pk,j)
    end;
    U_kk := s_k;
    for j := k + 1 step 1 until n do U_kj := A_kj - Σ_{ℓ=1}^{k-1} L_kℓ × U_ℓj;
    for i := k + 1 step 1 until n do L_ik := s_i/U_kk
end
```

Merk op dat we bij deze algoritme de elementen L_{kj} en $L_{pk,j}$ $(j = 1,\ldots,k-1)$ zo nodig wel mee verwisselen. Hierdoor bereiken we dat de verkregen matrices L en U behoren bij de matrix A' (in de zin dat A' = LU) die uit A verkregen wordt door op de rijen van A de diverse verwisselingen uit te voeren, dwz. door uitvoering van

```
for k := 1 step 1 until n do
if p[k] > k then wissel (k-de rij van A, p[k]-de rij van A)
```

Het gevolg is dat de reductie van een stelsel $\underline{A}\underline{x} = \underline{b}$ naar de vorm $\underline{U}\underline{x} = \underline{c}$ nu kan luiden: bepaal eerst $\underline{b}'$ uit $\underline{b}$ door de nodige verwisselingen uit te voeren en daarna $\underline{c}$ uit $\underline{L}\underline{c} = \underline{b}'$.
Uitgewerkt (we vlechten de twee stappen weer enigszins in elkaar):

```
for k := 1 step 1 until n do
begin if p[k] > k then wissel (b_k , b_p[k]);
      c_k := b_k - Σ_{ℓ=1}^{k-1} L_kℓ × c_ℓ
end
```

Opgave

Laat zien dan men, net als in opgave 3) op pag. 5.7, de elementen L_{ij} $(j < i)$ en U_{ij} $(j \geq i)$ zonder bezwaar op de plaatsen van de elementen A_{ij} kan schrijven, analoog voor c_i en b_i.

5.2.5. Foutenanalyse en gevoeligheidsanalyse.([2], 5.5; [17] sec.21)

We onderzoeken nu de invloed van afrondfouten op het resultaat van de decompositie A = LU en van het oplossen van een stelsel $A\underline{x} = \underline{b}$ indien de algoritmen van Gauss of Crout met partial pivoting uitgevoerd worden op een computer die met eindige precisie werkt.
In ca. 1960 liet Wilkinson zien dat men dit onderzoek het best in twee stappen kan uitvoeren:

1. Men toont aan dat de verkregen L en U horen bij de exacte decompositie van de matrix A + E die "dicht" bij A ligt; eveneens dat de verkregen $\underline{x}$ de exacte oplossing is van een "naburig" stelsel $(A + F)\underline{x} = \underline{b}$. Het blijkt dat men voor $\|E\| / \|A\|$ en $\|F\| / \|A\|$ theoretische bovengrenzen kan vinden die als regel niet veel groter zijn dan een aantal malen $n^2\eta$, waarin η de relatieve machinenauwkeurigheid is (in feite is de waarde van deze verhoudingen zelden groter dan $n\eta$).
   Hiermee is de numerieke stabiliteit (zie pag. 0.7) van de algoritmen aangetoond.
2. Men onderzoekt hoeveel de oplossing van een stelsel $A\underline{x} = \underline{b}$ verandert als A en $\underline{b}$ vervangen worden door naburige $A + \delta A$ en $\underline{b} + \delta\underline{b}$. Is dit weinig, dan is het stelsel goed geconditioneerd, anders slecht geconditioneerd (zie pag. 0.7). De mate van slecht geconditioneerd zijn kan worden uitgedrukt met behulp van een zg. conditiegetal dat ≥ 1 is en een bovengrens geeft voor de verhouding tussen de relatieve verandering in $\underline{x}$ en de relatieve veranderingen in A, resp. $\underline{b}$.

Combinatie van deze twee stappen levert dat bij een goed geconditioneerd stelsel de invloed van afrondfouten in de Crout- of Gauss-algoritme met partial pivoting gering is (in die zin dat de relatieve fout in $\underline{x}$ niet veel groter is dan een aantal malen $n\eta$). Bij een slecht geconditioneerd stelsel kan de invloed van afrondfouten aanzienlijk zijn, echter (omdat ook daar de algoritme wel numeriek stabiel is!) niet essentieel groter dan het effect van veranderingen in de elementen van A of $\underline{b}$ ter grootte van een aantal malen de relatieve machinenauwkeurigheid.
Voor veel practische situaties betekent dit: als de elementen van A en/of $\underline{b}$ niet exact bekende doch gemeten grootheden zijn, dan wordt de nauwkeurigheid van de oplossing geheel bepaald door de meetfouten (aangenomen dat $n\eta$ klein is ten opzichte van de relatieve fout in de metingen).

5.2.5.1. De invloed van afrondfouten.

We bekijken de in 5.2.4. besproken algoritmen ter oplossing van het stelsel $\underline{A}\underline{x} = \underline{b}$. We nemen daarbij aan dat geen verwisselingen nodig zijn (cq., dat de nodige verwisselingen reeds uitgevoerd zijn).
We merken nu op dat de berekening van de grootheden U_{kj}, L_{ik}, c_k en x_k volgens eenzelfde recept verloopt.
In de k-de slag van de algoritme van Crout worden U_{kj} $(j \geq k)$ en L_{ik} $(i > k)$ bepaald uit de formules

$$A_{kj} = \sum_{\ell=1}^{k} L_{k\ell}U_{\ell j} \ , \ A_{ik} = \sum_{\ell=1}^{k} L_{i\ell}U_{\ell k} \ , \tag{1}$$

waarin, behalve U_{kj}, resp. L_{ik}, alles bekend is.
In de k-de slag van het oplossen van $\underline{L}\underline{c} = \underline{b}$ wordt c_k bepaald uit

$$b_k = \sum_{\ell=1}^{k} L_{k\ell}c_\ell \ . \tag{2}$$

Op dezelfde wijze wordt $\underline{U}\underline{x} = \underline{c}$ opgelost door x_k te bepalen uit

$$c_k = \sum_{\ell=k}^{n} U_{k\ell}x_\ell \ . \tag{3}$$

Ook in (2), resp. (3), is behalve c_k, resp. x_k, alles bekend.

Dit zijn allemaal processen van de vorm:
bepaal z_k uit de vergelijking

$$\sum_{\ell=1}^{k} y_\ell z_\ell = a \ , \tag{4}$$

waarin a, $y_1,\ldots,y_k$, $z_1,\ldots,z_{k-1}$ bekend zijn en $y_k \neq 0$.
We zullen aan het eind van deze paragraaf aantonen dat, als we werken met een machine die in iedere bewerking een afrondfout maakt die in relatieve zin kleiner is dan de machine-nauwkeurigheid η, voor de verkregen z_k exact geldt dat

$$\sum_{\ell=1}^{k} y_\ell z_\ell(1 + \varepsilon_\ell) = a \tag{5}$$

met (als we afzien van termen van de orde η^2)

$$|\varepsilon_\ell| \leq \ell\eta \ .$$

Dat wil zeggen dat de verkregen z_k exact voldoet aan een "naburige" vergelijking.

Passen we dit toe op de formules (1) t/m (3) dan vinden we dat de verkregen L, U, $\underline{c}$ en $\underline{x}$ exact voldoen aan relaties

$$A_{ij} = \sum_{\ell=1}^{\min(i,j)} L_{i\ell}U_{\ell j}\,(1 + \varepsilon_{i\ell j}) \text{ met } |\varepsilon_{i\ell j}| \le \ell\eta \ ,$$

$$b_k = \sum_{\ell=1}^{k} L_{k\ell}c_\ell\,(1 + \varepsilon'_{k\ell}) \text{ met } |\varepsilon'_{k\ell}| \le \ell\eta \ ,$$

$$c_k = \sum_{\ell=k}^{n} U_{k\ell}x_\ell\,(1 + \varepsilon''_{k\ell}) \text{ met } |\varepsilon''_{k\ell}| \le (\ell - k + 1)\eta \ .$$

Hieruit volgt dat voor de verkregen L, U, $\underline{c}$ en $\underline{x}$ geldt

$$LU = A + E \text{ met } |E_{ij}| \le n\eta \sum_{\ell=1}^{\min(i,j)} |L_{i\ell}|\,|U_{\ell j}| \qquad (6)$$

$$\begin{aligned} &(L + E')\underline{c} = \underline{b} \text{ met } |E'_{ij}| \le n\eta\,|L_{ij}| \\ &(U + E'')\underline{x} = \underline{c} \text{ met } |E''_{ij}| \le n\eta\,|U_{ij}| \ . \end{aligned} \qquad (7)$$

Uit (6) volgt $|E_{ij}| \le n\eta(|L| \times |U|)_{ij}$ en hieruit volgt (in de maximum-norm)

$$\|E\| \le n\eta\|L\|.\|U\| \ .$$

Uit (7) volgt $\|E'\| \le n\eta\|L\|$, $\|E''\| \le n\eta\|U\|$. Uit (6), (7), (8) en (9) volgt derhalve

$$\underline{b} = (L + E')(U + E'')\underline{x} = (A + E + E'U + LE'' + E'E'')\underline{x} = (A + F)\underline{x} \qquad (8)$$

met (als we afzien van termen van de orde η^2)

$$\|F\| \le \|E\| + \|E'U\| + \|LE''\| + \|E'E''\| \le 3n\eta\|L\|\|U\| \ . \qquad (9)$$

We veronderstellen nu dat partial pivoting is toegepast. Dan is $|L_{ij}| \le 1$ en dus $\|L\| \le n$, zodat

$$\frac{\|F\|}{\|A\|} \le 3n^2\eta\,\frac{\|U\|}{\|A\|} \ . \qquad (10)$$

Het is duidelijk dat we hier een "worst case analysis" uitgevoerd hebben (alle afrondfouten maximaal ongunstig) en dat we daarna nog verschillende malen de bovengrenzen vervangen hebben door simpeler, maar grovere bovengrenzen. In de practijk blijkt dan ook dat als we na bepaling van L en U de matrix E := LU - A exact (of althans nauwkeurig genoeg) bepalen, de verhouding $\|E\| / \|U\|$ niet veel groter is dan $n\eta$ en vaak zelfs niet meer dan enkele malen η is. Iets dergelijks geldt voor $\|F\| / \|U\|$ als we na de bepaling van $\underline{x}$ het zg. residu $\underline{r} := \underline{b} - A\underline{x}$ nauwkeurig berekenen en dan geschreven denken als $\underline{r} = F\underline{x}$ met een F met zo klein mogelijke $\|F\|$ (het blijkt dat $\|F\| = \|\underline{r}\| / \|\underline{x}\|$ haalbaar is).

De conclusie uit het bovenstaande is: de algoritme van Crout met partial pivoting is numeriek stabiel in alle gevallen waarin $\|U\|$ niet veel groter is dan $\|A\|$.

Over de verhouding $\|U\| / \|A\|$ valt het volgende te zeggen

1. Men kan bewijzen dat bij partial pivoting altijd geldt dat $\|U\| / \|A\| \leq 2^{n-1}$ en dat er matrices zijn waarvoor deze grens willekeurig dicht benaderd wordt. Gelukkig blijkt echter dat bij de in de practijk voorkomende matrices $\|U\| / \|A\|$ zelden groter is dan n. Bij vele slecht geconditioneerde matrices is het zelfs zo dat de tweede en volgende rijen van U aanzienlijk kleiner zijn dan $\|A\|$, zodat dan $\|U\| / \|A\| \sim 1$.
2. Men kan na uitvoering van de decompositie de verhouding $\|U\| / \|A\|$ bepalen en kijken of hij verontrustend groot is.
3. Men kan in plaats van Gauss- of Crout-eliminatie met partial pivoting ook Gauss-eliminatie met zg. complete pivoting uitvoeren. Daarbij zoekt men in de eerste slag als pivot niet het grootste element uit de eerste kolom maar het grootste element uit de hele matrix en men verwisselt zo nodig rijen èn kolommen ; analoog bij de volgende slagen. Wilkinson bewees dat bij complete pivoting

$$\|U\| / \|A\| \leq 1.8n^{1+(\log n)/4}$$

en dat deze schatting steeds te pessimistisch is.

Appendix.

Formule (5) kan met behulp van de formules uit § 0.2, met name de formules (5) en (6), als volgt worden bewezen.
We beschouwen de vergelijking

$$\sum_{\ell=1}^{k} y_\ell z_\ell = a \ ,$$

waaruit, bij gegeven a, $y_1,\ldots,y_k$ en $z_1,\ldots,z_{k-1}$ de onbekende z_k bepaald moet worden.
Voeren we uit

$s := a;$
for $\ell := 1$ **step** 1 **until** $k - 1$ **do** $s := s - y_\ell \times z_\ell;$
$z_k := s/y_k$

dan geldt voor de verkregen waarden van s en z

$$s = a(1 + \beta_1) \ldots (1 + \beta_{k-1}) - x_1 y_1 (1 + \alpha_1)(1 + \beta_1) \ldots (1 + \beta_{k-1})$$
$$- \ldots - x_{k-1} y_{k-1} (1 + \alpha_{k-1})(1 + \beta_{k-1}) \ ,$$

en

$$s = y_k z_k (1 + \alpha_k) \ ,$$

waarbij $|\alpha_i| \leq \eta$ en $|\beta_i| \leq \eta$, $i = 1,2,\ldots,k$.

Na deling door $(1 + \beta_1)\ldots(1 + \beta_{k-1})$ volgt hieruit

$$a = y_1 z_1 (1 + \alpha_1) + y_2 z_2 \frac{1 + \alpha_2}{1 + \beta_1} + \ldots$$
$$+ y_k z_k \frac{1 + \alpha_k}{(1 + \beta_1)\ldots(1 + \beta_{k-1})}$$
$$= \sum_{\ell=1}^{k} y_\ell z_\ell (1 + \varepsilon_\ell)$$

met (als we termen met η^2 verwaarlozen)

$$|\varepsilon_\ell| \leq \ell\eta \ .$$

5.2.5.2. Gevoeligheidsanalyse.

Zij A een reguliere n×n matrix, $\underline{b}$ een n-vector en $\underline{x}$ de oplossing van

$$A\underline{x} = \underline{b} \ . \tag{1}$$

Zij δA en $\delta\underline{b}$ kleine storingen op A resp. $\underline{b}$ en $\underline{x} + \delta\underline{x}$ de oplossing van

$$(A + \delta A)(\underline{x} + \delta\underline{x}) = \underline{b} + \delta\underline{b} \ . \tag{2}$$

Kunnen we dan uitspraken doen over $\|\delta\underline{x}\| / \|\underline{x}\|$ in termen van $\|\delta A\| / \|A\|$ en $\|\delta\underline{b}\| / \|\underline{b}\|$?

Als $\delta A = 0$ dan geldt $\delta\underline{x} = A^{-1}\delta\underline{b}$ en dan is

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\|\,\|\underline{b}\|}{\|\underline{x}\|} \cdot \frac{\|\delta\underline{b}\|}{\|\underline{b}\|}$$

$$\leq \|A^{-1}\|\,\|A\| \cdot \frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \ ,$$

want $\underline{b} = A\underline{x}$, dus $\|\underline{b}\| \leq \|A\|\,\|\underline{x}\|$.
We noemen

$$c(A) := \|A^{-1}\|\,\|A\|$$

het conditiegetal van A, zodat we kunnen schrijven

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq c(A)\,\frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \ . \tag{3}$$

Het conditiegetal geeft dus aan hoeveel de relatieve verandering in $\underline{x}$ groter kan zijn dan de relatieve verandering in $\underline{b}$. Men kan bewijzen dat er bij iedere A een $\underline{b}$ en een $\delta\underline{b}$ zijn zo dat in (3) het gelijkteken geldt (neem nl. $\underline{b}$ zo dat $\|A\|\,\|\underline{x}\| = \|A\underline{x}\|$ en $\delta\underline{b}$ zo dat $\|A^{-1}\|\|\delta\underline{b}\| = \|A^{-1}\delta b\|$).
Daar $I = A^{-1}A$, dus $1 = \|I\| \leq \|A^{-1}\|\,\|A\|$ geldt steeds $c(A) \geq 1$.

Om de invloed van een verandering van A na te gaan bewijzen we eerst:
als $\|E\| < 1$ dan is $I + E$ regulier en

$$\|(I + E)^{-1}\| \leq \frac{1}{1 - \|E\|} \ . \tag{4}$$

Immers, als $I + E$ singulier is dan is er een $\underline{x} \neq \underline{0}$ zo dat $(I + E)\underline{x} = \underline{0}$, dus $\underline{x} = -E\underline{x}$, dus $\|\underline{x}\| \leq \|E\| \, \|\underline{x}\|$, dus (daar $\|\underline{x}\| > 0$) $\|E\| \geq 1$.
Zij nu $\|E\| < 1$ en $\underline{y} = (I + E)^{-1}\underline{x}$. Dan is $\underline{y} = \underline{x} - E\underline{y}$, dus $\|\underline{y}\| \leq \|\underline{x}\| + \|E\| \, \|\underline{y}\|$, dus $\|\underline{y}\| \leq \|\underline{x}\| \, / \, (1 - \|E\|)$. Voor iedere $\underline{x} \neq \underline{0}$ geldt dus

$$\frac{\|(I + E)^{-1}\underline{x}\|}{\|\underline{x}\|} \leq \frac{1}{1 - \|E\|} ;$$

met de definitie van de matrixnorm volgt hieruit (4).

Zij nu $\underline{x} + \delta\underline{x}$ de oplossing van (2) met $\delta\underline{b} = \underline{0}$. Dan is, als $\underline{x}$ aan (1) voldoet,

$$(A + \delta A)\delta\underline{x} = -\delta A\underline{x} ,$$

dus $$(I + A^{-1}\delta A)\delta\underline{x} = -A^{-1}\delta A\underline{x} .$$

Als $\|A^{-1}\delta A\| < 1$ dan volgt hieruit met (4)

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\delta A\|}{1 - \|A^{-1}\delta A\|} .$$

Als zelfs $\|A^{-1}\| \, \|\delta A\| < 1$ dan kunnen we hiervoor schrijven

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{\|A^{-1}\| \, \|\delta A\|}{1 - \|A^{-1}\| \, \|\delta A\|} = \frac{c(A)\|\delta A\| / \|A\|}{1 - c(A)\|\delta A\| / \|A\|} . \qquad (5)$$

In de practijk is meestal $c(A)\|\delta A\| / \|A\| << 1$, zodat dan ook bij storingen in A het conditiegetal $c(A)$ aangeeft hoeveel groter de relatieve verandering in $\underline{x}$ is dan de relatieve verandering in A. Men kan bewijzen dat er bij iedere A en $\underline{b}$ een δA is zo dat $\|\delta\underline{x}\| / \|\underline{x}\| = c(A)\|\delta A\| / \|A\|$.

In het geval dat zowel A als $\underline{b}$ variëren geldt (ga na)

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{c(A)}{1 - c(A)\|\delta A\|/\|A\|} \left\{ \frac{\|\delta A\|}{\|A\|} + \frac{\|\delta\underline{b}\|}{\|\underline{b}\|} \right\} \qquad (6)$$

De vraag rijst hoe we zonder veel moeite het conditiegetal $c(A)$ kunnen bepalen. De bepaling van A^{-1} is natuurlijk mogelijk maar duur (na de bepaling van de decompositie $A = LU$, die ca $\frac{1}{3}n^3$ vermenigvuldigingen en delingen kost, kost bepaling van $A^{-1} = U^{-1}L^{-1}$ nog eens ca $\frac{2}{3}n^3$ vermenigvuldigingen en delingen). Daarentegen is, als we L en U kennen, het oplossen van een extra stelsel $A\underline{z} = \underline{d}$ niet erg duur (ca n^2 vermenigvuldigingen en delingen). Voor ieder rechterlid $\underline{d}$ geldt dat $\|\underline{z}\| = \|A^{-1}\underline{d}\| \leq \|A^{-1}\| \, \|\underline{d}\|$ en dus

$$c(A) = \|A^{-1}\| \, \|A\| \geq \frac{\|\underline{z}\| \, \|A\|}{\|\underline{d}\|} . \qquad (7)$$

Als $\underline{d}$ een "random" karakter heeft dan is als regel de ongelijkheid $\|A^{-1}\underline{d}\| \leq \|A^{-1}\|\,\|\underline{d}\|$ niet zo erg ver van gelijkheid en dan wordt door (7) een redelijke schatting gegeven.

We bepalen nu met behulp van deze gevoeligheidsanalyse eerst een schatting voor de zg. onvermijdbare fout bij het oplossen van een stelsel $A\underline{x} = \underline{b}$, d.w.z. de fout die veroorzaakt kan worden door een relatieve verandering van de gegevens (de elementen van A en $\underline{b}$) ter grootte van de machineprecisie η (zie 0.3).
Veronderstel dus $|\delta A_{ij}| \leq \eta|A_{ij}|$, $|\delta b_i| \leq \eta|b_i|$. Dan is $\|\delta A\| \leq \eta\|A\|$, $\|\delta\underline{b}\| \leq \eta\|\underline{b}\|$ en dus geldt voor de bijbehorende verandering $\delta\underline{x}$ volgens (6)

$$\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{2\eta c(A)}{1 - \eta c(A)} \quad .$$

Hoewel het rechterlid in het algemeen niet gelijk is aan, maar een bovengrens is voor de relatieve verandering in de oplossing ten gevolge van relatieve veranderingen in de gegevens die niet groter zijn dan de machineprecisie, kunnen we dit rechterlid, of, eenvoudiger, de grootheid $\eta c(A)$, *toch als maat* voor de onvermijdbare fout beschouwen (daar als regel $c(A)$ wel enige malen *groter dan 1 is*, is het weinig belangrijk om, zoals in 0.3, ook nog de relatieve nauwkeurigheid waarmee het eindantwoord $\underline{x}$ voorgesteld kan worden in rekening te brengen).

Vervolgens bepalen we met (6) uit de resultaten van 5.2.5.1 een bovengrens voor de relatieve fout in de t.g.v. afrondfouten verkregen "oplossing" $\tilde{\underline{x}} = \underline{x} + \delta\underline{x}$ van het stelsel $A\underline{x} = \underline{b}$:

$$\left.\begin{aligned} &\frac{\|\delta\underline{x}\|}{\|\underline{x}\|} \leq \frac{c(A)\|F\|/\|A\|}{1 - c(A)\|F\|/\|A\|} \\ &\text{met} \\ &\frac{\|F\|}{\|A\|} \leq 3n^2\eta\,\frac{\|U\|}{\|A\|} \quad . \end{aligned}\right\} \qquad (8)$$

We constateren dat de schatting voor de relatieve fout in $\underline{x}$ tengevolge van afrondfouten ca. een factor $3n^2\,\|U\|/\|A\|$ groter is dan de hierboven gekozen maat voor de onvermijdbare fout. Hieruit volgt dat het oplossing van lineaire stelsels door middel van Gauss eliminatie met partial pivoting numeriek stabiel is in de zin van 0.3 voor die klasse van matrices waarvoor $\|U\|/\|A\|$ niet te groot is, bijv. $\|U\|/\|A\| \leq n$.

De bovengrens (8) is een theoretische bovengrens die als regel nogal pessimis-

tisch is. Beter is het, een zg. a posteriori foutenschatting te maken. Hiertoe bepaalt men, nadat een niet exacte oplossing $\tilde{\underline{x}}$ verkregen is, eerst het zg. residu

$$\underline{r} = \underline{b} - A\tilde{\underline{x}} \ . \qquad (9)$$

Dit moet in zg. dubbele lengte gebeuren, anders vindt men t.g.v. cijferverlies vaak geen enkel goed cijfer in $\underline{r}$.
Uit (9) volgt dat exact geldt

$$\underline{x} = A^{-1}\underline{b} = \tilde{\underline{x}} + A^{-1}\underline{r} \ .$$

Men bepaalt daarom (met behulp van de reeds bekende decompositie van A) de oplossing van het stelsel $A\underline{y} = \underline{r}$, waarmee een benadering $\tilde{\underline{y}}$ voor het verschil tussen $\underline{x}$ en $\tilde{\underline{x}}$ gevonden wordt. Vaak zal men als betere benadering voor de oplossing nu niet $\tilde{\underline{x}}$, doch $\tilde{\underline{x}} + \tilde{\underline{y}}$ accepteren (met helaas weer onbekende fout). Men noemt deze handelwijze (die eventueel enkele malen herhaald kan worden) naïtereren.
Als, zoals in 4.1.2.1. aangetoond is, $\tilde{\underline{x}}$ exact voldoet aan $(A+F)\tilde{\underline{x}} = \underline{b}$ dan is $\underline{r} = F\tilde{\underline{x}}$, zo dat $\underline{r}$ een min of meer random karakter heeft. Daaruit volgt dat, zoals in (7), $\|\tilde{\underline{y}}\|\,\|A\|/\|\underline{r}\|$ een redelijk scherpe ondergrens voor c(A) is.

Nb. Als $\underline{r} = F\tilde{\underline{x}}$ met $\|F\|/\|A\| \sim \eta$, dan is

$$\frac{\|\underline{r}\|}{\|\underline{b}\|} \leq \frac{\|F\|}{\|A\|} \cdot \frac{\|A\|\|\tilde{\underline{x}}\|}{\|\underline{b}\|}$$

en daar het kan voorkomen dat $\|\underline{b}\| \sim \|A\|\|\tilde{\underline{x}}\|$, kan $\|\underline{r}\|/\|\underline{b}\|$ van de orde η zijn terwijl toch $\|\tilde{\underline{x}} - \underline{x}\|/\|\underline{x}\|$ van de orde $c(A)\eta$ is. Men trekke dus geen optimistische conclusies uit de kleinheid van het residu!
Voorbeeld (afkomstig van Kahan):

$$A = \begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix} , \quad \underline{b} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix} .$$

Het residu bij $\tilde{\underline{x}} = (0.9911, -0.4870)^T$ is (exact) $\underline{r} = (-_{10}{}^{-8}, {}_{10}{}^{-8})$, de exacte oplossing is echter $\underline{x} = (2,-2)^T$! De matrix A moet dus extreem slecht geconditioneerd zijn, want $\underline{x} - \tilde{\underline{x}}$ is oplossing bij rechterlid $\underline{r}$, dus moet $\|A^{-1}\| \geq \|\underline{x} - \tilde{\underline{x}}\|/\|\underline{r}\| = 1.5130_{10}8$. Inderdaad blijkt dat (exact) $\det(A) = {}_{10}{}^{-8}$ en dus (regel van Cramer)

$$A^{-1} = {}_{10}{}^{8} \begin{pmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{pmatrix} ,$$

waaruit volgt $\|A^{-1}\| = 1.5130_{10}8$.

5.2.6. Lineaire stelsels met speciale matrices ([2], 5.4)

Hoewel de hierboven besproken methoden van Gauss en Crout altijd werken, is er een aantal typen matrices met een speciale structuur, waarvoor dank zij deze structuur de eliminatie op een voordeliger manier kan worden uitgevoerd.

5.2.6.1. Positief definiete matrices

Een matrix A heet symmetrisch als $A^T = A$ (dus $A_{ij} = A_{ji}$) en positief definiet als hij symmetrisch is en bovendien voor iedere $\underline{x} \neq \underline{0}$

$$\underline{x}^T A \underline{x} > 0 \ . \tag{1}$$

Dergelijke matrices komen in veel toepassingen voor, $\underline{x}^T A \underline{x}$ is dan meestal de toename van de potentiële energie van een systeem bij een afwijking $\underline{x}$ van een stabiele evenwichtsstand.
We merken nu de volgende zaken op:

a) Zij A symmetrisch. Stel dat A een LU-decompositie (zonder verwisselen) heeft (d.w.z. alle pivots zijn $\neq 0$). Dan is

$$LU = A = A^T = U^T L^T$$

en dus*)

$$UL^{-T} = L^{-1} U^T \ . \tag{2}$$

Nu is de inverse van een linksondermatrix L een linksondermatrix (ga na, merk op dat als $\underline{e}_k$ de k-de eenheidsvector is, de eerste $k-1$ componenten van de oplossing $\underline{x}$ van $L\underline{x} = \underline{e}_k$ nul zijn) en analoog voor rechtsbovenmatrices. Dus is in (2) het linkerlid rechtsboven en het rechterlid linksonder, beide zijn dus gelijk aan een diagonaalmatrix D. Derhalve $U = DL^T$ en

$$A = LDL^T \ . \tag{3}$$

b) Zij nu A positief definiet. Stel dat A een LU decompositie en dus een decompositie van de vorm (3) heeft. Dan zijn alle diagonaalelementen D_{kk} positief. Dit volgt uit (1) door te nemen $\underline{x} = L^{-T}\underline{e}_k$, waarvoor $\underline{x}^T A \underline{x} = \underline{e}_k^T D \underline{e}_k = D_{kk}$.

---

*) Daar $(L^T)^{-1} = (L^{-1})^T$ (ga na) schrijven we hiervoor L^{-T}.

Er is in dit geval dus ook een diagonaalmatrix $\tilde{D}$ met positieve diagonaalelementen zodat $D = \tilde{D}^2$ en als we stellen $\tilde{L} = L\tilde{D}$ dan is

$$A = \tilde{L}\tilde{L}^T \quad , \tag{4}$$

met $\tilde{L}$ linksonder (en $\tilde{L}_{ii} > 0$, maar in het algemeen niet $\tilde{L}_{ii} = 1$).

c) We bewijzen nu door volledige inductie dat elke positief definiete A een LU-decompositie heeft.

Het geval n = 1 is natuurlijk triviaal. Stel n > 1. Dat in dit geval de eerste slag van de Gauss-eliminatie lukt, volgt uit (1): neem daarin $\underline{x} = \underline{e}_1$, dan zien we $0 < \underline{e}_1^T A \underline{e}_1 = A_{11}$. We bewijzen nu dat de matrix $A_1^{(1)}$, die na de eerste slag van de eliminatie over blijft, weer positief definiet is. Zij

$$A = \left(\begin{array}{c|c} \alpha_1 & \underline{a}_1^T \\ \hline \underline{a}_1 & A_1 \end{array}\right)$$

(met $\alpha_1 = A_{11}$ een getal, $\underline{a}_1$ een (n-1)-vector, A_1 een (n-1) × (n-1)-matrix) dan ontstaat na één slag vegen de matrix

$$A^{(1)} = \left(\begin{array}{c|c} \alpha_1 & \underline{a}_1^T \\ \hline 0 & A_1^{(1)} \end{array}\right)$$

met

$$A_1^{(1)} = A_1 - \frac{1}{\alpha_1}\underline{a}_1\underline{a}_1^T$$

(waarin $\frac{1}{\alpha_1}\underline{a}_1\underline{a}_1^T$ de (n-1) × (n-1)-matrix met elementen $A_{i1}A_{1j}/A_{11}$, $i,j \geq 2$ is). $A_1^{(1)}$ is kennelijk symmetrisch. Om te bewijzen dat hij ook positief definiet is, nemen we in (1) $\underline{x} = \left(\begin{array}{c} \xi_1 \\ \hline \underline{x}_1 \end{array}\right)$. Dan is (ga na)

$$\underline{x}^T A \underline{x} = \xi_1^2 \alpha_1 + 2\xi_1 \underline{x}_1^T \underline{a}_1 + \underline{x}_1^T A_1 \underline{x}_1 = \alpha_1 \left(\xi_1 + \frac{\underline{x}_1^T \underline{a}_1}{\alpha_1}\right)^2 + \underline{x}_1^T A_1^{(1)} \underline{x}_1 \ .$$

Hieruit volgt (waarom?) dat $\underline{x}_1^T A_1^{(1)} \underline{x}_1 > 0$ voor $\underline{x}_1 \neq 0$, dus $A_1^{(1)}$ is positief definiet.

Uit de inductieveronderstelling volgt nu dat $A_1^{(1)}$ een LU-decompositie heeft: $A_1^{(1)} = L_1 U_1$. Derhalve is (ga na) de LU-decompositie van A

$$A = \left(\begin{array}{c|c} 1 & 0 \\ \hline \alpha_1^{-1}\underline{a}_1 & L_1 \end{array}\right) \left(\begin{array}{c|c} \alpha_1 & \underline{a}_1^T \\ \hline 0 & U_1 \end{array}\right) \ .$$

d) We schrijven de -nu voor elke positief definiete matrix geldende- splitsing maar weer als

$$A = LL^T \tag{5}$$

en vragen of we L ook rechtstreeks kunnen berekenen. Uit (5) volgt voor $i \geq k$

$$A_{ik} = \sum_{\ell=1}^{k} L_{i\ell}L_{k\ell} = \sum_{\ell=1}^{k-1} L_{i\ell}L_{k\ell} + L_{ik}L_{kk}$$

en dus

$$L_{kk} = (A_{kk} - \sum_{\ell=1}^{k-1} L_{k\ell}^2)^{\frac{1}{2}}$$

$$L_{ik} = (A_{ik} - \sum_{\ell=1}^{k-1} L_{i\ell}L_{k\ell})/L_{kk}, \quad i > k .$$

Hiermee kunnen we L dus kolomsgewijs bepalen. Dit is de zg. algoritme van Cholesky. Het aantal bewerkingen is van de orde van $\frac{1}{6}n^3$ vermenigvuldigingen en delingen en n worteltrekkingen, dus (dank zijn de symmetrie) ca. de helft van wat bij Gauss of Crout nodig is. En hoewel niet noodzakelijk $|L_{ik}|/L_{kk} \leq 1$ is, is de numerieke stabiliteit voortreffelijk. Voor de verkregen L geldt namelijk, analoog aan 5.2.5.1,

$$LL^T = A + E$$

met

$$E_{ij} = \sum_{\ell=1}^{\min(i,j)} L_{i\ell}L_{j\ell}\epsilon_{i\ell j}, \quad |\epsilon_{i\ell j}| \leq (\ell+1)\eta .$$

Hieruit volgt

$$|E_{ij}| \leq (n+1)\eta \sum_{\ell=1}^{\min(i,j)} |L_{i\ell}||L_{j\ell}| .$$

Daar voor de exacte waarden geldt

$$\sum_{\ell=1}^{i} L_{i\ell}^2 = A_{ii}$$

en dus, met de ongelijkheid van Cauchy,

$$(\sum_{\ell=1}^{\min(i,j)} |L_{i\ell}||L_{j\ell}|)^2 \leq \sum_{\ell=1}^{\min(i,j)} L_{i\ell}^2 \cdot \sum_{\ell=1}^{\min(i,j)} L_{j\ell}^2 \leq A_{ii}A_{jj} ,$$

geldt, afgezien van hogere orde termen,

$$|E_{ij}| \leq (n+1)\eta \sqrt{A_{ii}A_{jj}} \leq (n+1)\eta \|A\|_\infty, \quad \|E\|_\infty \leq n(n+1)\eta\|A\|_\infty .$$

5.2.6.2. Symmetrische matrices

Als A symmetrisch maar niet positief definiet is dan geldt niet altijd dat er een links ondermatrix L en een diagonaalmatrix D is zo dat $A = LDL^T$ (neem bv. het geval dat $A_{11} = 0$). En als het wel kan dan kunnen elementen van D willekeurig groot worden, hetgeen numerieke instabiliteit betekent. Voorbeeld:

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} 1 & \\ b/a & 1 \end{pmatrix} \begin{pmatrix} a & \\ & c-b^2/a \end{pmatrix} \begin{pmatrix} 1 & b/a \\ & 1 \end{pmatrix} ;$$

als $b^2 >> |ac|$ dan gaan bij de aftrekking $c - (b^2/a)$ de achterste cijfers van c verloren. Er bestaan echter wel numeriek stabiele algoritmen om in ca $\frac{1}{6} n^3$ bewerkingen A te splitsen als

$$A = BTB^T ,$$

waarin B een matrix is waarvoor de stelsels $B\underline{c} = \underline{b}$ en $B^T\underline{x} = \underline{d}$ eenvoudig oplosbaar zijn (B is bv. op rijverwisselingen na een linksondermatrix) en de matrix T een symmetrische tridiagonale matrix is (d.w.z. $T_{ij} = 0$ voor $|i - j| > 1$). Daar het numeriek stabiel oplossen van een stelsel $T\underline{d} = \underline{c}$ ca. 7n bewerkingen kost -zie 5.2.6.4- hebben we met behoud van de numerieke stabiliteit eenzelfde besparing verkregen als bij Cholesky. Nb. Natuurlijk hoeft men van een symmetrische matrix slechts het linksonderdeel (+ diagonaal) op te slaan; dat kan in een een-dimensionaal array A[1 : n × (n+1) ÷ 2] door het element A_{ij} op de plaats A[i × (i-1) ÷ 2 + j] te schrijven.

5.2.6.3. Bandmatrices

A heet bandmatrix met bandbreedte d als *)

$$A_{ij} = 0 \quad \text{voor } |i - j| > d .$$

Op de i-de rij zijn dus hoogstens de elementen A_{ij} met

$$i - d \leq j \leq i + d$$

ongelijk nul.
Iets algemener heet A een bandmatrix met linkerbandbreedte d_ℓ, resp. rechterbandbreedte d_r, als

$$A_{ij} = 0 \text{ voor } i - j > d_\ell \text{ resp. } A_{ij} = 0 \text{ voor } j - i > d_r .$$

*) Anderen noemen in dit geval de bandbreedte 2d + 1.

Nu geldt: als A linkerbandbreedte d_ℓ en A' linkerbandbreedte d'_ℓ heeft dan heeft AA' linkerbandbreedte $d_\ell + d'_\ell$. Immers, in

$$(AA')_{ij} = \sum_{\ell=1}^{n} A_{i\ell}A'_{\ell j}$$

kan een term $A_{i\ell}A'_{\ell j}$ alleen ongelijk nul zijn als zowel $i - \ell \le d_\ell$ als $\ell - j \le d'_\ell$ en hiervoor is nodig $i - j \le d_\ell + d'_\ell$. Voor $i - j > d_\ell + d'_\ell$ zijn dus alle termen rechts gelijk aan nul.
Een direct gevolg van deze stelling is: Als A linkerbandbreedte d_ℓ en rechterbandbreedte d_r heeft <u>en</u> A heeft een LU-decompositie (zonder verwisselen), dan heeft L linkerbandbreedte d_ℓ. Dit volgt uit $L = AU^{-1}$ en het feit dat U^{-1} rechtsboven is en dus linkerbandbreedte 0 heeft. Analoog: U heeft rechterbandbreedte d_r. Dit resultaat betekent dat we bij Gauss of Crout veel minder elementen van L en U uit hoeven te rekenen en dat de sommen $\Sigma L_{i\ell}U_{\ell j}$ ook vrij kort zijn. Het gevolg is dat de eliminatie nu ca. $nd_\ell(d_r + 1)$ bewerkingen kost. Als we terwille van de numerieke stabiliteit met rijverwisselingen werken dan hoeven we bij het zoeken naar een geschikte pivot slechts $d_\ell + 1$ elementen te bekijken en daaruit blijkt na enig denken dat de bandbreedte van L ook nu d_ℓ is en dat die van U hoogstens $d_r + d_\ell$ wordt. Gebruikmaking van een algoritme die rekening houdt met de bandstructuur is dus voordelig zodra d_r en d_ℓ kleiner dan ca. $\frac{1}{3}$ n zijn.

<u>Opmerkingen</u>

1) Behandeling van een rechterlid kost, als bij de eliminatie niet verwisseld wordt, ca. $n(d_\ell + d_r + 1)$ bewerkingen en als wel verwisseld wordt ca. $n(2d_\ell + d_r + 1)$ bewerkingen.
2) Men kan de niet triviale elementen van een bandmatrix opslaan in een array $A[1 : n, -d_\ell : d_r]$: het element A_{ij} staat dan in $A[i, j - i]$. Voor eliminatie met verwisselingen declareert men liefst een array $A[1 : n, -d_\ell : d_\ell + d_r]$. De niet-triviale elementen van L en U kunnen dan op de plaatsen van de corresponderende A_{ij} komen.

2.6.4. Tridiagonaalmatrices

Onder de bandmatrices nemen de tridiagonaalmatrices (bandmatrices met $d_\ell = d_r = 1$) een speciale plaats in omdat ze veel voorkomen en omdat de algoritmen bijzonder eenvoudig worden. Meestal noteert men een stelsel met een tridiagonale matrix als $\underline{Ax} = \underline{d}$ met

$$A = \begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & c_{n-1} & \\ & & b_n & a_n & \end{pmatrix} .$$

en bergt men de elementen a_i, b_i en c_i op in afzonderlijke een-dimensionale arrays. Eliminatie zonder verwisseling is bijzonder eenvoudig, zie § 3.5.2 voor een programma. We beschrijven nu eliminatie met verwisselen. Neem aan dat we na de k-1 ste slag als k-de vergelijking (dus als bovenste vergelijking van het stelsel met matrix $A^{(k-1)}$) hebben gekregen

$$p_k x_k + q_k x_{k+1} = r_k \qquad (1)$$

en dat de vergelijkingen k + 1 t/m n onveranderd zijn (deze bewering geldt als k = 1 met $p_1 = c_1$, $q_1 = c_1$, $r_1 = d_1$). Dan moet in de k-de slag x_k geëlimineerd worden met behulp van (1) of van de nog ongewijzigde k+1-ste vergelijking

$$b_{k+1} x_k + a_{k+1} x_{k+1} + c_{k+1} x_{k+2} = d_{k+1} \qquad (2)$$

(de vergelijkingen k+ 2 t/m n bevatten x_k niet!). Als $|p_k| \geq |b_{k+1}|$ dan wordt (1) de k-de vergelijking van het driehoeksstelsel en krijgen we met

$$\ell_{k+1} := b_{k+1} / p_k ; \; p_{k+1} := a_{k+1} - \ell_{k+1} \times q_k ,$$

$$q_{k+1} := c_{k+1}, \; r_{k+1} := d_{k+1} - \ell_{k+1} \times r_k$$

als nieuwe k + 1-ste vergelijking

$$p_{k+1} x_{k+1} + q_{k+1} x_{k+2} = r_{k+1} . \qquad (3)$$

Als $|b_{k+1}| > |p_k|$ dan wordt (2) de k-de vergelijking van het driehoeksstelsel en de nieuwe k + 1-ste vergelijking heeft de vorm (3) met

$$\ell_{k+1} = p_k/b_{k+1};\quad p_{k+1} := q_k - \ell_{k+1} \times a_{k+1};$$

$$q_{k+1} := -\ell_{k+1} \times c_{k+1};\quad r_{k+1} := r_k - \ell_{k+1} \times d_{k+1}\ .$$

Hiermee zijn we dus weer in de uitgangssituatie terug. Tevens zien we

- het rechterdriehoeksstelsel heeft hoogstens rechter bandbreedte 2.
- de eliminatie kost hoogstens ca. 3n, bewerking van een rechterlid hoogstens ca. 4n vermenigvuldigingen en delingen.

.2.6.5. IJl-bezette matrices (sparse matrices)

In de laatste jaren is er toenemende aandacht voor zeer grote matrices (n ~ 1000 à 10000) met slechts een kleine aantal niet-nullen per rij (bv. enkele tientallen), min of meer onregelmatig verdeeld over de matrix. In diverse toepassingen komen deze voor, bv. in netwerkproblemen en bij het oplossen van partiële differentiaalvergelijkingen met zg. elementenmethoden. Een eerste probleem is hoe men de niet-triviale elementen van zo'n matrix handzaam opslaat. De volgende methode is mogelijk. Zij N het aantal niet-nullen onder de matrixelementen. Declareer dan een real array A[1:N], een integer array index[1:N] en een integer array rowend[0:n]. Berg de niet-triviale elementen van A in "lexicografische" volgorde (d.w.z. rijsgewijs) op in A, noteer de kolomindices van deze elementen op de corresponderende plaatsen in index en noteer in rowend[i] het plaatsnummer van het laatste niet-triviale element van de i-de rij. Om te zien hoe men met een aldus opgeslagen matrix A werkt beschouwen we de matrix maal vector operatie $\underline{y} = A\underline{x}$. Hiervoor geldt voor i = 1,...,n (ga na)

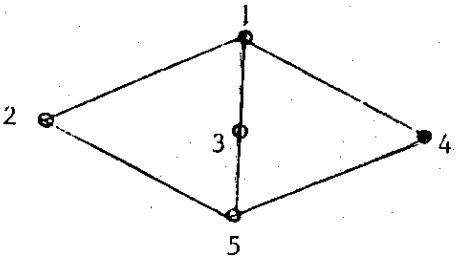
$$y_i = \sum_{\ell=\text{rowend}[i-1]+1}^{\text{rowend}[i]} A[\ell] \times x[\text{index}[\ell]]\ .$$

Voor het uitvoeren van ingewikkelder berekeningen is dit opbergschema wat minder geschikt. Een schema waarbij men aan de matrix gemakkelijk nieuwe niet-nul elementen kan toevoegen is het volgende. Declareer een real array A[1:N], integer arrays index, next[1:N] en rowstart, rownumber[1:n] met N voldoende groot (in een machine met "virtual storage" geeft het weinig of we N te groot nemen). Zet in A de matrixelementen in willekeurige volgorde en in de corresponderende plaatsen van index en next resp. de kolomindex van het element en de plaats in A waar het eerstvolgende (in lexicografische zin) niet-triviale element van de matrix staat; rowstart[i] bevat de plaats

in A waar het eerste niet-triviale element uit de i-de rij staat en row-number[i] is het aantal niet-triviale elementen in de i-de rij. Ga na hoe men bij dit schema gemakkelijk nieuwe elementen aan de matrix kan toevoegen.

Gaan we in een ijl-bezette matrix elimineren dan creëren we als regel een aantal nieuwe niet-triviale elementen (de matrix "loopt vol"). Als we in de eerste slag $A_{k\ell}$ als pivot kiezen (d.w.z. x_ℓ elimineren met behulp van de k-de vergelijking) dan geldt: voor alle i en j met $A_{i\ell} \neq 0$ en $A_{kj} \neq 0$ is $A_{ij}^{(1)} \neq A_{ij}$ en dus als regel $\neq 0$ (dat zo'n $A_{ij}^{(1)}$ ook weleens 0 wordt is "toevallig geluk" en wordt meestal genegeerd). Was $A_{ij} = 0$ dan ontstaat dus een nieuw niet-triviaal element dat moet worden opgeborgen en een rol speelt bij de volgende eliminatieslagen. Het is dus van belang de pivotkeuze zo te doen dat deze "fill-in" geminimaliseerd wordt (onder behoud van een zekere mate van numerieke stabiliteit). Hiervoor zijn vele strategieën ontwikkeld. Vaak gebruikt men daarbij graphen als hulpmiddel. Bijvoorbeeld kan men aan een symmetrische matrix A toevoegen een graph met n knooppunten die corresponderen met de diagonaalelementen van A en een tak tussen knooppunten i en j dan en slechts dan als $A_{ij} \neq 0$. Elimineert men nu met A_{kk} als pivot dan ontstaat een nieuwe tak tussen (i,j) dan en slechts dan als de punten i en j niet rechtstreeks maar wel via punt k met elkaar verbonden waren (namelijk als $A_{ij} = 0$ en $A_{ik}A_{kj} \neq 0$).
Voorbeeld:

$$A = \begin{pmatrix} x & x & x & x & 0 \\ x & x & 0 & 0 & x \\ x & 0 & x & 0 & x \\ x & 0 & 0 & x & x \\ 0 & x & x & x & x \end{pmatrix}$$



Begint men de eliminatie met A_{11} als pivot dan ontstaan nieuwe takken (2,3), (2,4) en (3,4), zodat $A^{(1)}$ helemaal vol is. Begint men met A_{22} als eerste pivot dan ontstaat een nieuwe tak (1,5) en als men daarna A_{33} en A_{44} als pivot kiest dan komen er geen nieuwe takken meer bij.

Een minder zuinig, maar voor Cholesky-decompositie erg handig opbergschema voor positief definiete matrices is het zg. profiel-opbergschema.
Hierbij gaan we uit van de opmerking dat bij de Cholesky-algorithme (zie 5.2.6.1) geldt (ga na):

als $A_{ij} = 0$ voor $1 \le j < i-p_i$, dan is ook $L_{ij} = 0$ voor $1 \le j < i-p_i$.

Dat wil zeggen, als in de i-de rij $A_{i,i-p_i}$ het meest linkse niet-triviale element is, dan kunnen bij de eliminatie hoogstens de plaatsen $A_{i,i-p_i+1}$ t/m $A_{i,i-1}$ "vollopen". Reserveren we dus voor het strict linksonder deel van de i-de rij p_i plaatsen in een een-dimensionaal array (waar aanvankelijk de elementen $A_{i,i-p_i}$ t/m $A_{i,i-1}$ opgeborgen worden, ook degene die nul zijn) en daarnaast een array met lengte n voor de diagonaal, dan hoeven we tijdens de eliminatie geen enkele nieuwe plaats meer te creëren. Men noemt de rij getallen $p_1,\ldots,p_n$ het profiel van de matrix. Men kan onder behoud van de positief definietheid het profiel van een matrix veranderen door de rijen en de kolommen op dezelfde wijze te permuteren, of anders: door de vergelijkingen en de bijbehorende (!) onbekenden anders te nummeren. Het is dus voordelig, vergelijkingen en onbekenden zo te nummeren, dat $\sum p_i$ min of meer minimaal is. Hiervoor bestaan algorithmen, o.a. het algorithme van Cuthill en McKee.

5.3. Iteratieve methoden ([2], 5.7)

5.3.1. De methoden van Jacobi en van Gauss-Seidel

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

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

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

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

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

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.

Om de convergentie van deze processen te onderzoeken schrijven we ze eerst in vectornotatie.
Zij

$$A = D - B = D - L - U \ ,$$

waarin

D = het diagonale deel van A,

-L = het strict linksonder deel van A,

-U = het strict rechtsboven deel van A,

-B = L + U = het buitendiagonale deel van A.

Dan kunnen we voor Jacobi schrijven

$$D\underline{x}^{(k+1)} = \underline{b} + B\underline{x}^{(k)} \ ,$$

en voor Gauss-Seidel

$$(D - L)\underline{x}^{(k+1)} = \underline{b} + U\underline{x}^{(k)} \ .$$

Als nu $\underline{x}$ de oplossing van het stelsel $A\underline{x} = \underline{b}$ is en

$$\underline{y}^{(k)} := \underline{x}^{(k)} - \underline{x}$$

de fout in de k-de benadering is, dan geldt (ga na) bij Jacobi, resp. Gauss-Seidel

$$\underline{y}^{(k+1)} = D^{-1}B\underline{y}^{(k)} \ ,$$

$$\underline{y}^{(k+1)} = (D - L)^{-1}U\underline{y}^{(k)} \ .$$

Convergentie betekent $\lim_{k\to\infty} \underline{y}^{(k)} = \underline{0}$ (voor iedere $\underline{y}^{(0)}$). Een voldoende voorwaarde hiervoor is kennelijk

$$\|D^{-1}B\| \leq \lambda < 1, \text{ resp. } \|(D - L)^{-1}U\| \leq \lambda < 1 \qquad (1)$$

want dan is $\|\underline{y}^{(k)}\| \leq \lambda^k \|\underline{y}^{(0)}\|$.
Men kan bewijzen dat een nodige en voldoende voorwaarde voor convergentie is

$$\rho(D^{-1}B) < 1, \text{ resp } \rho((D-L)^{-1}U) < 1 ,$$

waarin met $\rho(C)$ de in absolute waarde grootste eigenwaarde van C bedoeld wordt.

Opmerking

Beide processen zijn van de vorm $\underline{x}^{(k+1)} = \underline{f}(\underline{x}^{(k)})$, zoals behandeld in 1.4.2. Toepassing van de daar genoemde globale convergentiestelling met $L = \lambda$, $\underline{a} = \underline{0}$ en een geschikte R levert direct dat de voorwaarden (1) voldoende zijn voor convergentie (ga na!).

Stelling

Als de matrix A zg. diagonaaldominant is, dat wil zeggen:

$$\sum_{j\neq i} |A_{ij}| < |A_{ii}|, \quad i = 1,\ldots,n, \tag{2}$$

dan is A niet singulier en de processen van Jacobi en Gauss-Seidel convergeren.

Bewijs

Uit (2) volgt dat alle $A_{ii} \neq 0$, dus D is regulier en

$$\|D^{-1}B\|_\infty = \max_i \left(\frac{1}{|A_{ii}|} \sum_{j\neq i} |A_{ij}|\right) = \lambda < 1. \tag{3}$$

Hieruit volgt met een resultaat uit 5.2.5.2 dat $D^{-1}A = I - D^{-1}B$ regulier is, dus A ook. En tevens volgt uit (3) dat Jacobi convergent is.
De convergentie van Gauss-Seidel bewijzen we door aan te tonen dat uit (3) volgt dat

$$\|(D - L)^{-1}U\|_\infty \leq \lambda. \tag{4}$$

Zij voor gegeven $\underline{y} \neq 0$, $\underline{z} = (D-L)^{-1}U\underline{y}$. Dan is

$$\underline{z} = D^{-1}(L\underline{z} + U\underline{y})$$

en dus voor $i = 1,\ldots,n$

$$|z_i| \leq |D_{ii}|^{-1}\left(\sum_{j<i} |L_{ij}| + \sum_{j>i} |U_{ij}|\right) \cdot$$

$$\cdot \max(|z_1|,\ldots,|z_{i-1}||y_{i+1}|,\ldots,|y_n|),$$

waaruit volgt

$$\|\underline{z}\|_\infty \le \|D^{-1}(L+U)\|_\infty \max(\|\underline{z}\|_\infty, \|\underline{y}\|_\infty) = \lambda \max(\|\underline{z}\|, \|\underline{y}\|) \ .$$

De veronderstelling $\max(\|\underline{z}\|, \|\underline{y}\|) = \|\underline{z}\|$ leidt tot $0 < \|\underline{z}\| \le \lambda \|\underline{z}\|$, hetgeen onmogelijk is daar $\lambda < 1$. Derhalve geldt voor iedere $\underline{y}$ dat $\|\underline{z}\| \le \lambda \|\underline{y}\|$ en dat impliceert (4). □

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

a) λ 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 differentiatievergelijkingen) treden zeer grote stelsels op ($n \sim 1000$ 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.

5.3.2. Systematische overrelaxatie (S.O.R.)

Schrijf het Gauss-Seidel proces als

$$D(\underline{x}^{(k+1)} - \underline{x}^{(k)}) = \underline{b} - D\underline{x}^{(k)} + L\underline{x}^{(k+1)} + U\underline{x}^{(k)}$$

en merk op dat de i-de component van $\underline{x}^{(k+1)}$ verkregen wordt door de i-de component van het residu $\underline{b} - A\underline{x}$ te berekenen met de nieuwe waarden $x_1^{(k+1)}$ t/m $x_{i-1}^{(k+1)}$ en de oude waarden $x_i^{(k)}$ t/m $x_n^{(k)}$. De nieuwe waarde $x_i^{(k+1)}$ is dan zo dat als in de berekening van het (i-de) residu $x_i^{(k+1)}$ in plaats van $x_i^{(k)}$ gebruikt wordt, dit residu nul wordt (ga na). Men zei (bij bepaalde mechanica-toepassingen) dat de i-de vergelijking gerelaxeerd (ontspannen) wordt. En men constateerde dat de convergentie versneld kon worden door enigszins te overrelaxeren, d.w.z. door $\underline{x}^{(k+1)}$ te bepalen uit

$$D(\underline{x}^{(k+1)} - \underline{x}^{(k)}) = \omega(\underline{b} - D\underline{x}^{(k)} + L\underline{x}^{(k+1)} + U\underline{x}^{(k)})$$

met geschikte $\omega > 1$. Dat is het zogenaamde S.O.R. (systematische overrelaxatie) proces. Voor matrices A van een bepaalde structuur (die o.a. voorkomt bij discretisatie van partiële differentiaalvergelijkingen, zie 4.3) kan men theoretisch aangeven wat de beste waarde van ω en de te behalen convergentiesnelheid is. Ook hier gaat het om de grootste eigenwaarde van de matrix waarmee de fout $\underline{y}^{(k+1)}$ verkregen wordt uit $\underline{y}^{(k)}$:

$$\underline{y}^{(k+1)} = (D - \omega L)^{-1}((1 - \omega)D + \omega U)\underline{y}^{(k)} \quad .$$

We gaan daar hier niet verder op in, zie echter 4.3 voor een bijzonder geval.

6. Kleinste kwadraten aanpassing ([2], 5.7)

Zij A een m × n matrix met m > n, $\underline{b}$ een vector met m componenten. Dan heeft het stelsel

$$A\underline{x} = \underline{b} \tag{1}$$

als regel geen oplossing $\underline{x}$ (een vector met n componenten) omdat er meer vergelijkingen dan onbekenden zijn. Wel kunnen we vragen: is er een $\hat{\underline{x}}$ zo dat voor iedere $\underline{x}$ geldt

$$\|A\underline{x} - \underline{b}\| \geq \|A\hat{\underline{x}} - \underline{b}\|.$$

Men noemt $A\hat{\underline{x}}$ dan een beste approximatie voor $\underline{b}$. Kiezen we als norm de 2-norm dan spreken we van kleinste kwadraten-approximatie of -aanpassing omdat we dan minimaliseren de kwadratensom

$$\varphi(\underline{x}) = \|\underline{b} - A\underline{x}\|_2^2 = \sum_{i=1}^{m} (b_i - \sum_{j=1}^{n} A_{ij}x_j)^2. \tag{2}$$

(Men noemt $\hat{\underline{x}}$ in dit geval ook wel de kleinste kwadraten oplossing van (1) maar dat is misleidend, daar $\hat{\underline{x}}$ als regel geen oplossing van (1) is.)

Over de betekenis van kleinste kwadraten aanpassing voor het aanpassen van waarnemingen aan fysische en andere modellen spreken we in hoofdstuk 8. Nb. In dit hoofdstuk wordt uitsluitend de 2-norm gebruikt.

6.1. De normaalvergelijkingen

Een nodige voorwaarde voor het minimaal zijn van de functie $\varphi(\underline{x})$ uit (2) in $\underline{x} = \hat{\underline{x}}$ is dat daar de gradiënt van $\varphi(\underline{x})$ nul is. Dit leidt tot de vergelijkingen

$$\frac{\partial \varphi}{\partial x_k} = 2 \sum_{i=1}^{m} A_{ik}(b_i - \sum_{j=1}^{n} A_{ij}x_j) = 0 \ , \ k = 1,\ldots,n \ ,$$

of wel (ga na)

$$A^T(\underline{b} - A\underline{x}) = \underline{0}. \tag{3}$$

Dit zijn de zogenaamde normaalvergelijkingen van Gauss.
Hoewel zij niet het meest geschikt zijn voor de numerieke oplossing van het kleinste kwadratenprobleem, bespreken we een aantal theoretische conclusies.

1. Als de n kolommen van A onafhankelijk zijn dan is de matrix A^TA positief definiet. Want hij is duidelijk symmetrisch en $\underline{x}^TA^TA\underline{x} = 0$ impliceert $A\underline{x} = \underline{0}$, dus $\underline{x} = \underline{0}$. Het stelsel (3) heeft in dat geval dus voor iedere $\underline{b}$ een oplossing namelijk

$$\hat{\underline{x}} = (A^TA)^{-1}A^T\underline{b}.$$

Uit

$$\underline{b} - A\underline{x} = A(\hat{\underline{x}} - \underline{x}) + (\underline{b} - A\hat{\underline{x}})$$

en $A^T(\underline{b} - A\hat{\underline{x}}) = \underline{0}$ volgt met Pythagoras dat

$$\|\underline{b} - A\underline{x}\|^2 = \|A(\hat{\underline{x}} - \underline{x})\|^2 + \|\underline{b} - A\hat{\underline{x}}\|^2$$

en daaruit volgt dat $\|\underline{b} - A\underline{x}\|$ minimaal is dan en slechts dan als $A(\hat{\underline{x}} - \underline{x}) = \underline{0}$, dus als $\underline{x} = \hat{\underline{x}}$.

2. Als de kolommen van A niet onafhankelijk zijn dan heeft het stelsel (3) voor iedere $\underline{b}$ meerdere oplossingen, waarbij echter $A\underline{x}$ eenduidig bepaald is. We gaan daar hier niet verder op in (zie evenwel 9.1e).

Men kan de normaalvergelijkingen oplossen door $C := A^TA$, $\underline{c} := A^T\underline{b}$ te bepalen en het stelsel

$$C\underline{x} = \underline{c}$$

op te lossen met behulp van de methode van Cholesky. De invloed van afrondfouten hierbij wordt bepaald door het conditiegetal van C. We zullen zien dat dat het kwadraat is van het conditiegetal van A, als we dat verstandig definiëren voor een niet-verkante matrix.
Zij namelijk (voor het geval dat de kolommen van A onafhankelijk zijn)

$$c(A) := \max_{\underline{x}\neq\underline{0}} \frac{\|A\underline{x}\|}{\|\underline{x}\|} \Big/ \min_{\underline{x}\neq\underline{0}} \frac{\|A\underline{x}\|}{\|\underline{x}\|} \quad .$$

Daar voor vierkante matrices geldt: als de kolommen van A onafhankelijk zijn dan is A regulier en dan is

$$\|A^{-1}\| = \max_{\underline{y}\neq\underline{0}} \frac{\|A^{-1}\underline{y}\|}{\|\underline{y}\|} = \max_{\underline{x}\neq\underline{0}} \frac{\|\underline{x}\|}{\|A\underline{x}\|} = \left(\min_{\underline{x}\neq\underline{0}} \frac{\|A\underline{x}\|}{\|\underline{x}\|}\right)^{-1} ,$$

is deze definitie een generalisatie van de oude definitie. Gebruiken we de 2-norm dan geldt, als $C = A^TA$,

$$c^2(A) = \max \frac{\underline{x}^T C \underline{x}}{\underline{x}^T \underline{x}} \Big/ \min \frac{\underline{x}^T C \underline{x}}{\underline{x}^T \underline{x}}$$

$$= \text{grootste eigenwaarde van C/kleinste eigenwaarde van C}$$

$$= \|C\| \|C^{-1}\| = c(C) .$$

Dit is de reden dat bij het opstellen en oplossen van de normaalvergelijkingen meer nauwkeurigheidsverlies optreedt dan bij andere methoden om (2) te minimaliseren.

6.2. Orthogonale transformatie van het kleinste kwadraten probleem. ([2], 5.7.2)

Zij Q een zg. orthogonale m×m matrix, d.w.z.

$$Q^T Q = I .$$

Dan is voor iedere $\underline{y}$

$$\|Q\underline{y}\|_2^2 = \underline{y}^T Q^T Q \underline{y} = \underline{y}^T \underline{y} = \|\underline{y}\|^2 . \tag{1}$$

Stel nu dat we een orthogonale Q kunnen vinden zo dat

$$QA = R \tag{2}$$

met

$$R = \begin{pmatrix} R_{11} & \cdots & R_{1n} \\ & \ddots & \vdots \\ & & R_{nn} \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} R_1 \\ \hline 0 \end{pmatrix} , \tag{3}$$

waarin R_1 een n × n rechtsbovenmatrix is. Als de kolommen van A onafhankelijk zijn dan is R_1 regulier (dus alle $R_{ii} \neq 0$). We kunnen het kleinste kwadraten probleem dan als volgt oplossen. Zij

$$Q\underline{b} = \underline{c} = \begin{pmatrix} \underline{c}_1 \\ \hline \underline{c}_2 \end{pmatrix} \tag{4}$$

waarbij $\underline{c}_1$ bestaat uit de eerste n componenten van $\underline{c}$.
Met (1) volgt dan uit (2), (3), (4) en Pythagoras

$$\|\underline{b} - A\underline{x}\|^2 = \|Q(\underline{b} - A\underline{x})\|^2 = \|\underline{c} - R\underline{x}\|^2$$

$$= \|\underline{c}_1 - R_1\underline{x}\|^2 + \|\underline{c}_2\|^2 .$$

Het is duidelijk dat het rechterlid minimaal is dan en slechts dan als $\underline{x} = \hat{\underline{x}}$ waarin $\hat{\underline{x}}$ oplossing is van

$$R_1\hat{\underline{x}} = \underline{c}_1 \qquad (5)$$

(daar R_1 regulier is heeft dit n×n-stelsel een eenduidige oplossing). En de norm van de residu-vector

$$\underline{r} = \underline{b} - A\hat{\underline{x}} \qquad (6)$$

is dan gelijk aan $\|\underline{c}_2\|$.
Bepalend voor de nauwkeurigheid waarmee $\hat{\underline{x}}$ uit (5) opgelost kan worden is het conditiegetal van R_1. Daar voor iedere $\underline{x}$

$$\|A\underline{x}\| = \|QA\underline{x}\| = \|R\underline{x}\| = \|R_1\underline{x}\| ,$$

geldt (ga na)

$$c(R_1) = c(A) ,$$

zodat bij deze methode om het kleinste kwadraten probleem op te lossen geen kwadratering van het conditiegetal (en dus onnodig nauwkeurigheidsverlies) optreedt.
In 6.3 zullen we zien dat het inderdaad mogelijk is om een orthogonale transformatie Q te vinden waarvoor (2) geldt. Men geeft daarom tegenwoordig de voorkeur aan behandeling van het kleinste kwadraten probleem met de hier geschetste methode boven het opstellen en oplossen (met Cholesky) van de normaalvergelijkingen.

Opmerking
Er geldt $A^TA = R^TR = R_1^TR_1$ en $A^T\underline{b} = R^T\underline{c} = R_1^T\underline{c}_1$, zodat $\hat{\underline{x}} = R_1^{-1}\underline{c}_1$ inderdaad aan de normaalvergelijkingen voldoet.
Ook volgt hieruit dat de matrix $(A^TA)^{-1}$, die voor bepaalde statistische toepassingen nodig is, berekend kan worden uit R_1: $(A^TA)^{-1} = R_1^{-1}R_1^{-T}$.

6.3. De transformatie van Householder ([18], 3.4)

Zij $\underline{u}$ een gegeven m-vector met lengte 1 (d.w.z., $\|\underline{u}\|_2^2 = \underline{u}^T\underline{u} = 1$) en P de m×m-matrix

$$P = I - 2\underline{u}\underline{u}^T \qquad (1)$$

(dus $P_{ij} = \delta_{ij} - 2u_iu_j$).
Dan is P de matrix van de loodrechte spiegeling aan het vlak door de oorsprong loodrecht op $\underline{u}$, want $P\underline{u} = -\underline{u}$ en $P\underline{x} = \underline{x}$ als $\underline{x} \perp \underline{u}$.

We merken op dat P symmetrisch en ook orthogonaal is, want

$$P^TP = P^2 = I - 4\underline{u}\underline{u}^T + 4\,\underline{u}\underline{u}^T\underline{u}\underline{u}^T = I \ .$$

Daaruit (of uit de meetkundige interpretatie) volgt dat voor iedere $\underline{x}$ geldt $\|P\underline{x}\| = \|\underline{x}\|$.

Kunnen we bij gegeven vector $\underline{v}$ de vector $\underline{u}$ zo bepalen dat

$$P\underline{v} = \alpha\underline{e}_1 \ , \tag{2}$$

waarin $\underline{e}_1$ de eerste eenheidsvector is? Dan moet zeker

$$|\alpha| = \|P\underline{v}\| = \|\underline{v}\| \ . \tag{3}$$

We schrijven P **liever in de met (1) equivalente vorm**

$$P = I - \beta^{-1}\underline{w}\underline{w}^T \text{ met } \beta = \tfrac{1}{2}\underline{w}^T\underline{w} \tag{4}$$

(dit klopt met (1): neem $\underline{u} = (2\beta)^{-\frac{1}{2}}\underline{w}$).
Uit (2) en (4) volgt dan

$$\alpha\underline{e}_1 = P\underline{v} = \underline{v} - (\beta^{-1}\underline{w}^T\underline{v})\underline{w} \ .$$

en daaraan kunnen we voldoen door te nemen

$$\underline{w} = \underline{v} - \alpha\underline{e}_1 \ , \ \beta = \underline{w}^T\underline{v} \ . \tag{5}$$

Dit is een voordelige keus omdat $\underline{w}$ slechts in de eerste component verschilt van $\underline{v}$: $w_1 = v_1 - \alpha$, $w_i = v_i$ voor $i \geq 2$.
Uit (5) en (3) volgt

$$\beta = \underline{w}^T\underline{v} = \underline{v}^T\underline{v} - \alpha\underline{e}_1^T\underline{v} = \alpha^2 - \alpha v_1 = \alpha(\alpha - v_1) = -\alpha w_1 .$$

Daar dan

$$\underline{w}^T\underline{w} = \underline{v}^T\underline{v} - 2\alpha\underline{e}_1^T\underline{v} + \alpha^2 = 2\alpha^2 - 2\alpha v_1 = 2\,\beta \ ,$$

voldoen de gekozen $\underline{w}$ en β inderdaad aan de nevenconditie uit (4).
Rest nog de keuze van het teken van α. Omdat β bepaald wordt met behulp van $\alpha - v_1$ mag bij deze aftrekking geen cijferverlies optreden; dit wordt bereikt door te nemen

$$\alpha := \underline{\text{if}} \ \ v_1 > 0 \ \underline{\text{then}} \ -\|\underline{v}\| \ \underline{\text{else}} \ \|\underline{v}\| \ .$$

Het blijkt dat met deze keus de numerieke stabiliteit en ook de vrijwel-orthogonaliteit van P verzekerd is.

We hoeven P niet expliciet te berekenen en op te slaan in een m×m-array. Want om een vector $\underline{y}$ met P te vermenigvuldigen moeten we bepalen

$$\underline{z} = P\underline{y} = \underline{y} - (\beta^{-1}\underline{w}^T\underline{y})\underline{w} .$$

Het is dus voldoende om β en $\underline{w}$ te bewaren. Bovendien vergt de bepaling van $\underline{z}$ op deze wijze slechts 2m vermenigvuldigingen en één deling (tegen m^2 vermenigvuldigingen als we P als matrix hadden opgeslagen).
De transformatie P wordt Householder-transformatie genoemd.

Analoog aan (2) kunnen we van P eisen dat voor een gegeven vector

$$\underline{v} = \begin{pmatrix} \underline{v}_1 \\ \hline \underline{v}_2 \end{pmatrix}$$

waarin $\underline{v}_1$ k - 1 en $\underline{v}_2$ m - k + 1 componenten heeft, geldt

$$P\underline{v} = \begin{pmatrix} \underline{v}_1 \\ \hline \alpha\tilde{\underline{e}}_k \end{pmatrix} = \begin{pmatrix} \underline{v}_1 \\ \hline \underline{0} \end{pmatrix} + \alpha\underline{e}_k .$$

waarin $\tilde{\underline{e}}_k$ de k-de eenheidsvector is, waarvan de eerste k - 1 elementen zijn weggelaten (dus $\tilde{\underline{e}}_k$ in $\mathbb{R}^{m-k+1}$). We moeten dan nemen

$$\alpha = \pm \|\underline{v}_2\|, \quad \underline{w} = \begin{pmatrix} \underline{0} \\ \hline \underline{v}_2 - \alpha\tilde{\underline{e}}_k \end{pmatrix} , \quad \beta = -\alpha w_k = -\alpha(v_k - \alpha) .$$

Voor een willekeurige vector $\underline{y}$ geldt dan ook dat de eerste k - 1 componenten van $P\underline{y}$ dezelfde zijn als die van $\underline{y}$.

Op deze wijze kunnen we nu bij een gegeven m×n-matrix A achtereenvolgens Householdertransformaties $P_1,\ldots,P_n$ vinden zo dat

$$P_n \ldots P_1 A = R \tag{6}$$

waarbij R de vorm (3) uit 6.2 heeft.
Zij nl. $\underline{a}_k$ de k-de kolom van A. Neem P_1 zo dat $P_1\underline{a}_1 = R_{11}\underline{e}_1$. Daarna P_2 zo dat

$$P_2(P_1\underline{a}_2) = \begin{pmatrix} R_{12} \\ \hline R_{22}\tilde{\underline{e}}_2 \end{pmatrix}$$

en zo voort. In het algemeen nemen we P_k zo dat

$$P_k(P_{k-1} \ldots P_1\underline{a}_k) = \begin{pmatrix} R_{1k} \\ \ldots \\ R_{k-1,k} \\ \hline R_{kk}\tilde{\underline{e}}_k \end{pmatrix}$$

Daar voor iedere $\underline{y}$ de eerste k - 1 componenten van $P_k\underline{y}$ dezelfde zijn als die van $\underline{y}$ volgt nu direct dat het rechterlid van (6) de gewenste vorm heeft.

We kunnen de gegevens betreffende $P_1,\ldots,P_n$ en R grotendeels opslaan in het array waar A stond: boven de diagonaal de boven-diagonaalelementen van R, op en onder de diagonaal de niet-triviale elementen van $\underline{w}_1,\ldots,\underline{w}_n$. De diagonaal-elementen van R bergen we op in een apart een-dimensionaal array D; de getallen $\beta_1,\ldots,\beta_n$ zijn dan ook bekend: $\beta_k = -D_k \times A_{kk}$.

Ga nu na dat het volgende stukje programma eerst bij een gegeven A de transformaties $P_1,\ldots,P_n$ en de matrix R bepaalt en daarna bij gegeven $\underline{b}$ de oplossing $\underline{x}$ van het kleinste kwadraten probleem berekent.

```
for k := 1 step 1 until n do
begin comment bepaal de transformatie P_k;
      s := Σ_{i=k}^{m} A_ik^2;
      D_k := if A_kk > 0 then -sqrt(s) else sqrt(s);
      A_kk := A_kk - D_k; beta := -D_k × A_kk;
      comment de elementen A_1k tm A_k-1,k en D_k bevatten nu de k-de kolom R,
      de elementen A_kk tm A_mk bevatten de elementen van w_k;
      for j := k + 1 step 1 until n do
      begin comment pas P_k toe op de j-de kolom van A;
            s := (Σ_{i=k}^{m} A_ik × A_ij)/beta;
            for i := k step 1 until m do A_ij := A_ij - s × A_ik
      end
end;
for k := 1 step 1 until n do
begin comment pas P_k toe op de vector b;
      s := (Σ_{i=k}^{m} A_ik × b_i)/(-D_k × A_kk);
      for i := k step 1 until m do b_i := b_i - s × A_ik
end;
comment het array b bevat nu de vector c = Qb, los nu x op uit R_1x = c_1;
for k := n step -1 until 1 do x_k := (b_k - Σ_{j=k+1}^{n} A_kj × x_j)/D_k.
```

6.4. De Gram-Schmidt algorithme ([2], 5.7.2)

Uit 6.2 en 6.3 volgt: als A een m n-matrix is met onafhankelijke kolommen dan is er een orthogonale m×m-matrix Q zo dat

$$QA = R = \begin{pmatrix} R_1 \\ \hline 0 \end{pmatrix}$$

met R_1 n×n, rechtsboven en regulier.
Dit is equivalent met

$$A = Q^T \begin{pmatrix} R_1 \\ \hline 0 \end{pmatrix} .$$

Noem nu de eerste n kolommen van Q^T $\underline{q}_1,\ldots,\underline{q}_n$ en zij $Q_1 := (\underline{q}_1|\ldots|\underline{q}_n)$.
Dan geldt (omdat de kolommen van Q^T onderling orthonormaal zijn)

$$Q_1^T Q_1 = I_{11} \quad \text{en} \quad A = Q_1 R_1 \tag{1}$$

(met I_{11} de n×n eenheidsmatrix).

Kennis van Q_1 en R_1 is voldoende voor oplossing van het kleinste kwadraten-probleem, want uit (1) volgt

$$A^T A = R_1^T R_1 , \quad A^T \underline{b} = R_1^T Q_1^T \underline{b}$$

zodat de normaalvergelijkingen luiden

$$R_1^T R_1 \underline{x} = R_1^T Q_1^T \underline{b} ,$$

wat (omdat R_1^T regulier is) equivalent is met

$$R_1 \underline{x} = \underline{c}_1 , \quad \underline{c}_1 = Q_1^T \underline{b} .$$

(Merk op dat dit hetzelfde is als de formules (5) en (4) uit 6.2).

De vraag rijst of Q_1 en R_1 die aan (1) voldoen niet rechtstreekser dan via Householdertransformaties bepaald kunnen worden. Het antwoord is ja; men kan de kolommen van Q_1 en R_1 successievelijk uit (1) en de daaruit volgende relatie $R_1 = Q_1^T A$ bepalen. Dit leidt tot de zg. Gram-Schmidt algorithme (die meestal gepresenteerd wordt als orthogonalisatie-algorithme; uit (1) en het feit dat R_1 een rechts bovenmatrix is volgt dat voor k = 1,2,...,n geldt: $\underline{q}_1,\ldots,\underline{q}_k$ vormen een orthonormale basis voor de ruimte opgespannen door de eerste k kolommen van A).

De Gram-Schmidt algorithme is numeriek weinig stabiel, als het conditiegetal van A groot is dan kan de verkregen Q_1 zeer ver van orthogonaliteit afwijken. Veel beter is het zg. gemodificeerde Gram-Schmidt algorithme dat door een kleine wijziging uit het Gram-Schmidt algorithme verkregen wordt (en algebraïsch met Gram-Schmidt equivalent is). Past men deze wijziging ook toe op de bepaling van $Q_1^T \underline{b}$ dan ontstaat een betrouwbaar algorithme voor het kleinste kwadratenprobleem.

Opgemerkt zij nog dat voor de benodigde hoeveelheid vermenigvuldigingen en delingen geldt:

vorming normaalvergelijkingen + Cholesky	$\frac{1}{2} mn^2 + \frac{1}{6} n^3 + \mathcal{O}(mn)$
Householder	$mn^2 - \frac{1}{3} n^3 + \mathcal{O}(mn)$
Gram-Schmidt	$mn^2 + \mathcal{O}(mn)$.

Voor $m = n$ verhouden deze getallen zich ongeveer als $1 : 1 : \frac{3}{2}$, voor $m >> n$ als $1 : 2 : 2$. Hieruit volgt dat bij goed geconditioneerde problemen met $m >> n$ het gebruik van de normaalvergelijkingen niet in alle opzichten verwerpelijk is!

7. Minimalisering van sommen van kwadraten ([2], 10.5.4)

Bij de lineaire kleinste kwadraten aanpassing zoeken we bij gegeven m×n-matrix A en een vector $\underline{b} \in \mathbb{R}^m$ een vector $\underline{x} \in \mathbb{R}^n$ zo dat $\|\underline{b} - A\underline{x}\|$ minimaal is. Een generalisatie hiervan is: zij $\underline{f}$ een gegeven functie die aan een $\underline{x} \in \mathbb{R}^n$ een $\underline{y} = \underline{f}(\underline{x}) \in \mathbb{R}^m$ toevoegt; bepaal dan bij gegeven $\underline{b} \in \mathbb{R}^m$ de $\underline{x}$ waarvoor

$$\varphi(\underline{x}) := \|\underline{f}(\underline{x}) - \underline{b}\|^2$$

$$= \sum_{i=1}^{m} (f_i(\underline{x}) - b_i)^2$$

$$= [\underline{f}(\underline{x}) - \underline{b}]^T[\underline{f}(\underline{x}) - \underline{b}]$$

minimaal is.
We moeten nu dus een som van kwadraten minimaliseren.

Om de in 1.4.5 besproken algemene theorie van minimaliseringsmethoden toe te passen, bepalen we de gradient $\underline{F}$ en de Hessiaan G van φ:

$$F_i(\underline{x}) = \frac{\partial \varphi}{\partial x_i}(\underline{x}) = 2 \frac{\partial \underline{f}^T}{\partial x_i}(\underline{x})\,[\underline{f}(\underline{x}) - \underline{b}]$$

$$G_{ij}(\underline{x}) = \frac{\partial^2 \varphi}{\partial x_i \partial x_j}(\underline{x}) = 2 \frac{\partial \underline{f}^T}{\partial x_i}(\underline{x}) \frac{\partial \underline{f}}{\partial x_j}(\underline{x}) + 2 \frac{\partial^2 \underline{f}^T}{\partial x_i \partial x_j}(\underline{x})[\underline{f}(\underline{x}) - \underline{b}]$$

of

$$\underline{F}(\underline{x}) = 2J^T(\underline{x})[\underline{f}(\underline{x}) - \underline{b}] \quad \text{(een n-vector)},$$

$$G(\underline{x}) = 2J^T(\underline{x})J(\underline{x}) + 2 \sum_{\ell=1}^{m} [f_\ell(\underline{x}) - b_\ell]H_\ell(\underline{x}) \quad \text{(een n×n-matrix)},$$

waarin de m×n-matrix $J(\underline{x})$ en de n×n-matrices H_ℓ bepaald zijn door

$$J_{\ell j}(\underline{x}) = \frac{\partial f_\ell}{\partial x_j}(\underline{x}), \quad \ell = 1,\ldots,m, \quad j = 1,\ldots,n,$$

$$H_{\ell,ij}(\underline{x}) = \frac{\partial^2 f_\ell}{\partial x_i \partial x_j}(\underline{x}), \quad \ell = 1,\ldots,m, \quad i,j = 1,\ldots,n.$$

In 1.4.5 is aangetoond: als

$$\underline{F}(\hat{\underline{x}}) = \underline{0} \text{ en } G(\hat{\underline{x}}) \text{ is positief definiet}$$

dan is $\hat{\underline{x}}$ een (relatief) minimum van φ.

De methode van Newton om een nulpunt $\underline{\hat{x}}$ van $\underline{F}(\underline{x}) = \underline{0}$ te bepalen komt neer op: Zij $\underline{x}_{k-1}$ de laatst bepaalde benadering voor $\underline{\hat{x}}$; bepaal $\underline{d}_k$ uit

$$G(\underline{x}_{k-1})\underline{d}_k = -\underline{F}(\underline{x}_{k-1}) \ ;$$

neem

$$\underline{x}_k = \underline{x}_{k-1} + \underline{d}_k \ .$$

Een modificatie is:

neem $\underline{x}_k = \underline{x}_{k-1} + \lambda_k \underline{d}_k$

met λ_k zo dat

$$\varphi(\underline{x}_k) < \varphi(\underline{x}_{k-1}) \ ,$$

en eventueel zo dat $\varphi(\underline{x}_k)$ het minimum van $\varphi(\underline{x})$ is op de lijn $\underline{x} = \underline{x}_{k-1} + \lambda \underline{d}_k$ (Newton met lijnminimalisering).

In het geval dat φ een som van kwadraten is vervangt men in deze methoden de matrix $G(\underline{x})$ graag door

$$\tilde{G}(\underline{x}) := 2J^T(\underline{x})J(\underline{x})$$

met als argumenten

i) als $\underline{f}$ bijna lineair is (dus $\frac{\partial^2 \underline{f}}{\partial x_i \partial x_j}$ "klein") of als de residuen "klein" zijn (dus $\underline{f}(\underline{x}) - \underline{b}$ "klein") dan is $\tilde{G}$ dicht bij G.

ii) minder rekenwerk: het bepalen van $\frac{\partial^2 \underline{f}}{\partial x_i \partial x_j}$ is vaak veel duurder dan de bepaling van J.

iii) als de kolommen van J (dus de vectoren $\frac{\partial \underline{f}}{\partial x_1}, \ldots, \frac{\partial \underline{f}}{\partial x_n}$) onafhankelijk zijn dan is $\tilde{G}$ positief definiet en de methode een descent methode (zie 1.4.5).

De aldus ontstane methode wordt methode van Gauss-Newton genoemd. Als aan de voorwaarden i) en iii) voldaan is dan is de convergentie zeer bevredigend.

We merken nog op dat bij Gauss-Newton de vector $\tilde{\underline{d}}_k = -\tilde{G}(\underline{x}_{k-1})^{-1}\underline{F}(\underline{x}_{k-1})$ voldoet aan

$$J^T(\underline{x}_{k-1})J(\underline{x}_{k-1})\tilde{\underline{d}}_k = -J^T(\underline{x}_{k-1})[f(\underline{x}_{k-1}) - \underline{b}] \ .$$

Dat zijn de normaalvergelijking van het lineair kleinste kwadratenprobleem: minimaliseer

$$\| J(\underline{x}_{k-1})\tilde{\underline{d}} + \underline{f}(\underline{x}_{k-1}) - \underline{b} \|$$

als functie van $\tilde{\underline{d}}$.
Daar volgens Taylor (zie 1.4.2 formule (7))

$$\underline{f}(\underline{x}) - \underline{b} = \underline{f}(\underline{x}_{k-1}) - \underline{b} + J(\underline{x}_{k-1})(\underline{x} - \underline{x}_{k-1}) + \mathcal{O}(\| x - x_{k-1} \|^2)$$

is het verschil tussen Newton en Gauss-Newton ook te formuleren als:
Bij Newton wordt in de k-de slag de object functie $\varphi(\underline{x}) = \| \underline{f}(\underline{x}) - \underline{b} \|^2$ **vervangen door de kwadratische benadering**

$$\varphi(\underline{x}_{k-1}) + \underline{F}^T(\underline{x}_{k-1})(\underline{x} - \underline{x}_{k-1}) + \tfrac{1}{2}(\underline{x} - \underline{x}_{k-1})^T G(\underline{x}_{k-1})(\underline{x} - \underline{x}_{k-1})$$

en deze is minimaal in $\underline{x}_{k-1} + \underline{d}_k$.
Bij Gauss-Newton wordt de functie $\underline{f}(\underline{x}) - \underline{b}$ vervangen door de lineaire benadering

$$\underline{f}(\underline{x}_{k-1}) - \underline{b} + J(\underline{x}_{k-1})(\underline{x} - \underline{x}_{k-1})$$

en de norm van deze benadering is minimaal in $\underline{x}_{k-1} + \tilde{\underline{d}}_k$.

Een variant op de methode van Gauss-Newton is de tegenwoordig veel gebruikte methode van Marquardt. Hierbij wordt $\underline{d}_k^*$ bepaald uit

$$(2J^T(\underline{x}_{k-1})J(\underline{x}_{k-1}) + \mu_k I)\underline{d}_k^* = -2J^T(\underline{x}_{k-1})[\underline{f}(\underline{x}_{k-1}) - \underline{b}] .$$

Er wordt geen lijnminimalisering toegepast, maar $\mu_k > 0$ wordt zo gekozen dat $\varphi(\underline{x}_{k-1} + \underline{d}_k^*) < \varphi(\underline{x}_{k-1})$. Daar $2J^T(\underline{x})(\underline{f}(\underline{x}) - \underline{b})$ de gradient van $\varphi(\underline{x})$ is, is voor voldoende grote μ_k zeker aan deze voorwaarde voldaan.
Omdat $\mu_k = 0$ correspondeert met Gauss-Newton kiest men μ_k liefst zo klein mogelijk, b.v. met de volgende strategie:

1. Kies $\nu > 1$;
2. Kies $\underline{x}_0$, kies $\mu_1 > 0$, zet $k := 1$;
3. Bepaal $\underline{d}_k^*$ en $\underline{x}_k := \underline{x}_{k-1} + \underline{d}_k^*$;
4. Als $\varphi(\underline{x}_k) \geq \varphi(x_{k-1})$ neem dan $\mu_k := \nu\mu_k$ en ga terug naar 3 ;
5. Als niet aan een stopcriterium voldaan is neem dan $\mu_{k+1} = \mu_k/\nu$, zet $k := k+1$ en ga naar 3.

Mogelijke stopcriteria zijn bv. $\| x_k - x_{k-1} \| \leq \varepsilon$ of $\| \underline{F}(\underline{x}_k) \| \leq \varepsilon$ met geschikte waarden van ε (en daar zit de moeilijkheid want dit, evenals de keus van μ, hangt af van schaling e.d.).

8. Parameterschatting

Men spreekt van parameterschatting als een stel parameters in een model zo bepaald wordt dat de uitkomsten van het model zo goed mogelijk aansluiten bij een rij meetwaarden. Afhankelijk van hoe de parameters in het model voorkomen en van wat we onder zo goed mogelijk verstaan geeft dit probleem aanleiding tot zeer verschillende methodieken.

Voorbeeld

Zij een grootheid y afhankelijk van een instelbare of afleesbare parameter t (die continue of discrete waarden kan doorlopen) en van onbekende (maar van t onafhankelijke) parameters $x_1,\ldots,x_n$. Bijvoorbeeld

$$y = f(t,x_1,x_2) = x_1 e^{-\alpha_1 t} + x_2 e^{-\alpha_2 t} ,$$

$$y = f(t,x_1,x_2,x_3,x_4) = x_1 e^{-x_3 t} + x_2 e^{-x_4 t} .$$

In het eerste geval zijn α_1 en α_2 bekend, y hangt dus lineair van x_1 en x_2 af; in het tweede geval zijn x_3 en x_4 ook onbekend, y hangt daar niet-lineair van af.

Gevraagd wordt, uit metingen van y voor bekende parameterwaarden $t_1,\ldots,t_m$ de onbekende parameters $x_1,\ldots,x_n$ te bepalen. Als regel wordt bij de i-de meting een fout e_i gemaakt, zodat voor de meetvector $\underline{b} = (b_1,\ldots,b_m)^T$ geldt

$$b_i = f(t_i,x_1,\ldots,x_n) + e_i,\ i = 1,\ldots,m$$

of (met $f_i(\underline{x}) := f(t_i,\underline{x})$)

$$\underline{b} = \underline{f}(\underline{x}) + \underline{e} .$$

Als $\underline{e} = \underline{0}$ dan staan hier m vergelijkingen met n onbekenden die in het algemeen slechts een oplossing hebben als $m = n$. Is $\underline{e} \neq \underline{0}$ dan leert de statistiek dat, onder bepaalde veronderstellingen omtrent het statistisch gedrag van de e_i, de vector $\hat{\underline{x}}$ die bij een gegeven meetvector $\underline{b}$ de uitdrukking

$$\varphi(\underline{x}) = \| \underline{f}(\underline{x}) - b \|_2^2$$

minimaliseert, de "beste schatting" is voor de onbekende parameter-vector $\underline{x}$. In dit geval is het gunstig als m flink wat groter is dan n.

Deze probleemstelling leidt dus tot kleinste kwadraten aanpassing, hetzij lineair (als $\underline{f}(\underline{x}) = A\underline{x} + \underline{a}$) of niet-lineair, zodat de methoden van hoofdstukken

6 en 7 toepasbaar zijn. Indien y_i met onderling verschillende nauwkeurigheden σ_i gemeten worden dan moet men niet $\sum (f_i(\underline{x}) - b_i)^2$ maar $\sum \sigma_i^{-2}(f_i(\underline{x}) - b_i)^2$ minimaliseren.

Een voorbeeld van een geheel andere situatie is de volgende.

Van een onbekende functie f(t) kunnen waarden $f(t_i)$ in op te geven punten $t_0,\ldots,t_n$ uit $[-1,1]$ bepaald worden (door meting of berekening met een ontoegankelijk proces); gevraagd worden waarden van de afgeleide f'(t). Natuurlijk is dit probleem onbepaald als men verder niets over de functie f weet. Weet men dat f aan zekere gladheidseisen voldoet - bijv. dat de afgeleiden begrensd zijn - dan ligt de volgende methode voor de hand.
Kies een functie $g(t,\underline{x})$ die behalve van t afhangt van n+1 parameters $x_0,\ldots,x_n$. Bepaal deze parameters zo dat

$$g(t_i,\underline{x}) = f(t_i), \quad i = 0,\ldots,n , \tag{1}$$

d.w.z., zo dat $g(t,\underline{x})$ de functie f(t) <u>interpoleert</u> in de punten $t_0,\ldots,t_n$. En neem nu

$$\frac{\partial g}{\partial t}\,(t,\underline{x})$$

als benadering voor voor f'(t). Natuurlijk hangt het behalve van de gladheid van f, sterk af van de keuze van g en de punten $t_0,\ldots,t_n$ of we aldus een goed resultaat krijgen.

Een klassieke keuze voor g is: g is een polynoom in t met graad $\leq$ n:

$$g(t,\underline{c}) = \sum_0^n c_j t^j$$

(we schrijven nu maar c_j in plaats van x_j). Men kan eenvoudig bewijzen dat er bij ieder stel waarden $f(t_0),\ldots,f(t_n)$ precies één interpolerend polynoom g is. Immers, de vergelijkingen (1) vormen n+1 lineaire vergelijkingen voor de n+1 onbekenden $c_0,\ldots,c_n$. En het bijbehorende homogene stelsel heeft uitsluitend de nul-oplossing omdat een niet-triviaal polynoom met graad n niet meer dan n verschillende nulpunten heeft.
We gaan niet in op methoden om het interpolatiepolynoom effectief te bepalen. Vermeld zij slechts dat voor circa $n \geq 4$ de methoden bewerkelijk en gevoelig voor afrondfouten worden en dat toename van de nauwkeurigheid met toenemende n niet verzkerd is. Indien men vrijheid heeft in de keuze van de punten t_i

dan is een equidistante verdeling ($t_i = 1 - 2i/n$) niet de beste, veel beter is $t_i = \cos(\pi i/n)$ of $t_i = \cos(\pi(2i+1)/(2n+2))$, waarbij een verdichting van de punten t_i bij de eindpunten ± 1 optreedt.

Een moderne manier om een "gladde" functie door een aantal gegeven punten te leggen is de zg. spline-interpolatie. Onder zwakke voorwaarden voor de functie f worden hierbij ook de afgeleiden goed benaderd door de afgeleiden van de interpolant.
Een zeer veel gebruikt bijzonder geval van spline-interpolatie wordt in de volgende paragraaf behandeld.

8.1. Interpolatie met kubische splines ([2], 4.6)

Zij $x_0 < x_1 < \ldots < x_n$. Een kubische spline in het interval $[x_0,x_n]$ met als knooppunten $x_1,\ldots,x_{n-1}$ is een functie s die voldoet aan

i) s, s' en s" zijn continu in $[x_0,x_n]$,

ii) s"' bestaat in ieder der open intervallen $(x_0,x_1),\ldots,(x_{n-1},x_n)$ en is daar constant.

De functie s is dus stuksgewijs een derdegraads polynoom en deze polynomen sluiten, evenals hun eerste en tweede afgeleiden, in de knooppunten continu aan (de derde afgeleide is daar in het algemeen discontinu).
Daar we n intervallen en n-1 knooppunten hebben, een derdegraads polynoom lineair afhangt van vier parameters en in ieder knooppunt drie continuiteitseisen gelden die aanleiding geven tot lineaire relaties tussen de parameters hangt s lineair af van

$$4n - 3(n-1) = n + 3$$

parameters. We kunnen dus bij een gegeven functie f(x) vermoedelijk precies één kubische spline vinden die voldoet aan de interpolatie-eis

iii) $\quad s(x_j) = f(x_j), \quad j = 0,\ldots,n$

en aan twee randvoorwaarden, b.v.

iva) $\quad s'(x_0) = f'(x_0),\ s'(x_n) = f'(x_n)$

of

ivb) $\quad s''(x_0) = f''(x_0),\ s''(x_n) = f''(x_n)$

of

ivc) $$\frac{s''(x_2) - s''(x_1)}{x_2 - x_1} = \frac{s''(x_1) - s''(x_0)}{x_1 - x_0} ,$$

$$\frac{s''(x_n) - s''(x_{n-1})}{x_n - x_{n-1}} = \frac{s''(x_{n-1}) - s''(x_{n-2})}{x_{n-1} - x_{n-2}} .$$

Hieronder bewijzen we dat dit vermoeden juist is en we geven aan hoe s eenvoudig te berekenen is.
Men kan nu de volgende approximatiestelling bewijzen.

Zij bij gegeven $x_0,\ldots,x_n$

$$\text{hmin} := \min(x_{j+1} - x_j),\ \text{hmax} := \max(x_{j+1} - x_j) .$$

Dan geldt voor een viermaal continu differentieerbare f: er zijn constanten $C_0,\ldots,C_3$, alleen afhankelijk van f en **hmin/hmax, zo dat in het interval** $[x_0,x_n]$ **geldt**

$$|s^{(j)}(x) - f^{(j)}(x)| \le C_j(\text{hmax})^{4-j}, \quad j = 0,1,2,3 .$$

Om de interpolerende spline te bepalen stellen we

$$s'(x_j) = A_j ,\ s''(x_j) = B_j , \quad j = 0,\ldots,n .$$

Dan geldt (waarom) voor $x_j \le x \le x_{j+1}$

$$s''(x) = B_j \frac{x_{j+1}-x}{h_{j+\frac{1}{2}}} + B_{j+1} \frac{x-x_j}{h_{j+\frac{1}{2}}} ,$$

waarin $h_{j+\frac{1}{2}} = x_{j+1} - x_j$. Hieruit volgt (ga na) dat voor $x_j \le x \le x_{j+1}$

$$s(x) = \frac{x_{j+1}-x}{h_{j+\frac{1}{2}}} [f_j - \tfrac{1}{6}B_j(h^2_{j+\frac{1}{2}} - (x_{j+1} - x)^2)]$$

$$+ \frac{x-x_j}{h_{j+\frac{1}{2}}} [f_{j+1} - \tfrac{1}{6}B_{j+1}(h^2_{j+\frac{1}{2}} - (x - x_j)^2)] ,$$

waaruit blijkt dat

$$A_j = \frac{f_{j+1}-f_j}{h_{j+\frac{1}{2}}} - \frac{1}{6} h_{j+\frac{1}{2}} (2B_j + B_{j+1}) ,$$

$$A_{j+1} = \frac{f_{j+1}-f_j}{h_{j+\frac{1}{2}}} + \frac{1}{6} h_{j+\frac{1}{2}} (B_j + 2B_{j+1}) .$$

Vervangen we in de tweede relatie j+1 door j dan vinden we dat voor $j = 1,\ldots,n-1$ moet gelden

$$\frac{1}{6}h_{j-\frac{1}{2}} (B_{j-1} + 2B_j) + \frac{1}{6}h_{j+\frac{1}{2}} (2B_j + B_{j+1}) = \frac{f_{j+1}-f_j}{h_{j+\frac{1}{2}}} - \frac{f_j-f_{j-1}}{h_{j-\frac{1}{2}}} .$$

Dit zijn n-1 lineaire vergelijkingen voor $B_0,\ldots,B_n$. De randvoorwaarden voegen hier twee vergelijkingen aan toe, namelijk

$$\frac{1}{6}h_{\frac{1}{2}} (2B_0 + B_1) = \frac{f_1-f_0}{h_{\frac{1}{2}}} - f_0' ,$$

resp.

$$B_0 = f_0'' ,$$

resp.

$$\frac{B_2-B_1}{h_{3/2}} = \frac{B_1-B_0}{h_{\frac{1}{2}}}$$

en analoog aan de andere rand. Met behulp hiervan kunnen we B_0 en B_n elimineren en we houden dan een $(n-1) \times (n-1)$ stelsel over waarvan de matrix tridiagonaal en diagonaaldominant is. Hieruit volgt zowel de eenduidige existentie van s als de eenvoudige berekenbaarheid.

Opmerking

Zij $\hat{s}$ de hierboven eenduidig bepaalde spline die aan de eisen i), ii), iii) en iva) voldoet. Dan geldt voor iedere functie s die aan de eisen i), iii) en iva) voldoet

$$\int_{x_0}^{x_n} [s''(x)]^2 dx \geq \int_{x_0}^{x_n} [\hat{s}''(x)]^2 dx$$

met gelijkteken dan en slechts dan als $s = \hat{s}$.

Deze uitspraak geeft aan in welke zin $\hat{s}$ onder alle redelijk gladde (eis i)) functies s die interpoleren (eis iii)) en aan de randvoorwaarden voldoen (eis iva)) de gladste is.

9. Eigenwaarden en eigenvectoren van matrices ([2], 5.8; [16], ch. 14, [18])

9.1. Inleiding. Voorbeelden ([18], ch. 2)

Definitie. Een getal λ is eigenwaarde van een $n \times n$-matrix A als $A - \lambda I$ singulier is.
Als λ eigenwaarde van A is dan heten de vectoren $\underline{x} \neq \underline{0}$ waarvoor $A\underline{x} = \lambda\underline{x}$ eigenvectoren van A bij λ.

Nodig en voldoende opdat λ eigenwaarde is, is dat λ voldoet aan de zg. karakteristieke vergelijking

$$\det(\lambda I - A) = 0 .$$

Het linkerlid is een n-de graads polynoom in λ, heeft dus minstens één nulpunt (eventueel complex; als A reëel is dan komen complexe eigenwaarden steeds als toegevoegd complexe paren voor). En bij iedere eigenwaarde λ is er minstens één eigenvector. Is de nulruimte van $A - \lambda I$ m-dimensionaal, dan zijn er bij λ m onafhankelijke eigenvectoren te vinden, iedere lineaire combinatie daarvan (behalve de triviale) is dan ook eigenvector bij λ. Men noemt m de geometrische multipliciteit van λ. De algebraische multipliciteit is de multipliciteit van λ als oplossing van de karakteristieke vergelijking. Men kan bewijzen dat de algebraische multipliciteit niet kleiner is dan de geometrische. Als de geometrische multipliciteit van λ kleiner is dan de algebraische dan heet λ deficient, dergelijke eigenwaarden veroorzaken veel last.

Voorbeeld.

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} .$$

De karakteristieke vergelijking is

$$\lambda^2 = 0 .$$

$\lambda = 0$ is dus de enige eigenwaarde en heeft algebraische multipliciteit 2. Iedere eigenvector van A is echter een veelvoud van $(1,0)^T$ zodat de geometrische multipliciteit 1 is.

Als $\underline{x}_1,\ldots,\underline{x}_p$ eigenvectoren zijn bij onderling verschillende eigenwaarden dan zijn ze onafhankelijk. Want stel dat

$$\sum_1^p \gamma_j \underline{x}_j = \underline{0} .$$

Vermenigvuldig deze relatie met $(A-\lambda_2 I)\ldots(A-\lambda_p I)$. Dan volgt uit $A\underline{x}_j = \lambda_j \underline{x}_j$ dat

$$(\lambda_1 - \lambda_2)\ldots(\lambda_1 - \lambda_p)\gamma_1 \underline{x}_1 = \underline{0}$$

en daaruit (waarom?) dat $\gamma_1 = 0$. Analoog voor $\gamma_2,\ldots,\gamma_p$.
Hieruit volgt eenvoudig dat als A n onderling verschillende eigenwaarden heeft, er n onafhankelijke eigenvectoren zijn. Dit geldt ook als geen der eigenwaarden van A deficient is (want de som der geometrische multipliciteiten is dan gelijk aan de graad van de karakteristieke vergelijking, dus n).

Stel nu dat $\underline{u}_1,\ldots,\underline{u}_n$ onafhankelijke eigenvectoren zijn van A bij (al dan niet onderling verschillende) eigenwaarden $\lambda_1,\ldots,\lambda_n$. Zij de matrix U gedefinieerd door

$$U := (\underline{u}_1 | \ \ldots \ | \underline{u}_n) .$$

Daar $\underline{u}_1,\ldots,\underline{u}_n$ onafhankelijk zijn is U regulier.
Uit $A\underline{u}_j = \lambda_j \underline{u}_j$ volgt

$$AU = (\lambda_1 \underline{u}_1 | \ \ldots | \lambda_n \underline{u}_n) = U\Lambda$$

met

$$\Lambda := \mathrm{diag}(\lambda_1,\ldots,\lambda_n) .$$

We kunnen dit schrijven als

$$U^{-1}AU = \Lambda \qquad (1)$$

en ook als

$$A = U\Lambda U^{-1} . \qquad (2)$$

De eerste vorm zegt: A is gelijkvormig met Λ (A en B heten gelijkvormig als er een reguliere S is zo dat $S^{-1}AS = B$). De vorm (2) is de

zg. spectraalvoorstelling van A. Hij is nuttig als men machten of algemenere functies van A wil berekenen:

$$A^k = U\Lambda^k U^{-1} .$$

Als A deficient is, dan is er geen n-tal onafhankelijke eigenvectoren. Er geldt dan niet een spectraalvoorstelling (2) of anders: A is niet diagonaliseerbaar. Ook als A zeer naburige eigenwaarden heeft dan kunnen er numerieke moeilijkheden ontstaan omdat het conditiegetal van U groot kan zijn. Voorbeeld:

$$A = \begin{pmatrix} 0 & 1 \\ \alpha & 0 \end{pmatrix}$$

heeft eigenwaarden $\lambda_1 = \alpha^{\frac{1}{2}}$, $\lambda_2 = -\alpha^{\frac{1}{2}}$. Als $|\alpha| << 1$ dan zijn deze zeer naburig (want $|\lambda_1 - \lambda_2| << \|A\|$). Als $\alpha \neq 0$ dan kunnen we voor U nemen (ga na)

$$U = \begin{pmatrix} 1 & 1 \\ \alpha^{\frac{1}{2}} & -\alpha^{\frac{1}{2}} \end{pmatrix}$$

met inverse

$$U^{-1} = \tfrac{1}{2}\begin{pmatrix} 1 & \alpha^{-\frac{1}{2}} \\ 1 & -\alpha^{-\frac{1}{2}} \end{pmatrix} .$$

Ga na dat in de 2-norm

$$c(U) = \max(|\alpha|^{\frac{1}{2}}, |\alpha|^{-\frac{1}{2}}) .$$

Zonder bewijs (zie hiervoor Wiskunde 30, 1.5.4.) merken we op dat van een symmetrische matrix ($A^T = A$) de eigenwaarden steeds reëel zijn, dat eigenvectoren bij onderling verschillende eigenwaarden onderling orthogonaal zijn en dat hier een eigenwaarde λ nooit deficient is, zodat bij een m-voudig nulpunt van $\det(\lambda I - A)$ de nulruimte van $A - \lambda I$ m-dimensionaal is; hierin kunnen we een basis van m onderling orthogonale eigenvectoren kiezen. Een symmetrische matrix heeft dus steeds n onafhankelijke eigenvectoren; kiezen we ze onderling orthogonaal en met lengte 1 (in de 2-norm) dan geldt

$$\underline{u}_i^T \underline{u}_j = \delta_{ij} ,$$

dat wil zeggen

$$U^T U = I ,$$

dus U is orthogonaal en c(U) = 1 in de 2-norm. We kunnen in dit geval voor (2) ook schrijven

$$A = U\Lambda U^T . \tag{3}$$

We noemen enige gevolgen van (3).

a. Zij $\underline{x} = U\underline{c} = \sum \gamma_j \underline{u}_j$ (als $\underline{c} = (\gamma_1,\ldots,\gamma_n)^T$).
Dan volgt uit (3), als $\underline{x} \neq 0$,

$$\frac{\underline{x}^T A\underline{x}}{\underline{x}^T\underline{x}} = \frac{\underline{c}^T U^T AU\underline{c}}{\underline{c}^T U^T U\underline{c}} = \frac{\underline{c}^T\Lambda\underline{c}}{\underline{c}^T\underline{c}} = \frac{\sum \gamma_j^2\lambda_j}{\sum \gamma_j^2} . \tag{4}$$

Daar de λ's reëel zijn kunnen we ze zo nummeren dat $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.
Uit (4) volgt dan direct (ga na)

$$\max_{\underline{x}\neq\underline{0}} \frac{\underline{x}^T A\underline{x}}{\underline{x}^T\underline{x}} = \lambda_1 , \quad \min_{\underline{x}\neq\underline{0}} \frac{\underline{x}^T A\underline{x}}{\underline{x}^T\underline{x}} = \lambda_n .$$

b. A (symmetrisch) is positief definiet dan en slechts dan als alle eigenwaarden positief zijn (ga na).

c. Zij B een m×n-matrix en $A = B^T B$ (n×n, symmetrisch, met eigenwaarden $\lambda_1,\ldots,\lambda_n$ zo dat $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$, waarom kan dat?)
Uit a. volgt

$$\lambda_1 = \max_{\underline{x}\neq\underline{0}} \frac{\underline{x}^T A\underline{x}}{\underline{x}^T\underline{x}} = \max_{\underline{x}\neq\underline{0}} \frac{\|B\underline{x}\|_2^2}{\|\underline{x}\|_2^2}$$

en dus (vgl. 1.4, formule (4)):

$$\|B\|_2 = \max_{\underline{x}\neq\underline{0}} \frac{\|B\underline{x}\|_2}{\|\underline{x}\|_2} = \lambda_1^{\frac{1}{2}} .$$

Dit resultaat werd reeds in 1.4 genoemd en in 6.1 gebruikt.
Ook geldt

$$\min_{\underline{x}\neq\underline{0}} \frac{\|B\underline{x}\|^2}{\|\underline{x}\|^2} = \lambda_n .$$

Als de kolommen van B onafhankelijk zijn dan is er geen $\underline{x} \neq \underline{0}$ zodat $B\underline{x} = \underline{0}$ en dan is $\lambda_n > 0$. In de zin van de definitie in 6.1 geldt dan

$$c(B) = (\lambda_1/\lambda_n)^{\frac{1}{2}} \quad \text{(in de 2-norm)}.$$

Als B zelf symmetrisch is met eigenwaarden $\mu_1,\ldots,\mu_n$ dan is $A = B^2$, de eigenwaarden van A zijn $\mu_1^2,\ldots,\mu_n^2$ en dus is voor symmetrische B

$$\|B\|_2 = \max |\mu_j| \ .$$

d. Zij B een m×n-matrix met $m \geq n$. Zij $B^TB = U\Lambda U^T$ met $U^TU = I$ en $\Lambda = \text{diag}\,(\lambda_1,\ldots,\lambda_n)$, waarin $\lambda_1 \geq \ldots \geq \lambda_r > \lambda_{r+1} = \ldots = \lambda_n = 0$ (als $r = n$ dan is $\lambda_n > 0$). Zij $\Sigma_1 = \text{diag}\,(\sigma_1,\ldots,\sigma_r)$ met $\sigma_j = \lambda_j^{\frac{1}{2}}$ en U_1 de matrix bestaande uit de eerste r kolommen van U. Dan geldt ook $B^TB = U_1\Sigma_1^2U_1^T$, $U_1^TU_1 = I_1$ (r×r). Uit $(BU_1)^T(BU_1) = U_1^TB_1^TB_1U_1 = \Sigma_1^2$ volgt dat de kolommen van BU_1 onderling orthogonaal zijn en lengten $\sigma_1,\ldots,\sigma_r$ (> 0) hebben. Voor de m×r-matrix $V_1 := BU_1\Sigma_1^{-1}$ geldt dan $V_1^TV_1 = I_1$ (r×r).
Voor de matrix U_2, bestaande uit de laatste n-r-kolommen van U volgt uit $U^TB^TBU = \Lambda$ en het feit dat de n-r laatste λ's nul zijn dat $U_2^TB^TBU_2 = 0$ en dus ook (waarom?) $BU_2 = 0$. Breiden we nu nog V_1 uit tot een orthogonale matrix V (m×m) dan geldt

$$BU = B(U_1|U_2) = (V_1\Sigma_1|0) = (V_1|V_2)\left(\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right) = V\Sigma \ ,$$

waarin Σ m×n is. Na een analoog verhaal voor het geval $m \leq n$ is aldus bewezen: bij iedere m×n-matrix B zijn er orthogonale matrices U (n×n) en V (m×m) zo dat

$$B = V\Sigma U^T \ , \qquad (5)$$

waarin Σ m×n en "diagonaal" is met niet-triviale diagonaalelementen $\sigma_1 \geq \ldots \geq \sigma_r > 0$. De σ's heten de singuliere waarden (singular values) van B, (5) heet de singular value decompositie van B, het getal r is de rang van B (zie ook [2], 5.2.5).
Uit de singular value decompositie zien we

. B heeft een m-r dimensionale nulruimte (opgespannen door de kolommen van U_2) en een r dimensionale range (opgespannen door V_1)
. B^T heeft een n-r dimensionale nulruimte en een r dimensionale range
. $B^TB = U\Sigma^2U^T$, eigenwaarden $\sigma_1^2,\ldots,\sigma_r^2$ en 0 (n-r voudig)
. $BB^T = V\Sigma^2V^T$, eigenwaarden $\sigma_1^2,\ldots,\sigma_r^2$ en 0 (m-r voudig)
. $\|B\|_2 = \sigma_1$
. B is regulier als $r = m = n$, dan is $B^{-1} = U\Sigma^{-1}V^T$ en $\|B^{-1}\|_2 = \sigma_n^{-1}$
. als $r = n$ dan zijn de kolommen van B onafhankelijk, $\sigma_n > 0$ en

$$c(B) = \sigma_1/\sigma_n .$$

Met behulp van de singuliere waarden kunnen we goed uitmaken of een matrix B "bijna" singulier is (dicht ligt bij een matrix met rang $< \min(m,n)$). Er geldt namelijk: als $\sigma_{k+1}(B) \leq \varepsilon$ dan is er een m×n-matrix $\tilde{B}$ met rang k zo, dat $\|B - \tilde{B}\|_2 \leq \varepsilon$. Neem namelijk $\tilde{B} = \tilde{U}_1\tilde{\Sigma}_1\tilde{V}_1^H$, waarin $\tilde{U}_1$ en $\tilde{V}_1$ bestaan uit de eerste k kolommen van U, resp. V en $\tilde{\Sigma}_1 = \text{diag}(\sigma_1,\ldots,\sigma_k)$. De singuliere waarden van $B - \tilde{B}$ zijn dan (ga na) $\sigma_{k+1},\ldots,\sigma_r$ en dus is $\|B - \tilde{B}\| = \sigma_{k+1} \leq \varepsilon$.

e. Zij B als in (5) met

$$\Sigma = \left(\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right) \qquad (m \times n)$$

Definieer

$$\Sigma^+ := \left(\begin{array}{c|c} \Sigma_1^{-1} & 0 \\ \hline 0 & 0 \end{array}\right) \qquad (n \times m)$$

en

$$B^+ := U\Sigma^+V^T .$$

Dan is B^+ een n×m-matrix met als eigenschappen (ga na, bekijk eerst het geval dat $B = \Sigma$)

$$BB^+B = B$$

$$B^+BB^+ = B^+$$

$$BB^+ = V_1V_1^T \text{, symmetrisch}$$

$$B^+B = U_1U_1^T \text{, symmetrisch .}$$

Men kan bewijzen dat B^+ de enige matrix is met deze vier eigenschappen. B^+ heet pseudo-inverse van B (ook wel: Moore-Penrose generalized inverse). Er geldt:

. als $r = m = n$ (dus B regulier) dan is $B^+ = B^{-1}$

. als $r = n < m$ (dus kolommen van B onafhankelijk) dan is $B^+B = I$, $B^+ = (B^TB)^{-1}B^T$

. als $r = m < n$ (dus rijen van B onafhankelijk) dan is $BB^+ = I$, $B^+ = B^T(BB^T)^{-1}$.

De pseudo-inverse is o.a. van belang in verband met de volgende

Stelling

$\hat{\underline{x}}$ is oplossing van het kleinse kwadratenprobleem

$$\text{minimaliseer } \| B\underline{x} - \underline{b} \|_2$$

dan en slechts dan als

$$B(\hat{\underline{x}} - B^+\underline{b}) = \underline{0} .$$

Voor alle minimaliserende $\hat{\underline{x}}$ geldt $B\hat{\underline{x}} = BB^+\underline{b}$ en

$$\| \hat{\underline{x}} \| \geq \| B^+\underline{b} \|$$

met gelijkteken dan en slechts dan als $\hat{\underline{x}} = B^+\underline{b}$.

$B^+\underline{b}$ is dus de "kleinste kleinste kwadraten oplossing".
Merk op dat als de kolommen van B onafhankelijk zijn, $B\underline{z} = \underline{0}$ impliceert $\underline{z} = \underline{0}$. In dit geval is $B^+\underline{b} = (B^TB)^{-1}B^T\underline{b}$ de enige $\hat{\underline{x}}$ die $\| B\underline{x} - \underline{b} \|$ minimaliseert (vgl. 6.1).

Bewijs.

$$B\underline{x} - \underline{b} = B(\underline{x} - B^+\underline{b}) - (I - BB^+)\underline{b} .$$

Daar

$$(I - BB^+)^TB = (I - BB^+)B = 0$$

volgt hieruit (Pythagoras)

$$\| B\underline{x} - \underline{b} \|^2 = \| B(\underline{x} - B^+\underline{b}) \|^2 + \| (I - BB^+)\underline{b} \|^2 \geq \| (I - BB^+)\underline{b} \|^2$$

met gelijkteken dan en slechts dan als $B(\underline{x} - B^+\underline{b}) = \underline{0}$.

Dit bewijst het eerste deel van de stelling.

Zij nu $\underline{\hat{x}} = B^{+}\underline{b} + \underline{z}$ met $B\underline{z} = \underline{0}$. Dan is

$$(B^{+})^{T}\underline{z} = (B^{+}BB^{+})^{T}\underline{z} = (B^{+})^{T}(B^{+}B)^{T}\underline{z} = (B^{+})^{T}B^{+}B\underline{z} = \underline{0}$$

en dus (Pythagoras)

$$\|\underline{\hat{x}}\|^2 = \|B^{+}\underline{b}\|^2 + \|\underline{z}\|^2 \geq \|B^{+}\underline{b}\|^2$$

met gelijkteken dan en slechts dan als $\underline{z} = \underline{0}$. □

We noemen nu nog enige voor alle matrices geldende eigenschappen die van belang zijn voor de bepaling van de eigenwaarden.

a. Als A eigenwaarde $\lambda_1,\ldots,\lambda_p$ heeft dan heeft $A-\alpha I$ eigenwaarden $\lambda_1-\alpha,\ldots\lambda_p-\alpha$ met dezelfde multipliciteiten en dezelfde eigenvectoren, want $A\underline{x} = \lambda\underline{x}$ impliceert $(A-\alpha I)\underline{x} = (\lambda-\alpha)\underline{x}$.

b. Als S regulier is en $B = S^{-1}AS$ dan heten B en A gelijkvormig. Gelijkvormige matrices hebben dezelfde eigenwaarden (inclusief multipliciteiten) en de eigenvectoren hangen eenvoudig samen, want $\det(\lambda I - B) = \det(S^{-1}(A-\lambda I)S) = \det(S^{-1})\det(A-\lambda I)\det(S) = \det(A-\lambda I)$. En $A\underline{x} = \lambda\underline{x}$ impliceert $B\underline{y} = \lambda\underline{y}$ met $\underline{y} = S^{-1}\underline{x}$.

c. Als Q orthogonaal is dan is Q regulier, $Q^{-1} = Q^{T}$ en $B := Q^{T}AQ$ is dus gelijkvormig met A. Daar symmetrie van A symmetrie van B impliceert, werken we bij symmetrische matrices graag met deze orthogonale gelijkvormigheidstransformaties. Omdat orthogonale gelijkvormigheidstransformaties numeriek zeer stabiel uit te voeren zijn gebruiken we ze ook veel (maar op andere gronden!) bij niet symmetrische matrices.

We bespreken nu enige voorbeelden waaruit blijkt dat kennis van eigenwaarden en eigenvectoren van een matrix nuttig is.

a) Beschouw het stelsel van n homogene lineaire differentiaalvergelijkingen

$$\frac{d\underline{x}}{dt} = A\underline{x} \ . \tag{6}$$

De elementaire oplossingsmethode zoekt oplossingen van de vorm

$$\underline{x}(t) = e^{\lambda t}\underline{u} \ ,$$

Deze voldoet als λ eigenwaarde is van A en $\underline{u}$ een bijbehorende eigenvector. Heeft A nu n onafhankelijke eigenvectoren $\underline{u}_1,\ldots,\underline{u}_n$ bij eigenwaarden $\lambda_1,\ldots,\lambda_n$ dan is de algemene oplossing van (4)

$$\underline{x}(t) = \Sigma\gamma_j e^{\lambda_j t}\underline{u}_j = U \ \mathrm{diag}(e^{\lambda_1 t},\ldots,e^{\lambda_n t})\underline{c}$$

als $\underline{c} = (\gamma_1,\ldots,\gamma_n)^T$. Kennen we de beginwaarde $\underline{x}(0)$ dan volgt $\underline{c}$ uit

$$\underline{x}(0) = U\underline{c} \ .$$

Dus $\underline{c} = U^{-1}\underline{x}(0)$ en

$$\underline{x}(t) = U \ \mathrm{diag}(e^{\lambda_1 t},\ldots,e^{\lambda_n t})U^{-1}\underline{x}(0) = Ue^{t\Lambda}U^{-1}\underline{x}(0)$$

als we definiëren

$$e^{t\Lambda} := \mathrm{diag}(e^{\lambda_1 t},\ldots,e^{\lambda_n t}) \ .$$

Analoog geldt voor de oplossing van het inhomogene lineaire stelsel

$$\frac{d\underline{x}}{dt} = A\underline{x} + \underline{f}(t) \tag{7}$$

dat

$$\underline{x}(t) = Ue^{t\Lambda}U^{-1}\underline{x}(0) + U\int_0^t e^{(t-\tau)\Lambda}U^{-1}\underline{f}(\tau)d\tau \ . \tag{8}$$

Uit deze formules blijkt hoe kennis van eigenwaarden en eigenvectoren leidt tot oplossing van het stelsel. We kunnen dit nog iets anders toelichten, nl. met behulp van de spectraalvoorstelling van A. Substitutie van (2) in (7) levert

$$\frac{d\underline{x}}{dt} = U\Lambda U^{-1}\underline{x} + \underline{f}(t) .$$

Stel $\underline{x}(t) = U\underline{y}(t)$, $\underline{f}(t) = U\underline{g}(t)$. Dan geldt

$$\frac{d\underline{y}}{dt} = \Lambda\underline{y} + \underline{g}(t)$$

of wel

$$\frac{dy_j}{dt} = \lambda_j y_j + g_j(t), \quad j = 1,\ldots,n .$$

Doordat Λ een diagonaalmatrix is zijn nu de n differentiaalvergelijkingen ontkoppeld. Uit de oplossingen

$$y_j(t) = e^{\lambda_j t} y_j(0) + \int_0^t e^{\lambda_j (t-\tau)} g_j(\tau)d\tau$$

volgt eenvoudig de oplossing (8).

Behalve de expliciete vorm van de oplossingen kunnen uit eigenschappen van de eigenwaarden van A ook eigenschappen van de oplossingen gehaald worden. Bijvoorbeeld: als voor iedere eigenwaarde geldt

$$\mathrm{Re}(\lambda_j) < 0$$

dan gaan alle oplossingen van (6) naar 0 als $t \to \infty$; als bovendien $\|f(t)\| \leq M$ voor $0 \leq t < \infty$ dan zijn alle oplossingen van (7) begrensd in $0 \leq t < \infty$ (met een bovengrens die van M en $\|\underline{x}(0)\|$ afhangt). Omgekeerd, als er een eigenwaarde λ_j is met $\mathrm{Re}(\lambda_j) > 0$ dan heeft (6) onbegrensde oplossingen.
Als A niet n onafhankelijke eigenvectoren heeft, dan is de situatie ingewikkelder, we gaan daar niet op in.

b) Beschouw de differentievergelijking

$$\underline{x}_{k+1} = A\underline{x}_k + \underline{f}_k .$$

Eenvoudig blijkt dat

$$\underline{x}_{k+1} = A^{k+1}\underline{x}_0 + \sum_{j=0}^{k} A^{k-j}\underline{f}_j \ .$$

Invullen van de spectraalvoorstelling levert

$$\underline{x}_{k+1} = U(\Lambda^{k+1}U^{-1}\underline{x}_0 + \sum_{j=0}^{k} \Lambda^{k-j}U^{-1}\underline{f}_j)$$

en hieruit kunnen we weer uitspraken halen over het gedrag van $\underline{x}_k$ als $k \to \infty$ (doe dit zelf). Ook in dit geval werkt natuurlijk de transformatie $\underline{x}_k = U\underline{y}_k$, $\underline{f}_k = U\underline{g}_k$ ontkoppelend.

c) Voor veel lineaire trillende systemen zonder demping geldt een stelsel differentiaalvergelijkingen van de tweede orde:

$$M\frac{d^2\underline{x}}{dt^2} + K\underline{x} = 0 \ ,$$

waarin M en K positief definiete symmetrische matrices zijn (in het mechanische geval de zg. massa- en stijfheidsmatrices). We zoeken nu oplossingen van de vorm

$$\underline{x}(t) = e^{i\omega t}\underline{u} \ .$$

Deze voldoen als

$$K\underline{u} = \omega^2 M\underline{u} \ , \qquad (9)$$

d.w.z. als ω^2 eigenvector is van $M^{-1}K$ en $\underline{u}$ een bijbehorende eigenvector. Dit zgn. gegeneraliseerde eigenwaardeprobleem kan op fraaie wijze teruggebracht worden tot het standaard eigenwaardeprobleem met symmetrische matrix met behulp van de splitsing van Cholesky. Zij nl.

$$M = LL^T$$

met L linksonder (dit kan omdat M symmetrisch en positief definiet is). Stel nu $L^T\underline{u} = \underline{v}$, dan gaat (9) over in

$$A\underline{v} = \omega^2\underline{v}$$

met $A = L^{-1}KL^{-T}$, dus symmetrisch.

Heeft A de spectraalvoorstelling

$$A = V\Lambda V^T, \quad V^TV = I$$

en is $U := L^{-T}V$ dan geldt (ga na)

$$KU = MU\Lambda \quad (\text{of } K\underline{u}_j = \lambda_j M\underline{u}_j)$$

en

$$U^TMU = I \quad (\text{of } \underline{u}_i^T M\underline{u}_j = \delta_{ij}) \ .$$

Als K positief definiet is dan is A dat ook, de λ_j zijn dan positief. De getallen $\omega_j := \lambda_j^{\frac{1}{2}}$ zijn nu de zgn. eigenfrequenties van het systeem; de (onafhankelijke) oplossingen

$$\underline{x}_j = \underline{u}_j e^{i\omega_j t}$$

zijn de zgn. normaal-trillingen.

d) Bij trillende systemen met demping treedt als stelsel differentiaalvergelijkingen op

$$M\frac{d^2\underline{x}}{dt^2} + F\frac{d\underline{x}}{dt} + K\underline{x} = \underline{0} \ .$$

We kunnen dit stelsel herleiden tot een eerste orde stelsel door te stellen

$$\underline{y}_1 = \underline{x}, \ \underline{y}_2 = \frac{d\underline{x}}{dt} \ .$$

Dan geldt (ga na)

$$\frac{d}{dt}\begin{pmatrix}\underline{y}_1\\ \underline{y}_2\end{pmatrix} = \begin{pmatrix}0 & I\\ -M^{-1}K & -M^{-1}F\end{pmatrix}\begin{pmatrix}\underline{y}_1\\ \underline{y}_2\end{pmatrix} \ .$$

9.2. <u>De conditie van het eigenwaardeprobleem</u> ([18], ch. 2)

Een eenvoudig middel om iets over de ligging van de eigenwaarden van een matrix te weten te komen zijn de zg. cirkels van Gershgorin.

<u>Stelling</u>. Een eigenwaarde λ van A ligt in minstens één van de cirkelschijven (in het complexe vlak)

$$|z - A_{ii}| \leq \sum_{j\neq i} |A_{ij}|, \quad i = 1,\ldots,n \ .$$

Bewijs. Als voor alle i $|z - A_{ii}| > \sum_{j \neq i} |A_{ij}|$ dan is de matrix zI - A diagonaal-dominant en dus niet singulier (zie 5.3.1).

Men kan ook bewijzen: als de vereniging van p van de schijven disjunct is met de overige schijven dan bevat deze vereniging precies p eigenwaarden (waaronder eventueel meervoudige).

Voorbeeld.

$$A = \begin{pmatrix} 1 & 0.2 & 0.3 \\ 0.2 & 5 & 0.4 \\ 0.3 & 0.4 & 6 \end{pmatrix} .$$

De schijven zijn $|z - 1| \leq 0.5$, $|z - 5| \leq 0.6$, $|z - 6| \leq 0.7$. Omdat A symmetrisch is zijn de eigenwaarden reëel, er is dus één eigenwaarde in [0.5,1.5] en er zijn twee eigenwaarden in [4.4,6.7].
Door een truc kan men het eerste interval nog verkleinen. Vervang A door $A' := DAD^{-1}$ met D = diag(δ,1,1). Dan is A' gelijkvormig met A. Daar

$$A' = \begin{pmatrix} 1 & 0.2\delta & 0.3\delta \\ 0.2\delta^{-1} & 5 & 0.4 \\ 0.3\delta^{-1} & 0.4 & 6 \end{pmatrix}$$

heeft de eerste Gershgorin cirkel van A' als straal 0.5δ, en is gescheiden van de twee andere als

$$1 + 0.5\delta < 5 - 0.2\delta^{-1} - 0.4 ,$$
$$1 + 0.5\delta < 6 - 0.3\delta^{-1} - 0.4 .$$

Hieraan is nog net voldaan als $\delta = 0.056$. Voor de kleinste eigenwaarde λ_1 geldt nu dus $|\lambda_1 - 1| \leq 0.028$.

Uit de stelling van Gershgorin kunnen we een resultaat halen over de eigenwaarden van gestoorde matrices. We merken daartoe op dat als $A = \Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$, λ_k een enkelvoudige eigenwaarde van A is en

$$\|E\|_\infty < \tfrac{1}{2} \min_{i \neq k} |\lambda_i - \lambda_k| ,$$

A+E precies één eigenwaarde μ heeft die voldoet aan

$$|\mu - \lambda_k| \leq \|E\|_\infty .$$

Immers, de Gershgorin schijven van A + E zijn

$$|z - \lambda_i - E_{ii}| \leq \sum_{j \neq i} |E_{ij}|$$

en deze liggen binnen de schijven

$$|z - \lambda_i| \leq \|E\|_\infty .$$

En daar voor $i \neq k$ $|\lambda_i - \lambda_k| > 2\|E\|$ is de k-de van deze schijven disjunct met de rest, bevat dus precies één eigenwaarde.

Als nu A de spectraalvoorstelling $A = U\Lambda U^{-1}$ heeft en E een storing is, dan is $A + E = U(\Lambda + F)U^{-1}$ met $F = U^{-1}EU$. Uit het bovenstaande resultaat volgt nu:

Stelling. Zij A diagonaliseerbaar met spectraalvoorstelling

$$A = U\Lambda U^{-1} .$$

Als λ_k enkelvoudige eigenwaarde van A is en

$$c(U)\|E\| < \tfrac{1}{2} \min_{i \neq k} |\lambda_i - \lambda_k| ,$$

dan heeft A + E precies één eigenwaarde μ die voldoet aan

$$|\mu - \lambda_k| \leq c(U)\|E\|_\infty .$$

Opmerkingen.

1) Nadere beschouwing levert dat zelfs

$$\mu = \lambda_k + \frac{\underline{v}_k^T E \underline{u}_k}{\underline{v}_k^T \underline{u}_k} + O(\|E\|^2) ,$$

waarin $\underline{u}_k$ eigenvector van A en $\underline{v}_k$ eigenvector van A^T bij de eigenwaarde λ_k is (voor $\underline{v}_k^T$ kunnen we de k-de rij van U^{-1} nemen, er geldt dan $\underline{v}_k^T \underline{u}_k = 1$). Het getal

$$\frac{\|\underline{v}_k^T\| \, \|\underline{u}_k\|}{|\underline{v}_k^T \underline{u}_k|}$$

is dus een conditiegetal van λ_k (passend bij absolute fouten in A en λ_k). Dit conditiegetal kan in de buurt van c(U) liggen.

2) Als A symmetrisch is dan mogen we U orthogonaal veronderstellen en dan is (in de 2-norm) $c(U) = 1$, zodat de conditie van het eigenwaardeprobleem voor een symmetrische matrix altijd goed is.

3) In het geval dat A en E symmetrisch zijn kan men langs andere wegen veel meer bewijzen. Stel dat de eigenwaarden van A en A+E geordend zijn, resp. $\lambda_1 \geq \ldots \geq \lambda_n$ en $\mu_1 \geq \ldots \geq \mu_n$. Dan geldt

$$|\mu_k - \lambda_k| \leq \|E\|_2 \ , \ k = 1,\ldots,n \quad \text{(Stelling van Weyl)} \ ,$$

$$\sum_{k=1}^{n} (\mu_k - \lambda_k)^2 \leq \|E\|_E^2 := \sum_{i,j} E_{ij}^2 \quad \text{(stelling van Wielandt-Hoffman)}.$$

Hier blijkt weer de perfecte conditie van het eigenwaardeprobleem voor symmetrische matrices.

Hoe is het met de eigenvectoren? Zelfs bij symmetrische matrices is hier de zaak minder rooskleurig. Beschouw de matrix

$$A = \begin{pmatrix} 1 + \varepsilon \cos\varphi & \varepsilon \sin\varphi \\ \varepsilon \sin\varphi & 1 - \varepsilon \cos\varphi \end{pmatrix}.$$

Men rekent eenvoudig na dat de eigenwaarden en eigenvectoren zijn

$$\lambda_1 = 1 + \varepsilon, \ \underline{x}_1 = \begin{pmatrix} \cos \tfrac{1}{2}\varphi \\ \sin \tfrac{1}{2}\varphi \end{pmatrix}, \ \lambda_2 = 1 - \varepsilon, \ \underline{x}_2 = \begin{pmatrix} -\sin \tfrac{1}{2}\varphi \\ \cos \tfrac{1}{2}\varphi \end{pmatrix}.$$

Voor kleine ε zijn, ongeacht de waarde van φ, beide eigenwaarden dicht bij 1, terwijl de eigenvectoren, ongeacht de waarde van ε, de hele $\mathbf{R}^2$ doorzwaaien als φ van 0 naar 2π gaat. De oorzaak is natuurlijk het feit dat voor $\varepsilon = 0$ A twee gelijke eigenwaarden heeft met een twee-dimensionale ruimte van eigenvectoren. Toevoeging van de storing maakt de eigenwaarden onderling verschillend en (dus!) de richting van de eigenvectoren eenduidig bepaald, maar deze richtingen hangen geheel van de storing af. Beschouw nu de matrix

$$A = \begin{pmatrix} 1 + \varepsilon \cos\varphi & \varepsilon \sin\varphi & \alpha\varepsilon \\ \varepsilon \sin\varphi & 1 - \varepsilon \cos\varphi & \beta\varepsilon \\ \alpha\varepsilon & \beta\varepsilon & 2 + \gamma\varepsilon \end{pmatrix}$$

die voor $\varepsilon = 0$ een tweevoudige eigenwaarde $\lambda = 1$ heeft, met als eigenvectoren de hele ruimte opgespannen door $(1,0,0)^T$ en $(0,1,0)^T$, en een enkelvoudige eigenwaarde 2 met eigenvector $(0,0,1)^T$ heeft. Voor $\varepsilon \neq 0$ zijn de eigenwaarden en eigenvectoren

$$\lambda_1 = 1+\varepsilon+0(\varepsilon^2),\ \lambda_2 = 1-\varepsilon+0(\varepsilon^2),\ \lambda_3 = 2+\gamma\varepsilon+0(\varepsilon^2)\ ,$$

$$\underline{x}_1 = \begin{pmatrix}\cos\ \frac{1}{2}\varphi\\ \sin\ \frac{1}{2}\varphi\\ 0\end{pmatrix} + 0(\varepsilon),\ \underline{x}_2 = \begin{pmatrix}-\sin\ \frac{1}{2}\varphi\\ \cos\ \frac{1}{2}\varphi\\ 0\end{pmatrix} + 0(\varepsilon),\ \underline{x}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix} + 0(\varepsilon)\ .$$

Met name geldt dus: de door $\underline{x}_1$ en $\underline{x}_2$ opgespannen deelruimte verandert slechts met orde ε. Deze observatie geldt vrij algemeen. Als $(\lambda_1,\ldots,\lambda_p)$ een cluster van eigenwaarden is (d.w.z. $|\lambda_i - \lambda_j| << \|A\|$ voor i en $j \leq p$) dan zijn de individuele eigenvectoren $\underline{x}_1,\ldots,\underline{x}_p$ zeer gevoelig voor storingen in de matrix, de door $\underline{x}_1,\ldots,\underline{x}_p$ bepaalde deelruimte echter niet.

Tenslotte, als A een meervoudige eigenwaarde λ heeft met daarbij een deficiente eigenruimte, dan is de gevoeligheid van λ voor storingen groot. Voorbeeld: de eigenwaarden van

$$\begin{pmatrix}\alpha+\delta & 1\\ \varepsilon & \alpha-\delta\end{pmatrix}$$

zijn $\lambda = \alpha \pm \sqrt{\delta^2+\varepsilon}$. Er geldt

$$\text{als } \delta^2 << \varepsilon \text{ dan } \lambda = \alpha \pm \sqrt{\varepsilon} + O(\delta^2/\sqrt{\varepsilon})$$
$$\text{als } \delta^2 = \varepsilon \text{ dan } \lambda = \alpha \pm \sqrt{2\varepsilon} = \alpha \pm \delta\sqrt{2}$$
$$\text{als } \delta^2 >> \varepsilon \text{ dan } \lambda = \alpha \pm \delta + O(\varepsilon/\delta)\ .$$

Bij exact of vrijwel meervoudige eigenwaarden is de storing dus $O(\varepsilon^{\frac{1}{2}})$.

9.3. Methoden voor de bepaling van eigenwaarden en eigenvectoren

In de jaren 1955-1970 zijn door Rutishauser, Wilkinson en anderen een aantal methoden ontwikkeld die een zodanige perfectie hebben dat men het eigenwaardeprobleem voor niet extreem grote (en ijle) matrices als definitief opgelost mag beschouwen (uiteraard tot op de door de conditie van het probleem bepaalde nauwkeurigheid). De beste van deze methoden zijn opgenomen in het Handbook for automatic computation*) en beschikbaar in de programma bibliotheken van grote computerinstallaties. We bespreken daarom slechts enkele grondslagen van deze methoden.

*) Wilkinson, J.H. and C. Reinsch, Handbook for Automatic Computation, Vol. II, Linear Algebra, Springer-Verlag, Berlin etc., 1971.

9.3.1. De machtmethode en varianten ([2], 5.8.1, [18], ch. 4,5,6)

Als A diagonaliseerbaar is met spectraalvoorstelling $A = U\Lambda U^{-1}$ dan is voor een willekeurige vector $\underline{x} = \Sigma\gamma_j\underline{u}_j$

$$A^k\underline{x} = U\Lambda^k U^{-1}\underline{x} = \Sigma\gamma_j\lambda_j^k\underline{u}_j \ .$$

Als

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$$

en $\gamma_1 \neq 0$ dan volgt hieruit

$$A^k\underline{x} = \gamma_1\lambda_1^k[\underline{u}_1 + \sum_2^n (\frac{\gamma_j}{\gamma_1})(\frac{\lambda_j}{\lambda_1})^k\underline{u}_j] = \gamma_1\lambda_1^k[\underline{u}_1 + O(|\lambda_2/\lambda_1|^k)] \ .$$

De vector $A^k\underline{x}$ is op den duur dus vrijwel evenwijdig met $\underline{u}_1$ en de verhouding tussen de componenten van $A^{k+1}\underline{x}$ en die van $A^k\underline{x}$ nadert tot λ_1. We hebben hiermee dus een methode om de in absolute waarde grootste eigenwaarde met bijbehorende eigenvector te bepalen, de zg. machtmethode.
Natuurlijk wordt de methode uitgevoerd door, uitgaande van $\underline{x}_0 := \underline{x}$, de vectoren $\underline{x}_1, \underline{x}_2, \ldots$ te bepalen uit $\underline{x}_{k+1} = A\underline{x}_k$, eventueel met tussentijdse schaling om over- of underflow te voorkomen.
Gunstig voor de convergentie is dat γ_1 niet zeer klein is t.o.v. de overige γ's en dat $|\lambda_1|$ flink wat groter is dan $|\lambda_2|$. Dit laatste kan men soms enigszins verbeteren door niet met A maar met $A - \alpha I$ te werken (bijvoorbeeld: als A symmetrisch en positief definiet is en $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n > 0$, dan is de gunstigste waarde $\alpha = (\lambda_2 + \lambda_n)/2)$.

Een met de machtmethode zeer verwante methode is de zgn. inverse iteratie. Zij μ een goede benadering voor een eigenwaarde λ_k, zo goed dat $|\lambda_k - \mu| << |\lambda_i - \mu|$ voor alle $i \neq k$. De matrix $(A - \mu I)^{-1}$ heeft waarden $(\lambda_i - \mu)^{-1}$ en onder deze is $(\lambda_k - \mu)^{-1}$ in absolute waarde verreweg de grootste. De machtmethode, toegepast op $(A - \mu I)^{-1}$ convergeert dus zeer snel en produceert een eigenvector $\underline{y}$ van $(A - \mu I)^{-1}$ bij de eigenwaarde $(\lambda_k - \mu)^{-1}$. Maar dan is $\underline{y}$ ook eigenvector van A bij de eigenwaarde λ_k (ga na).
Inverse iteratie wordt veel gebruikt om eigenvectoren te berekenen bij eigenwaarden die berekend zijn met een methode die niet eenvoudig de eigenvectoren meebepaalt. De methode wordt uitgevoerd door uitgaande van $\underline{x}_0 := \underline{x}$ de vectoren $\underline{x}_1, \underline{x}_2, \ldots$ te bepalen door oplossing van $(A - \mu I)\underline{x}_{k+1} = \underline{x}_k$ (merk op dat maar een keer een LU-decompositie nodig is, $A-\mu I$ heeft een groot conditiegetal maar de richting waarin $\underline{x}_{k+1}$ slecht bepaald is, is juist de richting van de eigenvector!).

Moderne varianten van de machtmethode zijn de zgn. simultane vector iteratie methoden. Zij X een $n \times p$-matrix ($1 \leq p << n$) met onafhankelijke kolommen. Als $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$, dan zijn de kolommen van A^kX op den duur alle vrijwel evenwijdig met $\underline{u}_1$. Maar als $|\lambda_1| \geq \ldots \geq |\lambda_p| > |\lambda_{p+1}| \geq \ldots \geq |\lambda_n|$, dan nadert de ruimte opgespannen door de kolommen van A^kX tot de ruimte opgespannen door $\underline{u}_1,\ldots,\underline{u}_p$. Dit betekent dat, als A symmetrisch is, de eigenwaarden van de $p \times p$-matrix $(X^TA^kX)^{-1}X^TA^{k-1}X$ dicht bij $\lambda_1,\ldots,\lambda_p$ liggen. En de convergentiefactor waarmee λ_j benaderd wordt is $|\lambda_{p+1}/\lambda_j|$. Natuurlijk moet men zorgen dat geen nauwkeurigheid verloren gaat door het vrijwel evenwijdig worden van de kolommen van A^kX. Een goede manier hiervoor is om A^kX te schrijven als $A^kX = Q_kR_k$ waarin Q_k orthonormale kolommen heeft en R_k rechtsboven is. Dan moet Q_{k+1} bepaald worden uit de splitsing $AQ_k = Q_{k+1}P_{k+1}$ met P_{k+1} rechtsboven (ga na). Er geldt nu dat de eigenwaarden van $Q_k^TAQ_k$ naderen tot $\lambda_1,\ldots,\lambda_p$. Met deze methode kan men door af en toe een $p \times p$ eigenwaardeprobleem op te lossen, de p grootste eigenwaarden van A vinden. We merken nog op dat bij deze methode, evenals bij de machtmethode, de matrix A onveranderd blijft; dit is van belang als A een zeer grote maar ijle matrix is.

9.3.2. Voorbereidende transformaties ([18], ch. 3)

Voor diverse iteratieve methoden is het van groot belang dat de matrix A een speciale vorm heeft, b.v. Hessenberg-vorm ($A_{ij} = 0$ voor $i > j + 1$ of wel, in de terminologie van 5.2.6, linker bandbreedte 1) of tridiagonaalvorm.
Het blijkt dat een willekeurige matrix A met behulp van een orthogonale gelijkvormigheidstransformatie op Hessenberg-vorm gebracht kan worden:

$$Q^TAQ = H \ .$$

H heeft dan dezelfde eigenwaarden als A en als $\underline{y}$ eigenvector van H is, dan is $Q\underline{y}$ eigenvector van A.
We construeren Q als product van $n-1$ Householder transformaties (zie 6.3)

$$Q = P_1P_2 \ \ldots \ P_{n-1}, \quad Q^T = P_{n-1} \ \ldots \ P_1 \ .$$

Uit 6.3 volgt dat we een Householder transformatie P_1 kunnen bepalen zo dat

$$P_1A = \begin{pmatrix} A_{11} & x & \dots & x \\ \beta_1 & x & \dots & x \\ 0 & & & \\ & & & \\ 0 & x & & x \end{pmatrix}$$

P_1 is daarbij zo dat voor iedere $\underline{v}$ de eerste component van $P_1\underline{v}$ gelijk is aan die van $\underline{v}$. Hieruit volgt dat P_1AP_1 er uit ziet als

$$P_1AP_1 = \begin{pmatrix} A_{11} & x & \dots & x \\ \beta_1 & x & \dots & x \\ 0 & & & \\ & & & \\ 0 & x & \dots & x \end{pmatrix}$$

Ga nu door met een P_2 die van iedere $\underline{v}$ de eerste en de tweede component ongemoeid laat. Etc.

Natuurlijk geldt: als A symmetrisch is, dan is $H = H^T$, H is dan dus tridiagonaal. Daar de P's orthogonaal zijn is de invloed van afrondfouten gering. Voor de numeriek verkregen H geldt: er is een exact orthogonale $\tilde{Q}$ zo dat

$$H = \tilde{Q}^T(A + E)\tilde{Q}$$

met

$$\|E\| \leq \eta\|A\| .$$

9.3.3. <u>De QR-methode</u> ([2], 5.8.4; [18], ch. 10 en 13)

Deze methode is (als opvolger van de in 1958 door Rutishauser gepubliceerde LR-methode die zowel veelbelovende als onbevredigende aspecten had) in 1961 onafhankelijk door Francis en Kublanovskaya ontdekt en door verschillende auteurs (o.a. Wilkinson en Parlett) theoretisch verder doorgrond en vervolmaakt. Moderne varianten ervan zijn momenteel het beste algemene gereedschap voor volle matrices.

De grondvorm is aldus. Uit $A_0 := A$ construeren we achtereenvolgens $A_1, A_2, \dots$ met de volgende algoritme: splits $A_{k-1} = Q_kR_k$ met Q_k orthogonaal en R_k rechtsboven; neem $A_k := R_kQ_k$.

We merken op:

a) De splitsing kan uitgevoerd worden met behulp van Householder transformaties (zie § 6.3)

$$P^{(k)}_{n-1} \ldots P_1 A^{(k)}_{k-1} = R_k \ ,$$
$$A_k := R_k P^{(k)}_1 \ldots P^{(k)}_{n-1} \ .$$

b) Er geldt

$$A_k = Q_k^T A_{k-1} Q_k = Q_k^T \ldots Q_1^T A Q_1 \ldots Q_k \ ,$$

zodat A_k orthogonaal gelijkvormig is met A en dezelfde eigenwaarden heeft; dank zij de orthogonaliteit is de invloed van afrondfouten beperkt.

c) Er geldt

$$A^k = (Q_1 R_1)^k = Q_1 (R_1 Q_1)^{k-1} R_1 =$$

$$= Q_1 A_1^{k-1} R_1 = \ldots = (Q_1 \ldots Q_k)(R_k \ldots R_1) \ .$$

Daar $R_k \ldots R_1$ rechtsboven is geldt: als X de nxp-matrix is bestaande uit de eerste p kolommen van de eenheidsmatrix, dan vormen de eerste p kolommen van $Q_1 \ldots Q_k$ een orthonormale basis voor de ruimte opgespannen door de kolommen van $A^k X$; als $|\lambda_1| \geq \ldots \geq |\lambda_p| > |\lambda_{p+1}| \geq \ldots \geq |\lambda_n|$, dan nadert deze ruimte tot de ruimte opgespannen door $\underline{u}_1, \ldots, \underline{u}_p$.

d) Uit c) volgt met name dat $Q_1 \ldots Q_k \underline{e}_1$ een veelvoud is van $A^k \underline{e}_1$ en dus op de duur in de richting van $\underline{u}_1$ ligt; daaruit volgt dat $\lim A_k \underline{e}_1 = \lambda_1 \underline{e}_1$ met convergentiefactor $|\lambda_2/\lambda_1|$.

e) Analoog aan d) blijkt dat $\lim \underline{e}_n^T A_k = \lambda_n \underline{e}_n^T$ met convergentiefactor $|\lambda_n/\lambda_{n-1}|$.

f) Wat diepergaande beschouwingen leren dat als $|\lambda_1| > \ldots > |\lambda_n|$, A_k in het algemeen nadert tot bovendriehoeksvorm met als diagonaal $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$; als A symmetrisch is dan zijn ook alle A_k symmetrisch met als limiet $\mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

g) Uit e) volgt dat het voordelig is om zgn. shifts toe te passen: als A_{k-1} bekend is dan kiezen we een shift α_k en bepalen A_k uit

$$A_{k-1} - \alpha_k I = Q_k R_k \ ,$$

$$A_k := R_k Q_k + \alpha_k I \ ;$$

dan is wederom $A_k = Q_k^T A_{k-1} Q_k$ maar de convergentiefactor uit e) is nu $|(\lambda_n - \alpha)/(\lambda_{n-1} - \alpha)|$ (mits α dichter bij λ_n dan bij de overige eigenwaarden ligt). Er bestaan diverse strategieën om goede α's te kiezen, b.v. $\alpha_k := (A_{k-1})_{nn}$. De convergentie van de laatste rij wordt dan in het algemeen kwadratisch.

h) Zodra de laatste rij van A_k voldoende geconvergeerd is naar een veelvoud van $\underline{e}_n^T$ laten we hem en de laatste kolom van A_k weg en gaan we met de overblijvende $(n-1) \times (n-1)$-matrix verder.

Dank zij het feit dat het QR-proces met shifts een mengsel is van simultane vector-iteratie en inverse iteratie blijkt dat met een goede shift strategie als regel slechts 2n a 3n slagen nodig zijn om alle eigenwaarden te bepalen. De numerieke nauwkeurigheid is goed in de zin dat de verkregen λ's exacte eigenwaarden zijn van een matrix die niet meer dan $\eta\|A\|$ van A verwijderd is. De voor het proces benodigde hoeveelheid werk wordt zeer verminderd als we door een transformatie vooraf zorgen dat A_0 de Hessenberg vorm heeft. Uit $Q_1 = A_0 R_1^{-1}$ volgt dan met 5.2.6.3 dat Q_1 ook de Hessenberg vorm heeft en hetzelfde geldt voor $A_1 = R_1 Q_1$. Uitvoering van de splitsing $A = QR$ kost voor een Hessenberg matrix echter slechts $O(n^2)$ bewerkingen tegen $O(n^3)$ voor een volle matrix en hetzelfde geldt voor de bepaling van RQ. Als A_0 tridiagonaal is en symmetrisch, dan is A_1 Hessenberg en symmetrisch, dus weer tridiagonaal; een slag kost dan $O(n)$ bewerkingen.

Hoe bepalen we de eigenvectoren? Als na de laatste slag geldt

$$A_k = \Lambda_k + R_k + E_k$$

met Λ_k diagonaal, R_k strict rechtsboven, E_k strict linksonder en klein, dan bepalen we de eigenvectoren van $\Lambda_k + R_k$. Stel dat de diagonaalelementen van Λ_k onderling verschillend zijn. Dan is er een reguliere bovendriehoeksmatrix V waarvan de kolommen de eigenvectoren van $\Lambda_k + R_k$ zijn:

$$(\Lambda_k + R_k)V = V\Lambda_k \quad .$$

Er geldt dan

$$Q_k^T \ldots Q_1^T A Q_1 \ldots Q_k = V\Lambda_k V^{-1} + E_k \ ,$$

en dus, als $Q_1 \ldots Q_k = Q$, $QV = U$,

$$A - QE_kQ^T = U\Lambda U^{-1} \ ,$$

zodat we de spectraalvoorstelling gevonden hebben van een matrix die dicht bij A ligt. De eigenvectoren $\underline{u}_1, \ldots, \underline{u}_n$ van deze matrix volgen eenvoudig uit V als we het product $Q_1, \ldots, Q_k$ van de opvolgende transformaties bijgehouden hebben.

In het geval van een symmetrische A is $R_k = 0$ en dan kunnen we $V = I$ nemen.