Numerieke Wiskunde 1
2003-2004

M. de Jeu
Mathematisch Instituut
Universiteit Leiden

21 januari 2004

Voorwoord

Dit dictaat bevat de in het collegejaar 2002-2003 bij het college "Numerieke Wiskunde 1" behandelde stof. In het collegejaar 2003-2004 zullen we hier een deel van behandelen.

In de tekst is een aantal "Terzijdes" opgenomen. Deze behoren niet tot de tentamenstof, maar dienen om de hoofdtekst verder aan te vullen of in pespectief te plaatsen.

Ik ben Jaap van de Griend zeer erkentelijk voor zijn vele constructieve opmerkingen, en Gerard Westhoff voor zijn correctie van een aantal typfouten en vergissingen.

Marcel de Jeu.

Inhoudsopgave

Voorwoord **iii**

1 Inleiding **1**
1.1 Plaats en rol van de numerieke wiskunde 1
1.2 Criteria voor numerieke methodes 3

2 Rode draad en vooruitblik **9**
2.1 Rode draad 9
2.2 Vooruitblik 9

3 Polynomiale approximatie **13**
3.1 Motivatie 13
3.2 Interpolatieformule van Lagrange 14
3.3 Interpolatiemethode van Newton 16
3.4 Foutschatting bij interpolatie 17
3.5 Stuksgewijze polynomiale approximatie 20
3.6 Adaptieve methoden 23

4 Numerieke integratie en extrapolatie **27**
4.1 Inleiding en overzicht 27
4.2 Kwadratuurregels: algemeen 28
4.3 Enkelvoudige kwadratuurregels 30
4.4 Samengestelde kwadratuurregels 35
4.5 Asymptotiek voor de samengestelde trapeziumregel 38
4.6 Extrapolatie naar $h = 0$ en Romberg-integratie 43
4.7 Meerdimensionale integralen 49

5 Lineaire stelsels **51**
5.1 LU-ontbinding: algemene geval 52
5.2 Pivoting 58
5.3 LU-ontbinding: speciale gevallen 59
5.4 Iteratieve methoden voor lineaire stelsels 64

6 Eindige elementen methode **75**
6.1 Het randwaardeprobleem 75
6.2 Hilbertruimte-benadering: zwakke formulering 76
6.3 Eindige elementen methode 79

A Functionaalanalytisch kader **85**
A.1 Genormeerde lineaire ruimten 85
A.2 Begrensde lineaire afbeeldingen 88
A.3 Het eindigdimensionale geval 90
A.4 Stabiliteit en conditiegetal 94
A.5 Hilbertruimten 95

B Opgaven **103**
B.1 Nulpunten van reële functies 103
B.2 Fouten en hun doorwerking 108
B.3 Polynomiale approximatie 109
B.4 Numerieke integratie en extrapolatie 111
B.5 Lineaire stelsels 113
B.6 Eindige elementen methode 117

Hoofdstuk 1

Inleiding

1.1 Plaats en rol van de numerieke wiskunde

De plaats en de rol van de numerieke wiskunde kunnen we duiden door na te gaan hoe we (vaak) de wereld om ons kwantitatief proberen te beschrijven. Meestal kunnen we hier de volgende vier stappen bij onderscheiden:

1. We zijn geïnteresseerd in een bepaald deel van de werkelijkheid. Vb.: het weer, de economie, de temperatuur in een staaf, of de beweging van een massa die aan een veer hangt. Laten we dit laatste als concreet voorbeeld nemen.

2. We modelleren dit deel van de werkelijkheid in wiskundige termen. Dit is werk voor de betreffende vakspecialisten. Het model is in dit stadium nog algemeen, d.w.z. bevat nog bepaalde parameters.
   Voorbeeld: voor de massa aan de veer krijgen we $mu'' = -ku$.

3. We meten of schatten de waarden van de parameters in een specifieke situatie en krijgen een wiskundig model voor die specifieke situatie.
   Voorbeeld: na meting van de massa en bepaling van de veerconstante krijgen we $u'' = -1.7u$.

4. We maken gebruik van een *specifieke toepassing van een numerieke methode* om te komen tot een concrete getalsmatige uitspraak in ons specifieke geval.
   Voorbeeld: $u(2.2) = 3.08 \pm 0.1$. [1]

De numerieke wiskunde treedt hier dus in een specifiek geval op bij de overgang van (3) naar (4). Uiteraard is men in het algemeen wat ambitieuzer en wil men niet één specifiek model numeriek aankunnen, maar een hele klasse. Dat wil zeggen, de numerieke wiskunde houdt zich bezig met de vertaalslag van (2) naar (4), waarbij de parameters dan in de methode blijven staan en later in een concreet geval getalsmatig ingevuld kunnen worden.

Numerieke wiskunde is een breed gebied, en iedere definitie is dus bijna onvermijdelijkerwijs onvolkomen. Op grond van bovenstaande positionering wagen we toch een poging:
Numerieke wiskunde houdt zich bezig met het ontwerpen en analyseren van methoden om de oplossing van theoretisch opgeloste/oplosbare wiskundige problemen daadwerkelijk in getallen uit te drukken.

[1] Deze veer-vergelijking is exact op te lossen in termen van elementaire functies, maar dat doet voor de aard van het voorbeeld niet ter zake.

Theorie en numerieke methoden vullen elkaar op deze wijze dus aan. Methoden en technieken uit de numerieke wiskunde, en zeker de wat oudere, hebben veelal een oorsprong bij fysici en ingenieurs, wat gezien bovenstaande positie en rol ook wel voor de hand ligt.

Laten we het wederzijds aanvullen van "theorie" en "numerieke methoden" eens illustreren aan de hand van een paar voorbeelden.

1. De theorie geeft dat de reeks $\sum_{n=0}^{\infty} \frac{1}{n!}$ convergeert. Immers, de reeks is absoluut convergent en $\mathbb{R}$ is volledig. Getalsmatig hebben we daar niet veel aan. Beter is daarvoor de uitspraak: $e \in [2.71, 2.72]$.

2. Beschouw de differentiaalvergelijking (een zgn. beginwaardeprobleem):

$$\begin{cases} u'' & = -u \\ u(0) & = 0 \\ u'(0) & = 1. \end{cases}$$

   voor een functie $u : [0, \infty] \to \mathbb{R}$. De theorie leert ons dat er een unieke oplossing is. Voor de vraag wat $u(1.7)$ dan getalsmatig is, levert dat echter niet veel op. Een typische numerieke uitspraak, en resultaat van een numerieke methode om differentiaalvergelijkingen op te lossen, is dan bijv. dat $u(1.7) \in [0.991, 0.992]$.

3. De functie u in het vorige voorbeeld is de sinus, dus toepassing van een numerieke methode lijkt misschien niet nodig, hoewel het benaderen van $\sin 1.7$ met bijv. een machtreeks op zich natuurlijk al een numeriek probleem is. Het is echter gemakkelijk om voorbeelden op te schrijven die het karakter van de situatie scherper naar voren brengen. Basis hiervoor is de volgende theoretische stelling.

   Stelling 1.1.1. *Laat $p_{k-1}, \ldots, p_1$ en q continue functies op $[a, b]$ zijn. Dan heeft, voor alle $\alpha_0, \ldots, \alpha_{k-1}$, het beginwaardeprobleem*

$$\begin{cases} u^{(k)} + p_{k-1}u^{(k-1)} + \ldots p_1 u' + p_0 u = q \\ u(a) = \alpha_0 \\ \vdots \\ u^{(k-1)}(a) = \alpha_{k-1} \end{cases}$$

   precies 1 oplossing op $[a, b]$.

   Blijkbaar heeft ook het beginwaardeprobleem

$$\begin{cases} u'' + \frac{\sin\sqrt{x^2+1}}{e^{-\cos x}+8} u' + \frac{1}{x^4+3} u = 1 \\ u(0) = 0 \\ u'(0) = 1 \end{cases}$$

   een unieke oplossing op $[0, 1]$. Theoretisch is de zaak daarmee opgelost. Wanneer we nu echter bijvoorbeeld $u(0.6)$ willen weten, dan helpt de theorie niet. Met de bekende expliciete methoden voor speciale typen differentiaalvergelijkingen valt hier niets te beginnen. Toch zijn er numerieke methoden beschikbaar die ons $u(0.6)$ met willekeurig grote precisie opleveren—en dat is dan het beste wat bereikt kan worden.

4. Laat $f : [0, 1] \to \mathbb{R}$ continu zijn, met $f(0) = -1$ en $f(1) = 1$. Volgens de theorie heeft f tenminste één nulpunt op dit interval. Voor het concreet benaderen van zo'n nulpunt in een specifiek geval hebben we i.h.a. numerieke methoden nodig.

1.2 Criteria voor numerieke methodes

Vaak zijn er meerdere numerieke methodes denkbaar voor een bepaald probleem. Wanneer is de ene methode beter dan de andere? Belangijke criteria zijn de *efficiency* (d.w.z. rekentijd) en de *nauwkeurigheid* van een methode.

De *efficiency van een methode*: Soms bestaan er meerdere methoden die hetzelfde antwoord opleveren, maar waarbij de ene methode minder werk is dan de andere. Neem als voorbeeld de vergelijking $Ax = b$ met A een inverteerbare reële $n \times n$-matrix en een voorgegeven $b \in \mathbb{R}^n$. Deze vergelijking is uniek oplosbaar. Sterker nog: de theorie geeft ons via de regel van Cramer zelfs een expliciete uitdrukking voor de oplossing, in termen van determinanten, waarvoor dan ook weer gesloten uitdrukkingen bestaan als sommaties met alternerende tekens. Deze uitdrukking is van grote waarde, dat blijkt bijv. uit het wat verrassende inzicht dat de oplossing x gehele coëfficiënten heeft als A en b dat hebben en als bovendien $\det A = \pm 1$. Echter, al voor relatief kleine n, zeg $n = 100$, is het al niet meer mogelijk om de betrokken determinanten uit te rekenen, althans niet als som van de $n!$ termen in de definitie. Methoden als Gauß-eliminatie brengen dan uitkomst om het antwoord te bepalen. De berekening hiermee is veel minder werk: Gauß-eliminatie is, kortom, efficiënter dan de regel van Cramer.

Soms is echter zelfs Gauß-eliminatie nog teveel werk en past men benaderende methodes toe. Deze methodes leveren dan niet de exacte oplossing, maar in een aantal gevallen is een benadering al goed genoeg. Of dat daadwerkelijk het geval is, hangt af van de context van het probleem. Dat brengt ons op:

De *nauwkeurigheid van een methode*: een numerieke methode levert i.h.a. slechts een benadering van het—binnen het model—werkelijke antwoord. Is dat erg? Daar valt niet in zijn algemeenheid een antwoord op te geven. Uiteraard geldt: hoe nauwkeuriger, hoe beter, maar een nauwkeuriger antwoord kost i.h.a. ook meer rekenwerk. In iedere specifieke situatie zullen we ons dus af moeten vragen wat het ons waard is om de numerieke berekeningen nauwkeuriger uit te voeren. We moeten ons hierbij terdege realiseren dat er *meerdere* foutenbronnen zijn in de totale keten in Paragraaf 1.1. Deze grotere context maakt duidelijk hoe ernstig of niet ernstig de numerieke fout is, en dat zal van situatie tot situatie verschillen. We zullen ons daarom in het vervolg van deze paragraaf wijden aan het benoemen van de verschillende oorzaken van fouten en de mogelijke remedies.

Er zijn binnen de keten in Paragraaf 1.1 vier foutbronnen:

1. modelleerfout: modellen zijn bijna altijd een benadering.
2. meetfout: metingen/schattingen van parameters in een model zijn onderhevig aan fouten.
3. de numerieke methode zelf is i.h.a. per constructie een benaderende methode, en geen exacte berekening.
4. computers hebben een grote, maar toch eindige nauwkeurigheid. Afrondfouten zijn vrijwel altijd onvermijdelijk.

Wat kunnen we hieraan doen, om de precisie in de hele keten te verbeteren?

Ad 1: modelleerfout.
Verbeteren van de betreffende modellen kan alleen door deskundigen in het betreffende vak gebeuren.

Ad 2: meetfout.
Beter meten of schatten verhoogt uiteraard de nauwkeurigheid; dit vergt vakexpertise. Soms is

het wel nog mogelijk om aan te geven hoe onzekerheden in de parameters uiteindelijk via de numerieke methode doorwerken in het eindresultaat.

Ad 3: benaderend karakter van de methode.
Beter dan een benaderende methode is er vaak niet (denk aan het voorbeeld van de differentiaalvergelijking). Het beste wat we dan dus kunnen doen, is iets te zeggen over de blijkbaar onvermijdelijke fout van onze numerieke methode. Wat voor soort uitspraak zouden we hierover eigenlijk willen hebben?
Een voorbeeld: stel een numerieke methode voor een of ander probleem, zeg de nulpuntsbepaling van een ons gegeven functie, levert ons een rij $\{x_n\}_{n=1}^{\infty}$ op die de werkelijke oplosing $\hat{x}$ benadert, in de zin dat *we hebben kunnen aantonen* dat $|x_n - \hat{x}| \leq 2^{-n}$. We laten even in het midden hoe we dat dan zouden hebben kunnen aantonen, maar als dat eenmaal gebeurd is, dan zien we dat niet alleen $x_n \to \hat{x}$, maar beter nog: we kunnen $\hat{x}$ expliciet benaderen met iedere door ons gewenste nauwkeurigheid. Wanneer we bijvoorbeeld $\hat{x}$ op 10^{-6} nauwkeurig willen kennen, dan rekenen we x_{20} uit, en zijn dan klaar, daar immers met zekerheid geldt dat $|x_{20} - \hat{x}| \leq 2^{-20} < 10^{-6}$.
Een belangrijk onderdeel van de numerieke wiskunde is dan ook de *foutanalyse*, waarmee we de aan een methode inherente fout onder controle willen brengen. *De foutanalyse van een methode heeft tot doel een (bij sterke voorkeur) kwantitatieve uitspraak te doen over het verschil tussen de echte oplosing van een probleem en de benadering(en) die de methode levert.* De foutanalyse in bovenstaande voorbeeld heeft blijkbaar geleid tot de heel bruikbare uitspraak $|x_n - \hat{x}| \leq 2^{-n}$.
Het hangt overigens van de situatie af wat men als fout zal willen definiëren. Soms is dit het verschil tussen twee reële getallen, zoals boven, soms de integraal van de absolute waarde van het verschil van twee functies (de echte en de benaderende), soms de lengte van een verschilvector in $\mathbb{R}^n$—van alles is mogelijk.

Bij een foutanalyse van een methode wordt overigens i.h.a. aangenomen dat alle voor die methode noodzakelijke berekeningen *exact* worden uitgevoerd. We berekenen dus weliswaar slechts benaderingen, maar die benaderingen op zich berekenen we dan wel foutloos. Principieel is dit niet geheel juist, want in de praktijk wordt er vrijwel nooit exact gerekend, als gevolg van de intrinsieke onnauwkeurigheid van computers. De impliciete aanname is echter altijd dat de computerimplementatie zo verstandig is uitgevoerd, dat de praktische bruikbaarheid van onze theoretische schatting voor de fout niet door de eindige nauwkeurigheid van de berekeningen wordt aangetast. Een voorbeeld: in onze bovenstaande rij weten we door onze bewijsvoering zeker dat $|x_3 - \hat{x}| \leq \frac{1}{8}$. Stel nu, dat onze computerberekening $x_3 = 0.57136$ geeft, terwijl onze methode in werkelijkheid (maar dat weten we dus niet) als echte benadering $x_3 = 0.57137\ldots$ geeft. Wanneer we de uitspraak $\hat{x} \in [0.57136 - 0.125, 0.57136 + 0.125]$ doen, dan is dat strikt genomen dus niet juist—maar de *praktische* waarde van onze theoretische uitspraak zal toch niet vaak worden aangetast door zulke *relatief* zeer kleine onjuistheden, zeker ook gezien de twee andere praktische foutbronnen (modelleren en parameters) die we al noemden. Dit beeld streven we in zijn algemeenheid na: we willen er voor zorgen dat de consequenties van de computer-onnauwkeurigheid veel kleiner zijn dan de theoretische marge van de methode. In dat geval zullen we in de praktijk met vertrouwen aan onze theoretische foutschattingen willen vasthouden.

Dit brengt ons dus op het volgende punt, waar we wat langer bij stil zullen staan.

Ad 4: computer-nauwkeurigheid.
Voor het begrijpen van de computer-nauwkeurigheid moeten we weten hoe getallen in een computer worden gerepresenteerd. Een computer werkt binair, maar het begrip wordt niet gehinderd door de tientallige notatie die wij zullen gebruiken, en die voor ons uiteraard prettiger is.

Getallen $x \neq 0$ in een computer worden in *floating point* vorm gerepresenteerd, d.w.z. als

$$x = \pm 0.d_1 \dots d_k \cdot 10^n \quad (1 \leq d_1 \leq 9,\, 0 \leq d_i \leq 9\,(i = 2, \dots, k)).$$

Het deel $\pm 0.d_1 \dots d_k$ heet de *mantisse*, n is de *exponent*. De *precisie* k is het aantal decimalen waarmee getallen worden bewaard. Het getal 0 wordt op de voor de hand liggende manier weergegeven met exponent 0. De exponent kan gehele waarden aannemen in het interval $[-M, M]$ voor een of andere M. Typische waarden zijn $k = 16$ en $M = 1024$. Dit heeft zijn eigenaardigheden: een computer werkt met eindig veel getallen (allemaal rationaal), kent een grootste getal en kent ook een kleinste strikt positief getal.

Om de gedachten te bepalen, nemen we aan dat deze representatie geldt voor getallen in het werkgeheugen of op schijf, en dat de processor zelf met zoveel decimalen werkt dat de uitkomst van individuele processorberekeningen als exact beschouwd mag worden. Fouten worden dan geïntroduceerd doordat tussenresultaten moeten worden opgeslagen, of doordat niet alle decimalen van invoer gehandhaafd kunnen blijven. Een getal x wordt ter opslag of na invoer in zgn. *floating point representatie* $\mathrm{fl}(x)$ gebracht. Dat gaat als volgt (we nemen $x > 0$; voor $x < 0$ gaan we over op $-x$ en verdisconteren later het teken): Als $x = 0.d_1 \dots d_k \dots \cdot 10^n$, met $1 \leq d_1 \leq 9$, de decimale ontwikkeling van x is, definieer dan:

$$\mathrm{fl}(x) = \begin{cases} 0.d_1 \dots d_k \cdot 10^n & \text{als } d_{k+1} \leq 4, \\ 0.d_1 \dots d_k \cdot 10^n + 10^{-k+n} & \text{als } d_{k+1} \geq 5. \end{cases}$$

We ronden dus de $(k+1)$-de decimaal naar boven af en werken het gevolg helemaal door[2]. De aldus geïntroduceerde fout hangt van de grootte van x af, zoals men gemakkelijk ziet. De *relatieve* fout is echter onafhankelijk van de grootte van x, en hiervoor geven we een afschatting. We onderscheiden twee gevallen. Als $d_{k+1} \leq 4$, dan is

$$\begin{aligned} \left| \frac{x - \mathrm{fl}(x)}{x} \right| &= \left| \frac{0.0 \dots 0 d_{k+1} \dots \cdot 10^n}{0.d_1 \dots d_k d_{k+1} \dots \cdot 10^n} \right| \\ &\leq \frac{5 \cdot 10^{-k-1} \cdot 10^n}{10^{-1} \cdot 10^n} = 5 \cdot 10^{-k}. \end{aligned}$$

Wanneer $d_{k+1} \geq 5$ schatten we (de symbolen boven de mantissen geven de plaats van het cijfer aan):

$$\begin{aligned} \left| \frac{x - \mathrm{fl}(x)}{x} \right| &= \left| \frac{0.0 \dots 0 d_{k+1} \dots \cdot 10^n - 10^{-k+n}}{0.d_1 \dots d_k d_{k+1} \dots \cdot 10^n} \right| \\ &= \frac{(0.0 \dots \overset{k}{0}1 - 0.0 \dots \overset{k\ k+1}{0 d_{k+1}} \dots) \cdot 10^n}{0.d_1 \dots d_k d_{k+1} \dots \cdot 10^n} \\ &\leq \frac{(0.0 \dots \overset{k}{0}1 - 0.0 \dots \overset{k k+1}{0 0\ 5}) \cdot 10^n}{10^{-1} \cdot 10^n} = \frac{5 \cdot 10^{-k-1} \cdot 10^n}{10^{-1+n}} = 5 \cdot 10^{-k}. \end{aligned}$$

Conclusie: de relatieve fout is altijd ten hoogste $5 \cdot 10^{-k}$. Voor $k = 16$ betekent dit dus een zeer nauwkeurige representatie.

Zijn er dan, gegeven deze hoge mate van computer-nauwkeurigheid, toch nog problemen mogelijk die voortkomen uit het slechts eindig zijn van deze nauwkeurigheid? Het antwoord

[2] Een iets onnauwkeurigere methode zou bestaan uit het simpelweg verwijderen van de $(k+1)$-de decimaal en alle volgende.

hierop is bevestigend. Om dat te begrijpen formuleren we het problem iets abstracter, wat dan ook gelijk een kader biedt voor het bestuderen van de doorwerking van meetfouten (de tweede foutenbron).

Stel, we hebben een functie f, die we willen evalueren in een punt x. Onze functie f kunnen we exact uitreken voor alle waarden ("de processor werkt exact"), maar helaas hebben we niet de beschikking over x, maar slechts over een benadering $x + \Delta x$ van x. Kunnen we de fout $f(x+\Delta x) - f(x)$ dan relateren aan de fout Δx?[3] (Onze situatie valt binnen dit kader, wanneer de processor niet x als invoer voor de "exacte" berekening van $f(x)$ krijgt, maar slechts $x + \Delta x$ met $\Delta x = \mathrm{fl}(x) - x$.) Als f differentieerbaar is, dan is in eerste benadering:

$$\Delta f(x) \stackrel{\text{def.}}{=} f(x + \Delta x) - f(x) \simeq f'(x)\Delta x.$$

Voor de relatieve fouten geldt dus:

$$\frac{\Delta f(x)}{f(x)} \simeq \frac{xf'(x)}{f(x)} \cdot \frac{\Delta x}{x}.$$

W definiëren daarom de grootheid

$$\gamma \stackrel{\text{def.}}{=} \left| \frac{xf'(x)}{f(x)} \right|.$$

Soms schrijven we ook $\gamma(x)$. Deze grootheid[4] meet dus de factor waarmee de relatieve fout in x in eerste benadering doorwerkt in de relatieve fout in $f(x)$. We noemen γ of $\gamma(x)$ het *conditiegetal van* f (in het punt x). Als $\gamma \gg 1$, dan is de berekening slecht geconditioneerd. Hoe kleiner γ, hoe beter.

Voorbeeld 1.2.1. Stel, we willen $99 - 70\sqrt{2}$ berekenen, maar hebben slechts 1.4 als benadering voor $\sqrt{2}$ tot onze beschikking. We benaderen nu $99 - 70\sqrt{2}$ (de exacte waarde is $0.00505\ldots$) op twee manieren:

1. $99 - 70\sqrt{2} \simeq 99 - 70 \cdot 1.4 = 1$ (slecht).
2. $99 - 70\sqrt{2} \stackrel{\text{ga na!}}{=} \frac{1}{99+70\sqrt{2}}$, dus we nemen als benadering $\frac{1}{99+70\cdot 1.4} = \frac{1}{197}$ (goed).

Blijkbaar geven twee theoretisch dezelfde uitdrukkingen heel andere antwoorden, bij gebruik van toch dezelfde benadering. Dit kunnen we nu begrijpen met conditiegetallen. De functies

$$f_1(x) = 99 - 70x \quad \text{en} \quad f_2(x) = \frac{1}{99 + 70x}$$

hebben dezelfde waarde in $x = \sqrt{2}$, maar de conditiegetallen $\gamma_1(\sqrt{2})$ en $\gamma_2(\sqrt{2})$ zijn daar heel verschillend. Enig rekenwerk (ga na!) geeft

$$\frac{\gamma_1(\sqrt{2})}{\gamma_2(\sqrt{2})} = \frac{99 + 70\sqrt{2}}{99 - 70\sqrt{2}},$$

en dit quotiënt ligt in de orde van $200/(1/200) = 40.000$. M.a.w.: de relatieve fout in de benadering van $\sqrt{2}$ werkt voor f_1 ca. 40.000 keer zo sterk door als voor f_2.

[3] In onze computersituatie treedt nogmaals een fout op, wanneer de uitkomst van de "exacte" berekening van $f(x+\Delta x)$ door de processor als uitvoer gegeven moet worden, hetzij naar ons, hetzij naar het geheugen. Immers, dan wordt niet $f(x + \Delta x)$ uitgevoerd, maar slechts de benadering $\mathrm{fl}(f(x + \Delta x))$ daarvan. De relatieve fout in die laatste stap is echter, zoals we gezien hebben, typisch maximaal in de orde van $5 \cdot 10^{-16}$, d.w.z. zeer klein. We concentreren ons dus op $f(x + \Delta x)$ versus $f(x)$. De problemen daarmee kunnen veel ernstiger zijn, zoals we zullen zien.

[4] Conventies zonder de absolute waarden zijn ook in gebruik, en in sommige gevallen kan een analyse die de tekens in aanmerking neemt ook beter werken—wij beperken ons hier echter tot de absolute waarde.

We concluderen uit bovenstaand voorbeeld dat het loont om verschillende manieren van berekenen, die *in principe*, d.w.z. bij exacte data, hetzelfde zouden moeten opleveren, te vergelijken wat betreft hun conditiegetallen.

In meerdere variabelen is de manier van behandelen in principe hetzelfde: wanneer we $f(x_1, \ldots, x_n)$ willen uitrekenen, maar we kennen slechts benaderingen $x_1 + \Delta x_1, \ldots, x_n + \Delta x_n$ van de argumenten, dan is voor differentieerbare f in eerste benadering:

$$\Delta f(x_1, \ldots, x_n) \stackrel{\text{def.}}{=} f(x_1 + \Delta x_1, \ldots, x_n + \Delta x_n) - f(x_1, \ldots x_n) \simeq \sum_{i=1}^{n} \Delta x_i \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n),$$

en voor de relatieve fout hebben we

$$\frac{\Delta f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \simeq \sum_{i=1}^{n} \frac{x_i \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \cdot \frac{\Delta x_i}{x_i}.$$

We zien dat in eerste benadering de doorwerking van de relatieve fouten in de variabelen vastgelegd wordt door de conditiegetallen—hier zonder absolute waarden—m.b.t. ieder van de variabelen. We hebben de conservatieve afschatting

$$\left| \frac{\Delta f(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \right| \lesssim \sum_{i=1}^{n} \left| \frac{x_i \frac{\partial f}{\partial x_i}(x_1, \ldots, x_n)}{f(x_1, \ldots, x_n)} \right| \cdot \left| \frac{\Delta x_i}{x_i} \right|.$$

Een in computersituaties vaak voorkomend voorbeeld is het bepalen van het verschil van twee getallen x_1 en x_2, waarbij deze getallen eerder bepaald zijn in tussenberekeningen en zijn weggeschreven, maar helaas dus slechts als $\mathrm{fl}(x_1)$ en $\mathrm{fl}(x_2)$ i.p.v. als de exacte waarden. Hoe werken deze fouten door? In dit geval is $f(x_1, x_2) = x_1 - x_2$, en voor de relatieve fout hebben we (het kan hier exact en hoeft niet bij benadering)

$$\frac{\Delta f(x_1, x_2)}{x_1 - x_2} = \frac{\Delta x_1 - \Delta x_2}{x_1 - x_2} = \frac{x_1}{x_1 - x_2} \cdot \frac{\Delta x_1}{x_1} - \frac{x_2}{x_1 - x_2} \cdot \frac{\Delta x_2}{x_2}.$$

De conditiegetallen zijn dus $(1 - \frac{x_2}{x_1})^{-1}$ en $(1 - \frac{x_1}{x_2})^{-1}$. Schrijven we $x_2 = (1 + \epsilon)x_1$, dan is het eerste conditiegetal ϵ^{-1}. Voor ϵ zeer klein is deze berekening dus zeer slecht geconditioneerd. Blijkbaar geldt: *het bepalen van het verschil van twee vrijwel gelijke getallen is, wanneer er slechts benaderingen beschikbaar zijn, een onnauwkeurige operatie die bij voorkeur vermeden moet worden.*

Om een getallenvoorbeeld te geven: wanneer we in een computer $x_1 - x_2$ in een precisie van (voor de eenvoud van de notatie) drie cijfers gaan uitrekenen, met de exacte uitkomsten $x_1 = 1000.2$ en $x_2 = 1000.1$ van eerdere tussenberekeningen, dan is het resultaat 0. Deze relatieve fout van 100% wordt veroorzaakt doordat x_1 en x_2 beide als $0.100 \cdot 10^4$ worden opgeslagen, zodat de werkelijke waarde van het verschil $x_1 - x_2$ geheel in de computer-nauwkeurigheid verdwijnt.

Dit verschijnsel ligt ook ten grondslag aan de problemen bij de berekening van $99 - 70\sqrt{2}$ met f_1 in Voorbeeld 1.2.1. Daarbij wordt $f(x_1, x_2) = x_1 - x_2$ berekend in $(99, 70\sqrt{2})$. De waarde 99 kennen we exact, maar voor $70\sqrt{2}$ hebben we slechts 98 als benadering. De relatieve fout hiervan ten opzichte van de echte waarde $98.99\ldots$ valt dan misschien nog wel mee, maar doordat $98.99\ldots \simeq 99$ ontstaan er conditieproblemen bij het berekenen van het verschil.

Het hangt overigens af van de verdere structuur van de berekening in hoeverre grote conditiegetallen ook daadwerkelijk verstorend werken voor het eindresultaat. Zo moge het duidelijk zijn, dat het berekenen van $(x - y) + z$ met $x/y \simeq 1$ een slecht geconditioneerde deelberekening

bevat, nl. het berekenen van $x - y$, maar dat dit toch in het geheel niet ernstig is wanneer $z \gg x, y$.

Samenvattend kunnen we over computer-nauwkeurigheid dus het volgende zeggen:

1. de relatieve fout van de floating point representatie is zeer klein.

2. deze relatieve fout kan echter sterk uitvergroot raken door grote conditiegetallen van berekeningen. Dit kan al daadwerkelijk optreden bij schijnbaar onschuldige bewerkingen, zoals het bepalen van het verschil van twee vrijwel gelijke getallen.

De aanleiding voor de bovenstaande bestudering van computer-nauwkeurigheid was onze wens om de verstorende invloed van de eindige computer-precisie verwaarloosbaar klein te houden t.o.v. de theoretische marge in de numerieke methode. Uit bovenstaande concluderen we dus samenvattend, dat we dan—zo mogelijk—moeten vermijden dat deelberekeningen met grote conditiegetallen een sleutelrol krijgen in de berekening. Met dit als randvoorwaarde zullen we er verder uiteraard ook naar willen streven om zo weinig mogelijk operaties uit te voeren, wat voor de rekentijd ook nog goed uitpakt. De conditiegetallen hebben echter de prioriteit: het berekenen van $99 - 70\sqrt{2}$ kost weliswaar een operatie minder dan het berekenen van $(99 + 70\sqrt{2})^{-1}$, maar toch verdient, zoals we gezien hebben, de laatste berekeningswijze de voorkeur.

Tenslotte herhalen we nogmaals dat we er, bij het analyseren van numerieke methoden, in het vervolg steeds van zullen uitgaan dat de door de computerimplementatie geïntroduceerde fouten verwaarloosbaar klein zijn, waarbij we er ons overigens wel van bewust moeten zijn dat in voorkomende gevallen het realiseren van een dergelijke implementatie grondige expertise kan vergen.

Hoofdstuk 2

Rode draad en vooruitblik

2.1 Rode draad

We hebben gezien dat numerieke wiskunde veelal direct vanuit de praktijk gemotiveerd is. De diversiteit van praktische problemen heeft dan ook geresulteerd in een breed scala van numerieke methoden en technieken; een selectie is noodzakelijk. De keuze voor de in dit college te behandelen onderwerpen wordt als volgt gemotiveerd.

In veel praktijksituaties willen we veranderingen van grootheden in de tijd beschrijven. Vaak houdt dit in, dat in het betreffende model een afgeleide naar de tijd voorkomt (en mogelijk ook naar andere variabelen). In het bijzonder bevat het model dan dus een differentiaalvergelijking. Ook tijdsonafhankelijke problemen worden vaak met differentiaalvergelijkingen gemodelleerd, dit geldt bijv. in voor constructies belangrijke gevallen als de vorm van een kabel, staaf of plaat onder een voorgegeven belasting. De optredende differentiaalvergelijkingen zijn vaker niet dan wel expliciet op te lossen, zodat numerieke technieken de enige mogelijkheid bieden om antwoorden te genereren waar de praktijk iets aan heeft. Gezien deze relevantie van differentiaalvergelijkingen in de praktijk, en gezien de sleutelrol die de numerieke wiskunde dan speelt, kiezen we als rode draad voor dit college de vraag:

Hoe los je een (gewone) differentiaalvergelijking numeriek op?

Met "numeriek oplossen" bedoelen we dan het (kunnen) aangeven van een benadering met voorgegeven nauwkeurigheid, waarbij de fout in een nog nader op te geven zin wordt gemeten. Het woord "gewone" betekent dat de betreffende functie van slechts 1 variabele afhangt. Wanneer er meerdere variabelen in het spel zijn, spreken we van partiële differentiaalvergelijkingen. We beperken ons tot het gewone geval om redenen van ruimte en complexiteit, maar veel methoden en inzichten, die we voor gewone differentiaalvergelijkingen zullen tegenkomen, kunnen mutatis mutandis ook toegepast worden in het partiële geval.

2.2 Vooruitblik

Het kiezen van deze rode draad heeft als afgeleid voordeel, dat we op een natuurlijke manier een aantal numerieke technieken zullen tegenkomen, die ieder op zichzelf al belangrijk genoeg zijn om behandeld te worden. We zullen dit nu illustreren in deze vooruitblik, waar we—zeer informeel—zullen aangeven hoe we het oplossen van differentiaalvergelijkingen zullen benaderen. Voor alle duidelijkheid: het vervolg van deze paragraaf is bedoeld als motivatie en als duiding van de grote lijn. Het is niet wiskundig rigoreus, en we zullen het stramien hieronder uiteindelijk

ook wijzigen. Maar de basisideeën zijn er in terug te vinden, en in de loop van het college zullen de stappen steeds verder worden ingevuld en gepreciseerd.

We kiezen een concreet voorbeeld.

1. Stel, we willen het randwaardeprobleem

$$\begin{cases} (-a(x)u'(x))' = f(x) \\ u(0) = u(1) = 0 \end{cases} \tag{2.2.1}$$

numeriek oplossen op $[0,1]$. (Deze vergelijking is relevant voor de beschrijving van de vervorming van een aan de uiteinden gefixeerde staaf, die in de lengterichting belast wordt. De functie a, resp. f, beschrijft dan de elasticiteit, resp. de belasting.) De functies a en f zijn bekend verondersteld.

2. Als "deus" ex machina merken we op, dat dan blijkbaar voor alle functies v, zodanig dat $v(0) = v(1) = 0$, het volgende geldt:

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx \tag{2.2.2}$$

$$= -au'v|_0^1 + \int_0^1 au'v'\,dx \tag{2.2.3}$$

$$= \int_0^1 au'v'\,dx. \tag{2.2.4}$$

3. Laat nu, voor $q = 0, 1, \ldots$, V_0^q de verzameling van zijn van alle polynomen p van graad ten hoogste q, zodanig dat $p(0) = p(1) = 0$. We willen, voor iedere q, de oplossing u zo goed mogelijk gaan benaderen met een polynoom U^q in V_0^q. Hiertoe eisen we dat

$$\int_0^1 fv\,dx = \int_0^1 a(U^q)'v'\,dx \quad \text{voor alle } v \in V_0^q. \tag{2.2.5}$$

Vergelijk dit met (2.2.2): we verlangen van onze onbekende functie U^q nu dat deze in V_0^q is, wat dus minder keuze laat, maar (2.2.5) is zwakker dan (2.2.2), doordat nu gelijkheid wordt geëist voor $v \in V_0^q$ in plaats van voor alle v.

4. Veronderstel nu, dat we kunnen aantonen dat de rij polynomen $\{U^q\}_{q=0}^{\infty}$ naar een functie U^∞ convergeert, in de een of andere zin die alle formeel plausibele manipulaties hieronder valideert. Fixeer dan een q en een $v \in V_0^q$. Voor alle $\tilde{q} \geq q$ is $V_0^q \subset V_0^{\tilde{q}}$, dus zeker is

$$\int_0^1 fv\,dx = \int_0^1 a(U^{\tilde{q}})'v'\,dx.$$

Laat nu $\tilde{q} \to \infty$ en "concludeer" dat

$$\int_0^1 fv\,dx = \int_0^1 a(U^\infty)'v'\,dx. \tag{2.2.6}$$

Maar $v \in V_0^q$ was willekeurig gekozen, dus (2.2.6) is waar voor alle polynomen v met $v(0) = v(1) = 0$.

5. "En dus" is (2.2.6) waar voor alle functies v met $v(0) = v(1) = 0$.

6. Daaruit "volgt", "door de redenering van stap (1) naar stap (2) om te keren", dat U^∞ de oplossing is van (2.2.1).

7. *Dan hebben we de differentiaalvergelijking numeriek opgelost: de rij polynomen* $\{U^q\}_{q=0}^\infty$ *levert blijkbaar de gezochte benaderingen van de oplossing* U^∞.

De stappen hierboven staan bekend als de Galerkin methode. Of, meer precies: als de globale Galerkin methode, waarbij "globaal" dan aangeeft dat we steeds werken met benaderende functies, die ieder voor zich op het *hele* interval $[0,1]$ gegeven zijn door eenzelfde polynoom. Later zullen we een belangrijke verbetering van deze methode behandelen, de zgn. eindige elementen methode.

Even aannemend, dat de stappen hierboven allemaal zinvol zijn, wat hebben we dan (blijkbaar) nodig om dit schema concreet te kunnen uitvoeren? De crux ligt in (2.2.5). Hoe vinden we hieruit U^q? De enige redelijkerwijs in aanmerking komende methode is de volgende.

Merk op, dat V_0^q voor $q \geq 1$ een $(q-1)$-dimensionale reële vectoruimte is. Bepaal een basis $\{\phi_i\}_{i=1}^{q-1}$ en schrijf $U^q(x) = \sum_{i=1}^{q-1} \xi_i \phi_i(x)$ in termen van de basis. We zoeken $\xi_1, \ldots, \xi_{q-1}$. Het is nu, vanwege de lineariteit, voor (2.2.5) nodig en voldoende is om (2.2.5) op te leggen voor de basiselementen, d.w.z. we moeten

$$\int_0^1 f\phi_j \, dx = \int_0^1 a \left(\sum_{i=1}^{q-1} \xi_i \phi_i' \right) \phi_j' \, dx \quad (j = 1, \ldots, q-1)$$

oplossen. Anders geschreven:

$$\sum_{i=1}^{q-1} \left\{ \int_0^1 a(x)\phi_i'(x)\phi_j'(x) \, dx \right\} \xi_i = \int_0^1 f(x)\phi_j(x) \, dx \quad (j = 1, \ldots, q-1).$$

We herkennen dit als een eindigdimensionaal lineair stelsel in de onbekenden $\xi_1, \ldots, \xi_{q-1}$. Wanneer we dit kunnen oplossen voor alle q, dan kunnen we onze benaderende rij $\{U^q\}_{q=0}^\infty$ bepalen en hebben we blijkbaar de differentiaalvergelijking numeriek opgelost.

Alles overziend hebben we het volgende nodig om bovenstaande methode te kunnen toepassen:

- *Kennis over (benadering met) polynomen*—die hebben we immers een centrale plaats toegemeten.
- *Methoden om numeriek integralen uit te rekenen.* Immers: de coëfficiënten van de eindigdimensionale stelsels zijn gegeven als integralen.
- *Oplossingsmethoden voor (grote) eindigdimensionale lineaire stelsels.*
- *Validatie van deze Galerkin methode, foutschattingen en verbeteringen van de methode.*

Deze vier onderwerpen vormen de hoofdthema's van het college.

We zullen onze motiverende vraag "hoe een gewone differentialvergelijking numeriek op te lossen" overigens slechts zeer gedeeltelijk beantwoorden. In feite behandelen we een illustratie van de eindige elementen methode aan de hand van een eenvoudig eendimensionaal voorbeeld van een zgn. elliptisch randwaardeprobleem. Andere typen gewone differentiaalvergelijkingen en andere numerieke methoden blijven buiten beschouwing. Uiteraard is dit een inperking, maar door behandeling van dit geval, inclusief het bijbehorend functionaalanalytisch kader, wordt de basis gelegd voor numerieke methoden voor fysisch belangrijke hoger dimensionale analoga.

Hoofdstuk 3

Polynomiale approximatie

3.1 Motivatie

Het woord "approximeren" betekent "benaderen". We hebben in het vorige hoofdstuk geprobeerd om de oplossing van de differentiaalvergelijking met een rij polynomen te benaderen, in een verder nog niet nader ingevulde geschikte zin. De impliciete hoop daarbij was, dat dit ook daadwerkelijk *kon*. Is er reden om dit te veronderstellen? Waarom kiezen we voor polynomen, en niet voor iets anders? Drie punten spelen hierbij een rol:

1. Polynomen hebben inderdaad goede globale benaderende eigenschappen. Een bekende stelling op dat gebied is de stelling van Weierstraß:

   Stelling 3.1.1. *Laat $f \in C[a,b]$ en zij $\epsilon > 0$. Dan is er een polynoom p zodanig dat*

   $$\max_{x\in[a,b]} |f(x) - p(x)| < \epsilon$$

   De grafiek van p ligt dus in een band met de grafiek van f als midden, en met in ieder punt van de grafiek van f een verticale breedte ϵ boven en onder.

   Deze stelling is typisch voor de situatie: in veel vectorruimten (van functies), die de polynomen bevatten, en waarin met een norm een afstandsbegrip is gedefinieerd, ligt er willekeurig dicht bij ieder element van die vectorruimte een polynoom. Anders gezegd: de afsluiting van de polynomen is de gehele vectorruimte, ofwel, de polynomen liggen dicht in die ruimte. De stelling van Weierstraß hierboven is equivalent met te zeggen dat de polynomen dicht liggen in $C[a,b]$, voorzien van het afstandsbegrip dat voortvloeit uit de maximumnorm $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$. We zullen in het vervolg overigens herhaaldelijk resultaten formuleren in termen van genormeerde lineaire ruimten. Zie Paragraaf A.1 voor de relevante definities.

   Terzijde 3.1.2. Wanneer we een rij $\{\epsilon_n\}_{n=1}^\infty$ kiezen, z.d.d. $\epsilon_n \to 0$, en we passen op iedere ϵ_n de stelling toe, dan vinden we een rij polynomen $\{p_n\}_{n=0}^\infty$ die uniform naar f convergeert op $[a,b]$, d.w.z.

   $$\lim_{n\to\infty} \|f - p_n\|_\infty = 0. \tag{3.1.1}$$

   Een dergelijke rij kan men ook expliciet aangeven in de vorm van de zgn. *Bernstein-polynomen*. Voor het geval $a = 0$ en $b = 1$—alle andere intervallen zijn door translatie en schaling hierop te herleiden—blijkt aan (3.1.1) voldaan te zijn met

   $$p_n(x) = \sum_{k=0}^{n} \binom{n}{k} f(\frac{k}{n}) x^k (1-x)^{n-k} \quad (n = 0, 1, 2, \ldots).$$

2. Met polynomen kan gemakkelijk gerekend worden: ze zijn algebraïsch (optellen, vermenigvuldigen) gemakkelijk te manipuleren en ze zijn exact te integreren en te differentiëren.

3. Een ander type benaderingsprobleem dan het vinden van een uniforme benadering, zoals onder (1), is het vinden van een polynoom dat in een eindig aantal punten voorgeschreven waarden aanneeemt. Dit zgn. *interpolatieprobleem* treedt in de praktijk vaak op, een typisch voorbeeld is het volgende. Men meet van een afkoelend voorwerp de temperatuur T_j op t_j seconden, voor $j = 1, \ldots, 10$. Gevraagd wordt een schatting van de temperatuur $T(3.2)$ op $t = 3.2$ seconden. Strikt genomen valt daar zonder verdere informatie niets over te zeggen. Immers: het is duidelijk dat de temperatuur wiskundig in principe alles kan zijn op de niet-waargenomen tijdstippen. In de praktijk echter blijkt *interpolatie* (meer precies: polynomiale interpolatie) vaak bruikbare resultaten te geven in dit soort situaties. Dit houdt in, dat men een polynoom p vindt dat:

   (a) voldoet aan $p(t_j) = T_j$ voor $j = 1, \ldots, 10$, en

   (b) onder deze voorwaarde een zo laag mogelijke graad heeft.

   Als schatting voor $T(3.2)$ hanteert men dan $p(3.2)$.
   Het is overigens duidelijk dat een dergelijke schatting voor grote t zeker onbetrouwbaar is: het interpolerend polynoom gaat dan, tenzij het door een uitzonderlijk toeval constant is, naar $\pm\infty$, terwijl dat voor de temperatuur in dit voorbeeld toch zeker niet zo zal zijn. Dit geeft de principiële aard aan van het risico van *extrapolatie*, d.w.z. van het gebruiken van de waarde van een interpolerend polynoom als benadering voor de waarde van de functie, in een punt dat (ver) buiten het interval ligt waarin de punten liggen waarin men daadwerkelijk de waarde van de functie kent.

We zullen in dit hoofdstuk eerst de benaderende eigenschappen van polynomen nader bestuderen. Daartoe zullen we achtereenvolgens:

- laten zien dat een interpolatieprobleem (d.w.z. het benaderen in eindig veel punten) als hierboven altijd een unieke oplossing heeft ,
- aangeven hoe we die oplossing efficiënt kunnen berekenen, en
- de fout bestuderen tussen een polynoom, dat een voorgegeven functie f in een voorgegeven stel punten interpoleert, en die functie zelf. Uiteraard hebben we hierbij extra aannamen over f nodig—zonder dat valt er niets nuttigs over te zeggen.

Daarna zullen we benadering bekijken met functies die stuksgewijs polynomiaal zijn.

3.2 Interpolatieformule van Lagrange

In het vervolg is steeds $q \in \{0, 1, 2, \ldots\}$ en is $[a, b]$ met $a < b$ een interval in $\mathbb{R}$. Laat $P^q[a, b]$ de collectie van (reële) polynomen op $[a, b]$ zijn van graad ten hoogste q, d.w.z. de collectie van alle functies $p : [a, b] \mapsto \mathbb{R}$ van de vorm $p(x) = \sum_{i=0}^{q} c_i x^i$ met $c_i \in \mathbb{R}\,(i = 0, \ldots, q)$.

De verzameling $P^q[a, b]$ wordt met puntsgewijze operaties een vectorruimte (over $\mathbb{R}$):

1. $(p + r)(x) \stackrel{\text{def.}}{=} p(x) + r(x), \quad p, r \in P^q[a, b],\ x \in [a, b]$

2. $(\alpha p)(x) \stackrel{\text{def.}}{=} \alpha p(x), \quad p \in P^q[a, b],\ x \in [a, b].$

De dimensie van $P^q[a,b]$ is $q+1$: als basis kunnen we $\{1, x, \ldots, x^q\}$ nemen (waarom is dit trouwens een basis?).

Een andere basis, die voor de theorievorming veel handiger is dan de bovenstaande, is de *Lagrange-basis* $\{\lambda_0, \lambda_1, \ldots, \lambda_q\}$ die men construeert bij punten $\xi_0 < \xi_1 < \ldots < \xi_q$.[1] De Lagrange-basis hangt van de ξ_i af, maar we onderdrukken dat in de notatie. De betreffende elementen λ_i van de basis, de zgn. *Lagrange-polynomen*, worden voor $i = 0, 1, \ldots, q$ gedefinieerd door:

$$\lambda_i(x) \overset{\text{def.}}{=} \frac{(x-\xi_0)(x-\xi_1)\cdot\ldots\cdot(x-\xi_{i-1})(x-\xi_{i+1})\cdot\ldots\cdot(x-\xi_q)}{(\xi_i-\xi_0)(\xi_i-\xi_1)\cdot\ldots\cdot(\xi_i-\xi_{i-1})(\xi_i-\xi_{i+1})\cdot\ldots\cdot(\xi_i-\xi_q)} \quad (i = 0, 1, \ldots, q).$$

Een leeg produkt treedt op voor $q = 0$, we volgen dan de gebruikelijke conventie voor lege produkten en definiëren $\lambda_0(x) = 1$.

We zien dat $\deg \lambda_i = q$, en dat verder $\lambda_i(\xi_j) = \delta_{i,j}$ (Kronecker delta) voor $i, j \in \{0, 1, \ldots, q\}$:

$$\lambda_i(\xi_j) = \begin{cases} 1 & \text{als } i = j; \\ 0 & \text{als } i \neq j. \end{cases} \tag{3.2.1}$$

Voor $q = 0$ bestaat de basis dus alleen uit de constante functie 1, voor $q = 1$ zijn er twee elementen λ_0 en λ_1, gegeven door (teken zelf de grafieken)

$$\lambda_0(x) = \frac{x-\xi_1}{\xi_0-\xi_1} \quad , \quad \lambda_1(x) = \frac{x-\xi_0}{\xi_1-\xi_0}.$$

De Lagrange-basis *is* inderdaad een basis voor $P^q[a,b]$. Immers: als $\sum_{i=0}^q \alpha_i\lambda_i(x) = 0$, dan geeft evaluatie in ξ_j onmiddellijk dat $\alpha_j = 0$, en dat voor alle j. Op grond van het juiste aantal elementen $q+1$ vormen de blijkbaar lineair onafhankelijke Lagrange-polynomen inderdaad een basis van $P^q[a,b]$.

Terzijde 3.2.1. Voor wie met de terminologie bekend is: wanneer we voor $i = 0, 1, \ldots, q$ een evaluatie-afbeelding $\mathrm{ev}_i : P^q[a,b] \to \mathbb{R}$ definiëren door $\mathrm{ev}_i(p) = p(\xi_i)$, dan is $\{\mathrm{ev}_0, \mathrm{ev}_1, \ldots, \mathrm{ev}_q\}$ een basis van de duale vectorruimte van $P^q[a,b]$. Deze basis is volgens (3.2.1) in dualiteit met de Lagrange-basis $\{\lambda_0, \lambda_1, \ldots, \lambda_q\}$ van $P^q[a,b]$.

Deze Lagrange-basis stelt ons in staat om het interpolatieprobleem op te lossen, waarvan we een voorbeeld hadden bij het afkoelend voorwerp in Paragraaf 3.1.

Stelling 3.2.2. *Er is, voor gegeven punten $a \leq \xi_0 < \xi_1 \ldots < \xi_q \leq b$ en waarden $\alpha_0, \alpha_1, \ldots, \alpha_q$, precies één $p \in P^q[a,b]$ zodanig dat $p(\xi_i) = \alpha_i$ voor $i = 0, 1, \ldots, q$. Dit polynoom is*

$$p(x) = \sum_{i=0}^{q} \alpha_i\lambda_i(x) \in P^q[a,b] \quad \text{(Interpolatieformule van Lagrange)}, \tag{3.2.2}$$

met $\{\lambda_0, \lambda_1, \ldots, \lambda_q\}$ de Lagrange-basis van $P^q[a,b]$, behorend bij de punten $\xi_0, \xi_1, \ldots, \xi_q$.

Bewijs. Uniciteit. Wanneer p en $\tilde{p}$ beide de eigenschappen in de stelling hebben, dan is $p - \tilde{p}$ een polynoom van graad ten hoogste q met een nulpunt in de $q+1$ verschillende punten $\xi_0, \xi_1, \ldots, \xi_q$. Dus $p - \tilde{p} = 0$.
Existentie. Met p als in (3.2.2), is volgens (3.2.1) inderdaad $p(\xi_i) = \alpha_i$ voor alle i. Verder is ook $p \in P^q[a,b]$, zoals vereist. □

[1] Het is hier en in het vervolg overigens niet noodzakelijk dat de ξ_i op grootte geordend zijn; voor het mentale beeld is het echter wat overzichtelijker en we zullen het daarom bijna altijd zo noteren.

3.3 Interpolatiemethode van Newton

De Lagrange-polynomen in de vorige paragraaf zijn voor de theorievorming prettig om mee te werken vanwege (3.2.1), maar ze hebben een belangrijk bezwaar, wanneer we ze via vergelijking (3.2.2) als bouwstenen zouden gebruiken voor het vinden van interpolerende polynomen in concrete situaties. Stel immers, dat we (zeg, in een experimentele situatie) voor gegeven punten $\xi_0, \xi_1, \ldots, \xi_q$ de Lagrange-polynomen hebben uitgerekend, en dat we vervolgens op basis van (3.2.2) met de (gemeten) waarden $\alpha_0, \alpha_1, \ldots, \alpha_q$ het interpolerend polynoom $p(x) = \sum_{i=0}^{q} \alpha_i \lambda_i(x)$ in deze punten hebben bepaald. We willen nu een extra punt ξ_{q+1} toevoegen, met bijbehorende waarde α_{q+1} (bijvoorbeeld omdat we een extra meting hebben gedaan die we ook willen verwerken). Kunnen we dan ons oude rekenwerk nog gebruiken om het nieuwe interpolerende polynoom uit te rekenen? Helaas is dit niet het geval: het is niet eenvoudig om de nu benodigde $q+2$ Lagrange-polynomen uit de $q+1$ al berekende polynomen te verkrijgen.

De *interpolatiemethode van Newton* ondervangt dit bezwaar. Gegeven punten $\xi_0, \xi_1, \ldots, \xi_q$ en waarden $\alpha_0, \alpha_1, \ldots, \alpha_q$, berekent men met deze *recursieve* methode allereerst een polynoom p_0 van graad ten hoogste 0 dat in ξ_0 interpoleert, daarna een polynoom $p_{0,1}$ van graad ten hoogste 1 dat in ξ_0 en ξ_1 interpoleert, daarna een polynoom $p_{0,1,2}$ van graad ten hoogste 2 dat in ξ_0, ξ_1 en ξ_2 interpoleert,....., en tenslotte een polynoom $p_{0,1,2,\ldots,q}$ van graad ten hoogste q dat in alle gegeven punten interpoleert. Het toevoegen van 1 extra punt levert dan, vanwege het recursieve karakter van de methode, simpelweg 1 extra stap op die gebruik maakt van het al verrichte rekenwerk. De methode werkt als volgt.

- Kies $p_0(x) = \alpha_0$. Dan interpoleert p_0 in ξ_0, en inderdaad is de graad van p_0 maximaal 0.

- We willen $p_{0,1}$ vinden door aan p_0 een term toe te voegen die de al bereikte interpolatie in ξ_0 manifest niet verstoort, en zodanig dat ook nog interpolatie gaat plaatsvinden in ξ_1. Merk hiertoe op, dat $p_0(x) + \eta_1(x - \xi_0)$ inderdaad nog steeds in ξ_0 interpoleert voor alle (!) η_1. Deze vrijheid in η_1 benutten we, door te eisen dat interpolatie ook nog in ξ_1 gaat plaatsvinden, d.w.z. door te eisen dat $\alpha_1 = p_0(\xi_1) + \eta_1(\xi_1 - \xi_0)$. Dit legt η_1 dan vast (want $\xi_1 - \xi_0 \neq 0$) in termen van bekende grootheden (ook p_0 kennen we immers inmiddels!). Nadat η_1 hierdoor is vastgelegd, stellen we vast dat blijkbaar $p_{0,1}(x) = p_0(x) + \eta_1(x - \xi_0)$, omdat het rechterlid interpoleert in ξ_0, ξ_1 en graad maximaal 1 heeft.

- We willen $p_{0,1,2}$ vinden door aan $p_{0,1}$ een term toe te voegen die de al bereikte interpolatie in ξ_0, ξ_1 manifest niet verstoort, en zodanig dat ook nog interpolatie gaat plaatsvinden in ξ_2.
  Merk hiertoe op, dat $p_{0,1}(x) + \eta_2(x - \xi_0)(x - \xi_1)$ inderdaad nog steeds in ξ_0, ξ_1 interpoleert voor alle (!) η_2. Deze vrijheid in η_2 benutten we door te eisen dat interpolatie ook nog in ξ_2 gaat plaatsvinden, d.w.z. door te eisen dat $\alpha_2 = p_{0,1}(\xi_2) + \eta_2(\xi_2 - \xi_0)(\xi_2 - \xi_1)$. Dit legt η_2 dan vast (want $(\xi_2 - \xi_0)(\xi_2 - \xi_1) \neq 0$) in termen van bekende grootheden (ook $p_{0,1}$ kennen we immers inmiddels al!). Nadat η_2 hierdoor is vastgelegd, stellen we vast dat blijkbaar $p_{0,1}(x) = p_0(x) + \eta_2(x - \xi_0)(x - \xi_1)$, omdat het rechterlid interpoleert in ξ_0, ξ_1, ξ_2 en graad maximaal 2 heeft.

- Zo vervolgend, willen we uiteindelijk $p_{0,1,2,\ldots,q}$ vinden door aan $p_{0,1,2,\ldots,q-1}$ een term toe te voegen die de al bereikte interpolatie in $\xi_0, \xi_1, \ldots, \xi_{q-1}$ manifest niet verstoort, en zodanig dat ook nog interpolatie gaat plaatsvinden in ξ_q.
  Merk hiertoe op, dat $p_{0,1,2,\ldots,q-1}(x) + \eta_q(x - \xi_0)(x - \xi_1)\ldots(x - \xi_{q-1})$ inderdaad nog steeds in $\xi_0, \xi_1, \ldots, \xi_{q-1}$ interpoleert voor alle (!) η_q. Deze vrijheid in η_q benutten we door te eisen dat interpolatie ook nog in ξ_q gaat plaatsvinden, d.w.z. door te eisen dat $\alpha_q =$

$p_{0,1,2,\ldots,q-1}(\xi_q) + \eta_q(\xi_q - \xi_0)(\xi_q - \xi_1)\cdots(\xi_q - \xi_{q-1})$. Dit legt η_q dan vast (want $(\xi_q - \xi_0)(\xi_q - \xi_1)\ldots(\xi_q - \xi_{q-1}) \neq 0$) in termen van bekende grootheden (ook $p_{0,1,2,\ldots,q-1}$ kennen we immers inmiddels al!). Nadat η_q hierdoor is vastgelegd, stellen we tenslotte vast dat blijkbaar $p_{0,1,2,\ldots,q}(x) = p_0(x) + \eta_q(x - \xi_0)(x - \xi_1)\cdots(x - \xi_{q-1})$, omdat het rechterlid interpoleert in $\xi_0, \xi_1, \ldots, \xi_q$ en graad maximaal q heeft.

Uiteindelijk vindt men aldus een uitdrukking voor het interpolerend polynoom van de vorm

$$p_{0,\ldots,q} = \sum_{i=0}^{q} \eta_i(x - \xi_0)\cdots(x - \xi_{i-1}).$$

3.4 Foutschatting bij interpolatie

Laat $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$, en zij $f : [a, b] \to \mathbb{R}$ gegeven. We definiëren

$$\pi_q f(x) = \sum_{i=0}^{q} f(\xi_i)\lambda_i(x).$$

Blijkens het voorgaande is $\pi_q f$ het unieke polynoom van graad ten hoogste q dat in de punten $\xi_0, \xi_1, \ldots, \xi_q$ met f overeenstemt. Deze *interpolant van f* (ook wel: dit *interpolerend polynoom van f*) *in de punten $\xi_0, \xi_1, \ldots, \xi_q$* van f lost dus het interpolatieprobleem van f op voor de punten $\xi_0, \xi_1, \ldots, \xi_q$, zoals we dat bij het afkoelend voorwerp in Paragraaf 3.1 zagen optreden.

In deze paragraaf houden we ons voornamelijk bezig met het verschil $f(x) - \pi_q f(x)$, voor $x \in [a, b]$. Dit verschil is uiteraard 0 in ieder van de interpolatiepunten, maar behalve dat kunnen we er, zonder verdere aannamen over f, niet zo veel over zeggen. Tussen de punten in kan f immers sterk oscillerend gedrag vertonen, wat dan leidt tot een groot verschil. Deze observatie leidt wel tot de hoop, dat het misschien mogelijk is om iets over het verschil te zeggen, wanneer we de oscillatie maar in de een of andere zin kunnen controleren. Dit is inderdaad het geval, blijkens de volgende stelling.

Stelling 3.4.1. *Laat $f \in C^{q+1}[a, b]$ en $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$. Laat $\pi_q f$ de interpolant van f zijn in de punten $\xi_0, \xi_1, \ldots, \xi_q$. Dan is er voor iedere $x \in [a, b]$ een $\tau_x \in [a, b]$, zodanig dat:*

$$f(x) - \pi_q f(x) = \frac{1}{(q+1)!}(x - \xi_0)(x - \xi_1)\ldots(x - \xi_q)f^{(q+1)}(\tau_x). \tag{3.4.1}$$

Voor het bewijs gebruiken we het volgende hulpresultaat (een generalisatie van de stelling van Rolle):

Lemma 3.4.2. *Laat $n \geq 0$, zij $f \in C^n[a, b]$, en laat $a \leq \eta_1 < \eta_2 < \ldots < \eta_{n+1} \leq b$. Veronderstel dat $f(\eta_i) = 0$ voor $i = 1, \ldots, n + 1$. Dan heeft $f^{(n)}$ tenminste één nulpunt op $[a, b]$.*

Bewijs. Met inductie. Voor $n = 0$ voldoet η_1 als nulpunt. Voor $n \geq 1$ passen we op ieder van de n intervallen $[\eta_1, \eta_2], [\eta_2, \eta_3], \ldots, [\eta_n, \eta_{n+1}]$ de stelling van Rolle toe. We concluderen, dat er in het inwendige van deze respectievelijke intervallen respectievelijke punten $\tilde{\eta}_1, \tilde{\eta}_2, \ldots, \tilde{\eta}_n$ zijn, zodanig dat $f'(\tilde{\eta}_i) = 0$ voor $i = 1, \ldots, n$. Daar $f' \in C^{n-1}[\eta_1, \eta_n]$ volgt nu uit de inductieveronderstelling, toegepast op f' en het interval $[\eta_1, \eta_n]$, dat $(f')^{(n-1)}$ een nulpunt heeft op $[\eta_1, \eta_n] \subset [a, b]$. □

We voltooien nu het bewijs van Stelling 3.4.1.

Bewijs. Fixeer x. Als $x = \xi_i$ voor een of andere i, dan zijn in (3.4.1) linker- en rechterlid beide 0 en is de stelling duidelijk. Neem dus aan, dat $x \neq \xi_i$ voor $i = 0, 1, 2, \ldots, q$, en beschouw de hulpfunctie $g : [a, b] \mapsto \mathbb{R}$, gegeven door

$$g(s) = f(s) - \pi_q f(s) - K_x(s - \xi_0)(s - \xi_1) \ldots (s - \xi_q), \tag{3.4.2}$$

waarbij we K_x zo kiezen dat $g(x) = 0$ (waarom kan dit inderdaad?). Merk op, dat g een nulpunt heeft in de $q + 2$ verschillende punten $x, \xi_0, \xi_1, \ldots, \xi_q$. Omdat $g \in C^{q+1}[a, b]$, is er volgens Lemma 3.4.2 een $\tau_x \in [a, b]$ zodanig dat $g^{(q+1)}(\tau_x) = 0$. Neem nu de $q+1$-de afgeleide in (3.4.2), en merk op dat de $q+1$-de afgeleide van $\pi_q f$ (een polynoom van graad ten hoogste q) identiek 0 is. Net zo is, op grond van de graad, de term s^{q+1} de enige term uit $(s-\xi_0)(s-\xi_1)\ldots(s-\xi_q)$, die niet verdwijnt na $q+1$ maal differentiëren. Blijkbaar is $0 = g^{(q+1)}(\tau_x) = f^{(q+1)}(\tau_x) - 0 - (q+1)!K_x$. Omdat $g(x) = 0$ per constructie, concluderen we uit (3.4.2) dat

$$0 = g(x) = f(x) - \pi_q f(x) - \frac{1}{(q+1)!}(x - \xi_0)(x - \xi_1) \ldots (x - \xi_q) f^{(q+1)}(\tau_x). \tag{3.4.3}$$

□

Terzijde 3.4.3. Voor $n = 1$ hervinden we uit Lemma 3.4.2 de stelling van Rolle onder de eis dat $f \in C^1[a, b]$, hetgeen een sterkere eis is dan de reeds voldoende voorwaarde dat f continu is op $[a, b]$ en differentieerbaar op (a, b). Het is mogelijk om de voorwaarden in Lemma 3.4.2 zodanig te verzwakken, dat deze *scherpe* versie van de stelling van Rolle dan inderdaad precies correspondeert met het geval $n = 1$. Dit kan men laten doorwerken in resultaten, zoals Stelling 3.4.1, die op Lemma 3.4.2 voortbouwen. De praktische consequenties zijn echter gering.

Gevolg 3.4.4. *Onder de voorwaarden van Stelling 3.4.1 is*

$$|f(x) - \pi_q f(x)| \leq \frac{1}{(q+1)!}|(x - \xi_0)(x - \xi_1) \ldots (x - \xi_q)| \, \|f^{(q+1)}\|_{[a,b],\infty}.$$

De in Gevolg 3.4.4 optredende uitdrukking $|(x - \xi_0)(x - \xi_1) \ldots (x - \xi_q)|$ heeft een maximum op $[a, b]$, dat van $\xi_0, \xi_1, \ldots, \xi_q$ afhangt. Als bovengrens voor dit maximum kunnen we $(b - a)^{q+1}$ nemen, zodat blijkbaar

$$\|f - \pi_q f\|_{[a,b],\infty} \leq \frac{(b-a)^{q+1}}{(q+1)!}\|f^{(q+1)}\|_{[a,b],\infty} \quad (f \in C^{(q+1)}[a, b]). \tag{3.4.4}$$

Indien $f \in C^\infty[a, b]$, dan kunnen we dit voor alle q toepassen: als (maar dat hoeft niet zo te zijn) het rechterlid in (3.4.4) naar 0 gaat, dan vinden we blijkbaar een rij polynomen $\{\pi_q f\}_{q=0}^\infty$ die op $[a, b]$ uniform naar f convergeert, ongeacht de keuze van de interpolatiepunten. Dat een uniform naar f convergente rij polynomen bestaat, wisten we al uit Stelling 3.1.1, maar blijkbaar kunnen we, wanneer de afgeleiden van f niet te hard groeien, zo'n rij ook expliciet *door interpolatie* vinden. Dit treedt bijv. op voor e^x, $\sin x$ en $\cos x$: alle afgeleiden zijn daar zelfs begrensd door eenzelfde getal. Bij iedere keuze van een rij $\xi_0^{(0)}, \xi_0^{(1)} < \xi_1^{(1)}, \xi_0^{(2)} < \xi_1^{(2)} < \xi_2^{(2)}, \ldots$ van verzamelingen interpolatiepunten in $[a, b]$, convergeert de bijbehorende rij $\{\pi_q f\}_{q=0}^\infty$ van interpolerende polynomen dan dus uniform op $[a, b]$ naar de betreffende functie.

Terzijde 3.4.5. De volgende drie punten beogen de bovenstaande resultaten over puntsgewijze en uniforme foutschatting meer in perspectief te plaatsen:

1. Beschouw de functie $f(x) = \frac{1}{x^2+1}$ op $[-5, 5]$. Laat $\pi_n f$ het polynoom zijn dat f interpoleert in de $n + 1$ equidistante punten $-5 = \xi_0 < \xi_1 < \ldots < \xi_n = 5$. Dan is voor *iedere* $x \in (-5, 5)$ de rij $\{p_n(x)\}_{n=0}^\infty$ *divergent.*

2. Voor iedere tweemaal continu differentieerbare functie kan men een uniform benaderende rij polynomen vinden met behulp van interpolatie. Meer precies: er is expliciet een rij van verzamelingen interpolatiepunten $\xi_0^{(0)}, \xi_0^{(1)} < \xi_1^{(1)}, \xi_0^{(2)} < \xi_1^{(2)} < \xi_2^{(2)}, \ldots$ in $[a,b]$ bekend, zodanig dat voor iedere $f \in C^2[a,b]$ geldt dat
$$\lim_{n\to\infty} \|f(x) - \pi_n f\|_{[a,b],\infty} = 0,$$
waarbij $\pi_n f$ bepaald wordt met de n-de collectie interpolatiepunten.
Overigens, dat geldt dan dus ook voor de functie onder 1. Waarom is dit geen tegenspraak?

3. De Bernstein-polynomen in Terzijde 3.1.2 maken, net als in (1), gebruik van equidistante punten. Toepassing van de Bernstein-constructie, met verdiscontering van translatie en schaling, geeft een op equidistante punten in $[-5,5]$ gebaseerde rij polynomen, die uniform op dat interval naar de functie in 1 convergeert. Waarom is dit geen tegenspraak met 1?

Het volgende resultaat, dat we later nodig hebben, behandelt de fout tussen de *afgeleide* van een lineaire interpolant en de afgeleide van de geïnterpoleerde functie.

Stelling 3.4.6. *Laat $f \in C^2[a,b]$ en $a \le \xi_0 < \xi_1 \le b$. Zij $\pi_1 f$ de lineaire interpolant vam f in ξ_0 en ξ_1. Dan is voor $x \in [a,b]$*
$$|f'(x) - (\pi_1 f)'(x)| \le \frac{1}{2}\frac{(x-\xi_0)^2 + (x-\xi_1)^2}{(\xi_1 - \xi_0)} \|f''\|_{[a,b],\infty}.$$

Bewijs. De Lagrange-polynomen zijn voor $q = 1$ gegeven door
$$\lambda_0(x) = \frac{x-\xi_1}{\xi_0 - \xi_1} \quad , \quad \lambda_1(x) = \frac{x-\xi_0}{\xi_1 - \xi_0},$$
waaruit men nagaat dat
$$\lambda_0(x) + \lambda_1(x) = 1 \quad , \quad (\xi_0 - x)\lambda_0'(x) + (\xi_1 - x)\lambda_1'(x) = 1.$$
(De eerste van deze vergelijkingen is vanuit de eigenschappen van de Lagrange-polynomen overigens zonder rekenwerk in te zien. Hoe?) We weten uit Stelling 3.2.2 dat $\pi_1 f(x) = f(\xi_0)\lambda_0(x) + f(\xi_1)\lambda_1(x)$, zodat $(\pi_1 f)'(x) = f(\xi_0)\lambda_0'(x) + f(\xi_1)\lambda_1'(x)$.
We passen de stelling van Taylor toe met basispunt x en vinden voor $i = 0, 1$ een $\eta_i \in [a,b]$, zodanig dat
$$f(\xi_i) = f(x) + f'(x)(\xi_i - x) + \frac{1}{2}f''(\eta_i)(\xi_i - x)^2.$$
We berekenen dan vervolgens
$$\begin{aligned}
f'(x) - (\pi_1 f)'(x) &= f'(x) - f(\xi_0)\lambda_0'(x) - f(\xi_1)\lambda_1'(x) \\
&= f'(x) - \left\{ f(x) + f'(x)(\xi_0 - x) + \frac{1}{2}f''(\eta_0)(\xi_0 - x)^2 \right\} \lambda_0'(x) \\
&\quad - \left\{ f(x) + f'(x)(\xi_1 - x) + \frac{1}{2}f''(\eta_1)(\xi_1 - x)^2 \right\} \lambda_1'(x) \\
&= f'(x) - f(x)\left\{\lambda_0'(x) + \lambda_1'(x)\right\} - f'(x)\left\{(\xi_0 - x)\lambda_0'(x) + (\xi_1 - x)\lambda_1'(x)\right\} \\
&\quad - \frac{1}{2}f''(\eta_0)(\xi_0 - x)^2\lambda_0'(x) - \frac{1}{2}f''(\eta_1)(\xi_1 - x)^2\lambda_1'(x) \\
&= f'(x) - f(x)\cdot 1' - f'(x)\cdot 1 - \frac{1}{2}\frac{f''(\eta_0)(\xi_0 - x)^2}{\xi_1 - \xi_0} - \frac{1}{2}\frac{f''(\eta_1)(\xi_1 - x)^2}{\xi_0 - \xi_1} \\
&= -\frac{1}{2}\frac{f''(\eta_0)(\xi_0 - x)^2}{\xi_1 - \xi_0} - \frac{1}{2}\frac{f''(\eta_1)(\xi_1 - x)^2}{\xi_0 - \xi_1}.
\end{aligned}$$
Het nemen van absolute waarden uiterst links en uiterst rechts voltooit het bewijs. □

Terzijde 3.4.7. Het type interpolatie dat we bekeken hebben staat ook wel bekend als Lagrange-interpolatie, hetgeen betekent dat de *waarden* van het interpolerend polynoom voorgeschreven worden in een eindig aantal punten. Over (hogere orde) afgeleiden wordt niets geëist, maar het ligt voor de hand om in die richting te generaliseren. Dit leidt tot *Hermite-interpolatie*, in het Engels ook wel "osculating interpolation" genoemd. Men wil bij Hermite-interpolatie voor een polynoom in een eindig aantal punten de waarden voorschrijven van alle afgeleiden, van de 0-de orde afgeleide tot en met de afgeleide van een mogelijk hogere zekere orde, waarbij die mogelijk hogere orden dan weer van het punt mag afhangen. Meer precies: gegeven $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$, niet-negatieve gehele getallen $s_0, s_1, \ldots, s_q$, en waarden $\alpha_i^{(k)}$ voor $0 \leq q$ en $0 \leq k \leq s_i$, wordt een polynoom p gezocht, dat voldoet aan $p^{(k)}(\xi_i) = \alpha_i^{(k)}$ voor alle $0 \leq i \leq q$ en $0 \leq k \leq s_i$. Lagrange-interpolatie correspondeert dus met het geval $s_0 = s_1 = \ldots = s_q = 0$. De algemene stelling luidt, dat er een uniek polynoom van graad ten hoogste $q + \sum_{i=0}^{q} s_i$ bestaat dat aan de vereisten bij Hermite-interpolatie voldoet.

Uitgaande van een gegeven voldoende gladde functie f kan men uiteraard $\alpha_i^{(k)} = f^{(k)}(\xi_i)$ kiezen, voor alle i en k. Van het resulterende interpolerende polynoom stemt dan dus in ξ_i de waarde van iedere afgeleide, van de 0-de orde afgeleide tot en met de afgeleide van de s_i-de orde, overeen met de corresponderende waarde voor f, en dat voor alle i.

Er is ook in dit algemenere geval een foutuitdrukking bekend: indien $f \in C^{q+1+\sum_{i=0}^{q} s_i}[a,b]$, dan is er voor iedere $x \in [a,b]$ een $\tau_x \in [a,b]$, zodanig dat

$$f(x) - p(x) = \frac{1}{(q+1+\sum_{i=0}^{q} s_i)!}(x-\xi_0)^{s_0+1}(x-\xi_1)^{s_1+1}\ldots(x-\xi_q)^{s_q+1} f^{(q+1+\sum_{i=0}^{q} s_i)}(\tau_x).$$

Niet alleen Lagrange-interpolatie is een bekend speciaal geval van Hermite-interpolatie. Voor $q = 0$ geeft Hermite-interpolatie van een functie eveneens vertrouwde resultaten. Welke?

3.5 Stuksgewijze polynomiale approximatie

In de vorige paragrafen hebben we functies in C[a,b] benaderd met polynomen. Gegeven een aantal interpolatiepunten $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$, geldt blijkens (3.4.1) voor $f \in C^{q+1}[a,b]$ de afschatting

$$\|f - \pi_q f\|_\infty \leq \frac{1}{(q+1)!}\|(x-\xi_0)(x-\xi_1\ldots)(x-\xi_q)\|_\infty \, \|f^{(q+1)}\|_\infty \tag{3.5.1}$$

$$\leq \frac{(b-a)^{q+1}}{(q+1)!}\|f^{(q+1)}\|_\infty, \tag{3.5.2}$$

waarbij we kortheidshalve het interval $[a,b]$ in de notatie van de norm hebben onderdrukt.

Uit Stelling 3.4.6 volgt verder dat voor $f \in C^2[a,b]$

$$\|f' - (\pi_1 f)'\|_\infty \leq (b-a)\|f''\|_\infty. \tag{3.5.3}$$

Wanneer voor $f \in C^\infty[a,b]$ de rij $\{\|f^{(q+1)}\|\}_{q=0}^{\infty}$ niet te hard groeit, dan vinden we op basis van (3.5.1) blijkbaar eenvoudig een rij polynomen die uniform naar f convergeert. We hebben al eerder opgemerkt dat het helaas voor algemene f niet zo eenvoudig ligt, zie bijvoorbeeld het eerste punt in Terzijde 3.4.5. Een ander voorbeeld is de functie $\exp(-x^2)$ op $[-2,2]$: naarmate men meer equidistante interpolatiepunten op dit interval neemt, wordt de benadering in $\|\,.\,\|_\infty$ steeds *slechter*.

We zoeken daarom naar een alternatief benaderingsprocédé, dat nog steeds met eenvoudige functies werkt, maar pathologieën als hierboven vermijdt. We zullen hiertoe overgaan op een constructie, waarbij we het interval steeds verder opdelen, terwijl we werken met (stuksgewijze) polynomen van een *vaste* maximum graad. In de constructie tot nu toe was het interval vast,

maar waren we bereid om de graad steeds verder laten toenemen. De nieuwe constructie is als volgt.

Laat $a = x_0 < x_1 < \ldots < x_{m+1} = b$ een partitie van $[a, b]$ zijn, voor zekere $m \in \{0, 1, \ldots\}$. Zij $I_i = [x_{i-1}, x_i]$ het i-de subinterval, voor $i = 1, \ldots, m+1$. Definieer de bijbehorende *maasfunctie* (ook wel, en misschien zelfs beter: maaswijdtefunctie; Engels: "mesh function") $h : [a, b] \mapsto \mathbb{R}_{\geq 0}$, als

$$h(x) = \begin{cases} 0 & \text{als } x = x_i \text{ voor een of andere } i \in \{0, 1, \ldots, m+1\}; \\ x_i - x_{i-1} & \text{als } x \in (x_{i-1}, x_i) \text{ voor een of andere } i \in \{0, 1, \ldots, m+1\}. \end{cases}$$

Met uitzondering van de punten van de partitie zelf, geeft h dus in ieder punt de lengte van het subinterval waar het betreffende punt in ligt. Teken zelf een voorbeeld van een partitie met bijbehorende maasfunctie—er verschijnen vierkantjes. De maasfunctie legt de partitie vast; we zullen daarom vaak de letter h gebruiken om aan te geven dat iets van een partitie afhangt.

Voor $q \geq 1$ en een partitie met bijbehorende maasfunctie h zij

$$V_h^q[a, b] = \{v \in C[a, b] \mid v|_{I_i} \in P^q(I_i) \text{ voor } i = 1, 2, \ldots, m+1\}.$$

$V_h^q[a, b]$ bestaat dus uit continue functies die op ieder van de subintervallen gegeven worden door een polynoom van graad ten hoogste q. Die polynomen mogen van de subintervallen afhangen, maar moeten op elkaar aansluiten in de gemeenschappelijke randpunten. Teken zelf een plaatje voor $q = 1$ en $q = 2$. We nemen $q \geq 1$ omdat V_h^0 nooit goede approximanten voor algemene functies kan opleveren—waarom trouwens?

Het idee is nu, om een functie $f \in C[a, b]$ te gaan benaderen met functies in $V_h^q[a, b]$. We kiezen hiertoe, bij h en q vast, in ieder subinterval I_i interpolatiepunten $x_i = \xi_0^{(i)} < \xi_1^{(i)} < \ldots < \xi_q^{(i)} = x_{i+1}$, en construeren op die manier voor iedere i een polynoom $\pi_q(f|_{I_i}) \in P^q(I_i)$. Maak zelf een tekening voor $q = 1$ en $q = 2$, bij dezelfde functie f en dezelfde partitie van $[a, b]$.

De functies $\pi_q(f|_{I_i})$ zijn de restricties van een functie in $V_h^q[a, b]$, deze globale ("aan elkaar geplakte") functie noteren we als $\pi_{q,h} f \in V_h^q[a, b]$. (Deze functie $\pi_{q,h} f$ hangt af van de interpolatiepunten in ieder van de subintervallen, maar dat onderdrukken we in de notatie.) Merk overigens op, dat het inderdaad mogelijk is om aldus $\pi_{q,h} f$ te definiëren: in de punten van de partitie (afgezien van in a en b) zou er een definitie-probleem kunnen optreden, omdat die punten in twee subintervallen liggen. Dat probleem treedt echter niet werkelijk op, omdat de beide interpolanten daar dezelfde waarde aannemen, nl. de waarde die f daar aanneemt. Het is daarmee dan tegelijk ook duidelijk dat $\pi_{q,h} f$ continu is.

We zullen nu $\|f - \pi_{q,h} f\|_\infty$ afschatten. Op ieder interval I_i hebben we voor $f \in C^{q+1}[a, b]$ op grond van (3.5.1):

$$\|f|_{I_i} - \pi_q(f|_{I_i})\|_\infty \leq \frac{(x_i - x_{i-1})^{q+1}}{(q+1)!} \|f^{(q+1)}|_{I_i}\|_\infty = \frac{1}{(q+1)!} \|h^{q+1}|_{I_i} f^{(q+1)}|_{I_i}\|_\infty.$$

Deze lokale schattingen gebruiken bij de tweede stap in de volgende globale schatting:

$$\begin{aligned} \|f - \pi_{q,h} f\|_\infty &= \max_{i=1,\ldots,m+1} \|f|_{I_i} - \pi_q(f|_{I_i})\|_\infty && (3.5.4) \\ &\leq \frac{1}{(q+1)!} \max_{i=1,\ldots,m+1} \|h^{q+1}|_{I_i} f^{(q+1)}|_{I_i}\|_\infty && (3.5.5) \\ &= \frac{1}{(q+1)!} \|h^{q+1} f^{(q+1)}\|_\infty. && (3.5.6) \end{aligned}$$

We kunnen nog een stap verder gaan en hieruit concluderen dat blijkbaar

$$\|f - \pi_{q,h} f\|_\infty \leq \frac{1}{(q+1)!} \|h\|_\infty^{q+1} \|f^{(q+1)}\|_\infty. \tag{3.5.7}$$

Hieruit zien we dat, zoals beoogd, deze constructie inderdaad bij vaste q voor *alle* $f \in C^{q+1}[a,b]$ werkt, in de zin dat $\lim_{\|h\|_\infty \to 0} \|f - \pi_{q,h} f\|_\infty = 0$: we kunnen willekeurig precieze benaderingen krijgen door de maximum maaswijdte maar voldoende te verkleinen.

De bovengrens in (3.5.7) heeft de eigenschap dat, voor vaste $q_1 > q_2$ en vaste $f \in C^{q_1+1}[a,b]$, deze bovengrens voor $\|f - \pi_{q_1,h} f\|_\infty$ *uiteindelijk* altijd kleiner zal zijn dan die voor $\|f - \pi_{q_2,h} f\|_\infty$ wanneer $\|h\|_\infty \to 0$. Met andere woorden: hogere graad levert hier (bij voldoende kleine maaswijdte) wél een beter resultaat—vgl. de gesignaleerde pathologie die bij globale polynomiale interpolatie kan optreden.

Een in de praktijk belangrijk (want eenvoudig) geval treedt op voor $q = 1$. Er zijn dan in ieder subinterval $[x_{i-1}, x_i]$ slechts twee interpolatiepunten, namelijk x_{i-1} en x_i zelf. De interpolant op ieder subinterval is lineair en vastgelegd door de waarden van f in de randpunten. De grafiek van $\pi_{1,h}$ op $[a,b]$ bestaat simpelweg uit de koorden die de punten $(x_i, f(x_i))$ van de grafiek van f op $[a,b]$ met elkaar verbinden. Bovenstaande resultaten voor $q = 1$ geven dan een bovengrens voor $\|f - \pi_{1,h} f\|_\infty$ wanneer $f \in C^2[a,b]$. De daarin voorkomende factor $\frac{1}{2!}$ kan in feite nog iets verbeterd worden tot $\frac{1}{8}$, door de term $\|(x-a)(x-b)\|_\infty$ in de basisuitdrukking (3.5.1) niet ruw af te schatten op $(b-a)^2$, maar op $\frac{1}{4}(b-a)^2$ (waarom is dat trouwens zo?). Dit werkt dan door in de rest van de ongelijkheden en we vinden:

Stelling 3.5.1. *Laat $f \in C^2[a,b]$, en zij $a = x_0 < x_1 < \ldots < x_{m+1} = b$ een partitie van $[a,b]$. Dan geldt voor de bijbehorende continue stuksgewijs lineaire approximant $\pi_{1,h} f \in V_h^1[a,b]$:*

$$\|f - \pi_{1,h} f\|_\infty \leq \frac{1}{8} \|h^2 f^{(2)}\|_\infty. \tag{3.5.8}$$

De ruimte V_h^1 is eindigdimensionaal: we noemen (zonder bewijs) dat de dimensie van V_h^q gelijk is aan $mq + q + 1$. Voor het geval $q = 1$ kunnen we expliciet een basis aangeven als in de volgende stelling.

Stelling 3.5.2. *Laat $a = x_0 < x_1 < \ldots < x_{m=1} = b$ een partitie van $[a,b]$ zijn. Dan is $\{\phi_0, \phi_1, \ldots, \phi_{m+1}\}$ een basis van $V_h^1[a,b]$, wanneer deze functies gegeven worden door:*

$$i = 0: \quad \phi_0(x) = \begin{cases} \frac{x - x_1}{x_0 - x_1} & x \in [x_0, x_1]; \\ 0 & x \notin [x_0, x_1]; \end{cases}$$

$$i = 1, \ldots, m: \quad \phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}} & x \in [x_{i-1}, x_i]; \\ \frac{x - x_{i+1}}{x_i - x_{i+1}} & x \in [x_i, x_{i+1}]; \\ 0 & x \notin [x_{i-1}, x_{i+1}]; \end{cases}$$

$$i = m+1: \quad \phi_{m+1}(x) = \begin{cases} \frac{x - x_m}{x_{m+1} - x_m} & x \in [x_m, x_{m+1}]; \\ 0 & x \notin [x_m, x_{m+1}]. \end{cases}$$

Er geldt $\phi_i(x_j) = \delta_{i,j}$ voor $i, j \in \{0, 1, \ldots, m+1\}$. Voor $v \in V_h^1[a,b]$ is $v = \sum_{i=0}^{m+1} v(x_i)\phi_i$. Indien $f \in C[a,b]$, dan is $\pi_{1,h} f = \sum_{i=0}^{m+1} f(x_i)\phi_i$.

Teken zelf een plaatje voor ieder van de drie gevallen (doen!). De functies in deze basis worden om voor de hand liggende redenen wel *dakfuncties* (Engels: "hat functions") genoemd. Er zijn m hele en twee halve daken.

Bewijs. De eigenschap m.b.t. de Kronecker delta volgt uit de definitie van de ϕ_i. De lineaire onafhankelijkheid is dan onmiddellijk duidelijk, vgl. het bewijs bij Lagrange-polynomen. Merk op, dat elementen in $V_h^1[a,b]$ worden vastgelegd door hun waarden in de x_i. Aangezien $\sum_{i=0}^{m+1} v(x_i)\phi_i$ en v in de punten van de partitie inderdaad overeenstemmen, zijn ze dus gelijk en spannen de ϕ_i blijkbaar $V_h^1[a,b]$ op. De uitdrukking voor $\pi_{1,h}f$ is nu ook duidelijk, omdat per constructie $\pi_{1,h}f(x_i) = f(x_i)$. □

3.6 Adaptieve methoden voor stuksgewijze polynomiale approximatie

In de vorige paragraaf hebben we, gegeven een partitie $a = x_0 < x_1 < \ldots < x_{m+1} = b$, $q \geq 1$, en een keuze van interpolatiepunten $x_i = \xi_0^{(i)} < \xi_1^{(i)} < \ldots < \xi_q^{(i)} = x_{i+1}$ in ieder subinterval, voor $f \in C[a,b]$ een continue interpolant $\pi_{q,h}f \in V_{q,h}$ geconstrueerd, die stuksgewijs polynomiaal van graad ten hoogste q is. Voor $f \in C^{q+1}[a,b]$ hebben we hiervoor de foutschatting

$$\|f - \pi_{q,h}\|_\infty \leq \frac{1}{(q+1)!}\|h^{q+1}f^{(q+1)}\|_\infty \tag{3.6.1}$$

afgeleid, in termen van de maasfunctie h. Voor $q = 1$ kan dit, zoals we al opmerkten, nog iets verbeterd worden tot

$$\|f - \pi_{1,h}\|_\infty \leq \frac{1}{8}\|h^2 f^{(2)}\|_\infty \quad (f \in C^2[a,b]). \tag{3.6.2}$$

In dit geval $q = 1$ is $\pi_{1,h}f$ een continue stuksgewijs lineaire interpolant, die volkomen wordt vastgelegd door de waarden van f in de interpolatiepunten x_i. De grafiek ervan wordt gevonden door simpelweg de opeenvolgende punten van de grafiek van f, die corresponderen met opeenvolgende punten van de partitie, met lijnstukjes te verbinden.

We willen nu onze partitie zó gaan kiezen, dat een ons bekende functie f door de bijhorende interpolant $\pi_{q,h}f$ in de norm $\|\,.\,\|_\infty$ (d.w.z. uniform) zeker niet slechter dan TOL benaderd wordt, waarbij TOL > 0 een voorgegeven tolerantie is. We willen dit dan concluderen op grond van het rechterlid in (3.6.1). M.a.w.: we zoeken een maasfunctie h, zodanig dat

$$\frac{1}{(q+1)!}\|h^{q+1}f^{(q+1)}\|_\infty \leq \text{TOL}. \tag{3.6.3}$$

Er zijn i.h.a. twee methoden om dit te doen, die we nu behandelen. We nemen hierbij $q = 1$ (een in de praktijk veel voorkomende situatie); voor hogere q kan het analoog. Voor $q = 1$ gebruiken we de verscherping (3.6.2) van (3.6.1) om te komen tot de eis

$$\frac{1}{8}\|h^2 f^{(2)}\|_\infty \leq \text{TOL}, \tag{3.6.4}$$

en we stellen ons tot doel een partitie met bijbehorende maasfunctie h te bepalen die hieraan voldoet. We willen dit doen met liefst zo weinig mogelijk partitiepunten, om vervolgberekeningen, die met deze partitiepunten werken, zo eenvoudig (d.w.z. snel) mogelijk te houden.

Eerste methode: equidistante punten ("globale schatting"). Bij deze—wellicht het meest voor de hand liggende—methode kiezen we de partitiepunten equidistant. De maasfunctie h neemt dan de waarden 0 en $\frac{b-a}{m+1}$ aan, zodat (3.6.4) overgaat in

$$\frac{1}{8}\left(\frac{b-a}{m+1}\right)^2 \|f^{(2)}\|_\infty \leq \text{TOL}. \tag{3.6.5}$$

Hieruit kan men dan aflezen wat m bij deze benadering minimaal moet zijn, en dus ook wat de minimale waarde voor het aantal partitiepunten $(m+2)$ is. Merk op, dat in (3.6.5) het *globale* maximum (d.w.z. het maximum op *heel* $[a,b]$) van $|f^{(2)}|$ voorkomt.

Voorbeeld 3.6.1. Beschouw $f(x) = \exp(-x^2)$ op $[-10,10]$. Men gaat na dat $f''(x) = 2(2x^2 - 1)\exp(-x^2)$, en $f'''(x) = -4x(2x^2-3)\exp(-x^2)$. Teken zelf het plaatje voor f'': deze functie heeft nulpunten in $\pm\frac{1}{\sqrt{2}}$, een globaal minimum -2 in $x=0$, lokale maxima in $x = \pm\sqrt{\frac{3}{2}}$ ter grootte $\frac{4}{e\sqrt{e}}$, en lokale minima in $x = \pm 10$ ter grootte $398e^{-100}$. Men vindt m nu uit de eis $\frac{100}{(m+1^2)} \leq \text{TOL}$. Merk op, dat $\|f\|_\infty = 2$, maar dat $|f''|$ voorbij de extrema in $\pm\sqrt{\frac{3}{2}}$ zéér snel afvalt.

Het voordeel van deze methode is, dat hij eenvoudig is. Het nadeel is, dat hij i.h.a. leidt tot meer punten in de partitie dan er eigenlijk nodig zijn. Dat laatste is niet onmiddellijk duidelijk, maar zal blijken uit de alternatieve tweede methode hieronder.

Tweede methode: adaptieve maaswijdte ("lokale schattingen"). Merk op, dat de eis in (3.6.4) een puntsgewijs karakter heeft: waar $|f^{(2)}|$ klein is, kan h groot zijn, en waar $|f^{(2)}|$ groot is, moet h klein zijn. In (3.6.5) is deze wisselwerking tussen $|f^{(2)}|$ en h niet meer terug te vinden. Men zou zelfs kunnen zeggen, dat aan deze vergelijking in feite een worst-case scenario in (3.6.4) ten grondslag ligt, waarin $|f^{(2)}|$ overal gelijk is aan zijn maximumwaarde. Immers, als dat laatste zo zou zijn, dan gaat (3.6.4) over in $\frac{1}{8}\|h\|_\infty^2\|f\|_\infty \leq \text{TOL}$. Zoekend naar een partitie met zo weinig mogelijk punten, zullen we hieraan dan willen voldoen met h identiek gelijk aan een zo groot mogelijke constante (afgezien van de waarde 0 in de partitiepunten). Bepaling van deze constante geeft dan een bijbehorend (i.h.a. niet geheeltallig) aantal partitiepunten, dat na afronding juist de minimale m geeft, zoals die met (3.6.5) eveneens(!) wordt bepaald.

Dit is uiteraard een te pessimistische benadering, want $f^{(2)}$ is i.h.a. niet constant. De tweede methode buit dit dan ook uit, door de partitie aan te passen (vandaar de toevoeging "adaptief") aan de *lokale* situatie m.b.t. $|f^{(2)}|$. Men realiseert zich hierbij dat (3.6.4) equivalent is met te eisen dat (in voor de hand liggende notatie) voor ieder i-de interval $[x_{i-1}, x_i]$ moet gelden, dat

$$\frac{1}{8}(x_i - x_{i-1})^2 \|f^{(2)}\|_{[x_{i-1},x_i],\infty} \leq \text{TOL}. \tag{3.6.6}$$

Even aannemend dat x_{i-1} al bekend is, zien we het linkerlid in bovenstaande vergelijking als functie van x_i, voor $x_i \in [x_{i-1}, \infty)$. Voor $x_i = x_{i-1}$ is deze functie 0; verder is duidelijk dat deze functie monotoon niet-dalend is. Laat nu x_i van x_{i-1} tot b toenemen, totdat in bovenstaande ongelijkheid gelijkheid wordt bereikt, en leg op dat moment het partitiepunt x_i vast. (Als er nooit gelijkheid wordt bereikt, dan kiezen we $x_i = b$ en zijn we klaar). Nadat x_i eenmaal bepaald is, bepaalt men analoog x_{i+1}, etc. Aldus wordt de partitie recursief opgebouwd. Deze methode verdisconteert ten duidelijkste het lokale gedrag van $|f^{(2)}|$: in een gebied waar deze functie klein is, worden op deze manier lange intervallen gegenereerd—en dat geeft daar dus weinig extra partitiepunten. Korte intervallen, d.w.z. veel partitiepunten, worden slechts ingevoegd in gebieden waar dat blijkbaar nodig is, d.w.z. waar $|f^{(2)}|$ groot is.

Het zal overigens in het algemeen een hele exercitie zijn om dit stramien op deze ideale manier uit te voeren. In Voorbeeld 3.6.1 leidt het op deze manier bepalen van $x_1 \simeq -10$ bijv. tot de vergelijking $\frac{1}{8}(x_1+10)^2 f''(x_1) = \text{TOL}$. Hierbij gebruiken we dan, dat $|f''|$ stijgend is op $[-10, x_1]$, zolang $x_1 \simeq 10$ (wat bij voldoend kleine TOL inderdaad zo zal moeten zijn). Deze vergelijking voor x_1 is niet exact oplosbaar, en men zal methoden als Newton-Raphson moeten gebruiken om dit bij benadering op te lossen. Vervolgens vindt men dan x_2 etc. op dezelfde manier, totdat de rij partitiepunten $-\sqrt{\frac{3}{2}}$ passeert. Op de intervallen daarna neemt $|f''|$ zijn maximum juist in de beginpunten aan, wat dan een andere (overigens dan eenvoudige) vergelijking voor de nieuwe partitiepunten geeft. De situatie klapt dan weer om bij passage van $-\frac{1}{\sqrt{2}}$, etc. Bij de exacte passage van de omslagpunten moet men zelfs weer iets gedetailleerder kijken. Uiteindelijk heeft men dan echter wel een geschikte partitie geconstrueerd die alleen rond $x = 0$ (d.w.z. rond die waarden van x waar $|f(x)| = \|f\|_{[-10,10],\infty}$) een maaswijdte heeft die (vrijwel) gelijk is aan de maaswijdte die men vindt met de equidistante methode. In andere gebieden, en met name in de staarten (waar $|f^{(2)}|$ immers uiterst klein is), is de maaswijdte zeer veel groter dan men met de uniforme methode vindt.

Het voordeel van deze "ideale" manier van vaststellen van de partitie is uiteraard het kleinere aantal partitiepunten dan met de equidistante methode wordt gevonden. Als evident nadeel valt het rekenintensieve karakter op, en in de praktijk zal men een dergelijke gedetailleerde partitievorming daarom lang niet altijd willen uitvoeren. Sterker nog: vaak is dit zelfs simpelweg onuitvoerbaar, omdat gedetailleerde kennis over $|f^{(2)}|$ ontbreekt of de analyse daarvan te complex wordt. Wel wordt uit bovenstaande lokale beschouwing duidelijk, *dat het de nauwkeurigheid verbetert om, gegeven een aantal te verdelen partitiepunten, relatief veel punten te plaatsen in gebieden waar $|f^{(2)}|$ relatief groot is (d.w.z. de afgeleide relatief snel verandert, d.w.z. de functie relatief snel fluctueert), en (dus) relatief weinig punten in gebieden waar $|f^{(2)}|$ relatief klein is (d.w.z. de afgeleide relatief langzaam verandert, d.w.z. de functie relatief langzaam fluctueert).* Met deze observatie kan men dan ook minder gedetailleerde informatie over $|f^{(2)}|$ toch nog benutten of, wat belangrijker is, kan men in het algemeen "met timmermansoog" ook *zonder* gedetailleerd rekenwerk al een behoorlijke winst in nauwkeurigheid boeken, ten opzichte van de eerste methode van equidistributie van het aantal te verdelen partitiepunten.

Terzijde 3.6.2. Bovenstaande aanpak is een illustratie van het principe van "equidistributie van fout". Dit is een empirische vuistregel, die inhoudt dat het, bij het vaststellen van partities, i.h.a. gunstig is om ervoor te zorgen dat de fout op ieder subinterval (ongeveer) *gelijk* is. Hoe groot die fout dan mag zijn, wordt dan bepaald door de toegestane tolerantie. In onze situatie, waar de fout wordt gemeten met behulp van $\|\,.\,\|_\infty$, houdt dit principe dus in dat we ernaar moeten streven dat $\frac{1}{8}(x_i - x_{i-1})^2 \|f^{(2)}\|_{[x_{i-1},x_i],\infty} \simeq$ TOL, zoals we gedaan hebben. Wanneer de fout bijvoorbeeld gemeten zou worden m.b.v. een integraal, dan zou de praktische uitwerking van het principe inhouden om de partitie zo in te richten, dat de integraal over ieder subinterval (ongeveer) gelijk is aan $\frac{\text{TOL}}{m+1}$, waarbij $m+1$ het aantal subintervallen is.

Hoofdstuk 4

Numerieke integratie en extrapolatie

4.1 Inleiding en overzicht

We stellen ons in deze sectie tot doel, om praktische methoden te ontwikkelen waarmee we, voor gegeven $f \in C[a,b]$, de integraal $\int_a^b f(x)\,dx$ willekeurig dicht kunnen benaderen. De theorie van de Riemann-integraal vertelt ons, dat we dit altijd kunnen doen m.b.v. Riemann-sommen, maar dat is i.h.a. een rekenintensieve en langzaam convergerende methode. Onder milde gladheidseisen op f zullen we aanmerkelijk betere methoden ontwikkelen.

De strategie die we uiteindelijk zullen kiezen is de volgende. We zullen een rij $\{f_n\}_{n=1}^{\infty}$ construeren (in feite kunnen we zelfs kiezen uit meerdere rijen), zodanig dat:

- $\|f_n - f\|_\infty \to 0$ als $n \to \infty$, en
- de integralen $\int_a^b f_n(x)\,dx$ exact zijn uit te rekenen, en zelfs op een eenvoudige manier.

Aannemend, dat we zo'n rij hebben gevonden, zien we dat (vgl. een soortgelijk argument aan het einde van Paragraaf A.1)

$$\begin{aligned} |\int_a^b f_n(x)\,dx - \int_a^b f(x)\,dx| = |\int_a^b f_n(x) - f(x)\,dx| &\le \int_a^b |f_n(x) - f(x)|\,dx \\ &\le \int_a^b \|f_n - f\|_\infty\,dx = (b-a)\|f_n - f\|_\infty \to 0, \end{aligned}$$

m.a.w. de ons bekende rij $\{\int_a^b f_n(x)\,dx\}_{n=1}^{\infty}$ van integralen van de benaderende functies levert de gezochte benaderingen van de integraal $\int_a^b f(x)\,dx$.

De constructie zal overigens zodanig zijn, dat de rij $\{f_n\}_{n=1}^{\infty} \subset C[a,b]$ uiteindelijk zelfs uit beeld verdwijnt: wat er overblijft is een eenvoudig recept voor het berekenen van benaderingen van de integraal. Die benaderingen worden geparametriseerd door een stapgrootte h: bij het nemen van de limiet $h \to 0$ verkrijgt men dan de gezochte waarde van de integraal.

In een laatste stap zullen we, voor één specifiek recept, de convergentie voor $h \to 0$ naar de gezochte limietwaarde nog aanzienlijk versnellen, als toepassing van een algemene extrapolatie-techniek.

Tenslotte zullen we aangeven hoe de methodes voor eendimensionale integralen gebruikt kunnen worden als bouwstenen van methodes voor meerdimensionale integralen.

4.2 Kwadratuurregels: algemeen

Het wat verouderde woord "kwadratuur" betekent "het berekenen van oppervlakte". Het wordt binnen de wiskunde nog gebruikt in de specifieke context van numerieke integratie, en i.h.b. spreekt men daar van "kwadratuurregels". De volgende propositie dient o.a. ter motivatie van de definitie van een kwadratuurregel, zoals we die in deze paragraaf zullen geven.

Propositie 4.2.1. *Laat een vaste $q \in \{0, 1, 2, \ldots\}$ en een vaste keuze van $q+1$ punten $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$ gegeven zijn. Dan zijn er unieke $c_0, c_1, \ldots, c_q \in \mathbb{R}$, zodanig dat*

$$\int_a^b p(x)\,dx = \sum_{i=0}^{q} c_i p(\xi_i) \tag{4.2.1}$$

voor alle $p \in P^q[a, b]$. Deze uniek vastgelegde c_i worden gegeven door

$$c_i = \int_a^b \lambda_i(x)\,dx \quad (i = 0, 1, \ldots, q), \tag{4.2.2}$$

in termen van de Lagrange-basis $\{\lambda_i\}_{i=0}^q$ van $P^q[a, b]$, behorend bij de punten $\xi_0, \xi_1, \ldots, \xi_1$.

Bewijs. Uniciteit en waarde. Toepassing van (4.2.1) op λ_j, $(j = 0, 1, \ldots, q)$ geeft, gebruikmakend van $\lambda_j(x_i) = \delta_{i,j}$, dat (4.2.2) geldt.
Existentie. Schrijf, op basis van de interpolatieformule van Lagrange, $p = \sum_{i=0}^q p(\xi_i)\lambda_i$. Integratie geeft

$$\int_a^b p(x)\,dx = \sum_{i=0}^{q} \left\{ \int_a^b \lambda_i(x)\,dx \right\} p(\xi_i),$$

□

Terzijde 4.2.2. Een andere manier—vanuit de lineaire algebra—om hier tegenaan te kijken is de volgende. In Terzijde 3.2.1 hadden we al voor $i = 0, 1, \ldots, q$ een evaluatie-afbeelding $\mathrm{ev}_i : P^q[a, b] \to \mathbb{R}$ gedefinieerd door $\mathrm{ev}_i(p) = p(\xi_i)$, en we hadden opgemerkt dat $\{\mathrm{ev}_0, \mathrm{ev}_1, \ldots, \mathrm{ev}_q\}$ een basis van de duale vectorruimte van $P^q[a, b]$ is, in dualiteit met de Lagrange-basis $\{\lambda_0, \lambda_1, \ldots, \lambda_q\}$ van $P^q[a, b]$. Het is duidelijk dat de afbeelding $\mathrm{Int} : P^q[a, b] \to \mathbb{R}$, gegeven door $\mathrm{Int}(p) = \int_a^b p(x)\,dx$, eveneens een element van de duale is. Blijkbaar is $\mathrm{Int} = \sum_{i=0}^q c_i\,\mathrm{ev}_i$, voor uniek bepaalde c_i. Toepassing op de Lagrange-basis geeft dan de waarden van de c_i.

Definitie 4.2.3. Een *kwadratuurregel* op $[a, b,]$ is het gebruiken, bij vaste $q \geq 0$, vaste punten $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$, en vaste getallen $c_0, c_1, \ldots, c_q$, van

$$\widetilde{I}(f) \stackrel{\text{def.}}{=} \sum_{i=0}^{q} c_i f(\xi_i) \tag{4.2.3}$$

als benadering voor de integraal

$$I(f) \stackrel{\text{def.}}{=} \int_a^b f(x)\,dx,$$

voor *willekeurige* $f \in C[a, b]$.

De punten ξ_i heten de *kwadratuurpunten* (of, kortweg, de *punten*) van de regel, de c_i heten de *kwadratuurgewichten* (of, kortweg, de *gewichten*) van de regel

Zowel $I(f)$ als $\widetilde{I}(f)$ hangen lineair van f af.

Het is duidelijk dat er geen kwadratuurregel bestaat die het juiste resultaat geeft voor *alle* continue f (waarom trouwens?). Als een maatstaf voor het realiteitsgehalte van een regel hanteert men daarom wel het begrip "precisie", dat gerelateerd is aan het wél berekenen van de juiste waarde van de integraal voor een deelruimte van de polynomen.

Definitie 4.2.4. Een regel heeft *precisie minstens* q, als $\widetilde{I}(p) = I(p)$ voor alle $p \in P^q[a,b]$ (d.w.z. alle polynomen van graad ten hoogste q worden door de regel exact geïntegreerd). Een regel heeft *precisie* q, als de regel precisie minstens q heeft, maar er een $p \in P^{q+1}[a,b]$ bestaat, z.d.d. $\widetilde{I}(p) \neq I(p)$.

Een regel, die zelfs de constanten niet exact integreert, heeft geen precisie.

Opmerking 4.2.5. Blijkens Propositie 4.2.1 bestaat er, gegeven $q+1$ punten $a \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq b$ in $[a,b]$, precies één kwadratuurregel op $[a,b]$ met precisie minstens q, die de voorgegeven punten gebruikt. Dit is de regel

$$\widetilde{I}(f) \stackrel{\text{def.}}{=} I(\pi_q f) \left(= \int_a^b \sum_{i=0}^{q} f(\xi_i)\lambda_i(x)\,dx = \sum_{i=0}^{q} \left\{ \int_a^b \lambda_i(x)\,dx \right\} f(\xi_i) \right).$$

Immers, als $\widetilde{I}(f) = \sum_{i=0}^{q} c_i f(\xi_i)$ de gezochte regel is, dan legt de eis van precisie minstens q, d.w.z. de eis (4.2.1), blijkens de Propositie de gewichten vast als in (4.2.2). Blijkbaar is dan

$$\widetilde{I}(f) = \sum_{i=0}^{q} \left\{ \int_a^b \lambda_i(x)\,dx \right\} f(\xi_i) = \int_a^b \sum_{i=0}^{q} f(\xi_i)\lambda_i(x)\,dx = \int_a^b \pi_q f\,dx = I(\pi_q f)$$

Voorbeeld 4.2.6. Neem $q = 0$. Voor de kwadratuurregel van precisie minstens 0 op $[a,b]$, met kwadratuurpunt $\xi_0 \in [a,b]$, berekent men $c_0 = \int_a^b \lambda_0(x)\,dx = \int_a^b 1\,dx = b-a$. Dit gewicht hangt blijkbaar niet van ξ_0 af, wat overigens niet de algemene situatie weerspiegelt. De (uiteraard wel altijd van ξ_0 afhankelijke) regel luidt dus

$$\widetilde{I}(f) = (b-a)f(\xi_0).$$

Teken zelf het plaatje—het is ook duidelijk dat het gewicht $b-a$ moet zijn om constanten exact te integreren.

Is het mogelijk dat een regel met één kwadratuurpunt een grotere precisie dan 0 heeft? Het gewicht moet dan in ieder geval $b-a$ zijn, op grond van het voorgaande. Dit aannemend is, vanwege de lineariteit van I en $\widetilde{I}$, de eis van precisie minstens 1 dan equivalent (ga na!) met het exact integreren van de functie x, d.w.z. met

$$(b-a)\xi_0 = \int_a^b x\,dx.$$

Er volgt $\xi_0 = \frac{a+b}{2}$. Het is grafisch trouwens ook duidelijk dat dit de uitkomst moet zijn—teken zelf het plaatje.

Is het mogelijk dat de precisie van zo'n éénpuntsregel zelfs minstens 2 is? Het antwoord is negatief: dan zou, op grond van precisie minstens één, het kwadratuurpunt $\xi_0 = \frac{a+b}{2}$ moeten zijn, met gewicht $b-a$, en men kan nagaan dat x^2 dan door de regel niet exact wordt geïntegreerd. Sneller is echter, om op te merken dat $p(x) = (x-\xi_0)^2$ onmiddellijk laat zien dat een dergelijke regel niet kan bestaan (waarom?).
Conclusie: de regel $\widetilde{I}(f) = (b-a)f(\xi_0)$ heeft precisie 1 voor $\xi_0 = \frac{a+b}{2}$, en 0 anders. Voor alle andere gewichten dan $b-a$ heeft de regel geen precisie, ongeacht de keuze van ξ_0.

Terzijde 4.2.7. Voor alle mogelijke gegeven $q+1$ punten $a \le \xi_0 < \xi_1 < \ldots < \xi_q \le b$ zien we m.b.v. het polynoom $(x-\xi_0)^2(x-\xi_1)^2\cdots(x-\xi_q)^2$ onmiddellijk in, dat een regel met precisie tenminste $2q+2$ met deze punten nooit haalbaar is. Men kan echter bewijzen, dat een precisie van $2q+1$ wél altijd haalbaar is met $q+1$ kwadratuurpunten, maar slechts voor een *unieke* keuze van die punten. De bijbehorende regel staat bekend als een *Gauß-kwadratuurregel*. Voor iedere $q \in \{0,1,2,\ldots,\}$ is er dus een Gauß-kwadratuurregel op $[a,b]$ met precisie $2q+1$. De erin voorkomende gewichten blijken strikt positief te zijn.

Voor $q=0$ is, blijkens dit resultaat, de maximaal haalbare precisie 1, met een door die eis dan uniek bepaald kwadratuurpunt. Dit hadden we in het voorbeeld hierboven inderdaad al geconstateerd, met een rechttoe-rechtaan berekening. Het bewijs voor $q \ge 1$ is echter niet elementair meer, en maakt gebruik van de theorie van zgn. *orthogonale polynomen*. Het is binnen dit kader dan mogelijk om ook meer algemene Gauß-kwadratuurregels te formuleren voor integralen van het type

$$I(f) = \int_a^b f(x)w(x)\,dx \quad (f \in C[a,b]),$$

waarbij w een vaste positieve gewichtsfunctie is. Voor $w=1$ hervinden we ons geval hierboven.

4.3 Enkelvoudige kwadratuurregels

Een *enkelvoudige kwadratuurregel* is een kwadratuurregel als in Propositie 4.2.1, d.w.z. (zie ook Opmerking 4.2.5) een regel van de vorm

$$\widetilde{I}(f) \stackrel{\text{def.}}{=} I(\pi_q f) = \int_a^b \sum_{i=0}^{q} f(\xi_i)\lambda_i(x)\,dx = \sum_{i=0}^{q} \left\{ \int_a^b \lambda_i(x)\,dx \right\} f(\xi_i),$$

voor (vaste) gegeven punten $a \le \xi_0 < \xi_1 < \ldots < \xi_q \le b$. M.a.w.: men gebruikt, als benadering voor de integraal van f, de waarde van de integraal van de interpolant van f in de vast gekozen interpolatiepunten ξ_i. De ξ_i worden dan automatisch de kwadratuurpunten van de regel, waarvan de gewichten vastliggen als integralen van de betreffende Lagrange-polynomen. De regel heeft precisie minstens q.

Opmerking 4.3.1. De toevoeging "enkelvoudig" houdt in, dat we één globale polynomiale interpolant van f op het hele interval $[a,b]$ kiezen, en die integreren. Later zullen we het interval opdelen en op ieder van de subintervallen een enkelvoudige kwadratuurregel toepassen. Op die manier verkrijgen we dan meervoudige (ook wel: samengestelde) kwadratuurregels, die erop neerkomen dat de integraal wordt berekend van een stuksgewijs polynomiale interpolant. Deze regels worden ook nog wel uitgebreide regels genoemd, waarbij dan dus gedacht wordt vanuit een (sub)interval dat uitgebreid wordt.

We behandelen nu een aantal bekende enkelvoudige regels. Voor de bijbehorende foutschattingen zullen we o.a. gebruikmaken van het resultaat in Gevolg 3.4.4:

$$|f(x) - \pi_q f(x)| \le \frac{1}{(q+1)!}|(x-\xi_0)(x-\xi_1)\ldots(x-\xi_q)|\,\|f^{(q+1)}\|_\infty \quad (f \in C^{q+1}[a,b]).$$

Voorbeeld 4.3.2. $q=0$.
De regel is hier (zie de vorige paragraaf)

$$\widetilde{I}(f) = I(\pi_0 f) = (b-a)f(\xi_0),$$

waarbij $\xi_0 \in [a,b]$ het kwadratuurpunt is. Deze regel heet (om grafische redenen) de *rechthoeksregel*. Voor het speciale geval $\xi_0 = \frac{a+b}{2}$ is de term *middelpuntsregel* in omloop. We leiden een foutschatting af, voor $f \in C^1[a,b]$.

- Voor algemene $\xi_0 \in [a,b]$ is

$$\begin{aligned}|\widetilde{I}(f) - I(f)| &= |I(f) - I(\pi_0 f)| \\ &= |\int_a^b f(x) - \pi_0 f(x)\, dx| \\ &\leq \int_a^b |f(x) - \pi_0 f(x)|\, dx \\ &\leq \int_a^b \frac{1}{1!}|x - \xi_0|\, \|f'\|_\infty\, dx \\ &\leq \frac{1}{2}\{(\xi_0 - a)^2 + (b - \xi_0)^2\}\, \|f'\|_\infty \\ &\leq \frac{1}{2}(b-a)^2 \|f'\|_\infty.\end{aligned}$$

  Teken zelf het plaatje en zie de oppervlakteberekening in, die aan de voorlaatste stap ten grondslag ligt.

  Conclusie:
  Voor de rechthoeksregel

$$\widetilde{I}(f) = (b-a) f(\xi_0) \quad (\xi_0 \in [a,b])$$

  geldt de foutschatting

$$|\widetilde{I}(f) - I(f)| \leq \frac{1}{2}(b-a)^2 \|f'\|_\infty \quad (f \in C^1[a,b]). \tag{4.3.1}$$

- De middelpuntsregel heeft precisie 1, maar dit is niet manifest uit bovenstaande foutschatting. Dat zou het wel zijn, indien de tweede afgeleide er in voorkwam. Een dergelijke foutschatting bestaat inderdaad, en we leiden die nu af, voor $f \in C^2[a,b]$.

  Volgens Taylor is er voor $x \in [a,b]$ een $\tau_x \in [a,b]$, zodanig dat

$$f(x) = f(\frac{a+b}{2}) + f'(\frac{a+b}{2})(x - \frac{a+b}{2}) + \frac{1}{2} f''(\tau_x)(x - \frac{a+b}{2})^2.$$

  We zien daarmee dat

$$\begin{aligned}|\widetilde{I}_f - I_f| &= |I(\pi_0 f) - I(f)| \\ &= |\int_a^b f(\frac{a+b}{2})\, dx - \int_a^b f(x)\, dx| \\ &= |\int_a^b f'(\frac{a+b}{2})(x - \frac{a+b}{2}) + \frac{1}{2} f''(\tau_x)(x - \frac{a+b}{2})^2\, dx|.\end{aligned}$$

  De eerste term in de integraal geeft 0, dus we zien dat

$$|\widetilde{I}_f - I_f| \leq \frac{1}{2}\int_a^b (x - \frac{a+b}{2})^2\, dx\, \|f''\|_\infty = \frac{1}{24}(b-a)^3 \|f''\|_\infty.$$

  Conclusie:
  Voor de middelpuntsregel

$$\widetilde{I}(f) = (b-a) f(\frac{a+b}{2})$$

geldt de foutschatting

$$|\widetilde{I}(f) - I(f)| \leq \frac{1}{24}(b-a)^3 \|f''\|_\infty \quad (f \in C^2[a,b]). \tag{4.3.2}$$

Men kan, door de middelpuntsregel iets preciezer te analyseren, laten zien dat er een $\tau \in [a,b]$ bestaat, z.d.d.

$$\widetilde{I}(f) - I(f) = -\frac{1}{24}(b-a)^3 f''(\tau) \quad (f \in C^2[a,b]). \tag{4.3.3}$$

Dit impliceert de door ons afgeleide bovengrens voor de fout. Het geeft ook preciezere resultaten, zoals we later in een numeriek voorbeeld zullen zien. Het geeft ook aan, dat de middelpuntsregel de integraal van een functie f *altijd zal onderschatten, zodra* $f'' > 0$ *op* $[a,b]$. Dit is grafisch ook duidelijk (ga na!).

Voorbeeld 4.3.3. q=1.
Hier kan men de twee kwadratuurpunten ξ_0 en ξ_1 beide variëren. We beperken ons tot het geval $\xi_0 = a$, $\xi_1 = b$. De bijbehorende gewichten zijn dan beide gelijk aan $\frac{b-a}{2}$, zoals men eenvoudig vindt door berekening van de integralen van de Lagrange-polynomen. Als geheugensteuntje voor deze waarden van de gewichten kan men denken aan de situatie waarin $f(a), f(b) \geq 0$. Voor een lineaire functie f (die door de regel exact geïntegreerd wordt) is dan de integraal gelijk aan de oppervlakte tussen de grafiek van f en de x-as. Deze oppervlakte—de oppervlakte van een trapezium—is $(b-a) \cdot \frac{f(a)+f(b)}{2}$. Dit verklaart de naam van de *trapeziumregel*

$$\widetilde{I}(f) = I(\pi_1 f) = \frac{b-a}{2}(f(a) + f(b)).$$

Analoog aan het geval $q = 0$ leidt men af, dat

$$\begin{aligned} |\widetilde{I}(f) - I(f)| &= |\int_a^b \frac{1}{2!}(x-a)(x-b)f''(\tau_x)\,dx| \\ &\leq \frac{1}{2}\int_a^b |(x-a)(x-b)|\,\|f''\|_\infty\,dx \\ &= \frac{1}{12}(b-a)^3\|f''\|_\infty. \end{aligned}$$

Conclusie:
Voor de trapeziumregel

$$\widetilde{I}(f) = \frac{b-a}{2}(f(a) + f(b))$$

geldt de foutschatting

$$|\widetilde{I}(f) - I(f)| \leq \frac{1}{12}(b-a)^3 \|f''\|_\infty \quad (f \in C^2[a,b]). \tag{4.3.4}$$

Men kan, door de trapeziumregel iets preciezer te analyseren, laten zien dat er een $\tau \in [a,b]$ bestaat, z.d.d.

$$\widetilde{I}(f) - I(f) = \frac{1}{12}(b-a)^3 f''(\tau) \quad (f \in C^2[a,b]). \tag{4.3.5}$$

Dit impliceert de door ons afgeleide bovengrens voor de fout. Net als bij de middelpuntsregel geeft ook dit preciezere resultaten, zoals we later in een numeriek voorbeeld zullen zien. Het geeft ook aan, dat de trapeziumregel de integraal van een functie f *altijd zal overschatten, zodra* $f'' > 0$ *op* $[a,b]$. Dit is grafisch ook duidelijk (ga na!).

Opmerking 4.3.4. De bovengrens voor de trapeziumregel is dus *slechter* dan die voor de middelpuntsregel, hoewel er meer punten gebruikt worden.

Terzijde 4.3.5. De trapeziumregel heeft precisie 1. De Gauß-kwadratuurregel voor $q = 1$, met de maximale precisie 3, heeft als kwadratuurpunten $\frac{a+b}{2} \pm \frac{\sqrt{3}}{6}(b-a)$, met beide gewichten gelijk aan $\frac{b-a}{2}$.

Voorbeeld 4.3.6. q=2.
We kiezen hier a, $\frac{a+b}{2}$ en b als de drie kwadratuurpunten. Berekening van de betreffende gewichten geeft de zgn. *Simpson-regel*

$$\widetilde{I}(f) = I(\pi_2 f) = \frac{b-a}{6}\left(f(a) + 4f(\frac{a+b}{2}) + f(b)\right)$$

Hiervoor geldt de foutschatting

$$|\widetilde{I}(f) - I(f)| \leq \frac{(b-a)^5}{2880}\|f^{(4)}\|_\infty \quad (f \in C^4[a,b]). \tag{4.3.6}$$

We zullen dit niet bewijzen.

Meer precies is er een $\tau \in [a,b]$, z.d.d.

$$\widetilde{I}(f) - I(f) = \frac{(b-a)^5}{2880} f^{(4)}(\tau) \quad (f \in C^4[a,b]). \tag{4.3.7}$$

Opmerking 4.3.7. Op grond van de definitie $\widetilde{I}(f) = I(\pi_2 f)$ mag men van de Simpson-regel een precisie van minstens 2 verwachten. De precisie is in feite 3. Terzijde: de Gauß-kwadratuurregel voor $q = 2$, waarvan we de punten en gewichten kortheidshalve niet zullen geven, heeft de maximale precisie 5.

Zo voortgaand, zouden we bijvoorbeeld het algemene geval kunnen beschouwen, waarbij we het interval $[a,b]$ in subintervallen van gelijke lengte opdelen, en vervolgens de enkelvoudige kwadratuurregel analyseren die alle zo ontstane randpunten van de intervallen als kwadratuurpunten gebruikt. Deze kwadratuurregels zijn inderdaad bestudeerd, en heten Newton–Cotes-regels, of, meer precies, gesloten Newton–Cotes-regels, waarbij "gesloten" aanduidt dat de punten a en b zelf ook kwadratuurpunt zijn. De trapeziumregel en de Simpsonregel zijn gesloten Newton–Cotes-regels. Bij de open Newton–Cotes-regels gebruikt men a en b niet als kwadratuurpunten, maar de overige ontstane randpunten wél. De middelpuntsregel is hier een (in feite: het eenvoudigste) voorbeeld van. Voor deze Newton–Cotes-regels bestaan algemene foutschatttingen. Helaas bevatten deze schattingen termen als $\|f^{(n)}\|_\infty$, waarbij n dan (ongeveer) het aantal kwadratuurpunten is. Dit geeft aan, dat men—analoog aan de situatie bij polynomiale approximatie—niet mag verwachten, dat verhogen van het aantal equidistante kwadratuurpunten automatisch een convergerend proces oplevert om integralen te berekenen. Als praktisch bezwaar bij het verhogen van het aantal punten geldt verder nog, dat er geen eenvoudige uitdrukkingen voor de gewichten bekend zijn. Men kan deze wel weer bepalen, maar dat betekent weer meer rekenwerk geeft.

We zullen dus, om integralen te benaderen, een andere weg moeten inslaan dan het verhogen van het aantal punten in de enkelvoudige kwadratuurregels.

We sluiten deze paragraaf af met een voorbeeld van de toepassing van drie enkelvoudige kwadratuurregels op hetzelfde probleem, met foutschatting.

Voorbeeld 4.3.8. We nemen $f(x) = \exp(-x^2)$, $a = 0$ en $b = 1$, en zoeken een benadering van $I(f) = \int_0^1 \exp(-x^2)$. Er geldt $-2 \leq f''(x) \leq 1$ en $-20 \leq f^{(4)}(x) \leq 12$ voor $0 \leq x \leq 1$, zodat $\|f''\|_\infty \leq 2$ en $\|f^{(4)}\|_\infty \leq 20$.

1. De *middelpuntsregel geeft*:

$$\tilde{I}(f) = (1-0)e^{-\left(\frac{1}{2}\right)^2} = 0.7788008\cdots .$$

Volgens de grove afschatting (4.3.2) hebben we dan

$$|\tilde{I}(f) - I(f)| \leq \frac{(1-0)^3}{24} \cdot 2 = 0.0833333\cdots ,$$

zodat

$$\tilde{I}(f) \in [0.695, 0.863].$$

De preciezere uitdrukking (4.3.3) geeft:

$$-0.0416666\cdots = -\frac{(1-0)^3}{24} \cdot 1 \leq \tilde{I}(f) - I(f) \leq -\frac{(1-0)^3}{24} \cdot (-2) = 0.0833333\cdots ,$$

zodat

$$\tilde{I}(f) \in [0.737, 0.863].$$

2. De *trapeziumregel geeft*:

$$\tilde{I}(f) = \frac{(1-0)}{2}\left\{e^{-0^2} + e^{-1^2}\right\} = 0.6839397\cdots .$$

Volgens de grove afschatting (4.3.4) hebben we dan

$$|\tilde{I}(f) - I(f)| \leq \frac{(1-0)^3}{12} \cdot 2 = 0.1666666\cdots ,$$

zodat

$$\tilde{I}(f) \in [0.516, 0.851].$$

De preciezere uitdrukking (4.3.5) geeft:

$$-0.1666666\cdots = \frac{(1-0)^3}{12} \cdot (-2) \leq \tilde{I}(f) - I(f) \leq \frac{(1-0)^3}{12} \cdot 1 = 0.0833333\cdots ,$$

zodat

$$\tilde{I}(f) \in [0.599, 0.851].$$

3. De *Simpson-regel geeft*:

$$\tilde{I}(f) = \frac{(1-0)}{6}\left\{e^{-0^2} + 4e^{-\left(\frac{1}{2}\right)^2} + e^{-1^2}\right\} = 0.7471804\cdots .$$

Volgens de grove afschatting (4.3.6) hebben we dan

$$|\tilde{I}(f) - I(f)| \leq \frac{(1-0)^5}{2880} \cdot 20 = 0.0069444\cdots ,$$

zodat

$$\tilde{I}(f) \in [0.740, 0.755].$$

De preciezere uitdrukking (4.3.7) geeft:

$$-0.0069444\cdots = \frac{(1-0)^5}{2880}\cdot(-20) \le \tilde{I}(f) - I(f) \le \frac{(1-0)^5}{2880}\cdot 12 = 0.0041667\cdots,$$

zodat

$$\tilde{I}(f) \in [0.743, 0.755].$$

Opmerking 4.3.9. We zien in Voorbeeld 4.3.8 de praktische betekenis van het verschil tussen grovere afschattingen als (4.3.2) en preciezere uitspraaken als (4.3.3). In bijv. het geval van de middelpuntsregel heeft het toepassen van (4.3.2) tot gevolg, dat een interval met breedte $\frac{1}{12}(b-a)^3\cdot 2\|f''\|_\infty$ wordt opgegeven, terwijl toepassing van (4.3.3) leidt tot een interval met breedte $\frac{1}{12}(b-a)^3\cdot(\max_{x\in[a,b]} f''(x) - \min_{x\in[a,b]} f''(x))$. Dit kan een aanzienlijk smaller interval zijn: denk maar aan een voorbeeld waarin f'' op $[a,b]$ weinig varieert en uitsluitend zeer grote positieve waarden aanneemt.

4.4 Samengestelde kwadratuurregels

Zoals in de vorige paragraaf al is opgemerkt, worden benaderingen van integralen niet noodzakelijkerwijs beter, door het aantal kwadratuurpunten in een enkelvoudige regel te verhogen. Net als bij polynomiale approximatie zullen we een andere weg kiezen, en het interval opdelen in subintervallen. Op ieder van deze subintervallen afzonderlijk passen we vervolgens een enkelvoudige kwadratuurregel toe. We sommeren daarna de resultaten van deze deelberekeningen en verklaren de som tot benadering van de integraal over het hele interval. Een dergelijke procedure heet een *samengestelde* kwadratuurregel. De precisie van een dergelijke regel is i.h.a. gelijk aan de minimale precisie van de afzonderlijke enkelvoudige regels (d.w.z. (veel) kleiner dan het aantal kwadratuurpunten).

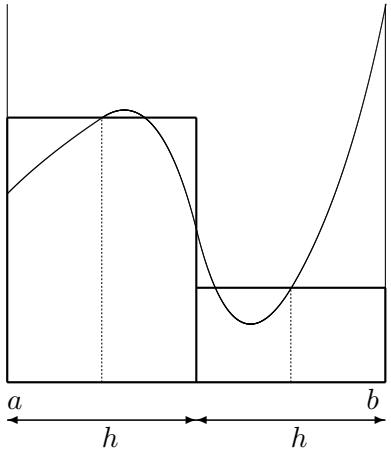
Men kan hierbij natuurlijk de opdeling variëren en/of de enkelvoudige kwadratuurregel van het subinterval laten afhangen. Op die manier ontstaat een grote variatie van samengestelde kwadratuurregels. Overzichtelijker, gemakkelijker te behandelen (zowel in de theorie als in computer-implementaties) en nauwkeurig genoeg is het, wanneer de subintervallen alle even lang zijn, en de enkelvoudige regel voor ieder subinterval dezelfde is. Men spreekt dan over *meervoudige* kwadratuurregels, of ook over (van een klein naar een groot interval) *uitgebreide* regels. Met "samengestelde" kwadratuurregels wordt overigens in de literatuur soms ook alleen deze (verreweg de belangrijkste) "uniforme" situatie aangeduid. De terminologie is niet geheel uniform, maar het is uit de context altijd duidelijk wat er bedoeld wordt.

We behandelen een tweetal samengestelde regels.

Samengestelde middelpuntsregel. Men kiest hierbij een equidistante partitie $a = x_0 < x_1 < \ldots < x_{m+1} = b$ van $[a,b]$. De lengte van ieder subinterval is $h \overset{\text{def.}}{=} \frac{b-a}{m+1}$. Volgens de samengestelde middelpuntsregel nemen we dan

$$M_h(f) = \sum_{i=0}^{m} \tilde{I}_{[x_i,x_{i+1}]}(f) = \sum_{i=0}^{m} (x_{i+1}-x_i) f(\frac{x_i+x_{i+1}}{2}) = h\sum_{i=0}^{m} f(\frac{x_i+x_{i+1}}{2})$$

als benadering voor $I(f) = \int_a^b f(x)\,dx$. Zie de figuur hieronder voor $m = 1$.



Samengestelde middelpuntsregel (voor twee subintervallen)

Voor de foutschatting baseren we ons op (4.3.2). Gebruikmakend van $(m+1)h = (b-a)$, zien we voor $f \in C^2[a,b]$, in voor de hand liggende notatie:

$$\begin{aligned}
|M_h(f) - I(f) &= |\sum_{i=0}^{m} \tilde{I}_{[x_i,x_{i+1}]}(f) - I_{[x_i,x_{i+1}]}(f)| \\
&\leq \sum_{i=0}^{m} |\tilde{I}_{[x_i,x_{i+1}]}(f) - I_{[x_i,x_{i+1}]}(f)| \\
&\leq \sum_{i=0}^{m} \frac{(x_{i+1}-x_i)^3}{24} \|f''\|_{[x_i,x_{i+1}],\infty} \\
&\leq \sum_{i=0}^{m} \frac{h^3}{24} \|f''\|_{[a,b],\infty} \\
&= \frac{(b-a)}{24} \|f''\|_\infty \cdot h^2.
\end{aligned}$$

Conclusie:
Voor de samengestelde middelpuntsregel

$$M_h(f) = h \sum_{i=0}^{m} f(\frac{x_i + x_{i+1}}{2}) \tag{4.4.1}$$

geldt de foutschatting

$$|M_h(f) - I(f)| \leq \frac{(b-a)}{24} \|f''\|_\infty \cdot h^2 \quad (f \in C^2[a,b]). \tag{4.4.2}$$

In feite is er een $\tau \in [a,b]$, zodanig dat

$$M_h(f) - I(f) = -\frac{(b-a)}{24} f''(\tau) \cdot h^2 \quad (f \in C^2[a,b]), \tag{4.4.3}$$

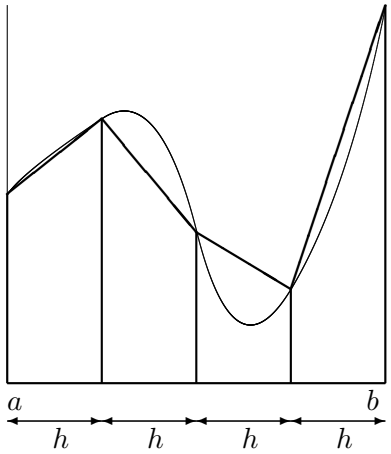
wat, net als bij enkelvoudige regels, i.h.a. nauwkeurigere uitspraken geeft.

Samengestelde trapeziumregel. Men kiest ook hierbij een equidistante partitie $a = x_0 < x_1 < \ldots < x_{m+1} = b$ van $[a,b]$. De lengte van ieder subinterval is weer $h \stackrel{\text{def.}}{=} \frac{b-a}{m+1}$. Volgens de

samengestelde trapeziumregel nemen we dan

$$\begin{aligned} T_h(f) =& \sum_{i=0}^{m} \tilde{I}_{[x_i,x_{i+1}]}(f) \\ =& \frac{h}{2}\{f(a)+f(x_1)\} + \frac{h}{2}\{f(x_1)+f(x_2)\} + \cdots \\ & \cdots + \frac{h}{2}\{f(x_{m-1})+f(x_m)\} + \frac{h}{2}\{f(x_m)+f(b)\} \\ =& h\left\{\frac{1}{2}f(a) + \sum_{i=1}^{m} f(x_i) + \frac{1}{2}f(b)\right\} \end{aligned}$$

als benadering voor $I(f) = \int_a^b f(x)\,dx$. Zie de figuur hieronder voor $m = 3$.



Samengestelde trapeziumregel (voor vier subintervallen)

Analoog aan de afleiding van de foutschatting voor de samengestelde middelpuntsregel vindt men:
Voor de samengestelde trapeziumregel

$$T_h(f) = h\left\{\frac{1}{2}f(a) + \sum_{i=1}^{m} f(x_i) + \frac{1}{2}f(b)\right\} \tag{4.4.4}$$

geldt de foutschatting

$$|T_h(f) - I(f)| \leq \frac{(b-a)}{12}\|f''\|_\infty \cdot h^2 \quad (f \in C^2[a,b]). \tag{4.4.5}$$

In feite is er een $\tau \in [a,b]$, zodanig dat

$$T_h(f) - I(f) = \frac{(b-a)}{12} f''(\tau) \cdot h^2 \quad (f \in C^2[a,b]), \tag{4.4.6}$$

Opmerking 4.4.1. De samengestelde kwadratuurregels vormen de uitwerking van de strategie, zoals genoemd in de inleiding van dit hoofdstuk, om f uniform te benaderen met functies waarvan de integraal eenvoudig en exact is uit te rekenen. Zo correspondeert de samengestelde middelpuntsregel met stuksgewijs constante functies als benaderende functies, en correspondeert de samengestelde trapeziumregel met continue stuksgewijs lineaire interpolanten als benaderende

functies. De samengestelde Simpsonregel zou corresponderen met continue stuksgewijs kwadratische functies als benaderende functies. Omdat de integralen van dergelijke benaderende functies receptmatig uit te rekenen zijn, geeft dit dus ook een eenvoudig recept voor de benaderingen van de integraal van f. De polynomiale approximatietheorie is uit het uiteindelijke recept daarmee verdwenen, maar vormt dus wel de onderliggende basis ervoor. Ook de foutschattingen voor de samengestelde regels komen, zoals we gezien hebben, uit deze theorie voort.

4.5 De sommatieformule van Euler–Maclaurin en asymptotiek voor de samengestelde trapeziumregel

In de vorige paragraaf behandelden we o.a. de foutschatting van de samengestelde trapeziumregel voor een C^2-functie f. Naarmate de functie gladder is, kan men, zo blijkt, deze foutschatting voor $T_h(f)$ verder preciseren m.b.v. de zgn. sommatieformule van Euler–Maclaurin. We zullen deze formule in deze paragraaf afleiden. Uit deze formule volgt op zijn beurt dan weer een asymptotische ontwikkeling voor $T_h(f)$ rond $h = 0$. Deze asymptotische ontwikkeling tenslotte zal dan, samen met extrapolatietechnieken voor dat soort ontwikkelingen, in de volgende paragraaf leiden tot een i.h.a. bijzonder efficiënte manier van numerieke integratie, de zgn. Romberg-integratie.

Om deze Romberg-integratie is het ons te doen, en we zullen daarvoor de nodige voorbereidingen moeten treffen. In de huidige paragraaf behandelen we daarom de formule van Euler–Maclaurin en de daarbij behorende asymptotische ontwikkeling voor de samengestelde trapeziumregel. In de volgende paragraaf passen we een extrapolatie-techniek toe die ons dan deze Romberg-integratie oplevert.

Het bewijs van de sommatieformule van Euler–Maclaurin, dat hieronder volgt, lijkt nogal technisch, maar in feite is het niet veel meer dan op een geschikte manier herhaaldelijk partiële integratie toepassen—en nauwkeurig de boeken bijhouden. We beginnen met een hulpresultaat.

Lemma 4.5.1. *Er is een unieke rij polynomen $\{p_n\}_{n=1}^{\infty}$ zodanig dat*

$$\begin{cases} p_1(x) & = x; \\ p'_{n+1}(x) & = p_n(x) \quad (n \geq 1); \\ p_{2l+1}(1) & = p_{2l+1}(-1) = 0 \quad (l \geq 1). \end{cases}$$

Bewijs. De rij wordt, uitgaande van p_1, in stappen van twee opgebouwd. We bekijken eerst de rij tot en met de derde term. Het is duidelijk, dat er moet gelden dat $p_2(x) = \frac{1}{2}x^2 + c$ voor zekere $c \in \mathbb{R}$ en dus $p_3(x) = \frac{1}{6}x^3 + cx + d$ voor zekere $d \in \mathbb{R}$. De eis $p_3(1) = p_3(-1) = 0$ leidt dan tot $c + d =$ `iets bekends` en $-c + d =$ `iets anders bekends`, zodat we inzien dat c en d inderdaad geschikt gekozen kunnen worden (en op een unieke manier), zodanig dat p_1, p_2 en p_3 voldoen aan $p_1(x) = x$, $p'_{n+1}(x) = p_n(x)\,(1 \leq n \leq 2)$ en $p_3(\pm 1) = 0$. Dit is het begin van een inductief proces, als volgt. Neem aan dat voor zekere $l \geq 1$ de polynomen $p_1, p_2, \ldots, p_{2l+1}$ al uniek geconstrueerd zijn, zodanig dat $p_1(x) = x$, $p'_{n+1}(x) = p_n(x)\,(1 \leq n \leq 2l)$, $p_3(\pm 1) = p_5(\pm 1) = \cdots = p_{2l+1}(\pm 1) = 0$. (Het geval $l = 1$ hebben we zojuist afgehandeld.) Het is dan duidelijk, dat moet gelden dat $p_{2l+2}(x) =$ `een bekend polynoom` $+ c$ voor zekere $c \in \mathbb{R}$ en dus $p_{2l+3}(x) =$ `een ander bekend polynoom` $+ cx + d$ voor zekere $d \in \mathbb{R}$. De eis $p_{2l+3}(1) = p_{2l+3}(-1) = 0$ legt dan c en d uniek vast, waarmee de rij twee stappen verder is opgebouwd, d.w.z. voor $l + 1$ i.p.v. voor l. Zo voortgaande definieert men op een unieke manier p_n voor alle n. □

Het is uit het bewijs verder duidelijk dat iedere p_n van graad n is en rationale coëfficiënten heeft.

Lemma 4.5.2. *Voor de rij $\{p_n\}_{n=1}^{\infty}$ uit Lemma 4.5.1 geldt, dat p_n een even functie is voor even n. Voor oneven n is p_n een oneven functie.*

Bewijs. Laat $q_n(x) = (-1)^n p_n(-x)$ voor $n \geq 1$. De q_n zijn polynomen, waarvoor geldt:

- $q_1(x) = -1 \cdot -x = x$;
- $q'_{n+1}(x) = (-1)^{n+1} \cdot -1 \cdot p'_{n+1}(-x) = (-1)^n p_n(-x) = q_n(x) \quad (n \geq 1)$;
- $q_{2l+1}(\pm 1) = (-1)^{2l+1} p_{2l+1}(\mp 1) = 0 \quad (l \geq 1)$.

Met andere woorden: de rij $\{q_n\}_{n=1}^{\infty}$ voldoet aan de definiërende eigenschappen in Lemma 4.5.1. Volgens de uniciteitsuitspraak in dat Lemma geldt dan blijkbaar dat $p_n = q_n$ voor alle $n \geq 1$, en dit is juist wat we moesten bewijzen. □

In het bijzonder is dus $p_{2l}(1) = p_{2l}(-1)$ voor $l \geq 1$; we zullen zien dat dit later zal zorgen voor een telescoopsom. We noteren

$$c_{2l} \overset{\text{def.}}{=} p_{2l}(1) = p_{2l}(-1) \quad (l \geq 1).$$

Deze c_{2l} komen terug in de uiteindelijke sommatieformule. Een eenvoudige gesloten uitdrukking hiervoor is niet bekend[1], maar het is duidelijk uit de constructie van de polynomen p_n dat c_{2l} voor iedere l een expliciet berekenbaar rationaal getal is.

De polynomen p_n gaan we gebruiken in een herhaalde partiële integratie, als volgt. Laat $\psi : [-1, 1] \mapsto \mathbb{R}$ een gegeven functie zijn, die "voldoende glad" is—de precieze benodigde conditie hiervoor zal volgen uit de afleiding. Gebruikmakend van het verdwijnen van randtermen als gevolg van $p_{2l+1}(\pm 1) = 0\,(l \geq 1)$, zien we dan dat

$$\begin{aligned}
&\int_{-1}^{1} \psi(x)\,dx = \\
&= \int_{-1}^{1} p'_1(x)\psi(x)\,dx \\
&= p_1(x)\psi(x)|_{-1}^{1} - \int_{-1}^{1} p_1(x)\psi^{(1)}(x)\,dx \\
&= \psi(1) + \psi(-1) - \int_{-1}^{1} p'_2(x)\psi^{(1)}(x)\,dx \\
&= \psi(1) + \psi(-1) - p_2(x)\psi^{(1)}(x)|_{-1}^{1} + \int_{-1}^{1} p_2(x)\psi^{(2)}(x)\,dx \\
&= \psi(1) + \psi(-1) - c_2\left\{\psi^{(1)}(1) - \psi^{(1)}(-1)\right\} + \int_{-1}^{1} p'_3(x)\psi^{(2)}(x)\,dx \\
&\overset{*}{=} \psi(1) + \psi(-1) - c_2\left\{\psi^{(1)}(1) - \psi^{(1)}(-1)\right\} + p_3(x)\psi^{(2)}(x)|_{-1}^{1} - \int_{-1}^{1} p_3(x)\psi^{(3)}(x)\,dx \\
&= \psi(1) + \psi(-1) - c_2\left\{\psi^{(1)}(1) - \psi^{(1)}(-1)\right\} - \int_{-1}^{1} p'_4(x)\psi^{(3)}(x)\,dx \\
&= \psi(1) + \psi(-1) - c_2\left\{\psi^{(1)}(1) - \psi^{(1)}(-1)\right\} - p_4(x)\psi^{(3)}(x)|_{-1}^{1} + \int_{-1}^{1} p_4(x)\psi^{(4)}(x)\,dx
\end{aligned}$$

[1] de c_{2l} zijn gerelateerd aan de *Bernouilli-getallen*

$$
\begin{aligned}
=&\psi(1)+\psi(-1)-c_2\left\{\psi^{(1)}(1)-\psi^{(1)}(-1)\right\}-c_4\left\{\psi^{(3)}(1)-\psi^{(3)}(-1)\right\}\\
&+\int_{-1}^{1}p_5'(x)\psi^{(4)}(x)\,dx\\
\stackrel{*}{=}&\psi(1)+\psi(-1)-c_2\left\{\psi^{(1)}(1)-\psi^{(1)}(-1)\right\}-c_4\left\{\psi^{(3)}(1)-\psi^{(3)}(-1)\right\}\\
&+p_5(x)\psi^{(4)}(x)|_{-1}^{1}-\int_{-1}^{1}p_5(x)\psi^{(5)}(x)\,dx.
\end{aligned}
$$

De rechterleden van de vergelijkingen met $\stackrel{*}{=}$ hebben dezelfde structuur. Er is blijkbaar in deze keten van partiële integraties een periodiciteit met periode 4, met dien verstande, dat het aantal stappen dat gezet kan worden uiteraard afhangt van de mate van differentieerbaarheid van ψ.

Wanneer we eenvoudigheidshalve $-c_2$ als γ_{2l} noteren, dan zien we als conclusie dat er constanten $\gamma_2,\gamma_4,\gamma_6,\ldots$ bestaan, zodanig dat voor alle $k\in\{0,1,2,\ldots\}$ en $\psi\in C^{2k+1}[-1,1]$ geldt:

$$
\begin{aligned}
\int_{-1}^{1}\psi(x)\,dx=&\psi(1)+\psi(-1)+\sum_{l=1}^{k}\gamma_{2l}\left\{\psi^{(2l-1)}(1)-\psi^{(2l-1)}(-1)\right\}\\
&-\int_{-1}^{1}p_{2k+1}(x)\psi^{(2k+1)}(x)\,dx.
\end{aligned}
\tag{4.5.1}
$$

Indien $\psi\in C^{2k+2}[-1,1]$, d.w.z. als ψ nog tenminste eenmaal extra continu differentieerbaar is, dan kunnen we de restterm nog herschrijven als

$$
\begin{aligned}
&-\int_{-1}^{1}p_{2k+1}(x)\psi^{(2k+1)}(x)\,dx=\\
&=-\int_{-1}^{1}\left(p_{2k+2}(x)-c_{2k+2}\right)'\psi^{(2k+1)}(x)\,dx\\
&=-\left(p_{2k+2}(x)-c_{2k+2}\right)\psi^{(2k+1)}(x)|_{-1}^{1}+\int_{-1}^{1}\left(p_{2k+2}(x)-c_{2k+2}\right)\psi^{(2k+2)}(x)\,dx\\
&=\int_{-1}^{1}\left(p_{2k+2}(x)+\gamma_{2k+2}\right)\psi^{(2k+2)}(x)\,dx.
\end{aligned}
\tag{4.5.2}
$$

De eerste twee termen samen in het rechterlid van (4.5.1) herkennen we al als de trapeziumregel voor ψ op het interval $[-1,1]$. Om te komen tot een vergelijkbaar resultaat voor de *samengestelde* trapeziumregel moeten we deze formule op ieder van de betreffende subintervallen gaan toepassen. Laat dus $a=x_0<x_1<\ldots<x_{m+1}=b$ een equidistante partitie van $[a,b]$ zijn, met stapgrootte $h=\frac{b-a}{m+1}$. Zij $f\in C^{(2k+2)}[a,b]$. Fixeer $i\in\{0,1,\ldots,m\}$ en definieer $\psi_i:[-1,1]\mapsto\mathbb{R}$ door

$$
\psi_i(x)=f(x_i+\frac{h}{2}(x+1))\quad x\in[x_i,x_{i+1}].
$$

Dan is $\psi_i\in C^{(2k+2)}[-1,1]$, $\psi_i(-1)=x_i$ en $\psi_i(1)=x_{i+1}$. Meer in het algemeen is $\psi_i^{(n)}(x)=\left(\frac{h}{2}\right)^n f^{(n)}(x_i+\frac{h}{2}(x+1))$ en i.h.b. is dus $\psi_i^{(n)}(-1)=\left(\frac{h}{2}\right)^n f^{(n)}(x_i)$ en $\psi_i^{(n)}(1)=\left(\frac{h}{2}\right)^n f^{(n)}(x_{i+1})$.

Volgens (4.5.1) en (4.5.2), toegepast op ψ_i, is dan

$$\begin{aligned}\int_{-1}^{1} f(x_i + \frac{h}{2}(x+1))\,dx = f(x_i) + f(x_{i+1}) \\ + \sum_{l=1}^{k} \gamma_{2l} \left(\frac{h}{2}\right)^{2l-1} \left\{ f^{(2l-1)}(x_{i+1}) - f^{(2l-1)}(x_i) \right\} \\ + \int_{-1}^{1} (p_{2k+2}(x) + \gamma_{2k+2}) \left(\frac{h}{2}\right)^{2k+2} f^{(2k+2)}(x_i + \frac{h}{2}(x+1))\,dx.\end{aligned}$$

Een overgang op een nieuwe integratievariabele $t \in [x_i, x_{i+1}]$, via de substitutie $t = x_i + \frac{h}{2}(x+1)$, levert op dat

$$\begin{aligned}\int_{x_i}^{x_{i+1}} f(t)\,dt =& \frac{h}{2}\left\{f(x_i) + f(x_{i+1})\right\} + \sum_{l=1}^{k} \gamma_{2l} \left(\frac{h}{2}\right)^{2l} \left\{ f^{(2l-1)}(x_{i+1}) - f^{(2l-1)}(x_i) \right\} \\ &+ \left(\frac{h}{2}\right)^{2k+2} \int_{x_i}^{x_{i+1}} \left\{ p_{2k+2}(\frac{2}{h}(t-x_i) - 1) + \gamma_{2k+2} \right\} f^{(2k+2)}(t)\,dt.\end{aligned} \tag{4.5.3}$$

Om de restterm onder controle te brengen definiëren we de van h afhankelijke continue functie $R_{2k+2,h} : [a,b] \mapsto \mathbb{R}$ als $R_{2k+2,h}(x) = p_{2k+2}(\frac{2}{h}(x - x_i) - 1) + \gamma_{2k+2}$ voor $x \in [x_i, x_{i+1}]$. Dan is $\|R_{2k+2,h}\|_{[a,b],\infty} \leq \|p_{2k+2}\|_{[-1,1]} + \gamma_{2k+2}$, onafhankelijk van h. Tenslotte sommeren we (4.5.3) over de verschillende subintervallen, genummerd door $i = 0, 1, \ldots, m$. Er onstaat dan o.a. de volgende dubbelsommatie, die door een telescoopwerking vereenvoudigt:

$$\begin{aligned}&\sum_{i=0}^{m} \left\{ \sum_{l=1}^{k} \gamma_{2l} \left(\frac{h}{2}\right)^{2l} \left\{ f^{(2l-1)}(x_{i+1}) - f^{(2l-1)}(x_i) \right\} \right\} = \\ &\sum_{l=1}^{k} \left\{ \sum_{i=0}^{m} \gamma_{2l} \left(\frac{h}{2}\right)^{2l} \left\{ f^{(2l-1)}(x_{i+1}) - f^{(2l-1)}(x_i) \right\} \right\} = \\ &\sum_{l=1}^{k} \gamma_{2l} \left(\frac{h}{2}\right)^{2l} \left\{ f^{(2l-1)}(b) - f^{(2l-1)}(a) \right\}.\end{aligned}$$

De sommatie over de subintervallen levert dus uiteindelijk het volgende resultaat:

Stelling 4.5.3 (Sommatieformule van Euler–Maclaurin). *Er zijn universele, expliciet berekenbare en rationale constanten $\gamma_2, \gamma_4, \gamma_6, \ldots$, zodanig dat voor ieder interval $[a,b]$, voor alle $k \in \{0, 1, 2, \ldots\}$, en voor alle $f \in C^{2k+2}[a,b]$ geldt:*

$$\begin{aligned}\int_a^b f(x)\,dx =& T_h(f) + \sum_{l=1}^{k} \gamma_{2l} \left(\frac{h}{2}\right)^{2l} \left\{ f^{(2l-1)}(b) - f^{(2l-1)}(a) \right\} \\ &+ \left(\frac{h}{2}\right)^{2k+2} \int_a^b R_{2k+2,h}(t) f^{(2k+2)}(t)\,dt,\end{aligned} \tag{4.5.4}$$

waarin $R_{2k+2,h} : [a,b] \mapsto \mathbb{R}$ een van h afhankelijke continue functie op $[a,b]$ is, die in absolute waarde begrensd is door een constante die niet van de stapgrootte h, het interval $[a,b]$ of de functie f afhangt, maar uitsluitend van k.

Opmerking 4.5.4. Men kan laten zien, dat de restterm in feite geschreven kan worden als

$$\gamma_{2k+2}\left(\frac{h}{2}\right)^{2k+2}(b-a)\,f^{(2k+2)}(\tau)$$

voor zekere $\tau \in [a,b]$. Een korte berekening leert verder dat $\gamma_2 = -\frac{1}{3}$, zodat de met deze uitdrukking voor de restterm aangescherpte Stelling 4.5.3 voor $k = 0$ dan juist het eerdere resultaat (4.4.6) geeft.

Opmerking 4.5.5. Wanneer we $f(x) = x^s$ nemen voor $s \in \{0,1,2,\ldots\}$ en $a = 1$, $b = n \in \mathbb{N}$ kiezen, dan volgt uit Stelling 4.5.3 met $2k+2 \geq s+1$, dat er rationale constanten $\alpha_0, \alpha_1, \ldots, \alpha_s$ zijn, zodanig dat $\sum_{i=1}^{n} i^s = \frac{1}{s+1}n^{s+1} + \sum_{j=0}^{s}\alpha_j n^j$ voor alle $n \in \mathbb{N}$. (Waarom?) Er zijn overigens ook wel eenvoudigere manieren om dit in te zien.

Als onmiddellijk gevolg van Stelling 4.5.3 verkrijgen we:

Stelling 4.5.6 (Asymptotiek voor de samengestelde trapeziumregel). *Laat $f \in C^{2k+2}[a,b]$ voor zekere $k \in \{0,1,2,\ldots\}$. Dan zijn er van f en het interval $[a,b]$ afhankelijke constanten $\tau_2, \tau_4, \tau_6 \ldots, \tau_{2k}$, zodanig dat*

$$T_h(f) = \int_a^b f(x)\,dx + \sum_{l=1}^{k}\tau_{2l}h^{2l} + O(h^{2k+2}), \tag{4.5.5}$$

voor alle $h \in \{\frac{b-a}{1}, \frac{b-a}{2}, \frac{b-a}{3}, \ldots\}$. Hierin staat de notatie $O(h^{2k+2})$ voor een functie $R(h)$, gedefinieerd voor de toegelaten waarden van h, die voldoet aan $|R(h)| \leq C|h|^{2k+2}$ voor zekere $C \geq 0$ en alle toegelaten waarden van h.

Bovenstaande stelling bevat veel minder informatie dan de Euler–Maclaurin sommatieformule in Stelling 4.5.3. Verrassend genoeg is echter het simpele feit, dat er voor de samengestelde trapeziumregel een dergelijke asymptotische ontwikkeling *bestaat*, al voldoende informatie om uiteindelijk mee verder te werken. De precieze waarde van de coëfficiënten in die ontwikkeling is i.h.a. minder relevant, zo zal in het vervolg duidelijk worden.

Wat voor toepassingen wel van belang is, is het al dan niet 0 zijn van coëfficiënten in de ontwikkeling. Hiervoor geldt:

Propositie 4.5.7. *In de situatie als in Stelling 4.5.6 geldt, voor $1 \leq l \leq k$, dat*

$$\tau_{2l} = 0 \iff f^{(2l-1)}(a) = f^{(2l-1)}(b).$$

Bewijs. Men kan laten zien dat $\gamma_{2l} \neq 0$ voor alle $l = 1, 2, \ldots$. De Propositie volgt dus onmiddellijk uit Stelling 4.5.3. □

In extreme mate treedt het verdwijnen van de coëfficiënten op voor een oneindig vaak differentieerbare functie $f : \mathbb{R} \mapsto \mathbb{R}$ die *periodiek* is met periode $b-a$. Dan is blijkbaar $\tau_{2l} = 0$ voor *alle* $l \geq 1$, zodat $T_h(f) = \int_a^b f(x)\,dx + O(h^{2k+2})$ voor *alle* $k \geq 0$. Blijkbaar convergeert voor dergelijke functies de samengestelde trapeziumregel zeer snel naar de waarde van de integraal, als $h \to 0$.

Opmerking 4.5.8. Het is verleidelijk om te veronderstellen dat $T_h(f)$ zich als een machtreeks laat ontwikkelen voor alle voldoende kleine toegelaten waarden van h. Er zijn echter functies waarvoor de bijbehorende reeks divergent is voor alle $h \neq 0$—net zoals er Taylor-reeksen zijn die divergeren in ieder punt anders dan het basispunt.

4.6 Extrapolatie naar $h = 0$ en Romberg-integratie

De convergentiesnelheid van de samengestelde trapeziumregel wordt, blijkens Stelling 4.5.6, bepaald door de eerste coëfficiënt in het rijtje $\tau_2, \tau_4, \ldots$ die niet 0 is (even aannemend dat zo'n coëfficiënt bestaat). Met een extrapolatietechniek kan deze convergentiesnelheid echter vaak aanmerkelijk worden verhoogd. Het idee erachter laat zich samenvatten als het "elimineren van machten van h". Deze techniek is in een bredere context bruikbaar, en om dat laatste te benadrukken illustreren we het idee aan de hand van iets anders, namelijk numerieke differentiatie.

Voorbeeld 4.6.1. Veronderstel dat $f \in C^\infty(\mathbb{R})$ en dat we, voor zekere vaste $x_0 \in \mathbb{R}$, $f'(x_0)$ willen benaderen op grond van de functiewaarden van f, die we voor alle waarden van het argument exact berekenbaar veronderstellen. Als eerste stap hiertoe kan men bijvoorbeeld de functie $N_1(h)$ bekijken, gedefinieerd door

$$N_1(h) = \frac{f(x_0 + h) - f(x_0)}{h} \quad (h \neq 0).$$

Duidelijk is dat $\lim_{h\to 0} N_1(h) = f'(x_0)$. In feite is, op grond van een Taylor-ontwikkeling,

$$N_1(h) = f'(x_0) + \frac{h}{2} f''(x_0) + O(h^2) \quad (h \neq 0).$$

Indien $f''(x_0) \neq 0$, dan is de convergentie $N_1(h) \to f'(x_0)$ blijkbaar van orde 1. We hoeven hier echter niet in te berusten: er is namelijk een manier om die orde van convergentie te verbeteren. Merk hiertoe eerst op dat voor alle $h \neq 0$ blijkbaar

$$N_1(\frac{h}{2}) = f'(x_0) + \frac{h}{4} f''(x_0) + O(h^2) \quad (h \neq 0).$$

Hiermee kunnen we de eerste orde term in h elimineren. Merk immers op dat

$$2N_1(\frac{h}{2}) - N_1(h) = f'(x_0) + O(h^2) \quad (h \neq 0).$$

Definieer dus, voor $h \neq 0$,

$$N_2(h) = 2N_1(\frac{h}{2}) - N_1(h) \quad \left(= \frac{4f(x_0 + \frac{h}{2}) - 3f(x_0) - f(x_0 + h)}{h} \right).$$

Dan convergeert $N_2(h)$ nog steeds naar $f'(x_0)$, maar de convergentie is nu van orde tenminste 2.

Dit proces kunnen we voortzetten. Een Taylor-ontwikkeling met één extra term laat bijvoorbeeld, bij doorwerking in de analyse hierboven, zien dat er een constante α bestaat, zodanig dat $N_2(h) = f'(x_0) + \alpha h^2 + O(h^3)$ voor $h \neq 0$. Blijkbaar is $2^2 N_2(\frac{h}{2}) - N_2(h) = (2^2 - 1)f'(x_0) + O(h^3)$, dus de definitie

$$N_3(h) = \frac{2^2 N_2(\frac{h}{2}) - N_2(h)}{2^2 - 1}$$

levert een rij op die convergent is naar $f'(x_0)$ van orde minstens 3. Op grond van $N_3(h)$ vindt men dan weer $N_4(h)$ die convergent is naar $f'(x_0)$ van orde minstens 4, enzovoorts.

Trouwens, numerieke differentiatie is gemakkelijker op papier beschreven dan het op een computer daadwerkelijk uitgevoerd kan worden. Wat zal men in dat geval immers vinden als waarde voor $\lim_{h\to 0} N_k(h)$, voor $k = 1, 2, \ldots$?

We zullen dit elimineren van machten van h wat verder onderzoeken, in een algemene context. We stellen ons hierbij voor, dat we geïnteresseerd zijn in de waarde c_0 van een of andere grootheid (hierboven was dat een afgeleide, later zal het een integraal zijn), waarvan we slechts met een of andere numerieke methode een benadering $N_1(h)$ kunnen berekenen, die naar de gezochte waarde c_0 convergeert als $h \to 0$. We weten ook, dat $N_1(h)$ een asymptotische expansie heeft rond $h = 0$, waaruit we een minimale orde van convergentie kunnen aflezen. Vervolgens proberen we de orde van convergentie te verbeteren op de manier zoals we dat hierboven bij numerieke differentiatie al deden. Dze situatie is de achtergrond van het volgende Lemma.

Lemma 4.6.2. *Laat* $N_1(h) = c_0 + \sum_{i=p}^{q} c_i h^i + O(h^{q+1})$ *voor zekere* $q \geq p > 0$ *en* $c_p, \ldots, c_q$. *Laat* $\theta \neq 1$ *en zij*

$$N_2(h) \stackrel{\text{def.}}{=} \frac{\theta^p N_1(h) - N_1(\theta h)}{\theta^p - 1}.$$

Dan is $N_2(h) = c_0 + \sum_{i=p+1}^{q} \tilde{c}_i h^i + O(h^{q+1})$, *waarbij voor* $i = p+1, \ldots, q$ *geldt dat* $\tilde{c}_i = 0 \Longleftrightarrow c_i = 0$.

Bewijs. Dit is duidelijk uit de volgende uitdrukking voor $N_2(h)$, die men rechtstreeks uit de definitie berekent:

$$N_2(h) = c_0 + \sum_{i=p}^{q} c_i \frac{\theta^p - \theta^i}{\theta^p - 1} h^i + O(h^{q+1})$$

□

Merk op dat de laagste orde term h^p, zo die al "echt" (d.w.z. met coëfficiënt ongelijk 0) voorkwam in de ontwikkeling van $N_1(h)$, in de ontwikkeling van $N_2(h)$ blijkbaar *zeker* niet meer staat. Andere machten dan h^p, die er "echt" stonden in de ontwikkeling van $N_1(h)$, blijven daarentegen ook "echt" staan. Met deze constructie wordt blijkbaar een *eventuele* term h^p verwijderd, maar niet meer dan dat. Wanneer dus de term met h^p er "echt" stond, dan—en dat is de waarde van deze constructie—hebben we een nieuwe en expliciet berekenbare benadering $N_2(h)$ geconstrueerd, die nog steeds naar c_0 convergeert, maar met een convergentie van een grotere orde dan bij $N_1(h)$. Als de term h^p er niet stond, maar een term $h^{p'}$ met $p' > p$ stond er wél, dan is de orde van convergentie ongewijzigd.

Op $N_2(h)$ kan men dan het proces ook weer toepassen, leidend tot $N_3(h)$. Zo zet men dit dan voort, voortdurend de orde van convergentie bevorderend, net zo lang als blijkens de asympotische ontwikkeling van $N_1(h)$ zinvol is. Dit proces staat bekend als *Richardson-extrapolatie*[2]. We hebben hierbij dan nog keuzevrijheid voor de parameter $\theta \neq 1$, die men zelfs kan laten variëren voor de verschillende stappen, zonder dat die keuze invloed heeft op de convergentiesnelheid.

Voorbeeld 4.6.3. Neem aan dat $N_1(h) = c_0 + c_2 h^2 + c_3 h^3 + c_5 h^5 + O(h^7)$, met $c_2, c_3, c_5 \neq 0$. We construeren dan vervolgens:

- $$N_2(h) = \frac{\theta^2 N_1(h) - N_1(\theta h)}{\theta^2 - 1}.$$

  Nu is h^2 geëlimineerd, de termen h^3, h^5 en $O(h^7)$ staan er nog en de convergentie is dus van orde 3.

[2]De reden van het gebruik van de term "extrapolatie" bij deze techniek van convergentieversnelling wordt uitgelegd in Terzijde 4.6.7

- $$N_3(h) = \frac{\theta^3 N_2(h) - N_2(\theta h)}{\theta^3 - 1}.$$

  Nu is h^3 geëlimineerd, de termen h^5 en $O(h^7)$ staan er nog en de convergentie is dus van orde 5.

- $$N_4(h) = \frac{\theta^5 N_3(h) - N_3(\theta h)}{\theta^5 - 1}.$$

  Nu is h^5 geëlimineerd, $O(h^7)$ is de enige overgebleven term en de convergentie is dus van orde ≥ 7.

Voorbeeld 4.6.4. Veronderstel, dat we voor de grootheid $N_1(h)$ uit het vorige voorbeeld slechts weten dat $N_1(h) = c_0 + \sum_{i=1}^{6} c_i h^i + O(h^7)$, waarbij het ons onbekend is of de c_i gelijk aan nul zijn of niet, terwijl in werkelijkheid alleen c_2, c_3 en c_5 ongelijk zijn aan 0. We kunnen dan, als we met de ons bekende informatie de convergentie willen bevorderen, niets anders doen dan achtereenvolgens *eventuele* termen h, $h^2, h^3, \ldots$ elimineren. Dit leidt dan tot de constructie van andere geassocieerde grootheden, namelijk als volgt:

- $$\tilde{N}_2(h) = \frac{\theta N_1(h) - N_1(\theta h)}{\theta - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 2, de echte orde is 2.

- $$\tilde{N}_3(h) = \frac{\theta^2 \tilde{N}_2(h) - \tilde{N}_2(\theta h)}{\theta^2 - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 3, de echte orde is 3.

- $$\tilde{N}_4(h) = \frac{\theta^3 \tilde{N}_3(h) - \tilde{N}_3(\theta h)}{\theta^3 - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 4, de echte orde is 5.

- $$\tilde{N}_5(h) = \frac{\theta^4 \tilde{N}_4(h) - \tilde{N}_4(\theta h)}{\theta^4 - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 5, de echte orde is 5.

- $$\tilde{N}_6(h) = \frac{\theta^5 \tilde{N}_5(h) - \tilde{N}_5(\theta h)}{\theta^5 - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 6, de echte orde is ≥ 7.

- $$\tilde{N}_7(h) = \frac{\theta^6 \tilde{N}_6(h) - \tilde{N}_6(\theta h)}{\theta^6 - 1}$$

  Er is gegarandeerd convergentie van orde ≥ 7, de echte orde is ≥ 7.

In het tweede voorbeeld hierboven moest, zoals we konden verwachtten, meer rekenwerk worden verricht om de facto tot dezelfde orden van convergentie te komen: sommige stappen zijn verloren moeite omdat de te elimineren termen er simpelweg niet staan en die stap dan helaas ook niets doet voor de andere termen. Het loont dus inderdaad de moeite om a priori kennis over het niet voorkomen van machten van h te gebruiken.

We passen dit alles nu toe op de asymptotiek voor de samengestelde trapeziumregel, om de convergentie naar de werkelijke waarde van de integraal te verbeteren. Laat dus $f \in C^{2k+2}[a,b]$ voor zekere $k \in \{0,1,2,\ldots\}$. Dan (zie Stelling 4.5.6) zijn er constanten $\tau_2, \tau_4, \tau_6 \ldots, \tau_{2k} \in \mathbb{R}$, zodanig dat

$$T_h(f) = \int_a^b f(x)\,dx + \sum_{l=1}^{k} \tau_{2l} h^{2l} + O(h^{2k+2}),$$

Propositie 4.5.7 geeft ook nog aan wanneer de coëfficiënten in deze ontwikkeling 0 zijn. Het belang daarvan begrijpen we nu in het licht van het bovenstaande: wanneer dit optreedt, dan kan men zich een stap besparen. Het is echter, een uitzondering daargelaten, vrijwel nooit praktisch haalbaar om uit te maken voor welke coëfficiënten dit het geval is. Een uitzondering hierop noemden we al eerder: bij het toepassen van de samengestelde trapeziumregel op periode functies, geïntegreerd over een periode, zijn alle coëfficiënten 0 en hebben we dus geen enkele garantie dat er convergentieversnelling optreedt. Wel weten we op voorhand in alle gevallen dat oneven machten nooit voorkomen, en in ons ontwerp voor algemene functies nemen we dat dus zeker wel mee, om werk te besparen.

We nemen aan dat $k \geq 1$ (anders valt er niets te elimineren) en definiëren $N_1(h) = T_h(f)$. We kiezen $\theta = 2$ (de reden voor deze keuze zal later duidelijk worden). Elimineer eerst h^2 om convergentie van orde ≥ 4 te verkrijgen:

$$N_2(h) = \frac{2^2 N_1(h) - N_1(2h)}{2^2 - 1};$$

vervolgens h^4 om convergentie van orde ≥ 6 te verkrijgen:

$$N_3(h) = \frac{2^4 N_2(h) - N_2(2h)}{2^4 - 1};$$

en zo vervolgens. De algemene omschrijving is, dat, voor $j = 1, \ldots, k$, de term h^{2j} wordt geëlimineerd (zodat convergentie van orde $\geq 2j+2$ wordt verkregen) door te definiëren:

$$N_{j+1}(h) = \frac{4^j N_j(h) - N_{j-1}(2h)}{4^j - 1}$$

Wat hebben we nu nodig om $N_j(h)$ daadwerkelijk te kunnen *berekenen* voor een gegeven h? De definitie hierboven laat zien dat dan $N_1(h), N_1(2h), \ldots, N_1(2^{j-1}h)$ blijkbaar bekend moeten zijn, m.a.w. we moeten $T_h(f), T_{2h}(f), \ldots, T_{2^{j-1}h}(f)$ kennen. De constante 2 in deze uitdrukkingen is de waarde van onze keuze voor θ, en nu blijkt waarom dat een gunstige is. Immers, bij het voor een of andere h_0 berekenen van $T_{\frac{h_0}{2}}(f)$ kan het werk, dat bij het berekenen van $T_{h_0}(f)$ is verricht, opnieuw gebruikt worden (ga na!). Dit soort efficiënte halveringen van de stapgrootte zit door onze keuze van θ nu ingebouwd.

Samenvattend is blijkbaar datgene, wat er in een concrete berekening gedaan moet worden, het toepassen van de samengestelde trapeziumregel op f voor stapgrootten van de vorm $h = \frac{h_0}{2^i}$, voor een of andere vaste h_0 (die we meestal gelijk aan $b-a$ zullen kiezen). Dit geeft ons dan

$N_1(h) = T_h(f)$ en daarvan afgeleid $N_2(h)$, $N_3(h), \ldots$, met dien verstande, dat $N_j(h)$ slechts gedefinieerd is voor h van de speciale vorm $h = \frac{h_0}{2^{j-1+i}}$ voor $i \geq 0$. Dat laatste fenomeen, waarbij de variabele alleen discrete waarden kan aannemen, hebben we niet verdisconteerd in onze bespreking van de extrapolatietechniek, waar we min of meer impliciet aannamen dat h continu kan variëren. De lezer zal zich er echter van overtuigen dat de resultaten over convergentieversnelling daar niet van afhangen en in de huidige situatie gewoon doorgaan.

In feite is onze integratiemethode, de zgn. *Romberg-integratie*, hiermee nu klaar. Hij kan gekarakteriseerd worden als het toepassen van extrapolatietechnieken op de asymptotische ontwikkeling voor de samengestelde trapeziumregel, zoals die volgt uit de sommatieformule van Euler–Maclaurin.

We voeren, afrondend, alleen nog een notatie in die het mechanische karakter van de methode benadrukt en een kant-en-klaar recept oplevert, waarin de achterliggende theorie niet meer zichtbaar is (behalve natuurlijk in de effectiviteit van dat recept!).

Laat $T_{i,0} = N_1(\frac{h_0}{2^i}) \left(= T_{\frac{h_0}{2^i}}(f)\right)$ voor $i \geq 0$ en, meer algemeen, laat

$$T_{i,j} = N_{j+1}(\frac{h_0}{2^i}) \quad (i \geq j),$$

zodat

$$\begin{aligned} T_{i,j} &= \frac{4^j N_j(\frac{h_0}{2^i}) - N_j(\frac{h_0}{2^{i-1}})}{4^j - 1} \\ &= \frac{4^j T_{i,j-1} - T_{i-1,j-1}}{4^j - 1} \quad (i \geq j). \end{aligned}$$

We hebben dan uiteindelijk het volgende schema.

Schema voor Romberg-integratie

h	i	$T_{i,0} = T_h(f)$	$T_{i,1} = \frac{4^1 T_{i,0} - T_{i-1,0}}{4^1 - 1}$	$T_{i,2} = \frac{4^2 T_{i,1} - T_{i-1,1}}{4^2 - 1}$	$T_{i,3} = \frac{4^3 T_{i,2} - T_{i-1,2}}{4^3 - 1}$	$\cdots$
h_0	0	$T_{0,0}$				
$\frac{h_0}{2^1}$	1	$T_{1,0}$	$T_{1,1}$			
$\frac{h_0}{2^2}$	2	$T_{2,0}$	$T_{2,1}$	$T_{2,2}$		
$\frac{h_0}{2^3}$	3	$T_{3,0}$	$T_{3,1}$	$T_{3,2}$	$T_{3,3}$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Hierin geldt:

- *In de meest linkse kolom wordt $T_{i,0}$ berekend met de samengestelde trapeziumregel voor f, d.w.z. als $T_{i,0} = T_{\frac{h_0}{2^i}}(f)$;*

- *De andere kolommen worden voor $j \geq 1$ berekend via*

$$T_{i,j} = \frac{4^j T_{i,j-1} - T_{i-1,j-1}}{4^j - 1} \quad (i \geq j).$$

- *Als $f \in C^{2k+2}[a,b]$ voor $k \in \{0,1,2,\ldots\}$, dan is, voor $j = 0,1,\ldots,k$, de kolom-rij $\{T_{i,j}\}_{i=j}^{\infty}$ convergent naar $\int_a^b f(x)\,dx$ van orde tenminste $2j+2$.*

De theorie, zoals we die ontwikkeld hebben, vertelt ons dat de kolommen "verticaal" convergeren, en dat de gegarandeerde minimale orde van convergentie toeneemt naarmate de kolom meer naar rechts is geplaatst[3]. Men kan echter laten zien dat voor een C^∞-functie de diagonaalrij $\{T_{i,i}\}_{i=0}^{\infty}$ ook naar de integraal convergeert, en dat men i.h.a. daarvan snelle convergentie mag verwachten. Ook deze keuze voor de diagonaalrij als benaderingen wordt vaak onder Romberg-integatie verstaan.

Helaas is het niet goed mogelijk om een praktisch bruikbare afschatting voor de verschillende fouten te geven, zodat men vaak pragmatisch met het schema omgaat wat betreft de keuze van een stopcriterium. Een wel gebruikt criterium is bijv. het nog slechts binnen de verlangde tolerantie verschillen van twee opeenvolgende diagonaalelementen $T_{i,i}$ en $T_{i+1,i+1}$, om dan vervolgens $T_{i+1,i+1}$ als benadering van de integraal te hanteren. Analoog voor het verschil tussen $T_{i,i-1}$ en $T_{i,i}$, d.w.z. de twee laatste elementen in een rij. Absolute zekerheid bieden deze criteria echter niet.

Opmerking 4.6.5.

1. Een getal in het schema wordt, behalve in de meest linkse kolom, gedefinieerd in termen van het getal links ervan en schuin links erboven, zodat na het beschikbaar komen van één nieuwe waarde van de samengestelde trapeziumregel weer een hele rij kan worden toegevoegd.

2. De theorie van de Riemann-integraal leert ons, dat de eerste kolom voor $f \in C[a,b]$ altijd naar $\int_a^b f(x)\,dx$ convergeert. De andere kolommen doen dat dan ook, zoals dat met inductie onmiddellijk volgt uit de recursieve definitie. Het is echter slechts voor *gladde* functies dat men zeker weet dat schema convergentieversnellend werkt.

Voorbeeld 4.6.6. Beschouw $I(f) = \int_1^2 \frac{1}{x}\,dx = \log 2 = 0.6931472$, in 7 cijfers precisie (de precisie waarin we in dit voorbeeld zullen werken). We vullen het schema in tot en met de zesde rij (ga als oefening na dat dit klopt):

h	i	$T_{i,0} = T_h(f)$	$T_{i,1}$	$T_{i,2}$	$T_{i,3}$	$T_{i,4}$
1	0	0.7500000				
$\frac{1}{2}$	1	0.7083333	0.6944444			
$\frac{1}{4}$	2	0.6970238	0.6932540	0.6931747		
$\frac{1}{8}$	3	0.6941219	0.6931546	0.6931480	0.6931476	
$\frac{1}{16}$	4	0.6933912	0.6931476	0.6931472	0.6931472	0.6931472

[3] In de praktijk blijkt ook het meer naar rechts gaan in een rij vaak betere benaderingen te geven.

Het tweede gesuggereerde stopcriterium geeft hier inderdaad de integraal in 7 decimalen nauwkeurig, gebaseerd op slechts 17 functie-evaluaties. Merk trouwens op dat $T_{3,0}$ (direct uit de samengestelde trapeziumregel, gebaseerd op 9 functiewaarden) nog een ca. 35 keer zo grote fout heeft als $T_{2,2}$ (geëxtrapoleerd, gebaseerd op slechts 5 functiewaarden). Er zijn nog wel extremere voorbeelden bekend: Romberg-integratie geldt als een zeer efficiënte manier van numeriek integreren, die met weinig functie-evaluaties een zeer precies resultaat bereikt.

Terzijde 4.6.7. Waarom staat de techniek van convergentieversnelling, zoals we die gebruikt hebben, nu bekend als Richardson-*extrapolatie*? Er lijkt zelfs geen polynoom "in zicht" te zijn dat men überhaupt zou kunnen gebruiken voor extrapolatie. Toch is er een verband, en wel als volgt.

Veronderstel, dat we een grootheid $N_1(h)$ hebben bepaald voor waarden $h_0, \theta h_0, \theta^2 h_0, \ldots, \theta^q h_0$ van h, voor zekere $h_0 > 0$, $\theta > 0$ ($\theta \neq 1$) en $q \in \{0, 1, 2, \ldots\}$, en dat we vervolgens geïnteresseerd zijn in een benadering van $\lim_{h\to 0} N_1(h)$ op grond van deze data. De voor de hand liggende manier om dit te doen, is het bepalen van het interpolerend polynoom bij deze data. We bepalen dus $p_1(h) = \sum_{i=0}^{q} \tilde{c}_i h^i$ zodanig dat $p_1(\theta^i h_0) = N_1(\theta^i h_0)$ voor $i = 0, \ldots, q$, en vervolgens kiezen we $p_1(0) = \tilde{c}_0$ als benadering van de limiet. Omdat 0 niet door de interpolatiepunten wordt ingeklemd, is hier inderdaad sprake van extrapolatie (naar $h = 0$). Het blijkt nu dat we $\tilde{c}_0$ op een eenvoudige manier kunnen berekenen, en wel door machten van h te elimineren in het polynoom p_1, zoals we dat al eerder deden voor de asymptotische ontwikkeling van N_1. Definieer immers $p_2(h) = \frac{\theta p_1(h) - p_1(\theta h)}{\theta - 1}$, zodat h is geëlimineerd. Daarna volgt $p_2(h) = \frac{\theta^2 p_2(h) - p_2(\theta h)}{\theta^2 - 1}$, zodat nu ook h^2 is geëlimineerd. Zo voortgaand zijn in $p_{q+1}(h)$ uiteindelijk *alle* positieve machten van h geëlimineerd die in het polynoom p_1 voorkwamen, zodat blijkbaar $p_{q+1}(h) = \tilde{c}_0$ voor alle h. De crux hiervan is, dat in het bijzonder dan dus ook $p_{q+1}(h_0) = \tilde{c}_0$ geldt, waarbij het linkerlid eenvoudig berekenbaar is. Immers: de definitie van de p_j maakt duidelijk dat we voor die berekening van het linkerlid de waarden van $p_1(\theta^i h_0)$ moeten kennen voor $i = 0, 1, \ldots, q$. Maar die kennen we inderdaad ook, want $p_1(\theta^i h_0) = N_1(\theta^i h_0)$ vanwege de keuze van p_1 als interpolerend polynoom van N_1 in die punten. Blijkbaar levert deze berekeningswijze ons inderdaad de geëxtrapoleerde waarde $\tilde{c}_0$ op.

Anderzijds kunnen we natuurlijk ook, zonder extrapolatie in gedachten te hebben, $N_2(h), N_3(h), \ldots$ op basis van $N_1(h)$ definiëren zoals in de hoofdtekst. Vanwege de gelijke structuur van de definities van de p_j en de N_j zien we dan dat $p_{q+1}(h_0) = N_{q+1}(h_0)$. Blijkbaar is ook $N_{q+1}(h_0) = \tilde{c}_0$. Samengevat zien we dat $N_{q+1}(h_0)$ hetzelfde is als de waarde bij extrapolatie naar $h = 0$, zoals die gevonden wordt op basis van $N_1(h_0), N_1(\theta h_0), \ldots, N_1(\theta^q h_0)$. Dit verklaart waarom de term extrapolatie verbonden is aan de techniek van convergentieversnelling.

De waarden in het Romberg-schema hebben (dus) ook een interpretatie in termen van extrapolatie naar $h = 0$. Voor de diagonaal-elementen $T_{q,q}$ ($q \geq 0$) is deze als volgt. Laat p het polynoom van graad ten hoogste q zijn, zodanig dat $p((\frac{h_0}{2^i})^2) = T_{\frac{h_0}{2^i}}(f)$ voor $i = 0, 1, 2 \ldots, q$. Dan is $T_{q,q} = p(0)$.

4.7 Meerdimensionale integralen

Net als bij "exacte" berekening van meerdimensionale integralen, wordt ook bij het numeriek benaderen daarvan het probleem herleid tot dezelfde vraag in 1 dimensie. We zullen dit niet uitgebreid behandelen, maar ons beperken tot een karakteristiek voorbeeld.

Veronderstel dat $D = \{(x, y) \,|\, 0 \leq x \leq 2, 0 \leq y \leq 1\}$ en dat $f \in C(D)$. We zullen $\int_D f(x, y)\, d(x, y)$ benaderen, als volgt.

Uiteraard is $\int_D f(x, y)\, d(x, y) = \int_{x=0}^{2} \left\{ \int_{y=0}^{1} f(x, y)\, dy \right\} dx$. Kies nu allereerst een kwadratuurregel voor x op $[0, 2]$, met kwadratuur-punten $x_1, \ldots, x_m$ en bijbehorende gewichten $w(x_1), \ldots, w(x_m)$. Dan zullen we als benadering voor de buitenintegraal dus

$$\sum_{i=0}^{m} w(x_i) \int_{y=0}^{1} f(x_i, y)\, dy$$

willen nemen. Kies nu een kwadratuurregel voor y op $[0, 1]$, met kwadratuur-punten $y_1, \ldots, y_n$ en bijbehorende gewichten $\tilde{w}(y_1), \ldots, \tilde{w}(y_n)$. Hiermee wordt de integraal in de i-de sommand

dan benaderd als

$$\sum_{j=0}^{n} \tilde{w}(y_j) f(x_i, y_j),$$

met als totaalresultaat de dubbelsom

$$\sum_{i=0}^{m} \sum_{j=0}^{n} w(x_i)\tilde{w}(y_j) f(x_i, y_j)$$

als benadering voor de tweedimensionale integraal. Het is duidelijk hoe dit idee naar hogere dimensies gegeneraliseerd kan worden: ook daar splitst men de integraal op in een herhaling van eendimensionale integralen. Het is ook duidelijk dat er grote variatie in de keuze van de benaderingen van de 1-dimensionale integralen mogelijk is (variatie in regels, desnoods ook nog een andere regel per variabele, Romberg-integratie). Alle mogelijke keuzes op dit punt leveren echter uiteindelijk altijd een herhaalde som van gewogen functiewaarden op als benadering voor de integraal, net als dat in het 1-dimensionale geval zo was. Bij die constatering laten we het hier.

Hoofdstuk 5

Lineaire stelsels

In deze sectie bekijken we verschillende oplossingsstrategieën voor een lineair stelsel van de vorm $Ax = b$. Hierbij is A een $n \times n$-matrix met reële of complexe coëfficiënten en regulier, d.w.z. $\det A \neq 0$.

Er zijn verschillende theoretische manieren om dit stelsel op concreet te lossen, maar de ene manier is voor praktische doeleinden beter dan de andere. De regel van Cramer bijvoorbeeld, geeft een antwoord in een vorm die in principe wel berekenbaar is, maar in de praktijk is dat alleen voor zeer kleine waarden van n ook nog echt te doen, althans wanneer we de erin voorkomende determinanten willen bepalen door $n!$ termen te berekenen en te sommeren. Het aantal termen wordt simpelweg te groot om zelfs met computers nog te kunnen behandelen. Gauß-eliminatie daarentegen, zoals die bij lineaire algebra wordt behandeld, is een veel efficiëntere methode, die ook bij grotere matrices nog werkbaar is. We zullen de theorie van de Gauß-eliminatie nog verder ontwikkelen en verfijnen, met als primair doel het ontwerpen van een efficiënte manier om *herhaaldelijk* een stelsel van de vorm $Ax = b$ op te lossen, met steeds dezelfde A, maar wisselende b. De betreffende inzichten zullen ons zelfs in staat stellen om in een aantal gevallen ook één stelsel al sneller op te lossen dan we tot nu toe konden.

Gauß-eliminatie is een voorbeeld van een zgn. *directe* methode om een stelsel op te lossen. Hiermee wordt een methode bedoeld die in een eindig aantal stappen het antwoord exact oplevert (even afgezien van machine-precisie). De naamgeving suggereert dat er een alternatieve aanpak bestaat met de naam "indirecte methoden", maar dat is niet zo. De alternatieve benadering van het stelsel bestaat uit het niet per sé meer willen kennen van de exacte oplossing, maar het tevreden zijn met benaderingen. De resulterende zgn. *iteratieve methoden* zijn er dan ook op gericht om een rij van almaar beter wordende benaderingen van de echte oplossing te produceren. Zeker in een context, waarin bijv. de coëfficiënten van het stelsel als gevolg van meetfouten sowieso niet exact zijn vast te stellen, of waarin het stelsel voortkomt uit een model dat de werkelijkheid alleen maar benadert, is de praktische waarde van zo'n benaderde oplossing eigenlijk niet kleiner dan die van de exacte uitkomst. In zekere zin is er zelfs in de praktijk ook geen echt principieel verschil tussen een door iteratie benaderde oplossing en een exacte oplossing: de waarden van een exacte oplossing zijn i.h.a. niet representeerbaar, en dus sowieso niet exact door een computer weer te geven. Ook aan deze iteratieve methoden (die gericht zijn op één stelsel tegelijk) zullen we aandacht besteden.

5.1 Herhaalde stelsels: LU-ontbinding (algemene geval)

In deze paragraaf ontwikkelen we de theorie van Gauß-eliminatie verder, primair gemotiveerd door de wens om op een efficiënte manier *herhaaldelijk* een stelsel van de vorm $Ax_i = b_i$ op te lossen, voor $i = 1, \ldots, p$.

Ter herinnering: een matrix U heet een *bovendriehoeksmatrix* ("upper triangular") als hij van de vorm

$$U = \begin{pmatrix} u_{11} & \ldots & u_{1n} \\ & \ddots & \vdots \\ \mathbf{0} & & u_{nn} \end{pmatrix}$$

is, d.w.z. wanneer alle coëfficiënten die niet 0 zijn zich op of boven de hoofddiagonaal bevinden. De determinant is het produkt van de coëfficiënten op de hoofddiagonaal, zodat U regulier is dan en slechts dan als er zich geen 0 op de hoofddiagonaal bevindt.

Het product van twee bovendriehoeksmatrices is weer een bovendriehoeksmatrix. Wanneer de inverse van een bovendriehoeksmatrix bestaat, dan is dat ook weer een bovendriehoeksmatrix. De eenheidsmatrix is uiteraard ook een bovendriehoeksmatrix, en we concluderen dat de verzameling van inverteerbare bovendriehoeksmatrices een groep vormt.

Analoog definieert men een onderdriehoeksmatrix L ("lower triangular"), met een analoog resultaat voor regulariteit en groepsstructuur.

Het idee zal nu zijn om een reguliere matrix A te ontbinden als het product $A = LU$ van een boven- en een onderdriehoeksmatrix[1]. Deze ontbinding is, zoals zal blijken, gemakkelijk te verkrijgen door in een Gauß-eliminatie de gebruikte quotiënten te bewaren. We passen deze eliminatie ("vegen", maar zonder rijverwisseling of vermenigvuldigen van een rij met een constante) daarbij dan alleen op A toe en vergeten b even. Een vergelijking $Ax = b$ lossen we dan daarna m.b.v. de ontbinding in twee slagen op: eerst lossen we y op uit $Ly = b$, en vervolgens lossen we x op uit $Ux = y$. Dan is $LUx = b$, dus x is de gevraagde oplossing.

We zullen nu eerst de theorie van deze ontbinding ontwikkelen en naderhand de efficiency van deze oplossingsmethode vergelijken met die van Gauß-eliminatie zoals die bij lineaire algebra wordt behandeld, d.w.z. "vegen met medeneming van b".

De te ontwikkelen theorie is gebaseerd op een beschrijving van Gauß-eliminatie in termen van matrixvermenigvuldiging. Dit gebeurt als volgt.

Beschouw de matrix

$$M = \begin{pmatrix} 1 & 0 & & & & \ldots & & & 0 \\ 0 & 1 & 0 & & & & & & 0 \\ & \ddots & 1 & \ddots & & 0 & & & \\ \vdots & & & & & & & & \vdots \\ & 0 & 0 & 0 & \ddots & 0 & & & \\ & 0 & \lambda_{ij} & 0 & \ddots & 1 & & & \\ & 0 & 0 & 0 & & & \ddots & & \\ & & & & & & 0 & 1 & 0 \\ 0 & & \ldots & & & & & 0 & 1 \end{pmatrix}$$

De linksvermenigvuldiging MA van A met M heeft tot gevolg dat λ_{ij} maal de j-de rij van A bij de i-de rij wordt opgeteld. De andere rijen van A blijven ongemoeid. Dit is precies het

[1] Dat kan niet altijd, maar die obstructie is onschuldig en makkelijk te omzeilen—zie Opmerking 5.1.5

soort operaties dat bij Gauß-eliminatie plaatsvindt, en we verzamelen dus wat informatie over dit soort matrices. Definieer hiertoe E_{ij} $(i,j=1,\ldots,n)$ als de $n\times n$-matrix die uitsluitend uit nullen bestaat, met uitzondering van de coëfficiënt met indices (i,j), die 1 is. De matrix M hierboven is dan dus te schrijven als $\mathbf{I}+\lambda_{ij}E_{ij}$.

Lemma 5.1.1. *1.* $E_{ij}E_{kl}=\delta_{jk}E_{il}$ $(i,j,k,l=1,\ldots,n)$.

2. $(\mathbf{I}+\lambda_{ij}E_{ij})(\mathbf{I}-\lambda_{ij}E_{ij})=\mathbf{I}$ *als* $i\neq j$ $(i,j=1,\ldots,n)$.

3. $(\mathbf{I}+\lambda_{ij}E_{ij})(\mathbf{I}+\lambda_{kj}E_{kj})=(\mathbf{I}+\lambda_{ij}E_{ij}+\lambda_{kj}E_{kj})$ *als* $j\neq k$ $(i,j,k=1,\ldots,n)$.

Bewijs. Deel 1 volgt door uitrekenen en de andere onderdelen zijn dan duidelijk. □

De inverse van de matrix M hierboven wordt blijkbaar gevonden door λ_{ij} te vervangen door $-\lambda_{ij}$. Het product van twee van dergelijke matrices, die corresponderen met dezelfde kolom, wordt blijkbaar gevonden door de beide niet-triviale coëfficiënten simpelweg over te nemen.

Neem nu aan dat de reguliere matrix A zonder rijverwisselingen naar bovendriehoeksvorm geveegd kan worden. Noteer

$$A^{(1)}=A=\begin{pmatrix}a_{11}^{(1)} & \ldots & a_{1n}^{(1)}\\ \vdots & & \vdots\\ a_{n1}^{(1)} & \ldots & a_{nn}^{(1)}\end{pmatrix}$$

Zij $\lambda_{i1}=\dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ voor $2\leq i\leq n$. Dan laat zich het van boven naar onder schoonvegen van de eerste kolom van $A^{(1)}$ beschrijven door de vergelijking

$$(\mathbf{I}-\lambda_{n1}E_{n1})\cdots(\mathbf{I}-\lambda_{31}E_{31})(\mathbf{I}-\lambda_{21}E_{21})A^{(1)}=A^{(2)}=\begin{pmatrix}a_{11}^{(2)} & a_{12}^{(2)} & \ldots & a_{1n}^{(2)}\\ 0 & a_{22}^{(2)} & \ldots & a_{2n}^{(2n)}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & a_{n2}^{(2)} & \ldots & a_{2n}^{(nn)}\end{pmatrix}$$

In feite is de eerste rij van $A^{(2)}$ dezelfde als de eerste rij van $A^{(1)}$, maar dat hebben we niet nodig. Definieer daarna $\lambda_{i2}=\dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}}$ voor $3\leq i\leq n$, dan is

$$(\mathbf{I}-\lambda_{n2}E_{n2})\cdots(\mathbf{I}-\lambda_{42}E_{43})(\mathbf{I}-\lambda_{32}E_{32})A^{(2)}=A^{(3)}=\begin{pmatrix}a_{11}^{(3)} & a_{12}^{(3)} & a_{13}^{(3)} & \ldots & a_{1n}^{(3)}\\ 0 & a_{22}^{(3)} & a_{23}^{(3)} & \ldots & a_{2n}^{(2n)}\\ 0 & 0 & a_{33}^{(3)} & \ldots & a_{3n}^{(3)}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & a_{n3}^{(3)} & \ldots & a_{nn}^{(3)}\end{pmatrix}$$

Als dit zonder rijverwisselingen kan blijven voortgaan, dan resulteert dit uiteindelijk in een bovendriehoeksmatrix U (het eindstadium van het veegproces), gerelateerd aan A via de vergelijking $L_{n-1}L_{n-2}\cdots L_2L_1A=U$, waarbij

$$L_k=(\mathbf{I}-\lambda_{nk}E_{nk})\cdots(\mathbf{I}-\lambda_{k+1,k}E_{k+1,k})\quad(1\leq k\leq n-1)$$

het vegen van de k-de kolom beschrijft. Blijkbaar is $A = L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1}U$. Nu is, vanwege het eerste deel van Lemma 5.1.1:

$$\begin{aligned} L_k^{-1} &= (\mathbf{I} + \lambda_{k+1,k}E_{k+1,k})\cdots(\mathbf{I} + \lambda_{nk}E_{nk}) \\ &= (\mathbf{I} + \lambda_{k+1,k}E_{k+1,k} + \ldots + \lambda_{nk}E_{nk}). \end{aligned}$$

Om het produkt van de verschillende L_k^{-1} verder te begrijpen kijken we naar een produkt $L_l^{-1}L_k^{-1}$ met $l < k$. Vanwege het voorgaande is

$$L_l^{-1}L_k^{-1} = (\mathbf{I} + \lambda_{l+1,l}E_{l+1,l} + \ldots + \lambda_{nl}E_{nl})(\mathbf{I} + \lambda_{k+1,k}E_{k+1,k} + \ldots + \lambda_{nk}E_{nk}).$$

Bij uitwerking hiervan komen termen van de vorm $E_{al}E_{bk}$ voor. Deze zijn echter alle 0 op grond van het eerste deel van Lemma 5.1.1, want bij deze termen is $b \geq k+1 > l+1$, zodat in het bijzonder $l \neq b$. Blijkbaar is voor $l > k$

$$L_l^{-1}L_k^{-1} = (\mathbf{I} + \lambda_{l+1,l}E_{l+1,l} + \ldots + \lambda_{nl}E_{nl} + \lambda_{k+1,k}E_{k+1,k} + \ldots + \lambda_{nk}E_{nk}).$$

Dit verschijnsel is typisch voor het stapsgewijs uitwerken van het totale produkt $L_1^{-1}L_2^{-1}\cdots L_{n-1}^{-1}$: alle "kwadratische termen" zijn voortdurend 0 op grond van de indices. Het produkt heeft blijkbaar uiteindelijk een verrassend eenvoudige vorm, die verkregen wordt door simpelweg alle voorkomende termen $\lambda_{ij}E_{ij}$ uit alle matrices L_k^{-1} bij de eenheidsmatrix op te tellen. Met deze constatering is het volgende resultaat bewezen.

Stelling 5.1.2 (LU-ontbinding). *Laat A een reguliere $n \times n$-matrix zijn. Neem aan dat A zonder rijverwisselingen door vegen in een bovendriehoeksmatrix U kan worden overgevoerd. Dan is*

$$A = LU,$$

met

$$L = (\mathbf{I} + \sum_{i>j} \lambda_{ij}E_{ij}),$$

waarin λ_{ij} het quotient is dat in de j-de stap (d.w.z. het vegen van de j-de kolom) gebruikt wordt met betrekking tot de i-de rij.

Het is dus gemakkelijk om in bovenstaande situatie de LU-ontbinding van A te bepalen: dat gebeurt door te vegen en gaandeweg de matrix L elementsgewijs onder de hoofddiagonaal op te bouwen: eerst de eerste kolom (van boven naar onder invullen), dan de tweede kolom (van boven naar onder invullen), etc. Op de hoofddiagonaal van L staan enen; daarboven uiteraard alleen nullen.

Voorbeeld 5.1.3. Laat

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix}$$

We starten met de kennis dat

$$L = \begin{pmatrix} 1 & 0 & 0 \\ & 1 & 0 \\ & & 1 \end{pmatrix}$$

De in de eerste stap te gebruiken quotiënten zijn $\frac{0}{1} = 0$ en $\frac{2}{1} = 2$, zodat blijkbaar

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \mathbf{0} & 1 & 0 \\ \mathbf{2} & & 1 \end{pmatrix}.$$

Na de eerste stap is A geveegd tot

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & 1 \end{pmatrix}.$$

In de tweede en laatste stap is het te gebruiken quotiënt gelijk aan $\frac{-1}{1} = -1$, zodat blijkbaar

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}.$$

Na deze laatste stap is A geveegd tot het eindprodukt

$$U = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}$$

Blijkbaar is

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}$$

de ontbinding $A = LU$ uit de Stelling 5.1.2.

Voorbeeld 5.1.4. In dit voorbeeld geven we een schematische aanpak van de ontbinding. Laat

$$A = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix}.$$

We bepalen voor de matrix A (die inderdaad zonder rijverwisselingen door vegen in bovendriehoeksvorm kan worden overgevoerd) de LU-ontbinding uit Stelling 5.1.2. Hieronder staan, steeds naast elkaar, de al zo ver mogelijk ingevulde L en de geveegde vorm van A, zoals die er aan het begin van iedere stap uitzien.

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ & 1 & 0 & 0 \\ & & 1 & 0 \\ & & & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ \mathbf{2} & 1 & 0 & 0 \\ \mathbf{3} & & 1 & 0 \\ \mathbf{-1} & & & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & -4 & -1 & -7 \\ 0 & 3 & 3 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & \mathbf{4} & 1 & 0 \\ -1 & \mathbf{-3} & & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & \mathbf{0} & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}$$

Uit de laatste regel lezen we af dat

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}$$

blijkbaar de ontbinding $A = LU$ uit Stelling 5.1.2 is.

Opmerking 5.1.5. Niet alle reguliere matrices A zijn zonder rijverwisselingen door vegen in een bovendriehoeksmatrix over te voeren. Het is in zijn algemeenheid daarom nodig om eerst de rijen van A te permuteren, zó, dat dit voor de matrix met de gepermuteerde rijen wél mogelijk is. Dit permuteren wordt gerealiseerd door links te vermenigvuldigen met een permutatiematrix P, d.w.z. een matrix die in iedere kolom en iedere rij precies één 1 heeft staan, en verder alleen nullen. Met andere woorden: voor wilekeurige reguliere A is er altijd een permutatiematrix P zodanig dat $PA = LU$, met L een onderdriehoeksmatrix met enen op de diagonaal, en U een bovendriehoeksmatrix.

Opmerking 5.1.6. Veronderstel dat de reguliere matrix A zonder rijverwisselingen door vegen in een bovendriehoeksmatrix is over te voeren. Dan bestaat er in het bijzonder een decompositie $A = LU$ met L een onderdriehoeksmatrix, en U een bovendriehoeksmatrix. De vraag is in hoeverre die decompositie uniek is.

Wanneer D een reguliere diagonaalmatrix is, dan vormen $L_* = LD$ en $U_* = D^{-1}U$ samen weer een dergelijke decompositie. Op deze manier kan men decomposities vinden waarbij de coëfficiënten op de hoofddiagonaal van L willekeurige voorgeschreven waarden ongelijk nul hebben.

Anderzijds, als $A = L_*U_*$ een andere decompositie is in onder- en bovendriehoeksfactoren, dan volgt uit $L_*U_* = LU$ dat $L^{-1}L_* = UU_*^{-1}$. Blijkbaar zijn deze produkten zowel een onder- als een bovendriehoeksmatrix, m.a.w., ze zijn gelijk aan een diagonaalmatrix D, zodat $L_* = LD$ en $U_* = D^{-1}U$. Deze D ligt dan uniek vast door de diagonaalelementen van L en L_*.

Voor een reguliere matrix A, die zonder rijverwisselingen in een bovendriehoeksmatrix over te voeren is, is de eindconclusie dus dat er, voor iedere keuze (ongelijk aan nul) voor ieder van de coëfficiënten op de hoofddiagonaal van L, precies 1 ontbinding $A = LU$ in een bovendriehoeksmatrix L en een onderdriehoeksmatrix U mogelijk is. De ontbinding uit Stelling 5.1.2 is de (blijkbaar unieke) ontbinding behorend bij de keuze van enen voor alle coëfficiënten op de hoofddiagonaal.

Later zal het volgende criterium handig blijken te zijn. In de formulering ervan komen de hoofdminoren van een $n \times n$ matrix voor: dit zijn de matrices $A_1, \ldots, A_n$ die verkregen worden door voor $k = 1, \ldots, n$ een $k \times k$-vierkant uit A te lichten, te beginnen bij de linkerbovenhoek.

Propositie 5.1.7. *Laat A regulier zijn. A is zonder rijverwisselingen in een bovendriehoeksmatrix over te voeren, dan en slechts dan, als iedere hoofdminor van A ook regulier is.*

Bewijs. Veronderstel dat er geen rijverwisselingen nodig zijn. Merk dan op, dat dan blijkbaar de determinant van iedere hoofdminor in iedere stap van het veegproces ongewijzigd blijft (de minor zelf verandert uiteraard wel). Uiteindelijk blijkt zo'n minor dus het produkt te zijn van een aantal elementen op de diagonaal van U, en die zijn alle ongelijk nul, want A is regulier.

Anderzijds, als alle hoofdminoren regulier zijn, dan bewijzen we voor $k = 1, \ldots, n-1$ met inductie dat er k stappen van het veegproces kunnen worden uitgevoerd, zonder dat er rijverwisseling nodig is. In het bijzonder kunnen er dus $n-1$ stappen worden uitgevoerd, en dat

correspondeert met het hele veegproces. Het regulier zijn van A_1, d.w.z. het gegeven dat $a_{11} \neq 0$, geeft onmiddellijk het begin van de inductie voor $k = 1$. Neem vervolgens aan dat er reeds k stappen zonder rijverwisseling uitgevoerd konden worden, met $k \leq n-2$. De matrix ziet er na die k stappen als volgt uit:

$$\begin{pmatrix} a_{11}^{(k+1)} & \dots & \dots & a_{1,k+1}^{(k+1)} & \dots & a_{1,n}^{(k+1)} \\ 0 & a_{22}^{(k+1)} & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & \dots & 0 & a_{k+1,k+1}^{(k+1)} & \dots & a_{k+1,n}^{(k+1)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & a_{n,k+1}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{pmatrix}$$

Omdat er nog steeds geen rijverwisselingen zijn geweest, is de determinant van de $(k+1)$-de hoofdminor ongewijzigd gebleven, en i.h.b. is die determinant nog steeds ongelijk aan nul. Blijkbaar is $a_{11}^{(k+1)} a_{22}^{(k+1)} \cdots a_{k+1,k+1}^{(k+1)} \neq 0$, zodat i.h.b. $a_{k+1,k+1}^{(k+1)} \neq 0$. Ook de $k+1$-de stap kan dus zonder rijverwisseling plaatsvinden. □

Alles bij elkaar hebben we:

Stelling 5.1.8. *Laat A een reguliere matrix zijn. Dan zijn equivalent:*

1. *A is zonder rijverwisselingen door vegen in een bovendriehoeksmatrix over te voeren.*
2. *A kan geschreven worden als $A = LU$, met L een onderdriehoeksmatrix en U een bovendriehoeksmatrix.*
3. *De determinant van iedere hoofdminor van A is ongelijk aan nul.*

Bewijs. De implicatie (1) $\Longrightarrow$ (2) is een gevolg van Stelling 5.1.2. De equivalentie van (1) en (3) is Propositie 5.1.7. Om de implicatie (2) $\Longrightarrow$ (3) in te zien, merkt men op dat iedere hoofdminor van A het produkt is van de corresponderende hoofdminoren van L en U. Ieder van die laatste hoofdminoren is regulier omdat L en U dat zijn, als gevolg van de regulariteit van A. □

Na deze bestudering van de ontbinding zullen we nu nader ingaan op het oplossen van stelsels m.b.v. de verkregen LU-ontbinding, en daarna deze aanpak vergelijken met de klassieke Gauß-eliminatie.

Het oplossen van de vergelijking $Ax = b$, waarbij $A = LU$ als in Stelling 5.1.2, gaat, zoals boven al vermeld, als volgt. Eerst lossen we y op uit $Ly = b$, en vervolgens lossen we x op uit $Ux = y$. Dan is $LUx = b$, dus x is de gevraagde oplossing.

Voorbeeld 5.1.9. Uitgaande van

$$A = \begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix}$$

en $b = (8, 7, 14, -7)^t$ lossen we eerst

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & 4 & 1 & 0 \\ -1 & -3 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 14 \\ -7 \end{pmatrix}$$

op. Dit is een stelsel in onderdriehoeksvorm, en met zgn. voorwaartse substitutie (d.w.z. beginnend met y_1) vinden we achtereenvolgens $y_1 = 8$, $y_2 = -9$, $y_3 = 26$ en $y_4 = -26$. Vervolgens lossen we

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 0 & -1 & -1 & 5 \\ 0 & 0 & 3 & 13 \\ 0 & 0 & 0 & -13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ -9 \\ 26 \\ -26 \end{pmatrix}$$

op. Dit stelsel is in bovendriehoeksvorm, en met zgn. achterwaartse substitutie (d.w.z. beginnend met x_4) vinden we voor de uiteindelijke oplossing achtereenvolgens $x_4 = 2$, $x_3 = 0$, $x_2 = -1$ en $x_1 = 3$.

We zullen nu de hoeveelheid werk van Gauß-eliminatie en de benadering met de LU-ontbinding met elkaar vergelijken Het verschil bij het oplossen van $Ax = b$ zit hem in de behandeling van b. In het klassieke proces wordt b tijdens het veegproces meegenomen, en moet er uiteindelijk een achterwaartse substitutie worden uitgevoerd. Bij de aanpak met de LU-ontbinding (even aannemend dat er geen rijverwisselingen nodig zijn) wordt b in eerste instantie ongemoeid gelaten, en worden pas later uitgaande van b een voorwaartse en een achterwaartse substitutie uitgevoerd. Wat betreft de bepaling van de LU-ontbinding is er eigenlijk niet eens een verschil: dat is gewoon een kwestie van het al dan niet opschrijven van de bij het vegen gebruikte quotiënten.

Voor het oplossen van één stelsel kosten—zo gaat men na—de klassieke methode en de aanpak met de LU-ontbinding precies evenveel operaties. Bij het herhaaldelijk oplossen van een stelsel $Ax_i = b_i$, voor $i = 1, \ldots, p$ (p vast), is de aanpak via de LU-ontbinding echter de beste:

- De klassieke methode, met medeneming van b, blijkt voor één stelsel $\frac{2}{3}n^3 + \text{l.o.t.}$ operaties te kosten. Met het oplossen van p stelsels zijn dus bij een herhaalde klassieke benadering (d.w.z. p keer vegen, steeds met medeneming van de betreffende b_i) in totaal $p(\frac{2}{3}n^3 + \text{l.o.t.})$ operaties nodig.

- Het aantal operaties voor het bepalen van de LU-ontbinding is eveneens van de vorm $\frac{2}{3}n^3 + \text{l.o.t.}$ (met kleinere lagere orde termen dan boven). Het voor één stelsel uitvoeren van de voorwaartse en achterwaartse substitutie kost samen $2n^2 + \text{l.o.t.}$ operaties. Het éénmaal bepalen van de LU-onbinding, en het daarna p maal toepassen van een voorwaartse en achterwaartse substitutie, kost dus samen $\frac{2}{3}n^3 + l.o.t.$ operaties

Voor grote n is de methode met de LU-ontbinding dus ruwweg een factor p sneller! De crux zit hem in het verschil tussen het vegen van een matrix, waarbij het aantal operaties van orde n^3 is, en het oplossen van een stelsel in driehoeksvorm, waarbij het aantal operaties slechts van orde n^2 is: dat laatste is van een orde lager, en kost dus "niets" extra.

5.2 Pivoting

In iedere stap van het eliminatie-proces wordt er een (deel van) kolom geveegd. In de noemer van de daarbij gebruikte quotiënten komt daarbij een (voor iedere stap vast) element voor, nl. het eerste element in de rij waarmee geveegd wordt (en dat dus niet nul is). Deze coëfficiënt heet de *pivot* ("spil") van die stap. Het zoeken van een (geschikte) pivot heet *pivoting*. Bij lineaire algebra verkrijgt men bijvoorbeeld, wanneer de "natuurlijke" kandidaat pivot toevallig nul is, een geschikte pivot door twee rijen te verwisselen. Propositie 5.1.7 kan men interpreteren als de beschrijving van een nodig en voldoende voorwaarde waaronder er geen pivoting nodig is, omdat het element linksboven in het nog te behandelen blok automatisch steeds ongelijk nul is en dus als

pivot gebruikt kan worden. Voor een theoretische beschrijving is dit natuurlijk overzichtelijk, maar voor computer-implementaties is het soms toch noodzakelijk om door verwisseling een andere pivot te kiezen, zelfs al is de "natuurlijke" kandidaat ongelijk aan nul. Dit komt door de eindige machine-precisie. Als voorbeeld hiervoor kiezen we het stelsel:

$$\begin{aligned} 0.0001\,x_1 &+ x_2 = 1 \\ x_1 &+ x_2 = 2. \end{aligned}$$

De exacte oplossing is $x_1 = 1.001\cdots$, $x_2 = 0.9999\cdots$. Bij berekening in drie cijfers precisie wordt dit, bij gebruik van 0.0001 als pivot, in één stap in de volgende bovendriehoeksvorm gebracht:

$$\begin{aligned} 0.0001\,x_1 &+ x_2 = 1 \\ -10000\,x_2 &= -10000 \end{aligned}$$

Achterwaartse substitutie geeft dan $x_2 = 1$ en $x_1 = 0$. De fout in x_2 is nog acceptabel, maar de fout in x_1 is 100%. Door rijverwisseling gaat het stelsel echter over in

$$\begin{aligned} x_1 &+ x_2 = 2 \\ 0.0001\,x_1 &+ x_2 = 1. \end{aligned}$$

Wanneer we nu 1 als pivot gebruiken, dan gaat (nog steeds in drie cijfers precisie rekenend) het stelsel over in

$$\begin{aligned} x_1 + x_2 &= 2 \\ x_2 &= 1. \end{aligned}$$

Achterwaartse substitutie geeft hier dan $x_2 = 1$ en $x_1 = 1$, hetgeen veel beter is. De oorzaak van het fenomeen ligt hier in het grote quotiënt $\frac{1}{0.0001}$ in de eerste berekeningswijze: dit laat bij doorwerking andere informatie verdwijnen in de machine-precisie.

Een veel gebruikte strategie om dit verschijnsel zo veel mogelijk te vermijden is *column pivoting*, d.w.z. het opzoeken van een (in absolute waarde) grootste element in de kolom die geveegd gaat worden, en het vervolgens verwisselen van de "natuurlijke" veegrij en de rij waarin het gevonden element staat. Dit maakt de gebruikte quotiënten dan immers zo klein mogelijk. De meest vergaande aanpak in die richting is *total pivoting*, d.w.z. het zoeken van een (in absolute waarde) grootste element in het hele nog te behandelen blok, en het daarna achtereenvolgens verwisselen van twee rijen en twee kolommen om dit element op de pivot-plek te krijgen. In ons voorbeeld komen beide strategieën toevallig op hetzelfde neer. Iedere strategie heeft zijn voordelen en nadelen, en het is niet bekend of er een strategie bestaat die in zijn algemeenheid het beste resultaat geeft.

5.3 LU-ontbinding: speciale gevallen

We hebben in het voorgaande gezien dat het, bij het herhaaldelijk oplossen van een stelsel met steeds dezelfde coëfficiënten, loont om eerst een LU-ontbinding van de betreffende matrix te bepalen (eventueel met inbegrip van een permutatie-matrix om noodzakelijke rijverwisselingen te verdisconteren). Daarna kan men snel door voorwaartse en achterwaartse substitutie de stelsels achter elkaar afhandelen. De grootste hoeveelheid werk hierbij zit hem in het bepalen van de ontbinding: die kost $\sim \frac{2}{3}n^3$ operaties. Het loont dus om te kijken of op het bepalen van die ontbinding misschien in speciale gevallen bezuinigd kan worden. Dat is inderdaad het geval, zoals hieronder zal blijken. De bijbehorende algoritmen staan bekend als *compacte methoden*, waarmee wordt aangeduid dat een ontbinding wordt bepaald op een manier die veel sneller is dan het uitvoeren van het volledige veegproces.

We zullen dit nu onderzoeken. Het is goed om daarbij in het achterhoofd te houden, dat het er in feite niet ertoe doet welke LU-ontbinding men vindt. De ontbinding in Stelling 5.1.2 is de unieke ontbinding waarin L uitsluitend enen op de hoofddiagonaal heeft, maar iedere andere ontbinding in een onder- en een bovendriehoeksmatrix laat zich net zo goed gebruiken in de methodiek van voorwaartse en achterwaartse substitutie.

Een belangrijke klasse van matrices, waarvoor er een compacte mtehode bekend is, bestaat uit de strikt positief definiete reële matrices. Ter herinnering: een matrix A is *strikt positief definiet* als $A = A^t$ en als $(Ax, x) > 0$ voor alle $0 \neq x \in \mathbb{R}^n$, waarbij $(\,.\,,\,.\,)$ het standaard inwendige produkt op $\mathbb{R}^n$ is. Voor deze klasse geldt:

Stelling 5.3.1 (Sylvester). *Een symmetrische matrix A is strikt positief definiet, dan en slechts dan, als de determinant van iedere hoofdminor strikt positief is.*

In het bijzonder heeft een strikt positief definiete matrix A dus volgens Stelling 5.1.8 een LU-ontbinding. In feite zijn er dan zeer veel van dat soort ontbindingen (vgl. Opmerking 5.1.6), en blijkens de volgende Stelling kunnen we uit die veelheid van ontbindingen een bijzonder fraaie kiezen. Het bewijs laten we achterwege.

Stelling 5.3.2 (Cholesky-decompositie). *Laat A een strikt positief definiete $n \times n$-matrix zijn. Dan is er een unieke onderdriehoeksmatrix L, zodanig dat $A = LL^t$ en $l_{ii} > 0$ voor $i = 1, \ldots, n$.*

Opmerking 5.3.3. Wanneer B regulier is, dan is $A \overset{\text{def.}}{=} BB^t$ strikt positief definiet (ga na). De Cholesky-decompositie laat dus zien, dat alle strikt positief definiete matrices op deze manier verkregen kunnen worden, en zelfs met een (dan unieke) B van een zeer eenvoudige vorm.

We blijken de factor L in een Cholesky-decompositie via hetzelfde "ritme" te kunnen vinden als bij het bepalen van de factor L in een veegproces: eerst de eerste kolom (van boven naar onder invullen), dan de tweede kolom (van boven naar onder invullen), etc. Hierbij moeten we de diagonaalelementen nu wél uitrekenen (bij het veegproces zijn die altijd 1). De berekeningswijze kan onthouden worden als het principe, dat we voor het bepalen van ieder van de elementen in de j-de kolom van L uitsluitend gebruik maken van eveneens de j-de kolom in L^t.

Voorbeeld 5.3.4. De matrix

$$\begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}$$

is strikt positief definiet. Laat

$$\begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{23} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}$$

met $l_{11}, l_{22}, l_{33} > 0$ de Cholesky-deompositie zijn. Er volgt $l_{11}^2 = 1$, dus $l_{11} = 1$. Blijkbaar is

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & l_{21} & l_{31} \\ 0 & l_{22} & l_{23} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}.$$

Nu volgt $l_{21} \cdot 1 = 1$ en $l_{31} \cdot 1 = -2$, dus $l_{21} = 1$ en $l_{31} = -2$. Blijkbaar is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & l_{22} & 0 \\ -2 & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & 1 & -2 \\ 0 & l_{22} & l_{23} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}.$$

Nu volgt $1^2 + l_{22}^2 = 5$, dus $l_{22} = 2$. Blijkbaar is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ -2 & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & 1 & -2 \\ 0 & 2 & l_{23} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}.$$

Nu volgt dat $-2 \cdot 1 + l_{32} \cdot 2 = -4$, dus $l_{32} = -1$. Blijkbaar is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ -2 & -1 & l_{33} \end{pmatrix} \begin{pmatrix} 1 & 1 & -2 \\ 0 & 2 & -1 \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}.$$

Nu volgt dat $(-2)^2 + (-1)^2 + l_{33}^2 = 14$, dus $l_{33} = 3$. Blijkbaar is

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ -2 & -1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 & -2 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -2 \\ 1 & 5 & -4 \\ -2 & -4 & 14 \end{pmatrix}$$

de Cholesky-decompositie van het rechterlid.

Het op deze manier bepalen van de Cholesky-decompositie kost $\sim \frac{1}{3}n^3$ operaties, zo kan men nagaan. Een alternatieve methode zou zijn, het vaststellen van een LU-ontbinding van A m.b.v. het veegproces, en het daarna bepalen van een diagonaalmatrix D z.d.d. LD en $D^{-1}U$ samen de Cholesky-decompositie vormen—we weten immers dat dat kan. De betreffende LU-ontbinding kost dan echter $\sim \frac{2}{3}n^3$ operaties, zodat het bovenstaande algoritme een factor 2 sneller is ("compact"). Het algoritme geeft, zo blijkt, ook een praktisch criterium voor het strikt positief definiet zijn van een symmetrische matrix: dit is precies dan het geval, als bovenstaand algoritme geheel doorlopen kan worden, met strikt positieve l_{ii}.

We zullen ons nu richten op de hieronder te definiëren zgn. bandstructuur van matrices. De bandstructuur van een reguliere matrix en die van de factoren in een eventuele LU-ontbinding blijken nauw gerelateerd te zijn. Dit zal ons a priori kennis over de factoren geven, waarop we dan compacte algoritmen kunnen baseren.

Definitie 5.3.5. Laat A een reguliere matrix zijn. Dan heeft A *bandstructuur* (p, q), waarbij p en q niet-negatieve gehele getallen zijn, als:

1. er op de p-de onderdiagonaal van A nog een element ongelijk aan nul staat, maar op alle lager gelegen onderdiagonalen uitsluitend nullen staan, en
2. er op de q-de bovendiagonaal van A nog een element ongelijk aan nul staat, maar op alle hoger gelegen bovendiagonalen uitsluitend nullen staan.

De hoofddiagonaal zelf wordt hierbij als 0-de bovendiagonaal en als 0-de onderdiagonaal geteld. De *bandbreedte* van A is $p + q + 1$.

Voorbeeld 5.3.6.

1. Een reguliere diagonaalmatrix heeft bandstructuur $(0, 0)$ en bandbreedte 1.
2. De reguliere matrices

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \text{ en } \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

hebben bandstructuur $(1, 2)$ en bandbreedte 4, resp. bandstructuur (2,1) en bandbreedte 4.

3. Een reguliere $n \times n$-onderdriehoeksmatrix heeft bandstructuur $(p, 0)$ en bandbreedte $p+1$, voor zekere $p \in \{0, \ldots, n-1\}$. Een dergelijke bandstructuur karakteriseert onderdriehoeksmatrices. Net zo heeft een reguliere $n \times n$-bovendriehoeksmatrix bandstructuur $(0, q)$ en bandbreedte $q+1$, voor zekere $q \in \{0, \ldots, n-1\}$; dit karakteriseert bovendriehoeksmatrices.

4. Als A bandstructuur (p, q) heeft, dan heeft A^t bandstructuur (q, p). De bandbreedten van A en A^t zijn gelijk.

Een gerelateerd, maar nauwkeuriger, begrip wordt gegeven in de volgende definitie.

Definitie 5.3.7. Laat A een $n \times n$-matrix zijn. Dan is de *skyline* van A de n-vector $\mathrm{skl}(A)$, met als componenten $\mathrm{skl}(A)_1 = 0$ en, voor $j = 2, \ldots, n$,

$$\mathrm{skl}(A)_j = \begin{cases} 0 & \text{als } a_{j-k,j} = 0 \text{ voor alle } k \in \{1, \ldots, j-1\}; \\ k & \text{als } a_{j-k,j} \neq 0 \text{ en } a_{j-l,j} = 0 \text{ voor alle } l \in \{k+1, \ldots, j-1\}. \end{cases}$$

Met andere woorden: $\mathrm{skl}(A)$ beschrijft de hoogte van de wolkenkrabbers die op de verschillende diagonaalelementen staan, waarbij het dak bepaald wordt door het hoogst staande element dat niet nul is (zo dit er is). De skyline van een onderdriehoeksmatrix is de nulvector, en omgekeerd. De matrix

$$\begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 3 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & -1 & 1 \end{pmatrix}$$

heeft skyline $(0, 1, 0, 2)^t$. Het is duidelijk dat een reguliere matrix A bandstructuur (p, q) heeft, dan en slechts dan, als $\max_j \mathrm{skl}(A^t)_j = p$ en $\max_j \mathrm{skl}(A)_j = q$.

De skyline is in zekere zin een behouden grootheid in LU-ontbindingen, getuige het volgende Lemma.

Lemma 5.3.8. *Laat A een reguliere matrix zijn en veronderstel dat A een ontbinding $A = LU$ in een onder- en een bovendriehoeksmatrix heeft. Dan is*

1. $\mathrm{skl}(A) = \mathrm{skl}(U)$*;*

2. $\mathrm{skl}(A^t) = \mathrm{skl}(L^t)$.

Bewijs. Merk op dat de diagonaalelementen van L ongelijk aan nul zijn, want A is regulier. Kijk nu naar de j-de kolom van LU en bezie, wat de eerste coëfficiënt is in die kolom (bovenaan beginnend), die ongelijk aan nul is. Een moment van nadenken laat zien dat dit—omdat L een onderdriehoeksmatrix is met diagonaalelementen ongelijk aan nul—juist correspondeert met de eerste coëfficiënt in de j-de kolom van U (bovenaan beginnend), die ongelijk aan nul is. Met andere woorden: de wolkenkrabbers boven het j-de diagonaalelement in A en U hebben het dak op dezelfde hoogte, en zijn dus ook even hoog. Dit bewijst het eerste deel, waar we nog niet hebben gebruikt dat U een bovendiehoeksmatrix is. Dit echter is nodig voor het tweede deel, dat dan door transpositie uit het eerste deel volgt. □

Ga eens na dat deze gemeenschappelijke skylines inderdaad te zien zijn in de ontbindingen die we tot nu toe als voorbeeld gegeven hebben.

Gevolg 5.3.9. *Zij A een reguliere matrix en veronderstel dat A een ontbinding $A = LU$ in een onder- en een bovendriehoeksmatrix heeft. Dan zijn de volgende twee beweringen equivalent:*

1. *A heeft bandstructuur* (p, q).

2. *L heeft bandstructuur* $(p, 0)$ *en U heeft bandstructuur* $(0, q)$.

Bewijs.

$$
\begin{aligned}
A \text{ heeft bandstructuur } (p,q) &\iff \begin{cases} \max_j \operatorname{skl}(A^t)_j &= p \\ \max_j \operatorname{skl}(A)_j &= q \end{cases} \\
&\overset{\text{Lemma 5.3.8}}{\iff} \begin{cases} \max_j \operatorname{skl}(L^t)_j &= p \\ \max_j \operatorname{skl}(U)_j &= q \end{cases} \\
&\iff \begin{cases} L & \text{heeft bandstructuur } (p,0) \\ U & \text{heeft bandstructuur } (0,q). \end{cases}
\end{aligned}
$$

□

Bovenstaand gevolg geeft dus aan, dat er in een veegproces ter bepaling van een LU-ontbinding van een matrix met een niet-triviale bandstructuur, hele blokken in L en U op voorhand al als nul kunnen worden ingevuld. Het nul zijn van de betreffende coëfficiënten hoeft niet meer door berekening geconstateerd te worden, en op voorhand weet men ook al dat bepaalde veegoperaties niet behoeven te worden uitgevoerd. Al met al kan dit, zeker voor kleine p en q, een zeer aanzienlijke hoeveelheid werk besparen.

Men kan in dit verband laten zien dat, indien een LU-decompositie van een reguliere matrix met bandstructuur (p, q) bestaat, het bepalen van zo'n decompositie met behulp van dit aangepaste "slimme" veegproces, samen met een voorwaartse en achterwaartse substitutie, ten hoogste $(pq + 2p + q + 1)n$ operaties kost. Dit is lineair in n (!). Een bijna curieus gevolg hiervan is, dat het op deze manier oplossen van een lineair stelsel $Ax = b$ dan beter is dan het uitvoeren van de matrixvermenigvuldiging $A^{-1}b$, zelfs wanneer de matrix A^{-1} al expliciet bekend is. De matrix A^{-1} heeft nl. in het algemeen geen bandstructuur (en kan zelfs alle coëfficiënten ongelijk aan nul hebben), zodat het uitvoeren van de vermenigvuldiging dan i.h.a. $\sim 2n^2$ operaties kost.

Voor een reguliere matrix A met bandstructuur $(1, 1)$—een zgn. (echt) *tridiagonale* matrix—waarvoor een LU-ontbinding bestaat, kan men een dergelijke ontbinding gemakkelijk vinden zonder het veegproces te doorlopen, als volgt. We mogen aannemen dat de diagonaal van L uit enen bestaat. Laat

$$
\begin{pmatrix} 1 & 0 & & & & \mathbf{0} \\ l_{21} & 1 & 0 & & & \\ 0 & l_{32} & 1 & 0 & & \\ & 0 & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 & 0 \\ \mathbf{0} & & & 0 & l_{n,n-1} & 1 \end{pmatrix}
\begin{pmatrix} u_{11} & u_{12} & 0 & & & \mathbf{0} \\ 0 & u_{22} & u_{23} & 0 & & \\ & 0 & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & 0 \\ & & & \ddots & \ddots & u_{n-1,n} \\ \mathbf{0} & & & & 0 & u_{n,n} \end{pmatrix}
=
\begin{pmatrix} a_{11} & a_{12} & 0 & & \mathbf{0} \\ a_{21} & a_{22} & a_{23} & 0 & \\ 0 & a_{32} & a_{33} & a_{34} & \ddots \\ & 0 & \ddots & \ddots & \ddots \\ \mathbf{0} & & & & \end{pmatrix}
$$

De vergelijkingen voor de eerste bovendiagonaal van A laten zien dat deze eerste bovendiagonaal identiek is aan de eerste bovendiagonaal voor U. Daarna kan men, zo zal in het voorbeeld hierna duidelijk worden, de diagonaalelementen van U en de elementen op de eerste onderdiagonaal van L bepalen door alternerend te kijken naar de inwendige produkten

(laatste bekende rij van L, eerste onbekende kolom van U)

en

(eerste onbekende rij van L, laatste bekende kolom van U).

Men begint dan bij een inprodukt van het eerste type. Deze methode, die inclusief navolgende voorwaartse en achterwaartse substitutie $8n + 7$ operaties kost, heet de *double sweep method*.

Voorbeeld 5.3.10. De matrix

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

is strikt positief definiet, en heeft bandstructuur $(1, 1)$. Er is dus volgens bovenstaande een LU-ontbinding van de vorm

$$\begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ 0 & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & -1 & 0 \\ 0 & u_{22} & -1 \\ 0 & 0 & u_{33} \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

waarbij de eerste bovendiagonaal van U alvast uit A is overgenomen. Volgens ons alternerend schema van de "double sweep method" kijken we nu naar:

- $1 \cdot u_{11} = 2$, dus $u_{11} = 2$;
- $l_{21} \cdot u_{11} = l_{21} \cdot 2 = -1$, dus $l_{21} = -\frac{1}{2}$;
- $l_{21} \cdot -1 + 1 \cdot u_{22} = -\frac{1}{2} \cdot -1 + 1 \cdot u_{22} = 2$, dus $u_{22} = \frac{3}{2}$;
- $l_{32} \cdot u_{22} = l_{32} \cdot \frac{3}{2} = -1$, dus $l_{32} = -\frac{2}{3}$;
- $l_{32} \cdot -1 + 1 \cdot u_{33} = -\frac{2}{3} \cdot -1 + 1 \cdot u_{33} = 2$, dus $u_{33} = \frac{4}{3}$.

Blijkbaar is

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{2}{3} & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 0 \\ 0 & \frac{3}{2} & -1 \\ 0 & 0 & \frac{4}{3} \end{pmatrix}$$

de ontbinding van A uit Stelling 5.1.2.

5.4 Iteratieve methoden voor lineaire stelsels

Een alternatieve benadering van lineaire stelsels bestaat uit het niet langer exact willen oplossen van het stelsel, maar het genoegen willen nemen met een benadering van de oplossing. Dit kan door praktische overwegingen worden ingegeven, bijvoorbeeld doordat een stelsel (zonder enige extra structuur) dermate groot is, dat zelfs Gauß-eliminatie niet meer uitvoerbaar is. Een exacte oplossing ligt dan niet meer binnen bereik en een benadering is het best haalbare. Dat hoeft ook niet altijd een bezwaar te zijn: wanneer het stelsel bijvoorbeeld voortkomt uit een benaderend model voor een probleem uit de praktijk, geeft zelfs een exacte oplossing van het stelsel i.h.a. slechts een benadering van de werkelijke praktische situatie. Wanneer de coëfficiënten onderhevig zijn aan meetfouten (of eventuele afrondfouten, wanneer ze uit andere berekeningen afkomstig zijn), is het stelsel sowieso al een benadering van het echte stelsel, zodat ook hier een exacte oplossing slechts een schijnzekerheid biedt. Tenslotte moeten we ons realiseren, dat de computerberekening van de exacte oplossing altijd aan afrondfouten onderhevig zal zijn. De

exacte oplossing is i.h.a. zelfs niet representeerbaar! De praktische waarde van een (goed) benaderde oplossing is, kortom, vaak net zo groot als die van de exacte oplossing. We zullen in deze paragraaf dan ook kijken naar methodes, waarmee benaderde oplossingen voor (sommige) stelsels gegenereerd kunnen worden. Deze methodes zijn iteratief.

We maken gebruik van de definities en resultaten uit Paragraaf A.3.

Het idee is als volgt. Laat A een reguliere matrix zijn. We zijn geïnteresseerd in het oplossen van het stelsel $Ax = b$, voor algemene b. Kies een splitsing $A = N - P$, met N regulier. De vergelijking $Ax = b$ is dan equivalent met de vergelijking $x = N^{-1}Px + N^{-1}b$. Definieer nu een iteratief proces, waarbij de rij $\{x^k\}_{k=0}^{\infty}$ in $\mathbb{R}^n$ (of $\mathbb{C}^n$) geconstrueerd wordt door x^0 willekeurig te kiezen en vervolgens

$$x^{k+1} = N^{-1}Px^k + N^{-1}b \tag{5.4.1}$$

te nemen. *Als we geluk hebben*, en de rij $\{x^k\}_{k=0}^{\infty}$ convergeert, zeg naar x^{∞}, dan is blijkbaar $x^{\infty} = N^{-1}Px^{\infty} + N^{-1}b$, m.a.w. x^{∞} is de gezochte oplossing van $Ax = b$ en de rij $\{x^k\}_{k=0}^{\infty}$ levert de gezochte benaderingen. De convergentie van deze rij $\{x^k\}_{k=0}^{\infty}$ hangt echter uiteraard van de gekozen splitsing $A = N - P$ af: voor een willekeurig gekozen splitsing zal er geen convergentie optreden en we zullen dus een verstandige splitsing moeten proberen te kiezen. De keuze van de splitsing wordt echter niet alleen beperkt door de eis van convergentie, maar ook nog door de eis dat de iteratiestappen goedkoop uit te rekenen zijn. Het zou bijvoorbeeld slecht zijn om de inverse van een volle matrix N expliciet te moeten bepalen: dat kost zelfs nog meer operaties dan Gauß-eliminatie gekost zou hebben. Een gunstig geval daarentegen treedt bijv. op wanner N een diagonaalmatrix is: dan kost iedere iteratiestap $\sim 2n^2$ bewerkingen. Voor grote n kan men dan dus $n/3$ maal (d.w.z. "vaak") itereren, voordat er evenveel werk gedaan is als er bij Gauß–eliminatie nodig zou zijn geweest.

Er zijn inderdaad splitsingen bekend, die aan beide vereisten voldoen, d.w.z. waarvoor het bijbehorende proces voor belangrijke klassen van matrices convergeert en waarvoor de iteratiestappen ook nog goedkoop zijn. We behandelen twee van die splitsingen, corresponderend met de zgn. de methode van Jacobi en de methode van Gauß–Seidel.

Als voorbereiding bekijken we, gemotiveerd door bovenstaande, het meer algemene problem van de convergentie van een rij $\{x^k\}_{k=0}^{\infty}$ die, voor $c \in \mathbb{R}^n$ (of $c \in \mathbb{C}^n$), gegeven wordt door zijn startwaarde x^0 en door een recursie van de vorm

$$x^{k+1} = Mx^k + c.$$

De matrix M heet om voor de hand liggende redenen de *iteratiematrix* van het proces.

Lemma 5.4.1. *Laat $M - \mathbf{I}$ regulier zijn. Dan zijn voor een proces van de vorm $x^{k+1} = Mx^k + c$ equivalent:*

1. *Er bestaat een c, zodanig dat voor iedere startwaarde x^0 de bijbehorende rij $\{x^k\}_{k=0}^{\infty}$ convergent is.*
2. *Voor alle c is voor iedere startwaarde x^0 de bijbehorende rij $\{x^k\}_{k=0}^{\infty}$ convergent.*
3. *$\rho(M) < 1$.*

Wanneer hieraan voldaan is, dan convergeren de rijen onder (1) *en* (2) *alle naar de unieke oplossing x^{∞} van $x = Mx + c$.*

Bewijs.

$1 \Longrightarrow 3$ Kies een startwaarde x^0 en laat dan $x^\infty = \lim_{k\to\infty} x^k$. Blijkbaar is $x^\infty = Mx^\infty + c$, waaruit we zien dat $x^{k+1} - x^\infty = M(x^k - x^\infty)$. Deze vergelijking herhaaldelijk toepassend, zien we dat $x^{k+1} - x^\infty = M^{k+1}(x^0 - x^\infty)$. Hierin gaat het linkerlid naar 0, dus blijkbaar is $\lim_{k\to\infty} M^{k+1}(x^0 - x^\infty) = 0$.

Merk vervolgens op dat $(\mathbf{I} - M)$ inverteerbaar is, zodat $x^\infty = (\mathbf{I} - M)^{-1}c$. Blijkbaar hangt x^∞ niet van de keuze van x^0 af, en dus kan $x^0 - x^\infty$ alle waarden aannemen door x^0 geschikt te kiezen.

We concluderen dat $\lim_{k\to\infty} M^{k+1}y = 0$ voor all y, en volgens Stelling A.3.12 impliceert dit dat $\rho(M) < 1$.

$3 \Longrightarrow 2$ Fixeer c en kies een startwaarde x_0. Men ziet eenvoudig dat $x^{k+1} = M^{k+1}x^0 + \sum_{i=0}^{k} M^i c$. Er geldt dat $\lim_{k\to\infty} M^{k+1}x^0 = 0$ volgens Stelling A.3.12. Merk verder op dat $(\mathbf{I} - M)\sum_{i=0}^{k} M^i c = (\mathbf{I} - M^{k+1})c$, zodat $\sum_{i=0}^{k} M^i c = (\mathbf{I} - M)^{-1}c - (\mathbf{I} - M)^{-1}M^{k+1}c$. Blijkbaar is, wederom volgens Stelling A.3.12, $\lim_{k\to\infty} \sum_{i=0}^{k} M^i c = (\mathbf{I} - M)^{-1}c$. We concluderen dat $\lim_{k\to\infty} x^{k+1}$ inderdaad bestaat (en gelijk is aan $(\mathbf{I} - M)^{-1}c$).

$2 \Longrightarrow 1$ Triviaal.

De toevoeging is duidelijk uit het overige deel van het bewijs. □

Voor dit proces hebben we de volgende foutschattingen.

Lemma 5.4.2. *Laat $\|\,.\,\|$ een norm op $\mathbb{R}^n$ (of $\mathbb{C}^n$) zijn, zodanig dat $\|M\| < 1$. Laat c willekeurig zijn en kies een startwaarde x^0 voor het proces gegeven door $x^{k+1} = Mx^k + c$. Dan convergeert de rij $\{x^k\}_{k=0}^{\infty}$ naar de unieke oplossing x^∞ van $x = Mx + c$, en bovendien:*

1. $\|x^k - x^\infty\| \le \|M\|\|x^{k-1} - x^\infty\|$;
2. $\|x^k - x^\infty\| \le \frac{\|M\|}{1-\|M\|}\|x^k - x^{k-1}\|$;
3. $\|x^k - x^\infty\| \le \frac{\|M\|^k}{1-\|M\|}\|x^1 - x^0\|$.

Bewijs. We weten dat $\rho(M) \le \|M\|$. Blijkbaar is $\rho(M) < 1$ en i.h.b. is 1 dus geen eigenwaarde van M. Lemma 5.4.1 is dus van toepassing en levert de convergentie van de rij. Voor de foutschattingen merken we allereerst op dat $x^k - x^\infty = M(x^{k-1} - x^\infty)$, hetgeen onmiddellijk de eerste schatting geeft. Voor de tweede schatting merken we op dat

$$\|x^k - x^\infty\| \le \|M\|\|x^{k-1} - x^\infty\| \le \|M\|\|x^{k-1} - x^k\| + \|M\|\|x^k - x^\infty\|,$$

waaruit de tweede schatting volgt. Uit $x^k - x^{k-1} = M(x^{k-1} - x^{k-2}) = \cdots = M^{k-1}(x^1 - x^0)$ zien we dat $\|x^k - x^{k-1}\| \le \|M\|^{k-1}\|x^1 - x^0\|$, zodat de derde schatting uit de tweede volgt. □

Opmerking 5.4.3. De foutschattingen in Lemma 5.4.2 hebben ieder een aparte waarde en toepassing. De derde, meer van theoretisch belang, laat zien dat het proces convergent is van orde tenminste 1. Het is een a priori schatting. Van groter praktisch belang zijn de a posteriori schattingen onder (1) en (2). Veronderstel immers dat we, om convergentie van het proces na te gaan, hebben weten vast te stellen dat $\|M\| \le \theta < 1$, waarbij θ expliciet bekend is. Dan geldt (ga na) op basis van Lemma 5.4.2 dat

$$\|x^k - x^\infty\| \le \frac{\theta}{1-\theta}\|x^k - x^{k-1}\| \tag{5.4.2}$$

en

$$\|x^k - x^\infty\| \leq \frac{\theta^k}{1-\theta}\|x^1 - x^0\|. \tag{5.4.3}$$

De schatting in (5.4.3) geeft ons, na berekening van x^1, de mogelijkheid om te beslissen hoeveel iteraties er nodig zijn om een voorgegeven nauwkeurigheid met zekerheid te bereiken. Het rechterlid in (5.4.2) wordt gedurende het iteratieproces expliciet berekenbaar en geeft dus voortdurend een nieuwe bovengrens voor de fout. Deze bovengrens wordt hierbij steeds kleiner, daar $\|x^k - x^{k-1}\| \leq \|M\|\|x^{k-1} - x^{k-2}\| < \|x^{k-1} - x^{k-2}\|$. Uiteraard geeft (5.4.3) ook nog een bovengrens voor de fout. Uit $x^k - x^{k-1} = M(x^{k-1} - x^{k-2}) = \cdots = M^{k-1}(x^1 - x^0)$ volgt $\|x^k - x^{k-1}\| \leq \theta^{k-1}\|x^1 - x^0\|$, zodat de bovengrens in (5.4.2) nooit groter zal zijn dan die in (5.4.3). Met andere woorden: voor het geven van een zo klein mogelijke bovengrens voor de fout gebruikt men bij voorkeur (5.4.2), en wel voor een zo groot mogelijke index k.

We combineren het voorgaande nu in de volgende basisstelling voor iteratieve methoden.

Stelling 5.4.4. *Laat A een reguliere matrix zijn, en zij $A = N - P$ een splitsing met N regulier. Beschouw voor $b \in \mathbb{R}^n$ (of $b \in \mathbb{C}^n$) het iteratieve proces*

$$x^{k+1} = N^{-1}Px^k + N^{-1}b.$$

Dan zijn equivalent

1. *Er bestaat een b, zodanig dat voor iedere startwaarde x^0 de bijbehorende rij $\{x^k\}_{k=0}^\infty$ convergent is.*
2. *Voor alle b is voor iedere startwaarde x^0 de bijbehorende rij $\{x^k\}_{k=0}^\infty$ convergent.*
3. *$\rho(N^{-1}P) < 1$.*

Wanneer hieraan voldaan is, dan convergeren de rijen onder (1) *en* (2) *alle naar de unieke oplossing x^∞ van $Ax = b$.*

Veronderstel dat $\|\,.\,\|$ een norm op $\mathbb{R}^n$ (of $\mathbb{C}^n$) is, zodanig dat $\|N^{-1}P\| \leq \theta < 1$. Dan is voldaan aan de conditie $\rho(N^{-1}P) < 1$ onder (3) *hierboven, en voor de convergentie naar de oplossing geldt:*

$$\|x^k - x^\infty\| \leq \frac{\theta}{1-\theta}\|x^k - x^{k-1}\| \quad (\text{“beste bovengrens voor de fout”})$$

en

$$\|x^k - x^\infty\| \leq \frac{\theta^k}{1-\theta}\|x^1 - x^0\| \quad (\text{“benodigde aantal iteraties”}).$$

Bewijs. Merk op dat $N^{-1}P - \mathbf{I}$ regulier is. Wanneer immers $(N^{-1}P - \mathbf{I})x = 0$, volgt dat $(P - N)x = 0$, d.w.z. $Ax = 0$. A is echter regulier verondersteld, dus $x = 0$. Het eerste deel van de stelling volgt dus uit Lemma 5.4.1, en de foutschattingen volgen uit Lemma 5.4.2 en Opmerking 5.4.3. □

De methode van Jacobi verkrijgt men, voor een reguliere matrix A, door het kiezen van de splitsing $A = N - P$, met $N = D$ (het diagonaaldeel van A) en $P = -L - U$, waarbij L (resp. U) het strikte onderdriehoeksgedeelte (resp. het strikte bovendriehoeksgedeelte) van A is[2]. Hierbij willen we, vanwege het algemene stramien, dat N regulier is, d.w.z. we moeten veronderstellen dat $a_{ii} \neq 0$ voor $i = 1, \ldots, n$.

[2]NB: dit zijn niet de L en U uit een eventuele LU-ontbinding van A!

Voorbeeld 5.4.5. Als

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \\ 3 & 4 & 1 & 2 \\ 4 & 1 & 2 & 3 \end{pmatrix},$$

dan is—zo blijkt—A regulier en uiteraard heeft A geen nul op de diagonaal. De decompositie voor de methode van Jacobi is hier dan $A = N - P$ met

$$N = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \text{ en } P = -\begin{pmatrix} 0 & 2 & 3 & 4 \\ 2 & 0 & 4 & 1 \\ 3 & 4 & 0 & 2 \\ 4 & 1 & 2 & 0 \end{pmatrix}.$$

De spectraalstraal van de iteratiematrix $M = -N^{-1}P$ blijkt hier 5.38 (in 3 cijfers precisie) te zijn, zodat de methode voor deze matrix A helaas niet convergeert.

Volgens het algemene stramien wordt de iteratiestap in de methode van Jacobi, voor gegeven b:

$$x^{k+1} = -D^{-1}(L+U)x^k + D^{-1}b.$$

We zouden nu, met Stelling 5.4.4 in het achterhoofd, bij voorkeur voor algemene A een algemene uitspraak over $\rho(-D^{-1}(L+U)) = \rho(D^{-1}(L+U))$ willen doen, die ons in staat stelt om eenvoudig convergentie of divergentie van de methode uit A af te lezen. Dat is echter in die algemeenheid niet haalbaar. Wel zullen we uiteindelijk eenvoudig convergentie kunnen concluderen voor de klasse van matrices in de volgende definitie.

Definitie 5.4.6. Laat A een complexe $n \times n$-matrix zijn. Dat heet A *strikt diagonaal dominant* als $|a_{ii}| > \sum_{j=1,j\neq i}^{n} |a_{ij}|$ voor $i = 1, \ldots, n$.

Met andere woorden: de diagonaalelementen zijn overheersend t.o.v. de andere coëfficiënten in dezelfde rij. Om te onthouden of het nu om kolommen of om rijen gaat in de definitie, kan men als ezelsbruggetje hanteren dat een strikt diagonaal dominante matrix op een manier op een vector werkt die "lijkt" op de actie van alleen zijn diagonaal.

Voorbeeld 5.4.7. De matrix

$$A = \begin{pmatrix} 4 & 3 \\ 1 & 2 \end{pmatrix}$$

is strikt diagonaal dominant. De matrix A^t is dat niet.

Lemma 5.4.8.

1. *Als A strikt diagonaal dominant is, dan is A regulier en heeft A geen nul op de diagonaal.*
2. *Als A^t strikt diagonaal dominant is, dan is A regulier en heeft A geen nul op de diagonaal.*

Bewijs. We bewijzen het eerste deel; het tweede deel volgt daaruit. Uit de definitie volgt onmiddellijk dat $|a_{ii}| > 0$ voor alle i. Wat betreft de regulariteit: veronderstel dat $Ax = 0$. Kies i zodanig dat $|x_i| = \|x\|_\infty$. Dan is i.h.b. $(Ax)_i = 0$, d.w.z. $a_{ii}x_i = -\sum_{j=1,j\neq i}^{n} a_{ij}x_j$. Hieruit volgt, door het nemen van absolute waarden, dat $|a_{ii}|\|x\|_\infty \leq \sum_{j=1,j\neq i}^{n} |a_{ij}||x_j| \leq \sum_{j=1,j\neq i}^{n} |a_{ij}|\|x\|_\infty$. Indien $\|x\|_\infty \neq 0$, dan leidt dit tot de tegenspraak $|a_{ii}| \leq \sum_{j=1,j\neq i}^{n} |a_{ij}|$. Dus $\|x\|_\infty = 0$ en blijkbaar $x = 0$. □

De methode van Jacobi is voor strikt diagonaal dominante matrices dus in ieder geval welgedefinieerd. De methode is dan zelfs convergent, blijkens de volgende stelling.

Stelling 5.4.9 (Voldoende voorwaarden voor convergentie van de methode van Jacobi). *Laat A een matrix zijn met geen enkele nul op de diagonaal. Schrijf $A = D + L + U$ waarbij D het diagonaaldeel van A is, en L (resp. U) het strikte onderdriehoeksgedeelte (resp. het strikte bovendriehoeksgedeelte) van A is. Definieer, voor gegeven b en startwaarde x^0, de rij $\{x^k\}_{k=0}^{\infty}$ conform de methode van Jacobi door*

$$x^{k+1} = -D^{-1}(L+U)x^k + D^{-1}b.$$

Veronderstel dat A strikt diagonaal dominant is. Dan geldt het volgende:

1. *De rij $\{x^k\}_{k=0}^{\infty}$ convergeert naar de oplossing x^{∞} van $Ax = b$;*
2. *Laat*

$$\theta = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|}.$$

Dan is $\theta < 1$, en de volgende foutschattingen gelden:

$$\|x^k - x^{\infty}\|_{\infty} \leq \frac{\theta}{1-\theta}\|x^k - x^{k-1}\|_{\infty} \quad (\text{“beste bovengrens voor de fout”})$$

en

$$\|x^k - x^{\infty}\|_{\infty} \leq \frac{\theta^k}{1-\theta}\|x^1 - x^0\|_{\infty} \quad (\text{“benodigde aantal iteraties”}).$$

Bewijs. De iteratiematrix is hier $-D^{-1}(L+U)$. Volgens Propositie A.3.6 is $\|-D^{-1}(L+U)\|_{\infty} = \theta$. Omdat $\theta < 1$ als gevolg van de aanname dat A strikt diagonaal dominant is, volgt de stelling dus uit Stelling 5.4.4. □

Voorbeeld 5.4.10. Beschouw het stelsel

$$\begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \end{pmatrix},$$

waarvan de exacte oplossing $x^{\infty} = (\frac{4}{5}, \frac{3}{5})^t$ is. De betreffende matrix A is strikt diagonaal dominant, dus volgens Stelling 5.4.9 is de methode van Jacobi hier convergent. We zullen drie iteraties van de methode toepassen en de foutschattingen uitwerken. Schrijf

$$\begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix} = A = N - P = D - (-L - U) = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix} - \begin{pmatrix} 0 & -1 \\ -2 & 0 \end{pmatrix}.$$

De iteratiematrix M is hier dus

$$M = -\begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{3} \\ \frac{1}{2} & 0 \end{pmatrix},$$

zodat $\theta = \frac{1}{2}$. De vector $D^{-1}b$ is hier

$$\begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

De iteratiestap ziet er dus als volgt uit:

$$\begin{pmatrix} x_1^{k+1} \\ x_2^{k+1} \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{3} \\ \frac{1}{2} & 0 \end{pmatrix}\begin{pmatrix} x_1^{k} \\ x_2^{k} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We kiezen als startwaarde $x^0 = 0$ en vinden dan achtereenvolgens $x^1 = (1,1)^t$, $x^2 = (\frac{2}{3}, \frac{1}{2})^t$ en $x^3 = (\frac{5}{6}, \frac{2}{3})^t$. We zien dat $x^1 - x^0 = (1,1)^t$ en $x^3 - x^2 = (\frac{1}{6}, \frac{1}{6})^t$. De werkelijke fout na de derde iteratie kunnen we hier uiteraard uitrekenen: $x^3 - x^\infty = (\frac{1}{30}, \frac{1}{15})^t$. We gaan na wat de foutschattingen uit de stelling opleveren:

- Allereerst is volgens de stelling:

$$\|x^3 - x^\infty\|_\infty \leq \frac{\frac{1}{2}}{1-\frac{1}{2}}\|x^3 - x^2\|_\infty = 1 \cdot (\frac{1}{6} + \frac{1}{6}) = \frac{1}{3}.$$

  In werkelijkheid is $\|x^3 - x^\infty\|_\infty = \frac{1}{30} + \frac{1}{15} = \frac{1}{10}$.

- Wanneer men een fout van bijv. ten hoogste $\frac{1}{1000}$ in de supremumnorm toe wil staan, d.w.z. $\|x^k - x^\infty\|_\infty \leq \frac{1}{1000}$ wil garanderen, dan zal daaraan zeker voldaan zijn indien

$$\frac{(\frac{1}{2})^k}{1-\frac{1}{2}}\|x^1 - x^0\|_\infty = 2(\frac{1}{2})^{k-1} \cdot 2 \leq \frac{1}{1000},$$

  d.w.z. als $k \geq 13$.

De Gauß–Seidel-methode kan men heuristisch motiveren vanuit die van Jacobi, als volgt. Voor de methode van Jacobi ziet de iteratie er per coördinaat als volgt uit:

$$x_i^{k+1} = \frac{1}{a_{ii}}\left\{ -\sum_{j=1, j\neq i} a_{ij}x_j^k + b_i \right\}. \tag{5.4.4}$$

Wanneer—voor een of ander iteratief proces—de rij x^k naar de oplossing van het stelsel $Ax = b$ convergeert, dan kan men hopen dat het een goed idee is om de coördinaten van het nieuw berekende element x^{k+1} al te gebruiken, zodra die in de iteratiestap, die van x^k naar x^{k+1} leidt, beschikbaar zijn gekomen. Die waarden van de coördinaten van x^{k+1} worden immers verhoopt "beter" te zijn dan de coördinaten van x^k, dus hoe eerder die gebruikt worden, hoe beter het waarschijnlijk is. Wanneer men deze gedachte toepast op de methode van Jacobi, dan wordt de bovenstaande uitdrukking gewijzigd in

$$x_i^{k+1} = \frac{1}{a_{ii}}\left\{ -\sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + b_i \right\}, \tag{5.4.5}$$

waarbij dus t.o.v. (5.4.4) de x_j^k voor $j < i$ (die immers al bekend zijn bij berekening van x_i^{k+1}) alvast zijn vervangen door de x_j^{k+1}. Het iteratief proces dat door (5.4.5) beschreven wordt, heet de methode van Gauß–Seidel. De achterliggende heuristiek dient ter motivatie van deze methode (of als ezelsbruggetje), maar is helaas niet meer dan dat: er zijn gevallen waarin de methode van Jacobi convergeert, maar die van Gauß–Seidel niet. Het omgekeerde komt ook voor.

Wanneer we weer $A = D + L + U$ schrijven, dan laat (5.4.5) zich uitdrukken als

$$x^{k+1} = D^{-1}(-Lx^{k+1} - Ux^k + b), \tag{5.4.6}$$

d.w.z.

$$(D+L)x^{k+1} = -Ux^k + b,$$

ofwel

$$x^{k+1} = -(D+L)^{-1}Ux^k + (D+L)^{-1}b. \tag{5.4.7}$$

We zien—dat was niet op voorhand duidelijk—dat deze methode zich eveneens laat formuleren in termen van een splitsing $A = N - P$ als boven, en wel met $N = D + L$ en $P = -U$. Het proces is gedefinieerd zodra A geen nul op de diagonaal heeft, en er treedt blijkbaar convergentie op, dan en slechts dan, als $\rho(-(D+L)^{-1}U) = \rho((D+L)^{-1}U) < 1$.

Opmerking 5.4.11. Waarschuwing: hoewel (5.4.7) de schrijfwijze van het proces is, die duidelijk maakt welke spectraalstraal convergentiebepalend is, is dit *niet* de wijze waarop de iteratiestap in de praktijk berekend moet worden. Berekening via (5.4.7) vooronderstelt immers het inverteren van $D+U$, wat $\sim n^3$ operaties kost. Voor het berekenen van de iteratiestap gebruikt men derhalve (5.4.5).

De analyse van de Gauß–Seidel-methode is wat lastiger dan die van de methode van Jacobi. Het bewijs van de volgende stelling laten we daarom achterwege.

Stelling 5.4.12 (Voldoende voorwaarden voor convergentie van de methode van Gauß–Seidel). *Definieer, voor een algemene matrix A met geen enkele nul op de diagonaal, een gegeven b en een startwaarde x_0, de rij $\{x_k\}_{k=0}^{\infty}$ conform de methode van Gauß–Seidel door*

$$x_i^{k+1} = \frac{1}{a_{ii}} \left\{ -\sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + b_i \right\}.$$

Indien

1. *A strikt diagonaal dominant is, en/of*
2. *A strikt positief definiet is,*

dan convergeert de rij $\{x_k\}_{k=0}^{\infty}$ naar de oplossing x^{∞} van $Ax = b$.

Voorbeeld 5.4.13. Voor de (strikt positief definiete) matrix

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

en de vector $b = (2,2,2)^t$ wordt de iteratiestap

$$\begin{aligned} x_1^{k+1} &= \frac{1}{2}(\quad x_2^k \quad + 2) \\ x_2^{k+1} &= \frac{1}{2}(x_1^{k+1} \quad + x_3^k + 2) \\ x_3^{k+1} &= \frac{1}{2}(\quad x_2^{k+1} \quad + 2). \end{aligned}$$

Men berekent dan, uitgaande van de startwaarde $x^0 = 0$, achtereenvolgens $x^1 = (1, \frac{3}{2}, \frac{7}{4})^t$, $x^2 = (\frac{7}{4}, \frac{11}{4}, \frac{19}{8})^t$, $x^3 = (\frac{19}{8}, \frac{27}{8}, \frac{43}{16})$, $x^4 = (\frac{43}{16}, \frac{59}{16}, \frac{91}{32})$ en $x^5 = (\frac{91}{32}, \frac{123}{32}, \frac{187}{64})^t$. Vergelijk dit met de exacte oplossing van $Ax = b$, nl. $(3,4,3)^t$.

Opmerking 5.4.14. De convergentie van de Gauß–Seidel-methode voor strikt positieve matrices is een belangrijk gegeven, want deze klasse van matrices komt vaak voor bij het numeriek oplossen van differentiaalvergelijkingen. Er is in dat geval nog een uitbreiding van de theorie mogelijk, die bekend staat als de methode van "herhaalde overrelaxatie". Het idee is om de vrijheid in het kiezen van een splitsing uit te buiten, door te werken met een geparametriseerde splitsing $A = N(\omega) - P(\omega)$. Door het kiezen van een geschikte waarde voor ω kan men de convergentiesnelheid dan vaak nog aanmerkelijk verhogen.

Opmerking 5.4.15. Laten we, voor de startwaarde $x^0 = 0$, eens proberen om voor een convergent iteratief proces van de vorm

$$x^{k+1} = Mx^k + c.$$

(met dus $\rho(M) < 1$) een indruk te krijgen van het benodige aantal iteraties waarmee men een nauwkeurigheid verkrijgt die in dezelfde orde ligt als de inherente machine(on)nauwkeurigheid, waarmee ook de exacte oplossing gerepresenteerd zou worden. Beter dan dat laatste is überhaupt niet haalbaar, dus een iteratieve methode die dat met dezelfde hoeveelheid werk eveneens weet te bereiken, is concurrerend.

We nemen hierbij aan—dat is bijvoorbeeld zowel bij de methode van Jacobi als die van Gauß–Seidel het geval—dat iedere iteratiestap $\sim 2n^2$ operaties vergt. We nemen eveneens aan, dat de norm $\|\,.\,\|$ zodanig is dat $\|M\| < 1$.[3] Voor de startwaarde $x^0 = 0$ ziet men dat $x^k - x^\infty = M^k(x^0 - x^\infty) = -M^k x^\infty$, zodat we voor de relatieve fout na k iteraties de afschatting

$$\frac{\|x^k - x^\infty\|}{\|x^\infty\|} \leq \|M\|^k$$

hebben. Laten we aannemen dat we werken op een machine met 16 cijfers precisie. De exacte oplossing wordt in de machine gerepresenteerd als $\tilde{x}^\infty$. Er zijn constanten C_1 en C_2 z.d.d. $C_1\|x\| \leq \|x\|_2 \leq C_2\|x\|$ voor alle x; dit gebruikend zien we dat voor de relatieve fout

$$\frac{\|\tilde{x}^\infty - x^\infty\|}{\|x^\infty\|} \leq \frac{C_2}{C_1}\frac{\|\tilde{x}^\infty - x^\infty\|_2}{\|x^\infty\|_2} \overset{\text{ga na}}{\leq} \frac{C_2}{C_1} \cdot 5 \cdot 10^{-16}$$

geldt. Wanneer men dit ook voor de iteratieve methode wil kunnen garanderen, dan zal men dus k zo groot kiezen dat

$$\|M\|^k \leq \frac{C_2}{C_1} \cdot 5 \cdot 10^{-16}.$$

Hieruit volgt een minimaal benodigd aantal iteraties; de hoeveelheid operaties per iteratiestap in aanmerking nemend zien we dat de iteratieve methode de nauwkeurigheid van de representatie van de exacte oplossing zeker evenaart in

$$\sim 2 \cdot \frac{\log(5C_2 \cdot 10^{-16}) - \log C_1}{\log \|M\|} \cdot n^2.$$

operaties.

Interessant genoeg is deze uitdrukking kwadratisch in n, terwijl de hoeveelheid werk voor het uitvoeren van een algemene Gauß-eliminatie kubisch in n is. De iteratieve methode wordt dus op een zeker moment zelfs aantrekkelijker! Waar ligt het omslagpunt? Het bepalen van een exacte

[3] Indien het proces convergeert, d.w.z. wanneer $\rho(M) < 1$, dan—maar dat is niet zonder meer duidelijk—bestaat een dergelijke norm ook inderdaad.

oplossing met Gauß-eliminatie kost $\sim \frac{2}{3}n^3$ operaties. De lagere orde termen verwaarlozend, vinden we dus als eindconclusie dat voor

$$n \geq \frac{3\log(5C_2 \cdot 10^{-16}) - 3\log C_1}{\log \|M\|}$$

de iteratieve methode dezelfde nauwkeurigheid bereikt, die voor de exacte oplossing haalbaar is, maar wel met *minder* werk dan het "bepalen" van die exacte oplossing met een Gauß-eliminatie zou kosten. Werkend in de norm $\|\,.\,\|_2$ (zodat $C_1 = C_2 = 1$)) is dit bijvoorbeeld het geval zodra

$$n \geq \frac{-106}{\log \|M\|_2}. \tag{5.4.8}$$

Dit geeft aan dat een iteratieve methode ook voor relatief kleine stelsels al lonend kan zijn: wanneer bijv. $\|M\|_2 = \frac{1}{2}$, dan is aan (5.4.8) al voldaan voor $n \geq 153$.

Bedenk overigens wel, dat we er in bovenstaande berekening van uit zijn gegaan dat Gauß-eliminatie $\sim \frac{2}{3}n^3$ operaties kost, d.w.z. dat we werkten met het meest algemene geval. Voor belangrijke speciale klassen van matrices (bijv. bandmatrices) is het aantal operaties voor de (compacte) directe methoden veel kleiner, bijv. lineair in n. Daarvoor gaat bovenstaande berekening dan niet op en de directe methoden blijven voor die klassen superieur.

Hoofdstuk 6

Eindige elementen methode

In het begin van dit college hebben we als motiverende leidraad ons de vraag gesteld "hoe we een differentiaalvergelijking numeriek kunnen oplossen". Die vraag is niet een-twee-drie te beantwoorden, gezien de veelheid van methoden die er beschikbaar is, ieder weer met zijn eigen specifieke voor- en nadelen voor de verschillende typen van differentiaalvergelijkingen die in de praktijk voorkomen. Een keuze is noodzakelijk.

We zullen ons in deze sectie concentreren op één van de bestaande methodes, nl. op de zgn. *eindige elementen methode*. We behandelen een typische toepassing van die methode voor het numeriek oplossen van een randwaardeprobleem in één dimensie. De methode laat zich goed generaliseren naar hogere dimensies, en ook naar vaak voorkomende vergelijkingen, anders dan randwaardeproblemen. Hoewel we op de keper beschouwd slechts één type gewone differentiaalvergelijkingen bekijken, staat de aanpak in dat specifieke geval dus model voor de aanpak in een hele familie van verwante problemen. Het stramien van de eindige elementen methode blijft in die familie op hoofdlijnen steeds hetzelfde, zij het dat de techniek vaak wat gecompliceerder wordt.

Deze eindige elementen methode heeft een grote praktische waarde, omdat hij toepasbaar is op allerlei vergelijkingen die in de praktijk nu eenmaal vaak voorkomen. Grofweg gezegd komt dat, doordat vele fysische vergelijkingen lineair zijn en, als door een merkwaardig soort toeval, vaak de Laplaciaan in die vergelijkingen voorkomt. Dergelijke vergelijking liggen nu eenmaal vaak binnen het toepassingsgebied van de eindige elementen methode. Voor een algemene, mogelijk ook niet-lineaire vergelijking, zal men andere methoden moeten gebruiken, bijv. *eindige differentie methodes*. We gaan daar hier niet verder op in. De numerieke analyse van differentiaalvergelijkingen is een discipline op zich, met sterke banden met de abstracte algemene theorie van differentiaalvergelijkingen en met de functionaalanalyse.

We zullen in dit hoofdstuk gebruik maken van de theorie uit Bijlage A.5.

6.1 Het randwaardeprobleem

Het probleem waar we naar zullen kijken is het volgende.

$$\begin{cases} -(au')' = f \\ u(0) = u(1) = 0 \end{cases} \tag{6.1.1}$$

Hierbij is $a \in C^2[0,1]$ en $f \in C[0,1]$; gezocht wordt een oplossing $u \in C^2[0,1]$. Het probleem heeft uiteraard een analogon op een algemeen interval $[c,d]$, dat door translatie en schaling vertaald kan worden naar een probleem op $[0,1]$ van bovenstaand type.

Het randwaardeprobleem (6.1.1) komt in de natuurkunde naar voren in zowel de warmteleer als in de elasticiteitsleer. In het eerste geval is a dan een warmtegeleidingscoëfficiënt en beschrijft f een extern aangebrachte energieflux. Fysisch is het in die gevallen duidelijk dat er in dergelijke gevallen een unieke oplossing moet zijn, en men kan dit ook wiskundig aantonen. Zonder bewijs vermelden we:

Stelling 6.1.1. *Als a strikt positief is op $[0,1]$, dan heeft probleem (6.1.1) een unieke oplossing.*

We zullen dan ook steeds aannemen dat

$$\alpha_1 \leq a(x) \leq \alpha_2 \quad \text{voor alle } x \in [0,1], \tag{6.1.2}$$

waarbij $\alpha_1 > 0$.

6.2 Hilbertruimte-benadering: zwakke formulering

We willen nu de theorie in Paragraaf A.5 gaan gebruiken om een numeriek proces te definiëren dat convergeert naar de oplossing u van (6.1.1); zie Stelling A.5.14 voor een samenvatting van deze zgn. Galerkin-methode. Het zal uit Stelling A.5.14 duidelijk zijn, dat we dan op zoek moeten naar een typering van onze oplossing u in termen van een geschikte Hilbertruimte V, een begrensde elliptische vorm a op $V \times V$ en een element $l \in V'$. Een dergelijke typering, zoals we die hieronder zullen geven, is voor een grote klasse van vergelijkingen mogelijk (ook in hogere dimensie) en staat bekend als de *zwakke formulering*[1] (ook wel *variationele formulering*) van het betreffende probleem. De oorspronkelijke formulering (6.1.1) heet wel de *klassieke formulering*.

De Galerkin-methode veronderstelt ook nog de keuze van eindigdimensionale deelruimten van V. De speciale keuze die we daarvoor zullen maken behandelen we in de de hierop volgende paragraaf; voorlopig concentreren we ons op het vinden van V, een begrensde elliptische vorm a op $V \times V$ en een element $l \in V'$ waarmee we de zwakke formulering van (6.1.1) kunnen geven.

De Hilbertruimte V waarin we zullen werken is de volgende. Als verzameling bestaat V uit alle functies $f : [0,1] \to \mathbb{R}$, die voldoen aan:

1. $f \in C[0,1]$;
2. $f(0) = f(1) = 1$;
3. f is "bijna overal" differentieerbaar;
4. $\int_0^1 (f'(x))^2\,dx < \infty$.

De vectorruimtestructuur wordt gegeven door puntsgewijze operaties. Als inwendig produkt op V nemen we:

$$(f,g) = \int_0^1 f(x)g(x) + f'(x)g'(x)\,dx.$$

Men kan laten zien (we zullen dit niet bewijzen) dat V met dit inwendig produkt een Hilbertruimte is[2].

De term "bijna overal" in bovenstaande definite heeft binnen de maattheorie een precieze betekenis, waarop we hier niet in detail kunnen ingaan. Het is in ieder geval zo, dat functies

[1] De wellicht bevreemdende term "zwak" stamt uit de functionaalanalyse, waar het een vaker gebruikt adjectief is.

[2] Het inwendig produkt op V is een voorbeeld van een zgn. Sobolev-inprodukt. De ruimte V staat bekend als de Sobolevruimte $H_0^1(0,1)$.

die op eindig veel punten na differentieerbaar zijn (bijv. stuksgewijs polynomiale functies), in maattheoretische zin bijna overal differentieerbaar zijn. Dergelijke functies (en "erger dan dat" zullen we ze niet tegenkomen) zijn dus in ieder geval elementen van V. Onze oplossing $u \in C^2[0,1]$ van ons randwaardeprobleem (6.1.1) is zeker ook een element van V.

Voor $f \in V$ geldt (we zullen dit niet bewijzen) de hoofdstelling van de integraalrekening, d.w.z. dat $f(x) = \int_0^x f'(t)\,dt$ voor $x \in [0,1]$ (merk op dat $f(0) = 0$), waarbij men dan $f'(t) = 0$ definieert voor die punten t waarin f niet differentieerbaar is.

We zullen, zoals al gezegd, naderhand de Galerkin-methode in V toepassenen. Dit levert ons dan een rij in V op die in de norm $\|\,.\,\|$ van V naar de oplossing u van (6.1.1) convergeert. Visueel (en ook fysisch) zal men echter eerder geïnteresseerd zijn in puntsgewijze convergentie naar u, of zelfs in *uniforme* convergentie, d.w.z. convergentie in de norm $\|\,.\,\|_\infty$. Het blijkt nu, dat convergentie in V inderdaad uniforme convergentie impliceert. Immers, voor $f \in V$ en $x \in [0,1]$ is (we maken gebruik van de Schwartz-ongelijkheid in $L_2([0,1], dx)$)

$$\begin{aligned} |f(x)| = |\int_0^x f'(t)\,dt| \le \int_0^x |f'(t)|\,dt \le \int_0^1 |f'(t)|\,dt = \int_0^1 |f'(t)| \cdot 1\,dt \qquad (6.2.1)\\ \overset{\text{Schwartz}}{\le} \left\{\int_0^1 |f'(t)|^2\,dt\right\}^{\frac{1}{2}} \cdot \left\{\int_0^1 1^2\,dt\right\}^{\frac{1}{2}} \le \|f\|. \end{aligned}$$

Dit geldt voor alle $x \in [0,1]$, dus $\|f\|_\infty \le \|f\|$ voor $f \in V$. We concluderen hieruit dat een convergente rij in V inderdaad ook automatisch uniform convergeert.

De ongelijkheden in (6.2.1) hebben nog een ander gevolg. Beschouw de bilineaire vorm op V die gegeven wordt door

$$(f,g)' = \int_0^1 f'(x)g'(x)\,dx \quad (f,g \in V).$$

Deze vorm is in feite een inwendig produkt op V. Als $(f,f)' = 0$, dan is namelijk $f' = 0$, dus f is constant[3]. Daar echter $f(0) = 0$ is blijkbaar zelfs $f = 0$. We zien hieruit dat $(\,.\,,\,.\,)$ inderdaad een inwendig produkt op V is.

Laten we de bij het inprodukt $(\,.\,,\,.\,)'$ op V behorende norm aangeven met $\|\,.\,\|'$. Het is duidelijk dat $\|f\|' \le \|f\|$ voor alle $f \in V$. Daar echter ook voor alle $f \in V$ geldt dat (zie (6.2.1) direct na toepassing van de Schwartz-ongelijkheid):

$$(f,f) = \int_0^1 f(x)^2 + f'(x)^2\,dx \le \int_0^1 \left\{\int_0^1 |f'(t)|^2\,dt\right\} + f'(x)^2\,dx = 2(f,f)',$$

is blijkbaar $\|f\| \le \sqrt{2}\|f\|'$. *We concluderen dat de normen* $\|\,.\,\|$ *en* $\|\,.\,\|'$ *op* V *equivalent zijn.*

Na deze beschrijving van de basiseigenschappen van de Hilbertruimte V zullen we nu aangeven wat de vorm a en $l \in V'$ zijn. Neem hiertoe $v \in V$ willekeurig en merk op dat uiteraard

$$-\int_0^1 (a(x)u'(x))'v(x)\,dx = \int_0^1 f(x)v(x)\,dx. \qquad (6.2.2)$$

Het linkerlid herschrijven we met partiële integratie als

$$-a(x)u'(x)v(x)|_0^1 + \int_0^1 a(x)u'(x)v'(x)\,dx \overset{v(0)=v(1)=0}{=} \int_0^1 a(x)u'(x)v'(x)\,dx.$$

[3] De exacte redenering op dit punt ligt maattheoretisch wat subtieler dan we hier aangeven.

We concluderen dat u voldoet aan:

$$\int_0^1 a(x)u'(x)v'(x)\,dx = \int_0^1 f(x)v(x)\,dx \text{ voor alle } v \in V. \tag{6.2.3}$$

Het ligt nu voor de hand wat onze a en l zullen worden:

$$a(v_1, v_2) = \int_0^1 a(x)v_1'(x)v_2'(x)\,dx \quad (v_1, v_2 \in V),$$
$$l(v) = \int_0^1 f(x)v(x)\,dx \quad (v \in V).$$

We verifiëren dat a en l aan de voorwaarden voor toepassing van de Galerkin-methode voldoen:

- De vorm a is duidelijk bilinear. Verder geeft onze aanname dat

  $$\alpha_1 \le a(x) \le \alpha_2 \quad \text{voor alle } x \in [0,1].$$

  voor zekere $\alpha_1 > 0$ ons de begrensdheid en ellipticiteit van a. Immers, gebruikmakend van de ongelijkheid van Schwartz in $L_2([0,1], dx)$ zien we dat

  $$\begin{aligned}|a(v_1, v_2)| &\le \int_0^1 |a(x)||v_1'(x)||v_2'(x)|\,dx \le \alpha_2 \int_0^1 |v_1'(x)||v_2'(x)|\,dx \\ &\overset{\text{Schwartz}}{\le} \alpha_2 \left\{\int_0^1 |v_1'|^2\,dx\right\}^{\frac{1}{2}} \left\{\int_0^1 |v_2'|^2\,dx\right\}^{\frac{1}{2}} \\ &= \alpha_2 \|v_1\|' \|v_2\|' \le \alpha_2 \|v_1\| \|v_2\|.\end{aligned}$$

  Inderdaad is a dus begrensd. Voor de ellipticiteit maken we gebruik van de afschatting $\|v\|' \ge \frac{1}{\sqrt{2}}\|v\|$ voor $v \in V$:

  $$a(v,v) = \int_0^1 a(x)v'(x)v'(x)\,dx \ge \alpha_1 \int_0^1 v'(x)v'(x)\,dx = \alpha_1(\|v\|')^2 \ge \frac{\alpha_1}{2}\|v\|^2.$$

  Inderdaad is a dus elliptisch.

- De vorm l is duidelijk lineair. Voor de begrensheid merken we op, alweer gebruikmakend van de ongelijkheid van Schwartz in $L_2([0,1], dx)$, dat:

  $$\begin{aligned}|l(v)| = |\int_0^1 f(x)v(x)\,dx| &\le \int_0^1 |f(x)v(x)|\,dx \le \|f\|_\infty \int_0^1 |v(x)|\,dx \\ &= \|f\|_\infty \int_0^1 |v(x)| \cdot 1\,dx \overset{\text{Schwartz}}{\le} \|f\|_\infty \left\{\int_0^1 |v(x)|^2\,dx\right\}^{\frac{1}{2}} \le \|f\|_\infty \|v\|.\end{aligned}$$

  Inderdaad is l dus begrensd, m.a.w. $l \in V'$.

We hebben nu de zwakke formulering van het randwaardeprobleem bereikt: *De oplossing u van het randwaardeprobleem* (6.1.1) *voldoet aan*

$$a(u, v) = l(v) \text{ voor alle } v \in V,$$

waarbij de begrensde elliptische vorm a gegeven wordt door

$$a(v_1, v_2) = \int_0^1 a(x)v_1'(x)v_2'(x)\,dx \quad (v_1, v_2 \in V),$$

en waarbij $l \in V'$ gegeven wordt door

$$l(v) = \int_0^1 f(x)v(x)\,dx \quad (v \in V).$$

Opmerking 6.2.1. We hebben tot nu toe geredeneerd *uitgaand* van de bestaande klassieke oplossing $u \in C^2[0,1]$ van (6.1.1). Deze u lost blijkbaar ook het boven geformuleerde zwakke probleem $a(u,v) = l(v)$ (voor alle $v \in V$) op. Zouden we de redenering ook om kunnen keren? Het Lax–Milgram-Lemma (Stelling A.5.8) garandeert ons immers dat er een u in V bestaat die voldoet aan $a(u,v) = l(v)$ (voor alle $v \in V$). Voldoet deze u dan misschien ook aan ons oorspronkelijke probleem (6.1.1), en kunnen we dan op deze manier het bestaan van een klassieke oplossing aantonen? Dat is inderdaad het geval, maar het is niet eenvoudig om dit direct in te zien. Het probleem zit hem in de gladheid van u. Een klassieke oplossing u is C^2, maar de elementen van V hoeven niet eens overal differentieerbaar te zijn, laat staan tweemaal continu differentieerbaar. Het cruciale extra ingrediënt is dan ook een zgn. regulariteitsstelling, die a priori garandeert dat een zwakke oplossing u in feite automatisch C^2 is, d.w.z. veel gladder is dan je eigenlijk mocht verwachten op grond van alleen het feit dat $u \in V$. Dit wetend is het verder niet lastig meer om in te zien dat een zwakke oplossing inderdaad in feite een klassieke oplossing is. Immers, het feit dat $a(u,v) = l(v)$ (voor alle $v \in V$) is niets anders dan een andere formulering van (6.2.3). Men kan dan (gebruikmakend van de gladheid van u!) de partiële integratie omkeren, en inzien dat u voldoet aan (6.2.2) (voor alle $v \in V$). Daaruit volgt dan weer dat $-(au')' = f$ (kies een klein "heuveltje" voor v rond een punt waar het linkerlid en het rechterlid eventueel niet aan elkaar gelijk zouden zijn). De zwakke formulering is daarmee dus, gebruikmakend van de regulariteitsstelling, equivalent met het randwaardeprobleem (6.1.1).

Bovenstaande omkering illustreert in feite een belangrijke moderne methode waarmee men, ook in hogere dimensies, de existentie van klassieke oplossingen van belangrijke differentiaalvergelijkingen kan aantonen. Men vertaalt het oorspronkelijke probleem in een zwakke formulering binnen een Hilbertruimte, waarbij men, gebruikmakend van een (i.h.a. niet-triviale) regulariteitsstelling, weet dat deze zwakke formulering in feite equivalent is met de oorspronkelijke vraagstelling. Het Lax–Milgram-Lemma laat dan vervolgens onmiddellijk zien dat zo'n zwakke oplossing in die Hilbertruimte inderdaad bestaat, waarmee de klassieke existentie dan is aangetoond.

6.3 Eindige elementen methode

In de vorige paragraaf hebben we het randwaardeprobleem (6.1.1) vertaald in een zwakke formulering ervan binnen een Hilbertruimte V. Daarmee is de Galerkin-methode, zoals die in Stelling A.5.14 samengevat staat, beschikbaar gekomen. Deze methode maakt gebruik van een stijgende rij eindigdimensionale lineaire deelruimte van V, en van bases in die lineaire deelruimten om concreet mee te rekenen. In de keuze van lineaire deelruimten is men vrij: iedere keuze van een stijgende rij gesloten lineaire deelruimten levert, wanneer aan een zekere dichtheidseis voldaan is, een convergent proces op. Na keuze van de lineaire deelruimten kan men in principe daarbinnen ook met alle mogelijke bases concreet rekenen. Er is dus nogal wat keuzevrijheid, maar de ene keuze is beter dan de andere. Een voor de hand liggende keuze zou bijvoorbeeld zijn, om te werken met de rij $P^0[0,1] \subset P^1[0,1] \subset P^2[0,1] \subset \cdots$, d.w.z. met polynomen van toenemende maximale graad. Daarin kan men dan bijv. $\{1, x, x^2, \ldots, x^q\}$ als basis van $P^q[0,1]$

nemen. Het blijkt echter, dat de bij deze keuze behorende stijfheidsmatrices dan grote conditiegetallen[4] kunnen krijgen, waardoor de numerieke fouten in de berekening van de belastingsvector onacceptabel kunnen gaan doorwerken in het eindresultaat. Liever neemt men dan ook lineaire deelruimten van continue stuksgewijs polynomiale functies (in ons geval: continue stuksgewijs lineaire functies) als uitgangspunt: hierbij treden deze grote conditiegetallen i.h.a. niet op. Deze keuze heeft overigens nog een paar andere belangrijke voordelen, waarvan later duidelijk zal worden wat dat zijn.

Laat daarom $0 = x_0 < x_1 < \ldots < x_{m+1} = 1$ een partitie $\mathcal{P}$ zijn van $[0,1]$, met bijbehorende maasfunctie h. Bij deze partitie introduceren we (deze notatie wijkt, kortheidshalve, enigszins af van die in Paragraaf 3.5) de vectorruimte V_h, die bestaat uit alle continue functies v op $[0,1]$ die lineair zijn op alle intervallen $[x_i, x_{i+1}]$ $(i = 0, \ldots, m)$, en die voldoen aan de randwaardecondities $v(0) = v(1) = 0$. Merk op dat $V_h \subset V$. Merk ook op, dat een verfijning $\widetilde{\mathcal{P}}$ van deze partitie $\mathcal{P}$, met bijbehorende maasfunctie $\tilde{h}$, een vectoruimte $V_{\tilde{h}}$ geeft z.d.d. $V_h \subset V_{\tilde{h}}$. Indien $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \cdots$ een rij van successievelijke verfijningen is, met bijbehorende vectorruimten V_{h_i}, dan geeft dit dus aanleiding tot een stijgende rij $V_{h_1} \subset V_{h_2} \subset \cdots$ van lineaire deelruimten van V. Zodra $\lim_{i\to\infty} \|h_i\|_\infty = 0$, d.w.z. zodra de maximum maaswijdte naar nul gaat, is verder $\overline{\bigcup_{i=1}^{\infty} V_{h_i}} = V$. Dit laatste is een standaard dichtheidsresultaat, dat we hier zonder bewijs zullen gebruiken. Een ruimte V_h als boven is eindigdimensionaal (en dus automatisch gesloten). Aan alle voorwaarden voor de abstracte theorie van de Galerkin-methode, zoals die staat samengevat in Stelling A.5.14, is dus voldaan.

Een basis van een ruimte V_h wordt gevormd door (vgl. Paragraaf 3.5) de dakfuncties ϕ_j^h $(j = 1, \ldots, m)$, gedefinieerd door

$$\phi_j^h(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & x \in [x_{j-1}, x_j]; \\ \frac{x - x_{j+1}}{x_j - x_{j+1}} & x \in [x_j, x_{j+1}]; \\ 0 & x \notin [x_{j-1}, x_{j+1}]. \end{cases}$$

Om de oplossing u van ons randwaarde probleem numeriek te benaderen hoeven we blijkbaar alleen de volgende stappen uit te voeren:

1. We kiezen een rij $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \cdots$ van verfijnende partities, met bijbehorende m_i-dimensionale vectorruimten V_{h_i}, zodanig dat $\lim_{i\to\infty} \|h_i\|_\infty = 0$.

2. We bepalen daarna voor iedere i, met behulp van de basis van dakfuncties $\{\phi_j^i\}_{j=1}^{m_i}$ van V_{h_i}, de stuksgewijs lineaire functie $u_{h_i} \in V_{h_i}$ die voldoet aan $a(u_{h_i}, v) = l(v)$ voor alle $v \in V_{h_i}$.

Automatisch geldt dan dat $u_{h_i} \to u$ in V, waarvan we al gezien hebben dat dit de meer aansprekende uniforme convergentie op $[0,1]$ impliceert.

Opmerking 6.3.1. De intervallen van de partities in het bovenstaande recept worden in dit verband wel "elementen" genoemd. Ook voor de bijbehorende dakfuncties in de bases wordt die term wel gebruikt; uit de context is altijd duidelijk wat er bedoeld wordt. Deze hele aanpak staat daarom bekend als de *eindige elementen methode*, waarmee dan bedoeld wordt de toepassing van de Galerkin-methode op de zwakke formulering van de differentiaalvergelijking, gebruikmakend van een opdeling van het onderliggende gebied in eenvoudige deelgebiedjes, en gebruikmakend van functies die op een eenvoudige manier in termen van die deelgebiedjes gedefinieerd zijn (in de praktijk vaak stuksgewijs polynomiaal).

[4] De precieze definitie van het conditiegetal van een matrix laten we achterwege.

Kunnen we nu ook iets zeggen over de snelheid van de convergentie $u_{h_i} \to u$? Volgens Stelling A.5.14 is er een constante $C \geq 0$ z.d.d.

$$\|u - u_{h_i}\| \leq C \inf_{v \in V_{h_i}} \|u - v\|.$$

Dit lijkt niet erg te helpen, omdat de precieze afstand $\inf_{v \in V_{h_i}} \|u - v\|$ van u tot V_{h_i}, zelfs als we u expliciet zouden kennen, i.h.a. niet uit te rekenen is. Resultaten uit de interpolatietheorie (die voor de foutschattingen bij de eindige elementen methode altijd een belangrijke rol spelen), helpen ons hier echter verder. Immers, in ieder geval is

$$\inf_{v \in V_{h_i}} \|u - v\| \leq \|u - \pi_{h_i} u\|,$$

waarbij $\pi_{h_i} u$ de stuksgewijs lineaire interpolant van u is, die in de punten van de i-de partitie met u overeenstemt. En over $\|u - \pi_{h_i} u\|$ kunnen we vanuit de interpolatieheorie wél iets zeggen. Uit vergelijking (3.5.3) in Paragraaf 3.5 zien we nl. dat

$$\|u' - (\pi_{h_i} u)'\|_\infty \leq \|h_i u''\|_\infty.$$

Hieruit volgt dat

$$(u - \pi_{h_i} u, u - \pi_{h_i} u)' = \int_0^1 (u'(x) - (\pi_{h_i} u)'(x))^2 \, dx \leq \|h_i u''\|_\infty^2,$$

waaruit we zien dat $\|u - \pi_{h_i} u\|' \leq \|h_i u''\|_\infty$. Blijkbaar is $\|u - \pi_{h_i} u\| \leq \sqrt{2}\|u - \pi_{h_i} u\|' \leq \sqrt{2}\|h_i u''\|_\infty$. We concluderen dat er een constante C_1 bestaat, zodanig dat

$$\|u - u_{h_i}\| \leq C_1 \|h_i u''\|_\infty.$$

We hebben op deze manier een a priori schatting verkregen in termen van de functie u''. Hoewel we deze functie niet kennen[5], leren we er toch een en ander van:

1. Wanneer we op de een of andere manier a priori informatie over het gedrag van u'' zouden hebben, leent het blijkbaar de moeite om de partitie fijn te kiezen in de gebieden waar $|u''|$ groot is (adaptieve methode).

2. Wanneer we de partitie equidistant zouden kiezen, met constante maaswijdte $\tilde{h}_i$, dan is blijkbaar
$$\|u - u_{h_i}\| \leq C_1 \tilde{h}_i \|u''\|,$$
Het proces is dan dus convergent van orde tenminste 1.

Opmerking 6.3.2. Zoals wel vaker met a priori schattingen het geval is, weten we nu nog steeds niets over de numerieke waarde van de fout. Men kan echter laten zien dat er een constante C_2 bestaat, expliciet te beschrijven in termen van alleen de functie a, zodanig dat

$$\|u - u_{h_i}\| \leq C_2 \|h_i (f + (a u'_{h_i})')\|.$$

Deze a posteriori foutschatting geeft ons in principe wél een concrete bovengrens voor de fout. Immers, na bepaling van u_{h_i} berekent men allereerst het residu $f + (a u'_{h_i})'$. Op grond daarvan volgt dan weer de waarde van de norm $\|h_i (f + (a u'_{h_i})')\|$, waarna men een expliciete bovengrens weet voor $\|u - u_{h_i}\|$.

[5] De constante C_1 is overigens, door de boeken bij te houden, wel expliciet te bepalen; deze hangt alleen van de functie a in het randwaardeprobleem af

We vatten de hoofdpunten van de uitvoering van de eindige elementen methode voor het randwaardeprobleem (6.1.1) samen in de volgende stelling. De formulering van het recept in deze stelling is geheel elementair: de functionaalanalytische context van Hilbertruimten is niet meer zichtbaar. De verklaring van het feit dat het ook echt *werkt* is natuurlijk wel degelijk gelegen in de abstractere theorie zoals we die hebben ontwikkeld.

Stelling 6.3.3. *Beschouw het randwaardeprobleem*

$$\begin{cases} -(au')' = f; \\ u(0) = u(1) = 0, \end{cases} \tag{6.3.1}$$

waarbij $a \in C^2[0,1]$ en $f \in C[0,1]$. Gezocht wordt $u \in C^2[0,1]$. Neem aan dat a strikt positief is op $[0,1]$—dan is er een unieke oplossing.

Kies een rij van opeenvolgende verfijningen van partities $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \cdots$ van $[0,1]$, zodanig dat voor de bijbehorende maasfuncties h_i $(i = 1, 2, \ldots)$ geldt dat $\lim_{i\to\infty} \|h_i\|_\infty = 0$. Laat V_{h_i} $(i = 1, 2, \ldots)$ de vectorruimte zijn die bestaat uit alle continue functies v op $[0,1]$, die lineair zijn op alle intervallen van de partitie $\mathcal{P}_i$ en die voldoen aan de randwaardecondities $v(0) = v(1) = 0$.

Bij een partitie $\mathcal{P}_i$, met partitiepunten $0 = x_0 < x_1 < \ldots < x_{m_i+1} = 1$, kiezen we voor V_{h_i} de basis $\{\phi_j^{(i)}\}_{j=1}^{m_i}$ bestaande uit de dakfuncties

$$\phi_j^{(i)}(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & x \in [x_{j-1}, x_j]; \\ \frac{x - x_{j+1}}{x_j - x_{j+1}} & x \in [x_j, x_{j+1}]; \\ 0 & x \notin [x_{j-1}, x_{j+1}]. \end{cases}$$

Introduceer nu de bilineaire vorm a en de lineaire vorm l door

$$a(v_1, v_2) = \int_0^1 a(x) v_1'(x) v_2'(x)\, dx$$
$$l(v) = \int_0^1 f(x) v(x)\, dx,$$

en bepaal voor $i = 1, 2, \ldots$ het unieke element $u_{h_i} \in V_{h_i}$ dat voldoet aan $a(u_{h_i}, v) = l(v)$ voor alle $v \in V_{h_i}$. Dit element bepalen we concreet als $u_{h_i} = \sum_{k=1}^{m_i} \xi_k \phi_k^{(i)}$, waarbij we $\xi_k^{(i)}$ $(k = 1, \ldots, m_i)$ vinden door het lineaire stelsel

$$\sum_{k=1}^{m_i} a(\phi_k^{(i)}, \phi_j^{(i)}) \xi_k = l(\phi_j^{(i)}) \quad (j = 1, \ldots, m_i).$$

op te lossen.

Dan is er een constante C, alleen afhankelijk van de functie a, zodanig dat $\|u - u_{h_i}\|_\infty \leq C \|h_i u''\|_\infty$. In het bijzonder convergeren de continue stuksgewijs lineaire functies u_{h_i} dus uniform naar u.

Opmerking 6.3.4. De gebruikte bases $\{\phi_j^{(i)}\}_{j=1}^{m_i}$ van dakfuncties hebben een tweetal voor de praktijk prettige eigenschappen:

1. De resulterende stijfheidsmatrices met coëfficiënten $a(\phi_k^{(i)}, \phi_j^{(i)})$ zijn ijl ("sparse"), d.w.z. bestaan grotendeels uit nullen. Immers, $\phi_k^{(i)} \cdot \phi_j^{(i)} = 0$ en $(\phi_k^{(i)})' \cdot (\phi_j^{(i)})' = 0$ zodra $|j - k| \geq 2$; de bijbehorende daken staan voor dergelijke indices namelijk los van elkaar. Dit impliceert

dat ook $a(\phi_k^{(i)}, \phi_j^{(i)}) = 0$ voor $|j-k| \geq 2$, d.w.z. de stijfheidsmatrices bestaan inderdaad overwegend uit nullen. Bedenk hierbij nu, dat iedere coëfficiënt $a(\phi_k^{(i)}, \phi_j^{(i)})$ berekend moet worden met een numerieke integratie. Hoe ijler de stijfheidsmatrices zijn, hoe goedkoper dit uitrekenen dus is.

2. Het vorige punt laat in feite zien dat de stijfheidsmatrices tridiagonaal zijn. Omdat ze ook nog strikt positief definiet zijn (het zijn immers Gram-matrices), hebben ze blijkbaar een LU-ontbinding. De lineaire stelsels in de stelling zijn dus zeer goedkoop met de double-sweep-method op te lossen.

Afrondend merken we op dat in Stelling 6.3.3 alle belangrijke thema's uit het college uiteindelijk samenkomen:

- Approximatie en interpolatie: om de foutschattingen bij de eindige elementen methode concreter te maken, en ook als basis voor:
- Numerieke integratie: om de stijfheidsmatrices en de belastingsvectoren te kunnen benaderen;
- De theorie van lineaire stelsels: om de functies u_{h_i} daadwerkelijk te kunnen bepalen uit de stijfheidsmatrices en de belastingsvectoren;
- Functionaalanalytisch kader: om convergentie van de eindige elementen methode te concluderen.

Bijlage A

Functionaalanalytisch kader

Veel resultaten over convergentie van numerieke methoden kennen hun meest natuurlijke formulering in de context van genormeerde lineaire ruimten, één van de basisbegrippen uit de functionaalanalyse. In deze Bijlage formuleren we eerst de relevante definities en geven we een aantal praktische voorbeelden van dergelijke ruimtes. Vervolgens bestuderen we het (relatief eenvoudige) eindigdimensionale geval. Tenslotte ontwikkelen we binnen een speciale klasse van genormeerde lineaire ruimten, de zgn. Hilbertruimten, de abstracte theorie van een numeriek proces dat bekend staat als de Galerkin-methode.

A.1 Genormeerde lineaire ruimten

Zij V een reële (resp. complexe) vectorruimte. Een *norm* op V is een afbeelding $\|\,.\,\| : V \mapsto \mathbb{R}$, zodanig dat

1. $\|x\| \geq 0$ voor alle $x \in V$;
2. $\|x\| = 0$ dan en slechts dan als $x = 0$;
3. $\|x + y\| \leq \|x\| + \|y\|$ voor alle $x, y \in V$ (driehoeksongelijkheid);
4. $\|\alpha x\| = |\alpha|\|x\|$ voor alle $\alpha \in \mathbb{R}$ (resp. voor alle $\alpha \in \mathbb{C}$) en alle $x \in V$.

Het paar $(V, \|\,.\,\|)$ heet een *genormeerde lineaire ruimte*. Vaak zijn er op een vectorruimte V meerdere plausibele normen mogelijk; wanneer het echter duidelijk is welke er bedoeld wordt, dan spreken we vaak over V zelf als de genormeerde lineaire ruimte.

Een *metrische ruimte* is een paar (S, d), waarbij S een niet-lege verzameling is, en d een afstandsfunctie (ook wel *metriek* genoemd), d.w.z. een afbeelding $\mathrm{d} : V \times V \mapsto \mathbb{R}$, zodanig dat

1. $\mathrm{d}(x, y) \geq 0$ voor alle $x, y \in S$;
2. $\mathrm{d}(x, y) = 0$ dan en slechts dan als $x = y$;
3. $\mathrm{d}(x, y) = \mathrm{d}(y, x)$ voor alle $x, y \in S$;
4. $\mathrm{d}(x, z) \leq \mathrm{d}(x, y) + \mathrm{d}(y, z)$ voor alle $x, y, z \in S$.

Ook hier zullen we vaak de metriek weglaten in de terminologie.

Een genormeerde lineaire ruimte geeft aanleiding tot een metrische ruimte, wanneer we de metriek $\mathrm{d} : V \times V \mapsto \mathbb{R}$ definiëren door $\mathrm{d}(x, y) = \|x - y\|$. Ga dit zelf na (de driehoeksongelijkheid voor de metriek volgt uit die voor de norm, wanneer we opmerken dat $x - z = (x - y) + (y - z)$).

Als V een genormeerde lineaire ruimte is, dan heet een rij $\{x_n\}_{n=0}^{\infty} \subset V$ *convergent naar* $x^* \in V$ als $\lim_{n\to\infty} \|x^* - x_n\| = 0$. Notatie: $\lim_{n\to\infty} x_n = x^*$, of $x_n \to x^*$. Als S een metrische ruimte is, dan heet een rij $\{x_n\}_{n=0}^{\infty} \subset S$ *convergent naar* $x^* \in S$ als $\lim_{n\to\infty} \mathrm{d}(x^*, x_n) = 0$. De notatie is weer $\lim_{n\to\infty} x_n = x^*$, of $x_n \to x^*$ Een rij in de genormeerde lineaire ruimte V is precies dan convergent, wanneer hij convergent is in de metrische ruimte V (dit volgt direct uit de definitie van de metriek in termen van de norm). Een rij in een genormeerde lineaire ruimte, of, meer in het algemeen, in een metrische ruimte, heeft ten hoogste één limiet.

Het belang van deze begrippen is er voor ons in gelegen, dat *een numerieke methode zich vaak laat interpreteren als het construeren van een expliciet berekenbare rij* $\{x_n\}_{n=0}^{\infty}$ *in een of andere genormeerde lineaire ruimte* V*, met als consequentie van de foutanalyse o.a. de uitspraak dat* $x_n \to x^*$*, waarbij* x^* *de te benaderen grootheid is.*

Voorbeeld A.1.1. Neem $V = \mathbb{R}^n$ (resp. $V = \mathbb{C}^n$). De meest gebruikte normen op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$) zijn, als $x = (x_1, \ldots, x_n)$:

- $\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$ (Euclidische norm);
- $\|x\|_1 = \sum_{i=0}^n |x_i|$ (somnorm);
- $|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$ (supremumnorm of maximumnorm).

Voor iedere $1 \le p < \infty$ definieert $\|x\| = \{\sum_{i=0}^n |x_i|^p\}^{1/p}$ een norm, maar het is voor $p \neq 1, 2$ dan niet meer heel direct in te zien dat aan de driehoeksongelijkheid voldaan is. De gevallen $p = 1$ en $p = 2$ staan hierboven. Er geldt $\lim_{p\to\infty} \|x\|_p = \|x\|_\infty$, hetgeen de op het eerste gezicht wat merkwaardige notatie $\|x\|_\infty$ (in plaats van iets als $\|x\|_{\max}$ of $\|x\|_{\sup}$) verklaart.

Voorbeeld A.1.2. Laat S een niet-lege verzameling zijn, en zij V de collectie van alle begrensde reële functies op S, met vectorruimtestructuur gegeven door puntsgewijze operaties. Definieer $\|f\|_\infty = \sup_{s\in S} |f(s)|$ (supremumnorm). Dan is $\|\,.\,\|$ inderdaad een norm op V. Merk op, dat het puntsgewijze produkt van twee functies in V weer in V is en dat $\|fg\| \le \|f\|_\infty \|g\|_\infty$ voor alle $f, g \in V$.

Wanneer we voor S een verzameling van n elementen nemen, dan is de bijbehorende genormeerde lineaire ruimte te identificeren met $(\mathbb{R}^n, \|\,.\,\|_\infty)$, door de waarden van f op te vatten als de coördinaten van een punt in $\mathbb{R}^n$ ten opzichte van de standaardbasis.

Voorbeeld A.1.3. ("uniforme convergentie") Als speciaal geval van het vorige voorbeeld nemen we $V = C[a, b]$ en voorzien we deze ruimte van de supremumnorm: $\|f\|_\infty = \sup_{x\in[a,b]} |f(x)|$. In dit geval is de term maximumnorm ook gerechtvaardigd, omdat het supremum van de continue functie $|f|$ op $[a, b]$ ook daadwerkelijk als maximum ergens wordt aangenomen: $\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$. Een rij $\{f_n\}$ in $(C[a, b], \|\,.\,\|_\infty)$ is dus convergent met limiet $f \in C[a, b]$, als $\lim_{n\to\infty} \|f - f_n\|_\infty = 0$, d.w.z. als $\lim_{n\to\infty} \max_{x\in[a,b]} |f(x) - f_n(x)| = 0$. Convergentie in deze norm is (dus) hetzelfde als uniforme convergentie.

Als generalisatie hiervan kunnen we $C^n[a, b]$ voorzien van de norm, die wordt gedefinieerd door $\|f\|_\infty = \sum_{i=0}^n \|f^{(i)}\|_\infty$. Convergentie in deze norm is equivalent met uniforme convergentie op $[a, b]$ van alle afgeleiden tot en met de afgeleide van orde n. Ga dit zelf na.

De poging $\|f\| \stackrel{\text{def.}}{=} \|f'\|_\infty$ definieert *geen* norm op $C^1[a, b]$. Er is aan alle vereisten voor een norm voldaan, op één na (welke?).

Voorbeeld A.1.4. ("convergentie in een gemiddelde") Laat weer $V = C[a, b]$. Voor $1 \le p < \infty$ en $f \in C[a, b]$ definiëren we

$$\|f\|_p = \left\{ \int_a^b |f(x)|^p \, dx \right\}^{\frac{1}{p}}.$$

Dit definieert (maar dat is voor algemene p niet heel direct in te zien) inderdaad een norm op $C[a,b]$. Een rij $\{f_n\}_{n=0}^{\infty}$ convergeert in $(C[a,b], \|\,.\,\|_p)$ naar $f \in C[a,b]$, dan en slechts dan, als

$$\lim_{n\to\infty} \int_a^b |f(x) - f_n(x)|^p \, dx = 0,$$

hetgeen een soort gemiddelde convergentie uitdrukt. Bekende speciale gevallen zijn $p = 1$:

$$\|f\|_1 = \int_a^b |f(x)| \, dx,$$

en $p = 2$:

$$\|f\|_2 = \left\{ \int_a^b |f(x)|^2 \, dx \right\}^{\frac{1}{2}}.$$

We zien trouwens dat eenzelfde vectorruimte, hier $C[a,b]$, voorzien kan worden van verschillende normen met een heel verschillend convergentiebegrip. Een voorbeeld: Laat $f_n(x) = x^n$, dan is $\lim_{n\to\infty} f_n = 0$ in $(C[0,1], \|\,.\,\|_1)$, maar dit is *niet* waar in $(C[0,1], \|\,.\,\|_\infty)$. Ga dit na![1]

Analoog aan $\mathbb{R}^n$ geldt

$$\lim_{p\to\infty} \|f\|_p = \|f\|_\infty,$$

voor alle $f \in C[a,b]$, wat weer de notatie verklaart.

Net als bij de klasse van "uniforme" voorbeelden kunnen ook hier voor $C^n[a,b]$ afgeleiden worden meegenomen, zelfs ieder met een eigen gewichtsfunctie. Men verkrijgt dan de zgn. gewogen Sobolevnormen, die in de theorie van differentiaalvergelijkingen een belangrijke rol spelen. De—misschien niet erg voor de hand liggende—definitie is:

$$\|f\| = \left\{ \sum_{i=0}^{n} \int_a^b |f^{(i)}(x)|^p w_i(x) \, dx \right\}^{\frac{1}{p}}, \tag{A.1.1}$$

waarbij de w_i continue en strikt positieve functies op $[a,b]$ zijn. Een variant hierop is

$$\|f\|' = \sum_{i=0}^{n} \left\{ \int_a^b |f^{(i)}(x)|^p w_i(x) \, dx \right\}^{\frac{1}{p}} \quad (1 \le p < \infty, f \in C^n[a,b]).$$

In feite maakt het voor convergentie niet uit welke van de twee men kiest: een rij is convergent in de ene norm, dan en slechts dan, als hij convergent is in de andere norm (dit is niet heel direct in te zien).

Merk op dat er nuttige ongelijkheden kunnen bestaan tussen normen. Zo is bijvoorbeeld $\|f\|_1 \le (b-a)\|f\|_\infty$ voor $f \in C[a,b]$ (ga dit na).

Voorbeeld A.1.5. Een *inwendig produkt* op een reële vectorruimte V is een afbeelding $(\,.\,,\,.\,) : V \times V \mapsto \mathbb{R}$, zodanig dat

1. $(x,x) \ge 0$ voor alle $x \in V$ (strikt positief definiet);
2. $(x,x) = 0$ dan en slechts dan als $x = 0$;
3. $(x,y) = (y,x)$ voor alle $x, y \in V$;

[1] Terzijde, voor wie de terminologie kent: de rij $\{f_n\}_{n=0}^{\infty}$ heeft geen limiet in $(C[a,b], \|\,.\,\|_\infty)$: het is zelfs geen Cauchyrij.

4. $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ voor alle $\alpha, \beta \in \mathbb{R}$ en $x, y, z \in V$.

Vanwege de symmetrie houdt de lineariteit in de eerste variabele automatisch ook lineariteit in de tweede variabele in. De Cauchy-Schwartz-ongelijkheid $|(x, y)|^2 \leq (x, x)(y, y)$ geldt; deze impliceert dat de definitie $\|x\| = \sqrt{(x, x)}$ een norm op V geeft, zodat V hiermee een genormeerde lineaire ruimte wordt. De (dan) bijbehorende metriek wordt dus gegeven door $\mathrm{d}(x, y) = \sqrt{(x - y, x - y)}$.

Het bekendste voorbeeld is uiteraard het standaard inwendig produkt op $\mathbb{R}^n$. Op $C^n[a, b]$ kan men, voor strikt positieve continue functies w_i, een gewogen Sobolev inwendig produkt definiëren door

$$(f, g) = \sum_{i=0}^{n} \int_a^b f^{(i)}(x) g^{(i)}(x) w_i(x)\, dx \quad (f, g \in C^n[a, b]).$$

De bijbehorende norm is dan die voor $p = 2$ in (A.1.1). Het feit dat het gewogen Sobolev inwendig produkt inderdaad ook strikt positief definiet is, zoals vereist voor een inwendig produkt, vergt trouwens nog een klein bewijs. Ga dit zelf eens na voor $n = 0$ en $w = 1$.

Het hangt van de context van een numeriek probleem af, wat de meest geschikte norm is om mee te werken. Voor het benaderen van een getal of vector zal men goed met $(\mathbb{R}^n, \|\,.\,\|_2)$ of $(\mathbb{C}^n, \|\,.\,\|_2)$ uit de voeten kunnen. Voor het benaderen van functies is het minder goed mogelijk om een alle gevallen afdekkend antwoord te geven. Wanneer men de functie puntsgewijs wil benaderen, en in alle punten tegelijk even goed, dan liggen normen die corresponderen met uniforme convergentie voor de hand. Het andere type normen hierboven, corresponderend met convergentie in een gemiddelde, is echter ook wel degelijk praktisch relevant, al lijkt dat op het eerste gezicht misschien niet zo. Stel bijv. maar eens, dat de functie f de energiestroom per tijdseenheid over een of ander grensvlak beschrijft. Als men de netto uitwisseling over de periode tussen $t = 0$ en $t = 1$ seconden wil benaderen, d.w.z. $\int_0^1 f(t)\, dt$ wil benaderen, dan is $\|\,.\,\|_1$ een goede norm om mee te werken. Als we namelijk een rij $\{f_n\}_{n=1}^{\infty}$ zouden kunnen berekenen, met de eigenschap dat $f_n \to f$ in $\|\,.\,\|_1$, dan geldt

$$\left| \int_0^1 f_n(t)\, dt - \int_0^1 f(t)\, dt \right| = \left| \int_0^1 f_n(t) - f(t)\, dt \right| \leq \int_0^1 |f_n(t) - f(t)|\, dt = \|f_n - f\|_1 \to 0,$$

dus $\int_0^1 f_n(t)\, dt \to \int_0^1 f(t)\, dt$. We krijgen blijkbaar de gewenste benaderingen van de netto uitwisseling, door simpelweg de ons bekende benaderingen (in de norm $\|\,.\,\|_1$) f_n van f te integreren.

A.2 Begrensde lineaire afbeeldingen

Een ander basisbegrip in de functionaalanalyse is het volgende.

Definitie A.2.1. Laten $(V, \|\,.\,\|_V)$ en $(W, \|\,.\,\|_W)$ genormeerde lineaire ruimte zijn. Dan heet een lineaire afbeelding $A : V \mapsto W$ *begrensd* als er een $M \in \mathbb{R}$ bestaat, zodanig dat $\|Ax\|_W \leq M\|x\|_V$ voor alle $x \in V$.[2]

[2]Zoals in de vorige paragraaf aangegeven is, kunnen V en W op een natuurlijke wijze worden voorzien van een metriek. Daardoor is het zinvol om te spreken over continue afbeeldingen van V naar W, d.w.z. afbeeldingen $\phi : V \mapsto W$ die de eigenschap hebben dat voor iedere convergent rijtje $x_n \to x_\infty$ in V automatisch ook weer $\phi(x_n) \to \phi(x_\infty)$ geldt. Zonder bewijs vermelden we, dat voor een *lineaire* afbeelding $A : V \mapsto W$ tussen twee begrensde lineaire ruimten, de begrippen continuïteit en begrensdheid dezelfde zijn.

We zullen in het vervolg $\|\,.\,\|_V$ en $\|\,.\,\|_W$ beide als $\|\,.\,\|$ noteren, en slechts subscripts toevoegen als dat nodig is. De verzameling begrensde lineaire afbeeldingen van $(V, \|\,.\,\|)$ naar $(W, \|\,.\,\|)$ geven we aan met $B(V, W)$, waarbij ook hier dus de normen in de notatie worden onderdrukt. Men gaat eenvoudig na dat $B(V, W)$ weer een vectorruimte is onder puntsgewijze operaties.

Als $A \in B(V, W)$, dan zijn de mogelijke "oprekfactoren" $\frac{\|Ax\|}{\|x\|}$ $(x \neq 0)$ dus begrensd. De kleinste bovengrens hiervan nemen we als een maat voor de grootte van A:

Definitie A.2.2. Laten $(V, \|\,.\,\|)$ en $(W, \|\,.\,\|$ genormeerde lineaire ruimten zijn en zij $A \in B(V, W)$. Dan heet

$$\|A\| \stackrel{\text{def.}}{=} \sup\left\{ \frac{\|Ax\|}{\|x\|} \mid x \in V,\, x \neq 0 \right\}$$

de *(operator)norm* van A.

Merk op dat $\|Ax\| \leq \|A\|\|x\|$ voor alle $x \in V$. Merk ook op dat $\|A\| \leq M$ voor alle M met de eigenschap dat $\|Ax\| \leq M\|x\|$ voor alle $x \in V$.

De naamgeving "(operator)norm" is consistent met het eerdere begrip norm, want inderdaad voldoet de afbeelding $\|\,.\|\, : B(V, W) \to \mathbb{R}$ aan de vereisten van een norm op een vectorruimte. Voor dit laatste moet men een aantal routineverficaties uitvoeren, waarvan we slechts de driehoeksongelijkheid laten zien. Laat dus $A, B \in B(V, W)$. Dan is voor $x \in V$:

$$\|(A + B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\| \leq \|A\|\|x\| + \|B\|\|x\| = (\|A\| + \|B\|)\|x\|,$$

zodat inderdaad $\|A + B\| \leq \|A\| + \|B\|$.

Voor samenstellingen gedraagt de norm zich submultiplicatief:

Lemma A.2.3. *Laat $A_{12} \in B(V_1, V_2)$ en $A_{23} \in B(V_2, V_3)$. Dan is $A_{23} \circ A_{12} \in B(V_1, V_3)$ en $\|A_{23} \circ A_{12}\| \leq \|A_{23}\| \cdot \|A_{12}\|$.*

Bewijs dit zelf.

Voorbeeld A.2.4. We zullen later zien dat *iedere* lineaire afbeelding $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ altijd continu is, ongeacht de (mogelijk verschillende) keuze voor de norm op $\mathbb{R}^n$ als domeinruimte en als beeldruimte. Voor $\mathbb{C}^n$ is dit eveneens waar.

Voorbeeld A.2.5. Laat $k \in C([a, b] \times [a, b])$. Definieer met behulp van deze zgn. integraalkern de zgn. integraaloperator $K : C[a, b] \to C[a, b]$ als

$$(Kf)(t) = \int_a^b k(s, t) f(s)\, ds \quad (t \in [a, b]).$$

Men kan nagaan (we zullen dit hier niet doen) dat Kf inderdaad weer continu is. Een routineverificatie leert dat K lineair is. In feite is K zelfs begrensd, wanneer we $C[a, b]$ als beeld- en domeinruimte van $\|\,.\,\|_\infty$ voorzien. Immers (we schatten vrij grof af):

$$\begin{aligned} |(Kf)(t)| &= \left| \int_a^b k(s, t) f(s)\, ds \right| \\ &\leq \int_a^b |k(s, t)||f(s)|\, ds \\ &\leq \int_a^b |k(s, t)|\, ds \cdot \|f\|_\infty \\ &\leq (b - a)\|k\|_{[a,b]\times[a,b],\infty} \cdot \|f\|_\infty. \end{aligned}$$

Blijkbaar is $\|Kf\|_\infty \leq (b - a)\|k\|_{[a,b]\times[a,b],\infty} \cdot \|f\|_\infty$. We concluderen dat K inderdaad begrensd is, en dat $\|K\| \leq (b - a)\|k\|_{[a,b]\times[a,b],\infty}$.

Twee normen $\|\,.\,\|$ en $\|\,.\,\|'$ op een vectorruimte V heten *equivalent* als er $\alpha, \beta > 0$ bestaan, zodanig dat $\alpha\|x\| \leq \|x\|' \leq \beta\|x\|$ voor alle $x \in V$. Dan geldt uiteraard $\frac{1}{\beta}\|x\|' \leq \|x\| \leq \frac{1}{\alpha}\|x\|'$ voor alle $x \in V$, dus er is symmetrie. Het betreft hier in feite een equivalentierelatie op de verzameling van alle normen op V.[3] Men kan een norm vervangen door een equivalente, zonder dat de verzameling van begrensde lineaire afbeeldingen verandert:

Propositie A.2.6. *Laten $(V, \|\,.\,\|_V)$ en $(W, \|\,.\,\|_W)$ genormeerde lineaire ruimten zijn, en veronderstel dat $A : (V, \|\,.\,\|_V) \mapsto (W, \|\,.\,\|_W)$ begrensd is met norm $\|A\|$. Indien $\|\,.\,\|'_V$ equivalent is met $\|\,.\,\|_V$, en $\|\,.\,\|'_W$ equivalent is met $\|\,.\,\|_W$, dan is ook $A : (V, \|\,.\,\|'_V) \mapsto (W, \|\,.\,\|'_W)$ begrensd.*

De door de paren $\|\,.\,\|_V$, $\|\,.\,\|_W$ en $\|\,.\,\|'_V$, $\|\,.\,\|'_W$ geïnduceerde operatornormen op $B(V, W)$ zijn equivalent.

Bewijs. Er zijn, als gevolg van de equivalentie van normen, $c_1, c_2 > 0$, zodanig dat voor alle $x \in V$ de ongelijkheden $\|Ax\|'_W \leq c_1\|Ax\|_W \leq c_1\|A\|\|x\|_V \leq c_1\|A\|c_2\|x\|'_V$ gelden. Dit laat zien dat de collectie van begrensde lineaire afbeeldingen dezelfde is. Het bewijs voor de equivalentie van de operatornormen is vergelijkbaar. □

A.3 Het eindigdimensionale geval

Waneer de onderliggende vectorruimte van een genormeerde lineaire ruimte *eindig*dimensionaal is, is alles redelijk eenvoudig. Dit berust op het volgende fundamentele resultaat, waarvan we het bewijs achterwege laten.

Stelling A.3.1. *Alle normen op $\mathbb{R}^n$ (of $\mathbb{C}^n$) zijn equivalent.*[4]

Gevolg A.3.2. *Voor iedere norm $\|\,.\,\|$ op $\mathbb{R}^n$ is een rij convergent in $(\mathbb{R}^n, \|\,.\,\|)$ (resp. in $(\mathbb{C}^n, \|\,.\,\|)$), dan en slechts dan, als voor iedere coördinaat de bijbehorende rij van coördinaten convergeert in $\mathbb{R}$ (resp. in $\mathbb{C}$).*

Bewijs. Het is een direct gevolg van Stelling A.3.1 dat een rij in de norm $\|\,.\,\|$ convergeert, dan en slechts dan als de rij convergeert in de norm $\|\,.\,\|_\infty$, en voor die laatste norm is het duidelijk. □

Voor $\mathrm{End}\,(\mathbb{R}^n)$ (resp. $\mathrm{End}\,(\mathbb{C}^n)$), de vectorruimte van alle lineaire afbeeldingen van $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) naar zichzelf, geldt het volgende.

Gevolg A.3.3. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ lineair zijn. Kies twee normen $\|\,.\,\|$ en $\|\,.\,\|'$ op $\mathbb{R}^n$. Dan is $A : (\mathbb{R}^n, \|\,.\,\|) \mapsto (\mathbb{R}^n, \|\,.\,\|')$ begrensd, ongeacht de keuze. Iedere keuze voor $\|\,.\,\|$ en $\|\,.\,\|'$ geeft een bijbehorende norm op $\mathrm{End}\,(\mathbb{R}^n)$, maar alle op deze manier verkregen normen op $\mathrm{End}\,(\mathbb{R}^n)$ zijn equivalent. Voor $\mathbb{C}^n$ geldt een analoog resultaat.*

Bewijs. We bewijzen het voor $\mathbb{R}^n$; het bewijs voor $\mathbb{C}^n$ is analoog. Als gevolg van Propositie A.2.6 en Stelling A.3.1 is het voldoende om te laten zien dat $A : (\mathbb{R}^n, \|\,.\,\|_\infty) \mapsto (\mathbb{R}^n, \|\,.\,\|_\infty)$ begrensd is. Hiertoe merken we op dat voor $i = 1, \ldots, n$ geldt:

$$|(Ax)_i| = \left|\sum_{j=1}^{n} a_{ij}x_j\right| \leq \sum_{j=1}^{n} |a_{ij}||x_j| \leq \sum_{j=1}^{n} |a_{ij}|\|x\|_\infty \leq \left\{\max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|\right\} \|x\|_\infty.$$

[3] Zonder bewijs vermelden we, dat twee normen op V precies dan equivalent zijn, wanneer ze dezelfde topologie op V definiëren.

[4] Er is dus precies één topologie op $\mathbb{R}^n$ (of $\mathbb{C}^n$) die door een norm wordt geïnduceerd.

Dus $\|Ax\|_\infty \leq \left\{\max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|\right\} \|x\|_\infty$ voor alle $x \in \mathbb{R}^n$. We concluderen dat A met deze keuze van normen inderdaad begrensd is, en verkrijgen zelfs als extra informatie dat dan

$$\|A\| \leq \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|.$$

De equivalentie van de operatornormen is een direct gevolg van Propositie A.2.6 (of, als alternatief bewijs, als gevolg van het noodzakelijkerwijs equivalent zijn van ieder tweetal normen op de eindigdimensionale vectorruimte $\mathrm{End}\,(\mathbb{R}^n)$). □

Opmerking A.3.4. Voor een lineaire afbeelding $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (of $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) kunnen we het supremum in de definitie van de operatornorm in feite vervangen door het maximum, maar dat zullen we niet nodig hebben.

Voor de veel gebruikte normen $\|\,.\,\|_1$, $\|\,.\,\|_2$ en $\|\,.\,\|_\infty$ op $\mathbb{R}^n$ (of op $\mathbb{C}^n$) zijn de onderlinge equivalenties eenvoudig direct in te zien:

Propositie A.3.5. *Voor alle $x \in \mathbb{R}^n$ (en alle $x \in \mathbb{C}^n$) geldt:*

1. $\frac{1}{\sqrt{n}}\|x\|_1 \leq \|x\|_2 \leq \|x\|_1$;
2. $\frac{1}{n}\|x\|_1 \leq \|x\|_\infty \leq \|x\|_1$;
3. $\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2$.

Bewijs. Alle ongelijkheden zijn triviaal (ga na), behalve de eerste in onderdeel 1. Hiertoe merken we op dat volgens de ongelijkheid van Schwartz

$$\|x\|_1 = \sum_{i=1}^n |x_i| = (\,(|x_1|,\ldots,|x_n|)^t\,,\,(1,\ldots,1)^t\,) \leq \|x\|_2 \cdot \sqrt{n}.$$

□

We noteren met $\|A\|_p$ de norm van A wanneer $\mathbb{R}^n$ (of $\mathbb{C}^n$) wordt voorzien van de norm $\|\,.\,\|_p$ ($1 \leq p \leq \infty$). Voor $\|A\|_p$ bestaan geen eenvoudige directe uitdrukkingen in termen van de coëfficiënten van A, behalve in de volgende twee gevallen.

Propositie A.3.6. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (of $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) lineair zijn. Dan is*

1. $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|$ *(maximale kolomsom);*
2. $\|A\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^n |a_{ij}|$ *(maximale rijsom).*

Bewijs. We bewijzen alleen onderdeel (1) voor het geval $\mathbb{R}^n$; alle andere situaties worden analoog behandeld. Merk in dat geval op dat, voor $x \in \mathbb{R}^n$,

$$\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^n |\sum_{j=1}^n a_{ij}x_j| \leq \sum_{i,j=1}^n |a_{ij}||x_j| \\
&= \sum_{j=1}^n \left\{\sum_{i=1}^n |a_{ij}|\right\} |x_j| \leq \sum_{j=1}^n \left\{\max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|\right\} |x_j| \\
&= \left\{\max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|\right\} \sum_{j=1}^n |x_j| = \left\{\max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|\right\} \|x\|_1.
\end{aligned}$$

Dus $\|A\|_1 \leq \{\max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|\}$. Om gelijkheid in te zien, kiezen we een k zodanig dat $\sum_{i=1}^n |a_{ik}| = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|$. Dan is $\|e_k\|_1 = 1$ en $\|Ae_k\|_1 = \|(a_{1k}, \ldots, a_{nk})^t\|_1 = \sum_{i=1}^n |a_{ik}| = \max_{j=1,\ldots,n} \sum_{i=1}^n |a_{ij}|$. □

Voor $\|\,.\,\|_2$ is de berekening van de operatornorm ingewikkelder. Een eerste indicatie van het algemene antwoord zien we in het volgende resultaat. Hierbij wordt gebruik gemaakt van het *spectrum* van een lineaire afbeelding $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (of $A : \mathbb{C}^n \mapsto \mathbb{C}^n$). Dit spectrum, genoteerd met $\sigma(A)$, is de verzameling van alle (eventueel complexe) nulpunten van het karakteristieke polynoom van A. In het geval van $A : \mathbb{C}^n \mapsto \mathbb{C}^n$ zijn dit dus precies de eigenwaarden van A, maar voor $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ hoeft dit niet zo te zijn. Voor $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ geldt $\sigma(A^t) = \sigma(A)$, en voor $A : \mathbb{C}^n \mapsto \mathbb{C}^n$ geldt $\sigma(A^*) = \overline{\sigma(A)}$. De *spectraalstraal* van A, genoteerd met $\rho(A)$, wordt gedefinieerd als $\rho(A) = \max_{\lambda \in \sigma(A)} |\lambda|$. Dan is $\rho(A) = \rho(A^t)$ in het reële geval, en $\rho(A) = \rho(A^*)$ in het complexe geval.

Propositie A.3.7. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (of $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) een orthonormale basis van eigenvectoren hebben. Dan is*

$$\|A\|_2 = \rho(A).$$

Bewijs. Laat $\{v_1, \ldots, v_n\}$ een orthonormale basis van eigenvectoren zijn, met $Av_i = \lambda_i v_i$ ($i = 1, \ldots, n$). Voor $x = \sum_{i=1}^n \xi_i v_i$ is dan

$$\|Ax\|_2 = \sqrt{\sum_{i=1}^n |\lambda_i \xi_i|^2} \leq \max_{\lambda \in \sigma(A)} |\lambda| \cdot \sqrt{\sum_{i=1}^n |\xi_i|^2} = \rho(A) \cdot \|x\|_2.$$

Dus $\|A\|_2 \leq \rho(A)$. Voor gelijkheid kiezen we een v_k, zodanig dat $|\lambda_k| = \rho(A)$. Dan is $\|Av_k\|_2 = |\lambda_k| \|v_k\|_2 = \rho(A)\|v_k\|_2$. □

Voor algemene matrices is het antwoord als volgt.

Stelling A.3.8. *Voor een lineaire afbeelding $A : \mathbb{R}^n \mapsto \mathbb{R}^n$, resp. $A : \mathbb{C}^n \mapsto \mathbb{C}^n$, is*

$$\|A\|_2 = \sqrt{\rho(A^t A)} \text{ , resp. } \|A\|_2 = \sqrt{\rho(A^* A)}.$$

Bewijs. We bewijzen eerst het reële geval. Laten $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ de eigenwaarden (met multipliciteiten) van de symmetrische matrix $A^t A$ zijn, met bijbehorende orthonormale basis van eigenvectoren $v_1, v_2, \ldots, v_n$. Dan is:

$$\begin{aligned}
\|A\|_2^2 &= \sup_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} \\
&= \sup_{x \neq 0} \frac{(Ax, Ax)}{(x, x)} \\
&= \sup_{x \neq 0} \frac{(A^t Ax, x)}{(x, x)} \\
&= \sup_{(\xi_1, \ldots, \xi_n) \neq (0, \ldots, 0)} \frac{(A^t A \sum_{i=1}^n \xi_i v_i, \sum_{i=1}^n \xi_i v_i)}{(\sum_{i=1}^n \xi_i v_i, \sum_{i=1}^n \xi_i v_i)} \\
&= \sup_{(\xi_1, \ldots, \xi_n) \neq (0, \ldots, 0)} \frac{\sum_{i=1}^n \lambda_i \xi_i^2}{\sum_{i=1}^n \xi_i^2} \\
&\leq \lambda_n.
\end{aligned}$$

Dus $\|A\|_2 \leq \sqrt{\lambda_n} = \sqrt{\rho(A^t A)}$. Door $\|Av_n\|_2^2$ te beschouwen zien we dat in feite gelijkheid geldt.

Voor het complexe geval gaat het bewijs analoog, wanneer men $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ de eigenwaarden (met multipliciteiten) van de Hermitese matrix A^*A laat zijn. □

Opmerking A.3.9.

1. Voor een matrix A, die een orthonormale basis van eigenvectoren bezit, reduceert Stelling A.3.8 tot Propositie A.3.7.
2. De getallen $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ in bovenstaand bewijs heten de *singuliere waarden* van A. Blijkbaar is $\|A\|_2$ gelijk aan de grootste singuliere waarde van A.

Lemma A.3.10. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (resp. $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) lineair zijn, en zij $\|\,.\,\|$ een norm op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$), met corresponderende norm $\|A\|$ van A. Indien $\lambda \in \sigma(A)$, dan is $|\lambda| \leq \|A\|$.*

Bewijs. Voor het complexe geval kiezen we een corresponderende eigenvector v_λ van A. Uit $\|Av_\lambda\| = |\lambda|\|v_\lambda\|$ volgt het gestelde dan onmiddellijk. In het reële geval is de meetkundige multipliciteit niet altijd groter dan nul, dus daar werkt dit niet. We herleiden het reële geval daarom tot het complexe, door allereerst A op natuurlijke wijze uit te breiden tot een complex lineaire afbeelding $A_{\mathbb{C}} : \mathbb{C}^n \mapsto \mathbb{C}^n$. We vermelden (zonder bewijs), dat het voor iedere norm $\|\,.\,\|$ op $\mathbb{R}^n$ mogelijk is om deze norm uit te breiden tot een norm $\|\,.\,\|_{\mathbb{C}}$ op $\mathbb{C}^n$, zodanig dat voor alle lineaire $B : \mathbb{R}^n \mapsto \mathbb{R}^n$ de norm $\|B_{\mathbb{C}}\|$ van $B_{\mathbb{C}} : (\mathbb{C}^n, \|\,.\,\|_{\mathbb{C}}) \mapsto (\mathbb{C}^n, \|\,.\,\|_{\mathbb{C}})$ gelijk is aan de norm $\|B\|$ van $B : (\mathbb{R}^n, \|\,.\,\|) \mapsto (\mathbb{R}^n, \|\,.\,\|)$. In het bijzonder is dus $\|A_{\mathbb{C}}\| = \|A\|$. Uit het resultaat voor het complexe geval volgt nu dat $|\lambda| \leq \|A_{\mathbb{C}}\| = \|A\|$. □

Gevolg A.3.11. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (resp. $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) lineair zijn, en zij $\|\,.\,\|$ een norm op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$), met corresponderende norm $\|A\|$ van A. Dan is $\rho(A) \leq \|A\|$.*

Bij de bestudering van iteratieve methoden van lineaire stelsels komen we machten A^k van een matrix A tegen. De volgende stelling is de functionaalanalytische basis voor de convergentie-eigenschappen van dergelijke methoden.

Stelling A.3.12. *Laat $A : \mathbb{R}^n \mapsto \mathbb{R}^n$ (resp. $A : \mathbb{C}^n \mapsto \mathbb{C}^n$) lineair zijn. Dan zijn equivalent:*

1. $\lim_{k\to\infty}(A^k)_{ij} = 0$ *voor alle* $i, j = 1, \ldots, n$;
2. *Er is een norm $\|\,.\,\|$ op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$), z.d.d. $\lim_{k\to\infty} \|A^k\| = 0$;*
3. *Voor iedere norm $\|\,.\,\|$ op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$) is $\lim_{k\to\infty} \|A^k\| = 0$;*
4. *Er is een norm $\|\,.\,\|$ op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$), z.d.d. $\lim_{k\to\infty} A^k x = 0$ voor alle $x \in \mathbb{R}^n$ (resp. voor alle $x \in \mathbb{C}^n$);*
5. *Voor iedere norm $\|\,.\,\|$ op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$) is $\lim_{k\to\infty} A^k x = 0$ voor alle $x \in \mathbb{R}^n$ (resp. voor alle $x \in \mathbb{C}^n$);*
6. $\rho(A) < 1$.

Al deze beweringen gelden in ieder geval wanneer er een norm op $\mathbb{R}^n$ (resp. op $\mathbb{C}^n$) bestaat, zodanig dat $\|A\| < 1$.[5]

Bewijs.

[5] Zonder bewijs vermelden we, dat er, zodra $\rho(A) < 1$ geldt, ook inderdaad een norm bestaat met $\|A\| < 1$. Het betreft hier dus in feite een zevende equivalente bewering

$1 \Longrightarrow 2$ Voor $\|\,.\,\|_1$ geldt $\|A^k\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^n |(A^k)_{ij}| \to 0$.

$2 \Longrightarrow 3$ Alle normen op $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) geven equivalente normen op $\mathrm{End}\,(\mathbb{R}^n)$ (resp. $\mathrm{End}\,(\mathbb{C}^n)$).

$3 \Longrightarrow 1$ De keuze $\|\,.\,\|_1$ levert op dat $\max_{j=1,\ldots,n} \sum_{i=1}^n |(A^k)_{ij}| = \|A^k\|_1 \to 0$. Blijkbaar geldt $\sum_{i=1}^n |(A^k)_{ij}| \to 0$ voor alle $j = 1, \ldots, n$, en dat impliceert dat $(A^k)_{ij} \to 0$ voor $i, j = 1, \ldots, n$.

$1 \Longrightarrow 4$ $\|\,.\,\|_1$ voldoet.

$4 \Longrightarrow 5$ Alle normen op $\mathbb{R}^n$ (resp. $\mathbb{C}^n$) zijn equivalent.

$5 \Longrightarrow 1$ Blijkbaar is dan, voor $j = 1, \ldots, n$, ook $\sum_{i=1}^n |(A^k)_{ij}| = \|A^k e_j\|_1 \to 0$; dit impliceert (1).

$2 \Longrightarrow 6$ Als $\lambda \in \sigma(A)$, dan $\lambda^k \in \sigma(A^k)$, dus volgens Lemma A.3.10 is $|\lambda^k| \leq \|A^k\| \to 0$. Dit impliceert $|\lambda| < 1$.

$6 \Longrightarrow 1$ We behandelen eerst het complexe geval. Een moment van nadenken leert, dat het voldoende is, om voor één geschikte basis aan te tonen dat de matrixcoëfficiënten naar 0 gaan, wanneer A wordt weergegeven t.o.v. die basis. We kiezen hiervoor een Jordanbasis, en merken op, dat het dan weer voldoende is om de bewering aan te tonen voor één Jordan-blok. Daar $\rho(A) < 1$, is de in zo'n blok voorkomende eigenwaarde in absolute waarde kleiner dan 1. We mogen, kortom, aannemen dat $A = \lambda \mathbf{I} + N$, met N het bekende nilpotente deel van A en $\lambda < 1$. Veronderstel dat $N^{m+1} = 0$. Voor $k \geq m$ is dan $A^k = \sum_{l=0}^m \binom{k}{l} \lambda^{k-l} N^l$. De coëfficiënten van A^k zijn dus van de vorm $\binom{k}{l}\lambda^{k-l}$, voor $l \in \{0, \ldots, m\}$. Merk nu op dat, voor l vast, $\binom{k}{l}\lambda^{-l}$ een polynoom in k is. Daar $|\lambda| < 1$, is dus $\lim_{k\to\infty} \binom{k}{l}\lambda^{-l}\lambda^k = 0$. Dit bewijst het complexe geval. Voor het reële geval merken we op dat $\rho(A_{\mathbb{C}}) = \rho(A)$. Volgens het voorgaande is dus $\lim_{k\to\infty}(A^k_{\mathbb{C}})_{ij} = 0$ voor alle $i, j = 1, \ldots, n$. Daar de matrices van $A_{\mathbb{C}}$ en A dezelfde zijn (bij kanonieke basiskeuze), volgt nu ook het reële geval.

De toevoeging is duidelijk op grond van Gevolg A.3.11. □

De toevoeging in bovenstaande stelling is van grote praktische waarde. De convergentie van iteratieve technieken voor het benaderend oplossen van lineaire stelsels hangt namelijk i.h.a. af van de spectraalstraal van een geschikt gekozen matrix. Wanneer die spectraalstraal kleiner dan 1 is, dan convergeert de methode. Het berekenen van een spectraalstraal is echter zonder verdere a priori kennis een dure aangelegenheid, omdat het bewerkelijk is om eigenwaarden van een matrix te berekenen (beter gezegd: te benaderen). Bovenstaande toevoeging geeft echter de mogelijkheid om in sommige gevallen toch zonder veel rekenwerk convergentie te kunnen concluderen, simpelweg omdat we "zien" dat er een norm op $\mathbb{R}^n$ of $\mathbb{C}^n$ is, zodanig dat de daarbij behorende relevante operatornorm al kleiner dan 1 is. Dat zal dan voor de spectraalstraal blijkbaar zeker zo zijn. De keuze van deze norm op $\mathbb{R}^n$ of $\mathbb{C}^n$ is hierbij vrij; men kan bijvoorbeeld $\|\,.\,\|_1$ of $\|\,.\,\|_\infty$ kiezen, die ieder een eenvoudige expliciete uitdrukking voor de operatornorm hebben (zie Propositie A.3.6).

A.4 Stabiliteit en conditiegetal

Een aantal numerieke methoden betreft het inverteren van een lineaire afbeelding $A : V \mapsto W$, waarbij V en W genormeerde lineaire ruimten zijn en $A \in B(V, W)$ een isomorfisme van abstracte vectorruimten is. Dan bestaat uiteraard A^{-1} als lineaire afbeelding. We kunnen ons

nu afvragen, in hoeverre fouten in de gegevens doorwerken bij het inverteren. Dat wil zeggen, we zijn geïnteresseerd in de oplossing x van de vergelijking $Ax = b$, maar we hebben slechts de beschikking over een benadering $\tilde{b}$ van b, op grond waarvan we (exact) de oplossing $\tilde{x}$ van de vergelijking $A\tilde{x} = \tilde{b}$ bepalen. In hoeverre wijkt $\tilde{x}$ nu van x af?

Een eerste pathologie treedt op wanneer A^{-1} niet begrensd is. Men kan laten zien, dat er dan een rij $\tilde{b}_n \to b$ bestaat, zodanig dat $\|A^{-1}\tilde{b}_n - A^{-1}b\| \to \infty$. Met andere woorden: het verbeteren van de nauwkeurigheid van de gegevens—d.w.z. het verkleinen van $\|\tilde{b} - b\|$—is zinloos, want dat geeft geen enkele garantie over het verbeteren van het antwoord. Men zegt dan wel dat het probleem $Ax = b$ *instabiel* is. Voor het inverteren van lineaire afbeeldingen tussen *eindig*dimensionale genormeerde lineaire ruimten komt dit dus niet voor: dat is altijd een stabiel probleem.

Laten we nu aannemen dat A^{-1} wel begrensd is, en dat dit evenzeer het geval is voor A.[6] Als maatstaf voor de ernst van de doorwerking van de fout zal men nu vaak de uitvergroting van de relatieve fout willen nemen, m.a.w. men kijkt, voor $b \neq \tilde{b}$ en $b \neq 0$, naar

$$\frac{\dfrac{\|\tilde{x} - x\|}{\|x\|}}{\dfrac{\|\tilde{b} - b\|}{\|b\|}} = \frac{\|A^{-1}(\tilde{b} - b)\|}{\|A^{-1}b\|}\frac{\|b\|}{\|\tilde{b} - b\|}.$$

De bovengrens voor deze uitvergroting is blijkbaar

$$\begin{aligned}\sup\left\{ \frac{\|A^{-1}(\tilde{b} - b)\|}{\|\tilde{b} - b\|}\frac{\|b\|}{\|A^{-1}b\|} \mid b, \tilde{b} \in W;\, b \neq 0, \tilde{b} \neq b \right\} \\ = \sup\left\{ \frac{\|A^{-1}w\|}{\|w\|}\frac{\|Av\|}{\|v\|} \mid v \in V,\, w \in W;\, v, w \neq 0 \right\} = \|A^{-1}\|\|A\|.\end{aligned}$$

Men noemt $\|A^{-1}\|\|A\|$ het *conditiegetal* $\kappa(A)$ van A (of van het bijbehorende probleem). Er geldt $\kappa(A) = \kappa(A^{-1})$. Vanwege $\mathbf{I} = A \circ A^{-1}$ is $\kappa(A) \geq 1$. Hoe kleiner het conditiegetal, hoe beter.

A.5 Hilbertruimten

De klasse van genormeerde lineaire ruimten, waarover verreweg het meeste bekend is, bestaat uit de zgn. Hilbertruimten. Deze ruimten sluiten, ook wanneer ze oneindigdimensionaal zijn, nog het beste aan bij onze ruimtelijke intuïtie in eindige dimensie. We zullen in deze paragraaf deze ruimten definiëren en een aantal eigenschappen ervan bestuderen. Hierbij werken we over $\mathbb{R}$—de theorie over $\mathbb{C}$ is overigens analoog.

Laat V een genormeerde lineaire ruimte zijn. Een *Cauchy-rij* in V is een rij die een Cauchy-rij is in de geassocieerde metrische ruimte $\{x_n\}_{n=1}^{\infty} \subset V$, d.w.z. een rij die de eigenschap heeft dat er voor iedere $\epsilon > 0$ een N is, zodanig dat $\|x_n - x_m\| < \epsilon$ voor alle $n, m \geq N$. Een rij $\{x_n\}_{n=1}^{\infty} \subset V$ *convergeert* naar zijn limiet x_∞, als er voor iedere $\epsilon > 0$ een N is, zodanig dat $\|x_n - x_\infty\| < \epsilon$ voor alle $n \geq N$, m.a.w. als de rij convergeert naar x_∞ in de geassocieerde metrische ruimte.

Definitie A.5.1. Een genormeerde lineaire ruimte $(V, \|\,.\,\|)$ heet een *Banachruimte* als iedere Cauchy-rij in V convergent is, m.a.w. wanneer de geassocieerde metrische ruimte volledig is.

[6] Er bestaat een abstracte stelling die zegt dat dit laatste in vele gevallen automatisch volgt uit de begrensdheid van A^{-1}, dus in zekere zin is dit geen zware extra eis.

Voorbeeld A.5.2.

1. $\mathbb{R}^n$ en $\mathbb{C}^n$ zijn Banachruimten in willekeurig welke norm. Immers: alle normen zijn hier equivalent, dus de bijbehorende genormeerde lineaire ruimten zijn óf alle Banachruimten, óf geen van alle, zoals men gemakkelijk nagaat. We weten echter dat de norm $\|\,.\,\|_2$ een Banachruimte oplevert, dus dit is blijkbaar voor alle normen het geval.

2. $(C[a,b], \|\,.\,\|_\infty)$ is een Banachruimte (dit is niet direct in te zien en behoeft bewijs).

Veronderstel nu dat V een ruimte is met een inwendig produkt $(\,.\,,\,.\,)$. Met de norm $\|x\| = \sqrt{(x,x)}$ kan de genormeerde lineaire ruimte $(V, \|\,.\,\|)$ al dan niet een Banachruimte zijn. Dit leidt tot de volgende definitie.

Definitie A.5.3. Laat V een ruimte zijn met een inwendig produkt $(\,.\,,\,.\,)$. Dan heet $(V, (\,.\,,\,.\,))$ een *Hilbertruimte* als $(V, \|\,.\,\|)$ een Banachruimte is, waarbij $\|x\| = \sqrt{(x,x)}$ voor $x \in V$.

Hilbertruimten zijn dus een speciaal soort Banachruimten, nl. die Banachruimten waarvan de norm geïnduceerd wordt door een inwendig produkt.

Voorbeeld A.5.4.

1. $\mathbb{R}^n$ is, voorzien van het standaard inwendig produkt, een Hilbertruimte.

2. Laat $l_2 = \{(x_1, x_2, \ldots \mid \sum_{i=1}^{\infty} x_i^2 < \infty)\}$ (de kwadratisch sommeerbare rijtjes). Onder coördinaatsgewijze operaties blijkt l_2 een vectorruimte te zijn. Definieer een inwendig produkt op l_2, door
$$((x_1, x_2, \ldots), (y_1, y_2, \ldots)) \stackrel{\text{def.}}{=} \sum_{i=1}^{\infty} x_i y_i.$$
Het rechterlid blijkt dan automatisch (absoluut) convergent te zijn, waarmee men dan gemakkelijk nagaat dat deze uitdrukking inderdaad een inwendig produkt oplevert. Het is mogelijk om te laten zien dat l_2 in de bijbehorende metriek volledig is, m.a.w. dat l_2 met bovenstaand inwendig produkt een Hilbertruimte is.

3. Laat $w : [a,b] \mapsto \mathbb{R}$ een strikt positieve continue gewichtsfunctie zijn. Definieer dan
$$L_2([a,b], w(x)\,dx) = \{\text{“alle functies } f : [a,b] \to \mathbb{R}\text{”} \mid \int_a^b f(x)^2 w(x)\,dx < \infty\}.$$
Onder puntsgewijze operaties is dit voorbeeld van een zgn. Lebesgueruimte een vectorruimte. Voorzien van het inwendig produkt
$$(f,g) = \int_a^b f(x)g(x)w(x)\,dx$$
blijkt het zelfs een Hilbertruimte te zijn. De definitie is strikt genomen overigens niet helemaal juist, omdat de elementen van deze ruimte in werkelijkheid geen echte functies zijn, maar equivalentieklassen van functies—vandaar de wat onprecieze uitdrukking “alle functies...” in de definitie. Omdat we hier verder niet mee zullen werken, en omdat het maattheoretisch kader ontbreekt om hier in meer detail op in te gaan, zullen we dit niet verder beschrijven.

Bij iedere genormeerde lineaire ruimte hoort op een natuurlijke manier weer een nieuwe genormeerde lineaire ruimte, die een belangrijke rol speelt in de theorie:

Definitie A.5.5. Laat V een genormeerde lineaire ruimte zijn. Dan is de *duale vectorruimte van* V per definitie gelijk aan $B(X, \mathbb{R})$, d.w.z. de begrensde lineaire afbeeldingen van V naar $\mathbb{R}$.

Merk op, dat V' vanuit de algemene theorie op een natuurlijke manier van een norm voorzien is, nl.

$$\|f\| = \sup_{v \in V, v \neq 0} \frac{|f(v)|}{\|v\|} \quad (f \in V').$$

Het is niet zo gemakkelijk, gegeven een genormeerde lineaire ruimte, om diens duale concreet te beschrijven. Soms lukt dat wél, en dergelijke zgn. representatiestellingen vergemakkelijken de analyse van zo'n ruimte dan aanzienlijk. De duale van een Hilbertruimte is echter verrassend eenvoudig te beschrijven. Merk allereerst op dat het eenvoudig mogelijk is om elementen van de duale V' van een Hilbertruime V aan te geven (iets wat op zich al uitzonderlijk is). Dat gebeurt als volgt. Kies en fixeer $v \in V$ en associeer hiermee de lineaire afbeelding $f_v : V \mapsto \mathbb{R}$, gedefinieerd door

$$f_v(x) = (x, v).$$

Men ziet uit de Schwartz-ongelijkheid dat $|f_v(x)| \leq \|x\|\|v\|$, zodat f_v inderdaad begrensd is. Dit leert ons ook nog dat $\|f_v\| \leq \|v\|$. In feite geldt zelfs gelijkheid. Immers, deze gelijkheid is duidelijk als $v = 0$, en anders is

$$\|f_v\| = \sup_{x \in V, x \neq 0} \frac{|f_v(x)|}{\|x\|} \geq \frac{|f_v(v)|}{\|v\|} = \|v\|.$$

Op deze manier krijgt men, door v te laten variëren, zelfs *geheel* V': dat is de strekking van de volgende stelling.

Stelling A.5.6 (Riesz representatiestelling[7]). *Laat V een Hilbertruimte zijn, en veronderstel $f \in V'$. Dan is er precies één $v \in V$, zodanig dat $f = f_v$, m.a.w. zodanig dat $f(x) = (x, v)$ voor alle $x \in V$. Er geldt dan dat $\|f\| = \|v\|$.*

Bewijs. De gelijkheid van normen hadden we boven al gezien, en de uniciteit is eenvoudig. Stel immers dat $f = f_v$ en $f = f_{v'}$, m.a.w. dat $(x, v) = (x, v')$ voor alle x, ofwel $(x, v - v') = 0$ voor alle x. De keuze $x = v - v'$ laat dan zien dat $v = v'$.

De existentie is wat lastiger, en berust op het volgende fundamentele feit, dat we niet bewijzen: *Als $L \subset V$ een gesloten lineaire deelruimte is, dan is $V = L \oplus L^\perp$, waarbij $L^\perp = \{x \in V | (x, l) = 0$ voor alle $l \in L\}$ het orthoplement van L is.* In $\mathbb{R}^3$ zegt dit resultaat dan bijvoorbeeld dat $\mathbb{R}^3$ de orthogonale directe som is van een vlak en de lijn gegeven door de normaalvector van dat vlak. Voor Hilbertruimten is iets dergelijks blijkbaar altijd het geval.

Hiervan uitgaand bewijzen we de existentie. Die is duidelijk als $f = 0$ (neem $v = 0$), dus we nemen aan dat $f \neq 0$. Kies dan een v_0 z.d.d. $f(v_0) \neq 0$. De lineaire deelruimte $L \overset{\text{def.}}{=} \operatorname{Ker} f$ is gesloten (omdat f continu is), dus we kunnen schrijven $v_0 = l + l^\perp$, met $l \in L$ en $l^\perp \in L^\perp$. Daar $f(v_0) \neq 0$ en $f(l) = 0$, is blijkbaar $f(l^\perp) \neq 0$. Voor $x \in V$ willekeurig is dan

$$f\left(x - \frac{f(x)}{f(l^\perp)} l^\perp\right) = 0,$$

m.a.w.

$$x - \frac{f(x)}{f(l^\perp)} l^\perp \in \operatorname{Ker} f = L,$$

[7] Er zijn, enigszins verwarrend, meerdere stellingen met de naam "Riesz representatiestelling".

zodat i.h.b.

$$\left(x - \frac{f(x)}{f(l^\perp)} l^\perp, l^\perp\right) = 0.$$

Uitwerken geeft

$$f(x) = \left(x, \frac{f(l^\perp) l^\perp}{(l^\perp, l^\perp)}\right).$$

Blijkbaar voldoet $v = \frac{f(l^\perp) l^\perp}{(l^\perp, l^\perp)}$. □

Opmerking A.5.7. Abstracter geformuleerd kan men dus zeggen dat de afbeelding $v \mapsto f_v$ een isometrisch isomorfisme tussen V en V' is: een Hilbertruimte "is" (d.w.z.: laat zich identificeren met) zijn eigen duale.

Een beroemde generalisatie van de Riesz representatiestelling is het volgende resultaat. Het is de verbinding bij uitstek is tussen hele klassen van differentiaalvergelijkingen en de zgn. Hilbertruimte-benadering van die vergelijkingen.

Stelling A.5.8 (Lax–Milgram-Lemma). *Laat V een Hilbertruimte zijn. Veronderstel dat $a : V \times V \mapsto \mathbb{R}$ een afbeelding is die*

- *bilineair is;*
- *begrensd is, d.w.z. dat er een $M \geq 0$ bestaat zodanig dat $|a(x, y)| \leq M\|x\|\|y\|$ voor alle $x, y \in V$.*
- *elliptisch (ook wel: coërcief) is, d.w.z. dat er een $c > 0$ bestaat zodanig dat $a(x, x) \geq c\|x\|^2$ voor alle $x \in V$.*

Dan is er voor iedere $l \in V'$ precies één $u \in V$ zodanig dat $a(u, v) = l(v)$ voor alle $v \in V$.

Bewijs. We zullen ons wat betreft het bewijs beperken tot het geval dat voor ons het belangrijkste is, nl. het geval wanneer a ook nog eens symmetrisch is, d.w.z. dat $a(x, y) = a(y, x)$ voor alle $x, y \in V$. Men gaat in dat geval eenvoudig na, dat de vorm $a(\,.\,,\,.\,)$ dan een inwendig produkt is (de ellipticiteit geeft het positief definiet zijn). Laat $\|x\|_a = \sqrt{a(x, x)}$ de bijbehorende norm zijn. Dan is voor alle $x \in V$:

$$c\|x\|^2 \leq a(x, x) = \|x\|_a^2 \leq M\|x\|^2,$$

waaruit blijkt dat $\|\,.\,\|$ en $\|\,.\,\|_a$ equivalent zijn. Aangezien V een Banachruimte is in de norm $\|\,.\,\|$, is V dat blijkbaar ook in de norm $\|\,.\,\|_a$. We concluderen dat $(V, a(\,.\,,\,.\,))$ een Hilbertruimte is. Merk verder op dat, vanwege de equivalentie van normen, we ook de gelijkheid $(V, a(\,.\,,\,.\,))' = (V, (\,.\,,\,.\,))'$ hebben, zodat $l \in (V, a(\,.\,,\,.\,))'$. Pas nu de Riesz representatiestelling toe op de Hilbertruimte $(V, a(\,.\,,\,.\,))$ en $l \in (V, a(\,.\,,\,.\,))'$. □

Opmerking A.5.9. Het inwendig produkt zelf levert een voorbeeld van een dergelijke vorm $a(\,.\,,\,.\,)$. In dat geval is de uitspraak van het Lax–Milgram-Lemma juist de Riesz representatiestelling.

In de situatie van het Lax–Milgram-Lemma is het, ook wanneer a en l expliciet gegeven zijn, i.h.a. geen eenvoudige zaak om de bijbehorende u expliciet te beschrijven. Wel is het mogelijk om een benaderingsproces aan te geven, waarmee we een rij in V kunnen construeren die naar u convergeert en die wél expliciet berekenbaar is. We zullen daarbij dan zelfs iets kunnen zeggen over de fout.

Het idee van de abstracte constructie is als volgt. Veronderstel dat we een gesloten lineaire deelruinte $V_N \subset V$ hebben. Vervang nu eens het oorspronkelijke probleem in V:

$$\text{bepaal } u \in V \text{ z.d.d. } a(u,v) = l(v) \text{ voor alle } v \in V, \tag{A.5.1}$$

door het corresponderende probleem in V_N:

$$\text{bepaal } u_N \in V_N \text{ z.d.d. } a(u,v) = l(v) \text{ voor alle } v \in V_N. \tag{A.5.2}$$

We weten dat $u_N \in V_N$ inderdaad uniek bestaat, op basis van het Lax–Milgram-Lemma, toegepast op V_N. Even aannemend dat we deze u_N ook expliciet kunnen vinden, construeren we aldus bij iedere stijgende rij $V_{N_1} \subset V_{N_2} \subset \cdots$ van gesloten lineaire deelruimten een expliciet berekenbare rij $\{u_{N_i}\}_{i=1}^{\infty} \subset V$. Wanneer $\bigcup_{i=1}^{\infty} V_{N_i}$ nu maar "groot genoeg" is, mag men hopen dat de steeds zwaarder wordende eisen in (A.5.2) zullen resulteren in de convergentie van de u_{N_i} naar u. Dit blijkt ook inderdaad zo te zijn, wanneer we het "groot genoeg zijn" preciseren als dicht liggen, d.w.z. als we eisen dat $\overline{\bigcup_{i=1}^{\infty} V_{N_i}} = V$. We zullen deze convergentie later aantonen.

De praktische waarde van deze constructie, zo zal duidelijk zijn, staat of valt met het al dan niet concreet kunnen berekenen van de u_{N_i}—die moeten ons immers de benaderingen van u gaan leveren. Het cruciale laatste ingrediënt in deze hele methode (de zgn. *Galerkin-methode*) is daarom de keuze van *eindigdimensionale* deelruimten V_{N_i}. In dat geval zijn de bijbehorende u_{N_i} namelijk inderdaad zeker expliciet berekenbaar! Om die berekenbaarheid voor het eindigdimensionale geval in te zien, keren we terug naar het probleem in (A.5.2), waarbij we nu nog veronderstellen dat $\dim V_N = N < \infty$. Laat $\{\phi_j\}_{j=1}^{\infty}$ dan een basis van V_N zijn en schrijf $u_N = \sum_{j=1}^{N} \xi_j \phi_j$. Vanwege de lineariteit is (A.5.2) equivalent met de eisen dat $a(u_N, \phi_i) = l(\phi_i)$ voor $j = 1, \ldots, N$, ofwel

$$\sum_{j=1}^{N} a(\phi_j, \phi_i)\xi_j = l(\phi_i) \quad (i = 1, \ldots, N).$$

Blijkbaar heeft dit stelsel een unieke oplossing $(\xi_1, \ldots, \xi_N)^t$, want u_N is immers uniek. De crux is nu, dat we die oplossing $(\xi_1, \ldots, \xi_N)^t$ met methoden voor *eindig*dimensionale lineaire stelsels inderdaad ook echt concreet kunnen uitrekenen.

De $N \times N$ matrix met coëfficiënten $a(\phi_j, \phi_i)$ heet (vanwege de historische wortels van het onderwerp bij het berekenen van bouwkundige constructies) de *stijfheidsmatrix* ("stiffness matrix"). De vector met coördinaten $l(\phi_i)$ heet de *belastingsvector* ("load vector"). Wanneer a symmetrisch is, is de stijfheidsmatrix ook symmetrisch. In feite is deze dan zelfs strikt positief definiet, want het is dan de Gram-matrix van de basis $\{\phi_j\}_{j=1}^{N}$ van V_N t.o.v. het inwendig produkt $a(\,.\,,\,.\,)$ op V_N. Voor symmetrische a komen dan dus ook de methoden voor lineaire stelsels in aanmerking die specifiek zijn voor strikt positieve matrices.

Opmerking A.5.10. Wanneer de begrensde elliptische vorm a symmetrisch is, heeft de oplossing u_N van (A.5.2) een meer meetkundige interpretatie, als volgt. In dat geval is, zoals we al zagen, $(V, a(\,.\,,\,.\,))$ eveneens een Hilbertruimte. In deze Hilbertruimte is V_N eveneens gesloten, dus we kunnen schrijven $V = V_N \oplus_a V_N^{\perp_a}$, waarbij dit een splitsing is die orthogonaal is t.o.v. het inwendig produkt $a(\,.\,,\,.\,)$. De oplossing u van (A.5.1) laat zich dan dus schrijven als $u = v_N + v_N^{\perp}$, met $v_N \in V_N$ en $v_N \in V_N^{\perp_a}$. De vergelijkingen (A.5.1) en (A.5.2) laten dan zien dat i.h.b.

$$a(v_N + v_N^{\perp_a}, v) = a(u_N, v) \text{ voor all } v \in V_N.$$

Daar echter per constructie geldt dat $a(v_N^{\perp^a}, v) = 0$ voor alle $v \in V_N$, is blijkbaar

$$a(v_N, v) = a(u_N, v) \text{ voor all } v \in V_N,$$

en dat impliceert dat $v_N = u_N$.

We concluderen dat, voor symmetrische a, de gezochte u_N niets anders is dan de orthogonale projectie (m.b.t. het inwendig produkt $a(\,.\,,\,.\,)$) van u op V_N.

Na deze schets van de hoofdlijnen in de Galerkin-methode zullen we ons nu richten op de vraag naar de convergentie. Het primaire resultaat daarvoor is het volgende.

Propositie A.5.11 (Ongelijkheid van Céa). *Laat V een Hilbertruimte zijn, voorzien van een bilineaire, begrensde en elliptische vorm a met corresponderende constanten $M \geq 0$ en $c > 0$:*

1. *$|a(x,y)| \leq M\|x\|\|y\|$ voor alle $x, y \in V$.*
2. *$a(x,x) \geq c\|x\|^2$ voor alle $x \in V$.*

Dan geldt, voor iedere gesloten lineaire deelruimte $V_N \subset V$, en voor alle paren $(u, u_N) \in V \times V_N$ die voldoen aan

$$a(u,v) = a(u_N, v) \text{ voor alle } v \in V_N,$$

de afschatting

$$\|u - u_N\| \leq \frac{M}{c} \inf_{v \in V_N} \|u - v\|. \tag{A.5.3}$$

Bewijs. Voor $u = u_N$ zijn beide leden in (A.5.3) gelijk aan 0. Neem dus aan dat $u \neq u_N$. Voor $v \in V_N$ is

$$c\|u - u_N\|^2 \leq a(u - u_N, u - u_N) = a(u - u_N, u - v) + a(u - u_N, v - u_N).$$

Nu is echter per aanname $a(u - u_N, \tilde{v}) = 0$ voor alle $\tilde{v} \in V_N$, dus i.h.b. is $a(u - u_N, v - u_N) = 0$ voor alle $v \in V_N$. Blijkbaar is voor alle $v \in V_N$

$$c\|u - u_N\|^2 \leq a(u - u_N, u - v) \leq |a(u - u_N, u - v)| \leq M\|u - u_N\|\|u - v\|,$$

Delen door $\|u - u_N\|$ geeft nu dat

$$\|u - u_N\| \leq \frac{M}{c}\|u - v\|$$

voor alle $v \in V_N$. □

Opmerking A.5.12.

1. De uitdrukking $\inf_{v \in V_N} \|u - v\|$ is uiteraard niets anders dan de afstand van u to V_N.
2. Wanneer a symmetrisch is, dan is (vgl. het argument in Opmerking A.5.10) u_N de orthogonale projectie van u op V_N m.b.t. $a(\,.\,,\,.\,)$. Men kan (analoog aan het eindigdimensionale geval) in zijn algemeenheid bewijzen, dat een orthogonale projectie van een punt op een gesloten lineaire deelruimte juist het punt oplevert, dat de afstand tot de betreffende deelruimte realiseert. Dat betekent in dit geval dat

   $$\|u - u_N\|_a = \min_{v \in V} \|u - v\|_a.$$

   Omdat $\|\,.\,\|$ en $\|\,.\,\|_a$ equivalente normen zijn, volgt de ongelijkheid van Céa hier dan ook uit.

Gevolg A.5.13. *Laat V een Hilbertruimte zijn, voorzien van een bilineaire, begrensde en elliptische vorm a. Veronderstel dat $V_{N_1} \subset V_{N_2} \subset \cdots$ een stijgende rij gesloten lineaire deelruimten is, zodanig dat $\overline{\bigcup_{i=1}^{\infty} V_{N_i}} = V$. Laat $u \in V$ vast zijn, en laat $u_{N_i} \in V_{N_i}$ zodanig zijn dat*

$$a(u, v) = a(u_{N_i}, v) \text{ voor alle } v \in V_{N_i} \quad (i = 1, 2, \ldots).$$

Dan is $u = \lim_{i\to\infty} u_{N_i}$.

Merk op dat de u_{N_i} inderdaad bestaan, en uniek vastliggen, als gevolg van het Lax–Milgram-Lemma: de afbeelding $v \mapsto a(u, v)$ van V_N naar $\mathbb{R}$ is immers voor iedere V_{N_i} een element van V'_{N_i}, als gevolg van de begrensdheid van a.

Bewijs. Zij $\epsilon > 0$. De aanname over het dicht liggen van de vereniging impliceert dat er een index k_0 en een $\tilde{v}_{k_0} \in V_{k_0}$ bestaan, zodanig dat $\|u - \tilde{v}_{k_0}\| < \frac{\epsilon c}{M}$, waarbij c en M de constanten uit Propositie A.5.11 zijn. Een toepassing van Propositie A.5.11 in de eerste ongelijkheid hieronder geeft dan dat voor $k \geq k_0$:

$$\|u - u_k\| \leq \frac{M}{c} \inf_{v\in V_k} \|u - v\| \leq \frac{M}{c} \inf_{v\in V_{k_0}} \|u - v\| \leq \frac{M}{c}\|u - \tilde{v}_{k_0}\| < \frac{M}{c} \cdot \frac{\epsilon c}{M} = \epsilon.$$

□

Onze analyse van de Galerkin-methode is nu voltooid. We vatten alles samen in de volgende stelling. Deze stelling vormt de basis voor de numerieke oplossing van belangrijke klassen differentiaalvergelijkingen (gewoon en partieel) volgens de zgn. eindige elementen methode, zoals we die in de hoofdtekst behandelen.

Stelling A.5.14 (Galerkin-methode). *Laat V een Hilbertruimte zijn, voorzien van een afbeelding $a : V \times V \mapsto \mathbb{R}$ die*

1. *bilineair is;*
2. *begrensd is, d.w.z. dat er een $M \geq 0$ bestaat zodanig dat $|a(x, y)| \leq M\|x\|\|y\|$ voor alle $x, y \in V$.*
3. *elliptisch is, d.w.z. dat er een $c > 0$ bestaat zodanig dat $a(x, x) \geq c\|x\|^2$ voor alle $x \in V$.*

Veronderstel dat $V_{N_1} \subset V_{N_2} \subset \cdots$ een stijgende rij gesloten lineaire deelruimten in V is, zodanig dat $\overline{\bigcup_{i=1}^{\infty} V_{N_i}} = V$. Laat $l \in V'$.

Dan is er een unieke $u \in V$, zodanig dat

$$a(u, v) = l(v) \text{ voor alle } v \in V.$$

en voor $i = 1, 2, \ldots$ is er een unieke $u_{N_i} \in V_{N_i}$ zodanig dat

$$a(u_{N_i}, v) = l(v) \text{ voor alle } v \in V_{N_i}.$$

Er geldt dan dat

1. $\lim_{i\to\infty} u_{N_i} = u$;
2. $\|u - u_{N_i}\| \leq \frac{M}{c} \inf_{v\in V_{N_i}} \|u - v\|$ $(i = 1, 2, \ldots)$.

Wanneer de V_{N_i} eindigdimensionaal zijn, zeg $\dim V_{N_i} = N_i$, dan kan u_{N_i} expliciet berekend worden, als volgt:

1. *Kies een basis* $\{\phi_j^{(i)}\}_{j=1}^{N_i}$ *van* V_{N_i}*;*
2. *Schrijf* $u_{N_i} = \sum_{k=1}^{N_i} \xi_k \phi_k^{(i)}$*;*
3. *Dan is* $(\xi_1, \ldots, \xi_{N_i})^t$ *de unieke oplossing van het stelsel*

$$\sum_{k=1}^{N_i} a(\phi_k^{(i)}, \phi_j^{(i)})\xi_k = l(\phi_j^{(i)}) \quad (j = 1, \ldots, N_i).$$

Opmerking A.5.15. De Galerkin-methode kent variaties. Een mogelijke variatie is in sommige gevallen bijv. het (moeten) werken met een bilineaire vorm $a : V \times \widetilde{V} \mapsto \mathbb{R}$, waarbij V en $\widetilde{V}$ beide Hilbertruimten zijn. Men kiest dan een stijgende rij $V_{N_1} \subset V_{N_2} \subset \cdots$ van eindigdimensionale lineaire deelruimten in V, en een stijgende rij $\widetilde{V}_{N_1} \subset \widetilde{V}_{N_2} \subset \cdots$ van eindigdimensionale lineaire deelruimten in $\widetilde{V}$. Het probleem van het vinden van een $u \in V$, zodanig dat $a(u, v) = l(v)$ voor alle $v \in \widetilde{V}$, wordt dan vervangen door het voor iedere i vinden van een $u \in V_{N_i}$, zodanig dat $a(u_{N_i}, v) = l(v)$ voor alle $v \in \widetilde{V}_{N_i}$. De hoop is dan dat $u_{N_i} \to u$ in V. Men spreekt in dit verband wel over de V_{N_i} als de ruimten van benaderingsfuncties ("trial functions") en over de $\widetilde{V}_{N_i}$ als de ruimten van testfuncties ("test functions"). In onze uitwerking van de methode hadden we te maken met één Hilbertruimte, en vielen de benaderingsfuncties met de testfuncties samen.

Bijlage B

Opgaven

B.1 Nulpunten van reële functies

We ontwikkelen in deze opgaven een tweetal methoden om een nulpunt van een reële functie te benaderen.

Als eerste methode bekijken we de *bisectiemethode* in de Opgaven B.1.1 tot en met B.1.3. Deze methode is toepasbaar op een continue functie, wanneer we op voorhand al een interval kennen met grenzen waarin de functie verschillende tekens heeft.

Opgave B.1.1. Laat $[a, b] \subset \mathbb{R}$ en $f \in C[a, b]$ zijn, zodanig dat $f(a)f(b) < 0$. We willen een nulpunt van f op $[a, b]$ gaan benaderen.

(a) f heeft inderdaad tenminste één nulpunt op $[a, b]$. Waarom?

(b) We rekenen $f(a)$, $f(b)$ en $f(\frac{a+b}{2})$ uit. Als $f(\frac{a+b}{2}) = 0$, dan hebben we bij toeval exact een nulpunt gevonden en zijn we klaar. Ook als $f(\frac{a+b}{2}) \neq 0$ heeft ons rekenwerk ons toch iets geleerd over de nulpunten van f. Wat?

(c) Geef aan hoe op deze manier een rij intervallen $\{[a_k, b_k]\}_{k=0}^{\infty} \subset [a, b]$ geconstrueerd kan worden, zodanig dat ieder van deze intervallen tenminste één nulpunt van f bevat en zodanig dat $[a_0, b_0] = [a, b]$ en $b_k - a_k = 2^{-k}(b - a)$.

De constructie in deze opgave staat om voor de hand liggende reden bekend als de *bisectiemethode*.

Opgave B.1.2. Gegeven is de functie f met $f(x) = x^2 - 2$. We passen de bisectiemethode uit Opgave B.1.1 toe om het nulpunt $\sqrt{2}$ van f te benaderen. Als startinterval $[a, b]$ nemen we $[1.4, 1.5]$.

(a) Bepaal een zo klein mogelijke index k_0, zodanig dat $b_{k_0} - a_{k_0} < 0.04$.

(b) Bereken de bijbehorende a_{k_0} en b_{k_0}.

(c) Bepaal met het vorige onderdeel een $\zeta \in \mathbb{R}$, zodanig dat $|\zeta - \sqrt{2}| < 0.02$.

Opgave B.1.3. Gegeven is de functie f met $f(x) = x^3 + 4x^2 - 10$. Bepaal een $\zeta \in \mathbb{R}$, zodanig dat er een nulpunt x^* van f bestaat met $|\zeta - x^*| < 0.125$.
(Hint: bereken $f(1)$ en $f(2)$)

De volgende serie opgaven behandelt de *methode van Newton–Raphson*. Deze methode is toepasbaar op tweemaal continu differentieerbare functies, waarbij we als beginpunt voor de methode een getal nodig hebben dat "voldoende dicht" bij een (enkelvoudig) nulpunt ligt. De methode van Newton–Raphson convergeert veel sneller dan de bisectiemethode en vergt meer achterliggende theorie. We beginnen met een voorbereidende opgave.

Opgave B.1.4. Laat $g : [a, b] \mapsto [a, b]$ continu zijn.

(a) Bewijs dat er tenminste één $x^* \in [a, b]$ is, zodanig dat $g(x^*) = x^*$. Zo'n x^* heet een *dekpunt* van g (Engels: "fixed point").
(Hint: kijk naar $g(x) - x$)

(b) Veronderstel dat g ook nog differentieerbaar is op (a, b), zodanig dat

$$\max_{x\in(a,b)} |g'(x)| \leq \theta$$

voor een of andere $\theta < 1$. Laat zien dat het dekpunt x^* dan uniek is.
(Hint: gebruik de middelwaarde-stelling)

(c) Handhaaf de veronderstelling in het voorgaande onderdeel. Kies $x_0 \in [a, b]$ willekeurig en definieer iteratief $x_{k+1} = g(x_k)$ voor $k \geq 0$. Bewijs dat

$$|x_k - x^*| \leq \theta^k |x_0 - x^*| \leq \theta^k (b - a) \quad \text{voor } k \geq 0, \tag{B.1.1}$$

en

$$|x_k - x^*| \leq \frac{\theta}{1-\theta} |x_k - x_{k-1}| \quad \text{voor } k \geq 1. \tag{B.1.2}$$

(d) Volgens (B.1.1) convergeert de rij $\{x_k\}_{k=0}^{\infty}$ blijkbaar naar x^*. Sterker nog: zowel (B.1.1) als (B.1.2) vertelt ons in een concreet geval iets kwantitatiefs over de fout na k stappen, even aangenomen dat we θ kennen. Deze twee afschattingen hebben op een heel andere manier praktische waarde voor het itereren van de berekening, wanneer we een voorgegeven nauwkeurigheid willen bereiken. Wat is het verschil?
De ongelijkheden in (B.1.1) geven een zogenaamde *a-priori* afschatting; (B.1.2) is een *a-posteriori* afschatting. Wat voor type(n) schatting(en) kunnen we bij de bisectiemethode geven?

Laat nu f een functie zijn waarvan we een nulpunt willen vinden. Ter bepaling van de gedachten nemen we voorlopig even aan dat f gedefinieerd is op heel $\mathbb{R}$, dat f differentieerbaar is, en dat f' geen nulpunten heeft. Definieer dan $g : \mathbb{R} \mapsto \mathbb{R}$, door

$$g(x) = x - \frac{f(x)}{f'(x)} \quad (x \in \mathbb{R}). \tag{B.1.3}$$

Een nulpunt van f, dat we willen vinden, is evident hetzelfde als een dekpunt van g. We zullen later dan ook de resultaten van Opgave B.1.4 op g toepassen (waarbij we een geschikt interval $[a, b]$ zullen moeten weten te vinden). Die resultaten leveren ons dan een rij $\{x_k\}_{k=0}^{\infty}$ op die naar een dekpunt van g, d.w.z. een nulpunt van f, convergeert. De iteratiestap hierin wordt dan dus gegeven door

$$x_{k+1} = g(x_k) = x_k - \frac{f(x_k)}{f'(x_k)} \quad \text{voor } k \geq 0. \tag{B.1.4}$$

We stellen de analyse van de methode even uit en proberen heuristisch te begrijpen waarom dit wel eens zou kunnen werken. We zien dan ook waarom een goede startwaarde zo belangrijk is.

Vergelijking (B.1.3) heeft een meetkundige betekenis. Namelijk: $g(x)$ is de waarde die gevonden wordt door de raaklijn in $(x, f(x))$ aan de grafiek van f te snijden met de x-as.

Opgave B.1.5. Ga dit na.

Opgave B.1.6. De rij, die door (B.1.4) gedefinieerd wordt, heeft dus ook een eenvoudige meetkundige betekenis, die ons in staat stelt te begrijpen wat er kan gebeuren.

(a) Neem $f(x) = e^x - 1$. Hoe ontwikkelt de rij $\{x_k\}_{k=0}^{\infty}$ zich dan voor verschillende startwaarden x_0?

(b) Neem $f(x) = x^2 - 2$. Nu is $f'(0) = 0$, maar zolang we 0 maar niet als startwaarde nemen levert (B.1.4) toch een welgedefinieerde rij op. Hoe ontwikkelt deze rij zich dan voor de verschillende startwaarden $x_0 \neq 0$?

(c) Neem $f(x) = x(x^2 - 1)$. Kies $x_0 = \frac{1}{\sqrt{5}}$ en voer door berekening twee stappen van (B.1.4) uit. Wat gebeurt er? Schets het bijbehorende plaatje.

We zullen nu de analyse van de methode van Newton–Raphson uitvoeren, en nemen hiervoor aan dat $f \in C^2[a, b]$ een *enkelvoudig* nulpunt heeft in $x^* \in (a, b)$, d.w.z. dat $f(x^*) = 0$, maar $f'(x^*) \neq 0$. Voor een startwaarde $x_0 \in [a, b]$ willen we de functie g uit (B.1.3) itereren, als in (B.1.4), waarbij we hopen dat de zo geconstrueerde rij $\{x_k\}_{k=0}^{\infty}$ dan naar het dekpunt x^* van g, d.w.z. het nulpunt x^* van f, convergeert. We willen hierbij gebruik maken van de resultaten over dekpunten in Opgave B.1.4.

Een interatiestap kan ons in het algemeen buiten $[a, b]$ brengen, maar we zullen laten zien dat er altijd een interval $[x^* - \delta, x^* + \delta] \subset [a, b]$ bestaat (voor een of andere $\delta > 0$), zodanig dat, voor iedere startwaarde in dat interval, de bijbehorende rij $\{x_k\}_{k=0}^{\infty}$ binnen het interval blijft en naar het nulpunt x^* convergeert. Sterker nog: de convergentie zal uiterst snel blijken te zijn. We beginnen in de volgende opgave eerst met het interval en met de convergentie op zich.

Opgave B.1.7.

(a) Ga na dat de functie g in (B.1.3) als afgeleide

$$g'(x) = \frac{f(x)f''(x)}{(f'(x))^2}$$

heeft.

(b) Laat zien dat er een $\delta > 0$ is, zodanig dat

$$\begin{cases} [x^* - \delta, x^* + \delta] \subset [a, b], \\ f'(x) \neq 0 \text{ voor alle } x \in [x^* - \delta, x^* + \delta], \text{en} \\ \max_{x \in [x^* - \delta, x^* + \delta]} |g'(x)| < 1. \end{cases} \tag{B.1.5}$$

(c) Kies een δ als in (B.1.5). Laat zien, dat g het interval $[x^* - \delta, x^* + \delta]$ in zichzelf afbeeldt (hint: x^* is een dekpunt van g, gebruik de middelwaardestelling).

(d) Voor iedere δ als in (B.1.5) is blijkbaar de rij $\{x_k\}_{k=0}^{\infty}$, voor iedere startwaarde x_0 in $[x^* - \delta, x^* + \delta]$, welgedefinieerd door de iteraties in (B.1.4), en bevat in datzelfde interval . Het is nu, op grond van eerdere resultaten, ook duidelijk dat de rij dan naar x^* convergeert, en ook dat x^* het enige nulpunt van f in $[x^* - \delta, x^* + \delta]$ is. Hoe zit dit?

(e) Waar hebben we eigenlijk gebruikt dat het nulpunt x^* van f enkelvoudig is?

In de situatie van Opgave B.1.7 is er sprake van een rij $\{x_k\}_{k=0}^{\infty}$ die naar x^* convergeert, zodanig dat $|x_k - x^*| \leq \theta |x_{k-1} - x^*|$ $(k \geq 1)$, voor zekere $\theta < 1$. De fout in de benadering wordt dus iedere keer met tenminste een factor $\theta < 1$ verkleind. Een dergelijke vorm van convergentie heet *van orde tenminste* 1. Voor $p > 1$ zeggen we dat de convergentie van een naar $x^* \in \mathbb{R}$ convergente rij $\{x_k\}_{k=0}^{\infty}$ *van orde tenminste* p is, wanneer er een constante $C \geq 0$ is, zodanig dat $|x_k - x^*| \leq C|x_{k-1} - x^*|^p$ $(k \geq 1)$. Voor $p \geq 1$ heet de convergentie van een naar $x^* \in \mathbb{R}$ convergente rij $\{x_k\}_{k=0}^{\infty}$ *van orde* p, als de convergentie van orde p is, en voor alle $\epsilon > 0$ *niet* van orde $p + \epsilon$. Convergentie van orde 1 heet ook wel *lineaire convergentie*, convergentie van orde 2 wordt ook wel *kwadratische convergentie* genoemd. We nemen bij het bespreken van de orde van convergentie vaak stilzwijgend aan dat $x_k \neq x^*$ voor alle k.

Hoe groter de orde van convergentie, hoe beter. Zie de volgende opgave ter gedachtenbepaling.

Opgave B.1.8. Stel, dat twee rijen $\{x_k\}_{k=0}^{\infty}$ en $\{\tilde{x}_k\}_{k=0}^{\infty}$ beide naar $x^* \in \mathbb{R}$ convergeren. Neem aan dat voor $k \geq 1$ het volgende geldt:

$$|x_k - x^*| \leq \frac{1}{2}|x_{k-1} - x^*| \quad , \quad |\tilde{x}_k - x^*| \leq 1000|\tilde{x}_{k-1} - x^*|^2.$$

We weten verder dat $|x_0 - x^*| \leq \frac{1}{10^{1001}}$ en dat $|\tilde{x}_0 - x^*| \leq \frac{1}{1001}$.

(a) Schat $|x_k - x^*|$ en $|\tilde{x}_k - x^*|$ af voor alle k.

(b) Van welke rij mogen we—op basis van de beschikbare informatie—voor grote k de beste benadering van x^* verwachten?

Opgave B.1.9. Wat voor orde van convergentie zullen door de bisectiemethode gegenereerde rijen in het algemeen hebben?

De methode van Newton–Raphson geeft, voor startwaarden in het interval $[x^* - \delta, x^* + \delta]$ uit Opgave B.1.7, rijen die in feite tenminste kwadratisch convergent zijn, maar dit is op grond van wat we nu weten nog niet duidelijk. Uit de Opgaven B.1.7 en B.1.4 volgt slechts dat de convergentie tenminste lineair is. We zullen deze betere convergentie nu gaan aantonen.

Laat δ zijn als in (B.1.5), kies een startwaarde $x_0 \in [x^* - \delta, x^* + \delta]$ en laat $\{x_k\}_{k=0}^{\infty} \subset [x^* - \delta, x^* + \delta]$ de bijbehorende met iteratie (B.1.4) geconstrueerde rij zijn. Volgens Opgave B.1.7 is deze rij inderdaad goed gedefinieerd. We schrijven dan, met behulp van de stelling van Taylor, voor $k \geq 0$:

$$f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2}f''(\tau_k)(x^* - x_k)^2,$$

voor een of andere τ_k in het gesloten interval met grenzen x^* en x_k. Nu is $f(x^*) = 0$; wanneer we verder, op basis van (B.1.4), de term $x_k f'(x_k)$ uit bovenstaande vergelijking herschrijven, dan verkrijgen we

$$(x_{k+1} - x^*) = \frac{1}{2}\frac{f''(\tau_k)}{f'(x_k)}(x_k - x^*)^2 \quad (k \geq 0). \tag{B.1.6}$$

Opgave B.1.10. Ga dit na.

Op grond van (B.1.6) concluderen we dat

$$|x_{k+1} - x^*| \leq \frac{1}{2}\frac{\max_{x \in [x^* - \delta, x^* + \delta]} |f''(x)|}{\min_{x \in [x^* - \delta, x^* + \delta]} |f'(x)|} |x_k - x^*|^2 \quad (k \geq 0),$$

d.w.z. de methode van Newton–Raphson geeft inderdaad, voor startwaarden in $[x^* - \delta, x^* + \delta]$, rijen met orde van convergentie minstens 2.

Samengevat hebben we het volgende resultaat:

Stelling B.1.11 (Methode van Newton–Raphson). *Laat $f \in C^2[a,b]$ en veronderstel dat f in $x^* \in (a,b)$ een enkelvoudig nulpunt heeft. Kies (dit is mogelijk) een $\delta > 0$, zodanig dat*

$$\begin{cases} [x^*-\delta, x^*+\delta] \subset [a,b], \\ f'(x) \neq 0 \text{ voor alle } x \in [x^*-\delta, x^*+\delta], \text{en} \\ \max_{x\in[x^*-\delta,x^*+\delta]} \left| \frac{f(x)f''(x)}{(f'(x))^2} \right| < 1. \end{cases}$$

Dan is x^* het enige nulpunt van f in $[x^*-\delta, x^*+\delta]$. Voor iedere startwaarde x_0 in $[x^*-\delta, x^*+\delta]$ is de door

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (k \geq 1)$$

recursief gegeven rij $\{x_k\}_{k=0}^{\infty}$ welgedefinieerd en bevat in $[x^*-\delta, x^*+\delta]$. Deze rij convergeert naar x^*. De convergentie is van orde minstens 2: er geldt

$$|x_{k+1} - x^*| \leq \frac{1}{2} \frac{\max_{x\in[x^*-\delta,x^*+\delta]} |f''(x)|}{\min_{x\in[x^*-\delta,x^*+\delta]} |f'(x)|} |x_k - x^*|^2 \quad (k \geq 0).$$

De methode van Newton–Raphson convergeert dus veel sneller dan de bisectiemethode, maar vereist meer differentieerbaarheid en, vooral, vereist ook meer kennis van de functie om een geschikte startwaarde te kunnen kiezen. De bisectiemethode kent dit nadeel niet: als er eenmaal een tekenwisseling is geconstateerd, dan is dat voldoende om de iteraties te starten. In de praktijk past men dan ook, nadat men een nulpunt ingeklemd weet op basis van tekenwisseling, soms eerst een aantal iteraties met bijv. de bisectiemethode toe, in de hoop (of wetenschap) daarmee in ieder geval het nulpunt zo dicht te benaderen dat een geschikte startwaarde voor de methode van Newton–Raphson wordt bereikt—die dan vervolgens wordt uitgevoerd om snel een hoge precisie te bereiken.

Opgave B.1.12. (Voor liefhebbers). Onze eerste analyse van de methode van Newton–Raphson (Opgave B.1.7) gaf op zijn minst lineaire convergentie, maar dat was veel zwakker dan mogelijk: de orde van convergentie is blijkbaar minstens 2. Het is, net zo, denkbaar dat de orde van convergentie in bovenstaande stelling ook nog voor algemene f verder verbeterd kan worden. Dat is echter niet het geval: als $f''(x^*) \neq 0$, dan is de convergentie van de rij *precies* van orde 2, voor alle startwaarden in het interval $[x^*-\delta, x^*+\delta]$ in de stelling. Ga dit met behulp van (B.1.6) na.

Terzijde B.1.13. Wanneer f een meervoudig nulpunt in x^* heeft van eindige orde, dan kan men—onder gladheidseisen op f—overgaan op f/f'. Deze functie heeft een enkelvoudig nulpunt in x^*, zodat de methode van Newton–Raphson dan weer beschikbaar is.

Opgave B.1.14. (Oefenopgave over contracties) De methode van Newton–Raphson bepaalt (hopelijk) het nulpunt van een functie, door het te zien als dekpunt van een bepaalde contractie die iteratief wordt uitgevoerd. Soms, wanneer men het probleem handig kan herschrijven, zijn er ook alternatieve contracties mogelijk. Een voorbeeld is het bepalen van het nulpunt van $f(x) = x - e^{-x}$. Men construeert hiertoe een rij $\{x_k\}_{k=0}^{\infty}$, met startwaarde $x_0 \in [0,1]$, en recursief gedefinieerd door $x_{k+1} = \frac{1}{2}(x_k + e^{-x_k})$.

(a) Ga na dat dit *niet* de methode van Newton–Raphson is.

(b) Bewijs dat de rij $\{x_k\}_{k=0}^{\infty}$ bevat is in $[0,1]$, en convergeert naar het nulpunt x^* van f, onafhankelijk van de startwaarde x_0.

(c) Laat zien dat $|x_k - x^*| < 3^{-k}$ voor $k = 1, 2, 3, \ldots$.

(d) Bepaal, met hulp van het vorige onderdeel, een index k zodanig dat $|x_k - x^*| < 10^{-6}$.

B.2 Fouten en hun doorwerking

Opgave B.2.1. Laat $f(x) = 10(1 - x^2)^{-1}$ voor $x \neq \pm 1$. Bereken het conditiegetal $\gamma(x)$ van f in x, voor $x \neq \pm 1$. Waar treden er problemen op? Waarom juist daar?

Opgave B.2.2. (a) Laat $a \in \mathbb{R}$, $a \neq 0$, en zij $f_a(x) = ax$. Bereken het conditiegetal $\gamma_a(x)$ van f_a in $x \neq 0$.

(b) Laat $a \in \mathbb{R}$ en zij $f_a(x) = x^a$ voor $x > 0$. Bereken het conditiegetal $\gamma_a(x)$ van f_a in $x \neq 0$.

(c) Druk het conditiegetal van de samenstelling $f \circ g$ van twee functies (van één variable) uit in de conditiegetallen van f en g afzonderlijk.

(d) Kijk nu nog eens met andere ogen naar Opgave B.2.1.

Opgave B.2.3. Laat $f(x, y) = x^2 + \frac{1}{\sqrt{y^2+16}}$. Bereken de conditiegetallen (met tekens) van f m.b.t. x en y. We willen een benadering van $f(\tilde{x}, \tilde{y})$ geven in een punt $(\tilde{x}, \tilde{y})$, waarvan we helaas de exacte coördinaten niet kennen, maar alleen weten dat $|\tilde{x} - 2| \leq 0.001$ en $|\tilde{y} - 3| \leq 0.002$. Geef, op basis van de conditiegetallen, een benaderd interval op waarin $f(\tilde{x}, \tilde{y})$ ligt.

Opgave B.2.4. Voor alle $x > 0$ is

$$\sqrt{x+1} - \sqrt{x} = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

Toch is één van beide uitdrukkingen duidelijk beter geschikt om op de betreffende voor de hand liggende manier in een computer geïmplementeerd te worden. Welke uitdrukking is dit, en waarom?

Opgave B.2.5. Een resultaat van een tussenberekening wordt, in een computer met een precisie van 16 cijfers, naar het werkgeheugen weggeschreven. Nadien wordt het resultaat gebruikt in een berekening met een conditiegetal van 10.000, en vervolgens weer weggeschreven. Door een ongelukkige computer-implementatie van de betreffende berekening herhaalt zich dit, steeds het resultaat van de voorgaande stap gebruikmakend, nog driemaal. Wat voor ordegrootte van de relatieve fout in het eindresultaat mag je verwachten?

Opgave B.2.6. Een computer, die met een precisie van 4 cijfers werkt, kan $1000 + 0.4 + 0.4$ op twee voor de hand liggende manieren berekenen. Wat zijn de uitkomsten dan?
Conclusie: optelling is in een computer *niet* associatief. Hetzelfde geldt overigens voor vermenigvuldiging.

Opgave B.2.7. Laat $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$.

(a) Bereken $f(4.71)$ exact.

(b) Een hypothetische computer, die met een precisie van 3 cijfers rekent, voert de evaluatie van f in 4.71 ook uit. Dit gebeurt op de voor de hand liggende manier, waarbij eerst per term alle vermenigvuldigingen stapsgewijs worden uitgevoerd (x^3 wordt bijv. uitgerekend als $\mathrm{fl}(x \cdot \mathrm{fl}(x \cdot x))$, en vervolgens alle termen stapsgewijs van links naar rechts worden opgeteld. Wat is dan de uitkomst?

(c) Men kan ook $f(x) = ((x - 6.1)x + 3.2)x + 1.5$ schrijven, en dit op de voor de hand liggende manier met een precisie van 3 cijfers evalueren. Wat is dan de uitkomst?

Opgave B.2.8. (Zie ook Opgave B.2.7(c).) Laat $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ het algemene polynoom van graad n zijn.

(a) Hoeveel vermenigvuldigingen en hoeveel optellingen zijn er nodig om $p(x_0)$ voor een gegeven x_0 op de meest voor de hand liggende "naïeve" manier uit te rekenen?

(b) Men kan ook schrijven:

$$p(x) = ((\cdots(((a_n x + a_{n-1})x + a_{n-1})x + a_{n-2})x + \cdots a_2)x + a_1)x + a_0.$$

Dit geeft een recursieve berekening, gestart met $b_n = a_n$ en vervolgens gegeven (met dalende indices) door $b_k = b_{k+1}x_0 + a_k$ voor $k = n-1, n-2, \ldots, 1, 0$. Uiteindelijk is dan $b_0 = p(x_0)$. Hoeveel vermenigvuldigingen en hoeveel optellingen zijn er nu nodig?
Deze berekeningswijze staat bekend als de *methode van Horner*.

(c) Vergelijk de beide methoden van berekenen voor een polynoom van graad 100.

.

B.3 Polynomiale approximatie

Opgave B.3.1. Een metalen staaf is verwarmd tot 100 °C. Daarna laat men hem afkoelen, waarbij na resp. 25, 30 en 35 minuten een temperatuur van resp. 24, 17 en 13 °C gemeten wordt. Geef, gebruikmakend van al deze gegevens, een schatting van de temperatuur na 31 minuten, met behulp van de interpolatieformule van Lagrange.

Opgave B.3.2. Zij $t_1 = -1$, $t_2 = 0$, $t_3 = 1$, $t_4 = 2$ en $\alpha_1 = -1$, $\alpha_2 = 1$, $\alpha_3 = 7$ en $\alpha_4 = 35$.

(a) Bepaal (uitdrukkingen voor) de Lagrange-polynomen $\lambda_1, \ldots, \lambda_4$, behorend bij de punten $t_1, \ldots, t_4$.

(b) Bepaal (een uitdrukking voor) het polynoom p van graad ten hoogste 3, zodanig dat $p(t_i) = \alpha_i$ voor $i = 1, \ldots, 4$.

(c) Bereken $p(-2)$.

(d) Veronderstel, dat $\alpha_i = f(t_i)$ voor $i = 1, \ldots, 4$, waarbij $f \in C^4[-2, 2]$ en $\|f^{(4)}\|_{[-2,2],\infty} \leq 1$. Bepaal dan een expliciete bovengrens voor $|p(-2) - f(-2)|$.

(e) Bereken $\gamma \stackrel{\text{def.}}{=} \sum_{i=1}^{4} |\lambda_i(-2)|$.

(f) Laat de $\tilde{\alpha}_i$ verstoringen van de α_i zijn, waarvan we weten dat $|\tilde{\alpha}_i - \alpha_i| \leq \frac{1}{10}$ voor $i = 1, \ldots, 4$. Laat $\tilde{p}$ het interpolerend polynoom zijn, behorend bij deze verstoorde gegevens. Geef een expliciete bovengrens voor $|\tilde{p}(-2) - p(-2)|$.

Opgave B.3.3. Laten $\xi_0, \xi_1, \ldots, \xi_q$ verschillende punten in $\mathbb{R}$ zijn, met bijbehorende Lagrange-polynomen $\{\lambda_i\}_{i=0}^{q}$. Bewijs dat

$$\sum_{i=0}^{q} \xi_i^n \lambda_i(x) = x^n \quad \text{voor alle } x \in \mathbb{R}$$

voor alle $n \in \{0, 1, 2, \ldots, q\}$.

Opgave B.3.4. Zij $t_1 = 0$, $t_2 = 2$, $t_3 = 3$, $t_4 = 4$ en $\alpha_1 = -3$, $\alpha_2 = 5$, $\alpha_3 = 12$, $\alpha_4 = 10$. Bereken de interpolerende polynomen p_1, $p_{1,2}$, $p_{1,2,3}$ en $p_{1,2,3,4}$ bij deze gegevens.

Opgave B.3.5. Voor de staaf uit Opgave B.3.1 blijkt de temperatuur na 40 minuten 11 °C te bedragen. Bereken, nu met behulp van de methode van Newton, een nieuwe schatting van de temperatuur na 31 minuten, op basis van alle vier de meetpunten. Richt de berekening zo in, dat de uitkomst van Opgave B.3.1 onderweg geverifieerd kan worden.

Opgave B.3.6. De interpolatiemethode van Newton geeft een uitdrukking voor het interpolerend polynoom van de vorm

$$P_{0,1,2,\ldots,q}(x) = \sum_{i=0}^{q} \eta_i (x - \xi_0)(x - \xi_1) \cdots (x - \xi_{i-1}),$$

Het idee achter de methode van Horner om polynomen op een efficiënte manier te evalueren (zie Opgave B.2.8) laat zich ook hierop toepassen. Hoe?

Opgave B.3.7. Laat $-1 \leq \xi_0 < \xi_1 < \ldots < \xi_q \leq 1$. Zij $f(x) = e^x$ en laat $\pi_q f$ het bijbehorende interpolerende polynoom zijn. Bepaal een q, zodanig dat

$$\|f - \pi_q f\|_{[-1,1],\infty} \leq 0.0001,$$

onafhankelijk van de ligging van de ξ_i in $[-1, 1]$.

Opgave B.3.8. Laat $f \in C^2[a, b]$, en zij $\pi_1 f$ de lineaire interpolant van f in de punten a en b. Laat zien, uitgaande van Stelling 3.5 in de aantekeningen, dat

$$\|f - \pi_1 f\|_{[a,b],\infty} \leq \frac{1}{8}(a - b)^2 \|f^{(2)}\|_{[a,b],\infty}.$$

Opgave B.3.9. Voor zekere $f \in C^4[0, 6]$ geldt $0 \leq f^{(4)}(x) \leq \frac{1}{100}$ voor alle $x \in [0, 6]$. Verder is bekend dat $f(0) = 8$, $f(2) = 1$, $f(4) = \sqrt{3}$ en $f(6) = 10$. We hebben de beschikking over een benadering δ van $\sqrt{3}$, zodanig dat $-\frac{1}{60} \leq \sqrt{3} - \delta \leq \frac{1}{90}$. Zij p het polynoom van graad ten hoogste 3, zodanig dat $p(0) = 8$, $p(2) = 1$, $p(4) = \delta$ en $p(6) = 10$. Toon aan dat $-\frac{1}{100} < f(3) - p(3) \leq \frac{1}{100}$

Opgave B.3.10. Veronderstel, dat we een logaritmen-tabel willen maken voor het grondtal 10, d.w.z., we willen een lijst aanleggen van equidistante getallen $1 = x_0 < x_2 < \ldots < x_m = 10$, samen met hun logarithmen ${}_{10}\log x_i$ voor $i = 0, 1, 2, \ldots, m$. We nemen ter vereenvoudiging aan dat deze $m+1$ logaritmen exact getabelleerd kunnen worden. Het aantal punten in deze partitie willen we zo kiezen dat de fout, die bij lineaire interpolatie tussen twee getabelleerde waarden gemaakt wordt, nooit groter is dan 10^{-6}.

(a) Hoeveel punten moeten we dan minimaal in onze tabel opnemen?

(b) Veronderstel, dat onze tabel niet voor een equidistante partitie hoeft te worden opgesteld, maar dat we die partitie vrijelijk kunnen kiezen—wel nog steeds met dezelfde eis op de maximale fout bij lineaire interpolatie. Dit scheelt werk. Waarom? Vergelijk de benodigde maaswijdte rond 1 en 10 in dit scenario .

Opgave B.3.11. Voor een zekere $f \in C^2[0, 2]$ weten we slechts dat $\|f^{(2)}\|_{[0,1],\infty} \leq 1$ en $\|f^{(2)}\|_{[1,2],\infty} \leq 100$. We willen punten x_1 en x_2 in een partitie $0 = x_0 < x_1 < x_2 < x_3 = 2$ zodanig kiezen, dat we op basis van de beschikbare informatie over f de kleinst mogelijke bovengrens in de afschatting voor de uniforme fout van de bijbehorende continue stuksgewijs lineaire interpolant $\pi_1 f$ krijgen (zie Stelling 3.12 in de aantekeningen). We perken ons in en overwegen alleen de volgende mogelijkheden:

- We kiezen in ieder geval het punt 1 in onze partitie, en kiezen het resterende partitiepunt dan zo gunstig mogelijk.
- We kiezen een equidistante partitie, dus $x_1 = \frac{2}{3}$ en $x_2 = \frac{4}{3}$.

Wat levert het beste resultaat?

Opgave B.3.12. (Voor liefhebbers; lastig) Veronderstel dat we in de situatie van Opgave B.3.11 niet twee, maar tien punten te verdelen hebben, en dat dit bovendien helemaal vrij kan gebeuren. Bewijs dan, dat de optimale partitie gegeven wordt door de punten 1.0, 1.1, 1.2, ..., 1.8, 1.9 te kiezen.

B.4 Numerieke integratie en extrapolatie

Opgave B.4.1. Bepaal de precisie van de trapeziumregel.

Opgave B.4.2. Bepaal de punten $x_0, x_1 \in [0,1]$ en het gewicht c_1 waarvoor de kwadratuurregel $\tilde{I}(f) = \frac{1}{2}f(x_0) + c_1 f(x_1)$ op het interval $[0,1]$ de grootst mogelijke precisie heeft.

Opgave B.4.3. Laat $a = x_0 < x_1 < \ldots < x_m = b$ een partitie van $[a,b]$ in subintervallen zijn. We kiezen voor ieder subinterval een enkelvoudige kwadratuurregel, en definiëren daarna op de voor de hand liggende wijze een samengestelde kwadratuurregel voor het hele interval. Doe een uitspraak over de precisie van de samengestelde regel, in termen van de precisie van de enkelvoudige regels.

Opgave B.4.4. Beschouw de integraal $\int_0^1 6t^5\,dt$.

(a) Bereken met behulp van de trapeziumregel, de middenpuntsregel en de regel van Simpson een benadering van de integraal. Bereken in ieder van de gevallen de daadwerkelijke fout en vergelijk deze met de verschillende theoretische bovengrenzen voor de fout.

(b) Idem voor de samengestelde trapeziumregel en de samengestelde middenpuntsregel, beide met stapgrootte $\frac{1}{2}$ en $\frac{1}{4}$.

Opgave B.4.5. Beschouw de integraal $\int_1^2 x \log x\,dx = 2\log 2 - \frac{3}{4} = 0.636294$ (afgerond op zes decimalen).

(a) Bereken met behulp van de trapeziumregel, de middenpuntsregel en de regel van Simpson een benadering van de integraal. Bereken in ieder van de gevallen de daadwerkelijke fout en vergelijk deze met de verschillende theoretische bovengrenzen voor de fout.

(b) Idem voor de samengestelde trapeziumregel en de samengestelde middenpuntsregel, beide met stapgrootte $\frac{1}{2}$ en $\frac{1}{4}$.

Opgave B.4.6. Laat $I = \int_0^1 e^{-t^2}\,dt$. We willen m.b.v. de samengestelde trapeziumregel een benadering $\tilde{I}$ van I bepalen z.d.d. $|I - \tilde{I}| < 10^{-6}$. Geef een stapgrootte aan waarvoor dit zeker zo is.

Opgave B.4.7. Zij $f(t) = (1+t^2)^{-\frac{3}{4}}$ en veronderstel dat $\tilde{f} \in C[0,1]$ een benadering is van f, zodanig dat $\|f - \tilde{f}\|_{[0,1],\infty} \leq 0.001$. Zij $I(f) = \int_0^1 f(t)\,dt$ en laten $T_h(f)$ (resp. $T_h(\tilde{f})$) de waarden van de samengestelde trapeziumregel met stapgrootte h op $[0,1]$ zijn voor f (resp. $\tilde{f}$).

(a) Toon aan dat $|T_h(\tilde{f}) - T_h(f)| \leq 0.001$.

(b) Toon aan dat $|T_h(\tilde{f}) - I(f)| \leq 0.001 + \frac{h^2}{8}$.

(c) Bepaal een stapgrootte h zodanig dat $|T_h(\tilde{f}) - I(f)| \leq 0.01$.

Opgave B.4.8. Neem aan dat de machtreeks

$$N_1(h) = \gamma_0 + \gamma_1 h + \gamma_2 h^2 + \gamma_3 h^3 + \cdots$$

convergeert voor alle (voldoende kleine) h. Gebruik de waarden van $N_1(h)$, $N_1(\frac{h}{3})$ en $N_1(\frac{h}{9})$ om een $O(h^3)$-benadering van γ_0 te construeren.

Opgave B.4.9. Neem aan dat

$$N_1(h) = \gamma_0 + \gamma_2 h + \gamma_5 h^4 + O(h^5)$$

voor alle (voldoende kleine) h.

(a) Gebruik de waarden van $N_1(h)$, $N_1(\frac{h}{2})$ en $N_1(\frac{h}{4})$ om een $O(h^5)$-benadering van γ_0 te construeren.

(b) Idem met gebruik making van de waarden van $N_1(h)$, $N_1(\frac{h}{3})$, $N_1(\frac{h}{5})$ en $N_1(\frac{1}{15})$.

Opgave B.4.10. Welke moeilijkheid doet zich voor als we $\int_0^1 \sqrt{t}e^t\,dt$ willen benaderen met behulp van van Romberg-integratie? Hoe kan men deze moeilijkheid wegnemen, zodat men alsnog deze techniek kan toepassen?

Opgave B.4.11. Laat $f(x) = \frac{1}{x}$ voor $x \neq 0$ en beschouw $I(f) = \int_1^2 \frac{1}{x}\,dx = \log 2 = 0.6931472$ (afgerond op 7 decimalen). Het volgende schema voor Romberg-integratie geeft benaderingen voor deze integraal. Verifieer dit schema.

h	i	$T_{i,0} = T_h(f)$	$T_{i,1}$	$T_{i,2}$	$T_{i,3}$	$T_{i,4}$
1	0	0.7500000				
$\frac{1}{2}$	1	0.7083333	0.6944444			
$\frac{1}{4}$	2	0.6970238	0.6932540	0.6931747		
$\frac{1}{8}$	3	0.6941219	0.6931546	0.6931480	0.6931476	
$\frac{1}{16}$	4	0.6933912	0.6931476	0.6931472	0.6931472	0.6931472

Opgave B.4.12. Laat $f(x) = \frac{\sin x}{x}$ voor $x \neq 0$ en laat $f(0) = 1$. De waarde van de integraal $\int_0^{0.8} f(x)\,dx$ willen we benaderen met Romberg-integratie. We berekenen daartoe eerst $T_{0.8}(f) = 0.758680$, $T_{0.4}(f) = 0.768760$ en $T_{0.2}(f) = 0.771262$. Voltooi het schema voor Romberg-integratie, zoals dat op grond van de gegeven getallen kan worden ingevuld, en vergelijk de resultaten met de exacte waarde 0.772095 (afgerond op zes decimalen).

Opgave B.4.13. Pas, gebruikmakend van stapgrootten $1, \frac{1}{2}$ en $\frac{1}{4}$, Romberg-integratie toe op de integraal in Opgave B.4.4. Vergelijk het resultaat met de uitkomsten van die opgave. Doe hetzelfde met de integraal in Opgave B.4.5.

Opgave B.4.14. (Voor liefhebbers) Laat $f \in C[a,b]$ en zij $T_h(f)$ de waarde van de samengestelde trapeziumregel met stapgrootte h voor f. Bij Romberg-integratie construeert men o.a. $\frac{4T_h(f) - T_{2h}(f)}{4-1}$. Laat zien dat dit de samengestelde Simpsonregel is.

Opgave B.4.15. (Voor liefhebbers, ter illustratie van Terzijde 4.29). We bekijken in deze opgave een methode om π te benaderen m.b.v. extrapolatie. Laat $f(h) = \sin\frac{\pi h}{h}$ voor $h \neq 0$.

(a) Laat zien dat er getallen γ_2, γ_4, $\gamma_6, \ldots$ bestaan z.d.d.

$$f(h) = \pi + \gamma_2 h^2 + \gamma_4 h^4 + \gamma_{2p} h^{2p} + O(h^{2p+2}) \quad (h \neq 0).$$

(b) Toon, gebruikmakend van de verdubbelingsformules $\cos\alpha = 2\cos^2\frac{\alpha}{2} - 1$ en $\sin\alpha = 2\sin\frac{\alpha}{2}\cos\frac{\alpha}{2}$, aan dat

$$f(\frac{h}{2}) = \sqrt{2} f(h) \frac{1}{\sqrt{1 + \sqrt{1 - (hf(h))^2}}} \quad (0 < h \leq \frac{1}{2}),$$

(c) Verifieer de waarden (in 7 cijfers nauwkeurig) in de volgende tabel.

h	$f(h)$
$\frac{1}{2}$	2.000 000 0
$\frac{1}{4}$	2.828 427 1
$\frac{1}{8}$	3.061 467 5
$\frac{1}{16}$	3.121 445 2

(d) Zij p het polynomoom van graad ten hoogste 3 zodanig dat $p(h^2) = f(h)$ voor $h = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}$. Bepaal $p(0)$ en vergelijk het resultaat met de waarde 3.141 593 van π, afgerond op 7 decimalen.

B.5 Lineaire stelsels

Opgave B.5.1. Veronderstel dat de reguliere matrix A zonder rijverwisselingen door vegen in bovendriehoeksvorm kan worden gebracht. Laat zien, dat het oplossen van de vergelijking $Ax = b$ evenveel operaties kost, wanneer men dit doet door b mee te nemen in het veegproces, als wanneer men een LU-decompositie bepaalt (met enen op de diagonaal van L), gevolgd door een voorwaartse en een achterwaartse substitutie. (Hint: in beide strategieën bepaalt men in feite een LU-decompositie met enen op de diagonaal van L, en voert men een achterwaartse substitutie uit. Wat wordt er in ieder van de gevallen nog "extra" gedaan?)

Opgave B.5.2. Laat

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix}.$$

De reguliere matrix A heeft een LU-ontbinding waarbij de diagonaal van L uit enen bestaat. Bepaal deze decompositie, en voer een skyline-controle m.b.t. deze decompositie uit. Idem voor

$$\begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix}.$$

Opgave B.5.3. Laat

$$A = \begin{pmatrix} 1 & -1 & -3 \\ -1 & 2 & 4 \\ 1 & 1 & 0 \end{pmatrix} \text{ en } b = \begin{pmatrix} 3 \\ -5 \\ -2 \end{pmatrix}.$$

De reguliere matrix A heeft een LU-ontbinding waarbij de diagonaal van L uit enen bestaat. Bepaal deze decompositie, en los dan m.b.v. deze ontbinding het stelsel $Ax = b$ op. Voer een skyline-controle m.b.t. de decompositie uit.

Opgave B.5.4. Laat

$$A = \begin{pmatrix} 2 & 1 & 1 \\ 2 & 3 & 0 \\ 4 & 2 & 3 \end{pmatrix}, \; b_1 = \begin{pmatrix} 4 \\ 5 \\ 9 \end{pmatrix} \text{ en } b_2 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

De reguliere matrix A heeft een LU-ontbinding waarbij de diagonaal van L uit enen bestaat. Bepaal deze, en los dan m.b.v. deze ontbinding de stelsels $Ax_1 = b_1$ en $Ax_2 = b_2$ op. Voer een skyline-controle m.b.t. de decompositie uit.

Opgave B.5.5. De matrix

$$A = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

is strikt positief definiet. Bereken de Cholesky-decompositie van A en voer een skyline-controle m.b.t. de decompositie uit. Idem voor

$$A = \begin{pmatrix} 9 & -6 & 12 \\ -6 & 29 & -23 \\ 12 & -23 & 26 \end{pmatrix} \text{ en } A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 4 & 6 & 0 \\ 0 & 6 & 10 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}.$$

Opgave B.5.6. De reguliere matrix

$$A = \begin{pmatrix} -2 & 0 & 6 & 0 \\ 0 & 2 & 3 & 0 \\ -2 & -6 & -9 & 4 \\ 0 & 0 & 0 & -8 \end{pmatrix}$$

heeft een LU-decompositie waarbij de diagonaal van L uit enen bestaat. Bepaal deze decompositie en voer een skyline-controle m.b.t. de decompositie uit. Ga ook na of de bandtypes kloppen.

Opgave B.5.7. Laat

$$A = \begin{pmatrix} 3 & 1 & 0 \\ 2 & 4 & 1 \\ 0 & 2 & 5 \end{pmatrix} \text{ en } b = \begin{pmatrix} -1 \\ 7 \\ 9 \end{pmatrix}.$$

De reguliere matrix A heeft een LU-decompositie waarbij de diagonaal van L uit enen bestaat. Los met behulp van de "double sweep method" de vergelijking $Ax = b$ op. Voer een skyline-controle m.b.t de gebruikte decompositie uit. Idem voor

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -4 & 4 & 1 & 0 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & -3 & 15 \end{pmatrix} \text{ en } b = \begin{pmatrix} 3 \\ -10 \\ 6 \\ 36 \end{pmatrix}.$$

Opgave B.5.8. Bereken voor de vectoren

$$\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \text{ en } \begin{pmatrix} 2 \\ 3 \\ 6 \end{pmatrix}$$

de 1-norm, de 2-norm en de maximum-norm.

Opgave B.5.9. Bereken voor de matrix

$$A = \begin{pmatrix} 0 & 2 \\ \frac{1}{8} & 0 \end{pmatrix}$$

de operatornormen $\|A\|_1$, $\|A\|_2$ en $\|A\|_\infty$. Bereken tevens $\rho(A)$.
Idem voor

$$A = \begin{pmatrix} 1 & 1 \\ -2 & -2 \end{pmatrix}$$

en

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Opgave B.5.10. Beschouw het stelsel

$$\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}.$$

Laat zien dat de methode van Jacobi hier convergeert. Voer drie iteraties uit, uitgaande van de startwaarde $x^0 = 0$, en geef een bovengrens voor de supremumnorm van de fout na drie iteraties. Controleer deze bovengrens met behulp van de exacte oplossing $(x_1, x_2)^t = (1, 1)^t$. Bij welk aantal iteraties zal de fout in de supremumnorm met zekerheid tot maximaal 10^{-6} gereduceerd zijn?

Opgave B.5.11. Beschouw het stelsel

$$\begin{pmatrix} 2 & 1 \\ 2 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

Laat, door de 2-norm van de iteratiematrix uit te rekenen, zien dat de methode van Gauß–Seidel hier convergeert. Voer drie iteraties uit, uitgaande van de startwaarde $x^0 = 0$, en geef een bovengrens voor de 2-norm van de fout na twee iteraties. Controleer deze bovengrens met behulp van de exacte oplossing $(x_1, x_2)^t = (1, 1)^t$. Bij welk aantal iteraties zal de fout in de 2-norm met zekerheid tot maximaal 10^{-6} gereduceerd zijn?

Opgave B.5.12. Zij

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

Laten M_J resp. M_G de bijbehorende Jacobi- resp. Gauß–Seidel-iteratiematrices zijn. Toon aan dat $\rho(M_J) = \frac{\sqrt{5}}{2}$ en $\rho(M_G) = \frac{1}{2}$, zodat hier de methode van Jacobi hier niet convergeert, maar die van Gauß–Seidel wel.

Opgave B.5.13. Zij

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}.$$

Laten M_J resp. M_G de bijbehorende Jacobi- resp. Gauß–Seidel-iteratiematrices zijn. Toon aan dat $\rho(M_J) = 0$ en $\rho(M_G) = 2$, zodat hier de methode van Jacobi hier wel convergeert, maar die van Gauß–Seidel niet.

Opgave B.5.14. Laat A een strikt positief definiete $n \times n$-matrix zijn, met eigenwaarden $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Voor iedere $\alpha \in \mathbb{R}$, $\alpha \neq 0$, is er een splitsing $A = N_\alpha - P_\alpha$, waarbij $N_\alpha = \frac{1}{\alpha}\mathbf{I}$ en $P_\alpha = \frac{1}{\alpha}\mathbf{I} - A$. Voor het oplossen van de vergelijking $Ax = b$ leidt dit, na keuze van een verder vaste α, tot het iteratieve proces

$$x^{k+1} = N_\alpha^{-1}P_\alpha x^k + N_\alpha^{-1}b = (\mathbf{I} - \alpha A)x^k + \alpha b \quad (\alpha \neq 0).$$

Ga dit na.

(a) Voor welke α, uitgedrukt in termen van de eigenwaarden van A, is het bijbehorende proces convergent?

(b) Voor welke α is $\rho(\mathbf{I} - \alpha A)$ minimaal? Druk voor deze α de bijbehorende spectraalstraal uit in het conditiegetal van A.

Opgave B.5.15. Laat A een strikt positief definiete $n \times n$-matrix zijn en zij $b \in \mathbb{R}^n$. Beschouw de functie $q : \mathbb{R}^n \mapsto \mathbb{R}$, gedefinieerd door

$$q(y) = \frac{1}{2}(Ay, y) - (b, y).$$

(a) Laat zien dat $\nabla q(y) = Ay - b$ voor alle $y \in \mathbb{R}^n$.

(b) Laat zien dat q precies één lokaal extremum heeft, namelijk een globaal minimum, dat zich bevindt in de oplossing x van $Ax = b$.

(c) We herschrijven de iteratiestap in Opgave B.5.14 als

$$x^{k+1} = x^k - \alpha(Ax^k - b) = x^k - \alpha\nabla q(x^k).$$

Dit proces is, blijkens Opgave B.5.14, convergent voor een zekere collectie van strikt positieve α. Verklaar, met behulp van bovenstaande vergelijking, waarom voor een dergelijke α het bijbehorende iteratieve proces bekend staat als een *steepest descent method*.

Opgave B.5.16. Laat A een $n \times n$-matrix zijn met complexe coëfficiënten.

(a) Veronderstel dat $\lambda \in \mathbb{C}$ voldoet aan $|a_{ii} - \lambda| > \sum_{j=1, j\neq i} |a_{ij}|$ voor $i = 1, \ldots, n$. Laat zien dat $A - \lambda\mathbf{I}$ regulier is.

(b) Definieer, voor $i = 1, \ldots, n$, $D_i = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j\neq i} |a_{ij}|\}$. Laat zien dat de eigenwaarden van A bevat zijn in $\bigcup_{i=1}^n D_i$. De randen van de schijven D_i heten de *cirkels van Gershgorin*.

(c) Laat zien dat de eigenwaarden eveneens bevat zijn in $\bigcup_{i=1}^n D_i'$, waarbij $D_i' = \{z \in \mathbb{C} \mid |z - a_{ii}| \leq \sum_{j=1, j\neq i} |a_{ji}|\}$.

Opgave B.5.17. Zij

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}.$$

Laat, zonder eigenwaarden van A uit te rekenen, zien dat $\|A\|_2 \leq 6$ en $\|A^{-1}\|_2 \leq \frac{1}{2}$. (Hint: gebruik Opgave B.5.16.)

B.6 Eindige elementen methode

Opgave B.6.1. Beschouw het randwaardeprobleem

$$\begin{cases} -((x^2+1)u')' = \sin x \\ u(0) = u(1) = 0 \end{cases}$$

op $[0,1]$. Gegeven is hier, dat dit probleem op $[0,1]$ een unieke oplossing $u \in C^2[0,1]$ heeft.

Geef aan, hoe men met behulp van de eindige elementen methode een rij $\{u_n\}_{n=1}^{\infty}$ van continue stuksgewijs lineaire functies $u_n : [0,1] \mapsto \mathbb{R}$ kan construeren, zodanig dat

$$\lim_{n\to\infty} \|u - u_n\|_{[0,1],\infty} = 0.$$