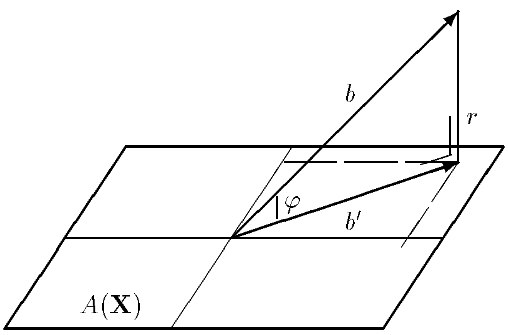
Numerieke Wiskunde C;

Numerieke lineaire algebra



Gerard Sleijpen

Mathematisch Instituut
Rijksuniversiteit Utrecht

Utrecht, juli 1989

NUMERIEKE WISKUNDE C;

Numerieke lineaire algebra

Gerard Sleijpen

Mathematisch Instituut
Rijksuniversiteit Utrecht

Utrecht, juli 1989

Voorwoord

In dit college bespreken we een aantal fundamentele numerieke oplosmethoden voor verschillende elementaire lineaire algebra problemen. Hoewel we ook methoden geven ligt onze nadruk op de wiskundige achtergrond. We besteden veel aandacht aan de stabiliteit en de efficiëntie van de methoden. Het inzicht dat we daardoor hopen te kunnen geven zal behulpzaam zijn bij diverse andere numerieke problemen.

Formeel veronderstellen we geen speciale wiskundige voorkennis. De lineaire algebra theorie (uit de eerste jaars colleges) die we nodig hebben herhalen we hier. Omdat dat dit nogal snel gebeurt zal een zekere routine in lineaire algebra manipulaties toch gewenst zijn.

Dit college is volledig onafhankelijk van het college Numeriek Wiskunde B. In dit college zullen we wel zo nu en dan een evaluatiefoutenanalyse uitvoeren zoals die behandeld is in het college Numerieke Wiskunde A. Ook als niet ieder detail van zo'n analyse begrepen wordt kan de rest van het college probleemloos gevolgd worden. In de inleiding vatten we kort de grondbeginselen voor zo'n analyse samen.

De voetnoten en de passage's die gemarkeerd zijn met een • hoeven niet voor het tentamen bestudeerd te worden.

Dit diktaat is een bewerking van het diktaat Numerieke Wiskunde C van Prof. Dr. A. van der Sluis. De volgorde waarin de onderwerpen hier behandeld worden is anders en een aantal kanttekeningen in het "oude" diktaat zijn in dit diktaat uitgewerkt. Veel van dit soort uitwerkingen staan in de ge-•-de passages.

Gerard Sleijpen

Mathematisch Instituut
Rijksuniversiteit Utrecht

juli 1989, Utrecht

Inhoudsopgave

Inleiding

In dit college zijn we geïnteresseerd in het numeriek oplossen van de volgende problemen.

Het oplossen van een *stelsel lineaire vergelijkingen* (of, in 'matrix taal', de *lineaire vergelijking*):

bereken de reële n vektor $\vec{x} = (\alpha_1, \cdots, \alpha_n)^T$ die voldoet aan

$$\begin{array}{ccccc} a_{11}\alpha_1 + a_{12}\alpha_2 + & \cdots & +a_{1n}\alpha_n & = & \beta_1 \\ a_{21}\alpha_1 + a_{22}\alpha_2 + & \cdots & +a_{2n}\alpha_n & = & \beta_2 \\ \vdots & & & & \\ a_{n1}\alpha_1 + a_{n2}\alpha_2 + & \cdots & +a_{nn}\alpha_n & = & \beta_n \end{array} \qquad \text{of, m.a.w.} \quad A\vec{x} = \vec{b},$$

waarbij $A = (a_{ij})$ een gegeven reële niet-singuliere $n \times n$ matrix is en $\vec{b} = (\beta_1, \ldots, \beta_n)^T$ een geven reële n vektor. (LV)

Het *eigenwaardenprobleem*:

bepaal de niet triviale complexe vektoren (*eigenvektoren*) $\vec{v}$ en bijbehorende skalairen (*eigenwaarden*) $\lambda \in \mathbf{C}$ waarvoor geldt

$$A\vec{v} = \lambda\vec{v},$$

waarbij A een gegeven reële $n \times n$ matrix is. (EP)

Terloops besteden we aandacht aan ondermeer *het kleinste kwadraten probleem*:

bereken de reële n vektor $\vec{x}$ die voldoet aan $A\vec{x} = \vec{b}$ in de kleinste kwadaten zin, d.w.z.

$$\|A\vec{x} - \vec{b}\|_2 = \min\{\|A\vec{y} - \vec{b}\|_2 \,\big|\, \vec{y} \in \mathbf{R}^n\},$$

waarbij met $m \geq n$, A een gegeven reële $m \times n$ matrix is en $\vec{b}$ een geven reële m vektor. (KK)

We konstrueren numerieke oplosmethoden voor bovenstaande problemen en analyseren de stabiliteit en de efficiëntie van die methoden.

Zuiver wiskundig zijn de problemen erg elementair. Het merendeel van de theorie is behandeld in eerste en tweede jaars colleges algebra. In het eerste hoofdstuk van dit diktaat herhalen we kort de theorie die we hier nodig hebben. Zo hier en daar plaatsen we in dat hoofdstuk ook al een numerieke kanttekening.

De problemen zijn echter numeriek wiskundig van fundamenteel belang. We komen ze in de praktijk nogal eens tegen. Bovendien worden in veel numerieke oplosmethoden voor niet-lineaire problemen deze niet-lineaire problemen benaderd door lineaire van bovenstaand type. Kontinue problemen die numeriek moeten worden opgelost worden gediskretizeerd, hetgeen ook resulteert in problemen als boven. Van dit laatste geven we een voorbeeld.

Voorbeeld. Beschouw op $\Omega := (0,1) \times (0,1)$ met rand $\partial\Omega$, voor $f \in C(\Omega)$ en $g \in C(\partial\Omega)$ de *Laplace vergelijking*

$$\left.\begin{array}{cccc} \Delta u & = & f & \text{op} & \Omega \\ u & = & g & \text{op} & \partial\Omega, \end{array}\right\}$$

waaruit $u \in C^{(2)}(\Omega) \cap C(\bar{\Omega})$ opgelost moet worden.

Voor $n \in \mathbf{N}$, $h := \frac{1}{n}$ is $\mathbf{Z}_h := \{jh \,\big|\, j \in \mathbf{Z}\}$. We benaderen de Laplace vergelijking op het rooster $\Omega_h := \Omega \cap (\mathbf{Z}_h \times \mathbf{Z}_h)$ met rand $\partial\Omega_h := \partial\Omega \cap (\mathbf{Z}_h \times \mathbf{Z}_h)$ door de volgende 'gediskretizeerde Laplace vergelijking'

$$\left.\begin{array}{cccc} \Delta_h u_h & = & f & \text{op} & \Omega_h \\ u_h & = & g & \text{op} & \partial\Omega_h, \end{array}\right\}$$

waarbij $\Delta_h u_h(j_1 h, j_2 h) :=$

$$\tfrac{1}{h^2}(u(j_1 h - h, j_2 h) + u(j_1 h + h, j_2 h) + u(j_1 h, j_2 h - h) + u(j_1 h, j_2 h + h) - 4u(j_1 h, j_2 h)).$$

Op Ω_h behoort u_h een goede benadering te zijn van u. De vraag 'hoe goed deze benadering is (als funktie van h)', is een analytische die behandeld wordt in het college Numerieke Wiskunde D.

Hier merken we op dat de gediskretizeerde Laplace vergelijking een stelsel lineaire vergelijkingen is als in (LV).

(Met $\quad p_1 := (0,0), \quad p_2 := (h,0), \quad \ldots, \quad p_{n+1} := (1,0),$

$\qquad p_{n+2} := (0,h), \quad p_{n+3} := (h,h), \quad \ldots$

$\qquad \vdots$

$\qquad \ldots, \quad p_{(n+1)^2} := (1,1),$ kunnen we funkties $u_h : \Omega_h \cup \partial\Omega_h \to \mathbf{R}$ identificeren met vektoren $\vec{u}_h := (u_h(p_1), \ldots, u_h(p_{(n+1)^2}))^T \in \mathbf{R}^{(n+1)^2}$ en de diskretizeerde Laplace vergelijking met een matrix–vektor vergelijking $A\vec{u}_h = \vec{f}_h$ (stel zelf de matrix A en de rechterlid vektor $\vec{f}_h$ op).)

De numerieke methoden die we in dit college behandelen zijn *niet* speciaal afgestemd op matrices A met een speciale struktuur aan nullen: we gaan ervan uit dat de matrices $A = (a_{ij})$ *vol* zijn (d.w.z. mogelijk $a_{ij} \neq 0$ voor alle i,j). (Ga na dat de boven aangegeven matrix A die de gediskretizeerde Laplace vergelijking beschrijft slechts op vijf co-diagonalen coëfficiënten $\neq 0$ heeft.) Numerieke methoden die afgestemd zijn op matrices met een dergelijke speciale struktuur worden behandeld in het vervolg college Numerieke Wiskunde D.

We analyseren in dit college de efficiëntie van algorithmische versies van de diverse methoden niet in alle details. Hoe efficiënt een methode precies is hangt af van het type van de computer waarmee gerekend wordt. Een algorithme of methode kan op bijvoorbeeld een 'vektor-computer' veel sneller (efficiënter) zijn dan een ander algorithme, terwijl dat op een 'skalaire-computer' precies andersom kan uitpakken
(als $(\alpha_1, \ldots, \alpha_n)^T, (\beta_1, \ldots, \beta_n)^T \in \mathbf{R}^n$, dan voert een *vektor-computer* ruwweg gesproken de n operaties $\alpha_1\beta_1, \ldots, \alpha_n\beta_n$ of de n operaties $\alpha_1 + \beta_n, \ldots, \alpha_n + \beta_n$ tegelijkertijd in een rekenslag, in een 'tik van de klok' uit, terwijl een *skalaire-computer* de n operaties achter elkaar uitvoert en n 'tikken van de klok' nodig heeft).
In dit college meten we de efficiëntie door eenvoudig het aantal floating point operaties in een algorithme te tellen. We hanteren de volgende maat: een *flop* (*fl*oating point *op*eration) bestaat uit een vermenigvuldiging of deling plus eventueel optelling of aftrekking en de benodigde geheugenbeheeraktities: dus α, β, γ uit het geheugen vissen, $\alpha \cdot \beta + \gamma$ berekenen en het resultaat opbergen kost 1 flop. In de numerieke lineaire algebra zijn de meeste skalaire operaties van dit type "$\alpha \cdot \beta + \gamma$". Een efficiëntie analyse die bestaat uit het tellen van het aantal flops vergt geen speciale wiskundige kennis of vaardigheid maar geeft voor skalaire machines wel een behoorlijk goede indruk over de efficiëntie van het algorithme in vergelijking tot een ander algorithme op zo'n zelfde machine. Voor vektor-computers kan men een soortgelijke informatieve grootheid invoeren die gemakkelijk te hanteren is; we doen dat hier niet.
1 flop noemt men ook wel 1AV (*accumulatieve vermenigvuldiging*).

Opgave. Een matrix-matrix vermenigvuldiging kost voor volle reële $n \times n$ matrices n^3 flop en en matrix-vektor vermenigvuldiging n^2 flop. Ga dat na.

In onze stabiliteitsanalyses is $\bar{\xi}$ de *relatieve machine precisie*: dit is de afstand tussen 1 en het dichtst bij gelegen ander getal dat in de machine gerepresenteerd kan worden, m.a.w.; $\bar{\xi}$ is de relatieve afstand tussen twee opeenvolgende computer getallen. ξ representeert telkens een getal tussen $-\bar{\xi}$ en $+\bar{\xi}$. Op verschillende plaatsen kan ξ telkens een ander getal uit $[-\bar{\xi}, +\bar{\xi}]$ representeren. De formules waarin ξvoorkomt dienen van links naar rechts gelezen te worden; dus $\xi = 2\xi$ is korrekt, terwijl $2\xi = \xi$ fout is (de eerste uitdrukking zegt in feite dat $[-\bar{\xi}, +\bar{\xi}\,] \subseteq 2[-\bar{\xi}, +\bar{\xi}\,] = [-2\bar{\xi}, +2\bar{\xi}\,]$, terwijl de tweede betoogt dat $[-2\bar{\xi}, +2\bar{\xi}\,] \subseteq [-\bar{\xi}, +\bar{\xi}\,]$). We nemen aan dat $\bar{\xi} \ll 1$ en we verwaarlozen vaak bijvoorbeeld ξ^2 ten opzichte van ξ; zo schrijven we $(1+\xi)^2 = 1 + 2\xi$.
We nemen aan dat iedere floating point operatie uitgevoerd wordt met hoogstens een relatieve fout ξ: als bijvoorbeeld '+' de machine optelling voorstelt dan

$$\alpha\text{'+'}\beta = (\alpha + \beta)(1 + \xi), \text{ etc..}$$

(We verwijzen naar het college diktaat Numerieke Wiskunde A voor meer details, voorbeelden en opgaven over de 'ξ-analyse', de analyse van de evaluatiefouten middels de axioma's en konventies die we zo juist aangetipt hebben.)

Op een paar plaatsen illustreren we onze beweringen door in bijvoorbeeld 8 *decimalen te rekenen*. Onze getallen zijn dan van de vorm $m10^p$ met $m, p \in \mathbf{Z}$ en $|m| < 10^8$. Iedere rekenoperatie, $+, -, \times, /$ tussen twee getallen van deze vorm voeren we exact uit en ronden we dan af op een getal van deze vorm.

Een algemene matrix theorie eenvoudig gepresenteerd kan men vinden in [11]. Een tamelijk volledige meer geavanceerde matrix theorie staat in [5]. Er zijn diverse boeken die een inleiding geven in de numerieke wiskunde en ook een paar hoofdstukken besteden aan numerieke lineaire algebra. We denken dan bijvoorbeeld aan []. Er zijn een paar elementaire boeken die speciaal geschreven zijn over matrixrekenen [10, 4, 3]. We bevelen speciaal het boek [3] van Golub en van Loan aan. Het boek van Wilkinson [15] is nog steeds een goede bron bij hoofdstuk 2. Een standaard werk bij de theorie in hoofdstuk 3 is [6]. Een goed boek over het materiaal in hoofdstuk 4 is [8].

De methoden die we in dit college bespreken zijn geïmplementeerd in diverse programma paketten. Een drietal paketten voor matrixrekenen wordt zeer veel gebruikt: EISPACK (zie [9, 2]), LINPACK (zie [1]) en MATLAB (zie [7]). Verder vindt men numerieke lineaire algebra routines in de NAG Fortran Library en voor PC, de NAG Fortran PC50 Library. De NAG Library is niet speciaal voor matrixrekenen opgesteld. Dit pakket bevat dan ook vele niet-algebra subroutines. MATLAB is een programma pakket dat op iedere computer (ook op iedere PC-achtige computer) draait. MATLAB is de gebruikers vriendelijke bewerking van EISPACK en LINPACK. Men kan in MATLAB progammeren; men zou MATLAB een vierde generatie programmeertaal kunnen noemen.

NUMERICAL RECIPES is een collectie korte FORTAN 77 routines (en elektronisch vertaalde PASCAL procedures) die op elkaar zijn afgestemd en bestaan uit één eenvoudig numeriek algorithme (zie [14, 12, 13]) maar waarbij niet echt geprobeerd is met alle numerieke problemen rekening te houden. In de kollektie zijn ook een aantal 'non sophisticated' lineaire algebra algorithmes opgenomen.

Referenties

[1] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart. *LINPACK Users' Guide.* SIAM Publications, Philadelphia, 1979.

[2] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler. *Matrix Eigensystem Routines: EISPACK Guide Extension.* Volume 51 of *Lectures Notes in Computer Science*, Springer-Verlag, New York, 1977.

[3] Gene H. Golub and Charles F. van Loan. *Matrix Computations.* North Oxford Academic, Oxford, 1983.

[4] A. R. Gourlay and G. A. Watson. *Computational Methods for Matrix Eigenproblems.* John Wiley, New York, 1973.

[5] R. A. Horn and C. A. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, etc., 1985.

[6] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problemes.* Prentice-Hall, Inc., Englewood Cliffs, New York, 1980.

[7] C. B. Moler. *MATLAB User's Guide.* Technical Report, Department of Computer Science, University of New Mexico, Albuquerque, New Mexico, 1980. Technical Report CS81-1.

[8] Beresford N. Parlett. *The Symmetric Eigenvalue Problem.* Prentice-Hall, Inc., Englewood Cliffs, New York, 1980.

[9] B. T. Smith, J. M. Boyle, J. J. Dongarra, B.S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler. *Matrix Eigensystem Routines: EISPACK guide.* Volume 6 of *Lecture Notes in Computer Science*, Springer-Verlag, New York, 1976.

[10] G. W. Stewart. *Introduction to Matrix Computations.* Academic Press, New York, 1973.

[11] G. Strang. *Linear Algebra and its Applications.* Academic Press, New York, 1976.

[12] W. T. Vetterling. *Numerical Recipes: example book (FORTRAN).* Cambridge University Press, Cambridge, etc., 1985.

[13] W. T. Vetterling. *Numerical Recipes: example book (PASCAL).* Cambridge University Press, Cambridge, etc., 1985.

[14] W. H. Press et al.. *Numerical Recipes: the Art of Sceintific Computing.* Cambridge University Press, Cambridge, etc., 1986.

[15] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

1 Notaties, konventies en eenvoudige eigenschappen

1.1 Lineaire ruimten

(1.1.1) Laat $\mathbf{K} = \mathbf{R}$ of $\mathbf{K} = \mathbf{C}$.
In de numerieke praktijk werken we het liefst met reële getallen. In theoretische analyses, met name die waarin eigenwaarden betrokken zijn, werken we vaak met complexe getallen.

$\mathbf{X}$ is een n dimensionale lineaire ruimte over $\mathbf{K}$.

De ruimte $\mathbf{K}^n$ is het standaard voorbeeld van een dergelijke ruimte $\mathbf{X}$; in zekere zin is dat zelfs de 'enige' ruimte van dit soort. De standaard basis vektoren in $\mathbf{K}^n$ noteren we met $e_1, e_2, \ldots, e_n$: $e_1 = (1, 0, 0, \ldots, 0)^T$, $e_2 = (0, 1, 0, \ldots, 0)^T$, ... Als $x_1, x_2, \ldots, x_n$ een basis is van $\mathbf{X}$, dan kunnen we $\mathbf{X}$ als volgt identificeren met $\mathbf{K}^n$. De afbeelding van $\mathbf{X}$ naar $\mathbf{K}^n$ die $x = \sum_{j=1}^n \alpha_j x_j$ overvoert in $\vec{x} := (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$, is een lineaire bijektie.

(1.1.2) Matrices. Laat $\mathbf{X}$ en $\mathbf{Y}$ lineaire ruimte zijn met $\dim \mathbf{X} = n$ en $\dim \mathbf{Y} = m$. De lineaire ruimte van alle lineaire afbeeldingen $\mathcal{A}$ van $\mathbf{X}$ naar $\mathbf{Y}$ noteren we met $\mathcal{L}(\mathbf{X}, \mathbf{Y})$. Als $x_1, \ldots, x_n$ een basis is van $\mathbf{X}$ en $y_1, \ldots, y_m$ een basis van $\mathbf{Y}$ dan kunnen we $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ als volgt identificeren met de ruimte $\mathbf{M}_{m\times n}(\mathbf{K})$ van alle $m \times n$ matrices met koëfficiënten in $\mathbf{K}$. De afbeelding van $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ naar $\mathbf{M}_{m\times n}(\mathbf{K})$ die $\mathcal{A}$ overvoert in $A = (a_{ij})$ is een lineaire bijektie als

$$(a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

zo is dat $\mathcal{A}x_j = \sum_{i=1}^m a_{ij} y_i \quad (j = 1, \ldots, n)$.

Merk op dat de correspondentie tussen een lineaire afbeelding en diens matrix afhangt van de gekozen basis.

Laat zowel $x_1, \ldots, x_n$ als $x'_1, \ldots, x'_n$ een basis van $\mathbf{X}$ zijn . Laat $\chi_{ij} \in \mathbf{K}$ zo zijn dat $x'_i = \sum_j \chi_{ij} x_j$ voor alle i. Beschouw de afbeelding in $\mathcal{L}(\mathbf{X}, \mathbf{X})$ die aan x_i de vektor x'_i toevoegt $(i = 1, \ldots, n)$. Dan is $X := (\chi_{ij})$ precies de matrix van deze afbeelding ten opzichte van de basis $x_1, \ldots, x_n$ in de origineelruimte $\mathbf{X}$ *en* in de beeldruimte $\mathbf{X}$. De i-de kolom van X 'is' de basis vektor x'_i geschreven ten opzichte van de basis $x_1, \ldots, x_n$.

Laat zowel $y_1, \ldots, y_m$ als $y'_1, \ldots, y'_m$ een basis van $\mathbf{Y}$ zijn. Laat de matrix $Y = (\upsilon_{ij})$ zo zijn dat $y'_i = \sum_j \upsilon_{ij} y_j$. Als $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$ de matrix A heeft t.o.v. $x'_1, \ldots, x'_n$ in $\mathbf{X}$ en $y'_1, \ldots, y'_m$ in $\mathbf{Y}$ dan is $Y^{-1}AX$ de matrix van $\mathcal{A}$ t.o.v. $x_1, \ldots, x_n$ en $y_1, \ldots, y_m$.

(1.1.3) Konventies. We schrijven $\mathcal{L}(\mathbf{X})$ in plaats van $\mathcal{L}(\mathbf{X}, \mathbf{X})$ en $\mathbf{M}_n(\mathbf{K})$ in plaats van $\mathbf{M}_{n\times n}(\mathbf{K})$. Als uit de kontekst duidelijk is wat $\mathbf{K}$ is of als geen rol speelt wat $\mathbf{K}$ is laten we de toevoeging '$(\mathbf{K})$' in $\mathbf{M}_{m\times n}(\mathbf{K})$ en $\mathbf{M}_n(\mathbf{K})$ achterwege.

Als uit de kontekst duidelijk is welke basis we in $\mathbf{X}$ en $\mathbf{Y}$ gekozen hebben, dan identificeren we elementen uit $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ zonder verdere toelichting met hun matrix; dit zal vaak het geval zijn als $\mathbf{X} = \mathbf{K}^n$, $\mathbf{Y} = \mathbf{K}^m$ en we met de standaard bases werken.

(1.1.4) Notaties. Zij $A = (a_{ij}) \in \mathbf{M}_{m\times n}$.
De *gespiegelde* of *getransponeerde* A^T van A de matrix (b_{ij}) met $b_{ij} = a_{ji}$. $\bar{A} := (\bar{a}_{ij})$ is de *gekonjugeerde* van A en de *geadjungeerde* A^H van A is gegeven door $A^H := \bar{A}^T$. We schrijven ook wel A^* in plaats van A^H (zie ook (1.5.9)). Als $A \in \mathbf{M}_{m\times n}(\mathbf{R})$ dan $A^T = A^H$.
(Merk op dat we vektoren in $\mathbf{K}^n$ kunnen opvatten als elementen van $\mathbf{M}_{n\times 1}(\mathbf{K})$; de notatie in (1.1.1) is konsistent met deze interpretatie.)

Op een paar plaatsen gebruiken we de volgende notatie.
$A \succeq 0$ wil zeggen dat $a_{ij} \geq 0$ voor alle i, j. We schrijven $A \succeq B$ voor een $B \in \mathbf{M}_n$ als $A - B \succeq 0$. $|A|$ is de matrix $(|a_{ij}|)$.

1.2 Normen

We wensen oplossingen van stelsels lineaire vergelijkingen numeriek te benaderen. Van zo'n numerieke benadering moeten we natuurlijk ook weten hoe nauwkeurig deze de echte oplossing benadert. We dienen dus te beschikken over een afstandsbegrip in 'n lineaire ruimte. Een norm is zo'n afstandsbegrip dat bovendien nog compatibel is met de lineaire struktuur.

(1.2.1) Definitie. Zij $\mathbf{X}$ een lineaire ruimte.
Een afbeelding $\|\cdot\| : \mathbf{X} \to [0,\infty)$ is een *norm* als deze de volgende drie eigenschappen heeft.

(a) $\|x\| = 0$ dan slechts dan als $x = 0$ $\quad (x \in \mathbf{X})$.
(b) $\|\lambda x\| = |\lambda|\|x\|$ $\quad (\lambda \in \mathbf{K}, x \in \mathbf{X})$.
(c) $\|x+y\| \leq \|x\| + \|y\|$ $\quad (x, y \in \mathbf{X})$.

(1.2.2) Voorbeelden. Zij $\mathbf{X} = \mathbf{K}^n$.
Kies $p \in [1,\infty)$. De volgende voorschriften definiëren normen op $\mathbf{X}$.

$$\|\vec{x}\|_\infty = \max\{|\alpha_i| \,\big|\, i = 1,\ldots,n\} \quad \text{voor alle} \quad \vec{x} = (\alpha_1,\ldots,\alpha_n)^T \in \mathbf{X}. \tag{1}$$
$$\|\vec{x}\|_2 = \sqrt{|\alpha_1|^2 + \cdots |\alpha_n|^2} \quad \text{voor alle} \quad \vec{x} = (\alpha_1,\ldots,\alpha_n)^T \in \mathbf{X}. \tag{2}$$
$$\|\vec{x}\|_p = \sqrt[p]{|\alpha_1|^p + \cdots |\alpha_n|^p} \quad \text{voor alle} \quad \vec{x} = (\alpha_1,\ldots,\alpha_n)^T \in \mathbf{X}. \tag{3}$$

De norm in (1) wordt *sup-norm, max-norm* of ∞*-norm* genoemd. De norm in (2) is de 2*-norm* en induceert de *Euclidische afstand* $\|\vec{x} - \vec{y}\|_2$ tussen $\vec{x}$ en $\vec{y}$ in $\mathbf{K}^n$. De 2-norm is een speciale p*-norm* of *Hölder norm* die in (3) gedefinieerd is. De sup-norm kan door $p \to \infty$ te nemen ook gezien worden als een speciale p-norm. De 1-norm, 2-norm en de ∞-norm zullen in dit college onze speciale aandacht hebben.

We zullen in dit college diverse processen beschouwen die iteratief een rij (x_k) van vektoren produceren die de oplossing x van een probleem beogen te benaderen. In zo'n geval willen we dus dat de rij (x_k) naar x *konvergeert*: we willen dat $\|x_k - x\| \to 0$ als $k \to \infty$ ten opzichte van een of andere norm $\|\cdot\|$ op $\mathbf{X}$. De volgende stelling laat zien dat ons konvergentie begrip niet van de gekozen norm afhangt.

(1.2.3) Stelling. *Zij* $\mathbf{X}$ *een n-dimensionale lineaire ruimte met basis* $x_1,\ldots,x_n$ *en voorzien van een norm* $\|\cdot\|$.
Er zijn $\Gamma, \gamma \in (0,\infty)$, $\gamma < \Gamma$ *zodat*

$$\gamma\|(\alpha_1,\ldots,\alpha_n)^T\|_\infty \leq \|x\| \leq \Gamma\|(\alpha_1,\ldots,\alpha_n)^T\|_\infty \quad \textit{voor alle} \quad x = \sum \alpha_j x_j \in \mathbf{X}. \tag{4}$$

Bewijs. $\|x\| \leq \sum\|\alpha_j x_j\| = \sum|\alpha_j|\|x_j\| \leq \max|\alpha_j| \sum\|x_j\|$, hetgeen met $\Gamma := \sum\|x_j\|$ de tweede ongelijkheid levert.

Beschouw nu de afbeelding $\Phi : \mathbf{K}^n \to [0,\infty)$ gedefiniëerd door $\Phi((\alpha_1,\ldots,\alpha_n)^T) := \|\sum \alpha_j x_j\|$. Φ is kontinu en heeft dus op de begrense gesloten verzameling $K := \{\vec{x} \in \mathbf{K}^n \,\big|\, \|\vec{x}\|_\infty = 1\}$ een minimum. Noem dit minimum γ. γ is precies het getal waarvoor de eerste ongelijkheid korrekt is; de verdere details laten we aan de lezer over. □

(1.2.4) Gevolg. *Op een eindig dimensionale ruimte* $\mathbf{X}$ *zijn alle normen* equivalent*: als* $\|\cdot\|$ *en* $\|\cdot\|_*$ *normen zijn op* $\mathbf{X}$ *dan zijn er getallen* $\tilde{\Gamma}, \tilde{\gamma} \in (0,\infty)$, $\tilde{\gamma} < \tilde{\Gamma}$ *zodat*

$$\tilde{\gamma}\|x\|_* \leq \|x\| \leq \tilde{\Gamma}\|x\|_* \quad \textit{voor iedere} \quad x \in \mathbf{X}.$$

Bewijs. Als Γ, γ als in de stelling zijn bij $\|\cdot\|$ en Γ_*, γ_* de anologe getallen bij $\|\cdot\|_*$ dan voldoen $\tilde{\Gamma} = \frac{\Gamma}{\gamma_*}$ en $\tilde{\gamma} = \frac{\gamma}{\Gamma_*}$. □

(1.2.5) Opgave. Laat $\mathbf{X}$ een eindig dimensionale ruimte zijn voorzien van een norm $\|\cdot\|$ en een norm $\|\cdot\|_*$. Laat (x_k) een rij in $\mathbf{X}$ zijn en laat $x \in \mathbf{X}$. Bewijs dat $\|x_k - x\| \to 0$ voor $k \to \infty$ dan en slechts dan als $\|x_k - x\|_* \to 0$ voor $k \to \infty$

We schrijven $x_k \to x \quad (k \to \infty)$ of $\lim_{k\to\infty} x_k = x$ in plaats van $\|x_k - x\| \to 0 \quad (k \to \infty)$.

(1.2.6) Gevolg. *Iedere eindig dimensionale ruimte* $\mathbf{X}$ *is* volledig*:*
als $\|\cdot\|$ *een norm op* $\mathbf{X}$ *is en* (x_k) *is een Cauchy rij in* $\mathbf{X}$ $\quad(\|x_k - x_l\| \to 0$ als $k, l \to \infty)$ *dan is er een* $x \in \mathbf{X}$ *zodat* $x_k \to x \quad (k \to \infty)$.

Bewijs. $\mathbf{K}$ is volledig en dus is $\mathbf{K}^n$ volledig (t.o.v. $\|\cdot\|_\infty$). De bewering volgt nu met behulp van (1.2.3). □

(1.2.7) Voorbeelden. Zij $\mathbf{X} = \mathbf{K}^n$. Dan geldt

$$\|\vec{x}\|_\infty \leq \|\vec{x}\|_2 \leq \sqrt{n}\|\vec{x}\|_\infty \qquad \text{voor alle} \quad \vec{x} \in \mathbf{K}^n \tag{5}$$

$$\|\vec{x}\|_1 \leq \sqrt{n}\|\vec{x}\|_2 \leq \sqrt{n}\|\vec{x}\|_1 \qquad \text{voor alle} \quad \vec{x} \in \mathbf{K}^n \tag{6}$$

(1.2.8) Opgave. Bewijs de ongelijkheden in (1.2.7) en laat zien dat de gelijktekens kunnen optreden.

Bewijs dat voor alle $p, q \in [1, \infty]$, $p \geq q$ en $\vec{x} \in \mathbf{K}^n$ geldt $\|\vec{x}\|_p \leq \|\vec{x}\|_q$ (Hint: differentiëer $t \rightsquigarrow \|\vec{x}\|_t$ voor $\vec{x} = (|\alpha_1|, \ldots, |\alpha_n|)^T$).

Vanuit een zuiver wiskundig oogpunt speelt het blijkbaar geen essentiële rol welke norm men op de eindig dimensionale ruimte $\mathbf{X}$ kiest.

Vanuit een numeriek standpunt is keuze van de norm wel degelijk belangrijk: de norm $\|\cdot\|$ en de norm $x \rightsquigarrow 10^{-10}\|x\|$ zijn equivalent maar van de mededeling dat x_k de oplossing x benadert met een fout 10^{-8} zal men toch graag willen weten in welke van die twee normen die fout $x - x_k$ gemeten is.

Als een eindig dimensionaal probleem een oneindig dimensionaal probleem benadert zal men ook willen dat norm op het eindig dimensionale probleem de norm op het oneindig dimensionale in zekere zin benadert
(stel dat men, voor $n \in \mathbf{N}$ met $h := \frac{1}{n-1}$, functies f in $C[0,1]$ wil benaderen door vektoren $\vec{f} := (f(0), f(h), f(2h), \ldots, f(nh - h))^T$ in $\mathbf{K}^n$. Als men de functies in $C[0,1]$ meet in de norm $\|f\|_2 := \sqrt{\int |f(t)|^2\,dt}$ dan zal men op $\mathbf{K}^n$ de 'geschaalde 2-norm' $\vec{x} \rightsquigarrow \sqrt{\frac{1}{n}}\|\vec{x}\|_2$ preferen, immers; met $f \equiv 1$ is $\sqrt{\int |f(t)|^2\,dt} = 1$ terwijl $\|(f(0), \ldots, f(1))^T\|_2 = \sqrt{n}$.)

(1.2.9) Bewering. *Laat* $\mathbf{X}, \mathbf{Y}$ *eindig dimensionale ruimten zijn.* $\mathbf{X}$ *is voorzien van de norm* $\|\cdot\|_{\mathbf{X}}$ *en* $\mathbf{Y}$ *van de norm* $\|\cdot\|_{\mathbf{Y}}$*. Voor* $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$ *definiëren we*

$$\|\mathcal{A}\| := \|\mathcal{A}\|_{\mathbf{Y}\leftarrow\mathbf{X}} := \sup\{\frac{\|\mathcal{A}x\|_{\mathbf{Y}}}{\|x\|_{\mathbf{X}}} \,\Big|\, x \in \mathbf{X}, x \neq 0\} = \{\|\mathcal{A}x\|_{\mathbf{Y}} \,\Big|\, x \in \mathbf{X}, \|x\|_{\mathbf{X}} = 1\}.$$

Dan $\|\cdot\|$ *is een norm op de lineaire ruimte* $\mathcal{L}(\mathbf{X}, \mathbf{Y})$*; dit is een zogenaamde* geassocieerde norm.

Een geassocieerde norm heeft de volgende eigenschappen:

(a) $\|\mathcal{A}x\|_{\mathbf{Y}} \leq \|\mathcal{A}\|\,\|x\|_{\mathbf{X}}$ *voor alle* $(x \in \mathbf{X})$.

(b) $\|\mathcal{I}\| = 1$ *als* $\mathbf{Y} = \mathbf{X}$ *en* $\mathcal{I}$ *de identiteit is.*

(c) $\|\mathcal{B} \circ \mathcal{A}\| \leq \|\mathcal{B}\|\,\|\mathcal{A}\|$ *voor alle* $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$, $\mathcal{B} \in \mathcal{L}(\mathbf{Y}, \mathbf{Z})$*, waarbij* $\mathbf{Z}$ *ook voorzien is van een norm: een geassocieerde norm is* multiplikatief.

(We laten het bewijs als een opgave aan de lezer over.) □

(1.2.10) Opgave. Bewijs bovenstaande bewering.

(1.2.11) Voorbeelden. Zij $\mathbf{X} = \mathbf{K}^n$ en $\mathbf{Y} = \mathbf{K}^m$.
We beschouwen geassocieerde normen op $\mathcal{L}(\mathbf{X}, \mathbf{Y})$. Ten opzichte van de standaard bases identificeren we $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ met $\mathbf{M}_{m\times n}$. De normen die daardoor op ruimten der matrices geïnduceerd worden noemen we ook geassocieerd.

We nemen nu op de origineel ruimte en op de beeld ruimte dezelfde norm.

Beschouw op $\mathbf{X}$ de ∞-norm. Noteer de geassocieerde norm ook met $\|\cdot\|_\infty$. Dan geldt voor iedere $A = (a_{ij}) \in \mathbf{M}_{m\times n}$

$$\|A\|_\infty = \sup\{\|A\vec{x}\|_\infty \,\big|\, \vec{x} \in \mathbf{K}^n, \|\vec{x}\|_\infty \le 1\} = \max_i \sum_j |a_{ij}|, \quad \textit{de maximale rij absoluut som}, \text{ en}$$

$$\|A\|_1 = \sup\{\|A\vec{x}\|_1 \,\big|\, \vec{x} \in \mathbf{K}^n, \|\vec{x}\|_1 \le 1\} = \max_j \sum_i |a_{ij}|, \quad \textit{de maximale kolom absoluut som.}$$

Voor de '2-norm' van een matrix is het veel moeilijker zoniet onmogelijk om een gemakkelijk berekenbare uitdrukking te geven. Bij schattingen kunnen de volgende ongelijkheden erg nuttig zijn (zie ook (1.2.17).

$$\|A\|_\infty \le \sqrt{n}\,\|A\|_2 \le n\|A\|_\infty \quad \text{voor alle} \quad A \in \mathbf{M}_{m\times n}.$$

(1.2.12) Opgave. Bewijs dat bovenstaande relaties korrekt zijn en laat zien dat ze *scherp* zijn (d.w.z. laat zien dat de gelijktekens kunnen optreden).

(1.2.13) Opgave. Zij $A, B \in \mathbf{M}_{m\times n}$ en $p \in [1,\infty]$. Bewijs
$\|A\|_p \le \|\,|A|\,\|_p$ en $A \succeq B \succeq 0 \;\Rightarrow\; \|A\|_p \ge \|B\|_p$.

We kunnen een lineaire bijektie definiëren tussen $\mathbf{M}_{m\times n}$ en $\mathbf{K}^{mn}$ (hoe?). Bijgevolg induceert ieder norm op $\mathbf{K}^{nm}$ een norm op $\mathbf{M}_{m\times n}$. Twee van zulke normen zijn voor ons van speciaal belang. Deze normen zijn voor konkrete matrices zeer gemakkelijk te berekenen en kunnen gebruikt worden middels schattingen als in (1.2.16).

(1.2.14) Definitie. Voor $A \in \mathbf{M}_{m\times n}$ definiëren we de norm $\|A\|_m$ en de *Frobenius norm* (of em Hilbert–Schmidt norm) $\|A\|_F$ door

$$\|A\|_m := \max\{|a_{ij}| \,\big|\, i = 1,\ldots,n,\ j = 1,\ldots,m\} \quad \text{voor alle} \quad A = (a_{ij}) \in \mathbf{M}_{m\times n}. \quad (7)$$

$$\|A\|_F := \sqrt{\textstyle\sum_{i=1}^n \|Ae_i\|_2^2} = \sqrt{\textstyle\sum_{i,j} |a_{ij}|^2} \quad \text{voor alle} \quad A = (a_{ij}) \in \mathbf{M}_{m\times n}. \quad (8)$$

(1.2.15) Opgave. Ga na dat geen van beide normen $\|\cdot\|_m$ en $\|\cdot\|_F$ geassocieerd is. (Hint: ga na dat geen van de normen alle eigenschappen (a), (b), (c) uit (1.2.9) heeft.)

(1.2.16) Bewering. *Voor iedere $A \in \mathbf{M}_{m\times n}$ en $B \in \mathbf{M}_{n\times l}$ geldt*

$$\|AB\|_F \le \|A\|_2\,\|B\|_F,$$

$$\|A\|_2 \le \|A\|_F \le \sqrt{n}\,\|A\|_2$$

Bewijs.

$$\|AB\|_F = \sqrt{\sum_{j=1}^{l} \|ABe_j\|_2^2} \le \sqrt{\sum_j \|A\|_2^2\,\|Be_j\|_2^2} = \|A\|_2\,\|B\|_F,$$

Verder is voor $\vec{x} = \sum_j \alpha_j e_j$, volgens de ongelijkheid van Cauchy–Schwarz

$$\|A\vec{x}\|_2 \le \sum_j |\alpha_j|\,\|Ae_j\|_2 \le \sqrt{\sum_j |\alpha_j|^2}\sqrt{\sum_j \|Ae_j\|_2^2} = \|\vec{x}\|_2\,\|A\|_F$$

en

$$\|A\|_F = \sqrt{\sum \|Ae_j\|_2^2} \le \sqrt{\sum \|A\|_2^2\,\|e_j\|_2^2} = \sqrt{n}\,\|A\|_2.$$

$\square$

(1.2.17) Opgave. Ga na dat voor iedere $A \in \mathbf{M}_n$ geldt

$$\frac{\|A\|_1}{\|A\|_2}, \frac{\|A\|_\infty}{\|A\|_2}, \frac{\|A\|_1}{\|A\|_F}, \frac{\|A\|_\infty}{\|A\|_F} \in [\frac{1}{\sqrt{n}}, \sqrt{n}] \quad \text{en} \quad \frac{\|A\|_2}{\|A\|_F} \in [\frac{1}{\sqrt{n}}, 1].$$

Ga na dat de grenzen kunnen worden aangenomen. Geef zulke scherpe schattingen ook voor $A \in \mathbf{M}_{m \times n}$.

1.3 Neumann reeksen

In deze paragraaf laten aan de hand van wat eenvoudige voorbeelden zien hoe normen funktioneren in numerieke toepassingen.

De struktuur van deze paragraaf is karakteristiek voor de struktuur van dit college: we behandelen een zuiver wiskundige "theorie", die we enerzijds gebruiken bij de konstruktie van numerieke oplosmethoden en anderzijds bij de analyse van de (evaluatie)fouten.

In het lemma worden meetkundige reeksen $((\zeta^k)$ met $\zeta \in \mathbf{C}$ en som $\frac{1}{1-\zeta}$ als $|\zeta| < 1)$ gegeven voor afbeeldingen in $\mathcal{L}(\mathbf{X})$ en dus ook voor matrices in $\mathbf{M}_n$ (merk op dat elementen ζ in $\mathbf{C}$ gezien kunnen worden als 1×1 matrices).

We introduceren vervolgens een klasse van methoden die voor zekere lineaire stelsels vergelijkingen iteratief een rij benaderende oplossingen produceren. Deze klasse van methoden speelt in de numerieke wiskunde een grote rol. Wij beschikken op dit moment echter nog niet over voldoende vergelijkings materiaal om de verdiensten van de methoden voldoende uit de verf te kunnen laten komen.

Verder laten we zien hoe de normen kunnen funktioneren in een evaluatie fouten analyse.

(1.3.1) Lemma. *Laat* $\|\cdot\|$ *een multiplikatieve norm zijn op* $\mathcal{L}(\mathbf{X})$. *Zij* $\mathcal{B} \in \mathcal{L}(\mathbf{X})$. *Als* $\rho := \|\mathcal{B}\| < 1$ *dan is* $\mathcal{I} - \mathcal{B}$ *inverteerbaar in* $\mathcal{L}(\mathbf{X})$ *en*

$$(\mathcal{I} - \mathcal{B})^{-1} = \sum_{j=0}^{\infty} \mathcal{B}^j, \qquad \|(\mathcal{I} - \mathcal{B})^{-1}\| \leq \frac{1}{1 - \|\mathcal{B}\|}$$

(waarbij $\mathcal{B}^0 := \mathcal{I}$. De reeks $\sum \mathcal{B}^k$ is een zogenaamde *Neumann reeks*).

Bewijs. Omdat $\|\sum_{j=k}^{l} \mathcal{B}^j\| \leq \sum_{j=k}^{\infty} \|\mathcal{B}\|^j \leq \frac{\|\mathcal{B}\|^k}{1-\|\mathcal{B}\|} = \frac{\rho^k}{1-\rho} \to 0 \qquad (k, l \to \infty)$,
is $(\sum_{j=0}^{l} \mathcal{B}^j)_l$ een Cauchy rij in $\mathcal{L}(\mathbf{X})$. $\mathcal{L}(\mathbf{X})$ is volledig (zie (1.2.6)). Dus is er een $\mathcal{C} \in \mathcal{L}(\mathbf{X})$ zodat $\mathcal{C} = \lim_{k \to \infty} \sum_{j=0}^{k} \mathcal{B}^j$.

We bewijzen dat $\|\mathcal{C}(\mathcal{I} - \mathcal{B}) - \mathcal{I}\| = 0$; dan $\mathcal{C}(\mathcal{I} - \mathcal{B}) = \mathcal{I}$ en dus $\mathcal{C} = (\mathcal{I} - \mathcal{B})^{-1}$.
Welnu, $\|\mathcal{C}(\mathcal{I} - \mathcal{B}) - \mathcal{I}\| \leq \|\sum_0^l \mathcal{B}^j(\mathcal{I} - \mathcal{B}) - \mathcal{I}\| + \|(\mathcal{C} - \sum_0^l \mathcal{B}^j)(\mathcal{I} - \mathcal{B})\| \leq$
$\|\mathcal{I} - \mathcal{B}^{l+1} - \mathcal{I}\| + \|(\mathcal{C} - \sum_0^l \mathcal{B}^j)(\mathcal{I} - \mathcal{B})\| \leq \|\mathcal{B}\|^{l+1} + \|\mathcal{C} - \sum_0^l \mathcal{B}^j\| \, \|\mathcal{I} - \mathcal{B}\| \to 0$ als $l \to \infty$,
immers $\|\mathcal{B}\| < 1$ en $\mathcal{C} = \sum \mathcal{B}^j$. □

Met behulp van het lemma kan men vaak eenvoudig aantonen dat een matrix *niet-singulier* (d.w.z. inverteerbaar) is. We geven een voorbeeld (in (1.3.3)). Eerst voeren we een notatie in.

(1.3.2) Zij $A = (a_{ij}) \in \mathbf{M}_n$.
Schrijf $A = D + L + U$, waarbij D het *'diagonaal'* deel is van A, L het *'strikt beneden driehoeks deel'* en U het *'strikt boven driehoeks deel'*; om precies te zijn
$D := \operatorname{diag}(A) := \operatorname{diag}(a_{11}, \ldots, a_{nn}) = (d_{ij})$ met $d_{ii} = a_{ii}$ en $d_{ij} = 0 \ (i, j = 1, \ldots, n, i \neq j)$,
met $L = (l_{ij})$ is $l_{ij} = a_{ij}$ als $i > j$ en $l_{ij} = 0$ als $i \leq j \ (i, j = 1, \ldots, n)$,
met $U = (u_{ij})$ is $u_{ij} = a_{ij}$ als $i < j$ en $u_{ij} = 0$ als $i \geq j \ (i, j = 1, \ldots, n)$.

(1.3.3) Bewering. *Stel dat* A strikt rij diagonaal dominant *is;*
d.w.z. $a_{ii} > \sum_{j \neq i} |a_{ij}|$ *voor iedere* $i = 1, \ldots, n$. *Dan is* A *niet-singulier en met* $K := D$ *en* $R := -L - U$ *is*

$$\|K^{-1}R\|_\infty < 1 \quad \text{en} \quad A^{-1} = \sum_{j=0}^{\infty} (K^{-1}R)^j K^{-1}$$

Bewijs. Uit de diagonaal dominantie volgt dat $d_{ii} \neq 0$ voor iedere i en dus dat D^{-1} bestaat. Omdat bovendien

$$\rho := \|K^{-1}R\|_\infty = \max_i \sum_{j\neq i} \frac{|a_{ij}|}{|a_{ii}|} = \max_i \frac{\sum_{j\neq i} |a_{ij}|}{|a_{ii}|} < 1$$

is $I - K^{-1}R$ niet-singulier en dus ook $A = K(I - K^{-1}R)$;
$A^{-1} = (I - K^{-1}R)^{-1}K^{-1} = \sum_j (K^{-1}R)^j K^{-1}$. □

(1.3.4) Opgave. Als A strikt kolom diagonaal dominant is en $D \succeq 0$, $-U \succeq 0$, $-L \succeq 0$ dan is A niet-singulier en $A^{-1} \succeq 0$. Bewijs dit.

Iteratieve oplosmethoden

Zij $A = (a_{ij})$ een niet-singuliere matrix in $\mathbf{M}_n$.
Voor $\vec{b} \in \mathbf{K}^n$ zijn we geïnteresseerd in de oplossing $\vec{x} \in \mathbf{K}^n$ van de vergelijking

$$A\vec{x} = \vec{b}. \tag{9}$$

Vaak is het erg 'duur' om de inverse A^{-1} van een matrix A konkreet uit te rekenen. Omdat we alleen geinteresseerd zijn in de oplossing $\vec{x}$ en niet zozeer in A^{-1} proberen we oplosmethoden te bedenken die $\vec{x}$ opleveren of goed benaderen zonder expliciet gebruik te maken van deze 'dure' A^{-1}. Voor sommige matrices A kunnen we met behulp van het lemma de oplossing $\vec{x}$ opschrijven in termen van de konkreet gegeven matrix A en de gegeven vektor $\vec{b}$. Het spreekt dan ook voor zich dat we na willen gaan of we met die uitdrukking 'goedkoper' $\vec{x}$ kunnen bepalen of benaderen.

Voor het lemma moet de matrix A van de vorm $I - B$ zijn met B zodat $\|B\| < 1$ voor een of ander norm op $\mathbf{M}_n$. Dit lijkt een zware restriktie op de toepasbaarheid van het lemma te leggen. Toch kunnen we vaak door wat met (deel)matrices te goochelen als in het bewijs van (1.3.3) een situatie kreëren waarop het lemma wel van toepassing is, terwijl dat aanvankelijk niet het geval was. We geven een paar voorbeelden.

(1.3.5) Zij weer $A = D + L + U$ als in (1.3.2).
Schrijf $K := D$ en $R := -(L + U)$. Dan $A = K - R$. Stel dat $\rho := \|K^{-1}R\|_\infty < 1$.
Dan is A niet-singulier en $A^{-1} = \sum_j (K^{-1}R)^j K^{-1}$. Schrijf $\vec{x}_k := \sum_{j=0}^k (K^{-1}R)^j K^{-1}\vec{b}$. Dan

$$\vec{x}_k \to \vec{x} \quad (k \to \infty) \quad \text{en} \quad \|\vec{x}_k - \vec{x}\|_\infty \le \frac{\rho^{k+1}}{1-\rho}\|K^{-1}\vec{b}\|_\infty \le \rho^{k+1}\frac{1+\rho}{1-\rho}\|\vec{x}\|_\infty.$$

We hebben hier een manier om de oplossing $\vec{x}$ van het stelsel (9) te benaderen. Bovendien hebben we een schatting voor de grootte van de fout $\vec{x}_k - \vec{x}$ in deze benadering.

We staan even stil bij de manier waarop we, gegeven de formule voor $\vec{x}_k$, de $\vec{x}_k$ zo efficiënt mogelijk kunnen berekenen. (In (1.3.8) draaien we de gedachtengang om: we geven daar een 'algorithme' en leiden dan een foutschatting af.)
Merk op dat

$$\vec{x}_{k+1} = \textstyle\sum_{j=0}^{k+1} (K^{-1}R)^j K^{-1}\vec{b} = (K^{-1}R)\vec{x}_k + K^{-1}\vec{b} =$$
$$K^{-1}(R\vec{x}_k + \vec{b}) = K^{-1}(K\vec{x}_k - A\vec{x}_k + \vec{b}) = \vec{x}_k + K^{-1}(\vec{b} - A\vec{x}_k).$$

Dit leidt tot het volgende algorithme (met variant).

Algorithme.

$\vec{x}_0 = 0$; $k = 0$;
(a) $r_k = \vec{b} - A\vec{x}_k$
(bereken het *residu* r_k);
`los` $\vec{y}_k$ `op uit` $K\vec{y}_k = r_k$;
$\vec{x}_{k+1} = \vec{x}_k + \vec{y}_k$;
$k := k + 1$; `ga naar` (a)

$\vec{x}_0 = 0$; $k = 0$;
(a') $\vec{y}_k{}' = R\vec{x}_k + \vec{b}$;
`los` $\vec{x}_{k+1}$ `op uit`
$K\vec{x}_{k+1} = \vec{y}_k{}'$;
$k := k + 1$; `ga naar` (a')

Het algorithme produceert *iteratief* (uit $\vec{x}_k$ wordt $\vec{x}_{k+1}$ berekend) een rij benaderingen $\vec{x}_k$ van

$\vec{x}$. Dit proces dient uiteraard bij een zekere k_0 te stoppen. We zouden de foutschatting die we boven gezien hebben kunnen gebruiken om een geschikte 'eind' k_0 te vinden. Er zijn voor de praktijk andere listige stopkriteria bedacht. We gaan daar verder hier niet meer op in.

Hoe '*duur*' het is, hoeveel flop we nodig hebben, om met deze methode een geschikte benadering $\vec{x}_{k_0}$ van de echte oplossing $\vec{x}$ te berekenen hangt af van

(i) de nauwkeurigheid waarmee we de echte oplossing wensen te benaderen,

(ii) het aantal flop dat in een iteratiestap (de berekening van $\vec{x}_{k+1}$ uit $\vec{x}_k$) moet worden uitgevoerd,

(iii) de konvergentiesnelheid (de faktor waarmee per iteratiestap de grootte $\|\vec{x}_k - \vec{x}\,\|$ van de fout min of meer gereduceerd wordt; zie ook (1.3.8) en het bewijs van die bewering).

Verderop in dit college zullen we '*direkte*' oplosmethoden voor (9) behandelen. Deze direkte methoden produceren niet een rij $(\vec{x}_k)$ van benaderingen voor $\vec{x}$ maar, afgezien van evaluatie fouten (waar we het overigens hier ook nog niet over gehad hebben), leveren zij de oplossing $\vec{x}$ en, indien gewenst, de inverse A^{-1}.

Als vrijwel alle koëfficiënten $a_{ij} \neq 0$, dan kost het berekenen van $\vec{x}$ met de goedkoopste direkte methode ongeveer $\frac{1}{3}n^3$ flop. (A^{-1} berekenen kost ongeveer $\frac{4}{3}n^3$ flop.) Omdat een matrix-vektor vermenigvuldiging n^2 flop kost (ga dit na) is het goedkoper om met bovenstaande iteratieve methode $\vec{x}_k$ voor $k < \frac{1}{3}n$ te berekenen. Afhankelijk van de precisie waarmee men $\vec{x}$ wil kennen en van de konvergentiesnelheid van de iteratieve methode zal men de voorkeur geven aan een direkte of aan een iteratieve methode.

Als men weet dat een matrix A *ijl* is (als relatief veel matrix-koëfficiënten $a_{ij} = 0$ zijn), kan men een matrix-vektor vermenigvuldiging relatief gekoop uitvoeren. Direkte methoden produceren dan vaak tijdens de oplosproceduren volle matrices en zijn daarom dan relatief duur. In zo'n situatie is een iteratieve methode aantrekkelijk.

(1.3.6) Opgave. Zij $S = (s_{ij}) \in \mathbf{M}_n$ zodat $Se_1 = 0$, $Se_i = e_{i-1}$ $(i = 2, \ldots, n)$ $(s_{ij} = 1$ als $j = i-1$ en $s_{ij} = 0$ anders). Zij nu $A = I - \frac{1}{4}(S + S^T)$. Bewijs dat $A^{-1} = (b_{ij})$ bestaat en toon aan dat $b_{ij} > 0$ voor alle i, j.

Stel dat de b_{ij} expliciet bekend zijn. Voor $\vec{b} = (\beta_1, \ldots, \beta_n)^T \in \mathbf{K}^n$ met $\beta_j \neq 0$ $(j = 1, \ldots, n)$ is $\vec{x} \in \mathbf{K}^n$ de oplossing van '$A\vec{x} = \vec{b}$'. We zijn tevreden als we $\tilde{x} \in \mathbf{K}^n$ kunnen berekenen zodat $\frac{\|\vec{x} - \tilde{x}\|_\infty}{\|\vec{x}\,\|_\infty} < 2\,10^{-3}$. Voor $n \geq 30$ is het efficiënter met het algorithme te werken dan met A^{-1}. Ga dit na.

(1.3.7) Een iteratieve methode als in (1.3.5) kunnen we formuleren voor iedere 'splitsing' $A = K - R$; we hoeven ons niet te beperken tot het geval $K = D$. Zo'n methode produceert een rij $(\vec{x}_k)$ die naar de oplossing $\vec{x}$ van (9) konvergeert als K niet-singulier en bijvoorbeeld $\|K^{-1}R\|_\infty < 1$. Voor de praktijk moet K ook nog zo zijn dat bv. '$K\vec{x}_{k+1} = \vec{y}_k$'' goedkoop oplosbaar is.

Door over te stappen op een andere splitsing kan men hopen een sneller konvergent proces te vinden (een proces met een kleinere fout reduktie factor per stap: zie (iii) in (1.3.5)). De kosten per iteratiestap kunnen dan wel duurder worden (het aantal flop nodig per stap kan groter zijn; zie (ii)). Men moet er voor zorgen dat de totale kosten om de oplossing in een gewenste precisie te benaderen omlaag gaan. In het algemeen kan men niet zeggen welke splitsing de efficiëntste methode geeft; de efficiëntie hangt af van de matrix A.
(Voor de praktijk maakt het ook nog verschil of men (9) voor een enkel rechterlid $\vec{b}$ wil oplossen of voor een groot aantal rechterleden. Bij een zeer groot aantal rechterleden zijn de eenmalige kosten nodig om bv. A^{-1} te bepalen verwaarloosbaar.)

Men past 'n iteratieve methode vaak toe, met name in de situaties waarin men over een heel goede benadering K^{-1} voor A^{-1} beschikt of als de matrix A ijl is (als A bijvoorbeeld een gediskretizeerde partiële differentiaaloperator representeert). In de praktijk zijn de volgende splitsingen populair. (Voor sommige splitsingen gebruikt men dan wel een efficiënte variant van het algorithme in (1.3.5); D, L en U zijn als in (1.3.2)).

K	naam	opmerkingen
$\frac{1}{\omega}$	Richardson-iteratie	*
D	Jacobi iteratie	besproken in (1.3.5)
$\frac{1}{\omega}D$	Jacobi overrelaxatie	*
$D+L$	Gauss–Seidel	even efficiënt als Jacobi, vaak sneller (zie (1.4.13))
$\frac{1}{\omega}D+L$	succesive overrelaxatie (SOR)	* Zie ook (1.3.11)
$(\frac{1}{\omega}D+L)(\frac{2-\omega}{\omega}D)^{-1}(\frac{1}{\omega}D+U)$	symmetrische SOR	* zie ook (1.3.12)

* ω is een zogenaamde relaxatie parameter en wordt gebruikt om 'konvergentie' te krijgen of om de konvergentiesnelheid te verhogen.

Of een proces konvergeert hang niet af van de gekozen norm. We geven twee varianten van bovenstaand verhaal. Voor de eerste variant geven we een alternatief bewijs.

(1.3.8) Bewering. *Zij* $\mathbf{X}$ *voorzien van een norm* $\|\cdot\|$.
Zij $\mathcal{A}\in\mathcal{L}(\mathbf{X})$. *Stel dat* $(\mathcal{K}_k)$ *een rij inverteerbare afbeeldingen is in* $\mathcal{L}(\mathbf{X})$ *zodat met* $\mathcal{R}_k:=\mathcal{K}_k-\mathcal{A}$ *geldt* $\rho:=\sup_k\|\mathcal{K}_k^{-1}\mathcal{R}_k\|<1$.
Voor $b\in\mathbf{X}$ *is* $x\in\mathbf{X}$ *zodat* $\mathcal{A}x=b$.
Kies $x_0\in\mathbf{X}$. *Laat* (x_k) *zo zijn dat* $\mathcal{K}_kx_{k+1}=\mathcal{R}_kx_k+b\quad(k\in\mathbf{N}_0)$.
Dan $\|x_k-x\|\leq\rho^k\|x_0-x\|\quad(k\in\mathbf{N}_0)$.
(Merk op dat met $r_k=b-Ax_k$ ook geldt $\mathcal{K}_k(x_{k+1}-x_k)=r_k$.)

Bewijs. Schrijf $f_k:=x_k-x$. Merk op dat $\mathcal{K}_kx=\mathcal{R}_kx+b$. Het verschil met $\mathcal{K}_kx_{k+1}=\mathcal{R}_kx_k+b$ levert $\mathcal{K}f_{k+1}=\mathcal{R}_kf_k$. Dus $\|f_{k+1}\|\leq\|\mathcal{K}_{k+1}^{-1}\mathcal{R}_{k+1}f_k\|\leq\rho\|f_k\|\leq\rho^{k+1}\|f_0\|$. □

Iteratieve methoden als aangetipt in (1.3.8) staan in de literatuur onder diverse namen bekend. De naamgeving hangt af van de historische kontekst of van de generalisatie voor niet-lineaire problemen die men in gedachten heeft. Men zou kunnen spreken over *succesive substitutie, naverfijning,* of *'defect correction'* (het residu kan men ook zien als een defect), etc..

(1.3.9) Bewering. *Laat* $\|\cdot\|$ *een norm zijn op* $\mathcal{L}(\mathbf{X})$. *Zij* $\mathcal{B}\in\mathcal{L}(\mathbf{X})$.
Als $\limsup_k\sqrt[k]{\|\mathcal{B}^k\|}<1$ *dan is* $\mathcal{I}-\mathcal{B}$ *inverteerbaar in* $\mathcal{L}(\mathbf{X})$ *en* $(\mathcal{I}-\mathcal{B})^{-1}=\sum_{j=0}^{\infty}\mathcal{B}^j$.

Bewijs. Kies $\rho\in(\limsup_k\sqrt[k]{\|\mathcal{B}^k\|},1)$. Merk op dat er een $k_0\in\mathbf{N}$ is waarvoor $\|\mathcal{B}^k\|\leq\rho^k$ voor iedere $k\geq k_0$. Ga verder te werk als in het bewijs van (1.3.1). □

Bewering (1.3.9) is, gegeven een multiplikatieve norm, op veel meer afbeeldingen $\mathcal{B}$ van toepassing dan (1.3.1). We geven een voorbeeld.

(1.3.10) Voorbeeld. Zij B de 2×2 matrix gegeven door

$$B=\begin{bmatrix}\frac{1}{2} & 2\\ 0 & \frac{1}{2}\end{bmatrix}.\quad\text{Dus}\quad B^k=\begin{bmatrix}(\frac{1}{2})^k & 4k(\frac{1}{2})^k\\ 0 & (\frac{1}{2})^k\end{bmatrix}\qquad(k\in\mathbf{N}).$$

We hebben nu dat $\|B\|_\infty=2\frac{1}{2}>1$ en $\limsup_k\sqrt[k]{\|B^k\|_\infty}=\lim_k\frac{1}{2}\sqrt[k]{1+4k}=\frac{1}{2}<1$.

(1.3.11) Opgave. Zij $A=D+L+U$ als in (1.3.2) met niet-singuliere D.
Bewijs dat $L^n=0$, en dat $D+L$ en $D+U$ niet-singulier zijn.
Laat zien dat de matrices $(D+L)^{-1}\succeq0$ en $-(D+L)^{-1}U\succeq0$ als $D\succeq0$ en $0\succeq U+L$.

(1.3.12) Opgave. Zij $A = D + L + U$ als in (1.3.2) met D niet-singulier.
Symmetrische Gauss–Seidel (of SSOR met $\omega = 1$) produceert voor $\vec{b} \in \mathbf{K}^n$ rijen $(\vec{x}_k)$ en $(\vec{x}_{k+\frac{1}{2}})$ als volgt:

$$\left\{ \begin{array}{lcl} (D+U)\vec{x}_{k+\frac{1}{2}} & = & -L\vec{x}_k + \vec{b} \\ (L+U)\vec{x}_{k+1} & = & -U\vec{x}_{k+\frac{1}{2}} + \vec{b} \end{array} \right. \qquad \text{voor alle} \quad k \in \mathbf{N}_0.$$

Met $K = (D + L)D^{-1}(D + U)$ en $R = UD^{-1}L$ is (theoretisch) $K\vec{x}_{k+1} = R\vec{x}_k + \vec{b}$ en dus $A = K - R$. Bewijs dit.
Als A *symmetrisch* is (d.w.z. $A^T = A$), dan zijn K en R dat ook. Ga dit na.

In de volgende paragraaf besteden we wat meer aandacht aan de grootheid $\limsup_k \sqrt[k]{\|\mathcal{B}^k\|}$. Maar we geven eerst nog een andere toepassing van (1.3.1).

De inverse van een geperturbeerde matrix

(1.3.13) Bewering. *Zij $\|\cdot\|$ een multiplikatieve norm op $\mathbf{M}_n$.*
Zij A een niet-singuliere matrix in $\mathbf{M}_n$. Beschouw voor 'n $C \in \mathbf{M}_n$ de matrix $A + C$.
Als $\delta := \|A^{-1}\|\,\|C\| < 1$ *dan is $A + C$ niet-singulier,*

$$\|(A+C)^{-1}\| \le \|A^{-1}\|\frac{1}{1-\delta} \qquad en \qquad \|(A+C)^{-1} - A^{-1}\| \le \|A^{-1}\|\frac{\delta}{1-\delta}.$$

Bewijs. $\|(A+C)^{-1}\| = \|(A(I + A^{-1}C))^{-1}\| = \|(I + A^{-1}C)^{-1}A^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}C\|}$
en $\|(A+C)^{-1} - A^{-1}\| = \|(A+C)^{-1}(A - (A+C))A^{-1}\| \le \|(A+C)^{-1}\|\,\|C\|\,\|A^{-1}\|$. □

1.4 De spektrale straal

(1.4.1) Definities. Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ met matrix A ten opzichte van een basis van $\mathbf{X}$.
$v \in \mathbf{X}$ is een *eigenvektor* van $\mathcal{A}$ als $v \neq 0$ en $\mathcal{A}v = \lambda v$ voor zekere $\lambda \in \mathbf{K}$: λ is een *eigenwaarde* van $\mathcal{A}$ bij de eigenvektor v.

Merk op dat $\det(A - \zeta I) = \det(T(A - \zeta I)T^{-1}) = \det(TAT^{-1} - \zeta I)$ voor iedere niet-singuliere matrix $T \in \mathbf{M}_n(\mathbf{C})$ en iedere $\zeta \in \mathbf{C}$. De waarde $\det(A - \zeta I)$ is blijkbaar onafhankelijk van de gekozen basis in $\mathbf{X}$ (in origineelruimte en in beeldruimte dezelfde basis!). We definiëren $\Psi_{\mathcal{A}} : \mathbf{C} \to \mathbf{C}$ het *karakteristieke polynoom* van $\mathcal{A}$ door $\Psi_{\mathcal{A}}(\zeta) := \det(A - \zeta I)$ $(\zeta \in \mathbf{C})$.
$\sigma(\mathcal{A}) := \{\lambda \in \mathbf{C} \,\big|\, \Psi_{\mathcal{A}}(\lambda) = 0\}$ is het *spektrum van $\mathcal{A}$* ($\sigma(\mathcal{A}) \subset \mathbf{C}$, ook als $\mathbf{K} = \mathbf{R}$!).
$\Psi_{\mathcal{A}}$ is een polynoom van graad n en heeft dus n nulpunten (eventueel samenvallende) in $\mathbf{C}$. De nulpunten van $\Psi_{\mathcal{A}}$ noteren we ook wel met $\lambda_1(A), \ldots, \lambda_n(A)$ (m-voudige nulpunten tellen we m maal: dus

$$\Psi_{\mathcal{A}}(\zeta) = (\lambda_1(A) - \zeta)(\lambda_2(A) - \zeta)(\cdots)(\lambda_n(A) - \zeta) \quad (\zeta \in \mathbf{C})\,).$$

De *spektrale straal* $\rho(\mathcal{A})$ is gedefinieerd door $\rho(\mathcal{A}) := \max\{|\lambda| \,\big|\, \lambda \in \sigma(\mathcal{A})\}$.

(1.4.2) Bewering. *Zij $\mathcal{A}, \mathcal{B} \in \mathcal{L}(\mathbf{X})$.*
Dan $\Psi_{\mathcal{AB}} = \Psi_{\mathcal{BA}}$ *en dus* $\sigma(\mathcal{AB}) = \sigma(\mathcal{BA})$ *en* $\rho(\mathcal{AB}) = \rho(\mathcal{BA})$.

Bewijs. Voor het voledige bewijs verwijzen we naar een wat algemenere variant van de bewering in een volgend hoofdstuk. Hier merken we op dat
$\det(AB - \zeta I) = \det((A - \zeta B^{-1})B) = \det(BA - \zeta I)$ als B niet-singulier is. □

(1.4.3) Bewering. *Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$. Stel* $\sigma(\mathcal{A}) \cap \mathbf{K} \neq \emptyset$. *Zij* $\lambda \in \sigma(\mathcal{A}) \cap \mathbf{K}$.
(a) *Dan λ is een eigenwaarde van $\mathcal{A}$: $\mathcal{A}v = \lambda v$ voor zekere* $v \in \mathbf{X}\backslash\{0\}$.
(b) *Zij* $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ *zodat* $\mathcal{AB} = \mathcal{BA}$ *en* $\sigma(\mathcal{B}) \subset \mathbf{K}$.
Dan is er een $v \in \mathbf{X}\backslash\{0\}$ *en een* $\mu \in \mathbf{K}$ *zodat* $\mathcal{A}v = \lambda v$ *én* $\mathcal{B}v = \mu v$.

Bewijs. Zij A de matrix van $\mathcal{A}$ en B van $\mathcal{B}$ ten opzichte van een basis $q_1, \ldots, q_n$ van $\mathbf{X}$.
(a) Merk op dat $\mathrm{Ker}(A - \zeta I) = \{0\} \Leftrightarrow \det(A - \zeta I) \neq 0$ $(\zeta \in \mathbf{C})$. Dus is er een $\vec{v} \in \mathbf{C}^n\backslash\{0\}$ zodat $A\vec{v} = \lambda\vec{v}$. Als $\vec{v} = (\nu_1, \ldots, \nu_n)^T$ dan $\mathcal{A}v = \lambda v$ met $v = \sum \nu_i q_i$. Dit bewijst (a) als $\mathbf{K} = \mathbf{C}$.
Stel $\mathbf{K} = \mathbf{R}$. Dan $A \in \mathbf{M}_n(\mathbf{R})$, $\lambda \in \mathbf{R}$ en dus $0 = (\bar{A} - \bar{\lambda}I)\bar{\vec{v}} = (A - \lambda I)\bar{\vec{v}}$. Omdat $\vec{v} + \bar{\vec{v}}$ of

$i(\vec{v}-\bar{\vec{v}})$ tot $\mathbf{R}^n\backslash\{0\}$ behoort heeft $\mathcal{A}$ ook in dit reëel geval een eigenvektor in $\mathbf{X}$.
(b) Beschouw de *eigenruimte* $\mathbf{V}_\lambda := \{v \in \mathbf{X} \,\big|\, \mathcal{A}v = \lambda v\}$ van $\mathcal{A}$ bij eigenwaarde λ. Dan $\mathcal{B}(\mathbf{V}_\lambda) \subseteq \mathbf{V}_\lambda$, immers: $\mathcal{A}\mathcal{B}v = \mathcal{B}(\mathcal{A}v) = \mathcal{B}(\lambda v) = \lambda\mathcal{B}v \quad (v \in \mathbf{V}_\lambda)$. Dus is $\mathcal{B}$, of formeler de restriktie $\mathcal{B}_{|\mathbf{V}_\lambda}$ van $\mathcal{B}$ tot $\mathbf{V}_\lambda$, een afbeelding in $\mathcal{L}(\mathbf{V}_\lambda)$ en $\sigma(\mathcal{B}_{|\mathbf{V}_\lambda}) \subseteq \sigma(\mathcal{B})$. De bewering volgt nu uit het feit dat het spectrum ook van afbeeldingen in $\mathcal{L}(\mathbf{V}_\lambda)$ niet leeg is: $\mathcal{B}_{|\mathbf{V}_\lambda}v = \mathcal{B}v = \mu v$ voor zekere $v \in \mathbf{V}_\lambda$ en $\mu \in \mathbf{K}$. □

(1.4.4) Opgave. Zij $A \in \mathbf{M}_n$ *diagonalizeerbaar*, d.w.z.: er is een niet-singuliere $T \in \mathbf{M}_n$ en een diagonaal $D = (d_{ij}) \in \mathbf{M}_n$ zodat $A = TDT^{-1}$. Ga na dat $\sigma(A) = \sigma(D) = \{d_{11}, \ldots, d_{nn}\}$ en zie in dat de kolommen Te_i van T eigenvektoren van A zijn (Te_i bij eigenwaarde d_{ii}).

De toevoeging $\mathcal{A} \rightsquigarrow \rho(\mathcal{A})$ definieert *geen* norm op $\mathcal{L}(\mathbf{X})$.

(1.4.5) Voorbeeld. Beschouw de 2×2 matrix $A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.
Dan $\Psi_{\mathcal{A}}(\zeta) = \zeta^2 \quad (\zeta \in \mathbf{C})$. Dus $\sigma(A) = \{0\}$ en $\rho(A) = 0$, terwijl $A \neq 0$.

(1.4.6) Opgave. Als $D = (d_{ij}) \in \mathbf{M}_n$ een diagonaal matrix is dan
$\rho(D) = \max_i |d_{ii}| = \|D\|_\infty = \|D\|_p$ alle $p \geq 1$. Bewijs dit.

(1.4.7) Opgave. Zij $A = (a_{ij}) \in \mathbf{M}_n$. Zij $\text{spoor}(A) = \sum_{i=1}^n a_{ii}$ het *spoor van* A.
Schrijf $\Psi_{\mathcal{A}}(\zeta) = (-\zeta)^n + \psi_{n-1}(-\zeta)^{n-1} + \cdots + \psi_0 \quad (\zeta \in \mathbf{C})$. Dan $\text{spoor}(A) = \psi_{n-1}$ en $\psi_0 = \lambda_1(A)\lambda_2(A)\cdots\lambda_n(A) = \det(A)$. Ga dit na. Konkludeer dat $\text{spoor}(AB) = \text{spoor}(BA)$ voor iedere $B \in \mathbf{M}_n$. In het bijzonder is $\text{spoor}(TAT^{-1}) = \text{spoor}(A)$ voor iedere niet-singuliere $T \in \mathbf{M}_n$.

(1.4.8) Stelling. *Zij $\|\cdot\|$ een norm op $\mathcal{L}(\mathbf{X})$. Dan*

$$\rho(\mathcal{A}) = \limsup_{k\to\infty} \sqrt[k]{\|\mathcal{A}^k\|} = \lim_{k\to\infty} \sqrt[k]{\|\mathcal{A}^k\|} \qquad (\mathcal{A} \in \mathcal{L}(\mathbf{X})\,).$$

Als de norm geassocieerd is dan geldt $\quad \rho(\mathcal{A}) \leq \sqrt[k]{\|\mathcal{A}^k\|} \quad$ *voor alle* $k \in \mathbf{N}$.

• *Bewijs.* Neem eerst even aan dat $\|\cdot\|$ een geassocieerde norm is.
Als v een eigenvektor is van $\mathcal{A}$ bij de eigenwaarde λ dan is $\mathcal{A}v = \lambda v$. Dus $\mathcal{A}^k v = \lambda^k v$ en $|\lambda|^k\|v\| = \|\mathcal{A}^k v\| \leq \|\mathcal{A}^k\|\,\|v\|$. Blijkbaar $|\lambda| \leq \sqrt[k]{\|\mathcal{A}^k\|}$ en dus $\rho(\mathcal{A}) \leq \sqrt[k]{\|\mathcal{A}^k\|}$.

We bewijzen nu de eerste bewering. Laat A de matrix zijn van $\mathcal{A}$ t.o.v een basis $q_1, \ldots, q_n$ van $\mathbf{X}$.
Er is een niet-singuliere matrix $T \in \mathbf{M}_n(\mathbf{C})$ zodat de matrix TAT^{-1} de *Jordan normaalvorm* heeft: $TAT^{-1} = D + N$, waarbij D een diagonaal matrix is, N een nilpotente ($N^n = 0$) en D en N kommuteren ($DN = ND$). Wat preciezer

$$D = \begin{bmatrix} \lambda_1 & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \lambda_n \end{bmatrix}, \quad N = \begin{bmatrix} 0 & \nu_1 & 0 & \ldots & 0 \\ 0 & 0 & \nu_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & \nu_{n-1} \\ 0 & 0 & \ldots & 0 & 0 \end{bmatrix},$$

met $\nu_i \in \{0, 1\}$ en $\lambda_i \in \mathbf{C}$ zodat $\lambda_i \neq \lambda_j$ als $\lambda_i \neq \lambda_{i+1}$ en $j > i$ (gelijke eigenwaarden λ_i staan onder elkaar) en $\nu_i = 0$ als $\lambda_i \neq \lambda_{i+1}$ (alle i, j).

Omdat $\Psi_A = \Psi_{D+N} = \Psi_D$ is $\rho(D) = \rho(D+N) = \rho(\mathcal{A})$.

Omdat $DN = ND$ en $N^n = 0$ is

$$(D+N)^k = D^k + kND^{k-1} + \binom{k}{2}N^2D^{k-2} + \cdots + \binom{k}{n-1}D^{k-n+1}N^{n-1}.$$

$\|\cdot\|$ induceert een norm $\|\cdot\|_*$ op $\mathbf{M}_n$ door $\|TBT^{-1}\|_* = \|\mathcal{B}\|$ voor $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ waarbij B de matrix is van $\mathcal{B}$ t.o.v. $q_1, \ldots, q_n$. Omdat op iedere eindig dimensionale ruimten normen equivalent zijn, zijn er $\gamma, \Gamma \in (0, \infty)$, $\gamma < \Gamma$ zodat $\gamma\|B\|_\infty \leq \|B\|_* \leq \Gamma\|B\|_\infty$ voor alle $B \in \mathbf{M}_n$.

Schrijf $\rho = \rho(D)$. Merk op dat $\rho = \|D\|_\infty$ en $\|N\|_\infty \le 1$. We zien dus dat (gebruik ook de tweede bewering in de stelling en het feit dat $\|\cdot\|_\infty$ geassocieerd is)

$$\gamma\rho(\mathcal{A})^k = \gamma\rho(D+N)^k \le \gamma\|(D+N)^k\|_\infty = \gamma\|(TAT^{-1})^k\|_\infty = \gamma\|TA^kT^{-1}\|_\infty \le \|\mathcal{A}^k\| \le$$

$$\le \Gamma\|(D+N)^k\|_\infty = \Gamma\|D^k + kND^{k-1} + \cdots + \tbinom{k}{n-1}D^{k-n+1}N^{n-1}\|_\infty \le$$

$$\le \Gamma(\rho^k + k\rho^{k-1} + \cdots + \tbinom{k}{n-1}\rho^{k-n+1}) = \Gamma\rho^{k-n}(\rho^n + \cdots + \tbinom{k}{n-1}\rho).$$

Dit bewijst de stelling voor het geval $\rho = 0$. Als $\rho \ne 0$ dan volgt de stelling uit het feit dat $\sqrt[k]{\gamma} \to 1$, $\sqrt[k]{\Gamma} \to 1$, $\sqrt[k]{\rho^{-n}} \to 1$, $\sqrt[k]{\rho^n + \cdots + \binom{k}{n-1}\rho} \to 0$ als $k \to \infty$. □

(1.4.9) Gevolg. *Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$.
$\rho(\mathcal{A}) < 1 \Leftrightarrow \lim_{k\to\infty} \mathcal{A}^k = 0$ *en* $\rho(\mathcal{A}) < 1 \Rightarrow (\mathcal{I} - \mathcal{A})^{-1}$ *bestaat.* □

(1.4.10) Opgave. Bewijs het bovenstaand gevolg (zie ook (1.3.9)).

(1.4.11) Opgave. Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ een inverteerbare afbeelding. Beschouw een splitsing $\mathcal{A} = \mathcal{K} - \mathcal{R}$ waarbij $\mathcal{K}$ inverteerbaar is. Beschouw het iteratieve proces (a) $\mathcal{K}x_{k+1} = \mathcal{R}x_k + b$.

Bewijs dat $\rho(\mathcal{K}^{-1}\mathcal{R}) < 1$ dan en slechts dan als het iteratieve proces (a) *konvergeert*, d.w.z. als voor iedere $b \in \mathbf{X}$ de rij (x_k) die voldoet aan (a) naar de oplossing x van $\mathcal{A}x = b$ konvergeert.

(1.4.12) Opgave. Als $A, B \in \mathbf{M}_n$ zo zijn dat $B \succeq A \succeq 0$ dan $\rho(B) \ge \rho(A)$. Bewijs dit.

(1.4.13) Opgave. De *stelling van Perron–Frobenius* zegt dat
$\rho(A) \in \sigma(A)$ voor alle $A \in \mathbf{M}_n$ waarvoor $A \succeq 0$.
U hoeft deze stelling niet te bewijzen, maar u mag hem wel toepassen.

Zij $A = D + L + U \in \mathbf{M}_n$ als in (1.3.2) en (1.3.11), $a_{ii} > 0$ en $a_{ij} \le 0$ voor alle i, j. Als $r := \rho(-(D+L)^{-1}U) \ge 1$ dan $1 \in \sigma(D^{-1}(-\frac{1}{r}U - L))$ en dus $\rho(D^{-1}(-U-L)) \ge 1$. Bewijs dit. Blijkbaar $\rho(-(D+L)^{-1}U) < 1$ als $\rho(D^{-1}(-U-L) < 1$.

Toon aan dat voor deze A het Gauss–Seidel proces *konvergeert* als het Jacobi proces konvergeert. Bepaal van het Gauss–Seidel – én van het Jacobi proces het aantal flop dat nodig is per iteratiestap. Aan welk proces geeft u de voorkeur?

1.5 Inprodukten

In het bewijs van (1.4.8) hebben we in $\mathbf{X}$ de basis genomen die ons geleverd wordt door de representatie van een $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ in Jordan normaalvorm. Zoals we zagen is het voor theoretische doeleinden weleens nuttig op zo'n basis over te stappen. In de praktijk veroorzaakt het gebruik van zo'n basis nogal eens grote evaluatiefouten. De fouten zijn dan groot omdat de basis erg 'scheefhoekig' is. We geven een voorbeeld en we 'vertellen' wat we onder scheve en rechthoekige (orthogonale) bases verstaan.

(1.5.1) Voorbeeld. Zij $\mathbf{X} = \mathbf{K}^2$. Zij $\varepsilon > 0$. Beschouw de standaard basis e_1, e_2 van $\mathbf{X}$ en de basis f_1, f_2 met $f_1 := e_1 + (1-\varepsilon)e_2$, $f_2 := e_1 + (1+\varepsilon)e_2$.

Beschouw $x := 2e_2 \in \mathbf{X}$. $(\alpha_1, \alpha_2)^T := (0, 2)^T$ representeert x ten opzichte van e_1, e_2 We schrijven x t.o.v. f_1, f_2 met representatie $(\beta_1, \beta_2)^T = (-\frac{1}{\varepsilon}, \frac{1}{\varepsilon})^T$: $x = \beta_1 f_1 + \beta_2 f_2$. Stel dat, door wat rekenwerk, de representatie $(\beta_1, \beta_2)^T$ belast raakt met een relatief klein foutje van de vorm $(\delta, \delta)^T$ (dus $\frac{\delta}{\max(|\beta_1|, |\beta_2|)} = \delta\varepsilon \ll 1$). Schrijven we tenslotte $x + d := (\beta_1 + \delta)f_1 + (\beta_2 + \delta)f_2$ weer op ten opzichte van e_1, e_2 dan $d = 2\delta(e_1 + e_2)$. Maar de relatieve fout $\frac{2\delta}{\max(|\alpha_1|, |\alpha_2|)} = \delta$ hoeft nu niet klein te zijn.

(1.5.2) Definitie. Zij $\mathbf{X}$ een lineaire ruimte..
Een afbeelding $(\cdot,\cdot) : \mathbf{X}\times\mathbf{X}\to\mathbf{K}$ is een *inprodukt* als het de volgende drie eigenschappen heeft.

(a) $(x,x)\in[0,\infty), \qquad (x,x)=0 \Leftrightarrow x=0 \qquad (x\in\mathbf{X})$.

(b) $(x,y)=\overline{(y,x)} \qquad (x,y\in\mathbf{X})$

(c) Voor iedere $y\in\mathbf{X}$ is $x\rightsquigarrow(x,y)$ een lineaire afbeelding van $\mathbf{X}$ naar $\mathbf{K}$.

Als $(x,y)=0$ voor $x,y\in\mathbf{X}$ dan zeggen we dat x *loodrecht* staat op y en we schrijven $x\perp y$.

(1.5.3) Voorbeeld. Op $\mathbf{K}^n$ definieert

$$(\vec{x},\vec{y}) := \sum_{j=1}^{n}\alpha_j\bar{\beta}_j \qquad (\vec{x}=(\alpha_1,\ldots,\alpha_n)^T, \vec{y}=(\beta_1,\ldots,\beta_n)^T\in\mathbf{K}^n)$$

een inprodukt, het zgn. *standaard inprodukt*.

In de volgende bewering verzamelen een paar bekende eigenschappen van inprodukten die we hier niet meer bewijzen. ((b) volgt gemakkelijk uit (a), (c) volgt door uitschrijven, (d) volgt met de Gram–Schmidt procedure).

(1.5.4) Bewering. *Laat $\mathbf{X}$ voorzien zijn van een inprodukt $(\cdot,\cdot)$. Dan:*

(a) $|(x,y)|\leq\sqrt{(x,x)}\sqrt{(y,y)} \qquad (x,y\in\mathbf{X})$ [De Cauchy–Schwarz ongelijkheid];

(b) $\|x\|:=\sqrt{(x,x)} \quad (x\in\mathbf{X})$ *definieert een norm op* $\mathbf{X}$;

(c) *Een stelsel vektoren $y_1,\ldots,y_l$ in $\mathbf{X}$ is* orthogonaal *als $y_i\perp y_j$ voor alle $i,j=1,\ldots,l$, $i\neq j$. Voor zo'n stelsel geldt* $\|\sum y_i\|^2=\sum\|y_i\|^2$ [De stelling van Pythagoras];

(d) *Als $\mathbf{X}$ n dimensionaal is, dan heeft $\mathbf{X}$ een orthonormale basis ($q_1,\ldots,q_l$ is een ortho*normaal *stelsel als het stelsel orthogonaal is en iedere q_i genormeerd is: $\|q_i\|=1$).* □

(1.5.5) Voorbeeld. De standaard basis $e_1,\ldots,e_n$ in $\mathbf{K}^n$ is orthonormaal ten opzichte van het standaard inprodukt.

(1.5.6) Opgave. Als $q_1,\ldots,q_n$ is een orthonormale basis is van $\mathbf{X}$ dan definieert
$(\vec{x},\vec{y}) := (\sum\alpha_i q_i, \sum\beta_i q_i) \qquad (\vec{x}=(\alpha_1,\ldots,\alpha_n)^T, \vec{y}=(\beta_1,\ldots,\beta_n)^T\in\mathbf{K}^n)$
het standaard inprodukt op $\mathbf{K}^n$. Ga dit na.

(1.5.7) Definities. Laat $\mathbf{X}$ voorzien zijn van een inprodukt $(\cdot,\cdot)$.
Zij $\mathcal{A}\in\mathcal{L}(\mathbf{X})$. $\mathcal{B}\in\mathcal{L}(\mathbf{X})$ is de *geadjungeerde van* $\mathcal{A}$ als $(\mathcal{A}x,y)=(x,\mathcal{B}y)$ voor iedere $x,y\in\mathbf{X}$ (ga na dat zo'n $\mathcal{B}$ bestaat in $\mathcal{L}(\mathbf{X})$ en dat er maar precies een is). We schrijven $\mathcal{A}^*$ in plaats van $\mathcal{B}$. Dus $(\mathcal{A}x,y)=(x,\mathcal{A}^*y) \quad (x,y\in\mathbf{X})$ definieert $\mathcal{A}^*$.

(1.5.8) Bewering. *Als $\mathcal{A},\mathcal{B}\in\mathcal{B}(\mathbf{X})$ en $\lambda\in\mathbf{K}$ dan $(\mathcal{A}\mathcal{B})^*=\mathcal{B}^*\mathcal{A}^*$, $(\lambda\mathcal{A})^*=\bar{\lambda}\mathcal{A}^*$, $\mathcal{A}^{**}=\mathcal{A}$.*
Als $\mathcal{A}$ ook nog inverteerbaar is dan is $\mathcal{A}^*$ inverteerbaar en $(\mathcal{A}^{-1})^*=(\mathcal{A}^*)^{-1}$. □

(1.5.9) Bewering. *Beschouw een orthonormale basis van $\mathbf{X}$. Zij $\mathcal{A}\in\mathcal{L}(\mathbf{X})$. Als A de matrix is van $\mathcal{A}$ en B de matrix $\mathcal{A}^*$ ten opzichte van deze orthonormale basis dan $B=A^H$.* □

(1.5.10) Definities. $\mathbf{X}$ is voorzien van een inprodukt $(\cdot,\cdot)$. Zij $\mathcal{A}\in\mathcal{L}(\mathbf{X})$.
$\mathcal{A}$ is *zelfgeadjungeerd* of *hermiets* als $\mathcal{A}^*=\mathcal{A}$.
$\mathcal{A}$ is *anti-hermiets* als $\mathcal{A}^*=-\mathcal{A}$.
$\mathcal{A}$ is *normaal* als $\mathcal{A}\mathcal{A}^*=\mathcal{A}^*\mathcal{A}$.
$\mathcal{Q}\in\mathcal{L}(\mathbf{X})$ heet *unitair* als $\mathcal{Q}^*\mathcal{Q}=\mathcal{Q}\mathcal{Q}^*=\mathcal{I}$.

(1.5.11) Bewering. *Laat* $q_1, \ldots, q_n$ *een orthonormale basis zijn in* $\mathbf{X}$.
Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ *met matrix* A *t.o.v.* $q_1, \ldots, q_n$.
Dan zijn $\mathcal{A}_h := \frac{1}{2}(\mathcal{A} + \mathcal{A}^*)$ *en* $\mathcal{A}_a := \frac{1}{2i}(\mathcal{A} - \mathcal{A}^*)$ *beiden hermiets en* $\mathcal{A} = \mathcal{A}_h + i\mathcal{A}_a$.
$\mathcal{A}$ *is normaal* $\Leftrightarrow$ $\mathcal{A}_h\mathcal{A}_a = \mathcal{A}_a\mathcal{A}_h$. $\mathcal{A}$ *is hermiets* $\Leftrightarrow$ $\mathcal{A}_a = 0$.
$\mathcal{A}$ *is hermiets* $\Rightarrow$ $\mathcal{A}$ *is normaal.*
Verder $\mathcal{A}^* = \mathcal{A} \Leftrightarrow A^H = A$ *en* $\mathcal{A}\mathcal{A}^* = \mathcal{A}^*\mathcal{A} \Leftrightarrow AA^H = A^HA$.
Zij $\mathcal{Q} \in \mathcal{L}(\mathbf{X})$ *met matrix* Q.
Dan: $\mathcal{Q}$ *is unitair* $\Leftrightarrow$ $\|\mathcal{Q}x\| = \|x\|$ *voor alle* $x \in \mathbf{X}$ $\Leftrightarrow$ $Q^HQ = I$ $\Leftrightarrow$
$\mathcal{Q}q_1, \ldots, \mathcal{Q}q_n$ *is een orthonormale basis voor* $\mathbf{X}$. □

Het is vaak van belang of een matrix $A \in \mathbf{M}_n$ diagonalizeerbaar is. In het algemeen is dat moeilijk aan een matrix af te zien. Voor hermietse matrices is het echter zeer gemakkelijk. Hoewel hermietse matrices nogal speciaal lijken komen ze in de praktijk nogal eens voor. We stellen het bewijs van de volgende stelling uit tot (1.5.16).

(1.5.12) Stelling. *Als* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ *hermiets is dan is er een orthonormale basis* $q_1, \ldots, q_n$ *in* $\mathbf{X}$ *van eigenvektoren van* $\mathcal{A}$ *en* $\sigma(\mathcal{A}) \subset \mathbf{R}$.

(1.5.13) Opgave. Als $A \in \mathbf{M}_n$ en $A^H = A$ dan $A = QDQ^H$ voor zekere *unitaire* matrix $Q \in \mathbf{M}_n$ $(Q^HQ = I)$ en zekere diagonaal matrix $D \in \mathbf{M}_n$. Ga dit na.

De Jordan normaalvorm doet in de theorie goed werk, maar kan in de praktijk om stabiliteits redenen beter niet gebruikt worden. De representatie in de Schur normaalvorm is 'stabiel' en in theoretische argumenten kan de Jordan normaalvorm vrijwel altijd vervangen worden door de Schur normaalvorm.

Twee kommuterende afbeeldingen kunnen simultaan op Schur normaalvorm gebracht worden (het bewijs van die bewering, (b) van de volgende stelling, wordt in (1.5.15) als een opgave aan de lezer overgelaten).

(1.5.14) Stelling. *Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ *zodat* $\sigma(\mathcal{A}) \subset \mathbf{K}$ (dus $\mathbf{K} = \mathbf{C}$ of $\sigma(\mathcal{A}) \subset \mathbf{R}$).
(a) *Dan kan* $\mathcal{A}$ *gerepresenteerd worden in de* Schur normaalvorm*: er is een orthonormale basis* $q_1, \ldots, q_n$ *van* $\mathbf{X}$ *ten opzichte waarvan de matrix* $U = (u_{ij}) \in \mathbf{M}_n(\mathbf{K})$ *van* $\mathcal{A}$ *boven driehoeks is* ($u_{ij} = 0$ als $i > j$).
(b) *Zij* $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ *zodat* $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$ *en ook* $\sigma(\mathcal{B}) \subset \mathbf{K}$.
Dan is er een orthonormale basis $q_1, \ldots, q_n$ *van* $\mathbf{X}$ *zodat de matrices* U *van* $\mathcal{A}$ *en* V *van* $\mathcal{B}$ *beiden boven driehoeks zijn.*

Bewijs. (a) Kies een orthonormale basis van $\mathbf{X}$ en laat A de bij $\mathcal{A}$ behorende matrix zijn.

Omdat $\sigma(A) \neq \emptyset$ is er een $\lambda_1 \in \mathbf{K}$ en een $\vec{q}_1 \in \mathbf{K}^n$ waarvoor $A\vec{q}_1 = \lambda_1\vec{q}_1$ en $\|\vec{q}_1\|_2 = 1$ (zie (1.4.3)). Vul $\vec{q}_1$ aan tot een orthonormale basis $\vec{q}_1, \vec{p}_2, \ldots, \vec{p}_n$ voor $\mathbf{K}^n$. Ten opzichte van deze basis heeft de afbeelding $A \in \mathcal{L}(\mathbf{K}^n)$ de volgende vorm.

$$\left[\begin{array}{c|ccc} \lambda_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{array}\right] \quad \text{voor een zekere} \quad A_2 \in \mathbf{M}_{n-1}. \quad \text{(Ga dit na.)}$$

Merk op dat $\Psi_A(\zeta) = (\lambda_1 - \zeta)\Psi_{A_2}(\zeta)$. Dus $\sigma(A_2) \subseteq \sigma(A) \subseteq \mathbf{K}$. Door bovenstaand argument herhaald toe toe passen op A_2, A_3, etc. volgt de stelling (als $(\beta_2, \ldots, \beta_n)^T \in \mathbf{K}^{n-1}$ de gekozen genormeerde eigenvektor is van A_2 beschouw dan
$\vec{q}_2 := \beta_2\vec{p}_2 + \cdots + \beta_n\vec{p}_n \in \mathbf{K}^n$, etc.). □

(1.5.15) Opgave. Bewijs (b) van bovenstaande stelling. (Hint: ga te werk als in het bewijs van (a) van deze stelling en gebruik (b) van (1.4.3).)

(1.5.16) Bewijs (1.5.12). Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ hermiets. Neem eerst aan dat $\mathbf{K} = \mathbf{C}$.

Laat $q_1, \ldots, q_n$ en U als in de vorige stelling zijn. Omdat $U^H = U$ (zie (1.5.11)) en U boven driehoeks is volgt dat U een reële diagonaal matrix is en $\sigma(\mathcal{A}) = \sigma(U) \subset \mathbf{R}$.

Als $\mathbf{K} = \mathbf{R}$ dan volgt door over $\mathbf{C}$ te werken (identificeer $\mathcal{A}$ met een reële matrix, vat die op als element van $\mathbf{M}_n(\mathbf{C})$) dat $\sigma(\mathcal{A}) = \sigma(U) \subset \mathbf{R}$ en pas weer de stelling toe. □

(1.5.17) Stelling. (zie ook (1.5.12)) *Als $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ normaal is en $\mathbf{K} = \mathbf{C}$ dan is er een orthonormale basis $q_1, \ldots, q_n$ in $\mathbf{X}$ van eigenvektoren van $\mathcal{A}$.*

Bewijs. Als $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ normaal is dan kommuteren de hermietse matrices $\mathcal{A}_h$ en $\mathcal{A}_a$ (zie (1.5.11)) en de stelling volgt uit (b) van (1.5.14) met een argument als in (1.5.16). □

In het algemeen is het moeilijk om de 2-norm van een matrix te berekenen. De relaties die we hieronder geven zijn in zo'n situatie vaak nuttig; we laten het bewijs aan de lezer over.

(1.5.18) Bewering. *Zij $x \in \mathbf{X}$. Dan*

$$\|x\|_2 := \sqrt{(x,x)} = \sup\{\tfrac{|(x,y)|}{\|y\|_2} \,\big|\, y \in \mathbf{X}, y \neq 0\} = \sup\{|(x,y)| \,\big|\, y \in \mathbf{X}, \|y\|_2 \leq 1\}.$$

Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$. Dan

$$\|\mathcal{A}\|_2 := \sup\{\|\mathcal{A}x\|_2 \,\big|\, x \in \mathbf{X}, \|x\|_2 \leq 1\} = \sup\{|(\mathcal{A}x, y)| \,\big|\, \|x\|_2 \leq 1, \|y\|_2 \leq 1\} = \|\mathcal{A}^*\|_2,$$

$$\|\mathcal{A}\|_2 = \sqrt{\|\mathcal{A}^*\mathcal{A}\|_2} = \sqrt{\|\mathcal{A}\mathcal{A}^*\|_2}.$$

Als A de matrix is van $\mathcal{A}$ ten opzichte van een orthonormale basis van $\mathbf{X}$ dan $\|\mathcal{A}\|_2 = \|A\|_2$ (waarbij de laatste norm de standaard 2-norm is). □

(1.5.19) Opgave. Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$. Bewijs dat $\|\mathcal{A}\|_\infty = \|\mathcal{A}^*\|_1$ en $\|\mathcal{A}\|_2 \leq \sqrt{\|\mathcal{A}\|_\infty \|\mathcal{A}\|_1}$.

(1.5.20) Bewering. *Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$.*
Als $\mathcal{A}$ normaal is dan $\rho(\mathcal{A}) = \|\mathcal{A}\|_2$. In het algemeen is $\|\mathcal{A}\|_2 = \sqrt{\rho(\mathcal{A}^*\mathcal{A})}$.

(In (1.4.5) is een $A \in \mathbf{M}_2$ met $\rho(A) = 0$ en $\|A\|_2 = 1$!)

Bewijs. Als $\mathcal{A}$ normaal is dan is er een orthonormale basis $q_1, \ldots, q_n$ in $\mathbf{X}$ van eigenvektoren van $\mathcal{A}$ en de matrix D van $\mathcal{A}$ ten opzichte van deze basis is dan diagonaal.
Dus $\|\mathcal{A}\|_2 = \|D\|_2 = \rho(D) = \rho(\mathcal{A})$.

De tweede bewering volgt uit het feit dat $\mathcal{A}^*\mathcal{A}$ hermiets is. □

(1.5.21) Opgave. Zij $A \in \mathbf{M}_n$. Bewijs dat $\|A\|_F = \sqrt{\mathrm{spoor}(A^H A)} = \sqrt{\sum_j \lambda_j(A^H A)}$.

(1.5.22) Opgave. Als $A \in \mathbf{M}_n$ dan geldt $\|A\|_2 \leq \|\,|A|\,\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2$. Ga dit na. Laat zien dat $\|\,|A|\,\|_2 \leq \sqrt{n}\,\|A\|_2$ scherp is (middels een voorbeeld voor $n = 2$).

We hebben representaties in Jordan normaalvorm en in Schur normaalvorm gezien van afbeeldingen in $\mathcal{L}(\mathbf{X})$ en van matrices in $\mathbf{M}_n$. In dit college komen we nog een paar andere representatie vormen tegen. Voor de volledigheid sommen wij die nu hier op voor matrices in $\mathbf{M}_n$ (van een paar representaties zullen we ook de $\mathbf{M}_{m\times n}$–versie tegenkomen).

(1.5.23) Stelling. *Zij $A \in \mathbf{M}_n(\mathbf{K})$. A kan op de volgende manieren geschreven worden.*
Jordan vorm *$A = T(D+N)T^{-1}$ met $T, D, N \in \mathbf{M}_n(\mathbf{C})$, T niet-singulier, D diagonaal en $DN = ND$, $N^n = 0$.*
Schur vorm *$A = QUQ^H$ met $Q, U \in \mathbf{M}_n(\mathbf{C})$, Q unitair en U boven driehoeks.*
Singuliere waarden dekompositie *$A = Q_1 D Q_2^H$ met $Q_1, Q_2, D \in \mathbf{M}_n(\mathbf{K})$, Q_1, Q_2 unitair en D diagonaal, $D \succeq 0$.* (de diagonaal elementen van D zijn de *singuliere waarden van A.*)
Gram–Schmidt of QU-decompositie *$A = QU$ met $Q, U \in \mathbf{M}_n(\mathbf{K})$, Q unitair en U boven driehoeks.*

Bewijs. In (1.5.14) hebben we gezien dat A in de Schur vorm gerepresenteerd kan worden. De Jordan vorm interesseert ons hier niet echt. Voor een bewijs verwijzen we de geïnteresseerde lezer dan ook naar een algebra boek of diktaat. De QU-dekompositie wordt later in dit college uitgebreid behandeld. We konstrueren dan zo'n dekompositie.

Omdat A^HA hermiets is zijn er een unitaire matrix Q_2 en een diagonaal matrix T zodat $A^HA = Q_2TQ_2^H$. Omdat $0 \leq \|A\vec{y}\|_2^2 = (Q_2TQ_2^H\vec{y}, \vec{y}) = (T\vec{x}, \vec{x}) \quad (\vec{y} = Q_2\vec{x} \in \mathbf{K}^n)$ volgt $\sigma(T) \subset [0, \infty)$ (zie ook de volgende bewering). Met $\vec{p}_j := AQ_2e_j$ is $\vec{p}_1, \dots, \vec{p}_n$ blijkbaar een orthogonaal stelsel (sommige $\vec{p}_j$ kunnen evenwel $\vec{0}$ zijn). Zij Q_1 unitair zodat $Q_1e_j = \frac{1}{\|\vec{p}_j\|_2}\vec{p}_j$ als $\vec{p}_j \neq \vec{0}$. Met $D = \mathrm{diag}(\|\vec{p}_1\|_2, \|\vec{p}_2\|_2, \dots, \|\vec{p}_n\|_2)$ is $Q_1D = AQ_2$ en hebben we de singuliere waarden dekompositie. □

In de numerieke wiskunde zijn de tweede en derde representatie populair; de derde representatie speelt met name een rol in de theorie over 'kleinste kwadraten problemen'. In veel gevallen kan de Schurm vorm stabiel bepaald worden en levert ons impliciet de eigenwaarden van A. De QU-dekompositie levert een stabiele direkte methoden om stelsels lineaire vergelijkingen en kleinste kwadraten problemen op te lossen. We komen op dit alles nog uitgebreid terug.

We besluiten deze paragraaf met wat eigenschappen van lineaire afbeeldingen van een speciaal type. Hoewel ze nogal speciaal lijken komen ze in de praktijk nogal eens voor.

(1.5.24) Bewering. *Zij* $\mathbf{X}$ *een ruimte over* $\mathbf{K} = \mathbf{C}$ *met orthonormale basis* $q_1, \dots, q_n$. *Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ *met bijbehorende matrix* A. *De volgende eigenschappen zijn equivalent.*

(a) $\mathcal{A}$ *is* positief definiet, *d.w.z.:* $(\mathcal{A}x, x) \geq 0$ *voor alle* $x \in \mathbf{X}$.

(b) $\mathcal{A}$ *is hermiets en* $\sigma(\mathcal{A}) \subset [0, \infty)$.

(c) $\mathcal{A} = \mathcal{B}^*\mathcal{B}$ *voor zekere* $\mathcal{B} \in \mathcal{L}(\mathbf{X})$.

(d) $\mathcal{A} = \mathcal{B}^*\mathcal{B}$ *voor zekere positief definiete* $\mathcal{B}$.

(e) $A = U^HU$ *voor zekere boven driehoeks matrix* $U \in \mathbf{M}_n(\mathbf{K})$.

Bewijs. Duidelijk is dat "(d) ⇒ (c) ⇒ (a)" en "(e) ⇒ (a)" (ga dit na).

Om "(a) ⇒ (b)" te bewijzen beschouwen we eerst een $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ waarvoor $(\mathcal{B}x, x) = 0$ voor alle $x \in \mathbf{X}$. Merk op dat $\sigma(\mathcal{B}) = \{0\}$ (vul eigenvektoren in voor x) en dus $\mathcal{B} = 0$ als $\mathcal{B}$ ook nog hermiets is.

Stel dat nu $\mathcal{A}$ positief definiet is (eigenschap (a) heeft). Dan (zie (1.5.11))

$$(\mathcal{A}_ax, x) = \tfrac{1}{2i}(\mathcal{A}x - \mathcal{A}^*x, x) = \tfrac{1}{2i}[(\mathcal{A}x, x) - (x, \mathcal{A}x)] = \tfrac{1}{2i}[(\mathcal{A}x, x) - \overline{(\mathcal{A}x, x)}] = 0.$$

Blijkbaar $\mathcal{A}_a = 0$ en $\mathcal{A} = \mathcal{A}_h$ is hermiets. Vullen we voor x eigenvektoren in dan zien we dat $\sigma(\mathcal{A}) \subset [0, \infty)$ en hebben we (b) bewezen.

Als $\mathcal{A}$ hermiets is dan is er een orthonormale basis $p_1, \dots, p_n$ van $\mathbf{X}$ van eigenvektoren van $\mathcal{A}$. Zij D de bijbehorende matrix van $\mathcal{A}$. Dan is $D = (d_{ij})$ diagonaal en als $\sigma(\mathcal{A}) \subset [0, \infty)$ dan $d_{ii} \geq 0$. Neem $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ zo dat $\mathcal{B}$ matrix $D^{\frac{1}{2}} := (\sqrt{d_{ij}})$ heeft t.o.v. $p_1, \dots, p_n$ en "(b) ⇒ (d)" volgt.

Tenslotte bewijzen we "(c) ⇒ (e)".

Stel dat $A = B^HB$ met B de matrix van $\mathcal{B}$ t.o.v. $q_1, \dots, q_n$. Het Gram–Schmidt proces leert ons dat $B = QU$ voor zekere unitaire $Q \in \mathbf{M}_n$ en zekere boven driehoeks matrix $U \in \mathbf{M}_n$. Blijkbaar $A = (QU)^H(QU) = U^HU$. □

(1.5.25) Opgave. Zij $\mathbf{K} = \mathbf{C}$.

Bewijs dat: $(\mathcal{A}x, x) > 0$ voor alle $x \in \mathbf{X}\backslash\{0\} \quad \Leftrightarrow \quad \mathcal{A}$ is hermiets en $\sigma(\mathcal{A}) \subset (0, \infty)$.

($\mathcal{A}$ is *strikt* positief definiet als $(\mathcal{A}x, x) > 0 \quad (x \in \mathbf{X}\backslash\{0\})$.)

(1.5.26) Opgave. Als we (a) in (1.5.24) vervangen door

(a) $\mathcal{A}$ is hermiets en $(\mathcal{A}x, x) \geq 0$ voor alle $x \in \mathbf{X}$

dan zijn de uitspraken (a) – (e) ook equivalent in geval $\mathbf{K} = \mathbf{R}$. Ga dat dat.

2 De lineaire vergelijking $Ax = b$

2.1 Inleiding

In dit hoofdstuk bespreken we direkte numerieke oplosmethoden voor (LV). We zullen A schrijven als $A = A_1A_2$ (of $A = A_1A_2A_3$) met A_i zodat de vergelijkingen "$A_1\vec{y} = \vec{b}$" en "$A_2\vec{x} = \vec{y}$ efficiënt oplosbaar zijn. We voeren wat notaties in en geven een paar voorbeelden van zulk soort vergelijkingen.

(2.1.1) Notaties. $\mathbf{L}_n(\mathbf{K}) := \{L = (l_{ij}) \in \mathbf{M}_n(\mathbf{K}) \,\big|\, l_{ij} = 0 \text{ als } j > i\}$ is de ruimte van de beneden driehoeks matrices. Verder bestaat $\mathbf{L}_n^0 := \{L \in \mathbf{L}_n \,\big|\, \mathrm{diag}(L) = 0\}$ uit alle strikt beneden driehoeks matrices en $\mathbf{L}_n^1 := \{I + M \,\big|\, M \in \mathbf{L}_n^0\}$ uit alle beneden driehoeks matrices met diagonaal elementen 1.

$\mathbf{U}_n := \{U = (u_{ij}) \in \mathbf{M}_n \,\big|\, u_{ij} = 0 \text{ als } j < i\}$ is de ruimte van alle boven driehoeks matrices.

$\mathbf{Q}_n := \{Q \in \mathbf{M}_n \,\big|\, Q^HQ = I\}$ is de ruimte van alle unitaire matrices.

(2.1.2) Voorbeeld. Als $L = (l_{ij}) \in \mathbf{L}_n$ niet-singulier is dan kunnen we voor $\vec{b} = (\beta_1, \ldots, \beta_n)^T$ in $\mathbf{K}^n$ de vektor $\vec{x} = (\alpha_1, \ldots, \alpha_n)^T$ in $\mathbf{K}^n$ efficiënt oplossen uit de vergelijking $L\vec{x} = \vec{b}$ door *voorwaartse eliminatie*:

$$\alpha_1 = \frac{1}{l_{11}}\beta_1, \alpha_2 = \frac{1}{l_{22}}(\beta_2 - l_{21}\alpha_1), \alpha_3 = \frac{1}{l_{33}}(\beta_n - \sum_{j=1}^{2} l_{3j}\alpha_j), \ldots, \alpha_n = \frac{1}{l_{nn}}(\beta_n - \sum_{j=1}^{n-1} l_{nj}\alpha_j).$$

Het berekenen van $\vec{x}$ kost op deze manier $\frac{1}{2}n(n+1)$ flop.
Als $\tilde{L} := L^{-1}$ dan is $\tilde{L}$ ook beneden driehoeks (ga dit na) en kost de matrix–vektor vermenigvuldiging $\tilde{L}\vec{b}$ ook $\frac{1}{2}n(n+1)$ flop; we kunnen ons hier dus de moeite van het expliciet bepalen van $\tilde{L}$ beter besparen. Voorwaartse eliminatie is hier de efficiënste oplosmethode.

Als $U \in \mathbf{U}_n$ niet-singulier is dan kunnen we analoog aan de voorwaartse eliminatie $\vec{x}$ efficiënt oplossen uit $U\vec{x} = \vec{b}$ door wat hier *terugwaartse substitutie* wordt genoemd (vul zelf de details in).

Als Q unitair is dan is $\vec{x} = Q^H\vec{b}$ de oplossing van $Q\vec{x} = \vec{b}$.

(2.1.3) Opgave. Zij $L = (l_{ij}) \in \mathbf{L}_n$ zodat $l_{ij} > 0$ als $j = i$, $l_{ij} < 0$ als $i = j+1$ en $l_{ij} = 0$ anders. Laat $\tilde{L} = L^{-1} = (\tilde{l}_{ij})$. Stel dat we de $\tilde{l}_{ij}$ expliciet tot onze beschikking hebben. We wensen de oplossing $\vec{x}$ van $L\vec{x} = \vec{b}$ te berekenen. Bepaal het aantal flop dat we nodig hebben om $\vec{x}$ te berekenen middels voorwaartse eliminatie em met behulp van $\tilde{L}$ door matrix–vektor vermenigvuldiging.

Bij de behandeling van de methoden schenken we met name aandacht aan de stabiliteit en de efficiëntie. Om de stabiliteit te kunnen analyseren hebben we nog wat meer wiskundige kennis en inzicht nodig dan we in hoofdstuk 1 opgedaan hebben. Om onze stabiliteitsopmerkingen te kunnen illustreren behandelen we echter eerst een bekende oplosmethode.

2.2 De LU-dekompositie

De dekompositie methode die we in deze paragraaf behandelen is al bekend van eerste jaars algebra colleges. We bekijken de methode hier nogmaals vanuit een numeriek perspektief.

Gauss eliminatie met pivotten

Het volgende 'getal' voorbeeldje geeft in essentie de methode.

(2.2.1) Voorbeeld. α_1, α_2 lossen we op uit het stelsel $\left\{ \begin{array}{ll} 8\alpha_1 + 2\alpha_2 & = 3 \\ 2\alpha_1 + \alpha_2 & = 7 \end{array} \right.$ door van de tweede rij $\frac{2}{8}\times$ de eerste af te trekken. De tweede vergelijking in $\left\{ \begin{array}{rl} 8\alpha_1 + 2\alpha_2 & = 3 \\ \frac{1}{2}\alpha_2 & = 6\frac{1}{4} \end{array} \right.$ levert $\alpha_2 = 12\frac{1}{2}$ en door substitutie van deze waarde in de eerste vinden we α_1.
Met $A = \begin{bmatrix} 8 & 2 \\ 2 & 1 \end{bmatrix}, \vec{x} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}$, en $\vec{b} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$, is $A\vec{x} = \vec{b}$ de matrix formulering van het stelsel. In feite hebben we A geschreven als $A = LU = \begin{bmatrix} 1 & 0 \\ \frac{2}{8} & 1 \end{bmatrix} \begin{bmatrix} 8 & 2 \\ 0 & \frac{1}{2} \end{bmatrix}$, vervolgens door voorwaartse eliminatie $\vec{y}$ opgelost uit $L\vec{y} = \vec{b}$ en daarna door terugwaartse substitutie $\vec{x}$ uit $U\vec{x} = \vec{y}$.

In het voorbeeldje hebben we de eliminatie methode van Gauss toegepast: uit een stelsel van n vergelijkingen met n onbekenden α_i elimineert deze methode α_1 uit de 2-de t/m de n-de vergelijking door geschikte skalaire veelvouden van de eerste vergelijking af te trekken van de volgenden. Vervolgens elimineert de methode α_2 uit de 3-de t/m n-de vergelijking van het nieuwe stelsel door geschikte veelvouden van de tweede vergelijking af te trekken van de volgenden, etc.. Uit het driehoekig stelsel dat aldus onstaat kunnen we tenslotte $\alpha_n, \alpha_{n-1}, \ldots$ berekenen door terugwaartse substitutie. Ook zonder modifikatie werkt dit eliminatie proces vaak. In matrix-taal zegt men dat het proces achtereenvolgens stukken van *kolommen veegt* . In de volgende alinea geven we details.

(2.2.2) De eliminatie methode van Gauss. Zij $A = (a_{ij}) \in \mathbf{M}_n$ niet-singulier. We proberen als boven A te schrijven als $A = LU$ middels de *eliminatie methode van Gauss.*
We berekenen rekursief

$$\begin{array}{ll} & A_1 := A, \\ M_1, & A_2 = (a_{ij}^{(2)}) := (I - M_1)A, \\ M_3, & A_3 = (a_{ij}^{(3)}) := (I - M_2)A_2, \\ \vdots & \end{array}$$

, als volgt in $n-1$ achtereenvolgende stappen:
(1) Stel $a_{11} \neq 0$.
Laat $m_{i1} := \frac{a_{i1}}{a_{11}}$. Dan vinden we met

$$M_1 = \left[\begin{array}{c|ccc} 0 & 0 & \cdots & 0 \\ \hline m_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ m_{n1} & 0 & \cdots & 0 \end{array}\right] \quad \text{dat} \quad A_2 := (I - M_1)A = \left[\begin{array}{c|ccc} a_{11} & a_{12} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{array}\right]$$

(dus $a_{i1}^{(2)} = 0$ voor $i = 2, \ldots, n$ en $a_{1i}^{(2)} = a_{1i}$ voor $i = 1, \ldots, n$).
(2) Stel dat $a_{22}^{(2)} \neq 0$.
Laat $m_{i2} := \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}$. Dan vinden we met

$$M_2 = \left[\begin{array}{c|c|ccc} 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & 0 & 0 & \cdots & 0 \\ \hline 0 & m_{32} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & m_{n2} & 0 & \ldots & 0 \end{array}\right] \quad \text{dat} \quad A_3 := (I - M_2)A_2 = \left[\begin{array}{c|c|ccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \hline 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ \hline 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{array}\right].$$

(3), ..., (n-1) Als al de $a_{kk}^{(k)}$ die we in de achtereenvolgende stappen tegenkomen $\neq 0$ zijn dan vinden we dat $U := A_n = (I - M_{n-1}) \cdots (I - M_2)(I - M_1)A$ een boven driehoeks matrix is.
Merk op dat $M_k M_j = 0$ als $k \leq j$. In het bijzonder zien we dat $(I - M_k)(I + M_k) = I$ en $(I + M_k)(I + M_j) = I + M_k + M_j$ als $k \leq j$. Blijkbaar $(I - M_k)^{-1} = I + M_k$ en

$L := I + M := ((I - M_{n-1})(\cdots)(I - M_1))^{-1} = (I + M_1)(\cdots)(I + M_{n-1}) = I + M_1 + \cdots + M_{n-1}$. M is strikt beneden driehoeks en $A = (I + M)U = LU$.

We vatten de wiskundige konsekwenties van bovenstaande resultaten samen in onderdeel (a) van de volgend stelling. We formuleren daar de voorwaarden "$a_{kk}^{(k)} \neq 0$" in termen van A.

(2.2.3) Definitie. $A \in \mathbf{M}_n$ heeft een *LU-dekompositie* als $A = LU$ voor zekere $L \in \mathbf{L}_n$ en $U \in \mathbf{U}_n$ (we eisen niet dat $L \in \mathbf{L}_n^1$).

(2.2.4) Stelling. *Zij $A \in \mathbf{M}_n$ niet-singulier.*
(a) *De volgende drie beweringen zijn equivalent.*
 (i) *A heeft een LU-dekompositie.*
 (ii) *De Gauss methode is uitvoerbaar* ($a_{kk}^{(k)} \neq 0$ voor alle k).
 (iii) $\det((a_{ij})_{i \leq k, j \leq k}) \neq 0$ *voor iedere* $k = 1, \ldots, n$.
(de determinant van 'n $k \times k$ bovenblok van A noemt men ook wel *leidende hoofdminor*).
(b) *Als $L, L' \in \mathbf{L}_n$ en $U, U' \in \mathbf{U}_n$ zodat $A = LU = L'U'$ en* $\mathrm{diag}(L) = \mathrm{diag}(L')$ *dan $L = L'$ en $U = U'$.*

Bewijs. (a) "(ii)$\Leftrightarrow$(iii)" Beschouw een $k \times k$ bovenblok van A. Passen we op dit blok het Gauss eliminatie proces toe dan zijn de optredende matrices precies de $k \times k$ bovenblokken van de korresponderende matrices in het Gauss eliminatie proces toegepast op A (ga dit na).

Stel nu dat $a_{ii}^{(i)} \neq 0$ voor $i = 1, \ldots n-1$ zodat het eliminatie proces tot en met de $n-1$-ste stap uitvoerbaar is. Omdat $a_{11}^{(1)} \cdot a_{22}^{(2)} \cdot \ldots \cdot a_{nn}^{(n)} = \det A_n = \det(I - M_{n-1})(\cdots)(I - M_1)A = \det A \neq 0$ zien we dat $a_{nn}^{(n)} \neq 0$.

Kombineren we deze twee opmerkingen in een induktie argument, dan zien we dat

"alle leidende hoofdminoren $\neq 0$" $\Leftrightarrow$ "alle optredende $a_{kk}^{(k)} \neq 0$".

"(ii)$\Rightarrow$(i)" is precies bewezen in (2.2.2).
"(i)$\Rightarrow$(iii)" Als $A = L'U'$ met $L' \in \mathbf{L}_n$ en $U' \in \mathbf{U}_n$, dan is het produkt van het $k \times k$ boven blok van L' met dat van U' precies het $k \times k$ boven blok van A en de k-de leidende hoofdminor is $\neq 0$ als L' en U' niet-singulier zijn.
(b) Omdat $0 \neq \det A = \det L \det U$ zijn L, L', U, U' allen niet-singulier. $L^{-1}L' \in \mathbf{L}_n^1$ (ga dit na) en $UU'^{-1} \in \mathbf{U}_n$. Als $LU = L'U'$ is $UU'^{-1} = L^{-1}L'$ en volgt $L^{-1}L' = I$, $UU'^{-1} = I$. □

Uit de uniciteit van de LU-dekompositie met voorgeschreven diagonaal volgt eenvoudig

(2.2.5) Gevolg. *Als een niet-singuliere $A \in \mathbf{M}_n$ een LU-dekompositie heeft dan $A = LU$ voor zekere $L \in \mathbf{L}_n^1$ en $U \in \mathbf{U}_n$; het Gauss eliminatie proces levert precies deze L en U.* □

(2.2.6) Opgave. Met $\vec{m}_k = (0, \ldots, 0, a_{k+1\,k}^{(k)}, \ldots, a_{nk}^{(k)})^T$ is $M_k = \frac{1}{a_{kk}^{(k)}} \vec{m}_k e_k^T$. Ga dit na.
M_k is de matrix van de zogenaamde *Gauss transformatie*. De coëfficiënten $\frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ zijn *vermenigvuldigingsfaktoren*. De noemers $a_{kk}^{(k)}$ zijn de *pivots* van het proces.

(2.2.7) Opgave. Als $M \in \mathbf{L}_n^0$ zo is dat $\|M\|_m \leq 1$ dan $\|(I + M)^{-1}\|_\infty \leq 2^{n-1}$. Bewijs dit en ga na dat de schatting scherp is.

Het Gauss eliminatie proces als boven beschreven loopt vast als we een $a_{kk}^{(k)} = 0$ tegenkomen. Door dan in het stelsel twee geschikte vergelijkingen om te wisselen of in de matrix twee rijen kunnen we proberen een nieuwe pivot te maken die $\neq 0$ is. Door zo te 'pivotten' op strategische momenten in het eliminatie proces krijgen we een proces dat (afgezien van mogelijke stabiliteits problemen) altijd werkt. We geven details.

(2.2.8) Definitie. $P \in \mathbf{M}_n$ is een *permutatie matrix* als $Pe_i = e_{\pi(i)}$ $(i = 1, \ldots, n)$ voor een permutatie π van $\{1, \ldots, n\}$. Linksvermenigvuldiging $A \leadsto PA$ met P $(A \in \mathbf{M}_n)$ permuteert de rijen van de matrix (de i-de rij van A is de $\pi(i)$-de rij van PA) en rechtsvermenigvuldiging met P permuteert de kolommen. Ga dit na.

$\mathbf{P}_n$ is de ruimte van alle permutatie matrices in $\mathbf{M}_n$.

Als P een permutatie matrix is dan is P^{-1} dat ook en $P^H = P^{-1}$. Ga dit na.

(2.2.9) Gauss eliminatie met pivotten. Zij $A = (a_{ij}) \in \mathbf{M}_n$ niet-singulier.

We berekenen rekursief

$$\begin{array}{llll} & & & A_1 := A, \\ P_1, & \tilde{A}_1 = (\tilde{a}_{ij}^{(1)}) := P_1 A_1, & M_1, & A_2 = (a_{ij}^{(2)}) := (I - M_1)\tilde{A}_1, \\ P_2, & \tilde{A}_2 = (\tilde{a}_{ij}^{(2)}) := P_2 A_2, & M_2, & A_3 = (a_{ij}^{(3)}) := (I - M_2)\tilde{A}_2, \\ \vdots & & & \end{array},$$

als volgt in $n-1$ achtereenvolgende stappen:

(1) Omdat A_1 niet-singulier is is er een i zodat $a_{i1} \neq 0$. Laat P_1 de permutatie matrix zijn die bij linksvermenigvulging precies de 1-ste en de i-de rij verwisselt: dus $\tilde{a}_{11}^{(1)} = a_{i1} \neq 0$. Met $m_{i1} := \frac{\tilde{a}_{i1}^{(1)}}{\tilde{a}_{11}^{(1)}}$ en M_1 verder als boven in (2.2.2) vinden we dat $A_2 = (a_{ij}^{(2)}) := (I - M_1)\tilde{A}_1$ van de volgende vorm is

$$A_2 = \left[\begin{array}{c|ccc} \tilde{a}_{11}^{(1)} & \tilde{a}_{12}^{(1)} & \cdots & \tilde{a}_{1n}^{(1)} \\ \hline 0 & & & \\ \vdots & & A^{(2)} & \\ 0 & & & \end{array}\right] \quad \text{met} \quad A^{(2)} = \begin{bmatrix} a_{11}^{(2)} & \cdots & a_{1n}^{(2)} \\ \vdots & & \vdots \\ a_{n1}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}.$$

(2) Omdat enerzijds $\det(I - M_1)P_1 A = \det(I - M_1)\det P_1 \det A = \det A \neq 0$ en anderzijds $\det A_2 = \tilde{a}_{11}^{(1)} \det A^{(2)}$ is $A^{(2)}$ niet-singulier en moet $a_{i2}^{(2)} \neq 0$ voor zekere $i \geq 2$. Als P_2 deze i-de rij verwisselt met de 2-de en $\tilde{A}_2 := P_2 A_2 = (\tilde{a}_{ij}^{(2)})$ is $\tilde{a}_{22}^{(2)} = a_{i2}^{(2)} \neq 0$ en we kunnen weer de vermenigvuldigins faktoren m_{i2} $(i > 2)$ definiëren. etc., etc..

Tenslotte hebben we $U := A_n = (I - M_{n-1})P_{n-1}(I - M_{n-2})P_{n-2} \cdots P_2(I - M_1)P_1 A$ een boven driehoeks matrix berekend.

De P_k's verwisselen slechts twee rijen. Dus $P_k P_k = I$. Omdat P_k de k-de rij verwisselt met een i-de voor een of andere $i \geq k$ is $M_j P_k = M_j$ voor iedere $j < k$ en heeft $P_k M_j$ voor iedere $j < k$ dezelfde vorm als M_j (ga dit na). Schrijf $P := P_{n-1} \cdots P_2 P_1$, $\tilde{M}_j := P_{n-1} \cdots P_{j+1} M_j$ $(j < n-1)$ en $\tilde{M}_{n-1} := M_{n-1}$. Dan heeft $\tilde{M}_j$ ook weer dezelde vorm als M_j. Bovendien is

$$(I - M_{n-1})P_{n-1}(I - M_{n-2})P_{n-2} \cdots P_2(I - M_1)P_1 = (I - \tilde{M}_{n-1})(I - \tilde{M}_{n-2}) \cdots (I - \tilde{M}_1)P.$$

Ook hier is $\tilde{M} := \tilde{M}_1 + \tilde{M}_2 + \cdots \tilde{M}_{n-1}$ strikt boven driehoeks en zien we met $L := I + \tilde{M}$ (als in (2.2.2)) dat $PA = LU$.

Het opsporen van de nieuwe pivot $\tilde{a}_{ik}^{(k)} \neq 0$ met $i \geq k$ (in dat gedeelte van de k-de kolom dat we wensen te vegen) en het verwisselen van de k-de met de i-de rij noemt men *pivotten*.

Om stabiliteits redenen spoort men liever telkens die $i \geq k$ op waarvoor $|\tilde{a}_{ik}^{(k)}| = \max\{|\tilde{a}_{jk}^{(k)}| \,\big|\, j \geq k\}$ (ook als $\tilde{a}_{kk}^{(k)}$ al $\neq 0$ mocht zijn!). Deze specifieke pivot keuze en verwisseling heet *partial pivotten*.

We vatten de wiskundige konsekwentie samen in een stelling.

(2.2.10) Stelling. *Zij $A \in \mathbf{M}_n$ niet-singulier.*
Dan $PA = LU$ *voor zekere* $P \in \mathbf{P}_n$, $L \in \mathbf{L}_n^1$ *en* $U \in \mathbf{U}_n$. □

De efficiëntie

Voor de praktijk is het proces in (2.2.9) veel interessanter dan de wiskundige konklusie in de stelling (2.2.10). Toch moet men het proces niet precies zo uitvoeren als in (2.2.9) beschreven is. We produceerden daar onderweg matrices M_k en P_k die vrijwel helemaal vol staan met nullen. Opslaan van deze matrices vergt zinloos een gigantische hoeveelheid geheugenruimte.

In de volgende alinea laten we zien dat Gauss eliminatie met (partial) pivotten zich kan afspelen binnen precies een $n \times n$ matrix-array uitgebreid met een vektor-array.

(2.2.11) Praktische Gauss eliminatie met pivotten. Als het eliminatie proces met pivotten afgelopen is hoeven we alleen de matrices $U = A_n$, P en $\tilde{M} = \tilde{M}_1 + \cdots \tilde{M}_{n-1}$ over te hebben.

In het volgende algorithme zijn de P_k en M_k als in (2.2.9). De U en de $\tilde{M}$ in het algorithme worden iedere stap herzien; de U en $\tilde{M}$ die we moeten hebben rollen uiteindelijk uit het algorithme.

Algorithme

$k = 1$, $U = A$, $\vec{p} = (1, 2, 3, \ldots, n)^T \in \mathbf{K}^n$, $\tilde{M} = 0 \in \mathbf{M}_n$;
`(a)` `bepaal` P_k, `bereken` $\vec{p} := P_k\vec{p}$, $U := P_kU$, $\tilde{M} := P_k\tilde{M}$;
`(b)` `bepaal` M_k, `bereken` $U := (I - M_k)U$, $\tilde{M} := \tilde{M} + M_k$;
`(c)` $k := k + 1$; `ga naar (a) als` $k < n$

In stap (a) wordt gepivot en in stap (b) geëlimineerd (of geveegd).
Het algorithme levert de U en $\tilde{M}$ die we moeten hebben (ga dat zorgvuldig na; gebruik de definitie van $\tilde{M}$ en $\tilde{M}_k$ in (2.2.9)). De vektor $\vec{p}$ vertelt ons precies wat de uiteindelijke permutatie matrix P is: als $\vec{p} = (\pi_1, \pi_2, \ldots, \pi_n)^T$ dan $Pe_i = e_{\pi_i}$ (ga dat na). De eindvektor $\vec{p}$ is net zo informatief als P, immers: $P\vec{b} = (\beta_{\pi_1}.\beta_{\pi_2}, \ldots, \beta_{\pi_n})^T$ voor ieder $\vec{b} = (\beta_1, \ldots, \beta_n)^T \in \mathbf{K}^n$ (ga dat na).
ad(a) Geen van de matrices P_k mag ooit geproduceerd worden en $\vec{p}$, U en $\tilde{M}$ mogen niet via een expliciete matrix vermenigvuldiging berekend worden. In het algorithme hierboven symbolizeert de P_k de pivot stap: het opsporen van een geschikte pivot en het daarop volgend verwisselen van twee coëfficiënten in $\vec{p}$ en twee rijen in U en $\tilde{M}$.
ad(b) De matrices M_k van de Gauss transformaties moeten natuurlijk ook niet expliciet opgesteld worden. De vermenigvuldigingsfaktoren m_{ik} moeten berekend worden, U moet worden bijgewerkt (maar ook niet door een expliciete matrix vermenigvuldiging), en de faktoren moeten worden "bijgezet" in $\tilde{M}$.

Omdat in iedere stap het niet-0-deel van U disjunkt ligt van het niet-0-deel van $\tilde{M}$ kunnen (de interessante delen van) deze matrices bijgehouden worden het array van de oorspronlijke A (of U). Het hele proces kan zich afspelen in een $n \times (n + 1)$ array `C`. We geven wat meer algorithmische details.
Beschouw het array `C` ($= (c_{ij})$) hieronder zoals het uitziet in de tweede stap ($k = 2$) na de pivot-slag (stap (a)).

$$
\mathtt{C} \quad = \quad
\begin{array}{cc|cccc|c}
\mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots & \mathtt{u} & \mathtt{p} \\
\mathtt{m} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots & \mathtt{u} & \mathtt{p} \\
\hline
\mathtt{m} & \mathtt{m} & \mathtt{u} & \mathtt{u} & \ldots & \mathtt{u} & \mathtt{p} \\
\mathtt{m} & \mathtt{m} & \mathtt{u} & \mathtt{u} & \ldots & \mathtt{u} & \mathtt{p} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots
\end{array}
$$

De in het figuur aangegeven elementen `u` stellen de coëfficiënten van $U = (u_{ij})$ $(= \tilde{A}_2)$ voor op de korresponderende plaats, `m` de coëfficiënten van $\tilde{M} = (\tilde{m}_{ij})$ en `p` van $\vec{p}$. De eliminatie-stap (b) wordt als volgt (als $k = 3$) uitgevoerd.

Algorithme; stap (b)

`voor` $j = 4$ `tot` n `bereken`
  $(\tilde{m}_{j3} =)$ c_{j3}`:=`$\frac{c_{j3}}{c_{33}}$ $(= \frac{u_{j3}}{u_{33}})$;
`voor` $j = 4$ `tot` n
  `voor` $i = 4$ `tot` n `bereken`
    $(u_{ji} =)$ c_{ji}`:=`$c_{ji} - c_{j3}c_{3i}$ $(= u_{ji} - \tilde{m}_{j3}u_{3i})$;

De pivot stap bestaat uit het opsporen van de pivot en het verwisselen van de twee gekozen rijen van het array `C` (inklusief `p`).

Het Gauss proces is efficiënt met betrekking tot de benodigde hoeveelheid geheugenruimte (als we niet pivotten hebben we precies evenveel geheugenruimte nodig voor de tussenresultaten als voor de oorsronkelijke matrix, met pivotten hebben we slechts een vektor-array extra nodig.) We berekenen nu hoeveel flop we nodig hebben om de LU-dekompositie op te stellen van een matrix A die vol is (waarvan mogelijk alle $a_{ij} \neq 0$ zijn).

(2.2.12) De kosten van een LU-dekompositie. Stel dat de niet-singuliere matrix $A \in \mathbf{M}_n$ een LU-dekompositie heeft. We bekijken hoeveel flop het berekenen van L met $\operatorname{diag}(L) = I$ en U kost met het Gauss proces in (2.2.2) (het proces in (2.2.11) zonder pivotten).

In de k-de stap berekenen we $m_{k+1\,k}, \ldots, m_{nk}$ met $n-k$ flop en $A_{k+1} = (I - M_k)A_k$ met $(n-k)^2$ flop (ga dat na). L en U opstellen kost dus in het totaal
$(n-1) + (n-1)^2 + \ldots + 1 + 1^2 = \frac{1}{3}n^3 - \frac{1}{3}n$ flop.

(2.2.13) Verdere kosten. Stel dat we L met $\operatorname{diag}(L) = I$ en U bepaald hebben. We moeten dan $\vec{x}$ oplossen uit $A\vec{x} = \vec{b}$ in twee stappen: eerst $\vec{y}$ uit $L\vec{y} = \vec{b}$ met voorwaartse eliminatie en daarna $\vec{x}$ uit $U\vec{x} = \vec{y}$ met terugwaartse substitutie. Samen kosten deze stappen, gegeven de L en U, precies n^2 flop (ga dit na; zie (2.1.2) en bedenk dat $\operatorname{diag}(L) = I$).

(2.2.14) Opgave. In het oorspronkelijke Gauss eliminatie proces voor stelsels worden de vermenigvuldigingsfaktoren m_{ij} niet 'onthouden', maar wordt in iedere eliminatie stap de rechterlid vektor 'bijgewerkt'. Ga na dat deze aanpak (bij 'volle' $\vec{b}$) precies evenveel flop vergt als bovenstaande matrix methode; eerst L en U bepalen en daarna twee driehoeks stelsels oplossen. Ga ook na dat beide methoden (als er niet gepivot wordt) evenveel geheugenruimte vergen.

(2.2.15) Opmerking. Stel dat L met $\operatorname{diag}(L) = I$ en U berekend zijn. Dan kunnen we voor iedere $\vec{b}$ de vektor $\vec{x}$ oplossen uit $A\vec{x} = \vec{b}$ in n^2 flop; dit is (i.h.a.) precies de prijs van een matrix-vektor vermenigvulding.

We zouden nu ook A^{-1} expliciet kunnen bepalen door $A\vec{b}_i = e_i$ voor $i = 1, \ldots, n$ op te lossen. De matrix-vektor vermenigvuldiging $\vec{x} = A^{-1}\vec{b}$ kost echter ook n^2 flop. Dus zelfs als men meerdere vergelijkingen met dezelfde A wil oplossen is het verspilde energie om expliciet A^{-1} te bepalen.

(2.2.16) Opgave. Zij $A \in \mathbf{M}_n$. Hoeveel flop kost het berekenen van $\det(A)$ met behulp van de formule $\det(A) = \sum_{j=1}^{n} (-1)^{j+1} a_{j1} \det(\hat{A}_j)$, waarbij $\hat{A}_j$ de matrix A is waaruit de eerste kolom en de j-de rij geschrapt zijn ("ontwikkelen naar de eerste kolom"). Laat zien de je $\det(A)$ ook in $\frac{1}{3}n^3 + \frac{2}{3}n$ flop kan berekenen.

Beschouw $A\vec{x} = \vec{b}$ met $\vec{x} = (\alpha_1, \ldots, \alpha_n)^T$ en $\vec{b} = (\beta_1, \ldots, \beta_n)^T$. Dan
$\alpha_j = \det(\Pi_j(A;\vec{b}))/\det(A)$ waarbij $\Pi_j(A;\vec{b})$ de matrix A is waarin de j-de kolom vervangen is door $\vec{b}$ (de *regel van Cramer*). Voor welke n ga je de belangstelling verliezen voor deze regel als praktische oplosmethode?

Varianten van de Gauss methode

Er zijn diverse andere methode die zich slechts van het Gauss eliminatie proces onderscheiden door de volgorde waarin de bewerkingen uitgevoerd worden, hetgeen soms voordelen oplevert. We bespreken een zo'n variant.

(2.2.17) De methode van Doolittle.

$$
\begin{array}{cc|ccccc}
\mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots \\
\mathtt{m} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots \\
\hline
\mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
\qquad
\begin{array}{ccc|cccc}
\mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots \\
\mathtt{m} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \mathtt{u} & \ldots \\
\hline
\mathtt{m} & \mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\mathtt{m} & \mathtt{m} & \mathtt{m} & \mathtt{a} & \mathtt{a} & \mathtt{a} & \ldots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}
$$

Stel dat in het eerste figuur de aangegeven elementen `u` en `m` reeds overeenstemmen met die van het array `C` uit (2.2.11), zoals ze zijn na het vegen van de eerste en tweede kolom. Laten de `a` nog de oorspronkelijke elementen van A zijn (afgezien van eventuele rijverwisselingen). De `u`'s zijn dus elementen van A_3. De ontbrekende elementen van A_3 zijn zonder meer te berekenen, immers:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - c_{i1}c_{1j} \quad \text{en} \quad a_{ij}^{(3)} = a_{ij}^{(2)} - c_{i2}c_{2j}$$

dus

$$a_{ij}^{(3)} = a_{ij}^{(1)} - c_{i1}c_{1j} - c_{i2}c_{2j}.$$

Maar we berekenen eerst $a_{33}^{(3)}, \ldots, a_{n3}^{(3)}$. Uit deze elementen kunnen we immers de pivot zoeken en aansluitend eventueel rijen verwisselen (zie (2.2.9)). We berekenen dan de resterende elementen van de 3-de rij van A_3 en dat zijn meteen elementen van A_4. Tenslotte delen we $c_{43}, \ldots, c_{n3}$ door c_{33} en hebben we dan de nieuwe `m`'s.
Dus uit de aangegeven elementen van `C` in het eerste figuur bepalen we de aangegen elementen van `C` in het tweede figuur.
Het zal duidelijk zijn hoe men het Doolittle proces van begin tot einde doorvoert.

Het proces is ook *numeriek* equivalent met Gauss eliminatie van (2.2.11), d.w.z. de afrond fouten in de berekende coëfficiënten van de L en U zijn bij beide exact gelijk (mits men bij beide methoden steeds dezelfde pivots kiest).

De coëfficiënten $a_{ij}^{(k)}$ berekent men bij Doolittle in feite door van a_{ij} een inprodukt af te trekken. Dat zelfde gebeurt bij Gauss kennelijk ook, omdat daar dezelfde bewerkingen uitgevoerd worden. Maar bij Gauss wordt dat inprodukt stapsgewijs opgebouwd (per stap 1 term erbij).
De Doolittle aanpak heeft hierdoor een tweetal voordelen:

- Vaak heeft men snelle routines voor het berekenen van inprodukten die bij Doolittle's methode bruikbaar zijn en bij Gauss niet (vektor-computers zijn speciaal afgestemd op dit soort vektor bewerkingen).
- Als men bij Gauss in dubbele precisie wenst te rekenen en de uiteindelijke matrices L en U weer op enkele precisie wil afronden, heeft men voor de hele matrix dubbele precisie geheugenplaats nodig (dit vergt in het algemeen twee maal zoveel ruimte als bij enkele precisie). Bij Doolittle hoeft men alleen de inprodukten in dubbele precisie uit te voeren. De verkregen resultaten kan men dan na afronding op enkele precisie in het geheugen opbergen.

LU-dekompositie methoden voor speciale matrices

Als de matrix A een speciale struktuur heeft kan men de dekompositie methode daar op afstemmen. We geven een voorbeeld.

Als A strikt positief definiet is dan is A niet-singulier en heeft A een LU-dekompositie (zie (1.5.24) en (1.5.25)). Het Gauss proces levert ons *zonder* pivotten de L met $\mathrm{diag}(L) = I$ en U met $A = LU$ (zie (2.2.5)). Omdat een positief definiete matrix zelf-geadjungeerd is kunnen we ons geheugen ruimte besparen door alleen de halve matrix (a_{ij} voor $i \geq j$) op te slaan. Het Choleski proces is een variant van de Gauss eliminatie die de zelf-geadjungeerdheid uitbuit en daardoor ongeveer een faktor twee aan geheugen spaart en een faktor twee aan snelheid wint.

(2.2.18) Choleski dekompositie. Zij $A \in \mathbf{M}_n(\mathbf{K})$ hermiets en strikt positief definiet (zie (1.5.24) en (1.5.25)). Dan is $A = U^H U$ voor zekere $U = (u_{ij}) \in \mathbf{U}_n(\mathbf{K})$ (zie (1.5.24) en (1.5.26)). Uitschrijven van het produkt $U^H U$ en de gelijkheid $A = U^H U$ levert het volgende rekenschema voor de $u_{ij} = \bar{u}_{ji}$:

$$\begin{aligned}
&u_{11} = \sqrt{a_{11}} \\
&\qquad u_{21} = \tfrac{a_{21}}{u_{11}}, u_{31} = \tfrac{a_{31}}{u_{11}}, \ldots, u_{n1} = \tfrac{a_{n1}}{u_{11}} \\
&u_{22} = \sqrt{a_{22} - |u_{21}|^2} \\
&\qquad u_{32} = \tfrac{1}{u_{22}}(a_{32} - u_{31}u_{12}), \ldots, u_{n2} = \tfrac{1}{u_{22}}(a_{n2} - u_{n1}u_{12}) \\
&\vdots \\
&u_{nn} = (a_{nn} - |u_{n1}|^2 - |u_{n2}|^2 - \ldots - |u_{n\,n-1}|^2)^{\frac{1}{2}}
\end{aligned}$$

(Als $\mathbf{K} = \mathbf{R}$ worden kennelijk alleen wortels uit positieve getallen getrokken!)

(2.2.19) Opgave. Ga na dat het Choleski proces inderdaad ongeveer de helft van de hoeveelheid werk en van de hoeveelheid geheugenruimte nodig heeft van het Gauss eliminatie proces zonder pivotten (als worteltrekken niet langer duurt dan een paar flop).

2.3 De stabiliteit van de vergelijking $Ax = b$

De L en de U die het Gauss eliminatie proces op de computer uitgevoerd ons levert zullen belast zijn met rekenfouten. We zullen verder moeten werken met een $\tilde{L}$ en een $\tilde{U}$. Van de oplossing van de vergelijking $A\vec{x} = \vec{b}$ die we met deze $\tilde{L}$ en $\tilde{U}$ met de computer berekenen kunnen we niet verwachten dat hij $\vec{x}$ beter benadert dan de oplossing $\tilde{x}$ van vergelijking $\tilde{L}\tilde{U}\tilde{x} = \vec{b}$. Geen enkele oplosmethode zal in het algemeen de exacte oplossing $\vec{x}$ van de vergelijking opleveren; van de berekende oplossing mogen we niet verwachten dat hij het beter doet dan exacte oplossing $\tilde{x}$ van een 'geperturbeerd' stelsel $\tilde{A}\tilde{x} = \tilde{b}$ waarbij $\tilde{A} = A + \Delta_A$ en $\tilde{b} = \vec{b} + \delta_{\vec{b}}$ voor zekere 'perturbatie' matrix Δ_A en 'perturbatie' vektor $\delta_{\vec{b}}$ (Δ_A en $\delta_{\vec{b}}$ zijn hopelijk klein in een of andere zin). (Als we een vergelijking met irrationale coëfficiënten invoeren in de computer hebben we al te maken met een geperturbeerd stelsel!) In deze paragraaf onderzoeken we de relatie tussen de grootte van de fout $\tilde{x} - \vec{x}$ en de grootte van de fouten Δ_A en $\delta_{\vec{b}}$; als kleine perturbaties op de vergelijking zeer grote perturbaties op de oplossing veroorzaken mogen we niet verwachten dat een numerieke oplosmethode instaat is om de oplossing zonder al te grote fouten te berekenen. We zullen alleen mogen verwachten, hopen, dat zo'n oplosmethode niet gevoeliger is voor perturbaties dan de vergelijking dat zelf is.

We zullen in een volgende paragraaf aantonen dat de numerieke oplossing $\tilde{x}$ die geleverd wordt door het Gauss proces (of door een andere numerieke methode) gezien kan worden als de exacte oplossing van een geperturbeerd stelsel.

Het geperturbeerde stelsel $\tilde{A}\tilde{x} = \tilde{b}$

We geven schattingen voor de fout in de oplossing $\tilde{x}$ van de geperturbeerde vergelijking en laten zien dat die schattingen scherp zijn in een zekere zin.

(2.3.1) Een foutschatting. Laat $\|\cdot\|$ een norm zijn op $\mathbf{X}$. Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ inverteerbaar. Voor $b \in \mathbf{X}$ voldoet $x \in \mathbf{X}$ aan

$$\mathcal{A}(x) = b. \tag{1}$$

Stel dat $\tilde{x} \in \mathbf{X}$ de oplossing is van de geperturbeerde vergelijking

$$(\mathcal{A} + \Delta_\mathcal{A})(\tilde{x}) = b + \delta_b, \tag{2}$$

voor zekere $\Delta_\mathcal{A} \in \mathcal{L}(\mathbf{X})$ en $\delta_b \in \mathbf{X}$. (2) minus (1) levert $(\mathcal{A} + \Delta_\mathcal{A})(\tilde{x} - x) + \Delta_\mathcal{A}(x) = \delta_b$. Stel dat $\|\mathcal{A}^{-1}\|\,\|\Delta_\mathcal{A}\| < 1$. Dan is $\mathcal{A} + \Delta_\mathcal{A}$ inverteerbaar (zie (1.3.13)). Dus

$$\tilde{x} - x = (\mathcal{A} + \Delta_\mathcal{A})^{-1}(\delta_b - \Delta_\mathcal{A} x) = \mathcal{A}^{-1}(\delta_b - \Delta_\mathcal{A} x) + ((\mathcal{A} + \Delta_\mathcal{A})^{-1} - \mathcal{A}^{-1})(\delta_b - \Delta_\mathcal{A} x). \tag{3}$$

Omdat $(\mathcal{A}+\Delta_\mathcal{A})^{-1} - \mathcal{A}^{-1} = (\mathcal{A}+\Delta_\mathcal{A})^{-1}\Delta_\mathcal{A}\mathcal{A}^{-1}$, is de tweede term in de rechter uitdrukkinging een faktor $(\mathcal{A} + \Delta_\mathcal{A})^{-1}\Delta_\mathcal{A}$ kleiner dan de eerste term. In de situaties die voor ons interessant zijn ($\|\mathcal{A}^{-1}\|\,\|\Delta_\mathcal{A}\|$ klein) zal de tweede term dan ook verwaarloosbaar zijn ten opzichte van de eerste. We bekijken daarom alleen de eerste term. Definiëer

$$\delta_x := \mathcal{A}^{-1}(\delta_b - \Delta_\mathcal{A}(x)); \tag{4}$$

we verwachten dat $\tilde{x} - x \approx \delta_x$. Er geldt

$$\begin{aligned} \|\delta_x\| &\leq \|\mathcal{A}^{-1}\|\,\|\delta_b - \Delta_\mathcal{A}(x)\| && (5)\\ &\leq \|\mathcal{A}^{-1}\|\,(\|\delta_b\| + \|\Delta_\mathcal{A}(x)\|) && (6)\\ &\leq \|\mathcal{A}^{-1}\|\,(\|\delta_b\| + \|\Delta_\mathcal{A}\|\,\|x\|). && (7) \end{aligned}$$

Gebruiken we dat $||b|| \leq ||\mathcal{A}||\,||x||$, dan krijgen we de volgende schatting voor de relatieve (of beter geschaalde) fout in de oplossing

$$\frac{||\delta_x||}{||x||} \leq ||\mathcal{A}^{-1}||\,||\mathcal{A}|| \left(\frac{||\Delta_{\mathcal{A}}||}{||\mathcal{A}||} + \frac{||\delta_b||}{||b||} \right) : \tag{8}$$

de geschaalde fout in de afbeelding plus de geschaalde fout in het rechterlid maal de faktor $||\mathcal{A}^{-1}||\,||\mathcal{A}||$ majoreerd blijkbaar de geschaalde fout in de oplossing. De faktor
$\mathcal{C}(\mathcal{A}) := ||\mathcal{A}^{-1}||\,||\mathcal{A}||$ wordt het *konditie-getal* van de afbeelding $\mathcal{A}$ (t.o.v. de norm $||\cdot||$) genoemd.

(2.3.2) Opgave. Bepaal t.o.v. $||\cdot||_\infty$ het konditie-getal van de matrix $A := \begin{bmatrix} 1 & 1 \\ 1-\varepsilon & 1+\varepsilon \end{bmatrix}$, met $\varepsilon = 10^{-5}$.

De schatting (8) van de geschaalde fout in de oplossing lijkt uiterst hanteerbaar. Het verbaasd ons niet dat de fout in b en de fout in $\mathcal{A}$ een rol spelen. Het konditie-getal van $\mathcal{A}$ kan echter erg groot zijn en we kunnen ons met recht afvragen of de schatting niet veel te grof is (we gebruikten immers driehoeksongelijkheden en dat soort bovenschattingen). We onderzoeken de scherpte van de eindschatting (8) door iedere stap (5) – (7) apart op zijn scherpte te beoordelen. We formuleren alvast de konklusie en vertellen daarmee gelijk wat we onder een scherpe schatting verstaan.

(2.3.3) Bewering. *Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$.
(a) *Schatting* (7) *is* scherp *in de volgende zin.*
Voor iedere $b \in \mathbf{X}$ *en alle gegeven toleranties* $\delta, \varepsilon > 0$ *zijn er perturbaties* $\Delta_{\mathcal{A}} \in \mathcal{L}(\mathbf{X})$ *en* $\delta_b \in \mathbf{X}$ *zodat met* δ_x *als in* (4) *geldt*

$$||\delta_b|| = \delta, \quad ||\Delta_{\mathcal{A}}|| = \varepsilon, \quad \textit{en} \quad ||\delta_x|| = ||\mathcal{A}^{-1}||\,(||\delta_b|| + ||\Delta_{\mathcal{A}}||\,||x||).$$

(b) *Schatting* (8) *is scherp in de volgende zin.*
Er is een $b \in \mathbf{X}$ *waarvoor er voor iedere* $\delta, \varepsilon > 0$ *perturbaties* $\Delta_{\mathcal{A}} \in \mathcal{L}(\mathbf{X})$ *en* $\delta_b \in \mathbf{X}$ *zijn zodat met* x *als in* (1) *en* δ_x *als in* (4) *geldt*

$$||\delta_b|| = \delta, \quad ||\Delta_{\mathcal{A}}|| = \varepsilon, \quad \textit{en} \quad \frac{||\delta_x||}{||x||} = ||\mathcal{A}^{-1}||\,||\mathcal{A}|| \left(\frac{||\Delta_{\mathcal{A}}||}{||\mathcal{A}||} + \frac{||\delta_b||}{||b||} \right)$$

(2.3.4) Opmerking. (a) Gewoonlijk zullen de schattingen (7) en (8) ernstige overschattingen zijn (als bijvoorbeeld $\Delta_{\mathcal{A}} = \varepsilon \cdot \mathcal{A}$ dan is $\frac{||\delta_x||}{||x||} = \frac{||\Delta_{\mathcal{A}}||}{||\mathcal{A}||}$ ongeacht de grootte van het konditie-getal van $\mathcal{A}$). Als echter alle perturbaties $\Delta_{\mathcal{A}}$ van gegeven norm mogelijk zijn dan is, volgens bovenstaand resultaat het konditie-getal een goede maat voor de mogelijke versterking van de geschaalde fout.
(b) In de praktijk zijn de perturbaties vaker van een bepaald type: als A bijvoorbeeld een beneden driehoeks matrix is dan zal de perturbatie matrix op A ten gevolge van rekenfouten (waarschijnlijk) ook een beneden driehoeks matrix zijn. Evenzo zijn de perturbaties op hermietse afbeelding ook vaak hermiets. Voor zo'n situatie zou het bovenstaand scherpheids begrip en de onderstaande analyse aangepast moeten worden. We doen dat hier echter niet.
(c) Laat $\mathcal{A}$ gegeven zijn.
Voor één gegeven b, voor één vergelijking (1) hoeft de schatting (8) met betrekking tot de term $\frac{||\delta_b||}{||b||}$ niet scherp te zijn.
Wensen we (1) in principe op te lossen voor iedere rechterlid vektor b dan zou $||b|| = ||\mathcal{A}||\,||x||$ kunnen zijn (zie (2.3.5) en (2.3.6)): we kunnen wel zeggen dat schatting (8) scherp is voor deze klasse van vergelijkingen.

Schatting (5) is scherp als $\delta_b - \Delta_{\mathcal{A}}(x)$ een maximalizerende vektor is van $\mathcal{A}^{-1}$.

(2.3.5) Definitie. $y \in \mathbf{X}$ is een *maximalizerende vektormaximalizerende vektor* van $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ (met betrekking tot $\|\cdot\|$) als $y \neq 0$ en $\|\mathcal{B}y\| = \|\mathcal{B}\|\,\|y\|$.

(2.3.6) Opgave. Bewijs dat iedere $\mathcal{B} \in \mathcal{L}(\mathbf{X})$ een maximalizerende vektor heeft. Als y maximalizerend is dan is αy voor $\alpha \in \mathbf{K}\backslash\{0\}$ dat ook. Ga dat na.

Laat $z \in \mathbf{X}$ een maximalizerende vektor zijn voor $\mathcal{A}^{-1}$ met $\|z\| = 1$. Dan is (5) scherp als $\delta_b - \Delta_{\mathcal{A}}(x) = \mu z$ voor zekere $\mu \in \mathbf{K}$. Schatting (6) is scherp als δ_b en $\Delta_{\mathcal{A}}(x)$ lineair afhankelijk zijn. Dus als $\delta_b = \mu_1 z$ en $\Delta_{\mathcal{A}}(x) = \mu_2 z$ voor zekere $\mu_1, \mu_2 \in \mathbf{K}$. Het is duidelijk dat $\delta_b = \delta z$ zou kunnen zijn. Het is echter niet duidelijk of er gegeven x en z een $\Delta_{\mathcal{A}} \in \mathcal{L}(\mathbf{X})$ is die x afbeelt in het opspansel van z. Met het oog op de scherpte van (7) zou x bovendien een maximalizerende vektor van $\Delta_{\mathcal{A}}$ moeten zijn. De volgende stelling van Hahn–Banach stelt ons in staat zo'n afbeelding te konstrueren. (Deze stelling is fundamenteel in de funktionaal analyse.)

(2.3.7) Stelling van Hahn–Banach. *Zij* $\mathbf{Y}$ *een lineaire deelruimte van* $\mathbf{X}$.
Zij $\Gamma \geq 0$ *en zij* $f_0 : \mathbf{Y} \to \mathbf{K}$ *een lineaire afbeelding zodat* $|f_0(y)| \leq \Gamma\|y\|$ *voor alle* $y \in \mathbf{Y}$.
Dan is er een lineaire afbeelding $f : \mathbf{X} \to \mathbf{K}$ *zodat*
(i) $f = f_0$ *op* $\mathbf{Y}$ *en* (ii) $|f(y)| \leq \Gamma\|y\|$ *voor alle* $y \in \mathbf{X}$.
(f is een *norm-bewarende uitbreiding* van f_0 op $\mathbf{X}$.) [1]

We kunnen nu $\Delta_{\mathcal{A}}$ definiëren en daarmee het bewijs van (2.3.3) voltooien. Door in (b) van (2.3.3) b zo te kiezen dat x maximalizerend is voor $\mathcal{A}$ volgt (b) eenvoudig uit (a).

(2.3.8) Gevolg. *Zij* $\varepsilon > 0$, $x, z \in \mathbf{X}\backslash\{0\}$ *met* $\|z\| = 1$.
Dan is er een $\Delta_{\mathcal{A}} \in \mathcal{L}(\mathbf{X})$ *zodat* $\|\Delta_{\mathcal{A}}x\| = \|\Delta_{\mathcal{A}}\|\,\|x\| = \varepsilon\|x\|$ *en* $\Delta_{\mathcal{A}}x = \mu z$ *voor zekere* $\mu \in \mathbf{K}$.

Bewijs. Neem $\mathbf{Y} = \{\alpha x \,\big|\, \alpha \in \mathbf{K}\}$ en definiëer $f_0(\alpha x) := \varepsilon\|x\|\alpha$ voor $\alpha \in \mathbf{K}$. Volgens de stelling van Hahn-Banach (met $\Gamma = \varepsilon$) is er een $f : \mathbf{X} \to \mathbf{K}$ met $f(x) = \varepsilon\|x\|$ en $|f(y)| \leq \varepsilon\|y\|$ voor iedere $y \in \mathbf{X}$. Definieer nu $\Delta_{\mathcal{A}} : \mathbf{X} \to \mathbf{X}$ door $\Delta_{\mathcal{A}}(y) := f(y)z$. Dan $\|\Delta_{\mathcal{A}}(y)\| \leq |f(y)|\|z\| \leq \varepsilon\|y\|$ voor iedere $y \in \mathbf{X}$ en $\|\Delta_{\mathcal{A}}(x)\| = \varepsilon\|x\|$. $\square$

(2.3.9) Opgave. Zij $\varepsilon > 0$, $\vec{x}, \vec{z} \in \mathbf{K}^n\backslash\{0\}$ met $\|\vec{z}\|_2 = 1$.
Definieer $\Delta_A(\vec{y}) := \frac{\varepsilon}{\|\vec{x}\|_2}(\vec{y}, \vec{x})\vec{z}$ voor iedere $\vec{y} \in \mathbf{K}^n$. Dan is $\vec{x}$ maximalizerend voor Δ_A met betrekking tot $\|\cdot\|_2$ en $\|\Delta_A\|_2 = \varepsilon$. Ga dat na.
Zie in dat zo'n eenvoudige konstruktie van Δ_A niet voorhanden is voor bijvoorbeeld de $\|\cdot\|_\infty$.

Als we een vergelijking representeren in de computer hebben we te maken met een geperturbeerd stelsel. We moeten ons dus neerleggen bij het feit dat we op zijn best een oplossing van een naburig stelsel te pakken kunnen krijgen. Als perturbaties op de vergelijking een groot effect hebben op de oplossing (groot konditie-getal) dan is dat de aard van het beestje en moeten we ons ook daar in eerste instantie bij neerleggen (zie echter ook pagina 29; door over te stappen op een hogere rekenprecisie kunnen we uiteraard de perturbaties op de vergelijking verkleinen).

[1] Bewijs. Stel eerst dat K = R. Stel er is een $y_1 \in \mathrm{X}\backslash\mathrm{Y}$ (anders is er niets te bewijzen).
Zij $\mathrm{Y}_1 := \{y + \alpha y_1 \,\big|\, \alpha \in \mathrm{R}, y \in \mathrm{Y}\}$. We laten zien dat er een norm-bewarende uitbreiding f_1 van f_0 op Y_1 bestaat; door vervolgens op deze manier weer norm-bewarend uit te breiden naar een 1 dimensie grotere ruimte Y_2, etc., etc., vinden we in hoogstens n stappen de gewenste uitbreiding f.
Definiëer $f_1(\alpha y_1 + y) = \alpha\gamma + f_0(y)$ $(\alpha \in \mathrm{R}, y \in \mathrm{Y})$ voor een nog te kiezen geschikte $\gamma \in \mathrm{R}$.
Omdat $f_0(y') - f_0(y'') = f_0(y' - y'') \leq \Gamma\|y' - y''\| \leq \Gamma\|y_1 + y'\| + \|y_1 + y''\|$ voor alle $y', y'' \in \mathrm{Y}$ is

$$\sup\{-\Gamma\|y_1 + y''\| - f_0(y'') \,\big|\, y'' \in \mathrm{Y}\} \leq \inf\{\Gamma\|y_1 + y'\| - f_0(y') \,\big|\, y' \in \mathrm{Y}\}.$$

Kies γ gelijk aan bijvoorbeeld het supremum. Dan $-\Gamma\|y_1 + y\| \leq \gamma + f_0(y) = f_1(y_1 + y) \leq \Gamma\|y_1 + y\|$ $(y \in \mathrm{Y})$.
Met α vermenigvuldigen levert tenslotte $|f_1(\alpha y_1 + y)| \leq \Gamma\|\alpha y_1 + y\|$ $(y \in \mathrm{Y})$ (onderscheid $\alpha > 0$ en $\alpha < 0$).
Stel K = C. De stelling volgt uit de 'reële stelling' door de volgende observatie uit te buiten.
X is op te vatten als een ruimte over R door alleen skalair-vermenigvuldiging met reële getallen toe te staan. Noem deze ruimte X_r.
Voor $g \in \mathcal{L}(\mathrm{X},\mathrm{C})$ definiëer $g_r(x) := \mathrm{Re}\, g(x)$ $(x \in \mathrm{X}_r)$. Dan $g_r \in \mathcal{L}(\mathrm{X}_r, R)$ en omdat
$\mathrm{Im}\, g(x) = \mathrm{Im}\,(-ig(ix)) = -\mathrm{Re}\, g(ix) = -g_r(ix)$ is $g(x) = g_r(x) - ig_r(ix)$. $\square$

Met behulp van de volgende stelling kan men achteraf kontroleren of de op een of andere manier berekende oplossing $\tilde{x}$ van vergelijking (1) voldoet aan een geperturbeerd stelsel met een perturbatie Δ_A en δ_b die we acceptabel vinden.
(Voor een $\vec{y} = (\beta_1, \ldots, \beta_n)^T \in \mathbf{K}^n$ schrijven we $\vec{y} \succeq 0$ als $\beta_i \geq 0$ voor alle i. Verder is $|\vec{y}\,| := (|\beta_1|, \ldots, |\beta_n|)^T$, etc..)

(2.3.10) Stelling van Prager–Oettli. *Zij $F \in \mathbf{M}_n$, $F \succeq 0$ en $\vec{f} \in \mathbf{K}^n$, $\vec{f} \succeq 0$. Voor een $\tilde{x} \in \mathbf{K}^n$ zijn de volgende twee uitspraken equivalent.*
(i) *Er zijn $\Delta_A \in \mathbf{M}_n$ en $\delta_{\vec{b}} \in \mathbf{K}^n$ zodat* $(A + \Delta_A)\tilde{x} = \vec{b} + \delta_{\vec{b}}$ *en* $|\Delta_A| \preceq F$ *en* $|\delta_{\vec{b}}\,| \preceq \vec{f}$.
(ii) *Voor* $\vec{r} := A\tilde{x} - \vec{b}$ *geldt* $|\vec{r}\,| \preceq F|\tilde{x}| + \vec{f}$.

Bewijs. Als (i) geldt dan $\vec{r} = (A\tilde{x} - \vec{b}\,) - ((A + \Delta_A)\tilde{x} - \vec{b} - \delta_{\vec{b}}\,) = -\Delta_A(\tilde{x}) + \delta_{\vec{b}}$ en (ii) volgt eenvoudig.

Stel dat (ii) geldt. We moeten laten zien dat $\vec{r} = -\Delta_A\tilde{x} + \delta_b$ voor zekere Δ_A en δ_b met $|\Delta_A| \preceq F$ en $|\delta_{\vec{b}}\,| \preceq \vec{f}$. Door $|\vec{r}\,| \preceq F|\tilde{x}| + \vec{f}$ koördinaatsgewijs uit te schrijven en de volgende konstatering te gebruiken volgt (i).

Als $|\rho| \leq \sum c_i|\tilde{\alpha}_i| + \varphi$ voor zekere $c_1, \ldots, c_n, \varphi \in [0, \infty)$ dan $\rho = \sum(\zeta_i c_i)\tilde{\alpha}_i + \zeta_0\varphi$ voor zekere $\zeta_i \in \mathbf{K}$, $|\zeta_i| \leq 1$. □

Als $\tilde{x}$ bekend is kost de verificatie van (ii) in bovenstaande stelling hoogstens $2n^2$ flop. Zeker als we voor maar één rechterlid vektor b de vergelijking (1) oplossen is het relatief goedkoop (t.o.v. de $\frac{1}{3}n^3$ flop die de L en U kost) om achteraf met de Prager-Oettli stelling te kontroleren of de berekende oplossing voldoende nauwkeurig is.
De berekening van de oplossing vergt gewoonlijk veel meer rekenoperaties dan nodig zijn voor de twee matrix-vektor vermenigvuldigingen in (ii); de evaluatiefouten in deze laatste resultaten zullen dan ook aanzienlijk kleiner zijn dan de fouten die we in $\tilde{x}$ verwachten.

Het konditie-getal van de vergelijking

Men noemt een vergelijking als (1) *stabiel* of *goed gekonditioneerd* als iedere kleine perturbatie op de vergelijking (op $\mathcal{A}$ en op b) slechts een klein effect heeft op de oplossing; anders spreekt men van een *instabiele* of *slecht gekonditioneerde* vergelijking. De begrippen 'klein' en 'niet klein' zijn subjektief: als we tevreden zijn met een berekende oplossing $\tilde{x}$ met een fout $\|\delta_x\| \leq 10^{-3}\|x\|$ dan is $\mathcal{C}(\mathcal{A}) = 10^6$ groot in geval $\|\Delta_{\mathcal{A}}\| \approx 10^{-6}\|\mathcal{A}\|$ en klein als $\|\Delta_{\mathcal{A}}\| \leq 10^{-12}\|\mathcal{A}\|$, $\|\delta_b\| \leq 10^{-12}\|b\|$. Het konditie-getal is een hanteerbare objektieve en betrouwbare maat die de kwalitatieve begrippen goed en slecht gekonditioneerd kwantificeert.

In de definitie van $\mathcal{C}(\mathcal{A})$ speelt de $\mathcal{A}^{-1}$ een rol die we niet expliciet kennen en ook niet wensen te kennen; om A^{-1} uit te rekenen met bijvoorbeeld een LU-dekompositie wordt het berekenen van $\mathcal{C}(A)$ 3 maal zo duur als het bepalen van L en U. Om $\|\mathcal{A}^{-1}\|$ te bepalen heeft men echter niet expliciet $\mathcal{A}^{-1}$ nodig. Bovendien is men gewoonlijk al tevreden met een (ruwe) schatting voor $\mathcal{C}(\mathcal{A})$ (als $\|\Delta_{\mathcal{A}}\| \approx 10^{-12}\|\mathcal{A}\|$ is men even tevreden met $\mathcal{C}(\mathcal{A}) \leq 10^6$ als met $\mathcal{C}(\mathcal{A}) \leq 10^5$, terwijl beide uitspraken $\mathcal{C}(\mathcal{A}) \geq 10^{15}$ en $\mathcal{C}(\mathcal{A}) \geq 10^{12}$ het ergste doen vrezen). Vaak schat men $\mathcal{C}(\mathcal{A})$ in een hanteerbare norm en en gebruikt men de resultaten in (1.2.17) om een uitspraak te krijgen voor de gewenste norm. Omdat men $\mathcal{A}$ expliciet kent is het niet moeilijk om $\|\mathcal{A}\|$ te schatten (via bv. $\|A\|_\infty$ of $\|A\|_F$); het probleem zit gewoonlijk alleen in het schatten van $\|\mathcal{A}^{-1}\|$.

In dit college zullen we (impliciet) een aantal schattings methoden voor $\mathcal{C}(\mathcal{A})$ tegenkomen. De relaties in deze subparagraaf zijn vaak behulpzaam.

In deze subparagraaf willen we op de eerste plaats inzien aan de hand van de matrix A of A een groot dan wel een klein konditie-getal heeft.

(2.3.11) Notatie. Als $\mathbf{X}$ voorzien is van een norm $\|\cdot\|$ dan is $\mathcal{C}(\mathcal{A}) = \mathcal{C}_{\|\cdot\|}(\mathcal{A}) = \|\mathcal{A}\|\,\|\mathcal{A}^{-1}\|$ het konditie-getal van de afbeelding $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ t.o.v. de norm $\|\cdot\|$.
($\mathcal{C}(\mathcal{A}) := \infty$ als $\mathcal{A}$ niet inverteerbaar is).

Verder is $\text{lub}(\mathcal{A}) := \|\mathcal{A}\| = \sup\{\|\mathcal{A}(x)\| \,\big|\, x \in \mathbf{X}, \|x\| = 1\}$ (de 'least upper bound' van $\mathcal{A}$)
en $\text{glb}(\mathcal{A}) := \inf\{\|\mathcal{A}(x)\| \,\big|\, x \in \mathbf{X}, \|x\| = 1\}$ (de 'greatest lower bound' van $\mathcal{A}$).

Voor $p \in [1, \infty]$ noteren we het konditie-getal van $A \in \mathbf{M}_n$ ten opzichte van $\|\cdot\|_p$ met $\mathcal{C}_p(A)$.

(2.3.12) Bewering. *Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$.
Dan $1 \leq \mathcal{C}(\mathcal{A})$, $\mathcal{C}(\alpha\mathcal{A}) = \mathcal{C}(\mathcal{A})$ $(\alpha \in \mathbf{K})$, $\mathcal{C}(\mathcal{A}\mathcal{B}) \leq \mathcal{C}(\mathcal{A})\mathcal{C}(\mathcal{B})$ $(\mathcal{B} \in \mathcal{L}(\mathbf{X}))$.
Verder $\mathcal{C}(\mathcal{A}) = \frac{\text{lub}(\mathcal{A})}{\text{glb}(\mathcal{A})} = \sup\{\frac{\|\mathcal{A}(u)\|}{\|\mathcal{A}(v)\|} \,\big|\, \|u\| = \|v\| \neq 0\}$. □

We laten het bewijs van de volgende twee beweringen aan de lezer over.

(2.3.13) Bewering. *Stel* $\mathbf{X}$ *is voorzien van een inprodukt en inprodukt-norm. Zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$.
Als $\mathcal{Q}, \mathcal{Q}_1, \mathcal{Q}_2 \in \mathcal{L}(\mathbf{X})$ *unitair zijn dan* $\mathcal{C}(\mathcal{Q}) = 1$ *en* $\mathcal{C}(\mathcal{Q}_1\mathcal{A}\mathcal{Q}_2^*) = \mathcal{C}(\mathcal{A})$.

$\sigma_1, \ldots, \sigma_n$ *noemt men* de singuliere waarden van $\mathcal{A}$ *als* $\sigma_i \in [0, \infty)$ *en* $\sigma_i^2 = \lambda_i(\mathcal{A}^*\mathcal{A})$ *de i-de eigenwaarde van* $\mathcal{A}^*\mathcal{A}$ *is* $(i = 1, \ldots, n)$. *Laat de singuliere waarden* $\sigma_1, \ldots, \sigma_n$ *zo geordend zijn dat* $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$. *Dan* $\text{lub}(\mathcal{A}) = \sigma_1$, $\text{glb}(\mathcal{A}) = \sigma_n > 0$ *en* $\mathcal{C}(\mathcal{A}) = \frac{\sigma_1}{\sigma_n}$. □

(2.3.14) Opgave. Bewijs (2.3.12) en (2.3.13).

(2.3.15) Voorbeelden. (a) Met $A =: \begin{bmatrix} 2^{18} & 2^{18} \\ -2^{-18} & 2^{-18} \end{bmatrix}$ is $A^{-1} = \begin{bmatrix} 2^{-19} & -2^{17} \\ 2^{-19} & 2^{17} \end{bmatrix}$ en $\mathcal{C}_2(A) = 2^{36}$; als in een matrix de rijen (of kolommen) onderlingen sterk in lengte verschillen dan is het konditie-getal groot

(b) Met $0 < \varepsilon \ll 1$ en $A := \begin{bmatrix} 1 & 1 \\ 1-\varepsilon & 1+\varepsilon \end{bmatrix}$ is $A^{-1} := \frac{1}{2\varepsilon}\begin{bmatrix} 1+\varepsilon & -1 \\ \varepsilon - 1 & 1 \end{bmatrix}$ en $\mathcal{C}_2(\mathcal{A}) \approx \frac{4}{\varepsilon}$; $\mathcal{C}_2(A)$ is ook groot als de lengten van de rijen min of meer gelijk zijn maar een van de hoeken tussen de kolom vektoren (of rij vektoren) klein is.

We besteden wat meer aandacht aan de situatie in voorbeeld (b); we onderzoeken wat nauwkeuriger het verband tussen de hoeken van de kolom vektoren van een matrix A en de grootte van het konditie-getal.

(2.3.16) Definitie (de hoek tussen een vektor en een deelruimte).
We noemen $\varphi \in [0, \pi)$ de *hoek* tussen u en v in $\mathbf{X}\backslash\{0\}$ als $\cos(\varphi) = \text{Re}\frac{(u,v)}{\|u\|_2\,\|v\|_2}$.
(Ga na dat deze definitie konsistent is met de gebruikelijke geometrische ook in geval $\mathbf{X} = \mathbf{C}$.)
Als $\|u\|_2 = \|v\|_2 = 1$ dan is $u = c_\varphi v + s_\varphi w$ voor zekere $w \in \mathbf{X}$ met $\|w\|_2 = 1$, $w \perp v$ en zekere $c_\varphi, s_\varphi \in \mathbf{K}$ met $|c_\varphi|^2 + |s_\varphi|^2 = 1$ en $\cos(\varphi) = \text{Re}(c_\varphi)$.

Als $\mathbf{Y} \subset \mathbf{X}$ en $u \in \mathbf{X}$ dan $\text{dist}_2(u, \mathbf{Y}) := \inf\{\|u - v\|_2 \,\big|\, v \in \mathbf{Y}\}$ is de *afstand* tussen u en $\mathbf{Y}$. Als $\|u\|_2 = 1$ en $\mathbf{Y}$ is lineair dan is $\text{dist}_2(u, \mathbf{Y}) = \sin(\varphi)$, waarbij φ de *hoek* is tussen u en $\mathbf{Y}$ ($\varphi \in [0, \frac{1}{2}\pi]$, en φ is de minimale hoek tussen u en vektoren in $\mathbf{Y}\backslash\{0\}$).

In een volgend hoofdstuk besteden we uitgebreid aandacht aan de hoek tussen twee lineaire deelruimten. Hier karakterizeren we nog de hoek tussen twee vektoren (voor het bewijs van stelling (2.3.21)) en de afstand tussen een vektor en een lineaire deelruimte.

(2.3.17) Eigenschap. *Als* φ *de hoek is tussen* u *en* v *in* $\mathbf{X}$ *met* $\|u\|_2 = \|v\|_2 \neq 0$ *dan*

$$\coth(\frac{1}{2}\varphi) = \frac{\|u + v\|_2}{\|u - v\|_2}.$$

Bewijs. We mogen zonder de algemeenheid te schaden aannemen dat $\|u\|_2 = \|v\|_2 = 1$. Schrijf $u = c_\varphi v + s_\varphi w$ als in de definitie (2.3.16). Dan volgt het lemma verder door eenvoudig uitschrijven. □

(2.3.18) Eigenschap. *Zij* $\mathbf{Y}$ *een lineaire deelruimte van* $\mathbf{X}$ *en zij* $u \in \mathbf{X}$*. Dan is* u *op precies een manier te schrijven als* $u = u' + u''$ *met* $u' \in \mathbf{Y}$ *en* $u'' \perp \mathbf{Y}$*. Verder is* $\mathrm{dist}_2(u, \mathbf{Y}) = \|u''\|_2$*.*

Bewijs. Kies een orthonormale basis $q_1, \ldots, q_m$ van $\mathbf{Y}$. Vul die aan tot een orthonormale basis $q_1, \ldots, q_n$ van $\mathbf{X}$. Met $u' := \sum_{i \le m}(u, q_i)q_i$ en $u'' := \sum_{i>m}(u, q_i)q_i$ is $u' \in \mathbf{Y}, u'' \perp \mathbf{Y}$ en $u = u' + u''$.

Als $v + w = v' + w'$ met $v, v' \in \mathbf{Y}, w, w' \perp \mathbf{Y}$ dan $v - v' = w' - w$ en $v - v' \perp w' - w$. Dus $v = v', w' = w$.

Als $y \in \mathbf{Y}$ dan $u' - y \perp u''$ en
$\|u - u'\|_2^2 = \|u''\|_2^2 \le \|u''\|_2^2 + \|u' - y\|_2^2 = \|u'' + u' - y\|_2^2 = \|u - y\|_2^2$. □

(2.3.19) Lemma. *Zij* $A \in \mathbf{M}_n$ *niet-singulier.*
Met $\mathbf{V}_i := \{\sum_j \alpha_j A e_j \,\big|\, \alpha_i = 0\}$ *is* $d_i := \mathrm{dist}_2(Ae_i, \mathbf{V}_i)$ *de afstand tussen de i-de kolom vektor van* A *en het opspansel* $\mathbf{V}_i$ *van de overige kolommen* $(i = 1, \ldots, n)$*.*
Dan $\frac{d_i}{\|A\|_2} \le \frac{d_i}{\|Ae_i\|_2} = \sin(\varphi_i)$*, waarbij* φ_i *de hoek is tussen* Ae_i *en* $\mathbf{V}_i$*.*
Noteer de getransponeerde van de i-de rij van de matrix A^{-1} *met* $\tilde{a}_i \in \mathbf{K}^n$*. Dan is* $d_i = \frac{1}{\|\tilde{a}_i\|_2}$*.*

Bewijs. Merk op dat $\tilde{a}_i = (A^{-1})^T e_i$. We bewijzen $d_1 = 1/\|\tilde{a}_1\|_2$.
Omdat $A^{-1}A = I$ is $(A^{-1})^* e_1 \perp Ae_j$ voor iedere $j = 2, \ldots, n$ en $((A^{-1})^* e_1, Ae_1) = 1$.
Als $Ae_1 = \alpha(A^{-1})^* e_1 + \sum_{j=2}^n \alpha_j A e_j$ dan $d_1 = \|\alpha(A^{-1})^* e_1\|_2$. Omdat $((A^{-1})^* e_1, Ae_1) = 1$ volgt dat $1 = \alpha\|(A^{-1})^* e_1\|_2^2$ en dus $d_1 = \frac{1}{\|(A^{-1})^* e_1\|_2} = \frac{1}{\|(A^{-1})^T e_1\|_2}$. □

(2.3.20) Bewering. *Zij* $A \in \mathbf{M}_n$ *niet-singulier.*
Met d_i *als in* (2.3.19) *geldt*

$$\frac{\|A\|_2}{d_i} \le \mathcal{C}_2(A) \le \sqrt{n} \max_j \frac{\|A\|_2}{d_j} \qquad \textit{voor alle} \quad i = 1, \ldots, n.$$

Bewijs. De bewering volgt uit de volgende ongelijkheden die korrekt zijn voor iedere i.
$\frac{1}{d_i} \le \|(A^{-1})^T e_i\|_2 \le \|(A^{-1})^T\|_2 \le \|A^{-1}\|_F \le \sqrt{n}\max_j \|(A^{-1})^T e_j\|_2 = \sqrt{n}\max_j \frac{1}{d_j}$. □

Noemen we een stelsel vektoren *bijna afhankelijk* als een van hen een kleine hoek maakt met het hyper-vlak opgespannen door de overige, dan heeft A een groot konditie-getal als de kolom vektoren van A bijna afhankelijk zijn. De omgekeerde bewering is niet juist: A kan een groot konditie-getal hebben terwijl de kolommen bijvoorbeeld onderling loodrecht staan (zie voorbeeld (a) in (2.3.15)). Als de kolommen min of meer gelijke lengte hebben dan impliceert een groot konditie-getal wel bija afhankelijkheid (als de kolommen gelijke lengte hebben dan is $\|A\|_2 \le \|A\|_F = \sqrt{n}\|Ae_i\|_2$ en dus is er volgens de twee laatste beweringen een j met $\mathcal{C}_2(A) \le \frac{n}{\sin(\varphi_j)}$).

Als $\vec{y} \in \mathbf{V}_i$ en $\vec{x} = Ae_i$ dan zijn $\vec{x}$ en $\vec{y}$ de beelden onder A van twee vektoren die onderling loodrecht op elkaar staan. De volgende stelling generalizeert daarom ook bovenstaande bewering.

(2.3.21) Stelling. *Zij* $\mathbf{X}$ *een ruimte met inprodukt en zij* $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ *inverteerbaar.*
Laat ϑ *de minimale hoek zijn tussen de beelden* $\mathcal{A}(u)$ *en* $\mathcal{A}(v)$*, waarbij het minimum genomen wordt over alle vektoren* $u, v \in \mathbf{X}\backslash\{0\}$ *waarvoor* $u \perp v$*.*
Dan

$$\mathcal{C}(\mathcal{A}) = \coth(\tfrac{1}{2}\vartheta)$$

Bewijs. "$\ge$" Beschouw $u, v \in \mathbf{X}\backslash\{0\}$ waarvoor $u \perp v$. Omdat we in hoeken geïnteresseerd zijn mogen we aannemen dat $\|\mathcal{A}(u)\| = \|\mathcal{A}(v)\|$. Laat φ de hoek tussen $\mathcal{A}(u)$ en $\mathcal{A}(v)$ zijn. Omdat $\|u + v\| = \|u - v\|$ (Pythagoras) is volgens lemma (2.3.17) en (2.3.12)

$$\coth(\tfrac{1}{2}\varphi) = \frac{\|\mathcal{A}(u+v)\|}{\|\mathcal{A}(u-v)\|} \le \mathcal{C}(\mathcal{A}).$$

"$\le$" Laat $q_1, \ldots, q_n$ een orthonormaal stelsel zijn zodat $\mathcal{A}^*\mathcal{A}q_i = \sigma_i^2 q_i$. (Zie (1.5.12). De σ_i zijn de singuliere waarden als in (2.3.13). De q_i zijn de *rechts-singuliere vektoren.*) Omdat $\sigma_i^2(q_i, q_j) = (\mathcal{A}^*\mathcal{A}(q_i), q_j) = (\mathcal{A}(q_i), \mathcal{A}(q_j))$ is $\mathcal{A}(q_1) \perp \mathcal{A}(q_n)$ en $\sigma_i = \|\mathcal{A}(q_i)\|$. Beschouw

$u := q_1 + q_n$ en $v := q_1 - q_n$. Dan $u \perp v$ en $\|\mathcal{A}(u)\|^2 = \|\mathcal{A}(v)\|^2 = \sigma_1^2 + \sigma_n^2 \neq 0$. Dus als ϑ de hoek is tussen $\mathcal{A}(u)$ en $\mathcal{A}(v)$ dan geldt volgens (2.3.17) en (2.3.13)

$$\coth(\tfrac{1}{2}\vartheta) = \frac{\|\mathcal{A}(u+v)\|}{\|\mathcal{A}(u-v)\|} = \frac{\|\mathcal{A}(q_1)\|}{\|\mathcal{A}(q_n)\|} = \frac{\sigma_1}{\sigma_n} = \mathcal{C}(\mathcal{A}). \qquad \square$$

Verbetering van de stabiliteit

Als in een vergelijking $\mathcal{A}(x) = b$ het konditie-getal $\mathcal{C}(\mathcal{A})$ zeer groot is mogen we niet verwachten dat we met welke numerieke methode dan ook de vergelijking in enige acceptabele nauwkeurigheid kunnen oplossen: de oplossing moet op een andere manier gezocht worden, mogelijk via een andere goed gekonditioneerde vergelijking met dezelde oplossing, of door de perturbaties op de vergelijking klein te maken (door nauwkeuriger te rekenen of, als het stelsel door metingen opgesteld is, door nauwkeuriger te meten). Beide aktiviteiten vallen in feite buiten het kader van de numerieke wiskunde. In de numerieke praktijk kennen we echter vaak aanvankelijk de vergelijking min of meer exact. Pas in het numerieke oplosproces verschijnen de grote perturbaties. We kunnen proberen de vergelijking te veranderen in een stabielere *voordat* we met het zware numerieke werk beginnen.

(2.3.22) Opgave. Bepaal $\mathcal{C}_\infty(A)$ en $\mathcal{C}_\infty(DA)$ met A als in (a) van (2.3.15) en $D = \text{diag}\,(2^{-18}, 2^{+18})$.

Als we $A \in \mathbf{M}_n$ en $\vec{b} \in \mathbf{K}^n$ exact kennen op eventuele afrondfouten na ($a_{ij}(1+\xi)$ i.p.v. a_{ij}) en de grote peturbaties op de vergelijking verschijnen pas bij het numeriek op lossen (door bijvoorbeeld middels Gauss eliminatie L en U te bepalen) dan mogen we zorgeloos $A\vec{x} = \vec{b}$ met bijvoorbeeld een niet-singuliere diagonaal matrix $D \in \mathbf{M}_n$ vermenigvuldigen, immers: $\vec{x}$ is ook de oplossing van $DA\vec{x} = D\vec{b}$ en DA en $D\vec{b}$ zijn nog steeds min of meer exact bekend (rekenen we $d_{ii}a_{ij}$ in de computer uit dan vinden we $d_{ii}a_{ij}(1+3\xi)$). We kunnen via een geschikte diagonaal D proberen de konditionering van de vergelijking te verbeteren[2]: we zoeken in deze subparagraaf een D waarvoor $\mathcal{C}(DA) \leq \mathcal{C}(A)$.

Als de vergelijking waarmee we aan de slag moeten al met grote fouten belast is (bv. meetfouten; groot t.o.v. de latere numerieke fouten) heeft het geen zin te 'prekonditioneren', immers: door met een diagonaal D te vermenigvuldigen kunnen we misschien wel het konditie-getal verkleinen maar de fouten worden evenzo met D vermenigvuldigd.
Om dezelde reden heeft het geen zin te 'prekonditioneren' door A rechts met D te vermenigvuldigen: als $AD\vec{y} = \vec{b}$ dan is het gemakkelijk zat om $\vec{x}$ uit $D\vec{y} = \vec{x}$ te bepalen, maar ook hier moeten we de fouten met D vermenigvuldigen (door naar $\vec{y}$ te kijken hebben we de fouten slechts tijdelijk prettig geschaald!). De fouten waarop we D moeten laten werken zijn die waarmee te hebben *na* het numeriek oplossen van het stelsel.

Het konditie-getal van een matrix A is groot als bijvoorbeeld de rijen onderling sterk in lengte verschillen. Door A met een geschikte diagonaal D_0 te vermenigvuldigen kunnen we er voor zorgen dat de rijen gelijke lengte krijgen; men noemt dit *rij equilibreren* . In zekere zin is dit de beste manier om te prekonditioneren met een diagonaal matrix. Om dit aan te tonen doen we de definitie van het konditie-getal een beetje geweld aan.

(2.3.23) Notatie. Schrijf $\mathcal{C}'(A) := \|A^{-1}\|_\infty \|A\|_m$ voor $A \in \mathbf{M}_n$ ($\|\cdot\|_m$ als op p. 4).
$\mathcal{C}'(A)$ is geen konditie-getal. Er geldt wel bv. $\mathcal{C}'(\alpha A) = \mathcal{C}'(A)$ voor $\alpha \in \mathbf{K}$.
Verder is bijvoorbeeld $\mathcal{C}_\infty(A) \leq \mathcal{C}'(A) \leq n\mathcal{C}_\infty(A)$.

(2.3.24) Stelling. *Zij $A = (a_{ij}) \in \mathbf{M}_n$ niet-singulier.*
Laat $D_0 \in \mathbf{M}_n$ een diagonaal matrix zijn zodat alle rijen van de matrix D_0A gelijke sup-norm hebben: $D = \text{diag}(d_{11}^0, \ldots, d_{nn}^0)$ met $d_{ii}^0 := \max_i \frac{1}{|a_{ij}|}$ voor iedere $i = 1, \ldots, n$.

[2] We hoeven voor de 'prekonditionerings matrix' D niet per-sé een diagonaal matrix te nemen. Wel moeten DA en $D\vec{b}$ vrijwel foutloos en goedkoop uitgerekend kunnnen worden, D moet efficiënt te bepalen zijn en de struktuur van DA moet niet 'moeilijker' zijn dan die van A (als A bv. tridiagonaal is dan is het niet wenselijk om DA vol te hebben).

Dan $\mathcal{C}'(D_0A) \leq \mathcal{C}'(DA)$ *voor iedere diagonaal matrix* $D \in \mathbf{M}_n$.

Bewijs. Beschouw een diagonaal $D = \mathrm{diag}(d_{11}, \ldots, d_{nn})$.

$$||(D_0A)^{-1}||_\infty = ||(DD_0A)^{-1}D||_\infty \leq ||(DD_0A)^{-1}||_\infty \, ||D||_\infty$$

$$||DD_0A||_m = \max_j ||d_{jj}(D_0A)^T e_j||_\infty = \max_j |d_{jj}| \, ||(D_0A)^T e_1||_\infty = ||D||_\infty \, ||D_0A||_m$$

Dus $||(D_0A)^{-1}||_\infty ||D_0A||_m \leq \mathcal{C}'(DD_0A)$. □

D_0A zal geen optimaal $||\cdot||_\infty$ konditie-getal hebben. Maar gezien schatting als $\mathcal{C}_\infty(A) \leq \mathcal{C}'(A) \leq n\mathcal{C}_\infty(A)$ zal deze prekonditionering ook m.b.t tot andere normen behoorlijk goed zijn. Als we al tevreden zijn met een min of meer optimaal konditie-getal dan kunnen we foutloos, door in iedere rij van de matrix de decimale punt op een geschikte manier te verplaatsen (per rij in iedere coëfficiënt hetzelfde aantal plaatsen), een matrix krijgen waarvan de rijen min of meer gelijke lengte hebben.

2.4 De stabiliteit van de oplosmethode

De stabiliteit van de methode

Een dekompositie methode als Gauss eliminatie levert in principe voor een matrix A matrices L en U zodat $A = LU$. Om de vergelijking $A\vec{x} = \vec{b}$ op te lossen moeten we dan nog de gesplitste vergelijking $L\vec{y} = \vec{b}$ en $U\vec{x} = \vec{y}$ op lossen. Perturbaties op deze vergelijkingen hebben hun effect op de oplossing.
Naast deze perturbaties hebben we ook nog te maken met het feit dat de dekompositie methode niet de exacte L en U levert. We stellen het bespreken van de invloed van dit type fouten nog even uit.

(2.4.1) Foutschatting voor een gesplitst stelsel. Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X})$ inverteerbaar. Stel dat we over een dekompositie $\mathcal{A} = \mathcal{B}\mathcal{C}$ van $\mathcal{A}$ beschikken. Voor $b \in \mathbf{X}$ is x de oplossing van $\mathcal{A}(x) = b$ precies dan als $\mathcal{B}(y) = b$ en $\mathcal{C}(x) = y$. Beschouw het geperturbeerde paar vergelijkingen $\left\{\begin{array}{rcl} (\mathcal{B}+\Delta_\mathcal{B})\tilde{y} & = & b+\delta_b \\ (\mathcal{C}+\Delta_\mathcal{C})\tilde{x} & = & \tilde{y} \end{array}\right.$. Met $\Delta_\mathcal{A} := \mathcal{B}\Delta_\mathcal{C} + \Delta_\mathcal{B}\mathcal{C}$ en $\delta_x := \mathcal{A}^{-1}(\delta_b - \Delta_\mathcal{A}(x))$ zal $\tilde{x} - x \approx \delta_x$ (als de perturbaties $\Delta_\mathcal{B}, \Delta_\mathcal{C}$ en δ_b voldoende klein zijn; ga dit na). Merk op dat

$$\frac{||\Delta_\mathcal{A}||}{||\mathcal{A}||} \leq \frac{||\mathcal{B}||\,||\mathcal{C}||}{||\mathcal{A}||}\left(\frac{||\Delta_\mathcal{B}||}{||\mathcal{B}||} + \frac{||\Delta_\mathcal{C}||}{||\mathcal{C}||}\right).$$

De gesplitste vergelijking zal niet slechter gekonditioneerd zijn dan de ongesplitste in geval $||\mathcal{B}||\,||\mathcal{C}|| \approx ||\mathcal{A}||$. In zekere zin is een schatting als $||\mathcal{A}^{-1}\Delta_\mathcal{B}\mathcal{C}|| \leq ||\mathcal{A}^{-1}||\,||\Delta_\mathcal{B}||\,||\mathcal{C}||$ scherp (zie opgave (2.4.2)). We moeten dan ook vrezen dat in geval $||\mathcal{A}|| \ll ||\mathcal{B}||\,||\mathcal{C}||$, de perturbaties op de gesplitste vergelijking een ernstiger effect hebben dan we op grond van het konditie getal van $\mathcal{A}$ verwachten. De opmerkingen in (2.3.4) zijn hier overigens ook van toepassing.

(2.4.2) Opgave. Bewijs de volgende beweringen.
Er zijn perturbaties $\Delta_\mathcal{B}$ en $\Delta_\mathcal{C}$ zodat

$$||\mathcal{B}\Delta_\mathcal{C}|| = ||\mathcal{B}||\,||\Delta_\mathcal{C}|| \neq 0 \qquad \text{en} \qquad ||\mathcal{A}^{-1}\Delta_\mathcal{B}\mathcal{C}|| = ||\mathcal{A}^{-1}||\,||\Delta_\mathcal{B}||\,||\mathcal{C}|| \neq 0.$$

Er is een b en een perturbatie $\Delta_\mathcal{B}$ zodat $||\mathcal{A}^{-1}\Delta_\mathcal{B}\mathcal{C}(x)|| = ||\mathcal{A}^{-1}||\,||\Delta_\mathcal{B}||\,||\mathcal{C}||\,||x|| \neq 0$.

(2.4.3) Voorbeeld. Als $0 < \varepsilon \ll 1$ is $A := \begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \frac{1}{\varepsilon} & 1 \end{bmatrix} \cdot \begin{bmatrix} \varepsilon & 1 \\ 0 & 1-\frac{1}{\varepsilon} \end{bmatrix}$
en $||L||_\infty ||U||_\infty = \varepsilon^{-2} - 1$, terwijl $||A||_\infty = 2$.

De stabiliteit van elementaire bewerkingen

Voor een paar elementaire bewerkingen geven we in deze subparagraaf aan dat een computer uitkomst gezien kan worden als de *exacte* uitkomst van een verstoorde bewerking (we laten de bewijzen aan de lezer over). Dit inzicht zal ons ondermeer in staat stellen de fout in de computer uitkomst te analyseren met behulp van de ontwikkelde perturbatie theorie uit de vorige paragraaf.

We verwaarlozen overal ξ^2-termen; zo schrijven we bijvoorbeeld ξA in plaats van $\xi \tilde{A}$ als $\tilde{A} - A = \mathcal{O}(\xi)$, etc..

De argumenten in de rest van dit hoofdstuk zijn korrekt voor zowel $\mathbf{K} = \mathbf{R}$ als $\mathbf{K} = \mathbf{C}$. Wel moet men in geval $\mathbf{K} = \mathbf{C}$ een andere relatieve machine precisie $\bar{\xi}$ hanteren dan in geval $\mathbf{K} = \mathbf{R}$. Een 'complexe flop' duurt ook langer dan een 'reële flop'.

Telkens als u een exact berekende uitkomst is stelt $u^\star$ het door computer berekende resultaat voor. Zo beweert het volgende lemma dat een door de computer uitgerekend inprodukt $(\vec{a}, \vec{b})^\star$ tussen twee vektoren $\vec{a}$ en $\vec{b}$ een exact inprodukt is tussen $\vec{b}$ en een licht gestoorde vector $\vec{a} + \delta_{\vec{a}}$.

(2.4.4) Lemma. *Laat $\vec{a}, \vec{b} \in \mathbf{K}^n$ en $\alpha, \beta \in \mathbf{K}$ exact gerepresenteerd zijn in de computer.*
(a) $(\vec{a}, \vec{b})^\star = (\vec{a} + \delta_{\vec{a}}, \vec{b})$ *voor zekere* $\delta_{\vec{a}} \in \mathbf{K}^n$ *met* $|\delta_{\vec{a}}| \preceq (n+1)\bar{\xi}|\vec{a}\,|$.
(b) $((\alpha - (\vec{a}, \vec{b}))/\beta)^\star = (\alpha - (\vec{a} + \delta_{\vec{a}}, \vec{b}))/(\beta + \delta_\beta)$ *voor zekere* $\delta_\beta \in \mathbf{K}$ *met* $|\delta_\beta| \leq 2\bar{\xi}|\beta|$. □

(2.4.5) Opmerking. Als het inprodukt in dubbele precisie uitgerekend is, is $|\delta_{\vec{a}}| \preceq \bar{\xi}|\vec{a}\,|$.

In dit hoofdstuk kunnen alle schattingen betreffende evaluatie fouten met een faktor $\frac{3}{n}$ of kleiner gereduceerd worden als alle voorkomende inprodukten in dubbele precisie uitgevoerd zijn.

(2.4.6) Lemma. *Zij $L \in \mathbf{L}_n^1$, $U \in \mathbf{U}_n$ niet-singulier $B \in \mathbf{M}_n$, $\vec{b}, \vec{y}, \vec{r} \in \mathbf{K}^n$, allen exact gerepresenteerd in de computer.*
(a) *Als $\vec{x}^\star$ de met voorwaartse eliminatie door de computer berekende uitkomst is van de vergelijking $L\vec{x} = \vec{b}$ dan voldoet $\vec{x}^\star$ exact aan*

$$(L + \Delta_L)\vec{x}^\star = \vec{b} \quad \textit{voor zekere} \quad \Delta_L \in \mathbf{L}_n \quad \textit{met} \quad |\Delta_L| \preceq n\bar{\xi}|L|.$$

(b) *Als $\vec{x}^\star$ de met terugwaartse substitutie door de computer berekende uitkomst is van de vergelijking $U\vec{x} = \vec{y}$ dan*

$$(U + \Delta_U)\vec{x}^\star = \vec{y} \quad \textit{voor zekere} \quad \Delta_U \in \mathbf{U}_n \quad \textit{met} \quad |\Delta_U| \preceq n\bar{\xi}|U|.$$

(c) *Als $\vec{x}^\star$ het matrix-vektor vermenigvuldigings resultaat $\vec{x} = B\vec{r}$ is dan*

$$(B + \Delta_B)\vec{r} = \vec{x}^\star \quad \textit{voor zekere} \quad \Delta_B \in \mathbf{M}_n \quad \textit{met} \quad |\Delta_B| \preceq (n+1)\bar{\xi}|B|.$$

(er komen geen perturbaties op $\vec{b}, \vec{y}$ en $\vec{r}$!) □

(2.4.7) Opgave. (a) Bewijs lemma (2.4.4).
(b) Bewijs lemma (2.4.6) met behulp van (2.4.4).
(c) Ga na dat $\|\Delta_L\|_\infty \leq n\bar{\xi}\,\|L\|_\infty$ en $\|\Delta_L\|_F \leq n\bar{\xi}\,\|L\|_F$. Geldt ook $\|\Delta_L\|_2 \leq n\bar{\xi}\,\|L\|_2$?

De stabiliteit van de LU-dekompositie

In deze subparagraaf koncentreren we ons weer op het Gauss eliminatie proces. We analyseren de fouten die bij het uitvoeren van de volgende oplosmethode optreden. We meten de fouten in de $\|\cdot\|_\infty$. Omdat we met speciale matrices, driehoeks matrices, te maken hebben stellen we hier norm-schattingen als in (2.4.1) zolang mogelijk uit.

(2.4.8) De oplosmethode. Zij $A \in \mathbf{M}_n$ niet-singulier, zij $\vec{b} \in \mathbf{K}^n$.
$\vec{x} \in \mathbf{K}^n$ is de oplossing van $A\vec{x} = \vec{b}$. De numerieke oplosmethode verloopt in drie stappen.

(0) Equilibreer de rijen door met een geschikte diagonaal matrix D_0 te vermenigvuldigen; $A_0 := D_0 A$, $\vec{b}_0 := D_0\vec{b}$.

(1) Zij $A_0 \in \mathbf{M}_n$. Bepaal $P \in \mathbf{P}_n$, $L \in \mathbf{L}_n^1$, $U \in \mathbf{U}_n$ met behulp van het Gauss eliminatie proces met pivotten zodat $PA_0 = LU$.
Het numeriek proces levert ons $\tilde{P} \in \mathbf{P}_n$, $\tilde{L} \in \mathbf{L}_n^1$, $\tilde{U} \in \mathbf{U}_n$.

(2) Zij $\tilde{L} \in \mathbf{L}_n^1$, $\tilde{U} \in \mathbf{U}_n$, $\vec{b}_1 \in \mathbf{K}^n$ ($\vec{b}_1 = \tilde{P}D_0\vec{b}$). Los $\vec{y}$ op uit $\tilde{L}\vec{y} = \vec{b}_1$ via voorwaartse eliminatie. Los $\tilde{x}_1$ op uit $\tilde{U}\tilde{x}_1 = \vec{y}$ via terugwaartse substitutie.
Numerieke uitvoering levert ons ($\tilde{y}$ en) $\tilde{x}$.

We analyseren de fouten in iedere stap. We gebruiken nergens dat de rijen van A_0 geëquilibreerd zijn; wat betreft de foutenanalyse mag D_0 iedere matrix zijn (b.v. ook $D_0 = I$).

We beginnen bij stap (2). We kijken eerst naar de fouten die optreden bij het numeriek oplossen van het gesplitste stelsel *gegeven* de $\tilde{L}$ en $\tilde{U}$. Aan de hand van de inzichten die we hiermee opdoen zouden we al kunnen bepalen hoe we m.b.v. het Gauss proces aan een zo stabiel mogelijk oplosproces kunnen komen.

(2.4.9) Lemma. *Beschouw stap (2) van bovenstaand proces.*
De berekende oplossingen $\tilde{y}$ en $\tilde{x}$ voldoen exact aan

$$(\tilde{L}+\Delta_L)\tilde{y}=\vec{b}_1 \quad \textit{en} \quad (\tilde{U}+\Delta_U)\tilde{x}=\tilde{y} \quad \text{(geen storing op } \vec{b}_1!\text{)},$$

voor zekere Δ_L *met* $|\Delta_L|\preceq n\bar{\xi}\cdot|\tilde{L}|$ *en* Δ_U *met* $|\Delta_U|\preceq n\bar{\xi}\cdot|\tilde{U}|$.
De berekende oplossing $\tilde{x}$ voldoet exact aan

$$(\tilde{L}\tilde{U}+\Delta_A^{(2)})\tilde{x}=\vec{b}_1 \quad \textit{voor zekere} \quad \Delta_A^{(2)} \quad \textit{met} \quad |\Delta_A^{(2)}|\preceq 2n\bar{\xi}\cdot|\tilde{L}|\,|\tilde{U}|. \tag{9}$$

Bewijs. Met $\Delta_A^{(2)}:=\tilde{L}\Delta_U+\Delta_L\tilde{U}$ volgt (9) eenvoudig uit (2.4.6). □

Het is niet verrassend dat $\tilde{L}\tilde{U}$ in (9) een rol speelt; in schattingen voor $\|\tilde{x}-\tilde{x}_1\|_\infty$ zullen we, door '$\mathcal{O}(\xi^2)$-termen' te verwaarlozen, $\tilde{L}\tilde{U}$ wel mogen vervangen door $\tilde{P}A_0$. De term $|\tilde{L}|\,|\tilde{U}|$ baart ons zorgen. Schatting (9) werp twee vragen op.
(1) Kan $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty$ veel groter zijn dan $\|\tilde{L}\tilde{U}\|_\infty$?
(2) (zo ja) Is de schatting (9) scherp?
Het volgende voorbeeld laat zien dat we beide vragen bevestigend kunnen beantwoorden. Omdat we hier met driehoeks matrices werken zou men kunnen hopen dat een schatting met een term als $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty$ betrouwbaarder is dan een schatting met een term als $\|\tilde{L}\|_\infty\,\|\tilde{U}\|_\infty$ (zie (2.4.1)). Het voorbeeldje laat zien dat deze hoop gerechtvaardigd is, maar onderschrijft toch ook wel de bewering dat een situatie waarin $\|\tilde{L}\tilde{U}\|_\infty \ll \|\tilde{L}\|_\infty\,\|\tilde{U}\|_\infty$ onze extra aandacht verdient.

(2.4.10) Voorbeeld. Laat $0<\varepsilon\ll 1$. Beschouw $A:=\begin{bmatrix}\varepsilon & 1\\ -1 & 0\end{bmatrix}=\begin{bmatrix}1 & 0\\ -\frac{1}{\varepsilon} & 1\end{bmatrix}\cdot\begin{bmatrix}\varepsilon & 1\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}$.
Met $\vec{b}:=(1,1)^T$ is $\vec{x}=(-1,1+\varepsilon)^T$. $\vec{y}=(1,1+\frac{1}{\varepsilon})^T$ is de oplossing van $L\vec{y}=\vec{b}$. Met $\varepsilon=10^{-8}$ vinden we, in 8 cijfers rekenend, dat $\tilde{y}=(1,\frac{1}{\varepsilon})^T$. $\tilde{y}$ voldoet exact aan $(L+\Delta_L)\tilde{y}=\vec{b}$ met $\Delta_L=\begin{bmatrix}0 & 0\\ 1 & 0\end{bmatrix}$. Verder is $\tilde{x}=(0,1)^T$ de oplossing van $U\tilde{x}=\tilde{y}$.
Dus $\|\tilde{x}-\vec{x}\|_\infty=1=\frac{1}{1+\varepsilon}\|\vec{x}\|_\infty$. De schatting $|\Delta_A|\preceq 2{\cdot}2{\cdot}\varepsilon{\cdot}|\tilde{L}|\,|\tilde{U}|=4\varepsilon\begin{bmatrix}\varepsilon & 1\\ 1-\varepsilon & \frac{2}{\varepsilon}\end{bmatrix}$ levert $\|\,|\Delta_A|\,\|_\infty\le 4\varepsilon(\frac{2}{\varepsilon}+1)$ en $\|\tilde{x}-\vec{x}\|_\infty\le\|A^{-1}\|_\infty\,\|\,|\Delta_A|\,\|_\infty\,\|\vec{x}\|_\infty\approx 8\|\vec{x}\|_\infty$, hetgeen op een faktor 8 na korrekt is. Hier zijn faktoren $\varepsilon,\frac{1}{\varepsilon},\ldots$ van belang: we mogen stellen dat de schatting in (9) scherp is.
Merk op dat $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty\approx\frac{2}{\varepsilon}\|A\|_\infty$, terwijl $\|\tilde{L}\|_\infty\,\|\tilde{U}\|_\infty\approx\frac{2}{\varepsilon^2}\|A\|_\infty$

Voordat we de fouten in stap (1) van het oplosproces analyseren merken we een drietal zaken op.
(i) Het Gauss eliminatie proces met pivotten uitgevoerd op A_0 is numeriek equivalent met het Gauss eliminatie proces zonder pivotten uitgevoerd op $\tilde{P}A_0$, mits het pivotten geschied als beschreven door $\tilde{P}$ ($\tilde{P}=\tilde{P}_{n-1}\cdots\tilde{P}_1$, een produkt van opeenvolgende verwisselingen); vooraf pivotten of onderweg pivotten levert ook numeriek dezelfde $\tilde{L}$ en $\tilde{U}$.
(ii) In het numeriek uitvoeren van stap (1), voeren we mischien niet dezelfde verwisselingen uit die we bij exact rekenen uitgevoerd zouden hebben (we vinden $\tilde{P}$ in plaats van P), maar het verwisselen zelf gebeurt foutloos. Bovendien interesseert het ons niet of we een LU-dekompositie vinden voor PA_0 of voor $\tilde{P}A_0$, zolang we maar een LU-dekompositie vinden.
(iii) We zijn *niet* geïnteresseerd in de fout $\tilde{L}-L$ of $\tilde{U}-U$, de fout in de afzonderlijke komponenten, maar in de fout $\tilde{L}\tilde{U}-A_1$ met $A_1:=\tilde{P}A_0$.

Het volgende lemma doet een uitspraak over Gauss eliminatie *zonder* pivotten toegepast op een willekeurige matrix A_1 die een LU-dekompositie heeft (en al in de computer zit). Wij passen het lemma toe met $A_1 = \tilde{P}A_0$ (zie opmerking (i) hierboven).

(2.4.11) Lemma. *Zij $A_1 \in \mathbf{M}_n$ een matrix die exact gerepresenteerd is in de computer. Als het Gauss eliminatie proces zonder pivotten $\tilde{L}$ en $\tilde{U}$ levert, dan*

$$|\tilde{L}\tilde{U} - A_1| \preceq 3n\bar{\xi}\cdot(|A_1| + |\tilde{L}|\,|\tilde{U}|).$$

(Zie stap (1) van het proces (2.4.8).)

•*Bewijs.*

Schrijf $A_1 = \begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix}$ voor geschilkte $\alpha \in \mathbf{K}, v, w \in \mathbf{K}^{n-1}, B \in \mathbf{M}_{n-1}$.

Dan is $\begin{bmatrix} \alpha & w^T \\ v & B \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ m & I \end{bmatrix}\begin{bmatrix} \alpha & w^T \\ 0 & A^{(2)} \end{bmatrix}$, met $m := \frac{1}{\alpha}v, A^{(2)} := B - mw^T$, de eerste stap van het Gauss eliminatie proces (ga dat na). We vinden $m^\star$ in plaats van m en $A^{(2)\star}$ in plaats van $A^{(2)}$. Volgens opgave (2.4.12) zijn er $f \in \mathbf{K}^{n-1}$ en $F \in \mathbf{M}_{n-1}$ zodat

$$m^\star = m + f \text{ met } |f| \preceq \bar{\xi}|m| \quad \text{en} \quad A^{(2)\star} = A^{(2)} + F \quad \text{met } |F| \preceq 3\bar{\xi}|A^{(2)}| + \bar{\xi}|B|.$$

We passen de stelling toe op $A^{(2)\star}$ (induktie hypothese). Dus voor zekere $\tilde{L}_2 \in \mathbf{L}^1_{n-1}$, $\tilde{U}_2 \in \mathbf{U}_{n-1}$ en $E \in \mathbf{M}_{n-1}$ geldt

$$A^{(2)\star} = \tilde{L}_2\tilde{U}_2 + E \quad \text{met} \quad |E| \preceq 3(n-1)\bar{\xi}(|A^{(2)}| + |\tilde{L}_2|\,|\tilde{U}_2|).$$

Dan $\tilde{L} = \begin{bmatrix} 1 & 0 \\ m^\star & \tilde{L}_2 \end{bmatrix}$ en $\tilde{U} = \begin{bmatrix} \alpha & w^T \\ 0 & \tilde{U}_2 \end{bmatrix}$. Uitwerken van $\tilde{L}\tilde{U} - A_1$ levert

$$\tilde{L}\tilde{U} - A_1 = \begin{bmatrix} 0 & 0 \\ \alpha f & F - E - fw^T \end{bmatrix} \quad \text{en} \quad |\tilde{L}|\,|\tilde{U}| = \begin{bmatrix} |\alpha| & |w|^T \\ |v| & |\tilde{L}_2|\,|\tilde{U}_2| + |m|\,|w|^T \end{bmatrix}.$$

Merk op dat $|\alpha f| \preceq \bar{\xi}|\alpha m| = \bar{\xi}|v|$. Het lemma volgt dus als we kunnen aantonen dat

$$|F| + |E| + |fw^T| \preceq 3n\bar{\xi}(|B| + |\tilde{L}_2|\,|\tilde{U}_2| + |m|\,|w|^T).$$

Merk op dat

$$\text{(i)} \quad |A^{(2)}| \preceq |B| + |m|\,|w|^T \quad \text{en} \quad \text{(ii)} \quad |A^{(2)}| \preceq |\tilde{L}_2|\,|\tilde{U}_2| + \mathcal{O}(\xi).$$

Volgens onze induktie aanname en (i) is het blijkbaar voldoende als

$$|F| + |fw^T| \preceq 3\bar{\xi}(|B| + |\tilde{L}_2|\,|\tilde{U}_2| + |m|\,|w|^T).$$

Welnu, (i) en vervolgens (ii) levert
$|F| = 3\bar{\xi}|A^{(2)}| + \bar{\xi}|B| \preceq \bar{\xi}|A^{(2)}| + 3\bar{\xi}|B| + 2\bar{\xi}|m|\,|w|^T \preceq \bar{\xi}|\tilde{L}_2|\,|\tilde{U}_2| + 3\bar{\xi}|B| + 2\bar{\xi}|m|\,|w|^T$. □

(2.4.12) Opgave. (a) Als $\vec{v} \in \mathbf{K}^m$, $\alpha \in \mathbf{K}$ en we berekenen $\vec{m} = \frac{1}{\alpha}\vec{v}$ met de computer dan de berekende $\vec{m}^\star$ gelijk aan $\vec{m}^\star = \vec{m} + \vec{f}$ voor zekere $\vec{f} \in \mathbf{K}^m$, met $|\vec{f}\,| \preceq \bar{\xi}|\vec{m}|$. Ga dit na.
(b) Als $\alpha, \beta, \gamma \in \mathbf{K}$ en $\mu := \alpha - (\gamma\beta)$ wordt in de computer uitgerekend dan is het berekend resultaat $\mu^\star$ gelijk aan $\mu^\star = \mu(1 + 2\xi) + \alpha\xi = \mu + 2\xi\mu + \xi\alpha$. Ga dit na.
(c) Als $G \in \mathbf{M}_m$,$\vec{m}, \vec{f}, \vec{w} \in \mathbf{K}^m$ met $|\vec{f}\,| \preceq \bar{\xi}|\vec{m}|$ en we berekenen $H = G - (\vec{m} + \vec{f}\,)(\vec{w})^T$ dan geldt voor de berekende $H^\star = H' + E$ met $H' = G - \vec{m}\vec{w}^{\,T}$ en $|E| \preceq 3\bar{\xi}\,|H| + \bar{\xi}|G|$. Bewijs dit.

Vatten we bovenstaande lemma's samen dan vinden we de volgende stelling die (vrijwel) volledig, op '$\mathcal{O}(\xi^2)$-termen' na, de fouten in het proces (2.4.8) beschrijft; rij equilibreren gaat (vrijwel) foutloos.

(2.4.13) Stelling. *De numerieke oplossing* $\tilde{x}$ *berekend volgens* (2.4.8) *voldoet exact aan*

$$(A_0 - \Delta_A)\tilde{x} = \vec{b}_0 \quad \textit{voor zekere} \quad \Delta_A \in \mathbf{M}_n \quad \textit{met} \quad |\Delta_A| \preceq n\bar{\xi}\cdot(3|\tilde{P}A_0| + 5|\tilde{L}|\,|\tilde{U}|). \qquad (10)$$

Verwaarlozen we weer $\mathcal{O}(\xi^2)$-termen, dan levert de stelling ons de volgende schatting voor de fout in $\tilde{x}$.

$$\frac{\|\tilde{x} - \vec{x}\,\|_\infty}{\|\vec{x}\,\|_\infty} \leq \mathcal{C}'(D_0A)\frac{\|\Delta_A\|_\infty}{\|A_0\|_m} \quad \text{met} \quad \frac{\|\Delta_A\|_\infty}{\|A_0\|_m} \leq 8n\bar{\xi}\frac{\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty}{\|A_0\|_m}. \qquad (11)$$

Op 'n faktor 8 en $\mathcal{O}(\xi^2)$-termen na is deze schatting scherp. Een term als $\mathcal{C}'(D_0A)8n\bar{\xi}$ in de schatting verbaasd ons niet en is min of meer de term die we in de ideale situatie als uiteindeljke bovenschatting verwachten. Gezien voorbeeld (2.4.10) kan, als we geen extra maatregelen nemen, $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty/\|A_0\|_m$ groot zijn. Tot nu toe hebben we nog geen strategie afgesproken voor het pivotten: we spoorden zomaar een pivot $\neq 0$ op. Door met wat meer beleid de pivots te kiezen kunnen we mogelijk voorkomen dat $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty/\|A_0\|_m$ groot wordt. Bedenken we dat $\|\,|\tilde{L}|\,|\tilde{U}|\,\|_\infty \leq \|\tilde{L}\|_\infty\,\|\tilde{U}\|_\infty \leq n\|\tilde{L}\|_\infty\,\|\tilde{U}\|_m$ dan lijkt partial pivotten (kiezen van het grootste element in de kolom die aan de beurt is om geveegd te worden) een goede strategie: dan is $\|\tilde{L}\|_\infty \leq n$. Als we in stap (1) partial pivotten dan geeft schatting (11) dat

$$\frac{\|\Delta_A\|_\infty}{\|A_0\|_m} \leq 8n^3\bar{\xi}\frac{\|\tilde{U}\|_m}{\|A_0\|_m}. \qquad (12)$$

In een eliminatie stap van het Gauss proces trekken we een skalair veelvoud van een rij af van een andere rij. Als deze skalair zeer groot is zal een rij-vektor uiterst klein zijn ten opzichte van de rij-vektor die eraf getrokken wordt. De oorspronkelijke rij-vektor zal in het verschil moeilijk terug te vinden zijn. Ook vanuit deze heuristische visie is het te begrijpen dat de skalaire veelvouden uit stabiliteits overwegingen kleine skalaire moeten betreffen.

Merk op dat $\|\tilde{U}\|_\infty \leq \|\tilde{L}^{-1}\|_\infty\,\|\tilde{L}\tilde{U}\|_\infty \leq 2^{n-1}n\|A_0\|_m$ (zie opgave (2.2.7)). Uit het volgend voorbeeld blijkt dat deze schatting scherp is.

(2.4.14) Voorbeeld.

$$\text{Met} \quad A = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 1 \\ -1 & 1 & 0 & \dots & 0 & 1 \\ -1 & -1 & 1 & \dots & 0 & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -1 & -1 & -1 & \dots & 1 & 1 \\ -1 & -1 & -1 & \dots & -1 & 1 \end{bmatrix} \quad \text{is} \quad U = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & 2 \\ 0 & 0 & \dots & 0 & 4 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 2^{n-2} \\ 0 & 0 & \dots & 0 & 2^{n-1} \end{bmatrix}.$$

In de praktijk blijkt gelukkig dat vrijwel altijd $\|\tilde{U}\|_m \leq 4\|A_0\|_m$ (ook als de rijen niet geëquilbreerd zijn: dus ook als $D_0 = I$). Dit praktijk fenomeen staat bekend als *het wonder van Wilkinson* en verklaart het grote succes van de methode in (2.4.8).

Wilkinson heeft de methode rond 1960 uitgebreid bestudeerd; hij heeft het zojuist vermelde fenomeen ontdekt. Zijn stabiliteits analyse levert het volgende enigszinds betere resultaat op (hoewel niet kwalitatief beter).

(2.4.15) Stelling van Wilkinson. *De numerieke oplossing* $\tilde{x}$ *berekend volgens* (2.4.8) *met partial pivotten voldoet exact aan*

$$(A_0 - \Delta_A)\tilde{x} = \vec{b}_0 \quad \textit{voor zekere} \quad \Delta_A \in \mathbf{M}_n$$

$$\textit{met} \quad \frac{\|\Delta_A\|_\infty}{\|A_0\|_m} \leq 1.01(n^3 + 3n^2)\bar{\xi}g(A_0), \quad \textit{waarbij} \quad g(A_0) := max_k\frac{\|A_k\|_m}{\|A_0\|_m}. \qquad (13)$$

Het resultaat is weer onafhankelijk van het feit of de rijen geëquilibreerd zijn.

De A_k zijn de matrices die we tijdens het Gauss eliminatie proces (stap (2); zie ook (2.2.9)) tegenkomen. De waarde g(A_0) kan al rekenend bijgehouden worden. g(A_0) kan 2^{n-1} zijn maar is hoogst zelden meer dan 4 (wonder van Wilkinson).

(2.4.16) Opgave. Bewijs dat, in geval we partial pivotten, en exact rekenen $\|A_0\|_m = \|\tilde{P}A_0\|_m = \|A_1\|_m \leq n\|U\|_m$. Laat nu op analoge wijze zien dat $\|A_k\|_m \leq (n+1-k)\|U\|_m$. Dus $\mathrm{g}(A_0) \leq n\frac{\|U\|_m}{\|A_0\|_m} \leq n\mathrm{g}(A_0)$.

In vrijwel 100% van de paktijk situaties wordt de geschaalde fout in $\tilde{x}$ begrensd door $\leq 5n^3\bar{\xi}\mathcal{C}'(A_0)$ (zelfs door $5\bar{\xi}\mathcal{C}'(A_0)$; de factor n^3 is gewoonlijk overdreven). Toch kan de fout ongewenst groot worden. Als het geëquilibreerd stelsel stabiel is (als $n^3\bar{\xi}\mathcal{C}'(A_0)$ aanvaarbaar groot is) dan kan men, door de faktor $\mathrm{g}(A_0)$ in de gaten te houden, zien of de methode een grote fout in $\tilde{x}$ veroorzaakt. Men hoeft pas over te stappen op een duurdere stabiele(re) methode als $\mathrm{g}(A_0)$ groot blijkt te zijn. Omdat dit zelden voorkomt is deze aanpak een stuk goedkoper dan die waarin men altijd met de duurdere stabiele methode werkt.
We geven twee van zulke stabiele methoden ter overweging.

(i) Gauss eliminatie met volledig pivotten.
(ii) Via een QU-dekompositie (ook wel QR-dekompositie genoemd).

Vaak kan men, ook als de grote fout niet veroorzaakt is door de oplosmethode, het resultaat verbeteren met behulp van de gemaakte (eventueel grove) dekompositie door

(iii) naverfijnen.

(2.4.17) Volledig pivotten. In het Gauss eliminatie proces zou men in iedere k-de pivot slag naast twee rijen ook twee kolommen (de k-de en 'n j-de met $j \geq k$) kunnen verwisselen. Dit resulteert in $P_1, P_2 \in \mathbf{P}_n$, $L \in \mathbf{L}_n^1$ en $U \in \mathbf{U}_n$ zodat $P_1AP_2 = LU$. Men spreekt van *volledig pivotten* als men op deze manier als nieuwe pivot telkens het in absolute waarde grootste element uit het nog te bewerken beneden blok kiest (in de k-de stap uit $A_k^{(2)}$; zie (2.2.9)). De foutenanalyse boven is ook van toepassing op deze modifikatie van het Gauss eliminatie proces. Wilkinson bewees dat de $\mathrm{g}(A_0)$ hier exact rekenend begrensd wordt door $\sqrt{n}(2\sqrt{3}\sqrt[3]{4}\ldots\sqrt[n-1]{n})$, dus door een grootheid die als funktie van n langzaam groeit. In de praktijk is de $\mathrm{g}(A_0)$ ook hier weer vrijwel altijd kleiner dan 4.

Volledig pivotten vergt echter $\approx \frac{1}{3}n^3$ 'max' operaties en is daarom nogal duur (partial pivotten vergt slechts $\approx \frac{1}{2}n^2$ 'max' operaties). Als er behoeft is aan stabielere oplosmethode, lost men in de praktijk de vergelijking liever op met een QU-dekompositie.

Als $A = QU$ met $Q \in \mathbf{Q}_n$ en $U \in \mathbf{U}_n$ dan is $\|Q\|_2\|U\|_2 = \|A\|_2$ zodat de gesplitste vergelijking in ieder geval even stabiel is als de oorspronkelijke (zie (2.4.1)). In de volgende paragraaf bespreken we uitgebreid een methode om zo'n QU-dekompositie te konstrueren. We kunnen de dekompositie in $\approx \frac{2}{3}n^3$ flop maken.

De naverfijnings aanpak bespreken we hieronder, in de volgende subparagraaf.

Naverfijnen

(2.4.18) Zij $A_0 \in \mathbf{M}_n$ en $\vec{b}_0 \in \mathbf{K}^n$ (stap (0) van (2.4.8)).
Laat $\tilde{P}$, $\tilde{L}$ en $\tilde{U}$ geproduceerd zijn uit A_0 (stap (1) van (2.4.8)); $A_1 = \tilde{P}A_0$, $\vec{b}_1 = \tilde{P}\vec{b}_0$.
Beschouw het volgende *naverfijnings* algorithme (zie ook p. 6).

Algorithme.

$\vec{x}_0$:=0, k:=0;
(a) r_k:=$\vec{b}_1 - A_1\vec{x}_k$;
(b) **Bereken de** $\vec{f}_{k+1}$ **die voldoet aan** $\tilde{L}\tilde{U}\vec{f}_{k+1} = r_k$
volgens stap (2) **van** (2.4.8);
(c) **Bereken** $\vec{x}_{k+1}$:=$\vec{x}_k + \vec{f}_{k+1}$;
(d) k:=$k+1$ **en ga naar** (a);

$\vec{x}_1 = \vec{f}_1$ is precies de $\tilde{x}$ uit stap (2) van (2.4.8). $\vec{f}_2$ poogt de fout $\vec{x} - \tilde{x}$ te benaderen (als $A_1 \approx \tilde{L}\tilde{U}$ dan $A_1\vec{f}_2 \approx \tilde{L}\tilde{U}\vec{f}_2 = \vec{b}_1 - A_1\tilde{x} = A_1(\vec{x} - \tilde{x})$, $\tilde{x} + \vec{f}_2 \approx \vec{x}$, etc.).

Volgens de stelling van Wilkinson (2.4.15) is er voor iedere $k \in \mathbf{N}_0$ een $E_k \in \mathbf{M}_n$ zodat

$$(A_1 + E_k)\vec{f}_k = r_k \quad \text{en} \quad \|E_k\|_\infty \leq \|A_1\|_m 1.01(n^3 + 3n^2)\bar{\xi}\mathrm{g}(A_1).$$

Als $\sigma := \|A_1^{-1}E_k\|_\infty \leq \mathcal{C}'(A_1)1.01(n^3+3n^2)\bar{\xi}\mathrm{g}\,(A_1) < \frac{1}{2}$ dan volgt uit bewering (1.3.8) dat

$$\|\vec{x}_k - \vec{x}\,\|_\infty \leq (\frac{\sigma}{1-\sigma})^k \|\vec{x}\,\|_\infty$$

(pas (1.3.8) toe met $A = A_1$, $K_k = A_1 + E_k$ en $R_k = E_k$,
$\rho = \|K_1^{-1}R_k\|_\infty = \|(I + A_1^{-1}E_k)^{-1}A_1^{-1}E_k\|_\infty \leq \frac{\sigma}{1-\sigma}$).

De uitspraken veronderstellen dat de stappen (a) en (c) exact worden uitgevoerd. Natuurlijk ten onrechte (zie (2.4.6)). Toch vinden we wel een verbetering als

(i) Gauss eliminatie veel instabieler is dan de vergelijking zelf ($\mathrm{g}\,(A_1)$ groot) of als

(ii) (a) en (c) in bijvoorbeeld dubbele precisie worden uitgevoerd.

Iedere stap in het naverfijnings algorithme is relatief goedkoop (de $\tilde{L}$ en $\tilde{U}$ zijn immers al à $\frac{1}{3}n^3$ flop berekend). Stap (b) kost n^2 flop en de stappen (a) en (c) kosten n^2 (dubbele precisie) flop.

Voor de naverfijning moet men naast de (D_0) $\tilde{L}$, $\tilde{U}$ en $\tilde{P}$ ook nog een kopie van de origenele matrix A (of A_1, berekend in dubbele precisie) bewaren.

2.5 De QR-dekompositie

De methode

De Gram–Schmidt procedure levert voor iedere niet-singuliere matrix $A \in \mathbf{M}_n$ matrices $Q \in \mathbf{Q}_n$ en $U \in \mathbf{U}_n$ zodat $A = QU$. Met behulp van zo'n dekompositie kunnen we gemakkelijk de vergelijking $A\vec{x} = \vec{b}$ op lossen en de gesplitste vergelijking is even stabiel als de ongesplitste (zie (2.4.1)). De Gram–Schmidt procedure is welliswaar konstruktief maar is duur (n^3 flop, 'n andere methode kost $\frac{2}{3}n^3$ flop) en, wat veel erger is, hij is instabiel.

(2.5.1) De Gram–Schmidt procedure en de instabiliteit ervan. Als $u, v, w, \ldots \in \mathbf{X}$ dan produceert Gram–Schmidt een orthonormaal stelsel $q_1, q_2, q_3, \ldots$ als volgt:

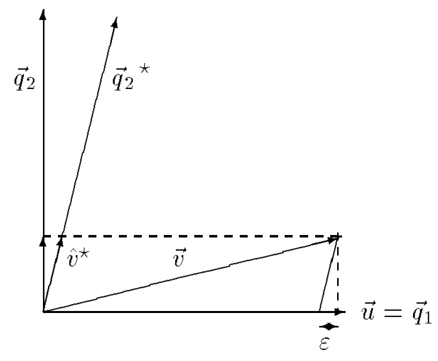
(i) $\hat{u} = u,$ $\qquad q_1 = \frac{1}{\|\hat{u}\|_2}\hat{u};$

(ii) $\hat{v} = v - (v, q_1)q_1,$ $\qquad q_2 = \frac{1}{\|\hat{v}\|_2}\hat{v}.$

(iii) $\hat{w} = w - (w, q_1)q_1 - (w, q_2)q_2,$ $\qquad q_3 = \frac{1}{\|\hat{w}\|_2}\hat{w}.$

$\vdots$

Beschouw nu $\vec{u}, \vec{v} \in \mathbf{R}^n$ met $\|\vec{u}\,\|_2 = \|\vec{v}\,\|_2 = 1$ en onderlinge hoek φ. Stel dat in de berekening van $(\vec{v}, \vec{q}_1) = (\vec{v}, \vec{u})$ een relatieve fout ε gemaakt wordt en dat de berekening verder foutloos verloopt. Dan is $\hat{v}^\star = \hat{v} - \varepsilon\cos\varphi\,\vec{u}$ en $|(\vec{q}_2{}^\star, \vec{q}_1)| = \frac{1}{\|\hat{v}^\star\|_2}|\varepsilon\cos\varphi| = \gamma(1+\gamma^2)^{-\frac{1}{2}}$ met $\gamma = |\varepsilon\coth\varphi|$. Als de hoek φ klein is zal de orthogonaliteit ernstig vertoort raken. Door verder te werken alsof $\vec{q}_1, \vec{q}_2{}^\star$ orthonormaal zijn wordt het effect in elke volgende stap van de Gram–Schmidt procedure er alleen maar erger van.



Er zijn modifikaties van de Gram–Schmidt procedure die deze gevoeligheid voor evaluatiefouten niet hebben ('modified Gram–Schmidt', 'twice is enough'). Deze modifikaties worden voor diverse problemen gebruikt. Maar voor onze situatie wordt de oplosmethode er alleen maar duurder door. Wij konstrueren de QU-dekompositie met een proces dat geheel analoog verloopt aan het eliminatie proces van Gauss; we vervangen de Gauss transformaties door geschikte unitaire spiegelingen uit de volgende stelling.

In de numerieke literatuur spreekt men overigens gewoonlijk niet over een QU-dekompositie maar over een QR-*dekompositie* (die men echter niet moet verwarren met het nog te behandelen 'QR-algorithme' ter berekening van eigenwaarden en eigenvektoren).

(2.5.2) Stelling. *Zij $\mathbf{X}$ een inprodukt ruimte. Zij $v \in \mathbf{X}\backslash\{0\}$.*
Beschouw de afbeelding $H_v \in \mathcal{L}(\mathbf{X})$ die gegeven is door

$$H_v(x) = x - \frac{2}{(v,v)}(x,v)v \quad voor \quad x \in \mathbf{X}.$$

H_v is een spiegeling *door het hypervlak $\perp$ op v: $H_v(\alpha v + b) = -\alpha v + b \quad (\alpha \in \mathbf{K},\ b \in \mathbf{X}, b \perp v)$.*
Verder is $H_v = H_v^*$, $H_v^* H_v = H_v H_v = I$.

Als $a, b \in \mathbf{X}$ zodat $\|a\| = \|b\| \neq 0$, $\mathrm{Im}(a,b) = 0$ en $v := a - b$ dan $H_v(a) = b$.

Bewijs. Alle beweringen volgen eenvoudig door invullen. We bewijzen de laatste bewering.

Merk op dat $\|a\| = \|b\|$ impliceert $a + b \perp a - b = v$. Dus
$$2H_v(a) = H_v((a-b) + (a+b)) = -(a-b) + (a+b) = 2b. \qquad \square$$

Met de spiegelingen hierboven kunnen we dus unitair een gegeven vektor afbeelden op een andere gegeven vektor. De volgende spiegelingen zijn voor ons belangrijk.

(2.5.3) Householder transformaties. Zij $\vec{a} \in \mathbf{K}^n$, $\vec{a} = \sum_{j \geq k} \alpha_j e_j$.
Zij τ het *teken* van α_k: $\tau := 1$ als $\alpha_k = 0$ en $\tau := \frac{\alpha_k}{|\alpha_k|}$ anders. Neem

$$\vec{v} := \vec{a} + \zeta \|\vec{a}\|_2 e_k = (\alpha_k + \zeta\|\vec{a}\|_2)e_k + \sum_{j>k} \alpha_j e_j \quad \text{met} \quad \zeta \in \{-\tau, +\tau\}.$$

Dan $H_{\vec{v}}(\vec{a}) = -\zeta\|\vec{a}\|_2 e_k$; $H_{\vec{v}}$ spiegelt $\vec{a}$ op een skalair veelvoud van een standaard basis vektor. In de numerieke wiskunde wordt dit soort spiegelingen *Householder transformaties* genoemd.

Om stabiliteits redenen kiest men $\zeta = \tau$: $\|\vec{a}\|_2$ en α_k zullen in de praktijk met (evaluatie)fouten belast zijn. Als deze getallen ongeveer even groot zijn kan het aftrekken een zéér grote relatieve fout tot gevolg hebben. Dit is precies het fenomeen dat we gezien hebben in het Gram–Schmidt proces hierboven. De som van twee ongeveer even grote getallen met fouten belast heeft een relatieve fout die ongeveer even groot is als de relatieve fouten in de oorspronkelijke getallen. Met Householder transformaties zitten we in in de komfortabele situaties dat we kunnen kiezen tussen optellen of aftrekken. Bij het Gram–Schmidt proces konden we dat niet.

Met $\mu := 2/(\vec{v},\vec{v})$ is $H_{\vec{v}} = I - \mu\vec{v}\vec{v}^*$. Het berekenen van $\vec{v}$ en μ kost $(n-k+3)$ flop en 1 worteltrekkingen

$$(\text{met} \quad \gamma := \sum_{j>k} |\alpha_j|^2, \quad \text{en} \quad \nu := \alpha_k + \zeta\sqrt{\gamma + |\alpha_k|^2} \quad \text{is} \quad \mu = 2/(v,v) = 2/(\gamma + |\nu|^2) \quad).$$

Als $\vec{v}$ en μ berekend zijn kost $H_{\vec{v}}(\vec{b}\,) = \vec{b} - \mu\vec{v}\vec{v}^*\vec{b} = \vec{b} - (\mu(\vec{b}\,,\vec{v}))\vec{v}$ berekenen $2(n-k+1)+1$ flop.
(De Gauss transformatie $I - \tilde{\mu}\tilde{v}e_k^*$ met $\tilde{v} = \sum_{j>k} \alpha_j e_j$ en $\tilde{\mu} = \frac{1}{\alpha_k}$ kost, om $\tilde{v}$ en $\tilde{\mu}$ op te stellen, bijna niets en om toe te passen op een vektor $n-k$ flop.)

Men werkt ook wel eens met $H_{\vec{v}} = I - \vec{w}\vec{y}^*$ waarbij $\vec{w} := \frac{\mu}{\nu}\vec{v}$ en $\vec{y} = \frac{1}{\nu}\vec{v}$ met ν zodat de eerste niet nul koördinaat van $\vec{y}$ de waarde 1 heeft. Men moet nu twee vektoren $\vec{y}$ en $\vec{w}$ onthouden in plaats van een vektor $\vec{v}$ en een skalair μ. Maar met name in geval maar een paar koördinaten van $\vec{v}$ niet nul zijn en men wil $H_{\vec{v}}\vec{b}$ voor vele $\vec{b}$'s berekenen, kan men hierdoor relatief veel rekenwerk besparen (ga na hoe dit precies werkt).

(2.5.4) De (eliminatie) methode van Householder. Zij $A = (a_{ij}) \in \mathbf{M}_n$ niet-singulier. We schrijven A als $A = QU$ middels de *methode van Householder*.

Berekenen rekursief
$$\begin{array}{ll} & A_1 := (a_{ij}^{(1)}) := A, \\ \vec{v}_1, & A_2 = (a_{ij}^{(2)}) := H_{\vec{v}_1}A, \\ \vec{v}_2, & A_3 = (a_{ij}^{(3)}) := H_{\vec{v}_2}A_2, \\ \vdots & \end{array}$$
, als volgt in $n-1$ achtereenvolgende stappen:

(1) Met $\tilde{a}_1 := (a_{11}, a_{21}, \ldots, a_{n1})^T$, $\tau_1 :=$ teken(a_{11}) is $\vec{v}_1 := \tilde{a}_1 + \tau_1\|\tilde{a}_1\|_2 e_1$ en vinden we

$$A_2 := H_{\vec{v}_1}A = (I - \mu_1\vec{v}_1\vec{v}_1{}^*)A = \left[\begin{array}{c|ccc} -\tau_1\|\tilde{a}_1\|_2 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} \\ \hline 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{array}\right] \quad \text{met } \mu_1 := \frac{2}{(\vec{v}_1, \vec{v}_1)}$$

(dus $a_{i1}^{(2)} = 0$ voor $i = 2, \ldots, n$ en $a_{11}^{(2)} = -\tau_1\|Ae_1\|_2$).

(2) Met $\tilde{a}_2 := (0, a_{22}^{(2)}, a_{32}^{(2)}, \ldots, a_{n2}^{(2)})^T$, $\tau_2 :=$ teken$(a_{22}^{(2)})$ is $\vec{v}_2 := \tilde{a}_2 + \tau_2\|\tilde{a}_2\|_2 e_2$ en vinden we

$$A_3 := H_{\vec{v}_2}A_2 = (I - \mu_2\vec{v}_2\vec{v}_2{}^*)A_2 = \left[\begin{array}{c|c|ccc} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} \\ \hline 0 & -\tau_2\|\tilde{a}_2\|_2 & a_{23}^{(3)} & \cdots & a_{2n}^{(3)} \\ \hline 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} \end{array}\right] \quad \text{met } \mu_2 := \frac{2}{(\vec{v}_2, \vec{v}_2)}.$$

(3), ..., (n-1) We vinden de boven driehoeks matrix $U := A_n = H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2}H_{\vec{v}_1}A$.
Met $Q := (H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2}H_{\vec{v}_1})^{-1} = H_{\vec{v}_1}H_{\vec{v}_2} \cdots H_{\vec{v}_{n-1}}$ is Q unitair en $A = QU$.

(2.5.5) De oplosmethode met Householder transformaties. Zij nu $\vec{b} \in \mathbf{K}^n$.
We lossen, nadat we de householder methode uitgevoerd hebben, de vergelijking $A\vec{x} = \vec{b}$ als volgt op.
We lossen $Q\vec{y} = \vec{b}$, om efficiëntie redenen (zie (2.5.7) en (2.5.8)), op met de *gefaktorizeerde* vorm van Q: $\vec{y} = H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2}H_{\vec{v}_1}\vec{b}$. Vervolgens bepalen we $\vec{x}$ uit $U\vec{x} = \vec{y}$ door terugwaartse substitutie.

De efficiëntie

(2.5.6) De kosten van een QU-dekompositie. In de k-de stap van het Householder proces berekenen we $\vec{v}_k$ en de schalings faktor μ_k in $(n - k + 3)$ flop. Om A_{k+1} te bepalen moeten we de Householder transformatie $H_{\vec{v}_k} = I - \mu_k\vec{v}_k\vec{v}_k{}^*$ op $n - k + 1$ vektoren laten werken, hetgeen $(n - k + 1)(2n - 2k + 3)$ flop kost (zie (2.5.3)). We zijn hier telkens $\approx$ een faktor 2 duurder uit dan in de korresponderende Gauss eliminatie stap. Totaal kost een QU-dekompositie dan ook $\approx \frac{2}{3}n^3$ flop (en n worteltrekkingen). We laten de Q in gefaktorizeerde vorm staan (zie (2.5.7) en (2.5.8))!

(2.5.7) Kosten oplosmethode. Om $\vec{y} = H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2}H_{\vec{v}_1}\vec{b}$ te berekenen moeten we $n-1$ Householder transformaties toepassen (van afnemende lengte), hetgeen ons $\approx n^2$ flop kost (zie (2.5.3)). Deze oplosmethode is vrijwel even duur als de matrix-vektor vermenigvuldiging $Q^*\vec{b}$ die we zouden kunnen uitvoeren als we Q expliciet bepaald zouden hebben. Het oplossen van $U\vec{x} = \vec{y}$ met terugwaartse substitutie vergt weer $\frac{1}{2}n(n+1)$ flop. In totaal kost het oplossen van $QU\vec{x} = \vec{b} \approx \frac{3}{2}n^2$ flop, $\approx$50% duurder dan met een LU-dekompositie.

(2.5.8) Praktische Householder proces. Als in het Gauss eliminatie proces kunnen we ook hier weer efficiënt omspringen met de geheugen ruimte. Op de nieuw gekreëerde nul plaatsen van A_{k+1} in de k-de stap kunnen we het niet-triviale deel van de $\vec{v}_k$ en de μ_k bewaren; we hebben in het totaal dan slechts twee rijen in het matrix array extra nodig voor de uiteindelijke U en alle $\vec{v}_k$ en μ_k. Omdat we $Q\vec{y} = \vec{b}$ kunnen oplossen met behulp van de $\vec{v}_k$, zonder de expliciete gedaante van Q, hebben we nooit meer geheugen ruimte nodig.

De stabiliteit

De eerste uitspraak vermelden we zonder bewijs.

(2.5.9) Lemma. *Zij* $\vec{a} = \sum_{j\geq k} \alpha_j e_j \in \mathbf{K}^n$ *en* $A \in \mathbf{M}_{n\times m}$.
Bereken $\vec{v}$ *en* μ *als in* (2.5.3). $\vec{v}^\star$ *en* $\mu^\star$ *zijn de computer resultaten. Schrijf* $\epsilon := (n+3)\bar{\xi}$.
Met $H^\star_{\vec{v}} := I - \mu^\star\vec{v}^\star(\vec{v}^\star)^*$ *is*
(a) $\quad \|H^\star_{\vec{v}} - H_{\vec{v}}\|_2 \leq 2(n-k+2)\bar{\xi} \leq 2\epsilon$ *en*
(b) $\quad \|H^\star_{\vec{v}}A - (H^\star_{\vec{v}}A)^\star\|_F \leq (n-k+3)\bar{\xi}\,\|A\|_F \leq \epsilon\|A\|_F$. □

(2.5.10) Opmerking. Bij de foutenanalyse werkt de Frobenius norm prettiger dan de 2-norm ondermeer omdat $\|\,|A|\,\|_F = \|A\|_F$ terwijl de schatting $\|\,|A|\,\|_2 \leq \;(\|A\|_F \leq)\; \sqrt{n}\|A\|_2$ scherp is.

Het berekende resultaat van een aantal Householder transformaties toegepast op een matrix A kan gezien worden als de bijna-unitair getransformeerde van een matrix uit de buurt van A.

(2.5.11) Lemma. *Zij* $A \in \mathbf{M}_{n\times m}$. *Zij* $\vec{a}_1, \ldots, \vec{a}_l \in \mathbf{K}^n$ *met bijbehorende exacte Housholder transformaties* $H_{\vec{v}_1}, \ldots, H_{\vec{v}_l}$ *en computer berekende* $H^\star_{\vec{v}_1}, \ldots, H^\star_{\vec{v}_l}$. $\epsilon := (n+3)\bar{\xi}$.
Zij $U_l := Q_l A$ *met* $Q_l := H_{\vec{v}_l}\ldots H_{\vec{v}_1}$ *en zij* $U^\star_l := (H^\star_{\vec{v}_l}U^\star_{l-1})^\star = \ldots$ *de computerberekende* U_l.
Dan $\quad U^\star_l = \tilde{Q}_l(A+E)$ *voor zekere* $E \in \mathbf{M}_{n\times m}$ *met* $\|E\|_F \leq (l+1)\epsilon\|A\|_F$,
waarbij $\tilde{Q}_l := H^\star_{\vec{v}_l}, \ldots, H^\star_{\vec{v}_1}$ *en* $\|Q_l - \tilde{Q}_l\|_2 \leq 2l\epsilon$
($\tilde{Q}_l$ is het exacte produkt van de berekende transformaties!).

• *Bewijs.* Schrijf $\delta := \epsilon$ en $\varepsilon := 2l\epsilon$.
We laten de schatting $\|Q_l - \tilde{Q}_l\|_2 \leq \varepsilon$ en $\|Q_l^{-1} - \tilde{Q}_l^{-1}\|_2 \leq \varepsilon(1+\varepsilon)$ aan de lezer over.

Volgens (2.5.9) is $U^\star_l = H^\star_{\vec{v}_l}U^\star_{l-1} + \Delta_l$ voor zekere $\Delta_l \in \mathbf{M}_{n\times m}$ met $\|\Delta_l\|_F \leq \delta\|U^\star_{l-1}\|_F$.

Stel $U^\star_{l-1} = \tilde{Q}_{l-1}(A+E')$ met $\|E'\|_F \leq l\delta(1+\varepsilon)^3(1+\delta(1+\varepsilon)^3)^{l-1}\|A\|_F$ (ind. hyp.).
Dan $\|U^\star_{l-1}\|_F \leq (1+\varepsilon)(\|A\|_F + \|E'\|_F)$ en $U^\star_l - \tilde{Q}_lA = H^\star_{\vec{v}_l}(U^\star_{l-1} - \tilde{Q}_{l-1}A) + \Delta_l = \tilde{Q}_lE' + \Delta_l$.
Neem nu $E := E' + \tilde{Q}_l^{-1}\Delta_l = E' + Q_l^*\Delta_l + (\tilde{Q}_l^{-1} - Q_l^*)\Delta_l$. Dan $U^\star_l = \tilde{Q}_l(A+E)$ en

$$\begin{aligned}
\|E\|_F &\leq \|E'\|_F + \|\Delta_l\|_F + \|\tilde{Q}_l^{-1} - Q_l^*\|_2\,\|\Delta_l\|_F \\
&\leq \|E'\|_F + \delta\|U^\star_{l-1}\|_F + (1+\varepsilon)\varepsilon\delta\|U^\star_{l-1}\|_F \\
&\leq \|E'\|_F + \delta(1+\varepsilon)^2\|U^\star_{l-1}\|_F \leq (1+\delta(1+\varepsilon)^3)\|E'\|_F + \delta(1+\varepsilon)^3\|A\|_F \\
&\leq (l+1)\delta(1+\varepsilon)^3(1+\delta(1+\varepsilon)^3)^l\|A\|_F \approx (l+1)\delta\|A\|_F \quad .
\end{aligned}$$
□

(2.5.12) • Opgave. Bewijs $\|Q_l - \tilde{Q}_l\|_2 \leq \varepsilon$ en $\|Q_l^{-1} - \tilde{Q}_l^{-1}\|_2 \leq \varepsilon(1+\varepsilon)$ met Q_l als in (2.5.11) en $\varepsilon := 2l\epsilon$.

(2.5.13) Opmerkingen. (a) We gebruiken alleen een produkt van Householder transformaties in gefaktorizeerde vorm. We zijn dus niet geïnteresseerd in een fout in een berekende $Q^\star_l$.
• (b) In de uitspraak hebben we $(1+\epsilon(1+2l\epsilon)^3)^l \approx 1$ verondersteld. Voor de l, n en $\bar{\xi}$ waar we mee te maken hebben is dit redelijk (ga dit na).

(2.5.14) Stelling. *Zij* $A \in \mathbf{M}_n$ *en* $\vec{b} \in \mathbf{K}^n$. $\vec{x} \in \mathbf{K}^n$ *is de oplossing van* $\quad A\vec{x} = \vec{b}$.
Als $\tilde{x}$ *de berekende oplossing is geproduceerd met de Householder methode als beschreven in*

(2.5.4) *en* (2.5.5) *dan voldoet* $\tilde{x}$ *voor zekere* $\Delta_A \in \mathbf{M}_n$ *en* $\delta_{\vec{b}} \in \mathbf{K}^n$ *exact aan*

$$(A+\Delta_A)\tilde{x} = \vec{b} + \delta_b \quad met \quad \|\Delta_A\|_F \leq (n+1)(n+3)\bar{\xi}\,\|A\|_F \quad en \quad \|\delta_b\|_2 \leq n(n+3)\bar{\xi}\,\|\vec{b}\,\|_2.$$

• *Bewijs.* Beschouw (2.5.4).
Voor $k = 1, \ldots n-1$, laat $H_{\vec{v}_k}$ de *exacte* Householder transformatie zijn uit de k-de stap bij de *berekende* kolom vektoren $\tilde{a}_k^\star$.

Schrijf $Q := H_{\vec{v}_{n-1}} \ldots H_{\vec{v}_1}$, $U := QA$, $\vec{y} := Q\vec{b}$, $\tilde{Q} := H^\star_{\vec{v}_{n-1}}, \ldots, H^\star_{\vec{v}_1}$ (als in (2.5.11) met $l = n-1$).
Volgens (2.5.11) is $U^\star = \tilde{Q}(A+E)$ met $\|E\|_F \leq n\epsilon\|A\|_F$
en omdat $\mathbf{K}^n = \mathbf{M}_{n\times 1}$ is ook volgens (2.5.11) $\vec{y}^{\,\star} = \tilde{Q}(\vec{b}+\delta_b)$ met $\|\delta_b\|_F = \|\delta_b\|_2 \leq n\epsilon\|\vec{b}\,\|_2$.

Los x' op uit $U^\star x' = \vec{y}^{\,\star}$. Dan voldoet $\tilde{x} := x'^\star$ aan (zie (2.4.6))

$$(U^\star + \Delta_U)\tilde{x} = \vec{y}^{\,\star} \quad \text{met} \quad \|\Delta_U\|_F \leq n\bar{\xi}\,\|U^\star\|_F \approx n\bar{\xi}\,\|U\|_F \leq \epsilon\|A\|_F.$$

Zij $\Delta_A := \tilde{Q}^{-1}\Delta_U + E$. Dan $(A+\Delta_A)\tilde{x} = \vec{b} + \delta_b$ en
$\|\Delta_A\|_F \leq \|Q^* - \tilde{Q}^{-1}\|_2\|\Delta_U\|_F + \|\Delta_U\|_F + \|E\|_F \leq (n+1)\epsilon\|A\|_F$. □

(2.5.15) Gevolg. $\dfrac{\|\tilde{x} - \vec{x}\,\|_2}{\|\vec{x}\,\|_2} \leq \mathcal{C}_2(A)(n+1)(n+3)\bar{\xi}(\sqrt{n}+1) \approx \mathcal{C}_2(A)n^2\sqrt{n}\,\bar{\xi}.$

Bewijs. Omdat $\|A\|_F \leq \sqrt{n}\,\|A\|_2$ volgt dit onmiddellijk uit de stelling. □

De schatting in het gevolg is aardig wat we er in de ideale situatie van verwachten. We mogen de Householder methode stabiel noemen.

(2.5.16) Opmerking. In deze opmerking willen we nogeens expliciet wijzen op een subtiliteit in het bewijs van stelling (2.5.14).

We vergeleken $\tilde{Q}$ met een unitaire matrix Q. Q is echter *niet* de matrix die uit het Householder proces gerold zou zijn als we het hele proces exact uitgevoerd zouden hebben. Q wordt opgebouwd uit exacte Householder transformaties en is daarom unitair, ook al behoren de exacte transformaties bij berekende kolom vektoren. Het feit dat $\tilde{Q}$ lijkt op een of andere unitaire Q (en het feit dat $\tilde{Q}^{-1}U^\star$ lijkt op A) is van belang; we hoeven niet te weten of $\tilde{Q}$ lijkt op de exacte unitaire transformatie.

(2.5.17) Opmerking. (a) Door in het bewijs van (2.5.14) te gebruiken dat de $\vec{v}_1, \ldots, \vec{v}_{n-1}$ een afnemende lengte hebben kunnen diverse schattingen met een faktor $\approx \frac{1}{2}$ verbeterd worden.

Voor het oplossen van vergelijkingen gebruikt men, zoals we al opgemerkt hebben, vrijwel altijd de Gauss eliminatie methode. De Householder methode speelt echter een zeer grote rol bij het oplossen van kleinste kwadraten problemen. We bespreken deze problemen in het volgend hoofdstuk.

3 Het kleinste kwadraten probleem $Ax = b$

Een belangrijk probleem in de lineaire algebra is het "zo goed mogelijk oplossen" van een *overbepaald* stelsel $A\vec{x} = \vec{b}$, dit is een stelsel waarin $A \in \mathbf{M}_{m\times n}$ met $m > n$ (meer vergelijkingen dan onbekenden; een dergelijk stelsel zal in het algemeen geen echte oplossing hebben). Dit probleem ontmoet men bijvoorbeeld bij het uitwerken van de resultaten van een experimenteel onderzoek, waar men voor een aantal onbekende grootheden $\alpha_1, \ldots, \alpha_n$ een groot aantal lineaire relaties $\sum_j a_{ij}\alpha_j = \beta_i$, $i = 1, \ldots, m$ kan bepalen waarvan met name de koëfficiënten β_i niet erg nauwkeurig zijn.
In NUMERIEKE WISKUNDE B benaderden we een funktie $f \in C[-1, +1]$ zo goed mogelijk door een polynoom van graad $n-1$. (De koëfficiënten van het polynoom zijn dan de onbekenden α_j en de β_i zijn funktiewaarden in geschikte steunpunten).
Onder een "zo goed mogelijk oplossing" verstaan we die $\vec{x} \in \mathbf{K}^n$ waarvoor $\|\vec{b} - A\vec{x}\|$ minimaal is ten opzichte van een of andere norm.

3.1 Het probleem

(3.1.1) Lemma [Hoofdstelling van de approximatie theorie].
Zij $\mathbf{Y}$ *een genormeerde ruimte,* $\mathbf{X}_A$ *een eindig dimensionale lineaire deelruimte van* $\mathbf{Y}$ *en* $b \in \mathbf{Y}$*. Dan is er een* $b' \in \mathbf{X}_A$ *zodat* $\|b - b'\|_{\mathbf{Y}} = \inf\{\|b - z\|_{\mathbf{Y}} \,\big|\, z \in \mathbf{X}_A\}$.

Bewijs. Omdat $\mathbf{X}_A$ eindig dimensionaal is kan $\mathbf{X}_A$ voorzien van de $\|\cdot\|_{\mathbf{Y}}$ norm in zekere zin (zie (1.2.3)) geïdentificeerd worden met een $\mathbf{K}^n$ voorzien van $\|\cdot\|_\infty$ norm.

De kontinue afbeelding $z \rightsquigarrow \|b - z\|_{\mathbf{Y}}$ van $\mathbf{X}_A$ naar $[0, \infty)$ neemt op de kompakte (gesloten en begrensde) verzameling $\mathbf{B} := \{z \in \mathbf{X}_A \,\big|\, \|z\|_{\mathbf{Y}} \le 2\|b\|_{\mathbf{Y}}\}$ zijn minimum aan, zeg in b'. Als $z \in \mathbf{X}_A \backslash \mathbf{B}$ dan $\|b - z\|_{\mathbf{Y}} \ge \|z\|_{\mathbf{Y}} - \|b\|_{\mathbf{Y}} > 2\|b\|_{\mathbf{Y}} - \|b\|_{\mathbf{Y}} = \|b - 0\|_{\mathbf{Y}} \ge \|b - b'\|_{\mathbf{Y}}$. Blijkbaar $\|b - b'\|_{\mathbf{Y}} \le \|b - z\|_{\mathbf{Y}}$ voor alle $z \in \mathbf{X}_A$, en het lemma is hiermee bewezen. □

(3.1.2) Stelling. $\mathbf{X}$ *is eindig dimensionaal en laat* $\mathbf{Y}$ *genormeerd zijn.*
Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$ *injektief. Zij* $b \in \mathbf{Y}$.
Er is een $x \in \mathbf{X}$ *waarvoor het residu* $b - \mathcal{A}(x)$ *geminimalizeerd wordt:*

$$\|b - \mathcal{A}(x)\|_{\mathbf{Y}} = \inf\{\|b - \mathcal{A}(z)\|_{\mathbf{Y}} \,\big|\, z \in \mathbf{X}\} = \min_z \|b - \mathcal{A}(z)\|_{\mathbf{Y}} : \tag{14}$$

x *is een zogenaamde* minimale residu oplossing *van* $\mathcal{A}(x) = b$.

Bewijs. Pas (3.1.1) toe met $\mathbf{X}_A = \mathcal{A}(\mathbf{X})$. Omdat $\mathcal{A}$ een bijektie is van $\mathbf{X}$ op $\mathcal{A}(\mathbf{X})$ is er een $x \in \mathbf{X}$ zodat $\mathcal{A}(x) = b'$. □

Het hangt van de norm op $\mathbf{Y}$ af of de oplossing van probleem (14) uniek is.

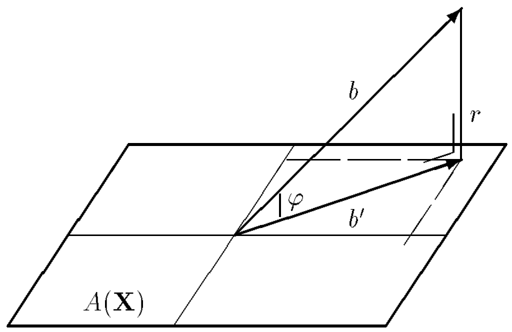
(3.1.3) Voorbeeld. Als $\vec{b} = (1, 0)^T \in \mathbf{R}^2$ en $A \in \mathbf{M}_{2\times 1}$ met $A := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ dan $\|\vec{b} - A\vec{x}\|_\infty = \min_z \|\vec{b} - A\vec{z}\|_\infty$ voor iedere $\vec{x} = (\alpha) \in \mathbf{R}^1$ met $|\alpha| \le 1$.

Minimalizerings problemen ten opzichte van bijvoorbeeld de sup-norm zijn voor de numerieke wiskunde interessant; zo waren we in NUMERIEKE WISKUNDE B geïnteresseerd in the situatie waarin $\mathbf{X} = \mathcal{P}_{n-1}$, de ruimte van polynomen van graad ten hoogste $n-1$, $\mathbf{Y} = C(J)$ voorzien van de sup-norm, met $J = [-1, +1]$ of $J = \{-1 + 2\frac{j-1}{m-1} \,\big|\, j = 1, \ldots m\}$ en $\mathcal{A}(p) = p$.

Hier beperken we ons ondermeer vanwege de volgende stelling echter tot situatie waarin $\mathbf{Y} = \mathbf{K}^m$ en voorzien is van de standaard 2-norm.

(3.1.4) Lemma. *Als in de hoofdstelling van de approximatie theorie de norm op* $\mathbf{Y}$ *afstamt van een inprodukt dan is de* b' *uniek en wordt* b' *gekarakterizeerd door de eigenschap*
$b' \in \mathbf{X}_A$ *en* $b - b' \perp \mathbf{X}_A$ (dus, $b - b' \perp \mathbf{X}_A$ en als $y \in \mathbf{X}_A$, $b - y \perp \mathbf{X}_A$ dan $y = b'$).
b' is de *orthonormale projektie* van b op $\mathbf{X}_A$.

Bewijs. Zie (2.3.16) en (2.3.18). □



Figuur 1: De kleinste kwadraten benadering.

(3.1.5) Opgave. Schrijf $\mathbf{Z} := \mathbf{X}_A$.
Definieer $P_\mathbf{Z} : \mathbf{Y} \to \mathbf{Y}$ door $P_\mathbf{Z}(b) = b'$: $P_\mathbf{Z}(b) \in \mathbf{Z}$, $b - P_\mathbf{Z}(b) \perp \mathbf{Z}$ voor iedere $b \in \mathbf{Y}$.
(a) Bewijs dat $P_\mathbf{Z} \in \mathcal{L}(\mathbf{Y})$, $P_\mathbf{Z}^* = P_\mathbf{Z} = P_\mathbf{Z} P_\mathbf{Z}$ en $P_\mathbf{Z}(\mathbf{Y}) = \mathbf{Z}$.

Zij $P \in \mathcal{L}(\mathbf{Y})$ een *projektor* (d.w.z. $P^2 = P$) met $P(\mathbf{Y}) = \mathbf{Z}$. P is een *orthonormale* projektor als Kern $P \perp \mathrm{Im}\, P$.
(b) Bewijs P is een orthonormale projektor $\Leftrightarrow P^* = P \Leftrightarrow P = P_\mathbf{Z}$.

(3.1.6) Stelling. $\mathbf{X}$ *is eindig dimensionaal en laat* $\mathbf{Y}$ *voorzien zijn van een inprodukt.*
Zij $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$ *injektief. Zij* $b \in \mathbf{Y}$.
Er is precies één $x \in \mathbf{X}$ *waarvoor het residu* $r := b - \mathcal{A}(x)$ *geminimalizeerd wordt.*
x *is de minimale residu oplossing van* $\mathcal{A}(x) = b \quad \Leftrightarrow \quad b - \mathcal{A}(x) \perp \mathcal{A}(\mathbf{X})$.

Bewijs. Pas bovenstaand lemma toe, weer met $\mathbf{X}_A = \mathcal{A}(\mathbf{X})$. □

(3.1.7) Bij minimale residu problemen ten opzichte van inprodukt-normen speelt in de theorie en in oplosprocessen (zie (3.4.8)) de singuliere waarden dekompositie vaak een grote rol.

Stel dat $\mathbf{X}$ ook voorzien is van een inprodukt. Dan zijn er een orthonormale basis $p_1, \ldots p_n$ van $\mathbf{X}$, een orthonormaal stelsel $q_1, \ldots, q_n$ in $\mathbf{Y}$ (dat tot een orthonormale basis aangevuld kan worden als $\dim\mathbf{Y} < \infty$) en skaleiren $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_n \geq 0$ zodat $\mathcal{A}(p_j) = \sigma_j q_j$ voor $j = 1, \ldots n$.

$\mathcal{A}$ is injektief $\Leftrightarrow \sigma_n > 0$.

Als $b = \sum_{j=1}^n \beta_j q_j + r \in \mathbf{Y}$ met $r \perp q_j$ $(j = 1, \ldots, n)$ dan is $x := \sum_{j=1}^n \frac{\beta_j}{\sigma_j} p_j$ de minimale residu oplossing van $\mathcal{A}(x) = b$ (ga dit na).

(3.1.8) Notatie. Als $\mathbf{X}$ en $\mathbf{Y}$ beide voorzien zijn van een inprodukt en $\mathcal{A} \in \mathcal{L}(\mathbf{X}, \mathbf{Y})$ dan is er precies één $\mathcal{B} \in \mathcal{L}(\mathbf{Y}, \mathbf{X})$ waarvoor $(\mathcal{A}(x), y)_\mathbf{Y} = (x, \mathcal{B}(y))_\mathbf{X}$ voor alle $x \in \mathbf{X}$ en $y \in \mathbf{Y}$. We schrijven $\mathcal{A}^*$ in plaats van $\mathcal{B}$: $(\mathcal{A}(x), y)_\mathbf{Y} = (x, \mathcal{A}^*(y))_\mathbf{X}$ voor alle $x \in \mathbf{X}$ en $y \in \mathbf{Y}$.
Als $\dim\mathbf{X} = n$, $\dim\mathbf{Y} = m$ en $A = (a_{ij}) \in \mathbf{M}_{m\times n}$ is de matrix van $\mathcal{A}$ ten opzichte van een orthonormale basis in $\mathbf{X}$ en een orthonormale basis in $\mathbf{Y}$ dan is $A^* = (b_{ij}) \in \mathbf{M}_{n\times m}$ met $b_{ij} = \bar{a}_{ji}$ de matrix van $\mathcal{A}^*$ ten opzichte van dezelfde bases.

$\mathcal{A}^*\mathcal{A} \in \mathcal{L}(\mathbf{X})$. Als $\mathcal{A}$ injektief is dan is $\mathcal{A}^*\mathcal{A}$ dat ook
(als $\mathcal{A}^*\mathcal{A}(x) = 0$ dan $0 = (\mathcal{A}^*\mathcal{A}(x), x)_\mathbf{X} = \|\mathcal{A}(x)\|_\mathbf{X}$ dus $\mathcal{A}(x) = 0$ en $x = 0$).

(3.1.9) Opgave. (a) Bewijs dat $\|\mathcal{A}\| = \|\mathcal{A}^*\| = \sqrt{\|\mathcal{A}^*\mathcal{A}\|} = \sigma_1$.
(Let op: de normen zijn allemaal verschillend gedefinieerd; we zouden eigenlijk $\|\mathcal{A}\|_{\mathbf{Y}\leftarrow\mathbf{X}}$ moeten schrijven in plaats van $\|\mathcal{A}\|$, etc.)
(b) Stel $\mathbf{X} \subset \mathbf{Y}$ en schrijf $\mathcal{B}_1 = \mathcal{A}^*$. Stel dat $\mathcal{A}(\mathbf{X}) \subset \mathbf{X}$.
Dan heeft $\mathcal{A}$ ook een geadjungeerde als afbeelding van $\mathbf{X}$ naar $\mathbf{X}$. Noteer deze met $\mathcal{B}_2$.
Bewijs dat $\mathcal{B}_1 = \mathcal{B}_2 P_\mathbf{X}$.

(3.1.10) Als $\mathbf{Y} = \mathbf{K}^m$ en voorzien is van het standaard inprodukt noemen we de minimale residu oplossing *de kleinste kwadraten oplossing.*

We bespreken deze situatie met $\mathbf{X} = \mathbf{K}^n$, $m \geq n$ in dit hoofdstuk.
Als $A \in \mathbf{M}_{m\times n}$ rang n heeft (A is injektief) dan is $A^*A \in \mathbf{M}_n$ niet-singulier.

(3.1.11) Stelling. *Zij $A \in \mathbf{M}_{m\times n}$ met $m \geq n$ en zij $\vec{b} \in \mathbf{K}^m$.*
(a) *Er zijn een $Q \in \mathbf{M}_{m\times n}$ en een $R \in \mathbf{M}_{n\times n}$ met $Q^*Q = I$, R boven driehoeks zodat $A = QR$.*
(b) *Voor een $\vec{x} \in \mathbf{K}^n$ zijn de volgende vier eigenschappen equivalent.*
 (i) *$\vec{x}$ is de kleinste kwadraten oplossing van $A\vec{x} = \vec{b}$.*
 (ii) *$A^*\vec{r} = 0$ met $\vec{r} = \vec{b} - A\vec{x}$.*
 (iii) *$\vec{x}$ voldoet aan de zgn.* normaalvergelijking (van Gauss) $A^*A\vec{x} = A^*\vec{b}$.
 (iv) *$\vec{x}$ voldoet aan $R\vec{x} = Q^*\vec{b}$.*

Bewijs. (a) Pas het Gram-Schmidt proces toe op het stelsel $Ae_1, \ldots, Ae_n$. We vinden dan een orthonormaal stelsel $\vec{q}_1, \ldots, \vec{q}_n$. Met $Q \in \mathbf{M}_{m\times n}$ zodat $Qe_j = \vec{q}_j$ is $Q^*Q = I$ en $A = QR$ met R boven driehoek in $\mathbf{M}_n$. (Zie ook (3.3.2) en (3.3.3).)
(b) De equivalentie van (i), (ii) en (iii) volgt onmiddellijk uit (3.1.6) en de definite van A^* in (3.1.8).
Omdat $\vec{q}_1, \ldots, \vec{q}_n$ een orthonormale basis is van $A(\mathbf{K}^n)$ kunnen we de orthonormale projektie $\vec{b}\,'$ van $\vec{b}$ op $A(\mathbf{K}^n)$ beschrijven: $\vec{b}\,' = \sum_{j=1}^n (\vec{b}, \vec{q}_j)\vec{q}_j$. Merk op dat $\vec{b}\,' = QQ^*\vec{b}$. Dus volgens (3.1.6) hebben we '(i) $\Leftrightarrow A\vec{x} = QQ^*\vec{b}$ $(= \vec{b}\,')$'. Omdat $Q^*Q = I$ is '$A\vec{x} = QQ^*\vec{b} \Leftrightarrow R\vec{x} = Q^*\vec{b}$' en het bewijs is geleverd. □

3.2 De stabiliteit van het probleem

In de stabiliteits uitspraken speelt het konditie-getal van de afbeelding $\mathcal{A}$ weer een rol. Omdat $\mathcal{A}$ niet inverteerbaar hoeft te zijn, moeten we de definitie opnieuw bezien.

(3.2.1) Definitie. Beschouw de situatie in (3.1.7).

$$\mathcal{C}(\mathcal{A}) := \sup \left\{ \frac{||\mathcal{A}(x)||_{\mathbf{Y}}}{||\mathcal{A}(y)||_{\mathbf{Y}}} \,\Big|\, x, y \in \mathbf{X}, ||x||_{\mathbf{X}} = ||y||_{\mathbf{X}} \neq 0 \right\} = \frac{\sigma_1}{\sigma_n}.$$

$\mathcal{C}(\mathcal{A})$ is het gebruikelijke konditie-getal van $\mathcal{A}$ als afbeelding van $\mathbf{X}$ op $\mathcal{A}(\mathbf{X})$.

(3.2.2) Opgave. Bewijs dat $\mathcal{C}(\mathcal{A}) = \mathcal{C}(\mathcal{A}^*)$ en $\mathcal{C}(\mathcal{A}^*\mathcal{A}) = \mathcal{C}(\mathcal{A})^2$.

(3.2.3) Stelling. *Beschouw de situatie in* (3.1.6) *met* $\mathbf{X}$ *voorzien van een inprodukt.*
$\tilde{x}$ is de minimale residu oplossing van het geperturbeerde probleem $(\mathcal{A} + \Delta_{\mathcal{A}})(\tilde{x}) = b + \delta_b$.
Met $\delta_x := (\mathcal{A}^*\mathcal{A})^{-1}(\mathcal{A}^*(\delta_b) - \mathcal{A}^*\Delta_{\mathcal{A}}(x) + \Delta_{\mathcal{A}}^*(r))$ is $\tilde{x} - x \approx \delta_x$.

$$||\delta_x||_{\mathbf{X}} \leq \frac{1}{\sigma_n}||\delta_b||_{\mathbf{Y}} + ||\Delta_{\mathcal{A}}|| \left(\frac{1}{\sigma_n}||x||_{\mathbf{X}} + \frac{1}{\sigma_n^2}||r||_{\mathbf{Y}} \right) \tag{15}$$

Laat φ de hoek tussen b en $\mathcal{A}(\mathbf{X})$ zijn.

$$\frac{||\delta_x||_{\mathbf{X}}}{||x||_{\mathbf{X}}} \leq \mathcal{C}(\mathcal{A}) \left(\frac{||\delta b||_{\mathbf{Y}}}{||b||_{\mathbf{Y}}} \frac{1}{\cos\varphi} + \frac{||\Delta_{\mathcal{A}}||}{||\mathcal{A}||} \right) + \mathcal{C}(\mathcal{A}^*\mathcal{A}) \frac{||\Delta_{\mathcal{A}}||}{||\mathcal{A}||} \tan\varphi. \tag{16}$$

Bewijs. Schrijf $\tilde{r} := b + \delta_b - (\mathcal{A} + \Delta_{\mathcal{A}})(\tilde{x})$.
Omdat $\mathcal{A}^*(r) = 0$ zien we dat

$$\begin{aligned} 0 &= (\mathcal{A} + \Delta_{\mathcal{A}})^*(\tilde{r}) \approx \mathcal{A}^*(\tilde{r} - r) + \Delta_{\mathcal{A}}^*(r) \\ &= \mathcal{A}^*(\mathcal{A}(x - \tilde{x}) + \delta_b - \Delta_{\mathcal{A}}(\tilde{x})) + \Delta_{\mathcal{A}}^*(r) \\ &\approx \mathcal{A}^*\mathcal{A}(x - \tilde{x}) + \mathcal{A}^*(\delta_b) - \mathcal{A}^*\Delta_{\mathcal{A}}(x) + \Delta_{\mathcal{A}}^* r, \end{aligned}$$

en we zien dat $\delta_x \approx \tilde{x} - x$.

Kombineren we (3.1.7) en (3.1.9) dan zien we dat $\|(\mathcal{A}^*\mathcal{A})^{-1}\mathcal{A}^*\| \leq \frac{1}{\sigma_n}$ en $\|(\mathcal{A}^*\mathcal{A})^{-1}\| \leq \frac{1}{\sigma_n^2}$, waaruit (15) gemakkelijk volgt.

Omdat $\sigma_1\|x\|_{\mathbf{X}} = \|\mathcal{A}\|\,\|x\|_{\mathbf{X}} \geq \|\mathcal{A}(x)\|_{\mathbf{Y}} = \|b\|_{\mathbf{Y}}\frac{\|b'\|_{\mathbf{Y}}}{\|b\|_{\mathbf{Y}}} = \|b\|_{\mathbf{Y}}\cos\varphi$ en $\frac{\|r\|_{\mathbf{Y}}}{\|x\|_{\mathbf{Y}}} \leq \|\mathcal{A}\|\frac{\|r\|_{\mathbf{Y}}}{\|\mathcal{A}(x)\|_{\mathbf{Y}}} = \|\mathcal{A}\|\tan\varphi$ volgt (16) uit (15). □

(3.2.4) Opmerking. (a) Als in (2.3.3) kan men aantonen dat de schatting (15) met betrekking tot de drie afzonderlijke termen in het rechterlid scherp is in de gebruikelijke zin (zie echter ook (3.2.5)). Evenzo is schatting (16) scherp met betrekking tot de afzonderlijke $\|\Delta_{\mathcal{A}}\|$-termen en ook met betrekking tot de $\|\delta_b\|_{\mathbf{Y}}$-term als iedere b met gegeven hoek φ met $\mathcal{A}(\mathbf{X})$ voor kan komen.
Uiteraard kunnen we hier weer dezelfde soort opmerkingen plaatsen als in (2.3.4).
(b) Als $\varphi = 0$ is $r = 0$ en zijn de schattingen (15) en (16) precies de schattingen die we met het oog op (7) en (8) verwachten.

Als b een niet onaanzienlijke hoek maakt met $\mathcal{A}(\mathbf{X})$ speelt met betrekking tot de perturbaties op $\mathcal{A}$ het *kwadraat* van het konditie-getal van $\mathcal{A}$ een rol. Onder omstandigheden, als $\mathcal{C}(\mathcal{A}) \approx \sqrt{\frac{1}{\xi}}$, is dit zeer vervelend.

(In NUMERIEKE WISKUNDE B hebben we er al op gewezen dat we met grote konditie-getallen te maken kunnen hebben bijvoorbeeld als we een kontinue funktie op $[-1,+1]$ in een kleinste kwadraten zin willen benaderen met polynomen en we hebben de basis in de poly-noomruimte onzorgvuldig gekozen; Chebychev polynomen waren prima maar de machten χ^j niet.)

(3.2.5) Opmerking. In onze opmerking betreffende de scherpte van bijvoorbeeld schatting (15) nemen we aan dat iedere mogelijke (kleine) perturbatie voor kan komen; we nemen bijvoorbeeld aan dat $\delta_b = \varepsilon q_n$ kan zijn, waarbij q_n de linker singuliere vektor uit (3.1.7) is.

In diverse praktische meettechnieken stelt men echter vaak een overbepaald stelsel op omdat men hoopt dat de diverse meetfouten door meerdere metingen min of meer tegen elkaar wegvallen. Het is dus met andere woorden zeer onwaarschijnlijk dat δ_b, als die afstamt van meetfouten, precies in de richting van q_n staat. Waarschijnlijker heeft δ_b in iedere q_j-richting een ongeveer even grote komponent. Dus als $\tilde{x}$ de minimale residu oplossing is van $\mathcal{A}(\tilde{x}) = b + \delta_b$, waarbij δ_b meetfouten zijn, zal

$$\|\delta_x\|_{\mathbf{X}} \leq \sqrt{\frac{1}{m}\sum_{j=1}^{n}\frac{1}{\sigma_j^2}}\;\|\delta_b\|_{\mathbf{Y}}$$

waarschijnlijk de bovengrens op de fout in $\tilde{x}$ zijn[3].

Evaluatie fouten hebben echter een totaal ander karakter dan meetfouten (het heeft geen zin een berekening tien keer uit te voeren in de hoop dat de gemiddelde uitkomst een kleinere fout heeft dan de uitkomst van één berekening). Met betrekking tot numerieke perturbaties zijn de schattingen in (3.2.3) dan ook reëel.

(3.2.6) Als σ_n heel klein is dan is de oplossing zeer gevoelig voor perturbaties (als $\Delta_A = 0$ dan is de schatting $\|\delta_x\|_{\mathbf{X}} \leq \frac{1}{\sigma_n}\|\delta_b\|_{\mathbf{Y}}$ scherp. Ga dit na). Er zijn diverse '*regularizerings*'-technieken om toch zinvolle benaderende oplossingen te krijgen: men kan bijvoorbeeld het volgende minimalizerings probleem oplossen

$$\text{met } \Phi(z) := \|b - \mathcal{A}(z)\|_{\mathbf{Y}}^2 + \gamma^2\|z\|_{\mathbf{X}}^2 \text{ bepaal } x \text{ zodat } \Phi(x) = \min\{\Phi(z) \,\big|\, z \in \mathbf{X}\}, \tag{17}$$

[3] Als de komponenten van δ_b ten opzichte van een orthonormale basis gemiddeld 0 zijn, allen dezelfde variantie σ^2 hebben, en onderling ongecorreleerd zijn dan kan men bewijzen dat

$$E\left((\|\delta_x\|_X)^2\right) \leq \sigma^2\sum_j\frac{1}{\sigma_j^2},$$

waarbij E de waarschijnlijkheidstheoretische verwachtingswaarde is.

waarbij γ een geschikt gewicht is.
De term $||z||_{\mathbf{X}}$ in Φ voorkomt dat de oplossing groot wordt (een grote p_n komponent heeft) ten gevolge van bijvoorbeeld een kleine singuliere waarde σ_n en perturbaties op het probleem.
Als de exacte oplossing van het minimale residu probleem een grote p_n komponent heeft zal die ook door de regularizering 'gedempt' worden. Het gewicht γ dient dus met beleid gekozen te worden. Door te regularizeren benaderen we het probleem door een ander. Naast de perturbaties (ten gevolge van meetfouten of evaluatiefouten) hebben we nu ook nog te maken met approximatiefouten. We moeten er voor zorgen dat de totale fout in de oplossing minimaal is.

We besteden aan dit soort regularizerings problemen in dit college verder geen aandacht meer (behalve in de volgende opgave).

(3.2.7) Opgave. Laat zien dat de oplossing van het geregularizeerd probleem (17) precies de minimale residu oplossing is van

$$\begin{bmatrix} \mathcal{A} \\ \gamma\mathcal{I} \end{bmatrix}(x) = \begin{bmatrix} b \\ 0 \end{bmatrix} \text{ met } \begin{bmatrix} \mathcal{A} \\ \gamma\mathcal{I} \end{bmatrix} \in \mathcal{L}(\mathbf{X}, \mathbf{Y}\times\mathbf{X}). \tag{18}$$

3.3 Een oplosmethode

(3.3.1) Zij $m \geq n$ en $A \in \mathbf{M}_{m\times n}$ een rang n matrix. $\vec{b} \in \mathbf{K}^m$. We wensen numeriek de kleinste kwadraten oplossing $\vec{x} \in \mathbf{K}^m$ van $A\vec{x} = \vec{b}$ te bepalen.

We zouden de normaalvergelijking $A^*A\vec{x} = A^*\vec{b}$ op kunnen lossen met behulp van bijvoorbeeld de Choleski methode.
De Choleski methode is goedkoop ($\approx \frac{1}{6}n^3$ flop tegen $\frac{1}{3}n^3$ voor Gauss en $\frac{2}{3}n^3$ voor Householder). Toch heeft deze aanpak twee nadelen en we raden deze methode dan ook af.

(i) A^*A moet berekend worden à mn^2 flop. Met een rechtstreekse methode zal dit vermeden kunnen worden.

(ii) Lossen we de normaalvergelijking op dan speelt bij fouten het konditie-getal $\mathcal{C}(A^*A) = \mathcal{C}(A)^2$ een rol. Ook als $\vec{b}$ een zeer kleine hoek met $A(\mathbf{K}^n)$ maakt, mogen we nu al slechte numerieke resultaten verwachten als $\mathcal{C}(A) \approx \sqrt{\frac{1}{\xi}}$. Op grond van onze perturbatie uitspraken verwachten we dan pas problemen als $\xi\mathcal{C}(A)$ niet klein is.

Aan de volgende methode die de oplossing van het kleinste kwadraten probleem bepaalt op grond van (3.1.11) (b.iv) kleven deze nadelen niet. Met het oog op onze ervaring in de paragraaf over de QR-dekompositie zal het duidelijk zijn waarom de Q en de R met Householder transformaties berekend worden.

(3.3.2) De methode van Householder. Bepaal (als in (2.5.4)) $\vec{v}_1, \ldots, \vec{v}_{n-1}$ zodat $U := H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2} H_{\vec{v}_1} A$ een boven driehoeks matrix is in $\mathbf{M}_{m\times n}$;
dus $U = \begin{bmatrix} R \\ 0 \end{bmatrix}$ voor zekere boven driehoeks matrix $R \in \mathbf{M}_n$ en nul-matrix 0 in $\mathbf{M}_{(m-n)\times n}$.

Bereken $\vec{y} = H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2} H_{\vec{v}_1} \vec{b}$.
Omdat de householder transformaties unitair zijn is $||\vec{b} - A\vec{z}||_2 = ||\vec{y} - U\vec{z}||_2$ voor iedere $\vec{z} \in \mathbf{K}^n$: $\vec{x}$ is dus precies de kleinste kwadraten oplossing van $\begin{bmatrix} R \\ 0 \end{bmatrix} \vec{x} = \vec{y}$.

Partitioneer $\vec{y}$ als $\vec{y} = \begin{bmatrix} \vec{y}\,' \\ \vec{y}\,'' \end{bmatrix}$ met $\vec{y}\,' \in \mathbf{K}^n$.
Omdat $\begin{bmatrix} \vec{y}\,' \\ 0 \end{bmatrix}$ de orthonormale projektie is van $\vec{y}$ op $U(\mathbf{K}^n)$ is $\vec{x}$ precies de oplossing van $\begin{bmatrix} R \\ 0 \end{bmatrix} \vec{x} = \begin{bmatrix} \vec{y}\,' \\ 0 \end{bmatrix}$ en dus van $R\vec{x} = \vec{y}\,'$.

Bereken $\vec{x}$ uit $R\vec{x} = \vec{y}\,'$ door terugwaartse substitutie.

(3.3.3) Opmerking. (a) Beschouw $P = \begin{bmatrix} I \\ 0 \end{bmatrix} \in \mathbf{M}_{m\times n}$, waarbij I de identiteit in $\mathbf{M}_n$ is. Dan is $U = PR$ en met $\hat{Q} := H_{\vec{v}_1}H_{\vec{v}_2}\cdots H_{\vec{v}_{n-1}}$ is $A = \hat{Q}U = \hat{Q}PR$. $Q := \hat{Q}P$ is orthonormaal: $Q^*Q = I \in \mathbf{M}_n$. We hebben hier een dekompositie $A = QR$ als vereist in (a) van stelling (3.1.11).

In principe hebben we dus hier weer een methode om een orthonormale basis $\vec{q}_1, \ldots, \vec{q}_n$ die uit het Gram-Schmidt proces rolt stabiel te berekenen (op een faktor $\zeta_j \in \mathbf{K}$ met $|\zeta_j| = 1$ is de kolomvektor Qe_j van Q de $\vec{q}_j$ uit het bewijs van (3.1.11); ga dit na). In het algemeen raden we het af om Q expliciet te bepalen (tenzij men de $\vec{q}_j$'s wil hebben): de gefaktorizeerde vorm werkt net zo efficiënt als de samengestelde Q (ga dit na). Het assembleren van de $H_{\vec{v}_j}$'s tot de Q is verspilde moeite.
(b) $A = QR$ impliceert $A^*A = R^*Q^*QR = R^*R$, hetgeen de Choleski dekompositie is van A^*A. Verder is $A^*\vec{b} = R^*Q^*\vec{b}$. We zien hiermee het theoretische verband tussen Choleski's methode toegepast op de normaalvergelijking en de QR-dekompositie aanpak.

(3.3.4) Opgave. Tel het aantal flop dat deze methode vereist en vergelijk de efficiëntie van deze methode met de Choleski methode toegepast op de normaalvergelijking. (Antwoord: U en de $H_{\vec{v}_j}$ bepalen kost $\approx n^2(m - \frac{1}{3}n)$ flop; vervolgens $\vec{x}$ berekenen kost nog eens $\approx 2mn$ flop.)

De volgende algemenere variant van stelling (2.5.14) (die we zonder bewijs geven) vertelt ons dat de methode een vrijwel optimale stabiliteit heeft.

(3.3.5) Stelling. *Als $\tilde{x}$ de berekende oplossing is geproduceerd met de zojuist beschreven methode van Householder dan voldoet $\tilde{x}$ voor zekere $\Delta_A \in \mathbf{M}_{m\times n}$ en $\delta_{\vec{b}} \in \mathbf{K}^m$ exact aan*

$$(A+\Delta_A)\tilde{x} = \vec{b}+\delta_{\vec{b}} \textit{ met } \|\Delta_A\|_2 \leq (6m-3n+41)n\bar{\xi}\,\|A\|_F \textit{ en } \|\delta_b\|_2 \leq (6m-3n+40)n\bar{\xi}\,\|\vec{b}\|_2. \quad \square$$

3.4 • Het minimale norm kleinste kwadraten probleem

Het probleem

(3.4.1) Als in (3.1.6) $\mathcal{A}$ niet injektief is dan is de minimale residu oplossing niet uniek: als x een minimale residu oplssing is dan is ieder element uit $x + \mathrm{Kern}(\mathcal{A})$ dat ook. Als $\mathbf{X}$ ook een inprodukt ruimte is kan men een element opsporen in $x + \mathrm{Kern}(\mathcal{A})$ met minimale norm. Zo'n element is uniek: schrijf $x = x' + x''$ met $x' \in \mathrm{Kern}(\mathcal{A})$ en $x'' \perp \mathrm{Kern}(\mathcal{A})$ (zie (3.1.1) en (3.1.4)) en x'' is **de** zogenaamde *minimale norm minimale residu oplossing* van $\mathcal{A}(x) = b$.

Als in (3.1.7) $\sigma_k > 0$ en $\sigma_{k+1} = 0$ dan is $x'' = \sum_{j=1}^{k} \frac{\beta_j}{\sigma_j} p_j$.

(3.4.2) Stelling. *Zij $A \in \mathbf{M}_{m\times n}$ met $m \geq n$ en rang$(A) = k < n$. Zij $\vec{b} \in \mathbf{K}^m$.*
(a) *Er is precies één $\vec{x} \in \mathbf{K}^n$ waarvoor*

$$\|\vec{x}\|_2 = \min\{\|\vec{z}'\|_2 \,\big|\, \|\vec{b} - A\vec{z}'\|_2 = \min_z \|\vec{b} - A\vec{z}\|_2\}.$$

(b) *Er zijn $P \in \mathbf{P}_n$, $Q \in \mathbf{Q}_m$, $V \in \mathbf{Q}_n$ en $U \in \mathbf{M}_{m\times n}$ zodat*

$$AP = QUV^* \quad \textit{met} \quad U = \begin{bmatrix} \tilde{R} & 0 \\ 0 & 0 \end{bmatrix},$$

waarbij $\tilde{R} \in \mathbf{U}_k$ niet-singuliere is en 0 triviale matrices zijn van geschikte afmetingen.
(c) *Laat $\tilde{x}$ de minimale norm kleinste kwadraten oplossing zijn van $U\tilde{x} = Q^*\vec{b}$.*
Dan $\vec{x} = PV\tilde{x}$ is de minimale norm minimale residu oplossing van $A\vec{x} = \vec{b}$.
(d) *Partioneer $Q^*\vec{b} = \begin{bmatrix} \vec{y}' \\ \vec{y}'' \end{bmatrix}$ met $\vec{y}' \in \mathbf{K}^k$ en $\tilde{x} = \begin{bmatrix} \vec{x}'' \\ \vec{x}' \end{bmatrix}$ met $\vec{x}'' \in \mathbf{K}^k$.*
Dan $\vec{x}' = 0$ en $\tilde{R}\vec{x}'' = \vec{y}'$.

Bewijs. Onderdeel (a) hebben we in (3.4.1) bewezen.

Voor een bewijs van (b) verwijzen we naar (3.4.4) en (3.4.7). Daar geven we een konstruktief bewijs.

(c) volgt eenvoudig uit het feit dat P, Q en V de 2-norm lengte bewarend.

$\begin{bmatrix} \vec{y}' \\ 0 \end{bmatrix}$ is de orthonormale projektie van $Q^*\vec{b}$ op $U(\mathbf{K}^n)$ en $\begin{bmatrix} 0 \\ \vec{x}' \end{bmatrix}$ is de is de orthonormale projektie van $\tilde{x}$ op Kern(U) hetgeen (d) bewijst volgens (3.4.1). $\square$

De stabiliteit van het probleem

(3.4.3) Het 'minimale norm minimale residu probleem' is *slecht gesteld*: de oplossing van dit probleem hangt niet kontinu af van de singuliere waarde σ_n! (Stel $k = n-1$. Laat Δ_A bijvoorbeeld gegeven zijn door $\Delta_A p_j = 0$ voor $j \neq n$ en $\Delta_A p_n = \varepsilon q_n$ met p_j en q_j als in (3.1.7). Laat x'' de minimale norm minimaal residu oplossing zijn van het ongestoord probleem. Dan is volgens (3.4.1) $\tilde{x}'' := x'' + \frac{\beta_n}{\varepsilon} p_n$ de oplossing van het gestoord probleem. En $\|\tilde{x}'' - x''\|_2 \not\to 0$ als $\varepsilon \to 0$ en $\beta_n \neq 0$.) De oplossing is niet stabiel te bepalen, tenzij men dim(kern($\mathcal{A}$)) kent of betrouwbaar kan schatten en die kennis in de berekening gebruiken kan. We komen hierop terug in (3.4.5).

Vaak lost men echter liever een geregularizeerd probleem op. Zo'n probleem is wel goedgesteld maar nu moet men uiteraard weer nagaan hoe goed de oplossing van het benaderend, geregularizeerd probleem het oorspronkelijke probleem oplost (zie ook (3.2.6)).

Een oplosmethode

We bewijzen onderdeel (b) van stelling (3.4.2) in twee stappen. Het bewijs is konstruktief en levert eenvoudig een praktisch algorithme. We laten dit soort details echter nu aan de geïnteresseerde lezer over.

(3.4.4) Lemma. *Zij $A \in \mathbf{M}_{m\times n}$ met $m \geq n$ en* rang$(A) = k < n$.
Dan is er een $P \in \mathbf{P}_n$ en een $Q \in \mathbf{Q}_n$ zodat

$$AP = QU_0 \quad \textit{met} \quad U_0 := \begin{bmatrix} R' & R'' \\ 0 & 0 \end{bmatrix},$$

waarbij $R' \in \mathbf{U}_k$ niet-singuliere is en $R'' \in \mathbf{M}_{k\times(n-k)}$.

Bewijs. We konstrueren de matrices middels de Householder methode met *kolom-pivotten*:

Stel dat we na de $l-1$-de stap de $m \times n$ matrix A_l hebben van de vorm $A_l := \begin{bmatrix} R_l & * \\ 0 & A^{(l)} \end{bmatrix}$, waarbij $R_l \in \mathbf{U}_{l-1}$ niet-singulier is en $*$ een of andere matrix voorstelt.

Verwissel (indien nodig) de l-de kolom met een j-de voor een $j \geq l$ zodat in de nieuwe matrix $\tilde{A}_l := A_l P_l = \begin{bmatrix} R_l & * \\ 0 & \tilde{A}^{(l)} \end{bmatrix}$ geldt $\|\tilde{A}^{(l)} e_l\|_2 \geq \|\tilde{A}^{(l)} e_i\|_2$ voor alle $i \geq l$ (spoor in de matrix $A^{(l)}$ de kolom met de grootste 2-norm op en zet die op de plaats van de eerste kolom van deze matrix; $P_l \in \mathbf{P}_n$ representeert de verwisseling). Bepaal nu $\vec{v}_l$ als in de standaard Householder methode zodat $A_{l+1} = H_{\vec{v}_l}\tilde{A}_l$ van de veronderstelde vorm is.

Dan heeft $U_0 := H_{\vec{v}_{n-1}} \cdots H_{\vec{v}_2} H_{\vec{v}_1} A P_1 P_2 \cdots P_{n-1}$ de aangekondigde vorm. $\square$

(3.4.5) Opmerking. Als rang$(A) = k$ dan is $A^{(k+1)} = 0$. Door evaluatie fouten zal het onderblok dat men berekent niet 0 zijn. Als men weet dat rang$(A) = k$ zal men het gevonden blok vervangen door 0.

Gewoonlijk kent men de rang van A niet en moet men die schatten.

Stel dat $A_l^\star$ de berekende matrix uit de l-de stap is. Dan is $A = Q_l(A_l^\star + \Delta_l)$ met $Q_l \in \mathbf{Q}_n$ en zekere Δ_l met $\|\Delta_l\|_2 \leq \varepsilon_1 \|A\|_2$ en $\varepsilon_1 = \mathcal{O}(\bar{\xi})$ (zie (2.5.11)). Als $A^{(l)\star}$ het onderblok is van $A_l^\star$ dan geldt

$$\sigma_{l+1} \leq \|\Delta_l\|_2 + \|A^{(l)\star}\|_2$$

(we zullen dit soort ongelijkheden pas in het volgende hoofdstuk afleiden). Men schat dat de rang van A nu k is als voor een of andere van te voren vastgestelde ε_2 $(= \mathcal{O}(\bar{\xi}\,))$ geldt $\|A^{(k)\star}\|_2 \leq \varepsilon_2\|A\|_2$ en $\|A^{(k-1)\star}\|_2 > \varepsilon_2\|A\|_2$. Dan is immers $\sigma_{k+1} \leq (\varepsilon_1+\varepsilon_2)\|A\|_2$ en is σ_{k+1} numeriek niet meer te onderscheiden van 0 (zie ook (3.2.6)).

Het kan echter voorkomen dat $\sigma_{l+1} \leq (\varepsilon_1+\varepsilon_2)\|A\|_2$, terwijl bv. $\|A^{(l)}\|_2 \approx \|A^{(l)\star}\|_2 \approx \|A\|_2$ (zie volgende opgave). De 'numerieke' rang kan dus te groot geschat worden. In de praktijk gebeurt dat echter zelden: de methode heeft een betrouwbaarheid die vergelijkbaar is met de betrouwbaarheid van de LU-dekompositie met partial pivotten.

(3.4.6) Opgave. Zij $M = (m_{ij}) \in \mathbf{M}_n$ strikt boven driehoeks met $m_{ij} = 1$ $(j > i)$. Voor $c, s \in (0,1)$ met $c^2+s^2 = 1$ beschouw

$$A := U_s := \operatorname{diag}(1, s, s^2, \ldots, s^{n-1})(I - cM).$$

Ga na dat de Householder methode met kolom pivotten als in het bewijs van (3.4.4) in iedere stap weer U_s oplevert (dus $A_l = U_s$ voor iedere l).
Geef een onderschatting voor $\mathcal{C}_\infty(A)$ (raadpleeg daartoe (2.2.7) en (2.4.14)).
Omdat $\frac{\sigma_1}{\sigma_n} = \mathcal{C}_2(A) \geq \frac{1}{n}\mathcal{C}_\infty(A)$ vinden we een bovenschatting voor σ_n. Ga dit na.
Bepaal nu $\|A^{(n-1)}\|_2$ en schat σ_n voor $n = 100$ en $c = 0.2$.

(3.4.7) Lemma. *Zij $R' \in \mathbf{U}_k$ een niet-singuliere boven driehoeks matrix en $R'' \in \mathbf{M}_{k\times(n-k)}$. Dan is er een $V \in \mathbf{Q}_n$ en een niet-singuliere $\tilde{R} \in \mathbf{U}_k$ zodat* $\begin{bmatrix} R' & R'' \end{bmatrix} = \begin{bmatrix} \tilde{R} & 0 \end{bmatrix} V^*$.
(Met U_0 uit (3.4.4) is $U := U_0V$ de in (3.4.2) gewenste U .)

Bewijs. Beschouw $J \in \mathbf{M}_k$ die gedefinieerd is door $Je_{j+1} = e_{k-j}$ voor $j = 0, \ldots, k-1$ (J nummert de standaard basis vektoren van achter naar voren). Dan $J^2 = J = J^*$ en als $C \in \mathbf{U}_k$ dan $C^* \in \mathbf{L}_k$, $JC^*J \in \mathbf{U}_k$ (ga dit na).

Zij $\hat{R} := J\begin{bmatrix} R' & R'' \end{bmatrix}\begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} JR'J & JR'' \end{bmatrix}$. Met behulp van de Householder methode volgt dat $\hat{R}^* = W\begin{bmatrix} C \\ 0 \end{bmatrix}$ voor zekere $W \in \mathbf{Q}_n$ en $C \in \mathbf{U}_k$ (ga dat na). Dus

$$\begin{aligned}\begin{bmatrix} R' & R'' \end{bmatrix} = J\hat{R}\begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} &= J\begin{bmatrix} C^* & 0 \end{bmatrix} W^* \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} JC^*J & 0 \end{bmatrix}\begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} W^* \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}.\end{aligned}$$

Met $\tilde{R} := JC^*J$ en $V := \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} W \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}$ volgt het lemma.
(Men moet uiteraard in een algorithme al dat gewissel niet expliciet uitvoeren.) □

(3.4.8) Opmerking. De methode werkt alleen als we de rang van A kennen of betrouwbaar kunnen schatten (zie (3.4.5)). Soms pakt zo'n schatting verkeerd uit. We schetsen een aanpak die de singuliere waarde dekompositie berekent en wel betrouwbaar is.

Door geschikte Householder transformaties beurtelings links en rechts op A toepassen vindt men een $Q \in \mathbf{Q}_n$, $V \in \mathbf{Q}_m$ en $B = (b_{ij}) \in \mathbf{U}_{m\times n}$ zodat bi-diagonaal is $(b_{ij} \neq 0 \Rightarrow 0 \leq j-i \leq 1)$ en $A = QBV^*$ (in een volgend hoofdstuk laten we zien hoe een matrix door er geschikte Householder transformaties beurtelings links en rechts op toe te passen op een eenvoudigere vorm gebracht kan worden; dan zal ook duidelijk zijn hoe dit bi-diagonalizeren hier precies in zijn werk gaat).

Via een iteratief proces kan men vervolgens de singuliere waarde dekompositie van B bepalen (ook deze iteratieve techniek wordt pas uitgebreid in het volgend hoofdstuk besproken (QR-algorithme)).

4 Eigenwaarden en eigenvektoren: theorie

4.1 Inleiding

In (1.4.1) hebben we een aantal begrippen gedefinieerd die te maken hebben met eigenwaarden en eigenvektoren. We hebben nog een paar extra begrippen nodig.

Als $\Psi_A(\zeta) = \det(\zeta I - A) \quad (\zeta \in \mathbf{C})$ het karakteristiek polynoom is van $A \in \mathbf{M}_n$,

$$\Psi_A(\zeta) = (\zeta - \lambda_1(A))(\zeta - \lambda_2(A)) \cdots (\zeta - \lambda_n(A)) \qquad (\zeta \in \mathbf{C})$$

zijn $\lambda_1(A), \ldots, \lambda_n(A)$ de eigenwaarden van de matrix A. Onze nummering van de eigenwaarden is echter *niet* in iedere subparagraaf dezelfde. Om mooiere uitspraken te krijgen hernummeren we de $\lambda_j(A)$ wel eens.

(4.1.1) Definitie. Beschouw de situatie in (1.4.1). Zij μ een eigenwaarde van $\mathcal{A}$.
Dan is $\#\{j \in \{1, 2, \ldots, n\} \,\big|\, \mu = \lambda_j(A)\}$ de *algebraïsche multipliciteit* van μ
en is $\dim\{v \in \mathbf{X} \,\big|\, \mathcal{A}(v) = \mu v\}$ de *geometrische multipliciteit.*
Eigenwaarde μ is *k-voudig* als μ algebraïsche multipliciteit k heeft.
Beschouw een $\mathcal{G} \subset \mathbf{C}$. Met de uitdrukking "$\lambda_1, \ldots, \lambda_k$ zijn de eigenwaarde van A in $\mathcal{G}$ *geteld naar multipliciteit*" bedoelen we dat voor iedere $\mu \in \sigma(A) \cap \mathcal{G}$ het aantal $\#\{j = 1, \ldots, k \,\big|\, \lambda_j = \mu\}$ precies de algebraïsche multipliciteit van μ is.

(4.1.2) Voorbeeld. Als $A := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ dan heeft de eigenwaarde 0 geometrische multipliciteit 1 en algebraische 2. $\sigma(A) = \{0\}$ en $\lambda_1(A) = \lambda_2(A) = 0$.

(4.1.3) Bewering. *Zij $m \geq n$ en $A \in \mathbf{M}_{m\times n}$, $B \in \mathbf{M}_{n\times m}$. Dan*

$$\Psi_{AB}(\zeta) = (-\zeta)^{m-n}\Psi_{BA}(\zeta) \quad \textit{voor alle} \quad \zeta \in \mathbf{C}.$$

Blijkbaar $\sigma(BA) = \sigma(AB) \cup \{0\}$, $\lambda_j(BA) = \lambda_j(AB)$ voor $j = 1, \ldots, n$, $\lambda_j(AB) = 0$ voor $j > n$.
Bewijs. Zij I_n de identieke matrix in $\mathbf{M}_n$.
Beschouw $C := \begin{bmatrix} 0 & B \\ A & 0 \end{bmatrix} \in \mathbf{M}_{(m+n)\times(m+n)}$. Voor iedere $\zeta \in \mathbf{C}$, $\zeta \neq 0$ geldt

$$\begin{aligned} C - \zeta I_{m+n} &\overset{\substack{\text{blok}\\ \text{LU}}}{=} \begin{bmatrix} I_n & 0 \\ -\frac{1}{\zeta}A & I_m \end{bmatrix} \begin{bmatrix} -\zeta I_n & 0 \\ 0 & E \end{bmatrix} \begin{bmatrix} I_n & -\frac{1}{\zeta}B \\ 0 & I_m \end{bmatrix} \\ &\overset{\substack{\text{blok}\\ \text{UL}}}{=} \begin{bmatrix} I_n & -\frac{1}{\zeta}B \\ 0 & I_m \end{bmatrix} \begin{bmatrix} F & 0 \\ 0 & -\zeta I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ -\frac{1}{\zeta}A & I_m \end{bmatrix} \end{aligned}$$

met $E := \frac{1}{\zeta}AB - \zeta I_m$ en $F := \frac{1}{\zeta}BA - \zeta I_n$. Dus

$$\begin{aligned} \Psi_C(\zeta) &= (-\zeta)^n \det(\frac{1}{\zeta}AB - \zeta I_m) = (-1)^n\zeta^{n-m}\det(AB - \zeta^2 I_m) = (-1)^n\zeta^{n-m}\Psi_{AB}(\zeta^2) \\ &= (-\zeta)^m \det(\frac{1}{\zeta}BA - \zeta I_n) = (-1)^m\zeta^{m-n}\det(BA - \zeta^2 I_n) = (-1)^m\zeta^{m-n}\Psi_{BA}(\zeta^2). \end{aligned}$$

Blijkbaar $(-1)^{m-n}\zeta^{2m-n}\Psi_{BA}(\zeta) = \Psi_{AB}(\zeta^2)$. Vervang nu ζ^2 door ζ en de bewering volgt. $\square$

Als $p(\zeta) = \sum_{j=0}^k \alpha_j\zeta^j \quad (\zeta \in \mathbf{C})$ dan $p(A) := \sum_{j=0}^k \alpha_j A^j$ voor $A \in \mathbf{M}_n$.
Omdat $p(T^{-1}AT) = T^{-1}p(A)T$ voor niet-singuliere $T \in \mathbf{M}_n$ volgt met behulp van de Jordan normaalvorm de volgende stelling.

(4.1.4) Stelling van Cayley–Hamilton.
$\psi_A(A) = (A - \lambda_1(A))(A - \lambda_1(A)) \cdots (A - \lambda_n(A)) = 0 \quad A \in \mathbf{M}_n).$ $\square$

4.2 Eigenwaarden van het geperturbeerd stelsel

Ons uitgangspunt is weer het resultaat in (1.3.1) over het bestaan de inverse van $\mathcal{I} - \mathcal{B}$ als $\mathcal{B}$ norm kleiner dan 1 heeft voor 'n norm.

Het volgende eenvoudige lemma levert in de navolgende toepassingen nuttige methoden om de ligging van de eigenwaarden in het complexe-vlak te schatten en gevoeligheid van de eigenwaarden voor perturbaties te beoordelen.

(4.2.1) Lemma. *Zij $\|\cdot\|$ een multiplikatieve norm op $\mathbf{M}_n(\mathbf{C})$.*
Zij $A \in \mathbf{M}_n(\mathbf{C})$. Als $K, R \in \mathbf{M}_n$ zodat $A = K - R$ dan

$$\sigma(A) \subset \{\lambda \in \mathbf{C}\backslash\sigma(K) \,\big|\, 1 \leq \|(\lambda I - K)^{-1} R\|\} \cup \sigma(K).$$

Bewijs. Beschouw een λ die niet in de rechter verzameling zit: $\lambda \notin \sigma(K)$ en $\|(\lambda I - K)^{-1} R\| < 1$. $\lambda I - K$ is inverteerbaar. Merk op dat $\lambda I - A = (\lambda I - K) + R = (\lambda I - K)(I + (\lambda I - K)^{-1} R)$. Volgens (1.3.1) is $I + (\lambda I - K)^{-1} R$ ook inverteerbaar en de existentie van $(\lambda I - A)^{-1}$ volgt. Blijkbaar $\lambda \notin \sigma(A)$. □

(4.2.2) Opmerking. Men kan diverse varianten bedenken. Zo geldt

$$\sigma(A) \subset \{\lambda \in \mathbf{C}\backslash\sigma(K) \,\big|\, 1 \leq \|R(\lambda I - K)^{-1}\|\} \cup \sigma(K)$$

(en als $K^* = K$ dan

$$\sigma(A) \subset \{\lambda \in \mathbf{C} \,\big|\, \lambda I - K > 0 \quad \text{en} \quad 1 \leq \|(\lambda I - K)^{-\frac{1}{2}} R (\lambda I - K)^{-\frac{1}{2}}\|\} \cup \{\lambda \,\big|\, \lambda I - K \not> 0\}. \quad)$$

Toepassingen: Schattings stellingen

In deze subparagraaf geven we een eenvoudige manier om de eigenwaarden te schatten. De schatting zijn welliswaar grof maar toch vaak nuttig.

(4.2.3) Voor een $\mu \in \mathbf{C}$ en $\rho \in [0, \infty)$ is $\mathcal{D}_\rho(\mu) := \{\lambda \in \mathbf{C} \,\big|\, |\lambda - \mu| \leq \rho\}$, de cirkelschijf in het complexe vlak met middelpunt μ en straal ρ. De cirkel $\mathcal{S}_\rho(\mu) := \{\lambda \in \mathbf{C} \,\big|\, |\lambda - \mu| = \rho\}$ is de rand van $\mathcal{D}_\rho(\mu)$.

(4.2.4) Bewering. *Zij $A = (a_{ij}) \in \mathbf{M}_n$. Dan $\sigma(A) \subset \bigcap_\zeta \mathcal{D}_{\|\zeta I - A\|}(\zeta)$.*

Bewijs. Pas het lemma toe met $K = \zeta I$: $\|(\lambda - \zeta)^{-1}(\zeta I - A)\| = \|\zeta I - A\| / |\lambda - \zeta|$. □

(4.2.5) Eerste Gershgorin cirkelstelling. *Zij $A = (a_{ij}) \in \mathbf{M}_n$.*

$$\sigma(A) \subset \bigcup_{j=1}^n \mathcal{D}_{\rho_j}(a_{jj}) \qquad \textit{met} \qquad \rho_j := \sum_{i \neq j} |a_{ji}| = \sum_{j=1}^n |a_{ji}| - |a_{jj}|,$$

$\mathcal{D}_{\rho_j}(a_{jj})$ is de cirkelschijf met middelpunt a_{jj} en straal ρ_j, de absoluut som van de overige elementen in de j-de rij.

Bewijs. Schrijf $A = D + B$ met $D = \operatorname{diag}(A)$. Neem $\|\cdot\| = \|\cdot\|_\infty$. Merk op dat

$$\|(\lambda I - D)^{-1}(-B)\|_\infty = \max_j \big(\sum_{i \neq j} |a_{ji}|\big) / |\lambda - a_{jj}|.$$

Dus $\sigma(A) \subset \{\lambda \in \mathbf{C}\backslash\{a_{ii}\} \,\big|\,$ er is een j zodat $|\lambda - a_{jj}| \leq \sum_{i \neq j} |a_{ji}|\} \cup \{a_{jj}\}$. □

(4.2.6) Met behulp van de $\|\cdot\|_1$-norm, en de eerste variant in (4.2.2), volgt $\sigma(A) \subset \bigcup_{j=1}^n \mathcal{D}_{\tilde{\rho}_j}(a_{jj})$ met $\tilde{\rho}_j = \sum_{i \neq j} |a_{ij}|$, j-de cirkelschijf heeft straal absoluut som van de j-de kolom elementen zonder diagonaal element.

(4.2.7) Voorbeeld.
$A := \begin{bmatrix} 2 & 1 & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & 1 & 9 \end{bmatrix}$. Dan $\sigma(A) \subset \mathcal{G} := \mathcal{D}_2(2) \cup \mathcal{D}_1(1) \cup \mathcal{D}_1(9) = \mathcal{D}_2(2) \cup \mathcal{D}_1(9)$.
Via de kolom sommen $\sigma(A) \subset \tilde{\mathcal{G}} := \mathcal{D}_{\frac{1}{2}}(2) \cup \mathcal{D}_2(1) \cup \mathcal{D}_{\frac{3}{2}}(9) = \mathcal{D}_2(1) \cup \mathcal{D}_{\frac{3}{2}}(9)$.
$\sigma(A)$ is bevat in doorsnede $(\mathcal{D}_2(2) \cap \mathcal{D}_2(1)) \cup \mathcal{D}_1(9)$ van deze twee verenigingen.

We kunnen op een eenvoudige manier nog meer verzamelingen kreëren die $\sigma(A)$ bevatten. Door naar de doorsnede van die verzamelingen te kijken krijgen we betere schattingen voor $\sigma(A)$. De volgende 'schalings truuk' wordt zeer veel gebruikt. Een cirkelschijf wordt klein gemaakt (ten koste van de andere):

$$A_\varepsilon := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ \frac{1}{2} & 1 & \frac{1}{2} \\ 0 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{\varepsilon} \end{bmatrix} = \begin{bmatrix} 2 & 1 & \frac{1}{\varepsilon} \\ \frac{1}{2} & 1 & \frac{1}{2\varepsilon} \\ 0 & \varepsilon & 9 \end{bmatrix}.$$

Omdat $\sigma(A) = \sigma(A_\varepsilon)$ vinden we $\sigma(A) \subset \mathcal{G}_\varepsilon := \mathcal{D}_{1+1/\varepsilon}(2) \cup \mathcal{D}_{\frac{1}{2}(1+1/\varepsilon)}(1) \cup \mathcal{D}_\varepsilon(9)$.
Omdat $\mathcal{G} \cap \tilde{\mathcal{G}} \cap \mathcal{G}_\varepsilon = (\mathcal{D}_2(2) \cap \mathcal{D}_2(1)) \cup \mathcal{D}_\varepsilon(9)$ voor $\varepsilon \in [\frac{1}{5}, 1]$ vinden we in de buurt van 9 een klein gebied waar $\sigma(A)$ 'mee te maken heeft'. (Waarom nemen we $\varepsilon \geq \frac{1}{5}$?)

We kijken nog even naar bovenstaand voorbeeld.
De eerste Gershgorin cirkelstelling leverde ons een gebied dat $\sigma(A)$ bevat. De stelling zelf sluit echter niet uit dat bijvoorbeeld 0 een eigenwaarde is van A met multipliciteit 3.
Omdat $\sum_j \lambda_j(A) = \text{spoor}(A) = \sum_j a_{jj} = 12$ en $\sigma(A) \subset (\mathcal{D}_2(2) \cap \mathcal{D}_2(1)) \cup \mathcal{D}_{\frac{1}{5}}(9)$ zien we dat $\mathcal{D}_{\frac{1}{5}}(9)$ precies één eigenwaarde bevat en de andere twee in $\mathcal{D}_2(2) \cap \mathcal{D}_2(1)$ liggen (zie dit in).

We zouden kunnen vermoeden dat iedere Gershgorin schijf $\mathcal{D}_{\rho_j}(a_{jj})$ een eigenwaarde bevat. Het volgende voorbeeld toont aan dat dit vermoeden niet korrekt is. De tweede Gershgorin cirkelstelling stelt ons in staat om de ligging van de eigenwaarden toch wat beter te bepalen. Het bewijs van deze stelling stellen we uit tot een volgende subparagraaf.

(4.2.8) Voorbeeld. Beschouw voor een $\varepsilon > 0$
$A := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \varepsilon & 0 & 0 & \dots & 0 \end{bmatrix}$. Omdat $\Psi_A(\lambda) = \lambda^n - \varepsilon$ $(\lambda \in \mathbf{C})$ is $\lambda_j(A) = \sqrt[n]{\varepsilon} \exp(\frac{2\pi i}{n} j)$.
Voor $\varepsilon \in (0,1)$ bevat de n-de Gershgorin cirkelschijf $\mathcal{D}_\varepsilon(0)$ geen eigenwaarde en voor $\varepsilon > 1$ bevatten de 1-ste t/m de $n-1$-ste cirkelschijf $\mathcal{D}_1(0), \dots, \mathcal{D}_1(0)$ geen eigenwaarde.

(4.2.9) Tweede Gershgorin cirkelstelling. *Schrijf* $\mathbf{n} := \{1, \dots, n\}$.
Zij $A = (a_{ij})$ *en* ρ_j *als in* (4.2.5).*Beschouw een deel* E *van* $\mathbf{n}$.

$$\textit{Als} \quad \left(\bigcup_{j \in E} \mathcal{D}_{\rho_j}(a_{jj})\right) \cap \left(\bigcup_{j \notin E} \mathcal{D}_{\rho_j}(a_{jj})\right) = \emptyset \quad \textit{dan} \quad \#E = \#\{l \in \mathbf{n} \,\big|\, \lambda_l(A) \in \bigcup_{j \in E} \mathcal{D}_{\rho_j}(a_{jj})\} :$$

als k Gershgorin cirkelschijven disjukt liggen van de overige dan bevat de vereniging van deze k schijven precies k eigenwaarden van A (geteld naar multipliciteit).

Toepassingen: Perturbatie resultaten

In de volgende stelling bekijken we een diagonalizeerbare matrix A en onderzoeken we hoe het spektrum verandert als we de matrix verstoren met een matrix R.

(4.2.10) Eerste Bauer–Fike cirkelstelling. *Laat* $\|\cdot\|$ *een multiplikatieve norm zijn op* $\mathbf{M}_n$ *waarvoor* $\|D\| = \max_j |d_{jj}|$ *voor iedere diagonaal* $D = (d_{ij}) \in \mathbf{M}_n$.
Zij $A \in \mathbf{M}_n$ *zodat* $A = VDV^{-1}$ *voor zekere niet-singuliere* $V \in \mathbf{M}_n$ *en diagonaal matrix* $D = \mathrm{diag}(\mu_1, \ldots, \mu_n) \in \mathbf{M}_n$
(A is diagonalizeerbaar, $\mu_1, \ldots, \mu_n$ zijn de eigenwaarden van A geteld naar multipliciteit).
Beschouw $\tilde{A} := A + R$ *voor zekere* $R \in \mathbf{M}_n$. *Dan*

$$\sigma(\tilde{A}) \subset \bigcup_{j=1}^{n} \mathcal{D}_\rho(\mu_j) = \bigcup_{\mu\in\sigma(A)} \mathcal{D}_\rho(\mu) \qquad \textit{met} \qquad \rho := \mathcal{C}_{\|\cdot\|}(V)\|R\|.$$

Bewijs. Merk op dat $(\lambda I - A)^{-1} = V(\lambda I - D)^{-1}V^{-1}$.
Definieer de norm $\|\cdot\|_V$ op $\mathbf{M}^n$ door $\|C\|_V := \|V^{-1}CV\|$ voor $C \in \mathbf{M}_n$.
$\|\cdot\|_V$ is een multiplikatieve norm (ga dit na).
Beschouw een $\lambda \in \sigma(\tilde{A})$. Als $\lambda \notin \sigma(A)$ geldt volgens (4.2.1) dat

$$\begin{aligned} 1 \leq \|(\lambda I - A)^{-1}(-R)\|_V &\leq \|(\lambda I - A)^{-1}\|_V \|R\|_V = \|(\lambda I - D)^{-1}\| \|R\|_V \\ &= \max_j \frac{1}{|\lambda - \mu_j|}\|R\|_V = \frac{1}{\min_j |\lambda - \mu_j|}\|R\|_V. \end{aligned}$$

Blijkbaar $\min_j |\lambda - \mu_j| \leq \|R\|_V = \|V^{-1}RV\| \leq \mathcal{C}(V)\|R\|$. □

(4.2.11) Opmerking. Iedere p-norm (met $p \in [1, \infty]$) heeft de eigenschap die vereist is in de Bauer–Fike stelling (ga dit na).

De Bauer–Fike stelling heeft ook een variant die wat meer over de ligging van de eigenwaarden in de schijven zegt. Ook het bewijs van deze stelling stellen we nog even uit.

(4.2.12) Tweede Bauer–Fike cirkelstelling. *Beschouw* (4.2.10). *Zij* $E \subset \mathbf{n}$. *Stel dat*

$$\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset \quad \textit{met} \quad \mathcal{K}_1 := \bigcup_{j\in E} \mathcal{D}_\rho(\mu_j) \quad \textit{en} \quad \mathcal{K}_2 := \bigcup_{j\notin E} \mathcal{D}_\rho(\mu_j).$$

Dan bevat $\mathcal{K}_1$ *geteld naar multipliciteit* $\#E$ *eigenwaarden van* $\tilde{A}$.

(4.2.13) Opgave. Beschouw (4.2.10). Bewijs dat er een nummering $\tilde{\mu}_1, \ldots, \tilde{\mu}_n$ van de eigenwaarden van $\tilde{A}$ is (geteld naar multipliciteit) zodat $|\tilde{\mu}_j - \mu_j| \leq (2n-1)\mathcal{C}(V)\|R\|$ voor iedere $j = 1, \ldots, n$.
(Hint: Als $\mathcal{K}_1$ in (4.2.12) samenhangend is, is $|\lambda - \mu_j| \leq (2n-1)\mathcal{C}(V)\|R\|$ voor alle $\lambda \in \mathcal{K}_1$ en $j \in E$.)

(4.2.14) De kolom vektoren van V zijn precies de eigenvektoren van A. De konditiegetal $\mathcal{C}(V)$ van de eigenvektoren is blijkbaar een (boven)maat voor de gevoeligheid van de eigenwaarden voor storingen. Als $\mathcal{C}(V)$ niet groot is kunnen we stellen dat het eigenwaarden probleem 'goed gekonditioneerd' is. Als A bijvoorbeeld normaal is, is er een orthonormale basis van eigenvektoren en is $\mathcal{C}_2(V) = 1$. Voor normale matrices is het eigenwaarden probleem goed gekonditioneerd.
Aan de hand van een voorbeeld laten we zien dat bij een slecht 'gekonditioneerde basis van eigenvektoren' de eigenwaarden zeer gevoelig kunnen zijn voor storingen.
Omdat $\mathcal{C}(V)$ in de praktijk moeilijk zoniet onmogelijk te bepalen is (we kennen V in het algemeen niet), is $\mathcal{C}(V)$ voornamelijk theoretisch van belang.

(4.2.15) Voorbeeld. Beschouw voor $\varepsilon \in \mathbf{R}$ de matrices $A, R \in \mathbf{M}_{20}$ en $\tilde{A} := A + R$ met

$$A := \begin{bmatrix} 1 & 20 & 0 & \dots & 0 & 0 \\ 0 & 2 & 20 & \dots & 0 & 0 \\ 0 & 0 & 3 & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & 19 & 20 \\ 0 & 0 & 0 & \dots & 0 & 20 \end{bmatrix} \quad \text{en} \quad R := \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 0 \\ \varepsilon & 0 & 0 & \dots & 0 & 0 \end{bmatrix}.$$

$$\text{Dan} \quad \sigma(A) = \{1, 2, \dots, 20\} \quad \text{en} \quad V = \begin{bmatrix} 1 & \frac{20}{1!} & \frac{20^2}{2!} & \dots & \frac{20^{18}}{18!} & \frac{20^{19}}{19!} \\ 0 & 1 & \frac{20}{1!} & \dots & \frac{20^{17}}{17!} & \frac{20^{18}}{18!} \\ 0 & 0 & 1 & \dots & \frac{20^{16}}{16!} & \frac{20^{17}}{17!} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & \frac{20}{1!} \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}.$$

Dus $\mathcal{C}_\infty(V) \geq \|V\|_m = \frac{20^{19}}{19!} \approx 4.3\,10^7$. Verder is $\Psi_{\tilde{A}}(\lambda) = p(\lambda) - 20^{19}\varepsilon$ met $p(\lambda) = (\lambda - 1)(\lambda - 2)\cdots(\lambda - 20)$ $(\lambda \in \mathbf{C})$. Voor $\varepsilon := 20^{-19}p(10.5) \approx 1.2\,10^{-15}$ is 10.5 een eigenwaarde van $\tilde{A}$ met algebraïsche multipliciteit 2.

De eigenwaarden van A liggen homogeen verdeelt tussen 1 en 20, de coëfficiënten van A zijn niet groot toch heeft een zeer kleine verstoring R (met $\|R\|_2 \leq 2\ 10^{-15}$) een niet onaanzienlijk effect op de eigenwaarden. Voor A is het eigenwaarden probleem inderdaad slecht gesteld.

De volgend resultaat maakt geen gebruik van lemma (4.2.1). Het bewijs verloopt langs dezelfde lijn als het bewijs van de Bauer–Fike stelling. Daarom vermelden we de schatting hier.

Met behulp van de schatting kunnen we gegeven een vektor en een skalair schatten hoever die skalair van 'n eigenwaarde van de matrix afligt. We moeten daartoe een 'residu' berekenen en het konditie-getal van de eigenvektor basis kennen.

(4.2.16) Bewering [Eisberg]. *Zij* $p \in [1, \infty]$.
Zij $A \in \mathbf{M}_n$ *zodat* $A = VDV^{-1}$ *voor zekere niet-singuliere* $V \in \mathbf{M}_n$ *en diagonaal matrix* D. *Zij* $\tilde{\lambda} \in \mathbf{C}$ *en* $\tilde{v} \in \mathbf{C}^n$. *Beschouw* $\vec{r} := (A - \tilde{\lambda}I)\tilde{v}$. *Dan*

$$\min_{\mu \in \sigma(A)} |\mu - \tilde{\lambda}| \leq \mathcal{C}_p(V) \frac{\|\vec{r}\|_p}{\|\tilde{v}\|_p}$$

Bewijs. Definieer als in het bewijs van (4.2.10) de norm $\|\cdot\|_V$ op $\mathbf{M}^n$ door $\|\vec{y}\|_V := \|V^{-1}\vec{y}\|_p$ voor $C \in \mathbf{M}_n$. Dan $\|C\|_V := \|V^{-1}CV\|_p$ voor $C \in \mathbf{M}_n$.

Ook als in het bewijs van (4.2.10) vinden we

$$\|\tilde{v}\|_V = \|(A - \tilde{\lambda}I)^{-1}\vec{r}\|_V \leq \|(A - \tilde{\lambda}I)^{-1}\|_V \, \|\vec{r}\|_V = \frac{1}{\min_{\mu \in \sigma(A)} |\mu - \tilde{\lambda}|} \|\vec{r}\|_V.$$

Blijkbaar

$$\min_{\mu \in \sigma(A)} |\mu - \tilde{\lambda}| \leq \frac{\|\vec{r}\|_V}{\|\tilde{v}\|_V} = \frac{\|V^{-1}\vec{r}\|_p}{\|V^{-1}\tilde{v}\|_p} \leq \mathcal{C}_p(V) \frac{\|\vec{r}\|_p}{\|\tilde{v}\|_p}. \quad \square$$

4.3 Kontinue eigenwaarde funkties

Voor numerieke berekening is het ook van belang te weten hoe de eigenwaarden (kwalitatief) afhangen van perturbaties.
Bekijk op een interval $J \subset \mathbf{R}$ een kontinue afbeelding $\tau \rightsquigarrow A(\tau)$ van J naar $\mathbf{M}_n(\mathbf{C})$ (de matrix-coëfficiënten zijn kontinue complex-waardige funkties op J). Als in het bewijs van de Bauer–Fike stelling kan men aantonen dat het spektrum $\sigma(A(\tau))$ kontinu van τ afhangt in de

volgende zin: voor $\tau_0 \in J$ geldt

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \text{zodat} \quad \tau \in J, |\tau_0 - \tau| < \delta \Rightarrow \sigma(A(\tau)) \subset \bigcup \{\mathcal{D}_\varepsilon(\mu) \,\big|\, \mu \in \sigma(A(\tau_0))\}.$$

In de buurt van iedere eigenwaarde ligt 'n eigenwaarde van de naburige matrix. De uitspraak vertelt niet of er in de buurt van een eigenwaarde met multipliciteit 3 ook weer 3 eigenwaarden (geteld naar multipliciteit) van de naburige matrix liggen. Met behulp van complexe funktie theorie kan men bewijzen dat dit zo is[4]. Vervolgens kan men bewijzen, maar dat is ook zeker niet triviaal, dat bij iedere τ de eigenwaarden van $A(\tau)$ zo genummerd kunnen worden dat kontinue 'eigenwaarde funkties' resulteren. We formuleren de stelling, maar geven geen bewijs.

(4.3.1) Stelling. *Zij* $\tau \rightsquigarrow A(\tau)$ *kontinu van een interval* $J \subset \mathbf{R}$ *naar* $\mathbf{M}_n(\mathbf{C})$. *Dan bestaan er kontinue funkties* $\lambda_1, \ldots, \lambda_n : J \to \mathbf{C}$ *zodat voor iedere* $\tau \in J$, $\lambda_1(\tau), \ldots, \lambda_n(\tau)$ *de eigenwaarden zijn van* $A(\tau)$ *geteld naar multipliciteit.* □

Tweede cirkelstellingen volgen uit deze kontinue afhankelijkheid van de eigenwaarden van perturbaties (of ook al rechtstreekser uit de eerste cirkelstelling en argumenten als in voetnoot 4). We bewijzen expliciet de tweede cirkelstelling van Bauer–Fike. Het bewijs van die van Gershgorin laten we aan de lezer over.

(4.3.2) Bewijs van de tweede cirkelstelling van Bauer–Fike.
Beschouw $A(\tau) := A + \tau R$ met A, R als in (4.2.10) en $\tau \in [0,1]$ zodat $A(0) = A$ en $A(1) = \tilde{A}$. Laat $E \subset \mathbf{n}$, $\mathcal{K}_1$ en $\mathcal{K}_2$ zijn als in (4.2.12).

Beschouw een kontinue eigenwaarde funktie $\lambda_k(\cdot)$ van $A(\cdot)$ als in (4.3.1).
Met $J_i := \{\tau \in [0,1] \,\big|\, \lambda_k(\tau) \in \mathcal{K}_i\}$ $(i = 1, 2)$ is J_i gesloten ($\mathcal{K}_i$ is gesloten en λ_k is kontinu) en $J_1 \cap J_2 = \emptyset$ $(\mathcal{K}_1 \cap \mathcal{K}_2 = \emptyset)$. Volgens de eerste cirkelstelling geldt voor iedere $\tau \in [0,1]$

$$\lambda_k(\tau) \in \sigma(A(\tau)) \subset \bigcup_{j \in \mathbf{n}} \mathcal{D}_{\mathcal{C}(V)\|\tau R\|}(\mu_j) \subseteq \bigcup_{j \in \mathbf{n}} \mathcal{D}_{\mathcal{C}(V)\|R\|}(\mu_j) = \mathcal{K}_1 \cup \mathcal{K}_2.$$

Blijkbaar $[0,1] \subseteq J_1 \cup J_2$. Dit kan alleen als óf $J_1 = \emptyset$ óf $J_1 = [0,1]$: als $\lambda_k(0) \in \mathcal{K}_1$ dan $\lambda_k(1) \in \mathcal{K}_1$, etc..

Blijkbaar heeft geteld naar multipliciteit $A(0)$ precies evenveel eigenwaarden in $\mathcal{K}_1$ als $A(1)$.□

(4.3.3) Opgave. Bewijs de tweede cirkelstelling van Gershgorin.
(Hint: Beschouw $A(\tau) := D - \tau B$ met $A = D - B$, $D = \mathrm{diag}(A)$ en $\tau \in [0,1]$.)

Als $\tau \rightsquigarrow A(\tau)$ gladder is zouden we misschien ook gladdere eigenwaarde funkties verwachten. We bekijken een aantal voorbeelden.

(4.3.4) Voorbeelden. De matrices $A(\tau)$ zijn gedefinieerd voor $\tau \in \mathbf{C}$. We geven eerst $A(\tau)$ en dan de eigenwaarde funkties.

(a) $\begin{bmatrix} 0 & \tau \\ \tau & 0 \end{bmatrix}$ $\begin{matrix} \tau \\ -\tau \end{matrix}$ (b) $\begin{bmatrix} 0 & 1 \\ \tau & 0 \end{bmatrix}$ $\begin{matrix} \sqrt{\tau} \\ -\sqrt{\tau} \end{matrix}$ (c) $\begin{bmatrix} 1 & \tau \\ \tau & -1 \end{bmatrix}$ $\begin{matrix} \sqrt{1+\tau^2} \\ -\sqrt{1+\tau^2} \end{matrix}$

(d) $\begin{bmatrix} 0 & \tau \\ 0 & 0 \end{bmatrix}$ $\begin{matrix} 0 \\ 0 \end{matrix}$ (e) $\begin{bmatrix} 1 & \tau \\ 0 & 0 \end{bmatrix}$ $\begin{matrix} 0 \\ 1 \end{matrix}$ (f) $\begin{bmatrix} \tau & 1 \\ 0 & 0 \end{bmatrix}$ $\begin{matrix} \tau \\ 0 \end{matrix}$

(g) $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & \tau \\ \tau & 0 & 0 \end{bmatrix}$ $\begin{matrix} |\tau|^{\frac{2}{3}} \exp\left(\frac{2\pi i}{3} 0\right) \\ |\tau|^{\frac{2}{3}} \exp\left(\frac{2\pi i}{3} 1\right) \\ |\tau|^{\frac{2}{3}} \exp\left(\frac{2\pi i}{3} 2\right) \end{matrix}$

---

[4]Beschouw $\varepsilon > 0$. Zij $\delta > 0$ als in de kontinuïteits uitspraak voor het spektrum. Beschouw een $\lambda_0 \in \sigma(A(\tau_0))$. Merk op dat $\mathcal{S}_\varepsilon(\lambda_0) \cap \sigma(A(\tau)) = \emptyset$ als $\tau \in J, |\tau - \tau_0| < \delta$.
Schrijf $p_\tau(\zeta) := \det(\zeta I - A(\tau)) = (\zeta - \lambda_1^\tau) \cdots (\zeta - \lambda_n^\tau)$ $(\zeta \in C)$. Voor $\tau \in J, |\tau - \tau_0| < \delta$ is

$$N_\tau := \frac{1}{2\pi i} \oint_{\mathcal{S}_\varepsilon(\lambda_0)} \frac{p_\tau'(\zeta)}{p_\tau(\zeta)} \, d\zeta = \frac{1}{2\pi i} \oint_{\mathcal{S}_\varepsilon(\lambda_0)} \sum_{j=1}^{n} \frac{1}{\zeta - \lambda_j^\tau} \, d\zeta$$

het aantal nulpunten van p_τ op $\mathcal{D}_\varepsilon(\lambda_0)$. Als funktie van τ is N_τ kontinu van $[\tau_0 - \delta, \tau_0 + \delta] \cap J$ naar N en dus konstant. Blijkbaar hebben p_τ en p_{τ_0} op $\mathcal{D}_\varepsilon(\lambda_0)$ evenveel nulpunten. □

Uit voorbeeld (b) blijkt dat er geen n *eenwaardige* kontinue funkties zijn van $\mathbf{M}_n(\mathbf{C})$ naar $\mathbf{C}$ die de eigenwaarden representeren (de wortel funktie is niet eenwaardig in 'n omgeving van 0 in $\mathbf{C}$).

In alle voorbeelden is $A(\tau)$ van de vorm $A+\tau R$. Toch is de wijze waarop voor 'n eigenwaarde funktie $\lambda_k(\cdot)$ de waarde $\lambda_k(\tau)$ naar $\lambda_k(0)$ gaat voor $\tau \to 0$ zeer verschillend, nl. als $\mathcal{O}(\tau^2)$ (in (c)), als $\mathcal{O}(\tau)$ (in (a) en (f)) en als $\mathcal{O}(\sqrt{\tau})$ (in (b)). Zelfs in geval $A(0)$ dezelfde is hangt het gedrag sterk af van de wijze waarop $A(\tau)$ naar $A(0)$ gaat (zie (f) en (b)).

In de buurt van enkelvoudige eigenwaarden (in (c) met $\tau \approx 0$ en in (a), (b), (e), (f), (g) met $\tau \not\approx 0$) zijn de eigenwaarde funkties glad.

(4.3.5) Opgave. Bepaal voor de voorbeelden (a), (b), (d) en (g) een eigenvektor basis en bereken het konditie getal van deze basis (als funktie van τ, $\tau \neq 0$).

Als $\tau_0 \in J$ en $\lambda_k(\tau_0)$ een enkelvoudige eigenwaarde is van $A(\tau_0)$ dan kan men met de impliciete funktiestelling[5] bewijzen dat in 'n δ-omgeving van τ_0 in J de eigenwaarde funktie $\lambda_k(\cdot)$ minstens zo glad is als de matrix-funktie $A(\cdot)$.
Als $\lambda_k(\tau_0)$ een enkelvoudige eigenwaarde is van $A(\tau_0)$ geldt in feite iets veel mooiers. λ_k en de bijbehorende eigenvektor hangen analytisch af van de *matrix*. Hangt de matrix bv. C^1 van τ af, dan hangt de eigenwaarde λ_k daardoor vanzelf ook C^1 af van τ. Analytisch betekent lokaal te schrijven als een machtreeks. In §4.6 stelt een expliciete konstruktie van de koëfficiënten van de machtreeks ons in staat het effect van perturbaties op een enkelvoudige eigenwaarde en bijbehorende eigenvektor te kwantificeren.

Door naar de lineaire matrix-funktie $A(\tau) := A+\tau R$ op $[0,1]$ te kijken krijgen we al interessante uitspraken voor geperturbeerde matrices (zie bv. het bewijs van tweede cirkelstellingen). Om nog meer al te technische details te vermijden vermelden we daarom nu alleen nog maar voor algemene matrices een resultaat voor deze lineaire matrix-funkties (hoewel de uitspraak in de volgende stelling in essentie korrekt is voor iedere analytische matrix-funktie). De matrix-funkties in (4.3.4) zijn alle lineair.

(4.3.6) Stelling. *Zij $A, R \in \mathbf{M}_n$.*
Beschouw $\tau \rightsquigarrow A(\tau) := A+\tau R$ voor $\tau \in \mathbf{C}$ met eigenwaarde funkties $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$.
Zij $\tilde{\mathcal{E}} := \{\tau \in \mathbf{C} \,\big|\, A(\tau)$ heeft 'n meervoudige eigenwaarde $\}$.
(a) *Als $\mathcal{J} \subset \mathbf{C}$ een 1-dimensionale C^1-variëteit is dan kunnen $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$ zo gekozen worden dat ze kontinu op $\mathcal{J}$ zijn.*
(b) *Het aantal $n(\tau)$ verschillende eigenwaarden ($n(\tau) := \#\sigma(A(\tau))$) is konstant op $\mathbf{C}\backslash\mathcal{E}$ voor een zekere eindige $\mathcal{E} \subset \tilde{\mathcal{E}}$.*
(c) *Op ieder enkelvoudig samenhangend gebied $\mathcal{G}_0 \subset \mathbf{C}\backslash\mathcal{E}$ kunnen $\lambda_1(\cdot), \ldots, \lambda_n(\cdot)$ analytisch gekozen worden.*
(d) *Stel dat $0 \in \mathcal{E}$ en $\lambda_0 \in \sigma(A(0))$ k-voudig is.*
Hernummer de $\lambda_j(\cdot)$ zodat $\lambda_j(0) = \lambda_0$ voor $j \leq k$ en $\lambda_j(0) \neq \lambda_0$ voor $j > k$.
Dan zijn er omgevingen $\mathcal{V}$ en $\mathcal{U}$ van 0 in $\mathbf{C}$,
analytische funkties $\phi_1, \ldots, \phi_q$ op $\mathcal{V}$ en $p_1, \ldots, p_q \in \mathbf{N}$ zodat $p_1 + \ldots + p_q = k$ en

$$\{\lambda_1(\tau), \ldots, \lambda_k(\tau)\} = \bigcup_{j=1}^{q} \phi_j(\mathcal{W}(\tau, p_j)) \text{ voor alle } \tau \in \mathcal{U}, \text{ met } \mathcal{W}(\tau,p) := \{\zeta \in \mathbf{C} \,\big|\, \zeta^p = \tau\}$$

$\mathcal{W}(\tau,p)$ is de verzameling van de p p-de machtswortels is van τ. □
(– Als $k=1$ is $q=1$, $p_1=1$ en $\lambda_1(\tau) = \phi_1(\tau)$: λ_1 is analytisch op $\mathcal{U}$ (konform (b)).
– λ_1 is voor 'n $p \in \{p_1, \ldots p_q\}$ en $m \in \{0, \ldots, p-1\}$ van de vorm

$$\lambda_1(\tau) = \lambda_0 + \alpha_1 \omega_p^m \eta_\tau + \alpha_2 \omega_p^{2m} \eta_\tau^2 + \ldots^{6} \quad \text{voor alle } \tau = re^{i\varphi} \in \mathcal{U},\ \eta_\tau = r^{\frac{1}{p}} \exp(\tfrac{i\varphi}{p}) \qquad (*)$$

[5] Beschouw $F(\zeta,\tau) := \det(\zeta I - A(\tau))$. Men kan nagaan dat $\frac{\partial F}{\partial \zeta}(\lambda_k(\tau_0), \tau_0) \neq 0$ als $\lambda_k(\tau_0)$ een enkelvoudig nulpunt is van $F(\cdot, \tau_0) = \Psi_{A(\tau_0)}$. Voor $\tau \approx \tau_0$ garandeert de impliciete funktiestelling nu dat $F(\lambda(\tau), \tau) = 0$ voor een zekere gladde $\lambda(\cdot)$ met $\lambda(\tau_0) = \lambda_k(\tau_0)$ Uit de enkelvoudigheid van $\lambda_k(\tau_0)$ en de kontinuiteit van alle $\lambda_j(\cdot)$ volgt $\lambda(\tau) = \lambda_k(\tau)$ voor $\tau \approx \tau_0$ en dus de gladheid van $\lambda_k(\cdot)$.

[6] Als $\sum_j \alpha_j \zeta^j$ een machtreeks is, is $\sum_j \alpha_j \zeta^{j/p}$ een *Puiseux reeks.*

met $\omega_p := \exp(\frac{2\pi i}{p})$, en $\sum \alpha_j \zeta^j$ een machtreeks; $\mathcal{W}(\tau, p) = \{\omega_p^j \eta_\tau \,\big|\, j = 0, \ldots, p-1\}$. Voor iedere $m \in \{0, \ldots, p-1\}$ is de puiseux reeks in $(*)$ een $\lambda_j(\tau)$ met $j \leq k$ (bij verschillende m's verschillende j's).
– Als $\tau_0 \in \mathcal{E}$ beschouw dan $\tilde{A} + \tilde{\tau} R = (A + \tau_0 R) + (\tau - \tau_0) R$ (in (c) van (4.3.4) is $\pm i \in \mathcal{E}$).
– In (g) van (4.3.4) is $\alpha_1 = 0$ en $p = 3$.)

Het 'wortel gedrag' in de voorbeelden (b), (c) en (g) in (4.3.4) is blijkbaar niet uitzonderlijk. Een meervoudige eigenwaarde is blijkbaar in het algemeen gevoelig voor perturbaties: een $\mathcal{O}(\bar{\xi})$ perturbatie op A kan een $\mathcal{O}(\sqrt[k]{\xi})$ effect op een k-voudige eigenwaarde hebben. In de voorbeelden hebben we gezien dat het effect van perturbaties niet altijd zo groot hoeft te zijn (zie (a), (d) en (f) (4.3.4)).

Voor hermiets-waardige analytische matrix-funkties kan men bewijzen dat de eigenwaarde funkties en eigenvektor funkties analytisch verondersteld mogen worden (zie (a) in (4.3.4)). We geven een stelling zonder bewijs.

(4.3.7) • Stelling. *Zij J een interval in $\mathbf{R}$ en $\tau \rightsquigarrow A(\tau)$ een analytische funktie van J naar $\mathbf{M}_n$ met $A(\tau)^* = A(\tau)$ alle $\tau \in J$. Dan zijn er analytische funkties $Q, D : J \to \mathbf{M}_n$ zodat $Q(\tau)$ unitair, $D(\tau)$ diagonaal en $A(\tau)Q(\tau) = Q(\tau)D(\tau)$ voor iedere $\tau \in J$.* □
Als $\vec{v}_j(\tau) := Q(\tau)e_j$ en $\lambda_j(\tau)$ zodat $D(\tau)e_j = \lambda_j(\tau)e_j$ dan is $\vec{v}_j(\cdot)$ een analytische funktie van eigenvektoren en $\lambda_j(\cdot)$ een analytische funktie van eigenwaarden.

4.4 Eigenwaarden van hermietse matrices

Hermietse matrices komt men in de praktijk nogal eens tegen.

(4.4.1) Notatie. Voor matrices $A \in \mathbf{M}_n$ spelen de *Rayleigh quotiënten*

$$\frac{(A\vec{x}, \vec{x}\,)}{(\vec{x}, \vec{x}\,)} \quad \text{met} \quad x \in \mathbf{C}^n$$

met name voor hermietse matrices A een grote rol. We noteren zo'n quotiënt met $\varrho_A(\vec{x}\,)$.

(4.4.2) De eigenwaarden van hermietse matrices A zijn telkens genummerd in opklimmende grootte:

$$\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_n(A).$$

In deze subparagraaf is $A \in \mathbf{M}_n$ hermiets. De stellingen voor A zijn natuurlijk ook korrekt voor $-A$. Om het resultaat voor $-A$ te interpreteren voor A bedenke men dat $\lambda_j(-A) = -\lambda_{n-j+1}(A)$ voor iedere j (ga dit na).

De stelling van Courant–Fischer of minimax-stelling heeft verschillende konsekwenties voor schattings– en perturbatie uitspraken.

(4.4.3) Stelling van Courant–Fischer. *Voor iedere $j \in \{1, \ldots, n\}$ is*

$$\min_{\mathbf{V}} \max \left\{ \frac{(A\vec{x}, \vec{x}\,)}{(\vec{x}, \vec{x}\,)} \,\Big|\, \vec{x} \in \mathbf{V},\ \vec{x} \neq 0 \right\} = \lambda_j(A),$$

waarbij het minimum genomen is over alle j-dimensionale lineaire deelruimte $\mathbf{V}$ van $\mathbf{C}^n$.

Bewijs. Laat $\vec{v}_1, \ldots, \vec{v}_n$ een orthonormale basis van eigenvektoren zijn: $A\vec{v}_i = \lambda_i(A)\vec{v}_i$.
Laat $\mathbf{V}$ j dimensionaal zijn. Beschouw $\mathbf{W} := \text{span}\,(\vec{v}_j, \ldots, \vec{v}_n\,)$.
Omdat $\dim \mathbf{W} + \dim \mathbf{V} > n$ is er een $\vec{y} \in \mathbf{W} \cap \mathbf{V}$, $\vec{y} \neq 0$. $\vec{y} = \sum_{i=j}^n \beta_i \vec{v}_i$ voor zekere β_i. Dus

$$\max_{\vec{x} \in \mathbf{V}} \frac{(A\vec{x}, \vec{x}\,)}{(\vec{x}, \vec{x}\,)} \geq \frac{(A\vec{y}, \vec{y}\,)}{(\vec{y}, \vec{y}\,)} = \frac{\sum_{i=j}^n |\beta_i|^2 \lambda_i(A)}{\sum_{i=j}^n |\beta_i|^2} \geq \frac{\sum_{i=j}^n |\beta_i|^2 \lambda_j(A)}{\sum_{i=j}^n |\beta_i|^2} = \lambda_j(A).$$

Neem nu $\mathbf{V} := \text{span}\,(\vec{v}_1, \ldots, \vec{v}_j\,)$. Met $\vec{x} := \vec{v}_j$ is $\vec{x} \in \mathbf{V}$, $(\vec{x}, \vec{x}\,) = 1$ en $(A\vec{x}, \vec{x}\,) = \lambda_j(A)$. □

(4.4.4) Scheidingsstelling van Cauchy. *Stel dat A als volgt gepartitioneerd is*

$$A = \begin{bmatrix} U & F \\ F^* & V \end{bmatrix} \quad \textit{met} \quad U \in \mathbf{M}_k,\ V \in \mathbf{M}_{n-k}.$$

Dan
$$\left.\begin{array}{rcl} \lambda_j(U) & \geq & \lambda_j(A) \\ \lambda_{n-j+1}(A) & \geq & \lambda_{k-j+1}(U) \end{array}\right\} \quad \textit{voor} \quad j = 1, \ldots, k.$$

Bewijs. Zij $\mathbf{V} \subseteq \mathbf{C}^k$ een j-dimensionale lineaire deelruimte waarvoor $\max_{\vec{y} \in \mathbf{V}} \varrho_U(\vec{y}\,) = \lambda_j(U)$. Definieer $\tilde{\mathbf{V}} := \{(\vec{y}, \vec{0}\,)^T \in \mathbf{C}^n \,\big|\, \vec{y} \in \mathbf{V}, \vec{0} \in \mathbf{C}^{n-k}\}$. Dan $\dim(\tilde{\mathbf{V}}) = j$ en volgens Courant–Fischer is $\max_{\vec{z} \in \tilde{\mathbf{V}}} \varrho_A(\vec{z}\,) \geq \lambda_j(A)$. Omdat voor $\vec{z} = (\vec{y}, \vec{0}\,)^T \in \tilde{\mathbf{V}}$ geldt $(A\vec{z}, \vec{z}\,) = (U\vec{y}, \vec{y}\,)$ en $(\vec{z}, \vec{z}\,) = (\vec{y}, \vec{y}\,)$ volgt de eerste ongelijkheid.

Uit $\lambda_j(-U) \geq \lambda_j(-A)$, $\lambda_j(-A) = -\lambda_{n-j+1}(A)$ en $\lambda_j(-U) = -\lambda_{k-j+1}(U)$ volgt de tweede ongelijkheid. □

(4.4.5) Gevolg. $\lambda_k(A) \leq \|U\|_2$. □

Schrappen we in A de j-de rij en de j-de kolom dan scheiden de eigenwaarden van de gereduceerde matrix de eigenwaarden van A.

(4.4.6) Gevolg. *Als in Cauchy's stelling* $k = n-1$ *dan*
$\lambda_1(A) \leq \lambda_1(U) \leq \lambda_2(A) \leq \ldots \leq \lambda_{n-1}(U) \leq \lambda_n(A)$. □

Hermietse perturbaties

In deze subparagraaf bekijken we hermietse perturbaties op hermietse matrices.

We hadden al gezien dat het hermiets eigenwaarden probleem goed gekonditioneerd is (iedere eigenwaarde van de geperturbeerde matrix zit in een schijf om de ongeperturbeerde eigenwaarde, een schijf met straal $\|R\|_2$, de norm van de perturbatie (Bauer–Fike)). De 'tweede cirkelstellingen' pakken voor hermietse matrices veel mooier uit: niet alleen bevat iedere schijf een 'geperturbeerde eigenwaarde' (ook al ligt hij niet disjunkt van de overige) maar de schijven en de geperturbeerde eigenwaarden kan men zo nummeren dat in de j-de schijf de j-de eigenwaarde zit (de stelling (4.4.7) van Weyl voor hermietse perturbaties H en stelling (4.4.13) voor willekeurige perturbaties R)

Omdat bij diverse berekeningsmethoden rekening wordt gehouden met het feit dat er aan een hermietse matrix gerekend wordt heeft dat vaak tot gevolg dat de perturbaties veroorzaakt door evaluatiefouten ook hermiets zijn (bij de Choleski methode hadden we bv. zo'n situatie).

In deze subparagraaf zijn $A, H \in \mathbf{M}_n$ hermiets en $\hat{A} := A + H$.

(4.4.7) Stelling van Weyl. *Voor iedere* $j = 1, \ldots, n$ *geldt*

$$\lambda_1(H) \leq \lambda_j(\hat{A}) - \lambda_j(A) \leq \lambda_n(H).$$

In het bijzonder is $|\lambda_j(\hat{A}) - \lambda_j(A)| \leq \|H\|_2$.

Bewijs. Zij $\mathbf{V} \subseteq \mathbf{C}^n$ j-dimensionaal zodat $\max_{\vec{x} \in \mathbf{V}} \varrho_A(\vec{x}\,) = \lambda_j(A)$.
Dan $\lambda_j(\hat{A}) \leq \max_{\vec{x} \in \mathbf{V}} \varrho_{A+H}(\vec{x}\,) \leq \max_{\vec{x} \in \mathbf{V}} \varrho_A(\vec{x}\,) + \max_{\vec{x} \in \mathbf{V}} \varrho_H(\vec{x}\,) \leq \lambda_j(A) + \lambda_n(H)$.
Dus ook $-\lambda_{n-j+1}(\hat{A}) = \lambda_j(-\hat{A}) \leq \lambda_j(-A) + \lambda_n(-H) = -\lambda_{n-j+1}(A) - \lambda_1(H)$. □

In de volgende opgave is de **monotoniciteits stelling** in zijn volledige algemeenheid geformuleerd.

(4.4.8) Opgave. Bewijs dat $\left.\begin{array}{rcl} \lambda_{k+m-n}(\hat{A}) & \leq & \lambda_m(A) + \lambda_k(H) \\ \lambda_{k+m-1}(\hat{A}) & \geq & \lambda_m(A) + \lambda_k(H) \end{array}\right\}$,
waarbij de k en m zo zijn dat de voorkomende λ_i's gedefiniëerd zijn.
(Hint: Als $A\vec{v}_j = \lambda_j(A)\vec{v}_j$ en $H\vec{w}_j = \lambda_j(H)\vec{w}_j$ beschouw $\mathbf{V}_1 := \operatorname{span}(\vec{v}_1, \ldots, \vec{v}_m)$ en $\mathbf{W}_1 := \operatorname{span}(\vec{w}_1, \ldots, \vec{w}_k)$ en merk op dat $\dim(\mathbf{V}) \geq k + m - n$ met $\mathbf{V} := \mathbf{V}_1 \cap \mathbf{W}_1$.)

(4.4.9) • **Opgave.** Bewijs de stelling van Weyl met behulp van (4.3.7).
(Hint: Beschouw $(A+\tau H)\vec{v}_j(\tau)=\lambda_j(\tau)\vec{v}_j(\tau)$ voor analytische $\vec{v}_j(\cdot)$ en $\lambda_j(\cdot)$ met $\|\vec{v}_j(\tau)\|_2=1$ en $\lambda_j(0)=\lambda_j(A)$. Dan $\lambda_j'(\tau)=(H\vec{v}_j(\tau),\vec{v}_j(\tau))$ en $\lambda_j(1)-\lambda_j(0)=\int_0^1\lambda'(\tau)\,d\tau$. Hernummer de $\lambda_j(1)$ om de $\lambda_j(\hat{A})$ te krijgen.)

Omdat de Frobenius norm gemakkelijk te schatten is in tegenstelling tot de 2-norm (zie ook (2.5.10)) schat men perturbaties vaak door middel van de Frobenius norm. Uit de stelling van Weyl volgt onmiddellijk zo'n schatting. De volgende is wat scherper. Het bewijs gebruikt technieken die we pas in een volgende paragraaf ontwikkelen. We bespreken ook dan pas het bewijs.

(4.4.10) **Stelling van Wielandt–Hoffman.** $\sqrt{\sum_{j=1}^{n}(\lambda_j(\hat{A})-\lambda_j(A))^2}\leq\|H\|_F.$ □

Stel een hermietse matrix A is hermiets gerepresenteerd in de computer door $\hat{A}$: dus $\hat{A}=A+\Delta_A$ met Δ_A hermiets en $|\Delta_A|\preceq\bar{\xi}|A|$.
Dan (volgens Weyl) $\quad |\lambda_j(\hat{A})-\lambda_j(A)|\leq\|\Delta_A\|_2\leq\sqrt{n}\,\bar{\xi}\,\|A\|_2=\sqrt{n}\,\bar{\xi}\max_k|\lambda_k(A)|$
of (volgens Wielandt–Hoffman) $\quad \sqrt{\sum_j|\lambda_j(\hat{A})-\lambda_j(A)|^2}\leq\bar{\xi}\,\|A\|_F=\bar{\xi}\sqrt{\sum_j|\lambda_j(A)|^2}.$
Voor een hermietse matrix worden de eigenwaarden dus zeer goed gerepresenteerd (vergelijk dit ook met voorbeeld (4.2.15)).

• Niet-hermietse perturbaties

Bij niet hermietse perturbaties van een hermietse matrix heeft de dimensie n (een milde) invloed. We zullen bewijzen (in (4.4.13)) dat de storing op de eigenwaarden gemajoreerd kan worden door ${}^2\!\log 8n\|R\|_2$ waarbij R de stoormatrix is. De schatting is (op een factor ≈ 1) scherp. Als $n\to\infty$ dan gaat ${}^2\!\log 8n$ naar ∞, maar langzaam (vul eens $n=100$ en $n=1000$ in; vergelijk met (4.2.13)).

(4.4.11) **Opgave.** Zij $B\in\mathbf{M}_n$ gepartitioneerd als
$B:=\begin{bmatrix}E&F\\G&K\end{bmatrix}\quad(=\begin{bmatrix}E&0\\0&K\end{bmatrix}+\begin{bmatrix}0&F\\G&0\end{bmatrix})$ met $E\in\mathbf{M}_k$ en $K\in\mathbf{M}_{n-k}$. Bewijs dat
$$\max(\|E\|_2,\|F\|_2,\|G\|_2.\|K\|_2)\leq\|B\|_2\leq\max(\|E\|_2,\|K\|_2)+\max(\|F\|_2,\|G\|_2).$$

(4.4.12) **Lemma.** *Voor* $A\in\mathbf{M}_n$ *geldt* $\|U-L\|_2\leq({}^2\!\log 2n)\,\|A\|_2,$ *waarbij* U *het strikt boven driehoeks deel van* A *is en* L *het strikt beneden driehoeks deel.*

Bewijs. Schrijf $\Pi(A):=U-L$ voor iedere $A\in\mathbf{M}_n$ met U strikt boven driehoeks deel van A en L strikt beneden driehoeks deel.
Zij $l\in\mathbf{N}$. Zij $\gamma_l\geq 0$ zodat $\|\Pi(A)\|_2\leq\gamma_l\|A\|_2$ voor iedere $n\leq 2^l$ en $A\in\mathbf{M}_n$.
Beschouw nu een $n\in(2^l,2^{l+1}]$ en een $B\in\mathbf{M}_n$. Schrijf $k:=2^l$. Dan $n-k\leq 2^l$. Partitioneer B als volgt $B:=\begin{bmatrix}E&F\\G&K\end{bmatrix}$ met $E\in\mathbf{M}_k$. Merk op dat $\Pi(B)=\begin{bmatrix}\Pi(E)&F\\-G&\Pi(K)\end{bmatrix}$.
Dus (volgens (4.4.11))
$$\begin{aligned}\|\Pi(B)\|_2&\leq\max(\|\Pi(E)\|_2,\|\Pi(K)\|_2)+\max(\|F\|_2,\|G\|_2)\\&\leq\gamma_l\max(\|E\|_2,\|K\|_2)+\max(\|F\|_2,\|G\|_2)\\&\leq(\gamma_l+1)\max(\|E\|_2,\|K\|_2,\|F\|_2,\|G\|_2)\leq(\gamma_l+1)\|B\|_2.\end{aligned}$$
Blijkbaar $\gamma_{l+1}\leq\gamma_l+1$. Omdat $\gamma_0=0$ is (induktie) $\gamma_{l+1}\leq l+1\leq{}^2\!\log 2n$. □

(4.4.13) **Stelling.** *Zij* $R\in\mathbf{M}_n$ *en* $\tilde{A}:=A+R$. *Laat* $\lambda_j(\tilde{A})$ *zo genummerd zijn dat*
$$\mathrm{Re}\lambda_1(\tilde{A})\leq\mathrm{Re}\lambda_2(\tilde{A})\leq\ldots\leq\mathrm{Re}\lambda_n(\tilde{A}).$$
Dan geldt voor iedere $j=1,\ldots,n$
$$\begin{cases}|\lambda_j(A)-\mathrm{Re}\lambda_j(\tilde{A})|&\leq\ {}^2\!\log 4n\,\|R\|_2\\ |\mathrm{Im}\lambda_j(\tilde{A})|&\leq\ \|R\|_2.\end{cases}$$

Bewijs. $\tilde{A} = Q(D+U)Q^*$ voor zekere $Q \in \mathbf{Q}_n$, U strikt boven driehoeks en $D = \operatorname{diag}(\lambda_1(\tilde{A}),\ldots,\lambda_n(\tilde{A}))$ (Stelling van Schur).
Schrijf $H := \frac{1}{2}(R+R^*)$, $V := \frac{1}{2}(U+U^*)$, $D_r := \frac{1}{2}(D+\bar{D})$.
(a) Omdat $A = \frac{1}{2}(\tilde{A}+\tilde{A}^*) - H = QD_rQ^* + QVQ^* - H$ volgt uit de stelling van Weyl

$$|\lambda_j(A) - \mathrm{Re}\lambda_j(\tilde{A})| = |\lambda_j(A) - \mathrm{Re}\lambda_j(D_r)| \le \|QVQ^* - H\|_2 \le \|V\|_2 + \|H\|_2.$$

Uit het vorige lemma volgt

$$\begin{aligned} 2\|V\|_2 &= \|U+U^*\|_2 \le {}^2\!\log 2n \|(D+U)-(D+U)^*\|_2 \\ &= {}^2\!\log 2n\|\tilde{A}-\tilde{A}^*\|_2 = {}^2\!\log 2n\|R-R^*\|_2 \le 2\,{}^2\!\log 2n\|R\|_2 \end{aligned}.$$

Blijkbaar $|\lambda_j(A) - \mathrm{Re}\lambda_j(\tilde{A})| \le (1 + {}^2\!\log 2n)\|R\|_2 = {}^2\!\log 4n\|R\|_2$.
(b) Voor iedere $B = (b_{ij}) \in \mathbf{M}_n$ is $|b_{ij}| \le \|B\|_2$. Dus

$$2|\mathrm{Im}\lambda_j(\tilde{A})| \le \|(D+U)-(D+U)^*\|_2 = \|\tilde{A}-\tilde{A}^*\|_2 \le 2\|R\|_2. \qquad \square$$

(4.4.14) Opgave. Bewijs het volgende. Met R en $\lambda_j(\tilde{A})$ genummerd als in (4.4.13) geldt

$$\sqrt{\sum_j |\lambda_j(A) - \mathrm{Re}\lambda_j(\tilde{A})|^2} \le 2\|R\|_F \quad \text{en} \quad \sqrt{\sum_j |\mathrm{Im}\lambda_j(\tilde{A})|^2} \le \|R\|_F.$$

(Hint: Loop het bewijs van bovenstaande stelling na en gebruik Wielandt–Hoffman.)

4.5 Nulpunten van polynomen

Eigenwaarden problemen duiken op diverse plaatsen op, ook op plaatsen die op het eerste gezicht niet direkt met matrices te maken hebben. Zo kan men een meerterms rekursie met konstante koëfficiënten beschrijven met behulp van een 'companion' matrix; de karakteristieke wortels van de karakteristieke vergelijking zijn dan precies de eigenwaarden van de companion matrix. Het karakteristieke polynoom van een matrix is een polynoom en omgekeerd kan ieder polynoom gezien worden als het karakteristiek polynoom van een 'companion' matrix. De nulpunten van het polynoom zijn weer eigenwaarden. Onze numerieke methoden om eigenwaarden te berekenen zijn dus impliciet methoden om nulpunten te berekenen.

(4.5.1) Opgave. Beschouw $p(\zeta) = \zeta^n + \alpha_1\zeta^{n-1} + \ldots + \alpha_{n-1}\zeta + \alpha_n$ ($\zeta \in \mathbf{C}$) een polynoom van graad n (zijn we geïnteresseerd in de nulpunten van p dan mogen we zonder de algemeenheid te schaden aannemen dat $\alpha_0 = 1$). Beschouw de *companion matrix* $A \in \mathbf{M}_n$

$$A := \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & \ldots & 0 & 0 \\ 0 & 0 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \ldots & -\alpha_2 & -\alpha_1 \end{bmatrix}$$

(a) Bewijs dat $p = \Psi_A$. Bepaal de eigenvektoren van A (in termen van de wortels van p).
Zij $\alpha \in \mathbf{C}$ zo dat $p(\alpha) = 0$.
(b) Bewijs dat
(i) $|\alpha| \le 1$ of $|\alpha + \alpha_1| \le \sum_{j=2}^n |\alpha_j|$ en
(ii) $|\alpha| \le 1 + \max_{2\le j\le n-1}|\alpha_j|$ of $|\alpha| \le |\alpha_n|$ of $|\alpha + \alpha_1| \le 1$.
(c) Zij $\gamma_j \in (0,1)$ zodat $\sum_{j=1}^n \gamma_j \le 1$.
Omdat $\frac{p(\alpha)}{\alpha^n} = 0$ is $1 \le |\frac{\alpha_1}{\alpha}| + |\frac{\alpha_2}{\alpha^2}| + \ldots + |\frac{\alpha_n}{\alpha^n}| = \sum_{j=1}^n |\frac{\alpha_j}{\alpha^j\gamma_j}|\gamma_j \le \max_j |\frac{\alpha_j}{\alpha^j\gamma_j}|$
en dus $|\alpha| \le \max_j \sqrt[j]{|\frac{\alpha_j}{\gamma_j}|}$.
Bewijs dit.
(i) Met $\gamma_j = 2^{-j}$ is $|\alpha| \le 2\max_j \sqrt[j]{|\alpha_j|}$.
(ii) Met $\gamma_j = |\alpha_j|/(\sum|\alpha_i|)$ is $|\alpha| \le \max(\sum|\alpha_j|, \sqrt[n]{\sum|\alpha_j|})$

4.6 Eigenvektoren van het geperturbeerd stelsel

We bekijken in deze subparagraaf twee situaties. Van een willekeurige matrix met een enkelvoudige eigenwaarde bestuderen we het effekt van perturbaties op die enkelvoudige eigenwaarde en de bijbehorende eigenvektor. Van een hermietse matrix bestuderen we ook het effekt van perturbaties op een meervoudige eigenwaarde en de bijbehorende eigenruimte.

Een eigenvektor bij een enkelvoudige eigenwaarde

(4.6.1) Zij $A_0 \in \mathbf{M}_n$.
Stel dat μ een enkelvoudige eigenwaarde is van A_0 met eigenvektor $\vec{v} \in \mathbf{K}^n \backslash \{0\}$.
Kies een basis $\vec{w}_2, \ldots, \vec{w}_n$ van $\mathbf{W} := \mathrm{Im}\,(A_0 - \mu I)$. Omdat μ enkelvoudig is, is $\vec{v} \notin \mathbf{W}$ (ga dat na met behulp van de Jordan normaalvorm) en is $\vec{v}, \vec{w}_2, \ldots, \vec{w}_n$ een basis voor $\mathbf{K}^n$.
Zij $V \in \mathbf{M}_n$ met in de kolommen deze basis vektoren; $Ve_1 = \vec{v}$ en $Ve_j = \vec{w}_j$ $(j > 1)$.
Dan heeft de matrix $A := V^{-1}A_0V$ de volgende vorm (ga dat na)

$$A = \begin{bmatrix} \mu & \vec{0}^* \\ \vec{0} & B \end{bmatrix} \quad \text{met} \quad B \in \mathbf{M}_{n-1}(\mathbf{C}), \quad \vec{0} = (0, \ldots, 0)^T \in \mathbf{C}^{n-1}.$$

Ook omdat μ enkelvoudig is volgt dat $\mu \notin \sigma(B)$.
(Als A_0 de matrix is van de afbeelding $\mathcal{A}$ ten opzichte van de standaard basis op $\mathbf{K}^n$ dan is A de matrix van $\mathcal{A}$ ten opzichte van de basis $\vec{v}, \vec{w}_2, \ldots, \vec{w}_n$.)

Kunnen we eigenvektoren berekenen dan stelt het volgende lemma ons in staat konstruktief vektoren in $\mathbf{W}$ aan te geven.

(4.6.2) Lemma. *Zij A_0 en μ als boven. Er is een eigenvektor $\vec{w}$ van A_0^* zodat $A_0^*(\vec{w}) = \bar{\mu}\vec{w}$.*
(Omdat $\vec{w}^*A_0 = \mu\vec{w}^*$ wordt $\vec{w}$ ook wel *links eigenvektor* genoemd.)
Verder is $\mathbf{W} = \mathrm{Im}\,(A_0 - \mu I) = \{\vec{x} \in \mathbf{K}^n \,\big|\, (\vec{x}, \vec{w}) = 0\}$.

Bewijs. Omdat $\Psi_{A_0^*} = \bar{\Psi}_{A_0}$ volgt dat $\bar{\mu} \in \sigma(A_0^*)$.
Als $\vec{x} = (A_0 - \mu I)(\vec{y})$ dan $(\vec{x}, \vec{w}) = (\vec{y}, (A_0^* - \bar{\mu}I)(\vec{w})) = 0$. □

(4.6.3) Opgave. Beschouw de situatie in (4.6.1) en (4.6.2).
Definieer $P \in \mathbf{M}_n$ door $P(\vec{x}\,) := \vec{x} - \frac{(\vec{x},\vec{w})}{(\vec{v},\vec{w})}\vec{v}$ voor $\vec{x} \in \mathbf{K}^n$: $P = I - \frac{1}{\vec{w}^*\vec{v}}\vec{v}\vec{w}^*$.
Bewijs de volgende beweringen.
(a) P is een projektor op $\mathbf{W}$
(dus $P^2 = P$ en $P(\mathbf{K}^n) \subseteq \mathbf{W}$; P is niet noodzakelijk orthonormaal!).
(b) $A_0P = PA_0 = A_0 - \frac{\mu}{\vec{w}^*\vec{v}}\vec{v}\vec{w}^*$.
(c) A_0 en A_0P hebben dezelfde eigenvektoren. $\sigma(A_0P) = (\sigma(A_0)\backslash\{\mu\}) \cup \{0\}$.
(d) $V^{-1}A_0PV = \begin{bmatrix} 0 & \vec{0}^* \\ \vec{0} & B \end{bmatrix}$ en $V^{-1}PV = I - e_1e_1^*$.
De techniek in (b) in feite om de matrix A_0 te projekteren op een kleinere ruimte met behoud van de eigenvektoren (en eigenwaarden) wordt *deflatie* genoemd.

(4.6.4) Opgave. Als $A_0^* = A_0$ dan $\vec{v} = \vec{w}$ en P is orthonormaal. Ga dit na.

(4.6.5) Opgave. Schaal $\vec{v}$ zo dat $\|\vec{v}\,\|_2 = 1$. Laat θ de hoek tussen $\vec{v}$ en $\mathbf{W}$ zijn (zie (2.3.16)).
Kies de links eigenvektor $\vec{w}$ zo dat $\|\vec{w}\,\|_2 = 1$ en $(\vec{v}, \vec{w}\,) \in [0, \infty)$.
Ga na dat $\sin\theta = (\vec{v}, \vec{w}\,)$. Bewijs dat $\mathcal{C}_2(V) = \cot\frac{1}{2}\theta$.
(Hint: kies $\vec{w}_2 := P(\vec{v}\,)/\|P(\vec{v}\,)\|_2$. Dan $\vec{v} = \sin\theta\vec{w} + \cos\theta\vec{w}_2$. Schrijf V ten opzichte van de orthonormale basis $\vec{w}, \vec{w}_2, \ldots, \vec{w}_n$.)

(4.6.6) Zij $R_0 \in \mathbf{M}_n$. Partitioneer $R := V^{-1}R_0V$ als volgt

$$R = \begin{bmatrix} \gamma & \vec{f}^* \\ \vec{g} & E \end{bmatrix} \quad \text{met} \quad \gamma \in \mathbf{C}, \quad \vec{f}, \vec{g} \in \mathbf{K}^{n-1} \quad \text{en} \quad E \in \mathbf{M}_{n-1}.$$

We zijn geïnteresseerd in de eigenwaarde van A_0+R_0 in de buurt van μ en de bijbehorende eigenvektor (in de buurt van $\vec{v}$) of equivalent hiermee de eigenwaarde van $A+R$ bij μ met eigenvektor bij e_1.

Stelling. *Zij* $\varrho := \|(B-\mu)^{-1}\|_2^{-1}$ *en* $\tilde{\varepsilon} := \max(\|E\|_2, \|\vec{f}\|_2, \|\vec{g}\|_2, |\gamma|)$. *Stel dat* $\tilde{\varepsilon} < \varrho$.
(a) *Als* $\zeta \in \mathbf{C}$, $|\zeta-\mu| < \varrho - \tilde{\varepsilon}$ *dan* $\zeta \notin \sigma(B+E)$ *en* $\|(B+E-\zeta I)^{-1}\|_2 \leq \frac{1}{\varrho-\tilde{\varepsilon}-|\zeta-\mu|}$.
(b) *Beschouw een* $\lambda \in \mathbf{C}\backslash\sigma(B+E)$. $\lambda \in \sigma(A+R)$ *dan en slechts dan als*

$$\lambda = \mu + \gamma - ((B+E-\lambda I)^{-1}\vec{g}, \vec{f}) \tag{19}$$

Als $\lambda \in \sigma(A+R)$ *dan* $(A+R)\vec{z} = \lambda\vec{z}$ *met* $\vec{z} := \begin{bmatrix} 1 \\ -(B+E-\lambda I)^{-1}\vec{g} \end{bmatrix}$.
(c) *Stel* $4\tilde{\varepsilon} < \varrho$. *Dan bestaat er een* $\lambda \in \sigma(A+R) \cap \mathcal{D}_{\frac{1}{2}\varrho-\tilde{\varepsilon}}(\mu+\gamma)$.
Laat φ *de hoek zijn tussen* e_1 *en* $\vec{z}$. *Dan is*

$$|\lambda-\mu-\gamma| \leq \frac{\alpha_0\tilde{\varepsilon}}{1-2\alpha_0^2} \quad \textit{en} \quad \tan\varphi \leq \frac{\alpha_0}{1-2\alpha_0^2} \quad \textit{waarbij} \quad \alpha_0 := \frac{\tilde{\varepsilon}}{\varrho-2\tilde{\varepsilon}}.$$

(Als $\tilde{\varepsilon} \ll \varrho$ dan $|\lambda-\mu-\gamma| \lesssim \frac{\tilde{\varepsilon}^2}{\varrho}$ en $\tan\varphi \lesssim \frac{\tilde{\varepsilon}}{\varrho}$.)

Bewijs. (a) $B+E-\lambda I = (B-\mu I)(I-(B-\mu I)^{-1}(\lambda I - \mu I - E))$ is inverteerbaar als $\|(B-\mu I)^{-1}(\lambda I-\mu I-E))\|_2 < 1$. De norm afschatting laten we aan de lezer over.
(b) Stel dat $\lambda \in \sigma(A+R)$ met eigenvektor $\vec{z} \neq 0$.
Schrijf $\vec{z} = \begin{bmatrix} \nu \\ \vec{z}_0 \end{bmatrix}$ met $\nu \in \mathbf{C}$ en $\vec{z}_0 \in \mathbf{C}^{n-1}$. Dan geldt

$$\left(\begin{bmatrix} \mu & \vec{0}^* \\ \vec{0} & B \end{bmatrix} + \begin{bmatrix} \gamma & \vec{f}^* \\ \vec{g} & E \end{bmatrix}\right)\begin{bmatrix} \nu \\ \vec{z}_0 \end{bmatrix} = \lambda \begin{bmatrix} \nu \\ \vec{z}_0 \end{bmatrix}.$$

We schrijven de vergelijking voor de eerste koördinaat uit en evenzo voor de laatste $n-1$-koördinaten. Als $\nu = 0$ dan blijkt $(B+E-\lambda I)\vec{z}_0 = \vec{0}$ en dus $\vec{z}_0 = \vec{0}$; immers $\lambda \notin \sigma(B+E)$. Blijkbaar $\nu \neq 0$. We mogen dus aannemen dat $\nu = 1$ (anders bekijk $\frac{1}{\nu}\vec{z}$ i.p.v. $\vec{z}$). Los nu $\vec{z}_0$ uit de laatste vergelijking en vul het resultaat in, in de eerste. Dit levert achtereenvolgens

$$\begin{array}{ll} \lambda = \mu + \gamma + (\vec{z}_0, \vec{f}) & \qquad \lambda = \mu+\gamma - ((B-\lambda I+E)^{-1}\vec{g}, \vec{f}) \\ (B-\lambda I+E)\vec{z}_0 = -\vec{g} & \qquad (B-\lambda I+E)\vec{z}_0 = -\vec{g}. \end{array}$$

Het bewijs van de andere bewering in (b) laten we aan de lezer over.
(c) We definiëren eerst een funktie ψ die nuttig is bij de afschattingen. Vervolgens bewijzen we in (i) de schattingen onder de aanname dat er zo'n λ bestaat. Pas in de laatste stap (ii) bewijzen we dat zo'n λ inderdaad bestaat.

Beschouw de funktie ψ op $[0, \varrho-2\tilde{\varepsilon})$ die gedefinieerd is door $\psi(t) := \frac{\tilde{\varepsilon}^2}{\varrho-2\tilde{\varepsilon}-t}$ $(t \in [0, \varrho-2\tilde{\varepsilon})\,)$.
Er is precies één $t_0 \in (0, \frac{1}{2}\varrho - \tilde{\varepsilon})$ waarvoor $\psi(t_0) = t_0$.
(Schets, om dit te bewijzen, in een prent de grafiek van ψ en van $t \rightsquigarrow t$ en merk op dat $\psi(0) > 0$, $\psi(\frac{1}{2}\varrho-\tilde{\varepsilon}) < \frac{1}{2}\varrho - \tilde{\varepsilon}$ en $\psi'' > 0$.)
Definieer $t_1 := \frac{\alpha_0\tilde{\varepsilon}}{1-2\alpha_0^2}$. Dan

$$\text{voor iedere } t \in [0, \tfrac{1}{2}\varrho-\tilde{\varepsilon}] \quad \text{met} \quad t \leq \psi(t) \quad \text{geldt} \quad t \leq t_0 < t_1 < \tfrac{1}{2}\varrho - \tilde{\varepsilon}.$$

(Beschouw, om $t_0 < t_1$ te bewijzen, $p(t) := \alpha_0\tilde{\varepsilon} + 2\alpha_0^2 t$. Dan $p(0) = \psi(0)$, $p(\frac{1}{2}\varrho-\tilde{\varepsilon}) = \psi(\frac{1}{2}\varrho-\tilde{\varepsilon})$ en $p > \psi$ op $(0, \frac{1}{2}\varrho-\tilde{\varepsilon})$. Omdat $p(t_1) = t_1$ is $t_0 < t_1$.)

(i) Stel dat er een $\lambda \in \sigma(A+R) \cap \mathcal{D}_{\frac{1}{2}\varrho-\tilde{\varepsilon}}(\mu+\gamma)$ is. Schrijf $\beta := |\lambda-\mu-\gamma|$.
Dan $|\lambda-\mu| \leq \beta + \tilde{\varepsilon} \leq \frac{1}{2}\varrho < \varrho - \tilde{\varepsilon}$. Dus volgens (b) voldoet λ aan (19) en volgens (a) is

$$\begin{array}{rcl} \beta & \leq & |((B+E-\lambda I)^{-1}\vec{g}, \vec{f})| \leq \|(B+E-\lambda I)^{-1}\|_2\,\|\vec{g}\|_2\,\|\vec{f}\|_2 \\ & \leq & \frac{\tilde{\varepsilon}^2}{\varrho-\tilde{\varepsilon}-|\lambda-\mu|} \leq \frac{\tilde{\varepsilon}^2}{\varrho-2\tilde{\varepsilon}-\beta} = \psi(\beta). \end{array}$$

Blijkbaar

$$\beta = |\lambda-\mu-\gamma| \leq t_0 < t_1 < \frac{1}{2}\varrho - \tilde{\varepsilon}. \tag{20}$$

(ii) Beschouw de kontinue eigenwaarde funktie $\lambda(\cdot)$ van $\tau \rightsquigarrow A + \tau R$ waarvoor $\lambda(0) = \mu$.

Dan $\lambda(1) \in \sigma(A+R)$. Stel dat $\lambda(1) \notin \mathcal{D}_{\frac{1}{2}\varrho-\tilde{\varepsilon}}(\mu+\gamma)$.
Dan $|\lambda(\tau)-\mu-\tau\gamma| = \frac{1}{2}\varrho - \tau\tilde{\varepsilon}$ voor zekere $\tau \in [0,1)$ (ga dit na). Dus, volgens (20) (pas (i) toe met $\tau\tilde{\varepsilon}$ i.p.v. $\tilde{\varepsilon}$, τR i.p.v. R, etc.) geldt $|\lambda(\tau)-\mu-\tau\gamma| < \frac{1}{2}\varrho - \tau\tilde{\varepsilon}$, hetgeen een tegenspraak is met de keuze van τ. Blijkbaar is $\lambda(1) \in \sigma(A+R) \cap \mathcal{D}_{\frac{1}{2}\varrho-\tilde{\varepsilon}}(\mu+\gamma)$. □

(4.6.7) Opmerking. ϱ is in zekere zin de afstand tussen de enkelvoudige eigenwaarde μ van A en de overige eigenwaarden van A.
(i) Stel dat A_0 diagonalizeerbaar is: $A_0 = V^{-1}AV$ met $A = \operatorname{diag}(\mu, \mu_2, \ldots, \mu_n)$.
Dan $B = \operatorname{diag}(\mu_2, \ldots, \mu_n)$ en $\varrho = \min_{i=2,\ldots,n} |\mu_i - \mu|$.
(ii) Als $B = W^{-1}\operatorname{diag}(\mu_2, \ldots, \mu_n)W$ voor zekere niet-singuliere $W \in \mathbf{M}_{n-1}$
dan $\frac{1}{\mathcal{C}(W)} \min_{i=2,\ldots,n} |\mu_i - \mu| \le \varrho \le \mathcal{C}(W) \min_{i=2,\ldots,n} |\mu_i - \mu|$ (ga dit na).

(4.6.8) Opmerking. λ kan geschreven worden als een machtreeks in termen van R en A. Om in te zien hoe dit in zijn werk gaat is het inzichtelijker de kontinue eigenfunktie $\tau \rightsquigarrow \lambda(\tau)$ van $\tau \rightsquigarrow A + \tau R$ met $\lambda(0) = \mu$ te bekijken.
Stel $\qquad \lambda(\tau) = \alpha_0 + \alpha_1\tau + \alpha_2\tau^2 + \alpha_3\tau^3 + \ldots$.
Omdat $\qquad \lambda(\tau) = \mu + \tau\gamma - \tau^2((B + \tau E - \lambda(\tau)I)^{-1}\vec{g}, \vec{f}\,)$,
levert uitschrijven van $(B+\tau E - \lambda(\tau)I)^{-1} = (I - (B-\mu I)^{-1}(\lambda(\tau)I - \mu I - \tau E))^{-1}(B-\mu I)^{-1}$ in een Neumann reeks, uitwerken van de machten van de matrices, en gelijkstellen van de koëfficiënten bij τ^k links en rechts van het gelijkteken een betrekking tussen α_k en α_j voor $j \le k-2$. Middels deze betrekking kunnen we de α_k voor iedere k berekenen.
(Ga na dat $\alpha_0 = \mu$, $\alpha_1 = \gamma$, $\alpha_2 = -((B-\mu I)^{-1}\vec{g}, \vec{f}\,)$. Bereken α_3. Dus

$$\lambda(\tau) = \mu + \tau\gamma - \tau^2((B-\mu I)^{-1}\vec{g}, \vec{f}\,) + \mathcal{O}(\tau^3) \quad \text{voor} \quad \tau \to 0.\ )$$

Deze aanpak levert ook een machtreeks voor $\vec{z}_0(\tau) := -(B + \tau E - \lambda(\tau)I)^{-1}\vec{g}$ (zie dit in).
De machtreeksen aanpak levert een $\mathcal{O}(\tau^3)$. De aanpak in stelling (4.6.6) levert een harde bovengrens.

(4.6.9) Opmerking. We gaven uitspraken in termen van de 2-norm op R. Ten koste van een faktor als $\mathcal{C}(V)$ kan men de schattingen in (4.6.6) vertalen naar schattingen in termen van de 2-norm op R_0 (vergelijk met (ii) in (4.6.7)). Ten koste van faktoren als $\sqrt{n}$ kan men ook schattingen geven in termen van $\|R_0\|_p$ voor $p \in [1,\infty]$. We geven die verder niet.

Als A hermiets is komen in (19) alleen reële termen voor. We kunnen dan bijvoorbeeld met behulp van de doorlopendheidsstelling rechtstreeks bewijzen dat (19) een oplossing λ heeft; we hoeven niet een kontinue eigenwaarde funktie $\lambda(\cdot)$ te gebruiken. In het hermietse geval krijgen we ook enigszinds betere schattingen onder wat mildere voorwaarden op de grootte van de perturbatie (voor het geval dat $\tilde{\varepsilon} \not\ll \varrho$). We geven een stelling voor de perturbatie op de eigenwaarde. De uitspraak voor de perturbatie op de eigenvektor stellen we nog even uit tot stelling (4.6.25) en gevolg (4.6.27).

(4.6.10) Lemma. *Beschouw een hermietse $H \in \mathbf{M}_n$ en een $\vec{x} \in \mathbf{C}^n \backslash \{0\}$. Zij $(c,d) \subset \mathbf{R}\backslash\sigma(H)$. De funktie $t \rightsquigarrow ((H-tI)^{-1}\vec{x}, \vec{x}\,)$ is strikt stijgend op (c,d).*

Bewijs. Zij $Q \in \mathbf{M}_n$ unitair zodat $H = QDQ^*$ met $D = \operatorname{diag}(\lambda_1(H), \ldots, \lambda_n(H))$.
Schrijf $Q^*\vec{x} = (\alpha_1, \ldots, \alpha_n)^T$. Dan

$$((H-tI)^{-1}\vec{x}, \vec{x}) = ((D-tI)^{-1}Q^*\vec{x}, Q^*\vec{x}\,) = \sum_{j=1}^{n} \frac{|\alpha_j|^2}{\lambda_j(H) - t}.$$

De bewering volgt nu uit het feit dat $t \rightsquigarrow \frac{1}{\lambda_j(H)-t}$ strikt stijgend is op $(-\infty, \lambda_j(H))$ en op $(\lambda_j(H), \infty)$.
(Alternatief bewijs:
met $\phi(t) := ((H-tI)^{-1}\vec{x}, \vec{x}\,)$ is $\phi'(t) = ((H-tI)^{-2}\vec{x}, \vec{x}\,) = \|(H-tI)^{-1}\vec{x}\,\|_2 > 0$.) □

(4.6.11) Stelling. *Stel dat* $A_0^* = A_0$, $A_0\vec{v} = \mu\vec{v}$ *met* $\|\vec{v}\|_2 = 1$ *en* μ *enkelvoudig. Zij* $\varrho := \min\{|\mu - \eta| \,\big|\, \eta \in \sigma(A_0)\backslash\{\mu\}\}$. *Verder is* R_0 *zodat* $R_0^* = R_0$ *en* $\varepsilon := \|R_0\|_2$. *Als* $2\varepsilon < \varrho$ *dan bestaat er een* $\lambda \in \sigma(A_0 + R_0) \cap \mathcal{D}_{\varrho-\varepsilon}(\mu)$. *Voor deze* λ *geldt*

$$|\lambda - \mu - \gamma| \leq \frac{\varepsilon^2}{\varrho - 2\varepsilon} \quad \textit{met} \quad \gamma = (R_0\vec{v}_1, \vec{v}_1).$$

Bewijs. We kunnen nu V unitair nemen (evenals de W in (ii) van (4.6.7)).

$B^* = B$, $\gamma \in \mathbf{R}$, en $\vec{g} = \vec{f}$. Merk op dat $\tilde{\varepsilon} = \max(\|E\|_2, \|\vec{g}\|_2, |\gamma|) \leq \varepsilon$ (zie (4.6.6)).

Zij $J := (\mu - \varrho + \varepsilon, \mu + \varrho - \varepsilon)$ zodat $J \cap \sigma(B + E) = \emptyset$ (uit (a) van (4.6.6)).

Beschouw $\phi(t) := \mu + \gamma - ((B + E - tI)^{-1}\vec{g}, \vec{g})$ voor $t \in J$.

ϕ is volgens (4.6.10) strikt dalend. Verder is $|\phi(\mu + \gamma) - \mu - \gamma| \leq \frac{\|\vec{g}\|_2^2}{\varrho - \|E - \gamma\|_2} \leq \frac{\varepsilon^2}{\varrho - 2\varepsilon}$.

Volgens de stelling van Weyl is er een $\lambda \in \sigma(A + R)$ met $|\lambda - \mu| \leq \|R\|_2 = \varepsilon$.

Als $2\varepsilon < \varrho$ dan is $\lambda \in J$ en is volgens (b) van (4.6.6) $\lambda = \phi(\lambda)$.

Stel $\mu + \gamma \leq \lambda$. Omdat ϕ dalend is volgt $\mu + \gamma \leq \lambda = \phi(\lambda) \leq \phi(\mu + \gamma) \leq \mu + \gamma + \frac{\varepsilon^2}{\varrho - 2\varepsilon}$.

Zo'n zelfde uitspraak geldt voor het geval $\lambda \leq \mu + \gamma$. □

(4.6.12) Opgave. Ga na dat de ϕ in het bewijs van (4.6.11) ook kontinu is. Bewijs nu (in geval $3\varepsilon \leq \varrho$) met behulp van de doorlopendheidsstelling (zonder gebruik te maken van de stelling van Weyl) dat er een $\lambda \in J$ is waarvoor $\lambda = \phi(\lambda)$.

Als in (4.6.6) $B = \mathrm{diag}(\mu_2, \ldots, \mu_n)$, $E = 0$ en alles reëel is kunnen we grafisch (19) analyseren met behulp van de grafiek van $t \rightsquigarrow \mu + \gamma + \sum_{j=2}^n \frac{\beta_j\gamma_j}{\mu_j - t}$ en van $t \rightsquigarrow t$ (hier is $\vec{f} = \sum \beta_j e_j$ en $\vec{g} = \sum \gamma_j e_j$). In de volgende opgave komen we zo'n soort situatie tegen.

(4.6.13) Opgave. Zij $A_0 \in \mathbf{M}_n(\mathbf{K})$ hermiets. Laat R_0 een hermietse rang 1 perturbatie zijn, d.w.z: $R_0 = \gamma\vec{p}\,\vec{p}^*$ voor zekere $\gamma \in \mathbf{R}$ en $\vec{p} \in \mathbf{K}^n$ met $\|\vec{p}\|_2 = 1$.

(a) Toon aan dat er een unitaire matrix $V \in \mathbf{M}_n$ is zodat
$R := V^*R_0V = \begin{bmatrix} \gamma & \vec{0}^* \\ \vec{0} & 0 \end{bmatrix}$. Schrijf $A := V^*A_0V = \begin{bmatrix} \mu & \tilde{g}^* \\ \tilde{g} & B \end{bmatrix}$

(b) $B = \tilde{W}D\tilde{W}^*$ voor zekere unitaire $\tilde{W} \in \mathbf{M}_{n-1}$ en diagonaal $D = \mathrm{diag}(\mu_2, \ldots, \mu_n)$. Toon dit aan en ga na dat $W^*V^*(A_0 + R_0)VW = \begin{bmatrix} \mu & \vec{g}^* \\ \vec{g} & D \end{bmatrix} + \begin{bmatrix} \gamma & \vec{0}^* \\ \vec{0} & 0 \end{bmatrix}$ met $W := \begin{bmatrix} 1 & \vec{0}^* \\ \vec{0} & \tilde{W} \end{bmatrix}$.

(c) Schrijf $\vec{g} = \sum_{j=2}^n \gamma_j e_j$. Laat zien dat voor iedere eigenwaarde λ van $A_0 + R_0$ geldt

$$\lambda \in \sigma(B) \quad \text{óf} \quad \lambda - \mu - \gamma = \phi(\lambda) \quad \text{met} \quad \phi(t) := \sum_{j=2}^n \frac{|\gamma_j|^2}{\mu_j - t}.$$

Schets de grafiek van ϕ en van $t \rightsquigarrow t - \mu - \gamma$. Ga na hoe de snijpunten van de grafieken veranderen als γ verandert.

Als de eigenwaarden zo genummerd zijn dat $\lambda_j(A_0) \leq \lambda_{j+1}(A_0)$ en $\lambda_j(A_0+R_0) \leq \lambda_{j+1}(A_0+R_0)$ alle $j = 1, \ldots, n-1$, dan

$\lambda_1(A_0) \leq \lambda_1(A_0+R_0) \leq \lambda_2(A_0) \leq \lambda_2(A_0+R_0) \leq \ldots \leq \lambda_{n-1}(A_0+R_0) \leq \lambda_n(A_0) \leq \lambda_n(A_0+R_0)$ als $\gamma \geq 0$.

Bewijs dit. (• Alternatief bewijs: Uit de stelling van Weyl volgt $\lambda_j(A_0) \leq \lambda_j(A_0 + R_0)$, uit de stelling van Courant–Fischer als in het bewijs van de scheidingsstelling van Cauchy volgt $\lambda_j(A_0 + R_0) \leq \lambda_{j+1}(A_0)$.)

Heeft $A_0 + R_0$ een eigenwaarde in $\sigma(B)$ als B $n-1$-verschillende eigenwaarden heeft?

Wat gebeurt er als B een dubbele eigenwaarde heeft?

• (d) Geef met behulp van bovenstaande argumenten een bewijs voor het gevolg (4.4.6).

De eigenruimte van hermietse matrices

Als $A \in \mathbf{M}_n$ hermiets is en de perturbatie H ook dan is er een basis van eigenvektoren voor $A + \tau H$ die analytisch van de perturbatie parameter τ afhangt (zie (4.3.7)). Als in (c) van stelling (4.6.6) wensen we de grootte van de perturbatie van een eigenvektor te schatten in termen van A en de *grootte* van de perturbatie τH op A. Het volgende voorbeeld laat zien dat dit voor een eigenvektor bij een meervoudige eigenwaarde geen kans van slagen heeft. Een eigenvektor bij meervoudige eigenwaarde hangt alleen kontinu af van de perturbatie voor een specifieke keuze van eigenvektoren. Die keuze hangt niet alleen af van de grootte maar ook van de vorm van de hermietse perturbatie.

(4.6.14) Voorbeeld. Zij $A \in \mathbf{M}_2$ de triviale matrix.
Beschouw $H_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ en $H_2 := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. $\frac{1}{2}\sqrt{2}(e_1 + e_2)$ (en $\frac{1}{2}\sqrt{2}(e_1 - e_2)$) is voor iedere τ een eigenvektor van $A + \tau H_1$, terwijl e_1 (en e_2) de eigenvektor is van $A + \tau H_2$. In beide gevallen is $\|\tau H_1\|_2 = \|\tau H_2\|_2 = |\tau|$. We kunnen echter voor A geen eigenvektoren kiezen zodat voor *iedere* hermietse perturbatie met norm $|\tau|$ een geperturbeerde eigenvektor een kleine hoek maakt met deze gekozen ongeperturbeerde eigenvektor.

We kunnen wel de hoek schatten tussen de *eigenruimte* bij een (eventueel) meervoudige eigenwaarde μ en het opspansel van de eigenvektoren bij *alle* eigenwaarden van het geperturbeerd stelsel die in de buurt liggen van μ.
Onze uitspraken zijn wat algemener: als $[\alpha, \beta] \subset \mathbf{R}$ en

$$\mathbf{V} := \bigoplus_{\mu \in [\alpha,\beta]} \text{Kern}\,(A - \mu I) := \text{span}\,\{\vec{v} \in \mathbf{C}^n \,\big|\, \vec{v} \in \text{Kern}\,(A - \mu I) \text{ voor zekere } \mu \in [\alpha, \beta] \cap \sigma(A)\}$$

$$\mathbf{W} := \bigoplus_{\mu \in [\alpha - \varepsilon, \beta + \varepsilon]} \text{Kern}\,(A + H - \mu I)$$

wensen we (als α, β en ε aan bepaalde voorwaarden voldoen) de hoek tussen $\mathbf{V}$ en $\mathbf{W}$ te schatten. Voor dat we dit kunnen doen moeten we eerst de hoek tussen twee deelruimten definiëren en karakterizeren.

(4.6.15) Definitie. Laat $\mathbf{V}$ en $\mathbf{W}$ lineaire deelruimten van $\mathbf{C}^n$ zijn met $\dim\mathbf{V} \leq \dim\mathbf{W}$.
Als $\vec{v} \in \mathbf{C}^n$ dan geven we de hoek tussen $\vec{v}$ en $\mathbf{W}$ aan met $\angle(\vec{v}, \mathbf{W})$
(zie (2.3.16) voor de definitie van de hoek tussen een vektor en een lineaire deelruimte).
De *hoek* $\angle(\mathbf{V}, \mathbf{W})$ tussen $\mathbf{V}$ en $\mathbf{W}$ is gedefiniëerd door

$$\angle(\mathbf{V}, \mathbf{W}) := \max\{\angle(\vec{v}, \mathbf{W}) \,\big|\, \vec{v} \in \mathbf{V}, \vec{v} \neq 0\} \; (= \max_{\vec{v} \in \mathbf{V}\setminus\{0\}} \; \min_{\vec{w} \in \mathbf{W}\setminus\{0\}} \angle(\vec{v}, \vec{w}\,)\,).$$

We laten in (4.6.19) zien dat $\angle(\mathbf{V}, \mathbf{W}) = \angle(\mathbf{W}, \mathbf{V})$ als $\dim\mathbf{V} = \dim\mathbf{W}$. We gebruiken daarbij lemma (4.6.17).

In (4.6.19) karakterizeren we de hoek tussen $\mathbf{V}$ en $\mathbf{W}$ ondermeer in termen van de orthonormale projektors uit definitie (4.6.18).

(4.6.16) Opgave. (a) Bepaal voor $\mathbf{W} := \{\alpha_1 e_1 + \alpha_2 e_2 \,\big|\, \alpha_i \in \mathbf{C}\}$ en $\mathbf{V} := \{\alpha(e_1 + e_3) \,\big|\, \alpha \in \mathbf{C}\}$ in $\mathbf{C}^3$ de hoek tussen $\mathbf{V}$ en $\mathbf{W}$. Bepaal ook $\max\{\angle(\vec{w}, \mathbf{V}) \,\big|\, \vec{w} \in \mathbf{W}\}$.
(b) Schrijf $\mathbf{V}^\perp := \{\vec{x} \in \mathbf{C}^n \,\big|\, \vec{x} \perp \mathbf{V}\}$. Als $\dim\mathbf{V} < \dim\mathbf{W}$ dan $\mathbf{V}^\perp \cap \mathbf{W} \neq \{0\}$ en dus $\max\{\angle(\vec{w}, \mathbf{V}) \,\big|\, \vec{w} \in \mathbf{W}\} = \frac{1}{2}\pi$. Bewijs dit.

(4.6.17) Lemma. *Zij $Q \in \mathbf{M}_n$ unitair.*
Met $E \in \mathbf{M}_k$, $H \in \mathbf{M}_{n-k}$ is Q gepartitioneerd als $Q = \begin{bmatrix} E & F \\ G & H \end{bmatrix}$. Dan $\|G\|_2 = \|F\|_2$.

Bewijs. Merk op dat

$$I_n = \begin{bmatrix} E^* & G^* \\ F^* & H^* \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} E^*E + G^*G & E^*F + G^*H \\ F^*E + H^*G & F^*F + H^*H \end{bmatrix}.$$

Dus $I_k = E^*E + G^*G$. Omdat ook $I_n = QQ^*$ volgt ook dat $I_k = EE^* + FF^*$ (ga dit na). Blijkbaar $\sigma(G^*G) = \sigma(I_k - E^*E) = 1 - \sigma(E^*E)$ en $1 - \sigma(EE^*) = \sigma(I_k - EE^*) = \sigma(FF^*)$. Omdat $\sigma(E^*E) = \sigma(EE^*)$ is $\|G\|_2^2 = \rho(G^*G) = \rho(FF^*) = \|F^*\|_2^2 = \|F\|_2^2$.
(Is het lemma ook korrekt als $E \in \mathbf{M}_{k\times l}$ en $l \neq k$?) □

(4.6.18) Definitie. Als $\mathbf{U}$ een lineaire deelruimte is van $\mathbf{C}^n$ dan is $P_{\mathbf{U}}$ de *orthonormale projektor* van $\mathbf{C}^n$ op $\mathbf{U}$: $P_{\mathbf{U}}(\vec{x}\,) \in \mathbf{U}$ en $\vec{x} - P_{\mathbf{U}}(\vec{x}\,) \perp \mathbf{U}$ voor iedere $\vec{x} \in \mathbf{C}^n$ (zie ook (3.1.5)).

(4.6.19) Lemma.
Laat $\mathbf{V}$ en $\mathbf{W}$ lineaire deelruimten van $\mathbf{C}^n$ zijn met $k := \dim\mathbf{V} \leq l := \dim\mathbf{W}$.
(a) $\|(I - P_{\mathbf{W}})P_{\mathbf{V}}\|_2 = \sin\angle(\mathbf{V},\mathbf{W})$.
(b) *Als $k = l$ dan $\angle(\mathbf{V},\mathbf{W}) = \angle(\mathbf{W},\mathbf{V})$.*
(c) *Zij $V, W \in \mathbf{M}_n$ unitair. Partitioneer V en W volgens*
$V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$ *en* $W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}$ *met $V_1 \in \mathbf{M}_{n\times k}$ en $W_1 \in \mathbf{M}_{n\times l}$.*
Als $\mathbf{V} := V_1(\mathbf{C}^k)$ en $\mathbf{W} := W_1(\mathbf{C}^l)$ dan $\sin\angle(\mathbf{V},\mathbf{W}) = \|W_2^*V_1\|_2$.

Bewijs. We mogen aannemen dat $\mathbf{V}$ en $\mathbf{W}$ optreden als beeld van orthonormale matrices V_1, respektievelijk W_1 als in (c) (in de volgende opgave vragen we de lezer te bewijzen dat deze aanname korrekt is).

Als $\vec{v} \in \mathbf{C}^n$ dan is $\vec{v}$ op precies een manier te schrijven als $\vec{v} = \vec{v}\,' + \vec{v}\,''$ met $\vec{v}\,' \in \mathbf{W}$ en $\vec{v}\,'' \perp \mathbf{W}$ (zie (3.1.4)). Dan is $\sin\angle(\vec{v},\mathbf{W}) = \frac{\|\vec{v}\,''\|_2}{\|\vec{v}\,\|_2}$.
Laat P de orthonormale afbeelding van $\mathbf{V}$ op $\mathbf{W}^\perp$ zijn: $P(\vec{v}\,) = \vec{v}\,''$ $(\vec{v} \in \mathbf{V})$. Dan

$$\sin\angle(\mathbf{V},\mathbf{W}) = \max\{\sin\angle(\vec{v},\mathbf{W}) \,\big|\, \vec{v} \in \mathbf{V}\} = \max\left\{\frac{\|\vec{v}\,''\|_2}{\|\vec{v}\,\|_2} \,\Big|\, \vec{v} \in \mathbf{V}\right\} = \|P\|_2.$$

Bovendien is P de restriktie van $P_{\mathbf{W}^\perp} = I - P_{\mathbf{W}}$ tot $\mathbf{V}$ en $\|P\|_2 = \|(I - P_{\mathbf{W}})P_{\mathbf{V}}\|_2$ (ga dit na). Dit bewijst (a).

Om (c) te bewijzen schrijven we $\vec{v}_j := Ve_j$ en $\vec{w}_j := We_j$.
Als $\vec{v} = \sum_{j=1}^k \alpha_j\vec{v}_j \in \mathbf{V}$ dan $P\vec{v} = \sum_{j=l+1}^n (\vec{v},\vec{w}_j)\vec{w}_j = \sum_{j=l+1}^n \sum_{i=1}^k \alpha_i(\vec{v}_i,\vec{w}_j)\vec{w}_j$. Blijkbaar is $W_2^*V_1$ de matrix van P ten opzichte van de orthonormale basis $\vec{v}_1,\ldots,\vec{v}_k$ van $\mathbf{V}$ en $\vec{w}_{l+1},\ldots,\vec{w}_n$ van $\mathbf{W}^\perp$ en $\|P\|_2 = \|W_2^*V_1\|_2$. Dit bewijst (c).

Passen we het voorgaande lemma toe met $Q = W^*V$ en $E = W_1^*V_1$ dan zien we dat $\|W_2^*V_1\|_2 = \|W_1^*V_2\|_2$ in geval $k = l$. Omdat we zojuist in (c) bewezen hebben dat $\sin\angle(\mathbf{W},\mathbf{V}) = \|V_2^*W_1\|_2$ (verwissel de rollen van $\mathbf{V}$ en $\mathbf{W}$) en $\|V_2^*W_1\|_2 = \|(V_2^*W_1)^*\|_2 = \|W_1^*V_2\|_2$ volgt (b). □

(4.6.20) Opgave. Als $\mathbf{V}$ een k-dimensionale lineaire deelruimte is van $\mathbf{C}^n$ dan is $\mathbf{V} = V(\{(\vec{x}^{\,T},\vec{0}^{\,T})^T \in \mathbf{C}^n \,\big|\, \vec{x} \in \mathbf{C}^k\})$ voor zekere unitaire $V \in \mathbf{M}_n$. Bewijs dit.

(4.6.21) Opgave. Beschouw de situatie als in (4.6.19).
(a) Ga na dat $\begin{bmatrix} 0 & W_1^*V_2 \\ W_2^*V_1 & 0 \end{bmatrix}$ de matrix is van $P_{\mathbf{W}} - P_{\mathbf{V}} = P_{\mathbf{W}}(I - P_{\mathbf{V}}) - (I - P_{\mathbf{W}})P_{\mathbf{V}}$ ten opzichte van de basis $\vec{v}_1,\ldots,\vec{v}_n$ in de origineel ruimte en $\vec{w}_1,\ldots,\vec{w}_n$ in de beeld ruimte.
(b) Konludeer dat $\|P_{\mathbf{W}} - P_{\mathbf{V}}\|_2 = \max(\|(I - P_{\mathbf{W}})P_{\mathbf{V}}\|_2, \|(I - P_{\mathbf{V}})P_{\mathbf{W}}\|_2)$.
(c) Als $k = l$ dan $\|P_{\mathbf{W}} - P_{\mathbf{V}}\|_2 = \sin\angle(\mathbf{V},\mathbf{W})$ anders $\|P_{\mathbf{W}} - P_{\mathbf{V}}\|_2 = 1$. Bewijs dat.

$\|P_{\mathbf{W}} - P_{\mathbf{V}}\|_2$ noemt men ook wel de *kloof* ('gap') tussen $\mathbf{V}$ en $\mathbf{W}$.

(4.6.22) Opgave. Beschouw de situatie als in (4.6.19).
Geldt $\cos\angle(\mathbf{V},\mathbf{W}) = \|W_1^*V_1\|$?
(Hint: bedenk een tegenvoorbeeld in $\mathbf{C}^3$ met $\dim\mathbf{V} = \dim\mathbf{W} = 2$.)

De uitspraak over de hoek tussen eigenruimten van de geperturbeerde en ongeperturbeerde hermietse matrix staat in feite in het volgende lemma. We interpreteren het resultaat in termen van perturbaties in de volgende stelling.

(4.6.23) Lemma. *Zij $A, H \in \mathbf{M}_n$ met $A^* = A$ en $H^* = H$. Laat*

$AV_1 = V_1\Lambda_1$ *met* $V_1 \in \mathbf{M}_{n\times k}$ *orthonormaal en* $\Lambda_1 \in \mathbf{M}_k$ *diagonaal en*
$(A+H)W_2 = W_2\tilde{\Lambda}_2$ *met* $W_2 \in \mathbf{M}_{n\times(n-l)}$ *orthonormaal en* $\tilde{\Lambda}_2 \in \mathbf{M}_{n-l}$ *diagonaal.*
Als $\sigma(\Lambda_1) \subset [\alpha,\beta]$ en $\sigma(\tilde{\Lambda}_2) \cap (\alpha-\eta, \beta+\eta) = \emptyset$ voor 'n $\eta > 0$ dan $\|W_2^* V_1\|_2 \le \frac{1}{\eta}\|H\|_2$.

Bewijs. Door links vermenigvuldiging met W_2^* en met V_1^* zien we dat

$$\begin{matrix} W_2^*AV_1 & = & W_2^*V_1\Lambda_1 \\ V_1^*(A+H)W_2 & = & V_1^*W_2\tilde{\Lambda}_2 \end{matrix}. \quad \text{Adjungeren levert} \quad \begin{matrix} W_2^*AV_1 & = & W_2^*V_1\Lambda_1 \\ W_2^*(A+H)V_1 & = & \tilde{\Lambda}_2W_2^*V_1 \end{matrix}.$$

Het verschil van deze laatste twee gelijkheden is $\tilde{\Lambda}_2W_2^*V_1 - W_2^*V_1\Lambda_1 = W_2^*HV_1$.
Voor $\tau \in \mathbf{R}$ is $(\tilde{\Lambda}_2 - \tau I)W_2^*V_1 = W_2^*V_1(\Lambda_1 - \tau I) + W_2^*HV_1$
en dus $W_2^*V_1 = (\tilde{\Lambda}_2 - \tau I)^{-1}(W_2^*V_1(\Lambda_1 - \tau I) + W_2^*HV_1)$ als $\tau \notin \sigma(\tilde{\Lambda}_2)$.
Blijkbaar $\|W_2^*V_1\|_2 \le \|(\tilde{\Lambda}_2 - \tau I)^{-1}\|_2(\|W_2^*V_1\|_2\,\|\Lambda_1 - \tau I\|_2 + \|W_2^*HV_1\|_2)$.
Merk op dat $\|W_2^*HV_1\|_2 \le \|H\|_2$. Als $\|(\tilde{\Lambda}_2 - \tau I)^{-1}\|_2^{-1} - \|\Lambda_1 - \tau I\|_2 > 0$ dan

$$\|W_2^*V_1\|_2 \le \frac{\|H\|_2}{\|(\tilde{\Lambda}_2 - \tau I)^{-1}\|_2^{-1} - \|\Lambda_1 - \tau I\|_2}.$$

Kies nu $\tau = \frac{1}{2}(\alpha+\beta)$ en schrijf $\delta := \frac{1}{2}(\beta - \alpha)$. Dan $\|(\tilde{\Lambda}_2 - \tau I)^{-1}\|_2 \le \frac{1}{\delta+\eta}$, $\|\Lambda_1 - \tau I\|_2 \le \delta$ en het lemma volgt. □

(4.6.24) Opgave. Het bewijs van bovenstaand lemma is een generalizering van het volgende bewijs dat een hermietse matrix een orthogonale basis van eigenvektoren heeft.
Als $A \in \mathbf{M}_n$ hermiets is en $AQ = Q\Lambda$ voor zekere $Q \in \mathbf{M}_n$ en zekere diagonaal $\Lambda \in \mathbf{M}_n$ dan volgt als in het bewijs van bovenstaand lemma dat $\Lambda Q^*Q - Q^*Q\Lambda = 0$. Als $\#\sigma(A) = n$ volgt hieruit dat Q^*Q diagonaal is. Ga dit na.

(4.6.25) Stelling van Davis–Kahan. *Zij $A, H \in \mathbf{M}_n$ met $A^* = A$ en $H^* = H$. $\varepsilon := \|H\|_2$.*
Stel dat $[\alpha,\beta] \subset \mathbf{R}$ en $\varrho > 0$ zodat $\sigma(A) \subset (-\infty, \alpha-\varrho] \cup [\alpha,\beta] \cup [\beta+\varrho, \infty)$.
Zij $\mathbf{V} := \bigoplus_{\mu\in[\alpha,\beta]} \mathrm{Kern}\,(A - \mu I)$ en $\mathbf{W} := \bigoplus_{\mu\in[\alpha-\varepsilon,\beta+\varepsilon]} \mathrm{Kern}\,(A + H - \mu I)$. Dan

$$\sin\angle(\mathbf{V},\mathbf{W}) \le \frac{\varepsilon}{\varrho - \varepsilon} \quad \textit{als} \quad \varepsilon < \varrho.$$

Bewijs. Omdat $\sin\varphi \le 1$ voor iedere $\varphi \in [0, \frac{1}{2}\pi]$ is de stelling triviaal als $\varepsilon < \varrho \le 2\varepsilon$. We mogen dus aannemen dat $2\varepsilon < \varrho$.

Nummer de eigenwaarden zodat

$$\lambda_1(A) \le \lambda_2(A) \le \ldots \le \lambda_n(A) \quad \text{en} \quad \lambda_1(A+H) \le \lambda_2(A+H) \le \ldots \le \lambda_n(A+H).$$

Dan is volgens Weyl's stelling $|\lambda_j(A) - \lambda_j(A+H)| \le \varepsilon$ en zien we dat

$$\begin{matrix} \lambda_j(A) \in [\alpha,\beta] & \Rightarrow & \lambda_j(A+H) \in [\alpha-\varepsilon, \beta+\varepsilon] & \Rightarrow & \lambda_j(A+H) \in (\alpha-\varrho+\varepsilon, \beta+\varrho-\varepsilon) \\ & \Rightarrow & \lambda_j(A) \in (\alpha-\varrho, \beta+\varrho) & \Rightarrow & \lambda_j(A) \in [\alpha,\beta] \end{matrix}.$$

Blijkbaar $\lambda_j(A) \in [\alpha,\beta] \Leftrightarrow \lambda_j(A+H) \in [\alpha-\varepsilon, \beta+\varepsilon]$.

Zij k het aantal eigenwaarden van A geteld naar multipliciteit in $[\alpha,\beta]$.

Zij $AV = V\Lambda$ en $(A+H)W = W\tilde{\Lambda}$ zodat $V, W \in \mathbf{M}_n$ unitair, $\Lambda = \mathrm{diag}\,(\mu_1, \ldots, \mu_n)$ en $\tilde{\Lambda} = \mathrm{diag}\,(\tilde{\mu}_1, \ldots, \tilde{\mu}_n)$, waarbij de eigenwaarden nu zo genummerd zijn dat $\mu_j \in [\alpha,\beta]$ voor $j \le k$ en $\tilde{\mu}_j \in [\alpha-\varepsilon, \beta+\varepsilon]$ voor $j \le k$. Met $\Lambda_1 := \mathrm{diag}\,(\mu_1, \ldots, \mu_k)$ en $\tilde{\Lambda}_2 := \mathrm{diag}\,(\tilde{\mu}_{k+1}, \ldots, \tilde{\mu}_n)$ is $\sigma(\Lambda_1) \subset [\alpha,\beta]$ en $\sigma(\tilde{\Lambda}_2) \cap (\alpha-\varrho+\varepsilon, \beta+\varrho-\varepsilon) = \emptyset$. De bewering volgt nu uit het vorige lemma en (c) van lemma (4.6.19). □

(4.6.26) Opmerking. Als $2\varepsilon < \varrho$ dan mag men volgens de stelling van Weyl de eigenwaarden van $A + H$ in $[\alpha-\varepsilon, \beta+\varepsilon]$ zien als de geperturbeerde eigenwaarden van A in $[\alpha,\beta]$. De eigenwaarden van $A + H$ buiten $[\alpha-\varepsilon, \beta+\varepsilon]$ zijn geen perturbaties van eigenwaarden van A in $[\alpha,\beta]$. Bovendien is $\dim\mathbf{V} = \dim\mathbf{W}$ (zie het bewijs van bovenstaande stelling).

(4.6.27) Gevolg. *Zij* $A, H \in \mathbf{M}_n$ *met* $A^* = A$ *en* $H^* = H$. $\varepsilon := \|H\|_2$.
Zij $\mu \in \sigma(A)$ *en* $\varrho := \min\{|\mu - \eta| \,\big|\, \eta \in \sigma(A)\backslash\{\mu\}\}$. ($\mu$ is eventueel meervoudig!)
Laat $\vec{z} \in \mathbf{C}^n\backslash\{0\}$ *en* $\lambda \in \mathbf{R}$ *zo zijn dat* $(A + H)\vec{z} = \lambda\vec{z}$ *en* $|\lambda - \mu| \leq \varepsilon$.
Als $2\varepsilon < \varrho$ *dan* $\sin \angle(\vec{z}, \mathrm{Kern}\,(A - \mu I)) \leq \frac{\varepsilon}{\varrho - \varepsilon}$. □

5 Eigenwaarden en eigenvektoren: praktijk

5.1 Inleiding

Men kan met behulp van de "a, b, c-formule" de wortels bepalen van het polynoom $a\chi^2 + b\chi + c$. Een resultaat in de Galois theorie zegt echter dat er voor $n > 4$ geen formules bestaan waarmee men in een eindig rekursief proces de wortels van ieder polynoom van graad n kan bepalen uit diens koëfficiënten. Omdat ieder polynoom kan optreden als karakteristiek polynoom zijn er geen formules waarmee we in een eindig rekursief proces voor iedere $n \times n$-matrix uit de matrix koëfficiënten de eigenwaarden kunnen bepalen. Zelfs voor bijvoorbeeld hermietse matrices bestaan er niet zo'n formules. In het algemeen (tenzij we met bijvoorbeeld een bovendriehoeks matrix te maken hebben) zullen we de eigenwaarden via een oneindig iteratief proces moeten benaderen.

Men kan twee type benaderings methoden onderscheiden:
– Een type methode past een nulpunts zoekproces voor funkties van $\mathbf{R}$ naar $\mathbf{R}$ of van $\mathbf{C}$ naar $\mathbf{C}$ toe op het karakteristieke polynoom dat al of niet expliciet bepaald is;
– Het andere type methode werkt rechtstreeks met de matrix.

De Hessenberg vorm

Vaak is het efficiënt om, voordat we met zo'n iteratief proces aan de slag gaan, de matrix A op *Hessenberg vorm* $B = (b_{ij})$ te brengen, d.w.z.: $B = QAQ^*$ voor zekere unitaire Q en $b_{ij} = 0$ als $i > j + 1$. Een *Hessenberg matrix* heeft in de beneden driehoek slechts een niet-triviale ko-diagonaal (een Hessenberg vorm is zo dicht bij de Schur vorm als konstruktief mogelijk is in een eindig proces). A en diens Hessenberg vorm B hebben hetzelfde karakteristieke polynoom. Onthouden we Q dan kunnen we uit de eigenvektoren van B die van A berekenen. We kunnen de Hessenberg vorm stabiel berekenen middels Householder transformaties.

(5.1.1) Stelling. *Zij* $A \in \mathbf{M}_n(\mathbf{K})$.
(a) *Er is een unitaire* $Q \in \mathbf{M}_n(\mathbf{K})$ *zodat* $B := QAQ^*$ *een* Hessenberg matrix *is, d.w.z.: met* $B = (b_{ij})$ *is* $b_{ij} = 0$ *als* $i > j + 1$.
(b) B *kan konstruktief à* $\approx \frac{4}{3}n^3$ *flop rekursief berekend worden middels Householder transformaties* $(H_{\vec{w}_1}, \ldots H_{\vec{w}_{n-2}}$: met $A_1 := A$ en $A_{k+1} = (H_{\vec{w}_k} A_k) H^*_{\vec{w}_k}$ voor $k = 1, \ldots, n-2$ is $B = A_{n-1}$ en $Q = H_{\vec{w}_{n-2}} \cdots H_{\vec{w}_1})$.
Als $B^\star$ *de door de computer berekende* B *is dan is* $B^\star = \tilde{Q}(A_0 + E)\tilde{Q}^*$ *voor zekere unitaire* $\tilde{Q} \in \mathbf{M}_n$ *en* $E \in \mathbf{M}_n$ *met* $\|E\|_F \leq 6n^2\bar{\xi}\,\|B\|_F$.
(c) *Als* $A^* = A$ *dan* $B^* = B$, B *is tri-diagonaal* (dus $b_{ij} = 0$ als $|i - j| > 1$) *en kan à* $\approx \frac{2}{3}n^3$ *flop middels Householder transformaties berekend worden.*

Bewijs. We beschrijven in detail de eerste stap in de konstruktie met Householder transformaties. De overigen stappen en de rest van het bewijs laten we aan de lezer over.

Laat $A_1 := A = (a^{(1)}_{ij})$.
Met $\tilde{a}_1 := (0, a_{21}, \ldots, a_{n1})^T$, $\tau_1 :=$ teken (a_{21}) is $\vec{w}_1 := \tilde{a}_1 + \tau_1\|\tilde{a}_1\|_2 e_2$ berekenen $\tilde{A}_2 := H_{\vec{w}_1} A_1$ en vervolgens $A_2 := \tilde{A}_2 H_{\vec{w}_1}$. Als in (2.5.4) en omdat de matrix $H_{\vec{w}_1}$ van de vorm
$H_{\vec{w}_1} = \begin{bmatrix} 1 & \vec{0}^* \\ \vec{0} & W \end{bmatrix}$ is met $W \in \mathbf{M}_{n-1}$ volgt dat de matrices $\tilde{A}_2$ en A_2 er als volgt uitzien

$$\tilde{A}_2 = \left[\begin{array}{c|ccc} a^{(1)}_{11} & a^{(1)}_{12} & \ldots & a^{(1)}_{1n} \\ \hline -\tau_1\|\tilde{a}_1\|_2 & \tilde{a}^{(2)}_{22} & \ldots & \tilde{a}^{(2)}_{2n} \\ \hline 0 & \tilde{a}^{(2)}_{32} & \ldots & \tilde{a}^{(2)}_{3n} \\ \vdots & \vdots & & \vdots \\ 0 & \tilde{a}^{(2)}_{n2} & \ldots & \tilde{a}^{(2)}_{nn} \end{array}\right] \quad \text{en} \quad A_2 = \left[\begin{array}{c|ccc} a^{(1)}_{11} & a^{(2)}_{12} & \ldots & a^{(2)}_{1n} \\ \hline -\tau_1\|\tilde{a}_1\|_2 & a^{(2)}_{22} & \ldots & a^{(2)}_{2n} \\ \hline 0 & a^{(2)}_{32} & \ldots & a^{(2)}_{3n} \\ \vdots & \vdots & & \vdots \\ 0 & a^{(2)}_{n2} & \ldots & a^{(2)}_{nn} \end{array}\right].$$

(c) Schrijf $\vec{w}_1 = (\omega_i)$. Bereken $A_2 = (I - \mu\vec{w}_1\vec{w}_1^*)A(I - \mu\vec{w}_1\vec{w}_1^*)$ à $\approx 2n^2$ flop als volgt: $\vec{a} = A\vec{w}_1$, $\vec{b} = (\beta_i) = \mu\vec{a}$, $\vec{c} = (\gamma_i) = \mu^2(\vec{a}, \vec{w}_1)\vec{w}_1$ en $a^{(2)}_{ij} = a^{(1)}_{ij} - \omega_i\bar{\beta}_j + (\gamma_i - \beta_i)\bar{\omega}_j$ als $j \leq i$

en $a^{(2)}_{ij} = \bar{a}^{(2)}_{ji}$ als $j > i$. $\square$

(5.1.2) Opgave. Als $B \in \mathbf{M}_n$ een Hessenberg matrix is, hoeveel flop heeft men dan nodig om $\det(\zeta I - B)$ te berekenen? Stel dat men voor een matrix A voor verschillende ζ de determinant $\det(\zeta I - A)$ wil berekenen. Bij welk aantal ζ's wordt het rendabel eerst de Hessenberg vorm van A te bepalen?

Voor het berekenen van de eigenwaarden (en eigenvektoren) van een Hessenberg matrix mogen we volgens de volgende bewering aannemen dat alle koëfficiënten van de beneden kodiagonaal $\neq 0$ zijn.

(5.1.3) Bewering. *Zij $B \in \mathbf{M}_n$. Stel $B = \begin{bmatrix} C & E \\ 0 & F \end{bmatrix}$ is met $C \in \mathbf{M}_k$ voor zekere $k < n$. Dan zijn $\lambda_1(C), \ldots, \lambda_k(C), \lambda_1(F), \ldots, \lambda_{n-k}(F)$ de eigenwaarden van B geteld naar multipliciteit.*

Bewijs. $\Psi_B = \Psi_C \Psi_F$. $\square$

(5.1.4) Opgave. Beschouw de situatie in (5.1.3). Bepaal de eigenvektoren van B bij eigenwaarde λ uit die van C en F in geval $\lambda \notin \sigma(C) \cap \sigma(F)$.
Analyseer ook de situatie in geval $\lambda \in \sigma(C) \cap \sigma(F)$.

Als een koëfficiënt van de beneden ko-diagonaal 0 is kan men blijkbaar het eigenwaarden probleem reduceren tot twee of meer sub-problemen. Het overstappen op de kleinere 'ontkoppelde' matrices noemt men ook wel deflatie.

(5.1.5) Definitie. Een Hessenberg matrix $B = (b_{ij}) \in \mathbf{M}_n$ is *reduceerbaar* als een van de beneden ko-diagonaalelementen 0 is: $b_{i+1\,i} = 0$ voor zekere $i \in \{1, \ldots, n-1\}$. Als de Hessenberg matrix niet reduceerbaar is noemt men hem *irreducibel*.

(5.1.6) Opmerking. In de numerieke praktijk stapt men over op kleinere problemen als in een Hessenberg matrix $B = (b_{ij})$ geldt $|b_{i+1\,i}| \leq \bar{\xi}(|b_{ii}| + |b_{i+1\,i+1}|)$ (B is *numeriek reduceerbaar*): men vervangt dan $b_{i+1\,i}$ door 0.
Ter argumentatie kan men aanvoeren dat men vanwege afrondingen en evaluatiefouten altijd relatieve fouten ter grootte ξb_{ij} moet verwachten.

Een Hessenberg matrix maakt niet alleen efficiënter rekenen mogelijk maar heeft ook een aantal theoretische eigenschappen die voor bepalen van de eigenwaarden en eigenvektoren van belang zijn. In de rest van deze subparagraaf verzamelen we de eigenschappen die we verderop nodig hebben.

(5.1.7) Definitie. Zij $A \in \mathbf{M}_n$ en $\vec{y} \in \mathbf{K}^n$.
Voor $l \in \mathbf{N}$ is $\mathrm{K}^l(A; \vec{y}) := \mathrm{span}(\vec{y}, A\vec{y}, \ldots, A^{l-1}\vec{y})$ een zogenaamde *Krylov ruimte* (voortgebracht door A en $\vec{y}$).

(5.1.8) Lemma.
Laat $B = (b_{ij}) \in \mathbf{M}_n$ nu een irreducibele Hessenberg matrix zijn.
Als $m < n$ en $\vec{y} \in \mathrm{span}(e_1, \ldots, e_{n-m}) \backslash \{0\}$ dan $\dim(\mathrm{K}^{m+1}(B; \vec{y})) = m + 1$.

Bewijs. Schrijf $\beta_{i+1} := b_{i+1\,i}$ en $\mathbf{V}_k := \mathrm{span}(e_1, \ldots, e_k)$.
Zij $\vec{y} = \alpha_1 e_1 + \ldots + \alpha_l e_l$ met $\alpha_l \neq 0$ en $l \leq n - m$.
Dan $B^j \vec{y} = \ldots + \gamma_{l+j} e_{l+j}$ met $\gamma_{l+j} := \beta_{l+j}\beta_{l+j-1} \cdots \beta_{l+1}\alpha_l \neq 0$ als $l + j \leq n$.
Dus $B^j \vec{y} \in \mathbf{V}_{l+j} \backslash \mathbf{V}_{l+j-1}$ als $l + j \leq n$ en de onafhankelijkheid van $\vec{y}, B\vec{y}, \ldots, B^{n-l}\vec{y}$ volgt. $\square$

(5.1.9) • Opgave. Zij $A \in \mathbf{M}_n$. Zij $\mu_1, \ldots, \mu_m \in \mathbf{C}$ de verschillende eigenwaarde van A ($\sigma(A) = \{\mu_1, \ldots, \mu_m\}$ en $\#\sigma(A) = m$).
Voor iedere i is l_i de lengte van het grootste Jordankastje bij de eigenwaarde μ_i.
Beschouw $\psi_A(\zeta) := (\zeta - \mu_1)^{l_1}(\zeta - \mu_2)^{l_2} \cdots (\zeta - \mu_m)^{l_m}$ voor $\zeta \in \mathbf{C}$.
(a) Bewijs dat $\psi_A(A) = 0$.
(b) Bewijs dat $\dim(\mathrm{K}^k(A; \vec{y})) \leq \sum_{i=1}^m l_i$ voor iedere Krylov ruimte $\mathrm{K}^k(A; \vec{y})$.
(c) Als $A \in \mathbf{M}_n$ een irreducibele Hessenberg matrix dan heeft iedere eigenwaarde precies een Jordankastje. Bewijs dit. (Hint: beschouw $\mathrm{K}^n(A; e_1)$.)
(d) Een irreducibele diagonalizeerbare Hessenberg matrix A heeft alleen enkelvoudige eigenwaarden. Bewijs dit.
(Voor het speciale geval dat A hermiets is zie (b) van stelling (5.2.3).)

(5.1.10) Stelling.
Zij $B \in \mathbf{M}_n$ een irreducibele Hessenberg matrix.
Zij $\mathbf{V}$ een lineaire deelruimte van $\mathbf{K}^n$ zodat $B(\mathbf{V}) \subset \mathbf{V}$.
Dan $\mathrm{span}(e_1, \ldots, e_{n-m}) \cap \mathbf{V} = \{0\}$ *waarbij* $m := \dim\mathbf{V}$.

Bewijs. Beschouw een $\vec{y} \in \mathrm{span}(e_1, \ldots, e_{n-m}) \cap \mathbf{V}$.
Omdat $B(\mathbf{V}) \subset \mathbf{V}$ is $\mathrm{K}^{m+1}(B; \vec{y}) \subseteq \mathbf{V}$. Lemma (5.1.8) vertelt ons dat $\dim(\mathrm{K}^{m+1}(B; \vec{y})) = m + 1$ als $\vec{y} \neq 0$. Omdat $\dim\mathbf{V} = m$ is blijkbaar $\vec{y} = 0$. □

(5.1.11) Gevolg. *Zij B als in* (5.1.10).
Als $\vec{v}_1, \ldots, \vec{v}_n$ een basis van eigenvektoren van B en $e_1 = \sum_{j=1}^n \nu_j \vec{v}_j$ dan $\nu_j \in \mathbf{K}\backslash\{0\}$ alle j.

Bewijs. Beschouw $j \in \{1, \ldots, n\}$. Met $\mathbf{V} = \{\sum_{i \neq j} \alpha_i \vec{v}_i\}$ en $m = n - 1$ volgt $e_1 \notin \mathbf{V}$ uit (5.1.10). □

De Hessenberg vorm is voor irreducibele Hessenberg matrices in zekere zin uniek.

(5.1.12) • Opgave. Zij $B \in \mathbf{M}_n$ een irreducible Hessenberg matrix. Als $Q \in \mathbf{M}_n$ unitair is, $Qe_1 = \zeta e_1$ met $\zeta \in \mathbf{C}$, $|\zeta| = 1$ en QBQ^* is ook weer een Hessenberg matrix dan is $|Q| = I$. Bewijs dit.

(5.1.13) • Opgave. Beschouw een $A \in \mathbf{M}_{m\times n}$. Laat zien dat er Householder transformaties $H_{\vec{v}_1}, H_{\vec{w}_1}, H_{\vec{v}_2}, H_{\vec{w}_2}, \ldots, H_{\vec{v}_{n-1}}$ bestaan zodat met $A_{k+1} := (H_{\vec{v}_k} A_k) H_{\vec{w}_k}$ voor $k = 1, \ldots n-1$ (met $\vec{w}_{n-1} := \vec{0}$) de uiteindelijke matrix $A_n = (a_{ij})$ een bi-diagonaal matrix is ($a_{ij} = 0$ als $j > i$ of $i > j+1$): $A_n = Q_1 A Q_2^*$ is bi-diagonaal voor zekere unitaire $Q_1 \in \mathbf{M}_n$ en unitaire $Q_2 \in \mathbf{M}_m$.

5.2 Karakteristieke polynoom methoden

Als $\Psi_A(\zeta) = \zeta^n + \alpha_1 \zeta^{n-1} + \ldots + \alpha_n$ het karakteristieke polynoom is kan men de koëfficiënten α_i van het karakteristieke polynoom (efficiënt) expliciet bepalen (zie de opgaven (5.2.1) en (5.2.2)). Hoewel we daarna efficiënt de polynoomwaarden kunnen bepalen (iedere waarde middels Horner's schema in n flop in plaats van de $\frac{1}{2}n^2$ die nodig zijn als A een Hessenberg matrix is) raden we toch deze aanpak om stabiliteitsredenen af: een representatie van een polynoom als lineaire kombinatie van machten ζ^i is, zoals we in NUMERIEKE WISKUNDE B gezien hebben, zeer gevoelig voor fouten op de koëfficiënten α_i.

(5.2.1) Opgave. (a) Zij $A \in \mathbf{M}_n(\mathbf{K})$.
Kies een $\vec{y} \in \mathbf{K}^n$. Bereken $\vec{b}_0, \vec{b}_1, \ldots, \vec{b}_n$ volgens
$$\vec{b}_0 := \vec{y} \quad \text{en} \quad \vec{b}_{k+1} = A\vec{b}_k \text{ voor } k = 0, 1, \ldots, n-1.$$
Zij nu $Y \in \mathbf{M}_n(\mathbf{K})$ met kolom vektoren $\vec{b}_0, \vec{b}_1, \ldots, \vec{b}_{n-1}$: $Ye_i = \vec{b}_{i-1}$.
Schrijf $\vec{x} := (\alpha_n, \alpha_{n-1}, \alpha_{n-2}, \ldots, \alpha_1)^T \in \mathbf{K}^n$.
Bewijs dat $$Y\vec{x} = -\vec{b}_n.$$
(Hint: gebruik de Cayley–Hamilton stelling.) Als Y niet-singulier is kunnen we hieruit de koëfficiënten α_i van het karakteristieke polynoom berekenen.

(b) Als A een irreducibele Hessenberg matrix is en $\vec{y} = e_1$ dan is Y een inverteerbare bovendriehoeks matrix. Bewijs dat.
Ga na dat Y bepalen gegeven de Hessenberg matrix $A \approx \frac{1}{6}n^3$ flop kost. Van een willekeurige $n \times n$ matrix kunnen we met $\approx 1\frac{1}{2}n^3$ flop alle koëfficiënten van het karakteristieke polynoom bepalen (ga dat na).
• (c) Stel dat $A = V\Lambda V^{-1}$ met V niet-singulier en $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$.
Stel dat $\vec{y} = \beta_1\vec{v}_1 + \ldots + \beta_n\vec{v}_n$ waarbij $\vec{v}_i := Ve_i$ en $\beta_i \neq 0$ alle i. Definieer $D = \mathrm{diag}\,(\beta_1, \ldots, \beta_n)$. Beschouw $X := D^{-1}V^{-1}Y$. Bewijs dat X een Vandermonde matrix is: met $X = (x_{ij})$ is $x_{ij} = \lambda_i^j$. Bewijs dat $\#\sigma(A) = n \Leftrightarrow Y$ is inverteerbaar.

(5.2.2) • **Opgave.** Bewijs dat $\alpha_1 = \mathrm{spoor}\,(A)$, $2\alpha_2 = (\mathrm{spoor}\,(A))^2 - \mathrm{spoor}\,(A^2)$ en $6\alpha_3 = (\mathrm{spoor}\,(A))^3 + 2\mathrm{spoor}\,(A^3) - 3\mathrm{spoor}\,(A^2)\,\mathrm{spoor}\,(A)$. Alle koëfficiënten $\alpha_1, \ldots, \alpha_n$ van het karakteristieke polynoom kunnen op dergelijke manier berekend worden uit $\mathrm{spoor}\,(A), \ldots, \mathrm{spoor}\,(A^n)$.

Van een reële matrix kunnen we de reële eigenwaarden iteratief benaderen door bijvoorbeeld de bisektie methode toe te passen op het karakteristieke polynoom, waarbij we de polynoomwaarden telkens berekenen door de determinant van $\zeta I - A$ met de LU-dekompositie te bepalen. De matrix kan eerst op Hessenberg vorm gebracht zijn. Deze aanpak noemt men “het in plakken snijden van het spektrum” (“slicing the spectrum”).

Voor hermietse matrices werken we deze methode wat verder uit. Omdat we bij het berekenen van een determinant ook impliciet de determinant van zekere ondermatrices berekenen hebben we in feite meer informatie dan alleen de karakteristieke polynoomwaarde. Voor hermietse matrices geven we aan hoe we deze extra informatie kunnen gebruiken.

(5.2.3) Zij $A \in \mathbf{M}_n$ een hermietse matrix.
Om de eigenwaarden te bepalen mogen we aannemen dat $A = A_n$, waarbij met $\gamma_i \in \mathbf{R}$ en $\beta_i \in \mathbf{C}\backslash\{0\}$ voor $k \leq n$ de matrices A_k gegeven zijn door

$$A_k := \begin{bmatrix} \gamma_1 & \bar{\beta}_2 & 0 & \ldots & 0 & 0 \\ \beta_2 & \gamma_2 & \bar{\beta}_3 & \ldots & 0 & 0 \\ 0 & \beta_3 & \gamma_3 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & \gamma_{k-1} & \bar{\beta}_k \\ 0 & 0 & 0 & \ldots & \beta_k & \gamma_k \end{bmatrix}$$

(A is een irreducibele tri-diagonaal; zie (5.1.1) en (5.1.3)).

De rij van karakteristieke polynomia van deze matrices A_k is een Sturm rij. We bekijken dit soort rijen in het volgende lemma.

(5.2.4) Lemma. *Zij $(p_k)_{k\leq n}$ een rij reële polynomia zodat voor iedere $k = 1, \ldots, n$ geldt:*
(i) *p_k is van graad k en heeft alleen reële nulpunten;*
(ii) *tussen twee opeenvolgende nulpunten van p_k ligt precies een nulpunt van p_{k-1};*
(iii) *$p_k(\zeta) = \zeta^k +$ lager graads termen, en $p_0 := 1$.*
Zij $V_k(\zeta)$ het aantal tekenwisselingen in de rij $p_0(\zeta), p_1(\zeta), \ldots, p_k(\zeta)$
($V_k(\zeta) := \frac{1}{2}\sum_{j=1}^{k} |\mathrm{teken}\, p_j(\zeta) - \mathrm{teken}\, p_{j-1}(\zeta)|$ waarbij $\mathrm{teken}\,\tau := +1$ als $\tau \geq 0$ en $\mathrm{teken}\,\tau := -1$ als $\tau < 0$). *Dan*

$$\#([\zeta, \infty) \cap \{\lambda \,\big|\, p_n(\lambda) = 0\}) = V_n(\zeta) \quad \textit{voor iedere} \quad \zeta \in \mathbf{R},\ p_n(\zeta) \neq 0.$$

Bewijs. Voor iedere k laat $\lambda_1^{(k)}, \ldots, \lambda_k^{(k)}$ de nulpunten van p_k zijn zo genummerd dat $\lambda_1^{(k)} < \lambda_2^{(k)} < \ldots < \lambda_k^{(k)}$ ((ii) impliceert dat de nulpunten van p_k allen enkelvoudig zijn). Schrijf $\lambda_0(A_k) := -\infty$, $\lambda_{k+1}(A_k) := +\infty$. Omdat p_k eigenschap (iii) heeft geldt dat voor iedere $\zeta \in \mathbf{R}$ $\mathrm{teken}\, p_k(\zeta) = (-1)^{k-j}$ als $\zeta \in (\lambda_j^{(k)}, \lambda_{j+1}^{(k)})$.
Uit (ii) volgt dat $\lambda_j^{(k-1)} \in (\lambda_j^{(k)}, \lambda_{j+1}^{(k)})$ voor iedere $j = 1, \ldots, k-1$.

Met induktie naar k bewijzen we dat voor iedere $\zeta \in \mathbf{R}$

$$V_k(\zeta) = k - j \quad \text{als} \quad \zeta \in [\lambda_j^{(k)}, \lambda_{j+1}^{(k)}). \qquad (*)$$

Het lemma volgt dan gemakkelijk (ga dat na).

$(*)$ is gemakkelijk te bewijzen voor $k = 1$. Stel $(*)$ is korrekt voor $k = n - 1$.
Beschouw een $\zeta \in \mathbf{R}$, $p_{n-1}(\zeta) \neq 0$ en $p_n(\zeta) \neq 0$.
Dan $\zeta \in (\lambda_j^{(n-1)}, \lambda_{j+1}^{(n-1)})$ voor zekere j. Dus $\quad$ teken $p_{n-1}(\zeta) = (-1)^{n-1-j}$.
Als $\zeta > \lambda_{j+1}^{(n)}$ dan teken $p_n(\zeta) = (-1)^{n-(j+1)} =$ teken $p_{n-1}(\zeta)$ en dus
$V_n(\zeta) = V_{n-1}(\zeta) = n - 1 - j = n - (j+1)$.
Als $\zeta < \lambda_{j+1}^{(n)}$ dan teken $p_n(\zeta) = (-1)^{n-j} \neq$ teken $p_{n-1}(\zeta)$ en dus
$V_n(\zeta) = V_{n-1}(\zeta) + 1 = n - 1 - j + 1 = n - j$.
Dus $(*)$ is korrekt voor $k = n$ als ζ geen nulpunt van p_n en p_{n-1}. Ga zelf na dat $(*)$ ook korrekt is voor $k = n$ en ζ wel een nulpunt van p_n of p_{n-1}. □

Een rij polynomia (p_k) met eigenschap (i) en (ii) is een zogenaamde *Sturm rij* . (iii) is slechts een schalings eigenschap.

(5.2.5) Stelling. *Zij* (A_k) *als in* (5.2.3).
Schrijf $p_{-1} := 0$, $p_0 := 1$ *en* $p_k := \Psi_{A_k}$ *voor* $k = 1, \ldots, n$. *Dan geldt:*
(a) *deze rij* (p_k) *van karakteristieke polynomia is een Sturm rij;*
(b) $p_k(\zeta) = (\zeta - \gamma_k)p_{k-1}(\zeta) - |\beta_k|^2 p_{k-2}(\zeta)$ *voor* $k = 1, \ldots n$ *en* $\zeta \in \mathbf{C}$.

Bewijs. De drie-terms rekursie in (b) volgt eenvoudig door de determinant van $\zeta I_k - A_k$ naar de laatste kolom te ontwikkelen.
Als $p_{j+1}(\zeta) = p_j(\zeta) = 0$ dan volgt, door de rekursie terugwaarts te doorlopen, dat $p_0(\zeta) = 0$ hetgeen in tegenspraak is met de definitie van p_0. Blijkbaar $\sigma(A_j) \cap \sigma(A_{j+1}) = \emptyset$. Omdat A_k hermiets is zijn de wortels van het karakteristiek polynoom p_k reëel. Eigenschap (i) en (iii) volgen nu gemakkelijk. (ii) volgt uit de scheidingsstelling van Cauchy (zie gevolg (4.4.6)). □

(5.2.6) Opgave. Bewijs de scheidingseigenschap (ii) ook uit het feit dat $p_j(\zeta) = 0$ impliceert $p_{j+1}(\zeta)p_{j-1}(\zeta) < 0$ voor iedere $j > 0$ (gebruik (b) van (5.1.10)).

Beschouw $\alpha, \beta \in \mathbf{R}\backslash\sigma(A_n)$, $\alpha < \beta$. Als bijvoorbeeld $V_n(\alpha) = 4$ en $V_n(\beta) = 2$ dan bevat $[\alpha, \beta]$ volgens de Sturm rij eigenschap in het lemma precies 2 eigenwaarden van A_n. Zouden we alleen $p_n(\alpha)$ en $p_n(\beta)$ berekend hebben en niet het aantal tekenwisselingen dan zouden we niet weten of we het intervalletje $[\alpha, \beta]$ in de bisektie methode nog verder moeten halveren, immers $p_n(\alpha)$ en $p_n(\beta)$ hebben hetzelfde teken (waarom?).

Het bereken van $p_n(\zeta)$ met behulp van de rekursie in (b) kost $\approx 2n$ flop ($|\beta_k|^2$ hoeft niet voor iedere ζ opnieuw uitgerekend te worden). Het bijhouden van de tekenwisselingen kost nauwelijks extra werk maar levert zeker in verband met de bisektie methode nuttige informatie.

Als we de eigenwaarden met behulp van de bisektie methode en de Sturm rij eigenschap in voldoende kleine intervalletjes gelokalizeerd hebben kunnen we eventueel overstappen op een sneller konvergente methode (die betere startwaarden verlangt dan de bisektiemethode) als Koorden Newton.

We evalueren $p_n(\zeta)$ zonder de representatie van het polynoom als lineaire kombinatie van machten ζ^j te gebruiken. De berekeningsmethode met behulp van de rekursie in (b) van (5.2.3) is, zoals uit het resultaat in de volgende opgave blijkt stabiel (in NUMERIEKE WISKUNDE B berekenden we ook om stabiliteitsredenen waarden van orthogonale polynomia liever met de drie-terms rekursie).

(5.2.7) Opgave. Zij A_n als in (5.2.3).
Laat $\zeta \in [-\|A_n\|_2, +\|A_n\|_2]$ exact in de computer gerepresenteerd zijn. $(p_k(\zeta)^\star)$ is door de computer met de rekursie in (b) van (5.2.5) geproduceerd. Bewijs dat de $p_k(\zeta)^\star$ exact voldoen aan $\quad p_{-1}(\zeta)^\star = 0, \quad p_0(\zeta)^\star = 1 \quad$ en

$$p_k(\zeta)^\star = (\zeta - \gamma'_k)p_{k-1}(\zeta)^\star - |\beta'_k|^2 p_{k-2}(\zeta)^\star \quad \text{voor} \quad k = 1, \ldots, n$$

met $\gamma'_k \in \mathbf{R}$, $\beta'_k \in \mathbf{C}$ zodat $\quad |\gamma_k - \gamma'_k| \leq 3\bar{\xi}(|\zeta| + |\gamma_k|) \quad$ en $\quad |\beta_k - \beta'_k| \leq \frac{3}{2}\bar{\xi}\,|\beta_k|$.

Als $\alpha, \beta \in [-\|A_n\|_2, +\|A_n\|_2]$, $\quad \alpha < \beta$ exact in de machine gerepresenteerd zijn en

$$V_n(\beta)^\star - V_n(\alpha)^\star \geq 1 \quad \text{dan} \quad \sigma(A_n) \cap [\alpha - 9\bar{\xi}\, \|A_n\|_2, \beta + 9\bar{\xi}\, \|A_n\|_2] \neq \emptyset.$$

Bewijs dit.

(5.2.8) • **Opgave.** Zij A_k als in (5.2.3), $A := A_n$. Zij $k < n$. Bewijs dat er voor iedere $j = 1, \ldots, k-1$ geldt

$$\sigma(A) \cap [\lambda_j(A_k), \lambda_{j+1}(A_k)] \neq \emptyset.$$

(Hint: Beschouw $P := I - e_{k+1}e_{k+1}^*$ en $A' := PAP$: A' is de matrix A waarin de koëfficiënten van de k-de rij en de k-de kolom door 0'en vervangen zijn. A' is hermiets en $\sigma(A_k) \subset \sigma(A')$. Bovendien volgt met argumenten als in het bewijs van de scheidingsstelling van Cauchy dat

$$\lambda_1(A) \leq \lambda_1(A') \leq \lambda_2(A) \leq \ldots \leq \lambda_{n-1}(A') \leq \lambda_n(A).$$
)

(5.2.9) • **Opgave.** Als $H \in \mathbf{M}_n$ hermiets is met eigenwaarden $\lambda_1(H) \leq \ldots \leq \lambda_n(H)$ dan is *het inertia* $\pi(H)$ *van* H is het drietal (k_-, k_0, k_+) gehele getallen dat gegeven is door $k_- := \#\{j \,|\, \lambda_j(H) < 0\}$, $k_0 := \#\{j \,|\, \lambda_j(H) = 0\}$ en $k_+ := \#\{j \,|\, \lambda_j(H) > 0\}$.
(a) Bewijs de *traagheidsstelling van Sylvester*:
$\pi(A) = \pi(L^*AL)$ voor iedere hermietse $A \in \mathbf{M}_n$ en niet-singuliere $L \in \mathbf{M}_n$.
(Hint: Met $\mathbf{V}_-(H) := \bigoplus_{\lambda_j(H)<0} \text{Kern}\,(H - \lambda_j(H)I)$, etc. is $k_- = \dim \mathbf{V}_-(H)$.
Als $x \in \mathbf{V}_-(H)\backslash\{0\}$ dan $(Hx, x) < 0$. Verder is $\mathbf{V}_-(A) \cap L^{-1}(\mathbf{V}_+(L^*AL)) = \{0\}$.)
(b) Als A een irreducibele hermietse tri-diagonaal matrix is kunnen we de eigenwaarden benaderend berekenen met behulp van de traagheidsstelling van Sylvester, de LU-dekompositie met de methode van Gauss, en de bisektiemethode analoog aan de boven geschetste aanpak met de Sturmrij eigenschap. Hoe gaat dit in zijn werk? Vergelijk de twee methoden.

5.3 De machtsmethode

In deze paragraaf behandelen we een methode die welliswaar praktisch bruikbaar is maar niet erg efficiënt. We behandelen deze methode, de machtsmethode, desondanks tamelijk uitgebreid om twee redenen.
Het zeer veel gebruikte en efficiënte QR-algorithme is er in wezen op gebaseerd. Inzicht in het QR-algorithme is gemakkelijk te verkrijgen via de konceptueel eenvoudigere machtsmethode.
Verder doen zich in diverse numerieke proceduren die niets met eigenwaarden problemen te maken hebben (onbedoeld) verschijnselen voor die in de theorie rond de machtsmethode een rol spelen. Door een goed begrip van de machtsmethode kunnen we die verschijnselen begrijpen en mogelijk onder kontrole houden.

(5.3.1) In deze paragraaf is $A \in \mathbf{M}_n$ en de eigenwaarden $\lambda_1, \ldots \lambda_n$ van A geteld naar multipliciteit zijn zo genummerd dat

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|.$$

(5.3.2) **Definitie.** Een eigenvektor $\vec{v}$ van A is *dominant* als de bijbehorende eigenwaarde λ enkelvoudig is en $|\lambda| > |\mu|$ voor iedere $\mu \in \sigma(A)\backslash\{\lambda\}$
(dus $\lambda = \lambda_1$ en $\rho_0 := \frac{|\lambda_2|}{|\lambda_1|} < 1$). De andere eigenvektoren worden dan *gedomineerd.*

De machtsmethode berust op het volgende idee. Als A een dominante eigenvektor heeft en $\vec{x}_0$ heeft een komponent in de richting van die dominante eigenvektor dan staat $A^k\vec{x}_0$ voor grote k min of meer in de richting van die dominante vektor (zie (5.3.6)).

Het algorithme van de machtsmethode kent diverse varianten. We geven er hier twee.

(5.3.3) **De machtsmethode.**
De *machtsmethode* produceert als volgt een rij $(\vec{x}_k)$ in $\mathbf{K}^n$ en een rij (μ_k) in $\mathbf{K}$:

Algorithme		**Algorithme (variant)**	
(a)	$k := 0$, $\vec{y}_0$ `is gegeven;`	(a')	$k := 0$, $\vec{y}_0$ `is gegeven;`
(b1)	`bereken` $\vec{x}_k = \vec{y}_k/\|\vec{y}_k\|_2$;	(b1')	`bereken` $\vec{x}_k = \vec{y}_k/(\vec{y}_k, e_1)$;
(b2)	`bereken` $\vec{y}_{k+1} = A\vec{x}_k$;	(b2')	`bereken` $\vec{y}_{k+1} = A\vec{x}_k$;
(b3)	$\mu_k := (\vec{y}_{k+1}, \vec{x}_k)$;	(b3')	$\mu_k := (\vec{y}_{k+1}, e_1)$;
(c)	$k := k+1$; `ga na naar` (b1);	(c')	$k := k+1$; `ga na naar` (b1');

(5.3.4) Opgave. Met $\vec{z}_k := A^k\vec{x}_0$ is $\vec{x}_k = \vec{z}_k/\|\vec{z}_k\|_2$ (of in de tweede variant $\vec{x}_k = \vec{z}_k/(\vec{z}_k, e_1)$). Ga dit na.

Als de rij $(\vec{x}_k)$ in richting naar een eigenvektor konvergeert dan konvergeert (μ_k) naar de bijbehorende eigenwaarde. Het volgende lemma schat de fout in μ_k in termen van de fout in de richting van $\vec{x}_k$.

(5.3.5) Lemma. *Stel* $\vec{v} \in \mathrm{Kern}\,(A-\lambda I)$, $\vec{v} \neq 0$. *Beschouw een* $\vec{x} = \beta\vec{v} + \vec{r}$ *met* $\|\vec{x}\|_2 = 1$. *Dan* $|(A\vec{x}, \vec{x}) - \lambda| \le \|A-\lambda I\|_2\,\|\vec{r}\|_2$. *Als* $A^* = A$ *dan* $|(A\vec{x}, \vec{x}) - \lambda| \le \|A-\lambda I\|_2\,\|\vec{r}\|_2^2$.

Bewijs. $(A\vec{x}, \vec{x}) = (\beta\lambda\vec{v} + A\vec{r}, \vec{x}) = \lambda(\vec{x}, \vec{x}) + ((A-\lambda I)\vec{r}, \vec{x})$ en de eerste schatting volgt uit de Cauchy–Schwartz ongelijkheid.
Als $A^* = A$ dan $((A-\lambda I)\vec{r}, \vec{x}) = (\vec{r}, (A-\lambda I)\vec{x}) = (\vec{r}, \beta(\lambda-\lambda)\vec{v} + (A-\lambda I)\vec{r})$. □

(5.3.6) Stelling.
Stel dat $\vec{v}_1$ *met eigenwaarde* λ_1 *een dominante eigenvektor van* A *is. Kies* $\rho_0 \in (\frac{|\lambda_2|}{|\lambda_1|}, 1)$.
Zij $\mathbf{W} := \mathrm{Im}\,(A - \lambda_1 I)$.
Zij $\alpha_1 \in \mathbf{K}$ *zodat* $\vec{x}_0 - \alpha_1\vec{v}_1 \in \mathbf{W}$. *Zij* (β_k) *in* $\mathbf{K}$ *zodat* $(\vec{x}_k - \beta_k\vec{v}_1)$ *in* $\mathbf{W}$.
Als $\alpha_1 \neq 0$ *dan is er voor iedere* $\rho > \rho_0$ *een* $\Gamma > 0$ *zodat*

$$\|\vec{x}_k - \beta_k\vec{v}_1\|_2 \le \Gamma\rho^k \quad \textit{en} \quad |\mu_k - \lambda_1| \le \|A - \lambda_1 I\|_2\Gamma{\rho_0}^k \quad \textit{voor alle } k. \tag{21}$$

(in variant: $\vec{x}_k \to \frac{1}{(\vec{v}_1, e_1)}\vec{v}_1$ als $k \to \infty$).

Bewijs. Schrijf $\vec{r}_k := \vec{x}_k - \beta_k\vec{v}_1$. De tweede schatting volgt onmiddellijk uit de eerste en (5.3.5).

$$A^k\vec{x}_0 = A^k(\alpha_1\vec{v}_1 + \vec{r}_0) = \alpha_1\lambda_1^k\vec{v}_1 + A^k\vec{r}_0 = \alpha_1\lambda_1^k(\vec{v}_1 + \tilde{r}_k) \quad \text{met} \quad \tilde{r}_k := \frac{1}{\alpha_1}(\frac{1}{\lambda_1}A)^k\vec{r}_0. \tag{22}$$

Omdat $A(\mathbf{W}) \subseteq \mathbf{W}$ en $\vec{r}_0 \in \mathbf{W}$ is $\tilde{r}_k \in \mathbf{W}$ en $\tilde{r}_k = \frac{1}{\alpha_1}(\frac{1}{\lambda_1}A_{|\mathbf{W}})^k\vec{r}_0$.
Omdat $\sigma(A_{|\mathbf{W}}) = \sigma(A)\backslash\{\lambda_1\}$ (λ_1 is enkelvoudig) is $\rho(\frac{1}{\lambda_1}A_{|\mathbf{W}}) = \frac{|\lambda_2|}{|\lambda_1|} = \rho_0 < 1$.
Dus er is een $\Gamma' > 0$ zodat voor iedere $k \in \mathbf{N}$ geldt $\|\tilde{r}_k\|_2 \le \Gamma'\rho^k$ (zie (1.4.8)).
(Als A diagonalizeerbaar is, $A\vec{v}_i = \lambda_i\vec{v}_i$ met basis $\vec{v}_1, \ldots, \vec{v}_n$, en $\vec{x}_0 = \sum\alpha_i\vec{v}_i$ dan

$$\tilde{r}_k = \frac{\alpha_2}{\alpha_1}(\frac{\lambda_2}{\lambda_1})^k\vec{v}_2 + \frac{\alpha_3}{\alpha_1}(\frac{\lambda_3}{\lambda_1})^k\vec{v}_3 + \ldots + \frac{\alpha_n}{\alpha_1}(\frac{\lambda_n}{\lambda_1})^k\vec{v}_n. \tag{23}$$

Dus $\|\tilde{r}_k\|_2 \le \Gamma'{\rho_0}^k$ met $\Gamma' := \sum_{i=2}^n |\frac{\alpha_i}{\alpha_1}|\,\|\vec{v}_i\|_2$.)
Merk op dat

$$\vec{x}_k = \frac{A^k\vec{x}_0}{\|A^k\vec{x}_0\|_2} = \frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|}\left(\frac{\vec{v}_1}{\|\vec{v}_1 + \tilde{r}_k\|_2} + \frac{\tilde{r}_k}{\|\vec{v}_1 + \tilde{r}_k\|_2}\right)$$

Blijkbaar

$$\vec{r}_k = \frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|}\frac{\tilde{r}_k}{\|\vec{v}_1 + \tilde{r}_k\|_2} \quad \text{en} \quad \beta_k = \frac{\alpha_1\lambda_1^k}{|\alpha_1\lambda_1^k|}\frac{1}{\|\vec{v}_1 + \tilde{r}_k\|_2}.$$

Dus er is een $\Gamma > 0$ zodat voor iedere k geldt $\|\vec{r}_k\|_2 \le \Gamma\rho^k$ ($\|\vec{r}_k\|_2 \le \Gamma'\rho_0^k/(\|\vec{v}\|_2 - \Gamma'\rho_0^k)$).□

(5.3.7) Opmerking. Omdat $\vec{v}_1$ dominant is overvleugelt voor grote k in $A^k\vec{x}_0$ de komponent in de richting van $\vec{v}_1$ de komponent in iedere andere richting. In een relatieve zin benadert $A^k\vec{x}_0$ een skalair veelvoud van $\vec{v}_1$. In wezen is dit de machtsmethode. Stap (b3) (of (b3')) schaalt slechts deze vektor $A^k\vec{x}_0$ om 'overflow' of 'underflow' te voorkomen. Het is duidelijk dat we de eigenwaarde kunnen bepalen als we de eigenvektor hebben. (b2) (of (b2')) is een manier waarop dit kan.

De tweede variant van het algorithme is efficiënter dan de eerst: we schalen alleen zodat de eerste koördinaat 1 is ((b2) en (b3) samen kosten $3n$ flop terwijl (b2') en (b3') samen n flop kost). De tweede variant werkt echter alleen als de $\vec{x}_k$ én de dominante eigenvektor een komponent in de richting e_1 hebben (waarom?).

(5.3.8) Opmerking. Omdat $\|\vec{x}_k - \beta_k\vec{v}_1\|_2 \to 0$ als $k \to \infty$ en $\vec{r}_k := \vec{x}_k - \beta\vec{v}_1 \in \mathbf{W}$ kunnen we zeggen dat $(\vec{x}_k)$ *in richting konvergeert* naar $\vec{v}_1$.
We kunnen dit wat verder precisiëren in termen van hoeken:
Als $\vec{v}_1 \perp \mathbf{W}$ dan is $\|\vec{r}_k\|_2 = \sin\angle(\vec{x}_k, \mathrm{kern}\,(A-\lambda_1 I))$ (en $\|\tilde{r}_k\|_2 = \tan\angle(\vec{x}_k, \mathrm{kern}\,(A-\lambda_1 I))$).
In het algemeen geldt $(1-\cos\theta)\|\vec{r}_k\|_2 \le \sin\angle(\vec{x}_k, \mathrm{kern}\,(A-\lambda_1 I)) \le (1+\cos\theta)\|\vec{r}_k\|_2$,
waarbij $\theta := \angle(\vec{v}_1, \mathbf{W})$ (ga dat na; zie ook (4.6.5)).

De konvergentie in stelling (5.3.6) is gemajoreerd lineair met konvergentie faktor $\approx \rho_0$. Deze konvergentie faktor hoeft nog niet zichtbaar te zijn voor kleine k. Pas als $\vec{x}_k$ al min of meer in de goede richting $\vec{v}_1$ staat is de konvergentie faktor waarneembaar.

(5.3.9) Voorbeeld. Laat $A \in \mathbf{M}_2$ gegeven zijn door $A\vec{v}_1 = 0.9\vec{v}_1$ en $A\vec{v}_2 = \vec{v}_2$ met $\vec{v}_1 = e_1$ en $\vec{v}_2 = e_1 + 0.1e_2$. Als $\vec{x}_0 = -e_1 + 0.1e_2 = -2\vec{v}_1 + \vec{v}_2$ dan neemt $\|A^k\vec{x}_0\|_2$ aanvankelijk af, vervolgens weer toe waarna de lengte stabilizeert bij ≈ 1.005 (reken dit na).

We beantwoorden achereenvolgens de volgende vragen.
– Wat gebeurt er als A geen dominante eigenvektor heeft?
– Hoe weet je of de startvektor $\vec{x}_0$ een komponent in de richting $\vec{v}_1$ heeft ($\alpha_1 \neq 0$)?
– Hoe bepaal je de overige eigenwaarden en eigenvektoren?
– Kan je de machtsmethode zo modificeren dat de geproduceerde rijen sneller konvergeren?

Geen dominante eigenvektor

Als A geen dominante eigenvektor heeft dan
(a) λ_1 is een meervoudige eigenwaarde van A en $|\lambda_1| > |\mu|$ voor alle $\mu \in \sigma(A)\backslash\{\lambda_1\}$ of
(b) $\lambda_1 \neq \lambda_i$ terwijl $|\lambda_1| = |\lambda_i|$ voor zekere $i > 1$.

In situatie (a) konvergeert de rij $(\vec{x}_k)$ toch in richting naar 'n eigenvektor bij de eigenwaarde λ_1 (mits $\vec{x}_0$ een komponent in de gegeneralizeerde eigenruimte bij eigenwaarde λ_1 heeft. De eigenruimte bij eigenwaarde λ_1 zou in deze situatie dominant genoemd kunnen worden). De konvergentie kan echter bijzonder langzaam zijn als de geometrische multipliciteit kleiner is dan de algebraïsche (zelfs als $\frac{|\mu|}{|\lambda_1|} \ll 1$ voor alle $\mu \in \sigma(A)\backslash\{\lambda_1\}$).

In situatie (b) zal de geproduceerde rij $(\vec{x}_k)$ niet konvergeren.

We zullen de beweringen niet verder theoretisch uitwerken. We volstaan met een tweetal illustratieve voorbeelden.

(5.3.10) Voorbeeld (a.i)
Met $A := \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix}$ is $A^k = \begin{bmatrix} 1 & k & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2^{-k} \end{bmatrix}$. Als $\vec{x}_0 = \sum_i \alpha_i e_i$ dan
$A^k\vec{x}_0 = (\alpha_1 + k\alpha_2)(e_1 + \frac{1}{\alpha_1 + k\alpha_2}(\alpha_2 e_2 + \alpha_3 e_3 + \alpha_4 2^{-k} e_4))$ en $\vec{x}_k \to e_1$ als $k \to \infty$ en $\alpha_2 \neq 0$
(in de k-de stap wordt de $\tan\varphi_k$ met $\varphi_k := \angle(A^k\vec{x}_0, e_1)$ gereduceerd met de faktor $\frac{\alpha_1 + (k-1)\alpha_2}{\alpha_1 + k\alpha_2}$.
Als $|\alpha_2| \ll |\alpha_1|$ duurt het lang voordat de $\alpha_3 e_3$ komponent verwaarloosbaar is).

Voorbeeld (a.ii)

Met $\quad A := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$ is $A^k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2^{-k} \end{bmatrix}$. Als $\vec{x}_0 = \sum_i \alpha_i e_i$ dan
$A^k \vec{x}_0 = \alpha_1 e_1 + \alpha_2 e_2 + \alpha_3 2^{-k} e_3$ en $\vec{x}_k \to \alpha_1 e_1 + \alpha_2 e_2$ als $k \to \infty$.

Als de "dominante" eigenwaarde λ_1 *half-enkelvoudig* is (d.w.z. de geometrische en algebraïsche multipliciteit zijn hetzelfde; λ_1 heeft alleen 1×1 Jordan kastjes) dan konvergeert $(\vec{x}_k)$ in richting naar de komponent van $\vec{x}_0$ in de eigenruimte van λ_1 (lineair met konvergentie faktor $\max\{\frac{|\mu|}{|\lambda_1|} \,\big|\, \mu \in \sigma(A)\backslash\{\lambda_1\}\}$). Bij grotere Jordan kastjes konvergeert $(\vec{x}_k)$ naar een eigenvektor met eigenwaarde λ_1 met maximaal Jordan kastje (maar de konvergentie is zeer traag).

Voorbeeld (b)
Met $\quad A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ en $\vec{x}_0 = \alpha_1 e_1 + \alpha_2 e_2$ is $A^k \vec{x}_0 = \alpha_1 e_1 + \alpha_2 (-1)^k e_2$ en konvergeert de rij $(\vec{x}_k)$ ook niet in richting naar een vektor (als $\alpha_i \neq 0$ voor $i = 1, 2$).

De situatie in (b) komt nogal eens voor.
Als $A \in \mathbf{M}_n(\mathbf{R})$ en $\lambda_1 \in \sigma(A)$ dan is ook $\bar{\lambda}_1 \in \sigma(A)$. Als deze λ_1 "dominant" is en niet reëel zitten we in situatie (b).

Heeft $\vec{x}_0$ een komponent in de dominante richting?

Stel dat A een dominante eigenvektor $\vec{v}_1$ heeft.
Afhankelijk van geloof zijn er drie antwoorden mogelijk op de vraag of $\vec{x}_0$ een komponent in de dominante richting heeft. Alle drie de antwoorden zijn positief.

(5.3.11) Opmerking. (a) De ruimte $\mathbf{W}$ heeft maat 0 in $\mathbf{K}^n$. Kiezen we random een startvektor $\vec{x}_0$ dan hebben we kans 0 dat $\vec{x}_0 \in \mathbf{W}$: $\alpha_1 \neq 0$ is het *generieke geval.*
(b) Als $\vec{x}_0 = \vec{r}_0 \in \mathbf{W}$ dan zal er door rekenfouten een komponentje in de richting $\vec{v}_1$ verschijnen: $\vec{x}_1^\star = \varepsilon \vec{v} + \vec{r}_1^\star$ voor een ε. (Omdat $\vec{v}_1$ dominant is zal de rekenfouten komponent in deze richting in het iteratief proces groeien. Op den duur (k groot) overvleugelt deze rekenfout komponent alle andere komponenten. In veel iteratieve processen gebeurt dit onbedoeld.)
(c) Als A een irreducibele Hessenberg matrix is dan heeft volgens (5.1.11) e_1 een komponent in de richting van alle eigenvektoren en dus ook in de richting van de dominante: $\vec{x}_0 = e_1$ is een goede startvektor.

Andere eigenwaarden en eigenvektoren

In deze subparagraaf behandelen we de vraag hoe we een volgende eigenwaarde met eigenvektor benaderend kunnen berekenen met behulp van een machtsmethode techniek.

(5.3.12) Stel dat de dominante eigenvektor $\vec{v}_1$ en bijbehorende eigenwaarde λ_1 (in zekere precisie) bepaald zijn.

Omdat we met een machtsmethode techniek de volgende eigenwaarde en eigenvektor willen berekenen nemen we aan dat $|\lambda_1| > |\lambda_2| > |\lambda_3|$.
Als $\vec{x}_0 \in \mathbf{W} = \mathrm{Im}\,(A - \lambda_1 I)$ en $\vec{x}_0$ heeft een komponent in de richting van $\vec{v}_2$, de eigenvektor bij eigenwaarde λ_2, dan produceert in theorie volgens stelling (5.3.6) de machtsmethode een rij vektoren die in richting naar $\vec{v}_2$ konvergeert
(omdat $A(\mathbf{W}) \subseteq \mathbf{W}$ hebben we theoretisch te maken met $A_{|\mathbf{W}}$ in plaats van A).

$\mathbf{W} = \{\vec{x} \in \mathbf{K}^n \,\big|\, \vec{x} \perp \vec{w}_1\}$, waarbij $\vec{w}_1$ de links eigenvektor is bij de eigenwaarde λ_1 ($\vec{w}_1^* A = \lambda_1 \vec{w}_1^*$; zie (4.6.2)). We kunnen met behulp van $\vec{w}_1$ als volgt iedere vektor $\vec{y} \in \mathbf{K}^n$ expliciet ontbinden als $\vec{y} = \alpha_1 \vec{v}_1 + \vec{r}_1$ met $\vec{r}_1 \in \mathbf{W}$. Dan $\alpha_1 := \frac{(\vec{y}, \vec{w}_1)}{(\vec{v}_1, \vec{w}_1)}$ en $\vec{r}_1 := P(\vec{y}) \in \mathbf{W}$, waarbij $P := I - \frac{1}{(\vec{v}_1, \vec{w}_1)} \vec{v}_1 \vec{w}_1^*$ (zie (4.6.3)). In het bijzonder lijkt $\vec{x}_0 := P(\vec{y})$ een geschikte startvektor.

Voor de praktijk zal deze aanpak echter gemodificeerd moeten worden:
– de dominante eigenvektor zal niet exact bekend zijn zodat een geprojekteerde startvektor $\vec{x}_0 := P(\vec{y})$ toch nog een komponent in de richting $\vec{v}_1$ zal hebben;

– zelfs als $\vec{v}_1$ exact bekend is zal de *berekende* iterant $\vec{x}_1^\star$ (en iedere volgende) door rekenfouten een komponentje in de richting $\vec{v}_1$ hebben. Deze 'rekenfout' komponent in de dominante richting zal na een aantal iteraties de andere komponenten volledig overvleugelen en de geproduceerde rij vektoren konvergeert in richting opnieuw naar $\vec{v}_1$.
(Dit overvleugelen gebeurt "vlugger" naarmate $|\lambda_2|/|\lambda_1|$ kleiner is, dus naarmate het konvergentie gedrag van de machtsmethode met betrekking tot de eerste eigenvektor gunstiger is.)

De ongewenste groei in de richting van de reeds berekende $\vec{v}_1$ kunnen we voorkomen door niet alleen bij de start maar na *iedere* slag het komponentje in de richting van $\vec{v}_1$ uit de iterant 'weg' te projecteren: $\vec{x}_k{}' := P(\vec{x}_k^\star)$, $\vec{y}_{k+1}^\star = A\vec{x}_k{}', \ldots$.
Deze methode, machtsmethode met projektie na iedere slag, is in feite de machtsmethode toegepast op een gedeflateerde matrix:

(5.3.13) Bewering. *Als $\vec{x}_0$ een komponent heeft in de richting van $\vec{v}_2$ dan produceert de machtsmethode toegepast met de matrix $A - \frac{\lambda_1}{(\vec{v}_1,\vec{w}_1)}\vec{v}_1\vec{w}_1^*$ een rij vektoren $(\vec{x}_k)$ die in richting konvergeert naar $\vec{v}_2$ en $((A\vec{x}_k,\vec{x}_k))$ konvergeert naar λ_2.*

Bewijs. Zie (b) van opgave (4.6.3) en opgave (5.3.14). □

(5.3.14) Opgave. $0, \lambda_2, \lambda_3, \ldots, \lambda_n$ zijn de eigenwaarden van $B := A - \frac{\lambda_1}{(\vec{v}_1,\vec{w}_1)}\vec{v}_1\vec{w}_1^*$.
Als $A\vec{v}_i = \lambda_i\vec{v}_i$ dan $B\vec{v}_i = \lambda_i\vec{v}_i$ voor $i > 1$ en $B\vec{v}_1 = 0$. Bewijs dit.

Door *deflatie* kunnen we dus met behulp van de machtsmethode de volgende eigenwaarden en eigenvektoren benaderend berekenen.

Om bovenstaande deflatie techniek toe te kunnen passen moeten we naast de dominante eigenvektor $\vec{v}_1$ ook nog de dominante links eigenvektor $\vec{w}_1$ bepalen (in de volgende subparagraaf laten we zien hoe dat enigszinds efficiënt kan als de eigenwaarde bekend is). Zijn we alleen geïnteresseerd in de eigenwaarden dan kunnen we ook onze deflatie techniek africhting op de Schur normaalvorm en hoeven we geen links eigenvektoren te bepalen: (we laten het bewijs van de volgende bewering aan de lezer over.)

(5.3.15) Bewering. *$A = USU^*$ met $S = (s_{ij})$ boven driehoeks zodat $s_{ii} = \lambda_i$ en U unitair. Schrijf $\vec{u}_i = Ue_i$. Dan $\vec{u}_1 = \vec{v}_1/\|\vec{v}_1\|_2$. Zij $\tilde{P} := I - \vec{u}_1\vec{u}_1^*$.*

Als $\vec{x}_0$ een komponent heeft in de richting van $\vec{u}_2$ dan produceert de machtsmethode toegepast met de matrix $\tilde{P}A$ een rij vektoren $(\vec{x}_k)$ die in richting konvergeert naar $\vec{u}_2$ en $((A\vec{x}_k,\vec{x}_k))$ konvergeert naar λ_2. □

(5.3.16) Opgave. Bewijs bovenstaande bewering.

(5.3.17) Opgave. Stel $|\lambda_i| > |\lambda_{i+1}|$ voor iedere i. Zij $m < n$.
(a) Laat $\vec{v}_i$ en $\vec{w}_i$ zo zijn dat $A\vec{v}_i = \lambda_i\vec{v}_i$, $\vec{w}_i^*A = \lambda_i\vec{w}_i^*$ en $(\vec{v}_i,\vec{w}_i) = 1$.
Beschouw de matrices $V_m, W_m \in \mathbf{M}_{n\times m}$ waarvoor $V_me_i = \vec{v}_i$ en $W_me_i = \vec{w}_i$ voor $i = 1,\ldots,m$.
Zij $P := I - V_mW_m^*$.
Bewijs dat de machtsmethode toegepast met de matrix PA een rij vektoren $(\vec{x}_k)$ produceert die in richting konvergeert naar $\vec{v}_{m+1}$ en $((A\vec{x}_k,\vec{x}_k))$ konvergeert naar λ_{m+1} als $\vec{x}_0$ een komponent heeft in de richting $\vec{v}_{m+1}$.
(b) Beschouw de matrix $U_m \in \mathbf{M}_{n\times m}$ waarvoor $U_me_i = Ue_i$ voor $i = 1,\ldots,m$.
Zij $\tilde{P} := (I - U_mU_m^*)$.
Bewijs dat de machtsmethode toegepast met de matrix $\tilde{P}A$ een rij vektoren $(\vec{x}_k)$ produceert die in richting konvergeert naar $\vec{u}_{m+1} = Ue_{m+1}$ en $((A\vec{x}_k,\vec{x}_k))$ konvergeert naar λ_{m+1} als $\vec{x}_0$ een komponent heeft in de richting $\vec{u}_{m+1}$.

Door de deflatie techniek herhaald toe te passen kunnen we alle eigenwaarden en eigenvektoren benaderend bepalen als $|\lambda_i| > |\lambda_{i+1}|$ alle i. Omdat de eigenvektoren niet exact bepaald zijn kan men zich met recht af vragen af de fouten niet met name de kleinere eigenwaarden ernstig verstoren. Als bijvoorbeeld $\vec{v}_1$ met de machtsmethode benaderd is dan heeft de benaderende vektor met name in de richting van de eigenvektoren bij in absolute waarde kleinste eigenwaarde een kleine komponent (zie bijvoorbeeld de uitdrukking (23) voor $\tilde{r}_k$ in het bewijs van (5.3.6)). Tengevolge hiervan valt de verstoring van de kleinere eigenwaarden nog wel mee. We werken dit hier niet verder meer uit.

Versnellingstechnieken: inverse machtsmethode met verschuiving

Onze versnellings technieken zijn gebaseerd op de volgende observatie.
Voor iedere $\kappa \in \mathbf{C}$ is $\vec{v}_i$ een eigenvektor van $A - \kappa I$ en van $(A - \kappa I)^{-1}$ (als $\kappa \notin \sigma(A)$) met eigenwaarde $\lambda_i - \kappa$, respektievelijk $\frac{1}{\lambda_i - \kappa}$ als $A\vec{v}_i = \lambda_i \vec{v}_i$ en $\vec{v}_i \in \mathbf{K}^n \backslash \{0\}$.

(5.3.18) Bewering.
Stel dat $\vec{x}_0$ een komponent heeft in de richting van $\vec{v}_1$. Beschouw een $\kappa \in \mathbf{C}$.
Pas de machtsmethode toe met $A - \kappa I$ in plaats van A. Als

$$\rho_\kappa := \frac{\max_{i>1} |\lambda_i - \kappa|}{|\lambda_1 - \kappa|} < 1$$

dan konvergeert $(\vec{x}_k)$ gemajoreerd lineair in richting naar $\vec{v}_1$ met konvergentie faktor $\approx \rho_\kappa$.

Bewijs. Pas stelling (5.3.6) met $A - \kappa I$ in plaats van A. □

De asymptotische konvergentie snelheid van de machtsmethode is $\frac{|\lambda_2|}{|\lambda_1|}$. Door met $A - \kappa I$ te itereren in plaats van A hebben we te maken met de konvergentie faktor $\approx \rho_\kappa$ (mits $\rho_\kappa < 1$) We verwachten de snelste konvergentie als κ de faktor ρ_κ minimalizeert.

Kunnen we de λ_i's schatten dan kunnen we een enigszinds gunstigere konvergentiesnelheid bewerkstelligen door de *machtsmethode met verschuiving* toe te passen.

(5.3.19) Opgave. Stel $\lambda_1 \approx 10$, $\lambda_2 \approx 8$, $\lambda_n = 0$ en $\lambda_i \in [0, 8]$ voor $i > 2$. Vergelijk de konvergentie snelheid van de machtsmethode met verschuiving en zonder verschuiving.

Hoewel we soms een gunstigere konvergentie kunnen bereiken met de machtsmethode met verschuiving is de winst niet echt spektakulair. Indrukwekkende snelheidswinst kunnen we boeken door over te stappen op de (duurdere) inverse machtsmethode met verschuiving.

(5.3.20) Inverse machtsmethode met verschuiving.
Zij (κ_k) een rij (nog nader te kiezen) verschuivingen ("shifts") in $\mathbf{C}$. De *inverse machtsmethode met verschuivingen* (κ_k) produceert als volgt een rij $(\vec{x}_k)$ in $\mathbf{K}^n$ en een rij μ_k in $\mathbf{K}$:

Algorithme

(a) $k := 0$, $\vec{y}_0$ `is gegeven;`
(b1) `bereken` $\vec{x}_k = \vec{y}_k / \|\vec{y}_k\|_2$;
(b2) $\mu_k := (A\vec{x}_k, \vec{x}_k)$;
(b3) `los` $\vec{y}_{k+1}$ `op uit` $(A - \kappa_k I)\vec{y}_{k+1} = \vec{x}_k$;
(c) $k := k + 1$; `ga na naar` (b1);

(als in (b2) $A - \kappa_k I$ niet inverteerbaar is, hebben we een eigenwaarde gevonden).

(5.3.21) Bewering.
Stel dat $\vec{x}_0$ een komponent heeft in de richting van $\vec{v}_1$. Beschouw een $\kappa \in \mathbf{C}$.
Produceer $(\vec{x}_k)$ met (5.3.20) en verschuivingen $\kappa_k = \kappa \quad (k \in \mathbf{N})$. Als

$$\tilde{\rho}_\kappa := \frac{\max_{i>1} \frac{1}{|\lambda_i - \kappa|}}{\frac{1}{|\lambda_1 - \kappa|}} = \frac{|\lambda_1 - \kappa|}{\min_{i>1} |\lambda_i - \kappa|} < 1$$

dan konvergeert $(\vec{x}_k)$ gemajoreerd lineair in richting naar $\vec{v}_1$ met konvergentie faktor $\approx \tilde{\rho}_\kappa$.

Bewijs. Pas stelling (5.3.6) met $(A - \kappa I)^{-1}$ in plaats van A. □

(5.3.22) Konvergentie snelheid en efficiëntie. (a) Theoretisch kan $\tilde{\rho}_\kappa$ willekeurig klein zijn, terwijl $\min_\kappa \rho_\kappa > 0$. We mogen dan ook verwachten dat we met de inverse machtsmethode sneller konvergente rijen $(\vec{x}_k)$ kunnen produceren dan met de gewone. Beschikken we over een goede schattting of een goede benadering $\tilde{\lambda}_1$ van λ_1 dan levert de inverse machtsmethode met verschuiving $\kappa = \tilde{\lambda}_1$ een snel konvergente rij.
(b) Bij de inverse machtsmethode moeten we echter in stap (b2) een vergelijking oplossen terwijl we bij de gewone machtsmethode slechts matrix–vektor vermenigvuldigingen hoeven uit te voeren. Lossen we de optredende vergelijkingen op met bijvoorbeeld een LU-dekompositie en hebben we steeds dezelfde verschuiving κ dan hoeven we die dekompositie in het iteratief proces slechts een maal te bepalen. Het verdere oplossen met de LU-dekompositie is weer even duur als de matrix-vektor vermenigvuldiging.
(c) Als A een irreducibele Hessenberg matrix is (de Hessenberg vorm was ook al aantrekkelijk in verband met de keuze van de startvektor; zie (5.3.11) (c)) dan kost voor $A - \kappa I$ het bepalen van bijvoorbeeld een LU-dekompositie $\approx \frac{1}{2}n^2$ flop, ongeveer de kosten van een matrix-vektor vermenigvuldiging.

(5.3.23) Andere eigenwaarden en eigenvektoren. Met de gewone machtsmethode met verschuiving kunnen we in principe de 'extreme' eigenwaarden (dus de eigenwaarden die aan de rand liggen van het konvexe omhulsel van het spektrum) en bijbehorende eigenvektoren benaderend bepalen: door geschikt te schuiven is immers iedere extreme eigenwaarde 'dominant te maken'. Niet-extreme eigenwaarden kunne we niet op zo'n manier dominant maken. Voor de gewone machtsmethode hebben we voor deze eigenwaarden dan ook deflatie technieken nodig. Met de inverse machtsmethode is daarentegen iedere eigenwaarde dominant te schuiven.

De inverse machtsmethode met veschuiving levert snel konvergente rijen als $\tilde{\rho}_\kappa \ll 1$. Omdat $\mathcal{C}(A - \kappa I) \geq 1/\tilde{\rho}_\kappa$ is dit een situatie waarin we ons zorgen behoren te maken over de stabiliteit van de vergelijking $(A - \kappa I)\vec{y} = \vec{x}$. De vektor $\vec{y}$ zal dan inderdaad met een fantastisch grote fout belast zijn maar, zoals uit het volgende lemma zal blijken, staat die fout vektor bijna volledig in de richting van de dominante eigenvektor. Dus *de richting* van $\vec{y}$ inklusief fout vektor heeft een kleine fout.

(5.3.24) Lemma. *Stel* $A\vec{v}_1 = \lambda_1\vec{v}_1$ *met* λ_1 *enkelvoudig en* $\|\vec{v}_1\|_2 = 1$.
Zij $V \in \mathbf{M}_n$ *de matrix met kolomvektoren* $\vec{v}_1, \vec{w}_2, \ldots, \vec{w}_n$*, waarbij* $\vec{w}_2, \ldots, \vec{w}_2$ *een orthonormale basis is voor* $\mathbf{W} := \mathrm{Im}\,(A - \lambda_1 I)$*. Dan* $V^{-1}AV = \begin{bmatrix} \lambda_1 & \vec{0}^* \\ \vec{0} & B \end{bmatrix}$ *voor 'n* $B \in \mathbf{M}_{n-1}$ (zie (4.6.1)).
Voor een $\vec{x} \in \mathbf{K}^n$ *beschouw de vergelijking* $(A + \Delta_A - \kappa I)\vec{y} = \vec{x}$.
Schrijf $\varrho := \|(B - \lambda_1 I)^{-1}\|_2^{-1}$, $\vec{x} = \alpha\vec{v}_1 + \vec{r}$ *en* $\vec{y} = \beta\vec{v}_1 + \vec{u}$ *met* $\vec{r}, \vec{u} \in \mathbf{W}$.
Verder is $t_0 := \|\frac{1}{\alpha}\vec{r}\|_2$, $t_1 := \|\frac{1}{\beta}\vec{u}\|_2$, $\eta := |\lambda_1 - \kappa|$ *en* $\varepsilon := \mathcal{C}_2(V)\|\Delta_A\|_2$*. Dan*

$$t_1 \leq \frac{t_0\eta + \varepsilon(1 + t_0)}{\varrho - \eta - \varepsilon(1 + t_0)} \quad \textit{als} \quad \varepsilon(1 + t_0) + \eta < \varrho. \tag{24}$$

Bewijs. Schrijf $\tilde{\eta} := \lambda_1 - \kappa$, $\tilde{B} := B - \kappa I$. Partioneer $V^{-1}\Delta_A V = \begin{bmatrix} \gamma & \vec{f}^* \\ \vec{g} & E \end{bmatrix}$ met $E \in \mathbf{M}_{n-1}$.
Schrijf $\tilde{\varepsilon} := \max(|\gamma|, \|\vec{f}\|_2, \|\vec{g}\|_2, \|E\|_2)$. Dan $\tilde{\varepsilon} \leq \varepsilon$ en $\|(\tilde{B} + E)^{-1}\|_2 \leq \frac{1}{\varrho - \eta - \tilde{\varepsilon}}$.
Met $\begin{bmatrix} \alpha \\ \tilde{r} \end{bmatrix} := V^{-1}\vec{x}$ en $\begin{bmatrix} \beta \\ \tilde{u} \end{bmatrix} := V^{-1}\vec{y}$ is $\left(\begin{bmatrix} \tilde{\eta} & \vec{0}^* \\ \vec{0} & \tilde{B} \end{bmatrix} + \begin{bmatrix} \gamma & \vec{f}^* \\ \vec{g} & E \end{bmatrix}\right)\begin{bmatrix} \beta \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} \alpha \\ \tilde{r} \end{bmatrix}$.
Schrijf $\hat{u} := \frac{1}{\beta}\tilde{u}$. Omdat $\vec{w}_2, \ldots, \vec{w}_n$ orthonormaal is, is $t_1 = \|\frac{\vec{u}}{\beta}\|_2 = \|\hat{u}\|_2$ en $t_0 = \|\frac{\tilde{r}}{\alpha}\|_2$.
Uitschrijven van de vergelijking levert $\tilde{\eta} + \gamma + (\hat{u}, \vec{f}\,) = \frac{\alpha}{\beta}$ en $(\tilde{B} + E)\hat{u} = \frac{1}{\beta}\tilde{r} - \vec{g}$.
Dus $|\frac{\alpha}{\beta}| \leq |\tilde{\eta}| + \tilde{\varepsilon} + \tilde{\varepsilon}\,\|\hat{u}\|_2$ en $t_1 = \|\hat{u}\|_2 \leq \frac{1}{\varrho - \eta - \tilde{\varepsilon}}(\frac{1}{|\beta|}\|\tilde{r}\|_2 + \tilde{\varepsilon}) = \frac{1}{\varrho - \eta - \tilde{\varepsilon}}(|\frac{\alpha}{\beta}|t_0 + \tilde{\varepsilon})$.
Blijkbaar $(\varrho - \eta - \tilde{\varepsilon})t_1 \leq |\frac{\alpha}{\beta}|t_0 + \tilde{\varepsilon} \leq \eta t_0 + \tilde{\varepsilon}t_0 + \tilde{\varepsilon}t_0\|\hat{u}\|_2 + \tilde{\varepsilon} = \eta t_0 + \tilde{\varepsilon}(1 + t_0) + \tilde{\varepsilon}t_0 t_1$
en de bewering volgt. □

We interpreteren het resultaat in bovenstaand lemma.

(5.3.25) De stabiliteit van de inverse machtsmethode. Beschouw (5.3.24).
Als $\eta \ll 1$ dan verwachten we dat de oplossing $\tilde{y}$ van de geperturbeerde vergelijking $(A + \Delta_A - \kappa I)\tilde{y} = \vec{x}$ sterk beïnvloed wordt door de perturbatie Δ_A. (Ga aan de hand van een 1×1 of 2×2 voorbeeld na dat de *grootte* van $\tilde{y}$ inderdaad sterk afhangt van de perturbatie.)

t_0 meet de hoek tussen $\vec{x}$ en $\vec{v}_1$ en t_1 meet de hoek tussen $\tilde{y}$ en $\vec{v}_1$ (als $A^* = A$ dan is t_i precies de tangens van die hoek; zie ook (5.3.8)). Als $\varepsilon \ll \varrho$, $\eta \ll \varrho$ en $t_0 \leq 1$ dan $t_1 \lessapprox t_0 \frac{\eta}{\varrho} + 2\frac{\varepsilon}{\varrho}$. Door de inverse iteratie slag wordt de hoek tussen de iteranden (voor niet al te kleine t_0) sterk gereduceerd. In de ongeperturbeerde situatie ($\varepsilon = 0$) met een faktor $\approx \frac{\eta}{\varrho}$. Ten gevolge van de perturbatie term wordt de richting van de resulterende vektor ten hoogste met $2\frac{\varepsilon}{\varrho}$ verstoord. Met betrekking tot het effect van de perturbatie kunnen we, gezien het resultaat in (c) van (4.6.6), geen beter resultaat verwachten.

De inverse machtsmethode met *stationaire verschuiving* als in (5.3.21) wordt vaak gebruikt om in een paar iteratie slagen de eigenvektor te benaderen van een eigenwaarde die al in grote precisie bekend is (κ_k is dan voor iedere k die nauwkeurige benaderende eigenwaarde). In zo'n geval spreekt men ook wel over *Wielandt iteratie*.

De inverse machtsmethode levert de snelst konvergente rij als we voor de verschuiving $\kappa \approx \lambda_1$ kunnen kiezen. We moeten dan wel over een schatting voor λ_1 beschikken. Als de inverse machtsmethode een konvergente rij produceert dan levert die methode zelf ons precies zo'n schatting. In het algorithme (5.3.20) lijkt $\kappa_k = \mu_k$ voor iedere k geen kwade keuze.

(5.3.26) Stelling. *Zij* A, λ_1, $\vec{v}_1$, $\mathbf{W}$ *en* ϱ *als in* (5.3.24). *Laat* $(\vec{x}_k)$ *door de inverse machtsmethode geproduceerd zijn met verschuivingen* $\kappa_k := (A\vec{x}_k, \vec{x}_k)$ $(k \in \mathbf{N})$.
Zij (β_k) *in* $\mathbf{K}$ *zodat* $\vec{r}_k := \vec{x}_k - \beta_k \vec{v}_1 \in \mathbf{W}$. *Schrijf* $t_k := \|\frac{1}{\beta_k}\vec{r}_k\|_2$ $(k \in \mathbf{N})$.
Als $t_0 < \frac{s\varrho}{4\|A\|_2}$ *met* $s := \sin\angle(\vec{v}_1, \mathbf{W})$ *dan* $t_{k+1} < t_k$,

$$t_{k+1} < \frac{4\|A\|_2}{s\varrho} t_k^2 \quad \text{en} \quad |(A\vec{x}_k, \vec{x}_k) - \lambda_1| \leq \frac{2}{s}\|A\|_2 t_k \quad \text{voor iedere } k.$$

Als $A^* = A$ *en* $t_0 < \sqrt{\frac{\varrho}{4\|A\|_2}}$ *dan* $t_{k+1} < t_k$,

$$t_{k+1} < \frac{4\|A\|_2}{\varrho} t_k^3 \quad \text{en} \quad |(A\vec{x}_k, \vec{x}_k) - \lambda_1| \leq 2\|A\|_2 t_k^2 \quad \text{voor iedere } k.$$

Bewijs. We analyseren de k-de stap.

Met $\vec{x} = \alpha\vec{v}_1 + \vec{r} := \beta_k \vec{v}_1 + \vec{r}_k = \vec{x}_k$ en $\vec{y} = \beta\vec{v}_1 + \vec{u} := \vec{y}_{k+1}$ is $(A - (A\vec{x}, \vec{x})I)\vec{y} = \vec{x}$. In (5.3.12) hebben we gezien dat $\alpha = \frac{(\vec{x}, \vec{w}_1)}{(\vec{v}_1, \vec{w}_1)}$ als $\vec{w}_1$ een links eigenvektor is bij de eigenwaarde λ_1. Dus $|\alpha| \leq \frac{\|\vec{w}_1\|_2}{|(\vec{v}_1, \vec{w}_1)|} = \frac{1}{s}$.
Nu is $t_k = \|\frac{1}{\alpha}\vec{r}\|_2$ en, omdat $\vec{y}_{k+1} = \|\vec{y}_{k+1}\|_2 \vec{x}_{k+1}$, is $t_{k+1} = \|\frac{1}{\beta}\vec{u}\|_2$. Schrijf $\eta := |(A\vec{x}, \vec{x}) - \lambda_1|$.

Neem in (5.3.24) $\kappa = (A\vec{x}, \vec{x})$ en $\Delta_A = 0$ zodat $\varepsilon = 0$. Dan zien we dat $t_{k+1} \leq \frac{\eta}{\varrho - \eta} t_k$.
Uit (5.3.5) volgt dat $\eta \leq \|A - \lambda_1 I\|_2 \|\vec{r}\|_2 \leq (\|A\|_2 + |\lambda_1|) t_k |\alpha| \leq 2\|A\|_2 \frac{t_k}{s}$.
Als $t_k < \frac{s\varrho}{4\|A\|_2}$ dan $\eta < \frac{1}{2}\varrho$ en dus $t_{k+1} < 2\frac{\eta}{\varrho} t_k \leq 4\frac{\|A\|_2}{s\varrho} t_k^2 \leq t_k$, waarmee we de eerste serie ongelijkheden bewezen hebben.

Als A hermiets is, is $|\alpha| \leq 1$ en volgt uit (5.3.5) dat $\eta \leq 2\|A\|_2 t_k^2$. De overige ongelijkheden volgen als in de vorige alinea. □

De inverse machtsmethode (5.3.20) met verschuivingen $\kappa_k := (A\vec{x}_k, \vec{x}_k)$ $(k \in \mathbf{N})$ noemt men ook wel *Rayleigh quotiënt iteratie*.

Als $\vec{x}_0$ een grote komponent heeft in de richting van $\vec{v}_1$ dan produceert de Rayleigh quotiënt iteratie een rij $(\vec{x}_k)$ die *gemajoreerd kwadratisch konvergeert in richting* naar $\vec{v}_1$. Als A hermiets is *konvergeert* $(\vec{x}_k)$ zelfs *gemajoreerd kubisch in richting* naar $\vec{v}_1$.

Bij de Wielandt iteratie hebben we in iedere iteratie slag te maken met dezelfde matrix: de kosten van stap (b2) in algorithme (5.3.20) zijn dan ook maar bij een stap extra groot (ten opzichte van een matrix-vektor vermenigvuldiging). Bij de Rayleigh quotiënt iteratie hebben we in iedere stap met een andere matrix te maken zodat iedere stap extra kosten vergt. Deze kosten vallen echter bijzonder mee in geval A een irreducibele Hessenberg matrix is (zie (5.3.22)).

De zeer snelle konvergentie betekent vanuit een praktisch standpunt dat de methode in feite eindig is: als we na een aantal iteratie slagen een benadering hebben van λ_1 in bijvoorbeeld

8 cijfers dan mogen we in de volgende iteratie (indien geschikt geschaald) een nauwkeurigheid van 16 cijfers tegemoed zien. Na een aantal iteraties kunnen we daarna voor nauwelijks meer werk iedere nauwkeurigheid bereiken die we wensen.

(5.3.27) Opmerking. De Rayleigh quotiënt iteratie kan, als de komponent van de startvektor in de richting van een eigenvektor niet voldoende groot is, stagneren, zelfs in geval A hermiets is en alle eigenwaarden enkelvoudig zijn (zie opgave (5.3.28)). Voor hermietse matrices is het onwaarschijnlijk dat dat gebeurt. Rekenfouten zullen de konvergentie weer op gang brengen.

(5.3.28) Opgave. Zij $A := \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Neem $\vec{x}_0 := e_1$. Passen we de inverse machtsmethode toe met verschuiving als in (5.3.26) dan $\vec{x}_{2k} = e_1$, $\vec{x}_{2k+1} = e_2$ en $\mu_k = 0$ voor iedere k: de methode produceert geen konvergente rij vektoren. Ga dit na.

(5.3.29) opgave.
(a) Bewijs dat $\|A\vec{x} - (A\vec{x}, \vec{x})\vec{x}\|_2 \le \|A\vec{x} - \kappa\vec{x}\|_2$ voor iedere $\kappa \in \mathbf{K}$, $\vec{x} \in \mathbf{K}^n$.
(b) Stel dat A hermiets is. Laat $(\vec{x}_k)$ geproduceerd zijn door de Rayleigh quotiënt iteratie. $\kappa_k := (A\vec{x}_k, \vec{x}_k)$. Bewijs dat $\|A\vec{x}_{k+1} - \kappa_{k+1}\vec{x}_{k+1}\|_2 \le \|A\vec{x}_k - \kappa_k\vec{x}_k\|_2$ voor iedere k.
(Hint: $\|(A - \kappa_{k+1}I)\vec{x}_{k+1}\|_2 \le \|(A - \kappa_k I)\vec{x}_{k+1}\|_2 = (x_k, (A - \kappa_k I)\vec{x}_{k+1})$.)

5.4 Het QR-algorithme

In de vorige paragraaf hebben we gezien hoe we met behulp van de machtsmethode, de inverse machtsmethode, deflatietechnieken en spektrale verschuivingen eigenwaarden en eigenvektoren kunnen benaderen. Het QR-algorithme kombineert in feite deze vier technieken in een zeer efficiënt algorithme. Voordat we het algorithme kunnen presenteren en doorzien moeten we nog wat meer theorie ontwikkelen. Zoals vaak met efficiënte algorithmen het geval is heeft het QR-algorithme ogenschijnlijk niets met zijn theoretische achtergrond te maken.

We presenteren en analyseren het QR-algorithme voor matrices met enkelvoudige eigenwaarden als in de volgende alinea. Het QR-algorithme werkt ook voor matrices uit een grotere klasse, maar daar besteden we hier slechts terloops aandacht aan.

(5.4.1) Zij $A \in \mathbf{M}_n$ met eigenwaarden $\lambda_1, \ldots, \lambda_n$ zodat
$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_{n-1}| > |\lambda_n|.$$
Laat $\vec{v}_i \in \mathbf{K}^n \backslash \{0\}$ de eigenvektor bij λ_i zijn.
Met $\Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ en $V \in \mathbf{M}_n$ zodat $Ve_i = \vec{v}_i \quad (i = 1, \ldots, n)$ is $AV = V\Lambda$.

Om stabiliteits en efficiëntie redenen bepaalt men liever de Schur normaalvorm van A dan de eigenvektoren (rechtstreeks). In het volgende lemma geven we (ondermeer om de notatie vast te leggen) een verband tussen de Schur normaalvorm en de basis van eigenvektoren; we laten het bewijs aan de lezer over. Met het QR-algorithme zullen we de Schur normaalvorm benaderen.

(5.4.2) Lemma. *Laat $U \in \mathbf{M}_n$ unitair en $\Phi \in \mathbf{M}_n$ bovendriehoek zijn zodat* $V = U\Phi$. *Dan is* $S := \Phi\Lambda\Phi^{-1}$ *boven driehoek,* $\operatorname{diag}(S) = \Lambda$ *en* $A = USU^*$. □

(5.4.3) Opgave. Stel dat S berekend is.
Ga na dat alle eigenvektoren van S in $\approx \frac{1}{6}n^3$ flop te bepalen zijn.
Als $A = A^*$ dan $S = \Lambda$. Ga dit na.

De machtsmethode voor deelruimten

Als bezwaar tegen de inverse machtsmethode met spektrale verschuiving kan men aanvoeren dat men het proces voor iedere eigenwaarde en eigenvektor weer opnieuw moet starten. Bovendien moet men met zorg telkens de startvektoren kiezen. Het QR-algorithme benadert terwijl het de eerste eigenvektor benadert impliciet ook de andere eigenvektoren.

De machtsmethode van de vorige paragraaf produceert een rij vektoren uit een startvektor. Het QR-algorithme is gebaseerd op de machtsmethode toegepast op $n \times l$ matrices (of, zo men wil, op l-dimensionale lineaire deelruimten; zie (5.4.6)).

(5.4.4) Lemma. *Zij A als in* (5.4.1). *Beschouw een $l \in \{1, \ldots, n\}$. Zij $X_0 \in \mathbf{M}_{n\times l}$. Voor iedere $k \in \mathbf{N}$ is $W_k \in \mathbf{M}_{n\times l}$ en $R_k \in \mathbf{M}_l$ zodat*

$$A^k X_0 = W_k R_k, \quad W_k^* W_k = I \quad \textit{en} \quad R_k \quad \textit{boven driehoek.}$$

Partitioneer $V = [\, V' \quad V'' \,]$ *en* $U = [\, U' \quad U'' \,]$ *met* $V', U' \in \mathbf{M}_{n\times l}$.
Schrijf $X_0 = V'C + V''D$ *en* $W_k = U'E_k + U''G_k$ *met* $C, E_k \in \mathbf{M}_l$ *en* $D, G_k \in \mathbf{M}_{(n-l)\times l}$

(dus $V^{-1}X_0 = \begin{bmatrix} C \\ D \end{bmatrix}$ en $U^*W_k = \begin{bmatrix} E_k \\ G_k \end{bmatrix}$).

Als C niet-singulier is dan is er een $\Gamma > 0$ zodat $\|G_k\|_2 \leq \Gamma |\frac{\lambda_{l+1}}{\lambda_l}|^k$ voor alle k.

Bewijs. Schrijf $\Phi' := U'^* V'$ (Φ' is de $l \times l$ bovenblok van Φ),
$\Lambda' := \mathrm{diag}\,(\lambda_1, \ldots, \lambda_l)$ en $\Lambda'' := \mathrm{diag}\,(\lambda_{l+1}, \ldots, \lambda_n)$.

$$A^k X_0 = V'(\Lambda')^k C + V''(\Lambda'')^k D = (U' + \Delta_k)\Phi'(\Lambda')^k C$$

met $\Delta_k := V''(\Lambda'')^k DC^{-1}(\Lambda')^{-k}(\Phi')^{-1}$. Merk op dat

$$\|\Delta_k\| \leq \Gamma' |\frac{\lambda_{l+1}}{\lambda_l}|^k \quad \text{waarbij} \quad \Gamma' := \|V\|_2 \, \|DC^{-1}\|_2 \, \|(\Phi')^{-1}\|_2. \tag{25}$$

Met $B_k := \Phi'(\Lambda')^k C$ is

$$U'E_k R_k + U''G_k R_k = W_k R_k = A^k X_0 = (U' + \Delta_k)B_k. \tag{26}$$

Merk op dat $(U'')^*U' = 0$, $(U'')^*U'' = I_l$. Dus volgt uit (26) dat $G_k = (U'')^*\Delta_k B_k R_k^{-1}$ en $E_k = (I_l + (U')^*\Delta_k)B_k R_k^{-1}$. Volgens (25) is, voor k voldoende groot, $\|\Delta_k\|_2 < 1$ en is $I_l + (U')^*\Delta_k$ inverteerbaar. Dan $B_k R_k^{-1} = (I_l + (U')^*\Delta_k)^{-1}E_k$ en we vinden dat $G_k = (U'')^*\Delta_k(I_l + (U')^*\Delta_k)^{-1}E_k$. Omdat U unitair is, W_k orthonormaal en $E = (U')^*W_k$ volgt dat $\|E_k\|_2 \leq 1$ en dus $\|G_k\|_2 \leq \frac{\|\Delta_k\|_2}{1-\|\Delta_k\|_2}$. Het lemma volgt nu. □

In feite is bovenstaand lemma een generalizering van stelling (5.3.6). Het werkt verhelderend om de argumenten in bovenstaand bewijs na te lopen voor het geval $X_0 \in \mathbf{M}_{n\times 1}$. In de volgende opgave vragen we dit stap voor stap te doen.

(5.4.5) Opgave. Zij $X_0 = \vec{x}_0 = c\vec{v}_1 + \sum_{i>1} \alpha_i \vec{v}_i$. Neem $\vec{v}_i$ zo dat $\|\vec{v}_i\|_2 = 1$.

Kontroleer de volgende beweringen, vergelijk die met die in het lemma en het bewijs en die in stelling (5.3.6).

$A^k\vec{x}_0 = (\vec{v}_1 + \vec{\delta}_k)c\lambda_1^k$ met $\vec{\delta}_k := \sum_{i>1} \frac{\alpha_i}{c}(\frac{\lambda_i}{\lambda_1})^k \vec{v}_i$ als $c \neq 0$. Met $\Gamma := \sum_{i>1} |\frac{\alpha_i}{c}|$ is $\delta_k := \|\vec{\delta}_k\|_2 \leq \Gamma|\frac{\lambda_2}{\lambda_1}|^k$. Met $r_k := \|A^k\vec{x}_0\|_2$ en $\vec{w}_k := \frac{1}{r_k}A^k\vec{x}_0$ is $A^k\vec{x}_0 = \vec{w}_k r_k$. Dus $\vec{w}_k = (\vec{v}_1 + \vec{\delta}_k)b$ waarbij $b := \frac{1}{r_k}c\lambda_1^k$ en $(\vec{w}_k, \vec{v}_1) = (1 + (\vec{\delta}_k, \vec{v}_1))b$. Omdat $\|\vec{w}_k\|_2 = \|\vec{v}_1\|_2 = 1$ is $|b| \leq \frac{1}{1-\delta_k}$. Schrijf $\vec{w}_k = \vec{v}_1 e_k + \vec{g}_k$ met $\vec{g}_k \perp \vec{v}_1$. Dan $\vec{g}_k = \vec{w}_k - (\vec{w}_k, \vec{v}_1)\vec{v}_1 = (\vec{\delta}_k - (\vec{\delta}_k, \vec{v}_1)\vec{v}_1)b$ en $\|\vec{g}_k\|_2 \leq \|\vec{\delta}_k\|_2 \, |b| \leq \frac{\delta_k}{1-\delta_k}$.

We interpreteren de uitspraak in lemma (5.4.4) in termen van deelruimten en hoeken daartussen.

(5.4.6) Opmerking. Beschouw de situatie in (5.4.4).
Beschouw de ruimte $\mathbf{V}'$ opgespannen door $\vec{v}_1, \ldots, \vec{v}_l$. Dan wordt $\mathbf{V}'$ ook opgespannen door het orthonormaal stelsel $Ue_1, \ldots, Ue_l$ (dit stelsel kan immers gezien worden als een resultaat van de Gram–Schmidt procedure toegepast op $\vec{v}_1, \ldots, \vec{v}_l$).

Beschouw de ruimte $\mathbf{W}_k$ opgespannen door de kolom vektoren van $A^k X_0$. $\mathbf{W}_k$ wordt ook opgespannen door het orthonormaal stelsel $W_k e_1, W_k e_2, \ldots, W_k e_l$.

Omdat $G_k = (U'')^* W_k$ vertelt (c) van (4.6.19) ons dat $\|G_k\|_2 = \sin\angle(\mathbf{W}_k, \mathbf{V}')$ is. We kunnen dus zeggen dat de machtsmethode een rij $(\mathbf{W}_k)$ van l-dimensionale lineaire deelruimten produceert die konvergeert naar de ruimte $\mathbf{V}'$ opgespannen door de l eigenvektoren $\vec{v}_1, \ldots, \vec{v}_l$.

Met betrekking tot de konditie "C niet-singulier" in lemma (5.4.4) kunnen we dezelfde opmerkingen maken als in (5.3.11).

In de volgende bewering geven we een drietal equivalente formuleringen en laten we zien hoe een irreducibele Hessenberg matrix een geschikt X_0 levert.

(5.4.7) Bewering. *Zij A als in* (5.4.1). *Zij* $X_0 \in \mathbf{M}_{n\times l}$. *Schrijf* $\vec{x}_i := X_0 e_i$.
Partitioneer $V = \begin{bmatrix} V' & V'' \end{bmatrix}$ *met* $V' \in \mathbf{M}_{n\times l}$.
Schrijf $X_0 = V'C + V''D$ *met* $C \in \mathbf{M}_l$ *en* $D \in \mathbf{M}_{(n-l)\times l}$.
C niet-singulier $\Leftrightarrow$ $\begin{bmatrix} X_0 & V'' \end{bmatrix}$ *niet-singulier* $\Leftrightarrow$ $\vec{x}_1, \ldots, \vec{x}_l, \vec{v}_{l+1}, \ldots, \vec{v}_n$ *is een basis voor* $\mathbf{K}^n$.
Als A een irreducibele Hessenberg matrix is en $X = \begin{bmatrix} I_l \\ 0 \end{bmatrix}$ *dan is C niet-singulier.*

Bewijs. $V^{-1}\begin{bmatrix} X_0 & V'' \end{bmatrix} = \begin{bmatrix} C & 0 \\ D & I_{n-l} \end{bmatrix}$, waaruit de eerste equivalenties volgen.

Met $\mathbf{V}'' := \operatorname{span}(\vec{v}_{l+1}, \ldots, \vec{v}_n)$ is $A(\mathbf{V}'') = \mathbf{V}''$ en dus is $\operatorname{span}(e_1, \ldots, e_l) \cap \mathbf{V}'' = \{0\}$ als A een irreducibele Hessenberg matrix is (zie stelling (5.1.10)). Blijkbaar is dan $e_1, \ldots, e_l, \vec{v}_{l+1}, \ldots, \vec{v}_n$ een basis voor $\mathbf{K}^n$. □

Beschouw voor een $n \times n$ matrix X_0 en een voldoend grote k de QU-dekompositie WR van $A^k X_0$, waarbij W unitair en R boven driehoeks. De volgende stelling zegt dat we mogen verwachten dat de j-de kolom van W op een teken na min of meer de j-de kolom van U is voor iedere j.

(5.4.8) Stelling. *Zij A als in* (5.4.1).
Zij $X_0 \in \mathbf{M}_n$ zodat ieder $l \times l$ links bovenblok van $V^{-1}X_0$ niet-singulier is.
(Als A een irreducibele Hessenberg matrix is, dan heeft $X_0 = I$ deze eigenschap; zie (5.4.7).)
Voor iedere $k \in \mathbf{N}$ is $W_k \in \mathbf{M}_n$ en $R_k \in \mathbf{M}_n$ zodat

$$A^k X_0 = W_k R_k, \quad W_k^* W_k = I \quad \text{en} \quad R_k \quad \text{boven driehoek.}$$

Dan is er een rij (J_k) diagonaal matrices in $\mathbf{M}_n$ met

$$|J_k| = I, \quad \lim_{k\to\infty} \|W_k - UJ_k\|_2 = 0 \quad \text{en} \quad \lim_{k\to\infty} \|W_k^* A W_k - J_k^* S J_k\|_2 = 0.$$

Bewijs. Schrijf $U^* W_k = \begin{bmatrix} E_k & F_k \\ G_k & H_k \end{bmatrix}$ met $E_k \in \mathbf{M}_l$. Volgens lemma (5.4.4) is er een Γ zodat $\|G_k\|_2 \le \Gamma |\frac{\lambda_{l+1}}{\lambda_l}|^k$ $(k \in \mathbf{N})$ (beschouw, om het lemma toe te passen, de eerste l kolommen van X_0, W_k, etc.). Omdat $U^* W_k$ unitair is geldt (zie lemma (4.6.17) en het bewijs van dat lemma)

$$\|I_l - E_k^* E_k\|_2 = \|G_k^* G_k\|_2 = \|G_k\|_2^2 = \|F_k\|_2^2 = \|I_{n-l} - H_k^* H_k\|_2.$$

Omdat voor iedere l de rijen (G_k) en (F_k) van buitendiagonaal blokken naar 0 konvergeren, konvergeert $(U^* W_k - \operatorname{diag}(U^* W_k))$ ook naar 0.
Laat nu $J_k := \operatorname{diag}(\theta_1, \ldots, \theta_n)$ waarbij $\theta_i := \frac{1}{|e_{ii}|} e_{ii}$ als $(e_{ij}) = U^* W_k$.
Omdat voor iedere l de rij $(I_l - E_k^* E_k)$ naar 0 konvergeert, konvergeert $(U^* W_k - J_k)$ naar 0.

Merk tenslotte op dat $W_k^* A W_k = W_k^* U S U^* W_k$ waaruit volgt dat

$$W_k^* A W_k - J_k^* S J_k \to 0 \quad (k \to \infty).$$ □

$(W_k^* A W_k)$ konvergeert (op tekens na) naar S. Merk op dat $\operatorname{diag} J_k^* S J_k = \operatorname{diag} S$. De konvergentie is gemajoreerd lineair. We geven de konvergentie faktoren.

(5.4.9) Bewering. *Beschouw de situatie in* (5.4.8).
Schrijf $\omega_j := |\frac{\lambda_{j+1}}{\lambda_j}|$ *voor* $j = 1, \ldots, n$. *Zij* $l \in \{1, \ldots, n\}$. *Voor* $k \in \mathbf{N}$ *is* $\lambda_l^{(k)} := e_l^* W_k^* A W_k e_l$
(het l-de diagonaal element van de benadering $W_k^* A W_k$ van S).
Dan $|\lambda_l^{(k)} - \lambda_l| = \mathcal{O}(\omega^k)$ $(k \to \infty)$ *met* $\omega := \max(\omega_{l-1}, \omega_l)$.

Als $A^* = A$ *dan* $|\lambda_l^{(k)} - \lambda_l| = \mathcal{O}(\tilde\omega^{2k})$ $(k \to \infty)$ *met* $\tilde\omega := \min(\omega_{l-1}, \omega_l)$.

Bewijs. Schrijf $U^*W_k = \begin{bmatrix} E_k & F_k \\ G_k & H_k \end{bmatrix} = \begin{bmatrix} \tilde E_k & \tilde F_k \\ \tilde G_k & \tilde H_k \end{bmatrix}$ met $E_k \in \mathbf{M}_l$ en $\tilde E_k \in \mathbf{M}_{l-1}$.
Schrijf $U^*W_k e_l = \beta_k e_l + \vec r_k$ met $\vec r_k \perp e_l$.

Dan geldt (ga dit na):
(a) $\|G_k\|_2 = \mathcal{O}(\omega_l^k)$ en $\|\tilde F_k\|_2 = \mathcal{O}(\omega_{l-1}^k)$ $(k \to \infty)$ (zie bewijs (5.4.8));
(b) $1 - |\beta_k|^2 \le \|I_l - E_k^* E_k\|_2 = \|G_k\|_2^2$ en $1 - |\beta_k|^2 \le \|I_{n-l+1} - \tilde H_k^* \tilde H_k\|_2 = \|\tilde F_k\|_2^2$;
(c) $|(Se_l, \vec r_k)| \le \|S\|_2 \|\tilde F_k\|_2$; (d) $|(S\vec r_k, \vec r_k)| \le \|S\|_2(\|\tilde F_k\|_2^2 + \|G_k\|_2^2)$;
(e) $\lambda_l^{(k)} = (S(\beta_k e_l + \vec r_k), \beta_k e_l + \vec r_k) = |\beta_k|^2\lambda_l + \beta_k(Se_l, \vec r_k) + \bar\beta_k(S\vec r_k, e_l) + (S\vec r_k, \vec r_k)$.
Uit (e), de Cauchy–Schwarz ongelijkheid, (b), (c), (d) en (a) volgt de bewering. □

(5.4.10) Opgave. Zij A en X_0 als in stelling (5.4.8). Kies een $\kappa \in \mathbf{K}$.
Voor iedere $k \in \mathbf{N}$ is $W_k \in \mathbf{M}_n$ en $R_k \in \mathbf{M}_n$ zodat
$$(A - \kappa I)^k X_0 = W_k R_k, \quad W_k^* W_k = I \quad \text{en} \quad R_k \ \text{boven driehoek.}$$
Als $|\lambda_{i+1} - \kappa| < |\lambda_i - \kappa|$ voor iedere $i = 1, \ldots, n-1$ dan is er een rij (J_k) diagonaal matrices in $\mathbf{M}_n$ met
$$|J_k| = I, \quad \lim_{k\to\infty} \|W_k - UJ_k\|_2 = 0 \quad \text{en} \quad \lim_{k\to\infty} \|W_k^* A W_k - J_k^* S J_k\|_2 = 0.$$
Bewijs dit.

(5.4.11) Opmerking. Zij A een irreducibele Hessenberg matrix.
Voor iedere k is W_k weer unitair zodat $W_k^* A^k$ boven driehoeks is als in stelling (5.4.8).
Laat $\lambda_1, \ldots, \lambda_n$ de eigenwaarden van A zijn geteld naar multipliciteit. Ook als sommige eigenwaarden gelijke absolute waarde hebben konvergeert $(W_k^* A W_k)$ in zekere zin:
Er is een blok boven driehoeks matrix S' waarbij er een $1-1$ korrespondentie is tussen de diagonaal blokken van S' en de elementen in $\sigma' := \{|\lambda_j| \,\big|\, j\}$. Het μ-blok, d.w.z. het diagonaal blok dat korrespondeert met $\mu \in \sigma'$, heeft afmeting $l_\mu := \#\{j \,\big|\, |\lambda_j| = \mu\}$. De diagonaal blokken korresponderen van boven naar beneden met 'dalende μ's'. Verder is er een rij blok diagonaal matrices J_k' waarvan ieder 'μ-blok' $l_\mu \times l_\mu$ unitair is en $W_k^* A W_k - J_k'^* S' J_k' \to 0$ als $k \to \infty$.

Het algorithme

We zullen nu een algorithme konstrueren uit de machtmethode in stelling (5.4.8).
De QU-dekompositie "$A^k X_0 = W_k R_k$" is de generalizering van de schalingsstap (b3) in algorithme (5.3.3).

Ter inleiding van het QR-algorithme bekijken we eerst even de rechtstreekse generalizering van algorithme (5.3.3): we schalen in iedere stap.

(5.4.12) Produceer als volgt een rijen (W_k) en (R_k) in $\mathbf{M}_n$ met W_k unitair en R_k boven driehoek:

(a1) $k := 0$, X_0 `is gegeven;`
(a2) `bepaal de QU-dekompositie` $X_0 = W_0 \tilde R_0$ `van` X_0`;`
(b1) `bereken` AW_k`;`
(b2) `bepaal de QU-dekompositie` $AW_k = W_{k+1}\tilde R_{k+1}$ `van` AW_k`;`
(c) $k := k+1$`; ga na naar` (b1)`;`

Dan $A^k X_0 = W_k(\tilde R_k \tilde R_{k-1} \cdots \tilde R_1 \tilde R_0) = W_k R_k$ met R_k als in stelling (5.4.8). We zijn geïnteresseerd in $A_k := W_k^* A W_k$ $(k \in \mathbf{N})$ immers $\lim_{k\to\infty} A_k - J_k^* S J_k = 0$ als X_0 geschikt is (zie (5.4.8)).
Definieer $Q_k := W_{k-1}^* W_k$ $(k \in \mathbf{N})$. Dan is Q_{k+1} unitair,

$$W_k = W_{k-1} W_{k-1}^* W_k = W_{k-1} Q_k = \ldots = W_0 Q_1 Q_2 \cdots Q_k \qquad (k \in \mathbf{N}) \tag{27}$$

en (volgens (b2))
$$A_k = W_k^* A W_k = Q_{k+1} \tilde R_{k+1}. \tag{28}$$

Dus $\tilde{R}_{k+1} = Q_{k+1}^* A_k$ en

$$A_{k+1} = W_{k+1}^* A W_{k+1} = W_{k+1}^* W_k A_k W_k^* W_{k+1} = Q_{k+1}^* A_k Q_{k+1} = \tilde{R}_{k+1} Q_{k+1}. \tag{29}$$

Is A een irreducibele Hessenberg matrix dan levert (28) en (29) een berekenings methode voor (A_k) die naar S "konvergeert". We kunnen nu het QR-algorithme geven.

(5.4.13) Het QR-algorithme met verschuiving. Zij (κ_k) een rij (nog nader te kiezen) verschuivingen ("shifts") in $\mathbf{C}$. Het *QR-algorithme met verschuivingen* (κ_k) produceert als volgt rijen (A_k) en (Q_k) in $\mathbf{M}_n$ met Q_k unitair:

QR-Algorithme

(a1) $k := 0$;
(a2) `bepaal` W_0 `unitair en` A_0 `Hessenberg zodat` $A_0 = W_0^* A W_0$;
(b1) `reduceer` A_k `indien mogelijk` (zie (5.1.6));
(b2) `kies` κ_k;
(b3) `bepaal de QU-dekompositie` $A_k - \kappa_k I = Q_{k+1} \tilde{R}_{k+1}$;
(b4) `bereken` $A_{k+1} := \tilde{R}_{k+1} Q_{k+1} + \kappa_k I$;
(c) $k := k+1$; `ga na naar` (b1);

Voor een keuze van κ_k zie (5.4.16). Voor enige toelichting bij de reduktie strategie zie (5.4.18).

Voor het QR-algorithme zonder verschuiving hebben we inmiddels het volgende bewezen (ga dat na).

(5.4.14) Gevolg. *Zij A als in (5.4.1). Zij $W_k := W_0 Q_1 \cdots Q_k$.*
Als $\kappa_k = 0$ $(k \in \mathbf{N})$ dan is er een rij (J_k) diagonaal matrices in $\mathbf{M}_n$ met

$$|J_k| = I, \quad \lim_{k\to\infty} \|W_k - U J_k\|_2 = 0 \quad \text{en} \quad \lim_{k\to\infty} \|A_k - J_k^* S J_k\|_2 = 0. \qquad \square$$

We hopen met geschikte verschuivingen de konvergentie te kunnen versnellen. De volgende stelling legt een verband met de inverse machtsmethode in (5.3.20). Stelling (5.3.26) suggereert ons daardoor een keuze voor de verschuivingen.

(5.4.15) Stelling. *Beschouw de situatie in* (5.4.13).
(a) *Met* $\vec{x}_k := W_k e_1$ *en* $\vec{y}_{k+1} := (A - \kappa_k I)\vec{x}_k$ *is*
$\vec{x}_{k+1} = \frac{\theta}{\|\vec{y}_{k+1}\|_2} \vec{y}_{k+1}$ *zekere* $\theta \in \mathbf{C}$, $|\theta| = 1$ *en* $(A\vec{x}_k, \vec{x}_k) = (A_k e_1, e_1)$ $(k \in \mathbf{N})$.
(b) *Met* $\vec{x}_k := W_k e_n$ *en* $\vec{y}_{k+1}$ *zodat* $(A^* - \bar{\kappa}_k I)\vec{y}_{k+1} := \vec{x}_k$ *is*
$\vec{x}_{k+1} = \frac{\theta}{\|\vec{y}_{k+1}\|_2} \vec{y}_{k+1}$ *zekere* $\theta \in \mathbf{C}$, $|\theta| = 1$ *en* $(A\vec{x}_k, \vec{x}_k) = (A_k e_n, e_n)$ $(k \in \mathbf{N})$.

Bewijs. (a)

$$(A - \kappa_k I)\vec{x}_k = (AW_k - \kappa_k W_k)e_1 \overset{(28)}{=} (W_k A_k - \kappa_k W_k)e_1 = W_k(A_k - \kappa_k I)e_1 \overset{(b.2)}{=}$$
$$W_k Q_{k+1} \tilde{R}_{k+1} e_1 = W_{k+1} \tilde{R}_{k+1} e_1 = \alpha W_{k+1} e_1 = \alpha \vec{x}_{k+1}$$

met $\alpha = (\tilde{R} e_1, e_1)$ (immers $\tilde{R}_k$ is boven driehoeks).
(b) Uit (b.3) volgt, door adjungatie, $A_{k+1}^* := Q_{k+1}^* \tilde{R}_{k+1}^* + \bar{\kappa}_k I$.

$$(A^* - \bar{\kappa}_k I)\vec{x}_{k+1} = (A^* W_{k+1} - \bar{\kappa}_k W_{k+1})e_n \overset{(29)}{=} (W_{k+1} A_{k+1}^* - \bar{\kappa}_k W_{k+1})e_n =$$
$$W_{k+1}(A_{k+1}^* - \bar{\kappa}_k I)e_n = W_{k+1} Q_{k+1}^* \tilde{R}_{k+1}^* e_n = W_k \tilde{R}_{k+1}^* e_n = \beta W_k e_n = \beta \vec{x}_k$$

met $\beta = (\tilde{R}_{k+1}^* e_n, e_n)$ (immers $\tilde{R}_{k+1}^*$ is beneden driehoeks). $\square$

(5.4.16) De keuze van de verschuivingen. We kunnen blijkbaar $((A_k e_1, e_1))$ zien als een rij benaderingen voor λ_1 die geproduceerd is door de gewone machtsmethode met verschuiving (zie (a)) en $((A_k e_n, e_n))$ als een rij benaderingen voor λ_n die geproduceerd is door de inverse machtsmethode met verschuiving toegepast met A^* en verschuivingen $(\bar{\kappa}_k)$ (zie (b)).
(a) Als $|\lambda_{n-1}| > |\lambda_n|$ levert, volgens (5.3.26) en (5.4.11), *de Rayleigh quotiënt verschuiving*

$$\kappa_k := (A_k e_n, e_n) \quad \text{voor } k \geq k_0 \tag{30}$$

en $\kappa_k = 0$ voor $k < k_0$, bij een geschikte k_0, een rij $((A_k e_n, e_n))$ die gemajoreerd kwadratisch konvergeert naar λ_n en zelfs kubisch als A hermiets is
(in feite levert de keuze $\bar{\kappa}_k = (A_k^* e_n, e_n)$ voor $k \geq k_0$ een rij $(\overline{(A_k^* e_n, e_n)})$ die gemajoreerd kwadratisch naar $\bar{\lambda}_n$ konvergeert. Echter $(A_k^* e_n, e_n) = (e_n, A_k e_n) = \overline{(A_k e_n, e_n)}$.

Als bijvoorbeeld $\lambda_1 > \lambda_2 > \ldots > \lambda_n \geq 0$ dan heeft de spektrale verschuiving in de richting van λ_n $(\kappa_k \approx \lambda_n)$ ook nog een positief effekt op de konvergentie naar de overige eigenwaarden (zie opgave (5.4.10)).)

(b) Zelfs in geval A hermiets is levert de keuze in (30) voor $k_0 = 0$ *niet* voor iedere A een rij $(A_k e_n, e_n)$ die naar een eigenwaarde van A konvergeert (zie opgave (5.3.28)).

De volgende verschuivings strategie van Wilkinson levert voor iedere hermietse matrix A een rij $((A_k e_n, e_n))$ die minstens kwadratisch konvergeert naar 'n eigenwaarde van A; de konvergentie is gewoonlijk zelfs sneller dan die met de Rayleigh quotiënt verschuiving. We geven hier geen bewijs.

Wilkinsons verschuiving κ_k: kies voor iedere k voor de verschuiving κ_k de eigenwaarde van het 2×2 beneden blok van A_k die het dichtst bij $(A_k e_n, e_n)$ ligt.

(5.4.17) Opmerking. Beschouw een $A \in \mathbf{M}_n(\mathbf{R})$ met eigenwaarden $\lambda_1, \ldots, \lambda_n$ zodat $\ldots \geq |\lambda_{n-2}| > |\lambda_{n-1}| = |\lambda_n|$ en $\lambda_n \notin \mathbf{R}$. Dus $\lambda_{n-1} = \bar{\lambda}_n$.

Om efficiëntie redenen is het aantrekkelijk om zolang mogelijk reëel te rekenen.

Rekenen we reëel $(W_0, Q_k, \tilde{R}_k \in \mathbf{M}_n(\mathbf{R}))$ met $\kappa_k = 0$ dan konvergeert (A_k) ook in zekere zin en wel naar een matrix $S' = \begin{bmatrix} S'' & E \\ 0 & L \end{bmatrix}$ met $L \in \mathbf{M}_2(\mathbf{R})$ (zie (5.4.11)). λ_n, λ_{n-1} zijn dan precies de eigenwaarden van het 2×2–blok L die pas in de laatste fase door komplex te rekenen bepaald hoeven te worden.

Door bijvoorbeeld $\kappa_k \approx \mathrm{Re}\,\lambda_n$ te kiezen (denk aan het reële deel van de eigenwaarden van het 2×2 beneden blok van A_k) kunnen we de konvergentie naar S' enigszins versnellen, maar spektakulair wordt de versnelling nooit als we voor κ_k reële getallen kiezen (waarom niet?). Door voor κ_k (voor k groot genoeg) 'n eigenwaarde μ_k van het 2×2 onderblok van A_k te nemen krijgen we een kwadratisch konvergent proces dat helaas komplex rekenen vereist. Door in stap $2k$ te schuiven over μ_{2k} en vervolgens in stap $2k+1$ over $\bar{\mu}_{2k}$ zouden we, als we exakt konden rekenen, na stap $2k+1$ weer in de reële situatie terug zijn en zouden we ook een kwadratisch konvergent proces hebben. We kunnen echter niet exakt rekenen en bovendien zouden we nog steeds komplex moeten rekenen.

Francis heeft laten zien hoe de twee zojuist geschetste stappen geassembleerd kunnen worden tot één stap die dan alleen reëel rekent.

De *Francis' verschuiving* levert dus met reële arithmetiek voor de situatie hier een rij (A_k) waarvan de 2×2 beneden blokken kwadratisch konvergeren naar een 2×2 matrix met eigenwaarden λ_n en λ_{n-1}.

(5.4.18) Reduktie naar kleinere matrices. Als de geproduceerde Hessenberg matrix bij $k = 0$ 'numeriek reducibel' is (zie (5.1.6)) dan stapt men over op kleinere problemen.

Het QR-algorithme is een oneindig iteratief proces. In de praktijk zullen we de berekening moeten stoppen.

Met behulp van de behandelde schattingsstellingen kunnen we tijdens het rekenproces nagaan hoe goed $(A_k e_n, e_n)$ de eigenwaarde λ_n benadert (zie opgave (5.4.19) voor het geval A hermiets is). Als voor een zekere k de benadering ons bevalt schrappen we in A_k de n-de rij en de n-de kolom en laten we het QR-algorithme verder werken met de gereduceerde $(n-1) \times (n-1)$ matrix (of met een nog kleinere matrix; zie (5.4.11) en (5.4.17)). (In de praktijk kontroleert men gewoonlijk ook weer of $|(A_k e_{n-1}, e_n)| \leq \xi(|(A_k e_{n-1}, e_{n-1})| + |(A_k e_n, e_n)|)$; zie (5.1.6).)

Onze spektrale verschuivingen richten we nu af op λ_{n-1}.

Terwijl we ons richtte op een snelle benadering van λ_n benaderde het $n-1$-ste diagonaal element van A_k ook al λ_{n-1}. In het QR-algorithme kunnen we na de reductie (gewoonlijk) onmiddellijk uit de gereduceerde matrix onze favoriete spektrale verschuiving berekenen en toepassen (zie (5.4.16)). Etc..

(5.4.19) Opgave. Stel dat A hermiets is. Zij $\varrho := \min_{j<n} |\lambda_j - \lambda_n|$.

Beschouw $A_k = \begin{bmatrix} \ddots & \ddots & \vdots & \vdots \\ \ddots & \gamma_{n-2} & \bar{\beta}_{n-1} & 0 \\ \dots & \beta_{n-1} & \gamma_{n-1} & \bar{\beta}_n \\ \dots & 0 & \beta_n & \gamma_n \end{bmatrix}$. Schrijf $e_n = \alpha \vec{v}_n + \vec{r}$ met $\vec{r} \perp \vec{v}_n$.

(a) Bewijs dat $\min_j |\gamma_n - \lambda_j| \leq |\beta_n|$.
(b) Stel dat $2|\beta_n| \leq \varrho$. Bewijs dat $\sin \angle(e_n, \mathrm{kern}\,(A - \lambda_n I)) = \|\vec{r}\|_2 \leq \frac{|\beta_n|}{\varrho - |\beta_n|}$
(Hint: gebruik de stelling van Davis–Kahan).
Toon aan dat $|\gamma_n - \lambda_n| \leq \frac{|\beta_n|^2}{\varrho - |\beta_n|}$.
(Hint: beschouw $R_0 := \beta_n(e_n e_{n-1}^* + e_{n-1} e_n^*)$ en $\tilde{A} := A_k - R_0$. Dan $\tilde{A} e_n = \gamma_n e_n$ en $(R_0 e_n, e_n) = 0$. Gebruik nu (4.6.11)).
(c) $\varrho \leq \min_{i<n} |\gamma_i - \gamma_n| - |\beta_i| - |\beta_{i-1}| - |\beta_n|$. Bewijs dit.

De efficiëntie

Stap (a2) en (b2) in het QR-algorithme voert men uit met Householder transformaties.

(5.4.20) Bewering. (a) *A_k is een Hessenberg matrix voor iedere k.*
A_{k+1} kan met Householder transformaties in $\approx 4n^2$ flop uit A_k berekend worden.
(b) *Als $A^* = A$ dan $A_k^* = A_k$ en A_k is tri-diagonaal voor iedere k.*
A_{k+1} kan dan met Householder transformaties in $\approx 16n$ flop uit A_k berekend worden.

Bewijs. We bekijken de rekendetails van A_k naar A_{k+1} in enig detail voor het geval $n = 5$.
(i) Stel dat $A_k - \kappa_k I$ van de volgende vorm is

$$A_k - \kappa_k I = \begin{bmatrix} a & a & a & a & a \\ a & a & a & a & a \\ 0 & a & a & a & a \\ 0 & 0 & a & a & a \\ 0 & 0 & 0 & a & a \end{bmatrix} \quad \text{zodat} \quad H_{\vec{w}_1} = \begin{bmatrix} h & h & 0 & 0 & 0 \\ h & h & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

waarin iedere a 'n matrix koëfficiënt van $A_k - \kappa_k I$ voorstelt en h 'n van $H_{\vec{w}_1}$ (de a's en de h's zijn op verschillende plaatsen mogelijk verschillend). Neem aan dat we na twee voorvermenigvuldigingen een matrix $A_k^{(2)} := H_{\vec{w}_2} H_{\vec{w}_1}(A_k - \kappa_k I)$ van de volgende vorm vinden

$$A_k^{(2)} = \left[\begin{array}{cc|cc|c} r & r & r & r & r \\ 0 & r & r & r & r \\ \hline 0 & 0 & r' & r' & r' \\ 0 & 0 & a & a & a \\ \hline 0 & 0 & 0 & a & a \end{array}\right]. \text{ Dan } H_{\vec{w}_3} = \left[\begin{array}{cc|cc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & h & h & 0 \\ 0 & 0 & h & h & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right] \text{ en } A_k^{(3)} = \left[\begin{array}{cc|cc|c} r & r & r & r & r \\ 0 & r & r & r & r \\ \hline 0 & 0 & r & r & r \\ 0 & 0 & 0 & r' & r' \\ \hline 0 & 0 & 0 & a & a \end{array}\right]$$

waarin a nog de 'oude' en r en r' al bewerkte koëfficiënten zijn. De r zullen koëfficiënten van $\tilde{R}_{k+1}$ blijken te zijn. (ga na dat de r in $A^{(2)}$ door voorvermenigvuldiging met $H_{\vec{w}_3}$ niet veranderen).
(ii) Dus $\tilde{R}_{k+1}$ is van de volgende vorm

$$\tilde{R}_{k+1} = \begin{bmatrix} r & r & r & r & r \\ 0 & r & r & r & r \\ 0 & 0 & r & r & r \\ 0 & 0 & 0 & r & r \\ 0 & 0 & 0 & 0 & r \end{bmatrix} \quad \text{zodat} \quad \tilde{R}_{k+1}^{(1)} := \tilde{R}_{k+1} H_{\vec{w}_1} = \begin{bmatrix} a & a' & r & r & r \\ a & a' & r & r & r \\ 0 & 0 & r & r & r \\ 0 & 0 & 0 & r & r \\ 0 & 0 & 0 & 0 & r \end{bmatrix}$$

waarin iedere a al matrix koëfficiënten van $A_{k+1} - \kappa_k I$ voorstellen. Neem aan dat we na twee navermenigvuldigingen een matrix $\tilde{R}_{k+1}^{(2)} := \tilde{R}_{k+1} H_{\vec{w}_1} H_{\vec{w}_2}$ van de volgende vorm vinden

$$\tilde{R}_{k+1}^{(2)} = \left[\begin{array}{cc|cc|c} a & a & a' & r & r \\ a & a & a' & r & r \\ \hline 0 & a & a' & r & r \\ 0 & 0 & 0 & r & r \\ \hline 0 & 0 & 0 & 0 & r \end{array}\right]$$

waarin a, a' 'nieuwe' en r oude koëfficiënten zijn. De a zullen koëfficiënten van $A_{k+1} - \kappa_k I$

blijken te zijn. Dan is

$$\tilde{R}^{(3)}_{k+1} = \left[\begin{array}{cc|cc|c} a & a & a' & r & r \\ a & a & a' & r & r \\ \hline 0 & a & a' & r & r \\ 0 & 0 & 0 & r & r \\ \hline 0 & 0 & 0 & 0 & r \end{array}\right] \left[\begin{array}{cc|cc|c} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & h & h & 0 \\ 0 & 0 & h & h & 0 \\ \hline 0 & 0 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cc|cc|c} a & a & a & a' & r \\ a & a & a & a' & r \\ \hline 0 & a & a & a' & r \\ 0 & 0 & a & a' & r \\ \hline 0 & 0 & 0 & 0 & r \end{array}\right]$$

(ga na dat de a in $\tilde{R}^{(2)}_{k+1}$ door navermenigvuldiging met $H_{\vec{w}_3}$ niet veranderen).
Hieruit blijkt dat A_{k+1} weer een Hessenberg matrix is.

Omdat $A_k = W_k^* A W_k$ is A_k hermiets als A dat is. Hermietse Hessenberg matrices zijn vanzelf tri-diagonaal.

We laten het tellen van de flops aan de lezer over. Wel wijzen we erop dat in geval A hermiets is de koëfficiënten op de $(i, i+2)$ plaatsen nooit berekend hoeven te worden: als er behoefte is aan een koëfficiënt op zo'n plaats kan men die behoefte bevredigen door een symmetrie argument toe te passen (ga da na voor bijvoorbeeld het $(1,3)$ element). □

(5.4.21) Efficiënt geheugengebruik. Beschouw voorgaande bewering en bewijs.

$A^{(2)}_k$, $\tilde{R}_{k+1}$ hebben dezelfde eerste kolom en tweede kolom. Omdat $H_{\vec{w}_1}$ bij navermenigvuldiging alleen de eerste en tweede kolom kombineert volgt dat $A^{(2)}_k H_{\vec{w}_1}$ en $\tilde{R}^{(1)}_{k+1} := \tilde{R}_{k+1} H_{\vec{w}_1}$ dezelfde eerste kolom (en tweede kolom) hebben. Omdat $\tilde{R}^{(1)}_{k+1}$ en $\tilde{R}^{(n-1)}_{k+1} = A_{k+1} - \kappa_k I$ dezelfde eerste kolom hebben, hebben $A^{(2)}_k H_{\vec{w}_1}$ en $A_{k+1} - \kappa_k I$ dezelfde eerste kolom.

Hebben we voorvermenigvuldigd met $H_{\vec{w}_2}$ dan kunnen we direkt daarna navermenigvuldigen met $H_{\vec{w}_1}$; na voorvermenigvuldigen met $H_{\vec{w}_{j+1}}$ volgt navermenigvuldigen met $H_{\vec{w}_j}$.

Merk op dat $W_{k+1} = W_k Q_{k+1} = W_k H_{\vec{w}_1} \cdots H_{\vec{w}_{n-1}}$. Zijn we geïnteresseerd in de W_i dan kunnen we na de navermenigvuldiging met $H_{\vec{w}_1}$ gelijk $W_k H_{\vec{w}_1}$ bereken en kunnen we $H_{\vec{w}_1}$ verder vergeten, etc..

Naast één $n \times n$ array voor de A_k's (of een Hessenberg matrix array) en één $n \times n$ array voor de W_k's (als we ook geïnteresseerd zijn in de U) hebben we maar een paar extra geheugen plaatsen nodig (≤ 8 voor de twee Householder transformaties, 1 voor κ_k).

In de volgende opgave suggereren we een andere manier waarmee we efficiënt in het QR-algorithme met geheugenruimte kunnen omspringen. De opgave geeft een manier waarop een QR-stap met verschuiving uitgevoerd kan worden zonder expliciet de verschuiving van de diagonaal elementen af te trekken.

(5.4.22) • Opgave. Zij A een Hessenberg matrix en zij $\kappa \in \mathbf{C}$.
(a) Zij $Q \in \mathbf{Q}_n$ en $R \in \mathbf{U}_n$ zo dat $A - \kappa I = QR$. Laat $\hat{A} := RQ + \kappa I$. Ga na dat $\hat{A} = Q^* A Q$.
(b) Zij $H_{\vec{w}}$ de Householder transformatie zodat $H_{\vec{w}}(Ae_1 - \kappa e_1) \in \operatorname{span}(e_1)$.
Bereken $A' := H_{\vec{w}} A H_{\vec{w}}$. Er is een $\tilde{Q} \in \mathbf{M}_n$, produkt van Householder transformaties zodat $\tilde{Q} e_1 = e_1$ zodat $\tilde{Q} A' \tilde{Q}^*$ een Hessenberg matrix is. Ga dit na. Bewijs dat $\tilde{Q} A' \tilde{Q}^* = J \hat{A} J^*$ voor zekere J met $|J| = I$ als $\hat{A}$ irreducibel is. (Hint: zie (5.1.12).)
(c) Als A' berekend is kan $\hat{A}$ (of $J\hat{A}J^*$) als volgt met geschikte Householder transformaties bepaald worden: $A^{(1)} = A'$, $A^{(i+1)} = H_{\vec{w}_i} A^{(i)} H_{\vec{w}_i}$ voor $i = 1, \ldots, n-1$ en $\hat{A} = A^{(n)}$, waarbij $\vec{w}_i \in \operatorname{span}(e_i, e_{i+1})$. Zie dit in.

(5.4.23) De hoeveelheid werk. Over de totale hoeveelheid werk zijn geen harde uitspraken te doen. De ervaring leert dat het QR-algorithme met de verschuivingen uit (30) nadat de Hessenberg vorm bepaald is (stap (a2)) in ≈ 5 slagen λ_n benadert en voor iedere volgende eigenwaarde ≈ 2 slagen nodig heeft. Om de Hessenberg vorm en daarna alle eigenwaarden te berekenen vergt dit dus ongeveer $\frac{4}{3}n^3 + 5 \cdot 4n^2 + 2(\cdot 4(n-1)^2 + \ldots + 4.2^2) \approx 4n^3$ flop.
Het QR-algorithme is zo efficiënt dat men wortels van polynomia bepaalt door het QR-algorithme toe te passen op de companion matrix.
(Omdat het QR-algorithme met verschuiving (30) de in absolute waarde kleinste eigenwaarde benadert met een kwadratisch konvergente rij hangt de hoeveelheid werk niet sterk af van de

verlangde nauwkeurigheid: hebben we in $\approx 4n^3$ flop alle eigenwaarden in 10 cijfers nauwkeurig, dan kunnen we door een extra iteratie slag à $\approx n^3$ flop 20 goede cijfers tegemoet zien.)

Voor een hermietse matrix A produceert het QR-algorithme een rij die kubisch konvergeert naar de in absolute waarde kleinste eigenwaarde. Ondertussen worden de overige eigenwaarden ook goed benaderd (zie (5.4.9)). In deze hermietse situatie konvergeert het QR-algorithme met verschuiving (30) en reduktie strategie (zie (5.4.18)) onvoorstelbaar snel. Het werk zit alleen in het bepalen van de tri-diagonaal vorm.

(5.4.24) De stabiliteit. Omdat we met Householder transformaties werken is de stabiliteit uitstekend. Uit (2.5.11) blijkt dat de berekende matrix A_k^* gelijk is aan $\tilde{Q}^*(A+E)\tilde{Q}$ voor een of ander unitaire $\tilde{Q}$ en een kleine E. De fout in de eigenwaarden van A_k^* is alleen groot als het eigenwaarden probleem voor A slecht gekonditioneerd is.

We besluiten deze paragraaf met een opgave waarin een snel algorithme om de singuliere waarde dekompositie te bepalen wordt gesuggereerd.

(5.4.25) • Opgave. (a) Als $R = (r_{ij}) \in \mathbf{M}_n$ een bi-diagonale boven driehoeks matrix is ($r_{ij} = 0$ als $j - i \notin \{0, 1\}$) dan is er een unitaire matrix V zodat $B := RV^*$ een bi-diagonale beneden driehoeks matrix is. Ga dit na en laat zien dat V op te bouwen is uit Householder transformaties.
(b) Zij nu R_0 een bi-diagonale boven driehoeks matrix met singuliere waarde dekompositie $R_0 = U\Sigma V$ waarbij U, V unitair en $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ zodat $\sigma_1 > \sigma_2 > \ldots > \sigma_{n-1} > \sigma_n$. Stel (Q_k) en (V_k) zijn rijen unitaire matrices en (B_k) is een rij bi-diagonale beneden driehoeks matrices zodat $R_k = B_k V_k$ en $R_{k+1} = Q_k B_k$. Bewijs dat dan $\lim_{k\to\infty} R_k = \Sigma$.
(Hint: $A_k := R_k R_k^* = B_k B_k^*$. Dan $A_k = Q_k^* R_{k+1} B_k^*$ en $R_{k+1} B_k^*$ is boven driehoeks. Verder is $A_{k+1} = R_{k+1} B_k^* Q_k^*$.)
(c) Ontwerp een algorithme om de singuliere waarde dekompositie te bepalen (gebruik (b) en opgave (5.1.13)). Geef aan hoe de spektrale verschuivingen gekozen moet worden.

5.5 De methode van Jacobi

In deze paragraaf behandelen we een methode die hermietse matrices diagonalizeert. De methode is aanzienlijk minder efficiënt (vergt $\approx 12n^3$ flop) dan het QR-algorithme met verschuiving. Maar werkt altijd zonder extra maatregelen (ook als eigenwaarden gelijke absolute waarde hebben) en is konceptueel eenvoudiger dan het QR-algorithme. Voor niet al te grote matrices is de methode daarom toch ook voor de praktijk interessant. Verder is de methode voor de praktijk interessant als de buiten diagonaal elementen van de te diagonalizeren hermietse matrix klein zijn. Bovendien is de methode 'vektorizeerbaar' en staat daarom tegenswoordig voor vektor-computers weer in de belangstelling.

Bij de behandeling van de methode komen we een numerieke transformatie techniek tegen die evenals de Gauss transformaties en de Householder transformaties veel gebruikt wordt en de moeite van het bestuderen waard is.

Givens transformaties

(5.5.1) Vlak rotaties of Givens transformaties. Voor een $\varphi \in [0, 2\pi)$ beschouw met $c = \cos\varphi$ en $s = \sin\varphi$ de unitaire matrix

$$J_{1\,2}(\varphi) := \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix} \in \mathbf{M}_n(\mathbf{R}), \text{ waarbij } \quad \tilde{Q} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \in \mathbf{M}_2.$$

De afbeelding waar $J_{1\,2}(\varphi)$ de matrix van is roteert in het (e_1, e_2)–vlak over een hoek φ.
Op analoge wijze is voor ieder (e_p, e_q)–vlak met $p, q \in \{1, \ldots, n\}$, $p \neq q$ de (e_p, e_q)*–vlak rotatie* of *Givens transformatie* $J_{pq}(\varphi)$ gedefiniëerd.

Givens transformaties en Householder transformaties worden voor hetzelfde type toepassing gebruikt. Afhankelijk van de precise toepassing heeft de ene transformatie voordelen boven de andere.

Als een householder transformatie $H_{\vec{w}}$ zich in een (e_p, e_q)–vlak afspeelt ($\vec{w} = \alpha e_p + \beta e_q$) dan is zuiver mathematisch die transformatie de samenstelling van een vlak rotatie $J_{pq}(\varphi)$ en een elementaire spiegeling ($e_i \rightsquigarrow e_i$ voor $i \neq q$ en $e_q \rightsquigarrow -e_q$).

In deze paragraaf zijn we vooral geïnteresseerd in similariteits transformaties $J_{pq}(\varphi) A J_{pq}(\varphi)^*$ van matrices A.

(5.5.2) Opmerking. Voor een matrix $A = (a_{ij}) \in \mathbf{M}_n$ geven we hieronder schematisch het effekt weer van de rotatie $J_{pq}(\varphi)$ in de berekening $J_{pq}(\varphi) A J_{pq}(\varphi)^* = (\hat{a}_{ij})$. Alle koëfficiënten van de matrix A in de p-de en de q-de rij en kolom worden door de rotatie bewerkt (in de schematische matrix hieronder, de elementen op de lijnen):
$\hat{a}_{pl} = ca_{pl} - sa_{ql}$, $\hat{a}_{ql} = ca_{ql} + sa_{pl}$ voor $l \notin \{p, q\}$, etc..

$$\begin{array}{c} \\ p \\ \\ q \\ \\ \end{array}\begin{bmatrix} & | & & | & \\ — & c & — & -s & — \\ & | & & | & \\ — & s & — & c & — \\ & | & & | & \end{bmatrix}\begin{bmatrix} & | & & | & \\ — & a_{pp} & — & a_{pq} & — \\ & | & & | & \\ — & a_{qp} & — & a_{qq} & — \\ & | & & | & \end{bmatrix}\begin{bmatrix} & | & & | & \\ — & c & — & s & — \\ & | & & | & \\ — & -s & — & c & — \\ & | & & | & \end{bmatrix} \tag{31}$$

(5.5.3) De berekening van de vlak rotatie.
De hoek φ die we in de definitie van de vlak rotaties gebruikt hebben moet natuurlijk niet expliciet bepaald worden. We werken met c en s. De manier waarop we c en s berekenen hangt natuurlijk af van wat wij van $J_{pq}(\varphi)$ verlangen. Het is echter verstandig eerst $t := \tan\varphi$ te berekenen en dan c en s als volgt te bepalen:

$$c = (1 + t^2)^{-\frac{1}{2}} \quad \text{en} \quad s = tc.$$

Als $t^\star$ de berekende waarde voor t is, dan is $t^\star = \tan\tilde{\varphi}$ voor zekere hoek $\tilde{\varphi}$. Zou de berekening van c en s uit $t^\star$ verder foutloos verlopen dan zouden we te maken hebben met $J_{pq}(\tilde{\varphi})$ in plaats van $J_{pq}(\varphi)$. Hoewel $J_{pq}(\tilde{\varphi})$ niet de vlak rotatie is die we wensten, is het wel een vlak rotatie die dus in een similariteits transformatie de eigenwaarden bewaart, zelfs als $t^\star$ met een grote fout belast is!

In de volgende alinea geven we zonder bewijs een schatting voor de evaluatiefouten in de vlak rotaties. (We verwaarlozen weer "$\mathcal{O}(\xi^2)$–termen". Als in (2.5.11) en (b) van (2.5.13) zijn de uitspraken die meer vlak rotaties betreffen alleen korrekt als het aantal vlak rotaties niet gigantisch groot is.) Door de computer berekende grootheden zijn weer gemarkeerd met een $\star$.

(5.5.4) De stabiliteit van de vlak rotaties. *Zij* $p, q \in \{1, \ldots, n\}$, $p \neq q$, $\varphi \in [0, 2\pi)$. *Laat* $Q \in \mathbf{M}_n$ *een produkt zijn van* l *vlak rotaties en zij* $A \in \mathbf{M}_n$.
Voor $t^\star \in \mathbf{R}$ *bereken* c *en* s *als in* (5.5.3). *Dan geldt het volgende.*
(a) $\|J_{pq}(\varphi)^\star - J_{pq}(\tilde{\varphi})\|_2 \leq 6\bar{\xi}$ *voor zekere* $\tilde{\varphi} \in [0, 2\pi)$.
(b) $\|(J_{pq}(\varphi)A)^\star - J_{pq}(\tilde{\varphi})A\|_2 \leq 9\bar{\xi}\|A\|_F$ *voor zekere* $\tilde{\varphi} \in [0, 2\pi)$.
(c) $\|(QA)^\star - \tilde{Q}A\|_F \leq 9l\bar{\xi}\,\|A\|_F$ *voor zekere unitaire* $\tilde{Q}$.
(d) *Is* Q *opgebouwd uit rotaties die in verschillende vlakken roteren* (in het bijzonder is dus $l \leq \frac{1}{2}n(n-1)$) *dan geldt zelfs* $\|(QA)^\star - \tilde{Q}A\|_F \leq 9n^{\frac{3}{2}}\bar{\xi}\,\|A\|_F$ *voor zekere unitaire* $\tilde{Q}$.

(5.5.5) Opmerking. Met behulp van de vlak rotaties kan men een matrix $B = (b_{ij}) \in \mathbf{M}_n(\mathbf{R})$ ook op Hessenberg vorm brengen.

Bepaal $J_{23}(\varphi_3)$ zodat de $(3, 1)$ koëfficiënt van $J_{23}(\varphi_3)B$ nul is $(cb_{31} + sb_{21} = 0)$. Dan is de eerste kolom van $J_{23}(\varphi_3)B$ gelijk aan de eerste van $J_{23}(\varphi_3)BJ_{23}(\varphi_3)^*$. Voor geschikte $J_{23}(\varphi_3)$, $J_{24}(\varphi_4)$, …,$J_{2n}(\varphi_n)$ zijn zo met $Q_1 := J_{2n}(\varphi_n) \cdots J_{23}(\varphi_3)$ de koëfficiënten van $Q_1 B Q_1^*$ op de $(3, 1)$, $(4, 1)$, …, $(n, 1)$ plaats allen nul te maken. Vervolgens kunnen we de tweede kolom bewerken, etc..

Deze benadering is enigszinds stabieler dan die met Householder transformaties (vergelijk (d) (5.5.4) met (5.1.1)), maar is, in geval B hermiets is, eens zo duur.

De Jacobi methode

(5.5.6) In deze paragraaf is $A \in \mathbf{M}_n(\mathbf{R})$ hermiets.

(5.5.7) Notaties. Zij $H \in \mathbf{M}_n$ hermiets. In deze paragraaf zijn de eigenwaarden $\lambda_1(H), \ldots, \lambda_n(H)$ zo genummerd dat $\lambda_1(H) \le \lambda_2(H) \le \ldots \le \lambda_{n-1}(H) \le \lambda_n(H)$. Verder schrijven we $\Lambda_H := \mathrm{diag}\,(\lambda_1(H), \ldots, \lambda_n(H))$,

$$\alpha(H) := \|H - \mathrm{diag}\,(H)\|_F \quad \text{en} \quad \beta(H) := \|H - \mathrm{diag}\,(H)\|_m. \tag{32}$$

Dan is volgens de stelling van Wielandt–Hoffman

$$\sqrt{\sum_{i=1}^{n} |\lambda_i(H) - \lambda_i(\mathrm{diag}\,(H))|^2} \le \alpha(H). \tag{33}$$

$\alpha(H)$ geeft aan hoe goed de diagonaal elementen van de matrix de eigenwaarden van de matrix H benaderen. Merk op dat

$$2\beta(H)^2 \le \alpha(H)^2 \le (n-1)n\beta(H)^2. \tag{34}$$

Tenslotte schrijven $\mathbf{n} := \{1, 2, \ldots, n\}$ en is $\tilde{n} := \frac{1}{2}n(n-1)$ het aantal strikt beneden driehoeks koëfficiënten.

Het volgende elementaire lemma gebruiken we in deze paragraaf een paar maal.

(5.5.8) Lemma. *Als $C \in \mathbf{M}_{k \times l}$ dan is $\|PCQ\|_F = \|C\|_F$ voor iedere unitaire matrix $P \in \mathbf{M}_k$ en iedere uniaire matrix $Q \in \mathbf{M}_l$.*

Bewijs. $\|C\|_F^2 = \mathrm{spoor}\,(C^*C) = \mathrm{spoor}\,(C^*P^*PCQQ^*) = \mathrm{spoor}\,(Q^*C^*P^*PCQ) = \|PCQ\|_F^2$. □

De methode van Jacobi uit 1846 produceert iteratief een rij (J_k) van vlak rotaties en een rij (A_k) van hermietse matrices zodat

$$A_0 = A, \quad A_{k+1} = J_k A_k J_k^* \quad (k \in \mathbf{N}) \quad \text{en} \quad \lim_{k\to\infty} \alpha(A_k) = 0.$$

Tijdens de berekening vertelt $\alpha(A_k)$ ons hoe nauwkeurig de diagonaal elementen van A_k de eigenwaarden van A_k en dus die van A benaderen.

(5.5.9) Lemma. *Zij $A = (a_{ij})$ hermiets. Zij $p, q \in \mathbf{n}$, $p < q$. Er is een vlak rotatie $J_{pq}(\varphi)$ zodat we met $J_{pq}(\varphi)AJ_{pq}(\varphi)^* = (\hat{a}_{ij})$ hebben dat $\hat{a}_{pq} = \hat{a}_{qp} = 0$. Voor deze vlak rotatie geldt $\alpha(J_{pq}(\varphi)AJ_{pq}(\varphi)^*)^2 = \alpha(A)^2 - 2|a_{pq}|^2$.*

Bewijs. We beschouwen voor de notationele overzichtelijkheid het geval dat $p = 1$ en $q = 2$.

Laat A als volgt gepartitioneerd zijn $A := \begin{bmatrix} U & F^* \\ F & V \end{bmatrix}$ met $U \in \mathbf{M}_2$.

Beschouw $Q = \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix} \in \mathbf{M}_n$ met $\tilde{Q} \in \mathbf{M}_2(\mathbf{R})$ unitair. Dan $QAQ^* = \begin{bmatrix} \tilde{Q}U\tilde{Q}^* & \tilde{Q}F^* \\ F\tilde{Q}^* & V \end{bmatrix}$.

Als $\tilde{Q} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}$ met $c, s \in \mathbf{R}$ zodat $c^2 + s^2 = 1$, $U = \begin{bmatrix} u & f \\ f & v \end{bmatrix}$ en $\tilde{U} := \tilde{Q}U\tilde{Q}^*$ dan

$$\tilde{U} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} u & f \\ f & v \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} = \begin{bmatrix} c^2u + s^2v - 2csf & cs(u-v) + f(c^2 - s^2) \\ cs(u-v) + f(c^2 - s^2) & s^2u + c^2v + 2csf \end{bmatrix}.$$

Kies nu c, s zo dat $cs(u-v) + f(c^2 - s^2) = 0$.
Met $\varphi \in [0, 2\pi)$ zo dat $e^{i\varphi} = c + is$ is $Q = J_{1\,2}(\varphi)$.

Merk tenslotte op dat met (5.5.8) volgt dat

$$\alpha(A)^2 = \alpha(U)^2 + \alpha(V)^2 + 2\|F\|_F^2 \quad \text{en} \quad \alpha(QAQ^*)^2 = \alpha(\tilde{Q}U\tilde{Q}^*)^2 + \alpha(V)^2 + 2\|F\|_F^2.$$

Blijkbaar $\alpha(A)^2 - \alpha(QAQ^*)^2 = \alpha(U)^2 - \alpha(\tilde{Q}U\tilde{Q}^*)^2 = 2|a_{12}|^2$. □

(5.5.10) De berekening van de tangens.
Bereken uit $t = \tan\varphi$ (met φ zo dat $e^{i\varphi} = c + is$) de c en s als aangeven in (5.5.3).
Berekenen t als volgt.

Neem aan dat $a_{pq} \neq 0$ (anders zet $c = 1$ en $s = 0$).
Bereken $\tau = (a_{qq} - a_{pp})/(2a_{pq})$ (merk op dat dan $\tau = \cot 2\varphi$).
Dan voldoet t aan $t^2 + 2\tau t - 1 = 0$. Met $t' := |\tau| + \sqrt{1+\tau^2}$ bereken $t = \text{teken}(\tau)/t'$.

Bekijken we alle vlak rotaties $J = J_{pq}(\varphi)$ die als in (5.5.9) een buiten diagonaal element 0 roteren dan zal de diagonaal van de resulterende matrix JAJ^* het best de eigenwaarden van A weergeven als $\alpha(JAJ^*)$ minimaal is; als J het in absolute waarde grootste buiten diagonaal element 0 roteert. Het Jacobi proces past deze rotatie methode iteratief toe.

(5.5.11) De methode van Jacobi. De methode produceert als volgt een rij (A_k) van hermietse matrices $A_k = (a_{ij}^{(k)})$ en een rij (J_k) van vlak rotaties :

Algorithme

(a) $k := 0$; $A_0 := A$;
(b) `bepaal` $(p,q) \in \mathbf{n}\times\mathbf{n}$, $p < q$ `zodat` $\beta(A_k) = |a_{pq}^{(k)}|$;
(c) `bereken` $c_k, s_k \in \mathbf{R}$ `als in` (5.5.12) `met` $\tau_k = (a_{qq}^{(k)} - a_{pp}^{(k)})/(2a_{pq}^{(k)})$;
(d) `Zij` $J_k := J_{pq}(\varphi_k)$ `met` $e^{i\varphi_k} = c_k + is_k$. `Bereken` $A_{k+1} = J_k A_k J_k^*$;
(e) $k := k+1$; `ga na naar` (b);

(5.5.12) Opmerking. Als een matrix koëfficiënt $a_{ij}^{(k)} = 0$ dan kan voor een $l > k$ weer gelden dat $a_{ij}^{(l)} \neq 0$ (zie opmerking (5.5.2))!

(5.5.13) Stelling. $\alpha(A_k) \leq \sqrt{1 - \frac{2}{n(n-1)}}\,\alpha(A_{k-1})$ *voor iedere* $k \in \mathbf{N}$.
$\lim_{k\to\infty} \alpha(A_k) = 0$ *en* $\lim_{k\to\infty} \lambda_j(\text{diag}\,(A_k)) = \lambda_j(A)$ *voor iedere* $j \in \mathbf{n}$.

Bewijs. Uit (34) volgt
$$\alpha(A_{k+1})^2 = \alpha(A_k)^2 - 2\beta(A_k)^2 \leq \alpha(A_k)^2(1 - \tfrac{2}{n(n-1)}) \leq \alpha(A)^2(1 - \tfrac{2}{n(n-1)})^{k+1}$$
en de stelling volgt uit de Wielandt–Hoffman stelling en de ordening van de diagonaal elementen van de A_k die door de rotaties opgelegd is. □

De konvergentie die door de stelling gegarandeerd wordt is zeer traag: na $\tilde{n} = \frac{1}{2}n(n-1)$ iteratie slagen garandeert de stelling dat de $\alpha(A_k)$ met een faktor $\approx e^{-\frac{1}{2}}$ gereduceerd is ($\alpha(A_{k+\tilde{n}}) \leq (1 - \frac{1}{\tilde{n}})^{\frac{1}{2}\tilde{n}}\alpha(A_k) \leq e^{-\frac{1}{2}}\alpha(A_k)$). Met $\tilde{\alpha}_k := \alpha(A_{k\tilde{n}})$ konvergeert de rij $(\tilde{\alpha}_k)$ gemajoreerd lineair naar 0 met konvergentie faktor $e^{-\frac{1}{2}}$. De volgende stelling vertelt echter dat rij gemajoreerd kwadratisch konvergeert.

(5.5.14) Stelling. *Er is een* $\Gamma > 0$ *zodat* $\tilde{\alpha}_{k+1} \leq \Gamma\tilde{\alpha}_k^2$ $(k \in \mathbf{N})$.

• *Bewijs.* Stel dat $\varrho := \min_{i>1}(\lambda_i(A) - \lambda_{i-1}(A)) > 0$.
We mogen wel aannemen dat $a_{j-1\,j-1}^{(k)} \leq a_{jj}^{(k)}$ voor $j = 2, \ldots, n$ en $k > 0$.

Er is een k_0 zo dat $2\alpha(A_k) < \varrho$ voor $k \geq k_0$. Dus $a_{j-1\,j-1}^{(k)} < a_{jj}^{(k)}$ voor $j > 1$ en $k \geq k_0$.
(a) Als $J_k = J_{qp}(\varphi)$ met $p < q$ en $l \notin \{q, p\}$ dan
$$|a_{ql}^{(k+1)}| \leq |\cos\varphi|\,|a_{ql}^{(k)}| + |\sin\varphi|\,|a_{pl}^{(k)}| \leq |a_{ql}^{(k)}| + |\sin\varphi|\beta(A_k).$$
Verder is $|\sin\varphi| \leq |\tan\varphi| \leq \frac{1}{2}|\tan 2\varphi| = |a_{qp}^{(k)}|/|a_{qq}^{(k)} - a_{pp}^{(k)}| \leq \beta(A_k)/(\varrho - 2\alpha(A_k))$ en dus
$$|a_{ql}^{(k+1)}| \leq |a_{ql}^{(k)}| + \frac{\beta(A_k)^2}{\varrho - 2\alpha(A_k)} \leq |a_{ql}^{(k)}| + \frac{\alpha(A_k)^2}{2(\varrho - 2\alpha(A_k))}. \tag{35}$$
Een zelfde schatting hebben we voor $|a_{lq}^{(k+1)}|$, $|a_{pl}^{(k+1)}|$ en $|a_{lp}^{(k+1)}|$.
(b) Laat $k_1, k_2 \in \mathbf{N}$, $k_0 \leq k_1 < k_2$ zo zijn dat J_{k_1} en J_{k_2} beiden in hetzelfde vlak roteren en

J_k in een ander vlak roteert dan J_l als $k_1 \leq k < l < k_2$. Stel dat J_{k_1} en J_{k_2} in het (e_q, e_p)–vlak roteren. Dan $a_{qp}^{(k_1+1)} = 0$ en $\beta(A_{k_2}) = |a_{qp}^{(k_2)}|$. Omdat van de k_1-stap tot de k_2-de in geen enkel vlak twee maal geroteerd wordt is de matrix koëfficiënt op de (q,p)-de plaats in ten hoogste $2(n-1)$ rotaties Q_l met $k_1 \leq l < k_2$ betrokken (zie (31)). Dus volgt uit (a), (35) en het feit dat de rij $(\alpha(A_k))$ dalend is dat

$$\beta(A_{k_2}) \leq 2(n-1)\frac{\alpha(A_{k_1})^2}{2(\varrho - 2\alpha(A_{k_1}))}.$$

(c) Als $m\tilde{n} \geq k_0$ dan zijn er k_1, k_2 als in (b) waarvoor $m\tilde{n} \leq k_1 < k_2 < (m+1)\tilde{n}$ (zie dat in). Dus

$$\tilde{\alpha}_{m+1} \leq \alpha(A_{k_2}) \leq n\beta(A_{k_2}) \leq n^2\frac{\alpha(A_{k_1})^2}{\varrho - 2\alpha(A_{k_1})} \leq n^2\frac{\tilde{\alpha}_m^2}{\varrho - 2\tilde{\alpha}_m}.$$

Dit bewijst de stelling voor het geval de eigenwaarden van A allen enkelvoudig zijn.

Als $\lambda_q = \lambda_{q+1}$ dan is voor k groot genoeg $|a_{q\,q+1}^{(k)}| < \beta(A_k)$ en wordt dus niet meer in het (e_q, e_{q+1})–vlak geroteerd. De argumenten in (a), (b) en (c) gaan daarom ook op voor het geval niet alle eigenwaarden enkelvoudig zijn (neem nu $\varrho := \min\{|\lambda - \mu| \,\big|\, \lambda, \mu \in \sigma(A), \lambda \neq \mu\}$). We laten het aan de lezer over in opgave (5.5.15) te bewijzen dat $|a_{q\,q+1}^{(k)}| < \beta(A_k)$ als $\lambda_q = \lambda_{q+1}$ en k groot. $\square$

(5.5.15) • Opgave. Stel dat $\lambda_1(A) = \ldots = \lambda_m(A) = \lambda_{m+1}(A) - \delta$ voor een $\delta > 0$. Stel dat $2\alpha(A_k) < \delta$ en $a_{j-1\,j-1}^{(k)} \leq a_{jj}^{(k)}$ voor $j = 2, \ldots, n$.

Laat A_k als volgt gepartitioneerd zijn $A_k = \begin{bmatrix} U & F^* \\ F & V \end{bmatrix}$ met $U \in \mathbf{M}_m$.

(a) Bewijs dat $V - \lambda_1 I$ niet-singulier is en dat $\|(V - \lambda_1 I)^{-1}\|_2 \leq 1/(\delta - \alpha_k(A))$.

(b) Met $X := U - \lambda_1 I - F^*(V - \lambda_1 I)^{-1}F \in \mathbf{M}_m$ is

$$A_k - \lambda_1(A) = \begin{bmatrix} I & F^*(V - \lambda_1 I)^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} X & 0 \\ 0 & V - \lambda_1 I \end{bmatrix} \begin{bmatrix} I & 0 \\ (V - \lambda_1 I)^{-1}F & I \end{bmatrix}.$$

Ga dit na en toon aan dat $X = 0$.

(c) Bewijs dat

$$\alpha(U) \leq \frac{\alpha(A_k)^2}{\delta - \alpha(A_k)}.$$

(d) Bewijs dat $|a_{qp}^{(k)}| < \beta(A_k)$ voor k voldoende groot en $p < q \leq m$.

Met behulp van de stellingen (5.5.13) en (5.5.14) kan men eventueel schatten hoeveel iteraties men nodig heeft om de eigenwaarden met een bepaalde nauwkeurigheid te benaderen (men raadplege het bewijs van stelling (5.5.14) voor kwantitatieve uitspraken). In de praktijk blijkt de konvergentie sneller te gaan dan de stellingen voorspellen.

(5.5.16) De hoeveelheid werk. De praktijk leert dat na $\approx 6\tilde{n}$ $(k \approx 6\tilde{n})$ iteratie stappen geldt $\beta(A_k) \leq 10^{-12}\|A\|_F$. Omdat 1 iteratie stap $\approx 4n$ flop kost hebben we in het totaal dus $\approx 12n^3$ flop nodig (omdat het proces asymptotisch zo snel konvergeert kunnen we een hogere nauwkeurigheid bereiken met nauwelijks meer werk). Willen we ook de eigenvektoren kennen dan moeten we ook nog in iedere stap $W_{k+1} := Q_k W_k$ uitrekenen, waarbij $W_0 := I$ (dan $A_k = W_k A W_k^*$) hetgeen per stap ook weer $4n$ flop kost.

Naast de rotaties moeten we per rekenstap uit de $\tilde{n}$ buiten diagonaal elementen echter ook nog de in absolute waarde grootste koëfficiënt zoeken (hetgeen duurder is dan de $\approx 4n$ rotatie–flops).

Omdat per iteratie slag slechts twee rijen en twee kolommen veranderen kunnen we het zoekwerk enigszins beperken door per rij bij te houden waar in die rij de in absolute waarde grootste koëfficiënt staat.

Om het zoek werk te vermijden past men ook wel het zogenaamde seriële Jacobi proces toe. In plaats van het opsporen van de in absolute waarde grootste buitendiagonaal koëfficiënt werkt men achtereenvolgens repeterend in een of andere vaste volgorde alle (signifikante) buitendiagonaal elementen af.

(5.5.17) De seriële Jacobi methode. De methode produceert als volgt een rij $(\hat{A}_k)$ van hermietse matrices $\hat{A}_k = (\hat{a}_{ij}^{(k)})$ en een rij $\hat{Q}_k$ van vlak rotaties :

Algorithme

(a) $k := 0$; $q := 1$; $p := 2$; $\hat{A}_0 := A$;
(b) `bereken` $c_k, s_k \in \mathbf{R}$ `als in` (5.5.12) `met` $\tau_k = (a_{qq}^{(k)} - a_{pp}^{(k)})/(2a_{pq}^{(k)})$;
(c) `Zij` $J_k := J_{pq}(\varphi_k)$ `met` $e^{i\varphi_k} = c_k + is_k$. `Bereken` $\hat{A}_{k+1} = J_k \hat{A}_k J_k^*$;
(d) $p := p + 1$; `als` $p > n$ `dan` $q := q + 1$ `en` $p := q + 1$;
`als` $q = n$ `dan` $q := 1$ `en` $p := 2$;
(e) $k := k + 1$; `ga na naar` (b);

(Als $|a_{pq}^{(k)}| \leq \bar{\xi}(|a_{pp}^{(k)}| + |a_{qq}^{(k)}|)$ vervangt men $a_{pq}^{(k)}$ ook wel door 0.)

De seriële Jacobi methode produceert een konvergente rij $(\hat{A}_k)$. Het is echter lastiger hiervan de konvergentie te bewijzen dan voor de klassieke Jacobi methode. Als alle eigenwaarden van A enkelvoudig zijn konvergeert de rij $(\hat{\alpha}_k)$ met $\hat{\alpha}_m := \alpha(\hat{A}_{m\tilde{n}})$ weer kwadratisch naar 0.

(5.5.18) • Opgave. Bewijs dat de seriële Jacobi methode een rij $(\hat{A}_k)$ produceert waarvoor $(\hat{\alpha}_m)$ kwadratisch konvergeert als de eigenwaarden van A allen enkelvoudig zijn en de rij $(\alpha(\hat{A}_k))$ naar 0 konvergeert. (Hint: inspekteer het bewijs van stelling (5.5.14).)

(5.5.19) De stabiliteit van de Jacobi methode.
Als $\hat{A}_{\tilde{n}}^\star$ de door de computer met de seriële Jacobi methode berekende $\tilde{n}$-de iterand is dan is er een unitaire matrix $\tilde{Q}$ en een matrix $E \in \mathbf{M}_n$ zodat $\hat{A}_{\tilde{n}}^\star = \tilde{Q}(A + E)\tilde{Q}^*$ met
$\|E\|_F \leq 18n^{\frac{3}{2}}\bar{\xi}\,\|A\|_F$ (dit volgt uit (5.5.4)).

Voor de door de computer berekende klassieke Jacobi iterand $A_{\tilde{n}}^\star$ hebben we een ietwat ongunstigere uitspraak: voor een unitaire $\tilde{Q}$ en een E geldt $A_{\tilde{n}}^\star = \tilde{Q}(A + E)\tilde{Q}^*$ met
$\|E\|_F \leq 18n^2\bar{\xi}\,\|A\|_F$.

Als we na $6\tilde{n}$ iteratie slagen door de computer $A_{6\tilde{n}}^\star = (\tilde{a}_{ij})$ opgeleverd krijgen en we nemen de $\tilde{a}_{jj}$ als benadering voor de eigenwaarden $\lambda_i(A)$ dan vinden we de volgende foutschatting.
$\|\tilde{Q}A\tilde{Q}^* - \operatorname{diag}(A_{6\tilde{n}}^\star)\|_F \leq \|\tilde{Q}A\tilde{Q}^* - A_{6\tilde{n}}^\star\|_F + \|A_{6\tilde{n}}^\star - \operatorname{diag}(A_{6\tilde{n}}^\star)\|_F \leq 108n^2\bar{\xi}\,\|A\|_F + \alpha(A_{6\tilde{n}}^\star)$,
de eerste term representeert de evaluatie fout en de tweede de aproximatie fout. De stelling van Wielandt–Hoffman vertelt ons dan dat

$$\frac{\sqrt{\sum_{j=1}^n |\lambda_j(A) - \lambda_j(\operatorname{diag}(A_{6\tilde{n}}^\star))|^2}}{\sqrt{\sum_{j=1}^n |\lambda_j(A)|^2}} \leq 108n^{\frac{3}{2}}\bar{\xi} + \frac{\alpha(A_{6\tilde{n}}^\star)}{\|A\|_F}.$$

(5.5.20) Opmerking. De Jacobi methode zoals we die hier besproken hebben is in essentie ook toepasbaar in geval A een normale matrix is. Er bestaat een algorithme dat iedere matrix $A \in \mathbf{M}_n$ transformeert naar een 'bijna normale' matrix TAT^{-1} waarop dan de Jacobi methode toepasbaar is.

• Het bewijs van de Wielandt–Hoffman stelling

We bewijzen tenslotte in deze paragraaf de stelling van Wielandt–Hoffman met behulp van de vlak rotaties.

(5.5.21) De stelling van Wielandt–Hoffman.
Als A en H hermiets zijn dan $\|\Lambda_A - \Lambda_H\|_F \leq \|A - H\|_F$.

Bewijs. Omdat iedere hermietse matrix H geschreven kan worden als $H = QDQ^*$ met Q unitair en D diagonaal als in het lemma 2 hieronder, volgt de stelling van Wielandt–Hoffman uit dit lemma 2 en (5.5.8). Lemma 1 gebruiken we in het bewijs van lemma 2. □

Lemma 1. *Zij* $A = \begin{bmatrix} u & f \\ f & v \end{bmatrix} \in \mathbf{M}_2$. *Zij* $D = \operatorname{diag}(d_1, d_2) \in \mathbf{M}_2$ *zodat* $d_1 < d_2$.
Als $f \neq 0$ *of* $u > v$ *dan* $\|\Lambda_A - D\|_F < \|A - D\|_F$.

Bewijs. Schrijf $a_i = \lambda_i(A)$ voor $i = 1, 2$. Laat $c, s \in \mathbf{R}$ zo zijn dat $c^2 + s^2 = 1$ en

$$\tilde{Q}^* A \tilde{Q} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} u & f \\ f & v \end{bmatrix} \begin{bmatrix} c & -s \\ s & c \end{bmatrix} = \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} = \Lambda_A.$$

Als $f \neq 0$ of $u > v$ dan is $s \neq 0$ en $a_1 < a_2$. Dus

$$\|A - D\|_F = \| \begin{bmatrix} c^2 a_1^2 - s^2 a_2 - d_1 & -cs(a_1 + a_2) \\ -cs(a_1 + a_2) & s^2 a_1 + c^2 a_2 - d_2 \end{bmatrix} \|_F =$$

$$(a_1 - d_1)^2 + (a_2 - d_2)^2 + 2s^2(d_2 - d_1)(a_1 - a_2) < (a_1 - d_1)^2 + (a_2 - d_2)^2 = \|\Lambda_A - D\|_F^2.$$ □

Lemma 2. *Laat A hermiets zijn en $D = \operatorname{diag}(d_1, \ldots, d_n)$ diagonaal met $d_1 \leq d_2 \leq \ldots \leq d_n$.*
Dan $\|\Lambda_A - D\|_F \leq \|A - D\|_F$.

Bewijs. We bewijzen dat $\|\Lambda_A - D\|_F \leq \|Q\Lambda_A Q^* - D\|_F$ voor iedere unitaire Q.
(a) Neem eerst aan dat $d_i < d_{i+1}$ voor iedere $i = 1, \ldots, n-1$.

Omdat $\mathbf{Q}_n := \{Q \in \mathbf{M}_n \,\big|\, Q^*Q = I\}$ kompakt is (begrensd en gesloten) in $\mathbf{M}_n$ en $Q \leadsto \|Q\Lambda_A Q^* - \Lambda_H\|_F$ kontinu is, is er een $Q_0 \in \mathbf{Q}_n$ zodat

$$\|Q_0 \Lambda_A Q_0^* - \Lambda_H\|_F \leq \|Q\Lambda_A Q^* - \Lambda_H\|_F \quad \text{voor alle} \quad Q \in \mathbf{Q}_n.$$

Partitioneer $Q_0 \Lambda_A Q_0^* = \begin{bmatrix} U & F^* \\ F & V \end{bmatrix}$ en $D = \begin{bmatrix} D' & 0 \\ 0 & D'' \end{bmatrix}$ met $U = \begin{bmatrix} u & f \\ f & v \end{bmatrix}$, $D' \in \mathbf{M}_2$.
Als $f \neq 0$ of $u > v$ dan is er volgens het voorgaande lemma een $\tilde{Q} \in \mathbf{Q}_2$ zodat
$\Lambda' := \tilde{Q}^* U \tilde{Q} = \operatorname{diag}(\lambda_1, \lambda_2)$ met $\lambda_1 \leq \lambda_2$ en $\|\Lambda' - D'\|_F < \|U - D'\|_F$.
Zij $Q := \begin{bmatrix} \tilde{Q} & 0 \\ 0 & I \end{bmatrix} \in \mathbf{Q}_n$. Dan

$$\begin{aligned} \|Q^* Q_0 \Lambda_A Q_0^* Q - \Lambda_H\|_F^2 &= \|\tilde{Q}^* U \tilde{Q} - D'\|_F^2 + 2\|F\|_F^2 + \|V - D''\|_F^2 \\ &< \|U - D'\|_F^2 + 2\|F\|_F^2 + \|V - D''\|_F^2 = \|Q_0 \Lambda_A Q_0^* - D\|_F^2 \end{aligned},$$

hetgeen de keuze van Q_0 tegenspreekt. Blijkbaar $f = 0$ en $u \leq v$.
Op analoge wijze volgt dat ieder buiten diagonaal element van $Q_0 \Lambda_A Q_0^*$ nul is en de diagonaal elementen geordend zijn in stijgende grootte. Dus $Q_0 \Lambda_A Q_0^* = \Lambda_A$ en de stelling is bewezen voor het geval alle d_i verschillend zijn.
(b) Als $d_i = d_{i+1}$ voor een i, dan beschouw $D + \varepsilon J$ met $J = \operatorname{diag}(1, 2, \ldots, n)$. Op eindig veel ε's na zijn alle diagonaal elementen van $D + \varepsilon J$ verschillend. Voor bijna alle ε volgt uit hetgeen we net bewezen hebben dat $\|\Lambda_A - (D + \varepsilon J)\|_F \leq \|A - (D + \varepsilon J)\|_F$. Nemen we de limiet voor $\varepsilon \to 0$ dan volgt het lemma. □

Index