Universiteit Utrecht

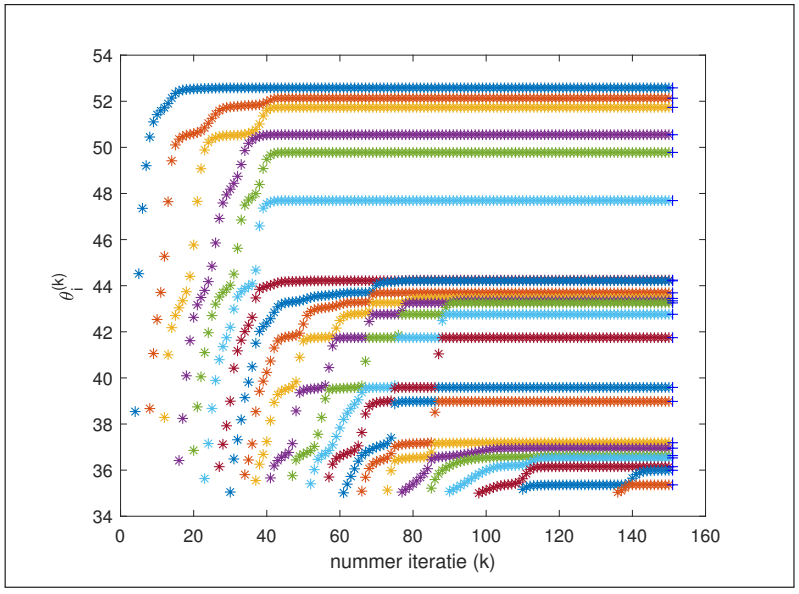
Faculteit Bètawetenschappen

Benaderen van enkele eigenwaarden en eigenvectoren van grote matrices met Krylov methodes

Bachelor Scriptie TWIN

Femke van Ieperen

Natuur- en Wiskunde



Begeleider:

Dr. S. W. Gaaf
Mathematisch Instituut

Datum van voltooiing:
11-06-2018

Samenvatting

We bekijken vier verschillende methoden om een aantal eigenparen van matrices van grote dimensies te benaderen. De machtsmethode is geschikt om het eigenpaar met de eigenwaarde met de grootste modulus te benaderen, indien deze bestaat. De inverse iteratie is te gebruiken om een eigenwaarde het dichtst bij een willekeurige complexe waarde te benaderen, indien deze uniek bepaald is. Als we meerdere eigenparen van een matrix willen bepalen zullen we ons tot de Lanczos methode moeten wenden voor een hermitische matrix en anders tot de Arnoldi methode. Ook bij deze methoden is het mogelijk om de eigenwaarde het dichtst bij een gekozen complex getal het eerst te laten convergeren.

Inhoudsopgave

1 Introductie

In veel toepassingsgebieden komen eigenwaardenprobleem voor. Bijvoorbeeld bij het analyseren van mechanische vibraties: het vinden van de resonantie frequenties van een systeem van door veren aan elkaar verbonden massa's [1]. Ook in de kwantum mechanica komen veel eigenwaardenproblemen voor: de mogelijke uitkomsten voor de meting van een operator zijn namelijk de eigenwaarden van deze operator. Het bekendste eigenwaardenprobleem uit de kwantum mechanica is de tijdonafhankelijke Schrödinger vergelijking. Dit is het eigenwaardenprobleem van de Hamiltoniaan, wat dus neerkomt op het bepalen van de mogelijke energieën van een systeem [2]. De directe algoritmen die gebruikt worden om het eigenwaardenprobleem op te lossen, bijvoorbeeld QR-iteratie en variaties hierop [3], werken prima voor kleine matrices. Echter deze algoritmen vereisen vaak een aantal bewerkingen van de orde $\mathcal{O}(n^3)$, met n de dimensie van de matrix [3]. Voor matrices met $n \geq 1000$ is de looptijd van deze algoritmen onwenselijke groot. Stel nu dat we slechts een paar eigenwaarden en eigenvectoren van een matrix willen weten, bijvoorbeeld omdat we alleen in de laagste energieniveaus van ons systeem geïnteresseerd zijn, dan willen we een graag een algoritme met een lager aantal bewerkingen aangezien we niet alle eigenparen hoeven te bepalen. In deze scriptie zullen we dergelijke algoritmen bekijken die gebaseerd zijn op Krylov deelruimten. De algoritmen die we bekijken zijn bestaande algoritmen die ook terug te vinden zijn in [1, 3, 6, 12].

In hoofdstuk 2 beginnen we met noodzakelijke achtergrond uit de lineaire algebra. De laatste paragraaf van dit hoofdstuk is gewijd aan eigenwaarde perturbatietheorie: wat is de invloed van een kleine verandering in de matrix op de eigenwaarden. Daarna bekijken we in hoofdstuk 3 het eerste algoritme: de machtsmethode. Een kleine variatie op de machtsmethode leidt tot de inverse iteratie die we in hoofdstuk 4 behandelen. Daarna bekijken we een aantal eigenschappen van een Krylov deelruimte in hoofdstuk 5. Hierna volgen twee methoden die de gehele Krylov deelruimte gebruiken om eigenparen te benaderen, de Lanczos methode, zie hoofdstuk 6, en de Arnoldi methode, zie hoofdstuk 7. Tot slot volgt de conclusie in hoofdstuk 8.

2 Achtergrond lineaire algebra & eigenwaarde perturbatietheorie

In de eerste drie paragrafen van dit hoofdstuk bespreken we een aantal basisconcepten uit de lineaire algebra die we later zullen gebruiken voor het analyseren van de algoritmen. Indien de lezer bekend is met de basis van lineaire algebra kunnen deze paragrafen overgeslagen worden. Voor verdere achtergrond uit de lineaire algebra verwijzen we naar [4, 5]. Tot slot bekijken we in paragraaf 2.4 kort de eigenwaarde perturbatietheorie. Algemeen zullen we alleen matrices $A \in \mathbb{C}^{n\times n}$ met $n \in \mathbb{N} = \{1, 2, 3, \cdots\}$ beschouwen en het element van A op rij i en kolom j aanduiden met a_{ij} voor $i, j \in \{1, 2, \cdots, n\}$.

2.1 Lineaire deelruimte en basissen

Aangezien alle matrices die we beschouwen een element zijn van $\mathbb{C}^{n\times n}$ komen, bespreken we alleen vectoren uit de vectorruimte $\mathbb{C}^n$. Met een vector bedoelen we hier een kolomvector. Echter een verzameling $\{v_1, v_2, \cdots v_k\}$ voor $k \in \mathbb{N}$ spant ook een deelvectorruimte op, die bevat is in $\mathbb{C}^n$.

Definitie 2.1. Laat $\{v_1, v_2, \cdots v_k\}$ een verzameling vectoren zijn met $v_i \in \mathbb{C}^n$ waarbij er geldt $k, n \in \mathbb{N}$, dan is de lineaire deelruimte opgespannen door deze vectoren als volgt gedefinieerd:

$$\text{span}\{v_1, v_2, \cdots v_k\} = \left\{ \sum_{i=1}^{k} \lambda_i v_i | \lambda_i \in \mathbb{C} \right\}. \tag{2.1}$$

Een vectorruimte V kan nu worden opgespannen door verschillende verzamelingen van vectoren. Indien voor elke vector $w \in V$ een unieke verzameling coëfficiënten $\lambda_1, \lambda_2, \cdots, \lambda_k \in \mathbb{C}$ bestaat, zodat er geldt $w = \sum_{i=1}^{k} \lambda_i v_i$, wordt $\{v_1, v_2, \cdots v_k\}$ een basis van V genoemd. De vectoren $\{v_1, v_2, \cdots v_k\}$ blijken een basis van V te vormen indien ze aan de volgende eisen voldoen:

- $\text{span}\{v_1, v_2, \cdots v_k\} = V$,
- $\sum_{i=1}^{k} \lambda_i v_i = 0 \Leftrightarrow \lambda_1 = \lambda_2 = \cdots \lambda_k = 0$ voor $\lambda_i \in \mathbb{C}$. Een verzameling vectoren die hieraan voldoet wordt ook wel lineair onafhankelijk genoemd.

We noemen dan k de dimensie van V. De dimensie van een vectorruimte is onafhankelijk van de basis. Later zullen we gebruiken dat voor een vectorruimte W van dimensie m het volstaat om een van de bovenstaande eisen te controleren voor verzameling van vectoren $\{w_1, w_2, \cdots, w_m\}$, om aan te tonen dat dit een basis is van W [4, hoofdstuk 2.C].

2.2 Eigenwaarden en eigenvectoren

In deze paragraaf bespreken we de definitie van eigenvectoren en eigenwaarden van een matrix. Ook bekijken we een aantal stellingen over de diagonaliseerbaarheid van matrices.
De eigenwaarden λ_i van de matrix zijn gedefinieerd als de nulpunten van het polynoom $p(\lambda) = \det(A - \lambda\mathbb{I})$. Uit de hoofdstelling van de Algebra [5, stelling 5.27], volgt dat er $c, \lambda_1, \cdots, \lambda_k \in \mathbb{C}$ en $m_1, \cdots m_k \in \mathbb{N}$, zodat er geldt $n = m_1 + m_2 + \cdots m_k$, bestaan waarvoor geldt $\det(A - \lambda\mathbb{I}) = c\,(\lambda - \lambda_1)^{m_1}\,(\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_k)^{m_k}$. Indien voor de eigenwaarde λ_i geldt $m_i = 1$, noemen we deze enkelvoudig, anders is λ_i een meervoudige eigenwaarde. Alle eigenwaarden van een matrix A te samen vormen het spectrum van een matrix, dat we noteren als

$$\Lambda(A) = \{\lambda \in \mathbb{C} | \det(A - \lambda\mathbb{I}) = 0\}.$$

Gegeven een eigenwaarde $\lambda_i \in \Lambda(A)$ worden de eigenvectoren van A behorend bij deze eigenwaarden gegeven door de vergelijking

$$Av_i = \lambda_i v_i.$$

Indien een combinatie (λ_i, v_i) aan de bovenstaande vergelijking voldoet, noemen we dit een eigenpaar van A. Alle eigenvectoren behorend bij een eigenwaarde spannen een lineaire deelruimte op, die we noteren als $\mathcal{E}(\lambda_i)$. De dimensie van deze deelruimte van de deelruimte is gelijk aan $d_i \leq m_i$. Indien er voor elke eigenwaarde van A geldt dat $d_i = m_i$ is A diagonaliseerbaar.

Definitie 2.2 (Diagonaliseerbare matrix)**.** Laat $A \in \mathbb{C}^{n\times n}$ een willekeurige matrix zijn. Dan noemen we deze matrix diagonaliseerbaar als er een inverteerbare matrix $P \in \mathbb{C}^{n\times n}$ en een diagonaalmatrix $D \in \mathbb{C}^{n\times n}$ bestaan, zodat er geldt

$$A = PDP^{-1}$$

Er blijkt zelfs te gelden:

Lemma 2.3 (Diagonaliseerbaarheid matrix & eigenvectoren (5.41 [4]))**.** *Neem een matrix $A \in \mathbb{C}^{n\times n}$. Er geldt, A is diagonaliseerbaar dan en slechts als voor alle $\lambda_i \in \Lambda(A)$ geldt $d_i = m_i$.*

De matrices P en D hebben een vorm die we uit kunnen drukken in termen van de eigenwaarden en eigenvectoren van A. Definieer de matrix $V = [v_1, v_2, \cdots v_n]$ met als kolommen de genormeerde eigenvectoren van A. Een vector heet genormeerd als deze voldoet aan $\|v_i\|_2 = 1$, zie definitie 2.12. Verder definiëren we de diagonaalmatrix Λ waarbij er geldt dat λ_{ii} gelijk is aan de eigenwaarde behorend bij v_i. Dan geldt er $A = V\Lambda V^{-1}$, wat ook wel de eigendecompositie van A wordt genoemd.

In het algemeen weten we niet op voorhand van een matrix niet of deze diagonaliseerbaar is. Echter van matrices met bepaalde eigenschappen is bekend dat deze altijd diagonaliseerbaar zijn. Deze eigenschappen maken gebruik van de hermitisch geconjugeerde van een matrix.

Definitie 2.4. (Hermitische conjugatie) Neem een matrix $A \in \mathbb{C}^{n\times n}$. De hermitisch geconjugeerde van A is dan gelijk aan:

$$A^H = A^{*T} = A^{T*}.$$

Hierbij staat A^T voor de getransponeerde van een matrix, wat wordt bepaald door ${a_{ij}}^T = a_{ji}$. Verder staat A^* voor het elementsgewijs complex conjugeren van de matrix A.

Zo blijkt een normale matrix altijd diagonaliseerbaar waarbij de matrix P uit definitie 2.2 unitair is.

Definitie 2.5 (Normale matrix)**.** Een matrix $A \in \mathbb{C}^{n\times n}$ is normaal als er geldt $A^H A = AA^H$.

Definitie 2.6 (Unitaire matrix)**.** Een matrix $A \in \mathbb{C}^{n\times n}$ noemen we unitair als er geldt $AA^H = \mathbb{I} = A^H A$ oftewel $A^{-1} = A^H$. Indien dit geldt voor een matrix $A \in \mathbb{R}^{n\times n}$ wordt A ook wel een orthogonale matrix genoemd.

Stelling 2.7. *(Spectraal stelling voor normale matrices,[6, Gevolg* 2.50*]) Zij $A \in \mathbb{C}^{n\times n}$ met $n \in \mathbb{N}$ een matrix. Dan geldt er dat A een normale matrix is dan en slechts dan als er een unitaire matrix $V \in \mathbb{C}^{n\times n}$ en een diagonaal matrix $\Lambda \in \mathbb{C}^{n\times n}$ bestaan zodat er geldt $A = V\Lambda V^{-1}$.*

Als een matrix ook hermitisch is, blijkt zelfs dat we op voorhand al weten dat alle eigenwaarden reëel zijn.

Definitie 2.8 (Hermitische matrix)**.** Een matrix $A \in \mathbb{C}^{n\times n}$ is een hermitische matrix als er geldt $A = A^H$.

Stelling 2.9. *(Spectraal stelling voor hermitische matrices,[6, Gevolg* 2.51*]) Zij $A \in \mathbb{C}^{n\times n}$ met $n \in \mathbb{N}$ een matrix. Dan geldt er dat A een hermitische matrix is dan en slechts dan als er een unitaire matrix $V \in \mathbb{C}^{n\times n}$ en een diagonaal matrix $\Lambda \in \mathbb{R}^{n\times n}$ bestaan zodat er geldt $A = V\Lambda V^{-1}$.*

2.3 Inproduct & norm

In de vorige paragraaf hebben we een vectornorm gebruikt. De vectornorm die wij zullen hanteren is afgeleid van een inproduct. Een functie van $\mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$ op $\mathbb{C}^n$ is een inproduct, indien deze aan de volgende vier eigenschappen voldoet:

1. $\langle .,. \rangle$ is lineair in de tweede component. Oftewel $\langle x, \lambda y + \mu z\rangle = \lambda\langle x, y\rangle + \mu\langle x, z\rangle$ voor $x, y, z \in \mathbb{C}^n$ en $\lambda, \mu \in \mathbb{C}$;
2. $\langle x, y\rangle = (\langle y, x\rangle)^*$ voor alle $x, y \in \mathbb{C}^n$;
3. $\langle x, x\rangle \geq 0$ voor alle $x \in \mathbb{C}^n$;
4. Er geldt $\langle x, x\rangle = 0$ dan en slechts dan als $x = \vec{0}$ voor $x \in \mathbb{C}^n$.

Uit deze lijst met eigenschappen volgt dat elk inproduct ook aan de volgende eigenschap zal voldoen

$$\langle \lambda x + \mu y, z\rangle = \lambda^*\langle x, z\rangle + \mu^*\langle y, z\rangle \qquad \text{voor alle } x, y, z \in \mathbb{C}^n \text{ en } \lambda, \mu \in \mathbb{C}.$$

Deze eigenschap noemen we de toegevoegde lineariteit van het inproduct.

Definitie 2.10 (Inproduct)**.** Laat $x, y \in \mathbb{C}^n$ twee vectoren zijn met $n \in \mathbb{N}$. Dan definiëren we het inproduct tussen deze twee vectoren als volgt:

$$\langle x, y\rangle = \sum_{i=1}^{n} (x_i)^* y_i = x^H y. \tag{2.2}$$

Het is eenvoudig na te gaan dat dit inproduct aan de bovengestelde eisen voldoet. Nu we een inproduct hebben kunnen we de notie van een verzameling orthogonale vectoren introduceren.

Definitie 2.11 (Orthogonaliteit vectoren)**.** Neem een verzameling vectoren $\{v_1, v_2, \cdots v_k\}$ met $v_i \in \mathbb{C}^n$. We noemen deze verzameling van vectoren orthogonaal indien er geldt

$$\langle v_i, v_j\rangle = 0 \,\text{voor alle}\, i, j \in \{1, 2, \cdots k\} \,\text{met}\, i \neq j.$$

We willen nu met behulp van ons inproduct een vectornorm definieëren. We weten dat een vectornorm een functie van $\mathbb{C}^n \to \mathbb{R}$ is die aan de volgende eigenschappen voldoet:

1. $||x|| \geq 0$ voor alle $x \in \mathbb{C}^n$;
2. $||x|| = 0$ dan en slecht dan als $x = \vec{0}$;
3. $||\lambda x|| = |\lambda|||x||$ voor alle $\lambda \in \mathbb{C}$ en $x \in \mathbb{C}^n$;

4. $||x + y|| \leq ||x|| + ||y||$ voor alle $x, y \in \mathbb{C}^n$.

We weten dat $||.|| = \sqrt{\langle ., . \rangle}$ een norm definieert indien $\langle ., . \rangle$ een inproduct vormt.[4, hoofdstuk 6].

Definitie 2.12 (2-norm vectoren)**.** Laat $x \in \mathbb{C}^n$ een vector zijn, dan definiëren we de norm van x als volgt

$$||x||_2 = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^{n} |x_i|}.$$

Nu kunnen we ook de notie van een orthonormale verzameling vectoren geven. De verzameling $\{v_1, v_2, \cdots, v_k\}$ met $v_i \in \mathbb{C}^n$ is orthonormaal indien deze orthogonaal is en er voor alle i geldt $||v_i||_2 = 1$.
We willen niet alleen een vectornorm hebben maar ook een matrixnorm. De norm van een matrix is namelijk een vaak gebruikt concept om de convergentie of nauwkeurigheid van een algoritme af te schatten. Wij zullen als matrixnorm de operatornorm afgeleid van de 2-norm voor vectoren, definitie 2.12 gebruiken.

Definitie 2.13 (2-norm matrices)**.** Laat $A \in \mathbb{C}^{n \times n}$ een matrix. De norm van A definiëren we nu als volgt

$$||A||_2 = \max_{x \in \mathbb{C}^n, x \neq \vec{0}} \frac{||Ax||_2}{||x||_2}.$$

Deze norm voldoet aan de volgende vijf eigenschappen [1]:

- $||A||_2 \leq 0$ voor alle $A \in \mathbb{C}^{n \times n}$;
- $||A||_2 = 0$ dan en slecht dan als $A = 0$;
- $||\lambda A||_2 = |\lambda| ||A||_2$ voor $\lambda \in \mathbb{C}$ en $A \in \mathbb{C}^{n \times n}$;
- $||A + B||_2 \leq ||A||_2 + ||B||_2$ voor alle $A, B \in \mathbb{C}^{n \times n}$;
- $||AB||_2 \leq ||A||_2 ||B||_2$ voor alle $A, B \in \mathbb{C}^{n \times n}$.

Het blijkt dat de 2-norm van een matrix uit te drukken is in termen van de spectrale radius van $A^H A$.

Definitie 2.14 (spectrale radius)**.** Laat $A \in \mathbb{C}^{n \times n}$ een matrix zijn. Dan wordt de spectrale radius van A gegeven door

$$\rho(A) = \max_i \{|\lambda_i| \,|\, \lambda_i \in \Lambda(A)\}.$$

Aangezien de matrix $A^H A$ per constructie een hermitische matrix is, volgt er uit stelling 2.9 dat de spectrale radius $A^H A$ gelijk is aan de wortel van absolute waarde van de grootste eigenwaarde.

Propositie 2.15. *Laat $A \in \mathbb{C}^{n \times n}$ een matrix zijn dan geldt er*

$$||A||_2 = \sqrt{\rho(A^H A)}.$$

Indien A hermitisch is volgt er bovendien

$$||A||_2 = \rho(A).$$

Bewijs. Er geldt

$$||A||_2^2 = \max_{x \in \mathbb{C}^n, x \neq \vec{0}} \frac{x^H A^H A x}{x^H x}.$$

Uit stelling 2.9 volgt er dat geldt $A^H A = V \Lambda V^H$,

$$= \max_{x \in \mathbb{C}^n, x \neq \vec{0}} \frac{x^H V \Lambda V^H x}{x^H V V^H x}.$$

Gebruik nu de basis transformatie $y = V^H x$,

$$= \max_{y\in\mathbb{C}^n, y\neq\vec{0}} \frac{y^H \Lambda y}{y^H y} = \max_{y\in\mathbb{C}^n, y\neq\vec{0}} \frac{\sum_i^n |\lambda_i||y_i|^2}{\sum_{i=1}^n |y_i|^2} = \max_{\lambda_i\in\Lambda(A^H A)} |\lambda_i| = \rho(A^H A).$$

Er volgt nu $||A||_2 = \sqrt{\rho(A^H A)}$.
Als A een hermitisch is, geldt er $A = V\Lambda V^H$ met $\Lambda \in \mathbb{R}^{n\times n}$ en dus $A^H A = V\Lambda^2 V^H$. Hieruit volgt dat geldt $||A||_2 = \sqrt{\rho(A^H A)} = \sqrt{\max_i\{|\lambda_i|^2 | \lambda \in \Lambda(A)\}} = \rho(A)$ voor een hermitische matrix. □

2.4 Eigenwaardenperturbatietheorie

We willen numeriek de eigenwaarden van een matrix $A \in \mathbb{C}^{n\times n}$ benaderen. Echter door de eindige precisie van de computer kan de numerieke implementatie $\tilde{A}$ van A verschillen van A zelf. Het algoritme benadert dus de eigenwaarden $\tilde{\lambda_i}$ van $\tilde{A}$ in plaats van de eigenwaarden λ_i van A. Maar eigenwaarden θ_i uit het algoritme zijn ook niet exact gelijk aan $\tilde{\lambda_i}$. De totale fout in de benadering is dus gelijk aan $|\theta_i - \lambda_i| \leq |\theta_i - \tilde{\lambda_i}| + |\tilde{\lambda_i} - \lambda_i|$. De term $|\theta_i - \tilde{\lambda_i}|$ is de fout van het algoritme. Dit zullen we later per algoritme bespreken. De term $|\tilde{\lambda_i} - \lambda_i|$ beschrijft het verschil tussen de eigenwaarden van A en $\tilde{A}$. en wordt beschreven door eigenwaardenperturbatietheorie. Dit is waar we ons in deze paragraaf over zullen buigen.

Definieer de matrix E zodat er geldt $\tilde{A} = A + E$. Voor een hermitische matrix volgt dan dat de eigenwaarden van A en $\tilde{A}$ niet meer dan $||E||_2$ afwijken.

Stelling 2.16. *(Stelling van Weyl [7, Stelling 1.1]) Laat $A, A+E \in \mathbb{C}^{n\times n}$ twee hermitische matrices zijn met eigenwaarden $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ en $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. Dan geldt er*

$$\max_i |\lambda_i - \mu_i| \leq ||E||_2. \tag{2.3}$$

Om voor een algemene matrix iets over het verschil tussen de eigenwaarden van A en $\tilde{A}$ te zeggen, definiëren we de linker eigenvectoren u_i van een matrix. Deze worden bepaald door de vergelijking

$$u_i^H A = \lambda_i u_i^H,$$

met $\lambda_i \in \Lambda(A)$. Om het onderscheid tussen deze eigenvectoren en die uit paragraaf 2.2 duidelijk te maken noemen we de v_i's de rechter eigenvectoren. Als er geen verwarring bestaat over dat we de rechter eigenvectoren bedoelen, noemen we deze simpelweg de eigenvectoren. Indien de matrix A n verschillende eigenwaarden heeft geldt er

$$|\lambda_1 - \tilde{\lambda_1}| \leq |t| \frac{||u_1||||v_1||}{|\langle u_1, v_1\rangle|} + \mathcal{O}(|t|^2), \tag{2.4}$$

met $t \in \mathbb{C}$ en $\tilde{A} = A + tF$ en $||F||_2 = 1$.[8, vergelijking (52.3)]. Indien $\mathcal{O}(|t|^2)$ verwaarloosbaar is volgt er nu dat de fout in $\tilde{\lambda_i}$ altijd gelijk is aan $\kappa(\lambda_i)|t|$ met

$$\kappa(\lambda_i) = \frac{||u_i||_2||v_i||_2}{|\langle u_i, v_i\rangle|}. \tag{2.5}$$

Dit wordt ook wel het conditiegetal van de eigenwaarde λ_i genoemd. Wegens de Cauchy-Schwarz vergelijking [4, 6.15] geldt er $\kappa(\lambda_i) \geq 1$. Ook volgt er $\kappa(\lambda_i) = 1$ dan en slechts dan als $u_i = \alpha v_i$ voor $\alpha \in \mathbb{C}$. Voor een normale matrix geldt dat voor alle eigenwaarden het conditiegetal gelijk is aan 1, voor een niet normale matrix is het alleen mogelijk dat dit voor enkele van de eigenwaarden geldt [8, §52]. Voor een eigenwaarde met $\kappa(\lambda_i) > 1$ wordt de fout t in $\tilde{\lambda_i}$ dus vergroot. We definiëren nu de spectrale projector $\mathcal{P}_i$ waarvoor geldt $||\mathcal{P}_i||_2 = \kappa(\lambda_i)$.

Definitie 2.17. (Spectrale projector) Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en λ_i een enkelvoudige eigenwaarde van A zijn. Neem u_i en v_i de genormeerde linker en rechter eigenvector behorend bij λ_i. De spectrale projector voor λ_i volgt nu als

$$\mathcal{P}_i = v_i u_i.$$

Deze projector projecteert een willekeurige vector op de ruimte span$\{v_i\}$. Verder geldt er $A\mathcal{P}_i = \mathcal{P}_i A = \lambda_i \mathcal{P}_i$. [8, §52]. Het spectrum van de matrix $\tilde{A}$ is onderdeel van het pseudospectrum van A.

Definitie 2.18 (ϵ-pseudospectrum)**.** Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en neem $\epsilon \in \mathbb{R}_{>0}$. Het ϵ-pseudospectrum van A volgt nu als:

$$\Lambda_\epsilon(A) = \{z \in \mathbb{C} | z \in \Lambda(A+E) \text{ met } ||E||_2 \leq \epsilon\}.$$

Dit is de definitie van een pseudospectrum uit [9, (3.2)], die equivalent is met de definitie uit [8], zie [9]. We kunnen nu analoog aan het spectrum van A en de spectrale radius ook de pseudo-spectrale radius van A definiëren.

Definitie 2.19 (pseudo-spectrale radius)**.** Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en neem $\epsilon \in \mathbb{R}_{>0}$. De pseudo-spectrale radius ρ_ϵ is gelijk aan

$$\rho_\epsilon(A) = \sup_{z\in\Lambda_\epsilon(A)} |z|.$$

Deze waarde zullen we in volgende hoofdstukken gebruiken om de convergentie en nauwkeurigheid van algoritmen te bepalen.

Als we nu stellen $||E||_2 = \epsilon$ volgt er dus $\Lambda(\tilde{A}) \subset \Lambda_\epsilon(A)$.

Stelling 2.20. *(Bauer-Fike Stelling [8, Stelling 52.2]) Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn met n verschillende eigenwaarden. Dan geldt er voor alle $\epsilon > 0$*

$$\Lambda_\epsilon(A) \subset \bigcup_{i=1}^{n} \left(\lambda_i + \Delta_{\epsilon n \kappa(\lambda_i)}\right),$$

waarbij geldt $\Delta_\beta = \{z \in \mathbb{C} | \ |z| < \beta\}$.

Er volgt nu voor een algemene matrix met n verschillende eigenwaarden dat $|\lambda_i - \tilde{\lambda_i}|$ altijd van boven begrensd is door $\max_i \Delta_{\epsilon n \kappa(\lambda_i)}$.

3 Machtsmethode

In dit hoofdstuk bespreken we de machtsmethode. Dit is het eenvoudigste algoritme dat we zullen beschouwen. De machtsmethode is een algoritme om de eigenwaarde met de grootste modulus en de bijbehorende eigenvector van een matrix te bepalen. We beginnen dit hoofdstuk met het uitwerken van de theorie waarop de machtsmethode zijn werking baseert in paragraaf 3.1. We kijken hierbij naar een diagonaliseerbare matrix met één eigenwaarde met de grootste modulus. Vervolgens introduceren we het algoritme voor de machtsmethode in paragraaf 3.2. Hierna analyseren we wat er gebeurt bij toepassing van de machtsmethode op een normale matrix met meerdere eigenwaarden met de grootste modulus in paragraaf 3.3. Daarna bekijken we de convergentie en de nauwkeurigheid van het algoritme in paragraaf 3.4.Tot slot bekijken we een aantal numerieke voorbeelden, zie paragraaf 3.5.

3.1 Theorie achter de machtsmethode

We zullen nu het idee achter de machtsmethode uitwerken. Stel dat we van een diagonaliseerbare matrix $A \in \mathbb{C}^{n\times n}$ de eigenwaarde met de grootste modulus en de bijbehorende eigenvector willen bepalen. Omdat deze matrix diagonaliseerbaar is, bestaan dus een inverteerbare matrix $V \in \mathbb{C}^{n\times n}$ en een diagonaalmatrix $\Lambda \in \mathbb{C}^{n\times n}$ zodanig dat geldt $A = V\Lambda V^{-1}$. Hierbij zijn de diagonaalelementen van Λ de eigenwaarden van A en de kolommen van V zijn de eigenvectoren van A. Oftewel als we met v_i de i^{de} kolom van V noteren geldt er dus $Av_i = \lambda_i v_i$. We nemen verder aan dat de eigenvectoren van A genormaliseerd zijn, oftewel dat voor alle $i \in \{1, 2, \cdots, n\}$ geldt $||v_i||_2 = 1$. Aangezien we nu een basis van eigenvectoren hebben, kunnen we elke vector

$x_0 \in \mathbb{C}^n$ uitdrukken in de basis van eigenvectoren. Voor x_0 bestaan er dus $\alpha_i \in \mathbb{C}$ zodat er geldt $x_0 = \sum_{i=1}^{n} \alpha_i v_i$. Als we nu de vector $A^k x_0$ in deze basis van eigenvectoren uitschrijven, voor $k \in \mathbb{N}$, vinden we:

$$A^k x_0 = A^k \sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} \lambda_i^k \alpha_i v_i. \tag{3.1}$$

Stel nu dat er voor de eigenwaarden van A geldt

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|. \tag{3.2}$$

Dan kunnen we λ_1^k voor de sommatie halen in vergelijking (3.1). Dit leidt tot de volgende uitdrukking voor $A^k x_0$:

$$A^k x_0 = \lambda_1^k \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_1}\right)^k \alpha_i v_i. \tag{3.3}$$

Aangezien er geldt dat $|\lambda_1| > |\lambda_i|$ voor $i \neq 1$, volgt nu $\left(\frac{\lambda_i}{\lambda_1}\right)^k \to \begin{cases} 0 & \text{voor } i \neq 1 \\ 1 & \text{voor } i = 1 \end{cases}$ voor $k \to \infty$. Indien ook geldt dat $\alpha_1 \neq 0$ volgt er dus

$$A^k x_0 \to \lambda_1^k \alpha_1 v_1 \quad \text{voor} \quad k \to \infty. \tag{3.4}$$

Oftewel als de matrix A één eigenwaarde λ_1 heeft met de grootste modulus en we bepalen $A^k x_0$, met een vector x_0 die een component ongelijk aan nul heeft in de richting van v_1, convergeert $A^k x_0$ naar de eigenvector behorend bij λ_1. Als we deze vector vervolgens normaliseren verkrijgen we v_1. Wanneer we eenmaal de genormaliseerde eigenvector v_1 hebben, kunnen we de bijbehorende eigenwaarde bepalen. Deze voldoet namelijk aan de vergelijking $\lambda_1 = \langle v_1, Av_1 \rangle$. Immers als v_1 een eigenvector is, geldt er dat $Av_1 = \lambda_1 v_1$. Ook we hebben v_1 genormaliseerd, dus geldt inderdaad $\langle v_1, Av_1 \rangle = \langle v_1, \lambda_1 v_1 \rangle = \lambda_1 \langle v_1, v_1 \rangle = \lambda_1$.

3.2 Algoritme machtsmethode

De machtsmethode is gebaseerd op de limiet uit vergelijking (3.4) en is weergegeven in algoritme 1. We zien in het algoritme in regel 3 en 7 dat we $y_k = \frac{A^k x_0}{\|A^k x_0\|_2}$ bepalen en niet $A^k x_0$. Dit doen we om te voorkomen dat $\lambda_1^k \alpha_1$ tegen de eindige precisie van de computer aanloopt. Als benadering van de eigenwaarde behorend bij y_k definiëren we $\theta_k = \langle y_k, Ay_k \rangle$, zie regel 5 en 9. De definitie van θ_k is zo gekozen omdat, als y_k wel een echte eigenvector is, dan θ_k de echte eigenwaarde behorend bij die eigenvector is. In regel 6 van het algoritme zien we dat het benaderde eigenpaar (θ_k, y_k) als geconvergeerd wordt beschouwd indien de norm van

$$r_k = Ay_k - \theta_k y_k, \tag{3.5}$$

kleiner is dan *tol*. Hierbij is *tol* de zelf te bepalen gewenste nauwkeurigheid van het benaderd eigenpaar. We noemen r_k het residu van het benaderd eigenpaar (θ_k, y_k). Er volgt nu dat $r_k = \vec{0}$ als (θ_k, y_k) een echt eigenpaar is van A. Het residu r_k geeft dus aan hoever het benaderde eigenpaar afligt van een echt eigenpaar van A en is dus een zinvol criterium voor de convergentie van het benaderde eigenpaar.

3.3 Normale matrices & de machtsmethode

In het algemeen is het voor een matrix A niet bekend of de eigenwaarden voldoen aan vergelijking (3.2). Nu rijst dus de vraag: wat er gebeurt met θ_k en y_k uit de machtsmethode als we een matrix A hebben waarvoor niet één eigenwaarde met de grootste modulus bestaat?
Hiertoe zullen we analyseren wat er gebeurt als we de machtsmethode toepassen op een normale matrix $B \in \mathbb{C}^{n \times n}$, waarvoor geldt

$$|\lambda_1| = |\lambda_2| = \cdots = |\lambda_j| > |\lambda_{j+1}| \geq \cdots \geq |\lambda_n|, \tag{3.6}$$

met $j \in \{1, 2, \cdots, n\}$. We willen nu weten wat de uitdrukkingen zijn voor θ_k en y_k na k iteraties van de machtsmethode uitgevoerd op B met een startvector $x_0 \in \mathbb{C}^n$. Ook willen we weten of θ_k en y_k convergeren

def *power*(A, x_0, tol)**:**

Input : -$A \in \mathbb{C}^{n\times n}$ een matrix.
- $x_0 \in \mathbb{C}^n$ de startvector van de iteratie.
- *tol* de gewenste nauwkeurigheid.

Output: -(θ, y) de benadering van het eigenpaar van A voor de eigenwaarde met de grootste modulus.

$w = x_0$
$y = \frac{w}{||w||_2}$
$w = Ay$
$\theta = y^H w$
while $||w - \theta y||_2 \geq tol$**:**
    $y = \frac{w}{||w||_2}$
    $w = Ay$
    $\theta = y^H w$
end

Algoritme 1: Machtsmethode

voor $k \to \infty$ en indien ze convergeren ook waarnaar ze convergeren. Aangezien B normaal is, bestaan er volgens Stelling 2.7 een unitaire matrix $V \in \mathbb{C}^{n\times n}$ en een diagonaalmatrix $\Lambda \in \mathbb{C}^{n\times n}$ zodat er geldt $B = V\Lambda V^{-1}$. Er bestaan dus, net zoals in paragraaf 3.1, $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{C}$ zodat geldt $x_0 = \sum_{i=1}^n \alpha_i v_i$. Er volgt nu dat θ_k na k iteraties gelijk is aan:

$$\begin{aligned}
\theta_k &= \left(\frac{B^k x_0}{||B^k x_0||_2}\right)^H \cdot \frac{B^{k+1}x_0}{||B^k x_0||_2} \\
&= \frac{1}{x_0^H \left(B^H\right)^k B^k x_0} \begin{pmatrix}\alpha_1^* & \alpha_2^* & \cdots & \alpha_n^*\end{pmatrix} V^H V \left(\Lambda^H\right)^k V^H V \Lambda^{k+1} V^H V \begin{pmatrix}\alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n\end{pmatrix} \\
&= \frac{1}{x_0^H V \left(\Lambda^H\right)^k V^H V \Lambda^k V^H x_0} \sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^{2k} \lambda_i \\
&= \frac{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^{2k} \lambda_i}{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^{2k}} \\
&= \frac{\sum_{i=1}^n |\alpha_i|^2 |\frac{\lambda_i}{\lambda_1}|^{2k} \lambda_i}{\sum_{i=1}^n |\alpha_i|^2 |\frac{\lambda_i}{\lambda_1}|^{2k}}.
\end{aligned}$$

Uit vergelijking 3.6 volgt er dat $|\frac{\lambda_i}{\lambda_1}| = \begin{cases} 1 & \text{voor } i \leq j \\ < 1 & \text{voor } i > j \end{cases}$. Hieruit volgt dat voor $k \to \infty$ geldt dat

$$\theta_k \to \frac{\sum_{i=1}^j |\alpha_i|^2 \lambda_i}{\sum_{i=1}^n |\alpha_i|^2}.$$

We hebben nu dus een uitdrukking voor θ_k gevonden. Ook weten we nu dat θ_k convergeert en waarnaartoe het convergeert voor $k \to \infty$. We zien ook dat in het geval dat $j = 1$ we inderdaad krijgen dat $\theta_k \to \lambda_1$ voor $k \to \infty$, zoals we in paragraaf 3.1 hebben afgeleid. Nu willen we bekijken waaraan de benadering van de

eigenvector y_k gelijk is na k iteraties. Er volgt dat:

$$\begin{aligned} y_k &= \frac{B^k x_0}{||B^k x_0||_2} = \frac{\sum_{i=1}^n \alpha_i \lambda_i^k v_i}{\sqrt{\sum_{i=1}^n |\alpha_i|^2 |\lambda_i|^{2k}}} \\ &= \frac{\lambda_1^k}{|\lambda_1|^k} \frac{\sum_{i=1}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^n |\alpha_i|^2 \left|\frac{\lambda_i}{\lambda_1}\right|^{2k}}} \\ &= \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\sum_{i=1}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^n |\alpha_i|^2 \left|\frac{\lambda_i}{\lambda_1}\right|^{2k}}}. \end{aligned}$$

Aangezien er nog steeds geldt dat $|\frac{\lambda_i}{\lambda_1}| = \begin{cases} 1 & \text{voor } i \leq j \\ < 1 & \text{voor } i > j \end{cases}$, volgt er dat voor $k \to \infty$

$$y_k \to \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\sum_{i=1}^j \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^j |\alpha_i|^2}}. \tag{3.7}$$

Er is nu niet noodzakelijk sprake van convergentie voor y_k voor $k \to \infty$. Immers we weten alleen dat er geldt $\left|\frac{\lambda_i}{\lambda_1}\right| = 1$, voor $i \in \{1, 2, \cdots, j\}$. Dus kunnen we niet zeggen of $\left(\frac{\lambda_i}{\lambda_1}\right)^k$ convergeert voor $k \to \infty$. Wat we wel met zekerheid kunnen zeggen is dat het element altijd op de eenheidscirkel $S^1 \subset \mathbb{C}$ zal liggen aangezien de modulus wel voor alle $k \in \mathbb{N}$ gelijk is aan 1. Wel geldt voor $k \to \infty$, $y_k \in \text{span}\{v_1, v_2, \cdots, v_j\}$.

We weten nu dus wat er met y_k en θ_k gebeurt gedurende de iteraties en welke waarde ze aannemen voor $k \to \infty$. Er volgt nu voor $k \to \infty$

$$B \cdot y_k = B \cdot \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\sum_{i=1}^j \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^j |\alpha_i|^2}} \qquad \text{en} \qquad \theta \cdot y = \left(\frac{\sum_{i=1}^j |\alpha_i|^2 \lambda_i}{\sum_{i=1}^n |\alpha_i|^2}\right) \cdot \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\sum_{i=1}^j \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^j |\alpha_i|^2}}.$$

$$= \left(\frac{\lambda_1}{|\lambda_1|}\right)^k \frac{\sum_{i=1}^j \alpha_i \lambda_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^j |\alpha_i|^2}},$$

We zien nu dat alleen indien $\lambda_1 = \lambda_2 = \cdots \lambda_j$ er geldt $||By_k - \theta_k y_k||_2 \to 0$ voor $k \to \infty$. In alle andere gevallen met $j \geq 2$ en minstens twee verschillende waarden $i \in \{1, 2, \cdots, j\}$ zodat geldt $\alpha_i \neq 0$ convergeert de machtsmethode voor B dus niet.

3.4 Convergentie & Nauwkeurigheid

We bestuderen nu hoe snel de machtsmethode convergeert en hoe nauwkeurig de benadering van (λ_1, v_1) door (θ, y) is. We zullen alleen matrices beschouwen waarvan de eigenwaarden voldoen aan vergelijking (3.2), tenzij anders vermeld.
Als maat voor de convergentie van de machtsmethode zullen we de sinus van de hoek tussen v_1 en y_k beschouwen.

Definitie 3.1. (Sinus van de hoek tussen een vector en een lineaire deelruimte)
Laat $x \in \mathbb{C}^n$ een vector zijn met $x \neq \vec{0}$. Laat $\{y_1, y_2, \cdots y_k\}$ met $y_i \in \mathbb{C}^n$ een basis zijn van de lineaire deelruimte W. Dan definiëren we de sinus van de hoek tussen x en W als:

$$\sin(x, W) = \min_{w \in W} \frac{||x - w||_2}{||x||_2}.$$

Als W deelruimte van dimensie één is, met y als mogelijke basis, noteren we $\sin(x, W)$ ook met $\sin(x, y)$ en spreken we over de sinus van de hoek tussen x en y.

Er volgt nu dat de $\sin(x, y)$ afneemt als de hoek tussen de twee vectoren afneemt en we de lengte van de vectoren constant houden. Immers als de hoek tussen de twee vectoren afneemt zal $||x - y_s||_2$ ook afnemen. Dus voor de machtsmethode om te convergeren moet $\sin(v_1, y_k)$ afnemen als k toeneemt. Indien er geldt dat $\sin(v_1, y_k) = 0$ is de hoek tussen v_1 en y_k nul en aangezien beide vectoren door normalisatie lengte één hebben,p geldt er dan $v_1 = cy_k$ voor zekere $c \in S^1$. Voor de waarde van $\sin(v_1, y_k)$ blijken we de volgende bovengrens te hebben:

Stelling 3.2 (Convergentie machtsmethode)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn waarvan de eigenwaarden voldoen aan* (3.2)*. Laat v_1 de genormeerde eigenvector zijn behorend bij de eigenwaarde λ_1. Neem verder y_k de benadering van de grootste eigenvector na k iteraties van de machtsmethode. Dan geldt er dat:*

$$\sin(v_1, y_k) \leq \frac{1}{\epsilon|\alpha_1|} \frac{\rho_\epsilon(A(\mathbb{I} - \mathcal{P}))^k}{|\lambda_1|^{k-1}}. \tag{3.8}$$

Hierbij is ρ_ϵ de pseudo-spectrale radius. Deze vergelijking geldt voor alle $\epsilon > 0$. Verder is $\mathcal{P}$ de spectrale projector behorend bij eigenwaarde λ_1.
In het geval dat A diagonaliseerbaar is vereenvoudigt deze grens tot

$$\sin(v_1, y_k) \leq \frac{||V||_2||V^{-1}||_2}{|\alpha_1|} \left|\frac{\lambda_2}{\lambda_1}\right|^{k-1}. \tag{3.9}$$

Hierbij is V de matrix met de basis van eigenvectoren van A.

Hierbij zijn $\mathcal{P}$ en ρ_ϵ zoals gedefinieerd in hoofdstuk 2. Deze stelling komt voort uit het combineren van het eerste deel van stelling 28.1, vergelijking (16.10) en (16.6) [8]. We vinden nu dus dat $\sin(v_1, y_{k+1}) \leq \frac{\rho_\epsilon(A(\mathbb{I}-\mathcal{P}))}{|\lambda_1|} \sin(v_1, y_k)$. Uit de definitie van $\rho_\epsilon(A(\mathbb{I} - \mathcal{P}))$ volgt dat er geldt $\rho_\epsilon(A(\mathbb{I} - \mathcal{P})) \to |\lambda_2|$ voor $\epsilon \to 0$. Voor een diagonaliseerbare matrix vinden we $\sin(v_1, y_{k+1}) \leq \frac{|\lambda_2|}{|\lambda_1|} \sin(v_1, y_k)$, en wordt de convergentiesnelheid van de machtsmethode dus bepaald door $\frac{|\lambda_2|}{|\lambda_1|}$. Voor een algemene matrix wordt deze bepaald door $\frac{\rho_\epsilon(A(\mathbb{I}-\mathcal{P}))}{|\lambda_1|}$. Voor een diagonaliseerbare matrix is er dus altijd sprake van convergentie omdat $|\lambda_1| > |\lambda_2|$ geldt. Echter bij een niet diagonaliseerbare matrix kan het zijn dat de machtsmethode niet convergeert indien er geldt $\rho_\epsilon(A(\mathbb{I} - \mathcal{P})) > |\lambda_1|$.

Nu we gezien hebben wat de convergentiesnelheid van de machtsmethode is, willen we ook weten hoe nauwkeurig deze is. De volgende stelling geeft een indicatie van het verschil in modulus tussen λ_1 en θ_k.

Stelling 3.3 (Nauwkeurigheid machtsmethode, [8] stelling 28.1)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn waarvan de eigenwaarden voldoen aan* (3.2)*, en laat v_1 de eigenvector zijn behorend bij λ_1. Neem θ_k de benadering van de grootste eigenwaarde en y_k de bijbehorende benaderde eigenvector na k iteraties van de machtsmethode. Dan geldt er dat*

$$|\lambda_1 - \theta_k| \leq 4 \frac{\sin(v_1, y_k)||A||_2}{\sqrt{1 - \sin^2(v_1, y_k)}}. \tag{3.10}$$

In combinatie met stelling 3.2 volgt hieruit dat $|\lambda_1 - \theta_k| \to 0$ voor $k \to \infty$ voor een diagonaliseerbare matrix. Dit komt overeen met onze eerdere afleiding over de werking van de machtsmethode. Of dit ook het geval is voor een niet diagonaliseerbare matrix hangt af van of $\frac{\rho_\epsilon(A(\mathbb{I}-\mathcal{P}))}{|\lambda_1|}$ kleiner is dan 1.

Tot slot hebben we in het geval dat A een hermitische matrix is, een afschatting voor $\min_{\lambda_i \in \Lambda(A)} |\lambda_i - \theta_k|$ en $||v_j - y_k||_2$, met $v_j \in \mathcal{E}(\lambda_j)$ voor $\lambda_j = \text{argmin}_{\lambda_i \in \Lambda(A)} |\lambda_i - \theta_k|$, in termen van het residu r_k. Dit geeft dus geen grens op de fout tussen het eigenpaar (λ_1, v_1) en het benaderde eigenpaar (θ_k, y_k). Echter stelt dit ons wel instaat om de fout tussen het benaderde eigenpaar en een echt eigenpaar van A dat het dichtstbij het benaderde eigenpaar ligt af te schatten. Dit wordt de lokale fout van het benaderde eigenpaar genoemd. Om deze afschatting te geven, definiëren we eerst nog de eigenwaarde kloof als volgt:

Definitie 3.4. (Eigenwaarde kloof [6]]Definitie 3.12]) Zij $A \in \mathbb{C}^{n\times n}$ een hermitische matrix. Voor alle $\lambda \in \Lambda(A)$ definiëren we de eigenwaarde kloof van λ en A als

$$\gamma_A(\lambda) = \inf\{|\mu - \lambda| \,|\mu \in \Lambda(A) \setminus \{\lambda\}\}. \tag{3.11}$$

De afschattingen van de fout van het benaderde eigenpaar met het werkelijke eigenpaar worden gegeven door de volgende stelling:

Stelling 3.5 (Lokale fout machtsmethode, hermitische matrix)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een hermitische matrix zijn. Neem nu θ_k en y_k de benadering van de grootste eigenwaarde en bijbehorende eigenvector van de machtsmethode na k iteraties. Laat r_k het residu zijn van (θ_k, y_k) zoals gedefinieerd in vergelijking* (3.5) *. Dan bestaan er een eigenwaarde λ van A en een vector $x \in \mathcal{E}(\lambda)$, zodat er geldt:*

$$|\lambda - \theta_k| \leq ||r_k||_2, \quad ||x - y_k||_2 \leq 2\frac{||r_k||_2}{\gamma_A(\lambda)}. \tag{3.12}$$

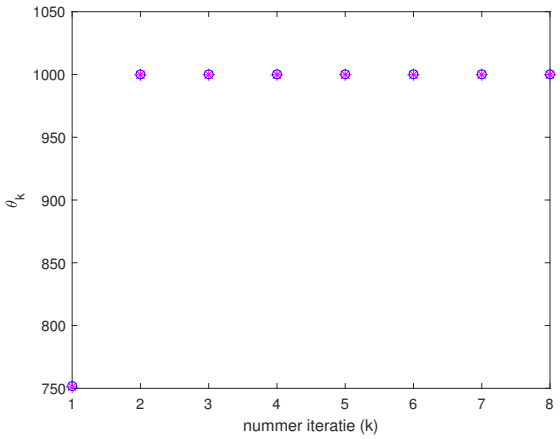
Deze stelling is de specifieke toepassing van Stelling 4.9 [6] op de machtsmethode. Deze stelling is geldig voor een willekeurige hermitische matrix. De eigenwaarden van A hoeven dus niet aan vergelijking (3.2) te voldoen voor deze stelling om toepasbaar te zijn. Uit deze stelling volgt dat als het algoritme van de machtsmethode een benaderd eigenpaar θ, y geeft, dat dan $|\lambda - \theta| < tol$ geldt. De waarde van tol is dus in ieder geval een bovengrens op het verschil tussen θ en en eigenwaarde van A.

3.5 Numerieke voorbeelden

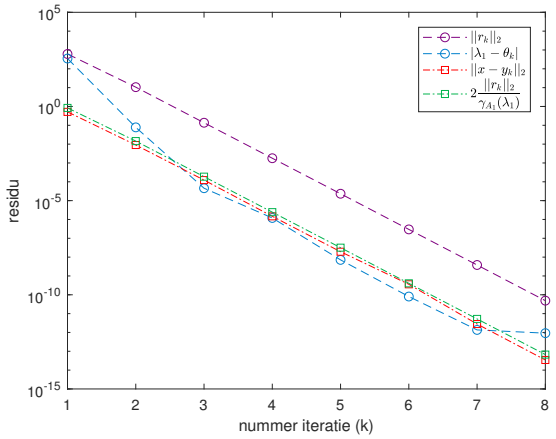
In deze paragraaf bekijken we een aantal numerieke voorbeelden van de machtsmethode. Alle numerieke voorbeelden zijn uitgevoerd met Matlab versie R2017b (9.3.0.713579). In het algemeen zullen we bij numerieke voorbeelden de eigenvectoren en eigenwaarden, bepaald met de functie eig uit Matlab, als de echte eigenwaarden en eigenvectoren van de matrix beschouwen. Bij alle voorbeelden in deze paragraaf is de startvector van de machtsmethode dezelfde vector en we noteren deze met $x_0 \in \mathbb{R}^{2000}$. De elementen van x_0 zijn bepaald uit uniforme verdeling op $(0, 1)$. In het vervolg zullen we een vector of een matrix met een dergelijke elementen aanduiden als een uniforme vector of matrix. Verder geldt er bij alle voorbeelden $tol = 10^{-10}$. Met $v(i)$ zullen we de i^{de} component van een vector aanduiden. We beginnen met de voorbeelden 3.6 en 3.7 waarvan de eigenwaarden voldoen aan vergelijking 3.2.

Voorbeeld 3.6. In dit voorbeeld bekijken de uniforme matrix $A_1 \in \mathbb{C}^{2000\times 2000}$, die niet hermitisch is. In figuur 1a is de waarde van θ_k op elke iteratie te zien. We zien hier dat de machtsmethode al na 8 iteraties convergeert. Deze snelle convergentie was ook te verwachten omdat er voor deze matrix geldt $\frac{\lambda_2(A)}{\lambda_1(A)} \simeq 0.01$. In figuur 1b zien we dat voor alle iteraties geldt $|\lambda_1 - \theta_k| < ||r_k||_2$ en $||x - y_k||_2 < 2\frac{||r_k||_2}{\gamma_A(\lambda_1)}$. Hierbij is x gelijk aan $\frac{|v_1(1)|y_k(1)}{v_1(1)|y_k(1)|}v_1$ met v_1 een genormeerde eigenvector behorend bij λ_1. Immers alle genormeerde eigenvectoren behorend bij λ_1 worden gegeven door $\{\omega v_1 | \omega \in S^1 \subset \mathbb{C}\}$ en x is nu die eigenvector in dezelfde richting als y_k. Ondanks dat Stelling 3.5 alleen geeft dat de machtsmethode voor hermitische matrices aan vergelijking (3.12) voldoet, blijkt dit ook voor A_1 te gelden. Verder zien we dat $||r_k||_2$ altijd een bovengrens vormt op de globale fout $|\lambda_1 - \theta_k|$ ondanks dat $||r_k||_2$ alleen een maat is voor de lokale fout. Echter door het grote verschil tussen λ_1 en λ_2 is de lokale fout voor A_1 al direct gelijk aan de globale fout. △

Voorbeeld 3.7. In dit voorbeeld bekijken een hermitische matrix $A_2 \in \mathbb{C}^{2000\times 2000}$ gevormd door $A_2 = \frac{Q^H DQ + (Q^H DQ)^H}{2}$ met Q een willekeurige unitaire matrix en D een diagonaalmatrix waarvan de elementen komen uit een normale verdeling met verwachtingswaarde $\mu = 0$ en standaarddeviatie $\sigma = 20$. In het vervolg zullen we dit aanduiden met een Gauss (μ, σ) diagonaalmatrix D. (De normale verdeling wordt ook wel de Gauss-verdeling genoemd.) In figuur 1a is de waarde van θ_k voor alle iteraties te zien. We zien dat de imaginaire component van de eigenwaarde naar 0 convergeert. Dit was ook te verwachten omdat A hermitisch is en dus wegens Stelling 2.9 alle eigenwaarden van A reëel zijn. Convergentie van de machtsmethode voor dezelfde waarde van tol kost nu veel meer iteraties dan in voorbeeld 3.6. Dit komt omdat er hier geldt $\frac{\lambda_2(A)}{\lambda_1(A)} \simeq 0.91$ terwijl dit in het vorige voorbeeld ongeveer 0.01 was en voor een hermitische matrix is dit een maat voor de convergentiesnelheid van de machtsmethode. Figuur 1b laat zien dat er al na een paar iteraties

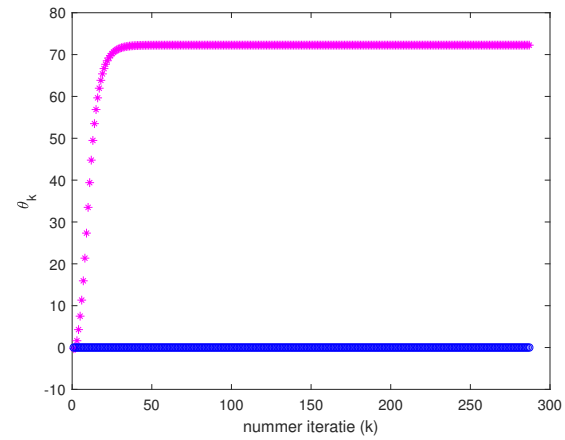


(a) Het reële deel (*) en imaginaire deel (∘) van θ_k weergegeven per iteratie.

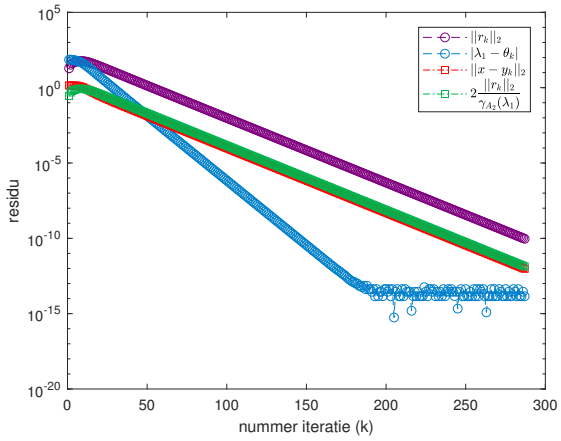


(b) De waarde van $||r_k||_2, |\lambda_1 - \theta_k|, ||x - y_k||_2, 2\frac{||r_k||_2}{\gamma_{A_1}(\lambda_1)}$ zijn weergegeven per iteratie.

Figuur 1: In deze figuren zien we voor elke iteratie van de machtsmethode, uitgevoerd op de matrix A_1, de waarde van $\theta_k, ||r_k||_2, |\lambda_1 - \theta_k|, ||x - y_k||_2, 2\frac{||r_k||_2}{\gamma_A(\lambda_1)}$. We zien hierbij dat voor alle iteraties geldt $|\lambda_1 - \theta_k| < ||r_k||_2$ en $||x - y_k||_2 < 2\frac{||r_k||_2}{\gamma_{A_1}(\lambda_1)}$.



(a) Het reële deel (*) en imaginaire deel (∘) van θ_k weergegeven per iteratie.



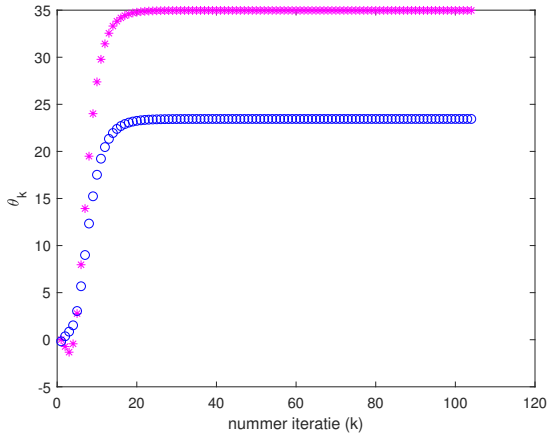
(b) De waarde van $||r_k||_2, |\lambda_1 - \theta_k|, ||x - y_k||_2, 2\frac{||r_k||_2}{\gamma_{A_2}(\lambda_1)}$ zijn weergegeven per iteratie.

Figuur 2: In deze figuren zien we voor elke iteratie van de machtsmethode, uitgevoerd op A_2, de waarde van $\theta_k, ||r_k||_2, |\lambda_1 - \theta_k|, ||x - y_k||_2, 2\frac{||r_k||_2}{\gamma_A(\lambda_1)}$. We zien hierbij dat al na een paar iteraties geldt $|\lambda_1 - \theta_k| < ||r_k||_2$ en $||x - y_k||_2 < 2\frac{||r_k||_2}{\gamma_{A_2}(\lambda_1)}$.

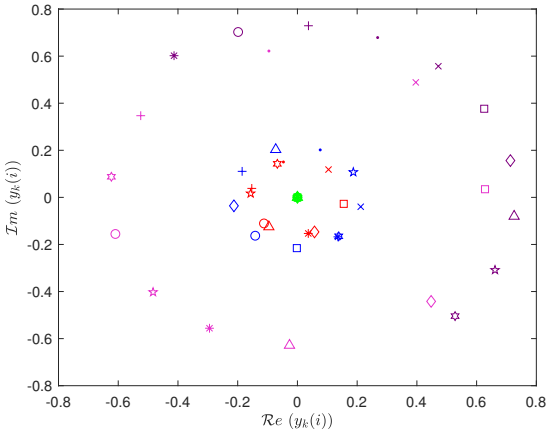
geldt $|\lambda_1 - \theta_k| < ||r_k||_2$ en $||x - y_k||_2 < 2\frac{||r_k||_2}{\gamma_A(\lambda_1)}$. Hierbij is x op dezelfde manier bepaald als in voorbeeld 3.6. Uit Stelling 3.5 volgt er nu dat al na een paar iteraties de globale fout gelijk is aan de lokale fout. Dat $|\lambda_1 - \theta_k|$ rond iteratie 200 sterk begint te schommelen komt doordat de machine precisie bereikt is. Opvallend in deze grafiek is dat de afschatting van $||x - y_k||_2$ veel scherper is dan die van $|\lambda_1 - \theta_k|$. △

We bekijken nu een normale matrix met eigenwaarden die voldoen aan vergelijking (3.6) om te zien of inderdaad θ_k convergeert en y_k het gedrag uit (3.7) gaat vertonen. Omdat we hebben afgeleid dat de machtsmethode uit algoritme 1 nu niet zal convergeren, vervangen we regel 6 uit dit algoritme door **while**$||\theta_k - \theta_{k-1}||_2 \geq tol$. Om duidelijk te kunnen zien of y_k zich volgens vergelijking (3.7) gaat gedragen bekijken we een

diagonaalmatrix. Immers voor een diagonaalmatrix zijn de eigenvectoren gelijk aan de standaard basis van $\mathbb{C}^{2000\times 2000}$ en de eigenwaarden gelijk aan de diagonaalelementen.



(a) Het reële deel (*) en imaginaire deel (∘) van θ_k weergegeven per iteratie.



(b) De vectorcomponenten $y_k(1)$, $y_k(2)$, $y_k(3)$, $y_k(4)$ en $y_k(5)$ zijn weergegeven voor $k \in \{conv+1\ (*), conv+2\ (\circ), conv+3\ (+), conv+4\ (.), conv+5\ (\times), conv+6\ (\square), conv+7\ (\diamond), conv+8\ (\triangle), conv+9\ (\star), \text{en } conv+10\ (\hexagram)\}$ in het complexe vlak. Hierbij is *conv* het aantal iteraties om convergentie voor θ te bereiken.

Figuur 3: In deze figuren zien is voor elke iteratie van de machtsmethode voor B de waarde van θ_k te zien. Ook zien we dat de componenten van de eigenvectoren met de eigenwaarde met gelijke grootste modulus inderdaad over een cirkel bewegen.

Voorbeeld 3.8. We bekijken een diagonaalmatrix $D \in \mathbb{C}^{2000\times 2000}$. De onderste 1996 diagonaalelementen van D zijn complexe waarden waarvan zowel het reële als het imaginaire deel uit een Gauss-verdeling met $\mu = 0$ en $\sigma = 20$ komt. De bovenste vier diagonaalelementen zijn complexe waarden met gelijke modulus, die groter is die van de andere 1996 diagonaalelementen. Figuur 3a toont dat θ_k inderdaad convergeert. In figuur 3b zijn de eerste vijf componenten van y_k te zien voor tien extra iteraties van de machtsmethode nadat θ_k is geconvergeerd. De punten van gelijke kleur zijn dezelfde componenten gedurende de verschillende iteraties. We zien hierbij dat $y_k(5)$ constant gelijk aan $0+0*i$ is. Dit was ook voorspeld omdat $y_k(5)$ de component van v_5 is in y_k. Wegens $\left|\frac{\lambda_1}{\lambda_5}\right| < 1$ dooft deze component inderdaad uit als θ_k convergeert. De overige weergegeven componenten van y_k zijn gelijk aan de component v_1, v_2, v_3 en v_4 in y_k. Omdat deze vier vectoren de eigenvectoren zijn behoren bij de eigenwaarden met de grootste modulus verwachtten we al uit vergelijking 3.7 dat deze over een cirkel bewegen. Dit zien we ook zien in figuur 3b. Het verschil in straal van de cirkels komt voort uit het verschil in waarden van $\frac{x_0(i)}{\sqrt{\Sigma_{l=1}^{4}|x_0(l)|^2}}$. Dat de cirkels met verschillende snelheden doorlopen worden, komt doordat de snelheid bepaald word door $\left(\frac{\lambda_i}{|\lambda_1|}\right)$. △

We zien nu dat de machtsmethode inderdaad het verwachte gedrag vertoont.

4 Inverse iteratie

In het vorige hoofdstuk hebben we gezien dat met de machtsmethode we het eigenpaar, waarvan de eigenwaarde de grootste modulus heeft, van een matrix A kunnen bepalen, indien er één eigenwaarde is met de grootste modulus. Wellicht zijn we niet geïntereseerd in de eigenwaarde met de grootste modulus van een matrix, maar willen we weten wat de eigenwaarde is die het dichtst bij een waarde $\sigma \in \mathbb{C}$ ligt. Dit kunnen we bepalen door de machtsmethode op een iets andere matrix toe te passen dan op A zelf. We krijgen wederom een benadering voor één eigenpaar. Deze methode heet de inverse iteratie. In paragraaf 4.1 zullen we eerst beschrijven hoe de inverse iteratie werkt, waarna we het algoritme bekijken in paragraaf 4.2. Vervolgens bekijken we in paragraaf 4.3 de convergentie van deze methode en zijn nauwkeurigheid. We besluiten dit hoofdstuk met een paar numerieke voorbeelden in paragraaf 4.4.

4.1 Werking inverse iteratie

De inverse iteratie werkt door de machtsmethode toe te passen op een matrix B waarvan we uit de eigenwaarde met de grootste modulus ν_1, de eigenwaarde λ_1 van A die het dichtst bij $\sigma \in \mathbb{C}$ ligt kunnen bepalen. Met de eigenwaarde die het dichtst bij σ ligt, bedoelen we $\lambda_1 = \min\{|\lambda_i - \sigma| | \lambda_i \in \Lambda(A)\}$, oftewel we meten de afstand tussen σ en een eigenwaarde met de modulus van het verschil. We nemen aan dat de eigenwaarden van A aan de volgende vergelijking voldoen:

$$|\lambda_1 - \sigma| \leq |\lambda_2 - \sigma| \leq \cdots \leq |\lambda_n - \sigma|. \tag{4.1}$$

Stel nu dat A diagonaliseerbaar is. Dan heeft A dus een eigendecompositie, oftewel er bestaan een diagonaalmatrix $D \in \mathbb{C}^{n\times n}$ met de eigenwaarden en een inverteerbare matrix $S \in \mathbb{C}^{n\times n}$ met de eigenvectoren als kolommen zodat er geldt $A = SDS^{-1}$. Aangezien we op zoek zijn naar de eigenwaarde λ_1 van A waarvoor geldt dat $|\lambda_1 - \sigma|$ minimaal is, is $\frac{1}{|\lambda_1 - \sigma|}$ maximaal. Om λ_1 te bepalen zullen we de volgende strategie hanteren.

- Construeer een matrix $B \in \mathbb{C}^{n\times n}$ met de eigenwaarden $\nu_i = \frac{1}{\lambda_i - \sigma}$ voor $\lambda_i \in \Lambda(A)$.
- Bepaal met de machtsmethode de eigenwaarde ν_1 met de grootste modulus van B en de bijbehorende eigenvector w.
- Bepaal uit ν_1 en w, λ_1 en v.

We willen nu dus een matrix B construeren met als eigenwaarden $\frac{1}{\lambda_i - \sigma}$ voor $\lambda_i \in \Lambda(A)$. Zolang er geldt dat $\lambda_i - \sigma \neq 0$ voor alle λ_i, kunnen we de diagonaalmatrix met als elementen $\frac{1}{\lambda_i - \sigma}$ creëren. Deze matrix is dan gelijk aan $(D - \sigma\mathbb{I})^{-1}$. Nu voldoet in principe elke matrix van de vorm $W(D - \sigma\mathbb{I})^{-1}W^{-1}$, met $W \in \mathbb{C}^{n\times n}$ inverteerbaar, aan de eis die we hierboven hebben gesteld op B. We gebruiken de volgende matrix voor B:

$$B = S(D - \sigma\mathbb{I})^{-1}S^{-1} = \left(S(D - \sigma\mathbb{I})S^{-1}\right)^{-1} = (A - \sigma\mathbb{I})^{-1}. \tag{4.2}$$

Er volgt nu direct dat de eigenvectoren van B ook gegeven worden door de kolommen van S. Oftewel de eigenvectoren van A en B zijn gelijk, wat ook de reden is voor onze keuze van B. Er volgt dus dat er geldt $v = w$ en $\lambda_1 = \sigma + \frac{1}{\nu_1}$. Echter we gebruiken de machtsmethode om ν_1 en w te bepalen. Dus de zojuist besproken methode kan alleen convergeren als er geldt

$$|\lambda_1 - \sigma| < |\lambda_2 - \sigma| \leq \cdots \leq |\lambda_n - \sigma|, \tag{4.3}$$

omdat anders de machtsmethode uitgevoerd op B niet convergeert naar ν_1. Dus indien $A - \sigma\mathbb{I}$ inverteerbaar is en de eigenwaarden van A voldoen aan vergelijking (4.3), hebben we nu een methode om λ_1 en v te bepalen. Indien $A - \sigma\mathbb{I}$ niet inverteerbaar is σ zelf een eigenwaarde van A. Dan hebben we dus wel de eigenwaarde het dichtst bij σ, namelijk σ zelf bepaald, echter geeft deze methode niet hoe we dan de bijbehorende eigenvector kunnen bepalen. Om toch de eigenvector behorend bij σ te bepalen kunnen we dan σ vervangen door $\tilde{\sigma}$, dat vlak bij σ in het complexe vlak ligt, zodat $A - \tilde{\sigma}\mathbb{I}$ wel inverteerbaar is en de methode naar σ convergeert.

1 **def** *inverseit*(A, x_0, σ, tol)**:**
  Input : -$A \in \mathbb{C}^{n\times n}$ een matrix.
  - $\sigma \in \mathbb{C}$ een complex getal.
  - $x_0 \in \mathbb{C}^n$ de start vector van de iteratie.
  - *tol* de gewenste nauwkeurigheid.
  Output: -(θ, y) de benadering van het eigenpaar van A voor de eigenwaarde het dichtst bij σ.
2 $w = x_0$
3 $y = \frac{w}{||w||_2}$
4 $w = (A - \sigma\mathbb{I})^{-1} y$
5 $\mu = y^H w$
6 **while** $||w - \mu y||_2 \geq tol$**:**
7 $y = \frac{w}{||w||_2}$
8 $w = (A - \sigma\mathbb{I})^{-1} y$
9 $\mu = y^H w$
10 **end**
11 $\theta = \sigma + \frac{1}{\mu}$

Algoritme 2: Inverse iteratie

4.2 Algoritme inverse iteratie

In de vorige paragraaf hebben we besproken hoe de inverse iteratie werkt. De inverse iteratie is weergegeven in algoritme 2.
In regel 4 en 8 wordt $(A - \sigma\mathbb{I})^{-1}$ gebruikt. Het is echter niet noodzakelijk om $(A - \sigma\mathbb{I})^{-1}$ ook daadwerkelijk te bepalen. Meestal wordt hiervoor de lineaire vergelijking $(A - \sigma\mathbb{I})\, w = y$ voor w opgelost. Ook zien we dat in regel 6 het convergentiecriterium bepaald hoe klein het residu van het benaderde eigenpaar (μ, y) van B moet zijn en niet het residu van (θ, y) van A.

4.3 Convergentie & Nauwkeurigheid

We bestuderen nu hoe snel de inverse iteratie convergeert en wat de nauwkeurigheid van de benadering is. Aangezien de inverse iteratie niets anders is dan toepassing van de machtsmethode op $(A - \sigma\mathbb{I})^{-1}$ volgt de convergentie van de inverse iteratie uit de toepassing van Stelling 3.2 op de matrix $(A - \sigma\mathbb{I})^{-1}$.

Stelling 4.1 (Convergentie inverse iteratie)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en de eigenwaarden van A voldoen aan vergelijking* (4.3) *voor een zekere $\sigma \in \mathbb{C}$. Laat verder $(A - \sigma\mathbb{I})$ inverteerbaar zijn en v_1 de genormeerde eigenvector zijn behorend bij de eigenwaarde λ_1 van A. Neem verder y_k de benadering van de eigenvector behorend bij λ_1 na k iteraties van de inverse iteratie. Dan geldt er dat :*

$$\sin(v_1, y_k) \leq \frac{1}{\epsilon|\alpha_1|}\rho_\epsilon((A - \sigma\mathbb{I})^{-1}(\mathbb{I} - \mathcal{P}))^k |\lambda_1 - \sigma|^{k-1}. \tag{4.4}$$

Hierbij is ρ_ϵ de pseudo-spectrale radius. Deze vergelijking geldt voor alle $\epsilon > 0$. Verder is $\mathcal{P}$ de spectrale projector behorend bij eigenwaarde ν_1 van de matrix $(A - \sigma\mathbb{I})^{-1}$. Tot slot is α_1 gegeven door $\mathcal{P}y_0 = \alpha_1 v_1$.
In het geval dat $(A - \sigma\mathbb{I})^{-1}$ diagonaliseerbaar is vereenvoudigt de grens op $\sin(v_1, y_k)$ tot

$$\sin(v_1, y_k) \leq \frac{||V||_2||V^{-1}||_2}{|\alpha_1|}\left|\frac{\lambda_1 - \sigma}{\lambda_2 - \sigma}\right|^{k-1}, \tag{4.5}$$

met V de matrix met de basis van eigenvectoren van A.

We hebben hierbij gebruikt dat de eigenvectoren van $(A - \sigma\mathbb{I})^{-1}$ en A gelijk zijn. Aangezien de eigenvectoren van deze matrices hetzelfde zijn, geldt er ook dat A diagonaliseerbaar is dan en slechts dan als $(A - \sigma\mathbb{I})^{-1}$ diagonaliseerbaar is. Er geldt nu dus dat voor een diagonaliseerbare matrix A er geldt dat de convergentiesnelheid

van de inverse iteratie bepaald wordt door $\left|\frac{\lambda_1-\sigma}{\lambda_2-\sigma}\right|$. Voor een algemene matrix wordt de convergentiesnelheid bepaald door $\rho_\epsilon((A-\sigma\mathbb{I})^{-1}(\mathbb{I}-\mathcal{P}))|\lambda_1-\sigma|$. We weten dat er voor $\epsilon \to 0$ geldt $\rho_\epsilon((A-\sigma\mathbb{I})^{-1}(\mathbb{I}-\mathcal{P})) \to \frac{1}{\lambda_2-\sigma}$ hieruit volgt echter niet noodzakelijk dat de inverse iteratie voor een willekeurige matrix convergeert, omdat het mogelijk is dat er $\rho_\epsilon((A-\sigma\mathbb{I})^{-1}(\mathbb{I}-\mathcal{P}))|\lambda_1-\sigma| > 1$ geldt.

Uit toepassing van stelling 3.3 op de matrix $(A-\sigma\mathbb{I})^{-1}$ volgt een afschatting voor $|\nu_1-\mu_k|$ afschatten. Hierbij is ν_1 de grootste eigenwaarde qua modulus van $(A-\sigma\mathbb{I})^{-1}$ en $\mu_k = y_k^H(A-\sigma\mathbb{I})^{-1}y_k$. Echter zijn we nu geïnteresseerd in de grens op $|\lambda_1-\theta_k|$ met $\theta_k = \sigma + \frac{1}{\mu_k}$ $\lambda_1 = \sigma + \frac{1}{\nu_1}$. Er volgt nu:

$$|\lambda_1-\theta_k| = |\frac{\mu_k-\nu_1}{\mu_k\nu_1}| = \frac{|\mu_k-\nu_1|}{|\mu_k|}|(\lambda_1-\sigma)|. \tag{4.6}$$

Ondanks dat $|(\lambda_1-\sigma)|$ niet bekend is, kunnen we nu wel het verloop van $|\lambda_1-\theta_k|$ gedurende de iteraties bepalen omdat onze onbekende een constante is. Het combineren van stelling 3.3 en vergelijking (4.6) geeft nu:

Stelling 4.2 (Nauwkeurigheid inverse iteratie)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn waarvan de eigenwaarden voldoen aan* (4.3)*, en laat v_1 de eigenvector zijn behorend bij λ_1. Neem (μ_k, y_k) de benadering van het eigenpaar met de grootste modulus van $(A-\sigma\mathbb{I})^{-1}$, (ν_1, v_1), na k iteraties van de inverse iteratie en $\theta_k = \sigma + \frac{1}{\mu_k}$. Dan geldt er*

$$|\lambda_1-\theta_k| \le \frac{|\lambda_1-\sigma|}{|\mu_k|}\frac{\sin(v_1,y_k)||A||_2}{\sqrt{1-\sin^2(v_1,y_k)}}. \tag{4.7}$$

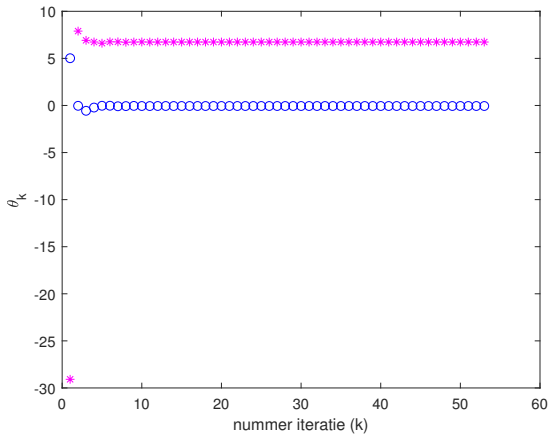
Als A een diagonaliseerbare matrix is volgt er nu uit stelling 4.1 dat $|\lambda_1-\theta_k| \to 0$ als $k \to \infty$ zolang $|\mu_k|$ niet naar 0 convergeert. Verder volgt er dat algemeen geldt, dat des te groter $|\mu_k|$ is des te nauwkeuriger de benadering van λ_1 met θ_k wordt.

4.4 Numerieke voorbeelden

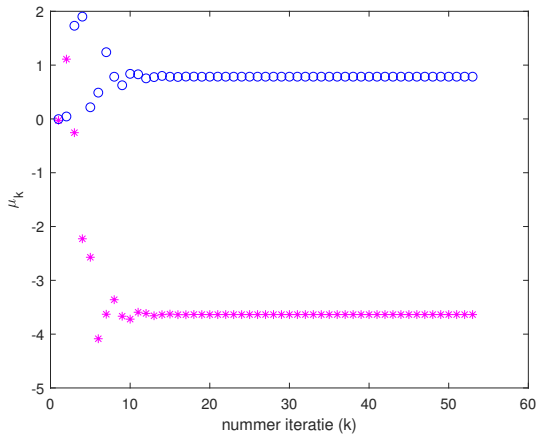
We bekijken nu een aantal numerieke voorbeelden van de inverse iteratie. In onze implementatie wordt regel 4 en 8 uit algoritme 2 uitgevoerd als $L\backslash U\backslash Py$ met $[L,U,P] = lu(A-\sigma\mathbb{I})$. Hierbij is lu de LU-decompositie van een matrix uit Matlab en is \ een Matlab functie die gegeven A en x de lineaire vergelijking $x = Ay$ oplost voor y. Voor beide voorbeelden is de startvector van de iteratie een uniforme vector $w \in \mathbb{R}^{2000}$ en geldt er $tol = 10^{-10}$. In beide voorbeelden voldoen de eigenwaarden van de matrix aan vergelijking (4.3) voor de gekozen waarde van σ.

Voorbeeld 4.3. In dit voorbeeld bekijken opnieuw de matrix $A_1 \in \mathbb{C}^{2000\times2000}$ uit voorbeeld 3.6. We passen nu de inverse iteratie toe op A_1 voor $\sigma = 7$. In figuur 4a is de waarde van θ_k op elke iteratie weergegeven en in figuur 4b die van μ_k. Opvallend is dat de inverse iteratie 53 iteraties nodig heeft om te convergeren terwijl het bij de machtsmethode maar 8 iteraties duurde. Dit komt omdat de convergentiesnelheid van de inverse iteratie door $\left|\frac{\lambda_1-\sigma}{\lambda_2-\sigma}\right| \simeq 0.61$ wordt begrensd, terwijl de convergentiesnelheid bij de machtsmethode door 0.01 is begrensd. Verder zien we dat μ_k gedurende een aantal iteraties redelijk verspringt terwijl θ_k dit niet doet. Dit komt doordat $\theta_k = \sigma + \frac{1}{\mu_k}$ geldt, waardoor we het verschil in θ_k niet zien omdat op de schaal van de plot het verschil tussen $\frac{1}{2}$ en $\frac{1}{3}$ niet zichtbaar is. Dat θ_1 veraf ligt van de overige waarde van θ_k komt doordat μ_1 klein is waardoor $\frac{1}{\mu_1}$ groot wordt. In figuur 1b zien we dat al na een paar iteraties geldt $|\lambda_1-\theta_k| < ||r_k||_2$ en $|\nu_1-\mu_k| < ||r_k||_2$, met $||r_k||_2 = ||(A-\sigma\mathbb{I})^{-1}y_k - \mu_k y_k||_2$. De vector x is wederom bepaald als $\frac{|v_1(1)|y_k(1)}{v_1(1)|y_k(1)|}v_1$ met v_1 een genormeerde eigenvector behorend bij λ_1. We zien dat $\|r_k\|_2$ al na een paar iteraties groter is dan $|\nu_1-\mu_k|$ oftewel dat een bovengrens op de lokale fout ook al snel een bovengrens op de globale fout vormt. Verder zien we dat er vanaf iteratie 2 geldt dat $|\lambda_1-\theta_k| < |\nu_1-\mu_k| < ||r_k||_2$, dus het convergentiecriterium zegt nu ook iets over de fout in θ_k. Ook zien we dat de eigenvector convergeert. △

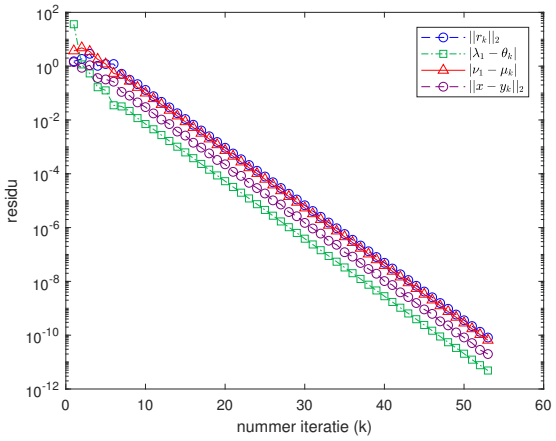
Voorbeeld 4.4. In dit voorbeeld bekijken opnieuw de hermitische matrix $A_2 \in \mathbb{C}^{2000\times2000}$ uit voorbeeld 3.7. We passen nu de inverse iteratie toe op A_2 met $\sigma = 11$. In figuur 4a is de waarde van θ_k op elke iteratie



(a) Het reële deel (*) en imaginaire deel (∘) van θ_k weergegeven per iteratie.



(b) Het reële deel (*) en imaginaire deel (∘) van μ_k weergegeven per iteratie.

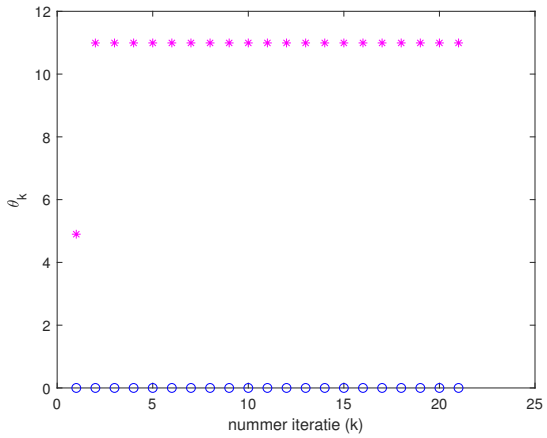


(c) De waarde van $||r_k||_2, |\lambda_1 - \theta_k|, |\nu_1 - \mu_k|, ||x - y_k||_2$ zijn weergegeven per iteratie.

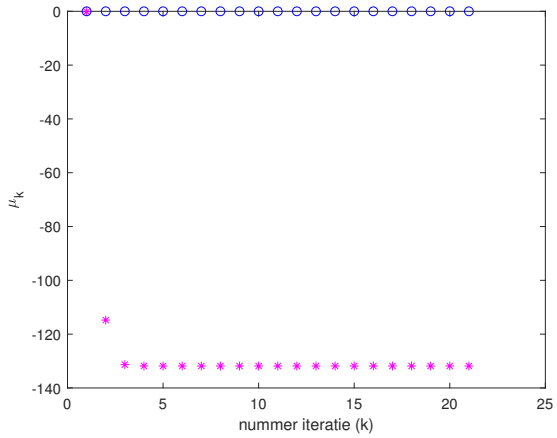
Figuur 4: In deze figuren zien voor elke iteratie van de inverse iteratie,uitgevoerd op de matrix A_1 met $\sigma = 7$, de waarde van $\theta_k, \mu_k, ||r_k||_2, |\lambda_1 - \theta_k|, |\nu_1 - \mu_k|$ en $||x - y_k||_2$.

weergegeven en in figuur 4b die van μ_k. Opvallend is dat de inverse iteratie maar 21 iteraties nodig heeft om te convergeren terwijl het bij de machtsmethode bijna 300 iteraties kostte. Dit komt omdat de convergentiesnelheid voor de inverse iteratie $\left|\frac{\lambda_1-\sigma}{\lambda_2-\sigma}\right| \simeq 0.23$ is, terwijl de convergentiesnelheid van de machtsmethode 0.91 was. Verder lijken θ_k en μ_k even snel te convergeren. Echter in figuur 1b zien we dat θ_k sneller convergeert dan μ_k. Dit komt doordat er geldt $\frac{|\lambda_1-\sigma|}{|\mu_k|} < 1$, zie vergelijking (4.6). Verder zien we dat vanaf iteratie 20 geldt $|\nu_1 - \mu_k| > ||r_k||_2 = ||\left(A_2 - \sigma\mathbb{I}\right)^{-1} y_k - \mu_k y_k||_2$, echter aangezien ν_1 dan toch de dichtstbijzijnde eigenwaarde van $\left(A_2 - \sigma\mathbb{I}\right)^{-1}$ voor μ_k is spreekt dit stelling 3.5 tegen (immers we voeren de machtsmethode uit op $\left(A_2 - \sigma\mathbb{I}\right)^{-1}$). Dit komt omdat rond iteratie 10 de waarde van μ_k, en daarmee ook die van θ_k, stagneert. Dit blijkt weer te komen omdat we niet exact de machtsmethode gebruiken maar lineaire vergelijkingen oplossen voor de LU-decompositie van $\left(A_2 - \sigma\mathbb{I}\right)$, immers bij directe toepassing van de machtsmethode op $\left(A_2 - \sigma\mathbb{I}\right)^{-1}$ treedt deze stagnatie niet op. Verder zien we in dit figuur $||x - y_k||_2$, met x op dezelfde manier bepaald als in voorbeeld 4.4, afnemen met ongeveer dezelfde snelheid als het residu. We zien ook dat $||r_k||_2$ hier groter is dan zowel $|\lambda_1 - \theta_k|$ als $||x - y_k||_2$ en dus convergentie van de inverse iteratie geeft hier ook een bovengrens op de globale fout van het benaderde eigenpaar. △

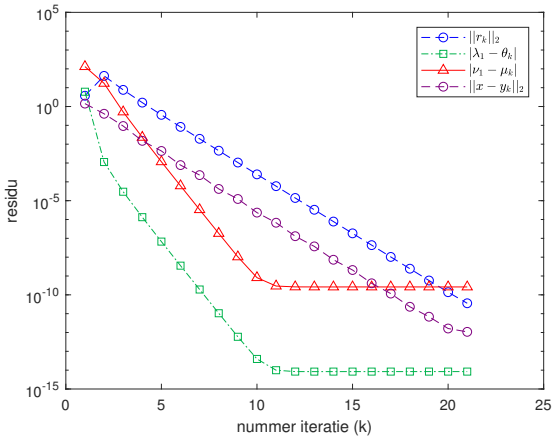
We zien dus dat convergentie van de inverse iteratie voor *tol* in onze voorbeelden een bovengrens op de fout



(a) Het reële deel (*) en imaginaire deel (∘) van θ_k weergegeven per iteratie.



(b) Het reële deel (*) en imaginaire deel (∘) van μ_k weergegeven per iteratie.



(c) De waarde van $||r_k||_2, |\lambda_1 - \theta_k|, |\nu_1 - \mu_k|, ||x - y_k||_2$ zijn weergegeven per iteratie.

Figuur 5: In deze figuren zien voor elke iteratie van de inverse iteratie,uitgevoerd op de matrix A_2 met $\sigma = 11$, de waarde van $\theta_k, \mu_k, ||r_k||_2, |\lambda_1 - \theta_k|, |\nu_1 - \mu_k|$ en $||x - y_k||_2$.

in het eigenpaar geeft.

5 Krylov deelruimte

In de afgelopen twee hoofdstukken ,3 en 4, hebben we twee methoden gezien die berusten op een reeks machten van de matrix A of $(A - \sigma\mathbb{I})^{-1}$ om een eigenpaar van A te benaderen. Hierbij bepalen we alle genormeerde vectoren van de vorm $v, Av, A^2v, \cdots A^kv$ voor een zekere $v \in \mathbb{C}^n$ totdat we $\frac{A^kv}{||A^kv||_2}$ als geconvergeerd naar de gewenste eigenvector beschouwen. Hierbij gebruiken we uiteindelijk alleen de informatie die in $\frac{A^kv}{||A^kv||_2}$ of $\frac{(A-\sigma\mathbb{I})^{-k}v}{||(A-\sigma\mathbb{I})^{-k}v||_2}$ zit. Nu rijst de vraag of we uit de verzameling van vectoren $v, Av, A^2v, \cdots A^kv$ niet meer informatie kunnen halen over de eigenparen van A dan uit alleen A^kv. Het idee dat dit mogelijk moet zijn wordt ondersteund door de volgende analyse. Stel dat we een matrix $B \in \mathbb{C}^{n\times n}$ hebben, B kan gelijk zijn aan A of $(A - \sigma\mathbb{I})^{-1}$, met de lineair onafhankelijke eigenvectoren $s_1, s_2, \cdots, s_k$. Aangezien we niet eisen dat B diagonaliseerbaar is, geldt er $k \leq n$. Als we nu een vector $v \in \mathbb{C}^n$ hebben zodat er geldt $v = \sum_{i=1}^{k} \alpha_i s_i$ voor $\alpha_i \in \mathbb{C}$ dan is elke vector van de vorm B^mv voor $m \in \mathbb{N}$ ook een lineaire combinatie van dezelfde eigenvectoren

als v. Het idee is nu dat met behulp van de lineaire deelruimte, opgespannen door $v, Bv, \cdots, B^k v$, meer eigenwaarden en eigenvectoren van B benaderd kunnen worden. Immers deze deelruimte bestaat dan uit lineaire combinaties van eigenvectoren. Deze lineaire deelruimte wordt een Krylov deelruimte genoemd. In paragraaf 5.1 zullen we een aantal eigenschappen van de Krylov deelruimten bespreken. De meeste eigenschappen die wij zullen afleiden zijn ook terug te vinden in [1, hoofdstuk 6] of hebben we hieruit overgenomen. Daarna zullen we in paragraaf 5.2 de Rayleigh-Ritz procedure bespreken. Deze methode zullen we gebruiken om uit een Krylov deelruimte benaderingen voor de eigenwaarden en eigenvectoren van A te bepalen. In de volgende hoofdstukken zullen we de Lanczos(hoofdstuk 6) en Arnoldi (hoofdstuk 7) methode bespreken. Deze methoden werken beide door toepassing van de Rayleigh-Ritz procedure op een Krylov deelruimte.

5.1 Eigenschappen Krylov deelruimten

Voordat we de eigenschappen van een Krylov deelruimte uitlichten, zullen we deze eerst definiëren.

Definitie 5.1 (Krylov deelruimte)**.** Neem een matrix $A \in \mathbb{C}^{n\times n}$ en een vector $v \in \mathbb{C}^n$. We definiëren nu de Krylov deelruimte $\mathcal{K}_m(A, v)$ voor $m \in \mathbb{N}$ als:

$$\mathcal{K}_m(A, v) = \text{span}\{v, Av, \cdots, A^{m-1}v\}. \tag{5.1}$$

Uit de definitie van de Krylov deelruimte $\mathcal{K}_m(A, v)$ volgt er nog een mogelijke karakterisatie van een Krylov deelruimte. We zien immers dat de Krylov deelruimte wordt opgespannen door alle vectoren $A^i v$ voor $i \in \{0, 1, \cdots m-1\}$. Oftewel elk element van de Krylov deelruimte kan geschreven worden als $\sum_{i=0}^{m-1} a_i A^i v = \left(\sum_{i=0}^{m-1} a_i A^i\right) v$ voor $a_i \in \mathbb{C}$. Deze sommatie is een polynoom van graad $m-1$ of lager geëvalueerd in A vermenigvuldigd met v. Aangezien alle mogelijke combinaties van waarden voor a_i toegestaan zijn, zien we dat er ook geldt:

$$\mathcal{K}_m(A, v) = \{p(A)v \mid p \text{ is een polynoom in } A \text{ over } \mathbb{C} \text{ van graad } m-1 \text{ of lager}\}. \tag{5.2}$$

Voor elke vector $v \in \mathbb{C}^n$ bestaat er een monisch polynoom q, oftewel een polynoom waarvan de coëfficiënt voor hoogste macht gelijk is aan 1, met minimale graad zodat er geldt dat $q(A)v = 0$, waarbij q niet het triviale polynoom is. Dit polynoom q heet het minimaal polynoom van v in A. De graad van dit polynoom q wordt ook wel de graad van v in A genoemd. Als het duidelijk is om welke matrix en vector het gaat wordt er vaak gesproken over het minimale polynoom van A en de graad van v. Stel dat α de graad is van v en q het minimaal polynoom. Neem nu vector $r(A)v$ waarbij r een willekeurig polynoom is. Dan verwachten we dat er geldt $r(A)v = p(A)v$ waarbij p een polynoom is van graad $\alpha - 1$ of lager. Dit vermoeden leidt tot de volgende propositie:

Propositie 5.2. *(Invariantie Krylov deelruimten, [1, proposite 6.2]) Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en $v \in \mathbb{C}^n$ een vector met α de graad van v. Dan geldt er dat $A \cdot \mathcal{K}_\alpha(A, v) \subset \mathcal{K}_\alpha(A, v)$, oftewel $\mathcal{K}_\alpha(A, v)$ is een invariante deelruimte van A. Er geldt zelfs dat $\mathcal{K}_\alpha(A, v) = \mathcal{K}_\beta(A, v)$ voor alle $\beta \geq \alpha$.*

Bewijs. We beginnen met het aantonen van $A \cdot \mathcal{K}_\alpha(A, v) \subset \mathcal{K}_\alpha(A, v)$. Neem een element $w \in A \cdot \mathcal{K}_\alpha(A, v)$. Uit vergelijking (5.2) volgt er dan $w = \sum_{i=1}^{\alpha} a_i A^i v$ met $a_i \in \mathbb{C}$. Echter was α de graad van v en dus bestaat het minimale polynoom $q(A)$ waarvoor geldt $q(A)v = \sum_{i=0}^{\alpha} b_i A^i v = 0$ voor $b_i \in \mathbb{C}$ met $b_\alpha = 1$. Er zijn nu twee opties voor w ofwel $a_\alpha = 0$, waaruit volgt dat $w \in \mathcal{K}_\alpha(A, v)$. Of er geldt $a_\alpha \neq 0$, maar dan volgt er $w = w - a_\alpha q(A)v = \sum_{i=1}^{\alpha} a_i A^i v - a_\alpha \sum_{i=0}^{\alpha} b_i A^i v = \sum_{i=0}^{\alpha-1} (a_i - a_\alpha b_i) A^i v$. Hierbij valt de term met A^α weg omdat er geldt $a_\alpha - a_\alpha b_\alpha = 0$. Dus nu is w opnieuw gerepresenteerd door een polynoom van graad $\alpha - 1$ of lager en dus geldt er $w \in \mathcal{K}_\alpha(A, v)$. Er geldt dus inderdaad dat $A \cdot \mathcal{K}_\alpha(A, v) \subset \mathcal{K}_\alpha(A, v)$.

Nu rest er dus te bewijzen $\mathcal{K}_\alpha(A, v) = \mathcal{K}_\beta(A, v)$ voor alle $\beta \geq \alpha$. Er volgt uit de definitie van de Krylov deelruimte direct dat er geldt $\mathcal{K}_\alpha(A, v) \subset \mathcal{K}_\beta(A, v)$. Het volstaat dus om te bewijzen $\mathcal{K}_\beta(A, v) \subset \mathcal{K}_\alpha(A, v)$. Hiertoe zullen we bewijs met volledige inductie gebruiken. Schrijf $\beta = \alpha + n - 1$ voor $n \in \mathbb{N}$ en pas nu inductie toe op n.
Inductie stap: Voor $n = 1$ volgt er $\beta = \alpha$ en volgt er dus direct dat er geldt $\mathcal{K}_\beta(A, v) \subset \mathcal{K}_\alpha(A, v)$.
Inductie hypothese: Stel het is waar voor n, dus er geldt $\mathcal{K}_\beta(A, v) \subset \mathcal{K}_\alpha(A, v)$ voor $\beta = \alpha + n - 1$.

We nu bewijzen dat er voor $\beta' = \alpha + (n+1) - 1$ geldt $\mathcal{K}_{\beta'}(A,v) \subset \mathcal{K}_\alpha(A,v)$. Echter uit de inductie hypothese volgt er dat $\mathcal{K}_\beta(A,v) \subset \mathcal{K}_\alpha(A,v)$. Als we dus $\mathcal{K}_{\beta'}(A,v) \subset \mathcal{K}_\beta(A,v)$ aantonen volgt uit de inductie hypothese dat $\mathcal{K}_{\beta'}(A,v) \subset \mathcal{K}_\beta(A,v) \subset \mathcal{K}_\alpha(A,v)$ geldt en zijn we ook klaar. Neem nu een willekeurige element $w \in \mathcal{K}_{\beta'}(A,v)$. Dan bestaat er dus een polynoom $p(A)$ van graad $\beta' - 1$ of lager zodat er geldt $w = \sum_{i=0}^{\beta'-1} a_i A^i v$. Echter hebben we ook het minimaal polynoom $q(A)$ van v van graad α. Er volgt nu $w = w - \alpha_{\beta'-1} A^n q(A) v = \sum_{i=0}^{\beta'-1} a_i A^i v - a_{\beta'-1} A^{n-1} \sum_{i=0}^{\alpha} b_i A^i v = \sum_{i=0}^{\beta'-2} a_i A^i v - a_{\beta'-1} A^{n-1} \sum_{i=0}^{\alpha-1} b_i A^i v$. We hebben nu dus γ geschreven als van graad $\beta - 1$ of lager en dus geldt er $\gamma \in \mathcal{K}_\beta(A,v)$. We hebben nu dus met inductie bewezen dat er geldt $\mathcal{K}_\beta(A,v) \subset \mathcal{K}_\alpha(A,v)$ voor alle $\beta \geq \alpha$. □

We zien inderdaad dat de Krylov deelruimte opgespannen door A en v maximaal dimensie α kan hebben, met α de graad van v. Nu willen we graag instaat zijn de dimensie van de Krylov deelruimte $\mathcal{K}_m(A,v)$ te bepalen. Uit Propositie 5.2 volgt dat de dimensie van $\mathcal{K}_m(A,v)$ vaststaat zodra m de waarde van α bereikt. De volgende propositie geeft de dimensie van een Krylov deelruimte voor alle waarden van m:

Propositie 5.3 (Dimensie Krylov ruimte)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een matrix zijn en $v \in \mathbb{C}^n$ een vector met α de graad van het minimaal polynoom van v in A. Dan geldt er dat*

$$\dim\left(\mathcal{K}_m(A,v)\right) = \begin{cases} m & \text{voor } m \leq \alpha \\ \alpha & \text{voor } m > \alpha \end{cases}.$$

Bewijs. Dat de dimensie van $\mathcal{K}_m(A,v)$ gelijk is aan de dimensie van $\mathcal{K}_\alpha(A,v)$ voor $m > \alpha$ volgt uit Propositie 5.2. Immers, er geldt $\mathcal{K}_\alpha(A,v) = \mathcal{K}_m(A,v)$, en als twee vectorruimten gelijk zijn, moeten hun dimensies ook gelijk zijn. Dat voor $m \leq \alpha$ geldt $\dim\left(K_m(A,v)\right) = m$ volgt uit dat de graad van v gelijk is aan α. Dat de graad van v gelijk is aan α is namelijk equivalent aan dat $v, Av, \cdots A^{\alpha-1}v$ lineair onafhankelijke vectoren zijn. Aangezien $\mathcal{K}_m(A,v)$ wordt opgespannen door de eerste m vectoren van deze verzameling vormen ze dus een basis voor $K_m(A,v)$ voor $m \leq \alpha$, waaruit volgt dat geldt $\dim\left(K_m(A,v)\right) = m$. □

Een zelfde propositie voor $m - 1 < \alpha$ is te vinden in [1, propositie 6.3]. We weten nu wat de dimensie van de Krylov deelruimte $\mathcal{K}_m(A,v)$ is als de graad van v bekend is.

5.2 Rayleigh-Ritz procedure

Nu hebben we een Krylov deelruimte, echter hebben we nog een methode nodig om uit deze deelruimte ook een benadering voor de eigenwaarden en eigenvectoren te bepalen, wat wel ons uiteindelijke doel is. De methode die hiertoe wordt toegepast in de algoritmen, die we in de komende hoofdstukken behandelen, gebruiken we de Rayleigh-Ritz procedure. Deze procedure is een algemene methode om uit de projectie van de oorspronkelijke matrix op een deelruimte eigenparen te benaderen. Neem nu een matrix A waar we eigenparen van willen benaderen door projectie van A op de deelruimte $\mathcal{W}$. De Rayleigh-Ritz procedure is nu als volgt:

1. Bepaal een orthonormale basis $\{w_1, w_2, \cdots w_k\}$ van $\mathcal{W}$ en zet dit in een matrix $W = [w_1, w_2, \cdots, w_k]$.
2. Bepaal $B = W^H A W$
3. Bepaal de eigenwaarden $\mu_1, \mu_2 \cdots, \mu_k$ van B.
4. Bepaal de eigenvectoren y_i van B behorend bij μ_i en bepaal $x_i = W y_i$.

Het eigenpaar (μ_i, x_i) is nu een benadering van het eigenpaar (λ, v) van A waarvoor er geldt $\mu_i \simeq \lambda_i$. We noemen nu μ_i een Ritz waarde en $W y_i$ een Ritz vector. Als we een deelruimte van dimensie k hebben, kunnen we met de Rayleigh-Ritz procedure k eigenparen van A benaderen. Er geldt nu niet noodzakelijk dat elk Ritz paar het eigenpaar waar het het dichtstbij ligt even goed benaderd.

We hebben nu alleen een benadering van een eigenpaar gevonden. Maar we hebben nog geen indicatie van de nauwkeurigheid van dit eigenpaar. Voor een algemene deelruimte $\mathcal{W}$ kunnen we in ieder geval het verschil tussen de Ritz waarden en de echte eigenwaarden van de matrix afschatten. Deze afschatting is een combinatie van Theorem 4.1 en Corrollary 4.2 uit Jia en Stewart [10]. Voordat we de afschatting geven,

definiëren we eerst de setting waarin deze stelling geldt. Stel dat we een eigenruimte $\mathcal{X}$ van A hebben, dus een lineaire deelruimte waarvoor geldt $A \cdot \mathcal{X} \subset \mathcal{X}$ en dat we een basis $X \in \mathbb{C}^{n\times l}$ hebben van $\mathcal{X}$, met l de dimensie van $\mathcal{X}$. Dan bestaat er een unieke matrix $L \in \mathbb{C}^{l\times l}$ zodat er geldt $AX = XL$. Stel nu dat de eigenwaarden van L verschillen van alle overige eigenwaarden van A. Stel verder dat we een deelruimte $\mathcal{W}$ hebben waarvoor er voor de matrix $W^H AW \equiv B \in \mathbb{C}^{k\times k}$ geldt dat $\Lambda(L) \simeq \Lambda(B)$ met $k \geq l$. Dan geldt er het volgende:

Stelling 5.4. *(Algemene convergentie Ritz waarden) Laat $\lambda_1, \lambda_2, \cdots, \lambda_l$ de eigenwaarden van L zijn en $\mu_1, \mu_2, \cdots, \mu_k$ de eigenwaarden van B. Dan bestaan er gehele getallen $j_1, \cdots, j_l$ zodat er geldt*

$$|\lambda_i - \mu_{j_i}| \leq 4\|A\|_2 \left(2 + \frac{\epsilon}{\sqrt{1-\epsilon^2}}\right)^{1-\frac{1}{k}} \left(\frac{\epsilon}{\sqrt{1-\epsilon^2}}\right)^{\frac{1}{k}} \qquad \textit{voor} \quad i = 1, 2, \cdots, l. \tag{5.3}$$

Hierbij geldt dat $\epsilon = \sin(X, W) = \|W_\perp^H X\|_2$.

We kunnen nu deze Stelling gebruiken voor de convergentie van de eigenwaarden voor methoden in de volgende hoofdstukken als we een afschatting voor ϵ kunnen bepalen. Deze stelling is ook de basis voor de afschatting van het verschil tussen λ_1 en θ in Stelling 3.3 voor de machtsmethode.

Er blijkt echter dat het ook mogelijk is dat de Ritz paren echte eigenparen van de matrix zijn. Dit is het geval als $\mathcal{W}$ een invariante deelruimte van A is.

Propositie 5.5. *(Invariante deelruimte) Laat $\mathcal{W}$ een invariante deelruimte zijn van A en laat W een orhtonormale basis vormen voor $\mathcal{W}$. Dan is elk Ritz paar (μ, x) gelijk aan een (λ, v) van A.*

Bewijs. Laat (μ, y) een Ritz paar zijn van A. Dan geldt er dus $W^H AWy - \mu y = 0$. Als we nu gebruiken dat er geldt $x = Wy$ volgt dat $W^H Ax - W^H \mu y = W^H (Ax - \mu x) = 0$. Uit de definitie van x volgt dat $x \in \mathcal{W}$ en omdat $\mathcal{W}$ een invariante deelruimte is van A volgt hieruit dat er ook geldt $Ax \in \mathcal{W}$. Daaruit volgt dat $W^H Ax$ niets anders is dan een basistransformatie van Ax en dus geldt er dat $Ax - \mu x = 0$ waaruit volgt dat het Ritz paar (μ, x) inderdaad een echt eigenpaar is van A. □

Aangezien uit propositie 5.2 volgt dat $\mathcal{K}_\alpha(A, v)$ een invariante deelruimte is van A, met α de graad van v, volgt er nu dat we uit $\mathcal{K}_\alpha(A, v)$ α exacte eigenparen van A kunnen bepalen.

6 Lanczos methode

In dit hoofdstuk zullen we de Lanczos methode bespreken. De werking van deze methode berust op toepassing van de Rayleigh-Ritz procedure op een Krylov deelruimte. Aangezien de dimensie van een Krylov deelruimte groter kan zijn dan één, kunnen we met deze methode meerdere eigenparen benaderen. Echter werkt de Lanczos methode alleen voor hermitische matrices. Voor algemene matrices bespreken we in hoofdstuk 7 de Arnoldi methode, die op een vergelijkbare manier werkt als de Lanczos methode.

We beginnen dit hoofdstuk met een theoretische bepaling van een orthonormale basis van de Krylov deelruimte en de projectie van de matrix op deze deelruimte in paragraaf 6.1. We bepalen dus uitdrukkingen om de eerste twee stappen van de Rayleigh-Ritz procedure uit te voeren, zie paragraaf 5.2, Daarna bespreken we het algoritme voor de Lanczos methode in paragraaf 6.2, waarna we in paragraaf 6.3 de convergentie en nauwkeurigheid van dit algoritme bekijken. Echter blijkt de theoretisch orthonormale basis van de Krylov uit paragraaf 6.1 numeriek niet orthogonaal te zijn. Manieren om te voorkomen dat de convergentie van de Ritz paren hier onder lijdt, bespreken we in paragraaf 6.4. In paragraaf 6.5 bespreken we de spectrale transformatie. Dit behelst toepassing van de Lanczos methode op de matrix $(A - \sigma\mathbb{I})^{-1}$ om de Ritz paren eerst naar de eigenparen in een gewenst gebied te laten convergeren. Tot slot bekijken we in paragraaf 6.6 een aantal numerieke voorbeelden.

6.1 Orthonormale basis Krylov deelruimte & projectie van A op deze deelruimte

We willen een orthonormale basis van de Krylov deelruimte $\mathcal{K}_m(A, v)$ en de projectie van A op $\mathcal{K}_m(A, v)$ bepalen voor een hermitische matrix $A \in \mathbb{C}^{n\times n}$. Voordat we een orthonormale basis van Krylov deelruimte

bepalen, zullen we eerst laten zien dat de orthonormale basis van $\mathcal{K}_m(A,v)$ als onderdeel van de orthonormale basis van $\mathcal{K}_n(A,v)$ kan dienen voor $m < n$. We nemen hierbij aan dat de dimensie van $\mathcal{K}_n(A,v)$ gelijk is aan n. Hiertoe hebben we de volgende propositie nodig.

Propositie 6.1 (Recursieve basis)**.** *Laat $\{a_1, a_2, \cdots, a_k\}$ en $\{b_1, b_2, \cdots, b_l\}$ twee verzamelingen vectoren zijn waarvoor er geldt $a_i, b_i \in \mathbb{C}^n$ met $n, k, l \in \mathbb{N}$. Neem verder nog een vector $c \in \mathbb{C}^n$. Als er dan geldt:*

$$span\{a_1, a_2, \cdots, a_k\} = span\{b_1, b_2, \cdots, b_l\}, \tag{6.1}$$

geldt er ook

$$span\{a_1, a_2, \cdots, a_k, c\} = span\{b_1, b_2, \cdots, b_l, c\}. \tag{6.2}$$

Bewijs. Neem een willekeurige vector $v \in \text{span}\{a_1, a_2, \cdots, a_k, c\}$. Uit definitie 2.1 volgt er dat geldt $v = w + \mu c$ voor zekere $w \in \text{span}\{a_1, a_2, \cdots, a_k\}$ en $\mu \in \mathbb{C}$. Echter uit de aanname, vergelijking 6.1, volgt nu $w \in \text{span}\{b_1, b_2, \cdots, b_l\}$. Echter uit definitie 2.1 volgt dan $v = w + \mu c \in \text{span}\{b_1, b_2, \cdots, b_l, c\}$. Aangezien v een willekeurige vector is hebben we dus bewezen dat er geldt $\text{span}\{a_1, a_2, \cdots, a_k, c\} \subset \text{span}\{b_1, b_2, \cdots, b_l, c\}$. Uit hetzelfde argument met de rol van de a_i's en de b_i's omgewisseld volgt dat ook geldt $\text{span}\{a_1, a_2, \cdots, a_k, c\} \supset \text{span}\{b_1, b_2, \cdots, b_l, c\}$. We hebben dus bewezen dat uit vergelijking 6.1 inderdaad volgt dat $\text{span}\{a_1, a_2, \cdots, a_k, c\} = \text{span}\{b_1, b_2, \cdots, b_l, c\}$. □

Uit deze propositie in combinatie met definitie 5.1 volgt nu dat indien $\{q_1, q_2, \cdots q_m\}$ een basis vormt van $\mathcal{K}_m(A,v)$, dat $\{q_1, q_2, \cdots q_m, A^m v, \cdots A^{n-1} v\}$ dan een basis vormt voor $\mathcal{K}_n(A,v)$. De basis $\{q_1, q_2, \cdots q_m, A^m v, \cdots A^{n-1} v\}$ kan omgevormd worden tot een orthonormale basis door toepassing van een Gram-Schimdt procedure. Uit de werking van de Gram-Schimdt procedure, volgt er dat de vectoren $q_1, q_2, \cdots q_m$ hierbij ongewijzigd blijven [5]. Een orthonormale basis van $\mathcal{K}_m(A,v)$ kan dus inderdaad een onderdeel zijn van een orthonormale basis van $\mathcal{K}_n(A,v)$.

We zullen nu een stelling geven die uitdrukkingen voor een orthonormale basis van de Krylov deelruimte en de projectie van A op deze deelruimte geeft. Hierbij zullen we de volgende definitie gebruiken.

Definitie 6.2 (Hessenberg matrices)**.** Een matrix $A \in \mathbb{C}^{n \times n}$ is

- een beneden-hessenbergmatrix als er geldt $a_{ij} = 0$ voor alle $j > i + 1$.
- een boven-hessenbergmatrix als er geldt $a_{ij} = 0$ voor alle $i > j + 1$.

Een matrix die zowel benden-hessenberg als boven-hessenberg is wordt tridiagonaal genoemd.

De rest van deze paragraaf omvat het bewijs van de volgende stelling.

Stelling 6.3. *Laat $A \in \mathbb{C}^{n \times n}$ een hermitische matrix zijn. Neem een vector $v \in \mathbb{C}^n$ en noem de graad van v, γ. Voor een Krylov deelruimte $K_m(A,v)$ met $m \leq \gamma$ geldt er dat $\{q_1, q_2, \cdots q_m\}$ een orthonormale basis van deze deelruimte vormt met:*

$$q_1 = \frac{v}{||v||_2}, \tag{6.3}$$

$$q_2 = \frac{Aq_1 - \alpha_1 q_1}{\beta_1}, \tag{6.4}$$

$$q_{i+1} = \frac{Aq_i - \beta_{i-1} q_{i-1} - \alpha_i q_i}{\beta_i} \quad voor\ i \in \{2, 3, \cdots m-1\}, \tag{6.5}$$

waarbij er geldt

$$\alpha_i = q_i^H A q_i, \qquad \beta_1 = ||Aq_1 - \alpha_1 q_1||_2, \qquad \beta_i = ||Aq_i - \beta_{i-1} q_{i-1} - \alpha_i q_i||_2. \tag{6.6}$$

Definieer verder de matrix $Q_m \in \mathbb{C}^{n\times m}$ als $Q_m = [q_1, q_2, \cdots q_m]$. Dan volgt er dat de projectie van A op de Krylov deelruimte, $Q_m^H A Q_m$, wordt gegeven door

$$T_m = \begin{pmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{m-1} \\ & & & \beta_{m-1} & \alpha_m \end{pmatrix}. \tag{6.7}$$

Hierbij zijn α_i en β_i gedefinieerd door vergelijking 6.6.

Bewijs. Voor het bewijs van deze stelling gebruiken we stelling 7.1 uit het hoofdstuk over de Arnoldi methode (hoofdstuk 7). Uit deze stelling volgt dat voor een algemene matrix A en een Krylov deelruimte $\mathcal{K}_m(A, v)$ van dimensie m, er een unitaire matrix $Q_m \in \mathbb{C}^{n\times m}$ en een boven-hessenbergmatrix $H_m \in \mathbb{C}^{m\times m}$ bestaan, zodat er geldt $Q_m^H A Q_m = H_m$. We beschouwen nu het geval dat A een hermitische matrix is. Hieruit volgt $Q_m^H A Q_m = \left(Q_m^H A Q_m\right)^H = H_m^H$, wat impliceert dat $H_m = H_m^H$ geldt. Maar een hermitische boven-hessenberg matrix moet ook een beneden-hessenberg matrix zijn. Uit definitie 6.2 volgt nu dat H_m een tridiagonale matrix is. Omdat H_m tridiagonaal is zullen we deze vanaf nu noteren met T_m.

Uit propositie 7.2 volgt er

$$AQ_m = Q_m T_m + \beta_m q_{m+1} e_m^H. \tag{6.8}$$

Hierbij geldt er $\beta_m = q_{m+1}^H A q_m$ en noteren we met q_i de i^{de} kolom van Q_m. Als we nu uit vergelijking (6.8) bepalen waar Aq_i aan gelijk is volgt er

$$Aq_1 = \alpha_1 q_1 + \beta_1 q_2, \tag{6.9}$$

$$Aq_i = \beta_{i-1} q_{i-1} + \alpha_i q_i + \beta_i q_{i+1}, \quad \text{voor } i \in \{2, 3, \cdots m\}. \tag{6.10}$$

Als we nu de orthogonaliteit van de q_i's gebruiken, immers Q_m is een unitaire matrix, volgt er:

$$\alpha_i = q_i^H A q_i, \qquad \beta_{i-1} = q_{i-1}^H A q_i, \qquad \beta_i = q_{i+1}^H A q_i. \tag{6.11}$$

Aangezien A hermitisch is volgt er nu dat α_i en β_i reëel zijn. Ook vormen de α_i's de diagonaal en de β_i's de nevendiagonaal van T_m. Aangezien T_m een hermitische tridiagonale matrix is, zijn dit ook gelijk alle elementen van T_m die ongelijk kunnen zijn aan nul. Dus T_m heeft inderdaad de vorm van vergelijking (6.7). Echter weten we nu alleen dat α_i en β_i voldoen aan (6.11) en niet of β_i ook voldoet aan vergelijking (6.6).

Verder volgt er uit vergelijking (6.9) en (6.10) inderdaad dat q_2 voldoet aan (6.4) en dat q_i voldoet aan (6.5) voor $i \in \{3, 4, \cdots, m-1\}$. Rest nu nog aan te tonen dat $\{q_1, q_2, \cdots, q_m\}$ ook een orthonormale basis vormt van $\mathcal{K}_m(A, v)$ en dat de uitdrukking voor β_i uit vergelijking (6.11) overeenkomt met die uit vergelijking (6.6).

Aangezien we de basis van $\mathcal{K}_m(A, v)$ recursief bepalen, zullen we nu met inductie bewijzen dat $\{q_1, q_2, \cdots, q_k\}$ een orthonormale basis vormt voor $\mathcal{K}_k(A, v)$ voor alle $k \leq m$. Voor $k = 1$ volgt er uit de definitie van q_1 dat er geldt $||q_1||_2 = 1$. Aangezien $\{q_1\}$ maar uit één genormeerde vector bestaat is dit een orthonormale verzameling van vectoren. Het rest nu aan te tonen dat $\{q_1\}$ ook $\mathcal{K}_1(A, v)$ opspant. Er geldt $\mathcal{K}_1(A, v) = \text{span}\{v\}$ en $\text{span}\{q_1\} = \text{span}\left\{\frac{1}{||v||_2} v\right\}$. Aangezien de ruimte opgespannen door één vector bestaat uit die vector maal een scalair, zie definitie 2.1, volgt er nu dat $\text{span}\{q_1\} = \mathcal{K}_1(A, v)$. Dus $\{q_1\}$ is inderdaad een orthonormale basis van $\mathcal{K}_1(A, v)$.

Neem nu aan dat we weten dat $\{q_1, q_2, \cdots, q_k\}$ een orthonormale basis is van $\mathcal{K}_k(A, v)$ en dat $k + 1 \leq m$. Dan volgt er met propositie 6.1 dat $\{q_1, q_2, \cdots, q_k, A^k v\}$ een basis vormt van $\mathcal{K}_{k+1}(A, v)$. Om dan aan te tonen dat $\{q_1, q_2, \cdots q_k, q_{k+1}\}$ een orthonormale basis vormt van $\mathcal{K}_{k+1}(A, v)$ volstaat het om de volgende drie punten aan te tonen:

- $||q_{k+1}||_2 = 1$;

- $\langle q_i, q_{k+1}\rangle = 0$ voor $i \in \{1, 2, \cdots, k\}$;
- $\text{span}\left\{q_1, q_2, \cdots q_k, A^k v\right\} = \text{span}\left\{q_1, q_2, \cdots q_k, q_{k+1}\right\}$.

Om dit aan te tonen definiëren we nu $\beta_0 = 0$ en $q_0 = \vec{0}$ zodat we nu vergelijking (6.10) en (6.5) ook kunnen gebruiken voor q_2.

We beginnen nu met aan te tonen dat geldt $||q_{k+1}||_2 = 1$. Echter aangezien er geldt $||q_{k+1}||_2 \geq 0$, volgt er dat $||q_{k+1}||_2 = 1$ equivalent is met $||q_{k+1}||_2^2 = 1$.

$$\begin{aligned} ||q_{k+1}||_2^2 &= \langle q_{k+1}, q_{k+1}\rangle \\ &= \frac{1}{\beta_k^2}\langle Aq_k - \beta_{k-1}q_{k-1} - \alpha_k q_k, Aq_k - \beta_{k-1}q_{k-1} - \alpha_k q_k\rangle. \end{aligned}$$

Als we nu gebruik maken van de vergelijkingen (6.11) en (6.10) en de (toegevoegde) lineariteit van het inproduct volgt er

$$\begin{aligned} &= \frac{1}{\beta_k^2}\left(\beta_{k-1}^2 + \alpha_k^2 + \beta_k^2 - 2\alpha_k^2 - 2\beta_{k-1}^2 + \alpha_k^2 + \beta_{k-1}^2\right) \\ &\frac{1}{\beta_k^2}\beta_k^2 = 1 \end{aligned}$$

Dus er geldt inderdaad $||q_{k+1}||_2 = 1$. Verder volgt hieruit dat geldt $\beta_i^2 = ||Aq_i - \beta_{i-1}q_{i-1} - \alpha_i q_i||_2^2$ wat impliceert dat $\beta_i = ||Aq_i - \beta_{i-1}q_{i-1} - \alpha_i q_i||_2$. We zien nu dus dat β_i gedefinieerd door vergelijking (6.11) inderdaad ook voldoet aan vergelijking (6.6).

Vervolgens controleren we de orthogonaliteit tussen q_{k+1} en q_i voor $i \in \{1, 2, \cdots, k\}$. Om $\langle q_i, q_{k+1}\rangle$ te bepalen gebruiken we vergelijking (6.5). Om na toepassing van deze vergelijking $\langle q_i, q_{k+1}\rangle$ te kunnen bepalen, bepalen we eerst $\langle q_i, Aq_k\rangle$. Hiertoe gebruiken we vergelijking (6.10) en dat geldt $A = A^H$:

$$\langle q_i, Aq_k\rangle = \langle Aq_i, q_k\rangle = \beta_{i-1}\langle q_{i-1}, q_k\rangle + \alpha_i\langle q_i, q_k\rangle + \beta_i\langle q_{i+1}, q_k\rangle. \tag{6.12}$$

Aangezien we al weten dat er geldt $\langle q_i, q_j\rangle = \delta_{ij}$ voor $i, j \in \{1, 2, \cdots, k\}$ volgt dus uit vergelijking (6.12) dat $\langle q_i, Aq_k\rangle = \begin{cases} 0 & i < k-1 \\ \beta_{i-1} & i = k-1 \\ \alpha_i & i = k \\ \beta_i & i = k+1 \end{cases}$. Omdat we hier alleen kijken naar het geval $i \in \{1, 2, \cdots, k\}$ volgt er nu

$$\langle q_i, q_{k+1}\rangle = \frac{1}{\beta_k}\langle q_i, Aq_k\rangle - \frac{\beta_{k-1}}{\beta_k}\langle q_i, q_{k-1}\rangle - \frac{\alpha_k}{\beta_k}\langle q_i, q_k\rangle = 0. \tag{6.13}$$

Er volgt dus dat geldt $\langle q_i, q_{k+1}\rangle = 0$ voor alle $i \in \{1, 2, \cdots, k\}$.

Er rest er nu te bewijzen dat $\text{span}\left\{q_1, \cdots, q_k, q_{k+1}\right\} = \text{span}\left\{q_1, \cdots, q_k, A^k v\right\}$. Als we nu vergelijking (6.10) herhaaldelijk toepassen op q_{k+1} vinden we dat er geldt $q_{k+1} = \nu A^k v + \sum_{i=1}^k \mu_i q_i$ voor $\mu_i, \nu \in \mathbb{C}$. Hierbij hebben we gebruik gemaakt van de definitie $q_1 = \frac{v}{||v||_2}$. Er volgt nu dat er geldt $\text{span}\left\{q_1, \cdots, q_k, q_{k+1}\right\} = \text{span}\left\{q_1, \cdots, q_k, \nu A^k v + \sum_{i=1}^k \mu_i q_i\right\}$. Echter omdat de ruimte opgespannen door een verzameling bestaat uit lineaire combinaties van vectoren uit die verzameling, volgt er nu: $\text{span}\left\{q_1, \cdots, q_k, \nu A^k v + \sum_{i=1}^k \mu_i q_i\right\} = \text{span}\left\{q_1, \cdots, q_k, A^k v\right\}$ en dus geldt er dat $\text{span}\left\{q_1, \cdots, q_k, q_{k+1}\right\} = \mathcal{K}_{k+1}(A, v)$. We hebben nu dus met inductie aangetoond dat $\{q_1, q_2 \cdots, q_m\}$ een orthonormale basis vormt voor $\mathcal{K}_m(A, v)$. □

We hebben nu uitdrukkingen gevonden om recursief een orthonormale basis van $\mathcal{K}_m(A, v)$ te bepalen en voor de projectie van A op deze Krylov deelruimte gevonden.

6.2 Het algoritme

De Lanczos methode werkt door het toepassen van de Rayleigh-Ritz procedure op $\mathcal{K}_m(A, v)$, met $A \in \mathbb{C}^{n\times n}, v \in \mathbb{C}^n$ en $m \in \mathbb{N}$ met m kleiner dan de graad van v. Het algoritme hiervoor is weergegeven in algoritme 3. Hierbij noteren we met q_i de $i+1^{\text{de}}$ kolom van Q_m. Met $B(i:j, k:l)$ noteren we de rijen i tot en met j en de kolommen k tot en met l van een matrix B. Indien $i:j$ dan wel $k:l$ is vervangen door $:$, nemen we alle rijen dan wel kolommen van de matrix. We zien hierbij dat in regel 3 niet alleen q_1 maar ook q_0 en β_0 gedefinieerd worden. Dit is zodat we in alle gevallen de recursie uit vergelijking (6.5) kunnen gebruiken en niet voor q_2 apart vergelijking (6.4) hoeven te nemen. Omdat q_0 verder geen deel uitmaakt van de basis voor de Krylov deelruimte, verwijderen we deze kolom weer uit Q in regel 16. In regel 3 tot en met 16 zien we de uitvoering van de eerste twee stappen van de Rayleigh-Ritz procedure. We zien hier dat de vergelijkingen (6.10) en (6.6) worden toegepast om de orthonormale basis van de Krylov deelruimte en T_m te bepalen. Aangezien deze beide vergelijkingen recursief werken, vormen de eerste k kolommen van Q_m en de $k \times k$ submatrix van T_m linksboven, een orthonormale basis van en een projectie op A voor $\mathcal{K}_k(A, v)$ voor $k \leq m$. Voor het bepalen

1 **def** *Lanczos*(A, x_0, k)**:**
  Input : -$A \in \mathbb{C}^{n\times n}$ een hermitische matrix.
  -$x_0 \in \mathbb{C}^n$ de vector waarmee we de Krylov deelruimte opspannen.
  -k de dimensie van de Krylov deelruimte.
  Output: -$Q \in \mathbb{C}^{n\times k+1}$ een unitaire matrix, waarvan de kolommen een orthonormale basis van de Krylov deelruimte vormen.
  -$T \in \mathbb{C}^{k+1\times k+1}$ een tridiagonale matrix, de projectie van A op de Krylov deelruimte.
  -$\Lambda \in \mathbb{C}^{k\times k}$ een diagonaalmatrix met de Ritz waarden als elementen.
  -$Y \in \mathbb{C}^{n\times k}$ een matrix met als kolommen de Ritz vectoren.
2 Creëer een $n \times k+2$ matrix Q en een $k+1 \times k+1$ matrix T.
3 Definieer $q_1 = \frac{x_0}{||x_0||_2}, \beta_0 = 0$ en $q_0 = \vec{0}$
4 **for** $i = 1:k$**:**
5 $z = Aq_i$
6 $\alpha_i = q_i^H z$
7 $z = z - \alpha_i q_i - \beta_{i-1} q_{i-1}$
8 $\beta_i = ||z||_2$
9 **if** $\beta_i = 0$**:**
10 break
11 **end**
12 $q_{i+1} = \frac{z}{\beta_i}$
13 $t_{ii} = \alpha_i$
14 $t_{i,i+1} = \beta_i$
15 $t_{i+1,i} = \beta_i$
16 **end**
17 Definieer $Q = Q(:, 2:end)$.
18 Bepaal de eigendecompositie $T_k = V\Lambda V^H$ van $T_k = T(1:k, 1:k)$.
19 Bepaal de Ritz vectoren $Y = Q(:, 1:k)V$

Algoritme 3: Lanczos methode

van de eigendecompositie van T, de laatste stappen van de Rayleigh-Ritz procedure, zullen wij de functie eig uit Matlab gebruiken.

In regel 9 wordt gecontroleerd of de norm van z ongelijk aan 0 is. Als deze gelijk is aan nul, zou de nieuwe basisvector gelijk zijn aan de nulvector. Echter kan de nulvector geen onderdeel zijn van een basis en dus stopt het algoritme. Het blijkt echter zo te zijn dat $\beta_i = 0$ geldt als i de waarde van de graad van v heeft bereikt.

Propositie 6.4 (afbraak Lanczos algoritme)**.** *Laat T gevormd worden door het Lanczos algoritme voor een*

hermitische matrix $A \in \mathbb{C}^{n\times n}$ en een vector $v \in \mathbb{C}^n$. Noem γ de graad van v. Dan geldt er $\beta_i = 0$ dan en slechts dan als geldt $i = \gamma$.

Bewijs. $\Rightarrow$ Stel dat $\beta_i = 0$. Dan volgt er dat $Aq_i - \alpha_i q_i - \beta_{i-1}q_{i-1} = 0$. Aangezien we alle vectoren q_l kunnen schrijven als lineaire combinaties van $A^k v$ voor $k \in \{0, 1, 2, \cdots, l-1\}$, hebben we dus een lineaire combinatie van vectoren van de vorm $A^j v$ voor $j = 0, 1, \cdots, i$ die gelijk is aan nul. Omdat we nu een polynoom p van graad i hebben zodat er geldt $p(A)v = 0$ volgt er $\gamma \leq i$. Maar er kan niet gelden $\gamma < i$. Immers als dit wel het geval was, zouden we meer dan γ orthonormale basisvectoren hebben gevonden, terwijl we uit propositie 5.3 weten dat de dimensie van de Krylov deelruimte niet groter kan zijn dan γ ↯.
$\Leftarrow$ Stel nu dat geldt $\gamma = i$. Als er dan zou gelden $\beta_i \neq 0$ zouden we in staat zijn om q_{i+1} te definiëren. Echter uit propositie 5.3 volgt dat de dimensie van $\mathcal{K}_l(A, v) = i$ voor $l \geq i$. Dan zouden we dus $i+1$ basisvectoren kunnen vinden voor een ruimte van dimensie i ↯. □

Uit het combineren van Proposities 5.2 ,5.5 en 6.4 volgt er nu: als β_i gelijk is aan nul, dan zijn alle Ritz paren echte eigenparen zijn van A.

6.3 Convergentie en nauwkeurigheid

In deze paragraaf bestuderen we de nauwkeurigheid en de convergentie van de Ritz paren en naar welk eigenpaar van A een Ritz paar convergeert. Gedurende dit hoofdstuk werken we met de Krylov deelruimte $K_k(A, v)$ met $v \in \mathbb{C}^n$ een vector en een hermitische matrix $A \in \mathbb{C}^{n\times n}$. Hierbij nemen we aan dar k voor alle Krylov deelruimten die we beschouwen kleiner is dan of gelijk aan de graad van v. De kolommen van $Q_k \in \mathbb{C}^{n\times k}$ vormen de orthonormale basis van de Krylov deelruimte en $T_k \in \mathbb{C}^{k\times k}$ is gedefinieerd door vergelijking (6.7). Met $\left(\theta_i^{(k)}, v_i^{(k)}\right)$ duiden we een eigenpaar van T_k aan. De i is hierbij de index van $\theta_i^{(k)}$ als we de Ritz waarden als volgt ordenen

$$\theta_1^{(k)} \leq \theta_2^{(k)} \leq \cdots \leq \theta_k^{(k)}. \tag{6.14}$$

Omdat hermitische matrices reële eigenwaarde hebben, zoals volgt uit Stelling 2.9, kunnen we de eigenwaarden op deze manier ordenen. Het bijbehorende Ritz paar is nu $\left(\theta_i^{(k)}, y_i^{(k)}\right)$ met $y_i^{(k)} = Q_k v_i^{(k)}$.

We beginnen met de analyse van naar welk eigenpaar van A het Ritz paar $\left(\theta_i^{(k)}, y_i^{(k)}\right)$ convergeert. Om dit te bepalen gebruiken we de Cauchy interlace Stelling die als volgt wordt gegeven:

Stelling 6.5. *(Cauchy interlace Theorem, [11]) Laat $A \in \mathbb{C}^{n\times n}$ een hermitische matrix zijn en neem $m \in \mathbb{N}, m < n$. Stel dat er een hermitische matrix $B \in \mathbb{C}^{m\times m}$, een matrix $C \in \mathbb{C}^{m\times n-m}$ en een matrix $D \in \mathbb{C}^{n-m\times n-m}$ bestaan zodat er geldt:*

$$A = \begin{bmatrix} B & C \\ C^H & D \end{bmatrix}.$$

Laat de eigenwaarden van A gegeven zijn door $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ en de eigenwaarden van B door $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_k$. Dan geldt er dat

$$\alpha_k \leq \beta_k \leq \alpha_{k+n-m} \quad voor k \in \{1, 2, \cdots, m\}.$$

Deze stelling kunnen we nu toepassen op T_{k+1}, er geldt immers dat $T_{k+1} = \begin{bmatrix} T_k & C \\ C^H & \alpha_{k+1} \end{bmatrix}$ met $C \in \mathbb{C}^k$ de vector met 0 als de eerste $k-1$ elementen en β_k als laatste element. Het toepassen van Stelling 6.5 op T_{k+1} met T_k als de hermitische submatrix geeft nu dat er geldt:

$$\theta_i^{(k+1)} \leq \theta_i^{(k)} \leq \theta_{i+1}^{(k+1)} \leq \theta_{i+1}^{(k)} \leq \theta_{i+2}^{(k+1)}. \tag{6.15}$$

Hierbij voldoen de eigenwaarden van T_k en T_{k+1} aan vergelijking (6.14). Uit vergelijking (6.15) volgt er nu, dat voor een vaste i, $\theta_i^{(k)}$ monotoon daalt als k toeneemt. Ook zien we dat de grootste eigenwaarde van T_{k+1}

altijd groter is dan of gelijk aan de grootste eigenwaarde van T_k. Immers vergelijking (6.15) geeft $\theta_k^{(k)} \leq \theta_{k+1}^{(k+1)}$ en T_k heeft k eigenwaarden en T_{k+1} heeft er $k+1$.
In het geval dat de graad van v gelijk is aan n, zijn de eigenwaarden van A gelijk aan die van T_n omdat deze matrices gelijkvormig zijn. We nemen aan dat de eigenwaarden $\lambda_i(A)$ zo zijn geordend zodat $\lambda_i(A) \leq \lambda_{i+1}(A)$ geldt. Omdat moet gelden $\lambda_i(A) = \theta_i^{(n)}$ volgt er dat $\theta_i^{(k)}$ monotoon dalend convergeert naar $\lambda_i(A)$ en dat $\theta_{k+1-i}^{(k)}$ monotoon stijgend convergeert naar $\lambda_{n+1-i}(A)$.
Echter als de graad van v, noem deze γ, kleiner is dan n kunnen we niet stellen naar welk eigenpaar van A een Ritz paar convergeert. Wel volgt er uit propositie 5.5 en 5.2 dat de eigenwaarden $Q_\gamma^H A Q_\gamma$ ook eigenwaarden zijn van A. Dus weten we wel dat $\theta_i^{(k)}$ monotoon dalend convergeert naar $\lambda_i(Q_\gamma^H A Q_\gamma)$ en dat $\theta_{k+1-i}^{(k)}$ monotoon dalend convergeert naar $\lambda_{\gamma+1-i}(Q_\gamma^H A Q_\gamma)$.

Verder volgt uit de Kaniel-Paige-Saad grenzen dat de Lanczos methode convergeert naar die eigenwaarden waarvan de vector v een component heeft in de richting van de eigenvectoren. Ook volgt eruit dat de eigenwaarden met de grootste absolute waarde het eerste convergeren. Tot slot is er ook bekend dat de Lanczos methode wel convergeert naar een meervoudige eigenwaarde, maar dat we maar één eigenvector voor de eigenwaarde vinden. Deze eigenvector is een willekeurige vector uit de eigenruimte behorend bij deze eigenwaarde [12, § 4.4.2].

Nu we gezien hebben waar een Ritz paar van T_k naartoe zal convergeren, willen we weten hoe nauwkeurig het Ritz paar het eigenpaar van A benaderd. Om een maat te hebben voor de afstand tussen een Ritz paar en een echt eigenpaar van A, en dus de nauwkeurigheid van het Ritz paar, definiëren we het residu van een Ritz paar. Dit lijkt sterk op het residu voor de machtsmethode, zie vergelijking (3.5). Het residu van het Ritz paar $\left(\theta_i^{(k)}, y_i^{(k)}\right)$ wordt gegeven door

$$r_i^{(k)} = A y_i^{(k)} - \theta_i^{(k)} y_i^{(k)}. \tag{6.16}$$

Echter zien we nu dat we de uitdrukking voor het residu, met behulp van vergelijking (6.8) als volgt kunnen omschrijven:

$$A y_i^{(k)} - \theta_i^{(k)} y_i^{(k)} = \left(A Q_k - Q_k \theta_i^{(k)}\right) v_i^{(k)} = \left(A Q_k - Q_k T_k\right) v_i^{(k)} = \beta_k q_{k+1} \cdot e_k^H \cdot v_i^{(k)}.$$

Als we met $v_i^{(k)}(j)$ het j^{de} element van de vector $v_i^{(k)}$ noteren volgt er dat het residu gelijk is aan:

$$r_i^{(k)} = \beta_k q_{k+1} v_i^{(k)}(k). \tag{6.17}$$

Dus de norm van het residu $\|r_i^{(k)}\|_2 = \beta_k |v_i^{(k)}(k)|$ is gegeven door de elementen van T_k en de eigenvectoren van T_k. Aangezien we deze beide bepalen met het Lanczos algoritme hebben we nu dus een maat voor de afstand tussen een Ritz paar en een echt eigenpaar van A die we gedurende het algoritme kunnen bepalen. Met het residu van het Ritz paar kunnen we nu de lokale fout van het Ritz paar afschatten op een vergelijkbare manier als bij de machtsmethode, zie Stelling 3.5.

Stelling 6.6 (Lokale fout Lanczos)**.** *Laat $A \in \mathbb{C}^{n\times n}$ een hermitische matrix zijn. Neem nu een Ritz paar $\left(\theta_i^{(k)}, y_i^{(k)}\right)$ bepaald uit de eigendecompositie van T_k en Q_k. Dan bestaan er een eigenwaarde λ van A en een vector $x \in \mathcal{E}(\lambda)$, zodat er geldt:*

$$|\lambda - \theta_i^{(k)}| \leq \beta_k |v_i^{(k)}(k)|, \quad \|x - y_i^{(k)}\|_2 \leq 2\frac{\beta_k |v_i^{(k)}(k)|}{\gamma_A(\lambda)}, \tag{6.18}$$

met $\gamma_A(\lambda)$ wederom de eigenwaarde kloof uit Definitie 3.4.

Deze Stelling is een toepassing van Stelling 4.9 uit [6] op de Lanczos methode. We zien voor de Lanczos methode dat de norm van $r_i^{(k)}$ een grens volgt op de lokale fout van de Ritz waarde. Hoe goed de Ritz vector ook een vector uit de eigenruimte van λ benadert is ook afhankelijk van de spreiding van de andere eigenwaarden van A die dicht bij λ liggen. Er volgt wel: hoe groter de afstand tussen λ en de rest van het

spectrum, hoe beter de benadering van de Ritz vector is bij dezelfde waarde van het residu.

Er is naast stelling 3.5 ook nog een afschatting voor de lokale fout van alle Ritz waarden van de matrix T_k.

Stelling 6.7. *(Lokale fout Ritz waarden) Laat $V^H T_k V = \Lambda$ de eigendecomopositie van T_k zijn en met $\Lambda = diag(\theta_1, \cdots, \theta_k)$. Dan zijn er eigenwaarden $\alpha_1, \cdots, \alpha_k$ van A zodat er geldt*

$$|\alpha_i - \theta_i| \leq \beta_k \quad voor\ i \in \{1, 2, \cdots, k\}. \tag{6.19}$$

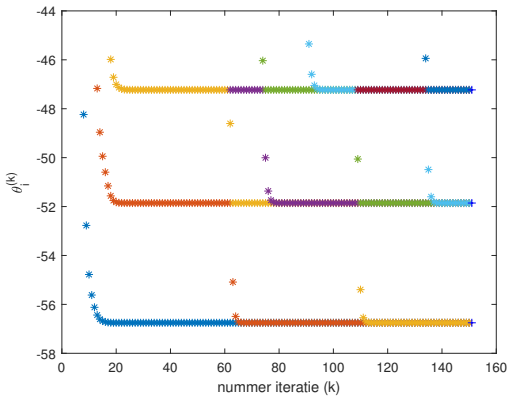
Deze Stelling is de veralgemening van Stelling 7.2.1[3]. Daar wordt deze stelling alleen bewezen voor symmetrische matrices. Het bewijs maakt echter alleen gebruik van dat de matrix symmetrisch is om de Stelling van Weyl toe te passen. Daar de Stelling van Weyl is ook geldig voor hermitische matrices, zie Stelling 2.16, is het bewijs uit [3] ook geldig voor Stelling 6.7.

We kunnen Stelling 6.7 ook zien als een gevolg van Stelling 6.6. Immers als we aannemen dat de eigenvectoren van T_k genormaliseerd zijn, volgt er $|v_i^{(k)}| \leq 1$ waardoor Stelling 6.7 inderdaad volgt uit Stelling 6.6. Uit Stelling 6.7 volgt nu dat β_k een bovengrens is voor de lokale fout voor alle Ritz waarden van T_k. Dit komt ook overeen met dat $\beta_k = 0$ impliceert dat alle Ritz paren echte eigenparen van A zijn.

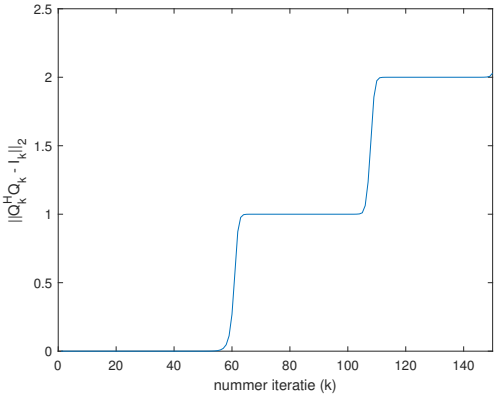
6.4 Orthogonalisatie

Het blijkt echter dat de Lanczos methode uit algoritme 3 niet altijd convergeert zoals besproken in de vorige paragraaf. Het volgende voorbeeld illustreert het probleem dat optreedt bij de convergentie van algoritme 3.

Voorbeeld 6.8. In dit voorbeeld bekijken we een symmetrische matrix $M \in \mathbb{R}^{2000\times2000}$. Deze is gevormd als $M = \frac{(QDQ^H)^H + QDQ^H}{2}$ waarbij $Q \in \mathbb{R}^{2000\times2000}$ een willekeurige orthogonale matrix en $D \in \mathbb{R}^{2000\times2000}$ een Gauss$(0, 15)$ diagonaalmatrix is. We passen algoritme 3 toe op M met $k = 150$ en $v \in \mathbb{R}^{2000}$ een willekeurige uniforme vector. De eigenwaarden van M zijn ook bepaald met de Matlab functie eig. De eigenwaarde met de grootste absolute waarde van M ligt rond -57. In grafiek 6a is het verloop van alle Ritz waarden kleiner dan



(a) Het verloop van alle Ritz waarden en alle eigenwaarden (+) kleiner dan -45 van M voor Lanczos zonder herorthogonalisatie. We zien hier dat er meerder Ritz waarden naar dezelfde eigenwaarde convergeren.



(b) De waarde van $||Q_k^H Q_k - \mathbb{I}_k||_2$ per iteratie. Theoretisch geldt er $||Q_k^H Q_k - \mathbb{I}_k||_2 = 0$. Zoals te zien in deze grafiek is het numeriek niet het geval dat deze norm gelijk is aan 0 maar heeft het verloop van deze norm gedurende de iteraties de vorm van een trap.

Figuur 6: In deze figuren zien we dat wanneer er een extra Ritz waarde ontstaat die convergeert naar de grootste eigenwaarde van M in absolute zin, de norm van $||Q_k^H Q_k - \mathbb{I}_k||_2$ dan of net daarvoor sterk stijgt.

-45 weergegeven. We zien hierbij dat er drie Ritz waarden naar de eigenwaarde rond -57 convergeren, ook

drie naar de eigenwaarde rond −52 en twee naar de eigenwaarde rond −47. Om het aantal Ritz waarden te bepalen dat naar een eigenwaarde convergeert, tellen we het aantal Ritz waarden dat naar deze eigenwaarde dalen. Figuur 6b toont de norm van $Q_k^H Q_k - \mathbb{I}_k$ voor alle iteraties. Theoretisch moet dit gelijk zijn aan nul omdat de kolommen van Q_k orthonormaal zijn. Echter zien we dat de norm $Q_k^H Q_k - \mathbb{I}_k$ twee keer in een paar iteraties met één toeneemt. In grafiek 6a zien we nu dat als de norm van $Q_k^H Q_k - \mathbb{I}_k$ bijna met één is toegenomen, er een extra Ritz waarde ontstaat die naar de grootste eigenwaarde van M, in absolute zin, convergeert. Uit de gelijktijdigheid van het ontstaan van een kopie van de Ritz waarde, die naar eigenwaarde van A met de grootste modulus convergeert, en het stijgen van de waarde $||Q_k^H Q_k - \mathbb{I}_k||_2$, zie figuur 6b, ontstaat het vermoeden dat het verlies van orthogonaliteit van de q_i's gelinkt kan worden aan het ontstaan van kopieën van Ritz waarden. △

In voorbeeld 6.8 zien we dat de kolommen van Q_k niet langer orthonormaal zijn en dat dit verband lijkt te houden met het ontstaan van kopieën van Ritz waarden. Het verlies van de orthonormaliteit van de q_i's, die volgt uit vergelijking (6.12), komt doordat de orthogonaliteit van q_{k+1} ten opzichte van $q_1, q_2, \cdots, q_k$ gewaarborgd wordt door de orthogonaliteit ten opzichte van q_{k-1} en q_k. Echter omdat we numeriek te maken hebben met afrondfouten, zal de orthogonaliteit van q_{k+1} ten opzichte van q_{k-1} en q_k niet exact zijn. Hierdoor is q_{k+1} ook niet meer exact orthogonaal ten opzichte van de overige vectoren. De waarde van $\langle q_i, q_{k+1}\rangle$ zal vanaf iteratie $k = i + 2$ steeds groter kunnebn worden, omdat voor toenemende waarde k elke iteratie een afrondfout in dit inproduct kan sluipen. Het blijkt dat zodra het residu van een Ritz paar klein wordt de nieuwe basis vectoren componenten in de richting van die Ritz vector krijgen, wat leidt tot een extra kopie van de Ritz waarde [12, §4.4.4] Om het ontstaan van kopieën van Ritz waarden te voorkomen moeten we dus de orthogonaliteit van de q_i's verbeteren .

Een manier om de orthogonaliteit van de vectoren $q_1, q_2, \cdots, q_k, q_{k+1}$ te verbeteren is door Aq_i te orthogonaliseren ten opzichte van alle al bepaalde basisvectoren, in plaats van alleen ten opzichte van q_i en q_{i-1}, door toepassing van Gram-Schimdt proces. Echter om er zeker van te zijn dat de fout nu klein genoeg is, dienen we twee keer Gram-Schmidt orthogonalisatie toe te passen [13]. Om dit te implementeren vervangen we regel 7 uit algoritme 3 door

$$z = z - Q(:, 2 : i+1)\left(Q(:, 2 : (i+1))^H z\right), z = z - Q(:, 2 : i+1)\left(Q(:, 2 : (i+1))^H z\right).$$

Dit wordt Lanczos met volledige herorthogonalisatie genoemd.

Echter kost Lanczos met volledige herorthogonalisatie veel extra bewerkingen. Daarom wordt in de praktijk vaak gebruik gemaakt van selectieve herorthogonalisatie. Hierbij wordt eerst vergelijking (6.5) gebruikt voor de orthogonalisatie. Daarna controleren we voor alle Ritz vectoren $y_i^{(k)}$ of de component van q_{k+1} in de richting de Ritz vector problemen oplevert voor de convergentie van de Ritz paren. Indien dat het geval is wordt q_{k+1} ook georthogonaliseerd ten opzichte van deze Ritz vector. Er geldt $y_i^{(k)} = \sum_{j=1}^{k} v_i^{(k)}(j) q_j$ waarbij $v_i^{(k)}(j)$ de j^{de} component van de eigenvector $v_i^{(k)}$ van T_k is. In feite orthogonaliseren we dus ook nu q_{k+1} ten opzichte van de vorige q_i's. Nu wordt alleen het belang van orthogonaliseren ten opzichte q_i aangegeven door $v_i^{(k)}(j)$.
Nu rest nog hoe we bepalen wanneer de component van q_{k+1} in de richting van $y_i^{(k)}$ problematisch is voor de convergentie. We hebben al genoemd dat dit gelinkt is aan het residu van het Ritz paar. De volgende Stelling geeft de orde van grootte van het inproduct $\langle y_i^{(k)}, q_{k+1}\rangle$.

Stelling 6.9. *(Paige [3, Stelling 7.3]) Laat T_k en $Q_k = [q_1, q_2, \cdots, q_k]$ de matrices zijn bepaald met het Lanczos algoritme voor een Krylov deelruimte van dimensie k. Neem aan dat we de exacte eigendecompositie van T_k hebben, die gegeven is door $V_k^H T_k V_k = \Lambda$ met $V_k = [v_1, v_2, \cdots v_k]$. Noteer de Ritz vectoren als $y_i^{(k)} = Q_k v_i^{(k)}$. Dan geldt er dat:*

$$\left(y_i^{(k)}\right)^H q_{k+1} = \frac{\mathcal{O}\left(\epsilon ||A||_2\right)}{\beta_k |v_i^{(k)}(k)|}. \tag{6.20}$$

Uit vergelijking (6.17) volgt er dat $\beta_k |v_i(k)|$ gelijk is aan de lengte van het residu van een Ritz paar.

1 **def** *Lanczosso*(A, x_0, k)**:**

Input : -$A \in \mathbb{C}^{n\times n}$ een hermitische matrix.
-$x_0 \in \mathbb{C}^n$ de vector waarmee we de Krylov deelruimte opspannen.
-k de dimensie van de Krylov deelruimte.

Output: -$Q \in \mathbb{C}^{n\times k+1}$ een unitaire matrix, waarvan de kolommen een orthonormale basis van de Krylov deelruimte vormen.
-$T \in \mathbb{C}^{k+1\times k+1}$ een tridiagonale matrix, de projectie van A op de Krylov deelruimte.
-$\Lambda \in \mathbb{C}^{k\times k}$ een diagonaalmatrix met de Ritz waarden als elementen.
-$Y \in \mathbb{C}^{n\times k}$ een matrix met als kolommen de Ritz vectoren.

2 Creëer een $n \times k+2$ matrix Q en een $k+1 \times k+1$ matrix T.
3 $q_1 = \frac{x_0}{||x_0||_2}, \beta_0 = 0, q_0 = \vec{0}$
4 **for** $i = 1 : k$**:**
5 $z = Aq_i$
6 $\alpha_i = q_i^H z$
7 $z = z - \alpha_i q_i - \beta_{i-1} q_{i-1}$
8 **for** $j = 1 : i-1$**:**
9 **if** $\beta_{i-1} |v_j^{(i-1)}(i-1)| \leq \sqrt{\epsilon} ||T_{i-1}||_2$**:**
10 $y_j^{(i-1)} = Q(:, 2:i) v_j^{(i-1)}$
11 $z = z - \left({y_j^{(i-1)}}^H z \right) y_j^{(i-1)}$
12 **end**
13 **end**
14 $\beta_i = ||z||_2$
15 **if** $\beta_i = 0$**:**
16 break
17 **end**
18 $q_{i+1} = \frac{z}{\beta_i}$
19 $t_{ii} = \alpha_i$
20 $t_{i,i+1} = \beta_i$
21 $t_{i+1,i} = \beta_i$
22 Bepaal de eigendecompositie $T(1:i, 1:i) = V^{(i)} \Lambda {V^{(i)}}^H$ van T
23 **end**
24 Definieer $Q = Q(:, 2 : end)$.
25 Bepaal de benaderde eigenvectoren $Y = Q(:, 1:k)V$

Algoritme 4: Lanczos met herorthogonalisatie

Bij selectieve herorthogonalisatie gebruiken dus vergelijking (6.20) om te bepalen wanneer q_{k+1} georthogonaliseerd moet worden ten opzichte van $y_i^{(k)}$. Het is aan te tonen dat het Lanczos algoritme de gewenste eigenschappen behoudt, indien Q_k semi-orthogonaal blijft. Hiermee word bedoelt dat er geldt $Q_k^H Q_k = \mathbb{I}_k + E$ met $||E||_2 < \sqrt{\epsilon}$ [12, §4.4.4]. We willen dus zorgen dat er geldt $\left(y_i^{(k)} \right)^H q_{k+1} < \sqrt{\epsilon}$. Uit Stelling 6.9 volgt dat hier meestal aan wordt voldaan indien er geldt $\beta_k |v_i(k)| < \sqrt{\epsilon} ||A||_2$. Aangezien $||A||_2$ wellicht onbekend is maar $||T_k||_2$ wel bekend is, vervangen we $||A||_2$ door $||T_k||_2$ [3, §7.4]. Omdat zowel A als T_k hermitische matrices zijn is de 2-norm van deze matrices gelijk aan hun de spectrale radius. Aangezien de spectrale radius gelijk is aan $\max_i |\lambda_i|$ voor $\lambda_i \in \Lambda(A)$ en dit de eigenwaarde is waar de Ritz waarden het eerst naar convergeren, is het vervangen van $||A||_2$ door $||T_k||_2$ redelijk.

Het algoritme voor selectieve herorthogonalisatie staat in algoritme 4 en is gebaseerd op algoritme 7.3 uit [3]. In regel 8 tot en met 12 zien we de uitvoering van de selectieve herorthogonalisatie. Dat hier β en de dimensie van T waar v een eigenvector van is $i-1$ is, komt omdat we i hier al wel met één verhoogd hebben maar nog niet T_i bepaald. Verder zien we nog dat de eigendecompositie van T_i nu elke iteratie bepaald wordt,

in plaats van pas als we T_k hebben bepaald, zoals in algoritme 3. Dit is noodzakelijk voor de selectieve herorthogonalisatie omdat we anders de grens uit Stelling 6.9 uit kunnen bepalen.

Op basis van Stelling 6.9 ontstaat ook een vermoeden voor de oorzaak achter het gedrag van de waarde $||Q_k^H Q_k - \mathbb{I}_k||_2$ uit voorbeeld 6.8. Het vermoeden is dat de waarde van deze norm samenhangt met de maximale kolomsom van $Q_k^H Q_k - \mathbb{I}_k$, wat gelijk is aan het maximale inproduct van één q_j met alle overige q_i's. Als de eerste Ritz waarde, die met de grootste modulus, convergeert zal er wegens Stelling 6.9 een nieuwe basisvector ontstaan, die ongeveer gelijk is aan de Ritz vector behorend bij deze Ritz waarde. De som van het inproduct van deze basisvector met alle overige basisvectoren is dan ongeveer gelijk aan 1. Immers een Ritz vector is een lineaire combinatie van q_i's. Als er daarna andere Ritz waarden convergeren en basisvectoren ongeveer gelijk worden aan die Ritz vectoren is de maximale som van het inproduct van een deze basisvectoren met alle andere basisvectoren nog steeds ongeveer gelijk aan 1. Echter als er nog een kopie van een Ritz waarde ontstaat, wat het eerst gebeurd voor de Ritz waarde met de grootste modulus, is er dus wederom een basisvector ongeveer gelijk aan de bijbehorende Ritz vector. Nu heeft deze basisvector een inproduct van twee met de overige basisvectoren omdat er al een andere basisvector bestond die een kopie van deze Ritz vector is.

6.5 Spectrale transformatie

Zoals we hebben gezien in hoofdstuk 6.3 convergeren bij de Lanczos methode de eigenwaarden met de grootste modulus van A het eerst. Echter als we nu opzoek zijn naar de eigenwaarden van een matrix A die het dichtst bij de waarde $\sigma \in \mathbb{R}$ liggen, kan het zijn dat het lang duurt voordat deze eigenwaarden convergeren. Om dit probleem te verhelpen kunnen we de Lanczos methode toepassen op een shift-and-invert operator, zodat de eigenwaarden het dichtst bij σ als eerste convergeren. Deze operator is gegeven door $B = (A - \sigma\mathbb{I})^{-1}$. Dit werkt op een vergelijkbare manier als de inverse iteratie, die de machtsmethode gebruikt om het eigenpaar het dichtst bij $\sigma \in \mathbb{C}$ te bepalen. Zoals we al hebben gezien in het hoofdstuk over de inverse iteratie (hoofdstuk 4) geldt er nu voor B:

- Voor alle $\mu \in \Lambda(B)$ bestaat er een $\lambda \in \Lambda(A)$ zodat $\mu = \frac{1}{\lambda - \sigma}$.
- Als er voor $v \in \mathbb{C}^n$ geldt $Bv = \mu v$ dan geldt ook er $Av = \lambda v$ met $\lambda = \sigma + \frac{1}{\mu}$.

Er moet nog altijd wel gelden dat $A - \sigma\mathbb{I}$ inverteerbaar is om dit uit te kunnen voeren. Als $A - \sigma\mathbb{I}$ niet inverteerbaar is hebben we voor σ een eigenwaarde van A gekozen. We kunnen dan natuurlijk wel een waarde $\tilde{\sigma}$ kiezen die iets van σ afligt om dit probleem te voorkomen.

Om dit te implementeren hoeven er maar een paar aanpassingen gedaan te worden aan de gewone Lanczos methode. Aangezien Lanczos met selectieve herorthogonalisatie het meest gebruikt wordt, zullen we aangeven hoe we spectrale transformatie kunnen implementeren uitgaande van algoritme 4. Hiertoe moet het volgende veranderd worden aan dat algoritme

- In regel 5 vervangen we $z = Aq_i$ door het oplossen van het lineaire stelsel $(A - \sigma\mathbb{I})\, z = q_i$ voor z.
- Voeg aan regel 21 toe dat de elementen van Λ, vervangen worden door $\theta_i = \sigma + \frac{1}{\lambda_i}$.

Zoals volgt uit de analyse van paragraaf 6.3 convergeren nu de eigenwaarden van B met de grootste modulus het eerst. Dit leidt ertoe dat de eigenwaarden $\lambda \in \Lambda(A)$ waarvoor $\left|\frac{1}{\lambda - \sigma}\right|$ het grootst is, en dus die eigenwaarden die het dichtst bij σ liggen, het eerst convergeren.

6.6 Numerieke voorbeelden

We zullen nu een numerieke voorbeeld van de Lanczos methode bekijken. We hanteren hierbij de volgende drie maten voor de fout van een Ritz waarde $\theta_i^{(k)}$ voor een matrix A.

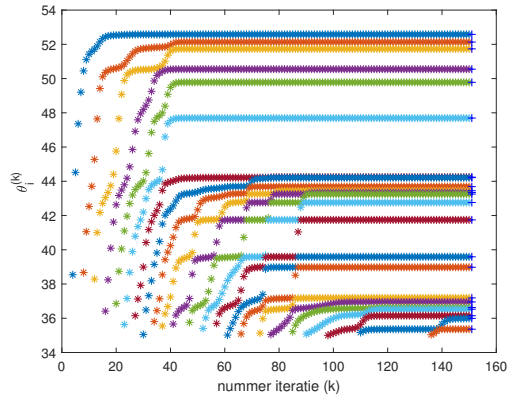
- De globale fout gegeven door $\frac{\left|\lambda_i(A) - \theta_i^{(k)}\right|}{|\lambda_i(A)|}$.

- De lokale fout, die gelijk is aan $\frac{\min_{\lambda \in \Lambda(A)} \left|\lambda - \theta_i^{(k)}\right|}{|\lambda_i(A)|}$.
- Het residu, bepaald door $\frac{\beta_k |v_i^{(k)}(k)|}{|\lambda_i(A)|}$.

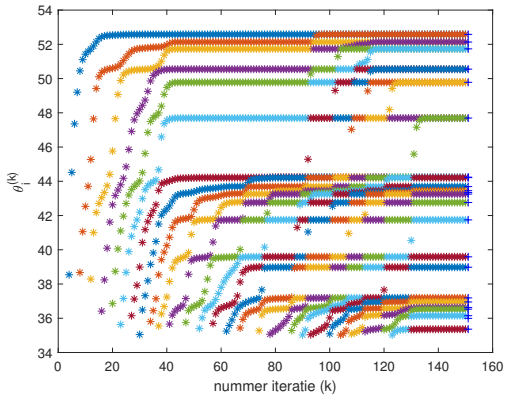
Hierbij is $\lambda_i(A)$ de eigenwaarde van A waar $\theta_i^{(k)}$ naartoe convergeert volgens de analyse uit paragraaf 6.3. Als de eigenwaarden van A en de Ritz waarden zijn geordend volgens 6.14, dan volgt voor negatieve Ritz waarde $\theta_i^{(k)}$ dat die naar $\lambda_i(A)$ convergeert en voor positieve Ritz waarde dat deze naar $\lambda_{n+1-l}(A)$ convergeert met $l = k + 1 - i$.[3, §7.3].

In het volgende voorbeeld vergelijken we de werking van de Lanczos methode met volledige, selectieve en zonder herorthogonalisatie.

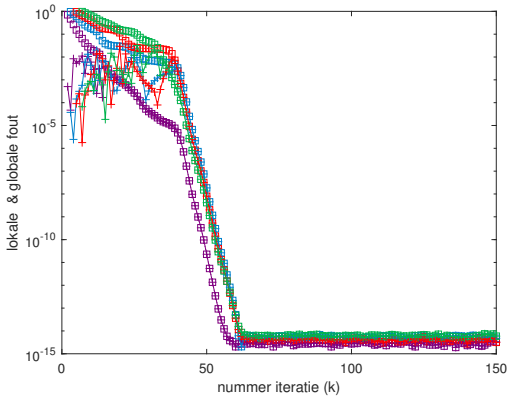
Voorbeeld 6.10. We bekijken opnieuw de matrix $M \in \mathbb{R}^{2000\times 2000}$ uit voorbeeld 6.8 met dezelfde vector $v \in \mathbb{C}^{2000}$ en $k = 150$. We voeren nu niet alleen de Lanczos zonder herorthogonalisatie uit, algoritme 3, maar ook met volledige en selectieve, algoritme 4, herorthogonalisatie. In figuur 7 zijn de Ritz waarden en eigenwaarden groter dan 35 te zien evenals de lokale en globale fout en het residu voor de grootste vier eigenwaarden, in reële zin. Er is voor gekozen om deze resultaten voor de selectieve herorthogonalisatie niet weer gegeven, omdat deze niet te onderscheiden zijn van de Lanczos met volledige herorthogonalisatie. We zien in de figuren 7a en 7b dat er bij Lanczos zonder herorthogonalisatie kopieën van Ritz waarden ontstaan en dat dit niet gebeurt bij de volledige herorthogonalisatie. Verder zien we in figuur 7c en 7d dat de lokale fout altijd kleiner is dan de globale fout. Dit volgt overigens ook al uit de definitie van deze twee fouten. Voor de volledige herorthogonalisatie zien we dat zodra deze fouten rond de machine precisie zitten, ze daar blijven schommelen. Voor Lanczos zonder herorthogonalisatie zien we dat, behalve voor de grootste Ritz waarde, de lokale en globale fout rond de iteraties 90 tot 100 terug schieten naar de orde 10^0. Dit komt omdat er dan een kopie van een Ritz waarde is ontstaan, die nu $\theta_2^{(k)}$ is geworden in de ordening van vergelijking 6.14. De fout schiet hier dus omhoog omdat we nu een kopie van Ritz waarde volgen. Omdat deze kopie nog niet geconvergeerd is vergroot ook de lokale fout, die afneemt als de kopie convergeert. De globale fout blijft van orde 10^0 omdat de kopie niet naar $\lambda_i(A)$ convergeert maar naar $\lambda_j(A)$ met $j > i$. In figuur 7e en 7f zien we dat het residu altijd groter is dan de lokale fout, wat theoretisch volgt uit stelling 6.6, totdat de machine precisie bereikt wordt. Daarna blijft het residu krimpen terwijl de lokale fout daar blijft hangen. Omdat we de machine precisie hebben bereikt is dit niet in tegenspraak met stelling 6.6 omdat waarden kleiner dan de machine precisie onbetrouwbaar zijn. Verder vertoont het residu hetzelfde gedrag als de lokale fout bij het ontstaan van kopieën van Ritz waarden omdat het residu ook alleen de afstand tot het dichtstbijzijnde eigenpaar aangeeft. In figuur 8 zijn per iteratie alle Ritz waarden weergeven waarvan de eigenvector $y_i^{(k)}$ gebruikt worden voor de herorthogonalisatie van q_{k+1}. We zien hierbij dat de Ritz waarden op het oog al geconvergeerd zijn, we eisen immers dat het residu kleiner is dan $\sqrt{\epsilon}||T_k||_2$ voordat we een Ritz vector selecteren. Ook zien we dat Ritz vectoren met een Ritz waarden met de grootste moduli het eerst voor herorthogonalisatie worden geselecteerd. Dit reflecteert dat de Ritz waarden het eerst convergeren naar eigenwaarden met de grootste moduli. △



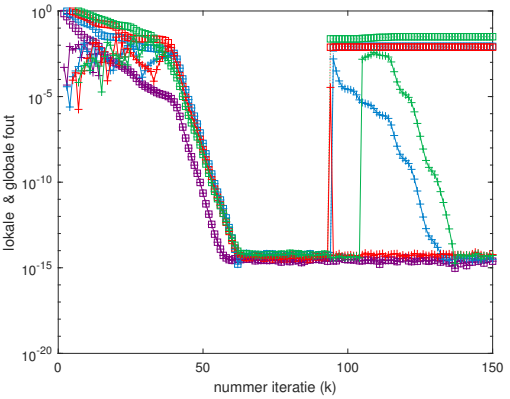
(a) De Ritz waarden en de eigenwaarden (+) groter dan 35 van M voor Lanczos met volledige herorthogonalisatie. We zien dat er geen kopieën van Ritz waarden ontstaan.



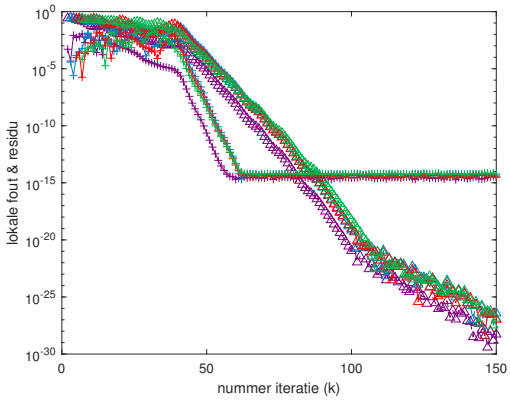
(b) De Ritz waarden en de eigenwaarden (+) groter dan 35 van M voor Lanczos zonder herorthogonalisatie. We zien dat er kopieën van Ritz waarden ontstaan.



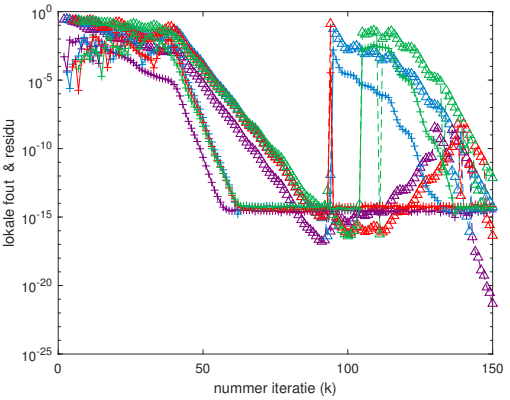
(c) De lokale (+) en globale fout (□) van de grootste vier positieve eigenwaarden van M met volledige herorthogonalisatie.



(d) De lokale (+) en globale fout (□) van de grootste vier positieve eigenwaarden van M zonder herorthogonalisatie.

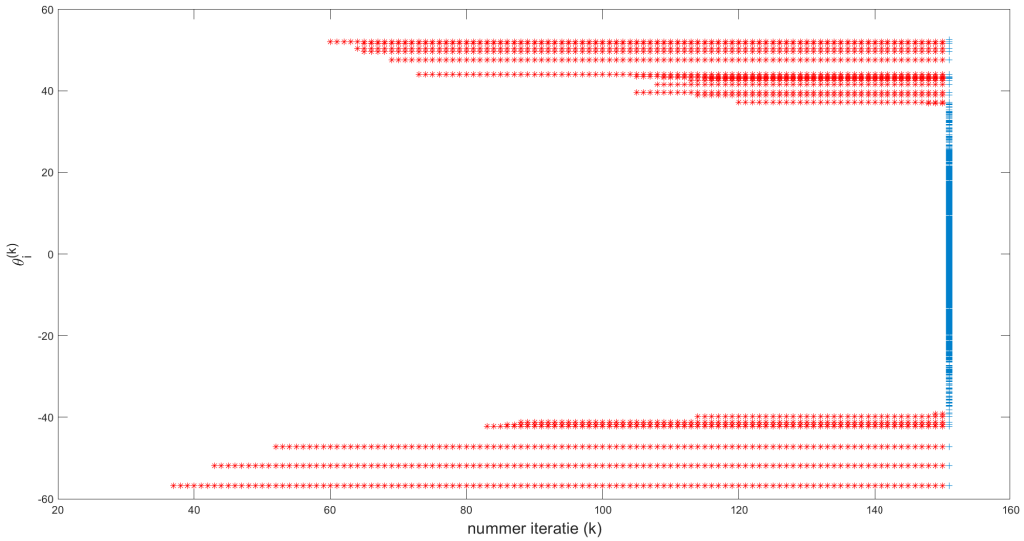


(e) De lokale (+) fout en het residu (△) van de grootste vier positieve eigenwaarden van M met volledige herorthogonalisatie.



(f) De lokale (+) fout en het residu (△) van de grootste vier positieve eigenwaarden van M zonder herorthogonalisatie.

Figuur 7: In deze figuren zijn de Ritz waarden en de eigenwaarden groter dan 35 van M te zien voor Lanczos met volledige en zonder herorthogonalisatie. Ook is de globale en lokale fout en het residu voor beide varianten van Lanczos voor de vier grootste eigenwaarden te zien. Hierbij zien we dat de fout voor Lanczos zonder herorthogonalisatie stijgt als er kopieën van Ritz waarden ontstaan.



Figuur 8: De Ritz waarde (+) waarvan de eigenvectoren $y_i^{(k)}$ geselecteerd worden voor de herorthogonalisatie van q_{k+1}. Ook alle eigenwaarden van M (+) worden getoond.

7 Arnoldi methode

In dit hoofdstuk bespreken we de Arnoldi methode. Net zoals de Lanczos methode uit hoofdstuk 6 berust deze methode op toepassing van de Rayleigh-Ritz procedure op een Krylov deelruimte. Deze methode is in staat om meerdere eigenparen van een willekeurige matrix $A \in \mathbb{C}^{n\times n}$ te benaderen. We beginnen het hoofdstuk met de afleiding van een methode om een orthogonale basis van de Krylov deelruimte te bepalen in paragraaf 7.1. In deze paragraaf leiden we ook de uitdrukkingen voor de projectie van A op de Krylov deelruimte af. Daarna bespreken we het standaard Arnoldi algoritme in paragraaf 7.2 en bekijken we de convergentie en de nauwkeurigheid van dit algoritme in paragraaf 7.3. In paragraaf 7.4 bekijken we verschillende variaties op het Arnoldi algoritme om convergentie te verkrijgen met een kleinere dimensie van de Krylov deelruimte of om gewenste eigenwaarden van A het eerst te laten convergeren. Tot slot bekijken we in paragraaf 7.5 een paar numerieke voorbeelden van de Arnoldi methode.

7.1 Bepaling orthonormale basis Krylov deelruimte & projectie van A

We gaan nu een orthonormale basis van de Krylov deelruimte $\mathcal{K}_m(A, v)$ bepalen voor een willekeurige matrix A. Wederom volgt uit Propositie 6.1: stel dat we een orthonormale basis $\{q_1, q_2, \cdots, q_m\}$ hebben voor $\mathcal{K}_m(A, v)$, dan vormt $\{q_1, q_2, \cdots, q_m, A^m v\}$ een basis van $\mathcal{K}_{m+1}(A, v)$. We nemen hierbij aan dat voor de graad van v, γ, geldt $m + 1 \leq \gamma$ zodat de dimensie van $\mathcal{K}_{m+1}(A, v)$ gelijk is aan $m + 1$, zie Propositie 5.3.

Door volgende iteratieve proces uit te voeren voor $k \in \{2, 3, \cdots, m + 1\}$ kunnen we een set genormaliseerde vectoren $\{q_1, q_2, \cdots, q_{m+1}\}$ bepalen. Hierbij geldt er $q_1 = \frac{v}{||v||_2}$.

$$
\begin{aligned}
q_{k+1} &= Aq_k \\
&\textbf{for } i \in \{1, 2, \cdots, k\} \\
q_{k+1} &= q_{k+1} - \langle q_i, q_{k+1}\rangle q_i \\
&\text{beëindig } \textbf{for} \\
q_{k+1} &= \frac{q_{k+1}}{||q_{k+1}||_2}.
\end{aligned}
$$

Dit proces is het modified Gram-Schimdt orthogonalisatieprocedé [13]. We zullen aantonen dat deze verzameling vectoren een orthonormale basis vormt van $\mathcal{K}_{m+1}(A,v)$ in Stelling 7.1. In dezelfde Stelling bepalen we ook de projectie van A op $\mathcal{K}_{m+1}(A,v)$. Het modified Gram-Schimdt orthogonalisatieprocedé kunnen we ook samenvatten in de volgende formule

$$q_{k+1} = \frac{Aq_k - \langle q_1, Aq_k\rangle q_1 - \langle Aq_k - \langle q_1, Aq_k\rangle q_1, q_2\rangle q_2 - \cdots - \langle Aq_k - \langle q_1, Aq_k\rangle q_1 - \cdots - \langle q_{k-1}, Aq_k\rangle q_{k-1}, q_k\rangle q_k}{\|Aq_k - \langle q_1, Aq_k\rangle q_1 - \langle Aq_k - \langle q_1, Aq_k\rangle q_1, q_2\rangle q_2 - \cdots - \langle Aq_k - \langle q_1, Aq_k\rangle q_1 - \cdots - \langle q_{k-1}, Aq_k\rangle q_{k-1}, q_k\rangle q_k\|_2}. \tag{7.1}$$

Stelling 7.1. *Laat $A \in \mathbb{C}^{n\times n}$ een willekeurige matrix zijn. Neem een vector $v \in \mathbb{C}^n$ en noem de graad van v, γ. Voor een Krylov deelruimte $K_m(A,v)$ met $m \leq \gamma$ geldt er dat $\{q_1, q_2, \cdots, q_m\}$ een orthonormale basis van deze deelruimte vormt met:*

$$q_1 = \frac{v}{\|v\|_2}, \tag{7.2}$$

en q_i voor $i \in \{2,3,\cdots,m\}$ gegeven door vergelijking (7.1). Definieer de matrix $Q_m \in \mathbb{C}^{n\times m}$ als $Q_m = [q_1, q_2, \cdot, sq_m]$. De projectie van A op de Krylov deelruimte, $Q_m^H A Q_m = H_m$ is een boven-hessenberg matrix met de volgende elementen:

$$h_{kl} = \begin{cases} \langle q_k, Aq_l\rangle & k \leq l \\ \|Aq_l - \sum_{i=1}^{l}\langle q_i, Aq_l\rangle q_i\|_2 & k = l+1 \\ 0 & k > l+1 \end{cases}. \tag{7.3}$$

Bewijs. Aangezien we de verzameling $\{q_1, q_2, \cdots, q_m\}$ recursief definiëren volstaat een bewijs met inductie op k om aan te tonen dat $\{q_1, q_2, \cdots, q_m\}$ een orthonormale basis vormt van $\mathcal{K}_m(A,v)$. Uit Stelling 6.3 volgt er voor $k = 1$ dat $\{q_1\}$ een orthonormale basis van $\mathcal{K}_1(A,v)$ vormt.

Neem nu aan aan dat $\{q_1, q_2, \cdots, q_k\}$ een orthonormale basis van $\mathcal{K}_k(A,v)$ vormt, met $k \leq m-1$. Als we nu de (toegevoegde) lineariteit van het inproduct en dat er geldt $\langle q_i, q_j\rangle = \delta_{ij}$ voor $i,j \in \{1,2,\cdots,k\}$ gebruiken, volgt er dat vergelijking 7.1 gelijk is aan:

$$q_{k+1} = \frac{Aq_k - \sum_{i=1}^{k}\langle q_i, Aq_k\rangle q_i}{\|Aq_k - \sum_{i=1}^{k}\langle q_i, Aq_k\rangle q_i\|_2}. \tag{7.4}$$

Uit de inductie hypothese volgt dat het voldoende is om de volgende drie punten aan te tonen om te laten zien dat $\{q_1, q_2, \cdots, q_{k+1}\}$ een orthonormale basis van $\mathcal{K}_{k+1}(A,v)$ vormt:

- $\|q_{k+1}\|_2 = 1$;
- $\langle q_{k+1}, q_i\rangle = 0$ voor $i \in \{1,2,\cdots,k\}$;
- $\operatorname{span}\left\{q_1, q_2, \cdots, q_k, A^k v\right\} = \operatorname{span}\{q_1, q_2, \cdots, q_k, q_{k+1}\}$.

Uit vergelijking (7.4) volgt er direct dat er aan het eerste punt is voldaan. Verder volgt er uit vergelijking (7.4), $\langle q_i, q_i\rangle = \delta_{ij}$ voor $i,j \in \{1,2,\cdots,k\}$. Als we dit combineren met de (toegevoegde) lineariteit van het inproduct volgt er:

$$\begin{aligned}
\langle q_{k+1}, q_i\rangle &= \frac{1}{\|Aq_k - \sum_{j=1}^{k}\langle q_j, Aq_k\rangle q_j\|_2}\langle Aq_k - \sum_{j=1}^{k}\langle q_j, Aq_k\rangle q_j, q_i\rangle \\
&= \frac{1}{\|Aq_k - \sum_{j=1}^{k}\langle q_j, Aq_k\rangle q_j\|_2}\left(\langle Aq_k, q_i\rangle - \sum_{j=1}^{k}\left(\langle q_j, Aq_k\rangle^*\langle q_j, q_i\rangle\right)\right) \\
&= \frac{\langle Aq_k, q_i\rangle - \langle q_i, Aq_k\rangle^*}{\|Aq_k - \sum_{j=1}^{k}\langle q_j, Aq_k\rangle q_j\|_2} = 0.
\end{aligned}$$

Dus de vectoren $\{q_1, q_2, \cdots, q_{k+1}\}$ vormen inderdaad een orthonormaal systeem. Het rest nu nog het laatste punt aan te tonen, oftewel dat het ook een basis van $\mathcal{K}_{k+1}(A, v)$ is. Uit het herhaald toepassen van vergelijking (7.4) volgt dat er $\nu, \mu_i \in \mathbb{C}$ bestaan zodat geldt $q_{k+1} = \nu A^k v + \sum_{i=1}^{k} \mu_i q_i$. Er volgt dus

$$\operatorname{span}\{q_1, q_2, \cdots, q_k, q_{k+1}\} = \operatorname{span}\left\{q_1, q_2, \cdots, q_k, \nu A^k v + \sum_{i=1}^{k} \mu_i q_i\right\}.$$

Uit Definitie 2.1 volgt er dat dit gelijk is aan

$$\operatorname{span}\{q_1, q_2, \cdots, q_k, q_{k+1}\} = \operatorname{span}\left\{q_1, q_2, \cdots, q_k, A^k v\right\}.$$

We hebben nu dus met inductie aangetoond dat $\{q_1, q_2, \cdots, q_m\}$ een orthonormale basis vormt van $\mathcal{K}_m(A, v)$ met $q_1 = \frac{v}{\|v\|_2}$ en q_i voor $i \geq 2$ gedefinieerd door vergelijking (7.1).

Nu rest het aan te tonen dat $H_m = Q_m^H A Q_m$ een boven-hessenberg matrix is met elementen die voldoen aan vergelijking (7.3). Uit de definitie van H_m volgt dat er geldt

$$h_{kl} = q_k^H A q_l. \tag{7.5}$$

Als we nu vergelijking (7.4) gebruiken en hier Aq_m uit vrij schrijven, krijgen we

$$h_{kl} = q_k^H \left(q_{l+1} \|Aq_l - \sum_{i=1}^{l} \langle q_i, Aq_l \rangle q_i\|_2 + \sum_{i=1}^{l} \langle q_i, Aq_l \rangle q_i \right) \tag{7.6}$$

$$= \begin{cases} \langle q_k, Aq_l \rangle & k \leq l \\ \|Aq_l - \sum_{i=1}^{l} \langle q_i, Aq_l \rangle q_i\|_2 & k = l+1 \\ 0 & k > l+1. \end{cases} \tag{7.7}$$

Dus H_m is inderdaad een boven-hessenberg matrix met elementen die voldoen aan vergelijking (7.3). □

We hebben nu uitdrukkingen gevonden voor een orthonormale basis van $\mathcal{K}_m(A, v)$ en de projectie $Q_m^H A Q_m = H_m$. Het blijkt echter dat we ook een uitdrukking voor AQ_m kunnen bepalen. Deze uitdrukking zullen we later gebruiken om het residu van een Ritz paar te bepalen.

Propositie 7.2. *Neem een willekeurige matrix $A \in \mathbb{C}^{n \times n}$ en een vector $v \in \mathbb{C}^n$. Noem de graad van v, γ. Voor een Krylov deelruimte $K_k(A, v)$ met $k \leq \gamma$ en H_k en Q_k zoals uit Stelling 7.1 geldt er*

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^H. \tag{7.8}$$

Bewijs. Uit de definitie van Q_k volgt er

$$AQ_k = [Aq_1, Aq_2, \cdots Aq_k].$$

Als we uit vergelijking (7.4) Aq_l vrijschrijven volgt er

$$= \left[q_2 \|Aq_1 - \langle q_1, Aq_1 \rangle q_1\|_2 + \langle q_1, Aq_1 \rangle q_1, \cdots\cdots\cdots, q_{k+1} \|Aq_l - \sum_{i=1}^{k} \langle q_i, Aq_k \rangle q_i\|_2 + \sum_{i=1}^{k} \langle q_i, Aq_k \rangle q_i \right].$$

Met gebruik van vergelijking (7.3) volgt er

$$= Q_k H_k + h_{k+1,k} q_{k+1} e_k^H.$$

□

De Arnoldi methode net als de Lanczos methode gebruik maakt van de Rayleigh-Ritz procedure op Krylov deelruimte. We zijn we nu instaat om eerste twee stappen van deze Rayleigh-Ritz procedure uit te voeren.

def *Arnoldi*(A, x_0, k)**:**

Input : -$A \in \mathbb{C}^{n\times n}$ een matrix waarvan we de eigenwaarde en eigenvectoren willen benaderen.
- $x_0 \in \mathbb{C}^n$ de vector waarmee we de Krylov deelruimte opspannen.
-k de gewenste dimensie van de Krylov deelruimte.

Output: -$Q \in \mathbb{C}^{n\times k+1}$ matrix waarvan de kolommen een orthonormale basis van de Krylov deelruimte vormen.
-$H \in \mathbb{C}^{k+1\times k}$ een boven-hessenberg matrix, de projectie van A op de Krylov deelruimte.
-$\Lambda \in \mathbb{C}^{k\times k}$ een diagonaalmatrix met de Ritz waarden.
-$Y \in \mathbb{C}^{n\times k}$ een matrix met als kolommen de Ritz vectoren.

Creëer een $n \times k+1$ matrix Q en een $k+1 \times k$ matrix H.
$q_1 = \frac{x_0}{||x_0||_2}$
for $j = 1 : k$**:**
    $w = Aq_j$
    for $i = 1 : j$**:**
        $h_{ij} = q_i^H \cdot w$
        $w = w - h_{ij}q_i$
    end
    $h_{j+1,j} = ||w||_2$
    if $h_{j+1,j} = 0$**:**
        break
    end
    $q_{i+1} = \frac{w}{h_{j+1,j}}$
end
Bepaal de eigendecompositie $H = V\Lambda V^H$ van H
Bepaal de benaderde Ritz vectoren $Y = Q(:, 1 : k)V$

Algoritme 5: Arnoldi methode

7.2 Algoritme Arnoldi

In algoritme 5 is het Arnoldi algoritme weergegeven. We zien hierbij dat h_{ij} in het algoritme door een net iets andere vergelijking bepaald wordt, dan we in vergelijking (7.3) hebben afgeleid. Dit komt doordat vergelijking (7.1), Gram-Schmidt, en (7.4), modified Gram-Schmidt, theoretisch wel hetzelfde zijn, maar niet numeriek. We gebruiken voor het algoritme het modified Gram-Schmidt orthogonalisatieprocedé omdat deze numeriek stabieler is dan de standard Gram-Schmidt. Immers we zien in vergelijking (7.1) dat we er bij modified Gram-Schmidt niet vanuit gaan dat $\langle q_i, q_j\rangle = \delta_{ij}$ bij het bepalen q_k voor $i < j < k$. Een fout in de waarde van $\langle q_i, q_j\rangle$ heeft nu dus minder invloed op de orthogonaliteit van q_j en q_k dan bij de standaard Gram-Schimdt waar we wel aannemen dat er geldt $\langle q_i, q_j\rangle = \delta_{ij}$ [1, §6.2.2].
Uit het omschrijven van vergelijking (7.1) volgt er dat de definitie van h_{ij} uit het algoritme overeenkomt met de q_i's bepaald met modified Gram-Schmidt. Onze implementatie van het Arnoldi algoritme is gebaseerd op algorimte 7.3 uit [12]. Het verschil in de bepaling van h_{ij} voor $i \leq j$ komt door verschillen in de definitie van het inproduct. Ons inproduct is gelijk aan $\langle x, y\rangle = \sum_{i=1}^n x_i^* y_i$, zie Definitie 2.10, terwijl ze in [12] gebruiken $\langle x, y\rangle = \sum_{i=1}^n x_i y_i^*$.

Voor het Arnoldi algoritme geldt eenzelfde relatie tussen de graad van v en de afbraak van het algoritme als voor het Lanczos algoritme, zie propositie 6.4.

Propositie 7.3 (afbraak Arnoldi algoritme)**.** *Laat H de matrix zijn uit Arnoldi algoritme voor een matrix $A \in \mathbb{C}^{n\times n}$ en een vector $v \in \mathbb{C}^n$. Noem de graad van v, γ. Er geldt nu $h_{j+1,j} = 0$ dan en slechts dan als j gelijk is aan γ.*

Bewijs. $\Rightarrow$ Stel dat $h_{j+1,j} = 0$. Dan volgt er dat

$$\frac{Aq_j - \langle q_1, Aq_j\rangle q_1 - \langle Aq_j - \langle q_1, Aq_j\rangle q_1, q_2\rangle q_2 - \cdots - \langle Aq_j - \langle q_1, Aq_j\rangle q_1 - \cdots - \langle q_{j-1}, Aq_j\rangle q_{j-1}, q_j\rangle q_j}{||Aq_j - \langle q_1, Aq_j\rangle q_1 - \langle Aq_j - \langle q_1, Aq_j\rangle q_1, q_2\rangle q_2 - \cdots - \langle Aq_j - \langle q_1, Aq_j\rangle q_1 - \cdots - \langle q_{j-1}, Aq_j\rangle q_{j-1}, q_j\rangle q_j||_2} = 0.$$

Aangezien we door herhaaldelijk toepassen van vergelijking (7.1) alle vectoren q_i en kunnen omschrijven tot een lineaire combinatie van $A^k v$ met $k \in \{1, 2, \cdots i\}$ volgt nu dat we een lineaire combinatie van vectoren van de vorm $A^i v$ voor $i \in \{0, 1, \cdots, j\}$ hebben die gelijk is aan nul. Oftewel we hebben nu dat $\gamma \leq j$ omdat we nu een polynoom p van graad j hebben zodat er geldt $p(A)v = 0$. Echter er kan niet gelden $\gamma < j$. Immers als dit wel het geval was zouden we meer dan γ orthonormale basisvectoren hebben gevonden, terwijl we uit Propositie 5.3 weten dat de dimensie van de Krylov deelruimte niet groter kan zijn dan γ↯.

⇐ Stel nu dat de graad van v gelijk is aan j. Als er dan zou gelden $h_{j+1,j} \neq 0$ zouden we in staat zijn om q_{j+1} te definiëren. Echter uit Propositie 5.3 volgt dat de dimensie van $\mathcal{K}_l(A, v) = j$ voor $j \leq l$. Dan zouden we dus $j + 1$ basisvectoren kunnen vinden voor een ruimte van dimensie j ↯. □

Er volgt dus indien er geldt $h_{j+1,j} = 0$ dat j dan de graad is van v in A. Uit Propositie 5.2 volgt er dat $\mathcal{K}_j(A, v)$ een invariante deelruimte is van A. In combinatie met Propositie 5.5 volgt er dat alle Ritz paren ook echte eigenparen zijn van A.

7.3 Convergentie en Nauwkeurigheid

We hebben nu het Arnoldi algoritme om eigenparen van A te benaderen gegeven. Echter weten we nog niets over de algemene nauwkeurigheid van dit algoritme. We hebben in paragraaf 7.2 alleen gezien dat indien de dimensie van de Krylov deelruimte bereikt is, het algoritme dan Ritz paren geeft die ook gelijk zijn aan eigenparen van A.

Als maat voor de convergentie van een Ritz paar, definiëren we opnieuw het residu van $\left(\theta_i^{(k)}, y_i^{(k)}\right)$ als

$$r_i^{(k)} = A x_i^{(k)} - \theta_i^{(k)} y_i^{(k)}. \tag{7.9}$$

Als we nu gebruiken dat $\theta_i^{(k)}$ een eigenwaarde is van H_k en Propositie 7.2 toepassen volgt er

$$A x_i^{(k)} - \theta_i^{(k)} x_i^{(k)} = \left(A Q_k - Q_k \theta_i^{(k)}\right) v_i^{(k)} = \left(A Q_k - Q_k H_k\right) v_i^{(k)} = h_{k+1,k} q_{k+1} \cdot e_k^H \cdot v_i^{(k)}.$$

We zien nu dat er geldt $\|r_i^{(k)}\|_2 = h_{k+1,k} |v_i^{(k)}(k)|$. Aangezien we nu meestal niet met hermitische matrices zullen werken, daarvoor gebruiken we de Lanczos methode, hebben we geen equivalent van Stelling 3.5. Echter indien een convergentie criterium voor een Ritz paar is gewenst, wordt hier meestal de waarde van het residu voor gebruikt. Het residu is immers nog altijd een maat voor de lokale fout van een Ritz paar.

Voor de convergentie van de Arnoldi methode van een enkelvoudige eigenwaarde is het volgende bekend:

Stelling 7.4 (Convergentie Arnoldi methode)**.** *Laat $A \in \mathbb{C}^{n \times n}$ een matrix zijn, waarop we de Arnoldi methode toepassen met als beginvector $x_0 \in \mathbb{C}^n$. Laat v_i een genormeerde eigenvector zijn behorend bij een willekeurige enkelvoudige eigenwaarde λ_i van A. Dan geldt er dat :*

$$\sin(v_i, \mathcal{K}_k(A, x_0)) \leq \frac{1}{|\alpha_i|} \min_{p \in \mathcal{P}_{k-1}, p(\lambda_i)=1} \|p(A)\left(\mathbb{I} - \mathcal{P}\right)\|_2. \tag{7.10}$$

Hierbij is $\mathcal{P}$ de spectrale projector behorend bij eigenwaarden λ_i en volgt α_i uit de vergelijking $\mathcal{P} x_0 = \alpha_i v_i$. In het geval dat A diagonaliseerbaar is met eigendecompositie $A = V \Lambda V^{-1}$ is de bovenstaande grens ook te schrijven als

$$\sin(v_i, \mathcal{K}_k(A, x_0)) \leq \frac{\|V\|_2 \|V^{-1}\|_2}{|\alpha_i|} \min_{p \in \mathcal{P}_{k-1}, p(\lambda_i)=1} \max_{\lambda \in \Lambda(A) \setminus \lambda_i} |p(\lambda)|. \tag{7.11}$$

Indien er geldt $Q_K^H v_i Q_k \neq 0$, bestaat er een Ritz waarde $\theta^{(k)} \in \Lambda(H_k)$ zodat er geldt:

$$|\lambda_i - \theta^{(k)}| \leq 4 \left(\frac{\delta \|A\|_2}{\sqrt{1 - \delta^2}}\right)^{\frac{1}{k}} \left(2\|A\|_2 + \frac{\delta \|A\|_2}{\sqrt{1 - \delta^2}}\right)^{1 - \frac{1}{k}}, \tag{7.12}$$

waarbij geldt $\delta = \sin(v_i, \mathcal{K}_k(A, x_0))$.

Deze stelling is een combinatie van Stelling 28.2 en opmerkingen uit §28 [8]. We herkennen ook Stelling 5.4 in de afschatting van $|\lambda_i - \theta^{(k)}|$. Er volgt uit deze stelling dat v_i gedurende de iteraties niet verder verwijderd kan raken van de Krylov deelruimte. Immers $\min_{p\in\mathcal{P}_{k-1}, p(\lambda_i)=1} ||p(A)(\mathbb{I}-\mathcal{P})||_2$ kan alleen maar afnemen als k toeneemt. We kunnen hieruit geen algemene convergentiesnelheid van het algoritme afleiden, echter kunnen we wel concluderen dat het lastig is om een eigenwaarde te bepalen die in de buurt ligt van andere eigenwaarden. Ook volgt nu niet noodzakelijk dat $|\lambda_i - \theta^{(k)}|$ elke iteratie afneemt. Als $\sin(v_i, \mathcal{K}_k(A,x)) \to 0$ voor $k \to \infty$ zal er ook gelden $|\lambda_i - \theta^{(k)}| \to 0$. Echter als $\sin(v_i, \mathcal{K}_k(A,x))$ nauwelijks afneemt ten opzichte van $\sin(v_i, \mathcal{K}_{k-1}(A,x))$, kan het zijn dat de afname in $\frac{\delta||A||_2}{\sqrt{1-\delta^2}}$ niet opweegt tegen de groter macht voor $2||A||_2 = 2\sqrt{\rho(A^H A)}$ waardoor de fout tussentijds de mogelijkheid heeft om te groeien.

7.4 Variaties op het Arnoldi algoritme

De Arnoldi methode uit algoritme 5 is de standaard variant van het Arnoldi algoritme. Er zijn ook vele variaties op dit algoritme al naar gelang de toepassing, voorbeelden hiervan zijn te vinden in [1, 14]. Een van de redenen waarom men variaties op dit algoritme bekijkt is de benodigde opslag en het benodigde aantal bewerkingen. Immers voor een Krylov deelruimte van dimensie m moeten we een $n \times m$ matrix en een $m \times m$ boven-hessenberg matrix opslaan en is een aantal bewerkingen van orde $\mathcal{O}(m^2(1+n))$ vereist. Indien we slechts k eigenparen willen bepalen, hebben we nu soms een grote dimensie van de Krylov deelruimte nodig om convergentie te bereiken. Een manier om toch convergentie te bereiken voor een kleinere dimensie van de Krylov deelruimte, is om Arnoldi met herstarts uit te voeren. Dit zullen we verder uitwerking in 7.4.1.

Ook zijn we soms niet geïnteresseerd in de eigenwaarden met de grootste modulus, waar de Ritz paren van Arnoldi meestal het eerst naar toe convergeren, maar in de eigenwaarden het dichtst bij de waarde $\sigma \in \mathbb{C}$. Dit kan wederom gerealiseerd worden door toepassing van het algoritme op de shift-and-invert operator. Dit werken we verder uit in paragraaf 7.4.2 [1].

7.4.1 Arnoldi met herstart

De methode die we in deze paragraaf zullen bespreken komt uit [14]. Stel nu dat we de k eigenwaarden van A willen bepalen met de grootste modulus. Een grote dimensie van de Krylov deelruimte kan noodzakelijk zijn voor de convergentie van Ritz waarden naar deze eigenwaarden voor algoritme 5. We nemen hierbij het residu van een Ritz paar als maat voor de convergentie van de Ritz waarden. We weten uit Stelling 7.4 dat de Ritz waarden verbeteren als de sinus van de eigenvector en de Krylov deelruimte, die nooit zal toenemen gedurende de iteraties, afneemt. Het idee is nu de dimensie van de Krylov deelruimte niet groter te laten worden dan m. Indien de Ritz waarden dan nog niet geconvergeerd zijn, beginnen we opnieuw een Arnoldi iteratie. Echter als startvector gebruiken we nu een lineaire combinatie van Ritz vectoren behorend bij de Ritz waarden waarvan we convergentie willen. Dit baseert zijn werking op dat de startvector zo al steeds beter de gewenste richtingen bevat. We sturen eigenlijk richting een invariante deelruimte, immers dan zijn de eigenwaarde die we vinden exact. De nieuwe startvector is dus gegeven door $x_{res} = \frac{\Sigma_{i=1}^k \rho_i y_i^{(m)}}{||\Sigma_{i=1}^k \rho_i y_i^{(m)}||_2}$. Echter een nadeel van deze methode is dat er geen systematisch methode is om ρ_i te bepalen om altijd convergentie te bereiken [14]. Wij zullen twee mogelijkheden voor ρ_i beschouwen:

1. $\rho_i = 1$ voor $i \in \{1, 2, \cdots, k\}$;
2. $\rho_i = \|r_i^{(m)}\|_2$ voor $i \in \{1, 2, \cdots, k\}$.

De eerste methode is de meest eenvoudige lineaire combinatie die we kunnen verzinnen. De tweede methode berust op het idee dat hoe kleiner het residu hoe kleiner de lokale fout en dus des te dichter de Ritz vector bij de Krylov deelruimte zal liggen. De Ritz vectoren die nog het verst van convergentie verwijderd zijn, zal de Krylov deelruimte het meest beïnvloeden en hopelijk dus het meest verbeteren tijdens een Arnoldi iteratie. Het algoritme voor Arnoldi met expliciete herstart is te vinden in algoritme 6.
We zien in regel 2 en 9 de uitvoering van het standaard Arnoldi voor de maximale dimensie m van de Krylov deelruimte. In regel 6 zien we de eis dat alle Ritz waarden geconvergeerd zijn, door een restrictie te zetten op het residu van al deze waarden. Tot slot zien we in regel 7 de twee opties voor de waarden van ρ_i.

1 **def** *Arnoldire(*A, x_0, k, m, tol*)*:

Input : -$A \in \mathbb{C}^{n\times n}$ matrix waarvan we de eigenwaarde en eigenvectoren willen benaderen.
- $x_0 \in \mathbb{C}^n$ de vector waarmee we de eerste Krylov deelruimte opspannen.
-k het aantal te convergeren eigenvectoren.
-m de maximale dimensie van de Krylov deelruimte
-*tol* de gewenste waarde van het residu van elke eigenpaar

Output: -$Q \in \mathbb{C}^{n\times m+1}$ matrix waarvan de kolommen een orthonormale basis van de Krylov deelruimte vormen.
-$H \in \mathbb{C}^{m+1\times m}$ een boven-hessenberg matrix, de projectie van A op de Krylov deelruimte.
-$\Lambda_k \in \mathbb{C}^k$ een vector met de geconvergeerde k Ritz waarden met het grootste reële deel
-$V_k \in \mathbb{C}^{n\times k}$ matrix met de geconvergeerde k Ritz vectoren behorend bij Ritz waarden uit Λ_k als kolommen.

2 $[Q, H] = Arnoldi(A, x_0, m)$
3 Bepaal de eigendecompositie $V\Lambda V^H$ van H
4 Sorteer de eigenwaarde zodat ze aflopen in modulus sorteer de eigenvectoren mee met de eigenwaarden.
5 $V_k = V(:, 1:k)$ en $L_k = (\lambda_1, \lambda_2, \cdots, \lambda_k)$
6 **while** $max_{i\in\{1,2,\cdots,k\}}|h_{m+1,m}V_k(m,i)| \geq tol$**:**
7 $x_{res} = \sum_{i=1}^{k} Q(:, 1:m)V_k(:,i)$ **of** $x_{res} = \sum_{i=1}^{k} ||h_{m+1,m}V_k(m,i)||_2 Q(:, 1:m)V_k(:,i)$
8 $x_{res} = \frac{x_{res}}{||x_{res}||_2}$
9 $[Q, H] = Arnoldi(A, x_{res}, m)$
10 Bepaal de eigendecompositie $H = V\Lambda V^H$ van H
11 Sorteer de eigenwaarde zodat ze aflopen in reële component en sorteer de eigenvectoren mee met de eigenwaarden.
12 $V_k = V(:, 1:k)$ en $L_k = (\lambda_1, \lambda_2, \cdots, \lambda_k)$
13 **end**
14 $V_k = Q(:, 1:m)V_k$

Algoritme 6: Arnoldi methode met expliciete herstart

7.4.2 Shift-and-invert Arnoldi

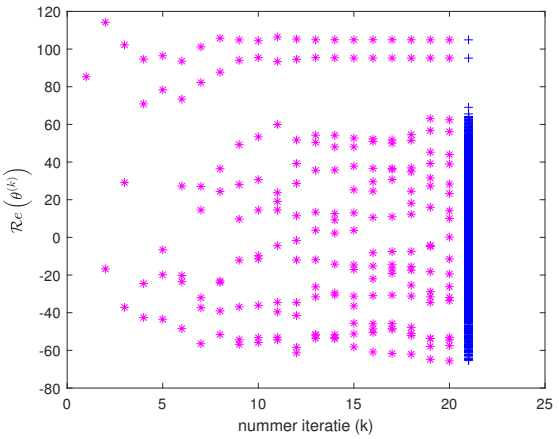
We nemen nu aan geïnteresseerd te zijn in de eigenwaarden van A die het dichtst bij een waarde $\sigma \in \mathbb{C}$ liggen. Hiertoe kunnen we de shift-and-invert operator voor σ op de matrix A toepassen, zoals we ook al deden voor de Lanczos methode (paragraaf 6.5). Immers uit de eigenwaarden van $(A - \sigma\mathbb{I})^{-1}$ met de grootste modulus kunnen we de eigenwaarde het dichtst bij σ bepalen, zie hoofdstuk 4. Dit zijn de eigenwaarden van $(A - \sigma\mathbb{I})^{-1}$ die het eerst zullen convergeren bij toepassing van de Arnoldi methode op deze matrix. [1, §8.1.3]

Om dit toe te kunnen passen moet nog steeds gelden dat $A - \sigma\mathbb{I}$ inverteerbaar is. Als $A - \sigma\mathbb{I}$ niet inverteerbaar is, hebben we gevonden dat σ een eigenwaarde is van A. Om dan de bijbehorende eigenvector en andere eigenwaarden in de buurt van σ te bepalen moeten we de shift-and-invert operator voor een net iets andere shift dan σ toepassen.
Om dit te implementeren hoeven er maar een paar aanpassingen gedaan te worden aan de Arnoldi methode uit algoritme 5:

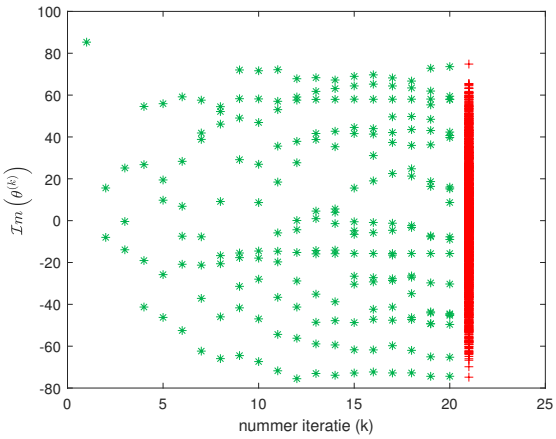
- In regel 5 vervangen we $w = Aq_i$ door het oplossen van het lineaire stelsel $(A - \sigma\mathbb{I})\, w = q_i$ voor w;
- Voeg aan regel 16 toe dat de elementen van Λ, vervangen worden door $\theta_i = \sigma + \frac{1}{\lambda_i}$.

7.5 Numerieke voorbeelden

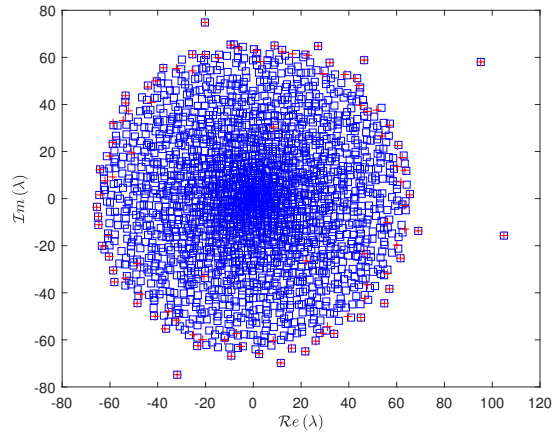
In deze paragraaf bekijken we een twee numerieke voorbeelden van de Arnoldi methode. We beginnen met een voorbeeld van de Arnoldi methode uit algoritme 5.



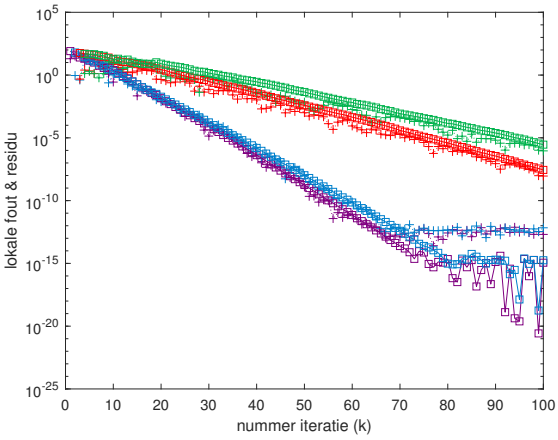
(a) Het reële deel (*) van $\theta_i^{(k)}$ voor de eerste 20 iteraties van de Arnoldi methode en het reële deel van de eigenwaarden(+) van N_1.



(b) Het imaginaire deel (*) van $\theta_i^{(k)}$ voor de eerste 20 iteraties van de Arnoldi methode en het imaginaire deel van de eigenwaarden(+) van N_1.



(c) De echte eigenwaarden (□) en de Ritz waarden (+) na 100 iteraties van Arnoldi voor N_1 in het complexe vlak.

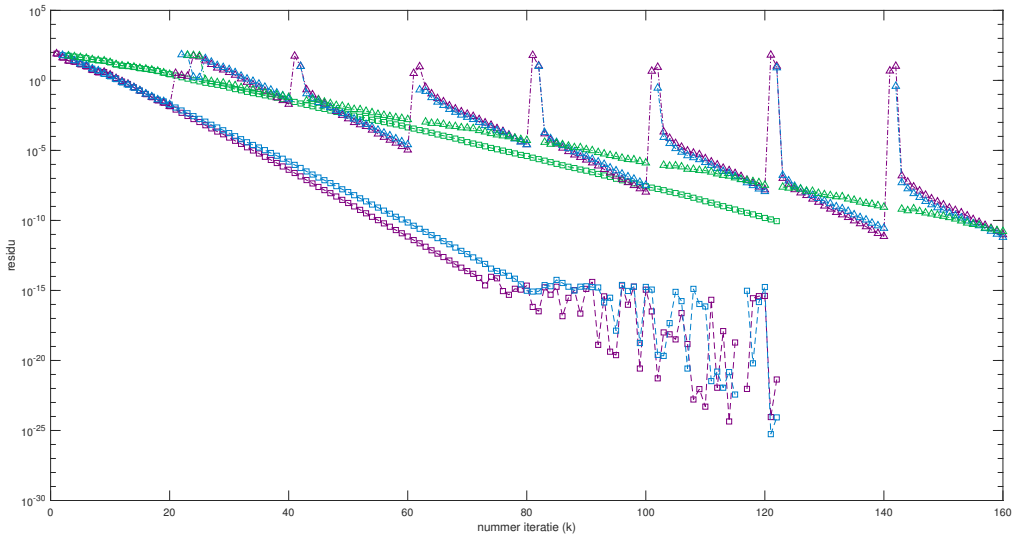


(d) De lokale fout (+) en het residu (□) van de vier Ritz waarden met de grootste modulus.

Figuur 9: In deze figuren zien we de waarde van $\theta_i^{(k)}$ voor $k \in \{1, 2, \cdots, 20, 100\}$ uit de Arnoldi iteratie en de eigenwaarde van N_1. Ook zien we de lokale fout en het residu van de vier Ritz waarden met de grootste modulus.

Voorbeeld 7.5. In dit voorbeeld bekijken de matrix $N_1 \in \mathbb{C}^{2000\times2000}$ gevormd als $N_1 = 60 * \frac{R}{||R||_2} * D$ met $R \in \mathbb{C}^{2000\times2000}$ een uniforme matrix en $D \in \mathbb{C}^{2000\times2000}$ een Gauss $(0, 60)$ diagonaalmatrix. Hierop voeren we de Arnoldi methode, algoritme 5, uit met een uniforme startvector $w_1 \in \mathbb{C}^{2000}$ gedurende 100 iteraties. In de figuren 9a en 9b zijn voor de eerste 20 iteraties het reële en imaginaire deel van alle Ritz waarden te zien. Tevens zijn alle eigenwaarden van N_1 weergegeven (+). We zien hierin niet een duidelijk patroon van de convergentie. Dit komt doordat de Ritz waarden bij de Arnoldi methode ongeveer convergeren op volgorde van de modulus. Hierdoor kan een eigenwaarden met niet het grootste reële/imaginaire deel eerder convergeren omdat de modulus in totaal wel groter is. Echter zien we niet hoe de reële en imaginaire delen van de Ritz waarden aan elkaar verbonden zijn. Daarom hebben we in figuur 9c de uiteindelijke Ritz (+) waarden in het complexe weergeven samen met de eigenwaarden (□) van N_1. We zien nu inderdaad dat de Ritz waarden vooral naar de eigenwaarden met een grote modulus convergeren. In figuur 9d zien we de lokale fout $\min_{\lambda\in\Lambda(N_1)} \left|\lambda - \theta_i^{(k)}\right|$ en het residu voor de vier Ritz waarden met de grootste modulus per iteratie. We zien hierbij dat de Ritz waarden met een grotere modulus eerder convergeren. Ook zien we dat de Ritz waarden (blauw en paars) die verder weg zitten van de grote cluster van eigenwaarden sneller convergeren dan de Ritz waarden (blauw en paars) die dichterbij de cluster zitten. Het residu blijkt hier een bovengrens op de lokale fout te zijn, tot het bereiken van de machine precisie. △

We bekijken nu voor dezelfde matrix hoe snel algoritme 6 convergeert voor de verschillende keuzen van ρ_i.



Figuur 10: Het residu van de drie Ritz waarden met de grootste modulus voor standaard Arnoldi (□) en Arnoldi met herstart ($\rho_i = ||r_i^{(m)}||_2$, △) totdat het residu voor al deze Ritz waarden kleiner is dan 10^{-10}.

Voorbeeld 7.6. We bekijken opnieuw de matrix N_1 uit voorbeeld 7.5. We passen hier nu de Arnoldi met herstart, algoritme 6 op toe voor $m = 20$, $k = 3$, $tol = 10^{-10}$ en dezelfde startvector w_1 als uit voorbeeld 7.5. We doen dit voor beide keuzen voor ρ_i, oftewel $\rho_i = 1$ en $\rho_i = ||r_i^{(m)}||_2$. Er blijken in beide gevallen 8 herstarts nodig te zijn om de gewenste convergentie te bereiken. Ter vergelijking bepalen we ook de benodigde dimensie van de Krylov deelruimte om het residu van deze drie Ritz waarden kleiner dan 10^{-10} te krijgen voor standaard Arnoldi. Deze dimensie blijkt gelijk te zijn aan 122. In figuur 10 is de waarde van het residu voor de standaard Arnoldi (□) en voor de Arnoldi met hertstart voor $\rho_i = ||r_i^{(m)}||_2$ (△) weergeven voor deze drie Ritz waarden. Hierbij is de herstart met $\rho_i = 1$ niet weergegeven omdat op deze schaal de beide herstarts nauwelijks te onderscheiden zijn. Bij de standaard Arnoldi is k gelijk aan de dimensie van de Krylov deelruimte. Voor de herstart is de dimensie op $1 + k20$ weer gelijk aan 1 en zijn we in de k^{de} herstart beland voor $k \in \mathbb{N}$. De pieken het residu voor de herstart komt doordat het residu voor een Krylov deelruimte van kleine dimensie hoog is, immers als $v_i^{(k)}$ genormaliseerd is en k is klein zullen de componenten groter zijn.

De gaten komen doordat pas voor een Krylov deelruimte van dimensie drie en hoger alle drie de residuen bestaan. We zien nu dat er minder iteraties nodig zijn voor de standaard Arnoldi dan voor Arnoldi met herstart. Ook de grootste twee Ritz waarden met standaard Arnoldi eerder convergeren dan met herstart. Echter als de rekentijd voor beide algoritmen vergelijken blijkt er te gelden de herstart maar 64% van de looptijd van de standaard Arnoldi nodig heeft. Dat Arnoldi met herstart toch sneller is, komt doordat voor een kleinere dimensie de rekentijd zodanig korter is dat iets meer iteraties bij een kleinere dimensies opweegt tegen minder iteraties voor grotere dimensies. △

We zien nu dus geen noemenswaardig verschil in de beide keuzen voor de ρ_i voor Arnoldi met herstart. We zien wel dat deze methode inderdaad de benodigde rekentijd en opslag voor de convergentie van een aantal Ritz waarden sterk kan verminderen.

8 Conclusie

Afhankelijk van welke eigenparen van een matrix A we willen benaderen, hebben we een methode om een deel van het spectrum van A te benaderen. Indien we de eigenwaarde met de grootste modulus willen bepalen volstaat de machtsmethode, zie hoofdstuk 3. Echter moet er wel één eigenwaarde met de grootste modulus bestaan om convergentie te verkrijgen. Als we opzoek zijn naar de eigenwaarde het dichtst bij $\sigma \in \mathbb{C}$ ligt, wenden we ons tot de inverse iteratie, zie hoofdstuk 4. Omdat deze methode werkt door de machtsmethode toe te passen op een shift-and-invert operator, is het noodzakelijk dat het maximum $\frac{1}{|\lambda_1 - \sigma|}$ maar voor één eigenwaarde $\lambda_1 \in \Lambda(A)$ wordt aangenomen om convergentie mogelijk te maken. Als er geldt $|\lambda_1| = |\lambda_2| = \cdots = |\lambda_j|$ blijkt voor een normale matrix de machtsmethode een convergente lineaire combinatie van deze eigenwaarde te geven als benadering van de eigenwaarde. Echter de eigenvector benadering, die geen eigenvector is, convergeert nu niet, waardoor het algoritme zelf ook niet convergeert. We kunnen alleen zeggen dat deze eigenvector zich in de ruimte opgespannen door de eigenvectoren van λ_1 tot en met λ_j bevindt. Indien we meerdere eigenparen van een matrix willen benaderen of het zojuist genoemde probleem willen omzeilen wenden we ons tot methoden gebaseerd op de toepassing van de Rayleigh-Ritz procedure op Krylov deelruimten. Voor een hermitische matrix gebruiken we de Lanczos methode, zie hoofdstuk 6, en anders wenden we ons tot de Arnoldi methode, zie hoofdstuk 7. Deze methoden zijn ook te gebruiken om de eigenwaarden het dichts bij $\sigma \in \mathbb{C}$ te bepalen door toepassing van de methode op de shift-and-invert operator voor σ. De convergentie naar een eigenwaarde wordt bij al deze methoden beïnvloedt door de afstand tussen de desbetreffende eigenwaarde en de overige eigenwaarden. Voor de Arnoldi methode kunnen we ook Arnoldi met herstart toepassen om de convergentie naar een paar eigenwaarden te versnellen. Bij de Lanczos methode nemen we het merkwaardige gedrag waar dat zonder herorthogonalisatie de norm van $||Q_i^H Q_i - \mathbb{I}||_2$ verhoogt met één indien vlak voordat er een kopie ontstaat van de Ritz waarde voor de eigenwaarde met de grootste modulus van A. Voor de oorzaak van dit fenomeen hebben we wel een vermoeden, zie in paragraaf 6.4, echter dient dit nog verder uitgezocht te worden.

Referenties

[1] Y. Saad, *Numerical methods for large eigenvalue problems* (Siam, 2011), 2nd ed., ISBN 978-1-611970-72-2.

[2] D. J. Griffiths, *Introduction to Quantum Mechanics*, Pearon New International Edition (Pearson, 2014), 2nd ed., ISBN 978-1-292-02408-0.

[3] J. W. Demmel, *Applied Numerical Linear Algebra* (Siam, 1997), 3rd ed., ISBN 0-89871-389-7.

[4] S. Axler, *Linear ALgebra Done Right*, Undergraduate Texts in Mathematics (Springer, Cham, 2015), 3rd ed., ISBN 978-3-319-11080-6.

[5] P. Igodt and W. Veys, *Lineaire algebra* (Universitaire Pers Leuven, 2015), 2nd ed., ISBN 978 94 6270 052 9.

[6] S. Börm and C. Mehl, *Numerical Methods for Eigenvalue Problems* (De Gruyter, Inc, 2012), ISBN 9783110250374.

[7] Y. Nakatsukasa, Linear Algebra and its Applications **432**, 242 (2010), ISSN 0024-3795, URL `http://www.sciencedirect.com/science/article/pii/S0024379509004078`.

[8] L. N. Trefethen and M. Embree, *Spectra and pseudopspectra: the behavior of nonnormal matrices and operators* (Princeton University Press, 2005), ISBN 0691119465 9780691119465.

[9] L. N. Trefethen, Acta Numerica **8**, 247295 (1999).

[10] G. Stewart and Z. Jia, Mathematics of computation **70**, 637 (2000).

[11] Y. Ikebe, T. Inagaki, and S. Miyamoto, The American Mathematical Monthly **94**, 352 (1987).

[12] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Patrical Guide* (Siam, 2000), 3rd ed., ISBN 9780898714715.

[13] L. S. J., B. ke, and G. Walter, Numerical Linear Algebra with Applications **20**, 492 (2012), URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/nla.1839`.

[14] Y. Saad, Tech. Rep. 88, Research Institute for Advanced Computer Science (1998).

A Code

In deze appendix geven we de implementatie van de algoritmen gebruikt om de numerieke voorbeelden te creëren. Deze code is geïmplementeerd in Matlab versie versie R2017b (9.3.0.713579).

A.1 Machtsmethode

```
function [lm,v,res] = machtit(A, y,tol)
% Input:
% -A in C^(n x n) een matrix
% - y in C^n een startvector
% - tol de gewenste waarde van het residu
% Output:
% -lm een vector met op positie i de benadering van de eigenwaarde
% met grootste modulus van A na i iteraties
% -v een vector met op kolom i de benadering van de eigenvector behorend bij lm(i)
% -res een vector met op positie i de waarde van het residu op die iteratie
% Deze functie voert de machtsmethode uit voor matrix A en startvector y totdat
% het residu  kleiner is dan tol. Het verloop van lm
% gedurende de iteraties wordt ook geplot.
i=1;
x=y;
v(:,i)=x*(1/norm(x));
x=A*v(:,i);
lm(i)=v(:,i)'*x;
res(i)=norm(x-lm(i)*v(:,i));
while res(i)>tol
        i=i+1;
        v(:,i)=x*(1/norm(x));
        x=A*v(:,i);
        lm(i)=v(:,i)'*x;
        res(i)=norm(x-lm(i)*v(:,i));
end
if isreal(lm)==1
        plot(lm,'m x')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
else
        pltr=real(lm);
        plti=imag(lm);
        plot(pltr,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
        hold on;
        plot(plti, 'o','Color',[0 0 1])
        hold off;
end

function [lm,v] = powerittoll(A,y,tol,it)
% Input:
%  -A in C^(n x n) een matrix
% - y in C^n een startvector
% -tol de gewenste waarde van het verschil tussen lm(i) en lm(i-1)
% -it het aantal gewenste iteraties na stabilisatie van de benadering van de eigenwaarde.
```

```
% Output:
% -lm een vector met op positie i de benadering van de eigenwaarde
% met grootste modulus van A na i iteraties
% -v een vector met op kolom i de benadering van de eigenvector behorend bij lm(i)
% Deze functie voert de machtsmethode uit voor matrix A en startvector y
% totdat de eigenwaarde benadering lm niet meer dan tol verschilt van
% die van de vorige iteratie. Daarna worden nog it
% iteraties uitgevoerd. Het verloop van lm wordt ook geplot.
i=1;
x=y;
v(:,i)=x*(1/sqrt(x'*x));
x=A*v(:,i);
lm(i)=v(:,i)'*x;
i=2;
v(:,i)=x*(1/sqrt(x'*x));
x=A*v(:,i);
lm(i)=v(:,i)'*x;
while norm(lm(i)-lm(i-1))>tol
        i=i+1;
        v(:,i)=x*(1/sqrt(x'*x));
        x=A*v(:,i);
        lm(i)=v(:,i)'*x;
end
for j = 1:it
        i=i+1;
        v(:,i)=x*(1/sqrt(x'*x));
        x=A*v(:,i);
        lm(i)=v(:,i)'*x;
end
if isreal(lm)==1
        plot(lm,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
else
        pltr=real(lm);
        plti=imag(lm);
        plot(pltr,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
        hold on;
        plot(plti,'b o')
        hold off;
end
```

A.2 Inverse iteratie

```
function [mu,lm,v,res] = invit(A,y,sig,tol)
% Input:
%- A in C^(n x n) een matrix
%- y in C^n een startvector
%- sig de waarde van de shift
% -tol de gewenste waarde van het residu
% Output:
% -mu  een vector met de benadering van de grootste eigenwaarde van de
% shift-and-inverrt matrix na i iteraties op positie i.
```

```
% -lm een vector met benadering van de eigenwaarde van A
% het dichtst bij sig na i iteraties op positie i
% -v een vector met op kolom i de benadering van de eigenvector behorend bij lm(i)
% Deze functie voert de inverse iteratite uit voor matrix A, startvector y
% en shift sig met als convergentiecriterium dat het residu kleiner is dan
% tol. Ook wordt het verloop van lm en mu  geplot.
[~,m]=size(A);
i=1;
x=y;
[L,U,P]=lu(A-sig*eye(m,m));
v(:,i)=x/norm(x);
x=U\(L\(P*v(:,i)));
mu(i)=v(:,i)'*x;
lm(i)=sig+(1/mu(i));
res(i)= norm(x-mu(i)*v(:,i));
while res(i)>tol
i=i+1;
v(:,i)=x*(1/norm(x));
x=U\(L\(P*v(:,i)));
mu(i)=v(:,i)'*x;
lm(i)=sig+(1/mu(i));
res(i)=norm(x-mu(i)*v(:,i));
end
if isreal(lm)==1
        plot(lm,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
else
        lmr=real(lm);
        lmi=imag(lm);
        plot(lmr,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\theta_k')
        hold; plot(lmi,'b o')
end
figure;
if isreal(mu)==1
        plot(mu,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\mu_k')
else
        mur=real(mu);
        mui=imag(mu);
        plot(mur,'m *')
        xlabel('nummer iteratie (k)')
        ylabel('\mu_k')
        hold; plot(mui,'b o')
end
```

A.3 Lanczos methode

```
function [Q,T,eigval]=Lanczosn(A,v,k)
%Input:
% -A in C^(n x n) een hermitische matrix
% -v in C^n een vector om de Krylov deelruimte mee op te spannen
```

```
% - k een natuurlijk getal (groter dan nul), de dimensie van de Krylov deelruimte
%Output:
% -Q in C^(n x k+1) een  matrix met de orthonormale basis van de Krylov deelruitme
% als kolommen
% -T in C^(k+1 x k+1) een tridiagonale matrix, de projectie van A op de Krylov deelruimte
% -eigval in C^(k bij k)  matrix met op kolom i en rij 1:i de Ritz
% waarden voor Krylov deelruimte van dimensie i
%Deze functie voert een Lanczos methode uit voor de matrix A met
%startvector v en k iteraties. Er is hierbij geen sprake van
%herorthagonalizatie.
[~,n]=size(A);
q1=v/norm(v);
Q=zeros(n,k+1);
Q(:,2)=q1;
a=zeros(n,1);
b=zeros(n,1);
eigval=zeros(k,k);
T=sparse(k,k);
for i = 1:k
        w=A*Q(:,i+1);
        a(i)=(Q(:,i+1))'*w;
        w=w-a(i)*Q(:,i+1)-b(i)*Q(:,i);
        b(i+1)=norm(w);
        if b(i+1)==0
                break
        end
        Q(:,i+2)=(w/(b(i+1)));
        T(i,i)=a(i);
        T(i,i+1)=b(i+1);
        T(i+1,i)=b(i+1);
        eigva=eig(full(T(1:i,1:i)));
        eigval(1:i,i)=eigva;
end
Q=Q(:,2:end)

function [Q,T,eigval]=Lanczosf(A,v,k)
%Input:
% -A in C^(n x n)een hermitische matrix
% -v in C^n een vector om de Krylov deelruimte mee op te spannen
% - k een natuurlijk getal (groter dan nul), de dimensie van de Krylov deelruimte
%Output:
% -Q in C^(n x k+1) een matrix met de orthonormale basis van de Krylov
% deelruitme als kolommen
% -T in C^(k+1 x k+1) een tridiagonale matrix, de projectie van A op de Krylov deelruimte
% -eigval in C^(k x k) een matrix met op kolom i en rij 1:i de Ritz
% waarden voor Krylov deelruimte dimensie i
%Deze functie voert een Lanczos methode  met volledige herorthogonalisatie uit voor
% de matrix A met startvector v en k de dimensie van de Krylov deelruimte.
[~,n]=size(A);
q1=v/norm(v);
Q=zeros(n,k+1);
Q(:,2)=q1;
a=zeros(n,1);
b=zeros(n,1);
```

```
eigval=zeros(k,k);
T=sparse(k,k);
for i = 1:k
        w=A*Q(:,i+1);
        a(i)=(Q(:,i+1))'*w;
        w=w-a(i)*Q(:,i+1)-b(i)*Q(:,i);
        w=w-Q(:,2:i)*((Q(:,2:i)')*w);
        b(i+1)=norm(w);
        if b(i+1)==0
                break
        end
        Q(:,i+2)=(w/(b(i+1)));
        T(i,i)=a(i);
        T(i,i+1)=b(i+1);
        T(i+1,i)=b(i+1);
        eigva=eig(full(T(1:i,1:i)));
        eigval(1:i,i)=eigva;
end
Q=Q(:,2:end)

function [Q,T,eigval,orth,res]=Lanczosso(A,v,k)
%Input:
% -A in C^(n x n) een hermitische matrix
% -v in C^n  vector om de Krylov deelruimte mee op te spannen
% - k een natuurlijk getal (groter dan nul), de dimensie van de Krylov deelruimte
%Output:
% -Q in C^(n x k+1) een matrix met de orthonormale basis van de Krylov deelruitme als kolom
% -T in C^(k+1 x k+1) een tridiagnoale matrix, de projectie van A op de Krylov deelruimte
% -eigval in C^(k x k) een matrix met op kolom i en rij 1:i de Ritz
% waarden voor Krylov deelruimte dimensie i
%-orth een matrix met twee rijen met op de eerste rij het nummer van de
%iteratie en op de tweede rij het nummer van de Ritz waarde die voor
%herorthogonalisatie is geselecteerd.
%-res in C^(k x k) een matrix met op kolom i en rij 1:i het residu voor
% de Ritz paren voor de Krylov deelruimte van dimensie i
%Deze functie voert een Lanczos methode  met selectieve herorthogonalisatie uit
%voor de matrix A met startvector v met k de dimensie van de Krylov deelruimte.
[~,n]=size(A);
q1=v/norm(v);
Q=zeros(n,k+1);
Q(:,2)=q1;
a=zeros(n,1);
b=zeros(n,1);
eigval=zeros(k,k);
res=zeros(k,k);
T=zeros(k,k);
orth=[];
for i = 1:k
        w=A*Q(:,i+1);
        a(i)=(Q(:,i+1))'*w;
        w=w-a(i)*Q(:,i+1)-b(i)*Q(:,i);
        if i >1
                for j = 1:i-1
                        if b(i)*abs(eigve(i-1,j)) <= sqrt(eps)*norm(T(1:i-1,1:i-1))
```

```
                                        orth=[orth,[i;j]];
                                        y=Q(:,2:i)*eigve(:,j);
                                        w=w-(y'*w)*y;
                                end
                        end
                end
        end
        b(i+1)=norm(w);
        if b(i+1)==0
                break
        end
        Q(:,i+2)=(w/(b(i+1)));
        T(i,i)=a(i);
        T(i,i+1)=b(i+1);
        T(i+1,i)=b(i+1);
        [eigve,eigva]=eig(full(T(1:i,1:i)));
        eigval(1:i,i)=diag(eigva);
        res(1:i,i)= b(i+1)*eigve(i,1:i);
end
Q=Q(:,2:end)
```

A.4 Arnoldi methode

```
function [Q,H,eigval,res,eigvec]=Arnoldi(A,v,k)
%Input:
% -A in C^(n x n) een matrix
% -v in C^n een vector om de Krylov deelruimte mee op te spannen
% - k een natuurlijk getal (groter dan nul), de dimensie van de Krylov deelruimte
%Output:
% -Q in C^(n x k+1) een matrix met de orthonormale basis
% van de Krylov deelruitme als kolommen
% -H in C^(k+1 x k) een boven-hessenbergmatrix, de projectie van A
% op de Krylov deelruimte
% -eigval in C^(k x k) een matrix met op kolom i en rij 1:i de Ritz
% waarden voor Krylov deelruimte dimensie i geordend op aflopende modulus
%- res in C^(k x k) een matrix met op kolom i en rij 1:i het residu van
% de Ritz paren voor een Krylov deelruimte van dimensie i
%-eigvec in C^(n x k) een matix met de Ritz vectoren als kolommen
%Deze functie voert een standaard Arnoldi methode toe op de matrix A met
%beginvector v met k iteraties.
[~,n]=size(A);
q1=v/norm(v);
Q=zeros(n,k+1);
Q(:,1)=q1;
H=zeros(k,k);
eigval=zeros(k,k);
res=zeros(k,k);
for j = 1:k
        w=A*Q(:,j);
        for i=1:j
                H(i,j)=(Q(:,i)')*w;
                w=w-H(i,j).*Q(:,i);
        end
        H(j+1,j)=norm(w);
        if H(j+1,j)==0
                break
```

```
        end
        Q(:,j+1)=(w/(H(j+1,j)));
        [eigvec,eigv]=eig(H(1:j,1:j));
        eigv=diag(eigv);
        [l,~]=size(eigv);
        [eigval(1:l,j),I]=sort(eigv,'descend','ComparisonMethod','abs');
        eigvec=eigvec(:,I);
        res(1:l,j)=H(j+1,j)*abs(eigvec(j,1:j).');
end

function [Q,H,eigval,res,eigvec]=Arnoldicon(A,v,m,tol)
%Input:
% -A in C^(n x n) een  matrix
% -v in C^n een vector om de Krylov deelruimte mee op te spannen
% - m  het aantal Ritz waarden met de grootste modulus waarvan we convergentie willen
%-tol de waarde van het residu om een Ritz paar als geconvergeerd te
%beschouwen.
%Output:
% -Q een matrix met de orthonormale basis van de Krylov deelruitme als kolommen
% -H een boven-hessenbergmatrix, de projectie van A op de Krylov deelruimte
% -eigval een matrix met op kolom i op rij 1:i de Ritz
% waarden voor Krylov deelruimte dimensie i geordend op aflopende modulus
%- res een  matrix met op kolom i op rij 1:i het residu van de Ritz
% paren voor een Krylov deelruimte van dimensie i
%-eigvec een matix met de Ritz vectoren als kolommen
%Deze functie voert een standaard Arnoldi methode toe op de matrix A met
%beginvector v totdat het residu van de m Ritz waarden
% met de grootste modulus kleiner is dan tol.
q1=v/norm(v);
Q(:,1)=q1;
for j = 1:m
        w=A*Q(:,j);
        for i=1:j
                H(i,j)=(Q(:,i)')*w;
                w=w-H(i,j).*Q(:,i);
        end
        H(j+1,j)=norm(w);
        if H(j+1,j)==0
                break
        end
        Q(:,j+1)=(w/(H(j+1,j)));
        [eigvec,eigv]=eig(H(1:j,1:j));
        eigv=diag(eigv);
        [l,~]=size(eigv);
        [eigval(1:l,j),I]=sort(eigv,'descend','ComparisonMethod','abs');
        eigvec=eigvec(:,I);
        res(1:l,j)=H(j+1,j)*abs(eigvec(j,1:j).');
end
con=abs(res(1:m,m));
while max(con)>=tol
        j=j+1;
        w=A*Q(:,j);
        for i=1:j
                H(i,j)=(Q(:,i)')*w;
```

```
                w=w-H(i,j).*Q(:,i);
        end
        H(j+1,j)=norm(w);
        if H(j+1,j)==0
                break
        end
        Q(:,j+1)=(w/(H(j+1,j)));
        [eigvec,eigv]=eig(H(1:j,1:j));
        eigv=diag(eigv);
        [l,~]=size(eigv);
        [eigval(1:l,j),I]=sort(eigv,'descend','ComparisonMethod','abs');
        eigvec=eigvec(:,I);
        res(1:l,j)=H(j+1,j)*abs(eigvec(j,1:j).');
        con = abs(res(1:m,j));
end

function [Q,H,vk,lk,eigval,it,res,l]=Arnoldiher(A,v,m,tol,k)
%Input:
% -A in C^(n x n) een  matrix
% -v in C^n een vector, startvector Krylov deelruimte eerste iteratie
% - m een positief geheel getal, de maximale dimensie van de Krylov deelruimte
% -tol een waarde waar het residu van een Ritz paar kleiner dan
% moet zijn om als geconvergeerd beschouwd te worden
%-k het aantal Ritz paren met de grootste modulus waarvan we convergentie wensen
%Output:
% -Q in C^(n x m+1) een matrix met de orthonormale basis van de Krylov deelruimte
% als kolommen van de laatste Arnoldi procedure
% -H in C^(m+1 x mk) een boven-hessenbergmatrix, de projectie van A
% op de Krylov deelruimte van de laatste Arnoldi procedure
% -vk de geconvergeerde gewenste Ritz vectoren
% -lk de geconvergeerde gewenste Ritz waarden
% -eigval in C^(m x m) een matrix met op rij i de Ritz waarden van
% iteratie i  van de laatste Arnoldi procedure
% -it het aantal herstart van de Arnoldi methode
% -res een matrix met op kolom i het residu van de te convergeren Ritz waarden
% -l geeft met de kolomen het verloop van de te convergeren Ritz waarden aan
%Deze functie voert Arnoldi met herstart  uit voor de matrix A met
%beginvector v met k te convergeren Ritz waarden een een maximale
%dimensie van m voor de Krylov deelruimte. De rho_i's zijn hier gelijk aan 1.
[n,~]=size(A);
[Q,H,eigval,resi,eigv]=Arnoldincon(A,v,m);
vk=Q(:,1:end-1)*eigv(:,1:k);
lk=eigval(1:k,end);
l=eigval(1:k,:);
res=resi(1:k,:);
it=1;
con=max(abs(res(:,end)));
while con>=tol
        it=it+1;
        vre=zeros(n,1);
        for i = 1:k
                vre=vk(:,i)+vre;
        end
        [Q,H,eigval,resi,eigv]=Arnoldincon(A,vre,m);
```

```
        vk=Q(:,1:end-1)*eigv(:,1:k);
        lk=eigval(1:k,end);
        res=[res,resi(1:k,:)];
        l=[l,eigval(1:k,:)];
        con=max(abs(res(:,end)));
end

function [Q,H,vk,lk,eigval,it,res,l]=Arnoldiherres(A,v,m,tol,k)
%Input:
% -A in C^(n x n) een matrix
% -v in C^n een  vector, startvector eerste iteratie Krylov deelruimte
% - m een geheel getal m>=1, de maximale dimensie van de Krylov deelruimte
% -tol een waarde waar het residu van een Ritz paar kleiner dan
% moet zijn om als geconvergeerd beschouwd te worden
%-k het aantal Ritz paren met de grootste modulus waarvan we convergentie wensen
%Output:
% -Q in C^(n x m+1) een matrix met de orthonormale basis van de Krylov deelruimte
% als kolommen van de laatste Arnoldi procedure
% -H in C^(m+1 x m) een boven-hessenbergmatrix, de projectie van A
% op de Krylov deelruimte van de laatste Arnoldi procedure
% -vk de geconvergeerde gewenste Ritz vectoren
% -lk de geconvergeerde gewenste Ritz waarden
% -eigval in C^(m x m) een matrix met op rij i de Ritz waarden van
% iteratie i  van de laatste Arnoldi procedure
% -it het aantal herstart van de Arnoldi methode
% -res een matrix met op kolom i het residu van de te convergeren Ritz waarden
% -l geeft met de kolomen het verloop van de te convergeren Ritz waarden aan
%Deze functie voert Arnoldi met herstart  uit voor de matrix A met
%beginvector v met k te convergeren Ritz waarden een een maximale
%dimensie van m voor de Krylov deelruimte. De rho_i's zijn hier gelijk aan
%het residu van het Ritz paar.
[n,~]=size(A);
[Q,H,eigval,resi,eigv]=Arnoldincon(A,v,m);
vk=Q(:,1:end-1)*eigv(:,1:k);
lk=eigval(1:k,end);
l=eigval(1:k,:);
res=resi(1:k,:);
it=1;
con=max(abs(res(:,end)));
while con>=tol
        it=it+1;
        vre=zeros(n,1);
        for i = 1:k
                vre=vk(:,i).*res(i,end)+vre;
        end
        [Q,H,eigval,resi,eigv]=Arnoldincon(A,vre,m);
        vk=Q(:,1:end-1)*eigv(:,1:k);
        lk=eigval(1:k,end);
        res=[res,resi(1:k,:)];
        l=[l,eigval(1:k,:)];
        con=max(abs(res(:,end)));
end
```