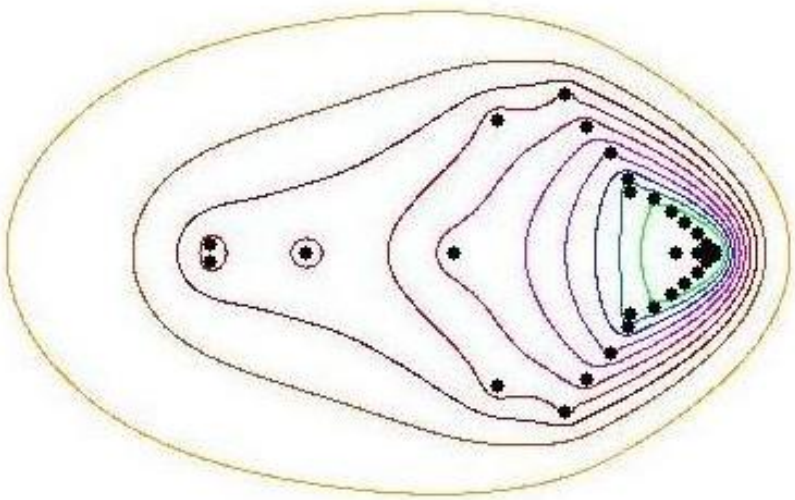
Pseudospectra en stabiliteit van Runge-Kuttamethoden

Mick Frido Kahmann

20 augustus 2008



Voorwoord

Dit verslag is gemaakt in het kader van het bachelorproject van de studie Technische Wiskunde aan de Technische Universiteit te Delft in de vakgroep Analyse. Het is voornamelijk ontwikkeld zodanig dat een derdejaars student Technische Wiskunde kan begrijpen wat er wordt besproken. Dit betekent zeker niet dat dit verslag oninteressant is voor de meer ervaren wiskundige; ik ken zelfs een aantal wiskundigen die niet bekend zijn met het eerste onderwerp van mijn onderzoek: pseudospectra. Het is een begrip dat nog vrij nieuw is. Pas sinds de jaren 90 word er door een selecte groep wiskundigen intensief onderzoek gedaan naar pseudospectra en zijn toepassingen.

In tegenstelling tot vele medestudenten had ik namelijk voor een analyse-onderzoek gekozen en deze vakgroep word onder mijn medestudenten toch als een meer fundamenteel wiskundig vak gezien. Voor ons 'toegepaste' wiskundigen klinkt dit toch wel een beetje eng. Toen ik aan dit onderzoek begon was ik er dus nog niet van onvertuigd dat ik een leuk onderwerp had gekozen. Het tegendeel is waar gebleken, want al gauw vatte het onderwerp van pseudospectra mij bij de kraag en ik heb dan ook met veel plezier en soms ook bewondering gewerkt aan dit onderzoek.

Bij alle voorbeelden die worden gegeven is gebruik gemaakt van het wiskundig ondersteunde programma MATLAB. Hoewel ik al voldoende bekend was met het programmeren in MATLAB, heb ik nog veel meer erbij geleerd. Dit soms na vele uren frustratie; sommige theorieën hadden meerdere 'aha'-erlebnissen nodig om de code goed te laten werken.

In dit verslag tracht ik de theorie concreet te maken met behulp van veel voorbeelden. Ook wil ik hiermee een bruggetje maken van de analyse naar de numerieke wiskunde.

Ten slotte wil ik graag een aantal mensen bedanken voor hun hulp bij mijn onderzoek. Als eerste wil ik Valérie Frayssé van het `Pseudospectra Gateway` bedanken voor haar e-mails aan het begin van mijn onderzoek. Ook Martin van Gijzen wil ik bedanken voor zijn hulp bij het numerieke gedeelte van dit onderzoek. In het bijzonder bedank ik mijn begeleider Birgit Jacob voor alle steun en hulp bij het tot stand komen van dit verslag.

Inhoudsopgave

1 Inleiding

Technisch wiskundigen komen in de praktijk problemen tegen waarbij we onze wiskundige kennis willen gebruiken om bijvoorbeeld een grafische weergave te geven van de situatie. Als ergens in de Noordzee een schip olie lekt, dan willen we weten waar deze olie terecht komt aan de kust. Als we in een ruimte met kamertemperatuur een warmtebron neerzetten, dan zijn we misschien geïnteresseerd in het temperatuurverloop op meerdere plaatsen in deze ruimte of juist in de stroming van de hitte. Veel van deze problemen kunnen we reduceren tot partiële differentiaalvergelijkingen. Helaas kunnen veel van deze problemen niet exact worden opgelost en daarom gebruiken we numerieke wiskunde om hiervan stelsels gewone differentiaalvergelijkingen te maken. Deze kunnen we dan oplossen met eenstapsmethoden zoals de Runge-Kuttamethoden.
Het oplossen van dit soort problemen kan misschien wel zo belangrijk dat er dierenlevens op het spel staan of dat het milieu ervan achteruit kan gaan. Dan is het ook belangrijk dat de numerieke oplossing dichtbij de werkelijkheid ligt. Er zijn welbekende methoden om te testen of een gekozen tijdstap voldoende klein is zodanig dat we een stabiel probleem hebben. Hierbij kijken we dan altijd naar het spectrum van de matrix die uit het probleem voortvloeit, maar er zijn problemen waarbij het spectrum niet kan verklaren waarom een oplossing instabiel is. De goed verborgen theorie hierachter is het principe van pseudospectra. Door het lezen van dit verslag zal duidelijk worden wat we bedoelen met het ε-pseudospectrum en uiteindelijk word aan de hand van een zelfbedacht voorbeeld uitgelegd wat voor effect de pseudospectra kunnen hebben op de stabiliteit van problemen die we oplossen met Runge-Kuttamethoden.

Dit verslag is grofweg in drie onderwerpen te verdelen die in drie afzonderlijke maar samenhangende hoofdstukken worden behandeld:

- Pseudospectra
- De norm van de matrixexponent en de norm van de macht van een matrix
- Toepassing van pseudospectra op het numeriek oplossen van stelsels gewone differentiaalvergelijkingen

In hoofdstuk 2 wordt het ε-pseudospectrum gedefinieerd en daarna bekijken we een aantal eigenschappen ervan. Vervolgens zullen er voorbeelden worden gegeven van pseudospectra bij verschillende soorten matrices. In de laatste paragraaf van dat hoofdstuk introduceren we het conditiegetal.
Hoofdstuk 3 bespreekt als eerste een aantal eigenschappen van de matrixexponent $e^{t\mathbf{A}}$ die later zullen worden gebruikt bij bewijzen voor de stellingen in dat hoofdstuk. In het resterende hoofdstuk word de relatie tussen de pseudospectra en de matrixexponent $||e^{t\mathbf{A}}||$ besproken. Ook de norm van de macht van een matrix $||\mathbf{A}^k||$ passeert de revue. Dit laatste doen we met het oog op hoofdstuk 4.
In hoofdstuk 4 bekijken we stelsels gewone differentiaalvergelijkingen waarvan we de oplossing numeriek willen oplossen met Runge-Kuttamethoden. Voor een dergelijk numeriek probleem willen we graag stabiliteit, want anders is de numerieke oplossing na enkele tijdstappen niet meer betrouwbaar. Normaal kijken we naar de eigenwaarden van de matrix die bij het probleem opduikt. Als de eigenwaarden onbetrouwbaar worden zullen we de theorie van pseudospectra gebruiken om te laten zien dat een schijnbaar stabiel probleem soms ook instabiel kan zijn.

2 Pseudospectra

Dit hoofdstuk gaat over bepaalde instrumenten van de lineaire algebra die bekend staan als pseudospectra. Deze pseudospectra zijn uitgevonden om informatie te geven over matrices die geen nette basis van eigenvectoren hebben. In dit verslag zal de norm $||\cdot||$ altijd de vector 2-norm of de matrix 2-norm voorstellen. De matrix 2-norm van een matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ definiëren we als

$$||\mathbf{A}|| = \sup_{\mathbf{x}\neq 0} \frac{||\mathbf{Ax}||}{||\mathbf{x}||} = \sup_{||\mathbf{x}||=1} ||\mathbf{Ax}||.$$

2.1 Definities en eigenschappen van ε-pseudospectra

Je kunt het idee van pseudospectra op de volgende manier motiveren. In de toegepaste wiskunde is de vraag 'Is $\mathbf{A}$ singulier?' niet sterk genoeg, want bij een kleine verandering verandert het antwoord van ja naar nee. Voor toegepaste doelen is 'Is $||\mathbf{A}^{-1}||$ groot?' een betere vraag. Als men vraagt 'Is z een eigenwaarde van $\mathbf{A}$?' is dat hetzelfde als dat men vraagt

'Is $zI - \mathbf{A}$ singulier?'

Daarom is de eigenschap om een eigenwaarde te zijn van een matrix ook niet sterk. Een betere vraag zou kunnen zijn

'Is $||(zI - \mathbf{A})^{-1}||$ groot?'

Deze manier van denken leidt ons naar het begrip van pseudospectra. Hieronder staan vier equivalente definities van het pseudospectrum.

Definitie 1 (ε-pseudospectrum) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$ en $\varepsilon > 0$ willekeurig. Het ε-pseudospectrum $\sigma_\varepsilon(\mathbf{A})$ van $\mathbf{A}$ is de verzameling $z \in \mathbb{C}$ zodanig dat*

1.1.1 $||(zI - \mathbf{A})^{-1}|| > \varepsilon^{-1}$, of

1.1.2 $z \in \sigma(\mathbf{A} + \mathbf{E})$ voor een matrix $\mathbf{E} \in \mathbb{C}^{n\times n}$ met $||\mathbf{E}|| < \varepsilon$, of

1.1.3 $||(zI - \mathbf{A})\mathbf{v}|| < \varepsilon$ voor elke vector $\mathbf{v} \in \mathbb{C}^n$ met $||\mathbf{v}|| = 1$, of

1.1.4 $s_{min}(zI - \mathbf{A}) < \varepsilon$

In stelling 2 wordt de equivalentie van definities 1.1.1, 1.1.2, 1.1.3 en 1.1.4 bewezen. Hier bedoelen we met s_{min} de kleinste singuliere waarde van een matrix. De matrix $(zI - \mathbf{A})^{-1}$ staat bekend als de *resolvente* van $\mathbf{A}$ op z.
In dit verslag bedoelen we met $\sigma(\mathbf{A})$ het *spectrum* (verzameling van eigenwaarden) van $\mathbf{A}$. Dan definiëren we hier ook de *resolvente verzameling* $\varrho(\mathbf{A})$ van de matrix $\mathbf{A}$, waarbij

$$\varrho(\mathbf{A}) := \{z \in \mathbb{C} : z \notin \sigma(\mathbf{A})\} \tag{2.1}$$

In dit verslag hanteren we de conventie

$$||(zI - \mathbf{A})^{-1}|| = \infty \text{ voor } z \in \sigma(\mathbf{A}). \tag{2.2}$$

In het bijzonder is het spectrum bevat in het ε-pseudospectrum voor elke $\varepsilon > 0$. U zou misschien kunnen denken dat $||(zI - \mathbf{A})^{-1}||$ alleen groot is precies als z dichtbij een eigenwaarde ligt van $\mathbf{A}$. Voor een normale matrix is deze intuïtie correct. Het belang van pseudospectra ontstaat voor matrices die verre van normaal zijn, waarvoor $||(zI - \mathbf{A})^{-1}||$ heel groot kan zijn zelfs als z ver van het spectrum ligt.

Merk op dat de pseudospectra met verschillende ε bevat in elkaar zijn. Immers, laat $0 < \varepsilon_1 < \varepsilon_2$ en $z \in \sigma_{\varepsilon_1}(\mathbf{A})$, dan

$$||(zI - \mathbf{A})^{-1}|| > \varepsilon_1^{-1} > \varepsilon_2^{-1}.$$

Dus dan $z \in \sigma_{\varepsilon_2}(\mathbf{A})$. Daarentegen zit niet elke $z \in \sigma_{\varepsilon_2}(\mathbf{A})$ in $\sigma_{\varepsilon_1}(\mathbf{A})$. Hiermee is het volgende aangetoond.

$$\sigma_{\varepsilon_1}(\mathbf{A}) \subseteq \sigma_{\varepsilon_2}(\mathbf{A}), \quad 0 < \varepsilon_1 < \varepsilon_2. \tag{2.3}$$

Merk ook op dat de doorsnede van alle pseudospectra gelijk is aan het spectrum,

$$\bigcap_{\varepsilon > 0} \sigma_\varepsilon(\mathbf{A}) = \sigma(\mathbf{A}). \tag{2.4}$$

Het getal z in onze definities is een *ε-pseudoeigenwaarde* van $\mathbf{A}$, en in de derde definitie is $\mathbf{v}$ de corresponderende *ε-pseudoeigenvector*. Met andere woorden, het *ε-pseudospectrum* is de verzameling van *ε-pseudoeigenwaarden*.

De vier definities van het ε-pseudospectrum lijken op het eerste gezicht verschillend, maar ze zijn in feite equivalent. In verschillende gevallen is de ene definitie handiger om te gebruiken dan een andere definitie, dus daarom tonen we de equivalentie van deze definities hieronder aan.

Stelling 1 *De vier gegeven definities van het ε-pseudospectrum zijn equivalent.*

Bewijs:

1.2 $\Rightarrow$ 1.3
Voor $z \in \sigma(\mathbf{A})$ is de gelijkheid triviaal, dus neem aan dat $z \notin \sigma(\mathbf{A})$, zodat $(zI - \mathbf{A})^{-1}$ bestaat (eindig is). Zij $||\mathbf{E}|| < \varepsilon$ en een $\mathbf{v} \in \mathbb{C}^n$, die we genormaliseerd kunnen nemen, $||\mathbf{v}|| = 1$. Neem aan dat $(\mathbf{A} + \mathbf{E})\mathbf{v} = z\mathbf{v}$. Dan $||(zI - \mathbf{A})\mathbf{v}|| = ||\mathbf{E}\mathbf{v}|| < \varepsilon$.

1.3 $\Rightarrow$ 1.1
Neem aan $(zI - \mathbf{A})\mathbf{v} = s\mathbf{u}$ voor $\mathbf{v}, \mathbf{u} \in \mathbb{C}^n$ met $||\mathbf{v}|| = ||\mathbf{u}|| = 1$ en $s < \varepsilon$. Dan $(zI - \mathbf{A})^{-1}\mathbf{u} = s^{-1}\mathbf{v}$, en dus $||(zI - \mathbf{A})^{-1}|| \geq ||(zI - \mathbf{A})^{-1}\mathbf{u}|| = s^{-1} > \varepsilon^{-1}$.

1.1 $\Rightarrow$ 1.2
Neem aan $||(zI-\mathbf{A})^{-1}|| > \varepsilon^{-1}$. Dan $(zI-\mathbf{A})^{-1}\mathbf{u} = s^{-1}\mathbf{v}$ voor $\mathbf{v},\mathbf{u} \in \mathbb{C}^n$ met $||\mathbf{v}|| = ||\mathbf{u}|| = 1$ en $s < \varepsilon$. Dit impliceert $z\mathbf{v} - \mathbf{A}\mathbf{v} = s\mathbf{u}$. We laten nu zien dat er een matrix $\mathbf{E} \in \mathbb{C}^{n\times n}$ bestaat met $||\mathbf{E}|| = s$ en $\mathbf{E}\mathbf{v} = s\mathbf{u}$, want dan $z\mathbf{v} - \mathbf{A}\mathbf{v} = \mathbf{E}\mathbf{v}$ en dus $(\mathbf{A}+\mathbf{E})\mathbf{v} = z\mathbf{v} \Rightarrow z \in \sigma(\mathbf{A}+\mathbf{E})$. Zo'n $\mathbf{E}$ kan worden opgesteld in de vorm $\mathbf{E} = s\mathbf{u}\mathbf{w}^*$ voor een $\mathbf{w} \in \mathbb{C}^n$ met $\mathbf{w}^*\mathbf{v} = 1$. We kunnen nu eenvoudig nemen $\mathbf{w} = \mathbf{v}$, zodat $\mathbf{E}\mathbf{v} = s\mathbf{u}\mathbf{v}^*\mathbf{v} = s\mathbf{u}||\mathbf{v}|| = s\mathbf{u}$.

1.1 $\Leftrightarrow$ 1.4
De norm van een matrix is gelijk aan zijn grootste singuliere waarde en de norm van de inverse is de inverse van de kleinste singuliere waarde. Dan,

$$||(zI-\mathbf{A})^{-1}|| = (s_{min}(zI-\mathbf{A}))^{-1} > \varepsilon^{-1} \;\Leftrightarrow\; s_{min}(zI-\mathbf{A}) < \varepsilon$$

Hiermee is de equivalentie van de vier definities van het ε-pseudospectrum bewezen. $\square$

In dit verslag zullen we veel aandacht besteden aan non-normale matrices. Er zijn immers veel meer non-normale matrices dan dat er normale matrices zijn. Voor de duidelijkheid geven we de volgende definitie.

Definitie 2 (normale matrix) *Een matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is normaal als het een complete verzameling orthogonale eigenvectoren heeft, ofwel, dat de matrix unitair diagonaliseerbaar is:*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^*.$$

Hierbij is $\mathbf{U}$ unitair en Λ is een diagonale matrix van eigenwaarden.

Een andere maar equivalente eigenschap van normale matrices is dat ze commuteren met hun hermites geadjungeerde: $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$ met $\mathbf{A}$ normaal.

Normale matrices gedragen zich heel netjes tegenover de norm. Merk als eerste op dat als $\mathbf{U}$ een unitaire matrix is (zodat $\mathbf{U}^* = \mathbf{U}^{-1}$), dan $||\mathbf{A}\mathbf{U}|| = ||\mathbf{A}||$ voor een willekeurige matrix $\mathbf{A}$. Dan

$$(zI - \mathbf{U}\mathbf{A}\mathbf{U}^*)^{-1} = [\mathbf{U}(zI-\mathbf{A})\mathbf{U}^*]^{-1} = \mathbf{U}(zI-\mathbf{A})^{-1}\mathbf{U}^*,$$

en dus

$$||(zI - \mathbf{U}\mathbf{A}\mathbf{U}^*)^{-1}|| = ||(zI-\mathbf{A})^{-1}|| \;\; \forall\, z \in \mathbb{C}.$$

Dus de resolvente norm is invariant onder unitaire basistransformaties, wat betekent dat ook het pseudospectrum invariant is onder unitaire basistransformaties. Voor een normale matrix $\mathbf{A}$ hebben we dus de mooie eigenschap:

$$\sigma_\varepsilon(\mathbf{A}) = \sigma_\varepsilon(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^*) = \sigma_\varepsilon(\mathbf{\Lambda}) \;\; \forall\, \varepsilon \geq 0.$$

Voor een normale matrix bestaat zijn ε-pseudospectrum uit de vereniging van de open cirkels met straal ε om de punten in het spectrum. Voordat we dit laten zien gaan we eerst het een en ander bespreken. Voor de volgende stelling definiëren we de open ε-cirkels.

$$\Delta_\varepsilon =: \{z \in \mathbb{C} :\; |z| < \varepsilon\}. \tag{2.5}$$

In deze stelling heeft de optelling van verzamelingen de gebruikelijke betekenis:

$$\sigma(\mathbf{A}) + \Delta_\varepsilon = \{z : \ z = z_1 + z_2, \ z_1 \in \sigma(\mathbf{A}), \ z_2 \in \Delta_\varepsilon\},$$

wat gelijk is aan $\{z : \ \text{dist}(z, \sigma(\mathbf{A})) < \varepsilon\}$, waarbij $\text{dist}(z, \sigma(\mathbf{A}))$ de afstand van een punt z tot het spectrum.

Stelling 2 *Voor elke* $\mathbf{A} \in \mathbb{C}^{n\times n}$,

$$\sigma_\varepsilon(\mathbf{A}) \supseteq \sigma(\mathbf{A}) + \Delta_\varepsilon \quad \forall\ \varepsilon > 0, \tag{2.6}$$

en als $\mathbf{A}$ *normaal is, dan*

$$\sigma_\varepsilon(\mathbf{A}) = \sigma(\mathbf{A}) + \Delta_\varepsilon \quad \forall\ \varepsilon > 0. \tag{2.7}$$

Bewijs Laat $\delta \in \mathbb{C}$ met $|\delta| < \varepsilon$ en $z \in \sigma(\mathbf{A})$, dan $z + \delta \in \sigma(\mathbf{A}) + \Delta_\varepsilon$. Omdat $||\delta I|| = |\delta| < \varepsilon$ weten we van de tweede definitie van het ε-pseudospectrum dat $z + \delta \in \sigma(\mathbf{A} + \delta I)$. Elk element van $\sigma(\mathbf{A}) + \Delta_\varepsilon$ zit dus ook in $\sigma(\mathbf{A} + \delta I)$ en dus $\sigma_\varepsilon(\mathbf{A}) \supseteq \sigma(\mathbf{A}) + \Delta_\varepsilon$.

Laat $\mathbf{A}$ normaal, dan kan worden aangenomen dat deze matrix diagonaal is zonder dat het effect heeft op de norm van $\mathbf{A}$, met diagonale elementen a_{jj} die gelijk zijn aan de eigenwaarden λ_j. In dit geval is de resolvente $(zI - \mathbf{A})^{-1}$ ook een diagonaalmatrix. Dit betekent:

$$||(zI - \mathbf{A})^{-1}|| = (s_{min}(zI - \mathbf{A}))^{-1} = (\min_j \ z - \lambda_j)^{-1} = \text{dist}(z, \sigma(\mathbf{A}))^{-1} = \varepsilon^{-1}.$$

en daarom $\sigma_\varepsilon(\mathbf{A}) = \sigma(\mathbf{A}) + \Delta_\varepsilon$. □

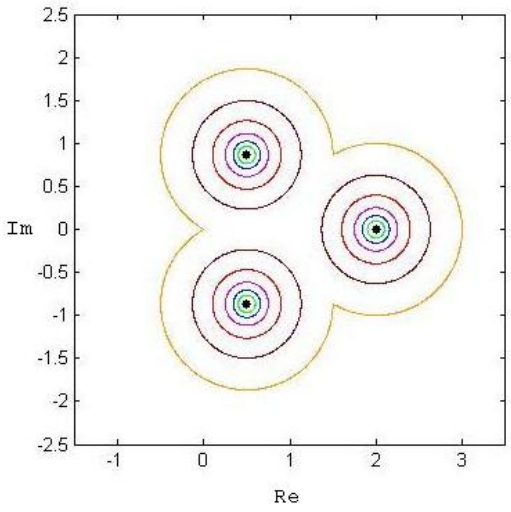
2.2 Voorbeelden van spectra

In figuur 1 op de volgende pagina zijn de randen van enkele ε-pseudospectra getekend van de matrix

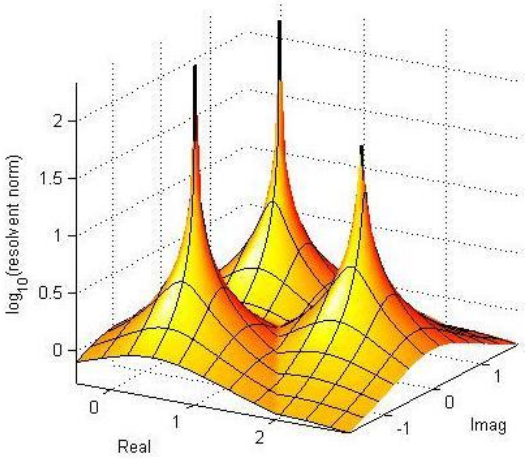
$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Deze matrix heeft eigenwaarden $\lambda_1 = 2, \ \lambda_2 = \frac{1}{2} + \frac{1}{2}i\sqrt{3}$ en $\lambda_3 = \frac{1}{2} - \frac{1}{2}i\sqrt{3}$.

In figuur 1 is goed te zien dat de pseudospectra bevat zijn in de ε-cirkels om de elementen van $\sigma(\mathbf{A})$. De cirkels zijn de grenzen voor de ε-pseudospectra met $\varepsilon = 10^{-1}, \ 10^{-0,8}, \ 10^{-0,6}$ $, 10^{-0,4}, \ 10^{-0,2}, \ 1$.
In figuur 2 is $\log_{10} ||(zI - \mathbf{A})^{-1}||$ getekend als een functie van $z \in \mathbb{C}$. We hebben hier voor het logaritme gekozen, omdat het een beter zicht geeft van deze grafiek. De lezer moet zich realiseren dat de grafiek een stuk hoger is dichtbij de eigenwaarden dan hier is weergegeven. Deze zijn te vinden bij de toppen van de grafiek.



Figuur 1: ε-pseudospectra van de normale matrix $\mathbf{A}$

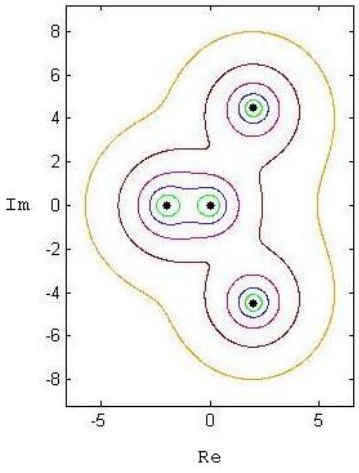


Figuur 2: De resolvente norm als een functie van $z \in \mathbb{C}$ voor normale matrix $\mathbf{A}$.

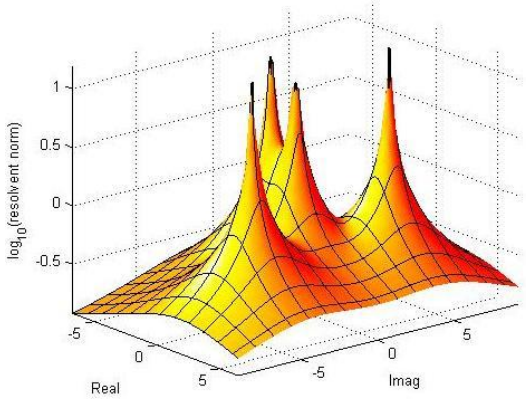
Voor een andere matrix $\mathbf{B}$ die niet normaal is hebben we ook enkele ε-pseudospectra getekend in figuur 3 en in figuur 4 de grafiek van $\log_{10} ||(zI - \mathbf{B})^{-1}||$ als een functie van $z \in \mathbb{C}$. Hierbij is

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 3 & 1 \\ 2 & -1 & 0 & -1 \\ -3 & 2 & 1 & -2 \\ 4 & 1 & 6 & 1 \end{bmatrix}$$

Deze matrix heeft eigenwaarden $\lambda_1 = 0,\ \lambda_2 = -2,\ \lambda_3 = 2 + 2i\sqrt{5}$ en $\lambda_4 = 2 - 2i\sqrt{5}$



Figuur 3: ε-pseudospectra van de niet normale matrix $\mathbf{B}$



Figuur 4: De resolvente norm als een functie van $z \in \mathbb{C}$ voor niet normale matrix $\mathbf{B}$.

De ε-pseudospectra zijn getekend voor $\varepsilon = 10^{-0,5},\ 10^{-0,25},\ 1,\ 10^{0,25},\ 10^{0,5}$. De spectra zijn hier niet cirkelvormig, maar meer in de vorm van een ellips. In deze figuur is ook te zien dat naarmate ε toeneemt, de randen van de spectra meer cirkelvormig lijken.
Merk op dat 0 een eigenwaarde is van $\mathbf{B}$ en dat $\mathbf{B}$ dus singulier is. Voor het tekenen van de grafiek van de resolvente norm is dit geen probleem omdat deze eenzelfde soort singulariteit heeft bij $z = 0$ als bij andere eigenwaarden. Voor meer voorbeelden van pseudospectra verwijzen we de lezer naar [3].

2.3 Het conditiegetal

In deze paragraaf wordt het *conditiegetal* van de basis van eigenvectoren gedefinieerd. Voordat we dit doen kijken we naar de norm van een inverse matrix. Het is bekend dat als $s_1, s_2, \ldots, s_n$ de singuliere waarden zijn van een matrix $\mathbf{A}$ met $s_1 \leq s_2 \leq \ldots \leq s_n$, dan $||A|| = s_{max}(\mathbf{A}) = s_n$. Ook weten we dan dat $\frac{1}{s_1}, \frac{1}{s_2}, \ldots, \frac{1}{s_n}$ de singuliere waarden zijn van $\mathbf{A}^{-1}$. Nu geldt dus dat $\frac{1}{s_1} \geq \frac{1}{s_2} \geq \ldots \geq \frac{1}{s_n}$ en dus dat

$$||\mathbf{A}^{-1}|| = s_{max}(\mathbf{A}^{-1}) = \frac{1}{s_1} = \frac{1}{s_{min}(\mathbf{A})}. \tag{2.8}$$

Dit gegeven gebruiken we bij de definitie van het conditiegetal van de basis van eigenvectoren.

Definitie 3 *Laat $\mathbf{V} \in \mathbb{C}^{n\times n}$ de matrix zijn met als kolommen de eigenvectoren van een diagonaliseerbare matrix $\mathbf{A}$. Het conditiegetal $\kappa(\mathbf{V})$ van de basis van eigenvectoren van $\mathbf{A}$ is*

$$\kappa(\mathbf{V}) \equiv ||\mathbf{V}||\ ||\mathbf{V}^{-1}|| = \frac{s_{max}(\mathbf{V})}{s_{min}(\mathbf{V})}, \tag{2.9}$$

Hier zijn $s_{max}(\mathbf{V})$ en $s_{min}(\mathbf{V})$ respectievelijk de grootste en kleinste singuliere waarde van $\mathbf{V}$. Aan de definitie is af te lezen dat $1 \leq \kappa(\mathbf{V}) < \infty$. Als $\mathbf{A}$ normaal is, dan is $\mathbf{V}$ unitair, dus $\kappa(\mathbf{V}) = ||\mathbf{V}||\ ||\mathbf{V}^{-1}|| = 1$. We kunnen nu dus een soort maat geven aan de mate waarop een matrix normaliteit benadert. Conditiegetallen dichtbij 1 geven namelijk aan dat de basis van eigenvectoren bijna orthogonaal en dat de bijbehorende matrix zich dus redelijk netjes gedragen ten opzichte van basistransformaties.
De volgende stelling, die ook wel bekend staat als het Bauer-Fike theorema geeft een onder- en bovengrens voor het ε-pseudospectrum.

Stelling 3 (Bauer-Fike theorema) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$ diagonaliseerbaar, $\mathbf{A} = \mathbf{V\Lambda V^{-1}}$. Dan geldt voor elke $\varepsilon > 0$,*

$$\sigma(\mathbf{A}) + \Delta_\varepsilon \subseteq \sigma_\varepsilon(\mathbf{A}) \subseteq \sigma(\mathbf{A}) + \Delta_{\varepsilon\kappa(\mathbf{V})}. \tag{2.10}$$

Bewijs De eerste inclusie is al bewezen, dus gaan we nu de tweede inclusie bewijzen. We berekenen

$$(zI - \mathbf{A})^{-1} = (zI - \mathbf{V\Lambda V}^{-1})^{-1} = [\mathbf{V}(zI - \mathbf{\Lambda})\mathbf{V}^{-1}]^{-1} = \mathbf{V}(zI - \mathbf{\Lambda})^{-1}\mathbf{V}^{-1}.$$

Dus

$$||(zI - \mathbf{A})^{-1}|| \leq ||\mathbf{V}||\ ||(zI - \mathbf{\Lambda})^{-1}||\ ||\mathbf{V}^{-1}|| = \kappa(\mathbf{V})||(zI - \mathbf{\Lambda})^{-1}|| = \frac{\kappa(\mathbf{V})}{\text{dist}(z, \sigma(\mathbf{A}))}.$$

Als $z \in \sigma_\varepsilon(\mathbf{A})$, dan weten we dat $||(zI - \mathbf{A})^{-1}|| > \varepsilon^{-1}$. Dus

$$\varepsilon^{-1} < \frac{\kappa(\mathbf{V})}{\text{dist}(z, \sigma(\mathbf{A}))} \quad \Rightarrow \quad \text{dist}(z, \sigma(\mathbf{A})) < \varepsilon\kappa(\mathbf{V}).$$

Dit wil zeggen dat z ligt op een van de open $\varepsilon\kappa(\mathbf{V})$-cirkels en dus $\sigma_\varepsilon(\mathbf{A}) \subseteq \sigma(\mathbf{A}) + \Delta_{\varepsilon\kappa(\mathbf{V})}$. $\square$

Het Bauer-Fike theorema zegt ons dat $\sigma_\varepsilon(\mathbf{A})$ relatief groot kan zijn als het conditiegetal van de basis van eigenvectoren van $\mathbf{A}$ groot is. Merk op dat we met deze stelling nogmaals bevestigt krijgen dat als $\mathbf{A}$ normaal is, dan

$$\sigma_\varepsilon(\mathbf{A}) = \sigma(\mathbf{A}) + \Delta_\varepsilon \quad \forall\ \varepsilon > 0.$$

Met deze stelling zien we ook dat we geen voorspelling kunnen doen over de grootte van het ε-pseudospectrum van een singuliere matrix $\mathbf{A}$. Dit komt doordat dan $\kappa(\mathbf{V}) = \infty$ en dus $\sigma_\varepsilon(\mathbf{A})$ bevat is in een ε-cirkel met oneindige straal.
In dit hoofdstuk hebben we het ε-pseudospectrum van een matrix op verschillende, maar equivalente manieren gedefinieerd. Ook hebben we een aantal eigenschappen ontdekt over deze spectra. Voordat we in hoofdstuk 4 het *Kreiss matrix theorema* gaan behandelen, hebben we meer theoretische ondersteuning nodig. Deze zal behandelt worden in hoofdstuk 3.

3 De matrixexponent

Bij veel wiskundige toepassingen is het nodig om systemen van differentiaalvergelijkingen op te lossen. Laat $\mathbf{x}_0 \in \mathbb{C}^n$ en $\mathbf{A} \in \mathbb{C}^{n\times n}$. Dan kunnen we het volgende beginwaardeprobleem oplossen:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{3.1}$$

De oplossing van dit beginwaardeprobleem is $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$. Voor een matrix $\mathbf{A}$ en $t \in \mathbb{R}$, definiëren we de matrixexponent $e^{t\mathbf{A}}$ als volgt:

$$e^{t\mathbf{A}} = I + t\mathbf{A} + \frac{t^2\mathbf{A}^2}{2!} + \frac{t^3\mathbf{A}^3}{3!} + \ldots = \sum_{n=1}^{\infty} \frac{t^n\mathbf{A}^n}{n!} \tag{3.2}$$

Deze reeks is convergent voor alle $\mathbf{A}$ en t, zodat $e^{t\mathbf{A}}$ een goed gedefinieerde $n \times n$ matrix is. In de eerste twee paragrafen van dit hoofdstuk word de relatie tussen pseudospectra en de e-macht behandelt. Hierbij hebben we het voornamelijk over de matrixexponent $||e^{t\mathbf{A}}||$. Dit is maar één van de vele toepassingen van pseudospectra. In de derde paragraaf zullen we het min of meer equivalente idee van $||\mathbf{A}^k||$ behandelen. Deze theorie is een belangrijke ondersteuning voor de theorie in hoofdstuk 4.

3.1 Eigenschappen van de matrixexponent

De matrixexponent heeft dezelfde eigenschappen als de e-macht e^{ta} met $a \in \mathbb{C}$. In deze paragraaf zullen deze eigenschappen gegeven en bewezen worden.

Stelling 4 (Eigenschappen $e^{t\mathbf{A}}$) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$. Dan geldt:*

1. $\frac{d}{dt}(e^{t\mathbf{A}}) = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A} \quad \forall\, t \in \mathbb{C}$.
2. $e^{t\mathbf{A}}e^{s\mathbf{A}} = e^{(t+s)\mathbf{A}} \quad \forall\, t, s \in \mathbb{C}$.
3. $(e^{t\mathbf{A}})^{-1} = e^{-t\mathbf{A}} \quad \forall\, t \in \mathbb{C}$.
4. *Laat $V \in \mathbb{C}^{n\times n}$ een inverteerbare matrix, dan $e^{t\mathbf{A}} = Ve^{t(V^{-1}\mathbf{A}V)}V^{-1}$.*

Bewijs

1) We bewijzen nu de eerste gelijkheid.

$$\begin{aligned}\frac{d}{dt}(e^{t\mathbf{A}}) &= \frac{d}{dt}\left[I + t\mathbf{A} + \frac{t^2\mathbf{A}^2}{2!} + \frac{t^3\mathbf{A}^3}{3!} + \ldots\right] = \mathbf{A} + t\mathbf{A}^2 + \frac{t^2\mathbf{A}^3}{2!} + \ldots \\ &= \mathbf{A}\left[I + t\mathbf{A} + \frac{t^2\mathbf{A}^2}{2!} + \frac{t^3\mathbf{A}^3}{3!} + \ldots\right] = \mathbf{A}e^{t\mathbf{A}}.\end{aligned}$$

Het mag nu duidelijk zijn dat $e^{t\mathbf{A}}$ de oplossing is van het systeem $X' = \mathbf{A}X$. Voor de tweede gelijkheid laten we zien dat $Y = e^{t\mathbf{A}}\mathbf{A}$ en $Z = \mathbf{A}e^{t\mathbf{A}}$ beide oplossingen zijn van hetzelfde beginwaardeprobleem $X' = \mathbf{A}X,\ X(0) = \mathbf{A}$. We hebben

$$\begin{aligned}Y' &= \frac{d}{dt}(e^{t\mathbf{A}})\mathbf{A} = (\mathbf{A}e^{t\mathbf{A}})\mathbf{A} = \mathbf{A}(e^{t\mathbf{A}}\mathbf{A}) = \mathbf{A}Y \\ Z' &= \mathbf{A}\frac{d}{dt}(e^{t\mathbf{A}}) = (\mathbf{A}e^{t\mathbf{A}}) = \mathbf{A}Z\end{aligned}$$

Ook hebben we $Y(0) = Z(0) = \mathbf{A}$. Vanwege uniciteit van oplossingen van beginwaardeproblemen weten we dat $Y = Z$. Dit is de gewenste gelijkheid.

2) Het systeem $X' = \mathbf{A}X$ is autonoom, dus voor elke constante s geldt dat $X_1(t) = e^{(t+s)\mathbf{A}}$ een oplossingsmatrix is voor dit systeem. Voor elke constante s is $X_2(t) = e^{t\mathbf{A}}e^{s\mathbf{A}}$ ook een oplossingsmatrix voor $X' = \mathbf{A}X$. Nu $X_1(0) = e^{s\mathbf{A}}$ en $X_2(0) = e^{0\mathbf{A}}e^{s\mathbf{A}} = I_n e^{s\mathbf{A}} = e^{s\mathbf{A}}$, dus door uniciteit van oplossingen voor beginwaardeproblemen geldt nu dat $X_1(t) = X_2(t)$.

3) Gebruik makende van het vorige bewijs, zien we dat $e^{t\mathbf{A}}e^{-t\mathbf{A}} = e^{(t-t)\mathbf{A}} = e^{0\mathbf{A}} = I_n$. Dit laat zien dat $e^{-t\mathbf{A}} = (e^{t\mathbf{A}})^{-1}$.

4) We laten zien dat $e^{t(V^{-1}\mathbf{A}V)} = V^{-1}e^{t\mathbf{A}}V$.

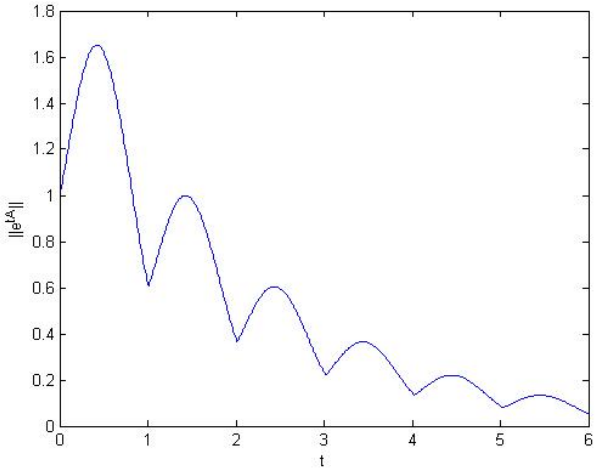
$$\begin{aligned} e^{t(V^{-1}\mathbf{A}V)} &= I + tV^{-1}\mathbf{A}V + \frac{t^2(V^{-1}\mathbf{A}V)^2}{2!} + \frac{t^3(V^{-1}\mathbf{A}V)^3}{3!} + \cdots \\ &= V^{-1}V + tV^{-1}\mathbf{A}V + \frac{t^2V^{-1}\mathbf{A}^2V}{2!} + \frac{t^3V^{-1}\mathbf{A}^3V}{3!} + \cdots \\ &= V^{-1}\left[I + t\mathbf{A} + \frac{t^2\mathbf{A}^2}{2!} + \frac{t^3\mathbf{A}^3}{3!} + \cdots\right]V = V^{-1}e^{t\mathbf{A}}V. \ \square \end{aligned}$$

3.2 De norm van de matrixexponent

Het bestuderen van ε-pseudospectra geeft meer inzicht in de eigenschappen van de matrixexponent $||e^{t\mathbf{A}}||$. We zijn nu dus geïnteresseerd in de relatie tussen $e^{t\mathbf{A}}$ en $(zI - \mathbf{A})^{-1}$. Voor het verdere onderzoek zijn we ook geïnteresseerd in gedragingen van de norm van de matrixexponent $||e^{t\mathbf{A}}||$. Zo worden er boven- en ondergrenzen gegeven en bewezen en zullen er ook voorbeelden worden gegeven waarbij de theorie een stuk concreter zal worden.

Om een idee te geven van hoe de grafiek van $||e^{t\mathbf{A}}||$ eruit kan zien, geven we in figuur 5 de grafiek ervan waarbij

$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ -4 & 2 \end{bmatrix}.$$



Figuur 5: De grafiek van $||e^{t\mathbf{A}}||$ voor $0 \leq t \leq 6$

We zien dat deze functie continu is, maar op sommige punten niet differentieerbaar. We kunnen uitzoeken waar dit door komt met behulp van een wiskundig programma als MAPLE. De matrixexponent is gemakkelijk uit te rekenen met dit programma. Deze zijn al gauw van een dusdanige vorm, dat het onplezierig is om dit met de hand uit te rekenen. We kunnen bijvoorbeeld de matrixexponent van de laatste matrix $\mathbf{A}$ uitrekenen. Dan krijgen we

$$e^{t\mathbf{A}} = e^{-\frac{1}{2}t}\begin{bmatrix} \cos(\frac{1}{2}t\sqrt{39}) - \frac{5}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39}) & \frac{8}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39}) \\ -\frac{8}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39}) & \cos(\frac{1}{2}t\sqrt{39}) + \frac{5}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39}) \end{bmatrix}.$$

De norm van deze matrix wordt dan

$$||e^{t\mathbf{A}}|| = \max\left\{e^{-\frac{1}{2}t}\Big(\frac{8}{\sqrt{39}}|\sin(\frac{1}{2}t\sqrt{39})| \pm |\cos(\frac{1}{2}t\sqrt{39}) - \frac{5}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39})|\Big)\right\}.$$

Het sinusoïdale gedrag van de grafiek is te herkennen aan de trigonometrische functies in deze uitdrukking en de afnemende amplitude is te herkennen aan de factor $e^{-\frac{1}{2}t}$. De niet-differentieerbaarheid voor sommige waarden van t valt nu ook uit te leggen. $||e^{t\mathbf{A}}||$ is een functie waarbij het maximum wordt genomen over twee redelijk onoverzichtelijke trigonometrische functies en dit verschijnsel zal zich voordoen bij allerlei matrices. In dit geval hebben we louter positieve termen vanwege de absoluutstrepen en dus zien we in deze grafiek eigenlijk alleen de grafiek van de functie

$$||e^{t\mathbf{A}}|| = e^{-\frac{1}{2}t}\Big(\frac{8}{\sqrt{39}}|\sin(\frac{1}{2}t\sqrt{39})| + |\cos(\frac{1}{2}t\sqrt{39}) - \frac{5}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39})|\Big).$$

Door de absoluutstrepen zien we in figuur 5 het zogenaamde 'stuitereffect'. De niet-differentieerbaarheid doet zich voor als $\frac{5}{\sqrt{39}}\sin(\frac{1}{2}t\sqrt{39}) = 0$. Dit gebeurt voor

$$t = \frac{n\pi}{\frac{1}{2}\sqrt{39}} \approx 1.00611n, \quad n = 1, 2, \ldots$$

Wat ook noemenswaardig is, is dat de eigenwaarden van deze matrix $\mathbf{A}$ gelijk zijn aan $\lambda = -\frac{1}{2} \pm \frac{1}{2}i\sqrt{39}$. Het reële deel is terug te vinden in de exponent van de e-macht en het imaginaire deel vinden we binnen de trigonometrische functies. Dit verschijnsel zien we over het algemeen bij 2×2-matrices en dus bepaalt het reële deel van de eigenwaarden de mate waarin de functie toe- of afneemt en de periode wordt bepaald door het imaginaire deel van de eigenwaarden. Als de norm van het imaginaire deel groot is, dan is de periode kleiner en zullen we dus meer hobbels zien.

In deze paragraaf gebruiken we de *spectrale abscis* en de *ε-pseudospectrale abscis* die te maken hebben met de ε-pseudoeigenwaarden. Deze zijn hieronder gedefinieerd.

Definitie 4 (abscissen) *De spectrale abscis $\alpha(\mathbf{A})$ is gedefinieerd als*

$$\alpha(\mathbf{A}) = \max_{z \in \sigma(\mathbf{A})} \text{Re } z. \tag{3.3}$$

De ε-pseudospectrale abscis $\alpha_\varepsilon(\mathbf{A})$ is gedefinieerd als

$$\alpha_\varepsilon(\mathbf{A}) = \max_{z \in \sigma_\varepsilon(\mathbf{A})} \text{Re } z. \tag{3.4}$$

De volgende stelling, waarbij de Laplacetransformatie wordt gebruikt, geeft een basis voor verdere bestudering.

Stelling 5 (Relatie tussen $e^{t\mathbf{A}}$ en $(zI-\mathbf{A})^{-1}$) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, dan bestaan er $\omega \in \mathbb{R}$ en $M \geq 1$ zodanig dat*

$$||e^{t\mathbf{A}}|| \leq Me^{\omega t} \quad \forall\ t \geq 0. \tag{3.5}$$

Elke $z \in \mathbb{C}$ met Re $z > \omega$ *is in de resolvente verzameling $\varrho(\mathbf{A})$ van $\mathbf{A}$, met*

$$(zI-\mathbf{A})^{-1} = \int_0^\infty e^{-zt}e^{t\mathbf{A}}dt. \tag{3.6}$$

Ook geldt

$$e^{t\mathbf{A}} = \frac{1}{2\pi i}\int_\Gamma e^{zt}(zI-\mathbf{A})^{-1}dz, \tag{3.7}$$

waarbij Γ een gesloten contour is met $\sigma(\mathbf{A})$ in zijn inwendige.

Bewijs Voor (3.5) merken we op dat $||e^{t\mathbf{A}}||$ continu is. Op elk gesloten interval heeft een continue functie een maximum, dus laat $M \geq 1$ met $||e^{t\mathbf{A}}|| \leq M$ voor $t \in [0,1]$. Laat nu $t \geq 0$, dan $n \leq t \leq n+1$ met $n \in \mathbb{N}$. Dan schatten we de matrixexponent als volgt af.

$$\begin{aligned}
||e^{t\mathbf{A}}|| &= ||e^{n\mathbf{A}}e^{(t-n)\mathbf{A}}|| \\
&\leq ||e^{n\mathbf{A}}||\ ||e^{(t-n)\mathbf{A}}|| \\
&\leq ||(e^{\mathbf{A}})^n|| \cdot M \\
&\leq ||e^{\mathbf{A}}||^n \cdot M \\
&\leq M^n \cdot M \\
&\leq M \cdot M^t \quad (e^\omega = M) \\
&= M(e^\omega)^t = Me^{\omega t}
\end{aligned}$$

(3.6) kan worden aangetoond door

$$\int_0^\infty e^{zIt}e^{t\mathbf{A}}dt = \lim_{R\to\infty}\left[(\mathbf{A}-zI)^{-1}e^{t(\mathbf{A}-zI)}\right]_0^R = (\mathbf{A}-zI)^{-1}\lim_{R\to\infty}\left[e^{R(\mathbf{A}-zI)} - I\right] = (zI-\mathbf{A})^{-1}.$$

De integraal (3.7) volgt direct uit de Cauchy integraalformule die gegeneraliseerd is naar matrices. □

In het geval dat $\mathbf{A}$ diagonaliseerbaar is, kunnen we een concrete bovengrens geven voor $||e^{t\mathbf{A}}||$. Laat $V, \Lambda \in \mathbb{C}^{n\times n}$ waarbij V de matrix is met als kolommen de eigenvectoren van $\mathbf{A}$ en Λ de diagonaalmatrix met de corresponderende eigenwaarden $\lambda_1, \lambda_2, \ldots, \lambda_n$. Dan hebben we

$$e^{t\mathbf{A}} = Ve^{t\Lambda}V^{-1}.$$

Nu kunnen we gemakkelijk een bovengrens bepalen voor $||e^{t\mathbf{A}}||$, namelijk

$$||e^{t\mathbf{A}}|| = ||Ve^{t\Lambda}V^{-1}|| \leq \kappa(V)\ ||e^{t\Lambda}|| = \kappa(V)e^{t\alpha(A)}.$$

We beweren overigens niet dat deze bovengrens optimaal is. De laatste uitdrukking geeft ons een leuke bijkomstigheid: als $\alpha(\mathbf{A}) < 0$, dan $\lim_{t\to\infty}||e^{t\mathbf{A}}|| = 0$. Stelling 6 brengt ons tot een interessante bovengrens voor de norm van de matrixexponent.

Stelling 6 (Bovengrens voor $||e^{t\mathbf{A}}||$) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, $\varepsilon > 0$ en Ω een samenhangende verzameling van pseudoeigenwaarden. We definiëren $\mathcal{L}_\varepsilon$ als de lengte van de rand $\partial\Omega$ de verzameling Ω. Dan*

$$||e^{t\mathbf{A}}|| \leq \frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon} \quad \forall\ t \geq 0. \tag{3.8}$$

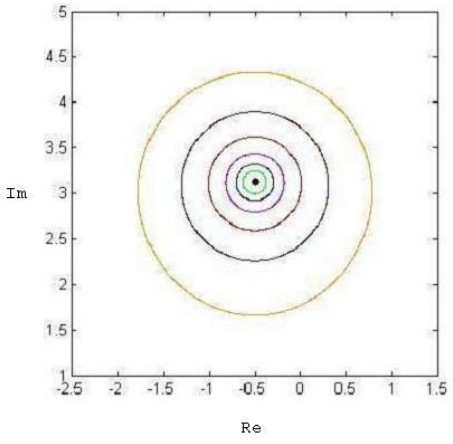
Bewijs Gelijkheid (6) geeft ons

$$e^{t\mathbf{A}} = \frac{1}{2\pi i}\int_{\partial\Omega} e^{zt}(zI-\mathbf{A})^{-1}dz$$

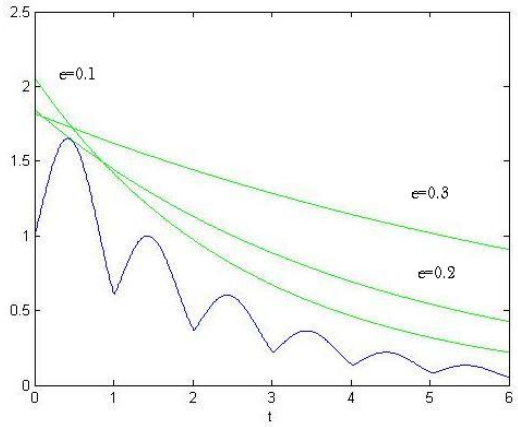
Het resultaat volgt uit het begrenzen van deze integraal.

$$\begin{aligned} ||e^{t\mathbf{A}}|| &= \frac{1}{2\pi}\left|\left|\int_{\partial\Omega} e^{zt}(zI-\mathbf{A})^{-1}dz\right|\right| \\ &\leq \frac{1}{2\pi}\int_{\partial\Omega} ||e^{zt}||\ ||(zI-\mathbf{A})^{-1}||\ |dz| \\ &= \frac{1}{2\pi\varepsilon}\int_{\partial\Omega} e^{zt}\ |dz| \\ &\leq \frac{\mathcal{L}_\varepsilon}{2\pi\varepsilon}\max_{z\in\partial\Omega} e^{zt} = \frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon} \quad \square \end{aligned}$$

Met behulp van numerieke wiskunde kunnen we de lengte van $\partial\Omega$ benaderen. Hierbij kiezen we een waarde voor ε en zoeken waarden z waarvoor $||(zI-\mathbf{A})^{-1}|| \approx \varepsilon^{-1}$. Door middel van interpolatie tussen de gevonden waarden, krijgen we een benadering van de gesloten contour $\partial\Omega$ waarvan we de lengte berekenen.



Figuur 6: $\partial\Omega$ voor verschillende ε rond de eigenwaarde $\lambda = \frac{1}{2} + \frac{1}{2}i\sqrt{39}$.



Figuur 7: De grafieken van $\frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon}$ met $\varepsilon = 0.1,\ 0.2,\ 0.3$.

Elke keer dat de lengte van $\partial\Omega$ werd berekend, bekeken we de rand van de pseudoeigenwaarden rond de eigenwaarde $\lambda = \frac{1}{2} + \frac{1}{2}i\sqrt{39}$. In figuur 6 hierboven staan deze contouren getekend voor $\varepsilon = 10^{-1},\ 10^{-0.8},\ 10^{-0.6},\ 10^{-0.4},\ 10^{-0.2},\ 1$.

In figuur 7 worden de drie grafieken weergegeven van $\frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon}$ bij waarden van $\varepsilon = 0.1,\ 0.2,\ 0.3$ en ook die van $||e^{t\mathbf{A}}||$. De drie grafieken blijven voor elke t netjes boven die van $||e^{t\mathbf{A}}||$, wat rijmt met de theorie van stelling 6. Bij $\varepsilon \approx 0.39$ hebben we dat $\alpha_\varepsilon(\mathbf{A}) = 0$, dus voor $\varepsilon \geq 0.4$ divergeren de functies $\frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon}$ naar ∞ als $t \to \infty$. Het lijkt erop dat deze bovengrens van betere kwaliteit wordt als $\varepsilon \to 0$. Dan hebben we

$$\frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon} \quad \to \quad M e^{t\alpha(\mathbf{A})}.$$

Hierbij is M zo groot dat de grafiek van de functie soms de grafiek van $||e^{t\mathbf{A}}||$ raakt. Een mooi voorbeeld van deze situatie is niet gevonden, omdat de zelfgemaakte software voor het vinden van de lengte van $\partial\Omega$ onbetrouwbaar wordt als ε erg klein gekozen is. Het vermoeden van de laatste bewering is wel sterk als we figuur 7 bekijken.

De volgende stelling geeft de relatie tussen $||e^{t\mathbf{A}}||$ en het spectrum aan.

Stelling 7 ($||e^{t\mathbf{A}}||$ en het spectrum) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, dan*

$$||e^{t\mathbf{A}}|| \geq e^{t\alpha(\mathbf{A})} \quad \forall\ t \geq 0 \tag{3.9}$$

en ook

$$\lim_{t\to\infty} t^{-1} \log ||e^{t\mathbf{A}}|| = \alpha(\mathbf{A}) \tag{3.10}$$

Bewijs Om (3.9) te krijgen maken we een bewijs met behulp van contradictie. Neem aan dat voor een $\tau > 0$, $||e^{\tau\mathbf{A}}|| = \nu < e^{\tau\alpha(\mathbf{A})}$. Voor ongelijkheid (3.5) zijn er twee gevallen:

1. $\omega \leq 0$
2. $\omega > 0$

Voor beide gevallen geldt een analoog bewijs, dus we bewijzen het voor het eerste geval. We nemen aan dat $\omega \leq 0$, zodat we weten van (3.5) dat $||e^{t\mathbf{A}}||$ begrensd is door M voor $0 \leq t < \tau$.
Voor $\tau \leq t \leq 2\tau$ hebben we dan $||e^{t\mathbf{A}}|| = ||e^{(t-\tau)\mathbf{A}} e^{\tau\mathbf{A}}|| \leq ||e^{(t-\tau)\mathbf{A}}||\ ||e^{\tau\mathbf{A}}|| \leq M\nu$. Dus $||e^{t\mathbf{A}}|| \leq M\nu^n < M e^{n\tau\alpha(\mathbf{A})}$ voor $n\tau \leq t \leq (n+1)\tau$. Dus voor alle $t \geq 0$ hebben we nu dat $||e^{t\mathbf{A}}|| \leq \hat{M} e^{n\tau\hat{\omega}} \leq \hat{M} e^{t\hat{\omega}}$ met $\hat{\omega} < \alpha(\mathbf{A})$. $\alpha(\mathbf{A})$ is het reële deel van een eigenwaarde van het spectrum van een matrix $\mathbf{A}$. Volgens stelling 5 betekent dit dat deze eigenwaarde in de resolvente verzameling ligt, wat duidelijk een tegenspraak is van het gestelde. Hiermee is (3.9) bewezen.

Om (3.10) te bewijzen merken we op dat vanwege (3.9),

$$\liminf_{t\to\infty} t^{-1} \log ||e^{t\mathbf{A}}|| \geq \liminf_{t\to\infty} t^{-1} \log e^{t\alpha(\mathbf{A})} = \alpha(\mathbf{A}).$$

En vanwege (3.8) hebben we

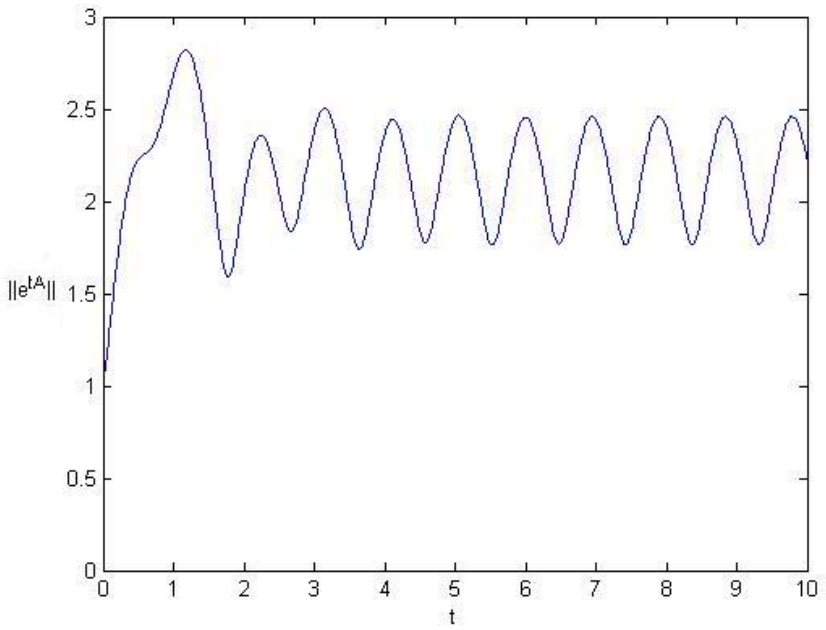
$$\limsup_{t\to\infty} t^{-1} \log ||e^{t\mathbf{A}}|| \leq \limsup_{t\to\infty} t^{-1} \log \left[\frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon}\right] = \limsup_{t\to\infty} \left[\alpha_\varepsilon(\mathbf{A}) + \frac{\log \mathcal{L}_\varepsilon - \log 2\pi\varepsilon}{t}\right].$$

Neem nu $\varepsilon \to 0$, dan $\limsup_{t\to\infty} t^{-1} \log ||e^{t\mathbf{A}}|| \leq \alpha(\mathbf{A})$. Dus dan hebben we uiteindelijk $\lim_{t\to\infty} t^{-1} \log ||e^{t\mathbf{A}}|| = \alpha(\mathbf{A})$. $\square$

Voor de 3×3-matrix $\mathbf{A}$ die is gegeven als

$$\mathbf{A} = \begin{bmatrix} -1 & -5 & -5 \\ 0 & 1 & 3 \\ 0 & -4 & -1 \end{bmatrix}$$

staat in figuur 8 de grafiek van $||e^{t\mathbf{A}}||$ getekend voor $0 \leq t \leq 10$. De eigenwaarden van $\mathbf{A}$ zijn $\lambda_1 = -1, \lambda_2 = i\sqrt{11}$ en $\lambda_3 = -i\sqrt{11}$ en dus $\alpha(\mathbf{A}) = 0$. Stelling 5 en stelling 7 zeggen ons nu dat deze functie naar beneden is begrensd door $e^0 = 1$ en naar boven begrensd is. Dit is duidelijk te zien in de plot. In feite convergeert deze functie ook niet als $t \to \infty$ en benadert uiteindelijk een trigonometrische functie.
Merk op dat hier niet mee wordt bedoeld dat elke functie $||e^{t\mathbf{A}}||$ met $\alpha(\mathbf{A}) = 0$ zich gedraagt als deze functie. Er zijn zelfs veel meer voorbeelden te vinden waarbij deze functie in deze situatie juist convergeert.



Figuur 8: De grafiek van $||e^{t\mathbf{A}}||$ met $\alpha(\mathbf{A}) = 0$.

Eerder in dit verslag hebben we gezien dat de ε-pseudospectra van een normale matrix exact te bepalen zijn. Nu we ook een verband hebben gelegd tussen het spectrum en de matrixexponent, kunnen we aantonen dat we een exact functievoorschrift hebben voor $||e^{t\mathbf{A}}||$ in het geval dat $\mathbf{A}$ normaal is.

Stelling 8 *Laat $A \in \mathbb{C}^{n\times n}$ een normale matrix zijn, dan*

$$||e^{t\mathbf{A}}|| = e^{t\alpha(A)}. \tag{3.11}$$

Bewijs Van stelling 6 weten we dat

$$||e^{t\mathbf{A}}|| \leq \frac{\mathcal{L}_\varepsilon e^{t\alpha_\varepsilon(\mathbf{A})}}{2\pi\varepsilon}.$$

Als $\mathbf{A}$ normaal is, dan bestaan zijn ε-pseudospectra uit de cirkels met straal ε om het spectrum. Hierbij weten we dat $\mathcal{L}_\varepsilon = 2\pi\varepsilon$ en dus $||e^{t\mathbf{A}}|| \leq e^{t\alpha_\varepsilon(\mathbf{A})}$. Bij deze stelling is ε willekeurig, dus we kiezen $\varepsilon \to 0$. Dan

$$||e^{t\mathbf{A}}|| \leq e^{t\alpha(A)}.$$

Samen met stelling 7 hebben we nu dat

$$e^{t\alpha(\mathbf{A})} \leq ||e^{t\mathbf{A}}|| \leq e^{t\alpha(\mathbf{A})} \quad \Rightarrow \quad ||e^{t\mathbf{A}}|| = e^{t\alpha(\mathbf{A})}. \quad \square$$

De verdere theorie behoeft het begrip van de *Kreiss constante*. Deze is als volgt gedefinieerd.

Definitie 5 (Kreiss constante) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, $f : (\mathbb{C}^{n\times n}, t) \to (\mathbb{C}^{n\times n}, t)$ en S het gebied met de volgende eigenschap: Als alle eigenwaarden van $f(\mathbf{A}, t)$ in S liggen, dan $\lim_{t\to\infty} ||f(\mathbf{A}, t)|| = 0$.*

Dan definiëren we de Kreiss constante $\mathcal{K}(\mathbf{A})$ t.o.v. S als

$$\mathcal{K}(\mathbf{A}) := \sup_{z\notin S} \operatorname{dist}(z, S)||(zI - \mathbf{A})^{-1}||.$$

Hierbij is $\operatorname{dist}(z, S)$ *de kleinste afstand tussen z en S.*

We kunnen $\mathcal{K}(\mathbf{A})$ beschouwen als een maat voor de snelheid waarmee de resolvente norm opblaast als $z \to S$, ofwel hoe ver de pseudospectra uit het gebied S steken.
Met de volgende stellingen kunnen we nog meer begrenzingen aangeven voor de matrixexponent.

Stelling 9 (Ondergrenzen voor $||e^{t\mathbf{A}}||$) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$ een matrix zijn. Als $||(zI - \mathbf{A})^{-1}|| = \frac{K}{\operatorname{Re} z}$ voor een $z \in \mathbb{C}$ met $\operatorname{Re} z > 0$ en $K > 1$, dan*

$$\sup_{t\geq 0} ||e^{t\mathbf{A}}|| \geq K. \tag{3.12}$$

De ε-pseudospectrale abscis $\alpha_\varepsilon(\mathbf{A})$ is eindig voor elke $\varepsilon > 0$. Als we z kiezen zodat $\operatorname{Re} z = \alpha_\varepsilon(\mathbf{A})$ en $||(zI - \mathbf{A})^{-1}|| = \varepsilon$, dan krijgen we

$$\sup_{t\geq 0} ||e^{t\mathbf{A}}|| \geq \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon} \quad \forall\ \varepsilon > 0, \tag{3.13}$$

en als we maximaliseren over ε krijgen we

$$\sup_{t\geq 0} ||e^{t\mathbf{A}}|| \geq \mathcal{K}(\mathbf{A}), \tag{3.14}$$

waarbij de Kreiss constante $\mathcal{K}(\mathbf{A})$ t.o.v. het linkerhalfvlak is gedefinieerd als

$$\mathcal{K}(\mathbf{A}) \equiv \sup_{\varepsilon>0} \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon} = \sup_{\operatorname{Re} z>0} (\operatorname{Re} z)||(zI - \mathbf{A})^{-1}||. \tag{3.15}$$

BEWIJS Als $\sup_{t\geq 0} ||e^{t\mathbf{A}}|| = M$ met $M > 0$, dan weten we van (3.6) dat voor elke z met Re $z > 0$,

$$\frac{K}{\text{Re } z} = ||(zI - \mathbf{A})^{-1}|| \leq \int_0^\infty |e^{-zt}| \, ||e^{t\mathbf{A}}|| dt \leq M \int_0^\infty |e^{-zt}| dt = \frac{M}{\text{Re } z}.$$

Hieruit volgt dat $K \leq M$ en dus $\sup_{t\geq 0} ||e^{t\mathbf{A}}|| = M \geq K$. Zo hebben we ongelijkheid (3.12).

We kiezen z zodanig dat Re $z = \alpha_\varepsilon(\mathbf{A})$ en $||(zI - \mathbf{A})^{-1}|| = \varepsilon^{-1}$, dan krijgen we voor alle $\varepsilon > 0$,

$$\sup_{t\geq 0} ||e^{t\mathbf{A}}|| \geq K = ||(zI - \mathbf{A})^{-1}|| \cdot \text{Re } z = \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon}.$$

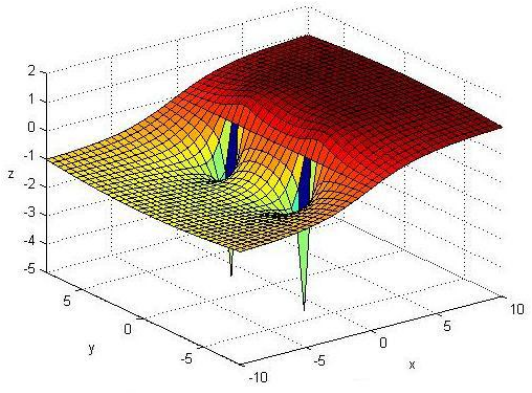
Dit geeft ons ongelijkheid (3.13). Als we daarna het supremum nemen over ε, dan krijgen we ongelijkheid (3.14). □

We bekijken nog eens de volgende matrix.

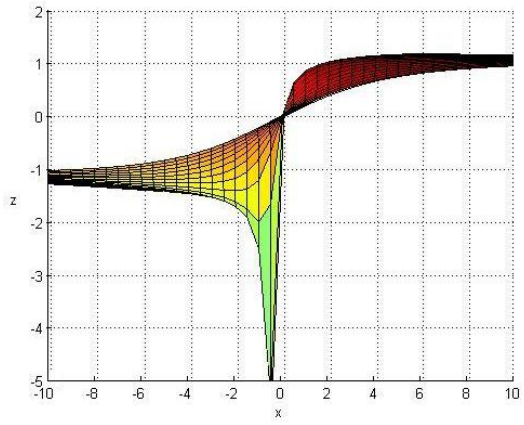
$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ -4 & 2 \end{bmatrix}.$$

Met MATLAB kan je berekenen dat $M = \sup_{t\geq 0} ||e^{t\mathbf{A}}|| \approx 1.6522$, wat ook af te lezen is uit de grafiek in figuur 5 op pagina 10. We gaan de verschillende ondergrenzen voor het supremum testen op deze matrix.

We weten van stelling 9 dat $\sup_{t\geq 0} ||e^{t\mathbf{A}}|| \geq K = \text{Re } z \cdot ||(zI - \mathbf{A})^{-1}||$ voor alle $z \in \mathbb{C}$. In figuren 9 en 10 op de volgende pagina zijn 3D-grafieken van Re $z \cdot ||(zI - \mathbf{A})^{-1}||$, maar beide vanuit een ander punt bekeken. Deze grafieken zijn getekend voor $-10 \leq x \leq 10, -8 \leq y \leq 8$ waarbij $z = x + iy$.



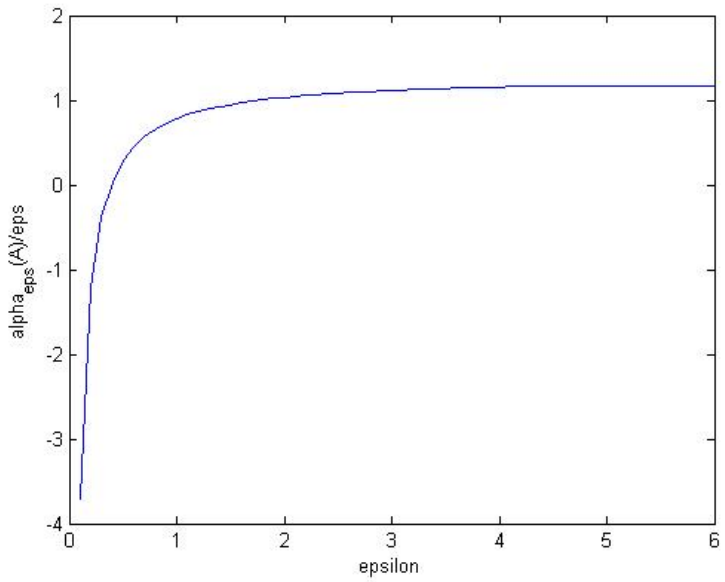
Figuur 9: Re $z \cdot ||(zI - \mathbf{A})^{-1}||$ voor $-10 \leq x \leq 10, -8 \leq y \leq 8$



Figuur 10: De resolvente norm als een functie van $z \in \mathbb{C}$ voor niet normale matrix $\mathbf{B}$.

Zoals we al eerder hebben gezien zijn de reële delen van beide eigenwaarden negatief en $||(zI - \mathbf{A})^{-1}|| > 0$ voor $z \in \mathbb{C}$. Dit verklaart waarom er twee pieken naar beneden zijn rond de eigenwaarden van $\mathbf{A}$.
Figuur 10 geeft duidelijk weer dat $K \leq 0$ voor alle z in het gesloten linkerhalfvlak, waardoor het maximum K_{max} nooit in het linkerhalfvlak kan liggen. Nog een voordeel van deze weergave is dat we ook min of meer kunnen aflezen hoe groot K_{max} is. Met MATLAB kan worden berekend dat $K_{max} \approx 1.1691 < \sup_{t \geq 0} ||e^{t\mathbf{A}}||$. Deze waarde wordt aangenomen op $z \approx 6.0937$.

In figuur 11 hieronder staat de grafiek van de functie $f(\varepsilon) = \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon}$ afgebeeld voor $0 \leq \varepsilon \leq 6$. Dit geeft een concreter beeld van deze functie. De grafiek van f lijkt te convergeren, maar deze functie daalt vanaf $\varepsilon \approx 5.3$. We weten van (3.13) en (3.15) dat $||e^{t\mathbf{A}}||$ groter is dan de maximale functiewaarde f_{max} op deze grafiek. Met MATLAB kan worden berekend dat $f_{max} \approx 1.1691 < \sup_{t \geq 0} ||e^{t\mathbf{A}}||$, dus deze ondergrens klopt.



Figuur 11: De grafiek van $f(\varepsilon) = \frac{\alpha_\varepsilon(\mathbf{A})}{\varepsilon}$ voor $0 \leq \varepsilon \leq 6$

Wat opvalt is dat $K_{max} = f_{max}$ en dus word hier ook aan ongelijkheid (12) voldaan. De waarden K_{max} en f_{max} zijn verkregen door te maximaliseren over respectievelijk z en ε en dus zullen we het vanaf dit punt steeds hebben over de Kreiss-constante $\mathcal{K}(\mathbf{A})$ als we op één van deze waarden doelen.

We kunnen ook kijken naar ondergrenzen van $||e^{t\mathbf{A}}||$ op een interval $0 \leq t \leq \tau$. De volgende stellingen gaan hierover.

Stelling 10 (Ondergrenzen voor $||e^{t\mathbf{A}}||$) *Als $a = \text{Re } z$, dan geldt voor elke $\tau > 0$,*

$$\sup_{0<t\leq\tau} ||e^{t\mathbf{A}}|| \geq \frac{e^{a\tau}}{1 + \frac{e^{a\tau}-1}{K}}, \tag{3.16}$$

en als $||e^{t\mathbf{A}}|| \leq M$ voor alle $t \geq 0$, dan geldt voor elke $\tau \geq 0$, met K gedefinieerd als hiervoor,

maar nu is $a < 0$ gepermitteerd en $-\infty < K/M \leq 1$,

$$||e^{\tau \mathbf{A}}|| \geq e^{a\tau} - \frac{e^{a\tau}-1}{K/M} = 1 - \frac{(e^{a\tau}-1)(1-K/M)}{K/M}. \tag{3.17}$$

In het bijzondere geval dat $a = K = 0$, wordt de laatste ongelijkheid gereduceerd tot

$$||e^{\tau \mathbf{A}}|| \geq 1 - \frac{\tau M}{||(zI-\mathbf{A})^{-1}||}. \tag{3.18}$$

Bewijs Om ongelijkheid (3.16) te bewijzen definiëren we $M_\tau = \sup_{0<t\leq\tau} ||e^{t\mathbf{A}}||$, waardoor $||e^{t\mathbf{A}}|| \leq M_\tau$ voor $0 < t \leq \tau$, $||e^{t\mathbf{A}}|| = ||e^{(t-\tau)\mathbf{A}} e^{\tau\mathbf{A}}|| \leq ||e^{(t-\tau)\mathbf{A}}||\,||e^{\tau\mathbf{A}}|| = M_\tau^2$ voor $\tau < t \leq 2\tau$ enzovoort. Met $a = \text{Re } z$ en (3.6) impliceert dit

$$\begin{aligned}
||(zI-\mathbf{A})^{-1}|| &= \left|\left| \int_0^\infty e^{-zt} e^{t\mathbf{A}} dt \right|\right| \\
&\leq \int_0^\infty |e^{-zt}|\;||e^{t\mathbf{A}}|| dt \\
&= \sum_{n=0}^\infty \int_{n\tau}^{(n+1)\tau} e^{-at} ||e^{t\mathbf{A}}|| dt \quad (\text{Re } z = a) \\
&\leq \sum_{n=0}^\infty \int_{n\tau}^{(n+1)\tau} e^{-at} M_\tau^{n+1} dt \\
&= \sum_{n=0}^\infty \int_0^\tau e^{-a(t+n\tau)} M_\tau^{n+1} dt \\
&= \int_0^\tau e^{-at} dt \sum_{n=0}^\infty e^{-an\tau} M_\tau^{n+1}.
\end{aligned}$$

Als $M_\tau \geq e^{a\tau}$, dan hebben we direct de gewenste ongelijkheid, want $e^{a\tau} \geq \frac{e^{a\tau}}{1+\frac{e^{a\tau}-1}{K}}$. We nemen nu dus aan dat $M_\tau \leq e^{a\tau}$. Stel $b < a$ en $b - a = -c$ met $b, c > 0$, dan

$$\sum_{n=0}^\infty e^{-an\tau} M_\tau^{n+1} = M_\tau \sum_{n=0}^\infty e^{-an\tau} e^{bn\tau} = M_\tau \sum_{n=0}^\infty (e^{-c\tau})^n = \frac{M_\tau}{1-e^{-a\tau}M_\tau}.$$

Dan kunnen we de resolvente norm afschatten door

$$||(zI-\mathbf{A})^{-1}|| \leq \left(\frac{1-e^{-a\tau}}{a}\right)\left(\frac{M_\tau}{1-e^{-a\tau}M_\tau}\right) = \frac{e^{a\tau}-1}{a\left(\frac{e^{a\tau}}{M_\tau}-1\right)}.$$

Als we deze formule inverteren, krijgen we

$$\frac{a}{K} = ||(zI-\mathbf{A})^{-1}||^{-1} \geq \frac{a\left(\frac{e^{a\tau}}{M_\tau}-1\right)}{e^{a\tau}-1},$$

en dus

$$\frac{e^{a\tau}}{M_\tau} - 1 \leq \frac{e^{a\tau}-1}{K} \Longrightarrow M_\tau = \sup_{0<t\leq\tau} ||e^{t\mathbf{A}}|| \geq \frac{e^{a\tau}}{1+\frac{e^{a\tau}-1}{K}}.$$

Hiermee hebben we ongelijkheid (3.16) bewezen.

Uiteindelijk bewijzen we nu ongelijkheden (3.17) en (3.18). We nemen nu aan dat $a \leq 0$ ook gepermitteerd is. Stel $||e^{\tau \mathbf{A}}|| = P$. Door (3.5) hebben we voor $0 \leq t \leq \tau$,

$$||e^{t\mathbf{A}}|| \leq M, \quad ||e^{(\tau+t)\mathbf{A}}|| \leq ||e^{\tau\mathbf{A}}||\,||e^{t\mathbf{A}}|| \leq PM, \quad ||e^{(2\tau+t)\mathbf{A}}|| \leq P^2 M,$$

enzovoort. We hebben nu twee mogelijkheden:

- $P = ||e^{\tau\mathbf{A}}|| \geq e^{a\tau}$
- $P = ||e^{\tau\mathbf{A}}|| < e^{a\tau}$

Als $P \geq e^{\tau\mathbf{A}}$, dan onderscheiden we de gevallen $a > 0$ en $a < 0$. Als $a > 0$, dan is $K/M > 0$ en $e^{a\tau} - 1 \geq 0$. Dus dan $||e^{t\mathbf{A}}|| \geq e^{a\tau} \geq e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$. Als $a < 0$, dan is $K/M < 0$ en $e^{a\tau} - 1 \leq 0$ en dus $||e^{t\mathbf{A}}|| \geq e^{a\tau} \geq e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$.
We nemen dus aan dat $P < e^{a\tau}$. We hebben $K = \text{Re } z \cdot ||(zI - \mathbf{A})^{-1}||$ en Re $z = a$. We bekijken eerst het volgende:

$$\begin{aligned}
||(zI-\mathbf{A})^{-1}|| &\leq \sum_{n=0}^{\infty} \int_{n\tau}^{(n+1)\tau} ||e^{t\mathbf{A}}|| e^{-at} dt \\
&= \sum_{n=0}^{\infty} \int_{0}^{\tau} ||e^{(n\tau+t)\mathbf{A}}|| e^{-a(t-n\tau)} dt \\
&\leq \sum_{n=0}^{\infty} P^n M e^{-an\tau} \int_0^{\tau} e^{-at} dt \\
&= M \sum_{n=0}^{\infty} (Pe^{-a\tau})^n \cdot \frac{1-e^{-a\tau}}{a} \\
&= Ma^{-1} \frac{1-e^{-a\tau}}{1-Pe^{-a\tau}}
\end{aligned}$$

De laatste gelijkheid geldt vanwege de aanname dat $P < e^{a\tau}$, waardoor $Pe^{-a\tau} < e^{a\tau}e^{-a\tau} = 1$. Nu vinden we eenvoudig de gewenste ongelijkheid:

$$\begin{aligned}
K = \text{Re } z \cdot ||(zI-\mathbf{A})^{-1}|| &\Rightarrow K \leq a \cdot Ma^{-1} \frac{1-e^{-a\tau}}{1-Pe^{-a\tau}} \\
&\Rightarrow \frac{K}{M} \leq \frac{1-e^{-a\tau}}{1-Pe^{-a\tau}} \\
&\Rightarrow 1 - Pe^{-a\tau} \leq \frac{1-e^{-a\tau}}{K/M} \\
&\Rightarrow P = ||e^{\tau\mathbf{A}}|| \geq e^{a\tau} - \frac{e^{a\tau}-1}{K/M}
\end{aligned}$$

Met elementair rekenwerk is aan te tonen dat

$$e^{a\tau} - \frac{e^{a\tau}-1}{K/M} = 1 - \frac{(e^{a\tau}-1)(1-K/M)}{K/M}.$$

Als we nu aannemen dat $a = K = 0$, dan zoeken we de volgende limiet.

$$\lim_{a\to 0} e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$$

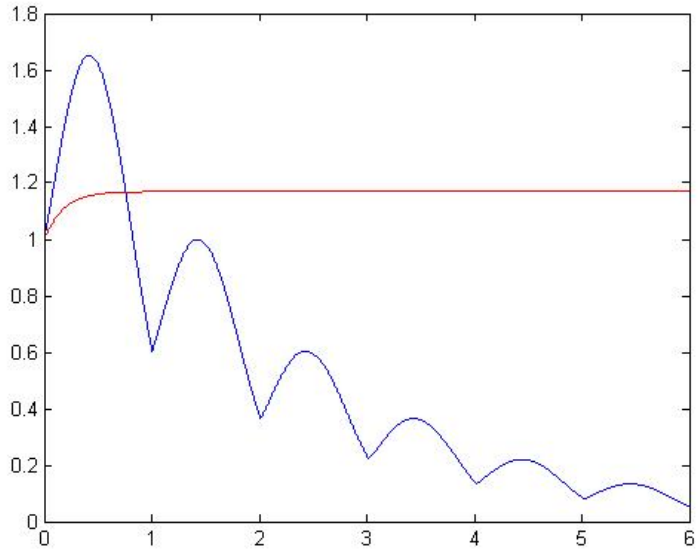
We gebruiken de regel van l'Hôpital, waarbij we differentiëren naar a en redeneren als volgt.

$$\begin{aligned}\lim_{a\to 0} e^{a\tau} - \frac{e^{a\tau}-1}{K/M} &= 1 - M \lim_{a\to 0} \frac{e^{a\tau}-1}{a\ ||(zI-\mathbf{A})^{-1}||} \\ &= 1 - M \lim_{a\to 0} \frac{\tau e^{a\tau}}{||(zI-\mathbf{A})^{-1}||} \\ &= 1 - \frac{\tau M}{||(zI-\mathbf{A})^{-1}||} \quad \square\end{aligned}$$

We gaan nu de grenzen van de vorige stelling testen op

$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ -4 & 2 \end{bmatrix}.$$

In figuur 12 zijn de twee grafieken getekend van $||e^{t\mathbf{A}}||$ in het blauw en van $\frac{e^{a\tau}}{1+\frac{e^{a\tau}-1}{K}}$ voor $0 \leq t, \tau \leq 6$ in het rood. Voor de laatste functie is gekozen voor $a = \text{Re } z = \text{Re}(6.0937) = 6.0937$ waar $K = \mathcal{K}(\mathbf{A}) = 1.1691$ wordt aangenomen. Voor elke τ geeft de rode grafiek een ondergrens voor $\sup_{0<t\leq\tau} ||e^{t\mathbf{A}}||$. De twee grafieken hebben een snijpunt bij $t \approx 0.758$. Voor $t > 0.758$ hebben we dat $||e^{t\mathbf{A}}|| < \frac{e^{a\tau}}{1+\frac{e^{a\tau}-1}{K}}$, maar aangezien het supremum van $||e^{t\mathbf{A}}||$ wordt aangenomen op $t \approx 0.421$, is in deze situatie aan ongelijkheid (3.16) voldaan.



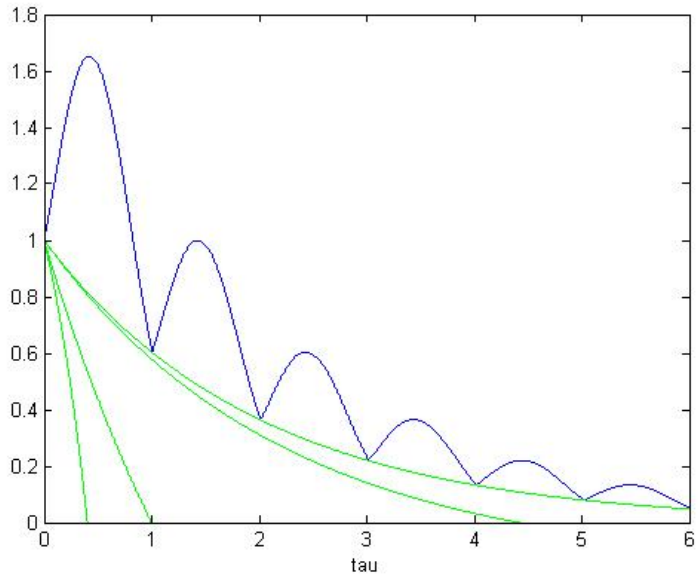
Figuur 12: De grafieken van $||e^{t\mathbf{A}}||$ en $\frac{e^{a\tau}}{1+\frac{e^{a\tau}-1}{K}}$ voor $0 \leq t, \tau \leq 6$

Het valt ook op dat de rode grafiek convergeert naar $K = \mathcal{K}(\mathbf{A})$. Dit is niet zo verwonderlijk, want voor elke $a > 0$,

$$\lim_{\tau\to\infty} \frac{e^{a\tau}}{1 + \frac{e^{a\tau}-1}{K}} = \lim_{\tau\to\infty} \frac{e^{a\tau}}{\frac{e^{a\tau}-1+K}{K}} = \lim_{\tau\to\infty} \frac{Ke^{a\tau}}{e^{a\tau} - 1 + K} = K \quad \forall\ a > 0.$$

Voor de optimale waarde van K in deze functie nemen we dus $\mathcal{K}(\mathbf{A})$. Als we $K > \mathcal{K}(\mathbf{A})$ kiezen, dan is ongelijkheid (3.16) niet meer betrouwbaar. We kunnen uit dit alles concluderen dat deze ondergrens voor het supremum slechte waarden geeft voor $z \in \mathbb{C}$ waarvoor K ver van de Kreiss constante $\mathcal{K}(\mathbf{A})$ ligt.

Dan bekijken we nu figuur 13 hieronder, waarin vijf grafieken zijn getekend. De blauwe lijn correspondeert met $||e^{\tau\mathbf{A}}||$ en de vier groene grafieken corresponderen met $e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$, waarbij van boven naar beneden gekozen is voor $z_1 = -0.5 + 3.122i, z_2 = -0.45 + 3.1i$,
$z_3 = -0.5 + 2.5i, z_4 = 2 + 4i$.



Figuur 13: De grafieken van $||e^{\tau\mathbf{A}}||$ en $e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$ voor $0 \leq \tau \leq 6$

In deze figuur is duidelijk te zien dat als $z \to \sigma(\mathbf{A})$, dat de benadering van $||e^{\tau\mathbf{A}}||$ steeds beter wordt. Zeker bij de keuze voor z_1 zien we dat $e^{a\tau} - \frac{e^{a\tau}-1}{K/M}$ de grafiek bijna snijdt op de punten waar $||e^{\tau\mathbf{A}}||$ niet differentieerbaar is. Dit verschijnsel kan als volgt worden uitgelegd. Als we z dichtbij $\sigma(\mathbf{A})$ kiezen, dan is $||(zI - \mathbf{A})^{-1}||$ erg groot. Daaruit volgt dat $|K| = |\text{Re } z| \cdot ||(zI - \mathbf{A})^{-1}||$ ook erg groot is (zie figuren 9 en 10) evenals K/M. Hieruit kunnen we concluderen dat

$$\lim_{z\to\sigma(\mathbf{A})} e^{a\tau} - \frac{e^{a\tau} - 1}{K/M} \to e^{a\tau}.$$

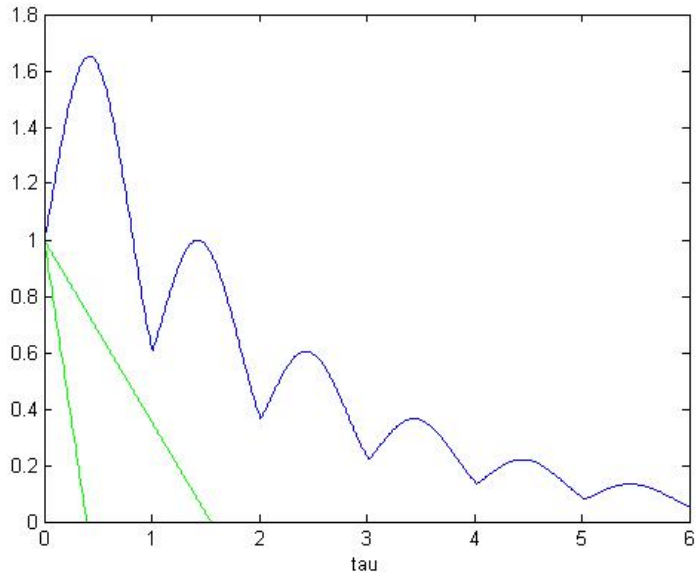
Onze puntsgewijze ondergrens is optimaal als we $a = \alpha(\mathbf{A})$ kiezen, want dan krijgen we $||e^{\tau\mathbf{A}}|| \geq e^{\alpha(\mathbf{A})\tau}$, wat gelijk is aan de ongelijkheid in stelling 7. Uit figuur 13 kunnen we in

ieder geval opmaken dat deze ondergrens over het algemeen belabberde waarden geeft, behalve als z dichtbij $\sigma(\mathbf{A})$ wordt gekozen.

Als laatste bekijken we het geval dat $a = K = 0$. Dat wil zeggen dat Re $z = 0$ en we ons dus beperken tot de imaginaire as. We bekijken dan de grafieken van

$$1 - \frac{\tau M}{||(zI - \mathbf{A})^{-1}||}.$$

Deze lineaire functies zijn volgens stelling 10 puntsgewijs kleiner dat $||e^{\tau \mathbf{A}}||$ voor $\tau > 0$. Hierbij hebben we (zoals eerder) $M = 1.6522$ en we kiezen daarna een z waarna $||(zI - \mathbf{A})^{-1}||$ kan worden berekend. Op $z = 3.0938i$ neemt de resolvente norm zijn maximum 2.5612 aan op de imaginaire as. Met deze waarde van z hebben we de kleinste absolute richtingscoëfficiënt gevonden voor deze lineaire functies en deze is gelijk aan 0.5340.



Figuur 14: De grafieken van $||e^{\tau \mathbf{A}}||$ en $1 - \frac{\tau M}{||(zI - \mathbf{A})^{-1}||}$ voor $0 \leq \tau \leq 6$

In figuur 14 hierboven staat de grafiek getekend bij $z = 3.0938i$ en iets lager de wat minder nauwkeurige ondergrens die we krijgen bij de keuze voor $z = 0$. Deze grafieken blijven puntsgewijs netjes onder de grafiek van $||e^{\tau \mathbf{A}}||$. Deze functies zijn minder goede puntsgewijze ondergrenzen dan de functies $e^{a\tau} - \frac{e^{a\tau} - 1}{K/M}$ van figuur 13. Dit feit volgt direct uit het vergelijken van de beste grafiek van figuur 13 en de beste grafiek uit figuur 14; de functie van figuur 13 is groter dan die van figuur 14 voor alle $t > 0$.

3.3 Machten van matrices en normen

In de §3.2 werd de continue functie $||e^{t\mathbf{A}}||$ behandeld. Voor elke waarde van t heeft deze functie een functiewaarde, maar er zijn ook gevallen waarin het niet mogelijk is om een dergelijke continue functie op te stellen. Dan kunnen we wel een discrete functie opstellen die correspondeert met de continue functie.
We bekijken in deze paragraaf de functie $||\mathbf{A}^k||$, $k \in \mathbb{N}$, waarbij $\mathbf{A} \in \mathbb{C}^{n\times n}$. Er zal niet zo breed worden uitgemeten over de theorie als in §3.2, want we kunnen $\mathbf{A}$ zodanig kiezen dat we het over dezelfde theorie hebben. Kies bijvoorbeeld $\mathbf{A} = e^{t\mathbf{B}}$, dan hebben we $||\mathbf{A}^k|| = ||e^{tk\mathbf{B}}||$ en hebben we dus de discrete versie te pakken van de norm van de matrixexponent. Verder zijn het maar een aantal theorema's die we nodig hebben om de theorie in hoofdstuk 4 uit te werken.

Als eerste bekijken we de relatie tussen $\mathbf{A}^k$ en $(zI - \mathbf{A})^{-1}$. Deze staat in de volgende stelling weergegeven.

Stelling 11 *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, dan bestaat er een $\gamma > 0$ en $M \geq 1$ zodanig dat*

$$||\mathbf{A}^k|| \quad \leq \quad M\gamma^k, \quad k \geq 0. \tag{3.19}$$

Elke $z \in \mathbb{C}$ met $|z| > \gamma$ is in de resolvente verzameling $\varrho(\mathbf{A})$ van $\mathbf{A}$ en de resolvente voor een dergelijke z is gegeven door de convergente reeks

$$(zI - \mathbf{A})^{-1} \quad = \quad z^{-1}(I + z^{-1}\mathbf{A} + (z^{-1}\mathbf{A})^2 + \ldots) = \frac{1}{z}\sum_{n=0}^{\infty} z^{-n}\mathbf{A}^n. \tag{3.20}$$

Daarentegen, voor elke $k \geq 0$,

$$\mathbf{A}^k = \frac{1}{2\pi i}\int_\Gamma z^k(zI - \mathbf{A})^{-1}dz, \tag{3.21}$$

waarbij Γ een willekeurige gesloten contour met $\sigma(\mathbf{A})$ in zijn inwendige.

Bewijs We weten dat $||\mathbf{A}^0|| = ||I|| = 1$, dus laat $M \geq 1$ zodanig dat $||\mathbf{A}|| \leq M$. Neem $k = m + 1, m \in \mathbb{N}$, zodat

$$\begin{aligned} ||\mathbf{A}^k|| &= ||\mathbf{A}^m\mathbf{A}^{k-m}|| \\ &\leq ||\mathbf{A}^m||\ ||\mathbf{A}|| \\ &\leq ||\mathbf{A}||^m \cdot M \\ &\leq M^m \cdot M \\ &\leq M\gamma^k. \end{aligned}$$

De laatste ongelijkheid geldt omdat we kiezen dat $M \leq \gamma$ en $k > m$. De rest volgt direct uit hun definitie. □

We definiëren nu de spectrale radius ρ als volgt.

$$\rho(\mathbf{A}) := \max_{\lambda\in\sigma(\mathbf{A})} |\lambda|. \tag{3.22}$$

De volgende stelling geeft een relatie weer tussen $||\mathbf{A}^k||$ en het spectrum van $\mathbf{A}$.

Stelling 12 *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$ en non-singulier, dan geldt*

$$\rho(\mathbf{A})^k \leq ||\mathbf{A}^k|| \leq ||\mathbf{A}||^k. \tag{3.23}$$

Bewijs Laat λ een eigenwaarde van $\mathbf{A}$ zijn en $\mathbf{v}$ zijn corresponderende eigenvector. We krijgen nu

$$|\lambda|^k \cdot ||\mathbf{v}|| = ||\lambda^k \mathbf{v}|| = ||\mathbf{A}^k \mathbf{v}|| \leq ||\mathbf{A}^k|| \cdot ||\mathbf{v}||.$$

Omdat $\mathbf{v} \neq 0$ voor elke corresponderende eigenwaarde λ, weten we nu dat

$$|\lambda|^k \leq ||\mathbf{A}^k|| \implies \rho(\mathbf{A})^k \leq ||\mathbf{A}^k||.$$

De tweede ongelijkheid volgt direct uit de definiërende eigenschappen van de matrixnorm en is dus triviaal. □

We willen nu ook graag een criterium voor het convergeren van $||\mathbf{A}^k||$ als $k \to \infty$ en als $k \to 0$. Voordat we hierover een uitspraak doen hebben we eerst een ander gegeven nodig. Deze wordt gegeven in stelling 13 en word gebruikt in het bewijs van stelling 14.

Stelling 13 *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$ en $\rho(\mathbf{A})$ zijn spectrale radius. Dan*

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0} \iff \rho(\mathbf{A}) < 1 \tag{3.24}$$

Verder geldt dat als $\rho(\mathbf{A}) > 1$, dan is $||\mathbf{A}^k||$ onbegrensd.

Bewijs ($\Rightarrow$) Laat $\mathbf{v}$ een eigenvector zijn van $\mathbf{A}$ met corresponderende eigenwaarde λ. Omdat $\mathbf{A}^k\mathbf{v} = \lambda^k\mathbf{v}$, hebben we

$$\begin{aligned} \mathbf{0} &= (\lim_{k\to\infty} \mathbf{A}^k)\mathbf{v} = \lim_{k\to\infty} \mathbf{A}^k\mathbf{v} \\ &= \lim_{k\to\infty} \lambda^k\mathbf{v} = \mathbf{v} \lim_{k\to\infty} \lambda^k. \end{aligned}$$

Omdat $\mathbf{v} \neq \mathbf{0}$ vanwege de aanname, moeten we hebben dat

$$\lim_{k\to\infty} \lambda^k = \mathbf{0}$$

wat betekent dat $|\lambda| < 1$. Omdat dit waar moet zijn voor alle eigenwaarden van $\mathbf{A}$, concluderen we dat $\rho(\mathbf{A}) < 1$.

($\Leftarrow$) Van het *Jordan-normaalvorm* theorema weten we dat voor elke $\mathbf{A} \in \mathbb{C}^{n\times n}$ er een non-singuliere matrix $V \in \mathbb{C}^{n\times n}$ en een blok-diagonaalmatrix $J \in \mathbb{C}^{n\times n}$ in Jordan-normaalvorm bestaan zodanig dat

$$\mathbf{A} = \mathbf{V}\mathbf{J}\mathbf{V}^{-1}.$$

Het is makkelijk te zien dat

$$\mathbf{A}^k = \mathbf{V}\mathbf{J}^k\mathbf{V}^{-1}.$$

Als $\mathbf{A}$ s werkelijk verschillende eigenwaarden heeft, dan is J^k de welbekende blok-diagonaalmatrix

$$\mathbf{J}^k = \begin{bmatrix} J^k_{m_1}(\lambda_1) & 0 & 0 & \dots & 0 \\ 0 & J^k_{m_2}(\lambda_2) & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & J^k_{m_{s-1}}(\lambda_{s-1}) & 0 \\ 0 & \dots & \dots & 0 & J^k_{m_s}(\lambda_s) \end{bmatrix}.$$

Hierbij is $J^k_{m_i}(\lambda_i) \in \mathbb{C}^{i\times i}$ voor alle $i \leq s$. Een standaardresultaat voor de k-de macht van Jordanblokken van grootte $m_i \times m_i$ is het volgende.

$$\mathbf{J}^k_{m_i}(\lambda_i) = \begin{bmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \binom{k}{2}\lambda_i^{k-2} & \dots & \binom{k}{m_i-1}\lambda_i^{k-m_i+1} \\ 0 & \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \dots & \binom{k}{m_i-2}\lambda_i^{k-m_i+2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} \\ 0 & 0 & \dots & 0 & \lambda_i^k \end{bmatrix}.$$

Dus als $\rho(\mathbf{A}) < 1$, dan $|\lambda_i| < 1$ voor alle i. Hieruit volgt

$$\lim_{k\to\infty} J^k_{m_i} = \mathbf{0} \quad \forall\ i \leq s.$$

Dit impliceert

$$\lim_{k\to\infty} J^k = \mathbf{0}.$$

Dus

$$\lim_{k\to\infty} \mathbf{A}^k = \lim_{k\to\infty} \mathbf{V}\mathbf{J}^k\mathbf{V}^{-1} = \mathbf{V}(\lim_{k\to\infty} \mathbf{J}^k)\mathbf{V}^{-1} = \mathbf{0}.$$

Daarentegen, als $\rho(\mathbf{A}) > 1$, dan is er tenminste één λ_i zodanig dat $|\lambda_i| > 1$. Dan is er tenminste één element in J die onbegrensd is als k toeneemt, wat de tweede bewering bewijst. □

De volgende stelling behandelt twee limieten van $||\mathbf{A}^k||$.

Stelling 14 (Gelfand's formule) *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$, dan geldt:*

$$\lim_{k\to\infty} ||\mathbf{A}^k|| = \rho(\mathbf{A})^k \tag{3.25}$$

en ook hebben we

$$\lim_{k\to 0} \frac{||\mathbf{A}^{k+1}||}{||\mathbf{A}^k||} = ||\mathbf{A}|| \tag{3.26}$$

Bewijs Kies $\varepsilon > 0$ willekeurig en beschouw de matrix

$$\tilde{\mathbf{A}} = (\rho(\mathbf{A}) + \varepsilon)^{-1}\mathbf{A}.$$

Dan is het duidelijk dat

$$\rho(\tilde{\mathbf{A}}) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) + \varepsilon} < 1.$$

Van stelling 13 weten we dat $\lim_{k\to\infty} \tilde{\mathbf{A}}^k = 0$. Dit betekent vanwege de limietdefinitie dat er een $N_1 \in \mathbb{N}$ bestaat zodanig dat voor alle $k \geq N_1$ geldt dat

$$||\tilde{\mathbf{A}}^k|| < 1 \;\Rightarrow\; ||\mathbf{A}^k|| < (\rho(\mathbf{A}) + \varepsilon)^k.$$

Bekijk nu de volgende matrix

$$\check{\mathbf{A}} = (\rho(\mathbf{A}) - \varepsilon)^{-1}\mathbf{A}.$$

Dan is het duidelijk dat

$$\rho(\check{\mathbf{A}}) = \frac{\rho(\mathbf{A})}{\rho(\mathbf{A}) - \varepsilon} > 1.$$

wat betekent dat $||\check{\mathbf{A}}^k||$ onbegrensd is. Dit betekent dat er een $N_2 \in \mathbb{N}$ bestaat zodanig dat voor alle $k \geq N_2$ geldt dat

$$||\check{\mathbf{A}}^k|| > 1 \;\Rightarrow\; ||\mathbf{A}^k|| > (\rho(\mathbf{A}) - \varepsilon)^k.$$

Neem $N := \max(N_1, N_2)$ en samen met de vorige resultaten krijgen we dat voor elke $\varepsilon > 0$ er een $N \in \mathbb{N}$ bestaat zodanig dat voor alle $k \geq N$ geldt dat

$$(\rho(\mathbf{A}) - \varepsilon)^k < ||\mathbf{A}^k|| < (\rho(\mathbf{A}) + \varepsilon)^k.$$

Dat per definitie betekent dat $\lim_{k\to\infty} ||\mathbf{A}^k|| = \rho(\mathbf{A})^k$. De tweede ongelijkheid is min of meer triviaal.

$$\lim_{k\to 0} \frac{||\mathbf{A}^{k+1}||}{||\mathbf{A}^k||} = \frac{||\mathbf{A}^1||}{||I||} = ||\mathbf{A}||. \;\square$$

Met de laatste stelling en in het bijzonder met (3.25) hebben we een interessant detail blootgelegd.

- Als $\rho(\mathbf{A}) > 1$, dan $\lim_{k\to\infty} ||\mathbf{A}^k|| = \infty$.
- Als $\rho(\mathbf{A}) = 1$, dan $\lim_{k\to\infty} ||\mathbf{A}^k|| = 1$.
- Als $0 \leq \rho(\mathbf{A}) < 1$, dan $\lim_{k\to\infty} ||\mathbf{A}^k|| = 0$.

Dus als alle eigenwaarden binnen de eenheidscirkel liggen, dan hebben we dat $\lim_{k\to\infty} ||\mathbf{A}^k|| = 0$. We gebruiken nu definitie 5 op pagina 16 en stellen de Kreiss constante van $\mathbf{A}$ op t.o.v. de eenheidscirkel.

$$\mathcal{K}(\mathbf{A}) = \sup_{|z|>1} (|z| - 1)||(zI - \mathbf{A})^{-1}|| \tag{3.27}$$

Tenslotte bekijken we nog de relatie tussen $||\mathbf{A}^k||$ en $\mathcal{K}(\mathbf{A})$.

Stelling 15 *Laat $\mathbf{A} \in \mathbb{C}^{n\times n}$. Als $||(zI - \mathbf{A})^{-1}|| = K/(|z| - 1)$ voor een z met $|z| = r > 1$ en $K > 1$, dan*

$$\sup_{k\geq 0} ||\mathbf{A}^k|| \geq rK - r + 1 > K. \tag{3.28}$$

Als $\mathcal{K}(\mathbf{A})$ de Kreiss constante t.o.v. de eenheidscirkel is, dan geldt

$$\sup_{k\geq 0} ||\mathbf{A}^k|| \geq \mathcal{K}(\mathbf{A}). \tag{3.29}$$

BEWIJS Neem aan dat $\sup_{k\geq 0}||\mathbf{A}^k|| = M$, dan weten we van (18) dat voor elke z met $|z| = r > 1$,

$$\frac{rK}{r-1} = r||(zI-\mathbf{A})^{-1}|| = r\cdot\left|\left|\frac{1}{z}\sum_{k=0}^{\infty} z^{-k}\mathbf{A}^k\right|\right| \leq ||I|| + M\sum_{k=1}^{\infty} z^{-k} = 1 + \frac{M}{r-1}.$$

Uitwerken van deze ongelijkheid geeft

$$M = \sup_{k\geq 0}||\mathbf{A}^k|| \geq rK - r + 1 = K + (r-1)(K-1) > K.$$

De laatste ongelijkheid volgt uit het gegeven dat $r, K > 1$, zodat $(r-1)(K-1) > 0$. Nu hebben we dus ongelijkheid (3.28). We weten van de definitie dat $K = (|z|-1)||(zI-\mathbf{A})^{-1}||$ en omdat dit geldt voor alle z met $z > 1$, geldt dit uiteraard voor zijn maximum. We vervolgen dus met

$$\sup_{k\geq 0}||\mathbf{A}^k|| > K = (|z|-1)||(zI-\mathbf{A})^{-1}|| \;\; \forall\, z : |z| > 1 \;\Rightarrow\; \sup_{k\geq 0}||\mathbf{A}^k|| \geq \mathcal{K}(\mathbf{A}).$$

Dit is precies ongelijkheid (3.29). □

4 Stabiliteit en pseudospectra

In de vorige hoofdstukken hebben we theorie opgebouwd over het ε-pseudospectrum, de matrixexponent en $||\mathbf{A}^k||$. In dit hoofdstuk gaan we kijken naar oplossingen van stelsels gewone differentiaalvergelijkingen. Met behulp van Runge-Kuttamethoden kunnen we de oplossingen van deze stelsels numeriek benaderen. Hierbij moeten we een tijdstap kiezen en het liefst zodanig dat de fout convergeert naar 0. Dit verschijnsel noemen we stabiliteit. Over het algemeen werken de bekende methoden om een dergelijke tijdstap te vinden goed, maar met behulp van pseudospectra laat dit hoofdstuk zien dat dit niet altijd werkt. Dit doen we aan de hand van zelfbedachte voorbeelden.
Eén van de Runge-Kuttamethoden is de methode van *Euler voorwaarts* waarbij de berekeningen snel kunnen worden gemaakt door een computer. Een groot nadeel van deze methode is dat er kleine tijdstappen Δt moeten worden genomen wil de benadering goed zijn. Runge-Kutta methoden van hogere orde zoals *modified Euler* of *Runge-Kutta 4* vereisen meer rekentijd, maar zijn nauwkeuriger en behoeven daardoor niet een dergelijk kleine tijdstap als bij Euler voorwaarts.

4.1 Eigenwaardenonderzoek

In het schema hieronder staat weergegeven hoe een stelsel differentiaalvergelijkingen kan worden gereduceerd tot een scalarprobleem.

1. $\mathbf{y}' = f(t, \mathbf{y})$	
$\downarrow$	Lineariseren
2. $\mathbf{u}' = \mathbf{A}(t)\mathbf{u}$	
$\downarrow$	Coëfficiënten bevriezen
3. $\mathbf{u}' = \mathbf{A}\mathbf{u}$	
$\downarrow$	Diagonaliseren
4. $\mathbf{u}' = \lambda\mathbf{u}$	

We beginnen met een stelsel van n eerste orde differentiaalvergelijkingen $\mathbf{y}' = f(t, \mathbf{y})$, waarbij $\mathbf{y} \in \mathbb{C}^n$ en $t \in \mathbb{R}$. We zijn geïnteresseerd in de oplossing $y_0(t)$ van dit systeem. Als we de volgende substitutie maken: $\mathbf{y}(t) = y_0(t) + \mathbf{u}(t)$, dan hangt instabiliteit van het probleem alleen af van de ontwikkeling van $\mathbf{u}(t)$.
Als eerste lineariseren we de vergelijking door aan te nemen dat $\mathbf{u}$ klein is. Als f differentieerbaar is, dan is

$$\mathbf{A}(t) = \frac{\partial f}{\partial \mathbf{y}}(t, y_0(t))$$

de Jacobiaan op tijdstip t met $\mathbf{A}(t) \in \mathbb{C}^{n\times n}$. Door de termen van orde u^2 en hoger weg te laten en met behulp van de identiteit $y_0'(t) = f(t, y_0(t))$ komen we aan bij $\mathbf{u}' = \mathbf{A}(t)\mathbf{u}$.
De volgende stap die we maken is het bevriezen van de coëfficiënten door te zeggen: $\mathbf{A} = \mathbf{A}(t_0)$ voor een $t = t_0$ waar wij in geïnteresseerd zijn. Het resultaat is lineaire probleem $\mathbf{u}' = \mathbf{A}\mathbf{u}$ met constante coëfficiënten.
De laatste stap die we nemen is het diagonaliseren van $\mathbf{A}$, aannemende dat $\mathbf{A}$ diagonaliseerbaar is. Dit verdeelt de laatste vergelijking in n onafhankelijke vergelijkingen $\mathbf{u}' = \lambda\mathbf{u}$ met

$\lambda \in \sigma(\mathbf{A})$. Met deze problemen kunnen we bepalen of het probleem stabiel is. Als alle eigenwaarden van $\mathbf{A}$ binnen het stabiliteitsgebied van de gebruikte methode liggen, dan noemen we het probleem stabiel en krijgen we dus een betrouwbare numerieke oplossing. We zullen zien dat naarmate we bij oplossingsmethoden zoals de Runge-Kuttamethoden de tijdstap Δt verkleinen, dat de stabiliteit van ons probleem wordt vergroot.

4.2 Stabiliteit van de numerieke oplossing

Eigenwaarden kunnen veel zeggen over de stabiliteit van een numerieke oplossing rond een bepaald tijdstip $t = t_0$. De geldigheid van de conclusies die we trekken naar aanleiding van deze eigenwaarden moet helaas in twijfel worden getrokken als de algebraïsche multipliciteit en de meetkundige multipliciteit van $\mathbf{A}$ niet gelijk aan elkaar zijn. In die situatie is er tenminste één eigenwaarde die meer dan eens voorkomt. In veel gevallen leidt dit tot het geval waarbij we geen volledige basis van eigenvectoren hebben, waardoor we de matrix niet meer kunnen diagonaliseren.
Als we bedenken dat eigenwaarden in 'normale' gevallen veel zeggen over de stabiliteit van een numerieke oplossing, lijkt het ook interessant om de relatie tussen de pseudospectra van $\mathbf{A}$ en de stabiliteit van de numerieke oplossing te bekijken. In dit onderzoek beperken wij ons tot het derde probleem in het schema.

$$\mathbf{u}' = \mathbf{A}\mathbf{u}$$

We zullen zien dat de pseudospectra ons uitsluitsel kunnen geven over de stabiliteit van de numerieke oplossing van dit probleem in het geval dat we geen volledige basis van eigenvectoren hebben.

Beschouw het lineaire, homogene probleem met constante coëfficiënten $\mathbf{u}' = \mathbf{A}\mathbf{u}$, waarbij $\mathbf{A} \in \mathbb{C}^{n\times n}$. Laat de oplossing van dit systeem worden benaderd met een Runge-Kuttamethode met constante tijdstap Δt. We willen een stabiele oplossing en dit betekent dat de afbreekfout bij elke stap afneemt. We definiëren hierbij de volgende foutrelatie voor de afbreekfout: $\mathbf{v}^{(n+1)} = Q(\mathbf{A}\Delta t)\mathbf{v}^{(n)}$, of wat algemener

$$\mathbf{v}^{(n)} = Q(\mathbf{A}\Delta t)^{n-n_0}\mathbf{v}^{(n_0)} \quad , n \in \mathbb{N}.$$

Hier is $Q(z) = 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^m}{m!}$ het m^{de}-machts Taylorpolynoom van e^z voor een Runge-Kuttamethode van orde m. We kunnen de relatie ook schrijven als

$$\mathbf{v}(t) = Q(\mathbf{A}\Delta t)^{(t-t_0)/\Delta t}\mathbf{v}(t_0),$$

als t en t_0 meervouden zijn van Δt. Doordat $Q(\mathbf{A}\Delta t)^{(t-t_0)/\Delta t} \approx e^{\mathbf{A}\Delta t(t-t_0)/\Delta t} = e^{(t-t_0)\mathbf{A}}$ zien we nu makkelijk dat het bovenstaande een discrete benadering is van

$$\mathbf{v}(t) = e^{(t-t_0)\mathbf{A}}\mathbf{v}(t_0).$$

Omdat de coëfficiënten constant zijn kunnen we voor het onderzoek gewoon aannemen dat $n_0 = t_0 = 0$ en krijgen uiteindelijk

$$v(t) = e^{t\mathbf{A}}v(0)$$

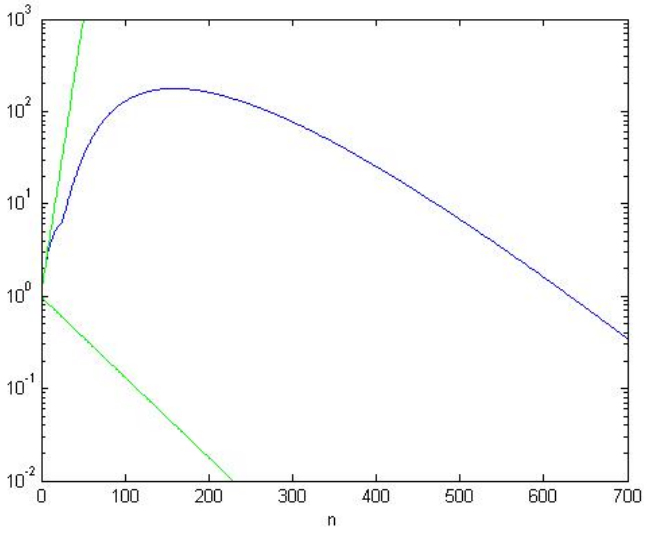
Omdat we graag willen dat de fout al direct erg klein is, zijn we erin geïnteresseerd hoe $Q(\mathbf{A}\Delta t)^n$ zich gedraagt als een functie van n. Het logische gevolg is dat we gaan kijken naar

de norm $||Q(\mathbf{A}\Delta t)^n||$. Deze norm hebben we al eerder afgeschat in §3.3 en we herhalen deze nog eens hieronder.

$$\rho(Q(\mathbf{A}\Delta t))^n \leq ||Q(\mathbf{A}\Delta t)^n|| \leq ||Q(\mathbf{A}\Delta t)||^n.$$

Helaas is het verschil tussen deze twee grenzen vaak erg groot als $\mathbf{A}$ geen normale matrix is. Als voorbeeld hierbij beschouwen we de matrix $\mathbf{C} \in \mathbb{R}^{4\times 4}$ die er als volgt uitziet.

$$\mathbf{C} = \begin{bmatrix} -20 & 2 & 2 & 0 \\ 0 & -20 & 2 & 2 \\ 0 & 0 & -20 & 2 \\ 0 & 0 & 0 & -20 \end{bmatrix}$$



Figuur 15: De grafieken van $||Q(\mathbf{C}\Delta t)^n||$, $\rho(Q(\mathbf{C}\Delta t))^n$ en $||Q(\mathbf{C}\Delta t)||^n$ voor $0 \leq n \leq 300$

Bij dit voorbeeld is de methode van Modified Euler gebruikt. Deze methode is stabiel als $|Q(\lambda\Delta t)| = |1 + \lambda\Delta t + \frac{(\lambda\Delta t)^2}{2}| < 1$ in het geval dat $\mathbf{C}$ een volledige basis van eigenvectoren heeft. Hierbij noemen we $Q(\lambda\Delta t)$ de versterkingsfactor van deze methode.
Deze matrix heeft eigenwaarde -20 met multipliciteit 4 en slechts één eigenvector $v_1 = (1, 0, 0, 0)'$. We gebruikten bij dit voorbeeld $\Delta t = 0.099$, zodat $|Q(\lambda\Delta t)| = 0.9802 < 1$. Het lijkt er dus op dat we de tijdstap Δt klein genoeg hebben gekozen, maar de grote bult in de grafiek wijst erop dat de afbreekfout erg groot is in het begin.
We hebben nu dat $\rho(Q(\mathbf{C}\Delta t)) = 0.9802$ en $||Q(\mathbf{C}\Delta t)|| \approx 1.1443$. In de figuur hierboven staat de grafiek van $||Q(\mathbf{C}\Delta t)^n||$ weergegeven met een blauwe lijn en de twee grenzen zijn in het groen gegeven.

Als we een dergelijk grote bult krijgen te zien zoals in figuur 15, dan zal de numerieke methode instabiel zijn vanwege de grote afbreekfout ondanks schijnbaar gunstige eigenwaarden van $Q(\mathbf{A}\Delta t)$. Als de bult klein is, dan zal de numerieke methode stabiel zijn ondanks een ongunstige norm. Het is dus de grootte van de bult die ertoe doet ofwel het gedrag van $||Q(\mathbf{A}\Delta t)^n|| = ||Q(\mathbf{A}\Delta t)^{t/\Delta t}||$ voor kleine, maar grotere waarden dan 0 van t.
Het valt de lezer misschien op dat de grafiek van $||Q(\mathbf{A}\Delta t)^n||$ in het begin even de bovengrens

volgt, maar later convergeert naar de ondergrens. Dit gegeven is in §3.3 al bewezen en het resultaat herhalen we hieronder nog kort.

- $\lim_{k\to\infty} ||\mathbf{A}^k|| = \rho(\mathbf{A})^k$
- $\lim_{k\to 0} \frac{||\mathbf{A}^{k+1}||}{||\mathbf{A}^k||} = ||\mathbf{A}||$

We kunnen dus de volgende conclusies trekken over het gedrag van $||Q(\mathbf{A}\Delta t)^n||$.

- Het gedrag voor $n \to \infty$ wordt bepaald door het spectrum van $Q(\mathbf{A}\Delta t)$ of van $\mathbf{A}\Delta t$.
- Het gedrag voor $n \to 0$ wordt bepaald door $||Q(\mathbf{A}\Delta t)||$.

Maar hoe zit dit voor eindige n? Het vorige voorbeeld suggereert dat er nog iets anders invloed heeft op het gedrag van $||Q(\mathbf{A}\Delta t)^n||$ dan de eigenwaarden of de norm van $Q(\mathbf{A}\Delta t)$. Al eerder is er een voorschot gegeven hiervoor; we gaan kijken naar de pseudospectra van $\mathbf{A}\Delta t$. De volgende stelling laat zien dat het gedrag van $||Q(\mathbf{A}\Delta t)^n||$ voor eindige n wordt bepaald door de pseudospectra van $\mathbf{A}\Delta t$. We gebruiken hierbij de Kreiss-constante t.o.v. de eenheidscirkel:

$$\mathcal{K}(\mathbf{A}) = \sup_{|z|>1} (|z| - 1)||(zI - \mathbf{A})^{-1}||.$$

Stelling 16 (Kreiss matrix theorema) *Laat $\mathbf{A} \in \mathbb{C}^{N\times N}$ en $\mathcal{K}(\mathbf{A})$ als hierboven, dan*

$$\mathcal{K}(\mathbf{A}) \leq \sup_{t\geq 0} ||\mathbf{A}^k|| \leq eN\mathcal{K}(\mathbf{A}) \tag{4.1}$$

Bewijs De eerste ongelijkheid hebben we al bewezen bij stelling 15 in §3.3, dus we gaan verder met de moeilijkere rechterongelijkheid. We beginnen dit bewijs met het volgende:

$$\mathbf{A}^k = \frac{1}{2\pi i} \int_\Gamma z^k (zI - \mathbf{A})^{-1} dz,$$

waarbij Γ een gesloten contour is met $\sigma(\mathbf{A})$ in zijn inwendige. We kunnen $\mathbf{u}, \mathbf{v} \in \mathbb{C}^N$ vinden met $||\mathbf{u}|| = ||\mathbf{v}|| = 1$ zodanig dat $||\mathbf{v}^*\mathbf{A}^k\mathbf{u}|| = ||\mathbf{A}^k||$. We doen het volgende:

$$\mathbf{v}^*\mathbf{A}^k\mathbf{u} = \frac{1}{2\pi i} \int_\Gamma z^k r(z) dz.$$

Hierbij gebruiken we $r(z) = \mathbf{v}^*(zI - \mathbf{A})^{-1}\mathbf{u}$, dat een rationale functie van orde N is. Een rationale functie van orde n is een quotïent van twee polynomen van elk graad $\leq N$. Partiële integratie geeft

$$\mathbf{v}^*\mathbf{A}^k\mathbf{u} = -\frac{1}{2\pi i(k+1)} \int_\Gamma z^{k+1} r'(z) dz.$$

We nemen nu $\Gamma = \{z \in \mathbb{C} : |z| = 1 + (k+1)^{-1}\}$. Op deze contour hebben we $|z^{k+1}| = |(1 + \frac{1}{k+1})^{k+1}| \leq e$, zodat we de volgende bovengrens hebben.

$$|\mathbf{v}^*\mathbf{A}^k\mathbf{u}| \leq \frac{e}{2\pi(k+1)} \int_\Gamma |r'(z)||dz|$$

De integraal kan worden gezien als de arclengte van het beeld van Γ onder r. In [1] op pagina 180 zijn de definitie en het bewijs van *Spijker's Lemma* te vinden. Volgens dit lemma voldoet deze arclengte aan

$$\int_\Gamma |r'(z)||dz| \leq 2\pi N \sup_{z\in\Gamma} |r(z)|.$$

Het supremum binnen de integraal schatten we op de volgende handige manier af.

$$\begin{aligned}
\sup_{z\in\Gamma} |r(z)| &= \sup_{z\in\Gamma} ||\mathbf{v}^*(zI-\mathbf{A})^{-1}\mathbf{u}|| \leq \sup_{z\in\Gamma} ||(zI-\mathbf{A})^{-1}|| \\
&= \sup_{z\in\Gamma} \frac{1}{|z|-1}(|z|-1)||(zI-\mathbf{A})^{-1}|| \\
&\leq \frac{1}{(1+(k+1)^{-1})-1} \sup_{|\alpha|>1} (|\alpha|-1)||(\alpha I-\mathbf{A})^{-1}|| \qquad (4.2) \\
&= (k+1)\mathcal{K}(\mathbf{A})
\end{aligned}$$

Ongelijkheid (4.2) is verkregen doordat $|z| = 1+(k+1)^{-1}$ op Γ en dus ook $|z|>1$ voor alle $k\in\mathbb{N}$. Alles bij elkaar hebben we nu

$$|\mathbf{v}^*\mathbf{A}^k\mathbf{u}| \leq \frac{e}{2\pi(k+1)} \cdot 2\pi N(k+1)\mathcal{K}(\mathbf{A}) = eN\mathcal{K}(\mathbf{A}).$$

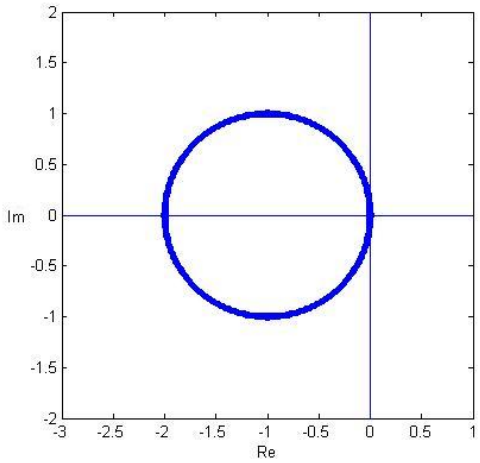
Omdat $|\mathbf{v}^*\mathbf{A}^k\mathbf{u}| = ||\mathbf{v}^*\mathbf{A}^k\mathbf{u}|| \leq ||\mathbf{v}^*||\,||\mathbf{A}^k||\,||\mathbf{u}|| = ||\mathbf{A}^k||$, is $||\mathbf{A}^k||$ het supremum van $|\mathbf{v}^*\mathbf{A}^k\mathbf{u}|$. Met dit gegeven hebben we de rechterongelijkheid in (4.1) bewezen. □

De lezers die meer willen weten over de scherpte van de bovengrens in (4.1) verwijzen we naar [1] op pagina's 176-179.
In §3.3 hebben we al besproken dat voor alle eigenwaarden λ van $\mathbf{A}$ moet gelden dat $|\lambda|<1$ als we willen dat $\lim_{k\to\infty}||\mathbf{A}^k||$. In de praktijk willen we uiteraard dat dit ook geldt voor $||Q(\mathbf{A}\Delta t)^n||$, want anders hebben we geen stabiele tijdstap. $Q(\mathbf{A}\Delta t)$ heeft voor elke methode ander functievoorschrift , dus het gebied waarin de eigenwaarden $\lambda\Delta t$ moeten liggen zodat $\lim_{n\to\infty}||Q(\mathbf{A}\Delta t)^n||$ is voor elke methode ook anders. Wij noemen dit het stabiliteitsgebied voor de betreffende methode en dit gebied geven we aan met Ω.
Voor Euler voorwaarts hebben we bijvoorbeeld de versterkingsfactor $|Q(\lambda\Delta t)| = |1+\lambda\Delta t|$. Het gebied dat wordt beschreven door $|1+\lambda\Delta t|<1$ is de cirkel met middelpunt -1 en straal 1. Dit stabiliteitsgebied staat in figuur 16 op de volgende pagina gegeven. Het stabiliteitsgebied van Modified Euler is moeilijker te beschrijven, maar staat figuur 17 gegeven.
We kunnen ook voor deze stabiliteitsgebieden de Kreiss constante opstellen, omdat $||Q(\mathbf{A}\Delta t)^n||$ daar convergeert naar 0. Dus noem Ω het stabiliteitsgebied behorende bij een bepaalde eenstapsmethode, dan definiëren we de Kreiss constante t.o.v. Ω als

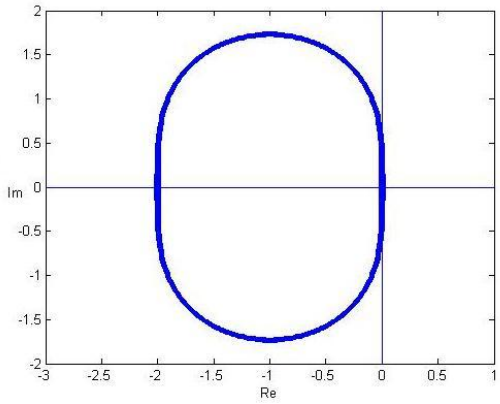
$$\mathcal{K}(\mathbf{A}) := \sup_{z\notin\Omega} \operatorname{dist}(z,\Omega)||(zI-\mathbf{A})^{-1}|| = \sup_{\varepsilon>0} \varepsilon^{-1}\overline{\operatorname{dist}}(\sigma_\varepsilon(\mathbf{A}),\Omega), \qquad (4.3)$$

waarbij $\overline{\operatorname{dist}}(A,B)$ hetzelfde betekent als $\sup_{z\in A}\operatorname{dist}(z,B)$. Dan hebben we volgens het Kreiss matrix theorema het volgende.

$$\mathcal{K}(\mathbf{A}\Delta t) \leq \sup_{k\geq 0} ||Q(\mathbf{A}\Delta t)^n|| \leq eN\mathcal{K}(\mathbf{A}\Delta t), \quad \mathcal{K}(\mathbf{A}\Delta t) = \sup_{\varepsilon>0} \varepsilon^{-1}\overline{\operatorname{dist}}(\sigma_\varepsilon(\mathbf{A}\Delta t),\Omega). \qquad (4.4)$$

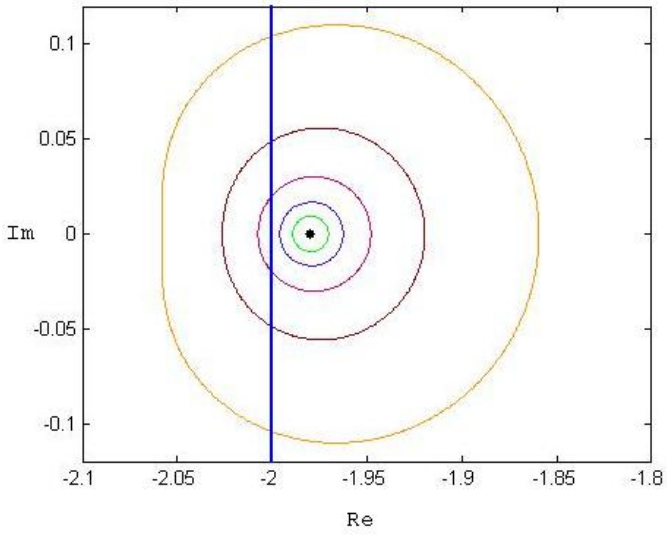


Figuur 16: Het stabiliteitsgebied in het $\lambda\Delta t$ vlak van Euler voorwaarts.



Figuur 17: Het stabiliteitsgebied in het $\lambda\Delta t$ vlak van Modified Euler.

We kunnen dit opmerkelijke resultaat als volgt verwoorden: de grootte van de bult (zoals in figuur 15) wordt bepaald door hoe ver de ε-pseudospectra van $\mathbf{A}\Delta t$ buiten het stabiliteitsgebied liggen.



Figuur 18: De ε-pseudospectra $\sigma_\varepsilon(\mathbf{C}\Delta t)$ met $\varepsilon = 10^{-2}, \ldots, 10^{-6}$ voor dezelfde matrix $\mathbf{C}$ als in figuur 15. De blauwe lijn is de rand van het stabiliteitsgebied Ω bij Modified Euler.

Figuur 18 illustreert deze theorie waarbij we dezelfde matrix $\mathbf{C}$ gebruiken als die bij figuur 15 is gebruikt. De blauwe lijn geeft de rand van het stabiliteitsgebied Ω aan bij Modified Euler. Verder zijn ook de randen van $\sigma_\varepsilon(\mathbf{C}\Delta t)$ getekend met $\varepsilon = 10^{-2}, \ldots, 10^{-6}$ en $\Delta t = 0.099$. De punt in het midden zijn de eigenwaarden $\lambda = -1.98$ weergegeven. Door de contouren te tellen is makkelijk te zien dat $\sigma_{10^{-5}}(\mathbf{C}\Delta t)$ nog volledig is bevat in Ω, maar $\sigma_{10^{-4}}(\mathbf{C}\Delta t)$ voor het eerst (met $\varepsilon \in \mathbb{N}$) niet. Dit betekent dat de bult in de grafiek van $||Q(\mathbf{C}\Delta t)^n||$ in figuur 15 van orde 10^4 is.

Nou lijkt het in figuur 15 alsof de fout eerder van orde 10^2 zou moeten zijn, maar dan moet er ook rekening worden gehouden met de grootte van $\sup_{\varepsilon>0} \overline{\text{dist}}(\sigma_\varepsilon(\mathbf{C}\Delta t), \Omega)$. Deze is bij $\varepsilon = 10^{-4}$ ongeveer gelijk aan 0.01, waardoor de orde van de grootte van de bult een factor 10^2 kleiner lijkt. Als we een matrix kunnen vinden van dezelfde grootte en met dezelfde eigenwaarden, maar waarvan de pseudospectra zich verder uitstrekken, dan zal deze bult groter zijn dan in figuur 15.

4.3 De convectie-diffusievergelijking

In deze pargraaf gaan we het Kreiss matrix theorema toepassen op de homogene convectie-diffusievergelijking met constante coëfficiënten. Deze partiële differentiaalvergelijking heeft over het algemeen geen exacte oplossing, maar is toch een veel voorkomend probleem waar technisch wiskundigen mee te maken hebben. In deze paragraaf bekijken we het volgende probleem voor de functie $u(x,t)$.

$$\frac{\partial u}{\partial t} = -\nu\frac{\partial u}{\partial x} + \epsilon\frac{\partial^2 u}{\partial x^2}, \quad u(0,t) = 0,\ u(1,t) = 1,\ u(x,0) = w(x). \tag{4.5}$$

Hierbij is ν de convectieconstante en ϵ de diffusieconstante. Met behulp van eem semi-discretisatie maken we hiervan een stelsel gewone differentiaalvergelijkingen, maar laten we eerst nog even kijken naar de betekenis van de termen convectie en diffusie.

4.3.1 Convectie

Convectie is de stroming van gas of vloeistof. Deze stroming kan plaats vinden onder invloed van onder meer verschillen in temperatuur, druk of dichtheid. Zo treedt convectie op in de aardatmosfeer waar warme lucht van de door de zon verwarmde bodem opstijgt. Op een andere plek daalt koude lucht juist af naar beneden. Convectiestromen zijn dus altijd gesloten. Convectie wordt in de techniek gebruikt in bijvoorbeeld de convectorput, een vorm van verzonken centrale verwarming. Bij conventionele kachels treedt ook vooral convectie op om een ruimte te verwarmen. Dit is ook het geval bij radiatoren. Ook al lijkt de naam radiator erop te wijzen dat er vooral warmtestraling geleverd wordt, dit is onjuist, er treden luchtstromingen op die de warmte door de ruimte verspreiden. De luchtstromingen treden op in cellen.
De convectieterm ν in (4.4) geeft de snelheid aan van het vloeistof of gas. Als $\nu > 0$, dan verplaatst het zich naar rechts en als $\nu < 0$, dan verplaatst het zich naar links. In het geval dat $\nu = 0$ vind er geen stroming plaats.

4.3.2 Diffusie

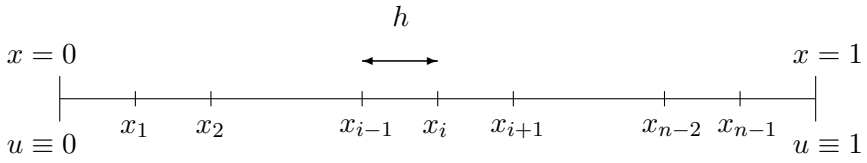
Diffusie is een proces als gevolg van de willekeurige beweging van deeltjes. Deze willekeurige beweging is het gevolg van de kinetische energie die deeltjes bezitten. Bij verschillen in concentratie leidt diffusie tot een netto verplaatsing van deeltjes van plaatsen met een hoge concentratie naar plaatsen met een lage concentratie. Denk bijvoorbeeld aan het maken van thee. Nadat het theezakje in het water is gedaan, zie je het thee-extract alle kanten op kruipen. Er vind dus een verplaatsing van hoge naar lage concentratie van het thee-extract plaats en dus vind er diffusie plaats en deze uitdeining vind plaats zonder een significante stroming.
Hoe groter ϵ, hoe sneller de diffusie plaatsvind. Als $\epsilon < 0$, dan is er een juist een stroming van lage naar hoge concentratie. Deze situatie behandelen we in dit onderzoek niet.

4.3.3 Convectie en diffusie

Een situatie waarbij duidelijk convectie en diffusie plaatsvind is bijvoorbeeld als een druppel kleurstof wordt gegoten in stromend water. De druppel wordt niet alleen meegesleurd met het water (convectie), maar zal ook uitdeinen in een andere richting dan de richting van de stroming.

4.4 Numeriek oplossen van de convectie-diffusievergelijking

Zoals eerder gezegd gaan we een semi-discretisatie maken van vergelijking (4.4). We gebruiken hiervoor centrale differentie in de eerste- en tweede afgeleide naar de plaats x. We maken een equidistant grid over het interval $[0, 1]$ met stapgrootte $h = 1/n$, zoals hieronder is weergegeven.



Hierbij hebben we $x_i = ih$ en $u_i = u(x_i)$.
Na gebruik te hebben gemaakt van centrale differentie, verandert (4.4) in de volgende semi-discretisatie op plaats x_i.

$$\begin{aligned}\left(\frac{\partial u}{\partial t}\right)_i &= -\nu\left(\frac{-u_{i-1}+u_{i+1}}{2h}\right)+\epsilon\left(\frac{u_{i-1}-2u_i+u_{i+1}}{h^2}\right)\\ &= \left(\frac{\nu}{2h}+\frac{\epsilon}{h^2}\right)u_{i-1}-\frac{2\epsilon}{h^2}u_i+\left(-\frac{\nu}{2h}+\frac{\epsilon}{h^2}\right)u_{i+1}\\ &= \left(\frac{\nu h+2\epsilon}{2h^2}\right)u_{i-1}-\frac{2\epsilon}{h^2}u_i+\left(\frac{2\epsilon-\nu h}{2h^2}\right)u_{i+1}\end{aligned} \tag{4.6}$$

Voor een makkelijkere notatie spreken we af dat $\left(\frac{\partial u}{\partial t}\right)_i = u_i'$. Als we de randvoorwaarden gebruiken in (4.5) bij de punten x_1 en x_{n-1}, dan krijgen we het volgende stelsel gewone differentiaalvergelijkingen.

$$\begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_{n-2}' \\ u_{n-1}' \end{bmatrix} = \begin{bmatrix} -\frac{2\epsilon}{h^2} & \frac{2\epsilon-\nu h}{2h^2} & 0 & \cdots & 0 \\ \frac{\nu h+2\epsilon}{2h^2} & -\frac{2\epsilon}{h^2} & \frac{2\epsilon-\nu h}{2h^2} & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & \frac{\nu h+2\epsilon}{2h^2} & -\frac{2\epsilon}{h^2} & \frac{2\epsilon-\nu h}{2h^2} \\ 0 & \cdots & 0 & \frac{\nu h+2\epsilon}{2h^2} & -\frac{2\epsilon}{h^2} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{2\epsilon-\nu h}{2h^2} \end{bmatrix}$$

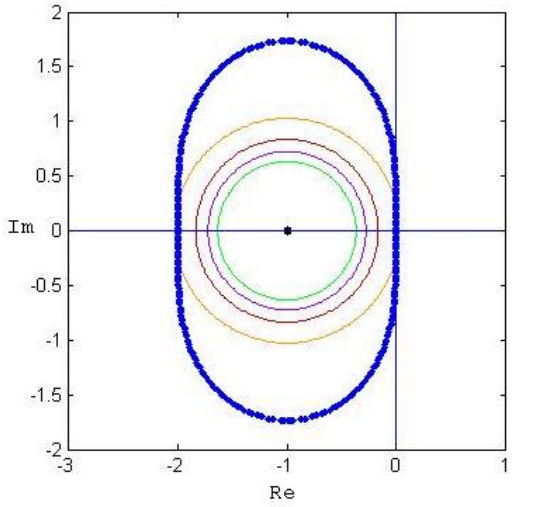
We korten dit af als $\mathbf{u'} = \mathbf{Au} + \mathbf{f}$. We kiezen ervoor om er een onderdiagonaalmatrix van te maken door $\nu = \frac{2\epsilon}{h}$ te kiezen, waardoor ook de vector $\mathbf{f}$ geen toevoeging geeft. Zo komen wij tot de gewenste vergelijking $\mathbf{u'} = \mathbf{Au}$, waarbij $\mathbf{A}$ de matrix is met $-\frac{2\epsilon}{h^2}$ op de hoofddiagonaal en $\frac{2\epsilon}{h^2}$ op de subdiagonaal.

4.4.1 Oplossen met Modified Euler

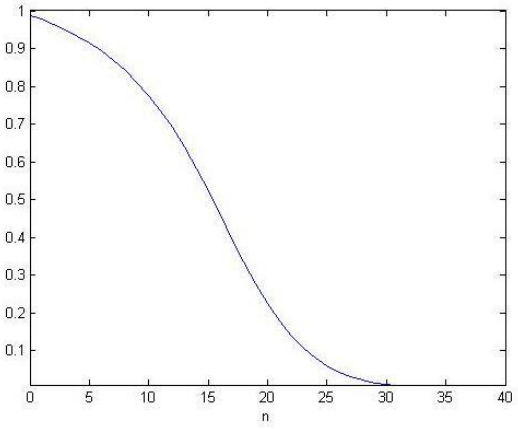
Om te beginnen kiezen we ervoor om 19 gridpunten te plaatsen tussen $x = 0$ en $x = 1$, zodat $h = \frac{1}{20}$. We kiezen ook $\epsilon = 0.1$, waardoor $\nu = 4$, en beginconditie $w(x_i) = 0.5$ voor $i = 1, \ldots, 19$. Dit brengt ons tot het volgende probleem.

$$\begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_{18}' \\ u_{19}' \end{bmatrix} = \begin{bmatrix} -80 & 0 & 0 & \cdots & 0 \\ 80 & -80 & 0 & \cdots & 0 \\ 0 & 80 & -80 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 80 & -80 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{18} \\ u_{19} \end{bmatrix}$$

Alle 19 eigenwaarden λ van $\mathbf{A}$ zijn gelijk aan -80. Voor Euler voorwaarts en Modified Euler kunnen we nu makkelijk zien hoe groot de tijdstap Δt op zijn hoogst mag zijn (willen we een stabiele tijdstap hebben), omdat voor reële λ gewoon Δt kiezen zodanig dat $-2 < \lambda \Delta t < 0$. De tijdstap mag op zijn hoogst gelijk zijn aan 0.025. Merk op dat als we meer gridpunten kiezen, dan neemt deze grootste tijdstap kwadratisch af. Laten we eerst een veilige tijdstap nemen zoals $\Delta t = 0.0125$, zodat de eigenwaarden van $\mathbf{A}\Delta t$ allemaal gelijk zijn aan -1.
Bij deze matrix hebben we $\kappa(\mathbf{V}) \approx 24.8$, dus de pseudopectra zullen zich een stuk verder uitstrekken rond de eigenwaarden dan bij een normale matrix. In figuur 19 op de volgende pagina is $\sigma_\varepsilon(\mathbf{A}\Delta t)$ getekend voor $\varepsilon = 10^{-4}, \ldots, 10^{-1}$ en de rand van het stabiliteistgebied Ω voor de methode van Modified Euler. Aangezien bij $\varepsilon = 10^{-1}$ voor het eerst de rand van Ω wordt overschreden en $\sup_{z \in \sigma_{10^{-1}}(\mathbf{A}\Delta t)} \text{dist}(z, \Omega)$ klein is, zal $||Q(\mathbf{A}\Delta t)^n||$ waarschijnlijk een kleine of zelfs geen bult hebben. Dit wordt bevestigt door figuur 20, waarin de grafiek hiervan is getekend. De fout neemt zelfs direct af.

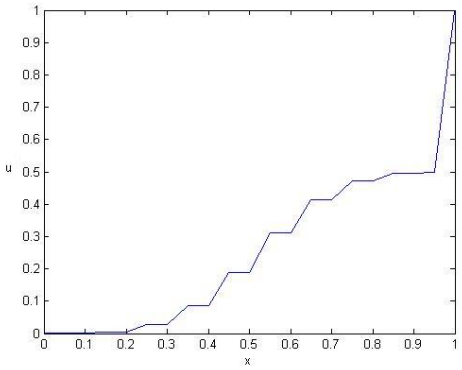


Figuur 19: Ω bij Modified Euler en pseudospectra van $\mathbf{A}\Delta t$.

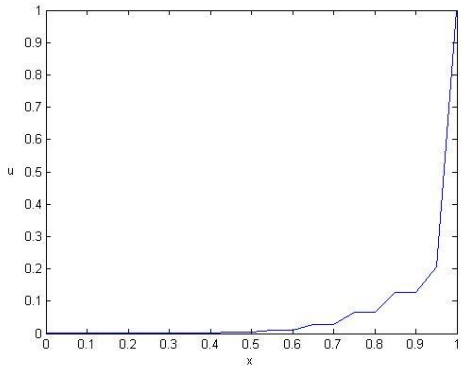


Figuur 20: $||Q(\mathbf{A}\Delta t)^n||$ voor $n = 0, \ldots, 40$.

In figuren 21 en 22 staat de numerieke oplossing van het probleem weergegeven na 10 en na 20 tijdstappen getekend. Dit is op $t = 0.125$ en $t = 0.25$. De numerieke oplossing convergeert vrij snel naar de stationaire oplossing en dit reflecteert ook de stabiliteit van de oplossing.



Figuur 21: De numerieke oplossing van (4.4) op $t = 0.125$.

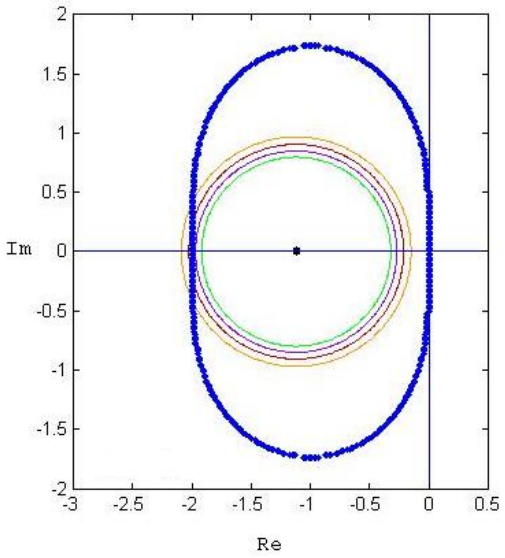


Figuur 22: De numerieke oplossing van (4.4) op $t = 0.25$.

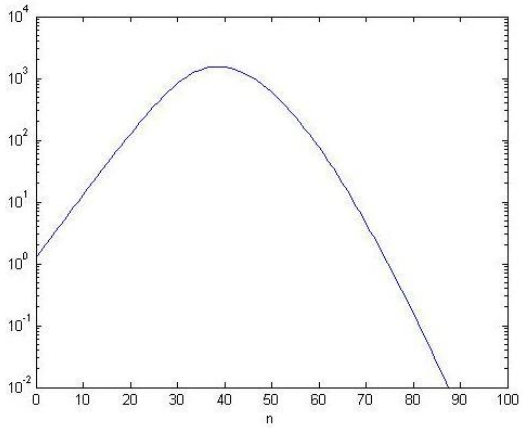
Doordat de stroming een positieve richting heeft, zien we dat de concentratie u het eerste afneemt aan de linkerkant. Uiteindelijk wordt de concentratie overal gelijk aan 0, behalve bij $x = 1$.
We gaan nu kijken naar een voorbeeld waarbij de afbreekfout erg groot wordt, ondanks schijnbaar gunstige eigenwaarden. Hierbij gebruiken we 39 gridpunten, zodat $h = 1/40$ en precies dezelfde convectieconstante ϵ, zodat $\nu = 8$. Het stelsel differentiaalvergelijkingen ziet er dan als volgt uit.

$$\begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_{38}' \\ u_{39}' \end{bmatrix} = \begin{bmatrix} -320 & 0 & 0 & \cdots & 0 \\ 320 & -320 & 0 & \cdots & 0 \\ 0 & 320 & -320 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 320 & -320 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{38} \\ u_{39} \end{bmatrix}$$

We kiezen $\Delta t = 0.0035$ zodat alle 39 eigenwaarden λ van $\mathbf{A}\Delta t$ gelijk zijn aan -1.12. Deze eigenwaarden liggen ver binnen het stabiliteitsgebied Ω van Modified Euler, zoals te zien is in figuur 23, maar in figuur 24 zien we dat de afbreekfout erg groot is voor de eerste 60 tijdstappen. Bij deze tijdstap Δt is het probleem dus niet stabiel.
Bij deze matrix hebben we $\kappa(\mathbf{V}) \approx 50.3$ en dat is een stuk groter dan bij het vorige voorbeeld. De pseudospectra zullen veel eerder problemen veroorzaken voor stabiliteit. In figuur 23 zijn de randen van $\sigma_\varepsilon(\mathbf{A}\Delta t)$ getekend voor $\varepsilon = 10^{-6}, \ldots, 10^{-3}$ Bij $\varepsilon = 10^{-5}$ overschrijdt deze grens de rand van Ω, dus rekening houdend met $\sup_{z \in \sigma_{10^{-5}}(\mathbf{A}\Delta t)} \operatorname{dist}(z, \Omega)$ komen we dus uit op een bult met zijn maximum groter dan 10^3. Ondanks de grote bult in de grafiek van $||Q(\mathbf{A}\Delta t)^n||$, wordt de afbreekfout na 70 tijdstappen weer klein. Dit is precies wat Gelfand's formule inhoudt. Uiteindelijk convergeert de numerieke oplossing dus naar de stationaire oplossing van het probleem.

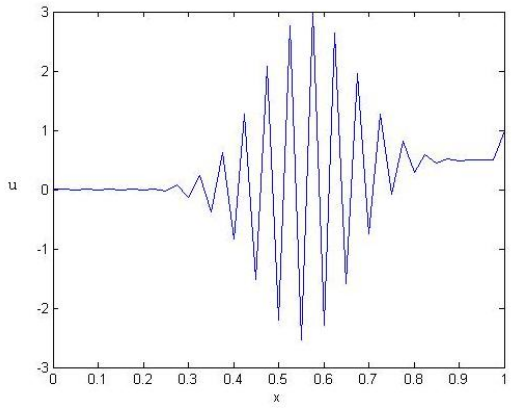


Figuur 23: Ω bij Modified Euler en pseudospectra van $\mathbf{A}\Delta t$.

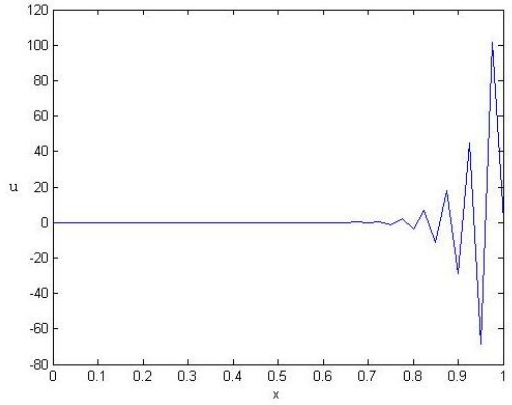


Figuur 24: $||Q(\mathbf{A}\Delta t)^n||$ voor $n = 0, \ldots, 100$.

In figuren 25 en 26 staan de numerieke oplossingen weergegeven na 20 en na 50 tijdstappen. Dit is op respectievelijk $t = 0.28$ en $t = 0.7$. De waarden van u worden onrealistisch groot en op veel punten ook negatief. Er is goed te zien dat de grote verschillen tussen functiewaarden die naast elkaar liggen, elkaar versterken; op $t = 0.7$ geeft de numerieke oplossing zelfs waarden die groter zijn dan 100, terwijl dat op $t = 0.28$ nog 'slechts' 3 was. Een leuke bijkomstigheid bij deze twee weergaven van de numerieke oplossing is dat heel goed te zien is dat de stroming een positieve richting heeft.



Figuur 25: De numerieke oplossing van (4.4) op $t = 0.28$.



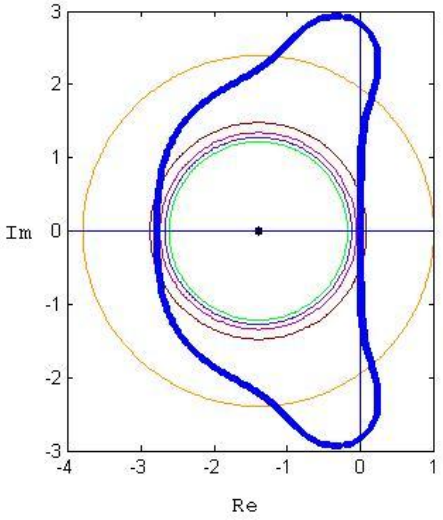
Figuur 26: De numerieke oplossing van (4.4) op $t = 0.7$.

4.4.2 Oplossen met Runge-Kutta 4

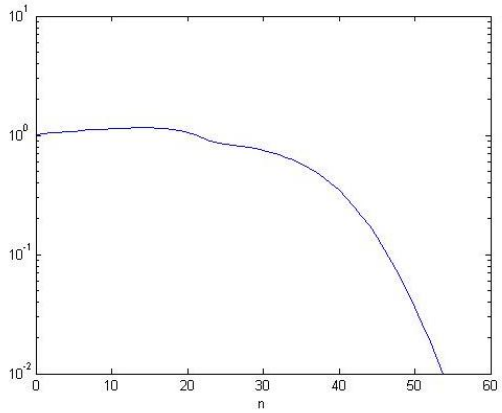
Als we figuur 19 en figuur 23 met elkaar vergelijken, dan zien we dat de grootte van de pseudospectra toenemen naarmate het aantal gridpunten (ofwel de grootte van de matrix) toeneemt. Het is te verwachten dat er vrij snel geen stabiele tijdstap bestaat voor Modified Euler als we het aantal gridpunten nog verder vergroten. Als we een stabiele oplossing willen hebben met een fijner grid, dan moeten we naar hogere orde Runge-Kutta methoden gebruiken. In deze subparagraaf zal de methode van *Runge-Kutta 4* worden gebruikt. Deze zullen in het vervolg afkorten als RK4.
We vergroten het aantal gridpunten naar 59, waardoor $h = 60$. We willen niet dat de diffusieconstante te groot wordt, dus we kiezen $\epsilon = 0.04$ zodat $\nu = 4.8$. Dan krijgen we de het volgende systeem van gewone differentiaalvergelijkingen.

$$\begin{bmatrix} u_1' \\ u_2' \\ \vdots \\ u_{58}' \\ u_{59}' \end{bmatrix} = \begin{bmatrix} -288 & 0 & 0 & \cdots & 0 \\ 288 & -288 & 0 & \cdots & 0 \\ 0 & 288 & -288 & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 288 & -288 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{58} \\ u_{59} \end{bmatrix}$$

We kiezen een tijdstap $\Delta t = 1/206 \approx 0.0049$ (bijna anderhalf keer zo groot als in het vorige voorbeeld) zodat alle 59 eigenwaarden ongeveer gelijk zijn aan -1.398. De tijdstap is gekozen zodanig dat $||Q(\mathbf{A}\Delta t)^n||$ het snelst kleiner is dan 10^{-2} als $\Delta t \in \{\frac{1}{n} : n \in \mathbb{N}\}$ en dat de grafiek niet boven de 2 uitkomt.

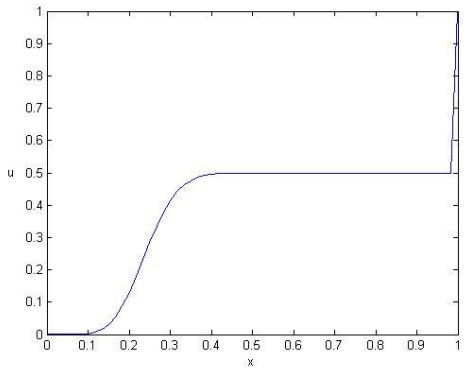


Figuur 27: Ω bij RK4 en pseudospectra van $\mathbf{A}\Delta t$.

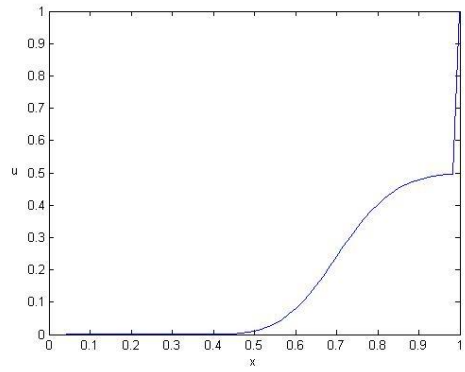


Figuur 28: $||Q(\mathbf{A}\Delta t)^n||$ voor $n = 0, \ldots, 60$.

Bij deze matrix hebben we $\kappa(\mathbf{V}) \approx 75.7$, dus de pseudospectra zullen zich nog verder uitstrekken dan bij het vorige voorbeeld. In figuur 27 staan de randen van $\sigma_\varepsilon(\mathbf{A}\Delta t)$ getekend voor $\varepsilon = 10^{-4}, \ldots, 10^{0}$. Bij $\varepsilon = 10^{-1}$ overschrijdt deze grens de rand van Ω, dus de bult is van de orde 10^1. Rekening houdend met $\sup_{z \in \sigma_{10^{-5}}(\mathbf{A}\Delta t)} \mathrm{dist}(z, \Omega)$, komen we op een maximum dat niet veel groter is dan 1. Dit zien we duidelijk in figuur 28, waarin de grafiek van $||Q(\mathbf{A}\Delta t)^n||$ is getekend. De rand van het pseudospectrum met $\varepsilon = 10^0 = 1$ staat getekend in de figuur om te laten zien hoe snel de pseudospectra zich uitbreiden.
Net zoals in de vorige voorbeelden geven we ook twee numerieke oplossingen. Deze staan in figuren 29 en 30 gegeven na respectievelijk 2060 en 6180 tijdstappen, dus voor $t = 10$ en $t = 30$. In figuur 28 is te zien dat de afbreekfout na 60 tijdstappen al verwaarloosbaar is, dus de numerieke oplossingen hieronder zijn betrouwbare resultaten. In dit voorbeeld duurt het veel langer voordat de functiewaarden op alle gridpunten convergeert naar 0 dan in de twee vorige voorbeelden. Dit komt doordat de grootte van de stromingssnelheid ϵ veel kleiner is gekozen. Ongetwijfeld zal de verandering van de diffusieconstante ν ook effect hebben op het verloop van de numerieke oplossing.
Dankzij het fijnere grid, maar vooral door de grotere nauwkeurigheid van RK4 zien deze twee grafieken er veel gladder uit dan in de twee vorige voorbeelden.



Figuur 29: De numerieke oplossing met RK4 van (4.4) op $t = 10$.



Figuur 30: De numerieke oplossing met RK4 van (4.4) op $t = 30$.

5 Conclusie

In dit verslag hebben we eerst kennis gemaakt met het ε-pseudospectrum en hebben verschillende eigenschappen bekeken. In hoofdstuk 3 hebben we de relatie gelegd tussen pseudospectra en de matrixexponent $||e^{t\mathbf{A}}||$. Dit is niet zozeer een toevoeging geweest voor de theorie in hoofdstuk 4, maar bevestigt wel dat pseudospectra eigenschappen hebben die we op het eerste gezicht niet verwachten. Dat is misschien wel de mooiste ontdekking geweest bij het maken van dit verslag; de intuïtie over wiskundige eigenschappen die met pseudospectra te maken hebben is zo slecht, dat eigenlijk alle theorema's in §3.2 wel verrassend zijn.
Het meest interessante van dit verslag is toch wel het Kreiss matrix theorema dat uitgebreid wordt behandeld in hoofdstuk 4. Veel wiskundigen die de stabiliteit van numerieke oplossingen van stelsels gewone differentiaalvergelijkingen bestuderen, bekijken alleen het gedrag als $t \to \infty$. Numerieke instabiliteit zijn juist eigenschappen van de transient, ofwel voor eindige t. Dit theorema laat dus zien dat er moet worden uitgekeken met de grootte van de tijdstap als het conditiegetal van de matrix groot is. De pseudospectra kunnen dan namelijk erg ver uitdeinen en daardoor hebben we snel een instabiel probleem.
Maar hoe belangrijk is het Kreiss matrix theorema in de wiskundige praktijk? Komt het vaak voor dat de Jacobiaan van een stelsel gewone differentiaalvergelijkingen zodanig ver van normaal is, dat we verschil moeten maken tussen conclusies die we trekken uit spectra en pseudospectra? Een duidelijk antwoord hierop heb ik niet gevonden en waarschijnlijk zal de wetenschap nog verder moeten groeien om een duidelijk antwoord hierop te krijgen. Naar mijn mening zal dit verschil in de overgrote meerderheid van de gevallen niet hoeven worden gemaakt en in een minderheid van de gevallen zal dit verschil wel belangrijk zijn. Niettemin blijft het een boeiend onderwerp en omdat het principe van pseudospectra nog vrij jong is (rond de 50 jaar), zullen er in de toekomst ongetwijfeld nog geweldige uitvindingen worden gedaan die gerelateerd zijn aan dit mooie onderwerp.

6 Bronvermelding

- [1] Trefethen, L.N. & Embree, M. (2005): Spectra and pseudospectra, the behavior of nonnormal matrices and operators.
- [2] Reddy, S.C. & Trefethen, L.N. (1992): Stability of the method of lines in Numer. Math. 62, 235-267 (1992).
- [3] Griffiths, D.F. & Watson, G.A. (1991): eds., Numerical Analysis 1991, 234-266.
- [4] Higham, D.J & Trefethen, L.N. (1993): Stiffness of ODE's.