



Technische Universiteit Delft
Faculteit Elektrotechniek, Wiskunde en Informatica
Delft Institute of Applied Mathematics

Het ontbinden van commutatormatrices
(Engelse titel: The commutator theorem)

Verslag ten behoeve van het
Delft Institute of Applied Mathematics
als onderdeel ter verkrijging

van de graad van

BACHELOR OF SCIENCE
in
TECHNISCHE WISKUNDE

door

TJEERD VAN AALST

Delft, Nederland
Juli 2014



BSc verslag TECHNISCHE WISKUNDE

"Het ontbinden van commutatormatrices"
(Engelse titel: The commutator theorem)

TJEERD VAN AALST

Technische Universiteit Delft

Begeleider

dr. ir. M. C. Veraar

Overige commissieleden

dr. B. van den Dries
dr. B. J. Meulenbroek
dr. J. Vermeer
F. Pei Msc.

Juni 2014,
Delft

Samenvatting

Sinds 1936 is het bekend dat een $m \times m$ complexwaardige matrix A geschreven kan worden als de commutator $A = [B, C] = BC - CB$ dan en slechts dan als $\text{spoor}(A) = 0$. In dit verslag wordt een bewijs van deze stelling gegeven en op basis daarvan een Matlab algoritme geconstrueerd dat voor gegeven A matrices B en C berekent zodat $A = BC - CB$. Tevens wordt onderzocht wat er voor algemene A over de matrices B en C en in het bijzonder hun normen $\|B\|$ en $\|C\|$ te zeggen valt. Er wordt bewezen dat er B en C bestaan waarvoor $\|B\| \|C\| \leq \frac{1}{2} (m-1)^2 \|A\|$, waarbij B bovendien een normale matrix is. Daarnaast wordt een bewijs uit een artikel van Johnson, Ozawa en Schechtman[6] uiteengezet waarin wordt aangetoond dat voor iedere $\varepsilon > 0$ er een K_ε bestaat zodat iedere commutator A te ontbinden is in B en C waarvoor $\|B\| \|C\| \leq K_\varepsilon m^\varepsilon \|A\|$ en B normaal is.

Voorwoord

Beste lezer,

Voor u ligt het eindverslag van mijn bacheloreindproject van de bachelor Technische Wiskunde aan de TU Delft. Dit verslag is geschreven voor lezers met een wiskundige achtergrond en niveau vergelijkbaar met een bachelor in de wiskunde. Om die reden wil ik graag dit voorwoord gebruiken om toch een poging te doen om de inhoud van dit project ook voor geïnteresseerde lezers zonder wiskundige achtergrond enigszins uit te leggen.

In de tak van de wiskunde die lineaire algebra wordt genoemd spelen zogenaamde matrices een grote rol. Een matrix is een wiskundig object dat in feite niets anders is dan een tabel met getallen (zoals bijvoorbeeld in voorbeeld 2.20 verderop in dit verslag). Wiskundigen hebben allerlei manieren bedacht om met deze matrices te spelen; zo kun je bijvoorbeeld twee matrices optellen of van elkaar aftrekken en ook kun je ze vermenigvuldigen. Matrices bij elkaar optellen of aftrekken gaat zoals je intuïtief zou verwachten: de som van twee matrices is de tabel (matrix) waarin de getallen de som (of het verschil) zijn van de twee getallen die op dezelfde plek staan in de matrices die je bij elkaar optelt. Het vermenigvuldigen van twee matrices is echter ingewikkelder. De "natuurlijke" vermenigvuldiging van twee matrices is zelfs zo opgezet dat het kan gebeuren dat het product BC is anders is dan CB . Als dit zo is dan zeggen wiskundigen dat de matrices B en C niet met elkaar commuteren en het verschil $BC - CB$ noemen ze de commutator van B en C . Dit zijn de commutatormatrices waar de titel van het verslag naar verwijst.

De vraag die in dit project centraal staat is nu: stel ik geef je een willekeurige matrix A , kun je mij dan matrices B en C geven zodat $A = BC - CB$? Daarbij kun je je grofweg drie dingen afvragen: (1) wat zijn precies de voorwaarden waar A aan moet voldoen zodat dit kan, (2) hoe kun je zulke B en C berekenen en (3) bestaan er altijd B en C die op een bepaalde manier "klein" zijn? Deze vragen komen allemaal aan bod in het verslag; in het bijzonder bestaat een groot deel van het project uit het bestuderen en doorgronden van een artikel waarin vraag (3) wordt beantwoord.

Tijdens dit project heb ik een hoop geleerd over hoe wiskundig onderzoek er in de praktijk aan toe gaat. Soms kan dit heel spannend en interessant zijn als een lang openstaand probleem plots wordt opgelost. Aan de andere kant is echt begrijpen en doorgronden van een wiskundig artikel helemaal niet zo eenvoudig, zeker voor een lezer van buiten het vakgebied. Een bewering die in een artikel in één zin gemaakt wordt kan soms heel wat uitleg gebruiken voordat een niet-ingewijde lezer de afleiding begrijpt. Het bestuderen van het artikel dat centraal staat in dit project heeft me dan ook een hoop geleerd; zowel inhoudelijke, wiskundige truukjes als kennis over wat er zoal komt kijken bij het structureren en vormgeven van een ingewikkelde wiskundige afleiding.

Ik wil graag mijn begeleider Mark Veraar bedanken, die mij goed op weg heeft geholpen en waarmee ik samen vele hobbels heb genomen. Regelmatig waren er subtiele struikelblokken waarop een bewijsje vastliep of zaken waar ik net iets te snel over heen stapte, maar altijd konden we samen wel een oplossing

vinden.

Tenslotte wens ik u veel plezier met het lezen van dit verslag.

Tjeerd van Aalst
juli 2014

Inhoudsopgave

1	Inleiding	5
1.1	Achtergrond en geschiedenis	5
1.2	Doel en vraagstellingen	6
1.3	Opbouw verslag	7
2	Over commutators, normen en de commutatorstelling	8
2.1	Het spoor van een matrix	8
2.2	Normale matrices	11
2.3	Commutators en de commutatorstelling	11
2.4	De norm van een matrix	14
2.4.1	De norm van een vector	14
2.4.2	Definitie matrixnorm	15
2.4.3	Eigenschappen matrixnorm	17
2.5	De commutatorstelling gekwantificeerd	21
2.6	Verband met het Kadison-Singer vermoeden	28
3	De commutatorstelling van Johnson, Ozawa en Schechtman	31
3.1	Wat opmerkingen vooraf	32
3.2	De commutatorstelling $\mathcal{O}\left((\log_4 m)^2 \sqrt{m}\right)$	34
3.3	2×2 blokmatrices	43
3.4	Een stelling van Bourgain en Tzafriri	50
3.5	De commutatorstelling $\mathcal{O}(m^\varepsilon)$	54
4	Een algoritme om commutators te ontbinden	60
4.1	Ontbindingsprocedure	60
4.2	Uitholprocedure	61
4.3	Gedrag van het programma	63
5	Conclusie en discussie	70
A	Matlab implementatie van het algoritme	73
A.1	Ontbindingsprocedure	73
A.2	Uitholprocedure	76
A.3	Willekeurige matrices	81

B De commutatorstelling voor reële matrices	82
Bibliografie	83

Hoofdstuk 1

Inleiding

1.1 Achtergrond en geschiedenis

Een commutatormatrix A is een matrix waarvoor matrices B en C bestaan zodat $A = BC - CB$. Het is eenvoudig om na te gaan dat in dat geval $\text{spoor}(A) = 0$. In 1937 bewees Shoda[11] dat het omgekeerde ook waar is (de zogenaamde commutatorstelling): iedere spoorloze matrix is te schrijven als een commutator, i.e. er bestaan B en C zodat $A = BC - CB$. Een logische vraag is nu: wat kunnen we zeggen over de matrices B en C ? Zijn ze diagonaal, normaal, unitair of van een andere interessante vorm? Zijn ze überhaupt uniek? Zo niet, bestaan er dan wel altijd B en C die van een bepaalde vorm zijn? En in het bijzonder: kunnen we iets zeggen over de grootte van (de getallen in) B en C , of specifieker: over de normen $\|B\|$ en $\|C\|$ van B en C ?

Vragen genoeg dus, maar gelukkig zijn er ook wat antwoorden. Een bijzonder antwoord is de stelling uit [6] die in dit project centraal staat. De auteurs Johnson, Ozawa en Schechtman bewijzen in dit artikel dat er voor iedere $\varepsilon > 0$ een constante K_ε bestaat zodat een $m \times m$ commutatormatrix A te ontbinden is in matrices B en C (dus $A = BC - CB$) waarvoor $\|B\| \|C\| \leq K_\varepsilon m^\varepsilon \|A\|$ en B ook nog eens een zogenaamde normale matrix is. Dat klinkt als een hele mond vol, dus laten we hem stapje voor stapje doornemen om te zien waarom de stelling deze vorm heeft.

In het algemeen blijkt dat de norm van B of C apart willekeurig groot of klein gemaakt kan worden. Om toch iets zinnigs te kunnen zeggen moeten we daarom kijken naar één of andere combinatie van de twee normen, bijvoorbeeld het product $\|B\| \|C\|$. Met een eenvoudige driehoeksongelijkheid kun je aantonen dat $\|B\| \|C\| \geq \frac{1}{2} \|A\|$. Nu kun je je afvragen: bestaan er altijd matrices B en C zodat $\|B\| \|C\| = \frac{1}{2} \|A\|$? We zullen zien dat dit voor 2×2 matrices inderdaad zo is, maar voor grotere matrices is dit niet zo eenvoudig te zeggen. Wel kunnen we ons afvragen of er dan tenminste een constante K is (die mogelijk nog van de matrixafmeting m afhangt) zodat er wel altijd matrices B en C bestaan waarvoor $\|B\| \|C\| \leq K \|A\|$. Dit is precies de vraag die op het internetforum

MathOverflow werd geplaatst (<http://mathoverflow.net/questions/27345>) en de aanleiding vormde voor het artikel [6].

Nu kun je je afvragen: wat heb je hieraan? Natuurlijk is de vraag zelf uit pure wiskundige nieuwsgierigheid al interessant, maar er zijn ook andere redenen om dit te onderzoeken. Als zo'n constante K bestaat die niet van m afhangt heeft dat allerlei interessante gevolgen. Een zo'n gevolg wordt door de auteurs genoemd; het zou een eenvoudige manier geven om de zogenaamde Wright factoren te classificeren. Een veel interessante gevolg is dat het ook zou betekenen dat het zogenaamde Kadison-Singer vermoeden waar is.

Het Kadison-Singer vermoeden is een beroemd open probleem in de wiskunde dat in 1959 voor het eerst werd geformuleerd door R. Kadison en I. Singer in [7]. In de loop der jaren is er bewezen dat vele problemen equivalent zijn met het KS-vermoeden, waaronder het zogenaamde paving vermoeden van J. Anderson [1] wat te maken heeft met matrices. De auteurs van [6] tonen in hun artikel aan dat een grens die niet van m afhangt zou betekenen dat het paving vermoeden en daarmee ook het KS-vermoeden waar zijn.

Nu wil het zo zijn dat er onlangs (2013) is aangetoond dat het KS-vermoeden waar is [9]. De grote vraag is dan natuurlijk: betekent dit ook dat er een K die niet van m afhangt bestaat? Dit is vooralsnog onbekend, al geven de auteurs van [6] in het artikel wel aan dat hun resultaat waarschijnlijk verbeterd kan worden als het paving vermoeden (equivalent met KS) waar is.

1.2 Doel en vraagstellingen

Het voornaamste doel van dit project is om het bewijs van [6] te begrijpen. Tijdens het bestuderen van dit artikel kwamen deelproblemen naar boven die met dit onderwerp te maken hebben en soms nodig zijn om het bewijs te begrijpen. De vragen waarop ik in dit verslag wil beantwoorden zijn:

- Hoe bewijs je dat de commutator matrices precies de spoorloze matrices zijn (de zogenaamde commutatorstelling)?
- Hoe kun voor een gegeven commutator matrix A matrices B en C construeren zodat $A = BC - CB$? Kunnen we een algoritme voor zo'n commutatorontbinding geven zodat een computer B en C kan berekenen?
- Als $A = BC - CB$, wat kun je dan in het algemeen zeggen over de normen van B en C ?
- Kun je op eenvoudige wijze aantonen dat er altijd een commutatorontbinding bestaat waarbij de normen van B en C op een zekere manier zijn af te schatten met de norm van A ?
- Hoe bewijzen Johnson, Ozawa en Schechtman in [6] dat er zo'n afschatting is van $\mathcal{O}(m^\varepsilon)$ voor willekeurig kleine $\varepsilon > 0$?
- Wat is het gevolg van de recente ontwikkelingen (bevestiging van het KS-vermoeden) voor dit bewijs? Hoe kan de afschatting - zoals de auteurs

beweren - verbeterd worden tot $\mathcal{O}\left((\log m)^k\right)$ als het paving vermoeden waar is?

1.3 Opbouw verslag

Ik heb geprobeerd om een natuurlijke opbouw in het verslag aan te brengen waarbij stellingen niet uit de lucht komen vallen maar eerst gemotiveerd worden. Om die reden kan het verslag het beste van voor naar achter worden gelezen. Waar mogelijk heb ik zoveel mogelijk geprobeerd te achterhalen waarom de auteurs van [6] bepaalde stappen nemen in hun bewijs.

In hoofdstuk 2 wordt de basis gelegd die nodig is om [6] te begrijpen. We bestuderen begrippen als het spoor en de norm van een matrix en bekijken enkele stellingen die ons meer inzicht geven in deze eigenschappen. Daarnaast bewijzen we de commutatorstelling en bekijken we hoe we aan de hand van dit bewijs de stelling kunnen kwantificeren en wat voor moeilijkheden we hierbij tegenkomen. In hoofdstuk 3 bekijken het bewijs van de hoofdstelling uit het artikel [6]. In hoofdstuk 4 werken tenslotte we het bewijs van de commutatorstelling uit hoofdstuk 2 verder uit, wat uiteindelijk resulteert in een computerprogramma dat commutators kan ontbinden. Een matlab implementatie van dit programma is opgenomen in de bijlage. Om inzicht te krijgen in het gedrag van het programma testen we het met een groot aantal willekeurig gegenereerde matrices.

Ik ben er van uit gegaan dat de lezer bekend is lineaire algebra en begrippen als gelijksoortigheid, unitaire matrices, blokmatrices, eigenwaarden en diagonaliseren kent (als naslagwerk voor lineaire algebra is [2] gebruikt). Ook enkele andere begrippen uit de wiskunde bachelor zullen voorbij komen, zoals convexiteit van een verzameling, het ordebegrip $\mathcal{O}$ en de stelling van Cauchy voor contourintegralen in het complexe vlak. We zullen over het algemeen werken met complexe getallen, waarbij de complex geconjugeerde van een x zal worden aangeduid met $\bar{x}$ en de hermitisch getransponeerde van een matrix (of vector) Y met Y^* . In het gehele verslag zullen we ons beperken tot vierkante matrices.

Hoofdstuk 2

Over commutators, normen en de commutatorstelling

In dit hoofdstuk leggen we de basis die we nodig zullen hebben om het bewijs van Johnson, Ozawa en Schechtman te begrijpen. Begrippen als het spoor en de norm van een matrix zullen worden uitgelegd en enkele nuttige eigenschappen die we later zullen gebruiken worden afgeleid. Met deze kennis zullen we de commutatorstelling bewijzen en eenvoudige afschatting afleiden waarmee de commutatorstelling gekwantificeerd kan worden.

2.1 Het spoor van een matrix

Definitie 2.1. Het *spoor* van een $m \times m$ matrix A is de som van de diagonaal-elementen van A :

$$\text{spoor}(A) = \sum_{i=1}^m a_{ii} .$$

Een matrix A heet *spoorloos* als $\text{spoor}(A) = 0$.

Definitie 2.2. Een vierkante matrix A heet *hol* als alle elementen op de diagonaal van A gelijk aan nul zijn.

Merk op dat holle matrices altijd spoorloos zijn, maar spoorloze matrices niet altijd hol - zoals het volgende voorbeeld illustreert:

Voorbeeld 2.3. Voor de matrices

$$X = \begin{pmatrix} 1 & 2 & 0 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \quad Y = \begin{pmatrix} 5 & 4 & 3 \\ 0 & -7 & -2 \\ -4 & 3 & 2 \end{pmatrix} \quad Z = \begin{pmatrix} 0 & 3 & 6 \\ 2 & 0 & 2 \\ 7 & 8 & 0 \end{pmatrix}$$

geldt

$$\begin{aligned}\text{spoor}(X) &= 1 + 5 + 9 = 15 \\ \text{spoor}(Y) &= 5 - 7 + 2 = 0 \\ \text{spoor}(Z) &= 0 + 0 + 0 = 0\end{aligned}$$

Zowel Y als Z zijn spoorloze matrices; Z is bovendien een holle matrix.

Uit de definitie van het spoor kunnen we gelijk een drietal eigenschappen afleiden die het berekenen van het spoor van een ingewikkeldere uitdrukking makkelijker maken:

Stelling 2.4. *Het spoor is een lineaire afbeelding, i.e. voor matrices A en B en constanten α en β geldt*

$$\text{spoor}(\alpha A + \beta B) = \alpha \text{spoor}(A) + \beta \text{spoor}(B).$$

Bewijs. Dit volgt direct uit de definitie en de lineariteit van matrixoptelling. $\square$

Stelling 2.5. *Voor matrices A en B geldt $\text{spoor}(AB) = \text{spoor}(BA)$.*

Bewijs. Zorgvuldig uitschrijven van het matrixproduct geeft een som die symmetrisch is in A en B :

$$\text{spoor}(AB) = \sum_{i,j} a_{ij}b_{ji} = \sum_{i,j} b_{ij}a_{ji} = \text{spoor}(BA).$$

$\square$

Stelling 2.6. *Het spoor van twee gelijksoortige matrices is gelijk.*

Bewijs. Voor een matrix A en een inverteerbare matrix P geldt

$$\text{spoor}(P^{-1}AP) = \text{spoor}((P^{-1}A)P) = \text{spoor}(P(P^{-1}A)) = \text{spoor}(A).$$

$\square$

Er is een bijzonder verband tussen het spoor en de eigenwaarden van een matrix:

Stelling 2.7. *Het spoor van een matrix gelijk aan de som van de eigenwaarden, d.w.z. voor een $m \times m$ matrix A geldt*

$$\text{spoor}(A) = \sum_{i=1}^m \lambda_i$$

waarbij $\lambda_1, \lambda_2, \dots, \lambda_m$ de eigenwaarden (inclusief algebraïsche multipliciteiten) van A zijn.

Bewijs. Indien A diagonaliseerbaar is volgt de bewering direct door toepassen van stelling 2.6, er geldt immers $\text{spoor}(A) = \text{spoor}(P^{-1}DP) = \text{spoor}(D) = \sum_i \lambda_i$. Ook als A niet diagonaliseerbaar is geldt de stelling; in plaats van de diagonaalmatrix D kan dan de Jordan normaalvorm J genomen worden die ook alle eigenwaarden van A op de diagonaal heeft staan. Dit vereist wel het (sterke) resultaat dat elke matrix een Jordan normaalvorm heeft. Gelukkig kunnen we ook zonder kennis van de Jordan normaalvorm deze stelling bewijzen, zij het wat minder eenvoudig:

De eigenwaarden van A zijn de nulpunten van het karakteristieke polynoom, gegeven door

$$p(\lambda) = \det(A - \lambda I) = c \prod_{i=1}^m (\lambda - \lambda_i) = c \left(\lambda^m + \left[-\sum_{i=1}^m \lambda_i \right] \lambda^{m-1} + \dots \right).$$

De laatste uitdrukking kun je eenvoudig nagaan met inductie naar m ; in het algemeen staat dit bekend als een van Viëta formules, die de coëfficiënten van een polynoom uitdrukken in sommen van de wortels van een polynoom. We kunnen de stelling bewijzen door een andere uitdrukking voor $\det(A - \lambda I)$ te bepalen en de coëfficiënten van λ^{m-1} te vergelijken, maar dan moeten we eerst weten wat c is. De constante c kunnen we bepalen door gebruik te maken van de uitdrukking voor de determinant van een matrix D als som over producten van d_{ij} 's die de corresponderen met een permutatie $\sigma \in S_m$ van m elementen:

$$\det(D) = \sum_{\sigma \in S_m} \text{sgn}(\sigma) \prod_{i=1}^m d_{i\sigma(i)}$$

($\text{sgn}(\sigma) = \pm 1$ is het teken van de permutatie). Omdat elke permutatie precies één element uit elke rij en kolom selecteert is term λ^m in $\det(A - \lambda I)$ volledig afkomstig van de identiteitspermutatie (met teken 1) die alle diagonaalelementen selecteert:

$$\prod_{i=1}^m (a_{ii} - \lambda) = (-1)^m \lambda^m + \dots$$

en dus $c = (-1)^m$. Maar ook de term λ^{m-1} moet volledig afkomstig zijn van de identiteitspermutatie, want als een permutatie $m - 1$ diagonaalelementen selecteert dan moet hij wel alle diagonaalelementen selecteren (selecties bevatten één element uit elke rij of kolom). Dus de term λ^{m-1} moet volledig afkomstig zijn van

$$\prod_{i=1}^m (a_{ii} - \lambda) = (-1)^m \prod_{i=1}^m (\lambda - a_{ii}) = (-1)^m \left(\lambda^m + \left[-\sum_{i=1}^m a_{ii} \right] \lambda^{m-1} + \dots \right)$$

waarbij we weer de formule van Viëta gebruiken. Daarmee volgt:

$$\text{spoor}(A) = \sum_{i=1}^m a_{ii} = \sum_{i=1}^m \lambda_i.$$

□

2.2 Normale matrices

Een belangrijke klasse matrices zijn de zogenaamde normale matrices, die zoals de naam al een beetje doet vermoeden een aantal mooie eigenschappen bezitten. Wij zullen ons hier beperken tot de definitie en een belangrijke eigenschap. Voor een bewijs van de stelling, zie bijvoorbeeld [2, stelling 9.21].

Definitie 2.8. Een matrix A heet *normaal* als $AA^* = A^*A$.

Stelling 2.9. Een matrix A is normaal dan en slechts dan als deze unitair diagonaliseerbaar is.

2.3 Commutators en de commutatorstelling

In deze paragraaf zullen we de commutatorstelling (de commutatormatrices zijn precies de spoorloze matrices) bewijzen. Het moeilijke werk zal zitten in het bewijzen van een lemma waarmee wordt aangetoond dat we ons in het bewijs van de commutatorstelling kunnen beperken tot holle matrices. Voor dit lemma zullen we een stelling uit de lineaire algebra nodig hebben die ons vertelt dat het zogenaamde numerieke bereik van een matrix convex is. Het is ook mogelijk om eerst de commutatorstelling voor 2×2 matrices te bewijzen en daarmee de convexiteit van het numeriek bereik aan te tonen [13]. Een alternatief (niet-constructief) bewijs voor de commutatorstelling wordt gegeven in [8].

We beginnen met een definitie:

Definitie 2.10. De *commutator* $[B, C]$ van twee vierkante matrices B en C is de matrix $BC - CB$. Als $[B, C] = 0$ dan $BC = CB$ en zeggen we ook wel dat B en C commuteren. Een matrix vierkante matrix A heet een commutator als er matrices B en C bestaan zodat $A = [B, C] = BC - CB$, en dan zeggen we ook wel dat de commutator A te *ontbinden* is in B en C .

Wat volgt zijn drie lemma's die het bewijzen van de commutatorstelling makkelijk gaan maken.

Lemma 2.11. Voor elke spoorloze matrix $A \in \mathbb{C}^{m \times m}$ bestaat er een eenheidsvector $x \in \mathbb{C}^m$ zodat $x^*Ax = 0$.

Bewijs. Voor een matrix A is de som van de eigenwaarden gelijk aan het spoor van A (stelling 2.7). Als A nu spoorloos is dan is de som van de eigenwaarden gelijk aan nul, en daarmee is ook het gemiddelde van de eigenwaarden nul. Volgens de stelling van Toeplitz-Hausdorff (zie b.v. [2, stelling 22.27]) is het numeriek bereik $W(A) = \{x^*Ax : |x| = 1\}$ van een matrix A een convexe verzameling. Het is eenvoudig om na te gaan dat de eigenwaarden van A in $W(A)$ zitten, neem voor x maar een bijbehorende eigenvector. Omdat $W(A)$ convex is ligt dan ook het gemiddelde van de eigenwaarden (nul) in $W(A)$, dus er bestaat een eenheidsvector x zodat $x^*Ax = 0$. $\square$

Lemma 2.12. Elke spoorloze matrix $A \in \mathbb{C}^{m \times m}$ is unitair equivalent met een holle matrix.

Bewijs. We bewijzen met inductie naar m . Voor $m = 1$ is de bewering triviaal; de 'matrix' 0 is immers de enige spoorloze 1×1 matrix en ook de enige holle 1×1 matrix. Stel nu dat de bewering geldt voor een zekere $m \geq 1$ en zij A een spoorloze $(m+1) \times (m+1)$ matrix. Wegens lemma 2.11 bestaat er een eenheidsvector $x \in \mathbb{C}^{m+1}$ zodat $x^*Ax = 0$. Neem nu een matrix $X \in \mathbb{C}^{(m+1) \times (m+1)}$ waarvan de eerste kolom gelijk is aan x en de overige kolommen samen met x een orthonormale basis vormen (dan is X unitair). Als we A nu unitair transformeren met X wordt het eerste diagonaalelement van A gelijk aan nul (ga na) en krijgen we een matrix van de vorm

$$X^{-1}AX = \begin{pmatrix} 0 & p \\ q & A' \end{pmatrix}$$

waarbij $p \in \mathbb{C}^{1 \times m}$ een rijvector is, $q \in \mathbb{C}^{m \times 1}$ een kolomvector en $A' \in \mathbb{C}^{m \times m}$ een $m \times m$ matrix. Omdat het spoor van een matrix behouden is onder unitaire transformaties (stelling 2.6) volgt

$$\text{spoor}(A') = \text{spoor}(X^{-1}AX) = \text{spoor}(A) = 0$$

en dus bestaat er wegens de inductiehypothese een unitaire matrix $U' \in \mathbb{C}^{m \times m}$ zodat $(U')^{-1}A'U'$ een holle matrix is. Neem nu

$$U = X \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix}$$

dan is U unitair (het product van twee unitaire matrices is unitair) en

$$\begin{aligned} U^{-1}AU &= \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix}^{-1} X^{-1}AX \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix}^{-1} \begin{pmatrix} 0 & p \\ q & A' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & (U')^{-1} \end{pmatrix} \begin{pmatrix} 0 & p \\ q & A' \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & U' \end{pmatrix} \\ &= \begin{pmatrix} 0 & pU' \\ (U')^{-1}q & (U')^{-1}A'U' \end{pmatrix} \end{aligned}$$

een holle matrix. □

Lemma 2.13. *Commutatierelaties zijn behouden onder unitaire transformaties.*

Bewijs. Stel U is een unitaire matrix en $A = U^{-1}\tilde{A}U$, $B = U^{-1}\tilde{B}U$ en $C =$

$U^{-1}\tilde{C}U$ waarvoor $\tilde{A} = \tilde{B}\tilde{C} - \tilde{C}\tilde{B}$. Dan geldt

$$\begin{aligned}
A &= U^{-1}\tilde{A}U \\
&= U^{-1}(\tilde{B}\tilde{C} - \tilde{C}\tilde{B})U \\
&= U^{-1}\tilde{B}\tilde{C}U - U^{-1}\tilde{C}\tilde{B}U \\
&= U^{-1}\tilde{B}(UU^{-1})\tilde{C}U - U^{-1}\tilde{C}(UU^{-1})\tilde{B}U \\
&= (U^{-1}\tilde{B}U)(U^{-1}\tilde{C}U) - (U^{-1}\tilde{C}U)(U^{-1}\tilde{B}U) \\
&= BC - CB
\end{aligned}$$

□

We hebben nu alle ingrediënten voor de commutatorstelling:

Stelling 2.14 (De commutatorstelling). *Een matrix A is een commutator dan en slechts dan als A spoorloos is.*

Bewijs. Stel A is een commutator, dan bestaan er matrices B en C zodat $A = BC - CB$. Met de eigenschappen van het spoor volgt:

$$\begin{aligned}
\text{spoor}(A) &= \text{spoor}(BC - CB) \\
&= \text{spoor}(BC) - \text{spoor}(CB) && \text{(stelling 2.4)} \\
&= \text{spoor}(BC) - \text{spoor}(BC) && \text{(stelling 2.5)} \\
&= 0
\end{aligned}$$

en dus is A spoorloos.

Stel A is spoorloze matrix, dan bestaat er wegens lemma 2.12 een unitaire matrix U zodat $\tilde{A} = UAU^{-1}$ een holle matrix is. Neem nu een diagonaalmatrix $\tilde{B}$ met op de diagonaal louter verschillende elementen en een matrix $\tilde{C}$ met elementen gegeven door

$$\tilde{c}_{ij} = \begin{cases} \frac{\tilde{a}_{ij}}{\tilde{b}_{ii} - \tilde{b}_{jj}} & i \neq j \\ 0 & i = j \end{cases}$$

($\tilde{a}_{ij}$ en $\tilde{b}_{ij}$ zijn de elementen van $\tilde{B}$ en $\tilde{C}$). Dan geldt $\tilde{A} = \tilde{B}\tilde{C} - \tilde{C}\tilde{B}$, immers

$$(\tilde{B}\tilde{C} - \tilde{C}\tilde{B})_{ij} = (\tilde{B}\tilde{C})_{ij} - (\tilde{C}\tilde{B})_{ij} = \tilde{b}_{ii}\tilde{c}_{ij} - \tilde{c}_{ij}\tilde{b}_{jj} = \tilde{a}_{ij}$$

en als we B en C terug transformeren met U vinden we $B = U^{-1}\tilde{B}U$ en $C = U^{-1}\tilde{C}U$ waarvoor wegens lemma 2.13 geldt $A = BC - CB$ en dus is A een commutator. □

De constructie die in bovenstaand bewijs gebruik wordt (met een holle $\tilde{A}$ en een diagonaalmatrix $\tilde{B}$) zullen we vaker terugzien. Merk op dat de B bij deze constructie altijd een normale matrix is, hij is immers unitair equivalent met de diagonaalmatrix $\tilde{B}$ (stelling 2.9).

2.4 De norm van een matrix

Om de commutatorstelling te kwantificeren (i.e. iets zeggen over de normen van B en C) moeten we natuurlijk eerst weten wat de norm van een matrix eigenlijk is en hoe je die kunt berekenen. Voordat we dit doen is het een goed idee om onze kennis over vectornormen op te frissen.

2.4.1 De norm van een vector

Definitie 2.15. Een *norm* op een vector in $\mathbb{C}^m$ is een functie $|\cdot| : \mathbb{C}^m \rightarrow \mathbb{R}$ die voldoet aan de volgende drie eigenschappen:

1. Voor alle $x \in \mathbb{C}^m$ geldt $|x| \geq 0$ en $|x| = 0 \iff x = 0$.
2. Voor alle $x \in \mathbb{C}^m$ en $\alpha \in \mathbb{C}$ geldt $|\alpha x| = |\alpha| |x|$.
3. Voor alle $x, y \in \mathbb{C}^m$ geldt $|x + y| \leq |x| + |y|$ (de driehoeksongelijkheid).

Een grote klasse van normen zijn de p -normen, gegeven door

$$|x|_p = \left(\sum_{i=1}^m x_i^p \right)^{1/p}.$$

De belangrijkste zijn de 1-norm

$$|x|_1 = \sum_{i=1}^m |x_i|$$

(de som van de absolute waarden van de elementen van de vector), de 2-norm

$$|x|_2 = \sqrt{\sum_{i=1}^m |x_i|^2}$$

(ook wel de Euclidische norm) en de ∞ -norm

$$|x|_\infty = \lim_{p \rightarrow \infty} |x|_p = \max_{i=1..m} |x_i|$$

(de max-norm). In het algemeen bedoelen we met de norm van een vector de Euclidische norm. Het subscript 2 wordt om die reden vaak weggelaten, dus $|x| = |x|_2$ (tenzij expliciet anders vermeld).

Voorbeeld 2.16. Voor de vector

$$x = (-5 \quad 2 \quad 0 \quad 4)^T$$

zijn de genoemde normen gelijk aan

$$\begin{aligned} |x|_1 &= 5 + 2 + 4 = 11 \\ |x|_2 &= \sqrt{5^2 + 2^2 + 4^2} = \sqrt{45} = 3\sqrt{5} \approx 6.7 \\ |x|_\infty &= 5 \end{aligned}$$

Merk op dat het nemen van de norm een continue afbeelding $|\cdot| : \mathbb{C}^m \rightarrow \mathbb{R}$ is; voor alle vectoren $a \in \mathbb{C}^m$ en $\varepsilon > 0$ kunnen we immers een vector $\xi \in \mathbb{C}^m$ nemen met $|\xi| < \varepsilon$ zodat

$$||a + \xi| - |a|| \leq |(a + \xi) - a| = |\xi| < \varepsilon$$

hier gebruiken we de omgekeerde driehoeksongelijkheid die eenvoudig uit de driehoeksongelijkheid kan worden afgeleid:

$$\left. \begin{aligned} |x| &= |(x - y) + y| \leq |x - y| + |y| \implies |x| - |y| \leq |x - y| \\ |y| &= |(y - x) + x| \leq |y - x| + |x| \implies |y| - |x| \leq |y - x| \end{aligned} \right\} ||x| - |y|| \leq |x - y|$$

2.4.2 Definitie matrixnorm

Matrixnormen zijn eigenlijk op geheel analoge wijze gedefinieerd:

Definitie 2.17. Een *norm* op een matrix in $\mathbb{C}^{m \times m}$ is een functie $\|\cdot\| : \mathbb{C}^{m \times m} \rightarrow \mathbb{R}$ die voldoet aan de volgende drie eigenschappen:

1. Voor alle $X \in \mathbb{C}^{m \times m}$ geldt $\|X\| \geq 0$ en $\|X\| = 0 \iff X = 0$.
2. Voor alle $X \in \mathbb{C}^{m \times m}$ en $\alpha \in \mathbb{C}$ geldt $\|\alpha X\| = |\alpha| \|X\|$.
3. Voor alle $X, Y \in \mathbb{C}^{m \times m}$ geldt $\|X + Y\| \leq \|X\| + \|Y\|$ (de driehoeksongelijkheid).

Door de analogie met vectornormen zal het je waarschijnlijk niet verbazen dat op eenvoudige wijze een norm kunt construeren door te doen alsof een $m \times m$ matrix een vector van grootte m^2 is. De *elementsgewijze norm* die je dan krijgt is in het Euclidische geval de zogenaamde *Frobenius norm*:

$$\|A\|_F = \sum_{i,j} |a_{ij}|^2$$

Andere normen krijg je bijvoorbeeld door te kijken naar het gedrag van een $m \times m$ matrix A als lineaire operator. Zo krijg je de *geïnduceerde p -norm*:

$$\|A\|_p = \max_{x \in \mathbb{C}^m} \frac{|Ax|_p}{|x|_p} = \max_{x : |x|_p = 1} |Ax|_p$$

De geïnduceerde p -norm geeft als het ware de maximale vergroting van de p -norm van een vector als je de matrix A toepast op deze vector, dus $|Ax|_p \leq \|A\|_p |x|_p$ voor alle vectoren x . Net als bij de vectornormen is de geïnduceerde 2-norm (ook wel de geïnduceerde Euclidische norm) het meest interessant omdat deze overeenkomt met ons normale idee van lengte. Vanaf nu bedoelen we met de norm van een matrix dan ook de geïnduceerde Euclidische norm, dus $\|A\| = \|A\|_2$.

Het maximum in de definitie van de matrixnorm wordt daadwerkelijk aangenomen, want het is een maximum van een continue functie op een compacte

verzameling. De verzameling $\{x : |x|_p = 1\}$ is een compacte (gesloten en begrensde) verzameling, immers $x \mapsto |x|_p$ is een continue afbeelding en dus is het inverse beeld van de gesloten verzameling $\{1\}$ ook gesloten. Bovendien is de matrixvermenigvuldiging $x \mapsto Ax$ een continue afbeelding en dus ook de samenstelling $x \mapsto |Ax|_p$.

Ook het nemen van de matrixnorm is een continue afbeelding $\|\cdot\| : \mathbb{C}^{m \times m} \rightarrow \mathbb{R}$, een feit dat om precies dezelfde wijze als voor vectornormen uit de driehoeksongelijkheid kan worden afgeleid.

Alle matrixnormen die we tot nu toe genoemd hebben zijn zogenaamde *submultiplicatieve normen*, wat betekent dat ze voldoen aan de extra eigenschap:

4. Voor alle $X, Y \in \mathbb{C}^{m \times m}$ geldt $\|XY\| \leq \|X\| \|Y\|$ (submultiplicativiteit).

Van de geïnduceerde norm kan dit makkelijk worden ingezien, immers

$$\|XY\| = \max_{|x|=1} |XYx| \leq \max_{|x|=1} \|X\| |Yx| \leq \|X\| \|Y\|$$

Een voorbeeld van een norm die niet submultiplicatief is is de elementsgewijze ∞ -norm, ook wel de *max-norm* genoemd:

$$\|A\|_{\max} = \max_{i,j} |a_{ij}|$$

(de waarde van het grootste element van A , in absolute zin).

De definitie van de geïnduceerde normen vertelt ons niet hoe we ze makkelijk kunnen berekenen. Gelukkig gelden voor de 1- en ∞ -norm de volgende mooie identiteiten [2, lemma 7.14]:

Stelling 2.18. *De geïnduceerde 1-norm is gelijk aan de maximale absolute kolomsum:*

$$\|A\|_1 = \max_j \sum_i |a_{ij}|$$

De geïnduceerde ∞ -norm is gelijk aan de maximale absolute rijsum:

$$\|A\|_{\infty} = \max_i \sum_j |a_{ij}|$$

Voor de geïnduceerde Euclidische norm ligt het wat ingewikkelder [2, lemma 7.14]:

Stelling 2.19. *De geïnduceerde Euclidische norm van een $m \times m$ matrix A is gelijk aan de grootste singuliere waarde van A :*

$$\|A\| = \|A\|_2 = \max_i \sqrt{\lambda_i(A^*A)} = \sigma_{\max}(A)$$

Voorbeeld 2.20. Voor de matrix

$$A = \begin{pmatrix} 1 & 2 & -3 \\ -4 & 5 & 6 \\ 7 & -8 & 9 \end{pmatrix}$$

zijn de genoemde normen gelijk aan

$$\begin{aligned}
\|A\|_{\max} &= 9 \\
\|A\|_F &= \sqrt{1^2 + 2^2 + \dots + 9^2} = \sqrt{285} \approx 16.9 \\
\|A\|_1 &= \max(1+4+7, 2+5+8, 3+6+9) = 18 \\
\|A\|_{\infty} &= \max(1+2+3, 4+5+6, 7+8+9) = 24 \\
\|A\|_2 &= \max(\sqrt{4.2}, \sqrt{78.8}, \sqrt{202}) \approx 14.2
\end{aligned}$$

Merk op dat van deze normen alleen de Euclidische norm verandert als we de tekens (of algemener: de complexe fase) van de elementen van A veranderen. Dit zorgt ervoor dat de Euclidische norm zeker het interessant is, maar ook lastig is af te schatten aan de hand van de elementen.

2.4.3 Eigenschappen matrixnorm

De volgende eigenschappen van de (Euclidische) matrixnorm zullen straks handig zijn als we de norm van een matrix die we geconstrueerd hebben willen afschatten.

Stelling 2.21. *Voor een matrix A geldt*

$$\|A\| \leq \sqrt{\|A\|_1 \|A\|_{\infty}}$$

Bewijs. Neem v een eigenvector bij de grootste eigenwaarde λ van A^*A , dan

$$\|A^*A\|_1 = \max_{|x|_1=1} |A^*Ax|_1 \geq \left| A^*A \frac{v}{|v|_1} \right|_1 = \left| \lambda \frac{v}{|v|_1} \right|_1 = \lambda$$

en met de singuliere-waarde uitdrukking voor $\|A\|$ (stelling 2.19) volgt:

$$\|A\| = \sqrt{\lambda} \leq \sqrt{\|A^*A\|_1} \leq \sqrt{\|A^*\|_1 \|A\|_1} = \sqrt{\|A\|_{\infty} \|A\|_1}$$

(merk op dat $\|A^*\|_1 = \|A\|_{\infty}$ wegens stelling 2.18, we verwisselen rij- en kolom-sommen). $\square$

Stelling 2.22. *Voor een $m \times m$ matrix A geldt*

$$\begin{aligned}
\|A\| &\leq m \|A\|_{\max} \\
\|A\| &\leq \sqrt{m} \|A\|_1 \\
\|A\| &\leq \sqrt{m} \|A\|_{\infty}
\end{aligned}$$

Bewijs. Met de uitdrukkingen voor $\|A\|_1$ en $\|A\|_{\infty}$ als maximale rij- en kolom-sommen (stelling 2.18) volgt eenvoudig dat:

$$\begin{aligned}
\|A\|_1 &\leq m \|A\|_{\max} && \leq m \|A\|_{\infty} \\
\|A\|_{\infty} &\leq m \|A\|_{\max} && \leq m \|A\|_1
\end{aligned}$$

en als we dit invullen in $\|A\| \leq \sqrt{\|A\|_1 \|A\|_{\infty}}$ dan volgen de genoemde ongelijkheden. $\square$

Stelling 2.23. Voor een symmetrische of hermitische matrix A geldt $\|A\|_1 = \|A\|_\infty$ en

$$\begin{aligned}\|A\| &\leq \|A\|_1 \\ \|A\| &\leq \|A\|_\infty\end{aligned}$$

Bewijs. In een symmetrische of hermitische matrix is er voor elke rij som een even grote kolomsom en vice versa en dus $\|A\|_1 = \|A\|_\infty$. Invullen in $\|A\| \leq \sqrt{\|A\|_1 \|A\|_\infty}$ geeft de ongelijkheden. $\square$

Stelling 2.24. Voor een holle $m \times m$ matrix A geldt $\|A\| \leq (m-1) \|A\|_{\max}$.

Bewijs. In dit geval bestaat de diagonaal uit nullen en hebben we

$$\begin{aligned}\|A\|_1 &\leq (m-1) \|A\|_{\max} \\ \|A\|_\infty &\leq (m-1) \|A\|_{\max}\end{aligned}$$

en daarmee volgt $\|A\| \leq \sqrt{\|A\|_1 \|A\|_\infty} \leq (m-1) \|A\|_{\max}$. $\square$

Stelling 2.25. Voor een matrix A met in elke rij en kolom maximaal één niet-nul element (bijvoorbeeld een diagonaalmatrix) geldt $\|A\| = \|A\|_{\max}$.

Bewijs. Voor zo'n matrix geldt $\|A\|_1 = \|A\|_\infty = \|A\|_{\max}$ (de maximale rij- en kolomsommen bestaan uit één element: het grootste element) en dus

$$\|A\| \leq \sqrt{\|A\|_1 \|A\|_\infty} = \|A\|_{\max}$$

Noemen we het grootste element $a_{\rho\kappa}$ (dus $\|A\|_{\max} = a_{\rho\kappa}$) en nemen we de eenheidsvector $x = e_\kappa$ zodat $|x| = x_\kappa = 1$ dan is $|Ax| = |a_{\rho\kappa}| = \|A\|_{\max}$ en dus $\|A\| \geq |Ax| = \|A\|$. Conclusie: $\|A\| = \|A\|_{\max}$. $\square$

Stelling 2.26. De norm van een vierkante matrix blijft hetzelfde bij het toevoegen of verwijderen van lege rijen en kolommen.

Bewijs. Dit volgt eenvoudig uit de definitie van de matrixnorm. Als we een lege rij aan een matrix A toevoegen dan krijgt de vector Ax een extra nulelement en verandert dus niet van lengte. Als we een lege kolom aan A toevoegen hebben we een extra element in de vector x nodig. Omdat we dit element met een lege kolom vermenigvuldigen zal het niets bijdragen aan Ax . Om de maximale lengte van Ax te vinden kunnen we dit nieuwe element dus het beste nulstellen - en dan vinden we precies de oorspronkelijke norm terug. $\square$

Voorbeeld 2.27. Noem $x = (x_1 \ x_2 \ x_3)^T$ dan geldt bijvoorbeeld

$$\left\| \begin{pmatrix} 1 & 0 & -3 \\ -4 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix} \right\| = \max_{|x|=1} \left| \begin{pmatrix} 1 & 0 & -3 \\ -4 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \right| = \max_{|x|=1} \begin{vmatrix} 1x_1 - 3x_3 \\ -4x_1 + 6x_3 \\ 0 \end{vmatrix}$$

en ook met $y = (y_1 \ y_2)^T$

$$\left\| \begin{pmatrix} 1 & -3 \\ -4 & 6 \end{pmatrix} \right\| = \max_{|y|=1} \left| \begin{pmatrix} 1 & -3 \\ -4 & 6 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right| = \max_{|y|=1} \begin{vmatrix} 1y_1 - 3y_2 \\ -4y_1 + 6y_2 \end{vmatrix}$$

en deze maxima zijn aan elkaar gelijk voor $x_{\max} = (x_1 \ x_2 \ x_3)^T = (y_1 \ 0 \ y_2)^T$ bij $y_{\max} = (y_1 \ y_2)^T$.

Stelling 2.28. *De norm van een deelmatrix X van A is niet groter dan $\|A\|$.*

Bewijs. Een deelmatrix X wordt gevormd door een aantal rijen en kolommen uit A te selecteren en de overige rijen en kolommen te verwijderen. Nemen we de matrix Y die even groot is als A en gevormd wordt door de betreffende rijen en kolommen uit A te selecteren en de overige rijen en kolommen met nullen te vullen (zodat $\|X\| = \|Y\|$, stelling 2.26), dan kunnen we Y uitdrukken als

$$Y = RAK$$

waarbij R en K diagonaalmatrices zijn met diagonaalelementen gelijk aan 1 als de betreffende rij (R) of kolom (K) geselecteerd moet worden. Er geldt dan $\|R\| = \|K\| = 1$ (stelling 2.25) en daarmee volgt met de submultiplicatieve eigenschap:

$$\|X\| = \|Y\| = \|RAK\| \leq \|R\| \|A\| \|K\| = \|A\|$$

□

In het bijzonder volgt met deze stelling dat

$$\|A\|_{\max} \leq \|A\|$$

als we de deelmatrix kiezen die alleen uit het grootste element van A bestaat.

Voorbeeld 2.29. Toepassen op de matrix

$$A = \begin{pmatrix} 1 & 2 & -3 \\ -4 & 5 & 6 \\ 7 & -8 & 9 \end{pmatrix}$$

van voorbeeld 2.20 geeft bijvoorbeeld

$$\begin{aligned} 7.83 \approx \|X\| &= \left\| \begin{pmatrix} 1 & -3 \\ -4 & 6 \end{pmatrix} \right\| \\ &= \left\| \begin{pmatrix} 1 & 0 & -3 \\ -4 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix} \right\| \\ &\leq \left\| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & 2 & -3 \\ -4 & 5 & 6 \\ 7 & -8 & 9 \end{pmatrix} \right\| \left\| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\| = \|A\| \approx 14.2 \end{aligned}$$

De volgende "majoranten"-stelling is zeer nuttig:

Stelling 2.30. Zij A een $nk \times nk$ blokmatrix bestaande uit n^2 deelmatrices A_{ij} ($1 \leq i, j \leq n$) van afmeting $k \times k$ en R een reële $n \times n$ matrix die bovengrenzen bevat voor de normen van alle blokjes:

$$\|A_{ij}\| \leq r_{ij}$$

dan geldt

$$\|A\| \leq \|R\|$$

Bewijs. Neem de vector $x \in \mathbb{C}^{nk}$ waarvoor $|Ax|$ maximaal is en neem $x_i \in \mathbb{C}^k$ zodat $x^T = (x_1^T, x_2^T, \dots, x_n^T)$. Met de definitie van de matrixnorm volgt:

$$\begin{aligned} \|A\| &= |Ax| \\ &= \sqrt{\begin{matrix} |A_{11}x_1 + A_{12}x_2 + \dots + A_{1n}x_n|^2 \\ + |A_{21}x_1 + A_{22}x_2 + \dots + A_{2n}x_n|^2 \\ \vdots \\ + |A_{n1}x_1 + A_{n2}x_2 + \dots + A_{nn}x_n|^2 \end{matrix}} \\ &\leq \sqrt{\begin{matrix} (|A_{11}x_1| + |A_{12}x_2| + \dots + |A_{1n}x_n|)^2 \\ + (|A_{21}x_1| + |A_{22}x_2| + \dots + |A_{2n}x_n|)^2 \\ \vdots \\ + (|A_{n1}x_1| + |A_{n2}x_2| + \dots + |A_{nn}x_n|)^2 \end{matrix}} \quad (\text{driehoeksongelijkheid}) \\ &\leq \sqrt{\begin{matrix} (r_{11}|x_1| + r_{12}|x_2| + \dots + r_{1n}|x_n|)^2 \\ + (r_{21}|x_1| + r_{22}|x_2| + \dots + r_{2n}|x_n|)^2 \\ \vdots \\ + (r_{n1}|x_1| + r_{n2}|x_2| + \dots + r_{nn}|x_n|)^2 \end{matrix}} \quad (\text{definitie } \|A_{ij}\|) \\ &= \sqrt{\begin{matrix} |r_{11}y_1 + r_{12}y_2 + \dots + r_{1n}y_n|^2 \\ + |r_{21}y_1 + r_{22}y_2 + \dots + r_{2n}y_n|^2 \\ \vdots \\ + |r_{n1}y_1 + r_{n2}y_2 + \dots + r_{nn}y_n|^2 \end{matrix}} \\ &= |Ry| \\ &\leq \|R\| |y| \\ &= \|R\| \end{aligned}$$

waarbij $y \in \mathbb{R}_+^n$ de vector gegeven door $y_i = |x_i| \geq 0$ is zodat $|y| = |x| = 1$. $\square$

In het bijzonder volgt uit deze stelling dat de norm van een matrix niet kleiner wordt als we de elementen ($k = 1$) vervangen door hun absolute waarden.

Stelling 2.31. De norm van een matrix is invariant onder unitaire transformaties.

Bewijs. Het vermenigvuldigen van een unitaire matrix U met een vector x levert een vector Ux die dezelfde lengte heeft als x ; het is immers precies dezelfde vector maar dan opgeschreven in een andere orthonormale basis. Daarmee volgt:

$$\begin{aligned}\|U^{-1}AU\| &= \max_{|x|=1} |U^{-1}Ax| \\ &= \max_{|y|=1} |U^{-1}Ay| \\ &= \max_{|y|=1} |Ay| \\ &= \|A\|\end{aligned}$$

voor unitaire U en willekeurige A . □

2.5 De commutatorstelling gekwantificeerd

Nu we heel wat weten over matrixnormen kunnen we deze kennis toe gaan passen op de commutatorstelling. We beginnen door de norm van de commutatierelatie $A = BC - CB$ te bekijken:

$$\begin{aligned}\|A\| &= \|BC - CB\| \\ &\leq \|BC\| + \|CB\| && \text{(driehoeksomgelijkheid)} \\ &\leq \|B\| \|C\| + \|C\| \|B\| && \text{(submultiplicativiteit)} \\ &= 2 \|B\| \|C\|\end{aligned}$$

daarmee volgt de ondergrens

$$\|B\| \|C\| \geq \frac{1}{2} \|A\| \quad (2.1)$$

Over de norm van B en C apart valt in het algemeen niets te zeggen, zij kunnen immers willekeurig groot of klein gemaakt worden. Stel maar $A = BC - CB$ en neem $\tilde{B} = \varepsilon B$ en $\tilde{C} = C/\varepsilon$ voor willekeurige $\varepsilon \neq 0$, dan geldt ook $A = \tilde{B}\tilde{C} - \tilde{C}\tilde{B} = \varepsilon B \frac{1}{\varepsilon} C - \frac{1}{\varepsilon} C \varepsilon B = BC - CB = A$. Om die reden beperken we ons tot het afschatten van het product $\|B\| \|C\|$. We kunnen daarbij slim gebruik maken van het feit dat het geheel zich netjes gedraagt onder unitaire transformaties:

Lemma 2.32. *Voor een spoorloze matrix A bestaan er matrices B en C zodat $A = BC - CB$ en $\|B\| \|C\| \leq K \|A\|$ dan en slechts dan als er matrices $\tilde{B}$ en $\tilde{C}$ bestaan zodat $\tilde{A} = \tilde{B}\tilde{C} - \tilde{C}\tilde{B}$ en $\|\tilde{B}\| \|\tilde{C}\| \leq K \|\tilde{A}\|$, waarbij $\tilde{A}$ een holle matrix is die unitair equivalent met A is.*

Bewijs. Omdat de norm van een matrix invariant is onder unitaire transformaties (stelling 2.31) volgt dit resultaat direct met lemma 2.13, waar $\|A\| = \|\tilde{A}\|$, $\|B\| = \|\tilde{B}\|$, $\|C\| = \|\tilde{C}\|$. □

Nu kunnen we de normen van de matrices B en C die in het bewijs van de commutatorstelling worden geconstrueerd gaan afschatten. Het is daarbij handig om eerst naar het eenvoudige geval van 2×2 matrices te kijken, met name omdat de norm dan relatief eenvoudig te berekenen is. Dit geeft gelijk een mooi resultaat:

Stelling 2.33. *Voor elke spoorloze matrix $A \in \mathbb{C}^{2 \times 2}$ bestaan er matrices $B, C \in \mathbb{C}^{2 \times 2}$ zodat $A = BC - CB$ en $\|B\| \|C\| = \frac{1}{2} \|A\|$.*

Bewijs. Het voldoet om de bewering te bewijzen voor holle A , met lemma 2.32 volgt dan automatisch het resultaat voor willekeurige A . Zij dus A een holle 2×2 matrix:

$$A = \begin{pmatrix} 0 & a_{12} \\ a_{21} & 0 \end{pmatrix}.$$

Neem nu B en C gegeven door

$$B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad C = \begin{pmatrix} 0 & \frac{1}{2}a_{12} \\ -\frac{1}{2}a_{21} & 0 \end{pmatrix}$$

zodat $A = BC - CB$. De normen van deze eenvoudige matrices zijn in dit geval gelijk aan het grootste element (stelling 2.25). Er volgt

$$\|B\| = \max\{|1|, |-1|\} = 1$$

$$\|C\| = \max\left\{\left|\frac{1}{2}a_{12}\right|, \left|-\frac{1}{2}a_{21}\right|\right\} = \frac{1}{2} \max\{|a_{12}|, |a_{21}|\} = \frac{1}{2} \|A\|$$

zodat $\|B\| \|C\| = \frac{1}{2} \|A\|$. Merk op dat dit optimaal is, vanwege de driehoeksongelijkheid geldt $\|B\| \|C\| \geq \frac{1}{2} \|A\|$ (vergelijking 2.1) en dus kunnen er geen B en C bestaan die nog kleiner zijn. $\square$

Voorbeeld 2.34. De matrices

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad C = \begin{pmatrix} 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 \end{pmatrix}$$

voldoen aan de commutatierelatie $A = BC - CB$ en $\|B\| \|C\| = 1 \cdot \frac{1}{2} = \frac{1}{2} \cdot 1 = \frac{1}{2} \|A\|$.

Het is niet zo dat $\|B\| \|C\| = \frac{1}{2} \|A\|$ voor alle 2×2 matrices A, B, C waarvoor $A = BC - CB$, neem immers

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad C = \begin{pmatrix} 10 & \frac{1}{2} \\ -\frac{1}{2} & 10 \end{pmatrix}$$

dan $A = BC - CB$ maar $\|B\| \|C\| \geq 1 \cdot 10 > \frac{1}{2} \|A\| = \frac{1}{2}$.

Wel heeft de constructie van voorbeeld 2.34 een interessant gevolg; de ondergrens die gegeven wordt door de driehoeksongelijkheid (vergelijking 2.1) is voor alle m de grootst mogelijke ondergrens voor de verzameling van alle paren B en C :

Opmerking 2.35. Voor alle $m \geq 2$ bestaan er matrices A , B en C zodat $\|B\| \|C\| = \frac{1}{2} \|A\|$.

Bewijs. Neem

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \quad C = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 \\ -\frac{1}{2} & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

dan $A = BC - CB$ en omdat het toevoegen van lege rijen of kolommen geen invloed heeft op de norm van een matrix (stelling 2.26) geldt $\|B\| \|C\| = 1 \cdot \frac{1}{2} = \frac{1}{2} \cdot 1 = \frac{1}{2} \|A\|$. $\square$

Deze constructie van stelling 2.33 kunnen we eenvoudig uitbreiden naar $m \times m$ matrices, maar we zullen zien dat dit niet zo'n strakke grens levert als voor 2×2 matrices:

Stelling 2.36 (Commutatorstelling $\mathcal{O}(m^2)$). *Voor elke spoorloze matrix $A \in \mathbb{C}^{m \times m}$ bestaan er matrices $B, C \in \mathbb{C}^{m \times m}$ zodat $A = BC - CB$ en $\|B\| \|C\| \leq \frac{1}{2} (m-1)^2 \|A\|$.*

Bewijs. Wederom volstaat het om de bewering te bewijzen voor holle A . Zij dus A een holle $m \times m$ matrix met elementen a_{ij} . Neem nu B een diagonaalmatrix met elementen die het interval $[-1, 1]$ gelijkmatig opvullen met onderlinge afstand $d = 2/(m-1)$:

$$b_{ii} = -1 + (i-1)d = -1 + \frac{2(i-1)}{m-1}, \quad 1 \leq i \leq m$$

(bijvoorbeeld $\{b_{ii}\} = \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$ voor $m = 5$). Dan liggen twee willekeurige elementen van B altijd minstens de scheidingsafstand d van elkaar af:

$$|b_{ii} - b_{jj}| \geq d = \frac{2}{m-1}, \quad i \neq j$$

(gelijkheid voor 'buren' $i = j \pm 1$). Omdat B een diagonaalmatrix is is de norm gelijk aan het grootste element:

$$\|B\| = \max_i |b_{ii}| = 1.$$

Neem nu ook C een holle matrix met elementen gegeven door

$$c_{ij} = \frac{a_{ij}}{b_{ii} - b_{jj}}$$

zodat $A = BC - CB$. Wegens de keuze de b_{ii} 's geldt voor de grootte van de niet-diagonaal elementen van C dat

$$|c_{ij}| = \frac{|a_{ij}|}{|b_{ii} - b_{jj}|} \leq \frac{m-1}{2} |a_{ij}|.$$

Bovendien zijn elementen van een matrix altijd kleiner dan de norm van de matrix, dus $|a_{ij}| \leq \|A\|$ en

$$|c_{ij}| \leq \frac{m-1}{2} \|A\|.$$

Nu kunnen we de norm van de holle matrix C afschatten (stelling 2.24) en volgt

$$\|B\| \|C\| = \|C\| \leq (m-1) \max_{i,j} |c_{ij}| \leq \frac{1}{2} (m-1)^2 \|A\|$$

Voor $m = 2$ vinden we het resultaat van stelling 2.33 weer terug. $\square$

Hoewel we alle elementen van C kunnen afschatten met de corresponderende elementen van A , i.e.

$$|c_{ij}| \leq K |a_{ij}|$$

voor een zekere $K \in \mathbb{R}$, kunnen we daaruit niet concluderen dat $\|C\| \leq K \|A\|$. Dit komt doordat we iets weten over de *grootte* van de elementen, maar niet over het *teken* (of in het complexe geval: de *fase*). Stel bijvoorbeeld

$$A = \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ -2 & 2 & 0 \end{pmatrix} \quad \text{en} \quad C = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

dan $|c_{ij}| = \frac{1}{2} |a_{ij}|$ maar $\|A\| = \sqrt{12} < 4$ en $\|C\| = 2$ en dus $\|C\| \not\leq \frac{1}{2} \|A\|$.

Stelling 2.36 kan nog wat aanscherpt worden. In de constructie gebruiken we een B waarvoor $\|B\| = 1$ en dus $|b_{ii}| \leq 1$. Omdat we complex werken hoeven we de elementen van B echter niet te beperken tot het interval $[-1, 1]$ maar kunnen we ze verdelen binnen de gehele eenheidscirkel (de eenheidsschijf) in het complexe vlak. Om een zo goed mogelijke grens te vinden zoeken we dus een verdeling van punten in de eenheidsschijf waarbij de minimale afstand tussen een paar punten maximaal is. Dit is op zichzelf een zeer interessant probleem dat al decennia lang bestudeerd wordt. Het is equivalent met het vinden van een verdeling van m zo groot mogelijke cirkels in de eenheidsschijf[3]. Voor kleine m zijn een aantal verdelingen bewezen optimaal, voor grotere m zijn alleen numerieke oplossingen gevonden. Tegenwoordig zijn op internet lijsten te vinden van de beste bekende oplossingen[12]. Wij zullen ons beperken tot een ondergrens voor de optimale verdeling.

Lemma 2.37. *Voor alle $m \geq 2$ bestaat er een verdeling van m punten $\{x_1, x_2, \dots, x_m\} \in \mathbb{C}$ binnen de eenheidscirkel ($|x_i| \leq 1$ voor alle $i = 1, 2, \dots, m$) zodat*

$$|x_i - x_j| \geq \sqrt{\frac{2}{m-1}}$$

voor alle $i \neq j \in 1, 2, \dots, m$.

Bewijs. Het is zeer waarschijnlijk mogelijk om $m \geq 2$ even grote, niet overlappende cirkels in de eenheidsschijf te plaatsen op zo'n manier dat minstens de helft van de oppervlakte van de eenheidsschijf bedekt is. Voor kleine m zijn zulke bedekkingen gevonden (zie [3] voor voorbeelden en getalwaarden voor deze 'packing density', voor $m = 2$ wordt in het optimale geval precies de helft van de eenheidsschijf bedekt). Voor grotere m is het intuïtief duidelijk dat het kan, al is niet eenvoudig om met een bewijs te komen. We zullen zo een andere manier bekijken om tot een vergelijkbaar resultaat te komen zonder deze aanname.

Noemen we de straal van deze cirkels r dan volgt

$$m\pi r^2 = A_{\text{vulcirkels}} \geq \frac{1}{2} A_{\text{eenheidsschijf}} = \frac{1}{2} \pi (1)^2$$

en dus

$$r \geq \sqrt{\frac{1}{2m}}.$$

Noemen we $\{m_1, m_2, \dots, m_m\}$ de middelpunten van deze cirkels, dan liggen deze allen in de cirkel met straal $1-r$ en geldt voor de afstand tussen twee willekeurige middelpunten

$$|m_i - m_j| \geq 2r = \sqrt{\frac{2}{m}}.$$

Nu kunnen het geheel nog opschalen met een factor $1/(1-r)$ zonder dat er punten buiten de eenheidscirkel komen te liggen, dus we nemen de punten $\{x_1, x_2, \dots, x_m\}$ gegeven door

$$x_i = \frac{m_i}{1-r}$$

zodat

$$|x_i - x_j| = \frac{|m_i - m_j|}{1-r} \geq \frac{2r}{1-r} = \frac{\sqrt{\frac{2}{m}}}{1 - \sqrt{\frac{1}{2m}}} = \frac{\sqrt{2}}{\sqrt{m} - \sqrt{\frac{1}{2}}} \geq \frac{\sqrt{2}}{\sqrt{m-1}} = \sqrt{\frac{2}{m-1}}$$

immers $\sqrt{m-1} \geq \sqrt{m} - \sqrt{\frac{1}{2}}$ voor $m \geq 2$ (ga na voor $m = 2$ er merk op dat een wortel sneller groeit voor kleinere waarden, dus $\sqrt{m-1}$ blijft groter dan $\sqrt{m} - \sqrt{\frac{1}{2}}$). $\square$

Zoals opgemerkt in bovenstaand bewijs gebruiken we de aanname dat het mogelijk om de helft van de oppervlakte van een cirkel te bedekken met m even grote cirkels die geheel in de cirkel liggen. Omdat dit niet zo eenvoudig aan te tonen is bekijken we nog een lemma over de verdeling van punten in de eenheidsschijf. Ditmaal doen we geen aannames en geven we een directe constructie voor de locatie van de punten. Dit heeft als voordeel dat we deze indeling eenvoudig kunnen gebruiken in bijvoorbeeld een computeralgoritme, maar daartegenover staat dat het resultaat ietsje minder mooi is.

Lemma 2.38. Voor alle $m \geq 2$ bestaat er een verdeling van m punten $\{x_1, x_2, \dots, x_m\} \in \mathbb{C}$ over concentrische cirkels binnen of op de eenheidscirkel ($|x_i| \leq 1$ voor alle $i = 1, 2, \dots, m$) zodat

$$|x_i - x_j| \geq \frac{1}{\sqrt{\frac{m-1}{2}} + 1}$$

voor alle $i \neq j \in 1, 2, \dots, m$.

Bewijs. Neem $d \in \mathbb{R}$ zodanig dat $\frac{1}{d}$ het kleinste gehele getal groter dan $\sqrt{\frac{m-1}{2}}$ is, dus

$$\frac{1}{d} = \left\lceil \sqrt{\frac{m-1}{2}} \right\rceil \leq \sqrt{\frac{m-1}{2}} + 1$$

en dus

$$d \geq \frac{1}{\sqrt{\frac{m-1}{2}} + 1}$$

We construeren een indeling van punten met onderlinge afstand minstens d op concentrische cirkels rond de oorsprong in het complexe vlak. We nemen cirkels met stralen

$$r_k = kd$$

voor $1 \leq k \leq \frac{1}{d}$, dus de eerste cirkel heeft straal d en de laatste komt overeen met de eenheidscirkel. De afstand tussen punten op verschillende cirkels is zo altijd minstens d . Vul nu elke cirkel met de punten

$$x_{kn} = r_k \exp(in\gamma(r_k, d))$$

voor $0 \leq n \leq \left\lfloor \frac{2\pi}{\gamma(r_k, d)} \right\rfloor - 1$ waarbij $\gamma(r, d)$ de hoek is tussen twee punten die op een cirkel met straal r op afstand precies d van elkaar liggen. Voor een afschatting van γ kunnen we het beste naar twee uitersten kijken. Voor hele kleine hoeken nadert d de booglengte γr . Aan de andere kant geldt dat als de twee punten tegenover elkaar op een cirkel liggen de hoek maximaal ($\gamma = \pi$) is en de afstand is in dat geval $2r$. Er geldt dus:

$$\frac{d}{r} \leq \gamma(r, d) \leq \frac{\pi d}{2r}$$

Nu volgt met $1 + 2 + \dots + n = \frac{1}{2}n(n+1)$ voor het aantal punten dat we op

deze manier over de ringen kunnen verdelen:

$$\begin{aligned}
N &= \sum_{k=0}^{\frac{1}{d}} \left\lfloor \frac{2\pi}{\gamma(r_k, d)} \right\rfloor \\
&\geq \sum_{k=1}^{\frac{1}{d}} \left\lfloor \frac{2\pi}{\frac{\pi d}{2r_k}} \right\rfloor \\
&= \sum_{k=1}^{\frac{1}{d}} \left\lfloor \frac{4r_k}{d} \right\rfloor \\
&= 4 \sum_{k=1}^{\frac{1}{d}} \left\lfloor \frac{kd}{d} \right\rfloor \\
&= 4 \sum_{k=0}^{\frac{1}{d}} k \\
&= 4 \left[\frac{1}{2} \left(\frac{1}{d} \right) \left(\frac{1}{d} + 1 \right) \right] \\
&= 2 \left(\frac{1}{d} \right)^2 + 2 \frac{1}{d} \\
&\geq 2 \left(\frac{1}{d} \right)^2 \\
&\geq 2 \left(\sqrt{\frac{m-1}{2}} \right)^2 \\
&= m-1
\end{aligned}$$

Omdat de kleinste cirkel straal d heeft kunnen we de oorsprong als extra punt toevoegen, en dus kunnen we zeker m punten kwijt op onderlinge afstand d . $\square$

Omdat het resultaat van lemma 2.37 wat fijner rekent dan dat van 2.38 zullen we met het eerste doorwerken. Merk op dat de resultaten van beide lemma's elkaar naderen voor grote m . De verdeling van punten stelt ons nu in staat om de commutatorstelling weer een beetje aan te scherpen:

Stelling 2.39 (Commutatorstelling $\mathcal{O}(m^{3/2})$). *Voor elke spoorloze matrix $A \in \mathbb{C}^{m \times m}$ bestaan er matrices $B, C \in \mathbb{C}^{m \times m}$ zodat $A = BC - CB$ en $\|B\| \|C\| \leq \frac{1}{\sqrt{2}} (m-1)^{3/2} \|A\|$.*

Bewijs. Het bewijs gaat op precies dezelfde wijze als stelling 2.36, maar nu met

$$d \geq \sqrt{\frac{2}{m-1}}$$

wegens lemma 2.37. $\square$

Merk op dat voor $m = 2$ stelling 2.39 niet meer de optimale grens geeft ($\frac{1}{\sqrt{2}}$ i.p.v. $\frac{1}{2}$). Dit komt doordat we een algemene afschatting voor de verdeling van punten in de eenheidsschijf hebben gemaakt, waardoor het optimum voor een bijzonder geval als $m = 2$ niet meer bereikt wordt. In ruil hiervoor hebben we wel een strakkere grens voor grote m gevonden. Dit zullen we verderop ook weer terugzien, we zullen dan de precieze grens voor kleine m links laten liggen en alleen kijken naar het gedrag van de grens voor grote m .

Opmerking 2.40. Je zou je nu kunnen afvragen of alle commutatierelaties $A = BC - CB$ op unitaire equivalentie na van de vorm $H = DX - XD$ zijn met H een holle matrix en D een diagonaalmatrix. Als B unitair equivalent is met een diagonaalmatrix is dit zo, want dan $DX - XD$ is altijd hol voor D diagonaal (ga na met een handige definitie van het matrixproduct). Ook als C unitair equivalent is met een diagonaalmatrix kunnen we een overeenkomende relatie van deze vorm vinden, want we kunnen altijd B en C verwisselen door $B' = -C$ en $C' = B$ te nemen zodat $B'C' - C'B' = (-C)B - B(-C) = BC - CB$. In het algemeen lukt het echter niet. Nemen we voor B een niet-normale matrix en $C = B^*$, bijvoorbeeld:

$$B = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \quad C = B^* = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$

dan

$$A = BC - CB = BB^* - B^*B = \begin{pmatrix} 0 & -4 \\ -4 & 0 \end{pmatrix}$$

en omdat B en $C = B^*$ niet normaal zijn bestaat er geen enkele unitaire transformatie die B of C diagonaal maakt.

In het bijzonder kunnen we hieruit concluderen dat beweringen over $\|B\| \|C\|$ met B diagonaal niet gebruikt kunnen worden om een weerlegging van de commutatorstelling te construeren. Stel dat we bewijzen dat de diagonaalconstructie aan een zekere ondergrens voldoet ($\|B\| \|C\| \geq f(m) \|A\|$ voor zekere f met B diagonaal) dan kunnen we daar niet uit afleiden dat alle commutatorontbindingen hier aan voldoen.

2.6 Verband met het Kadison-Singer vermoeden

Zoals in de inleiding al werd gezegd impliceert een sterke commutatorstelling (met een constante die niet van de dimensie afhangt) dat het Kadison-Singer vermoeden waar is. Het Kadison-Singer vermoeden is equivalent aan uiteenlopende andere vermoedens [5], waaronder het zogenaamde eindig dimensionale paving vermoeden (stelling 6.3 bewering 3 in [5]) van Anderson. Voordat we het paving vermoeden kunnen bestuderen moeten we eerste weten van een betegeling ('paving') eigenlijk is:

Definitie 2.41. Een *betegeling* (van lengte n) van een matrix $A \in \mathbb{C}^{m \times m}$ is een partitie $\{P_1, P_2, \dots, P_n\}$ van de getallen $1, 2, \dots, m$. Met de tegel T_k bedoelen

we de centrale deelmatrix van A die bestaat uit de rijen en kolommen met indices in P_i , dus $T_k = \{a_{ij} : i \in P_k \text{ en } j \in P_k\}$.

Voorbeeld 2.42. De 7×7 matrix

$$A = \begin{pmatrix} \mathbf{1} & 2 & 3 & 4 & 5 & 6 & 7 \\ 7 & \mathbf{6} & \mathbf{5} & 4 & \mathbf{3} & 2 & 1 \\ 4 & \mathbf{5} & \mathbf{8} & \mathbf{6} & \mathbf{2} & 1 & 7 \\ 7 & 4 & 4 & \mathbf{5} & \mathbf{9} & 8 & 8 \\ 4 & \mathbf{5} & \mathbf{6} & \mathbf{2} & \mathbf{1} & 3 & 6 \\ 7 & 8 & 9 & 5 & 4 & \mathbf{2} & \mathbf{1} \\ 4 & 5 & 4 & 5 & 9 & \mathbf{8} & \mathbf{7} \end{pmatrix}$$

kan bijvoorbeeld met $\{\{1\}, \{2, 3, 4, 5\}, \{6, 7\}\}$ betegeld worden, wat de volgende 3 tegels oplevert:

$$T_1 = (1) \quad T_2 = \begin{pmatrix} 6 & 5 & 4 & 3 \\ 5 & 8 & 6 & 2 \\ 4 & 4 & 5 & 9 \\ 5 & 6 & 2 & 1 \end{pmatrix} \quad T_3 = \begin{pmatrix} 2 & 1 \\ 8 & 7 \end{pmatrix}$$

(de tegels komen overeen met de dik gedrukte elementen van A).

Merk op dat een tegel niet per se uit aangesloten rijen en kolommen van A hoeft te bestaan, in bovenstaand voorbeeld zou $\{\{1, 6, 7\}, \{2, 3, 4, 5\}\}$ ook een geldige betegeling zijn. Het paving vermoeden zegt nu iets over de normen van de tegels:

Vermoeden 2.43 (Paving vermoeden). *Voor alle $\varepsilon > 0$ bestaat er een geheel getal n zodat iedere holle $m \times m$ matrix $A \in \mathbb{C}^{m \times m}$ te betegelen is met hoogstens n tegels met norm $\|T_k\| \leq \varepsilon \|A\|$ voor alle $1 \leq k \leq n$.*

Het bijzondere van dit vermoeden zit hem in het feit dat het aantal tegels niet van de matrixdimensie m afhangt. Dit betekent dat de afmeting van de tegels vanaf een zekere m steeds groter wordt naarmate m toeneemt, maar als het vermoeden klopt blijft de norm van de tegels desondanks klein. Zou het aantal tegels n wel van de matrixdimensie m af mogen hangen dan is het vermoeden triviaal: we kunnen dan immers de betegeling van lengte m nemen die alleen bestaat uit de diagonaalelementen van A en die zijn allen 0 (A is immers hol). Ook voor $\varepsilon > 1$ is het vermoeden triviaal: alle deelmatrices van A hebben immers norm $\leq \|A\|$ (stelling 2.28).

De volgende stelling uit het artikel dat in dit project centraal staat [6] laat zien dat een sterke commutatorstelling het paving vermoeden bevestigt:

Stelling 2.44. *Voor alle $0 < \varepsilon < 1$ en holle $A = BC - CB$ bestaat er een betegeling van A van lengte $\lceil \frac{2}{\varepsilon} \rceil^2$ en norm $\sqrt{2}\varepsilon \|B\| \|C\|$.*

Bewijs. Neem $\varepsilon' < \varepsilon$ zodat $N = \frac{2}{\varepsilon'} = \lceil \frac{2}{\varepsilon} \rceil$. Neem ook $B' = B/\|B\|$ en $C = \|B\|C$ zodat $A = B'C' - C'B'$ met $\|B'\| = 1$ en $\|C'\| = \|B\| \|C\|$. Verdeel

het interval $[-1, 1]$ in N intervallen I_i van lengte ε' en daarmee het complexe eenheidsvierkant $[-1, 1] \times [-\iota, \iota]$ in N^2 vierkanten $V_{ij} = I_i \times \iota I_j$. Noem $B'(i, j)$ de centrale (diagonale) deelmatrix van B' die alle elementen van B' in het vierkant V_{ij} bevat, en noem $A(i, j)$ en $C'(i, j)$ de centrale deelmatrices van A en C' die worden gevormd door de dezelfde rijen en kolommen als $B'(i, j)$. Dan geldt

$$A(i, j) = B'(i, j) C'(i, j) - C'(i, j) B'(i, j)$$

(zie het verband tussen de elementen van A , B en C in het bewijs van de commutatorstelling).

Omdat $\|B'\| = 1$ liggen alle elementen van B' in $[-1, 1] \times [-\iota, \iota]$ en dus in een zekere $B'(i, j)$ en daarmee vormen de $A(i, j)$ een betegeling van A bestaande uit hoogstens N^2 tegels (sommige $A(i, j)$ kunnen ook leeg zijn). Om de norm van deze tegels af te schatten noemen we v_{ij} het midden van het vierkant V_{ij} en merken we op dat de diagonaalmatrix $B'(i, j) - v_{ij}$ nu enkel elementen in $[-\frac{1}{2}\varepsilon', \frac{1}{2}\varepsilon'] \times [-\frac{1}{2}\iota\varepsilon', \frac{1}{2}\iota\varepsilon']$ bevat en dus $\|B'(i, j) - v_{ij}\| \leq \frac{1}{2}\sqrt{2}\varepsilon'$. Samen met de driehoeksongelijkheid volgt nu

$$\begin{aligned} \|A(i, j)\| &= \|B'(i, j) C'(i, j) - C'(i, j) B'(i, j)\| \\ &= \|(B'(i, j) - v_{ij}) C'(i, j) - C'(i, j) (B'(i, j) - v_{ij})\| \\ &\leq 2 \|(B'(i, j) - v_{ij})\| \|C'(i, j)\| \\ &\leq \sqrt{2}\varepsilon' \|C'(i, j)\| \\ &\leq \sqrt{2}\varepsilon \|C'\| \\ &\leq \sqrt{2}\varepsilon \|B\| \|C\| \end{aligned}$$

□

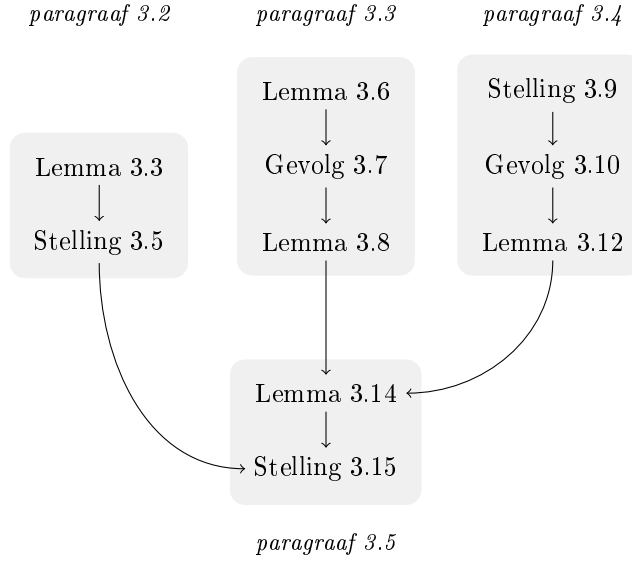
Als we er nu in slagen een sterke commutatorstelling te bewijzen, i.e. voor alle holle A bestaan er B en C waarvoor $\|B\| \|C\| \leq K \|A\|$ voor zekere K onafhankelijk van de matrixdimensie, dan volgt met bovenstaande stelling dat iedere holle A ook te betegelen is met een vast aantal tegels waarvan de norm niet van de matrixdimensie afhangt - en dat zou het paving vermoeden bevestigen.

Hoofdstuk 3

De commutatorstelling van Johnson, Ozawa en Schechtman

In dit hoofdstuk zullen we het bewijs van de commutatorstelling van William B. Johnson, Narutaka Ozawa en Gideon Schechtman uit [6] bestuderen. Deze stelling is een verbetering van de kwantitatieve versies van de commutatorstelling ($\mathcal{O}(m^2)$ en $\mathcal{O}(m^{3/2})$) die we in het vorige hoofdstuk hebben gezien; zij bewijzen een commutatorstelling $\mathcal{O}(m^\varepsilon)$ voor $\varepsilon > 0$ willekeurig klein. In woorden uitgeschreven betekent dat dat er voor iedere $\varepsilon > 0$ een constante K_ε bestaat zodat iedere $m \times m$ commutator A een ontbinding $A = BC - CB$ heeft waarbij $\|B\| \|C\| \leq K_\varepsilon m^\varepsilon \|A\|$. Het bewijs geeft geen expliciete uitdrukking voor de waarde van K_ε , er wordt slechts aangetoond dat zij bestaat.

Het doel van dit hoofdstuk is niet om het artikel letterlijk over te nemen, maar juist om zoveel mogelijk structuur en duidelijkheid in het (soms complexe) bewijs aan te brengen zodat iedere stap zonder veel moeite te begrijpen is. Om het geheel overzichtelijk te houden is in figuur 3.1 een diagram van de verschillende lemma's en stellingen die we in dit hoofdstuk zullen tegenkomen opgenomen. In paragraaf 3.1 zullen we eerst wat definities invoeren waarmee het werken aan het bewijs van een commutatorstelling (vaak een mond vol - zie vorige alinea) wat minder omslachtig wordt. Daarna gebruiken we in paragraaf 3.2 een slimme truuk met een complexe contourintegraal om een bepaalde recurrente betrekking af te leiden voor de bovengrens in de commutatorstelling. Deze recurrente betrekking levert ons al direct een verbetering van de commutatorstelling. In paragraaf 3.3 passen we hetzelfde truukje nogmaals in een iets andere vorm toe met de hoop een beter resultaat te krijgen. Dit lukt, maar we hebben er het bestaan van een bepaalde deelmatrix voor nodig. In paragraaf 3.4 zullen we een stelling van Bourgain en Tzafriri gaan gebruiken om aan te tonen dat zo'n deelmatrix ook daadwerkelijk bestaat. De resultaten van deze beide paragrafen combineren we in paragraaf 3.5 in een lemma dat zegt dat,



Figuur 3.1: Schematisch overzicht van de bewijsstructuur

gegeven een bepaalde grens, er altijd een betere grens is (in zekere zin). Beginnen we nu met het resultaat uit paragraaf 3.2 en nemen we willekeurig vaak zo'n betere grens dan verkrijgen we de commutatorstelling uit het artikel (hier stelling 3.15).

Ik heb geprobeerd om waar mogelijk de notatie die de auteurs in hun artikel gebruiken aan te houden, zodat het makkelijk is om de uitwerking in dit verslag te vergelijken met het oorspronkelijke artikel. In een aantal gevallen wordt er ook nieuwe notatie ingevoerd om bepaalde concepten overzichtelijker te maken; als dit gebruikt wordt zal het duidelijk worden aangegeven.

3.1 Wat opmerkingen vooraf

In het hele bewijs wordt net als bij de kwantitatieve commutatorstellingen in het vorige hoofdstuk steeds gebruik gemaakt van lemma 2.12 en 2.32: we kunnen zonder verlies van algemeenheid aannemen dat een commutator A een holle matrix is. De auteurs nemen voor B ook steeds een diagonaalmatrix, zodat C gegeven wordt als in het bewijs van stelling 2.36. Hierdoor kunnen ze gebruik maken van het volgende lemma om de bewijzen te beperken tot matrices van handige afmetingen, bijvoorbeeld machten van vier.

Lemma 3.1. *Als alle $m \times m$ commutators A te ontbinden zijn als $A = BC - CB$ met $\|B\| \|C\| \leq K(m) \|A\|$ en B een diagonaalmatrix, dan zijn ook alle $n \times n$ commutators A met $n \leq m$ te ontbinden als $A = BC - CB$ met $\|B\| \|C\| \leq K(m) \|A\|$ en B een diagonaalmatrix.*

Bewijs. Zij A een $n \times n$ commutator en $m \geq n$. Laat A' de $m \times m$ matrix zijn die verkregen wordt door aan A lege rijen en kolommen toe te voegen, dan is $\|A'\| = \|A\|$ (stelling 2.26). Wegens de aanname bestaat er een diagonaalmatrix B' en een matrix C' zodat $A' = B'C' - C'B'$ en $\|B'\| \|C'\| \leq K(m) \|A'\|$. Omdat de elementen van C' dan gegeven worden door

$$c'_{ij} = \frac{a'_{ij}}{b'_{ii} - b'_{jj}}$$

(zie ook het bewijs van stelling 2.36) zijn de rijen en kolommen die corresponderen met de lege rijen en kolommen in A' ook leeg in C' . Noem B en C de deelmatrices van B' en C' die gevormd worden door de rijen en kolommen corresponderend met A in A' , dan is $A = BC - CB$ (zie bovenstaand verband van matrixelementen) en $\|C\| = \|C'\|$ (stelling 2.26) en $\|B\| \leq \|B'\|$ (stelling 2.28) en dus

$$\|B\| \|C\| \leq \|B'\| \|C'\| \leq K(m) \|A'\| = K(m) \|A\|.$$

□

Daarnaast beperken de auteurs zich ook tot diagonaalmatrices B met elementen in het eenheidsvierkant $[-1, 1] \times [-\imath, \imath]$ in het complexe vlak, zodat $\|B\| \leq \sqrt{2}$ (stelling 2.25). Zo meteen zal blijken waarom dit handiger is dan een B met elementen in de eenheidscirkel zoals bij het bewijs van stelling 2.39 in het vorige hoofdstuk. Om de bewijzen overzichtelijk te houden gebruiken we de volgende notatie:

Definitie 3.2. Noem $\lambda(A)$ het kleinst mogelijke getal waarvoor een holle matrix A te ontbinden is in een diagonaalmatrix B met louter verschillende elementen in $[-1, 1] \times [-\imath, \imath]$ en een matrix C met $\|C\| \leq \lambda(A) \|A\|$, dus

$$\lambda(A) = \min \{ \|C\| : A = BC - CB, B \text{ diagonaal in } [-1, 1] \times [-\imath, \imath] \}.$$

Noem $\lambda(m)$ het kleinst mogelijke getal zodat $\lambda(A) \leq \lambda(m)$ voor alle $m \times m$ matrices A waarvoor $\|A\| = 1$, dus

$$\lambda(m) = \sup \{ \lambda(A) : \|A\| = 1, A \text{ } m \times m \text{ en hol} \}.$$

zodat voor algemene A geldt

$$\lambda(A) = \lambda\left(\frac{A}{\|A\|}\right) \|A\| \leq \lambda(m) \|A\|$$

Stel ook $\lambda(1) = 0$.

Het minimum in de definitie van $\lambda(A)$ wordt daadwerkelijk aangenomen, d.w.z. voor alle A bestaan er B en C die aan de eisen voldoen en $\|C\| = \lambda(A)$. Dit volgt in een aantal stappen. Wegens onze eerdere resultaten (stelling

2.36) kunnen we het minimum beperken tot het minimum over de (niet-lege) verzameling waarvoor $\|C\| \leq \frac{1}{2}(n-1)^2 \|A\|$. Voor zekere A is de verzameling

$$V = \left\{ (B, C) : A = BC - CB, B \text{ diagonaal in } [-1, 1] \times [-\iota, \iota], \|C\| \leq \frac{1}{2}(n-1)^2 \|A\| \right\}$$

van paren (B, C) die aan de eisen voldoen nu begrensd (bijvoorbeeld onder de norm $\|(B, C)\| = \|B\| + \|C\|$; herinner dat alle normen op een eindig-dimensionale ruimte equivalent zijn); immers $\|B\| \leq \sqrt{2}$ en $\|C\| \leq \frac{1}{2}(n-1)^2 \|A\|$. Bovendien is V gesloten, want elke convergente rij $(B_n, C_n) \rightarrow (B, C)$ heeft een limiet in V , i.e. $A = BC - CB$. Dit volgt uit het feit dat

$$\begin{aligned} \|BC - B_n C_n\| &= \|BC - BC_n + BC_n - B_n C_n\| \\ &\leq \|C_n\| \|B - B_n\| + \|B\| \|C - C_n\| \\ &\leq K \|(B, C) - (B_n, C_n)\| \\ \|CB - C_n B_n\| &= \|CB - CB_n + CB_n - C_n B_n\| \\ &\leq \|C\| \|B - B_n\| + \|B_n\| \|C - C_n\| \\ &\leq K \|(B, C) - (B_n, C_n)\| \end{aligned}$$

voor zekere K (want $\|B_n\|$ en $\|C_n\|$ begrensd en normen zijn continu, dus ook $\|B\|$ en $\|C\|$ begrensd). Daarmee

$$\|(BC - CB) - (B_n C_n - C_n B_n)\| \leq 2K \|(B, C) - (B_n, C_n)\| \rightarrow 0$$

voor $(B_n, C_n) \rightarrow (B, C)$ en dus is $(B, C) \mapsto BC - CB$ continu en volgt $A = B_n C_n - C_n B_n = BC - CB$ en dus $(B, C) \in V$ (aan de andere eisen voor B en C wordt op triviale wijze ook voldaan). V is dus compact (gesloten en begrensd) en bovendien is de functie $(B, C) \mapsto \|C\|$ continu, dus wordt het minimum $\lambda(A)$ aangenomen.

Als we nu een grens voor $\lambda(m)$ kunnen vinden volgt daarmee automatisch een commutatorstelling; er bestaan dan immers B en C waarvoor

$$\|B\| \|C\| \leq \sqrt{2} \|C\| \leq \sqrt{2} \lambda(A) \leq \sqrt{2} \lambda(m) \|A\|.$$

Merk op dat $\lambda(m)$ wegens lemma 3.1 een stijgende rij vormt, dus

$$\dots \leq \lambda(m-1) \leq \lambda(m) \leq \lambda(m+1) \leq \dots$$

3.2 De commutatorstelling $\mathcal{O}\left((\log_4 m)^2 \sqrt{m}\right)$

Nu we deze handige definities hebben kunnen we aan de slag gaan. In deze paragraaf zullen we een matrix op een slimme manier in blokjes opdelen en zo een recursierelatie voor λ afleiden. Omdat het complexe vlak eenvoudig in vier kwadranten op de delen is kunnen we dit het beste met 4×4 matrices doen; elk diagonaalblokje krijgt straks zijn eigen kwadrant.

Lemma 3.3. *Zij $\|A\|$ een holle $4^n \times 4^n$ matrix met $\|A\| = 1$ bestaande uit $4^{n-1} \times 4^{n-1}$ deelmatrices A_{ij} :*

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$

Dan geldt voor alle $0 < \varepsilon < 1$:

$$\lambda(A) \leq \frac{2}{1-\varepsilon} \max_{1 \leq i \leq 4} \lambda(A_{ii}) + \frac{6}{\varepsilon^2}$$

In het bijzonder volgt hieruit voor alle $n \geq 1$ de volgende recursierelatie:

$$\lambda(4^n) \leq \frac{2}{1-\varepsilon} \lambda(4^{n-1}) + \frac{6}{\varepsilon^2}$$

waarbij $\lambda(1) = 0$.

Bewijs. Zij $0 < \varepsilon < 1$ en A een willekeurige holle $4^n \times 4^n$ matrix met $\|A\| = 1$. We gaan $4^n \times 4^n$ matrices B' en C' construeren die voldoen aan $A = B'C' - C'B'$ en aan de te bewijzen relatie. Schrijf A als een 4×4 blokmatrix:

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{pmatrix}$$

waarbij de A_{ij} allemaal $4^{n-1} \times 4^{n-1}$ matrices zijn (de diagonaalblokjes A_{ii} zijn hol). Neem B_{ii} en C_{ii} de 'beste' $4^{n-1} \times 4^{n-1}$ matrices die voldoen aan $A_{ii} = B_{ii}C_{ii} - C_{ii}B_{ii}$; dus B_{ii} zijn diagonaalmatrices met elementen in $[-1, 1] \times [-\varepsilon, \varepsilon]$ waarbij C_{ii} zo klein mogelijk is:

$$\|C_{ii}\| = \lambda(A_{ii})$$

Willen we nu $A = BC - CB$ dan zouden we

$$B = \begin{pmatrix} B_{11} & 0 & 0 & 0 \\ 0 & B_{22} & 0 & 0 \\ 0 & 0 & B_{33} & 0 \\ 0 & 0 & 0 & B_{44} \end{pmatrix} \quad \text{en} \quad C = \begin{pmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & C_{23} & C_{24} \\ C_{31} & C_{32} & C_{33} & C_{34} \\ C_{41} & C_{42} & C_{43} & C_{44} \end{pmatrix}$$

kunnen nemen en C_{ij} die voldoen aan

$$A_{ij} = B_{ii}C_{ij} - C_{ij}B_{jj}$$

Helaas gaat dit niet zo makkelijk. Het zou namelijk zo kunnen zijn dat twee elementen van B_{ii} en B_{jj} aan elkaar gelijk zijn, dus $(B_{ii})_{kk} = (B_{jj})_{ll}$ voor zekere $k, l \in \{1, 2, \dots, 4^{n-1}\}$. Dan is het kl element noodzakelijk nul:

$$(B_{ii}C_{ij} - C_{ij}B_{jj})_{kl} = (C_{ij})_{kl} [(B_{ii})_{kk} - (B_{jj})_{ll}] = 0$$

terwijl $(A_{ij})_{kl}$ niet per se nul hoeft te zijn (alleen de diagonaal van de diagonaalblokjes A_{ii} is altijd nul).

Om dit probleem op te lossen kunnen we de B_{ii} een klein beetje te verschuiven zodat de waarden van verschillende B_{ii} in verschillende gebieden liggen. Dit is de reden dat we werken met B 's met waarden in het eenheidsvierkant i.p.v. de eenheidscirkel - we kunnen nu namelijk eenvoudig het eenheidsvierkant in vier stukken delen en elk van de vier B_{ii} 's verkleinen en naar zijn eigen kwadrant verschuiven:

$$\begin{aligned} B'_{11} &= \frac{1-\varepsilon}{2}B_{11} + \left(\frac{1+\varepsilon}{2} + \frac{1+\varepsilon}{2}\iota\right)I \\ B'_{22} &= \frac{1-\varepsilon}{2}B_{22} + \left(-\frac{1+\varepsilon}{2} + \frac{1+\varepsilon}{2}\iota\right)I \\ B'_{33} &= \frac{1-\varepsilon}{2}B_{33} + \left(-\frac{1+\varepsilon}{2} - \frac{1+\varepsilon}{2}\iota\right)I \\ B'_{44} &= \frac{1-\varepsilon}{2}B_{44} + \left(\frac{1+\varepsilon}{2} - \frac{1+\varepsilon}{2}\iota\right)I \end{aligned}$$

(zie ook figuur 3.2). Het gebied met waarden $([-1, 1] \times [-\iota, \iota])$ wordt hier steeds verkleind met een factor $\frac{1-\varepsilon}{2} = \frac{1}{2} - \frac{1}{2}\varepsilon$ en deze verkleinde versie wordt naar één van de vier hoeken van $[-1, 1] \times [-\iota, \iota]$ verschoven (zie figuur 3.2). Elk van de nieuwe B'_{ii} bevat nu waarden die in hun eigen kwadrant liggen en bovendien een afstand van minstens ε tot de reële of imaginaire as hebben. We kunnen niet $\varepsilon = 0$ kiezen, want dan zouden twee B'_{ii} waarden gemeenschappelijk kunnen hebben als die precies op een as liggen. ε mag ook niet gelijk 1 worden want dan houden we geen gebied mee over, dus $0 < \varepsilon < 1$.

Met deze nieuwe B'_{ii} bouwen we een B' en we zoeken een bijbehorende C'

$$B' = \begin{pmatrix} B'_{11} & 0 & 0 & 0 \\ 0 & B'_{22} & 0 & 0 \\ 0 & 0 & B'_{33} & 0 \\ 0 & 0 & 0 & B'_{44} \end{pmatrix} \quad C' = \begin{pmatrix} C'_{11} & C'_{12} & C'_{13} & C'_{14} \\ C'_{21} & C'_{22} & C'_{23} & C'_{24} \\ C'_{31} & C'_{32} & C'_{33} & C'_{34} \\ C'_{41} & C'_{42} & C'_{43} & C'_{44} \end{pmatrix}$$

die moeten voldoen aan

$$A_{ij} = B'_{ii}C'_{ij} - C'_{ij}B'_{jj}$$

zodat ook $A = B'C' - C'B'$. We kunnen deze vergelijking direct oplossen; de elementen van de blokjes worden gegeven door:

$$(C'_{ij})_{kl} = \begin{cases} 0 & k = l \text{ en } i = j \\ \frac{(A_{ij})_{kl}}{(B'_{ii})_{kk} - (B'_{jj})_{ll}} & k \neq l \text{ of } i \neq j \end{cases}$$

Hier hebben we alleen niet zoveel aan als we een goede afschatting van B' en C' willen krijgen; we zijn namelijk vooral geïnteresseerd in de normen van B'

en C' en niet in hun elementen. De afschatting die we kunnen krijgen door één van de stellingen uit het vorige hoofdstuk toe te passen zijn niet goed genoeg voor het resultaat dat we willen bewijzen. We passen daarom twee truukjes toe - een voor de diagonaalblokjes en een voor de niet-diagonaalblokjes - die ons direct iets over deze normen vertellen. Voor de diagonaalblokjes hebben we voor zekere $\beta_i \in \mathbb{C}$

$$\begin{aligned}
A_{ii} &= B'_{ii}C'_{ii} - C'_{ii}B'_{ii} \\
&= \left(\frac{1-\varepsilon}{2}B_{ii} + \beta_i I \right) C'_{ii} - C'_{ii} \left(\frac{1-\varepsilon}{2}B_{ii} + \beta_i I \right) \\
&= \frac{1-\varepsilon}{2} (B_{ii}C'_{ii} - C'_{ii}B_{ii}) + \beta_i C'_{ii} - \beta_i C'_{ii} \\
&= \frac{1-\varepsilon}{2} (B_{ii}C'_{ii} - C'_{ii}B_{ii})
\end{aligned}$$

de verschuiving heeft dus geen invloed op de commutatierelatie, maar de verkleining wel. Corrigeren we hiervoor door

$$C'_{ii} = \frac{2}{1-\varepsilon} C_{ii}$$

te nemen dan $A_{ii} = B'_{ii}C'_{ii} - C'_{ii}B'_{ii}$.

De uitdrukking voor de elementen van een niet-diagonaalblokjes

$$(C'_{ij})_{kl} = \frac{(A_{ij})_{kl}}{(B'_{ii})_{kk} - (B'_{jj})_{ll}}$$

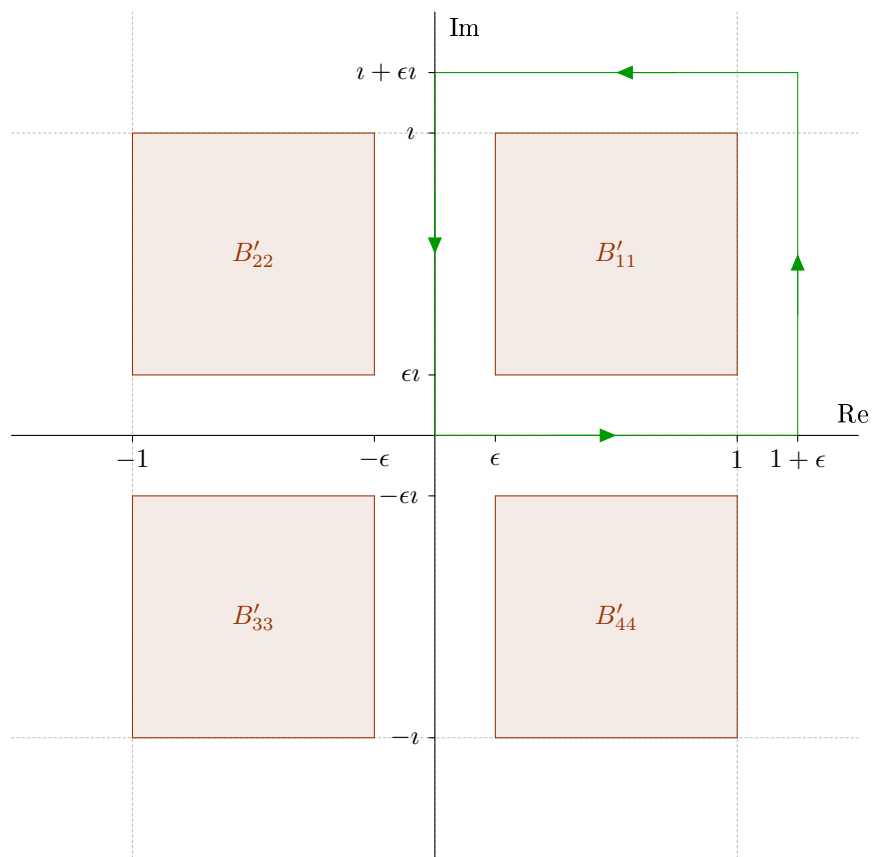
kunnen we ook schrijven als de complexe contourintegraal (herinner de formule van Cauchy: $f(x) = \frac{1}{2\pi i} \int_{\partial X} \frac{f(z)}{z-x} dz$ voor een contour ∂X die precies éénmaal tegen de klok in om het punt x heen wikkelt en een functie f die analytisch is binnen deze contour en continu op de contour):

$$(C'_{ij})_{kl} = \frac{1}{2\pi i} \int_{\partial D_{ii}} \frac{(A_{ij})_{kl}}{(z - (B'_{ii})_{kk})(z - (B'_{jj})_{ll})} dz$$

waarbij ∂D_{ii} een contour is (figuur 3.2) die precies om het gebied waar de waarden van B'_{ii} in liggen heen loopt (de waarden van B_{jj} liggen hier dus buiten). We kunnen de integralen die elk element $(C'_{ij})_{kl}$ geven samenvoegen tot één matrixintegraal die ons in één keer de hele matrix C'_{ij} geeft:

$$C'_{ij} = \frac{1}{2\pi i} \int_{\partial D_{ii}} (zI - B'_{ii})^{-1} A_{ij} (zI - B'_{jj})^{-1} dz$$

(De auteurs van het artikel gebruiken hier zelf het sterkere resultaat van Rosenblum [10] voor dat zegt dat deze formule ook geldt voor niet-diagonale B 's waarbij de eigenwaarden in plaats van de diagonaalwaarden in gescheiden gebieden liggen.) Nu kunnen we hier een mooie driehoeksongelijkheid voor integralen



Figuur 3.2: De contour ∂D_{11} (groen) en de vier gebieden waar de waarden van de B'_{ii} in liggen.

op loslaten waarmee we in één keer $\|C'_{ij}\|$ kunnen afschatten. Om dit goed te doen nemen we een vierkante contour ∂D_{ii} die samenvalt met de assen en zijde $1 + \varepsilon$ heeft (figuur 3.2). Dan liggen alle waarden in B'_{ii} in de contour en bovendien liggen alle waarden in B'_{ii} én B'_{jj} op een afstand van minstens ε van de contour af. Dan geldt

$$z - (B'_{ii})_{kk} \geq \varepsilon$$

voor alle i, k en $z \in \partial D_{ii}$ en dus

$$\|(zI - B'_{ii})^{-1}\| = \|(zI - B'_{ii})^{-1}\|_{\max} = \frac{1}{\|zI - B'_{ii}\|_{\max}} \leq \frac{1}{\varepsilon}$$

(nb: de matrix $zI - B'_{ii}$ is een diagonaalmatrix). De contour ∂D_{ii} heeft een lengte $4 + 4\varepsilon < 8$ en daarmee volgt

$$\begin{aligned} \|C'_{ij}\| &\leq \left| \frac{1}{2\pi i} \right| \left\| (zI - B'_{ii})^{-1} A_{ij} (zI - B'_{jj})^{-1} \right\| (4 + 4\varepsilon) \\ &\leq \frac{1}{2\pi} \left\| (zI - B'_{ii})^{-1} \right\| \|A_{ij}\| \left\| (zI - B'_{jj})^{-1} \right\| 8 \\ &\leq \frac{4}{\pi} \frac{1}{\varepsilon} \|A_{ij}\| \frac{1}{\varepsilon} \\ &\leq \frac{2}{\varepsilon^2} \|A_{ij}\| \end{aligned}$$

Nu kunnen we alle stukjes samenvoegen. We hebben een B' geconstrueerd met elementen in $[-1, 1] \times [-\varepsilon, \varepsilon]$ en een C' waarvoor $A = B'C' - C'B'$. De blokjes waar C' uit bestaat voldoen aan

$$\|C'_{ii}\| = \frac{2}{1 - \varepsilon} \|C_{ii}\| = \frac{2}{1 - \varepsilon} \lambda(A_{ii})$$

en

$$\|C'_{ij}\| \leq \frac{2}{\varepsilon^2} \|A_{ij}\| \leq \frac{2}{\varepsilon^2} \|A\| = \frac{2}{\varepsilon^2} \quad (i \neq j)$$

(stelling 2.28) en daarmee kunnen we de norm van deze blokmatrix afschatten (stelling 2.30):

$$\|C'\| \leq \left\| \begin{array}{cccc} \frac{2}{1-\varepsilon} \lambda(A_{11}) & \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} \\ \frac{2}{\varepsilon^2} & \frac{2}{1-\varepsilon} \lambda(A_{22}) & \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} \\ \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} & \frac{2}{1-\varepsilon} \lambda(A_{33}) & \frac{2}{\varepsilon^2} \\ \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} & \frac{2}{\varepsilon^2} & \frac{2}{1-\varepsilon} \lambda(A_{44}) \end{array} \right\|$$

en dus (stelling 2.23):

$$\lambda(A) \leq \|C'\| \leq \frac{2}{1 - \varepsilon} \max_{1 \leq i \leq 4} \lambda(A_{ii}) + \frac{6}{\varepsilon^2}$$

Met

$$\lambda(A_{ii}) \leq \lambda(4^{n-1}) \|A_{ii}\| \leq \lambda(4^{n-1})$$

(stelling 2.28) volgt ook

$$\lambda(A) \leq \frac{2}{1-\varepsilon} \lambda(4^{n-1}) + \frac{6}{\varepsilon^2}$$

en omdat A willekeurig was ook

$$\begin{aligned} \lambda(4^n) &= \sup_A \min_{B, C} \|C\| \\ &\leq \frac{2}{1-\varepsilon} \lambda(4^{n-1}) + \frac{6}{\varepsilon^2} \end{aligned}$$

□

Laten we eens kijken wat deze stelling ons vertelt. Voor 4×4 matrices ($n = 1$) zijn de blokmatrices in bovenstaand verhaal gewoon de elementen en kunnen we C hol nemen (vandaar $\lambda(1) = 0$), waarmee

$$\lambda(4) \leq \frac{6}{\varepsilon^2}$$

In feite niets nieuws, aangezien we met stelling 2.36 al wisten dat $\lambda(4) \leq \frac{1}{2}(4-1)^2 = 4\frac{1}{2}$. Voor $n = 2$ krijgen we

$$\lambda(16) \leq \frac{2}{1-\varepsilon} \lambda(4) + \frac{6}{\varepsilon^2} = \left(\frac{2}{1-\varepsilon} + 1 \right) \frac{6}{\varepsilon^2}$$

waarbij we ε nog vrij kunnen kiezen tussen 0 en 1. Kiezen we ε optimaal dan $\varepsilon \approx 0.697$ en

$$\lambda(16) \leq 93.9$$

een duidelijke verbetering van het eerste resultaat uit het vorige hoofdstuk ($\frac{1}{2}15^2 = 112.5$) maar nog niet van de tweede ($15^{3/2}/\sqrt{2} \approx 41$). Gaan we nog een stapje verder dan

$$\lambda(64) \leq \frac{2}{1-\varepsilon} \lambda(16) + \frac{6}{\varepsilon^2} = \left(\left(\frac{2}{1-\varepsilon} \right)^2 + \frac{2}{1-\varepsilon} + 1 \right) \frac{6}{\varepsilon^2} \leq 499.0$$

voor optimale $\varepsilon \approx 0.535$. Nog steeds is dit niet beter dan $63^{3/2}/\sqrt{2} \approx 354$, al wordt het verschil wel kleiner. Merk op dat de optimale ε veranderd is - we kunnen ε dus het beste zo lang mogelijk vrij laten om een zo goed mogelijke grens te bereiken. We zullen zo zien dat de grens voor grote n wel degelijk beter dan de resultaten uit het vorige hoofdstuk zal worden. Om die afchatting te maken hebben we het volgende lemma nodig:

Lemma 3.4. *Voor $n \geq 2$ geldt*

$$\left(1 - \frac{1}{n} \right)^n \geq \frac{1}{4}$$

Bewijs. In de bekende limiet $n \rightarrow \infty$ gaat de linkerhand van de deze vergelijking naar $e^{-1} \approx 0.3$. Het is mogelijk om aan te tonen dat deze uitdrukking voor de exponentiële functie altijd stijgt, maar wij zullen ons hier beperken tot een elementair bewijs voor dit bijzondere geval.

Voor $n = 2$ hebben we gelijkheid in de geclaimde bewering. Noem $f(n) = (1 - \frac{1}{n})^n$ en beschouw zijn logaritme

$$g(n) = \ln f(n) = n \ln \left(1 - \frac{1}{n}\right)$$

We tonen aan dat de afgeleide $g'(n) > 0$ waaruit volgt dat f stijgend is en de ongelijkheid blijft gelden voor $n > 2$. Bekijk dus

$$\begin{aligned} g'(n) &= \ln \left(1 - \frac{1}{n}\right) + n \cdot \frac{1}{1 - \frac{1}{n}} \cdot \frac{1}{n^2} \\ &= \ln \left(\frac{n-1}{n}\right) + \frac{1}{n-1} \end{aligned}$$

De gezochte ongelijkheid kunnen we een beetje omschrijven:

$$\begin{aligned} &g'(n) > 0 \\ \iff &\frac{1}{n-1} > -\ln \left(\frac{n-1}{n}\right) \\ \iff &\frac{1}{n-1} > \ln \left(\frac{n}{n-1}\right) \\ \iff &\exp \left(\frac{1}{n-1}\right) > \frac{n}{n-1} \\ \iff &(n-1) \exp \left(\frac{1}{n-1}\right) > n \end{aligned}$$

Nu weten we dat alle afgeleiden van $\exp(x)$ positief zijn voor positieve x , en dus $\exp(x) > 1 + x$ (denk b.v. aan de Taylorreeks voor $\exp(x)$). Daarmee

$$(n-1) \exp \left(\frac{1}{n-1}\right) > (n-1) \left(1 + \frac{1}{n-1}\right) = (n-1) + 1 = n$$

□

Nu kunnen we aan de slag:

Stelling 3.5 (Commutatorstelling $\mathcal{O}((\log_4 m)^2 \sqrt{m})$). *Voor iedere spoorloze matrix $A \in \mathbb{C}^{m \times m}$ met $m \geq 4$ bestaan er matrices $B, C \in \mathbb{C}^{m \times m}$ zodat $A = BC - CB$ en $\|B\| \|C\| \leq 48(1 + \log_4 m)^2 \sqrt{m} \|A\|$. Algemener bestaan er voor iedere spoorloze matrix $A \in \mathbb{C}^{m \times m}$ matrices $B, C \in \mathbb{C}^{m \times m}$ zodat $A = BC - CB$ en*

$$\|B\| \|C\| \leq \max \left(384, 48(1 + \log_4 m)^2 \sqrt{m}\right) \|A\|$$

Bewijs. Als we de recurrente betrekking van lemma 3.3 herhaaldelijk invullen vinden we

$$\begin{aligned}
\lambda(4^n) &\leq \frac{2}{1-\varepsilon}\lambda(4^{n-1}) + \frac{6}{\varepsilon^2} \\
&\leq \frac{2}{1-\varepsilon}\left(\frac{2}{1-\varepsilon}\lambda(4^{n-2}) + \frac{6}{\varepsilon^2}\right) + \frac{6}{\varepsilon^2} \\
&\leq \frac{2}{1-\varepsilon}\left(\frac{2}{1-\varepsilon}\left(\frac{2}{1-\varepsilon}\lambda(4^{n-3}) + \frac{6}{\varepsilon^2}\right) + \frac{6}{\varepsilon^2}\right) + \frac{6}{\varepsilon^2} \\
&\leq \dots \\
&\leq \frac{6}{\varepsilon^2} \sum_{i=0}^{n-1} \left(\frac{2}{1-\varepsilon}\right)^i
\end{aligned}$$

(herinner dat $\lambda(4^0) = 0$). De geometrische reeks kunnen we uitwerken (herinner dat $0 < \varepsilon < 1$):

$$\begin{aligned}
\sum_{i=0}^{n-1} \left(\frac{2}{1-\varepsilon}\right)^i &= \frac{1 - \left(\frac{2}{1-\varepsilon}\right)^n}{1 - \left(\frac{2}{1-\varepsilon}\right)} \\
&= (1-\varepsilon) \cdot \frac{1 - \left(\frac{2}{1-\varepsilon}\right)^n}{-1-\varepsilon} \\
&< \frac{1 - \left(\frac{2}{1-\varepsilon}\right)^n}{-1-\varepsilon} \\
&= \frac{\left(\frac{2}{1-\varepsilon}\right)^n - 1}{1+\varepsilon} \\
&< \left(\frac{2}{1-\varepsilon}\right)^n - 1 \\
&< \left(\frac{2}{1-\varepsilon}\right)^n
\end{aligned}$$

Nemen we $\varepsilon = \frac{1}{n}$ dan volgt

$$\begin{aligned}
\lambda(4^n) &< \frac{6}{\varepsilon^2} \left(\frac{2}{1-\varepsilon}\right)^n \\
&= 6n^2 2^n \left(\frac{1}{1-\frac{1}{n}}\right)^n \\
&\leq 24n^2 2^n
\end{aligned}$$

voor $n \geq 2$. Hierbij gebruiken we dat

$$\left(1 - \frac{1}{n}\right)^n \geq \frac{1}{4}$$

vanwege lemma 3.4.

Voor $m \geq 4$ kunnen we n het kleinste gehele getal groter dan $\log_4 m$ nemen, dan $n \geq 2$ en $n \leq \log_4 m + 1$ en dus (lemma 3.1):

$$\begin{aligned}
\lambda(m) &\leq \lambda(4^n) \\
&\leq 24n^2 2^n \\
&\leq 24(1 + \log_4 m)^2 2^{1 + \log_4 m} \\
&= 48(1 + \log_4 m)^2 2^{\log_4 m} \\
&= 48(1 + \log_4 m)^2 2^{\frac{1}{2} \log_2 m} \\
&= 48(1 + \log_4 m)^2 \sqrt{m}
\end{aligned}$$

De gevallen $m < 4$ kunnen we afschatten met $\lambda(4)$, invullen geeft:

$$\lambda(m) \leq \lambda(4) \leq 48 \cdot 4 \cdot 2 = 384$$

□

Voor grote m (waar we voornamelijk in geïnteresseerd zijn) groeit deze nieuwe grens langzamer dan die van het vorige hoofdstuk, i.e.

$$\mathcal{O}\left((\log m)^2 \sqrt{m}\right) < \mathcal{O}\left(m^{3/2}\right)$$

immers $\mathcal{O}(\log m) < \mathcal{O}(\sqrt{m})$ want $\mathcal{O}(m) < \mathcal{O}(\frac{1}{2} \exp m)$. Die laatste ongelijkheid laat duidelijk zien dat we hier echt een betere grens te pakken hebben.

De auteurs bewijzen in hun artikel een grens $\mathcal{O}\left((\log m)^3 \sqrt{m}\right)$ (derde in plaats van tweede macht van $\log m$) door de recursierelatie op een andere manier af te schatten; dit heeft verder geen invloed op de rest van het bewijs.

3.3 2×2 blokmatrixes

De constructie uit de vorige paragraaf is natuurlijk prachtig, maar kunnen we deze nog op een of andere manier verbeteren? In het bijzonder - zo vragen de auteurs zich in het artikel af - kunnen we misschien de 2 uit $\frac{2}{1-\varepsilon}$ wegmoffelen? Dan valt namelijk de 2^n weg uit de afchatting van $\lambda(4^n)$ en houden we een grens in $\log_4 m$ over.

De factor $\frac{2}{1-\varepsilon}$ is een noodzakelijk gevolg van de constructie die in lemma 3.3 gebruikt wordt: omdat we de B_{ii} met een factor $< \frac{1}{2}$ verkleinen wordt de norm van de C_{ii} uiteindelijk > 2 keer zo groot. Om deze vergroting kleiner te maken zullen we dus een nieuwe constructie moeten bedenken waarbij de B_{ii} niet al te klein worden. Maar daar hebben we dan wel meer ruimte voor nodig in het complexe eenheidsvierkant. Bovendien zullen we altijd een verkleining van de B_{ii} houden, simpelweg omdat we meerdere B'_{ii} in dezelfde oppervlakte als één B_{ii} willen proppen. De auteurs proberen dit op te lossen dit op door

te kijken naar 2×2 blokmatrixes zodat er slechts twee in plaats van vier B in het eenheidsvierkant gepast hoeven te worden. Daarnaast gebruiken ze een asymmetrische constructie, waarbij één van de B_{ii} willekeurig klein wordt en de andere met een factor willekeurig dicht bij 1 verkleind wordt. Dit resulteert in het volgende lemma, dat qua bewijs sterk lijkt op lemma 3.3:

Lemma 3.6. *Zij A een holle $2m \times 2m$ blokmatrix met $\|A\| \leq 1$ bestaande uit $m \times m$ deelmatrices A_{ij} :*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

dan geldt voor alle $0 < \delta < \frac{1}{2}$

$$\lambda(A) \leq \max \left\{ \frac{\lambda(A_{11})}{\delta}, \frac{\lambda(A_{22})}{1-2\delta} \right\} + \frac{2}{\delta^2}$$

Bewijs. Net als in het bewijs van lemma 3.3 nemen we weer de 'beste' matrices B_{11}, B_{22} met elementen in $[-1, 1] \times [-\iota, \iota]$ en C_{11}, C_{22} waarvoor $A_{11} = [B_{11}, C_{11}]$, $A_{22} = [B_{22}, C_{22}]$ en

$$\begin{aligned} \|C_{11}\| &= \lambda(A_{11}) \\ \|C_{22}\| &= \lambda(A_{22}) \end{aligned}$$

Om de elementen van B_{11} en B_{22} te scheiden verkleinen en verschuiven we ze elk naar hun eigen stukje van het 'eenheidsvierkant' $[-1, 1] \times [-\iota, \iota]$ (zie figuur 3.3):

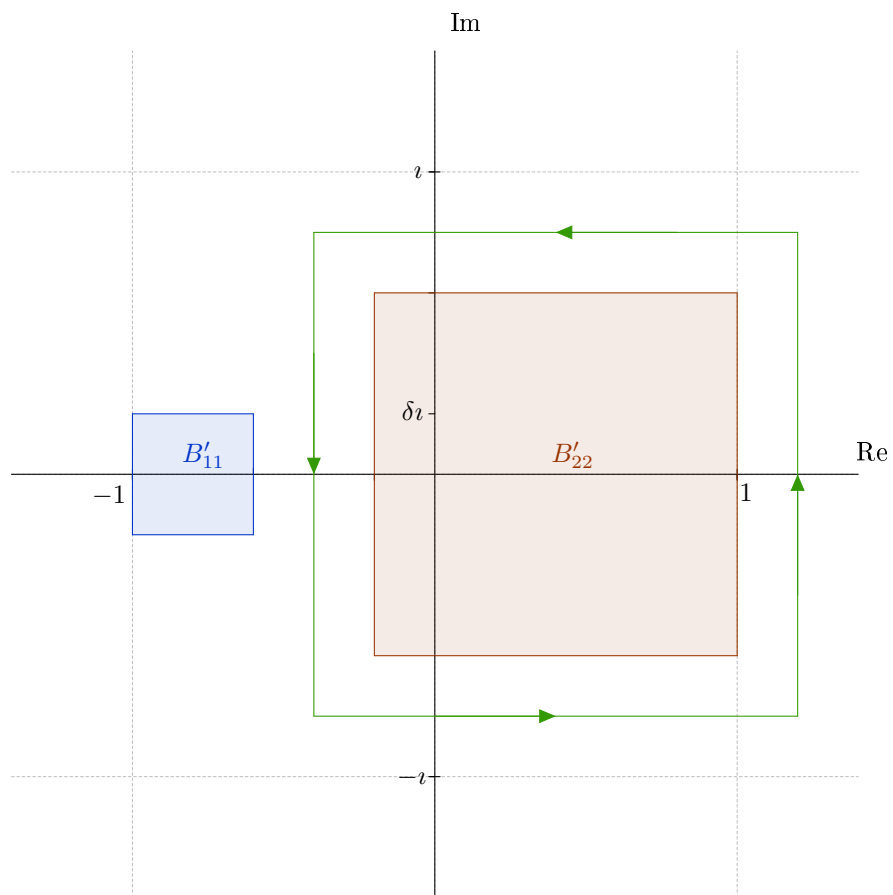
$$\begin{aligned} B'_{11} &= \delta B_{11} + (-1 + \delta)I \\ B'_{22} &= (1 - 2\delta)B_{22} + 2\delta I \end{aligned}$$

waarbij $\delta > 0$ een vergelijkbare rol als ε in lemma 3.3 speelt. Zoals in figuur 3.3 te zien is worden de gebieden met de waarden van B'_{11} en B'_{22} elk gecentreerd op de reële as en tegen de linker- en respectievelijk rechterzijkant van $[-1, 1] \times [-\iota, \iota]$ aangeplakt. Bekijken we het eenheidsvierkant langs het reële interval $[-1, 1]$ dan wordt dit opgedeeld in drie stukken: $[-1, -1 + 2\delta]$ met waarden van B'_{11} , $[-1 + 4\delta, 1]$ met waarden van B'_{22} en het tussenstuk $(-1 + 2\delta, -1 + 4\delta)$ zonder waarden van B'_{11} of B'_{22} . Voor $\delta = \frac{1}{2}$ verdwijnt B'_{22} , dus we eisen $\delta < \frac{1}{2}$.

Merk op dat deze indeling asymmetrisch is: voor kleinere δ wordt het gebied met waarden van B'_{11} kleiner en dat van B'_{22} juist groter; bovendien verschuift het tussengebied zonder waarden naar links. Dit is de essentie van deze constructie: als we beide B_{ii} met een factor δ zouden vermenigvuldigen zodat de gebieden op symmetrische verkleind worden dan vinden we dezelfde constructie als in lemma 3.3 terug met $\delta \triangleq \frac{1-\varepsilon}{2}$ (maar dan met twee i.p.v. vier matrices). Juist omdat we nu slechts twee matrices hebben kunnen de ruimte die overblijft gebruiken om B'_{22} groter te laten worden en zo een beter resultaat te verkrijgen.

Nu de waarden van de B 's gescheiden zijn kunnen we C'_{12} en C'_{21} nemen die voldoen aan

$$A = B'_{ii}C'_{ij} - C'_{ij}B'_{jj},$$



Figuur 3.3: De contour ∂D_{22} (groen) en de gebieden waar de waarden van B'_{11} en B'_{22} in liggen.

bijvoorbeeld door ze te schrijven als de complexe contourintegraal

$$C'_{ij} = \frac{1}{2\pi i} \int_{\partial D_{ii}} (zI - B'_{ii})^{-1} A_{ij} (zI - B'_{jj})^{-1} dz$$

waarbij ∂D_{ii} in dit geval de vierkante contour is die precies tussen beide gebieden in over de lijn $\Re\{z\} = -1 + 3\delta$ tegen de klok in om het gebied met de waarden van B'_{ii} heen loopt, zodat de afstand tot beide gebieden altijd minstens δ is (zie figuur 3.3), i.e.

$$z - (B'_{ii})_{kk} \geq \delta$$

voor alle i, k en $z \in \partial D_{ii}$ en dus

$$\|(zI - B'_{ii})^{-1}\| = \|(zI - B'_{ii})^{-1}\|_{\max} = \frac{1}{\|zI - B'_{ii}\|_{\max}} \leq \frac{1}{\delta}$$

De contour ∂D_{11} om het linkergebied heeft lengte $12\delta < 6 < 8$ en de contour ∂D_{22} om het rechter gebied heeft een lengte $4(2 - 2\delta) < 8$ (zie figuur 3.3) en daarmee volgt

$$\begin{aligned} \|C'_{ij}\| &\leq \left| \frac{1}{2\pi i} \right| \|(zI - B'_{ii})^{-1} A_{ij} (zI - B'_{jj})^{-1}\| 8 \\ &\leq \frac{1}{2\pi} \|(zI - B'_{ii})^{-1}\| \|A_{ij}\| \|(zI - B'_{jj})^{-1}\| 8 \\ &\leq \frac{4}{\pi} \frac{1}{\delta} \|A_{ij}\| \frac{1}{\delta} \\ &\leq \frac{2}{\delta^2} \|A_{ij}\| \\ &\leq \frac{2}{\delta^2} \end{aligned}$$

De verkleining van de B_{ii} moeten we compenseren door de C_{ii} te vergroten; we nemen dus

$$\begin{aligned} C'_{11} &= \frac{1}{\delta} C_{11} \\ C'_{22} &= \frac{1}{1 - 2\delta} C_{22} \end{aligned}$$

waarmee

$$\begin{aligned} \|C'_{11}\| &= \frac{1}{\delta} \|C_{11}\| = \frac{1}{\delta} \lambda(A_{11}) \\ \|C'_{22}\| &= \frac{1}{1 - 2\delta} \|C_{22}\| = \frac{1}{1 - 2\delta} \lambda(A_{22}) \end{aligned}$$

Nu kunnen we de bouwstenen bijeenrapen in twee grote matrices

$$B' = \begin{pmatrix} B'_{11} & 0 \\ 0 & B'_{22} \end{pmatrix} \quad C' = \begin{pmatrix} C'_{11} & C'_{12} \\ C'_{21} & C'_{22} \end{pmatrix}$$

die voldoen aan $A = B'C' - C'B'$ en B' is diagonaal met elementen in $[-1, 1] \times [-\iota, \iota]$. De norm van C' kunnen we weer afschatten met de blokjes waar hij uit op is gebouwd (stelling 2.30):

$$\|C'\| \leq \left\| \begin{pmatrix} \frac{1}{\delta} \lambda(A_{11}) & \frac{2}{\delta^2} \\ \frac{2}{\delta^2} & \frac{1}{1-2\delta} \lambda(A_{22}) \end{pmatrix} \right\|$$

en dus (stelling 2.23):

$$\lambda(A) \leq \max \left\{ \frac{\lambda(A_{11})}{\delta}, \frac{\lambda(A_{22})}{1-2\delta} \right\} + \frac{2}{\delta^2}$$

□

Op het eerste gezicht lijkt dit een verbetering van lemma 3.3: de vervelende 2 uit $\frac{2}{1-\varepsilon}$ lijkt immers verdwenen. Helaas kunnen we niet hetzelfde truukje uithalen als bij lemma 3.3, want

$$\begin{aligned} \max \left\{ \frac{\lambda(A_{11})}{\delta}, \frac{\lambda(A_{22})}{1-2\delta} \right\} &\leq \max \left\{ \frac{\lambda(4^n)}{\delta}, \frac{\lambda(4^n)}{1-2\delta} \right\} \\ &= \lambda(4^n) \max \left\{ \frac{1}{\delta}, \frac{1}{1-2\delta} \right\} \\ &\geq \lambda(4^n) \inf_{0 < \delta < \frac{1}{2}} \left\{ \frac{1}{\delta}, \frac{1}{1-2\delta} \right\} \\ &= 3\lambda(4^n) \end{aligned}$$

immers $\frac{1}{\delta}$ daalt in δ , $\frac{1}{1-2\delta}$ stijgt en $\delta = 1 - 2\delta \implies \delta = \frac{1}{3}$. Deze situatie is dus nog slechter dan bij lemma 3.5: zelfs als we δ optimaal kiezen hebben nog een vervelende factor 3 in plaats van 2.

Toch is dit resultaat niet geheel nutteloos. De kracht van lemma 3.6 zit hem in de asymmetrie: als we van één van de $\lambda(A_{11})$, $\lambda(A_{22})$ weten dat hij flink kleiner is dan de andere, dan kunnen we δ hier op afstemmen en zo met lemma 3.6 aantonen dat $\lambda(A)$ niet veel groter (als in $1 + \varepsilon$) dan de grootste $\lambda(A_{11})$, $\lambda(A_{22})$ zal zijn. Deze bewering maken we precies in de volgende twee lemma's:

Gevolg 3.7. *Zij A een holle $2m \times 2m$ blokmatrix met $\|A\| \leq 1$ bestaande uit $m \times m$ deelmatrices A_{ij}*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

waarvoor de bovengrenzen $\lambda(A_{11}) \leq c_1$ en $\lambda(A_{22}) \leq c_2$ gelden met $c_1 < \frac{1}{4}c_2$. Dan

$$\lambda(A) \leq \left(\frac{1}{1 - 2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{c_1} \right) c_2$$

Bewijs. Neem $\delta = \sqrt{\frac{c_1}{c_2}}$ dan volgt met $\frac{c_1}{c_2} < \frac{1}{4}$ dat $\delta < \frac{1}{2}$ en ook $\delta > \frac{c_1}{c_2}$ waarmee

$$\frac{c_1}{\delta} < \frac{c_1}{c_1/c_2} = c_2 < \frac{c_2}{1-2\delta}$$

zodat

$$\max \left\{ \frac{\lambda(A_{11})}{\delta}, \frac{\lambda(A_{22})}{1-2\delta} \right\} \leq \max \left\{ \frac{c_1}{\delta}, \frac{c_2}{1-2\delta} \right\} = \frac{c_2}{1-2\delta}$$

Met lemma 3.6 volgt dat

$$\begin{aligned} \lambda(A) &\leq \frac{c_2}{1-2\delta} + \frac{2}{\delta^2} \\ &= \frac{c_2}{1-2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{\left(\sqrt{\frac{c_1}{c_2}}\right)^2} \\ &= \frac{c_2}{1-2\sqrt{\frac{c_1}{c_2}}} + 2\frac{c_2}{c_1} \\ &= \left(\frac{1}{1-2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{c_1} \right) c_2 \end{aligned}$$

□

De auteurs zijn in hun artikel iets te voorbarig, zij claimen daar dat

$$\left(\frac{1}{1-2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{c_1} \right) c_2 \leq \left(1 + 2\sqrt{\frac{c_1}{c_2}} + \frac{2}{c_1} \right) c_2$$

wat niet altijd opgaat, b.v. voor $\frac{c_1}{c_2} = \frac{1}{9}$ geldt

$$3 = \frac{1}{1-\frac{2}{3}} = \frac{1}{1-2\sqrt{\frac{c_1}{c_2}}} \not\leq 1 + 2\sqrt{\frac{c_1}{c_2}} = 1 + \frac{2}{3}$$

Ook een iets vrijere interpretatie van de claim in het artikel (de auteurs zijn vaag over de constante waar ik hier 2 voor heb genomen) valt niet goed te praten, want er bestaat geen enkele constante K waarvoor

$$\left(\frac{1}{1-2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{c_1} \right) c_2 \leq \left(1 + K\sqrt{\frac{c_1}{c_2}} + \frac{2}{c_1} \right) c_2$$

altijd opgaat, we kunnen immers c_1 en c_2 zo kiezen dat $\frac{c_1}{c_2}$ willekeurig dicht bij $\frac{1}{4}$ komt (en dus $2\sqrt{\frac{c_1}{c_2}}$ willekeurig dicht bij 1), waarmee $\left(1 - 2\sqrt{\frac{c_1}{c_2}}\right)^{-1}$ ontploft terwijl $1 + K\sqrt{\frac{c_1}{c_2}}$ begrensd blijft.

De reden dat dit zo (slordig) in het artikel staat is waarschijnlijk dat we dit lemma nu gaan gebruiken met c_1 en c_2 waarvoor $\frac{c_1}{c_2}$ klein is (b.v. $< \frac{1}{256}$) en in dat geval kun je zo'n afchatting natuurlijk wel maken.

We zullen straks gaan werken met grenzen van de vorm $Kn^p 2^{\frac{n}{q}}$, dus daar gaan we nu lemma 3.7 op loslaten:

Lemma 3.8. *Zij A een holle $2m \times 2m$ matrix met $\|A\| \leq 1$*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

waarvoor

$$\lambda(A_{11}) \leq Kn^p 2^{\frac{n}{q}}$$

voor zekere $n \geq 4$ (niet noodzakelijk aan m gerelateerd), $p \geq 2$, $q \geq 1$ en constante $K > 1$ onafhankelijk van n . Dan

$$\lambda(A) \leq \left(1 + \frac{4}{n^2}\right) \max \left\{ Kn^{p+4} 2^{\frac{n}{q}}, \lambda(A_{22}) \right\}$$

Bewijs. We gaan lemma 3.7 toepassen met geschikte bovengrenzen. Neem voor c_2 een vergelijkbare bovengrens met een hogere macht van n , dus:

$$\begin{aligned} c_1 &= Kn^p 2^{\frac{n}{q}} \\ c_2 &= \max \left\{ Kn^{p+4} 2^{\frac{n}{q}}, \lambda(A_{22}) \right\} \end{aligned}$$

zodat $\lambda(A_{11}) \leq c_1$, $\lambda(A_{22}) \leq c_2$ en

$$\frac{c_1}{c_2} \leq \frac{Kn^p 2^{\frac{n}{q}}}{Kn^{p+4} 2^{\frac{n}{q}}} = \frac{1}{n^4} \leq \frac{1}{256} < \frac{1}{4}$$

(immers $n \geq 4$). Merk op

$$\sqrt{\frac{c_1}{c_2}} \leq \frac{1}{n^2}$$

$$1 - 2\sqrt{\frac{c_1}{c_2}} \geq 1 - \frac{2}{n^2}$$

$$\left(1 - \frac{2}{n^2}\right) \left(1 + \frac{3}{n^2}\right) = 1 + \frac{1}{n^2} - \frac{6}{n^4} = 1 + \frac{1}{n^2} \left(1 - \frac{6}{n^2}\right) > 1$$

voor $n \geq 4$ en dus

$$\frac{1}{1 - 2\sqrt{\frac{c_1}{c_2}}} \leq \frac{1}{1 - \frac{2}{n^2}} \leq 1 + \frac{3}{n^2}$$

Bovendien ook (n.b. $p \geq 2$):

$$\frac{2}{c_1} = \frac{2}{Kn^p 2^{\frac{n}{q}}} \leq \frac{1}{n^2}$$

Met lemma 3.7 en bovenstaande ongelijkheden volgt

$$\begin{aligned}
\lambda(A) &\leq \left(\frac{1}{1 - 2\sqrt{\frac{c_1}{c_2}}} + \frac{2}{c_1} \right) c_2 \\
&\leq \left(1 + \frac{3}{n^2} + \frac{1}{n^2} \right) c_2 \\
&= \left(1 + \frac{4}{n^2} \right) c_2 \\
&= \left(1 + \frac{4}{n^2} \right) \max \left\{ K n^{p+4} 2^{\frac{n}{q}}, \lambda(A_{22}) \right\}
\end{aligned}$$

□

De reden dat we hier $p+4$ nemen in plaats van bijvoorbeeld $p+2$ (wat ook prima aan de voorwaarden van lemma 3.7 voor c_1 en c_2 zou voldoen) is dat dit het kwadraat in $(1 + \frac{4}{n^2})$ oplevert en dat zullen we straks nodig hebben.

3.4 Een stelling van Bourgain en Tzafriri

Nu hebben we een mooi lemma dat waarmee we de norm van een 2×2 blokmatrix met één 'klein' diagonaalblokje kunnen afschatten, maar voor we deze kunnen toepassen hebben we eerst zo'n klein diagonaalblokje nodig. In het bewijs van het artikel gebruiken de auteurs daar een sterke stelling van Bourgain en Tzafriri uit [14] voor die iets zegt over het bestaan van een bepaalde centrale deelmatrix (ons 'diagonaalblokje'). We zullen deze stelling ook gebruiken (zonder deze te bewijzen):

Stelling 3.9 (Bourgain-Tzafriri). *Er bestaat een constante $K_{BT} > 0$ zodat iedere holle $m \times m$ matrix A voor alle $\varepsilon > 0$ een centrale deelmatrix $A_{\sigma \times \sigma}$ bevat van afmeting $[\varepsilon^2 m \times \varepsilon^2 m]$ en met norm $\|A_{\sigma \times \sigma}\| \leq \varepsilon K_{BT} \|A\|$.*

Een centrale deelmatrix $A_{\sigma \times \sigma}$ is een deelmatrix die wordt gevormd door dezelfde rijen als kolommen van A te selecteren; de diagonaal van $A_{\sigma \times \sigma}$ is dus een deelverzameling van de diagonaal van A . Een deelmatrix hoeft niet per se te bestaan uit rijen en kolommen die naast elkaar liggen; er mogen ook rijen en kolommen worden 'overgeslagen'. Preciezer geformuleerd: voor een deelverzameling $\sigma \subseteq \{1, 2, \dots, m\}$ bestaat de centrale deelmatrix $A_{\sigma \times \sigma}$ uit de elementen a_{ij} van A waarvoor $(i, j) \in \sigma \times \sigma$. Twee deelmatrices kunnen we disjunct noemen als ze geen enkele dezelfde rij of kolom uit A bevatten, dus als de bijbehorende σ 's disjunct zijn. Later zullen we onze matrix A zo herschikken dat de centrale deelmatrices die we hier vinden de gezochte diagonaalblokjes vormen.

Stelling 3.9 zullen we gebruiken in de vorm van het volgende gevolg:

Gevolg 3.10. Zij A een holle $2 \cdot 4^n \times 2 \cdot 4^n$ matrix met $\|A\| = 1$ en zij $0 \leq l \leq n$, dan bestaan er 4^l disjuncte centrale deelmatrices $A_{\sigma_i \times \sigma_i}$ ($1 \leq i \leq 4^l$) van afmeting $4^{n-l} \times 4^{n-l}$ met norm $\|A_{\sigma_i \times \sigma_i}\| \leq K_{BT} 2^{-l}$.

Bewijs. Neem $\varepsilon = \frac{2^{-l}}{\sqrt{2}}$ dan volgt met stelling 3.9 dat er een centrale $4^{n-l} \times 4^{n-l}$ deelmatrix $A_{\sigma_1 \times \sigma_1}$ bestaat met

$$\|A_{\sigma_1 \times \sigma_1}\| \leq \varepsilon K_{BT} \|A\| = \varepsilon K_{BT} = \frac{2^{-l}}{\sqrt{2}} K_{BT} < 2^{-l} K_{BT}$$

Laten we nu de rijen en kolommen van $A_{\sigma_1 \times \sigma_1}$ weg uit A dan krijgen we een nieuwe holle matrix $A_{\rho_1 \times \rho_1}$ (waarbij $\rho_1 = \{1, 2, \dots, 2 \cdot 4^n\} \setminus \sigma_1$) van afmeting $2 \cdot 4^n - 4^{n-l}$ en norm $\|A_{\rho_1 \times \rho_1}\| \leq \|A\| = 1$ (stelling 2.28). Op deze matrix kunnen we opnieuw stelling 3.9 toepassen en dit kunnen we zo op recursieve wijze blijven doen tot we 4^l matrices hebben. Voor $1 \leq i \leq 4^l$ beschouwen we dus de matrix $A_{\rho_{i-1} \times \rho_{i-1}}$ de we over houden als we de al gevonden matrices uit A weglaten:

$$\rho_{i-1} = \{1, 2, \dots, 2 \cdot 4^n\} \setminus \bigcup_{k=1}^{i-1} \sigma_k$$

De (vierkante) matrix $A_{\rho_{i-1} \times \rho_{i-1}}$ heeft dus afmeting

$$2 \cdot 4^n - (i-1) 4^{n-l} = 4^n (2 - (i-1) 4^{-l})$$

en norm $\|A_{\rho_{i-1} \times \rho_{i-1}}\| \leq \|A\| = 1$ (stelling 2.28). Passen we stelling 3.9 toe op $A_{\rho_{i-1} \times \rho_{i-1}}$ met

$$\varepsilon = \frac{2^{-l}}{\sqrt{2 - (i-1) 4^{-l}}} < 2^{-l}$$

(immers $i \leq 4^l$) dan vinden we een centrale $4^{n-l} \times 4^{n-l}$ deelmatrix $A_{\sigma_i \times \sigma_i}$ in $A_{\rho_{i-1} \times \rho_{i-1}}$ (dus in A en disjunct met alle $A_{\sigma_k \times \sigma_k}$, $k < i$) met norm:

$$\|A_{\sigma_i \times \sigma_i}\| \leq \varepsilon K_{BT} \|A_{\rho_{i-1} \times \rho_{i-1}}\| \leq \varepsilon K_{BT} < 2^{-l} K_{BT}$$

□

De oplettende lezer merkt misschien op dat we steeds nog één extra matrix kunnen vinden waarvoor het nog steeds goed gaat (in dat geval $\varepsilon = 2^{-l}$); in plaats van 4^l deelmatrices zou in bovenstaand gevolg dus ook $4^l + 1$ kunnen staan. We zullen dit echter niet nodig hebben.

Voorbeeld 3.11. Stel A is een holle 32×32 matrix met $\|A\| = 1$, dan zegt gevolg 3.10 ($n = 2$) dat er één centrale 16×16 deelmatrix is met norm niet groter dan K_{BT} ($l = 0$), dat er vier disjuncte centrale 4×4 deelmatrices zijn die elk een norm hebben niet groter dan $\frac{1}{2} K_{BT}$ ($l = 1$) en zestien disjuncte centrale 1×1 deelmatrices (elementen) met normen (absolute waarden) niet groter dan $\frac{1}{4} K_{BT}$ ($l = 2$). In elk van deze drie gevallen wordt precies de helft van de diagonaal (16 elementen) van A bedekt door de verzameling van centrale deelmatrices, maar in het algemeen zal dit niet steeds dezelfde helft zijn.

De stelling van Bourgain en Tzafriri kunnen we nu gebruiken om het gezochte kleine diagonaalblokje te vinden. Hierbij betekent klein: beter dan de beste grens voor $\lambda(4^n)$ die we tot nu toe hadden (i.e. $\lambda(4^n) \leq Kn^2 2^n$, lemma 3.5). We zullen dit in een algemene vorm doen zodat we de stelling straks herhaaldelijk kunnen toepassen om steeds betere grenzen te vinden.

Lemma 3.12. *Neem aan dat voor gehele $n \geq 1$, $p \geq 2$, $q \geq 1$ en constante $K_q > 1$ die niet van n afhangt geldt*

$$\lambda(4^n) \leq K_q n^p 2^{\frac{n}{q}}$$

en zij A een holle $2 \cdot 4^n \times 2 \cdot 4^n$ matrix met $n = l(q+1)$ voor $l \geq 2$ geheel en $\|A\| = 1$. Dan bevat A een centrale $4^n \times 4^n$ deelmatrix A' waarvoor

$$\lambda(A') \leq K n^p 2^{\frac{n}{q+1}}$$

met $K = 4K_{BT}K_q + 24$.

Bewijs. Omdat $q+1 \geq 2$ geldt $0 \leq l \leq n$ voor $l = \frac{n}{q+1}$ (we zullen straks zien waarom juist deze keuze voor l handig is). Wegens gevolg 3.10 bestaan nu er 4^l disjuncte centrale deelmatrices $A_{ii}^l = A_{\sigma_i^l \times \sigma_i^l}$ van afmeting 4^{n-l} waarvoor

$$\|A_{ii}^l\| \leq K_{BT} 2^{-l}$$

We kunnen de A_{ii}^l steeds in groepjes van 4 bijeenemen, dus voor $1 \leq s \leq l$ definiëren we 4^s disjuncte centrale deelmatrices $A_{ii}^s = A_{\sigma_i^s \times \sigma_i^s}$ van afmeting 4^{n-s} door

$$\sigma_i^{s-1} = \sigma_{4i-3}^s \cup \sigma_{4i-2}^s \cup \sigma_{4i-1}^s \cup \sigma_{4i}^s$$

en we noemen $A' = A_{11}^0$ de deelmatrix die bestaat uit dezelfde rijen en kolommen als alle A_{ii}^l samen. Elke A_{ii}^s is dus te schrijven als blokmatrix

$$A_{ii}^s = \begin{pmatrix} A_{4i-3}^{s+1} & \cdot & \cdot & \cdot \\ \cdot & A_{4i-2}^{s+1} & \cdot & \cdot \\ \cdot & \cdot & A_{4i-1}^{s+1} & \cdot \\ \cdot & \cdot & \cdot & A_{4i}^{s+1} \end{pmatrix}$$

Dan geldt met de recursierelatie van lemma 3.3 dat

$$\begin{aligned} \lambda(A') &= \lambda(A_{11}^0) \\ &\leq \frac{2}{1-\varepsilon} \max_{1 \leq i \leq 4} \lambda(A_{ii}^1) + \frac{6}{\varepsilon^2} \\ &\leq \frac{2}{1-\varepsilon} \left(\frac{2}{1-\varepsilon} \max_{1 \leq i \leq 16} \lambda(A_{ii}^2) + \frac{6}{\varepsilon^2} \right) + \frac{6}{\varepsilon^2} \\ &\leq \dots \\ &\leq \left(\frac{2}{1-\varepsilon} \right)^l \max_{1 \leq i \leq 4^l} \lambda(A_{ii}^l) + \frac{6}{\varepsilon^2} \sum_{i=0}^{l-1} \left(\frac{2}{1-\varepsilon} \right)^i \end{aligned}$$

Nu geldt

$$\lambda(A_{ii}^l) \leq \lambda(4^{n-l}) \|A_{ii}^l\| \leq \lambda(4^{n-l}) K_{BT} 2^{-l}$$

en dus kan de eerste term worden afgeschat met

$$\begin{aligned} \left(\frac{2}{1-\varepsilon}\right)^l \max_{1 \leq i \leq 4^l} \lambda(A_{ii}^l) &\leq \left(\frac{2}{1-\varepsilon}\right)^l \lambda(4^{n-l}) K_{BT} 2^{-l} \\ &= K_{BT} \left(\frac{1}{1-\varepsilon}\right)^l \lambda(4^{n-l}) \\ &\leq 4K_{BT} \lambda(4^{n-l}) \end{aligned}$$

waarbij we de laatste afchatting maken door $\varepsilon = \frac{1}{l}$ te kiezen zodat

$$\left(\frac{1}{1-\varepsilon}\right)^l \leq 4$$

net als in het bewijs van stelling 3.5 ($l \geq 2$ want we hebben aangenomen $n \geq 2(q+1)$). Merk op dat de factor 2^l hier wegvalt. Voor deze keuze van ε hebben we bij stelling 3.5 ook al bewezen dat

$$\frac{6}{\varepsilon^2} \sum_{i=0}^{l-1} \left(\frac{2}{1-\varepsilon}\right)^i \leq 24l^2 2^l$$

Als we dit samen nemen dan volgt

$$\lambda(A') \leq 4K_{BT} \lambda(4^{n-l}) + 24l^2 2^l$$

Herinneren we nu dat

$$\begin{aligned} l &= \frac{n}{q+1} \\ n-l &= n - \frac{n}{q+1} = \frac{nq}{q+1} \end{aligned}$$

(merk ook op $n-l \geq l$) en gebruiken we de aanname

$$\lambda(4^n) \leq K_q n^p 2^{\frac{n}{q}}$$

(waarbij $p \geq 2$) dan

$$\begin{aligned} \lambda(A') &\leq 4K_{BT} \lambda(4^{n-l}) + 24l^2 2^l \\ &= 4K_{BT} K_q (n-l)^p 2^{\frac{n-l}{q}} + 24l^2 2^l \\ &\leq 4K_{BT} K_q (n-l)^p 2^{\frac{n}{q+1}} + 24l^2 2^{\frac{n}{q+1}} \\ &\leq 4K_{BT} K_q (n-l)^p 2^{\frac{n}{q+1}} + 24(n-l)^p 2^{\frac{n}{q+1}} \\ &= (4K_{BT} K_q + 24) (n-l)^p 2^{\frac{n}{q+1}} \\ &= (4K_{BT} K_q + 24) \left(\frac{q}{q+1}\right)^p n^p 2^{\frac{n}{q+1}} \\ &\leq (4K_{BT} K_q + 24) n^p 2^{\frac{n}{q+1}} \end{aligned}$$

(hier zien we waarom $l = \frac{n}{q+1}$ een handige keuze is: we kunnen de tweemachten samennemen). $\square$

Voorbeeld 3.13. Van stelling 3.5 weten we dat

$$\lambda(4^n) \leq 24n^2 2^n$$

voor $n \geq 2$ en dus

$$\lambda(4^n) \leq 192n^2 2^n$$

voor alle $n \geq 1$ (want $\lambda(4^1) \leq \lambda(4^2)$). Toepassen van stelling 3.12 vertelt ons dat iedere holle $2 \cdot 4^n \times 2 \cdot 4^n$ matrix met $n = 2l$ voor $l \geq 2$ geheel en $\|A\| = 1$ een centrale $4^n \times 4^n$ deelmatrix A' bevat waarvoor

$$\lambda(A') \leq K n^2 2^{\frac{n}{2}}$$

voor zekere constante K onafhankelijk van n .

Merk op dat voor $n = \log_4 m$ deze grens overeenkomt met $\mathcal{O}\left((\log_4 m)^2 m^{1/4}\right)$; we hebben dus een soort van verbetering van stelling 3.5 ($\mathcal{O}\left((\log_4 m)^2 m^{1/2}\right)$) te pakken. Dit geldt echter nog alleen voor bepaalde deelmatrices.

3.5 De commutatorstelling $\mathcal{O}(m^\varepsilon)$

In deze paragraaf zullen we de resultaten van de vorige twee paragrafen combineren. Lemma 3.12 uit paragraaf 3.4 stelt ons in staat om bepaalde kleine deelmatrices te vinden en met lemma 3.8 uit paragraaf 3.3 kunnen we hieruit ook grenzen voor algemene matrices afleiden. Dit alles zullen we weer voor grenzen van een algemene vorm doen, zodat we dit uiteindelijk net zo vaak kunnen herhalen totdat onze grens willekeurig klein is geworden.

Vanaf nu zullen we wat losser met onze constanten omgaan, zoals de auteurs in hun artikel ook doen: K (zonder subscript) staat in deze sectie voor een willekeurige constante die onafhankelijk van n is maar wel kan veranderen. Meestal zullen we dit doen zodat grenzen die vanaf een bepaalde n gelden ook voor alle n gelden.

Lemma 3.14. *Neem aan dat voor gehele $p \geq 2$, $q \geq 1$ en $n \geq 1$ geldt*

$$\lambda(4^n) \leq K_q n^p 2^{\frac{n}{q}}$$

Dan ook

$$\lambda(4^n) \leq K_{q+1} n^{p+8(q+1)} 2^{\frac{n}{q+1}}$$

voor een zekere constante K_{q+1} (onafhankelijk van n).

Bewijs. Noem $\tilde{q} = q + 1$ en zij $n = k\tilde{q}$, $k \geq 2$ en A een $4^{n+\tilde{q}} \times 4^{n+\tilde{q}}$ blokmatrix met $\|A\| = 1$ bestaande uit $4^n \times 4^n$ deelmatrices A_{ij} :

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1,4^{\tilde{q}}} \\ A_{21} & A_{22} & \dots & A_{2,4^{\tilde{q}}} \\ \vdots & \vdots & \ddots & \vdots \\ A_{4^{\tilde{q}},1} & A_{4^{\tilde{q}},1} & \dots & A_{4^{\tilde{q}},4^{\tilde{q}}} \end{pmatrix}$$

Noteer $A_{[u,v]}$ voor de centrale deelmatrix gevormd uit de A_{ij} met $u \leq i, j \leq v$, b.v.

$$A_{[1,3]} = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

Wegens lemma 3.12 bevat $A_{[1,2]}$ een centrale $4^n \times 4^n$ deelmatrix $A_{11} = A_{\sigma \times \sigma}$ waarvoor

$$\lambda(A_{11}) \leq Kn^p 2^{\frac{n}{\tilde{q}}}$$

Noem $A_{22} = A_{\sigma^c \times \sigma^c}$ de centrale deelmatrix die de andere helft van de diagonaal van A vormt, dus $\sigma^c = \{1, 2, \dots, 2 \cdot 4^n\} \setminus \sigma$. Merk op dat dit niet noodzakelijk strijdig is met onze eerdere definitie van A_{11} en A_{22} , we kunnen immers altijd de rijen en kolommen van van A op de gewenste manier herschikken zonder dat $\lambda(A)$ verandert (lemma 2.13; het permuteren van de rijen en kolommen kan met een unitaire transformatie). Op dezelfde manier kunnen we lemma 3.12 toepassen op de overige centrale $2 \cdot 4^n \times 2 \cdot 4^n$ deelmatrices $A_{[2k-1, 2k]}$ met $1 \leq k \leq 4^{\tilde{q}}/2$, er volgt:

$$\lambda(A_{ii}) \leq Kn^p 2^{\frac{n}{\tilde{q}}}$$

voor alle oneven i . Met lemma 3.8 volgt nu voor alle centrale $2 \cdot 4^n \times 2 \cdot 4^n$ deelmatrices ($1 \leq k \leq 4^{\tilde{q}}/2$)

$$\begin{aligned} \lambda(A_{[2k-1, 2k]}) &\leq F \max \left\{ Kn^{p+4} 2^{\frac{n}{\tilde{q}}}, \lambda(A_{(2k)(2k)}) \right\} \\ &\leq F \max \left\{ Kn^{p+4} 2^{\frac{n}{\tilde{q}}}, \lambda(4^n) \right\} \end{aligned}$$

waarbij

$$F = F_k = 1 + \frac{4}{n^2} = 1 + \frac{4}{k^2 \tilde{q}^2}$$

en $\lambda(A_{(2k)(2k)}) \leq \lambda(4^n)$ omdat $\|A_{(2k)(2k)}\| \leq \|A\| = 1$.

Merk op dat zeker

$$\lambda(A_{[2k-1, 2k]}) \leq FKn^{p+4} 2^{\frac{n}{\tilde{q}}}$$

Om die reden kunnen we lemma 3.8 nogmaals toepassen, maar nu op de grotere

$4 \cdot 4^n \times 4 \cdot 4^n$ deelmatrices (b.v. $A_{[1,4]}$) waarmee volgt

$$\begin{aligned}
\lambda(A_{[4k-3,4k]}) &\leq F \max \left\{ FK n^{p+8} 2^{\frac{n}{q}}, \lambda(A_{[4k-1,4k]}) \right\} \\
&\leq F \max \left\{ FK n^{p+8} 2^{\frac{n}{q}}, FK n^{p+4} 2^{\frac{n}{q}}, F \lambda(4^n) \right\} \\
&\leq F \max \left\{ FK n^{p+8} 2^{\frac{n}{q}}, FK n^{p+8} 2^{\frac{n}{q}}, F \lambda(4^n) \right\} \\
&= F^2 \max \left\{ K n^{p+8} 2^{\frac{n}{q}}, \lambda(4^n) \right\}
\end{aligned}$$

voor $1 \leq k \leq 4^{\tilde{q}}/4$. Op dezelfde wijze kunnen we doorgaan: bij elke stap worden de matrices die we beschouwen tweemaal zo groot en krijgt de afschatting er een factor F en n^4 bij. Na $\log_2 4^{\tilde{q}} = 2\tilde{q}$ stappen volgt zo uiteindelijk

$$\lambda(A) \leq F^{2\tilde{q}} \max \left\{ K n^{p+8\tilde{q}} 2^{\frac{n}{q}}, \lambda(4^n) \right\}$$

maar A was willekeurig van afmeting $4^{n+\tilde{q}}$ en $\|A\| = 1$, dus ook

$$\lambda(4^{n+\tilde{q}}) \leq F^{2\tilde{q}} \max \left\{ K n^{p+8\tilde{q}} 2^{\frac{n}{q}}, \lambda(4^n) \right\}$$

Noemen we $\tilde{p} = p + 8\tilde{q}$ en $G_k = F_k^{2\tilde{q}}$ dan kunnen ook schrijven

$$\begin{aligned}
\lambda(4^{k\tilde{q}}) &\leq G_k \max \left\{ K n^{\tilde{p}} 2^{k-1}, \lambda(4^{(k-1)\tilde{q}}) \right\} \\
&\leq G_k \max \left\{ K n^{\tilde{p}} 2^k, \lambda(4^{(k-1)\tilde{q}}) \right\}
\end{aligned}$$

Herhaald invullen van deze recursierelatie (waarbij we goed op moeten letten dat F en G van k afhangen) geeft

$$\begin{aligned}
\lambda(4^{k\tilde{q}}) &\leq G_k \max \left\{ K n^{\tilde{p}} 2^k, \lambda(4^{(k-1)\tilde{q}}) \right\} \\
&\leq G_k \max \left\{ K n^{\tilde{p}} 2^k, \max \left\{ G_{k-1} K n^{\tilde{p}} 2^{k-1}, G_{k-1} \lambda(4^{(k-2)\tilde{q}}) \right\} \right\} \\
&= G_k \max \left\{ K n^{\tilde{p}} 2^k, G_{k-1} K n^{\tilde{p}} 2^{k-1}, G_{k-1} \lambda(4^{(k-2)\tilde{q}}) \right\} \\
&= G_k \max \left\{ K n^{\tilde{p}} 2^k, G_{k-1} K n^{\tilde{p}} 2^k, G_{k-1} \lambda(4^{(k-2)\tilde{q}}) \right\} \\
&\leq G_k \max \left\{ G_{k-1} K n^{\tilde{p}} 2^k, G_{k-1} K n^{\tilde{p}} 2^k, G_{k-1} \lambda(4^{(k-2)\tilde{q}}) \right\} \\
&= G_k G_{k-1} \max \left\{ K n^{\tilde{p}} 2^k, \lambda(4^{(k-2)\tilde{q}}) \right\} \\
&\dots \\
&\leq \left(\prod_{i=2}^k G_i \right) \max \left\{ K n^{\tilde{p}} 2^k, \lambda(4^{2\tilde{q}}) \right\}
\end{aligned}$$

Verder dan dit kunnen we niet gaan, want in lemma 3.12 nemen we aan dat $n \geq 2\tilde{q}$. Nu is $\lambda(4^{2\tilde{q}})$ een eindig getal (onafhankelijk van k of n), dus kunnen

we altijd de constante K vergroten zodat

$$\lambda(4^{k\tilde{q}}) \leq \left(\prod_{i=2}^k G_i \right) K n^{\tilde{p}} 2^k$$

De voorfactor kunnen we uitwerken tot

$$\prod_{i=2}^k G_i = \prod_{i=2}^k F_i^{2\tilde{q}} = \left(\prod_{i=2}^k F_i \right)^{2\tilde{q}}$$

en omdat $\tilde{q} = q + 1 \geq 2$ is

$$\begin{aligned} \prod_{i=2}^k F_i &= \prod_{i=2}^k \left(1 + \frac{4}{i^2 \tilde{q}^2} \right) \\ &\leq \prod_{i=2}^k \left(1 + \frac{2}{i^2} \right) \\ &\leq \prod_{i=1}^{\infty} \left(1 + \frac{2}{i^2} \right) \end{aligned}$$

en dit oneindig product convergeert, want

$$\ln \left(\prod_{i=1}^{\infty} \left(1 + \frac{2}{i^2} \right) \right) = \sum_{i=1}^{\infty} \ln \left(1 + \frac{2}{i^2} \right) \leq \sum_{i=1}^{\infty} \frac{2}{i^2} = 2 \sum_{i=1}^{\infty} \frac{1}{i^2} = 2 \frac{\pi^2}{6} = \frac{\pi^2}{3}$$

waarbij $\sum_{i=1}^{\infty} \frac{1}{i^2}$ de beroemde reeks is waarvan Euler aantoonde dat deze naar $\pi^2/6$ convergeert. Nu zien we ook waarom we de macht $p + 4$ moesten nemen in lemma 3.8, want als we daar bijvoorbeeld $p + 2$ hadden genomen dan krijgen zou $Q_i \approx 1 + \frac{1}{i}$ (geen kwadraat) en dan werkt deze afchatting niet:

$$\prod_{i=1}^{\infty} \left(1 + \frac{1}{i} \right) = \frac{2}{1} \cdot \frac{3}{2} \cdot \frac{4}{3} \cdot \frac{5}{4} \cdots = \infty$$

We kunnen de factor $\left(\prod_{i=2}^k Q_i \right)$ dus afschatten met een constante die niet van k afhangt. Als we die constante opnemen in onze K hebben we

$$\lambda(4^{k\tilde{q}}) \leq K n^{\tilde{p}} 2^k$$

Vullen de waarden van $\tilde{p}$, $\tilde{q}$ en $k = \frac{n}{\tilde{q}}$ in dan

$$\lambda(4^n) \leq K n^{\tilde{p}} 2^{\frac{n}{\tilde{q}}} = K n^{p+8(q+1)} 2^{\frac{n}{q+1}}$$

Als laatste stap merken we nog op dat we $\lambda(4^n)$ voor n ondeelbaar door $\tilde{q}$ altijd kunnen afschatten door een grotere $n' = \left\lceil \frac{n}{\tilde{q}} \right\rceil \tilde{q}$ te nemen; dit verandert niets aan de grens (op de constante na). Ook hebben we een aantal keer aangenomen dat $n \geq 4$, maar ook dat kunnen we opvangen door de constante groter te maken zodat de grens geldt voor alle n . $\square$

De auteurs zijn in hun artikel vrij slordig in dit deel van het bewijs. Ze vermelden bijvoorbeeld klakkeloos dat de procedure voor grotere q herhaald kan worden maar vergeten daarbij dat lemma 3.12 alleen kan worden toegepast op matrices van afmeting $2 \cdot 4^n$ met n deelbaar door $q+1$. Dit zorgt ervoor dat we stappen van 4^{q+1} moeten maken in bovenstaand bewijs in plaats van stappen van 4^n zoals de auteurs in hun artikel doen. Ook beweren zij hun versie van lemma 3.8 toe te kunnen passen om van $2 \cdot 4^n$ naar $4 \cdot 4^n$ matrices te gaan zonder dat de grens (in het bijzonder de exponent van n) hierbij verslechterd. Vermoedelijk zijn dit soort details de auteurs ontschoten tijdens het schrijven van het artikel; gelukkig heeft het geen invloed op de juistheid van het uiteindelijke resultaat.

Lemma 3.14 kunnen we nu willekeurig vaak herhalen en zo volgt:

Stelling 3.15 (Commutatorstelling $\mathcal{O}(m^\varepsilon)$). *Voor alle $\varepsilon > 0$ bestaat er een constante K_ε zodat voor iedere spoorloze matrix $A \in \mathbb{C}^{m \times m}$ er matrices $B, C \in \mathbb{C}^{m \times m}$ bestaan waarvoor $A = BC - CB$ en $\|B\| \|C\| \leq K_\varepsilon m^\varepsilon \|A\|$.*

Bewijs. Van lemma 3.5 weten we dat

$$\lambda(4^n) \leq K_1 n^2 2^n$$

Herhaald toepassen van lemma 3.14 geeft nu achtereenvolgens

$$\begin{aligned} \lambda(4^n) &\leq K_2 n^{18} 2^{\frac{n}{2}} \\ \lambda(4^n) &\leq K_3 n^{42} 2^{\frac{n}{3}} \\ &\vdots \\ \lambda(4^n) &\leq K_q n^{2+8(2+3+4+\dots+q)} 2^{\frac{n}{q}} \end{aligned}$$

met $1 + 2 + 3 + \dots + q = \frac{1}{2}q(q+1)$ volgt voor elke $q \geq 1$ een bovengrens van de vorm

$$\lambda(4^n) \leq K_q n^{4q(q+1)-6} 2^{\frac{n}{q}}$$

Zij nu $\varepsilon > 0$ gegeven en kies q zodat $\frac{1}{q} < \varepsilon$. We bewijzen

$$n^{4q(q+1)-6} \leq K_f 2^{\frac{n}{q}}$$

voor zekere constante K_f onafhankelijk van n . Bekijken we de logaritme van bovenstaande uitdrukking, dan willen we aantonen dat

$$[4q(q+1) - 6] \log_2 n \leq \log_2 K_f + \frac{n}{q}$$

Noem nu

$$[4q(q+1) - 6] \log_2 n - \frac{n}{q} \leq 4q(q+1) \log_2 n - \frac{n}{q} = f(n)$$

dan is de afgeleide

$$f'(n) = \frac{4q(q+1)}{\ln 2} \frac{1}{n} - \frac{1}{q}$$

en voor alle $n \geq N = \frac{4q(q+1)}{\ln 2}q$ geldt

$$f'(n) \leq f'(N) = \frac{1}{q} - \frac{1}{q} = 0$$

en dus $f(n) \leq f(N)$ voor alle $n \geq N$. Nemen we nu $K_f = 2^{f(N)}$ dan geldt voor alle $n \geq N$

$$\begin{aligned} [4q(q+1) - 6] \log_2 n - \frac{n}{q} &\leq 4q(q+1) \log_2 n - \frac{n}{q} \\ &= f(n) \\ &\leq f(N) \\ &= \log_2 K_f \end{aligned}$$

en dus

$$[4q(q+1) - 6] \log_2 n \leq \log_2 K_f + \frac{n}{q}$$

ofwel

$$n^{4q(q+1)-6} \leq K_f 2^{\frac{n}{q}}$$

Tenslotte kunnen we K_f altijd zo groot maken dat deze ongelijkheid ook geldt voor $n < N$.

Nemen we nu $K_\varepsilon = K_q K_f$ dan volgt

$$\lambda(4^n) \leq K_q n^{4q(q+1)-6} 2^{\frac{n}{q}} \leq K_q K_f 2^{\frac{n}{q}} 2^{\frac{n}{q}} = K_\varepsilon 2^{2\varepsilon n} = K_\varepsilon 4^{\varepsilon n}$$

en met $n = \log_4 m$

$$\lambda(m) \leq K_\varepsilon m^\varepsilon$$

en met de definitie van $\lambda(m)$ volgt dan de geclaimde commutatorstelling. $\square$

Hoofdstuk 4

Een algoritme om commutators te ontbinden

In dit hoofdstuk bekijken we een aantal procedures die samen een algoritme vormen waarmee commutator matrices ontbonden kunnen worden. We beperken ons hierbij tot een beknopte beschrijving van de te nemen stappen; aan de hand hiervan kan een implementatie in iedere gewenste programmeertaal gemaakt worden. Steeds zal worden aangegeven waar een bewijs van correctheid gevonden kan worden. In de bijlage is een Matlab implementatie van de procedures uit dit hoofdstuk opgenomen; in paragraaf 4.3 zullen we het gedrag van dit programma bekijken.

4.1 Ontbindingsprocedure

Het volgende algoritme is een direct implementatie van de commutatorstelling (lemma 2.13 en stelling 2.14):

Algoritme 4.1 (Commutator ontbinden, geïmplementeerd in alg. A.1 en A.3).

INVOER:

$m \times m$ matrix A met $\text{spoor}(A) = 0$.

UITVOER:

Matrices B en C waarvoor $A = BC - CB$;
 B een normale matrix.

PROCEDURE:

Bereken unitaire U waarvoor $H = U^*AU$ hol is (algoritme 4.2).
Zij B' een diagonaal matrix met louter verschillende elementen.

Zij C' een holle matrix met niet-diagonaalelementen gegeven door

$$c'_{ij} = \frac{h_{ij}}{b'_i - b'_j}$$

Dan voldoen $B = UB'U^*$ en $C = UC'U^*$ aan de eisen.

De elementen van de diagonaalmatrix B' kunnen willekeurig gekozen worden zolang er geen twee dezelfde waarden gebruikt worden. Dit kan bijvoorbeeld met waarden uit het interval $[-1, 1]$ zoals in stelling 2.36 (geïmplementeerd in algoritme A.1) of met waarden op ringen binnen de complexe eenheidsschijf zoals in stelling 2.39 (geïmplementeerd in algoritme A.3). In beide gevallen geeft dit $\|B\| = 1$.

4.2 Uitholprocedure

Het meeste werk bij het ontbinden van een commutator zit hem in het vinden van een geschikte holle matrix, in de vorm van het volgende algoritme (gebaseerd op het bewijs van lemma 2.12):

Algoritme 4.2 (Spoorloze matrix uithollen, geïmplementeerd in alg. A.6).

INVOER:

$m \times m$ matrix A met $\text{spoor}(A) = 0$.

UITVOER:

Unitaire matrix U waarvoor U^*AU hol is.

PROCEDURE:

Als $m = 1$ dan is $U = (1)$ de gezochte matrix. Zo niet dan:

Bereken eenheidsvector x waarvoor $x^*Ax = 0$ (algoritme 4.3)

Pas Gram-Schmidt toe op x en de collectie eenheidsvectoren e_i

Vorm de matrix X uit de resulterende orthonormale kolomvectoren

Noem A' de deelmatrix van X^*AX die ontstaat door de eerste rij en kolom weg te laten

Bereken U' waarvoor $U'^*A'U'$ hol is door dit algoritme (4.2) recursief toe te passen

$U = XU'$ is nu de gezochte matrix

Om dit werkend te krijgen moeten we de nulpunten van het numerieke bereik kunnen vinden, zoals in lemma 2.11:

Algoritme 4.3 (Nulpunt numeriek bereik, geïmplementeerd in alg. A.8).

INVOER:

$m \times m$ matrix A met $\text{spoor}(A) = 0$.

UITVOER:

Eenheidsvector x waarvoor $x^*Ax = 0$.

PROCEDURE:

Bereken de genormaliseerde eigenvectoren $\{v_i\}$ en bijbehorende eigenwaarden $\{\lambda_i\}$ van A (inclusief algebraïsche multipliciteit; voor defecte A moet een eigenvector meerdere malen voorkomen)

Als v_i een eigenvector is bij eigenwaarde 0 dan voldoet $x = v_i$ aan de eisen
Zo niet, berekende dan iteratief het gemiddelde $g = 0$ van de eigenwaarden:

$x \leftarrow v_1$

$g \leftarrow \lambda_1$

voor $2 \leq i \leq m$ doe

$t \leftarrow (i - 1) / i$

$g \leftarrow tg + (1 - t)\lambda_i$

$x \leftarrow$ de bijbehorende vector waarvoor $x^*Ax = g$,

berekend met x en v_i (algoritme 4.4)

Nu is x de gezochte vector.

De volgende twee algoritmes zijn gebaseerd op het bewijs van de stelling van Toeplitz-Hausdorff uit [2, stelling 22.27]. De convexiteit van het numerieke bereik wordt daar bewezen door aan te tonen dat $ta + (1 - t)b$ in het numerieke bereik zit voor alle $0 \leq t \leq 1$ (als a en b in het numerieke bereik zitten). Het bewijs zelf is vrij technisch en zullen we hier niet herhalen; we beperken ons tot het algoritme dat eruit afgeleid kan worden. Eerst wordt het probleem gereduceerd tot het eenvoudigere geval waarin $a = 1$ en $b = 0$:

Algoritme 4.4 (Interpoleren numeriek bereik, geïmplementeerd in alg. A.9).

INVOER:

$m \times m$ matrix A ,

eenheidsvectoren x en y waarvoor $x^*Ax = a$ en $y^*Ay = b$,

een reëel getal $0 \leq t \leq 1$.

UITVOER:

Eenheidsvector z waarvoor $z^*Az = ta + (1 - t)b$.

PROCEDURE:

Als $a = b$ dan voldoet $z = x$ (of $z = y$). Zo niet, neem

$$\alpha = \frac{1}{a-b}$$

$$\beta = \frac{-b}{a-b}$$

$$B = \alpha A + \beta I$$

Bereken z waarvoor $z^*Bz = t$ (algoritme 4.5)

Dan voldoet z aan de gestelde eis.

Het echte werk gebeurt in het volgende algoritme, gebaseerd op lemma 22.16 uit [2]:

Algoritme 4.5 (Interpoleren numeriek bereik ($a = 1$ en $b = 0$), geïmplementeerd in alg. A.10).

INVOER:

$m \times m$ matrix B ,
 eenheidsvectoren x en y waarvoor $x^* B x = 1$ en $y^* B y = 0$,
 een reëel getal $0 \leq \lambda \leq 1$.

UITVOER:

Eenheidsvector z waarvoor $z^* B z = \lambda$.

PROCEDURE:

$C = (B + B^*)/2$

$D = (B - B^*)/(2i)$

$c = y^* C x$

$d = y^* D x$

Als $d \neq 0$ dan $\gamma = i\bar{d}/|d|$ anders $\gamma = 1$

Los

$$\frac{t^2 + t(1-t)[\gamma c + \overline{\gamma c}]}{t^2 + t(1-t)[y^* \gamma x + x^* \overline{\gamma} y] + (1-t)^2} = \lambda$$

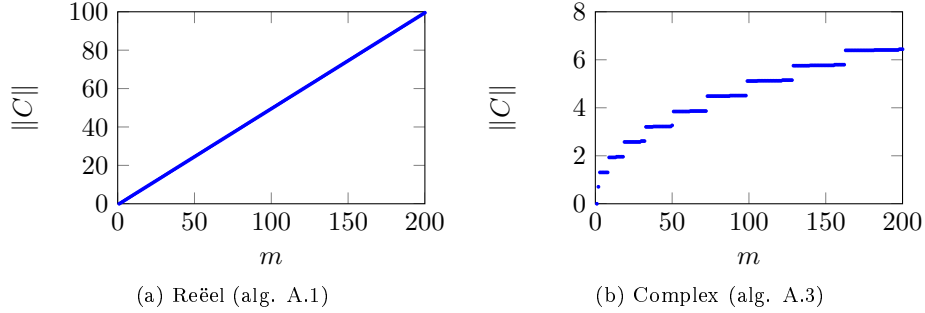
op voor $0 \leq t \leq 1$.

Dan voldoet $z = t\gamma x + (1-t)y$ aan $z^* B z = \lambda$.

4.3 Gedrag van het programma

Het eerste wat opvalt als we kijken naar het gedrag van het programma is natuurlijk dat de uitvoer slechts een eindige nauwkeurigheid heeft. Bij elke stap worden er wegens de eindige precisie van de computer afrondfouten gemaakt waardoor de berekende B en C niet exact A opleveren als we $BC - CB$ bekijken (zie voorbeeld A.2 en A.4).

Daarnaast gaat het meeste werk in het vinden van een geschikte holle matrix zitten. Dit geldt voor de hoeveelheid benodigde code, maar ook voor de benodigde rekentijd. Het ontbinden van 1000 willekeurige spoorloze 3×3 matrices kostte in één geval bijvoorbeeld 100 seconden, terwijl het ontbinden van 1000 willekeurige *holle* 3×3 matrices (die niet 'uitgehold' hoeven te worden) slechts 0.25 seconden kostte. Bij het uithollen gaat volgens de Matlab profiler vrijwel alle tijd zitten in het oplossen van de vergelijking voor t in de `interpNB10` routine. Om het algoritme sneller te maken zouden we deze vergelijking dus het beste op een andere manier kunnen oplossen, bijvoorbeeld door de oplossing expliciet uit te rekenen (het is een kwadratische vergelijking). Voor grotere matrices neemt de benodigde tijd ook nog eens rap toe; experimenteren leert dat de benodigde tijd ongeveer met het kwadraat van de matrixdimensie gaat. Dit kunnen we verklaren door te kijken naar de code. De routine `vindNulpuntNB` roept `interpNB10` $m-1$ keer aan voor een $m \times m$ matrix. De routine `maakHol` roept `vindNulpuntNB` één keer aan maar werkt recursief, dus het hol maken van een $m \times m$ matrix vereist in het slechtste geval



Figuur 4.1: $\|C\|$ als functie van matrixgrootte m voor A als in vergl. 4.1

$(m-1) + (m-2) + \dots + 1 = \frac{1}{2}m(m-1)$ aanroepen van `interpNB10`; wat zich vertaalt in een tijdscomplexiteit van $\mathcal{O}(m^2)$.

Bijzondere matrices

We kunnen ook kijken naar de norm van de berekende C ($\|B\|$ is altijd 1); voor het gemak nemen we $\|A\| = 1$. Omdat het algoritme B onafhankelijk van A kiest kunnen we proberen een A te bedenken waarvoor de uitvoer zo slecht mogelijk is (i.e. $\|C\|$ zo groot mogelijk). Dit kan bijvoorbeeld door de matrix

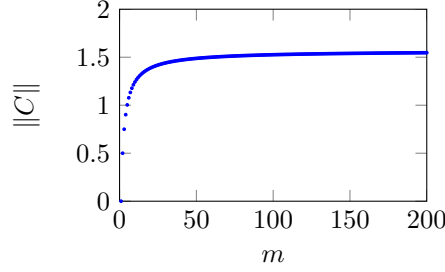
$$A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (4.1)$$

met 1 op de eerste zijdiagonaal te nemen, waarmee waarden in B die dicht bij elkaar liggen geselecteerd worden. Dit levert C 's op met normen die lineair of in de wortel van de dimensie groeien (zie figuur 4.1).

Als de waarden van A erg verschillen is het in theorie mogelijk om de waarden in B zo te herschikken dat C zo klein mogelijk wordt. Om die reden is het interessant om te kijken wat er gebeurt als alle elementen van $A \in \mathbb{C}^{m \times m}$ gelijk zijn:

$$A = \frac{1}{m-1} \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & 1 & \dots & 1 \\ 1 & 1 & 0 & \ddots & 1 \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \quad (4.2)$$

(de voorfactor $\frac{1}{m-1}$ zorgt ervoor dat $\|A\| = 1$). In dit geval lijkt de norm van C begrensd (zie figuur 4.2). Laten we de computer even rekenen op het



Figuur 4.2: Grootte van $\|C\|$ (alg. A.1) als functie van matrixgrootte m voor holle A met gelijke elementen (vergl. 4.2).

geval $m = 10.000$ dan vinden we $\|C\| \approx 1.5696$ en dat lijkt verdacht veel op $\frac{\pi}{2} \approx 1.5708$. We zullen nu aantonen dat $\|C\|$ inderdaad naar $\frac{\pi}{2}$ convergeert.

Als we algoritme A.1 of stelling 2.36 goed bekijken zien we dat de elementen van B en C in dit geval gegeven worden door:

$$b_i = -1 + 2 \frac{i-1}{m-1}$$

$$c_{ij} = \frac{a_{ij}}{b_i - b_j} = \frac{1/(m-1)}{2(i-j)/(m-1)} = \frac{1}{2(i-j)}$$

(met diagonaalelementen $c_{ii} = 0$). De $m \times m$ matrix C_m is dus van de volgende vorm

$$C_m = \frac{1}{2} \begin{pmatrix} 0 & -1 & -\frac{1}{2} & -\frac{1}{3} & \cdots & -\frac{1}{m-1} \\ 1 & 0 & -1 & -\frac{1}{2} & \ddots & \vdots \\ \frac{1}{2} & 1 & 0 & -1 & \ddots & -\frac{1}{3} \\ \frac{1}{3} & \frac{1}{2} & 1 & 0 & \ddots & -\frac{1}{2} \\ \vdots & \ddots & \ddots & \ddots & \ddots & -1 \\ \frac{1}{m-1} & \cdots & \frac{1}{3} & \frac{1}{2} & 1 & 0 \end{pmatrix}$$

Onze kennis over normen uit hoofdstuk 2 vertelt ons niet zoveel nuttigs, we kunnen enkel concluderen dat

$$\|C_m\| \geq \|C\|_{\max} = \frac{1}{2}$$

$$\|C_m\| \leq \|C_m\|_1 = \|C_m\|_{\infty} \geq \frac{1}{2} \sum_{k=1}^{m-1} \frac{1}{k} \rightarrow \infty$$

(voor $m \rightarrow \infty$). Om toch iets te kunnen zeggen over $\|C_m\|$ zullen we dus andere truukjes moeten gebruiken. Nu is C_m een zogenaamde Toeplitz matrix: op elke diagonaal van C_m zijn de elementen gelijk. We kunnen C_m dus volledig

beschrijven met de rij

$$c_k = \begin{cases} 0 & k = 0 \\ \frac{1}{2k} & k \neq 0 \end{cases}$$

($-m+1 \leq k \leq m-1$); de elementen van C_m worden zo gegeven door $c_{ij} = c_{i-j}$. Een hele mooie stelling van Grenander en Szegő [4] zegt nu grofweg het volgende: in de limiet $m \rightarrow \infty$ zijn van de eigenwaarden van C_m verdeelt als de waarden van de Fourierreeks f met Fourier coëfficiënten c_k . Preciezer gezegd: het aantal eigenwaarden in een zekere verzameling V is evenredig met de maat van het inverse beeld $f^{-1}(V)$. Deze stelling kunnen we het makkelijkst toepassen door naar de matrices $Z_m = C_m/\imath$ te kijken (die dezelfde norm hebben; $\|Z_m\| = \|C_m\|$). Z_m zijn hermitische matrices (i.e. $Z_m^* = Z_m$) en dus volgt dat de grootste singuliere waarde en daarmee $\|Z_m\|$ gelijk is aan de grootste eigenwaarde (in absolute zin). De bijbehorende Fourierreeks is

$$\begin{aligned} f(t) &= \sum_{k=-\infty}^{\infty} \frac{c_k}{\imath} \exp(\imath kt) \\ &= \sum_{k \neq 0} \frac{1}{2\imath k} \exp(\imath kt) \\ &= \sum_{k=1}^{\infty} \frac{1}{2\imath k} [\exp(\imath kt) - \exp(-\imath kt)] \\ &= \sum_{k=1}^{\infty} \frac{\sin(kt)}{k} \end{aligned}$$

Deze standaardreeks convergeert naar de zaagtandfunctie met bereik $[-\frac{\pi}{2}, \frac{\pi}{2}]$. De eigenwaarden van Z_m zijn dus asymptotisch uniform (want de zaagtandfunctie is lineair) verdeelt over het interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$, en daarmee volgt dat de grootste eigenwaarde van Z_m naar $\frac{\pi}{2}$ nadert voor $m \rightarrow \infty$. We kunnen dus concluderen dat

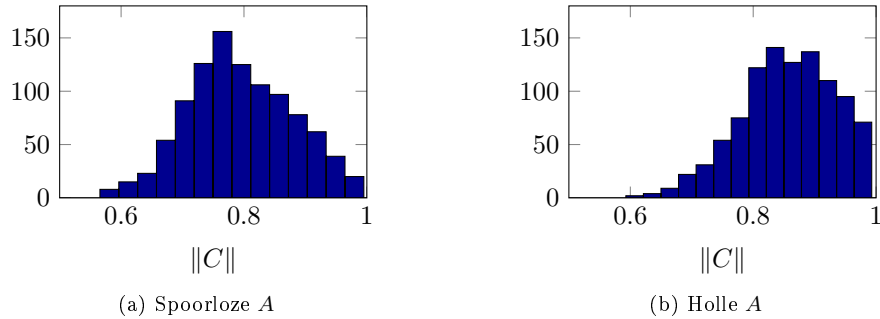
$$\lim_{m \rightarrow \infty} \|C_m\| = \lim_{m \rightarrow \infty} \|Z_m\| = \lim_{m \rightarrow \infty} \max |\lambda_i(Z_m)| = \frac{\pi}{2}$$

De observatie dat $\|C_m\|$ een stijgende rij vormt als functie van m kan eenvoudig verklaart worden met het feit dat iedere C_m een deelmatrix is van C_{m+1} (stelling 2.28).

Willekeurige testmatrices

Naast bijzondere A 's kunnen we ook kijken naar de uitvoer voor een 'doorsnee' A , d.w.z. een A waarvan de elementen volgens een bepaalde kansverdeling willekeurig zijn. In de bijlage is opgenomen hoe we dit soort matrices precies kunnen maken.

Als eerste onderzoeken we of de uitvoer van het algoritme vergelijkbaar is voor holle en spoorloze A . In figuur 4.3 zien we dat dit (voor 3×3) matrices

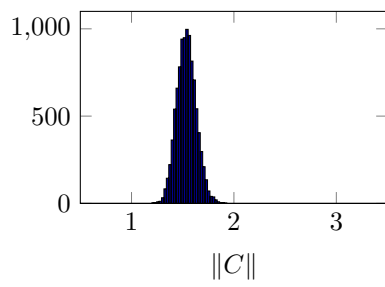


Figuur 4.3: Verdeling van $\|C\|$ (alg. A.1) bij 1000 willekeurige complexwaardige 3×3 matrices A .

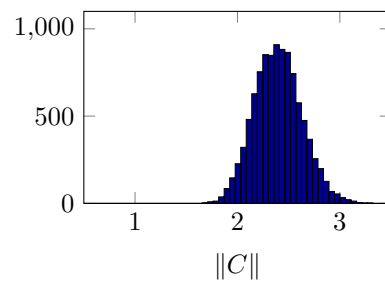
inderdaad zo is; de verdeling van $\|C\|$ is weliswaar anders maar het bereik is hetzelfde (beide $[\frac{1}{2}, 1]$). Vanaf nu zullen we dan ook alleen nog maar met holle matrices werken, wat als voordeel heeft dat de berekeningen vele malen sneller gaan. Hierbij dienen we wel op te merken dat het uithollen van een reëelwaardige matrix met het huidige algoritme een complexwaardige holle matrix op kan leveren.

In figuur 4.4 is de verdeling van de norm van de berekende C bij willekeurige holle A (ditmaal 10×10 matrices) weergegeven voor verschillende manieren om willekeurige matrices te genereren. Wat direct opvalt is dat geen van gebruikte methoden de waarde uit figuur 4.1 ($\|C\| = 4.5$ voor $m = 10$) weet te bereiken, noch het bekende minimum $\|C\| = \frac{1}{2}$ (opmerking 2.35). We zien ook $\|C\|$ in dit geval geconcentreerd is op $[2, 3]$, tenzij we A nemen met alleen niet-negatieve reële elementen.

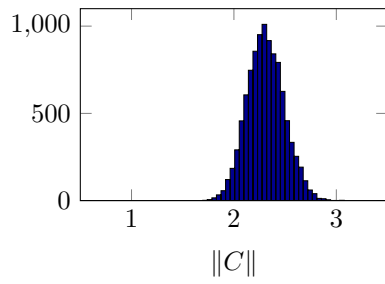
Interessanter wordt het als we kijken naar het gedrag van $\|C\|$ als functie van de matrixdimensie m . Daartoe is in figuur 4.5 steeds de grootste en de kleinste $\|C\|$ van een steekproef van 10 uitgezet. Deze minimale en maximale $\|C\|$ liggen nergens ver uit elkaar en geven samen dus een goede indicatie voor de verwachte $\|C\|$ (voor dit soort willekeurige A 's). Voor de matrices met niet-negatieve reële waarden lijkt $\|C\|$ begrensd; al is de grens bij (a) duidelijk groter dan de $\frac{\pi}{2}$ van het hierboven uitgewerkte geval van A met alle elementen gelijk aan $a_{ij} = \frac{1}{m-1}$. Voor complexwaardige A groeit $\|C\|$ bij het reële algoritme (A.1) ongeveer volgens $\sqrt{m}$, bij het complexwaardige algoritme (A.3) lijkt het echter veel langzamer te groeien; wellicht logaritmisch.



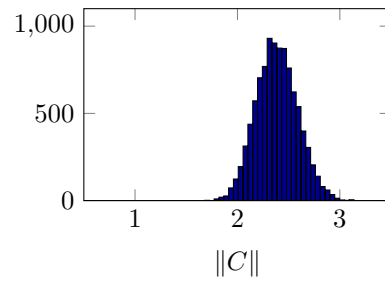
(a) $a_{ij} \in [0, 1]$



(b) $a_{ij} \in [-1, 1]$

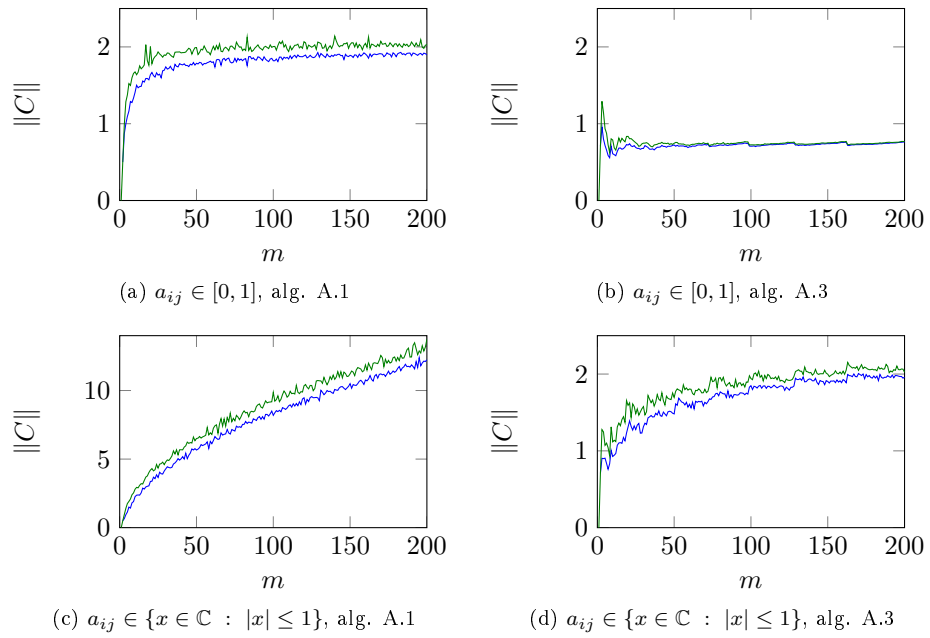


(c) $a_{ij} \in \{-1, 1\}$



(d) $a_{ij} \in \{x \in \mathbb{C} : |x| \leq 1\}$

Figuur 4.4: Verdeling van $\|C\|$ (alg. A.1) bij 10.000 willekeurige holle 10×10 matrices A .



Figuur 4.5: Minimale en maximale $\|C\|$ bij een steekproef van 10 willekeurige holle matrices A als functie van m .

Hoofdstuk 5

Conclusie en discussie

In hoofdstuk 2 hebben we gezien hoe de commutatorstelling bewezen kan worden door gebruik te maken van de stelling van Toeplitz-Hausdorff. Daarnaast hebben we een aantal eigenschappen van de geïnduceerde Euclidische matrixnorm bestudeerd en daarmee op eenvoudige wijze de commutatorstelling gekwantificeerd, zowel met een reële (bovengrens $\mathcal{O}(m^2)$) als complexwaardige (bovengrens $\mathcal{O}(m^{3/2})$) constructie. De bewijzen van deze kwantitatieve versies van de commutatorstelling konden volledig constructief gemaakt worden met de hulp van een constructief bewijs van de stelling van Toeplitz-Hausdorff uit [2]. Op die manier konden in hoofdstuk 4 beide versies in een computeralgoritme worden omgezet. Naar aanleiding van de uitvoer van deze algoritmes hebben we in het bijzonder gezien dat een holle matrix met enkel dezelfde niet-diagonaalelementen ontbonden kan worden in B en C met $\|B\| \|C\| < \frac{\pi}{2} \|A\|$.

In hoofdstuk 3 is het bewijs van de commutatorstelling met een grens $\mathcal{O}(m^\varepsilon)$ uit het artikel [6] uiteengezet. We hebben gezien dat dit een vrij technisch bewijs is waarbij de auteurs soms op ingenieuze wijze te werk gaan, bijvoorbeeld door een complexe contourintegraal te gebruiken om de norm van een matrix af te schatten. Om het uiteindelijke resultaat te bereiken gebruiken de auteurs een sterke stelling van Bourgain en Tzafriri [14]; verder is het bewijs volledig elementair.

Zoals wel vaker in de wiskunde eindigen we met meer vragen dan waarmee we begonnen. Het volgende zijn een aantal vragen die tijdens dit project naar voren zijn gekomen en nog niet konden worden beantwoord:

- *Zijn reëelwaardige commutatormatrices A altijd te ontbinden in twee reëelwaardige matrices B en C ?*

Voor holle matrices is dit zeker waar wegens de constructie in stelling 2.36. Algemener zou dit zo zijn als spoorloze reëelwaardige matrices met een orthogonale (in plaats van unitaire) matrixtransformatie in een holle matrix zijn om te zetten. De niet-lineariteit van de kwadratische vorm x^*Ax maakt dit niet-triviaal; toch wist de heer J. Vermeer dit op elegante wijze aan te tonen (zie bijlage B) - waarmee hij een positief antwoord op

deze vraag kon bevestigen.

- *Kan het algoritme zo worden aangepast dat reëelwaardige commutators altijd worden ontbonden in reëelwaardige matrices?*

In theorie zou dit mogelijk moeten zijn (zie het bewijs in bijlage B). In de praktijk kan dit waarschijnlijk het makkelijkst worden bewerkstelligd door de eigenwaarden en eigenvectoren van reële matrices in geconjugeerde paren af te handelen. Daarnaast zou de implementatie van de stelling van Toeplitz-Hausdorff nauwkeurig moeten worden nagegaan en (indien mogelijk) zo worden aangepast dat reële invoer ook altijd reële uitvoer oplevert.

- *Is er een directe manier om bij een spoorloze A een orthonormale collectie vectoren $\{x_i\}$ te vinden waarvoor $x_i^* A x_i = 0$?*

In hoofdstuk 2 gebruiken we hier een recursieve constructie voor die steeds één zo'n vector geeft. Een directe manier om in één keer een volledige orthonormale basis te vinden (zonder recursie op kleinere matrices) zou de algoritmes uit hoofdstuk 4 (in het bijzonder het vinden van een geschikte holle matrix) flink sneller maken.

- *Kunnen de normafschattingen $\mathcal{O}(m^2)$ en $\mathcal{O}(m^{3/2})$ uit stelling 2.36 en 2.39 op eenvoudige wijze aangescherpt worden (tot $\mathcal{O}(m)$ en $(m^{1/2})$)?*

De resultaten van de algoritmes uit hoofdstuk 4 doen vermoeden dat dit voor de gebruikte constructie inderdaad kan.

- *Is een ontbinding met B (of C) een normale matrix optimaal?*

We hebben overal gebruik gemaakt van de zogenaamde diagonaalconstructie waarbij B (op unitaire transformaties na) een diagonaalmatrix is. We weten ook dat er commutatormatrices zijn die kunnen worden ontbonden in niet-normale B en C (opmerking 2.40). Als aangetoond kan worden dat de diagonaalconstructie optimaal is (kleinst mogelijke $\|B\| \|C\|$) dan zou een eventuele ondergrens voor de diagonaalconstructie ook een algemeen geldende ondergrens zijn.

- *Kunnen alle niet-negatieve reëelwaardige holle A ontbonden worden zodat $\|B\| \|C\| \leq K \|A\|$ met K onafhankelijk van de matrixafmeting?*

De uitvoer van de algoritmes alsmede het feit dat dit waar is voor holle A met enkel dezelfde elementen (hoofdstuk 4) doet vermoeden dat dit inderdaad kan. Hierbij dient de matrix B in ieder geval wel afgestemd te worden op de waarden in A zodat de pathologische matrices als in figuur 4.1 geen kans hebben.

- *Wat kunnen we zeggen als we naar een andere matrixnorm kijken?*

In het artikel beweren de auteurs bijvoorbeeld dat het voor de Frobeniusnorm mogelijk is om een onder- en bovengrens van orde $(\log m)^{1/2}$ respectievelijk $(\log m)^{3/2}$ voor de constante K uit de commutatorstelling te bewijzen.

- *Hoe kan het paving vermoeden gebruikt worden om het resultaat uit het artikel te verbeteren?*

In het artikel geven de auteurs aan dat hun resultaat vereenvoudigd en verbeterd kan worden tot $\mathcal{O}\left((\log m)^k\right)$, mits het paving vermoeden waar is "with the right parameters". Wat deze voorwaarden precies zijn en hoe deze verbetering in zijn werk zou gaan vertellen ze er echter niet bij.

Bijlage A

Matlab implementatie van het algoritme

In deze bijlage is een Matlab implementatie van het algoritme uit hoofdstuk 4 opgenomen. In de code worden aannames getoetst met het `assert()` commando, waarbij afwijkingen (veroorzaakt door de eindige precieze van de berekeningen) kleiner dan 10^{-10} verwaarloost worden (een empirisch gekozen getal, afwijkingen waren typisch van de orde 10^{-12}). Tussen de algoritmes door zijn een aantal voorbeelden opgenomen die het gebruik van de verschillende routines laten zien.

A.1 Ontbindingsprocedure

Algoritme A.1 (`ontbindCommutatorReeel`).

```
1 function [ B, C ] = ontbindCommutatorReeel( A )
2 %ontbindCommutatorReeel Bereken matrices B en C zodat A = B*C - C*B
3 %   A is een spoorloze vierkante matrix
4 %   B is normaal met norm(B)=1 en eigenwaarden in [-1, 1].
5
6 assert(ismatrix(A) && ~isscalar(A), 'A moet een matrix zijn');
7 [n, m] = size(A);
8 assert(n == m, 'A moet vierkant zijn');
9 assert(abs(trace(A)) < 1e-10, 'A moet spoorloos zijn');
10
11 % Vind unitaire U zodat H = U'*A*U hol is
12 [H, U] = maakHol(A);
13
14 % Verdeel b_ii over [-1, 1]
15 Bval = linspace(-1, 1, n)';
16 Bh = diag(Bval);
17
18 % Bereken C als volgt:
19 % c_ij = a_ij / (b_ii - b_jj)
20 % c_ii = a_ii = 0
21 Ch = H ./ ( repmat(Bval, 1, n) - repmat(Bval', n, 1) + eye(n) );
```

```

22
23  assert(all(all( abs(H - Bh*Ch + Ch*Bh) < 1e-10 )));
24
25  B = U * Bh * U';
26  C = U * Ch * U';
27
28  assert(all(all( abs(A - B*C + C*B) < 1e-10 )));
29
30  end

```

Voorbeeld A.2.

```

>> A = [1 2 3
        4 5 6
        7 8 -6];

>> [B, C] = ontbindCommutatorReeel(A)

B =

    -0.8343    -0.2670     0.3099
    -0.2670    -0.0294     0.3333
     0.3099     0.3333     0.8638

C =

    -0.0925     0.5598    -0.8180
     0.8861     0.5513    -7.1573
     2.6894     7.5813    -0.4588

>> B*C-C*B

ans =

     1.0000     2.0000     3.0000
     4.0000     5.0000     6.0000
     7.0000     8.0000    -6.0000

```

Algoritme A.3 (ontbindCommutatorComplex).

```

1  function [ B, C ] = ontbindCommutatorComplex( A )
2  %ontbindCommutatorComplex Bereken matrices B en C zodat A = B*C - C*B
3  %   A is een spoorloze vierkante matrix
4  %   B is normaal met norm(B)=1 en eigenwaarden in de complexe ...
   eenheidscirkel.
5
6  assert(ismatrix(A) && ~isscalar(A), 'A moet een matrix zijn');
7  [n, m] = size(A);
8  assert(n == m, 'A moet vierkant zijn');
9  assert(abs(trace(A)) < 1e-10, 'A moet spoorloos zijn');
10
11 % Vind unitaire U zodat H = U'*A*U hol is
12 [H, U] = maakHol(A);
13

```

```

14 % Verdeel b_ii over complexe eenheidsschijf
15 Bval = vulEenheidscirkel(n).';
16 Bh = diag(Bval);
17
18 % Bereken C als volgt:
19 % c_ij = a_ij / (b_ii - b_jj)
20 % c_ii = a_ii = 0
21 Ch = H ./ ( repmat(Bval, 1, n) - repmat(Bval.', n, 1) + eye(n) );
22
23 assert(all(all( abs(H - Bh*Ch + Ch*Bh) < 1e-10 )));
24
25 B = U * Bh * U';
26 C = U * Ch * U';
27
28 assert(all(all( abs(A - B*C + C*B) < 1e-10 )));
29
30 end

```

Voorbeeld A.4.

```

>> A = [1 2 3
        4 5 6
        7 8 -6];

>> [B, C] = ontbindCommutatorComplex(A)

B =

    0.9454 + 0.1111i    0.0643 - 0.2027i   -0.1384 + 0.1715i
    0.0643 - 0.2027i    0.6839 + 0.6544i   -0.2071 + 0.1262i
   -0.1384 + 0.1715i   -0.2071 + 0.1262i    0.0778 + 0.9416i

C =

   -0.0832 + 0.1632i   -0.9121 + 0.1281i    0.4510 + 1.7963i
   -0.1642 - 1.7268i   -0.4043 - 0.6003i    8.2259 + 4.3739i
   -2.3761 - 3.6635i   -8.8049 - 4.6956i    0.4874 + 0.4371i

>> B*C-C*B

ans =

    1.0000 - 0.0000i    2.0000 - 0.0000i    3.0000 + 0.0000i
    4.0000 + 0.0000i    5.0000              6.0000 - 0.0000i
    7.0000 + 0.0000i    8.0000 - 0.0000i   -6.0000 - 0.0000i

```

De routine `vulEenheidscirkel` uit bovenstaand algoritme geeft een indeling van punten op cirkels binnen de complexe eenheidsschijf, zoals in lemma 2.38:

Algoritme A.5 (`vulEenheidscirkel`).

```

1 function [ x ] = vulEenheidscirkel(n)

```

```

2 %vulEenheidscirkel Geeft n punten binnen de eenheidscirkel in het ...
   complexe vlak
3 %   zodat de onderlinge afstand tussen de punten minimaal
4 %    $d = 1 / (\sqrt{n/2} + 1)$  is.
5
6 % Noem  $R = 1/d$  het aantal ringen
7  $R = \text{ceil}(\sqrt{n/2})$ ;
8  $d = 1/R$ ;
9
10 % Bereken stralen en hoeken
11  $k = 1:R$ ;
12  $r = k*d$ ;
13  $\text{gamma} = \pi/2 * d./r$ ;
14
15 % Plaats punten, buitenste ring eerst (zodat  $\max(\text{abs}(x))=1$ )
16  $x = \text{zeros}(1, n)$ ;
17  $\text{ind} = 1$ ;
18 for  $k = R:-1:1$ 
19     for  $j = 1:\text{floor}(2*\pi/\text{gamma}(k))$ 
20          $x(\text{ind}) = r(k)*\exp(1i * (j-1) * \text{gamma}(k))$ ;
21          $\text{ind} = \text{ind} + 1$ ;
22         if  $(\text{ind} > n)$ 
23             return;
24         end
25     end
26 end
27
28 end

```

A.2 Uitholprocedure

Algoritme A.6 (maakHol).

```

1 function [ H, U ] = maakHol( A )
2 %maakHol Bereken holle H en unitaire U zodat  $H = U'*A*U$ 
3 %   A is een spoorloze vierkante matrix
4
5 assert(ismatrix(A) && ~isscalar(A), 'A moet een matrix zijn');
6 [n, m] = size(A);
7 assert(n == m, 'A moet vierkant zijn');
8 assert(abs(trace(A)) < 1e-10, 'A moet spoorloos zijn');
9
10 % Als A al hol is zijn we snel klaar
11 if all( diag(A) < eps )
12     H = A;
13     U = eye(n);
14     return;
15 end
16
17 % We bouwen eerst een unitaire transformatie die het eerste
18 % diagonaalelement van A nul maakt. Daartoe nemen we eenheidsvector x
19 % waarvoor  $x'*A*x = 0$  en vullen we dit aan tot een orthonormale basis.
20
21 x = vindNulpuntNB(A);

```

```

22 I = eye(n);
23
24 % Gebruik Gram-Schmidt om de collectie {x, I(:,i)} uit te dunnen
25 % tot n orthonormale vectoren. Dit kan Matlab voor ons doen:
26 [Q, R] = qr([x I]);
27 U = Q;
28
29 % U is nu unitair
30 assert(all(all(U'*U - I < 1e-10)));
31
32 % Het eerste diagonaalelement is nu nul
33 H = U' * A * U;
34 assert(abs(H(1,1)) < 1e-10);
35
36 % Pas recursie toe om de rest v/d diagonaal nul te maken
37 if n > 2
38
39     Ap = H(2:n, 2:n);
40     [Hp, Up] = maakHol(Ap);
41
42     U = U * [1 zeros(1, n-1); zeros(n-1, 1) Up];
43     H = U' * A * U;
44
45 end
46
47 % H is nu hol
48 assert(all(diag(H) < 1e-10));
49
50 end

```

Voorbeeld A.7.

```

>> A = [1 2 3
        4 5 6
        7 8 -6];

>> [H, U] = maakHol(A)

H =

    0.0000    -0.0394   -6.2934
   -1.0777    -0.0000    6.1257
  -11.0797     6.2408   -0.0000

U =

   -0.9214   -0.3695    0.1207
   -0.3261    0.9037    0.2773
    0.2116   -0.2162    0.9532

>> U'*U

ans =

    1.0000   -0.0000     0

```

```

-0.0000    1.0000   -0.0000
      0   -0.0000    1.0000

>> U'*A*U

ans =

    0.0000   -0.0394   -6.2934
   -1.0777   -0.0000    6.1257
  -11.0797    6.2408   -0.0000

>> [norm(A), norm(H)]

ans =

    12.7669    12.7669

```

Algoritme A.8 (vindNulpuntNB). *algoritme*

```

1 function [ x ] = vindNulpuntNB( A )
2 %vindNulpuntNB Bereken kolomvector x zodat x'*A*x = 0 (nulpunt ...
   numeriek bereik)
3 % A is een spoorloze matrix
4
5 assert(ismatrix(A) && ~isscalar(A), 'A moet een matrix zijn');
6 [n, m] = size(A);
7 assert(n == m, 'A moet vierkant zijn');
8 assert(abs(trace(A)) < 1e-10, 'A moet spoorloos zijn');
9
10 % Bereken matrix van genormaliseerde eigenvectoren V en ...
   diagonaalvorm D
11 [V, D] = eig(A);
12 eigValues = diag(D);
13
14 % Voor spoorloze matrix is de som v/d eigenwaarden nul
15 assert(abs(sum(eigValues)) < 1e-10);
16
17 % Als nul een eigenwaarde is (tot op machine-precisie)
18 % volstaat de bijbehorende eigenvector
19 i = find(abs(eigValues) < 1e-10, 1, 'first');
20 if ~isempty(i)
21     x = V(:, i);
22     return;
23 end
24
25 % Als alle eigenwaarden niet nul zijn kunnen we gebruik maken van het
26 % numeriek bereik. Het gemiddelde van de eigenwaarden is nul (want
27 % spoorloos) en zit in het numeriek bereik (want convex).
28
29 % We berekenen het gemiddelde met bijbehorende vector stap voor stap,
30 % door steeds te interpoleren
31 % Voor defecte matrices blijft dit goed gaan omdat eig(A) dan twee
32 % maal dezelfde eigenvector bevat
33 gem = eigValues(1);
34 x = V(:, 1);
35 for i = 2:n

```

```

36     % Interpoleer tussen huidige waarde en volgende eigenwaarde
37     % NB: weegfactor
38     [x, gem] = interpNB(A, x, V(:, i), (i - 1)/i);
39 end
40
41 % Het gemiddelde zou nu nul moeten zijn
42 assert(abs(gem) < 1e-10);
43 assert(abs(x'*A*x) < 1e-10);
44
45 end

```

Algoritme A.9 (interpNB).

```

1  function [ z, c ] = interpNB( A, x, y, t )
2  %interpNB Interpoleer in het numeriek bereik van A
3  %
4  %   Bereken genormaliseerde vector z waarvoor
5  %        $z' * A * z = c = ta + (1-t)b$  in  $W(A)$ 
6  %   waarbij
7  %        $x'Ax=a$  in  $W(A)$ ,  $x$  genormaliseerd en
8  %        $y'Ay=b$  in  $W(A)$ ,  $y$  genormaliseerd en
9  %        $0 \leq t \leq 1$ .
10 %
11 %   Gebaseerd op het bewijs van de stelling van Toeplitz-Hausdorff in
12 %   'Linear Algebra in Action' van H. Dym.
13 %
14
15 assert(ismatrix(A) && ~isscalar(A), 'A moet een matrix zijn');
16 [n, m] = size(A);
17 assert(n == m, 'A moet vierkant zijn');
18 assert(all(size(x)==[n, 1]) && all(size(y)==[n, 1]));
19 assert(abs(norm(x)-1) < 1e-10, 'x moet genormaliseerd zijn');
20 assert(abs(norm(y)-1) < 1e-10, 'y moet genormaliseerd zijn');
21 assert(0 <= t && t <= 1, '0 <= t <= 1 moet gelden');
22
23 a = x'*A*x;
24 b = y'*A*y;
25 c = t*a + (1-t)*b;
26
27 % Indien a==b, t=0 of t=1 zijn we snel klaar
28 if c == a
29     z = x;
30     return;
31 end
32 if c == b
33     z = y;
34     return;
35 end
36
37 % Reduceer het probleem tot het geval waarin a=1 en b=0
38 alpha = 1/(a-b);
39 beta = -b/(a-b);
40 B = alpha*A + beta*eye(n);
41
42 z = interpNB10(B, x, y, t);
43

```

```

44 assert(abs(z'*A*z - c) < 1e-10);
45 assert(abs(norm(z)-1) < 1e-10);
46
47 end

```

Algoritme A.10 (interpNB10).

```

1 function z = interpNB10( B, x, y, lambda )
2 %interpNB Interpoleer in het numeriek bereik van B
3 %
4 % Bereken genormaliseerde vector z zodat
5 %    $z' * B * z = c = \text{lambda}$  in  $W(B)$ 
6 %   waarbij
7 %    $x' * B * x = 1$  in  $W(B)$ ,  $x$  genormaliseerd
8 %    $y' * B * y = 0$  in  $W(B)$ ,  $y$  genormaliseerd
9 %   en  $0 \leq t \leq 1$ .
10 %
11 % Gebaseerd op het bewijs van de stelling van Toeplitz-Hausdorff in
12 % 'Linear Algebra in Action' van H. Dym.
13 %
14
15 assert(ismatrix(B) && ~isscalar(B), 'B moet een matrix zijn');
16 [n, m] = size(B);
17 assert(n == m, 'B moet vierkant zijn');
18 assert(all(size(x)==[n, 1]) && all(size(y)==[n, 1]));
19 assert(abs(norm(x)-1) < 1e-10, 'x moet genormaliseerd zijn');
20 assert(abs(norm(y)-1) < 1e-10, 'y moet genormaliseerd zijn');
21 assert(abs(x'*B*x-1) < 1e-10);
22 assert(abs(y'*B*y) < 1e-10);
23 assert(0 <= lambda && lambda <= 1, '0 <= lambda <= 1 moet gelden');
24
25 % Bepaal gamma op de complexe eenheidscirkel waarvoor
26 %  $\text{gamma} * y' B x + \text{gamma}' * x' B y$  reeel is.
27
28 % Bereken het hermitische en anti-hermitische deel van B
29 % zodat  $B = C + iD$  (beide hermitische matrices)
30  $C = (B + B')/2$ ;
31  $D = (B - B')/(2i)$ ;
32
33 % Noem
34  $c = y' * C * x$ ;
35  $d = y' * D * x$ ;
36
37 % Neem
38  $\text{gamma} = 1$ ;
39 if d ~= 0
40      $\text{gamma} = i * \text{abs}(d)^{-1} * d'$ ;
41 end
42
43 % Zodat
44 assert(abs(imag(gamma * y'*B*x + gamma' * x'*B*y)) < 1e-10);
45
46 % Noem
47  $p = \text{real}(\text{gamma}*c + \text{gamma}'*c')$ ;
48  $q = \text{real}(y'*\text{gamma}*x + x'*\text{gamma}'*y)$ ;
49

```

```

50 % Bepaal t
51 syms t;
52 ts = vpasolve( (t^2 + t*(1-t)*p) / (t^2 + t*(1-t)*q + (1-t)^2) == ...
    lambda, t, [0 1] );
53 t_ = double(ts(1));
54
55 % De gezochte vector is nu
56 z = t_*gamma*x + (1-t_)*y;
57 z = z / norm(z);
58
59 assert(abs(z'*B*z - lambda) < 1e-10);
60
61 end

```

A.3 Willekeurige matrices

We kunnen willekeurige matrices genereren met één van de volgende uitdrukkingen

```

A = rand(m); % uniform op [0,1]
A = 2*rand(m)-ones(m); % uniform op [-1,1]
A = 2*round(rand(n))-ones(n); % 50-50 uit {-1,1}
A = rand(m)*exp(2*pi*i*rand(m)); % complex

```

vervolgens kunnen we deze spoorloos maken door het spoor van de diagonaalelementen af te trekken of hol door de diagonaalelementen nul te maken:

```

A = A - eye(m)*trace(A)/m; % spoorloos
A = A - diag(diag(A)); % hol

```

en vervolgens normalizeren met $A = A / \text{norm}(A)$.

Bijlage B

De commutatorstelling voor reële matrices

In het bewijs van de commutatorstelling van hoofdstuk 2 werken we met complexwaardige matrices. Een logische vraag is nu: kan een reëelwaardige commutator A ook ontbonden worden in reëelwaardige B en C ? Het volgende lemma van de heer J. Vermeer maakt dit mogelijk:

Lemma B.1. *Voor elke reëelwaardige spoorloze matrix $A \in \mathbb{R}^{m \times m}$ bestaat er een reëelwaardige eenheidsvector $y \in \mathbb{R}^m$ zodat $y^T A y = 0$.*

Bewijs. Wegens lemma 2.11 bestaat er een $x \in \mathbb{C}^m$ waarvoor $x^* A x = 0$. Schrijf $x = p + \imath q$ met $p, q \in \mathbb{R}^m$ dan geldt

$$\begin{aligned} 0 &= x^* A x \\ &= (p - \imath q)^T A (p + \imath q) \\ &= [p^T A p + q^T A q] + \imath [p^T A q - q^T A p] \end{aligned}$$

en dus

$$p^T A p + q^T A q = 0$$

Nu kunnen p en q niet beide de nulvector zijn (immers $x = p + \imath q$ is een eenheidsvector). Als $p = 0$ dan voldoet $y = q/|q|$ en als $q = 0$ voldoet $y = p/|p|$, dus neem aan dat $p \neq 0$ en $q \neq 0$. Als $p^T A p = 0$ dan voldoet $y = p/|p|$, dus neem aan $p^T A p \neq 0$. Dan ook $q^T A q \neq 0$ en bovendien hebben $p^T A p \in \mathbb{R}$ en $q^T A q \in \mathbb{R}$ tegengesteld teken. Neem nu een kromme K in $\mathbb{R}^m \setminus \{0\}$ van p naar q , dan neemt de continue functie $f : K \rightarrow \mathbb{R}$ gegeven door $k \mapsto k^T A k$ ergens een nulpunt aan. Er bestaat dus een $k_0 \in K$ waarvoor $k_0^T A k_0 = 0$ en dus voldoet $y = k_0/|k_0|$ (n.b. K vermijdt de nulvector). $\square$

Met dit lemma volgt de rest van het bewijs van de reële versie van de commutatorstelling op precies dezelfde manier als de complexwaardige versie, waarbij x^* steeds vervangen wordt door x^T en orthogonale matrices de rol van de

unitaire matrices overnemen. In het bijzonder leren we hieruit dat alle reëelwaardige spoorloze matrices equivalent zijn met een holle reëelwaardige matrix onder orthogonale transformaties.

Bibliografie

- [1] Joel Anderson. Extensions, restrictions, and representations of states on c^* -algebras. *Transactions of the American Mathematical Society*, 249(2):303, May 1979.
- [2] Harry Dym. *Linear Algebra in Action*. American Mathematical Society, 2007.
- [3] R. L. Graham, B. D. Lubachevsky, K. J. Nurmela, and P. R. J. Östergård. Dense packings of congruent circles in a circle. *Discrete Mathematics*, 181(1–3):139–154, February 1998.
- [4] Ulf Grenander and Gabor Szegő. *Toeplitz Forms and Their Applications*. University of California Press, 1958.
- [5] Nicholas J. A. Harvey. An introduction to the kadison-singer problem and the paving conjecture, July 2013.
- [6] William B. Johnson, Narutaka Ozawa, and Gideon Schechtman. A quantitative version of the commutator theorem for zero trace matrices. *Proceedings of the National Academy of Sciences*, 110(48):19251–19255, November 2013.
- [7] Richard V. Kadison and I. M. Singer. Extensions of pure states. *American Journal of Mathematics*, 81(2):383, April 1959.
- [8] W. Kahan. Only commutators have trace zero, June 1999.
- [9] Adam Marcus, Daniel A. Spielman, and Nikhil Srivastava. Interlacing families II: Mixed characteristic polynomials and the kadison-singer problem. *arXiv:1306.3969 [math]*, June 2013. arXiv: 1306.3969.
- [10] Marvin Rosenblum. On the operator equation $BX-XA=qI$. *Duke Mathematical Journal*, 23(2):263–269.
- [11] Von Kenjiro Shoda. Einige sätze über matrizen. *Japanese journal of mathematics :transactions and abstracts*, 13:361–365, 1936.
- [12] Eckard Specht. The best known packings of equal circles in a circle (complete up to $n = 2600$), June 2014.

- [13] V. S. Sunder. On trace zero matrices. *Resonance*, 7(6):14–26, June 2002.
- [14] L. Tzafriri and J. Bourgain. On a problem of kadison and singer. *Journal für die reine und angewandte Mathematik*, 420:1–44, 1991.