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

"Het effect van een anisotropiecoëfficiënt in de Poissonvergelijking op de convergentie van de CG-methode en de ICCG(0)-methode" ("The impact of an anisotropy coefficient in the Poisson equation on the convergence of the CG-method and the ICCG(0)-method")

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

WILBERT GORTER

Delft, Nederland
Augustus 2017

BSc verslag TECHNISCHE WISKUNDE

"Het effect van een anisotropiecoëfficiënt in de Poissonvergelijking op de convergentie van de CG-methode en de ICCG(0)-methode"
("The impact of an anisotropy coefficient in the Poisson equation on the convergence of the CG-method and the ICCG(0)-method")

WILBERT GORTER

Technische Universiteit Delft

Begeleider

Prof.dr.ir. C. Vuik

Overige commissieleden

Prof.dr.ir. A.W. Heemink

Dhr. B. van den Dries

Augustus, 2017

Delft

Samenvatting

Allerlei diffusieverschijnselen in de natuurkunde kunnen omschreven worden met de Poissonvergelijking

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

op een gebied Ω met te kiezen randvoorwaarden. Er zijn echter ook verschijnselen waarin er een gelaagdheid optreedt in Ω, zodat er in verticale richting minder makkelijk diffusie kan plaatsvinden. Dit leidt tot een vergelijking

$$-\frac{\partial^2 u}{\partial x^2} - \epsilon\frac{\partial^2 u}{\partial y^2} = f(x, y),$$

met $0 < \epsilon \ll 1$. We noemen dit anisotropie en ϵ is de anisotropiecoëfficiënt. Via (onder andere) de Finite Volume Method kunnen deze vergelijkingen gediscretiseerd worden, wat leidt tot een stelsel $A\mathbf{u} = \mathbf{f}$. Hierbij is A symmetrisch en positief definiet (alle eigenwaarden zijn positief). Er zijn talloze methoden om dit stelsel op te lossen. Een hiervan is de CG-methode, een iteratieve methode, met als uitbreiding de ICCG(0)-methode. De convergentie van de CG-methode blijkt af te hangen van de eigenwaarden van de matrix A. Hoe dichter de eigenwaarden bij elkaar liggen, hoe sneller de convergentie. De geschaalde eigenwaarden[1] van A liggen min of meer geconcentreerd rond 1. Het is voor een snelle convergentie van belang dat de kleinste eigenwaarden niet al te dicht bij 0 liggen. Dit leidt namelijk tot trage convergentie. Er zijn manieren om de eigenwaarden dichter bij elkaar te brengen. We noemen dat preconditionering. De ICCG(0)-methode is een vorm van preconditionering en zorgt er (onder andere) voor dat de kleinste eigenwaarden dichter bij 1 komen te liggen. In dit onderzoek wordt onderzocht wat er gebeurt met de eigenwaarden, zowel in de CG-methode als in de ICCG(0)-methode, als de anisotropiecoëfficiënt ϵ heel klein wordt. Het blijkt dat de ondergrens voor de eigenwaarden van A lineair afhankelijk is van ϵ. Dit leidt tot een zeer trage convergentie van de CG-methode in het geval ϵ klein is. Er blijkt echter ook dat preconditionering met behulp van de ICCG(0)-methode voldoende is om de ondergrens voor de eigenwaarden onafhankelijk te maken van ϵ. Dit betekent dat ook in situaties waarin anisotropie optreedt, ICCG(0) een geschikte methode is om de vergelijking $A\mathbf{u} = \mathbf{f}$ op te lossen.

[1] Met 'geschaalde eigenwaarden' bedoelen we de eigenwaarden die verkregen worden uitgaande van $\Delta x = \Delta y = 1$ in de discretisatie.

Inhoudsopgave

Voorwoord

Dit onderzoek heb ik uitgevoerd als Bacheloreindproject in het derde jaar van mijn studie BSc Technische Wiskunde aan de TU Delft. Het Bacheloreindproject is de afsluiting van de Bachelor door middel van een eigen onderzoek. Het doel hiervan is dat de student (een deel van) de in de Bachelor opgedane kennis toepast en laat zien dat hij/zij in staat is tot het doen van onderzoek.

Ik heb dit onderzoek uitgevoerd in het voorjaar van het jaar 2017. In augustus heb ik het verslag afgerond. Het onderwerp van het onderzoek behoorde tot de aangeboden onderwerpen in het jaar 2017. In de eerste fase ben ik voornamelijk bezig geweest met het bestuderen van literatuur, om kennis te vergaren met betrekking tot de CG-methode en ICCG(0)-methode, waar ik nog onbekend mee was. Hierna kon de probleemstelling vastgesteld worden en in de volgende fase ben ik vooral bezig geweest met het in Python programmeren van de methoden. Hierop volgde een fase van experimenteren met het geschreven programma en het ontwikkelen van vermoedens met betrekking tot de onderzoeksvraag. Tot slot heb ik de vermoedens theoretisch wiskundig hard gemaakt, mijn conclusies getrokken en de resultaten in verslag gezet.

Het beoogde lezerspubliek voor dit verslag zijn diegenen die voorkennis hebben van het vakgebied Numerieke Wiskunde en dan voornamelijk de Finite Volume Method. Ook is enige kennis vereist van Lineaire Algebra, in het bijzonder met betrekking tot eigenwaardenanalyse en de eigenschappen van SPD-matrices. Verder is wat algemene kennis van partiële differentiaalvergelijkingen noodzakelijk. Al met al moet het verslag op z'n minst leesbaar zijn voor derdejaars Technische Wiskundestudenten die het vak Numerieke Methoden 2 gevolgd hebben.

Tijdens mijn onderzoek ben ik begeleid door dhr. C. (Kees) Vuik. Hij heeft mij in de opstartfase de bedoeling van het onderzoek duidelijk gemaakt en mij voorzien van de benodigde literatuur met betrekking tot het onderwerp. Tijdens het proces heb ik meestal wekelijks een bespreking met hem gehad. Daarnaast heeft hij positieve en opbouwend kritische kanttekeningen geplaatst, die het onderzoek en de verslaglegging ten goede zijn gekomen. Verder is het op z'n plaats om hier dhr. J.A. (Koos) Meijerink te vermelden. Meijerink is werkzaam geweest bij Shell en heeft vanuit dat werk veel ervaring met het onderwerp van dit onderzoek. Hij heeft enkele meetings bijgewoond en mij van benodigde informatie voorzien.

Graag wil ik dhr. Vuik en dhr. Meijerink hartelijk bedanken voor hun begeleiding tijdens mijn Bacheloreindproject. Zonder hun begeleiding zou dit onderzoek zich niet hebben kunnen ontwikkelen tot wat het geworden is.

1. Inleiding

Deze inleiding heeft tot doel de natuurkundige achtergrond te beschrijven van de vergelijkingen die in dit onderzoek aan de orde zijn.

1.1. **Poissonvergelijking.** Diffusie is een veel voorkomend verschijnsel in de natuurkunde. Een voorbeeld waar sprake is van diffusie is de verdeling van de temperatuur in een warmtegeleidend voorwerp (bijvoorbeeld een ijzeren plaat). Vaak wordt de betreffende grootheid (in voorbeeld: temperatuur), aangegeven met $u(x, y)$ (in twee dimensies) of $u(x, y, z)$ (in drie dimensies). We gaan in dit onderzoek uit van twee dimensies. De differentiaalvergelijking die diffusie omschrijft in steady-state wordt gegeven door

$$-\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

gedefinieerd op een gebied $\Omega \subset \mathbb{R}^2$. Hierbij is f de bronterm. In eerder genoemd voorbeeld kan $f(x, y)$ bijvoorbeeld de warmteproductie in het punt (x, y) voorstellen. Op de rand $\partial\Omega$ van Ω gelden randvoorwaarden. Dit kunnen Dirichletvoorwaarden zijn, zoals $u(x, y) = 0$ (voor $(x, y) \in \partial\Omega$), maar ook Neumann- of Robinvoorwaarden. In dit onderzoek gaan we steeds uit van een rechthoekige Ω, met de Dirichletrandvoorwaarde $u = 0$ op de noordelijke rand (de bovenrand) en van de Neumannrandvoorwaarde $\frac{\partial u}{\partial n} = 0$ op de drie andere randen, tenzij anders aangegeven. Bovenstaande vergelijking wordt de Poissonvergelijking genoemd en kan afgekort worden geschreven als $-\Delta u = f$. Hierbij stelt $\Delta = \nabla \cdot \nabla$ de Laplace-operator voor. Andere voorbeelden waar deze vergelijking een rol speelt zijn zwaartekrachtvelden op elektrische potentialen.

1.2. **Discretisatie.** Via eindige differenties of eindige volumes kan men dergelijke differentiaalvergelijkingen discretiseren. Dit levert een stelsel vergelijkingen op van de vorm $A\mathbf{u} = \mathbf{f}$. Hierbij is de vector $\mathbf{u} = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix}^T$ de gediscretiseerde functie u op de gridpunten $\mathbf{x}_1, \dots, \mathbf{x}_n$. De vector $\mathbf{f} = \begin{bmatrix} f_1 & \dots & f_n \end{bmatrix}^T$ is dan vanzelfsprekend de gediscretiseerde functie f. De matrix A is de discretisatie van de Laplaceoperator. In dit onderzoek gebruiken we steeds de Finite Volume Method (FVM). Hoe de matrix A er dan uitziet, kunnen we in stencilnotatie weergeven, wat het volgende resultaat oplevert (voor inwendige punten):

$$\frac{1}{\Delta x \Delta y} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

We gaan in dit onderzoek om vereenvoudigingsredenen steeds uit van $\Delta x = \Delta y = 1$. Dit heeft geen wezenlijk effect op datgene dat onderzocht wordt. Het aantal gridpunten per richting noemen we m en we gaan uit van een gelijk aantal gridpunten per richting (dus het totaal aantal gridpunten is altijd m^2). Dit betekent dat een willekeurige rij van A (behorend bij een inwendig punt) er dan als volgt uitziet:

$$\begin{bmatrix} 0 & \dots & -1 & \dots & -1 & 4 & -1 & \dots & -1 & \dots & 0 \end{bmatrix}.$$

Op de noordelijke randpunten wordt de stencilnotatie als volgt:

$$\begin{bmatrix} -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Op de andere randen krijgen we het volgende stencil (we tonen alleen de westelijke rand, de zuidelijke en oostelijke werken analoog):

$$\begin{bmatrix} -1 & 0 \\ 3 & -1 \\ -1 & 0 \end{bmatrix}.$$

De matrix A is (bij gebruik van FVM) altijd symmetrisch en positief definiet (SPD).

1.3. **Anisotropie.** In sommige gevallen is er sprake van een gebied waar in horizontale richting gemakkelijk diffusie plaatsvindt, terwijl dit in verticale richting juist moeilijk gaat. Dit is het geval bij de zogenoemde poreuze media. Poreuze media bevatten zogenoemde poriën, waarin zich bijvoorbeeld water, olie of gas bevindt. Bekend zijn de zogenoemde olie- en gasreservoirs, zie hoofdstuk 1, Bear (1972). De poriën zorgen ervoor dat de vloeistoffen of gassen in horizontale richting weinig weerstand ondervinden, terwijl diffusie in verticale richting moeilijk gaat. Ook stroming van grondwater is een voorbeeld waarin zich deze situatie voordoet. Naast geologische toepassingen zijn er bijvoorbeeld ook medische toepassingen, zoals diffusie van bloed in een bot. Dergelijke situaties leiden tot een aangepaste differentiaalvergelijking

$$-\frac{\partial^2 u}{\partial x^2} - \epsilon \frac{\partial^2 u}{\partial y^2} = f(x, y),$$

waarbij ϵ positief maar klein is, dus $\epsilon > 0$ maar $\epsilon \ll 1$. We noemen ϵ de anisotropiecoëfficiënt (als deze gelijk is aan 1 krijgen we de normale Poissonvergelijking weer terug). Een kleine waarde van deze coëfficiënt geeft weer dat er in verticale richting (de y-richting) moeilijk diffusie kan plaatsvinden.

Er zijn talrijke methoden beschikbaar om de vergelijking $A\mathbf{u} = \mathbf{f}$ op te lossen. Een hiervan is de CG-methode, met als uitbreiding de ICCG(0)-methode. In dit onderzoek zullen wij onderzoeken hoe deze methoden reageren op een kleine anisotropiecoëfficiënt. Alvorens wij echter tot de eigenlijke probleemstelling kunnen komen, is het eerst nodig om uiteen te zetten wat deze methode eigenlijk inhoudt.

2. Oplossingsmethoden

Het doel van dit hoofdstuk is een inleiding te geven op de CG- en ICCG(0)-methode. Eerst zullen we deze echter in bredere context plaatsen door meer algemeen naar oplossingsmethoden voor de vergelijkging $A\mathbf{u} = \mathbf{f}$ te kijken. Vervolgens zullen we de CG-methode en daarna de ICCG(0)-methode definiëren. Ook zullen we definiëren wat in het licht van deze methoden convergentie betekent en waar de convergentie van afhankelijk is.

2.1. **Direct of iteratief.**

Er zijn twee klassen van methoden om het stelsel $A\mathbf{u} = \mathbf{f}$ op te lossen. De eerste klasse zijn de directe methoden. Feitelijk betekent dit dat de oplossing $\mathbf{u} = A^{-1}\mathbf{f}$ berekend wordt. Dit gebeurt door middel van Gauss-eliminatie. De tweede klasse methoden zijn de iteratieve methoden. Het idee van een iteratieve methode is om een rij $\{\mathbf{u}^k\}_{k\geq 0}$ van vectoren te creëren, waarbij $\mathbf{u}^k \to \mathbf{u}$ als $k \to \infty$. Hier zijn enkele belangrijke begrippen aan verbonden, die nu geïntroduceerd zullen worden. We noemen $\mathbf{u}^k$ de benadering na k iteraties. De foutvector $\mathbf{e}^k$ (Engels: error vector) na k iteratiestappen wordt gedefinieerd door $\mathbf{e}^k = \mathbf{u} - \mathbf{u}^k$, ofwel het verschil tussen de uiteindelijke oplossing en de benadering na k stappen. Merk op dat deze foutvector $\mathbf{e}^k$ in het algemeen onbekend is, immers $\mathbf{u}$ is onbekend. Het residu $\mathbf{r}^k$ na k iteratiestappen wordt gedefinieerd door $\mathbf{r}^k = \mathbf{f} - A\mathbf{u}^k$. Deze kan wél worden berekend, aangezien $\mathbf{f}, A$ en $\mathbf{u}^k$ steeds bekend zijn. Merk op $\mathbf{r}^k = \mathbf{f} - A\mathbf{u}^k = A\mathbf{u} - A\mathbf{u}^k = A(\mathbf{u} - \mathbf{u}^k) = A\mathbf{e}^k$. Op deze manier is er een verband tussen het residu en de foutvector. Het residu wordt zodoende als stopcriterium gebruikt voor de iteratieve methoden. Als het residu 'klein genoeg' is, wordt het algoritme stopgezet. Voordelen van iteratieve methoden ten opzichte van de directe methode zijn dat zij veel minder geheugen in beslag nemen tijdens het rekenproces en dat er veel meer mogelijkheden zijn om parallel te rekenen.

Klassieke iteratieve methoden gaan in het algemeen als volgt in hun werk: schrijf $A = B - N$, waarbij B inverteerbaar is. We kunnen $A\mathbf{u} = \mathbf{f}$ nu schrijven als $(B - N)\mathbf{u} = \mathbf{f}$ ofwel $B\mathbf{u} = N\mathbf{u} + \mathbf{f}$. Hierop gebaseerd definiëren we als volgt het iteratieve schema: $\mathbf{u}^{k+1} = B^{-1}N\mathbf{u}^k + B^{-1}\mathbf{f}$. Deze uitdrukking is om te schrijven tot $\mathbf{u}^{k+1} = \mathbf{u}^k + B^{-1}\mathbf{r}^k$. Twee bekende klassieke iteratieve methoden zijn de Jacobimethode en de Gauss-Seidelmethode. Deze kiezen respectievelijk $B = D$ en $B = D - E$. Hierbij is D de matrix is die op de diagonaal gelijk is aan A. De matrix E is onder de diagonaal gelijk aan A.

Naast klassieke iteratieve methoden zijn er ook andere, die gebruik maken van Krylovdeelruimten. Een voorbeeld hiervan is de Conjugate Gradient-methode (CG-methode). Omdat de ICCG(0)-methode een uitbreiding is op de CG-methode, zal de CG-methode nu eerst gedefinieerd worden.

2.2. **Conjugate Gradientmethode.**

In het vorige hoofdstuk zagen we dat in klassieke iteratieve methoden de benadering na stap $k+1$ wordt berekend via $\mathbf{u}^{k+1} = \mathbf{u}^k + B^{-1}\mathbf{r}^k$. Als we, gegeven een eerste schatting $\mathbf{u}^0$, nu $\mathbf{u}^1$ en

$\mathbf{u}^2$ uitschrijven, krijgen we:

$$\begin{aligned}
&\mathbf{u}^0, \\
\mathbf{u}^1 &= \mathbf{u}^0 + (B^{-1}\mathbf{r}^0) \\
\mathbf{u}^2 &= \mathbf{u}^1 + (B^{-1}\mathbf{r}^1) = \mathbf{u}^0 + B^{-1}\mathbf{r}^0 + B^{-1}(\mathbf{f} - A\mathbf{u}^0 - AB^{-1}\mathbf{r}^0) \\
&= \mathbf{u}^0 + 2B^{-1}\mathbf{r}^0 - B^{-1}AB^{-1}\mathbf{r}^0 \\
&\vdots
\end{aligned}$$

Dit suggereert dat

$$\mathbf{u}^k \in \mathbf{u}^0 + \text{ span}\{B^{-1}\mathbf{r}^0, B^{-1}A(B^{-1}\mathbf{r}^0), \ldots, (B^{-1}A)^{k-1}(B^{-1}\mathbf{r}^0)\}.$$

Naar aanleiding hiervan definiëren we de Krylovdeelruimte

$$K^k(A; \mathbf{r}^0) := \text{ span}\{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{k-1}\mathbf{r}^0\}.$$

Verder definiëren we het A-inwendig product[2]: $(\mathbf{y}, \mathbf{z})_A := \mathbf{y}^T A\mathbf{z}$ en de A-norm $||\mathbf{y}||_A := \sqrt{\mathbf{y}^T A\mathbf{y}}$.

Stel nu $A\mathbf{u} = \mathbf{f}$ met A symmetrisch en positief definiet. Voor het gemak gaan we vanaf nu uit van $B = I$ (zodat $B^{-1}A = A$ en $B^{-1}\mathbf{r}^0 = \mathbf{r}^0$). Stel verder $\mathbf{u}^0 = 0$ en (dus) $\mathbf{r}^0 = \mathbf{f}$. De Conjugate Gradientmethode construeert $\mathbf{u}^k$ in $K^k(A; \mathbf{r}^0)$ zodanig dat

$$||\mathbf{u} - \mathbf{u}^k||_A = \min_{\mathbf{y} \in K^k(A;\mathbf{r}^0)} ||\mathbf{u} - \mathbf{y}||_A.$$

De norm van de foutvector wordt dus geminimaliseerd over de k-Krylovdeelruimte $K^k(A; \mathbf{r}^0)$. De pseudocode voor het Conjugate Gradientalgoritme luidt als volgt:

$\mathbf{u}^0 = \mathbf{0}$	$\mathbf{r}^0 = \mathbf{f}$	initializering
for	$k = 1, 2, \ldots$ do	k is de iteratiestap
	if $k = 1$ do	
	$\quad \mathbf{p}^1 = \mathbf{r}^0$	
	else	
	$\quad \beta_k = \frac{(\mathbf{r}^{k-1})^T \mathbf{r}^{k-1}}{(\mathbf{r}^{k-2})^T \mathbf{r}^{k-2}}$	$\mathbf{p}^k$ is de richtingsvector om
	$\quad \mathbf{p}^k = \mathbf{r}^{k-1} + \beta_k \mathbf{p}^{k-1}$	$\mathbf{u}^{k-1}$ naar $\mathbf{u}^k$ te updaten
	end if	
	$\alpha_k = \frac{(\mathbf{r}^{k-1})^T \mathbf{r}^{k-1}}{(\mathbf{p}^k)^T A\mathbf{p}^k}$	
	$\mathbf{u}^k = \mathbf{u}^{k-1} + \alpha_k \mathbf{p}^k$	update iterand
	$\mathbf{r}^k = \mathbf{r}^{k-1} - \alpha_k A\mathbf{p}^k$	update residu

[2]Als A SPD is (zoals we aannemen), voldoet deze definitie aan de eisen van een inwendig product.

De volgende zaken kunnen bewezen worden:

Stelling 2.1

(1) Er geldt $\text{span}\{\mathbf{p}^1, \ldots, \mathbf{p}^k\} = \text{span}\{\mathbf{r}^0, \ldots, \mathbf{r}^{k-1}\} = K^k(A; \mathbf{r}^0)$.
(2) Er geldt $(\mathbf{r}^j)^T \mathbf{r}^i = 0$ voor $i = 0, \ldots, j-1; j = 1, \ldots, k$
(3) Er geldt $||\mathbf{u} - \mathbf{u}^k||_A = \min_{\mathbf{y} \in K^k(A;\mathbf{r}^0)} ||\mathbf{u} - \mathbf{y}||_A$.

Voor het bewijs hiervan verwijzen we, in navolging van Lahaye & Vuik (2015), naar Paragraaf 10.2 van Golub & van Loan (1996). Uit (1) en (2) volgt dat de vectoren $\mathbf{r}^0, \ldots, \mathbf{r}^{k-1}$ een orthogonale basis van $K^k(A; \mathbf{r}^0)$ vormen.

2.3. **Convergentie.** Bij goed werkende iteratieve methoden is er sprake van convergentie. Met convergentie bedoelen we informeel de 'snelheid' waarmee de benaderingen $\mathbf{u}^k$ naar de precieze oplossing $\mathbf{u}$ gaan als $k \to \infty$. Formeel kan dit gedefinieerd worden als het kleinste aantal stappen k waarvoor geldt dat de orde relatieve fout $\frac{||\mathbf{u}-\mathbf{u}^k||_2}{||\mathbf{u}||_2}$ kleiner is dan een te kiezen (geheel) getal p. Uiteraard dient deze p negatief te zijn, omdat er anders geen sprake is van nauwkeurigheid.

2.3.1. *Convergentie van de CG-methode.* Om de convergentie van de Conjugate Gradientmethode te beschouwen is er een belangrijke ongelijkheid bewezen:

$$||\mathbf{u} - \mathbf{u}^k||_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k ||\mathbf{u} - \mathbf{u}^0||_A.$$

Hierbij is $\kappa_2(A)$ het conditiegetal van A, dat gelijk is aan $\frac{\lambda_{max}}{\lambda_{min}}$, met λ_{max} de grootste eigenwaarde van A en λ_{min} de kleinste eigenwaarde van A. Als de eigenwaarden van A dicht bij elkaar liggen, dan is $\kappa_2(A) \approx 1$, dus is er snelle convergentie. Het komt echter ook voor dat de eigenwaarden verder uit elkaar liggen, waardoor $0 \leq \left(\frac{\sqrt{\kappa_2(A)}-1}{\sqrt{\kappa_2(A)}+1} \right)^k < 1$ vrij dicht bij 1 ligt, zodat de convergentie relatief langzaam is.

De bovenstaande begrenzing van de fout geldt echter alleen bij de eerste iteraties. Bij latere iteraties wordt de convergentie beter. Dit heet superlineaire convergentie. De reden hiervoor kan als volgt uiteengezet worden. Stel de $n \times n$-matrix A heeft eigenwaarden $\lambda_1, \ldots, \lambda_n$ met bijbehorende eigenvectoren $y_1, \ldots, y_n$, waarbij $||y_i||_2 = 1$ en $(y_i)^T y_j = 0$ als $i \neq j$. De fout na iteratie k kan dan geschreven worden als

$$u - u^k = \sum_{j=1}^{n} \gamma_j q_k(\lambda_j) y_j.$$

Stel nu dat $\gamma_j = 0$, dan heeft de eigenrichting y_j geen invloed meer op de fout. We definiëren nu $\alpha = \min\{\lambda_j | \gamma_j \neq 0\}$ en $\beta = \max\{\lambda_j | \gamma_j \neq 0\}$. Het quotiënt $\frac{\beta}{\alpha}$ heet het effectieve conditiegetal. In de eerder gegeven bovengrens voor de fout, kan nu het conditiegetal worden vervangen door het effectieve conditiegetal. Immers, α en β zijn respectievelijk de minimale en maximale eigenwaarde die nog daadwerkelijk invloed hebben op het proces. We krijgen

dus

$$||\mathbf{u}-\mathbf{u}^k||_A \leq 2\left(\frac{\sqrt{\frac{\beta}{\alpha}}-1}{\sqrt{\frac{\beta}{\alpha}}+1}\right)^k ||\mathbf{u}-\mathbf{u}^0||_A.$$

Dit verklaart waarom de convergentie zich superlineair gedraagt. Het effectieve conditiegetal zal namelijk na verloop van tijd kleiner worden dan het normale conditiegetal waardoor de bovengrens minder scherp wordt. Waarom dit zo is, zullen we uitleggen aan de hand van Ritzwaarden.

2.3.2. *Convergentie en Ritzwaarden.* Ritzwaarden, die we nu zullen definiëren, zijn een veelgebruikt concept bij het beschouwen van de convergentie van de CG-methode. Er geldt dat $\{\mathbf{r}^0,\ldots,\mathbf{r}^{k-1}\}$ een orthogonale basis is voor $K^k(A;\mathbf{r}^0)$. We schalen deze vectoren naar Euclidische norm 1, dus $\tilde{\mathbf{r}}^i = \frac{\mathbf{r}^i}{||\mathbf{r}^i||_2}$. Definieer R_k als de $n \times k$-matrix, die als kolommen de vectoren $\tilde{\mathbf{r}}^0,\ldots,\tilde{\mathbf{r}}^{k-1}$ heeft. We de definiëren vervolgens de Ritzmatrix T_k door $T_k = R_k^T A R_k$. Nu is T_k de projectie van A op $K^k(A;\mathbf{r}^0)$. De eigenwaarden θ_i van T_k worden Ritzwaarden van A genoemd. Als verder $T_k\mathbf{z}^i = \theta_i\mathbf{z}^i$ en $||\mathbf{z}^i||_2 = 1$, dan heet $R_k\mathbf{z}^i$ een Ritzvector van A. De Ritzwaarden en Ritzvectoren zijn een benadering voor de daadwerkelijke eigenwaarden en -vectoren van A. Naarmate k groter wordt, worden deze benaderingen beter (de Ritzwaarden en -vectoren convergeren dus naar de A-eigenwaarden en -vectoren). We kunnen vaststellen dat $R_k^T R_k = I$, waarbij I de $k \times k$-eenheidsmatrix is. In het bijzonder als $k = n$, dan geldt dat $R_n^T R_n = I$, dus $R_n^T = R_n^T R_n R_n^{-1} = R_n^{-1}$. Als nu $T_n\mathbf{z}_i = \theta_i\mathbf{z}_i$, dan krijgen we het volgende resultaat

$$\begin{aligned}
T_n\mathbf{z}_i &= \theta_i\mathbf{z}_i \\
R_n^T A R_n\mathbf{z}_i &= \theta_i\mathbf{z}_i \\
R_n R_n^T A R_n\mathbf{z}_i &= R_n\theta_i\mathbf{z}_i \\
R_n R_n^{-1} A R_n\mathbf{z}_i &= \theta_i R_n\mathbf{z}_i \\
A(R_n\mathbf{z}_i) &= \theta_i(R_n\mathbf{z}_i)
\end{aligned}$$

We zien dus dat in het geval $k = n$ de Ritzwaarden en -vectoren exact gelijk zijn aan de A-eigenwaarden en -vectoren. Wanneer $k < n$ is dit in het algemeen niet zo. Verder bestaat R_n niet altijd, want als de CG-methode stopt na $m < n$ iteraties dan is $\mathbf{r}^k = \mathbf{0}$ voor $k \geq m$, zodat er geen orthogonale basis $\{\mathbf{r}^0,\ldots,\mathbf{r}^{k-1}\}$ gemaakt kan worden voor $k > m$. Echter is wel het volgende bewezen.

- De convergentiesnelheid van een Ritzwaarde naar de bijbehorende A-eigenwaarde hangt af van de afstand van deze eigenwaarde tot de rest van het A-spectrum.
- In het algemeen convergeren de extreme Ritzwaarden het snelst en zij convergeren naar α en β.

2.3.3. *Ritzwaarden en superlineariteit.* Uit (1) en (2) van Stelling 1 volgt dat $\mathbf{r}^k$ loodrecht staat op $K^k(A,\mathbf{r}^0)$. Het is duidelijk dat voor alle Ritzwaarden $R^k\mathbf{z}^i$ dat $R^k\mathbf{z}^i \in K^k(A;\mathbf{r}^0)$. Immers, deze waarden zijn allemaal lineaire combinaties van $\mathbf{r}^0,\ldots,\mathbf{r}^k$ zodat het voorgaande volgt uit (1) van Stelling 1.

Stel nu dat de Ritzvector $R^k\mathbf{z}^j$ een goede benadering is van de eigenvector $\mathbf{y}^j$ van A, dan geldt dat $\mathbf{y}^j$ 'bijna' bevat is in $K^k(A;\mathbf{r}^0)$. Nu volgt dat $(\mathbf{r}^0)^T\mathbf{y}^j \approx 0$ ofwel $(A(\mathbf{u}-\mathbf{u}^k))^T\mathbf{y}^j \approx 0$. Hieruit volgt $\lambda_j(\mathbf{u}-\mathbf{u}^k)^T\mathbf{y}^j = A(\mathbf{u}-\mathbf{u}^k)^T\mathbf{y}^j = (A(\mathbf{u}-\mathbf{u}^k))^T\mathbf{y}^j \approx 0$. Dit betekent dat de fout $\mathbf{e}^k = \mathbf{u}-\mathbf{u}^k$ 'zo goed als loodrecht' staat op de eigenvector $\mathbf{y}^j$ van A, zodat deze fout een verwaarloosbaar kleine component heeft in de j-de eigenrichting. Dit verklaart precies de eerder genoemde superlineaire eigenschap. Deze superlineaire eigenschap zullen we later ook terugzien in de waarnemingen. We kunnen informeel zeggen, dat wanneer het CG-proces eenmaal de j-de eigenrichting 'uit de weg geruimd heeft', deze geen invloed meer heeft op de convergentie.

Het is mogelijk het bovenstaande formeel te maken, maar hiervoor verwijzen we naar Lahaye en Vuik (2015), Stelling 7.1.4.

2.4. **Preconditionering.** Zoals we hebben gezien hangt de convergentie van de CG-methode sterk af van de verdeling van de eigenwaarden van de matrix A. In veel gevallen is deze verdeling vrij slecht (dat wil zeggen, de eigenwaarden liggen ver uit elkaar). Een manier om in die gevallen de convergentie van de methode te verbeteren is preconditionering. Gegeven het stelsel $A\mathbf{u} = \mathbf{f}$ is het algemene idee van preconditionering als volgt. We vermenigvuldigen met een matrix M^{-1}, zodat we een nieuw stelsel

$$M^{-1}A\mathbf{u} = M^{-1}\mathbf{f}$$

verkrijgen. Matrix M dient zodanig gekozen te worden, dat de eigenwaarden van $M^{-1}A$ geclusterd zijn (zodat $\kappa_2(M^{-1}A) \approx 1$), maar ook zodanig dat $M^{-1}\mathbf{y}$ (met $\mathbf{y}$ een zekere vector) gemakkelijk uit te rekenen is.

2.4.1. *Precoditionering van de CG-methode.* We gaan nu in op preconditionering van de CG-methode. We definiëren een matrix P zodanig dat $M = PP^T$. Nu is $M^{-1}A = P^{-T}P^{-1}A$. We schrijven nu $P^{-1}AP^{-T}\tilde{\mathbf{u}} = P^{-1}\mathbf{f}$. Er geldt daarbij dat $\mathbf{u} = P^{-T}\tilde{\mathbf{u}}$. We lossen nu met behulp van de CG-methode $\tilde{\mathbf{u}}$ op uit dit stelsel. Vervolgens is uiteraard nog vermenigvuldiging met P^{-T} nodig om $\mathbf{u}$ te verkrijgen. Het algoritme wat zodoende verkregen wordt kan ook meteen in termen van $\mathbf{u}$ opgeschreven worden en dit noemen de Preconditioned Conjugate Gradient-methode (PCG). Het algoritme luidt als volgt:

$\mathbf{u}^0$	$\mathbf{r}^0 = \mathbf{f}$	initializering
for	$k = 1, 2, \ldots$ do	k is de iteratiestap
	$\mathbf{z}^{k-1} = M^{-1}\mathbf{r}^{k-1}$	
	if $k = 1$ do	
	$\mathbf{p}^1 = \mathbf{r}^0$	
	else	
	$\beta_k = \frac{(\mathbf{r}^{k-1})^T\mathbf{z}^{k-1}}{(\mathbf{r}^{k-2})^T\mathbf{z}^{k-2}}$	$\mathbf{p}^k$ is de richtingsvector om
	$\mathbf{p}^k = \mathbf{r}^{k-1} + \beta_k\mathbf{p}^{k-1}$	$\mathbf{u}^{k-1}$ naar $\mathbf{u}^k$ te updaten
	end if	
	$\alpha_k = \frac{(\mathbf{r}^{k-1})^T\mathbf{z}^{k-1}}{(\mathbf{p}^k)^TA\mathbf{p}^k}$	
	$\mathbf{u}^k = \mathbf{u}^{k-1} + \alpha_k\mathbf{p}^k$	update iterand
	$\mathbf{r}^k = \mathbf{r}^{k-1} - \alpha_kA\mathbf{p}^k$	update residu

Bij het toepassen van de CG-methode op een stelsel $A\mathbf{u} = \mathbf{f}$ hangt de convergentie af van het conditiegetal van A. Omdat in het PCG-algoritme wordt

$P^{-1}AP^{-T}\tilde{\mathbf{u}} = P^{-1}\mathbf{f}$ opgelost wordt, hangt hier de convergentie af van het conditiegetal van $P^{-1}AP^{-T}$. Precies geformuleerd geldt er

$$||\mathbf{u} - \mathbf{u}^k||_A \leq 2\left(\frac{\sqrt{\kappa_2(P^{-1}AP^{-T})} - 1}{\sqrt{\kappa_2(P^{-1}AP^{-T})} + 1}\right)^k ||\mathbf{u} - \mathbf{u}^0||_A.$$

Het is duidelijk dat de eigenwaarden van $P^{-T}P^{-1}A$ gelijk zijn aan de eigenwaarden van $P^{-1}AP^{-T}$. Immers, stel $P^{-T}P^{-1}A\mathbf{v} = \lambda\mathbf{v}$. Dan kunnen we het volgende afleiden:

$$\begin{aligned} P^{-T}P^{-1}A\mathbf{v} &= \lambda\mathbf{v} \\ P^{T}P^{-T}P^{-1}A\mathbf{v} &= P^T\lambda\mathbf{v} \\ P^{-1}A\mathbf{v} &= \lambda P^T\mathbf{v} \\ P^{-1}AP^{-T}(P^T\mathbf{v}) &= \lambda(P^T\mathbf{v}). \end{aligned}$$

Als het gaat om het conditiegetal, zijn de genoemde drie matrices dus uitwisselbaar.

2.4.2. *Incomplete Choleskydecompositie.* Er zijn verschillende manieren om de preconditioner $M = PP^T$ te kiezen. Eén van deze manieren staat bekend als de Incomplete Choleskydecompositie (ICCG-methoden). Een Cholesky decompositie is een onderdriekhoeksmatrix L, zodanig dat $A = LL^T$. Kiezen we nu $P = L$ ofwel $M = LL^T$, dan is $M^{-1} = A^{-1}$. Het is echter niet de bedoeling om de originele Choleskydecompositie als preconditioner te nemen. Immers, dan zou namelijk $M^{-1}A = A^{-1}A = I$, zodat de preconditionering direct de oplossing zou geven. Dit wilden we juist voorkomen, omdat het berekenen van de inverse matrix (directe oplossing) bij grote stelsels te veel geheugen vereist en veel rekentijd kost (zie paragraaf 2.1). Daarom wordt in de ICCG-methoden een iets aangepaste L gekozen, die lijkt op een Choleskydecompositie, maar niet helemaal hetzelfde is. Dit noemen we een Incomplete Choleskydecompositie. Er zijn meerdere Incomplete Choleskydecomposities denkbaar. Deze worden zo gekozen dat inversen met (relatief) weinig geheugen en rekentijd berekend kunnen worden. Een hiervan wordt gebruikt in de ICCG(0)-methode, waarbij L als volgt bepaald wordt. Schrijf matrix A als

$$A = \begin{bmatrix} a_1 & b_1 & & c_1 & & & & \\ b_1 & a_2 & b_2 & & c_2 & & & \\ & \ddots & \ddots & \ddots & & \ddots & & \\ c_1 & & b_m & a_{m+1} & b_{m+1} & & c_{m+1} & \\ & \ddots & & \ddots & \ddots & \ddots & & \ddots \\ & & & & & & & \end{bmatrix}.$$

Methode ICCG(0) definieert nu

$$L = \begin{bmatrix} \tilde{d}_1 & & & & & \\ b_1 & \tilde{d}_2 & & & & \\ & \ddots & \ddots & & & \\ c_1 & & b_m & \tilde{d}_{m+1} & & \\ & \ddots & & \ddots & \ddots & \\ & & & & & \end{bmatrix},$$

waarbij $\tilde{d}_i = a_i - \frac{b_{i-1}^2}{\tilde{d}_{i-1}} - \frac{c_{i-m}^2}{\tilde{d}_{i-m}}$ voor $i > m$ en $\tilde{d}_i = a_i - \frac{b_{i-1}^2}{\tilde{d}_{i-1}}$ voor $1 < i \leq m$ en $\tilde{d}_1 = a_1$. Het blijkt dat preconditionering met behulp van de ICCG(0)-methode duidelijke effect heeft op de convergentie van het proces. Nemen we bijvoorbeeld $\epsilon = 1$ (geen anisotropie) en $\mathbf{f} = \begin{bmatrix} 1 & \dots & 0 \end{bmatrix}^T$. Verder nemen we een stopcriterium van 10^{-10}, wat wil zeggen de het proces afgebroken wordt als $\frac{||\mathbf{r}_k||}{||\mathbf{f}||} < 10^{-10}$. We nemen $m = 60$. De CG-methode heeft dan 360 stappen nodig alvorens het stopcriterium bereikt is, terwijl de ICCG(0)-methode slechts 105 stappen nodig heeft. Verder tonnen Figuur 2 en Figuur 3 in paragraaf 4.2 de eigenwaarden van zowel A als $M^{-1}A$. Hoewel ook $M^{-1}A$ nog een wat kleinere eigenwaarde overhoudt, is duidelijk te zien dat de eigenwaarden toch meer geconcentreerd rond 1 liggen dan bij A.

3. Probleemstelling

In dit onderzoek gaat het erom wat de invloed is van anisotropie op de convergentie van de CG- en de ICCG(0)-methode. Het is gemakkelijk in te zien dat een kleine ϵ zorgt voor kleine eigenwaarden van A. Als we ϵ verwerken in ons stencil, dan krijgen we het volgende stencil (voor inwendige punten):

$$\begin{bmatrix} 0 & -\epsilon & 0 \\ -1 & 2+2\epsilon & -1 \\ 0 & -\epsilon & 0 \end{bmatrix}.$$

Een willekeurige rij van A, behorend bij een inwendig punt ziet er dan als volgt uit:

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & -1 & 2+2\epsilon & -1 & \dots & -\epsilon & \dots & 0 \end{bmatrix}.$$

Vanwege de kleine waarde van ϵ, is het waarschijnlijk dat A kleine eigenwaarden heeft. Juist ook vanuit natuurkundig perspectief rijst dit vermoeden. Wegens de anisotropie er moeilijk diffusie plaats vindt in verticale richting. Samen met de randvoorwaarden zoals gegeven in hoofdstuk 1 creëert dit dan een bijna gesloten systeem (omdat door de oostelijke en westelijke rand geen stroming is). Vervolgens is het nu de vraag of de kleine ϵ ook doorwerkt in de eigenwaarden van $M^{-1}A$ bij de ICCG(0)-methode en zo ja, in welke mate. Dit brengt ons tot de volgende onderzoeksvraag:

Wat is de invloed van een kleine anisotropicoëfficiënt in de Poissonvergelijking (met de gegeven randvoorwaarden) op de convergentie van de CG- en de ICCG(0)-methode?

De verdere opbouw van het verslag is als volgt: eerst zullen er een aantal waarnemingen gedaan worden door het CG- en ICCG(0)-proces uit te voeren voor verschillende waarden van ϵ. Met behulp van deze waarnemingen kunnen we een hypothese opstellen. Vervolgens zullen er middels eigenwaardenanalyses ondergrenzen voor de eigenwaarden opgesteld worden, om de hypothese hard te maken. Hieruit kunnen we dan conclusies trekken met betrekking tot de convergentie van het CG- en ICCG(0)-proces.

4. Waarnemingen

In dit hoofdstuk wordt voor diverse situaties bekeken hoe de CG- en ICCG(0)-methode reageren. Ook zullen we voor een aantal matrices de eigenwaarden berekenen. Voordat we naar de eigenlijke waarnemingen kijken, geven we eerst een verantwoording van de gebruikte werkwijze in de berekeningen.

4.1. **Verantwoording van de werkwijze.** De berekeningen worden uitgevoerd met behulp van een Pythonprogramma, waarvan de code in de bijlage te vinden is. Aan de grondslag van de berekeningen ligt de vergelijking

$$-\frac{\partial^2 u}{\partial x^2} - \epsilon \frac{\partial^2 u}{\partial y^2} = f(x, y).$$

We nemen $\mathbf{f}$ vast, namelijk $\mathbf{f} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$, om waarnemingen eerlijk te kunnen vergelijken. Dat deze keuze voor $\mathbf{f}$ verantwoord is, blijkt als volgt. Wanneer we $\mathbf{f}$ random laten genereren door Python en hier de CG- en ICCG(0)-methoden op loslaten, verschilt het aantal iteraties dat nodig is alvorens het stopcriterium bereikt is, voor de random gekozen $\mathbf{f}$ meestal weinig van de vaste $\mathbf{f}$.

We gebruiken voor de discretisatie een horizontale nummering. Om te laten zien dat er geen verschil is tussen horizontale en verticale nummering, nemen we de vergelijking

$$-\epsilon_1 u_{xx} - \epsilon_2 u_{yy} = f$$

en vergelijken het benodigde aantal iteraties voor $\epsilon_1 = 1$ en $0 < \epsilon_2 \ll 1$ met het benodigde aantal voor $0 < \epsilon_1 \ll 1$ en $\epsilon_2 = 1$. De aantallen blijken hierbij nauwelijks verschil te maken, waaruit we kunnen concluderen dat het verantwoord is om uit te gaan van horizontale nummering.

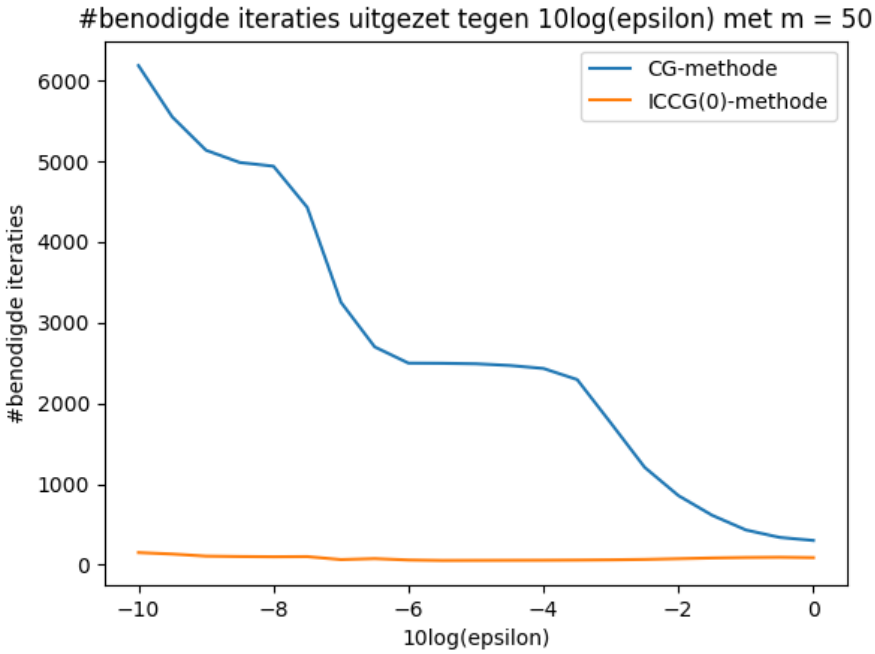
Een ander punt waar we aandacht aan moeten besteden is het verschil tussen het stopcriterium en de relatieve fout. Voor het stopcriterium gebruiken we het quotiënt $\frac{||\mathbf{r}_k||_2}{||\mathbf{f}||_2}$. Dit is echter niet hetzelfde als de relatieve fout $\frac{||\mathbf{e}_k||}{||\mathbf{u}||}$. Daarom berekenen we telkens na het uitvoeren van het proces, de relatieve fout, om te controleren dat deze niet te groot is. Daar echter bij het exact bepalen van $\mathbf{u}$ de inverse A^{-1} berekend moet worden, en deze berekening onnauwkeurig kan worden als A hele kleine eigenwaarden heeft, is er nog een andere validatie nodig. Dit doen we door Python een aantal keer een random $\mathbf{u}$ te laten generen, vervolgens $\mathbf{f} = A\mathbf{u}$ te bereken, en uit deze $\mathbf{f}$ met de CG- en ICCG(0)-methode een $\mathbf{u}_k$ te berekenen. Nu is $\mathbf{u}$ exact bekend en kan de relatieve fout betrouwbaar berekend worden. Uit deze werkwijze is gebleken dat wanneer we als stopcriterium nemen dat het eerder genoemde quotiënt kleiner is dan 10^{-10}, er geen gevaar bestaat voor de relatieve fout. Zie voor een uitwerking hiervan bijlage 1. Zodoende zullen we vanaf nu steeds dit stopcriterium gebruiken.

Om de rekentijd te bekorten, zullen we steeds bij het uitvoeren van het proces $m = 50$ gebruiken, maar voor het berekenen van de eigenwaarden $m = 30$.

4.2. **Enkele resultaten voor verschillende waarden.** We bekijken nu de resultaten voor een aantal waarden van ϵ. [3]

ϵ	# iteraties CG	# iteraties ICCG(0)
10^{-8}	4990	99
10^{-6}	2499	59
10^{-4}	2433	56
10^{-2}	857	75
1	301	88

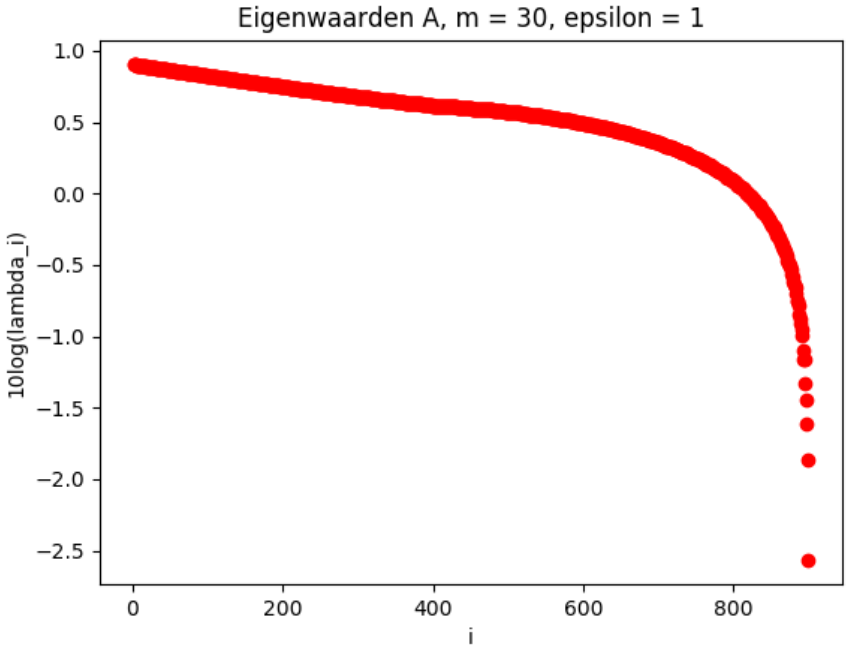
We zien dat het benodigde aantal iteraties in de CG-methode explodeert als ϵ kleiner wordt, terwijl het aantal iteraties dat nodig is in de ICCG(0)-methode weliswaar iets varieert, maar toch redelijk constant blijft. We kunnen dit ook grafisch weergeven, waarbij we duidelijk hetzelfde verloop zien, zie Figuur 1 op de volgende pagina. Aan de hand van deze resultaten stellen



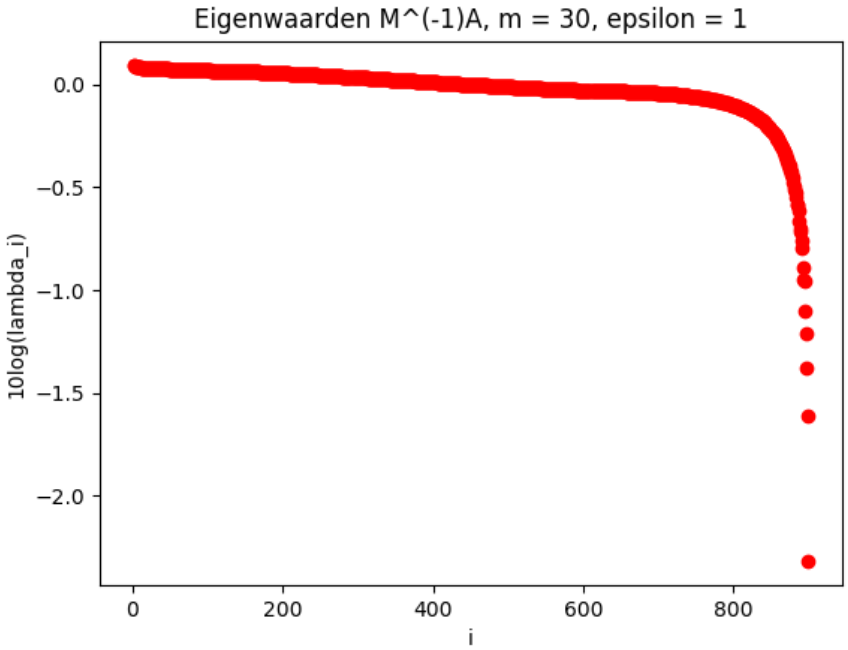
FIGUUR 1. Aantal benodigde iteraties uitgezet tegen $^{10}\log(\epsilon)$

we de hypothese dat een kleine ϵ de convergentie van de CG-methode vertraagt, maar niet de convergentie van de ICCG(0)-methode. Vertaald naar eigenwaarden betekent deze hypothese dat de kleinste eigenwaarden van A klein worden als ϵ verkleind wordt, terwijl de eigenwaarden van $M^{-1}A$ niet klein worden. We plotten de eigenwaarden van A en $M^{-1}A$ en zien dit beeld bevestigd (zie Figuren 2 t/m 5 op de volgende pagina's).

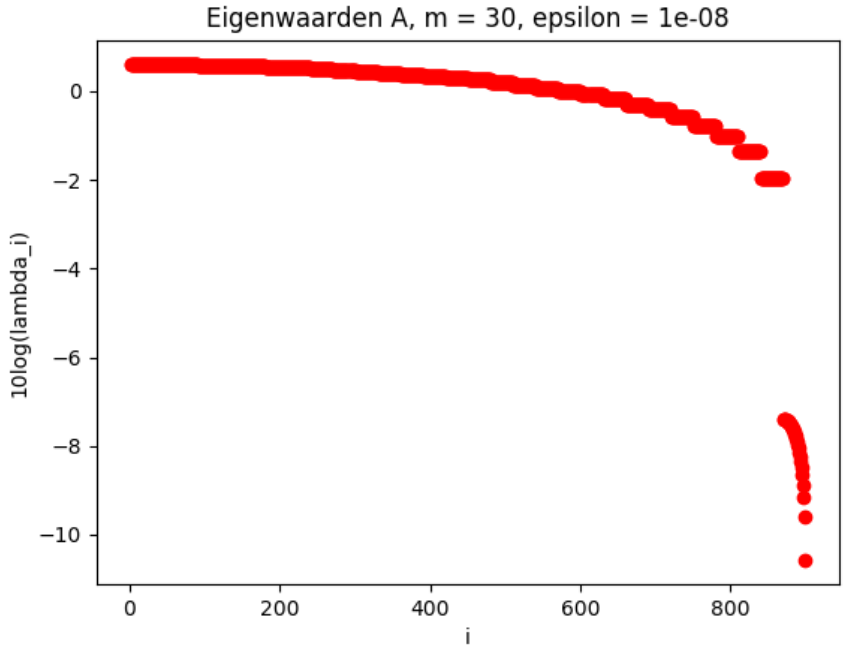
[3] In theorie zou het proces altijd maximaal $50^2 = 2500$ iteraties moeten kosten. Het feit dat bij gebruik van de CG-methode bij kleine ϵ het aantal iteraties meer is, wordt veroorzaakt door afrondfouten.



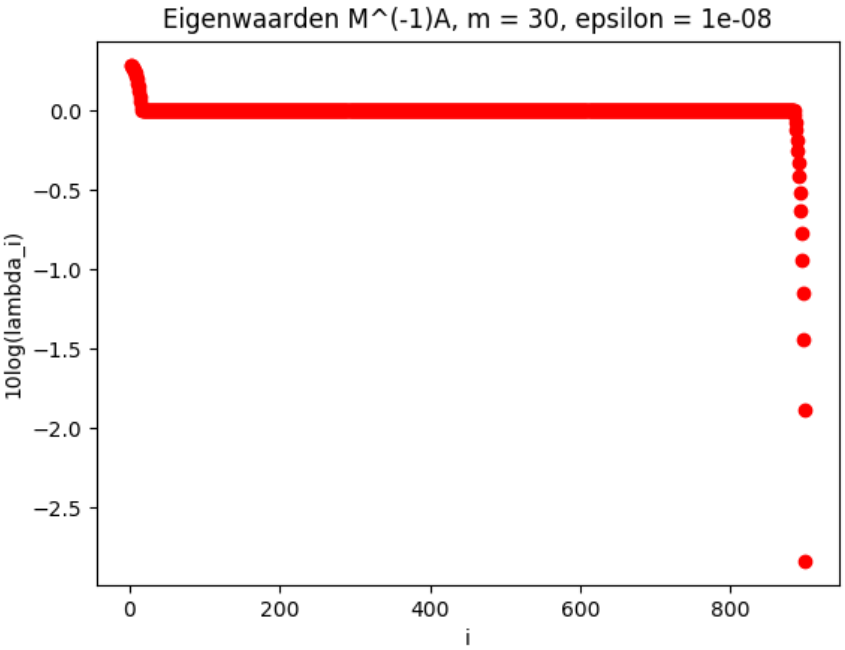
FIGUUR 2. Eigenwaarden van A met $\epsilon = 1$



FIGUUR 3. Eigenwaarden van $M^{-1}A$ met $\epsilon = 1$

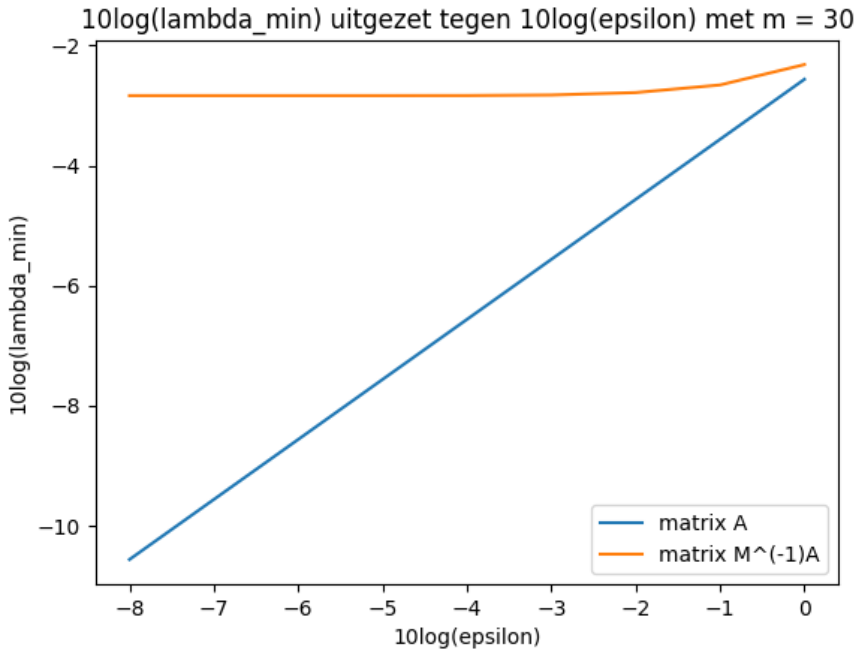


FIGUUR 4. Eigenwaarden van A met $\epsilon = 10^{-8}$



FIGUUR 5. Eigenwaarden van $M^{-1}A$ met $\epsilon = 10^{-8}$

We zien duidelijk dat er een aantal eigenwaarden van A zeer klein wordt (kleiner dan orde -7) als we $\epsilon = 10^{-8}$ kiezen, terwijl de kleinste eigenwaarden van $M^{-1}A$ ongeveer even groot blijven. Nog beter zien we dit, als we in een grafiek de kleinste eigenwaarden uitzetten tegen ${}^{10}\log(\epsilon)$, zie Figuur 6 (we noemen de kleinste eigenwaarden λ_{min}). We kunnen met het resultaat



FIGUUR 6. ${}^{10}\log(\lambda_{min})$ uitgezet tegen ${}^{10}\log(\epsilon)$

uit Figuur 6 onze hypothese nog wat specifieker maken. We kunnen als hypothese stellen dat de kleinste eigenwaarde van A lineair afhankelijk is van ϵ en dat de kleinste eigenwaarde van $M^{-1}A$ (zo goed als) onafhankelijk is van ϵ. De komende hoofdstukken zal blijken dat we deze hypothese wiskundig kunnen bewijzen en zodoende als hard resultaat kunnen stellen.

5. Eigenwaardenanalyses (1)

In dit hoofdstuk wordt met analyse van eigenwaarden aangetoond dat een kleine ϵ ervoor zorgt dat de matrix A kleine eigenwaarden kan krijgen, maar dat dit voor de matrix $M^{-1}A$ niet het geval is. We zullen aantonen dat $M^{-1}A$ weliswaar in bepaalde gevallen kleine eigenwaarden kan krijgen, maar dat dit niet veroorzaakt wordt door de kleine ϵ. We bouwen de analyse in een paar stappen op. Alvorens we de matrix A beschouwen die bij de Poissonvergelijking hoort, kijken we eerst naar de A die voortkomt uit een iets aangepast vergelijking, namelijk

$$\delta u - \frac{\partial^2 u}{\partial x^2} - \epsilon \frac{\partial^2 u}{\partial y^2} = f(x, y). \tag{1}$$

Hierbij is $0 < \delta << 1$, en we laten op alle randen (dus ook op de noordelijke rand!) de Neumannvoorwaarde $\partial u = 0$ gelden. Er is een tweevoudige reden om eerst naar deze vergelijking te kijken.

(1) Bij gebruik van de Poissonvergelijking (met een Dirichletvoorwaarde op één rand), is er een Von Neumannanalyse nodig om aan te tonen dat A geen nuleigenwaarden heeft (ofwel niet singulier is). Wanneer we echter bovenstaande δu invoeren, is de Stelling van Gershgorin al voldoende om dit aan te tonen. Dit maakt de eigenwaardenanalyse eenvoudiger en omdat de Poissonvergelijking overeenkomt met de limiet $\delta \to 0$, is dit een goede opstap naar ons doel. Behalve dat het een stap in de beoogde richting is, verschaft het ook meer inzicht in het hoe en waarom van de uiteindelijke resultaten.
(2) Niet alleen de Poissonvergelijking, maar ook vergelijking (1) komt in de praktijk voor, zodat met deze tussenstap de scope van het onderzoek vergroot wordt.

We voeren dus eerst een analyse uit voor deze vergelijking. We zullen eerst naar de eigenwaarden van A kijken en vervolgens naar de eigenwaarden van $M^{-1}A$. Vervolgens zullen we met een Von Neumannanalyse de analyse uitbreiden naar de beoogde Poissonvergelijking. Dit zullen we in het volgende hoofdstuk doen. In het huidige hoofdstuk hebben A en M dus steeds betrekking op vergelijking (1).

5.1. **De matrix A.** Voor het uitvoeren van eigenwaardenanalyses is het van belang te weten hoe de matrix A er precies uitziet. Het stencil voor een inwendig punt ziet er voor de vergelijking (1) als volgt uit:

$$\begin{bmatrix} 0 & -\epsilon & 0 \\ -1 & \delta + 2 + 2\epsilon & -1 \\ 0 & -\epsilon & 0 \end{bmatrix}.$$

Dit betekent dat een willekeurige (bij een inwendig punt behorende) rij van A er als volgt uitziet:

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & -1 & \delta + 2 + 2\epsilon & -1 & \dots & -\epsilon & \dots & 0 \end{bmatrix}.$$

De rijen behorend bij de zuidelijke randpunten (de eerste m rijen) zien er als volgt uit (niet uitgaande van een hoekpunt):

$$\begin{bmatrix} 0 & \dots & -1 & \delta + 2 + \epsilon & -1 & \dots & -\epsilon & \dots & 0 \end{bmatrix}.$$

De rijen behorend bij de noordelijke randpunten (de laatste m rijen) zien er als volgt uit (niet uitgaande van een hoekpunt):

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & -1 & \delta+2+\epsilon & -1 & \dots & 0 \end{bmatrix}.$$

De rijen behorend bij de oostelijke randpunten (de rijen $0 \mod m$) zien er als volgt uit (niet uitgaande van een hoekpunt):

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & -1 & \delta+1+2\epsilon & 0 & \dots & -\epsilon & \dots & 0 \end{bmatrix}.$$

De rijen behorend bij de westelijke randpunten (de rijen $1 \mod m$) zien er als volgt uit (niet uitgaande van een hoekpunt):

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & 0 & \delta+1+2\epsilon & -1 & \dots & -\epsilon & \dots & 0 \end{bmatrix}.$$

De rijen behorend bij een hoekpunt zijn combinaties van deze.

5.2. **De eigenwaarden van de matrix A.** Uit de Stelling van Gershgorin volgt dat voor eigenwaarden μ van A geldt

$$\begin{aligned} |\mu - 2 - 2\epsilon - \delta| = |\mu - a_{ii}| &\leq \sum_{j \neq i} |a_{ij}| = 2 + 2\epsilon \\ -2 - 2\epsilon \leq \mu - \delta - 2 - 2\epsilon &\leq 2 + 2\epsilon \\ \delta \leq \mu &\leq \delta + 4 + 4\epsilon \end{aligned}$$

Hetzelfde resultaat verkrijgen we voor rand- en hoekpunten. Omdat dit gemakkelijk is na te gaan, schrijven we dit hier niet uit. We concluderen nu dat voor de minimale eigenwaarde μ_{min} van A geldt dat $\mu_{min} \geq \delta$. We zien hier dat, dankzij de δ, in deze situatie ϵ nooit zo klein kan zijn dat de minimale eigenwaarde van A willekeurig klein kan worden. De verwachting is dan ook, dat ϵ ook op $M^{-1}A$ geen desatreuze werking kan hebben. Dit zullen we in de volgende paragraaf aantonen.

5.3. **De eigenwaarden van de matrix $M^{-1}A$.** In deze paragraaf wordt niet alleen aangetoond dat ϵ geen invloed heeft op $M^{-1}A$, maar ook zal het verband tussen A en $M^{-1}A$, wat ook weer bruikbaar zal zijn bij de analyse van de Poissonvergelijking. Er volgt nu eerst een kort intermezzo over normen.

5.3.1. *Matrixnormen.* Vanaf nu bedoelen we met $||\cdot||$ steeds de 2-norm $||\cdot||_2$. We zullen nu eerst de matrixnorm definiëren.

Definitie. De matrixnorm $||B||$ van een matrix B wordt gedefinieerd als $||B|| = \max\{||B\mathbf{x}|| : ||\mathbf{x}|| = 1\}$.

We kunnen nu direct een paar stellingen bewijzen.

Stelling 5.1

(1) Voor willekeurige matrix B en willekeurige vector $\mathbf{y}$ geldt $||B\mathbf{y}|| \leq ||B|| \cdot \mathbf{y}$;
(2) Voor willekeurige matrices B_1 en B_2 geldt $||B_1 B_2|| \leq ||B_1|| \cdot ||B_2||$.

(3) Definieer $S = \{|\lambda| : \lambda \text{ is een eigenwaarde van } B\}$. Zij $\lambda_{max} = \max S$ en $\lambda_{min} = \min S$. Dus λ_{max} en λ_{min} zijn respectievelijk de maximale en minimale waarde van de absolute waarden van de eigenwaarden van B.[4] Als B alleen reële eigenwaarden heeft, dan geldt dat $\lambda_{max} \leq ||B||$ en $\frac{1}{\lambda_{min}} \leq ||B^{-1}||$. Als B symmetrisch is, geldt in beide gevallen gelijkheid.

Bewijs. We bewijzen eerst (1) en gebruiken dan (1) om (2) en (3) te bewijzen.

(1) Er geldt

$$||B\mathbf{y}|| = \left|\left|B\left(\frac{||\mathbf{y}||}{||\mathbf{y}||}\mathbf{y}\right)\right|\right| = ||\mathbf{y}|| \cdot \left|\left|B\left(\frac{1}{||\mathbf{y}||}\mathbf{y}\right)\right|\right| \leq ||\mathbf{y}|| \cdot ||B|| = ||B|| \cdot ||\mathbf{y}||.$$

(2) Stel $||\mathbf{x}|| = 1$, dan

$$||B_1B_2\mathbf{x}|| \leq ||B_1|| \cdot ||B_2\mathbf{x}|| \leq ||B_1|| \cdot ||B_2|| \cdot ||\mathbf{x}|| = ||B_1|| \cdot ||B_2||.$$

Dit geldt voor willekeurige $\mathbf{x}$ met $||\mathbf{x}|| = 1$, dus ook

$$||B_1B_2|| = \max\{||B_1B_2\mathbf{x}||\} \leq ||B_1|| \cdot ||B_2||.$$

(3) Zij $\mathbf{v}$ een eigenvector ($\mathbf{v} \neq \mathbf{0}$) die behoort bij die eigenwaarde uit S, die in absolute waarde gelijk is aan λ_{max}. Dan is

$$\lambda_{max}||\mathbf{v}|| = ||\lambda_{max}\mathbf{v}|| = ||B\mathbf{v}|| \leq ||B|| \cdot ||\mathbf{v}||,$$

zodat $\lambda_{max} \leq ||B||$. Verder, als en slechts als λ een eigenwaarde is van B met eigenvector $\mathbf{w}$, dan is

$$B^{-1}\mathbf{w} = B^{-1}\left(\frac{\lambda}{\lambda}\mathbf{w}\right) = \frac{1}{\lambda}B^{-1}(\lambda\mathbf{w}) = \frac{1}{\lambda}B^{-1}B\mathbf{w} = \frac{1}{\lambda}\mathbf{w}.$$

Zodoende is $\frac{1}{\lambda_{min}}$ het maximum van de absolute waarden van de eigenwaarden van B^{-1}, zodat $\frac{1}{\lambda_{min}} \leq ||B^{-1}||$. Als B symmetrisch is, dan is B orthogonaal diagonliseerbaar. Stel de eigenparen van B zijn $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_n, \mathbf{v}_n)$ met $||\mathbf{v}_i|| = 1$. De $\mathbf{v}_i$ vormen nu de eigenbasis en we kunnen iedere $\mathbf{v} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$ schrijven als $\begin{bmatrix} c_1 & \ldots & c_n \end{bmatrix}^T$ waarbij $c_1^2 + \ldots + c_n^2 = ||\mathbf{v}||$. Als nu $||\mathbf{v}|| = 1$, dan geldt $B\mathbf{v} = \begin{bmatrix} \lambda_1c_1 & \ldots & \lambda_nc_n \end{bmatrix}^T$. Er volgt dat

$$\begin{aligned} ||B\mathbf{v}|| &= \sqrt{(\lambda_1c_1)^2 + \ldots + (\lambda_nc_n)^2} \\ &\leq \sqrt{(\lambda_{max})^2(c_1^2 + \ldots + c_n^2)} = \lambda_{max} \cdot \sqrt{1} = \lambda_{max}. \end{aligned}$$

Omdat $||B|| = ||B\mathbf{v}||$ voor zekere $\mathbf{v}$ met $||\mathbf{v}|| = 1$, dus $||B|| = \lambda_{max}$ zodat gelijkheid optreedt. Gelijkheid in het tweede geval volgt direct uit het eerste geval.

□

[4] Merk op dat S een verzameling van absolute waarden is en dus alleen niet-negatieve elementen bevat. Dit betekent dat λ_{max} en λ_{min} dus altijd niet-negatief zijn. Als B SPD is, zijn alle eigenwaarden positief zodat S gewoon de verzameling eigenwaarden is. In dat geval zijn λ_{max} en λ_{min} letterlijk de maximale en minimale eigenwaarde van B.

5.3.2. *Het verband tussen de eigenwaarden van de matrices A en $M^{-1}A$.* Er geldt bij toepassing van ICCG(0) dat $A = M - R$. In onze eigenwaardenanalyse willen we kijken naar de kleinste eigenwaarde van $M^{-1}A$. Uit Stelling 5.1.3 weten we dat voor de minimale eigenwaarde λ_{min} van $M^{-1}A$ geldt dat $\frac{1}{\lambda_{min}} \leq ||(M^{-1}A)^{-1}|| = ||A^{-1}M||$, ofwel

$$\lambda_{min} \geq \frac{1}{||A^{-1}M||}.$$

We richten ons daarom nu op het bepalen van een bovengrens voor $||A^{-1}M||$, ten einde een ondergrens voor λ_{min} te verkrijgen. Met $M = A + R$ en met behulp van de driehoeksongelijkheid vinden we

$$\begin{aligned} ||A^{-1}M|| &= ||A^{-1}(A+R)|| = ||I + A^{-1}R|| \leq ||I|| + ||A^{-1}R|| \\ &= 1 + ||A^{-1}R|| \\ &\leq 1 + ||A^{-1}|| \cdot ||R||. \end{aligned}$$

Uit Stelling 5.1.3 volgt dat $||A^{-1}|| = \frac{1}{\mu_{min}}$, ofwel

$$||A^{-1}M|| \leq 1 + \frac{1}{\mu_{min}}||R||.$$

We krijgen dus het volgende verband tussen de eigenwaarden μ van A en λ van $M^{-1}A$:

$$\lambda_{min} \geq \frac{1}{1 + \frac{1}{\mu_{min}}||R||}.$$

Dit verband kunnen we ook weer gebruiken in hoofdstuk 6. In paragraaf 5.1 zagen we dat $\mu_{min} \geq \delta$. We krijgen dus

$$\lambda_{min} \geq \frac{1}{1 + \frac{1}{\delta}||R||}.$$

Nu moeten we echter nog bepalen wat $||R||$ is.

5.3.3. *Het berekenen van een bovengrens voor de maximale eigenwaarde van de matrix R.* We weten dat $R = M - A$. Omdat M en A beide symmetrisch zijn[5], is ook R symmetrisch. Er geldt dus $||R|| = \xi_{max}$ met ξ_{max} de maximale eigenwaarde van R. Het doel van deze subparagraaf is om ξ_{max} te bepalen.

Voordat we de Stelling van Gershgorin kunnen gebruiken om de eigenwaarden van R te schatten, moeten we eerst weten hoe R eruitziet. We gebruiken de notatie voor A, L en D zoals eerder geïntroduceerd, namelijk

$$A = \begin{bmatrix} a_1 & b_1 & & c_1 & & & & \\ b_1 & a_2 & b_2 & & c_2 & & & \\ & \ddots & \ddots & \ddots & & \ddots & & \\ c_1 & & b_m & a_{m+1} & b_{m+1} & & c_{m+1} & \\ & \ddots & & \ddots & \ddots & \ddots & & \ddots \\ & & & & & & & \end{bmatrix}.$$

[5]Merk op $M^T = (LD^{-1}L^T)^T = (L^T)^T D^{-T} L^T = LD^{-1}L^T = M$

en

$$L = \begin{bmatrix} d_1 & & & & & \\ b_1 & d_2 & & & & \\ & \ddots & \ddots & & & \\ c_1 & & b_m & d_{m+1} & & \\ & \ddots & & \ddots & \ddots & \\ & & & & & \end{bmatrix}$$

en $D = \text{diag}(d_1, \ldots, d_{m^2})$. We bewijzen nu eerst een stelling.

Stelling 5.2 Er geldt $r_{i,i+m-1} = \frac{\epsilon}{d_{i-1}}$ $(i \not\equiv 1 \mod m)$, en $r_{ij} = 0$ voor alle andere (i, j).

Bewijs. Definiëren we $Q_0 := \{(i, j) : |i - j| \neq 0, 1, m\}$, dan geldt $r_{ij} = 0$ als $(i, j) \notin Q_0$ (Vuik, C. & Lahaye, D.J.P., 2015). We hoeven dus alleen r_{ij} te bepalen voor $(i, j) \in Q_0$. Dit komt overeen met het bepalen van m_{ij}. Immers, als $(i, j) \in Q_0$, dan is $a_{ij} = 0$, zodat $r_{ij} = m_{ij} - 0 = m_{ij}$. Vanwege de symmetrie beperken we ons tot de gevallen waarin $j \geq i$.

Omdat $M = LD^{-1}L^T$, geldt er dat

$$m_{ij} = \begin{bmatrix} \frac{l_{i1}}{d_1} & \frac{l_{i2}}{d_2} & \cdots & \frac{l_{i,m^2}}{d_{m^2}} \end{bmatrix} \begin{bmatrix} l_{j1} \\ l_{j2} \\ \vdots \\ l_{j,m^2} \end{bmatrix} = \sum_{k=1}^{m^2} \frac{l_{ik}}{d_k} \cdot l_{jk}.$$

Wegens definitie is $l_{ik} = 0$ voor $k \notin \{i-m, i-1, i\}$ en evenzo $l_{jk} = 0$ voor $k \notin \{j-m, j-1, j\}$. Dit betekent dat voorgenoemde som, (aangenomen dat $j \geq i$,) alleen termen ongelijk aan 0 krijgt als $i = j$, $i = j - 1$, $i = j - m$ of als $i - 1 = j - m$. De eerste drie situaties zijn niet relevant, want we weten al dat dan $r_{ij} = 0$. Stel nu dat $i - 1 = j - m$ ofwel $j = i + m - 1$. Dan is

$$r_{ij} = \sum_{k=1}^{n \cdot m} \frac{l_{ik}}{d_k} \cdot l_{jk} = \frac{l_{i,i-1}}{d_{i-1}} \cdot l_{j,j-m} = \frac{-1}{d_{i-1}} \cdot -\epsilon = \frac{\epsilon}{d_{i-1}},$$

als $i \not\equiv 1 \mod m$. Als $i \equiv 1 \mod m$, dan is $l_{i,i-1} = 0$, zodat $r_{ij} = 0$. Zodoende volgt $r_{i,i+m-1} = \frac{\epsilon}{d_{i-1}}$ $(i \not\equiv 1 \mod m)$, en $r_{ij} = 0$ voor alle andere (i, j). □

Uit Stelling 5.2 volgt dat

$$R = \begin{bmatrix} & 0 & & & & & & & \\ & & \frac{\epsilon}{d_1} & & & & & \emptyset & \\ & & & \frac{\epsilon}{d_2} & & & & & \\ & & & & \ddots & & & & \\ 0 & & & & & 0 & & & \\ & \frac{\epsilon}{d_1} & & & \emptyset & & & \frac{\epsilon}{d_{m+1}} & & \\ & & \ddots & & & & & & \ddots & \\ & & & \ddots & & & & & & 0 \\ & \emptyset & & & \ddots & & & & & \\ & & & & & \frac{\epsilon}{d_{(n-1)m-1}} & & & & \\ & & & & & & 0 & & & \end{bmatrix}.$$

Uit de Stelling van Gershgorin volgt nu voor de eigenwaarden ξ_i van R dat

$$|\xi - 0| \leq \begin{cases} 0 & \text{als} \quad 1 \leq i \leq nm \quad \text{en} \quad i \equiv 1 \mod m \\ \left|\frac{\epsilon}{d_{i-1}}\right| & \text{als} \quad 1 \leq i \leq m \quad \text{en} \quad i \not\equiv 1 \mod m \\ \left|\frac{\epsilon}{d_{i-1}}\right| + \left|\frac{\epsilon}{d_{i-1-m}}\right| & \text{als} \quad m < i \leq (n-1)m \quad \text{en} \quad i \not\equiv 1 \mod m \\ \left|\frac{\epsilon}{d_{i-1-m}}\right| & \text{als} \quad (n-1)m < i \leq nm \quad \text{en} \quad i \not\equiv 1 \mod m \end{cases}$$

We bewijzen nu de volgende stelling.

Stelling 5.3. Als $k \not\equiv 0 \mod m$, dan $d_k \geq 1 + \delta$.

Bewijs. We bewijzen de stelling met behulp van inductie. We weten dat $d_1 = 1 + \epsilon + \delta \geq 1 + \delta$ (inductiebasis). We kijken eerst naar $k < m$. Neem aan $d_{k-1} \geq 1 + \delta$ (inductiehypothese). Dan is $-\frac{1}{d_{k-1}} \geq -\frac{1}{1+\delta}$. Er volgt dat

$$\begin{aligned} d_k &= a_k - \frac{(b_{k-1})^2}{d_{k-1}} = 2 + \epsilon + \delta - \frac{1}{d_{k-1}} \geq 2 + \epsilon + \delta - 1 \geq 1 + \epsilon + \delta \\ &> 1 + \delta. \end{aligned}$$

We kijken nu naar $m < k < (n-1)m$. Hierin onderscheiden we twee gevallen, namelijk $k \equiv 1 \mod m$ en $k \not\equiv 1 \mod m$. Stel $k \equiv 1 \mod m$ en neem aan $d_{k-m} > 1 + \delta$ ofwel $\frac{1}{d_{k-m}} > \frac{1}{1+\delta}$ (inductiehypothese). Dan is

$$\begin{aligned} d_k &= a_k - \frac{(b_{k-1})^2}{d_{k-1}} - \frac{(c_{k-m})^2}{d_{k-m}} = 1 + 2\epsilon + \delta - \frac{0}{d_{k-1}} - \frac{\epsilon^2}{d_{k-m}} \\ &> 1 + 2\epsilon + \delta - \epsilon^2 \geq 1 + \epsilon + \delta \\ &> 1 + \delta. \end{aligned}$$

Stel $k \not\equiv 1 \mod m$ (en $k \not\equiv 0 \mod m$). Neem aan $d_{k-1} > 1+\delta$ en $d_{k-m} > 1+\delta$, ofwel $-\frac{1}{d_{k-1}} > -\frac{1}{1+\delta}$ en $-\frac{1}{d_{k-m}} > -\frac{1}{1+\delta}$ (inductiehypothese). Dan

$$\begin{aligned} d_k &= a_k - \frac{(b_{k-1})^2}{d_{k-1}} - \frac{(c_{k-m})^2}{d_{k-m}} = 2 + 2\epsilon + \delta - \frac{1}{d_{k-1}} - \frac{\epsilon^2}{d_{k-m}} \\ &> 1 + 2\epsilon + \delta - 1 - \epsilon^2 \geq 1 + \epsilon + \delta \\ &> 1 + \delta. \end{aligned}$$

Rest nog het geval $k > (n-1)m$. Dit gaat hetzelfde als wanneer $m < k < (n-1)m$, alleen in a_k verandert 2ϵ in ϵ. Het is echter gemakkelijk in te zien dat dan alsnog volgt dat $d_k > 1+\delta$. □

Als we Stelling 5.3 en de toepassing van de Stelling van Gershgorin nu combineren, dan krijgen we

$$\xi_{max} < \left|\frac{\epsilon}{1+\delta}\right| + \left|\frac{\epsilon}{1+\delta}\right| = \frac{2\epsilon}{1+\delta}.$$

Nu volgt dat $||R|| = \xi_{max} < \frac{2\epsilon}{1+\delta}$, zodat

$$||A^{-1}M|| \leq 1 + \frac{1}{\delta} \cdot \frac{2\epsilon}{1+\delta} = 1 + \frac{2\epsilon}{\delta(1+\delta)}.$$

Daaruit volgt tot slot een ondergrens voor de minimale eigenwaarde λ_{min} van $M^{-1}A$, namelijk

$$\lambda_{min} \geq \frac{1}{1 + \frac{2\epsilon}{\delta(1+\delta)}}.$$

Nu kunnen we enkele conclusies trekken ten aanzien van vergelijking (1).

5.4. **Deelconclusies.** De in de vorige paragraaf verkregen ondergrens is bijzonder interessant. Deze laat namelijk zien dat ϵ juist een positieve werking heeft op de convergentie van $M^{-1}A$ (in het geval $\delta > 0$). Er geldt namelijk als $\epsilon \to 0$, dat het quotiënt van de ondergrens 1 nadert. Als $\epsilon = 0$, geldt dat de ondergrens precies gelijk is aan 1. We kunnen uit de analyses in dit hoofdstuk concluderen dat ϵ bij aanwezigheid van een δu-term geen invloed heeft op de ondergrens van de eigenwaarden van A en een verbeterende invloed op de ondergrens van $M^{-1}A$. Wanneer δ erg klein is, kunnen er wel kleine eigenwaarden voorkomen. Deze worden dan echter niet veroorzaakt door de anisotropiecoëfficiënt ϵ, maar doordat $\delta \to 0$.

6. Eigenwaardenanalyses (2): Von Neumannanalyse

In dit hoofdstuk stellen we $\delta = 0$, zodat we de oorspronkelijk Poissonvergelijking (met anisotropiecoëfficiënt) terugkrijgen. Voor de volledigheid zullen we deze hier nog even noemen:

$$-\frac{\partial^2 u}{\partial x^2} - \epsilon \frac{\partial^2 u}{\partial y^2} = f(x, y).$$

Om singulariteit te voorkomen, gaan we weer uit van de Dirichletvoorwaarde op de noordelijke rand zoals in hoofdstuk 1 gedefinieerd. Als we in de ondergrens voor λ_{min} in subparagraaf 5.2.3. δ naar 0 laten gaan, zien we dat λ_{min} ook richting 0 gaat. Dit zou het vermoeden kunnen opleveren dat de Poissonvergelijking problemen gaat geven wat betreft de kleine eigenwaarden. In dit hoofdstuk zullen we zien dat dit echter niet het geval is. Om te laten zien dat A niet singulier is, kan men een Von Neumannanalyse gebruiken. Het is precies deze analyse, die ook voor $M^{-1}A$ laat zien, dat λ_{min} niet oneindig groot wordt. Zoals we zullen zien, is er echter wel een verschil tussen A en $M^{-1}A$. In het vorige hoofdstuk zagen we dat er een 'ϵ-orde' verschil zat tussen de ondergrens van de A-eigenwaarden en die van de $M^{-1}A$-eigenwaarden. Dit blijkt ook nu weer het geval te zijn. Waar namelijk nu (in tegenstelling tot hoofdstuk 5) ϵ een negatieve invloed heeft op de A-eigenwaarden, zullen we zien dat $M^{-1}A$ geen last heeft van de anisotropiecoëfficiënt. Echter, de positieve beïnvloeding, zoals in hoofdstuk 5, is wel verdwenen. Dit heeft tot gevolg dat $M^{-1}A$ in bepaalde gevallen wel kleine eigenwaarden kan krijgen.

6.1. **Wat is er anders?** In paragraaf 5.1 hebben we gezien hoe de matrix A eruitzag. Het uiterlijk van de matrix verandert nu iets ten opzichte van hoofdstuk 5. In alle rijen verdwijnt de δ op de diagonaal. Verder worden de laatste m rijen, behorend bij de noordelijke randpunten als volgt (niet uitgaande van een hoekpunt):

$$\begin{bmatrix} 0 & \dots & -\epsilon & \dots & -1 & \delta + 2 + 3\epsilon & -1 & \dots & 0 \end{bmatrix}.$$

Voor het overige blijft de matrix hetzelfde.

We gaan opnieuw kijken naar de kleinste eigenwaarde van $M^{-1}A$. Net als in het vorige geval kijken we daarvoor naar de norm $||(M^{-1}A)^{-1}|| \leq 1 + ||A^{-1}|| \cdot ||R||$. Voor de eigenwaarden van R geldt nog steeds dat $\xi \leq |\frac{\epsilon}{d_{i-1}}| + |\frac{\epsilon}{d_{i-1-m}}|$. We hebben in Stelling 5.3 bewezen dat voor alle $i \not\equiv 1 \mod m$ geldt dat $d_i > 1 + \delta$. Omdat nu $\delta = 0$, wordt dit dus $d_i > 1$. Verder heeft Dirichletrandvoorwaarde hier geen negatieve invloed op, want dit maakt de diagonaalelementen alleen maar groter. We kunnen dus stellen dat de eigenwaarden van R begrensd worden door $\xi \leq 2\epsilon$.

Het wordt problematischer als we kijken naar de eigenwaarden van A. Passen we nu de Stelling van Gershgorin toe op A, dan krijgen we dat $|\lambda - 2 - 2\epsilon| \leq |1| + |1| + |\epsilon| + |\epsilon| = 2 + 2\epsilon$, ofwel $0 \leq \lambda \leq 4 + 4\epsilon$. Dit brengt nu het probleem met zich mee dat A eigenwaarde 0 zou kunnen hebben. We zullen nu echter zien aan de hand van een Von Neumannanalyse dat dit niet het geval is.

6.2. **Von Neumannanalyse.** De Von Neumannanalyse legt een link met de eigenwaarden van het continue probleem. Stellen we dat $A\mathbf{v} = \mu\mathbf{v}$, met μ een eigenwaarde en $\mathbf{v}$ een bijbehorende eigenvector, dan stellen we $v_j = e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_j}$, naar analogie van het continue probleem. Hierbij zijn x_j en y_j de x- en y-coördinaat van het j-de punt. Uit $A\mathbf{v} = \mu\mathbf{v}$ volgt dat $(Av)_j = (\mu v)_j = \mu v_j$, ofwel $-\epsilon v_{j-m} - v_{j-1} + (2+2\epsilon)v_j - v_{j+1} - \epsilon v_{j+m} = \mu v_j$. Werken we dit nu verder uit.

$$\begin{aligned} -\epsilon v_{j-m} - v_{j-1} + (2+2\epsilon)v_j - v_{j+1} - \epsilon v_{j+m} &= \mu v_j \\ -\epsilon e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_{j-m}} - e^{i\sqrt{\lambda_1}x_{j-1}}e^{i\sqrt{\lambda_2}y_j} & \\ +(2+2\epsilon)e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_j} & \\ -e^{i\sqrt{\lambda_1}x_{j+1}}e^{i\sqrt{\lambda_2}y_j} - \epsilon e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_{j+m}} &= \mu e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_j}. \end{aligned}$$

Schrijven we $e^{i\sqrt{\lambda_2}y_{j-m}} = e^{i\lambda_2 y_j}e^{-i\sqrt{\lambda_2}}$ en $e^{i\sqrt{\lambda_1}x_{j-1}} = e^{i\sqrt{\lambda_1}x_j}e^{-i\sqrt{\lambda_1}}$, et cetera, en delen we vervolgens door $e^{i\sqrt{\lambda_1}x_j}e^{i\sqrt{\lambda_2}y_j}$, dan krijgen we

$$\begin{aligned} \mu &= -\epsilon e^{-i\sqrt{\lambda_2}} - e^{-i\sqrt{\lambda_1}} + 2 + 2\epsilon - e_{i\sqrt{\lambda_1}} - \epsilon e^{i\sqrt{\lambda_2}} \\ &= 2 + 2\epsilon - 2\cdot\frac{e^{i\sqrt{\lambda_1}} + e^{-i\sqrt{\lambda_1}}}{2} + 2\epsilon - 2\epsilon\cdot\frac{e^{i\sqrt{\lambda_2}} + e^{-i\sqrt{\lambda_2}}}{2} \\ &= 2 + 2\epsilon - 2\cos\sqrt{\lambda_1} - 2\cos\sqrt{\lambda_2}. \end{aligned}$$

Voor $\sqrt{\lambda_1}$ en $\sqrt{\lambda_2}$ moeten we dan de eigenwaarden nemen behorende bij het continue probleem. Hiertoe werken we eerste het continueprobleem iets uit. We hebben de vergelijking

$$-u_{xx} - \epsilon u_{yy} = f,$$

met $u_x(0,y) = u_x(L,y) = 0$ en $u_y(x,0) = u(x,H) = 0$. Na het toepassen van scheiding van variabelen krijgen we twee deelproblemen

$$g_1(x) = c_1\cos(\sqrt{\lambda_1}x) + c_2\sin(\sqrt{\lambda_1}y),$$

met $g_1'(0) = g_1'(L) = 0$ en

$$g_2(y) = c_3\cos(\sqrt{\lambda_2}x) + c_4\sin(\sqrt{\lambda_2}y),$$

met $g_2'(0) = g_2(H) = 0$. Wanneer we dit uitwerken, krijgen we

$$c_2 = 0 \text{ en } \sqrt{\lambda_1} = \frac{k\pi}{L} \text{ en } c_4 = 0 \text{ en } \sqrt{\lambda_2} = \frac{(l+\frac{1}{2})\pi}{H},$$

met $k, l = 0, 1, 2, \ldots$.

Omdat in het discrete geval de intervallen $[0, L]$ en $[0, H]$ opgedeeld zijn in respectievelijk m en m punten en we om wille van de vereenvoudiging hebben gekozen $L = H$, krijgen we voor het discrete geval $\sqrt{\lambda_1} = \frac{k\pi}{m}$ en $\sqrt{\lambda_2} = \frac{(l+\frac{1}{2})\pi}{m}$. De precieze eigenwaarden van A zijn dus gelijk aan[6]

$$\mu_{kl} = 2 + 2\epsilon - 2\cos(\frac{k\pi}{m}) - 2\epsilon\cos\left(\frac{(l+\frac{1}{2})\pi}{m}\right).$$

[6]Deze eigenwaarden kunnen indien gewenst getoetst worden aan de matrix A (voor algemene m) en blijken dan te kloppen (eigenvectoren zijn gemakkelijk te vinden met behulp van (co)sinussen).

Hierbij lopen k en l van 0 tot en met $m-1$ (er zijn immers maar m^2 verschillende eigenwaarden mogelijk). Deze eigenwaarden zijn minimaal als de cosinussen zo groot mogelijk zijn. Dit is het geval als $k = l = 0$. We krijgen dus

$$\mu_{min} = 2 + 2\epsilon - 2\cos(0) - 2\epsilon\cos\left(\frac{\left(\frac{1}{2}\right)}{m}\pi\right) = 2\epsilon\left(1 - \cos\left(\frac{1}{2m}\pi\right)\right).$$

Uit de gonioformule $1 - \cos(2x) = 2\sin^2 x$ volgt

$$\mu_{min} = 2\epsilon \cdot 2\sin^2\left(\frac{1}{4m}\right) = 4\epsilon\sin^2\left(\frac{1}{4m}\right). \quad (*)$$

Zodoende krijgen we dat $||A^{-1}|| = \frac{1}{4\epsilon\sin^2\left(\frac{1}{4m}\right)}$, waaruit volgt

$$||(M^{-1}A)^{-1}|| \leq 1 + ||A^{-1}|| \cdot ||R|| = 1 + \frac{2\epsilon}{4\epsilon\sin^2\left(\frac{1}{4m}\right)} = 1 + \frac{1}{2\sin^2\left(\frac{1}{4m}\right)}.$$

Zoals eerder reeds verklaard, volgt hieruit dat voor de kleinste eigenwaarde van $M^{-1}A$ geldt dat

$$\lambda_{min} \geq \frac{1}{1 + \frac{1}{2\sin^2\left(\frac{1}{4m}\right)}}. \quad (**)$$

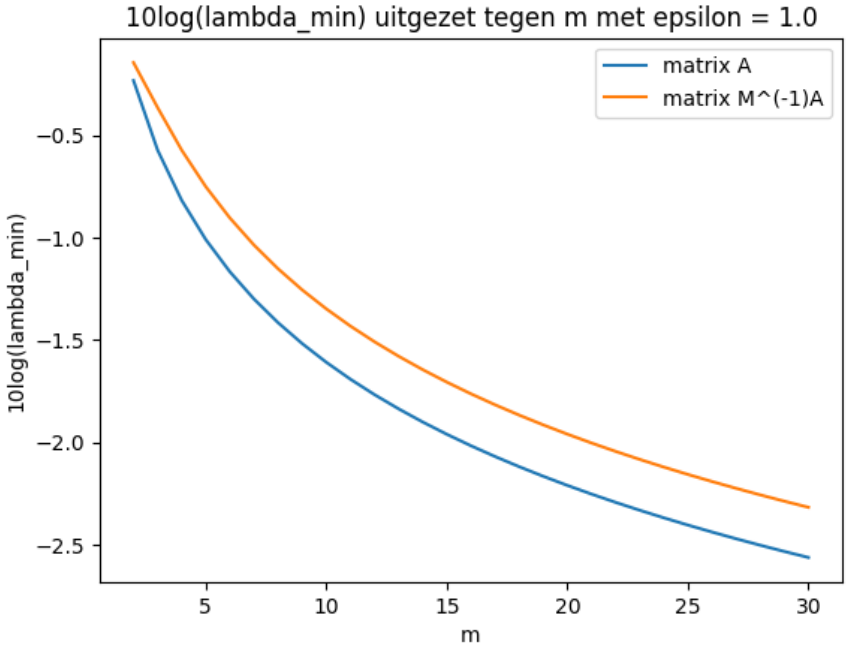
Hiermee hebben we een ondergrens gevonden voor de eigenwaarden van $M^{-1}A$.

6.3. **Deelconclusies.** We zien dus dat wanneer $\delta = 0$, met slechts op één rand een Dirichletrandvoorwaarde, de ondergrens voor de eigenwaarden van A afhankelijk wordt van ϵ. Om precies te zijn bestaat er een lineair verband tussen, zoals uit blijkt uit formule (*) in paragraaf 6.2. Dit verklaart de in hoofdstuk 4 waargenomen kleine eigenwaarden van A. We kunnen echter ook concluderen, dat de waarde van ϵ geen invloed heeft op de ondergrens voor de eigenwaarden van $M^{-1}A$. Echter zien we nu wel dat deze ondergrens afhankelijk is geworden van het aantal gridpunten in de discretisatie. De ondergrens voor de eigenwaarden heeft limiet 0 als $\sin^2\left(\frac{1}{4m}\right) \to 0$. Dit is het geval als $\frac{1}{4m} \to q\pi$ voor $q \geq 0$ geheel. Omdat $m \geq 1$ geldt altijd dat $\frac{1}{4m} \ll \pi$, dus alleen het geval $q = 0$ is aan de orde. Dit betekent dus dat de eigenwaarden klein worden als $m \to \infty$. We tonen nu voor diverse waarden van m wat de ondergrens voor de eigenwaarden van $M^{-1}A$, zoals weergegeven in (**), wordt.

m	orde van de ondergrens (**)
10	10^{-3}
100	10^{-5}
500	$5 \cdot 10^{-7}$
1000	10^{-7}

Het blijkt dat met name bij erg verfijnde roosters, hier een probleem onstaat. Zodoende kunnen er wel kleine eigenwaarden ontstaan in $M^{-1}A$, echter worden deze niet veroorzaakt door een kleine ϵ. De waarden in bovenstaande tabel gelden namelijk net zo goed voor $\epsilon = 1$ als voor hele kleine ϵ. Met behulp van ons programma kunnen we het bovenstaande resultaat ook controleren. We plotten weer de $^{10}\log$ van de kleinste eigenwaarden van

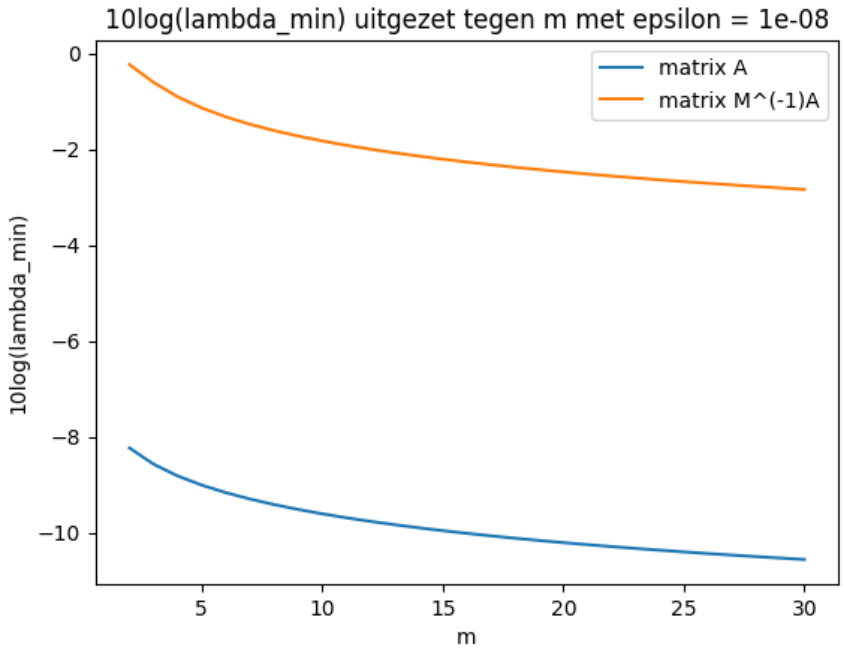
A en $M^{-1}A$ zoals in Figuur 6 in Paragraaf 4.2, echter nu niet uitgezet tegen ${}^{10}\log(\epsilon)$, maar tegen het aantal gridpunten m. Zie Figuur 7 en 8, waarin respectievelijk is gekozen voor $\epsilon = 1$ en $\epsilon = 10^{-8}$. We krijgen het verwachte resultaat.



FIGUUR 7. ${}^{10}\log(\lambda_{min})$ uitgezet tegen m met $\epsilon = 1$

7. Beantwoording van de onderzoeksvraag: conclusies

We kunnen nu een antwoord geven op de onderzoeksvraag: Wat is de invloed van een kleine anisotropicoëfficiënt in de Poissonvergelijking (met de gegeven randvoorwaarden) op de convergentie van de CG- en de ICCG(0)-methode? Er is gebleken dat een kleine ϵ ernstig doorwerkt in de eigenwaarden van A. Er bestaat een lineair verband tussen ϵ en de minimale eigenwaarde van A. Een kleine anisotropiecoëfficiënt heeft vertraagt dus de convergentie van de CG-methode. Verder is gebleken dat de ondergrens voor de eigenwaarden van $M^{-1}A$ niet afhankelijk is van ϵ. Dit betekent dat de invloed van een kleine anisotropiecoëfficiënt op de convergentie van de ICCG(0)-methode slechts beperkt is. Behalve deze hoofdconclusie hebben we nog enkele bijreresultaten verkregen. Er is gebleken dat de ondergrens voor de eigenwaarden van zowel A als $M^{-1}A$ bij de Poissonvergelijking afhangt van het aantal gridpunten in de discretisatie. Hoe groter het aantal gridpunten, hoe trager de convergentie. Ook hebben we tussentijds gezien dat toevoeging van een δu-term voorkomt dat de eigenwaarden van A en $M^{-1}A$ willekeurig dicht bij 0 kunnen komen als we ϵ en het aantal gridpunten variëren.



Figuur 8. ${}^{10}\log(\lambda_{min})$ uitgezet tegen m met $\epsilon = 10^{-8}$

8. Uitbreiding en aanbevelingen voor onderzoek

In het voorgaande onderzoek hebben we aangetoond, dat een kleine anisotropiecoëfficiënt de kleinste eigenwaarden van A omlaag trekt, maar dat preconditionering met behulp van ICCG(0) voldoende is om dit probleem op te lossen. Toch kan ook $M^{-1}A$ in bepaalde situaties kleine eigenwaarden krijgen. Dit is bijvoorbeeld het geval wanneer we in hoofdstuk 5 een hele kleine δ kiezen. Ook zagen we in hoofdstuk 6 dat het aantal gridpunten invloed heeft op de eigenwaarden. Ook hebben we in dit onderzoek niet heel uitgebreid gekeken naar de invloed van de gekozen randvoorwaarden op het proces. Het is daarom na dit onderzoek aanbevelenswaardig om andere situaties te onderzoeken, door te variëren in type vergelijking en in randvoorwaarden. Het is mogelijk dat men dan op situaties stuit waarin preconditionering niet afdoende blijkt te zijn. Naast preconditionering bestaat er dan nog een andere methode om te proberen kleine eigenwaarden te omzeilen en zo de convergentie te verbeteren. Deze methode heet deflatie. In dit hoofdstuk zullen we hier, als aanzet tot nader onderzoek, wat uitgebreider op ingaan.

8.1. **Deflatie.** We gaan weer uit van het stelsel $A\mathbf{u} = \mathbf{f}$ met A SPD en stel dat A een $n \times n$-matrix is. We definiëren eerst een deflatiematrix Z. De kolommen van deze matrix bestaan uit de zogenoemde deflatievectoren. Dit zijn de vectoren die corresponderen met de uit te schakelen eigenwaarden. De deflatiematrix is dus een $n \times m$-matrix (met $m < n$), waarbij m het aantal deflatievectoren is. Deze vectoren dienen onafhankelijk te zijn, ofwel de rang van Z moet gelijk zijn aan m. Het is belangrijk dat de deflatievectoren verstandig gekozen worden, ten einde het deflatieproces effectief te laten zijn. Op de keuze van de deflatievectoren komen we later terug.

We definiëren[7] nu de matrix E door $E := Z^T AZ$ en vervolgens de matrices $Q := I - ZE^{-1}Z^T A$ en $B := I - AZE^{-1}Z^T$. Merk op dat

$$BA = A - AZE^{-1}Z^T A = AQ.$$

Verder is

$$\begin{aligned} B^2 &= (I - AZE^{-1}Z^T)^2 = I - 2AZE^{-1}Z^T + AZE^{-1}Z^T AZE^{-1}Z^T \\ &= I - 2AZE^{-1}Z^T + AZE^{-1}EE^{-1}Z^T = I - 2AZE^{-1}Z^T + AZE^{-1}Z^T \\ &= I - AZE^{-1}Z^T \\ &= B. \end{aligned}$$

Zo volgt ook

$$Q^2 = (A^{-1}BA)^2 = A^{-1}BAA^{-1}BA = A^{-1}B^2A = A^{-1}BA = Q.$$

Definiëren we nu $\tilde{\mathbf{u}} := Q\mathbf{u}$, dan geldt

$$BA\tilde{\mathbf{u}} = BAQ\mathbf{u} = BBA\mathbf{u} = B^2A\mathbf{u} = BA\mathbf{u} = B\mathbf{f}.$$

We kunnen $B\mathbf{f}$ gemakkelijk uitrekenen. Vervolgens kunnen we $\tilde{\mathbf{u}}$ oplossen uit $BA\tilde{\mathbf{u}} = B\mathbf{f}$ met behulp van PCG. Hierna kunnen we $\mathbf{u}$ als volgt berekenen:

$$\mathbf{u} = (I - Q)\mathbf{u} + Q\mathbf{u} = ZE^{-1}Z^T A\mathbf{u} + Q\mathbf{u} = ZE^{-1}Z^T\mathbf{f} + \tilde{\mathbf{u}}.$$

[7] Deze E bestaat, aangezien A SPD is en Z een full-rankmatrix.

Nu wordt de PCG-methode dus toegepast op het stelsel $BA\tilde{\mathbf{u}} = B\mathbf{f}$ en wordt de convergentie dus bepaald door het effectieve conditiegetal (en dus de eigenwaarden) van de matrix BA.

8.2. **De invloed van deflatie op de eigenwaarden.** We zullen nu nader analyseren welke invloed deflatie heeft op de eigenwaarden. Om inzicht te verschaffen in de werking van deflatie, nemen we m orthogonale eigenparen $(\lambda_1, \mathbf{v}_1), \ldots, (\lambda_m, \mathbf{v}_m)$ van A. Er geldt dus $\mathbf{v}_i^T \mathbf{v}_j = 0$ voor alle $i \neq j$ (de eigenvectoren kunnen orthogonaal gekozen worden omdat A SPD is). [8] We nemen nu als deflatievectoren $\mathbf{v}_1, \ldots, \mathbf{v}_m$, dus Z is dan een $n \times m$-matrix, namelijk $Z = \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_m \end{bmatrix}$.

Stelling 8.1 Noteren we het spectrum van A als $\mathrm{Sp}(A)$, dan geldt dat $\mathrm{Sp}(BA) = (\ \mathrm{Sp}(A) \cup \{0\}) \backslash \{\lambda_1, \ldots, \lambda_m\}$. Voor de eigenvectoren $\mathbf{v}_1, \ldots, \mathbf{v}_m$ geldt $(BA)\mathbf{v}_i = \mathbf{0}$, voor de overige eigenvectoren geldt $(BA)\mathbf{v}_i = A\mathbf{v}_i$.

Anders geformuleerd geldt er dat BA de gekozen m eigenwaarden van A naar 0 stuurt, terwijl de overige eigenwaarden volledig intact blijven.

Bewijs. Er geldt dat

$$\begin{aligned}
E = Z^T A Z &= \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T \end{bmatrix} A \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_m \end{bmatrix} = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T \end{bmatrix} \begin{bmatrix} A\mathbf{v}_1 & \ldots & A\mathbf{v}_m \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T \end{bmatrix} \begin{bmatrix} \lambda_1\mathbf{v}_1 & \ldots & \lambda_m\mathbf{v}_m \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1\mathbf{v}_1^T\mathbf{v}_1 & \lambda_2\mathbf{v}_1^T\mathbf{v}_2 & \ldots & \lambda_m\mathbf{v}_1^T\mathbf{v}_m \\ \lambda_1\mathbf{v}_2^T\mathbf{v}_1 & \lambda_2\mathbf{v}_2^T\mathbf{v}_2 & \ldots & \lambda_m\mathbf{v}_2^T\mathbf{v}_m \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1\mathbf{v}_m^T\mathbf{v}_1 & \lambda_2\mathbf{v}_m^T\mathbf{v}_2 & \ldots & \lambda_m\mathbf{v}_m^T\mathbf{v}_m \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1||\mathbf{v}_1||^2 & 0 & \ldots & 0 \\ 0 & \lambda_2||\mathbf{v}_2||^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_m||\mathbf{v}_m||^2 \end{bmatrix}.
\end{aligned}$$

Hieruit volgt

$$E^{-1} = \begin{bmatrix} \frac{1}{\lambda_1||\mathbf{v}_1||^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{\lambda_2||\mathbf{v}_2||^2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{\lambda_m||\mathbf{v}_m||^2} \end{bmatrix}.$$

Er geldt

$$BA = (I - AZE^{-1}Z^T)A = A - AZE^{-1}Z^TA.$$

[8] Deflatie werkt ook als A niet SPD is, maar deze analyse wordt dan wat ingewikkelder. Om wille van de eenvoud beperken we ons nu tot een A die SPD is.

Nemen we nu een $i \in \{1, \ldots, m\}$, dan krijgen we $BA\mathbf{v}_i = A\mathbf{v}_i - AZE^{-1}Z^TA\mathbf{v}_i$. Uiteraard is $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$, dus we richten ons nu op de berekening van $AZE^{-1}Z^TA\mathbf{v}_i$.
Er geldt

$$\begin{aligned}
AZE^{-1}Z^TA\mathbf{v}_i &= AZE^{-1}Z^T\lambda_i\mathbf{v}_i \\
&= A\begin{bmatrix}\mathbf{v}_1 & \ldots & \mathbf{v}_m\end{bmatrix}E^{-1}\begin{bmatrix}\mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T\end{bmatrix}\lambda_i\mathbf{v}_i \\
&= \begin{bmatrix}\lambda_1\mathbf{v}_1 & \ldots & \lambda_m\mathbf{v}_m\end{bmatrix}E^{-1}\begin{bmatrix}\lambda_i\mathbf{v}_1^T\mathbf{v}_i \\ \vdots \\ \lambda_i\mathbf{v}_m^T\mathbf{v}_i\end{bmatrix} \\
&= \begin{bmatrix}\lambda_1\mathbf{v}_1 & \ldots & \lambda_m\mathbf{v}_m\end{bmatrix}\begin{bmatrix}\frac{1}{\lambda_1||\mathbf{v}_1||^2} & & \\ & \ddots & \\ & & \frac{1}{\lambda_m||\mathbf{v}_m||^2}\end{bmatrix}\begin{bmatrix}0 \\ \vdots \\ 0 \\ \lambda_i||\mathbf{v}_i||^2 \\ 0 \\ \vdots \\ 0\end{bmatrix} \\
&= \begin{bmatrix}\lambda_1\mathbf{v}_1 & \ldots & \lambda_m\mathbf{v}_m\end{bmatrix}\begin{bmatrix}0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0\end{bmatrix} = \lambda_i\mathbf{v}_i.
\end{aligned}$$

Nu volgt dus dat $BA\mathbf{v}_i = A\mathbf{v}_i - AZE^{-1}Z^TA\mathbf{v}_i = \lambda_i\mathbf{v}_i - \lambda_i\mathbf{v}_i = \mathbf{0}$. Dit geldt voor alle $i \in \{1, \ldots, m\}$, dus we zien dat de eigenwaarden behorend bij de deflatievectoren als het ware uitgeschakeld worden. Nemen we daarentegen een eigenvector $\mathbf{v}_i$ met $i \notin \{1, \ldots, m\}$, dan zien we dat

$$\begin{aligned}
AZE^{-1}Z^TA\mathbf{v}_i &= AZE^{-1}(AZ)^T\mathbf{v}_i \\
&= AZE^{-1}\begin{bmatrix}\lambda_1\mathbf{v}_1^T \\ \vdots \\ \lambda_m\mathbf{v}_m^T\end{bmatrix}\mathbf{v}_i = AZE^{-1}\begin{bmatrix}\lambda_1\mathbf{v}_1^T\mathbf{v}_i \\ \vdots \\ \lambda_m\mathbf{v}_m^T\mathbf{v}_i\end{bmatrix} \\
&= AZE^{-1}\mathbf{0} = \mathbf{0},
\end{aligned}$$

ofwel $BA\mathbf{v}_i = A\mathbf{v}_i - \mathbf{0} = A\mathbf{v}_i$. We zien dus dat in de overige eigenrichtingen geen aantasting van de eigenwaarden van A plaatsvindt. □

Deze stelling geeft een bijzonder interessant resultaat. Als we voor $\lambda_1, \ldots, \lambda_m$ precies de eigenwaarden kiezen die een zeer negatieve invloed hebben op de convergentie van PCG, dan hebben we bij PCG met deflatie geen last meer van deze eigenwaarden en zal de convergentie dus substantieel verbeteren.

8.3. **De keuze van de deflatievectoren.** We komen nu terug op de keuze van de deflatievectoren. Zoals in de vorige paragraaf is gebleken, is deflatie het meest effectief wanneer er als deflatievectoren precies de eigenvectoren worden gekozen die bij de slechte eigenwaarden horen. Echter, in de praktijk is het meestal erg kostbaar om deze eigenwaarden en -vectoren te vinden, aangezien er sprake is van een grote n. Daarom wordt er vaak een andere methode gebruikt om de deflatievectoren te kiezen.

Zij $\mathbf{u} = \begin{bmatrix} u(\mathbf{x}_1) & \dots & u(\mathbf{x}_n) \end{bmatrix}^T$, waarbij dus $\mathbf{x}_1, \dots, \mathbf{x}_n \in \Omega$. We verdelen het gebied Ω in d disjuncte deelgebieden $\Omega_1, \dots, \Omega_d$. We definiëren voor $1 \leq i \leq d$:

$$(\mathbf{z}_i)_j := \begin{cases} 0 & \text{als } \mathbf{x}_j \notin \Omega_i \\ 1 & \text{als } \mathbf{x}_j \in \Omega_i \end{cases}$$

Nu nemen we $\mathbf{z}_1, \dots, \mathbf{z}_d$ als deflatievectoren, dus $Z = \begin{bmatrix} \mathbf{z}_1 & \dots & \mathbf{z}_d \end{bmatrix}$. De decompositie $\Omega = \Omega_1 \cup \ldots \cup \Omega_d$ moet dan uiteraard verstandig gekozen worden. Wat een verstandige keuze is, verschilt per situatie. Het is daarom aanbevelenswaardig om situaties op te zoeken waarin preconditionering an sich niet afdoende is, zoals in de inleiding van dit hoofdstuk al genoemd is en vervolgens te onderzoeken wat er in die situaties bereikt kan worden met deflatie.

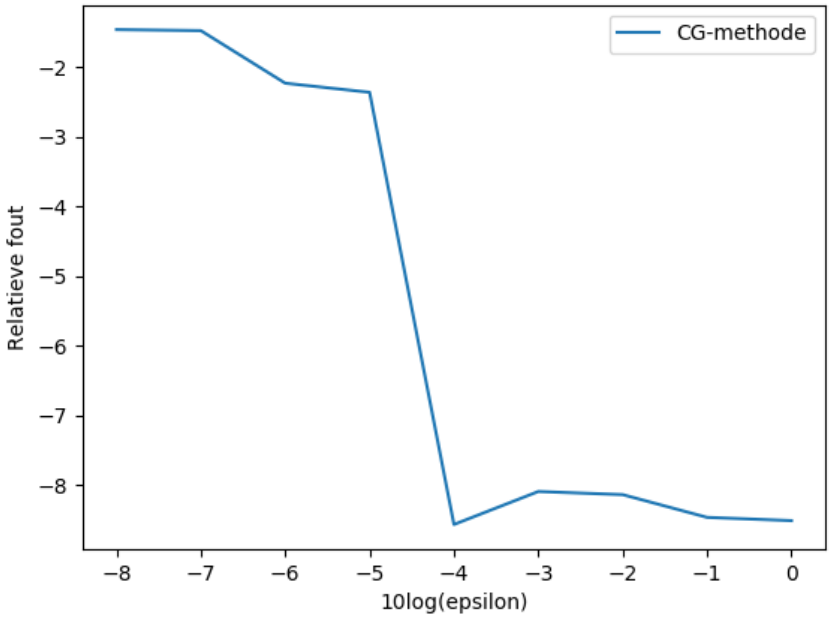
9. Discussie

Een aantal kritische overwegingen ter discussie:

- Dit onderzoek heeft slechts beperkt gekeken naar de rol van de randvoorwaarden in het proces. Wel is in hoofdstuk 3 vanuit natuurkundig perspectief toegelicht waarom juist deze randvoorwaarden gekozen zijn. Er is echter geen onderzoek gedaan naar de resultaten wanneer er in randvoorwaarden gevarieerd zou worden. Zoals in hoofdstuk 8 reeds aangegeven, zou dit een mogeljikheid zijn voor nader onderzoek.
- Het oorspronkelijke plan met dit onderzoek was om deflatie uitgebreider mee te nemen in het onderzoek. Omdat echter gebleken is dat preconditionering afdoende is om de invloed van ϵ teniet te doen, was hier eigenlijk geen aanleiding meer toe. Zodoende viel dit niet meer binnen de scope van het eigenlijke onderzoek. Dit is de reden dat het literatuuronderzoek naar deflatie is opgenomen in hoofdstuk 8, gepaard gaande met aanbevelingen voor nader onderzoek.
- Er is in het onderzoek voornamelijk gekeken naar de ondergrenzen van de eigenwaarden. Het conditiegetal hangt echter ook af van de bovengrens van de eigenwaarden. Omdat het te ver voerde om hier ook naar te kijken, is dit buiten beschouwing gelaten. Weliswaar doet dit enigszins af aan de volledigheid van het onderzoek, maar omdat anisotropie de eigenwaarden vooral verkleint en niet vergroot, is dit een overkomelijk bezwaar.
- Er is niet onderzocht hoe accuraat de gegeven ondergrens voor de eigenwaarden van $M^{-1}A$ is. Omdat de matrix A goede eigenschappen heeft, kan de minimale eigenwaarde van A gemakkelijk precies gemaakt worden, zoals ook gedaan is. Voor $M^{-1}A$ zijn echter diverse afschattingen gebruikt, zoals de driehoeksongelijkheid en de Stelling van Gershgorin. Er kan dan ook niet in strikte zin geconcludeerd worden dat ϵ géén invloed heeft op de eigenwaarden van $M^{-1}A$ (dit is ook vrij triviaal, want de waarden van d matrix veranderen met ϵ). Wel kan echter aan de hand van de ondergrens geconcludeerd worden dat de invloed beperkt is; de kleinste eigenwaarde blijft in ieder geval boven de ondergrens. Het feit dat dit uit de waarnemingen in hoofdstuk 4 duidelijk blijkt, ondersteunt dit. De nauwkeurigheid van de afschattingen zou echter dus nog uitgebreider onderzocht kunnen worden.

10. Bijlagen

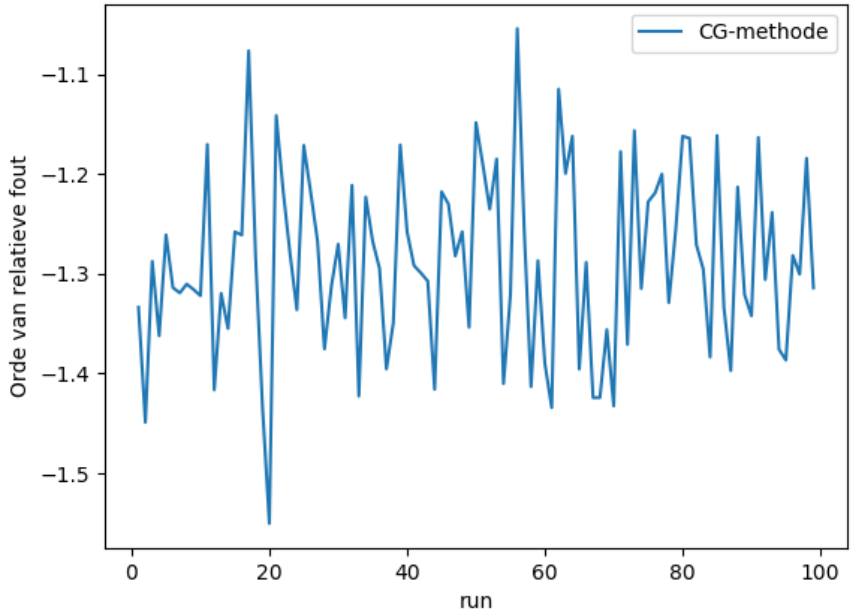
10.1. **Verantwoording ten aanzien van de relatieve fout.** In deze bijlage gaan we verder in op de relatieve fout zoals genoemd in paragraaf 4.1. We hebben het stelsel $A\mathbf{u} = \mathbf{f}$. We laten $\mathbf{u}$ willekeurig genereren door Python. Vervolgens berekenen we $\mathbf{f} = A\mathbf{u}$. We lossen met de CG-methode het stelsel $A\mathbf{u} = \mathbf{f}$ op met stopcriterium 10^{-10}, zodat we een oplossing $\mathbf{u}_k$ krijgen. Vervolgens berekenen we de relatieve fout RE $:= \frac{||\mathbf{u}-\mathbf{u}_k||}{||\mathbf{u}||}$. We doen dit eerst voor diverse waarden van ϵ en zetten de relatieve fout uit tegen de waarden van ϵ (met $m = 50$). Zie Figuur 9.



Figuur 9. ${}^{10}\log(RE)$ uitgezet tegen ${}^{10}\log(\epsilon)$

Uiteraard is Figuur 9 slechts de weergave van één run. Bij andere runs kunnen de waarden anders zijn, daar de $\mathbf{u}$ telkens random wordt gegenereerd. Daarom toont onderstaande figuur voor $\epsilon = 10^{-8}$ de relatieve fout voor 100 verschillende runs, zie Figuur 10.

Het blijkt dat de relatieve fout voor $\epsilon = 10^{-8}$ in de meeste gevallen rond de $10^{-1,3}$ ligt. Zoals in Figuur 9 bleek, stijgt deze waarde naarmate ϵ kleiner wordt. Omdat $\epsilon = 10^{-8}$ de kleinste waarde is die we gebruiken voor ϵ, is dit dus voldoende. Als we daarbij ook nog in ogenschouw nemen dat de relatieve fout voor de eigenwaardenanalyses geen verschil maakt (de eigenwaarden hangen niet af van het rekenproces), is aangetoond dat de relatieve fout geen probleem vormt voor de betrouwbaarheid van het onderzoek.



FIGUUR 10. $^{10}\log(RE)$ voor 100 verschillende runs met $\epsilon = 10^{-8}$

10.2. **Pythoncode.** Pythoncode is opvraagbaar bij de auteur van deze thesis.

11. Bibliografie

- Vuik, C. & Lahaye, D.J.P. (2015). Scientific Computing. Delft: Delft Institute of Applied Mathematics
- Mittal, A. (2016). Parallelization of An Experimental Multiphase Flow Algorithm (Master thesis). Geraadpleegd van http://ta.twi.tudelft.nl/nw/users/vuik/numanal/amittal_scriptie.pdf
- Bear, J. (1972). Dynamics of Fluids in Porous Media. Geraadpleegd van https://books.google.nl/books?id=fBMeVSZ_3u8C