Numerieke Partiële Differentiaal Vergelijkingen

Deel 1

H.A.van der Vorst

12 december 1997

Hoofdstuk 1

Elliptische Differentiaal vergelijkingen

1.1 Inleiding

Zij $d \in \mathbf{N}$. Beschouw een open gebied Ω in $\mathbf{R}^d$ met rand $\Gamma \equiv \partial\Omega \equiv \bar{\Omega}\backslash\Omega$.
Zij $\mathcal{L}$ een tweede orde differentiaal operator, voor reëel (of complex) waardige functies u op Ω, van de vorm

$$\mathcal{L}(u) = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}(a_{i,j}\frac{\partial}{\partial x_j}u) + \sum_{j=1}^{d} b_j \frac{\partial}{\partial x_j}u + cu, \tag{1.1}$$

waarbij $a_{i,j}, b_j$ en c gegeven coëfficiënt functies op Ω zijn.
De operator $\mathcal{L}$ heet *elliptisch* als voor zekere $\delta > 0$ geldt

$$\sum_{i,j=1}^{d} \xi_i a_{i,j}(x)\xi_j \geq \delta \sum_{j=1}^{d} |\xi_j|^2, \text{ voor alle } (\xi_1, \ldots, \xi_d) \in \mathbf{R}^d, x \in \Omega.$$

We zijn geïnteresseerd in de (gladde) reëelwaardige functie u op Ω die de oplossing is van het volgende elliptische probleem

$$\mathcal{L}(u) = f, \quad \text{op} \quad \Omega,$$

$$\mathcal{R}(u) = g, \quad \text{op} \quad \Gamma,$$

waarbij f een gegeven functie is op Ω, $\mathcal{R}$ een gegeven randwaarde operator die aan een functie u op Ω een functie op Γ toevoegt, en g een gegeven functie op Γ. Afhankelijk van de vorm van Ω, de gladheid van de rand Γ, de vorm van $\mathcal{R}$, de gladheid van de coëfficient functies en de termen f en g, is het probleem goed geformuleerd, uniek oplosbaar en hangt de oplossing (in of andere zin) continu af van de bepalende grootheden.
Deze problematiek van *existentie, uniciteit* en *goed gesteldheid* behoort tot de toegepaste analyse en valt buiten de scope van dit college. We zullen er slechts terloops aandacht aan besteden: we nemen aan dat de elliptische problemen die we in dit college tegen komen allemaal aan alle eisen voldoen.
In het bijzonder zullen we steeds aannemen dat $a_{i,j} \in C^1(\Omega)$ en $b_j, c \in C(\Omega)$.

Voorbeelden.

1. $\mathcal{L}$ is de Laplace operator, m.b.t. $\mathbf{R}^d$: $\mathcal{L} = \Delta$ met
$$\Delta v = -\sum_{i=1}^{d} \frac{\partial^2 v}{\partial x_i^2},$$
voor $v \in C^2(\Omega)$.
Voor $f \in C(\Omega)$ en $g \in C(\Gamma)$ luidt de *Poisson vergelijking*:
$$\Delta v = f \quad \text{op} \quad \Omega$$
$$v = g \quad \text{op} \quad \Gamma. \ \text{(Dirichlet randvoorwaarde)}$$
We zoeken naar een oplossing $v \in C(\bar{\Omega}) \cap C^2(\Omega)$.
In de fysica staat de Poisson vergelijking ook wel bekend onder de naam *potentiaal vergelijking*.

2. $d = 3$, met
$$\mathcal{L} \equiv -\sum_i \frac{\partial}{\partial x_i} a(x_1, x_2, x_3) \frac{\partial}{\partial x_i} + \sum_i b_i(x_1, x_2, x_3) \frac{\partial}{\partial x_i},$$
met $a > 0$ over Ω, is voor $f \in C(\Omega)$ en $g \in C(\Gamma)$,
$\mathcal{L}u = f$ op Ω,
en $u = g$ op Γ,
een voorbeeld van de *konvektie-diffusie vergelijking*. Gevraagd wordt een oplossing $u \in C(\bar{\Omega}) \cap C^2(\Omega)$. Men noemt de tweede orde term de *diffusie term* en de eerste orde term de *konvektie term*.

Natuurlijk gaat in de praktijk de belangstelling niet in de eerste plaats uit naar de vergelijkingen in de twee voorbeelden. Maar deze eenvoudige problemen en de analyse daaraan, geven vaak wel een zeer verhelderend inzicht in de numerieke mogelijkheden. We zullen deze problemen daarom nog een aantal keren tegenkomen.

Opmerkingen:

1. Iedere uniform kontinue funktie op Ω kan op precies een manier uitgebreid worden tot een kontinue funktie op $\bar{\Omega}$. Deze uitbreiding geeft een een-een verband tussen de ruimte van de uniform kontinue funkties op Ω en de ruimte $C(\bar{\Omega})$ van kontinue funkties op $\bar{\Omega}$. We kunnen dus $C(\bar{\Omega}) \cap C^2(\Omega)$ zien als een ruimte van gladde funkties op Ω. De *restrictie operator* $\mathcal{R}$ die aan een $u \in C(\bar{\Omega})$ diens beperking tot Γ toevoegt ($\mathcal{R}(u)(x) \equiv u(x)$ voor $x \in \Gamma$) is in deze optiek een operator voor die gladde funkties op Ω (dus in feite eerst uitbreiden en dan beperken).

2. In bovenstaande voorbeelden beperkt de restrictie operator $\mathcal{R}$ de functie u tot de rand Γ van het gebied Ω. De resulterende randvoorwaarde $\mathcal{R}(u) = g$ op Γ wordt dan de *Dirichlet randvoorwaarde* genoemd. Men is ook nogal eens geïnteresseerd in de operator $\mathcal{R}$ die aan een functie u op Ω in ieder punt x op de rand van Γ de afgeleide $\frac{\partial u(x)}{\partial n}$ van u in x in

de richting van de uitwaartse normaal $n(x)$ op Γ in x toevoegt (in feite ook weer pas nadat de afgeleiden van u zijn uitgebreid tot funkties op $\bar{\Omega}$). Deze restrictie operator legt de zogenaamde *Neumann randvoorwaarde* op. We zullen ook nog combinaties van beide typen randvoorwaarden tegenkomen.

1.2 Discretisatie methoden

In de numerieke praktijk worden verschillende methoden gebruikt om elliptische partiële differentiaal vergelijkingen te diskretiseren. De meest gebruikte zijn:

- de eindige differentie methode (ED methode)
- de eindige volume methode (of box methode)
- de eindige elementen methode (finite elements)

De eerste twee methoden zijn nauw aan elkaar verwant en worden meestal beide aangeduid als Finite Difference (FD) methoden. Zoals zal blijken heeft de box methode enige voordelen en verdient daarom de voorkeur. De eindige differentie methode ligt conceptueel wat meer voor de hand en is eenvoudiger te analyseren.

In de industrie wordt tegenwoordig meestal gewerkt met software pakketten die gebaseerd zijn op de Finite Element (FE) methode. Voorbeelden daarvan zijn PROTRAN, NASTRAN, DIANA, ABAQUS en SEPRAN. De FE methode laat zich beter toepassen op domeinen Ω met een grillige rand Γ. De daarvoor benodigde programmatuur, met name voor de zogenaamde gridgenerator, is echter zeer ingewikkeld en men moet dan denken aan pakketten van honderdduizenden regels code. Bovendien leidt deze aanpak veelal tot matrix vector vergelijkingen, waarvan de matrix een ingewikkeld niet-nul patroon heeft en de numerieke oplossing van deze stelsels kan tot efficiency problemen leiden. Om dit soort redenen zijn, vooral bij het werken op gebieden Ω met "regelmatige"randen (rechthoeken, balken, cylinders,...) en bij lokaal niet al te grote variaties in de oplossing u, de FD methoden zeer populair. Voorbeelden van toepassing zijn oliereservoir simulatie, halfgeleider simulatie, grondwater stromingen, reactor diffusie.

We zullen de 3 methoden presenteren aan de hand van een eenvoudig model probleem.

Model probleem:

$$-u'' + bu = f$$

op

$$\Omega = [0, 1], \tag{1.2}$$

$$u(0) = \alpha, u(1) = \beta, \alpha, \beta \in \mathbf{R},$$

met $b, f \in C(\Omega)$ en $b(x) \geq 0$, voor $x \in \Omega$.

De presentatie van de drie discretisatie aanpakken is ontleend aan [10].

1.2.1 De eindige differentie (ED) methode

We beschouwen voor $h = 1/(n+1)$ (de stapgrootte of roosterafstand) de n roosterpunten $\Omega_h = \{x_j = jh, j = 1, 2, \ldots, n\}$. Ons doel is om de oplossing u van de differentiaalvergelijking te benaderen met een discrete oplossing $u_h(x_j)$ in de roosterpunten x_j. Hierbij zal de benaderingsfout $|u(x_j) - u_h(x_j)|$ van h afhangen. Het idee is nu om de differentiaal operator $\mathcal{L}$, die werkt op functies op Ω, te vervangen door een differentie operator $\mathcal{L}_h$, die werkt op functies op Ω_h.

Een voor de hand liggende aanpak is om uit te gaan van Taylor ontwikkelingen. Aannemend dat de oplossing u over Ω voldoende vaak continue differentieerbaar is, geldt voor elk roosterpunt:

$$u(x_{j\pm1}) = u(x_j) \pm hu'(x_j) + \frac{h^2}{2!}u''(x_j) \pm \frac{h^3}{3!}u'''(x_j) + \frac{h^4}{4!}u^{(4)}(x_j + \xi_j^{\pm}h),$$

met $0 < |\xi_j^{\pm}| < 1$. Hieruit volgt direct

$$-u''(x_j) = \frac{2u(x_j) - u(x_{j-1}) - u(x_{j+1})}{h^2} + \frac{2h^2}{4!}u^{(4)}(x_j + \xi_j h),$$

met $0 \leq |\xi_j| < 1$.

Met $u(x_0) \equiv u(0) = \alpha$ en $u(x_{n+1}) \equiv u(1) = \beta$ krijgen we aldus voor de n onbekenden $u(x_j), j = 1, \ldots, n$ een stelsel lineaire vergelijkingen dat de differentiaal vergelijking vervangt:

$$A\bar{u} = \bar{f} + \tau(u),$$

met

$$A = \frac{1}{h^2}\begin{pmatrix} 2+b(x_1)h^2 & -1 & & & \\ -1 & 2+b(x_2)h^2 & & & \\ & \ddots & \ddots & \ddots & \\ & & & -1 & 2+b(x_n)h^2 \end{pmatrix},$$

$$\bar{u} = (u(x_1), u(x_2), \ldots, u(x_n))^T$$

$$\bar{f} = (f(x_1) + \frac{\alpha}{h^2}, f(x_2), \ldots, f(x_{n-1}), f(x_n) + \frac{\beta}{h^2})^T$$

$$\tau(u) = -\frac{h^2}{12}(u^{(4)}(x_1 + \xi_1 h), \ldots, u^{(4)}(x_n + \xi_n h))^T.$$

Als discrete benadering $u_h(x_j)$ voor de oplossing $u(x_j)$ nemen we de oplossing van het stelsel

$$A\bar{u}_h = \bar{f}, \tag{1.3}$$

dat wil zeggen, we laten uit het oorspronkelijke stelsel de term $\tau(u)$ gewoon weg.

Opgave: Toon aan dat de symmetrische matrix A niet-singulier is en dat alle elementen van A^{-1} positief zijn, te noteren als $A^{-1} > 0$.
(Hint: schrijf $A = LL^T$, met L beneden bidiagonaal; toon aan dat de diagonaal elementen van

L positief zijn en de nevendiagonaal elementen negatief; laat tenslotte zien dat $L_{ij}^{-1} > 0$ voor $i \geq j$).

We zullen ons nu buigen over de vraag of de oplossing u_h convergeert naar u, op de punten x_i, voor $h \to 0$. De volgende stelling doet daar een uitspraak over.

Stelling 1 *Zij $u(x)$ de oplossing van het model probleem, en zij u_h de discrete oplossing verkregen uit het stelsel $A\bar{u}_h = \bar{f}$. Indien*

$$\max_{x \in [0,1]} |u^{(4)}(x)| \leq M,$$

dan geldt

$$\max_{1 \leq i \leq n} |u(x_i) - u_h(x_i)| \leq \frac{Mh^2}{96}.$$

Bewijs: $\bar{u} - \bar{u}_h = A^{-1}\tau(u)$.
Zij $|v|$ de vector met componenten $|v_i|$. Er geldt

$$|\bar{u} - \bar{u}_h| \leq A^{-1}|\tau(u)|,$$

wegens $A^{-1} > 0$.
Hieruit volgt $|\bar{u} - \bar{u}_h| \leq \frac{h^2}{12} M A^{-1} e$,
met $e = (1, 1, \ldots, 1)^T$.
Men gaat nu eenvoudig na dat $A^{-1} \leq A_0^{-1}$, waarbij A_0^{-1} de matrix is die men verkrijgt bij discretisatie van het modelprobleem met $b(x) = 0$. Voorts kan men verifiëren dat met $w(x) = \frac{1}{2}x(1-x)$ de vector $\bar{w} = (w(x_1), w(x_2), \ldots, w(x_n))^T$ oplossing is van het stelsel $A_0 w = e$.
Dus

$$\max_{1 \leq i \leq n} |u(x_i) - u_h(x_i)| \leq \frac{h^2}{12} M \cdot \max_{x_i} |\frac{1}{2}x(1-x)| \leq M\frac{h^2}{96}. \quad \square$$

Deze manier van discretiseren heeft er toe geleid dat elke onbekende $u_h(x_i)$ middels een lineaire vergelijking met zijn twee nabuur waarden $u_h(x_{i-1})$ en $u_h(x_{i+1})$ verbonden is. Men spreekt daarom wel van een 3-punts schema ter onderscheiding van de (meerpunts) schema's die men verkrijgt door met de voortgezette Taylor reeksen via meerdere buren ook nog de 4-de en/of hogere afgeleiden te elimineren.

Uit $A^{-1} > 0$ volgt voor de elementen $a_{i,j}^{(-1)}$ van A^{-1} dat $\sum_j |a_{i,j}^{(-1)}| = (A^{-1}e)_i$, zodat we uit het bewijs van stelling 1 en de cirkelstelling van Gershgorin concluderen dat $\rho(A^{-1}) \leq \frac{1}{8}$ ($\rho(A) \equiv$ het maximum der absolute waarden der eigenwaarden van A). Omdat

$$\sum_j |a_{i,j}| \leq \max_i \{\frac{1}{h^2}(4 + |b(x_i)|h^2)\},$$

volgt dat $\rho(A) = \mathcal{O}(h^{-2})$. Voor het conditiegetal $\kappa_2(A) \equiv \|A\|_2 \|A^{-1}\|_2$, dat een belangrijke rol speelt bij de foutgevoeligheid van lineaire stelsels, alsmede bij de convergentie snelheid van relevante iteratieve methoden voor het oplossen van die stelsels, geldt dan $\kappa_2(A) = \mathcal{O}(h^{-2})$.

1.2.2 De box methode of eindige volume methode

Bij de tweede aanpak beschouwt men een geschikt gekozen deeldomein ω_i (een "box") rond elk roosterpunt x_i en men integreert de differentiaal vergelijking over dit deeldomein.
Voor ons model probleem kiezen we als deeldomein bij x_i het segment $\omega_i = [x_i - \frac{h}{2}, x_i + \frac{h}{2}]$. Integratie van (1.2) over ω_i levert dan:

$$\int_{\omega_i} -u''dx + \int_{\omega_i} budx = \int_{\omega_i} fdx. \tag{1.4}$$

Hieruit volgt meteen

$$-u'(x_i + \frac{h}{2}) + u'(x_i - \frac{h}{2}) + \int_{\omega_i} budx = \int_{\omega_i} fdx. \tag{1.5}$$

De beide afgeleiden worden benaderd met centrale differenties, bijvoorbeeld

$$u'(x_i + \frac{h}{2}) \rightarrow \frac{u(x_{i+1}) - u(x_i)}{h},$$

en de beide integralen worden benaderd met, bijvoorbeeld, de midpunt regel:

$$\int_{\omega_i} budx = hb(x_i)u(x_i).$$

Om symmetrie in het resulterende stelsel te waarborgen is het wel essentieel dat afgeleiden op de gemeenschappelijke rand van domeinen in elk der domeinen op precies dezelfde manier benaderd worden.

Men gaat eenvoudig na dat deze aanpak leidt tot $h^2 A\bar{u} = h^2 \bar{f}$ met precies dezelfde A, $\bar{u}$ en $\bar{f}$ als in (1.3) en dus is de approximatie fout hier gegeven door $h^2\tau(u)$ (In feite wordt er hier een stelsel opgeleverd dat met een factor h^2 vermenigvuldigd is t.o.v. ED-discretisatie).

Men zou zich op dit punt kunnen afvragen wat het voordeel is van de box methode boven de ED aanpak. In de praktijk werkt men vaak met niet-equidistante roosters en na enig rekenwerk blijkt dat ook dan de box methode, althans voor een klasse van problemen waartoe (1.2) behoort, leidt tot symmetrische matrices, in tegenstelling tot de ED aanpak. Voor de numerieke oplossing van de resulterende lineaire stelsels blijkt dat van enorm voordeel te zijn. Meer in het algemeen kan men verifiëren dat de box methode ook nog symmetrische stelsels oplevert bij discretisatie van

$$-(a(x)u'(x))' + c(x)u(x) = f(x), x \in [p, q] \equiv \Omega.$$

Tenslotte komt het in de praktijk veel voor dat men rekent met functies $a(x), c(x)$ en $f(x)$ die stuksgewijs continue zijn. Zulke discontinuïteiten doen zich voor indien men bijvoorbeeld in diffusie problemen met verschillende materialen te maken krijgt. Indien we aannemen dat $u(x)$ en $a(x)u'(x)$ beide continue zijn op $[p, q]$, dan kan men met de box methode nog tot bruikbare discretisaties komen, met nog steeds symmetrische stelsels, *mits* de discontinuïteits punten allen ook roosterpunt zijn. Voor details, zie [10].

1.2.3 De eindige elementen methode (FE)

We zullen deze methode in dit college slechts zeer summier behandelen.
De FE methode is gebaseerd op een variationele formulering. Toegepast op het model probleem (1.2) ontstaat het volgende beeld. Laat $v(x)$ voorlopig een willekeurige functie zijn op $[0,1]$ met stuksgewijs continue eerste afgeleide en zodanig dat $v(x)$ voldoet aan de randvoorwaarden: $v(0) = \alpha, v(1) = \beta$.

Beschouw

$$F(v) = \frac{1}{2}\int_0^1 \{(\frac{dv(x)}{dx})^2 + b(x)v^2(x) - 2v(x)f(x)\}dx \tag{1.6}$$

Met $v(x) = u(x) + \epsilon(x)$, waarin $u(x)$ de oplossing van (1.2) voorstelt, ontstaat

$$F(v) = F(u) + \int_0^1 \{\frac{du}{dx}\frac{d\epsilon}{dx} + bu\epsilon - \epsilon f\}dx + \frac{1}{2}\int_0^1 \{(\frac{d\epsilon}{dx})^2 + b\epsilon^2\}dx. \tag{1.7}$$

Partiële integratie voor de eerste integraal in (1.7) geeft:

$$\epsilon(1)u'(1) - \epsilon(0)u'(0) + \int_0^1 \epsilon(x)\{-u'' + bu - f\}dx = 0.$$

Dus kan (1.7) herschreven worden en we vinden

$$F(v) = F(u) + \frac{1}{2}\int_0^1 \{(\frac{d\epsilon}{dx})^2 + b\epsilon^2\}dx \geq F(u). \tag{1.8}$$

Blijkbaar minimaliseert de oplossing $u(x)$, van de model vergelijking, de functionaal $F(v)$ voor alle voldoend vaak differentieerbare functies $v(x)$ die aan de rand voorwaarden voldoen.
Een populaire methode om dit minimaliseringsprobleem numeriek op te lossen is de klasse van toelaatbare functies te beperken tot continue stuksgewijs lineaire functies $v(x)$ die aan de randvoorwaarden voldoen. Het idee hierachter is dat elke continue functie willekeurig dicht (in een geschikt gekozen norm) benaderd kan worden door een continue stuksgewijze lineaire functie.

Met parameters $v_1, \ldots, v_n$ en geschikt gekozen vaste punten $x_j \in [0,1]$ kunnen we zo'n toelaatbare functie voorstellen door

$$v(x) = v_j + \frac{v_{j+1} - v_j}{x_{j+1} - x_j}(x - x_j), \quad x_j \leq x \leq x_{j+1}$$

(hierbij wordt aangenomen dat $x_0 = 0, x_{n+1} = 1, v_0 = \alpha$ en $v_{n+1} = \beta$).

De functionaal $F(v)$ kan nu geschreven worden als een kwadratische vorm in de parameters $v_1, \ldots, v_n$, en deze vorm kan geminimaliseerd worden door de vergelijkingen

$$\frac{\partial F(v)}{\partial v_i} = 0, \quad 1 \leq i \leq n \tag{1.9}$$

op te lossen. Indien we voor het gemak maar aannemen dat de roosterpunten x_j equidistant over Ω verspreid liggen, met $h \equiv x_{j+1} - x_j$, dan kan men deze vergelijkingen herschrijven tot

$$\frac{v_j - v_{j+1}}{h} + \frac{v_j - v_{j-1}}{h} + \int_{x_{j-1}}^{x_{j+1}} b(x)v(x)t_j(x)dx = \int_{x_{j-1}}^{x_{j+1}} f(x)t_j(x)dx, \quad 1 \leq j \leq n \tag{1.10}$$

met
$t_j(x)$ de continue stuksgewijs lineaire functie die gedefinieerd wordt door

- $t_j(x) = 0$ voor $x \leq x_{j-1}$ en $x \geq x_{j+1}$
- $t_j(x_j) = 1$.

Merk op dat de functies $v(x)$ uit de toelaatbare klasse geschreven kunnen worden als

$$v(x) = \sum_{j=0}^{n+1} v_j t_j(x),$$

waarbij $t_0(x)$ en $t_{n+1}(x)$ op voor de hand liggende wijze gedefinieerd worden als de stuksgewijs lineaire functies:
$t_0(x) = 0$ voor $x \geq x_1$, $t_0(0) = 1$,
$t_{n+1}(x) = 0$ voor $x \leq x_n, t_{n+1}(1) = 1$.

M.a.w. de set $\{t_j(x)\}$ kunnen we beschouwen als een basis voor de beperkte functie klasse waaruit we de $v(x)$ gekozen hebben.

Indien we de beide integralen in (1.10) benaderen via de midpunt regel, dan krijgen we uiteindelijk voor de v_j's hetzelfde lineaire stelsel als bij de eerder beschreven discretisatie technieken, maar we zouden de beide integralen in voorkomende gevallen eventueel ook exact kunnen berekenen en dat leidt dan tot andere stelsels.

De eindige elementen methode komt er in de hier beschreven vorm op neer dat men een kwadratische functionaal tracht te minimaliseren binnen een beperkte functie klasse, waarvoor men als basis een set functies kiest waarbij elk van deze functies slechts ongelijk aan nul is op een beperkt deelgebied van Ω. Dit laat in de praktijk grote mogelijkheden toe bij het overdekken van ingewikkelde gebieden Ω en bij het kiezen van geschikte basis functies $t_i(x)$. Bijvoorbeeld in het 2-dimensionale geval is het gebruikelijk een gebied te overdekken met driehoeken en/of vierhoeken.
Als basis functies zijn, behalve stuksgewijs lineaire, ook stuksgewijs kwadratische functies populair. Voor theoretische aspecten van deze methode zie bijvoorbeeld [3],[9]. Practische aspecten worden uitgebreid behandeld in [1].

Het herformuleren van zelfgeadjungeerde elliptische partiële differentiaal vergelijkingen tot een kwadratische vorm heeft vaak als voordeel dat het eenvoudiger is om positief definietheid van de resulterende matrix vast te stellen. Het laatste is van belang voor diverse numerieke oplosmethoden voor het stelsel vergelijkingen.
In tegenstelling tot de box methode is de variationele aanpak niet zonder meer geschikt voor problemen die niet zelf geadjungeerd zijn.

Toepassing van de besproken discretisatie methoden op het 1-dimensionale model probleem levert weinig boeiende stelsels vergelijkingen op. Ook de verwerking van de randvoorwaarden geeft in deze situatie geen problemen. Daarom zullen we nog wat uitgebreider ingaan op een geschikt tweedimensionaal probleem.

1.3 De box methode voor een twee dimensionaal probleem

We beschouwen in deze paragraaf de tweede orde zelf-geadjungeerde elliptische partiële differentiaal vergelijking

$$-(P(x,y)u_x)_x - (P(x,y)u_y)_y + Q(x,y)u(x,y) = F(x,y), \qquad (1.11)$$

met $Q, F \in C(\bar{\Omega})$, $Q(x,y) \geq 0$ en $P \in C^1(\bar{\Omega})$.
Als randvoorwaarden zijn opgelegd:

$$\alpha(x,y)u + \beta(x,y)\frac{\partial u}{\partial n} = \gamma(x,y), \quad (x,y) \in \Gamma.$$

Hierin is $\frac{\partial u}{\partial n}$ de afgeleide in de richting van de uitwendige normaal langs Γ. De functies $\alpha, \beta, \gamma \in C(\Gamma)$ en voldoen aan

$$\alpha(x,y) \geq 0, \beta(x,y) \geq 0, \alpha + \beta > 0, \quad (x,y) \in \Gamma,$$

waarbij we aannemen dat tenminste langs een gedeelte van Γ geldt dat $\beta(x,y) = 0$ (Dirichlet randvoorwaarde).
Voor het gemak nemen we voor Ω het eenheidsvierkant in R^2: $[0,1] \times [0,1]$.

Net als in het 1-dimensionale geval leggen we eerst een rechthoekig rooster over Ω aan, dat wil zeggen we kiezen een set $\{x_i | i = 0, \ldots, n\}$ met $0 = x_0 < x_1 < \ldots < x_n = 1$, en evenzo een set $\{y_j\}$, $j = 0, 1, \ldots, m$, met $0 = y_0 < y_1 < y_2 < \ldots < y_m = 1$.
De rooster afstanden $h_i \equiv x_{i+1} - x_i$ en $h_{n+j} \equiv k_j \equiv y_{j+1} - y_j$ hoeven niet uniform te zijn in elk der coördinaat richtingen. Het rooster Ω_h bestaat uit alle paren (x_i, y_j). We gaan nu weer de differentiaal vergelijking vervangen door een differentie vergelijking over Ω_h via de box methode.

We associëren met elk roosterpunt (x_i, y_j) een lokaal domein $\omega_{i,j}$, gedefinieerd als het vierkant

$$\omega_{i,j} = \{(x,y) | x_i - \frac{h_{i-1}}{2} \leq x \leq x_i + \frac{h_i}{2},$$
$$y_j - \frac{k_{j-1}}{2} \leq y \leq y_j + \frac{k_j}{2}\}.$$

Conventie: $h_{-1} = k_{-1} = h_n = k_m = 0$.

Over elk lokaal domein $\omega_{i,j}$ rond een roosterpunt (x_i, y_j) integreren we (1.11):

$$-\int\int_{\omega_{i,j}} \{(Pu_x)_x + (Pu_y)_y\} dxdy + \int\int_{\omega_{i,j}} Qudxdy = \int\int_{\omega_{i,j}} Fdxdy. \qquad (1.12)$$

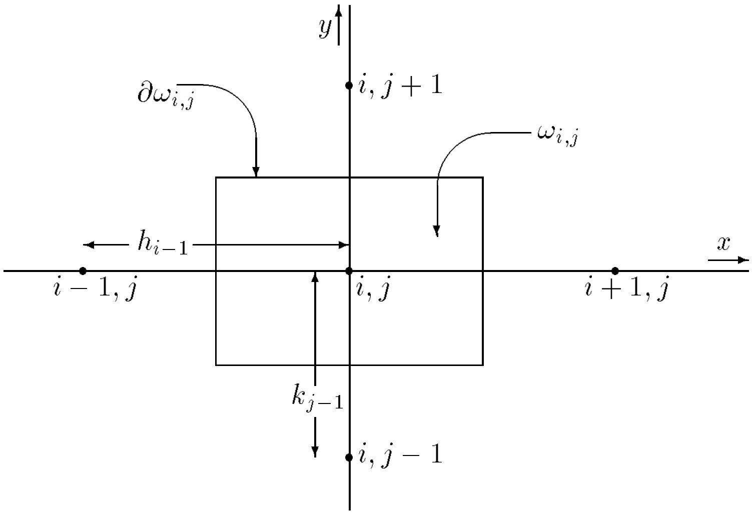
De divergentie stelling van Gauss, voor partiële integratie in meer dimensies:

$$\int_{\omega} div\vec{v} = \int_{\partial\omega} (\vec{v}, \vec{n}),$$

zegt nu dat voor elk paar functies $S(x,y), T(x,y) \in C^1(\bar{\omega}_{i,j})$ geldt:

$$\int\int_{\omega_{i,j}} (S_x + T_y) dxdy = \int_{\partial\omega_{i,j}} (-Tdx + Sdy),$$

waarbij $\partial\omega_{i,j}$ de rand van $\omega_{i,j}$ is en waarbij de lijnintegraal wordt genomen tegen de wijzers van de klok in:



Met $S = Pu_x, T = Pu_y$ kunnen we vergelijking (1.12) schrijven als

$$-\int_{\partial\omega_{i,j}} \{Pu_x dy - Pu_y dx\} + \int\int_{\omega_{i,j}} Qudxdy = \int\int_{\omega_{i,j}} Fdxdy. \tag{1.13}$$

De twee oppervlakte integralen worden weer simpelweg met een midpunt regel benaderd, bijvoorbeeld

$$\int\int Qudxdy \to Q(x_i, y_j)u(x_i, y_j)\mathcal{O}_{i,j},$$

met $\mathcal{O}_{i,j} = (h_i + h_{i-1})(k_j + k_{j-1})/4$ (de oppervlakte van het lokaal domein). Vervolgens worden in de lijn integralen de partiële afgeleiden benaderd met centrale differenties en de midpuntregel. Uiteindelijk leidt dat tot een lineair stelsel, hetgeen ons, indien we discretisatie fouten verder veronachtzamen, een benaderende oplossing in de roosterpunten zal opleveren die we zullen noteren als $u_{i,j}$.
De lijnintegraal over de rand van het lokaal domein in (1.13) wordt nu dus als volgt benaderd (leid dit zelf ook af):

$$\begin{aligned}
&-\int_{\partial\omega_{i,j}} \{Pu_x dy - Pu_y dx\} \to \qquad (1.14)\\
&\frac{k_{j-1}+k_j}{2}\{P_{i+1/2,j}(\frac{u_{i,j}-u_{i+1,j}}{h_i}) + P_{i-1/2,j}(\frac{u_{i,j}-u_{i-1,j}}{h_{i-1}})\} +\\
&\frac{h_{i-1}+h_i}{2}\{P_{i,j+1/2}(\frac{u_{i,j}-u_{i,j+1}}{k_j}) + P_{i,j-1/2}(\frac{u_{i,j}-u_{i,j-1}}{k_{j-1}})\},
\end{aligned}$$

met $P_{i+1/2,j} \equiv P(x_i + \frac{h_i}{2}, y_j)$, etcetera.
Ook hier is het voor het verkrijgen van symmetrie in het gediscretiseerde stelsel weer van wezenlijk belang dat de afgeleiden in (1.14), bij integratie over een gemeenschappelijke rand van

twee aangrenzende lokale domeinen (alsook de randintegralen zelf), steeds op dezelfde manier gediscretiseerd worden.

Als $u_{i,j}$ een punt op de rand vertegenwoordigt dan wordt in geval van een Dirichlet randvoorwaarde ($\beta_{i,j} = 0$) de bekende waarde van $u_{i,j}$ ingevuld en deze bekende termen verhuizen naar het rechterlid van de onstane lineaire vergelijkingen. In zo'n geval wordt over het boxje om de rand niet geïntegreerd. In het geval van een Neumann randvoorwaarde ($\alpha_{i,j} = 0$), of Robin randvoorwaarde ($\alpha_{i,j} \neq 0, \beta_{i,j} \neq 0$), kan langs het betreffende deel $\Gamma_{i,j} = \partial\omega_{i,j} \cap \partial\Omega$ van de rand van het lokale domein het overeenkomstige stuk van de lijnintegraal benaderd worden via

$$\int_{\Gamma_{i,j}} P \frac{\partial u}{\partial n} dt \to P_{i,j} \frac{\gamma_{i,j} - \alpha_{i,j} u_{i,j}}{\beta_{i,j}} \ell,$$

waarbij ℓ de lengte van het stukje rand $\Gamma_{i,j}$ van het lokale domein is.

Het hier afgeleide schema is een zogenaamd 5-punts schema, dat wil zeggen dat de onbekende $u_{i,j}$ in elk roosterpunt (in het inwendige van Ω_h) is verbonden met de onbekenden in de 4 nabuur punten. Uiteraard kan men tot meer-punts schema's komen door nauwkeuriger te integreren en door de afgeleiden nauwkeuriger te bepalen. We zullen in dit dictaat verder op het 5-punts schema ingaan, maar merken nu vast op dat veel van de bewijsvoering zonder meer ook door te voeren is voor schema's met meer punten.

De vergelijking bij het roosterpunt (i,j) zullen we noteren als

$$\sum_{k=0}^{4} a_{i,j}^k u_{i\pm1,j\pm1} \equiv a_{i,j}^0 u_{i,j} + (a_{i,j}^1 u_{i-1,j} + a_{i,j}^2 u_{i+1,j} + a_{i,j}^3 u_{i,j-1} + a_{i,j}^4 u_{i,j+1}) = f_{i,j}. \tag{1.15}$$

De volgende conventies vereenvoudigen het schrijfwerk:
(1) als $u_{i\pm1,j\pm1}$ 'over de rand ligt' (of een Dirichlet randpunt) is dan ontbreekt de bijbehorende term in de vergelijking (of is verwerkt in het rechterlid $f_{i,j}$).
(2) het gedeelte van de vergelijking tussen haken zullen we ook noteren als $\sum_{k=1}^4 a_{i,j}^k u_{i\pm1,j\pm1}$.

Voor het gemak nemen we verder maar aan dat als ergens langs een van de vier zijden van Ω geldt $\beta(x,y) = 0$, dat dat dan langs de gehele betreffende zijde geldt. Dat leidt tot lineaire stelsels met een wat overzichtelijker niet-nul structuur.

Opgave. Ga na dat de beschreven manier van discretiseren voor (1.11) leidt tot een stelsel $Au = f$, waarbij A een symmetrische matrix van de orde $N = n_x \cdot n_y$ is (n_x is het aantal onbekenden in x-richting, n_y is het aantal onbekenden in y-richting).
Als we de onbekenden $u_{i,j}$ lexicografisch ordenen volgens hun indices, dat wil zeggen $u = (u_{1,1}, u_{2,1}, \ldots, u_{n_x,1}, u_{1,2}, u_{2,2}, \ldots, u_{n_x,n_y})^T$, laat dan zien dat de matrix $A = (a_{i,j})$, met $i = 1, \ldots, N$, $j = 1, \ldots, N$ (let op: dit is een andere indexering dan die van het rooster) de volgende eigenschappen heeft

1. $a_{i,j} \leq 0$ als $i \neq j$
2. $a_{i,i} \geq \sum_j |a_{i,j}|$, met strikt groter teken als $Q(x,y) > 0$ over heel Ω ,of als een der corresponderende rooster buurpunten een Dirichlet randpunt is.

3. de matrix A is blok tridiagonaal:

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & & & \\ A_{2,1} & A_{2,2} & A_{2,3} & & \\ & \ddots & \ddots & \ddots & \\ & & & A_{n_y,n_y-1} & A_{n_y,n_y} \end{pmatrix},$$

waarbij $A_{i,j}$, met $|i-j|=1$, diagonaal matrices zijn van orde n_x en $A_{i,i}$ zijn tridiagonaal matrices van orde n_x.

Opmerking 1: Voor $k = \nu h$ zijn de elementen van A van de orde 1 en de elementen van f zijn $\mathcal{O}(h^2)$. Door de foutterm met Taylor reeksen te bepalen ziet men dat deze van $\mathcal{O}(h^4)$ is (vergelijk dit met het resultaat in paragraaf 1.2.2). Men bedenke dat bij de box methode steeds geïntegreerd is over lokale domeinen met oppervlakte $\mathcal{O}(h^2)$, zodat de benaderingsfout tussen de op de juiste wijze geschaalde differentie operator en de differentiaal operator (beperkt tot het rooster) van $\mathcal{O}(h^2)$ is.

Opmerking 2: voor minder regelmatige gebieden Ω, of bij Dirichlet randvoorwaarden langs gedeelten van zijden van Ω, kunnen we ook een tridiagonale blok structuur van A tot stand brengen, echter de optredende blokken kunnen dan van wisselende orde zijn. De overige eigenschappen uit de vorige opgave gelden onveranderd.
Voor de behandeling van randvoorwaarden bij wat algemenere gebieden zie [3] of [10].

1.3.1 Eigenschappen van het gediscretiseerde stelsel

In deze paragraaf gaan we nader in op de lineaire stelsels die we door discretisatie verkregen hebben. De eigenschappen die we zojuist de revue lieten passeren zijn meer algemeen geldig en het zal handig blijken ze van namen te voorzien.

Definitie 1 *Een n bij n matrix $A = (a_{i,j})$ heet* zwak diagonaal dominant *als geldt $|a_{i,i}| \geq \sum_{j \neq i} |a_{i,j}|$ voor alle i. Indien voor alle i het $>$-teken geldt, dan heet de matrix A* diagonaal dominant.

Blijkbaar geldt voor het modelprobleem (1.11) dat de matrix A van het met de box methode gediscretiseerde stelsel zwak diagonaal dominant is. Indien er langs een gedeelte van de rand Γ Dirichlet randvoorwaarden zijn opgelegd dan geldt voor de waarde van i, passend bij een roosterpunt dat zo'n Dirichlet randpunt als nabuur heeft, het $>$ teken. Wanneer de functie $Q(x,y) > 0$, over het hele domein Ω, dan is A diagonaal dominant.

In het rooster Ω_h voor het model probleem (1.11) zijn alle roosterpunten via een traplijn van roosterpunten binnen Ω_h met elkaar verbonden. Dat heeft als gevolg dat elke onbekende van elke andere onbekende afhangt. Nog weer anders gezegd: we kunnen niet een deelgroep $u_{i,j}$'s oplossen zonder de overige u-waarden te kennen. Lineaire stelsels met die eigenschap noemt men *irreducibel*.

Definitie 2 *Een matrix A van orde n, met $n \geq 2$, heet* reducibel *als er een permutatie matrix P bestaat zodat*

$$PAP^T = \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix},$$

met $A_{1,1}$ een r bij r matrix en $A_{2,2}$ een $(n-r)$ bij $(n-r)$ matrix. Als zo'n P niet bestaat dan heet A irreducibel.

Voor het door ons gehanteerde rooster ziet men eenvoudig in dat elke onbekende (door zijn nabuur relaties) via roosterlijntjes met elke andere onbekende te verbinden is en dat er dus door hernummering der onbekenden (gevolg van P^T), en door herschikking der vergelijkingen (gevolg van P) geen groep van onbekenden te formeren is die onafhankelijk van de rest opgelost kan worden. Dit geldt algemeen voor rechthoekige roosters over samenhangende gebieden, mits de roosterafstand h (en ook k) klein genoeg is.

Om vervolgens te kunnen bewijzen dat het door discretisatie ontstane stelsel niet singulier is komt het volgende Lemma goed van pas.

Lemma 1 *Zij $Au^h = f$ het met de box methode gediscretiseerde stelsel dat past bij de modelvergelijking (1.11) met $F(x,y) = 0$, en zij A irreducibel. Dan geldt*

$$\max_{i,j \in \Omega_h} |u^h_{i,j}| = \max_{i,j \in \Gamma_h} |u^h_{i,j}|.$$

Bewijs: Zij $|u^h_{i,j}|$ de maximale waarde, waarbij (i,j) niet correspondeert met een roosterpunt op de rand. Dan volgt uit

$$|a^0_{i,j} u^h_{i,j}| = |\sum_{k=1}^{4} a^k_{i,j} u^h_{i\pm1,j\pm1}|$$

dat

$$|u^h_{i,j}| \leq \sum_{k=1}^{4} \frac{|a^k_{i,j}|}{|a^0_{i,j}|} |u^h_{i\pm1,j\pm1}| \leq |u^h_{i,j}|.$$

Wegens de (zwak) diagonaal dominantie van A kan dit alleen indien voor alle naburen geldt: $|u^h_{i\pm1,j\pm1}| = |u^h_{i,j}|$.
Omdat A irreducibel is kunnen we alle roosterpunten met elkaar verbinden en dus zijn alle $|u^h_{i,j}|$ in het inwendige van Ω_h aan elkaar gelijk. Echter, zodra we bij een onbekende uitkomen die een Dirichlet randpunt als buur heeft, dan geldt

$$|u^h_{i,j}| \leq \sum_{k=1}^{4} \frac{|a^k_{i,j}|}{|a^0_{i,j}|} \max |u^h_{i\pm1,j\pm1}| < \max |u^h_{i\pm1,j\pm1}| \leq |u^h_{i,j}|,$$

en dit leidt tot een tegenspraak. □

Lemma 1 zegt dat het maximum principe voor homogene elliptische differentiaalvergelijkingen ook geldt voor het bijbehorende gediscretiseerde stelsel. Uit Lemma 1 volgt meteen dat het gediscretiseerde homogene probleem $\mathcal{L}_h = 0$, met Dirichlet randvoorwaarde $u^h_{i,j} = 0$ langs de rand Γ_h, alleen de triviale oplossing $u^h_{i,j} = 0$, voor alle i,j toelaat. Blijkbaar is de matrix A van het gediscretiseerde stelsel dus *niet-singulier*. We kunnen deze resultaten samenvatten in de volgende stelling.

Stelling 2 *Zij $A = (a_{i,j})$ een n bij n matrix, die verkregen is met de box methode. Dan geldt:*
(a) A is niet-singulier als A diagonaal dominant is.
(b) A is niet-singulier als aan de volgende 3 voorwaarden voldaan is:

1. *A is zwak diagonaal dominant*
2. *er is een i zodat $|a_{i,i}| > \sum_{j \neq i} |a_{i,j}|$*
3. *de matrix A is irreducibel.*

Opgave. Bewijs stelling 2 voor algemene n bij n matrices.

Een onmiddellijk gevolg van stelling 2 is de volgende stelling.

Stelling 3 *Zij $A = (a_{i,j})$ een symmetrische n bij n matrix met $a_{i,i} > 0$ voor alle i. Dan is A positief definiet indien A diagonaal dominant is of indien voldaan is aan de drie voorwaarden onder (b) in stelling 2.*

Bewijs: Als A symmetrisch is dan zijn al haar eigenwaarden reëel. Vanwege de cirkelstelling van Gershgorin zijn alle eigenwaarden niet-negatief en omdat A niet singulier is, is 0 als eigenwaarde uitgesloten. $\square$

Het opgeleverde discrete stelsel is dus eenduidig oplosbaar, maar er rest nog aan te tonen dat de opgeleverde oplossing in een of andere zin dicht bij de oplossing van het continue probleem ligt. Voor het 1-dimensionale model probleem is zo'n resultaat verkregen in stelling 1. Daarbij is gebruik gemaakt van het feit dat de matrix A een niet-negatieve inverse heeft. Dat kunnen we ook weer voor algemenere gevallen bewijzen. Daarbij komt het volgende lemma van pas.

Lemma 2 *Zij $Au^h = f$ het gediscretiseerde stelsel dat past bij de model vergelijking (1.11). Laat v zo zijn dat $Av \geq 0$ (d.w.z. de componenten van de vector Av zijn niet-negatief), dan geldt $v \geq 0$.*

Bewijs: Laat $v_{i,j}$ het element zijn waarvan de waarde $M \equiv v_{i,j}$ minimaal is en zij $M < 0$. Dan geldt

$$0 \leq \sum_{k=0}^{4} a^k_{i,j}(v_{i\pm1,j\pm1} - M) = \sum_{k=1}^{4} a^k_{i,j}(v_{i\pm1,j\pm1} - M).$$

Echter, $a^k_{i,j} < 0$, voor $k \neq 0$, en dus moet of minstens een der met $v_{i,j}$ verbonden naburen kleiner zijn dan M of alle naburen hebben de waarde M. Dit leidt tot een tegenspraak met de zwak diagonaal dominantie van A. Hieruit concluderen we dat $M > 0$. $\square$

Door de stelsels $Av = e_i$, $i = 1, \ldots, N$, te beschouwen en daar Lemma 2 op toe te passen, concluderen we dat $A^{-1} \geq 0$. Deze eigenschap zal van groot belang blijken voor de numerieke oplossing van de opgeleverde lineaire stelsels. We introduceren daarom de volgende definitie.

Definitie 3 *Een reële n bij n matrix $A = (a_{i,j})$ met $a_{i,j} \leq 0$ voor alle $i \neq j$ heet een M-matrix als A niet singulier is en $A^{-1} \geq 0$.*

Blijkbaar zijn de door de box methode, voor het model probleem, opgeleverde matrices *M-matrices.*

Stelling 4 *De diagonaal elementen van een M-matrix zijn positief.*

Bewijs: Zij $A = (a_{i,j})$ een M-matrix en stel $a_{i,i} \leq 0$. Dan volgt dat $A^{-1}Ae_i \leq 0$ hetgeen een tegenspraak inhoudt. □

Er rest nu nog te bewijzen dat de verkregen oplossing u^h ïn de roosterpunten convergeert"naar u indien $\max_i h_i \to 0$.
We geven een schets van het bewijs. Om onnodig schrijfwerk te vermijden zullen we aannemen dat we te maken hebben met een regelmatig rooster ($h_i = k_j = h$). De oppervlakten der lokale domeinen zijn dus maximaal h^2.

De volgens de box methode gediscretiseerde modelvergelijking $\mathcal{L}(u) = f$ kunnen we schrijven als

$$A\bar{u} = \bar{f} + \tau(\bar{u}).$$

Door met Taylor reeksen te werken volgt $|\tau_i(\bar{u})| \leq Mh^4$, voor zekere $M > 0$, indien we nog aannemen dat $u^{(4)}$ bestaat en begrensd is over Ω. De benaderde oplossing u^h wordt verkregen door oplossing van de vergelijking $Au^h = \bar{f}$ en dus: $\bar{u} - u^h = A^{-1}\tau(\bar{u})$, ofwel

$$|\bar{u} - u^h| \leq Mh^4 A^{-1} e.$$

Om grenzen te verkrijgen voor $A^{-1}e$ beschouwen we de oplossing $v(x)$ van het probleem $\mathcal{L}(v) = 1$, met $v = 0$ langs Γ. We veronderstellen dat v continu en voldoende vaak differentieerbaar is. Dan geldt $A\bar{v} = h^2 e^h$, met $e_i^h = 1 + \nu_i(\bar{v})$, $|\nu_i(\bar{v})| \leq Nh^2$. Dus voor h klein genoeg:

$$\frac{1}{2}h^2 e \leq A\bar{v} \leq 2h^2 e,$$

waaruit volgt

$$A^{-1}e \leq \frac{2}{h^2}\bar{v} \leq \frac{2}{h^2}\sup_{x\in\Omega}|v(x)|e. \tag{1.16}$$

Hieruit concluderen we dat, voor h klein genoeg, geldt

$$\max_i |\bar{u}_i - u_i^h| \leq 2Mh^2 \sup_{x\in\Omega}|v(x)|,$$

en omdat zowel M als $\sup|v(x)|$ niet van h afhangen, is hiermee, althans voor het modelprobleem (1.11) met regelmatig rooster, convergentie (van orde 2 in h) aangetoond. Merk op dat de convergentie orde in dit geval gelijk is aan de consistentie orde (de lokale fout in de differentie vergelijking).

Aannemende dat $\bar{u}$ minstens 6 keer continue differentieerbaar is, kunnen we ook nog aantonen dat, voor homogene roosters, de discretisatiefout $\bar{u} - u^h$ zich op $\mathcal{O}(h^4)$ na gedraagt als $h^2 w$, met

w een gladde functie.
Er geldt namelijk $A\bar{u} = \bar{f} + h^4\mu(\bar{u}) + \mathcal{O}(h^6)$, waarbij $\mu(\bar{u})$ niet van h afhangt ($\mu(\bar{u})$ bestaat uit 4-de afgeleiden van $\bar{u}$).
Wegens $Au^h = \bar{f}$, geldt voor de discretisatiefout

$$A(\bar{u} - u^h) = h^4\mu(\bar{u}) + \mathcal{O}(h^6).$$

Beschouw nu het probleem $\mathcal{L}(w) = \mu(\bar{u})$ met $w = 0$ langs Γ. Dan geldt

$$Aw = h^2\mu(\bar{u}) + \mathcal{O}(h^4).$$

Dus kennelijk

$$\begin{aligned} \bar{u} - u^h &= h^2 w + A^{-1}\mathcal{O}(h^6) \\ &= h^2 w + \mathcal{O}(h^4) \end{aligned}$$

(de laatste overgang is gebaseerd op (1.16)).

Met dit inzicht gewapend kan men zich in actuele situaties experimenteel een indruk verschaffen omtrent de kwaliteit van de gekozen stapgrootte h, door (lokaal) voor verschillende stapgrootten de oplossingen te vergelijken en te verifiëren of hun onderlinge verschillen zich gedragen volgens h^2.

Evenals voor het 1-dimensionale model probleem kunnen we ook nu weer voor het wat algemenere 2-dimensionale model probleem aantonen dat het conditie getal $\kappa_2(A)$ zich gedraagt als $\mathcal{O}(h^{-2})$ (ga dit zelf na). Dit blijkt ook voor meer-dimensionale problemen zo te zijn. Zoals reeds opgemerkt in paragraaf 1.2.1, is $\kappa_2(A)$ van belang voor de convergentie snelheid van sommige populaire iteratieve methoden. Het is opmerkelijk dat, wanneer we meerdimensionale problemen volgens de hier beschreven methoden gaan aanpakken, de orde van de daaruit voortvloeiende lineaire stelsels zich gedraagt als $\mathcal{O}(h^{-d})$, terwijl het conditiegetal zich, onafhankelijk van de dimensie d, blijft gedragen als $\mathcal{O}(h^{-2})$. Dit is een belangrijk argument om voor hoger dimensionale problemen, $d \geq 2$ zeg, het gebruik van iteratieve methoden voor het oplossen van de resulterende lineaire stelsels te rechtvaardigen. We zullen hier in de volgende paragraaf nog op terug komen.

Opmerking 1: Indien in de modelvergelijking (1.11) geldt $Q(x, y) = 0$ over Ω, en als $\alpha(x, y) = 0$ langs de hele rand Γ (dus uitsluitend Neumann randvoorwaarden), dan geldt voor de matrix van het resulterende stelsel dat alle rijsommen gelijk zijn aan 0 (ga na!). Het stelsel is dan singulier, hetgeen geheel in overeenstemming is met het feit dat in dat geval de model vergelijking ook niet eenduidig oplosbaar is (bij iedere oplossing $u(x, y)$ voldoet namelijk ook $u(x, y) + c$). Als men de nulruimte van de betreffende operatoren expliciet kent, dan kan het soms voordelig zijn het singuliere systeem op te lossen, in plaats van het niet-singuliere systeem dat ontstaat door het vastprikken van de oplossing ergens op Γ.

Opmerking 2: De door ons beschouwde model problemen hebben gemeen dat de eerste afgeleide termen (convectie termen) er in ontbreken. Veel van de resultaten is zonder meer ook van toepassing op die gevallen, men dient zich evenwel rekenschap te geven van een aantal essentiële verschillen. We zullen die kort toelichten aan de hand van een eenvoudig voorbeeld (ontleend aan [3]):

$$-u_{xx} - u_{yy} + \frac{1}{\epsilon}(c_1 u_x + c_2 u_y) = 0 \text{ op } \Omega \text{ en } u = g \text{ op } \Gamma. \qquad (1.17)$$

Op $\Omega = [0,1]\times[0,1]$ leggen we een homogeen rooster aan met rooster afstand h en we discretiseren (1.17) via de box methode. De integraal over het subdomein $\omega_{i,j}$ van de convectie termen benaderen we door deze te vervangen door een centraal differentie quotiënt in het centrum punt en dit te vermenigvuldigen met de oppervlakte van het subdomein. Dit leidt voor een inwendig roosterpunt (i,j) tot de volgende differentie vergelijking

$$(-1 - \frac{1}{2\epsilon}c_2 h)u_{i,j-1} + (-1 - \frac{1}{2\epsilon}c_1 h)u_{i-1,j} + 4u_{i,j} + (-1 + \frac{1}{2\epsilon}c_1 h)u_{i+1,j} + (-1 - \frac{1}{2\epsilon}c_2 h)u_{i,j+1} = 0.$$

Merk op dat de symmetrie nu verloren is gegaan, hetgeen voornamelijk consequenties heeft voor het numeriek oplossen van het resulterende lineaire stelsel. De matrix A is nu alleen zwak diagonaal dominant indien $\max(|\frac{1}{2\epsilon}hc_1|, |\frac{1}{2\epsilon}hc_2|) < 1$. Wanneer aan deze ongelijkheid niet is voldaan, dan gaat weliswaar de zwak diagonaal dominantie verloren, maar het lineaire stelsel blijft uniek oplosbaar. Echter in dat geval verliest wel het discrete maximum principe haar geldigheid en naarmate de verhouding h/ϵ toeneemt neemt ook de instabiliteit in de numerieke oplossing toe.
In de volgende tabel geven we waarden van $u^h(x, \frac{1}{2})$ voor $h = \frac{1}{32}$, $c_1 = 1$, $c_2 = 0$, $g(x,y) = (1-x)\sin(\pi y)$ en verschillende waarden van ϵ.

$\Downarrow \epsilon\backslash x \Rightarrow$	$\frac{1}{32}$	$\frac{2}{32}$	$\frac{3}{32}$	$\frac{4}{32}$	$\frac{30}{32}$	$\frac{31}{32}$
0.1	0.97	0.95	0.92	0.89	0.22	0.13
0.02	0.99	0.99	0.98	0.98	0.82	0.73
0.01	1.00	0.99	0.99	0.99	0.87	1.11
10^{-3}	1.03	0.99	1.04	0.99	0.23	1.89
10^{-4}	4.95	0.95	5.00	0.90	0.08	5.88
10^{-5}	48.6	0.94	48.7	0.88	0.06	49.6
10^{-7}	4859.5	0.94	4859.6	0.88	0.06	4860.4

Als u_ϵ de exacte oplossing is van het continue probleem, dan geldt voor iedere $(x,y) \in \Omega$ dat $\lim_{\epsilon\to 0} u_\epsilon = \sin(\pi y)$.
In geval $\epsilon > h/2$, de eerste twee rijen in de tabel, passen de getallen bij deze limiet uitspraak. Zodra echter $\epsilon < h/2$ dan laten de getallen zien dat het discrete maximum principe niet meer geldt $(\max|g(x,y)| \leq 1)$. We worden daarom gedwongen om soms met een onwerkbaar kleine h het probleem op te lossen. We kunnen dat evenwel vermijden door de convectie termen iets anders te discretiseren zodat we wel weer keurig M-matrices krijgen. In ons voorbeeld (1.17) krijgen we dat voor elkaar met geschikt gekozen eenzijdige differenties: voorwaarts danwel achterwaarts, maar zodanig dat de bijdrage aan de diagonaal term van A positief is. Men spreekt dan van *upwind* schema's: de keuze van de eenzijdige differentie wordt bepaald door het teken van c_1, resp. c_2. Deze wijze van discretisatie heeft een konsistentie orde 1 (ga na).
Als $u \in C^4(\Omega)$ dan geldt voor deze diskretizering

$$Au = \mathcal{L}(u) - h(\frac{|c_1|}{\epsilon}\frac{\partial^2}{\partial x^2}u + \frac{|c_2|}{\epsilon}\frac{\partial^2}{\partial y^2}u) + \mathcal{O}(h^2), \quad (h \to 0).$$

De extra term in het rechterlid "versterkt"de diffusie term. Men noemt deze term ook wel *numerieke diffusie* of numerieke viscositeit.
Voorbeelden - Opgaven.

1. $d = 1$, $-u_{xx} = f$, $\Omega = (0, 1)$, $u(0) = \alpha$, $u(1) = \beta$.
Ga na dat, by uniforme roosterafstand $h = \frac{1}{N+1}$, het gediscretiseerde stelsel, bij zowel de ED methode als de box methode, een matrix oplevert van orde N die er als volgt uitziet:

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix},$$

Laat zien dat de vectoren $x^{(k)}$, $k = 1, \ldots, N$, met

$$x_i^{(k)} = \sin(\frac{k\pi i}{N+1}),$$

een volledig stel eigenvectoren voor A vormen. Bepaal hiermee de eigenwaarden λ_k en laat zien dat

$$\frac{\max_k \lambda_k}{\min_k \lambda_k} = (\kappa_2(A) =)\mathcal{O}(h^{-2}).$$

2. $d = 2$, $-u_{xx} - u_{yy} = f$, $\Omega = (0, 1) \star (0, 1)$, $u(x) = g(x)$, $x \in \bar{\Omega}/\Omega$.
Bij uniforme roosterafstand, $h = \frac{1}{N+1}$ in beide richtingen, is de discrete benadering voor de differentiaal operator (volgens zowel de Taylor- als de box methode) een blok tridiagonale matrix van de orde N^2. Ga na dat de diagonaal elementen van A allen gelijk zijn aan 4 en dat de niet-nul elementen van A buiten de hoofddiagonaal allen gelijk zijn aan -1.
De eigenvectoren van A zijn gegeven door $\{x^{k,\ell}\}$, $k = 1, \ldots, N, \ell = 1, \ldots, N$. De coördinaat $x_{i,j}^{k,\ell}$, overeenkomend met roosterpunt (i, j) wordt gegeven door

$$x_{i,j}^{k,\ell} = \sin(\frac{k\pi i}{N+1}) \sin(\frac{\ell\pi j}{N+1}), 1 \le i, j \le N.$$

Bepaal weer de eigenwaarden van A en bepaal de waarde van

$$\frac{\max_{k,\ell} \lambda_{k,\ell}}{\min_{k,\ell} \lambda_{k,\ell}},$$

waarbij $\lambda_{k,\ell}$ de bij $x^{k,\ell}$ passende eigenwaarde is.
Vergelijk het antwoord met het antwoord verkregen voor $d = 1$.

1.4 Oplossen van het gediscretiseerde stelsel

In deze paragraaf zullen we ons bezighouden met de numerieke methoden die in aanmerking komen om de lineaire stelsels, die voortvloeien uit de discretisatie van elliptische zelfgeadjungeerde partiële differentiaal vergelijkingen, op te lossen. Aangezien het oplossen van lineaire stelsels hoofdthema is van het college Numerieke Wiskunde C, richten we hier de aandacht speciaal op die kenmerken van de stelsels die het oplossen van deze gediscretiseerde stelsels beïnvloeden.

Uiteraard zal men graag gebruik maken van het feit dat A slechts weinig elementen ongelijk aan 0 heeft per rij. De in het 1-dimensionale geval opgeleverde bandmatrices kunnen zonder meer met geschikte directe methoden worden opgelost.
Als we voor het gemak maar even aannemen dat in elke coördinaatrichting ongeveer evenveel discretisatie punten (onbekenden) zijn, dan geldt voor directe oplosmethoden, voor de N dimensionale A, dat de hoeveelheid rekenwerk grofweg evenredig is met $N(N^{1-\frac{1}{d}})^2$ (zie [2] of het collegedictaat Numerieke Wiskunde C).
Bij iteratieve methoden is de hoeveelheid rekenwerk evenredig met $n_{it} \cdot N$. Voor, met name, de geconjugeerde gradienten methode wordt n_{it} , nodig voor een vaste foutreductie, begrensd door een van het conditiegetal afhangende grootheid (zie (1.23)). Voor het conditiegetal geldt $\kappa_2(A) \cong h^{-2} \approx N^{\frac{2}{d}}$, en daaruit kan worden afgeleid dat $n_{it} \cong N^{\frac{1}{d}}$ (en soms zelfs nog gunstiger, zie bijv. [8]).
Dus de totale hoeveelheid rekenwerk voor de geconjugeerde gradienten methode is ruwweg evenredig met $N^{1+\frac{1}{d}}$, versus $N^{3-\frac{2}{d}}$ voor de directe oplosmethoden. Voor $d = 2, 3$, en met name voor grotere waarden van N, wordt het dus gaandeweg aantrekkelijker om iteratief te werk te gaan.

De theorie voor iteratieve oplosmethoden, die we hier zullen gaan behandelen, steunt in belangrijke mate op ([10]: hoofdstuk 3.6).
Bij iteratieve oplossing van het stelsel $Au = f$, waarbij A een n bij n niet-singuliere matrix is, gaan we uit van een geschikte *splitsing* van A in de vorm

$$A = M - N,$$

waarbij M en N eveneens n bij n matrices zijn, en zodanig dat M niet-singulier is. Deze splitsing geeft aanleiding tot een basis iteratie methode:

$$Mu^{i+1} = Nu^i + f, \quad i \geq 0,$$

die ook geschreven kan worden als

$$u^{i+1} = M^{-1}Nu^i + M^{-1}f. \tag{1.18}$$

Omdat de oplossing u moet voldoen aan

$$u = M^{-1}Nu + M^{-1}f,$$

volgt voor de iteratie fout $\Delta u^i \equiv u^i - u$ de relatie

$$\Delta u^{i+1} = M^{-1}N\Delta u^i.$$

Dit leidt tot $\Delta u^{i+1} = (M^{-1}N)^{i+1}\Delta u^0$ of in norm:

$$\|\Delta u^{i+1}\| \leq \|(M^{-1}N)^{i+1}\|\|\Delta u^0\|.$$

Voor convergentie van de rij $\{u^i\}$ naar u, voor elke start vector u^0, is het dus noodzakelijk dat $\lim_{i\to\infty}(M^{-1}N)^i = 0$.

Definitie 4 *Een n bij n matrix A heet convergent als* $\lim_{i\to\infty} A^i = 0$.

Stelling 5 *Zij A een $n\star n$ complexe matrix, dan is A convergent dan en slechts dan als* $\rho(A) \equiv \max_{1\leq i\leq n}|\lambda_i| < 1$ *(met λ_i de eigenwaarden van A).*

Bewijs: Als A hermiets is dan geldt $\rho(A) = \|A\|_2$. Met orthogonale Q is A op diagonaal gedaante te brengen: $QAQ^{-1} = D$ en $\|A\|_2 = \|D\|_2$. Dan volgt uit $\|A^i\|_2 = \|D^i\|_2 = \|D\|_2^i$ dat convergentie naar 0 dan en slechts dan geldt als $\rho(A) < 1$.
Wanneer A niet hermiets is dan is de bewijsvoering ingewikkelder. In dat geval is A met niet-singuliere S naar *Jordan normaalvorm* te transformeren:

$$SAS^{-1} = \tilde{A} = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_r \end{pmatrix},$$

met

$$J_\ell = \begin{pmatrix} \lambda_\ell & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix},$$

een $n_\ell \star n_\ell$ Jordan blok.
Als we de elementen van J_ℓ^m noteren met $d_{i,j}^{(m)}(\ell)$ dan geldt

$$d_{i,j}^{(m)}(\ell) = \begin{cases} 0 & j < i \\ \begin{pmatrix} m \\ j-i \end{pmatrix} \lambda_\ell^{m-j+i} & i \leq j \leq \min(n_\ell, m+i) \\ 0 & m+i < j \leq n_\ell \end{cases}$$

Men kan nu weer aantonen dat $\lim_{m\to\infty} J_\ell^m = 0$ dan en slechts dan als $|\lambda_\ell| < 1$.
We laten het aan de lezer over om het bewijs verder rond te maken (of na te zien in ([10]:blz 13). $\square$

Opmerking: Via een soortgelijke bewijsvoering als hiervoor kan men ook laten zien dat

$$\rho(A) = \lim_{i\to\infty} \sqrt[i]{\|A^i\|_2}. \tag{1.19}$$

Wil men verzekerd zijn van convergentie van de rij $\{u^i\}$ dan moet er dus gezocht worden naar splitsingen van A waarvoor geldt dat $\rho(M^{-1}N) < 1$. Een klasse van splitsingen die aan die eis voldoen wordt gevormd door de zogenaamde *reguliere splitsingen*, die we nu ten tonele zullen voeren.

Definitie 5 *Voor een $n \star n$ reële matrix A is $A = M - N$ een reguliere splitsing als M niet singulier is, $M^{-1} \geq 0$ en $N \geq 0$.*

Uit de lineaire algebra theorie voor matrices met een niet-negatieve inverse komt de volgende stelling ([10]:blz.89; we geven hier een elementairder bewijs).

Stelling 6 *Zij $A = M - N$ een reguliere splitsing is van A. Dan is $A^{-1} \geq 0$ dan en slechts dan als $\rho(M^{-1}N) < 1$.*

Bewijs: Stel $NM^{-1}x = \lambda x$ met $|\lambda| \geq 1$.
Met $y = |x|$ geldt dan

$$y \leq |\lambda x| = |NM^{-1}x| \leq NM^{-1}|x| = NM^{-1}y.$$

Hieruit volgt

$$e \equiv (I - NM^{-1})(-y) \geq 0,$$

of

$$(I - NM^{-1})^{-1}e = -y \leq 0.$$

Nu volgt dus dat $M^{-1}(I - NM^{-1})^{-1}e = -M^{-1}y \leq 0$, hetgeen in tegenspraak is met $A^{-1}e = M^{-1}(I - NM^{-1})^{-1}e \geq 0$. Dus geldt $\lambda < 1$.
Uit $NM^{-1}x = \lambda x \Leftrightarrow M^{-1}NM^{-1}x = \lambda M^{-1}x \Leftrightarrow M^{-1}Ny = \lambda y$ volgt dan dat $\rho(M^{-1}N) < 1$.
Omgekeerd, als $\rho(M^{-1}N) < 1$, dan geldt

$$A^{-1} = (I - M^{-1}N)^{-1}M^{-1} = \left(\sum_{j=0}^{\infty}(M^{-1}N)^j\right) M^{-1}.$$

Aangezien $M^{-1} \geq 0$ en $M^{-1}N \geq 0$ volgt daaruit dat $A^{-1} \geq 0$. □

Voor een reguliere splitsing geldt dus dat de matrix $M^{-1}N$ convergent is en de iteratieve methode (1.18) convergeert dan voor iedere startvector u^0.

We staan nu voor de taak om voor de uit discretisatie ontstane lineaire stelsels geschikte reguliere splitsingen te construeren. Een wel zeer eenvoudige splitsing ontstaat door de keuze $M = diag(A)$ (zie stelling 4). De daarmee corresponderende iteratieve methode staat bekend als *Gauss Jacobi methode*.
Met behulp van stelling 6 kunnen we nu bewijzen dat een symmetrische M-matrix positief definiet is. Voor de uit discretisatie van het model probleem (1.11) ontstane M-matrices hadden we dat reeds ingezien. De eigenschap is van belang bij het oplossen van de betreffende lineaire stelsels.

Stelling 7 *Een symmetrische M-matrix A is positief definiet.*

Bewijs: Zij $M = diag(A)$, met $m_{i,i} = a_{i,i} > 0$ (zie stelling 4)
Uit $M^{-1}Ax = \lambda x$ volgt $M^{-1/2}AM^{-1/2}y = \lambda y$ met $y = M^{-1/2}x$. Omdat $M^{-1/2}AM^{-1/2}$ symmetrisch is, is λ reëel en dus heeft $M^{-1}A$ uitsluitend reële eigenwaarden.
Omdat M een reguliere splitsing van A definieert, geldt (stelling 6):

$$\rho(M^{-1}N) = \rho(I - M^{-1}A) < 1.$$

Hieruit volgt dat alle eigenwaarden van $M^{-1}A$ positief zijn en dus is $M^{-1/2}AM^{-1/2}$ positief definiet. Laat nu zelf zien dat dan ook A positief definiet is. $\square$

Intuïtief is duidelijk dat we moeten proberen een zo groot mogelijk deel van A af te splitsen, d.w.z., N moet in een of andere zin klein zijn zodat $\rho(M^{-1}N)$ klein is t.o.v. 1. De ideale keuze zou $M = A$ zijn, omdat in dat geval (1.18) in een iteratieslag convergeert. Evenwel dient M zo gekozen te worden dat stelsels van de vorm $Mx = z$ eenvoudig, dat wil zeggen met weinig rekenwerk en weinig extra computer geheugenruimte, kunnen worden opgelost. De splitsing $M = A$, die neerkomt op direct oplossen, komt daarom voor de wat grotere problemen doorgaans niet in aanmerking.

Een bredere klasse van reguliere splitsingen is gebaseerd op de volgende stelling.

Stelling 8 *Zij $A = (a_{i,j})$ een reële $n \times n$ $M-$matrix en zij $C = (c_{i,j})$ een reële $n \times n$ matrix zo dat $c_{i,i} \geq a_{i,i}$ en $a_{i,j} \leq c_{i,j} \leq 0$, voor $i \neq j$, dan is C ook een M-matrix.*

Bewijs: Stel $A = D - B$ en $C = \tilde{D} - \tilde{B}$, met $D = diag(A) > 0$ en $\tilde{D} = diag(C) > 0$. Uit de voorwaarden van de stelling volgt $\tilde{D}^{-1}\tilde{B} \leq D^{-1}B$. Hieruit volgt dan weer dat

$$\|(\tilde{D}^{-1}\tilde{B})^i\|_2 \leq \|(D^{-1}B)^i\|_2.$$

Uit (1.19) volgt dan $\rho(\tilde{D}^{-1}\tilde{B}) \leq \rho(D^{-1}B) < 1$.
Omdat $\tilde{D} - \tilde{B}$ een reguliere splitsing is van C, is C volgens stelling 6 een M-matrix. $\square$

Met behulp van stelling 8 kan men, voor M-matrices A, zelf nagaan dat de volgende splitsingen regulier zijn (we houden matrices zoals voortvloeiend uit de box methode in gedachten):

- $M =$ benedendriehoeksgedeelte van A (inclusief diagonaal). De resulterende iteratieve methode heet *Gauss Seidel methode*
- $M =$ tridiagonaal gedeelte van A. De resulterende methode heet *blok Gauss Jacobi methode*
- $M =$ blok benedendriehoeksgedeelte van A (inc. blok diagonaal). Dit leidt tot de *blok Gauss Seidel methode.*

De Gauss Seidel methodes kunnen door introductie van een parameter nog belangrijk versneld worden, hetgeen leidt tot de vroeger uitermate populaire *SOR*-methoden (de blok varianten heten ook wel BSOR of SLOR). We gaan hier niet verder op in, voor uitgebreide behandelingen zie [10] of [11].

Een interessante klasse van reguliere splitsingen is gebaseerd op de observatie dat bij het direct oplossen van $Au = f$, via standaard Gauss eliminatie, relatief kleine elementen in A ontstaan door het 'vegen'. Door het weglaten van deze elementen tijdens het eliminatie proces ontstaat zo een onvolledig Gauss eliminatie proces. Voor het preciezer begrijpen van dit proces is de volgende stelling van belang.

Stelling 9 *Gauss eliminatie voert een M-matrix over in een M-matrix.*

Bewijs: We beschouwen een stap in het Gauss eliminatie proces, nl. het vegen van de eerste kolom van A beneden de diagonaal. In matrix notatie $A^{(1)} = L^{(1)}A$ met

$$L^{(1)} = \begin{pmatrix} 1 & & & \\ -\frac{a_{2,1}}{a_{1,1}} & 1 & & \\ \vdots & & \ddots & \\ -\frac{a_{n,1}}{a_{1,1}} & & & 1 \end{pmatrix} \geq 0 \quad .$$

De elementen van $A^{(1)}$ zijn gedefinieerd door $a_{i,j}^{(1)} = a_{i,j} - \frac{a_{1,j}}{a_{1,1}} a_{i,1}$, voor $i > 1$, en $a_{1,j}^{(1)} = a_{1,j}$.
Er geldt blijkbaar $a_{i,j}^{(1)} \leq 0$ voor $i \neq j$.
Rest nog te bewijzen dat $(A^{(1)})^{-1} \geq 0$. Beschouw daartoe $(A^{(1)})^{-1}e_i = A^{-1}(L^{(1)})^{-1}e_i$. Voor $i = 1$ geldt $A^{-1}(L^{(1)})^{-1}e_1 = \frac{1}{a_{1,1}}e_1$. Verder geldt dat $(A^{(1)})^{-1}e_i = A^{-1}e_i \geq 0$, voor $i \neq 1$. Men kan nu het bewijs zelf afmaken. □

Na $n-1$ keren vegen, steeds in de eerste kolom van de resterende ondermatrix, krijgt men

$$A^{(n-1)} = L^{(n-1)} \cdots L^{(1)} A,$$

ofwel $A = \{L^{(n-1)} \cdots L^{(1)}\}^{-1} A^{(n-1)} = LU$, waarbij L beneden driehoeks is, met $\ell_{i,i} = 1$ en $U = A^{(n-1)}$ bovendriehoeks.
De werkwijze bij het onvolledig vegen is nu als volgt. Men bepaalt, naar keuze, vooraf of tijdens het vegen welke elementen men wenst te vegen (dus welke elementen men wenst mee te nemen in L) en welke correcties (bv. $-\frac{a_{k,j}^{(k)}}{a_{k,k}^{(k)}} a_{j,k}^{(k)}$) men nog wenst te handhaven in U. Er ontstaat zo een *incomplete decompositie* van A:

$$A = \tilde{L}\tilde{U} - N.$$

De bijbehorende iteratieve methode, die ontstaat door $M = \tilde{L}\tilde{U}$ te kiezen, convergeert wegens $(\tilde{L}\tilde{U})^{-1} \geq 0$ (ga na).

De volgende stelling vat de voorgaande resultaten samen voor de situatie waarin we vooraf vastleggen welke elementen in $\tilde{L}$ en $\tilde{U}$ we in het eliminatie proces wensen te negeren. De indices van die elementen leggen we vast in de set P:

$$P \subset P_n \equiv \{(i,j) | i \neq j, 1 \leq i \leq n, 1 \leq j \leq n\}. \tag{1.20}$$

Stelling 10 *Zij $A = (a_{i,j})$ een $n \times n$ M-matrix, dan bestaat er voor elke $P \subset P_n$ een benedendriehoeksmatrix $\tilde{L} = (\ell_{i,j})$, met $\ell_{i,i} = 1$, een bovendriehoeksmatrix $\tilde{U} = (u_{i,j})$ en een matrix $N = (n_{i,j})$ met*

- $\ell_{i,j} = 0$, $u_{i,j} = 0$, *als* $(i,j) \in P$
- $n_{i,j} = 0$ *als* $(i,j) \notin P$,

zodat de splitsing $A = \tilde{L}\tilde{U} - N$ een reguliere splitsing is.
De factoren $\tilde{L}$ en $\tilde{U}$ zijn uniek bepaald.

Een speciaal geval is, voor symmetrische M-matrices, de zogenaamde incomplete Choleski decompositie, waarvan het bestaan gegarandeerd wordt door de volgende stelling.

Stelling 11 *Zij A een symmetrische M-matrix, dan bestaat er voor elke $P \in P_n$ met de eigenschap dat met $(i,j) \in P$ ook $(j,i) \in P$, een uniek bepaalde benedendriehoeksmatrix $L = (\ell_{i,j})$ en een symmetrische niet-negatieve matrix $N = (n_{i,j})$, met $\ell_{i,j} = 0$ als $(i,j) \in P$ en $n_{i,j} \notin P$, zodanig dat de splitsing $A = LL^T - N$ regulier is.*

Een voorbeeld van incomplete decompositie

We beschouwen weer het symmetrische bloktridiagonale stelsel dat ontstaat na discretisatie van het modelprobleem (1.11, zie paragraaf 1.3). Ter beperking van het rekenwerk per iteratieslag willen we graag een incomplete decompositie hebben waarvan de componenten ijl gevuld zijn. Een geschikte keuze voor de set P blijkt te zijn

$$P^\star = \{(i,j) \mid |i-j| \neq 0, 1, n_x\}$$

dat wil zeggen, we volgen voor L het ijlheidspatroon van A. De elementen van L (en dus ook van L^T) volgen uit de eis dat $n_{i,j} = 0$ als $(i,j) \notin P^\star$, dus blijkbaar geldt voor het (i,j)-de element van A, mits $(i,j) \in P^\star$:

$$(LL^T)_{i,j} = a_{i,j}.$$

Ga na dat hierdoor de elementen van L zijn vastgelegd.

Dit soort regelmatige gestructureerde matrices wordt doorgaans op zuinige wijze in een computergeheugen opgeslagen. In A bevatten slechts 5 (neven)diagonalen elementen ongelijk aan nul. We kunnen daarom A representeren door drie van deze diagonalen op te slaan (de andere twee volgen uit symmetrie):
- De elementen van de (hoofd)diagonaal noteren we als a_i,
- de elementen van de eerste bovennevendiagonaal noteren we als b_i,
- en de elementen van de n_x-de bovennevendiagonaal noteren we met c_i.

$(i = 1, \ldots, n)$, evt. de arrays aanvullen met nullen.
Op dezelfde wijze noteren we de elementen van L^T met $\tilde{a}_i$, $\tilde{b}_i$ en $\tilde{c}_i$. Leid zelf betrekkingen af voor de elementen $\tilde{a}_i$, $\tilde{b}_i$ en $\tilde{c}_i$ in termen van de elementen van A.

We zullen nu gaan zien dat het handig is deze speciale incomplete decompositie te herschrijven in de vorm $\bar{L}\bar{D}\bar{L}^T$, met $\bar{D}$ een diagonaal matrix. Ga zelf na dat de elementen van $\bar{L}^T$ en $\bar{D}$ vastgelegd worden door de relaties

$$\left\{ \begin{array}{ll} \bar{b}_i = b_i & \bar{c}_i = c_i \\ \bar{a}_i = \bar{d}_i^{-1} = a_i - \bar{b}_{i-1}^2 \bar{d}_{i-1} - \bar{c}_{i-n_x}^2 \bar{d}_{i-n_x} & \end{array} \right. \quad i = 1, 2, \ldots, n$$

(waarbij niet gedefinieerde elementen gewoon door nullen worden vervangen).
Het opmerkelijke is dat de incomplete decompositie in dit geval dus volledig vastgelegd wordt door de elementen van $\bar{D}$. De bijbehorende basis iteratieve methode ziet er dan uit als

$$\bar{L}\bar{D}\bar{L}^T u_{i+1} = f + N u_i$$

of

$$\bar{L}\bar{D}\bar{L}^T u_{i+1} = f + (\bar{L}\bar{D}\bar{L}^T - A)u_i$$
$$u_{i+1} = u_i + (\bar{L}\bar{D}\bar{L}^T)^{-1}(f - Au_i) = u_i + \Delta u_i.$$

De correctie Δu_i wordt opgelost uit het stelsel

$$\bar{L}\bar{D}\bar{L}^T \Delta u_i = f - Au_i \tag{1.21}$$

$$\begin{aligned} \text{via } \bar{L}w &= f - Au_i \Rightarrow w \\ \bar{D}v &= w \Rightarrow v \\ \text{en } \bar{L}^T \Delta u_i &= v \Rightarrow \Delta u_i. \end{aligned} \tag{1.22}$$

Men gaat eenvoudig na dat de hoeveelheid rekenwerk, nodig voor het oplossen van deze 3 stelsels (waarvan de laatste twee efficiënt gecombineerd kunnen worden), ongeveer net zo groot is als die voor het berekenen van Au_i (voor gegeven u_i). Per iteratieslag is deze splitsing dus ongeveer 2 keer zo duur als bijvoorbeeld de Gauss Jacobi en Gauss Seidel iteratie methode.

1.4.1 Incomplete decomposities in de praktijk

Incomplete decomposities geven zelf vaak al aanleiding tot aanzienlijk snellere convergentie van de basis iteratie methode in vergelijking met bijvoorbeeld de Gauss Jacobi methode. Zij worden tegenwoordig echter hoofdzakelijk gebruikt als zg. *preconditionering* voor optimale iteratie methoden als *geconjugeerde gradienten methode* en *GMRES*. Deze methoden worden in het college Numerieke Analyse C behandeld, maar we gaan er voor de volledigheid hier kort op in.

Deze optimale, of beter *projectie methoden*, gaan in wezen uit van de basis iteratie methode (1.18) toegepast op de splitsing $A = I - (I - A)$. Dit is *Richardson iteratie* en, als we aannemen dat de M-matrix A (symmetrisch) geschaald is zodat $diag(A) = I$, dan is de splitsing regulier. Met $r_0 = f - Au_0$ gaat men met volledige inductie eenvoudig na dat na i slagen geldt

$$u_i = u_0 + y, \qquad y \in \{r_0, Ar_0, \ldots, A^{i-1}r_0\} \equiv K^i(A, r_0).$$

Als A symmetrisch positief definiet is dan blijkt het mogelijk te zijn om met relatief weinig extra rekenwerk, in plaats van deze specifieke y, dat element $y \in K^i(A, r_0)$ te bepalen waarvoor $\|(u_0 + y) - u\|_A$ mimimaal is ($\|v\|_A^2 \equiv (v, Av)$). Het algoritme waar dit mee tot stand gebracht wordt heet de geconjugeerde gradienten methode [2]. Wanneer$A$ niet positief definiet is dan kan men bijvoorbeeld die $y \in K^i(A, r_0)$ bepalen waarvoor $\|A(u_0 + y) - b\|$ minimaal is. Een methode die dat tot stand brengt is GMRES [7].

Voor de iteratiefout bij de geconjugeerde gradienten methode geldt

$$\|u_i - u\|_A \leq 2\left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1}\right)^i \|u_0 - u\|_A. \tag{1.23}$$

Ook de eigenwaardenverdeling speelt nog een rol, in die zin dat bij een betere spreiding van de uiterste eigenwaarden (bij gelijkblijvende $\kappa_2(A)$) de bovengrens in (1.23) vervangen kan worden door een, voor toenemende i, lagere bovengrens (zie [8]).

Opgave: Voor het modelprobleem (1.11) kan men zelf nagaan dat de Gauss Jacobi methode een foutreductie factor heeft die er uit ziet als $1-r$, met $r>0$ en $r=\mathcal{O}(h^2)$. De geconjugeerde gradienten methode leidt tot een reductiefactor van tenminste $1-s$, met $s>0$ en $s=\mathcal{O}(h)$ (ga na).

Opgave: Bepaal voor de twee voorbeelden in paragraaf 1.3.1 de foutreductie factoren voor zowel Gauss Jacobi als de geconjugeerde gradienten methode.

Door nu de geconjugeerde gradienten methode toe te passen op het (gepreconditioneerde) stelsel $\tilde{A}\tilde{u} \equiv L^{-1}AL^{-T}\tilde{u} = L^{-1}b$, met $u = L^{-T}\tilde{u}$, krijgt men te maken met $\kappa_2(\tilde{A})$ en een andere eigenwaarden spreiding. Uit analyse aan model problemen, zowel als uit ervaringen met praktijk problemen, blijkt dat $\kappa_2(\tilde{A})$ vaak veel kleiner is dan $\kappa_2(A)$ en de eigenwaarde verdeling is eveneens veel gunstiger.

Opmerking: In het college numerieke analyse C wordt getoond hoe de gepreconditioneerde geconjugeerde gradienten methode elegant kan worden uitgevoerd met $\bar{L}\bar{D}\bar{L}^T$, zonder dat het nodig is $\bar{D}$ over beide factoren te splitsen, en waarbij de terugtransformatie voor $\tilde{u}$ vermeden wordt.

Om een indruk te geven van de soms formidabele rekenwinsten die men kan behalen citeren we een resultaat uit [5], voor het oplossen van een zeker elliptisch probleem. We geven slechts de iteratie aantallen, nodig voor het reduceren van de beginfout met een factor 10^{-6}, voor een 3-tal methoden.

methode	aantal iteraties
Gauss Seidel	208000
BSOR	765
geconjugeerde gradienten met $\bar{L}\bar{D}\bar{L}^T$	25

Bedenk hierbij nog wel dat het rekenwerk voor de laatste methode grofweg twee keer zo groot is per iteratieslag als het rekenwerk voor de beide andere methoden.

Hoofdstuk 2

Parabolische differentiaal vergelijkingen

2.1 Inleiding

De in het vorige hoofdstuk behandelde elliptische differentiaal vergelijkingen spelen ook een grote rol in andere partiële differentiaal vergelijkingen:

- De *warmte vergelijking:*

$$\begin{cases} U_t = \mathcal{L}(U) & op & (0,\infty) \star \Omega \\ U(0,x) = \phi(x) & voor & x \in \Omega \\ \mathcal{R}(U(t,\cdot)) = g(t,\cdot) & op & \Gamma, \text{ voor iedere } t > 0, \end{cases}$$

  waarbij $(t,x) \rightsquigarrow U(t,x)$ een reëelwaardige functie is ($t \in [0,\infty)$, $x \in \bar{\Omega}$), en waarbij de elliptische operator $\mathcal{L}$ en de restriktie operator $\mathcal{R}$ beide voor iedere $t > 0$ werken op de functie $x \rightsquigarrow U(t,x)$, $(x \in \Omega)$. ϕ is een functie op Ω die de *beginvoorwaarde* beschrijft.
  De eerste variabele t noemt men de tijd-variabele en de variabelen $x = (x_1, \ldots, x_d)$ de ruimte-variabele(n). Zoals we aanstonds zullen zien beschrijft de vergelijking onder andere de temperatuursverdeling in een lichaam Ω. In de wiskunde is de warmtevergelijking het standaard voorbeeld van een parabolische partiële differentiaalvergelijking (merk op dat $\tilde{\mathcal{L}}(U) \equiv \mathcal{L}(U) - U_t$ niet elliptisch is).

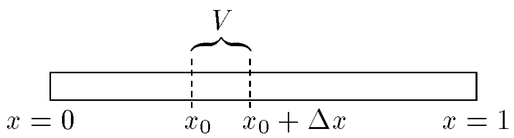
- De *golf vergelijking:*

$$\begin{cases} U_{tt} = \mathcal{L}(U) & op & (0,\infty) \star \Omega \\ U(0,x) = \phi(x), U_t(0,x) = \psi(x) & voor & x \in \Omega \\ \mathcal{R}(U(t,\cdot)) = g(t,\cdot) & op & \Gamma, \text{ voor iedere } t > 0. \end{cases}$$

  ϕ en ψ zijn functies op Ω die de beginvoorwaarden beschrijven. De golf vergelijking treedt op in diverse fysische modellen waarin golven beschreven worden (electrodynamica, akoustiek, stromingsleer, etc.). In de wiskunde is de golf vergelijking het standaard voorbeeld van een hyperbolische partiële differentiaalvergelijking.

2.2 De warmte vergelijking

Als voorbeeld van een parabolische differentiaal vergelijking zullen we de vereenvoudigde *warmte-vergelijking* in 1 (ruimte) dimensie beschouwen. We geven eerst een korte weergave van de fysische herkomst van deze vergelijking. Het uitgangspunt is een dunne geïsoleerde homogene staaf:



Van deze staaf beschouwen we een segmentje V ter lengte Δx. Over V wordt de wet van behoud van energie toegepast. Deze wet zegt dat de energie die V binnenstroomt, via de randen, aanleiding geeft tot een zekere temperatuurstijging binnen V.

Laat verder $T(t,x)$ de temperatuur weergeven en $H(t,x)$ de warmtestroom. De netto warmte I, die V via de randen binnenstroomt van tijdstip t_0 tot $t_0 + \Delta t$ wordt beschreven door

$$I = \int_{t_0}^{t_0+\Delta t} \{H(t, x_0 + \Delta x) - H(t, x_0)\} dt.$$

Deze warmte I is, via de wet van behoud van energie, gerelateerd aan de temperatuur toename binnen V in het tijdsinterval Δt:

$$I = \int_{x_0}^{x_0+\Delta x} \rho c \{T(t_0 + \Delta t, x) - T(t_0, x)\} dx.$$

Hierin is ρ de dichtheid van de staaf en c de specifieke warmte.
Door gelijkstelling van beide expressies voor I en na deling door $\Delta t \Delta x$, volgt

$$\frac{1}{\Delta x} \int_{x_0}^{x_0+\Delta x} \rho c \frac{T(t_0 + \Delta t, x) - T(t_0, x)}{\Delta t} dx = \frac{1}{\Delta t} \int_{t_0}^{t_0+\Delta t} \frac{H(t, x_0 + \Delta x) - H(t, x_0)}{\Delta x} dt.$$

Via een limiet overgang $\Delta t, \Delta x \to 0$ volgt dan voor ieder punt (t_0, x_0):

$$\rho c \frac{\partial T}{\partial t}(t_0, x_0) = \frac{\partial H}{\partial x}(t_0, x_0). \tag{2.1}$$

De (empirische) wet van Fourier relateert de warmtestroom aan de temperatuur:

$$H = k \frac{\partial T}{\partial x}. \tag{2.2}$$

Hierin is k de geleidingscoëfficient. Substitutie van (2.2) in (2.1) leidt dan tot de vereenvoudigde warmtevergelijking:

$$\rho c T_t = \frac{\partial}{\partial x} k \frac{\partial T}{\partial x}.$$

Voor constante ρ, c en k reduceert, na geschikte schaling, deze vergelijking tot de modelvergelijking:

$$\begin{cases} U_t = U_{xx} & op \quad (0,\infty) \star (0,1) \\ U(0,x) = U_0(x) & voor \quad x \in [0,1] \\ U(t,0) = U_\ell(t), U(t,1) = U_r(t) & op \quad \Gamma, \text{ voor iedere } t > 0, \end{cases}$$

e discretisatie van de ruimtelijke afgeleiden in de warmtevergelijking (2.3) kan gebeuren met de in het vorige hoofdstuk besproken discretisatie technieken. Met verwerking van de randvoorwaarden krijgen we zo, na weglating van de discretisatie fouten, voor de benaderde onbekenden $u_i(t)$, in de roosterpunten op $[0,1]$, het stelsel gewone differentiaal vergelijkingen:

$$\frac{du}{dt} = -Au + f + \tau(u), \tag{2.3}$$

met A als in paragraaf 1.2.1. Met name is hier relevant dat $A^{-1} \geq 0$ en dat alle eigenwaarden van A positief zijn.
Men noemt dit semi-discretisatie. Het resulterende stelsel kan in de tijd geïntegreerd worden met geschikte methoden uit het eerste deel van dit college. De combinatie van semi-discretisatie en aparte tijdsintegratie staat bekend als de *methode der lijnen.*

Als $|u^{(4)}|$ begrensd is dan geldt $|\tau(u)| \leq Mh^2$. De numerieke oplossing u^h wordt verkregen door het stelsel op te lossen dat ontstaat door in (2.3) de foutterm $\tau(u)$ te verwaarlozen:

$$u^h_t = -Au^h + f. \tag{2.4}$$

Voor de discretisatiefout $\Delta u \equiv u - u^h$ geldt dus

$$(\Delta u)_t = -A\Delta u + \tau(u), \quad \Delta(u)_{|t=0} = 0.$$

Uit de theorie voor differentiaal ongelijkheden (zie bijv. [4]:blz 30) volgt dat

$$v^- \leq \Delta u \leq v^+,$$

met

$$v^+_t = -Av^+ + Mh^2 e, \quad v^+_{|t=0} = 0,$$

$$v^-_t = -Av^- - Mh^2 e, \quad v^-_{|t=0} = 0.$$

Deze vergelijkingen zijn eenvoudig oplosbaar en men gaat gemakkelijk na dat hieruit volgt:

$$-A^{-1}Mh^2(1 - e^{-At})e \leq \Delta u \leq A^{-1}Mh^2(1 - e^{-At})e.$$

De vector e is de vector met alle elementen gelijk aan 1.
Voor elke t is de fout in u^h, tengevolge van de ruimte discretisatie, dus $\mathcal{O}(h^2)$ (zie bewijs van stelling 1: $A^{-1}h^2 e = \mathcal{O}(h^2)e$). Soortgelijke resultaten zijn eenvoudig af te leiden, voor meerdimensionale problemen, met de in Hoofdstuk 1 gepresenteerde theorie.

Voor het oplossen van het stelsel gewone differentiaalvergelijkingen (2.4) kunnen we in principe methoden gebruiken die in het eerste deel van dit college behandeld zijn. Bedenk hierbij wel dat

$$\frac{\lambda_{max}(A)}{\lambda_{min}(A)} = \mathcal{O}(h^{-2}),$$

zodat bij kleinere h het stelsel *stijf* is. Dat betekent dat expliciete methoden, zoals voorwaarts Euler, een zeer kleine tijdstap vereisen en men dus veelal de voorkeur geeft aan een geschikt gekozen impliciete methode.

Bibliografie

[1] C.Cuvelier, A.Segal and A.A.Van Steenhoven, *Finite Element Methods and Navier-Stokes Equations*, Reidel Publishing Company, Dordrecht, 1986

[2] G.H.Golub and C.F.Van Loan, *Matrix Computations*, North Oxford Academic, Oxford, 1983

[3] W.Hackbusch, *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner Studienbücher, Mathematik, B.G.Teubner, Stuttgart, 1986

[4] J.K.Hale, *Ordinary differential equations*, Interscience Series on Pure and Applied Mathematics, **Vol 21**, Wiley-Interscience, New York, 1969

[5] D.S.Kershaw, *The incomplete choleski-conjugate gradient method for the iterative solution of systems of linear equations*, J.of Comp.Phys., **26**, 1978, pp43-65

[6] J.A.Meijerink and H.A.van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math.of Comp., **31**, 1977, pp148-162

[7] Y.Saad and M.H.Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J.Sci.Statist.Comput., **7**, 1986, pp856-869

[8] A.van der Sluis and H.A.van der Vorst, *The rate of convergence of conjugate gradients*, Numer.Math., **48**, 1986, pp543-560

[9] G.Strang and G.J.Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973

[10] R.S.Varga, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962

[11] D.M.Young, *Iterative solution of large linear systems*, Academic Press, New York, 1971