# Numerieke Oplosmethoden

voor

# **Gewone Differentiaalvergelijkingen;**

## **Beginwaarde Problemen**

Gerard L.G. Sleijpen

Mathematisch Instituut
Universiteit Utrecht

Utrecht, mei 1991 (kor: januari 1996)

# Numerieke Oplosmethoden

voor

## Gewone Differentiaalvergelijkingen;

### Beginwaarde Problemen

Gerard L.G. Sleijpen

Mathematisch Instituut
Universiteit Utrecht

Utrecht, mei 1991 (kor: januari 1996)

# Voorwoord

We behandelen in dit college de theorie over het numerieke oplossen van beginwaarde problemen van gewone differentiaalvergelijkingen. Deze theorie is uiteraard van belang voor het numeriek benaderd oplossen van dit soort problemen, maar speelt ook een essentiële rol bij het oplossen van partiële tijdsafhankelijke differentiaalvergelijkingen. In college "Numerieke Partiële Differentiaalvergelijkingen" worden echter pas numerieke methoden voor deze partiële differentiaalvergelijkingen in detail besproken.

We bespreken een aantal fundamentele numerieke oplosmethoden voor gewone differentiaalvergelijkingen. Hoewel we ook methoden geven ligt onze nadruk op de wiskundige en numeriek wiskundige achtergrond. We besteden veel aandacht aan de stabiliteit en de efficiëntie van de methoden. Het inzicht dat we daardoor hopen te kunnen geven zal behulpzaam zijn ook bij problemen en methoden die we hier niet behandelen.

Formeel veronderstellen we geen geavanceerde wiskundige voorkennis. We gebruiken wat analyse en algebra uit de eerste jaars colleges. We manipuleren echter herhaaldelijk op eenvoudige manier met normen op vektoren en de bijbehorende normen op matrices. In het college "Numerieke Lineaire Algebra" (Numerieke Wiskunde C) is routine opgedaan met dit soort manipulaties. Het college gaat over het numeriek oplossen van gewone differentiaalvergelijkingen. Op een enkele stelling na, over het bestaan van de oplossing (die we overigens uitgebreidt zullen citeren) en het oplossen van lineaire differentiaalvergelijkingen, gebruiken we geen voorkennis uit de theorie van de differentiaalvergelijkingen. Voor een goede motivatie is het wel prettig om enig gevoel te hebben voor differentiaalvergelijkingen en de bijbehorende oplossingen.
Een zelfde opmerking geldt met betrekking tot de andere numerieke wiskunde colleges: dit college is er volledig onafhankelijk van, maar een gevoel voor de numerieke problematiek (van stabiliteit en efficiëntie en weten van evaluatiefouten, etc.) zal dit vak interessanter maken.

De methoden die we behandelen kunnen eenvoudig op een PC geïmplementeerd worden. De lezer zal meer schik in het vak hebben als hij/zij zo nu en dan de beweringen middels een numeriek experiment toetst aan de werkelijkheid. Het programma MATLAB laat zeer eenvoudige codes toe en maakt het mogelijk de resultaten grafisch te representeren.

De passage's die gemarkeerd zijn met een $\circ$ hoeven niet voor het tentamen bestudeerd te worden.

Gerard Sleijpen

Mathematisch Instituut
Universiteit Utrecht

Utrecht, januari 1996

# Inhoudsopgave

## Notaties en conventies

We herhalen hier wat notaties uit ander colleges die wellicht niet dagelijks gebruikt worden. Een index voor de nieuwe notaties van dit diktaat vindt je op pagina 88.

$\mathbf{N}$, $\mathbf{N}_0$, $\mathbf{Z}$, $\mathbf{R}$, $\mathbf{C}$ is de verzameling van, respektievelijk, de natuurlijke getallen ($\geq 1$), de gehele getallen $\geq 0$, de gehele getallen, de reële getallen en de complexe getallen.
Met $\#$ van een verzamekling bedoelen het aantal elementen in die verzameling.

Onze funkties zijn vektor waardig: $C(\mathcal{J}, \mathbf{R}^d)$ is bijvoorbeeld de ruimte van alle kontinue funktie van een reeel interval $\mathcal{J}$ naar $\mathbf{R}^d$, waarbij $\mathbf{R}^d$ de ruimte van kolom vektoren $(x_1, \ldots, x_d)^T$ met reële coëfficiënten. Met $u'$ voor $u$ in de ruimte $C^1(\mathcal{J}, \mathbf{R}^d)$ van kontinu differenteerbare funkties uit $C(\Omega, \mathbf{R}^d)$ bedoelen koördinaatsgewijze differentiatie.

Vektoren meten we met een norm $\|\cdot\|$. Bij $\|x\|$ can je aan je favoriete norm denken, aan bijvoorbeeld een $p$-norm:

$$\|x\|_p := (|x_1|^p + \cdots + |x_d|^p)^{\frac{1}{p}} \quad \text{voor } p \in [1, \infty) \quad \text{of} \qquad \|x\|_\infty := \max_j |x_j| \ \text{ voor } p = \infty.$$

Slechts op een paar plaatsen in dit diktaat is het van belang te weten welke norm precies bedoeld wordt. Gewoonlijk is dat dan de 2-norm $\|\cdot\|_2$ omdat die gerelateerd aan het *standaard inprodukt* $<\cdot, \cdot>$:

$$\|x\|_2 = \sqrt{<x, x>},$$

waarbij

$$<x, y> := x_1 y_1 + \ldots + x_d y_d \quad \text{voor } x = (x_1, \ldots, x_d)^T,\ y = (y_1, \ldots, y_d)^T \in \mathbf{R}^d.$$

Inprodukten hebben prettige eigenschappen (Cauchy-Schwartz, etc..) die in de theorie soms nuttig zijn. In de praktijk speelt de *max-norm* $\|\cdot\|_\infty$ een grotere rol.

$\mathbf{M}_d(\mathbf{R})$ is de ruimte van alle reële $d \times d$ matrices.
Voor $A \in \mathbf{M}_d(\mathbf{R})$ is $\|A\|$ de *norm van* $A$ die geassocieerd is met de norm $\|\cdot\|$ op $\mathbf{R}^d$:

$$\|A\| := \max\left\{ \|Ax\| \,\Big|\, \|x\| = 1 \right\}.$$

Het *conditie getal* van $A$

$$\mathcal{C}(A) := \|A\| \, \|A^{-1}\|$$

geeft vaak aan hoe gevoelig berekeningen met $A$ zijn voor fouten.
In geval van een $p$-norm op $\mathbf{R}^d$ schrijven we $\|A\|_p$ en $\mathcal{C}_p(A)$ i.p.v. $\|A\|$ , resp. $\mathcal{C}(A)$.
Voor een paar specifieke normen hebben we relaties die berekeningen kunen vergemakkelijken:

$$\|A\|_\infty = \max_i \sum_j |\alpha_{ij}|, \quad \|A\|_1 = \max_j \sum_i |\alpha_{ij}| \quad \text{voor} \quad A = (\alpha_{ij}) = \begin{bmatrix} \alpha_{11} & \ldots & \alpha_{1d} \\ \vdots & & \vdots \\ \alpha_{d1} & \ldots & \alpha_{dd} \end{bmatrix},$$

$$\|A\|_2 = \max\{ |\lambda| \,\Big|\, \lambda \text{ eigenwaarde } A^T A \}.$$

Partiële afgeleiden geven we compact weer: $u_x$ in plaats van $\frac{\partial u}{\partial x}$, etc..

# 1 Gewone differentiaalvergelijkingen

## 1.1 Het kontinue probleem

In deze paragraaf beschrijven we precies het probleem waarvan we numerieke oplosmethoden zullen bestuderen. Verder geven we een stelling die ons vertelt wanneer het probleem één oplossing heeft.

We noemen het probleem "bepaal de (differentieerbare) oplossing van de differentiaalvergelijking" *kontinu* ter onderscheiding van het *diskrete* probleem "bepaal een aantal funktiewaarden van een zekere benaderende oplossing".

### 1.1.1 Het beginwaarde probleem.

Zij $d \in \mathbf{N}$, $t_0 \in \mathbf{R}$ en $T > 0$. Zij $\mathcal{J} := [t_0, t_0 + T]$ of $\mathcal{J} := [t_0, \infty)$
Zij $\Omega$ een open deel van $\mathbf{R} \times \mathbf{R}^d$. Laat $f \in C(\Omega, \mathbf{R}^d)$ en $u_0 \in \mathbf{R}^d$ zodat $(t_0, u_0) \in \Omega$.

We zijn geïnteresseerd in een funktie $u \in C^1(\mathcal{J}, \mathbf{R}^d)$ waarvoor geldt

$$\begin{cases} u'(t) := \dfrac{d(u)}{dt}(t) = f(t, u(t)) \quad \text{voor alle} \quad t \in \mathcal{J}, \\ u(t_0) = u_0 \end{cases} \tag{1}$$

Het is niet duidelijk of probleem (1) een oplossing $u$ in $C^1(\mathcal{J}, \mathbf{R}^d)$ heeft met *grafiek* $\{(t, u(t)) \big| t \in \mathcal{J}\}$ binnen het definitie gebied $\Omega$ van $f$. Het is ook niet duidelijk of er hoogstens een oplossing is. Voordat we een stelling formuleren over het bestaan van zo'n oplossing bekijken we eerst een tweetal illustratieve voorbeeldjes.

**1.1.2 Voorbeeld.** Beschouw, met $f(t, x) := x^2$ voor $(t, x) \in \Omega := \mathbf{R} \times \mathbf{R}$ en $u_0 = \alpha > 0$, het probleem

$$u'(t) = u^2(t) \quad \text{voor} \quad t \geq 0 \quad \text{en} \quad u(0) = \alpha.$$

Voor $t \in [0, \frac{1}{\alpha})$ heeft dit probleem als enige oplossing $u(t) = \frac{\alpha}{1 - \alpha t}$.

Voor geen $T > \frac{1}{\alpha}$ vinden we, met $\mathcal{J} = [0, T]$, een oplossing op de hele $\mathcal{J}$!

**1.1.3 Voorbeeld.** Beschouw, met $f(t, x) := 2\sqrt{|x|}$ voor $(t, x) \in \Omega := \mathbf{R} \times \mathbf{R}$ en $u_0 = 0 \in \mathbf{R}$, het probleem

$$u'(t) = 2\sqrt{|u(t)|} \quad \text{voor} \quad t \geq 0 \quad \text{en} \quad u(0) = 0.$$

Dit probleem heeft de oplossingen $u(t) = 0 \quad (t \geq 0)$ en $u(t) = t^2 \quad (t \geq 0)$.

Voor geen enkele $\varepsilon > 0$ is, met $\mathcal{J} = [0, \varepsilon]$, de oplossing op $\mathcal{J}$ uniek!

De volgende stelling geeft ons voorwaarden waaronder (1) in ieder geval in de buurt van $t_0$ een unieke oplossing. Men kan de stelling bewijzen met behulp van zogenaamde Picard iteratie; we geven hier geen bewijs. We verklaren eerst nog een frase die we in de stelling gebruiken.

Als $\tau_{-1}, \tau_1 \in \mathbf{R}$, $\tau_{-1} < t_0 < \tau_1$ en $u : (\tau_{-1}, \tau_1) \to \mathbf{R}^d$ dan *loopt* de grafiek van $u$ op $(\tau_{-1}, \tau_1)$ binnen $\Omega$ *tot de rand* $\partial\Omega := \bar{\Omega} \backslash \Omega$ van $\Omega$ als $\{(t, u(t)) \big| t \in (\tau_{-1}, \tau_1)\} \subset \Omega$ en

$$\tau_1 = \infty \quad \text{of} \quad \lim_{t \uparrow \tau_1} (t, u(t)) \in \partial\Omega,$$
$$\tau_{-1} = -\infty \quad \text{of} \quad \lim_{t \downarrow \tau_{-1}} (t, u(t)) \in \partial\Omega.$$

**1.1.4 Stelling.** *Beschouw* 1.1.1.
*Stel dat* $f$ uniform Lipschitz kontinu is in de tweede variabele:

(Dif.1) $\qquad \exists L > 0 \quad zodat \quad \|f(t,x) - f(t,y)\| \leq L\,\|x-y\| \quad \forall (t,x),(t,y) \in \Omega.$

*Dan geldt het volgende.*
(a) *Er zijn* $\tau_{-1}, \tau_1 \in \mathbf{R}$, $\tau_{-1} < t_0 < \tau_1$ *en een* $u \in C^1((\tau_{-1},\tau_1),\mathbf{R}^d)$ *waarvan de grafiek op* $(\tau_{-1},\tau_1)$ *loopt binnen* $\Omega$ *tot de rand* $\partial\Omega$ *en waarvoor*

$$u'(t) = f(t,u(t)) \quad \textit{voor alle} \quad t \in (\tau_{-1},\tau_1) \quad \textit{en} \quad u(t_0) = u_0.$$

*Als* $\mathcal{J} \times \mathbf{R}^d \subset \Omega$ *dan* $\mathcal{J} \subset (\tau_{-1},\tau_1)$.
(b) *De grafiek van* $u$ *vertakt zich niet op* $(\tau_{-1},\tau_1)$
(als $\tilde{t}_0 \in (\tilde{\tau}_{-1},\tilde{\tau}_1) \subset (\tau_{-1},\tau_1)$ en $v \in C((\tilde{\tau}_{-1},\tilde{\tau}_1),\mathbf{R}^d)$ differentieerbaar zodat

$$v'(t) = f(t,v(t)) \quad \text{voor alle} \quad t \in (\tilde{\tau}_{-1},\tilde{\tau}_1) \quad \text{en} \quad v(\tilde{t}_0) = u(\tilde{t}_0),$$

dan $v = u$ op $(\tilde{\tau}_{-1},\tilde{\tau}_1)$). □

**1.1.5 Opgave.** a. Beschouw 1.1.2.
Ga na dat $f$ niet uniform Lipschitz kontinu is in de tweede variabele op $\mathbf{R}\times\mathbf{R}$. Ga na dat $f$ dat wel is, voor $\alpha,\beta \in \mathbf{R}$, $\alpha < \beta$, op $\Omega = \mathbf{R} \times (\alpha,\beta)$. Ga na hoe de stelling van toepassing is.
b. Voor geen enkel open deel $\Omega$ van $\mathbf{R}\times\mathbf{R}$ met $(0,0) \in \Omega$ is de $f$ in 1.1.3 uniform Lipschitz kontinu in de tweede variabele op $\Omega$. Ga dat na.

**1.1.6 Opmerking.** De stelling geeft voldoende voorwaarden voor existentie en uniciteit van de oplossing op $\mathcal{J}$, maar geen noodzakelijke! In 1.1.2 heeft het probleem $u' = u^2, u(0) = \alpha$ precies een oplossing op $[0,\infty)$ als $\alpha < 0$ nl.: $u(t) = \alpha/(1-\alpha t)$.

We zullen verder betreffende probleem (1) het volgende aannemen.

**1.1.7 Aanname.**
We nemen aan dat $\Omega$, $f$ en $u_0$ in 1.1.1 zo zijn dat (1) precies één differentieerbare oplossing $u$ in $C(\mathcal{J},\mathbf{R}^d)$ heeft. We noteren deze unieke oplossing met $u^*$. Verder nemen we aan dat
(Dif.0) Voor een zekere $\tilde{r} > 0$ is de $\tilde{r}$-buis $\{(t,x) \,\big|\, \|x - u^*(t)\| \leq \tilde{r}, t \in \mathcal{J}\}$ rond de grafiek van $u^*$ binnen $\Omega$.

**1.1.8 Opmerkingen.** (a) Stelling 1.1.4 geeft ons voorwaarden waaronder de aanname korrekt is; eventueel moeten we, om de stelling van toepassing te laten zijn, $\mathcal{J}$ verkorten.
(b) In ieder numeriek proces zullen we fouten maken. We kunnen het probleem zeker niet oplossen als zo'n fout ons uit het definitie gebied van $f$ zet. De aanname (Dif.0) geeft ons wat ruimte.
(c) De notatie $u^*$ van wat we de *exakte oplossing* van (1) zullen noemen (ter onderscheid van numerieke oplossingen) kan in een zekere kontext verwarrend zijn: $*$ word ook gebruikt om bijvoorbeeld adjungatie van matrices aan te geven, terwijl $\star$ aangeeft dat we met een berekende grootheid te maken hebben (dus belast is met rekenfouten). We kiezen toch voor de $*$ notatie omdat deze de gebruikelijke is in de numerieke literatuur over het oplossen van gewone differentiaalvergelijkingen.

Bij iedere numerieke methode maken we fouten. In de volgende paragraaf zullen we zien dat, in de analyse van die fouten, vanzelf lineaire differentiaalvergelijkingen opduiken. Het lineaire probleem dat hieronder geformuleerd is zullen we dan ook in de fouten analyse nogal eens tegenkomen.

**1.1.9 Lineaire problemen.**
Beschouw, voor $J \in C(\mathcal{J}, \mathbf{M}_d(\mathbf{R}))$ en $g \in C(\mathcal{J}, \mathbf{R}^d)$ het probleem

$$u'(t) = J(t)u(t) + g(t) \quad (t \in \mathcal{J}) \quad \text{en} \quad u(t_0) = u_0. \tag{2}$$

We schrijven vaak $Ju(t)$ in plaats van $J(t)u(t)$.
Hier is $f(t, x) = J(t)x + g(t) \quad (t \in \mathcal{J}, x \in \mathbf{R}^d)$. $L = \sup_t \|J(t)\|$.

**1.1.10 Greense funkties en variatie van konstanten.** Laat, voor iedere $\tau \in \mathcal{J}$, $t \to G(t, \tau)$ de kontinu differentieerbare funktie van $[\tau, \infty) \cap \mathcal{J}$ naar $\mathbf{M}_d$ zijn waarvoor

$$\frac{\partial}{\partial t} G(t, \tau) = J(t)G(t, \tau) \quad \text{voor} \quad t \geq \tau \quad \text{en} \quad G(\tau, \tau) = I.$$

$G$ is de *Greense funktie* van dit lineaire beginwaarde probleem.
Op $[t_0, t)$ is $\tau \to G(t, \tau)$ kontinu en

$$u^*(t) = G(t, t_0)u_0 + \int_{t_0}^{t} G(t, \tau)g(\tau)\, d\tau \quad \text{voor alle} \quad t \in \mathcal{J}. \tag{3}$$

Bij deze laatste formule spreekt men ook wel over *variatie van konstanten* (bedenk dat $G(t, \tau) = G(t, t_0)G(\tau, t_0)^{-1}$ en dus $u^*(t) = G(t, t_0)[u_0 + \int G(\tau, t_0)^{-1} g(\tau)\, d\tau]$. De uitdrukking tussen [] kan men zien als variatie van de konstante vektor $u_0$: $G(t, t_0)u_0$ is de oplossing van het homogene beginwaarde probleem $u' = Ju$, $u(t_0) = u_0$.).

In de praktijk zal men lineaire problemen met konstante koëfficiënten niet numeriek willen oplossen: de exakte oplossing kan gemakkelijk analytisch gegeven worden. In de numerieke theorie spelen dit soort problemen wel een belangrijke rol. Ondermeer omdat ze zo gemakkelijk te analyseren zijn. Ze kunnen daardoor inzicht verschaffen in eigenschappen van numerieke oplosmethode. We zullen een paar keer naar de twee onderstaande problemen kijken.

**1.1.11 Een lineaire probleem: $d = 1$.**
Beschouw, voor $d = 1$, $\eta \in \mathbf{R}$ en $g \in C(\mathcal{J})$ het probleem

$$u'(t) = \eta u(t) + g(t) \quad (t \in \mathcal{J}) \quad \text{en} \quad u(t_0) = u_0. \tag{4}$$

Hier is $G(t, \tau) = e^{\eta(t-\tau)}$ voor $t \geq \tau$ en $L = |\eta|$.

**1.1.12 Een lineair probleem: $d > 1$.**
Beschouw, voor $d > 1$, $J_0 \in \mathbf{M}_d(\mathbf{R})$ en $g \in C(\mathcal{J}, \mathbf{R}^d)$ het probleem

$$u'(t) = J_0 u(t) + g(t) \quad (t \in \mathcal{J}) \quad \text{en} \quad u(t_0) = u_0. \tag{5}$$

Stel dat $J_0$ diagonalizeerbaar is: er is een niet-singuliere matrix $V$ zodat
$V^{-1} J_0 V = D := \mathrm{diag}\,(\eta_1, \ldots, \eta_d)$ met $\eta_j \in \mathbf{C}$ de eigenwaarden van $J_0$.
Dan $G(t, \tau) = V \exp(D(t - \tau))\, V^{-1} \quad (t \geq \tau)$.
Als $\|\cdot\| = \|\cdot\|_p$ voor zekere $p \in [1, \infty]$ en $\mathcal{C}_p(V) := \|V\|_p \, \|V^{-1}\|_p$ is het *conditie getal* van $V$ dan geldt

$$\mathcal{C}_p(V)^{-1} \max_{j \leq d} |\eta_j| \leq L \leq \mathcal{C}_p(V) \max_{j \leq d} |\eta_j|.$$

**Voorbeelden**

**1.1.13 Semi-gediskretiseerde partiële differentiaalvergelijkingen.**
Beschouw, voor $\phi_0, \phi_1 \in C([0,\infty))$, $\psi \in C([0,1])$, de funktie $u \in C^{1,2}([0,\infty)\times[0,1])$ die de oplossing is van het volgende partiële differentiaalvergelijkingsprobleem.

$$\begin{cases} u_t(t,x) := \frac{\partial}{\partial t}u(t,x) = u_{xx}(t,x) := \frac{\partial^2}{\partial x^2}u(t,x) & \text{voor} \quad t\in[0,\infty), x\in[0,1] \\ u(t,0)=\phi_0(t),\ u(t,1)=\phi_1(t) & \text{voor} \quad t\in[0,\infty) \\ u(0,x)=\psi(x) & \text{voor} \quad x\in[0,1] \end{cases}$$

Zij $d\in\mathbf{N}$. Schrijf $h := \frac{1}{d+1}$.
Met $U(t) := (u(t,h), u(t,2h), \ldots, u(t,dh))^T \in \mathbf{R}^d$ is $U\in C^1([0,\infty),\mathbf{R}^d)$ en geldt (ga dit na)

$$\begin{cases} U'(t) = \frac{1}{h^2}Q_hU(t) + \frac{1}{h^2}F(t) - \delta(t) \quad (t\in[0,\infty)) \\ U(0) = (\psi(h),\ldots,\psi(dh))^T, \end{cases} \tag{6}$$

met $Q_h \in \mathbf{M}_d$, $F(t), \delta(t) \in \mathbf{R}^d$ zodat

$$Q_h := \begin{bmatrix} -2 & 1 & & & 0 \\ 1 & -2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & -2 & 1 \\ 0 & & & 1 & -2 \end{bmatrix}, \quad F(t) := \begin{bmatrix} \phi_0(t) \\ 0 \\ \vdots \\ 0 \\ \phi_1(t) \end{bmatrix}, \quad \delta(t) := \frac{h^2}{12}\begin{bmatrix} u_{xxxx}(t,\xi_1) \\ u_{xxxx}(t,\xi_2) \\ \vdots \\ u_{xxxx}(t,\xi_{d-1}) \\ u_{xxxx}(t,\xi_d) \end{bmatrix}.$$

Diskretiseren we de “ruimte richting” dan houden we een gewoon beginwaarde probleem over; deze diskretisatie techniek noemt men ook wel *semi–diskretisatie*. Merk op dat de dimensie $d$ groter wordt naarmate $h\to 0$.

**1.1.14 Een 2–de orde probleem.** Beschouw, voor $d=1$, $\eta\in\mathbf{R}$ en $\varepsilon>0$ het probleem

$$\varepsilon u''(t) + u'(t) = \eta u(t) + g(t) \quad \text{voor} \quad t\geq t_0 \quad \text{en} \quad u(t_0)=u_0, \quad u'(t_0)=0.$$

$u\in C^2(\mathcal{J})$ is de oplossing van dit probleem precies dan als $v=(u,u')^T \in C^1(\mathcal{J},\mathbf{R}^2)$ de oplossing is van het volgende probleem

$$\begin{cases} v'(t) = \begin{bmatrix} 0 & 1 \\ \frac{\eta}{\varepsilon} & -\frac{1}{\varepsilon} \end{bmatrix} v(t) + \begin{bmatrix} 0 \\ \frac{1}{\varepsilon}g(t) \end{bmatrix} \quad (t\in\mathcal{J}) \\ v(t_0) = (u_0, 0)^T. \end{cases}$$

**1.1.15 Hogere orde problemen.** In het laatste voorbeeld zagen we dat we een beginwaarde problemen met hogere orde afgeleiden in de differentiaalvergelijking ook kunnen schrijven als een probleem met alleen eerste orde afgeleiden. Dit kan voor ieder hoger orde beginwaarde probleem:

Voor $k\in\mathbf{N}$ en voor $u_0^{(0)},\ldots,u_0^{(k)}\in\mathbf{R}^d$ voldoet $u\in C^1(\mathcal{J},\mathbf{R}^d)$ aan

$$\begin{cases} u^{(k+1)}(t) = f(t,u(t),u^{(1)}(t),\ldots,u^{(k)}(t)) \quad \text{voor iedere} \quad t\in\mathcal{J}, \\ u(t_0)=u_0^{(0)},\ u^{(1)}(t_0)=u_1^{(1)},\ \ldots,\ u^{(k)}(t_0)=u_0^{(k)} \end{cases}$$

precies dan als $v(t) := (u(t)^T, u^{(1)}(t)^T, \ldots, u^{(k)}(t)^T)^T$ voldoet aan

$$\begin{cases} v'(t) = F(t,v(t)) \quad (t\in\mathcal{J}), \text{ met } F(t, \begin{bmatrix} x_0 \\ \vdots \\ x_{k-1} \\ x_k \end{bmatrix}) = \begin{bmatrix} x_1 \\ \vdots \\ x_k \\ f(t,x_0,\ldots,x_k) \end{bmatrix} \quad (x_j\in\mathbf{R}^d, t\in\mathcal{J}) \\ v(t_0) = (u_0^{(0)^T}, u_0^{(1)^T}, \ldots, u_0^{(k)^T})^T. \end{cases}$$

**1.1.16 Autonome problemen.** We zeggen dat probleem (1) *autonoom* is als $f$ niet expliciet van $t$ afhangt: $f(t,x) = f(s,x) \quad ((t,x),(s,x) \in \Omega)$.

Probleem (1) is equivalent met een autonoom probleem: $u \in C^1(\mathcal{J}, \mathbf{R}^d)$ voldoet aan (1) precies dan als $v(t) := (t, u(t)^T)^T \in C^1(\mathcal{J}, \mathbf{R}^{d+1})$ voldoet aan

$$\left\{ \begin{array}{l} v'(t) = F(v(t)) \qquad (t \in \mathcal{J}) \quad \text{met} \quad F(\left[ \begin{array}{c} s \\ x \end{array} \right]) := \left[ \begin{array}{c} 1 \\ f(s,x) \end{array} \right] \qquad ((s,x) \in \Omega) \\ v(t_0) = (t_0, u_0^T)^T \end{array} \right. .$$

## 1.2 De konditionering van het probleem

Als het probleem zelf iedere fout onbeperkt laat groeien (zie voorbeeld 1.2.1) dan maken we met geen enkele numerieke methode kans de exakte oplossing uniform op bijvoorbeeld $[t_0, \infty)$ te kunnen benaderen.

**1.2.1 Voorbeeld.** Beschouw, voor $d = 1$, het probleem

$$u'(t) = u(t) + (\cos t - \sin t) \quad \text{voor} \quad t \in [0, \infty) \quad \text{en} \quad u(0) = 0,$$

met exakte oplossing $u^* = \sin$. Als voor $\tau > 0$ de funktie $\tilde{u}$ ook aan de differentiaalvergelijking voldoet en, voor $\varepsilon \in \mathbf{R}$ is $\tilde{u}(\tau) = u^*(\tau) + \varepsilon$, dan

$$\tilde{u}(t) = u^*(t) + \varepsilon e^{t-\tau} \quad \text{voor iedere} \quad t \geq \tau:$$

de lokale perturbatie $\varepsilon$ gemaakt op tijdstip $\tau$ resulteert in de globale fout $\varepsilon e^{t-\tau}$ voor $t \geq \tau$ die onbegrensd is als $t \to \infty$.

Tijdens het numerieke proces maken we in iedere rekenstap fouten. We mogen alleen hopen een geschikte numerieke oplosmethode te kunnen vinden als het probleem zelf niet al te "gevoelig" is voor perturbaties. Het probleem moet "goed gekonditioneerd" zijn: de oplossing van de geperturbeerde differentiaalvergelijking met geperturbeerde beginwaarde moet lijken op de exakte oplossing als de perturbaties voldoende klein zijn. We geven de formele definitie.

**1.2.2 De konditionering van het beginwaarde probleem.**
Beschouw een $\delta \in C(\mathcal{J}, \mathbf{R}^d)$ en een $\delta_0 \in \mathbf{R}^d$.
We zijn geïnteresseerd in de funktie $u \in C^1(\mathcal{J}, \mathbf{R}^d)$ die voldoet aan

$$\left\{ \begin{array}{l} u'(t) = f(t, u(t)) + \delta(t) \qquad \text{voor alle} \quad t \in \mathcal{J} \\ u(t_0) = u_0 + \delta_0. \end{array} \right. \tag{7}$$

$\delta$ is een perturbatie op de oorspronkelijke differentiaalvergelijking en $\delta_0$ op de oorspronkelijke beginwaarde.

**A.** Het probleem (1) heeft een *eindig konditie getal* als er een $C > 0$ is zodat het geperturbeerd probleem (7) een oplossing $\tilde{u}$ heeft en er geldt

$$\sup_{t \in \mathcal{J}} \|u^*(t) - \tilde{u}(t)\| \leq C \left( \|\delta_0\| + \int_{\mathcal{J}} \|\delta(\tau)\| \, d\tau \right) \tag{8}$$

zodra $\|\delta_0\| + \int \|\delta(\tau)\| \, d\tau$ voldoende klein is.
Het infimum $\mathcal{C}(u^*)$ van de $C$'s waarvoor (8) korrekt is voor voldoend kleine perturbaties is het *konditie getal van het probleem* (1).

Het konditie getal vertelt hoe de kumulatie van kleine lokale absolute perturbaties maximaal doorwerkt. Als het konditie getal niet te groot is (denk aan $\leq 10^4$) noemen we het probleem (1) *goed gekonditioneerd.*

Met name in geval $\mathcal{J}$ onbegrensd is ($\mathcal{J} = [t_0, \infty)$) is de kumulatie $\int_{\mathcal{J}} \|\delta(\tau)\| d\tau$ van de lokale absolute perturbaties nogal eens onbegrensd terwijl de resulterende perturbatie $u^* - \tilde{u}$ op de oplossing van het beginwaarde probleem toch niet al te groot is. In zo'n geval is het konditie

getal niet zo informatief. We werken dan liever met de volgende grootheid.
**B.** Probleem (1) heeft een *eindig sterk konditie getal* als er een $C > 0$ is zodat het geperturbeerd probleem (7) een oplossing $\tilde{u}$ heeft en er geldt

$$\sup_{t\in\mathcal{J}} \|u^*(t) - \tilde{u}(t)\| \leq C \left( \|\delta_0\| + \sup_{\tau\in\mathcal{J}} \|\delta(\tau)\| \right) \tag{9}$$

zodra $\|\delta_0\| + \sup_\tau \|\delta(\tau)\|$ voldoende klein is.
Het infimum $\tilde{\mathcal{C}}(u^*)$ van de $C$'s waarvoor (8) korrekt is voor voldoend kleine perturbaties is het *sterke konditie getal van het probleem* (1).

Het sterke konditie getal vertelt hoe de maximale absolute waarde van kleine lokale pertubaties maximaal doorwerkt.

### 1.2.3 De afhankelijkheid van de norm.

De keuze van de norm op $\mathbf{R}^d$ speelt i.h.a. geen essentiële rol[1]. Stappen we over op een andere norm dan verandert alleen de konstante $\mathcal{C}(u^*)$. Als bijvoorbeeld $\mathcal{C}_2(u^*)$ het konditie getal is ten opzichte van de 2-norm en $\mathcal{C}_\infty(u^*)$ ten opzichte van de sup-norm dan is $\mathcal{C}_\infty(u^*) \leq \sqrt{d}\mathcal{C}_2(u^*) \leq d\mathcal{C}_\infty(u^*)$. (Bewijs dit.)

### 1.2.4 Minder gladde perturbaties.

In de praktijk is de perturbatie $\delta$ natuurlijk niet altijd een kontinue afbeelding. Onze definities en uitspraken zijn echter ook toepasbaar voor mindere gladde perturbaties. Definiëren we bijvoorbeeld het konditie getal voor links kontinue perturbaties $\delta$ (d.w.z. $\lim_{h\to 0, h>0} \delta(t-h) = \delta(t)$) en kontinue links differentiëerbare oplossingen $\tilde{u}$ (nu met $u'(t) := \lim_{h\to 0,h>0} \frac{1}{h}(u(t) - u(t-h))$ voor $t \in \mathcal{J}, t > t_0$) dan zijn onze uitspraken ook korrekt.

### 1.2.5 Modifikaties: relatieve fouten, gewogen fouten, etc..

De definitie van het konditie getal geeft een exakte mathematische inhoud aan de begrippen "goed gekonditioneerd", "kleine pertubatie" en "lijken op". Hiermee is echter niet gezegd dat de definitie de wens verwoordt van degene die met de definitie wil werken.

Als de exakte oplossing $u^*$ bijvoorbeeld sterk groeiend is dan zal het niet bezwaarlijk zijn als de absolute fout $\|\tilde{u}(t) - u^*(t)\|$ ook groeit. Of als $\mathcal{J} = [t_0, \infty)$ en de oplossing $|u^*(t)| \to 0$ voor $t \to \infty$ dan zal men gewoonlijk liever zien dat de absolute fout ook afneemt. Gewoonlijk is men meer geïnteresseerd in een kleine relatieve absolute fout dan in een kleine absolute fout. Slingert de exakte oplossing sterk dan mag men niet verwachten dat $u^*(t)$ voor $t$ in de buurt van een nulpunt $\tau$ in goede relatieve precisie te berekenen is: men zal wel graag zien dat de absolute fout klein is ten opzichte van funktiewaarden $u^*(t)$ voor $t$ in de buurt van $\tau$. Wil men bijvoorbeeld de grafiek van $u^*$ plotten dan zal men graag zien dat de absolute fout in dat gedeelte van de grafiek dat geplot wordt kleiner is dan het oplossend vermogen van het scherm of de plotter. Kortom, men wil graag zien dat de *gewogen fout* $\sup_t \|\omega(t)[\tilde{u}(t) - u^*(t)]\|_2$ klein is, waarbij $\omega$ een geschikte *gewichtsfunktie* is. Voor $\gamma > 0$ zou men, met $\mathcal{U}(t) := \{u^*(s) \,\big|\, s \in \mathcal{J}, |t-s| \leq \gamma\}$, kunnen denken aan $\omega$ zodat $\frac{1}{\omega(t)} = \sup\{|x| \,\big|\, x \in \mathcal{U}(t)\}$ of $\frac{1}{\omega(t)} = \sup \mathcal{U}(t) - \inf \mathcal{U}(t)$.
Men zal de perturbaties meten met de norm die het beste past bij het type perturbaties dat men verwacht. Als $u^*$ sterk groeit zal $\delta$ ook wel groeien: de lokale fouten zullen wellicht in een of andere relatieve zin klein zijn.
(De exakte oplossing $t \to u^*(t) = e^{2t}$ van het probleem $u'(t) = u(t) + e^{2t}$ voor $t \in [0, 100]$, $u(0) = 1$ is slecht gekonditioneerd volgens bovenstaande definitie: iedere lokale fout $\varepsilon$ in $\tau \geq 0$ groeit volgens $\varepsilon e^{t-\tau}$. Men kan stellen dat $u^*$ wel goed gekonditioneerd is ten opzichte van de gewichtsfunktie $\omega(t) = e^{-2t}$ op de perturbatie en op de fout: met $v^*(t) := e^{-2t}u^*(t)$ is $v^*$ immers de exakte oplossing van $v'(t) = -v(t) + 1$, $v(0) = 1$. Zie (a) in 1.2.12.)

[1] Tenzij $d$ groot is, hetgeen met name het geval is in situaties als in voorbeeld 1.1.13, waarin de gewone differentiaalvergelijking afstamt van een semigediskretiseerde partiële differentiaalvergelijking.

Onze definities en resultaten in dit hoofdstuk sporen met de wensen die we zojuist geformuleerd hebben in geval $\omega$ op $\mathcal{J}$ min of meer konstant is. Wil men resultaten voor een gewichtsfunktie $\omega$ die niet min of meer konstant is dan zal men onze verdere resultaten opnieuw moeten bezien.

Lokale perturbaties worden nog al eens "exponentiëel" voortgeplant. De volgende uitspraak levert dan een uitspraak over de konditionering van het probleem.

**1.2.6 Bewering.** *Stel er is een* $\mu \in \mathbf{R}$ *en een* $C > 0$ *zodat het geperturbeerd probleem* (7) *een oplossing* $\tilde{u}$ *heeft en er geldt*

$$\varepsilon(t) := \|\tilde{u}(t) - u^*(t)\| \leq Ce^{\mu(t-t_0)}\|\delta_0\| + C\int_{t_0}^{t} e^{\mu(t-\tau)}\|\delta(\tau)\|\,d\tau \quad \textit{voor} \quad t \in \mathcal{J},$$

*als de majorant in het rechterlid voor iedere* $t \in \mathcal{J}$ *kleiner is dan* $\tilde{r}$*. Dan is*

$$\varepsilon(t) \leq C(1+e^{\mu T})(\|\delta_0\| + \int_{\mathcal{J}} \|\delta(\tau)\|\,d\tau) \leq C(1+e^{\mu T})(1+T)(\|\delta_0\| + \sup_{\tau\in\mathcal{J}} \|\delta(\tau)\|)$$

*en heeft probleem* (1) *een eindig konditie getal:*

$$\mathcal{C}(u^*) \leq C(1+e^{\mu T}).$$

*Als* $\mu < 0$ *dan is zelfs*

$$\varepsilon(t) \leq C(1+\frac{1}{|\mu|})(\|\delta_0\| + \sup_{\tau} \|\delta(\tau)\|)$$

*en heeft probleem* (1) *zelfs een eindig sterk konditie getal.*

$$\begin{array}{lll} \mu \geq 0 \quad \& \quad T < \infty & \Rightarrow & \tilde{\mathcal{C}}(u^*) \leq C(1+e^{\mu T})(1+T) \\ \mu < 0 & \Rightarrow & \tilde{\mathcal{C}}(u^*) \leq C(1+\frac{1}{|\mu|}). \end{array} \qquad \square$$

Als $\mathcal{J}$ begrensd is en (Dif.1) geldt dan heeft probleem (1) een eindig konditie getal. In het bewijs van dit resultaat gebruiken we het volgende lemma.

**1.2.7 Lemma.**
*Zij* $\bar{\delta}_0 \in [0,\infty)$, $\eta \in \mathbf{R}$ *en* $\bar{\delta} \in C(\mathcal{J},[0,\infty))$. *Als* $\epsilon \in C^1(\mathcal{J},[0,\infty))$ *zodat*

$$\epsilon'(t) \leq 2\eta\epsilon(t) + 2\bar{\delta}(t)\sqrt{\epsilon(t)} \quad (t \in \mathcal{J}), \quad \textit{en} \quad \epsilon(t_0) = \bar{\delta}_0$$

*dan*

$$\sqrt{\epsilon(t)} \leq \bar{\delta}_0 e^{\eta(t-t_0)} + \int_{t_0}^{t} e^{\eta(t-\tau)}\bar{\delta}(\tau)\,d\tau.$$

∘ *Bewijs.* Beschouw een $\tilde{\delta} \in C(\mathcal{J})$ zodat $\tilde{\delta}(t) > \bar{\delta}(t)$ voor alle $t \in \mathcal{J}$ en $\phi \in C^1(\mathcal{J})$ zodat $\phi'(t) = \eta\phi(t) + \tilde{\delta}(t)$ $(t \in \mathcal{J})$ en $\phi(t_0) > \bar{\delta}_0$.
Stel er is een $t_1 > t_0$ zodat $\epsilon < \phi^2$ op $[t_0, t_1)$ en $\epsilon(t_1) = \phi(t_1)^2$. Dan is $\epsilon'(t_1) \geq (\phi^2)'(t_1)$.
Dus

$$\begin{aligned} 2\eta\epsilon(t_1) + 2\bar{\delta}(t_1)\sqrt{\epsilon(t_1)} &\geq \epsilon'(t_1) \geq 2\phi(t_1)\phi'(t_1) = \\ 2\eta\phi^2(t_1) + 2\tilde{\delta}(t_1)\phi(t_1) &= 2\eta\epsilon(t_1) + 2\tilde{\delta}(t_1)\sqrt{\epsilon(t_1)}, \end{aligned}$$

hetgeen, gezien onze keuze van $\tilde{\delta}$, niet kan. Blijkbaar $\epsilon(t) < \phi(t)^2$ voor alle $t \in \mathcal{J}$.
Door $\phi$ op te lossen (zie 1.1.11 en 1.1.9) en het infimum te nemen over alle $\tilde{\delta} > \bar{\delta}$, $\phi(t_0) > \bar{\delta}_0$ volgt het lemma. $\square$

**1.2.8 Stelling.** *Stel* (Dif.0) *en* (Dif.1) *gelden.*
*Dan is* 1.2.6 *van toepasing met de* 2*-norm,* $\mu = L$ *en* $C = 1$. *Dus*

$$\mathcal{C}(u^*) \leq 1 + e^{LT}.$$

*Bewijs.*
Met $e(t) := \tilde{u}(t) - u^*(t)$ en $\epsilon(t) := \|e(t)\|_2^2$ voor $t \in \mathcal{J}$ is

$$\epsilon'(t) = 2 < e'(t), e(t) > \leq 2\|e'(t)\|_2 \sqrt{\epsilon(t)} \leq 2L\epsilon(t) + 2\|\delta(t)\|_2 \sqrt{\epsilon(t)}.$$

Het resultaat volgt nu eenvoudig uit bovenstaand lemma. □

Voor lineaire problemen laat de fout zich beschrijven in termen van de Greense funktie en de lokale perturbaties. De volgende bewering volgt onmiddellijk uit de beschrijving van de oplossing voor dit soort problemen in 1.1.9. De tweede volgt uit de eerste middels een voor de handliggende norm schatting.

**1.2.9 Bewering.** *Beschouw het probleem in* 1.1.9.
*Het geperturbeerde probleem* (7) *heeft een oplossing* $\tilde{u}$ *en er geldt*

$$\tilde{u}(t) - u^*(t) = G(t, t_0)\delta_0 + \int_{t_0}^{t} G(t, \tau)\delta(\tau)\, d\tau \qquad \textit{voor iedere} \quad t \in \mathcal{J}. \quad \square \tag{10}$$

**1.2.10 Bewering.** *Beschouw het probleem in* 1.1.9. *Stel er zijn* $\mu \in \mathbf{R}$ *en* $C > 0$ *zodat* $\|G(t, \tau)\| \leq Ce^{\mu(t-\tau)}$ *voor alle* $t, \tau \in \mathcal{J}$, $t \geq \tau$. *Dan is* 1.2.6 *van toepassing.* □

**1.2.11 Interpretatie.** De globale fout $\tilde{u} - u^*$ kan volgens (10), in het tijdstip $t$, gezien worden als een superpositie van de effekten $G(t, \tau)\delta(\tau)$ van de lokale pertubaties $\delta(\tau)$ en het effekt van de startfout. $G(t, \tau)$ is de faktor waarmee de lokale perturbatie $\delta(\tau)$ in het tijdstip $t$ gegroeid is. Als het effekt van de lokale fouten op den duur exponentiëel uitsterft—als, in de laatste bewering, $\mu < 0$; $C$ laat een aanvankelijke groei toe—dan is het lineaire probleem goed gekonditioneerd ook in geval $\mathcal{J}$ het half-oneindige interval $[t_0, \infty)$ is. De globale fout wordt dan niet alleen gemajoreerd door een veelvoud van de kumulatie van de lokale absolute fouten maar zelfs door een veelvoud van de maximale lokale absolute fout[2].

We formuleren hieronder bovenstaand resultaat voor de twee speciale gevallen in 1.1.11 en 1.1.12. Voor deze gevallen zijn de schattingen min of meer scherp. We laten het bewijs over aan de lezer.

**1.2.12 Gevolg.**
(a) *Als* $\mathcal{C}(u^*)$ *het konditie getal is van het probleem in* 1.1.11 *dan geldt*

$$e^{\eta T} \leq \mathcal{C}(u^*) \leq 1 + e^{\eta T}.$$

(b) *Als* $\mathcal{C}_2(u^*)$ *het konditie getal is van het probleem in* 1.1.12 *t.o.v. de* 2*-norm dan geldt met* $\eta := \max_{j \leq d} Re(\eta_j)$ *dat*

$$\mathcal{C}_2(V)^{-1} e^{\eta T} \leq \mathcal{C}_2(u^*) \leq \mathcal{C}_2(V)(1 + e^{\eta T}).$$

(c) *Beide problemen hebben, in geval* $\mathcal{J} = [t_0, \infty)$, *een eindig konditie getal precies dan als* $\eta < 0$. *Als* $\eta < 0$ *dan hebben beide problemen zelfs een eindig sterk konditie getal en geldt*

$$\tilde{\mathcal{C}}_2(u^*) \leq (1 + \frac{1}{|\eta|}). \qquad \square$$

[2] In de analytische literatuur noemt men probleem (2) *uniform stabiel* als $\mu = 0$ en *exponentiël asymptotisch stabiel* als $\mu < 0$.

**1.2.13 Opgave.** Bewijs 1.2.12.

De problemen in 1.1.11 en 1.1.12 mogen we goed gekondioneerd noemen als $\eta T < 10$ $(e^{10} \approx 2.2\,10^4)$ en slecht als $\eta T > 40$ $(e^{40} \approx 2.3\,10^{17})$ (zie 1.2.13).

Als in 1.1.11 bijvoorbeeld $\eta = -10^5$ dan is het probleem uitstekend gekonditioneerd terwijl $e^{LT}$ voor $T \geq 1$ waanzinnig groot is. De schatting voor het konditie getal in stelling 1.2.8 is dus op zijn zachtst gezegd niet altijd even informatief.

**1.2.14 Opgave.** Ga, voor de problemen in 1.1.11 en 1.1.12, na in welke gevallen de schatting $1 + e^{LT}$ in 1.2.8 voor het konditie getal niet te grof is (zie 1.2.12).

### Het gelineairiseerde probleem

Als $f \in C^2(\Omega, \mathbf{R}^d)$ en de fout in de oplossing $\tilde{u}$ van het geperturbeerde probleem (7) is klein dan voldoet die fout min of meer aan een lineaire differentiaalvergelijking (zie (a) van 1.2.17). Iedere kleine lokale perturbatie wordt door de differentiaalvergelijking min of voortgeplant volgens een oplossing van een homogene lineaire vergelijking (zie (b) van 1.2.17). Deze lineaire vergelijkingen ontstaan door in de oorspronkelijke $f$ als volgt te lineariseren.

**1.2.15 Notatie.** Stel $f \in C^2(\Omega, \mathbf{R}^d)$.
Voor iedere $t \in \mathcal{J}$ bestaat dan de totale afgeleide $D_x(f)(t, u^*(t))$ van $f$ naar de tweede variabele ( $D_x(f)(t, x)$ is in $\mathbf{M}_d$ met $(i, j)$-de matrix koëfficiënt $\frac{\partial}{\partial x_j} f_i(t, x)$ ). We noteren deze funktionaalmatrix in het vervolg met $J(t)$.

Merk op dat $t \to J(t)$ in $C^1(\mathcal{J}, \mathbf{M}_d(\mathbf{R}))$.
Verder is er een $a \in C(\mathcal{J}, [0, \infty))$ zodat voor iedere $t \in \mathcal{J}$, $e \in \mathbf{R}^d$ met $\|e\|$ voldoende klein geldt

$$\|f(t, u^*(t) + e) - f(t, u^*(t)) - J(t)e\| \leq a(t)\|e\|^2. \tag{11}$$

De funktionaalmatrix $J(t)$ en de relatie (11) speelt een belangrijke bij de analyse van fouten in zowel het kontinue probleem als in de numerieke oplosmethode.

**1.2.16 Stelling.** *Stel* $\bar{a} := \int_{\mathcal{J}} a(\tau)\, d\tau < \infty$. *Zij* $g \in C(\mathcal{J}, \mathbf{R}^d)$.
*Dan is het konditie getal* $\mathcal{C}(u^*)$ *van probleem* (1) *precies gelijk aan dat van het probleem*

$$u'(t) = Ju(t) + g(t) \quad \textit{voor alle} \quad t \in \mathcal{J} \quad \textit{en} \quad u(t_0) = 0.$$

*Als* $G$ *de Greense funktie is van dit lineaire probleem dan* $\|G(t, \tau)\| \leq \mathcal{C}(u^*)$ $(t \geq \tau)$.

*Bewijs.* Schrijf $A(t, x) := f(t, u^*(t) + x) - f(t, u^*(t)) - J(t)x$ voor $t \in \mathcal{J}$ en $x \in \mathbf{R}^d$ met $\|x\|$ voldoende klein.
∘ Beschouw een $v \in C(\mathcal{J}, \mathbf{R}^d)$ waarvoor geldt $v'(t) = Jv(t) + \delta(t)$ en $v(t_0) = \delta_0$.
Met $\tilde{u} := u^* + v$ is $\tilde{u}'(t) = f(t, \tilde{u}(t)) + \delta(t) - A(t, v(t))$.
Ga dit na. De overige details laten we ook over aan de geïnteresseerde lezer. □

**1.2.17 Stelling.** *Stel* $\bar{a} := \int_{\mathcal{J}} a(\tau)\, d\tau < \infty$. *Zij* $C > \mathcal{C}(u^*)$.
(a) *Als* $\bar{\delta} := \|\delta_0\| + \int_{\mathcal{J}} \|\delta(\tau)\|\, d\tau$ *voldoende klein is dan heeft het geperturbeerde probleem* (7) *een oplossing* $\tilde{u}$ *en er geldt*

$$\|\tilde{u}(t) - u^*(t)\| \leq C\bar{\delta} \quad \textit{en} \quad \|\tilde{u}(t) - u^*(t) - w(t)\| \leq \bar{a}C(C\bar{\delta})^2 \quad (t \in \mathcal{J}),$$

*waarbij* $w \in C^1(\mathcal{J}, \mathbf{R}^d)$ *de oplossing is van het lineaire probleem*

$$w'(t) = Jw(t) + \delta(t) \quad (t \in \mathcal{J}) \quad \textit{en} \quad w(t_0) = \delta_0. \tag{12}$$

(b) *Als* $\tau \geq t_0$, $\delta_\tau \in \mathbf{R}^d$ *met* $\bar{\delta} := \|\delta_\tau\|$ *voldoende klein, dan heeft het probleem*

$$u'(t) = f(t, u(t)) \quad \textit{voor alle} \quad t \in \mathcal{J}, t \geq \tau \quad \textit{en} \quad u(\tau) = u^*(\tau) + \delta_\tau,$$

*een oplossing* $\tilde{u}$ *en er geldt*

$$\|\tilde{u}(t) - u^*(t)\| \leq C\bar{\delta} \quad \textit{en} \quad \|\tilde{u}(t) - u^*(t) - w(t)\| \leq \bar{a}C(C\bar{\delta})^2 \quad (t \in \mathcal{J}, t \geq \tau),$$

*waarbij* $w = G(\cdot, \tau)\delta_\tau \in C^1(\mathcal{J}, \mathbf{R}^d)$ *de oplossing is van het homogene probleem*

$$w'(t) = Jw(t) \quad (t \in \mathcal{J}, t \geq \tau) \quad \textit{en} \quad w(\tau) = \delta_\tau.$$

*Bewijs.* (a) De eerste schatting volgt uit de definitie van het konditie getal.
Zij $A(t,x)$ als in het bewijs van 1.2.16. Met $e := \tilde{u} - u^*$ voldoet $e$ dan aan

$$e'(t) = Je(t) + \delta(t) + A(t, e(t)) \quad (t \in \mathcal{J}) \quad \text{en} \quad e(t_0) = \delta_0.$$

Met $\tilde{e} := e - w$ geldt $\tilde{e}'(t) = J\tilde{e}(t) + A(t, e(t))$, $\tilde{e}(t_0) = 0$.
Dus $\|e(t)\| \leq C \int \|A(\tau, e(\tau))\| \leq C\bar{a}\sup_\tau \|e(\tau)\|^2 \leq \bar{a}C(C\bar{\delta})^2$.

Het bewijs (b) laten we aan de lezer over (bedenk dat $\tilde{u}(t) - u^*(t) = G(t,\tau)\delta_\tau$). □

**Samentrekkende differentiaalvergelijkingen**

In opgave 1.2.14 hebben we gezien dat de schatting voor het konditie getal in 1.2.8 veel te grof kan zijn. Als $f \in C^2(\Omega, \mathbf{R}^d)$ is en de funktionaalmatrix $J$ niet sterk varieert als funktie van $t$ dan zouden we kunnen hopen dat de resultaten in 1.2.12 ook in dat geval nog min of meer van toepassing zijn (als $\mathrm{Re}(\eta_j(t)) \leq \eta < 0$ voor iedere $t$ en iedere eigenwaarde $\eta_j(t)$ van $J(t)$ dan zal het konditie getal wel niet veel groter zijn dan 1, etc.). Het is echter verre van duidelijk hoe dit precies uitpakt[3] (zie voorbeeld 1.2.21). Verder werkt deze benadering zeker niet als $J$ snel varieert of als $f$ niet differentieerbaar is.

De volgende stelling geeft ook een interessante uitspraak voor een grote klasse van dit soort "wilde" funkties.

**1.2.18 Definitie.** Voor $\eta \in \mathbf{R}$ is $f$ *$\eta$-samentrekkend* als,

$$< f(t,x) - f(t,y), x - y > \leq \eta\|x - y\|_2^2 \quad \text{voor alle} \quad (t,x), (t,y) \in \Omega.$$

$\eta$ wordt ook wel een *eenzijdige* Lipschitz konstante genoemd.

**1.2.19 Voorbeeld.** De $f$ in 1.1.9 is $\eta$-samentrekkend als voor iedere $t \in \mathcal{J}$ iedere eigenwaarde van $\frac{1}{2}[J(t)^* + J(t)]$ in $(-\infty, \eta]$ zit.
(Met $f(t,x) = J(t)x + g(t)$ en $e = x - y$ is

$$< f(t,x) - f(t,y), x - y > = < J(t)e, e > = < \tfrac{1}{2}[J(t)^* + J(t)]e, e > \leq \lambda_m(t)\|e\|_2^2,$$

waarbij $\lambda_m(t)$ de maximale eigenwaarde is van de symmetrische matrix $\frac{1}{2}[J(t)^* + J(t)]$.)
Merk op dat $\eta \leq L = \sup_t \|J(t)\|$, maar dat ook $\eta \ll L$ kan zijn.

**1.2.20 Stelling.** *Stel $f$ is $\eta$-samentrekkend.*
*Als* $v, w \in C^1(\mathcal{J}, \mathbf{R}^d)$ *oplossingen zijn van de differentiaalvergelijking* $u'(t) = f(t, u(t))$ *dan*

$$\|v(t) - w(t)\|_2 \leq e^{\eta(t-\tau)}\|v(\tau) - w(\tau)\|_2 \quad \textit{voor alle} \quad t, \tau \in \mathcal{J}, t \geq \tau.$$

*Verder is* 1.2.6 *van toepassing met* $\mu = \eta$ *en* $C = 1$*:*

$$\begin{aligned} T < \infty \quad &\Rightarrow \quad \mathcal{C}(u^*) \leq 1 + e^{\eta T} \\ \eta < 0 \quad &\Rightarrow \quad \tilde{\mathcal{C}}(u^*) \leq 1 + \tfrac{1}{|\eta|}. \end{aligned}$$

[3] In geval $t \to J(t)$ periodiek is met periode $\tau$ (dus $J(t+\tau) = J(t)$ voor alle $t$) is $t \to Y(t) := G(t, t_0)$ periodiek met periode $\tau$. Dus, met $C := Y(t_0 + \tau)$ is $Y(t+\tau) = Y(t)C$ voor iedere $t$. Voor een niet al te groot konditie getal moeten de eigenwaarden van $C$ in modulus $< 1$ zijn.

*Bewijs.* Met $e(t) := \tilde{u}(t) - u^*(t)$ en $\epsilon(t) := \|e(t)\|_2^2$ voor $t \in \mathcal{J}$ is $\epsilon'(t) = 2 < e'(t), e(t) > \leq 2 < f(t, \tilde{u}(t)) - f(t, u^*(t)), e(t) > + < (\delta(t), e(t) > \leq 2\eta\epsilon(t) + 2\|\delta(t)\|_2 \sqrt{\epsilon(t)}$. Het resultaat volgt nu eenvoudig uit lemma 1.2.7 (zie ook het bewijs van 1.2.8). $\square$

Het volgende voorbeeldje (van Markus en Yamabe) laat zien dat we voorzichtig moeten zijn met aan te nemen dat de resultaten in 1.2.12 ook nog min of meer van toepassing zijn als $f \in C^2(\Omega, \mathbf{R}^d)$ is en de funktionaalmatrix $J$ niet sterk varieert als funktie van $t$.

**1.2.21 Voorbeeld.** Laat, voor ieder $t \in \mathcal{J} := [0, T]$,

$$J(t) := \begin{bmatrix} -1 & 4 \\ -4 & -1 \end{bmatrix} + 3 \begin{bmatrix} \cos t & -\sin t \\ -\sin t & -\cos t \end{bmatrix}. \tag{13}$$

Dan heeft $J(t)$ voor iedere $t \geq 0$ eigenwaarden $\eta_1(t) = -1 + i\sqrt{7}$ en $\eta_2(t) = -1 - i\sqrt{7}$. We zien dat $\eta = \max_t \max_i \mathrm{Re}\eta(t) = -1$. Door in te vullen gaat men echter eenvoudig na dat

$$e(t) := \delta_0(\cos\frac{1}{2}t, -\sin\frac{t}{2})^T \exp(\frac{1}{4}t) \qquad (t \in [0, T])$$

de oplossing is van het probleem

$$u'(t) = J(t)u(t) \quad \text{voor} \quad t \in [0, T] \quad \text{en} \quad u(0) = (\delta_0, 0)^T.$$

Ook al varieert $J(t)$ niet al te wild en is $\eta < 0$ toch kan het probleem slecht gekonditioneerd zijn: met betrekking tot de oplossing $u^* \equiv 0$ geldt in bovenstaand probleem $\mathcal{C}(u^*) \geq \exp\frac{T}{4}$.

Merk op dat in het onderhavige voorbeeld voor iedere $t$, $+2$ en $-4$ de eigenwaarden zijn van $\frac{1}{2}[J(t)^* + J(t)]$: de bijbehorende $f$ is 2–samentrekkend en *niet* 0– of $-1$–samentrekkend!

# 2 Numerieke oplosmethoden: inleiding

## 2.1 Het diskrete probleem

We zullen methoden bestuderen waarmee we numeriek de oplossing van het probleem (1) in 1.1.1 kunnen benaderen. In deze paragraaf voeren we wat notaties en konventies in en we formuleren een paar vragen die ons hier in de numerieke theorie zullen bezig houden.

**2.1.1** Verder gaan we ervan uit dat $\mathcal{J} = [t_0, t_0+T]$ (begrensd) is, tenzij we expliciet zeggen dat $\mathcal{J} = [t_0, \infty)$.

**2.1.2 Roosterfunkties.** Voor $h > 0$ zullen we geïnteresseerd zijn in $\mathbf{R}^d$-waardige funkties $u_h$ die gedefiniëerd zijn op het *rooster* (tijdsrooster) $\mathcal{J}_h := \{t_0 + nh \in \mathcal{J} \,\big|\, n \in \mathbf{N}_0\}$ en die op dat rooster de funktie $u^*$ in een of andere zin redelijk benadert (zie ook 2.1.4).

De ruimte van funkties $v_h$ van $\mathcal{J}_h$ naar $\mathbf{R}^d$ noteren we met $C(\mathcal{J}_h, \mathbf{R}^d)$.
We schrijven $C(\mathcal{J}_h)$ in plaats van $C(\mathcal{J}_h, \mathbf{R})$.

Als $v \in C(\mathcal{J}, \mathbf{R}^d)$ of als $v$ een funktie is op het rooster $\mathcal{J}_H$ met $h \in H\mathbf{N}$ dan kunnen we $v$ beperken tot het rooster $\mathcal{J}_h$. We noteren deze funktie ook met $v$.

**2.1.3 Notaties en konventies.** Voor iedere $h > 0$ hebben we te maken met een tijdsrooster en een benaderende oplossing $u_h$. Om de notatie niet te "zwaar" te laten zijn zullen we de afhankelijkheid van $h$ in de diverse grootheden niet telkens expliciet in de notatie aangeven. Hou die afhankelijkheid wel goed voor ogen! We gebruiken de volgende notatie.

Beschouw een $h > 0$.
We schrijven $t_n := t_0 + nh$ voor $n \in \mathbf{N}$, $nh \leq T$. Verder is

$$u_n^* := u^*(t_n) \quad \text{en} \quad f_n^* := f^*(t_n) := f(t_n, u^*(t_n)) \quad \text{voor} \quad n \in \mathbf{N}, nh \leq T.$$

Als $u_h : \mathcal{J}_h \to \mathbf{R}^d$ een roosterfunktie is dan schrijven we

$$u_n := u_h(t_n) \quad \text{en} \quad f_n := f_h(t_n) := f(t_n, u_n) = f(t_n, u_h(t_n)) \quad \text{voor} \quad n \in \mathbf{N}, nh \leq T.$$

We zullen numerieke benaderings methoden bestuderen die voor iedere $h > 0$ (of voor iedere $h$ uit een positieve naar 0 dalende rij; bv. $h = \frac{1}{n}$ $(n \in \mathbf{N})$ ) een $u_h$ op $\mathcal{J}_h$ produceert die $u^*$ benadert. We zullen ons ondermeer bezighouden met de volgende vragen.

- "Konvergeert" de rij $(u_h)$ naar $u^*$, waarbij we geïnteresseerd zijn in de konvergentie zoals die hier beneden in 2.1.4 gedefiniëerd is?
- Is $u_h$ met behulp van de methode "stabiel" te berekenen: is het effekt van rekenfouten zo klein dat de berekende benadering $u_h^*$, berekend met de methode, 'goed lijkt' op de beoogde benadering $u_h$?
- Is de methode "efficiënt": krijgen we al met grotere $h$ een redelijke benadering $u_h$ van $u^*$?

We geven hieronder al een exacte mathematische inhoud aan het begrip "konvergentie". We laten nog even in het midden wat we precies bedoelen met "stabiel" en "efficiënt". Met name van het begrip "stabiel" hangt de exacte betekenis sterk af van het type gewone differentiaalvergelijking die men wenst op te lossen en van bijvoorbeeld de hoeveelheid werk die men wil verrichten om de oplossing benaderend te berekenen.

**2.1.4 Konvergentie.** Zij $\mathbf{H} \subset (0,\infty)$ met verdichtingspunt 0 (bv. $\mathbf{H} = \{\frac{1}{n} \,\big|\, n \in \mathbf{N}\}$). Voor iedere $h \in \mathbf{H}$ is $v_h$ een $\mathbf{R}^d$–waardige funktie op $\mathcal{J}_h$. $v$ is een $\mathbf{R}^d$–waardige funktie op $\mathcal{J}$. De familie $(v_h)$ van rooster funkties *konvergeert* op $\mathcal{J}$ naar $v$ als[4]

$$\sup_{t_n \in \mathcal{J}_h} \|v_h(t_n) - v(t_n)\| \to 0 \quad \text{als} \quad h \to 0.$$

De *konvergentie is van orde* $l$ voor een $l \in \mathbf{N}$ als voor zekere $\bar{C} > 0$ en $h_0 > 0$ geldt

$$\sup_{t_n \in \mathcal{J}_h} \|v_h(t_n) - v(t_n)\| \leq \bar{C}h^l \quad \text{voor alle} \quad h \in \mathbf{H}, h \leq h_0;$$

we schrijven dan

$$\begin{array}{llll} v_h(t_n) = v(t_n) + \mathcal{O}(h^l) & \text{uniform} & (\text{op } \mathcal{J}) & (h \to 0) \quad \text{of} \\ v_h = v + \mathcal{O}(h^l) & \text{uniform} & (\text{op } \mathcal{J}) & (h \to 0). \end{array}$$

**2.1.5 Opmerking.** Merk op dat de konvergentie *uniform* is op $\mathcal{J}$.

Als $(v_h)$ konvergeert naar $v$ dan kunnen we voor iedere $t \in \mathcal{J}$ de funktiewaarde $v(t)$ benaderen met roosterfunktiewaarden: dan $\lim_{h\to 0} v_h(t_n) = v(t)$ mits $n$ voor iedere $h$ zo gekozen is dat $t_n = t_0 + nh \to t$ als $h \to 0$ (in het bijzonder gaat $n$ dan naar $\infty$).

## 2.2 Elementaire oplosmethoden

### 2.2.1 Eindige differenties.

Beschouw, voor $h > 0$ en voor funkties $v \in C(\mathcal{J}_h)$, de volgende eindige differenties.

**Voorwaartse differentie**

$$\partial_h v(t_n) = \partial_h^+ v(t_n) := \frac{1}{h}[v(t_{n+1}) - v(t_n)] \qquad (t_n \in \mathcal{J}_h \text{ zodat ook } t_{n+1} \in \mathcal{J}_h) \tag{14}$$

**Terugwaartse differentie**

$$\partial_h v(t_n) = \partial_h^- v(t_n) := \frac{1}{h}[v(t_n) - v(t_{n-1})] \qquad (t_n \in \mathcal{J}_h \text{ zodat ook } t_{n-1} \in \mathcal{J}_h) \tag{15}$$

**Centrale differentie**

$$\partial_h v(t_n) = \partial_h^0 v(t_n) := \frac{1}{2h}[v(t_{n+1}) - v(t_{n-1})] \qquad (t_n \in \mathcal{J}_h \text{ zodat ook } t_{n-1}, t_{n+1} \in \mathcal{J}_h) \tag{16}$$

'n hogere orde differentie

$$\begin{array}{l} \partial_h v(t_n) := \frac{1}{12h}[v(t_{n+2}) - 8v(t_{n+1}) + 8v(t_{n-1}) - v(t_{n-2})] \\ (t_n \in \mathcal{J}_h \quad \text{zodat ook} \quad t_{n+j} \in \mathcal{J}_h,\ j = -2, \ldots, 2) \end{array} \tag{17}$$

De differenties $\partial_h v(t_n)$ benaderen voor funkties $v$ die glad genoeg zijn in $t_n$ de afgeleide $v'(t_n)$ van $v$ (door $v$ te beperken tot $\mathcal{J}_h$ hebben we een funktie op $\mathcal{J}_h$).
Voor $v \in C^1(\mathcal{J})$ en $t_n \in \mathcal{J}_h$ [5] is

$$\delta_n := \delta_h(v)(t_n) := \partial_h v(t_n) - \frac{dv}{dt}(t_n)$$

de *lokale diskretisatie fout* (van de diskretisatie $\partial_h$ van $\frac{d}{dt}$ m.b.t $v$ in $t_n$).

[4] Zij $k \in \mathrm{N}$. $\tilde{\mathcal{J}}_h$ is een deelverzameling van $\mathcal{J}_h$ die iedere $t_n$ uit $\mathcal{J}_h$ bevat waarvoor $t_{n-k}, t_{n+k} \in \mathcal{J}$. We gebruiken dezelfde terminologie in geval de $v_h$'s slechts gedefiniëerd zijn op $\tilde{\mathcal{J}}_h$.

[5] In feite alleen voor $t_n$ in het rooster $\tilde{\mathcal{J}}_h$ waarvoor $\partial_h v(t_n)$ gedefiniëerd is.

Voor bovenstaande diskretisaties geldt voor voldoende gladde funkties $v$ achtereenvolgens:

$$\delta_n = +\frac{1}{2}hv''(\xi_n) \quad \text{voor zekere} \quad \xi_n \in (t_{n-1}, t_n) \quad \textit{(voorwaarts)} \tag{18}$$

$$\delta_n = -\frac{1}{2}hv''(\xi_n) \quad \text{voor zekere} \quad \xi_n \in (t_n, t_{n+1}) \quad \textit{(terugwaarts)} \tag{19}$$

$$\delta_n = +\frac{1}{6}h^2 v^{(3)}(\xi_n) \quad \text{voor zekere} \quad \xi_n \in (t_{n-1}, t_{n+1}) \quad \textit{(centraal)} \tag{20}$$

$$|\delta_n| \leq h^4 \frac{1}{18} \max_{\xi} |v^{(5)}(\xi)| \tag{21}$$

We zeggen dat de *differentie operator* $\partial_h$ een *konsistente* diskretizering is van $\frac{d}{dt}$ als voor iedere $v \in C^1(\mathcal{J})$ geldt

$$\sup_{t_n \in \tilde{\mathcal{J}}_h} |\partial_h v(t_n) - \frac{dv}{dt}(t_n)| \to 0 \quad \text{als} \quad h \to 0.$$

De diskretizering is voor een $l \in \mathbf{N}$ *met betrekking tot* $v$ *konsistent van orde* $l$ als

$$\partial_h v = \frac{dv}{dt} + \mathcal{O}(h^l) \quad \text{uniform} \quad (h \to 0).$$

De diskretizering is *konsistent van orde* $l$ als hij dat is met betrekking tot iedere voldoende gladde $v$.
De lokale diskretiesatiefouten zijn dan dus over het hele interval uniform klein.

De diskretizeringen die we hierboven expliciet gedefiniëerd hebben zijn konsistent en met betrekking tot voldoende gladde funkties van orde 1, 1, 2, 4 respektievelijk.

De differentie operatoren en de diskretiesatie fouten zijn hier toegelicht aan de hand van $\mathbf{R}$–waardige funkties. Door koördinaatsgewijs te werken vindt men onmiddellijk uitspraken voor $\mathbf{R}^d$–waardige funkties.

De afgeleide operator $v \to v'(t)$ is gedefiniëerd voor funkties $v$ op $\mathcal{J}$ die voldoende glad zijn. Het ligt voor de hand om voor roosterfunkties deze afgeleide operator te benaderen door differentie operatoren.

We wensen dus op het rooster $\mathcal{J}_h$ de volgende differentievergelijking met beginvoorwaarde te onderzoeken.

### 2.2.2 Een differentievergelijking.

We zijn geïnteresseerd in de oplossing $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ van de volgende *eindige differentievergelijking* met beginvoorwaarde.

$$\begin{cases} \partial_h u_h(t_n) = f(t_n, u_h(t_n)) & \text{voor alle} \quad t_n \in \tilde{\mathcal{J}}_h \\ u_h(t_0) = u_0, \end{cases} \tag{22}$$

waarbij $\partial_h$ een differentie operator is.

Door de differentievergelijking op een voor de hand liggende manier te herschrijven krijgt men een rekursie voor de roosterfunktiewaarden $u_n = u_h(t_n)$ waarmee men door iteratieve toepassing, startend met $u_0$, de roosterfunktiewaarden $u_n$ kan berekenen.

Voor de differenties in 2.2.1 geven we expliciet hieronder die rekursie voor roosterfunktiewaarden $u_n$ en de gerelateerde rekursie voor de exacte funktiewaarden $u_n^* = u^*(t_n)$.

### 2.2.3 Elementaire oplosmethoden.

**Euler forward** (E.f.)

$$\begin{cases} u_n = u_{n-1} + hf_{n-1} \\ u_n^* = u_{n-1}^* + hf_{n-1}^* + h\delta_n \quad \text{met} \quad \delta_n = +\frac{1}{2}hu^{*\prime\prime}(\xi_n) \quad \text{zekere} \quad \xi_n \in (t_{n-1}, t_n) \end{cases} \tag{23}$$

voor $n = 1, 2, \ldots,$ zodat $nh \leq T$ en $u_0 = u_0^*$

**Euler backward** (E.b.)

$$\begin{cases} u_n = u_{n-1} + hf_n \\ u_n^* = u_{n-1}^* + hf_n^* + h\delta_n \quad \text{met} \quad \delta_n = -\frac{1}{2}hu^{*\prime\prime}(\xi_n) \quad \text{zekere} \quad \xi_n \in (t_{n-1}, t_n) \end{cases} \tag{24}$$
$$\text{voor} \quad n = 1, 2, \ldots, \quad \text{zodat} \quad nh \leq T \quad \text{en} \quad u_0 = u_0^*$$

**Midpunt regel** (M.p.)

$$\begin{cases} u_n = u_{n-2} + 2hf_{n-1} \\ u_n^* = u_{n-2}^* + 2hf_{n-1}^* + 2h\delta_n \quad \text{met} \quad \delta_n = \frac{1}{6}h^2u^{*(3)}(\xi_n) \quad \text{zekere} \quad \xi_n \in (t_{n-1}, t_n) \end{cases} \tag{25}$$
$$\text{voor} \quad n = 2, 3, \ldots, \quad \text{zodat} \quad nh \leq T \quad \text{en} \quad u_0 = u_0^*$$

(x)

$$\begin{cases} u_n = 8u_{n-1} - 8u_{n-3} + u_{n-4} + 12hf_{n-2} \\ u_n^* = 8u_{n-1}^* - 8u_{n-3}^* + u_{n-4}^* + 12hf_{n-2}^* + 12h\delta_n \quad \text{met} \quad \delta_n \leq \frac{h^4}{18}\max_\xi |u^{*(5)}(\xi)| \end{cases} \tag{26}$$
$$\text{voor} \quad n = 4, 5, \ldots, \quad \text{zodat} \quad nh \leq T \quad \text{en} \quad u_0 = u_0^*$$

Bezien we bovenstaande formules, dan vallen ons een aantal zaken op:

• Met de Euler forward methode kunnen we uit $u_0$ onmiddellijk $u_1$ bepalen en vervolgens, met $u_1$, kunnen we $u_2$ bepalen, etc.

• Met Euler backward kunnen we evenzo iteratief iedere $u_n$ berekenen. Bij deze methode komt $u_n$ ook in het rechterlid voor en moeten we om $u_n$ uit $u_{n-1}$ te bepalen $u_n$ uit de vergelijking $u_n = u_{n-1} + hf(t_n, u_n)$ oplossen. Als $d > 1$ en $f$ niet lineair is kan dit bezwaarlijk zijn. We zullen verderop zien dat dit oplossen met succesieve substitutie of Newton Raphson onder milde voorwaarde succesvol zal verlopen. $u_n$ is dus impliciet gegeven als $u_k$ voor $k < n$ bekend is. We noemen de methode *impliciet.* Euler backward heeft, zoals zal blijken, betere stabiliteits eigenschappen dan Euler forward.

• De midpunt regel is evenals Euler forward *expliciet*: zijn de $u_k$ voor $k < n$ bekend dan kunnen we onmiddellijk $u_n$ bepalen. Helaas vertelt de midpuntregel ons niet wat $u_1$ is. De midpunt regel is een zogenaamde *twee staps methode*: "we hebben twee stappen uit het verleden nodig om een stap in de toekomst te kunnen zetten". Om $u_1$ te bepalen hebben we een aparte *startprocedure* nodig. We zullen verderop zien waaraan een geschikte startprocedure moet voldoen. De midpunt regel heeft een hogere orde lokale nauwkeurigheid dan Euler methoden (lokale diskretisatiefout is van hogere orde). De gedachte dringt zich op dat men met een grotere $h$, dus met minder rekenwerk, een benadering kan krijgen die $u^*$ even nauwkeurig benadert als de Euler forward oplossing met kleinere $h$.

• Hoewel de naamloze methode een vier staps methode is lijkt hij mischien aantrekkelijk omdat hij expliciet is en een zeer hoge orde lokale nauwkeurigheid heeft. De methode is echter zoals we hier beneden zullen zien zo gevoelig voor fouten (lokale diskretisatie fouten en rekenfouten) dat hij absoluut onbruikbaar is.

**2.2.4 Instabiliteits voorbeeld.** We onderzoeken wat het effekt is van foutjes als we het probleem

$$\begin{cases} u'(t) = 0 \quad \text{voor} \quad t \geq 0 \\ u(0) = 0 \end{cases}$$

benaderend oplossen met de naamloze methode: $\Omega = \mathbf{R} \times \mathbf{R}$, $f = 0$. De differentiaalvergelijking is welliswaar uiterst banaal, maar als we zelfs deze differentiaalvergelijking niet redelijk kunnen oplossen met de naamloze methode kunnen we niet verwachten dat we ingewikkeldere vergelijkingen enigszinds nauwkeurig benaderend kunnen oplossen.

Met $\mu = 4 + \sqrt{15}$ is $\mu^4 - 8\mu^3 + 8\mu - 1 = (\mu^2 - 1)(\mu^2 - 8\mu + 1) = 0$.
Beschouw een $\varepsilon > 0$, $\varepsilon \ll 1$. Stel dat $u_j = \varepsilon\mu^j$ voor $j = 0, 1, 2, 3$: de startwaarden zijn welliswaar niet exact bekend, maar de fout is klein. Dan voldoet $u_n := \varepsilon\mu^n$ aan (zie (26))

$$u_n = 8u_{n-1} - 8u_{n-3} + u_{n-4} + 12hf_{n-2} = 8u_{n-1} - 8u_{n-3} + u_{n-4}.$$

| | voordeel | nadeel | bruikbaar |
|---|---|---|---|
| E.f. | zelf startend<br>expliciet | lagere orde<br>nauwkeurigheid | ja |
| E.b. | zelf startend<br>"stabieler" | lagere orde<br>nauwkeurigheid<br>impliciet | ja |
| M.p. | hogere orde<br>nauwkeurigheid<br>expliciet | zwak stabiel<br>niet zelfstartend | ja |
| x | hogere orde<br>nauwkeurigheid<br>expliciet | instabiel<br>niet zelfstartend | nee |

Tabel 1: eenvoudige methoden

Als $n$ zo is dat $nh \to 1 \quad (h \to 0)$ dan $\lim_{h\to 0} u_n - u^*(1) = \lim_{h\to 0} u_n = \infty$.
Voor niet al te grote $n$ is al $|u_n| \gg |u_3|$ (zo is bv. $|u_{14}| \geq 10^9 |u_3|$).

Men zou nog kunnen hopen dat het slechte foutvoortplantings gedrag, dat we hierboven gezien hebben, voortvloeit uit de speciale 'startfouten'. We zullen in de volgende paragraaf zien dat deze hoop ijdel is.

De midpunt regel heeft slechtere stabiliteits eigenschappen dan Euler forward, maar is niet zo instabiel dat hij onbruikbaar is. We zullen verderop uitgebreid aandacht besteden aan de diverse stabiliteits eigenschappen.

Tabel 1 vat de vermelde voor– en nadelen samen.

**2.2.5 Kwadratuur.** Men kan, voor $h > 0$, de differentiaalvergelijking ook omschrijven tot een integraalvergelijking:

$$\begin{cases} u^*(t_n) = u^*(t_{n-k}) + \int_{t_{n-k}}^{t_n} f(t, u^*(t))\, dt \qquad \text{voor iedere} \quad n \geq k, \\ u^*(t_0) = u_0 \end{cases}$$

Door de integraal te vervangen door een kwadratuurformule kunnen we deze vergelijking ook diskretiseren.
Met bijvoorbeeld de trapezium regel $\int_{t_{n-1}}^{t_n} g(t)\, dt = \frac{1}{2}h[g(t_n) + g(t_{n-1})] - \frac{1}{12}h^3 g''(\xi_n)$ vinden we de volgende rekurrente betrekkingen voor de roosterfunktiewaarden $u_n$ en de exacte funktiewaarden $u_n^*$.

**Trapezium regel** Voor $n = 1, 2, \ldots,$ zodat $nh \leq T$ geldt

$$\begin{cases} u_n = u_{n-1} + \frac{1}{2}h(f_n + f_{n-1}) \\ u_n^* = u_{n-1}^* + \frac{1}{2}h(f_n^* + f_{n-1}^*) + h\delta_n \\ \quad \text{met} \quad \delta_n = -\frac{1}{12}u^{*(3)}(\xi_n)\, h^2 \quad \text{zekere} \quad \xi_n \in (t_{n-1}, t_n) \\ u_0 = u_0^*. \end{cases} \tag{27}$$

Met bijvoorbeeld de Simpson regel $\int_{t_{n-2}}^{t_n} g(t)\, dt = \frac{1}{3}h[g(t_n) + 4g(t_{n-1}) + g(t_{n-2})] - \frac{1}{90}h^5 g^{(4)}(\xi_n)$ vinden we de volgende rekurrente betrekkingen voor de roosterfunktiewaarden $u_n$ en de exacte funktiewaarden $u_n^*$.

**Simpson regel** Voor $n = 1, 2, \ldots,$ zodat $nh \leq T$ geldt

$$\begin{cases} u_n = u_{n-2} + \frac{1}{3}h(f_n + 4f_{n-1} + f_{n-2}) \\ u_n^* = u_{n-2}^* + \frac{1}{3}h(f_n^* + 4f_{n-1}^* + f_{n-2}^*) + 2h\delta_n \\ \quad \text{met} \quad \delta_n = -\frac{1}{180}h^4 u^{*(5)}(\xi_n) \quad \text{zekere} \quad \xi_n \in (t_{n-2}, t_n) \\ u_0 = u_0^*. \end{cases} \tag{28}$$

# 3 Intermezzo: rekursies

In de stabiliteits analyses van de diverse methoden komen we rekursies tegen (als bijvoorbeeld in 2.2.4). In deze paragraaf geven we een paar elementaire resultaten over de 'groei' van de oplossing van zekere voor ons interessante rekursies. Deze resultaten zullen we herhaaldelijk gebruiken.

**3.1.1 Rekursies.** Zij $k \in \mathbf{N}$. Laat, voor iedere $j = 1, \ldots, k$, $(c_{jn})_n$ een rij in $\mathbf{C}$ gegeven zijn. Verder is $(g_n)$ ook een gegeven rij in $\mathbf{C}$.

We zijn geïnteresseerd in de rijen $(u_n)$ in $\mathbf{C}$ waarvoor geldt

$$u_{n+k} + c_{1n}u_{n+k-1} + \cdots + c_{kn}u_n = g_n \quad \text{voor alle} \quad n \in \mathbf{N}_0. \tag{r}$$

$(u_n)$ voldoet dan aan een *rekursieve relatie* of *rekursie.*
De $g_n$ zijn de *inhomogene termen* van de rekursie (r) en de $c_{jn}$ de *koëfficiënten.*

De rekursie (r) is *homogeen* als $g_n = 0$ voor alle $n \in \mathbf{N}_0$.
De rekursie heeft *konstante koëfficiënten* als
voor iedere $j = 1, \ldots, k$ er een $c_j$ is zodat $c_{jn} = c_j$ voor alle $n \in \mathbf{N}_0$.

**3.1.2 Bewering.** *De oplossingsruimte*

$$\mathcal{S} := \{(u_n)_{n\in\mathbf{N}_0} \,\big|\, u_{n+k} + c_{1n}u_{n+k-1} + \cdots + c_{kn}u_n = 0 \quad \textit{alle} \quad n \in \mathbf{N}_0\}$$

*van de homogene rekursie is linear en $k$ dimensionaal.*

*Bewijs.* Iedere $k$-tal $u_0, u_1, \ldots, u_{k-1}$ in $\mathbf{C}$ bepaalt precies een oplossing $(u_n)$ van de homogene rekursie. □

**3.1.3 Definitie.** Stel de rekursie (r) heeft konstante koëfficiënten. Beschouw

$$\psi(\zeta) := \zeta^k + c_1\zeta^{k-1} + \ldots + c_k = (\zeta - \lambda_1)(\zeta - \lambda_2)\cdots(\zeta - \lambda_k) \qquad (\zeta \in \mathbf{C}).$$

$\psi$ is het zogenaamde *karakteristieke polynoom* van de rekursie (r) en de wortels $\lambda_1, \ldots, \lambda_k$ van $\psi$ zijn de zogenaamde *karakteristieke wortels* van de rekursie (r).

Merk op dat $\psi^{(j)}(\lambda) = 0$ voor $j = 0, \ldots, l-1$ als $l := \#\{j \,\big|\, \lambda = \lambda_j\} > 0$; $\lambda$ is dan een *$l$–voudige wortel* van $\psi$.

**3.1.4 Stelling.** *Stel dat de rekursie* (r) *konstante koëfficiënten heeft.*
(a) *Als $\lambda \in \mathbf{C}$ een $l$–voudige wortel van $\psi$ is dan*

$$(\lambda^n)_{n\in\mathbf{N}_0}, (n\lambda^{n-1})_{n\in\mathbf{N}_0}, \ldots, (n(n-1)\cdots(n-l+2)\lambda^{n-l+1})_{n\in\mathbf{N}_0} \in \mathcal{S}.$$

(b) *De verzameling van de oplossingen die in* (a) *aangegeven zijn vormen een basis voor $\mathcal{S}$.*

*Bewijs.* (a) Schrijf $c_0 := 0$. Merk op dat
$\sum_{j=0}^k c_{k-j}\lambda^{n+j} = \lambda^n\psi(\lambda)$,
$\sum_{j=0}^k c_{k-j}(n+j)\lambda^{n+j-1} \begin{array}{l} = n\sum_{j=0}^k c_{k-j}\lambda^{n+j-1} + \sum_{j=0}^k jc_{k-j}\lambda^{n+j-1} \\ = \lambda^{n-1}\psi(\lambda) + \lambda^n\psi'(\lambda) \end{array}$, etc.. (a) volgt hier eenvoudig uit.
(b) We moeten laten zien dat de aangegeven oplossingen lineair onafhankelijk zijn; uit een dimensie argument volgt dan dat ze een basis vormen.

We bewijzen lineair onafhankelijkheid voor het geval $k = 4$ en $\lambda := \lambda_1$, $\mu := \lambda_2 = \lambda_3 = \lambda_4$; een generalizering van de argumenten laten we aan de lezer over.
Beschouw de matrix $V$

$$V := \begin{bmatrix} 1 & \lambda & \lambda^2 & \lambda^3 \\ 1 & \mu & \mu^2 & \mu^3 \\ 0 & 1 & 2\mu & 3\mu^2 \\ 0 & 0 & 2 & 6\mu \end{bmatrix}.$$

De rijen van $V$ zijn precies de beginstukken van de rijtjes waarin we geïnteresseerd zijn. Zijn de rijen van $V$ lineair onafhankelijk, dan zijn de oneindig lange rijen dat ook. We tonen aan dat $V$ een triviale kern heeft; dan zijn de rijen van $V$ immers lineair onafhakelijk.

Beschouw $\vec{a} = (\alpha_0, \ldots, \alpha_3)^T \in \mathbf{C}^4$. Beschouw ook het polynoom $p(\zeta) := \alpha_0 + \alpha_1\zeta + \alpha_2\zeta^2 + \alpha_3\zeta^3 \quad (\zeta \in \mathbf{C})$. Merk op dat $V\vec{a} = (p(\lambda), p(\mu), p'(\mu), p''(\mu))^T$. Als $V\vec{a} = \vec{0}$ dan is $p(\zeta) = q(\zeta)(\zeta - \lambda)(\zeta - \mu)^3$ voor zekere polynoom $q$ (hoofdstelling van de algebra). Omdat $p$ van graad 3 is moet $q \equiv 0$ en dus $p \equiv 0$ en $\vec{a} = \vec{0}$. □

**3.1.5 Waarschuwing.** De oplossingen van de **homogene** rekursie (r) kan men alleen in geval van **konstante koëfficiënten** beschrijven met behulp van de karakteristieke wortels.

**3.1.6 Opgave.** Stel (r) heeft konstante koëfficiënten. Als $\lambda \in \mathbf{C}$ een $l$–voudige karakteristieke wortel is dan $(n^j\lambda^n) \in \mathcal{S}$ voor $j = 0, \ldots, l-1$. Ga dit na en zie in dat dit soort oplossingen ook een basis voor $\mathcal{S}$ vormen in geval $\psi(0) \neq 0$.

**3.1.7 Voorbeeld.** Beschouw de rekursie in 2.2.4:

$$u_{n+4} - 8u_{n+3} + 8u_{n+1} - u_n = 0 \qquad (n \in \mathbf{N}_0).$$

Laat $(u_n)$ aan de rekursie voldoen. De stelling vertelt ons dat er $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbf{C}$ zodat

$$u_n = \alpha_1 + \alpha_2(-1)^n + \alpha_3(4 - \sqrt{15})^n + \alpha_4(4 + \sqrt{15})^n \qquad (n \in \mathbf{N}_0).$$

De skalairen $\alpha_i$ hangen af van de waarden $u_0, u_1, u_2, u_3$. Als $\alpha_4 \neq 0$ dan $\lim_{n\to\infty} |u_n| = \infty$. Kiezen we $u_0, \ldots, u_3$ 'random' dan hebben we 100% kans dat $\alpha_4 \neq 0$!

De *$k$-staps* of *$k+1$-terms skalaire rekursie* kan men in verband brengen met een 1-staps matrix–vektor rekursie, de zogenaamde companion rekursie.

**3.1.8 Companion matrices.** Laat, voor iedere $j = 1, \ldots, k$, de rij $(c_{jn})_n$ zijn als 3.1.1. $(g_n)$ is ook als in 3.1.1. Beschouw voor iedere $n \in \mathbf{N}_0$ de vektoren $\vec{g}_n \in \mathbf{C}^k$ en de *companion* matrices $A_n \in \mathbf{M}_k(\mathbf{C})$ die geven zijn door

$$\vec{g}_n := \begin{bmatrix} 0 \\ \vdots \\ 0 \\ g_n \end{bmatrix} \quad \text{en} \quad A_n := \begin{bmatrix} 0 & 1 & 0 & \ldots & & 0 \\ 0 & 0 & 1 & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & & \\ \vdots & & & \ddots & \ddots & \\ 0 & 0 & 0 & \ldots & 0 & 1 \\ -c_{kn} & -c_{k-1\,n} & -c_{k-2\,n} & \ldots & -c_{2n} & -c_{1n} \end{bmatrix}.$$

We zijn nu geïnteresseerd in de rijen $(\vec{v}_n)$ in $\mathbf{C}^k$ waarvoor geldt

$$\vec{v}_{n+1} = A_n\vec{v}_n + \vec{g}_n \quad \text{voor alle} \quad n \in \mathbf{N}_0. \tag{R}$$

Het resultaat in de volgende bewering legt het verband tussen de oplossingen $(u_n)$ van de skalaire rekursie (r) in 3.1.1 en de oplossingen $(\vec{v}_n)$ van de vektor rekursie (R) in 3.1.8; we laten de verifikatie van de bewering aan de lezer over. Vaak is het gemakkelijker om de een staps rekursie (R) te analyseren dan de meer staps rekursie (r).

**3.1.9 Bewering.** *Zij* $v_0, \ldots, v_{k-1} \in \mathbf{C}$.
*Laat* $u_j = v_j$ *zijn voor* $j = 0, \ldots, k-1$ *en* $\vec{v}_0 = (v_0, \ldots, v_{k-1})^T$.
(a) *Als* $(u_n)$ *in* $\mathbf{C}$ *aan* (r) *voldoet en* $\vec{v}_n := (u_n, \ldots, u_{n+k-1})^T$, *dan voldoet* $(\vec{v}_n)$ *aan* (R).
(b) *Als* $(\vec{v}_n)$ *in* $\mathbf{C}^k$ *aan* (R) *voldoet en*[6] $u_n := (\vec{v}_n, \vec{e}_1)$, *dan voldoet* $(u_n)$ *aan* (r). $\square$

In de stabiliteitsanalyse van de diverse numerieke oplosmethoden van differentiaalvergelijkingen lopen we nog al eens tegen rekursies aan. We zijn dan geïnteresseerd in de vraag of de oplossingen van de rekursie begrensd zijn of, als ze dat niet zijn, hoe "hard ze groeien". We geven hieronder kriteria waarmee men een majorant van de oplossing kan schatten.

**3.1.10 Bewering.** *Stel* (r) *is homogeen* ($g_n = 0$ alle $n$).
*Dan zijn de volgende uitspraken* (a) *t/m* (d) *equivalent.*
(a) *Voor iedere oplossing* $(u_n)$ *van* (r) *is er een* $K > 0$ *zodat*

$$\max\{|u_n| \,\Big|\, n \in \mathbf{N}_0\} \leq K \max_{0 \leq j < k} |u_j|.$$

(b) *Er is een* $K > 0$ *zodat voor iedere oplossing* $(u_n)$ *van* (r) *geldt*

$$\max\{|u_n| \,\Big|\, n \in \mathbf{N}_0\} \leq K \max_{0 \leq j < k} |u_j|.$$

(c) *Er is een* $K > 0$ *zodat voor iedere oplossing* $(\vec{v}_n)$ *van* (R) *geldt*

$$\max\{\|\vec{v}_n\| \,\Big|\, n \in \mathbf{N}_0\} \leq K \max_{0 \leq j < k} \|\vec{v}_j\|.$$

(d) *Er is een* $K > 0$ *zodat* $\max\{\|A_n A_{n-1} \cdots A_0\| \,\Big|\, n \in \mathbf{N}_0\} \leq K.$

*Als* (r) *ook nog konstante koëfficiënten heeft dan zijn de uitspraken* (a) *t/m* (d) *equivalent met de volgende uitspraak* (e).
(e) *Voor iedere* $\lambda \in \mathbf{C}$ *waarvoor* $\psi(\lambda) = 0$ *is òf* $|\lambda| < 1$ *òf* $|\lambda| = 1$ *en* $\psi'(\lambda) \neq 0$*:*
$\psi$ *voldoet aan* het wortel kriterium (karakteristieke wortels zijn in absolute waarde $\leq 1$ en de wortels op de eenheidscirkel zijn enkelvoudig).

*Bewijs.* De equivalentie van (b), (c) en (d) volgt envoudig uit de vorige bewering. De equivalentie van (a) en (b) volgt uit het feit dat $\mathcal{S}$ lineair en eindig dimensionaal is (de $K$ in (b) is $\leq$ de som van de $K$'s in (a) die horen bij de $k$ basis oplossingen).
De equivalentie van (e) en (a) in geval (r) homogeen is en konstante koëfficiënten heeft volgt eenvoudig uit stelling 3.1.4. $\square$

(e) vertelt ons of iedere oplossing van een homogene rekursie met konstante koëfficiënten begrensd is.
De stelling van Kreiss hieronder schat de "groei" van de oplossing van een rekursie met "bijna konstante koëfficiënten", mits de homogene variant van de "naburinge" rekursie met konstante koëfficiënten alleen begrensde oplossingen heeft. Om te zien wanneer dat laatste het geval is, gebruiken we nogal eens het wortel kriterium in (e) van bewering 3.1.10. In het bewijs van de stelling van Kreiss maken we (om aan te tonen dat de rekursie een oplossing heeft) gebruik van het bekende kontraktie lemma. Voor de volledigheid formuleren we dit lemma hier.

**3.1.11 Kontraktie Lemma.** *Zij* $\mathbf{B} \subset \mathbf{R}^d$, $\delta \in [0,1)$ *en* $\Phi : \mathbf{B} \to \mathbf{R}^d$ *zodat*

$$\|\Phi(x) - \Phi(y)\| \leq \delta \|x - y\| \quad \textit{voor alle} \quad x, y \in \mathbf{B}.$$

(a) *Er is hoogstens een* $z \in \mathbf{B}$ *waarvoor* $\Phi(z) = z$.
(b) *Als* $z \in \mathbf{B}$ *en* $\Phi(z) = z$ *dan is* $\|z\| \leq r := \frac{1}{1-\delta}\|\Phi(0)\|$.

[6] $\vec{e}_i$ zijn de standaard basis vektoren in $\mathrm{R}^k$ en $(\cdot,\cdot)$ is het standaard inprodukt.

(c) *Stel dat* $\mathbf{B}_r := \{y \in \mathbf{R}^d \,\big|\, \|y\| \leq r\} \subset \mathbf{B}$. *Dan*

(i) *bestaat* $z \in \mathbf{B}_r$ *met* $z = \Phi(z)$,

(ii) *is de rij* $(y_i)$, *waarvoor* $y_0 = 0$ *en* $y_i = \Phi(y_{i-1})$ *voor alle* $i \in \mathbf{N}$, *in* $\mathbf{B}_r$ *en*

(iii) *geldt dat* $\|z - y_i\| \leq \delta^i \|z\| \to 0$ *voor* $i \to \infty$.

*Bewijs.* (a) Als $y = \Phi(y)$ en $z = \Phi(z)$ dan $\|y - z\| = \|\Phi(y) - \Phi(z)\| \leq \delta\|y - z\|$ en dus $\|y - z\| = 0$.
(b) $\|z\| = \|\Phi(z)\| \leq \|\Phi(z) - \Phi(0)\| + \|\Phi(0)\| \leq \delta\|z\| + \|\Phi(0)\|$.
(c) $\|y_{i+1} - y_i\| = \|\Phi(y_i) - \Phi(y_{i-1})\| \leq \delta\|y_i - y_{i-1}\| \leq \ldots \leq \delta^i \|\Phi(0)\|$.
Dus $\|y_n\| = \| \sum_{i=0}^{n-1} y_{i+1} - y_i\| \leq \frac{1}{1-\delta}\|\Phi(0)\| = r$ en $y_n \in \mathbf{B}_r$.
Omdat $\|y_{i+k} - y_i\| = \| \sum_{j=1}^{k} y_{i+j} - y_{i+j-1}\| \leq \delta^i r$ is $(y_i)$ een Cauchy rij en konveregeert naar, zeg $y_\infty \in \mathbf{B}_r$. Omdat $\Phi$ kontinu is volgt $y_\infty = \Phi(y_\infty)$.
Tenslotte $\|y_\infty - y_i\| = \|\Phi(y_\infty) - \Phi(y_{i-1})\| \leq \delta\|y_\infty - y_{i-1}\| \leq \ldots \leq \delta^i \|y_\infty\|$. □

**3.1.12 De stelling van Kreiss.** *Zij* $A \in \mathbf{M}_k(\mathbf{C})$. *Zij* $(\vec{g}_n)$ *een rij in* $\mathbf{C}^k$. *Zij* $r > 0$, *en* $N \in \mathbf{N}$. *Voor iedere* $n \in \mathbf{N}_0$, $n \leq N$, *laat*

$$E_n, F_n : \mathbf{B}_r := \{\vec{w} \in \mathbf{C}^k \,\big|\, \|\vec{w}\,\| \leq r\} \to \mathbf{C}^k \quad \textit{met} \quad E_n(\vec{0}\,) = F_n(\vec{0}\,) = \vec{0}.$$

*Voor* $\vec{v}_0 \in \mathbf{C}^k$ *zijn we geïnteresseerd in de rij* $(\vec{v}_n)$ *die voldoet aan*

$$(I + E_n)\vec{v}_{n+1} = (A + F_n)\vec{v}_n + \vec{g}_n \quad \textit{voor alle} \quad n \in \mathbf{N}_0, n < N. \tag{R}$$

*Stel dat* $K > 0$, $\varepsilon \geq 0$ *en* $\delta \in [0, 1)$ *zo zijn dat voor iedere* $n \leq N$ *en* $x, y \in \mathbf{B}_r$ *geldt*

$$\|A^n\| \leq K, \quad \|F_n(\vec{x}\,) - F_n(\vec{y}\,)\| \leq \varepsilon\|\vec{x} - \vec{y}\,\|, \quad \|E_n(\vec{x}\,) - E_n(\vec{y}\,)\| \leq \delta\|\vec{x} - \vec{y}\,\|.$$

*Schrijf* $\tilde{K} := \frac{K}{1-\delta}$ *en* $\bar{r} := \tilde{K}(\|\vec{v}_0\| + \sum_{j=0}^{N-1} \|\vec{g}_j\|) \exp(N(\varepsilon + \delta)\tilde{K})$.
*Stel dat* $\bar{r} \leq r$. *Dan is er een rij* $(\vec{v}_n)_{n\leq N}$ *in* $\mathbf{B}_r$ *en er geldt voor iedere* $n \leq N$

$$\|\vec{v}_n\| \leq \tilde{K}(\|\vec{v}_0\| + \sum_{j=0}^{n-1} \|\vec{g}_j\|)(1 + (\varepsilon + \delta)\tilde{K})^n \leq \tilde{K}(\|\vec{v}_0\| + \sum_{j=0}^{n-1} \|\vec{g}_j\|)e^{n(\varepsilon+\delta)\tilde{K}} \leq \bar{r}.$$

*Bewijs.* Zij $n < N$. Stel dat de bewering korrekt is voor deze $n$ (ind. hyp.).
(i) Beschouw eerst de rij $(\tilde{v}_n)$ die voldoet aan

$$\tilde{v}_0 = \vec{v}_0 \quad \text{en} \quad \tilde{v}_{j+1} = A\tilde{v}_j + \vec{g}_j. \tag{$\tilde{\text{R}}$}$$

Merk op dat $\tilde{v}_{n+1} = A^{n+1}\vec{v}_0 + \sum_{j=0}^{n} A^{n-j}\vec{g}_j$ en dus

$$\|\tilde{v}_{n+1}\| \leq \|A^{n+1}\vec{v}_0\| + \sum_{j=0}^{n} \|A^{n-j}\vec{g}_j\| \leq M := K(\|\vec{v}_0\| + \sum_{j=0}^{n} \|\vec{g}_j\|).$$

(ii) Schrijf $\tilde{M} = \frac{M}{1-\delta}$.
We zoeken $\vec{v}_{n+1}$ in $\mathbf{B}_r$ waarvoor

$$(I + E_n)\vec{v}_{n+1} = (A + F_n)\vec{v}_n + \vec{g}_n \tag{R}$$

of met $\vec{f}_j := \vec{v}_j - \tilde{v}_j$ $(j \leq n+1)$, de $\vec{f}_{n+1}$ waarvoor geldt $\vec{f}_{n+1} = A\vec{f}_n + F_n\vec{v}_n - E_n(\tilde{v}_{n+1} + \vec{f}_{n+1})$ (verschil (R) en ($\tilde{\text{R}}$)). Voor $j < n$ zien we, met $\vec{e}_j := F_j\vec{v}_j$, $\vec{d}_j := E_{j-1}\vec{v}_j = E_{j-1}(\tilde{v}_j + \vec{f}_j)$, dat $\vec{f}_{j+1} = A\vec{f}_j + \vec{e}_j - \vec{d}_{j+1}$. Blijkbaar $\vec{f}_n = \sum_{j=1}^{n} A^{n-j-1}(\vec{e}_{j-1} - \vec{d}_j)$.
Met $\Phi(\vec{x}\,) := A\vec{f}_n + \vec{e}_n - E_n(\tilde{v}_{n+1} + \vec{x}\,)$ zoeken we $f_{n+1}$ waarvoor $\vec{f}_{n+1} = \Phi(\vec{f}_{n+1})$.
De induktie hypothese en het resultaat in (i) leveren ons dat

$$\|\Phi(0)\| \;\; \leq \;\; \|A\vec{f}_n + \vec{e}_n\| + \|E_n(\tilde{v}_{n+1})\| \leq \|\sum_{j=1}^{n} A^{n-j}(\vec{e}_{j-1} - \vec{d}_j) + \vec{e}_n\| + \delta M$$

$$
\begin{aligned}
&\leq \; K(\varepsilon+\delta)\sum_{j=0}^{n}\|\vec{v}_j\|+\delta M \leq K(\varepsilon+\delta)\tilde{M}\sum_{j=0}^{n}[1+(\varepsilon+\delta)\tilde{K}]^j+\delta M \\
&= \; M([1+(\varepsilon+\delta)\tilde{K}]^{n+1}-1)+\delta M \\
&= \; (\tilde{M}[1+(\varepsilon+\delta)\tilde{K}]^{n+1}-M)(1-\delta) \leq (1-\delta)\bar{r}.
\end{aligned}
$$

Omdat $\|\Phi(\vec{x})-\Phi(\vec{y})\| \leq \delta\|\vec{x}-\vec{y}\|$ is er volgens (c) van het kontraktie lemma een $\vec{f}_{n+1}$ met $\vec{f}_{n+1}=\Phi(\vec{f}_{n+1})$. (b) van dit lemma, (i) en de schatting voor $\|\Phi(0)\|$ leert tenslotte dat

$$\|\vec{v}_{n+1}\| = \|\tilde{v}_{n+1}+\vec{f}_{n+1}\| \leq \|\tilde{v}_{n+1}\| + \tfrac{1}{1-\delta}\|\Phi(0)\| \leq \tilde{M}(1+(\varepsilon+\delta)\tilde{K})^{n+1}. \qquad \square$$

**3.1.13 Opmerking.** Beschouw de situatie in de stelling van Kreiss $(\bar{r} \leq r)$. Stel dat $\vec{v}_n$ berekend is. Om $\vec{v}_{n+1}$ te bepalen moeten we $\vec{v}_{n+1}$ oplossen uit de vergelijking

$$(I+E_n)\vec{v}_{n+1} = \vec{x}_0 := (A+F_n)\vec{v}_n + \vec{g}_n.$$

Laat $\Psi(\vec{x}) := \vec{x}_0 - E_n(\vec{x})$ $(= \tilde{v}_{n+1} + \Phi(\vec{x}-\tilde{v}_{n+1})$ met $\Phi$ als in het bewijs hierboven).
Dan bestaat de rij $(\vec{y}^{(i)})$ waarvoor $\vec{y}^{(0)} = \vec{v}_n$ en $\vec{y}^{(i)} = \Phi(\vec{y}^{(i-1)})$ $(i \in \mathbf{N})$
en geldt $\|\vec{y}^{(i)} - \vec{v}_{n+1}\| \leq \delta^i \|\vec{y}^{(0)} - \vec{v}_{n+1}\| \to 0$ voor $i \to \infty$:
we kunnen dus $\vec{v}_{n+1}$ oplossen middels een voor de hand liggend succesief substitutie proces.

**3.1.14 Opmerking.** In zekere speciale gevallen worden een aantal voorwaarden in de stelling triviaal (zie hieronder 3.1.15). Geef een formulering van de stelling van Kreiss voor het geval $E_n$ en $F_n$ lineair zijn voor alle $n$ en ook voor het geval $E_n = 0$ voor alle $n$.

**3.1.15 Gevolg.** *Stel dat de rekursie* (r) *konstante koëfficiënten heeft.*
(a) *Als $\psi$ aan het wortel kriterium voldoet dan is er een* $K > 0$ *zodat voor iedere oplossing* $(u_n)$ *van* (r) *geldt*

$$|u_n| \leq K(\max_{0\leq j<k}|u_j| + \sum_{j=0}^{n}|g_j|) \qquad (n \in \mathbf{N}).$$

(b) *Als $|\lambda| < 1$ voor iedere wortel $\lambda$ van $\psi$ dan is er een* $\tilde{K} > 0$ *zodat voor iedere oplossing* $(u_n)$ *van* (r) *geldt*

$$|u_n| \leq \tilde{K}(\max_{0\leq j<k}|u_j| + \max_{j\leq n}|g_j|) \qquad (n \in \mathbf{N}).$$

*Bewijs.* (a) Volgens (d) van 3.1.10 is $K := \sup_n \|A^n\| < \infty$. De schatting volgt nu onmiddellijk uit de stelling van Kreiss (met $E_n = F_n = 0$ voor alle $n$, $\varepsilon = \delta = 0$).
(b) Kies $a \in (0,1)$ zodat $|\lambda| < a$ voor iedere wortel $\lambda$ van $\psi$.

Beschouw een oplossing $(u_n)$ van (r). Definieer $\tilde{u}_n := a^{-n}u_n$ en $\tilde{g}_n := a^{-n}g_n$ voor iedere $n$. Dan

$$a^k\tilde{u}_{n+k} + c_1a^{k-1}\tilde{u}_{n+k-1} + \ldots + c_k\tilde{u}_n = \tilde{g}_n \quad \text{voor alle} \quad n \in \mathbf{N}_0.$$

Met $\tilde{\psi}(\zeta) := \psi(a\zeta)$ is $\tilde{\psi}$ het karakteristiek polynoom van deze rekursie. Als $\tilde{\lambda}$ een wortel is van $\tilde{\psi}$ dan is $a\tilde{\lambda}$ een wortel van $\psi$. Dus $|\tilde{\lambda}| < 1$ voor iedere wortel $\tilde{\lambda}$ van $\tilde{\psi}$: $\tilde{\psi}$ voldoet aan het wortel kriterium. Met $K$ als in (a) volgt nu uit (a) en het feit dat $a < 1$ voor iedere $n$ dat

$$
\begin{aligned}
|u_n| &= a^n|\tilde{u}_n| \leq a^nK(\max_{0\leq j<k}|\tilde{u}_j| + \sum_{j=0}^{n}|\tilde{g}_j|) \\
&\leq K(\max_{0\leq j<k}a^{n-j}|u_j| + \sum_{j=0}^{n}a^{n-j}|g_j|) \leq K(\max_{0\leq j<k}|u_j| + \sum_{j=0}^{n}a^{n-j}\max_{j\leq n}|g_j|) \\
&\leq K(\max_{0\leq j<k}|u_j| + \frac{1}{1-a}\max_{j\leq n}|g_j|) \leq \frac{K}{1-a}(\max_{0\leq j<k}|u_j| + \max_{j\leq n}|g_j|). \qquad \square
\end{aligned}
$$

**3.1.16 Opmerking.** We zullen ook geïnteresseerd zijn in rijen $(\vec{u}_n)$ in $\mathbf{C}^d$ die voldoen aan een k-staps matrix-vektor rekursie: in (r) zijn dan de $c_{jn}$ in $\mathbf{M}_d(\mathbf{C})$ en de $g_n$ in $\mathbf{C}^d$. De companion matrices $A_n$ zijn in dat geval $kd \times kd$-matrices. In geval van konstante koëfficiënten (dan $A_n = A$ alle $n$) zijn de karakteristieke wortels van de rekursie de eigenwaarden (karakteristieke wortels) van $A$. De stellingen en beweringen hierboven in deze paragraaf laten zich eenvoudig generalizeren voor deze vektoriële rekursies: we laten de details aan de lezer over.

# 4 Numerieke oplosmethoden: multistep methoden

## 4.1 Konsistentie, stabiliteit en konvergentie

De diskretisatie methoden in 2.2, die gebaseerd zijn op eindige differentie benaderingen en op kwadratuur, zijn beide voorbeelden van numerieke oplosmethoden uit de grotere klasse van multistep methoden die we hieronder introduceren.

### 4.1.1 Multistep methoden.

Voor $k \in \mathbf{N}$, laat $\alpha_0, \ldots, \alpha_k$ en $\beta_0, \ldots, \beta_k$ in $\mathbf{R}$ zijn zodat $\alpha_0 \neq 0$ en $|\alpha_k| + |\beta_k| > 0$.

Voor $h > 0$ zijn we geïnteresseerd in de roosterfunkties $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ waarvoor geldt

$$\begin{array}{l} \frac{1}{h}(\alpha_0 u_{n+k} + \ldots + \alpha_k u_n) = \beta_0 f(t_{n+k}, u_{n+k}) + \ldots + \beta_k f(t_n, u_n) \\ \text{voor} \quad n = 0, 1, \ldots \quad \text{met} \quad (n+k)h \leq T, \quad \text{waarbij} \quad u_n = u_h(t_n). \end{array} \tag{M}$$

We zien (M) als een diskretizering van de differentiaalvergelijking in 1.1.1. Deze diskretisatie methode is een zogenaamde *$k$-staps methode* (multistep). De methode is *expliciet* als $\beta_0 = 0$, anders is de methode *impliciet.*

Voor verdere analyse representeren we (M) als volgt.
Definieer de polynomen $\rho$ en $\sigma$ door

$$\rho(\zeta) := \alpha_0 \zeta^k + \ldots + \alpha_{k-1}\zeta + \alpha_k \quad \text{en} \quad \sigma(\zeta) := \beta_0 \zeta^k + \ldots + \beta_k \qquad (\zeta \in \mathbf{C}).$$

Voor iedere $h > 0$, laat $T_h$ de *schuif operator* op $C(\mathcal{J}_h, \mathbf{R}^d)$ gedefiniëerd zijn door

$$T_h(v)(t_n) = v(t_{n+1}) \quad \text{voor} \quad v \in C(\mathcal{J}_h, \mathbf{R}^d), \quad t_n \in \mathcal{J}_h \text{ met } t_{n+1} \in \mathcal{J}_h.$$

Merk op dat $T_h^j(v)(t_n) = v(t_{n+j})$ voor $j = 0, 1, \ldots$.
Met $f_h(t_n) := f(t_n, u_h(t_n))$ voor $t_n \in \mathcal{J}_h$, kunnen we (M) nu schrijven als

$$\tfrac{1}{h}\rho(T_h)(u_h) = \sigma(T_h)(f_h) \quad \text{op} \quad \tilde{\mathcal{J}}_h := \{t_n \in \mathcal{J}_h \,\big|\, (n+k)h \leq T\}. \tag{M}$$

De polynomen $\rho$ en $\sigma$ karakteriseren het rekenschema (M) voor de multistep. We spreken dan ook over de multistep methode met *schema* $(\rho, \sigma)$.

**Konsistentie**

**4.1.2 Lokale diskretisatie fout en konsistentie.** Voor iedere $v \in C^1(\mathcal{J}, \mathbf{R}^d)$ en $h > 0$ is de *lokale diskretisatie fout* $\delta_h(v)(t_n)$ in $t_n \in \tilde{\mathcal{J}}_h$ gedefiniëerd door

$$\delta_h(v)(t_n) = \frac{1}{h}\rho(T_h)(v)(t_n) - \sigma(T_h)(\frac{dv}{dt})(t_n).$$

We zeggen dat de multistep methode *konsistent* is als voor iedere $v \in C^1(\mathcal{J}, \mathbf{R}^d)$ geldt

$$\sup_{t_n \in \tilde{\mathcal{J}}_h} \|\delta_h(v)(t_n)\| \to 0 \quad \text{als} \quad h \to 0.$$

Voor $l \in \mathbf{N}$ is de multistep *konsistent van orde $l$ met betrekking tot $v$* als

$$\delta_h(v) = \mathcal{O}(h^l) \quad \text{uniform} \quad (h \to 0).$$

Is de multistep konsistent van orde $l$ met betrekking tot iedere voldoend gladde funktie $v$ dan is de multistep *konsistent van orde* $l$indexkonsistent van orde $l$.

**4.1.3 Opmerking.** Bovenstaande 'konsistentie definities' sporen met die in 2.2.1: daar is immers[7] $\sigma = \chi^j$ voor 'n $j$ tussen 0 en $k$ en dus $\sigma(T_h)(\frac{dv}{dt})(t_n) = \frac{dv}{dt}(t_{n+j})$.

Als $\sigma = \chi^j$ kan men $\frac{1}{h}(\sigma(T_h))^{-1}\rho(T_h)(v)(t_n)$ interpreteren als $\frac{1}{h}(\alpha_0 v(t_{n+k-j}) + \ldots + \alpha_k v(t_{n-j}))$ en dus als een eindige differentie benadering van $\frac{dv}{dt}(t_n)$. Ook als $\sigma$ een ingewikkelder polynoom is kan men $\frac{1}{h}\sigma(T_h)^{-1}\rho(T_h)(v)(t_n)$ zien als een benadering van $\frac{dv}{dt}(t_n)$; nu wellicht als een '*on*eindige differentie benadering'. We werken deze visie hier echter niet verder uit: de theoretische kontekst waarin $\sigma(T_h)^{-1}$ fatsoenlijk gedefiniëerd is interesseerd ons in dit college volstrekt niet. Vanuit deze visie is het wel duidelijk waarom in de konsistentie definities het rechterlid $f$ van de differentiaalvergelijking geen rol speelt: we diskretizeren in feite $\frac{d}{dt}$.

**4.1.4 Bewering** *We zeggen dat de multistep voor een* $l \in \mathbf{N}$ exact is voor polynomen van graad $\leq l$ als een van de volgende vier equivalente beweringen geldt.

(a) $\delta_h(p)(t_n) = 0$ *voor alle*[8] $p \in \mathcal{P}_l$ *en* $h > 0$, $t_n \in h\mathbf{Z}$.

(b) $\delta_h(p)(0) = 0$ *voor alle* $p \in \mathcal{P}_l$ *en* $h > 0$.

(c) $\delta_h(\chi^j)(0) = 0$ *voor alle* $j = 0, \ldots, l$, $h > 0$.

(d) $\delta_1(\chi^j)(0) = 0$ *voor alle* $j = 0, \ldots, l$.

*Bewijs.* Het is duidelijk dat (a) $\Rightarrow$ (d). (d) $\Rightarrow$ (c) volgt uit het feit dat

$$\delta_h(\chi^j) = h^{j-1}\delta_1(\chi^j) \quad \text{voor iedere} \quad j \in \mathbf{N}_0. \tag{29}$$

Uit de lineairiteit van $v \to \delta_h(v)(t_n)$ volgt (c) $\Rightarrow$ (b).

(b) $\Rightarrow$ (a) volgt uit het feit dat $q(\zeta) := p(\zeta + t_n)$ een polynoom in $\mathcal{P}_l$ definiëert als $p$ dat is en dat $\delta_h(q)(0) = \delta_h(p)(t_n)$. □

De volgende stellingen laten zien dat de konsistentie uitspraak over $\mathbf{R}^d$–waardige funkties equivalent is met een eenvoudige uitspraak over $\mathbf{R}$–waardige polynomen en met een eenvoudige uitspraak over het $\rho$ en $\sigma$ polynoom.

**4.1.5 Stelling.** *Voor een* $l \in \mathbf{N}$ *zijn de volgende drie uitspraken equivalent.*

(a) *De multistep methode is konsistent van orde* $l$.

(b) *De multistep is exact voor polynomen van graad* $\leq l$.

(c) *Met* $\phi(\zeta) := \rho(\zeta) - \sigma(\zeta)\log\zeta$ *is* $\phi^{(j)}(1) = 0$ *voor* $j = 0, 1, \ldots, l$

*Bewijs.* (b) $\Leftarrow$ (a). Neem $d = 1$. Omdat $\delta_h(\chi^j) = h^{j-1}\delta_1(\chi^j)$ (zie (29)) is $\delta_h((\chi - t_n)^j)(t_n) = h^{j-1}\delta_1(\chi^j)(0) = \mathcal{O}(h^l)$ voor $j < l+1$ maar alleen als $\delta_1(\chi^j)(0) = 0$ en (b) volgt uit bewering 4.1.4.

We bewijzen (b) $\Rightarrow$ (a) voor $v \in C^{l+1}(\mathcal{J})$; door koördinaatsgewijs te werken volgt onmiddellijk de implikatie voor $\mathbf{R}^d$–waardige funkties. Beschouw

$$p(\zeta) := v(t_n) + (\zeta - t_n)v'(t_n) + \ldots + (\zeta - t_n)^l \frac{1}{l!} v^{(l)}(t_n) \qquad (\zeta \in \mathbf{C})$$

het $l$-de graads Taylor polynoom van $v$ rond $t_n$. Met $M := \max\{|v^{(l+1)}(\xi)| \,\big|\, \xi \in \mathcal{J}\}$ is

$$|v(t) - p(t)| \leq |t - t_n|^{l+1}\frac{M}{(l+1)!} \quad \text{en} \quad |v'(t) - p'(t)| \leq |t - t_n|^l \frac{M}{l!} \qquad (t \in [t_n, t_{n+k}])$$

(stel de Taylor reeks met restterm op rond $t_n$ voor $v - p$ en voor $v' - p'$; differentieer *niet* de Taylor reeks met restterm van $v - p$). Omdat $|t_{n+j} - t_n| = jh$ volgt

$$|\delta_h(v)(t_n)| \;=\; |\delta_h(v - p)(t_n)| \leq \frac{1}{h}\sum |\alpha_{k-j}| j^{l+1} h^{l+1}\frac{M}{(l+1)!} + \sum |\beta_{k-j}| j^l h^l \frac{M}{l!}$$

[7] Met $\chi$ geven we de funktie $\zeta \to \zeta$ $(\zeta \in \mathrm{C})$ aan.

[8] $\mathcal{P}_l$ is de ruimte van alle polynomen van graad $\leq l$

$$= \quad \bar{C}_{l+1} h^l M \quad \text{met} \quad \bar{C}_{l+1} := \sum |\alpha_{k-j}| \frac{j^{l+1}}{(l+1)!} + \sum |\beta_{k-j}| \frac{j^l}{l!}.$$

Merken we tenslotte op dat $\bar{C}$ niet afhangt van $h$ en $t_n$ dan zien we dat (a) geldt.
(a) $\Rightarrow$ (c). Neem voor $v$, $v(t) := e^{t-t_n}$. Dan

$$\delta_h(v)(t_n) = \delta_h(\exp)(0) = \frac{1}{h}\left(\rho(e^h) - h\sigma(e^h)\right) = \mathcal{O}(h^l).$$

Vervangen we $e^h$ door $t$ dan zien we dat deze laatste uitspraak equivalent is met

$$\phi(t) := \rho(t) - \sigma(t)\log t = \mathcal{O}((t-1)^{l+1}) \qquad (t \to 1)$$

en (c) volgt.

Uitschrijven leert dat $\delta_1(\chi^j)(0)$ een lineaire kombinatie is van $\phi^{(i)}(1) = 0$ met $i = 0, \ldots, j$ en toont (d) in 4.1.4 aan en dus (c) $\Rightarrow$ (b). We laten de details hiervan aan de lezer over. $\square$

**4.1.6 Opmerking.** Bovenstaande stelling zegt dat de konsistentie orde van de multistep hoger is naarmate de funkties $\rho/\sigma$ en log in het punt 1 beter op elkaar lijken: als $\sigma(1) \neq 0$ (de standaard situatie) dan is (c) van stelling 4.1.5 equivalent met

$$\left(\frac{\rho}{\sigma}\right)^{(j)}(1) = \log^{(j)} 1 \quad \text{voor} \quad j = 0, \ldots, l.$$

De volgende stelling is in feite een speciaal geval van stelling 4.1.5. We staan stil bij de implikatie (b) $\Rightarrow$ (a); de overige details van het bewijs laten we aan de lezer over.

**4.1.7 Stelling.** *De volgende drie uitspraken zijn equivalent.*
(a) *De multistep methode is konsistent.*
(b) *De multistep is exact voor polynomen van graad* $\leq 1$.
(c) $\rho(1) = 0$ *en* $\rho'(1) = \sigma(1)$.

*Bewijs.* We bewijzen (b) $\Rightarrow$ (a) voor $v \in C^1(\mathcal{J})$. Zij $\varepsilon > 0$.
Omdat $\mathcal{J}$ kompakt is en $v'$ kontinu is er een $h_0 > 0$ zodat $|v'(s) - v'(t)| \leq \varepsilon$ voor $s, t \in \mathcal{J}$, $|s - t| \leq h_0$. Beschouw nu een $h \in (0, h_0]$. Met $p(\zeta) := v(t_n) + (\zeta - t_n)v'(t_n)$ is, voor $t \in [t_n, t_{n+k}]$, $|v'(t) - p'(t)| \leq k\varepsilon$ en

$$|v(t) - p(t)| = |\int_{t_n}^{t} v'(s) - p'(s)|\, ds| \leq \int_{t_n}^{t_{n+k}} |v'(s) - p'(s)|\, ds \leq hk^2\varepsilon.$$

We vinden nu met $\tilde{C}_2 := \sum_{j \leq k}(|\alpha_{k-j}|k^2 + |\beta_{k-j}|k)$,

$$|\delta_h(v)(t_n)| = |\delta_h(v-p)(t_n)| \leq \frac{1}{h}\sum |\alpha_{k-j}|hk^2\varepsilon + \sum |\beta_{k-j}|k\varepsilon \leq \tilde{C}_2\varepsilon. \qquad \square$$

**4.1.8 Bewering.** *Stel dat de multistep konsistent is van orde l. Schrijf*

$$C_j := \frac{1}{j!}\delta_1(\chi^j)(0) \quad \text{en} \quad \bar{C}_j := \frac{1}{j!}\sum_{i=0}^{k} i^j(|\alpha_{k-i}| + \frac{j}{i}|\beta_{k-i}|) \quad \text{voor iedere } j \in \mathbf{N}.$$

*Stel dat* $v \in C^{m+2}(\mathcal{J}, \mathbf{R}^d)$. *Dan geldt met* $M := sup_{t\in\mathcal{J}}\|v^{(m+2)}(t)\|$ *dat*[9]

$$\delta_h(v)(t_n) = \sum_{j=l}^{m} h^j C_{j+1} v^{(j+1)}(t_n) + \bar{\delta}_h(v)(t_n) \quad \text{met} \quad \sup_{t_n \in \tilde{\mathcal{J}}_h} \|\bar{\delta}_h(v)(t_n)\| \leq \bar{C}_{m+2}h^{m+1}M.$$

*Bewijs.* Beschouw het $m+1$-ste graads Taylor polynoom van $v$ rond $t_n$. De bewering volgt nu met de argumenten in het bewijs van stelling 4.1.5 van de implikatie "(b) $\Rightarrow$ (a)". $\square$

[9] Als $l > m$ dan $\sum_{j=l}^{m} \ldots := 0$.

**4.1.9 Opmerking.** In 4.1.8 zou men misschien, voor $d = 1$ en $v \in C^{(l+2)}(\mathcal{J})$, de relatie $\delta_h(v)(t_n) = h^l C_{l+1} v^{(l+1)}(\xi_n)$ voor zekere $\xi_n \in [t_n, t_{n+k}]$, verwachten op grond van de ervaringen in 2.2. Deze relatie geldt echter alleen voor speciale schema's $(\rho, \sigma)$!

**4.1.10 Schaling.** Schalen van (M), het linker– en het rechterlid in (M) vermenigvuldigen met een skalair, heeft geen effekt heeft op de oplossing $u_h$. Om rekentechnische redenen is het handig om (M) zo te schalen dat $\alpha_0 = 1$. Om die redenen zal men overigens ook (M) met $h$ vermenigvuldigen. Als $\sigma(1) \neq 0$ (zie 4.1.34) dan zou het om theoretische redenen elegant zijn om (M) zo te schalen dat $\sigma(1) = 1$—de lokale diskretisatie fout hangt dan niet meer af van de toevallige schaling en is, in geval $\sigma = \beta_{k-j}\chi^j$, gelijk aan die in 2.2.1 (zie ook stelling 4.3.7).

Als de multistep konsistent van orde $l$ is en $C_{l+1}$ is als in bovenstaande bewering, dan noemt men $\frac{1}{\sigma(1)}C_{l+1}$ de *foutkonstante* van de multistep.

Voor de rest van dit diktaat maken we de volgende

**Aanname.** $$\alpha_0 = 1$$

**4.1.11 Opmerking.** Met behulp van de uitspraak (c) in 4.1.5 of met behulp van (b) in 4.1.5 en de equivalente uitspraak (d) in 4.1.4 kan men nu gemakkelijk multistep methoden konstrueren van iedere gewenste konsistentie orde.

**4.1.12 Gevolg.** *Kies $\kappa \in \{0, 1, \ldots, k\}$.*
*Stel dat het reële polynoom $\rho$, met $\rho(1) = 0$ van graad precies $k$, gegeven is.*
*Dan bestaat er precies een $k + 1 - \kappa$-tal $\beta_\kappa, \ldots, \beta_k$ zodat, met $\sigma = \beta_\kappa\chi^{k-\kappa} + \ldots + \beta_k$, de multistep met schema $(\rho, \sigma)$ konsistentie orde ten minste $k + 1 - \kappa$ heeft.*

*Bewijs.* Schrijf $\sigma = \sum_{j=0}^{k-\kappa} \gamma_j(\chi - 1)^j$. De eis $\phi^{(1)}(1) = 0$ in (c) van stelling 4.1.5 (konsistentie orde 1) legt $\gamma_0$ vast, etc.. Met $l = k + 1 - \kappa$ liggen de $\gamma_0, \ldots, \gamma_{k-\kappa}$ en dus de $\beta_\kappa, \ldots, \beta_k$ vast. $\square$

**4.1.13 Opgave.** Formuleer en bewijs een analoog resultaat voor het geval het polynoom $\sigma$ van graad $k$ met $\sigma(1) \neq 0$ gegeven is en $\rho$ gekonstrueerd moet worden.

**4.1.14 Schema's van Adams**
In de *schema's van Adams* is $\rho = \chi^k - \chi^{k-1}$ en is $\sigma$ als voorgeschreven in bovenstaand gevolg met òf $\kappa = 1$ (expliciete Adams) òf $\kappa = 0$ (impliciete Adams).
De expliciete 1-stap Adams methode is precies de Euler forward methode en de impliciete 1-staps Adams methode is de trapezium regel. $(\chi^2 - \chi, \frac{3}{2}\chi - \frac{1}{2})$ is het schema voor de expliciete 2-stap Adams methode. Deze methode wordt ook wel *Adams–Bashforth* genoemd.
$(\chi^2 - \chi, \frac{1}{12}[5\chi^2 + 8\chi - 1])$ karakterizeert de impliciete 2-stap Adams methode. Deze heet ook wel *Adams–Moulton*. etc.

**4.1.15 Schema's van Milne**
In de *schema's van Milne* is $\rho = \chi^k - \chi^{k-2}$ en is $\sigma$ als voorgeschreven in bovenstaand gevolg.
De expliciete 2-stap Milne methode is precies de midpunt regel en de impliciete 2-staps is de regel van Simpson.

**4.1.16 Achterwaartse differentie schema's.**
In *achterwaartse differentie schema's* is $\sigma = \chi^k$ en is $\rho$ zo gekozen dat de konsistentie orde maximaal is. In dat geval is $\frac{1}{h}\rho(T_h)(v)(t_n)$ dus een eindige differentie benadering van $v'(t_{n+k})$ van orde $k$. Deze schema's, ook wel BDF-methoden (backward difference formulas) genoemd, zijn voor de praktijk van bijzonder belang voor $k = 1, 2, \ldots, 6$ (voor $k > 6$ zijn de schema's instabiel—in de volgende subparagraaf vertellen we wat een stabiel schema is).

Het resultaat in de volgende bewering laat zien dat we met een niet konsistente multistep methode zelfs de eenvoudigste differentiaalvergelijkingen niet redelijk benaderend kunnen oplossen: iedere acceptabele multistep moet dus konsistent zijn. We laten het bewijs van de bewering over aan de lezer.

**4.1.17 Bewering.** *Beschouw voor* $\alpha, \beta \in \mathbf{R}$ *het probleem*

$$\begin{cases} u'(t) = \beta & \text{voor} \quad t \in [0, T] \\ u(0) = \alpha \end{cases} .$$

*Voor* $h > 0$ *voldoet* $u_h \in C(\mathcal{J}_h)$ *aan* (M) *en is* $u_h(t_j) = u^*(t_j) = \alpha + \beta jh$ *voor* $j < k$.

*De volgende twee uitspraken zijn equivalent.*

(a) *De multistep is konsistent.*

(b) *Voor zowel* $\beta = 0$, $\alpha = 1$ *en* $\beta = 1$, $\alpha = 0$ *geldt* $\lim_{h\to 0} \sup_{t_n \in \mathcal{J}_h} |u_h(t_n) - u^*(t_n)| = 0$. □

Een multistep is blijkbaar alleen konsistent als voor de twee meest banale differentiaalvergelijkingen de benaderende oplossingen konvergeren.

**4.1.18 Opgave.** Bewijs bewering 4.1.17.

### Stabiliteit

**4.1.19 Stabiliteit.** De multistep heet *stabiel* als

$$\text{óf} \quad |\lambda| < 1 \quad \text{óf} \quad |\lambda| = 1 \text{ en } \rho'(\lambda) \neq 0 \qquad \text{voor iedere } \lambda \in \mathbf{C} \text{ met } \rho(\lambda) = 0$$

($\rho$ voldoet aan het *wortel kriterium*: iedere wortel van het $\rho$-polynoom ligt binnen de eenheidsschijf en is enkelvoudig als hij op de rand ervan ligt).

De schema's van Adams en Milnes geven stabiele multistep methoden. Voor $k \leq 6$ zijn de BDF-methoden ook stabiel, voor $k > 6$ zijn ze instabiel (dit laatste is niet zonder meer duidelijk: we bewijzen dat hier echter niet).

Onderstaande bewering generalizeert voorbeeld 2.2.4. We laten het bewijs ervan aan de lezer over (zie 3.1.10).

**4.1.20 Bewering.** *Beschouw het volgende probleem*

$$\begin{cases} u'(t) = 0 & \text{voor iedere} \quad t \in [0, T] \\ u(0) = 0. \end{cases}$$

*Voor* $h > 0$ *voldoet* $u_h \in C(\mathcal{J}_h)$ *aan* (M).

*De volgende twee uitspraken zijn equivalent.*

(a) *Er is een* $K > 0$ *zodat voor iedere* $h > 0$ *geldt*

$$\sup_{t_n \in \mathcal{J}_h} |u_h(t_n) - u^*(t_n)| = \sup_{t_n} |u_h(t_n)| \leq K \max_{j=0,\ldots,k-1} |u_h(t_j)|.$$

*De kleinste* $K$ *die hieraan voldoet noemen we de* stabiliteits konstante *van de multistep.*

(b) *De multistep is stabiel.* □

**4.1.21 Opgave.** Bewijs bovenstaande bewering.

**4.1.22 Opmerking.** De stabiliteits definitie sluit niet (expliciet) aan bij onze heuristische stabiliteits opvatting die verlangt dat (evaluatie)fouten het uiteidelijk resultaat niet domineren. In bewering 4.1.20 hebben we gezien dat, met betrekking tot de triviale differentiaalvergelijkingen, het gedefinieerde stabiliteitsbeprip wel spoort met onze intuïtie. De bewering vertelt dat startfouten ten hoogste met een faktor $K$ worden "opgeblazen". In de onderstaande stabiliteits stelling 4.1.26 zullen we zien dat in een stabiele multistep lokale fouten inderdaad niet onbeperkt kunnen groeien.

Het stabiliteits kriterium in de definitie heeft als voordeel dat het zeer eenvoudig te kontroleren is. Bovendien kunnen we, gezien het resultaat in 4.1.20, stellen dat iedere acceptabele multistep eraan moet voldoen.

**4.1.23 Multistep iteratie.** Voor $h > 0$, $v_0, \ldots, v_{k-1} \in \mathbf{R}^d$ en $(\epsilon_n)$ een rij in $\mathbf{R}^d$ zijn we geïnteresseerd in de $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ waarvoor geldt, met $u_n = u_h(t_n)$,

$$\begin{cases} u_{n+k} + \alpha_1 u_{n+k-1} + \ldots + \alpha_k u_n = h(\beta_0 f(t_{n+k}, u_{n+k}) + \ldots + \beta_k f(t_n, u_n)) + \epsilon_n \\ \qquad\qquad\qquad \text{voor} \quad n = 0, 1, \ldots \quad \text{met} \quad (n+k)h \leq T. & (\tilde{\mathrm{M}}) \\ u_j = v_j \quad \text{voor} \quad j = 0, \ldots, k-1 & (\mathrm{S}) \end{cases}$$

In de praktijk zullen we de funktiewaarden $u_n = u_h(t_n)$ berekenen door $(\tilde{\mathrm{M}})$ iteratief te gebruiken startend met de $u_0, \ldots, u_{k-1}$ die gegeven zijn in (S). De grafiek van $u_h$ moet dan wel binnen $\Omega$ liggen en als $\beta_0 \neq 0$ moet $u_{n+k}$ op te lossen zijn uit de vergelijking in $(\tilde{\mathrm{M}})$.

$v_0, \ldots, v_{k-1}$ zijn de *startwaarden* en $u^*(t_0) - v_1, \ldots, u^*(t_{k-1}) - v_{k-1}$ de *startfouten.*
$\epsilon_n$ representeert de *lokale fout*, de fout (evaluatie fout of diskretisatie fout of beide) die in de $n$-de iteratie stap gemaakt wordt.

Onder zekere voorwaarden kunnen we bewijzen dat $(\tilde{\mathrm{M}})$ & (S) een oplossing heeft en dat het effekt van iedere fout in het rekenproces binnen de perken blijft.
We formuleren eerst die voorwaarden. Andere resultaten kunnen we onder soortgelijke voorwaarden bewijzen. Voor de overzichtelijkheid sommen we hieronder alvast al die voorwaarden op. Een drietal voorwaarden heeft betrekking alleen op het kontinue probleem, twee alleen op het schema $(\rho, \sigma)$ van de multistep en drie op de multistep iteratie, met name op de start ervan. We voeren ook wat notatie in die we verder in dit hoofdstuk zonder verdere referentie zullen gebruiken.

**4.1.24 Voorwaarden en notaties.**
Schrijf $\tilde{\sigma} := \sum_{j=0}^{k} |\beta_j|$. Zij $\tilde{h} > 0$, en $l, m \in \mathbf{N}$.

**De differentiaalvergelijking.**
(Dif.0) Voor een zekere $\tilde{r} > 0$ is de $\tilde{r}$-buis $\{(t,x) \,\big|\, \|x - u^*(t)\| \leq \tilde{r}, t \in \mathcal{J}\}$ rond de grafiek van $u^*$ binnen $\Omega$ (als in 1.1.7).
(Dif.1) $f$ is Lipschitz kontinu op $\Omega$ in de tweede variabele met Lipschitz konstant $L$
(als in 1.1.4). $\tilde{h}$ is zo dat $\tilde{h}L|\beta_0| < 1$.
(Dif.2) $f \in C^{l+1}(\Omega, \mathbf{R}^d)$.
(Dif.3) $f \in C^{l+m}(\Omega, \mathbf{R}^d)$.

**De multistep.**
(Mul.1) De multistep is stabiel met stabiliteitskonstante $K$.
(Mul.2) De multistep is konsistent van orde $l$.

**De multistep iteratie.**
Zij $\mathbf{H}$ een deelverzameling van $(0, \infty)$ waar 0 een verdichtingspunt van is
(bv. $\mathbf{H} = \{\frac{1}{n} \,\big|\, n \in \mathbf{N}\}$): $h$ doorloopt deze verzameling $\mathbf{H}$.
Voor iedere $h \in \mathbf{H}$, laat $v_{h\,0}, \ldots, v_{h\,k-1}$ in $\mathbf{R}^d$ de startwaarden zijn.
(Start.1) $\max_{j<k} \|u^*(t_j) - v_{h\,j}\| = \mathcal{O}(h^l)$ voor $h \to 0$.
(Start.2) $\max_{j<k} \|u^*(t_j) - v_{h\,j}\| = \mathcal{O}(h^{l+1})$ voor $h \to 0$.

(Start.3) Er zijn $c_{ij} \in \mathbf{R}^d$ zodat voor iedere $j < k$ geldt

$$v_{h\,j} = u^*(t_j) + \sum_{i=1}^{m-1} h^{l+i} c_{ij} + \mathcal{O}(h^{l+m}) \qquad \text{voor} \qquad h \to 0.$$

In geval (Dif.1) en (Mul.1) gelden schrijven we

$$K' := qKe^{TL\tilde{\sigma}qK} \qquad \text{met} \qquad q = \frac{1}{1 - \tilde{h}L|\beta_0|}.$$

Met behulp van het kontraktie lemma zien we gemakkelijk het volgende.

**4.1.25 Bewering.** *Als* (Dif.1) *geldt en* $h \in (0, \tilde{h}]$ *dan heeft* ($\tilde{\mathrm{M}}$) & (S) *hoogstens een oplossing* (in $C(\mathcal{J}_h, \mathbf{R}^d)$ met grafiek binnen $\Omega$). □

**4.1.26 Stabiliteitsstelling.** *Stel* (Dif.1) *en* (Mul.1) *gelden. Zij* $h \in (0, \tilde{h}]$ *en* $r > 0$. *Stel dat* ($\tilde{\mathrm{M}}$) & (S) *een oplossing* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *heeft met r-buis* $\{(t_n, x) \,\big|\, \|x - u_h(t_n)\| \le r,\, t_n \in \mathcal{J}_h\}$ *om de grafiek binnen* $\Omega$.
*Laat* $\tilde{v}_0, \ldots, \tilde{v}_{k-1}$ *in* $\mathbf{R}^d$ *zijn en laat* $(\tilde{\epsilon}_n)$ *een rij in* $\mathbf{R}^d$ *zijn.*
($\tilde{\mathrm{M}}$) & (S), *waarbij* $\epsilon_n$ *door* $\tilde{\epsilon}_n$ *en* $v_j$ *door* $\tilde{v}_j$ *vervangen is, heeft precies een oplossing* $\tilde{u}_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *en, met* $K_h := \frac{K}{1 - hL|\beta_0|}$, *geldt dat*

$$\|u_h(t_n) - \tilde{u}_h(t_n)\| \le K_h \left( \max_{j<k} \|v_j - \tilde{v}_j\| + \sum_{j=0}^{n-k} \|\epsilon_j - \tilde{\epsilon}_j\| \right) e^{nhL\tilde{\sigma}K_h} \qquad \textit{alle} \qquad n,\, nh \le T$$

*als de uitdrukking in het rechterlid voor iedere* $n,\, nh \le T$ *kleiner is dan* $r$.

*Bewijs.* Voor het schrijfgemak nemen we aan dat $d = 1$: het bewijs voor het geval $d > 1$ laten we over aan de lezer.

Definiëer voor iedere $n,\, nh \le T$ de (niet lineaire) afbeelding

$$d_n : \mathbf{B}_{\tilde{r}} \to \mathbf{R} \quad \text{door} \quad d_n(x) := f(t_n, u_n) - f(t_n, u_n + x) \qquad x \in \mathbf{B}_{\tilde{r}}.$$

Dan is $|d_n(x) - d_n(y)| \le L|x - y|$ en $d_n(0) = 0$.

Het verschil tussen de twee betrekkingen ($\tilde{\mathrm{M}}$) leert ons dat, met $\mu_n := \epsilon_n - \tilde{\epsilon}_n$, de *globale fouten* $e_n := u_h(t_n) - \tilde{u}_h(t_n)$ aan de volgende rekursieve relatie voldoen

$$\begin{cases} e_{n+k} + \ldots + \alpha_k e_n = h(\beta_0 d_{n+k}(e_{n+k}) + \ldots + \beta_k d_n(e_n)) - \mu_n & \text{als} \quad (n+k)h \le T, \\ e_j = v_j - \tilde{v}_j \quad \text{voor} \quad j = 0, \ldots, k-1. \end{cases}$$

Stellen we de "companion vorm" op, dan vinden we

$$(I + hE_n)\vec{e}_{n+1} = (A + hD_n)\vec{e}_n + \vec{\mu}_n,$$

waarbij $\vec{e}_n = (e_n, \ldots, e_{n+k-1})^T \in \mathbf{R}^k$, $A \in \mathbf{M}_k$ de companion matrix van de "$h = 0$–rekursie" is, $\vec{\mu}_n \in \mathbf{R}^k$ zodat $\|\vec{\mu}_n\|_\infty \le |\mu_n|$, en $E_n$, $D_n$ niet–lineaire afbeeldingen zijn die de $\beta_{k-j} d_{n+j}$–termen representeren. Er geldt $\sup_n \|A^n\|_\infty \le K$ (waarom?).
Verder is $\|E_n(\vec{x}) - E_n(\vec{y})\|_\infty \le |\beta_0| L \|\vec{x} - \vec{y}\|_\infty$, $\|D_n(\vec{x}) - D_n(\vec{y})\|_\infty \le \sum_{j=1}^k |\beta_j| L \|\vec{x} - \vec{y}\|_\infty$.
Merk op dat de oplossing $\vec{e}_n$ van de companion vorm onmiddellijk de oplossing $\tilde{u}_n$ geeft.
De stelling van Kreiss vertelt ons dat de $\vec{e}_n$ bestaan en geeft een schatting voor $\|\vec{e}_n\|_\infty$. De existentie van $\tilde{u}_h(t_n)$ en onze schatting volgt hieruit. □

**4.1.27 Diskussie.** Berekenen we benaderend $u^*$ door vanaf de start $t_0$ (M) iteratief te gebruiken dan maken we startfouten (we zullen $u_0^*, \ldots u_{k-1}^*$ of een benadering ervoor moeten hebben) en in iedere iteratie stap lokale fouten (evaluatie fouten en diskretisatie fouten). Iedere startfout en iedere lokale fout $\epsilon_j$ levert een bijdrage tot de globale fout $\tilde{u}_n - u_n$. Als $h \in (0, \tilde{h}]$ dan is, volgens bovenstaand resultaat, deze bijdrage ten hoogste $K'$ maal groter dan die lokale fout, een konstante is die verder niet meer van $h$ afhangt. Misschien is $K'$ wel groot, maar door lokaal nauwkeurig te werken blijft het effekt van een lokale fouten toch binnen de perken. Hoewel we met kleinere $h$ meer stappen moeten zetten om het eind van het interval $\mathcal{J}$ te bereiken wordt het effekt van 'n lokale fout daardoor niet vervelender. Omdat de bijdragen van de lokale fouten kumuleren in de globale fouten zal de totale fout toch groeien bij afnemende $h$, tenzij bij afnemende $h$ de grootte van de lokale fouten evenzo afneemt (hetgeen met betrekking tot de lokale diskretisatie fout het geval is).

**4.1.28 Opmerking.** In geval de multistep stabiel is "groeien" de lokale fouten niet onbeperkt. Als de oplossing van de differentiaalvergelijking daalt kunnen de fouten toch nog de exacte oplossing volledig overvleugelen. In zo'n situatie willen we dat het effekt van iedere lokale fouten kleiner wordt bij het iteratief doorlopen van (M): we hebben dan behoefte aan een ander stabiliteitsbegrip. We komen hierop uitgebreid terug in volgende paragrafen.

Stelling 4.1.5 is de diskrete variant van het perturbatie resultaat voor het kontinue probleem in 1.2.8. Voor een groot aantal problemen was de schatting in 1.2.8 veel te grof (zie 1.2.14). In zo'n geval hebben we ook hier behoefte aan een beter resultaat.

### Konvergentie

We willen dat de multistep de ingewikkeldere vektoriële problemen, die we in stelling 1.1.4 besproken hebben, goed benaderend oplost.

**4.1.29 Konvergentie.**
We zeggen de multistep *konvergent is met betrekking tot het probleem in* 1.1.1 als
(i) er een $h_0 > 0$ is en een $\bar{\nu} > 0$ zodat er, voor iedere $h \in (0, h_0]$ en $v_0, \ldots, v_{k-1}$ in $\mathbf{R}^d$ met $\|v_j - u_0\| \leq \bar{\nu}$ $(j < k)$, een oplossing $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ van (M) & (S) is, en
(ii) als $h$ een naar 0 dalende rij doorloopt en voor iedere $h$ uit die rij voldoet $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ aan (M) dan konvergeert $(u_h)$ naar $u^*$ als $\max_{j<k} \|u_h(t_j) - u^*(t_j)\| \to 0$ voor $h \to 0$.

(voor voldoend kleine $h$ en voldoende nauwkeurige start moet de benaderende oplossing bestaan en de rij benaderende oplossingen moet konvergeren als de *start konsistent* is).

De multistep is *konvergent van orde $l$ met betrekking tot het onderhavige probleem* als ook
(ii$'$) $(u_h)$ met orde $l$ naar $u^*$ konvergeert in geval
$\max_{j=0,\ldots,k-1} \|u_h(t_j) - u^*(t_j)\| = \mathcal{O}(h^l)$ voor $h \to 0$.

De multistep is *konvergent* als hij konvergent is met betrekking tot alle problemen in 1.1.4 en is *konvergent van orde $l$* als hij dat is met betrekking tot alle problemen in 1.1.4 waarvan de oplossing $u^*$ voldoende glad is.

De volgende stelling "**stabiliteit + konsistentie = konvergentie**" vertelt dat uit de eenvoudige voorwaarden "stabiliteit" en "konsistentie" het bijzonder krachtig resultaat "konvergentie" volgt. De uitspraak in de stelling is kwalitatief en volgt onmiddellijk uit diens kwantitatieve variant in 4.1.31 en uit de beweringen 4.1.17 en 4.1.20.

**4.1.30 Hoofdstelling.** *Stel* (Dif.0) *en* (Dif.1) *gelden. Dan:*
(a) *de multistep is konvergent dan slechts dan als hij stabiel en konsistent is;*

(b) *voor een $l \in \mathbf{N}$ is de multistep konvergent van orde $l$ dan en slechts dan als hij stabiel is en konsistent van orde $l$.* $\square$

**4.1.31 Konvergentie stelling.** *Stel* (Dif.0), (Dif.1) *en* (Mul.1) *gelden.*
*Laat* $h \in (0, \tilde{h}]$, $v_0, \ldots, v_{k-1} \in \mathbf{R}^d$ *en* $(\epsilon_n)$ *een rij in* $\mathbf{R}^d$ *zijn.*
*Er is een* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *met grafiek in* $\Omega$, *die aan* $(\tilde{\mathrm{M}})$ & (S) *voldoet en, met* $K_h := \frac{K}{1-hL|\beta_0|}$, *geldt voor alle* $t_n \in \mathcal{J}_h$, *dat*

$$\|u^*(t_n) - u_h(t_n)\| \leq K_h \left( \max_{j<k} \|u^*(t_j) - v_j\| + \sum_{j=0}^{n-k} \|\epsilon_j\| + h \sum_{j=0}^{n-k} \|\delta_h(u^*)(t_j)\| \right) e^{(t_n - t_0) L \tilde{\sigma} K_h}$$

*als de majorant in het rechterlid voor iedere* $n$, $nh \leq T$ *kleiner is dan* $\tilde{r}$.

*Bewijs.* Pas de stabiliteits stelling 4.1.26 toe met $\epsilon_n = h\delta_h(u^*)(t_n)$, $v_j = u^*(t_j)$, $\tilde{\epsilon}_n = \epsilon_n$ en $\tilde{v}_j = u_h(t_j)$ $(j < k,\, nh \leq T)$. $\square$

**4.1.32 Opmerking.** Naast de lokale diskretisatie fouten maken we bij het iteratief doorlopen van (M) ook nog *lokaal evaluatie fouten* $\epsilon_n$ en startfouten $u^*(t_j) - v_j$ $(j < k)$.

Stel dat (Mul.2) geldt, $u^* \in C^{l+1}(\mathcal{J}, \mathbf{R}^d)$ en dat we $M := sup_{t \in \mathcal{J}} \|u^{*(l+1)}(t)\|$ kunnen schatten. Dan (zie 4.1.8)

$$\sup_{n, nh \leq T} h \sum_{j=0}^{n-k} \|\delta_h(u^*)(t_j)\| \leq T \sup_{t_j \in \tilde{\mathcal{J}}_h} \|\delta_h(u^*)(t_j)\| \leq T \bar{C}_{l+1} h^l M.$$

Als (Dif.0-1), (Mul.1) ook nog gelden en $u_h$ bestaat, dan is dus, met $K'$ uit 4.1.24,

$$\|u^*(t_n) - u_h(t_n)\| \leq K' \left( \max_{j<k} \|u^*(t_j) - v_j\| + \sum_{j=0}^{n-k} \|\epsilon_j\| + T\bar{C}_{l+1} M h^l \right). \tag{30}$$

Willen we $u^*$ met een zekere precisie benaderend berekenen, dan kunnen we met deze schatting vaststellen met welke stapgrootte $h$ en met welke startnauwkeurigheid we die precisie kunnen bereiken (à priori foutschatting). De startwaarden moeten op zijn minst dezelfde orde van nauwkeurigheid hebben als de konsistentie orde van de multistep ((Start.1) moet gelden, anders worden de globale fouten in essentie bepaald door de startfouten en zouden we net zo goed met een multistep met lagere orde konsistentie kunnen werken). In de praktijk zijn de evaluatie fouten $\epsilon_n$ vaak beduidend kleiner dan de bijdrage $h\delta_h(u^*)(t_n)$ van de diskretisatie fouten en van geen belang in de foutschatting.

**4.1.33 Opmerking.** Als de multistep impliciet is, is bij het iteratief doorlopen van (M) $u_{n+k}$ telkens impliciet gedefiniëerd en moeten we $u_{n+k}$ oplossen.

Herschrijven van het resultaat in 3.1.13 leert ons dat we dat als volgt zouden kunnen doen. Met

$$y_n := \sum_{j=0}^{k-1} (h\beta_{k-j} f(t_{n+j}, u_{n+j}) - \alpha_{k-j} u_{n+j}) \quad \text{en} \quad \Phi(x) := h\beta_0 f(t_{n+k}, x) + y_n \tag{31}$$

(voor $x$ in de buurt van $u_{n+k}$) bestaat de rij $(x^{(i)})_i$ met $x^{(0)} = u_{n+k-1}$ en

$$x^{(i+1)} = \Phi(x^{(i)}) \qquad (i \in \mathbf{N}_0)$$

en konvergeert naar $u_{n+k}$ als $h < \tilde{h}$: er geldt (zie 3.1.13)

$$\|x^{(i)} - u_{n+k}\| \leq (h|\beta_0|L)^i \|x^{(0)} - u_{n+k}\| \quad \text{voor alle} \quad i \in \mathbf{N}.$$

Is $f$ voldoende vaak differentieerbaar dan zouden we uiteraard ook kunnen overwegen $u_{n+k}$ uit $u_n, \ldots, u_{n+k-1}$ te berekenen met bijvoorbeeld het Newton–Raphson proces. Met een goede startwaarde $x^{(0)}$ hoeven we echter maar een paar "goedkope" succesieve substitutie stappen te zetten en is het niet nodig het "dure" Newton–Raphson proces te gebruiken (in het Newton–Raphson proces moeten ook nog afgeleiden berekend worden! Zie 4.4.7)).

Berekenen we $u_{n+k}$ middels een iteratief proces dan moeten we het proces na eindig veel iteraties stoppen. De fout die hierdoor onstaat kunnen we opnemen in de $\epsilon_n$-term.

**4.1.34 Opmerking.** Als de multistep stabiel en konsistent is dan is $\sigma(1) \neq 0$ $\quad(\rho(1) = 0$ en $\rho'(1) = \sigma(1)$ vanwege de konsistentie (zie 4.1.7) en $\rho'(1) \neq 0$ vanwege de stabiliteit (zie 4.1.19)). In dat geval is $\rho(t) - \sigma(t)\log t = \sigma(t)(\frac{\rho}{\sigma}(t) - \log t) = \mathcal{O}((1-t)^{l+1})$ $\quad (t \to 1)$ precies dan als $\frac{\rho}{\sigma}(t) - \log t = \mathcal{O}((t-1)^{l+1})$ $\quad (t \to 1)$. Blijkbaar is voor een stabiele multistep (c) in stelling 4.1.5 equivalent met

(c') $\quad \rho(1) = 0 \quad$ en $\quad (\frac{\rho}{\sigma} - \log)^{(j)}(1) = 0 \quad$ voor $\quad j = 1, \ldots, l$.

Zuiver polynomiale eigenschappen van het $\rho$ en $\sigma$ polynoom bepalen of een multistep stabiel en konsistent van orde $l$ is en daarmee konvergent van orde $l$. Andere eigenschappen van de multistep methoden zijn evenzo equivalent met zekere polynomiale eigenschappen van het $\rho$ en $\sigma$ polynoom. De vraag bijvoorbeeld wat de maximale konsistentie orde is die een $k$-staps methode kan hebben laat zich vertalen naar de vraag naar de maximale orde waarmee $\frac{\rho}{\sigma}$ de funktie log kan benaderen in $t = 1$ als $\rho, \sigma \in \mathcal{P}_l$ en $\rho$ aan het wortel ktiterium voldoet. Zonder bewijs vermelden het volgende resultaat van Dahlquist.

**De eerste barrière stelling.** *Als de $k$-staps multistep stabiel is en $l$ is de orde dan is*
$l \leq k+1$ *als $k$ oneven is,* $\quad l \leq k+2$ *als $k$ even is en* $\quad l \leq k$ *als* $\beta_0 \leq 0$.
*Als* $\quad l = k+2$ *dan is* $\quad \alpha_j = -\alpha_{k-j}$ *en* $\quad \beta_j = \beta_{k-j}$ *voor* $\quad j = 0, \ldots, k$. □

**4.1.35 ∘ Cyclische multistep methoden.** De barrière in de stelling kan "doorbroken" worden door verschillende multistep methoden cyclisch te gebruiken. We geven hier een nauwkeurige beschrijving van zo'n cyclische methode en we vermelden (zonder bewijs) wat resultaten.

Zij $m \in \mathbf{N}$. Voor $\quad j = 0, \ldots, m-1 \quad$ hebben we een $k$-staps multistep met schema $(\rho_j, \sigma_j)$. De *$m$-cyclische $k$-staps multistep methode* met deze schema's berekent, voor stapgrootte $h > 0$, de numerieke oplossing $\quad u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$, $\quad$ met $\quad f_h(t_n) := f(t_n, u_h(t_n))$, $\quad$ als volgt

$$\rho_j(T_h)(u_h)(t_n) = h\sigma_j(T_h)(f_h)(t_n) \quad \text{met} \quad j = n \bmod m \quad \text{voor alle} \quad t_n \in \tilde{\mathcal{J}}_h.$$

De $m$-cyclische multistep is *stabiel* als er een $K > 0$ bestaat zodat voor iedere $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ geldt

$$\sup_{t_n \in \mathcal{J}_h} \|u_h(t_n)\| \leq K \max_{j<k} \|U_h(t_j)\| \quad \text{als} \quad \rho_j(T_h)(u_h)(t_n) = 0 \quad j = n \bmod m \quad (t_n \in \mathcal{J}_h).$$

Stabiliteit en konsistentie impliceert weer konvergentie.

**Stelling.** *Als de cyclische multistep stabiel is, konsistentie orde $l$ heeft (d.w.z iedere afzonderlijke multistep is konsistent van orde $l$) en de start orde $l$ is, dan konvergeert de multistep met orde $l$.* □

Men kan stabiele cyclische multistep methoden konstrueren waarvan de afzonderlijke multisteps niet stabiel zijn. De extra vrijheid (de verzwakking van de stabiliteitseis) kan men gebruiken om de konsistentie orde te verhogen van ieder van de afzonderelijke $k$-staps methoden.

**Stelling.** *Zij* $\quad k \in \mathbf{N}$.
(a) *Er is een stabiele $k-1$-cyclische $k$-staps multistep methode met konsistentie orde $2k-1$.*
(b) *Er is een stabiele $k$-cyclische $k$-staps multistep methode met konsistentie orde $2k$.* □

## 4.2 Start procedures voor multistep methoden

Zoals we al opmerkten in 4.1.32 levert een stabiele en konsistente multistep methode een goede benadering van $u^*$ door met voldoend kleine $h$ de multistep rekursie iteratief te doorlopen; de start moet dan wel voldoende nauwkeurig zijn. In deze pargraaf staan we even stil bij de vraag hoe we aan 'n nauwkeurige start kunnen komen.

We merken allereerst op dat we slechts één keer starten. De hoeveelheid werk die we in een goede start moeten steken zal verwaarloosbaar zijn ten opzichte van het werk dat gaat zitten in het iteratief doorlopen van de multistep rekursie. (Willen we na een aantal iteratie stappen de grootte van de stapgrootte herzien—in bijvoorbeeld een automatische stapgrootte besturings procedure—dan kunnen we in de $t_j$ waar we gearriveerd zijn opnieuw starten. Gewoonlijk zal het aantal iteratie stappen dat we dan al gemaakt hebben relatief groot zijn: ook dan is de hoeveelheid werk om te starten klein ten opzichte van het overige werk.)

**4.2.1 Taylor reeks.** Met behulp van een Taylor reeks benadering kan men, als $f$ voldoende glad is, $u^*(t_j)$ voor $j < k$, nauwkeurig benaderen;
met $p(t) := u^*(t_0) + (t - t_0)u^{*(1)}(t_0) + \ldots + (t - t_0)^m \frac{1}{m!} u^{*(m)}(t_0)$ is
$v_j := p(t_j) = u^*(t_j) + \mathcal{O}(h^{m+1})$ $(h \to 0)$ voor $j < k$. Bovendien kennen we de beginwaarde $u^*(t_0) = u_0$ en kunnen we, met behulp van de differentiaalvergelijking en de beginwaarde, $u^{*(1)}(t_0) = f(t_0, u_0)$, $u^{*(2)}(t_0) = \frac{\partial}{\partial t} f(t_0, u_0) + f(t_0, u_0) \frac{\partial}{\partial x} f(t_0, u_0), \ldots$ berekenen. Hiermee is $v_j$ berekenbaar.

Bepalen we $v_1$ als boven met het eerste graads Taylor polynoom dan hebben we in feite een stap gezet van het Euler forward proces.

Het bepalen van de diverse partiële afgeleiden van $f$ is niet altijd even praktisch. We geven ook nog een andere methode die deze berekening vermijdt.

**4.2.2 Multistep.** $v_0 = u_0$ is gegeven. Om de overige startwaarden $v_1, \ldots, v_{k-1}$ met nauwkeurigheid $\mathcal{O}(h^{m+1})$ te bepalen kunnen we een $k-1$–tal "multistep methoden" van orde $m$ bedenken waarmee we de $v_j$ $(j < k)$ als volgt met $m$ slagen van een succesive substitutie proces zouden kunnen bepalen.

Stel $(\rho_1, \sigma_1), \ldots, (\rho_{k-1}, \sigma_{k-1})$ zijn paren polynomen van graad $\leq k-1$ met de volgende eigenschappen:
- Als $\rho_j(\zeta) = \sum_{i=0}^{k-1} \alpha_{j\,k-1-i} \zeta^i$ en $A = (a_{ij}) \in \mathbf{M}_{k-1}(\mathbf{R})$ met $a_{ij} = \alpha_{i\,j-1}$ dan is $A$ niet-singulier;
- $\rho_j(T_h)(p)(0) - h\sigma_j(T_h)(\frac{d}{dt}p)(0) = 0$ voor alle $p \in \mathcal{P}_m$ en alle $j = 1, \ldots, k-1$

(de "multisteps" zijn konsistent van orde $m$: hoewel we hier niet met echte multisteps te maken zoals die gedefiniëerd zijn in 4.1.1—de $\alpha_{j\,0}$ hoeven niet $\neq 0$ te zijn—geldt stelling 4.1.7 ook hier).

Als (Dif.1) geldt en $h$ klein genoeg is dan kunnen we, met $\vec{u}_0 := (\alpha_{1\,k-1}u_0, \ldots, \alpha_{k-1\,k-1}u_0)^T$, de vektor $\vec{v} := (v_{k-1}, \ldots, v_1)^T$ oplossen uit de vergelijking

$$A\vec{v} + \vec{u}_0 = \begin{bmatrix} \alpha_{1\,0}v_{k-1} & + & \ldots & +\alpha_{1\,k-2}v_1 & +\alpha_{1\,k-1}u_0 \\ \vdots & & & & \\ \alpha_{k-1\,0}v_{k-1} & + & \ldots & +\alpha_{k-1\,k-2}v_1 & +\alpha_{k-1\,k-1}u_0 \end{bmatrix} =$$

$$\Phi(\vec{v}\,) := h \begin{bmatrix} \beta_{1\,0}f(t_{k-1}, v_{k-1}) & + & \ldots & +\beta_{1\,k-2}f(t_1, v_1) & +\beta_{1\,k-1}f(t_0, u_0) \\ \vdots & & & & \\ \beta_{k-1\,0}f(t_{k-1}, v_{k-1}) & + & \ldots & +\beta_{k-1\,k-2}f(t_1, v_1) & +\beta_{k-1\,k-1}f(t_0, u_0) \end{bmatrix};$$

voor $h$ klein genoeg is $A^{-1}\Phi$ immers een kontraktie. Met $\vec{v}^{\,(0)} = (u_0, \ldots, u_0)^T$ en $A\vec{v}^{\,(i)} + \vec{u}_0 = \Phi(\vec{v}^{\,(i-1)})$ konvergeert $(\vec{v}^{\,(i)})_i$ naar $\vec{v}$ (zie het kontraktie lemma).

Met $\vec{u}^{\,*} := (u^*(t_{k-1}), \ldots, u^*(t_1))^T$, geldt dat $A\vec{u}^{\,*} + \vec{u}_0 = \Phi(\vec{u}^{\,*}) + \mathcal{O}(h^{m+1})$ Voor $h$ klein genoeg volgt hieruit dat $\|\vec{u}^{\,*} - \vec{v}\,\| = \mathcal{O}(h^{m+1})$. Omdat $\|\vec{u}^{\,*} - \vec{v}^{\,(0)}\| = \mathcal{O}(h)$ en $\|\vec{v}^{\,(i)} - \vec{v}\,\| = \mathcal{O}(h^i)\|\vec{v}^{\,(0)} - \vec{v}\,\|$ (zie het kontraktie lemma) zien we dat $\|\vec{v}^{\,(m)} - \vec{u}^{\,*}\| = \mathcal{O}(h^{m+1})$.

Men kan het oplossen van een matrix-vektor vergelijking met matrix $A$ eenvoudig houden door $\rho_j$ precies van graad $j$ te nemen: $A$ is dan een driehoeks matrix.

**4.2.3 Voorbeeld.** De trapezium regel levert $v_1$ met nauwkeurigheid $\mathcal{O}(h^3)$. Vervolgens levert de midpunt regel $v_2$ ook met nauwkeurigheid $\mathcal{O}(h^3)$.

In hoofdstuk 6 zullen we Runge–Kutta methoden bespreken. Deze methoden worden ook veel gebruikt om differentiaalvergelijkingen benaderend op te lossen. Ze zijn bovendien uitstekend geschikt om een multistep methode te starten. We komen hierop terug.

### 4.3 De fout in multistep methoden nader bekeken

Is de start voldoende nauwkeurig, zijn de evaluatie fouten verwaarloosbaar, is de multistep stabiel en konsistent van orde $l$ en is $u^*$ in $C^{l+1}(\mathcal{J},\mathbf{R}^d)$ dan geeft (30) een bovenschatting voor $\|u^*(t_n) - u_h(t_n)\|$. Deze foutschatting kan bijzonder grof zijn (zie voorbeeld 4.3.1). Bovendien kunnen we op grond van deze foutschatting geen procedure bedenken waarmee we al rekenend de fout kunnen schatten, de benadering kunnen korrigeren, etc.. Voor dit soort procedures hebben we een nauwkeurigere beschrijving nodig van de fout. In stelling 4.3.7 in deze pargraaf zal blijken dat, onder zekere voorwaarden,

$$u_h = u^* - h^l \frac{1}{\sigma(1)} C_{l+1} w^* + \mathcal{O}(h^{l+1}) \qquad (h \to 0),$$

waarbij $w^*$ een gladde funktie is op $\mathcal{J}$ die onafhankelijk is van de multistep en van $h$.

**4.3.1 Voorbeeld.** Wensen we met $d = 1$ de oplossing $u^*$ in $t = 12$ van de differentiaalvergelijking $u' = -u$ op $[0,12]$ met beginwaarde $u(0) = 1$ te benaderen met een relatieve precisie van 1%, dan garandeert de schatting in (30) ons, als we met de trapezium regel werken, alleen die nauwkeurigheid met $h < 10^{-9}$. In de praktijk blijkt die nauwkeurigheid al te halen te zijn met $h \approx 10^{-1}$. De schatting verlangt van ons $10^8$ maal de hoeveelheid werk die noodzakelijk is.

**4.3.2 Bewering.** *Als* (Dif.0–2) *en* (Mul.2) *gelden dan*
$u^* \in C^{l+2}(\mathcal{J},\mathbf{R}^d)$ *en* $\delta_h(u^*) = h^l C_{l+1} u^{*(l+1)} + \mathcal{O}(h^{l+1})$ *uniform* $(h \to 0)$. □

Bij de nauwkeurigere analyse in deze pargraaf van de numerieke fouten komen we weer de funktionaalmatrix $J(t)$ tegen en de relatie (11) uit 1.2.15.

We beschrijven eerst, in de volgende bewering, de funktie $w^*$ die voor kleine $h$ de globale fout $e_h$ in de multistep oplossing bepaalt. De globale fout $e_h$ wordt niet bepaald door één lokale fout maar door alle lokale fouten die gemaakt worden in de opeenvolgende rekenstappen. $w^*$ is dan ook een oplossing van een inhomogene lineaire differentiaalvergelijking; in de inhomogene term komt de opeenvolging van de lokale fouten tot uiting (zie opmerking 4.3.8).

**4.3.3 Bewering.** *Stel* (Dif.2) *geldt.*
*Laat $w^* \in C^1(\mathcal{J},\mathbf{R}^d)$ de oplossing zijn van het volgende beginwaarde probleem van een gewone lineaire differentiaalvergelijking.*

$$\begin{cases} w'(t) = J(t)w(t) + u^{*(l+1)}(t) & \textit{voor alle} \quad t \in \mathcal{J} \\ w(t_0) = 0. \end{cases} \tag{32}$$

*Dan* $w^* \in C^2(\mathcal{J},\mathbf{R}^d)$ *en* $\max_{j<k} \|w^*(t_j)\| = \mathcal{O}(h)$.
*Als de multistep konsistent is dan geldt* $\delta_h(w^*) = \mathcal{O}(h)$ *uniform* $(h \to 0)$.

*Bewijs.* We laten het bewijs van de eerste twee beweringen over aan de lezer. De derde bewering volgt onmiddellijk uit de eerste en uit 4.1.8. □

Voordat we in stelling 4.3.7 laten zien dat $e_h$ inderdaad bepaald wordt door $w^*$ geven we eerst drie lemma's die het bewijs van de stelling overzichtelijker maken. Het eerste lemma berust ook weer op (11).

Als $v_h \in C(\mathcal{J}, \mathbf{R}^d)$ dan schrijven we $Jv_h(t_n)$ in plaats van $J(t_n)v_h(t_n)$.

**4.3.4 Lemma.** *Stel* (Dif.2) *geldt.*
*Voor iedere* $h \in \mathbf{H}$ *is* $u_h, e_h \in C(\mathcal{J}, \mathbf{R}^d)$ *zodat* $u_h = u^* - e_h$.
*Schrijf* $f_h(t) = f(t, u_h(t))$ *en* $f^*(t) = f(t, u^*(t))$. *Als* $e_h = \mathcal{O}(h^l)$ *uniform* $(h \to 0)$ *dan*

$$\sigma(T_h)(f_h) = \sigma(T_h)(f^*) + \sigma(T_h)(Je_h) + \mathcal{O}(h^{2l}) \quad \textit{uniform} \quad (h \to 0)$$

*Bewijs.* Met $\mu_h(t) := f(t, u^*(t)) - f(t, u_h(t)) - J(t)e_h(t) \quad (t \in \mathcal{J}_h)$ is, volgens (11),

$$\sup_{t \in \mathcal{J}_h} \|\mu_h(t)\| = \mathcal{O}((\sup_t \|e_h(t)\|)^2) = \mathcal{O}(h^{2l}) \qquad (h \to 0).$$

Omdat $\sup_{t_n} \|\sigma(T_h)(\mu_h)(t_n)\| \leq \sup_{t_n} \tilde{\sigma}\|\mu_h(t_n)\| = \mathcal{O}(h^{2l})$ volgt het lemma. □

**4.3.5 Lemma.** *Voor* $v \in C^1(\mathcal{J}, \mathbf{R}^d)$ *is* $\sigma(T_h)(v) - \sigma(1)(v) = \mathcal{O}(h)$ *uniform* $(h \to 0)$.

*Bewijs.* Met $M := \sup_{t \in \mathcal{J}} \|v'(t)\|$ is $\|\sigma(T_h)(v)(t_n) - \sigma(1)v(t_n)\| =$
$\|\sum_{j=0}^{k} \beta_{k-j}(v(t_{n+j}) - v(t_n))\| \leq \tilde{\sigma} \max_{j<k} \|v(t_{n+j}) - v(t_n)\| \leq kh\tilde{\sigma}M.$ □

**4.3.6 Lemma.** *Stel dat* (Dif.2) *en* (Mul.1) *gelden. Zij* $m \in \mathbf{N}$.
*Voor iedere* $h \in \mathbf{H}$ *en voor* $i = 1, 2$ *is* $\epsilon_h^{(i)} \in C(\mathcal{J}_h, \mathbf{R}^d)$ *en* $v_{h\,0}^{(i)}, \ldots, v_{h\,k-1}^{(i)}$ *in* $\mathbf{R}^d$. *Stel dat*

$$\max_{j<k} \|v_{h\,j}^{(1)} - v_{h\,j}^{(2)}\| = \mathcal{O}(h^m) \quad \textit{en} \quad \epsilon_h^{(1)} - \epsilon_h^{(2)} = \mathcal{O}(h^{m+1}) \qquad (h \to 0).$$

*Voor iedere* $h$ *uit die rij en voor* $i = 1, 2$ *is* $v_h^{(i)} \in C(\mathcal{J}_h, \mathbf{R}^d)$ *zodat*

$$\rho(T_h)(v_h^{(i)}) = h\sigma(T_h)(Jv_h^{(i)}) + \epsilon_h^{(i)} \quad \textit{op } \tilde{\mathcal{J}}_h \quad \textit{en} \quad v_h^{(i)}(t_j) = v_{h\,j}^{(i)} \quad (j < k).$$

*Dan is*

$$v_h^{(1)} - v_h^{(2)} = \mathcal{O}(h^m) \quad \textit{uniform} \quad (h \to 0).$$

*Bewijs.* Pas de stabiliteits stelling 4.1.26 toe nu met $f(t, x) := J(t)x$ en bedenk dat $\sup_t \|J(t)\| \leq L$. □

**4.3.7 Stelling.** *Stel dat* (Dif.0), (Dif.2), (Mul.1–2) *en* (Start.2) *gelden.*
*Laat voor iedere* $h \in \mathbf{H}$, *h klein genoeg,* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *zo zijn dat*

$$\begin{cases} \rho(T_h)(u_h) = h\sigma(T_h)(f_h) \quad \textit{met} \quad f_h(t) := f(t, u_h(t)) \quad \textit{op} \quad \tilde{\mathcal{J}}_h \\ u_h(t_j) = v_{h\,j} \quad \textit{voor} \quad j < k. \end{cases}$$

*Dan geldt, met de* $w^* \in C^2(\mathcal{J}, \mathbf{R}^d)$ *uit bewering* 4.3.3, *dat*

$$u_h = u^* - h^l \frac{1}{\sigma(1)} C_{l+1} w^* + \mathcal{O}(h^{l+1}) \quad \textit{uniform} \quad (h \to 0). \tag{33}$$

$\frac{1}{\sigma(1)}C_{l+1}$ *wordt de* foutkonstante *van de multistep genoemd* (zie ook 4.1.10).

*Bewijs.* Laat $e_h := u^* - u_h$ in $C(\mathcal{J}_h, \mathbf{R}^d)$ de globale fout zijn. Zoals we gezien hebben in de konvergentie stelling 4.1.31 (en de hoofdstelling) is

$$e_h = \mathcal{O}(h^l) \quad \text{uniform} \quad (h \to 0). \tag{34}$$

Schrijven we $\quad f^*(t) := f(t, u^*(t)) \quad (t \in \mathcal{J}) \quad$ dan vertelt 4.3.2 dat

$$\rho(T_h)(u^*) = h\sigma(T_h)(f^*) + h^{l+1}C_{l+1}u^{*(l+1)} + \mathcal{O}(h^{l+2}) \quad \text{uniform} \quad (h \to 0).$$

Omdat $\quad \rho(T_h)(u_h) = h\sigma(T_h)(f_h) \quad$ leert (34) en 4.3.4 ons dat

$$\rho(T_h)(e_h) = h\sigma(T_h)(Je_h) + h^{l+1}C_{l+1}u^{*(l+1)} + \mathcal{O}(h^{l+2}) \quad \text{uniform} \quad (h \to 0). \tag{35}$$

Beschouw nu de situatie in bewering 4.3.3.
Omdat $\quad w^* \in C^2(\mathcal{J}, \mathbf{R}^d) \quad$ en $\quad l \geq 1 \quad$ is, volgens 4.1.8,

$$\rho(T_h)(w^*) = h\sigma(T_h)(Jw^* + u^{*(l+1)}) + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0).$$

Lemma 4.3.5 vertelt ons dat
$h\sigma(T_h)(C_{l+1}u^{*(l+1)}) = h\sigma(1)C_{l+1}u^{*(l+1)} + \mathcal{O}(h^2) \quad$ uniform $\quad (h \to 0)$. Dus

$$\rho(T_h)(\frac{1}{\sigma(1)}C_{l+1}w^*) = h\sigma(T_h)(J\frac{1}{\sigma(1)}C_{l+1}w^*) + hC_{l+1}u^{*(l+1)} + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0).$$

Vermenigvuldigen met $h^l$ levert, uniform voor $h \to 0$,

$$\rho(T_h)(h^l\frac{1}{\sigma(1)}C_{l+1}w^*) = h\sigma(T_h)(Jh^l\frac{1}{\sigma(1)}C_{l+1}w^*) + h^{l+1}C_{l+1}u^{*(l+1)} + \mathcal{O}(h^{l+2}). \tag{36}$$

Verder is, voor $j < k$ volgens 4.3.3, $\quad e_h(t_j) - h^l\frac{1}{\sigma(1)}C_{l+1}w^*(t_j) = e_h(t_j) + \mathcal{O}(h^{l+1}) =$
$u^*(t_j) - v_{h\,j} + \mathcal{O}(h^{l+1}) = \mathcal{O}(h^{l+1})$. Vergelijken we (35) en (36) dan zien we, door toepassing van lemma 4.3.6, dat $\quad e_h - h^l\frac{1}{\sigma(1)}C_{l+1}w^* = \mathcal{O}(h^{l+1}) \quad$ uniform $(h \to 0)$ en de stelling volgt. □

**4.3.8 Opmerking.** Vergelijking (36) is in feite, op $\mathcal{O}(h^{l+2})$–termen na, de multistep toegepast op het beginwaarde probleem

$$w'(t) = J(t)w(t) + h^l\frac{1}{\sigma(1)}C_{l+1}u^{*(l+1)}, \qquad (t \in \mathcal{J}) \quad \text{en} \quad w(t_0) = 0.$$

$h^l\frac{1}{\sigma(1)}C_{l+1}w^*$ is hier de oplossing van. De vergelijking in (35), die de rekursieve relatie legt tussen de waarden van globale fout, kan dus geïnterpreteerd worden als een diskretisatie met de al gebruikte multistep van 'n differentiaalvergelijking. Men kan natuurlijk ook stellen dat $e_h$ in essentie de oplossing is van een lineaire differentiaalvergelijking met als inhomogene term in essentie de lokale diskretisatiefout (zie 4.3.2. $\sigma(1)$ is slechts een schalings faktor zie 4.1.10). We werken deze visie voor een lineaire vergelijking wat konkreter uit in 4.3.10.

**4.3.9 Opmerking.** De fout $e_h$ in de multistep benadering $u_h$ van $u^*$ is dus $\approx \frac{h^l}{\sigma(1)}C_{l+1}w^*$ waarin $w^*$ een gladde funktie is die *niet* afhangt van de stapgrootte $h$ en zelfs niet van het multistep schema. De startfouten moeten dan wel een orde nauwkeuriger zijn dan de konsistentie orde van de multistep. Zijn we alleen geïnteresseerd in een majorant van de fout $e_h$ en niet in de struktuur ervan dan konden we al bevredigend werken met startfouten van dezelfde orde als de multistep (zie 4.1.32).

**4.3.10 Opmerking.**
Beschouw, met $\quad d = 1 \quad$ en $\quad f(t, x) := J(t)x + g(t), \quad$ de lineaire differentiaalvergelijking

$$u'(t) = J(t)u(t) + g(t) \quad \text{voor} \quad t \in \mathcal{J} \quad \text{en} \quad u(t_0) = u_0. \tag{37}$$

Als we de exacte oplossing $u^*$ op tijdstip $t = \tau$ storen met $\varepsilon$ en $\tilde{u}^*$ is de resulterende oplossing (precieser: $\quad \tilde{u}^*(\tau) = u^*(\tau) + \varepsilon \quad$ en $\quad \frac{d\tilde{u}^*}{dt} = J(t)\tilde{u}^* + g \quad$ op $\quad \mathcal{J}$) dan is de fout

$$\tilde{u}^*(t) - u^*(t) = \varepsilon G(t, \tau) \quad \text{met} \quad G(t, \tau) := \exp(\int_\tau^t J(s)\, ds) \quad \text{voor} \quad t \geq \tau : \tag{38}$$

de verstoring $\varepsilon$ wordt voortgeplant volgens een oplossing van de homogene differentiaalvergelijking $u'(t) = J(t)u(t) \quad (t \in \mathcal{J})$ (zie (b) in 1.2.17)[10]
(als $J(t) = \eta$ voor alle $t$ dan is $\tilde{u}^*(t) - u^*(t) = \varepsilon G(t,\tau) = \varepsilon e^{\eta(t-\tau)}$).

In iedere numerieke oplosmethode waarmee we de oplossing in de tijd volgen maken we in zekere tijdstippen fouten (diskretisatie fouten en evaluatie fouten). Op zijn gunstigs zal zo'n fout in het numeriek rekenproces voortgeplant worden volgens een oplossing van de homogene differentiaalvergelijking.

Met betrekking tot (37) kunnen we $w^*$ die aan (32) voldoet expliciet oplossen (zie (3)):

$$h^l \frac{1}{\sigma(1)} C_{l+1} w^*(t) = \frac{1}{\sigma(1)} h^l C_{l+1} \int_{t_0}^{t} \exp\left(\int_{\tau}^{t} J(s)\,ds\right) u^{*(l+1)}(\tau)\,d\tau. \tag{39}$$

De funktie $t \to \exp(\int_\tau^t J(s)\,ds)$ is voor $t \geq \tau$ de oplossing van de homogene differentiaalvergelijking $u'(t) = J(t)u(t)$ met waarde 1 in $\tau$. Vervangen we in (39) de $\int_{t_0}^t$ door een Riemann som met stapgrootte $h$ dan zien we dat de lokale diskretisatie fout $h\delta_h(u^*)(\tau) \approx C_{l+1}h^{l+1}u^{*(l+1)}(\tau)$ gedeeld door $\sigma(1)$ zich ongeveer volgens een oplossing van de homogene differentiaalvergelijking voortplant, precies zoals we hierboven gehoopt hadden en voor speciale gevallen in het college Numerieke Wiskunde A al gezien hadden. Dit resultaat geldt, voor voldoend kleine $h$ en op orde $\mathcal{O}(h^{l+1})$ termen na, ook voor het geval $d > 1$ of $f$ niet lineair is: ook dan wordt iedere lokale diskretisatie fout voortgeplant ongeveer volgens een oplossing van de homogene differentiaalvergelijking $u'(t) = J(t)u(t)$.

De schatting in 4.1.31 "houdt geen rekening" met het feit dat oplossingen van de homogene differentiaalvergelijking dalend kunnen zijn en dat daardoor lokale fouten gedempt kunnen worden (de $L$ voor het probleem in voorbeeld 4.3.1 is $+1$, een waarde die we ook voor de vergelijking $u' = +u$, met stijgende oplossingen, vinden! Zie ook 1.2.12 en 1.2.14). De schatting in 4.3.7 doet dat blijkbaar wel.

**4.3.11 Opmerking.** Door parallel aan de iteratieve berekening van $u_h$ ook de differentiaalvergelijking in bewering 4.3.3 benaderend op te lossen met een eenvoudige multistep (als Euler forward) kan men een indruk krijgen over de fout $e_h$. Voor de waarde $u^*(t_n)$ die men in deze eenvoudige methode nodig heeft kan men de recent bepaalde $u_h(t_n)$ nemen, $u^{*(l+1)}(t_n)$ kan men met een eindige differentie benaderen.

**4.3.12 Richardson extrapolatie.** In het college Numerieke Wiskunde A hebben we gezien dat we, in geval de fout $e_h \approx h^l \frac{1}{\sigma(1)} C_{l+1} w^*$ is, al rekenend de fout kunnen schatten, de benadering korrigeren, etc..

Kies $h$ klein genoeg. Beschouw een $\tau \in \mathcal{J}_h$. Als $\frac{1}{\sigma(1)} C_{l+1} w^*(\tau) \neq 0$ (groot in absolute waarde t.o.v. de $\mathcal{O}(h^{l+1})$–term) dan zal $2^l e_{\frac{1}{2}h}(\tau) \approx e_h(\tau)$ zijn en dus $u_h(\tau) - u_{\frac{1}{2}h}(\tau) \approx (2^l - 1)e_{\frac{1}{2}h}(\tau)$: berekenen we een benaderende oplossing $u_h$ met stapgrootte $h$ én een $u_{\frac{1}{2}h}$ met stapgrootte $\frac{1}{2}h$ dan geeft het veelvoud $(2^l-1)^{-1}(u_h - u_{\frac{1}{2}h})$ van het verschil een $\mathcal{O}(h^{l+1})$–schatting voor de fout $e_{\frac{1}{2}h}$ en zal de *Richardson extrapolant* $u_{\frac{1}{2}h} - (2^l-1)^{-1}(u_h - u_{\frac{1}{2}h})$ een $\mathcal{O}(h^{l+1})$ benadering zijn voor $u^*$.

De vraag dringt zich op of we wellicht een Romberg schema (zie Numerieke Wiskunde A diktaat) kunnen opstellen. De fout zal daartoe van de vorm

$$e_h = h^l \tfrac{1}{\sigma(1)} C_{l+1} w^* + h^{l+1} w_1 + \ldots + h^{l+m-1} w_{m-1} + \mathcal{O}(h^{l+m}) \quad \text{uniform} \quad (h \to 0),$$

moeten zijn met $w_j : \mathcal{J} \to \mathbf{R}^d$ onafhankelijk van $h$. Met het oog op het resultaat in 4.1.8 lijkt de hoop gerechtvaardigd dat een induktieve aanpak met argumenten als in het bewijs van stelling

[10] $\tilde{u}^*(t) - u^*(t) = \varepsilon G(t,\tau)$ in (38) is ook korrekt in geval $d > 1$; echter $G(t,\tau) := \exp(\int_\tau^t J(s)\,ds)$ geldt alleen onder 'n extra voorwaarde als " $J(t)J(s) = J(s)J(t)$ voor alle $t, s \in \mathcal{J}$".

4.3.7 leidt tot zo'n vorm voor de fout. Echter als het $\rho$ polynoom naast 1 ook nog andere wortels op de eenheidscirkel heeft dan zullen de "startfouten" $h^l \frac{1}{\sigma(1)} C_{l+1} w^*(t_1), \ldots, h^l \frac{1}{\sigma(1)} C_{l+1} w^*(t_{k-1})$ een bijdrage leveren aan de $\mathcal{O}(h^{l+1})$–term in $e_h$ die verre van glad is. We laten dit laatste zien aan de hand van een illustratief voorbeeld.

**4.3.13 Voorbeeld.** Beschouw, met $d = 1$, de differentiaalvergelijking

$$u'(t) = 3t^2 \quad \text{voor} \quad t \in [0, T] \quad \text{en} \quad u(0) = 0$$

met exacte oplossing $u^*(t) = t^3$.
We berekenen $u = u_h$ met de midpunt regel en we starten exact. Dus

$$\begin{cases} u^*_{n+2} = u^*_n + 2h3t^2_{n+1} + 2h^3 \\ u_0 = 0, u_1 = u^*(t_1) = h^3 \\ u_{n+2} = u_n + 2h3t^2_{n+1}. \end{cases}$$

$e_n = e_h(t_n) = u^*_n - u_n$ voldoet aan

$$\begin{cases} e_{n+2} = e_n + 2h^3 \\ e_0 = 0, e_1 = 0. \end{cases}$$

Hieruit volgt dat $e_{2n} = n2h^3 = h^2 t_{2n}$ en $e_{2n+1} = n2h^3 = h^2 t_{2n+1} - h^3$.
Met $w^*(t) = t$ is blijkbaar

$$e_h(t_n) = h^2 w^*(t_n) - h^3 [\frac{1}{2}(1 - (-1)^n)] \quad \text{voor alle} \quad t_n \in \mathcal{J}_h :$$

de $\mathcal{O}(h^3)$–term is niet van de vorm $h^3 w_1$ met $w_1 : [0, T] \to \mathbf{R}$ onafhankelijk van $h$.
De faktor $(-1)^n$ "stamt af" van de wortel $-1$ op de eenheids cirkel van het $\rho$ polynoom: $\rho(\zeta) = \zeta^2 - 1$. Merk op dat

$$h^3[\frac{1}{2}(1 - (-1)^n)] = h^2 w^*(0)[\frac{1}{2}(1 + (-1)^n)] + h^2 w^*(t_1)[\frac{1}{2}(1 - (-1)^n)].$$

Uit dit voorbeeld blijkt dat de struktuur van de globale fout ondermeer bepaald wordt door de wortels van $\rho$. Voordat we in 4.3.18 een resultaat formuleren over de struktuur van de $\mathcal{O}(h^{l+1})$–term in $e_h$ kijken we, in 4.3.14–4.3.17, wat nauwkeuriger naar de wijze waarop de wortels van $\rho$ (in feite van $\rho - \tilde{\eta}\sigma$) de fout beïnvloeden.

**4.3.14 Parasitaire wortels.**
Beschouw, voor $d = 1$, $\eta \in \mathbf{R}$ en $g \in C(\mathcal{J})$ het probleem

$$u'(t) = \eta u(t) + g(t) \quad \text{voor} \quad t \in \mathcal{J} \quad \text{en} \quad u(t_0) = u_0. \tag{40}$$

Los (40), met stapgrootte $h$, op met een stabiele konsistente multistep methode.
De fout $e_h$ in de multistep oplossing van (40) voldoet, met $e_n = e_h(t_n)$, aan

$$\sum_{j=0}^{k} \alpha_{k-j} e_{n+j} = h\eta \sum_{j=0}^{k} \beta_{k-j} e_{n+j} + \mu_n \quad \text{met} \quad \mu_n := h\delta_n - \epsilon_n,$$

waarin $\delta_n := \delta_h(u^*)(t_n)$ de lokale diskretisatie fout is en $\epsilon_n$ de lokale evaluatie fout.
Beschouw de rij $(G_n)_{n \in \mathbf{Z}}$ waarvoor geldt

$$\begin{cases} G_j = 0 \quad \text{voor} \quad j \in \mathbf{Z}, j < 0, \quad G_0 = 1 \\ \sum_{j=0}^{k} (\alpha_{k-j} - h\eta\beta_{k-j}) G_{n-k+j} = 0 \quad \text{voor alle} \quad n \in \mathbf{Z}, n \neq 0. \end{cases}$$

Met $\mu_{-j} := (\alpha_0 - h\eta\beta_0)e_{k-j} + \ldots + (\alpha_{k-j} - h\eta\beta_{k-j})e_0$ voor $j = 1, \ldots, k$ geldt (ga dit door invullen na)

$$\begin{aligned} e_n &= \sum_{\nu=0}^{k-1} G_{n-\nu}\frac{\mu_{\nu-k}}{\alpha_0 - h\eta\beta_0} + \sum_{\nu=k}^{n} G_{n-\nu}\frac{\mu_{\nu-k}}{\alpha_0 - h\eta\beta_0} \\ &= \sum_{\nu=0}^{n} G_{n-\nu}\frac{\mu_{\nu}}{\alpha_0 - h\eta\beta_0} = \sum_{\nu=0}^{\infty} G_{n-\nu}\frac{\mu_{\nu}}{\alpha_0 - h\eta\beta_0} \quad \text{voor alle} \quad n \in \mathbf{N}_0. \end{aligned}$$

De lokale fout $\mu_{\nu-k}/(\alpha_0 - h\eta\beta_0)$, gemaakt bij de berekening van $u_h(t_\nu)$, vindt men terug in de globale fout op tijdstip $t = t_n$ "opgeblazen" met de faktor $G_{n-\nu}$. (Vergelijk dit met (12) en (10).) Gezien onze overwegingen in 4.3.10 vragen wij ons af in hoeverre $G_{n-\nu}$ lijkt op $G(t_n, t_\nu) = e^{\eta(t_n - t_\nu)}$: is de *diskrete Greense funktie* $(G_{n-\nu})$ een benadering van de kontinue?
Schrijf $\tilde\eta := h\eta$. Laat $\lambda_1(\tilde\eta), \ldots, \lambda_k(\tilde\eta)$ de wortels zijn van $\rho - \tilde\eta\sigma$ geteld naar multipliciteit. De wortels van $\rho$ op de eenheidscirkel zijn enkelvoudig. Om te technische details te vermijden nemen we aan dat alle wortels van $\rho$ enkelvoudig zijn. Voor voldoend kleine $\tilde\eta$ zijn de wortels van $\rho - \tilde\eta\sigma$ dat dan ook. We nemen aan dat $\tilde\eta$ inderdaad zo klein is, dus: $\lambda_i(\tilde\eta) \neq \lambda_j(\tilde\eta)$ als $i \neq j$. Verder schrijven we $\lambda_j := \lambda_j(0)$ en we nemen aan dat de nummering zo is dat $\lambda_1 = 1$.
Omdat $(G_n)_{n>0}$ een oplossing is van een homogene rekursie met konstante koëfficiënten zijn er (ga dit na, zie ook 4.3.15) $\gamma_1(\tilde\eta), \ldots, \gamma_k(\tilde\eta)$ in $\mathbf{C}$ zodat

$$G_{n-\nu} = \gamma_1(\tilde\eta)\lambda_1(\tilde\eta)^{n-\nu} + \ldots + \gamma_k(\tilde\eta)\lambda_k(\tilde\eta)^{n-\nu} \quad \text{voor} \quad n > \nu.$$

Merk op dat $\lambda_j(t) = \lambda_j(0) + t\lambda_j'(0) + \mathcal{O}(t^2)$ $(t \to 0)$. Verder is $\lambda_j'(0) = \frac{\sigma(\lambda_j)}{\rho'(\lambda_j)}$ ( gebruik dat $\rho(\lambda_j(t)) - t\sigma(\lambda_j(t)) = 0$ voor alle $t$ en differentieer $t \to \rho(\lambda_j(t)) - t\sigma(\lambda_j(t))$ in $t = 0$. In feite passen we de impliciete funktie stelling toe op de funktie $(\zeta, t) \to \rho(\zeta) - t\sigma(\zeta)$. De veronderstelling dat de funkties $t \to \lambda_j(t)$ glad zijn is ook met het oog op deze stelling gerechtvaardigd). Dus

$$\lambda_j(\tilde\eta) = \lambda_j + \tilde\eta\frac{\sigma(\lambda_j)}{\rho'(\lambda_j)} + \mathcal{O}(\tilde\eta^2) \quad (\tilde\eta \to 0).$$

Uit de konsistentie volgt dat $\lambda_1(\tilde\eta) = 1 + \tilde\eta + \mathcal{O}(\tilde\eta^2)$, dus
$\lambda_1(\tilde\eta)^{n-\nu} = e^{(n-\nu)\tilde\eta}(1 + \mathcal{O}(\tilde\eta)) = e^{\eta(t_n - t_\nu)}(1 + \mathcal{O}(h))$ uniform $(h \to 0)$.

- De wortel 1 van $\rho$, de *hoofdwortel*, "werkt" dus mee aan de voortplanting van de lokale fout $\mu_{\nu-k}$ in de globale fout volgens $\approx \gamma_1(\tilde\eta)\mu_{\nu-k}e^{\eta(t_n-t_\nu)}$; een voortplanting zoals we die, op de faktor $\gamma_1(\tilde\eta)$ na, verwachten. (De wortel 1, of nauwkeuriger $1 + \eta h + \mathcal{O}(h^2)$, hebben we überhaupt nodig om de oplossing numeriek te kunnen volgen).

De overige wortels, de *parasitaire*, werken op een totaal andere manier mee.

- Als $|\lambda_j| < 1$ dan is $|\lambda_j(\tilde\eta)| < 1$ voor voldoende kleine $h$. Voor iedere $N \in \mathbf{N}$ is $\lambda_j(\tilde\eta)^{n-\nu} = \mathcal{O}(h^N)$ als $t_0 + nh \to t$ voor 'n $t > t_\nu$ en $h \to 0$ (dus $n \to \infty$): voor $h$ klein is de invloed van dit soort wortels al vlug verwaarloosbaar.
- $\lambda_j$ is een *essentiële wortel* als $|\lambda_j| = 1$. Met *groeiparameter* $\kappa := \frac{\sigma(\lambda_j)}{\lambda_j\rho'(\lambda_j)}$ is dan $\lambda_j(\tilde\eta) = \lambda_j(1 + \kappa\tilde\eta + \mathcal{O}(\tilde\eta^2))$ en $\lambda_j(\tilde\eta)^{n-\nu} = \lambda_j^{n-\nu}e^{\eta\kappa(t_n-t_\nu)}(1 + \mathcal{O}(h))$. De term $\lambda_j^{n-\nu}$ zorgt voor een slingerend gedrag in deze komponent van de foutbijdrage, een komponent die bovendien niet uitdempt (met betrekking tot zekere type fouten kan het slingeren er wel voor zorgen dat verschillende foutbijdrage min of meer tegen elkaar wegvallen; zie 4.3.17).

De koëfficiënten $\gamma_j(\tilde\eta)$ hangen van $\tilde\eta$ af. Maar dit feit heeft geen essentiële invloed op de wijze waarop de lokale fouten worden voortgeplant. Er geldt $\gamma_j(\tilde\eta) = \gamma_j(0) + \mathcal{O}(h)$ voor $h \to 0$ (zie de volgende opgave).

**4.3.15 Opgave.** Beschouw de situatie in 4.3.14. Beschouw de vergelijking

$$V(\tilde\eta)\begin{bmatrix} \tilde\gamma_1 \\ \vdots \\ \tilde\gamma_{k-1} \\ \tilde\gamma_k \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad \text{met} \quad V(\tilde\eta) := \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \lambda_1(\tilde\eta) & \lambda_2(\tilde\eta) & \ldots & \lambda_k(\tilde\eta) \\ \vdots & \vdots & & \vdots \\ \lambda_1(\tilde\eta)^{-k+1} & \lambda_2(\tilde\eta)^{-k+1} & \ldots & \lambda_k(\tilde\eta)^{-k+1} \end{bmatrix}.$$

Ga na dat $\gamma_j(\tilde{\eta}) = \tilde{\gamma}_j \lambda_j(\tilde{\eta})^{k-1} \quad (j = 1, \ldots, k)$.
Met $\Delta(\tilde{\eta}) := V(\tilde{\eta}) - V(0)$ is $\|\Delta(\tilde{\eta})\| = \mathcal{O}(h)$. Ga dit na. Ga na dat de Vandermonde matrix $V(0)$ inverteerbaar is en laat zien dat $\|V(\tilde{\eta})^{-1} - V(0)^{-1}\| = \|V(0)^{-1}\Delta(\tilde{\eta})V(\tilde{\eta})^{-1}\| = \mathcal{O}(h)$.
Konludeer dat $\gamma_j(\tilde{\eta}) = \gamma_j(0) + \mathcal{O}(h)$.

**4.3.16 Definitie.** We noemen een multistep met schema $(\rho, \sigma)$ *sterk stabiel* als hij stabiel is en $\{1\} = \{\lambda \in \mathbf{C} \,\big|\, |\lambda| = 1, \rho(\lambda) = 0\}$: 1 is de enige wortel van $\rho$ op de eenheidscirkel. Een multistep die stabiel maar niet sterk stabiel is noemen we *zwak stabiel*: het $\rho$ polynoom van een zwak stabiele multistep heeft naast 1 ook nog andere wortels op de eenheidscirkel.

Konsistente 1–staps methoden zijn sterk stabiel evenals de methoden van Adams. Iedere $k$-staps methode met $k > 1$ en $\rho = \chi^k - 1$ of $\rho = \chi^k - \chi^{k-2}$ is zwak stabiel; de methoden van Milne zijn dus zwak stabiel. BDF-methoden zijn, voor $k \leq 6$, sterk stabiel.

**4.3.17 Opmerking.** We hebben in 4.3.14 gezien dat door de parasitaire wortels op de eenheidscirkel in een zwak stabiele multistep het dempend karakter van de homogene differentiaalvergelijking in de multistep om zeep geholpen kan worden. Om de mate waarin dit kan gebeuren beter te begrijpen bekijken we ter illustratie een konkrete zwak stabiele multistep methode.

Beschouw weer 4.3.14. We berekenen, met stapgrootte $h$, de benaderende oplossing van de vergelijking in (40) met behulp van de midpunt regel.
Dan is $\lambda_1(\tilde{\eta}) = 1 + \tilde{\eta} + \mathcal{O}(\tilde{\eta}^2)$ en $\lambda_2(\tilde{\eta}) = -1 + \tilde{\eta} + \mathcal{O}(\tilde{\eta}^2) = (-1)(1 - \tilde{\eta} + \mathcal{O}(\tilde{\eta}^2))$. Dus $\lambda_1(\tilde{\eta})^{n-\nu} = e^{\eta(t_n - t_\nu)}(1 + \mathcal{O}(h))$ en $\lambda_2(\tilde{\eta})^{n-\nu} = (-1)^{n-\nu}e^{-\eta(t_n-t_\nu)}(1 + \mathcal{O}(h))$ uniform $(h \to 0)$.
Verder is $G_{n+1} = \frac{1}{\lambda_1(\tilde{\eta}) - \lambda_2(\tilde{\eta})}(\lambda_2(\tilde{\eta})^n - \lambda_2(\tilde{\eta})^n)$. Omdat $\lambda_1(\tilde{\eta}) - \lambda_2(\tilde{\eta}) = 2 + \mathcal{O}(\tilde{\eta}^2)$ is

$$G_{n-\nu} = \frac{1}{2}\left(e^{\eta(t_n-t_\nu)} - (-1)^{n-\nu}e^{-\eta(t_n-t_\nu)}\right)(1 + \mathcal{O}(h)) \quad \text{uniform} \quad (h \to 0).$$

- Als $\eta > 0$ heeft de parasitaire wortel een dempende effect: de fout groeit toch, nu door de “werking van de hoofdwortel”, volgens de nu groeiende oplossing van de homogene differentiaalvergelijking (bedenk wel dat de in het kontinue probleem ook iedere fout zal groeien).
- Als $\eta < 0$ dan “zorgt” de parasitaire wortel voor een slingerende groei met een in absolute waarde grote groeifaktor (als $\eta = -1$ is deze groeifaktor van $t_\nu$ tot $t_n = t_\nu + 12$ gelijk aan $1.6\,10^5$!).

Uit het resultaat in stelling 4.3.7 mogen we echter konkluderen dat de parasitaire wortels op de eenheidscirkel de gladde en niet al te grote bijdrage van de lokale diskretisatie fout in de $\mathcal{O}(h^l)$–term van $e_h$ (de $h^l\frac{1}{\sigma(1)}C_{l+1}w^*$–term) toch niet verstoren. We kunnen dit als volgt verklaren.
Stel weer dat $\lambda_2 = -1$. Dan $\lambda_2(\tilde{\eta})^{p-\nu}u^{*(l+1)}(t_\nu) + \lambda_2(\tilde{\eta})^{p-\nu-1}u^{*(l+1)}(t_{\nu+1}) \approx$

$$\lambda_2^{p-\nu}u^{*(l+1)}(t_\nu) + \lambda_2^{p-\nu-1}u^{*(l+1)}(t_{\nu+1}) = (-1)^{p-\nu}(u^{*(l+1)}(t_\nu) - u^{*(l+1)}(t_{\nu+1})) = \mathcal{O}(h)$$

omdat $u^{*(l+1)} \in C^1(\mathcal{J})$. In het algemeen, als $\lambda_j$ een parasitaire wortel is op de eenheidscirkel, komt door het alternerend gedrag van de rij $(\lambda_j(\tilde{\eta})^{p-\nu})_\nu$ en de gladheid van $u^{*(l+1)}$ de gezamelijke bijdrage van de lokale diskretisatie fouten ten gevolge van deze wortel in een hogere orde term terecht. Door de alternerende foutvoortplanting en de gladheid van de lokale fout heffen de verschillende bijdrage tot de fout elkaar min of meer op. Dit gelukkige opheffings fenomeen zal zich echter niet voordoen met betrekking tot lokale evaluatie fouten of in de $\mathcal{O}(h^{l+1})$–termen van $e_h$ (zie ook voorbeeld 4.3.13). Voor $h$ klein genoeg zal $e_h \approx h^l\frac{1}{\sigma(1)}C_{l+1}w^*$ zijn (dat hebben we immers bewezen) maar voor zwak stabiele methoden en een “stabiele” differentiaalvergelijking (een differentiaalvergelijking waarvan de oplossingen van $w'(t) = J(t)w(t)$ dalen) zal de $\mathcal{O}(h^{l+1})$–term in $e_h$ maar alleen voor erg kleine $h$ verwaarloosbaar zijn.

Het zal duidelijk zijn dat men in de praktijk, als men het resultaat uit stelling 4.3.7 wil uitbuiten, aan de slag zal gaan met een sterk stabiele multistep methode.

Zonder bewijs vermelden we het volgende resultaat waarin de struktuur van de $\mathcal{O}(h^{l+1})$–term van $e_h$ beschreven wordt.

**4.3.18 Bewering.** *Stel* (Dif.0), (Dif.3), (Mul.1–2) *en* (Start.3) *gelden.*
*Laat voor* $h \in \mathbf{H}$, *h klein genoeg,* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *zo zijn dat*

$$\begin{cases} \rho(T_h)(u_h) = h\sigma(T_h)(f_h) \quad \textit{met} \quad f_h(t) := f(t, u_h(t)) \\ u_h(t_j) = v_{h\,j} \quad \textit{voor} \quad j < k \quad (h \to 0). \end{cases}$$

(a) *Als de multistep sterk stabiel is, dan zijn er* $w^*, w_1, \ldots, w_{m-1} \in C(\mathcal{J}, \mathbf{R}^d)$ *zodat, voor iedere* $\varepsilon > 0$, *voor* $h \to 0$ *uniform op* $[t_0 + \varepsilon, t_0 + T]$, *geldt dat*[11]

$$u_h = u^* - h^l \frac{1}{\sigma(1)} C_{l+1} w^* - h^{l+1} w_1 - \ldots - h^{l+m-1} w_{m-1} + \mathcal{O}(h^{l+m}).$$

(b) *Als de multistep zwak stabiel is,* $m > 1$, *en* $\lambda_1 = 1, \lambda_2, \ldots, \lambda_\kappa$ *zijn de verschillende wortels van* $\rho$ *op de eenheidscirkel, dan zijn er* $w^*, w_1, \ldots, w_\kappa \in C(\mathcal{J}, \mathbf{R}^d)$ *zodat,*
*voor iedere* $\varepsilon > 0$, *voor* $h \to 0$ *uniform op* $[t_0 + \varepsilon, t_0 + T]$, *geldt dat*

$$u_h(t_n) = u^*(t_n) - h^l \frac{1}{\sigma(1)} C_{l+1} w^*(t_n) - h^{l+1}(\lambda_1^n w_1(t_n) + \ldots + \lambda_\kappa^n w_\kappa(t_n)) + \mathcal{O}(h^{l+2}). \quad \square$$

**4.3.19 Opmerking.** Stel dat, in bovenstaande bewering de start van orde $l$ is:
$v_{h\,j} = u^*(t_j) + \sum_{i=0}^{m-1} h^{l+i} c_{ij} + \mathcal{O}(h^{l+m})$ voor $j < k$ $(h \to 0)$ (zie (Start.3)).
Dan is de uitspraak in (a) ook korrekt als we de foutkonstante $\frac{1}{\sigma(1)} C_{l+1}$ maar vervangen door een geschikte andere konstante.
Ook uitspraak (b) heeft in dit geval een analogon: de "$\mathcal{O}(h^l)$–term" van de fout $u^* - u_h$ is al opgebouwd uit termen $h^l \lambda_j^n w_j(t_n)$.

In de volgende opgave geven we aan hoe we de doorwerking van de startfouten in de "asymptotische ontwikkeling" van de fout kunnen analyseren. Bovenstaande bewering kunnen we bewijzen door (in essentie) de argumenten in de volgende opgave te kombineren met die in het bewijs van 4.3.7. Om het bewijs helemaal goed te maken heeft men echter met vele vervelende en technische details te maken.

**4.3.20 ∘ Opgave.** Stel dat $\rho$ en $\sigma$ geen gemeenschappelijke faktoren hebben en dat alle wortels van $\rho$ enkelvoudig zijn.
Zij $\lambda \neq 0$ een wortel van $\rho$. Laat, voor iedere $h > 0$, $v_h \in C(\mathcal{J}, \mathbf{R}^d)$ zo zijn dat

$$\begin{cases} \rho(T_h)(v_h) = h\sigma(T_h)(Jv_h) \quad \text{op} \quad \tilde{\mathcal{J}}_h \\ v_h(t_j) = \lambda^j \quad \text{voor} \quad j = 0, \ldots, k-1. \end{cases}$$

Laat $\lambda' := \frac{\sigma(\lambda)}{\lambda\rho'(\lambda)}$ de groeiparameter zijn van $\lambda$. Laat $w_\lambda \in C^1(\mathcal{J}, \mathbf{R}^d)$ voldoen aan

$$\begin{cases} w'_\lambda(t) = \lambda' J(t) w_\lambda(t) \quad \text{voor} \quad t \in \mathcal{J} \\ w_\lambda(t_0) = 1. \end{cases}$$

We bewijzen dat $v_h(t_n) = \lambda^n w_\lambda(t_n) + \mathcal{O}(h)$ uniform $(h \to 0)$.
(dat wil zeggen $\sup_{n, nh \leq T} \|v_h(t_0 + nh) - \lambda^n w_\lambda(t_0 + nh)\| = \mathcal{O}(h)$ voor $h \to 0$.)

[11] Hier in (a) en in (b) is, voor iedere $\varepsilon > 0$, de orde uitspraak uniform op $[t_0 + \varepsilon, t_0 + T]$: om een uitspraak te krijgen die uniform is op $\mathcal{J}$ moeten er in de machtreeksontwikkeling nog termen bij van de vorm $h^{l+j} \lambda_i^n w_{ij}(t_n)$, waarbij $\lambda_i$ een wortel is van $\rho$ met $|\lambda_i| < 1$ en $w_{ij}$ een gladde funktie (zie opgave 4.3.20). Voor iedere $t > t_0$ en voor iedere $N \in \mathrm{N}$ is echter $\lambda_i^n w_{ij}(t_n) = \mathcal{O}(h^N)$ als $t_0 + nh \to t$ en $h \to 0$ (zie ook 4.3.14).

a. Definiëer $\rho_\lambda(\zeta) := \rho(\lambda\zeta)$ en $\sigma_\lambda(\zeta) := \frac{1}{\lambda'}\sigma(\lambda\zeta)$ voor $\zeta \in \mathbf{C}$.
Dan $\rho_\lambda(1) = 0, \rho'_\lambda(1) = \sigma_\lambda(1)$, $\rho_\lambda(T_h)(w_\lambda) = h\sigma_\lambda(T_h)(\lambda' J w_\lambda) + \mathcal{O}(h^2)$ uniform $(h \to 0)$ en $w_\lambda(t_j) = 1 + \mathcal{O}(h)$ voor $j < k$. Bewijs dit.
b. Met $\tilde{w}_\lambda(t_n) := \lambda^n w_\lambda(t_n)$ geldt $\rho(T_h)(\tilde{w}_\lambda) = h\sigma(J\tilde{w}_\lambda) + \mathcal{O}(h^2)$ uniform, en $\tilde{w}_\lambda(t_j) = \lambda^j + \mathcal{O}(h)$ $(h \to 0)$ voor $j < k$. Bewijs dit.
c. Toon aan dat $v_h(t_n) = \tilde{w}_\lambda(t_n) + \mathcal{O}(h) = \lambda^n w_\lambda(t_n) + \mathcal{O}(h)$ uniform $(h \to 0)$.
d. Schrijf $\Lambda := \{\lambda \in \mathbf{C} \,\big|\, \rho(\lambda) = 0\}$.
Als $v_0, \ldots, v_{k-1} \in \mathbf{R}^d$ en voor iedere $h > 0$ is $v_h \in C(\mathcal{J}, \mathbf{R}^d)$ zodat

$$\rho(T_h)(v_h) = h\sigma(T_h)(Jv_h) \quad \text{op} \quad \tilde{\mathcal{J}}_h, \quad v_h(t_j) = v_j \quad \text{voor} \quad j < k$$

dan zijn er $\gamma_\lambda \in \mathbf{R}$ zodat

$$v_h(t_n) = \sum_{\lambda \in \Lambda} \gamma_\lambda \lambda^n w_\lambda(t_n) + \mathcal{O}(h) \quad \text{uniform} \quad (h \to 0).$$

Bewijs dit. (Hint: de vektoren $(1, \lambda, \ldots, \lambda^{k-1})^T$, met $\lambda \in \Lambda$, vormen een basis voor $\mathbf{R}^k$.)

## 4.4 Stapgrootte besturing

**4.4.1 Het nut van stapgrootte besturing.** Voor zekere niet al te kleine stapgrootte $\Delta t$ en zekere nauwkeurigheid $\bar{\epsilon}$, wenst men wat betreft het beginwaarde probleem gewoonlijk alleen waarden te kennen die de exacte waarden $u^*(\tau_j)$ in de *steunpunten* $\tau_j := t_0 + j\Delta t$ met nauwkeurigheid $\bar{\epsilon}$ benaderen (met $T = 10$ kan men bijvoorbeeld aan $\Delta t = 0.1$ denken en $\bar{\epsilon}$ zodat $\bar{\epsilon} \leq 0.01 \max_j \|u^*(\tau_j)\|$). Wil men $u^*$ ook benaderen in tussenliggende punten, omdat men bijvoorbeeld de grafiek wil plotten, dan berekent men die benaderingen uit de benaderingen in de $\tau_j$ door stuksgewijze polynoom interpolatie of spline benadering (omdat niet alleen de benadering van $u^*(\tau_j)$ bekend is maar ook van diens afgeleide $f(\tau_j, u^*(\tau_j))$ gebruikt men daarvoor nogal eens Hermite interpolatie).

Om de benadering in de $\tau_j$ te krijgen met de precisie $\bar{\epsilon}$ moet men gewoonlijk de multistep rekursie doorlopen met een stapgrootte $h$ die beduidend kleiner is dan $\Delta t$ (denk bijvoorbeeld aan $\Delta t = 0.1$ en $h = 0.01$). Om efficiëntie redenen is het gewenst $h$ zo groot mogelijk te nemen. Verder is het prettig als men de computer de stapgrootte kan laten bepalen. In geval de evaluatie fouten verwaarloosbaar zijn, wordt, voor problemen met een gladde oplossing $u^*$, de fout in essentie bepaald door $h^l$ en $u^{*(l+1)}$ (zie 4.3.7). Hieruit is duidelijk dat de maximaal toelaatbare stapgrootte $h$ nogal per tijdstip kan verschillen (nl. als $u^{*(l+1)}$ dat doet). We zijn dus geïnteresseerd in procedures die telkens of na een aantal rekenstappen nagaan of de stapgrootte $h$ herzien moet worden: in procedures die tijdens het iteratief doorlopen van de multistep rekursie de fout schatten.

Bij de konstruktie van dit soort procedures gaat men uit van de foutvoorstelling in stelling 4.3.7 en neemt men aan dat de rekennauwkeurigheid zo groot is dat de evaluatie fouten verwaarloosbaar zijn ten opzichte van de start– en diskretisatie fouten. Wij zullen dat verder ook doen.

Wil men de stapgrootte $h$ in bijvoorbeeld $\tau_p$ veranderen in $\tilde{h}$, dan kan men, om weer "op gang te komen", een van de volgende drie alternatieve kiezen.

- Start opnieuw in $\tau_p$ (met een of andere startprocedure).
- Bereken de benodigde $u$-waarden in punten $\tau_p - i\tilde{h}$ $(i = 1, \ldots, k-1)$ met bijvoorbeeld Hermite interpolatie uit de berekende $u$- en $f$-waarden in de punten $\tau_p - ih$ $(i \geq 0)$.
- Werk in de eerste nieuwe $k-1$ stappen met een zogenaamde *variabele stapgrootte multistep methode.*

(Beschouw $s_0 < s_1 < \ldots < s_k$ in $\mathcal{J}$. Stel men heeft benaderingen voor $u^*(s_0), \ldots, u^*(s_{k-1})$. Dan kan men $u^*(s_k)$, met behulp van deze benaderingen en de differentiaalvergelijking of de equivalente integraalvergelijking, als volgt benaderen.
Vervang $\frac{d}{dt}u^*(s_k)$ door $\frac{d}{dt}p(s_k)$, waarbij $p$ het Lagrange interpolatie polynoom is voor de benaderende $u^*$ op de steunpunten $s_0, \ldots, s_k$.

Of diskretiseer $\int_{s_j}^{s_k} f(t, u^*(t))\, dt$, voor zekere $j \in \{0, \ldots, k-1\}$, met een interpolatoire of extrapolatoire kwadratuurformule met als steunpunten weer $s_0, \ldots, s_k$.)

**4.4.2 De toelaatbare fout.** Beschouw $t_0 = \tilde{\tau}_0 < \tilde{\tau}_1 < \ldots < \tilde{\tau}_{M+1} = t_0 + T$. Laat $u$ de multistep oplossing zijn waarbij in de berekening, het iteratief doorlopen van (M), eventueel gewerkt zou kunnen zijn met verschillende stapgrootten op de verschillende intervallen $[\tilde{\tau}_j, \tilde{\tau}_{j+1}]$.

Wensen we dat in $\tilde{\tau}_n$ de benadering $u(\tilde{\tau}_n)$ van $u^*(\tilde{\tau}_n)$ een nauwkeurigheid tenminste $\bar{\epsilon}$ heeft dan zal in tijdstippen $\tilde{\tau}_\nu < \tilde{\tau}_n$ de benadering $u(\tilde{\tau}_\nu)$ van $u^*(\tilde{\tau}_\nu)$ een beduidend kleinere fout moeten hebben: bij verdere berekeningen, in de $t_j$ tussen $\tilde{\tau}_\nu$ en $\tilde{\tau}_n$, leveren immers, naast de fout $u^*(\tilde{\tau}_\nu) - u(\tilde{\tau}_\nu)$, ook de lokale fouten in de $t_j$ tussen $\tilde{\tau}_\nu$ en $\tilde{\tau}_n$ hun bijdrage tot de globale fout in de benadering van $u^*(\tilde{\tau}_n)$.
Laat $\mu_{j+1}$ de totale fout zijn die gemaakt is op het interval $[\tilde{\tau}_j, \tilde{\tau}_{j+1})$: als $\tilde{u}^*$ de exacte oplossing van de differentiaalvergelijking is nu met "beginwaarde" $\tilde{u}^*(\tilde{\tau}_j) = u(\tilde{\tau}_j)$ dan
$\mu_{j+1} := \tilde{u}^*(\tilde{\tau}_{j+1}) - u(\tilde{\tau}_{j+1})$.
Neem even aan dat eenmaal gemaakte fouten hoogstens slechts een kumulatieve bijdrage leveren tot de globale fout: neem aan dat $\|u^*(\tilde{\tau}_n) - u(\tilde{\tau}_n)\| \le \sum_{j=1}^{n} \|\mu_j\|$. Een procedure die $\mu_j$ schat en er voor zorgt dat

$$\|\mu_j\| \le \frac{1}{T}(\tilde{\tau}_{j+1} - \tilde{\tau}_j)\bar{\epsilon} \tag{41}$$

levert dan een benadering $u(\tilde{\tau}_n)$ met $\|u^*(\tilde{\tau}_n) - u(\tilde{\tau}_n)\| \le \bar{\epsilon}$.
Helaas worden eenmaal gemaakte fouten min of meer voortgeplant volgens een oplossing van de homogene differentiaalvergelijking $w' = Jw$ (zie (b) van 1.2.17). In het tijdstip $\tilde{\tau}_j$ kunnen we in het algemeen nog niet voorzien hoe die voortplanting er verder ongeveer zal uitzien (tenzij we—met een goedkope methode—al $u^*$ en de oplossing van de vergelijking $w' = Jw$ grof benaderd hebben). Daarom doet men in de praktijk alsof eenmaal gemaakte fouten hoogstens slechts een kumulatieve bijdrage leveren en probeert men ervoor te zorgen dat (41) vervuld is.[12] De (meeste) programma pakketten die in de handleiding beweren de oplossing te produceren met een door de gebruiker opgegeven precisie $\bar{\epsilon}$ gaan zonder enige verdere kontrôle uit van zo'n kumulatieve foutbijdrage of zij adviseren de gebruiker het programma voor verschillende $\bar{\epsilon}$ oplossingen te laten produceren. Uit het verschil in de uitkomsten, zo is de suggestie, kan de gebruiker de wijze waarop de fout wordt voortgeplant schatten. (Kontrôles, waarin getest wordt of de stapgrootte klein genoeg is om de gebruikte schattingsprocedures te rechtvaardigen, of de $\mathcal{O}(h^{l+1})$–term inderdaad verwaarloosbaar is, ontbreken ook vaak. Men hoopt dat een kleinere tolerantie—bv. $\frac{1}{5}\bar{\epsilon}$ in plaats van $\bar{\epsilon}$—ongewenste afwijkingen opvangt.)

In de praktijk schat men gewoonlijk niet de totale fout die gemaakt is op een interval $[\tilde{\tau}_j, \tilde{\tau}_{j+1})$, maar schat men (op gezette tijden $t_n$) de lokale diskretisatie fout $\delta_h(u^*)(t_n)$ (of, zo men wil, de totale fout op het interval $[t_{n+k-1}, t_{n+k})$) en probeert men er voor te zorgen dat $\delta_h(u^*)(t_n) \le \frac{1}{T}\bar{\epsilon}$ (dan is immers $h\delta_h(u^*)(t_n) \le \frac{1}{T}(t_{n+k} - t_{n+k-1})\bar{\epsilon}$).

In de volgende alinea sommen we een paar schattingsmethoden op die gezien onze overwegingen en resultaten in de vorige paragraaf en op grond van onze ervaringen in eerdere numerieke wiskunde colleges voor de hand liggen. Uit efficiëntie overwegingen zijn deze methoden in de praktijk toch niet populair. Alle schattings methoden, ook de populaire, werken dankzij het feit dat de globale fout in essentie glad is (berusten dus op stelling 4.3.7)!

**4.4.3 Eenvoudige schattingsmethoden.** Stel dat (Dif.0-2) gelden.
Beschouw een stabiele multistep van konsistentie orde $l$ en met foutkonstante
$C := \frac{1}{\sigma(1)} C_{l+1}$ (zie 4.3.7). Stel dat de startfouten van $\mathcal{O}(h^{l+1})$ zijn.
• $e_h = u^* - u_h = \frac{1}{2^l - 1}(u_h - u_{2h}) + \mathcal{O}(h^{l+1})$ uniform $(h \to 0)$. Als voor een $h$, klein genoeg, $u_h$ en $u_{2h}$ berekend zijn kan men $e_h$ schatten door $\frac{1}{2^l-1}(u_h - u_{2h})$.

[12] Als het probleem heel goed gekonditioneerd is, is deze aanpak in ieder geval niet zo slecht.

• Door parallel aan de berekening van de $u_h(t_j)$–waarden de differentiaalvergelijking in bewering 4.3.3 benaderend op te lossen levert $e_h \approx h^l C w^*$ een schatting voor $e_h$.
• Beschouw ook een andere stabiele multistep methode van konsistentie orde $l$ met foutkonstante $C_c \neq C$. Als $u_h^c \in C(\mathcal{J}_h, \mathbf{R}^d)$ een benaderende oplossing is geproduceerd met die andere multistep en startfouten van $\mathcal{O}(h^{l+1})$, dan is
$u_h - u_h^c = e_h^c - e_h = (C_c - C)h^l w^* + \mathcal{O}(h^{l+1}) = \frac{C_c - C}{C} e_h + \mathcal{O}(h^{l+1})$ uniform $(h \to 0)$.
Als, voor $h$ klein genoeg, $u_h$ en $u_h^c$ berekend zijn kan men $e_h$ schatten door $\frac{C}{C_c - C}(u_h - u_h^c)$.
• Met een differentie kotiënt kan men ${u^*}^{l+1}(t_n)$ benaderen door een lineaire kombinatie van de $u^*(t_j)$'s, $f(t_j, u^*(t_j))$'s of een kombinatie van beide. Vervangen we deze waarden door de berekende benaderende waarden $u_h(t_j)$ en $f(t_j, u_h(t_j))$ dan hebben we een berekenbare schatting voor ${u^*}^{l+1}(t_n)$ en daarmee voor de lokale diskretisatie fout. Ook hier is het weer essentieel dat de fout glad is.

(*Voorbeeld.* Schrijf $D_h^2(v)(t_n) := v(t_{n+1}) - 2v(t_n) + v(t_{n-1})$ voor $v_h \in C(\mathcal{J}_h, \mathbf{R}^d)$.
Dan is $v''(t_n) - \frac{1}{h^2} D_h^2(v)(t_n) = \mathcal{O}(h^2)$ uniform $(h \to 0)$ voor $v \in C^4(\mathcal{J}, \mathbf{R}^d)$.
Dus bij de trapezium regel kunnen we, als $f$ glad genoeg is, het volgende gebruiken.
Met $f^*(t) := f(t, u^*(t)) = u^{*(1)}$ is $h^2 u^{*(3)} = D_h^2(f^*) + \mathcal{O}(h^4)$ uniform $(h \to 0)$.
Als $u^*$ glad genoeg is, is $u_h = u^* - h^2 \frac{1}{12} w^* + v_h$ met $v_h = \mathcal{O}(h^3)$ (zie 4.3.7) en dus is, met $f_h(t) := f(t, u_h(t))$, $D_h^2(f_h) = D_h^2(f^*) - h^2 \frac{1}{12} D_h^2(Jw^*) + D_h^2(Jv_h) + \mathcal{O}(h^4)$ (zie (11)).
Omdat $v_h = \mathcal{O}(h^3)$ is $D_h^2(Jv_h) = \mathcal{O}(h^3)$.
Omdat $Jw^*$ glad is, is $D_h^2(Jw^*) = h^2 (Jw^*)'' + \mathcal{O}(h^3) = \mathcal{O}(h^2)$.
Blijkbaar $D_h^2(f_h) = h^2 u^{*(3)} + \mathcal{O}(h^3)$ uniform $(h \to 0)$: hiermee is $\frac{1}{12} D_h^2(f_h)(t_n)$ een berekenbare schatting voor de lokale diskretisatie fout in de trapezium regel.
Merk op dat de mededeling dat $e_h = u^* - u_h = \mathcal{O}(h^2)$ uniform $(h \to 0)$ slechts leert dat $D_h^2(Je_h) = \mathcal{O}(h^2)$ uniform en dus slechts dat $D_h^2(f_h) = \mathcal{O}(h^2)$ uniform.)

### 4.4.4 Prediktor–korrektor methoden.

Beschouw een multistep met schema $(\rho_p, \sigma_p)$ en een met schema $(\rho_c, \sigma_c)$. Het paar multistep methoden noemt men een *prediktor–korrektor* methode, of PC-methode, als
(i) beide methoden dezelfde konsistentie orde $l$ hebben,
(ii) de multistep met schema $(\rho_p, \sigma_p)$, de *prediktor*, expliciet is, en
(iii) de multistep met schema $(\rho_c, \sigma_c)$, de *korrektor*, sterk stabiel is.

In de praktijk werkt men met korrektors waarvan de stabiliteitseigenschappen sterker zijn dan die van de prediktor. De sterkere stabiliteitseigenschappen die we in een volgende paragraaf zullen bespreken eisen dat de korrektor impliciet is. Sommige auteurs noemen iedere stabiele konsistente multistep methode een korrektor als hij impliciet is en een prediktor als hij expliciet is.

Laat $C_p$ en $C_c$ de konstanten zijn zodat voor de lokale fout diskretisatie fouten het volgende geldt (zie 4.1.8). Voor iedere $v \in C^{l+2}(\mathcal{J}, \mathbf{R}^d)$ is

$$\begin{aligned}
\delta_h^p(v) &:= \tfrac{1}{h}\rho_p(T_h)(v) - \sigma_p(T_h)(v') = h^l C_p v^{(l+1)} + \mathcal{O}(h^{l+1}) \quad \text{uniform} \quad (h \to 0),\\
\delta_h^c(v) &:= \tfrac{1}{h}\rho_c(T_h)(v) - \sigma_c(T_h)(v') = h^l C_c v^{(l+1)} + \mathcal{O}(h^{l+1}) \quad \text{uniform} \quad (h \to 0).
\end{aligned}$$

**4.4.5 Stelling.** *Stel dat* (Dif.0), (Dif.3) *en* (Start.3)[13] *gelden.*
*Beschouw een prediktor–korrektor methode als in* 4.4.4.
*Laat voor iedere* $h \in \mathbf{H}$, *$h$ klein genoeg,* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *zo zijn dat*

$$\begin{cases} \rho_c(T_h)(u_h) = h\sigma_c(T_h)(f_h) \quad \textit{op} \quad \tilde{\mathcal{J}}_h \quad \textit{met} \quad f_h(t_n) := f(t_n, u_h(t_n)) \\ u_h(t_j) = v_{hj} \quad \textit{voor} \quad j < k. \end{cases}$$

[13] Met $m$ als in (Start.3) is $m > 1$ als $\rho_p \neq \rho_c$.

*Dan*[14]

$$
\begin{aligned}
&\rho_p(T_h)(u_h) - h\sigma_p(T_h)(f_h) \\
&= h^{l+1}\left(C_p - \frac{\sigma_p(1)}{\sigma_c(1)}C_c\right) u^{*(l+1)} + \mathcal{O}(h^{l+2}) \\
&= \left(\frac{C_p}{C_c} - \frac{\sigma_p(1)}{\sigma_c(1)}\right) h\delta_h^c(u^*) + \mathcal{O}(h^{l+2}) \quad \textit{uniform} \quad (h \to 0).
\end{aligned}
$$

*Bewijs.* Volgens 4.3.7 is $u_h = u^* - h^l \frac{1}{\sigma_c(1)} C_c w^* + v_h$ waarbij $v_h = \mathcal{O}(h^{l+1})$ uniform $(h \to 0)$. Merk op dat hierdoor $h\sigma_p(T_h)(Jv_h) = \mathcal{O}(h^{l+2})$ uniform $(h \to 0)$. We vullen dit in en gebruiken 4.1.30 en 4.3.4. Vervolgens vullen we $\sigma_p(T_h)(u^{*(l+1)}) = \sigma_p(1)u^{*(l+1)} + \mathcal{O}(h)$ (zie 4.3.5) in, en tenslotte $\delta_h^p(w^*) = \mathcal{O}(h)$ (konsistentie van de prediktor en gladheid $w^*$). We vinden dan achtereenvolgens

$$
\begin{aligned}
&\rho_p(T_h)(u_h) - h\sigma_p(T_h)(f_h) \\
&= \rho_p(T_h)(u^*) - h\sigma_p(T_h)(u^{*\prime}) - h^l\frac{1}{\sigma_c(1)}C_c\left(\rho_p(T_h)(w^*) - h\sigma_p(T_h)(Jw^*)\right) \\
&\quad + \rho_p(T_h)(v_h) - h\sigma_p(T_h)(Jv_h) + \mathcal{O}(h^{2l+1}) \\
&= h\delta_h^p(u^*) - h^l\frac{1}{\sigma_c(1)}C_c\left(\rho_p(T_h)(w^*) - h\sigma_p(T_h)(\frac{d}{dt}w^*)\right) \\
&\quad - h^{l+1}\frac{1}{\sigma_c(1)}C_c\sigma_p(T_h)(u^{*(l+1)}) + \rho_p(T_h)(v_h) + \mathcal{O}(h^{l+2}) \\
&= h^{l+1}C_p u^{*(l+1)} - h^{l+1}\frac{1}{\sigma_c(1)}C_c\delta_h^p(w^*) - h^{l+1}\frac{\sigma_p(1)}{\sigma_c(1)}C_c u^{*(l+1)} + \rho_p(T_h)(v_h) + \mathcal{O}(h^{l+2}) \\
&= h^{l+1}(C_p - \frac{\sigma_p(1)}{\sigma_c(1)}C_c)u^{*(l+1)} + \rho_p(T_h)(v_h) + \mathcal{O}(h^{l+2}) \quad \text{uniform} \quad (h \to 0).
\end{aligned}
$$

Als boven volgt (neem de prediktor even gelijk aan de korrektor. Dan $C_c = C_p, \ldots$)

$$0 = \rho_c(T_h)(u_h) - h\sigma_c(T_h)(f_h) = \rho_c(T_h)(v_h) + \mathcal{O}(h^{l+2}).$$

Blijkbaar is $\rho_c(T_h)(v_h) = \mathcal{O}(h^{l+2})$. De stelling volgt nu onmiddellijk voor het geval dat $\rho_p = \rho_c$.
Als $v_h = h^{l+1}w_1 + \mathcal{O}(h^{l+2})$ uniform $(h \to 0)$ voor een $w_1 \in C^1(\mathcal{J}, \mathbf{R}^d)$ (zie (a) in 4.3.18) dan volgt uit het feit dat $\rho_p(1) = 0$ dat $\rho_p(T_h)(v_h) = h^{l+1}\rho_p(T_h)(w_1) + \mathcal{O}(h^{l+2}) = \mathcal{O}(h^{l+2})$.
□

**4.4.6 Opmerkingen.** De stelling is ook korrekt in geval de korrektor zwak stabiel is en de essentiële parasitaire wortels van $\rho_c$ ook wortels van $\rho_p$ zijn ($\rho_p(\lambda) = 0$ als $|\lambda| = 1$ en $\rho_c(\lambda) = 0$; voor een sterk stabiele korrektor is deze eis triviaal). In geval van een zwak stabiele korrektor met zo'n *aangepaste* prediktor moeten we in het laatste argument van het bewijs van de stelling gebruik maken van (b) in 4.3.18.

De eerste uitspraak $\rho_p(T_h)(u_h) - h\sigma_p(T_h)(f_h) = h^{l+1}[C_p - \frac{\sigma_p(1)}{\sigma_c(1)}C_c]u^{*(l+1)} + \mathcal{O}(h^{l+2})$ in de stelling is ook korrekt als (i) in de definitie van de PC-methode niet geldt maar we hier wel toestaan dat $C_p = 0$ of $C_c = 0$: neem dan voor $l$ het minimum van de orde van de prediktor en de orde van de korrektor.

De stelling heeft twee interessante toepassingen. De kombinatie van de twee toepassingen maakt de prediktor–korrektor methode voor de praktijk zo interessant. De eerste toepassing verklaart de naam "prediktor–korrektor".

[14]Als $\rho_p \neq \rho_c$ is, voor iedere $\varepsilon > 0$, de orde uitspraak uniform op $[t_0 + \varepsilon, t_0 + T]$ (zie 4.3.18 en de voetnoot daarbij).

**4.4.7 Toepassing.** Beschouw de situatie als in bovenstaande stelling 4.4.5 [15].
Stel dat $u_h(t_j)$ berekend is voor $j < n+k$. We wensen $u_h(t_{n+k})$ te bepalen. Als de (korrektor) methode impliciet is (zoals we zullen zien is dat het geval voor iedere methode met aantrekkelijke stabiliteitseigenschappen) moeten we $u_{n+k}$ oplossen uit een mogelijk niet–lineaire vergelijking. Met een succesief substitutie proces kunnen we $u_{n+k}$ in iedere gewenste nauwkeurigheid benaderen (zie 4.1.32 en 3.1.13). Het scheelt werk (iteratie slagen) als we goedkoop een goede start voor het succesief substitutie proces kunnen vinden. De prediktor levert zo'n start.

Bereken $u^p_{n+k}$ zodat met $u_j = u_h(t_j)$ en $f_h(t_j) = f(t_j, u_h(t_j))$ geldt

$$u^p_{n+k} + \alpha^p_1 u_{n+k-1} + \ldots + \alpha^p_k u_n = h\beta^p_1 f(t_{n+k-1}, u_{n+k-1}) + \ldots + h\beta^p_k f(t_n, u_n) \qquad \text{(P)}$$

($\alpha^p_j$ de koëfficiënten van $\rho_p$ en $\beta^p_j$ die van $\sigma_p$; bedenk dat $\alpha^p_0 = \alpha^c_0 = 1!$): een multistep slag met de prediktor waarbij de waarden uit de korrektor gebruikt worden. Tijdens de hele berekening bepalen we de $u_h(t_j)$ en onthouden deze waarden; de $u^p_{n+k}$ spelen slechts een rekentechnische rol en worden na gebruik onmiddellijk "vergeten". Omschrijven van (P) leert dat

$$u^p_{n+k} - u_{n+k} + \rho_p(T_h)(u_h)(t_n) - h\sigma_p(T_h)(f_h)(t_n) = 0.$$

Stelling 4.4.5 vertelt ons nu dat uniform voor $h \to 0$ geldt

$$u^p_{n+k} - u_{n+k} = -\left(\frac{C_p}{C_c} - \frac{\sigma_p(1)}{\sigma_c(1)}\right) h\delta^c_h(u^*) + \mathcal{O}(h^{l+2}) = \mathcal{O}(h^{l+1}). \qquad (42)$$

Blijkbaar is de gemakkelijk te berekenen $u^p_{n+k}$—bedenk dat de $f(t_{n+j}, u_{n+j})$ voor $j < k$ al in een eerdere stap uitgerekend zijn—een uitstekende start voor het succesief substitutie proces dat $u_{n+k}$ moet benaderen.

We gaan als volgt te werk.
Stel $\tilde{u}_{n+j}$ en $\tilde{f}_{n+j}$ zijn voor $j < k$ berekend. We berekenen dan $\tilde{u}_{n+k}$ en $\tilde{f}_{n+k}$ als volgt.

$$u^{(0)}_{n+k} := \textstyle\sum_{j=1}^k (h\beta^p_j \tilde{f}_{n+k-j} - \alpha^p_j \tilde{u}_{n+k-j}); \qquad \text{(P)}$$

$$i = 0;\ y_n := \textstyle\sum_{j=1}^k (h\beta^c_j \tilde{f}_{n+k-j} - \alpha^c_j \tilde{u}_{n+k-j});$$

$$f^{(i)}_{n+k} := f(t_{n+k}, u^{(i)}_{n+k}); \qquad \text{(E)}$$

$$u^{(i+1)}_{n+k} := h\beta^c_0 f^{(i)}_{n+k} + y_n; \qquad \text{(C)}$$

$$i := i+1; \quad \texttt{ga naar (E)}$$

Als $\tilde{u}_{n+j} = u_{n+j}$, $\tilde{f}_{n+j} = f(t_{n+j}, u_{n+j})$ voor $j < k$ dan $u^{(0)}_{n+k} := u^p_{n+k}$ en $u^{(i)}_{n+k} - u_{n+k} = \mathcal{O}(h^{l+1+i})$ (zie 3.1.13). De prediktor levert een nauwkeurige start. Met de korrektor korrigeren we tot de gewenste precisie bereikt is. Met $i = 1$ of $i = 2$ is $u^{(i)}_{n+k}$ al een voldoend nauwkeurige benadering (zie 4.1.33).

Kies een $N \in \mathbf{N}$.
Werken we telkens, ook in verdere berekeningen, met de $N$-de iteranden— $\tilde{u}_{n+k} = u^{(N)}_{n+k}$ en $\tilde{f}_{n+k} = f^{(N)}_{n+k}$ —en niet met de beoogde $u_{n+k}$ en $f(t_{n+k}, u_{n+k})$ dan spreekt men over een P(EC)$^{\mathrm{N}}$E-methode: men stopt de succesieve substitutie na de evaluatie van de $f$-waarde en men werkt verder met de $u$- en $f$-waarden die op dat moment berekend zijn.
Werken we telkens, ook in verdere berekeningen, met $\tilde{u}_{n+k} = u^{(N)}_{n+k}$ en $\tilde{f}_{n+k} = f^{(N-1)}_{n+k}$ dan spreekt men over een P(EC)$^{\mathrm{N}}$-methode: men stopt de succesieve substitutie na de korrektie stap.[16]

[15] Alleen om notationele redenen nemen we hier aan dat zowel de prediktor als de korrektor een $k$-staps methode is. De beweringen zijn evenzo korrekt voor het geval het aantal termen in de prediktor rekursie verschilt van die in de korrektor.

[16] Men moet hierbij in de gaten houden dat de stabiliteitseigenschappen van deze P(C)$^{\mathrm{N}}$-methoden niet die van de korrektor methode zullen zijn; naarmate $N$ groter is zullen die wel, naar verwachting, op elkaar gaan lijken.

**4.4.8** ∘ **Opgave.** Laat $\tilde{u}_n$ berekend zijn met de $\mathrm{P(EC)}^{\mathrm{N}}$E-methode. Bewijs dat

$$\tilde{u}_n = u^*(t_n) - h^l \frac{1}{\sigma_c(1)} C_c w^*(t_n) + \mathcal{O}(h^{l+1}) \quad \text{uniform} \quad (h \to 0).$$

In de vierde methode in 4.4.3 hebben we aangegeven dat we door middel van lineaire kombinaties van $u_h(t_j)$'s en $f(t_j, u_h(t_j))$'s de lokale diskretiesatie fout kunnen schatten. In de volgende toepassing geven we een speciale lineaire kombinatie, die we—vanwege het werk dat al verricht is—bijzonder efficiënt kunnen evalueren.

**4.4.9 Toepassing.**
Laat $u_{n+k}^p$ en $u_{n+k}$ berekend zijn als in 4.4.7. Stel dat $\frac{1}{\sigma_c(1)} C_c \neq \frac{1}{\sigma_p(1)} C_p$.
Dan volgt uit (42) in 4.4.7 dat

$$h\delta_h^c(u^*)(t_n) = \frac{\sigma_c(1)C_c}{\sigma_c(1)C_p - \sigma_p(1)C_c}(u_{n+k}^p - u_{n+k}) + \mathcal{O}(h^{l+2}).$$

Hiermee hebben we een berekenbare schatting voor de lokale diskretisatie fout $\delta_h^c(u^*)$ in de korrektor multistep.

Berekenen we de $u_{n+k}$ middels de prediktor–korrektor techniek uit 4.4.7 dan vergt deze schatting vrijwel geen extra werk (slechts 1 flop). Merk op dat we een schatting met dezelfde orde van nauwkeurigheid krijgen als we in bovenstaande schatting voor $h\delta_h^c(u^*)(t_n)$ de waarde $u_{n+k}$ door $u_{n+k}^{(i)}$ vervangen met $i > 0$.

De programma pakketten die pretenderen de oplossing in de gewenste nauwkeurigheid te berekenen schatten de lokale diskretiesatie fout in de korrektor methode door op de aangegeven manier het korrektor resultaat te vergelijken met het resultaat uit zo'n lokale prediktor stap. Ze passen eventueel de stapgrootte aan of, mocht die te klein of te groot worden, de orde (ze stappen over op een prediktor–korrektor methode met een andere konsistentie orde); ze besturen dus automatisch de stapgrootte én de orde[17].
De schattings methode is theoretisch gefundeerd in 4.4.5 onder de aanname dat de globale fout glad is. In onze stelling over de gladheid van de fout namen we onder meer aan aan dat we met dezelfde stapgrootte en dezelfde multistep werkten (de stapgrootte $h$ moest ook nog klein genoeg zijn). Het is maar de vraag of we deze gladheid ook nog hebben als we vaak van stapgrootte of van methode veranderen. Veranderen we niet van stapgrootte binnen $k$ stappen dan zijn er wel theoretische argumenten aan te geven waarom de aanpak toch verantwoord is; we werken dat hier niet verder uit.

## 4.5 Multistep methoden op een half-oneindig tijdsinterval

Tot nu toe namen we aan dat $\mathcal{J}$ een begrensd tijdsinterval was. In deze paragraaf bekijken we de situatie waarin $\mathcal{J}$ het half-oneindig tijdsinterval $[t_0, \infty)$ is. We nemen dus hier aan dat (1) precies een oplossing $u^*$ heeft op $\mathcal{J} = [t_0, \infty)$ met grafiek in $\Omega$.

We zouden ook in deze situatie graag zien dat de multistep oplossingen $u_h$ voor $h \to 0$ uniform op $[t_0, \infty)$ konvergeren naar de exacte oplossing $u^*$. Zo'n uniforme konvergentie volgt niet uit de voorgaande resultaten door die toe te passen op $[t_0, t_0 + T]$ en vervolgens de limiet te nemen voor $T \to \infty$. Ze volgen zelfs niet als de afgeleiden $u^{*(j)}$ voor $j \leq l$ bestaan en op $[t_0, \infty)$ uniform begrensd zijn. De foutschatting in bijvoorbeeld (30) bevat naast $\sup_{t \in \mathcal{J}} \|u^{*(l+1)}(t)\|$ ook de konstante $T$. Voor niet al te grote $T$ kan de $K'$ uit 4.1.24 al onbehoorlijk groot zijn. Deze $T$ en $K'$ zit ook in de $\mathcal{O}(h^{l+1})$–term in 4.3.7: het is dan ook niet duidelijk of we zo'n uniforme $\mathcal{O}(h^{l+1})$–term hebben voor het geval $\mathcal{J} = [t_0, \infty)$.

[17] De pakketten werken met prediktor-korrektor methoden van orde 1,2,…,6. De korrektors zijn de BDF-methoden. Ze starten met kleine stapgrootte en met de methode van orde 1 (Euler forward-Euler backward) en al itererend voeren ze de orde op en vergroten ze de stapgrootte; ze hebben dus geen extra startprocedure nodig.

We behandelen hier deze problematiek niet uitputtend: we geven voorwaarden aan waaronder de $u_h$'s uniform konvergeren, we geven geen uniforme foutschattingen. Door te werk te gaan als in het bewijs van stelling 4.3.7 kan men, onder zekere voorwaarde, met behulp van het aangekondigde uniforme konvergentie resultaat, bewijzen dat weer op een uniforme $\mathcal{O}(h^{l+1})$-term na de globale fout glad is (zie stelling 4.5.6).

In de praktijk zal men uiteraard niet door itereren tot $\infty$. Wel gaat men vaak itereren zonder de eindtijd te specificeren: men plot bijvoorbeeld al rekenend de oplossing en stopt het rekenproces pas als men een goede indruk heeft over het verloop van de oplossing. Zonder eindtijd $t_0+T$ kan men op grond van de foutschattingen die we tot dusver gezien hebben niet beslissen met welke stapgrootte $h$ men moet rekenen. Alle schattingen tot dusver verlangen een kleinere stapgrootte $h$ bij een grotere $T$. We weten zelfs nog niet of, als de iteratie lang duurt, de numerieke oplossing üperhaupt wel kan lijken op de exacte.

### Totaal stabiliteit

Als het probleem zelf iedere fout onbeperkt laat groeien (zie voorbeeld 1.2.1) dan maken we met geen enkele numerieke methode kans de exacte oplossing uniform op $[t_0,\infty)$ te kunnen benaderen. We hebben hier echter het sterkere goed gekondioneerdheids begrip nodig uit **B** van 1.2.2: een kumulatie van lokale perturbaties zoals we dat in **A** van 1.2.2 toestonden is op een half-oneindig tijdsinterval ontoelaatbaar.

**4.5.1 Definitie.** $u^*$ is een *totaal stabiele* oplossing van (1) als er voor iedere $\bar{\varepsilon}>0$ een $\bar{\delta}>0$ is zodat het geperturbeerd probleem (7) een oplossing $\tilde{u}$ heeft en er geldt

$$\sup_{t\in\mathcal{J}}\|u^*(t)-\tilde{u}(t)\|\leq\bar{\varepsilon}$$

zodra $\|\delta_0\|+\sup_t\|\delta(t)\|\leq\bar{\delta}$.

Als er een $C>0$ is zodat we, voor iedere $\bar{\varepsilon}$ die voldoende klein is, $\bar{\delta}\geq\frac{1}{C}\bar{\varepsilon}$ kunnen kiezen dan heeft het probleem een eindig *sterk konditie getal* (zie **B** van 1.2.2).

Totaal stabiliteit verwoordt het feit dat het beginwaarde probleem *goed gesteld* is t.o.v. de normen $\sup_t\|v(t)\|$: de oplossing hangt t.o.v. deze normen kontinu af van de startwaarde $u_0$ en de funktie $f$. Het sterke konditie getal vertelt hoe de maximale lokale fout doorwerkt in de globale fout.

**4.5.2 Opmerkingen.**
De definitie kan ook weer geformuleerd worden voor minder gladde perturbaties (zie 1.2.4).

In de definitie gaan we er weer impliciet van uit dat de gewichtsfunktie $\omega\equiv 1$ het adekwate gewicht is om fouten mee te wegen (zie de diskussie in 1.2.5).

In de meeste resultaten in paragraaf 2 is bewering 1.2.6 van toepassing. Deze bewering vertelt onder meer dat probleem (1) een sterk konditie getal $\leq C(1+\frac{1}{|\mu|})$ heeft als $\mu<0$.

Als $f\in C^2(\Omega)$ en alle tweede orde afgeleiden van $f$ uniform begrensd zijn op de $\tilde{r}$-buis rond de grafiek van $u^*$ dan is het sterke konditie getal van probleem (1) ook gelijk aan dat van het gelineairiseerde probleem (12) (zie 1.2.16. $\circ$ Dit kan men bewijzen, door te werk als in het bewijs van 1.2.16, nu met $\bar{a}:=\sup_\tau a(\tau)\,d\tau<\infty$).

### Zwak stabiel multistep methoden

Niet iedere stabiele konsitentie multistep methode produceert numerieke oplossingen die konvergeren naar de exacte oplossing zelfs niet als die totaal stabiel is. We geven een voorbeeld.

**4.5.3 Voorbeeld.** Beschouw, met $d=1$, het probleem

$$u'(t)=-u(t)\quad\text{voor}\quad t\in[0,\infty)\quad\text{en}\quad u(0)=0.$$

Dit probleem is totaal stabiel (met sterk konditie getal 2; zie 1.2.12).

We lossen de totaal stabiele oplosing $u^* \equiv 0$ benaderend op met behulp van de midpunt regel met stapgrootte $h$. De fout $e_h$ voldoet dan, met $e_n = e_h(t_n)$ aan de rekursie

$$e_{n+1} = e_{n-1} - 2he_n \quad \text{voor} \quad n \in \mathbf{N}$$

We nemen aan dat $e_0 = 0$, maar dat $e_1 \neq 0$. Dan is

$$e_n = e_1 \frac{\lambda_1(h)^n - \lambda_2(h)^n}{\lambda_1(h) - \lambda_2(h)} \quad \text{met} \quad \lambda_i(h) = -h - (-1)^i \sqrt{1+h^2} \quad (i = 1, 2).$$

Omdat $\lambda_1(h) < 1$ gaat $\lambda_1(h)^n \to 0$ als $n \to \infty$ ( $\lambda_1(h)^n = e^{-t_n}(1 + \mathcal{O}(h))$ uniform $(h \to 0)$, volgens verwachting). Echter $-\lambda_2(h) > 1 + h$ en $(-\lambda_2(h))^n > (1+h)^n \to \infty$ als $n \to \infty$ ( $(-\lambda_2(h))^n > e^{t_n(1-h)}$ voor alle $h < 1$ en alle $n \in \mathbf{N}$ ). Blijkbaar $(-1)^{n-1} e_n \approx \frac{1}{2} e^{t_n} e_1$.

Als, in een lineair probleem of in het gelineairiseerde probleem (12), voor zekere $\mu < 0$ en $C > 0$ geldt $\|G(t,\tau)\| \leq Ce^{\mu(t-\tau)}$ voor alle $t, \tau \in \mathcal{J}$, $t \geq \tau$ (zie 1.2.10) dan wordt iedere lokale fout en iedere startfout door de differentiaalvergelijking gedempt voortgeplant (zie 1.2.11 en (b) van 1.2.17). Parasitaire wortels op de eenheidscirkels kunnen, gezien het bovenstaand voorbeeld, het exponentiëel uitdempend karakter van fouten in het kontinue probleem om zeep helpen en in het numeriek proces zelfs omzetten in een exponentiële groei van lokale fouten zodat op den duur de fouten de ware oplossing overvleugelen. Met zwak stabiele methoden verwachten we niet dat we oplossingen op half-oneindige tijdsintervallen uniform kunnen benaderen.

### Sterk stabiele multistep methoden

In deze subparagraaf laten we zien dat sterk stabiele multistep methoden wel in staat zijn een totaal stabiele oplossing $u^*$ uniform te benaderen, in geval $f$ uniform Lipschitz kontinu is in beide variabelen en uniform begrensd op de $\tilde{r}$-buis rond de grafiek van $u^*$. Om didaktische redenen bewijzen we onderstaand konvergentie resultaat eerst voor Adams methoden.

**4.5.4 Stelling.** *Stel dat $u^*$ een totaal stabiele oplossing is van* (1).
*Stel dat* (Dif.0) *en* (Dif.1$'$) *gelden, waarbij, met* $\Omega_{\tilde{r}} := \{(t,x) \,\big|\, \|u^*(t) - x\| \leq \tilde{r}\}$,

(Dif.1$'$) $\quad \exists L > 0$ *zodat* $\|f(t,x) - f(s,y)\| \leq L \max(|t-s|, \|x-y\|) \qquad ((t,x),(s,y) \in \Omega)$
$\qquad\qquad \exists M > 0$ *zodat* $\|f(t,x)\| \leq M \qquad ((t,x) \in \Omega_{\tilde{r}})$

*Stel dat de $k$-staps multistep konsistent en sterk stabiel is.*
*Voor* $h \in \mathbf{H}$ *is* $v_{h\,0}, \ldots, v_{h\,k-1} \in \mathbf{R}_d$ *zodat* $v_{h\,j} = u^*(t_0) + \mathcal{O}(h)$ *voor* $j < k$ *en* $h \to 0$.
*Dan is er voor iedere* $h \in \mathbf{H}$, *die voldoende klein is, een* $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *waarvoor*

$$\begin{cases} \rho(T_h)(u_h) = h\sigma(T_h)(f_h) & \text{op} \quad \tilde{\mathcal{J}}_h \quad \text{met} \quad f_h(t) = f(t, u_h(t)) \\ u_h(t_j) = v_{h\,j} & \text{voor} \quad j < k \end{cases}$$

*en er geldt dat* $\lim_{h \to 0} \sup_{t_n \geq t_0} \|u_h(t_n) - u^*(t_n)\| = 0$.

*Bewijs voor het geval* $\rho = \chi^k - \chi^{k-1}$ (een Adams methode).

Neem eerst aan dat $u_h$ bestaat met grafiek in $\Omega_{\tilde{r}}$.
Met $u_n := u_h(t_n)$ is $u_{n+k} = u_{n+k-1} + h\sigma(T_h)(f_h)(t_n) \quad (n \geq 0)$
en uit (Dif.1$'$) volgt dat $\|u_{n+k} - u_{n+k-1}\| \leq h\tilde{\sigma}LM$. Definiëer

$$\begin{cases} z(t) := u_h(t_{n+k-1}) + (t - t_{n-1})\sigma(T_h)(f_h)(t_n) & \text{voor} \quad t \in [t_{n+k-1}, t_{n+k}] \\ z(t) := u^*(t) - u^*(t_{k-1}) + v_{h\,k-1} & \text{voor} \quad t \in [t_0, t_{k-1}]. \end{cases}$$

$z$ is kontinu, links differentieerbaar en $z(t_n) = u_h(t_n)$ voor iedere $n \geq k-1$.
Merk op dat

$$\|u_h(t_{n+j}) - z(t)\| \leq hk\tilde{\sigma}LM \quad \text{als} \quad |t - t_{n+j}| \leq kh.$$

Verder is (met $z'(t)$ de links afgeleide; zie 1.2.4)

$$z'(t) = f(t, z(t)) + \delta(t) \quad \text{voor} \quad t \geq t_0,$$

waarbij

$$\begin{aligned} \delta(t) &:= \sigma(T_h)(f_h)(t_n) - f(t, z(t)) && \text{voor} \quad t \in (t_{n+k-1}, t_{n+k}] \quad \text{en} \\ \delta(t) &:= f(t, u^*(t)) - f(t, z(t)) && \text{voor} \quad t \in [t_0, t_{k-1}]. \end{aligned}$$

Omdat $\sigma(1) = \rho'(1) = 1$ is
$$\|\delta(t)\| = \|\sigma(T_h)(f_h - f(t, z(t))(t_n)\| \leq \tilde{\sigma} L(hk\tilde{\sigma}LM) \quad \text{voor} \quad t \geq t_{k-1}$$
$$\|\delta(t)\| \leq L\|u^*(t) - z(t)\| = L\|u^*(t_{k-1}) - v_{h\,k-1}\| \quad \text{voor} \quad t \leq t_{k-1}.$$
Tenslotte is $\|z(t_0) - u^*(t_0)\| = \|u^*(t_{k-1}) - v_{h\,k-1}\|$.
Zij $\bar{\varepsilon} \in (0, \frac{1}{2}\tilde{r})$. Laat $\bar{\delta}$ zijn als in definitie 4.5.1 van totaal stabiliteit (die we nu toepassen met links kontinue $\delta$ en links afgeleiden; zie 4.5.2 en 1.2.4). We zien dat, door $h$ maar klein genoeg te nemen, $\|z(t_0) - u^*(t_0)\| + \sup_{t \geq t_0} \|\delta(t)\| \leq \bar{\delta}$. Dus $\sup_t \|z(t) - u^*(t)\| \leq \bar{\varepsilon}$ en in het bijzonder is $\sup_{t_n} \|u_h(t_n) - u^*(t_n)\| \leq \bar{\varepsilon}$

Tot dusver namen we aan dat $u_h$ bestaat en grafiek in $\Omega_{\tilde{r}}$ heeft. Met een induktie argument, zoals we dat in het bewijs van de stelling van Kreiss gegeven hebben, volgt de korrektheid van deze aanname. We laten deze details aan de lezer over.

Voordat we de stelling kunnen bewijzen voor een willekeurige sterk stabiele multistep methode herformuleren we eerst een "stabiliteitsresultaat" voor inhomogene rekursies met konstante koëfficiënten waarvan de karakteristieke wortels binnen de eenheidscirkel liggen.

**4.5.5 Lemma.** *Zij* $\bar{\rho} = \bar{\alpha}_0\chi^{k-1} + \bar{\alpha}_1\chi^{k-2} + \ldots + \bar{\alpha}_{k-1}$ *een polynoom van graad* $k-1$ *waarvan alle wortels in absolute waarde strikt kleiner dan 1 zijn. Dan is er een* $\bar{K} > 0$ *zodat voor iedere* $v_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ *geldt*

$$\sup_{t_n \geq t_0} \|v_h(t_n)\| \leq \bar{K}\left(\max_{j<k-1} \|v_h(t_j)\| + \sup_{t_n \geq t_0} \|\bar{\rho}(T_h)(v_h)(t_n)\|\right).$$

*Bewijs.* Beschouw een $v_h \in C(\mathcal{J}_h, \mathbf{R}^d)$. Met $\theta_n := \bar{\rho}(T_h)(v_h)(t_n)$ en $v_n := v_h(t_n)$ kan $(v_n)$ gezien worden als de oplossing van de volgende inhomogene rekursie met konstante koëfficiënten.
$$\bar{\alpha}_0 v_{n+k-1} + \ldots + \bar{\alpha}_{k-1} v_n = \theta_n \qquad (n \in \mathbf{N}_0).$$
Het lemma volgt nu onmiddellijk uit (b) van 3.1.15. □

*Vervolg bewijs van de stelling.* Zij $\bar{\rho} := \frac{\rho}{\chi - 1}$. Merk op dat $\bar{\rho}(1) = \rho'(1) = \sigma(1)$. Omdat de methode sterk stabiel is liggen alle wortels $\lambda \neq 1$ van $\rho$ binnen de eenheidscirkel. Omdat dit precies de wortels van $\bar{\rho}$ zijn heeft $\bar{\rho}$ alle wortels binnen de eenheidscirkel.
Voor het gemak nemen we hier aan dat de multistep zo geschaald is dat $\sigma(1) = 1$.

We nemen weer aan dat $u_h$ bestaat met grafiek in $\Omega_{\tilde{r}}$.
Beschouw nu $\bar{u}_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ die gedefinieerd is door $\bar{u}_h(t_n) := \bar{\rho}(T_h)(t_n)$. Merk op dat dan

$$\bar{u}_h(t_{n+1}) - \bar{u}_h(t_n) = \rho(T_h)(u_h)(t_n) = h\sigma(T_h)(f_h)(t_n) \quad \text{voor alle} \quad t_n \in \tilde{\mathcal{J}}_h.$$

We bewijzen dat $v_h := \bar{u}_h - u_h = \mathcal{O}(h)$ uniform $(h \to 0)$: dan kan men de stelling verder bewijzen door het bewijs voor de Adams methoden te kopiëren
$(z(t) := \bar{u}_h(t_n) + (t - t_n)\sigma(T_h)(f_h)(t_n)$, etc.. We laten deze details verder aan de lezer over).

Uit (Dif.1′) volgt weer $\|\bar{u}_h(t_{n+1}) - \bar{u}_h(t_n)\| \leq h\tilde{\sigma}LM$.
Zij $\theta_h := \bar{\rho}(T_h)(v_h) = \bar{\rho}(T_h)(\bar{u}_h - u_h) = \bar{\rho}(T_h)(\bar{u}_h) - \bar{u}_h$. Gebruiken we het feit dat $\bar{\rho}(1) = 1$ dan zien we dat

$$\begin{aligned}\|\theta_h(t_n)\| &= \|\bar{\rho}(T_h)(\bar{u}_h)(t_n) - \bar{u}_h(t_n)\| = \|\bar{\rho}(T_h)(\bar{u}_h - \bar{u}_h(t_n))(t_n)\| \\ &\leq \sum_{j=0}^{k-1} |\bar{\alpha}_j| \max_{j<k} \|\bar{u}_h(t_{n+j}) - \bar{u}_h(t_n)\| \leq \sum_{j=0}^{k-1} |\bar{\alpha}_j| kh\tilde{\sigma}LM = \mathcal{O}(h).\end{aligned}$$

Met andere woorden $\theta_h = \mathcal{O}(h)$ uniform $(h \to 0)$.
Ook omdat $\bar{\rho}(1) = 1$ geldt voor $j < k-1$ dat
$$\bar{u}_h(t_j) = \bar{u}_h(t_0) + \mathcal{O}(h) = \bar{\rho}(T_h)(u_h)(t_0) + \mathcal{O}(h) = u^*(t_0) + \mathcal{O}(h) \quad \text{als} \quad h \to 0.$$
Dus $\bar{u}_h(t_j) - u_h(t_j) = \mathcal{O}(h)$ voor $h \to 0$. Uit bovenstaand lemma volgt nu dat
$\sup_{t_n} \|v_h\| \leq \bar{K}(\max_{j<k-1} \|v_h(t_j)\| + \sup_{t_n} \|\theta_h(t_n)\|) = \mathcal{O}(h)$ uniform $(h \to 0)$. □

**4.5.6 Stelling.** *Beschouw de situatie in stelling* 4.5.4.
*Stel bovendien dat het probleem* (1) *een sterk eindig konditie getal heeft, alle tweede orde afgeleiden van* $f$ *op* $\Omega_{\tilde{r}}$ *uniform begrensd zijn,* $u^* \in C^{l+1}(\mathcal{J}, \mathbf{R}^d)$ *en de* $l+1$*-ste afgeleide van* $u^*$ *uniform begrensd is. Als de multistep konsistentie orde* $l$ *heeft dan*

$$\sup_{t_n} \|h^{-l} e_h(t_n) - \frac{1}{\sigma(1)} C_{l+1} w^*(t_n)\| \to 0 \quad \textit{als} \quad h \to 0.$$

$\circ$ *Bewijs.* Het probleem $w'(t) = J(t)w(t) + u^{*l+1}(t), \quad w(t_0) = 0$ heeft een sterk konditie getal $C < \infty$ (zie 4.5.2). $h^{-l} \frac{\sigma(1)}{C_{l+1}} e_h$ kan gezien worden als de oplossing van de multistep toegepast op deze lineaire differentiaalvergelijking (zie 4.3.7). Uit het bewijs van bovenstaande stelling volgt, de bewering met behulp van $C$; we laten de details over aan de geïnteresseerde lezer. $\square$

## 4.6 Stabiliteit van multistep methoden bij grotere stapgrootte

De resultaten in de vorige paragrafen zijn korrekt onder de voorwaarde dat de stapgrootte $h$ klein genoeg is (zie 4.3.7 en 4.5.6) of ze zijn onder omstandigheden nogal grof (zie 4.1.31, 4.3.1 en 4.3.10). Ze zijn met name grof in geval een lineair probleem of het gelineairiseerde probleem iedere lokale fout en iedere startfout exponentieel dempend voortplant en de multistep sterk stabiel is. In deze paragraaf gaan we na of we ook met grotere stapgrootte aan de slag kunnen en onder welke voorwaarden dat kan.

In het numeriek proces maken we telkens lokale fouten. We beschrijven eerst hoe lokale fouten in het numeriek proces worden voortgeplant: we generalizeren de beschrijving in 4.3.14 en geven de diskrete variant van (10).

**4.6.1 De foutvoortplanting in een multistep.** Beschouw, voor $h \in \mathbf{H}$, de berekende multistep oplossing $u_h^* \in C(\mathcal{J}_h, \mathbf{R}^d)$. Laat $\tilde{e}_h := u^* - u_h^*$ de totale fout zijn. We zijn geïnteresseerd in de situatie waarin $\sup_{t_n} \|\tilde{e}_h(t_n)\|$ klein is. Dan voldoet $\tilde{e}_h$ aan

$$\rho(T_h)(\tilde{e}_h)(t_n) = h\sigma(T_h)(J\tilde{e}_h)(t_n) + \mu_n \qquad \text{voor iedere} \qquad t_n \in \tilde{\mathcal{J}}_h,$$

waarbij $\mu_n$ de lokale fout is die gemaakt is bij de berekening van $u_h(t_{n+k})$ ($\mu_n = h\delta_n - \epsilon_n + \mathcal{O}(\|\tilde{e}_h(t_n)\|^2)$ met $\delta_n$ de lokale diskretisatie fout en $\epsilon_n$ de lokale evaluatie fout).

Laat $G_h$ de funktie zijn van $\mathcal{J}_h \times \mathcal{J}_h$ naar $\mathbf{M}_d$ waarvoor, voor iedere $t_\nu \in \mathcal{J}_h$, geldt dat

$$\begin{aligned} &G_h(t_j, t_\nu) = 0 \quad \text{voor} \quad j < \nu, \quad G_h(t_\nu, t_\nu) = I \\ &\textstyle\sum_{j=\max(0,k-n)}^{k} \Big(\alpha_{k-j} - h\beta_{k-j} J(t_{n-k+j})\Big) G_h(t_{n-k+j}, t_\nu) = 0 \quad \text{voor} \quad n > \nu. \end{aligned}$$

Hierbij is 0 de triviale matrix in $\mathbf{M}_d$ en $I$ de eenheids matrix. We veronderstellen dat de matrices $\alpha_0 - h\beta_0 J(t_{n+k})$ inverteerbaar zijn: als ze dat niet zijn dan zal $u_h(t_{n+k})$, gegeven $u_h(t_n), \ldots, u_h(t_{n+k-1})$, niet oplosbaar zijn uit de multistep rekursie.

Dan geldt

$$\tilde{e}_h(t_n) = \sum_{\nu=0}^{n} G_h(t_n, t_\nu)\, (\alpha_0 - h\beta_0 J(t_\nu))^{-1} \mu_{\nu-k}, \qquad \text{voor iedere} \qquad t_n \in \mathcal{J}_h,$$

waarbij de $\mu_{-i} := \sum_{j=i}^{k} (\alpha_{k-j} - h\beta_{k-j} J(t_{j-i})) \tilde{e}_h(t_{j-i}) \quad (i = 1, \ldots, k)$ de bijdrage van de startfouten beschrijven.
(Ga dit na. Vergelijk dit met de konstruktie in 1.2.9 en 4.3.14: $G_h(t_n, t_\nu) = G_{n-\nu}$.)
De lokale fout $(\alpha_0 - h\beta_0 J(t_\nu))^{-1} \mu_{\nu-k}$ gemaakt bij de berekening van $u_h(t_\nu)$, vinden we terug in de totale fout in tijdstip $t_n$. De lokale fout is dan wel vermenigvuldigd met de matrix $G_h(t_n, t_\nu)$.

De differentiaalvergelijking plant iedere lokale fout, als die klein genoeg is, min of meer voort volgens een oplossing van de homogene vergelijking $u' = Ju$ (zie 1.2.11). We kunnen niet verwachten dat het foutvoortplantingsgedrag van de multistep gunstiger is. Voor zeer kleine $h$ (en niet te grote $T$ (zie 4.3.7) of sterk stabiele multistep (zie 4.5.6)) is het niet veel slechter. Voor grotere $h$ mogen we dit echter niet meer verwachten. Maar, in feite interesseert ons dat ook niet echt. Als $\sup_t \|\tilde{e}_h(t)\|$ een goede manier is om de totale fout te meten (zie 1.2.5) dan zijn we al tevreden als iedere lokale fout niet of nauwelijks groeit—als er een $K > 0$ is, niet al te groot (denk aan $K < 10$) zodat $\|G_h(t_n, t_\nu)\| \leq K$ voor iedere $t_n > t_\nu$ of $\|G_h(t_n, t_\nu)\| \leq Ke^{t_n - t_\nu}$.[18] De totale absolute fout wordt dan gemajoreerd door de som van de norm van de lokale fouten maal $K$ of $Ke^T$. Uiteraard is het gunstiger als iedere lokale fout uitdempt—als er bijvoorbeeld een $K > 0$ is, niet al te groot, en een $\lambda \in [0, 1)$ zodat $\|G_h(t_n, t_\nu)\| \leq K\lambda^{n-\nu}$ voor iedere $t_n > t_\nu$. De totale absolute fout wordt dan gemajoreerd door het maximum van de norm van de lokale fouten maal $K\frac{1}{1-\lambda}$. In deze situaties blijft de totale absolute fout klein als de lokale absolute fouten klein zijn.

Merk op dat de lokale absolute fout $\|h\delta_h(t_n)\|$ klein kan zijn ook al voor grotere $h$ (als bijvoorbeeld $u^{*(l+1)}$ klein is of, als deze funktie niet al te groot is, als $l$ wat groter is).

We merkten al op dat het foutvoortplantingsgedrag van de multistep niet gunstiger zal zijn dan dat van de differentiaalvergelijking. We mogen dus alleen hopen dat we multistep methoden kunnen vinden die lokale fout begrensd of gedempt voortplant als we de multistep toepassen op een differetiaalvergelijking die fouten ook zo voortplant.

**4.6.2 Opgave.** Zij $d \in \mathbf{N}$. Beschouw $\eta_1, \ldots, \eta_d \in \mathbf{C}$.
Stel $\eta := \mathrm{Re}(\eta_1) \geq \mathrm{Re}(\eta_j)$ voor $j = 2, \ldots, d$.
Beschouw, voor $\tau \geq t_0$, $\mathbf{R}^d$-waardige funkties $\varphi$ op $[\tau, \infty)$ van de vorm

$$\varphi(t) = \varepsilon_1 e^{\eta_1(t-\tau)} + \ldots + \varepsilon_d e^{\eta_d(t-\tau)} \quad \text{voor} \quad t \geq \tau,$$

waarbij $\varepsilon_j \in \mathbf{R}^d$.
a. Ga na dat, voor zekere $C > 0$ geldt

$$\|\varphi(t)\| \leq Ce^{\eta(t-\tau)} \quad \text{voor} \quad t \geq \tau.$$

b. Stel dat voor een $r > 0$ geldt $\eta \geq r + \mathrm{Re}(\eta_j)$ voor $j = 2, \ldots, d$ en $\varepsilon_1 \neq 0$.
Laat zien dat er een $c \in (0, \infty)$ is zodat

$$\|\varphi(t)\| = ce^{\eta(t-\tau)} \left(1 + \mathcal{O}(e^{r(t-\tau)})\right) \quad \text{voor} \quad t \to \infty.$$

c. Stel dat $\eta_1 = \bar{\eta}_2$, $\eta > \mathrm{Re}(\eta_j)$ voor $j = 3, \ldots, d$ en $\varepsilon_1 = \bar{\varepsilon}_2 \neq 0$.
Bewijs dat, voor $\gamma > 0$ er een $c > 0$ is zodat

$$\frac{1}{2\gamma} \int_{t-\gamma}^{t+\gamma} \|\varphi(s)\| \, ds \geq ce^{\eta(t-\tau)} \quad \text{voor } t \geq \tau.$$

(Kiezen we de $\varepsilon_j$ random dan hebben we 100% kans dat $\varepsilon_1 \neq 0$.) We mogen stellen dat bijna iedere $\|\varphi\|$, met $\varphi$ van bovenstaande vorm, op den duur (gemiddeld) min of meer groeit volgens $t \to e^{\eta(t-\tau)}$.

**4.6.3 Opgave.** Zij $p \in \mathbf{N}$. Beschouw $\lambda_1, \ldots, \lambda_p \in \mathbf{C}$.
Stel $\lambda := |\lambda_1| \geq |\lambda_j|$ voor $j = 2, \ldots, p$.
Beschouw de rij $(\varphi_n)$ in $\mathbf{R}^d$ van de vorm

$$\varphi_n = \varepsilon_1 \lambda^n + \ldots + \varepsilon_p \lambda_p^n \quad \text{voor} \quad n \in \mathbf{N}_0,$$

waarbij $\varepsilon_j \in \mathbf{R}^d$.
Ga na dat, voor zekere $C > 0$ geldt

$$\|\varphi_n\| \leq C\lambda^n \quad \text{voor} \quad n \in \mathbf{N}_0.$$

Ga na dat deze schatting ook weer in zekere zin scherp is (als in 4.6.2): we mogen stellen dat bijna iedere $(\|\varphi_n\|)$, met $(\varphi_n)$ van bovenstaande vorm, op den duur min of meer (gemiddeld) groeit volgens $(\lambda^n)$.

Om de volgende redenen onderzoeken we eerst onder welke voorwaarden een multistep toegepast op een lineaire vergelijkingen met exponentieel dempend foutvoortplantingsgedrag ook zo'n

[18] Willen we de fout $\tilde{e}_h$ klein houden ten opzichte van een gewichtsfunktie $\omega$ dan zouden we wellicht graag zien dat $\|\omega(t_n)G_h(t_n, t_\nu)\| \leq K$.

foutvoortplantingsgedrag heeft.

• Zeker in geval de (lokale) fouten klein zijn en fouten door de differentiaalvergelijking gunstig (bv. gedempt) worden voortgeplant, wordt iedere fout min of meer voortgeplant volgens een oplossing van een lineaire vergelijking.
• Lineaire vergelijkingen laten zich gemakkelijk analyseren.
• Als een multistep methode slechte eigenschappen heeft met betrekking tot lineaire vergelijkingen dan zal hij zeker diezelfde slechte eigenschappen hebben met betrekking tot algemenere niet-lineaire problemen.

We bekijken eerst de vergelijking met konstante koëfficiënten voor het geval $d = 1$.

**Lineaire vergelijkingen:** $d = 1$

**4.6.4 Het probleem.** Beschouw, met $d = 1$, het lineaire probleem in 1.1.11.
**Foutvoortplanting in de differentiaalvergelijking.** Iedere lokale fout wordt door de differentiaalvergelijking over een afstand $h$ voortgeplant met een faktor $e^{h\eta}$.
Wegen we de fouten met gewichtsfunktie $\omega \equiv 1$ (zie 1.2.5) dan is het beginwaarde probleem goed gekonditioneerd alleen als $\eta T < \kappa$ voor zekere $\kappa$ niet al te groot (denk aan $\kappa = 10$; zie 1.2.12).
**Foutvoortplanting in de multistep.** In de multistep oplossing wordt, met $\tilde{\eta} := h\eta$, zoals we in 4.3.14 gezien hebben, iedere lokale fout in $t_\nu$ naar $t_n$ voortgeplant met een faktor

$$G_h(t_n, t_\nu) = G_{n-\nu} = \sum_{j=1}^{k} \gamma_j(\tilde{\eta}) \lambda_j(\tilde{\eta})^{n-\nu}$$

waarin $\lambda_j(\tilde{\eta})$ de wortels zijn van $\rho - \tilde{\eta}\sigma$ (we nemen weer aan dat deze allen enkelvoudig zijn). Voor $n \gg \nu$ neemt $t_n \to |G_h(t_n, t_\nu)|$ over een afstand $h$—van $t_n$ naar $t_{n+1}$—gemiddeld ongeveer toe met $\lambda := \max_{j \le k} |\lambda_j(\tilde{\eta})|$ (zie 4.6.3). In geval $\eta < 0$ zouden we graag zien dat

$$\lambda = \max_{j \le k} |\lambda_j(\tilde{\eta})| < 1. \tag{43}$$

Omdat

$$\tilde{K} := \sup \left\{ \sum_{j=1}^{k} |\gamma_j(\tilde{\eta})| \left( \frac{|\lambda_j(\tilde{\eta})|}{\lambda} \right)^n \,\middle|\, n \in \mathbf{N} \right\} < \infty$$

is dan

$$|G_{n-\nu}| \le \tilde{K} \lambda^{n-\nu} \le \tilde{K} \quad \text{voor alle} \quad n \in \mathbf{N}.$$

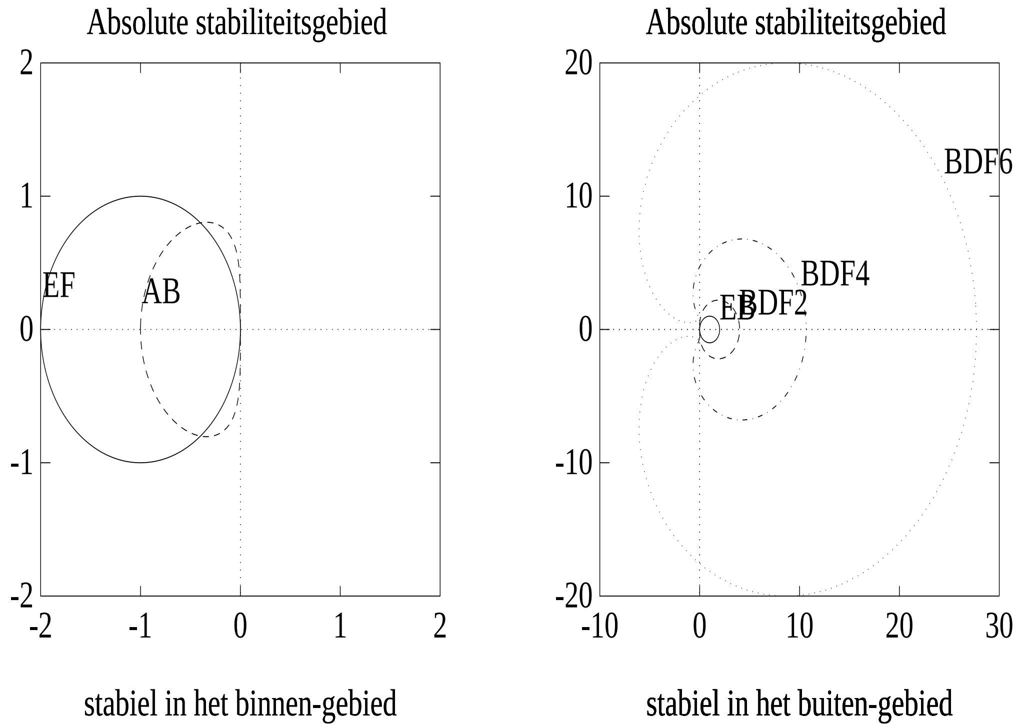
$\tilde{K}$ zou wellicht erg groot kunnen zijn, maar het is in ieder geval duidelijk dat de fouten in de multistep te hard groeien als $\lambda > 1$ [19]. Aan de hand van wat voorbeeldjes bekijken we wat (43) betekent. We geven eerst het gebied van de $\tilde{\eta}$ die voor ons van belang lijken te zijn een naam. Met name met het oog op toepassingen in de $d > 1$ situatie bekijken we ook complexe irreële $\tilde{\eta}$.

**4.6.5 Definitie.** Voor $\lambda \in [0, \infty)$ is

$$\mathcal{S}(\rho, \sigma)_\lambda := \left\{ \tilde{\eta} \in \mathbf{C} \,\middle|\, |\lambda_j| < \lambda \quad \text{voor alle} \quad \lambda_j \in \mathbf{C} \quad \text{waarvoor} \quad \rho(\lambda_j) - \tilde{\eta}\sigma(\lambda_j) = 0 \right\}$$

het *stabiliteitsgebied* van de multistep ten opzichte van de *groeifaktor* $\lambda$.
$\mathcal{S}(\rho, \sigma) := \mathcal{S}(\rho, \sigma)_1$ is het *absolute stabiliteits gebied* van de multistep.

[19] We zijn hier geïnteresseerd in de situatie waarin $h$ niet echt klein is. De mededeling bijvoorbeeld dat welliswaar $\lambda > 1$ maar dat wel $\lambda = 1 + \mathcal{O}(h)$ spreekt ons hier nu niet zo aan. Zie ook 4.6.6.




Figuur 1: Absolute stabiliteits gebieden voor verschillende methoden.
Links: expliciete methoden, Euler forward (EF), Adams–Bashford (AB).
Rechts: impliciete methoden, Euler backward (EB), BDF van orde $N$ (BDF$N$).

We kunnen de stabiliteits gebieden ook als volgt representeren. Er geldt

$$\mathbf{C}\backslash\mathcal{S}(\rho,\sigma)_\lambda = \left\{ \frac{\rho}{\sigma}(\zeta) \,\Big|\, \zeta \in \mathbf{C},\, |\zeta| \geq \lambda \right\}.$$

De stabiliteits gebieden zijn open delen van $\mathbf{C}$.

#### 4.6.6 Voorbeeld.

Met de Euler forward methode is $k = 1$ en $\lambda_1(\tilde{\eta}) = 1 + \tilde{\eta}$.
Het absolute stabiliteitsgebied van de Euler forward methode is een open cirkelschijf in het complexe vlak met straal 1 en middelpunt $-1$.
De eis dat $|1 + h\eta| < 1$ betekent voor $\eta \ll 0$ ook $0 < h \ll 1$.

$u^*(t) = \frac{1}{1+\varepsilon^2}(\cos t + \varepsilon \sin t)$ is de oplossing van het probleem

$$\varepsilon u'(t) = -u(t) + \cos(t) \quad (t \geq 0) \quad \text{en} \quad u(0) = \frac{1}{1+\varepsilon^2}. \tag{44}$$

Stel nu dat $\varepsilon = 10^{-6}$. Dus $\eta := -\frac{1}{\varepsilon} = -10^6$. Dan $|\delta_h(u^*)(t_n)| \leq 0.001$ als $h = 0.002$.
• De eis $|1 + h\eta| < 1$ verlangt echter dat $h < 2\varepsilon = 2\,10^{-6} \ll 2\,10^{-3}$.
• Als $h = 2.001\,10^{-6}$ dan is $|1 + h\eta| = 1.001$ en in $t_n = 0.02$ is $n \approx 10^4$ en $|1 + h\eta|^n \approx e^{10} \approx 2.2\,10^4$. In $t_n = 0.1$ is $|1 + h\eta|^n \approx 5.2\,10^{21}$.
• Staan we een lichte groei toe in de foutvoortplanting, bv. $|1+h\eta| < 1+h$ zodat $|1+h\eta|^n < e^{nh}$, dan moet $h < 2.000001\,10^{-6}$ zijn.
• Willen we de foutgroei klein hebben ten opzichte van bijvoorbeeld de gewichtsfunktie $\omega(t) = e^{-t}$ dan moet $|e^{-h}(1 + h\eta)| < 1$, hetgeen vereist dat $h < 2.00001\,10^{-6}$.
( $u^*(t) = t^2$ is de oplossing van het probleem $u'(t) = \eta(u(t) - t^2) + 2t$, $u(0) = 0$. Dus $|\delta_h(t_n)| \leq$

0.001 als $h = 0.001$. Omdat $u^*$ groeit zijn we nu wellicht tevreden als $|t_n^{-2}(1+h\eta)^n| < 5$. Echter om $|t_n^2(1+h\eta)^n|$ onder de veel grotere bovengrens $t_n^{-2}e^{t_n}$ te houden moet $|1+h\eta| \leq e^h$.)

We zien dat de eis $|1+h\eta| < 1$ een vele kleinere $h$ verlangt dan de $h = 0.002$ waarmee we zouden willen werken. Verder zien we dat de redelijke pogingen om de eis $|1+h\eta| < 1$ te verlichten geen echte invloed heeft op de grootte van $h$.

### 4.6.7 Voorbeeld.

Met de Euler backward methode is $k = 1$ en $\lambda_1(\tilde{\eta}) = \frac{1}{1-\tilde{\eta}}$.
Voor iedere $\eta < 0$ en voor iedere $h > 0$ is nu $0 < |\lambda_1(\tilde{\eta})| < 1$.
Het stabiliteitsgebied is het complement van de gesloten cirkelschijf met straal 1 rond +1.
De foutdemping in de multistep is beter naarmate $\eta$ kleiner is. Met $h = 0.002$ en $\eta = -10^6$ is $\lambda_1(h\eta) \approx 5\,10^{-4}$. De demping door de differentiaalvergelijking is aanzienlijk beter (met een faktor $e^{-2000}$ over een afstand $h = 0.002$), maar over de demping in de multistep mogen we toch ook dik tevreden zijn.

Beschouw, ter illustratie, de multistep oplossing $u_h$ van (44) met $e_0 = 0$ (en $\epsilon_n = 0$). Met $\eta := -\frac{1}{\varepsilon}$ en $e_h := u^* - u_h$ is $e_h(t_{n+1}) = \frac{1}{1-h\eta}(e_h(t_n) + h\delta_n)$ voor iedere $t_n \in \tilde{\mathcal{J}}_h$. Blijkbaar

$$e_h(t_n) = \sum_{j=1}^{n}(\frac{1}{1-h\eta})^j h\delta_{n-j} \quad \text{en} \quad |e_h(t_n)| \leq \frac{\varepsilon}{h}\frac{1}{2}h^2 \sup_{t\leq t_n}|u^{*(2)}(t)| \leq \frac{1}{2}\varepsilon h(1+\varepsilon).$$

Als $\varepsilon \leq h \leq 1$ dan is $|e_h(t_n)| \leq h^2$ voor iedere $t_n$.
Als $h \leq \varepsilon \leq 1$ dan is $|e_h(t_n)| \leq h$ voor iedere $t_n$.

**4.6.8 Konklusies** $h\eta$ lijkt beter informatie te geven dan $h$ over de stabiliteit van de multistep toegepast op het beginwaarde probleem, terwijl $\eta T$ de grootheid is die de konditionering van het beginwaarde probleem bepaalt (zie 1.2.12).

Zij $0 < \epsilon < E$ en $0 < \kappa < K$ waarin $\epsilon$ klein (denk aan $\epsilon < 0.2$), $E$ niet klein (denk aan $E > 0.9$), $\kappa$ niet groot (denk aan $\kappa < 10$) en $K$ groot (denk aan $K > 1000$).

In het volgende schema vatten we de stabiliteits voorwaarden samen. We gaan ervan uit dat de totale absolute fout uniform klein moet zijn (zie 1.2.5).

| | $\lvert\eta T\rvert < \kappa$ | $\lvert\eta T\rvert > K$ |
|---|---|---|
| $\lvert h\eta\rvert < \epsilon$ | stabiele multistep | sterk stabiele multistep<br>$\eta < 0$ |
| $\lvert h\eta\rvert > E$ | | $h\eta \in \mathcal{S}(\rho,\sigma)$ |

Als $|h\eta|$ klein is worden zeker de gladde lokale diskretisatie fouten ongeveer voortgeplant volgens een oplossing van het homogene deel van de lineaire vergelijking. Als het aantal iteratie stappen zeer groot is zouden totale fout toch nog groot kunnen zijn—door een kumulatief effekt van lokale fouten of doordat in zwak stabiele methoden minder gladde fouten langzaam groeien. Als het probleem goed sterk gekonditioneerd is kunnen we dit voorkomen door met een sterk stabiele methode te werken (Als $|\eta T|$ groot is dan is dit lineaire probleem goed konditioneerd precies dan als $\eta < 0$. Zie 1.2.12).

Als $|h\eta|$ niet klein is en $|\eta T|$ groot dan moet $h\eta \in \mathcal{S}(\rho,\sigma)$: lokale fouten sterven dan snel uit en de totale absolute fout $|\tilde{e}_h(t_n)|$ in $t_n$ wordt dan gemajoreerd door een niet al te groot veelvoud van de majorant $\max_{n-p\leq\nu\leq n}(|h\delta_\nu)| + |\epsilon_\nu|)$ van de laatste $p$ paar lokale fouten (of, wat veiliger, door de majorant $\max_{j<k}|\tilde{e}_h(t_j)| + \sup_{\nu\leq n}(|h\delta_\nu)| + |\epsilon_\nu|)$ van de startfouten en van de lokale fouten (diskretisatie fouten en evaluatie fouten). Als $|h\eta|$ niet klein is dan valt aan de eis "$h\eta \in \mathcal{S}(\rho,\sigma)$" niet te tornen, zelfs niet als we een moderate groei in de totale fout toestaan (als we de fouten wegen met bijvoorbeeld de gewichtsfunktie $\omega(t) = e^{-\theta(t-t_0)}$ met $|\theta| \ll |\eta|$).

Als $|h\eta| > E$ en $|\eta T| < \kappa$ dan zijn we met een beperk aantal iteratie stappen door het integratie interval. Als $h\eta$ niet in het stabiliteitsgebied zit groeit de fout. In de paar iteratie stappen zou de groei echter nog acceptabel kunnen zijn. De voorwaarde $h\eta \in \mathcal{S}(\rho,\sigma)$ is in dit geval niet zo strikt. Voor de praktijk is deze situatie echter gewoonlijk niet zo bar interessant.

We zouden de verzameling van de $h\eta$ in $\mathbf{C}$ waarvoor we een acceptabel foutvoortplantingsgedrag hebben het stabiliteitsgebied kunnen noemen. De grens van dit (heuristisch gedefiniëerd) stabiliteitsgebied is minder scherp naarmate $|h\eta|$ kleiner is. Als $|h\eta|$ niet klein is, is die grens zeer scherp en valt samen met de rand van $\mathcal{S}(\rho,\sigma)$: we kunnen ook nog met $h$ werken als $|h\eta| \in (\varepsilon, E)$ en $h\eta$ enigszinds buiten het absolute stabiliteitsgebied ligt.

**4.6.9 Opgave.**
a. Bewijs dat $\{\zeta \in \mathbf{C} \,\big|\, \mathrm{Re}(\zeta) < 0\}$ het absolute stabiliteitsgebied is van de trapezium regel.
b. Beschouw het schema $(\rho,\sigma) = (\chi^2 - \chi, \frac{1}{2}(3\chi^2 - 1))$ van de expliciete 2-staps Adams methode. Bewijs dat $(-\infty,-1] \cap \mathcal{S}(\rho,\sigma) = \emptyset$ en dat $(-1,0) \subset \mathcal{S}(\rho,\sigma)$.
c. Ga na dat het absolute stabiliteitsgebied van de midpunt regel leeg is. Ga na dat

$$\mathbf{C}\backslash\{\tfrac{\rho}{\sigma}(\zeta) \,\big|\, |\zeta| > 1\} = \{ix \,\big|\, x \in [-1,+1]\}.$$

d. Bepaal het absolute stabiliteitsgebied van de methode

$$u_h(t_{n+2}) = u_h(t_n) + 2hf(t_{n+2}, u_h(t_{n+2})) \quad \text{voor} \quad t_n \in \tilde{\mathcal{J}}_h.$$

**4.6.10** $\circ$ **Opgave.** Het absolute stabiliteitsgebied representeert men vaak grafisch. Men tekent daartoe in het complexe vlak eerst de kromme $\{\frac{\rho}{\sigma}(e^{i\phi}) \,\big|\, \phi \in [0,2\pi)\}$. Ga na wat het verband is tussen deze kromme en het absolute stabiliteits gebied.

Het absolute stabiliteitsgebied van sommige zwak stabiele methoden is leeg (zie c. in opgave 4.6.9). Voor sterk stabiele methoden is dat gelukkig niet het geval. Zonder bewijs vermelden we het volgende.

**4.6.11 Bewering.**
*Als de multistep met schema $(\rho,\sigma)$ sterk stabiel is dan is $\mathcal{S}(\rho,\sigma) \neq \emptyset$.* □

**Lineaire vergelijkingen: $d > 1$**

**4.6.12 Het probleem.**
Beschouw het lineaire $d > 1$ probleem in 1.1.12, weer met een diagonalizeerbare $J_0$ met eigenwaarden $\eta_1,\ldots,\eta_d$: $J_0 = V\mathrm{diag}(\eta_1,\ldots,\eta_d)V^{-1}$.
**Foutvoortplanting in de differentiaalvergelijking.** Zij $\eta := \max_{i\leq k}\mathrm{Re}(\eta_i)$.
Een lokale fout $\varepsilon \in \mathbf{R}^d$ in tijdstip $\tau$ wordt door de differentiaalvergelijking voortgeplant volgens $e(t) := V\exp(D(t-\tau))V^{-1}\varepsilon$ voor $t \geq \tau$. Deze fout wordt als volgt gemajoreerd

$$\|e(t)\| \leq \mathcal{C}(V) \max_{i=1,\ldots,d} |\, e^{\eta_i(t-\tau)}| \, \|\varepsilon\| \leq \mathcal{C}(V)\, e^{\eta(t-\tau)}\, \|\varepsilon\|,$$

met $\mathcal{C}(V) = \|V\|\,\|V^{-1}\|$. Behoudens wat zeer speciale gevallen (bv. $J_0 = D$ en $\varepsilon = (\varepsilon_1, 0, \ldots, 0)^T$), groeit $t \to \|e(t)\|$ gemiddeld op den duur volgens $t \to e^{\eta(t-\tau)}$ (zie 4.6.2).
Wegen we de fouten met gewichtsfunktie $\omega \equiv 1$ dan is het beginwaarde probleem alleen goed gekondioneerd als $\eta T < \kappa$ voor zekere $\kappa$ niet al te groot.
**Foutvoortplanting in de multistep.** In de multistep oplossing wordt, met $\tilde{\eta}_i := h\eta_i$ $(i = 1,\ldots,d)$, iedere lokale fout in $t_\nu$ naar $t_n$ voortgeplant met de faktor

$$G_h(t_n,t_\nu) = V\mathrm{diag}\,(G_n^{(1)}, G_n^{(2)}, \ldots, G_n^{(d)})V^{-1} \quad \text{met} \quad G_n^{(i)} = \sum_{j=1}^{k} \gamma_j(\tilde{\eta}_i)\lambda_j(\tilde{\eta}_i)^{n-\nu},$$

waarin $\lambda_j(\tilde{\eta}_i)$ de wortels zijn van $\rho - \tilde{\eta}_i\sigma$ (we nemen weer aan dat deze allen enkelvoudig zijn). Ga dit na. Met $\lambda := \max_{i,j}|\lambda_j(\tilde{\eta}_i)|$ is er een $\tilde{K} > 0$ zodat

$$\|G_h(t_n,t_\nu)\| \leq \|V\|\,\tilde{K}\lambda^{n-\nu}\|V^{-1}\| \leq \mathcal{C}(V)\tilde{K}\lambda^{n-\nu} \quad \text{voor alle} \quad n \in \mathbf{N}.$$

Behoudens wat zeer speciale gevallen neemt $t_n \to \|G_h(t_n, t_\nu)\|$ op den duur, over een afstand $h$, gemiddeld ongeveer toe met de faktor $\lambda$ (zie 4.6.3). We zouden dus weer graag zien dat

$$\lambda = \max_{i,j} |\lambda_j(\tilde{\eta}_i)| < 1.$$

**4.6.13** $\circ$ **Opgave.** Beschouw weer bovenstaande lineaire differentiaalvergelijking waarin nu $J_0$ niet per sé diagonalizeerbaar is. Stel dat $h\eta \in \mathcal{S}(\rho, \sigma)$ voor iedere $\eta$ in het spektrum van $J_0$. Bewijs dat er een $\lambda \in [0, 1)$ zodat $|\lambda_j(h\eta_i)| < \lambda$ (alle $i, j$) en een $K > 0$ (ook in geval niet alle wortels van $\rho - h\eta\sigma$ enkelvoudig zijn) zodat $\|G_h(t_n, t_\nu)\| \leq K\lambda^{n-\nu}$ voor alle $n, \nu \in \mathbf{N}$, $n > \nu$.

**4.6.14** **Opgave.** Beschouw de eerste orde formulering van het probleem in 1.1.14. Ga na dat voor dit twee dimensionale probleem geldt dat $\eta_1, \eta_2 \in \mathbf{R}$ en $\eta_2 \ll \eta_1 \approx \eta$.

**4.6.15** **Konklusies.** De konklusies in 4.6.8 laten zich gemakkelijk generaliseren. Voor iedere eigenwaarde $\eta_i$ van $J_0$ moet $|h\eta_i|$ klein zijn of $h\eta_i$ moet in $\mathcal{S}(\rho, \sigma)$ liggen. Wegen we de fouten met $\omega \equiv 1$, dan moet het probleem goed gekonditioneerd zijn: voor zekere $\kappa$ niet te groot moet $\mathrm{Re}(\eta_i)T \leq \kappa$ voor iedere eigenwaarde $\eta_i$. Is voor een eigenwaarde $\eta_i$ van $J_0$, $|h\eta_i|$ klein en is $\mathrm{Re}(\eta_i)T \leq -\kappa$ dan moet de multistep ook nog sterk stabiel zijn.

Als $u^*$ niet al te snel varieert maar $J_0$ heeft eigenwaarden $\eta_j$ met $\mathrm{Re}(\eta_j) \ll -1$ dan wordt, zeker in geval de multistep een begrensd absoluut stabiliteitsgebied heeft, de grootte van $h$ gedikteerd door deze $\eta_i$. Voor de exacte oplossing $u^*$ is deze eigenwaarde van geen enkel belang: de differentiaalvergelijking dempt de komponent van iedere lokale perturbatie op $u^*$ in de richting van de eigenvektoren bij deze eigenwaarden $\eta_j$ zeer snel uit.
(Schrijf, in 4.6.12, $e(\tau) = \sum_i \varepsilon_i v_i$ met $\varepsilon_i \in \mathbf{C}$ en $v_i \in \mathbf{C}^d$ de eigenvektor van $J_0$ bij eigenwaarde $\eta_i$—de $i$-de kolom vektor van $V$. Dan $e(t) = \sum_i \varepsilon_i e^{\eta_i(t-\tau)} v_i$ voor $t \geq \tau$.)
Het zou bijzonder prettig zijn en de efficiëntie van de methode ten goede komen als we ons niet hoefden te bekommeren om deze eigenwaarden die in feite geen essentiële rol behoren te spelen. Voor geen enkele methode is het hele complexe vlak het absolute stabiliteitsgebied (waarom is dat zo?). Onze beschouwingen hier zijn alleen van belang als $|h\eta_i|$ niet klein is en $|\eta_i T|$ groot voor zekere eigenwaarde $\eta_i$ van $J_0$. Voor zo'n situatie willen we echter $u^*$ alleen oplossen als voor deze eigenwaarde $\mathrm{Re}(\eta_i)T$ niet groot is. We hebben dus niet direkt behoefde aan een multistep methode waarvan het absolute stabiliteitsgebied het hele complexe vlak is. We zijn al tevreden als $h\eta_i \in \mathcal{S}(\rho, \sigma)$ voor iedere $\eta_i \in \mathbf{C}$ met $\mathrm{Re}(\eta_i) < 0$.

**4.6.16** **Definitie.** We zeggen dat de multistep met schema $(\rho, \sigma)$ A*–stabiel* is als $\{\zeta \in \mathbf{C} \,\big|\, \mathrm{Re}(\zeta) < 0\} \subset \mathcal{S}(\rho, \sigma)$.

Als het probleem goed gekonditioneerd is, als $J_0$ geen eigenwaarde $\eta_i$ heeft waarvoor $|\eta_i T| > \kappa$ en $0 \leq \mathrm{Re}(\eta_i)T < \kappa$ en als de multistep A-stabiel is[20], dan kiezen we $h$ zodat $|h\eta_i|$ klein is voor die eigenwaarde $\eta_i$ van $J_0$ waarvoor $|\eta_i T| \leq \kappa$. Verder kiezen we $h$ zodat de lokale diskretisatiefouten klein zijn.

**4.6.17** **Voorbeelden.** De trapezium regel en de Euler backward methode zijn A-stabiel.

**4.6.18** **Opgave.** Bewijs dat voor een scheme $(\rho, \sigma)$ geldt

$$\text{A-stabiel} \quad \Leftrightarrow \quad \{\zeta \in \mathbf{C} \,\big|\, |\zeta| \geq \mathbf{1}\} \subset \{\zeta \in \mathbf{C} \,\big|\, \mathrm{Re}\left(\tfrac{\rho}{\sigma}(\zeta)\right) \geq \mathbf{0}\}.$$

[20] De situatie waarin $J_0$ eigenwaarden $\eta_i$ heeft met klein reëel deel en groot imaginair deel verdient extra aandacht, maar die kunnen we hier niet aan dit probleem schenken.

### Het algemeen probleem

**4.6.19** Voor het probleem $u(t) = J(t)u(t) + g(t) \quad (t \in \mathcal{J}), \quad u(t_0) = u_0$ waarin $J$ een niet-konstante kontinue $\mathbf{M}_d(\mathbf{R})$-waardige funktie is de stabiliteitsvraag aanzienlijk moeilijker te beantwoorden dan in het konstante matrix geval. Toch zijn bovenstaande inzichten en konklusies ook hier van toepassing. Als $J(t)$ ten opzichte van $h$ weinig varieert (als bijvoorbeeld $J$ kontinu differentieerbaar is en $\|hJ'(t)\|$ klein is) dan lijkt het probleem lokaal op een konstant koëfficiënten probleem. We zullen daarom willen werken met een $h$ waarvoor voor iedere $t \in \mathcal{J}$ en iedere eigenwaarde $\eta_i(t)$ van $J(t)$ of $|h\eta_i(t)|$ klein is of $h\eta_i(t) \in \mathcal{S}(\rho, \sigma)$ ligt. Om de veranderingen in $J(t)$ in diens spektrum op te vangen zal het wenselijk zijn dat, voor een $\bar{\varepsilon} > 0$, voor de grotere $h\eta_i(t)$ ook $\zeta \in \mathcal{S}(\rho, \sigma)$ als $|h\eta_i - \zeta| \leq \bar{\varepsilon}$. Wegen we de fouten met $\omega \equiv 1$, dan moet het probleem goed gekonditioneerd zijn. Is voor een eigenwaarde $\eta_i(t)$, $|h\eta_i(t)|$ klein en is $\mathrm{Re}(\eta_i(t))T \leq -\kappa$ dan moet de multistep ook nog sterk stabiel zijn.

Passen we de multistep methode toe op (1), is de totale fout $\tilde{e}_h$ klein—het geval dat ons interesseert—en is $f \in C^2(\mathcal{J}, \mathbf{R}^d)$, dan voldoet de fout min of meer aan de multistep toegepast op het lineaire probleem met $J(t) = \frac{\partial}{\partial x} f(t, u^*(t))$ (zie 1.2.16). Dus ook voor deze situatie zijn de inzichten en konklusies uit de voorgaande subparagrafen van toepassing.

We zullen een multistep methode alleen toepassen op goed gekonditioneerde problemen. In de praktijk zullen we, zeker voor niet-lineaire problemen, de eigenwaarden van $J(t)$ niet bepalen. Toch zijn de inzichten die we opgedaan hebben waardevol. We hebben gezien dat, met name voor *stijve* problemen, problemen waarvan de differentiaalvergelijking zekere komponenten van lokale perturbaties snel uitdempt, de stabiliteitseis de maximaal toelaatbare $h$ scherp bepaalt en dat voor sommige multistep methoden die $h$ bijzonder klein is. De vraag met welke $h$ we nu in de praktijk aan de slag moeten beantwoordt zich vanzelf: de lokale diskretisatiefouten bepalen de nauwkeurigheid of de instabiliteit is zo groot dat dat al vlug te zien is. Verder hebben we gezien dat voor sommige multistep methoden de stapgrootte $h$ niet zo klein hoeft te zijn, zelf niet als we die toepassen op stijve problemen. Zeker met betrekking tot stijve problemen verdienen deze methoden de voorkeur.

Het is duidelijk dat er nog flink wat vragen te beantwoorden zijn.
Het is niet duidelijk wat er gebeurt als $J$ snel varieert (maar voor iedere $t$ wel een spektrum heeft in het linker complexe halfvlak) of als $f$ niet differentieerbaar is. In onze beschouwingen in 4.6.4 en 4.6.12 komen konstanten $\tilde{K}$ voor waarvan het niet duidelijk is of die niet te groot zijn. Het bijzonder fraaie resultaat van Dahlquist beantwoordt een aantal van die vragen.

Als in probleem (1) de rechterlid funktie $f$ *samentrekkend* (d.w.z 0-samentrekkend) is (zie 1.2.18) dan laat de differentiaalvergelijking lokale fouten niet groeien (zie 1.2.20). Het zou bijzonder fraai zijn als de multistep methode toegepast op dit probleem in zo'n geval de lokale fouten ook niet laat groeien.
We voeren eerst een notatie in en formuleren dan precies wat we van de multistep verlangen.

**4.6.20 Notatie.** Voor $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ definiëren we $\vec{u}_h \in C(\tilde{\mathcal{J}}_h, (\mathbf{R}^d)^k)$ door
$$\vec{u}_h(t_n) := (u_h(t_{n+k-1}), \ldots, u_h(t_n)) \qquad (t_n \in \tilde{\mathcal{J}}_h).$$

Voor een strikt positief definiete matrix $G = (g_{ij}) \in \mathbf{M}_k$ definiëren we m.b.v. het inprodukt $<\cdot,\cdot>$:

$$\|\vec{x}\,\|_G := \sqrt{\sum_{i,j=1}^{k} g_{ij} <x_i, x_j>} \quad \text{voor alle} \quad \vec{x} = (x_1, \ldots, x_k) \quad \text{met} \quad x_j \in \mathbf{R}^d.$$

Omdat $G$ strikt positief definiet is, is $\|\cdot\|_G$ een norm op $(\mathbf{R}^d)^k$.

**4.6.21 Opgave.** Bewijs dat $\|\cdot\|_G$ een norm is op $(\mathbf{R^d})^\mathbf{k}$.

**4.6.22 Definitie.** De multistep met schema $(\rho, \sigma)$ dat zo geschaald is dat $\sigma(1) = 1$ is *G-stabiel* als er een strikt positief definiete matrix $G \in \mathbf{M}_k$ is zodat voor iedere 0-samentrekkende funktie $f$ geldt

$$\|\vec{v}_h(t_n) - \vec{w}_h(t_n)\|_G \leq \|\vec{v}_h(t_\nu) - \vec{w}_h(t_\nu)\|_G \quad \text{voor alle} \quad t_n, t_\nu \in \mathcal{J},\, t_n > t_\nu,$$

en voor iedere $v_h, w_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ met grafiek in $\Omega$ waarvoor met $u_h = v_h$ en met $u_h = w_h$

$$\rho(T_h)(u_h)(t_n) = hf(\sigma(T_h)(\chi)(t_n), \sigma(T_h)(u_h)(t_n)) \quad \text{voor alle} \quad t_n \in \tilde{\mathcal{J}}_h.$$

**4.6.23 Opmerking.** De uitvoering van de multistep methode zoals in die in de definitie voorgesteld is noemt men de *een-been variant* van de multistep: we maken in een iteratie stap gebruik van een funktie waarde $f$ die uitgerekend is in het gemiddelde tijdstip
$\sigma(T_h)(\chi)(t_n) = \sum_{j=0}^{k} \beta_{j-k} t_{n+j}$ en de gemiddelde $u_h$-waarde $\sum_{j=0}^{k} \beta_{j-k} u_h(t_{n+j})$.

De een-been variant heeft betere stabiliteitseigenschappen dan zijn meer-benige verwant (zie 4.6.27 en 4.6.28). Bij de berekening hoeft geen $f$-waarde onthouden te worden. Verder is een een-been variant onmiddellijk toepasbaar op de differentiaalvergelijking van de vorm
$F(t, u(t), u'(t)) = 0 \quad (t \in \mathcal{J})$: toepassing resulteert in de differentievergelijking

$$F(\sigma(T_h)(\chi)(t_n), \sigma(T_h)(u_h)(t_n), \tfrac{1}{h}\rho(T_h)(u_h)(t_n)) = 0 \qquad (t_n \in \tilde{\mathcal{J}}_h).$$

**4.6.24 Opgave.** Vergelijk de efficiëntie van een een-been methode met die van zijn meerbeense variant.

**4.6.25 Opmerking.** Een G-stabiele multistep methode laat lokale fouten niet groeien—mits adekwaat gemeten—als die toegepast wordt in de een-been variant op een differentiaalvergelijkingen met een 0-samentrekkende rechterlid funktie $f$. Zoals we gezien hebben in 1.2.20, laat zo'n differentiaalvergelijking lokale perturbaties ook niet groeien. De funktie $f$ moet 0-samentrekkend zijn, maar hoeft verder, wat betreft de stabiliteit, aan geen enkele kontinuïteits– of differentieerbaarheids voorwaarde te voldoen. De totale fout $\tilde{e}_h$ in de multistep wordt, hoe groot $h$ ook is, in zo'n geval gemajoreerd door de som van de norm van de lokale fouten en de startfouten; dus, afgezien van de evaluatie fouten, hebben we

$$\max_{t_n \in \tilde{\mathcal{J}}_h} \|\vec{e}_h(t_n)\|_G \leq \|\vec{e}_h(t_0)\|_G + T \max_{t_n} \|\vec{\delta}_h(u^*)(t_n)\|_G. \tag{45}$$

De matrix $G$ hangt alleen van de methode af en niet van de differentiaalvergelijking. Willen we liever een schatting in bijvoorbeeld de $\|\cdot\|_2$–norm op $\mathbf{R}^d$ dan krijgen we die onmiddellijk omdat voor iedere $v_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ geldt

$$\begin{aligned} \max\nolimits_{t_n} \|v_h(t_n)\|_2^2 &\leq \max\nolimits_{t_n} \|\vec{v}_h(t_n)\|_2^2 \leq \|G^{-1}\|_2 \max\nolimits_{t_n} \|v_h(t_n)\|_G^2 \\ &\leq d\mathcal{C}_2(G) \max\nolimits_{t_n} \|v_h(t_n)\|_2^2 \end{aligned} \tag{46}$$

(ga dit na).

**4.6.26 Opgave.** Bewijs (45).
(Hint: Beschouw $\Phi : \mathbf{R^d} \to \mathbf{R^d}$ waarvoor $\|\Phi(x) - \Phi(y)\|_G \leq \|x - y\|_G$ alle $x, y$.
Laat, voor $(\delta_n)$, $(u_n)$ en $(u_n^*)$ zo zijn dat

$$u_{n+1} = \Phi(u_n), \quad u_{n+1}^* = \Phi(u_n^*) + \delta_n, \quad \text{alle } n, \quad u_0 = u_0^*.$$

Om $\|u_n^* - u_n\|_G$ af te schatten, beschouw $(u_n^{(k)})_{n \geq k}$ waarvoor

$$u_{n+1}^{(k)} = \Phi(u_n^{(k)}), \quad \text{alle } n > k, \quad u_k^{(k)} = u_k^*.$$

Dan $u_k^{(k)} - u_k^{(k-1)} = \delta_{k-1}$ en $u_n^* - u_n = u_n^{(n)} - u_n^{(n-1)} + u_n^{(n-1)} - u_n^{(n-2)} + \ldots + u_n^{(1)} - u_n^{(0)}$.)

Voor een zeer grote klasse van differentiaalvergelijkingen doen G-stabiele multistep methoden precies wat we graag zouden willen. De A-stabiele methoden zijn in feite afgericht op differentiaalvergelijkingen uit een zeer kleine klasse, namelijk op de zeer eenvoudige 1-dimensionale lineaire differentiaalvergelijkingen met een konstante koëfficiënt $\eta$ met $\mathrm{Re}(\eta) < 0$. Het volgende resultaat is daarom erg indrukwekkend; we geven geen bewijs.

**4.6.27 Stelling [Dahlquist].** *Beschouw een multistep met schema $(\rho, \sigma)$ zodat $\sigma(1) = 1$. De multistep is* A*-stabiel dan en slechts dan als de multistep* G*-stabiel is.* □

**4.6.28 Opgave.** Beschouw, voor $d = 1$ en $g \in C^1(\mathcal{J})$, het probleem

$$u'(t) = \lambda(t)(u(t) - g(t)) + g'(t) \quad \text{voor} \quad t \in \mathcal{J} \quad \text{en} \quad u(t_0) = g(t_0),$$

waarbij, voor $\eta \in (0, \infty)$, $\lambda(t) := -\frac{\eta}{t - t_0 + 1}$ $(t \in \mathcal{J})$.
a. Lossen we het probleem benaderend op met de (twee-been) trapezium regel dan voldoen de fouten $e_n := u^*(t_n) - u_h(t_n)$ aan de rekursie van de vorm $e_{n+1} = g_n e_n + h\tilde{\delta}_n$ en is $G_h(t_n, t_\nu) = g_{n-1} \cdot \ldots \cdot g_\nu$ voor $n > \nu$. Ga dit na en geef een uitdrukking voor $g_n$.
Bepaal de grootste $h_0 > 0$ (in termen van $\eta$) zodat voor iedere $h \in (0, h_0)$ geldt $|g_n| < 1$ $(n \in \mathbf{N}, nh < T)$.
b. Beantwoord de vragen uit a. ook voor de een-been trapezium regel. Vergelijk de twee "maximale" $h_0$.

Het enthousiasme over het fraaie resultaat wordt wellicht getemperd door de volgende barrière stelling, ook weer van Dahlquist, die we ook niet bewijzen.

**4.6.29 De tweede barrière stelling.** *Iedere* A*-stabiele multistep methode heeft hoogstens konsistentie orde 2. De foutkonstante $\frac{C_3}{\sigma(1)}$ van een* A*-stabiele methode van orde 2 is in absolute waarde groter dan of gelijk aan $\frac{1}{12}$, de foutkonstante van de trapezium regel.* □

We zullen in de volgende paragraaf situaties tegenkomen waarin we toch blij zijn met het feit dat er naast de trapezium regel ook nog andere A-stabiele methoden bestaan.

We staan even stil bij een klasse van multistep methoden waarvan het absolute stabiliteitsgebied niet het hele linker complexe halfvlak bevat, maar die toch voor een grote klasse van differentiaalvergelijkingen uitstekende stabiliteitseigenschappen heeft.

**4.6.30 Definitie.** Beschouw een multistep met schema $(\rho, \sigma)$.
Voor een $\alpha \in (0, \frac{\pi}{2}]$ noemen we de multistep A$(\alpha)$-*stabiel* als

$$\mathbf{C}(\alpha) := \{\zeta = re^{i\varphi} \in \mathbf{C} \,\big|\, r \in (0, \infty), |\pi - \varphi| < \alpha\} \subset \mathcal{S}(\rho, \sigma).$$

We zeggen dat de multistep is A$(0)$-*stabiel* als hij A$(\alpha)$-stabiel is voor zekere $\alpha > 0$.
De multistep noemen we A$_0$-*stabiel* als $\{\zeta \in \mathbf{C} \,\big|\, \mathrm{Im}(\zeta) = 0, \mathrm{Re}(\zeta) < 0\} \subset \mathcal{S}(\rho, \sigma)$.

A$(\frac{\pi}{2})$-stabiel is precies A-stabiel. A-stabiel $\Rightarrow$ A$(\alpha)$-stabiel $\Rightarrow$ A$_0$-stabiel.

Beschouw een lineaire vergelijking met konstante matrix $J_0$. Stel dat, voor zekere $\alpha \in [0, \frac{\pi}{2}]$, het spektrum van $J_0$ een deel is van $\mathbf{C}(\alpha)$. Dan is $h\eta \in \mathcal{S}(\rho, \sigma)$ voor *iedere* $h$ en iedere eigenwaarde $\eta$ van $J_0$.

We hopen uiteraard dat er A$(\alpha)$-stabiele methoden bestaan waarvan de konsistentie orde groter dan 2 is. Zonder bewijs vermelden we het volgende.

**4.6.31 Bewering.**
*Voor iedere $k = 1, \ldots, 6$ is de* BDF*-methode, voor zekere $\alpha > 0$,* A$(\alpha)$*-stabiel:*

| $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $\alpha$ | $\frac{\pi}{2}$ | $\frac{\pi}{2}$ | 1.544 | 1.278 | 0.905 | 0.328 |

*Voor iedere $\alpha \in [0, \frac{\pi}{2})$ en iedere $l \in \mathbf{N}$ bestaat er een* A$(\alpha)$*-stabiele $k$-staps methode met konsistentie orde $l$. Als $l \in \{3, 4\}$ dan is er zo'n methode met $k = l$.* □

**4.6.32 Bewering.** *Iedere* $A_0$*-stabiele multistep methode is impliciet.* □

**4.6.33** ∘ **Opgave.** Bewijs deze laatste bewering.

A($\alpha$)-stabiele methoden zijn impliciet en de stappen zijn dus moeilijker uitvoerbaar dan van menig andere expliciete multistep. Omdat we in zo'n andere methode vaak met veel kleinere stapgrootte zullen moeten werken kunnen A($\alpha$)-stabiele methoden toch veel efficiënter zijn.

**4.6.34 Opmerking.** A($\alpha$)-stabiele methoden zijn impliciet. We zouden de impliciet gedefiniëerde waarde $u_h(t_{n+k})$ willen oplossen met een succesief substitutie proces, met een prediktor–korrektor methode. Echter in de situatie waarin we met name geïnteresseerd zijn in dit soort stabiele methoden is $hL$ groot (in geval van een lineaire differentiaalvergelijking is de Lipschitz konstante $L$ het supremum van de $|\eta(t)|$ waarbij $t \in \mathcal{J}$ en $\eta(t)$ een eigenwaarde van $J(t)$). We mogen niet verwachten dat het succesief substitutie proces zal konvergeren. In zo'n situatie gaat men aan de slag met bijvoorbeeld het Newton-Raphson proces of een aanverwant proces (Koorden Newton, stationaire Newton-Raphson, etc.).

**4.6.35 Wie de schoen past ....** Voor iedere $l \in \mathbf{N}$ bestaan er expliciete en impliciete sterk stabiele multistep methoden (bv. de Adams methoden) van konsistentie orde $l$. Voor niet stijve goed gekonditioneerde problemen zijn deze methoden uitstekend toepasbaar. Het absolute stabiliteitsgebied van de expliciete methode is niet erg groot zodat voor stijve problemen de stapgrootte $h$ zeer klein moet zijn. Omdat in een expliciete methode iedere stap de evaluatie vereist van slechts een $f$-waarde vindt men het, voor het geval $d$ niet te groot is (denk aan $d < 5$), geen bezwaar om hier ook met kleine $h$ aan de slag te gaan: moderaat stijve laag dimensionale problemen worden ook numeriek opgelost met deze expliciete methoden.

Voor hoog dimensionale problemen is de tijd die een iteratie stap kost niet langer meer verwaarloosbaar. Omdat de impliciete methoden een groter stabiliteitsgebied zullen hebben worden die methoden interessanter. In kombinatie met een prediktor gebruikt men ze voor niet al te stijve goed gekonditioneerde problemen.

Voor stijve goed gekonditioneerde hoog dimensionale problemen gebruikt men A($\alpha$)-stabiele methoden: men lost liever de impliciet gedefinieerde $u$-waarden op dan de zeer vele $f$-waarden te evalueren die een expliciet schema zou eisen.

**4.6.36 Opgave.** Beschouw het lineaire probleem in 1.1.9. Zij $K \in (0, \infty)$.
Stel dat, voor iedere $t \in \mathcal{J}$, $J(t)^* = J(t)$ en $\eta(t) \in [-K, 0)$ voor iedere eigenwaarde $\eta(t)$ van $J(t)$.

In $\mathbf{R}^d$ kost een matrix–vektor vermenigvuldiging $d^2$ flop. Het oplossen van een matrix–vektor vergelijking kost $\approx \frac{1}{3}d^3$ flop.

Stel dat de lokale absolute diskretisatie fout van de trapezium regel voldoende klein is als $h \leq 0.1$ en van de expliciete 2-staps Adams–Bashforth als $h \leq 0.055$. Men wil zo efficiënt mogelijk $u^*(10)$ benaderen. Aan welke methode geef je, afhankelijk van $K$ en $d$, de voorkeur, aan de trapezium regel of aan de Adams–Bashforth methode? (Zie ook a. en b. in 4.6.9.)

**4.6.37** ∘ **Het stabiliteiten kabinet.** Stabiliteit ($\rho$ voldoet aan het wortel kriterium) noemt in de literatuur ook wel D-stabiliteit—ter ere van Dahlquist, de grote man in de multistep theorie—of asymptotische stabiliteit—omdat de uitspraken korrekt zijn voor $h$ klein genoeg. Naast deze stabiliteitsbegrippen, sterke en zwakke stabiliteit, A($\alpha$)-stabiliteit en G-stabiliteit komen er in de literatuur nog een groot aantal andere voor. We vermelden er een paar.

We beschouwen een multistep met schema $(\rho, \sigma)$ en met groeifaktor
$$\lambda(\tilde{\eta}) := \max\{|\lambda_j| \,\big|\, \rho(\lambda_j) - \tilde{\eta}\sigma(\lambda_j) = 0\} \quad \text{(zie 4.6.5).}$$
De multistep is L-*stabiel* als de methode A-stabiel is en $\lim_{\mathrm{Re}(\tilde{\eta}) \to -\infty} \lambda(\tilde{\eta}) = 0$.

De multistep is *stijf stabiel* als hij sterk stabiel is en er een $D > 0$ is zodat

$$\mathbf{C}_D := \{\zeta \in \mathbf{C} \,\big|\, \mathrm{Re}(\zeta) < -D\} \subset \mathcal{S}(\rho, \sigma).$$

Voor $\alpha \in (0, \frac{\pi}{2}]$, $D > 0$ is de methode A$(\alpha, D)$-*stabiel* als $\mathbf{C}_D \cup \mathbf{C}(\alpha) \subset \mathcal{S}(\rho, \sigma)$.

## 4.7 Exponentieel fitten

Onze inspanningen in de vorige paragraaf waren erop gericht voorwaarden te vinden waaronder een multistep lokale fouten fouten gedempt voortplant als het differentiaalvergelijking probleem dat toeliet. In zo'n situatie wordt de totale absolute fout $\|\tilde{e}_h(t_n)\|$ gemajoreerd door een niet al te groot veelvoud van de som van de norm van de lokale fouten en de startfouten (zie 4.6.23) of zelfs door zo'n veelvoud van het maximum van de norm van de laatste paar lokale fouten (zie 4.6.8): als alle lokale absolute fouten klein zijn is de totale absolute fout ook klein. In deze paragraaf vragen we ons af hoe we efficiënt te werk kunnen gaan als voor een zekere $h$ slechts een paar lokale diskretisatie fouten niet klein zijn. We beschouwen eerst een illustratief voorbeeld.

**4.7.1 Voorbeeld.** Beschouw, voor $\gamma \in \mathbf{R}$, $\varepsilon > 0$, $\varepsilon \ll 1$, de differentiaalvergelijking

$$\varepsilon u'(t) = -u(t) + \cos(t) \qquad (t \geq 0) \quad \text{uit (44) maar nu met} \quad u(0) = 1 + \gamma.$$

Met $w^*(t) := \frac{1}{1+\varepsilon^2}(\cos t + \varepsilon \sin t)$ is $w^*$ de exacte oplossing voor het geval $\gamma = 0$. Met $\eta := -\frac{1}{\varepsilon}$ en $v^*(t) := \gamma e^{\eta t}$ $(t \geq 0)$ is de exacte oplossing $u^* = v^* + w^*$. $u^*$ is opgebouwd uit een langzaam variërende funktie $w^*$ en een, voor kleine $t$, zeer snel variërende $v^*$.
Bovendien is, voor zowel de funktie als diens afgeleiden, de invloed van $v^*$ alleen van betekenis voor zeer kleine $t$:

$$u^{*(2)}(t) = \gamma \eta^2 e^{\eta t} + w^{*(2)}(t).$$

Voor $t \in [0, \varepsilon]$ is $u^{*(2)}(t) \approx v^{*(2)}(t) = \gamma \eta^2 e^{\eta t}$.
Voor $t \geq \sqrt{\varepsilon}$ is $|u^{*(2)}(t)| \approx |w^{*(2)}(t)| \leq 1 + \varepsilon$: met bv. $\varepsilon = 10^{-4}$ is $|u^{*(2)}(t)| \geq \frac{1}{3}|\gamma|\, 10^8$ als $t \in [0, 10^{-4}]$, terwijl $|\gamma \eta^2 e^{\eta t}| \leq \gamma 10^{-40}$ als $t \geq 10^{-2}$.
We kunnen stellen dat buiten de *"grenslaag"* $[0, \sqrt{\varepsilon}]$ —op een "inschakelverschijnsel" na— de oplossing $u^*$ gelijk is aan de langzaam variërende funktie $w^*$.

Stel we wensen $u^*$ te benaderen met een absolute fout $\leq 10^{-4}$.
Om stabiliteits redenen passen we Euler backward toe: we hebben, in 4.6.7, gezien dat we, in geval $\gamma = 0$, de gewenste nauwkeurigheid halen als $\varepsilon h \leq 10^{-4}$.

Laat $w_h$ de multistep oplossing zijn voor het geval $\gamma = 0$. Met $e_{2\,h} := w^* - w_h$ is, volgens 4.6.7, $|e_{2\,h}(t_n)| \leq \frac{1}{2}\varepsilon h$.
Als $u_h$ de Euler backward oplossing is voor het algemene geval en $v_h$ van het probleem $u' = \eta u$ op $\mathcal{J}$, $u(0) = \gamma$, dan is $u_h = v_h + w_h$. We zijn dus geïnteresseerd in $e_{1\,h} := v^* - v_h$ op $\mathcal{J}_h$. Omdat $v_h(t_n) = \frac{1}{1-h\eta} v_h(t_{n-1}) = (\frac{1}{1-h\eta})^n \gamma$ is $e_{1\,h}(t_n) = \gamma(e^{nh\eta} - (\frac{1}{1-h\eta})^n)$.
• Als, voor $p \in \mathbf{N}$, $h|\eta| = \frac{1}{p} \ll 1$ dan $t_p = \varepsilon$ en $\frac{1}{1-h\eta} = \exp(h\eta + (h\eta)^2 + \mathcal{O}((h\eta)^3))$. Dus

$$e_{1\,h}(t_n) = \gamma e^{t_n \eta} n (h\eta)^2 (1 + \mathcal{O}(h\eta)) \approx h t_n v^{*(2)}(t_n).$$

Blijkbaar $e_{1\,h}(\varepsilon) = \gamma e^{-1} h\eta(1 + \mathcal{O}(h\eta))$. Om in de roosterpunten in $[0, \varepsilon]$ de gewenste nauwkeurigheid te halen als $\gamma = 1$, moet $h \leq 10^{-4}\varepsilon e$ zijn. Dit is een waanzinnig veel zwaardere eis dan $h\varepsilon \leq 10^{-4}$. (De eis $h \leq 10^{-4}\varepsilon e$ is ook al aanzienlijk zwaarder dan onze aanname $|h\eta| = \frac{1}{p}$ voor bv. $p = 10$!) Voor $t_n \geq \sqrt{\varepsilon}$ is $v^{*(2)}(t_n)$ verwaarloosbaar.
• Als $1 \ll h|\eta|$ dan $|e_{1\,h}(t_n)| \approx \gamma(\frac{1}{1-h\eta})^n \leq \gamma(\frac{\varepsilon}{h})^n$. Als bv. $\varepsilon \leq 10^{-4}h$ dan is zelfs $|e_{1\,h}(t_1)| \leq 10^{-4}$ en hoeven we slechts $h \leq 1$ te kiezen om er ook nog voor te zorgen dat $h\varepsilon \leq 10^{-4}$. Als $\varepsilon \approx \frac{1}{10}h$ dan halen we de gewenste nauwkeurigheid in $t_n$ voor $n \geq 4$ en kunnen we met $h \leq .035$ aan de slag.
(ga na wat de nauwkeurigheid is als we $h$ zo kiezen dat $|h\eta| = 1$.)

De konklusies die uit bovenstaand voorbeeld getrokken kunnen worden zijn ook van toepassing op andere multistep methoden met een groot absoluut stabiliteits gebied en op andere stijve en vektorwaardige problemen. We formuleren die konklusies hieronder.

**4.7.2 Konklusie.** De oplossing $u^*$ van een stijve (vektor waardige) differentiaalvergelijking is opgebouwd uit een langzaam variërende funktie $w^*$ en een snel variërende funktie $v^*$, die verwaarloosbaar is buiten een zeer smalle grenslaag.

Lossen we benaderend op met een multistep methode dan halen we buiten deze grenslaag min of meer de gewenste nauwkeurigheid met iedere stapgrootte $h$ die goed genoeg is voor het langzaam variërende deel $w^*$. Werken we met een $\mathrm{A}(\alpha)$-stabiele methode dan hoeft die stapgrootte $h$ buiten de grenslaag niet klein te zijn, zelfs niet in geval het probleem zeer stijf is ($h\varepsilon \leq 10^{-4}$ in ons voorbeeld). In geval van een niet kleine stapgrootte zijn we al met één stap door de grenslaag heen en hebben we, in alle punten $t_n$ waarin we $u^*(t_n)$ benaderen, de gewenste nauwkeurigheid. Het heeft echter geen zin $u^*(t)$ voor $t$ in de grenslaag te benaderen door bv. lineair interpoleren in de berekende goede benaderingen (waarom niet?).

Is het snel variërende gedrag van $u^*$ in de grenslaag niet slechts een oninteressant "inschakelverschijnsel", maar wensen we $u^*$ ook in de grenslaag in goede nauwkeurigheid te benaderen, dan kunnen we met een kleine stapgrootte aan de slag. We zouden dan echter wel eens een zeer kleine stapgrootte opgedrongen kunnen krijgen. In zo'n geval heeft het geen voordeel om met een 'dure' impliciete methode aan de slag te gaan: in de grenslaag kunnen we, als we $u^*$ ook daar nauwkeurig willen benaderen, aan de slag met een 'goedkope' expliciete methode met diezelfde (zeer) kleine stapgrootte $h$. Als we de grenslaag gepasseerd zijn kunnen we overstappen op de $\mathrm{A}(\alpha)$-stabiele (impliciete) methode en verder rekenen met een veel grotere stapgrootte.

**4.7.3 Opgave.** Ga na dat deze konklusies inderdaad ook korrekt zijn voor het probleem in 1.1.14 en de Euler backward methode (zie ook 4.6.14).

In bovenstaand voorbeeld zitten we, in geval $\varepsilon = 10^{-4}$ is, met de volgende situatie. Met $h = 1$ vinden we, met de Euler backward methode, in alle tijdstippen $t_n$ een multistep benadering met een nauwkeurigheid $\leq 10^{-4}$. Werken we met kleinere $h$—om bijvoorbeeld ook benaderingen in tussenpunten, met name die in de grenslaag, te krijgen—dan moeten we in de grenslaag aan de slag met $h = 2.7\,10^{-8}$. Om door de grenslaag te komen vergt dit zo'n $10^5$ stappen[21], terwijl we met bv. $h = 10^{-5}$ in $10^2$ stappen de grenslaag al bevredigend in beeld zouden kunnen brengen (met $h = 10^{-5}$ geven de $u^*(nh)$ een goed beeld van $u^*$). In de volgende bewering geven we een methode aan om met zo'n grotere stapgrootte $h$ $u^*$ nauwkeurig te kunnen benaderen.

**4.7.4 Bewering.** *Zij $h\eta = \tilde{\eta} \in \mathbf{C}$. Voor een multistep met schema $(\rho, \sigma)$ en konsistentie orde $l$ zijn de volgende twee uitspraken equivalent.*

(a) $\rho(e^{\tilde{\eta}}) = \tilde{\eta}\sigma(e^{\tilde{\eta}})$.

(b) *Voor iedere* $\gamma \in \mathbf{C}$ *en* $p \in \mathcal{P}_l$ *is* $\delta_h(u^*) = 0$, *waarbij* $u^*(t) = \gamma e^{\eta t} + p(t) \quad (t \in \mathcal{J})$.

*Bewijs.* (a) volgt onmiddellijk uit (b) door $u^*(t) = e^{\eta t}$ te nemen en $\delta_h(u^*)(0)$ uit te schrijven.

De operator $v \to \delta_h(v)$ is lineair. Met het oog op stelling 4.1.5 is het daarom voldoende (b) uit (a) te bewijzen voor het geval $u^*(t) = v(t) := e^{\eta t}$. Welnu $\rho(T_h)(v)(t_n) = v(t_n)\rho(e^{h\eta})$ en $h\sigma(T_h)(v')(t_n) = hv'(t_n)\sigma(e^{h\eta}) = h\eta v(t_n)\sigma(e^{h\eta})$. □

Men kan deze bewering zien als een analogon van bewering 4.1.4. Stelling 4.1.5 heeft uiteraard ook hier een analogon; we laten de formulering en het bewijs ervan over aan de geïnteresseerde lezer.

**4.7.5 Opgave.** Beschouw voor $\alpha, \beta \in \mathbf{R}$ de methode met schema $(\chi - 1, \alpha\chi + \beta)$.

a. De methode is konsistent precies dan als $\alpha + \beta = 1$. De methode heeft konsistentie orde 2 precies dan als $\alpha = \beta = \frac{1}{2}$ (de trapezium regel). Bewijs dit.

[21] Het gebeurt nogal eens dat een standaard programma pakket voor het oplossen van beginwaarde problemen óf de grenslaag niet "opmerkt"—er met een stap overheen is—óf met zo'n waanzinnig kleine stapgrootte aan de slag gaat dat al voordat de grenslaag doorlopen is de beschikbare computer tijd op is. Omdat de methoden impliciet zijn zijn de stappen relatief duur!

Stel nu dat $\alpha + \beta = 1$.

b. Bewijs dat A-stabiel $\Leftrightarrow$ A(0)-stabiel $\Leftrightarrow$ $\alpha \in [\frac{1}{2}, 1]$.

c. Voor iedere $\tilde{\eta} \in (-\infty, 0)$ is er een $\alpha \in [\frac{1}{2}, 1]$ zodat $\rho(e^{\tilde{\eta}}) = \tilde{\eta}\sigma(e^{\tilde{\eta}})$. Bewijs dit.

**4.7.6 Exponentieel fitten.** Stel dat $u^*(t) = v^*(t) + w^*(t)$ waarbij $w^*$ en $\phi(t) := e^{-t\eta}v^*(t)$ langzaam variëren (d.w.z. $w^*$, $\phi$ en de eerste paar afgeleiden zijn uniform begrensd men een niet al te grote grens). In deze situatie is de lokale diskretisatie fout klein ook voor $\eta$ met $-\mathrm{Re}(\eta)$ groot en niet al te kleine $h$ als de konsistente multistep *exponentieel fit* (als $\rho(e^{h\eta}) = h\eta\sigma(e^{h\eta})$). Met zo'n multistep kunnen we in ook in de grenslaag met niet al te kleine stapgrootte de oplossing goed benaderen. Buiten de grenslaag kunnen we, als de multistep ook nog A($\alpha$)-stabiel is, met een grotere stapgrootte verder werken.

Als $u^*(t) = v_1^*(t) + \ldots + v_p^*(t) + w^*(t)$ zodat $w^*$ en $e^{-h\eta_j}v_j^*$ $(j = 1, \ldots, p)$ langzaam variëren dan zou men met stapgrootte $h$ aan de slag kunnen met een A($\alpha$)-stabiele konsitentie multistep methode die de $h\eta_j$ exponentieel fit $(\rho(e^{h\eta_j}) = h\eta_j\sigma(e^{h\eta_j})$ voor $j = 1, \ldots, p)$.
Omdat de $k$ in zo'n exponentieel fittende $k$-staps multistep methode groter zal zijn naarmate $p$ groter is, is het niet aantrekkelijk te fitten voor $p$ groter dan 3 of 4.

**4.7.7 Opmerking.** We hebben hier, om een oplossing van een stijf probleem ook in de grenslaag nauwkeurig te benaderen, twee alternatieven aangegeven: in de grenslaag werken met zeer kleine $h$ en een expliciete methode of werken met grotere $h$ en een exponentieel fittende A($\alpha$)-stabiele methode. In beide aanpakken kunnen we daarna, buiten de grenslaag, aan de slag met een grotere $h$ en een A($\alpha$)-stabiele methode.

Het exponentieel fitten heeft als voordeel dat we met de grotere stapgrootte kunnen werken. Een bezwaar is echter dat we de $\eta_j$, met $-\mathrm{Re}(\eta_j)$ groot, enigszins moeten kennen. Bovendien werkt exponentieel fitten alleen als $J$ in de grenslaag vrijwel konstant is en een of een paar (klusters van) eigenwaarden $\eta_j$ van $J(t_0)$ duidelijk gescheiden liggen van de overige (voor zekere $\zeta_j \in \mathbf{C}$ met $-\mathrm{Re}(\zeta_j)$ groot en $\kappa > 0$ niet al te groot moeten de eigenwaarden van $J(t_0)$ liggen in $\{\zeta \in \mathbf{C} \,\big|\, |\zeta| \leq \kappa\} \cup \bigcup_j \{\zeta \,\big|\, |\zeta - \zeta_j| \leq \kappa\}$ ).

# 5 Multistep voor tweede orde problemen

## 5.1 Konsistentie, stabiliteit en konvergentie

Tweede orde problemen zijn belangrijk en verdienen daarom speciale aandacht. Bovendien kunnen we hier aan de hand van tweede orde problemen een paar ideeën demonstreren die ook bruikbaar zijn voor de hogere orde problemen in 1.1.15.

Beschouw, voor $u_0, v_0 \in \mathbf{R}^d$ het tweede orde probleem

$$u''(t) = f(t, u(t), u'(t)) \quad \text{voor} \quad t \in \mathcal{J},\ u(t_0) = u_0,\ u'(t_0) = v_0. \tag{47}$$

Schrijven we deze vergelijking als een eerste orde systeem en passen we een multistep toe met schema $(\rho, \sigma)$ dan zijn we geïnteresseerd in de $u_h, v_h \in C(\mathcal{J}, \mathbf{R}^d)$ waarvoor

$$\begin{aligned} &\rho(T_h)(u_h) = h\sigma(T_h)(v_h) \\ &\rho(T_h)(v_h) = h\sigma(T_h)(f_h) \quad \text{op} \quad \tilde{\mathcal{J}}_h \\ &\quad \text{waarbij} \quad f_h(t_n) = f(t_n, u_h(t_n), v_h(t_n)) \quad \text{voor} \quad t_n \in \mathcal{J}_h. \end{aligned} \tag{48}$$

Als de rechterlidfunktie $f$ in (47) niet afhangt van $u'$, het geval waarin we in deze pargraaf speciaal geïnteresseerd zijn,

$$u''(t) = f(t, u(t)) \quad \text{voor} \quad t \in \mathcal{J}, u(t_0) = u_0,\ u'(t_0) = v_0 \tag{49}$$

dan ligt het voor de hand om te zoeken naar numerieke methoden waarin niet gebruik wordt gemaakt van $u'$ of van een benadering $v_h$ van $u'$. Elimineren we, voor dit geval, in (48), de roosterfunktie $v_h$ dan zien we dat

$$\rho^2(T_h)(u_h) = h^2\sigma^2(T_h)(f_h) \quad \text{met} \quad f_h(t_n) = f(t_n, u_h(t_n)).$$

De polynomen $\rho^2$ en $\sigma^2$ zijn hier nogal van een speciale vorm. Men mag hopen, door algemenere polynomen te bekijken, met lager graads polynomen een zelfde orde van nauwkeurigheid te kunnen krijgen.

**5.1.1 Multistep methoden voor tweede orde problemen.** We wensen (49) numeriek benaderend op te lossen met een multistep methode. Beschouw daartoe een tweetal polynomen $\rho$ en $\sigma$ van graad $k$. Voor $h > 0$ zijn we geïnteresseerd in de funktie $u_h \in C(\mathcal{J}, \mathbf{R}^d)$ waarvoor, met $f_h(t_n) = f(t_n, u_h(t_n))$, geldt

$$\rho(T_h)(u_h) = h^2\sigma(T_h)(f_h) \quad \text{op} \quad \tilde{\mathcal{J}}_h. \tag{50}$$

### Konsistentie

De konsistentie definities en resultaten in §1.6 voor eerste orde problemen laten zich gemakkelijk vertalen naar resultaten voor de tweede orde problemen in (49). We geven alleen de definitie en het analogon van stelling 4.1.7. We laten de verdere details aan de lezer over.

**5.1.2 Lokale diskretisatie fout en konsistentie.** De *lokale diskretisatie fout* is gegeven door $\delta_h(u^*) := h^{-2}\rho(T_h)(u^*) - \sigma(T_h)(u^{*(1)})$. De methode is *konsistent van orde $l$* als voor iedere probleem (49) waarin $f$ voldoende glad is geldt $\delta_h = \mathcal{O}(h^l)$ uniform $(h \to 0)$.

**5.1.3 Stelling.**
*De methode is konsistent dan en slechts dan als $\rho(1) = 0$, $\rho'(1) = 0$ en $\rho''(1) = \sigma(1)$.* □

**Stabiliteit**

De konsistentie eis $\rho(1)=\rho'(1)=0$ lijkt strijdig te zijn met onze stabiliteits wens: voor stabiliteit moesten immers de wortels op de eenheidscirkel enkelvoudig zijn! In de stabiliteits stelling 5.1.6, het analogon van 4.1.26, verliezen we inderdaad een extra faktor $h$: de som van de abslute lokale fouten en de startfout wordt vermenigvuldigd met $\frac{K}{h}$ in plaats van $K$. Omdat de lokale diskretisatie fout in de multistep rekursie gerepresenteerd wordt door de term $h^2\delta_h$ levert het verlies van de faktor $h$ wat deze term betreft hier geen probleem op:

$$\tfrac{1}{h}\textstyle\sum_{j=0}^{n} h^2\|\delta_h(u^*)(t_j)\| \leq T\max_{j\leq n}\|\delta_h(u^*)(t_j)\|$$

Verder moeten we om probleem (49) benaderend te kunnen oplossen niet alleen een goede benadering hebben voor $u^*(t_0)$ maar ook voor $u^{*(1)}(t_0)$. Een *konsistente start* eist dat

$$\max_{j<k}\|u_h(t_j)-u_0\|\to 0 \quad \text{en} \quad \max_{j<k}\|\tfrac{u_h(t_j)-u_h(t_0)}{jh}-v_0\|\to 0 \qquad (h\to 0).$$

Als $u_h(t_0)=u_0$ dan

$$\tfrac{1}{h}\|u^*(t_j)-u_h(t_j)\|\leq j\|\tfrac{u^*(t_j)-u^*(t_0)}{jh}-v_0\|+j\|\tfrac{u_h(t_j)-u_h(t_0)}{jh}-v_0\|,$$

zodat de extra faktor $h$ ook met betrekking tot de startfout geen probleem geeft. We kunnen daarom toch, met behulp van de navolgende stabiliteits definitie bewijzen dat ook hier weer "**stabiliteit + konsistentie = konvergentie**" (zie 5.1.8). Men kan het stabiliteits resultaat bewijzen door het bewijs van 4.1.26 hier en daar aan te passen: vervang $h$ door $h^2$ en $K$ door $\frac{K}{h}$. We geven hier geen verdere details.

**5.1.4 Bewering.** *Beschouw, voor $d=1$, het probleem*

$$u''(t)=0 \quad \textit{voor} \quad t\in[0,T] \quad \textit{en} \quad u(0)=u'(0)=0.$$

*De volgende twee beweringen zijn equivalent.*

(a) *Er is een $K>0$ zodat, voor iedere $h\in\mathbf{H}$, en iedere multistep oplossing $u_h$* $\quad(\rho(T_h)(u_h)=0)$ *geldt*

$$\sup_{t_n\in\mathcal{J}_h}|u_h(t_n)-u^*(t_n)|=\sup_{t_n}|u_h(t_n)|\leq K\max_{j<k}\max(|u_h(t_j)|,\frac{1}{h}|u_h(t_j)|).$$

(b) *De wortels van $\rho$ liggen op de eenheidsschijf en de wortels die op de eenheidscirkel liggen zijn of enkelvoudig of dubbel.* □

**5.1.5 Stabiliteit.** We zeggen dat de multistep in 5.1.1 *stabiel* is hij als een van de twee equivalente eigenschappen in 5.1.4 heeft. $K$ is dan de *stabiliteits konstante.*

**5.1.6 Stabiliteits stelling.** *Stel dat* (Dif.1) *geldt. Beschouw de multistep in* 5.1.1. *Schrijf weer $\tilde{\sigma}:=\sum_j|\beta_j|$. Stel dat de multistep stabiel is met stabiliteits konstante $K$. Laat, voor $h>0$, $u_h$ en $\tilde{u}_h$ multistep oplossingen zijn met lokale fout $(\epsilon_n)$, respektievelijk $(\tilde{\epsilon}_n)$. Dan geldt, met $K_h:=\frac{K}{1-h^2L|\beta_0|}$, voor iedere $t_n\in\mathcal{J}_h$ dat*

$$\|u_h(t_n)-\tilde{u}_h(t_n)\|\leq K_h\frac{1}{h}\left(\max_{j<k}\|u_h(t_j)-\tilde{u}_h(t_j)\|+\sum_{j=0}^{n-k}\|\epsilon_j-\tilde{\epsilon}j\|\right)e^{nh\tilde{\sigma}LK_h}. \quad \square$$

**5.1.7 Stabiliteit met betrekking tot evaluatie fouten.** We kunnen de extra faktor $h$ in de bijdrage van de evaluatie fout vermijden door het $\rho$ polynoom te splitsen.

Splits $\rho$ in een produkt $\rho=\rho_1\rho_2$ van twee polynomen $\rho_1$ en $\rho_2$ zodat beide polynomen slechts enkelvoudige wortels op de eenheidscirkel hebben. Dan geldt voor de oplossing $u_h$ en zekere $v_h$, beide in $C(\mathcal{J},\mathbf{R}^d)$,

$$\rho_1(T_h)(u_h)=hv_h \quad \text{en} \quad \rho_2(T_h)(v_h)=h\sigma(T_h)(f_h).$$

Neem voor het gemak aan dat de methode expliciet is.

Schrijf $u_n=u_h(t_n)$, $v_n=v_h(t_n)$. Laat $m_i$ de graad van $\rho_i$ zijn. Merk op dat $m_1+m_2=k$. Als $v_0,\ldots,v_{n+m_2-1}$ en $u_0,\ldots,u_{n+k-2}$ berekend zijn kan men $v_{n+m_2}$ en $u_{n+k-1}$ berekenen met behulp van deze relaties:

$\sum_{i=0}^{m_2} \alpha_{2\,m_2-i} v_{n+i} := \rho_2(T_h)(v_h)(t_n) = h \sum_{i=0}^{k-1} \beta_{k-i} f(t_{n+i}, u_{n+i}) := h\sigma(T_h)(f_h)(t_n)$

en $\sum_{i=0}^{m_1} \alpha_{1\,m_1-i} u_{n+i+m_2-1} := \rho_1(T_h)(u_h)(t_{n+m_2-1}) = hv_{n+m_2-1}$.

De benodigde startwaarden $v_0, \ldots, v_{m_2-1}$ volgen uit de gegeven startwaarden $u_0, \ldots, u_{k-1}$ en de relatie $\rho_2(T_h)(u_h) = hv_h$.

Zuiver mathematisch is deze methode equivalent met die in 5.1.1. Met betrekking tot rekenfouten echter niet. Deze laatste methode is stabieler: de foutvoortplanting in de rij van de paren $(v_{n+m_2}, u_{n+k-1})$ wordt in de eerste komponent bepaald door de karakteristieke wortels van het $\rho_2$ polynoom en in de tweede komponent door die van het $\rho_1$ polynoom. Beide polynomen hebben slechts enkelvoudige wortels op de eenheidscirkel. Deze methode is dus stabiel in de zin van 4.1.19.

### Konvergentie

De volgende steling volgt eenvoudig uit de stabiliteits stelling.

**5.1.8 Konvergentie stelling.** *Stel dat* (Dif.0) *en* (Dif.1) *gelden.*
*Beschouw de multistep in* 5.1.1. *Stel dat de multistep stabiel is en konsistentie orde $l$ heeft.*
*Laat, voor iedere $h \in \mathbf{H}$, $u_h \in C(\mathcal{J}_h, \mathbf{R}^d)$ een oplossing zijn van* (50) *zodat*

$$\max_{j<k} \|u_h(t_j) - u^*(t_j)\| = \mathcal{O}(h^{l+1}) \quad \textit{als} \quad h \to 0.$$

*Dan konvergeert $(u_h)$ met orde $l$ naar $u^*$.* $\square$

# 6 Numerieke oplosmethoden: Runge-Kutta methoden

## 6.1 Konsistentie, stabiliteit en konvergentie

We kunnen de integraal $\int_{\tilde{t}}^{\tilde{t}+h} g(t)\,dt$ numeriek benaderen met behulp van een kwadratuur formule als de integrant $g$ expliciet gegeven is. Een Runge-Kutta methode is een gegeneraliseerd soort kwadratuur formule waarmee we de integraal $\int_{\tilde{t}}^{\tilde{t}+h} f(t,u(t))\,dt$ numeriek kunnen benaderen, ook als de integrant, voor zekere $(\tilde{t},x)\in\Omega$, impliciet gedefinieerd is door de beginwaarde $u(\tilde{t}\,)=x$ en de differentiaalvergelijking

$$u'(t)=f(t,u(t)) \qquad (t\in\mathcal{J}). \tag{51}$$

De oplossing $u\in C^1([\tilde{t},\tilde{t}+h],\mathbf{R}^d)$ is de *oplossing van* differentiaalvergelijking (51) *door* $(t,x)$. Omdat

$$u(\tilde{t}+h)=x+\int_{\tilde{t}}^{\tilde{t}+h} f(t,u(t))\,dt. \tag{52}$$

levert de Runge-Kutta integratie methode onmiddellijk een een-staps numerieke oplosmethode op van probleem (1).

We beschrijven eerst in 6.1.1 precies de integratie procedure, die we vervolgens in 6.1.2, ten behoeve van een verdere analyse, wat abstrakter representeren. Een een-staps methode vergt geen aparte startprocedure en kan daarom uitstekend gebruikt worden in een variabele stapgrootte procedure. In 6.1.6 beschrijven we de Runge-Kutta methode met variabele stapgrootte.

**6.1.1 Runge-Kutta integratie: inleiding.** Laat $c_1,c_2,\ldots,c_k\in[0,1]$ steunpunten zijn en $b_1,\ldots,b_k$ gewichten van een kwadratuurformule.
Laat $l\in\mathbf{N}$ en $C\in\mathbf{R}$ zo zijn dat voor $g\in C^{l+2}(\mathcal{J})$, $\tilde{t}\in\mathcal{J}$ geldt

$$\int_{\tilde{t}}^{\tilde{t}+h} g(t)\,dt = h\sum_{i=1}^{k} b_i g(\tilde{t}+hc_i)+Ch^{l+1}g^{(l)}(\tilde{t}\,)+\mathcal{O}(h^{l+2}) \quad \text{uniform op } \mathcal{J};(h\to 0).$$

Zij $u$ de oplossing van (51) door $(\tilde{t},x)$.
Omdat (52) geldt, is, uniform op $\mathcal{J}$, voor $h\to 0$,

$$u(\tilde{t}+h)=x+h\sum_{i=1}^{k} b_i f(\tilde{t}+hc_i,u(\tilde{t}+hc_i))+Ch^{l+1}u^{(l+1)}(\tilde{t}\,)+\mathcal{O}(h^{l+2}). \tag{53}$$

Wegstrepen van de diskretisatiefout levert nog geen benaderings methode: we moeten nog de waarden $u(\tilde{t}+hc_j)$ zien te benaderen. We kunnen dit ook weer doen met (interpolatoire of extrapolatoire) kwadratuur formules weer met steunpunten $c_j$.
Voor $i=1,\ldots,k$ benaderen we, met gewichten $\alpha_{ij}$, de integraal $\int_{\tilde{t}}^{\tilde{t}+hc_i} g(t)\,dt$ door $h\sum_{j=1}^{k}\alpha_{ij}g(\tilde{t}+hc_j)$ en passen dit toe op (52) met $hc_j$ in plaats van $h$.
Om de benadering $\tilde{u}(\tilde{t}+h)$ van $u(\tilde{t}+h)$ te bepalen gaan we dus als volgt te werk.

Bepaal, met behulp van de "interne" kwadratuur formules, de $\tilde{u}_1,\ldots,\tilde{u}_k\in\mathbf{R}^d$ zodat

$$\tilde{u}_i=x+h\textstyle\sum_{j=1}^{k}\alpha_{ij}f(\tilde{t}+hc_j,\tilde{u}_j) \quad \text{voor} \quad i=1,\ldots,k.$$

Bereken dan, met de "externe" kwadratuur formule,

$$\tilde{u}(\tilde{t}+h)=x+h\textstyle\sum_{i=1}^{k} b_i f(\tilde{t}+hc_i,\tilde{u}_i).$$

**6.1.2 Runge-Kutta integratie.** Zij $k\in\mathbf{N}$.
Een *k-traps Runge-Kutta methode* (RK methode) wordt gegeven door een steunpuntenrij $c_1,\ldots,c_k$ (gewoonlijk in $[0,1]$), een gewichten vector $\vec{b}:=(b_1,\ldots,b_k)^T\in\mathbf{R}^k$ en een gewichten matrix $A:=(\alpha_{ij})\in\mathbf{M}_k(\mathbf{R})$.

Voor $h>0$, $\tilde{t}\in\mathcal{J}$ en $U\in\mathbf{M}_{d\times k}$ definiëren we $F(\tilde{t},U;h)\in\mathbf{M}_{d\times k}$ door

$$F(\tilde{t},U;h)e_j:=f(\tilde{t}+hc_j,Ue_j) \quad \text{voor} \quad j=1,\ldots,k.$$

Verder schrijven we $\vec{\mathbf{1}} := (1,1,\ldots,1)^T \in \mathbf{R}^k$.

De *lokale Runge-Kutta oplossing* $\tilde{u}$ van (51) *door* $(\tilde{t},x) \in \Omega$ is nu, voor iedere $h > 0$, gegeven door

$$\begin{aligned} &\tilde{u}(\tilde{t}+h) = x + hF(\tilde{t},U;h)\vec{b}, \\ &\quad \text{waarbij} \quad U \in \mathbf{M}_{d\times k}(\mathbf{R}) \quad \text{zo is dat} \quad U = x\vec{\mathbf{1}}^T + hF(\tilde{t},U;h)A^T. \end{aligned} \tag{54}$$

Als $u$ de exacte oplossing is van (51) door $(\tilde{t},x)$, dan benadert $\tilde{u}(\tilde{t}+h)$ de waarde $u(\tilde{t}+h)$.

De RK methode representeert men vaak door middel van de

*genererende* $(k+1)\times(k+1)$*-matrix* $\mathbf{A} := \left[\begin{array}{c|c} \vec{c} & A \\ \hline 0 & \vec{b}^T \end{array}\right]$, waarbij $\vec{c} := (c_1,\ldots,c_k)^T$.

Als $\alpha_{ij} = 0$ voor iedere $i \le j$ dan noemen we de methode *expliciet* (een ERK methode), anders noemen we de methode *impliciet* (een IRK methode). Als $\alpha_{ij} = 0$ voor iedere $i < j$ en $\alpha_{ii} \neq 0$ voor zekere $i$ dan is de methode *diagonaal-impliciet* (een DIRK methode).

### 6.1.3 Voorbeelden.

(a) **De midpunt regel.** Bereken $u_{n+1}$ uit $u_n$ als volgt

$$\begin{cases} x_1 = u_n \\ x_2 = u_n + \frac{1}{2}hf(t_n,x_1) \\ u_{n+1} = u_n + hf(t_n+\frac{1}{2}h, x_2), \end{cases} \quad \text{dus met genererende matrix} \quad \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline 0 & 0 & 1 \end{array}\right].$$

(b) **De expliciete trapezium regel.** Bereken $u_{n+1}$ uit $u_n$ als volgt

$$\begin{cases} x_1 = u_n \\ x_2 = u_n + hf(t_n,x_1) \\ u_{n+1} = u_n + \frac{1}{2}h[f(t_n,x_1) + f(t_n+h,x_2)], \end{cases} \quad \text{dus met genererende matrix} \quad \left[\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline 0 & \frac{1}{2} & \frac{1}{2} \end{array}\right].$$

(c) **Heun.** $\left[\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & 0 & \frac{2}{3} & 0 \\ \hline 0 & \frac{1}{4} & 0 & \frac{3}{4} \end{array}\right]$ (d) **"De" Runge-Kutta.** $\left[\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline 0 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}\right]$

De volgende bewering geeft voorwaarden waaronder een RK integratie stap uitvoerbaar is. De bewering is het analogon van 4.1.25.

### 6.1.4 Bewering.

*Stel*

(Dif.1)$_1$ $\quad \|f(t,x) - f(t,y)\|_1 \le L\|x-y\|_1 \quad ((t,x),(t,y)\in\Omega).$

*Voor* $h > 0$ *en* $(\tilde{t},x) \in \Omega$ *heeft de vergelijking* $U = x\vec{\mathbf{1}}^T + hF(\tilde{t},U;h)A^T$ *precies één oplossing* $U \in \mathbf{M}_{d\times k}(\mathbf{R})$ *als de methode expliciet is of als* $hL\|A\|_\infty < 1$.

*Bewijs.* Het is duidelijk dat het expliciete geval uniek oplosbaar is.

Definieer $\Phi(U) := x\vec{\mathbf{1}}^T + hF(\tilde{t},U;h)A^T$ voor $U \in \mathbf{M}_{d\times k}$.

Merk op dat $\|B\|_1 = \max_j \|Be_j\|_1$ $(B \in \mathbf{M}_{d\times k})$. We zien hiermee dat

$$\begin{aligned} \|\Phi(X)-\Phi(Y)\|_1 &\le h\|F(\tilde{t},X;h) - F(\tilde{t},Y;h)\|_1\|A^T\|_1 \\ &= h\max_j \|f(\tilde{t}+hc_j, Xe_j) - f(\tilde{t}+hc_j, Ye_j)\|_1\|A\|_\infty \\ &\le hL\max_j \|Xe_j - Ye_j\|_1\|A\|_\infty = hL\|A\|_\infty\|X-Y\|_1, \end{aligned}$$

zodat het kontraktie lemma toepasbaar is als $hL\|A\|_\infty < 1$. □

### 6.1.5 De hoeveelheid werk.

Beschouw 6.1.2.

In de Runge-Kutta integratie stap (54) moeten we een aantal funktie waarden evalueren. Het is van belang dit aantal zo klein mogelijk te houden.

In een ERK methode gaat men daarom als volgt te werk.
Bereken achtereenvolgens $y_1, \ldots, y_k \in \mathbf{R}^d$ en $\tilde{u}(\tilde{t}+h)$ als volgt

$$\begin{aligned}
y_1 &= hf(\tilde{t}+hc_1, u(\tilde{t}\,)) \\
y_2 &= hf(\tilde{t}+hc_2, u(\tilde{t}\,)+\alpha_{21}y_1) \\
&\vdots \\
y_k &= hf(\tilde{t}+hc_k, u(\tilde{t}\,)+\textstyle\sum_{j=1}^{k-1}\alpha_{kj}y_j) \\
\tilde{u}(\tilde{t}+h) &= u(\tilde{t}\,)+b_1y_1+\ldots b_ky_k.
\end{aligned}$$

Als $U = u(\tilde{t}\,)\vec{\mathbf{1}}^T + hF(\tilde{t}, U; h)A^T$ dan is $y_j = hF(\tilde{t}, U; h)e_j$. (Ga dit na.)

Bij een $k$-traps DIRK methode moeten $k$ (in het algemeen niet-lineaire) vergelijkingen elk met $d$ onbekenden worden opgelost en bij een algemenere $k$-traps IRK methode een (niet-lineaire) vergelijking met $kd$ onbekenden. Zeker de impliciete methoden zijn dus relatief duur. We zullen echter zien dat, om stabiliteitsredenen, zekere impliciete methoden toch interessant zijn (bij multistep methoden gaven we ook, om stabiliteistredenen, voor zekere type problemen de voorkeur aan impliciete methoden: zie 4.6.32). De interne "hellingen" $y_j = Ue_j$ kan men oplossen met behulp van een succesief substitutie proces (zie het bewijs van 6.1.4). De impliciete methoden zal men echter gewoonlijk toepassen op stijve problemen. In zo'n geval zal $hL\|A\|_\infty$ niet kleiner dan 1 zijn en het is dubieus of het succesieve substitutie proces konvergeert. Men past dan Newton-Raphson of een aanverwant proces toe.

**6.1.6 Runge-Kutta iteratie.** Beschouw de RK methode in 6.1.2.
Zij $\mathbf{h} = (h_n)$ een *stapgrootterij* in $(0,\infty)$. De rij $(t_n)$ in $\mathcal{J}$ is nu gegeven door

$$t_{n+1} := t_n + h_n = t_0 + \textstyle\sum_{j=0}^{n} h_j.$$

Verder is $\mathcal{J}_{\mathbf{h}} := \{t_n \,\big|\, t_n \in \mathcal{J}\}$ en definiëren we de *maaswijdte* $|\mathbf{h}|$ van de stapgrootterij $\mathbf{h}$ door $|\mathbf{h}| := \max_n h_n$.

$u_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ is een *(exacte) Runge-Kutta oplossing* van (51) als, voor iedere $t_n \in \mathcal{J}_{\mathbf{h}}$, met $u_n := u_{\mathbf{h}}(t_n)$, geldt

$$u_{n+1} = u_n + h_nF(t_n, U_n; h_n)\vec{b} \quad \text{met } U_n \text{ zodat} \quad U_n = u_n\vec{\mathbf{1}}^T + h_nF(t_n, U_n; h_n)A^T.$$

Voor een $\epsilon_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ is $u_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ een *Runge-Kutta oplossing met lokale fout* $\epsilon_{\mathbf{h}}$ als, voor iedere $t_n \in \mathcal{J}_{\mathbf{h}}$, met $u_n := u_{\mathbf{h}}(t_n)$, geldt

$$u_{n+1} = u_n + h_nF(t_n, U_n; h_n)\vec{b} + \epsilon_{\mathbf{h}}(t_n) \quad \text{met } U_n \text{ zodat} \quad U_n = u_n\vec{\mathbf{1}}^T + h_nF(t_n, U_n; h_n)A^T.$$

### Konsistentie

**6.1.7 De lokale diskretisatie fout.** Beschouw 6.1.6.
De *lokale diskretisatie fout* $\delta_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ in de RK methode toegepast op het probleem (1) is, voor iedere $t_n \in \mathcal{J}_{\mathbf{h}}$, met $t_n + h_n \in \mathcal{J}$ gegeven door

$$\begin{aligned}
\delta_{\mathbf{h}}(t_n) &:= \tfrac{1}{h_n}[u^*(t_n+h_n) - u^*(t_n)] - F(t_n, U_n; h_n)\vec{b}, \\
&\text{met } U_n \text{ zodat} \quad U_n = u^*(t_n)\vec{\mathbf{1}}^T + hF(t_n, U_n; h_n)A^T.
\end{aligned}$$

De RK methode heeft *konsistentie orde* $l$ als $\delta_{\mathbf{h}} = \mathcal{O}(|\mathbf{h}|^l)$ uniform $(|\mathbf{h}| \to 0)$ voor iedere probleem (1) waarin $f$ voldoende glad genoeg.

**6.1.8 Voorbeelden.**
De methoden in voorbeeld 6.1.3 zijn respektievelijk van orde 2, 2, 3, en 4.

**6.1.9 Opmerking.** Stel dat de RK methode konsistentie orde $l$ heeft. Dan heeft de externe kwadratuur formule een fout van $\mathcal{O}(h^{l+1})$ (als in (53); neem maar voor $f(t,x) = g(t)$). Omdat bij de berekening van $hF(\tilde{t}, U; h)\vec{b}$ fouten in de interne kwadratuur formules tegen elkaar kunnen wegvallen hoeven de fouten in de interne kwadratuur formules niet $\mathcal{O}(h^l)$ te zijn. Omdat "interne" fouten bij de "externe" berekening tegen elkaar kunnen wegvallen bestaan er toch, voor iedere $l$, expliciete $k$-traps RK methoden van konsistentie orde $l$. (Zie 6.2.3.)

**Stelling.** *Voor iedere $l \in 2\mathbf{N}$ is er een expliciete $k$-traps* RK *methode die konsistentie orde $l$ heeft en $k = \frac{1}{4}l^2 + 1$.* □

Uit efficiëntie overwegingen (zie 6.1.5) is het van belang $k$ zo klein mogelijk te houden. De minimale $k$ van zo'n methode wordt echter wel bepaald door $l$. De volgende barrière stelling is van Butcher. Het resultaat is scherp (zie 6.1.14).

**Barrière stelling.** *Als de $k$-traps* RK *methode expliciet is en konsistentie orde $l$ heeft dan $k > l$ als $l > 4$ $\qquad k > l+1$ als $l > 6$ $\qquad k > l+2$ als $l > 7$.* □

Een kwadratuur formule met $k$ steunpunten heeft maximaal een fout als in (53) van $\mathcal{O}(h^{2k+1})$ (Gauß-Legendre kwadratuur). Met een expliciete $k$-traps RK methode kunnen we die nauwkeurigheid volgens de barrière stelling niet bereiken, met een IRK methode daarentegen wel (ook weer volgens Butcher).

**Stelling.** *Voor iedere $k$ bestaat er een impliciete $k$-traps* RK *methode van orde $2k$.* □

**6.1.10 Opmerking.** Door van de voorkomende funkties Taylor reeksen rond $\tilde{t}$ te beschouwen kan men vaststellen of een RK methode konsistentie orde $l$ heeft.

Ook door middel van Taylor reeksen kan men genererende matrices konstrueren en lokale diskretisatie fouten beschrijven voor RK methoden van orde $l$. Met behulp van de kettingregel en de andere standaard regels voor differentiëren stelt men Taylor reeksen op rond $t_n$ in termen van partiële afgeleiden van $f$. Vervolgens bepaalt men de koefficiënten zodat de termen met $h_n^j$ in $\delta_{\mathbf{h}}(t_n)$ voor $j = 0, \ldots, l-1$ nul zijn. In feite is dit een rechttoe rechtaan karwei. De praktische komplikaties zijn echter immens. Ter illustratie konstrueren we een expliciete 2-traps RK methode van orde 2.

Beschouw, met $\tilde{t} = t_n$ en $h = h_n$,

$$\tilde{u}(\tilde{t}+h) = u(\tilde{t}) + h[b_1 f(\tilde{t}, u(\tilde{t})) + b_2 f(\tilde{t}+hc_2, \tilde{u}_2)]$$
$$\text{met} \quad \tilde{u}_2 = u(\tilde{t}) + h\alpha_{21} f(\tilde{t}, u(\tilde{t})).$$

We schrijven $f, f_t, f_x, \ldots$ in plaats van $f(\tilde{t}, u(\tilde{t}))$, $\frac{\partial f}{\partial t}(\tilde{t}, u(\tilde{t}))$, $\frac{\partial f}{\partial x}(\tilde{t}, u(\tilde{t})), \ldots$.
Omdat

$$f(\tilde{t}+hc_2, \tilde{u}_2) = f + h[c_2 f_t + \alpha_{21} f_x f] + \tfrac{1}{2}h^2[c_2^2 f_{tt} + 2c_2\alpha_{21} f f_{tx} + \alpha_{21}^2 f^2 f_{xx}] + \mathcal{O}(h^3)$$

is

$$\begin{aligned}\tilde{u}(\tilde{t}+h) = u(\tilde{t}) &+ h[b_1 f + b_2 f] + h^2[b_2 c_2 f_t + b_2\alpha_{21} f_x f] \\ &+ \tfrac{1}{2}h^3[b_2 c_2^2 f_{tt} + 2b_2 c_2 \alpha_{21} f f_{tx} + b_2\alpha_{21}^2 f^2 f_{xx}] + \mathcal{O}(h^3)\end{aligned}.$$

Verder is

$$u(\tilde{t}+h) = u(\tilde{t}) + hf + \tfrac{1}{2}h^2[f_t + f f_x] + \tfrac{1}{6}h^3[f_{tt} + 2f f_{tx} + f^2 f_{xx} + f_t f_x + f f_x^2] + \mathcal{O}(h^3).$$

Eisen we dat $\alpha_{21} = c_2$ en dat zo veel mogelijk termen in $\delta_{\mathbf{h}}(\tilde{t}) = \frac{1}{h}[u(\tilde{t}+h) - \tilde{u}(\tilde{t}+h)]$ verdwijnen dan zien we dat $b_1 + b_2 = 1$ en $2b_2 c_2 = 1$.
In dat geval is

$$\delta_{\mathbf{h}}(\tilde{t}) = \tfrac{1}{6}h^2[(\tfrac{3}{2}c_2 - 1)f_{tt} + (3c_2 - 2)f f_{tx} + (\tfrac{3}{2}c_2 - 1)f^2 f_{xx} - f_t f_x - f f_x^2] + \mathcal{O}(h^3).$$

Het is duidelijk dat, voor hogere orde methoden, de uitdrukkingen met de partiële afgeleiden van $f$ al vlug onoverzichtelijk worden. In de literatuur vindt men een representatie van deze partiële afgeleiden door middel van zekere diagrammen die het opstellen en weergeven van de Taylor reeksen uitdrukkingen enigszins overzichtelijk houdt. We gaan hier niet verder op in. Als de Taylor reeksen opgesteld zijn vinden we de (niet-lineaire) vergelijkingen, "beperkingen", waaraan de koëfficiënten van de genererende matrix moeten voldoen opdat de methode konsistentie orde $l$ heeft. Aan de hand van de volgende tabel zal het duidelijk zijn dat het niet gemakkelijk is met name hogere orde expliciete $k$-traps RK methoden te konstrueren. De tabel geeft het aantal beperkingen op de koëfficiënten van een RK methode van orde $l$.

| $l$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| no. vergl. | 1 | 2 | 4 | 8 | 17 | 37 | 85 | 200 | 486 | 1205 |

In de volgende stelling formuleren we de eerste vier beperkingen.

**6.1.11 Stelling.** *Beschouw, voor de* RK *methode in* 6.1.2, *de voorwaarden*

$$(1)\quad \sum_i b_i = 1 \qquad\qquad (2)\quad 2\sum_{i,j} b_i\alpha_{ij} = 1$$
$$(3)\quad 3\sum_{i,j,m} b_i\alpha_{ij}\alpha_{im} = 1 \quad \text{en} \quad 6\sum_{i,j,m} b_i\alpha_{ij}\alpha_{jm} = 1$$

*Voor $l \in \{1,2,3\}$ heeft de methode konsistentie orde $l$ dan en slechts dan als* (1) *t/m* ($l$) *vervuld is.* □

Het aantal beperkingen reduceert drastisch als, voor zekere $m_1, m_2 \in \mathbf{N}$ de volgende relaties gelden.

$$\mathrm{C}(m_1):\quad q\sum_{j=1}^k \alpha_{ij}c_j^{q-1} = c_i^q \qquad \text{voor } i = 1,\ldots,k,\ q = 1,\ldots,m_1$$

$$\mathrm{D}(m_2):\quad q\sum_{i=1}^k b_i c_i^{q-1}\alpha_{ij} = b_j(1-c_j^q) \quad \text{voor } j = 1,\ldots,k,\ q = 1,\ldots,m_2$$

De externe kwadratuur formule heeft een fout van $\mathcal{O}(h^{l+1})$ (als in (53)) dan en slechts dan als

$$\mathrm{B}(l):\quad q\sum_{j=1}^k b_j c_j^{q-1} = 1 \quad \text{voor } q = 1,\ldots,l.$$

(dit volgt door de formule exact te praten voor de polynomen $1,\chi,\ldots,\chi^{l-1}$; zie het college Numerieke Wiskunde A.) De relaties in $\mathrm{C}(m_1)$ veronderstellen dus dat de interne kwadratuur formules een zekere nauwkeurigheid hebben.

In de literatuur eist men nogal eens in de definitie van een RK methode dat C(1) geldt.

Ter illustratie van de wijze waarop de relaties $\mathrm{C}(m_1)$ en $\mathrm{D}(m_2)$ het aantal "konsistentie" voorwaarden reduceren vermelden we het volgende lemma. Dit lemma is vooral nuttig bij de konstruktie van hogere orde IRK methoden.

**6.1.12 Lemma.** *Beschouw de* RK *methode in* 6.1.2. *Zij $l \in \mathbf{N}$.*
*Als de methode konsistentie orde $l$ heeft dan geldt* B($l$).
*Als de relaties* B($l$), C($m_1$) *en* D($m_2$) *gelden, $l \le m_1 + m_2 + 1$ en $l \le 2m_1 + 2$ dan heeft de* RK *methode konsistentie orde $l$.* □

**6.1.13 Opgave.** Stel we wensen een 10-de orde RK methode te konstrueren. We maken daarbij gebruik van het lemma. Hoeveel beperkingen zullen we dan opleggen aan de koëfficiënten? Vergelijk dit aantal met het aantal in de tabel.
Voor de externe kwadratuur formule gebruiken we nu 5-punts Gauß-Legendre kwadratuur zodat de $b_j$ en $c_j$ bekend zijn. Hoeveel beperkingen zullen we dan nog verder opleggen aan de koëfficiënten?

**6.1.14 Opmerking.** Er zijn expliciete $k$-traps RK methoden gekonstrueerd van orde $l$ met $k = l < 5$, $k-1 = l < 7$, $k-2 = l < 8$ en met $k-3 = l = 8$. De hoogste orde expliciete $k$-traps RK methode die daadwerkelijk gekonstrueerd is een 18-traps methode van orde 10 (Guiness Book of Records).

Door te werk te gaan met Taylor reeksen als in 6.1.10 kan men het volgende resultaat bewijzen (zie ook 4.1.8).

**6.1.15 Bewering.** *Beschouw* 6.1.6. *Zij $m \in \mathbf{N}$.*
*Is de konsistentie orde $l$ en is $f \in C^{m+1}(\Omega, \mathbf{R}^d)$ dan zijn er $\phi_l,\ldots,\phi_m \in C(\mathcal{J},\mathbf{R}^d)$ zodat*

$$\delta_{\mathbf{h}}(t_n) = \sum_{j=l}^m h_n^j\phi_j(t_n) + \mathcal{O}(|\mathbf{h}|^{m+1}) \quad \textit{uniform} \quad (|\mathbf{h}| \to 0).$$ □

**6.1.16 Opmerking.** De lokale diskretisatie fout in een multistep methode van orde $l$ is, voor voldoend gladde $f$ en voldoend kleine stapgrootte, min of meer gelijk is aan $h^l C_{l+1} u^{*(l+1)}$. De lokale diskretisatie fout in een RK methode is daarentegen niet slechts een skalair veelvoud van 'n afgeleide van de exacte oplossing $u^*$. In het algemeen hangt ieder van de $\phi_j$ niet alleen af van de oplossing $u^*$ van (1) en diens afgeleiden maar is iedere $\phi_j$ een niet-lineaire kombinatie van $f$ en diens partiële afgeleiden (zie 6.1.10 en opgave 6.1.17). De kombinatie hangt bovendien essentieel af van de RK methode. ($\delta_{\mathbf{h}}$ hangt niet alleen af van de exacte oplossing $u^*$ maar van $f$ en van $(t_0, u_0)$!)

**6.1.17 Opgave.** a. Pas de expliciete trapezium regel toe
op het probleem $\quad u'(t) = u(t) \quad$ voor $\quad t \in [0,1] \quad$ en $\quad u(0) = 1$
en op het probleem $\quad u'(t) = e^t \quad$ voor $\quad t \in [0,1] \quad$ en $\quad u(0) = 1$.
Bepaal in beide gevallen de lokale diskretisatie fout en vergelijk de twee fouten met elkaar.
b. Bewijs 6.1.15 voor de expliciete trapezium regel.

**Stabiliteit**

RK methoden zijn een-staps methoden. Het "$\rho$-polynoom"heeft slechts een wortel en wel in 1. Om voor voldoende kleine stapgrootte een beperkte fout groei te krijgen hoeven we dan ook geen extra stabiliteits eisen op te leggen. Om het "$f$-deel" af te schatten moeten we nu wel wat meer werk doen dan in het multistep geval. We doen dat in het volgende lemma.

**6.1.18 Lemma.** *Stel* $(\text{Dif.1})_1$ *geldt. Beschouw* 6.1.2.
*Als* $\; V_i = v_i \vec{\mathbf{1}}^T + hF(t, V_i; h)A^T \;$ *voor* $\; i = 1,2 \;$ *dan* $\; \|V_1 - V_2\|_1 \leq q(h)\|v_1 - v_2\|_1$,

*waarbij* $\begin{cases} q(h) := \frac{1}{1 - hL\|A\|_\infty} & \textit{als} \quad hL\|A\|_\infty < 1 \\ q(h) := \sum_{j=0}^{k-1} (hL\|A\|_\infty)^j & \textit{als de methode expliciet is.} \end{cases}$

*Bewijs.* Merk op dat $\|\vec{\mathbf{1}}^T\|_1 = \|\vec{\mathbf{1}}\|_\infty = 1$
Voor $\; B = (b_{ij}) \in \mathbf{M}_{p\times q} \;$ schrijven we $\; |B| := (|b_{ij}|)$.
Voor $B = (b_{ij}), C = (c_{ij}) \in \mathbf{M}_{p\times q}(\mathbf{R})$ schrijven we $B \preceq C$ als $b_{ij} \leq c_{ij}$ voor alle $i, j$.
Merk op dat $\; \|B\| \leq \|C\|_\infty \;$ als $\; 0 \preceq B \preceq C \;$ en dat $\; \|B^T\|_1 = \|B\|_\infty = \|\,|B|\,\|_\infty$.
Als $0 \preceq B \preceq C$ en $E \in \mathbf{M}_p$ is $0 \preceq E$ dan $\; 0 \preceq EB \preceq EC$.
Met $\; \phi := (\|(V_1 - V_2)e_1\|_1, \ldots, \|(V_1 - V_2)e_k\|_1) \;$ is $\; \phi \preceq \|v_1 - v_2\|_1 \vec{\mathbf{1}}^T + hL\phi|A|^T$.
Verder is $\; \|\phi^T\|_\infty = \|V_1 - V_2\|_1$.
Omdat $\; \|\phi^T\|_\infty \leq \|v_1 - v_2\|_1 + hL\|A\|_\infty \|\phi^T\|_\infty \;$ volgt de bewering voor het geval dat $\; hL\|A\|_\infty < 1$. Als $A$ strikt beneden driehoeks is, dan is $(I - hL|A|)^{-1} = \sum_{j=0}^{k-1} (hL|A|)^j \succeq 0$.
Omdat $\; (I - hL|A|)\phi^T \preceq \|v_1 - v_2\|_1 \vec{\mathbf{1}} \;$ volgt $\; \phi^T \preceq \|v_1 - v_2\|_1 (I - hL|A|)^{-1}\vec{\mathbf{1}} \;$ en dus $\|\phi^T\|_\infty \leq \sum_{j=0}^{k-1} (hL\|A\|_\infty)^j \|v_1 - v_2\|_1$. □

**6.1.19 De stabiliteits stelling.** *Beschouw* 6.1.6. *Stel* $(\text{Dif.1})_1$ *geldt.*
*Zij* $\mathbf{h}$ *een stapgrootterij met voldoende kleine maaswijdte.*
*Schrijf, met* $q(h)$ *als in* 6.1.18, $\quad Q := \max_n q(h_n) \quad$ *en* $\quad \tilde{L} := QL\|\vec{b}\|_1$.
*Stel dat* $u_{\mathbf{h}}$ *en* $\tilde{u}_{\mathbf{h}}$ *in* $C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ *oplossingen zijn van de* RK *methode met lokale fout* $\epsilon_{\mathbf{h}}$*, repektievelijk* $\tilde{\epsilon}_{\mathbf{h}}$ (zie 6.1.6). *Dan geldt, voor iedere* $t_n \in \mathcal{J}_{\mathbf{h}}$*, dat*

$$\|u_{\mathbf{h}}(t_n) - \tilde{u}_{\mathbf{h}}(t_n)\|_1 \leq e^{\tilde{L}(t_{n-1} - t_0)} \left( \|u_{\mathbf{h}}(t_0) - \tilde{u}_{\mathbf{h}}(t_0)\|_1 + \sum_{j=0}^{n-1} \|\epsilon_{\mathbf{h}}(t_j) - \tilde{\epsilon}_{\mathbf{h}}(t_j)\|_1 \right).$$

*Bewijs.* Met behulp van het lemma volgt eenvoudig dat
$\|u_{\mathbf{h}}(t_{n+1}) - \tilde{u}_{\mathbf{h}}(t_{n+1})\|_1 \leq (1 + h_n L\|\vec{b}\|_1 Q)\|u_{\mathbf{h}}(t_n) - \tilde{u}_{\mathbf{h}}(t_n)\|_1 + \|\epsilon_{\mathbf{h}}(t_n) - \tilde{\epsilon}_{\mathbf{h}}(t_n)\|_1$.
Op de gebruikelijke manier volgt nu de stelling. □

**6.1.20 Opgave.** Formuleer en bewijs zelf een stelling waarin de existentie van de Runge-Kutta oplossing $u_{\mathbf{h}}$ met lokale fout $\epsilon_{\mathbf{h}}$ de existentie impliceert van de Runge-Kutta oplossing $\tilde{u}_{\mathbf{h}}$ met lokale fout $\tilde{\epsilon}_{\mathbf{h}}$ (zie de stabiliteits stelling 4.1.26 voor multistep methoden).

### Konvergentie

Bij een RK methode hebben we, als de maaswijdte klein genoeg is, geen stabiliteits problemen: dus "**konvergentie = konsistentie**". De volgende stelling volgt weer onmiddellijk uit diens kwantitatieve variant in 6.1.22. We laten het bewijs van deze stellingen over aan de geïnteresseerde lezer.

**6.1.21 Hoofdstelling.** *Stel* (Dif.0) *en* (Dif.1) *gelden.*
(a) *Een* RK *methode is konvergent dan en slechts dan als hij konsistent is.*
(b) *De konvergentie is van orde $l$ dan en slechts dan als de methode konsistent van orde $l$ is.* □

**6.1.22 Konvergentie stelling.** *Stel* (Dif.0) *en* $(\text{Dif.1})_1$ *gelden.*
*Zij* $\tilde{u}_0 \in \mathbf{R}^d$ *en* $\epsilon_{\mathbf{h}} \in C(\mathcal{J}, \mathbf{R}^d)$. *Als* $|\mathbf{h}|$, $\sup_t \|\epsilon_{\mathbf{h}}(t)\|_1$ *en* $\|u_0 - \tilde{u}_0\|_1$ *voldoende klein zijn dan heeft de* RK *methode in* 6.1.6 *een oplossing* $\tilde{u}_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$ *met* $\tilde{u}_{\mathbf{h}}(t_0) = \tilde{u}_0$ *en lokale fout* $\epsilon_{\mathbf{h}}$. *Met* $Q$ *en* $\tilde{L}$ *als in* 6.1.19 *geldt*

$$\|u^*(t_n) - \tilde{u}_{\mathbf{h}}(t_n)\|_1 \leq \left(\|u_0 - \tilde{u}_0\|_1 + \sum_{j=0}^{n-1}(\|\epsilon_{\mathbf{h}}(t_j)\|_1 + h_j\|\delta_{\mathbf{h}}(t_j)\|_1)\right) e^{\tilde{L}(t_n - t_0)}.$$

*Als, voor iedere* $t_n \in \mathcal{J}_{\mathbf{h}}$, $Q\times$ *de uitdrukking in het rechterlid* $\leq \tilde{r}$ *dan zijn de startfout, de lokale fout* $\epsilon_{\mathbf{h}}$ *en* $|\mathbf{h}|$ *voldoende klein: dan bestaat* $\tilde{u}_{\mathbf{h}}$. □

**6.1.23 Gevolg.** *Voor* $|\mathbf{h}|$ *voldoende klein heeft de* RK *methode met konsistentie orde* $l \geq 1$ *een exacte oplossing* $u_{\mathbf{h}}$ *met* $u_{\mathbf{h}}(t_0) = u_0$ *en er geldt dat*

$$\sup_{t_n} \|u^*(t_n) - u_{\mathbf{h}}(t_n)\|_1 \leq Te^{\tilde{L}T} \max_j \|\delta_{\mathbf{h}}(t_j)\|_1 \leq Te^{\tilde{L}T}|\mathbf{h}|^l \sup_t \|\phi_l(t)\|_1 + \mathcal{O}(|\mathbf{h}|^{l+1}).$$

*Als* $Q\times$ *de uitdrukking in het middelste lid* $\leq \tilde{r}$ *dan is* $\mathbf{h}$ *voldoende klein.* □

## 6.2 De struktuur van de globale fout in een RK methode

Werken we telkens met een *uniforme stapgrootterij* $\mathbf{h} = (h_n)$ (d.w.z. $h_n = |\mathbf{h}|$ voor iedere $n$), is $f$ glad genoeg dan is ook hier, bij RK methoden, de globale diskretisatie fout voor kleine $|\mathbf{h}|$ min of meer glad. Bovendien kunnen we (voor zeer gladde $f$) een asymptotische ontwikkeling van die de fout geven. Bij een RK methode kunnen we exact starten. Verder heeft het "$\rho$-polynoom" geen parasitaire wortels. Dit vereenvoudigt het bewijs voor RK methoden enigszins. De behandeling van de $f$-term is daarentegen gekompliceerder dan bij multistep methoden. Zonder bewijs vermelden we de volgende stelling (zie 4.3.7 en 4.3.18).

**6.2.1 Stelling.** *Stel* (Dif.0) *en* (Dif.3) *gelden.*
*Beschouw* 6.1.6 *telkens met stapgrootterij* $\mathbf{h}$ *zodat* $h := h_n = |\mathbf{h}|$ *voor iedere* $n$.
*Laat, voor iedere* $\mathbf{h}$ *met* $|\mathbf{h}|$ *voldoende klein,* $u_{\mathbf{h}}$ *de exacte Runge-Kutta oplossing zijn met* $u_{\mathbf{h}}(t_0) = u_0$. *Als de* RK *methode konsistentie orde* $l$ *heeft dan zijn er, met* $m$ *als in* (Dif.3), *voor* $j = 0, \ldots, m-1$, *funkties* $w_j \in C^{m-j}(\mathcal{J}, \mathbf{R}^d)$ *met* $w_j(t_0) = 0$ *zodat*

$$u_{\mathbf{h}} = u^* - h^l w_0 - \ldots - h^{l+m-1} w_{m-1} + \mathcal{O}(h^{m+l}) \quad \textit{uniform} \quad (h \to 0).$$

*Met* $\phi_l$ *als in* 6.1.15 *is* $w_0$ *de oplossing van het probleem*

$$w' = Jw + \phi_l \quad \textit{op} \quad \mathcal{J} \quad \textit{en} \quad w(t_0) = 0.$$ □

### 6.2.2 Symmetrische RK methoden.

Beschouw een $k$-traps RK methode als in 6.1.2.
Definieer de afbeelding $J : \mathbf{R}^k \to \mathbf{R}^k$ door $Je_j = e_{k+1-j}$ voor $j = 1, \ldots, k$. De *geadjungeerde* RK methode heeft genererende matrix

$$\left[\begin{array}{c|c} J\vec{c} & J(\vec{b}^T\,\vec{\mathbf{1}} - A)J \\ \hline 0 & (J\vec{b})^T \end{array}\right].$$

(De geadjungeerde methode is in feite de oorspronkelijke methode met stapgrootte $-h$ waarbij we toch voorwaarts itereren: $u_n = u_{\mathbf{h}}(t_0 + T - nh)$ en we bepalen $u_n$ uit $u_{n+1}$.)
Als de genererende matrices van de methode en van diens geadjungeerde samenvallen zeggen we dat de RK methode *symmetrisch* is.

**Voorbeelden.** De geadjungeerde van de Euler forward methode is de Euler backward methode. De impliciete trapezium regel is symmetrisch.

**Bewering.** *De lokale en de globale fout in een symmetrische* RK *methode heeft een asymptotische ontwikkeling in machten van* $h^2$. □

### 6.2.3 Romberg schema's.

Zij $(\tilde{t}, x) \in \Omega$. Laat $u_h$ de Runge-Kutta oplossing zijn door $(\tilde{t}, x)$ met uniforme stapgrootterij $\mathbf{h}$ met maaswijdte $h := |\mathbf{h}|$. De RK methode heeft konsistentie orde $l$.

Kies een $\mu \in \mathbf{N}_0$ en kies een stijgende rij $n_0 = 1, n_1, n_2, \ldots$ in $\mathbf{N}$. Beschouw een $h > 0$ (voldoende klein). Schrijf $h_j := \frac{1}{n_j} h \quad (j \in \mathbf{N}_0)$ en stel een Romberg schema op:

$$\begin{array}{lllll}
T_{00} := u_{h_0}(\tilde{t} + h) & & & & \\
T_{10} := u_{h_1}(\tilde{t} + h) & T_{11} & & & \\
T_{20} := u_{h_2}(\tilde{t} + h) & T_{21} & T_{22} & & \\
\qquad\qquad\quad \vdots & \vdots & \vdots & \ddots & \\
T_{\mu 0} := u_{h_\mu}(\tilde{t} + h) & T_{\mu 1} & T_{\mu 2} & \ldots & T_{\mu\mu}
\end{array}$$

waarbij de $T_{ij}$ als volgt door extrapolatie gegeven zijn.
Voor iedere $i, j$ is $T_{ij} = p(0)$ waarbij

$$p(\zeta) = \gamma_0 + \gamma_1 \zeta^l + \ldots + \gamma_j \zeta^{l-1+j} \qquad (\zeta \in \mathbf{C})$$

en de $\gamma_\iota \in \mathbf{R}^d$ bepaald zijn door $p(h_\iota) = T_{\iota 0}$ voor $\iota = i - j, \ldots, i$
De $T_{ij}$ zijn efficiënt rekursief te berekenen. Als $l = 1$ kan dat als volgt:

$$T_{j\,i+1} = T_{ji} + \frac{T_{ji} - T_{j-1\,i}}{(n_j / n_{j-i}) - 1}.$$

Als $u$ de exacte oplossing is van (51) door $(\tilde{t}, x)$ dan volgt uit stelling 6.2.1 dat

$$\|u(\tilde{t} + h) - T_{\mu\mu}\| \leq C h^{l+\mu} \|w_\mu(\tilde{t} + h)\| + \mathcal{O}(h^{l+\mu+1}) = \mathcal{O}(h^{l+\mu+1}) \quad (h \to 0),$$

waarbij $C \in (0, \infty)$ alleen afhangt van $l$, $\mu$ en de rij $(n_j)$; bedenk voor de laatste schatting dat $w_\mu(\tilde{t}) = 0$ en dus $w_\mu(\tilde{t} + h) = \mathcal{O}(h)$.

Bij gegeven $\mu$ en rij $(n_j)$ definieert het Romberg schema een lokale benadering $\tilde{u}$ van $u$. Men kan $\tilde{u}$ ook zien als een lokale Runge-Kutta oplossing met een $\mu k$-traps methode. Deze nieuwe RK methode is expliciet als de oorspronkelijke dat is en heeft konsistentie orde $l + \mu$. (Hiermee is gelijk bewezen dat er ERK methoden bestaan van orde $l$ voor iedere $l$; zie 6.1.9.)

Voor de rij $(n_j)$ gebruikt men nogal eens de *Romberg rij* $1, 2, 4, 8, 16, \ldots$ of de *Bulirsch rij* $1, 2, 3, 4, 6, 8, 12, 16, 24, 32, \ldots$. Voor hogere orde benaderingen heeft de Bulirsch rij minder funktie evaluaties nodig dan de Romberg rij.

### 6.2.4 Opmerking.

Als $f$ voldoende glad is, is de globale diskretisatie fout in een stabiele multistep methode van orde $l$, voor voldoende kleine $h$, min of meer gelijk aan $h^l C w^*$, waarbij de foutkonstante $C$ afhangt van de methode. De funktie $w^*$ hangt alleen af van $l$ en van het probleem

(1), niet van de stapgrootte $h$ en afgezien van $l$ ook *niet* van de methode. Dit resultaat volgde in 4.3.7 uit het feit dat de lokale diskretisatie fout in een multistep methode min of meer gelijk is aan $h^l C_{l+1} u^{*(l+1)}$. De lokale diskretisatie fout in een RK methode hangt echter fundamenteler af van de methode (zie 6.1.16). In het algemeen zal de globale diskretisatie fout in twee RK methoden van konsistentie orde $l$ dan ook *niet* op een skalair veelvoud na min of meer aan elkaar gelijk zijn ook niet als we dezelfde stapgrootterij gebruiken en de maaswijdte zeer klein is.

## 6.3 Stapgrootte besturing bij RK methoden

In deze paragraaf laten we zien hoe we in de benadering $u_{\mathbf{h}}$ de fout, die gemaakt is op het interval $[t_n, t_{n+1})$, kunnen schatten. Wat de bijdrage van deze "lokale fout" tot de globale fout is hangt weer af van de stabiliteit van het probleem (1) (zie het betoog in 4.4.2).

Door met stapgrootte $h$ te werken én met stapgrootte $2h$ kunnen we al rekenend de fout schatten.

**6.3.1 Rekenen, ook met dubbele stapgrootte.** Beschouw 6.1.6, waarbij de RK methode konsistentie orde $l$ heeft. Zij $\mathbf{h} = (h_n)$ een stapgrootterij met $h_{2n} = h_{2n+1} \quad (n \in \mathbf{N}_0)$.

Laat $u_{\mathbf{h}}$ de exacte RK oplosingen zijn bij deze stapgrootterij.
We kunnen de fout die gemaakt is op het intervalletje $[t_{2n}, t_{2n+2})$ in de benadering $u_{\mathbf{h}}$ van $u^*$, voor voldoend kleine maaswijdte, schatten door de volgende grootheid[22].

$$\frac{1}{2^l - 1}[u_{\mathbf{h}}(t_{2n+2}) - \tilde{u}(t_{2n+2})],$$

hierin is $\tilde{u}$ de lokale Runge-Kutta oplossing door $(t_{2n}, u_{\mathbf{h}}(t_{2n}))$, als in 6.1.2, met $h = 2h_{2n}$.

Berekenen we $u_{\mathbf{h}}$ en schatten we de fout op deze manier in iedere stap dan kost ons dat 50 % meer dan het berekenen van $u_{\mathbf{h}}$ zonder foutschatting.

De lokale diskretisatie fout in een sterk stabiele multistep methode van orde $l$ konden we schatten met behulp van een andere multistep methode van orde $l$ (met een prediktor-korrektor methode; zie 4.4.9). We konden dit omdat de lokale diskretisatie fout in de ene methode min of meer een skalair veelvoud is van die fout in de andere. We mogen, gezien onze observaties in 6.1.16 en 6.2.4, niet verwachten dat deze strategie werkt voor RK methoden. We kunnen voor deze methoden wel de lokale diskretisatie fout schatten met behulp van een RK methode die een orde nauwkeuriger is.

**6.3.2 Al rekenend de lokale fout schatten.** Beschouw 6.1.6. Laat $l$ de konsistentie orde zijn van de RK methode. Laat $u_{\mathbf{h}}$ de exacte Runge-Kutta oplossing zijn.
Beschouw ook een tweede RK methode die konsistentie orde $l+1$ heeft.

Dan kunnen we de fout die gemaakt is op het intervalletje $[t_n, t_{n+1})$ in de benadering $u_{\mathbf{h}}$ van $u^*$ als volgt schatten.
Laat $\hat{u}$ de lokale Runge-Kutta benadering zijn met de tweede RK methode door $(t_n, u_{\mathbf{h}}(t_n))$.
Omdat de tweede methode konsistentie orde $l+1$ heeft is

$$\hat{u}(t_n + h) - u_{\mathbf{h}}(t_n + h) = u(t_n + h) - u_{\mathbf{h}}(t_n + h) + \mathcal{O}(h^{l+2}) \quad (h \to 0),$$

waarbij $u$ de exacte oplossing is van (51) door $(t_n, u_{\mathbf{h}}(t_n))$. We schatten de "lokale" fout dan ook door $\hat{u}(t_{n+1}) - u_{\mathbf{h}}(t_{n+1})$.

Alleen al het schatten van de lokale fout in iedere stap op de aangegeven manier is, als we zomaar een RK methode van orde $l+1$ nemen, minstens even duur als het berekenen van

[22] De fout die gemaakt wordt op het interval $[t_{2n}, t_{2n+1})$ is $h_{2n}\times$ de lokale diskretisatie fout in $t_{2n}$ ten opzichte van de exacte oplossing door $(t_{2n}, u_{2\mathrm{h}}(t_{2n}))$. In feite schatten we deze lokale fout. Zie ook 6.2.3.

de Runge-Kutta oplossing $u_{\mathbf{h}}$ zelf. Het zou bijzonder kosten besparend zijn als de genererende matrix van de tweede RK methode gegeven is door $\left[\begin{array}{c|c} \vec{c} & A \\ \hline 0 & \vec{d}^T \end{array}\right]$, waarbij $\left[\begin{array}{c|c} \vec{c} & A \\ \hline 0 & \vec{b}^T \end{array}\right]$ de genererende matrix is van de eerste RK methode. We willen dus dat de interne kwadratuurformules van de twee methoden identiek zijn. Alleen de gewichten van de externe kwadratuurformules mogen verschillend zijn. De “dure” $f$-waarden, $F(t_n, U; h_n)$, hoeven dan per stap maar een keer bepaald te worden: met $u_n := u_{\mathbf{h}}(t_n)$ is

$$u_{n+1} = u_n + h_n F(t_n, U_n; h_n)\vec{b} \quad \text{met } U_n \text{ zodat } \; U_n = u_n \vec{\mathbf{1}}^T + h_n F(t_n, U_n; h_n) A^T$$

en

$$\hat{u}_{n+1} - u_{n+1} := h_n F(t_n, U_n; h_n)(\vec{d} - \vec{b})$$

is de schatting van de fout in $u_{\mathbf{h}}$ die gemaakt is op $[t_n, t_{n+1})$.

Zo’n paar van RK methoden noemt men een *bevattende* RK *methode* (“embedded RK method”). Men representeert zo’n bevattende RK methode door middel van de genererende matrix

$$\left[\begin{array}{c|c} \vec{c} & A \\ \hline 0 & \vec{b}^T \\ \hline 0 & \vec{d}^T \end{array}\right].$$

**Voorbeeld [Ceschino (1962)].**

$$\left[\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 1 & -2 & 2 & 0 \\ \hline 0 & 1 & -2 & 2 & 0 \\ \hline 0 & \frac{1}{6} & 0 & \frac{2}{3} & \frac{1}{6} \end{array}\right]$$

representeert een bevattende RK methode van orde 2 waarin de foutschatter orde 4 heeft. Merk op dat de laatste interne kwadratuur formule gelijk is aan de externe kwadratuurformule; dit bespaart tijdens de berekening een matrix vektor vermenigvuldiging.

De gedachte ligt voor de hand om de RK methode waarmee we rekenen zo te willen kiezen dat de koëfficiënten in de lokale diskretisatie fout zo klein mogelijk zijn. De koëfficiënten in de diskretisatie fout van de schatter zullen dan wat groter zijn. Omdat we hier met orde uitspraken te maken hebben lopen we hierdoor echter het risiko dat we de fout onderschatten: we nemen aan dat de hogere orde termen verwaarloosbaar zijn terwijl ze dat in werkelijkheid nog niet zijn.

Als de foutschatter een orde nauwkeuriger is dan de RK methode waarmee we itereren dan is het aantrekkelijker om de rollen van de twee methoden om te wisselen; we mogen dan hopen dat we daardoor een betere benadering bereken—de foutschatting zal ons dat evenwel niet vertellen.

**Voorbeeld [Dormand-Prince (1980)].**

$$\left[\begin{array}{c|ccccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3}{10} & \frac{3}{40} & \frac{9}{40} & 0 & 0 & 0 & 0 & 0 \\ \frac{4}{5} & \frac{44}{45} & -\frac{56}{15} & \frac{32}{9} & 0 & 0 & 0 & 0 \\ \frac{8}{9} & \frac{19372}{6561} & -\frac{25360}{2187} & \frac{64448}{6561} & -\frac{212}{729} & 0 & 0 & 0 \\ 1 & \frac{9017}{3168} & -\frac{355}{33} & \frac{46732}{5247} & \frac{49}{176} & -\frac{5103}{18656} & 0 & 0 \\ 1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline 0 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & -\frac{2187}{6784} & \frac{11}{84} & 0 \\ \hline 0 & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & \frac{92097}{339200} & \frac{187}{2100} & \frac{1}{40} \end{array}\right]$$

representeert een RK methode van orde 5 met een schatter van orde 4.

## 6.4 Stabiliteit van RK methoden bij grotere stapgrootte

In deze paragraaf vragen we ons af onder welke omstandigheden lokale fouten gedempt worden voortgeplant zodat de globale absolute fout min of meer gemajoreerd wordt door de som van de lokale absolute fouten of zelfs door het maximum van de laatste paar lokale absolute fouten. De inspiratie, motivaties en konklusies in deze paragraaf voor RK methoden zijn vaak analoog aan die voor de multistep methoden in §1.11. Onze tekst is hier dan ook beknopt.

In de volgende alinea geven we aan dat de foutvoortplanting ook hier weer beschreven kan worden met Greense funkties.

**6.4.1 De foutvoortplanting in een RK methode.** Beschouw de situatie in 6.1.6. Laat $\tilde{u}_{\mathbf{h}}$ een Runge-Kutta oplossing zijn met lokale fout $\epsilon_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$. Dan is $\tilde{e}_{\mathbf{h}} := u^* - \tilde{u}_{\mathbf{h}}$ op $\mathcal{J}_{\mathbf{h}}$ de totale fout. De rooster funktie $\tilde{e}_{\mathbf{h}}$ is de Runge-Kutta oplossing van het gelineariseerde probleem $u' = Ju$ op $\mathcal{J}$ met $\tilde{e}_{\mathbf{h}}(t_0) = \mu_{-1} := u^*(t_0) - \tilde{u}_{\mathbf{h}}(t_0)$ (de startfout) en met lokale fout in $t_n$ die gegeven is door $\mu_n := h_n\delta_{\mathbf{h}}(t_n) - \epsilon_{\mathbf{h}}(t_n) + \mathcal{O}(\|\tilde{e}_{\mathbf{h}}(t_n)\|^2)$.

Laat $G_{\mathbf{h}}$ de funktie zijn van $\mathcal{J}_{\mathbf{h}} \times \mathcal{J}_{\mathbf{h}}$ naar $\mathbf{M}_d$ zodat (zie ook 4.6.1)
$G_{\mathbf{h}}(t_n, t_\nu) = 0$ voor $t_n < t_\nu$, $G_{\mathbf{h}}(t_\nu, t_\nu) = I$,
en voor iedere $\nu$ is $t_n \to G_{\mathbf{h}}(t_n, t_\nu)$ de exacte Runge-Kutta oplossing in $C(\{t_n \in \mathcal{J}_{\mathbf{h}} \,\big|\, t_n \geq t_\nu\}, \mathbf{M}_d)$ van het probleem $U' = JU$.
Dan geldt weer als in 4.6.1 dat

$$\tilde{e}_{\mathbf{h}}(t_n) = \sum_{\nu=0}^{n} G_{\mathbf{h}}(t_n, t_\nu)\mu_{\nu-1} \quad (t_n \in \mathcal{J}_{\mathbf{h}}).$$

**Lineaire vergelijkingen:** $d = 1$

**6.4.2 Het probleem.** Beschouw, met $d = 1$, het lineaire probleem in 1.1.11 met $g = 0$.
**Foutvoortplanting in de RK methode.** Beschouw de situatie in 6.1.6. Als $\tilde{u}_{\mathbf{h}}$ de Runge-Kutta oplossing is met lokale fout $\epsilon_{\mathbf{h}}$ dan is, met $u_n := \tilde{u}_{\mathbf{h}}(t_n)$ en $\epsilon_n = \epsilon_{\mathbf{h}}(t_n)$,
$u_{n+1} = u_n + h_n F(t_n, U; h_n)\vec{b} + \epsilon_n$ met $U$ zodat $U = u_n \vec{\mathbf{1}}^T + h_n F(t_n, U; h_n)A^T$.
Omdat $F(t_n, U; h_n) = \eta U$ is $U^T = (I - h_n\eta A)^{-1}\vec{\mathbf{1}}u_n$ en dus
$u_{n+1} = [1 + h_n\eta\,(\vec{b}, [I - h_n\eta A]^{-1}\vec{\mathbf{1}}\,)\,]u_n + \epsilon_n$. Schrijf

$$\lambda(\zeta) := 1 + \zeta\,(\vec{b}, [I - \zeta A]^{-1}\vec{\mathbf{1}}\,) \quad \text{voor} \quad \zeta \in \mathbf{C}. \tag{55}$$

Dus $u_{n+1} = \lambda(h_n\eta)u_n + \epsilon_n$. Omdat evenzo $u^*(t_{n+1}) = \lambda(h_n\eta)u^*(t_n) + h_n\delta_{\mathbf{h}}(t_n)$ volgt dat $\tilde{e}_{\mathbf{h}}(t_{n+1}) = \lambda(h_n\eta)\tilde{e}_{\mathbf{h}}(t_n) + \mu_n$, met $\mu_n := h_n\delta_{\mathbf{h}}(t_n) - \epsilon_n$. Voor dit $d = 1$ geval met konstante koëfficiënt volgt nu eenvoudig dat

$$G_{\mathbf{h}}(t_n, t_\nu) = \prod_{j=\nu}^{n-1} \lambda(h_j\eta) \quad \text{voor} \quad n > \nu.$$

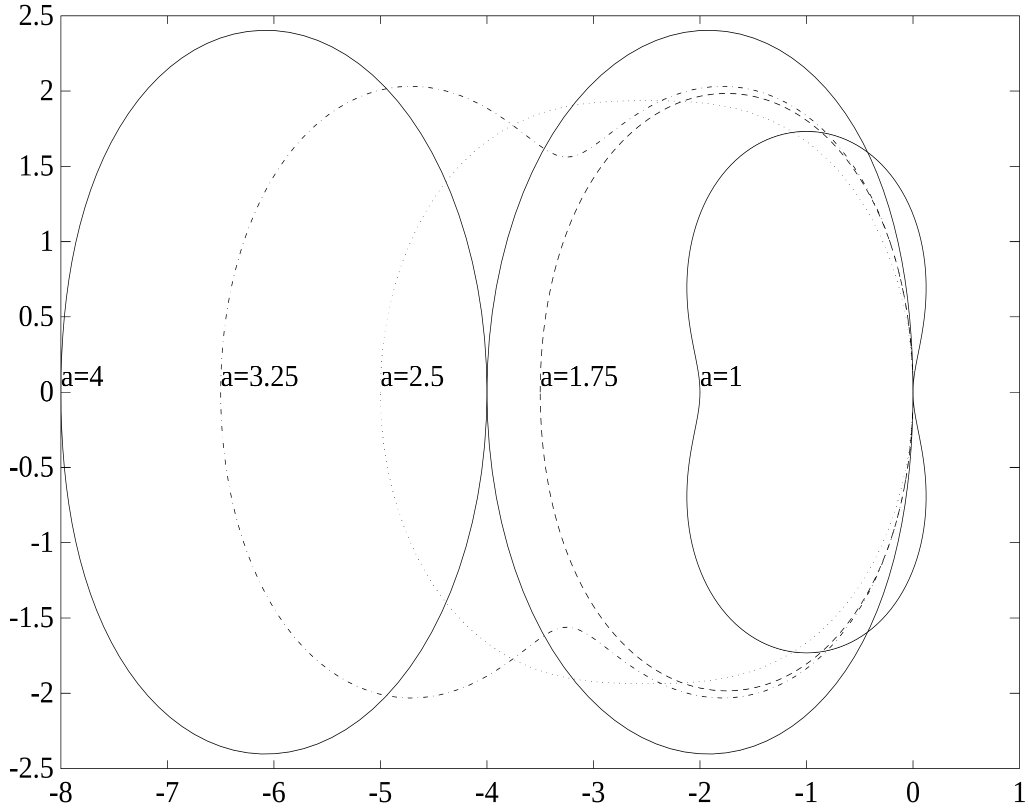
Zeker als $\eta < 0$ zouden we ook hier weer graag zien dat voor de *groeifaktor* $\lambda(h_n\eta)$ geldt (vergelijk met (43))

$$|\lambda(h_n\eta)| < 1 \quad \text{voor iedere} \quad n. \tag{56}$$

**6.4.3 Opmerking.** Als de methode van orde $l$ is dan is $\lambda^{(j)}(0) = \frac{1}{j!}$ voor $j = 0, \ldots, l$ (pas de RK methode toe als in 6.1.2 met $t = 0$ op het probleem $u' = u$, $u(0) = 1$. Dan $\tilde{u}(h) = \lambda(h)$ en $u(h) = e^h$): het $l$-de graads Taylor polynoom van $\lambda$ rond 0 en van $e^\chi$ rond 0 vallen samen.

Als de $k$-traps RK methode expliciet is dan is $\lambda(\zeta)$ een polynoom in $\zeta$ van graad $k$, immers: $(I - \zeta A)^{-1} = \sum_{j=0}^{k-1}(\zeta A)^j$. Dus als de expliciete $k$-traps konsistentie orde $k$ heeft dan is

$$\lambda(\zeta) = \sum_{j=0}^{k} \frac{1}{j!}\zeta^j \quad (\zeta \in \mathbf{C}).$$




Figuur 2: Absolute stabiliteits gebieden voor verschillende (a =)$\alpha$'s (zie 6.4.5).

(Dit beschrijft de $\lambda$ voor de vier RK methoden in voorbeeld 6.1.3.)

Als de RK methode impliciet is dan is $\lambda(\zeta)$ een rationale funktie in $\zeta$ (Waarom?).

De foutvoortplanting bij een ERK methode toegepast op het probleem in 1.1.11 wordt blijkbaar bepaald door een polynoom *waarde*, bij een IRK methode door de *waarde* van een rationale funkties. Bij een mulitstep methode wordt de foutvoortplanting bepaald door de *wortels* van een rationale funktie (van $\frac{\rho}{\sigma} - \zeta$. Het Taylor polynoom van $\frac{\rho}{\sigma}$ rond 1 en van log vallen samen!).

**6.4.4 Definitie.** Voor $\mu \in [0, \infty)$ is

$$\mathcal{S}(\mathbf{A})_\mu := \{\, \zeta \in \mathbf{C} \,\big|\, |\lambda(\zeta)| < \mu \,\}$$

het *stabiliteits gebied* van de RK methode met genererende matrix $\mathbf{A}$ ten opzichte van de *groeifaktor* $\mu$. $\mathcal{S}(\mathbf{A}) := \mathcal{S}(\mathbf{A})_1$ is het *absolute stabiliteits gebied* van de RK methode.

**6.4.5 Voorbeelden.** Beschouw, voor $\alpha \in \mathbf{R}$, de expliciete 2-traps RK methode met genererende matrix

$$\left[\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline 0 & 0.5 & 0.5 \end{array}\right].$$

Dan $\lambda(\zeta) = 1 + \zeta + \frac{1}{2}\alpha\zeta^2$.
Met $\alpha = \frac{1}{2}$ is het absolute stabiliteits gebied een cirkelschijf in complexe vlak rond -2 met straal 2. Met $\alpha = \frac{1}{4}$ bevat het stabiliteits gebied $(-8, 0)$. (zie prent 1 met $\alpha = \frac{m}{2}$, $m = 2, 3, 4, \ldots, 8$.)

RK methoden waarvan het absolute stabiliteits gebied het linker complexe halfvlak omvat zijn weer van speciaal belang bij het oplossen van met name stijve differentiaalvergelijkingen. Omdat de groeifaktor $\lambda(\zeta)$ van een ERK methode een polynoom is in $\zeta$ is het absolute stabiliteits gebied van deze methode begrensd. A-stabiele RK methoden zijn dus impliciet.

**6.4.6 Definitie.** De RK methode met genererende matrix $\mathbf{A}$ is A-*stabiel* als

$$\mathbf{C}_- := \{\zeta \in \mathbf{C} \,\big|\, \mathrm{Re}(\zeta) < 0\} \subset \mathcal{S}(\mathbf{A}).$$

**6.4.7 Bewering.** *Iedere* A-*stabiele* RK *methode is impliciet.*
*Er bestaan* A-*stabiele* DIRK *methoden* (zie 6.4.16). $\square$

**6.4.8 Voorbeeld.** Beschouw de impliciete trapezium regel met genererende matrix

$$\left[\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 0.5 & 0.5 \\ \hline 0 & 0.5 & 0.5 \end{array}\right].$$

Dan $\lambda(\zeta) = \frac{2+\zeta}{2-\zeta}$. De impliciete trapezium regel is A-stabiel.

Een A-stabiele RK methode is er op afgericht $d = 1$ problemen met een konstante koëfficiënt $\eta$ in $\mathbf{C}_-$ stabiel op te lossen voor iedere stapgrootterij $\mathbf{h}$. Als $f$ een gladde funktie is, $t \to J(t) = \frac{\partial f}{\partial x}(t, u^*(t))$ varieert langzaam en het spektrum van $J(t)$ ligt, voor iedere $t$, in $\mathbf{C}_-$ dan zal zo'n A-stabiele methode het probleem ook wel stabiel oplossen met niet al te kleine stapgrootte $|\mathbf{h}|$. Het is echter onduidelijk of we ook met een niet al te kleine stapgrootte aan de slag kunnen als $f$ of $J$, met spektrum in $\mathbf{C}_-$, sterk varieert. De volgende beschouwing laat zien dat we dat in het algemeen niet kunnen.

**6.4.9 Notatie.** Als $\vec{x} = (x_1, \ldots, x_k)^T \in \mathbf{R}^k$ dan is $\operatorname{diag}(\vec{x})$ de diagonaal matrix $D = (d_{ij}) \in \mathbf{M}_k$ met $d_{ii} = x_i \quad (i = 1, \ldots, k)$.

**6.4.10 Het probleem.** We bekijken nogmaals het $d = 1$ probleem in 1.1.11, waarin nu $\eta \in C(\mathcal{J}, \mathbf{C}_-)$ (en $\eta$ mag sterk variëren: $h \ll \sup\{|\eta(t) - \eta(s)| \,\big|\, |t - s| \leq h\}$).
**Foutvoortplanting in de RK methode.** Beschouw een $t_n \in \mathcal{J}_\mathbf{h}$. Schrijf $\tilde{t} = t_n$ en $h = h_n$.
Verder is $\vec{\eta} = (\eta_1, \ldots, \eta_k)^T := (\eta(\tilde{t} + hc_1), \ldots, \eta(\tilde{t} + hc_k))^T \in \mathbf{R}^k$.
Als $u_{n+1} = u_n + hF(\tilde{t}, U; h)\vec{b}$ met $U$ zodat $U = u_n\vec{\mathbf{1}}^T + hF(\tilde{t}, U; h)A^T$ dan $F(\tilde{t}, U; h) = U \operatorname{diag}(\vec{\eta})$. Dus

$$U^T = u_n[I - hA\operatorname{diag}(\vec{\eta})]^{-1}\vec{\mathbf{1}} \quad \text{en} \quad u_{n+1} = [1 + (h\operatorname{diag}(\vec{\eta})\vec{b}, [I - hA\operatorname{diag}(\vec{\eta})]^{-1}\vec{\mathbf{1}})]u_n.$$

Definieer

$$\Lambda(\vec{\zeta}) := 1 + (\operatorname{diag}(\vec{\zeta})\vec{b}, [I - A\operatorname{diag}(\vec{\zeta})]^{-1}\vec{\mathbf{1}}) \quad \text{voor alle} \quad \vec{\zeta} \in \mathbf{C}^k. \tag{57}$$

In de $n$-de RK iteratie stap wordt de fout voortgeplant met de faktor $\Lambda(h\vec{\eta})$. Als

$$|\Lambda(\vec{\zeta})| < 1 \quad \text{voor iedere} \quad \vec{\zeta} \in (\mathbf{C}_-)^k \tag{58}$$

dan wordt de globale absolute fout in bovenstaand $d = 1$ probleem gemajoreerd door

$$T \max_n |\delta_\mathbf{h}(t_n)|.$$

**6.4.11 Voorbeeld.** Voor de impliciete trapezium regel (zie 6.4.8) is

$$\Lambda((\zeta_1, \zeta_2)^T) = \frac{2+\zeta_1}{2-\zeta_2} \qquad ((\zeta_1, \zeta_2)^T \in \mathbf{C}^2).$$

We zien dat $|\Lambda((\zeta_1, \zeta_2)^T)| > 1$ als bijvoorbeeld $\zeta_1, \zeta_2 \in (-\infty, 0)$ en $\zeta_2 < -4 + \zeta_1$.

Niet voor iedere A-stabiele RK methode geldt (58) (zie 6.4.8 en 6.4.11). We kunnen dus niet met iedere A-stabiele RK methode het probleem (1) met 0-samentrekkende $f$ nauwkeurig oplossen met een niet al te kleine stapgrootte. De relatief eenvoudige algebraïsche eis (58) vertelt ons, volgens de volgende stelling 6.4.14, echter wel of 'n RK methode dat kan. Stelling 6.4.14 is de RK variant van de mulitstep stelling 4.6.27. Om voor multistep methoden te bewijzen dat een A-stabiele methode ook problemen met een 0-samentrekkende funktie $f$ nauwkeurig oplost met niet al te kleine stapgrootte moesten we ook een modifikatie aanbrengen. Hier, bij RK methoden, verzwaren we de algebraïsche eis, bij multistep methoden pasten we de methode enigsins aan: we gingen daar aan de slag met een-been varianten.

**6.4.12 Definitie.** Beschouw 6.1.6.
Beschouw een tweetal exacte Runge-Kutta oplossing $w_{\mathbf{h}}, v_{\mathbf{h}} \in C(\mathcal{J}_{\mathbf{h}}, \mathbf{R}^d)$. Stel dat

$$\|w_{\mathbf{h}}(t_n) - v_{\mathbf{h}}(t_n)\|_2 \leq \|w_{\mathbf{h}}(t_\nu) - v_{\mathbf{h}}(t_\nu)\|_2 \quad \text{voor alle } \; t_n, t_\nu \in \mathcal{J}_{\mathbf{h}}, t_n \geq t_\nu.$$

We zeggen dat de RK methode B-*stabiel* is als de methode deze samentrekkings eigenschap heeft voor ieder probleem (1) waarin $f$ 0-samentrekkend is, voor iedere stapgrootte rij $\mathbf{h}$ en ieder tweetal exacte Runge-Kutta oplossingen $w_{\mathbf{h}}$ en $v_{\mathbf{h}}$.

**6.4.13 Notaties.** Beschouw 6.1.2. We schrijven

$$M := \operatorname{diag}(\vec{b})A + A^T \operatorname{diag}(\vec{b}) - \vec{b}\vec{b}^T. \tag{59}$$

**6.4.14 Stelling.** *Beschouw* 6.1.6. *Beschouw de volgende vier uitspraken.*
(a) *De methode is* A-*stabiel:* $|\lambda(\zeta)| < 1$ *voor alle* $\zeta \in \mathbf{C}_-$.
(b) $|\Lambda(\vec{\zeta})| < 1$ *voor alle* $\vec{\zeta} \in (\mathbf{C}_-)^k$.
(c) *De methode is* B-*stabiel.*
(d) $M$ *is positief definiet en* $b_j \geq 0$ *voor iedere* $j = 1, \ldots, k$.
*Er geldt* (a) $\Leftarrow$ (b) $\Longleftrightarrow$ (c) $\Longleftrightarrow$ (d).
*Bewijs.* Het is duidelijk dat (a) $\Leftarrow$ (b).
In het bewijs verder gebruiken we herhaaldelijk het volgende.
Als $B \in \mathbf{M}_{d\times k}$ en $C \in \mathbf{M}_{k\times d}$ dan $\operatorname{spoor}(BC) = \operatorname{spoor}(CB)$.
Als $B \in \mathbf{M}_k$ dan $\operatorname{spoor}(B)^- = \operatorname{spoor}(B^*)$ en dus $2\mathrm{Re}(\operatorname{spoor}(B)) = \operatorname{spoor}(B + B^*)$.
Stel nu dat $\tilde{w} = w + hF(t, W; h)\vec{b}$ met $W = w\vec{\mathbf{1}}^* + hF(t, W; h)A^*$
$\tilde{v} = v + hF(t, V; h)\vec{b}$ met $V = v\vec{\mathbf{1}}^* + hF(t, V; h)A^*$.
Met $\tilde{x} := \tilde{w} - \tilde{v}$, $x := w - v$, $X := W - V$ en $Y := F(t, W; h) - F(t, V; h)$ volgt

$$\tilde{x} = x + hY\vec{b} \quad \text{en} \quad X = x\vec{\mathbf{1}}^* + hYA^*. \tag{60}$$

De eerste relatie levert

$$\|\tilde{x}\|_2^2 = \|x\|_2^2 + 2h\mathrm{Re}(x^*Y\vec{b}) + h^2\vec{b}^*Y^*Y\vec{b}. \tag{61}$$

Voordat we de verschillende equivalenties bewijzen schrijven we eerst deze laatste gelijkheid om. Uit de tweede relatie in (60) volgt, door te adjungeren en te vermenigvuldigen met $Y \operatorname{diag}(\vec{b})$, dat

$$\vec{\mathbf{1}}\,x^*Y \operatorname{diag}(\vec{b}) = X^*Y \operatorname{diag}(\vec{b}) - hAY^*Y \operatorname{diag}(\vec{b}).$$

Merk op dat $\operatorname{spoor}(\vec{\mathbf{1}}\,x^*Y \operatorname{diag}(\vec{b})) = \operatorname{spoor}(x^*Y \operatorname{diag}(\vec{b})\vec{\mathbf{1}}) = x^*Y\vec{b}$. Vullen we de uitdrukking voor $x^*Y\vec{b}$ in in (61) dan zien we dat

$$\begin{aligned}
&\|\tilde{x}\|_2^2 - \|x\|_2^2 - 2h\mathrm{Re}(\operatorname{spoor}(X^*Y \operatorname{diag}(\vec{b})))\\
&= h^2\vec{b}^*Y^*Y\vec{b} - 2h^2\mathrm{Re}(\operatorname{spoor}(AY^*Y \operatorname{diag}(\vec{b})))\\
&= h^2[\operatorname{spoor}(\vec{b}^*Y^*Y\vec{b}) - 2\mathrm{Re}(\operatorname{spoor}(Y \operatorname{diag}(\vec{b})AY^*))]\\
&= h^2[\operatorname{spoor}(Y\vec{b}\vec{b}^*Y^*) - \operatorname{spoor}(Y[\operatorname{diag}(\vec{b})A + A^* \operatorname{diag}(\vec{b})]Y^*)]\\
&= -h^2\operatorname{spoor}(YMY^*).
\end{aligned}$$

Blijkbaar

$$\|\tilde{x}\|_2^2 - \|x\|_2^2 = 2h\mathrm{Re}(\operatorname{spoor}(X^*Y \operatorname{diag}(\vec{b}))) - h^2\operatorname{spoor}(YMY^*) \tag{62}$$

(d) $\Rightarrow$ (c). Als $f$ 0-samentrekkend is, is

$$(Xe_j, Ye_j) = (We_j - Ve_j, f(t + hc_j, We_j) - f(t + hc_j, Ve_j)) \leq 0.$$

Als ook nog $b_j \geq 0$ dan is $\operatorname{spoor}(X^*Y \operatorname{diag}(\vec{b})) = \sum_j b_j(Ye_j, Xe_j) \leq 0$. Is $M$ bovendien positief definiet, dan volgt uit (62) dat $\|\tilde{x}\|_2^2 \leq \|x\|_2^2$. Dit bewijst (c); neem maar $\tilde{w} = w_{n+1}$, $\tilde{v} = v_{n+1}$, $w = w_n$, $t = t_n$, etc..
In 6.4.10 hebben we gezien dat (c) $\Rightarrow$ (b).

Beschouw, om (b) $\Rightarrow$ (d) te bewijzen, de situatie in 6.4.10. We knopen weer aan bij de overwegingen in het eerste deel van het bewijs. Merk op dat nu $\tilde{x}, x \in \mathbf{C}$.
Voor de $Y$ geldt nu $Y = X \operatorname{diag}(\vec{\eta})$. Schrijven we $\vec{u} := (I - hA \operatorname{diag}(\vec{\eta})^*)^{-1}\vec{\mathbf{1}}$ en $\vec{z} := \operatorname{diag}(\vec{\eta})^*\vec{u}$ dan zien we dat $X = x\vec{u}^*$ (gebruik (60)) en dus $Y = x\vec{z}^*$. Omdat $\tilde{x} = \Lambda(h\vec{\eta})x$ volgt uit (62) dat

$$|\Lambda(h\vec{\eta})|^2 = 1 + 2h\mathrm{Re}(\operatorname{diag}(\vec{b})\vec{u}, \vec{z}) - h^2(M\vec{z}, \vec{z}). \tag{63}$$

Stel dat $|\Lambda(h\vec{\eta})|^2 < 1$ voor iedere $h > 0$ en iedere $\vec{\eta} \in (\mathbf{C}_-)^k$.
Merk op dat $\vec{u} = [I - hA \operatorname{diag}(\vec{\eta})^*]^{-1}\vec{\mathbf{1}} = \vec{\mathbf{1}} + \mathcal{O}(h)$ als $h \to 0$ en dus $\vec{z} = \vec{\eta} + \mathcal{O}(h)$ als $h \to 0$. Dus

$$|\Lambda(h\vec{\eta})|^2 = 1 + 2h\mathrm{Re}(\operatorname{diag}(\vec{b})\vec{u}, \vec{z}) + \mathcal{O}(h^2) = 1 + 2h\textstyle\sum_j b_j \mathrm{Re}(\eta_j) + \mathcal{O}(h^2) \quad (h \to 0).$$

Blijkbaar $\sum_j b_j \mathrm{Re}(\eta_j) \leq 0$. Dit kan alleen als $b_j \geq 0$ voor iedere $j = 1, \ldots, k$.
Neem nu een $\vec{\zeta} \in \mathbf{R}^k$ en beschouw $\vec{\eta} := i\vec{\zeta}$. Een kontinuiteits argument leert ons dat in dit geval $1 \geq |\Lambda(h\vec{\eta})|^2 = 1 - h^2(M\vec{z}, \vec{z})$. Omdat $\vec{z} = \vec{\eta} + \mathcal{O}(h)$ $(h \to 0)$, zien we ten slotte dat $(M\vec{\eta}, \vec{\eta}) = (M\vec{\zeta}, \vec{\zeta}) \geq 0$. Omdat $M = M^*$ volgt hieruit dat $M$ positief definiet is. $\square$

**6.4.15 Opmerking.** In de literatuur noemt men de eigenschap in (b) ook wel AN-*stabiliteit* en de eigenschap in (c) *algebraïsche stabiliteit*. B-stabiliteit, AN-stabiliteit en algebraïsche stabiliteit zijn dus equivalente eigenschappen.

**6.4.16 Opgave.** Beschouw voor $\alpha \in [0, 1]$ de DIRK methode met genererende matrix

$$\left[\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1-\alpha & 1-2\alpha & \alpha \\ \hline 0 & 0.5 & 0.5 \end{array}\right].$$

a. Bewijs dat de methode algebraïsch stabiel dan slechts dan als $\alpha \geq 0.25$.
b. Toon aan dat de methode orde 3 heeft als $\alpha = \frac{1}{6}(3 + \sqrt{3})$.

**6.4.17 Runge-Kutta versus multistep.** RK methoden zijn een-staps methoden en hebben dus geen aparte startprocedure nodig. Bovendien zijn ze hierdoor uitstekend te gebruiken in een variabele stapgrootte procedure. Een expliciete $k$-traps RK methode vergt per stap echter een $k$-tal funktie evaluaties. In een expliciete $k$-staps multistep hoeft in iedere iteratie stap slechts een funktie geëvalueerd te worden. In een impliciete $k$-staps multistep methode moet een (niet-lineaire) $d$-dimensionale vergelijking worden opgelost terwijl in een impliciete $k$-traps RK methode een $kd$-dimensionale vergelijking opgelost moet worden.

## 6.5 ◦ Kollokatie methoden

In de numerieke wiskunde konstrueert men nogal eens benaderingsprocedures door te "co-localiseren". Voor gewone differentiaalvergelijkingen betekent dit dat men een polynoom zoekt van graad $k$ waarvan de afgeleide in een gegeven $k$-tal punten samenvalt met het vektorveld van de differentiaalvergelijking (zie 6.5.1). De resulterende methode blijkt equivalent te zijn met een IRK methode (zie 6.5.2 en 6.5.3).

**6.5.1 Kollokatie.** Een *kollokatie methode met k kollokatie punten* wordt gegeven door een rij steunpunten of *kollokatie punten* $c_1 < c_2 < \ldots < c_k$ (gewoonlijk in $[0, 1]$).

De *lokale kollokatie oplossing* $\tilde{u}$ van (51) *door* $(\tilde{t}, x) \in \Omega$ is nu, voor iedere $h > 0$, gegeven door $\tilde{u}(\tilde{t} + h) = p(h)$, waarbij iedere koördinaat van $p$ een $k$-de graads polynoom is zodat geldt

$$\begin{aligned} &p(\tilde{t}) = x \\ &p'(\tilde{t} + c_j h) = f(\tilde{t} + c_j h, p(\tilde{t} + c_j h)) \quad \text{voor } j = 1, \ldots, k. \end{aligned} \tag{64}$$

**6.5.2 Stelling** *De kollokatie methode in* 6.5.1 *is equivalent met een impliciete $k$-traps* RK *methode als in* 6.1.2.

*Bewijs.* Voor $j = 1, \ldots, k$ is $l_j$ een basis funktie voor $k$-de graads Langrange interpolatie:

$$l_j(\zeta) := \prod_{i \neq j} \frac{\zeta - c_i}{c_j - c_i} \qquad (\zeta \in \mathbf{C}).$$

Met $\quad y_j := p(\tilde{t} + c_j h) \quad$ is

$$p'(\tilde{t} + \zeta h) = \sum_{j=1}^{k} y_j l_j(\zeta) \qquad (\zeta \in \mathbf{C}).$$

Omdat

$$p(\tilde{t} + h) = x + h \int_0^1 p'(\tilde{t} + sh)\, ds, \;\; p(\tilde{t} + c_j h) = x + h \int_0^{c_j} p'(\tilde{t} + sh)\, ds \;\; (j = 1, \ldots, k)$$

volgt de bewering verder eenvoudig door $b_j$ en $\alpha_{ij}$ als volgt te kiezen (zie ook 6.1.1)

$$b_j = \int_0^1 l_j(s)\, ds, \qquad \alpha_{ij} = \int_0^{c_j} l_j(s)\, ds \qquad (i, j = 1, \ldots, k). \quad \Box \tag{65}$$

De volgende stelling volgt eenvoudig uit de vorige stelling ondermeer door de relaties in (65) te gebruiken; we laten de details aan de geïnteresseerde lezer over.

**6.5.3 Stelling.** *Beschouw een $k$-traps* RK *methode als in* 6.1.2 *die konsistentie orde minstens $l$ heeft en waarbij $c_i \neq c_j$ als $i \neq j$. Deze* RK *methode is een kollokatie methode dan en slechts dan als iedere interne kwadratuur formule exact is voor alle polynomen van graad $\leq k-1$.* $\Box$

**6.5.4 Opgave.** Bewijs stelling 6.5.3.

De volgende stelling is van belang voor het bepalen van de konsistentie orde van een kollokatie metode; we bewijzen de stelling niet.

**6.5.5 Stelling.** *Schrijf* $q(\zeta) = \prod_{j=1}^{k}(\zeta - c_j) \quad (\zeta \in \mathbf{C})$. *Zij $m \in \mathbf{N}$.*
*De kollokatie methode in* 6.5.1 *heeft konsistentie orde $k + m$ als $q$ loodrecht staat op $\mathcal{P}_{m-1}$:*

$$\int_0^1 q(s) s^j\, ds = 0 \qquad \textit{voor iedere} \quad j = 0, \ldots, m-1. \quad \Box$$

# 7 Aantekeningen

## 7.1 Historische opmerkingen.

De oudste methoden stammen van Euler (1768).

Adams ontwierp in de tweede helft van de vorige eeuw (1855) multistep methoden die nauwkeuriger waren dan die van Euler. Hij berekende hiermee voor Bashford de theoretische vorm van een druppel. De BDF formules stammen uit 1952. De formules voor die tijd (van o.a. Milnes) berusten op kwadratuur. Serieuse theorie met nauwkeurige definities en konvergentie stellingen is pas door Dahlquist in 1956 (algemene konvergentie stellingen) en 1959 (eerste barrière stelling) ontwikkeld. Stabiliteits en konvergentie resultaten voor multistep methoden met variabele stapgrootte zijn door onder andere Grigorieff verkregen (1983).
In 1963 introduceerde Dahlquist het begrip A-stabiliteit en bewees de tweede barrière stelling. Zijn stelling waarin de equivalentie van G-stabilitiet en A-stabiliteit bewezen wordt stamt uit 1978. Andere belangrijke bijdrage in de multistep theorie voor stijve differentiaalvergelijkingen zijn onder meer van Widlund (1967, A($\alpha$)-stabiliteit), Cryer (1973) en Grigorieff (1983, existentie stellingen).
Een veel gebruikt programma pakket waarin variabele stapgrootte en variabele orde geimplimenteerd is, is ontwikkeld door Gear (1969-...).

Runge introduceerde in 1895 de expliciete midpuntregel door de midpuntregel en Euler methode te kombineren. Kutta generaliseerde dit idee in 1901 en formuleerde het algemene schema voor de Runge-Kutta methoden. De methoden waarmee tegenswoordig de konsistentie orde van een RK methode wordt vast gesteld en waarmee methoden gekonstrueerd worden zijn voornamelijk van Butcher (1963). Barrière stellingen werden door Butcher bewezen (1964, 1965, 1985). Gragg bewees in 1964 de stellingen over de asymptotische ontwikkeling van de globale fout. Merson stelde in 1957 voor om met bevattende RK methoden fouten te schatten. Een aantal populaire schema's van dit type zijn ontworpen door Fehlberg (1964-1970). Belangrijke bijdragen in de RK theorie voor stijve differentiaalvergelijkingen werden geleverd door Butcher (1975, B-stabiliteit) en Burrage en Butcher (1979, equivalente van diverse sterke stabiliteits begrippen).

## 7.2 Kanttekeningen bij het literatuur lijstje.

In [4] laat Stetter geen enkel theoretisch detail onbesproken. Hij besteed in dit boek echter weinig aandacht aan praktische zaken. Naast een behandeling van de stabiliteits theorie voor gewone differentiaalvergelijkingen geeft hij de volledige theorie van de Runge-Kutta methoden en de multistep methoden zoals die tot 1972 ontwikkeld was. Omdat een aantal belangrijke ontwikkeling met name voor stijve differentiaalvergelijkingen van latere datum zijn is dit boek voor dit onderwerp niet zo'n interessante bron. Dat is het wel voor de situatie waarin $h \to 0$. Omdat de auteur geen ruimte tot misverstand wil laten is de notatie nogal zwaar en is het boek niet, als men er al niet mee vertrouwd is, geschikt als naslag werk.

[1] is een luchtig werk waarin de Runge-Kutta methoden en multistep methoden in theorie en praktijk voor stijve en niet stijve differentiaalvergelijkingen besproken worden. Het boek is uitstekend leesbaar maar roept hier en daar, omdat het te oppervlakkig is, onnodig vragen op.

Het boek van Hairier, Nørsett en Wanner [2] en het vervolg van Hairier en Wanner [3] zijn bijzonder prettig leesbaar terwijl ze allesbehalve oppervlakkig zijn. De theorie wordt met de nodige diepgang behandeld (niet zo gedetailleerd als in [4]) en gelardeerd met historische aantekeningen en anekdotes. Verder worden er een groot aantal praktische beschouwingen gehouden en praktijk voorbeelden en programma's gegeven. De stabiliteit van het kontinue probleem met moderne ontwikkelingen (bifurkaties en chaos), Runge-Kutta methoden en multistep methoden komen uitgebreid aan de orde. Voor niet stijve problemen geeft deel I, de "state of the art" anno 1987 goed weer. In deel II worden methodes voor stijve problemen behandeld.

## Referenties


[1] S. O. Fatunla. *Numerical Methods for Initial Value Problems in Ordinary Differential Equations. Computer Science and Scientific Computing*, Academic Press, Inc., Boston, etc., 1988.

[2] E. Hairier, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I.* in *Springer Series in Computational Mathematics*, Springer-Verlag, Berlin, etc., 1987.

[3] E. Hairier and G. Wanner. *Solving Ordinary Differential Equations II.* in *Springer Series in Computational Mathematics*, Springer-Verlag, Berlin, etc., 1991.

[4] H.J. Stetter. *Analysis of Discretization Methods for Ordinary Differential Equations.* Volume 23 of *Springer Tracts in Natural Philosophy*, Springer-Verlag, Berlin, etc., 1973.

# Index

# Notaties

# Gewone differentiaalvergelijkingen: opgaven

## Euler methoden

Zij $u^*$ de exacte oplossing van de differentiaalvergelijking

$$\begin{cases} u'(t) = f(t, u(t)) & \text{voor } t \in \mathcal{J} \equiv [t_0, t_0 + T], \\ u(t_0) = u_0, \end{cases} \tag{1}$$

waarbij $f \in C^{(1)}(\Omega, \mathbf{R}^d)$ Lipschitz kontinu is in de tweede variabele met Lipschitz konstante $L$.

Soms beperken we ons tot het 1-dimensionale lineaire probleem

$$\begin{cases} u'(t) = \eta u(t) + g(t) & \text{voor } t \in \mathcal{J}, \\ u(t_0) = u_0, \end{cases} \tag{2}$$

waarbij $\eta \in \mathbb{C}$ en $g \in C^\infty(\mathcal{J})$. Het is al instruktief te weten hoe een numerieke oplosmethode zich gedraagt voor dit probleem. Ter gedachte bepaling nemen we zo nu en dan $t_0 = 0$, $\eta = -1/\epsilon$, met $0 < \epsilon \ll 1$, $g(t) = -\eta \cos t - \sin t$ en $u_0 = 1$, d.w.z. we beschouwen het probleem

$$\begin{cases} \epsilon\, u'(t) = -u(t) + (\cos t - \epsilon \sin t) & \text{voor } t \in [0, T], \\ u(0) = 1, \end{cases} \tag{3}$$

met exacte oplossing $u^*(t) = \cos t$.

Voor $h > 0$ schrijven we $t_n \equiv t_0 + nh$, $u_n^* \equiv u^*(t_n)$ en $f_n^* \equiv f(t_n, u_n^*)$.
Verder is $u_h$ de numerieke oplossing op $\{t_n\}$ en schrijven we $u_n \equiv u_h(t_n)$, $f_n \equiv f(t_n, u_n)$.

### Opgave 1 [Stabiliteit & konvergentie Euler forward]

**a** [*Konsistentie*] Bewijs dat, voor het 1-d. geval ($d = 1$) geldt

$$u_{n+1}^* = u_n^* + h f_n^* + h\delta_n \quad \text{met} \quad \delta_n \equiv \frac{1}{2} h (u^*)''(\xi_n) \quad \text{zekere} \quad \xi_n \in [t_n, t_{n+1}].$$

**b** [*Majorerende rekursie*] We lossen probleem (1) numeriek op middels de Euler forward methode. Dan geldt, voor zekere lokale evaluatiefouten $\epsilon_n$,

$$u_{n+1} = u_n + h f_n + \epsilon_n.$$

Bewijs dat voor de globale fout $e_n \equiv u_n^* - u_n$ geldt

$$\|e_n\| \leq \widetilde{e}_n$$

waarbij, met $E \equiv \max_n \|\epsilon_n\|$ en $D \equiv h\,\frac{1}{2} \max_\xi \|(u^*)''(\xi)\|$, de majorant $\widetilde{e}_n$ gedefinieerd is door

$$\widetilde{e}_{n+1} = \widetilde{e}_n + hL\widetilde{e}_n + E + hD, \quad \widetilde{e}_0 = \|e_0\|.$$

**c** [*Stabiliteit*] We nemen aan dat $\widetilde{e}_0 = 0$ (is dit redelijk?). Toon aan dat

$$\widetilde{e}_n \leq \sum_{j=0}^{n-1} (1 + hL)^j (E + hD) \leq e^{(t_n - t_0)L} (nE + nhD) \leq n\, e^{TL} E + T e^{TL} D \quad \text{alle n.}$$

**d** [*Konvergentie*] Als $\epsilon_n = 0$ voor alle $n$ dan geldt

$$\sup_n \|u^*(t_n) - u_n\| \leq h\, \frac{1}{2} T e^{TL} \max_\xi \|(u^*)''(\xi)\| \quad \text{voor iedere} \quad h > 0 \tag{4}$$

en dus

$$u_h = u^* + \mathcal{O}(h) \quad \text{uniform} \quad (h \to 0). \tag{5}$$

Bewijs dit.

**e**. Schat, met behulp van (4), met welke $h$ Euler forward een oplossing produceert van (3) met een fout $\leq 10^{-3}$. Neem daarbij voor $\epsilon$ en $T$ kombinaties van $\epsilon = 1$, $\epsilon = 0.01$, $T = 1$ en $T = 10$. Zijn de problemen voor de verschillende kombinaties goed gesteld?

**Opgave 2 [De fout is in essentie glad]**

In deze opgave nemen we aan dat de evaluatiefouten verwaarloosbaar zijn ($\epsilon_n = 0$) en dat we starten exact ($e_0 = 0$). Verder veronderstellen we dat $f \in C^{(3)}(\Omega, \mathbf{R}^d)$.

Zij $w$ de oplossing van

$$\begin{cases} w'(t) = J(t)w(t) + \frac{1}{2}(u^*)''(t) & \text{voor } t \in \mathcal{J}, \\ w(t_0) = 0, \end{cases}$$

waarbij $J(t)$ de Jacobi matrix is van $f$ in $(t, u^*(t))$:

$$J(t) \equiv \left[ \frac{\partial f_j}{\partial x_i}(t, u^*(t)) \right].$$

**a**. Ga na dat, met $J_n \equiv J(t_n)$ geldt

$$hw_{n+1} = hw_n + h\left[J_n\, h\, w_n + \frac{1}{2}h\,(u^*)''(t_n)\right] + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0).$$

**b**. Toon aan dat voor de globale fouten $e_n \equiv u_n^* - u_n$ in de Euler-forward-oplossing van (1) geldt

$$e_{n+1} = e_n + hJ_ne_n + \frac{1}{2}h^2\,(u^*)''(t_n) + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0).$$

**c** [*De fout is in essentie glad*] Toon nu aan dat

$$u_h = u^* - hw + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0). \tag{6}$$

(Hint: beschouw $v_n \equiv e_n - hw_n$ en ga te werk als in Opgave˜1.) Zie ook in dat

$$u_h = u^* - hw + h^2 w_2 + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0), \tag{7}$$

voor een zekere gladde funktie $w_2$.

**d**. Beschouw $\widetilde{w} \equiv h\,w$. Dan geldt

$$\begin{cases} \widetilde{w}'(t) = J(t)\widetilde{w}(t) + \widetilde{\delta}(t) & \text{met} \quad \widetilde{\delta} \equiv h\,\frac{1}{2}\,(u^*)'', \\ \widetilde{w}(t_0) = 0, \end{cases}$$

en kunnen we, als de $\mathcal{O}(h^2)$-term in (6) verwaarloosbaar is, de globale fout schatten in termen van (i) het konditiegetal van de differentiaalvergelijking en (ii) een majorant van de lokale fouten. Geef zo'n schatting en konkludeer dat

$$\|u_n^* - u_n\| \lesssim \mathcal{C}(u^*) \int_0^{t_n} \|\widetilde{\delta}(\tau)\|\,\mathrm{d}\tau \lesssim \mathcal{C}(u^*) \sum_{j=0}^{n} h\,\|\delta_j\| : \tag{8}$$

op een faktor na (nl. $\approx$ konditie getal diff.verg.) kan de norm van de globale fout gemajoreerd worden door de som van de norm van de lokale fouten.

**Opgave 3** Laat, voor stapgrootte $h > 0$, $u_h$ de Euler-forward-oplossing zijn van (1). Definieer de funktie $v$ op $\mathcal{J}$ door stuksgewijs lineaire interpolatie van $u_h$:

$$v(t) \equiv u_n + sf(t_n, u_n) \quad \text{voor} \quad t = t_n + s \quad \text{met } s \text{ zodat} \quad 0 \leq s < h.$$

Dan geldt voor zekere funktie $\delta$ (waarom?)

$$\begin{cases} v'(t) = f(t, v(t)) + \delta(t) & \text{voor} \quad t \in \mathcal{J}, \\ v(t_0) = u_0. \end{cases}$$

**a**. Stel dat $f$ Lipschitz kontinu is in twee variabelen: voor zekere $L > 0$ geldt

$$\|f(t,u) - f(s,v)\| \leq L \max(|t-s|, \|u-v\|) \quad \text{alle} \quad (t,u),\ (s,v);$$

en stel dat $f$ begrensd is: voor zekere $M > 0$ geldt

$$\|f(t,u)\| \leq M \quad \text{alle} \quad (t,u).$$

Bewijs dat dan

$$\delta = \mathcal{O}(h) \quad \text{uniform} \quad (h \to 0).$$

**b**. Wat betekent dit voor sterk stabiele differentiaalvergelijkingen (druk de Euler-forward-fout uit in termen van het sterke konditie getal)? Wat betekent dit voor de schattingen in Opgave˜1e?

**Opgave 4 [Foutschatten en stapgrootte besturing]**
We noteren de Euler-backward-oplossing en $u_h^{\mathrm{b}}$ van (1) en de Euler-forward-oplossing met $u_h^{\mathrm{f}}$ ($= u_h$ in Opgave˜1).

**a** [<u>*Globale-fout schatting I*</u>] Leg uit hoe je met behulp van $u_h^{\mathrm{f}}$ en $u_H^{\mathrm{f}}$, met bv $H = 2h$, de fout kunt schatten in een vast tijdstip $\tau \in \mathcal{J}$ (gebruik (6)). Je moet dan aannamen maken over de invloed van de $\mathcal{O}(h^2)$-term. Hoe kan je enige zekerheid krijgen dat die aannamen redelijk zijn in een konkrete situatie?

**b** [<u>*Globale-fout schatting II*</u>] Ga na dat, met $w$ als in Opgave˜2, geldt

$$u_h^{\mathrm{b}} = u^* + hw + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0) \tag{9}$$

en dus

$$u^* - u_h^{\mathrm{f}} = \frac{1}{2}\left(u_h^{\mathrm{b}} - u_h^{\mathrm{f}}\right) + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0).$$

Hoe kan je dit resultaat gebruiken om de fout te schatten in een vast tijdstip $\tau$?

**c** [<u>*Lokale-fout schatting*</u>] Bereken voor 'n $n$ de waarde $\widetilde{u}_{n+1}$ volgens

$$\widetilde{u}_{n+1} = u_n^{\mathrm{f}} + hf(t_{n+1}, u_{n+1}^{\mathrm{f}})$$

(een Euler-backward-stap met Euler-forward-resultaten). Als $f \in C^{(2)}(\Omega)$ dan geldt

$$u_{n+1}^{\mathrm{f}} - \widetilde{u}_{n+1} = -\frac{1}{2}\,h^2\,(u^*)''(t_n) + \mathcal{O}(h^3) = -h\,\delta_n^{\mathrm{f}} + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0). \tag{10}$$

Bewijs dit (Hint: Gebruik (7)). Kan je dit resultaat ook bewijzen door alleen gebruik te maken van (5) of van (6)?

**d** [<u>*Stapgrootte besturing*</u>] We gebruiken de schatting in c als volgt.
Stel het interval $\mathcal{J}$ is verdeeld in $K$ subintervalletjes $\mathcal{J}_i = [\tau_{i-1}, \tau_i]$, $(i = 1, \ldots, K)$ waarbij

$$t_0 = \tau_0 < \tau_1 < \ldots < \tau_K = t_0 + T$$

en de subintervalletjes zo zijn dat de funktie $(u^*)''$ per subinterval niet veel varieert. Per subinterval gebruiken we een stapgrootte $h_i$ met $h_i$ zodat $(\tau_i - \tau_{i-1})/h_i \in \mathbb{N}$ en zo gekozen dat voor een zekere $\bar{\epsilon}$ geldt

$$h_i \|\delta_n^{\mathrm{f}}\| \approx h_i\,\bar{\epsilon} \quad \text{in} \quad t_n = \tau_{i-1} \quad \text{en iedere} \quad i = 1, \ldots, K.$$

Geef een schatting voor globale fout in termen van $\bar{\epsilon}$ en het konditie getal van de diff.vergl.. Stel dat $h_{i-1}$ zo is dat $\|u_n^{\mathrm{f}} - \widetilde{u}_n\| = q\,h_{i-1}\bar{\epsilon}$ in $t_n = \tau_{i-1}$, voor zekere $q > 0$. Hoe kies je dan $h_i$ (in termen van $h_{i-1}$ en $q$) om te bewerkstelligen dat $h_i\|\delta_n^{\mathrm{f}}\| \approx h_i\,\bar{\epsilon}$?

In de praktijk kiest men gewoonlijk $\tau_i = t_i$ voor alle $i$, d.w.z. in iedere stap wordt de lokale fout geschat en wordt de stapgrootte eventueel aangepast. Waarom is het ongewenst de stapgrootte te vaak te veranderen? Om te voorkomen dat dit toch gebeurt, verandert men daarom alleen als $\|u_n^{\mathrm{f}} - \widetilde{u}_n\| \notin [\frac{1}{2}\,h_{i-1}\bar{\epsilon}, 2\,h_{i-1}\bar{\epsilon}]$.

**e**. Bespreek de voor– en nadelen van de schattingsmethoden in a, b en c.

### Opgave 5 [Stabiliteit bij grotere stapgrootte]

In deze opgave onderzoeken we wanneer (8) een betrouwbare schatting geeft.

**a**. De schattingen voor $h$ in Opgave˜1d waren, voor grotere $T$ en kleinere $\epsilon$, nogal teleurstellend. Schattingen met (8) pakken gunstiger uit. Speelt de grote konstante $T e^{TL}$ in de afleiding van (8) geen rol? Wat betekent dit voor de betrouwbaarheid van schatting (8)?

We beschouwen probleem (2) met $\eta < 0$, verwaarlozen evaluatiefouten en lossen in eerste instantie op met Euler forward.

**b**. Bepaal, voor probleem (3) met $\epsilon = 10^{-6}$, $h \approx$ zo groot mogelijk zodat $|\delta_n| \leq 0.001$ voor alle $n$.

**c**. Laat zien dat $\mathcal{C}(u^*) = 1$.

Is, voor de situatie in b. (8) betrouwbaar? (vergelijk $\|u_n^* - u_n\|$ met $\sum h\|\delta_j\|$ voor bv. $h = 10^{-3}$)

**c**. Ga na dat

$$e_{n+1} = e_n + h\,\eta\,e_n + h\,\delta_n = (1 + h\,\eta)e_n + h\,\delta_n:$$

de fouten worden telkens voortgeplant met een factor $(1 + h\,\eta)$. Konkludeer dat

$$e_n = (1 + h\eta)^n e_0 + \sum_{j=0}^{n-1} (1 + h\,\eta)^j\,h\,\delta_{n-1-j}.$$

Dit impliceert dat (vergelijk met (8))

$$|e_n| \leq |e_0| + \sum_{j<n} h\,|\delta_j| \qquad \left(\leq |e_0| + T \max_{j<n} |\delta_j|\right)$$

als

$$|1 + h\eta| \leq 1.$$

Ga dit na. Wat betekent dit voor $h$ voor probleem (3) met $\epsilon = 10^{-6}$? Vergelijk de gevonden schatting met die in b.

**d**. We onderzoeken of we de voorwaarde $|1 + h\,\eta| \leq 1$ kunnen afzwakken. We bekijken weer probleem (3) met $\epsilon = 10^{-6}$.

Beschouw $h \equiv 1.001 \times h_0$ met $h_0$ zodat $|1 + h_0\eta| = 1$. Beschrijf de bijdrage van de startfout $e_0$ in de totale fout $e_n$ in het punt $t_n = t_0 + 1$. Mogen we van deze "grotere" $h$ goede resultaten verwachten?

Kies nu $h$ zodat $|1 + h\,\eta| = e^h$. Schat de globale fout af in termen van de som van de lokale fouten. Een kleine foutgroei lijkt wel acceptabel. Wat betekent dit echter voor $h$ in vergelijking met $h_0$?

Als we, voor zekere $\kappa$ (met $|\kappa| = \mathcal{O}(1)$), de differentiaalvergelijking

$$\epsilon u'(t) = -u(t) + (1 + \kappa\,\epsilon)e^{\kappa t} \quad \text{voor} \quad t \in [0,T] \quad \text{en} \quad u(0) = 1$$

(met exacte oplossing $u^*(t) = \exp(\kappa t)$) willen oplossen met Euler forward lijkt het verstandig $h$ zo te kiezen dat $|1 + h\eta| \leq e^{\kappa h}$. Waarom?

**e.** Voor de globale fout $e_n^{\rm b}$ in Euler backward geldt

$$e_n^{\rm b} = \left(\frac{1}{1 - h\,\eta}\right)^n e_0 + \sum_{j=1}^{n} \left(\frac{1}{1 - h\,\eta}\right)^j h\,\delta_{n-j}^{\rm b}.$$

Ga dit na en konkludeer dat

$$|e_n| \leq |e_0| + \frac{1}{2}\frac{1}{|h\eta|}h^2 \max_{\xi \leq t_n} |(u^*)''(\xi)|.$$

Neem $e_0 = 0$. Hoe groot kies je $h$ nu, voor probleem (3) met $\epsilon = 10^{-6}$, als je een globale fout $\leq 0.001$ wilt hebben?

**f.** Bespreek de voor– en nadelen van Euler forward in vergelijking met de Euler backward.

**g.** Bespreek ook de voor– en nadelen in geval $\epsilon = -1$ in (3).

**Opgave 6** In Opgave˜5 hebben we gezien dat, in ieder geval voor 1-d lineaire problemen (2), Euler-forward-fouten begrensd voortplant als

$$h\eta \in \mathcal{S} \equiv \left\{\zeta \in \mathbb{C} \,\Big|\, |1 + \zeta| \leq 1\right\}.$$

$\mathcal{S}$ is het *stabiliteits gebied* van Euler forward: $\mathcal{S}$ is een cirkel in het complexe vlak met straal 1 en middelpunt $-1$. Beschijf het stabiliteitsgebied van Euler backward. In Opgave˜5 hebben we ook gezien dat de grens van dit stabiliteitsgebied (zeker voor $|h\eta| \not\ll 1$) "hard" is. De inzichten zijn ook van belang voor meer dimensionale – of niet-lineaire problemen en voor partiële differentiaalvergelijkingen.

Beschouw de 1-d. warmtevergelijking

$$\frac{\partial\, u}{\partial\, t} = \frac{\partial^2 u}{\partial\, x^2} + f(x,t) \qquad \text{voor } x \in [0,\pi] \text{ en } t \geq 0,$$

met begin- en randvoorwaarde

$$\left\{\begin{array}{l} u(x,0) = \varphi(x), \\ u(0,t) = u(\pi,t) = 0. \end{array}\right. \tag{11}$$

**a.** Ga na dat voor $f(x,t) = \sin\,x\,(1 + 2\sin\,t)$ en $\varphi(x) = 0$ de exakte oplossing $u^*$ gegeven wordt door

$$u^*(x,t) = (\sin\,x)\,(1 - \cos\,t\, + \sin\,t).$$

We diskretiseren eerst in de $x$-richting (*semi-diskretisatie*) en vervolgens in de $t$-richting. Voor de diskretisatie in de $x$-richting verdelen we $[0,\pi]$ in $N$ deelintervallen van gelijke lengte. De stapgrootte wordt dus $\Delta x \equiv \pi/(N+1)$ en voor (voorlopig vaste) $t \geq 0$ zoeken we een rijtje $(u_n)_{0\leq n \leq N+1}$ dat de waarden van $u^*$ geevalueerd in $0$, $\Delta x$, $2\Delta x$, …, $(N+1)\Delta x = \pi$ op tijdstip $t$ benadert. We gebruiken de volgende notatie.

$$u_j(t) \equiv u(j\Delta x, t), \quad u_j^*(t) \equiv u^*(j\Delta x, t) \qquad (0 \leq\, j \,\leq\, N+1),$$

$$\vec{u} \equiv (u_1, \ldots u_N)^T, \quad \vec{u^*} \equiv (u_1^*, \ldots u_N^*)^T, \quad \vec{f} \equiv (f_1, \ldots f_N)^T.$$

Voor de plaatsafgeleide gebruiken we de volgende relatie:

$$\frac{\partial^2 u}{\partial x^2}(j\Delta x, t) = \frac{u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)}{(\Delta x)^2} + \delta_j^x(t).$$

**b**. Toon aan dat voor de plaatsdiskretisatiefout $\delta_j^x$ geldt:

$$\delta_j^x(t) = \frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi, t) \quad \left(= \mathcal{O}((\Delta x)^2)\right).$$

Door nu tegelijkertijd in alle inwendige diskretisatiepunten de relatie

$$u_j' = \frac{u_{j+1} - 2u_j + u_{j-1}}{(\Delta x)^2}$$

op te leggen $\left(u_j' = \frac{\partial u}{\partial t}(j\Delta x, t) = \frac{\partial^2 u}{\partial x^2}(j\Delta x, t) = \frac{\partial^2 u_j}{\partial x^2}\right)$ ontstaat een gewone differentiaalvergelijking in $\vec{u}$ van de vorm

$$\begin{cases} \vec{u}' = \dfrac{1}{(\Delta x)^2} A\vec{u} + \vec{f}, \\ \vec{u}(0) = (\varphi_1, \ldots, \varphi_N)^T. \end{cases} \tag{12}$$

**c**. Bepaal $A$ en $\vec{f}$. Reken na dat $(\vec{v}_\ell)_{1\leq \ell \leq N}$ met

$$\vec{v}_\ell = (\sin \ell\Delta x, \sin 2\ell\Delta x, \ldots, \sin N\ell\Delta x)^T$$

eigenvectoren van $A$ zijn en bepaal de bijbehorende eigenwaarden $\eta_\ell$: $A\vec{v}_\ell = \eta_\ell \vec{v}_\ell$. Konkludeer dat $-4 \leq \eta_\ell \leq 0$ voor alle $\ell$ waarvoor $1 \leq \ell \leq N$.

**d**. We lossen (12) op met Euler forward met tijdsstapgrootte $\Delta t$. Om stabiliteitsredenen moeten we $\Delta t$ zo kiezen dat

$$\frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}. \tag{13}$$

Laat dit zien. Bepaal $\Delta t$ en $\Delta x$ zo dat de *lokale* diskretisatiefouten $\leq 0.01$. Hoe pakt dit uit voor de eis dat de *globale* fouten $\leq 0.01$ voor alle $t \in [0, 1]$? Beantwoord deze vragen ook voor Euler backward.

De restriktie in (13) heet de *CFL-konditie* (Courant-Friedrichs-Lewy); 0.5 is het CFL getal.

**Opgave 7** Beschouw, voor zekere $a \in \mathbb{R}$, het 1-dimensionale hyperbolisch probleem

$$\frac{\partial u}{\partial} = a\frac{\partial u}{\partial x} + f(x,t) \quad \text{voor } x \in [0, \pi] \text{ en } t \geq 0, \tag{14}$$

met begin- en randvoorwaarde

$$\begin{cases} u(x, 0) = \phi(x), \\ u(0, t) = u(\pi, t), \quad , \frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(\pi, t). \end{cases}$$

De randvoorwaarde zijn *periodiek* (ga na dat dan dat de oplossing $u^*$ in plaats "periodiek" is, $u^*(x,t) = u^*(x + \pi, t)$, waarbij $u^*$ de oplossing is van de differentiaal vergelijking voor alle $x \in \mathbb{R}$ met dezelfde randvoorwaarden, met periodieke $f$ en $\phi$).

We diskretiseren weer eerst in de $x$-richting met stapgrootte $\Delta x = \pi/(N + 1)$. We schrijven

$$\vec{u} \equiv (u_0, u_1, \ldots, u_N)^T, \quad \vec{u}^* \equiv (u_0^*, u_1^*, \ldots, u_N^*)^T, \quad \vec{f} \equiv (f_0, f_1, \ldots, f_N)^T$$

met $u_j(t)$, $u_j^*(t)$ en $f_j(t)$ als in Opgave˜6. Merk op dat hier $u_{N+1}(t) = u_0(t)$, etc..

Voor de plaatsafgeleide bekijken we de volgende twee relaties

$$\frac{\partial u}{\partial t}(j\Delta x, t) = \frac{u_{j+1}(t) - u_j(t)}{\Delta x} + \delta_j^u(t) \tag{15}$$

en

$$\frac{\partial u}{\partial t}(j\Delta x, t) = \frac{u_{j+1}(t) - u_{j-1}(t)}{\Delta x} + \delta_j^c(t). \tag{16}$$

Voor $j = 0$ en $j = N$ maken we gebruik van de periodiciteit, d.w.z. we werken met $u_N(t)$ in plaats van $u_{-1}(t)$, etc..

**a**. Toon aan dat voor de plaatsdiskretisatiefouten geldt

$$\delta_j^u = \mathcal{O}(\Delta x) \quad \text{en} \quad \delta_j^c = \mathcal{O}((\Delta x)^2).$$

Door in de plaats richting op de gesuggereerde manier te diskrestiseren onstaan gewone differentiaal vergelijkingen van de vorm

$$\begin{cases} \vec{u}\,' = \dfrac{a}{\Delta x} A_u \vec{u} + \vec{f}, \\ \vec{u}(0) = (\phi_0, \ldots, \phi_N)^T \end{cases} \tag{17}$$

en

$$\begin{cases} \vec{u}\,' = \dfrac{a}{\Delta x} A_c \vec{u} + \vec{f}, \\ \vec{u}(0) = (\phi_0, \ldots, \phi_N)^T. \end{cases} \tag{18}$$

**b**. Bepaal $A_u$ en $A_c$. Reken na dat $(\vec{v}_\xi)$ met

$$\vec{v}_\xi \equiv (1, e^{\xi i}, e^{2\xi i}, \ldots, e^{N\xi i})^T \qquad (\text{met } i = \sqrt{-1})$$

voor $\xi = \ell\Delta x$, $\ell = 0, 1, \ldots, N$, de eigenvectoren van $A_u$ en $A_c$ zijn. Bepaal de bijbehorende eigenwaarden en laat zien dat de eigenvectoren een orthogonaal stelsel vormen (met konditiegetal 1).

**c**. We lossen de stelsels gewone differentiaalvergelijkingen (17) en (18) op met Euler forward met tijdsstapgrootte $\Delta t$. Om stabiliteitsredenen kunnen we $\Delta t$ niet willekeurig groot kiezen. Laat zien dat we voor (17) voor het geval $a > 0$, voor een zekere $r > 0$, $\Delta t$ zo kunnen kiezen dat

$$a\frac{\Delta t}{\Delta x} \leq r \tag{19}$$

en bepaal $r$.

Bestaat er ook zo'n $r > 0$ voor (18)? Kan je voor (18) wel stabiliteit garanderen onder een restriktie van de vorm $\Delta t/(\Delta x)^2 \leq r$? Hoe veel wordt de startfout opgeblazen in $t = 1$ met $a = 1$, $\Delta t/(\Delta x)^2 = 1$ en $N = 50$?

**d**. Hoe diskretiseer je (14) met Euler forward met een eenzijdig schema voor het geval $a < 0$? Beschrijf ook je aanpak voor het geval $a$ van $x$ afhangt, bijvoorbeeld $a(x) = \cos x$.

Of je de plaatsdifferenties links– of rechtszijdig kiest zal hangen af van het teken van $a$: de zogenaamde "upwind" differentie leidt tot een stabiel schema ($a$ in de advektie term $a\frac{\partial u}{\partial x}$ vertelt in sommige toepassingen waar de wind vandaan komt).

**e**. Stel nu (17) onstaat door plaatsdikretisatie van (14) met randvoorwaarden (11). Hoe ziet $A_u$ er dan uit? Wat zijn de eigenwaarden?

Als we alleen naar de eigenwaarden kijken kunnen we met een $\Delta t$ toe waarvoor $\Delta t/\Delta x \leq r/2$, met $r$ als in c. Ga dat na. Toch blijkt in de praktijk de stabiliteitsrestriktie in c beter. Kun je inzien hoe dat komt?

In de praktijk wordt de stabiliteit van rekenschema's als boven vaak geanalyseerd door te doen alsof de vektoren $\vec{v}_\xi$ eigenvektoren zijn van het gediskretiseerde ruimtelijke deel, d.w.z. effekten van de randvoorwaarden worden genegeerd en koëfficiënten die van plaats of tijd afhangen worden"bevroren" (vervangen door konstanten met waarde gelijk aan'n koëfficiënt waarde). Zo'n analyse heet von Neumann analyse of local mode analysis (er wordt lokaal met Fourier modes gewerkt).

**Opgave 8 [Deferred correction]**

We beschouwen weer de situatie als in Opgave˜2.

**a**. Ga na dat

$$\delta_n = \frac{1}{2}h(u^*)''(t_n) + \mathcal{O}(h^2) = \frac{u^*_{n+1} - 2u^*_n + u^*_{n-1}}{2h} + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0). \tag{20}$$

**b**. Bewijs dat

$$\delta_n = \frac{u_{n+1} - 2u_n + u_{n-1}}{2h} + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0)$$

(Hint: gebruik (7)). Kan je dit resultaat ook bewijzen door alleen gebruik te maken van (5) of van (6)?

**c**. We lossen nu de differentiaalvergelijking als volgt op. Met stapgrootte $h > 0$ berekenen we de Euler forward oplossing $u_h$ voor alle $n$, $n \leq T/h$, en vervolgens berekenen we $v_h$ volgens

$$\begin{cases} v_{n+1} = v_n + hf(t_n, v_n) + \frac{1}{2}\left(u_{n+1} - 2u_n + u_{n-1}\right), \\ v_0 = u_0. \end{cases}$$

Bewijs dat (vergelijk met (6))

$$v_h = u^* + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0)$$

**d**. Het is verleidelijk (waarom?) om $v_h$ te berekenen volgens

$$\begin{cases} v_{n+1} = v_n + hf(t_n, v_n) + (v_{n+1} - 2v_n + v_{n-1})/2, \\ v_0 = u_0, \quad v_1 = v_0 + hf(t_n, v_0) + (u_2 - u_1 + u_0)/2. \end{cases} \tag{21}$$

We zullen (21) uitgebreid analyseren in dit college. Verlopig merken we op dat het gevaarlijk is dit soort ad-hoc aanpassingen uit te voeren. Immers, omdat

$$\delta_n = \frac{1}{2}h(u^*)''(t_{n+1}) + \mathcal{O}(h^2) = \frac{u_{n+2} - 2u_{n+1} + u_n}{2h} + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0)$$

(zie dit in), zou met evenveel recht $v_h$ berekend kunnen worden volgens

$$v_{n+1} = v_n + hf(t_n, v_n) + (v_{n+2} - 2v_{n+1} + v_n)/2. \tag{22}$$

Pas dit eens toe op de diff. vergl. $u' = 0$, $u(0) = 1$, neem daarbij $v_0 = 1 + \epsilon$, $v_1 = 1 + 3\epsilon$ en bereken $v_n$ voor $n$ waarvoor $t_n = 1$.

**Opgave 9 [Prediktor-korrektor techniek]**

In Opgave˜5 hebben we gezien dat Euler backward gunstigere stabiliteitsvoorwaarden heeft dan Euler forward. Euler backward is echter impliciet: $u_{n+1}$ moet, gegeven $u_n$, opgelost worden uit $u_{n+1} = u_n + hf(t_{n+1}, u_{n+1})$. Dit oplossen moet niet al te duur zijn omdat we anders weer het stabiliteitsvoordeel verspelen.

Bekijk de volgende strategie.

$$\begin{cases} \widetilde{u}_{n+1} = u_n + hf(t_n, u_n), \\ u_{n+1} = u_n + hf(t_{n+1}, \widetilde{u}_{n+1}) : \end{cases} \tag{23}$$

gebruik Euler forward als “voorzegger” (predictor) voor Euler backward (“corrector”).

**a**. Laat zien dat voor de lokale diskretisatiefout $\delta_n$ geldt

$$\delta_n = \delta_n^{\mathrm{b}} + \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0) \tag{24}$$

(hoe definieer je hier lokale diskretisatiefout?). De lokale diskretisatiefout $\delta_n^{\rm b}$ van Euler backward (en ook van Euler forward) hangt alleen af van de exacte oplossing $u^*$: $f$ (d.w.z. de differentiaalvergelijking) speelt expliciet geen enkele rol. Hoe zit dat voor methode (23)?

**b**. Methode (23) is een mengsel van Euler forward en Euler backward. We mogen hopen dat de methode "stabieler" is dan Euler forward, maar of hij even stabiel is al Euler backward is de vraag.

Laat zien dat voor probleem (2) de fouten worden voortgeplant met een factor

$$1 + h\,\eta + (h\,\eta)^2$$

(zie c van Opgave˜5). Hoe groot moet $h$ gekozen worden voor het geval $\eta = -1/10^{-6}$ (zie Opgave˜5)? Hoe groot is de lokale diskretisatiefout in dit geval? Mogen we de $\mathcal{O}(h^2)$-term in (24) verwaarlozen? Zijn er situaties waarin deze gemengde aanpak toch nog stabiliteits voordelen biedt boven Euler forward? (Hint: bekijk een probleem met $\eta$ zuiver imaginair, bv. $h\eta = i$).

**c**. We kunnen de strategie in (23) proberen te verbeteren.
Bereken, voor een $N \geq 1$, $u_h$ volgens de zogenaamde P(EC)$^N$E-methode (Predict, $N\times$(Evaluate-Correct), and Evaluate):

$$\begin{cases} u_{n+1}^{(0)} = u_n + hf_n, \\ f_{n+1}^{(i)} \equiv f(t_{n+1}, u_{n+1}^{(i)}), \quad u_{n+1}^{(i+1)} = u_n + hf_{n+1}^{(i)} \quad \text{voor} \quad i = 0, \ldots, N-1, \\ u_{n+1} = u_{n+1}^{(N)}, \quad f_{n+1} \equiv f_{n+1}^{(N)}). \end{cases} \tag{25}$$

Laat zien dat, met $\quad u_{n+1}^{(\infty)} \equiv \lim_{i\to\infty} u_{n+1}^{(i)}, \quad$ geldt $\quad u_{n+1}^{(\infty)} = u_{n+1}^{\rm b} \quad$ (als $u_n = u_n^{\rm b}$) en

$$\|u_{n+1}^{(i)} - u_{n+1}^{(\infty)}\| \leq (hL)^i \|u_{n+1}^{(0)} - u_{n+1}^{(\infty)}\| = \mathcal{O}(h^{i+2}).$$

Bepaal de foutvoortplantingsfaktor van de P(EC)$^N$E-methode in (25). Hoe groot moet $h$ zijn voor probleem (3) met $\epsilon = 10^{-6}$?

**d**. Bij grote $L$ is de P(EC)$^N$E-methode niet zo'n succes. Het zetten van de impliciete stap middels 'n Newton-Raphson achtige aanpak pakt dan gunstiger uit

$$\begin{cases} u_{n+1}^{(0)} = u_n \quad \text{en} \quad J \equiv \left[\dfrac{\partial f_j}{\partial x_j}\Big(t_{n+1}, u_{n+1}^{(0)}\Big)\right] \\ u_{n+1}^{(i+1)} = u_{n+1}^{(i)} - (I - hJ)^{-1}\left(u_{n+1}^{(i)} - hf\Big(t_{n+1}, u_{n+1}^{(i)}\Big) - u_n\right) \quad \text{voor} \quad i = 1, \ldots, N-1, \\ u_{n+1} = u_{n+1}^{(N)}. \end{cases}$$

Analyseer deze aanpak voor $N = 1$ en achtereenvolgens het probleem (3) en het probleem:

$$\begin{cases} \epsilon\, u'(t) = 3u(t) - u^2(t) \quad \text{voor } t \in [0, T], \\ u(0) = 1, \end{cases} \tag{26}$$

Verwacht je verbetering van de keuze $u_{n+1}^{(0)} = u_n + hf(t_n, u_n)$?

### Opgave 10 [Inschakelverschijnselen]

Beschouw, voor $0 < \epsilon \ll 1$ (met bv. $\epsilon = 10^{-4}$), de differentiaalvergelijking

$$\begin{cases} \epsilon\, u'(t) = -u(t) + (\cos t - \epsilon\, \sin t) \quad \text{voor } t \in [0, T], \\ u(0) = 1 + \gamma, \end{cases} \tag{27}$$

(zie (3)). De exacte oplossing $u^*$ bestaat uit twee komponenten: $u^* = w + v$, met

$$w(t) = \cos(t) \quad \text{en} \quad v(t) = \gamma \exp\left(-t/\epsilon\right);$$

$w$ is de exacte oplossing voor $\gamma = 0$ en $w$ varieert langzaam ($w''$ is relatief klein) en $v$ is 'n oplossing van het homogene deel en varieert snel voor kleine $t$. De invloed van $v$ op de oplossing verdwijnt al snel: ga na dat $v(t)$ verwaarloosbaar is voor $t > \sqrt{\epsilon}$. Ga na dat dit ook het geval is voor de tweede orde afgeleiden.
Het effekt van $v(t)$, voor $t$ in de *grenslaag* ($t \leq \sqrt{\epsilon}$), is een "*inschakelverschijnsel*".

We wensen $u^*$ te berekenen met een absolute fout $\leq 10^{-4}$, ook in de grenslaag.

**a**. Bepaal $h$, in termen van $\epsilon$, zodat buiten de grenslaag de Euler-backward-oplossing voldoende nauwkeurig is.

**b**. Schrijf $u_h^{\mathrm{b}} = w_h + v_h$. Ga na dat, met $\eta = -1/\epsilon$,

$$v - v_h(t_n) = \gamma \left( e^{nh\eta} - \left( \frac{1}{1 - h\eta} \right)^n \right).$$

**c**. Om de oplossing in de grenslaag informatief te kunnen weergeven lijkt het redelijk $h$ in ieder geval zo te kiezen dat $|h\,\eta| \leq 1$. Waarom?

Bepaal $h$, in termen van $\epsilon$, zodat de fout in $v$ ook in de grenslaag de vereiste nauwkeurigheid haalt. Vergelijk deze $h$ met die in a en met de "Euler-forward $h$". Heeft het zin om in de grenslaag met de "dure" Euler-backward-methode te werken in plaats van met de goedkope Euler forward?

**d** [Exponentieel fitten] Beschouw, voor een $\alpha > 0$, de volgende kombinatie van Euler forward en Euler backward

$$u_{n+1} = u_n + h\ [(1 - \alpha) f(t_n, u_n) + \alpha f(t_{n+1}, u_{n+1})].$$

Bepaal de lokale diskretisatiefout. Bewijs dat, voor probleem (27), de foutvoortplantingsfaktor van deze methode in absolute waarde $\leq 1$ is voor iedere $h > 0$. Bepaal $\alpha$ zodat ook in de grenslaag met een $h$ gewerkt kan worden die alleen afhangt van de lengte van de grenslaag (zie aanhef c) (Hint: zorg ervoor dat $v - v_h = 0$).

## Midpunt regel en trapezium regel

In de Opgave~11 tot en met Opgave~17 bewijzen we voor een konkrete eenvoudige situatie een groot aantal van de belangrijkste resultaten uit de multistep theorie. De aanpak is hier en daar wat anders dan in de theorie van het diktaat. We gebruiken deze theorie niet expliciet. Door deze andere aanpak en de eenvoudige situatie hopen we een beter inzicht in de theorie te verschaffen.

We bekijken in Opgave~11 t/m Opgave~17 de volgende situatie.
Zij $p, q \in C^\infty(\mathcal{J})$ met $\mathcal{J} = [0, 1]$ en $c \in \mathbf{R}$. Laat $u^*$ de exacte oplossing zijn van het probleem

$$u'(t) = p(t)u(t) + g(t) \quad \text{voor alle} \quad t \in \mathcal{J} \quad \text{en} \quad u(0) = c. \tag{28}$$

Met $N \in \mathbf{N}$ is $h \equiv \frac{1}{N}$, $t_n \equiv nh$, $u_n^* \equiv u^*(t_n)$, $p_n \equiv p(t_n)$ en $q_n \equiv q(t_n)$. Op $\mathcal{J}_h \equiv \{t_n \,\big|\, n = 0, \ldots, N\}$ benaderen we de exacte $u^*$ door $u_h$, waarbij $u_n \equiv u_h(t_n)$ berekend is (met de midpunt regel) volgens

$$\begin{cases} u_0 = c, \quad u_1 = u^*(t_1) + \mathcal{O}(h^2) \\ u_{n+2} = u_n + 2h(p_{n+1}u_{n+1} + q_{n+1}) \quad \text{voor} \quad n = 0, \ldots, N-2. \end{cases} \tag{29}$$

Voor $u_n^*$ geldt

$$u_{n+2}^* = u_n^* + 2h(p_{n+1}u_{n+1}^* + q_{n+1}) + 2h\delta_n$$

waarbij $\delta_n \equiv \delta_h(u^*)(t_n)$ de lokale diskretisatie fout is. We zijn geïnteresseerd in de globale fout $e_h$ in de benadering $u_h$ van $u^*$; $e_n \equiv e_h(t_n) \equiv u_n^* - u_n$ op $\mathcal{J}_h$.

**Opgave 11 [Stabiliteit midpunt regel]**
In deze opgave is

$$\Delta_n \equiv \max\{|u^{*(3)}(\xi)| \Big| \xi \in [t_n, t_{n+2}]\} \qquad \text{en} \qquad P \equiv \max_n |p_n|.$$

**a**. Met $u_1 = u_0 + h(p_0u_0 + q_0)$ is $e_1 = \mathcal{O}(h^2)$ en
met $u_1 = u_0 + \frac{1}{2}h(p_0u_0 + q_0 + p_1u_1 + q_1)$ is $e_1 = \mathcal{O}(h^3)$. Bewijs dit.
**b**. Geef een rekurrente betrekking voor de rij fouten $e_0, \ldots, e_N$.
**c**. Beschouw de rij $\widetilde{e}_0, \ldots, \widetilde{e}_N$ in $[0, \infty)$ die gegeven wordt door

$$\begin{cases} \widetilde{e}_0 = e_0 = 0, \ \widetilde{e}_1 = |e_1|, \\ \widetilde{e}_{n+2} = \widetilde{e}_n + 2hP\widetilde{e}_{n+1} + \frac{1}{3}h^3\Delta_n \quad \text{voor} \quad n = 0, \ldots, N-2. \end{cases}$$

Bewijs dat $|e_n| \leq \widetilde{e}_n$ voor iedere $n$.
**d**. Bereken de karakteristieke wortels $\lambda_1$ en $\lambda_2$ van de *majorerende* rekursie in c.
**e**. Toon aan dat

$$\widetilde{e}_n = G_{n-1}\widetilde{e}_1 + \sum_{j=2}^{n} G_{n-j}\,\frac{1}{3}h^3\Delta_{j-2} \quad (n = 2, \ldots, N),$$

waarbij

$$G_j = \frac{\lambda_1^{j+1} - \lambda_2^{j+1}}{\lambda_1 - \lambda_2}.$$

**f**. Laat zien dat $|\lambda_1| \leq e^{hP}$ en $|\lambda_2| \leq e^{hP}$.
Bewijs nu dat

$$\widetilde{e}_n \leq \frac{1}{3}h^2e^P \max_j \Delta_j + \widetilde{e}_1e^P.$$

**g**. Toon ten slotte aan dat

$$e_h = \mathcal{O}(h^2) \quad \text{uniform} \quad (h \to 0) \quad \text{als} \quad e_1 = \mathcal{O}(h^2).$$

**Opgave 12 [Waarom de fout in essentie glad is]**
Neem in deze opgave voor $p$ de konstante funktie, $p \equiv -1$, en $u_1$ zodat $u_1 = u_1^* + \mathcal{O}(h^3)$.
**a**. Druk, met behulp van b uit de vorige opgave, de fouten $e_n$, voor geschikte positieve konstanten $\lambda$, $\alpha$ en $\beta$ (welke?), uit in de lokale diskretisie fout $\delta_n$ en $G_j = \alpha\lambda^j + \beta(-\frac{1}{\lambda})^j$.
**b**. Toon aan dat voor iedere $n = 0, \ldots, N$ en $i = 1, \ldots, N$ geldt

$$(-\frac{1}{\lambda})^n\delta_i + (-\frac{1}{\lambda})^{n+1}\delta_{i-1} = \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0)$$

**c**. Laat nu met behulp van a en b zien dat

$$e_n = \frac{1}{3}\alpha h^2 \left( \sum_{i=2}^{n} h[u^{*(3)}(t_i)\exp(-t_n + t_i) + \mathcal{O}(h)] \right) + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0).$$

Konkludeer dat

$$\begin{aligned} e_n &= \frac{1}{3}\alpha h^2 \int_0^{t_n} e^{-t_n+\tau}u^{*(3)}(\tau)\,d\tau + \mathcal{O}(h^3) \quad \text{uniform } (h \to 0) \\ &= \frac{1}{6}h^2w^*(t_n) + \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0), \end{aligned}$$

waarbij $w^*(t) \equiv \int_0^t \exp(-t + \tau)u^{*(3)}(\tau)\,d\tau \qquad (t \in \mathcal{J})$.

**d**. We kunnen zeggen dat, bij voldoende kleine $h$, de lokale diskretisatie fouten zich ongeveer voortplanten volgens (dalende) oplossingen van de homogene vergelijking (zie §4.3.10 van het diktaat). Stel dat, naast de diskretisatie fout, in iedere stap ook nog een evaluatie fout $\epsilon_n$ gemaakt wordt (tel in het rechterlid van (29) nog $\epsilon_n$ op). Verwacht je dat deze fouten ook zo "stabiel" voortgeplant worden? Motiveer je antwoord.

**e**. Op tijdstip $\tau_0 \in \mathcal{J}$ (met $\frac{\tau_0}{h} \in \mathbf{N}$) is $x_1 \equiv u_h(\tau_0)$ en $x_2 \equiv u_{\frac{1}{2}h}(\tau_0)$ de benadering van $u^*(\tau_0)$ verkregen met de midpunt regel met stapgrootte $h$, respektievelijk $\frac{1}{2}h$.
Toon aan dat

$$x_2 - u^*(\tau_0) = \frac{1}{3}(x_1 - x_2) + \mathcal{O}(h^3)$$

gesteld dat de evaluatie fouten verwaarloosbaar zijn.
Beschouw de tabellen hieronder. Hierin zijn de midpunt regel benaderingen voor een exacte oplossing $u^*$ afgedrukt. In de "$h_1$-tabel" is in de berekening van $u_h$ in iedere stap een ekstra fout $\varepsilon$ gemaakt. Denk je dat, in deze tabel, de evaluatie fout verwaarloosbaar is? Schat, als je denk dat dat betrouwbaar is, de fout in de kolommen die daarvoor in aanmerking komen.
(In de tabel staat de midpunt regel benaderingen van de exacte oplossing $u^*$ van het probleem

$$u'(t) = -u(t) + \cos(\pi t) - \pi \sin(\pi t) \quad \text{op}\,[0,1] \qquad \text{en} \quad u(0) = 1,$$

met stapgrootte $h_0 = \frac{1}{12}$. Met $h_1 = \frac{1}{6}$. is de perturbatie $\epsilon$ zodat $|\epsilon(t)| \leq 10^{-5}$.)

| $\tau$ | $\frac{1}{6}$ | $\frac{1}{3}$ | $\frac{1}{2}$ |
|---|---|---|---|
| $u^*(\tau)$ | .8660254038 | .5000000000 | .1570796327 $10^{-11}$ |
| $h_0$ | .8823242180 | .5328337740 | .5412464990 $10^{-1}$ |
| $\frac{1}{2}h_0$ | .8702171849 | .5084843042 | .1407234968 $10^{-1}$ |
| $\frac{1}{4}h_0$ | .8670790442 | .5021307745 | .3529201436 $10^{-2}$ |
| $\frac{1}{8}h_0$ | .8662814026 | .5005035059 | .8023047936 $10^{-3}$ |
| $\frac{1}{16}h_0$ | .8660616260 | .5000344041 | −.1785831150 $10^{-4}$ |
| $h_1$ | .8664904544 | .5009331434 | .1528700299 $10^{-2}$ |
| $\frac{1}{2}h_1$ | .8661249420 | .5001735028 | .2298564530 $10^{-3}$ |
| $\frac{1}{4}h_1$ | .8659995984 | .4999042421 | −.2484150841 $10^{-3}$ |

Vertrouwensgetallen $V_h$:

in $\tau$ is $\quad V_h(\tau) \equiv \dfrac{u_h(\tau) - u_{\frac{1}{2}h}(\tau)}{u_{\frac{1}{2}h}(\tau) - u_{\frac{1}{4}h}(\tau)}$

| $\tau$ | $\frac{1}{6}$ | $\frac{1}{3}$ | $\frac{1}{2}$ |
|---|---|---|---|
| $h_0$ | 3.86 | 3.83 | 3.80 |
| $\frac{1}{2}h_0$ | 3.93 | 3.90 | 3.87 |
| $\frac{1}{4}h_0$ | 3.63 | 3.47 | 3.32 |
| $h_1$ | 2.92 | 2.82 | 2.72 |

In Opgave˜12 zagen we een fout $e_h$ van de vorm $\frac{1}{6}h^2 w^* + \mathcal{O}(h^3)$ uniform: de fout is in essentie glad. De term $\frac{1}{6}h^2 w^*$ is, zeker voor kleinere $h$, het belangrijkste deel van de fout; men noemt deze funktie daarom ook wel de *principal error function*. In de volgende opgave tonen we het bestaan aan van zo'n principal error function voor het geval de midpunt regel wordt toegepast op de algemenere lineaire differentiaalvergelijking in (28). Zie ook het bewijs van stelling 4.3.7.

**Opgave 13** Laat $w^* \in C^1(\mathcal{J})$ de exacte oplossing zijn van

$$w' = pw + u^{*(3)} \quad \text{op}\,[0,1] \qquad \text{en} \quad w(0) = 0.$$

**a**. Schrijf $\widetilde{w} \equiv +\frac{1}{6}h^2 w^*$. Ga na da

$$\widetilde{w}' = p\widetilde{w} + \frac{1}{6}h^2 u^{*(3)} \quad \text{op} \quad [0,1] \quad \text{en} \quad \widetilde{w}(0) = 0.$$

Verder is $\widetilde{w}^{(3)} = \mathcal{O}(h^2)$ uniform en $\widetilde{w}(h) = \mathcal{O}(h^3)$ als $h \to 0$. Ga dit na.
We berekenen $u_h$ weer met de midpunt regel, nu met $u_0 = u_0^*$ en $u_1 = u_1^* + \mathcal{O}(h^3)$ (in de volgende redenering is het niet voldoende dat $u_1 = u_1^* + \mathcal{O}(h^2)$; ga dit telkens na in iedere redeneerstap). Beschouw naast de fout $e_h$ ook het verschil $v_n \equiv v_h(t_n) \equiv e_h(t_n) - \widetilde{w}(t_n)$ op $\mathcal{J}_h$.
**b**. Bewijs dat

$$\begin{cases} v_0 = 0, \ v_1 = \mathcal{O}(h^3) \\ v_{n+2} = v_n + 2hp_{n+1}v_{n+1} + \mathcal{O}(h^4) \quad \text{uniform} \quad (h \to 0). \end{cases}$$

(Hint. Schrijf de rekursie op voor de $e_n$ en pas een geschikte diskretisatie toe op de differentiaalvergelijking in a)
**c**. Toon nu aan dat $v_h = \mathcal{O}(h^3)$ uniform $(h \to 0)$.
(Hint. Ga te werk als in Opgave~11, c t/m g.)
**d**. Stel dat $e_1 = \mathcal{O}(h^4)$. Merk op dat dan

$$v_1 = -\frac{1}{6}h^2 w^*(h) + \mathcal{O}(h^4) = -\frac{1}{6}h^3 w^{*(1)}(0) + \mathcal{O}(h^4) = -\frac{1}{6}h^3 u^{*(3)}(0) + \mathcal{O}(h^4).$$

Gemakshalve beschouwen we nu het geval dat $p(t) = \eta$ $(t \in [0,1])$.
Laat zien dat $G_{n-1}$ in de bijdrage $G_{n-1}v_1$ in de verschilterm $v_n$ voldoet aan

$$G_{n-1} = \frac{1}{2}\left(\lambda_1^n - \lambda_2^n\right)(1 + \mathcal{O}(h^2)) = \frac{1}{2}\left(e^{\eta t_n} - (-1)^n e^{-\eta t_n}\right)(1 + \mathcal{O}(h)),$$

waarbij $\lambda_1$ en $\lambda_2$ de karakteristieke wortels zijn van de rekursie in b.
Schrijf $w_{21}(t) \equiv \frac{1}{12}u^{*(3)}(0)e^{\eta t}$ en $w_{22}(t) \equiv \frac{1}{12}u^{*(3)}(0)e^{-\eta t}$ $(t \in [0,1])$. De bijdrage van $v_1$ in de fout $e_n$ is dus van de vorm $h^3[w_{21}(t_n) - (-1)^n w_{22}(t_n)] + \mathcal{O}(h^4)$.
Overtuig je ervan dat er geen $w_2 \in C^\infty([0,1])$ is waarvoor

$$e_h = \frac{1}{6}h^2 w^* + h^3 w_2 + \mathcal{O}(h^4) \quad \text{uniform} \quad (h \to 0).$$

**e**. Beschouw ook de funktie $\widetilde{u}_h$ die gegeven is door

$$\widetilde{u}_h(t_n) \equiv \frac{1}{4}\left[u_h(t_{n-1}) + 2u_h(t_n) + u_h(t_{n+1})\right] \qquad (t_n \in \mathcal{J}_h, 0 < t_n < 1).$$

Men kan laten zien dat er funkties $w_1, w_2, \ldots \in C^\infty([0,1])$ zijn zodat

$$u^*(t_n) - \widetilde{u}_h(t_n) = h^2 w_1(t_1) + h^3 w_2(t_n) + \ldots + h^m w_{m-1}(t_n) + \mathcal{O}(h^{m+1}) \quad \text{uniform} \quad h \to 0.$$

Dit hoeft je niet te laten zien; laat wel zien dat er zo'n funktie $w_1$ is en dat de bijdrage van $v_1$ in $\widetilde{u}_h$ in essentie wel glad is (van de vorm $h^3\widetilde{w}_2 + \mathcal{O}(h^4)$ met $\widetilde{w}_2 \in C^\infty([0,1])$ ).

In Opgave~12e zagen we hoe de principal error function gebruikt kan worden om een methode te ontwerpen die al rekenend een foutschatting geeft. Zo'n foutschatting kan dan natuurlijk ook gebruikt worden om de gevonden benadering te korrigeren (ga dit na, ook middels de getallen in de tabel). Het uitvoeren van zo'n verbetering noemt men *Richardson extrapolatie*. Natuurlijk rijst de vraag of in deze verbetering ook weer de fout al rekenend geschat kan worden.
(Wat abstrakter kan men het gebeuren als volgt beschrijven:
Voor $d, c, I \in \mathbf{R}$ is

$$B(h) \equiv I + ch^2 + dh^3 + \mathcal{O}(h^4) \quad \text{voor} \quad h \geq 0, \, h \to 0.$$

Voor $h > 0$ kunnen we $B(h)$ berekenen. We wensen $I = \lim_{h\to 0} B(h)$ te benaderen. Omdat

$$B(h) - I \approx \frac{1}{3}(3ch^2 + 7dh^3 + \mathcal{O}(h^4)) \approx \frac{1}{3}[B(2h) - B(h)]$$

kunnen we de fout al rekenend schatten. Met

$$\widetilde{B}(h) \equiv B(h) - \frac{1}{3}h[B(2h) - B(h)] = I - \frac{4}{3}dh^4 + \mathcal{O}(h^4)$$

hebben we, als $h \to 0$, een betere benadering van $I$ dan $B(h)$. Bovendien geeft

$$\widetilde{B}(h) - I \approx -\frac{4}{3}dh^3 + \mathcal{O}(h^4) \approx \frac{1}{7}[\widetilde{B}(2h) - \widetilde{B}(h)]$$

een schatting voor de fout in de nieuwe benadering $\widetilde{B}(h)$. Deze schatting kan al rekenend bepaald worden.)
In Opgave˜13d zagen we dat dit, in geval van de midpunt regel, *niet* kan. (Om een Romberg schema te kunnen opzetten met een multistep benadering moet de startfout een asymptotische ontwikkeling hebben — $e_1$ speelt een rol in $v_1$ en dus in de $\mathcal{O}(h^3)$-term in c — én de multistep moet sterk stabiel zijn.) Deze moeilijkheden doen zich bij de trapezium regel niet voor (we kunnen dan in het algemeen foutloos starten en de trapezium regel is sterk stabiel).

In Opgave˜14, Opgave˜15 en Opgave˜16 benaderen we $u^*$ op $\mathcal{J}_h$ ook nog door $x_h \in C(\mathcal{J}_h)$ die, met $x_n \equiv x_h(t_n)$, berekend is (met de trapezium regel) volgens

$$\begin{cases} x_0 = u_0^* \\ x_{n+1} = x_n + \frac{1}{2}h\,[p_{n+1}x_{n+1} + q_{n+1} + p_n x_n + q_n] \qquad \text{voor} \quad n = 0, \ldots, N-1. \end{cases} \tag{30}$$

**Opgave 14 [Trapezium regel]**
**a**. Bewijs dat voor een $g \in C^5([0,1])$ geldt

$$\int_0^h g(\tau)\,d\tau = \frac{1}{2}h\,[g(h) + g(0)] - \frac{1}{12}h^2\left(\frac{1}{2}h\,[g''(h) + g''(0)]\right) + \mathcal{O}(h^5) \quad \text{als} \quad (h \to 0).$$

(Hint. Beschouw Taylor reeksen rond $\frac{1}{2}h$.)
**b**. Konkludeer dat

$$u_{n+1}^* = u_n^* + \frac{1}{2}h\,[f_{n+1}^* + f_n^*] - \frac{1}{12}h^2\left(\frac{1}{2}h\left[u^{*(3)}(t_{n+1}) + u^{*(3)}(t_n)\right]\right) + \mathcal{O}(h^5),$$

waarbij $f_j^* \equiv p_j u_j^* + q_j$.
**c**. Ga nu als in Opgave˜13, a, b en c, te werk en bewijs dat

$$x_h = u^* + \frac{1}{12}h^2 w^* + \mathcal{O}(h^4) \quad \text{uniform} \quad (h \to 0)$$

met $w^*$ als in de aanhef van Opgave˜13.

Het resultaat in Opgave˜14c kan bijvoorbeeld gebruikt worden om in iedere willekeurig tijdstip $t_n$ de lokale diskretisatie fout te schatten door in dat tijdstip bijvoorbeeld één midpunt iteratie uit te voeren met als startwaarden $u_{n+1}$ en $u_n$ (zie de volgende opgave).

**Opgave 15 [Schatten van de lokale fout]**

**a.** Met $w^*$ als in Opgave~13 en $\widehat{w} \equiv \frac{1}{12}h^2 w^*$ geldt $\widehat{w}' = p\widehat{w} + \frac{1}{12}h^2 u^{*(3)}$. Bewijs dit en toon aan dat

$$\widehat{w}(t_{n+2}) = \widehat{w}(t_n) + 2h\left[p_{n+1}\widehat{w}(t_{n+1}) + \frac{1}{12}h^2 u^{*(3)}(t_{n+1})\right] + \mathcal{O}(h^5).$$

**b.** Toon nu aan dat

$$\begin{aligned} x_{n+2} - (x_n + 2h[p_{n+1}x_{n+1} + q_{n+1}]) &= \tfrac{1}{3}h^3 u^{*(3)}(\xi_n) + \tfrac{1}{6}h^3 u^{*(3)}(t_n) + \mathcal{O}(h^4) \\ &= \tfrac{1}{2}h^3 u^{*(3)}(t_n) + \mathcal{O}(h^4). \end{aligned}$$

(Hint. Gebruik Opgave~14c.)

**c.** Met

$$\widetilde{x}_{n+2} \equiv x_n + 2h[p_{n+1}x_{n+1} + q_{n+1}]$$

is

$$\widetilde{x}_{n+2} - x_{n+2} = -\frac{1}{2}h^3 u^{*(3)}(t_n) + \mathcal{O}(h^4).$$

Toon dit aan en ga na dat voor de lokale diskretisatie fout in de trapezium regel geldt

$$-\frac{1}{12}h^3 u^{*(3)}(\xi_n) = \frac{1}{6}(\widetilde{x}_{n+2} - x_{n+2}) + \mathcal{O}(h^4) \quad \text{uniform} \quad (h \to 0).$$

**d.** Verwacht je dat je met behulp van $\widetilde{u}_{n+1} \equiv u_n + \frac{1}{2}h[p_{n+1}u_{n+1} + q_{n+1} + p_n u_n + q_n]$ (met $u_n$ de midpunt regel benadering) de lokale diskretisatie fout in de midpunt regel kan schatten? Motiveer je antwoord.

Benaderen we naast elkaar $u^*$ door de trapezium regel oplossing $x_h$ en de midpunt regel oplosing $u_h$ dan kunnen we (als $u_1 = u_1^* + \mathcal{O}(h^3)$), dankzij het feit dat de globale fouten in essentie glad zijn, uit beide benaderingen de globale fout schatten. We doen dat in de volgende opgave.

**Opgave 16 [Schatten van de globale fout]**

**a.** Neem nu $u_0 = x_0 = u_0^*$ en $u_1 = x_1$. Bewijs (m.b.v. Opgave~13c en Opgave~14c) dat

$$2(x_n - u_n^*) + (u_n - u_n^*) = \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0).$$

**b.** Hoe kan je uit de waarden $x_n$ en $u_n$ in de benedenstaande tabel de fout $e_n$ in de benadering $x_n$ van $u_n^*$ schatten? Kan je deze schatting gebruiken om de benadering $x_n$ van $u_n^*$ te korrigeren? Zo ja, hoe pakt dat hier uit? Wat kan je zeggen over de fout in deze gekorrigeerde benadering?

| $n$ | $u_n^*$ | $x_n$ | $u_n$ |
|---|---|---|---|
| 20 | 0.30902 | 0.30939 | 0.30828 |
| 40 | −0.80902 | −0.80695 | −0.81314 |
| 60 | −0.80902 | −0.80164 | −0.82368 |
| 80 | 0.30902 | 0.33424 | 0.25895 |
| 100 | 1.00000 | 1.08814 | 0.82527 |

Het is nogal dubieus om op basis van één berekening een fout te schatten in $t_n$. Men weet niet of de hogere orde termen verwaarloosbaar zijn. Desondanks gebruikt men in de praktijk dit soort foutschattingen nogal eens; vooral die van het type in Opgave~15c is populair.

Merk op dat bij de schatting van de lokale diskretisatie fout in $t_n$, zoals in Opgave~15c, men slechts een midpunt stap hoeft uit te voeren. Voor de schattingen in $t_n$ van de globale fout moet men daarentegen de midpunt-benadering in *alle* voorafgaande tijdstippen $t_j$ (j=0,...,n) berekend hebben.

In de volgende opgave geven we een laatste toepassing van de principal error function. In Opgave˜12 hebben we een methode gezien om een gevonden benadering te korrigeren (via Richardson extrapolatie). De methode die we in de volgende opgave tegenkomen verbetert ook een gevonden benadering. We schatten de lokale diskretisatie fout middels de al gevonden benadering en gebruiken de schatting om de multistep te korrigeren. Na korrektie wordt de multistep opnieuw doorlopen. We spreken van *deferred correction.*

**Opgave 17 [Deferred correction]**

**a**. Toon aan dat voor het geval $u_1 = u_1^* + \mathcal{O}(h^3)$ geldt

$$\widetilde{\delta}_n \equiv \frac{1}{h^2}[f_{n+1} - 2f_n + f_{n-1}] = u^{*(3)}(t_n) + \mathcal{O}(h) \quad \text{uniform} \quad (h \to 0),$$

waarbij $f_n \equiv p_n u_n + q_n$.

**b**. Stel nu dat $z_h$ met $z_n \equiv z_h(t_n)$, berekend is volgens

$$\begin{cases} z_0 = u_0 = u_0^*, \; z_1 = u_1 \\ z_{n+2} = z_n + 2h[p_{n+1}z_{n+1} + q_{n+1}] + \frac{1}{3}h^3\widetilde{\delta}_n. \end{cases}$$

Zij $v_n \equiv v_h(t_n) \equiv u_n^* - z_n$ op $\mathcal{J}_h$. Toon aan dat

$$v_{n+2} = v_n + 2hp_{n+1}v_{n+1} + \mathcal{O}(h^4) \quad \text{uniform} \quad (h \to 0).$$

**c**. Ga te werk als in Opgave˜11 om te laten zien dat

$$v_n = \mathcal{O}(h^3) \quad \text{uniform} \quad (h \to 0).$$

Konkludeer dat $z_n = u_n^* + \mathcal{O}(h^3)$ uniform $(h \to 0)$.

**d**. Vergelijk de hoeveelheid werk die extra nodig is om de $(z_n)$ te berekenen met de hoeveelheid werk die nodig is om de $(u_n)$ te berekenen.

## Stabiliteit

**Opgave 18** Beschouw, voor $\beta \in [0, \infty)$, met $u_n \equiv u_h(t_n)$, de multistep

$$u_{n+2} = u_{n+1} + h\left[\beta f(t_{n+2}, u_{n+2}) + \left(\frac{3}{2} - 2\beta\right) f(t_{n+1}, u_{n+1}) + \left(\beta - \frac{1}{2}\right) f(t_n, u_n)\right].$$

**a**. Bewijs dat deze multistep stabiel en konsistent is.

**b**. Bepaal de orde en de foutkonstante.

**c**. Beschouw, voor $\eta \in \mathbf{R}\backslash\{-1\}$, het probleem

$$u'(t) = \eta u(t) + e^{-t} \quad \text{voor} \quad t \in [0, 1] \quad \text{en} \quad u(0) = 0.$$

Ga na dat

$$u^*(t) = \frac{1}{1+\eta}\left(e^{\eta t} - e^{-t}\right)$$

de exacte oplossing beschrijft.

Neem nu verder $\eta = -100$. We gaan na hoe groot we $h$ moeten kiezen opdat $u^*$ op $[0, 1]$ benaderd wordt met een absolute fout $\leq 10^{-4}$.

**d**. Waarom mag stelling 4.3.7 worden toegepast? Ga na dat de oplossing van (32), voor $\beta = \frac{1}{2}$, gegeven wordt door

$$\left(\frac{1}{1+\eta}\right)^2 \left[(1+\eta)\eta^3 t e^{\eta t} + e^{\eta t} - e^{-t}\right].$$

Het maximum hiervan wordt aangenomen voor $t \approx 10^{-2}$.
Bepaal nu $h$ zodanig dat de absolute fout op [0,1] uniform kleiner is dan $10^{-4}$.

**e**. Maak duidelijk dat er voor $h\eta = -1$ een $\beta$, zeg $\beta^*$, te vinden is zodat de daar bij passende multistep, voor iedere $u_0 \in \mathbf{R}$, de oplossing van het probleem $u' = \eta u$ op $[0,1]$ en $u(0) = u_0$ exact benadert in de punten $t_n = nh$.

Neem verder $h = 10^{-2}$ en $\beta = \beta^*$.

**f**. Wat is de lokale diskretisatie fout bij het eerste probleem?

**g**. Geef een redelijke schatting voor de absolute fout uniform op $[0,1]$ m.b.v. stelling 4.3.7; ga er van uit dat $u^{*(l+1)}(t)$ in (30) vervangen mag worden door $\frac{1}{99}e^{-t}$ (waarom is dit redelijk?).

**Opgave 19** De trapezium regel is een multistep methode van orde 2.

**a**. Bewijs dat de trapezium regel A-stabiel is.

**b**. Neem nu de multistep met schema $(\chi^2 - \chi, \frac{1}{2}[3\chi - 1])$ (Adams–Bashforth) als prediktor en de trapezium regel als korrektor. Als men een korrektor stap uitvoert is de orde nog steeds 2. Toon dit aan.

**c**. Is het proces in b nog steed A-stabiel?

**d**. Onderzoek algemener hoe het met de orde van het verkregen proces gesteld is als men een "prediktor" van orde $l_p$ neemt en een "korrektor" van orde $l_c$ en met de korrektor $N$ korrektieslagen per stap uitvoert.

**Opgave 20**

**a**. Bepaal het stabiliteitsgebied van de midpunt regel en van de trapezium regel.

**b**. Gebruik Euler forward als prediktor voor de trapezium regel. Analyseer de stabiliteit en de lokale diskretisatiefout van de P(EC)E-methode.

**c**. Beantwoord dezelfde vraag als in b, maar nu met de midpunt regel als prediktor.

**Opgave 21 [Sterk stabiel, A-stabiel]**

Beschouw, voor $\alpha, \beta, \gamma \in \mathbf{R}$, met $u_n \equiv u_h(t_n)$, de multistep

$$u_{n+2} = u_{n+1} + h\left[\alpha f(t_{n+2}, u_{n+2}) + \beta f(t_{n+1}, u_{n+1}) + \gamma f(t_n, u_n)\right].$$

**a**. Bewijs dat deze multistep stabiel is.

**b**. Toon aan dat orde minstens 2 is precies dan als $\gamma = \alpha - \frac{1}{2}$ én $\beta = \frac{3}{2} - 2\alpha$.
We nemen verder aan dat $\beta$ en $\gamma$ zo gekozen zijn.

**c**. Beschouw de rekursie $u_{n+2} = a u_{n+1} + b u_n$ $(n \in \mathbf{N}_0)$.
Als $|a| + |b| \leq 1$ dan geldt $|u_n| \leq (|a| + |b|)^{\frac{n}{2}} \max(|u_0|, |u_1|)$ $(n \in \mathbf{N}_0)$. Toon dit aan.

**d**. Zij, voor $\eta \in (-\infty, 0)$ en gegeven $u_0, u_1 \in \mathbf{R}$, $u_h$ de multistep oplossing van het probleem $u' = \eta u$ op $[0, \infty)$ met $u_h(0) = u_0$ en $u_h(t_1) = u_1$.
Als $\alpha \in [\frac{1}{2}, 1)$ dan geldt voor iedere stapgrootte $h$ dat $\lim_{n\to\infty} u_h(t_n) = 0$. Bewijs dit.

**e**. Beschouw nu voor $d = 2$ het probleem $u'(t) = f(t, u(t))$ op $[0, \infty)$ en $u(0) = u_0$ en laat de funktionaal matrix $J(t) = \frac{\partial f}{\partial x}(t, u^*(t))$ gegeven worden door

$$J(t) = \begin{bmatrix} -1000 & \sin(t) \\ \cos(t) & -2 \end{bmatrix}.$$

Welke beperking wordt door de stabiliteit (absolute stabiliteit) aan de stapgrootte h opgelegd als men de multistep gebruikt met $\alpha \in [\frac{1}{2}, 1)$?

**f.** Vaak past men bij multistep methoden de prediktor-korrektor techniek toe (zie 4.4.7 van het diktaat). Noem *twee* redenen waarom men dit niet doet bij stijve differentiaalvergelijkingen. (Hint. Kies bv. de midpunt regel als prediktor: $u^p_{n+2} = u_n + 2hf(t_{n+1}, u_{n+1})$. Ga, als $f(t,x) = \eta x$, na wat de fout in de voorspelde waarde is als $u_n$ en $u_{n+1}$ met kleine foutjes behept zijn. Ga ook de korrektie iteraties na.)

**Opgave 22 [A-stabiel]**

Beschouw, met $u_n = u_h(t_n)$ de multistep

$$u_{n+2} = \frac{4}{5}u_{n+1} + \frac{1}{5}u_n + h\frac{2}{5}\left[f(t_{n+2}, u_{n+2}) + 2f(t_{n+1}, u_{n+1})\right].$$

**a.** Laat zien dat de multistep konsistent is en dat de orde $\geq 2$ is (in feite is de orde 3).

**b.** Laat zien dat de multistep stabiel is.

**c.** Zij $f(t,x) \equiv \eta x$ waarbij $\eta \in (-\infty, 0)$. Schrijf $\widetilde{\eta} \equiv h\eta$.
Ga na dat $(u_n)$ voldoet aan de rukursie

$$\left(1 - \frac{2}{5}\widetilde{\eta}\right) u_{n+2} - \frac{4}{5}(1+\widetilde{\eta})\, u_{n+1} - \frac{1}{5}u_n = 0 \quad \text{voor} \quad n \in \mathbf{N}_0.$$

Laat $\lambda_1(\widetilde{\eta})$ en $\lambda_2(\widetilde{\eta})$ de karakteristieke wortels van de rekursie zijn, zo genummerd dat $\lambda_1(0) = 1$. Bereken $\lambda_1$ en $\lambda_2$ in $\widetilde{\eta} = 0, -1, -4$ en schets de grafiek van de funkties $\lambda_1$ en $\lambda_2$ op $[-4, 0]$.
Leidt hieruit af dat de groeifaktor $\lambda \equiv \lambda(\widetilde{\eta}) \equiv \max_i |\lambda_i(\widetilde{\eta})|$ (zie (43) in het diktaat) gegeven wordt door

$$\lambda = \lambda_1(\widetilde{\eta}) \quad \text{voor} \quad \widetilde{\eta} \in [-1, 0] \quad \text{en} \quad \lambda = |\lambda_2(\widetilde{\eta})| \quad \text{voor} \quad \widetilde{\eta} \in [-4, -1].$$

Voor welke waarden van $h$ zal de berekende oplossing nog naar nul gaan ($\lim_{n\to\infty} u_h(t_n) = 0$)? Stel dat $\lambda_1(\widetilde{\eta})$ in redelijke mate lijkt op $e^{\widetilde{\eta}}$. Welke eis zul je dan redelijkerwijs opleggen op de stapgrootte $h$?

**d.** Zij verdere $f(t,x) \equiv \eta x + e^{-t}$. De oplossing $u^*$ van de differentiaalvergelijking met $u(0) = u_0$ wordt, als $\eta \neq -1$, gegeven door (ga dit na)

$$u^*(t) = \left(u_0 + \frac{1}{1+\eta}\right) e^{\eta t} - \frac{1}{1+\eta} e^{-t}.$$

Stel dat men het probleem benaderend oplost met de gegeven multistep. Zij $\delta_1$ de diskretisatie fout gemaakt in de eerste stap. Toon aan dat het effekt van deze ene fout in de $n$-de stap gegeven wordt door $(\gamma_1\lambda_1(\widetilde{\eta})^n + \gamma_2\lambda_2(\widetilde{\eta})^n)\delta_1$, waarbij $\gamma_i \in \mathbf{R}\backslash\{0\}$ wel van $\widetilde{\eta}$ afhangen maar niet van $n$.

**e.** Zij nu $\eta \ll -1$. Maak duidelijk (in hoogstens 5 regels) waarom het wenselijk is de stapgrootte $h$ zo te laten zijn dat $\lambda(\widetilde{\eta}) \leq e^{-h}$.

**Opgave 23 [Konvergentie]**

Beschouw, voor $\beta_0, \beta_1 \in \mathbf{R}$, met $u_n = u_h(t_n)$, de multistep

$$u_{n+2} = 2u_{n+1} - u_n + h[\beta_0 f(t_{n+2}, u_{n+2}) + \beta_1 f(t_{n+1}, u_{n+1})].$$

**a.** Bepaal $\beta_0$ en $\beta_1$ zo dat de orde maximaal is.

**b.** Voor $g \in C([0,T])$ en $u_0 \in \mathbf{R}$, passen we de multistep toe op het probleem

$$u'(t) = u(t) + g(t) \quad \text{voor}\, t \in [0, T] \qquad \text{en} \quad u(0) = u_0.$$

Bepaal de orde van de globale diskretisatie fout.
(Hint. Zet de 2-staps rekursie voor de globale diskretisatie fout om in een 1-staps matrix-vektor rekursie $\vec{e}_{n+1} = A\vec{e}_n$ en toon aan dat $\|A^n\| = \mathcal{O}(\frac{1}{h})$ op $[0, T]$.)

**c**. Geef aan welke orde de startfouten moeten hebben als men nog konvergentie wenst; geef ook aan hoe men startwaarden zou kunnen bepalen met fouten van zodanige orde dat de bijdrage van de startfouten tot de globale diskretisatie fouten van hogere orde is dan het totale effekt van de lokale diskretiesatie fouten.

**Opgave 24 [Stapgrootte besturing]**
Beschouw, voor $a_1, a_2, \alpha_1, \alpha_2 \in \mathbf{R}$ en $b_1, b_2, \beta_0, \beta_2 \in \mathbf{R}$, $\beta_0 \neq 0$, het prediktor-korrektor paar

$$u_{n+2} = a_1 u_{n+1} + a_2 u_n + h[b_1 f(t_{n+1}, u_{n+1}) + b_2 f(t_n, u_n)] \qquad \text{(P)}$$
$$u_{n+2} = \alpha_1 u_{n+1} + \alpha_2 u_n + h[\beta_0 f(t_{n+2}, u_{n+2}) + \beta_2 f(t_n, u_n)]. \qquad \text{(C)}$$

**a**. Stel dat de prediktor (P) konsistent is. Toon aan dat hij instabiel is als $a_1 < 0$ is of als $a_1 > 2$ is. Is het van belang of een prediktor stabiel is?

**b**. Kies de konstanten in (P) zodat de prediktor exact is voor polynomen van graad $\leq 2$. Druk de koëfficiënten uit in $a_2$.

**c**. Met $a_2 = \frac{1}{2}$ wordt de prediktor

$$u_{n+2} = \frac{1}{2}u_{n+1} + \frac{1}{2}u_n + \frac{1}{4}h[7f(t_{n+1}, u_{n+1}) - f(t_n, u_n)].$$

Bepaal de lokale diskretisatie fout door aan te nemen dat deze de gedaante $ch^l u^{*(l+1)}(\xi)$ heeft.

**d**. Maak de korrektor van dezelfde orde als de prediktor. Er is nog een vrije parameter. Waarom lijkt $\alpha_1 = 1$ een goede keuze? Bepaal voor deze keuze de lokale diskretisatie fout.

**Opgave 25 [Stabiliteit prediktor-korrektor]**
Beschouw het probleem

$$u'(t) = -2u(t) + 2\cos(t) - \sin(t) \quad \text{voor} \quad t \in [0, 100] \quad \text{en} \quad u(0) = 0.$$

We onderzoeken de stabiliteits eigenschappen van de volgende "prediktor-korrektor methode" toegepast op dit probleem met stapgrootte $h$.

$$u_{n+2} = -4u_{n+1} + 5u_n + h[4f(t_{n+1}, u_{n+1}) + 2f(t_n, u_n)] \qquad \text{(P)}$$
$$u_{n+2} = u_n + \tfrac{1}{3}h[f(t_{n+2}, u_{n+2}) + 4f(t_{n+1}, u_{n+1}) + f(t_n, u_n)]. \qquad \text{(C)}$$

**a**. Ga na dat $u^*(t) = -e^{-2t} + \cos(t)$ de exacte oplossing is van het probleem.

**b**. Toon aan dat de prediktor (P) orde 3 heeft. Is de prediktor stabiel?

**c**. De korrektor (C) heeft orde 4 (dit hoef je niet na te gaan).
Als $u_h$ de korrektor oplossing is met startfouten van orde 4 en $u_n = u_h(t_n)$ dan geldt

$$u_{n+2} + 4u_{n+1} - 5u_n - h[4f(t_{n+1}, u_{n+1}) + 2f(t_n, u_n)] = h^4 C_p u^{*(4)}(t_n) + \mathcal{O}(h^5)$$

($C_p$ is de foutkonstante van de prediktor. Vergelijk dit met stelling 4.4.5 in het diktaat; deze bewering volgt door het bewijs van de stelling na te lopen. Dit hoef je niet te doen.)
Ga wel na of de korrektor stabiel is.

**d**. Is het voor de prediktie-korrektie van belang (als $h$ willekeurig klein is) of de prediktor stabiel is?

**e**. Laat zien dat het absolute stabiliteits gebied van de korrektor geen negatief reëel element bevat. (Aarzel niet de formule voor de vierkantsvergelijking te gebruiken). Wat betekent dit voor het probleem?

**f**. Beschouw nu, met bovenstaande prediktor en korrektor, de P(EC)-methode (zie 4.4.7 in het diktaat). Van welke orde zullen de lokale diskretisatie fouten zijn?

**g**. Zij $\mathcal{S}$ het absolute stabiliteits gebied van deze P(EC)-methode. Men kan bewijzen dat $\mathcal{S} \cap (-\infty, 0] = [-1, 0]$.
Geef kort weer wat dit betekent voor een stijve differentiaalvergelijking. Welke stapgrootte is nog, bij het probleem in de aanhef van de opgave, toelaatbaar om voor deze P(EC)-methode stabiliteit te hebben?

**Opgave 26** Beschouw de volgende methode om $u_n = u_h(t_n)$ te berekenen

$$\begin{cases} u_0 = u_0^* \\ x_{n+1} = u_n + hf(t_n, u_n) \\ u_{n+1} = u_n + \frac{1}{2}h[f(t_{n+1}, x_{n+1}) + f(t_n, u_n)]. \end{cases}$$

Dit is een Runge-Kutta methode (of, zo men wil, een P(EC)-methode).
Laat $\delta_n^*$ de lokale diskretisatie fout zijn, dus:

$$\begin{aligned} &\text{met} \quad x_{n+1}^* \equiv u_n^* + hf(t_n, u_n^*) \quad \text{geldt} \\ &u_{n+1}^* = u_n^* + \tfrac{1}{2}h[f(t_{n+1}, x_{n+1}^*) + f(t_n, u_n^*)] + h\delta_n^*. \end{aligned}$$

Neem verder $\mathcal{J} = [0, 1]$ en $f(t, x) \equiv p(t)x + q(t)$, waarbij $p, q \in C^\infty(\mathcal{J})$.
**a**. Omdat $x_{n+1}^* = u_{n+1}^* - \frac{1}{2}h^2 u^{*(2)}(\xi_n)$ voor zekere $\xi_n \in [t_n, t_{n+1}]$ is

$$h\delta_n^* = -\frac{1}{12}h^3 u^{*(3)}(\eta_n) + \frac{1}{4}h^3 p_{n+1} u^{*(2)}(\xi_n) = \mathcal{O}(h^3) \quad \text{zekere} \quad \eta_n \in [t_n, t_{n+1}].$$

Bewijs dit. Er is een funktie $g \in C(\mathcal{J})$ zodat, voor iedere $h$, geldt $h\delta_n^* = h^3 g(t_n) + \mathcal{O}(h^4)$ uniform $(h \to 0)$. Ga na dat $g$ niet alleen van $u^*$ afhangt, maar ook van $f$.
(Hint. Beschouw bv. $u'(t) = 3t^2$ en $u'(t) = u(t) - t^3 + 3t^2$.)
**b**. Schrijf $e_n \equiv e_h(t_n) \equiv u^*(t_n) - u_n$. Met $G_n \equiv 1 + \frac{1}{2}h[p_{n+1} + p_n] + \frac{1}{2}p_{n+1}p_n h^2$ is

$$e_{n+1} = G_n e_n + h\delta_n^* \quad \text{voor} \quad n = 0, \ldots, N-1.$$

Bewijs dit. Toon aan dat $e_n = \mathcal{O}(h^2)$ uniform $(h \to 0)$.

**Opgave 27** Zij $\mu, \beta, \varepsilon \in (0, \infty)$ en $u_0, w_0 \in \mathbf{R}$. We zijn geïnteresseerd in de oplossing $u \in C^2([0, \infty))$ van de volgende *van der Pol* vergelijking

$$\begin{cases} \varepsilon u'' = \mu(u' + \beta) - u - (u' + \beta)^3 \quad \text{op} \quad [0, \infty) \\ u(0) = u_0 \\ u'(0) = w_0. \end{cases}$$

Schrijf deze tweede orde vergelijking, met behulp van een tweede funktie $v \equiv u' + \beta$, om in een stelsel eerste orde vergelijkingen.

We wensen dit eerste orde stelsel numeriek op te lossen.
Onderzoek numerieke eigenschappen van de eerste en tweede orde expliciete Adams methode en de tweede orde impliciete Adams methode toegepast op dit stelsel.
(Neem voor $w_0 = -\beta$, $u_0 = 0.1$, $\epsilon = \frac{1}{19}$, $\mu = 1$ en voor $\beta = 0.573380 : 0.000002 : 0.573390$.
Werk met $h = 0.01$ (en $t \in [0, 20]$. Plot de grafiek van $v$; $v(t)$ behoort tot $[-1, 1.7]$). Hoe gevoelig hangt de oplossing af van $h$ en van $\beta$? Hangt het een en ander af van het kontinue probleem of van de diskretisatiemethode? Welke $h$ is maximaal "toelaatbaar" (i.v.m. de stabiliteit)? Heeft een van de methoden een duidelijke voorkeur? Wat valt er theoretisch over te zeggen?)
Onderzoek een procedure, gebaseerd op de expliciete en de impliciete Adams methode van orde twee, waarin automatisch de stapgrootte bestuurd wordt.