Numerička Matematika

Studeni 2010.

Sadržaj

Dio I

Diferencijalne jednadžbe

Poglavlje 1

Numeričko rješavanje inicijalnog problema za ODJ

U mnogim primjenama u prirodnim i inženjerskim znanostima se vremenska promjena stanja promatranog sustava (mehanički sustav, kemijska reakcija) opisuje diferencijalnim jednadžbama koje matematički opisuju zakone (npr. zakoni mehanike) koji

$$\begin{aligned} y'(t) &= f(t, y(t)),\ t \geq t_0 \\ y(t_0) &= y_0 \end{aligned} \tag{1.0.1}$$

Pri tome je $f : \mathcal{I} \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$, gdje je $\mathcal{I} \subseteq \mathbb{R}$ otvoren interval, $t_0 \in \mathcal{I}$, $d \geq 1$ i $y_0 \in \mathbb{R}^d$. Po komponentama problem (1.0.1) glasi

$$\underbrace{\begin{pmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_d'(t) \end{pmatrix}}_{y'(t)} = \underbrace{\begin{pmatrix} f_1(t, y_1(t), y_2(t), \ldots, y_d(t)) \\ f_2(t, y_1(t), y_2(t), \ldots, y_d(t)) \\ \vdots \\ f_d(t, y_1(t), y_2(t), \ldots, y_d(t)) \end{pmatrix}}_{f(t,y(t))}, \quad \begin{pmatrix} y_1(t_0) \\ y_2(t_0) \\ \vdots \\ y_d(t_0) \end{pmatrix} = \begin{pmatrix} (y_0)_1 \\ (y_0)_2 \\ \vdots \\ (y_0)_d \end{pmatrix}.$$

Kažemo da je (1.0.1) inicijalni problem za sustav običnih diferencijalnih jednadžbi. Ako je

$$f(t, x_1, \ldots, x_d) = A(t) \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \begin{pmatrix} b_1(t) \\ \vdots \\ b_d(t) \end{pmatrix}$$

onda kažemo da je sustav ODJ linearan.

1.1 Rješivost

Pod određenim uvjetima problem (1.0.1) ima jedinstveno rješenje. Pokazuje se da je ključni element Lipschitzovo svojstvo u varijabli $x = (x_1, \ldots, x_d)^T$ funkcije $f(t, x)$.

Definicija 1.1.1. Za funkciju $f : \mathcal{I} \times \mathcal{D} \ni (t, x) \longrightarrow f(t, x) \in \mathbb{R}^d$ kažemo da je Lipschitzova u varijabli x ako postoji konstanta $L_f \in (0, \infty)$ takva da za sve $(t, v), (t, w) \in \mathcal{I} \times \mathcal{D}$ vrijedi

$$\|f(t, w) - f(t, v)\| \leq L_f \|w - v\|.$$

Lema 1.1.1. *Neka funkcija $f = (f_1, \ldots, f_d)^T : \mathcal{I} \times \mathcal{D} \longrightarrow \mathbb{R}^d$ ima neprekidne sve parcijalne derivacije $\partial_j f_k \equiv \partial f_k / \partial x_j$, $k, j = 1, \ldots, d$, i neka su one omeđene na $\mathcal{I} \times \mathcal{D}$. Neka je $\mathcal{D}$ konveksan. Tada je f Lipschitzova u varijabli x.*

Dokaz: Neka su $t \in \mathcal{I}$, $z_1, z_2 \in \mathcal{D}$ proizvoljni. Primjenom teorema srednje vrijednosti je za svaku fukciju $f_k : \mathcal{I} \times \mathcal{D} \longrightarrow \mathbb{R}$ (uz oznake $\partial_j = \partial / \partial x_j$)

$$\begin{aligned}
|f_k(t, z_2) - f_k(t, z_1)| &= |\sum_{j=1}^{d} \partial_j f_k(t, z_1 + \xi_*(z_2 - z_1))((z_2)_j - (z_1)_j)| \\
&\leq \|z_2 - z_1\|_\infty \sum_{j=1}^{d} \max_{\xi \in [0,1]} |\partial_j f_k(t, z_1 + \xi(z_2 - z_1))| \\
&\leq d \max_{j=1:d} \sup_{(t,x) \in \mathcal{I} \times \mathcal{D}} |\partial_j f_k(t, x)| \, \|z_2 - z_1\|_\infty \\
&\leq d \, M_f \|z_2 - z_1\|_\infty, \quad M_f \equiv \max_{k=1:d} \max_{j=1:d} \sup_{(t,x) \in \mathcal{I} \times \mathcal{D}} |\partial_j f_k(t, x)| \, .
\end{aligned}$$

Dakle, sve koordinatne funkcije f_k su Lipschitzove pa se lako uvjerimo da f mora imati isto svojstvo. Na primjer, neka je k indeks za kojeg je

$$\|f(t, z_2) - f(t, z_1)\|_\infty = |f_k(t, z_2) - f_k(t, z_1)|.$$

Iz upravo dokazanog je $\|f(t, z_2) - f(t, z_1)\|_\infty \leq d \, M_f \|z_2 - z_1\|_\infty$. $\boxplus$

Slijedi najjednostavnija varijanta teorema o rješivosti problema (1.0.1).

Teorem 1.1.2. *Neka je f definirana i neprekidna na*

$$S = \{(t, x) \ : \ (t, x) \in [a, b] \times \mathbb{R}^d\}, \ a < b < \infty$$

i neka vrijedi Lipschitzov uvjet: postoji $L_f \in (0,\infty)$ *tako da za sve* $t \in [a,b]$ *i* $u, v \in \mathbb{R}^d$ *vrijedi*

$$\|f(t,u) - f(t,v)\| \leq L_f \|u - v\|.$$

Tada za svaki $t_0 \in [a,b]$ *i svaki* $y_0 \in \mathbb{R}^d$ *postoji točno jedna funkcija* y*, derivabilna na* $[a,b]$*, koja zadovoljava* $y'(t) = f(t, y(t))$*,* $t \in [a,b]$*, i* $y(t_0) = y_0$*.*

Primjer 1.1.1. Postojanje rješenja u teoremu 1.1.2 je osigurano ako je f samo neprekidna. Dodatno Lipschitzovo svojstvo osigurava da je rješenje jedinstveno. Za ilustraciju, pogledajmo sljedeći problem sa $f(t,x) = x^{1/3}$:

$$y'(t) = y(t)^{1/3}$$

Ako je zadano i $y(0) = 0$ onda je $y(t) \equiv 0$ očito jedno rješenje, a lako se provjeri da je još jedno rješenje dano s

$$y(t) = \begin{cases} \left(\frac{2}{3}t\right)^{3/2} & \text{za } t \geq 0 \\ 0 & \text{za } t \leq 0 \end{cases}.$$

Uočavamo da f u okolini nule nije Lipschitzova jer npr. za $x > 0$ imamo

$$\left| \frac{f(t,x) - f(t,-x)}{x - (-x)} \right| = \frac{1}{x^{2/3}}$$

i gornji kvocijent se ne može uniformno omeđiti.

U primjenama je važno znati koliko se mijenja rješenje $y(t)$ problema (1.0.1) ako se promijeni početni uvjet $y(t_0) = y_0$.

Teorem 1.1.3. *Neka vrijede pretpostavke Teorema 1.1.2. Ako je* $\tilde{y}(t)$ *rješenje problema* $y'(t) = f(t, y(t))$*,* $y(t_0) = \tilde{y_0}$ *onda za svaki* $t \in [a,b]$ *vrijedi*

$$\|\tilde{y}(t) - y(t)\| \leq \mathbf{e}^{L_f(t-t_0)} \|\tilde{y_0} - y_0\|.$$

1.2 Jednokoračne metode

Riješiti problem (1.0.1) numerički znači izračunati aproksimacije vrijednosti $y(t_i)$ u konačno točaka $t_1, \ldots t_n$ u zadanom intervalu $[t_0, T]$.

1.2.1 Eulerova metoda

Ključni element numeričkog rješavanja inicijalnog problema (1.0.1) je kako iskoristiti diferencijalnu jednadžbu i od poznate vrijednosti $y(t_0)$ dobiti što bolju aproksimaciju za $y(t_1)$ i tako dalje, $t_0 \curvearrowright t_1 \curvearrowright \cdots \curvearrowright t_i \curvearrowright t_{i+1} \curvearrowright \cdots$ sve do zadnje točke t_n.

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(\tau)d\tau = \int_{t_i}^{t_{i+1}} f(\tau, y(\tau))d\tau \tag{1.2.1}$$

Sada uočimo da relaciju

$$y(t_{i+1}) = y(t_i) + \int_{t_i}^{t_{i+1}} f(\tau, y(\tau))d\tau \tag{1.2.2}$$

možemo iskoristiti za aproksimaciju $y(t_{i+1})$ tako što ćemo aproksimirati numeričku vrijednost integrala. Istina, podintegralnu funkciju poznamo samo u lijevom rubu intervala $[t_i, t_{i+1}]$ pa je ovaj čas najjednostavnije što možemo

$$\int_{t_i}^{t_{i+1}} f(\tau, y(\tau))d\tau \approx (t_{i+1} - t_i)f(t_i, y(t_i)), \quad y(t_{i+1}) \approx y(t_i) + (t_{i+1} - t_i)f(t_i, y(t_i)).$$

Geometrijski, iz točke $(t_i, y(t_i))$ se gibamo duž pravca $y(t) = y(t_i)+f(t_i, y(t_i))(t-t_i)$ sve do $t = t_{i+1}$. Naravno, osim u $t = t_0$ kada je $y(t_0)$ poznata zadana vrijednost, sve vrijednosti $y(t_i)$ su zamijenjene aproksimacijama y_i. Diskretne vrijednosti $t_0, t_1, \ldots$ možemo birati sa varijabilnim koracima $h_i = t_{i+1} - t_i$ ili, jednostavnije, ekvidistantno $h = t_{i+1} - t_i$ za sve $i = 0, 1, \ldots$.

$[y] = \text{Euler}(f, y_0, h, n)$	$[y] = \text{Euler}(f, y_0, t, n)$
for $i = 0, \ldots, n-1$ $\quad y_{i+1} = y_i + hf(t_i, y_i)$ *end*	*for* $i = 0, \ldots, n-1$ $\quad y_{i+1} = y_i + (t_{i+1} - t_i)f(t_i, y_i)$ *end*

Eulerova metoda je najjednostavniji primjer *jednokoračne eksplicitne metode* – vrijednost y_{i+1} je dobivena eksplicitnim izrazom koji koristi informaciju samo iz koraka t_i.

Definicija 1.2.1. Jednokoračna eksplicitna metoda za numeričko rješavanje inicijalnog problema (1.0.1) računa niz aproksimacija

$$y_{i+1} = y_i + h_i\phi(t_i, y_i, h_i),\ i = 0, 1, 2, \ldots \tag{1.2.3}$$

u kojem je $y_0 = y(t_0)$ zadana inicijalna vrijednost a $\phi : [t_0, \infty) \times \mathbb{R}^d \times \mathbb{R}_+ \longrightarrow \mathbb{R}^d$ je preslikavanje koje definira metodu.

1.2.2 Konvergencija

Naravno, zanima nas kada možemo garantirati da će metoda (1.2.3) dati konvergentan niz aproksimacija. Pri tome prvo trebamo definirati konvergenciju.

Za početak, važna nam je pogreška jednog koraka: ako metodi na intervalu $[t_i, t_{i+1}] \equiv [t, t + \delta t]$ damo egzaktnu vrijednost rješenja u lijevom rubu, koliko je odstupanje aproksimacije u t_{i+1}?

Definicija 1.2.2. Lokalna pogreška diskretizacije metode (1.2.3) u točki $t + h$ je definirana s

$$\boldsymbol{\epsilon}(t, h) = y(t + h) - y(t) - h\phi(t, y(t), h). \tag{1.2.4}$$

Primijetimo da (1.2.3) daje $y_{i+1} - y_i - h_i\phi(t_i, y_i, h_i) = 0$, pa lokalna pogreška diskretizacije mjeri kako dobro egzaktno rješenje zadovoljava rekurziju jednokoračne metode (1.2.3).

$$\begin{aligned} 0 &= \frac{y_{i+1} - y_i}{h_i} - \phi(t_i, y_i, h_i) \\ \frac{\boldsymbol{\epsilon}(t, h)}{h} &= \underbrace{\frac{y(t + h) - y(t)}{h}}_{\longrightarrow y'(t)(h \to 0)} - \phi(t, y(t), h). \end{aligned}$$

Odavde odmah izvodimo nužan uvjet da bi metoda imala smisla – metoda treba biti konzistentna u smislu sljedeće definicije:

Definicija 1.2.3. Metoda (1.2.3) je konzistentna ako za sve $t \in [t_0, T]$ vrijedi

$$\lim_{h \to 0} \phi(t, y(t), h) = f(t, y(t)).$$

Kažemo da je metoda reda konzistentnosti p ako je za sve $t, t + h \in [t_0, T]$

$$\|\boldsymbol{\epsilon}(t, h)\| \leq Ch^{p+1},$$

gdje je C konstanta neovisna o t i h.

Propozicija 1.2.1. *Ako $f : [t_0, T]\times\mathbb{R}^d \longrightarrow \mathbb{R}^d$ ima neprekidne parcijalne derivacije, onda Eulerova metoda ima red konzistentnosti $p = 1$.*

Dokaz: Naime, koristeći Taylorov razvoj

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}\begin{pmatrix} y_1''(\xi_1) \\ \vdots \\ y_d''(\xi_d) \end{pmatrix}$$

lako dobijemo

$$y(t+h) - y(t) - hf(t, y(t)) = \frac{h^2}{2}\begin{pmatrix} y_1''(\xi_1) \\ \vdots \\ y_d''(\xi_d) \end{pmatrix}$$

pa je $\|\boldsymbol{\epsilon}(t,h)\|_\infty \leq Ch^2$ sa $C = \max_{\tau\in[t_0,T]} \|y''(\tau)\|_\infty$. $\boxplus$

Definicija 1.2.4. *Globalna pogreška diskretizacije* metode (1.2.3) u točki t_i je definirana s

$$\boldsymbol{\gamma}_i = y_i - y(t_i). \tag{1.2.5}$$

Kažemo da je metoda *reda konvergencije* p ako postoji konstanta K tako da je

$$\max_{i=0:n} \|\boldsymbol{\gamma}_i\| \leq K\overline{h}^p,\ \overline{h} = \max_{i=0:n-1}(t_{i+1} - t_i).$$

Dokazi konvergencije su bazirani na dva ključna elementa: redu konzistentnosti i Lipschitzovom svojstvu fukncije ϕ koja generira metodu. Pri tome su, naravno, teoremi koji pretpostavljaju jače uvjete jednostavniji za dokazati od onih koji dozvoljavaju i slabije polazne pretpostavke.

Teorem 1.2.2. *Neka jednokoračna metoda (1.2.3) ima red konzistentnosti $p \geq 1$, i neka je funkcija ϕ Lipschitzova u drugoj varijabli: postoji konstanta $L_\phi \in (0, \infty)$ tako da je za sve $t \in [t_0, T]$, $h > 0$ i $z_1, z_2 \in \mathbb{R}^d$*

$$\|\phi(t, z_1, h) - \phi(t, z_2, h)\| \leq L_\phi\|z_1 - z_2\|.$$

Tada vrijedi

$$\max_{i=0:n} \|y_i - y(t_i)\| \leq \frac{C}{L_\phi}(\mathbf{e}^{L_\phi(T-t_0)} - 1)\left(\max_{i=0:n-1}(t_{i+1} - t_i)\right)^p \tag{1.2.6}$$

Dokaz: Za globalnu pogrešku $\boldsymbol{\gamma}_i = y_i - y(t_i)$, $i = 0, \ldots, n$, oduzimanjem relacija

$$\begin{aligned} y_{i+1} &= y_i + h_i\phi(t_i, y_i, h_i) \\ y(t_{i+1}) &= y(t_i) + h_i\phi(t_i, y(t_i), h_i) + \boldsymbol{\epsilon}(t_i, h_i) \end{aligned}$$

dobijemo

$$\begin{aligned} \boldsymbol{\gamma}_{i+1} &= \boldsymbol{\gamma}_i + h_i(\phi(t_i, y_i, h_i) - \phi(t_i, y(t_i), h_i)) - \boldsymbol{\epsilon}(t_i, h_i) \\ \|\boldsymbol{\gamma}_{i+1}\| &\leq \|\boldsymbol{\gamma}_i\| + h_i\|\phi(t_i, y_i, h_i) - \phi(t_i, y(t_i), h_i)\| + \|\boldsymbol{\epsilon}(t_i, h_i)\| \\ &\leq \|\boldsymbol{\gamma}_i\| + h_i L_\phi\|\boldsymbol{\gamma}_i\| + \|\boldsymbol{\epsilon}(t_i, h_i)\| = (1 + L_\phi h_i)\|\boldsymbol{\gamma}_i\| + \|\boldsymbol{\epsilon}(t_i, h_i)\| \end{aligned}$$

Sada stavimo $\overline{h} = \max_i h_i$, $\overline{\boldsymbol{\epsilon}} = \max_i \|\boldsymbol{\epsilon}(t_i, h_i)\|$. Pokazali smo da vrijedi

$$\|\boldsymbol{\gamma}_{i+1}\| \leq (1 + L_\phi h_i)\|\boldsymbol{\gamma}_i\| + h_i C\overline{h}^p, \ i = 0, \ldots, n-1.$$

Odavde se može zaključiti (tehnička lema) da je

$$\|\boldsymbol{\gamma}_i\| \leq \frac{C}{L_\phi}(\mathbf{e}^{L_\phi \sum_{j=0}^{i-1} h_j} - 1)\overline{h} + \|\boldsymbol{\gamma}_0\|\mathbf{e}^{L_\phi \sum_{j=0}^{i-1} h_j}.$$

⊞

Korolar 1.2.3. *Neka vrijede uvjeti iz Teorema 1.2.2 i neka je $t \in (t_0, T]$. Neka je $h_n = (t - t_0)/n$ i neka y_{i,h_n}, $i = 1, \ldots, n$, označava aproksimacije dobivene s fiksnim korakom h_n. Tada je*

$$\lim_{n\to\infty} \|y_{n,h_n} - y(t)\| = 0.$$

Korolar 1.2.4. *Ako $f : [t_0, T] \times \mathbb{R}^d \longrightarrow \mathbb{R}^d$ ima neprekidne i omeđene parcijalne derivacije, onda Eulerova metoda ima red konvergencije $p = 1$.*

Primjer 1.2.1. (*Poboljšanje Eulerove metode.*) Eulerova metoda je jednostavan proces i, kako je i za očekivati, ima mali red konvergencije. Zanimljivo je vidjeti kako čak i tako jednostavnu metodu možemo jednostavnim trikom popraviti. Ako je rješenje $y(t)$ glatka funkcija, onda pogrešku u i–tom koraku možemo analitički zapisati u obliku

$$y_i = y(t_i) + \eta_1 h + \eta_2 h^2 + \cdots \tag{1.2.7}$$

pri čemu koeficijenti $\eta_1, \eta_2, \ldots$ ne ovise o h. Ako sada Eulerovom metodom do momenta t_i dođemo prvo korakom h a zatim korakom $\tilde{h} < h$ (npr. $\tilde{h} = h/2$) dobijemo dvije aproksimacije (s korakom $\tilde{h}$ do t_i stižemo u $\tilde{i}$ koraka)

$$\begin{aligned} y_{i,h} &= y(t_i) + \eta_1 h + \eta_2 h^2 + \cdots \\ y_{\tilde{i},\tilde{h}} &= y(t_i) + \eta_1 \tilde{h} + \eta_2 \tilde{h}^2 + \cdots \end{aligned}$$

Sada od dvije dobivene aproksimacije $y_{i,h}$ i $y_{\tilde{i},\tilde{h}}$ pokušajmo složiti linearnu kombinaciju $\alpha y_{i,h} + \beta y_{\tilde{i},\tilde{h}}$ koja će biti bolja od obe. Prirodno je zahtijevati $\alpha + \beta = 1$. Lako izračunamo da je

$$\alpha y_{i,h} + \beta y_{\tilde{i},\tilde{h}} = y(t_i) + \eta_1(\alpha h + \beta\tilde{h}) + \eta_2(\alpha h^2 + \beta\tilde{h}^2) + O(h^3) + O(\tilde{h}^3)$$

i odmah uočavamo da odabir α i β tako da je još i $\alpha h + \beta\tilde{h} = 0$ daje

$$\alpha y_{i,h} + \beta y_{\tilde{i},\tilde{h}} = y(t_i) + O(h^2).$$

Ako je na primjer $\tilde{h} = h/2$ onda je $\alpha = -1$, $\beta = 2$ i imamo novu aproksimaciju

$$y_i = 2y_{2i,h/2} - y_{i,h}.$$

Praktična izvedba ove nove sheme se sastoji u tome da se dva Eulerova procesa kombiniraju odmah *u hodu*.

$$\begin{aligned} y_{i+1,h} &= y_i + hf(t_i, y_i) \\ y_{i+1/2,h/2} &= y_i + \frac{h}{2} f(t_i, y_i) \\ y_{i+1,h/2} &= y_{i+1/2,h/2} + \frac{h}{2} f(t_i + \frac{h}{2}, y_{i+1/2,h/2}) \\ y_{i+1} &= 2y_{i+1,h/2} - y_{i+1,h} \end{aligned}$$

Algoritam je jednostavan:

$[y] = \text{Euler2}(f, y_0, h, n)$
for $i = 0, \ldots, n$ $\quad y_{i+1} = y_i + hf(t_i + h/2, y_i + (h/2)f(t_i, y_i))$ *end*

Sada ćemo se uvjeriti da je lokalna pogreška diskretizacije poboljšane Eulerove metode za red veličine manja (u odnosu na klasičnu Eulerovu metodu). Računamo

$$\begin{aligned} d_{i+1} &= y(t_{i+1}) - y(t_i) - hf(t_i + h/2, y(t_i) + (h/2)f(t_i, y(t_i))) \\ &= y'(t_i)h + \frac{y''(t_i)}{2}h^2 + \frac{y'''(t_i)}{6}h^3 + \end{aligned}$$

1.2.3 Trapezna metoda

Ideja trapezne formule je jednostavna: U relaciji (**??**) treba integral numerički aproksimirati trapeznom formulom da dobijemo

$$y(t_{i+1}) \approx y(t_i) + \frac{t_{i+1} - t_i}{2}(f(t_i, y(t_i) + f(t_{i+1}, y(t_{i+1})))) \tag{1.2.8}$$

Trapezna metoda je najjednostavniji primjer *jednokoračne implicitne metode* – vrijednost $y_{i+1} \approx y(t_{i+1})$ se dobije koristeći samo informaciju iz koraka t_i i pri tome je y_{i+1} definiran implicitno kao rješenje jednadžbe.

1.3 Runge–Kuttine metode

Željeli bismo poboljšati jednostavne jednokoračne metode. Polazimo od osnovne relacije

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(\tau, y(\tau))d\tau,$$

i u njoj pokušajmo numeričku vrijednost integrala aproksimirati koristeći dodatne čvorove u $[t_k, t_{k+1}]$ i dodatne slobodne parametre koje ćemo naknadno fino naštimati. Konkretno, uzmimo na primjer tri čvora ξ_1, ξ_2, ξ_3, tri dodatna parametra c_1, c_2, c_3 i potražimo formulu oblika

$$y_{k+1} = y_k + h[c_1 f(\xi_1, y(\xi_1)) + c_2 f(\xi_2, y(\xi_2)) + c_3 f(\xi_3, y(\xi_3))], \tag{1.3.1}$$

pri čemu ćemo slobodne parametre iskoristiti da dobijemo što veći red konzistentnosti, tj. što je moguće manju lokalnu pogrešku diskretizacije.

Odaberimo $\xi_1 = t_k$, $\xi_2 = t_k + a_2 h$, $\xi_3 = t_k + a_3 h$ sa $0 < a_2, a_3 < 1$. Kako nam vrijednosti $y(\xi_i)$ nisu dostupne, zamijenit ćemo ih prediktorima – aproksimacijama, ali opet sa slobodnim parametrima A_{21}, A_{31}, A_{32}:

$$\begin{aligned} &y(\xi_1) \approx y_k \\ &y(\xi_2) \approx y_k + hA_{21} f(t_k, y_k) \\ &y(\xi_3) \approx y_k + hA_{31} f(t_k, y_k) + hA_{32} f(t_k + a_2 h, y_k + hA_{21} f(t_k, y_k)) \end{aligned} \tag{1.3.2}$$

Odavde dobivamo opći oblik k–tog koraka:

$$\begin{aligned} \Psi_1 &= f(t_k, y_k) \\ \Psi_2 &= f(t_k + a_2 h, y_k + hA_{21}\Psi_1) \\ \Psi_3 &= f(t_k + a_3 h, y_k + h(A_{31}\Psi_1 + A_{32}\Psi_2)) \\ y_{k+1} &= y_k + h(c_1\Psi_1 + c_2\Psi_2 + c_3\Psi_3) \end{aligned}$$

Sada osam slobodnih parametara treba odrediti tako da dobivena shema bude što je moguće bolja. Za početak, osigurajmo npr. da prediktori (1.3.2) u nekim jednostavnim primjerima rade perfektno. Recimo, ako rješavamo $y'(t) = 1$ na $(0, \infty)$ sa $y(0) = 0$, onda je rješenje dano s $y(t) = t$. Ako zahtijevamo da su u ovom jednostavnom primjeru[1] prediktori egzaktni, onda dobivamo uvjete

$$\begin{aligned} a_2 &= A_{21} \\ a_3 &= A_{31} + A_{32} \end{aligned}$$

Ostale uvjete na slobodne koeficijente ćemo izvesti iz oblika lokalne pogreške diskretizacije

$$\begin{aligned} d_{k+1} &= y(t_{k+1}) - [y(t_k) + h(c_1\Phi_1 + c_2\Phi_2 + c_3\Phi_3)], \\ \text{gdje je:} \qquad \Phi_1 &\equiv f(t_k, y(t_k)), \\ \Phi_2 &\equiv f(t_k + a_2 h, y(t_k) + A_{21} h \Phi_1), \\ \Phi_3 &\equiv f(t_k + a_3 h, y(t_k) + h(A_{31}\Phi_1 + A_{32}\Phi_2)). \end{aligned}$$

Sada sve Φ_i razvijemo u Taylorov red oko $(t_k, y(t_k))$. Dobijemo

$$\begin{aligned} \Phi_2 &= f(t_k, y(t_k)) + \partial_t f(t_k, y(t_k)) \cdot a_2 h + \partial_y f(t_k, y(t_k)) \cdot a_2 h f(t_k, y(t_k)) \\ &+ \frac{1}{2}\partial^2_{tt} f(t_k, y(t_k)) \cdot a_2^2 h^2 + \partial^2_{ty}(t_k, y(t_k)) \cdot a_2^2 h^2 f(t_k, y(t_k)) \\ &+ \frac{1}{2}\partial^2_{yy} f(t_k, y(t_k)) a_2^2 h^2 f^2(t_k, y(t_k)) + O(h^3) \\ \Phi_3 &= f(t_k, y(t_k)) + a_3 h F(t_k, y(t_k)) + h^2(a_2 A_{32} F(t_k, y(t_k)) \partial_y f(t_k, y(t_k)) \\ &+ \frac{1}{2} a_3^2 G(t_k, y(t_k))) + O(h^3), \quad \text{gdje je} \\ F(t_k, y(t_k)) &= \partial_t f(t_k, y(t_k)) + \partial_y f(t_k, y(t_k)) f(t_k, y(t_k)) \\ G(t_k, y(t_k)) &= \partial^2_{tt} f(t_k, y(t_k)) + 2\partial^2_{ty} f(t_k, y(t_k)) f(t_k, y(t_k)) \\ &+ \partial^2_{yy} f(t_k, y(t_k)) f^2(t_k, y(t_k)) \end{aligned}$$

Također, imamo razvoj za $y(t_{k+1})$ oko t_k, pri čemu je

$$\begin{aligned} y'(t_k) &= f(t_k, y(t_k)) \\ y''(t_k) &= F(t_k, y(t_k)) \\ y'''(t_k) &= G(t_k, y(t_k)) + F(t_k, y(t_k)) f(t_k, y(t_k)). \end{aligned}$$

[1] Razumno je zahtijevati da se metoda koju pokušavamo konstruirati dobro ponaša u jednostavnim slučajevima.

Ako sada sve navedene razvoje iskoristimo u izrazu za d_{k+1} i uredimo ga po potencijama od h dobijemo

$$\begin{aligned} d_{k+1} &= hf(t_k, y(t_k))(1 - c_1 - c_2 - c_3) + h^2 F(t_k, y(t_k))(\frac{1}{2} - a_2c_2 - a_3c_3) \\ &+ h^3 \Big[F(t_k, y(t_k))\partial_y f(t_k, y(t_k))(\frac{1}{6} - a_2c_3A_{32}) \\ &+ G(t_k, y(t_k))(\frac{1}{6} - \frac{1}{2}a_2^2c_2 - \frac{1}{2}a_3^2c_3)\Big] + O(h^4). \end{aligned}$$

Vidimo da možemo postići $d_{k+1} = O(h^4)$ ako koeficijenti zadovoljavaju sljedeći sustav jednadžbi:

$$\begin{aligned} c_1 + c_2 + c_3 &= 1 \\ a_2c_2 + a_3c_3 &= \tfrac{1}{2} \\ a_2A_{32}c_3 &= \tfrac{1}{6} \\ a_2^2c_2 + a_3^2c_3 &= \tfrac{1}{3} \end{aligned}$$

Šest parametara, četiri jednadžbe. Uzmimo da su $a_2, a_3 \in (0,1)$ slobodni parametri i riješimo uz uvjete $a_2 \neq a_3$, $a_2 \neq 2/3$:

$$\begin{aligned} c_1 &= \frac{6a_2a_3 + 2 - 3(a_2 + a_3)}{6a_2a_3},\ c_2 = \frac{3a_3 - 2}{6a_2(a_3 - a_2)},\ c_3 = \frac{2 - 3a_2}{6a_3(a_3 - a_2)}, \\ A_{21} &= a_2,\ A_{32} = \frac{a_3(a_3 - a_2)}{a_2(2 - 3a_2)},\ A_{31} = a_3 - A_{32}. \end{aligned}$$

Rješenje kompaktno zapisujemo u obliku

$$\begin{array}{c|ccc} a_1 & & & \\ a_2 & A_{21} & & \\ a_3 & A_{31} & A_{32} & \\ \hline & c_1 & c_2 & c_3 \end{array}.$$

Tako na primjer imamo Kuttinu i Heunovu metodu reda 3, koje su dane, redom,

Kutta:
$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

Heun:
$$\begin{array}{c|ccc} 0 & & & \\ 1/3 & 1/3 & & \\ 1 & 0 & 2/3 & \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

Popularna je Runge–Kuttina metoda stupnja i reda 4:

$$\begin{array}{r|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \tag{1.3.3}$$

Nekoliko komentara:

- Prvo što uočavamo je da je procedura izvođenja formula komplicirana. Za metodu još višeg stupnja su izrazi s parcijalnim derivacijama još kompliciraniji.
- Povećavanjem stupnja metode, tj. uvođenjem više slobodnih parametara postižemo veći red metode. Na žalost, stupanj 4 je najveći za kojeg je metoda reda 4. Maksimalni red za stupanj 5 je 4, za stupanj 6 je maksimalni red metode 5, za stupnjeve 7 i 8 je maksimalni red jednak 6, dok metoda stupnja 9 ima maksimalni red jednak 7.
- Metoda je nelinearna i u svakom koraku treba nekoliko puta računati vrijednosti funckije f.

1.4 Višekoračne metode

Kada pogledamo npr. Eulerovu metodu koja počevši od zadane vrijednosti $y(t_0) = y_0$ generira niz aproksimacija u čvorovima $t_1, t_2, \ldots, t_{k-3}, t_{k-2}, t_{k-1}, t_k$ onda nam se čini razumnim da pri računanju aproksimacije u t_{k+1} uzmemo u obzir ne samo t_k, y_k nego i dio *iz prošlosti*, tj. vrijednosti $y_{k-1}, y_{k-2}, \ldots$ iz nekoliko prethodnik koraka. Kako možemo iskoristiti tu informaciju?

Rezultat su takozvane višekoračne metode. Primjere takvih metoda ćemo konstruirati koristeći jednostavnu heuristiku baziranu na interpolacijskom polinomu.

1.4.1 Adams–Bashforth–ove metode

Prisjetimo se, u jednom koraku metoda koje smo do sada proučavali koristimo relaciju

$$y(t_{k+1}) = y(t_k) + \int_{t_k}^{t_{k+1}} f(\tau, y(\tau))d\tau,$$

i u njoj na različite načine numerički računamo integral. Sada pogledamo informaciju nekoliko koraka unazad, recimo u vremenima t_{k-3}, t_{k-2}, t_{k-1}, i zajedno sa informacijom u vremenu t_k pokušajmo dobiti bolju procjenu integrala.

Konstruirajmo Lagrangeov interpolacijski polinom

$$P_3(t) = \sum_{j=0}^{3} f(t_{k-j}, y_{k-j}) L_{k-j}(t), \quad L_{k-j}(t) = \frac{\prod_{i \neq k-j}(t - t_i)}{\prod_{i \neq k-j}(t_{k-j} - t_i)}$$

i iskoristimo ga za aproksimaciju podintegralne funkcije na $[t_k, t_{k+1}]$. Time smo definirali metodu

$$y_{k+1} = y_k + \sum_{j=0}^{3} f(t_{k-j}, y_{k-j}) \underbrace{\int_{t_k}^{t_{k+1}} L_{k-j}(\tau) d\tau}_{\ell_j}. \tag{1.4.1}$$

Ostaje izračunati koeficijente ℓ_j. Inegrale ćemo računati zamjenom varijabli $t = t_k + sh$, $dt = hds$, gdje je $s \in [0, 1]$. Idemo redom:

$$\begin{aligned}
\ell_0 &= \int_{t_k}^{t_{k+1}} L_k(\tau) d\tau = \int_{t_k}^{t_{k+1}} \frac{(t - t_{k-3})(t - t_{k-2})(t - t_{k-1})}{(t_k - t_{k-3})(t_k - t_{k-2})(t_k - t_{k-1})} dt \\
&= h \int_0^1 \frac{h \cdot (3+s) \cdot h \cdot (2+s) \cdot h \cdot (1+s)}{3 \cdot h \cdot 2 \cdot h \cdot 1 \cdot h} ds \\
&= \frac{h}{6} \int_0^1 (s^3 + 6s^2 + 11s + 6) ds = \frac{55}{24} h;
\end{aligned}$$

$$\begin{aligned}
\ell_1 &= \int_{t_k}^{t_{k+1}} L_{k-1}(\tau) d\tau = \int_{t_k}^{t_{k+1}} \frac{(t - t_{k-3})(t - t_{k-2})(t - t_k)}{(t_{k-1} - t_{k-3})(t_{k-1} - t_{k-2})(t_{k-1} - t_k)} dt \\
&= h \int_0^1 \frac{h \cdot (3+s) \cdot h \cdot (2+s) \cdot h \cdot s}{2 \cdot h \cdot 1 \cdot h \cdot (-1) \cdot h} ds \\
&= -\frac{h}{2} \int_0^1 (s^3 + 5s^2 + 6s) ds = -\frac{59}{24} h;
\end{aligned}$$

$$\begin{aligned}
\ell_2 &= \int_{t_k}^{t_{k+1}} L_{k-2}(\tau)d\tau = \int_{t_k}^{t_{k+1}} \frac{(t-t_{k-3})(t-t_{k-1})(t-t_k)}{(t_{k-2}-t_{k-3})(t_{k-1}-t_{k-1})(t_{k-2}-t_k)}dt \\
&= h\int_0^1 \frac{h\cdot(3+s)\cdot h\cdot(1+s)\cdot h\cdot s}{1\cdot h\cdot(-1)\cdot h\cdot(-2)\cdot h}ds \\
&= \frac{h}{2}\int_0^1 (s^3+4s^2+3s)ds = \frac{37}{24}h;
\end{aligned}$$

$$\begin{aligned}
\ell_3 &= \int_{t_k}^{t_{k+1}} L_{k-3}(\tau)d\tau = \int_{t_k}^{t_{k+1}} \frac{(t-t_{k-2})(t-t_{k-1})(t-t_k)}{(t_{k-3}-t_{k-2})(t_{k-3}-t_{k-1})(t_{k-3}-t_k)}dt \\
&= h\int_0^1 \frac{h\cdot(2+s)\cdot h\cdot(1+s)\cdot h\cdot s}{(-1)\cdot h\cdot(-2)\cdot h\cdot(-3)\cdot h}ds \\
&= -\frac{h}{6}\int_0^1 (s^3+3s^2+2s)ds = -\frac{9}{24}h.
\end{aligned}$$

Iako su gornji integrali elementarni, dali smo detalje računa s ciljem da se uoči da u tim formulama ima dosta strukture i da bi se u principu mogle očekivati zatvorene formule za računanje koeficijenata ℓ_i, bez da idemo kroz sve korake kao u gornjem primjeru. To je uistinu moguće, ali za sada nećemo ulaziti u te detalje.

Algoritam 1.4.1. Adams–Bashforthova eksplicitna metoda reda 4.

$[y] = \text{Adams\_Bashforth\_4}(f, y_{0:3}, f_{0:3}, t, n)$
for $i = 3, \ldots, n-2$
$\quad y_{i+1} = y_i + \frac{h}{24}(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3})$
$\quad f_{i+1} = f(t_{i+1}, y_{i+1})$
end
$y_n = y_{n-1} + \frac{h}{24}(55f_{n-1} - 59f_{n-2} + 37f_{n-3} - 9f_{n-4})$

Ako želimo odmah dobiti osjećaj koliko je ovakva shema dobra, pogledajmo

lokalnu pogrešku diskretizacije:

$$\begin{aligned}
d_{k+1} &= y(t_{k+1}) - y(t_k) - \frac{h}{24}\{55f(t_k, y(t_k)) - 59f(t_{k-1}, y(t_{k-1})) \\
&+ 37f(t_{k-2}, y(t_{k-2})) - 9f(t_{k-3}, y(t_{k-3}))\} = \boxed{\text{jer je } f(t, y(t)) = y'(t)} \\
&= y(t_{k+1}) - y(t_k) - \frac{h}{24}\{55y'(t_k) - 59y'(t_{k-1}) + 37y'(t_{k-2}) - 9y'(t_{k-3})\} \\
&= \boxed{\text{Taylorov razvoj oko } t_k} = hy'(t_k) + \frac{1}{2}h^2y''(t_k) + \frac{1}{6}h^3y'''(t_k) \\
&+ \frac{1}{24}h^4y^{(4)}(t_k) + \frac{1}{120}h^5y^{(5)}(t_k) + O(h^6) - \frac{h}{24}\{55y'(t_k) \\
&- 59[y'(t_k) - hy''(t_k) + \frac{1}{2}h^2y'''(t_k) - \frac{1}{6}h^3y^{(4)}(t_k) + \frac{1}{24}h^4y^{(5)}(t_k) + O(h^5)] \\
&+ 37[y'(t_k) - 2hy''(t_k) + 2h^2y'''(t_k) - \frac{4}{3}h^3y^{(4)}(t_k) + \frac{2}{3}h^4y^{(5)}(t_k) + O(h^5)] \\
&- 9[y'(t_k) - 3hy''(t_k) + \frac{9}{2}h^2y'''(t_k) - \frac{9}{3}h^3y^{(4)}(t_k) + \frac{27}{8}h^4y^{(5)}(t_k) + O(h^5)] \\
&= \frac{251}{720}h^5 + O(h^6).
\end{aligned}$$

Slično, koristeći pet uzastopnih koraka dobijemo:

Algoritam 1.4.2. Adams–Bashforthova eksplicitna metoda reda 5.

$[y] = \text{Adams\_Bashforth\_5}(f, y_{0:4}, f_{0:4}, t, n)$

for $i = 4, \ldots, n-2$
$\quad y_{i+1} = y_i + \frac{h}{720}(1901f_i - 2774f_{i-1} + 2616f_{i-2} - 1274f_{i-3} + 251f_{i-4})$
$\quad f_{i+1} = f(t_{i+1}, y_{i+1})$
end
$y_n = y_{n-1} + \frac{h}{720}(1901f_{n-1} - 2774f_{n-2} + 2616f_{n-3} - 1274f_{n-4} + 251f_{n-5})$

Sažetak:

- Nova aproksimacija se konstruira koristeći informaciju iz prethodnih m koraka. Kažemo da je metoda m–koračna.
- Za početak trebamo m vrijednosti y_0, y_1, y_2, …, y_{m-1}, te m primjena funkcije f za računanje vrijednosti $f_0, f_1, f_2, \ldots, f_{m-1}$.
- U svakom koraku trebamo samo jednu novu primjenu funkcije f.

- Lokalna pogreška diskretizacije npr. reda $O(h^5)$ za $m = 4$ je bitno poboljšanje u odnosu na npr. Eulerovu metodu.

Prije nego krenemo na sustavnu analizu višekoračnih metoda, pogledajmo još jedan primjer: Adams–Moultonove metode.

1.4.2 Adams–Moulton–ove metode

Adams–Moultonove metode polaze od iste ideje kao i Adams–Bashfortove, ali se ovdje u čvorove inteporlacije uključuje i čvor t_{i+1}. Što smo time dobili? Ukupno se gleda isti broj prethodnih koraka, ali s jednim čvorem više možemo konstruirati interpolacijski polinom višeg stupnja. Nadalje, interval nad kojim ćemo integrirati pomoću interpolacijskog polinoma je uključen u interval nad kojim interpoliramo.

Dakle, imamo

$$\begin{aligned} P_3(t) &= \sum_{j=-1}^{3} f(t_{k-j}, y_{k-j}) L_{k-j}(t), \quad L_{k-j}(t) = \frac{\prod_{i \neq k-j}(t - t_i)}{\prod_{i \neq k-j}(t_{k-j} - t_i)} \\ y_{k+1} &= y_k + \sum_{j=-1}^{3} f(t_{k-j}, y_{k-j}) \underbrace{\int_{t_k}^{t_{k+1}} L_{k-j}(\tau) d\tau}_{\ell_j} . \end{aligned}$$

Uočavamo da se na desnoj strani također pojavljuje nepoznata vrijednost y_{k+1}, što znači da je jedan korak metode zadan kao rješenje jednadžbe.

Ako izračunamo koeficijente, dobijemo

$[y] = \text{Adams\_Moulton4}(f, y_{0:3}, f_{0:3}, t, n)$

for $k = 3, \ldots, n-2$
  $y_{k+1} \coloneqq y_k + \frac{h}{720}(251 f(t_{k+1}, y_{k+1}) + 646 f_k - 264 f_{k-1} + 106 f_{k-2} - 19 f_{k-3})$
  $f_{k+1} = f(t_{k+1}, y_{k+1})$
end
$y_n = y_{n-1} + \frac{h}{720}(251 f(t_n, y_n) + 646 f_{n-1} - 264 f_{n-2} + 106 f_{n-3} - 19 f_{n-4})$

Vrijede svi komentari kao kod Adams–Bashfortovih metoda, plus jedan novi: U svakom koraku metode trebamo riješiti jednadžbu čije rješenje je nova vrijednost u Adams–Moultovoj metodi. Odmah se, naravno, postavlja pitanje rješivosti: da li je u svakom koraku y_{k+1} dobro definiran i kako ga efikasno izračunati?

1.4.3 Linearne diferencijske jednadžbe

Višekoračna metoda sa m koraka generira niz diskretnih vrijednosti koje, uz odgovarajućih m početnih vrijednosti, zadovoljavaju jednadžbe oblika

$$\Phi_k(y_k, y_{k+1}, \ldots, y_{k+m}) = 0, \quad k = k_0, k_0+1, k_0+2, \ldots. \tag{1.4.2}$$

Gornja relacija definira diferencijsku jednadžbu reda m. Sa $\{y_k\}$ ćemo označavati rješenje od (1.4.2) tj. niz $y_k, y_{k+1}, \ldots$, koji za svaki $k \geq k_0$ zadovoljava (1.4.2). Nas će posebno zanimati linearne diferencijske jednadžbe

$$\alpha_m^{(k)} y_{k+m} + \alpha_{m-1}^{(k)} y_{k+m-1} + \cdots + \alpha_1^{(k)} y_{k+1} + \alpha_0^{(k)} y_k = \gamma_{k+m}, \ \alpha_m^{(k)} \neq 0. \tag{1.4.3}$$

Uz svaku takvu jednadžbu vežemo njenu pripadnu homogenu jednadžbu

$$\alpha_m^{(k)} y_{k+m} + \alpha_{m-1}^{(k)} y_{k+m-1} + \cdots + \alpha_1^{(k)} y_{k+1} + \alpha_0^{(k)} y_k = 0, \tag{1.4.4}$$

koja ima važnu ulogu u opisu skupa rješenja od (1.4.3).

Za očekivati je da ćemo do korisne informacije o rješenju doći oponašanjem rješavanja sustava linearnih algebarskih jednadžbi i običnih diferencijalnih jednadžbi. Odmah vidimo da je trivijalni nul–niz $\{0\}$ jedino rješenje homogene jednadžbe (1.4.4) sa zadanim $y_0 = \cdots = y_{m-1} = 0$. Očito je da je sa svakim zadanim $y_0, \ldots, y_{m-1}$ jedinstveno zadano rješenje $\{y_k\}$.

Neka su sada $\{y_k^{(1)}\}, \ldots, \{y_k^{(\ell)}\}$ nekih ℓ rješenja od (1.4.4). Kažemo da su $\{y_k^{(i)}\}$, $i = 1, \ldots, \ell$, linearno nezavisni u k_0 ako su linearno nezavisni retci matrice

$$Y_{k_0} = \begin{pmatrix} y_{k_0}^{(1)} & y_{k_0+1}^{(1)} & \cdots & y_{k_0+m-1}^{(1)} \\ y_{k_0}^{(2)} & y_{k_0+1}^{(2)} & \cdots & y_{k_0+m-1}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k_0}^{(\ell)} & y_{k_0+1}^{(\ell)} & \cdots & y_{k_0+m-1}^{(\ell)} \end{pmatrix},$$

tj. ako $(\eta_1, \ldots, \eta_\ell) Y_{k_0} = \mathbf{0} \iff \eta_1 = \cdots = \eta_\ell = 0$. Odavde odmah slijedi da je maksimalni broj linearno nezavisnih rješenja u bilo kojoj točki $k = k_0$ jednak m.

Definicija 1.4.1. Fundamentalni sistem rješenja jednadžbe (1.4.4) u točki $k = k_0$ je svaki skup od m linearno nezavisnih rješenja u točki $k = k_0$.

Teorem 1.4.1. *Svaki niz $\{y_k\}$ koji je za $k \geq k_0$ rješenje linearne homogene diferencijske jednadžbe (1.4.4) možemo napisati kao linearnu kombinaciju rješenja iz fundamentalnog sistema u $k = k_0$.*

Posebno nas zanimaju fundamentalni sistemi u $k = 0$ linearne homogene diferencijske jednadžbe sa konstantnim koeficijentima

$$\alpha_m y_{k+m} + \alpha_{m-1} y_{k+m-1} + \cdots + \alpha_1 y_{k+1} + \alpha_0 y_k = 0, \;\; \alpha_0 \neq 0, \; \alpha_m \neq 0. \qquad (1.4.5)$$

Ako rješenje potražimo u obliku $y_n = \mathbf{e}^{\lambda n} \equiv \zeta^n$, $\zeta = \mathbf{e}^{\lambda} \neq 0$, onda uvrštavanjem u (1.4.5) dobijemo

$$\zeta^n(\alpha_m \zeta^m + \alpha_{m-1}\zeta^{m-1} + \cdots + \alpha_1 \zeta + \alpha_0) = 0.$$

Dakle, $\{\zeta^n\}$ je rješenje od (1.4.5) ako i samo ako je ζ nultočka polinoma

$$\rho(\zeta) = \alpha_m \zeta^m + \alpha_{m-1}\zeta^{m-1} + \cdots + \alpha_1 \zeta + \alpha_0. \qquad (1.4.6)$$

Lema 1.4.2. *Neka polinom $\rho(\zeta)$ ima m jednostrukih nultočki $\zeta_1, \ldots, \zeta_m$. Tada $\{\zeta_1^k\}, \ldots, \{\zeta_m^k\}$ čine fundamentalni sistem rješenja u $k = 0$.*

Dokaz: Zbog $\alpha_0 \neq 0$ je $\rho(0) \neq 0$, tj. niti jedno rješenje $\{\zeta_i^k\}$ nije trivijalno. Linearnu nezavisnost u $k = 0$ vidimo iz činjenice da je determinanta matrice

$$\begin{pmatrix} 1 & \zeta_1 & \cdots & \zeta_1^{m-1} \\ 1 & \zeta_2 & \cdots & \zeta_2^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta_1 & \cdots & \zeta_1^{m-1} \end{pmatrix} \;\text{ jednaka }\; \prod_{i<j} (\zeta_i - \zeta_j) \neq 0.$$

⊞

Ako su neke nultočke polinoma $\rho(\zeta)$ višestruke, sa ukupno $\ell < m$ međusobno različitih, onda konstrukcija iz prethodne leme ne daje dovoljno rješenja za fundamentalni sistem. Trebamo konstruirati dodatna rješenja. Za motivaciju, neka je ζ_1 dvostruka i zamislimo je kao rezultat kolizije u limesu dvije jednostruke $\zeta_1, \zeta_1 + \epsilon$, $\epsilon \to 0$. Zbog linearnosti je (u toj zamišljenoj situaciji) za svaki $\epsilon > 0$ rješenje dano i formulom

$$\frac{1}{\epsilon}((\zeta_1 + \epsilon)^k - \zeta_1^k), \;\; \text{pri čemu je } \lim_{\epsilon \to 0} \frac{1}{\epsilon}((\zeta_1 + \epsilon)^k - \zeta_1^k) = \frac{d}{d\zeta}\zeta^k_{|\zeta=\zeta_1} = k\zeta_1^{k-1}.$$

Teorem 1.4.3. *Neka su $\zeta_1, \ldots, \zeta_\ell$ sve međusobno različite nultočke polinoma (1.4.6), te neka je κ_i algebarska kratnost nultočke ζ_i. Tada m nizova*

$$\left.\begin{array}{l} \{y_k\} = \{\zeta_i^k\} \\ \{y_k\} = \{k\zeta_i^{k-1}\} \\ \{y_k\} = \{k(k-1)\zeta_i^{k-2}\} \\ \cdots \\ \{y_k\} = \{k(k-1)\cdots(k-\kappa_i+2)\zeta_i^{k-\kappa_i+1}\} \end{array}\right\} \; i = 1, \ldots, \ell \qquad (1.4.7)$$

čine fundamentalni sistem rješenja od (1.4.5) u $k = 0$.

Primjer 1.4.1. Nultočke polinoma $\rho(\zeta)$ su općenito kompleksni borjevi pa su rješenja opisana u Teoremu 1.4.3 kompleksna. Ako trebamo realna rješenja jednadžbe (1.4.5) sa realnim koeficijentima, onda sve kompleksne nultočke dolaze u kolpleksno–konjugiranim parovima i svaki takav par možemo zamijeniti sa realnim i imaginarnim dijelom. Na primjer, $y_{n+2} + y_n = 0$ ima $\rho(\zeta) = \zeta^2 + 1$ sa $\zeta_1 = \mathfrak{i}$, $\zeta_2 = -\mathfrak{i}$, pa su rješenja $\{\mathfrak{i}^k\}$, $\{(-\mathfrak{i})^k\}$. Umjesto njih, za realan fundamentalni sistem možemo uzeti $\{\cos(n\pi/2)\}$ i $\{\sin(n\pi/2)\}$.

1.4.4 Konvergencija višekoračnih metoda

Analiziramo konvergenciju m–koračne metode na intervalu $[a, b]$, kojeg smo podijelili na n jednakih podintervala duljine $h = (b-a)/n$, oblika $[t_k, t_{k+1}]$, $t_k = a+kh$, $k = 0, 1, \ldots$. Zanima nas kako dobro izračunate vrijednosti y_k aproksimiraju vrijednosti točnog rješenja $y(t_k)$ kada $n \to \infty$, tj. kada $h \to 0$. Općenito se može promatrati proizvoljne podjele, ali takve da $\max_k(t_{k+1} - t_k)$ konvergira u nulu kada $n \to \infty$. Takva razmatranja su tehnički nešto kompliciranija i nećemo ih provoditi. Kod nas će niz $t_0, \ldots, t_n$ uvijek definirati ekvidistantnu subdiviziju sa konstantnim korakom $h = t_{k+1} - t_k$.

Lako vidimo da pomakom indeksa svaku m–koračnu metodu iz prethodnih sekcija možemo zapisati u generičkom obliku kao

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h\phi(t_k, y_k, \ldots, y_{k+m}, h),\ \alpha_m \neq 0, \tag{1.4.8}$$

gdje su $y_0 \approx y(t_0), \ldots, y_{m-1} \approx y(t_{m-1})$ zadane početne aproksimacije, preslikavanje ϕ koje definira metodu je definirano na

$$\phi : [a, b] \times \underbrace{\mathbb{R}^d \times \mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{m+1} \times \mathbb{R}_+ \longrightarrow \mathbb{R}^d.$$

Metoda je linearna m–koračna ako je oblika

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h \sum_{j=0}^{m} \beta_j f(t_{k+j}, y_{k+j}),\ \alpha_m \neq 0. \tag{1.4.9}$$

Dijeljenjem (1.4.8) s α_m možemo definicijsku relaciju normirati tako da bez smanjenja općenitosti uzimamo $\alpha_m = 1$.
Iz analize linearnih diferencijskih jednadži u §1.4.3 znamo da m–koračna generira više rješenja (rezultat zamjene diferencijalne jednadžbe prvog reda diferencijskom m–tog reda) i da će za uspjeh višekoračne metode morati postojati mehanizam koji će prigušiti ta *parazitska rješenja*.

1.4.4.1 Red metode

Kao i kod jednokoračnih metoda, uvodimo pojam lokalne pogreške diskretizacije:

Definicija 1.4.2. Za m–koračnu metodu (1.4.8) i $t + jh \in [a, b]$, $j = 0, \ldots, m$, je lokalna pogreška diskretizacije

$$\tau_m(t, h) = \sum_{j=0}^{m} \alpha_j y(t + jh) - h\phi(t, y(t), \ldots, y(t + mh), h).$$

Kažemo da je metoda konzistentna reda konzistentnosi p ako postoje $h_* > 0$ i $D > 0$ tako da, za sve $h \leq h_*$ i $t + jh \in [a, b]$, $j = 0, \ldots, m$, vrijedi ocjena $\|\tau_m(t, h)\| \leq Dh^{p+1}$.

Kod linearnih višekoračnih metoda je lokalna pogreška diskretizacije oblika

$$\tau_m(t, h) \equiv \mathcal{L}_m(y(t), h) = \sum_{j=0}^{m} \alpha_j y(t + jh) - h \sum_{j=0}^{m} \beta_j y'(t + jh) \tag{1.4.10}$$

i možemo je shvatiti i kao linearni diferencijski operator koji djeluje na (ovisno o situaciji i primjeni) dovoljno glatke funkcije $y(\cdot)$. Koristeći Taylorove razvoje

$$\begin{aligned} y(t + jh) &= y(t) + y'(t)jh + \frac{y''(t)}{2!}j^2h^2 + \frac{y'''(t)}{3!}j^3h^3 + \cdots \\ y'(t + jh) &= y'(t) + y''(t)jh + \frac{y'''(t)}{2!}j^2h^2 + \frac{y^{(4)}(t)}{3!}j^3h^3 + \cdots \end{aligned}$$

i grupiranjem elemenata po potencijama od h izraz za $\mathcal{L}_m(y(t), h)$ možemo zapisati u obliku

$$\mathcal{L}_m(y(t), h) = C_0 y(t) + C_1 h y'(t) + C_2 h^2 y''(t) + \cdots + C_\ell h^\ell y^{(\ell)}(t) + \cdots$$

gdje su koeficijenti C_ℓ dani s

$$\begin{aligned} C_0 &= \sum_{j=0}^{m} \alpha_j, \quad C_1 = \sum_{j=1}^{m} j\alpha_j - \sum_{j=0}^{m} \beta_j, \quad \text{i općenito} \\ C_\ell &= \frac{1}{\ell!} \sum_{j=1}^{m} j^\ell \alpha_j - \frac{1}{(\ell - 1)!} \sum_{j=1}^{m} j^{\ell-1} \beta_j, \quad \ell = 2, 3, \ldots \end{aligned}$$

Definicija 1.4.3. Kažemo da je linearni diferencijski operator (1.4.10) reda p ako je $C_0 = \ldots = C_p = 0$ i $C_{p+1} \neq 0$.

1.4.4.2 Ocjena globalne pogreške

Želimo ocijeniti globalnu pogrešku $e_k = y_k - y(t_k)$, $k = 0, \ldots, n$, te vidjeti što se događa kada $h \to 0$. Za početak, slijedimo shemu koju smo primijenili kod jednokoračnih metoda. Komplikacije koje nastaju zbog složenije strukture višekoračne metode ćemo rješavati u hodu.

Naravno, u realnim uvjetima primjene metode ne možemo izbjeći pogreške konačne (strojne) aritmetike i druge aproksimacije pri računanju vrijednosti funkcije ϕ, rješavanja jednadži u implicitnim metodama itd. To znači da za realističnu analizu metode, u k–tom koraku (1.4.8) trebamo uključiti i tu dodatnu pogrešku – označimo je s ξ_{k+m}.

Oduzimanjem relacija

$$\begin{aligned} \sum_{j=0}^{m} \alpha_j y_{k+j} &= h\phi(t_k, y_k, \ldots, y_{k+m}, h) + \xi_{k+m} \\ \sum_{j=0}^{m} \alpha_j y(t_{k+j}) &= h\phi(t_k, y(t_k), \ldots, y(t_{k+m}), h) + \tau_m(t_k, h) \end{aligned}$$

i uz oznake $\tau_{k+m} = \tau_m(t_k, h)$, $\delta_k = \phi(t_k, y_k, \ldots, y_{k+m}, h) - \phi(t_k, y(t_k), \ldots, y(t_{k+m}), h)$ dobijemo

$$\sum_{j=0}^{m} \alpha_j e_{k+j} = \underbrace{h\delta_k + \xi_{k+m} - \tau_{k+m}}_{\omega_k}, \ k = 0, \ldots, n-m. \tag{1.4.11}$$

Izvedene relacije možemo kompaktno zapisati kao

$$\underbrace{\begin{pmatrix} e_{k+1} \\ e_{k+2} \\ \vdots \\ e_{k+m-1} \\ e_{k+m} \end{pmatrix}}_{E_{k+1}} = A \underbrace{\begin{pmatrix} e_k \\ e_{k+1} \\ \vdots \\ e_{k+m-2} \\ e_{k+m-1} \end{pmatrix}}_{E_k} + \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \omega_k \end{pmatrix}}_{\Omega_k}, \ A = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{m-2} & -\alpha_{m-1} \end{pmatrix}}_{A_\alpha} \otimes \mathsf{I}_d.$$

Sada zamjenama $E_k = A(AE_{k-2} + \Omega_{k-2}) + \Omega_{k-1}$ itd. dobijemo

$$E_k = A^k E_0 + \sum_{j=0}^{k-1} A^{k-j-1} \Omega_j, \ k = 0, \ldots, n-m+1, \tag{1.4.12}$$

gdje $E_0 = (e_0, \ldots, e_{k-1})^T$ sadrži pogreške inicijalnih m vrijednosti $y_0, \ldots, y_{m-1}$.

Odmah uočavamo da postoji opasnost nekontroliranog rasta pogreške sa $k \to \infty$. Zato pretpostavimo da je

$$\sup_k \|A^k\|_\infty \leq C < \infty. \tag{1.4.13}$$

Tada je

$$\begin{aligned} \|E_k\|_\infty &\leq \|A\|_\infty^k \|E_0\|_\infty + \sum_{j=0}^{k-1} \|A\|_\infty^{k-j-1} \|\Omega_j\|_\infty & (1.4.14)\\ &\leq C(\|E_0\|_\infty + \sum_{j=0}^{k-1} \|\Omega_j\|_\infty); & (1.4.15)\\ \|\Omega_j\|_\infty &= \|\omega_j\|_\infty \leq h\|\delta_j\|_\infty + \|\xi_{j+m}\|_\infty + \|\tau_j\|_\infty. \end{aligned}$$

Sada pretpostavimo da je ϕ Lipschitzova:

$$\|\phi(t_k, y_k, \ldots, y_{k+m}, h) - \phi(t_k, y(t_k), \ldots, y(t_{k+m}), h)\|_\infty \leq L_\phi \sum_{j=0}^{m} \|y_{k+j} - y(t_{k+j})\|_\infty \tag{1.4.16}$$

i odmah zaključimo da je $\|\delta_j\|_\infty \leq L_\phi \sum_{i=0}^{m} \|e_{j+i}\|_\infty$ pa je

$$\begin{aligned} \|\Omega_j\|_\infty &\leq \max_{i=0:n-m} \|\tau_i\|_\infty + hL_\phi \sum_{i=0}^{m} \|e_{j+i}\|_\infty + \|\xi_{j+m}\|_\infty \\ &\leq \max_{k=0:n-m} \|\tau_k\|_\infty + hL_\phi(m\|E_j\|_\infty + \|E_{j+1}\|_\infty) + \|\xi_{j+m}\|_\infty \\ \sum_{j=0}^{k-1} \|\Omega_j\|_\infty &\leq \underbrace{(n \max_i \|\tau_i\|_\infty + n \max_i \|\xi_i\|_\infty)}_{\eta} + hL_\phi \sum_{i=0}^{k-1} (m\|E_i\|_\infty + \|E_{i+1}\|_\infty) \\ &\leq \eta + hL_\phi(m+1) \sum_{i=0}^{k-1} \|E_i\|_\infty + hL_\phi \|E_k\|_\infty. \end{aligned}$$

Sada u (1.4.15) vrijedi

$$\|E_k\|_\infty \leq C(\|E_0\|_\infty + \eta + hL_\phi(m+1) \sum_{i=0}^{k-1} \|E_i\|_\infty + hL_\phi \|E_k\|_\infty)$$

Neka je sada $h_* > 0$ odabran tako da je $h_* L_\phi C < 1$, i neka je $h \in (0, h_*]$. Prethodna nejednakost daje

$$\begin{aligned}\|E_k\|_\infty &\leq \boxed{\tfrac{C}{1-ChL_\phi}(\|E_0\|_\infty + n\max_i \|\tau_i\|_\infty + n\max_i \|\xi_i\|_\infty)} \\ &\quad + \boxed{\tfrac{(m+1)CL_\phi}{1-ChL_\phi}} h \sum_{i=0}^{k-1} \|E_i\|_\infty \\ &\equiv \alpha + \beta \sum_{i=0}^{k-1} h\|E_i\|_\infty, \quad k = 1, \ldots, n-m+1.\end{aligned}$$

U ovoj situaciji primjenjujemo standardnu nejednakost koja je opisana u sljedećoj lemi:

Lema 1.4.4. *Neka su zadani $h_0, \ldots, h_{r-1} > 0$ i $\alpha \geq 0$, $\beta \geq 0$. Ako $v_1, \ldots, v_r$ zadovoljavaju*

$$|v_0| \leq \alpha, \quad |v_i| \leq \alpha + \beta \sum_{j=0}^{i-1} h_j |v_j|, \ i = 1, \ldots, r,$$

onda vrijedi $|v_i| \leq \mathbf{e}^{\beta \sum_{j=0}^{i-1} h_j}$, $i = 1, \ldots, r$.

Dakle, imamo

$$\begin{aligned}\|E_k\|_\infty &\leq \frac{\max(C,1)}{1-ChL_\phi}(\|E_0\|_\infty + n\max_i \|\tau_i\|_\infty + n\max_i \|\xi_i\|_\infty)\mathbf{e}^{\frac{(m+1)CL_\phi}{1-ChL_\phi}(n-m)h} \\ &\leq K(\|E_0\|_\infty + \frac{1}{h}\max_i \|\tau_i\|_\infty) + K\frac{1}{h}\max_i \|\xi_i\|_\infty, \\ &\quad \text{sa } K = \frac{(b-a)\max(C,1)}{1-Ch_*L_\phi}\mathbf{e}^{\frac{(m+1)CL_\phi}{1-Ch_*L_\phi}(b-a)}.\end{aligned}$$

1.4.4.3 Omeđenost niza A^k i Dahlquistov uvjet stabilnosti

Jedna od ključnih pretpostavki u prethodnim razmatranjima je bila da je niz potencija $\|A^k\|_\infty$ omeđen odozgor. Zapravo, iz relacije (1.4.12) se dade naslutiti da je ta omeđenost i nužna za dobro ponašanje globalne pogreške. Zato nam je važno znati koji su uvjeti na samu metodu pa da niz $\|A^k\|_\infty$ bude omeđen. Sljedeća propozicija to pitanje svodi na pitanje omeđenosti potencija matrice A_α.

Propozicija 1.4.5. *Vrijedi* $(A_\alpha \otimes \mathsf{I}_d)^k = A_\alpha^k \otimes \mathsf{I}_d$ *i za* $\|\cdot\| \in \{\|\cdot\|_1, \|\cdot\|_\infty, \|\cdot\|_2\}$ *je* $\|(A_\alpha \otimes \mathsf{I}_d)^k\| = \|A_\alpha^k\|$.

Dokaz: Direktna provjera. ⊞

Propozicija 1.4.6. *Karakteristični polinom matrice*

$$A_\alpha = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{m-2} & -\alpha_{m-1} \end{pmatrix}$$

je ujedno i njen minimalni polinom i glasi $\chi_{A_\alpha}(\lambda) = \lambda^m + \sum_{j=0}^{m-1} \alpha_j \lambda^j$.

Dokaz: Lako provjerimo, razvojem determinante po zadnjem retku, da je zbilja $\det(\lambda \mathsf{I} - A_\alpha) = \lambda^m + \sum_{j=0}^{m-1} \alpha_j \lambda^j$. Nadalje, ako je $p(\lambda) = \sum_{j=0}^{\ell} \beta_j \lambda^j$ polinom stupnja $\ell < m$ ($\beta_\ell \neq 0$), pokazat ćemo da je $p(A_\alpha) \neq \mathbf{0}$. Primijetimo da je $e_1^T A_\alpha^j = e_{j+1}^T$, $j = 1, \ldots, m-1$. Dakle,

$$e_1^T p(A_\alpha) = \sum_{j=0}^{\ell} \beta_j e_1^T A_\alpha^j = (\beta_0, \beta_1, \ldots, \beta_\ell, \underbrace{0, \ldots, 0}_{m-\ell-1}) \neq \mathbf{0}.$$

⊞

Teorem 1.4.7. *Neka je* A *proizvoljna* $n \times n$ *matrica. Tada je niz potencija* $(A^k,\ k = 0, 1, 2, \ldots)$ *omeđen ako i samo ako su ispunjena sljedeća dva uvjeta:*

1. *Spektralni radijus matrice* A *je najviše jedan:* $\mathrm{spr}(A) \leq 1$.
2. *Svaka svojstvena vrijednost* λ *za koju je* $|\lambda| = 1$ *je jednostruka nultočka minimalnog polinoma od* A. *(Ovo je ekvivalentno zahtjevu da u Jordanovoj normalnoj formi od* A *toj svojstvenoj vrijednosti* λ *pripadaju* 1×1 *blokovi.)*

Dokaz: Neka je $A = TJT^{-1}$ dekompozicija sa Jordanovom normalnom formom $J = \oplus_i J_i$, gdje je $J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_i & 1 \\ & & & \lambda_i \end{pmatrix}$ blok dimenzije $n_i \times n_i$. Ako je $n_i = 1$, onda je $J_i = (\lambda_i)$.

Ako pretpostavimo dva uvjeta teorema, onda je, za svaki λ_i, $|\lambda_i| \leq 1$ i $|\lambda_i| = 1 \Longrightarrow n_i = 1$. Specijalno je za $n_i \geq 2$ ispunjeno $|\lambda_i| < 1$ pa $\varepsilon = 1 - \max_{n_i \geq 2} |\lambda_i|$ zadovoljava $|\lambda_i| + \varepsilon \leq 1$ kad god je $n_i \geq 2$.

Ako stavimo $D_\varepsilon = \mathrm{diag}(1, \varepsilon, \ldots, \varepsilon^{n-1})$ i definiramo $J_\varepsilon = D_\varepsilon^{-1} J D_\varepsilon$, onda je $(J_\varepsilon)_{ij} = \varepsilon^{j-i} J_{ij}$, tj. $J_\varepsilon = \oplus J_{\varepsilon,i}$, gdje je $J_{\varepsilon,i} = \begin{pmatrix} \lambda_i & \varepsilon & & \\ & \ddots & \ddots & \\ & & \lambda_i & \varepsilon \\ & & & \lambda_i \end{pmatrix}$, ili $J_{\varepsilon_i} = (\lambda_i)$. Očito je $\|J_\varepsilon\|_\infty \leq 1$, pa je onda i za svaki $k = 0, 1, \ldots$, $\|J_\varepsilon^k\|_\infty \leq 1$. Konačno, kako je $A^k = T D_\varepsilon J_\varepsilon^k D_\varepsilon^{-1} T^{-1}$, zaključujemo

$$\|A^k\|_\infty \leq \frac{\|T\|_\infty \|T^{-1}\|_\infty}{(1 - \max_{n_i \geq 2} |\lambda_i|)^{n-1}}.$$

Sada pokažimo da su uvjeti teorema i nužni. Pa recimo da ne vrijede oba. Ako ne vrijedi prvi uvjet, onda je $\mathrm{spr}(A) > 1$, pa postoji svojstvena vrijednost λ_i od A koja je po modulu strogo veća od jedan, $|\lambda_i| > 1$. Tada je $\mathrm{spr}(A^k) \geq |\lambda_i|^k \longrightarrow \infty$, $(k \to \infty)$, pa A^k ne može biti omeđeno. Ako ne vrijedi drugi uvjet, onda mora postojati svojstvena vrijednost λ_i od A modula jedan i sa $n_i \geq 2$. Sada se sjetimo da $f(J_i) = J_i^k$ ima strukturu npr. ako je $n_i = 5$

$$f(J_i) = \begin{pmatrix} f(\lambda_i) & \frac{f^{(1)}(\lambda_i)}{1!} & \frac{f^{(2)}(\lambda)}{2!} & \frac{f^{(3)}(\lambda)}{3!} & \frac{f^{(4)}(\lambda)}{4!} \\ 0 & f(\lambda_i) & \frac{f^{(1)}(\lambda_i)}{1!} & \frac{f^{(2)}(\lambda)}{2!} & \frac{f^{(3)}(\lambda)}{3!} \\ 0 & 0 & f(\lambda_i) & \frac{f^{(1)}(\lambda_i)}{1!} & \frac{f^{(2)}(\lambda)}{2!} \\ 0 & 0 & 0 & f(\lambda_i) & \frac{f^{(1)}(\lambda)}{1!} \\ 0 & 0 & 0 & 0 & f(\lambda_i) \end{pmatrix},$$

odakle zaključujemo da je $(J_i^k)_{12} = k\lambda_i^{k-1}$, tj. da J_i^k pa onda i J^k neograničeno raste kada $k \to \infty$, $\|J^k\|_\infty \geq k$. Naravno, onda, iz $J^k = T^{-1} A^k T$, je $\|A^k\|_\infty \geq \|J^k\|_\infty / (\|T\|_\infty \|T^{-1}\|_\infty) \to \infty$. $\boxplus$

1.4.4.4 Konvergencija = stabilnost + konsistentnost

Sada prethodna razmatranja možemo formalizirati i iskazati u obliku ključnih teorema u analizi višekoračnih metoda.

Definicija 1.4.4. m–koračna metoda

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h\phi(t_k, y_k, \ldots, y_{k+m}, h)$$

je nul–stabilna (kažemo i Dahlquist–stabilna) ako njen pripadni polinom

$$\rho(\zeta) = \alpha_m \zeta^m + \alpha_{m-1}\zeta^{m-1} + \ldots + \alpha_1 \zeta + \alpha_0$$

ima sve nultočke unutar kruga $\{\zeta \in \mathbb{C} \ : \ |\zeta| \leq 1\}$ i ako je svaka nultočka sa ruba tog kruga jednostruka ($\rho(\zeta) = 0 \ \ \& \ \ |\zeta| = 1 \Longrightarrow \rho'(\zeta) \neq 0$).

Dakle, dokazali smo sljedeći teorem:

Teorem 1.4.8. *Neka je m–koračna metoda (1.4.8) nul–stabilna i neka je $C = \sup_k \|A^k\|_\infty$. Nadalje, neka ϕ zadovoljava Lipschitzov uvjet: postoji konstanta $L_\phi > 0$ tako da za sve $t \in [a,b]$, $h > 0$ i sve $v_i, w_i \in \mathbb{R}^d$ vrijedi*

$$\|\phi(t, v_0, \ldots, v_m, h) - \phi(t, w_0, \ldots, w_m, h)\|_\infty \leq L_\phi \sum_{j=0}^{m} \|v_j - w_j\|_\infty.$$

Tada za $h_* < 1/(CL_\phi)$ i svaki $n \in \mathbb{N}$ za kojeg je $h = (b-a)/n \in (0, h_*]$ vrijedi

$$\max_{k=0:n} \|y_k - y(t_k)\|_\infty \leq K(\max_{k=0:m-1} \|y_k - y(t_k)\|_\infty + \frac{1}{h} \max_i \|\tau_i\|_\infty + \frac{1}{h} \max_i \|\xi_i\|_\infty)$$

gdje je $K = \dfrac{(b-a)\max(C,1)}{1 - Ch_* L_\phi} \mathbf{e}^{\frac{(m+1)CL_\phi}{1-Ch_*L_\phi}(b-a)}$. *Ako je metoda konsistentna reda p onda imamo*

$$\max_{k=0:n} \|y_k - y(t_k)\|_\infty \leq K(\max_{k=0:m-1} \|y_k - y(t_k)\|_\infty + Dh^p + \frac{1}{h} \max_i \|\xi_i\|_\infty).$$

Uočavamo dvije stvari:

(i) Ako je metoda reda konzistentnosti p, onda je u idelanim teorijskim uvjetima ($\xi_i = \mathbf{0}$ za sve i) njen potencijal aproksimiranja točnog rješenja s greškom $O(h^p)$ uvjetovan odgovarajućom ocjenom za $\max_{k=0:m-1} \|y_k - y(t_k)\|_\infty$, tj. inicijalnih m vrijednosti $y_0, \ldots, y_m$ moraju biti odgovarajuće kvalitete. Drugim riječima, za startati m–koračnu metodu ćemo trebati jednokoračnu metodu odgovarajućeg reda. Za to nam mogu poslužiti npr. Runge–Kuttine metode.

(ii) U praktičnom računanju, kada ne možemo izbjeći pogreške $\xi_i \neq \mathbf{0}$, vidimo da postoji efekt njihovog akumuliranja i faktor $1/h = n/(b-a)$ sprječava da taj dio pogreške ide u nulu kada $h \to 0$. To znači da valja biti oprezan sa izborom koraka h.

Definicija 1.4.5. Kažemo da je linearna višekoračna metoda (1.4.9) konvergentna ako za sve inicijalne probleme za koje vrijedi Teorem 1.1.2 vrijedi

$$\lim_{\substack{h\to 0\\ x=a+nh}} y_n = y(x)$$

za sve $x \in [a,b]$ i niz $\{y_n\}$ generiran metodom s inicijalnim uvjetima $y_i = y_i(h)$ za koje je $\lim_{h\to 0} y_i(h) = y_0$, $i = 0,\ldots,m-1$.

Teorem 1.4.9. *(Dahlquits) Linearna višekoračna metoda je konvergentna ako i samo ako je konzistentna i nul–stabilna.*

1.5 Kruti sistemi i $\mathcal{A}$–stabilnost

Promotrimo linearni sustav diferencijalnih jednadžbi $y'(t) = Ay(t) + \varphi(t)$ u kojem $d \times d$ matrica A ima jednostruke svojstvene vrijednosti $\lambda_1,\ldots,\lambda_d$, sa pripadnim, međusobno nezavisnim, svojstvenim vektorima $v_1,\ldots,v_d$. Znamo da je tada opće rješenje oblika

$$y(t) = \sum_{i=1}^{d} c_i \mathbf{e}^{\lambda_i t} v_i + \psi(t).$$

U primjenama, na primjer kemiji, jednadžbe opisuju kemijsku reakciju i zanima nas stacionarno rješenje, pri čemu je $\mathrm{Re}(\lambda_i) < 0$ za sve i. Sama kemijska reakcija se sastoji od više procesa koji se odvijaju u različitim vremenskim skalama. Za transijentni dio rješenja vrijedi

$$\lim_{t\to\infty} \sum_{i=1}^{d} c_i \mathbf{e}^{\lambda_i t} v_i = 0,$$

pri čemu svaki $\mathbf{e}^{\lambda_i t}$ trne u nulu brzinom ovisnom o λ_i. Ako želimo numerički dobiti stacionarno rješenje, onda numerička metoda mora u vremenu napredovati tako daleko dok cijeli transijentni dio ne utrne. Brzina najsporije padajuće komponente je određena sa $\min_i |\,\mathrm{Re}(\lambda_i)|$ – što je taj broj manji (tj. što je neka svojstvena vrijednost bliže imaginarnoj osi) to će trebati više koraka metode. S druge strane, ako je $\max_i |\,\mathrm{Re}(\lambda_i)|$ veliki, to nas tjera na uzimanje manjeg koraka diskretizacije h – što je taj broj veći trebamo manji korak. Sve zajedno, u najgoroj situaciji smo ako moramo numerički rješavati jednadžbu na velikom intervalu i to sa jako malim korakom.

Definicija 1.5.1. Linearni sustav diferencijalnih jednadžbi $y'(t) = Ay(t) + \varphi(t)$ je krut ako su sve svojstvene vrijednosti $\lambda_1, \ldots, \lambda_d$ matrice A u lijevoj otvorenoj poluravnini ($\mathrm{Re}(\lambda_i) < 0$, $i = 1, \ldots, d$) i ako je omjer

$$\frac{\max_i |\,\mathrm{Re}(\lambda_i)|}{\min_i |\,\mathrm{Re}(\lambda_i)|} \gg 1.$$

Primjer 1.5.1. Promotrimo na $(0, \infty)$ sustav $y'(t) = Ay(t)$ sa inicijalnim uvjetom $y(0) = y_0$, gdje pretpostavljamo da je matrica A ima samo jednostruke svojstvene vrijednosti $\lambda_1, \ldots, \lambda_d$, te da je $\max_j \mathrm{Re}(\lambda_j) < 0$. Tada je A dijagonalizabilna i njenu spektralna dekompozicija je $A = V\Lambda V^{-1}$, gdje su stupci $v_1, \ldots, v_d$ svojstveni vektori, $Av_i = \lambda_i v_i$, $\Lambda = \mathrm{diag}(\lambda_i)_{i=1}^n$. Rješenje $y(t)$ možemo pisati kao

$$y(t) = \mathbf{e}^{tA} y_0 = V\mathbf{e}^{t\Lambda}V^{-1}y_0 = \sum_{j=1}^{d} \mathbf{e}^{\lambda_j t} \underbrace{(V^{-1}y_0)_j}_{\gamma_j} v_j \equiv \sum_{j=1}^{d} \mathbf{e}^{\lambda_j t}\gamma_j v_j.$$

Kako su sve svojstvene vrijednosti po pretpostavci u lijevoj poluravnini, vrijedi $\lim_{t\to\infty} y(t) = \mathbf{0}$. Pri tome brzina padanja rješenja u smjeru svojstvenog vektora v_j ovisi o veličini $\mathrm{Re}(\lambda_j)$. Dakle, ako je neki $\mathrm{Re}(\lambda_j) < 0$ blizu nuli, onda će trebati odgovarajuće veliki $t > 0$ dok ta komponenta ne bude dovoljno mala.

Ako ovaj problem pokušamo riješiti Eulerovom metodom, u k–tom koraku imamo

$$y_k = (\mathsf{I}_d + hA)^k y_0 = V(\mathsf{I}_d + h\Lambda)^k V^{-1} y_0 = \sum_{j=1}^{d} (1 + h\lambda_j)^k \gamma_j v_j.$$

Odmah uočavamo problem: ako je za neki j $|1 + h\lambda_j| > 1$, onda će duž smjera v_j niz y_k rasti u beskonačno brzinom $|1+h\lambda_j|^k$. Kažemo da imamo *parazitsko rješenje* koje je generirala metoda i koje je očito u suprotnosti sa ponašanjem rješenja koje u beskonačnosti trne u nulu.

Sada na istom primjeru primijenimo trapeznu metodu

$$y_{k+1} = y_k + \frac{h}{2}(Ay_k + Ay_{k+1}).$$

U k–tom koraku je

$$y_k = (\mathsf{I}_d - \frac{h}{2}A)^{-k}(\mathsf{I}_d + \frac{h}{2}A)^k y_0 = \sum_{j=1}^{d} \left(\frac{1 + \frac{h}{2}\lambda_j}{1 - \frac{h}{2}\lambda_j}\right)^k \gamma_j v_j.$$

Sada odmah uočimo da je, zbog $\mathrm{Re}(\lambda_j) < 0$,

$$\left| \frac{1 + \frac{h}{2}\lambda_j}{1 - \frac{h}{2}\lambda_j} \right| < 1, \text{ za proizvoljan } h > 0.$$

Dakle, trapezna formula će u beskonačnosti reproducirati ponašanje egzaktnog rješenja i to sa proizvoljnim korakom $h > 0$, neovisno o svojstvenim vrijednostima.

Za nastavak proučavanja fenomena opisanog u prethodnom primjeru promatramo model–problem

$$y'(t) = \lambda y(t), t \geq 0, y(0) = 1, \tag{1.5.1}$$

sa rješenjem $y(t) = \mathbf{e}^{\lambda t}$. Uzimamo $\mathrm{Re}(\lambda) < 0$, tako da je $\lim_{t\to\infty} y(t) = 0$. Za metodu koja sa korakom $h > 0$ rješava ovaj problem definiramo domenu (linearne) stabilnosti kao skup

$$\mathcal{D} = \{h\lambda \in \mathbb{C} \ : \ \lim_{n\to\infty} y_n = 0\}.$$

U gornjim primjerima je dakle

$$\begin{aligned} \mathcal{D}_{\text{Euler}} &= \{z \in \mathbb{C} \ : \ |1 + z| < 1\}, \\ \mathcal{D}_{\text{Trapez}} &= \{z \in \mathbb{C} \ : \ \left| \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z} \right| < 1\} = \{z \in \mathbb{C} \ : \ \mathrm{Im}(z) < 0\}. \end{aligned}$$

Definicija 1.5.2. Numerička metoda za (1.5.1) je $\mathcal{A}$–stabilna ako njena domena stabilnosti $\mathcal{D}$ sadrži cijelu otvorenu lijevu poluravninu $\{z \in \mathbb{C} \ : \ \mathrm{Im}(z) < 0\}$.

Sada želimo proučiti kako izgleda numeričko rješenje problema (1.5.1), ako koristimo m–koračnu metodu

$$\sum_{j=0}^{m} \alpha_j y_{k+j} = h\lambda \sum_{j=0}^{m} \beta_j y_{k+j}, \ k = 0, 1, 2, \ldots$$

Zahtijevamo stabilnost za svaki odabir inicijalnih vrijednosti $y_0, \ldots, y_{m-1}$. Aproksimacije generirane metodom zadovoljavaju linearnu diferencijsku jednadžbu

$$\sum_{j=0}^{m} (\alpha_j - h\lambda\beta_j) y_{k+j} = 0, \ k = 0, 1, 2, \ldots \tag{1.5.2}$$

Stavimo

$$\sigma(\zeta) = \sum_{j=0}^{m} \beta_j \zeta^j, \;\; \omega(\zeta;\xi) = \rho(\zeta) - \xi\sigma(\zeta).$$

Prema §1.4.3, rješenje od (1.5.2) je određeno nultočkama pripadnog polinoma

$$\sum_{j=0}^{m} (\alpha_j - h\lambda\beta_j)\zeta^j = \rho(\zeta) - h\lambda\sigma(\zeta) = \omega(\zeta; h\lambda).$$

Propozicija 1.5.1. *Neka je $\omega(\zeta;\xi)$ polinom pridružen linearnoj m–koračnoj metodi (1.5.2) i neka su $z_1(\xi), \ldots, z_{\ell(\xi)}(\xi)$ sve njegove međusobno različite nultočke. Tada je metoda (1.5.2) $\mathcal{A}$–stabilna ako i samo ako vrijedi da je za svaki $\xi \in \mathbb{C}$ sa* $\mathrm{Re}(\xi) < 0$ *ispunjeno*

$$\max_{i=1:\ell(\xi)} |z_i(\xi)| < 1.$$

Navedimo još jedan, donekle razočaravajući rezultat.

Teorem 1.5.2. *(Dahlquist) Eksplicitna linearna višekoračna metoda ne može biti $\mathcal{A}$–stabilna. $\mathcal{A}$–stabilna implicitna višekoračna metoda može biti reda najviše dva.*

Takvih rezultata ima još:

Teorem 1.5.3. *Niti jedna eksplicitna Runge–Kuttina metoda ne može biti $\mathcal{A}$–stabilna.*

Naravno, ovi negativni rezultati ne znače da su te metode beskorisne, već nas upozoravaju da metode u konkretnim primjenama treba uvijek primijenjivati *cum grano salis*.

Poglavlje 2

Rubni problem za ODJ

U primjenama se često javljaju tzv. rubni problemi, gdje uz diferencijalnu jednadžbu koja definira funkciju na nekoj domeni imamo i zadano ponašanje rješenja jednadžbe na rubovima domene.

Jednostavni primjer je rubni problem za običnu diferencijalnu jednadžbu drugog reda

$$\begin{aligned} &u''(x) = f(x, u(x), u'(x)),\ x \in [a,b] \subset \mathbb{R} \\ &u(a) = \alpha,\ u(b) = \beta. \end{aligned}$$

Mi ćemo osnovne elemente numeričkog rješavanja rubnih problema ilustrirati na jednom specijalnom primjeru iz klase Sturm–Liouvilleovih rubnih problema:

$$\begin{aligned} &-y''(x) + r(x)u(x) = f(x),\ x \in [a,b] \\ &y(a) = \alpha,\ y(b) = \beta. \end{aligned} \tag{2.0.1}$$

Teorem 2.0.4. *Ako su funkcije $r, f : [a,b] \longrightarrow \mathbb{R}$ neprekidne na $[a,b]$ i ako je $r(x) \geq 0$ za sve $x \in [a,b]$ onda problem (2.0.1) ima jedinstveno rješenje $y \in \mathcal{C}^2[a,b]$.*

2.1 Rješavanje konačnim diferencijama

Segment $[a,b]$ diskretiziramo na standardni način: odaberemo $n \in \mathbb{N}$, stavimo

$$h = \frac{(b-a)}{n+1},\ \ x_i = a + i \cdot h,\ \ i = 0, \ldots, n+1. \tag{2.1.1}$$

Koristeći centralne diferencije, lijevu stranu jednadžbe izračunatu u unutarnjim čvorovima možemo zapisati kao

$$-u''(x_i) = \frac{-u(x_i - h) + 2u(x_i) - u(x_i + h)}{h^2} - u^{(4)}(x_i + \theta_i h)\frac{h^2}{12}. \tag{2.1.2}$$

(Naravno, u ovom momentu pretpostavljamo da je u klase $\mathcal{C}^4$.) Znači da rješenje problema u čvorovima x_i, $i = 1, \ldots, n$, zadovoljava

$$\frac{-u(x_i - h) + 2u(x_i) - u(x_i + h)}{h^2} + r(x_i)u(x_i) = f(x_i) + u^{(4)}(x_i + \theta_i h)\frac{h^2}{12}. \quad (2.1.3)$$

Aproksimacije $u_i \approx u(x_i)$ ćemo odrediti tako da zadovoljavaju slične jednadžbe i to tako da zanemarimo nepoznati $O(h^2)$ član i u_i definiramo (uz oznake $r_i = r(x_i)$, $f_i = f(x_i)$) s

$$\frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} + r_i u_i = f_i, \; i = 1, \ldots, n, \quad (2.1.4)$$

gdje je $u_0 = u(a) = \alpha$, $u_{n+1} = u(b) = \beta$. Vidimo da zadane rubne vrijednosti možemo iskoristiti u prvoj i u zadnjoj jednadžbi. Tako na primjer u prvoj, koja glasi

$$\frac{-u_0 + 2u_1 - u_2}{h^2} + r_1 u_1 = f_1$$

možedmo iskorititi $u_0 = \alpha$ i dobiti

$$\frac{2u_1 - u_2}{h^2} + r_1 u_1 = f_1 + \frac{\alpha}{h^2}.$$

Kada analogono modificiramo i zadnju jednadžbu, dobiveni sustav jednadžbi možemo zapisati u matričnom obliku kao

$$\frac{1}{h^2}\begin{pmatrix} 2+r_1h^2 & -1 & & & & \\ -1 & 2+r_2h^2 & -1 & & & \\ & -1 & 2+r_3h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2+r_{n-1}h^2 & -1 \\ & & & & -1 & 2+r_nh^2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-1} \\ u_n \end{pmatrix} = \begin{pmatrix} f_1+\alpha/h^2 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-1} \\ f_n+\beta/h^2 \end{pmatrix} \quad (2.1.5)$$

tj. $\frac{1}{h^2}Av = f$. Da bi aproksimacija bila dobro definirana, trebamo npr. imati osiguranu regularnost matrice A.

Teorem 2.1.1. *Ako su $r_i \geq 0$, $i = 1, \ldots, n$, onda je matrica*

$$A = \begin{pmatrix} 2+r_1h^2 & -1 & & & & \\ -1 & 2+r_2h^2 & -1 & & & \\ & -1 & 2+r_3h^2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2+r_{n-1}h^2 & -1 \\ & & & & -1 & 2+r_nh^2 \end{pmatrix}$$

pozitivno definitna i $0 \leq A^{-1} \leq \tilde{A}^{-1}$*, gdje je*

$$\tilde{A} = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

Specijalno je

$$\|A^{-1}\|_\infty \leq \|\tilde{A}^{-1}\|_\infty \leq \frac{(b-a)^2}{8h^2}.$$

Dokaz: Koristeći Geršgorinove krugove zaključujemo da je $\mathfrak{S}(\tilde{A}) \subset [0,4]$. Lako vidimo da je $\tilde{A}$ regularna, pa $0 \notin \mathfrak{S}(\tilde{A})$. Ako bi $\lambda = 4$ bila svojstvena vrijednost od $\tilde{A}$, onda bi $\tilde{A} - 4\mathsf{I}$ bila singularna, što je nemoguće jer je $\tilde{A} - 4\mathsf{I} = -\Omega\tilde{A}\Omega^{-1}$ sa $\Omega = \mathrm{diag}(\pm 1)$. Dakle, $\mathfrak{S}(\tilde{A}) \subset (0,4)$.

Specijalno je $\tilde{A}$ pozitivno definitna, pa je i $A = \tilde{A} + h^2\mathrm{diag}(r_i)_{i=1}^n$ pozitivno definitna: za $x \neq \mathbf{0}$ je

$$x^*Ax = x^*\tilde{A}x + h^2\sum_{i=1}^n r_i|x_i|^2 > 0.$$

Matricu A napišimo kao produkt

$$A = D(\mathsf{I} - J),\ J = \begin{pmatrix} 0 & \frac{1}{2+r_1h^2} & & & & \\ \frac{1}{2+r_2h^2} & 0 & \frac{1}{2+r_2h^2} & & & \\ & \frac{1}{2+r_3h^2} & 0 & \frac{1}{2+r_3h^2} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \frac{1}{2+r_{n-1}h^2} & 0 & \frac{1}{2+r_{n-1}h^2} \\ & & & & \frac{1}{2+r_nh^2} & 0 \end{pmatrix},$$

$D = \mathrm{diag}(2 + r_ih^2)_{i=1}^n$. Slično je $\tilde{A} = \tilde{D}(I - \tilde{J})$, $\tilde{D} = 2\mathsf{I}$. Odmah uočavamo da je

$$0 \leq \tilde{D} \leq D,\ \ 0 \leq J \leq \tilde{J} = \begin{pmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ & \frac{1}{2} & 0 & \frac{1}{2} & \\ & & \ddots & \ddots & \ddots \\ & & & \frac{1}{2} & 0 & \frac{1}{2} \\ & & & & \frac{1}{2} & 0 \end{pmatrix}.$$

Kako je $\tilde{J} = \frac{1}{2}(2\mathsf{I} - \tilde{A})$, zaključujemo da je $\mathfrak{S}(\tilde{J}) \subset (-1,1)$ pa je $\mathrm{spr}(\tilde{J}) < 1$. Dakle, red $\sum_{k=0}^\infty \tilde{J}^k$ konvergira i

$$0 \leq \mathsf{I} + \tilde{J} + \tilde{J}^2 + \tilde{J}^3 + \cdots = (\mathsf{I} - \tilde{J})^{-1}.$$

Nadalje, iz $0 \leq J \leq \tilde{J}$ slijedi da i $\sum_{k=0}^{\infty} J^k$ konvergira i

$$0 \leq \mathsf{I} + J + J^2 + J^3 + \cdots = (\mathsf{I} - J)^{-1} \leq (\mathsf{I} - \tilde{J})^{-1}.$$

Konačno, iz $0 \leq D^{-1} \leq \tilde{D}^{-1}$ imamo

$$0 \leq A^{-1} = (\mathsf{I} - J)^{-1} D^{-1} \leq (\mathsf{I} - \tilde{J})^{-1} \tilde{D}^{-1} = \tilde{A}^{-1}.$$

Sada izračunajmo gornju ogradu za $\|\tilde{A}\|_\infty$. Prvo se prisjetimo:

$$\|S\|_\infty = \max_{\|x\|_\infty = 1} \|Sx\|_\infty = \max_{k=1:n} \sum_{j=1}^{n} |s_{kj}|.$$

Ako je $S \geq 0$, onda imamo

$$\|S\|_\infty = \max_{k=1:n} \sum_{j=1}^{n} s_{kj} = \|S \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \|_\infty$$

Odavde imamo $T \geq S \geq 0 \Longrightarrow \|T\|_\infty \geq \|S\|_\infty$.

Sada promotrimo problem

$$\begin{aligned} &-z''(x) = 1,\ a < x < b \\ &z(a) = z(b) = 0, \end{aligned}$$

kojem je egzaktno rješenje parabola $z(x) = \frac{1}{2}(x-a)(b-x)$. Kako je $z^{(4)}(x) \equiv 0$, centralne diferencije perfektno aproksimiraju drugu derivaciju od z u čvorovima pa je

$$\frac{1}{h^2} \tilde{A} \begin{pmatrix} z(x_1) \\ \vdots \\ z(x_n) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \Longrightarrow \tilde{A}^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{h^2} \begin{pmatrix} z(x_1) \\ \vdots \\ z(x_n) \end{pmatrix} \tag{2.1.6}$$

Kako je

$$\max_{x \in [a,b]} \frac{1}{2}(x-a)(b-x) = \frac{(b-a)^2}{8}$$

zaključujemo da je

$$\|\tilde{A}^{-1}\|_\infty = \frac{1}{h^2} \| \begin{pmatrix} z(x_1) \\ \vdots \\ z(x_n) \end{pmatrix} \|_\infty \leq \frac{(b-a)^2}{8h^2}.$$

⊞

Teorem 2.1.2. *Neka rubni problem 2.0.1 ima rješenje $u \in \mathcal{C}^4[a,b]$. Neka je $u = (u_1, \ldots, u_n)$ diskretna aproksimacija rješenja na mreži čvorova (2.1.1), dobivena iz (2.1.5). Tada vrijedi*

$$|u(x_i) - u_i| \leq \frac{\|u^{(4)}\|_\infty}{24} h^2 (x_i - a)(b - x_i) \leq \frac{(b-a)^2}{96} \|u^{(4)}\|_\infty h^2.$$

Dokaz: Oduzimanjem relacija (2.1.3) i (2.1.5) dobijemo

$$\frac{1}{h^2} A \begin{pmatrix} u_1 - u(x_1) \\ \vdots \\ u_n - u(x_n) \end{pmatrix} = -\frac{h^2}{12} \begin{pmatrix} u^{(4)}(x_1 + \theta_1 h) \\ \vdots \\ u^{(4)}(x_n + \theta_n h) \end{pmatrix}$$

odakle je

$$\begin{pmatrix} u_1 - u(x_1) \\ \vdots \\ u_n - u(x_n) \end{pmatrix} = -\frac{h^4}{12} A^{-1} \begin{pmatrix} u^{(4)}(x_1 + \theta_1 h) \\ \vdots \\ u^{(4)}(x_n + \theta_n h) \end{pmatrix}.$$

Ako sada pogledamo apsolutne vrijednosti po komponenetama i uvažimo da je $0 \leq A^{-1} \leq \tilde{A}^{-1}$, $\max_i |u^{(4)}(x_i + \theta_i h)| \leq \|u^{(4)}\|_\infty$, dobijemo (koristeći (2.1.6))

$$\begin{pmatrix} |u_1 - u(x_1)| \\ \vdots \\ |u_n - u(x_n)| \end{pmatrix} \leq \frac{\|u^{(4)}\|_\infty}{12} h^4 \tilde{A}^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{\|u^{(4)}\|_\infty}{24} h^2 \begin{pmatrix} (x_1 - a)(b - x_1) \\ \vdots \\ (x_n - a)(b - x_n) \end{pmatrix}.$$

⊞

2.1.1 Varijacijska formulacija

Promatramo problem

$$\begin{aligned} &-(p(x)u'(x))' + q(x)u(x) = g(x),\ x \in [a,b] \\ &u(a) = \alpha,\ u(b) = \beta \end{aligned} \tag{2.1.7}$$

uz pretpostavke da je

$$\begin{aligned} &p \in \mathcal{C}^1[a,b],\ p(x) \geq p_0 > 0 \\ &q \in \mathcal{C}[a,b],\ \ q(x) \geq 0 \\ &f \in \mathcal{C}[a,b] \end{aligned} \tag{2.1.8}$$

Uočimo da je

$$\mathcal{C}^2[a,b] \ni v \mapsto L(v) \equiv -(pv')' + qv \in \mathcal{C}[a,b]$$

linearno preslikavanje pa gornju diferencijalnu jednadžbu možemo zapisati u obliku $L(v) = f$. Sada pogledajmo dodatne rubne uvjete. Ako je $\alpha = \beta = 0$, onda je

$$\mathcal{D}(L) \equiv \{v \in \mathcal{C}^2[a,b] \ : \ v(a) = v(b) = 0\} \subset \mathcal{C}^2[a,b]$$

vektorski potprostor, i ako onda L definiramo kao

$$L \ : \ \mathcal{C}^2[a,b] \supset \mathcal{D}(L) \longrightarrow \mathcal{C}[a,b] \tag{2.1.9}$$

onda naš polazni problem (2.1.7) postaje

$$L(u) = f, \ u \in \mathcal{D}(L) \tag{2.1.10}$$

Uvedimo skalarni produkt i pripadnu normu:

$$(u,v) \equiv \int_a^b u(x)v(x)dx, \quad \|u\|_2 = \sqrt{(u,u)}. \tag{2.1.11}$$

Teorem 2.1.3. *Operator L je u skalarnom produktu $(\cdot,\cdot)$ simetričan na $\mathcal{D}(L)$: za sve $u, v \in \mathcal{D}(L)$ vrijedi $(L(u), v) = (u, L(v))$.*

Dokaz: Provjera direktnim računom koristeći parcijalnu integraciju i definiciju prostora $\mathcal{D}(L)$. Vrijedi

$$\begin{aligned}
(u, L(v)) &= \int_a^b u(x)[-(p(x)v'(x))' + q(x)v(x)]dx \\
&= -u(x)p(x)v'(x)|_a^b + \int_a^b (p(x)v'(x)u'(x) + q(x)u(x)v(x))dx \\
&= \int_a^b (p(x)v'(x)u'(x) + q(x)u(x)v(x))dx = (v, L(u)) = (L(u), v),
\end{aligned}$$

jer je zadnji izraz simetričan u u i v. $\boxplus$

Iz prethodnog računa vidimo da je izraz $(L(u), v)$ definiran na većoj klasi funkcija – na primjer, za razliku od $L(u)$ uopće ne koristi drugu derivaciju.

Prije nego nastavimo, trebamo malu digresiju:

Definicija 2.1.1. Kažemo da je $f : [a, b] \to \mathbb{R}$ apsolutno neprekidna na $[a, b]$ ako za svaki $\epsilon > 0$ postoji $\delta > 0$ tako da za svaki konačan niz podintervala $[a_i, b_i]$ sa $a \leq a_1 < b_1 < a_2 < b_2 < a_3 < \cdots < a_n < b_n \leq b$ vrijedi

$$\sum_{i=1}^{n} |b_i - a_i| < \delta \Longrightarrow \sum_{i=1}^{n} |f(b_i) - f(a_i)| < \epsilon.$$

Komentar 2.1.1. Ako je f apsolutno neprekidna onda je ona i neprekidna i $f'(x)$ postoji gotovo svuda. Vrijedi i $f(x) = f(a) + \int_a^x f'(t)dt$. Ako su f i g apsolutno neprekidne onda vrijedi formula parcijalne integracije.

Definirajmo prostor funkcija

$$\mathcal{K}^m[a, b] = \{f : [a, b] \longrightarrow \mathbb{R} \ : \ f^{(m-1)} \text{ apsolutno neprekidna}, \ f^{(m)} \in \mathbf{L}^2[a, b]\}. \tag{2.1.12}$$

Vidimo da je $(L(u), v)$ dobro definiran i na skupu

$$\mathcal{D} = \{u \in \mathcal{K}^1[a, b] \ : \ u(a) = u(b) = 0\}$$

na kojem možemo definirati simetričnu bilinearnu formu

$$[u, v] = \int_a^b (p(x)u'(x)v'(x) + q(x)u(x)v(x))dx \tag{2.1.13}$$

Teorem 2.1.4. *Postoje pozitivne konstane $\gamma > 0$ i $\Gamma > 0$ tako da za svaki $u \in \mathcal{D}$ vrijedi*

$$\gamma \|u\|_\infty^2 \leq [u, u] \leq \Gamma \|u'\|_\infty^2. \tag{2.1.14}$$

Specijalno je, za svaki $u \in \mathcal{D}(L) \setminus \{\mathbf{0}\}$, $[u, u] \equiv (L(u), u) > 0$.

Dokaz: Uzmimo $u \in \mathcal{D}$. Vrijedi

$$\begin{aligned} u(x) &= u(a) + \int_a^x u'(t)dt = \int_a^x u'(t)dt \\ u(x)^2 &= \left(\int_a^x 1 \cdot u'(t)dt \right)^2 \leq \int_a^x 1^2 dt \int_a^x u'(t)^2 dt \leq (b - a) \int_a^b u'(t)^2 dt \end{aligned}$$

i zaključujemo da je

$$\|u\|_\infty^2 \leq (b-a)\int_a^b u'(t)^2 dt \leq (b-a)^2\|u'\|_\infty^2.$$

Sada koristeći $p(x) \geq p_0 > 0$ i $q(x) \geq 0$ imamo

$$[u,u] \;=\; \int_a^b (p(x)u'(x)^2 + q(x)u(x)^2)dx \geq p_0\int_a^b u'(x)^2dx \geq \frac{p_0}{b-a}\|u\|_\infty^2.$$

Također,

$$\begin{aligned}[u,u] &\leq \|p\|_\infty(b-a)\|u'\|_\infty^2 + \|q\|_\infty(b-a)\|u\|_\infty^2 \\ &\leq (\|p\|_\infty(b-a) + \|q\|_\infty(b-a)^3)\|u'\|_\infty^2\end{aligned}$$

pa (2.1.14) vrijedi sa

$$\gamma = \frac{p_0}{b-a}, \quad \Gamma = \|p\|_\infty(b-a) + \|q\|_\infty(b-a)^3.$$

⊞

Odmah vidimo da ovakav pristup daje elegantne dokaze. Na primjer, ako su $y_1, y_2 \in \mathcal{D}(L)$ dva rješenja problema (2.1.7), vrijedi $L(y_1) = L(y_2) = f$, pa je odmah $L(y_1 - y_2) = \mathbf{0}$. Iz Teorema 2.1.4 je

$$0 = (y_1 - y_2, L(y_1 - y_2)) \geq \gamma\|y_1 - y_2\|_\infty^2 \geq 0, \quad \text{pa je} \quad y_1 = y_2.$$

Sada definiramo funkcional

$$F : \mathcal{D} \longrightarrow \mathbb{R}, \quad F(u) = \frac{1}{2}[u,u] - (f,u) \tag{2.1.15}$$

Teorem 2.1.5. *Neka je $y \in \mathcal{D}(L)$ rješenje rubnog problema (2.1.7). Tada za svaki $u \in \mathcal{D}$, $u \neq y$ povlači $F(u) > F(y)$.*

Dokaz: Uzmimo $u \in \mathcal{D}$ i računajmo

$$\begin{aligned}F(u) &= \frac{1}{2}[u,u] - (L(y),u) = \frac{1}{2}([u,u] - 2[y,u] + [y,y] - [y,y]) \\ &= \frac{1}{2}([u-y,u-y] - [y,y]) > -\frac{1}{2}[y,y] = F(y).\end{aligned}$$

⊞

Gornji rezultat daje ideju da rješenje y pokušamo dobiti minimizacijom funkcionala F po $\mathcal{D}$. Uočimo da je $\mathcal{D} \supset \mathcal{D}(L)$.

Samu minimizaciju ćemo provesti aproksimativno, na sljedeći način: Odaberimo n–dimenzionalni potprostor $\mathcal{S} \subset \mathcal{D}$, $n < \infty$, i pokušajmo izračunati

$$u_{\mathcal{S}} = \arg\min_{u \in \mathcal{S}} F(u).$$

Da bi račun bio praktično i konkretno provediv, $\mathcal{S}$ zadajemo bazom $u_1, \ldots, u_n$, u kojoj je svaki $u \in \mathcal{S}$ prikazan s $u = \sum_{i=1}^{n} \xi_i u_i$. Na taj način je $\mathcal{S} \simeq \mathbb{R}^n$ pa minimiziramo

$$\begin{aligned} \Phi(\xi_1, \ldots, \xi_n) &= F(\sum_{j=1}^{n} \xi_j u_j) = \frac{1}{2}[\sum_{j=1}^{n} \xi_j u_j, \sum_{k=1}^{n} \xi_k u_k] - (f, \sum_{k=1}^{n} \xi_k u_k) \\ &= \sum_{j=1}^{n} \sum_{k=1}^{n} [u_k, u_j] \xi_k \xi_j + \sum_{k=1}^{n} \xi_k (f, u_k). \end{aligned}$$

Uz oznake

$$x = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}, \; A = \begin{pmatrix} [u_1, u_1] & \cdots & [u_1, u_n] \\ \vdots & \ddots & \vdots \\ [u_n, u_1] & \cdots & [u_n, u_n] \end{pmatrix}, \; b = \begin{pmatrix} (f, u_1) \\ \vdots \\ (f, u_n) \end{pmatrix} \tag{2.1.16}$$

imamo problem minimizacije na $\mathbb{R}^n$

$$\Phi(x) = \frac{1}{2} x^{\mathrm{T}} A x - x^{\mathrm{T}} b \longrightarrow \min$$

Teorem 2.1.6. *Problem minimizacije $F(u) \longrightarrow \min$, $u \in \mathcal{S}$, ima jedinstveno rješenje $u_{\mathcal{S}} = \sum_{i=1}^{n} \xi_i u_i$ u kojem su koeficijenti $x_* = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ određeni kao jedinstveno rješenje sustava jednadžbi $Ax_* = b$, sa b kao u (2.1.16).*

Ako je $y \in \mathcal{D}(L)$ rješenje problema (2.1.7), tj. $L(y) = f$, onda je

$$[u_{\mathcal{S}} - y, u_{\mathcal{S}} - y] = \min_{u \in \mathcal{S}} [u - y, u - y]. \tag{2.1.17}$$

Dokaz: Dokaz ide u četiri poteza:

- Matrica $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ definirana kao u (2.1.16) je pozitivno definitna. To vidimo iz $x^{\mathrm{T}} A x = [u, u]$.

- $\nabla\Phi(x) = Ax - b$. To se lako dobije računanjem parcijalnih derivacija.
- Neka je $Ax_* = b$ i neka je $\delta x \in \mathbb{R}^n$ proizvoljan, $x = x_* + \delta x$.

$$\begin{aligned}\Phi(x) &= \frac{1}{2}(x_* + \delta x)^{\mathrm{T}} A(x_* + \delta x) - (x_* + \delta x)^{\mathrm{T}} b \\ &= \frac{1}{2}x_*^{\mathrm{T}} A x_* + \frac{1}{2}\delta x^{\mathrm{T}} A x_* + \frac{1}{2}x^{\mathrm{T}} A\delta x + \frac{1}{2}\delta x^{\mathrm{T}} A\delta x - x_*^{\mathrm{T}} b - \delta x^{\mathrm{T}} b \\ &= \Phi(x_*) + \frac{1}{2}\delta x^{\mathrm{T}} A\delta x.\end{aligned}$$

  Dakle, ako je $\delta x \neq \mathbf{0}$, $\Phi(x) > \Phi(x_*)$.

- I na kraju, primijetimo da vrijedi

$$[u - y, u - y] = 2F(u) + [y, y].$$

⊞

Teorem 2.1.7. *Uz oznake prethodnog teorema imamo*

$$\|u_{\mathcal{S}} - y\|_\infty \leq C\|u' - y'\|_\infty,\ u \in \mathcal{S} \text{ proizvoljan.} \tag{2.1.18}$$

Dokaz: Uzmimo $u \in \mathcal{S}$ i iskoristimo (2.1.14) da dobijemo

$$\gamma\|u_{\mathcal{S}} - y\|_\infty^2 \leq [u_{\mathcal{S}} - y, u_{\mathcal{S}} - y] \leq [u - y, u - y] \leq \Gamma\|u' - y'\|_\infty^2,$$

tj. $\|u_{\mathcal{S}} - y\|_\infty \leq \sqrt{\Gamma/\gamma}\|u' - y'\|_\infty$. ⊞

2.1.1.1 Primjer prostora $\mathcal{S}$: kubični splineovi

Na segmentu $[a, b]$ načinimo podjelu

$$\Delta: \quad a = x_0 < x_1 < x_2 < x_3 < \cdots < x_{n-1} < x_n = b. \tag{2.1.19}$$

Kažemo da je funkcija $s \equiv s_\Delta : [a, b] \to \mathbb{R}$ kubični spline ako je

- $s \in \mathcal{C}^2[a, b]$;
- Restrikcija od s na svaki $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, n-1$, je polinom stupnja najviše tri.

Neka $\mathcal{S}_\Delta$ označava prostor kubičnih splineova.

Teorem 2.1.8. *Neka je* $f \in \mathcal{C}^4[a,b]$ *i* $\|f^{(4)}\|_\infty \leq K_f$ *i neka je u podjeli (2.1.19)* $h = \max_j(x_{j+1} - x_j)$, $\kappa_\Delta = h/\min_j(x_{j+1} - x_j)$. *Ako je s interpolacijski spline za funckiju f* ($s(x_i) = f(x_i)$, $s'(x) = f'(x)$ *za* $x \in \{a,b\}$) *onda postoje konstante* $c_k \leq 2$ *tako da za svaki* $x \in [a,b]$ *vrijedi*

$$|f^{(k)}(x) - s^{(k)}(x)| \leq c_k K_f h^{4-k},\ k = 0,1,2,3$$

Pri tome je $c_0 = 5/384$, $c_1 = 1/24$, $c_2 = 3/8$, $c_3 = (\kappa_\Delta + 1/\kappa_\Delta)/2$

Odaberimo sada $\mathcal{S} = \{s \in \mathcal{S}_\Delta \ :\ s(a) = s(b) = 0\}$. Očito je $\mathcal{S} \subset \mathcal{D}(L) \subset \mathcal{D}$.

Teorem 2.1.9. *Neka je egzaktno rješenje y problema (2.1.7) sa (2.1.8) klase* $\mathcal{C}^4[a,b]$. *Neka je* $\mathcal{S} = \{s \in \mathcal{S}_\Delta \ :\ s(a) = s(b) = 0\}$ *i* $u_\mathcal{S} = \arg\min_{u\in\mathcal{S}} F(u)$. *Tada je*

$$\|u_\mathcal{S} - y\|_\infty \leq \frac{1}{24}\sqrt{\frac{\Gamma}{\gamma}}\|y^{(4)}\|_\infty h^3.$$

Dokaz: Tvrdnju dokazujemo kombinacijom Teorema 2.1.7 i Teorema 2.1.8. Ako odaberemo spline u koji interpolira točno rješenje na način da je $u(x_i) = y(x_i)$, $i = 0,\ldots,n$, i $u'(a) = y'(a)$, $u'(b) = y'(b)$ onda prvo uočavamo da je takav $u \in \mathcal{S}$ pa je prema Teoremu 2.1.8

$$\|u' - y'\|_\infty \leq c_1\|y^{(4)}\|_\infty h^3.$$

Sada ocjena iz Teorema 2.1.7 daje tvrdnju. ⊞

Poglavlje 3

Numeričko rješavanje parcijalnih diferencijalnih jednadžbi

3.1 Paraboličke jednadžbe

U ovoj sekciji proučavamo numeričko rješavanje parcijalne diferencijalne jednadžbe

$$\frac{\partial u}{\partial t}(x,t) - a(x,t)\frac{\partial^2 u}{\partial x^2}(x,t) - b(x,t)\frac{\partial u}{\partial x}(x,t) - c(x,t)u(x,t) = f(x,t) \tag{3.1.1}$$

sa početnim uvjetom

$$u(x,0) = g(x), \quad x \in [0,1], \tag{3.1.2}$$

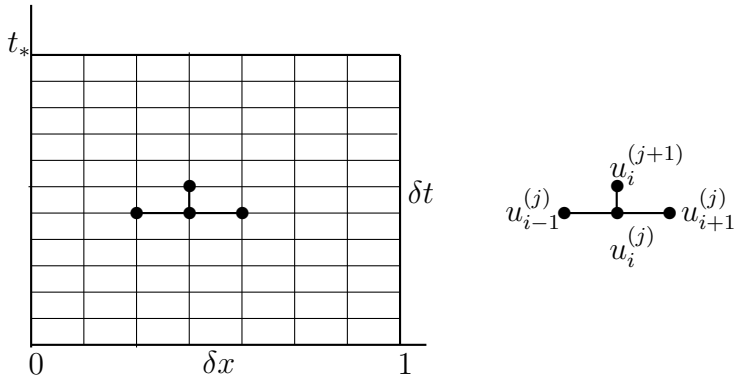
te rubnim uvjetima

$$\left.\begin{aligned} \alpha_0(t)\frac{\partial u}{\partial x}(0,t) + \beta_0(t)u(0,t) &= \rho_0(t) \\ \alpha_1(t)\frac{\partial u}{\partial x}(1,t) + \beta_1(t)u(1,t) &= \rho_1(t) \end{aligned}\right\} \; t \geq 0. \tag{3.1.3}$$

Mi ćemo zbog jednostavnosti veći dio posla napraviti na jednostavnijem obliku jednadžbe,

$$\frac{\partial u}{\partial t}(x,t) = \frac{\partial^2 u}{\partial x^2}(x,t), \tag{3.1.4}$$

a za detaljniju analizu ćemo odabrati samo dva specijalna tipa rubnih uvjeta. Pokazat će se da i ovako jednostavna inačica problema generira dovoljno zanimljivih problema koji motoviraju razvoj kompliciranе teorijske analize.



Slika 3.1: Disretizacija domene $[0,1] \times [0,t_*]$ s koracima δx, δt.

3.1.1 Jednostavna diskretizacija

Analogno rješavanju ODJ, diferencijalnu jednadžbu ćemo napisati u određenim diskretnim čvorovima u domeni, koristeći konačne diferencije kao aproksimacije odgovarajućih parcijalnih derivacija. Za početak, prisjetimo se aproksimacija parcijalnih derivacija

$$\frac{\partial}{\partial x}u(x,t) = \frac{u(x+\delta x,t)-u(x-\delta x,t)}{2\delta x} + O((\delta x)^2), \tag{3.1.5}$$

$$\frac{\partial^2}{\partial x^2}u(x,t) = \frac{u(x-\delta x,t)-2u(x,t)+u(x+\delta x,t)}{(\delta x)^2} + O((\delta x)^2) \tag{3.1.6}$$

$$\frac{\partial}{\partial t}u(x,t) = \frac{u(x,t+\delta t)-u(x,t)}{\delta t} + O(\delta t). \tag{3.1.7}$$

Dakle, jednadžbu (3.1.4) u točki (x,t) iz unutrašnjosti domene možemo zapisati kao

$$\frac{u(x,t+\delta t)-u(x,t)}{\delta t} = \frac{u(x-\delta x,t)-2u(x,t)+u(x+\delta x,t)}{(\delta x)^2} + O((\delta x)^2) + O(\delta t). \tag{3.1.8}$$

Analogno možemo izvesti za jednadžbu (3.1.1).

Ako domenu $[0,1] \times [0,t_*]$ diskretiziramo mrežom (x_i,t_j), gdje je $x_i = i\delta x$ $(i = 0,1,\ldots,n)$, $t_j = j\delta t$ $(j = 0,1,\ldots,m)$, te ako sa $u_i^{(j)}$ označimo aproksimaciju vrijednosti $u(x_i,t_j)$ točnog rješenja, onda relaciju (3.1.8) u točki (x_i,t_j) približno

zapisujemo kao

$$\frac{u_i^{(j+1)} - u_i^{(j)}}{\delta t} = \frac{u_{i+1}^{(j)} - 2u_i^{(j)} + u_{i-1}^{(j)}}{(\delta x)^2}. \tag{3.1.9}$$

Komentar 3.1.1. Ako u jednadžbi (3.1.1) stavimo $a_i^{(j)} = a(x_i, t_j)$, $b_i^{(j)} = b(x_i, t_j)$, $c_i^{(j)} = c(x_i, t_j)$, te ako sve parcijalne derivacije aproksimiramo konačnim diferencijama, dobijemo

$$\frac{u_i^{(j+1)} - u_i^{(j)}}{\delta t} - a_i^{(j)} \frac{u_{i+1}^{(j)} - 2u_i^{(j)} + u_{i-1}^{(j)}}{(\delta x)^2} - b_i^{(j)} \frac{u_{i+1}^{(j)} - u_{i-1}^{(j)}}{2\delta x} - c_i^{(j)} = f_i^{(j)}. \tag{3.1.10}$$

Stavimo

$$\mathbf{c} = \frac{\delta t}{(\delta x)^2} \quad (\mathbf{c} \text{ zovemo Courantov broj.}) \tag{3.1.11}$$

Vidimo da imamo evoluciju aproksimacija u diskretnim vremenskim koracima

$$u_i^{(j+1)} = \mathbf{c}u_{i-1}^{(j)} + (1 - 2\mathbf{c})u_i^{(j)} + \mathbf{c}u_{i+1}^{(j)} \tag{3.1.12}$$

Sada moramo uključiti dodatne uvjete.

Inicijalni uvjet (3.1.2) daje

$$u_i^{(0)} \equiv u(x_i, 0) = g_i \equiv g(x_i), \quad i = 0, 1, \ldots, n. \tag{3.1.13}$$

Sada pogledajmo neke specijalne oblike rubnih uvjeta.
Neka je zadano da je $u(0,t) = \rho_0(t)$, $u(1,t) = \rho_1(t)$. Tada je $u_0^{(j)} \equiv u(0, t_j) = \rho_0(t_j)$ i $u_n^{(j)} \equiv u(1, t_j) = \rho_1(t_j)$ za sve j. To znači da u svakom vremenskom koraku j imamo $n-1$ nepoznanicu $u_1^{(j)}, \ldots, u_{n-1}^{(j)}$, te da su dva vremenska koraka za $j = 0, 1, 2, \ldots$ povezana relacijom

$$\begin{pmatrix} u_1^{(j+1)} \\ u_2^{(j+1)} \\ \vdots \\ u_{n-2}^{(j+1)} \\ u_{n-1}^{(j+1)} \end{pmatrix} = \underbrace{\begin{pmatrix} 1-2\mathbf{c} & \mathbf{c} & 0 & \cdots & 0 \\ \mathbf{c} & 1-2\mathbf{c} & \mathbf{c} & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & \mathbf{c} & 1-2\mathbf{c} & \mathbf{c} \\ 0 & & 0 & \mathbf{c} & 1-2\mathbf{c} \end{pmatrix}}_{T} \begin{pmatrix} u_1^{(j)} \\ u_2^{(j)} \\ \vdots \\ u_{n-2}^{(j)} \\ u_{n-1}^{(j)} \end{pmatrix} + \mathbf{c} \begin{pmatrix} \rho_0(t_j) \\ 0 \\ \vdots \\ 0 \\ \rho_1(t_j) \end{pmatrix} \tag{3.1.14}$$

u kojoj su vrijednosti za $j = 0$ dane inicijalnim uvjetom (3.1.13).

Neka je sada uvjet u lijevom rubu kao i ranije, $u(0,t) = \rho_0(t)$, a u desnom rubu domene neka je zadano da je $\partial_x u(1,t) = 0$, za $t > 0$. Dakle, vrijednosti $u_n^{(j)}$ nisu zadane pa ih kao nepoznanice moramo izračunati koristeći relaciju (3.1.12) sa $i = n$. Uočavamo da tada imamo varijablu $u_{n+1}^{(j)}$ koja nije definirana. U našoj shemi indeksu $n+1$ odgovara točka $x_{n+1} = x_n + \delta x \equiv 1 + \delta x$ koja je izvan domene $[0,1]$ i simetrično pozicionirana prema $x_{n-1} = 1 - \delta x$. Ako iskoristimo (3.1.5), onda rubni uvjet u desnom rubu glasi

$$0 = \frac{\partial}{\partial x} u(1,t) = \frac{u(1+\delta x, t) - u(1-\delta x, t)}{2\delta x} + O((\delta x)^2)$$

pa relaciju (3.1.12) sa $i = n$ možemo zapisati u obliku

$$u_n^{(j+1)} = 2\mathbf{c} u_{n-1}^{(j)} + (1 - 2\mathbf{c}) u_n^{(j)}.$$

Sve relacije zajedno glase

$$\begin{pmatrix} u_1^{(j+1)} \\ u_2^{(j+1)} \\ \vdots \\ u_{n-1}^{(j+1)} \\ u_n^{(j+1)} \end{pmatrix} = \underbrace{\begin{pmatrix} 1-2\mathbf{c} & \mathbf{c} & 0 & \cdots & 0 \\ \mathbf{c} & 1-2\mathbf{c} & \mathbf{c} & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & \mathbf{c} & 1-2\mathbf{c} & \mathbf{c} \\ 0 & & 0 & 2\mathbf{c} & 1-2\mathbf{c} \end{pmatrix}}_{\tilde{T}} \begin{pmatrix} u_1^{(j)} \\ u_2^{(j)} \\ \vdots \\ u_{n-1}^{(j)} \\ u_n^{(j)} \end{pmatrix} + \mathbf{c} \begin{pmatrix} \rho_0(t_j) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} \tag{3.1.15}$$

Obadvije metode (3.1.14) i (3.1.15) možemo generički zapisati u obliku

$$u^{(j+1)} = Au^{(j)} + b^{(j)},\ j = 0, 1, 2, \ldots \tag{3.1.16}$$

Matrica A ovisi o parametrima diskretizacije δt i δx. Vidimo da joj se prvi i zadnji redak mijenjaju ovisno o rubnim uvjetima.

Propozicija 3.1.1. *Neka je A bilo koja od matrica iz (3.1.14) i (3.1.15). Tada je $\|A\|_\infty = |1 - 2\mathbf{c}| + 2\mathbf{c}$. Ako je $\mathbf{c} \leq 1/2$, onda je $\mathrm{spr}(A) < 1$ i $\|A\|_\infty = 1$.*

Dokaz: Ako je $A = T$, onda je spektar od A realan jer je T simetrična. Ako je $A = \tilde{T}$ onda je spektar od A realan jer je $\tilde{T}$ slična simetričnoj matrici $D^{-1}\tilde{T}D$, $D = \mathrm{diag}(1, \ldots, 1, \sqrt{2})$. U oba slučaja su Geršgorinovi krugovi centrirani u $1 - 2\mathbf{c}$, i također su u oba slučaja radijusi krugova $\mathbf{c}$ ili $2\mathbf{c}$, tako da je u oba slučaja krugovima pokriven isti dio kompleksne ravnine. Zaključujemo da su u oba slučaja svojstvene vrijednosti matrice A sadržane u $[1-4\mathbf{c}, 1]$. Sada ćemo isključiti mogućnost

da su $1 - 4\mathbf{c}$ i 1 svojstvene vrijednosti od A. Naime, niti jedna od te dvije vrijednosti nije u (otvorenoj) unutrašnjosti niti jednog Geršgorinovog kruga. Ako bi neka od te dvije vrijednosti bila uistinu svojstvena vrijednost od A, onda bi, zbog toga što je A očito ireducibilna, ta svojstvena vrijednost morala biti na presjeku rubova svih Geršgorinovih krugova od A – a to očito nije slučaj. Dakle, za svaku svojstvenu vrijednost λ matrice A vrijedi $1 - 4\mathbf{c} < \lambda < 1$. Sada još primijetimo da je $\mathbf{c} \leq 1/2$ ekvivalentno s $1 - 4\mathbf{c} \geq -1$. $\boxplus$

Komentar 3.1.2. Dokaz prethodne propozicije pokazuje praktičnu korist, jednostavnost i eleganciju teorije Geršgorinovih krugova.

Definicija 3.1.1. Kažemo da je metoda (3.1.16) Lax–Richtmyer stabilna ako za svako zadano vrijeme $t_* > 0$ postoji konstanta $\gamma(t_*)$ sa svojstvom da za svaki indeks j, $j\delta t \leq t_*$ povlači $\|A^j\| \leq \gamma(t_*)$.

Komentar 3.1.3. U prethodnoj definiciji nismo specificirali matričnu normu. Zbog ekvivalentnosti normi je jasno da je u danoj situaciji dovoljno odabrati neku normu sa kojom su tehnički detalji najjednostavniji. Primijetimo također da za traženu omeđenost nije nužno (ali je poželjno) da bude $\|A\| \leq 1$. Npr. $\|A\| \leq 1 + \eta\delta t$ povlači

$$\|A^j\| \leq \|A\|^j \leq (1 + \eta\delta t)^j \leq \mathbf{e}^{j\eta\delta t} \leq \mathbf{e}^{\eta t_*}.$$

Komentar 3.1.4. Promotrimo metodu (3.1.16) i uzmimo zbog jednostavnosti $b^{(j)} \equiv \mathbf{0}$ za sve j. tada je $u^{(j)} = A^j u^{(0)}$. Zamislimo sada malu perturbaciju inicijalnih uvjeta, $u^{(0)} \rightsquigarrow \tilde{u}^{(0)} = u^{(0)} + \delta u^{(0)}$. Novo rješenje je $\tilde{u}^{(j)} = A^j \tilde{u}^{(0)} = u^{(j)} + A^j \delta u^{(0)}$. Sada vidimo zašto nam je važna omeđenost niza potencija A^j.

Komentar 3.1.5. Valja primijetiti da uvjet omeđenosti niza A^j kada $\delta t \to 0$, $\delta x \to 0$, primjenjujemo na sve veće potencije ($j \to \infty$) te da dimenzije matrica $A = A(\delta t, \delta x)$ također rastu u beskonačno.

3.1.1.1 Konvergencija

Naravno, da bi numerička metoda bila korisna u primjenama, mora biti konvergentna – aproksimacije moraju težiti točnom rješenju kada δt, δx teže u nulu. Kako ćemo uskoro vidjeti, konvergencija može ovisiti i o tome kojim brzinama δt i δx teže u nulu.

Stavimo $\hat{u}_i^{(j)} = u(x_i, t_j)$, gdje je $u(x, t)$ egzaktno rješenje problema. Zanima nas lokalna pogreška diskretizacije – to je ostatak koji dobijemo kada u relaciju koja definira metodu uvrstimo točno rješenje,

$$\epsilon_i^{(j+1)} = u(x_i, t_{j+1}) - \mathbf{c}u(x_{i-1}, t_j) - (1 - 2\mathbf{c})u(x_i, t_j) - \mathbf{c}u(x_{i+1,j}).$$

Koristeći Taylorov razvoj dobijemo

$$\begin{aligned}
\epsilon_i^{(j+1)} &= u(x_i,t_j) + (\delta t)\partial_t u(x_i,t_j) + \frac{1}{2}(\delta t)^2\partial_{tt}^2 u(x_i,t_j) + \cdots \\
&- \mathbf{c}\left\{u(x_i,t_j) - (\delta x)\partial_x u(x_i,t_j) + \frac{1}{2}(\delta x)^2\partial_{xx}^2 u(x_i,t_j) - \frac{1}{6}(\delta x)^3\partial_{xxx}^3 u(x_i,t_j)\right. \\
&+ \left.\frac{1}{24}(\delta x)^4\partial_{xxxx}^4 u(x_i,t_j) - \cdots\right\} - (1-2\mathbf{c})u(x_i,t_j) - \\
&- \mathbf{c}\left\{u(x_i,t_j) + (\delta x)\partial_x u(x_i,t_j) + \frac{1}{2}(\delta x)^2\partial_{xx}^2 u(x_i,t_j) + \frac{1}{6}(\delta x)^3\partial_{xxx}^3 u(x_i,t_j)\right. \\
&+ \left.\frac{1}{24}(\delta x)^4\partial_{xxxx}^4 u(x_i,t_j) + \cdots\right\} = (\delta t)(\partial_t u(x_i,t_j) - \partial_{xx}^2 u(x_i,t_j)) \\
&+ \frac{(\delta t)^2}{2}\partial_{tt}^2 u(x_i,t_j) - \frac{(\delta t)(\delta x)^2}{12}\partial_{xxxx}^4 u(x_i,t_j) + \cdots \\
&= \frac{(\delta t)^2}{2}\partial_{tt}^2 u(x_i,t_j) - \frac{(\delta t)(\delta x)^2}{12}\partial_{xxxx}^4 u(x_i,t_j) + \cdots
\end{aligned}$$

Dakle, uz $\delta t = \mathbf{c}(\delta x)^4$, vrijedi

$$\hat{u}_i^{(j+1)} = \mathbf{c}\hat{u}_{i-1}^{(j)} + (1-2\mathbf{c})\hat{u}_i^{(j)} + \mathbf{c}\hat{u}_{i+1}^{(j)} + \epsilon_i^{(j)} \approx O((\delta t)^2) + O((\delta x)^4) = O((\delta x)^4). \tag{3.1.17}$$

Drugim riječima, ako bismo u vremenskom koraku j metodi dali egzaktne vrijednosti rješenja u svim čvorovima, pogreška u $(j+1)$–vom koraku bi bila reda veličine $(\delta t)\cdot O((\delta x)^2)$.
Uočimo da npr. u slučaju diskretizacije (3.1.14) možemo pisati, u matričnoj notaciji, za $j = 0, 1, \ldots$

$$\hat{u}^{(j+1)} = T\hat{u}^{(j)} + \underbrace{\begin{pmatrix}\rho_0(t_j)\\0\\\vdots\\0\\\rho_1(t_j)\end{pmatrix}}_{r^{(j)}} + \epsilon^{(j+1)}, \quad \hat{u}^{(j)} = \begin{pmatrix}\hat{u}_1^{(j)}\\\hat{u}_2^{(j)}\\\vdots\\\hat{u}_{n-1}^{(j)}\end{pmatrix}, \; \epsilon^{(j+1)} = \begin{pmatrix}\epsilon_1^{(j+1)}\\\epsilon_2^{(j+1)}\\\vdots\\\epsilon_{n-1}^{(j+1)}\end{pmatrix}. \tag{3.1.18}$$

Pri tome je, za svaki indeks j,

$$\|\epsilon^{(j)}\|_\infty \leq (\delta t)K(\delta x)^2.$$

Od sada uvijek pretpostavljamo da je odnos δt i δx konstantan i zadan Courantovim brojem $\mathbf{c}$. Zanima nas kako se ponašaju razlike $e_i^{(j)} = u_i^{(j)} - \hat{u}_i^{(j)}$ kada $\delta x \to 0$.

Teorem 3.1.2. *Promotrimo diskretizacijsku shemu (3.1.14) u kojoj je Courantov broj* $\mathbf{c}$ *konstantan kada* $\delta t \to 0$, $\delta x \to 0$, *i još je* $\mathbf{c} \leq 1/2$. *Neka je rješenje aproksimirano na vremenskom intervalu* $[0, t_*]$ *s korakom* $\delta t = \mathbf{c}(\delta x)^2$, *te neka je* $j_* = \lfloor t_*/\delta t \rfloor$. *Tada metoda (3.1.14) konvergira:*

$$\lim_{\delta x \to 0} \max_{j=0:j_*} \|e^{(j)}\|_\infty = 0.$$

Dokaz: Stavimo $e^{(j)} = u^{(j)} - \hat{u}^{(j)}$. Oduzimanjem relacija (3.1.14) i (3.1.18) dobivamo vezu $e^{(j+1)} = Te^{(j)} - \epsilon^{(j+1)}$ pa onda lako induktivno zaključimo da je za svaki j

$$e^{(j)} = T^j e^{(0)} - \sum_{i=0}^{j-1} T^i \epsilon^{(j-i)}.$$

Sada uzimanjem norme dobijemo

$$\begin{aligned} \|e^{(j)}\|_\infty &\leq \|T^j\|_\infty \|e^{(0)}\|_\infty + j \max_{i=0:j-1} \|T^i\|_\infty \max_{i=1:j} \|\epsilon^{(i)}\|_\infty \\ &\leq \gamma(t_*)\|e^{(0)}\|_\infty + j_*\gamma(t_*)\delta t \cdot K \cdot (\delta x)^2 \leq t_*\gamma(t_*) \cdot K \cdot (\delta x)^2. \end{aligned}$$

Ovdje smo iskoristili činjenicu da je inicijalna vrijednost rješenja (u vremenu $t = 0$) zadana pa je $e^{(0)} = 0$. Kako za $\mathbf{c} \leq 1/2$ vrijedi $\|T\|_\infty = 1$, imamo $\gamma(t_*) = 1$. ⊞

Komentar 3.1.6. Iz dokaza prethodnog teorema su jasno vidljive dvije komponente koje osiguravaju konvergenciju: (i) *Konzistentnost*: Lokalna pogreška diskretizacije zadovoljava $\|\epsilon^{(j)}\|/(\delta t) \to 0$ kada $\delta x \to 0$. (ii) *Stabilnost*: Niz $\|T^j\|$ je omeđen.

3.1.2 Veza sa ODJ: Metoda linija

Ako diferencijalnu jednadžbu (3.1.4) diskretiziramo samo po varijabli x, dobijemo na svim unutarnjim čvorovima i za svaki $t > 0$

$$\frac{\partial}{\partial t} u(x_i, t) = \frac{\partial^2}{\partial x^2} u(x_i, t) \approx \frac{u(x_{i-1}, t) - 2u(x_i, t) + u(x_{i-1}, t)}{(\delta x)^2}.$$

Za $i = 0, \ldots, n$ definirajmo funkcije $y_i(t) = u(x_i, t)$, $t > 0$.

Rubni uvjet $u(0, t) = \rho_0(t)$ odmah daje $y_0(t) = u(x_0, t) = u(0, t) = \rho_0(t)$. Desni rubni uvjet $u(1, t) = \rho_1(t)$ bi dao $y_n(t) = \rho_1(t)$, dok bismo npr. iz rubnog uvjeta $\partial_x u(1, t) = 0$ kao i ranije definirali $y_{n+1}(t) = y_{n-1}(t)$. Time smo dobili sustav

običnih diferencijalnih jednadžbi. U prvom slučaju,

$$\begin{aligned} y_1'(t) &= \frac{1}{(\delta x)^2}(-2y_1(t) + 2y_2(t) + \rho_0(t)) \\ y_i'(t) &= \frac{1}{(\delta x)^2}(y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)), \quad i = 2, \ldots, n-2 \\ y_{n-1}'(t) &= \frac{1}{(\delta x)^2}(y_{n-2}(t) - 2y_{n-1}(t) + \rho_1(t)) \end{aligned}$$

tj., matrično,

$$\begin{pmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_{n-2}'(t) \\ y_{n-1}'(t) \end{pmatrix} = \underbrace{\frac{1}{(\delta x)^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & 1 & -2 & 1 \\ 0 & & 0 & 1 & -2 \end{pmatrix}}_{T_0} \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{n-2}(t) \\ y_{n-1}(t) \end{pmatrix} + \frac{1}{(\delta x)^2}\begin{pmatrix} \rho_0(t) \\ 0 \\ \vdots \\ 0 \\ \rho_1(t) \end{pmatrix}. \tag{3.1.19}$$

U drugom slučaju imamo

$$\begin{aligned} y_1'(t) &= \frac{1}{(\delta x)^2}(-2y_1(t) + 2y_2(t) + \rho_0(t)) \\ y_i'(t) &= \frac{1}{(\delta x)^2}(y_{i-1}(t) - 2y_i(t) + y_{i+1}(t)), \quad i = 2, \ldots, n-1 \\ y_n'(t) &= \frac{1}{(\delta x)^2}(2y_{n-1}(t) - 2y_n(t)), \end{aligned}$$

matrično zapisano kao

$$\begin{pmatrix} y_1'(t) \\ y_2'(t) \\ \vdots \\ y_{n-1}'(t) \\ y_n'(t) \end{pmatrix} = \underbrace{\frac{1}{(\delta x)^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & 1 & -2 & 1 \\ 0 & & 0 & 2 & -2 \end{pmatrix}}_{\tilde{T}_0} \begin{pmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{n-1}(t) \\ y_n(t) \end{pmatrix} + \frac{1}{(\delta x)^2}\begin{pmatrix} \rho_0(t) \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{3.1.20}$$

U svakom slučaju dobijemo sustav ODJ oblika

$$y'(t) = My(t) + b(t) \tag{3.1.21}$$

sa inicijalnim uvjetima $y_i(0) = u(x_i, 0) = g(x_i)$ iz (3.1.2). Dobili smo još jedan primjer logične težnje matematičara da problem svedu na jedan kojeg su već naučili rješavati. Ako imamo dobar software za rješavanje sustava linearnih ODJ, možemo ga odmah uključiti u ovu shemu.

Nama je cilj i više od te ekonomičnosti – želimo dublje shvatiti problem. Sjetimo se svih metoda, analiza i problema koji se mogu javiti kada numerički rješavamo ODJ. Jasno je da pri rješavanju PDJ možemo očekivati samo više samo težih problema.

Komentar 3.1.7. Ovdje valja uočiti da sustav diferencijalnih jednadžbi ovisi o δx i to ne samo kroz množenje s $1/(\delta x)^2$. Kada $\delta x \to 0$ onda dimenzija matrice M raste zajedno sa brojem nepoznatih funkcija $y_i(t)$ tako da u analizi konvergencije treba očekivati više poteškoća nego kod rješavanja sustava ODJ.

Za početak, riješimo (3.1.21) Eulerovom metodom na diskretnoj mreži točaka $t_j = j\delta t$. Stavimo $y_i^{(j)} \approx y_i(t_j)$. Kako je $y_i(t_j) = u(x_i, t_j)$, možemo poistovijetiti $y_i^{(j)} \equiv u_i^{(j)}$. Eulerova metoda glasi

$$u^{(j+1)} = u^{(j)} + \delta t(Mu^{(j)} + b(t_j)) \equiv (I + \delta t M)u^{(j)} + \delta t\, b(t_j).$$

U slučaju sustava (3.1.19) imamo

$$I + \delta t M = I + \frac{\delta t}{(\delta x)^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & 1 & -2 & 1 \\ 0 & & 0 & 1 & -2 \end{pmatrix} = T, \quad \delta t\, b(t_j) = \frac{\delta t}{(\delta x)^2} \begin{pmatrix} \rho_0(t_j) \\ 0 \\ \vdots \\ 0 \\ \rho_1(t_j) \end{pmatrix}$$

dok je odgovarajuća shema za (3.1.20) definirana matricom

$$I + \delta t M = I + \frac{\delta t}{(\delta x)^2} \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & 1 & -2 & 1 \\ 0 & & 0 & 2 & -2 \end{pmatrix} = \tilde{T}.$$

Zaključujemo da je pristup opisan u §3.1.1 zapravo diskretizacija u varijabli x komponirana sa rješavanjem sustava običnih diferencijalnih jednadžbi pomoću Eulerove metode. To odmah inicira dvije stvari:

i) Problemi stabilnosti Eulerove metode se moraju manifestirati u metodi (3.1.16).

ii) Sustav ODJ možemo rješavati metodama boljim od Eulerove i tako dobiti bolju metodu za rješavanje polaznog problema.

3.1.2.1 Crank–Nicolsonova metoda

Prethodni komentari motiviraju sljedeći pristup: pokušajmo npr. (3.1.19) riješiti trapeznom metodom. Imali bismo

$$u^{(j+1)} = u^{(j)} + \frac{\delta t}{2}(T_0 u^{(j)} + \frac{1}{(\delta x)^2}\begin{pmatrix}\rho_0(t_j)\\0\\\vdots\\0\\\rho_1(t_j)\end{pmatrix} + T_0 u^{(j+1)} + \frac{1}{(\delta x)^2}\begin{pmatrix}\rho_0(t_{j+1})\\0\\\vdots\\0\\\rho_1(t_{j+1})\end{pmatrix})$$

pa je $u^{(j+1)}$ zadan kao rješenje sustava

$$\begin{pmatrix} 1+\mathbf{c} & -\mathbf{c}/2 & 0 & \cdots & 0 \\ -\mathbf{c}/2 & 1+\mathbf{c} & -\mathbf{c}/2 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & -\mathbf{c}/2 & 1+\mathbf{c} & -\mathbf{c}/2 \\ 0 & & 0 & -\mathbf{c}/2 & 1+\mathbf{c} \end{pmatrix} u^{(j+1)} = \begin{pmatrix} 1-\mathbf{c} & \mathbf{c}/2 & 0 & \cdots & 0 \\ \mathbf{c}/2 & 1-\mathbf{c} & \mathbf{c}/2 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & & \mathbf{c}/2 & 1-\mathbf{c} & \mathbf{c}/2 \\ 0 & & 0 & \mathbf{c}/2 & 1-\mathbf{c} \end{pmatrix} u^{(j)} + \frac{\mathbf{c}}{2}\begin{pmatrix}\rho_0(t_j)+\rho_0(t_{j+1})\\0\\\vdots\\0\\\rho_1(t_j)+\rho_1(t_{j+1})\end{pmatrix} \tag{3.1.22}$$

kojeg kompaktno zapisujemo u matričnom obliku kao

$$(I - \frac{\delta t}{2}T_0)u^{(j+1)} = (I + \frac{\delta t}{2}T_0)u^{(j)} + \begin{pmatrix}\rho_0(t_j)+\rho_0(t_{j+1})\\0\\\vdots\\0\\\rho_1(t_j)+\rho_1(t_{j+1})\end{pmatrix} \tag{3.1.23}$$

Dobili smo tzv. Crank–Nicolsonovu metodu, koja je implicitna i svaka nova iteracija je rješenje sustava linearnih jednadžbi. Za potrebe teorijske analize ćemo gornju relaciju zapisati i kao

$$u^{(j+1)} = (I - \frac{\delta t}{2}T_0)^{-1}(I + \frac{\delta t}{2}T_0)u^{(j)} + (I - \frac{\delta t}{2}T_0)^{-1}\begin{pmatrix}\rho_0(t_j)+\rho_0(t_{j+1})\\0\\\vdots\\0\\\rho_1(t_j)+\rho_1(t_{j+1})\end{pmatrix} \tag{3.1.24}$$

Propozicija 3.1.3. *Za proizvoljne $\delta t > 0$ i $\delta x > 0$ je spektralni radijus $\mathfrak{S}(C)$ matrice $C = (I - \frac{\delta t}{2}T_0)^{-1}(I + \frac{\delta t}{2}T_0)$ strogo manji od jedan.*

Dokaz: Sve svojstvene vrijednosti matrice C su oblika

$$\lambda(C) = \frac{1 + \frac{\delta t}{2}\lambda(T_0)}{1 - \frac{\delta t}{2}\lambda(T_0)}$$

gdje $\lambda(T_0)$ redom prolazi sve svojstvene vrijednosti od T_0. S druge strane, pomoću Geršgorinovih krugova možemo zaključiti da je $\mathfrak{S}(T_0) \subset (-4, 0)$ pa je onda sigurno $|\lambda(C)| < 1$. $\boxplus$

3.2 Eliptičke jednadžbe

Promatramo Poissonovu dvodimenzionalnu parcijalnu diferencijalnu jednadžbu

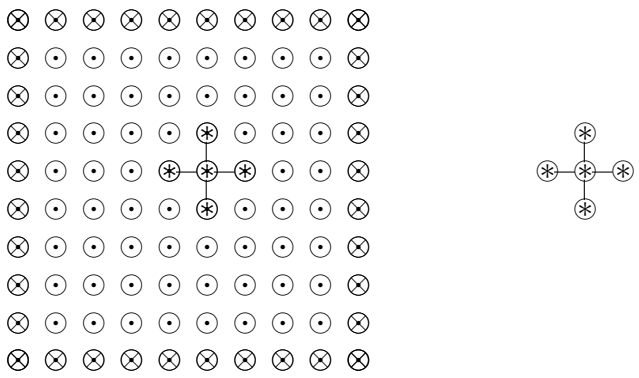
$$\begin{aligned} -u_{xx}(x,y) - u_{yy}(x,y) &= f(x,y), \quad (x,y) \in \Omega \\ u(x,y) &= 0, \quad (x,y) \in \partial\Omega \end{aligned}$$

gdje je $\Omega = \{(x,y) \,:\, 0 < x, y < 1\} \subset \mathbb{R}^2$ jedinični kvadrat, a $\partial\Omega$ njegov rub.

Diskretizaciju konstruiramo na sljedeći način: Odaberimo $n \in \mathbb{N}$ i stavimo

$$\begin{aligned} h &= \frac{1}{n+1}, \quad x_i = ih, \quad y_j = jh, \quad i,j = 0,1,2,\ldots,n+1 \\ \Omega_h &= \{(x_i,y_j) \,:\, i,j = 1,\ldots,n\}, \\ \partial\Omega_h &= \{(x_i,0),(x_i,1),(0,y_j),(1,y_j) \,:\, i,j = 0,1,\ldots,n+1\} \end{aligned}$$

Zamislimo diskretne točke $\Omega_h \bigcup \partial\Omega_h$ bačene kao mreža na zatvoreni kvadrat $\Omega \bigcup \partial\Omega$. Stavimo $u_{ij} = u(x_i,y_j)$, $i,j = 0,1,\ldots,n+1$. Sada $-u_{xx}(x_i,y_j)$ i $-u_{yy}(x_i,y_j)$ lako aproksimiramo centralnim diferencijama.



Slika 3.2: $-\triangle u(x_i,y_j) \approx \frac{4u_{ij}-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}}{h^2}$.

Koristeći prethodni primjer, za $(x_i,y_j) \in \Omega_h$ je

$$\begin{aligned} -u_{xx}(x_i,y_j) - u_{yy}(x_i,y_j) &= \frac{4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} \\ &+ e_{ij} \end{aligned}$$

gdje je e_{ij} pogreška diskretizacije koju možemo samo ocijeniti s $O(h^2)$. Zato e_{ij} zanemarimo (činimo pogrešku diskretizacije) i, s $f_{ij} = f(x_i,y_j)$, rješavamo sustav

linearnih jednadžbi

$$4v_{ij} - v_{i-1,j} - v_{i+1,j} - v_{i,j-1} - v_{i,j+1} = h^2 f_{ij}, \quad i,j = 1,\ldots,n \tag{3.2.1}$$
$$v_{0j} = v_{n+1,j} = v_{i0} = v_{i,n+1} = 0 \tag{3.2.2}$$

čije rješenje $V = (v_{ij})_{i,j=1}^{n}$ aproksimira $U = (u_{ij})$. Primijetimo da je prirodno vrijednosti u_{ij}, v_{ij} držati u matrici jer to odražava 2D strukturu. Ipak, za operativno računanje je koristan i generički zapis "$Ax = b$".

Elemente matrice V stavimo u vektor stupac po stupac, isto napravimo s $F = (f_{ij})$:

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nn} \end{pmatrix} \mapsto v = \begin{pmatrix} v_{11} \\ \vdots \\ v_{n1} \\ v_{12} \\ \vdots \\ v_{n2} \\ \vdots \\ v_{1n} \\ \vdots \\ v_{nn} \end{pmatrix} \equiv \mathrm{vec}(V)$$

Ovdje linearni operator $\mathrm{vec}(\cdot)$ poistovjećuje vektorske prostore $\mathbb{R}^{n\times n}$ i $\mathbb{R}^{n^2}$. Sada naš sustav jednadžbi možemo zapisati kao

$$\mathsf{T}_{\otimes n} v = h^2 \mathrm{vec}(F)$$

gdje je

$$\mathsf{T}_{\otimes n} = \left(\begin{array}{cccc|cccc|cccc|cccc}
4 & -1 & & & -1 & & & & & & & & & & & \\
-1 & 4 & \ddots & & & -1 & & & & & & & & & & \\
 & \ddots & \ddots & -1 & & & \ddots & & & & & & & & & \\
 & & -1 & 4 & & & & -1 & & & & & & & & \\
\hline
-1 & & & & 4 & -1 & & & \ddots & & & & & & & \\
 & -1 & & & -1 & 4 & \ddots & & & \ddots & & & & & & \\
 & & \ddots & & & \ddots & \ddots & -1 & & & \ddots & & & & & \\
 & & & -1 & & & -1 & 4 & & & & \ddots & & & & \\
\hline
 & & & & \ddots & & & & \ddots & & & & -1 & & & \\
 & & & & & \ddots & & & & \ddots & & & & -1 & & \\
 & & & & & & \ddots & & & & \ddots & & & & \ddots & \\
 & & & & & & & \ddots & & & & \ddots & & & & -1 \\
\hline
 & & & & & & & & -1 & & & & 4 & -1 & & \\
 & & & & & & & & & -1 & & & -1 & 4 & \ddots & \\
 & & & & & & & & & & \ddots & & & \ddots & \ddots & -1 \\
 & & & & & & & & & & & -1 & & & -1 & 4
\end{array}\right) \tag{3.2.3}$$

Par komentara:

- $\begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix} = \mathsf{T}_n + 2\mathsf{I}_n$
- $\mathsf{T}_{\otimes n} = \mathsf{I}_n \otimes \mathsf{T}_n + \mathsf{T}_n \otimes \mathsf{I}_n$, gdje je $\otimes$ Kroneckerov produkt.
- Ako uzmemo da korak subdivizije h bude samo reda veličine 10^{-2} trebamo uzeti $n = 100$ i dobivamo $n^2 = 10000$ jednadžbi i isto toliko nepoznanica. Ako je problem trodimenzionalan, onda broj nepoznanica postaje $n^3 = 10^6$.
- U Matlab–u se $\mathsf{T}_{\otimes n}$ dobije pomoću `delsq(numgrid('S',n+2))`.

Definicija 3.2.1. Kažemo da je realna matrica A TST matrica (tridijagonalna simetrična Toeplitzova) ako je tridijagonalna, simetrična i konstantna duž svojih dijagonala,

$$\mathsf{A} = \begin{pmatrix} \alpha & \beta & & \\ \beta & \ddots & \ddots & \\ & \ddots & \ddots & \beta \\ & & \beta & \alpha \end{pmatrix}$$

Propozicija 3.2.1. *Neka je* A *TST matrica reda* n*. Tada su njene svojstvene vrijednosti dane formulama*

$$\lambda_i = \alpha + 2\beta \cos \frac{i\pi}{n+1},\ i = 1, \ldots, n.$$

Pripadni ortonormalni vektori $\mathsf{v}_1, \ldots, \mathsf{v}_n$ *su dani formulama*

$$\mathsf{v}_{ji} = \sqrt{\frac{2}{n+1}} \sin \frac{ji\pi}{n+1},\ \ j = 1, \ldots, n. \tag{3.2.4}$$

(Ovdje v_{ji} *označava* j*–tu komponentu od* v_i*.) Sve TST matrice međusobno komutiraju.*

Korolar 3.2.2. *Svojstvene vrijednosti matrice* T_n *su*

$$\lambda_i = 2(1 - \cos \frac{i\pi}{n+1}),\ i = 1, \ldots, n,$$

a pripadni svojstveni vektori su dani formulama (3.2.4).

Definicija 3.2.2. Matricu $S = \mathsf{I}_m \otimes A + B \otimes \mathsf{I}_n$ zovemo Kroneckerov zbroj matrica $A \in \mathbb{M}_n$ i $B \in \mathbb{M}_n$, u oznaci $S = A \oplus B$.

Propozicija 3.2.3. *Ako su* $Au_j = \alpha_j u_j$*,* $Bv_i = \beta_i v_i$ *svojstvene vrijednosti i vektori, onda su svojstvene vrijednosti od* $A \otimes B$ *svi produkti* $\alpha_i \cdot \beta_j$*, a pripadni svojstveni vektori su* $u_i \otimes v_j$*.*

Dokaz: Koristimo svojstva Kroneckerovog produkta. Lako se provjeri da općenito vrijedi $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Sada imamo

$$\begin{aligned}(A \otimes B)(u_i \otimes v_j) &= (Au_i) \otimes (Bv_j) = (\alpha_i u_i) \otimes (\beta_j v_j) \\ &= (\alpha_i \beta_j)(u_i \otimes v_j).\end{aligned}$$

⊞

Propozicija 3.2.4. *Ako su* $Au_j = \alpha_j u_j$*,* $Bv_i = \beta_i v_i$ *svojstvene vrijednosti i vektori, onda su svojstvene vrijednosti od* $A \oplus B$ *svi zbrojevi* $\alpha_j + \beta_i$*, a pripadni svojstveni vektori su* $v_i \otimes u_j$*.*

Dokaz:

$$\begin{aligned}&(\mathsf{I} \otimes A + B \otimes \mathsf{I})(v_i \otimes u_j) = (\mathsf{I} \otimes A)(v_i \otimes u_j) + (B \otimes \mathsf{I})(v_i \otimes u_j) \\ &= (v_i \otimes \alpha_j u_j) + (\beta_i v_i \otimes u_j) = \alpha_j v_i \otimes u_j + \beta_i v_i \otimes u_j \\ &= (\alpha_j + \beta_i)(v_i \otimes u_j).\end{aligned}$$

⊞

Korolar 3.2.5. *Svojstvene vrijednosti matrice* $\mathsf{T}_{\otimes n} = \mathsf{I}_n \otimes \mathsf{T}_n + \mathsf{T}_n \otimes \mathsf{I}_n$ *su*

$$\begin{aligned} \lambda_{ij} &= 4 - 2(\cos\frac{i\pi}{n+1} + \cos\frac{j\pi}{n+1}) \\ &= 4(\sin^2\frac{i\pi}{2(n+1)} + \sin^2\frac{j\pi}{2(n+1)}),\ i,j = 1,\ldots,n, \end{aligned}$$

tj. $\lambda_{ij} = \lambda_i + \lambda_j$*, gdje su* $\lambda_1,\ldots,\lambda_n$ *svojstvene vrijednosti matrice* T_n*.*

Dokaz:
⊞

3.3 Hiperboličke jednadžbe

Dio II

Aproksimacija funkcija

3.4 Diskretna Fourierova transformacija

3.4.1 Trigonometrijska interpolacija

U primjenama često imamo uzorke (mjerenja) veličine koja je periodička, poznatog perioda T. Neka su podaci (x_k, f_k) dobiveni nad ekvidistantnom mrežom realnih čvorova $x_k = kT/n \in [0, T]$, $k = 0, 1, \ldots, n$, pri čemu su vrijednosti f_k općenito kompleksni brojevi. Razumno je takve podatke pokušati analitički reprezentirati periodičkom fukcijom $\Phi(x)$, perioda T, pri čemu za princip aproksimacije možemo uzeti interpolaciju,

$$\Phi(x_k) = f_k,\ k = 0, 1, \ldots, n-1. \tag{3.4.1}$$

Zbog pretpostavljene periodičnosti je $f_n = f_0$ i $\Phi(x_n) = \Phi(x_0)$. Ako su podaci f_k realni brojevi, onda zahtijevamo da je i funkcija Φ realna.

Ako funkciju Φ pretpostavimo kao linearnu kombinaciju sinusa i kosinusa

$$\Phi(x) = \frac{a_0}{2} + \sum_{j=1}^{m} (a_j \cos jx\frac{2\pi}{T} + b_j \sin jx\frac{2\pi}{T}), \tag{3.4.2}$$

onda imamo $2m + 1$ slobodnih parametara s kojima treba zadovoljiti n uvjeta interpolacije (3.4.1) koji glase

$$\frac{a_0}{2} + \sum_{j=1}^{m} (a_j \cos jk\frac{2\pi}{n} + b_j \sin jk\frac{2\pi}{n}) = f_k,\ \ k = 0, 1, \ldots, n-1. \tag{3.4.3}$$

Ako je n neparan i $m = \lfloor n/2 \rfloor$, tj. $n = 2m + 1$, onda nam ovakav Φ odgovara i pokušat ćemo odrediti parametre a_j, b_j. Ako je n paran i $m = n/2$, tj. $n = 2m$, onda u (3.4.2) stavljamo npr. $b_m \equiv 0$, tako da i dalje imamo n slobodnih parametara.

Ako uzmemo npr. $n = 2m$ i sa funkcijom (3.4.2) napišemo uvjete interpolacije (3.4.1) dobijemo, uz oznaku $\tau_k = x_k 2\pi/T = k2\pi/n$, $n \times n$ linearni sustav

$$\begin{pmatrix}
\frac{1}{2} & \cos\tau_0 & \sin\tau_0 & \cos 2\tau_0 & \sin 2\tau_0 & \cdots & \cos(m-1)\tau_0 & \sin(m-1)\tau_0 & \cos m\tau_0 \\
\frac{1}{2} & \cos\tau_1 & \sin\tau_1 & \cos 2\tau_1 & \sin 2\tau_1 & \cdots & \cos(m-1)\tau_1 & \sin(m-1)\tau_1 & \cos m\tau_1 \\
\frac{1}{2} & \cos\tau_2 & \sin\tau_2 & \cos 2\tau_2 & \sin 2\tau_2 & \cdots & \cos(m-1)\tau_2 & \sin(m-1)\tau_2 & \cos m\tau_2 \\
\frac{1}{2} & \cos\tau_3 & \sin\tau_3 & \cos 2\tau_3 & \sin 2\tau_3 & \cdots & \cos(m-1)\tau_3 & \sin(m-1)\tau_3 & \cos m\tau_3 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{1}{2} & \cos\tau_{n-2} & \sin\tau_{n-2} & \cos 2\tau_{n-2} & \sin 2\tau_{n-2} & \cdots & \cos(m-1)\tau_{n-2} & \sin(m-1)\tau_{n-2} & \cos m\tau_{n-2} \\
\frac{1}{2} & \cos\tau_{n-1} & \sin\tau_{n-1} & \cos 2\tau_{n-1} & \sin 2\tau_{n-1} & \cdots & \cos(m-1)\tau_{n-1} & \sin(m-1)\tau_{n-1} & \cos m\tau_{n-1}
\end{pmatrix}
\begin{pmatrix}
a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \\ a_{m-1} \\ b_{m-1} \\ a_m
\end{pmatrix}
=
\begin{pmatrix}
f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ \vdots \\ f_{n-2} \\ f_{n-1}
\end{pmatrix}, \tag{3.4.4}$$

čije rješenje bi trebalo dati nepoznate koeficijente. Analogno dobijemo u slučaju $n = 2m + 1$:

$$\begin{pmatrix} \frac{1}{2} & \cos\tau_0 & \sin\tau_0 & \cos 2\tau_0 & \sin 2\tau_0 & \cdots & \cos m\tau_0 & \sin m\tau_0 \\ \frac{1}{2} & \cos\tau_1 & \sin\tau_1 & \cos 2\tau_1 & \sin 2\tau_1 & \cdots & \cos m\tau_1 & \sin m\tau_1 \\ \frac{1}{2} & \cos\tau_2 & \sin\tau_2 & \cos 2\tau_2 & \sin 2\tau_2 & \cdots & \cos m\tau_2 & \sin m\tau_2 \\ \frac{1}{2} & \cos\tau_3 & \sin\tau_3 & \cos 2\tau_3 & \sin 2\tau_3 & \cdots & \cos m\tau_3 & \sin m\tau_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{2} & \cos\tau_{n-2} & \sin\tau_{n-2} & \cos 2\tau_{n-2} & \sin 2\tau_{n-2} & \cdots & \cos m\tau_{n-2} & \sin m\tau_{n-2} \\ \frac{1}{2} & \cos\tau_{n-1} & \sin\tau_{n-1} & \cos 2\tau_{n-1} & \sin 2\tau_{n-1} & \cdots & \cos m\tau_{n-1} & \sin m\tau_{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \\ a_{m-1} \\ b_{m-1} \\ a_m \\ b_m \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ f_3 \\ f_4 \\ \vdots \\ f_{n-2} \\ f_{n-1} \end{pmatrix}. \tag{3.4.5}$$

Naravno, želimo da su koeficijenti jedinstveno određeni za sve desne strane i poželjno je da ih možemo jednostavno i efikasno računati. Ključ je u specijalnoj strukturi matrice koeficijenata sustava.

Propozicija 3.4.1. *Neka A označava bilo koju od matrica u (3.4.4) i (3.4.5). Tada je $A = QD$, gdje je Q ortogonalna a D regularna dijagonalna matrica, pa je $A^{-1} = D^{-2}A^{\mathrm{T}}$. Euklidski skalarni produkti parova stupaca od A se računaju prema sljedećim formulama:*

- $\displaystyle\sum_{k=0}^{n-1} \cos j\tau_k = \begin{cases} 0, & za\ \dfrac{j}{n} \notin \mathbb{Z} \\ n, & za\ \dfrac{j}{n} \in \mathbb{Z} \end{cases}$; • $\displaystyle\sum_{k=0}^{n-1} \sin j\tau_k = 0$ *za sve* $j \in \mathbb{Z}$.

- $\displaystyle\sum_{k=0}^{n-1} \cos j\tau_k \cos \ell\tau_k = \begin{cases} 0, & za\ \dfrac{j+\ell}{n} \notin \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \notin \mathbb{Z} \\ \dfrac{n}{2} & za\ ili\ \dfrac{j+\ell}{n} \in \mathbb{Z}\ ili\ \dfrac{j-\ell}{n} \in \mathbb{Z} \\ n & za\ \dfrac{j+\ell}{n} \in \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \in \mathbb{Z} \end{cases}$

- $\displaystyle\sum_{k=0}^{n-1} \sin j\tau_k \sin \ell\tau_k = \begin{cases} 0 & za\ \dfrac{j+\ell}{n} \notin \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \notin \mathbb{Z}\ ili\ \dfrac{j+\ell}{n} \in \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \in \mathbb{Z} \\ -\dfrac{n}{2} & za\ \dfrac{j+\ell}{n} \in \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \notin \mathbb{Z} \\ \dfrac{n}{2} & za\ \dfrac{j+\ell}{n} \notin \mathbb{Z}\ i\ \dfrac{j-\ell}{n} \in \mathbb{Z} \end{cases}$

- $\displaystyle\sum_{k=0}^{n-1} \cos j\tau_k \sin \ell\tau_k = 0$ *za sve* $j, \ell \in \mathbb{N}_0$

Dokaz: Lako izračunamo

$$\sum_{k=0}^{n-1}(\cos j\tau_k + \mathrm{i}\sin j\tau_k) = \sum_{k=0}^{n-1}(\mathbf{e}^{\mathrm{i}\tau_k})^j = \sum_{k=0}^{n-1}(\mathbf{e}^{\mathrm{i}2\pi\frac{j}{n}})^k = \begin{cases} n,\ j/n \in \mathbb{Z} \\ 0,\ j/n \notin \mathbb{Z} \end{cases}.$$

Time su dokazane prve dvije formule. Za preostale koristimo dobro poznate formule

$$\begin{aligned} \cos j\phi\cos\ell\phi &= \frac{1}{2}(\cos(j+\ell)\phi + \cos(j-\ell)\phi) \\ \sin j\phi\sin\ell\phi &= \frac{1}{2}(\cos(j-\ell)\phi - \cos(j+\ell)\phi) \\ \cos j\phi\sin\ell\phi &= \frac{1}{2}(\sin(j+\ell)\phi - \sin(j-\ell)\phi). \end{aligned}$$

⊞

Korolar 3.4.2. *Koeficijenti interpolacije su dani formulama:*
Za $n = 2m+1$*:*

$$a_j = \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos jx_k\frac{2\pi}{T} = \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos\frac{2\pi jk}{n},\ j = 0,\dots,m, \tag{3.4.6}$$

$$b_j = \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin jx_k\frac{2\pi}{T} = \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin\frac{2\pi jk}{n},\ j = 1,\dots,m. \tag{3.4.7}$$

$$\tag{3.4.8}$$

Za $n = 2m$*:*

$$a_j = \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos jx_k\frac{2\pi}{T} = \frac{2}{n}\sum_{k=0}^{n-1} f_k \cos\frac{2\pi jk}{n},\ j = 0,\dots,m-1, \tag{3.4.9}$$

$$a_m = \frac{1}{n}\sum_{k=0}^{n-1} f_k \cos jx_k\frac{2\pi}{T} = \frac{1}{n}\sum_{k=0}^{n-1} f_k \cos\frac{2\pi jk}{n}, \tag{3.4.10}$$

$$b_j = \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin jx_k\frac{2\pi}{T} = \frac{2}{n}\sum_{k=0}^{n-1} f_k \sin\frac{2\pi jk}{n},\ j = 1,\dots,m. \tag{3.4.11}$$

$$\tag{3.4.12}$$

Dokaz: Matrica D iz Propozicije 3.4.1 ima sljedeću strukturu:

$$\begin{aligned} D^2 &= \left[\frac{n}{4}\right] \oplus \frac{n}{2}\mathsf{I}_{n-2} \oplus [n],\ \text{ za } n = 2m; \\ D^2 &= \left[\frac{n}{4}\right] \oplus \frac{n}{2}\mathsf{I}_{n-1},\ \text{ za } n = 2m+1, \end{aligned}$$

pa računanje koeficijenata pomoću inverza $A^{-1} = D^{-2}A^{\mathrm{T}}$ daje navedene formule. $\boxplus$

3.4.2 Računanje kompleksnom aritmetikom

Najelegantniji račun koeficijenata a_j, b_j se dobije prelaskom na kompleksnu aritmetiku. Naime, i sa sinusnim i sa kosinusnim članovima možemo računati istovremeno ako iskoristimo vezu sa eksponencijalnom funkcijom,

$$\cos\varphi = \frac{\mathbf{e}^{\mathrm{i}\varphi} + \mathbf{e}^{-\mathrm{i}\varphi}}{2}, \quad \sin\varphi = \frac{\mathbf{e}^{\mathrm{i}\varphi} - \mathbf{e}^{-\mathrm{i}\varphi}}{2\mathrm{i}},$$

iz koje odmah slijedi

$$\begin{aligned}
\Phi(x) &= \frac{a_0}{2} + \sum_{j=1}^{m}(a_j \cos jx\frac{2\pi}{T} + b_j \sin jx\frac{2\pi}{T}) \\
&= \frac{a_0}{2} + \sum_{j=1}^{m}(a_j \frac{\mathbf{e}^{\mathrm{i}jx2\pi/T} + \mathbf{e}^{-\mathrm{i}jx2\pi/T}}{2} + b_j \frac{\mathbf{e}^{\mathrm{i}jx2\pi/T} - \mathbf{e}^{-\mathrm{i}jx2\pi/T}}{2\mathrm{i}}) \\
&= \frac{a_0}{2} + \sum_{j=1}^{m}\frac{a_j - \mathrm{i}b_j}{2}\mathbf{e}^{\mathrm{i}jx2\pi/T} + \sum_{j=1}^{m}\frac{a_j + \mathrm{i}b_j}{2}\mathbf{e}^{-\mathrm{i}jx2\pi/T} \\
&= \sum_{j=-m}^{m} c_j \mathbf{e}^{\mathrm{i}jx2\pi/T}, \quad \text{gdje je} \quad \begin{cases} c_0 = a_0/2, \\ c_j = (a_j - \mathrm{i}b_j)/2, \ j = 1, \ldots, m, \\ c_{-j} = (a_j + \mathrm{i}b_j)/2, \ j = 1, \ldots, m. \end{cases}
\end{aligned}$$

Pri tome, u slučaju $n = 2m$, zbog $b_m = 0$, vrijedi $c_m = c_{-m} = a_m$. Primijetimo i da $a_j, b_j \in \mathbb{R}$ povlači $c_{-j} = \overline{c_j}$, te $a_j = 2\,\mathrm{Re}(c_j)$, $b_j = -2\,\mathrm{Im}(c_j)$.

Dobiveni izraz za $\Phi(x)$ ima dodatnu strukturu u čvorovima $x_k = kT/n$. Naime, vrijedi

$$\mathbf{e}^{-\mathrm{i}jx_k 2\pi/T} = \mathbf{e}^{-\mathrm{i}jk2\pi/n}\mathbf{e}^{\mathrm{i}nk2\pi/n} = \mathbf{e}^{\mathrm{i}(n-j)x_k 2\pi/T} \tag{3.4.13}$$

pa je

$$\Phi(x_k) = \sum_{j=0}^{m} c_j \mathbf{e}^{\mathrm{i}jx_k 2\pi/T} + \sum_{j=1}^{m} c_{-j}\mathbf{e}^{\mathrm{i}(n-j)x_k 2\pi/T}, \quad k = 0, 1, \ldots, n-1. \tag{3.4.14}$$

Dakle, uvjeti interpolacije u novim varijablama glase

$$\Phi(x_k) = \sum_{j=0}^{n-1}\gamma_j(\mathbf{e}^{\mathrm{i}x_k 2\pi/L})^j \equiv \sum_{j=0}^{n-1}\gamma_j(\mathbf{e}^{\mathrm{i}k2\pi/n})^j \equiv \sum_{j=0}^{n-1}\gamma_j\omega^{kj} = f_k, \quad k = 0, 1, \ldots, n-1. \tag{3.4.15}$$

gdje je $\omega = \mathbf{e}^{\mathrm{i}2\pi/n}$ i koeficijenti γ_j su definirani s:

$$\text{za } n = 2m+1: \quad \gamma_j = \begin{cases} c_j, & j = 0, 1, \ldots, m \\ c_{-(n-j)} & j = m+1, \ldots, n-1 \end{cases} \tag{3.4.16}$$

$$\text{za } n = 2m: \quad \gamma_j = \begin{cases} c_j & j = 0, 1, \ldots, m-1 \\ a_m & j = m \\ c_{-(n-j)} & j = m+1, \ldots, n-1 \end{cases} \tag{3.4.17}$$

Gornje relacije lako invertiramo, tj. iz danih koeficijenata γ_j lako izračunamo sve a_j i b_j:

$$\text{za } n = 2m+1: \quad \begin{cases} a_0 = 2\gamma_0 & a_j = \gamma_j + \gamma_{n-j} \quad j = 1, \ldots, m \\ & b_j = \mathrm{i}(\gamma_j - \gamma_{n-j}) \quad j = 1, \ldots, m \end{cases} \tag{3.4.18}$$

$$\text{za } n = 2m: \quad \begin{cases} a_0 = 2\gamma_0 & a_m = \gamma_m \\ a_j = \gamma_j + \gamma_{n-j} & j = 1, \ldots, m-1 \\ b_j = \mathrm{i}(\gamma_j - \gamma_{n-j}) & j = 1, \ldots, m-1 \end{cases} \tag{3.4.19}$$

Primijetimo da je (3.4.15) zapravo sustav linearnih jednadžbi kojeg matrično pišemo kao $\Omega\gamma = f$:

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-2} & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(n-2)} & \omega^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \omega^{n-2} & \omega^{2(n-2)} & \cdots & \omega^{(n-2)^2} & \omega^{(n-2)(n-1)} \\ 1 & \omega^{n-1} & \omega^{2(n-1)} & \cdots & \omega^{(n-1)(n-2)} & \omega^{(n-1)^2} \end{pmatrix}}_{\Omega} \underbrace{\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n-2} \\ \gamma_{n-1} \end{pmatrix}}_{\gamma} = \underbrace{\begin{pmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{n-2} \\ f_{n-1} \end{pmatrix}}_{f}. \tag{3.4.20}$$

Također uočimo polinom u varijabli $\zeta = \mathbf{e}^{\mathrm{i}x2\pi/T}$, $\varphi(\zeta) = \sum_{j=0}^{n-1} \gamma_j \zeta^j$, te da naš problem možemo interpretirati kao polinomijalnu interpolaciju $\varphi(\zeta_k) = f_k$ sa $\zeta_k = \mathbf{e}^{\mathrm{i}x_k 2\pi/T}$, $k = 0, \ldots, n-1$.

Propozicija 3.4.3. *Za matricu Ω iz (3.4.20) vrijedi $\Omega^*\Omega = \overline{\Omega}\Omega = n\mathrm{I}_n$. Specijalno je $\Omega^{-1} = \dfrac{1}{n}\overline{\Omega}$.*

Dokaz: Iz $\Omega_{kj} = \omega^{kj}$ slijedi da je Ω simetrična matrica, pa je $\Omega^* = \overline{\Omega^{\mathrm{T}}} = \overline{\Omega}$. Nadalje,

$$(\Omega^*\Omega)_{k\ell} = \sum_{j=0}^{n-1} \omega^{(\ell-k)j} = \frac{\omega^{(l-k)n} - 1}{\omega^{\ell-k} - 1} = 0$$

⊞

Korolar 3.4.4. *Koeficijenti γ_k su dani relacijom $(\gamma_k)_{k=0}^{n-1} = \frac{1}{n}\overline{\Omega}(f_k)_{k=0}^{n-1}$, tj.*

$$\gamma_k = \frac{1}{n}\sum_{j=0}^{n-1} f_j \mathbf{e}^{-\mathrm{i}jk2\pi/n}, \quad k = 0, 1, \ldots, n-1.$$

Pri tome vrijedi: Ako su svi $f_k \in \mathbb{R}$, onda je $\gamma_0 \in \mathbb{R}$ i $\gamma_{n-k} = \overline{\gamma_k}$, $k = 1, \ldots, n-1$.

Korolar 3.4.5. *Ako su koeficijenti γ_k izračunati kao u Korolaru 3.4.4, onda koeficijente a_k, b_k računamo prema formulama (3.4.18), (3.4.19), koje u slučaju realnih podataka f_k glase:*

- *Za $n = 2m+1$:*

$$\begin{aligned} a_k &= 2\,\mathrm{Re}(\gamma_k), \quad k = 0, \ldots, m \\ b_k &= -2\,\mathrm{Im}(\gamma_k), \quad k = 1, \ldots, m \end{aligned}$$

- *Za $n = 2m$:*

$$\begin{aligned} a_k &= 2\,\mathrm{Re}(\gamma_k), \quad k = 0, \ldots, m-1;\ a_m = \gamma_m \equiv \mathrm{Re}(\gamma_m) \\ b_k &= -2\,\mathrm{Im}(\gamma_k), \quad k = 1, \ldots, m-1;\ b_m = 0. \end{aligned}$$

Primjer 3.4.1. Primjer trigonometrijeske interpolacije (prema navedenim formulama):

Definicija 3.4.1. Preslikavanje $\mathcal{F} : \mathbb{C}^n \longrightarrow \mathbb{C}^n$ koje n–torci $f = (f_0, \ldots, f_{n-1})^{\mathrm{T}}$ pridružuje $\gamma = (\gamma_0, \ldots, \gamma_{n-1})^{\mathrm{T}} = \mathcal{F}(f)$ definiran s

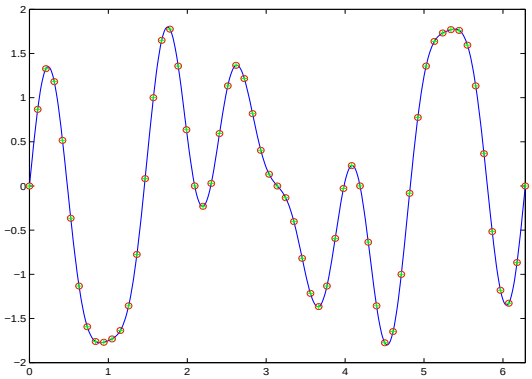
$$\gamma_k = \frac{1}{n}\sum_{j=0}^{n-1} f_j \mathbf{e}^{-\mathrm{i}jk2\pi/n}, \quad k = 0, 1, \ldots, n-1.$$

zovemo diskretna Fourierova transformacija (DFT). Matrični prikaz DFT je $\gamma = \frac{1}{n}\overline{\Omega} f$, gdje je Ω matrica definirana s (3.4.20).

Propozicija 3.4.6. *DFT je bijekcija na $\mathbb{C}^n$ i njen inverz, inverzna diskretna Fourierova transformacija (IDFT), je dana s $f = \mathcal{F}^{-1}\gamma = \Omega\gamma = \overline{\overline{\Omega}\overline{\gamma}}$, tj.*

$$f_k = \sum_{j=0}^{n-1} \gamma_j \mathbf{e}^{\mathrm{i}jk2\pi/n}, \quad k = 0, 1, \ldots, n-1.$$

Drugim riječima, $\mathcal{F}^{-1}(\gamma) = n\overline{\mathcal{F}(\overline{\gamma})}$.



Slika 3.3: Trigonometrijska interpolacija

Bibliografija