

1

Obične diferencijalne jednađbe

1.1 Uvod

U uvodnom dijelu ponavljamo neke definicije i teoreme iz teorije običnih diferencijalnih jednađbi.

Neka je $I_0 \subset \mathbb{R}$ interval i $\mathbf{f}: I_0 \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ neprekidna funkcija. Za zadano $t_0 \in I_0$ i $\mathbf{x}_0 \in \mathbb{R}^n$ treba naći derivabilnu funkciju $\mathbf{x}: I_0 \rightarrow \mathbb{R}^n$ koja zadovoljava

$$\forall t \in I_0, \quad \dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)), \quad (1.1)$$

$$\mathbf{x}(t_0) = \mathbf{x}_0. \quad (1.2)$$

Zadaću (1.1), (1.2) nazivamo Cauchyjeva zadaća za običnu diferencijalnu jednađbu (1.1). Funkcija $\mathbf{x}(t)$ je rješenje zadaće na intervalu I_0 .

U mnogim fizikalnim zadaćama varijabla t ima značenje vremena, pa stoga (1.2) nazivamo početnim uvjetom, a t_0 početnim trenutkom. Najčešće promatramo slučaj $I_0 = [t_0, T)$ ($T > t_0$) ili $I_0 = [t_0, \infty)$. Time se ne smanjuje općenitost problema jer se slučaj $I_0 = (t_0 - T, t_0]$ svodi na slučaj $I_0 = [t_0, t_0 + T)$ zamjenom varijabli $t \mapsto 2t_0 - t$. Slično, problem u kojem je $I_0 = (t_0 - T, t_0 + T)$ može se promatrati kao dva odvojena problema na intervalima $I_1 = (t_0 - T, t_0]$ i $I_2 = [t_0, t_0 + T)$.

Zadaća (1.1), (1.2) nema nužno rješenje na cijelom intervalu I_0 , na kojem je funkcija \mathbf{f} dobro definirana i neprekidna. Stoga se uvodi pojam **lokalnog rješenja**: Kažemo da je par (I, \mathbf{x}) , koji se sastoji od intervala $I \subset I_0$, $t_0 \in I$, i funkcije $\mathbf{x}: I \rightarrow \mathbb{R}^n$, lokalno rješenje Cauchyjeve zadaće (1.1), (1.2), ako je \mathbf{x} rješenje na intervalu I .

Za lokalno rješenje (J, \mathbf{x}^1) kažemo da proširuje lokalno rješenje (I, \mathbf{x}) , ako je $I \subset J$ i $\mathbf{x}^1(t) = \mathbf{x}(t)$ za svako $t \in I$. Ako je $J \neq I$, onda govorimo o strogom proširenju.

Lokalno rješenje (I, \mathbf{x}) je **maksimalno rješenje** zadaće (1.1), (1.2) ako ne postoji lokalno rješenje koje ga strogo proširuje. Maksimalno rješenje jednostavno zovemo rješenje.

Lokalno rješenje (I, \mathbf{x}) je globalno rješenje zadaće (1.1), (1.2) na intervalu I_0 , ako je $I = I_0$.

Existenciju lokalnog rješenja osigurava sljedeći teorem.

Teorem 1.1 (Cauchy-Péano) Neka je funkcija \mathbf{f} neprekidna na nekoj okolini točke $(t_0, \mathbf{x}_0) \in \mathbb{R}^{n+1}$. Tada postoji barem jedno lokalno rješenje zadaće (1.1), (1.2).

Pokazuje se da vrijedi:

Lema 1.1 Ako Cauchyjeva zadaća ima lokalno rješenje, onda ona ima barem jedno maksimalno rješenje koje ga proširuje.

Primjer 1. Zadaća

$$\begin{aligned}\dot{x}(t) &= -2tx(t)^2 \\ x(0) &= 1,\end{aligned}$$

ima jedinstveno globalno rješenje na \mathbb{R} ; $x(t) = 1/(1+t^2)$.

Primjer 2. Zadaća

$$\begin{aligned}\dot{x}(t) &= 2tx(t)^2 \\ x(0) &= 1,\end{aligned}$$

ima jedinstveno maksimalno rješenje na \mathbb{R} : $((-1, 1), x(t) = 1/(1-t^2))$. Globalno rješenje na \mathbb{R} ne postoji.

Primjer 3. Zadaća

$$\begin{aligned}\dot{x}(t) &= \sqrt[3]{x(t)^2} \\ x(0) &= 0,\end{aligned}$$

ima dva globalna rješenja na \mathbb{R} : $x \equiv 0$ i $x(t) = t^3/27$.

Kažemo da zadaća (1.1), (1.2) ima jedinstveno rješenje ako ima jedinstveno globalno rješenje i ako je svako lokalno rješenje restrikcija tog globalnog rješenja.

Jedinstvenost rješenja se osigurava uvjetom lipšicovosti.

Teorem 1.2 (Cauchy-Lipschitz) Neka je funkcija \mathbf{f} neprekidna na $I_0 \times \mathbb{R}^n$ i neka postoji konstanta L takva da je

$$\forall (t, \mathbf{x}^1), (t, \mathbf{x}^2) \in I_0 \times \mathbb{R}^n, \quad |\mathbf{f}(t, \mathbf{x}^1) - \mathbf{f}(t, \mathbf{x}^2)| \leq L|\mathbf{x}^1 - \mathbf{x}^2|.$$

Tada problem (1.1), (1.2), ima jedinstveno rješenje.

Uvjet iz teorema nazivamo uvjet lipšicovosti, a L Lipschitzovom konstantom.

1.2 Eulerova metoda

Zadana je skalarna Cauchyjeva zadaća

$$\dot{x}(t) = f(t, x(t)), \quad t > 0, \quad (1.3)$$

$$x(0) = x_0. \quad (1.4)$$

Pretpostavit ćemo da je $f: [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ neprekidna funkcija, Lipschitzova s konstantom L po drugoj varijabli, tj.,

$$\forall (t, x^1), (t, x^2) \in [0, T] \times \mathbb{R}, \quad |f(t, x^1) - f(t, x^2)| \leq L|x^1 - x^2|.$$

U nizu ekvidistantnih točaka $t_0 = 0$, $t_1 = t_0 + h$, $t_2 = t_1 + h$, ... želimo generirati niz vrijednosti x_i , $i = 0, 1, 2, \dots$ koje aproksimiraju rješenje Cauchyjeve zadaće (1.3), (1.4) u točkama t_i ($x_i \approx x(t_i)$).

Najjednostavniji način *diskretizacije* problema je taj da se derivacija zamijeni diferencijskim kvocijentom. Time dolazimo do postupka

$$\frac{x_{i+1} - x_i}{h} = f(t_i, x_i), \quad i = 0, 1, \dots$$

Početna vrijednost x_0 je zadana. To je Eulerova metoda.

Korak metode h ne mora nužno biti konstantan. Moguće je prema nekim kriterijima dinamički povećavati i smanjivati korak metode kako bi se postigla tražena točnost približnog rješenja. Metoda u tom slučaju ima oblik

$$x_{i+1} = x_i + h_i f(t_i, x_i), \quad i = 0, 1, 2, \dots$$

gdje je $t_{i+1} = t_i + h_i$ te $x_i \approx x(t_i)$. Za početak, mi ćemo se baviti metodama s konstantnim korakom.

Eulerova se metoda neposredno generalizira na sustave diferencijalnih jednadžbi. Ako je zadana funkcija $\mathbf{f}: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ i Cauchyjev problem

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)), \quad t > 0, \quad (1.5)$$

$$\mathbf{x}(0) = \mathbf{x}_0, \quad (1.6)$$

onda metoda glasi:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{f}(t_i, \mathbf{x}_i), \quad i = 0, 1, 2, \dots$$

Da bi se metoda primijenila na diferencijalne jednadžbe višeg reda, kao npr.

$$x^{(n)}(t) = f(t, x(t), \dot{x}(t), \dots, x^{(n-1)}(t)),$$

potrebno je prvo jednadžbu zapisati u obliku sustava prvog reda:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\dots\dots \\ \dot{x}_{n-1} &= x_n \\ \dot{x}_n &= f(t, x_1, x_2, \dots, x_n).\end{aligned}$$

Slično se postupa i sa sustavima višeg reda.

Napomena. Eulerovu smo metodu izveli aproksimacijom derivacije. Umjesto toga, mogli smo početi od integralne jednadžbe

$$x(t_{i+1}) - x(t_i) = \int_{t_i}^{t_{i+1}} f(t, x(t)) dt,$$

i aproksimirati integral:

$$\int_{t_i}^{t_{i+1}} f(t, x(t)) dt \approx hf(t_i, x(t_i)).$$

Time dobivamo isti numerički postupak. \square

Analizu Eulerove metode nije teško provesti. Pretpostavit ćemo da je točno rješenje Cauchyjevog problema $x \in C^2([0, T])$, za neki $T > 0$. Razvojem u Taylorov red dobivamo

$$\begin{aligned}x(t_{i+1}) &= x(t_i) + h\dot{x}(t_i) + \frac{h^2}{2}\ddot{x}(\tau) \\ &= x(t_i) + hf(t_i, x(t_i)) + \frac{h^2}{2}\ddot{x}(\tau),\end{aligned}\tag{1.7}$$

za neki $\tau \in (t_i, t_{i+1})$. Izraz

$$\varepsilon_i = x(t_{i+1}) - x(t_i) - hf(t_i, x(t_i))$$

nazvamo lokalna greška diskretizacije. Ona je ostatak koji se dobiva kada se točno rješenje Cauchyjevog problema uvrsti u numeričku metodu. U slučaju Eulerove metode iz (1.7) vidimo da je lokalna greška diskretizacije proporcionalna s h^2 . Preciznije, ako je $M = \max\{|\ddot{x}(t)| : t \in [0, T]\}$, onda je

$$|\varepsilon_i| \leq \frac{1}{2}Mh^2, \quad \forall i.\tag{1.8}$$

Pogledajmo sada kako evoluirala greška metode. U trenutku t_i greška metode je

$$e_i = x(t_i) - x_i,$$

gdje je $x(t)$ egzaktno rješenje Cauchyjeve zadaće. Imamo

$$\begin{aligned} e_{i+1} &= x(t_{i+1}) - x_{i+1} \\ &= \varepsilon_i + x(t_i) + hf(t_i, x(t_i)) - x_i - hf(t_i, x_i) \\ &= \varepsilon_i + e_i + h[f(t_i, x(t_i)) - hf(t_i, x_i)]. \end{aligned}$$

Treći član koji se pojavljuje u ovoj jednakosti možemo ocijeniti na osnovu pretpostavke uniformne lipšicovosti funkcije f po drugom argumentu. Time dobivamo

$$|e_{i+1}| \leq |\varepsilon_i| + (1 + Lh)|e_i| \leq (1 + Lh)|e_i| + \frac{1}{2}Mh^2. \quad (1.9)$$

Lema 1.2 (Diskretna Gronwallova lema) Neka je $h > 0$, $L > 0$ i $b \geq 0$ te neka je (x_i) , $i = 0, 1, 2, \dots$ niz nenegativnih brojeva koji zadovoljava

$$x_{i+1} \leq (1 + Lh)x_i + b.$$

Tada je za sve $n \geq 0$

$$x_n \leq x_0 e^{Lnh} + \frac{e^{Lnh} - 1}{Lh} b.$$

Dokaz. Jedostavnim iteriranjem i korištenjem nejednakosti $1 + a \leq e^a$, koja vrijedi za $a \geq 0$. \square

Primjenom Gronwallove leme na (1.8) dobivamo,

$$|e_n| \leq |e_0| e^{Lnh} + \frac{e^{Lnh} - 1}{2Lh} Mh^2 = \frac{e^{Lnh} - 1}{2L} Mh.$$

Time smo dobili zaključak:

Teorem 1.3 Neka je funkcija $f: [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ neprekidna i Lipschitzova s konstantom L po drugoj varijabli. Neka je $x \in C^2([0, T])$ rješenje Cauchyjeve zadaće (1.3), (1.4) i neka je $M = \max\{|\ddot{x}(t)|: t \in [0, T]\}$. Ako je $h = T/N$ i ako su x_i , $i = 0, 1, 2, \dots, N$ vrijednosti generirane Eulerovom metodom, onda je za sve $i = 1, 2, \dots, N$

$$|x(t_i) - x_i| \leq \frac{Mh}{2L} (e^{Lt_i} - 1).$$

Teorem 1.3 pretpostavlja da se računanje vrši u egzaktnoj aritmetici. Ako su prisutne greške zaokruživanja, onda računamo vrijednosti

$$\begin{aligned} \xi_0 &= x_0 + \delta_0 \\ \xi_{i+1} &= \xi_i + hf(t_i, \xi_i) + \delta_{i+1}, \quad i = 0, 1, 2, \dots \end{aligned}$$

gdje je δ_i greška zaokruživanja u i -tom koraku. Greška metode je sada

$$\begin{aligned} x(t_{i+1}) - \xi_{i+1} &= x(t_i) + h\dot{x}(t_i) + \frac{1}{2}h^2\ddot{x}(\tau_i) - [\xi_i + hf(t_i, \xi_i) + \delta_{i+1}] \\ &= x(t_i) - \xi_i + h[f(t_i, x(t_i)) - f(t_i, \xi_i)] + \frac{1}{2}h^2\ddot{x}(\tau_i) - \delta_{i+1}. \end{aligned}$$

Uvedimo oznaku $e_i = x(t_i) - \xi_i$, iskoristimo lipšicovost funkcije f i uniformu ograničenost druge derivacije rješenja. Izlazi

$$|e_{i+1}| \leq (1 + Lh)|e_i| + \frac{1}{2}Mh^2 + |\delta_{i+1}|.$$

Ako pretpostavimo da su greške zaokruživanja uniformno ograničene, odnosno da postoji $\delta > 0$, takav da je za svako $i \geq 0$

$$|\delta_i| \leq \delta,$$

onda ponovo možemo primijeniti Gronwallovu lemu i dobivamo

$$|e_i| \leq |e_0|e^{Lt_i} + \frac{e^{Lt_i} - 1}{Lh} \left(\frac{1}{2}Mh^2 + \delta \right).$$

Vidimo da se red metode ne mijenja ako je $\delta = O(h^2)$. Funkcija

$$h \mapsto \frac{Mh}{2L} + \frac{\delta}{Lh}$$

ima svoj minimum u $h^* = (2\delta/M)^{1/2}$, i manje korake ne treba birati jer tada dominiraju greške zaokruživanja.

Zadatak. Dokažite sljedeću verziju diskretne Gronwallove leme:

Lema 1.3 Neka je $h > 0$, $L > 0$ te neka su (x_i) , (b_i) $i = 0, 1, 2, \dots$ nizovi nenegativnih brojeva koji zadovoljavaju

$$x_{i+1} \leq (1 + Lh)x_i + b_i.$$

Tada je za sve $n \geq 0$

$$x_n \leq x_0 e^{Lnh} + \sum_{i=0}^{n-1} e^{L(n-i-1)h} b_i.$$

1.3 Runge-Kutta metode

Promatrajmo, radi jednostavnosti, Cauchyjev problem za skalarnu diferencijalnu jednadžbu

$$\dot{x}(t) = f(t, x(t)) \tag{1.10}$$

$$x(0) = x_0. \tag{1.11}$$

Eulerovu je metodu moguće izvesti polazeći od razvoja rješenja $x(t)$ u Taylorov red

$$x(t+h) = x(t) + h\dot{x}(t) + \frac{h^2}{2}\ddot{x}(t) + \frac{h^3}{6}\dddot{x}(t) + \dots$$

Uzimajući aproksimaciju $x(t+h) \approx x(t) + h\dot{x}(t)$ i uvažavajući da je $\dot{x}(t) = f(t, x(t))$, dobivamo

$$x(t+h) \approx x(t) + hf(t, x(t)),$$

što vodi na numerički postupak

$$x_{i+1} = x_i + hf(t_i, x_i).$$

To je Eulerova metoda. Istim postupkom možemo generirati metode višeg reda točnosti; potrebno je jedino zadržati veći broj članova u Taylorovom razvoju. Pokažimo to na primjeru metode drugog reda. Zadržat ćemo prva tri člana u razvoju, pa stoga koristimo jednakost

$$\begin{aligned} \ddot{x}(t) &= \frac{d}{dx}f(t, x(t)) = f_t(t, x(t)) + f_x(t, x(t))\dot{x}(t) \\ &= f_t(t, x(t)) + f_x(t, x(t))f(t, x(t)), \end{aligned}$$

gdje je $f_t = \frac{\partial f}{\partial t}$ itd. Time smo dobili aproksimaciju

$$x(t+h) \approx x(t) + hf(t, x(t)) + \frac{h^2}{2}[f_t(t, x(t)) + f_x(t, x(t))f(t, x(t))],$$

iz koje slijedi metoda

$$x_{i+1} = x_i + hf(t_i, x_i) + \frac{h^2}{2}[f_t(t_i, x_i) + f_x(t_i, x_i)f(t_i, x_i)]. \quad (1.12)$$

Slično bi se izvodile metode trećeg i višeg reda.

Metode ovog tipa ponekad se nazivaju **Taylorove metode**¹. Iz njihovog izvoda je jasno koji je red lokalne greške diskretizacije metode. Tako je za metodu (1.12) LGD reda $O(h^3)$. Preciznije,

$$\begin{aligned} x(t+h) &= x(t) + hf(t, x(t)) \\ &\quad + \frac{h^2}{2}[f_t(t, x(t)) + f_x(t, x(t))f(t, x(t))] + O(h^3). \end{aligned} \quad (1.13)$$

Nedostatak metode je što moramo znati parcijalne derivacije funkcije f . On se može ukloniti na ovaj način: uočimo da Taylorovim razvojem dobivamo

$$f(t+h, x(t) + hf(t, x(t))) = f(t, x(t)) + [f_t(t, x(t)) + f_x(t, x(t))f(t, x(t))]h + O(h^2),$$

¹Ili metode na osnovi Taylorovog razvoja.

odnosno,

$$h^2[f_t(t, x(t)) + f_x(t, x(t))f(t, x(t))] = h[f(t+h, x(t) + hf(t, x(t))) - f(t, x(t))] + O(h^3).$$

Sada vidimo da izraz s parcijalnim derivacijama funkcije f možemo, bez smanjenja reda LGD, zamijeniti izrazom bez parcijalnih derivacija:

$$x(t+h) = x(t) + hf(t, x(t)) + \frac{h}{2}[f(t+h, x(t) + hf(t, x(t))) - f(t, x(t))] + O(h^3).$$

To vodi na novu metodu

$$x_{i+1} = x_i + \frac{h}{2}[f(t_i, x_i) + f(t_i + h, x_i + hf(t_i, x_i))] \quad (1.14)$$

koja se naziva **modificirana Eulerova metoda** i zapisuje se u obliku

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + h, x_i + hm_1) \\ x_{i+1} &= x_i + \frac{h}{2}(m_1 + m_2). \end{aligned}$$

Dobivena metoda je jedna iz porodice Runge-Kutta metoda s lokalnom greškom diskretizacije trećeg reda.

Drugi pristup izvođenju metoda ovog tipa je putem generalizacije dobivenih formula. Tako, na primjer, formulu (1.14) možemo generalizirati na sljedeći način:

$$x_{i+1} = x_i + w_1 hf(t_i, x_i) + w_2 hf(t_i + \alpha h, x_i + \beta hf(t_i, x_i)),$$

gdje su w_1 , w_2 , α i β neke konstante. Pitamo se koje uvjete moraju zadovoljavati ove konstante da bi lokalna greška diskretizacije metode bila reda tri (skup takvih koeficijenata je očito neprazan). Preciznije, ako je $x(t)$ točno rješenje, želimo imati

$$x(t+h) = x(t) + w_1 hf(t, x(t)) + w_2 hf(t + \alpha h, x(t) + \beta hf(t, x(t))) + O(h^3).$$

Koristeći Taylorov razvoj, dobivamo

$$f(t + \alpha h, x(t) + \beta hf(t, x(t))) = f(t, x(t)) + [\alpha f_t(t, x(t)) + \beta f_x(t, x(t))f(t, x(t))]h + O(h^2),$$

pa prethodni izraz možemo zapisati u obliku

$$\begin{aligned} x(t+h) &= x(t) + (w_1 + w_2)hf(t, x(t)) \\ &\quad + w_2 h^2[\alpha f_t(t, x(t)) + \beta f_x(t, x(t))f(t, x(t))] + O(h^3) \end{aligned}$$

Usporedbom s (1.13) dobivamo sljedeće uvjete:

$$w_1 + w_2 = 1, \quad w_2\alpha = \frac{1}{2}, \quad w_2\beta = \frac{1}{2}.$$

Time dobivamo jednoparametarsku familiju metoda

$$w_1 = 1 - \frac{1}{2\alpha}, \quad w_2 = \frac{1}{2\alpha}, \quad \beta = \alpha.$$

Modificiranu Eulerovu metodu dobivamo uz izbor $\alpha = 1$. Drugi zanimljivi izbor je $\alpha = 1/2$, što vodi na metodu

$$x_{i+1} = x_i + hf(t_i + \frac{1}{2}h, x_i + \frac{1}{2}f(t_i, x_i)),$$

odnosno

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + \frac{h}{2}, x_i + \frac{h}{2}m_1) \\ x_{i+1} &= x_i + hm_2. \end{aligned}$$

To je midpoint metoda.

Zadatak. Nađite geometrijsku interpretaciju modificirane Eulerove i midpoint metode. \square

Generalizacijom gornjeg postupka dolazimo do ove definicije: Runge-Kutta metode su metode oblika

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + \alpha_2h, x_i + h\beta_{2,1}m_1) \\ &\vdots \\ m_n &= f(t_i + \alpha_nh, x_i + h\sum_{j=1}^{n-1} \beta_{n,j}m_j) \\ x_{i+1} &= x_i + h\sum_{j=1}^n w_jm_j. \end{aligned}$$

Na primjer, za $n = 3$ imamo

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + \alpha_2h, x_i + h\beta_{2,1}m_1) \\ m_3 &= f(t_i + \alpha_3h, x_i + h\beta_{3,1}m_1 + h\beta_{3,2}m_2) \\ x_{i+1} &= x_i + h(w_1m_1 + w_2m_2 + w_3m_3). \end{aligned}$$

U skladu s našim osnovnim primjerom ($n = 2$) prirodno je očekivati da se koeficijenti Runge-Kutta metode n -tog reda mogu odabrati tako da LGD bude reda $O(h^{n+1})$. Dapače, za svako n očekujemo da postoji čitava familija takvih metoda.

Zadatak. Pokažite da Runge-Kutta metoda trećeg reda ima lokalnu grešku diskretizacije reda $O(h^4)$ ako njezini koeficijenti zadovoljavaju sljedeće uvjete:

$$\begin{aligned} w_1 + w_2 + w_3 &= 1 \\ w_2\alpha_2 + w_3\alpha_3 &= \frac{1}{2} \\ w_2\alpha_2^2 + w_3\alpha_3^2 &= \frac{1}{3} \\ w_2\beta_{2,1} + w_3(\beta_{3,1} + \beta_{3,2}) &= \frac{1}{2} \\ w_2\alpha_2\beta_{2,1} + w_3\alpha_3(\beta_{3,1} + \beta_{3,2}) &= \frac{1}{3} \\ w_2\beta_{2,1}^2 + w_3(\beta_{3,1} + \beta_{3,2})^2 &= \frac{1}{3} \\ w_3\alpha_2\beta_{3,2} &= \frac{1}{6}, \quad w_3\beta_{2,1}\beta_{3,2} = \frac{1}{6}. \end{aligned}$$

Uvjerite se da je skup rješenja tog sustava neprazan. \square

Najčešće primijenjivana Runge-Kutta metoda je tzv. klasična Runge-Kutta metoda:

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f\left(t_i + \frac{1}{2}h, x_i + \frac{1}{2}hm_1\right) \\ m_3 &= f\left(t_i + \frac{1}{2}h, x_i + \frac{1}{2}hm_2\right) \\ m_4 &= f(t_i + h, x_i + hm_3) \\ x_{i+1} &= x_i + \frac{h}{6}(m_1 + 2m_2 + 2m_3 + m_4). \end{aligned}$$

Osnovna prednost metoda višeg reda je u tome što dozvoljavaju upotrebu većeg koraka h prilikom računanja rješenja. Na taj se način izbjegavaju greške zaokruživanja. S druge strane, u praksi se pokazuje da su metode višeg reda (posebno metoda četvrtog reda) numerički efikasnije. Na primjer, iako za RK4 metodu treba 4 računanja funkcije u svakom koraku, dok Eulerova metoda treba samo jedno, RK4 metoda će za istu točnost trebati daleko manji broj koraka.

Zadatak. Modificiranom Eulerovom, midpoint i klasičnom Runge-Kutta metodom izračunati rješenje zadaće:

$$\begin{aligned} \dot{x}(t) &= 5(t-1)x(t) \\ x(0) &= 5. \end{aligned}$$

Izračunajte točno rješenje. Za svaku pojedinu metodu eksperimentalno pronađite broj koraka potreban da se u $t = 1$ postigne točnost od 10^{-6} . \square

Zadatak. Klasičnom Runge-Kutta metodom riješiti problem gibanja harmonijskog oscilatora u polju sile teže:

$$\begin{aligned} m\ddot{x} + kx &= 0 \\ m\ddot{y} + ky &= -mg \\ x(0) &= x_0, \quad y(0) = y_0, \\ \dot{x}(0) &= 0, \quad \dot{y}(0) = 0. \end{aligned}$$

Prikažite grafički stazu materijalne točke za neki izbor parametara. \square

1.4 Jednokoračne metode

Promatramo Cauchijevu zadaću

$$\dot{x}(t) = f(t, x(t)) \quad (1.15)$$

$$x(0) = x_0, \quad (1.16)$$

gdje je $f: [0, T] \times \mathbb{R} \rightarrow \mathbb{R}$ neprekidna funkcija, Lipschitzova po drugoj varijabli, s konstantom lipšicovosti L .

Sve do sada uvedene numeričke metode su jednokoračne u smislu da za računanje vrijednosti x_{i+1} koriste samo prethodnu vrijednost x_i . Sve se takve metode mogu prikazati u obliku

$$x_{i+1} = x_i + h\Phi(t_i, x_i; h). \quad (1.17)$$

Pri tome zahtijevamo da je

$$\Phi: [0, T] \times \mathbb{R} \times [0, h^*] \rightarrow \mathbb{R}$$

neprekidna funkcija za neki $h^* > 0$.

Definicija 1.1 Kažemo da je metoda (1.17) konzistentna s jednadžbom (1.15) ako za svako rješenje $x = x(t)$ jednadžbe (1.15) vrijedi

$$\lim_{h \rightarrow 0} \sum_{i=0}^{N-1} |x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h)| = 0,$$

gdje je $T = Nh$.

Uočimo da je izraz

$$\varepsilon_i = x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h),$$

lokalna greška diskretizacije u točki x_i .

M. JURAK 31. svibnja 2006.

Definicija 1.2 Za metodu (1.17) kažemo da je stabilna ako postoji konstanta M , neovisna o h , takva da za svaka dva niza x_i i y_i , $i = 0, 1, \dots, N$, koji zadovoljavaju

$$\begin{aligned}x_{i+1} &= x_i + h\Phi(t_i, x_i; h), \\y_{i+1} &= y_i + h\Phi(t_i, y_i; h) + \varepsilon_i,\end{aligned}$$

vrijedi

$$\max_{0 \leq i \leq N} |x_i - y_i| \leq M \left[|x_0 - y_0| + \sum_{i=0}^{N-1} |\varepsilon_i| \right].$$

Stabilnost implicira da male promjene podataka povlače male promjene rješenja.

Definicija 1.3 Kažemo da je metoda (1.17) konvergentna ukoliko

$$\lim_{h \rightarrow 0} \max_{0 \leq i \leq N} |x(t_i) - x_i| = 0,$$

gdje je $x = x(t)$ rješenje zadatce (1.15)-(1.16), a x_i je dano metodom (1.17), uz početni uvjet x_0 i $T = Nh$.

Teorem 1.4 Ako je metoda (1.17) stabilna i konzistentna, onda je ona i konvergentna.

Dokaz. Uočimo da $y_i = x(t_i)$ zadovoljava jednadžbu

$$y_{i+1} = y_i + h\Phi(t_i, y_i; h) + \varepsilon_i,$$

gdje je

$$\varepsilon_i = x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h),$$

lokalna greška diskretizacije u točki x_i . Zbog istog početnog uvjeta, stabilnost nam daje

$$\max_{0 \leq i \leq N} |x_i - x(t_i)| \leq M \sum_{i=0}^{N-1} |\varepsilon_i|,$$

a desna strana teži u nulu po pretpostavci konzistentnosti. \square

Lema 1.4 Nužan i dovoljan uvjet da bi metoda (1.17) bila konzistentan je

$$\forall t \in [0, T], \forall x \in \mathbb{R}, \quad \Phi(t, x; 0) = f(t, x).$$

Dokaz. Lokalnu grešku diskretizacije

$$\varepsilon_i = x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h),$$

možemo zapisati pomoću Lagrangeovog teorema srednje vrijednosti u obliku

$$\varepsilon_i = h[f(c_i, x(c_i)) - \Phi(t_i, x(t_i); h)],$$

gdje je $c_i \in (t_i, t_{i+1})$ neka točka. Nadalje, možemo pisati

$$\varepsilon_i = \alpha_i + h\beta_i,$$

gdje je

$$\begin{aligned}\alpha_i &= h[f(c_i, x(c_i)) - \Phi(c_i, x(c_i); 0)], \\ \beta_i &= \Phi(c_i, x(c_i); 0) - \Phi(t_i, x(t_i); h).\end{aligned}$$

Zbog uniformne neprekidnosti funkcije

$$t \mapsto f(t, x(t)) - \Phi(t, x(t); 0),$$

ona je Riemann integrabilna, pa imamo

$$\lim_{h \rightarrow 0} \sum_{i=0}^{N-1} |\alpha_i| = \int_0^T |f(t, x(t)) - \Phi(t, x(t); 0)| dt.$$

S druge strane je

$$|\beta_i| \leq \beta(h) = \max_{\substack{|t-t'| \leq h' \\ 0 \leq h' \leq h}} |\Phi(t, x(t); 0) - \Phi(t', x(t'); h')|.$$

Po pretpostavci je funkcija Φ neprekidna na kompaktu $[0, T] \times [a, b] \times [0, h^*]$ (za svako $a < b$) i ako je $x(t)$ uniformno neprekidna na $[0, T]$, dobivamo

$$\lim_{h \rightarrow 0} \beta(h) = 0.$$

Time dobivamo

$$\sum_{i=0}^{N-1} h|\beta_i| \leq T\beta(h) \quad \Rightarrow \quad \lim_{h \rightarrow 0} \sum_{i=0}^{N-1} h|\beta_i| = 0.$$

Stoga je

$$\lim_{h \rightarrow 0} \sum_{i=0}^{N-1} |\varepsilon_i| = \int_0^T |f(t, x(t)) - \Phi(t, x(t); 0)| dt. \quad (1.18)$$

Ako je metoda konzistentna, onda na svakom rješenju $x(t)$ jednadžbe $\dot{x}(t) = f(t, x(t))$ vrijedi

$$\forall t \in [0, T], \quad f(t, x(t)) = \Phi(t, x(t); 0).$$

Koristeći teorem jedinstvenosti i egzistencije rješenja, zaključujemo da je $f(t, x) = \Phi(t, x; 0)$ za svako $t \in [0, T]$ i $x \in \mathbb{R}$. Obrat je evidentan iz formule (1.18). \square

Lema 1.5 Jedan dovoljan uvjet stabilnosti metode (1.17) je da postoji konstanta Λ takva da je

$$\forall t \in [0, T], \forall x, y \in \mathbb{R}, \forall h \in [0, h^*], \quad |\Phi(t, x; h) - \Phi(t, y; h)| \leq \Lambda|x - y|;$$

Tada je $M = e^{\Lambda T}$.

Dokaz. U oznakama iz definicije stabilnosti, koristeći lipšicovost, dobivamo

$$|x_{i+1} - y_{i+1}| \leq (1 + h\Lambda)|x_i - y_i| + |\varepsilon_i|.$$

Koristeći Lemu 1.3 dobivamo

$$\begin{aligned} |x_n - y_n| &\leq e^{nh\Lambda}|x_0 - y_0| + \sum_{i=0}^{n-1} e^{(n-i-1)h\Lambda}|\varepsilon_i| \\ &\leq e^{\Lambda T}(|x_0 - y_0| + \sum_{i=0}^{n-1} |\varepsilon_i|). \quad \square \end{aligned}$$

Iz dobivenoga možemo dobiti ocjenu greške posve analognu onoj kod Eulerove metode. Uvedemo li oznaku za maksimalnu lokalnu grešku diskretizacije

$$\varepsilon(h) = \max_{0 \leq i < N} |\varepsilon_i|,$$

onda dobivamo

$$\begin{aligned} |x_n - x(t_n)| &\leq e^{nh\Lambda}|x_0 - x(t_0)| + \sum_{i=0}^{n-1} e^{(n-i-1)h\Lambda}|\varepsilon_i| \\ &\leq e^{\Lambda T}|x_0 - x(t_0)| + \varepsilon(h) \sum_{i=0}^{n-1} e^{(n-i-1)h\Lambda} \\ &= e^{\Lambda T}|x_0 - x(t_0)| + \varepsilon(h) \frac{e^{nh\Lambda} - 1}{e^{h\Lambda} - 1} \\ &\leq e^{\Lambda T}|x_0 - x(t_0)| + \varepsilon(h) \frac{e^{nh\Lambda} - 1}{h\Lambda} \end{aligned}$$

gdje smo iskoristili nejednakost $e^x \geq 1 + x$, ($x \in \mathbb{R}$). Time smo dobili

$$\max_{0 \leq n \leq N} |x_n - x(t_n)| \leq e^{\Lambda T}|x_0 - x(0)| + \frac{\varepsilon(h)}{h\Lambda}(e^{\Lambda T} - 1).$$

Ova ocjena greška kaže da za sve jednokoračne metode vrijedi pravilo: *ako je LGD reda $n + 1$, onda je točnost metode reda n .*

1.5 Višekoračne metode

Želimo li točnije diskretizirati vremensku derivaciju u jednadžbi $\dot{x} = f(t, x)$ možemo se poslužiti centralnom diferencijom $\dot{x}(t_i) \approx (x_{i+1} - x_{i-1})/2h$. Time dolazimo do sljedeće metode:

$$x_{i+1} = x_{i-1} + 2hf(t_i, x_i).$$

Osnovna razlika prema prethodnim metodama je u tome što je ova metoda višekoračna: za računanje vrijednosti x_{i+1} potrebni su nam x_i i x_{i-1} .

Lokalna greška diskretizacije se lako dobiva na osnovu Taylorovog razvoja:

$$\begin{aligned} x(t_{i+1}) - [x_{i-1} + 2hf(t_i, x_i)] &= [x(t_i) + h\dot{x}(t_i) + \frac{h^2}{2}\ddot{x}(t_i) + \frac{h^3}{6}\ddot{x}(\tau_i)] \\ &- [x(t_i) - h\dot{x}(t_i) + \frac{h^2}{2}\ddot{x}(t_i) - \frac{h^3}{6}\ddot{x}(\tau_{i+1})] - 2hf(t_i, x_i) \\ &= 2h\dot{x}(t_i) + \frac{h^3}{6}[\ddot{x}(\tau_i) + \ddot{x}(\tau_{i+1})] \\ &= \frac{h^3}{3}\ddot{x}(\tau_{i+2}), \end{aligned}$$

gdje su τ_i, τ_{i+1} i τ_{i+2} neke točke iz intervala (t_{i-1}, t_{i+1}) . Vidimo dakle da je lokalna greška diskretizacije reda $O(h^3)$.

Sustavniji način izvođenja višekoračnih metoda oslanja se na integralnu formulaciju problema. Diferencijalna jednadžba se integrira na intervalu (t_i, t_{i+1}) :

$$x(t_{i+1}) = x(t_i) + \int_{t_i}^{t_{i+1}} \dot{x}(\tau) d\tau.$$

Provucimo Lagrangeov interpolacijski polinom kroz točke $(t_{i-1}, \dot{x}(t_{i-1}))$ i $(t_i, \dot{x}(t_i))$, te *ekstrapolirajmo* tu vrijednost na interval (t_i, t_{i+1}) . Dobivamo:

$$p_2(t) = \frac{t_i - t}{h}\dot{x}(t_{i-1}) + \frac{t - t_{i-1}}{h}\dot{x}(t_i)$$

i tu ćemo vrijednost koristiti kao aproksimaciju za $\dot{x}(t)$ na intervali (t_i, t_{i+1}) ; stoga je

$$\begin{aligned} x(t_{i+1}) &\approx x(t_i) + \int_{t_i}^{t_{i+1}} p_2(\tau) d\tau \\ &= x(t_i) + \frac{1}{h} \int_{t_i}^{t_{i+1}} (t_i - \tau) d\tau \dot{x}(t_{i-1}) + \frac{1}{h} \int_{t_i}^{t_{i+1}} (\tau - t_{i-1}) d\tau \dot{x}(t_i) \\ &= x(t_i) - \frac{1}{h} \int_0^h u du \dot{x}(t_{i-1}) + \frac{1}{h} \int_h^{2h} u du \dot{x}(t_i) \\ &= x(t_i) - \frac{h}{2}\dot{x}(t_{i-1}) + \frac{3h}{2}\dot{x}(t_i). \end{aligned}$$

Time dolazimo do Adams-Bashforthove metode

$$x_{i+1} = x_i + \frac{h}{2}[3f(t_i, x_i) - f(t_{i-1}, x_{i-1})]. \quad (1.19)$$

Prednost ovakve metode prema odgovarajućoj Runge-Kutta metodi je u tome što ona treba samo jedno računanje funkcije po koraku. Naime, jednu funkcijsku vrijednost možemo iskoristiti iz prethodnog koraka.

Prirodniji postupak bi bio interpolirati \dot{x} na intervalu (t_i, t_{i+1}) . Opisani postupak vodi na aproksimaciju

$$x(t_{i+1}) \approx x(t_i) + \frac{h}{2}[\dot{x}(t_{i+1}) + \dot{x}(t_i)].$$

što vodi na Adams-Moultonovu metodu

$$x_{i+1} = x_i + \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x_{i+1})]. \quad (1.20)$$

Zadatak. Pokažite da su lokalne greške diskretizacije za Adams-Bashfortovu i Adams-Moultonovu metodu dane formulama:

$$\begin{aligned} x_{i+1} &= x_i + \frac{h}{2}[3f(t_i, x_i) - f(t_{i-1}, x_{i-1})] + \frac{5}{12}\ddot{x}(\tau_i)h^3, \\ x_{i+1} &= x_i + \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x_{i+1})] + \frac{1}{12}\ddot{x}(\tau_i)h^3. \quad \square \end{aligned}$$

Adams-Moultonova metoda je implicitna. To znači da u svako koraku moramo rješavati nelinearnu jednadžbu kako bismo došli do aproksimacije na sljedećem vremenskom sloju. S tim u vezi javlja se pitanje da li jednadžba ima rješenje. Odgovor je pozitivan za dovoljno male vrijednosti koraka h . Da bismo to pokazali uočimo da u svakom koraku Adams-Moultonove metode rješavamo problem fiksne točke

$$\Phi(x) = x,$$

gdje je

$$\Phi(x) = x_i + \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x)].$$

Kako je

$$|\Phi(x) - \Phi(y)| = \frac{h}{2}|f(t_{i+1}, x) - f(t_{i+1}, y)| \leq \frac{h}{2}L|x - y|,$$

gdje je L konstanta lipšicovosti funkcije f , vidimo da je Φ kontrakcija za $hL/2 < 1$, i stoga problem ima jedinstveno rješenje.

U praksi se za rješavanje nelinearne jednadžbe u svakoj iteraciji najčešće koristi Newtonova metoda. Razlog je taj što za male h iz prethodne iteracije već imamo dobru početnu iteraciju za Newtonov algoritam. Najčešće su zatim dovoljne dvije do tri Newtonove iteracije.

Da bismo primijenili Newtonovu metodu napišimo Adams-Moultonovu metodu u obliku

$$G(x_{i+1}) = x_{i+1} - x_i - \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x_{i+1})] = 0.$$

$$G'(x) = 1 - \frac{h}{2} \frac{\partial f}{\partial x}(t_{i+1}, x).$$

Stoga Newtonove iteracije glase

$$x_{i+1}^{(n+1)} = x_{i+1}^{(n)} - \frac{G(x_{i+1}^{(n)})}{G'(x_{i+1}^{(n)})}, \quad n = 0, 1, 2, \dots$$

Za početnu iteraciju možemo uzeti $x_{i+1}^{(0)} = x_i$, ili

$$x_{i+1}^{(0)} = x_i + hf(t_i, x_i).$$

Iteracije se zaustavljaju kad je $|x_{i+1}^{(n+1)} - x_{i+1}^{(n)}| < \varepsilon$ i/ili $G(x_{i+1}^{(n+1)}) < \varepsilon$, gdje je ε zadana tolerancija. Zadnja iteracija se uzima za x_{i+1} .

Jedan način da se izbjegne rješavanje nelinearne jednadžbe je da se formira metoda tipa **prediktor-korektor**. To se čini tako da se spoje dvije metode istog reda točnosti: jedna implicitna i jedna eksplicitna. Eksplicitna metoda se koristi kao prediktor - ona računa međuvrijednost koja ulazi u implicitnu metodu, koja funkcionira kao korektor. Na primjer, Adams-Bashfortovu i Adams-Moultonovu metodu možemo povezati u jednu prediktor-korektor metodu na sljedeći način:

$$\begin{aligned} x_{i+1}^p &= x_i + \frac{h}{2}[3f(t_i, x_i) - f(t_{i-1}, x_{i-1})] \\ x_{i+1} &= x_i + \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x_{i+1}^p)] \end{aligned}$$

Koristeći u implicitnoj metodi vrijednost x_{i+1}^p , koju je izračunala prediktor metoda, izbjegavamo rješavanje nelinearne jednadžbe.

Zadatak. *Izračunajte lokalnu grešku diskretizacije Adamsove prediktor-korektor metode. Pokažite da je ona ostala reda $O(h^3)$. \square*

Napomena. *Modificirana Eulerova metoda može se interpretirati kao metoda tipa prediktor-korektor. Metoda ima oblik*

$$x_{i+1} = x_i + \frac{h}{2}[f(t_i, x_i) + f(t_i + h, x_i + hf(t_i, x_i))],$$

što se može zapisati u obliku

$$\begin{aligned} x_{i+1}^p &= x_i + hf(t_i, x_i) \\ x_{i+1} &= x_i + \frac{h}{2}[f(t_i, x_i) + f(t_{i+1}, x_{i+1}^p)] \end{aligned}$$

Prediktor je, dakle, Eulerova metoda, a korektor Adams-Moultonov metoda. \square

Adamsove metode (Adams-Bashfortova i Adams-Moultonova) mogu se poopćiti do metoda proizvoljno visokog reda. Potrebno je samo povećati broj točaka u kojima se interpolira derivacije \dot{x} .

Adams-Bashfortova metoda s n točaka dobiva se tako da se \dot{x} aproksimira Lagrangeovim interpolacijskim polinomom p_{n-1} , $(n-1)$ -og stupnja, kroz točke

$$(t_{i-n+1}, \dot{x}(t_{i-n+1})), \dots, (t_{i-1}, \dot{x}(t_{i-1})), (t_i, \dot{x}(t_i)).$$

Zatim se koristi aproksimacija

$$x(t_{i+1}) \approx x(t_i) + \int_{t_i}^{t_{i+1}} p_{n-1}(\tau) d\tau,$$

i $\dot{x}(t_j)$ se zamijeni s $f(t_j, x_j)$. Evidentno, takvim postupkom dobivamo eksplicitnu metodu.

Adams-Moultonova metoda s n točaka dobiva se tako da se \dot{x} aproksimira Lagrangeovim interpolacijskim polinomom $(n-1)$ -og stupnja p_{n-1} , kroz točke

$$(t_{i-n+2}, \dot{x}(t_{i-n+2})), \dots, (t_i, \dot{x}(t_i)), (t_{i+1}, \dot{x}(t_{i+1})).$$

i zatim se koristi aproksimacija

$$x(t_{i+1}) \approx x(t_i) + \int_{t_i}^{t_{i+1}} p_{n-1}(\tau) d\tau.$$

Time se dobiva implicitna metoda.

Zadatak. *Izvedite sljedeće dvije Adamsove metode s 4 točke:*

$$x_{i+1} = x_i + \frac{h}{24}[55f(t_i, x_i) - 59f(t_{i-1}, x_{i-1}) + 37f(t_{i-2}, x_{i-2}) - 9f(t_{i-3}, x_{i-3})],$$

$$x_{i+1} = x_i + \frac{h}{24}[9f(t_{i+1}, x_{i+1}) + 19f(t_i, x_i) - 5f(t_{i-1}, x_{i-1}) + f(t_{i-2}, x_{i-2})].$$

Pokažite da metode respektivno imaju greške diskretizacije

$$\frac{251}{720}x^{(5)}(\tau_i)h^5, \quad -\frac{19}{720}x^{(5)}(\tau_i)h^5. \quad \square$$

Pomoću Adamsovih metoda 4. reda možemo konstruirati prediktor-korektor metodu 4. reda. Ona pred klasičnom Runge-Kutta metodom ima tu prednost da treba samo dva računanja funkcije po iteraciji:

$$x_{i+1}^p = x_i + \frac{h}{24}[55f(t_i, x_i) - 59f(t_{i-1}, x_{i-1}) + 37f(t_{i-2}, x_{i-2}) - 9f(t_{i-3}, x_{i-3})],$$

$$x_{i+1} = x_i + \frac{h}{24}[9f(t_{i+1}, x_{i+1}^p) + 19f(t_i, x_i) - 5f(t_{i-1}, x_{i-1}) + f(t_{i-2}, x_{i-2})].$$

Zadatak. *Usporedite točnost klasične Runge-Kutta metode i Adamsove prediktor-korektor metode 4. reda na test-primjeru. Pokažite da je Runge-Kutta metoda preciznija.* \square

1.6 Metode varijabilnog koraka

Metode varijabilnog koraka dinamički mijenjaju korak h kako bi držale grešku metode ispod zadane tolerancije ε . Metode tog tipa moraju na neki način procijeniti grešku metode.

Kao osnovu metode varijabilnog koraka koristit ćemo jednu Runge-Kutta metodu reda n . Greška takve metode je oblika Ch^n , gdje je C neka konstanta koja ovisi o egzaktnom rješenju, pa nam je stoga nepoznata.

Postavimo se u situaciju kada imamo aproksimaciju x_t točnog rješenja $x(t)$ u trenutku t . Ta je vrijednost izračunata s vremenskim korakom kojeg ćemo označiti s h_0 . Koristeći odabranu Runge-Kutta metodu računamo dvije aproksimacije $x_{t+h_0}^{(1)}$ i $x_{t+h_0}^{(2)}$ točne vrijednosti $x(t+h_0)$: $x_{t+h_0}^{(1)}$ računamo s korakom h_0 , a $x_{t+h_0}^{(2)}$ s korakom $h_0/2$. Drugim riječima, prvu aproksimaciju računamo s jednim korakom Runge-Kutta metode, a drugu s dva. Nakon toga računamo veličinu

$$E = |x_{t+h_0}^{(1)} - x_{t+h_0}^{(2)}|,$$

koja nam služi za procjenu greške. Na osnovu naše pretpostavke da je greška Runge-Kutta metode oblika Ch^n , gdje je C približno konstanta, možemo procijeniti C :

$$\begin{aligned} E &= |x_{t+h_0}^{(1)} - x_{t+h_0}^{(2)}| \\ &= |(x_{t+h_0}^{(1)} - x(t+h_0)) - (x_{t+h_0}^{(2)} - x(t+h_0))| \\ &\approx |Ch_0^n - C\left(\frac{h_0}{2}\right)^n| = (1 - 2^{-n})Ch_0^n. \end{aligned}$$

Imamo prema tome

$$C \approx \frac{E}{(1 - 2^{-n})h_0^n}.$$

Mi bismo htjeli da je $Ch_0^n < \varepsilon$. Ukoliko je taj uvjet ispunjen smatramo da je korak h_0 dovoljno mali i uzimamo točniju vrijednost $x_{t+h_0}^{(2)}$ kao aproksimaciju u trenutku $t+h_0$. Ukoliko je $Ch_0^n \geq \varepsilon$, onda trebamo smanjiti korak h_0 kako bismo postigli traženu točnost. Novi korak h_ε mora biti takav da je

$$\varepsilon \approx Ch_\varepsilon^n.$$

Uočimo da mi ne želimo prevelik, ali niti premalen korak. Stoga novi korak računamo iz

$$\varepsilon \approx \frac{E}{(1 - 2^{-n})h_0^n} h_\varepsilon^n,$$

što daje

$$h_\varepsilon = \left[\frac{(1 - 2^{-n})\varepsilon}{E} \right]^{1/n} h_0.$$

Naš postupak će biti sljedeći: U svakom slučaju računamo optimalni korak h_ε .

- Ako je $E/(1 - 2^{-n}) < \varepsilon$ prihvaćamo aproksimaciju $x_{t+h_0}^{(2)}$ i za korak h_0 na sljedećem vremenskom sloju koristimo h_ε .
- Ako je $E/(1 - 2^{-n}) \geq \varepsilon$ ponavljamo račun s novim korakom $h_0 = h_\varepsilon$.

Uočimo da se uzimanjem h_ε za novi korak u slučaju kad je aproksimacija zadovoljavajuća osiguravamo od premalenog koraka. U tom slučaju je $h_\varepsilon > h_0$.

Prelazimo na konstrukciju algoritma. Uočimo da metoda smanjuje korak ako je to potrebno pa stoga treba staviti neku minimalnu vrijednost koraka h_{\min} ispod koje ne želimo smanjivati korak. Pretjerano smanjivanje koraka može biti znak približavanja kraju domene egzistencije rješenja.

Algoritam 1 daje rješenje kroz polja t i x koja sadrže niz vremenskih trenutaka (t) i niz aproksimacija rješenja u tim trenucima (x). Uzet ćemo jednostavan pristup u kojem se ta polja dimenzioniraju izvan rutine. Rutina će se zaustaviti ako su polja suvuše mala.

U Algoritmu 1 koristimo rutinu $RK(n, t_0, x_0, t, x)$ koja računa Runge-Kutta aproksimaciju x rješenja $x(t)$, uz početni uvjet $x(t_0) = x_0$, u n koraka. Za n ćemo uzimati 1 ili 2. Rutine $RK()$ poziva funkciju $f(t, x)$ koja mora biti kodirana kao potprogram.

Zadatak. Programirati Algoritam 1 i testirati ga na donjim primjerima. Zadažite si $T > t_0$ te $\varepsilon > 0$ i provjerite da li algoritam postiže zadanu točnost. Eksperimentalno odredite korak h potreban da metoda s fiksnim korakom postigne istu točnost u krajnjem trenutku. Odredite broj poziva funkcije f u oba slučaja.

Test primjeri. 1. Aproksimirati $x(1.25)$ za

$$\dot{x} = 5(t - 1)x, \quad x(0) = 5.$$

Rješenje:

$$x(t) = 5e^{(5/2)t^2 - 5t}.$$

2. Aproksimirati $x(1.5)$ za

$$\dot{x} = 1 + x^2, \quad x(0) = 0.$$

Rješenje:

$$x(t) = \tan x.$$

3. Aproksimirati $x(30)$ za

$$\dot{x} = \cos \frac{\pi t}{12} - x, \quad x(0) = 50.$$

Rješenje:

$$x(t) = \frac{\cos(\pi t/12) + (\pi/12) \sin(\pi t/12)}{1 + (\pi/12)^2} + \left[50 - \frac{1}{1 + (\pi/12)^2}\right] e^{-t}.$$

4. Aproximirati $x_1(3)$, $x_2(3)$ i $x_3(3)$ za

$$\begin{aligned}\dot{x}_1 &= -2x_1 - x_2 + e^{-3t}, & x_1(0) &= 1 \\ \dot{x}_2 &= 2x_1 - x_2 + x_3, & x_2(0) &= 0 \\ \dot{x}_3 &= 2x_2 - 2x_3 - 2e^{-3t}, & x_3(0) &= 0.\end{aligned}$$

Rješenje:

$$\begin{aligned}x_1(t) &= -2e^{-t} + (4 + 2t)e^{-2t} - e^{-3t} \\ x_2(t) &= 2e^{-t} - 2e^{-2t} \\ x_3(t) &= 4e^{-t} - (6 + 4t)e^{-2t} + 2e^{-3t}.\end{aligned}$$

5. Aproximirati $x_1(20)$ i $x_2(20)$ za

$$\begin{aligned}\ddot{x}_1 &= -2x_1 - \frac{1}{2}x_2, & x_1(0) &= 0, & \dot{x}_1(0) &= 0 \\ \ddot{x}_2 &= 2x_1 - 2x_2 + 10 \cos(2t), & x_2(0) &= 0, & \dot{x}_2(0) &= 0\end{aligned}$$

Rješenje:

$$\begin{aligned}x_1(t) &= \frac{5}{3} \cos(2t) + \frac{5}{6} \cos t - \frac{5}{2} \cos(\sqrt{3}t) \\ x_2(t) &= -\frac{20}{3} \cos(2t) + \frac{5}{3} \cos t + 5 \cos(\sqrt{3}t).\end{aligned}$$

□

Efikasnost metode varijabilnog koraka može se povećati ako se umjesto jedne Runge-Kutta metode upotrijebe dvije takve metode različitog reda. Uzmimo, na primjer, da imamo jednu RK metodu reda n i jednu reda $n + 1$. Polazimo od vrijednosti x_t koja predstavlja aproksimaciju za $x(t)$ i od koraka h_0 . Računamo aproksimaciju rješenja u trenutku $t + h_0$ pomoću obje metode: prva daje vrijednost $x_{t+h_0}^1$, s greškom proporcionalnom s h^n , a druga vrijednost $x_{t+h_0}^2$, s greškom proporcionalnom s h^{n+1} . Za dovoljno malo h_0 imamo približno

$$\begin{aligned}E &= |x_{t+h_0}^1 - x_{t+h_0}^2| = |(x_{t+h_0}^1 - x(t+h_0)) - (x_{t+h_0}^2 - x(t+h_0))| \\ &\approx |C_1 h_0^n - C_2 h_0^{n+1}| \approx C_1 h_0^n.\end{aligned}$$

Time smo dobili procjenu za konstantu C_1 :

$$C_1 \approx \frac{E}{h_0^n}.$$

Naša je želja raditi s optimalnim korakom h_ε , za zadanu toleranciju ε . Takav korak treba zadovoljavati

$$C_1 h_\varepsilon^n \approx \varepsilon,$$

Algoritam 1 kRK metoda varijabilnog koraka

Ulaz: t_0, x_0 , početni podaci $x(t_0) = x_0$;
 Krajnje vrijeme T ;
 Tolerancija ε ;
 Minimalni korak $h_{\min} > 0$;
 N dimenzija polja t i x ;
 Polja t i x ;
 $t_0=t_0, t_1=T, x_0=x_0, k = 0, x(0)=x_0, t(0)=t_0, h_0=t_1-t_0, \text{flag}=\text{FALSE}$
repeat
 repeat
 RK(1, t_0,x_0,t_1,x_1), RK(2, t_0,x_0,t_1,x_2)
 $E = |x_1 - x_2|$
 if ($E < \varepsilon$) **then**
 $t_0=t_1, x_0=x_2, \text{flag}=\text{TRUE}$
 if $k \geq N - 1$ **then**
 Izlaz: x, t . Dimenzija polja je suviše mala.
 end if
 $k = k + 1$
 $t(k) = t_0, x(k) = x_0$
 end if
 {Računanje novog koraka}
 if $E > 0$ **then**
 $h_0 = h_0 \sqrt[n]{(1 - 2^{-n})\varepsilon/E}$
 if ($\text{flag} = \text{FALSE} \ \& \ h_0 < h_{\min}$) **then**
 Izlaz: x, t . Korak je suviše mali.
 end if
 $t_1 = t_0 + h_0$
 if $t_1 > T$ **then**
 $h_0 = T - t_0, t_1 = T$
 end if
 else
 $h_0 = T - t_0, t_1 = T$
 end if
 until ($\text{flag} = \text{TRUE}$)
until ($t_0 \geq T$)
 Izlaz: Polja t i x sadrže aproksimativno rješenje.

pa stoga za optimalni korak dobivamo procjenu

$$h_\varepsilon \approx \sqrt[n]{\frac{\varepsilon}{E}} h_0.$$

Mi ćemo uzimati

$$h_\varepsilon = \sigma \sqrt[n]{\frac{\varepsilon}{E}} h_0, \quad 0 < \sigma \leq 1,$$

gdje je $\sigma \approx 0.9$.

Metoda kontrole koraka je ista kao i ranije. Ukoliko je $E < \varepsilon$, prihvaćamo aproksimaciju izračunatu s korakom h_0 ; pri tome uzimamo točniju aproksimaciju $x_{t+h_0}^2$. Aproksimaciju na sljedećem vremenskom sloju računamo s korakom $h_0 = h_\varepsilon$. Ako je $E \geq \varepsilon$ onda odbacujemo izračunatu aproksimaciju i ponavljamo postupak s $h_0 = h_\varepsilon$. Uočimo da je u tom slučaju h_ε manji od prethodnog h_0 .

Pogledajmo sada kako ovakav postupak omogućava optimizaciju broja računskih operacija po koraku metode. Pri tome ćemo kao značajnu operaciju uzimati samo poziv funkcije f . Uzmimo jednu metodu drugog reda i jednu trećeg reda. Metoda 2. reda:

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + \alpha_2^{(1)} h, x_i + h\beta_{2,1}^{(1)} m_1) \\ x_{i+1} &= x_i + h(w_1^{(1)} m_1 + w_2^{(1)} m_2). \end{aligned}$$

Metoda 3. reda:

$$\begin{aligned} m_1 &= f(t_i, x_i) \\ m_2 &= f(t_i + \alpha_2^{(2)} h, x_i + h\beta_{2,1}^{(2)} m_1) \\ m_3 &= f(t_i + \alpha_3^{(2)} h, x_i + h\beta_{3,1}^{(2)} m_1 + h\beta_{3,2}^{(2)} m_2) \\ x_{i+1} &= x_i + h(w_1^{(2)} m_1 + w_2^{(2)} m_2 + w_3^{(2)} m_3). \end{aligned}$$

Ovdje smo te metode zapisali u općenitoj formi. Da bismo imali korektan red metode koeficijenti moraju zadovoljavati:

$$w_1^{(1)} + w_2^{(1)} = 1, \quad w_2^{(1)} \alpha_2^{(1)} = \frac{1}{2}, \quad w_2^{(1)} \beta_{2,1}^{(1)} = \frac{1}{2}.$$

M. JURAK 31. svibnja 2006.

za metodu drugog reda i

$$\begin{aligned}
 w_1^{(2)} + w_2^{(2)} + w_3^{(2)} &= 1 \\
 w_2^{(2)}\alpha_2^{(2)} + w_3^{(2)}\alpha_3^{(2)} &= \frac{1}{2} \\
 w_2^{(2)}(\alpha_2^{(2)})^2 + w_3^{(2)}(\alpha_3^{(2)})^2 &= \frac{1}{3} \\
 w_2^{(2)}\beta_{2,1}^{(2)} + w_3^{(2)}(\beta_{3,1}^{(2)} + \beta_{3,2}^{(2)}) &= \frac{1}{2} \\
 w_2^{(2)}\alpha_2^{(2)}\beta_{2,1}^{(2)} + w_3^{(2)}\alpha_3^{(2)}(\beta_{3,1}^{(2)} + \beta_{3,2}^{(2)}) &= \frac{1}{3} \\
 w_2^{(2)}(\beta_{2,1}^{(2)})^2 + w_3^{(2)}(\beta_{3,1}^{(2)} + \beta_{3,2}^{(2)})^2 &= \frac{1}{3} \\
 w_3^{(2)}\alpha_2^{(2)}\beta_{3,2}^{(2)} = \frac{1}{6}, \quad w_3^{(2)}\beta_{2,1}^{(2)}\beta_{3,2}^{(2)} &= \frac{1}{6}.
 \end{aligned}$$

za metodu trećeg reda.

Usporedimo li metodu drugog i trećeg reda vidimo da se vrijednost m_1 treba računati najviše jednom. Najviše što možemo postići je da se i m_2 računa samo jednom. To nas vodi do dodatnog uvjeta

$$\alpha_2^{(1)} = \alpha_2^{(2)}, \quad \beta_{2,1}^{(1)} = \beta_{2,1}^{(2)}. \quad (1.21)$$

Zadatak. *Krenite od modificirane Eulerove metode*

$$\begin{aligned}
 m_1 &= f(t_i, x_i) \\
 m_2 &= f(t_i + h, x_i + hm_1) \\
 x_{i+1} &= x_i + \frac{h}{2}(m_1 + m_2).
 \end{aligned}$$

($\alpha_2^{(1)} = \beta_{2,1}^{(1)} = 1$ i $w_1^{(1)} = w_2^{(1)} = 1/2$.) Uvjeti (1.21) koje treba zadovoljiti su sada

$$\alpha_2^{(2)} = 1, \quad \beta_{2,1}^{(2)} = 1.$$

Pokažite da metoda

$$\begin{aligned}
 m_1 &= f(t_i, x_i) \\
 m_2 &= f(t_i + h, x_i + hm_1) \\
 m_3 &= f(t_i + h/2, x_i + (h/4)m_1 + (h/4)m_2) \\
 x_{i+1} &= x_i + \frac{h}{6}(m_1 + m_2 + 4m_3).
 \end{aligned}$$

zadovoljava traženi uvjet kao i sve uvjete kompatibilnosti koeficijenata za metodu trećeg reda. \square

Ocjenu greške za ovaj par metoda možemo računati na sljedeći način ($t = t_i$ pa je $x_t = x_i$):

$$\begin{aligned} E &= |x_{t+h}^1 - x_{t+h}^2| \\ &= |(x_t + \frac{h}{2}(m_1 + m_2)) - (x_t + \frac{h}{6}(m_1 + m_2 + 4m_3))| \\ &= \frac{h}{3}|2m_3 - m_1 - m_2|. \end{aligned}$$

Vidimo da $x_{t+h_0}^1$ uopće ne trebamo računati. Time smo došli do metode koja se naziva Runge-Kutta-Felbergova (2)3 metoda. Ona može biti zapisana na sljedeći način:

$$\begin{aligned} m_1 &= f(t, x_t) \\ m_2 &= f(t + h, x_t + hm_1) \\ m_3 &= f(t + h/2, x_t + (h/4)m_1 + (h/4)m_2) \\ E &= \frac{h}{3}|2m_3 - m_1 - m_2| \\ x_{t+h} &= x_t + \frac{h}{6}(m_1 + m_2 + 4m_3). \end{aligned}$$

Iz izvoda Runge-Kutta-Felbergova (2)3 metode je jasno da se upotrebom metoda višeg reda mogu postići veće uštede. Ovdje, bez izvoda, dajemo primjer Runge-Kutta-Felbergove (4)5 metode.

$$\begin{aligned} m_1 &= f(t, x_t) \\ m_2 &= f(t + \frac{2}{9}h, x_t + \frac{2}{9}hm_1) \\ m_3 &= f(t + \frac{1}{3}h, x_t + \frac{1}{12}hm_1 + \frac{1}{4}hm_2) \\ m_4 &= f(t + \frac{3}{4}h, x_t + \frac{69}{128}hm_1 - \frac{243}{128}hm_2 + \frac{135}{64}hm_3) \\ m_5 &= f(t + h, x_t - \frac{17}{12}hm_1 + \frac{27}{4}hm_2 - \frac{27}{5}hm_3 + \frac{16}{15}hm_4) \\ m_6 &= f(t + \frac{5}{6}h, x_t + \frac{65}{432}hm_1 - \frac{5}{16}hm_2 + \frac{13}{16}hm_3 + \frac{4}{27}hm_4 + \frac{5}{144}hm_5) \\ E &= h| -\frac{1}{150}m_1 + \frac{3}{100}m_3 - \frac{16}{75}m_4 - \frac{1}{20}m_5 + \frac{6}{25}m_6| \\ x_{t+h} &= x_t + h(\frac{47}{450}m_1 + \frac{12}{25}m_3 + \frac{32}{225}m_4 + \frac{1}{30}m_5 + \frac{6}{25}m_6). \end{aligned}$$

Implementacija Runge-Kutta-Felbergove (4)5 metode dana je u Algoritmu 2. **Zadatak.** Testirajte Algoritam 2 i usporedite ga s Algoritmom 1. \square

Algoritam 2 kRKF 4(5) metoda

Ulaz: t_0, x_0 , početni podaci $x(t_0) = x_0$;
 Krajnje vrijeme T ;
 Tolerancija ε ;
 Minimalni korak $h_{\min} > 0$;
 N dimenzija polja t i x ;
 Polja t i x ;
 $t_0=t_0, t_1=T, x_0=x_0, k = 0, x(0)=x_0, t(0)=t_0, h_0=t_1-t_0, \text{flag}=\text{FALSE}$
repeat
 repeat
 RK(1, t_0,x_0,t_1,x_1), RK(2, t_0,x_0,t_1,x_2)
 $E = |x_1 - x_2|$
 if ($E < \varepsilon$) **then**
 $t_0=t_1, x_0=x_2, \text{flag}=\text{TRUE}$
 if $k \geq N - 1$ **then**
 Izlaz: x, t . Dimenzija polja je suviše mala.
 end if
 $k = k + 1$
 $t(k) = t_0, x(k) = x_0$
 end if
 {Računanje novog koraka}
 if $E > 0$ **then**
 $h_0 = h_0 \sqrt[n]{(1 - 2^{-n})\varepsilon/E}$
 if ($\text{flag} = \text{FALSE} \ \& \ h_0 < h_{\min}$) **then**
 Izlaz: x, t . Korak je suviše mali.
 end if
 $t_1 = t_0 + h_0$
 if $t_1 > T$ **then**
 $h_0 = T - t_0, t_1 = T$
 end if
 else
 $h_0 = T - t_0, t_1 = T$
 end if
 until ($\text{flag} = \text{TRUE}$)
until ($t_0 \geq T$)
 Izlaz: Polja t i x sadrže aproksimativno rješenje.

1.7 O stabilnosti. Kruti sustavi

Matematička zadaća je dobro postavljena (ili korektna) ukoliko ima jedno i samo jedno rješenje, koje neprekidno ovisi o ulaznim podacima. Na primjer, u uvjetima Cauchy-Lipschitzovog teorema, Cauchyjeva zadaća za običnu diferencijalnu jednadžbu je dobro postavljena.

Na diskretu zadaću primijenjuje se ista definicija korektnosti. Zadaća je korektna ako ima jedinstveno rješenje koje neprekidno ovisi o zadanim podacima. To znači da male promjene ulaznih podataka proizvode male promjene rješenja.

U diskretnim zadaćama *malost* promjene rješenja se uvijek mjeri u odnosu na diskretizacijski parametar. Stoga korektna kontinuirana zadaća može postati nekorektna u fp-sustavu određene preiznosti. Uzmimo, na primjer, Cauchyjev problem

$$\begin{aligned}\dot{x}(t) &= 3x(t) - 3t, & t \in [0, 5] \\ x(0) &= 1/3.\end{aligned}$$

Točno rješenje je $\dot{x}(t) = 1/3 + t$. Budući da broj $1/3$ nije moguće reprezentirati u računalu, stvarni početni uvjet u diskretnom problemu je $x(0) = 1/3 + \varepsilon$, gdje je ε greška reda veličine strojnog epsilon. Pripadno egzaktno rješenje je $x_1(t) = 1/3 + \varepsilon e^{3t}$, a razlika između dva rješenja u trenutku $t = 5$ je $\varepsilon e^{15} \approx 3\varepsilon 10^6$. Primijenimo li, dakle, bilo koju numeričku metodu u sustavu fp-brojeva čija je preciznost 10^{-6} , nećemo moći postići zadovoljavajuću aproksimaciju u $t = 5$; problem je numerički nekorektan. S druge strane, u fp-sustavu s preciznošću 10^{-16} dobivamo numerički korektan problem.²

Pokazali smo da su jednokoračne metode, primijenjene na Cauchyjevu zadaću, stabilne. Konstanta koja se pojavljuje u ocjeni stabilnosti ovisi, između ostalog, o duljini vremenskog intervala na kojem se jednadžba integrira i o diskretizacijskom parametru h . Ukoliko je ta konstanta vrlo velika kažemo da je numerička zadaća *slabo uvjetovana*. U krajnjim slučajevima zadaća može postati numerički nekorektna. Pogledajmo, na primjer, Cauchyjevu zadaću

$$\dot{x}(t) = -2tx(t), \quad t \in [0, 15] \tag{1.22}$$

$$x(0) = 1. \tag{1.23}$$

Egzaktno rješenje je $x(t) = e^{-t^2}$, koje vrlo brzo konvergira u nulu kada $t \rightarrow \infty$. Primijenimo li Eulerovu metodu s korakom 0.2 na ovaj problem dobit ćemo zadovoljavajuće rješenje sve do trenutka $t \approx 10$, kada se javljaju snažne oscilacije. Eulerova metoda je na ovom primjeru nestabilna.

Nestabilnost se pojavila za velika vremena t pa možemo pretpostaviti da je konstanta u ocjeni stabilnosti, koja ovisi o krajnjem vremenu T , postala suviše

²Ovo je ekstreman primjer u kojem se umjesto diskretizacijskog parametra pojavljuje preciznost fp-sustava kao faktor koji određuje je li numerička zadaća korektna ili nije.

velika. Da bismo bolje razumjeli pojavu nestabilnosti, pogledajmo jednostavniju zadaću:

$$\begin{aligned}\dot{x}(t) &= \lambda x(t), \\ x(0) &= x_0.\end{aligned}$$

Primjenom Eulerove metode dobivamo

$$\begin{aligned}x_1 &= x_0 + h\lambda x_0 = (1 + \lambda h)x_0 \\ x_2 &= (1 + \lambda h)x_1 = (1 + \lambda h)^2 x_0 \\ \dots &\dots\end{aligned}$$

što daje $x_n = (1 + \lambda h)^n x_0$. Za $\lambda < 0$ rješenje $x(t) = x_0 e^{\lambda t}$ teži u nulu kada $t \rightarrow \infty$. Eulerova metoda će imati isto asimptotsko ponašanje ako je

$$|1 + \lambda h| < 1. \quad (1.24)$$

Taj se uvjet svodi na $-2 < \lambda h < 0$, odnosno dobivamo ograničenje na korak h :

$$h < \frac{2}{|\lambda|}.$$

Na osnovu ovog primjera možemo bolje razumijeti nestabilnost Eulerove metode u prethodnom primjeru. U njemu λ nije konstantan već imamo $\lambda = -2t$. Primijenim li formalno uvjet stabilnosti (1.24) na problem (1.22), (1.23), dobivamo

$$-2 < -2ht < 0.$$

Uz $h = 0.2$ uvjet će biti narušen za $t > 5$. Možemo zaključiti da ova analiza daje kvalitativno dobro objašnjenje pojave nestabilnosti, pa ju stoga generaliziramo na diferencijalnu jednadžbu oblika

$$\dot{x}(t) = f(t, x(t)).$$

Pri tome moramo pretpostaviti da diferencijalna jednadžba ima jedno stacionarno rješenje i da druga rješenja, barem za bliske početne uvjete, konvergiraju k stacionarnom, kada $t \rightarrow \infty$. Kriterij egzistencije stacionarnog rješenja je vrlo jednostavan: $x(t) = \bar{x}$ je stacionarno rješenje ako i samo ako je $f(t, \bar{x}) = 0$, za sva vremena t . U ovoj diskusiji mi bez smanjenja općenitosti možemo uzeti da je $\bar{x} = 0$, tj. $f(t, 0) = 0$.

Jednadžbu ćemo najprije *linearizirati* tako da desnu stranu razvijemo u Taylorov red oko $x = \bar{x} = 0$,

$$f(t, x) = f(t, 0) + \frac{\partial f}{\partial x}(t, 0)x + O(x^2) = \frac{\partial f}{\partial x}(t, 0)x + O(x^2).$$

Sada u slučaju da je $\frac{\partial f}{\partial x}(t, 0) < 0$ zaključujemo da će Eulerova metoda biti stabilna ako je

$$-2 < h \frac{\partial f}{\partial x}(t, 0) < 0.$$

Ovi primjeri nas navode na sljedeću definiciju.

Definicija 1.4 Interval (a, b) nazivamo interval apsolutne stabilnosti numeričke metode ako za $h\lambda \in (a, b)$ metoda daje aproksimaciju (x_n) problema

$$\begin{aligned} \dot{x}(t) &= \lambda x(t), \\ x(0) &= x_0. \end{aligned}$$

sa svojstvom

$$\lim_{n \rightarrow \infty} x_n = 0.$$

Uočimo da mora vrijediti $\lambda < 0$, tako da je $a, b < 0$.

Interval apsolutne stabilnosti za Eulerovu metodu je $(-2, 0)$.

Zadatak. *Odredite interval apsolutne stabilnosti za modificiranu Eulerovu metodu.*

$$x_{i+1} = x_i + \frac{h}{2} [f(t_i, x_i) + f(t_i + h, x_i + hf(t_i, x_i))].$$

Rješenje. Za $f(t, x) = \lambda x$ dobivamo

$$\begin{aligned} x_{i+1} &= x_i + \frac{h}{2} [\lambda x_i + \lambda(x_i + h\lambda x_i)] \\ &= (1 + h\lambda + \frac{h^2 \lambda^2}{2}) x_i. \end{aligned}$$

Oдавде slijedi uvjet

$$|1 + h\lambda + \frac{h^2 \lambda^2}{2}| < 1,$$

što nakon jednostavnog računa daje interval apsolutne stabilnosti $(-2, 0)$. \square

Kod Runge-Kutta metoda općenito se dobiva jednačba oblika

$$x_n = P(h\lambda)^n x_0,$$

gdje je P neki polinom. Uvjet apsolutne stabilnosti je tada $|P(h\lambda)| < 1$.

Zadatak. *Pokažite da je za klasičnu Runge-Kutta metodu*

$$P(h\lambda) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2 + \frac{1}{6}(h\lambda)^3 + \frac{1}{24}(h\lambda)^4,$$

što daje interval apsolutne stabilnosti $(-2.78, 0)$, približno.

Izračunajmo interval apsolutne stabilnosti implicitne Eulerove metode. Dobivamo rekuriju

$$x_{i+1} = x_i + h\lambda x_{i+1},$$

što daje

$$x_i = \frac{1}{(1 - h\lambda)^i} x_0.$$

Vidimo da je metoda apsolutno stabilna za sve $h\lambda < 0$. Drugim riječima, interval apsolutne stabilnosti je $(-\infty, 0)$. To je osnovna prednost implicitnih metoda: One ne postavljaju dodatna ograničenja na korak metode.

Kod višekoračnih metoda uvjet apsolutne stabilnosti je složenije izračunati. Uzmimo kao primjer metodu

$$x_{i+1} = x_{i-1} + 2hf(t_i, x_i).$$

Primijenjena na linearizirani problem, ona daje

$$x_{i+1} = x_{i-1} + 2h\lambda x_i.$$

Ovdje se radi o diferencijskoj jednadžbi drugog reda. Njena rješenja možemo tražiti u obliku $x_i = r^i$. Uvrštavanjem u jednadžbu i skraćivanjem dobivamo kvadratnu jednadžbu za r :

$$r^2 = 1 + 2h\lambda r,$$

koja ima općenito dva rješenja r_1 i r_2 . Opće rješenje diferencijske jednadžbe će biti oblika $x_i = C_1 r_1^i + C_2 r_2^i$, pa stoga za stabilnost moramo zahtijevati

$$|r_1| < 1, \quad |r_2| < 1.$$

Jednostavno je izračunati

$$r_1 = h\lambda + \sqrt{h^2\lambda^2 + 1}, \quad r_2 = h\lambda - \sqrt{h^2\lambda^2 + 1},$$

i vidjeti da za $h\lambda < 0$ ne možemo zadovoljiti $|r_2| < 1$. Interval apsolutne stabilnosti je stoga prazan.

Zadatak. Pokažite da je Adams-Moultonova metoda apsolutno stabilna.

Metode tipa prediktor-korektor općenito ne zadržavaju svojstvo apsolutne stabilnosti svog implicitnog dijela.

Stabilnost metoda za sustave običnih diferencijalnih jednadžbi definira se na sličan način. Neka je zadana konstantna matrica \mathbf{A} reda N i Cauchyjev problem

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} \\ \mathbf{x}(0) &= \mathbf{x}_0. \end{aligned}$$

Promatrat ćemo jednostavnu situaciju u kojoj matrica \mathbf{A} ima N jednostrukih svojstvenih vrijednosti λ_i i bazu svojstvenih vektora \mathbf{v}^i :

$$\begin{aligned} \sigma(\mathbf{A}) &= \{\lambda_1, \lambda_2, \dots, \lambda_N\}, \\ \mathbf{A}\mathbf{v}^i &= \lambda_i \mathbf{v}^i, \quad i = 1, 2, \dots, N. \end{aligned}$$

Opće rješenje diferencijalne jednadžbe ima oblik

$$\mathbf{x}(t) = C_1 \mathbf{v}^1 e^{\lambda_1 t} + \dots + C_N \mathbf{v}^N e^{\lambda_N t}.$$

Promatramo one sustave čija se rješenja stabiliziraju u nulu (koja je stacionarno rješenje), tj. uzimamo matricu \mathbf{A} čije sve svojstvene vrijednosti imaju negativan realan dio.

Primijenom Eulerove metode dobivamo:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + h\mathbf{A}\mathbf{x}_i, \quad i = 0, 1, \dots \quad (1.25)$$

Rješenje diferencijalne jednadžbe tražimo u obliku $\mathbf{x}_i = r^i \mathbf{v}^p$ ($p = 1, \dots, N$). Uvrštavanjem u (1.25) dobivamo $r = 1 + h\lambda_p$, pa stoga opće rješenje ima oblik

$$\mathbf{x}_i = C_1 \mathbf{v}^1 (1 + h\lambda_1)^i + \dots + C_N \mathbf{v}^N (1 + h\lambda_N)^i$$

Uvjet apsolutne stabilnosti je tada

$$|1 + h\lambda_p| < 1, \quad p = 1, 2, \dots, N.$$

Odavde vidimo da je uvjet stabilnosti zadovoljen ako se korak h nalazi u presjeku N diskova. Taj je uvjet stoga teže zadovoljiti u slučaju sustava, nego u skalarnom slučaju.

Primjer.

$$\begin{aligned} \ddot{x} + (b+1)\dot{x} + bx &= 0 \\ x(0) &= 1, \quad \dot{x}(0) = 0, \end{aligned}$$

Zapisano kao sustav

$$\begin{aligned} \dot{x}_1 &= x_2, \quad x_1(0) = 1 \\ \dot{x}_2 &= -bx_1 - (b+1)x_2, \quad x_2(0) = 0. \end{aligned}$$

Točno rješenje:

$$\begin{aligned} x_1(t) &= \frac{b}{b-1}e^{-t} - \frac{1}{b-1}e^{-bt} \\ x_2(t) &= -\frac{b}{b-1}e^{-t} + \frac{1}{b-1}e^{-bt}. \end{aligned}$$

Matrica sustava je

$$\begin{bmatrix} 0 & 1 \\ -b & -(b+1) \end{bmatrix}$$

a svojstvene vrijednosti su $\lambda_1 = -1$ i $\lambda_2 = -b$. Uvjet stabilnosti se svodi na

$$h < 2, \quad hb < 2.$$

U slučaju kad je b velik, drugi uvjet postaje ograničavajući. U tom slučaju rješenje ima dvije komponente (e^{-t} i e^{-bt}) koje se mijenjaju na dvije bitno različite vremenske skale. Za takav sustav kažemo da je **krut**.

Kod krutih sustava potrebno je koristiti implicitne metode kako bi se izbjegli mali vremenski koraci. Ako se traži metoda višeg reda može se primijeniti Richardsonova aproksimacija.