Numeeriset Menetelmät, 031022P

Keijo Ruotsalainen
Teknillinen tiedekunta
matematiikan jaos

11. helmikuuta 2010

Sisältö

Luku 1

Johdanto

1.1 Varoitus

Kädessäsi oleva luentomoniste perustuu teknillisessä tiedekunnassa pitämiini numeeristen menetelmien luentoihin. Monisteesta puuttuvat luennoilla käydyt esimerkit ja lauseiden todistukset. Koska moniste ei ole tarkoitettu itseopiskelumateriaaliksi, niin suosittelen osallistumista luennoille.

Edelleen pidätän oikeuden muutoksiin koskien kurssin sisältöä.

1.2 Numeeriset menetelmät insinööritietieteissä

Insinöörit valmistavat tuotteita ja palveluja inhimillisen elämän parantamiseksi. Insinööritieteet perustuvat luonnonlakien, -materiaalien ja -energialähteiden hyväksikäyttöön. Näitä tuotteita ja palveluja ovat energiansiirto, tietoliikenne, teollinen toiminta, kuljetus ja tiedon hallinta.

Teknillinen toiminta perinteisesti sisältää seuraavat asiat:

- tutkimus
- suunnittelu
- testaus
- valmistaminen

- koulutus

Suunnittelu on oleellinen osa insinöörin toimintaa. Hyvällä suunnittelulla voidaan pienentää ympäristön kuormitusta, tehdä turvallisempia ja luotettavampia tuotteita.

Luonnonkieli on matematiikka. Insinöörin on ymmärrettävä matematiikkaa voidakseen mallintaa matemaattisesti teknilliset ongelmat yhtälöiden muodossa. Matemaattisten yhtälöiden ratkaisujen tulee olla fysikaalisesti mielekkäitä, jotta niistä olisi jotain hyötyä. Yksinkertaisemmissa malleissa matemaattiset yhtälöt voidaan ratkaista analyyttisesti suljetussa muodossa, tai sarjakehitelmien avulla. Näissä apuvälineeksi riittää kynä, paperi ja hyvä sohva. Usein ongelmat ovat niin komplisoituja, että tarvitaan laskentakonetta ongelmien ratkaisemiseen. Esimerkiksi lineaarisen yhtälöryhmän ratkaiseminen voidaan periaatteessa suorittaa (jopa käytännössä) suorittaa kynällä ja paperilla kunhan tehtävän suorittajan pinna on riittävän pitkä. Gaussin menetelmä on ns. suora ratkaisumenetelmä, jonka avulla yhtälöryhmän ratkaiseminen voidaan suorittaa hyvin nopeasti ja erheettömästi; pyöristysvirheitä lukuunottamatta.

Varsin usein matemaattinen ongelma ei ole ratkaistavissa suljetussa muodossa. Tällöin yhtälön matemaattinen ratkaiseminen korvataan diskreetillä ratkaisuprosessilla, jonka avulla haetaan ratkaisulle riittävän hyvä approksimaatio l. likiratkaisu. Newtonin menetelmä epälineaaristen yhtälöiden ratkaisemiseksi on yksi tälläinen ratkaisumenetelmä. Näille menetelmille on tyypillistä, että ne etenevät askelittain alkuarvausta kohti ratkaisua. Nämä ovat iteratiivisia menetelmiä.

1.3 Mitä on numeerinen analyysi

Luonnontieteen ja tekniikan matemaattisten ongelmien numeeristen laskentamenetelmien

- johtaminen;
- analysointi;
- matemaattisten ongelmien konstruktiivinen ratkaiseminen

Esimerkki 1.1 *Tyypillisesti tekniikan ongelma mallinnetaan osittaisdifferentiaaliyhtälöiden avulla.*

- *Reaalimaailma;*
- *Matemaattinen malli; (puutteellinen informaatio)*
- *diskretisointi;*
- *Numeerinen malli ja algoritmi;*
- *tulos;*
- *virheanalyysi ja vertailu muihin tuloksiin*

1.3.1 Numeerinen algoritmi

Esimerkki 1.2 *Etsitään luvun neliöjuuri puolitusmenetelmällä: Olkoon* $a > 1$ *ja* $\epsilon > 0$.

1. *Alkuarvaus:* $x_0 = 1,\ x_1 = a$
2. $x = \frac{x_1+x_0}{2}$
   - *Kysymys: Onko* $x > \sqrt{a}$*?*
   - *Testi:* $x^2 > a$ *vai* $x^2 < a$*?*
3. *Jos* $x^2 > a$*, niin asetetaan* $x_0 = x$ *ja* $x_1 = a$*. Muussa tapauksessa* $x_0 = 1$ *ja* $x_1 = x$
4. *Testi: Onko* $x_1 - x_0 < 2\epsilon$*? Jos vastaus on kyllä, niin seuraava* x *on riittävän lähellä juurta. Muussa tapauksessa*
5. *Palaa kohtaan 2*

Tyypillisesti numeerinen algoritmi tuottaa jonon lukuja x_n (vektoreita, funktioita, jne.). Algoritmi on konvergoiva, jos jono suppenee kohti ongelman ratkaisua:

$$x_n \to x$$

Algoritmin analyysissa meidän on tutkittava lasketun approksimaation virhe l. suoritettava virheanalyysi.

Virhelähteet:

- lukujen pyöristys

- äärettömän jonon katkaiseminen
- inhimillinen erehdys

Virheiden tyypit:

- a priori-virhe ei ole riippuvainen lasketusta approksimaatiosta. Se voidaan arvioida ennen varsinaista laskentaa.
- a posteriori-virhe voidaan päätellä, kun approksimaatio on laskettu tai riittävä määrä termejä jonosta on laskettu.

Virheanalyysissa analysoidaan

- absoluuttista virhettä $|a - a^*|$
- suhteellista virhettä $\frac{|a-a^*|}{|a|}$

1.3.2 Hyvin asetettu ongelma

Oletetaan, että $S(d)$ esittää ongelman ratkaisua annetulle datalle d. Olkoon $S(d + \delta d)$ häirityn ongelman ratkaisu, missä δd on häiriö (Esimerkiksi mittausvirhe). Määritellään ei-negatiivinen luku

$$\|S(d + \delta d) - S(d)\|$$

ilmoittamaan ratkaisujen erotuksen suuruutta. Numeerinen ratkaisu on hyvin asetettu, jos seuraavat kaksi ehtoa ovat voimassa:

1. Jokaiselle datalle on olemassa yksikäsitteinen ratkaisu;
2. Ratkaisu $S(d)$ riippuu jatkuvasti datasta, ts.

$$\|S(d + \delta d) - S(d)\| \to 0, \text{ kun } \|\delta d\| \to 0$$

Numeerinen ratkaisu on **hyvin käyttyvä** (well posed), jos pieni häiriö datassa aiheuttaa pienen muutoksen ratkaisussa. Muutoin numeerinen prosessi on **huonosti käyttäytyvä** (ill posed).

Määritelmä 1.3.1 *Numeerinen prosessi on Lipschitz-jatkuva, jos on olemassa vakio $L > 0$ siten, että kaikille $\epsilon > 0$*

$$\|S(d + \delta d) - S(d)\| \leq L\|\delta d\|,$$

kun $\|\delta d\| < \epsilon$.

Jos numeerinen prosessi $S(d)$ on Lipschitz-riippuva datasta, niin se on hyvin käyttäytyvä.

1.3.3 Virhelähteet

Tavallisesti probleema P ei voida ratkaista suoraan; vaan sen sijaan ratkaistaan approksimaatio ongelma P_1. Tästä aiheutuu ns. globaali virhe E_s. Approksimaatio-ongelma P_1 vuorostaan ratkaistaan numeerisesti. Tässä prosessissa aiheutetaan typistysvirhe E_p. Numeerisissa menetelmissä ratkaisumenetelmien kehittelyn lisäksi on osattava arvioida ratkaisumenetelmän virhelähteet.

Tavallisesti numeeriset algoritmit perustuvat rekursiokaavaan

$$y_{n+1} = F(y_n, y_{n-1}, \ldots, y_0),\ n = m, m+1, m+2, \ldots,$$

missä $y_0, \ldots, y_{m-1}$ on probleeman alkudata.

Virheen analysoinnin lisäksi tutkitaan numeerisen ratkaisuprosessin stabiilisuutta: pieni virhe alkudatassa aiheuttaa pienen virheen ratkaisussa.

Luku 2

Numeerista lineaarialgebraa

2.1 Matriisialgebran kertaus

Matriisi on lukukaavio, jossa on m riviä ja n saraketta

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},$$

missä luvut a_{ij} ovat joko reaali- tai kompleksilukuja. Indeksi i alkion riviindeksi, ja j sarakeindeksi. Matriisi, jolla on vain yksi rivi tai sarake, sanotaan rivivektoriksi tai sarakevektoriksi. Mikäli matriisin rivien ja sarakkeiden lukumäärä on sama, niin matriisi on neliömatriisi. Tällöin matriisin *päädiagonaali* on $\mathrm{diag}(A) = (a_{11}, a_{22}, \ldots, a_{nn})$.

Matriisien laskuoperaatiot Olkoon $A = (a_{ij})$, $B = (b_{ij})$ kaksi $m \times n$ matriisia. Matriisit ovat samat, jos $a_{ij} = b_{ij}$ kaikille i ja j. Lisäksi matriiseille voidaan määritellä seuraavat laskuoperaatiot:

- *Summa*: $A + B = (a_{ij} + b_{ij}$;
- *Skalaarilla kertominen*: $kA = (ka_{ij})$;

- *Matriisitulo*: Olkoon A $(m \times p)$-matriisi ja B $(p \times n)$-matriisi. Tällöin tulomatriisi
$$AB = (c_{ij}) = (\sum_{k=1}^{p} a_{ik} b_{kj}).$$

Matriisitulo on liitännäinen ja assosiatiivinen; mutta ei yleensä vaihdannainen l kommutatiivinen. Siis yleensä matriiseille on voimassa $AB \neq BA$.

Yksikkömatriisi
$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$
on matriisikertolaskun suhteen yksikköalkio, ts. $AI = IA = A$.

Käänteismatriisi

Määritelmä 2.1.1 *Neliömatriisi A on kääntyvä (säännöllinen tai ei-singulaarinen), jos on olemassa matriisi B siten, että $AB = BA = I$. Matriisia B kutsutaan A:n käänteismatriisiksi, ja merkitään $B = A^{-1}$*

Tulomatriisin AB käänteismatriisi, jos se on olemassa, on $AB)^{-1} = B^{-1}A^{-1}$.

Matriisin säännöllisyydelle on seuraava hyödyllinen ominaisuus:

Lause 2.1.1 *Neliömatriisi on kääntyvä jos ja vain jos sen sarakevektorit ovat lineaarisesti riippumattomat.*

Määritelmä 2.1.2 *Matriisin $A \in \mathbb{R}^{m \times n}$ transpoosi on $(n \times m$-matriisi A^T, joka saadaan vaihtamalla matriisin A rivit matriisin A^T sarakkeiksi.*

Transpoosille on voimassa seuraavat ominaisuudet:
$$\begin{aligned} (A^T)^T &= A \\ (A + B)^T &= A^T + B^T \\ (AB)^T &= B^T A^T \\ (\alpha A)^T &= \alpha A^T \\ (A^T)^{-1} &= (A^{-1})^T. \end{aligned}$$

Matriisi on symmetrinen, jos $A^T = A$. Lopuksi matriisi on ortogonaalinen, jos $A^T A = AA^T = I$, ts. $A^{[-1} = A^T$.

Matriisin jälki ja determinantti Olkoon A neliömatriisi kertalukua n. Matriisin *jälki* on matriisin diagonaalialkioide summa:

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

Matriisin determinantti on skalaari

$$\det(A) = \sum_{\sigma \in P} \mathrm{sign}(\sigma) a_{i,\sigma(1)} a_{2,\sigma(2)} \cdots a_{n,\sigma(n)},$$

missä P indeksivektorin $i = (1, 2, 3 \ldots, n)$ kaikkien permutaatioiden l. järjestysten joukko. Permutaation merkki $\mathrm{sign}(\sigma)$ on 1 (-1), jos $\sigma(i)$ saadaan i:stä parillisella (parittomalla) määrällä paikanvaihtoja.

Matriisin determinantille pätee ominaisuudet:

$$\begin{aligned}
&\det(A) = \det(A^T) \\
&\det(AB) = \det(A)\det(B) \\
&\det(A^{-1} = \frac{1}{\det(A)} \\
&\det(kA) = k^n \det(A).
\end{aligned}$$

Merkitään matriisilla A_{ij} sellaista kertalukua $n-1$ olevaa matriisia, joka saadaan matriisista A poistamalla i:s rivi ja j:s sarake. Määritellään luku $\Delta_{ij} = (-1)^{i+j)} \det(A_{ij})$, jota kutsutaan matriisialkion a_{ij} kofaktoriksi. Tällöin matriisin determinantin laskemiseen voidaan käyttää ns. Laplacen sääntöä:

$$\det(A) = \sum_{j=1}^{n} \Delta_{ij} a_{ij} = \sum_{i=1}^{n} \Delta_{ij} a_{ij}.$$

Matriisin A käänteimatriisi, jos se on olemassa, voidaan laskea seuraavasti:

$$A^{-1} = \frac{1}{\det(A)} [\Delta_{ji}].$$

Näin ollen matriisi on kääntyvä, jos ja vain jos sen determinantti on nollasta eroava.

Matriisin aste ja ydin Olkoon matriisi A tyyppiä $m \times n$. Matriisin kuva-avaruus on

$$R(A) = \{y \in \mathbb{R}^m | \, y = Ax \text{ jollain } x \in \mathbb{R}^n\}.$$

Matriisin A aste $\text{rank}(A)$ on kuva-avaruuden dimensio, l. lineaarisesti riippumattomien sarakevektoreiden lukumäärä. Mikäli matriisi on neliömatriisi ja matriisin aste on sama kuin sarakkeiden lukumäärä, on vastaavalla yhtälöryhmällä $Ax = f$ yksikäsitteinen ratkaisu jokaisella vektorilla f.

Matriisin ydin on niiden vektoreiden joukko, joille $Ax = 0$:

$$\text{Ker}(A) = \{x \in \mathbb{R}^n | \; Ax = 0\}.$$

Matriisin asteelle on voimassa seuraavat ominaisuudet:

1. $\text{rank}(A) = \text{rank}(A^T)$;
2. $\text{rank}(A) + \dim(\ker(A) = n$.

Jos matriisi on säännöllinen, niin $\text{rank}(A) = n$ ja $\dim \ker(A) = 0$.

Lopuksi toteamme, että neliömatriislle seuraavat ominaisuudet ovat yhtäpitäviä:

1. A on säännöllinen l. A^{-1} on olemassa;
2. $\det(A) \neq 0$;
3. $\ker(A) = \{0\}$;
4. $\text{rank}(A) = n$;
5. A:n sarake- ja rivivektorit ovat lineaarisesti riippumattomia;
6. YHtälöryhmällä $Ax = f$ on yksikäsitteinen ratkaisu jokaiselle f.

Jos ratkaisujoukko on ääretön, niin tavallisesti etsitään sellaista ratkaisua, joka täyttää annetun side-ehdon. Jos taas ratkaisujoukko on tyhjä (ylimääräytynyt tehtävä), niin haetaan sellaista vektoria x, joka on mahdollisimman lähellä ratkaisua, ts. minimoidaan vektorin $Ax - f$ pituus, jonkin normin mielessä.

Yksikäsitteisen ratkaisun yhteydessä pyritään kehittelemään sellaisia ratkaisumenetelmiä jotka soveltuvat mahdollisimman hyvin tietokoneella suoritettavaksi. Ratkaisumenetelmä ei saisi olla herkkä liukulukujen pyöristysvirheille, ja ratkaisun laskeminen ei

saisi viedä liikaa koneen kapasiteettia eikä se saisi kestää kohtuuttoman kauan aikaa.

Esimerkiksi Stealth-hävittäjien (tutkassa näkymättön lentokone) suunnittelussa joudutaan ratkomaan lineaarisia yhtälöryhmiä, joissa on jopa 10^8 tuntematonta muuttujaa. Matriisin koko on tällöin 10^{16} alkiota. Klassinen Gaussin menetelmä antaa kyllä yhtälölle tarkan ratkaisun, jota lapsenlapsemme voivat sitten analysoida. Mutta Stealth-hävittäjiä on. Kuinka yhtälöt on ratkaistu?

Erikoismatriisit Usein sovelluksissa törmätään erikoistyyppisiin matriiseihin, joiden käsittely on helppoa (?!). Tällaisia matriiseja ovat ylä- ja alakolmiomatriisit. Näissä matriiseissa kullakin sarakkeella lävistäjäalkion ala- tai yläpuolella on pelkkiä nollia:

$$L = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix}, \; U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & l_{n3} & \cdots & u_{nn} \end{bmatrix}.$$

Ylä- ja alakolmiomatriisien ominaisuuksia ovat:

- Deerminantti on diagonaalialkioide tulo;
- yläkolmio- ja alakolmiomatriisin käänteismatriisi on myös yläkolmio- ja alakolmiomatriisi.
- Kahden alakolmiomatriisin tulo on edelleen alakolmiomatriisi; vastaava pätee yläkölmiomatriiseille.

Ominaisarvot Luku $\lambda \in \mathbb{C}$ on **ominaisarvo**, jos se toteuttaa karaktersitisen yhtälön $\det(A - \lambda I) = 0$, ja vektori $u_\lambda \in \mathbb{R}^n$ on ominaisarvoon λ liittyvä **ominaisvektori**, jos

$$Au_\lambda = \lambda u_\lambda.$$

Karakteristinen yhtälö on n-asteinen polynomiyhtälö, jolla on algebran peruslauseen perusteeella aina n kappaletta ominaisarvoja (, kun A on $n \times n$-matriisi), ja matriisin detrminantti ja jälki ovat

$$\det(A) = \Pi_{i=1}^n \lambda_i(A), \; \mathrm{tr}(A) = \sum_{i=1}^n \lambda_i.$$

Matriisin **spektraalisäde** on

$$\rho(A) = \max_{1 \leq i \leq} |\lambda_i(A)|.$$

Ominaisarvojen laskeminen tuottaa tosin käytännössä suuria hankaluuksia. Ensinnäkin voidaan osoittaa, että viisi ja sitä korkeampi asteisille polynomeille ei ole nollakohtien ratkaisukaavaa. Nollakohtien määräminen on siksi suoritettava numeerisesti, joka on numeerisesti varsin epästabiili operaatio. Tarkastelemme tätä ongelmaa hieman epälineaaristen yhtälöiden ratkaisumenetelmien yhteydessä luvussa 6. Tässä luvussa tarkastelemme joitain matriisialgebran menetelmiä ominaisarvojen määräämiseen ilman karakteristisen yhtälön ratkaisemista.

Esimerkkinä olkoon reaaliset ja symmetriset matriisit ($A^T = A$). Reaalisen ja symmetrisen matriisin kaikki ominaisarvot ovat aina reaalisia. Lisäksi matriisin ominaisvektoreista voidaan muodostaa vektoriavaruudelle $\mathbb{R}^n$ ortonormaalikanta. Edelleen symmetrisen matriisin ominaisvektoreista voidaan muodostaa ortogonaalinen matriisi Q, joka diagonalisoi matriisin A:

$$Q^T A Q = D,$$

missä matriisi D on diagonaalimatriisi, jonka lävistäjäalkioina ovat matriisin A ominaisarvot.

Yleisesti neliömatriisi on diagonalisoituva, jos on olemassa matriisi U siten, että

$$U^{-1} A U = D.$$

Edelleen on voimassa seuraava

Lause 2.1.2 *jokaisella matriisilla on nk. singulaariarvohajotelma*

$$U^T A V = \Lambda,$$

missä diagonaalimatriisin Λ diagonaalialkiot ovat matriisin $A^T A$ ominaisarvojen positiiviset neliöjuuret, ja matriisit U ja V ovat sarakeortogonaalisia.

Matriisialgebran kurssilla osoitettiin vielä, että jokaisella matriisilla on QR-hajotelma: $A = QR$, missä Q on sarakeortogonaalinen ja R on yläkolmiomatriisi.

Symmetrinen matriisi on **positiivisesti definiitti**, jos kaikilla $x \neq 0$

$$x^T A x > 0.$$

Tällöin ominaisarvot ovat reaalisia ja positiivisia.

Vektori- ja matriisinormit

Vektorinormi $\|\cdot\|$ on kuvaus $\mathbb{R}^n \to \mathbb{R}_+$ siten että

1. $\|x\| > 0,\ x \neq 0$ ja jos $\|x\| = 0$, niin $x = 0$;
2. $\|\lambda x\| = |\lambda|\|x\|$;
3. $\|x + y\| \leq \|x\| + \|y\|$.

Tavallisimpia vektorinormeja ovat:

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i|, \\ \|x\|_2 &= [\sum_{i=1}^{n} |x_i|^2]^{\frac{1}{2}}, \\ \|x\|_\infty &= \max_{1 \leq i \leq n} |x_i|. \end{aligned}$$

Vektoreiden välinen euklidinen sisätulo on kuvaus $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ siten, että

$$x^T y = (x, y) = \sum_{i=1}^{n} x_i y_i$$

Selvästi sisätulolle on voimassa seuraavat ominaisuudet:

1. Se on lineaarinen: $(ax + by, z) = a(x, z) + b(y, z)$;
2. Symmetrisyys: $(x, y) = (y, x)$;
3. Positiviisuus: $(x, x) > 0$ kaikille $x \neq 0$, ja $(x, x(= 0$, jos ja vain jos $x = 0$.

Lause 2.1.3 *Cauchy-Schwarz Kaikille pareille x, y:*

$$|(x, y)| = |x^T y| \leq \|x\|_2 \|y\|_2.$$

Yhtälö on voimassa, jos ja vain jos $x = ky$ jollain $k \in \mathbb{R}$.

Vektorijono $\{x^{(k)}\}_{k\in\mathbb{N}}$ suppenee kohti vektoria x, jos

$$\lim k \to \infty x_i^{(k)} = x_i, \ \forall i = 1, 2, 3 \ldots, n.$$

Helposti todetaan, että

$$\lim_{k\to\infty} x^{(k)} = x \ \Leftrightarrow \ \lim_{k\to\infty} \|x - x^{(k)}\| = 0$$

jokaiselle vektroinormille, joka on määritelty $\mathbb{R}^n n$:ssä.

Matriisinormi toteuttaa seuraavat määrittelevät ehdot

1. $\|A\| > 0, \ A \neq 0$ ja jos $\|A\| = 0$, niin $A = 0$;
2. $\|\lambda A\| = |\lambda|\|A\|$;
3. $\|A + B\| \leq \|A\| + \|B\|$;
4. $\|AB\| \leq \|A\|\|B\|$.

Indusoitu matriisinormi määritellään vektorinormin avulla:

$$\|A\| = \max_{x=1} \ \|Ax\|.$$

Esimerkiksi seuraavat matriisinormit ovat vastaavien vektorinormien indusoimia:

$$\begin{aligned} \|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \\ \|A\|_2 &= \text{A:n suurin singulaariarvo} \\ \|A\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \end{aligned}$$

Frobenius-normi

$$\|A\|_F = [\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2]^{\frac{1}{2}}$$

ei ole minkään vektorinormin indusoima matriisinormi.

Vektori- ja matriisinormin yhteensopivuus : Vektori- ja matriisinormi ovat yhteensopivia, jos

$$\|Ax\| \leq \|A\|\|x\|.$$

Huom! Vaikka Frobenius-normi ei ole l_2-normin indusoima, niin silti on voimassa

$$\|Ax\|_2 \leq \|A\|_F\|x\|_2.$$

2.2 Suorat ratkaisumenetelmät

2.2.1 Gaussin menetelmä ja LU-hajotelma

Yhtälöryhmä: $Ax = f$, missä kerroinmatriisi on $A = [a_{ij}]_{i,j=1,\ldots,n}$ on neliömatriisi ja datavektori

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}.$$

Alkeisrivimuunnoksilla (yhtälöryhmässä yksi yhtälö voidaan kertoa luvulla ja lisätä muihin yhtälöihin ei muuta yhtälöryhmän ratkaisujoukkoa) matriisiyhtälö pyritään muuntamaan sellaiseen yhtäpitävään muotoon, josta yhtälöryhmän ratkaisu on helppo määrittää.

Merkitään jatkossa matriisilla $A^{(k)} = [a_{ij}^{(k)}]$ kerroinmatriisia, joka on saatu k:n eliminaatioaskeleen jälkeen. Vastaavasti $f^{(k)}$ on vastaava oikeanpuolen vektori. Lisäksi asetetaan $A^{(1)} = A$ ja $f^{(1)} = f$.

Olkoon $A^{(k-1)}$ määrätty ja vastaava oikeanpuolen vektori. Kun on suoritettu k-2 eliminaatioaskelta on muodostettu yhtälöryhmä $A^{(k-1)}x = f^{(k-1)}$. Yhtälöryhmän kerroinmatriisin $A^{(k-1)}$ sarakkeilla $j = 1, \ldots, k-2$ on pelkkiä

nollia lävistäjäalkion alapuolella:

$$A^{(k-1)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1,k-2}^{(1)} & a_{1,k-1}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2,k-2}^{(2)} & a_{2,k-1}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3,k-2}^{(3)} & a_{3,k-1}^{(3)} & \cdots & a_{3n}^{(3)} \\ 0 & 0 & 0 & & & & & \\ & & & \cdots & a_{k-2,k-2}^{(k-2)} & a_{k-2,k-1}^{(k-2)} & \cdots & a_{k-2,n}^{(k-2)} \\ & & & \cdots & 0 & a_{k-1,k-1}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n,k-1}^{(k-1)} & \cdots & a_{n,n}^{(k-1)} \end{bmatrix}.$$

Oikeanpuolen vektori on tällöin

$$f = \begin{bmatrix} f_1^{(1)} \\ f_2^{(2)} \\ f_3^{(3)} \\ \vdots \\ f_{k-2}^{(k-2)} \\ f_{k-1}^{(k-1)} \\ \vdots \\ f_n^{(k-1)} \end{bmatrix}.$$

Oletetaan lisäksi, että jokaisella askeleella $a_{k,k}^{(k)} \neq 0$. Suoritetaan uusi eliminaatioaskel seuraavasti: kun $k = 2, 3, \ldots, n$

$$a_{i,j}^{(k)} = \begin{cases} a_{i,j}^{(k-1)}, \ i \leq k-1 \\ 0, \ i \geq k, \ j \leq k-1 \\ a_{i,j}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)}, \ i \geq k, \ j \geq k \end{cases}$$

$$f_i^{(k)} = \begin{cases} f_i^{(k-1)}, \ i \leq k-1 \\ f_i^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} f_{k-1}^{(k-1)} \end{cases}$$

Suorittamalla n-1 eliminaatioaskelta päädytään lopulta yhtälöryhmään, jossa kerroinmatriisina U on yläkolmiomatriisi. Olkoon matriisit $A^{(k)}, k = 1, \ldots, n$ kuten yllä tällöin on voimassa seuraava

Lause 2.2.1 (LU-hajotelma) *Oletetaan, että matriisille A suoritetuissa eliminaatioaskelissa kertoimet $a_{k,k}^{(k)} \neq 0,\ k = 1, \ldots, n$. Tällöin matriisin A determinantti on*

$$\det(A) = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)},$$

ts. yhtälöryhmällä on yksikäsitteinen ratkaisu. Lisäksi matriisille A on voimassa hajotelma $A = LU$, missä $U = A^{(n)}$ on yläkolmiomatriisi ja L on alakolmiomatriisi, jonka alkiot ovat

$$m_{i,k} = \begin{cases} 0, & i < k \\ 1, & i = k \\ \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, & i > k \end{cases}.$$

Tod.: Mikäli matriisille on voimassa LU-hajotelma, niin

$$\det(A) = \det(L)\det(U) = 1 \cdot a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)},$$

sillä ylä- ja alakolmiomatriisin determinantit ovat niiden diagonaalialkioiden tulo.

Osoitetaan seuraavaksi, että A:lle on voimassa LU-hajotelma: Olkoon siten matriisitulo $LU = [c_{i,j}]$. Koska matriisi L on alakolmiomatriisi ja vastaavasti U yläkolmiomatriisi, niin alkio

$$c_{i,j} = \sum_{k=1}^{n} m_{i,k} a_{k,j}^{(k)} = \sum_{k=1}^{\min(i,j)} m_{i,k} a_{k,j}^{(k)}.$$

Eliminaatioaskeleen nojalla matriisin $A^{(k)}$ alkioille on voimassa

$$a_{i,j}^{(k)} = a_{i,j}^{(k-1)} - m_{i,k-1} a_{k-1,j}^{(k-1)},\ 2 \leq k \leq i,\ k \leq j.$$

Joten jos $i \leq j$, niin

$$\begin{aligned} c_{i,j} &= \sum_{k=1}^{i} m_{i,k} a_{k,j}^{(k)} \\ &= [\sum_{k=1}^{i-1} m_{i,k} a_{k,j}^{(k)}] + a_{i,j}^{(i)} \\ &= \sum_{k=1}^{i-1} (a_{i,j}^{(k)} - a_{i,j}^{(k+1)}) + a_{i,j}^{(i)} \\ &= a_{i,j}^{(1)} = a_{i,j}. \end{aligned}$$

Vastaavasti jos $i > j$, niin $a_{i,j}^{(j+1)} = 0$ ja siten

$$\begin{aligned} c_{i,j} &= \sum_{k=1}^{j} m_{i,k} a_{k,j}^{(k)} \\ &= \sum_{k=1}^{j} (a_{i,j}^{(k)} - a_{i,j}^{(k+1)}) \\ &= a_{i,j}^{(1)} - a_{i,j}^{(j+1)} = a_{i,j}^{(1)}. \end{aligned}$$

□

Yhtälöryhmä voidaan nyt ratkaista helposti käyttäen kerroinmatriisin LU-hajotelmaa:

1. Ratkaise g yhtälöryhmästä $Lg = f$ eteenpäin sijoituksella;

$$\begin{aligned} g_1 &= f_1 \\ g_i &= f_i - \sum_{k=1}^{i-1} m_{i,k} f_k, \ i = 2, \ldots, n. \end{aligned}$$

2. Ratkaise x yhtälöryhmästä $Ux = g$ taaksepäin sijoituksella;

$$\begin{aligned} x_n &= \frac{1}{u_{nn}} g_n, \\ x_i &= \frac{1}{u_{ii}} [g_i - \sum_{j=i+1}^{n} u_{i,j} x_j]. \end{aligned}$$

Pivotisointistrategiat

Oletetaan, että on suoritettu $k = 1, ..., n-1$ eliminointisaskelta jolloin yhtälöryhmän kerroinmatriisi on saatettu rivioperaatioilla muotoon

$$A^{(k-1)} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1,k-2} & a_{1,k-1} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2,k-2}^{(1)} & a_{2,k-1}^{(1)} & \cdots & a_{2,n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3,k-2}^{(2)} & a_{3,k-1}^{(2)} & \cdots & a_{3,n}^{(2)} \\ 0 & 0 & 0 & & & & & \\ & & & \cdots & a_{k-1,k-1}^{(k-2)} & a_{k-1,k-1}^{(k-2)} & \cdots & a_{k-1,n}^{(k-2)} \\ & & & \cdots & 0 & a_{k,k}^{(k-1)} & \cdots & a_{k,n}^{(k-1)} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n,k-1}^{(k-1)} & \cdots & a_{n,n}^{(k-1)} \end{bmatrix}.$$

Seuraava muuttuja, joka eliminoidaan jäljellä jäävästä yhälöryhmästä (n-k:sta viimeisestä yhtälöstä) voidaan valita käyttäen joko yksinkertaista tai osittaista pivot-strategiaa.

Yksinkertainen pivotisointi Pivot-alkioksi valitaan diagonaalialkio $a_{k,k}^{(k-1)}$, mikäli se on nollasta eroava. Jos ko. alkio on nolla, niin valitaan seuraava alkio.

Osittainen pivotisointi Pivot-alkioksi valitaan

$$|a_{l,k}^{(k-1)}| = \max_{j=k,\ ...,\ n} |a_{j,k}^{(k-1)}|.$$

2.2.2 Stabiilisuusanalyysi

Matriisin ehtoluku Matriisin A ehtoluku

$$K(A) = \|A\|\|A^{-1}\|,$$

missä $\|A\|$ on joku indusoitu matriisinormi. Ehtoluku riippuu valittavsta normista. Tavallisesti kuitenkin käytetään $\|A\|_1,\ \|A\|_2,\ \|A\|_\infty$ normeja ehtoluvun määrittelyyn. Tällöin merkitään ehtoluvulle alaindeksi $K_1(A),\ K_2(A)$ tai $K_\infty(A)$.

Koska

$$1 = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = K(A),$$

niin aina $K(A) \geq 1$, matriisn ja sen käänteismatriisin ehtoluvut ovat yhtä suuria.

Kun $p = 2$, niin

$$K_2(A) = \frac{\sigma_1(A)}{\sigma_n(A)},$$

missä σ_1 on matriisin suurin ja σ_n pienin singulaariarvo.

Jos matriisi A on symmetrinen ja positiivisesti definiitti, niin tällöin sen ominaisarovt ovat samalla sen singulaariarvoja. Tällöin matriisin A ehtoluku

$$K_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Virheanalyysi Erilaisten pyöristysten ja typistysten takia yleensä yhtälöryhmän

$$Ax = b \tag{2.1}$$

sijasta lasketaankin häiritty yhtälöryhmä

$$(A + \delta A)(x + \delta x) = b + \delta b. \tag{2.2}$$

Seuraavan lauseen avulla voidaan arvioida matriisin A ja vektorin b pienistä häiriöistä johtuva suhteellinen virhe ratkaisussa.

Lause 2.2.2 *Olkoon A säännöllinen matriisi ja δA pieni häiriö siten, että*

$$\|A^{-1}\|\|\delta A\| < 1.$$

Tällöin, jos x on yhtälöryhmän (2.1) ratkaisu ja δx toteuttaa yhtälön (2.2), niin ratkaisun suhteellinen virhe

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\delta a\|}{\|A\|}}\left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right).$$

Lauseen todistuksessa tarvitaan seuraavaa aputulosta:

Lemma 2.2.1 *Olkoon matriisin B normi $\|B\| < 1$. Tällöin matriisi $I + B$ on säännöllinen ja sen käänteismatriisi on*

$$(I + B)^{-1} = \sum_{k=0}^{\infty}(-1)^k B^k$$

ja käänteismatriisin normille on voimassa arvio

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Tod.: Potenssisarja $\sum_{k=0}^{\infty}(-1)^k B^k$ on hyvin määritelty, l. sarja suppenee, koska

$$\|\sum_{k=0}^{\infty}(-1)^k B^k\| \leq \sum_{k=0}^{\infty}\|B\|^k < \infty.$$

Jokaiselle $n \in \mathbb{N}$

$$(I + B)\sum_{k=0}^{n}(-1)^k B^k = I + (-1)^n B^n.$$

Näin ollen

$$\lim_{n=1}(I+B)\sum_{k=0}^{n}(-1)^k B^k = I.$$

Kirjoittamalla yksikkömatriisi muodossa

$$I = I - B + B$$

ja kertomalla se oikealta matriisilla $(I+B)^{-1}$ saadaan identiteetti

$$(I+B)^{-1} = -B(I+B)^{-1} + I.$$

Matriisinormin kolmioepäyhtälön nojalla on siten voimassa epäyhtälö

$$\|(I+B)^{-1}\| \leq 1 + \|B\|\|(I+B)^{-1}\|,$$

josta helposti saadaan väite:

$$\|(I+B)^{-1}\| \leq \frac{1}{1-\|B\|}.$$

□

Tod.:[Lauseen 2.2.1 todistus] Matriisille $B = A^{-1}\delta A$ on voimassa edellisen lemman ehdot, sillä

$$\|A^{-1}\delta A\| \leq \|A^{-1}\|\|\delta A\| < 1.$$

Näin ollen on voimassa normiepäyhtälö

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1-\|A^{-1}\|\|\delta A\|}. \tag{2.3}$$

Toisaalta ratkaisun virhe δx voidaan kirjoittaa muodossa

$$\delta x = (I + A^{-1}\delta A)^{-1}A^{-1}(\delta b - \delta A x).$$

Käyttämällä epäyhtälöä (2.3) ja vektori- ja matriisinormin yhteensopivuutta saadaan

$$\|\delta x\| \leq \frac{1}{1-\|A^{-1}\|\|\delta A\|}\|A^{-1}\|[\|\delta b\| + \|\delta A\|\|x\|],$$

josta jakamalla puolittain ratkaisun normilla $\|x\|$ saadaan väite. □

Lause 2.2.3 *Olkoon edellisen lauseen ehdot voimassa ja $\delta A = 0$. Tällöin*

$$\frac{1}{K(A)}\frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq K(A)\frac{\|\delta b\|}{\|b\|}.$$

Edellisten lauseiden valossa matriisin ehtoluku on ratkaiseva suure häiriöiden δA ja δb aiheuttamaan muutokseen ratkaisussa.

Oletetaan, että d-paikkaisessa liukulukuaritmetiikassa matriisin A ja oikean puolen vektorin b häiriöiden suhteellinen virhe on luokkaa $5 \cdot 10^{-d}$, ts.

$$\frac{\|\delta A\|}{\|A\|} \approx 5 \cdot 10^{-d}, \ \frac{\|\delta b\|}{\|b\|} \approx 5 \cdot 10^{-d},$$

ja että matriisin ehtoluku $K(A) = 10^{\alpha}$, missä α on siten, että $5 \cdot 10^{\alpha - d} < 1$. Tällöin edellisen lauseen nojalla ratkaisun muutoksen suhteellinen virhe

$$\frac{\|\delta x\|}{\|x\|} \leq 10^{\alpha - d + 1}.$$

Näin ollen käytännössä voidaan käyttää seuraavaa peukalosääntöä: *Jos matriisin A ehtoluku on suuruusluokkaa 10^{α}, niin d-paikkaisessa liukulukuaritmetiikassa korkeintaan $d - \alpha - 1$ desimaalia on oikein.*

Esimerkki 2.1 *Lineaarisen yhtälöryhmän*

$$\begin{bmatrix} 0.99 & 0.98 \\ 0.98 & 0.97 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.97 \\ 1.95 \end{bmatrix}$$

ratkaisu on $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$*. Laske matriisin ehtoluku $K_2(A)$. Arvioi ratkaisun suhteellisen virheen herkkyyttä yhtälöryhmän kertoimien pienille muutoksille. Yhtälöryhmän sijasta ratkaistaan häiritty yhtälö*

$$\begin{bmatrix} 0.990005 & 0.979996 \\ 0.979996 & 0.970004 \end{bmatrix} \begin{bmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \end{bmatrix} = \begin{bmatrix} 1.969967 \\ 1.950035 \end{bmatrix}$$

Arvioi suhteellista virhettä ehtoluvun avulla ja laske todellinen suhteellinen virhe.

2.3 QR-hajotelma

2.3.1 Householderin muunnos

Olkoon vektori $w \in \mathbb{R}^n$ sellainen, että sen l_2-normi $\|w\|_2 = 1$, ts. $w^T w = 1$, missä w^T on matriisin w transponoitu vektori. Matriisi

$$H = I - 2ww^T$$

on Householderin muunnos. Sille on voimassa seuraavat ominaisuudet:

Lause 2.3.1 *Jos $H = I - 2ww^T$ on Householderin muunnos, niin H on symmetrinen ja ortogonaalinen matriisi. Siten $H^{-1} = H$.*

Householder-muunnosta sovelletaan tavallisesti matriisien käsittelyyn. Se muuttaa matriisin A matriisiksi

$$\tilde{A} = HA = A - 2ww^T A = A - 2wu^T.$$

Mikäli A on $n \times k$-matriisi, niin Householder-muunnos voidaan tehdä seuraavalla algoritmilla:

1. $u^T = w^T A$;
2. $\tilde{A} = A - 2wu^T$.

Algoritmin suorittamiseksi tarvitaan $2kn$ kertolaskua.

Olkoon $x \in \mathbb{R}^n$ nollasta eroava vektori siten, että sen L^2-normi

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2 = 1}.$$

Määritellään vektori

$$w = \frac{1}{\sqrt{2}} \frac{x \pm e_1}{\sqrt{1 \pm x_1}},$$

missä vektori e_1 on $\mathbb{R}^n$:n ensimmäinen kantavektori. Vektorin w määritelmässä valitaan merkki sen mukaan, onko vektorin x ensimmäinen koordinaatti positiivinen tai negatiivinen. Jos $x_1 < 0$, niin valitaan $-x_1$. Tällä valinnalla nimittäjä ei ole koskaan nolla. Nyt vektori w on myös yksikkövektori:

$$\begin{aligned} w^T w &= \frac{1}{2} \frac{1}{1 \pm x_1} (x \pm e_1)^T (x \pm e_1) \\ &= \frac{1}{2(1 \pm x_1)} [x^T x \pm 2x^T e_1 + e_1^T e_1] \\ &= \frac{2(1 \pm x_1)}{2(1 \pm x_1)} = 1. \end{aligned}$$

Vektorin x Householder-muunnos on siten

$$Hx = x - 2(w^T x)x,$$

missä

$$w^T x = \frac{(x \pm e_1)^T x}{\sqrt{2(1 \pm x_1)}} = \sqrt{\frac{1 \pm x_1}{2}}.$$

Näin ollen vektorin x Householder-muunnos on

$$\begin{aligned} Hx &= x - 2\sqrt{\frac{1 \pm x_1}{2}} \frac{x \pm e_1}{\sqrt{2(1 \pm x_1)}} \\ &= x - (x \pm e_1) = \pm e_1. \end{aligned}$$

Yleisesti on voimassa

Lemma 2.3.1 *Olkoon $x \in \mathbb{R}^n$, $x \neq 0$. Tällöin vektorin x Householder-muunnos on*

$$Hx = (I - 2ww^T)x = \pm \|x\|_2 e_1,$$

kun muunnosvektori on

$$w = \frac{x}{\|x\|_2}.$$

2.3.2 QR-hajotelma

Pienimmän neliösumman menetelmässä ratkaistaan *ylimääräytynyt lineaarinen yhtälöryhmä* $Ax = f$, missä yhtälöiden lukumäärä n on suurempi kuin tuntemattomien lukumäärä k. Yleensä yhtälöryhmällä ei ole ratkaisua lainkaan. Tällöin haetaan "ratkaisua"x, joka minimoi L^2-normin neliön:

$$\min_{x \in \mathbb{R}^n} \|Ax - f\|_2^2 = \sum_{i=1}^{n} |\sum_{j=1}^{k} a_{ij} x_j - f_i|^2.$$

Vektorin $Ax - f$ pituus on pienimmillään, kun se on kohtisuorassa A:n virittämää kuva-avaruutta vasten.

Seuraavassa tarkastellaan ratkaisumenetelmää, joka perustuu Householderin muunnoksen käyttöön, jonka avulla matriisille A muodostetaan hajotelma ortogonaalisen ja yläkolmiomatriisin avulla.

Lause 2.3.2 *Olkoon A $n \times k$-matriisi, jonka aste on k. Silloin on olemassa ortogonaalinen matriisi Q siten, että*

$$Q^T A = R,$$

missä R on yläkolmiomatriisi, jonka lävistäjäalkiot ovat positiivisia.

Edellisessä kappaleessa muodostettiin vektorin x QR-hajotelma Householderin muunnoksen avulla. Kohdistetaan Householderin muunnosta rekursiivisesti matriisin A sarakkeisiin.

Olkoon matriisin A sarakevektorit $\underline{a}_j$, $j = 1, \ldots, k$. Aluksi suoritetaan Householderin muunnos

$$H_1 = I - 2\underline{w}_1 \underline{w}_1^T$$

matriisin A ensimmäisen sarakevektorin avulla. Normitetaan ensimmäinen sarake jakamalla se L^2-normilla:

$$\underline{x_1} = \frac{\underline{a}_1}{\|\underline{a}_1\|_2}.$$

Määritellään sitten vektori vektori $\underline{w}_1$ asettamalla

$$\underline{w}_1 = \frac{\underline{x}_1 \pm \underline{e}_1}{\sqrt{2(1 \pm x_{11})}}$$

riippuen vektorin $\underline{x}_1$ ensimmäisen koordinaatin x_{11} merkistä. Kertomalla matriisi A Householderin matriisilla H_1 saadaan matriisi

$$A^{(2)} = H_1 A = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & \cdots & a_{1k}^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2k}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3k}^{(2)} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & a_{n2}^{(2)} & a_{n3}^{(2)} & \cdots & a_{nk}^{(2)} \end{bmatrix}.$$

Oletetaan sitten, että on määrätty matriisi $A^{(m)}$:

$$A^{(k)} = \begin{bmatrix} a_{11}^{(m)} & a_{12}^{(m)} & a_{13}^{(m)} & \cdots & a_{1m}^{(m)} & \cdots & a_{1k}^{(m)} \\ 0 & a_{22}^{(m)} & a_{23}^{(m)} & \cdots & a_{2m}^{(m)} & \cdots & a_{2k}^{(m)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3m}^{(m)} & \cdots & a_{3k}^{(m)} \\ \vdots & \vdots & \vdots & \cdots & a_{mm}^{(m)} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nm}^{(m)} & \cdots & a_{nk}^{(m)} \end{bmatrix}.$$

Suoritetaan matriisin $A^{(m)}$ Householderin muunnos matriisilla

$$H_m = \begin{bmatrix} I_{m-1} & 0 \\ 0 & H_{mm} \end{bmatrix},$$

missä H_{mm} on $(n-m)$-ulotteinen Householderin matriisi

$$H_{mm} = I_{n-m} - 2\underline{w}_m \underline{w}_m^T.$$

Vektori $\underline{w}_m$ muodostetaan matriisin $A^{(m)}$ m:nnen sarakevektorin $(n-m)$:n viimeisen koordinaatin avulla. Määritellään sitä varten $(n-m)$-ulotteinen yksikkövektori

$$\underline{z} = \sqrt{\frac{1}{\sum_{i=m}^{n} |a_{im}^{(m)}|^2}} \begin{bmatrix} a_{mm}^{(m)} \\ a_{m+1,m}^{(m)} \\ \vdots \\ a_{nm}^{(m)} \end{bmatrix}.$$

Tämän jälkeen vektori

$$\underline{w}_m = \frac{1}{\sqrt{2(1 \pm z_1)}} \begin{bmatrix} z_1 \pm 1 \\ z_2 \\ \vdots \\ z_{n-m} \end{bmatrix}.$$

Tällöin matriisin $A^{(m+1)} = H_m A^{(m)}$ m:nnellä sarakkeella lävistäjäalkion alapuolella on pelkkiä nollia. Näin jatkamalla lopulta k:n askeleen jälkeen matriisi $A^{(k)}$ on yläkolmiomatriisi. Ja siten edellinen lause on todistettu. (Huom! Miten saadaan lävistäjäalkiot positiivisiksi?).

2.4 Konjugaattigradienttimenetelmä

Oletetaan, että yhtälöryhmän kerroinmatriisi on symmetrinen ja positiivisesti definiitti. Tällöin yhtälöryhmä $Ax = b$ on yhtäpitävä seuraavan minimointiongelman kanssa:

$$\min_x \ \frac{1}{2} x^T A x - b^T x.$$

Konjugaattigradienttimenetelmän peerusominaisuudet ovat

Ratkaisu $x = A^{-1}b$ saavutetaan iteratiivisesti n:llä iteraatiolla (eksaktissa aritmetiikassa).

Jokainen välivaihe x_k on minimointiongelman ratkaisu.

Jokainen muutos "$x_k \to x_{k+1}$"on konjugoitu kaikkiin edellisiin muutoksiin nähden.

Jokaisella iteraatiolla lasketaan ns. hakusuunta d_{k+1} ja jäännösvektori $r_k = b - Ax_k$.

Määritelmä 2.4.1 *Vektorit $d_1, \ldots, d_k$ ovat konjugoituja l. A-ortogonaalisia, jos*

$$d_i^T A d_j = 0, \ i \neq j.$$

Funktion $f(x) = \frac{1}{2}x^T Ax - b^T x$ gradientti

$$\nabla f(x) = Ax - b.$$

Initialisointi Aloituspisteeksi valitaan $x_0 = 0$, hakusuunnaksi funktion f(x) gradientti ko. pisteessä: $d_1 = r_0 = b - Ax_0$.

Funktion $g(\alpha) = f(x_0 + \alpha d_1)$ minimi löytyy derivaatan nollakohdasta:

$$\phi(\alpha_1) = \nabla f(x_0 + \alpha_1 d_1)^T d_1 = \alpha_1 d_1^T A d_1 - b^T d_1 = 0.$$

Joten

$$\alpha_1 = \frac{b^T d_1}{d_1^T A d_1}.$$

Yleinen iteraatio Olkoon edellesillä iteraatioilla määrätty pisteet $x_0, \ x_1, \ \ldots, \ x_k$, jäännösvektorit $r_0, \ r_1, \ \ldots, \ r_k = b - Ax_k$ ja konjugoidut suunnat $d_1, \ d_2, \ \ldots, \ d_k$.

Uusi konjugoitu suunta $d_{k+1} = r_k + \beta_{k+1} d_k$. Kerroin β_{k+1} määräätään A-ortogonaalisuuden avulla:

$$d_{k+1}^T A d_k = r_k^T A d_k + \beta_{k+1} d_k^T A d_k = 0.$$

Näin ollen kerroin

$$\beta_{k+1} = -\frac{r_k^T A d_k}{d_k^T A d_k}.$$

Uusi piste $x_{k+1} = x_k + \alpha_{k+1} d_{k+1}$ löydetään minimoimalla funktio

$$g(\alpha) = f(x_k + \alpha d_{k+1}).$$

Minimi löytyy derivaatan nollakohdasta

$$f'(\alpha_{k+1}) = \alpha_{k+1} d_{k+1}^T A d_{k+1} - r_k^T d_{k+1} = 0$$

. Näin ollen

$$\alpha_{k+1} = \frac{r_k^T d_{k+1}}{d_{k+1}^T A d_{k+1}}.$$

Seuraavien lemmojen avulla iteraatiossa laskettavat kertoimet voidaan esittää hieman toisessa muodossa.

Lemma 2.4.1 *Jäännösvektori pisteessä x_k on kohtisuorassa vektoria d_k vasten, ts. $d_k^T r_k = 0$.*

Tod.: Koska funktion $\phi(\alpha) = f(x_k + \alpha d_{k+1})$ minimi löytyy arvolla α_{k+1} derivaatan nollakohdassa, niin

$$0 = \nabla f(x_k + \alpha_{k+1} d_{k+1})^T d_{k+1} = r_{k+1}^T d_{k+1},$$

sillä funktion f(x) gradientti pisteessä x_{k+1} on jäännösvektori r_{k+1}. □

Lemma 2.4.2 *Kaikilla k: $r_k^T r_{k-1} = 0$.*

Tod.: Konjugoitujen suuntien konstruktion perusteella

$$d_k = r_{k-1} + \beta_k d_{k-1}$$

ja jäännösvektorin $r_k = b - Ax_k$ määritelmän nojalla

$$r_k = r_{k-1} - \alpha_k A d_k.$$

Näin ollen käyttäen edellisen lemman saadaan

$$0 = r_k^T d_k = r_k^T (r_{k-1} + \beta_k d_{k-1}) = r_k^T r_{k-1}.$$

□ Konjugoidun suunnan ja gradientin ortogonaalisuuden nojalla

$$\alpha_{k+1} = \frac{r_k^T d_{k+1}}{d_{k+1}^T A d_{k+1}} = \frac{r_k^T (r_k + \beta_{k+1} d_k}{d_{k+1}^T A d_{k+1}} = \frac{\|r_k\|^2}{d_{k+1}^T A d_{k+1}}.$$

Vastaavasti koska

$$\|r_{k+1}\|^2 = r_{k+1}^T r_k - \alpha_{k+1} r_{k+1}^T A d_{k+1} = -\frac{\|r_k\|^2}{d_{k+1}^T A d_{k+1}} r_{k+1}^T A d_{k+1},$$

niin

$$\beta_{k+2} = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Nyt voidaan konjugaattigradienttimenetelmän lopullinen algoritmi:

Algoritmi 2.4.1 *Initialisointi:* $x_0 = 0,\ r_0 = d_1 = b - Ax_0$*;*

Kaikille k=0,...,n-1

$$\begin{aligned}
\beta_{k+1} &= \frac{\|r_k\|^2}{\|r_{k-1}\|^2},\ \beta_1 = 0 \\
d_{k+1} &= r_k + \beta_{k+1} d_k,\ d_1 = r_0 \\
\alpha_{k+1} &= \frac{\|r_k\|^2}{d_{k+1}^T A d_{k+1}} \\
x_{k+1} &= x_k + \alpha_{k+1} d_{k+1} \\
r_{k+1} &= r_k - \alpha_{k+1} A d_{k+1}
\end{aligned}$$

2.5 Iteratiiviset menetelmä

2.5.1 Yleinen iteraatiomenetelmä

Kaikki yhtälöryhmien iteratiiviset ratkaisumenetelmät voidaan esittää muodossa:

- Anna alkuarvaus: $\underline{x_0} \in \mathbb{R}^n$;
- Jos $\underline{x_k}$ on määrätty, niin

$$\underline{x}_{k+1} = B\underline{x}_k + \underline{c},$$

missä B on menetelmän iteraatiomatriisi ja $\underline{c}$ kiinteä vakiovektori, joka riippuu alkuperäisen yhtälöryhmän oikeanpuolen vektorista.

Iteratiivisen menetelmän suppenemista varten tarvitaan muutamia matriisialgebran aputuloksia.

Matriisin A luonnollisella normilla tarkoitetaan sitä normia, joka on määritelty jonkin vektorinormin avulla:

$$\|A\| = \sup_{\underline{x} \neq 0} \frac{\|A\underline{x}\|}{\|\underline{x}\|}.$$

Kaikki matriisin p-normit ovat ns. luonnollisia normeja. Matriisin normin avulla voidaan arvioida ominaisarvojen suuruutta.

Lemma 2.5.1 *Jokaiselle matriisin A luonnolliselle normille matriisin spektraalisäde*

$$\rho(A) \leq \|A\|.$$

Tod.: Olkoon x_s matriisin A ominaisarvoa λ_s vastaava ominaisvektori. Tällöin

$$\|A\| = \sup \frac{\|Ax\|}{\|x\|} \geq \frac{\|Ax_s\|}{\|x_s\|} = \frac{\|\lambda_s x_s\|}{\|x_s\|} = |\lambda_s|.$$

Näin ollen kaikille ominaisarvoille

$$|\lambda| \leq \|A\|,$$

ja siten spektraalisäde $\rho(A) \leq \|A\|$. □

Lause 2.5.1 *Jokaiselle $\epsilon > 0$ on olemassa luonnollinen normi siten, että*

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon.$$

Tod.: Edellisen lemman nojalla tarvitsee osoittaa vain jälkimmäinen epäyhtälö. Jokaiselle matriisille on voimassa Jordanin normaalimuoto (kts. Matriisialgebra), ts. on olemassa säännöllinen matriisi P siten, että

$$PAP^{-1} = \Lambda + U,$$

missä Λ on diagonaalimatriisi alkioina A:n omianisarvot, ja U on yläkolmiomatriisi, jonka diagonaalialkiot ovat nollia.

Määritellään jokaiselle $\delta > 0$ diagonaalimatriisi

$$D = \mathrm{diag}(1, \delta^{-1}, \delta^{-2}, \ldots, \delta^{1-n}).$$

Kerrotaan $\Lambda + U$ matriisilla D molemmin puolin. Näin saadaan matriisi

$$C = D(\Lambda + U)D^{-1} = \Lambda + E,$$

missä matriisi E on yläkolmiomatriisi ja sen alkiot ovat

$$e_{ij} = \begin{cases} 0, \ i \geq j \\ u_{ij}\delta^{j-i}, \ j > i. \end{cases}$$

Matriisialkiot e_{ij} voidaan tehdä kuinka pieneksi tahansa δ:n valinnalla.

Näin ollen matriisille A on voimassa esitys

$$A = P^{-1}D^{-1}ADP.$$

Määritellään vektorinormi

$$\|x\| = \sqrt{x^T P^T D^T DPx},$$

ja sitä vastaava luonnollinen matriisinormi. Nyt vektorin Ay normi tämän uuden vektorinormin suhteen on

$$\begin{aligned} \|Ay\|^2 &= y^T P^T D^T C^T (D^{-1^T} P^{-1^T} P^T D^T)(DPP^{-1}D^{-1})CDPy \\ &= y^T P^T D^T C^T CDPy = z^T C^T Cz, \end{aligned}$$

missä $z = DPy$.

Matriisi $C^T C$ on olennaisesti diagonaalimatriisi. Nimittäin

$$C^T C = (\Lambda + E)^T(\Lambda + E) = \Lambda^T\Lambda + M(\delta),$$

missä matriisi $M(\delta)$ lähestyy nollamatriisia tasaisesti, kun $\delta \to 0$.

Näin ollen

$$\begin{aligned} z^T C^T Cz &= z^T\Lambda^T\Lambda z + z^T M(\delta)z \\ &\leq (\max|\lambda|^2 + C\delta)z^T z \\ &= (\rho(A)^2 + C\delta)z^T z. \end{aligned}$$

Nyt vektorin y normi on $\|y\|^2 = z^T z$. Joten edellisen normiepäyhtälön nojalla

$$\|Ay\|^2 \leq [\rho(A)^2 + C\delta]\|y\|^2,$$

mistä väite seuraa. □

Lause 2.5.2 *Seuraavat väittämät ovat yhtäpitäviä:*

1. $\lim_{k\to\infty} B^k = 0$;
2. $\lim_{k\to\infty} B^k v = 0,\ \forall\ v \in \mathbb{R}^n$;
3. *spektraalisäde* $\rho(B) < 1$;
4. *Ainakin yhdelle matriisinormille* $\|B\| < 1$.

Tod.: (1) $\Longrightarrow$ (2) : Väite seuraa epäyhtälöstä

$$\|B^k v\| \leq \|B^k\|\|v\|.$$

(2) $\Longrightarrow$ (3) : Jos $\rho(B) \geq 1$, niin on vektori u ja luku $\lambda \geq 1$ siten, että

$$Bu = \lambda u.$$

Näin ollen $B^k u = \lambda^k u$ ja siten reaalilukujono

$$\|B^k u\| = |\lambda|^k \|u\|$$

ei suppene kohti nollaa vastoin oletusta: $\|B^k v\| \to 0$ kaikille $v \in \mathbb{R}^n$.
(3) $\Longrightarrow$ (4) : Väite seuraa suoraan edellisestä lauseesta.
(4) $\Longrightarrow$ (1) : Tämä seuraa epäyhtälöstä:

$$\|B^k\| \leq \|B\|^k.$$

□ Oletetaan, että yhtälöllä

$$\underline{x} = B\underline{x} + \underline{c}$$

on yksikäsitteinen ratkaisu. Iteraatiot suppenevat mikäli

$$\lim_{k\to\infty} \underline{x}_k = \underline{x}$$

kaikilla alkuarvauksilla $\underline{x}_0$. Iteratiivisaten menetelmien suppenemiselle on olennaista seuraava lause:

Lause 2.5.3 *Seuraavat väittämät ovat yhtäpitäviä:*

1. *Iteratiivinen menetelmä on suppeneva;*
2. $\rho(B) < 1$;

3. Ainakin yhdelle matriisinormille $\|B\| < 1$.

Tod.: Määritellään virhevektori $\underline{e}_k = \underline{x}_k - \underline{x}$. Koska kaikille $k \in \mathbb{N}$

$$\underline{e}_k = \underline{x}_k - \underline{x} = B(\underline{x}_{k-1} - \underline{x}) = B\underline{e}_{k-1},$$

niin

$$\underline{e}_k = B^k \underline{e}_0.$$

Näin ollen lauseen väittämä on tosi edellisen lauseen nojalla. □ Iteratiivisen menetelmän virheelle on voimassa seuraava virhe-arvio:

Lause 2.5.4 *Jos iteratiivinen menetelmä suppenee, niin jonkin vektorinormin suhteen on voimassa a posteriori-arvio:*

$$\|\underline{x} - \underline{x}_k\| \leq \frac{\|B\|}{1 - \|B\|} \|\underline{x}_k - \underline{x}_{k-1}\|,$$

missä matriisinormi on vektorinormin kanssa yhteensopiva ja $\|B\| < 1$.

Tod.: Virheelle pätee edellisen lauseen todistuksen nojalla

$$\underline{x}_k - \underline{x} = B(\underline{x}_{k-1} - \underline{x}).$$

Lisäämällä ja vähentämällä yhtälön oikealla puolella vektori $B\underline{x}_k$ virhe voidaan esittää muodossa

$$\underline{x}_k - \underline{x} = B(\underline{x}_{k-1} - \underline{x}_k) + B(\underline{x}_k - \underline{x}).$$

Näin ollen on voimassa yhtälö

$$(I_n - B)(\underline{x}_k - \underline{x}) = B(\underline{x}_{k-1} - \underline{x}_k).$$

Väite seuraa, jos osoitetaan, että matriisilla $I_n - B$ on käänteismatriisi ja sen normi

$$\|(I_n - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Matriisialgebran perusteella matriisilla on käänteismatriisi, jos sen ydin $N(B) = \{0\}$. Oletetaan, että $\underline{z} \neq 0$ on matriisin $I_n - B$ ytimen alkio. Silloin on voimassa

$$\|\underline{z}\| = \|B\underline{z}\| \leq \|B\|\|\underline{z}\| < \|\underline{z}\|,$$

mikä on mahdottomuus. Näin ollen matriisilla $I_n - B$ on käänteismatriisi, joka voidaan esittää muodossa

$$(I_n - B)^{-1} = I_n + B(I_n - B)^{-1}.$$

Näin ollen matriisinormin ominaisuuksien nojalla

$$\|(I_n - B)^{-1}\| \leq 1 + \|B\|\|(I_n - B)^{-1}\|,$$

josta ratkaisemalla saadaan käänteismatriisin normille arvio

$$\|(I_n - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Eli lauseen väittämä on tosi. □

2.5.2 Jacobin ja Gauss-Seidelin iteraatiot

Jacobin menetelmä perustuu matriisin A summahajotelmaan $A = L + D + U$, missä L on matriisin alakolmiomatriisi, U A:n yläkolmiomatriisi ja D A:n diagonaalimatriisi. Tällöin yhtälöryhmä $A\underline{x} = \underline{b}$ voidaan esittää muodossa

$$D\underline{x} = \underline{b} - (L + U)\underline{x}.$$

Jos matriisin A kaikki diagonaalialkiot ovat nollasta eroavia, niin yo. yhtälö voidaan esittää kiintopisteyhtälönä

$$\underline{x} = -D^{-1}(L + U)\underline{x} + D^{-1}\underline{b}.$$

Jacobin menetelmässä kiintopistematriisi on

$$B_J = -D^{-1}(L + U)$$

ja vakiovektori $\underline{c} = D^{-1}\underline{b}$.

Gauss-Seidelin menetelmässä yhtälöryhmä $A\underline{x} = \underline{b}$ esitetään muodossa

$$\underline{x} = -(D + L)^{-1}U\underline{x} + (D + L)^{-1}\underline{b}.$$

Matriisin $D + L$ käänteismatriisi on taas olemassa ainakin silloin kun A:n diagonaalialkiot ovat nollasta eroavia. Siten Gauss-Seidelin menetelmän iteraatiomatriisi ja vakiovektori ovat

$$B_G = -(D + L)^{-1}U,\ \underline{c} = (D + L)^{-1}\underline{b}.$$

Menetelmien suppenemisesta Jacobin ja Gauss-Seidelin menetelmät suppenevat, mikäli matriisi A on aidosti diagonaalidominantti, ts. joko on voimassa

$$|a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij},$$

tai

$$|a_{ii} > \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ji}.$$

Tällöin iteraatiomatriisille joko $\|B\|_\infty < 1$ tai $\|B\|_1 < 1$ kummassakin tapauksessa.

On tilanteita, joissa Jacobin menetelmä suppenee; mutta Gauss-Seidelin menetelmä ei. Tällöin matriisi A ei luonnollisesti ole aidosti diagonaalidominantti. Mutta jos molemmat menetelmät suppenevat, niin yleensä Gauss-Seidelin konvergenssinopeus on huomattavasti nopeampi.

Luku 3

Epälineaariset yhtälöt ja yhtälöryhmät

3.1 Funktion nollakohdat

Olkoon $f : [a,\ b] \to \mathbf{R}$ jatkuva funktio. Väli $[a,\ b]$ on valittu siten, että $f(a)f(b) < 0$. Tällöin jatkuvuuden nojalla funktiolla on ainakin yksi nolla kohta välillä $[a,\ b]$, ts. on olemassa $x \in [a,\ b]$ siten, että

$$f(x) = 0.$$

Puolitusmenetelmä

on yksinkertaisin nollakohdan määräämismenetelmistä ja se toimii aina.

Algoritmi 3.1.1 *Olkoon funktio $f(x)$ jatkuva välillä $[a,\ b]$ siten, että*

$$f(a)f(b) < 0.$$

Tällöin

1. **Laske** $x_{mid} = \frac{1}{2}(a + b)$.
2. **Jos**

$$f(x_{mid})f(a) < 0,$$

niin

$$\begin{aligned} a &= a \\ b &= x_{mid}, \end{aligned}$$

muutoin

$$\begin{aligned} a &= x_{mid} \\ b &= b. \end{aligned}$$

3. **Jos** $|b-a| < \epsilon$, **niin** *STOP; muutoin palaa kohtaan (2).*

Koska puolitusmenetelmässä väli puolintuu joka askeleella, niin haluttuun tarkkuuteen vaadittavien askelien lukumäärä n:

$$n > log_2(\frac{b-a}{\epsilon}),$$

missä ϵ on haluttu tarkkuus. Puolitusmenetelmä konvergoi hitaasti, koska menetelmässä ei käytetä varsinaisesti informaatiota funktion f ominaisuuksista.

Kiintopisteiteraatiot

Tavallisesti nollakohdan määrääminen voidaan palauttaa iteratiiviseksi menetelmäksi, jossa etsitään jonkin sopivan funktion $\Phi(x)$ kiintopiste: $x = \Phi(x)$. Oletetaan esimerkiksi, että $g(x) \neq 0$ kaikilla $x \in [a,\ b]$. Tällöin $x \in [a,\ b]$ on funktion $f(\cdot)$ nollakohta täsmälleen silloin, kun se on funktion

$$\Phi(x) = x - g(x)f(x)$$

kiintopiste.

Kiintopistettä voidaan hakea ns. **kiintopisteiteraatiolla**:

Algoritmi 3.1.2 *Määritellään lukujono* (x_n) *seuraavasti:*

1. $x^{(0)} \in [a,\ b]$

2. **Kun** $x^{(k)}$ *on annettu,* **niin** $x^{(k+1)} = \Phi(x^{(k)})$

3. STOP, jos $|x^{(k+1)} - x^{(k)}| < \ \epsilon$.

Seuraavien ehtojen vallitessa menetelmä suppenee:

Lause 3.1.1 *Olkoon funktio* $\Phi(x)$ *jatkuva ja*

- *oletetaan, että* Φ *toteuttaa* **Lipschitz-ehdon:**

$$|\Phi(x) - \Phi(y)| \leq L\,|x - y|,\ 0 < L < 1,$$

suljetussa ja rajoitetussa joukossa $[a,\ b]$.

- *Lisäksi oletetaan, että*

$$\Phi([a,\ b]) \subset [a,\ b].$$

Tällöin on olemassa yksikäsitteisesti määrätty kiintopiste $x \in\ [a,\ b]$, *ja kiintopisteiteraatiot suppenevat kohti kiintopistettä jokaisella alkuarvauksella* $x_0 \in\ [a,\ b]$.

Kiintopistelauseen todistuksessa tarvitaan seuraavaa reaalilukujonoja koskevaa tulosta:

Lause 3.1.2 *Jokaisella Cauchy-jonolla* $\{x_n|\ n = 0, 1, 2, 3, \dots\}$ *on raja-arvo.*

Kiintopistelauseen todistus: Olkoon $\{x_n|\ n = 0, 1, 2, 3, \dots\}$ funktion $\phi(x)$ kiintopisteiteraatiot.

1. askel: Jokaiselle $n, j \in \mathbf{N}:\ |x_{n+j+1} - x_{n+j}| \leq\ L^{n+j}|x_1 - x_0|$:

Oletuksen (2) nojalla

$$|x_{n+j+1} - x_{n+j}| = |\phi(x_{n+j}) - \phi(x_{n+j-1})| \leq L|x_{n+j} - x_{n+j-1}|.$$

Toistamalla arvio n+j kertaa saadaan väite.

2. askel: Kirjoitetaan erotus $x_{n+m} - x_n$ teleskooppisummana

$$x_{n+m} - x_n = \sum_{j=0}^{m-1} [x_{n+j+1} - x_{n+j}].$$

Kolmioepäyhtälön nojalla pätee

$$|x_{n+m} - x_n| \leq \sum_{j=0}^{m-1} |x_{n+j+1} - x_{n+j}|.$$

Käyttäen edellisen kohdan arvioita saadaan

$$\sum_{j=0}^{m-1} |x_{n+j+1} - x_{n+j}| \leq \sum_{j=0}^{m-1} L^{n+j}|x_1 - x_0|.$$

Geometrisen sarjan summa on

$$\sum_{j=0}^{m-1} L^{n+j} = L^n \frac{L^m - 1}{L - 1}.$$

Näin ollen

$$|x_{n+m} - x_n| \leq L^n \frac{L^m - 1}{L - 1}|x_1 - x_0|.$$

Epäyhtälön oikean puolen lauseke suppenee kohti nollaa, sillä

$$\lim_{n\to\infty} L^n = 0.$$

Siten kiintopisteiteraatiojono on Cauchy-jono, ja siksi jonolla on raja-arvo.

Todetaan vielä, että raja-arvo on kiintopiste. Funktion $\phi(x)$ jatkuvuuden ja kiintopisteiteraation nojalla

$$x = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} \phi(x_n) = \phi(\lim_{n\to\infty} x_n) = \phi(x).$$

Yksikäsitteisyys: Olkoon x ja y kaksi kiintopistettä. Tällöin

$$|x - y| = |\phi(x) - \phi(y)| \leq L|x - y|.$$

Induktiivisesti jatkamalla saadaan, että kaikille n

$$|x - y| \leq L^n|x - y|.$$

Koska $\lim\ L^n = 0$, niin $x = y$.

Kiintopisteiteraatiolle voidaan osoittaa **virhe-arviot**

Lause 3.1.3 *Olkoon (x_n) suppeneva kiintopisteiteraatiojono. Tällöin jonon alkioille on voimassa seuraavat virhe-arviot:*

- **A priori-arvio**

$$|x^{(k)} - x| \leq \frac{L^k}{1-L}|x^{(1)} - x^{(0)}|$$

- **A posteriori-arvio**

$$|x^{(k)} - x| \leq \frac{L}{1-L}|x^{(k)} - x^{(k-1)}|$$

Vaihtoehtoisesti Lipschitz-ehdon sijaan voidaan vaatia, että

$$|\Phi'(x)| \leq L < 1,\ x \in [a,\ b].$$

Tällöin kiintopisteiteraatiot myös suppenevat ja edellä esitetyt virhearviot ovat voimassa.

3.1.1 Aitkenin δ^2-prosessi

Olkoon $\{x_n|\ n \in \mathbf{N}\}$ funktion $\phi(x)$ kiintopisteiteraatiot, ja x kiintopiste. Tällöin

$$\lim_{n\to\infty} \frac{x_{n+2} - x}{x_{n+1} - x} = \lim_{n\to\infty} \frac{\phi(x_{n+1}) - \phi(x)}{x_{n+1} - x} = \phi'(x).$$

Riittävän suurilla n:n arvoilla

$$\begin{aligned} \frac{x_{n+2} - x}{x_{n+1} - x} &\approx \phi'(x) \\ \frac{x_{n+1} - x}{x_n - x} &\approx \phi'(x), \end{aligned}$$

ja siten on voimassa likipitäen

$$\frac{x_{n+2} - x}{x_{n+1} - x} \approx \frac{x_{n+1} - x}{x_n - x}.$$

Käytetään tätä yhtälöä korjatun likiarvon laskemiseen. Ratkaistaan yhtälöstä

$$\frac{x_{n+2} - x^*}{x_{n+1} - x^*} = \frac{x_{n+1} - x^*}{x_n - x^*}$$

uusi likiarvo

$$x^* = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

Tämä proseduuri on *Aitkenin δ^2-prosessin* perusta:

Algoritmi 3.1.3 *1. x_0 alkuarvaus;*

2. Lasketaan kiintopisteiteraatiolla lukujono $(x_n)_{n\geq 0}$;

3. Korjataan Aitkenin δ^2-prosessilla uudet likiarvot

$$z_n = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

Aitkenin δ^2-prosessi suppenee nopeammin kuin kiintopisteiteraatio.

Lause 3.1.4 *Oletetaan, että jono $(x_n)_{n\geq 0}$ suppenee lineaarisesti, ts. virheelle $e_n = x_n - x$ on voimassa:*

$$e_{n+1} \approx \ qe_n, \ q < 1.$$

Tällöin Aitkenin δ^2-prosessilla konstruoidulle jonolle on voimassa

$$\lim_{n\to\infty} \frac{z_n - x}{x_n - x} = 0.$$

3.1.2 Konvergenssiaste

Määritellään aluksi kiintopisteiteraation konvergenssiaste:

Määritelmä 3.1.1 *Iteraatiojonon (x_n) konvergenssiaste on vähintäin p, jos*

$$\limsup_{k\to\infty} \frac{|x_{n+1} - x|}{|x_n - x|^p} = K,$$

missä

$$\begin{aligned} 0 &< K < \infty, \ p > 1 \\ K &< 1, \ p = 1. \end{aligned}$$

Olkoon sitten jono (x_n) funktion $\phi(x)$ iteraatiojono ja virhe $e_n = x_n - x$. Tällöin

$$x_{n+1} = x + e_{n+1} = \phi(x_n) = \phi(x + e_n).$$

Mikäli kiintopistefunktio on riittävän säännöllinen, niin Taylorin kehitelmän nojalla

$$\begin{aligned} x + e_{n+1} &= \phi(x) + \phi'(x)e_n + \frac{1}{2}\phi^{(2)}(x)e_n^2 + \cdots + \frac{1}{(k-1)!}\phi^{(k-1)}(x)e_n^{k-1} \\ &\quad + \frac{1}{k!}\phi^{(k)}(\zeta)e_n^k. \end{aligned}$$

Koska $x = \phi(x)$, niin edellisestä saadaan virheelle e_{n+1} asymptoottinen kehitelmä

$$e_{n+1} = \phi'(x)e_n + \frac{1}{2}\phi^{(2)}(x)e_n^2 + \cdots + \frac{1}{(k-1)!}\phi^{(k-1)}(x)e_n^{k-1} + \frac{1}{k!}\phi^{(k)}(\zeta)e_n^k.$$

Tämän kehitelmän nojalla seuraava lause on ilmeinen:

Lause 3.1.5 *Kiintopisteiteraation konvergenssiaste on vähintäin k, jos kiintopisteessä on voimassa*

$$\phi^{(j)}(x) = 0,\ j = 1, \ldots, k-1$$

ja

$$\phi^{(k)}(x) \neq 0.$$

3.1.3 Newtonin menetelmä

Oletetaan, että funktio $f(x)$ on ainakin kaksi kertaa jatkuvasti differentioituva. Olkoon sitten laskettu Newtonin menetelmällä nollakohdan likiarvot $x_k,\ k = 0, 1, \ldots, n$, missä tietysti x_0 on alkuarvaus. Funktion $f(x)$ Taylorin kehitelmällä pisteen x_n ympäristössä on

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2}f''(\zeta)(x - x_n)^2.$$

Jos $|x - x_n|^2 << 1$, niin funktiota voidaan approksimoida ensimmäisen asteen Taylorin polynomilla, jonka nollakohta on uusi "tarkempi"likiarvo $f(x)$:n nollakohdalle:

$$f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0 \rightarrow x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Algoritmi 3.1.4

1. *Valitse alkuarvaus* $x_0 \in [a,\ b]$;
2. $n = 0, 1, 2, \cdots : x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$;
3. $|x_n - x_{n+1}| < \epsilon \rightarrow$ *Lopeta;*

Newtonin menetelmä on siten kiintopisteiteraatio, jonka iteraatiofunktio

$$F(\xi) = \xi - \frac{f(\xi)}{f'(\xi)}.$$

Seuraava lause takaa, että Newtonin menetelmä suppenee kvadraattisesti useimmissa tapauksissa:

Lause 3.1.6 *Olkoon $f : [a, b] \to \mathbf{R}$ kolme kertaa jatkuvasti differentioituva välillä, ja $s \in [a, b]$ funktion nollakohta siten, että $f'(s) \neq 0$. Silloin on olemassa väli $I_\delta = [s - \delta, s + \delta],\ \delta > 0$, jossa Newtonin menetelmän iteraatiofunktio $F : I_\delta \to I_\delta$ on kontraktio, ja siten Newtonin menetelmä suppenee jokaisella alkuarvauksella $x_0 \in I_\delta$. Lisäksi konvergenssiaste on ainakin kaksi.*

3.2 Yhtälöryhmät

3.2.1 Kiintopisteiteraatiot yhtälöryhmälle

Etsitään kuvauksen

$$\Phi(x) = \begin{bmatrix} \Phi_1(x_1, x_2, \cdots, x_n) \\ \vdots \\ \Phi_n(x_1, x_2, \ldots, x_n) \end{bmatrix}$$

kiintopistettä, ts

$$x = \Phi(x).$$

Algoritmi 3.2.1 *Määritellään kiintopisteiteraatiot seuraavasti:*

1. *Alkuarvaus $x^{(0)} \in \mathbf{R^n}$;*
2. *Kaikille $k \geq 0 :\ x^{(k+1)} = \Phi(x^{(k)}) \in \mathbf{R^n}$;*
3. *If $\|x^{(k+1)} - x^{(k)}\| < \epsilon$, then STOP.*

Suppenemisehto on sama kuin yhden muuttujan funktion kiintopistelauseessa:

Lause 3.2.1 *Oletetaan, että seuraavat ehdot ovat voimassa*

1. *Suljettu ja rajoitettu joukko $A \subset \mathbf{R^n}$ s.e.*

$$\Phi(A) \subset A;$$

2. *Lipschitz-ehto:*

$$\|\Phi(x) - \Phi(y)\| \leq L\ \|x - y\| < \|x - y\|$$

kaikilla $x,\ y \in A$

Tällöin joukossa A on olemassa yksikäsitteisesti määrätty kiintopiste $x \in A$.

Lipschitz-ehto on tosi, jos Φ:n funktionaalimatriisille l. derivaatalle

$$\Phi'(x) = \begin{bmatrix} \nabla\Phi_1(x) \\ \vdots \\ \nabla\Phi_n(x) \end{bmatrix}$$

on voimassa

$$\|\Phi'(x)\| \leq L < 1, \ x \in A,$$

jonkin matriisinormin suhteen (kts. liite A). Tämä on yhtäpitävä sen ehdon kanssa, että funktionaalimatriisin spektraalisäde

$$\rho(\Phi'(x)) < 1.$$

3.2.2 Newton-Raphson-menetelmä

Yhtälöryhmä:

$$F(x) = \begin{bmatrix} F_1(x_1, \ldots, x_n) \\ \vdots \\ F_n(x_1, \ldots, x_n) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Funktionaalimatriisi l. Jakobiaani pisteessä x on funktion F derivaatta:

$$F'(x) = \begin{bmatrix} \nabla F_1(x) \\ \vdots \\ \nabla F_i(x) \\ \vdots \\ \nabla F_n(x) \end{bmatrix}$$

missä

$$\nabla F_i = [\frac{\partial F_i}{\partial x_1}, \ldots' \frac{\partial F_i}{\partial x_n}].$$

Oletus 3.2.1 *Yhtälöryhmän ratkaisulle $\zeta \in \mathbf{R^n}$:*

$$\det(F'(\zeta)) \neq 0.$$

Algoritmi 3.2.2 *1. Alkuarvaus $x^{(0)} \in \mathbf{R^n}$;*

2. Ratkaise $\delta x \in \mathbf{R^n}$*:*

$$F'(x^{(k)})\delta x = -F(x^{(k)});$$

3.

$$x^{(k+1)} = x^{(k)} + \delta x;$$

4. **Lopetuskriteerio:** $\|\delta x\| < \epsilon$ *ja* $\|F(x^{(k+1)})\| < \rho$.

Konvergenssiaste:

Lause 3.2.2 *Olkoon funktion* $F(x)$ *koordinaattifunktiot kolmesti jatkuvasti differentioituvia suorakaiteessa*

$$A = \{x \in \mathbf{R}^n |\ a_i \leq x_i \leq b_i\},$$

joka sisältää F*:n nollakohdan, ja funktionaalimatriisi* $F'(x)$ *on säännöllinen matriisi nollakohdassa. Silloin Newtonin menetelmä suppenee kvadraattisesti kohti nollakohtaa, jos alkuarvaus on riittävän hyvä:*

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - \zeta\|}{\|x^{(k)} - \zeta\|^2} = \alpha < \infty.$$

Yksinkertaistettu Newtonin menetelmä

Newtonin menetelmässä joudutaan ratkaisemaan yhtälöryhmä jokaisella iteraatiokierroksella. Mikäli iteraatiojono $(x^{(k)};\ k = 0, 1, 2, \dots)$ suppenee ja funktio $F(x)$ on riittävän sileä, niin

$$\lim_{k\to\infty} F'(x^{(k)}) = F'(x)$$

ja siten riittävän suurilla k:n arvoilla

$$F'(x^{(m)}) \approx F'(x^{(k)}),\ m = k+1, k+2, \dots.$$

Näin ollen seuraavan algoritmi käyttö on perusteltua yhtälöryhmän numeeriseen ratkaisemiseen:

Algoritmi 3.2.3 *1. Alkuarvaus* $x^{(0)} \in \mathbf{R^n}$*;*

2. Ratkaise $\delta x \in \mathbf{R^n}$*:*

$$F'(x^{(0)})\delta x = -F(x^{(k)});$$

3.

$$x^{(k+1)} = x^{(k)} + \delta x;$$

4. **Lopetuskriteerio:** $\|\delta x\| < \epsilon$ *ja* $\|F(x^{(k+1)})\| < \rho$.

3.2.3 Kvasi-Newton-menetelmä

Kvasi-Newton-menetelmä on yksinkertaistetun ja varsinaisen Newtonin menetelmän välissä. Se suppenee nopeammin kuin yksinkertaistettu menetelmä; mutta on yksinkertaisempi käyttää kuin Newtonin menetelmä. Menetelmässä approksimoidaan derivaatan käänteismatriisia. Olkoon $x^{(0)}$ kvasi-Newton menetelmän alkuarvaus ja $A_0 = F'(x^{(0)})$ derivaatta kyseisessä pisteessä. Lasketaan uusi likiarvo $x^{(1)}$ ja vastaava funktion arvo $F(x^{(1)})$. Määritellään derivaataan approksimaatio A_1 vaatimalla, että

$$A_1(x^{(1)} - x^{(0)}) = F(x^{(1)}) - F(x^{(0)}).$$

Tämä saavutetaan, kun

$$A_1 = A_0 + \frac{(\Delta F - A_0\Delta x)(\Delta x)^T}{(\Delta x)^T(\Delta x)},$$

missä

$$\Delta x = x^{(1)} - x^{(0)},\ \Delta F = F(x^{(1)}) - F(x^{(0)}).$$

Toistetaan iteraatio jokaisella askeleella. Voidaan osoittaa, että

$$\lim_{k\to\infty} A_k = F'(x),$$

missä x on yhtälöryhmän ratkaisu.

Algoritmi 3.2.4 *1. Alkuarvaus $x^{(0)} \in \mathbf{R^n}$;*

2. *Olkoon $x^{(j)}$, $F(x^{(j)})$, kun $j = 0, 1, \ldots, k$ ja A_j määrätty, kun $j = 0, \ldots, k-1$;*

3. *Päivitä matriisi A_k seuraavasti:*

$$A_k = A_{k-1} + \frac{(\Delta F - A_{k-1}\Delta x)(\Delta x)^T}{(\Delta x)^T(\Delta x)},$$

missä

$$\Delta x = x^{(k)} - x^{(k-1)},\ \Delta F = F(x^{(k)}) - F(x^{(k-1)}).$$

4. *Laske*

$$x^{(k+1)} = x^{(k)} + A_k^{-1}F(x^{(k)})\ ja\ F(x^{(k+1)});$$

5. **Lopetuskriteerio:** $\|\Delta x\| < \epsilon$ *ja* $\|F(x^{(k+1)})\| < \rho$.

Kvasi-Newton menetelmää käytetään yhtenä monista menetelmistä adaptiivisten suodattimien suunnittelussa (IIR-suodattimet).

Luku 4

Funktion approksimointi ja interpolointi

4.1 Interpolointi

4.1.1 Taylorin polynomi

Funktio $f(x)$ on $(n+1)$-kertaa jatkuvasti differentioituva funktio, ts. $f^{(n+1)}(x)$ on jatkuva funktio välillä $I = [a,\ b]$. Funktion Taylorin polynomi $p(x)$ on sellainen, että

$$p^{(j)}(x_0) = f^{(j)}(x_0),\ 0 \leq j \leq n$$

pisteessä $x_0 \in I$. Ilmeisesti $p(x)$ on n-asteinen polynomi ja sen esitysmuoto on

$$p(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n.$$

Taylorin polynomin määrittelevästä derivaattaehdosta voidaan määrittää vakiot a_j:

$$a_j = \frac{f^{(j)}(x_0)}{j!}.$$

Näin ollen funktion $f(x)$ Taylorin polynomi pisteen x_0 ympäristössä on

$$p(x) = \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x - x_0)^j.$$

Lause 4.1.1 *Olkoon $f^{(n+1)}(x)$ jatkuva välillä $[a,\ b]$ ja $P_n(x)$ funktion Taylorin polynomi pisteessä $x_0 \in [a,\ b]$. Tällöin Taylorin polynomin virheelle on*

voimassa lauseke

$$R_n(x) = f(x) - P_n(x) = \frac{(x - x_0)^{(n+1)}}{(n+1)!} f^{(n+1)}(\xi_x),$$

missä ξ_x *on pisteiden* x *ja* x_0 *välillä oleva piste.*

Lause 4.1.2 *Oletetaan, että funktio* $f(x, y)$ *on* $(n+1)$*-kertaa jatkuvasti differentioituva pisteen* (x_0, y_0) *ympäristössä. Tällöin riittävän pienille* h *ja* k *on voimassa Taylorin kehitelmä*

$$\begin{aligned} f(x_0 + h, y_0 + k) &= f(x_0, y_0) + [(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y})f](x_0, y_0) \\ &+ \cdots + \\ &+ \frac{1}{n!}[(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y})^n f](x_0, y_0) \\ &+ \frac{1}{(n+1)!}[(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y})^{n+1} f](x_0 + \theta h, y_0 + \theta k), \end{aligned}$$

missä $0 < \theta < 1$.

4.1.2 Polynomi-interpolaatio

Polynomi-interpolaatio-ongelma on seuraava:

Annetuille koordinaattitason pisteille $(x_0, f_0),\ (x_1, f_1), \ldots, (x_n, f_n)$, missä luvut x_i ovat erisuuria, määrää polynomi P, joka toteuttaa seuraavat ehdot

1. $\deg(P) \leq n$,
2. $P(x_i) = f_i,\ i = 0, \ldots, n.$

Oletetaan, että funktion $f(x)$ (yhden muuttujan funktio) arvot tunnetaan pisteissä $x_0, x_1, \ldots, x_n$. Tavoitteena on konstruoida polynomi $P_n(x)$ siten, että

$$P_n(x_j) = f(x_j),\ j = 0, \ldots, n$$

Lause 4.1.3 *Interpolaatiotehtävällä on yksikäsitteisesti määrätty n-asteinen interpolaatiopolynomi*

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x),$$

missä $L_i(x)$ on Lagrange'n kantapolynomi:

$$L_i(x) = \prod_{\substack{j=0 \\ i \neq j}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Tod.:
Lauseen väittämä on ilmeinen seuraus tosiasiasta:

$$L_i(x_k) = \begin{cases} 1, & i = k \\ 0, & i \neq k. \end{cases}$$

Nimittäin tällöin summassa

$$P(x_j) = \sum_{i=0}^{n} f_i L_i(x_j)$$

kaikki muut termit ovat nollia paitsi, kun $i = j$. Näin ollen

$$P(x_j) = f_j.$$

Yksikäsitteisyys seuraa algebran peruslauseen nojalla, jonka mukaan n-asteinen polynomi on nollapolynomi, jos sillä on n+1 nollakohtaa.

Olettaen, että $P(x)$ ja $Q(x)$ on kaksi interpolaatiopolynomia, joiden asteluku on korkeintaan n, niin polynomilla $R(x) = P(x) - Q(x)$ on n+1 nollakohtaa ja asteluku on korkeintaan n. Edellä esitetyn algebran peruslauseen nojalla $R(x)$ on nollapolynomi. □

Huomioita: Lagrangen interpolaatiopolynomi on vaikea evaluoida (ts. laskea sen arvoja), sillä muotoa

$$(x_0 - x_i) \cdots (x_{i-1} - x_i)(x_{i+1} - x_i) \cdots (x_n - x_i)$$

olevien tulojen laskut johtavat helposti yli- tai alivuotoon. Siksi seuraavassa kappaleessa esitettävä interpolaatiopolynomin esitystapa on suositeltavampi.

4.1.3 Newtonin interpolaatio:

Johdetaan seuraavaksi Newtonin esitys interpolaatiopolynomille. Sitä varten tarvitaan tekninen laskenta apuväline l. **jaetut erotukset**.

Määritelmä 4.1.1 *Olkoon $x = (x_1, \ldots, x_n)$ ja $f = (f_1, \ldots, f_n)$ kaksi vektoria. Jaettu erotus $f[x_i, \ldots, x_{i+k}]$ määritellään rekursiivisesti:*

$$\begin{aligned} f[x_i] &= f_i \\ f[x_i, \ldots, x_j] &= \frac{f[x_{i+1}, \ldots, x_j] - f[x_i, \ldots, x_{j-1}]}{x_j - x_i}, \ j = i+1, \ldots, i+k. \end{aligned}$$

Newtonin erotustaulukko

$$\begin{array}{llllll} x_0 & f_0 & & & & \\ & & f[x_0, x_1] & & & \\ x_1 & f_1 & & f[x_0, x_1, x_2] & & \\ & & f[x_1, x_2] & & f[x_0, x_1, x_2, x_3] & \\ x_2 & f_2 & & f[x_1, x_2, x_3] & & f[x_0, x_1, x_2, x_3, x_4] \\ & & f[x_2, x_3] & & f[x_1, x_2, x_3, x_4] & \\ x_3 & f_3 & & f[x_2, x_3, x_4] & & \\ & & f[x_3, x_4] & & & \\ x_4 & f_4 & & & & \end{array}$$

Newtonin interpolaatiossa interpolaatiopolynomia haetaan muodossa

$$\begin{aligned} P_n(x) &= a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) \\ &+ \cdots + a_n(x - x_0) \cdots (x - x_{n-1}). \end{aligned}$$

Polynomin kertoimet ratkaistaan interpolaatioehdosta

$$P_n(x_j) = f_j.$$

Tällöin saadaan yhtälöryhmä

$$\begin{aligned} f_0 &= a_0 \\ f_1 &= a_0 + (x_1 - x_0)a_1 \\ f_2 &= a_0 + a_1(x_2 - x_1) + a_2(x_2 - x_0)(x_2 - x_1) \\ \vdots &= \vdots \\ f_n &= a_0 + \cdots + a_n(x_n - x_0) \cdots (x_n - x_{n-1}) \end{aligned}$$

Ratkaisemalla tämä alakolmiomuodossa oleva yhtälöryhmä, jolla ilmeisesti on yksikäsitteinen ratkaisu, niin saadaan Newtonin interpolaatiopolynomin kertoimet

Lause 4.1.4 *Newtonin interpolaatiopolynomin kertoimet ovat*

$$a_j = f[x_0, x_1, \cdots, x_j], \ j = 0, 1, 2, \cdots, n.$$

Tod.: Edellisestä yhtälöryhmästä helposti nähdään, että $a_0 = f_0$ ja $a_1 = \frac{f_1 - f_0}{x_1 - x_0} = f[x_0, x_1]$. Vastaavasti, kerroin a_2 ratkaistaan yhtälöstä

$$a_2(x_2 - x_0)(x_2 - x_1) = f_2 - f_0 - \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_0).$$

Kirjoitetaan edellinen yhtälömuodossa

$$\begin{aligned} a_2(x_2 - x_0)(x_2 - x_1) &= f_2 - f_0 - \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_1 + x_1 - x_0) \\ &= f_2 - f_0 - \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_1) - f_1 + f_0 \\ &= f_2 - f_1 - \frac{f_1 - f_0}{x_1 - x_0}(x_2 - x_1), \end{aligned}$$

josta jakamalla puolittain termillä $x_2 - x_0)(x_2 - x_1)$ saadaan lauseke

$$a_2 = \frac{\frac{f_2 - f_1}{x_2 - x_1} - \frac{f_1 - f_0}{x_1 - x_0}}{x_2 - x_0} = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = f[x_0, x_1, x_2].$$

Muut kertoimet voidaan osoittaa induktiolla. □

Newtonin esityksen kertoimien laskeminen suoritetaan seuraavalla algoritmilla:

```
for (j = 2;     j <= n; j++)
      for       (i = n; i >= j; i--)
                c[i] = (c[i] - c[i-1])/(x[i] - x[i-j])
```

Kertoimien laskemiseen tarvitaan n^2 kertolaskua ja $\frac{1}{2}n^2$ yhteenlaskua.

Newtonin interpolaatiopolynomin laskeminen Oletetaan, että Newtonin polynomin kertoimet c_n on määrätty. Se voidaan kirjoittaa teleskooppitulona:

$$\begin{aligned} P(t) = ((\cdots \quad &(\quad (c_n(t - x_{n-1}) \\ &+ \quad c_{n-1})(t - x_{n-2}) + c_{n-2})\cdots)(t - x_1) + c_1(t - x_0) + c_0. \end{aligned}$$

Sama pseudoalgoritmina:

$$\begin{aligned} p &= c[n]; \\ \text{for} &\quad (i = n-1;\ i >= 0;\ i--) \\ &\quad p = p * (t - x[i]) + c[i]; \end{aligned}$$

Algoritmissa suoritetaan 2n yhteenlaskua ja n kertolaskua.

4.1.4 Interpolaatiovirhe

Lause 4.1.5 *Olkoon $[a,\ b]$ väli sisältäen interpolaatiopisteet $\{x_0, \ldots, x_n\}$, ja funktio f(x) tällä välillä $(n+1)$-kertaa jatkuvasti differentioituva funktio. Silloin jokaisella $x \in [a,\ b]$ on olemassa $\xi_x \in [a,\ b]$ siten, että*

$$f(x) - P_n(x) = (x - x_0) \cdots (x - x_n) \frac{f^{(n+1)}(\xi_x)}{(n+1)!}$$

Tod.: Funktiolla

$$g(x) = f(x) - P_n(x) + \lambda(x - x_0) \cdots (x - x_n),\ \lambda \in \mathbf{R}.$$

on nollakohdat $x_0, \ldots, x_n$. Olkoon $\alpha \in [a,\ b] \setminus \{x_0, \ldots, x_n\}$ mielivaltainen. Valitaan

$$\lambda_\alpha = -\frac{f(\alpha) - P_n(\alpha)}{\prod_{i=0}^{n}(\alpha - x_i)}.$$

Tällöin funktiolla $g_\alpha(x)$ on välillä $[a,\ b]$ $n+2$ nollakohtaa. Voidaan olettaa, että nollakohdat ovat seuraavanlaisessa järjestyksessä:

$$x_0 < x_1 < \ldots < x_n < \alpha = x_{n+1}.$$

Rolle'n lauseen nojalla derivaattafunktioilla $g_\alpha^{(j)}(x)$ on $n+2-j$ nollakohtaa $x_i^{(j)},\ i = 0, 1, \ldots, n+1-j$ välillä $[a, b]$ siten, että

$$x_i^{(j-1)} < x_i^{(j)} < x_i^{(j-1)}.$$

Siten funktiolla $g_\alpha^{(n+1)}(x)$ on nollakohta $\xi_\alpha \in [a,\ b]$:

$$0 = g_\alpha^{(n+1)}(\xi_\alpha) = f^{(n+1)}(\xi_\alpha) - P_n^{(n+1)}(\xi_\alpha) + \lambda_\alpha (n+1)!.$$

Koska n-asteisen polynomin $n+1$-kertainen derivaatta häviää identtisesti, niin välttämättä

$$\lambda_\alpha = -\frac{f^{(n+1)}(\xi_\alpha)}{(n+1)!}.$$

Näin ollen on saatu pisteessä $\alpha \in [a,b] \setminus \{x_0, \ldots, x_n\}$:

$$0 = g_\alpha(\alpha) = f(\alpha) - P_n(\alpha) + \lambda_\alpha(\alpha - x_0)\cdots(\alpha - x_n).$$

Sijoittamalla tähän identiteettiin λ_α saadaan väittämä. Väite on triviaalisti tosi pisteissä $x_0, \ldots, x_n$. □

Huomiota konvergenssista Interpolaatiopisteitä ei kannata valita aina tasavälisesti eikä ainakaan kannata lisätä niiden lukumäärää määrättömästi kuten seuraava Runge'n esittämä esimerkki osoittaa.

Tarkastellaan funktiota

$$f(t) = \frac{1}{1+t^2}$$

välillä $[-5,\ 5]$. Interpoloidaan sitä tasavälisellä hilalla n-asteisella polynomilla. Jos interpolaatiopisteiden lukumäärä n kasvaa rajatta, niin voidaan (ei kovin) helposti osoittaa, että myös interpolaation maksimivirhe kasvaa rajatta. Nimittäin n-asteisen interpolaatiopolynomin virheelle on voimassa

$$\lim_{n\to\infty} \max_{x\in[-5,5]} |P_n(x) - f(x)| = \infty.$$

4.1.5 Tschebyscheffin interpolaatiopisteet

Runge'n esimerkissä ongelmana on polynomi

$$\omega(t) = (t - x_0)(t - x_1)\cdots(t - x_n),$$

joka esiintyy virhelausekkeessa. Jos pisteet valitaan tasavälisesti, niin välin päätepisteiden lähellä $\omega(t)$ saa suuria arvoja interpolaatiopisteiden lukumäärän n kasvaessa. Tätä epämiellyttävää ilmiötä voidaan välttää valitsemalla interpolaatiopisteet sopivasti.

Tschebyscheffin pisteet valitaan siten, että ko. virhepolynomi on tasaisesti mahdollisimman pieni. Oletetaan, että interpoloitavan funktion n+1 ensimmäinen derivaatta on rajoitettu välillä $[-1,1]$:

$$|f^{(n+1)}(t)| \le M.$$

Näin ollen polynomi-interpolaation virhelauseketta voidaan arvioida ylöspäin kuten

$$|f(t) - P(t)| \leq \frac{M}{(n+1)!} \max_{x \in [-1,1]} |\omega(x)|.$$

Virhepolynomin johtavan termin kerroin on yksi. Määritellään sellainen n+1-asteinen polynomi, jonka maksimi välillä $[-1, 1]$ on pienin mahdollinen. Tällöin tämän polynomin nollakohdat ovat siten optimaalisia interpolaatiopisteitä.

Tätä varten meidän on määriteltävä nk. Tschebyscheffin polynomit $T_n(x)$. Ne määritellään asettamalla

$$T_n(x) = T_n(\cos(\phi)) = \cos(n\phi), \ x = \cos(\phi) \in [-1, 1].$$

Trigonometrisen identiteetin

$$\cos((n+1)\phi) + \cos((n-1)\phi) = 2\cos(\phi)\cos(n\phi)$$

perusteella Tschebyscheffin funktioille on voimassa rekursiokaava

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Kaksi ensimmäistä funktiota ovat

$$T_0(x) = 1, \ T_1(x) = x.$$

Induktiivisesti voidaan päätellä, että funktio $T_n(x)$ on n-asteinen polynomi. Edelleen suoraan T-polynomien määritelmästä seuraa, että niiden maksimiarvot ovat

$$\max_{x \in [-1,1]} |T_n(x)| = 1,$$

ja ne saavutetaan derivaatan nollakohdissa:

$$\begin{aligned} \frac{d}{dx}T_n(x) &= \frac{d\phi}{dx}\frac{d}{d\phi}T_n(\cos(\phi)) \\ &= -n\phi'(x)\sin(n\phi) = 0. \end{aligned}$$

Näin ollen derivaatan nollakohdat ovat

$$x_k = \cos(\frac{k\pi}{n}), \ k = 0, 1, 2, \ldots, n.$$

Koska rekursiossa jokainen alempiasteinen T-polynomi kerrotaan funktiolla $2x$, niin T-polynomin johtavan termin (siis x^n:n) kerroin on 2^{n-1}. Näin ollen on voimassa

Lause 4.1.6 *Kaikille* $x \in [-1,1]$*:*

$$|2^{-n+1}T_n(x)| \leq \frac{1}{2^{n-1}}$$

Nyt voidaan osoittaa, että polynomin $\omega(x) = 2^{-n}T_n(x)$ nollakohdat ovat optimaalinen valinta interpolaatiopisteille. Nimittäin on voimassa lause:

Lause 4.1.7 (Tschebyscheff) *Jos polynomi* $P_n(x)$ *on n-asteinen ja jonka johtavan termin kerroin on yksi, niin*

$$\max_{x\in[-1,1]} |P_n(x)| \geq \frac{1}{2^{n-1}} \max_{x\in[-1,1]} |T_n(x)| = \frac{1}{2^{n-1}}.$$

Tod.: Oletetaan, että on olemassa polynomi $P_n(x)$, jonka asteluku on n ja

$$|P_n(x)| < \frac{1}{2^{n-1}},\ \forall x \in [-1,1].$$

Tällöin polynomin $T_n(x)$ ääriarvokohdissa $x_k,\ k = 0,\ldots,n$ on voimassa epäyhtälöt

$$\begin{aligned} P_n(x_0) &< \frac{1}{2^{n-1}} \\ P_n(x_1) &> -\frac{1}{2^{n-1}} \\ &\vdots \end{aligned}$$

Siten polynomilla

$$Q(x) = P_n(x) - \frac{1}{2^{n-1}}T_n(x)$$

on T-polynomin ääriarvokohtien välissä jatkuvuuden nojalla ainakin n eri nollakohtaa. Toisaalta polynomin $Q(x)$ asteluku on korkeintaan n-1, sillä molempien polynomien $P_n(x)$ ja $\frac{1}{2n-1}T_n(x)$ johtavat termit ovat x^n. Algebran päälauseen nojalla $Q(x)$ on identtisesti nolla, vastoin oletusta. Näin ollen oletus, että olisi olemassa polynomi $P_n(x)$, joka on kaikille $x \in [-1,1]$ itseisarvoltaan pienempi kuin

$$\frac{1}{2^{n-1}}$$

ja jonka johtavan termin kerroin olisi yksi, on väärä. Mikä oli todistettava. □

Mutta, mutta. Epäonneksi on olemassa funktioita, joita interpoloitaessa Tschebyscheffin pisteissä, interpolaatiopolynomit hajaantuvat, kun pisteiden lukumäärää kasvatetaan. Tämä ilmiö taasen johtuu, että n:nnen derivaatan maksimiarvot kasvavat rajatta.

4.1.6 Käänteisinterpolaatio:

Etsitään funktion $f(x)$ nollakohtaa välillä $[a, b]$ käyttäen hyväksi funktion $f(x)$ "käänteisfunktion"interpolaatiopolynomia. Oletetaan, että tarkasteltava funktio on välillä $[a, b]$ aidosti monotoninen. Valitaan väliltä $[a, b]$ pisteet $x_i,\ i = 0, \ldots, n$ siten, että

$$x_i < x_{i+1},\ i = 0, \ldots, n-1.$$

Tällöin funktio saa arvot $y_i = f(x_i),\ i = 0, \ldots, n$ pisteissä x_i. Lisäksi y_i:t ovat pareittain erisuuria. Määritellään y-muuttujan suhteen interpolaatiopolynomi $Q_n(y)$ ehdosta:

$$Q_n(y_i) = x_i,\ i = 0, \ldots, n.$$

Nyt funktion $f(x)$ nollakohdan approksimaatio on polynomin $Q_n(y)$ arvo nollassa l. $\tilde{x} = Q_n(0)$.

Luku 5

Paras approksimaatio

5.1 Johdanto

Ongelman määrittely

Kokeellisissa tieteissä joudutaan usein ratkaisemaan seuraavanlainen ongelma. Kahden tai useamman fysikaalisen suureen välistä riippuvuutta mitataan koejärjestelyllä. Olkoon $x_i,\ 1 = 1, \ldots, n$ säädettävät mittauspisteet (paikka, aika, paine, etc.) ja y_i:t vastaavat mittaustulokset. Mittausaineistoon on sovitettava lineaarinen funktio $p(x) = a_0 + a_1 x$ minimoimalla *pienin neliösumma*:

$$\sum_{i=1}^{n} |y_i - p(x_i)|^2. \tag{5.1}$$

Edellinen summalauseke määrittelee kahden muuttujan funktion

$$\begin{aligned} f(a_0, a_1) &= \sum_{i=1}^{n} [y_i - a_0 - a_1 x_i]^2 \\ &= \sum_{i=1}^{n} y_i^2 - 2a_0 \sum_{i=1}^{n} y_i - 2a_0 a_1 \sum_{i=1}^{n} x_i \\ &- 2a_1 \sum_{i=1}^{n} x_i y_i + a_1^2 \sum_{i=1}^{n} x_i^2 + n a_0^2. \end{aligned}$$

Funktio on määrittelynsä nojalla aina ei-negatiivinen. Lisäksi se on jatkuvasti derivoituva; ja kun $a_0^2 + a_1^2 \to \infty$ funktion raja-arvo on

$$\lim\ f(a_0, a_1) = \infty.$$

Tämän nojalla funktiolla on olemassa yksikäsitteinen minimi, joka löytyy gradientin nollakohdasta:

$$\begin{aligned}\frac{\partial f}{\partial a_0} &= -2\sum_{i=1}^{n}[y_i - a_1 x_i - a_0] = 0 \\ \frac{\partial f}{\partial a_1} &= -2\sum_{i=1}^{n} x_i[y_i - a_0 - a_1 x_i] = 0.\end{aligned}$$

Yhtälöä

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \tag{5.2}$$

kutsutaan *normaaliyhtälöksi*.

Normi l. funktioiden etäisyys

Funktioiden etäisyyttä voidaan mitata samalla tavalla kuin mitataan pisteiden etäisyyttä koordinaattiavaruudessa $\mathbf{R}^2$. Tarkastellaan lähinnä jatkuvia; mutta myös derivoituvia funktioita ja niiden approksimointia jonkin normin suhteen.

Määritelmä 5.1.1 *Funktion $f(x)$ normi $\|f\|$ on ei-negatiivinen luku, jolle on voimassa seuraavat ominaisuudet:*

1. *$\|f\| = 0$ jos ja vain jos $f = 0$;*
2. *Kaikille $\lambda \in \mathbf{R}$: $\|\lambda f\| = |\lambda|\|f\|$;*
3. *Kolmioepäyhtälö: $\|f + g\| \leq \|f\| + \|g\|$.*

Esimerkki 5.1 *Suljetulla ja rajoitetulla välillä määritellyn jatkuvan funktion $f(x)$ maksimi-normi*

$$\|f\|_\infty = \max_{a \leq x \leq b} |f(x)|.$$

Esimerkki 5.2 *Neliöintegroituville funktioille määritellään L^2-normi*

$$\|f\|_2 = \sqrt{\int_a^b |f(x)|^2 dx}.$$

Esimerkki 5.3 *Yleisesti voidaan määritellä jatkuville funktioille L^pnormi, kun $1 \leq p < \infty$ asettamalla*

$$\|f\|_p = [\int_a^b |f(x)|^p]^{\frac{1}{p}}.$$

Edellä mainittujen normien avuilla voidaan määritellä normi säännöllisempien funktioiden luokkaan.

Esimerkki 5.4 *Kuvaus*

$$\|f\|_{1,p} = [\|f(x)\|_p^p + \|f'(x)\|_p^p \, dx]^{\frac{1}{p}},$$

missä $f'(x)$ on funktion derivaatta, määrittelee jatkuvasti derivoituvien funktioiden luokkaan normin.

Edelleen funktioille voidaan määritellä pisteittäinen l. diskreetti seminormi:

Määritelmä 5.1.2 *Olkoon $x_i,\ i = 1, \ldots, n$ määritelty pistejoukko ja $f(x)$ jatkuva funktio välillä $[a,\ b]$, joka sisältää pisteet x_i. Tällöin*

$$|f|_p = [\sum_{i=1}^{n} |f(x_i)|^p]^{\frac{1}{p}},\ 1 \leq p \leq \infty$$

määrittelee seminormin, jolla on ominaisuudet

1. $|f|_p \geq 0$;
2. $\lambda f|_p = |\lambda||f|_p$, *kaikilla* $\lambda \in \mathbf{R}$;
3. $|f + g|_p \leq |f|_p + |g|_p$.

Huom! Jos funktion seminormi $|f|_p = 0$, niin siitä ei seuraa, että funktio olisi nollafunktio.

Tällä kurssilla approksimoidaan funktioita lähinnä pienimmän neliösumman ja L^2-normin suhteen. Mutta on hyvä tietää eksoottisimmistakin normiapproksimaatioista. Näitä tarvitaan esimerkiksi kuvankäsittelyssä.

5.1.1 Paras L^2-approksimaatio

Olkoon $\{\phi_j(x)|\ j = 1, \ldots, n\}$ approksimaatioavaruuden kantafunktiot, ja $f(x)$ välillä $[a,\ b]$ neliöintegroituva funktio (mielellään jatkuva). Ratkaistaan seuraava **a***pproksimaatio − ongelma*:

Probleema 5.1.1.1 *Määrää kertoimet* $a_j \in \mathbf{R},\ j = 1, \ldots, n$ *siten, että funktio*

$$f_n(x) = \sum_{j=1}^{n} a_j \phi_j(x)$$

minimoi L^2*-normin*

$$\|f - f_n\| = \{\int_a^b |f(x) - f_n(x)|^2 dx\}^{\frac{1}{2}}.$$

Koska yhtä hyvin voidaan minimoida L^2-normin neliötä, niin approksimaatio-ongelmassa haetaan n:n muuttujan funktion

$$F(a_1, \ldots, a_n) = \int_a^b |f(x) - \sum_{j=1}^{n} a_j \phi_j(x)|^2 dx$$

minimiä. Funktio $F(a_1, \ldots, a_n)$ on alhaalta rajoitettu, jatkuvasti differentioituva ja koersiivinen, ts.

$$\lim_{\|\vec{a}\| \to \infty} F(a_1, \ldots, a_n) = \infty.$$

Tällöin funktiolla on minimi ja se löytyy gradientin nollakohdasta:

$$\begin{aligned}
\frac{\partial F}{\partial a_1} &= -2 \int_a^b [f(x) - \sum_{j=1}^{n} a_j \phi_j(x)] \phi_1(x) dx = 0 \\
\vdots &= \vdots \\
\frac{\partial F}{\partial a_n} &= -2 \int_a^b [f(x) - \sum_{j=1}^{n} a_j \phi_j(x)] \phi_n(x) dx = 0.
\end{aligned}$$

Näin ollen kertoimet a_j ratkaistaan *normaaliyhtälöstä*:

$$\begin{bmatrix} \int_a^b \phi_1(x)\phi_1(x)dx & \cdots & \int_a^b \phi_1(x)\phi_n(x)dx \\ \int_a^b \phi_2(x)\phi_1(x)dx & \cdots & \int_a^b \phi_2(x)\phi_n(x)dx \\ \vdots & \ddots & \vdots \\ \int_a^b \phi_n(x)\phi_1(x)dx & \cdots & \int_a^b \phi_n(x)\phi_n(x)dx \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \int_a^b f(x)\phi_1(x)dx \\ \vdots \\ \int_a^b f(x)\phi_n(x)dx \end{bmatrix}$$

Tarkastellaan myöhemmin ehtoja, joiden vallitessa normaaliyhtälöllä on yksikäsitteinen ratkaisu. Nyt todetaan vain, että approksimaatio-avaruus tulisi valita siten, että yhtälöryhmän ratkaisu olisi numeerisesti stabiili, ts. kerroinmatriisin ehtoluku olisi siedettävä. Lisäksi olisi toivottavaa, että approksimaatioavaruuden funktioilla olisi voimassa approksimaatio-ominaisuus:

$$\lim_{n\to\infty} \|f - f_n\| = 0.$$

Esimerkki 5.5 *Tarkastellaan funktion approksimointia m-asteisella polynomifunktiolla välillä* $[0,\ 1]$*. Polynomifunktioiden kantafunktioiksi voidaan valita esimerkiksi funktiot*

$$\phi_j(x) = x^{j-1},\ j = 1, 2, 3, \ldots, m+1.$$

Tällöin normaaliyhtälön kerroinmatriisin alkiot ovat

$$A_{i,j} = \int_0^1 x^{i+j-2} dx = \frac{1}{i+j-1},\ i, j = 1, \ldots,\ n$$

ja oikean puolen vektorin alkiot ovat

$$b_i = \int_0^1 f(x) x^{i-1} dx,\ i = 1, \ldots, n.$$

5.1.2 Pienimmän neliösumman menetelmä

Tarkastellaan seuraavaksi funktion approksimointia *pienimmän neliösumman menetelmällä*. Sitä varten olkoot pisteet $\{x_k|\ k = 1, \ldots, m\}$ valittu väliltä $[a,\ b]$ ja funktio $f(x)$ määrittelyjoukossaan ainakin jatkuva.

Probleema 5.1.2.1 *Määrää kertoimet* $a_1, \ldots, a_n \in \mathbf{R}$ *siten, että funktio*

$$f_n(x) = \sum_{j=1}^{n} a_j \phi_j(x)$$

minimoi l^2*-normin l. neliösumman*

$$|f - f_n|^2 = \sum_{k=1}^{m} |f(x_k) - f_n(x_k)|^2.$$

Funktio

$$F(a_1, \ldots, a_n) = \sum_{k=1}^{m} |f(x_k) - \sum_{j=1}^{n} a_j \phi_j(x_k)|^2$$

on alhaalta rajoitettu, jatkuvasti differentioituva ja koersiivinen, ts.

$$\lim_{\|\vec{a}\| \to \infty} F(a_1, \ldots, a_n) = \infty.$$

Tällöin sillä on minimi ja se löytyy gradientin nollakohdasta:

$$\begin{aligned}
\frac{\partial F}{\partial a_1} &= -2 \sum_{k=1}^{m} [f(x_k) - \sum_{j=1}^{n} a_j \phi_j(x_k)] \phi_1(x_k) = 0 \\
\vdots &= \vdots \\
\frac{\partial F}{\partial a_n} &= -2 \sum_{k=1}^{m} [f(x_k) - \sum_{j=1}^{n} a_j \phi_j(x_k)] \phi_n(x_k) = 0.
\end{aligned}$$

Näin ollen kertoimet a_j ratkaistaan *normaaliyhtälöstä*:

$$\begin{bmatrix}
\sum_{k=1}^{m} \phi_1(x_k)\phi_1(x_k) & \cdots & \sum_{k=1}^{m} \phi_1(x_k)\phi_n(x_k) \\
\sum_{k=1}^{m} \phi_2(x_k)\phi_1(x_k) & \cdots & \sum_{k=1}^{m} \phi_2(x_k)\phi_n(x_k) \\
\vdots & \ddots & \vdots \\
\sum_{k=1}^{m} \phi_n(x_k)\phi_1(x_k) & \cdots & \sum_{k=1}^{m} \phi_n(x_k)\phi_n(x_k)
\end{bmatrix}
\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} =
\begin{bmatrix} \sum_{k=1}^{m} f(x_k)\phi_1(x_k) \\ \vdots \\ \sum_{k=1}^{m} f(x_k)\phi_n(x_k) \end{bmatrix}$$

Esimerkki 5.6 *Tarkastellaan funktion approksimointia m-asteisella polynomifunktiolla. Polynomifunktioiden kantafunktioiksi voidaan valita esimerkiksi funktiot*

$$\phi_j(x) = x^{j-1}, \ j = 1, 2, 3, \ldots, m+1.$$

Tällöin normaaliyhtälön kerroinmatriisin alkiot ovat

$$A_{i,j} = \sum_{k=1}^{m} x_k^{i+j-2},$$

ja oikean puolen vektorin alkiot ovat

$$b_i = \sum_{k=1}^{m} f(x_k) x_k^{i-1}.$$

5.1.3 Approksimaatio-ominaisuus

Seuraavassa joitain tärkeimpiä approksimaatioon käytettäviä funktioita, joille voidaan osoittaa approksimaatio-ominaisuus todeksi:

- polynomit muodostavat tiheän joukon;
- trigonometriset polynomit;
- spline-funktiot;
- wavelet-approksimaatio (tavallisesti käytetään spline-pohjaisia wavelet-teja);

Normaaliyhtälö on yksikäsitteisesti ratkeava mikäli funktiot

$$\{\phi_j(x),\ j = 1, \ldots, n\}$$

ovat lineaarisesti riippumattomia.

Määritelmä 5.1.3 *Funktiot*

$$\{\phi_j(x),\ j = 1, \ldots, n\}$$

ovat lineaarisesti riippumattomat, jos ehdosta

$$\sum_{j=1}^{n} a_j \phi_j(x) = 0,\ \forall\ x \in [a,\ b]$$

seuraa, että välttämättä kertoimet $a_j = 0$ *kaikille* $j = 1, \ldots, n$.

Ilmeisesti ainakin polynomifunktiot ja trigonometriset funktiot ovat lineaarisesti riippumattomia. Seuraavaksi tarkastelemmekin näiden funktioidn ominaisuuksia hieman tarkemmin.

5.2 Fourier-approksimaatio

5.2.1 Jatkuva Fourier-approksimaatio

Approksimoitava funktio:

- periodisuus; $f(x + 2\pi) = f(x),\ \forall x \in \mathbf{R}$.

- paloittain jatkuvuus: Funktiolla on äärellinen määrä epäjatkuvuuskohtia ja epäjatkuvuuskohdassa on toispuoleiset raja-arvot, ts.

$$\lim_{h\to 0^+} f(x_0 \pm h) = y_0^{\pm},\ -\infty < y_0^{\pm} < \infty.$$

Approksimaatio-ongelma: Etsi kertoimet a_k ja b_k s.e.

$$g_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^{n}\{a_k \cos(kx) + b_k \sin(kx)\}$$

minimoi lausekkeen

$$\min \|g_n - f\|^2 = \int_{-\pi}^{\pi} |g_n(x) - f(x)|^2 dx.$$

Lause 5.2.1 (Ortogonaalisuus) *Trigonometriset funktiot ovat L^2-sisätulon suhteen pareittain ortogonaaliset:*

$$\begin{aligned}
\int_{-\pi}^{\pi} \cos(jx)\cos(kx)dx &= \begin{cases} 0,\ j \neq k \\ 2\pi,\ j = k = 0 \\ \pi,\ j = k > 0 \end{cases} \\
\int_{-\pi}^{\pi} \sin(jx)\sin(kx)dx &= \begin{cases} 0,\ j \neq k \\ \pi,\ j = k > 0 \end{cases} \\
\int_{-\pi}^{\pi} \sin(jx)\cos(kx)dx &= 0,\ \forall j, k > 0.
\end{aligned}$$

Ortogonaalisuusrelaation nojalla pienimmän neliösumman menetelmä päätyy lineaariseen yhtälöryhmään, jonka ratkaisu on

Lause 5.2.2

$$\begin{aligned}
a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx \\
b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx.
\end{aligned}$$

Lause 5.2.3 *Olkoon $f(x)$ paloittain jatkuva $2\pi-$periodinen funktio. Silloin sen Fourier-approksimaatio suppenee pisteittäin kohti funktioita f sen jatkuvuuskohdissa, ja epäjatkuvuuskohdissa*

$$\lim_{n\to\infty} g_n(x_0) = \frac{1}{2}[y_0^+ + y_0^-].$$

Lause 5.2.4 *Olkoon $f(x)$ 2π-periodinen ja jatkuvasti derivoituva funktio. Tällöin funktion Fourier-approksimaation neliöllinen virhe on*

$$\|f - g_n\|_{L^2(-\pi,\pi)} \equiv \{\int_{-\pi}^{\pi} |f(x) - g_n(x)|^2 dx\}^{\frac{1}{2}} \leq \sqrt{\frac{2}{\pi}\frac{1}{n}} \int_{-\pi}^{\pi} |f'(x)|^2 dx.$$

Lause 5.2.5 *Olkoon $f(x)$ 2π-periodinen ja m-kertaa jatkuvasti derivoituva. Silloin Fourier-approksimation virhe on*

$$\|f - g_n\|_{L^2(-\pi,\pi)} \leq \frac{c}{n^m} \int_{-\pi}^{\pi} |f^{(m)}(x)|^2 dx.$$

5.2.2 Diskreetti Fourier-muunnos ja FFT

Approksimoidaan tasavälisellä hilalla $\Delta = \{t_i = \frac{2\pi}{N} i |\ i = 0, \ldots, N-1\}$ 2π-periodista funktiota trigonometrisillä polynomeilla kuten edellisessä paragraafissa käyttäen pienimmän neliösumman menetelmää. Approksimaation kantafunktioina käytetään kompleksisia funktioita

$$\phi_k(t) = e^{jkt},\ k = -M, -M+1, \ldots, 0, \ldots, M-1, M,$$

missä $j = \sqrt{-1}$ on imaginaariyksikkö. Tällöin funktiota $f(t)$ approksimoidaan M:nnen asteen trigonometrisellä polynomilla

$$f_M(x) = \sum_{k=-M}^{M} d_k e^{jkt}.$$

Neliösumman

$$\sum_{i=0}^{N-1} |f(t_i) - f_M(t_i)|^2$$

pienin arvo löytyy gradientin nollakohdasta. Näin saadaan *normaaliyhtälö*

$$\sum_{i=0}^{N-1} e^{-j\frac{2\pi l i}{N}} [f(t_i) - \sum_{k=-M}^{M} d_k e^{j\frac{2\pi k i}{N}}] = 0,\ l = -M, \ldots, M,$$

joka voidaan kirjoittaa muodossa

$$\sum_{i=0}^{N-1} \sum_{k=-M}^{M} d_k e^{-j\frac{2\pi(l-k)i}{N}} = \sum_{i=0}^{N-1} e^{-j\frac{2\pi l i}{N}} f(t_i).$$

Lemma 5.2.1 *Kantafunktioille on voimassa seuraava ortogonaalisuusrelaatio:*

$$\sum_{i=0}^{N-1}[e^{-j\frac{2\pi(l-k)}{N}}]^i = \begin{cases} N, \ k = l + vN \\ 0, \ k \neq l + vN \end{cases} \quad v \in \mathbb{Z}$$

Lemman nojalla normaaliyhtälö on

$$N \sum_{l+vN=-M}^{M} d_{l+vN} = \sum_{i=0}^{N-1} f(t_i)e^{-j\frac{2\pi li}{N}}, \ v \in \mathbb{Z}, \ l = -M, \ldots, M.$$

Jotta jokainen 2M+1:stä yhtälöstä sisältää täsmälleen yhden kertoimen d_k täytyy olla voimassa ehto:

$$M < \frac{N}{2}.$$

Tällöin jokaiselle $v > 0$ on voimassa

$$\begin{aligned} l + vN &> l + 2vM > M, \ \forall l \geq -M \\ l - vN &< l - 2vM < -M, \ \forall l \leq M. \end{aligned}$$

Näin ollen yo. summissa $v = 0$ ja summattavana on vain yksi termi d_l. Trigonometrisen approksimaation kompleksiset kertoimet saadaan siis kaavasta

$$d_k = \frac{1}{N} \sum_{i=0}^{N-1} f(t_i)e^{-j\frac{2\pi ki}{N}}, \ k = -M, \ldots, M.$$

Trigonometrisen approksimaation reaalinen esitys kosini- ja sinifunktioiden avulla saadaan käyttämällä Eulerin kaavoja

$$\begin{aligned} \cos(kt) &= \frac{e^{jkt} + e^{-jkt}}{2} \\ \sin(kt) &= \frac{e^{jkt} - e^{-jkt}}{2j}. \end{aligned}$$

Sijoittamalla nämä trigonometrisen polynomin

$$f_M(t) = a_0 + \sum_{k=1}^{M} \{a_k \cos(kt) + b_k \sin(kt)\}$$

lausekkeeseen saadaan kompleksinen muoto

$$f_M(t) = a_0 + \sum_{k=1}^{M} \frac{a_k - jb_k}{2} e^{jkt} + \sum_{k=-1}^{-M} \frac{a_k + jb_k}{2} e^{-jkt}.$$

Vertaamalla kertoimia kompleksisen Fourier-sarjan kertoimiin havaitaan, että reaalisen Fourier-sarjan kertoimet ovat

$$\begin{aligned} a_0 &= d_0, \\ a_k &= 2\mathcal{R}e(d_k), \\ b_k &= -2\mathcal{I}m(d_k). \end{aligned}$$

Näin olemme todistaneet seuraavan

Lause 5.2.6 *Olkoon funktio $f(t)$ 2π-periodinen jatkuva funktio ja $2M < N$. Tällöin funktio*

$$f_M(x) = a_0 + \sum_{k=1}^{M} \{a_k \cos(kt) + b_k \sin(kt)\},$$

on funktion $f(t)$ paras approksimaatio pienimmän neliösumman mielessä:

$$\sum_{j=1}^{N} |f_M(x_j) - f(x_j)|^2 = \min!,$$

missä kertoimet $a_k,\ b_k$ ovat

$$\begin{aligned} a_0 &= \frac{1}{N} \sum_{i=0}^{N-1} f(t_i) \\ a_k &= \frac{2}{N} \sum_{i=0}^{N-1} f(t_i) \cos(kt_i),\ \ k = 0, 1, 2, \ldots \\ b_k &= \frac{2}{N} \sum_{i=0}^{N-1} f(t_i) \sin(kt_i),\ \ k = 0, 1, 2, \ldots, \end{aligned}$$

Trigonometrisessa interpolaatiossa etsitään trigonometrista polynomia $f_M(x)$ siten, että $f_M(t_i) = f(t_i),\ i = 0, \ldots, N-1$, missä pisteet t_i ovat kuten edellä.

Koska yhtälöitä on N kappaletta, pitää tuntemattomien kertoimien määrä olla sama. Näin ollen trigonometrista interpolaatiota etsitään muodossa

$$f_M(t) = a_0 + \sum_{k=1}^{M-1} \{a_k \cos(kt) + b_k \sin(kt)\} + \frac{1}{2} a_M \cos(Mt),$$

missä $2M = N$. Tällöin yhtälöryhmässä on sama määrä tuntemattomia kuin yhtälöitä. Trigonometrinen interpolaatio voidaan nyt ratkaista olennaisesti samalla tavalla. Nimittäin voidaan osoittaa seuraava lause:

Lause 5.2.7 *Olkoon funktio $f(t)$ kuten edellä 2π-periodinen ja riittävän sileä. Tällöin on yksikäsitteisesti määrätty trigonometrinen interpolaatiofunktio $f_M(t)$ pisteiden $t_i = \frac{2\pi}{N} i$ suhteen (N=2M):*

$$f_M(t_i) = f(t_i), \ i = 0, \ldots, N-1.$$

Interpolaatiofunktion esitys on

$$f_M(t) = a_0 + \sum_{k=1}^{M-1} \{a_k \cos(kt) + b_k \sin(kt)\} + \frac{1}{2} a_M \cos(Mt),$$

missä kertoimet lasketaan kuten edellisessä lauseessa.

Matriisiesitys Diskreetit Fourier-kertoimet voidaan laskea matriisikertolaskuna. Sitä varten määritellään luku $w = e^{-\frac{2\pi j}{N}}$ ja Fourier-matriisi

$$W = [w^{ki}] \begin{array}{l} k = -M, \ldots, M \\ i = 0, \ldots, N-1 \end{array},$$

missä indeksi k on rivi-indeksi ja i sarakeindeksi. Tällöin diskreetin Fourier-approksimaation kertoimet lasketaan kertomalla Fourier-matriisilla datavektori $\underline{f} = [f(t_0) \ \cdots \ f(t_{N-1})]^T$:

$$\begin{bmatrix} d_{-M} \\ d_{-M+1} \\ \vdots \\ d_M \end{bmatrix} = \underline{d} = \frac{1}{N} W \underline{f}.$$

Kysessä oleva matriisikertolasku voidaan suorittaa tehokkaasti käyttämällä hyväksi nopeata Fourier-muunnosta, josta hieman enemmän seuraavassa kappaleessa.

5.2.3 Nopea Fourier-muunnos (FFT)

Lasketaan diskreetin Fourier-muunnoksen kertoimet

$$\begin{aligned} a'_k &= \sum_{j=0}^{N-1} f(x_j)\cos(kx_j) \\ b'_k &= \sum_{j=0}^{N-1} f(x_j)\sin(kx_j) \end{aligned}$$

nopeasti. Sitä varten oletetaan, että hilapisteiden lukumäärä on $N = 2^p$. Määritellään kompleksinen datavektori:

$$y_j = f(x_{2j}) + \imath f(x_{2j+1}),\ j = 0, 1, \ldots, \frac{N}{2} \text{ ja } \imath = \sqrt{-1}.$$

Diskreetti, kompleksinen Fourier-muunnos ($n = \frac{N}{2}$):

$$c_k = \sum_{j=0}^{n-1} y_j e^{-\imath jk\frac{2\pi}{n}}.$$

Seuraavan lauseen avulla voidaan kompleksisista kertomista laskea reaalisen Fourier-muunnoksen kertoimet:

Lause 5.2.8 *Asetetaan apukertoimet $b'_0 = b'_n = 0$ ja $c_0 = c_n$. Tällöin kaikille $k = 0, \ldots, n$:*

$$\begin{aligned} a'_k - \imath b'_k &= \frac{1}{2}[c_k + \overline{c}_{n-k}] + \frac{1}{2\imath}[c_k - \overline{c}_{n-k}]e^{-\frac{\imath k\pi}{n}} \\ a'_{n-k} - \imath b'_{n-k} &= \frac{1}{2}[\overline{c}_k + c_{n-k}] + \frac{1}{2\imath}[\overline{c}_k - c_{n-k}]e^{\frac{\imath k\pi}{n}}. \end{aligned}$$

Nopean Fourier-muunnoksen reduktioaskel:

Olkoon jatkossa $n = 2m$. Tällöin parilliset kertoimet voidaan kirjoittaa muodossa

$$\begin{aligned} c_{2l} &= \sum_{j=0}^{2m-1} y_j e^{-\frac{\imath jl2\pi 2}{2m}} = \sum_{j=0}^{2m-1} y_j e^{-\frac{\imath jl2\pi}{m}} \\ &= \sum_{j=0}^{m-1} y_j e^{-\frac{\imath jl2\pi}{m}} + \sum_{j=0}^{m-1} y_{j+m} e^{-\frac{\imath (j+m)l2\pi}{m}} \\ &= \sum_{j=0}^{m-1} y_j e^{-\frac{\imath jl2\pi}{m}} + \sum_{j=0}^{m-1} y_{j+m} e^{-\frac{\imath jl2\pi}{m}} e^{-\imath 2\pi l} \\ &= \sum_{j=0}^{m-1} (y_j + y_{j+m}) e^{-\frac{\imath jl2\pi}{m}}. \end{aligned}$$

Vastaavasti parittomille kertoimille pätee:

$$c_{2l+1} = \sum_{j=0}^{m-1} (y_j - y_{j+m}) e^{-\imath \frac{2\pi j}{n}} e^{-\frac{\imath jl2\pi}{m}}.$$

Määrittelemällä apusuureet

$$\begin{aligned} z_j &= y_j + y_{j+m},\ j = 0, \ldots, m-1 \\ z_{j+m} &= (y_j - y_{j+m}) e^{-\imath \frac{2\pi j}{n}},\ j = 0, \ldots, m-1, \end{aligned}$$

voidaan n-ulotteiset Fourier-muunnokset palauttaa kahdeksi m-ulotteiseksi Fourier-muunnokseksi:

$$\begin{aligned} c_{2l} &= \sum_{j=0}^{m-1} z_j e^{-\imath \frac{jl2\pi}{m}} \\ c_{2l+1} &= \sum_{j=0}^{m-1} z_{m+j} e^{-\imath \frac{jl2\pi}{m}}. \end{aligned}$$

5.3 Ortogonaaliset polynomit ja approksimaatio

5.3.1 Tschebyscheffin approksimaatio

Määritelmä 5.3.1 *Määritellään jokaiselle* $x = \cos(\phi),\ \phi \in [0,\ \pi]$,

$$T_k(x) = \cos(k\phi).$$

Funktio $T_k(x)$ on ns. Tschebyscheffin polynomi. Määritelmän nojalla

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x. \end{aligned}$$

Vaikka alunperin T-polynomit ovat määritelty vain välillä $[-1,\ 1]$, niin ilmeisesti funktiot ovat määriteltyjä kaikille x:n arvoille.

T-polynomit on helppo generoida seuraavan rekursiokaavan avulla

Lause 5.3.1 *Tschebyscheffin polynomeille on voimassa rekursiokaava:*

$$T_{k+1}(x) + T_{k-1}(x) = 2xT_k(x).$$

Tod. Rekursiokaava perustuu trigonometriseen identiteettiin:

$$\cos((k+1)\phi) + \cos((k-1)\phi) = 2\cos(\phi)\cos(k\phi)$$

.

Seuraavassa muutamia T-polynomeja:

$$\begin{aligned} T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x^2 \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ \vdots &= \vdots \end{aligned}$$

Lause 5.3.2 *Tschebyscheffin polynomeille on voimassa seuraavat ominaisuudet:*

1. $|T_k(x)| \leq 1$ *kaikille* $x \in [-1,\ 1]$ *ja kaikille* $k \geq 0$*;*

2. $T_k(x)$*:n nollakohdat ovat*

$$x_p = \cos(\frac{(2p-1)\pi}{2k}),\ p = 1, \ldots, k;$$

3. $T_k(-x) = (-1)^k T_k(x),\ k \geq 0$*;*

4. Ortogonaalisuusrelaatio:

$$\int_{-1}^{1} T_k(x)T_l(x)\frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & jos\ k \neq l \\ \frac{\pi}{2}, & jos\ k = l \neq 0 \\ \pi, & jos\ k = l = 0 \end{cases}$$

Nyt jokainen m-asteinen polynomi voidaan lausua Tschebyscheffin polynomien avulla (T-polynomit muodostavat kannan polynomiavaruuteen):

$$P_n(x) = \frac{c_0}{2} + \sum_{k=1}^{n} c_k T_k(x).$$

Probleema 5.3.1.1 *Minimoidaan painoitettu* L^2*-normin neliö*

$$\|f - P_n\|_{2,\omega}^2 = \int_{-1}^{1} [f(x) - \frac{c_0}{2} - \sum_{k=1}^{n} c_k T_k(x)]\frac{dx}{\sqrt{1-x^2}},$$

missä painofunktio $\omega(x) = \frac{1}{\sqrt{1-x^2}}$.

Koska Tschebyscheffin polynomit ovat ortogonaalisia painotetun L^2-sisätulon suhteen, niin normaaliyhtälön kerroinmatriisi on diagonaalinen (diagonaalialkio $d = \frac{\pi}{2}$). Siten seuraava lause on helposti todennettavissa:

Lause 5.3.3 *Olkoon* $f(x)$ *jatkuva funktio välillä* $[-1,\ 1]$*. Tällöin approksimaatio-ongelmalla on yksikäsitteinen ratkaisu, ja kertoimet*

$$c_k = \frac{2}{\pi}\int_{-1}^{1} f(x)T_k(x)\frac{dx}{\sqrt{1-x^2}}.$$

5.3.2 Ortogonaaliset polynomit

Ortogonaalisuus L^2-sisätulon suhteen

Tarkastellaan aluksi tapausta, jossa painofunktio $w(x) \equiv 1$. Ortogonaaliset polynomit muodostavat jonon $\{p_i\}_{i=0}^{\infty}$, jonka alkiot ovat polynomeja siten, että

1. $\deg(p_i) = i$;
2. $\int_a^b p_i(x)p_j(x)dx = 0$.

Ortogonaaliset polynomit normitetaan joko vaatimalla että

$$\int_a^b |p_i(x)|^2 dx = 1$$

tai

$$p_i(x) = x^i + \text{alemman asteen polynomi.}$$

Lause 5.3.4 *Olkoon $\{p_i\}_{i=0}^{\infty}$ joukko ortogonaalisia polynomeja siten, että*

$$deg(p_i) = i.$$

Tällöin jokaiselle polynomille

$$q(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$$

on olemassa yksikäsitteisesti kertoimet b_i siten, että

$$q = b_n p_n + b_{n-1}p_{n-1} + \cdots + b_0 p_0.$$

Tod.: Oletetaan, että ortogonaalisten polynomien johtavien termien kertoimet ovat 1. Tällöin vakiopolynomille

$$q(x) = a_0 = a_0 \cdot 1 = a_0 p_0(x).$$

Oletetaan, että väite on tosi kaikille polynomeille $q(x)$, joiden aste on pienempi kuin n-1.

Olkoon sitten $q_n(x)$ n-asteinen polynomi. Koska polynomin $p_n(x)$ johtavan termin kerroin on 1, niin polynomin $q_n(x) - a_n p_n(x)$ asteluku on n-1. Tällöin induktio-oletuksen nojalla on olemassa yksikäsitteisesti luvut $b_{n-1}, \ldots, b_0$ siten, että

$$q_n(x) - a_n p_n(x) = b_{n-1}p_{n-1}(x) + \cdots + b_0 p_0,$$

mikä todistaakin väittämän. □

Edellisen lauseen seurauksena voimme päätellä

Lause 5.3.5 *Polynomi $p_{n+1}(x)$ on ortogonaalinen kaikkia alempiasteisia polynomeja vasten.*

Ortogonaalisten polynomien nollakohdat

Lause 5.3.6 (Stewart, 1998) *Ortogonaalisen polynomin $p_n(x)$ nollakohdat ovat reaalisia, yksinkertaisia ja ovat kaikki välillä $[a,\ b]$.*

Tod.: Olkoon pisteet $x_0, x_1, \ldots, x_k$ ortogonaalisen polynomin $p_n(x)$ ne nollakohdat välillä $[a,\ b]$, jossa funktio vaihtaa merkkiä. Jos $k+1 = n$, niin väite on tosi.

Oletetaan siksi, että $k + 1 < n$. Polynomin

$$r(x) = (x - x_0)(x - x_1) \cdots (x - x_k)$$

asteluku on $k + 1 < n$ ja siten lauseen 3.3.5 nojalla

$$\int_a^b p_n(x)q(x)dx = 0.$$

Toisaalta polynomi $p_n(x)q(x)$ ei vaihda merkkiä välillä $[a,\ b]$, sillä jokainen polynomin $p_n(x)$ merkinvaihto kompensoidaan vastaavalla $q(x)$:n merkin vaihdolla. Näin ollen

$$\int_a^b q(x)p_n(x)dx \neq 0,$$

mikä on ristiriita oletuksen kanssa. □

Yleiset ortogonaaliset polynomit

Tavoitteena on kehittää menetelmä, jonka avulla mielivaltaisella rajoitetulla ja suljetulla välillä painotetun sisätulon suhteen voidaan määrätä ortogonaaliset polynomit. Ortogonaalisuusehto on silloin

$$\int_a^b w(x)p_m(x)p_n(x)dx = \begin{cases} 0,\ m \neq n \\ c_n,\ n = m. \end{cases}$$

Kuten edellisessä kappaleessa voidaan yleisesti osittaa, että ortogonaalinen polynomi $p_n(x)$ on ortogonaalinen kaikkia alempiasteisia polynomeja vasten. Näin ollen määrittelevä ehto voidaan myös kirjoittaa muodossa

$$\int_a^b w(x)p_n(x)x^m dx = 0,\ \forall\ m < n.$$

Valitaan polynomi $p_n(x)$ siten, että se on muotoa

$$p_n(x) = \frac{1}{w(x)} \frac{d^n U_n(x)}{dx^n}.$$

Tällöin funktio $U_n(x)$ toteuttaa differentiaaliyhtälön

$$\frac{d^{n+1}}{dx^{n+1}}[\frac{1}{w(x)} \frac{d^n U_n(x)}{dx^n}] = 0,$$

sillä n-asteisen polynomin kertalukua n+1 oleva derivaatta häviää identtisesti.

Ortogonaalisuusehdon nojalla jokaiselle m-asteiselle (m<n) polynomille $q_m(x)$

$$\int_a^b \frac{d^n U_n(x)}{dx^n} q_m(x) dx = 0.$$

Osittaisintegroimalla lauseke m+1 kertaa saadaan identiteetti

$$0 = \sum_{j=1}^{m+1} (-1)^{j+1} \Big/_a^b \frac{d^{n-j} U_n}{dx^{n-j}} q_m^{(j-1)}.$$

Näin ollen funktion $U_n(x)$ on toteutettava reunaehdot

$$U_n^{(j)}(a) = U_n^{(j)}(b) = 0, \ j = 0, \ldots, n-1.$$

Siten funktio $U_n(x)$ on reuna-arvotehtävän

$$\begin{aligned} \frac{d^n}{dx^n}[\frac{1}{w(x)} \frac{d^n U_n(x)}{dx^n}] &= A_n \\ U_n^{(j)}(a) = U_n^{(j)}(b) &= 0, \ j = 0, \ldots, n-1 \end{aligned}$$

ratkaisu, missä vakio A_n on mielivaltainen reaaliluku. Siten ortogonaaliset polynomit ovat normeerauskerrointa vaille yksikäsitteisesti määrätyt.

Ortogonaaliset polynomit ja approksimaatio Oletetaan, että ortogonaaliset polynomit $p_k(x), \ k = 0, 1, \ldots,$ on määrätty. Välillä [a, b] määritellyn funktion $f(x)$ paras polynomiapproksimaatio L^2-normin suhteen voidaan esittää ortogonaalisten polynomien avulla:

$$Q_n(x) = \sum_{k=0}^{n} b_k p_k(x).$$

Kertoimet määrätään minimoimalla interaali

$$\int_a^b w(x)|f(x) - Q_n(x)|^2 dx.$$

Tämä johtaa normaaliyhtälöön, jonka kerroinmatriisi on diagonaalinen, koska polynomit $p_k(x)$ ovat pareittain ortogonaalisia painotetun sisätulon suhteen. Näin ollen kertoimet voidaan ratkaista yhtälöistä

$$b_k \int_a^b w(x)p_k(x)^2 dx = \int_a^b w(x)f(x)p_k(x)dx.$$

Ongelma on siten täydellisesti ratkaistu, jos kertoimet

$$c_k = \int_a^b w(x)p_k(x)^2 dx$$

osataan määrätä.

Ne voidaan laskea edeltä käsin riippumatta funktiosta $f(x)$. Jokainen polynomi $p_k(x)$ voidaan esittää muodossa

$$p_k(x) = a_0 + a_1 x + \cdots + a_{k-1}x^{k-1} + a_k x^k.$$

Koska $p_k(x)$ on ortogonaalinen alempiasteisten polynomien kanssa, niin

$$c_k = a_k \int_a^b w(x)x^k p_k(x)dx.$$

Legendren approksimaatio Sovelletaan edellä esitettyä menetelmää polynomiapproksimaatioon välillä [-1, 1], kun painofunktio $w(x) \equiv 1$.

Tällöin $U_n(x)$ ratkaistaan reuna-arvotehtävästä

$$\begin{aligned} \frac{d^n U_n}{dx^n} &= B_n \\ U_n^{(k)}(\pm 1) &= 0,\ k = 0, \ldots, n-1. \end{aligned}$$

Yhtälön ratkaisu on 2n-asteinen polynomi. joka toteuttaa kaikki 2n reunaehtoa. Helposti havaitaan, että ko. funktio on

$$U_n(x) = B_n(x^2 - 1)^n.$$

Tavallisesti funktio normitetaan valitsemalla $B_n = \frac{1}{2^n n!}$. Näin saadaan ns. Rodrigues'n kaavat:

$$p_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2-1)^n].$$

Legendre'n polynomien konstruoiminen käyttäen Rodriguesin kaavoja on kuitenkin aika vaivalloista. Ne voidaan laskea varsin sujuvasti käyttämällä rekursiokaavoja (kolmen termin rekursio): $P_0(x) = 1,\ P_1(x) = x$ ja

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x),\ n = 1, 2, \ldots\ .$$

Kertoimen c_k laskeminen Legendre'n polynomeille tapahtuu seuraavasti. Polynomin $p_k(x)$ johtava kerroin saadaan kehitelmästä

$$\frac{d^k}{dx^k}[(x^2-1)^k] = \frac{(2k)!}{k!}x^k - \cdots\ .$$

Näin ollen

$$\begin{aligned} c_k &= \int_{-1}^{1} p_k(x)^2 dx = \frac{(2k)!}{2^{2k}(k!)^2}\int_{-1}^{1} x^k p_k(x)dx \\ &= \frac{(2k)!}{2^{2k}(k!)^2}\frac{1}{2^k k!}\int_{-1}^{1} x^k \frac{d^k}{dx^k}[(x^2-1)^k]dx. \end{aligned}$$

Osittaisintegroimalla k kertaa saadaan lopulta kertoimelle c_k lauseke

$$c_k = \frac{(2k)!}{2^{2k}(k!)^2}\int_{-1}^{1}(1-x^2)^k dx = \frac{2}{2k+1}.$$

Paras polynomiapproksimaatio Funktion $f(x)$ paras polynomiapproksimaatio L^2-normin suhteen välillä [-1,1] on tällöin

$$Q_n(x) = \sum_{k=0}^{n} b_k p_k(x),$$

missä kertoimet

$$b_k = \frac{2k+1}{2}\int_{-1}^{1} f(x)p_k(x)dx.$$

Luku 6

Numeerinen differentiaalilaskenta

6.1 Numeerinen integrointi

6.1.1 Interpolaatiokaavat

Approksimoitava integraali

$$I = \int_a^b f(x)dx.$$

Määritellään tasavälisellä hilalla funktion $f(x)$ interpolaatiopolynomi

$$P_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}).$$

Integraalin I approksimaatio on tällöin

$$I_n = \int_a^b P_n(x)dx.$$

Approksimaation virhe on tällöin

$$I - I_n = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi_x) \Pi_{j=0}^n (x - x_j)dx.$$

Keskipistekaava

Interpoloidaan funktiota vakiofunktiolla integroimisvälin keskipisteessä

$$x_0 = \frac{a+b}{2}.$$

Integroimiskaava on tällöin

$$I_0 = \int_a^b f(x_0)dx = (b-a)f(x_0).$$

Keskipistekaavan virhettä varten kirjoitetaan vakiointerpolaation virhelauseke muodossa

$$f(x) - f(x_0) = f[x_0, x_1](x - x_0) + \frac{1}{2}f^{(2)}(\xi_x)(x - x_0)(x - x_1),$$

missä $x_1 \in [a,\ b]$ on mielivaltainen. Funktion $x - x_0$ integraali välin $[a,\ b]$ yli häviää sillä x_0 on välin keskipiste:

$$\int_a^b f[x_0, x_1](x - x_0)dx = 0,$$

kaikilla $x_1 \in [a,\ b]$. Näin ollen keskipistekaavan virhe on

$$E_0(f) = \frac{1}{2}\int_a^b f^{(2)}(\xi_x)(x - x_0)(x - x_1)dx.$$

Valitaan nyt $x_1 = x_0$. Tällöin virhelauseke on

$$E_0(f) = \frac{1}{2}\int_a^b f^{(2)}(\xi_x)(x - x_0)^2dx.$$

Sovelletaan oikeanpuolen integraaliin integraalilaskennan väliarvolausetta:

Lause 6.1.1 *Olkoon $f(x)$ jatkuva funktio, ja $g(x)$ funktio, joka ei vaihda merkkiä integroimisvälillä $[a,\ b]$. Tällöin on olemassa piste $\zeta \in [a,\ b]$ siten, että*

$$\int_a^b f(x)g(x)dx = f(\zeta)\int_a^b g(x)dx.$$

Funktio $f^{(2)}(\xi_x)$ on oletettava jatkuvaksi. Mutta funktio $g(x) = (x - x_0)^2$ on koko välillä ei-negatiivinen, joten väliarvolauseen nojalla on $\zeta \in [a,\ b]$ siten, että

$$E_0(f) = \frac{1}{2}f^{(2)}(\zeta)\int_a^b (x - x_0)^2dx.$$

Näin ollen olemme johtaneet keskipistekaavan virhelausekkeen

Lause 6.1.2 *Olkoon funktio $f(x)$ kaksi kertaa jatkuvasti derivoituva funktio välillä $[a,\ b]$. Tällöin keskipistekaavan*

$$I_0(f) = (b-a)f(x_0)$$

virhe on

$$E_0(f) = \frac{1}{24}(b-a)^3 f^{(2)}(\zeta),$$

missä $\zeta \in [a,\ b]$.

Tavallisesti käytetään nk. summattua keskipistekaavaa. Jaetaan integroimisväli $[a,\ b]$ pistevieraisiin osaväleihin $[t_i, t_{i+1}],\ i = 0, \ldots, n-1$, missä

$$t_i = a + ih,\ h = \frac{b-a}{n}$$

ja n on osavälien lukumäärä. Integraali I voidaan jakaa summaksi

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} \int_{t_i}^{t_{i+1}} f(x)dx.$$

Sovelletaan jokaisella osavälillä keskipistekaavaa

$$\int_a^b f(x)dx = h\sum_{i=0}^{n} f(t_i + \frac{h}{2}) + \frac{h^3}{24}\sum_{i=0}^{n-1} f^{(2)}(\xi_i),$$

missä $\xi_i \in [t_i\ t_{i+1}]$.

Jos oletetaan, että integroitava funktio on kaksi kertaa jatkuvasti derivoituva, niin kaikilla $i = 0, \ldots, n-1$

$$\min_{x\in[a,\ b]} f^{(2)}(x) \le f^{(2)}(\xi_i) \le \max_{x\in[a,\ b]} f^{(2)}(x).$$

Näin ollen

$$n \min_{x\in[a,\ b]} f^{(2)}(x) \le \sum_{i=0}^{n-1} f^{(2)}(\xi_i) \le n \max_{x\in[a,\ b]} f^{(2)}(x).$$

Koska jatkuva funktio saa kaikki arvot suurimman ja pienimmä arvonsa väliltä, niin on olemassa $\xi \in [a,\ b]$ siten, että

$$\frac{1}{n}\sum_{i=0}^{n-1} f^{(2)}(\xi_i) = f^{(2)}(\xi).$$

Näin ollen voimassa:

Lause 6.1.3 *Kaksi kertaa jatkuvasti derivoituvan funktion integraali voidaan laskea summatulla keksipistekaavalla*

$$I_{0,n} = h\sum_{i=0}^{n} f(t_i + \frac{h}{2}),$$

jonka virhelauseke on

$$E_{0,n} = \frac{h^2(b-a)}{24} f^{(2)}(\xi).$$

Puolisuunnikassääntö

Interpoloidaan funktiota integroimisvälin päätepisteissä $x_0 = a$ ja $x_1 = b$. Funktio voidaan kirjoittaa silloin muodossa

$$f(x) = f(x_0) - f[x_0, x_1](x - x_0) + \frac{1}{2} f^{(2)}(\xi_x)(x - x_0)(x - x_1.$$

Integroimiskaava on tällöin

$$I_1 = \frac{b-a}{2}[f(a) + f(b)]$$

ja puolisuunnikassäännön virhe on

$$E_1(f) = \frac{1}{2}\int_a^b f^{(2)}(\xi_x)(x-a)(x-b)dx.$$

Väliarvolauseen nojalla, kun asetetaan $g(x) = (x-a)(x-b) \leq 0$ ja oletetaan funktio $f^{(2)}(\xi_x)$ jatkuvaksi, on $\zeta \in [a,\ b]$ siten, että

$$E_1(f) = \frac{1}{2} f^{(2)}(\zeta) \int_a^b (x-a)(x-b)dx.$$

Oikean puolen integraalin suoralla laskulla saadaan todistettua

Lause 6.1.4 *Olkoon funktio $f(x)$ kaksi kertaa jatkuvasti derivoituva funktio välillä $[a,\ b]$. Tällöin puolisuunnikassäännön*

$$I_1(f) = \frac{b-a}{2}[f(a) + f(b)]$$

virhe on

$$E_1(f) = -\frac{1}{12}(b-a)^3 f^{(2)}(\zeta),$$

missä $\zeta \in [a,\ b]$.

Kuten keskipistekaavaakin puolisuunnikassääntöä sovelletaan summatussa muodossa. Integroimisväli $[a, b]$ jaetaan pistevieraisiin osaväleihin

$$[t_i, t_{i+1}],\ i = 0, \ldots, n-1,$$

missä

$$t_i = a + ih,\ h = \frac{b-a}{n}$$

ja n on osavälien lukumäärä. Sovelletaan jokaisella osavälillä puolisuunnikassääntöä. Tällöin saadaan summattu puolisuunnikassääntö:

Lause 6.1.5 *Kaksi kertaa jatkuvasti derivoituvan funktion integraali voidaan laskea summatulla puolisuunnikassäännöllä*

$$I_{1,n} = \frac{h}{2}[f(a) + f(b) + 2\sum_{i=1}^{n-1} f(t_i)],$$

jonka virhelauseke on

$$E_{1,n} = -\frac{h^2(b-a)}{12} f^{(2)}(\xi),$$

missä $\xi \in [a,\ b]$ on väliarvolauseen nojalla määritelty piste.

Simpsonin sääntö

Interpoloidaan funktiota $f(x)$ pisteiden $x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b$ suhteen toisen asteen polynomilla:

$$P_2(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1).$$

Suoraviivaisesti polynomia $P_2(x)$ integroimalla saadaan Simpsonin sääntö:

$$I_2(f) = \frac{b-a}{6}[f(a) + 4f(\frac{a+b}{2}) + f(b)].$$

Polynomi-interpolaation virhelauseke voidaan kirjoittaa kuten keskipistekaavan tapauksessa muodossa

$$f(x) - P_2(x) = f[x_0, x_1, x_2](x - x_0)(x - x_1)(x - x_2) + \frac{f^{(4)}(\xi_x)}{4!}\Pi_{j=0}^{3}(x - x_j),$$

missä $x_3 = x_1$ on ylimääräinen piste. Oikean puolen kolmannen asteen polynomi on pariton välin $[x_0,\ x_2]$ keskipisteen suhteen. Joten sen integraali on nolla. Jäljelle jäävään virhetermiin voidaan soveltaa väliarvolausetta olettaen, että integroitava funktio on neljä kertaa jatkuvasti derivoituva. Näin ollen virhetermi on

$$E_2(f) = \frac{f^{(4)}(\zeta)}{4!}\int_a^b \Pi_{j=0}^3 (x - x_j)dx = -\frac{(b-a)^5}{2880}f^{(4)}(\zeta),$$

missä ζ on joku piste integroimisväliltä.

Summatussa Simpsonin säännössä valitaan aina parillinen määrä osavälejä, ts n=2m. Sovelletaan Simpsonin sääntöä osaväleillä $[t_{2i},\ t_{2i+2}]$, $i = 0,\ \ldots,\ m-1$. Tällöin on voimassa

Lause 6.1.6 *Neljästi jatkuvasti derivoituvalle funktiolle on voimassa summattu Simpsonin sääntö: (n=2m)*

$$I_{2,n} = \frac{h}{3}[f(a) + f(b) + 4\sum_{i=1}^{m} f(t_{2i-1}) + 2\sum_{i=1}^{m-1} f(t_{2i})],$$

jolle on voimassa virhelauseke

$$E_{2,n} = -\frac{h^4(b-a)}{180}f^{(4)}(\zeta)$$

jollain $\zeta \in [a,\ b]$.

6.1.2 Ekstrapolaatio

Euler-Maclaurinin summakaava

Olkoon funktiolla $f(x)$ välillä $[a,\ b]$ suppeneva Taylorin kehitelmä:

$$f(x+h) = f(x) + \sum_{k=1}^{\infty} \frac{(h\frac{d}{dx})^k}{k!}f(x).$$

Muodollisesti kehitelmä voidaan kirjoittaa differentiaalioperaattorin

$$e^{hD} = \sum_{k=0}^{\infty} \frac{(hD)^k}{k!},\ D = \frac{d}{dx}$$

avulla. Näin ollen

$$f(x+h) - f(x) = [e^{hD} - 1]f(x). \tag{6.1}$$

Funktion $\frac{x}{(e^x-1)}$ Taylorin kehitelmän avulla:

$$\frac{x}{e^x - 1} = \sum_{k=0}^{\infty} \frac{B_k}{k!} x^k.$$

Kehitelmässä olevat luvut B_k ovat Bernoullin lukuja:

$$\begin{aligned} B_0 &= 1,\ B_1 = -\frac{1}{2},\ B_2 = \frac{1}{6},\ B_3 = B_5 = B_7 = \cdots = 0, \\ B_4 &= -\frac{1}{30},\ B_6 = \frac{1}{42},\ B_8 = -\frac{1}{30}, \ldots \end{aligned}$$

Näin ollen

$$\begin{aligned} f(x) &= [e^{hD} - 1]^{-1}(f(x+h) - f(x)) \\ &= \frac{1}{h} D^{-1}[f(x+h) - f(x)] - \frac{1}{2}[f(x+h) - f(x)] \\ &+ \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} h^{2k-1}[f^{(2k-1)}(x+h) - f^{(2k-1)}(x)] \end{aligned} \tag{6.2}$$

Koska differentiaalioperaattorin käänteisoperaatio on integrointi, niin

$$\begin{aligned} D^{-1}[f(x+h) - f(x)] &= D^{-1}[\int_a^{x+h} f'(t)dt + \int_a^x f'(t)dt] \\ &= \int_a^{x+h} D^{-1} f'(t)dt - \int_a^x D^{-1} f'(t)dt \\ &= \int_a^{x+h} f(t)dt - \int_a^x f(t)dt = \int_x^{x+h} f(t)dt. \end{aligned}$$

Soveltamalla edellistä relaatiota ja kertomalla yhtälö (2) puolittain h:lla saadaan

$$\frac{h}{2}[f(x+h) + f(x)] = \int_x^{x+h} f(t)dt + \sum_{k=1}^{\infty} \frac{B_{2k}}{(2k)!} h^{2k}[f^{(2k-1)}(x+h) - f^{(2k-1)}(x)].$$

Soveltamalla edellistä kehitelmää summatun puolisuunnikassäännön jokaisella osavälillä saadaan *Euler-Maclaurinin summakaava*:

$$I_{1,n}(f) = \int_a^b f(x)dx + \sum_{k=1}^{\infty} c_k h^{2k}.$$

Rombergin menetelmä

Koska Euler-Maclaurinin summakaava on voimassa kaikille $n \in \mathbf{N}$, niin kaksinkertaistetaan osavälien lukumäärä. Tällöin

$$\begin{aligned} I_{1,2n} &= \int_a^b f(x)dx + \sum_{k=1}^{\infty} c_k (\frac{h}{2})^{2k} \\ &= I + c_2\frac{h^2}{4} + c_4\frac{h^4}{16} + c_6\frac{h^6}{64} + \dots \end{aligned}$$

Näin ollen saadaan integraalin approksimaatio

$$\frac{4I_{1,2n} - I_{1,n}}{3} = I + c_2'h^4 + c_4'h^6 + c_3'h^8 + \dots,$$

jonka virheen asymptoottisen kehitelmän johtava termi on h:n neljäs potenssi.

Merkitään jatkossa integraalin approksimaation arvoa summatulla puolisuunnikassäännöllä laskettuna

$$R_{k,1} = I_{1,2^{k-1}},\ k = 1, 2, 3, 4, \dots.$$

Kun indeksi kasvaa $k-1 \to k$, niin osavälin pituus puoliintuu: $h \to \frac{h}{2}$.

Siten kun $k \geq 2$ on integraalille $I = \int_a^b f(x)dx$ kaksi asymptoottista kehitelmää:

$$\begin{aligned} I &= R_{k-1,1} + c_2h^2 + c_4h^4 + c_6h^6 + c_8h^8 + \cdots \\ I &= R_{k,1} + c_2\frac{h^2}{4} + c_4\frac{h^4}{16} + c_6\frac{h^6}{64} + c_8\frac{h^8}{256} + \cdots. \end{aligned}$$

Eliminoidaan alemmasta kehitelmästä h:n toinen potenssi seuraavasti:

$$I = \frac{4I - I}{3} = \frac{4R_{k,1} - R_{k-1,1}}{3} + c_4h^4 + c_6h^6 + c_8h^8 + \cdots$$

Tässä asymptoottisessa kehitelmässä olemme käyttäneet geneerisiä vakioita c_{2m}, jotka siis muuttuvat arvoltaan eri kehitelmissä. Olennaista on, että ne ovat vakioita ja samoja eri h:n arvoille. Lisäksi edellä mainitussa ekstrapolaatioaskeleessa vakioiden arvot pienenevät.

Merkitään jatkossa jokaiselle $k \geq 2$:

$$R_{k,2} = \frac{4R_{k,1} - R_{k-1,1}}{3},$$

jolle on voimassa edellisen nojalla kehitelmä

$$I = R_{k,2} + c_4 h^4 + c_6 h^6 + c_8 h^8 + \cdots .$$

Puolittamalla osavälien pituus saadaan kehitelmät

$$\begin{aligned} I &= R_{k-1,2} + c_4 h^4 + c_6 h^6 + c_8 h^8 + \cdots \\ I &= R_{k,2} + c_4 \frac{h^4}{16} + c_6 \frac{h^6}{64} + c_8 \frac{h^8}{256} + \cdots \end{aligned}$$

jokaiselle $k \geq 3$.

Suoritetaan uudelleen ekstrapolaatio kertomalla alempi yhtälö luvulla 16 ja vähentämällä siitä ylempi yhtälö, saadaan integraalin approksimaatio

$$R_{k,3} = \frac{4^2 R_{k,2} - R_{k-1,2}}{4^2 - 1},$$

jolle on voimassa virhekehitelmä

$$I = R_{k,3} + c_6 h^6 + c_8 h^8 + c_{10} h^{10} + \cdots .$$

Näin jatkamalla induktiivisesti voidaan johtaa Rombergin menetelmän yleinen askel:

$$\begin{aligned} R_{k,1} &= I_{1,2^{k-1}} \\ R_{k,j} &= \frac{4^{j-1} R_{k,j-1} - R_{k-1,j-1}}{4^{j-1} - 1}, \ k \geq j \geq 2. \end{aligned}$$

Näin ollen termin $R_{n,n}$ virhekehitelmän johtava termi on

$$O((\frac{1}{2})^{2n-2}).$$

Käytännön laskennassa virhettä kontrolloidaan seuraavalla säännöllä:

Lause 6.1.7 (sääntö) *Jos* $|R_{n,n} - R_{n-1,n-1}| < \frac{1}{2}10^{-d}$, *niin approksimaatiossa* $R_{n,n}$ *on* d *oikeata desimaalia.*

Edellinen väittämä tietysti edellyttää, että integroitava funktio on sileä funktio.

Singulariteettien käsittely Usein sovellutuksissa on integroitava funktioita, jotka ovat singulaarisia integroimisvälin päätepisteissä. Esimerkiksi funktio

$$f(x) = \frac{1}{\sqrt{x}}$$

kasvaa rajatta, kun x lähestyy origoa. Silti integraali $\int_0^1 f(x)dx$ on olemassa. Kuitenkaan Simpsonin tai puolisuunnikassääntöä ei voida käyttää integraalin numeerisen approksimaation laskemiseen.

Kerrotaan funktio $f(x)$ singulariteetin käänteisfunktiolla:

$$g(x) = \sqrt{x}f(x)(= 1).$$

Tämä funktio on hyvin säännöllinen. Näin ollen funktion $f(x)$ integraali voidaan kirjoittaa muodossa

$$\int_0^1 g(x)x^{-\frac{1}{2}}dx.$$

Muodostetaan sellainen integroimiskaava, joka integroi tarkasti muotoa $x^k\sqrt{x}$ olevia funktioita ennalta määrättyyn potenssiin n asti. Valitaan integroimiskaavalle sopivat tukipisteet $x_0, x_1, \ldots, x_m$. Näin ollen on valittava kertoimet $A_0, A_1, \ldots, A_m$ siten, että

$$\int_0^1 x^k\sqrt{x}dx = A_0x_0^k + A_1x_1^k + \cdots + A_mx_m^k.$$

Esimerkki 6.1 *Määrää integroimiskaava, joka integroi tarkasti ensimmäisen asteen polynomeja pisteiden* $x_0 = \frac{1}{4},\ x_1 = \frac{3}{4}$ *suhteen.*

Ratkaisu: Ratkaistaan kertoimet $A_0,\ A_1$ yhtälöparista

$$\begin{aligned} A_0 + A_1 &= \int_0^1 1 \cdot x^{-\frac{1}{2}}dx = 2 \\ \frac{1}{4}A_0 + \frac{3}{4}A_1 &= \int_0^1 x \cdot x^{-\frac{1}{2}}dx = \frac{2}{3}. \end{aligned}$$

Yhtälöparin ratkaisu on $A_0 = \frac{5}{3}$ ja $A_1 = \frac{1}{3}$. Näin ollen integroimiskaavaksi saadaan

$$\int_0^1 g(x)x^{-\frac{1}{2}}dx = \frac{5}{3}g(\frac{1}{4}) + \frac{1}{3}g(\frac{3}{4}).$$

6.1.3 Gaussin kvadratuurit ja ortogonaaliset polynomit

Gaussin kvadratuurin ideana on samanaikaisesti määrätä integroimiskaavan painokertoimet $\omega_i,\ i = 0, 2, \ldots, n$ ja tukipisteet $x_i,\ i = 0, \ldots, n$ siten, että integroimiskaava

$$\omega_0 f(x_0) + \cdots \omega_n f(x_n)$$

integroi tarkasti mahdollisimman korkea-asteisia polynomeja.

Interpolaatiokaavoissa valitaan *tukipisteet* $x_0, x_1, \ldots, x_n$, joiden suhteen integroitavaa funktiota interpoloidaan. Vastaava interpolaatiopolynomi on

$$F_n(x) = \sum_{k=0}^{n} f(x_k) L_k(x),$$

kun käytetään polynomille Lagrangen esitystä. Tällöin numeerinen integroimiskaava on

$$\int_a^b f(x)dx = \sum_{k=0}^{n} \omega_k f(x_k) + E_n(f),$$

missä $E_n(f)$ on integroimiskaavan virhe ja painokertoimet

$$\omega_k = \int_a^b L_k(x)dx.$$

Koska interpolaation virhe on verrannollinen funktion $f(x)$ n+1:seen derivaattaan, niin ilmeisesti integroimiskaava integroi tarkasti ainakin n-asteisia polynomeja, koska sillon $f^{(n+1)}(\xi) \equiv 0$.

Toisaalta interpolaatiokaava integroi tarkasti korkeintaan 2n+1-asteisia polynomeja. Nimittäin funktio

$$f(x) = [\Pi_{k=0}^{n}(x - x_k)]^2$$

on 2n+2-asteinen polynomi, jolle integraalin arvo

$$\int_a^b f(x)dx > 0,$$

sillä funktio on positiivinen funktio välillä $[a,\ b]$. Toisaalta kvadratuurin arvo on

$$\sum_{k=0}^{n} \omega_k f(x_k) = 0,$$

sillä interpolaatiopisteissä on polynomin nollakohdat.

Kysymys jatkossa kuuluu, voidaanko interpolaatiopisteet valita siten, että integrointikaava integroi tarkasti 2n+1-asteisia polynomeja. Vastaus kysymykseen on myönteinen. Tällaisen kvadratuurin konstruktio perustuu ortogonaalisten polynomien käyttöön. Sitä varten tarvitaan interpolaatiopolynomille vaihtoehtoinen virhelauseke.

Interpolaatiovirhe Vaihtoehtoinen virhetermi perustuu interpolaatiopolynomin Newtonin esitykseen.

Olkoon Newtonin interpolaatiopolynomi $F_k(x),\ k = 0, \ldots, n$. Tällöin

$$F_k(x) = F_{k-1}(x) + a_k \Pi_{j=0}^{k-1}(x - x_j).$$

Koska

$$F_k(x_k) = f(x_k) = F_{k-1}(x_k) + a_k \Pi_{j=0}^{k-1}(x_k - x_j),$$

niin k:s jaettu erotus voidaan lausua muodossa

$$a_k = \frac{f(x_k) - F_{k-1}(x_k)}{\Pi_{j=0}^{k-1}(x_k - x_j)} = f[x_0, \ldots, x_k].$$

Tämän nojalla interpoloitaessa n+2:n pisteen $x_0, \ldots, x_n, x$ suhteen interpolaatiopolynomin kerroin

$$a_{n+1} = \frac{f(x) - F_n(x)}{\Pi_{j=0}^{n}(x - x_j)} = f[x_0, \ldots, x_n, x].$$

Siten interpolaatiopolynomin virheelle saadaan lauseke

$$f(x) - F_n(x) = f[x_0, \ldots, x_n, x] \Pi_{j=0}^{n}(x - x_j).$$

Merkitään jatkossa

$$q(x) = \Pi_{j=0}^{n}(x - x_j).$$

Näin ollen integroimiskaavan virheen lauseke on

$$E_n(f) = \int_a^b q(x) f[x_0, \ldots, x_n, x] dx.$$

Edellä esitetty tarkastelu on riippumaton interpolaatiopisteiden valinnasta. Seuraavaksi valitaan pisteet $x_o, \ldots, x_n$ siten, että $E_n(f) = 0$ kaikille n+m-astetta oleville polynomeille, missä m on mahdollisimman suuri.

Jos $f(x)$ on astetta n+m+1 oleva polynomi, niin jaettu erotus

$$f[x_0, \ldots, x_n, x]$$

on m-asteinen polynomi. Näin ollen interpolaatiovirhe

$$E_n(f) = 0$$

kaikille polynomeille, joiden asteluku on korkeintaan n+m+1, jos ja vain jos

$$\int_a^b q(x)x^r dx = 0, \ \forall r = 0, \ldots, m.$$

Huomaa, että polynomin $q(x)$ asteluku on n+1.

Ortogonaalinen polynomi p_{n+1} on kohtisuorassa kaikkia alempi asteisia polynomeja vastaan (Lauseen 3.3.5 nojalla). Näin ollen optimaalinen valinta interpolaatiopisteiksi ovat ortogonaalisen polynomin $p_{n+1}(x)$ nollakohdat. Tällöin myös $q(x) = p_{n+1}(x)$ ja m=n. Näin saimme todistettua seuraaan lauseen.

Lause 6.1.8 *Integroimiskaava on tarkka 2n+1-asteisille polynomeille, jos interpolaatiokaavan tukipisteiksi valaitaan ortogonaalisen polynomin $p_{n+1}(x)$ nollakohdat ja painokertoimiksi*

$$\omega_k = \int_a^b L_k(x)dx, \ k = 0, \ldots, n.$$

Gauss-Legendren kaavat Perustuvat Legendre'n polynomien nollakohtien käyttöön. Legendre'n polynomit toteuttavat seuraavan kolmen termin rekursiokaavan Perusvälillä $[-1, 1]$:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ (n+1)P_{n+1}(x) &= (2n+1)xP_n(x) - nP_{n-1}(x), \ n = 1, 2, \ldots \ . \end{aligned}$$

Toisen asteen Legendren polynomi on

$$p_2(x) = \frac{3}{2}x^2 - \frac{1}{2}.$$

Silloin toisen asteen Legendre'n polynomin nollakohdat ovat

$$x_0 = \frac{1}{\sqrt{3}},\ x_1 = -\frac{1}{\sqrt{3}}.$$

Pisteeseen x_0 liittyvä integroimiskaavan painokerroin on

$$\omega_1 = \int_{-1}^{1} L_1(x)dx = \int_{-1}^{1} \frac{\sqrt{3}}{2}x + \frac{1}{2}dx = 1.$$

Symmetrian nojalla ilmeisesti $\omega_1 = \omega_2$. Kahden pisteen Gaussin kaava:

$$Q_2(f) = f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}).$$

Integraali $I = \int_a^b f(x)dx$ voidaan palauttaa perusmuotoon integroimismuuttujan vaihdolla

$$x = \frac{b-a}{2}t + \frac{a+b}{2},\ -1 \leq t \leq 1.$$

Tällöin integraali on

$$I = \frac{b-a}{2}\int_{-1} 1f(\frac{b-a}{2}t + \frac{a+b}{2})dt,$$

jonka approksimaatio kahden pisteen Gaussin kaavalla on siten

$$Q_2(f) = \frac{b-a}{2}[f(-\frac{b-a}{2\sqrt{3}} + \frac{a+b}{2}) + f(\frac{b-a}{2\sqrt{3}} + \frac{a+b}{2})].$$

Gauss-Legendren kaavojen virhe on

$$E_n(f) = \frac{2}{(2n+1)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2 f^{(2n)}(\xi),\ -1 < \xi < 1.$$

6.2 Numeerinen derivointi

1. kertaluvun derivaatan approksimaatio Peruskurssien nojalla suoraan derivaatan määritelmästä saadaan yksinkertaisin derivointikaava ns. *eteenpäin differenssikaava*:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + O(h).$$

Näin ollen derivaatan approksimaatio on funktion arvojen lineaarikombinaatio pisteen x ympäristössä, kun diskretisointiparametri h oletetaan riittävän pieneksi.

Keskeisdifferenssikaava Muodostetaan funktion arvojen $f(x+h)$ ja $f(x-h)$ Taylorin kehitelmät pisteen x ympäristössä

$$\begin{aligned} 1: \ f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f^{(3)}(\xi_+) \\ 0: \ f(x) &= f(x) \\ -1: \ f(x-h) &= f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f^{(3)}(\xi_-) \end{aligned}$$

Ratkaistaan yhtälöryhmästä 1. kertaluvun derivaatta kertomalla kolmas kehitelmä luvulla -1 ja lasketaan ensimmäinen ja viimeinen yhtälö keskenään yhteen. Näin me saamme *keskeisdifferenssikaavan*

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}\frac{f^{(3)}(\xi_+) + f^{(3)}(\xi_-)}{2}.$$

Jos oletetaan, että derivoitavan funktion kolmas derivaatta on jatkuva, niin differentiaalilaskennan väliarvolauseen nojalla on olemassa $\xi \in [x-h, x+h]$ siten, että

$$f^{(3)}(\xi) = \frac{f^{(3)}(\xi_+) + f^{(3)}(\xi_-)}{2}.$$

Joten keskeisdifferenssikaava voidaan kirjoittaa muodossa

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}f^{(3)}(\xi).$$

Keskeisdifferenssikaava on huomattavasti tarkempi kuin eteenpäin differenssikaava, sillä virhetermi on verrannollinen diskretisointiparametrin toiseen potenssiin.

Päätepistekaavoissa muodostetaan funktion $f(x)$ Taylorin kehitelmät pisteissä $x+h$ ja $x+2h$:

$$\begin{array}{l} -3: \ f(x) = f(x) \\ +4: \ f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f^{(3)}(\xi_1) \\ -1: \ f(x+2h) = f(x) + 2hf'(x) + 2h^2f''(x) + \frac{8h^3}{6}f^{(3)}(\xi_2) \\ \hline -f(x+2h) + 4f(x+h) - 3f(x) = 2hf'(x) + \frac{2h^3}{3}f'''(\xi_1) - \frac{4h^3}{3}f'''(\xi_2) \end{array}$$

Näin ollen päätepistekaavaksi saadaan

$$f'(x) = \frac{-3f(x) + 4f(x+h) - f(x+2h)}{2h} + \frac{h^2}{3}f^{(3)}(\xi),$$

missä piste ξ on valittu siten, että

$$f^{(3)}(\xi) = 2f'''(\xi_2) - f'''(\xi_1).$$

Tämä on mahdollista väliarvolauseen nojalla, jos funktion kolmas derivaatta on jatkuva.

Ensimmäisen kertaluvun derivaatalle voidaan johtaa myös korkeamman kertaluvun (lue: tarkempia) derivointikaavoja; mutta niissä tarvitaan funktion arvot useammassa kuin kolmessa pisteessä.

Toisen kertaluvun derivaatta Samalla tavalla kuin edellä ratkaistaan Taylorin kehitelmistä

$$\begin{aligned}
1: &\ f(x+h) = f(x) + hf'(x) + \tfrac{h^2}{2}f''(x) + \tfrac{h^3}{6}f^{(3)}(x) + \tfrac{h^4}{24}f^{(4)}(\xi_+) \\
-2: &\ f(x) = f(x) \\
1: &\ f(x-h) = f(x) - hf'(x) + \tfrac{h^2}{2}f''(x) - \tfrac{h^3}{6}f^{(3)}(x) + \tfrac{h^4}{24}f^{(4)}(\xi_-)
\end{aligned}$$

toisen derivaatan lauseke yhtälöryhmän oikealta puolelta. Tällöin saadaan

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi),$$

missä $\xi \in [x-h, x+h]$. Sen olemassaolo riippuu neljännen derivaatan jatkuvuudesta. Tällöin kuten 1. kertaluvun derivaattojen virhearvioiden johtamisessa sovelletaan differentiaalilaskennan väliarvolausetta.

Numeerisen derivoinnin stabiilisuus: Kaikki numeeriset derivointikaavat ovat herkkiä pyöristysvirheille. Olkoon f_{-1} ja f_1 funktion $f(x)$ pyöristetyt arvot pisteissä x_{-1} ja x_1:

$$f_i - f(x_i) \approx \frac{1}{2}10^{-k},\ i = -1, 1.$$

Tällöin keskeisdifferenssikaavan todellinen virhe

$$f'(x_0) - \frac{f_1 - f_{-1}}{2h} = E_R - \frac{1}{6}h^2 f^{(3)}(\xi),$$

missä

$$E_R = \frac{f(x_1) - f_1}{2h} + \frac{f_{-1} - f(x_{-1})}{2h}.$$

Tällöin todellisen virheen yläraja on pienempi kuin

$$\frac{1}{2h}10^{-k} + \frac{1}{6}h^2 \max|f^{(3)}(\xi)| \equiv E(h).$$

Kiinteällä desimaalitarkkuudella k kokonaisvirheen $E(h)$ maksimi kasvaa, kun hilapisteiden välinen erotus h lähestyy nollaa. Toisin sanoen pyöristysvirheet alkavat dominoimaan laskentaa.

Merkitään

$$M_3 = \max_{\xi\in[x_{-1},x_1]} |f^{(3)}(\xi)|.$$

Tällöin numeerisen derivoinnin virhettä voidaan arvioida ylöspäin kuten

$$|D_h f(x_0) - f'(x_0)| \leq \frac{\epsilon}{2h} + \frac{M_3}{6}h^2,$$

missä $\epsilon = 10^{-k}$. Kokonaisvirheen minimi saavutetaan, kun $\frac{\epsilon}{2h} = \frac{M_3}{6}h^2$. Optimaalinen diskretisointiparametri numeeriselle derivoinnille on silloin

$$h = \{\frac{3\epsilon}{M_3}\}^{\frac{1}{3}}.$$

Tällä diskretisoinnilla virheen yläraja on

$$\frac{\epsilon}{2}\{\frac{M_3}{3\epsilon}\}^{\frac{1}{3}} = \{\frac{\epsilon^2 M_3}{24}\}^{\frac{1}{3}}.$$

Tarkastellaan yksinkertaisen esimerkin valossa numeerisen derivoinnin stabiilisuutta:

Esimerkki 6.2 *Laske MATLABilla derivaatan approksimaatio funktiolle*

$$f(x) = \sin(x)$$

pisteessä $x = \frac{\pi}{3.2}$ *eteenpäin differenssikaavalla diskretisointiparametreillä* $h = 10^{-1}, \ldots, 10^{-14}$.

Ratkaisu: MATLAB-koodi on

```
for   i = 1 : 14
      x(i) = (sin(pi/3.2 + 10^-i) - sin(pi/3.2))/10^(-i);
      y(i) = x(i) - cos(pi/3.2);
end
```

Laskennan tulos on annettu seuraavassa taulukossa:

$D_h f(x)$	Virhe
0.51310589790214	-0.04246433511746
0.55140366014496	-0.00416657287465
0.55515440565301	-0.00041582736659
0.55552865861230	-0.00004157440730
0.55556607565510	-0.00000415736450
0.55556981726212	-0.00000041575748
0.55557019096319	-0.00000004205641
0.55557023426189	-0.00000000124229
0.55557025646635	-0.00000002344675
0.55557003442175	-0.00000019859785
0.55556670375267	-0.00000352926693
0.55555560152243	-0.00001463149717

Tuloksista havaitaan, että derivointitarkkuus paranee, kun diskretisointiparametri pienenee 10^{-8} asti. Tämän jälkeen virhe kasvaa vaikka pisteiden välinen etäisyys pienenee.

Luku 7

Alkuarvotehtävien numeerinen ratkaisu

7.1 Tavalliset 1. kertaluvun yhtälöt

7.1.1 Johdatus aiheeseen

Tässä luvussa tarkastellaan differentiaaliyhtälön alkuarvotehtävää

$$\begin{aligned} \frac{dy}{dt} &= f(t,y),\ a \le t \le b \\ y(t_0) &= y_0,\ a \le t_0 \le b \end{aligned} \tag{7.1}$$

tai yleisemmin 1. kertaluvun differentiaaliyhtälösysteemin alkuarvotehtävää

$$\begin{aligned} y_1'(t) &= f_1(t, y_1, \ldots, y_n) \\ y_2'(t) &= f_2(t, y_1, \ldots, y_n) \\ &\vdots \\ y_n'(t) &= f_n(t, y_1, \ldots, y_n) \end{aligned} \tag{7.2}$$

alkuehdoin

$$\begin{aligned} y_1(a) &= \alpha_1 \\ &\vdots \\ y_n(a) &= \alpha_n \end{aligned} \tag{7.3}$$

Lause 7.1.1 *Olkoon f ja $\frac{\partial f}{\partial y}$ jatkuvia välillä $[a,\ b]$. Silloin differentiaaliyhtälön alkuarvotehtävällä on yksikäsitteinen ratkaisu välillä $[a,\ b]$.*

7.1.2 Taylorin menetelmä

Määritellään välillä $[a,\ b]$ pistejoukko $\{t_0, t_1, \ldots, t_n\}$, missä

$$t_i = a + ih,\ i = 0, ..., n$$

ja

$$h = \frac{b-a}{n}.$$

Oletus 7.1.1 *Ratkaisun toinen derivaatta $y''(t)$ on jatkuva funktio suljetulla välillä $[a, b]$.*

Tällöin Taylorin lauseen nojalla

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i),$$

jollain $\xi_i \in [t_i, t_{i+1}]$. Koska $h = t_{i+1} - t_i$, niin

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(\xi_i).$$

Edelleen koska $y(t)$ on alkuarvotehtävän ratkaisu, niin

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i).$$

Merkitään funktion $y(t)$ approksimaatiota pisteessä t_i y_i:llä. Eulerin menetelmässä "unohdetaan"Taylorin kehitelmän jäännöstermi $\frac{1}{2}h^2y''(\xi_i)$.

Eulerin menetelmä

$$\begin{aligned} y_0 &= \alpha \\ y_{i+1} &= y_i + hf(t_i, y_i), \end{aligned} \tag{7.4}$$

missä $i = 0, \ldots, n-1$. Menetelmän lokaalivirhe on $\frac{1}{2}h^2y''(\xi_i)$.

Eulerin menetelmän virhe Olkoon $y(t)$ alkuarvotehtävän

$$\begin{aligned} \frac{dy}{dt} &= f(t,y), a \leq t \leq b \quad (7.5) \\ y(t_0) &= y_0 \quad (7.6) \end{aligned}$$

ratkaisu ja $y_0, y_1, \ldots, y_n$ Eulerin menetelmällä generoidut approksimaatiot.

Lause 7.1.2 *Jos funktio $f(t,y)$ on jatkuva kaikilla $t \in [a,\ b]$, $y \in]-\infty,\ \infty[$ ja on olemassa vakiot $L,\ M$ siten, että*

$$|\frac{\partial f(t,y(t))}{\partial y}| \leq L,\ |y''(t)| \leq M,$$

silloin jokaiselle $i = 0, 1, \ldots, n$

$$|y(t_i) - y_i| \leq \frac{hM}{2L}[e^{L(t_i-a)} - 1].$$

N:nnen kertaluvun Taylorin menetelmä

Olettaen, että alkuarvotehtävän ratkaisun $(n+1)$:n derivaatta $y^{(n+1)}(t)$ on jatkuva, niin Taylorin kehitelmä pisteen t_i suhteen on

$$\begin{aligned} y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y^{(2)}(t_i) + \ldots \\ + \frac{h^n}{n!}y^{(n)}(t_i) + \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi_i),\ \xi_{\in}[t_i,\ t_{i+1}] \quad (7.7) \end{aligned}$$

Olettamalla jäännöstermi pieneksi saadaan ratkaisun approksimaatioille laskentakaava

$$\begin{aligned} y_0 &= \alpha \quad (7.8) \\ y_{i+1} &= y_i + hT_n(t_i, y_i), \end{aligned}$$

missä

$$T_n(t_i, y_i) = f(t,y_i) + \frac{h}{2}f'(t_i, y_i) + \cdots + \frac{h^{n-1}}{n!}f^{(n-1)}(t_i, y_i)$$

ja menetelmän lokaalivirhe on tällöin

$$\frac{1}{(n+1)!}y^{(n+1)}(\xi_i)h^{n+1}.$$

7.2 Runge-Kutta menetelmät

Differentiaaliyhtälön alkuarvotehtävä voidaan kirjoittaa välillä $[t_i,\ t_{i+1}]$ muodossa

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(t)dt = \int_{t_i}^{t_{i+1}} f(t, y(t))dt.$$

Muodostetaan seuraavaksi integrointikaava, joka integroi oikean puolen integraalin mahdollisimman tarkasti. Yleisesti valitaan väliltä $[t_i, t_{i+1}]$ pisteet $\xi_1, \ldots, \xi_n$ ja painokertoimet $c_i,\ i = 1, \ldots, n$ siten, että lokaalin virheen

$$|y(t_{i+1}) - y(t_i) - h\sum_{i=1}^{n} c_i f(\xi_i, y(\xi_i))| = O(h^p)$$

kertaluku p on mahdollisimman suuri.

7.2.1 2-vaiheinen Runge-Kutta menetelmä

Integroimispisteet: $\xi_1 = t_i,\ \xi_2 = t_i + ah$.

Funktion $f(t, y)$ arvot kyseisissä pisteissä:

$$\begin{aligned} f(\xi_1, y(\xi_1) &= f(t_i, y(t_i)) \\ f(\xi_2, y(\xi_2) &= f(t_i + ah, y(t_i + ah)) \end{aligned}$$

Ongelma: Mikä on ratkaisun arvo pisteessä $\xi = t_i + ah$?

Valitaan lisäparametri b siten, että

$$y(t_i + ah) = y_i + bhf(t_i, y(t_i)).$$

Tällöin kaksivaiheinen Runge-Kutta menetelmä on seuraavanlainen:

$$\begin{aligned} k_1 &= f(t_i, y(t_i)); \\ k_2 &= f(t_i + ah, y(t_i) + bhf(t_i, y(t_i))); \\ y_{i+1} &= y_i + h(c_1k_1 + c_2k_2). \end{aligned}$$

Oletetaan, että $y_i = y(t_i)$ ja määrätään parametrit a, b, c_1, c_2 siten, että ratkaisumenetelmän lokaalivirhe

$$d_i = y_{i+1} - y(t_{i+1}) = O(h^p)$$

on mahdollisimman korkeata astetta.

Sovelletaan funktioon $f(t_i, y_i + bhf(t_i, y_i))$ Taylorin kehitelmää

$$f(t_i + ah, y_i + bhf(t_i, y_i)) = f(t_i, y_i) + ah\frac{\partial f}{\partial t}(t_i, y_i) + hbk_1\frac{\partial f}{\partial y}(t_i, y_i) + O(h^2).$$

Näin ollen ($k_1 = f(t_i, y_i)$)

$$y_{i+1} = y_i + (hc_1 + hc_2)f(t_i, y_i) + ac_2h^2\frac{\partial f}{\partial t}(t_i, y_i) + bc_2h^2 f\frac{\partial f}{\partial y}(t_i, y_i) + O(h^3).$$

Toisaalta funktion $y(t)$ Taylorin kehitelmä pisteessä t_{i+1} on

$$y(t_{i+1}) = y_i + hf(t_i, y_i) + \frac{1}{2}h^2[\frac{\partial f}{\partial t}(t_i, y_i) + f\frac{\partial f}{\partial y}(t_i, y_i)] + O(h^3).$$

Lokaalivirhe on tällöin

$$\begin{aligned} d_{i+1} &= y(t_{i+1}) - y_{i+1} \\ &= h(1 - c_1 - c_2)f(t_i, y_i) + (\frac{1}{2} - ac_2)h^2\frac{\partial f}{\partial t}(t_i, y_i) \\ &\quad +(\frac{1}{2} - bc_2)h^2 f\frac{\partial f}{\partial y}(t_i, y_i) + O(h^3) \end{aligned}$$

Nyt lokaalivirhe on kertalukua $O(h^3)$, mikäli

$$\begin{aligned} c_1 + c_2 &= 1 \\ ac_2 &= \frac{1}{2} \\ bc_2 &= \frac{1}{2}. \end{aligned}$$

Yhtälöryhmällä on useita eri ratkaisuja; mutta vakiintuneimmat valinnat ovat seuraavanlaiset:

Modifioitu Eulerin menetelmä

$$\begin{aligned} k_1 &= f(t_i, y(t_i)); \\ k_2 &= f(t_i + \frac{1}{2}h, y(t_i) + \frac{1}{2}hf(t_i, y(t_i))); \\ y_{i+1} = y_i &+ hk_2. \end{aligned}$$

Yksinkertainen Runge-Kutta menetelmä

$$\begin{aligned} k_1 &= f(t_i, y(t_i)); \\ k_2 &= f(t_i + h, y(t_i) + hf(t_i, y(t_i))); \\ y_{i+1} = y_i &+ \frac{1}{2}h(k_1 + k_2). \end{aligned}$$

7.2.2 Kolmivaiheinen Runge-Kutta-menetelmä

Valitaan

$$\xi_1 = t_k,\ \xi_2 = t_k + a_2 h,\ \xi_3 = t_k + a_3 h,\ 0 < a_2, a_3 < 1.$$

Ennustus:

$$\begin{aligned} y(\xi_2) : y_2 &= y_k + hb_{2,1} f(t_k, y_k) \\ y(\xi_3) : y_3 &= y_k + hb_{3,1} f(t_k, y_k) + hb_{3,2} f(t_k + a_2 h, y_k) \end{aligned} \tag{7.9}$$

Eksplisiittinen kolmivaiheinen yksiaskelmenetelmä

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + a_2 h, y_k + hb_{2,1} k_1) \\ k_3 &= f(t_k + a_3 h, y_k + hb_{3,1} k_1 + hb_{3,2} k_2) \\ y_{k+1} &= y_k + h(c_1 k_1 + c_2 k_2 + c_3 k_3). \end{aligned} \tag{7.10}$$

Parametrit $a_2, a_3, b_{2,1}, b_{3,1}, b_{3,2}, c_1, c_2, c_3$ valitaan siten, että menetelmän virheen kertaluku on mahdollisimman korkea. Menetelmän tulee olla tarkka esimerkiksi alkuarvotehtävälle

$$y'(t) = 1.$$

Tästä ehdosta saadaan lisäehdot

$$a_2 = b_{2,1},\ a_3 = b_{3,1} + b_{3,2}.$$

Menetelmän lokaalivirhe on siis

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h(c_1 \tilde{k}_1 + c_2 \tilde{k}_2 + c_3 \tilde{k}_3),$$

missä

$$\begin{aligned} \tilde{k}_1 &= f(t_k, y(t_k)) = f \\ \tilde{k}_2 &= f(t_k + a_2 h, y(t_k) + a_2 h f(t_k, y(t_k))) \\ \tilde{k}_3 &= f(t_k + a_3 h, y(t_k) + h(b_{3,1} \tilde{k}_1 + b_{3,2} \tilde{k}_2)). \end{aligned} \tag{7.11}$$

Merkintöjä:

$$\begin{aligned} F &= \frac{\partial f}{\partial t} + f\frac{\partial f}{\partial y} \\ G &= \frac{\partial^2 f}{\partial t^2} + 2f\frac{\partial^2 f}{\partial t \partial y} + f^2\frac{\partial^2 f}{\partial y^2} \end{aligned} \tag{7.12}$$

Kaksiulotteisen Taylorin kehitelmän nojalla pisteen $(t_k, y(t_k))$ suhteen saadaan

$$\begin{aligned} \tilde{k}_2 &= f + a_2 hF + \frac{1}{2}a_2^2 h^2 G + O(h^2) \\ \tilde{k}_3 &= f + a_3 hF + h^2[a_2 b_{3,2} F\frac{\partial f}{\partial y} + \frac{1}{2}a_3^2 G] + O(h^3). \end{aligned} \tag{7.13}$$

Tällöin lokaalivirhe

$$\begin{aligned} d_{k+1} &= hf[1 - c_1 - c_2 - c_3] + h^2 F[\frac{1}{2} - a_2 c_2 - a_3 c_3] + \\ &\quad + h^3\{F\frac{\partial f}{\partial y}[\frac{1}{6} - a_2 c_3 b_{3,2}] + G[\frac{1}{6} - \frac{1}{2}a_2^2 c_2 - \frac{1}{2}a_3^2 c_3]\} + \\ &\quad + O(h^4) \end{aligned} \tag{7.14}$$

Lokaalivirhe on verrannollinen askelpituuden neljänteen potenssiin mikäli

$$\begin{aligned} c_1 + c_2 + c_3 &= 1 \\ a_2 c_2 + a_3 c_3 &= \frac{1}{2} \\ a_2^2 c_2 + a_3^2 c_3 &= \frac{1}{3} \\ a_2 c_3 b_{3,2} &= \frac{1}{6} \end{aligned} \tag{7.15}$$

Parametrit c_1, c_2, c_3 voidaan ratkaista kolmesta ensimmäisestä yhtälöstä (lineaarisella yhtälöryhmällä on yksikäsitteinen ratkaisu, mikäli kerroinmatriisi on säännöllinen ts. $a_2 \neq a_3, a_2 \neq 0, a_3 \neq 0$). Yhtälöryhmän ratkaisu on

$$\begin{aligned} c_1 &= \frac{6a_2 a_3 + 2 - 3(a_3 + a_2)}{6a_3 a_2} \\ c_2 &= \frac{3a_3 - 2}{6a_2(a_3 - a_2)} \\ c_3 &= \frac{2 - 3a_2}{6a_3(a_3 - a_2)} \end{aligned} \tag{7.16}$$

Lopulta neljännestä yhtälöstä voidaan $b_{3,2}$ ratkaista mikäli $a_2 \neq \frac{2}{3}$, ja

$$\begin{aligned} b_{3,2} &= \frac{a_3(a_3 - a_2)}{a_2(2 - 3a_2)} \\ b_{3,1} &= a_3 - \frac{a_3(a_3 - a_2)}{a_2(2 - 3a_2)} \\ b_{2,1} &= a_2 \end{aligned}$$

Näin olemme konstruoineet äärettömän määrän alkuarvotehtävän ratkaisumenetelmiä, jotka riippuvat parametrien a_2 ja a_3 valinnasta.

Heunin 3. kertaluvun menetelmä Valitsemalla $a_2 = \frac{1}{3}, a_3 = \frac{2}{3}$ saadaan

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{1}{3}h, y_k + \frac{1}{3}hk_1) \\ k_3 &= f(t_k + \frac{2}{3}h, y_k + \frac{2}{3}hk_2)) \\ y_{k+1} &= y_k + \frac{h}{4}(k_1 + 3k_3). \end{aligned}$$

Kutta'n 3. kertaluvun menetelmä Valitsemalla $a_2 = \frac{1}{2}, a_3 = 1$ saadaan

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_1) \\ k_3 &= f(t_k + h, y_k - hk_1 + 2hk_2)) \\ y_{k+1} &= y_k + \frac{h}{6}(k_1 + 4k_2 + k_3). \end{aligned}$$

7.2.3 Klassinen Runge-Kutta-menetelmä

Kuten edellä määritellään parametrit

$$a_2, a_3, a_4, c_1, c_2, c_3, c_4, b_{2,1}, b_{3,1}, b_{3,2}, b_{4,1}, b_{4,2}, b_{4,3}$$

siten, että

$$a_k = \sum_{j=1}^{k-1} b_{k,j}, \;\; k = 2, 3, 4.$$

Edelleen asetetaan

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_j &= f(t_k + a_j h, y_k + h\sum_{l=1}^{j-1} b_{j,l} k_l),\ j = 2, 3, 4. \end{aligned}$$

Kuten edellisessä kappaleessa määrätään parametrit siten, että lokaalivirhe

$$d_{k+1} = y(t_{k+1}) - y(t_k) - h\sum_{j=1}^{4} c_j k_j$$

on mahdollsimman pieni. Epälineaarisessa yhtälöryhmässä on kahdeksan yhtälöä ja 10 tuntematonta parametria (jätetään harjoitustehtäväksi!!!). Historiallisesti vanhin on ns. Klassinen Runge-Kutta-menetelmä (4. kertaluvun menetelmä):

$$\begin{aligned} k_1 &= f(t_k, y_k) \\ k_2 &= f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_1) \\ k_3 &= f(t_k + \frac{1}{2}h, y_k + \frac{1}{2}hk_2) \\ k_4 &= f(t_k + h, y_k + hk_3) \\ y_{k+1} &= y_k + \frac{h}{6}[k_1 + 2k_2 + 2k_3 + k_4]. \end{aligned}$$

7.3 Korkeamman kertaluvun yhtälöt ja systeemit

Tarkastellaan esimerkiksi toisen kertaluvun differentiaaliyhtälöä

$$\begin{aligned} y''(t) &= f(t, y(t), y'(t)) \\ y(t_0) &= \alpha_1 \\ y'(t_0) &= \alpha_2 \end{aligned}$$

Yhtälö on ekvivalentti ensimmäisen kertaluvun differentiaaliyhtälön kanssa, jos asetetaan

$$y_1(t) = y(t),\ y_2(t) = y'(t).$$

Tällöin nimittäin on voimassa yhtälösysteemi

$$\begin{aligned} y_1'(t) &= y_2(t), \\ y_2'(t) &= f(t, y_1(t), y_2(t)) \end{aligned}$$

alkuehdoin

$$\begin{aligned} y_1(t_0) &= \alpha_1, \\ y_2(t_0) &= \alpha_2. \end{aligned}$$

Vastaavasti neljännen kertaluvun alkuarvotehtävä

$$y^{(4)}(t) = f(t, y'(t), y^{(2)}(t), y^{(3)}(t))$$

muunnetaan 1. kertaluvun differentiaaliyhtälösysteemiksi määrittelemällä

$$\begin{aligned} y_1(t) &= y(t) \\ y_2(t) &= y'(t) \\ y_3(t) &= y^{(2)}(t) \\ y_4(t) &= y^{(3)}(t). \end{aligned}$$

Funktiot $y_1,\ y_2,\ y_3,\ y_4$ toteuttavat silloin yhtälösysteemin

$$\begin{aligned} y_1'(x) &= y_2(x) \\ y_2'(x) &= y_3(x) \\ y_3'(x) &= y_4(x) \\ y_4'(x) &= f(t, y_1(t), y_2(t), y_3(t)). \end{aligned}$$

Teknillisissä sovellutuksissa matemaattiset mallit ovat usein toisen kertaluvun differentiaaliyhtälösysteemejä

$$\begin{aligned} u''(t) &= f(t, u(t), u'(t), v(t), v'(t)) \\ v''(t) &= g(t, u(t), u'(t), v(t), v'(t)) \end{aligned}$$

alkuehdoin

$$u(t_0) = a_1,\ u'(t_0) = a_2,\ v(t_0) = a_3,\ v'(t_0) = a_4.$$

Määritellään kuten aikaisemmin

$$\begin{aligned} y_1(t) &= u(t) \\ y_2(t) &= u'(t) \\ y_3(t) &= v(t) \\ y_4(t) &= v'(t). \end{aligned}$$

Yhtälösysteemi

$$\begin{aligned} y_1'(t) &= y_2(t) \\ y_2'(t) &= f(t, y_1, y_2, y_3, y_4) \\ y_3'(t) &= y_4(t) \\ y_4'(t) &= g(t, y_1, y_2, y_3, y_4) \end{aligned}$$

7.4 Implisiittiset menetelmät

Implisiittinen Eulerin menetelmä Approksimoidaan pisteessä t_{i+1} derivaattaa taaksepäin differenssimenetelmällä. Tällöin derivaatan lauseke on likipitäen

$$y'(t_{i+1}) \approx \frac{y_{i+1} - y_i}{h}.$$

Toisaalta mikäli $y(t)$ on alkuarvotehtävän ratkaisu, niin

$$y'(t_{i+1}) = f(t_{i+1}, y(t_{i+1}).$$

Yhdistämällä yo. lausekkeet saadaan implisiittisen Eulerin menetelmän Eulerin menetelmän yleinen askel:

$$y_{i+1} = y_i + hf(t_{i+1}, y_{i+1}).$$

Yhtälössä uutta approksimaatiota ei saada suoraan iteraatiosta; vaan se on ratkaistava iteratiivisesti:

$$y_{i+1}^{(k+1)} = y_i + hf(t_{i+1}, y_{i+1}^{(k)}),$$

missä alkuarvauksena voidaan käyttää eksplisiittisen Eulerin menetelmän askelta.

$$y_{i+1}^{(0)} = y_i + hf(t_i, y_i).$$

Voidaan osoittaa, että yksi iteraatio riittää, sillä tällöin menetelmä on sama kuin modifioitu Eulerin menetelmä ja tarkkuus ei olennaisesti parane lisäiteraatioilla.

Implisiittisessä puolisuunnikassäännössä sovelletaan samaa ajatusta kuin Runge-Kutta menetelmien tapauksessa:

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{y_{i+1}} f(s, y(s))ds.$$

Approksimoidaan oikean puolen integraalia puolisuunnikassäännöllä. Tällöin saadaan implisiittisen puolisuunnikassäännön yleinen iteraatioaskel askelpituudella $h = t_{i+1} - t_i$:

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

Tässäkin menetelmässä yleensä jokaisella askeleella y_{i+1} joudutaan ratkaisemaan iteratiivisesti.

7.5 Stabiilisuus, Konsistenssi ja Konvergenssi

Tarkastellaan yksiaskelmenetelmiä:

$$y_{k+1} = y_k + h\Phi(t_k, y_k; h).$$

Määritelmä 7.5.1 *Yksiaskelmenetelmä on konvergoiva, jos kiinteälle pisteelle $t = t_0 + kh$ on voimassa*

$$\lim_{h \to 0} |y(t) - y_k| = 0.$$

Huomaa, että edellisessä määritelmässä kh on vakio; vaikka askelpituus lähestyy nollaa. Pisteeseen t pääsemiseksi on suoritettava useampia aaskelia ko. menetelmällä.

Määritelmä 7.5.2 *Numeerinen ratkaisumenetelmä on konsistentti differentiaaliyhtälön kanssa, mikäli kiinteälle $t = t_0 + kh$*

$$\lim_{h \to 0}[\frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k); h)] = 0.$$

Edelleen sanotaan, että menetelmä on konsistentti kertalukua p, jos jollain $N > 0$

$$\sup_{t_0 \le t \le b} |\frac{y(t_{k+1}) - y(t_k)}{h} - \Phi(t_k, y(t_k); h)| \le \ Nh^p.$$

Määritelmä 7.5.3 *Olkoon* $\{\delta_0, \delta_1\}$ *ja* $\{\delta_0^*, \delta_1^*\}$ *differentiaaliyhtälön alkuarvotehtävän ratkaisumenetelmän häiriöitä ja* $\{z_k; k = 0, 1, 2, \dots\}$ *sekä* $\{z_k^*; k = 0, 1, 2, \dots\}$ *vastaavat ratkaisut, ts.*

$$\begin{aligned} z_{k+1} &= z_k + h\phi(t_k, z_k; h) + \delta_1, \\ z_0 &= \alpha_1 + \delta_0, \end{aligned}$$

ja

$$\begin{aligned} z_{k+1}^* &= z_k^* + h\phi(t_k, z_k^*; h) + \delta_1^*, \\ z_0^* &= \alpha_1 + \delta_0^*. \end{aligned}$$

Ratkaisumenetelmä on $\mathbf{n}olla - stabiili$*, jos*

$$|z_k - z_k^*| < \ S\epsilon$$

aina kun $|\delta_0 - \delta_0^*| < \epsilon$ *ja* $|\delta_1 - \delta_1^*| < \epsilon$.

Lause 7.5.1 *Jos ratkaisumenetelmä on nolla-stabiili ja konsistentti, niin silloin se on myös konvergoiva.*

7.6 Käytännön esimerkki: kierteinen pallo

Tarkastellaan ilmassa lähellä maan pintaa liikkuvaa palloa, jonka

- **massa** on m
- **halkaisija** on d
- **kulmanopeus** on $\vec{\omega}$

Vaikuttavat voimat

- Painovoima $\vec{G} = m\vec{g},\ \vec{g} = (0, 0, -g)$
- Ilmanvastus $\vec{D} = -D_L(v)\frac{\vec{v}}{v}$
- Magnus-voima $\vec{M} = M_L\frac{\vec{\omega}}{\omega} \times \frac{\vec{v}}{v}$

$$\begin{aligned} D_L(v) &= C_D \tfrac{1}{2} \tfrac{\pi d^2}{4} \rho v^2 \\ M_L(v) &= C_M \tfrac{1}{2} \tfrac{\pi d^2}{4} \rho v^2 \end{aligned}$$

missä ρ on ilmantiheys. Kertoimet C_D ja C_M ovat lisäksi nopeuden ja kulmanopeuden funktioita:

$$\begin{aligned} C_D &= 0.508 + (22.053 + 4.196(\frac{v}{\omega})^{\frac{5}{2}})^{-\frac{2}{5}} \\ C_M &= \frac{1}{2.022 + 0.981(\frac{v}{\omega})}. \end{aligned}$$

Newtonin laki $\Rightarrow$ alkuarvotehtävä

$$m\frac{d^2\vec{r}(t)}{dt^2} = -m\vec{g} - D_L\frac{\vec{v}}{v} + M_L\frac{\vec{\omega}}{\omega} \times \frac{\vec{v}}{v}$$

$$\vec{r}(0) = \vec{r}_0,\ \frac{d\vec{r}}{dt}(0) = \vec{v}_0.$$

Kuningas jalkapallossa tavallisesti

$$\begin{aligned} \vec{\omega} \cdot \vec{v} &= 0,\ t > 0 \\ \vec{\omega} &\ || \ \vec{k}. \end{aligned}$$

Tällöin pallon radan yhtälö xy-tasossa

$$\begin{aligned} \ddot{x} &= -C_D\alpha v\dot{x} - C_M\alpha v\dot{y} \\ \ddot{y} &= -C_D\alpha v\dot{y} + C_M\alpha v\dot{x} \end{aligned}$$

missä $v = \sqrt{\dot{x}^2 + \dot{y}^2},\ \alpha = (\rho\pi^2 d)/(8m)$.

Alkuehdot:

$$x(0) = 0,\ y(0) = 0,\ \dot{x}(0) = v_0\cos(\phi),\ \dot{y}(0) = v_0\sin(\phi).$$

Ongelma palautuu differentiaaliyhtälösysteemiksi määrittelemällä funktiot

$$\begin{aligned} u_1 &= x \\ u_2 &= y \\ u_3 &= \dot{x} \\ u_4 &= \dot{y}. \end{aligned}$$

Tällöin ratkaistavaksi saadaan alkuarvotehtävä

$$\begin{aligned}
\dot{u}_1 &= u_3 \\
\dot{u}_2 &= u_4 \\
\dot{u}_3 &= C_D\alpha\sqrt{u_3^2+u_4^2}u_3 - C_M\alpha\sqrt{u_3^2+u_4^2}u_4 \\
\dot{u}_4 &= -C_D\alpha\sqrt{u_3^2+u_4^2}u_3 + C_M\alpha\sqrt{u_3^2+u_4^2}u_4.
\end{aligned}$$

alkuehdoin

$$\begin{aligned}
u_1 &= 0 \\
u_2 &= 0 \\
u_3 &= v_0\cos(\phi) \\
u_4 &= v_0\sin(\phi).
\end{aligned}$$

Kyseinen alkuarvotehtävä voidaan ratkaista numeerisesti esimerkiksi Runge-Kutta-menetelmillä.

Luku 8

Finite Difference Method

8.1 1-ulotteinen reuna-arvotehtävä

Differentiaaliyhtälö

$$-y''(x) + p(x)y'(x) + q(x)y(x) = r(x).$$

Reunaehdot: $y(a) = y_0,\ y(b) = y_n$.

Ratkaisun olemassaolo: Oletetaan, että

1. $p(x), q(x), r(x)$ ovat jatkuvia välillä $[a, b]$;
2. $q(x) \geq \gamma > 0,\ x \in [a, b]$.

Näiden oletusten vallitessa reuna-arvotehtävällä on yksikäsitteinen ratkaisu, joka on määritelty koko välillä $[a, b]$. Numeerisen ratkaisun virhearvioita varten on kuitenkin oletettava, että ratkaisulla on C^4-säännöllisyys, ts.

$$\max_{a \leq x \leq b} |y^{(4)}(x)| \leq M < \infty.$$

Diskretisointi: Valitaan hilaparametri $h = \frac{b-a}{n}$. Tällöin ratkaisun approksimaatio lasketaan solmupisteissä

$$x_i = a + ih,\ i = 1, \ldots, n-1.$$

Solmupiste approksimaatio on siten $y_i \approx y(x_i)$. Korvataan sisäpisteissä x_i derivaatat differenssiapproksimaatioilla

$$\begin{aligned} y''(x_i) &\approx \frac{y(x_{i-1}) - 2y(x_i) + y(x_{i+1})}{h^2} \\ y'(x_i) &\approx \frac{y(x_{i+1}) - y(x_{i-1})}{2h}. \end{aligned}$$

Tällöin saadaan differenssiyhtälö

$$\frac{-y(x_{i-1}) + 2y(x_i) - y(x_{i+1})}{h^2} + p_i \frac{y(x_{i+1}) - y(x_{i-1})}{2h} + q_i y(x_i) = r_i + O(h^2),$$

missä $O(h^2)$ on derivaattojen approksimaatiovirhe sekä $p_i = p(x_i)$, $q_i = q(x_i)$, $r_i = r(x_i)$. Solmupisteapproksimaatiot ratkaistaan approksimaatioyhtälöstä

$$\frac{-y_{i-1} + 2y_i - y_{i+1}}{h^2} + p(x_i)\frac{y_{i+1} - y_{i-1}}{2h} q(x_i) y_i = r(x_i),$$

joka voidaan kirjoittaa kompaktimmassa muodossa

$$\begin{aligned} -(1 + \frac{p_i h}{2}) y_{i-1} + (2 + q_i h^2) y_i - (1 - \frac{p_i h}{2}) y_{i+1} &= r_i h^2 \\ y_0 &= y(a) \\ y_n &= y(b) \end{aligned}$$

Yhtälöryhmän kerroinmatriisi on diagonaalisesti dominantti, sillä

$$a_{ii} = 2 + q_i h^2 > 2 = |-(1 + \frac{p_i h}{2})| + |-(1 - \frac{p_i h}{2})| = |a_{i,i-1}| + |a_{i,1+1}|.$$

Näin ollen yhtälöryhmällä on yksikäsitteinen ratkaisu kaikilla oikean puolen vektoreilla $[r_1 h^2 \ \ldots \ r_{n-1} h^2]^T$, kun

$$\frac{p_i h}{2} < 1.$$

Lause 8.1.1 *Jos reuna-arvotehtävän ratkaisulle on voimassa*

$$\max |y^{(4)}| \leq M < \infty,$$

niin solmupisteapproksimaatiolle on voimassa virhe

$$|y_i - y(x_i)| \leq \frac{h^2}{12} M.$$

Neumannin reunaehto: Tarkastellaan seuraavaksi reuna-arvotehtävää, jossa toinen reunaehdoista on korvattu derivaattaehdolla

$$y'(a) = \alpha.$$

Määritellään fiktiivinen solmupiste $x_{-1} = a-h$. Approksimoidaan reunaehtoa pisteessä $x_0 = a$ keskeisdifferenssikaavalla

$$y'(x_0) \approx \frac{y(x_1) - y(x_{-1})}{2h} \approx \frac{y_1 - y_{-1}}{2h}.$$

Vaaditaan lisäksi, että solmupisteissä $x_i,\ i = 0, \ldots, n-1$, differenssiapproksimaatio toteutuu kuten edellisessä kohdassa. Tällöin saadaan lineaarinen yhtälöryhmä

$$\begin{aligned} y_1 - y_{-1} &= 2\alpha h \\ -(1 + \frac{p_i h}{2})y_{i-1} + (2 + q_i h^2)y_i - (1 - \frac{p_i h}{2})y_{i+1} &= r_i h^2, i = 0, \ldots, n-1 \\ y_n &= y(b) \end{aligned}$$

Edelleen yhtälöryhmä on yksikäsitteisesti ratkeava ja edellisen kohdan virhearviot ovat voimassa.

8.2 Epälineaarinen reuna-arvotehtävä

Reuna-arvotehtävä:

$$\begin{aligned} -y''(x) &= f(x, y(x), y'(x)) \\ y(a) &= y_0 \\ y(b) &= y_n \end{aligned}$$

Diskretisointi: Kuten lineaarisessa tapauksessa hilaparametri on $h = \frac{b-a}{n}$ ja solmupisteet ovat $x_i = a + ih, i = 0, \ldots, n$. Suorittamalla solmupisteissä derivaattojen differenssiapproksimaatio saadaan epälineaarinen yhtälöryhmä

$$\begin{aligned} -y_{i-1} + 2y_i - y_{i+1} &= h^2 f(x_i, y_i, \frac{y_{i+1} - y_{i-1}}{2h}),\ i = 1, \ldots, n-1 \\ y_0 &= y(a) \\ y_n &= y(b), \end{aligned}$$

josta solmupisteapproksimaatiot voidaan ratkaista. Yhtälöryhmän matriisiesitys on

$$A\vec{y} = h^2\vec{f}(\vec{y}),$$

missä A on vakiomatriisi, ja vektorin $\vec{y}$ koordinaatit ovat ratkaisun solmupisteapproksimaatiot.

Oletetaan, että funktion $\vec{f}(\vec{y})$ koordinaattifunktiot ovat jatkuvasti differentioituvia. Tällöin riittävän pienille h:n arvoilla oikean puolen funktion derivaatta (matriisi) on normiltaan pienempi kuin 1:

$$\|h^2\vec{f}'(\vec{y})\| \leq L < 1.$$

Näin ollen yhtälöryhmä voidaan ratkaista kiintopisteiteraatiolla:

$$\begin{aligned} \vec{y}_0 &= \text{alkuarvaus} \\ A\vec{y}_{n+1} &= h^2\vec{f}(\vec{y}_n). \end{aligned}$$

Myös Newtonin menetelmää voidaan käyttää ongelman ratkaisemiseen. Mutta usein kiintopisteiteraation käyttäminen on nopeampaa. Lisäksi differentiaaliyhtälön diskretisoinnisssa tehdään suurempi virhe kuin mitä kiintopisteiteraatiossa.

8.3 Poissonin yhtälö

Olkoon

$$R = \{(x, y) \in \mathbb{R}^2 |\ a \leq x \leq b,\ c \leq y \leq d\}$$

suorakaide tasossa. Tarkastellaan Poissonin yhtälöä

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

suorakaiteessa R Dirichlet'n reunaehdolla

$$u(x, y) = g(x, y),\ (x, y) \in \partial R,$$

missä ∂R on suorakaiteen reuna.

Määritellään suorakaiteeseen suorakaiteenmuotoinen diskreetti pistehila

$$\Theta(n, m) = \{(x_i, y_j) |\ \left\{ \begin{array}{l} x_i = a + i\frac{b-a}{n},\ i = 0, \ldots, n \\ y_j = c + j\frac{d-c}{m},\ j = 0, \ldots, m \end{array} \right. \}.$$

Merkitään jatkossa hilaparametrejä

$$h_1 = \frac{b-a}{n},\ h_2 = \frac{d-c}{m}.$$

Laplace-operaattorin approksimaatio diskreetillä pistehilalla saadaan toisen derivaatan differenssiapproksimaatiosta:

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x_i, y_j) &= \frac{-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}}{h_1^2} \\ -\frac{\partial^2 u}{\partial y^2}(x_i, y_j) &= \frac{-u_{i,j-1} + 2u_{i,j} - u_{i,j+1}}{h_2^2}, \end{aligned}$$

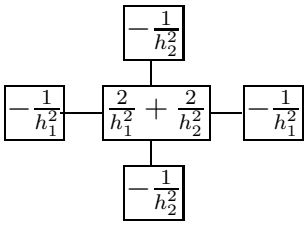
missä $u_{i,j}$ on ratkaisun approksimaatio solmupisteessä $P_{i,j} = (x_i, y_j)$. Näin ollen "diskreetti"Laplace-operaattori on

$$\Delta_h u(x_i, y_j) = \frac{1}{h_1^2}u_{i-1,j} + \frac{1}{h_2^2}u_{i,j-1} - [\frac{2}{h_1^2} + \frac{2}{h_2^2}]u_{i,j} + \frac{1}{h_1^2}u_{i+1,j} + \frac{1}{h_2^2}u_{i,j+1}.$$

Siten solmupisteapproksimaatiot ratkaistaan lineaarisesta yhtälöryhmästä

$$\begin{aligned} -\Delta_h u(x_i, y_j) &= f(x_i, y_j),\ i = 1, \ldots, n-1,\ j = 1, \ldots, m-1. \\ u_{0,j} &= g(a, y_j),\ u_{n,j} = g(b, y_j),\ j = 1, \ldots, m-1 \\ u_{i,0} &= g(x_i, c),\ u_{i,m} = g(x_i, d),\ i = 1, \ldots, n-1. \end{aligned}$$

Maski Yhtälöryhmä voidaan helposti muodostaa siirtämällä jokaisen sisäsolmupisteen päälle "maski", joka sisältää diskreetin Laplace-operaattorin kertoimet annetulle pistehilalle.



Uudelleen numerointi Tietokoneella laskettaessa on tarpeen numeroida solmupisteet uudelleen. Solmut numeroidaan kuten luettaisiin tekstiä -

"vasemmalta ylhäältä oikealle alas". Tällöin solmupisteen (x_i, y_j) järjestysnumero on

$$k = i + (m - 1 - j)(n - 1), \begin{cases} i \in \{i, \ldots, n-1\}, \\ j \in \{1, \ldots, m-1\}. \end{cases}$$

Vastaavaa solmupistearvoa merkitään tällöin $u_k = u_{i,j}$ ja sen lähimmät solmupistearvot ovat (nearest neihgbour)

$$\begin{aligned} u_{k-1} &= u_{i-1,j} \\ u_{k+1} &= u_{i+1,j} \\ u_{k-n+1} &= u_{i,j+1} \\ u_{k+n-1} &= u_{i,j-1} \end{aligned}$$

Uudelleen numeroiduille solmupisteille on voimassa yhtälöryhmä

$$-\frac{h_1^2}{h_2^2}u_{k-n+1} - u_{k-1} + 2[1 + (\frac{h_1}{h_2})^2]u_k - u_{k+1} - (\frac{h_1}{h_2})^2 u_{k+n-1} = b_k,$$

missä oikean puolen arvot b_k riippuvat funktion $f(x, y)$ solmupistearvoista ja mahdollisista reunaehdoista $g(x, y)$.

Yhtälöryhmän kerroinmatriisi on "blokkitridiagonaalinen"

$$A = \begin{bmatrix} B & -k\mathbb{I} & \mathbb{O} & \cdots & \mathbb{O} \\ -k\mathbb{I} & B & -k\mathbb{I} & \cdots & \mathbb{O} \\ \mathbb{O} & -k\mathbb{I} & B & \ddots & \mathbb{O} \\ \vdots & \ddots & \ddots & \ddots & -k\mathbb{I} \\ \mathbb{O} & \cdots & \cdots & -k\mathbb{I} & B \end{bmatrix},$$

missä $k = (\frac{h_1}{h_2})^2$, matriisi B on tridiagonaalimatriisi

$$B = \begin{bmatrix} 2(1+k) & -1 & 0 & \cdots & 0 \\ -1 & 2(1+k) & -1 & \cdots & 0 \\ 0 & -1 & 2(1+k) & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & -1 & 2(1+k) \end{bmatrix}.$$

ja $\mathbb{I}$ on $n \times n$-yksikkömatriisi ja $\mathbb{O}$ on $n \times n$-nollamatriisi. Matriisi A sisältää $m \times m$-blokkia.

Oikean puolen vektori muodostuu "varaustiheydestä"f(x,y) ja "reunapotentiaalista"g(x,y).

Varaustiheyden kontribuutio on vektori

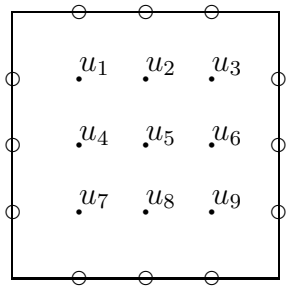
$$-h_1^2\vec{f} = -h_1^2 \begin{bmatrix} f(P_1) \\ f(P_2) \\ \vdots \\ f(P_{(m-1)(n-1)} \end{bmatrix}.$$

Reunaehtovektorista aiheutuu lisäys oikean puolen vektoriin, jos sisäsolmupiste $P_l, l = i + (m-1-j)(n-1)$ on reunasolmupisteen naapuri. Tämä on mahdollista vain silloin kun $i = 1, n-1$ tai $j = 1, m-1$.

Esimerkki 8.1 *Ratkaise Poissonin yhtälö*

$$\begin{aligned} -\Delta u(x,y) &= 0,\ 0 < x, y < \frac{1}{2} \\ u(0,y) = 0,\ u(\frac{1}{2}, y) &= 200y,\ 0 \le y \le \frac{1}{2} \\ u(x,0) = 0,\ u(x, \frac{1}{2}) &= 200x,\ 0 \le x \le \frac{1}{2} \end{aligned}$$

differenssimenetelmällä, kun $h_1 = h_2 = \frac{1}{8}$.



Yhtälöryhmän kerroinmatriisi on 3×3-blokkitridiagonaalinen matriisi

$$A = \begin{bmatrix} B & -\mathbb{I} & \mathbb{O} \\ -\mathbb{I} & B & -\mathbb{I} \\ \mathbb{O} & -\mathbb{I} & B \end{bmatrix}.$$

Matriisi B on tridiagonaalinen:

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

Matriisi $\mathbb{I}$ on 3×3-yksikkömatriisi ja $\mathbb{O}$ nollamatriisi.

Valkoisilla ympyröillä merkityistä solmupisteistä saadaan informaatio yhtälöryhmän oikealle puolelle, joka on

$$\vec{b} = [25,\ 50,\ 150,\ 0,\ 0,\ 50,\ 0,\ 0,\ 25]^T.$$

Yhtälöryhmän ratkaisu

$$\vec{u} = A^{-1}\vec{b} = [\frac{75}{4},\ \frac{75}{2},\ \frac{225}{4},\ \frac{25}{2},\ 25,\ \frac{75}{2},\ \frac{25}{4},\ \frac{25}{2},\ \frac{75}{4}]^T.$$

8.4 Lämpöyhtälön numeerinen ratkaisu

Yleinen diffuusioyhtälö on muotoa

$$\frac{\partial u}{\partial t} - a\frac{\partial^2 u}{\partial x^2} = f(x,t),\ 0 < t,\ 0 < x < 1,$$

missä $a > 0$ on diffuusiokerroin. Yleisesti diffuusiokerroin riippuu sekä ajasta että paikasta, ja joskus myös itse ratkaisusta, jolloin kyseessä on epälineaarinen diffuusioyhtälö. Nyt tarkastelemme yksinkertaisuuden vuoksi lineaarista ja vakiokertoimista diffuusioyhtälöä.

Reunaehdot: Toisella tai molemmilla reunoilla voi olla toinen seuraavista reunaehdoista:

- Dirichlet'n reunaehto: $u(0,t) = u_0(t)$ tai/ja $u(1,t) = u_n(t)$.
- Neumannin reunaehto: $u'(0,t) = q_0(t)$ ja/tai $u'(1,t) = q_1(t)$.
- Säteilylaki: $-u'(0,t) = F(u(0,t))$.

Huomaa reunaehdoista kullakin reunanosalla vain yksi ehto kerrallaan on voimassa.

Alkuehto: $u(x,0) = h(x)$.

8.4.1 Differenssimenetelmä

Tarkastellaan lämpöyhtälön numeerista ratkaisemista käyttäen differenssimenetelmää. Yksinkertaisuuden vuoksi tarkastellaan vain Dirichlet'n reunaehtoa.

Diskretisoinnissa valitaan aika- ja paikkamuuttujalle diskreettijoukko hilapisteitä (x_i, t_j), missä ko. pisteet muodostavat uniformihilan (säännöllisen suorakaidehilan) (x,t)-tasoon. Toisin sanoen

$$\begin{aligned} \Delta t &= t_{j+1} - t_j \\ h &= x_{i+1} - x_i \end{aligned}$$

ovat vakioita. Tällöin solmupisteet ovat

$$(x_i, t_j) = (ih, j\Delta t),\ i = 0, \ldots, n,\ j = 0, 1, 2, \ldots.$$

Numeerisen ratkaisumenetelmän tavoitteena on määrätä lämpöyhtälön ratkaisufunktion $u(x,t)$ solmupistearvojen $u(ih, j\Delta t)$ **approksimaatiot** $u_{i,j}$.

Alku- ja reunaehtojen avulla osa solmupistearvoista ovat tunnettuja:

alkuehto $u(x_i, 0) = u_{i,0} = h(x_i),\ i = 1, \ldots, n-1$;

reunaehto

$$\begin{aligned} u_{0,j} &= u(0, t_j) = g_0(t_j),\ j \in \mathbb{N} \\ u_{n,j} &= u(1, t_j) = g_1(t_j),\ j \in \end{aligned}$$

Sovelletaan derivaattojen approksimointiin differenssikaavoja. Aikaderivaatta solmupisteessä (x_i, t_j) voidaan approksimoida joko eteenpäin tai taaksepäin differenssikaavoilla:

$$\begin{aligned} \frac{\partial u}{\partial t}(x_i, t_j) &\approx \frac{u_{i,j+1} - u_{i,j}}{\Delta t} \text{ (eteenpäin diff.kaava)} \\ \frac{\partial u}{\partial t}(x_i, t_j) &\approx \frac{u_{i,j} - u_{i,j-1}}{\Delta t} \text{ (taaksepäin diff.kaava)} \end{aligned}$$

Paikkaderivaatan approksimaatioon käytämme keskeisdifferenssikaavaa:

$$\frac{\partial^2 u}{\partial x^2}(x_i, t_j) \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{h^2}.$$

Jokaisessa solmupisteessä diskretisoinnissa suoritetaan approksimaatiovirhe:

Aikaderivaatalle $\frac{1}{2}\Delta t \frac{\partial^2 u}{\partial t^2}(x_i, \theta_j)$;

paikkaderivaatalle $\frac{1}{24}h^2 \frac{\partial^4 u}{\partial x^4}(\xi_i, t_j)$.

Eteenpäin differenssimenetelmä Solmupisteessä (x_i, t_j) lämpöyhtälön approksimaatio on

$$\frac{u_{i,j+1} - u_{i,j}}{\Delta t} - \frac{a}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) = f_{i,j}.$$

Ratkaistaan differenssiyhtälöstä aikatasolla t_{j+1} olevat termit saadaan iteraatio ajansuhteen:

$$u_{i,j+1} = \frac{a\Delta t}{h^2}u_{i-1,j} + (1 - 2\frac{a\Delta t}{h^2})u_{i,j} + \frac{a\Delta t}{h^2}u_{i+1,j} + \Delta t f_{i,j},\ i = 1, \ldots, n-1.$$

Merkitään jatkossa $\lambda = \frac{a\Delta t}{h^2}$.

Eteenpäin differenssimenetelmä voidaan esittää kompaktisti matriisiesityksen avulla määrittelemällä jokaisella $j \in \mathbb{N}$ vektorit

$$\vec{u}_j = \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{n-1,j} \end{bmatrix},\ \vec{g}_j = \begin{bmatrix} \lambda g_0(t_j) \\ 0 \\ \vdots \\ \lambda g_1(t_j) \end{bmatrix},\ \vec{f}_j = \Delta t \begin{bmatrix} f_{1,j} \\ f_{2,j} \\ \vdots \\ f_{n-1,j} \end{bmatrix}$$

ja matriisi

$$A = \begin{bmatrix} 1-2\lambda & \lambda & 0 & \cdots & 0 \\ \lambda & 1-2\lambda & \lambda & \cdots & 0 \\ 0 & \lambda & 1-2\lambda & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & \lambda & 1-2\lambda \end{bmatrix}.$$

Tällöin eteenpäin differenssimenetelmän iteraatio on

$$\vec{u}_{j+1} = A\vec{u}_j + \vec{g}_j + \vec{f}_j;\ \vec{u}_0 = \vec{h},$$

missä

$$\vec{h}$$

on pisteiden x_i alkulämpötilavektori.

Taaksepäin differenssimenetelmä Aikaderivaatta korvataan taaksepäin differenssiapproksimaatiolla. Tällöin diskretisoitu lämpöyhtälö solmupisteessä on

$$\frac{u_{i,j} - u_{i,j-1}}{\Delta t} - \frac{a}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) = f_{i,j}.$$

Kerrotaan yhtälö Δt:llä ja siirretään $u_{i,j-1}$ yhtälön oikealle puolelle. Näin jokaisella aika-askeleella on ratkaistava tridiagonaalinen yhtälöryhmä

$$-\frac{a\Delta t}{h^2}u_{i-1,j} + (1 + 2\frac{a\Delta t}{h^2})u_{i,j} - \frac{a\Delta t}{h^2}u_{i+1,j} = u_{i,j-1} + \Delta t f_{i,j}.$$

Edellisen kohdan merkinnöillä yhtälöryhmä voidaan kirjoittaa muodossa

$$B\vec{u}_j = \vec{u}_{j-1} + \vec{g}_j + \vec{f}_j,$$

missä iteraatiomatriisi

$$B = \begin{bmatrix} 1+2\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & 1+2\lambda & -\lambda & \cdots & 0 \\ 0 & -\lambda & 1+2\lambda & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\lambda \\ 0 & \cdots & 0 & -\lambda & 1+2\lambda \end{bmatrix}.$$

Crank-Nicholson-menetelmässä lämpöyhtälöä approksimoidaan eteenpäin ja taaksepäin differenssimenetelmillä ja muodostetaan approksimatioiden keskiarvo. Tällöin solmupisteissä approksimaatioksi saadaan

$$\begin{aligned} \frac{u_{i,j+1} - u_{i,j}}{\Delta t} &- \frac{a[(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + (u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1})]}{2h^2} \\ &= \frac{1}{2}[f_{i,j} + fi, j+1]. \end{aligned}$$

Approksimaatio johtaa lineaariseen yhtälöryhmään

$$\begin{aligned} -\frac{\lambda}{2}u_{i-1,j+1} + (1+\lambda)u_{i,j+1} - \frac{\lambda}{2}u_{i+1,j+1} &= -\frac{\lambda}{2}u_{i-1,j} + (1-\lambda)u_{i,j} + \frac{\lambda}{2}u_{i+1,j} \\ &+ \frac{\Delta t}{2}(f_{i,j+1} + f_{i,j}), \end{aligned}$$

jonka matriisimuoto on

$$B\vec{u}_{j+1} = A\vec{u}_j + \frac{\lambda}{2}(\vec{g}_j + \vec{g}_{j+1}) + \frac{\Delta t}{2}(\vec{f}_j + \vec{f}_{j+1}),$$

Matriisit A ja B ovat samat kuin eteenpäin ja taaksepäin differenssimenetelmissä kunhan korvataan niissä λ luvulla $\frac{\lambda}{2}$.

8.4.2 Stabiilisuus

Lämpöyhtälön ratkaisumenetelmät, jotka perustuvat derivaattojen differenssi-approksimaatioon, on esitettävissä matriisi-iteraationa

$$\vec{u}_{j+1} = G\vec{u}_j + \vec{b}_j.$$

Matriisi G riippuu käytetystä numeerisesta approksimaatiosta. Vektori $\vec{b}_j$ sisältää informaation yhtälön reunaehdoista ja yhtälön oikean puolen funktiosta.

Olkoon $\{\vec{w}_j |\ j \in \mathbb{N}\}$ toinen approksimaatio lämpöyhtälön ratkaisulle poiketen approksimaatiosta $\vec{u}_j$ vain alkuehdossa:

$$\|\vec{u}_0 - \vec{w}_0\| \leq \delta.$$

Tällöin erotus vektori $\vec{w}_j - \vec{u}_j,\ j \in \mathbb{N}$ toteuttaa iteraation

$$\vec{e}_{j+1} = G\vec{e}_j.$$

Erityisesti induktion nojalla

$$\vec{e}_j = G^j\vec{e}_0.$$

Määritelmä 8.4.1 *Ratkaisumenetelmä on stabiili, jos*

$$\|\vec{e}_j\| \leq M,\ \forall j \in \mathbb{N},$$

kun $\|\vec{e}_0\| \leq \delta$.

Lämpöyhtälön ratkaisumenetelmä on stabiili yllä olevan määritelmän puitteissa mikäli matriisijono $\{G^j |\ j \in \mathbb{N}\}$ on rajoitettu normiltaan. Erityisesti näin on asian laita, kun matriisin G spektraalisäde

$$\rho(G) < 1. \text{ (vrt. Matriisialgebra)}$$

Spektraalisäde määriteltiin asettamalla

$$\rho(G) = \max\{|\mu| \ \mu \text{ on G:n omin.arvo}\}.$$

Siksi on tutkittava matriisin G ominaisarvoja.

Fourier'n menetelmässä oletetaan, että alussa on oskilloiva häiriö

$$\vec{e}_j = [e^{ijkh}]_{k=1,\ldots,n-1}.$$

Eteenpäin differenssimenetelmän stabiilisuus Eteenpäin differenssimenetelmälle iteraatiomatriisi on

$$G(\lambda) = \begin{bmatrix} 1-2\lambda & \lambda & 0 & \cdots & 0 \\ \lambda & 1-2\lambda & \lambda & \cdots & 0 \\ 0 & \lambda & 1-2\lambda & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \lambda \\ 0 & \cdots & 0 & \lambda & 1-2\lambda \end{bmatrix}.$$

Soveltamalla matriisia häiriöön $\vec{e}_j$ saadaan

$$G\vec{e}_j = (1 - 2\lambda + 2\lambda\cos(kh))\vec{e}_j,$$

toisin sanoen oskilloiva häiriö on matriisin ominaisvektori ja ominaisarvoina ovat luvut

$$\mu_k = 1 - 2\lambda + 2\lambda\cos(kh).$$

Koska $|\cos(kh)| \leq 1$, niin $1 - 4\lambda \leq \mu_k \leq 1$. Tällöin ratkaisumenetelmä on stabiili, mikäli $-1 < 1 - 4\lambda$. Tämän nojalla on voimassa

Lause 8.4.1 *Eteenpäin differenssimenetelmä on stabiili, jos*

$$\lambda = \frac{a\Delta t}{h^2} \leq \frac{1}{2}.$$

Taaksepäin differenssimenetelmän stabiilisuus

Lause 8.4.2 *Taaksepäin differenssimenetelmä on stabiili jokaisella*

$$\lambda = \frac{a\Delta t}{h^2} > 0.$$

Tod.: Taaksepäin differenssimenetelmälle iteraatiomatriisi on $G = B^{-1}$, missä

$$B(\lambda) = \begin{bmatrix} 1+2\lambda & -\lambda & 0 & \cdots & 0 \\ -\lambda & 1+2\lambda & -\lambda & \cdots & 0 \\ 0 & -\lambda & 1+2\lambda & \cdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\lambda \\ 0 & \cdots & 0 & -\lambda & 1+2\lambda \end{bmatrix}.$$

Kuten eteenpäin differenssimenetelmän yhteydessä matriisin B ominaisvektorit ovat

$$\vec{e_j} = [e^{ijkh}]_{k\in\{1,\ldots,n-1\}},$$

ja ominaisarvot ovat

$$\mu_k = 1 + 2\lambda - 2\lambda\cos(kh).$$

Näin ollen B:n käänteismatriisin ominaisarvot ovat B:n ominaisarvojen käänteislukuja

$$\frac{1}{1 + 2(1 - \cos(kh))},\ k = 1, \ldots, n-1.$$

Koska $1 - \cos(kh) \geq 0$, niin taaksepäin differenssimenetelmän ominaisarvot ovat itseisarvoltaan pienempiä kuin 1. Näin ollen menetelmä on stabiili kaikille $\lambda > 0$. □

Crank-Nicholson-menetelmän stabiilisuus