INDHOLDSFORTEGNELSE

Kapitel I. REGNING OG NUMERISK ANALYSE.

§ 1. Regning med tilnærmelser og afrundingsfejl.

Numerisk analyse handler om praktisk regning. Hvordan vi bærer os ad, og hvordan vi holder styr på de fejl, vi gør undervejs. Allerede når vi skal angive et reelt tal på lommeregneren, gør vi den fejl at afrunde tallet til de cifre, lommeregneren nu engang rummer. Og selv de simple aritmetiske operationer indfører nye fejl, som vi også må holde styr på.

Man interesserer sig for to former af fejlen, nemlig den absolutte fejl og den relative fejl.

ABSOLUT FEJL = SAND VÆRDI - TILNÆRMET VÆRDI

RELATIV FEJL = ABSOLUT FEJL / SAND VÆRDI

Fejlen er naturligvis altid ubekendt, men vi kan have en vurdering af den, f. eks. $|\text{abs fejl}| \leq \epsilon$ og dermed $|\text{rel fejl}| \leq \frac{\epsilon}{|\text{tiln værdi}| - \epsilon}$.

Ved addition af tal af samme størrelsesorden bliver fejlen højst adderet, og dermed forbliver den relative fejl uændret. Men ved addition af et stort og et lille tal går informationsværdien af det lille tal tabt.

Eksempel. $(1 + \frac{1}{n})^n \longrightarrow e$ for $n \longrightarrow \infty$. Prøv at beregne en følge af disse approksimander for n løbende gennem potenser af 2 eller 10 på lommeregner. Fra et vist trin er $1 + \frac{1}{n} = 1$ og dermed $(1 + \frac{1}{n})^n = 1$.

Ved subtraktion af næsten lige store tal går den relative fejl betragteligt op.

Eksempel. Løs en andengradsligning f. eks.

$$x^2 - 885\,x + 1 = 0$$

med den sædvanlige formel

$$x = \frac{885 \pm \sqrt{885^2-4}}{2} = \frac{885 \pm 884.998}{2} = \begin{cases} 884.999 \\ 0.001. \end{cases}$$

Her er den absolutte fejl naturligvis numerisk den samme, fordi summen er 885 eksakt; men den relative fejl er af størrelsesordenerne hhv. $5\cdot10^{-7}$ og $5\cdot10^{-1}$.

Ved multiplikation og division er det lettest at se på den relative fejl. Er $x = \hat{x} + \epsilon$ og $y = \hat{y} + \eta$, så er $xy = \hat{x}\hat{y} + \epsilon\hat{y} + \eta\hat{x} + \epsilon\eta$, hvorfor den relative fejl er ca. $\frac{\epsilon}{x} + \frac{\eta}{y}$ dvs. summen af de relative fejl.

Eksempel. Hvis vi ønskede at løse andengradsligningen ovenfor med lille relativ fejl på begge rødder, kunne vi benytte formlen for røddernes produkt $x_1x_2 = 1$. Heraf fås

$$x_2 = \frac{1}{x_1} = \frac{1}{884.999} = 0.001129945,$$

som har relativ fejl $5\cdot10^{-7}$ og dermed absolut fejl $5\cdot10^{-10}$.

Denne fremgangsmåde kan give approksimationer til begge rødder uden roduddragning. Har vi indsat 0 og 1 og dermed set, at den ene rod opfylder

$$0 < x_2 < 1,$$

får vi af $x_1 + x_2 = 885$, at den anden opfylder

$$884 < x_1 < 885.$$

Vi kan så sætte $x_1^0 = 885$ med $|AF| < 1$ og derfor $|RF| < \frac{1}{885} < 10^{-2}$. Men så er $x_2^0 = \frac{1}{x_1^0} = \frac{1}{885} = 0.00113$ med $|RF| < 10^{-2}$ og derfor $|AF| < 10^{-5}$.

Af summen finder vi nu $x_1^1 = 885 - x_2^0 = 884.99887$ med $|AF| < 10^{-5}$ og derfor $|RF| < 10^{-7}$. Og igen af produktet findes $x_2^1 = \frac{1}{x_1^1} = 0.001129949$ med $|RF| < 10^{-7}$ og derfor $|AF| <$

10^{-9}. Endelig findes $x_1^2 = 885 - x_2^1 = 884.9988700551$ med $|AF| < 10^{-9}$ og dermed $|RF| < 10^{-11}$ osv., osv.

§ 2. Fejlanalyser ved regning.

Summerer vi n tal med samme vurdering af den absolutte fejl, $|AF| < \epsilon$, kan vi med sikkerhed fastslå summens absolutte fejl til at være højst $n\epsilon$. Men er tallene dannet ved afrunding, er der dog en chance, for at nogen af fejlene har hævet hinanden.

Er et tal opnået ved afrunding, kan vi antage, at den sande værdi er lige fordelt med middelværdi i den tilnærmede værdi over et interval af længde 2ϵ. En sum af n sådanne tal er nu fordelt som en sum af n sådan fordelte stokastiske variable. Middelværdien bliver altså summen af de tilnærmede værdier, og variansen bliver summen af de n ens varianser, der hver er $\frac{\epsilon^2}{12}$, altså $\epsilon^2 \cdot \frac{n}{12}$.

Den centrale grænseværdisætning siger endda, at summen er approksimativt normalt fordelt med nævnte middelværdi og varians.

Vi kan altså efter smag og formål sige, at med

sandsynligheden	er	fejlen højst
$\frac{1}{2}$		$\epsilon \cdot 0.6745 \cdot \sqrt{\frac{n}{12}}$
95 %		$\epsilon \cdot 1.96 \cdot \sqrt{\frac{n}{12}}$
100 %		$\epsilon \cdot \frac{1}{2} \cdot n$

Dette sandsynlighedsteoretiske synspunkt kan forfølges for alle regningsarter, men det vil ikke blive gjort her.

§ 3. De generelle problemer i numerisk analyse.

I dette kursus interesserer vi os for problemer i forbindelse med en operator, (funktional, funktion) af et Banach rum (funktionsrum, talrum) ind i sig selv eller et andet:

$$T: X \longrightarrow Y.$$

Det <u>direkte</u> problem er:

Givet $x \in X$, find $y = Tx$.

Eksempel. $T = \int_a^b$, $X = C[a,b]$, $f \in X$, find $\int_a^b f(x)dx$.

Det <u>omvendte</u> problem er:

Givet $y \in Y$, find $x \in X$, så $Tx = y$.

Eksempel 1. $T = A$, en $n \times n$-matrix og $X = Y = \mathbb{R}^n$, løs ligningssystemet $Ax = y$.

Eksempel 2. $T = f: \mathbb{R} \longrightarrow \mathbb{R}$, en kontinuert funktion, find et nulpunkt for f, dvs. $x \in \mathbb{R}$, så $f(x) = 0$.

Det omvendte problem føres tilbage til det direkte på to måder. 1) Man finder $T^{-1}: Y \longrightarrow X$. 2) Man gætter x og beregner direkte Tx til sammenligning med y.

<u>Egenværdiproblemet</u>: For T lineær og $X = Y$ søges $\lambda \in \mathbb{C}$ og $x \in X$, så $Tx = \lambda x$.

Eksempel. $T = A$, en $n \times n$-matrix og $X = Y = \mathbb{C}^n$.

Egenværdiproblemet kan føres tilbage til det omvendte problem. Lad $S = \{x \in X \mid ||x|| = 1\}$ være enhedskuglen i X og $\hat{T}: S \times \mathbb{C} \longrightarrow X$ være defineret ved $\hat{T}(x,\lambda) = Tx - \lambda x$. Så er $\hat{T}(x,\lambda) = 0$, netop når x er en egenvektor med λ som egenværdi. Er X et Hilbertrum, kan vi definere $\hat{T}: S \longrightarrow X$ ved

$\hat{T}(x) = Tx - (Tx,x)x$, som er 0 netop når x er egenvektor med (Tx,x) som egenværdi.

§ 4. Fejlvurderinger i de generelle problemer.

I mere interessante matematiske problemer kan vi ikke umiddelbart klare os med aritmetik. Der optræder grænseovergange og uendelighed i form af kontinuerte funktioner på intervaller, der indeholder uendelig mange punkter.

Hvis et tal er bestemt som grænseværdi for en følge, $x_n \longrightarrow x$, og vi benytter x_n som approksimation til x, kalder vi den derved opståede fejl, $x - x_n$, for TRUNKERINGSFEJLEN. F. eks. ved en rækkeudvikling af én funktion, hvor vi summerer n led af rækken, opstår en sådan fejl.

I det direkte problem, hvor T kan beregnes, men x må tilnærmes med x_n, får vi altså en trunkeringsfejl, $x - x_n$, nogle afrundingsfejl ved beregningen af Tx_n, samt fejlen på resultatet $Tx - Tx_n$. Er T differentiabel, giver middelværdisætningen

$$Tx - Tx_n = T'(\xi)(x-x_n),$$

hvorefter fejlvurderingen er

$$||Tx - Tx_n|| \leq K \cdot ||x-x_n||$$

for passende normer, hvor altså $||T'(\xi)|| \leq K$ for ξ i en omegn af x og x_n.

I det tilsvarende omvendte problem fås trunkeringsfejlen på y, $y - y_n$, og vurderingen omvendt

$$||x - x_n|| \leq K \cdot ||y - y_n||,$$

hvor $||T'(\xi)^{-1}|| \leq K$ for ξ i omegnen.

I begge tilfælde kaldes konstanten K <u>konditionstallet</u> for problemet. Det skal vi vende tilbage til i kapitel II. Er konditionstallet meget stort, siges problemet at være <u>dårligt konditioneret</u>. Sagen er, at trunkeringsfejlen bliver ganget med K.

I problemer, der involverer funktionsrum, må vi ofte approksimere operatoren T. Er X og Y Banach rum, har vi i disse tilfælde underrum $X_n \subseteq X$ og $Y_n \subseteq Y$, som er endelig- (n-) dimensionale vektorrum, der indlejres ved en <u>interpolation</u> $i_n: X_n \longrightarrow X$ og $j_n: Y_n \longrightarrow Y$, således at der også findes <u>projektioner</u> herpå, $p_n: X \longrightarrow X_n$, $q_n: Y \longrightarrow Y_n$, f. eks. funktionsværdierne i visse punkter. Der gælder da, at

$$p_n i_n = id_{X_n} \qquad (q_n j_n = id_{Y_n}).$$

I almindelighed forlanger vi yderligere om disse operatorer, at de er lineære, at der findes p, q, i, j, så

$$||p_n|| \leq p < \infty, \quad ||q_n|| \leq q < \infty,$$
$$||i_n|| \leq i < \infty, \quad ||j_n|| \leq j < \infty$$

for alle n, samt at $||i_n p_n x - x|| \longrightarrow 0$, $||j_n q_n x - x|| \longrightarrow 0$ for $n \longrightarrow \infty$.

Man kan definere konvergens af $x_n \longrightarrow x$ på to måder, <u>diskret</u> ved $||x_n - p_n x|| \longrightarrow 0$ og <u>globalt</u> ved $||x - i_n x_n|| \longrightarrow 0$. Men med antagelserne ovenfor er de to begreber ækvivalente. Thi

$$x_n - p_n x = p_n i_n x_n - p_n x = p_n(i_n x_n - x),$$

så $||x_n - p_n x|| \leq ||p_n|| \cdot ||i_n x_n - x|| \leq p \cdot ||x - i_n x_n||$;

og $x - i_n x_n = x - i_n p_n x + i_n p_n x - i_n x_n$,

så $||x - i_n x_n|| \leq ||x - i_n p_n x|| + i \cdot ||p_n x - x_n||$.

Problemet er nu at definere $T_n: X_n \longrightarrow Y_n$, så $T_n \longrightarrow T$ for $n \longrightarrow \infty$ i den forstand, at for <u>enhver</u> følge $x_n \longrightarrow x$ gælder $T_n x_n \longrightarrow Tx$ (diskret eller globalt).

Vi vil dele konvergensproblemet i to; et <u>stabilitets-problem</u> for operatorfølgen T_n, der går ud på, at T_n er ufølsom over for <u>fejl</u> $(x_n - p_n x)$, og et <u>konsistensproblem</u> for T_n med T, der går ud på konvergens i specialtilfældet

$x_n = p_n x$.

Definition. Følgen af operatorer T_n kaldes stabil, hvis der findes en konstant K, så for alle $x_n^1, x_n^2 \in X_n$ gælder:

$$\forall n \in \mathbb{N}: ||T_n x_n^1 - T_n x_n^2|| \leq K \cdot ||x_n^1 - x_n^2||.$$

Definition. Følgen af operatorer T_n kaldes konsistent med operatoren T, hvis diskretiseringsfejlen

$$\delta_n(x) = ||q_n Tx - T_n p_n x|| \longrightarrow 0 \quad \text{for} \quad n \longrightarrow \infty.$$

Sætning: T_n stabil og T_n konsistent med T $\Rightarrow$ $T_n \longrightarrow T$.

Bevis: $||T_n x_n - q_n Tx|| \leq ||T_n x_n - T_n p_n x|| + ||T_n p_n x - q_n Tx||$

$$\leq K \cdot ||x_n - p_n x|| + \delta_n(x). \qquad \square$$

§ 5. Hovedsætningen i numerisk analyse.

Den vigtigste, men ikke den eneste, metode i numerisk analyse går ud på at omformulere sit problem til et fixpunktproblem, som så forsøges løst ved iteration.

Hovedsætning: Er f en afstandsformindskende afbildning af et Banach rum ind i sig selv, så findes netop ét fixpunkt $x^* = f(x^*)$ for f, og for ethvert punkt x gælder, at følgen $f^n(x)$ konvergerer mod x^* for $n \longrightarrow \infty$

Bevis: Der findes $\alpha < 1$, så for vilkårlige x og y gælder $||f(x) - f(y)|| < \alpha \cdot ||x - y||$. Altså er

$$||f^{n+1}(x) - f^n(x)|| < \alpha ||f^n(x) - f^{n-1}(x)||$$

og derfor

$$||f^{n+1}(x) - f^n(x)|| < \alpha^n ||f(x) - x||.$$

Men så er for p vilkårlig

$$||f^{n+p}(x) - f^n(x)|| < \sum_{i=0}^{p-1} \alpha^{n+i} ||f(x) - x||$$

$$= \alpha^n \cdot \frac{1-\alpha^p}{1-\alpha} \cdot ||f(x) - x||$$

$$< \alpha^n \cdot \frac{||f(x) - x||}{1-\alpha},$$

som er uafhængig af p og går mod 0 for $n \longrightarrow \infty$. Derfor er $f^n(x)$ en fundamentalfølge, som i et Banach rum er konvergent. Sæt $x^* = \lim f^n(x)$, da gælder, når f jo er kontinuert, at $f(x^*) = \lim f(f^n(x)) = \lim f^n(x) = x^*$. Altså findes et fixpunkt, og enhver itereret følge konvergerer mod et fixpunkt. Men er nu x^* og y^* to fixpunkter, gælder $||x^* - y^*|| = ||f(x^*) - f(y^*)|| \leq \alpha \cdot ||x^* - y^*||$, hvoraf $(1-\alpha)||x^* - y^*|| \leq 0$, altså $||x^* - y^*|| = 0$, hvoraf $x^* = y^*$. □

Corollar: Er f en kontinuert funktion af et metrisk rum ind i sig selv, og er en itereret følge $f^n(x)$ konvergent med grænseværdi x^*, så er $f(x^*) = x^*$.

Kapitel II. LØSNING AF LINEÆRE LIGNINGER VED DIREKTE METODER.

§1. Problemformuleringen.

En lineær afbildning $f: \mathbb{R}^n \to \mathbb{R}^n$ eller $f: \mathbb{C}^n \to \mathbb{C}^n$ er givet ved en matrix $A = (a_{ij})$, hvor $a_{ij} \in \mathbb{R}$ eller $a_{ij} \in \mathbb{C}$ for $i,j = 1,\cdots,n$, og $f(x) = A \cdot x$ med sædvanlig matrixmultiplikation. Har vi givet et punkt $b \in \mathbb{R}^n$ eller $b \in \mathbb{C}^n$, ønsker vi at bestemme $x \in \mathbb{R}^n$, hhv. $\mathbb{C}^n$, så

$$f(x) = b .$$

Ligningen skrives også

$$A \cdot x = b ,$$

eller udførligt

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n . \end{aligned}$$

Ved en *direkte* metode til løsning af problemet forstås en metode, der fører til den nøjagtige løsning efter endelig mange skridt.

Dvs. at fejlen kun består af ophobede afrundingsfejl.

Herved findes en kandidat, $\hat{x}$, til problemets løsning. Gøres prøve i ligningen, findes RESIDUET

$$b - A\hat{x},$$

og heraf fås

$$\frac{||x - \hat{x}||}{||x||} \leq K \cdot \frac{||b - A\hat{x}||}{||b||},$$

hvor K er konditionstallet (her for de relative fejl).

§ 2. Normer og konditionstal for matricer.

Normer.

Har vi en norm på $\mathbb{R}^n$ eller $\mathbb{C}^n$, f. eks.

$$||x||_1 = |x_1| + |x_2| + \ldots + |x_n|,$$

$$||x||_2 = \sqrt{|x_1|^2 + |x_2|^2 + \ldots + |x_n|^2}$$

$$= \sqrt{(x,x)},$$

$$||x||_\infty = \max_i |x_i|,$$

kan vi definere den tilsvarende matrix-norm ved

$$(*) \qquad ||A|| = \sup \left\{ ||Ax|| \mid ||x|| \leq 1 \right\}.$$

Disse betegnes hhv. $||A||_1$, $||A||_*$, $||A||_\infty$, fordi vi reserverer "2" til den euklidiske matrix-norm

$$||A||_2 = \sqrt{\sum_{i,j} |a_{ij}|^2} \ .$$

En norm, der er defineret ved (*), opfylder de sædvanlige normegenskaber, og dertil ulighederne

$$(1) \qquad ||Ax|| \leq ||A|| \cdot ||x||,$$

$$(2) \qquad ||AB|| \leq ||A|| \cdot ||B||.$$

En matrix-norm, der opfylder (1), kaldes fordragelig med rumnormen i (1).

Bevis: (1). $||x|| = 0 \Rightarrow ||Ax|| = 0$. Ellers er $\lambda = \frac{1}{||x||}$, og

$$\lambda \cdot ||Ax|| = ||\lambda Ax|| = ||A(\lambda x)|| \leq \sup \left\{ ||Ax|| \mid ||x|| \leq 1 \right\} \leq ||A||,$$

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

(2). Følger af (*) og (1):

$$||(AB)x|| = ||A(Bx)|| \leq ||A|| \cdot ||Bx|| \leq ||A|| \cdot ||B|| \cdot ||x||,$$

så for $||x|| \leq 1$ gælder $||(AB)x|| \leq ||A|| \cdot ||B||$. Altså

$$||AB|| = \sup \left\{ ||(AB)x|| \mid ||x|| \leq 1 \right\} \leq ||A|| \cdot ||B||. \qquad \square$$

Også den euklidiske matrix-norm opfylder (1) og (2), hvor (1) gælder sammen med den euklidiske vektornorm.

Bevis: (1). Cauchy-Schwartz' ulighed giver os

$$||Ax||_2^2 = \sum_j \left| \sum_k a_{jk}x_k \right|^2 \leq \sum_j \left(\sum_k |a_{jk}|^2 \right) \left(\sum_i |x_i|^2 \right)$$

$$= \sum_i |x_i|^2 \sum_{j,k} |a_{jk}|^2 = ||x||_2^2 \cdot ||A||_2^2,$$

(2). $$||AB||_2^2 = \sum_{i,k} \left| \sum_j a_{ij}b_{jk} \right|^2$$

$$\leq \sum_{i,k} \left(\sum_j |a_{ij}|^2 \right) \left(\sum_l |b_{lk}|^2 \right)$$

$$= \left(\sum_{i,j} |a_{ij}|^2 \right) \cdot \left(\sum_{l,k} |b_{lk}|^2 \right) = ||A||_2^2 \cdot ||B||_2^2. \quad \square$$

Corollar: For en vilkårlig matrix A gælder

$$||A||_* \leq ||A||_2.$$

Bevis: Af (1) fås for $||x||_2 \leq 1$, at

$$||Ax||_2 \leq ||A||_2,$$

hvoraf $$||A||_* = \sup \left\{ ||Ax||_2 \mid ||x||_2 \leq 1 \right\} \leq ||A||_2. \quad \square$$

Konditionstal.

For problemet med A regulær

$$Ax = y$$

med tilnærmelsesløsningen

$$A\hat{x} = \hat{y},$$

fås $$A(x - \hat{x}) = y - \hat{y},$$ og

$$x - \hat{x} = A^{-1}(y - \hat{y}).$$

(1) giver for en vilkårlig norm

$$||x - \hat{x}|| \leq ||A^{-1}|| \cdot ||y - \hat{y}||$$

og $$||y|| \leq ||A|| \cdot ||x||,$$

hvoraf

$$\frac{||x - \hat{x}||}{||x||} \leq ||A|| \cdot ||A^{-1}|| \cdot \frac{||y - \hat{y}||}{||y||}.$$

Altså er $||A^{-1}||$ og $||A|| \cdot ||A^{-1}||$ konditionstal for problemet med hensyn til de valgte normer og hhv. den absolutte og den relative fejl.

Bemærk, at (2) siger:

$$||E|| = ||AA^{-1}|| \leq ||A|| \cdot ||A^{-1}||.$$

($||E||_1 = ||E||_* = ||E||_\infty = 1$, $||E||_2 = \sqrt{n}$.)

Konditionstallet er altså altid mindst 1.

Til beregning af konditionstal har vi foruden formlen for $||A||_2$ også formler for $||A||_1$ og $||A||_\infty$.

$$||A||_\infty = \max_i \sum_j |a_{ij}|$$

$$||A||_1 = \max_j \sum_i |a_{ij}|.$$

Med andre ord, max-normen svarer til største numeriske rækkesum og sumnormen til største numeriske søjlesum.

Bevis: $||A||_\infty$.

$$||A||_\infty = \sup \left\{ \max_i \left| \sum_j a_{ij}x_j \right| \;\middle|\; \max_j |x_j| \leq 1 \right\}$$

$$= \max_i \sup \left\{ \left| \sum_j a_{ij}x_j \right| \;\middle|\; |x_j| \leq 1 \right\}$$

$$= \max_i \sum_j |a_{ij}|,$$

idet vi dels har $\left|\sum_j a_{ij}x_j\right| \leq \sum_j |a_{ij}x_j| \leq \sum_j |a_{ij}|$, når $|x_j| \leq 1$, og dels for valget af x_j, så $|x_j| = 1$ og $a_{ij}x_j \geq 0$, får for hvert i for sig, at

$$\sum_j |a_{ij}| = \left|\sum_j a_{ij}x_j\right| \in \left\{ \left|\sum_j a_{ij}x_j\right| \;\middle|\; |x_j| \leq 1 \right\}.$$

$||A||_1$.

$$||A||_1 = \sup\left\{ \sum_i \left|\sum_j a_{ij}x_j\right| \;\middle|\; \sum_j |x_j| \leq 1 \right\}.$$

$$\sum_i \left|\sum_j a_{ij}x_j\right| \leq \sum_i \sum_j |a_{ij}x_j| = \sum_j \sum_i |a_{ij}| \cdot |x_j| \leq$$

$$= \sum_j |x_j| \sum_i |a_{ij}| \leq \sum_j |x_j| \cdot \max_k \sum_i |a_{ik}| =$$

$$= \left(\max_k \sum_i |a_{ik}|\right) \cdot \sum_j |x_j| \leq \max_k \sum_i |a_{ik}|,$$

når $\sum_j |x_j| \leq 1$. Altså gælder

$$||A||_1 \leq \max_j \sum_i |a_{ij}|.$$

Lad nu m være valgt, så $\sum_i |a_{im}| = \max_k \sum_i |a_{ik}|$. Vi sætter $x = (0,0,\ldots,1,\ldots,0)$, hvor 1-tallet står på plads nr. m. For dette valg af x fås

$$\sum_i \left|\sum_j a_{ij}x_j\right| = \sum_i |a_{im}| = \max_k \sum_i |a_{ik}|,$$

samtidig med at x opfylder

$$\sum_j |x_j| \leq 1.$$

Derfor sluttes, at

$$\max_k \sum_i |a_{ik}| \in \left\{ \sum_i \left|\sum_j a_{ij}x_j\right| \;\middle|\; \sum_j |x_j| \leq 1 \right\}.$$

Men heraf følger åbenbart

$$||A||_1 \geq \max_j \sum_i |a_{ij}|$$

□

Bemærkning om egenværdier.

For en vilkårlig fordragelig matrix-norm og en vilkårlig egenværdi λ for matricen A, gælder

$$|\lambda| \leq ||A||.$$

Bevis: Der findes en norm på vektorrummet, så (1) er opfyldt. Lad x være en egenvektor ($\neq 0$) for egenværdien λ. Da gælder

$$\lambda x = Ax,$$

hvoraf $$|\lambda| \cdot ||x|| = ||\lambda x|| = ||Ax|| \leq ||A|| \cdot ||x||,$$

så da $||x|| > 0$, fås umiddelbart den ønskede ulighed.

□

§ 3. Elimination.

Afrundingsfejlen vokser med matricens størrelse som kvadratroden af antallet af regneoperationer. (Se kapitel I § 2.) Ved determinantberegning efter definitionen og løsning af lineære ligninger med Cramers formler, vokser antallet af regninger eksponentielt som funktion af dimensionen. For 10 ligninger og regning med 5 betydende cifre, dvs. relativ fejl 10^{-5}, bliver den relative fejl 10^{-1}. (Når vi så ganger med konditionstallet, bliver der ikke megen information tilbage.)

Ved elimination vokser regneoperationstallet som n^3, altså afrundingsfejlen som $n^{3/2}$.

Vi betragter et ligningssystem med regulær koefficientmatrix A:

$$(1)\qquad \begin{aligned} a_{11}x_1 + \ldots + a_{1n}x_n &= a_{1(n+1)} \\ \vdots \qquad\quad & \quad \vdots \quad\; \vdots \\ a_{n1}x_1 + \ldots + a_{nn}x_n &= a_{n(n+1)} \end{aligned}$$

Ved ombytning af ligninger kan vi opnå, at $|a^o_{11}| = \max_j |a_{j1}|$. Ved division af ligningen med a^o_{11} ($\neq 0$, da A er regulær), fås en ny ligning

$$(2)\qquad x_1 + a^1_{12}x_2 + \ldots + a^1_{1n}x_n = a^1_{1(n+1)}$$

hvor $a_{1k} = \dfrac{a^o_{1k}}{a^o_{11}}$.

Nu ganges (2) med a^o_{j1} og fratrækkes den j'te ligning for $j = 2,\ldots,n$. Herved fremkommer et reduceret ligningssystem

$$
\begin{aligned}
x_1 + a^1_{12}x_2 + \ldots + a^1_{1n}x_n &= a^1_{1(n+1)} \\
a^1_{22}x_2 + \ldots + a^1_{2n}x_n &= a^1_{2(n+1)} \\
\vdots \qquad\qquad \vdots & \qquad \vdots \\
a^1_{n2}x_2 + \ldots + a^1_{nn}x_n &= a^1_{n(n+1)}
\end{aligned} \tag{3}
$$

hvor $a^1_{jk} = a^o_{jk} - a^o_{j1}a^1_{1k}$, $k = 2,\ldots,n+1$.

Herefter ombyttes ligninger, så $|a^1_{22}| = \max_{2\leq k\leq n} |a^1_{k2}|$; lad os tænke os, at det er gjort. Igen findes en ligning svarende til (2) ved division med a^1_{22}

$$x_2 + a^2_{23}x_3 + \ldots + a^2_{2n}x_n = a^2_{2(n+1)}, \tag{4}$$

hvor $a^2_{2k} = \dfrac{a^1_{2k}}{a^1_{22}}$.

Igen ganges (4) med a^1_{j2} for $j \neq 2$ og fratrækkes den j'te ligning. Herved fremkommer

$$
\begin{aligned}
x_1 + 0\cdot x_2 + a^2_{13}x_3 + \ldots + a^2_{1n}x_n &= a^2_{1(n+1)} \\
x_2 + a^2_{23}x_3 + \ldots + a^2_{2n}x_n &= a^2_{2(n+1)} \\
a^2_{33}x_3 + \ldots + a^2_{3n}x_n &= a^2_{3(n+1)} \\
\vdots \qquad\qquad \vdots & \qquad \vdots \\
a^2_{n3}x_3 + \ldots + a^2_{nn}x_n &= a^2_{n(n+1)},
\end{aligned} \tag{5}
$$

hvor $a^2_{jk} = a^1_{jk} - a^1_{j2}a^2_{2k}$, $j \neq 2$, $k > 2$.

Således bliver man ved, indtil man har elimineret alle ubekendte, hvorved løsningen bliver

$$x_j = a^n_{j(n+1)}, \quad j = 1,\ldots,n. \tag{6}$$

Bemærk, at man uden videre løser flere ligningssystemer simultant. F. eks. kan man også invertere matricen ved at betragte højresiderne

$$\begin{pmatrix}1\\0\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}, \ldots, \begin{pmatrix}0\\0\\0\\\vdots\\1\end{pmatrix},$$

benævnt i formlerne $a_{1k},\ldots,a_{nk}$, $k = n+2,\ldots,2n+1$.

Man ender herved med systemet

$$\begin{aligned}
x_1 &= a^n_{1(n+2)}\ a^n_{1(n+3)} \cdots a^n_{1(2n+1)}\\
x_2 &= a^n_{2(n+2)}\ a^n_{2(n+3)} \cdots a^n_{2(2n+1)}\\
&\ \ \vdots \qquad\ \vdots \qquad\qquad\ \vdots\\
x_n &= a^n_{n(n+1)}\ a^n_{n(n+2)} \cdots a^n_{n(2n+1)},
\end{aligned}$$

som ved tilbagepermutation af ligningerne til den oprindelige orden giver A^{-1} på højre side. Dette kan være nyttigt til beregning af konditionstallet. Hertil er det nok, at A^{-1} er beregnet med lille nøjagtighed.

Bemærk, at alt er lige gyldigt, hvad enten vi har reelle eller komplekse tal i matricen A.

§ 4. Ortogonalisering og overbestemte ligningssystemer.

Vi skriver ligningssystemet (1) i § 3 som

$$x_1\begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix} + x_2\begin{pmatrix} a_{12} \\ \vdots \\ a_{n2} \end{pmatrix} + \ldots + x_n\begin{pmatrix} a_{1n} \\ \vdots \\ a_{nn} \end{pmatrix} = \begin{pmatrix} a_{1(n+1)} \\ \vdots \\ a_{n(n+1)} \end{pmatrix}$$

Opfattes søjlerne som vektorer $a^{(1)},\ldots,a^{(n+1)}$, er opgaven at fremstille $a^{(n+1)}$ som linearkombination af de øvrige. Lad os antage $a^{(j)} \in \mathbb{C}^n$ og søge $x_j \in \mathbb{C}$.

I $\mathbb{C}^n$ har vi det indre produkt

$$(x,y) = x_1\bar{y}_1 + \ldots + x_n\bar{y}_n,$$

og den hertil svarende norm

$$||x||_2 = \sqrt{(x,x)}.$$

Med hensyn til dette produkt ortonormaliseres basen $a^{(1)},\ldots,a^{(n)}$:

$$v_1 = a^{(1)}, \qquad w_1 = \frac{v_1}{||v_1||};$$

$$v_j = a^{(j)} - \sum_{k=1}^{j-1}(a^{(j)},w_k)w_k, \qquad w_j = \frac{v_j}{||v_j||},$$

for $j = 2,\ldots,n$;

$$v_{n+1} = a^{(n+1)} - \sum_{k=1}^{n}(a^{(n+1)},w_k)w_k = 0,$$

fordi $w_1,\ldots,w_n$ er en ortonormal basis for $\mathbb{C}^n$. Herefter findes $x_1,\ldots,x_n$ af formlerne

$$x_n = \frac{(a^{(n+1)},v_n)}{||v_n||_2^2}$$

(*)

$$x_j = \frac{1}{||v_j||_2^2}\left((a^{(n+1)},v_j) - \sum_{k=j+1}^{n} x_k(a^{(k)},v_j)\right).$$

Bevis. Vektorsættene $(a^{(1)},\ldots,a^{(j)})$ og $(w_1,\ldots,w_j)$ udspænder samme underrum af $\mathbb{C}^n$. Derfor er

$$a^{(n+1)} = \sum_{k=1}^{j-1}\beta_k w_k + \sum_{k=j}^{n} x_k a^{(k)},$$

for $j = 1,\ldots,n$, for visse entydigt bestemte β_k. Heraf findes, da $(w_k,v_j) = 0$ for $k \neq j$,

$$(a^{(n+1)},v_j) = \sum_{k=j}^{n} x_k (a^{(k)},v_j)$$

$$= x_j(a^{(j)},v_j) + \sum_{k=j+1}^{n} x_k (a^{(k)},v_j).$$

Af definitionen på v_j findes endvidere

$$(v_j,v_j) = (a^{(j)},v_j) - \sum_{k=1}^{j-1}(a^{(j)},w_k)(w_k,v_j) = (a^{(j)},v_j),$$

hvoraf (*) følger. □

Denne metode anvendes også ved overbestemte ligningssystemer. F. eks. kan vektorerne $a^{(j)}$ være m-dimensionale, hvor $m > n$. Da er formlerne de samme bortset fra, at

$$v_{n+1} \neq 0,$$

og at nu er

$$\sum_{k=1}^{n}(a^{(n+1)},w_k)w_k$$

projektionen af $a^{(n+1)}$ på det n-dimensionale underrum, der udspændes af $(a^{(1)},\ldots,a^{(n)})$, og dermed den nærmeste vektor i euklidisk norm til $a^{(n+1)}$ med afstanden hertil lig med

$$||v_{n+1}||_2.$$

Bevis. $(w_1,\ldots,w_n)$ er en ortonormal basis for underrummet, så en vilkårlig vektor heri kan skrives

$$\lambda_1 w_1 + \ldots + \lambda_n w_n.$$

Afstanden til $a^{(n+1)}$ er (kvadreret)

$$||a^{(n+1)} - (\lambda_1 w_1 + \ldots + \lambda_n w_n)||_2^2 =$$

$$(a^{(n+1)}, a^{(n+1)}) - (a^{(n+1)}, \lambda_1 w_1 + \ldots + \lambda_n w_n) -$$

$$(\lambda_1 w_1 + \ldots + \lambda_n w_n, a^{(n+1)}) + (\lambda_1 w_1 + \ldots + \lambda_n w_n, \lambda_1 w_1 + \ldots + \lambda_n w_n)$$

$$= ||a^{(n+1)}||_2^2 + \sum_j |\lambda_j|^2 -$$

$$\sum_j (\bar{\lambda}_j (a^{(n+1)}, w_j) + \lambda_j (w_j, a^{(n+1)})).$$

Denne funktion af $(\lambda_1, \ldots, \lambda_n)$ er reel og har minimum, hvor alle afledede med hensyn til realdele og imaginærdele er 0. Med $\lambda_j = \alpha_j + i\beta_j$ fås, idet $(a^{(n+1)}, w_j) = a_j + ib_j$, formen

$$||a^{(n+1)}||_2^2 + \sum_j (\alpha_j^2 + \beta_j^2) - \sum_j (2\alpha_j a_j + 2\beta_j b_j),$$

hvis afledede med hensyn til α_j er $2\alpha_j - 2a_j$ osv., så den har minimum for $\alpha_j = a_j$, $\beta_j = b_j$, dvs.

$$\lambda_j = a_j + ib_j = (a^{(n+1)}, w_j).$$

Indsættes disse λ_j i formlen for afstandens kvadrat, fås

$$||a^{(n+1)}||_2^2 - \sum_j (a^{(n+1)}, w_j)(w_j, a^{(n+1)}) = ||v_{n+1}||_2^2.$$

□

Kapitel 3. Iterative metoder til løsning af lineære ligninger.

§1. Problemformuleringen.

Vi havde en lineær funktion f og ønskede at finde x, så

$$f(x) = b .$$

Ved elimination kunne vi bestemme x nogenlunde godt. Men vi kunne ønske os metoder, der var egnede til at forbedre en given approximation til løsningen. Derfor ønsker vi at omformulere problemet til et fixpunktproblem. Vi ønsker en funktion g, der opfylder, at ligningen

$$g(x) = x$$

har én og kun en løsning, x, der samtidig tilfredsstiller ligningen

$$f(x) = b .$$

Hvis g samtidig er afstandsformindskende i en eller anden metrik, dvs. at der findes $\alpha < 1$, så for alle y og z gælder

$$\text{dist}(g(y),g(z)) \leq \alpha \, \text{dist}(y,z) ,$$

så kan vi finde x ved at vælge et punkt x^0 og derefter definere en følge (x^k) ved

$$x^{k+1} = g(x^k) \quad \text{for} \quad k = 0,1,2,\cdots;$$

thi da vil følgen (x^k) konvergere mod x .

En metode, der benytter en sådan funktion, g , kaldes en *iterativ* metode. Hvis g tillige er affin, kaldes metoden en *lineær iterativ metode*.

§2. Lineære iterative metoder.

Lad os tænke os matricen A skrevet som en sum af to matricer B og C, hvor B er invertibel, og vi endda let kan invertere B. Vi skriver nu ligningen

$$A \cdot x = b$$

som

$$(B + C) \cdot x = b$$

eller

$$B \cdot x + Cx = b .$$

Ganger vi med B^{-1}, står der efter omflytning

$$x = B^{-1} b - B^{-1} Cx .$$

Funktionen

$$g : y \to B^{-1} b - B^{-1} Cy$$

har altså x som fixpunkt. Den er endda lineær, så vi mangler bare, at den skal være afstandsformindskende.

Har vi en norm på $\mathbb{C}^n$ eller $\mathbb{R}^n$, kan vi sætte

$$\|-B^{-1}C\| = \max\{\|-B^{-1}Cy\| \mid \|y\| = 1\} .$$

hvis så $\|-B^{-1}C\| < 1$, vil g være afstandsformindskende.

Nu er vi interesserede i, at følgen (x^k) defineret ved x^0 og

$$x^{k+1} = g(x^k) \quad k = 0,1,2,\cdots$$

konvergerer mod en løsning x til ligningen

$$Ax = b .$$

Normbetingelsen ville sikre dette, men vi kan klare os med en svagere betingelse. Der gælder følgende.

Sætning III,2,1 *Givet et punkt* $b \in \mathbb{C}^n$ *og* $n \times n$*-matricer* A, B *og* C, *således at* $A = B + C$, *og* B *er regulær.*
Da er følgende ensbetydende

1) $\forall \lambda \in \mathbb{C} : \det(\lambda B + C) = 0 \Rightarrow |\lambda| < 1.$

2) A *er regulær, der findes netop et* $x \in \mathbb{C}^n$, *så* $A \cdot x = b$, , *og for ethvert* $x^0 \in \mathbb{C}^n$ *vil følgen* (x^k) *defineret ved* $x^{k+1} = B^{-1}b - B^{-1}Cx^k$, $k = 0,1,2,\cdots$ *konvergere mod* x.

Bevis.

2) ⇒ 1).

Vi vil vise, at hvis A er regulær og 1) ikke gælder, så findes $x^0 \in \mathbb{C}^n$, så den følge, der defineres ved

$$x^{k+1} = B^{-1}b - B^{-1}Cx^k , \quad k = 0,1,2,\cdots,$$

divergerer. Når 1) ikke gælder, findes $\lambda \in \mathbb{C}$, så $\det(\lambda B + C) = 0$ og $|\lambda| \geq 1$. Da A er regulær, er $\det(B + C) = \det A \neq 0$. Altså er $\lambda \neq 1$. Da

$$\lambda B + C = B\lambda E + C = B(\lambda E + B^{-1}C) ,$$

og $$\det B \neq 0 ,$$

må $$\det(-B^{-1}C - \lambda E) = 0 .$$

λ er altså egenværdi for $-B^{-1}C$. Til λ findes en egenvektor $\varepsilon^0 \in \mathbb{C}^n$, $\varepsilon^0 \neq 0$. Da A er regulær, findes netop ét x, så $Ax = b$. Lad nu $x^0 = x + \varepsilon^0$. Da bliver

$$x^1 = B^{-1}b - B^{-1}Cx^0 = B^{-1}b - B^{-1}Cx - B^{-1}C\varepsilon^0$$
$$= x + \lambda\varepsilon^0$$

Heraf fås, at

$$x^k = x + \lambda^k\varepsilon^0 .$$

Da $|\lambda| \geq 1$, $\lambda \neq 1$ og $\varepsilon^0 \neq 0$, må følgen (x^k) divergere.

1) $\Rightarrow$ 2).

Af 1) følger for $\lambda = 1$, at $\det A = \det(B + C) \neq 0$, og dermed at A er regulær. Heraf følger, at der findes ét og kun ét $x \in \mathbb{C}^n$, så $Ax = b$, og dermed $x = B^{-1}b - B^{-1}Cx$. Lad nu $x^0 \in \mathbb{C}^n$ være givet, og sæt

$$x^{k+1} = B^{-1}b - B^{-1}Cx^k , \quad k = 0,1,2,\cdots .$$

Vi sætter nu

$$\varepsilon^k = x^k - x ,$$

og får, idet $x = B^{-1}b - B^{-1}Cx$, at

$$\begin{aligned}\varepsilon^k &= B^{-1}b - B^{-1}Cx^{k-1} - B^{-1}b + B^{-1}Cx \\ &= -B^{-1}C(x^{k-1} - x) = -B^{-1}C\,\varepsilon^{k-1}\end{aligned}$$

Ved gentagne anvendelse af formlen fås

$$\varepsilon^k = (-B^{-1}C)^k\,\varepsilon^0 .$$

At vise, at $x^k \to x$, er det samme som at vise, at $\varepsilon^k \to 0$. Vi vælger nu en basis for $\mathbb{C}^n$, så $-B^{-1}C$ får Jordans normalform med hensyn til den ny basis. Vi skriver nu matricen som

$$\Lambda + \Psi ,$$

hvor Λ er diagonalmatricen med egenværdierne for $-B^{-1}C$ i diagonalen, og Ψ er resten af Jordans matrix, altså en matrix med 0'er i diagonalen og 1'taller på visse pladser i øvre bidiagonal. Da er $\Psi^n = 0$. Der gælder, at Λ og Ψ kommuterer

$$\Lambda\Psi = \Psi\Lambda .$$

Thi

$$\begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & & \lambda_n \end{pmatrix} \begin{pmatrix} 0 & \varepsilon_1 & & 0 \\ & 0 & \varepsilon_2 & \\ & & \ddots & \ddots \\ & & & \varepsilon_{n-1} \\ 0 & & & 0 \end{pmatrix} = \begin{pmatrix} 0 & \lambda_1\varepsilon_1 & & 0 \\ & 0 & \lambda_2\varepsilon_2 & \\ & & \ddots & \ddots \\ & & & \lambda_{n-1}\varepsilon_{n-1} \\ 0 & & & 0 \end{pmatrix}$$

og

$$\begin{pmatrix} 0 & \varepsilon_1 & & \mathbf{0} \\ & 0 & \varepsilon_2 & \\ & & \ddots & \ddots \\ & & & \varepsilon_{n-1} \\ \mathbf{0} & & & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 & & & \mathbf{0} \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} = \begin{pmatrix} 0 & \varepsilon_1\lambda_2 & & \mathbf{0} \\ & 0 & \varepsilon_2\lambda_3 & \\ & & \ddots & \ddots \\ & & & \varepsilon_{n-1}\lambda_n \\ \mathbf{0} & & & 0 \end{pmatrix}$$

Om Jordans normalform ved vi netop, at enten er $\varepsilon_i = 0$, eller $\varepsilon_i = 1$, og da er $\lambda_i = \lambda_{i+1}$. Altså stemmer matrixprodukterne overens. Udtrykt i de nye koordinater får vi altså

$$\varepsilon^k = (\Lambda + \Psi)^k \varepsilon^0 = \left(\sum_{j=0}^{k} \binom{k}{j} \Lambda^{k-j} \Psi^j \right) \varepsilon^0 \quad .$$

Men da $\Psi^j = 0$ for $j \geq n$, kan vi skrive

$$\varepsilon^k = \left(\sum_{j=0}^{n-1} \binom{k}{j} \Lambda^{k-j} \Psi^j \right) \varepsilon^0 \quad .$$

For at vise, at $\varepsilon^k \to 0$ for $k \to \infty$, er det derfor nok at vise, at hvert af de n led går mod 0. Vi betragter for fast $j < n$

$$\binom{k}{j} \Lambda^{k-j} \Psi^j \varepsilon^0 \quad .$$

$\Psi^j \varepsilon^0$ er uafhængig af k. Binomialkoefficienten er

$$\binom{k}{j} = \frac{k(k-1)\cdots(k-j+1)}{j!} = P_j(k)$$

et polynomium i k af grad j. Da

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} ,$$

hvor $\lambda_1, \cdots, \lambda_n$ er egenværdier for $-B^{-1}C$ og derfor opfylder $|\lambda_i| < 1$ for alle i

ifølge 1), så er

$$\Lambda^k = \begin{pmatrix} \lambda_1^k & & 0 \\ & \lambda_2^k & \\ 0 & & \lambda_n^k \end{pmatrix} ,$$

hvor λ_i^k går eksponentielt mod 0 for $k \to \infty$. Derfor vil $P_j(k)\lambda_i^k \to 0$ for $k \to \infty$ og alle i. Følgelig må

$$\varepsilon^k = (P_j(k)\Lambda^k)\Psi^j\varepsilon^0 \to 0 \quad \text{for} \quad k \to \infty ,$$

hvilket var, hvad vi behøvede at vise.

§ 3. Jacobis metode og Gauss-Seidel.

En $n \times n$-matrix A skrives som en sum af tre matricer

$$A = L + D + U \ ,$$

hvor L har samme elementer som A under diagonalen (L for "lower") og 0 ellers, D har samme elementer som A i diagonalen (D for "diagonal") og 0 ellers, og U har samme elementer som A over diagonalen (U for "upper") og 0 ellers. L kaldes A's *nedre trekantsmatrix*, U den *øvre* og D for A's *diagonal*.

Jacobis metode fås ved at sætte

$$B = D$$

$$C = L + U,$$

Den kaldes også SIMULTANEOUS REPLACEMENTS. Formlerne bliver med $A = (a_{ij})$ **og** $b = (b_i)$**:**

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^k\right).$$

Metoden konvergerer, når egenværdierne for $-D^{-1}(L+U)$ **er mindre end 1, hvilket f. eks. er tilfældet, når en af matrixnormerne opfylder**

$$||D^{-1}(L + U)|| < 1.$$

Bruges $||\ ||_1$ **eller** $||\ ||_\infty$**, betyder det, at diagonalen i A skal dominere række- eller søjlesummerne i matricen. Altså f. eks.**

$$|a_{ii}| \geq \sum_{j\neq i} |a_{ij}|, \quad \text{for alle } i.$$

En lineær iterativ metode fås ved at sætte

$$B = L + D$$

$$C = U$$

i den generelle form fra §2. Den derved definerede metode kaldes *Gauss-Seidel* eller *successive replacements*. Den sidste betegnelse skyldes, at ved brug af metoden finder vi ikke $(L + D)^{-1}$, men løser ligningerne i x^{k+1}

$$(L + D)x^{k+1} = b - Ux^k .$$

Dette er let i rækkefølgen x_1^{k+1}, x_2^{k+1}, $\cdots$ Vi får for $A = (a_{ij})$, $b = (b_i)$ formlen

$$x_i^{k+1} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right)$$

til successiv anvendelse for $i = 1,2,\cdots,n$.

Naturligvis har vi hverken sikkerhed for, at B er regulær, eller at egenværdierne for $-B^{-1}C$ er numerisk mindre end 1. Men det gælder i et vigtigt specialtilfælde.

Sætning III,3,1 *Lad matricen* A *være spaltet som ovenfor i* $A = L + D + U$. *Hvis* A *er reel, symmetisk og positiv definit, så er* $L + D$ *regulær og samtlige egenværdier for* $-(L+D)^{-1}U$ *er numerisk mindre end* 1.

Bevis.

At A er reel og positiv definit betyder, at den kvadratiske form $\overline{z}\,Az$ for $z \in \mathbb{C}^n$ er reel og har minimum i 0. Specielt for $z = (0,\cdots,0,1,0,\cdots,0)$ hvor 1-tallet står på den i-te plads, gælder altså

$$\overline{z}\,Az > 0 .$$

Men $\overline{z}\,Az = a_{ii}$, det i-te diagonalelement. Derfor er D regulær og positiv definit, og dermed er $(L+D)$ regulær; thi

$$\det(L+D) = \det D = \prod_{i=1}^{n} a_{ii} > 0 .$$

Lad nu $\lambda \in \mathbb{C}$ være en egenværdi for $(L+D)^{-1}U$. Da findes $z \in \mathbb{C}^n \setminus \{0\}$ egenvektor for λ, så

$$(L+D)^{-1}Uz = \lambda z ,$$

heraf fås

$$Uz = \lambda(L+D)z .$$

Nu er

$$(*) \qquad Az = (L+D+U)z = (L+D)z + Uz = (1+\lambda)(L+D)z .$$

Konjugering giver

$$A\overline{z} = (1+\overline{\lambda})(L+D)\overline{z} ,$$

og transponering, da $L' = U$ og A er symmetrisk, giver

$$\overline{z}A = (1+\overline{\lambda})\overline{z}(U+D) .$$

Ganges fra højre med z , fås

$$\overline{z}\, Az = (1+\overline{\lambda})\overline{z}\, Uz + (1+\overline{\lambda})\overline{z}\, Dz$$

$$= (1+\overline{\lambda})\overline{z}\, \lambda(L+D)z + (1+\overline{\lambda})\overline{z}\, Dz \ ,$$

Heraf fås

$$(1+\lambda)\overline{z}\, Az = (1+\lambda)(1+\overline{\lambda})\lambda\overline{z}(L+D)z + |1+\lambda|^2\overline{z}\, Dz \ .$$

På den anden side fås af (*) ved at gange fra venstre med $\overline{z}$

$$(1+\overline{\lambda})\lambda\ \overline{z}\, Az = (1+\lambda)(1+\overline{\lambda})\lambda\overline{z}(L+D)z \ .$$

Herefter kan $L + D$ elimineres ved subtraktion

$$(1+\lambda-(1+\overline{\lambda})\lambda)\ \overline{z}\, Az = |1+\lambda|^2\ \overline{z}\, Dz \ ,$$

hvoraf

$$(1-|\lambda|^2)\ \overline{z}\, Az = |1+\lambda|^2\ \overline{z}\, Dz \ .$$

Da $\overline{z}\, Dz > 0$ og $\overline{z}\, Az > 0$, må enten

$$\lambda = -1$$

eller

$$1 - |\lambda|^2 > 0$$

hvis $\lambda = -1$, følger af (*) , at $Az = 0$, og dermed

$$\overline{z}\, Az = 0$$

i modstrid med, at $z \neq 0$ og A positiv definit.

Altså må

$$|\lambda| < 1 \ .$$

Bemærkning. Er A en vilkårlig reel $n \times n$-matrix, og er A' dens transponerede, så er $A'A$ reel, symmetrisk og positiv semidefinit. Hvis A er regulær, så er $A'A$ positiv definit. Lad $C = A'A$ have elementer c_{ij}, $i,j = 1,\cdots,n$. Så er

$$c_{ij} = \sum_{k=1}^{n} a_{ki}\, a_{kj} = c_{ji} \ ,$$

hvorfor $A'A$ er symmetrisk. Specielt er

$$c_{ii} = \sum_{k=1}^{n} a_{ki}^2 \geq 0 \ ,$$

og hvis $c_{ii} = 0$, må samtlige $a_{ki} = 0$ for $k = 1,\cdots,n$. Altså en hel række i matricen A er 0. Er A regulær, så er ingen rækker 0. For $z \in \mathbb{C}^n$ er

$$\overline{z}\, A'A\, z = (A\overline{z})'(Az) = \langle Az, Az\rangle \geq 0 \ ,$$

hvor lighedstegnet gælder, når $Az = 0$. Altså er $A'A$ positiv semidefinit, og positiv definit, når $Az = 0$ kun har løsningen $z = 0$, altså når A er regulær.

Skal vi løse et vilkårligt reelt ligningssystem med regulær koefficientmatrix, A,

$$Ax = b \ ,$$

kan vi ved at gange med A' få et ligningssystem,

$$A'Ax = A'b \ ,$$

hvorpå Gauss-Seidel kan anvendes.

§4. SOR = Successive Overrelaxation.

I stedet for at lade hele diagonalen , D , gå til matricen B i opspaltningen af A , kan man lade en passende del af D gå til B ; resten må så gå til C . Lad $\omega \in \mathbb{R}$, $\omega \neq 0$. Sæt nu

$$B = (\frac{1}{\omega} D+L) \ , \ C = U + (1 - \frac{1}{\omega}) D \ .$$

Ved passende valg af ω opnår man, at konvergensen går væsentlig hurtigere end for Gauss-Seidel.

Løsning af ligningerne

$$Bx^{k+1} = b - Cx^k$$

bliver

$$x_i^{k+1} = (1-\omega)x_i^k + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k \right) \ ,$$

til successiv anvendelse for $i = 1,2,\cdots,n$.
For $\omega = 1$ får vi Gauss-Seidel igen.

Hvis A er reel, symmetrisk og positiv definit, så er B regulær. Foretager vi en beregning svarende til beviset for sætning III,3,1 får vi

$$(1-|\lambda|^2) \ \overline{z} \ Az = (\frac{2}{\omega} - 1)|1+\lambda|^2 \ \overline{z} \ Dz \ .$$

Da vi ønsker at slutte, at

$$|\lambda| < 1 \ ,$$

må vi altså forudsætte, at

$$\frac{2}{\omega} - 1 > 0 .$$

Det er netop tilfældet, når

$$0 < \omega < 2 .$$

Erfaringen tyder på, at det bedste ω ligger mellem 1 og 2, nærmest ved 1, når n er lille, og nærmere 2, når n er stor. Desuden står man sig ved at vælge ω lidt større end den bedste værdi frem for lidt mindre.

Lad os antage, at matricen A kan skrives på formen

$$A = \begin{pmatrix} D_1 & G \\ F & D_2 \end{pmatrix} ,$$

hvor D_1 og D_2 er diagonalmatricer og F og G vilkårlige (rektangulære) matricer. Denne form er ikke så speciel, som den umiddelbart kan give indtryk af. Ved diskretisering af Laplace-operatoren til løsning af et Dirichlet-problem kan man ordne ligningerne i en rækkefølge, hvorved koefficientmatricen får netop denne form. Hertil kommer, at det er under denne og analoge former for koefficientmatricer, man kan analysere virkningen af ω. Lad nu μ være en egenværdi for **Jacobis metode, dvs. for** $-D^{-1}(L+U)$, **altså lad**

$$\det(-D^{-1}(L+U) - \mu E) = 0 .$$

Så er $$\det(L + U + \mu D) = 0$$

eller
$$\det\begin{pmatrix} \mu D_1 & G \\ F & \mu D_2 \end{pmatrix} = 0.$$

Da er μ^2 egenværdi for Gauss-Seidel, thi ved multiplikation af de p første søjler og de q sidste rækker med μ fås
$$\det\begin{pmatrix} \mu^2 D_1 & G \\ \mu^2 F & \mu^2 D_2 \end{pmatrix} = 0$$

eller $\det(U + \mu^2(D + L)) = 0,$

dvs. $\det(-(D+L)^{-1}U - \mu^2 E) = 0.$

Da $|\mu^2| < 1$, må $|\mu| < 1$ og Jacobi konvergent.

Lad nu λ være en egenværdi for givet ω for
$$-(\frac{1}{\omega} D + L)^{-1}(U + (1 - \frac{1}{\omega})D) .$$

Dvs.
$$\det(-(\frac{1}{\omega} D + L)^{-1}(U + (1 - \frac{1}{\omega})D) - \lambda E) = 0 ,$$

eller
$$\det(\omega U + (\omega-1)D + \lambda(\omega L+D)) = 0.$$

Altså
$$\det\begin{pmatrix} (\lambda+\omega-1)D_1 & \omega G \\ \omega\lambda F & (\lambda+\omega-1)D_2 \end{pmatrix} = 0 .$$

Som ovenfor får vi heraf
$$\lambda^{\frac{n}{2}}\,\omega^n \det\begin{pmatrix} \frac{\lambda+\omega-1}{\omega\sqrt{\lambda}} D_1 & G \\ F & \frac{\lambda+\omega-1}{\omega\sqrt{\lambda}} D_2 \end{pmatrix} = 0 .$$

Altså er enten $\lambda = 0$, eller λ må opfylde, at

$$(*) \qquad \frac{\lambda+\omega-1}{\omega\sqrt{\lambda}} = \mu ,$$

hvor μ er egenværdi for **Jacobi.** Med matricen A er μ givet, så ligningen (*) bestemmer implicit λ som funktion af ω . For $\omega = 1$ fås åbenbart $\lambda = \mu^2$. Har vi λ som funktion af ω, kan vi minimalisere $|\lambda|$. Herved bestemmes det optimale ω som funktion af μ .

$\sqrt{\lambda}$ **findes af** (*) **ved omformningen**

$$(**) \qquad (\sqrt{\lambda})^2 - \omega\mu\sqrt{\lambda} + \omega - 1 = 0.$$

Denne ligning i $\sqrt{\lambda}$ **har diskriminant** $\omega^2\mu^2 - 4\omega + 4$, **som er** 0 **for**

$$\omega = \frac{4 \pm \sqrt{16 - 16\mu^2}}{2\mu^2} = \frac{2 \pm 2\sqrt{1-\mu^2}}{\mu^2},$$

hvoraf kun

$$\omega_0 = \frac{2-2\sqrt{1-\mu^2}}{\mu^2} = \frac{\frac{4}{\mu^2}}{\frac{2+2\sqrt{1-\mu^2}}{\mu^2}} = \frac{2}{1+\sqrt{1-\mu^2}}$$

(**røddernes produkt er** $\frac{4}{\mu^2}$) **ligger i intervallet** $[0,2]$.

Det betyder, at for $2 > \omega > \omega_0$ **er der to komplekse konjugerede rødder,** $\sqrt{\lambda} = \alpha$ **og** $\sqrt{\lambda} = \bar{\alpha}$, **hvoraf**

$$|\lambda| = \sqrt{\lambda\bar{\lambda}} = \sqrt{\lambda}\cdot\sqrt{\bar{\lambda}} = \alpha\cdot\bar{\alpha} = \omega - 1,$$

altså to rødder med samme norm lig med $\omega - 1 < 1$. **For** $\omega = \omega_0$ **er der kun én reel dobbeltrod, som må være** $\sqrt{\lambda} = \sqrt{\omega_0 - 1}$, **altså** $\lambda = \omega_0 - 1$.

For $\omega \leq \omega_0$ **kan** (**) **skrives som**

$$\omega = \frac{\lambda - 1}{\mu\sqrt{\lambda} - 1},$$

som er interessant for $0 \leq \lambda \leq 1$. **Den har toppunkt for**

$$0 = \frac{d\omega}{d\lambda} = \frac{(\mu\sqrt{\lambda} - 1) - (\lambda - 1)\cdot\mu\cdot\frac{1}{2\sqrt{\lambda}}}{(\mu\sqrt{\lambda} - 1)^2},$$

dvs.

$$\mu\sqrt{\lambda} - 1 - \mu\frac{\sqrt{\lambda}}{2} + \frac{\mu}{2\sqrt{\lambda}} = 0$$

$$\mu\frac{\sqrt{\lambda}}{2} + \mu\frac{1}{2\sqrt{\lambda}} - 1 = 0$$

$$\mu\cdot\lambda - 2\sqrt{\lambda} + \mu = 0$$

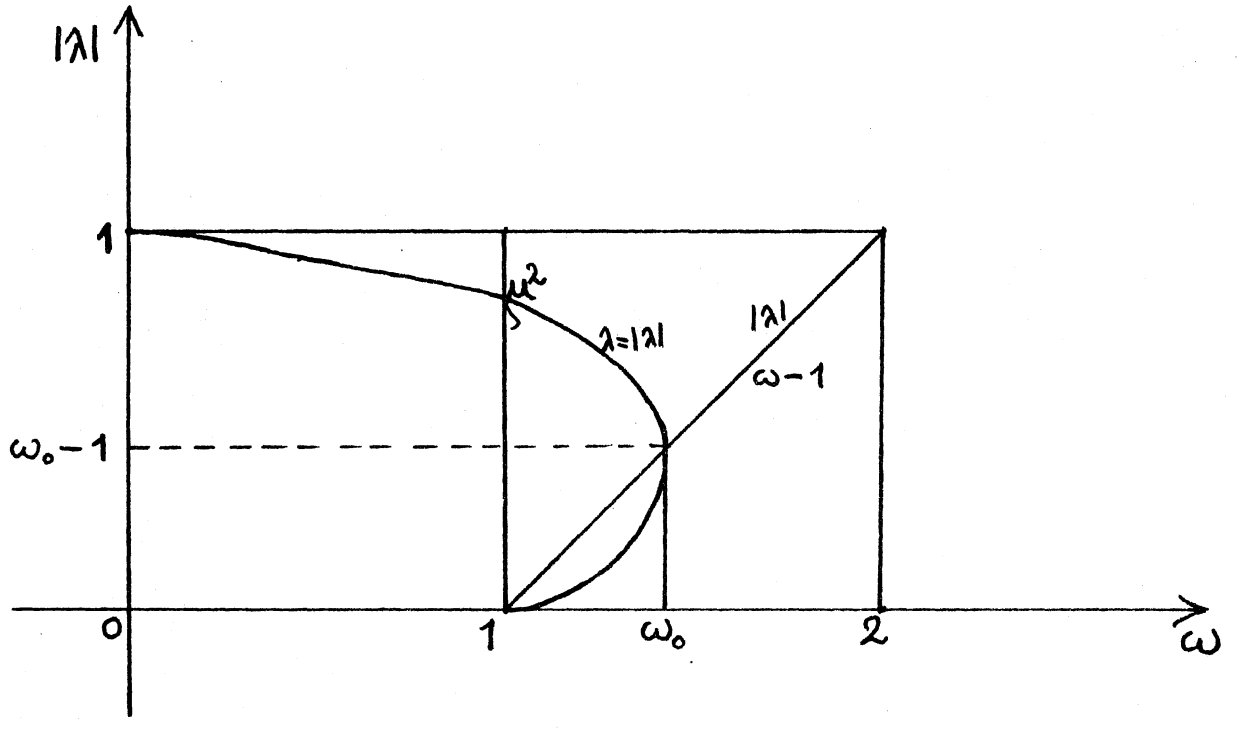
$$\sqrt{\lambda} = \frac{2 \pm \sqrt{4 - 4\mu^2}}{2\mu} = \frac{1 \pm \sqrt{1 - \mu^2}}{\mu},$$

hvoraf

$$\lambda = \left(\frac{1 \pm \sqrt{1 - \mu^2}}{\mu}\right)^2 = \frac{1+1-\mu^2 \pm 2\sqrt{1-\mu^2}}{\mu^2}$$

$$= \frac{2 \pm 2\sqrt{1-\mu^2}}{\mu^2} - 1 = \begin{cases} \omega_0 - 1 \\ \text{rod} > 1. \end{cases}$$

Sammenhængen mellem $|\lambda|$ og ω givet ved (**) ser grafisk således ud



Vælger vi nu ω_0 svarende til den største forekommende Jacobi-egenværdi μ, bliver egenværdierne, der svarer til de andre værdier af μ, komplekse, parvis konjugerede med norm $\omega_0 - 1$. Men egenværdien $\lambda = \omega_0 - 1$ bliver en dobbelt rod, så konvergenshastigheden for den bliver som $k \cdot \lambda^{k-1}$ (jvf. § 2). Det kan derfor som regel betale sig at vælge $\omega > \omega_0$, hvorved alle rødderne bliver komplekse, men enkelte, så hastigheden bliver som $(\omega - 1)^k$.

Bemærkning. En let omskrivning giver

$$\omega_0 = \frac{2}{1 + \sqrt{1 - \mu^2}} = 1 + \frac{\mu^2}{\left(1 + \sqrt{1 - \mu^2}\right)^2},$$

hvoraf

$$\lambda = \left(\frac{\mu}{1 + \sqrt{1 - \mu^2}}\right)^2,$$

som kaldes Youngs formler efter D. M. Young (1950).

Kapitel IV.

Nulpunktsbestemmelse.

Iterative metoder til løsning af ikke lineære ligninger.

§1. Problemformuleringen.

Vi har givet en funktion

$$f: \mathbb{R}^n \to \mathbb{R}^n$$

og ønsker at bestemme et punkt $x^* \in \mathbb{R}^n$, så

$$f(x^*) = 0,$$

med andre ord en *løsning* til ligningen eller *rod* i ligningen

$$f(x) = 0 .$$

Problemet er ækvivalent med fixpunktproblemet, hvor der er givet en funktion

$$g: \mathbb{R}^n \to \mathbb{R}^n ,$$

og vi ønsker at bestemme et *fixpunkt* $x^* \in \mathbb{R}^n$, så

$$g(x^*) = x^* .$$

Vi kan sætte $g(x) = x-f(x)$, eller omvendt $f(x) = x-g(x)$, alt efter det givne problem.

Er nu g afstandsformindskende i en konveks omegn af x^*, kan problemet løses ved iteration af g. Hvis x^0 ligger i omegnen, vil følgen $x^{k+1} = g(x^k)$, $k = 0,1,2,\ldots$ konvergere mod x^*.

Problemerne er næsten ækvivalente med problemet at finde et *minimum* for en funktion

$$h: \mathbb{R}^n \to \mathbb{R}.$$

Er nemlig h differentiabel, vil vi ved

$$f_i(x) = \frac{\partial h(x)}{\partial x_i}$$

definere en funktion $f = (f_1,\ldots,f_n)$,

$$f: \mathbb{R}^n \to \mathbb{R}^n ,$$

som er 0 i h's minima.

Og omvendt, hvis vi blot har en funktion

$$\varphi: \mathbb{R}^n \to \mathbb{R},$$

som har netop ét minimum i 0, f.eks.

$$\varphi(x) = x_1^2+\ldots+x_n^2 ,$$

så vil funktionen

$$h = \varphi \circ f$$

have minima netop i f's nulpunkter.

§2. De parallelle korders metode.

Givet f og nulpunktsproblemet. For en vilkårlig matrix A sættes

$$g(x) = x - A^{-1}f(x).$$

Vi håber, at for passende valg af A er g afstandsformindskende, så iteration af g løser problemet.

Geometrisk er x^{k+1} løsning til ligningen

$$0 = A(x-x^k) + f(x^k),$$

altså skæringen mellem x-aksen (hyperplan) og hyperplanen $y = A(x-x^k) + f(x^k)$.

For $n = 1$ ser det således ud med $A = (\alpha)$:

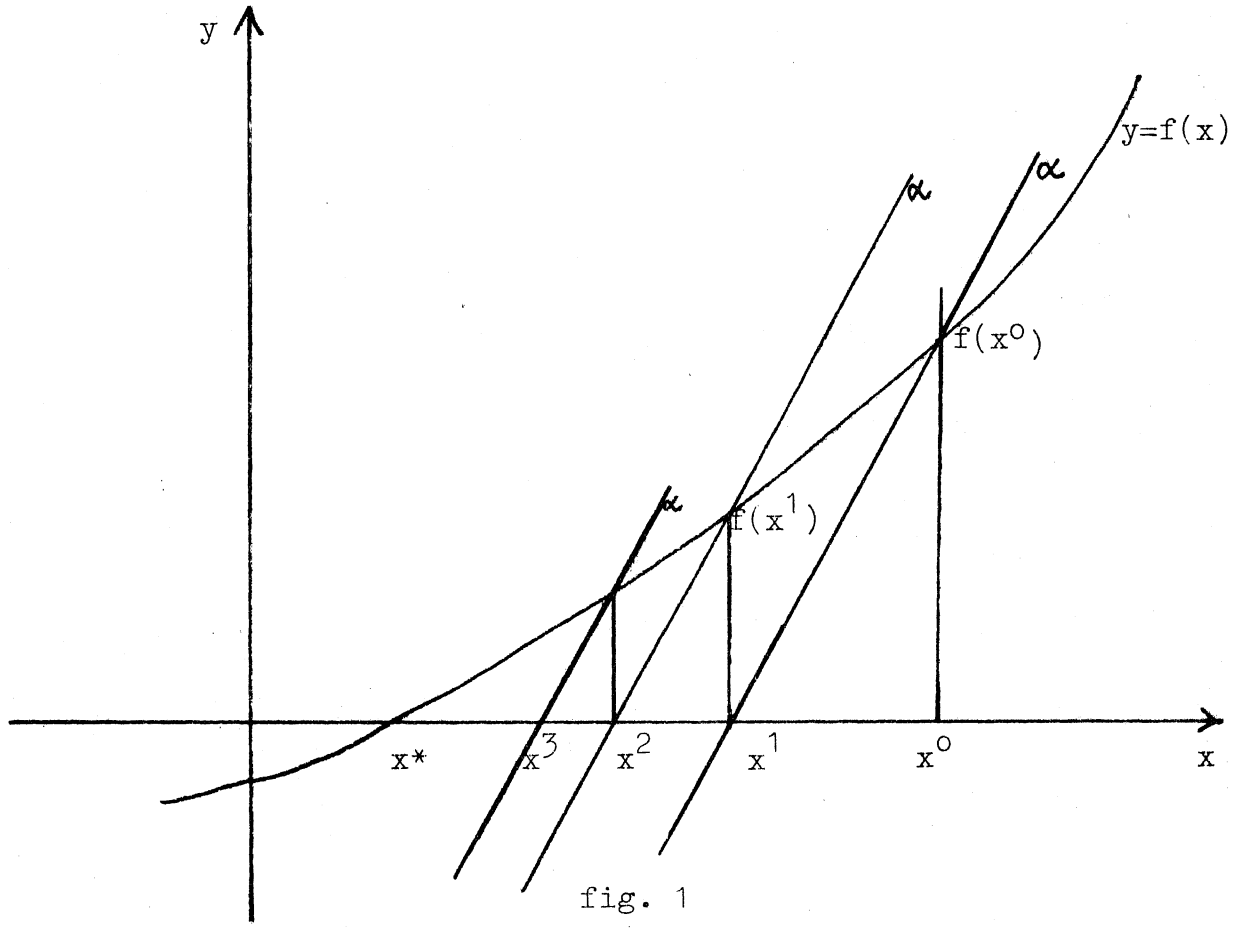


fig. 1

§3. Newtons metode.

Af fig. 1 synes at fremgå, at det vil være fordelagtigt, om α er nær differentialkvotienten for f.

Vi sætter derfor

$$g(x) = x - f'(x)^{-1} f(x),$$

hvor $f'(x)$ er funktionalmatricen for f. Også g har f's nulpunkter som fixpunkter.

For $n = 1$ ser det således ud

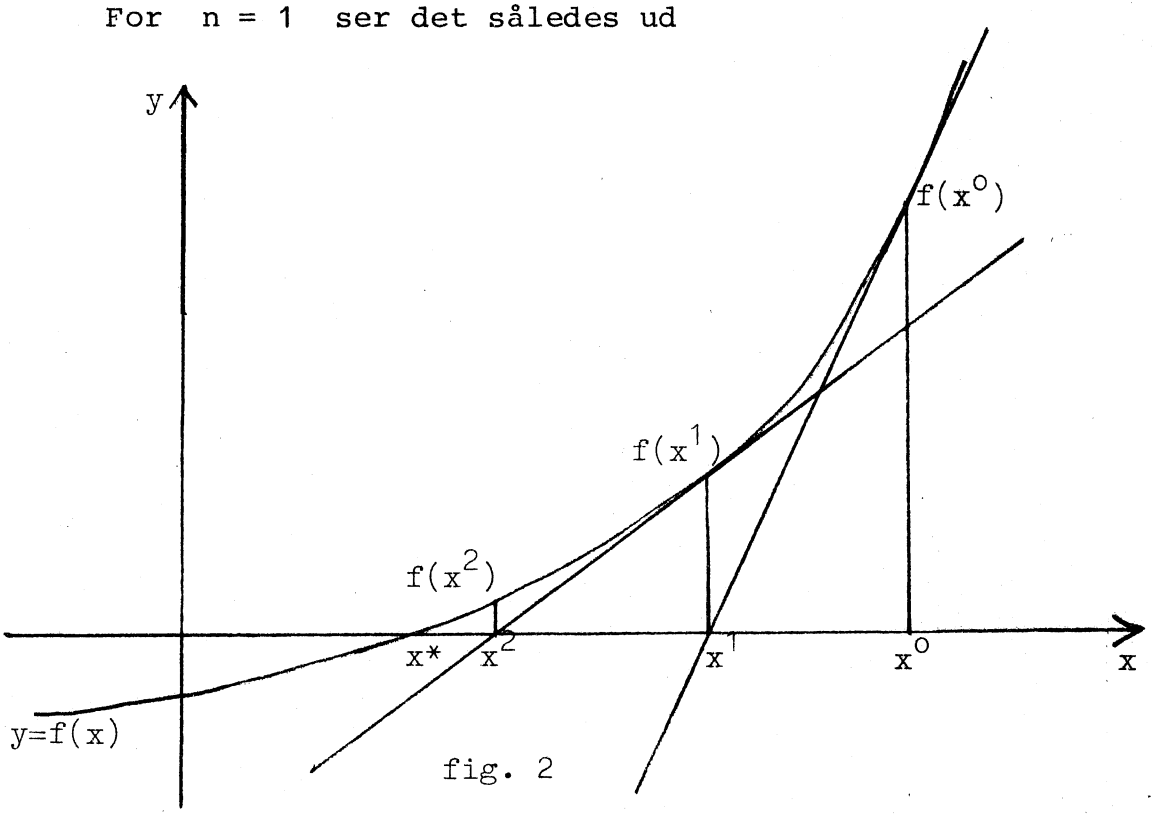


fig. 2

Da g er afstandsformindskende, hvis $|g'| < 1$, finder vi konvergensbetingelsen, stadig for $n = 1$,

$$|g'(x)| = \left| 1 - \frac{f'(x)^2 - f(x)\,f''(x)}{f'(x)^2} \right| < 1$$

eller

$$\left|\frac{f(x)f''(x)}{f'(x)^2}\right| < 1.$$

Denne ulighed er opfyldt i omegnen af x^*, hvis $f'(x^*) \neq 0$ og $f''(x)$ begrænset i omegnen af x^*, f.eks. $|f''(x)| < M$ for $|x-x^*| < \delta$. Thi da skal gælde

$$|f(x)| < \frac{f'(x)^2}{|f''(x)|} ,$$

som er opfyldt for $|x-x^*|$ så lille, at $|f'(x)| \geq \frac{|f'(x^*)|}{2}$, og $|f''(x)| < M$, og

$$|f(x)| < \frac{f'(x^*)^2}{4M} .$$

I højere dimensioner er det praktiske problem enten at invertere matricen $f'(x^k)$ eller at løse ligningssystemet i x^{k+1}:

$$(*) \qquad f'(x^k)x^{k+1} = f'(x^k)x^k - f(x^k) .$$

På grund af besværet ved både at beregne $f'(x^k)$ og løse ligningssystemet (*) findes utallige modifikationer, der dels diskretiserer $f'(x^k)$, altså br uger differenskvotienter, dels nøjes med grove tilnærmelser til løsningen for (*).

P.S. I litteraturen betegnes Newtons metode ofte som "Newton-Raphson".

§ 4. Eksempel på Newtons metode: Invers matrix.

Lad A være en regulær matrix. Så er A^{-1} nulpunkt for funktionen

$$f(X) = X^{-1} - A.$$

$f'(X)$ er en funktionalmatrix, der afbilder vektorrummet af kvadratiske matricer $\mathbb{R}^{n^2}$ eller $\mathbb{C}^{n^2}$ ind i sig selv. Den kan defineres ved

$$K = f'(X)H = - X^{-1}HX^{-1},$$

thi $$f'(X)\frac{\Delta X}{||\Delta X||} \simeq \frac{(f(X+\Delta X) - f(X))}{||\Delta X||} = \frac{((X+\Delta X)^{-1}-X^{-1})}{||\Delta X||} =$$

$$= \frac{-(X+\Delta X)^{-1}(\Delta X \cdot X^{-1})}{||\Delta X||} \longrightarrow -X^{-1}HX^{-1}$$

for $\Delta X \longrightarrow 0$ mens $\frac{\Delta X}{||\Delta X||} \longrightarrow H$; $(||H|| = 1)$.

Men så er $H = f'(X)^{-1}K = -XKX$, så Newtons metode bliver iteration af

$$\begin{aligned} g(X) &= X - f'(X)^{-1}f(X) \\ &= X + X(X^{-1}-A)X \\ &= X + X(E - AX) \\ &= 2X - XAX. \end{aligned}$$

Iterationen konvergerer, når X er i nærheden af A^{-1}, f. eks. hvis X er fundet ved elimination som i kapitel II § 3.

Mere præcist gælder følgende

Sætning IV,4,1. Hvis X_o opfylder uligheden

$$q := || E - AX_o || < 1$$

for en vilkårlig matrix-norm og for A regulær, så konvergerer følgen X_t defineret ved

$$X_{t+1} = 2X_t - X_tAX_t, \qquad t = 0,1,2,...$$

mod A^{-1} med fejlvurderingerne

$$|| X_t - A^{-1} || \leq \frac{|| X_o||}{1 - q} || E - AX_t || \leq \frac{|| X_o ||}{1-q} q^{(2^t)}$$

$$|| X_t - A^{-1} || \leq \frac{|| X_t ||\cdot|| E - AX_t ||}{1 - || E - AX_t ||}$$

Bevis. Vi kan skrive

$$X_{t+1} = X_t(E + R_t),$$

hvor $$R_t = E - AX_t.$$

Da gælder

$$\begin{aligned} R_t &= E - AX_{t-1}(E + R_{t-1}) \\ &= E - AX_{t-1} - AX_{t-1}R_{t-1} \\ &= R_{t-1} - AX_{t-1}R_{t-1} \\ &= (E - AX_{t-1})R_{t-1} = R_{t-1}^2. \end{aligned}$$

Ved iteration fås

$$R_t = R_o^{(2^t)}, \qquad t = 0,1,2,...$$

Da $|| R_o || = q < 1$, må $R_t \longrightarrow 0$ for $t \longrightarrow \infty$, så

$$X_t = A^{-1}(E - R_t) \longrightarrow A^{-1} \quad \text{for} \quad t \longrightarrow \infty.$$

Af $X - A^{-1} = -A^{-1}(E - AX)$ fås

$$(*) \qquad || X - A^{-1}|| \leq ||A^{-1}||\cdot|| E - AX ||,$$

og af $A^{-1} = X + A^{-1} - X$ fås

$$\begin{aligned} || A^{-1} || &\leq || X || + || A^{-1} - X || \\ &\leq || X || + || A^{-1} ||\cdot|| E - AX ||, \end{aligned}$$

hvoraf

$$|| A^{-1} || \leq \frac{|| X ||}{1 - || E - AX ||}$$

(for $|| E - AX || < 1$), samt ved indsættelse i (*)

$$|| X - A^{-1} || \leq \frac{|| X ||}{1 - || E - AX ||} || E - AX ||.$$

Med $X = X_t$ er dette den anden vurdering i sætningen. Sættes $X = X_o$ i vurderingen af A^{-1}, og indsættes denne vurdering i (*), med det sidste $X = X_t$, fås den første vurdering i sætningen.

Af

$$X_t - A^{-1} = - A^{-1} R_t$$

fås

$$\begin{aligned} || X_t - A^{-1} || &\leq || A^{-1} || \cdot || R_t || \\ &\leq \frac{|| X_o ||}{1 - q} \cdot || R_o^{(2^t)} || \\ &\leq \frac{|| X_o ||}{1 - q} q^{(2^t)}. \end{aligned}$$

□

§5. Regule falsi og sekantmetoden.

Lad os for overskuelighedens skyld indskrænke os til dimension 1. Lad $(h^k) \to 0$, for $k \to \infty$, og sæt differenskvotienten til erstatning af $f'(x^k)$ til

$$\frac{f(x^k+h^k)-f(x^k)}{h^k} .$$

Vælges specielt $h^k = \overline{x} - x^k$, hvor $\overline{x}$ er fast, kaldes metoden *regula falsi*, og vælges $h^k = x^{k-1}-x^k$, kaldes metoden *sekantmetoden*. Der er dog dem, der kalder sekantmetoden for "regula falsi".

Geometrisk består metoderne i sukcessiv løsning af ligningerne

$$\frac{f(x^k+h^k)-f(x^k)}{h^k}\,(x-x^k) + f(x^k) = 0$$

hvilket illustreres på figurerne 3 og 4.

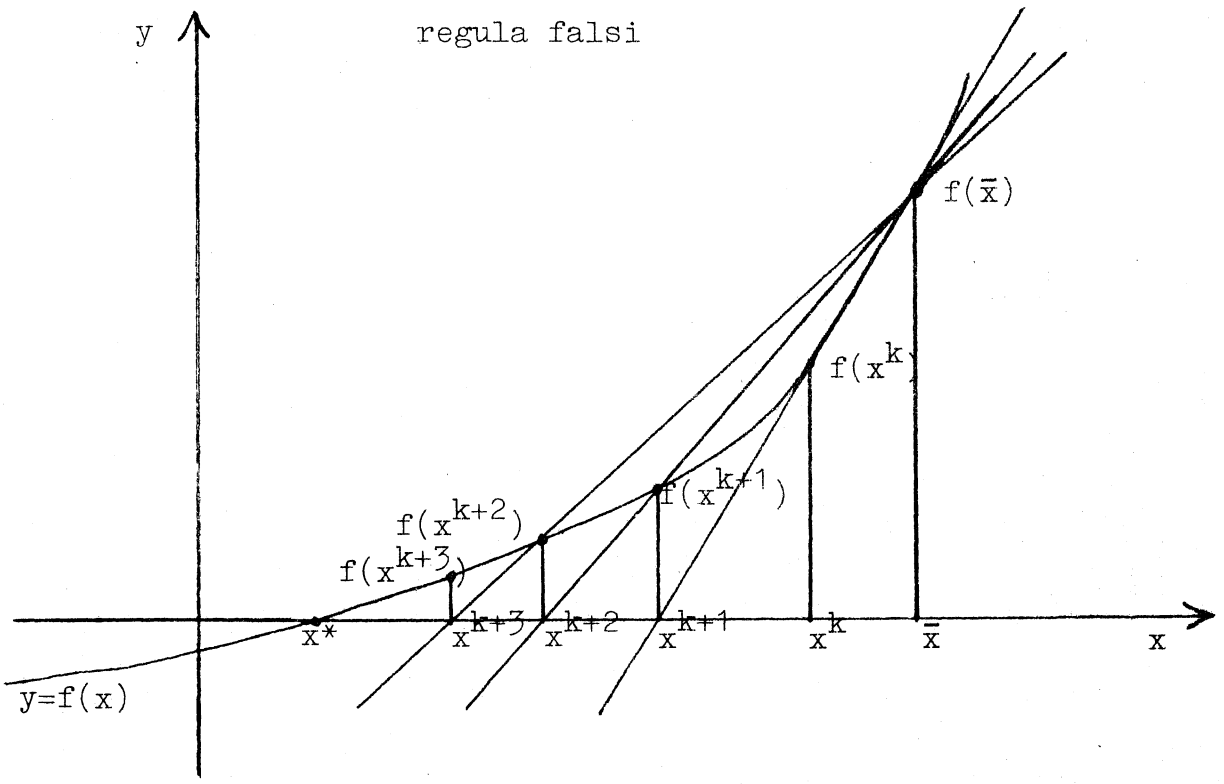

y
regula falsi
f(x̄)
f(x^k)
f(x^{k+1})
f(x^{k+2})
f(x^{k+3})
x*
x^{k+3}
x^{k+2}
x^{k+1}
x^k
x̄
x
y=f(x)

fig. 3

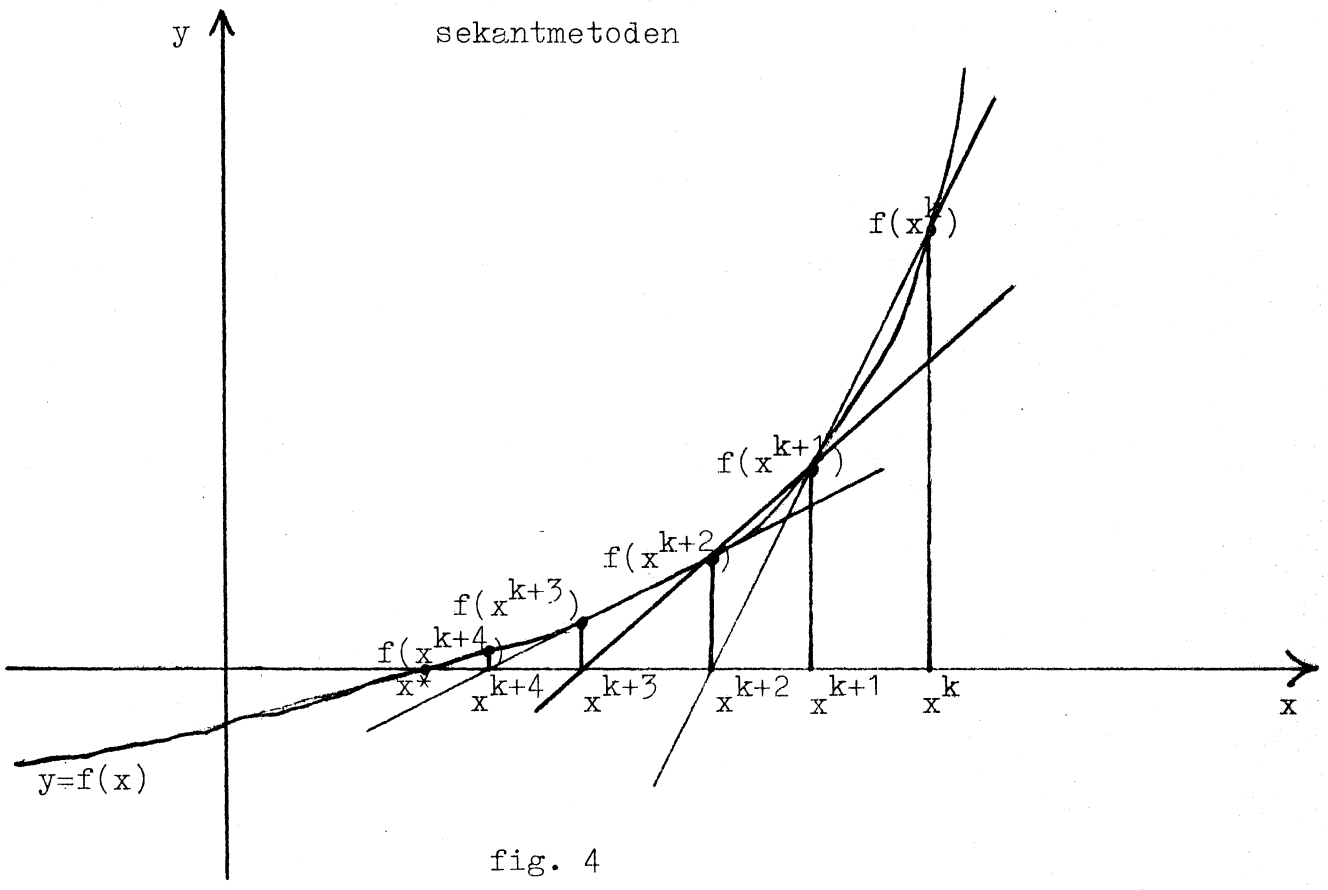

y
sekantmetoden
f(x^k)
f(x^{k+1})
f(x^{k+2})
f(x^{k+3})
f(x^{k+4})
x*
x^{k+4}
x^{k+3}
x^{k+2}
x^{k+1}
x^k
x
y=f(x)

fig. 4

Kapitel V. EGENVÆRDIER FOR REELLE, SYMMETRISKE MATRICER.

§ 1. Egenværdiernes kontinuitet.

Sætning V,1,1. *En reel, symmetrisk matrix har kun reelle egenværdier.*

Bevis. Lad A være matricen, $\lambda \in \mathbb{C}$ en egenværdi og $z \in \mathbb{C}^n$ en tilhørende egenvektor, ↄ:

$$Az = \lambda z \ .$$

Konjugering og transponering giver

$$\overline{z}A = \overline{\lambda z} \ ,$$

hvorefter $\overline{z}Az$ kan beregnes efter den associative lov:

$$\overline{\lambda} \, \| z \|_2^2 = (\overline{\lambda}\overline{z})z = \overline{z}Az = \overline{z}(\lambda z) = \lambda \, \| z \|_2^2$$

Da z som egenvektor ikke er 0 , sluttes at $\lambda = \overline{\lambda}$ eller at $\lambda \in \mathbb{R}$.

Bevis slut.

Lad A være en reel, symmetrisk matrix. Vi indicerer egenværdierne efter størrelse, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Lad $w_1, \cdots, w_n$ være et hertil svarende ortonormalsystem af egenvektorer. For en vilkårlig vektor $x \neq 0$ defineres *Rayleigh-kvotienten*

$$\rho_A(x) = \frac{(Ax,x)}{\| x \|_2^2} \ ,$$

hvor $(\ ,\ )$ er det sædvanlige indre produkt i $\mathbb{R}^n$ og $\|x\|_2^2 = (x,x)$.

Skriver vi $x = \sum_j (x,w_j)w_j$, finder vi umiddelbart $Ax = \sum_j \lambda_j (x,w_j)w_j$. Heraf fås

$$\rho_A(x) = \frac{(\sum_j \lambda_j (x,w_j)w_j,\ \sum_j (x,w_j)w_j)}{(\sum_j (x,w_j)w_j,\ \sum_j (x,w_j)w_j)} = \frac{\sum_j \lambda_j (x,w_j)^2}{\sum_j (x,w_j)^2} .$$

Af denne fremstilling ses, at

$$\max_{x\neq 0} \rho_A(x) = \lambda_1 .$$

$\leqq$ er oplagt, og $\geqq$ fås for $x = w_1$.

Ved hjælp af Rayleigh-kvotienten kan vi endda finde samtlige egenværdier. Lad M_m være underrummet i $\mathbb{R}^n$ udspændt af $w_j,\cdots,w_n$ hvor $m = n+1-j$. M_m har da dimension m . Som ovenfor fås

$$\max_{x\in M_m} \rho_A(x) = \lambda_j .$$

Vælger vi et andet underrum i $\mathbb{R}^n$ af dimension m , M , bliver det tilsvarende maximum ikke mindre

$$\max_{x\in M_m} \rho_A(x) \leqq \max_{x\in M} \rho_A(x) .$$

Vi har derfor *Courant-Weyls Minimax-princip* til karakterisering af egenværdierne

$$\lambda_j = \min_{\dim M=m} \max_{x\in M} \rho_A(x) \quad \text{for} \quad m = n+1-j .$$

For at finde de numeriske værdier af egenværdierne bemærker vi, at A^2 har de samme egenvektorer som A, men egenværdierne er netop $\lambda_1^2,\cdots,\lambda_n^2$. Thi

$$A^2w_j = A(\lambda_j w_j) = \lambda_j(Aw_j) = \lambda_j(\lambda_j w_j) = \lambda_j^2 w_j .$$

For eksempel ønsker vi at finde den numerisk største egenværdi. Rayleigh-kvotienten for A^2 er

$$\rho_{A^2}(x) = \frac{(A^2x,x)}{\|x\|_2^2} = \frac{(Ax,Ax)}{\|x\|_2^2} = \frac{\|Ax\|_2^2}{\|x\|_2^2} ,$$

hvor andet lighedstegn skyldes, at A er symmetrisk. Vi får altså

$$\max_j \lambda_j^2 = \max_x \rho_{A^2}(x) = \max_{\mathbf{x}} \frac{\|Ax\|_2^2}{\|x\|_2^2} ,$$

hvoraf

$$\max_j |\lambda_j| = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_* .$$

Vi har altså også en smuk karakteristik af $\|\ \|_*$.

Lad nu A og B være reelle, symmetriske matricer med egenværdier $\lambda_j(A)$ og $\lambda_j(B)$, $j = 1,\cdots,n$.

Sætning V.1.2. *For* $j = 1,\cdots,n$ *og* $p = *,1,2,\infty$ *gælder*

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A-B\|_p .$$

Med andre ord, ikke alene afhænger egenværdierne kontinuert af matricen, men vi kan vurdere afvigelserne bekvemt ved matrixnormerne

Bevis.

$$\rho_A(x) - \rho_B(x) = \frac{(Ax,x)-(Bx,x)}{\|x\|_2^2} = \frac{((A-B)x,x)}{\|x\|_2^2} \leq$$

$$\lambda_1(A-B) \leq \max_j |\lambda_j(A-B)| \leq \|A-B\|_p ,$$

hvor sidste ulighed følger af kap. II, § 2.

Er nu M et underrum i $\mathbb{R}^n$, gælder altså

$$\max_{x\in M} \rho_A(x) \leq \max_{x\in M} \rho_B(x) + \|A-B\|_p .$$

Har M dimension m, fås heraf

$$\min_{\dim N=m} \max_{x\in N} \rho_A(x) \leq \max_{x\in M} \rho_B(x) + \|A-B\|_p .$$

Da denne ulighed gælder for hvert M med dimension m, gælder også

$$\min_{\dim N=m} \max_{x\in N} \rho_A(x) \leq \min_{\dim M=m} \max_{x\in M} \rho_B(x) + \|A-B\|_p ;$$

med kortere skrivemåde og $j = n+1-m$

$$\lambda_j(A) \leq \lambda_j(B) + \|A-B\|_p .$$

Ved ombytning af A og B, idet $\|A\|_p = \|-A\|_p$, fås

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A-B\|_p .$$

Bevis slut.

§2. Potensmetoden.

Lad $n\times n$-matricen A være reel, symmetrisk og yderligere positiv semidefinit, og lad egenværdierne være

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

Den tilsvarende afbildning vil afbilde en kugle på en ellipsoide, og ved iteration vil ellipsoiden normalt blive fladere og fladere. En vilkårlig vektor, der ikke er ortogonal på nogen egenvektor, vil ved iterationen komme nærmere og nærmere til den egenvektor, w_1 , der svarer til den største egenværdi, λ_1. Denne kendsgerning kan vi udnytte til at approximere egenvektoren w_1 med $A^m x$, x vilkårlig.

Er egenværdierne negative eller komplekse, vil iterationerne forstyrres af roterende tendenser. For negative egenværdier går det an, men i det komplekse tilfælde går vi glip af imaginærdelen af egenværdien. Det er derfor, vi begrænser os til positiv semidefinitte matricer.

Vi betragter altså iterationsfølgen

$$x^{(t+1)} = Ax^{(t)} \quad , \quad t = 0,1,2,\cdots$$

$x^{(0)} = x \in \mathbb{R}^n$ vilkårlig, og skal undersøge, hvorledes den approximerer en egenvektor.

Hvis $Ax = 0$, er 0 egenværdi og x egenvektor for 0 . Vi kan derfor se bort fra $Ax = 0$ i det følgende.

Lad $w_1,\cdots,w_n$ være en ortonormalbasis svarende til egenværdierne $\lambda_1,\cdots,\lambda_n$. Da gælder

$$x^{(t)} = A^t x = \sum_j \lambda_j^t (x,w_j) w_j \quad , \quad t = 0,1,2,\cdots$$

Da $0 \neq Ax = \sum_j \lambda_j (x,w_j) w_j$, må $\|Ax\|_2^2 \neq 0$, altså

$$\sum_j \lambda_j^2 (x,w_j)^2 > 0 \ .$$

Lad nu m være det mindste tal, så $\lambda_m(x,w_m) \neq 0$. Da er $(x,w_j) = 0$ for $j = 1,\cdots,m-1$, altså x er ortogonal på $w_1,\cdots,w_{m-1}$. Da er $\lambda = \lambda_m$ den største egenværdi i spil, den vi vil approximere ved hjælp af $x^{(t)}$.

Vi spalter nu $x^{(t)}$ i to dele; en del, der er egenvektor svarende til λ, og en del, der er ortogonal på egenrummet hørende til λ. Med

$$z = \sum_{\substack{j=m \\ \lambda_j=\lambda}}^{n} (x,w_j)w_j \quad \text{og} \quad r^{(t)} = \sum_{\substack{j=m \\ \lambda_j<\lambda}}^{n} \left(\frac{\lambda_j}{\lambda}\right)^t (x,w_j)w_j$$

fås opspaltningen

$$x^{(t)} = \lambda^t(z + r^{(t)}).$$

Da z og $r^{(t)}$ er ortogonale, får vi af Pythagoras

$$\|x^{(t)}\|_2 = \lambda^t\sqrt{\|z\|_2^2 + \|r^{(t)}\|_2^2}.$$

Vi vil nu vise, at $r^{(t)} \to 0$ for $t \to \infty$. Sæt $\lambda' = \max_{\lambda_j<\lambda} \lambda_j$, så er $\frac{\lambda_j}{\lambda} \leq \frac{\lambda'}{\lambda} = q < 1$ for $\lambda_j < \lambda$. Da gælder

$$\|r^{(t)}\|_2 \leq q^t \|r^{(0)}\|_2 \to 0 \quad \text{for} \quad t \to \infty,$$

fordi $q < 1$. Vi får derfor for $x^{(t)}$ normeret

$$y^{(t)} = \frac{x^{(t)}}{\|x^{(t)}\|_2} = \frac{z+r^{(t)}}{\sqrt{\|z\|_2^2+\|r^{(t)}\|_2^2}} \to \frac{z}{\|z\|_2} = w \text{ for } t \to \infty.$$

$y^{(t)}$ konvergerer altså mod en egenvektor, w, hørende til egenværdien λ. Rayleigh-kvotienten konvergerer mod egenværdien:

$$\rho_A(y^{(t)}) = (Ay^{(t)},y^{(t)}) \to (Aw,w) = \lambda \quad \text{for} \quad t \to \infty,$$

idet $\|y^{(t)}\|_2 = 1$, $\|w\|_2 = 1$ og A er kontinuert.

Raylaigh-kvotienten for $x^{(t)}$ er den samme:

$$\rho_A(x^{(t)}) = \frac{(Ax^{(t)},x^{(t)})}{\|x^{(t)}\|_2^2} = (Ay^{(t)},y^{(t)}) = \rho_A(y^{(t)}).$$

Vi vil nu vurdere konvergenshastigheden for de to grænseovergange mod hhv. λ og w.

$$(Ax^{(t)},x^{(t)}) = \sum_{j=m}^{n} \lambda_j^{2t+1}(x,w_j)^2 =$$

$$\sum_{\substack{j=m \\ \lambda_j=\lambda}}^{n} \lambda^{2t+1}(x,w_j)^2 + \sum_{\substack{j=m \\ \lambda_j<\lambda}}^{n} \lambda_j^{2t+1}(x,w_j)^2 \geq \lambda^{2t+1}\|z\|_2^2 ,$$

idet første led er lig med højre side, og andet led er ikke negativt.

Ved på den anden side at vurdere én faktor λ_j op til λ for hvert j og erindre, at $\|x^{(t)}\|_2^2 = \sum_{j=m}^{n} \lambda_j^{2t}(x,w_j)^2$, fås

$$(Ax^{(t)},x^{(t)}) \leq \sum_{j=m}^{n} \lambda\cdot\lambda_j^{2t}(x,w_j)^2 = \lambda\|x^{(t)}\|_2^2$$

Af den sidste ulighed fås

$$\rho(y^{(t)}) = \rho(x^{(t)}) = \frac{(Ax^{(t)},x^{(t)})}{\|x^{(t)}\|_2^2} \leq \lambda \quad , \quad \text{eller}$$

$$0 \leq \lambda - \rho(y^{(t)}) .$$

Medens vi af den første ulighed får

$$\lambda - \rho(y^{(t)}) = \lambda - \frac{(Ax^{(t)},x^{(t)})}{\|x^{(t)}\|_2^2} \leq$$

$$\lambda - \frac{\lambda^{2t+1}\|z\|_2^2}{\|x^{(t)}\|_2^2} =$$

$$\lambda - \frac{\lambda^{2t+1}\|z\|_2^2}{\lambda^{2t}(\|z\|_2^2+\|r^{(t)}\|_2^2)} =$$

$$\lambda \frac{\|z\|_2^2 + \|r^{(t)}\|_2^2 - \|z\|_2^2}{\|z\|_2^2 + \|r^{(t)}\|_2^2} \leq$$

$$\lambda \frac{\|r^{(t)}\|_2^2}{\|z\|_2^2} \leq \lambda q^{2t} \cdot \gamma^2 ,$$

hvor $\gamma = \frac{\|r^{(0)}\|_2}{\|z\|_2}$. Vi har altså også

$$0 \leq \frac{1}{\lambda} (\lambda - \rho(y^{(t)})) \leq \gamma^2 q^{2t} .$$

Rayleigh-kvotienten konvergerer altså mod λ fra neden med en relativ fejlforbedring med en faktor højst q^2 .

For egenvektoren $w = \frac{z}{\|z\|}$ gælder

$$\|y^{(t)} - w\|_2 \leq \gamma q^t \quad \text{for} \quad t = 0,1,2,\cdots$$

Bevis.

$$\|y^{(t)} - w\|_2^2 = \|y^{(t)}\|_2^2 + \|w\|_2^2 - 2(y^{(t)}, w)$$

$$= 2 - \frac{2}{\|z\|_2} (y^{(t)}, z)$$

$$= 2 - \frac{2}{\|z\|_2} \frac{(x^{(t)}, z)}{\|x^{(t)}\|_2}$$

$$= 2 - \frac{2}{\|z\|_2} \frac{\lambda^t \|z\|_2^2}{\lambda^t \|z+r^{(t)}\|_2}$$

$$= 2 \cdot \frac{\|z+r^{(t)}\|_2 - \|z\|_2}{\|z+r^{(t)}\|_2}$$

$$= 2 \cdot \frac{\|z+r^{(t)}\|_2^2 - \|z\|_2^2}{\|z+r^{(t)}\|_2 (\|z+r^{(t)}\|_2 + \|z\|_2)}$$

$$\leq \frac{\|r^{(t)}\|_2^2}{\|z\|_2^2}$$

$$\leq \gamma^2 q^{2t} .$$

Bevis slut.

Selv om vi begynder med en vektor x, der er ortogonal på vektoren w_1, kan vi ikke stole på at $A^m x$ forbliver ortogonal på w_1. Vi risikerer, at afrundingsfejl giver små bidrag i w_1-retningen, der efterhånden vokser sig store og til sidst bliver dominerende. Vi må derfor regne med i praksis at finde λ_1 og w_1 uanset startvektor.

Man kan afhjælpe denne ulempe, hvis man har fundet w_1 og søger w_2 ved at ortogonalisere hen ad vejen:

$$v^{(t)} = Ax^{(t)} \; ; \; x^{(t+1)} = v^{(t)} - (v^{(t)}, w_1)w_1 .$$

Mere almindeligt, vi har fundet egenvektorerne $w_1, \cdots, w_m$ og søger w_{m+1}. Vi sætter så

$$v^{(t)} = Ax^{(t)} \; ; \; x^{(t+1)} = v^{(t)} - \sum_{l=1}^{m} (v^{(t)}, w_l)w_l .$$

Endelig bemærker vi, at er p er polynomium,

$$p(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_m x^m ,$$

kan vi definere matricen

$$p(A) = \alpha_0 E + \alpha_1 A + \cdots + \alpha_m A^m .$$

Denne matrix har samme egenvektorer som A, thi

$$\begin{aligned} p(A)w &= \alpha_0 Ew + \alpha_1 Aw + \cdots + \alpha_m A^m w \\ &= \alpha_0 w + \alpha_1 \lambda w + \cdots + \alpha_m \lambda^m w \\ &= (\alpha_0 + \alpha_1 \lambda + \cdots + \alpha_m \lambda^m) w \\ &= p(\lambda) w , \end{aligned}$$

når w er egenvektor for A svarende til egenværdi λ. Samtidig ser vi, at egenværdierne for $p(A)$ er $p(\lambda)$. Og tilsvarende, hvis A er regulær og derfor alle egenværdier for A forskellige fra 0, gælder

$$A^{-1}w = \frac{1}{\lambda} A^{-1}(\lambda w)$$

$$= \frac{1}{\lambda} A^{-1}(Aw)$$

$$= \frac{1}{\lambda} w .$$

Altså A og A^{-1} har samme egenvektorer og inverse egenværdier.

Disse betragtninger kan benyttes. Er A en reel, symmetrisk matrix med egenværdier $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, vil A^2 være reel, symmetrisk med egenværdier $\lambda_1^2, \lambda_2^2, \ldots, \lambda_n^2$, der alle er ikke-negative.

Med andre ord, benytter vi kun de lige potenser af matricen, får vi i dette tilfælde konvergens mod en egenvektor og kvadratet på den tilsvarende egenværdi. Ved til sidst at anvende A ses, om egenværdien er positiv eller negativ.

Denne metode slår fejl, hvis λ og $-\lambda$ er egenværdier for A, så egenrumemt svarende til λ^2 for A^2 er mindst 2-dimensionalt. Vi risikerer da at finde en egenvektor for A^2, der ikke er egenvektor for A. Men erstatter vi A med $A + \alpha E$, $\alpha \neq 0$, bliver egenværdierne flyttet til $\lambda+\alpha$ og $-\lambda+\alpha$, hvis kvadrater også er forskellige.

Er α ikke egenværdi for A, kan vi betragte matricen $(A-\alpha E)^{-1}$. Den har egenværdier $\frac{1}{\lambda-\alpha}$, altså den numeriske største egenværdi for $(A-\alpha E)^{-1}$ svarer til den egenværdi λ for A, der ligger nærmest α. Til gengæld forudsætter denne variant, at ligningssystemet

$$(A-\alpha E)x^{(t+1)} = x^{(t)}$$

er overkommeligt at løse med hensyn til $x^{(t+1)}$.

Til gengæld vil konvergensen jo gå hurtigere, når α ligger tættere på λ, da $|\frac{1}{\lambda-\alpha}|$ derved bliver forholdsvis større end $|\frac{1}{\mu-\alpha}|$, hvor μ er den næst-nærmeste egenværdi. Dette kan vi udnytte ved successivt at erstatte α med den sidste tilnærmelse til λ. Dette giver forskriften

$$\rho_t = (Ay^{(t)}, y^{(t)});\ (A-\rho_t E)\nu^{(t)} = y^{(t)};\ y^{(t+1)} = \frac{\nu^{(t)}}{\|\nu^{(t)}\|_2}.$$

Denne metode føres man også til ved anvendelse af Newton på

$$f(x) = Ax - \frac{(Ax,x)}{(x,x)}x = Ax - \varrho(x)x,$$

der netop har A's egenvektorer som nulpunkter. Man finder

$$f'(x) = A - \varrho(x)E - P(x),$$

hvor $P(x) = (x_j \frac{\partial \varrho}{\partial x_k}(x))_{j,k}$. Newton giver derfor iterationen

$$f(x^t) + f'(x^t)(x^{t+1} - x^t) = 0,$$

dvs. $Ax^t - \varrho(x^t)x^t + (A-\varrho(x^t))x^{t+1} - Ax^t + \varrho(x^t)x^t - P(x^t)(x^{t+1}-x^t) = 0$

$\Leftrightarrow$ $(A - \varrho(x^t))x^{t+1} = P(x^t)(x^{t+1} - x^t)$

$\Leftrightarrow$ $(A - \varrho(x^t))x^{t+1} = \sigma_t x^t,$

hvor $$\sigma_t = \sum_{k=1}^{n} \frac{\partial \varrho}{\partial x_k}(x^t)(x_k^{t+1}-x_k^t).$$

Ved iterationen bliver σ_t uden betydning, da vi normerer x^{t+1} til y^{t+1}, der netop er $y^{(t+1)}$ ovenfor.

§3. Jacobis metode.

Lad n×n-matricen A være reel og symmetrisk. Lad S være matricen med A's egenvektorer som søjler, valgt så egenvektorerne er ortonormale. Da er $S' = S^{-1}$ og $S'AS = \Lambda$ en diagonalmatrix med A's egenværdier i diagonalen. Jacobis metode søger at approximere S og Λ simultant.

Idet vi erindrer om sætning V.1.2 er det godt at foretage en ortonormal transformation af A, så elementerne uden for diagonalen alle er små. Da vil diagonalen approximere Λ. Dertil benyttes en successiv udlugning af elementerne uden for diagonalen med specielle transformationer. Denne proces er dog ikke endelig, da allerede udlugede elementer kommer igen, dog altid mindre.

Lad U være matricen

$$\begin{pmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \cos\varphi & -\sin\varphi & \\ & & \sin\varphi & \cos\varphi & \\ 0 & & & & 1 \end{pmatrix}$$

som er enhedsmatricen på nær pladserne (σ,σ), (σ,τ), (τ,σ), (τ,τ).

$$U = (u_{jk}) ,$$

$$u_{jk} = \delta_{jk} \quad \text{for} \quad j \neq \sigma,\tau \vee k \neq \sigma,\tau$$

$$u_{\sigma\sigma} = u_{\tau\tau} = \cos\varphi$$

$$u_{\sigma\tau} = -u_{\tau\sigma} = -\sin\varphi .$$

Lad $A = (a_{jk})$ være en matrix og $B = U'AU$, $B = (b_{jk})$. Da gælder relationerne mellem a_{jk} og b_{jk} :

$$b_{jk} = a_{jk} \quad \text{for} \quad j \neq \sigma,\tau \wedge k \neq \sigma,\tau$$

$$b_{j\sigma} = b_{\sigma j} = a_{j\sigma}\cos\varphi + a_{j\tau}\sin\varphi \quad \text{for} \quad j \neq \sigma,\tau$$

$$b_{j\tau} = b_{\tau j} = -a_{j\sigma}\sin\varphi + a_{j\tau}\cos\varphi \quad \text{for} \quad j \neq \sigma,\tau$$

$$b_{\sigma\sigma} = a_{\sigma\sigma}\cos^2\varphi + a_{\sigma\tau}\sin 2\varphi + a_{\tau\tau}\sin^2\varphi$$

$$b_{\tau\tau} = a_{\sigma\sigma}\sin^2\varphi - a_{\sigma\tau}\sin 2\varphi + a_{\tau\tau}\cos^2\varphi$$

$$b_{\sigma\tau} = b_{\tau\sigma} = a_{\sigma\tau}\cos 2\varphi + \tfrac{1}{2}(a_{\tau\tau} - a_{\sigma\sigma})\sin 2\varphi$$

Vi kan vurdere A's afvigelse fra at være diagonal ved 2-normen af A's forskel fra sin diagonal:

$$|A| = \|A - \operatorname{diag}(a_{jj})\|_2 = \sqrt{\sum_{j\neq k} a_{jk}^2} .$$

<u>Sætning V.3.1</u>. *Er* $B = U'AU$, *hvor* U *er defineret ovenfor, gælder*

$$|B|^2 = |A|^2 + 2(b_{\sigma\tau}^2 - a_{\sigma\tau}^2) .$$

Bevis. Sporet af A^2 er kvadratet på normen af A:

$$\operatorname{Sp}(A^2) = \sum_{j=1}^{n}\sum_{l=1}^{n} a_{jl}a_{lj} = \sum_{j,l} a_{jl}^2 = \|A\|_2^2 .$$

Da $B^2 = U'A^2U$, har B^2 og A^2 samme spor,

$$|B|^2 + \sum_{j=1}^{n} b_{jj}^2 = \|B\|_2^2 = \|A\|_2^2 = |A|^2 + \sum_{j=1}^{n} a_{jj}^2 .$$

Nu var $a_{jj} = b_{jj}$ for $j \neq \sigma,\tau$, så vi får heraf

$$|B|^2 = |A|^2 + a_{\sigma\sigma}^2 + a_{\tau\tau}^2 - b_{\sigma\sigma}^2 - b_{\tau\tau}^2 .$$

For matricerne

$$A_0 = \begin{pmatrix} a_{\sigma\sigma} & a_{\sigma\tau} \\ a_{\tau\sigma} & a_{\tau\tau} \end{pmatrix} , \quad B_0 = \begin{pmatrix} b_{\sigma\sigma} & b_{\sigma\tau} \\ b_{\tau\sigma} & b_{\tau\tau} \end{pmatrix} , \quad U_0 = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} .$$

gælder

$$B_0 = U_0' \, A_0 \, U_0 .$$

Da derfor $Sp(A_0^2) = Sp(B_0^2)$ eller $\|A_0\|_2^2 = \|B_0\|_2^2$, fås formlen

$$a_{\sigma\sigma}^2 + a_{\tau\tau}^2 + 2a_{\sigma\tau}^2 = b_{\sigma\sigma}^2 + b_{\tau\tau}^2 + 2b_{\sigma\tau}^2 ,$$

hvoraf

$$a_{\sigma\sigma}^2 + a_{\tau\tau}^2 - b_{\sigma\sigma}^2 - b_{\tau\tau}^2 = 2(b_{\sigma\tau}^2 - a_{\sigma\tau}^2) .$$

Bevis slut.

Vi skal nu disponere over σ,τ,φ med henblik på at gøre $|B|$ mindst mulig. Først vælges σ,τ , så $|a_{\sigma\tau}|$ er størst; siden vælges φ, så $|b_{\sigma\tau}|$ bliver mindst, nemlig 0. Til bestemmelse af φ har vi ligningen

$$b_{\sigma\tau} = 0 , \quad \text{ɔ:}$$

$a_{\sigma\tau}\cos 2\varphi + \frac{1}{2}(a_{\tau\tau}-a_{\sigma\sigma})\sin 2\varphi = 0$. Er $a_{\sigma\tau} = 0$ er vi allerede

færdige; sæt $\varphi = 0$. Hvis $a_{\sigma\sigma} = a_{\tau\tau}$, fås $\varphi = \frac{\pi}{4}$, $\cos\varphi = \sin\varphi = \frac{\sqrt{2}}{2}$. Ellers er

$$\operatorname{tg}2\varphi = \frac{2a_{\sigma\tau}}{a_{\sigma\sigma}-a_{\tau\tau}} ,$$

hvoraf $\sin\varphi$ og $\cos\varphi$ kan findes ved brug af formlerne

$$\operatorname{tg}\varphi = \frac{\operatorname{tg}2\varphi}{1+\sqrt{1+(\operatorname{tg}2\varphi)^2}}$$

$$\cos\varphi = \frac{1}{\sqrt{1+\operatorname{tg}^2\varphi}}$$

$$\sin\varphi = \cos\varphi \cdot \operatorname{tg}\varphi .$$

Vi har herved en iterationsmetode. Ud fra matricen A_t bestemmes U_t som ovenfor og $A_{t+1} = U_t' A_t U_t$. Sætning V.3.1 siger da

$$|A_{t+1}|^2 = |A_t|^2 - 2 \max_{j\neq k} a_{jk}^2 .$$

Da altså $|A_{t+1}| \leq |A_t|$, må følgen $|A_t|$ være konvergent. Og da der for $j \neq k$ gælder

$$|a_{jk}^{(t)}| \leq \max_{i\neq h} |a_{ih}^{(t)}| = \sqrt{\frac{|A_t|^2-|A_{t+1}|^2}{2}} \to 0$$

for $t \to \infty$, kan der kun gælde, at

$$|A_t| \to 0 \quad \text{for} \quad t \to \infty.$$

Det er endvidere klart, at

$$|A_t|^2 \leq (n^2-n) \max_{j\neq k} (a_{jk}^{(t)})^2 ,$$

så vi finder, at

$$|A_{t+1}|^2 \leq |A_t|^2 - \frac{2}{n^2-n} |A_t|^2$$

$$= (1 - \frac{2}{n^2-n}) |A_t|^2 .$$

Så $|A_t|$ konvergerer mindst så hurtigt som

$$\left(\sqrt{1 - \frac{2}{n^2-n}}\right)^t .$$

Er $\lambda_1,\cdots,\lambda_n$ egenværdierne for A i passende orden, får vi, da A og A_t har samme egenværdier, af sætning V.1.2 vurderingerne

$$|\lambda_j - a_{jj}^{(t)}| \leq \|A_t - \mathrm{diag}(a_{jj})\|_p \ , \ p = 1,2,\infty .$$

Altså enten $|A_t|$ eller $\max_j \sum_{k\neq j} |a_{jk}^{(t)}|$.

Jacobis metode giver således simultan approximation af egenværdierne. Sætter vi $Y_0 = E$, $Y_{t+1} = Y_t U_t$, $t = 0,1,2,\cdots$ har vi

$$A_t = Y_t' A Y_t , \qquad (\text{når } Y_0 = E,)$$

hvor A_t næsten er en diagonalmatrix, så Y_t næsten har A's egenvektorer som søjler. I tilfældet hvor A har lutter forskellige egenværdier, vil vi vurdere denne approximation nærmere.

Lad $\lambda_1,\cdots,\lambda_n$ være de parvis forskellige egenværdier for A og $w_1,\cdots,w_n$ et ortonormalsystem af tilsvarende egenvektorer, og lad $a_{jj}^{(t)} \to \lambda_j$ for $t \to \infty$ og $j = 1,\cdots,n$.

Sæt

$$\delta_k = \min_{j\neq k} |\lambda_j - \lambda_k| .$$

Vektoren

$$\nu_k^{(t)} = Y_t' w_k$$

er egenvektor for A_t svarende til λ_k ; thi

$$A_t \nu_k^{(t)} = (Y_t' A Y_t)(Y_t' w_k)$$

$$= Y_t' A w_k$$

$$= Y_t' \lambda_k w_k$$
$$= \lambda_k \nu_k^{(t)} \quad .$$

Vi vælger i det følgende t så stor, at $|A_t| \leq \frac{\delta_k}{2}$;

herved bliver specielt

$$|\lambda_k - a_{kk}^{(t)}| \leq |A_t| \leq \frac{\delta_k}{2} \quad .$$

For $j \neq k$ gælder modsvarende

$$|\lambda_j - a_{kk}^{(t)}| \geq |\lambda_j - \lambda_k| - |\lambda_k - a_{kk}^{(t)}| \geq \delta_k - \frac{\delta_k}{2} = \frac{\delta_k}{2} \quad .$$

Lad nu e_k være enhedsvektoren $(\delta_{jk})_j$; så er e_k egenvektor for $\mathrm{diag}(a_{jj}^{(t)})$ svarende til egenværdien $a_{kk}^{(t)}$. Der gælder

$$\| A_t e_k - a_{kk}^{(t)} e_k \|_2 = \sqrt{\sum_{\substack{j=1 \\ j \neq k}}^{n} a_{jk}^2} \leq |A_t| \leq \frac{\delta_k}{2} \quad .$$

Vi ønsker at approximere $w_k = Y_t \nu_k^{(t)}$ med $y_k^{(t)} = Y_t e_k$. Da Y_t er ortonormal, kan vi vurdere approximationen ved

$$\| y_k^{(t)} - (y_k^{(t)}, w_k) w_k \|_2 = \| e_k - (e_k, \nu_k^{(t)}) \nu_k^{(t)} \|_2 \quad .$$

Nu er $\nu_j^{(t)}$ egenvektor for $A_t - a_{kk}^{(t)} E$ for $j = 1, \cdots, n$, så

$$\| (A_t - a_{kk}^{(t)} E)(e_k - (e_k, \nu_k^{(t)}) \nu_k^{(t)}) \|_2^2 =$$
$$\sum_{\substack{j=1 \\ j \neq k}}^{n} |\lambda_j - a_{kk}^{(t)}|^2 |(e_k - (e_k, \nu_k^{(t)}) \nu_k^{(t)} , \nu_j^{(t)})|^2 \geq$$
$$\min_{\substack{j=1,\cdots,n \\ j \neq k}} |\lambda_j - a_{kk}^{(t)}|^2 \sum_{j=1}^{n} |(e_k - (e_k, \nu_k^{(t)}) \nu_k^{(t)}, \nu_j^{(t)})|^2$$
$$\geq \left(\frac{\delta_k}{2}\right)^2 \cdot \| e_k - (e_k , \nu_k^{(t)}) \nu_k^{(t)} \|_2^2 \quad ,$$

da $(e_k - (e_k,v_k^{(t)})v_k^{(t)},v_k^{(t)}) = 0$.

Til vurderingen opad sættes $B = A_t - a_{kk}^{(t)}E$, som er symmetrisk med $\nu_k^{(t)}=\nu$ som egenvektor, og $e = e_k$. Med disse betegnelser fås

$$||B(e - (e,\nu)\nu)||^2$$

$$||$$ da B er lineær

$$||Be - (e,\nu)B\nu||^2$$

$$||$$ da ν er egenvektor

$$||Be - (e,B\nu)\nu||^2$$

$$||$$ da B er symmetrisk

$$||Be - (Be,\nu)\nu||^2$$

$$||$$ ifølge Pythagoras

$$||Be||^2 - ||(Be,\nu)\nu||^2$$

$$|\wedge$$

$$||Be||^2 \leq |A_t|^2.$$

Sammenfattende fås vurderingen

$$||e_k - (e_k,v_k^{(t)})v_k^{(t)}||_2 \leq \frac{2}{\delta_k}|A_t| \quad ,$$

og derfor også

$$||y_k^{(t)} - (y_k^{(t)},w_k)w_k||_2 \leq \frac{2}{\delta_k}|A_t| \ .$$

Variant.

Hvis man finder det besværligt at opsøge det numerisk største element uden for diagonalen i A_t , kan man vælge at tage alle elementer uden for diagonalen efter tur. Man taler da om *den cykliske Jacobis metode*.

Kapitel VI. EGENVÆRDIER FOR KOMPLEKSE MATRICER.

§1. Minimalpolynomier.

For en vilkårlig kompleks matrix A har *det karakteristiske polynomium*

$$p_A(x) = \det(xE-A) = x^n + \beta_{n-1}x^{n-1} + \cdots + \beta_0$$

netop A's egenværdier som rødder.

Vi kan tage et polynomiums værdi i en matrix og derved få en ny matrix. Specielt gælder

Sætning VI.1.1. *Er* p_A *det karakteristiske polynomium for matricen* A, *da er*

$$p_A(A) = 0 .$$

Bevis. Lad w være en egenvektor for A med tilsvarende egenværdi λ. Da er

$$\begin{aligned} p_A(A)w &= (A^n + \beta_{n-1}A^{n-1} + \cdots + \beta_0 E)w \\ &= A^n w + \beta_{n-1}A^{n-1}w + \cdots + \beta_0 w \\ &= \lambda^n w + \beta_{n-1}\lambda^{n-1}w + \cdots + \beta_0 w \\ &= (\lambda^n + \beta_{n-1}\lambda^{n-1} + \cdots + \beta_0)w \\ &= 0 \cdot w \end{aligned}$$

Matricen $p_A(A)$ giver altså anledning til en 0-afbildning, så den må være 0-matricen.

Bevis slut.

Lad a være en vilkårlig vektor, så $Aa \neq 0$. Vi betragter vektorerne $a^{(t)} = A^t a$ for $t = 0,1,\cdots,n$. Disse $n+1$ vektorer er åbenbart lineært afhængige; der findes derfor et tal m, så

$a^{(0)},\cdots,a^{(m-1)}$ er lineært uafhængige, og
$a^{(0)},\cdots,a^{(m)}$ er lineært afhængige.

Altså findes ét og kun ét talsæt $\alpha_0,\cdots,\alpha_{m-1}$, så

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

Definition. Polynomiet $p_a(x) = x^m + \alpha_{m-1}x^{m-1} +\cdots+ \alpha_1 x + \alpha_0$ kaldes *minimalpolynomiet* for vektoren a (med hensyn til matricen A).

Sætning VI.1.2. *Minimalpolynomiet er divisor i det karakteristiske polynomium;*

$$p_a(x) \mid p_A(x) .$$

Bevis.
$$\begin{aligned} p_a(A)a &= (A^m + \alpha_{m-1}A^{m-1} +\cdots+ \alpha_1 A + \alpha_0 E)a \\ &= a^{(m)} + \alpha_{m-1}a^{(m-1)} +\cdots+ \alpha_1 a^{(1)} + \alpha_0 a^{(0)} \\ &= 0 \end{aligned}$$

Euklids algoritme tillader os at finde polynomier q af grad $n-m$ og r af grad $< m$, så

$$p_A(x) = q(x)p_a(x) + r(x) .$$

Lad $r(x) = \gamma_{m-1}x^{m-1} + \cdots + \gamma_0$, hvor vi skal vise, at $\gamma_0, \cdots, \gamma_{m-1}$ alle er 0.
Indsættes A ovenfor, fås af Sætning VI.1.1.

$$0 = p_A(A) = q(A)p_a(A) + r(A) .$$

Anvendes matricen på vektoren a, står der

$$0 = 0\cdot a = q(A)p_a(A)a + r(A)a = r(A)a ,$$

det samme som

$$\begin{aligned} 0 &= (\gamma_{m-1}A^{m-1} + \cdots + \gamma_0 E)a \\ &= \gamma_{m-1}a^{(m-1)} + \cdots + \gamma_0 a^{(0)} , \end{aligned}$$

som sammen med uafhængigheden af $a^{(0)}, \cdots, a^{(m-1)}$ giver $\gamma_0 = \cdots = \gamma_{m-1} = 0$.

Bevis slut.

Corollar VI.1.2. *Rødderne i minimalpolynomiet er egenværdier for matricen.*
Klart.

Minimalpolynomiet giver os også egenvektorerne til de egenværdier, der er rødder deri. Lad λ være en rod i $p_a(x)$

og dermed egenværdi for A. Vi kan da skrive $p_a(x) = (x-\lambda)q(x)$, hvor q har grad m-1 , vi skriver $q(x) = x^{m-1} + \delta_{m-2}x^{m-2} + \cdots + \delta_0$.

Sætning VI.1.3. *Vektoren*

$$z = q(A)a$$

er egenvektor for A *svarende til egenværdien* λ.

Bevis. $z \neq 0$, da $z = q(A)a = (A^{m-1} + \delta_{m-2}A^{m-2} + \cdots + \delta_0 E)a = a^{(m-1)} + \delta_{m-2}a^{(m-2)} + \cdots + \delta_0 a^{(0)}$, og $a^{(0)}, \cdots, a^{(m-1)}$ er uafhængige og koefficienten til $a^{(m-1)}$ er $1 \neq 0$. Da endvidere

$$p_a(A) = (A-\lambda E)q(A) ,$$

fås

$$Az - \lambda z = (A-\lambda E)z = (A-\lambda E)q(A)a = p_a(A)a = 0 ,$$

som i beviset for sætning VI.1.2.

Altså er

$$Az = \lambda z$$

Bevis slut.

§2. Krylovs metode.

Resultaterne i §1 hjælper os til at løse egenværdiproblemet for en vilkårlig matrix. Vi skal blot kende en metode til at finde minimalpolynomier, så kan vi finde egenværdier og egenvektorer ved at finde nulpunkter i et polynomium som i kapitel IV.

Til dette formål skal vi løse problemet at bestemme, hvor længe $a^{(0)},\dots,a^{(m)},\dots$ er uafhængige. Dertil foretager vi en successiv ortogonalisering af vektorerne $a^{(0)},\dots,a^{(m)},\dots$ indtil en sådan proces giver 0 . Vi sætter

$$\nu_0 = a^{(0)} , \qquad w_0 = \frac{\nu_0}{\|\nu_0\|_2} ;$$

$$\nu_j = a^{(j)} - \sum_{k=0}^{j-1} (a_j, w_k) w_k , \quad w_j = \frac{\nu_j}{\|\nu_j\|} .$$

Denne proces ender med, at

$$\nu_m = a^{(m)} - \sum_{k=0}^{m-1} (a_m, w_k) w_k = 0 .$$

Herefter kan $\alpha_{m-1},\dots,\alpha_0$ beregnes ved tilbageregning:

$$\alpha_{m-1} = - \frac{(a^{(m)}, \nu_{m-1})}{\|\nu_{m-1}\|_2^2} ,$$

(*)

$$\alpha_j = - \frac{1}{\|\nu_j\|_2^2} \left((a^{(m)}, \nu_j) + \sum_{k=j+1}^{m-1} \alpha_k (a^{(k)}, \nu_j) \right) .$$

Der blot er formlen (*) fra § II,4.

Kapitel VII. Interpolation.

§1. Lagrange polynomier.

Er der givet $m+1$ punkter $(x_0,y_0),\cdots,(x_m,y_m)$ i (x,y)-planen, så $x_0,\cdots,x_m$ er parvis forskellige, da findes ét og kun et polynomium $p(x)$ af grad højst m, så $p(x_j) = y_j$ for $j = 0,1,\cdots,m$. Den simpleste interpolationsopgave går ud på at finde p.

Lagrange's løsning består i at løse opgaven i det tilfælde, hvor $y_j = 0$ for $j \neq k$ og $y_k = 1$. Kaldes løsningen hertil $L_K(x)$, fås

$$p(x) = \sum_{k=0}^{m} y_k L_k(x).$$

Da Lagrange polynomiet har grad m og rødder $x_0,\cdots,x_m$ på nær x_k, kan det åbenbart skrives

$$L_k(x) = \alpha(x-x_0)\cdots(x-x_m) \not/ (x-x_k)$$

$$= \alpha \prod_{\substack{j=0 \\ j\neq k}}^{m} (x-x_j).$$

Til bestemmelse af α har vi

$$L_k(x_k) = 1, \quad \text{d:}$$

$$\alpha = \frac{1}{\prod_{\substack{j=0 \\ j\neq k}}^{m} (x_k-x_j)}.$$

Herefter har vi

$$L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{m} \frac{(x-x_j)}{(x_k-x_j)} .$$

§2. Newton polynomier.

Da Lagrange polynomier er upraktiske, f.eks. ved tilføjelse af et nyt punkt, hvorved alle L_k skal ændres, vil vi se på rekursionen i systemet.

Lad os betragte alle delmængder af indexmængden,

$$\{j_0,\cdots,j_\ell\} \subseteq \{0,\cdots,m\} \quad \text{for} \quad \ell = 0,\cdots,m.$$

Vi kalder interpolationspolynomiet for punkterne $(x_{j_0},y_{j_0}),\cdots,(x_{j_\ell},y_{j_\ell})$ for $p_{j_0\cdots j_\ell}(x)$. Da er specielt

$$p_{0\cdots m}(x) = p(x).$$

For disse gælder rekursionsformlen

$$p_{j,\cdots,j+\ell}(x) = \frac{1}{x_j - x_{j+\ell}}((x-x_{j+\ell})p_{j,\cdots,j+\ell-1}(x)-(x-x_j)p_{j+1,\cdots,j+\ell}(x))$$

for $j = 0,\cdots,m-\ell$, $\ell = 1,\cdots,m$.

Bevis. I punkterne x_k, $k = j+1,\cdots,j+\ell-1$ stemmer polynomierne uden videre overens med værdierne y_k. I x_j bliver højre side $\frac{x_j-x_{j+\ell}}{x_j-x_{j+\ell}}p_{j,\cdots,j+\ell-1}(x_j) - 0 = y_j$ og i $x_{j+\ell}$: $\frac{(x_{j+\ell}-x_j)}{x_j-x_{j+\ell}}(0-p_{j+1,\cdots,j+\ell}(x_{j+\ell})) = y_{j+\ell}$. Da polynomierne begge har grad højst ℓ, stemmer de overens.

Bevis slut.

Formlerne viser, at polynomial interpolation kan fås ved sammensætning af lineære interpolationer.

$$y[x_j,\cdots,x_{j+\ell}] = \frac{1}{\ell!}\,\frac{d^\ell p_{j,\cdots,j+\ell}}{dx^\ell}$$

$$= \frac{1}{\ell!}\,\frac{1}{x_j - x_{j+\ell}}\left(\ell\frac{d^{\ell-1}p_{j,\cdots,j+\ell-1}}{dx^{\ell-1}} - \ell\frac{d^{\ell-1}p_{j+1,\cdots,j+\ell}}{dx^{\ell-1}}\right)$$

$$= \frac{1}{x_j - x_{j+\ell}}\left(y[x_j,\cdots,x_{j+\ell-1}] - y[x_{j+1},\cdots,x_{j+\ell}]\right)$$

Bevis slut.

§3. Hermite polynomier.

Har vi en funktion på et interval, f, kan vi uden videre anvende Newtons fremstilling på punkterne (x_j,y_j), hvor $y_j = f(x_j)$. Vi skriver nu

$$f[x_{j_o},\cdots,x_{j_\ell}] = y[x_{j_o},\cdots,x_{j_\ell}],$$

og har fremstillingen af polynomiet

$$p(x) = f[x_o] + (x-x_o)f[x_o,x_1]+\cdots+(x-x_o)\cdots(x-x_{m-1})f[x_o,\cdots,x_m]$$

Vi kan opfatte p som en approximation til f på intervallet, som er nøjagtig i $x_o,\cdots,x_m$. Vi skriver

$$f(x) = p(x) + R(x),$$

hvor altså R er nul i $x_o,\cdots,x_m$. Lad os nu tilføje punktet $x = x_{m+1}$. Da gælder

$$f(x) = f[x_o] + (x-x_o)f[x_o,x_1] +\cdots+(x-x_o)\cdots(x-x_{m-1})f[x_o,\cdots,x_m]$$
$$+ (x-x_o)\cdots(x-x_m)f[x_o,\cdots,x_m,x]$$

når blot $x \neq x_o,\cdots,x_m$. Vi har altså

$$R(x) = (x-x_o)\cdots(x-x_m)f[x_o,\cdots,x_m,x]$$

for $x \notin \{x_o,\cdots,x_m\}$.

Hvis nu f antages differentiabel i $x_o,\cdots,x_m$, så bliver også R differentiabel i disse punkter. Derved kan $f[x_o,\cdots,x_m,x]$ som funktion af x udvides til punkterne $x_o,\cdots,x_m$ ved definitionen

$$f[x_o,\cdots,x_m,x_j] = \frac{R'(x_j)}{\prod_{\substack{k=0\\k\neq j}}^{m}(x_j-x_k)}$$

idet vi ved, at $R(x_j) = 0$. Med denne udvidelse gælder formlen for $R(x)$ uden indskrænkning.

Mere generelt har vi:

Sætning VII3.1. *Lad* f *være* $m+1$ *gange differentiabel på intervallet* I. *For vilkårlige tal* $x_0,\cdots,x_{m+1}$, *og* $0 \le \ell \le m+1$, *defineres*

$$f[x_0,\cdots,x_\ell] = \int_0^1\int_0^{t_1}\cdots\int_0^{t_{\ell-1}} f^{(\ell)}(x_0+t_1(x_1-x_0)+\cdots+t_\ell(x_\ell-x_{\ell-1}))dt_\ell\cdots dt_1.$$

Da gælder rekursionsformlen for $x \in I$

$$f[x_0,\cdots,x_{\ell-1},x] = f[x_0,\cdots,x_{\ell-1},x_\ell] + (x-x_\ell)f[x_0,\cdots,x_\ell,x]$$

og Hermite's interpolationsformel for $x \in I$

$$f(x) = f[x_0] + (x-x_0)f[x_0,x_1]+\cdots+(x-x_0)\cdots(x-x_{m-1})f[x_0,\cdots,x_m]$$
$$+ (x-x_0)\cdots(x-x_m)f[x_0,\cdots,x_m,x].$$

Bevis.

$$f(x) = f(x_0) + \int_0^1 \frac{df(x_0+t(x-x_0))}{dt}dt$$
$$= f(x_0) + (x-x_0)\int_0^1 f'(x_0+t(x-x_0))dt$$
$$= f[x_0] + (x-x_0)f[x_0,x].$$

Lad nu for $x_0,\cdots,x_\ell,x$ vilkårlige

$$\varphi(t \quad\quad) = f^{(\ell)}(x_0+t_1(x_1-x_0)+\cdots+t_{\ell+1}(x-x_\ell)).$$

Da er

$$\int_0^{t_\ell} \frac{\partial\varphi}{\partial t_{\ell+1}}(t)dt \quad\quad = \varphi(\quad\quad,t_\ell) - \varphi(\quad\quad,0),$$

Sætning VII2.1. *Interpolationspolynomiet* $p(x) = p_{0,\dots,m}(x)$ *for punkterne* $(x_0,y_0),\dots,(x_m,y_m)$ *har fremstillingen*

$$p(x) = y[x_0] + (x-x_0)y[x_0,x_1]+\dots+(x-x_0)\cdots(x-x_{m-1})y[x_0,\dots,x_m]$$

(kaldet Newtons fremstilling), hvor

$$y[x_0,\dots,x_\ell] = \frac{1}{\ell!}\,\frac{d^\ell p_{0,\dots,\ell}}{dx^\ell}$$

kan beregnes rekursivt efter forskriften

$$y[x_j] = y_j \qquad j = 0,\dots,m$$

$$y[x_j,\dots,x_{j+\ell}] = \frac{1}{x_j-x_{j+\ell}}\left(y[x_j,\dots,x_{j+\ell-1}] - y[x_{j+1},\dots,x_{j+\ell}]\right)$$

for $j = 0,\dots,m-\ell$; $\ell = 1,\dots,m$.

Bevis. $r_\ell(x) = p_{0,\dots,\ell}(x) - p_{0,\dots,\ell-1}(x)$

er et polynomium af grad højst ℓ, som er 0 i punkterne $x_0,\dots,x_{\ell-1}$. Derfor findes γ_ℓ, så

$$r_\ell(x) = \gamma_\ell(x-x_0)\cdots(x-x_{\ell-1}).$$

Den ℓ-te afledede af $p_{0,\dots,\ell-1}$ er 0, så den ℓ-te afledede af r_ℓ og $p_{0,\dots,\ell}$ er den samme (og konstant):

$$\gamma_\ell = \frac{1}{\ell!}\,\frac{d^\ell r_\ell}{dx^\ell} = \frac{1}{\ell!}\,\frac{d^\ell p_{0,\dots,\ell}}{dx^\ell} = y[x_0,\dots,x_\ell].$$

Altså er

$$p_{0,\dots,m}(x) = y_0 + \sum_{\ell=1}^{m} r_\ell(x) = y_0 + \sum_{\ell=1}^{m}(x-x_0)\cdots(x-x_{\ell-1})y[x_0,\dots,x_\ell].$$

Endelig fås af rekursionsformlen for $p_{j,\dots,j+\ell}$ at

hvoraf

$$\int_o^{t_\ell}(x-x_\ell)f^{(\ell+1)}(x_o+t_1(x_1-x_o)+\cdots+t_{\ell+1}(x-x_\ell))dt_{\ell+1}$$

$$= f^{(\ell)}(x_o+t_1(x_1-x_o)+\cdots+t_\ell(x_\ell-x_{\ell-1}) + t_\ell(x-x_\ell))$$

$$- f^{(\ell)}(x_o+t_1(x_1-x_o)+\cdots+t_\ell(x_\ell-x_{\ell-1}) + 0\cdot(x-x_\ell)).$$

Ved yderligere integration ℓ gange fås

$$\int_o^1\cdots\int_o^{t_{\ell-1}}\int_o^{t_\ell}(x-x_\ell)f^{(\ell+1)}(x_o+\cdots+t_{\ell+1}(x-x_\ell))dt_{\ell+1}\cdots dt_1$$

$$= \int_o^1\cdots\int_o^{t_{\ell-1}}f^{(\ell)}(x_o+\cdots+t_\ell(x-x_{\ell-1}))dt_\ell\cdots dt_1$$

$$- \int_o^1\cdots\int_o^{t_{\ell-1}}f^{(\ell)}(x_o+\cdots+t_\ell(x_\ell-x_{\ell-1}))dt_\ell\cdots dt_1;$$

det samme som

$$(x-x_\ell)f[x_o,\cdots,x_\ell,x] = f[x_o,\cdots,x_{\ell-1},x] - f[x_o,\cdots,x_\ell].$$

Dette er rekursionsformlen, hvoraf Hermite's interpolationsformel følger ved induktion.

Bevis slut.

Bemærkning 1. Hvis $x_o,\cdots,x_m$ er forskellige, så er rekursionsformlen den samme som Newton. Derfor er $f[x_o,\cdots,x_m]$ i overensstemmelse med den tidligere definition.

Bemærkning 2. Hvis $x_0 = x_1 = \cdots = x_m$, da er

$$f[x_0,\cdots,x_0] = \int_0^1 \cdots \int_0^{t_{\ell-1}} f^{(\ell)}(x_0)\,dt_\ell \cdots dt_1 = \frac{1}{\ell!} f^{(\ell)}(x_0),$$

så Hermite's interpolationsformel bliver

$$f(x) = f(x_0) + (x-x_0)f'(x_0) + \cdots + \frac{1}{m!}(x-x_0)^m f^{(m)}(x_0) + R_m(x),$$

hvor $R_m(x) = (x-x_0)^{m+1} \int_0^1 \cdots \int_0^{t_m} f^{(m+1)}(x_0+t(x-x_0))\,dt \cdots dt_1$

$$= \frac{1}{m!}(x-x_0)^{m+1} \int_0^1 f^{(m+1)}(x_0+t(x-x_0))(1-t)^m dt$$

også kendt som Taylors formel.

Corollar. VII*3.1.*

$$f[x_0,\cdots,x_\ell] = \frac{1}{\ell!} f^{(\ell)}(\xi), \quad \min_{0\le j\le \ell} x_j \le \xi \le \max_{0\le j\le \ell} x_j$$

$$R(x) = \frac{(x-x_0)\cdots(x-x_m)}{(m+1)!} f^{(m+1)}(\xi), \quad \min_j x_j \le \xi \le \max_j x_j.$$

Bevis. Følger umiddelbart af middelværdisætningen for funktioner. Thi hvis $\mu_0 \le f^{(\ell)}(x) \le \mu_1$, når x løber i intervallet, så fås vurderingen

$$\int_0^1 \int_0^{t_1} \cdots \int_0^{t_{\ell-1}} \mu_0 dt \cdots dt_1 \le f[x_0,\cdots,x_\ell] \le \int_0^1 \int_0^{t_1} \cdots \int_0^{t} \mu_1 dt_\ell \cdots dt_1,$$

det samme som

$$\frac{\mu_0}{\ell!} \le f[x_0,\cdots,x_\ell] \le \frac{\mu_1}{\ell!}.$$

Altså findes ξ i intervallet, så

$$f^{(\ell)}(\xi) = \ell! f[x_0,\cdots,x_\ell].$$

Bevis slut.

Kapitel VIII Numerisk integration eller quadratur.

§1. Quadraturformler.

Lad $f: I \to \mathbb{R}$ være en integrabel funktion. Ved en *quadraturformel* for $J(f) = \int_a^b f(x)dx$ forstås et udtryk af formen

$$Q(f) = (b-a)\sum_{j=0}^{m} \alpha_j f(x_j),$$

hvor $x_0,\cdots,x_m \in I$, $[a,b] \subseteq I$, $\alpha_0,\cdots,\alpha_j \in \mathbb{R}$. Restleddet $E(f) = J(f) - Q(f)$ måler, hvor godt quadraturformlen approximerer integralet.

En sådan formel kan f.eks. findes ved at interpolere $f(x)$ og derefter integrere interpolationspolynomiet. Af

$$f(x) = p_{0,\cdots,m}(x) + R(x)$$

fås

$$\int_a^b f(x)dx = \int_a^b p_{0,\cdots,m}(x)dx + \int_a^b R(x)dx.$$

Er $x_0,\cdots x_m$ parvis forskellige, og er $p_{0,\cdots,m}$ udtrykt ved Lagrange-polynomier

$$p_{0,\cdots,m}(x) = \sum_{j=0}^{m} f(x_j)L_j(x),$$

fås quadraturformlen

$$Q(f) = \int_a^b p_{0,\cdots,m}(x) = (b-a)\sum_{j=0}^{m} \alpha_j f(x_j), \quad \text{hvor}$$

$$\alpha_j = \frac{1}{b-a}\int_a^b L_j(x)dx, \qquad j = 0,\cdots,m.$$

Vi havde restleddet på formen (Kap.VII§3)

$$R(x) = (x-x_o)\cdots(x-x_m)f[x_o,\cdots,x_m,x];$$

hvoraf

$$E(f) = \int_a^b R(x)dx = \int_a^b (x-x_o)\cdots(x-x_m)f[x_o,\cdots,x_m,x]dx.$$

Bruger vi Hermite-polynomier, når f er tilstrækkelig ofte differentiabel, kan $x_o,\cdots,x_m$ være vilkårlige. Af Hermites interpolationsformel i sætning VII.3.1 fås quadraturformlen

$$Q(f) = \int_a^b p_{o,\ldots,m}(x)dx = \sum_{j=0}^{m} \beta_j (b-a)^{j+1} f[x_o,\cdots,x_j],$$

hvor

$$\beta_o = 1,$$

$$\beta_j = \frac{1}{(b-a)^{j+1}} \int_a^b (x-x_o)\cdots(x-x_{j-1})dx = \int_o^1 (z-z_o)\cdots(z-z_{j-1})dz,$$

$$z_j = \frac{x_j-a}{b-a}, \quad j = 0,\ldots,m,$$

og restleddet

$$E(f) = \int_a^b R(x)dx = \int_a^b (x-x_o)\cdots(x-x_m)f[x_o,\cdots x_m,x]dx.$$

Er nu I et interval, så $a,b,x_o,\cdots,x_m \in I$, får vi af corollar VII.3.1. vurderingen

$$|E(f)| \le \frac{(b-a)^{m+2}}{(m+1)!} \int_o^1 |(z-z_o)\cdots(z-z_m)|dz \cdot \max_{x\in I} |f^{(m+1)}(x)|.$$

Hvis $(x-x_o)\cdots(x-x_m)$ ikke skifter fortegn i $[a,b]$, vil jo

$$\int_o^1 |(z-z_o)\cdots(z-z_m)|dz = \left|\int_o^1 (z-z_o)\cdots(z-z_m)dz\right| = |\beta_{m+1}|$$

hvoraf

$$|E(f)| \leq \frac{|\beta_{m+1}|}{(m+1)!}(b-a)^{m+2} \max_{x\in I}|f^{(m+1)}(x)|,$$

og

$$E(f) = \beta_{m+1}(b-a)^{m+2} f[x_o,\cdots,x_m,\xi], \quad \xi \in I,$$

eller

$$E(f) = \frac{\beta_{m+1}}{(m+1)!}(b-a)^{m+2} f^{(m+1)}(\xi), \quad \xi \in I.$$

Benyttes formlen på funktionen $f(x) = x^{m+1}$, fås

$$\beta_{m+1} = \frac{E(x^{m+1})}{(b-a)^{m+2}} .$$

Er f specielt et polynomium af grad $\leq m$, bliver $f = p_{o,\ldots,m}$ og restleddet 0. Derfor bliver quadraturformlerne eksakte i disse tilfælde.

Når det således gælder, at

$$E(1) = E(x) = \cdots = E(x^n) = 0, \quad E(x^{n+1}) \neq 0,$$

siger man, at quadraturformlen er af nøjagtighedsgrad n.
Bemærk, at $E(1) = 0 \Leftrightarrow \sum_{j=0}^{m} \alpha_j = 1$.

§2. Specielle quadraturformler.

1. *Kordetrapezformlen.*

$$Q(f) = \frac{b-a}{2}(f(a)+f(b)).$$

Vi har valgt $x_o = a$, $x_1 = b$, $m = 1$, og finder

$$\beta_o = 1, \quad \beta_1 = \int_o^1 z\,dz = \frac{1}{2}, \quad \beta_2 = \int_o^1 z(z-1)\,dz = -\frac{1}{6}.$$

$$f[a] = f(a), \quad f[a,b] = \frac{f(b)-f(a)}{b-a}$$

$$Q(f) = (b-a)f(a) + \frac{1}{2}(b-a)^2 f[a,b]$$

$$|E(f)| \leq \frac{1}{12}(b-a)^3 \max_{a\leq x\leq b} |f''(x)|$$

$$E(f) = -\frac{1}{12}(b-a)^3 f''(\xi) \qquad a \leq \xi \leq b.$$

2. *Tangenttrapezformlen.*

$$Q(f) = (b-a)f(c), \quad c = \frac{a+b}{2}.$$

Vi har valgt $x_o = x_1 = c$, $m = 1$, og finder

$$\beta_o = 1, \quad \beta_1 = \int_o^1 (z-\tfrac{1}{2})\,dz = 0, \quad \beta_2 = \int_o^1 (z-\tfrac{1}{2})^2 dz = \frac{1}{12}.$$

$$Q(f) = (b-a)f(c) + 0$$

$$|E(f)| \leq \frac{1}{24}(b-a)^3 \max_{a\leq x\leq b} |f''(x)|$$

$$E(f) = \frac{1}{24}(b-a)^3 f''(\xi), \qquad a \leq \xi \leq b.$$

1. + 2. Hvis f er konveks, kan vi endda slutte, at integralet ligger mellem de to quadraturer:

$$(b-a)f(c) \leq \int_a^b f(x)dx \leq \frac{(b-a)}{2}(f(a)+f(b)).$$

Er f konkav, vendes ulighederne.

3. *Simpsons formel eller Keplers tønderegel.*

$$Q(f) = \frac{1}{6}(b-a)(f(a)+4f(c)+f(b)), \quad c = \frac{a+b}{2}$$

Vi har valgt $x_o = a$, $x_1 = b$, $x_2 = x_3 = c$, $m = 3$, og finder med $z_o = 0$, $z_1 = 1$, $z_2 = z_3 = \frac{1}{2}$, at

$$\beta_o = 1, \quad \beta_1 = \int_o^1 z dz = \frac{1}{2}, \quad \beta_2 = \int_o^1 z(z-1)dz = -\frac{1}{6}$$

$$\beta_3 = \int_o^1 z(z-1)(z-\frac{1}{2})dz = 0, \quad \beta_4 = \int_o^1 z(z-1)(z-\frac{1}{2})^2 dz = -\frac{1}{120}.$$

Med $h = b-a$ bliver quadraturformlen

$$Q(f) = (b-a)(f(a)+\frac{h}{2}f[a,b]-\frac{h^2}{6}f[a,b,c]),$$

hvor $f[a,b] = \frac{1}{h}\cdot(f(b)-f(a))$, $f[a,b,c] = \frac{2}{h^2}(f(a)+f(b)-2f(c))$.
For restleddet får vi vurderingen

$$|E(f)| \leq \frac{1}{2880}(b-a)^5 \max_{a\leq x\leq b} |f^{(4)}(x)|$$

og fremstillingen

$$E(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \qquad a \leq \xi \leq b.$$

4. *Bessels formel.*

$$Q(f) = \frac{b-a}{24}(-f(a-h)+13f(a)+13f(b)-f(b+h)),\ h = b-a\ .$$

Vi har valgt $x_o = a$, $x_1 = b$, $x_2 = a-h$, $x_3 = b+h$, $m = 3$. Da interpolationspolynomiet $p(x)$ er uafhængigt af rækkefølgen af støttepunkterne, er

$$p(x) = p_{0,1,2,3}(x) = p_{1,0,3,2}(x)\ .$$

Med $c = \frac{a+b}{2}$ får vi derfor

$$p(x) = \frac{1}{2}(p_{0,1,2,3}(x)+p_{1,0,3,2}(x)) =$$

$$\frac{1}{2}(f(a)+f(b))+(x-c)f[a,b]+\frac{1}{2}(x-a)(x-b)(f[a,b,a-h]+f[a,b,b+h])$$

$$+(x-a)(x-b)(x-c)f[a,b,a-h,b+h],$$

som ved integration fra a til b giver

$$Q(f) = \int_a^b p(x)dx = \frac{h}{2}(f(a)+f(b))-\frac{h^3}{12}(f[a,b,a-h]+f[a,b,b+h]),$$

hvor $f[a,b,a-h]+f[a,b,b+h] = -\frac{1}{2h^2}(f(a)+f(b)-f(a-h)-f(b+h))$.

Til fejlvurderingen beregnes

$$\beta_4 = \int_o^1 z(z-1)(z+1)(z-2)dz = \frac{11}{30}\ .$$

Restleddet vurderes således

$$|E(f)| \leq \frac{11}{720}(b-a)^5 \max_{a-h\leq x\leq b+h} |f^{(4)}(x)|$$

og fremstilles

$$E(f) = \frac{11}{720}(b-a)^5 f^{(4)}(\xi),\quad a-h \leq \xi \leq b+h\ .$$

5. *Hermites formel.*

$$Q(f) = \frac{b-a}{2}(f(a)+f(b))+\frac{(b-a)^2}{12}(f'(a)-f'(b)) .$$

Vi har valgt $x_o = x_2 = a$, $x_1 = x_3 = b$, $m = 3$.

Med $z_o = z_2 = 0$ og $z_1 = z_3 = 1$ findes

$$\beta_o = 1,\ \beta_1 = \frac{1}{2},\ \beta_2 = -\frac{1}{6},\ \beta_3 = -\frac{1}{12},\ \beta_4 = \frac{1}{30} .$$

Med $h = b-a$ fås

$$Q(f) = (b-a)(f(a)+\frac{h}{2}f[a,b]-\frac{h^2}{6}f[a,b,a]-\frac{h^3}{12}f[a,b,a,b]) ,$$

hvor sætning VII,3,1giver formlerne

$$f[a,b,a,b] = \frac{1}{h}(f[a,b,b]-f[a,b,a])$$

$$f[a,b,a]+f[a,b,b] = \frac{1}{h}(f'(b)-f'(a)) .$$

Den første umiddelbart. Den anden ved

$$f[a,a] = f[a,b]+(a-b)f[a,b,a]$$

$$f[b,b] = f[b,a]+(b-a)f[b,a,b]$$

og $f[a,b] = f[b,a]$ samt $f[x,x] = f'(x)$.

Restleddet vurderes ved

$$|E(f)| \leq \frac{1}{720}(b-a)^5 \max_{a\leq x\leq b} |f^{(4)}(x)|$$

og fremstilles

$$E(f) = \frac{1}{720}(b-a)^5 f^{(4)}(\xi), \quad a \leq \xi \leq b .$$

6. *Euler-MacLaurins formel.*

$$Q(f) = \frac{b-a}{2}(f(a)+f(b)) + \sum_{j=1}^{n} \frac{B_{2j}}{(2j)!}(b-a)^{2j}(f^{(2j-1)}(a)-f^{(2j-1)}(b)),$$

hvor B_{2j} er Bernoulli-tallene. Selv om formlen åbenbart generaliserer kordetrapez-formlen ($n = 0$) og Hermites formel ($n = 1$), fås den ikke umiddelbart af interpolationsformlerne.

Bernoulli-polynomierne og Bernoulli-tallene,

$B_p(t)$ og B_p, for $p = 0,1,2,\ldots$ defineres ved formlerne

$B_o(t) = 1, \quad B_o = 1,$

$B_p(t) = p\int_o^t B_{p-1}(\tau)d\tau + B_p$ så $\int_o^1 B_p(t)dt = 0$.

Herved bliver graden af B_p netop p og $\frac{1}{p}(B_p(t)-B_p)$ en stamfunktion til $B_{p-1}(t)$. Man finder

$$B_1(t) = t - \frac{1}{2} \qquad B_1 = -\frac{1}{2}$$

$$B_2(t) = t^2 - t + \frac{1}{6} \qquad B_2 = \frac{1}{6}$$

$$B_3(t) = t^3 - \frac{3}{2}t^2 + \frac{1}{2}t \qquad B_3 = 0$$

$$B_4(t) = t^4 - 2t^3 + t^2 - \frac{1}{30} \qquad B_4 = -\frac{1}{30} .$$

For p lige, f.eks. $p = 2,4$, gælder, at $B_p(t)$ er symmetrisk om linien $t = \frac{1}{2}$ i intervallet $[0,1]$, samt (for $p > 0$) at $\int_o^1 B_p(t) = 0$. Heraf følger, at en stamfunktion f til $B_p(t)$ opfylder $f(x) = -f(1-x)$ samt for $p > 0$ at $f(0) = f(1) = f(\frac{1}{2}) = 0$. Thi

$$f(x) = \int_o^x B_p(t)dt = \int_{1-x}^1 B_p(t)dt = -\int_o^{1-x} B_p(t)dt \quad \text{for} \quad p > 0$$

og $\int_0^{\frac{1}{2}} B_p(t)dt = \int_{\frac{1}{2}}^1 B_p(t)dt = \frac{1}{2}\cdot\int_0^1 B_p(t)dt = 0$

hvoraf følger, at $f(0) = f(1) = 0 = f(\frac{1}{2})$ og af antisymmetrien $f(x) = -f(1-x)$ følger, at

$$\int_0^1 f(x)dx = \int_0^{\frac{1}{2}} f(x)dx + \int_{\frac{1}{2}}^1 f(x)dx = \int_0^{\frac{1}{2}} f(x)dx + \int_0^{\frac{1}{2}} f(1-x)dx = 0 .$$

Altså er $B_p = 0$ for p ulige og $p > 1$.

Når $B_p(t)$ for p ulige er symmetrisk om <u>punktet</u> $t = \frac{1}{2}$ og opfylder $B_p(0) = B_p(\frac{1}{2}) = B_p(1) = 0$ som for $p \geq 3$, så må

$$\int_x^{1-x} B_p(t)dt = 0 \quad \text{for alle} \quad x \in [0,1],$$

specielt gælder

$$\int_0^x B_p(t)dt = \int_0^{1-x} B_p(t)dt,$$

så denne stamfunktion er symmetrisk om linien $t = \frac{1}{2}$. Dette gælder så også for $B_{p+1}(t)$, som tillige opfylder $\int_0^1 B_{p+1}(t)dt = 0$.

Hvis $B_p(t)$ for p ulige har et nulpunkt i $[0,1]$ foruden $0, \frac{1}{2}, 1$, så må den have endnu et på grund af symmetrien, altså ialt mindst 5. Men så har $B_p(t)$ mindst 4 ekstremer i $]0,1[$, og da $B_{p-1}(t)$ er proportional med $B_p'(t)$, har $B_{p-1}(t)$ altså mindst 4 nulpunkter i $]0,1[$. Tilsvarende må $B_{p-2}(t)$ have mindst 3 nulpunkter i $]0,1[$, og hertil 0 og 1, altså ialt 5. Men da $B_3(t)$ har grad 3 og derfor kun de 3 nulpunkter $0, \frac{1}{2}, 1$, kan dette ikke være rigtigt. Altså gælder, at $B_p(t)$ for p ulige har højst 3 nulpunkter og for $p \geq 3$ netop de 3 nulpunkter $0, \frac{1}{2}, 1$.

Men heraf følger af

$$B_p(t)-B_p = p\int_o^t B_{p-1}(\tau)\,d\tau$$

for p lige, at $B_p(t)-B_p$ kun har nulpunkterne 0 og 1. Bemærk, at $B_p(0) = B_p(1) = B_p$.

Ved delt integration fås

$$\int_a^b f(x)\,dx = \int_a^b B_o \cdot f(x)\,dx = [(b-a)B_1(\tfrac{x-a}{b-a})f(x)]_a^b - \int_a^b (b-a)B_1(\tfrac{x-a}{b-a})f'(x)dx$$

$$= \frac{b-a}{2}(f(a)+f(b)) - (b-a)\int_a^b B_1(\tfrac{x-a}{b-a})f'(x)\,dx\ ,$$

og ved delt integration af andet led fås

$$-(b-a)\int_a^b B_1(\tfrac{x-a}{b-a})f'(x)\,dx = -(b-a)[\frac{(b-a)}{2}B_2(\tfrac{x-a}{b-a})f'(x)]_a^b$$

$$+(b-a)\int_a^b \frac{(b-a)}{2}B_2(\tfrac{x-a}{b-a})f''(x)\,dx =$$

$$-\frac{B_2}{2}\cdot(b-a)^2(f'(b)-f'(a))+\frac{(b-a)^2}{2}\int_a^b B_2(\tfrac{x-a}{b-a})f''(x)\,dx\ .$$

Således går det videre, den almindelige formel er

$$(-1)^n\frac{(b-a)^n}{n!}\int_a^b B_n(\tfrac{x-a}{b-a})f^{(n)}(x)\,dx =$$

$$(-1)^n\frac{(b-a)^n}{n!}[\frac{b-a}{n+1}B_{n+1}(\tfrac{x-a}{b-a})f^{(n)}(x)]_a^b$$

$$-(-1)^n\frac{(b-a)^{n+1}}{(n+1)!}\int_a^b B_{n+1}(\tfrac{x-a}{b-a})f^{(n+1)}(x)\,dx =$$

$$(-1)^n\frac{(b-a)^{n+1}}{(n+1)!}B_{n+1}(f^{(n)}(b)-f^{(n)}(a))$$

$$+(-1)^{n+1}\frac{(b-a)^{n+1}}{(n+1)!}\int_a^b B_{n+1}(\tfrac{x-a}{b-a})f^{(n+1)}(x)\,dx\ .$$

Ved summation af ligningerne og $B_n = 0$ for n ulige fås Euler-MacLaurins formel med restleddet

$$E(f) = \frac{(b-a)^{2n}}{(2n)!}\int_a^b B_{2n}\left(\frac{x-a}{b-a}\right) f^{(2n)}(x)\,dx$$

$$= -\frac{(b-a)^{2n+1}}{(2n+1)!}\int_a^b B_{2n+1}\left(\frac{x-a}{b-a}\right) f^{(2n+1)}(x)\,dx\ .$$

Vi laver nu en delt integration med stamfunktionen $\frac{1}{2n+2}(B_{2n+2}(t)-B_{2n+2})$ til $B_{2n+1}(t)$, som er 0 i endepunkterne.

$$E(f) = -\frac{(b-a)^{2n+1}}{(2n+1)!}\left[\frac{b-a}{2n+2}\left(B_{2n+2}\left(\frac{x-a}{b-a}\right)-B_{2n+2}\right) f^{(2n+1)}(x)\right]_a^b$$

$$+ \frac{(b-a)^{2n+2}}{(2n+2)!}\int_a^b \left(B_{2n+2}\left(\frac{x-a}{b-a}\right)-B_{2n+2}\right) f^{(2n+2)}(x)\,dx\ .$$

Da første led er 0, fås altså $E(f)$ lig med sidste led. Nu er $B_{2n+2}\left(\frac{x-a}{b-a}\right)-B_{2n+2}$ af konstant fortegn i $[a,b]$. Derfor findes $\xi \in [a,b]$, så

$$E(f) = \frac{(b-a)^{2n+2}}{(2n+2)!} f^{(2n+2)}(\xi)\int_a^b \left(B_{2n+2}\left(\frac{x-a}{b-a}\right)-B_{2n+2}\right)dx$$

$$= -\frac{(b-a)^{2n+3}}{(2n+2)!} B_{2n+2} f^{(2n+2)}(\xi),$$

da $\int_a^b B_{2n+2}\left(\frac{x-a}{b-a}\right)dx = 0$.

§3. Summerede quadraturformler.

Til beregning af integralet

$$J(f) = \int_a^b f(x)dx$$

deler vi intervallet i små delintervaller

$$[a,b] = \bigcup_{\nu=1}^{N} [a_\nu, b_\nu],$$

hvor $a_1 = a$, $b_N = b$, $a_{\nu+1} = b_\nu$ for $\nu = 1,\ldots,N-1$ og

$a_1 < a_2 < \cdots < a_{N+1} < b$, hvorefter vi anvender quadraturformler på hvert interval

$$\int_{a_\nu}^{b_\nu} f(x)dx = Q_\nu(f) + E_\nu(f), \quad \nu = 1,\ldots,N$$

og finder

$$J(f) = \sum_{\nu=1}^{N} \int_{a_\nu}^{b_\nu} f(x)dx = \sum_{\nu=1}^{N} Q_\nu(f) + \sum_{\nu=1}^{N} E_\nu(f) .$$

I almindelighed vælges småintervallerne af samme længde

$$h = b_\nu - a_\nu = \frac{b-a}{N}, \quad \nu = 1,\ldots,N,$$

og der bruges samme quadraturformel på hvert interval. Da kan restleddet fremstilles for f tilstrækkelig differentiabel

$$E(f) = \sum_{\nu=1}^{N} E_\nu(f) = \frac{\beta_{m+1}}{(m+1)!} h^{m+2} \sum_{\nu=1}^{N} f^{(m+1)}(\xi_\nu)$$

$$= (b-a)\frac{\beta_{m+1}}{(m+1)!} h^{m+1} f^{(m+1)}(\xi) .$$

Thi $f^{(m+1)}$ er kontinuert og antager derfor sin middelværdi

$$\frac{1}{N} \sum_{\nu=1}^{N} f^{(m+1)}(\xi_\nu)$$

i intervallet $[a,b]$.

1. *Den summerede kordetrapez-formel.*

$$Q(f) = \sum_{\nu=1}^{N} \frac{b_\nu - a_\nu}{2}(f(a_\nu)+f(b_\nu)) = h(\frac{1}{2}f(a)+\sum_{\nu=1}^{N-1} f(b_\nu)+\frac{1}{2}f(b))$$

med restleddet

$$E(f) = -(b-a)\frac{h^2}{12}f''(\xi) .$$

2. *Den summerede tangenttrapezformel.*

$$Q(f) = \sum_{\nu=1}^{N} (b_\nu - a_\nu) f(\frac{a_\nu+b_\nu}{2}) = h \sum_{\nu=1}^{N} f(\frac{b_{\nu-1}+b_\nu}{2}) ,$$

hvor $b_o = a$, med restleddet

$$E(f) = (b-a)\frac{h^2}{24}f''(\xi) .$$

3. *Den summerede Simpson-formel.*

$$\begin{aligned} Q(f) &= \sum_{\nu=1}^{N} \frac{b_\nu - a_\nu}{6}(f(a_\nu)+4f(\frac{a_\nu+b_\nu}{2})+f(b_\nu)) \\ &= \frac{h}{3}\left\{\frac{1}{2}f(a)+\sum_{\nu=1}^{N-1} f(b_\nu)+\frac{1}{2}f(b)+2\sum_{\nu=1}^{N} f(\frac{b_{\nu-1}+b_\nu}{2})\right\} \\ &= \frac{1}{3}(Q_K(f)+2Q_T(f)) , \end{aligned}$$

hvor Q_K er kordetrapezformlen og Q_T tangenttrapezformlen.

Restleddet er

$$E(f) = -(b-a)\frac{h^4}{2880}f^{(4)}(\xi) .$$

4. *Euler-MacLaurins sumformel.*

$$\begin{aligned} Q(f) = {} & h(\frac{1}{2}f(a)+\sum_{\nu=1}^{N-1} f(b_\nu)+\frac{1}{2}f(b))+ \\ & \sum_{j=1}^{n-1} \frac{B_{2j}}{(2j)!}h^{2j}(f^{(2j-1)}(a)-f^{(2j-1)}(b)) . \end{aligned}$$

Denne formel fremgår af den summerede kordetrapezformel ved et korrektionsled, der kun afhænger af intervallets endepunkter. Det tilhørende restled er

$$E(f) = -(b-a)\frac{B_{2n}}{(2n)!}h^{2n}f^{(2n)}(\xi).$$

Bemærkning. Er f en periodisk funktion med periode $b-a$, bliver korrektionsleddet åbenbart 0. I dette tilfælde har vi altså en række vurderinger på fejlen efter den summerede kordetrapezformel, så langt f er differentiabel.

5. *Romberg-integration.*

Vi betegner med $K(f)$ den summerede kordetrapezformel. Euler-MacLaurins sumformel lader sig skrive

$$Q(f) = K(f) + c_1h^2 + \cdots + c_{n-1}h^{2n-2},$$

hvor

$$c_j = \frac{B_{2j}}{(2j)!}(f^{(2j-1)}(a)-f^{(2j-1)}(b)), \quad j = 1,\ldots,n-1,$$

ikke afhænger af skridtlængden h. Det tilhørende restled er åbenbart $O(h^{2n})$.

Vi foretager nu en sukcessiv halvering af skridtlængden, $h_j = \frac{h}{2^j}$, $N_j = 2^jN$, og betegner de tilsvarende kordetrapezformler

$$K_j(f) = h_j(\tfrac{1}{2}f(a)+\sum_{\nu=1}^{N_j-1}f(a+\nu\cdot h_j)+\tfrac{1}{2}f(b)).$$

For j, $j+1$ fås altså ligningerne

$$\int_a^b f(x)dx = K_j(f)+c_1h_j^2+\ldots+c_{n-1}h_j^{2n-2}+E_j(f),$$

$$\int_a^b f(x)dx = K_{j+1}(f)+c_1(\tfrac{h_j}{2})^2+\cdots+c_{n-1}(\tfrac{h_j}{2})^{2n-2}+E_{j+1}(f).$$

Ganger vi anden ligning med 4, trækker den første fra og deler med 3, får vi, idet h_j^2 går ud,

$$\int_a^b f(x)dx = \frac{1}{3}(4K_{j+1}(f)-K_j(f))+c_2^{(1)}h_j^4+\cdots+c_{n-1}^{(1)}h_j^{2n-2}+E_j^{(1)}(f) .$$

Vi har herved opnået en quadraturformel, hvis fejl er af orden h^4. Deri er der ikke noget nyt, vi har blot genfundet den summerede Simpson-formel

$$K_j^{(1)} = \frac{1}{3}(4K_{j+1}(f)-K_j(f))$$
$$= \frac{h_j}{3}(\frac{1}{2}f(a)+\sum_{\nu=1}^{N_j-1} f(a+\nu h_j)+\frac{1}{2}f(b)+2\sum_{\nu=1}^{N_j} f(a+(\nu-\frac{1}{2})h_j)) .$$

Det nye er, at vi kan gentage processen og eliminere fejlens hovedled, $c_2^{(1)}h_j^4$ ved at gange $K_{j+1}^{(1)}$ med 16 etc. Vi får derved den generelle fremstilling

$$\int_a^b f(x)dx = K_j^{(k)}(f)+\sum_{l=k+1}^{n-1} c_l^{(k)}h_j^{2l}+E_j^{(k)}(f),$$

hvor $E_j^{(k)}(f) = O(h_j^{2n})$ og $K_j^{(k)}(f)$ findes af rekursionsformlerne

$$K_j^{(o)}(f) = K_j(f), \quad K_j^{(k+1)} = \frac{1}{4^{k+1}-1}(4^{k+1}K_{j+1}^{(k)}(f)-K_j^{(k)}(f))$$

for $j = 0,1,2,\ldots$, $k = 0,1,2,\ldots,n-1$, hvis f er $2n$ gange differentiabel.

Vi giver ikke en egentlig fejlvurdering, men udtrykker kun fejlens orden i h_j. Der gælder

$$\int_a^b f(x)dx = K_j^{(k)}(f)+O(h_j^{2(k+1)}) \; ; \; h_j = \frac{h}{2^j} ,$$

$$k = 0,\ldots,n-1 \; ; \quad j = 0,1,2,\ldots$$

Processen kan udmærket fortsættes ud over det n, hvor f er $2n$ gange differentiabel, men så kan vi ikke sige noget om fejlens orden. Kun at for f kontinuert vil $K_j^{(k)}$ konvergere

som Riemann sum mod integralet for $j \to \infty$.

Når $K_{j+1}(f)$ skal beregnes, har vi jo allerede regnet funktionsværdierne ud i halvdelen af punkterne til $K_j(f)$. Dette kan udnyttes, f.eks. ved at bruge formlen

$$K_{j+1}(f) = \frac{1}{2}(K_j(f)+T_j(f)) \; ; \quad T_j(f) = h_j \sum_{\nu=0}^{N_j-1} f(a+(\nu+\tfrac{1}{2})h_j),$$

hvor T_j altså er den j'te tangenttrapez-sumformel.

Romberg-integrationen fremstilles overskueligt i skemaet

$$\begin{array}{ccccccc}
K_o^{(o)} & & & & & & \\
K_1^{(o)} & K_o^{(1)} & & & & & \\
K_2^{(o)} & K_1^{(1)} & K_o^{(2)} & & & & \\
K_3^{(o)} & K_2^{(1)} & K_1^{(2)} & K_o^{(3)} & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & & \\
\cdot & \cdot & \cdot & \cdot & & \cdot & \\
\cdot & \cdot & \cdot & \cdot & & & \cdot \\
K_{j-1}^{(o)} & K_{j-2}^{(1)} & K_{j-3}^{(2)} & \cdot & \cdot & \cdot & K_o^{(j-1)} \\
K_j^{(o)} & K_{j-1}^{(1)} & K_{j-2}^{(2)} & \cdot & \cdot & K_1^{(j-1)} & K_o^{(j)}
\end{array}$$

der beregnes linie for linie med $K_o^{(j)}$ som "facit".

(Hvis ellers $h < 1$).

§4. Gauss' quadraturformler.

Quadraturformlerne i § 1 kan karakteriseres ved at for givne n+1 støttepunkter i intervallet vælges koefficienterne, så formlerne er eksakte for alle polynomier af grad $\leq n$. Gauss' idé er, at vi ved at vælge støttepunkterne optimale skal kunne øge nøjagtighedsgraden til 2n+1.

Er quadraturformlen på intervallet [a,b]

$$Q(f) = \sum_{k=o}^{n} \alpha_k f(x_k),$$

er vores krav, at de 2n+2 ligninger

$$\int_a^b x^j dx = \sum_{k=o}^{n} \alpha_k x_k^j \quad , \quad j = 0,1,\ldots,2n+1,$$

skal opfyldes af de 2n+2 ubekendte $\alpha_o,\ldots,\alpha_n,x_o,\ldots,x_n$. Dette ønske er ikke for ambitiøst, som skal ses nedenfor.

Ortogonale polynomier.

Lad μ være et mål på $\mathbb{R}$, **hvis masse ikke er koncentreret i endelig mange punkter,** så alle polynomier er integrable med hensyn til μ. F.eks. kan μ være

$$\mu(A) = \int_A 1_{[a,b]} dm \ ,$$

hvor m er Lebesque-målet og $-\infty < a \leq b < \infty$, eller

$$\mu(A) = \int_A e^{-x^2} dx \ , \text{ eller}$$

$$\mu(A) = \int_A 1_{[o,\infty[} \cdot e^{-x} dx \ .$$

Lad nu $\mathcal{P}_n$ betegne rummet af polynomier af grad $\leq n$ og $\mathcal{P}$ rummet af samtlige polynomier. Da er $\mathcal{P} \subseteq \mathcal{L}_2(\mu)$ og dermed forsynet med det indre produkt

$$(p,q) = \int_{\mathbb{R}} p \cdot q \, d\mu \ .$$

Rummene $\mathcal{P}_n$ har dimension n+1 over $\mathbb{R}$ eller $\mathbb{C}$ som det passer, og $\mathcal{P}_0 \subseteq \mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \cdots \subseteq \mathcal{P}$. Lad $p_0,p_1,p_2,\ldots$ være et ortogonalsystem af polynomier, så $\{p_0,p_1,\ldots,p_n\}$ er en basis for $\mathcal{P}_n$ og så alle har højestegradskoefficient 1.

Disse polynomier kan beregnes sukcessivt. $p_0 = 1$, og $p_1 = x-\beta_0$, hvor β_0 findes af ligningen

$$(p_0,p_1) = 0 \quad \text{d:} \quad (1,x-\beta_0) = 0,$$

$$(1,x) = \beta_0(1,1), \quad \text{hvoraf}$$

$$\beta_0 = \frac{(x,1)}{\|1\|^2} = \frac{(xp_0,p_0)}{\|p_0\|^2} .$$

For at finde p_{j+1} er det nok at vælge β_j,γ_j så

$$p_{j+1} = (x-\beta_j)p_j-\gamma_j p_{j-1} ;$$

thi for vilkårlige β_j,γ_j er p_{j+1} ortogonal på p_k, $k < j-1$, og af $(p_{j+1},p_j) = 0$ og $(p_{j+1},p_{j-1}) = 0$ fås

$$((x-\beta_j)p_j,p_j) = 0 \quad \text{og} \quad (\gamma_j \cdot p_{j-1},p_{j-1}) = (xp_j,p_{j-1})$$

hvoraf

$$\beta_j = \frac{(xp_j,p_j)}{\|p_j\|^2} \quad \text{og} \quad \gamma_j = \frac{(xp_j,p_{j-1})}{\|p_{j-1}\|^2} = \frac{\|p_j\|^2}{\|p_{j-1}\|^2} ,$$

som åbenbart gør p_{j+1} ortogonal på p_j og p_{j-1}. Den sidste ligning $(xp_j,p_{j-1}) = \|p_j\|^2$ følger af, at $(xp_j,p_{j-1}) = (p_j,xp_{j-1})$ og $\|p_j\|^2 = (p_j,p_j)$, så af $(p_j,xp_{j-1})-(p_j,p_j) = (p_j,xp_{j-1}-p_j) = 0$ følger påstanden. Da $xp_{j-1}-p_j$ har grad < j, er p_j forudsat ortogonal herpå.

Bemærk, at $p_0,p_1,p_2,\ldots$ i alle tilfælde får reelle koefficienter.

Sætning. *Et polynomium* p_n, *der er ortogonalt på* $\mathcal{P}_{n-1}$ *og har grad* n, *har lutter forskellige reelle rødder.*

Bevis. Indirekte. Har p_n en reel dobbeltrod eller et par af komplekse konjugerede rødder, så kan p_n skrives

$$p_n(x) = ((x-\mu)^2+\nu^2)r(x) ,$$

hvor r har grad $n-2$ og $(\mu,\nu) \neq (0,0)$, $\mu,\nu \in \mathbb{R}$, og r er et reelt polynomium. Men så er

$$0 = (p_n,r) = (((x-\mu)^2+\nu^2)r,r) =$$

$$((x-\mu)^2r,r)+\nu^2(r,r) =$$

$$\|(x-\mu)r\|^2+\nu^2\|r\|^2 .$$

Men så er $r = 0$ og dermed $p_n = 0$.
Bevis slut.

Vi kan altså skrive $p_n(x) = (x-\lambda_1)\cdots(x-\lambda_n)$, hvor $\lambda_1,\ldots,\lambda_n$ er reelle og forskellige. Med $q_k(x) = \prod_{j\neq k}(x-\lambda_j)$ gælder

$$0 = (p_n,q_k) = ((x-\lambda_k)q_k,q_k) = (xq_k,q_k)-\lambda_k\|q_k\|^2 ,$$

hvoraf $$\lambda_k = \frac{(xq_k,q_k)}{\|q_k\|^2} .$$

Nu er q_k proportionalt med Lagranges interpolationspolynomium L_k, så vi kan skrive

$$\lambda_k = \frac{(xL_k,L_k)}{\|L_k\|^2} .$$

Vi vender nu tilbage til problemet, om vi kan finde en quadraturformel, der er eksakt af grad $2n+1$. Hertil vælger vi først $x_0,\ldots,x_n$ som nulpunkter i p_{n+1}.

Valget af punkterne $x_0,\ldots,x_n$ som nulpunkterne i p_{n+1} er den eneste løsning. Thi lad

$$Q(f) = \sum_{j=0}^{n} \alpha_j f(x_j)$$

være en quadraturformel, der er eksakt op til grad $2n+1$. Da gælder for $q(x) = (x-x_0)\cdots(x-x_n)$ og $p(x) = x^k q(x)$, $k \leq n$, at $p \in \mathcal{P}_{2n+1}$, så

$$(x^k, q(x)) = (p,1) = Q(p) = \sum_{j=0}^{n} \alpha_j x_j^k (x_j-x_0)\cdots(x_j-x_n) = 0 ,$$

altså at q er ortogonal på $\mathcal{P}_n$.

Vi skal nu vise, at vælger vi $x_0,\ldots,x_n$ som rødderne i p_{n+1}, så kan vi finde $\alpha_0,\ldots,\alpha_n$, så den tilsvarende quadraturformel er eksakt for alle polynomier i $\mathcal{P}_{2n+1}$.

Lad derfor $p \in \mathcal{P}_{2n+1}$. Da findes $q,r \in \mathcal{P}_n$, så

$$p = q\cdot p_{n+1} + r .$$

For $x_0,\ldots,x_n$ gælder $p(x_j) = r(x_j)$, da $p_{n+1}(x_j) = 0$. Altså gælder

$$r(x) = \sum_{j=0}^{n} r(x_j) L_j(x) = \sum_{j=0}^{n} p(x_j) L_j(x) ,$$

og da p_{n+1} er ortogonal på $\mathcal{P}_n$, får vi

$$(p,1) = (q\cdot p_{n+1}+r,1) = (q,p_{n+1})+(r,1) = (r,1) =$$

$$\left(\sum_{j=0}^{n} p(x_j) L_j(x),1\right) = \sum_{j=0}^{n} p(x_j)(L_j(x),1) = Q(p) ,$$

hvor

$$Q(p) = \sum_{j=0}^{n} \alpha_j p(x_j) ,$$

som altså er nøjagtig op til grad $2n+1$, hvis blot

$$\alpha_j = \int_a^b L_j(x)\,d\mu ,$$

når vi ønsker at finde $(p,1) = \int_a^b p(x)\,d\mu$.

Vi kan finde et andet udtryk for α_j. Thi $(L_k(x))^2 \in \mathcal{P}_{2n}$, så Q er eksakt for dette polynomium. Altså gælder

$$\|L_k\|^2 = \int_a^b L_k(x)^2 d\mu = Q(L_k(x)^2) = \sum_{j=0}^{n} \alpha_j L_k(x_j) = \alpha_k .$$

Men så er også

$$\alpha_k = \|L_k\|^2 .$$

Restledsvurdering. Er f en funktion, og ønsker vi at beregne

$$J(f) = \int_a^b f(x)\,d\mu ,$$

interpolerer vi først f med Hermites interpolationsformel i punkterne $x_o, x_1, \ldots, x_n, x_o, x_1, \ldots, x_n$, og finder

$$f(x) = p(x) + R(x) \; ; \; R(x) = (x-x_o)^2 \cdots (x-x_n)^2 f[x_o, \ldots, x_n, x].$$

Nu er p af grad $2n+1$, så $J(p) = Q(p) = Q(f)$. Altså gælder

$$J(f) = J(p) + J(R) = Q(f) + \int_a^b f[x_o, \ldots, x_n, x] p_{n+1}^2 d\mu ,$$

og da $p_{n+1}^2(x) \geq 0$ for alle x, finder vi for restleddet

$$E(f) = \frac{1}{(2n+2)!} \int_a^b p_{n+1}^2(x)\,d\mu \cdot f^{(2n+2)}(\xi)$$
$$= \frac{\|p_{n+1}\|^2}{(2n+2)!} f^{(2n+2)}(\xi) .$$

Eksempel 1.

Lad $a = -1$, $b = 1$ og målet givet ved $1_{[a,b]}$. De tilsvarende ortogonale polynomier, *Legendre-polynomierne*, er

$$p_0(x) = 1 \qquad \|p_0\|^2 = 2$$

$$p_1(x) = x \qquad \|p_1\|^2 = \frac{2}{3}$$

$$p_2(x) = x^2 - \frac{1}{3} \qquad \|p_2\|^2 = \frac{8}{45}$$

$$p_3(x) = x^3 - \frac{3}{5}x \qquad \|p_3\|^2 = \frac{8}{175}$$

$$p_4(x) = x^4 - \frac{6}{7}x^2 + \frac{1}{7} \qquad \|p_4\|^2 = \frac{128}{11025}$$

p_2 har nulpunkterne $\pm \frac{1}{\sqrt{3}}$, p_3 har nulpunkterne $0, \pm \sqrt{\frac{3}{5}}$. Med variabeltransformation til et vilkårligt interval $[a,b]$ og $c = \frac{a+b}{2}$, $h = \frac{b-a}{2}$ fås første og anden Gauss' quadraturformel

$$\int_a^b f(x)dx = \frac{b-a}{2}\left(f\left(c-\frac{h}{\sqrt{3}}\right)+f\left(c+\frac{h}{\sqrt{3}}\right)\right)+\frac{(b-a)^5}{4320}f^{(4)}(\xi_1),$$

$$\int_a^b f(x)dx = \frac{b-a}{18}\left(5f\left(c-h\frac{\sqrt{15}}{5}\right)+8f(c)+5f\left(c+h\frac{\sqrt{15}}{5}\right)\right)+\frac{(b-a)^7}{201600}f^{(6)}(\xi_2).$$

Eksempel 2.

Lad $a = 0$, $b = \infty$ og målet være givet ved tætheden e^{-x}, $x \in [0,\infty[$. De tilsvarende ortogonale polynomier, Laguerre-polynomierne, er

$$L_0(x) = 1 \qquad \|L_0\|^2 = 1$$
$$L_1(x) = x - 1 \qquad \|L_1\|^2 = 1$$
$$L_2(x) = x^2 - 4x + 2 \qquad \|L_2\|^2 = 4$$
$$L_3(x) = x^3 - 9x^2 + 18x - 6 \qquad \|L_3\|^2 = 36$$
$$L_4(x) = x^4 - 16x^3 + 72x^2 - 96x + 24 \qquad \|L_4\|^2 = 576$$

L_2 har nulpunkterne $2 \pm \sqrt{2}$, og L_3 har nulpunkterne 0.4157745568, 2.29428036, 6.289945083. Heraf fås Gauss' Quadraturformel for $n = 1$

$$\int_0^\infty f(x)e^{-x}\,dx = \frac{2+\sqrt{2}}{4} f(2-\sqrt{2}) + \frac{2-\sqrt{2}}{4} f(2+\sqrt{2}) + \frac{f^{(4)}(\xi)}{6}.$$

For $n = 2$ fås knap så pænt

$$\int_0^\infty f(x)e^{-x}\,dx = 0.711093099\cdot f(0.4157745568) + 0.2785177336\cdot f(2.29428036) + 0.0103892565\cdot f(6.289945083) + \frac{f^{(6)}(\xi)}{20}.$$

Eksempel 3.

Lad $a = -\infty$ og $b = +\infty$ og målet være givet ved tætheden e^{-x^2}, $x \in \mathbb{R}$. De tilsvarende ortogonale polynomier, Hermite-polynomierne, er

$$H_0(x) = 1 \qquad ||H_0||^2 = \sqrt{\pi}$$

$$H_1(x) = x \qquad ||H_1||^2 = \tfrac{1}{2}\sqrt{\pi}$$

$$H_2(x) = x^2 - \tfrac{1}{2} \qquad ||H_2||^2 = \tfrac{1}{2}\sqrt{\pi}$$

$$H_3(x) = x^3 - \tfrac{3}{2}x \qquad ||H_3||^2 = \tfrac{3}{4}\sqrt{\pi}$$

$$H_4(x) = x^4 - 3x^2 + \tfrac{3}{4} \qquad ||H_4||^2 = \tfrac{3}{2}\sqrt{\pi}$$

$$H_5(x) = x^5 - 5x^3 + \tfrac{15}{4}x \qquad ||H_5||^2 = \tfrac{15}{4}\sqrt{\pi}$$

med nulpunkterne

H_2: $\pm \frac{\sqrt{2}}{2} = \pm\, 0.7071067812$

H_3: $0, \pm\sqrt{\frac{3}{2}} = \pm\, 1.224744871$

H_4: $\pm\sqrt{\frac{3 \pm \sqrt{6}}{2}} = \pm\, 0.5246476233, \pm\, 1.650680124$

H_5: $0, \pm\sqrt{\frac{5 \pm \sqrt{10}}{2}} = \pm\, 0.9585724646, \pm\, 2.02018287$

Heraf fås Gauss' quadraturformler for $n = 1$

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx = \frac{\sqrt{\pi}}{2}\left(f\left(-\frac{\sqrt{2}}{2}\right) + f\left(\frac{\sqrt{2}}{2}\right)\right) + \frac{\sqrt{\pi}}{48}f^{(4)}(\xi),$$

og for $n = 2$

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx = \frac{\sqrt{\pi}}{6}\left(f\left(-\sqrt{\tfrac{3}{2}}\right) + 4f(0) + f\left(\sqrt{\tfrac{3}{2}}\right)\right) + \frac{\sqrt{\pi}}{960}f^{(6)}(\xi),$$

og for $n = 3$

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx =$$

$$\frac{\sqrt{\pi}}{4}\left((1+\sqrt{\frac{2}{3}})(f(-\sqrt{\frac{3-\sqrt{6}}{2}}) + f(\sqrt{\frac{3-\sqrt{6}}{2}})) + (1-\sqrt{\frac{2}{3}})(f(-\sqrt{\frac{3+\sqrt{6}}{2}}) + f(\sqrt{\frac{3+\sqrt{6}}{2}}))\right) + \frac{\sqrt{\pi}}{3360}f^{(8)}(\xi).$$

samt for $n = 4$

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}dx = \frac{\sqrt{\pi}}{15}\left(8f(0) + \frac{7+2\sqrt{10}}{4}(f(-\sqrt{\frac{5-\sqrt{10}}{2}}) + f(\sqrt{\frac{5-\sqrt{10}}{2}})) + \frac{7-2\sqrt{10}}{4}(f(-\sqrt{\frac{5+\sqrt{10}}{2}}) + f(\sqrt{\frac{5+\sqrt{10}}{2}}))\right) + \frac{\sqrt{\pi}}{967680}f^{(10)}(\xi).$$

Kapitel IX. Numerisk løsning af sædvanlige differentialligninger.

§ 1. Problemformulering.

Lad $I = [a,b]$ være et reelt interval og $f: I \times \mathbb{R} \to \mathbb{R}$ en kontinuert funktion, og lad $\alpha \in \mathbb{R}$. Vi søger nu en funktion $u: I \to \mathbb{R}$ med kontinuert differentialkvotient, så

(*) $u(a) = \alpha$ og $u'(x) = f(x,u(x))$ for $x \in I$.

Vi deler nu intervallet I i N_h stykker med delepunkterne $I_h = \{a+jh \mid j = 0,\ldots,N_h\}$. I hvert punkt i $I_h' = \{a+jh \mid j = 0,\ldots,N_h-1\}$ kan vi approximere $u'(x)$ med differenskvotienten $\frac{u(x+h)-u(x)}{h}$. Disse differenskvotienter opfylder ligningen

$$\frac{1}{h}(u(x+h)-u(x)) = \frac{1}{h}\int_x^{x+h} f(t,u(t))dt \ , \quad x \in I_h' \ , \ u(a)=\alpha,$$

hvis u er en løsning til (*).

De metoder, vi her ser på, approximerer ovenstående integral med en quadraturformel og finder de nødvendige funktionsværdier ved hjælp af en Taylorudvikling af integranden og en elimination af u's afledede.

Kalder vi approximationen af højre side $f_h(x,u(x))$, får vi altså ligningen erstattet af

(★) $\frac{1}{h}(y_{j+1}-y_j) = f_h(x_j,y_j), \quad j = 0,\ldots,N_h-1$

$y_o = \alpha$.

Vi definerer nu løsningen til (✪) som $u_h(x_j) = y_j$, $x_j \in I_h'$. Derved begås for hvert j en *trunkeringsfejl*

$$T_h(x) = \frac{1}{h}\int_x^{x+h} f(t,u(t))dt - f_h(x,u(x)), \quad x \in I_h' ,$$

som ophobes for j voksende til N_h-1 til den samlede *diskretiseringsfejl*. Hertil kommer sædvanlige *afrundingsfejl*.

§ 2. Euler-metoder.

1. Det simpleste er approximation af $\frac{1}{h}\int_x^{x+h} f(t,u(t))dt$ med $f(x,u(x))$. Derved fås Euler metode

$$u_h(a) = \alpha \ , \quad \frac{1}{h}(u_h(x+h)-u_h(x)) = f(x,u_h(x)), \quad x \in I_h' \ .$$

2. En forbedring vil være at anvende tangenttrapezformlen til quadratur af integralet. Vi skal hertil bruge $u(x+\frac{h}{2})$, som findes af Taylors formel

$$u(x+\tfrac{h}{2}) = u(x)+\tfrac{h}{2}u'(x)+R(x) = u(x)+\tfrac{h}{2}f(x,u(x))+R(x),$$

så man får ligningerne

$$u_h(a) = \alpha \ , \quad \frac{1}{h}(u_h(x+h)-u_h(x)) = f(x+\tfrac{h}{2},\tilde{u}_h(x+\tfrac{h}{2})),$$

hvor $\tilde{u}_h(x+\frac{h}{2}) = u_h(x)+\frac{h}{2}f(x,u_h(x))$.

3. En anden forbedring vil være at anvende kordetrapezformlen til quadratur af integralet. Hertil behøves $u(x+h)$; vi får

$$u_h(a) = \alpha \ , \quad \frac{1}{h}(u_h(x+h)-u_h(x)) = \frac{1}{2}(f(x,u_h(x))+f(x+h,\tilde{u}_h(x+h)),$$

hvor $\tilde{u}_h(x+h) = u_h(x)+hf(x,u_h(x))$.

§ 3. Konsistens.

Differensligningerne (✯) siges at være *konsistente* med differentialligningen (*), hvis trunkeringsfejlen $T_h(x)$ for en løsning u til (*) går ligeligt mod 0 med h. Hvis der endvidere findes et $p > 0$, og $K > 0$, så

$$\max_{x\in I_h'} |T_h(x)| \leq Kh^p ,$$

siges (✯) at have *konsistensorden* p.

Konsistenssætningen:

$$\max_{x\in I_h'} |T_h(x)| \to 0 \quad \textit{for} \quad h \to 0 \iff$$

$$\max_{x\in I_h'} |f_h(x,u(x))-f(x,u(x))| \to 0 \quad \textit{for} \quad h \to 0,$$

når u *er en løsning til* (*).

Bevis. Af (*) og definitionen på $T_h(x)$ fås

$$T_h(x) = \frac{1}{h}(u(x+h)-u(x))-u'(x)+f(x,u(x))-f_h(x,u(x)),$$

og videre gælder

$$\frac{1}{h}(u(x+h)-u(x))-u'(x) = \frac{1}{h}\int_0^h (u'(x+t)-u'(x))dt, \quad a\leq x\leq b-h .$$

Da u løser (*) er u' kontinuert i [a,b], altså ligelig kontinuert, så

$$\max_{x\in I_h'} |\frac{1}{h}(u(x+h)-u(x))-u'(x)| \leq \max_{\substack{a\leq x\leq b-h \\ 0\leq t\leq h}} |u'(x+t)-u'(x)| \to 0 \quad \text{for } h \to 0 .$$

Bevis slut.

§ 4. Konsistens af Euler-metoderne.

1. I dette tilfælde er $f_h(x,u(x)) = f(x,u(x))$, så konsistenssætningen kan anvendes. Endda fås af Taylors formel

$$T_h(x) = \frac{1}{h}(u(x+h)-u(x))-u'(x) = h\int_0^1 u''(x+th)(1-t)dt$$

hvoraf

$$\max_{x\in I_h'} |T_h(x)| \leq K\cdot h, \quad \text{hvor} \quad K = \frac{1}{2}\max_{x\in I}|u''(x)|,$$

så metoden har konsistensorden 1.

2. Vi antager at f har begrænsede partielle afledede og at løsningerne u har begrænsede afledede af op til 3. orden. Lad $\|\ \|$ være max-normen og

$$\max_{x,y} \left|\frac{\partial f(x,y)}{\partial y}\right| \leq L.$$

Vi har da

$$T_h(x) = \frac{1}{h}\int_x^{x+h} u'(t)dt - f_h(x,u(x)),$$

hvor $f_h(x,u(x)) = f(x+\frac{h}{2},u(x)+\frac{h}{2}f(x,u(x))$.

Vi indskyder nu $u'(x+\frac{h}{2})$; thi tangenttrapezformlen opfylder vurderingen

$$\left|\frac{1}{h}\int_x^{x+h} u'(t)dt - u'(x+\frac{h}{2})\right| \leq \frac{h^2}{24}\|u'''\|.$$

Og på den anden side fås ved hjælp af Taylors formel for $u(x+\frac{h}{2})$

$$|u'(x+\tfrac{h}{2}) - f_h(x,u(x))| =$$

$$|f(x+\tfrac{h}{2},u(x+\tfrac{h}{2}))-f(x+\tfrac{h}{2},u(x)+\tfrac{h}{2}u'(x))| \leq$$

$$L\cdot|u(x+\tfrac{h}{2})-u(x)-\tfrac{h}{2}u'(x)| =$$

$$L\cdot|\tfrac{h^2}{4}\int_0^1 u''(x+\tfrac{h}{2}t)\,(1-t)\,dt| \leq$$

$$L\cdot\tfrac{h^2}{8}\cdot\|u''\| .$$

Denne forbedring af Eulers metode er altså konsistent af orden 2.

§ 5. Runge-Kutta-metoden.

Ved denne metode anvendes Simpsons formel til approximation af integralet. Funktionsværdierne til brug i Simpsons formel kaldes k_o, $\frac{k_1+k_2}{2}$, k_3, så vi skriver

$$f_h(x,y) = \frac{1}{6}(k_o+2k_1+2k_2+k_3)(x,y) ,$$

hvor vi har sat

$$k_o(x,y) = f(x,y)$$

$$k_1(x,y) = f(x+\tfrac{h}{2},y+\tfrac{h}{2}k_o)$$

$$k_2(x,y) = f(x+\tfrac{h}{2},y+\tfrac{h}{2}k_1)$$

$$k_3(x,y) = f(x+h,y+hk_2) .$$

Denne metode har konsistensorden 4.

Bevis.

Vi betragter

$$T_h(x) = \frac{1}{h}\int_x^{x+h} u'(t)dt - f_h(x,u(x)) .$$

Vi approximerer integralet med Simpsons formel, så idet vi sætter $u_j = u(x_j)$, $u_j' = u'(x_j)$ for $j = 0,1,2,3$ fås

$$\frac{1}{h}\int_x^{x+h} u'(t)dt = \frac{1}{6}(u_o' + 4u_1' + u_3') + 0(h^4) .$$

Sæt $x_o = x$, $y_o = y$, $x_1 = x_2 = x+\frac{h}{2}$, $y_1 = y+\frac{h}{2}k_o$, $y_2 = y+\frac{h}{2}k_1$, $x_3 = x+h$, $y_3 = y+hk_2$. Vi mangler blot at vise, at

$$f_h(x,u(x)) = \sum_{j=0}^{3} \gamma_j f(x_j,y_j) = \frac{1}{6}(u_o'+4u_1'+u_3')+0(h^4) .$$

Hertil bemærkes, at

$$u_o' = f(x_o,y_o) = k_o.$$

Dernæst udvikles u og u' efter Taylors formel fra $x_1 = x_2 = x+\frac{h}{2}$:

$$u_o = u(x) = u_1 - \frac{h}{2}u_1' + \frac{h^2}{8}u_1'' - \frac{h^3}{48}u_1''' + 0(h^4),$$

$$u_o' = u'(x) = u_1' - \frac{h}{2}u_1'' + \frac{h^2}{8}u_1''' + 0(h^3),$$

$$y_1 = u_o + \frac{h}{2}u_o' = u_1 - \frac{h^2}{8}u_1'' + \frac{h^3}{24}u_1''' + 0(h^4).$$

Ved Taylorudvikling af f som funktion af y ud fra u_1 fås, idet $y_1 = u_1 + 0(h^2)$,

$$k_1(x,u(x)) = f(x_1,y_1) = f(x_1,u_1)+(y_1-u_1)f_y(x_1,u_1)+0(h^4)$$
$$= u_1' - (\frac{h^2}{8}u_1'' - \frac{h^3}{24}u_1''')f_y(x_1,u_1)+0(h^4).$$

Videre findes

$$y_2 = u_o+\frac{h}{2}f(x_1, y_1) = u_1+\frac{h^2}{8}u_1'' -\frac{h^3}{48}u_1''' -\frac{h^3}{16}u_1''f_{y1} +0(h^4).$$

Ved Taylorudvikling af f som funktion af y ud fra u_1 fås, idet $y_2 = u_1 + 0(h^2)$,

$$k_2(x,u(x)) = f(x_2,y_2) = f(x_1,u_1)+(y_2-u_1)f_{y1}+0(h^4)$$
$$= u_1'+(\frac{h^2}{8}u_1'' - \frac{h^3}{48}u_1''' - \frac{h^3}{16}u_1''f_{y1})f_{y1} + 0(h^4).$$

Sidst udvikles u fra $x_1 = x_2$

$$u_3 = u(x+h) = u_1+\frac{h}{2}u_1'+\frac{h^2}{8}u_1'' + \frac{h^3}{48}u_1''' + 0(h^4).$$

Heraf fås med udviklingen til u_o, at

$$y_3 = u_o+hf(x_2,y_2) = u_o+hu_1'+\frac{h^3}{8}u_1''f_{y1} + 0(h^4)$$
$$= u_3 - \frac{h^3}{24}u_1''' + \frac{h^3}{8}u_1''f_{y1} + 0(h^4).$$

§ 6. Konvergens.

I virkeligheden er vi mest interesserede i, at løsningerne til differensligningerne (✡) konvergerer mod en løsning til differentialligningen (*). Hertil betragter vi diskretiseringsfejlen

$$r_h(x) = u_h(x) - u(x), \quad x \in I_h.$$

Fejlen må opfylde ligningen

$$\frac{1}{h}(r_h(x+h) - r_h(x)) = f_h(x,u_h(x)) - f_h(x,u(x)) - T_h(x), \quad x \in I_h'.$$

Her er trunkeringsfejlen styret, hvis metoden er konsistent. Lad os i det følgende antage, at f_h opfylder en Lipschitz-betingelse i y, ↔: $\exists\, L>0$ (Lipschitz-konstant):

$$|f_h(x,y) - f_h(x,y')| \leq L\cdot|y - y'|$$

for alle x og y i interessesfæren.

Konvergenssætning: *Hvis en approximation* f_h *til en differentialligning* (*) *opfylder en Lipschitz-betingelse i* y *og er konsistent af orden* p, *da vil løsningen* u_h *til* (✡) *være konvergent mod løsningen til* (*), u, *og*

$$|u_h(x) - u(x)| \leq K\cdot h^p (x-a) e^{L(x-a)}, \quad x \in I_h.$$

Bevis. Vi viser ved induktion efter $x \in I_h$, at

$$(\nabla) \quad |r_h(x)| \leq (x-a) \max_{x\in I_h'} |T_h(x)| e^{L(x-a)}, \quad x \in I_h.$$

Heraf følger sætningen umiddelbart.

For $x = a$ er intet at vise. Med $x' = x + h$ fås

$$|r_h(x')| \leq |r_h(x)| + h|f_h(x,u_h(x)) - f_h(x,u(x))| + h|T_h(x)|$$

$$\leq (1+hL)|r_h(x)| + h|T_h(x)|.$$

Da $r_h(x)$ opfylder (∇), fås

$$|r_h(x')| \leq (1+hL)(x-a) \max_{x\in I_h'} |T_h(x)| e^{L(x-a)} + h|T_h(x)|$$

$$\leq (e^{hL}\cdot(x-a)\cdot e^{L(x-a)} + h) \max_{x\in I_h'} |T_h(x)|$$

$$\leq (x' - a)e^{L(x'-a)} \max_{x\in I_h'} |T_h(x)|.$$

Bevis slut.

Bemærkning. Heraf følger ikke, at det mindste h er det bedste. Thi det forudsætter, at vi for et mindre valg af h også kan vælge en større nøjagtighed. Har vi en given nøjagtighed at regne med, bliver vurderingen af afrundingsfejlen proportional med antallet af skridt og dermed omvendt proportional med h. Den samlede fejlvurdering ved summen af de to fejlvurderinger ser derfor således ud: