TMA671

Ruben Seyer <rubense@student.chalmers.se>

31 maj 2019

Satser markerade med * är på listan. I slutet ingår även korta förklaringar, formler och teori för några områden. Materialet presenteras med reservation för eventuella fel – se över, tänk igenom och jämför själv för bästa möjliga förståelse!

Förteckning över satser

Definition (Linjärt rum och underrum) . . . 3
Sats* (L1.4, lösningsmängden till lin. avb.) . . . 4
Sats* (L1.5, bassatsen (del)) . . . 4
Sats* (L1.9, bassatsen (del)) . . . 4
Anmärkning (L1.18) . . . 4
Definition (Skalärprodukt) . . . 5
Sats* (L2.1, Cauchy-Schwarz olikhet) . . . 5
Sats* (L2.2, triangelolikheten) . . . 5
Sats* (L2.11, fyra fundamentala underrummen) . . . 6
Sats* (L4.6, oberoende av egenvektorer) . . . 7
Sats* (L4.10, symmetriska matriser har reella egenvärden) . . . 7
Definition (Felanalys och numerik) . . . 8

Innehåll

1 Projektioner **9**
1.1 Gram-Schmidts ortogonaliseringsprocess, L2.4 . . . 9

2 Numerisk ekvationslösning **9**
2.1 Newtons metod . . . 9
2.2 Sekantmetoden . . . 10
2.3 Fixpunktsiterationer . . . 10

3 Polynominterpolation **10**
3.1 Splines . . . 11

4 Numerisk integration — kvadraturformler **11**
4.1 Trapetsregeln . . . 11
4.2 Simpsons regel . . . 12

5 Faktoriseringar **12**
5.1 LU-faktorisering 12
5.2 QR-faktorisering 12
5.3 SVD (singular value decomposition) 13

6 Numerisk beräkning av derivator och lösning av ODE **13**
6.1 Begynnelsevärdesproblem 13

Definition (Linjärt rum och underrum)**.** Ett *linjärt rum* är en mängd V med tal (skalärer) K om:

(i) $\forall u,v \in V$ finns entydigt $u \oplus v \in V$

(ii) $\forall u \in V$, $\alpha \in K$ finns entydigt $\alpha \odot u \in V$

(iii) $u \oplus v = v \oplus u \quad \forall u,v \in V$ (kommutativ lag)

(iv) $(u \oplus v) \oplus w = u \oplus (v \oplus w) \quad \forall u,v,w \in V$ (associativ lag)

(v) det finns $0 \in V$ (nollelement) s.a. $0 \oplus u = u \oplus 0 = u \quad \forall u \in V$ (medför V icke-tom)

(vi) för varje $u \in V$ finns motsvarande $-u \in V$ (additiv invers) s.a. $u \oplus -u = 0$

(vii) $\alpha \odot (\beta \odot u) = (\alpha\beta) \odot u \; \forall u \in V, \quad \forall \alpha,\beta \in K$ (associativ lag)

(viii) $\alpha \odot (u \oplus v) = (\alpha \odot u) \oplus (\alpha \odot v) \quad \forall u,v \in V$, $\alpha \in K$ (distributiv lag 1)

(ix) $(\alpha + \beta) \odot u = (\alpha \odot u) \oplus (\beta \odot u) \quad \forall u \in V$, $\alpha,\beta \in K$ (distributiv lag 2)

(x) $\exists 1 \in K$ så att $1 \odot u = u \quad \forall u \in V$.

En icke-tom delmängd $M \subset V$ kallas ett underrum av det linjära rummet V om M är ett linjärt rum m.a.p. samma operationer.

En delmängd $M \subset V$ kallas affin om det finns en vektor $u_0 \in V$ och ett underrum $U \subset V$ så att $M = \{u_0 \oplus u : u \in U\}$.

Sats* (L1.4, lösningsmängden till lin. avb.)**.** *Låt U, V vara linjära rum och $F : U \to V$ vara en linjär avbildning. Antag $\boldsymbol{u}_p$ är en lösning till $F(\boldsymbol{u}) = \boldsymbol{v}$, där $\boldsymbol{v} \in V$ är känd men vi söker $\boldsymbol{u} \in U$. Då är $\boldsymbol{u}$ en lösning till ekv. $F(\boldsymbol{u}) = \boldsymbol{v}$ om och endast om $\boldsymbol{u} = \boldsymbol{u}_p + \boldsymbol{u}_h$ där $\boldsymbol{u}_h \in N(F)$ (nollrummet).*

Bevis. Då $F(\boldsymbol{u}_p) = \boldsymbol{v}$ gäller ekvivalenserna

$$F(\boldsymbol{u}) = \boldsymbol{v} = F(\boldsymbol{u}_p) \iff F(\boldsymbol{u} - \boldsymbol{u}_p) = 0 \iff (\boldsymbol{u} - \boldsymbol{u}_p) \in N(F)$$

Kalla $\boldsymbol{u} - \boldsymbol{u}_p$ för $\boldsymbol{u}_h$. □

Sats* (L1.5, bassatsen (del))**.** *Antag att $\dim V = n > 0$. Då finns minst en uppsättning av n st. linjärt oberoende vektorer i V. Varje sådan uppsättning är en bas för V.*

Bevis. Låt $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ godt. uppsättning med n linjärt oberoende vektorer (finns p.g.a definitionen av dimension). Låt $\boldsymbol{v} \in V$ godtyckligt. Om $v \notin \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ så är $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \boldsymbol{v}$ linjärt oberoende. Men max antal linjärt oberoende vektorer är n — motsägelse! Alltså godtyckligt $\boldsymbol{v} \in \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\} \implies V = \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\} \implies \boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ bas för V. □

Sats* (L1.9, bassatsen (del))**.** *Antag att $\dim V = n > 0$ och $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in V$ linjärt oberoende, med $m < n$. Då finns vektorer $\boldsymbol{u}_{m+1}, \ldots, \boldsymbol{u}_n$ s.a. $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ är en bas för V.*

Bevis. Låt $U_m = \operatorname{span}\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$ så $\dim U_m = m < n$, och U_m är därmed inte hela V. Då $\exists \boldsymbol{u}_{m+1} \in V$ s.a. $\boldsymbol{u}_{m+1} \notin U_m$. Lemma 1.2 ger att $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m, \boldsymbol{u}_{m+1}$ är linjärt oberoende. Vi upprepar detta tills vi har n vektorer, vilka då utgör en bas enligt sats 1.5. □

Anmärkning (L1.18). Ekivalenta påståenden om en $n \times n$-matris A:

(i) $A\boldsymbol{x} = \boldsymbol{0}$ har endast lösn. $\boldsymbol{x} = \boldsymbol{0}$ $(N(A) = \{\boldsymbol{0}\})$

(ii) $A\boldsymbol{x} = \boldsymbol{b}$ lösbar $\forall \boldsymbol{b} \in \mathbb{R}^n$ $(V(A) = \mathbb{R}^n)$

(iii) $A\boldsymbol{x} = \boldsymbol{b}$ entydigt lösbar $\forall \boldsymbol{b} \in \mathbb{R}^n$

(iv) rang $A = n$

(v) kolonnerna i A är linjärt oberoende

(vi) A är inverterbar

(vii) $\det A \neq 0$

Definition (Skalärprodukt)**.** En skalärprodukt (inre produkt) i ett reellt linjärt rum V är en reellvärd funktion $\langle u,v\rangle$ av $u,v \in V$ som uppfyller

(i) $\langle u,v\rangle = \langle v,u\rangle \quad \forall u,v \in V$

(ii) $\langle \alpha u,v\rangle = \alpha \langle u,v\rangle \quad \forall u,v \in V,\ \alpha \in \mathbb{R}$

(iii) $\langle u_1 + u_2,v\rangle = \langle u_1,v\rangle + \langle u_2,v\rangle \quad \forall u_1,u_2,v \in V$

(iv) $\langle u,u\rangle \geq 0 \quad \forall u \in V$ och $\langle u,u\rangle = 0 \iff u = 0$

Sats* (L2.1, Cauchy-Schwarz olikhet)**.**

$$|\langle u,v\rangle| \leq \|u\|\,\|v\| \quad \forall u,v \in V \qquad \textit{likhet omm } u,v \textit{ lin. beroende}$$

Bevis. Om $v = 0$ gäller olikheten och u,v lin. beroende trivialt. Om $v \neq 0$: Betrakta

$$0 \leq \|u - \alpha v\|^2 = \|u\|^2 + 2\langle u, -\alpha v\rangle + \|-\alpha v\|^2 = \|u\|^2 - 2\alpha\langle u,v\rangle + \alpha^2\|v\|^2$$

Låt nu $\alpha = \langle u,v\rangle / \|v\|^2$. Då får vi

$$0 \leq \|u\|^2 - 2\frac{\langle u,v\rangle^2}{\|v\|^2} + \frac{\langle u,v\rangle^2}{\|v\|^4}\|v\|^2 = \|u\|^2 - \frac{\langle u,v\rangle^2}{\|v\|^2}$$

Förläng med $\|v\|^2$:

$$0 \leq \|u\|^2\|v\|^2 - \langle u,v\rangle^2 \iff \langle u,v\rangle^2 \leq \|u\|^2\|v\|^2$$

Kvadratroten ur båda sidor ger olikheten i påst. □

Sats* (L2.2, triangelolikheten)**.**

$$\|u + v\| \leq \|u\| + \|v\| \quad \forall u,v \in V \qquad \textit{likhet omm } v = 0 \textit{ eller } u = \alpha v,\ \alpha \geq 0$$

Bevis.

$$\|u + v\|^2 = \|u\|^2 + 2\langle u,v\rangle + \|v\|^2 \leq \|u\|^2 + 2|\langle u,v\rangle| + \|v\|^2 \overset{C-S}{\leq}$$

$$\overset{C-S}{\leq} \|u\|^2 + 2\|u\|\,\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2$$

Likhet på båda ställen omm $v = 0$ eller $u = \alpha v$, $\alpha \geq 0$. Kvadratroten ur båda sidor ger olikheten i påst. □

Sats* (L2.11, fyra fundamentala underrummen)**.** *För godt. $m \times n$-matris A gäller*

$$N(A) = V(A^\mathsf{T})^\perp \qquad N(A)^\perp = V(A^\mathsf{T})$$
$$N(A^\mathsf{T}) = V(A)^\perp \qquad N(A^\mathsf{T})^\perp = V(A)$$

Bevis. Låt

$$A = \begin{bmatrix} — & r_1 & — \\ — & r_2 & — \\ & \vdots & \\ — & r_m & — \end{bmatrix} \implies A\boldsymbol{x} = \begin{bmatrix} — & r_1 \bullet \boldsymbol{x} & — \\ — & r_2 \bullet \boldsymbol{x} & — \\ & \vdots & \\ — & r_m \bullet \boldsymbol{x} & — \end{bmatrix}$$

$x \in N(A) \iff A\boldsymbol{x} = 0 \iff x$ ortogonal mot $r_i^\mathsf{T} \quad \forall i = 1, \dots n$
$\iff x$ ortogonal mot kolonner i A^T
$\iff x$ ortogonal mot $V(A^\mathsf{T})$ dvs. underrummet som spänns upp av kolonnerna
$\iff x \in V(A^\mathsf{T})^\perp \implies N(A) = V(A^\mathsf{T})^\perp \implies N(A)^\perp = (V(A^\mathsf{T})^\perp)^\perp = V(A^\mathsf{T})$
Applicera p.s.s. för A^T för övriga två likheter. □

Sats* (L4.6, oberoende av egenvektorer)**.** *Om $\lambda_1,\ldots,\lambda_k$ är olika egenvärden och $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_k$ motsvarande egenvektorer så är $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_k\}$ linjärt oberoende.*

Bevis. Motsägelsebevis. Antag $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_k\}$ linjärt *beroende*. Låt m vara max antal lin. ober. vektorer bland dessa. Efter ev. omnumrering har vi då att $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_m\}$ lin. ober. och enligt Sats 1.8

$$\boldsymbol{e}_k = \alpha_1\boldsymbol{e}_1 + \alpha_2\boldsymbol{e}_2 + \cdots + \alpha_m\boldsymbol{e}_m \qquad (\star)$$

Vi har dessutom att $A\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i\ \forall i$ ty de är egenvektorer. Betrakta

$$\begin{aligned} A(\star) &\implies \lambda_k\boldsymbol{e}_k = \alpha_1\lambda_1\boldsymbol{e}_1 + \alpha_2\lambda_2\boldsymbol{e}_2 + \cdots + \alpha_m\lambda_m\boldsymbol{e}_m \\ \lambda_k(\star) &\implies \lambda_k\boldsymbol{e}_k = \alpha_1\lambda_k\boldsymbol{e}_1 + \alpha_2\lambda_k\boldsymbol{e}_2 + \cdots + \alpha_m\lambda_k\boldsymbol{e}_m \end{aligned}$$

Subtrahera de två leden med varandra

$$\boldsymbol{0} = \alpha_1(\lambda_1-\lambda_k)\boldsymbol{e}_1 + \cdots + \alpha_m(\lambda_m-\lambda_k)\boldsymbol{e}_m$$

Då $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_m\}$ är lin. ober. måste $\alpha_i(\lambda_i-\lambda_k)=0\ \forall i=1,\ldots,m$, men egenvärdena är skilda så $\alpha_1=\cdots=\alpha_m=0$.

Detta innebär att $(\star)$ ger $\boldsymbol{e}_k=0$, men $\boldsymbol{e}_k$ var ju en egenvektor (per def. nollskild) $\implies$ motsägelse! □

Sats* (L4.10, symmetriska matriser har reella egenvärden)**.** *Låt A vara en reell symmetrisk $n\times n$-matris. Då är egenvärdena till A reella.*

Bevis. Låt λ vara egenvärde med egenvektor $\boldsymbol{x}$ (a priori kan detta vara komplext), dvs. det gäller

$$A\boldsymbol{x} = \lambda\boldsymbol{x}$$

Om vi tar komplex konjugat får vi i stället

$$\overline{A}\overline{\boldsymbol{x}} = \overline{\lambda}\overline{\boldsymbol{x}}$$

Eftersom A är reell är $\overline{A}=A$. Transponera nu ($A^{\mathsf{T}}=A$ ty A symmetrisk)

$$\text{VL: } (A\overline{\boldsymbol{x}})^{\mathsf{T}} = \overline{\boldsymbol{x}}^{\mathsf{T}}A^{\mathsf{T}} = \overline{\boldsymbol{x}}^{\mathsf{T}}A \quad \text{HL: } (\overline{\lambda}\overline{\boldsymbol{x}})^{\mathsf{T}} = \overline{\lambda}\overline{\boldsymbol{x}}^{\mathsf{T}} \quad \implies \overline{\boldsymbol{x}}^{\mathsf{T}}A = \overline{\lambda}\overline{\boldsymbol{x}}^{\mathsf{T}}$$

Låt oss multiplicera med $\boldsymbol{x}$ från höger på båda sidor (minns $A\boldsymbol{x}=\lambda\boldsymbol{x}$)

$$\overline{\boldsymbol{x}}^{\mathsf{T}}A\boldsymbol{x} = \lambda\overline{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{x} = \overline{\lambda}\overline{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{x}$$

Flytta över mellan de sista två leden

$$(\lambda-\overline{\lambda})\overline{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{x} = 0$$

Men $\overline{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{x}$ är större än noll om $\boldsymbol{x}$ är nollskild (vilket $\boldsymbol{x}$ är ty det är en egenvektor), så $\lambda=\overline{\lambda} \implies \lambda\in\mathbb{R}$. □

Definition (Felanalys och numerik)**.** Låt $\hat{x}$ och $\hat{f}$ approx. till x resp. $f(x)$.

absoluta felet $\delta x := \hat{x} - x$

relativa felet $\dfrac{\delta x}{x} = \dfrac{\hat{x} - x}{x} \approx \dfrac{\delta x}{\hat{x}}$

n **korrekta decimaler** om $|\delta x| \leq 0.5 \cdot 10^{-n}$

n **signifikanta siffror** om $|\delta x / x| < 0.5 \cdot 10^{-n}$

felfortplantning hur fortplantar sig fel δx i indata sig till fel δf i utdata

$$\delta f(x) = f'(x + \theta\delta x)\delta x \approx f'(\hat{x})\delta x \qquad \theta \in (0,1)$$

$$|\delta f(x)| \lesssim |f'(\hat{x})||\delta x| \qquad \left|\frac{\delta f(x)}{f(\hat{x})}\right| \lesssim \left|\frac{f'(\hat{x})}{f(\hat{x})}\right| |\delta x|$$

För högre ordningens approx. Taylorutv. man $f(\hat{x}) = f(x + \delta x)$ kring x.

felfortplantning (flera variabler) helt analogt:

$$\delta f(\boldsymbol{x}) \approx \nabla f(\hat{\boldsymbol{x}}) \bullet \delta \boldsymbol{x}$$

$$|\delta f(\boldsymbol{x})| \lesssim |\nabla f(\hat{\boldsymbol{x}}) \bullet \delta \boldsymbol{x}| \leq \|\nabla f(\hat{\boldsymbol{x}})\| \, \|\delta \boldsymbol{x}\| \qquad \left|\frac{\delta f(\boldsymbol{x})}{f(\hat{\boldsymbol{x}})}\right| \lesssim \frac{\|\nabla f(\hat{\boldsymbol{x}})\| \, \|\delta \boldsymbol{x}\|}{|f(\hat{\boldsymbol{x}})|}$$

konditionstal

$$\kappa(x) = \lim_{\delta \to 0^+} \max_{\|\delta x\| \leq \delta} \frac{|f(x + \delta x) - f(x)|/|f(x)|}{\|\delta x\| \,/\, \|x\|} \approx \frac{\|\text{rel. fel ut}\|}{\|\text{rel. fel in}\|}$$

$$\kappa(x) \lesssim \frac{\|\nabla f(x)\| \, \|\delta x\| \,/|f(x)|}{\|\delta x\| \,/\, \|x\|} = \|x\| \, \frac{\|\nabla f(x)\|}{|f(x)|}$$

framåtfel $\hat{f}(x) - f(x)$

bakåtfel $\hat{x} - x = f^{-1}(\hat{f}(x)) - x$

stabilitet (algoritmer) En numerisk algoritm $\hat{f} \approx f$ sägs vara stabil för $x \neq 0$ om rel. bakåtfelets storlek är liten:

$$\frac{|f^{-1}(\hat{f}(x)) - x|}{\|x\|}$$

1 Projektioner

Projektionen av $\boldsymbol{u}$ på ett rum som spänns av ON-basen $\boldsymbol{e}_1,\ldots,\boldsymbol{e}_n$ är

$$\boldsymbol{u}' = \langle \boldsymbol{u},\boldsymbol{e}_1\rangle\,\boldsymbol{e}_1 + \cdots + \langle \boldsymbol{u},\boldsymbol{e}_n\rangle\,\boldsymbol{e}_n$$

Den klassiska formeln kommer av fallet då basen $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ ej är normaliserad:

$$\boldsymbol{u}' = \frac{\langle \boldsymbol{u},\boldsymbol{v}_1\rangle}{\langle \boldsymbol{v}_1,\boldsymbol{v}_1\rangle}\boldsymbol{v}_1 + \cdots + \frac{\langle \boldsymbol{u},\boldsymbol{v}_n\rangle}{\langle \boldsymbol{v}_n,\boldsymbol{v}_n\rangle}\boldsymbol{v}_n$$

Projektionen är entydig så länge rummet är ändligdimensionellt, och $\|\boldsymbol{u}-\boldsymbol{u}'\|$ är det minsta avståndet från rummet till $\boldsymbol{u}$. Detta tillämpas för minstakvadratlösningar.

1.1 Gram-Schmidts ortogonaliseringsprocess, L2.4

Varje ändligdimensionellt inre produktrum V med $\dim V > 0$ har en ON-bas. Låt $n = \dim V$. Låt $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ vara en bas för V.

Låt $\boldsymbol{e}'_1 = \boldsymbol{v}_1$. Antag att vi konstruerat $\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_k$ parvis ortogonala, så att $\operatorname{span}\{\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_k\} = \operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k\}$. Låt

$$\boldsymbol{e}'_{k+1} = \boldsymbol{v}_{k+1} - \sum_{j=1}^{k} \frac{\langle \boldsymbol{v}_{k+1},\boldsymbol{e}'_j\rangle}{\langle \boldsymbol{e}'_j,\boldsymbol{e}'_j\rangle}\boldsymbol{e}'_j$$

Denna nya vektor är också parvis ortogonal med de övriga (visas ej här), och $\operatorname{span}\{\boldsymbol{e}'_1,\ldots,\boldsymbol{e}'_k,\boldsymbol{e}'_{k+1}\} = \operatorname{span}\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_k,\boldsymbol{v}_{k+1}\}$. Upprepa detta tills dess att vi har n baselement. Normera sedan för att få en ON-bas.

2 Numerisk ekvationslösning

2.1 Newtons metod

Antag att vi vill lösa ekvationen $f(x) = 0$ med rot $x = x^*$.

Newtons metod för numerisk ekvationslösning bygger på en linearisering av ekvationen kring den approximativa roten.

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

Samma princip kan appliceras analogt på flervariabelsfallet, men vi delar i stället upp det i ett linjärt ekvationssystem med en sökriktingsvariabel:

$$\begin{cases} \mathcal{J}(x_k)s_k = -f(x_k) \\ x_{k+1} = s_k + x_k \end{cases}$$

Vi visar kort konvergensordningen för det enkla fallet då vi har konvergens ($x_k \to x^*$, $k \to \infty$) mot en enkelrot $f(x^*) = 0$, $f'(x^*) \neq 0$. Taylorutv. kring x_k, där ξ_k mellan x_k,x^*:

$$0 = f(x^*) = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2}f''(\xi_k)(x^* - x_k)^2 \qquad (\star)$$

Betrakta en iteration

$$x_{k+1} - x_k + \frac{f(x_k)}{f'(x_k)} = 0 \iff 0 = (x_{k+1} - x_k)f'(x_k) + f(x_k) \qquad (\star\star)$$

$$(\star) - (\star\star) \implies 0 = f'(x_k)(x^* - x_{k+1}) + \frac{1}{2}f''(\xi_k)(x^* - x_k)^2$$

Men f' är nollskild för tillräckligt stora k ty $f'(x^*) \neq 0$ och vi har konvergens. Således, pga. kontinuitet gäller

$$|x_{k+1} - x^*| = \frac{|f''(\xi_k)||x^* - x_k|^2}{2|f'(x_k)|} \implies \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} = \frac{|f''(\xi_k)|}{2|f'(x_k)|} \xrightarrow[k\to\infty]{} \frac{|f''(x^*)|}{2|f'(x^*)|}$$

Vi har lokal kvadratisk konvergens mot enkelrötter, om metoden konvergerar. Mot multipelrötter med mult. m får vi linjär konvergens. (Vi kan anpassa den andra termen i envariabelsfallet med en faktor m för kvadratisk konvergens.)

2.2 Sekantmetoden

Saknar vi uttryck för f' kan vi inte tillämpa Newtons metod. Vi kan använda approximationen

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$$

$$\implies x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}, \quad k = 1,2,\ldots$$

I gengäld krävs två startapproximationer, samt att konvergensordningen blir det något sämre $q = 1{,}618$ (superlinjär men inte kvadratisk).

2.3 Fixpunktsiterationer

Skriv om $f(x) = 0$ på formen $x = g(x)$ och iterera enligt $x_{k+1} = g(x_k)$. Om metoden konvergerar gäller då, pga. kontinuitet, att $x^* = g(x^*)$ dvs. roten är en fixpunkt till g. Konvergenshastigheten avgörs av $|g'(x^*)|$, som minst måste vara < 1. Förfarandet är helt analogt i flervariabelsfallet, fast vi studerar i stället Jacobianen.

Vi kan göra en metodoberoende feluppskattning genom en Taylorutveckling.

$$|\hat{x} - x^*| \approx \frac{f(\hat{x})}{f'(\hat{x})}$$

3 Polynominterpolation

En funktion f approximeras av ett polynom $\hat{f} = p_n$ så att de stämmer överens i vissa givna punkter. Det finns ett entydigt interpolationspolynom $p_n(x)$ av $\deg p_n \leq n$ som går genom (distinkta) punkterna $(x_i, f(x_i))$, $i = 0,\ldots,n$. På Newtons form:

$$p_n(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \cdots + c_n(x - x_0)(x - x_1)\cdots(x - x_{n-1})$$

där villkoren $p_n(x_i) = f(x_i)$ bestämmer c_i.

Trunkeringsfelet från interpolationen fås genom argument med Rolles sats

$$R_T = p_n(x) - f(x) = \frac{-f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n)$$

där ξ ligger mellan $x_0,\ldots,x_n,x$. Notera att felet är 0 i interpolationspunkterna. Om vi bara känner till funktionsvärden kan vi begränsa felet med nästa term (den första försummade termen) genom att försöka approximera $f^{(n+1)}(x)$ med $p_{n+1}^{(n+1)}(x)$.

$$|R_T(x)| \leq |c_{n+1}(x-x_0)(x-x_1)\cdots(x-x_n)|$$

Till sist anmärker vi att interpolation med högt gradtal kan ge mycket stora fel mellan interpolationspunkterna, i synnerhet om punkterna är jämnt utspridda. Detta kallas *Runges fenomen*. Ett alternativ för att undvika detta är s.k. splines.

3.1 Splines

En spline $s(x)$ av grad k är ett styckvis polynom av grad k med kontinuerliga derivator av ordn. $k-1$ över interpolationspunkterna. Den bestäms av villkoren (för n punkter), där $i = 1,\ldots,n-1$:

$$\begin{cases} s_1(x_0) = f(x_0) \\ s_i(x_i) = s_{i+1}(x_i) = f(x_i) \quad s_i'(x_i) = s_{i+1}'(x_i) \\ s_n(x_n) = f(x_n) \end{cases}$$

I allmänhet räcker detta inte för att entydigt bestämma s, utan det krävs också någon form av randvillkor. Det finns olika typer beroende på vår kunskap om funktionen, t.ex. *rätta randvillkor* $s'(x_0) = f'(x_0)$, $s'(x_n) = f'(x_n)$ och *naturliga randvillkor* $s''(x_0) = s''(x_n) = 0$.

4 Numerisk integration — kvadraturformler

4.1 Trapetsregeln

Vi approximerar integranden med en linjär funktion över hela intervallet.

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\left(f(a)+f(b)\right) - \underbrace{\frac{f''(\xi)}{12}(b-a)^3}_{R_T} \quad \xi \in (a,b)$$

Bättre approximation fås genom att betrakta n delintervall med längd $h = (b-a)/n$ och $x_i = a + ih$, $i = 0,\ldots,n$

$$T(h) = h\left(\frac{f(x_0)}{2} + \sum_{i=1}^{n-1} f(x_i) + \frac{f(x_n)}{2}\right) \qquad R_T = \frac{b-a}{12}h^2 f''(\xi) \quad \xi \in (a,b)$$

4.2 Simpsons regel

Vi approximerar nu integranden med ett andragradspolynom, interpolerat från jämnt fördelade punkter.

$$\int_a^b f(x)\,dx = \frac{b-a}{6}\left(f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right) - \underbrace{\frac{f^{(4)}(\xi)}{2880}(b-a)^5}_{R_T} \quad \xi \in (a,b)$$

På samma sätt kan vi i stället tillämpa regeln på delintervall (n jämnt)

$$S(h) = \frac{h}{3}\left(f(x_0)+4\sum_{i=1}^{n/2} f(x_{2i-1})+2\sum_{i=1}^{n/2-1} f(x_{2i})+f(x_n)\right)$$

$$R_T = \frac{b-a}{180}h^4 f^{(4)}(\xi) \quad \xi \in (a,b)$$

Önskar vi ännu bättre approx. kan vi använda Richardsonextrapolation, som eliminerar några feltermer av lägre ordning

$$T^{(2)}(h) = T(h) + \frac{T(h)-T(2h)}{3}$$

$$S^{(2)}(h) = S(h) + \frac{S(h)-S(2h)}{15}$$

5 Faktoriseringar

5.1 LU-faktorisering

LU-faktorisering ses enklast som ekvivalent med Gausselimination. Vi delar upp en $m \times n$-matris $A = LU$ där L $m \times m$ nedåt triangulär och U $m \times n$ uppåt triangulär (i blockavseende). Av numeriska skäl kan vi även införa pivotering enligt $PA = LU$ så att vi sorterar raderna med störst element (till beloppet) överst.

Gausseliminera A som vanligt så erhålls U. Spara information i L om operationerna som genomfördes — i mitt tycke är det enklast att tänka sig detta om resultatet om samma operationer fast med omvänt tecken utfördes på enhetsmatrisen.

För att lösa $Ax = b$ löser man då $Ly = Pb$ (framåtsubstitution) följt av $Ux = y$ (bakåtsubstitution). (Utan pivotering är $P = I$).

Vi har följande feluppskattningar:

$$\frac{\|\delta x\|}{\|x\|} \lesssim \kappa(A)\frac{\|\delta b\|}{\|b\|} \qquad \frac{\|\delta x\|}{\|x\|} \lesssim \kappa(A)\frac{\|\delta A\|}{\|A\|}$$

5.2 QR-faktorisering

QR-faktorisering ses enklast som ekvivalent med Gram-Schmidt processen. Vi delar upp en $m \times n$-matris $A = QR$ där Q $m \times m$ ortogonal och R $m \times n$ uppåt triangulär (i blockavseende).

Bestäm en ON-bas till $V(A)$ och för in denna i Q i kolonnernas ordning. Det är möjligt att spara information i R under processen men det enklaste är att utnyttja att Q är ortogonal och lösa $R = Q^\mathsf{T} A$ (i det reella fallet skall anmärkas). Observera att detta inte är tillvägagångssättet som används på "riktigt" med datorstöd, utan man använder då Householderreflektioner ty de är mer stabila.

För att lösa $Ax = b$ löser man då $Rx = Q^\mathsf{T} b$ (bakåtsubstitution). Detta ger minstakvadratlösningen om problemet vore överbestämt (man kan behöva anpassa nollrader o.dyl.).

5.3 SVD (singular value decomposition)

Singulärvärdesuppdelning är en slags generaliserad diagonalisering. Vi delar upp en $m \times n$-matris $A = U\Sigma V^\mathsf{T}$ där U $m \times m$ ortogonal, Σ $m \times n$ "diagonal" sorterad i fallande storleksordning till beloppet och V $n \times n$ ortogonal. (Med "diagonal" menas att beroende på rangen av A kan diagonalen i Σ åtföljas av ett antal nollrader under.)

Bestämma detta är komplicerat och uppdelningen är inte entydig. De singulära värdena på "diagonalen" i Σ är roten ur egenvärdena till $A^\mathsf{T} A$ (de är alla positiva eftersom den produkten är positivt semidefinit). På samma sätt inser man att kolonnerna i V och U utgörs av normerade egenvektorer till $A^\mathsf{T} A$ respektive AA^T.

Med SVD kan man införa Moore-Penrose pseudoinvers $A^+ = V_1 \Sigma_1^{-1} U_1^\mathsf{T}$. Notera att $x = A^+ b$ löser minstakvadratproblemet. Av numeriska skäl kan man minska felet och approximera lösningen genom att ersätta mindre singulära värden med nollor, eftersom det visar sig att $\kappa(A) = \sigma_{max}/\sigma_{min}$. Ju större minimum på de singulära värdena desto mindre konditionstal.

6 Numerisk beräkning av derivator och lösning av ODE

Det finns tre enkla fall av differenskvoter, nämligen framåt-, bakåt- respektive centraldifferens:

$$\frac{f(x+h) - f(x)}{h} \qquad \frac{f(x) - f(x-h)}{h} \qquad \frac{f(x+h) - f(x-h)}{2h}$$

Vi nöjer oss med att notera att centraldifferensen ger trunkeringsfel med ledande term h^2 medan de övriga ger trunkeringsfel med ledande term h. För litet h bidrar emellertid till kancellation i täljaren.

6.1 Begynnelsevärdesproblem

Vi önskar lösa

$$\begin{cases} y'(t) = f(t, y(t)) & a \le t \le b \\ y(a) = c \end{cases}$$

där f är given ocn a,b,c konstanter. Det är möjligt att y är vektorvärd. Skulle vi ha ett högre ordningens system kan vi då alltid skriva om det på denna formen.

Vår strategi för att lösa detta med s.k. differensmetoder är att approximera genom att stega fram mellan diskreta tidpunkter s.a. $a = t_0 < t_1 < \cdots < t_N = b$

och bestämma $y_k \approx y(t_k)$. För enkelhets skull låter vi $t_k + 1 = t + h$ för något fixt h.

Eulers framåtmetod	$y_{k+1} = y_k + hf(t_k,y_k)$
Eulers bakåtmetod	$y_{k+1} = y_k + hf(t_{k+1},y_{k+1})$
Mittpunktsmetoden	$y_{k+1} = y_{k-1} + 2hf(t_k,y_k)$
Trapetsmetoden	$y_{k+1} = y_k + \frac{h}{2}\left(f(t_k,y_k) + f(t_{k+1},y_{k+1})\right)$