

# Sammanfattning av SF1544 Numeriska metoder, grundkurs

Yashar Honarmandi  
yasharh@kth.se

12 mars 2019

## **Sammanfattning**

Detta är en sammanfattning av SF1544 Numeriska metoder, grundkurs.

## Innehåll

1	Användbar matte	1
2	Grundläggande koncept för numeriska metoder	5
3	Lösning av ekvationer	7
4	Interpolation	12
5	Optimering	13
6	Derivator	13
7	Integration	15
8	Lösning av ordinarie differentialekvationer	19
9	Partiella differentialekvationer	22
10	Stabilitet	22
11	Egenvärdesberäkningar	22

# 1 Användbar matte

**Hjälpsats från Grönwalls lemma** Låt  $f(x, y)$  vara Lipschitzkontinuerlig i  $y$  med Lipschitzkonstant  $L$ , och  $Y$  och  $Z$  lösa  $\frac{dy}{dx} = f$  med något initialvillkor i  $x = a$ . Då är

$$|Y(t) - Z(t)| \leq e^{L(t-a)}|Y(a) - Z(a)|.$$

**Bevis** Bilda  $u = Y - Z$ . Vi antar att  $u(a) \neq 0$  eftersom entydighet annars skulle ge satsen trivialt. Vi vet även att  $u$  är strikt positiv eller negativ, så vi kan anta att  $u$  är strikt positiv. Eftersom  $f$  är Lipschitzkontinuerlig, uppfyller  $u$  då

$$\frac{du}{dx} = |f(t, Y) - f(t, Z)| \leq L|Y - Z| = Lu,$$

och lösning av detta med integration från  $a$  till  $t$  fullför beviset.

**Allmän begränsning av globalt fel** Betrakta begynnelsevärdesproblemet

$$\begin{aligned}\frac{dy}{dt}(t) &= f(t, y(t)), \\ y(a) &= b,\end{aligned}$$

löst på  $[a, T]$ , där  $f$  är Lipschitzkontinuerlig med Lipschitzkonstant  $L$ . Betrakta en numerisk lösning med lokalt fel begränsad av  $Mh^{p+1}$ . Då begränsas det globala felet av

$$|y(T) - y_N| \leq \frac{e^{L(T-a)}M}{L}h^p.$$

**Bevis** Vi inför  $y(t; t_n)$  som den exakta lösningen som startar i  $(t_n, y_n)$  (alltså är den exakta lösningen  $y(t; t_0)$ ). Det globala felet ges då av

$$\begin{aligned}|y(T) - y_N| &= |y(T) - y(T; t_{N-1}) + y(T; t_{N-1}) + \cdots - y(T; t_1) + y(T; t_1) - y_N| \\ &\leq |y(T) - y(T; t_{N-1})| + |y(T; t_{N-1}) - y(T; t_{N-2})| + \cdots + |y(T; t_1) - y_N|.\end{aligned}$$

Vi noterar att  $y(T) = y(T; t_{N-1})$ .

Den första termen ges simpelthen av det lokala felet. Hjälpsatsen ovan ger

$$|y(T; t_i) - y(T; t_{i-1})| \leq e^{L(T-t_i)}|y(t_i; t_i) - y(t_i; t_{i-1})|.$$

Det som står kvar i absolutbeloppstecknet är det lokala felet, eftersom den vänstra termen är exakt och den högra kommer från en iteration. Detta ger

$$|y(T; t_i) - y(T; t_{i-1})| \leq e^{L(T-t_i)}Mh^{p+1} = e^{L(N-i)h}Mh^{p+1}$$

och vidare

$$\begin{aligned}
|y_N - y(T)| &\leq Mh^{p+1} + Mh^{p+1}e^{Lh} + \dots + Mh^{p+1}e^{L(N-1)h}. \\
&= Mh^{p+1} \frac{1 - e^{LNh}}{1 - e^{Lh}} \\
&= Mh^{p+1} \frac{e^{LNh} - 1}{e^{Lh} - 1} \\
&\leq Mh^{p+1} \frac{e^{LNh}}{Lh} \\
&= \frac{e^{LNh}M}{L} h^p,
\end{aligned}$$

och beviset är klart.

**Möjlighet för polynominterpolation** Givet  $n + 1$  punkter  $(x_i, y_i)$ , där  $x_i \neq x_j$  om  $i \neq j$ , finns det ett entydigt polynom  $p$  av grad (högst)  $n$  så att  $p(x_i) = y_i \ \forall i = 1, \dots, n + 1$ .

**Bevis** För att visa existensen av  $p$ , kan vi välja det som

$$p(x) = \sum_{i=1}^{n+1} y_i \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Denna metoden kallas för Lagrangeinterpolation. Vi ser att  $p$  har grad  $n$  och

$$p(x_k) = \sum_{i=1}^{n+1} y_i \prod_{j \neq i} \frac{x_k - x_j}{x_i - x_j}.$$

För alla termer i summan där  $i \neq k$  kommer det finnas en faktor  $x_k - x_k$  i nämnaren, och dessa ger inget bidrag. För  $i = k$  blir produkten bara 1, och  $p(x_k) = y_k$ .

För att visa att  $p$  är entydig, antag att  $q$  är ett annat interpolationspolynom och bilda  $v = q - p$ , med grad högst  $n$ . Detta ger att  $v$  har  $n + 1$  nollställen. Detta är endast möjligt om  $v = 0$ , och därmed måste  $p$  vara unikt.

**Fel för linjär interpolation** Antag att  $f \in C^2$ . Låt  $p$  vara det linjära polynomet som interpolerar punkterna  $(x_0, f(x_0))$  och  $(x_0 + h, f(x_0 + h))$ . Då gäller  $|f(x) - p(x)| \leq Ch^2$ ,  $x_0 \leq x \leq x_0 + h$ , där

$$C = \max_{x_0 \leq z \leq x_0 + h} \frac{\left| \frac{d^2 f}{dx^2}(z) \right|}{8}.$$

**Bevis** Antag  $x_0 = 0$ , och bilda  $g = f - p$ .  $g$  är kontinuerlig på  $[0, h]$ , och antar därmed ett största och minsta värde på detta intervallet. Antag att den antar sitt största värde i  $x = a$ . Taylorutveckling kring  $a$  ger

$$g(x) = g(a) + \frac{dg}{dx}(a)(x-a) + \frac{1}{2} \frac{d^2g}{dx^2}(c)(x-a)^2$$

för något  $c \in [a, x]$ . Vi ser att andra termen måste bli 0, ty  $a$  är en maxpunkt. Evaluering av Taylorutvecklingen i  $x = 0$  eller  $x = h$  (den som är närmast  $a$ ) ger  $g(x) = 0$  och

$$g(a) = -\frac{1}{2} \frac{d^2g}{dx^2}(c)(x-a)^2,$$

$$|g(a)| \leq \frac{1}{2} \max_{x_0 \leq z \leq x_0+h} \left| \frac{d^2g}{dx^2}(z) \right| \left( \frac{1}{2}h \right)^2.$$

Vi vet att  $\frac{d^2f}{dx^2}(x) = \frac{d^2g}{dx^2}(x)$  eftersom  $p$  är linjär, vilket ger

$$|g(a)| \leq \max_{x_0 \leq z \leq x_0+h} \frac{\left| \frac{d^2f}{dx^2}(z) \right|}{8} h^2.$$

**Konvergens för potensmetoden** Antag för en given matris  $A$  att dens egenvärden uppfyller  $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$  och att dens egenvektorer  $v_1, \dots, v_n$  spänner upp  $R^n$ . Då konvergerar potensmetoden för nästan alla startvektorer  $x^0$  till egenvektorn  $v_1$  och egenvärdet  $\lambda_1$ , med maximal konvergensrate  $\left| \frac{\lambda_2}{\lambda_1} \right|$ .

**Bevis** Vi kan skriva

$$x^0 = \sum_{i=1}^n c_i v_i.$$

Om  $c_1 \neq 0$  ger  $k$  multiplikationer med  $A$

$$A^k x^0 = \sum_{i=1}^n \lambda_i^k c_i v_i.$$

Potensmetoden ger oss nu

$$x^{k+1} = \frac{1}{\left\| \sum_{i=1}^n \lambda_i^k c_i v_i \right\|} \sum_{i=1}^n \lambda_i^k c_i v_i.$$

Om vi antar att både  $c_1$  och  $\lambda_1$  är positiva, kan detta skrivas som

$$x^{k+1} = \frac{1}{\left\| \sum_{i=1}^n \left( \frac{\lambda_i}{\lambda_1} \right)^k \frac{c_i}{c_1} v_i \right\|} \sum_{i=1}^n \left( \frac{\lambda_i}{\lambda_1} \right)^k \frac{c_i}{c_1} v_i.$$

Om  $\lambda_1$  är negativ, kan man dividera med dens belopp. Om  $c_1$  är negativ, får du fortfarande konvergens mot en egenvektor motsvarande egenvärdet  $\lambda_1$ .

Enligt antagandet kommer iterationen för stora  $k$  att konvergera mot  $\frac{v_1}{\|v_1\|}$ . Vi noterar även att enligt antagandet är det termen motsvarande  $i = 2$  som konvergerar långsammast, och att konvergensen sker med en faktor  $\left| \frac{\lambda_2}{\lambda_1} \right|$ .

**Eigenvärden för tridiagonala matriser** Betrakta  $n \times n$ -matrisen  $T$  med 1 på diagonalen och  $-1$  på över- och underdiagonalen. Dens egenvektorer  $v_i$  uppfyller

$$v_{i,j} = \sin\left(\frac{ij}{n+1}\pi\right)$$

och har motsvarande egenvärden

$$\lambda_i = 1 - 2 \cos\left(\frac{i}{n+1}\pi\right).$$

**Bevis** Låt  $x_i = \frac{i}{n+1}\pi$

$$(Tv_i)_j = -\sin((j-1)x_i) + \sin(jx_i) - \sin((j+1)x_i).$$

Trigonometriska identiteter ger

$$\sin((j \pm 1)x_i) = \sin(jx_i) \cos(x_i) \pm \cos(jx_i) \sin(x_i),$$

vilket implicerar

$$(Tv_i)_j = (1 - 2 \cos(x_i)) \sin(jx_i).$$

**Stabilitet för numerisk lösning av värmeledningsekvationen** Betrakta värmeledningsekvationen

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}$$

löst numerisk med en framåt differens i tid, med steglängd  $h$  i  $x$  och  $k$  i  $t$ . Lösningen är stabil om

$$\frac{Dk}{h^2} < \frac{1}{2}.$$

**Bevis** Att lösa detta numeriskt innebär att göra iterationen

$$y_{i+1} = Ay_i + s_j,$$

där  $y_j$  är en vektor med värdet av lösningen i de olika punkterna i diskretiseringen vid tidssteg  $j$ . Man kan ställa upp ekvationerna och få att  $A$  är en diagonalmatris med  $1 - 2\sigma$  på diagonalen,  $\sigma$  på över- och underdiagonalen och 0 överallt annars, där  $\sigma = \frac{Dk}{h^2}$ . Man får även att  $s_j$  är en vektor vars alla element är 0 förutom de två sista, som har värdet  $\sigma y_{0,j}$  och  $\sigma y_{N,j}$ , alltså värdena på randen.

Låt nu  $y_j$  vara en exakt lösning som uppfyller  $y_{j+1} = Ay_j + s_j$  och  $w_j$  vara den numeriska lösningen som erhålls. Skillnaden  $e_j$  mellan dessa uppfyller  $e_{j+1} = Ae_j$ . Om inte felen skall förstöras, måste alla egenvärden ha belopp mindre än 1.

Låt nu  $T$  vara  $n \times n$ -matrisen med 1 på diagonalen och  $-1$  på över- och underdiagonalen. Då kan vi skriva matrisen  $A$  som  $A = (1 - \sigma)I - \sigma T$ . Vi känner egenvärdena till  $T$ , och vet därmed att  $A$  har egenvärden

$$\begin{aligned}\lambda_i &= 1 - \sigma - \sigma \left( 1 - 2 \cos \left( \frac{i}{n+1} \pi \right) \right) \\ &= 2\sigma \left( \cos \left( \frac{i}{n+1} \pi \right) - 1 \right) + 1.\end{aligned}$$

Eftersom  $j$  går från 1 till  $n$ , ligger alla egenvärden mellan  $1 - 4\sigma$  och 1. Vi vet att  $\sigma$  är positiv, så vi kräver  $1 - 4\sigma > -1$ . Om man sätter in uttrycket för  $\sigma$ , fås slutligen

$$\frac{Dk}{h^2} < \frac{1}{2},$$

och beviset är klart.

## 2 Grundläggande koncept för numeriska metoder

**Lokal konvergens** En numerisk metod säjs vara lokalt konvergent om den konvergerar mot ett givet numeriskt värde för startgissningar som är tillräckligt nära det rätta värdet.

**Linjär konvergens** En numerisk metod vars felterm uppfyller

$$\lim_{n \rightarrow \infty} \frac{e_{n+1}}{e_n} = S$$

är linjärt konvergent med hastighet  $S$ .

**Kvadratisk konvergens** En numerisk metod vars feltermer uppfyller

$$\lim_{n \rightarrow \infty} \frac{e_{n+1}}{e_n^2} = M$$

är kvadratisk konvergent.

**Noggrannhetsordning** Om  $u_h$  är en approximation av en storhet  $u$ , där  $h$  är positionsparametern, och det finns tall  $p, C$  och  $h_0$  sådana att

$$|u_h - u| \leq Ch^p, h \leq h_0,$$

sägs approximationen ha noggrannhetsordning  $p$ .

**Absolut och relativt fel** Låt  $\tilde{x}$  vara ett approximativt värde av storheten  $x$ . Då definieras det absoluta felet som

$$e_x = \tilde{x} - x$$

och det relativa felet som

$$r_x = \frac{e_x}{x}.$$

**Felgränser** Felgränserna definieras som

$$|e_x| \leq E_x, |r_x| \leq R_x.$$

**Felförplantning** Låt  $y = f(x)$  och  $\tilde{y} = f(\tilde{x})$  vara en approximation av  $y$ . Detta ger

$$\begin{aligned} e_y &= \tilde{y} - y \\ &= f(\tilde{x}) - f(x) \\ &\approx f(\tilde{x}) - \left( f(\tilde{x}) + \frac{df}{dx}(\tilde{x})(x - \tilde{x}) \right) \\ &= -\frac{df}{dx}(\tilde{x})(x - \tilde{x}) \\ &= \frac{df}{dx}(\tilde{x})e_x, \end{aligned}$$

och felgränsen ges av

$$E_y \approx \left| \frac{df}{dx}(\tilde{x}) \right| E_x.$$

Den motsvarande relationen i högre dimensioner är

$$E_y \approx \sum \left| \frac{\partial f}{\partial x_i}(\tilde{\mathbf{x}}) \right| E_{x_i}.$$



Om vi nu utgår från ett approximativt värde  $\tilde{y} = f(\tilde{\mathbf{x}})$  och approximerar de partiella derivatorna, till exempel genom mätningar, med

$$\frac{\partial f}{\partial x_i}(\tilde{\mathbf{x}}) = \frac{f(\tilde{\mathbf{x}} + E_{x_i} \mathbf{e}_{x_i}) - f(\tilde{\mathbf{x}})}{E_{x_i}},$$

kan man definiera

$$\tilde{y}_i = f(\tilde{\mathbf{x}} + E_{x_i} \mathbf{e}_{x_i})$$

och få

$$E_y \approx \sum |\tilde{y}_i - \tilde{y}|.$$

**Kancellation** Om man på en dator subtraherar två nästan lika stora tal, förlorar man precision. Det suger.

**Utskiftning** Om man på en dator adderar två tal som är mycket olika stora, kan datorn ignorera information från den minre siffran.

**Konditionstal** Konditionstalet definieras som

$$\kappa = \lim_{\delta \rightarrow 0} \max_{|e_x| \leq \delta} \frac{\left| \frac{e_y}{y} \right|}{\left| \frac{e_x}{x} \right|}.$$

### 3 Lösning av ekvationer

**Fixpunktsmetoden** Betrakta ekvationen

$$x = g(x).$$

Fixpunktsmetoden är en enkel iterationsmetod för att lösa denna ekvationen, med den enkla iterationsformeln

$$x_{n+1} = g(x_n).$$

En pseudokod-beskrivning av en lösning med startvillkor  $\mathbf{x}_0$  där det itereras tills lösningen stämmer med en tolerans  $\mathbf{t}$  är:

```
define g(x)
input x0
input t
while abs(x - g(x)) > t
    x = g(x)
end
return x
```

**Konvergens** Om  $g \in C^1$ ,  $\left|\frac{dg}{dx}(\alpha)\right| < 1$  och  $\alpha$  är en fixpunkt. finns det en omgivning till  $\alpha$  så att om  $x_0$  är i denna omgivningen, går  $x_n \rightarrow \alpha$ . Metoden konvergerar linjärt med reduktionsfaktor  $S = \left|\frac{dg}{dx}(\alpha)\right|$ .

För att visa detta, skriver vi

$$x_{n+1} - \alpha = g(x_n) - g(\alpha) = \frac{dg}{dx}(c)(x_n - \alpha),$$

där vi har använt medelvärdesatsen och det faktum att  $\alpha$  är en fixpunkt. Vidare, eftersom  $g \in C^1$  finns det en omgivning till  $\alpha$  så att  $\left|\frac{dg}{dx}(x)\right| \leq \frac{S+1}{2}$ , dvs. ett tal som är mindre än 1, men större än  $S$ . Om  $x_0$  är i denna omgivningen, är

$$e_{n+1} \leq \frac{S+1}{2} e_n.$$

Detta implicerar att  $x_n \rightarrow \alpha$  och att

$$\frac{e_{n+1}}{e_n} \rightarrow S.$$

**Fixpunktsmetoden för system** Betrakta ekvationssystemet

$$\mathbf{x} = \mathbf{g}(\mathbf{x}).$$

Vi löser även detta med fixpunktsmetoden, och använder iterationsformeln

$$\mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n).$$

En pseudokod-beskrivning av en lösning med startvillkor  $\mathbf{x}_0$  där det itereras tills lösningen stämmer med en tolerans  $\mathbf{t}$  är:

```
define g(x)
input x0
input t
while abs(x - g(x)) > t
    x = g(x)
end
return x
```

där allt nu är listor. Toleransen ser kanske lite annorlunda ut, vafan vet jag.

**Intervallhalveringsalgoritmen** Betrakta ekvationen

$$f(x) = 0.$$

Intervallhalveringsalgoritmen utgår från två punkter  $a, b$  så att  $f(a)f(b) < 0$ , och gör följande:

1. Beräkna funktionsvärdet i punkten  $m = \frac{b+a}{2}$ .
2. Om  $f(a)f(m) < 0$ , sätt  $a = m$ . Annars, sätt  $b = m$ .
3. Sluta iterationen när intervallbredden  $\frac{b-a}{2}$  är mindre än den givna toleransen.
4. Returnera  $\frac{b+a}{2}$ .

En pseudokod-beskrivning av en lösning med startvillkor **a** och **b** där det itereras tills lösningen stämmer med en tolerans **t** är:

```

define f(x)
input a, b
input t
while (b - a)/2 > t
    m = (a + b)/2
    if f(a)f(m) < 0
        a = m
    else
        b = m
    end
end
return (a + b)/2

```

**Newton-Rhapsonsmetoden** Betrakta ekvationen

$$f(x) = 0.$$

Newton-Rhapsons metod utgår från tangenten till  $f$ . Om man startar i  $x_0$ , har tangenten ekvation  $t(x) = \frac{df}{dx}(x_0)(x - x_0) + f(x_0)$ . Dens nollställe ges av

$$x = x_0 - \frac{f(x_0)}{\frac{df}{dx}(x_0)}.$$

Från detta gör vi iterationen

$$x_{n+1} = x_n - \frac{f(x_n)}{\frac{df}{dx}(x_n)}$$

som avslutas när  $|x_{n+1} - x_n|$  är mindre än någon tolerans. Vi ser att detta är en variant av fixpunktsmetoden med  $g(x) = x - \frac{f(x)}{\frac{df}{dx}(x)}$ .

En pseudokod-beskrivning av en lösning med startvillkor **x\_0** där det itereras tills lösningen stämmer med en tolerans **t** är:

```

define f(x)
define fderiv(x)
input x_0
input t
while f(x_0) > t
    x = x - f(x)/fderiv(x)
end
return x

```

**Konvergens** Om  $f \in C^2$  och  $\frac{df}{dx}(\alpha) \neq 0$ , är Newton-Rhasons metod kvadratisk konvergent med konstant  $M = \frac{\frac{d^2f}{dx^2}(\alpha)}{2\frac{df}{dx}(\alpha)}$ .

För att visa detta, notera först att om  $\alpha$  är ett nollställe till  $f$ , är det även ett nollställe till  $\frac{dg}{dx}$ , ty för  $f \in C^2$  och  $\frac{df}{dx}(\alpha) \neq 0$  gäller att

$$\frac{dg}{dx}(\alpha) = 1 - \frac{\frac{df}{dx}(\alpha)}{\frac{df}{dx}(\alpha)} + \frac{f(\alpha)\frac{d^2f}{dx^2}(\alpha)}{\frac{df}{dx}(\alpha)^2} = 0.$$

Därmed följer lokal konvergens av beviset som gjordes för fixpunktsmetoden. Vi Taylorutvecklar vidare  $f$  nära  $x_n$  och får

$$f(\alpha) = f(x_n) + \frac{df}{dx}(x_n)(\alpha - x_n) + \frac{1}{2}\frac{d^2f}{dx^2}(c_n)(\alpha - x_n)^2,$$

$$\frac{f(\alpha)}{\frac{df}{dx}(x_n)} = \frac{f(x_n)}{\frac{df}{dx}(x_n)} + \alpha - x_n + \frac{1}{2}\frac{\frac{d^2f}{dx^2}(c_n)}{\frac{df}{dx}(x_n)}(\alpha - x_n)^2 = 0,$$

under antagandet  $\frac{df}{dx}(x_n) \neq 0$ . Eftersom  $\alpha$  är ett nollställe till  $f$ , skriver vi om detta till

$$x_n - \frac{f(x_n)}{\frac{df}{dx}(x_n)} - \alpha = \frac{1}{2}\frac{\frac{d^2f}{dx^2}(c_n)}{\frac{df}{dx}(x_n)}(\alpha - x_n)^2.$$

Vi känner igen de två första termerna på vänstersidan som  $x_{n+1}$ , och detta implicerar därmed

$$\frac{e_{n+1}}{e_n^2} = \left| \frac{\frac{d^2f}{dx^2}(c_n)}{2\frac{df}{dx}(x_n)} \right| \rightarrow M,$$

och beviset är klart.

**Newton-Rhasons metod för system** Betrakta ekvationen

$$\mathbf{f}(\mathbf{x}) = 0.$$

Metoden är den samma för ett enda system. Den utgår från den linjära approximationen

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + d\mathbf{f}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

och ger iterationen

$$\mathbf{x}_{n+1} = \mathbf{x}_n - d\mathbf{f}(\mathbf{x}_n)^{-1}f(\mathbf{x}_n).$$

Notera att beräkningsmässigt är det svårt att hitta en inversmatris, så det är smartare att hitta en vektor  $\mathbf{h}$  så att  $d\mathbf{f}(x_n)\mathbf{h} = \mathbf{f}(x_n)$  och använda den i stället.

**Linjära system** Lösning av ett  $n \times n$  linjärt system  $Ax = b$  görs med ett antal beräkningar som är  $O(n^p)$  för något  $p$ . För Gausselimination är  $p = 3$ .

**Störningsanalys** Antag att man försöker lösa ett system  $Ax = b$ . Givet approximativ indata  $\tilde{b}$  med en given felgräns, dvs. en given gräns för  $|b - \tilde{b}|$ , hur stor felgräns fås för den approximativa lösningen  $\tilde{x}$ ?

Man kan skriva

$$|x - \tilde{x}| = A^{-1}|b - \tilde{b}|,$$

och normen av detta blir

$$\|x - \tilde{x}\| \leq \|A^{-1}\| \|b - \tilde{b}\|,$$

där vi har infört matrisnormen

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

som uppfyller

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

Uttrycket ovan innehåller absoluta fel, men ofta vet vi bara relativa fel på formen  $\frac{\|b - \tilde{b}\|}{\|b\|}$ . Med hjälp av matrisnormen kan vi dock skriva

$$\|b\| \leq \|A\| \|x\|,$$

vilket kan kombineras med uttrycket ovan för att ge

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \|A^{-1}\| \|A\| \frac{\|b - \tilde{b}\|}{\|b\|}.$$

Vi kan då definiera konditionstalet  $\kappa = \|A^{-1}\| \|A\|$ , vilket ger

$$R_x \leq \kappa R_b.$$

## 4 Interpolation

Det fundamentala interpolationsproblemet går ut på att hitta en kurva som bäst möjligt passar med vissa datapunkter. Kom i håg satsen om möjlighet för polynominterpolation.

**Polynominterpolation - första försök** Vi gör först en naiv ansats

$$p(x) = \sum_{i=0}^n c_i x^i$$

och anpassar konstanterna så att  $p$  antar rätt värden i datapunkterna. Detta ger oss ett linjärt system

$$X\mathbf{c} = \mathbf{y},$$

där  $\mathbf{c}$  är en vektor med alla koefficienter,  $\mathbf{y}$  är en vektor med alla  $y$ -värden och  $X_{ij} = x_j^{i-1}$ .  $X$  kallas för en Vandermonde-matris. Om man har många datapunkter, kan detta dock ge upphov till ett illakonditionerat system.

**Lagrangeinterpolation** Som vi så innan är ett möjligt interpolationspolynom

$$p(x) = \sum_{i=1}^{n+1} y_i \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

**Newtons interpolationsmetod** Vi gör en ny ansats

$$\begin{aligned} p(x) &= d_0 + d_1(x - x_0) + d_2(x - x_0)(x - x_1) + \dots \\ &= \sum_{i=0}^n d_i \prod_{j < i} (x - x_j). \end{aligned}$$

Vi ser att  $p(x_0) = y_0$ , och man kan från detta få de nästa koefficienterna.

**Minsta kvadratmetoden** Minsta kvadratmetoden är en metod för approximation av överbestämda ekvationssystem, dvs. system med fler ekvationer än obekanta. Sådana system har inget interpolationspolynom.

Sådana ekvationssystem kan formuleras som  $A\mathbf{x} = \mathbf{b}$ . Minsta kvadratlösningen är den lösningen som minimerar  $\|A\mathbf{x} - \mathbf{b}\|$ . Mer specifikt, om vi söker en funktion  $f(x) = \sum_{i=1}^n c_i \phi_i(x)$ , där funktionerna  $\phi_i$  kan väljas som vi vill, är  $A_{ij} = \phi_j(x_i)$ ,  $\mathbf{x}_i = c_i$  och  $\mathbf{b}_i = y_i$ .

Minstakvadratlösningen till detta systemet löser normalekvationerna

$$A^T A \mathbf{x} = A^T \mathbf{b}.$$

Om kolumnerna i  $A$  är linjärt oberoende, har detta en lösning.

## 5 Optimering

Denna delen kommer handla om att försöka hitta minima till en funktion över ett rum. Att hitta maxpunkter är det samma som att byta tecken och hitta minimum, så vi fokuserar därför på att hitta minima.

**Gyllene snitt-sökning** Antag att  $f$  är kontinuerlig och har endast ett lokalt minimum på  $[a, b]$ , och betrakta två punkter  $x_1$  och  $x_2$ . Om  $f(x_1) \leq f(x_2)$  finns minimumspunkten på  $[a, x_2]$ , och om  $f(x_1) \geq f(x_2)$  finns minimumspunkten på  $[x_1, b]$ . Detta gäller eftersom derivatan endast byter tecken en gång på  $[a, b]$ , och vi med hjälp av funktionsvärdena i två punkter får information om derivatan imellan dessa punkterna. Man kan få en ganska säker metod vid att välja både  $x_1$  och  $x_2$  nära mitten av intervallet.

Metoden kan förbättras vid att återanvända funktionsvärden i kommande iterationer. Sätt  $a = 0, b = 1$ . Vi väljer  $x_1$  och  $x_2$  symmetriskt och likformigt, så att  $x_2 = 1 - x_1$  och

$$\frac{x_1}{x_2} = \frac{x_2}{1}.$$

Vi kan lösa detta och få

$$x_1 = \frac{3 \pm \sqrt{5}}{2}, \quad x_2 = \frac{\sqrt{5} - 1}{2}.$$

Vi ser att  $x_2$  är det gyllene snitt. Med detta val gäller att intervalllängden avtar med en faktor  $g$  per iteration.

**Newtons metod** Newtons metod för att hitta minima till en funktion  $f(\mathbf{x})$  är att använda Newtons metod för att lösa ekvationssystemet  $\vec{\nabla} f(\mathbf{x}) = \mathbf{0}$ .

**Gradientmetoden** Vi vet att  $f$  avtar snabbast i riktningen  $-\vec{\nabla} f(\mathbf{x})$ . Vi gör därför iterationen

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \vec{\nabla} f(\mathbf{x}_n).$$

$\gamma_n$  kan väljas konstant eller så att den minimerar  $f(\mathbf{x}_{n+1})$ . Jämförd med Newtons

## 6 Derivator

**Framåtdifferens** En derivata kan approximeras med framåtdifferensen

$$\frac{df}{dx}(x) \approx \frac{f(x+h) - f(x)}{h}.$$

**Fel** För att få en skattning av felet, använder vi Taylorutveckling för att få

$$f(x+h) - f(x) \approx \frac{df}{dx}(x)h + \frac{1}{2} \frac{d^2f}{dx^2}(c)h^2,$$

och

$$\frac{f(x+h) - f(x)}{h} - \frac{df}{dx}(x) = \frac{1}{2} \frac{d^2f}{dx^2}(c)h.$$

**Val av steglängd** Antag att vi approximerar  $f$  med  $\tilde{f}$ , där  $\tilde{f}$  uppfyller  $|\tilde{f}(x) - f(x)| \leq \varepsilon \forall x$ . Felet vi gör i skattning av derivatan ges då av

$$\begin{aligned} \left| \frac{df}{dx}(x) - \frac{\tilde{f}(x+h) - \tilde{f}(x)}{h} \right| &\leq \left| \frac{df}{dx}(x) - \frac{f(x+h) - f(x)}{h} \right| \\ &\quad + \left| \frac{\tilde{f}(x+h) - f(x+h) - \tilde{f}(x) + f(x)}{h} \right| \\ &\leq Ch + \frac{2\varepsilon}{h}. \end{aligned}$$

Detta felet har sitt minimum för  $h = \sqrt{\frac{2\varepsilon}{C}}$ .

**Central differens** Man kan alternativt försöka med en central differens

$$\frac{df}{dx}(x) \approx \frac{f(x+h) - f(x-h)}{2h}.$$

**Fel** Vi Taylorutvecklar igen och får

$$\begin{aligned} f(x+h) &\approx f(x) + \frac{df}{dx}(x)h + \frac{1}{2} \frac{d^2f}{dx^2}(x)h^2 + \frac{1}{6} \frac{d^3f}{dx^3}(c_1)h^3, \\ f(x-h) &\approx f(x) - \frac{df}{dx}(x)h + \frac{1}{2} \frac{d^2f}{dx^2}(x)h^2 - \frac{1}{6} \frac{d^3f}{dx^3}(c_2)h^3. \end{aligned}$$

Då blir centrala differensen

$$\begin{aligned} \frac{f(x+h) - f(x-h)}{2h} - \frac{df}{dx}(x) &\approx \frac{1}{12} h^2 \left( \frac{d^3f}{dx^3}(c_1) + \frac{d^3f}{dx^3}(c_2) \right) \\ &\approx \frac{1}{6} \frac{d^3f}{dx^3}(c)h^2, \end{aligned}$$

där vi i sista raden har använt medelvärdesatsen.

**Val av steglängd** Antag att vi approximerar  $f$  med  $\tilde{f}$ , där  $\tilde{f}$  uppfyller  $|\tilde{f}(x) - f(x)| \leq \varepsilon \forall x$ . Felet vi gör i skattning av derivatan minimeras då för  $h \propto \sqrt[3]{\varepsilon}$ .



## 7 Integration

**Eulers metod** Antag att vi vill integrera en funktion  $f$  över  $[a, b]$ . En enkel metod är att dela upp  $[a, b]$  i  $N$  steg, låta  $x_0 = a, x_N = b$ . Eulers metod ger

$$\begin{aligned} y_N &= y_{N-1} + hf(x_{N-1}) \\ &= y_{N-2} + hf(x_{N-2}) + hf(x_{N-1}) \\ &\vdots \\ &= y_0 + h \sum_{i=0}^{N-1} f(x_i). \end{aligned}$$

Vi får slutligen approximationen

$$\int_a^b f(x) dx \approx h \sum_{i=0}^{N-1} f(x_i).$$

**Fel** Utgå från differentialekvationen

$$\frac{dy}{dx} = f, y(a) = y_0,$$

och antag att vi vill integrera  $f$  över  $[a, b]$ . Analysens huvudsats ger

$$\int_a^b f(x) dx = y(b) - y(a).$$

Att beräkna denna integralen är alltså helt ekvivalent med Eulers metod. Vi försöker därför först att hitta det lokala felet på  $[x_i, x_{i+1}]$ , som enligt argumentet från Eulers metod ges av

$$|y_{i+1} - y(x_{i+1})| \leq \frac{1}{2} \left| \frac{d^2 y}{dt^2}(c) \right| h^2,$$

där  $c \in [x_i, x_{i+1}]$ . Enligt sats ges det globala felet då av

$$|y_N - y(x_N)| \leq \frac{e^{L(a-b)}}{2L} \max_{c \in [a, b]} \left| \frac{d^2 y}{dt^2}(c) \right| h,$$

där vi har ändrat derivatan för att vara säker på att den anger en övre begränsning av det lokala felet. Detta kan skrivas om till

$$|y_N - y(x_N)| \leq \frac{e^{L(a-b)}}{2L} \max_{c \in [a, b]} \left| \frac{df}{dt}(c) \right| h.$$

**Trapetsmetoden** Betrakta integralen av en funktion  $f$  över  $[a, b]$ . Dela nu upp  $[a, b]$  i  $N$  lika långa delintervall, och låt  $x_0 = a, x_N = b$ . I trapetsmetoden interpolerar vi  $f$  i  $x_i$  och  $x_{i+1}$  och integrerar denna rätta linjen. Integralen är lika med arean av en trapets, vilket vi enkelt kan beräkna. En approximation av integralen över  $[x_i, x_{i+1}]$  är därmed

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

där  $h = \frac{b-a}{N}$ . En approximation av hela integralen är då

$$\begin{aligned} \int_a^b f(x) \, dx &\approx h \sum_{i=0}^{N-1} \frac{f(x_{i+1}) + f(x_i)}{2} \\ &= h \left( \frac{f(x_0) + f(x_N)}{2} + \sum_{i=1}^{N-1} f(x_i) \right). \end{aligned}$$

**Fel** Låt  $T_h$  vara skattningsfelet av integralen av  $C^2$ -funktionen  $f$  med trapetsmetoden med steglängd  $h$ . Felet ges då av

$$\left| \int_a^b f(x) \, dx - T_h \right| \leq \max_{x \in [a, b]} \frac{1}{12} \left| \frac{d^2 f}{dx^2}(x) \right| (b-a) h^2.$$

För att visa detta, försöker vi först hitta det lokala felet och sedan skatta det globala felet. Vi söker alltså först det lokala felet

$$e_i = \left| \frac{h}{2} (f(x_{i+1}) - f(x_i)) - \int_{x_i}^{x_{i+1}} f(x) \, dx \right|.$$

Partiell integration ger

$$e_i = \left| \frac{h}{2} (f(x_{i+1}) - f(x_i)) - x_{i+1} f(x_{i+1}) + x_i f(x_i) + \int_{x_i}^{x_{i+1}} x f'(x) \, dx \right|.$$

Vi inför  $c = \frac{x_{i+1} + x_i}{2}$  och får

$$e_i = \left| \frac{h}{2} (f(x_{i+1}) - f(x_i)) - (x_{i+1} - c) f(x_{i+1}) + (x_i - c) f(x_i) + \int_{x_i}^{x_{i+1}} (x - c) f'(x) \, dx \right|.$$

Eftersom  $c$  ligger mitt i mellan  $x_i$  och  $x_{i+1}$ , som ligger ett avstånd  $h$  från varandra, kan detta skrivas om till

$$e_i = \left| \int_{x_i}^{x_{i+1}} (x - c) f'(x) \, dx \right|.$$

En ny delvis integration ger

$$\begin{aligned} e_i &= \left| \frac{1}{8} h^2 (f'(x_{i+1}) - f'(x_i)) - \frac{1}{2} \int_{x_i}^{x_{i+1}} (x - c)^2 f''(x) \, dx \right| \\ &= \left| \frac{1}{2} \int_{x_i}^{x_{i+1}} \left( \left( \frac{h}{2} \right)^2 - (x - c)^2 \right) f''(x) \, dx \right|. \end{aligned}$$

Om  $|f''(x)| \leq M_i$  på intervallet, fås

$$\begin{aligned} e_i &\leq \frac{1}{2} M_i \left| \int_{x_i}^{x_{i+1}} \left( \left( \frac{h}{2} \right)^2 - (x - c)^2 \right) \, dx \right| \\ &= \frac{1}{2} M_i \left| \frac{h^3}{4} - \frac{2}{3} \left( \frac{h}{2} \right)^3 \right| \\ &= \frac{1}{12} M_i h^3. \end{aligned}$$

För att skatta det globala felet  $E$ , skriver vi

$$E \leq \sum_{i=0}^{N-1} e_i$$

där integrationsintervallet har delats i  $N + 1$  steg. Om  $|f''(x)| \leq M$  på hela integrationsintervallet, kan detta skrivas som

$$E \leq MN \frac{1}{12} h^3.$$

Sambandet  $h = \frac{b-a}{N}$  gör att detta kan skrivas som

$$E \leq \frac{1}{12} M (b - a) h^2.$$

**Simpsons metod** Använder kvadratisk interpolation.

**Fel** Låt  $S_h$  vara skattningsfelet av integralen med Simpsons metod med steglängd  $h$ . Felet ges då av

$$\left| \int_a^b f(x) \, dx - S_h \right| \leq \max_{x \in [a, b]} \frac{1}{80} \left| \frac{d^4 f}{dx^4}(x) \right| (b - a) h^4.$$

**Monte Carlo-integration** Medelvärdesatsen för integraler ger att

$$\frac{1}{b-a} \int_a^b f(x) \, dx = f(c),$$

där  $f(c)$  är funktionens medelvärde på  $[a, b]$ . Vi kan skatta medelvärdet som

$$\frac{1}{N} \sum_{i=1}^N f(x_i),$$

där alla  $x_i$  är dragna oberoende från en likformig fördelning på  $[a, b]$ . Detta kallas för Monte Carlo-integration.

Metoden är helt analog i högre dimensioner.

**Fel** Om vi antar att alla  $x_i$  dras från en likformig sannolikhetstäthet, kan vi skriva

$$\frac{1}{b-a} \int_a^b f(x) \, dx = \int_a^b f(x)p(x) \, dx,$$

där  $p$  är sannolikhetstätheten. Vänstersidan kan även skrivas som  $E(f(X))$ .

Vi inför den stokastiska variabeln

$$\varepsilon_n = \frac{1}{N} \sum_{i=1}^N f(X_i) - E(f(X)).$$

Dens väntevärde ges av

$$\begin{aligned} E(\varepsilon_n) &= \frac{1}{N} \sum_{i=1}^N E(f(X_i)) - E(f(X)) \\ &= E(f(X)) - E(f(X)) \\ &= 0. \end{aligned}$$

Vi har vidare

$$\begin{aligned}
E(\varepsilon_n^2) &= E\left(\left(\frac{1}{N} \sum_{i=1}^N f(X_i) - E(f(X))\right)^2\right) \\
&= \frac{1}{N^2} E\left(\left(\sum_{i=1}^N (f(X_i) - E(f(X)))\right)^2\right) \\
&= \frac{1}{N^2} E\left(\left(\sum_{i=1}^N (f(X_i) - E(f(X)))\right)^2\right) \\
&= \frac{1}{N^2} E\left(\left(\sum_{i=1}^N (f(X_i) - E(f(X)))\right)^2\right) \\
&= \frac{1}{N^2} E\left(\sum_{i=1}^N (f(X_i) - E(f(X)))^2 + \sum_{n \neq m} (f(X_n) - E(f(X)))(f(X_m) - E(f(X)))\right) \\
&= \frac{1}{N^2} \sum_{i=1}^N E((f(X_i) - E(f(X)))^2) + E\left(\sum_{n \neq m} (f(X_n) - E(f(X)))(f(X_m) - E(f(X)))\right) \\
&= \frac{\sigma_f^2}{N} + 0.
\end{aligned}$$

Korstermerna har väntevärde

$$\begin{aligned}
E((f(X_n) - E(f(X)))(f(X_m) - E(f(X)))) &= E(f(X_n) - E(f(X))) E(f(X_m) - E(f(X))) \\
&= 0,
\end{aligned}$$

då de olika  $X_i$  är oberoende. Därmed är standardavvikelsen av  $\varepsilon_N$  lika med  $\frac{\sigma}{\sqrt{N}}$ , och felet i metoden är proportionellt mot  $\frac{1}{\sqrt{N}}$ .

Vi noterar att centrala gränsvärdesatsen ger att  $\nu = \frac{\sqrt{N}}{\sigma} \varepsilon_N$  är standard normalfördelat.

I  $d$  dimensioner är felet proportionellt mot  $N^{-\frac{d}{2}}$ , och vi ser att den är bättre än trapetsmetoden för  $d > 4$ .

## 8 Lösning av ordinarie differentialekvationer

**Eulers metod (framåt)** Betrakta begynnelsevärdesproblemet

$$\begin{aligned}
\frac{dy}{dt}(t) &= f(t, y(t)), \\
y(a) &= b.
\end{aligned}$$

Eulers metod går ut på att

1. Dela upp intervallet vi vill lösa problemet på i diskreta steg, där steg  $n$  finns i punkten  $t_n = a + nh$ , där  $h$  är steglängden.
2. Linjarisera problemet till

$$y_{n+1} - y_n = f(t_n, y_n)h,$$

där  $y_n = y(t_n)$ .

En pseudokod-beskrivning av en lösning med startvillkor  $t_0$  och  $y_0$ , steglängd  $h$  och  $N$  steg är:

```

define f(t, y)
input t0 and y0
input h and N
t = t0
y = y0
for 1 < i < N
    y = y + f(t, y)*h
    t = t + h
end

```

**Felanalys** Om vi betraktar det första steget i iterationen, har man lokalt

$$\begin{aligned}
 y(t_1) &= y(t_0) + \frac{dy}{dt}(t_0)(t_1 - t_0) + \frac{1}{2} \frac{d^2y}{dt^2}(\alpha)(t_1 - t_0)^2 \\
 &= y(t_0) + hf(t_0, y_0) + \frac{1}{2} \frac{d^2y}{dt^2}(\alpha)h^2.
 \end{aligned}$$

Om andraderivatan av  $y$  är begränsad, ger detta

$$|y_1 - y(t_1)| \leq Mh^2,$$

och det lokala felet är  $O(h^2)$ .

Det globala felet kan nu uppskattas som det lokala felet multiplicerat med antal steg. Om vi försöker lösa ekvationen på intervallet  $[a, T]$  med  $N$  steg, har man

$$Nh = T - a,$$

och det globala felet kan uppskattas som

$$|y_N - y(t_N)| \approx h^2 \frac{T - a}{h} = Ch.$$

**Eulers metod för system av differentialekvationer** Betrakta begynnelsevärdesproblemet

$$\begin{aligned}\frac{d\mathbf{y}}{dt}(t) &= \mathbf{f}(t, \mathbf{y}(t)), \\ \mathbf{y}(a) &= \mathbf{b}.\end{aligned}$$

Eulers metod går ut på att

1. Dela upp intervallet vi vill lösa problemet på i diskreta steg, där steg  $n$  finns i punkten  $t^n = a + nh$ , där  $h$  är steglängden.
2. Linjarisera problemet till

$$y^{n+1} - y_n = \mathbf{f}(t^n, y^n)h,$$

där  $\mathbf{y}^n = \mathbf{y}(t^n)$ .

En pseudokod-beskrivning av en lösning med startvillkor  $\mathbf{t0}$  och  $\mathbf{y0}$ , där denna är en lista med  $M$  element, steglängd  $h$  och  $N$  steg är:

```
define f(t, y)
input t0 and y0
input h and N
t = t0
y = y0
for 1 < i < N
    for 1 < j < M
        y[j] = y[j] + f[j](t, y)*h
    end
    t = t + h
end
```

Observera att  $\mathbf{f}$  nu är en lista av  $M$  funktioner, och kom ihåg att högre ordningens ekvationer med en funktion kan skrivas som ett system av differentialekvationer.

**Eulers metod bakåt** Betrakta begynnelsevärdesproblemet

$$\begin{aligned}\frac{dy}{dt}(t) &= f(t, y(t)), \\ y(a) &= b.\end{aligned}$$

Eulers metod bakåt går ut på att

1. Dela upp intervallet vi vill lösa problemet på i diskreta steg, där steg  $n$  finns i punkten  $t_n = a + nh$ , där  $h$  är steglängden.

2. Linjarisera problemet till

$$y_{n+1} - y_n = f(t_{n+1}, y_{n+1})h,$$

där  $y_n = y(t_n)$ . Detta ger en ekvation i  $y_{n+1}$  som måste lösas numeriskt.

## 9 Partiella differentialekvationer

**Finita differensmetoden** I finita differensmetoden approximerar vi derivator med finita differenser. Detta är helt analogt med vad som har gjorts för ordinarie differentialekvationer.

## 10 Stabilitet

**Absolutstabilitet** För allmänna numeriska problem säjs en numerisk metod vara absolutstabil om effekten av små initiala störningar försvinner när antal iterationer blir stort.

## 11 Egenvärdesberäkningar

**Potensmetoden** Potensmetoden är ett sätt att hitta det största egenvärdet till en matris på. Algoritmen går ut på att välja en startvektor  $x^0$  och iterera enligt

$$x^{i+1} = \frac{1}{\|Ax^i\|} Ax^i, \lambda^{i+1} = (x^{i+1})^T Ax^{i+1}.$$

Detta kan alternativt skrivas som

$$v^{i+1} = Ax^i, \lambda^i = (x^i)^T v^{i+1}, x^{i+1} = \frac{1}{\|v^{i+1}\|} v^{i+1}.$$