ETH Zürich D-MATH, D-PHYS 04.08.2008

Numerische Methoden

Prof. D. Kressner

Wichtige Hinweise:

- Bitte deutlich schreiben und alle Ergebnisse begründen. Ergebnisse ohne Begründung werden nicht gewertet.
- Wenn nicht anders vermerkt, sollen Rechnungen auf 4 Dezimalstellen genau ausgeführt werden.
- Die Argumente von trigonometrischen Funktionen sind in Radian angegeben.
- Alle Aufgaben werden gleich gewichtet.

1. [**Lineare Gleichungssysteme**] Gegeben ist die Matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0.05 & 3 \\ 2 & 0.1001 & 1 \\ 1 & 4 & 1 \end{pmatrix}$$

a) Berechnen Sie die **LR**-Zerlegung von **A** in $\mathbb{F}(10, 4, -10, 10)$ (*ohne* Pivotsuche).

b) Berechnen Sie die Lösung $\underline{x}$ des linearen Gleichungssystems $\mathbf{A}\underline{x} = \underline{b}$ durch Vor- und Rückwärtseinsetzen (in $\mathbb{F}(10, 4, -10, 10)$), wobei

$$\underline{b} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^\top$$

Berechnen Sie ebenfalls die 2-Norm des Residuums $\underline{r}$.

c) Berechnen Sie das Residuum $\underline{r}_d$ in *doppelter* Genauigkeit: Nehmen Sie dazu den Vektor $\underline{x}$ aus b) und berechnen Sie $\underline{b} - \mathbf{A}\underline{x}$ in $\mathbb{F}(10, 2 \cdot 4, -10, 10)$.

d) Führen Sie, ausgehend von $\underline{r}_d$ einen Schritt der Nachiteration ebenfalls in doppelter Genauigkeit durch. Benutzen Sie dazu die berechneten **L** und **R** aus Teil a).

Lösung

a) Mit Algorithmus 2.10 berechnet man

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 39500 & 1 \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & 0.05 & 3 \\ 0 & 0.0001 & -5 \\ 0 & 0 & 197500 \end{pmatrix}.$$

b) Wir lösen zuerst das System $\mathbf{L}\underline{y} = \underline{b}$ und dann $\mathbf{R}\underline{x} = \underline{y}$ und erhalten

$$\underline{y} = \begin{pmatrix} 1 \\ -1 \\ 39500 \end{pmatrix} \quad \underline{x} = \begin{pmatrix} 0.4 \\ 0 \\ 0.2 \end{pmatrix} \Longrightarrow \underline{r} = \underline{b} - \mathbf{A}\underline{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 0.6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0.4 \end{pmatrix}$$

Somit ist $\|\underline{r}\|_2 = 0.4$.

c) Es ändert sich nichts in 8-stelliger Genauigkeit, da bei der Matrixmultiplikation $\mathbf{A}\underline{x}$ und -Subtraktion nicht gerundet werden musste. Das heisst also

$$\underline{r}_d = \begin{pmatrix} 0 \\ 0 \\ 0.4 \end{pmatrix} \qquad \Longrightarrow \|\underline{r}_d\|_2 = 0.4.$$

d) Wir müssen das System $\mathbf{A}\underline{x}_{up} = \underline{r}_d$ lösen und die verbesserte Lösung ist dann

$$\hat{\underline{x}} = \underline{x} + \underline{x}_{up}.$$

Die **LR**-Zerlegung wurde schon in a) berechnet, also brauchen wir nur noch Vorwärts- und Rückwärtseinzusetzen:

$$\mathbf{L}\underline{y}_{up} = \underline{r}_d \Longrightarrow \underline{y}_{up} = \begin{pmatrix} 0 \\ 0 \\ 0.4 \end{pmatrix}$$

und somit

$$\underline{x}_{up} \approx \begin{pmatrix} -0.50693669 \cdot 10^{-2} \\ 0.101265820 \\ 0.20253164 \cdot 10^{-5} \end{pmatrix}$$

Die verbesserte Lösung ist somit

$$\hat{\underline{x}} = \underline{x} + \underline{x}_{up} = \begin{pmatrix} 0.4 \\ 0 \\ 0.2 \end{pmatrix} + \begin{pmatrix} -0.50693669 \cdot 10^{-2} \\ 0.101265820 \\ 0.20253164 \cdot 10^{-5} \end{pmatrix} \approx \begin{pmatrix} 0.39493063 \\ 0.101265820 \\ 0.20000202 \end{pmatrix}$$

2. **[Interpolation]**

a) Gesucht ist eine Polynominterpolation $p_4(x)$ 4. Grades zur Funktion $f(x) = \sin(x)$, welche an den Stützstellen $x_i = i \cdot \frac{\pi}{4}$, $i = 0, \ldots, 4$, dieselben Werte wie $f(x)$ hat. Geben Sie die Newton-Darstellung von $p_4(x)$ an.

b) Berechnen Sie den exakten Fehler $|p_4(x) - f(x)|$ bei $x = 1$ und vergleichen Sie diesen mit der in der Vorlesung angegebenen Fehlerabschätzung.

c) Welche Art von Spline (vollständig, periodisch, oder natürlich) bietet sich am besten an, um dieselbe Funktion in den selben Stützstellen zu interpolieren. Stellen Sie das entsprechende Gleichungssystem auf (ohne es zu lösen).
Wie kann man ein solches System effizient lösen, insbesondere bei einer grossen Anzahl von Stützstellen? Begründen Sie.

d) Geben Sie eine Abschätzung für den Fehler bei der Spline-Interpolation an.

Lösung:

a) Wir bestimmen die Koeffizienten des Newtonschen Interpolationspolynoms

$$p(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + f[x_0, x_1, x_2, x_3]...$$
$$\cdot (x - x_0)(x - x_1)(x - x_2) + f[x_0, x_1, x_2, x_3, x_4](x - x_0)(x - x_1)(x - x_2)(x - x_3)$$

mit dem Dividierte-Differenzen-Schema.

x_i	$f[x_i]$	$f[x_i, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$,	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	$f[x_0, \ldots, x_4]$
0	0				
$\frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{2\sqrt{2}}{\pi}$			
$\frac{\pi}{2}$	1	$\frac{2(2-\sqrt{2})}{\pi}$	$\frac{8}{\pi^2}(1 - \sqrt{2})$		
$\frac{3\pi}{4}$	$\frac{\sqrt{2}}{2}$	$\frac{2(\sqrt{2}-2)}{\pi}$	$\frac{8}{\pi^2}(\sqrt{2} - 2)$	$\frac{32}{3\pi^3}(2\sqrt{2} - 3)$	
π	0	$\frac{-2\sqrt{2}}{\pi}$	$\frac{8}{\pi^2}(1 - \sqrt{2})$	$\frac{32}{3\pi^3}(3 - 2\sqrt{2})$	$\frac{32}{3\pi^4}(6 - 4\sqrt{2})$

Damit erhalten wir

$$p(x) = \frac{2\sqrt{2}}{\pi}x + \frac{8}{\pi^2}(1 - \sqrt{2})x\left(x - \frac{\pi}{4}\right) + \frac{32}{3\pi^3}(2\sqrt{2} - 3)x\left(x - \frac{\pi}{4}\right)\left(x - \frac{\pi}{2}\right)$$
$$+ \frac{32}{3\pi^4}(6 - 4\sqrt{2})x\left(x - \frac{\pi}{4}\right)\left(x - \frac{\pi}{2}\right)\left(x - \frac{3\pi}{4}\right)$$

b) Der exakte Fehler ist

$$|p(1) - \sin(1)| \approx |0.841736 - 0.841471| \approx 0.00026537.$$

Die Fehlerdarstellung aus dem Skript sieht folgendermassen aus

$$E_n[f](x_\star) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x_\star),$$

wobei $\xi \in [x_0, x_n]$ eine (unbekannte) Zwischenstelle ist.
Der Betrag $|E_n[f](x_\star)|$ kann jetzt folgendermassen abgeschätzt werden:

$$|E_n[f](x_\star)| \leq \frac{\|f^{(n+1)}(\xi)\|_\infty}{(n+1)!}|\omega_{n+1}(x_\star)|.$$

Also haben wir mit $x_\star = 1$:

$$|E_n[\sin](1)| \leq \frac{1}{(n+1)!}\underbrace{|\omega_{n+1}(1)|}_{\approx 0.3558} \approx 0.2965 \cdot 10^{-2}.$$

c) Es kommen nur der vollständige (man kennt die Ableitungen von f am linken und rechten Endpunkt) oder der natürliche (die 2. Ableitungen von f sind links und rechts 0) Spline, nicht aber der periodische, da wir hier nur die halbe Periode des Sinus betrachten und die Ableitungen bei $x = 0$ und $x = \pi$ nicht übereinstimmen.

Das Gleichungssystem für den vollständigen kubischen Spline entnehmen wir der Formel (3.40) auf p.89 des Skripts. Wir wissen, dass $c_0 = f'(x_0) = 1$, $c_n = f'(x_n) =$

-1, also kann man die Koeffizienten der Matrix und der rechten Seite $\underline{r}$ berechnen:

$$\begin{aligned}
\alpha_j &= \frac{2}{h_j} + \frac{2}{h_{j+1}} = 2 \cdot \frac{2}{\pi/4} = \frac{16}{\pi},\ j = 1, \dots 3 \\
\beta_j &= \frac{1}{h_j} = \frac{4}{\pi},\ j = 2, 3 \\
d_j &= 3\left(\frac{y_j - y_{j-1}}{h_j^2} + \frac{y_{j+1} - y_j}{h_{j+1}^2}\right) = \frac{3 \cdot 16}{\pi^2}(y_{j+1} - y_{j-1}) \\
d_1 &= \frac{3 \cdot 16}{\pi^2}, \quad d_2 = 0, \quad d_3 = -\frac{3 \cdot 16}{\pi^2} \\
\underline{r} &= \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix} \qquad r_1 = d_1 - \frac{4}{\pi},\ r_2 = d_2,\ r_3 = d_3 + \frac{4}{\pi}
\end{aligned}$$

Die Koeffizientenmatrix $\mathbf{A}$ ist sicher symmetrisch. Dass sie auch positiv definit ist, sieht man wie folgt:

$$\begin{aligned}
\underline{x}^\top \mathbf{A}\underline{x} = \frac{16}{\pi}(x_1^2 + x_2^2 + x_3^2) + \frac{8}{\pi}(x_1x_2 + x_2x_3) = \underbrace{(\frac{16}{\pi} - \frac{8}{\pi})}_{>0}(x_1^2 + x_2^2 + x_3^2) \\
+ \frac{4}{\pi}(x_1^2 + x_3^2 + (x_1 + x_2)^2 + (x_2 + x_3)^2) > 0
\end{aligned}$$

Bei einer SPD-Matrix bietet sich das CG-Verfahren (die Matrix ist tridiagonal, also kann die Matrix-Vektor Multiplikation schnell ausgeführt werden) oder das SOR-Verfahren an.

d) Der Fehler für den vollständigen Spline kann mithilfe von Satz 3.27 abgeschätzt werden. Damit gilt

$$\|f - s\|_\infty \le \frac{5}{384} h^4 \|f^{(4)}\|_\infty = \frac{5}{384}\left(\frac{\pi}{4}\right)^4 \cdot 1 \approx 0.4954 \cdot 10^{-2}$$

3. **[Numerische Integration]**

a) Approximieren Sie das Integral

$$\int_{-1}^{2} x \exp(x^2) dx = \frac{1}{2} e(e^3 - 1) \approx 25.93993410234259692 1$$

mithilfe der folgenden summierten Quadraturformeln mit jeweils 1, 2 und 4 Teilintervallen: i) Trapezregel, ii) Simpsonregel.

b) Berechnen Sie den Fehler für jede der obigen Approximationen und vergleichen Sie mit der theoretisch zu erwartenden Konvergenzrate.

c) Wie lässt sich das obige Integral numerisch besser berechnen, *ohne* die Anzahl der Funktionsauswertungen zu erhöhen? Begründen Sie.

Lösung:

a) Die Trapez- und Simpsonregeln liefern die folgenden Zahlen:

(a) $T1 = \frac{3+1}{2}(f(-1) + f(2)) \approx 159.72$
(b) $T2 = \frac{3+1}{4}(f(-1) + 2(f(1/2) + f(2)) \approx 80.822$
(c) $T4 = \frac{3+1}{8}(f(-1) + 2f(-1/4) + 2f(1/2) + 2f(5/4) + f(2)) \approx 44.684$
(d) $S1 = \frac{3+1}{6}(f(-1) + 4f(1/2) + f(2)) \approx 54.523$
(e) $S2 = \frac{3+1}{12}(f(-1) + 4f(-1/4) + 2f(1/2) + 4f(5/4) + f(2)) \approx 32.638$
(f) $S4 = \frac{3+1}{24}(f(-1) + 4f(-5/8) + 2f(-1/4) + 4f(1/8) + 2f(1/2) + 4f(7/8) + 2f(5/4) + 4f(13/8) + f(2)) \approx 26.830$

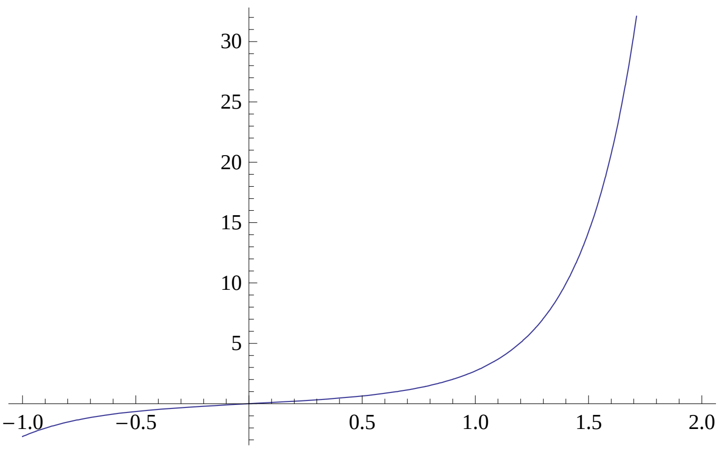
b) Für die Fehler gilt (I bezeichne hier den exakten Wert des Integrals)

(a) $|T1 - I| \approx 133.78$
(b) $|T2 - I| \approx 54.882$
(c) $|T4 - I| \approx 18.744$
(d) $|S1 - I| \approx 28.583$
(e) $|S2 - I| \approx 6.6979$
(f) $|S4 - I| \approx 0.88990$

Von den theoretischen Fehlerschranken (Satz 5.4) erwartet man, dass die Fehlerschranke bei der Trapezregel bei Halbierung der Intervalls um einen Faktor 4 kleiner wird, bei der Simpsonregel um einen Faktor $16 = 2^4$.
Die beobachteten Fehler nehmen folgendermassen ab:

$$\frac{|T1 - I|}{|T2 - I|} \approx 2.4376$$
$$\frac{|T2 - I|}{|T4 - I|} \approx 2.9280$$
$$\frac{|S1 - I|}{|S2 - I|} \approx 4.2675$$
$$\frac{|S2 - I|}{|S4 - I|} \approx 7.5265$$

Die beobachteten Fehler nehmen also deutlich weniger schnell ab als die theoretische Fehlerschranke.



c) Der Integrand ist eine ungerade Funktion. Deshalb gilt

$$\int_{-1}^{1} f(x)dx = 0$$

und somit liesse sich das Integral auch nur auf dem Intervall $[1,2]$ berechnen. Andernfalls könnte man auch eine adaptive Quadratur oder Gauss-Quadratur wählen.

4. **[Ausgleichsrechnung]**

In dieser Aufgabe soll die Funktion $f(x) = e^{-x^2}$ durch ein Polynom der Form $p(x) = a + bx + cx^2$ in den Punkten

$$x = [-1\ ,-0.5\ ,0\ ,0.5\ ,1]$$

im Sinne der kleinsten Quadrate approximiert werden, d.h. man sucht Werte für a, b und c so dass

$$\sum_{i=1}^{5} |f(x_i) - p(x_i)|^2$$

minimal ist.

a) Erklären Sie, weshalb das den Fehler minimierende Polynom $b = 0$ erfüllen muss.

b) Formulieren Sie das entsprechende Ausgleichsproblem. Vereinfachen Sie das Ausgleichsproblem so weit wie möglich. (Hinweis: Benutzen Sie dazu a)).

c) Lösen Sie das Ausgleichsproblem mithilfe der Normalengleichungen. Begründen Sie die Wahl der Methode, mit welcher Sie das resultierende lineare Gleichungssystem lösen.

Lösung:

a) Die Funktion $f(x)$ ist gerade und die Stützstellen x_i sind symmetrisch um Null verteilt. Aus diesem Grund muss auch die Parabel, welche f in den x_i am besten approximiert, eine gerade Funktion sein, und somit $b = 0$ gelten.
Ein richtiger Beweis dafür geht wie folgt:
Die Punkte x_i sind symmetrisch bezüglich der 0, also betrachten wir die Summe der Fehler die von x_1 und $-x_1$ kommen:

$$\begin{aligned}(f(x_1) - p(x_1))^2 + (\underbrace{f(-x_1)}_{=f(x_1)} - p(-x_1))^2 &= (f(x_1) - (a + bx_1 + cx_1^2))^2 + (f(x_1) \\ &\quad - (a - bx_1 + cx_1^2))^2 \\ &= \underbrace{2f(x_1)^2 - 4f(x_1)(a + cx_1^2) + 2(a + cx_1^2)^2}_{=2(f(x_1)-(a+cx_1^2))^2} + \underbrace{2b^2x_1^2}_{\geq 0}\end{aligned}$$

b) Das Ausgleichsproblem lautet:
Finde $\underline{x} = \begin{pmatrix} a \\ c \end{pmatrix}$ so dass der Vektor der Residuen $\underline{r}$ definiert durch

$$\underline{r} := \mathbf{A}\underline{x} - \underline{b} = \begin{pmatrix} \vdots & \vdots \\ 1 & x_i^2 \\ \vdots & \vdots \end{pmatrix} \underline{x} - \begin{pmatrix} \vdots \\ f(x_i) \\ \vdots \end{pmatrix}$$

minimale 2-Norm hat.
Das Problem lässt sich insofern vereinfachen, als dass für x_i und x_j mit gleichem

Betrag zwei Mal die selbe Gleichung auftaucht. Dies muss man bei den Normalengleichungen insofern berücksichtigen, als diejenigen Matrixelemente, welche von einem $|x_i| > 0$ herrühren, mit einem Faktor $\sqrt{2}$ gewichtet werden müssen (dasselbe gilt für die rechte Seite). Das reduzierte Problem heisst also:

$$\underline{r}_r := \mathbf{A}_r \underline{x} - \underline{b}_r = \begin{pmatrix} 1 & 0 \\ \sqrt{2} & \sqrt{2}\,0.5^2 \\ \sqrt{2} & \sqrt{2}\,1^2 \end{pmatrix} \underline{x} - \begin{pmatrix} f(0) \\ \sqrt{2}f(0.5) \\ \sqrt{2}f(1) \end{pmatrix}$$

soll minimale 2-Norm haben. Dieses Problem wird mithilfe der Normalengleichungen gelöst.

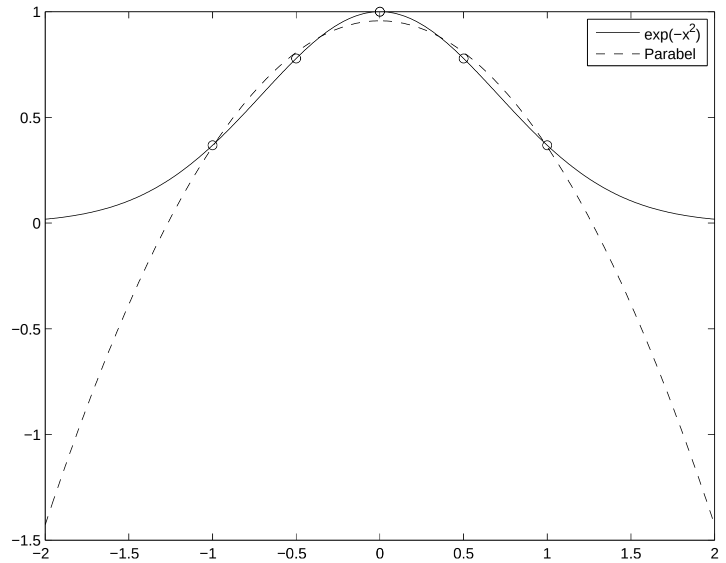
c) Da die Matrix $\mathbf{A}_r^\top \mathbf{A}_r$ SPD ist, bietet sich die Cholesky-Zerlegung an (spart Speicherplatz bei grossen Systemen).

```
function [x]=ausgleich

st = [0:0.5:1]'

f = exp(-st.^2);
f(2:end)=sqrt(2)*f(2:end);

n = size(f);

A = [ones(n,1) st.^2];
A(2:end,:)=sqrt(2)*A(2:end,:);

A
A'*A

% Loese Normalengleichungen


y = A'*f


%GLS mit Cholesky loesen
c=chol(A'*A)

y_c = c'\y
x_c = c\y_c

x_fine = [-2:0.01:2];

y = exp(-x_fine.^2);

y_approx = x_c(1)+x_c(2)*x_fine.^2;

f = exp(-st.^2);
plot(x_fine,y,x_fine,y_approx,'--',[-st' st'],[f' f'],'yo')
legend('exp(-x^2)','Parabel')
print -deps 'ausgleich.eps'
```

```
% A =
%
%    1.000000000000000                   0
%    1.414213562373095   0.353553390593274
%    1.414213562373095   1.414213562373095
%
%
% ans =
%
%    5.000000000000001   2.500000000000000
%    2.500000000000000   2.125000000000000
%
%
% y =
%
%    3.293360448485695
%    1.125159273878587
%
%
% c =
%
%    2.236067977499790   1.118033988749895
%                    0   0.935414346693485
%
%
% y_c =
%
%    1.472835567444641
%   -0.557529347510800
%
%
% x_c =
%
%    0.956684061333859
%   -0.596023943273440


% Noch einmal ohne Vereinfachung

st = [-1:0.5:1]'

f = exp(-st.^2);

n = size(f);

A = [ones(n,1) st.^2];

A
```

```
A'*A

% Loese Normalengleichungen

y = A'*f

%GLS mit Cholesky loesen
c=chol(A'*A)

y_c = c'\y
x_c = c\y_c



% A =
%
%    1.000000000000000   1.000000000000000
%    1.000000000000000   0.250000000000000
%    1.000000000000000                   0
%    1.000000000000000   0.250000000000000
%    1.000000000000000   1.000000000000000
%
%
% ans =
%
%    5.000000000000000   2.500000000000000
%    2.500000000000000   2.125000000000000
%
%
% y =
%
%    3.293360448485695
%    1.125159273878587
%
%
% c =
%
%    2.236067977499790   1.118033988749895
%                    0   0.935414346693485
%
%
% y_c =
%
%    1.472835567444641
%   -0.557529347510800
%
%
% x_c =
%
%    0.956684061333859
```

```
%   -0.596023943273440
```



5. **[Eigenwerte, MATLAB]**

Schreiben Sie eine MATLAB-Funktion, welche die Potenzmethode benutzt, um die Konditionszahl bezüglich der 2-Norm einer quadratischen Matrix **A** zu berechnen. Nutzen Sie dazu den Zusammenhang zwischen Singulärwerten und der 2-Norm einer Matrix.

Input soll die Matrix **A** sein, Output eine Approximation der Konditionszahl von **A** in der 2-Norm.

Verwenden Sie ein geeignetes Abbruchkriterium. Als Startvektor für die Potenzmethode sollen Zufallsvektoren verwendet werden. **Lösung:**

```
function [c]=Potenz(A)

% Implementiert die Potenzmethode zur Berechnung der Konditionszahl von A
% bezueglich der 2-Norm

A_sym = A'*A;

n = size(A_sym,1);
x=rand(n,1);
x=x/norm(x);
i=0;

while( norm(A_sym*x-(x'*A_sym*x)*x)>1e-10 && i<1e2)
    x=A_sym*x;
    x=x/norm(x);
    i=i+1;
end
i
```

```
lambda1 = x'*A_sym*x;

sig1 = sqrt(lambda1);

% Inverse Iteration fuer kleinsten EW
y=rand(n,1);
y=y/norm(y);
i=0;

while (norm(A_sym*y-(y'*A_sym*y)*y)>1e-10 &&i<1e2)
    y=A_sym\y;
    y=y/norm(y);
    i=i+1;
end

i
lambdan=y'*A_sym*y;

sign = sqrt(lambdan);

c = sig1/sign

fehler = abs(c-cond(A))
```