2 Direkte Lösungsverfahren für lineare Gleichungen

Sei $A \in \mathbb{R}^{N\times N}$ invertierbar und $b \in \mathbb{R}^N$. Löse $Ax = b$ genau und effizient.

Die LR-Zerlegung

Wir berechnen eine Zerlegung $A = LR$ mit $L, R \in \mathbb{R}^{N\times N}$ und den folgenden Eigenschaften:

$$L[n,n] = 1,\ R[n,n] \neq 0,\ L[n,k] = R[k,n] = 0 \quad \text{für} \quad n = 1,\dots,N,\ k > n.$$

L ist eine untere normierte Dreiecksmatrix: $L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix}$.

R ist eine obere reguläre Dreiecksmatrix: $R = \begin{pmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{pmatrix}$.

Um $Ax = b$ zu lösen, löse zunächst $Ly = b$, $Rx = y$ und damit $Ax = LRx = Ly = b$.

(2.1) Satz

(a) *Sei $L \in \mathbb{R}^{N\times N}$ eine normierte untere Dreiecksmatrix (d.h.* $\operatorname{diag} L = I_N$ *und $L[1:n, n+1] = 0_n$ für $n = 1,\dots,N-1$), und sei $b \in \mathbb{R}^N$. Dann ist L regulär und das Lineare Gleichungssystem (LGS) $Ly = b$ ist mit $O(N^2)$ Operationen lösbar.*

(b) *Für eine reguläre obere Dreiecksmatrix $R \in \mathbb{R}^{N\times N}$, (d.h. $R[n,n] \neq 0$ für alle n und $R[n+1, 1:n] = 0_n^T$ für $n < N$) ist das LGS $Rx = y$ in $O(N^2)$ Operationen lösbar.*

Beweis. Zu (a): Wir lösen das LGS $L[1:n, 1:n]x[1:n] = y[1:n]$ induktiv für $n = 1,\dots,N$ durch *Vorwärtssubstitution*. Es gilt $\det(L[1:n, 1:n]) = 1$, also ist $L[1:n, 1:n]$ regulär.
Für $n = 1$ gilt $y[1] = b[1]$.
Nun sei für $n > 1$ bereits $y[1:n-1]$ mit $L[1:n-1, 1:n-1]y[1:n-1] = b[1:n-1]$ berechnet. Aus dem Ansatz

$$\begin{pmatrix} L[1:n-1, 1:n-1] & 0_{n-1} \\ L[n, 1:n-1] & 1 \end{pmatrix} \begin{pmatrix} y[1:n-1] \\ y[n:n] \end{pmatrix} = \begin{pmatrix} b[1:n-1] \\ b[n:n] \end{pmatrix}$$

folgt $L[n, 1:n-1]y[1:n-1] + y[n:n] = b[n:n]$, also ist

$$y[n:n] = b[n:n] - L[n, 1:n-1]y[1:n-1]$$

mit $O(N)$ Operationen berechenbar. Zusammen werden somit $O(N^2)$ Operationen benötigt.
Zu (b): Umgekehrt lösen wir LGS $R[n:N, n:N]y[n:N] = b[n:N]$ induktiv für $n = N,\dots,1$ durch *Rückwärtssubstitution*. Da R regulär ist, gilt $\det(R) = \prod_{n=1}^{N} R[n,n] \neq 0$, also $R[n,n] \neq 0$ für alle $n = 1,\dots,N$.
Für $n = N$ setze $x[N] = y[N]/R[N,N]$.
Nun sei für $n < N$ bereits $x[n:N]$ mit $R[n:N, n:N]x[n:N] = y[n:N]$ berechnet. Aus

$$\begin{pmatrix} R[n-1, n-1] & R[n-1, n:N] \\ 0_{N-n} & R[n:N, n:N] \end{pmatrix} \begin{pmatrix} x[n-1:n-1] \\ x[n:N] \end{pmatrix} = \begin{pmatrix} y[n-1:n-1] \\ y[n:N] \end{pmatrix}$$

folgt $R[n-1,1:n-1]x[n-1:n-1]+R[n-1,n:N]x[n:N]=y[n-1:n-1]$, also ist

$$x[n-1:n-1]=\Big(y[n-1:n-1]-R[n-1,n:N]x[n:N]\Big)R[n-1,1:n-1]$$

mit $O(N)$ Operationen berechenbar. Zusammen werden somit $O(N^2)$ Operationen benötigt.

□

(2.2) Satz
a) *Die normierten unteren Dreiecksmatrizen bilden eine Gruppe.*

b) *Die regulären oberen Dreiecksmatrizen bilden eine Gruppe.*

Beweis. Das Produkt von Dreiecksmatrizen ist wieder eine Dreiecksmatrix.
Wie zeigen induktiv, dass $L[1:n,1:n]^{-1}$ normierte untere Dreiecksmatrix ist.
Für $n=1$ gilt $L[1,1]^{-1}=1$.
Nun betrachte $n>1$. Der Ansatz

$$\begin{pmatrix} L[1:n,1:n] & 0_n \\ L[n+1,1:n] & 1 \end{pmatrix}\begin{pmatrix} L[1:n,1:n]^{-1} & 0_n \\ \ell^T & 1 \end{pmatrix} = \begin{pmatrix} I_N & 0_n \\ 0_n^T & 1 \end{pmatrix}$$

mit $\ell\in\mathbb{R}^n$ ergibt $\ell^T=-L[n+1,1:n]L[1:n,1:n]^{-1}$. Also besitzt $L[1:n+1,1:n+1]$ eine Inverse. Wenn $L[1:n,1:n]^{-1}$ normierte untere Dreiecksmatrix ist, dann ist damit auch $L[1:n+1,1:n+1]^{-1}$ normierte untere Dreiecksmatrix.
Wenn R reguläre obere Dreiecksmatrix ist, dann ist $L=\operatorname{diag}(R)^{-1}R^T$ normierte untere Dreiecksmatrix. Also ist L^{-1} normierte untere Dreiecksmatrix und somit

$$R^{-1}=\Big(L^T\operatorname{diag}(R)\Big)^{-1}=\operatorname{diag}(R)^{-1}L^{-T}$$

obere Dreiecksmatrix.

□

Diagonalskalierung

Die Multiplikation einer Matrix $A\in\mathbb{R}^{M,N}$ mit einer Diagonalmatrix $D_M=\operatorname{diag}(d_1,...,d_M)$ von links entspricht einer Zeilenskalierung (bzw. mit $D_N=\operatorname{diag}(d_1,...,d_N)$ von rechts einer Spaltenskalierung):

$$D_MA=\begin{pmatrix} d_1A[1,1:N] \\ \vdots \\ d_MA[M,1:N] \end{pmatrix},\qquad AD_N=\Big(d_1A[1:M,1]\Big|\ \cdots\ \Big|d_NA[1:M,N]\Big).$$

(2.3) Satz
Wenn eine Matrix $A\in\mathbb{R}^{N\times N}$ eine LR-Zerlegung $A=LR$ besitzt, dann ist A regulär und das LGS $Ax=b$ ist mit $O(N^2)$ Operationen lösbar.

Beweis. Es gilt nach Voraussetzung $\det(L)=1$ und $\det(R)\neq 0$, also $\det(A)=\det(LR)=\det(L)\det(R)=\det(R)\neq 0$. Also ist A regulär.
Weiterhin gilt $b=Ax=LRx=Ly$ für $y=Rx$. Also löse zunächst $Ly=b$ und dann $Rx=y$.
□

(2.4) Satz
Eine Matrix $A \in \mathbb{R}^{N\times N}$ besitzt genau dann eine LR-Zerlegung, wenn alle Hauptuntermatrizen $A[1:n,1:n]$, $(n=1,\ldots,N)$ regulär sind. Die LR-Zerlegung ist eindeutig und lässt sich mit $O(N^3)$ Operationen berechnen.

Beweis. (i) Existenz:
Wir berechnen die LR-Zerlegung $A[1:n,1:n] = L[1:n,1:n]R[1:n,1:n]$ induktiv für $n = 1,...,N$.
Für $n=1$ gilt $L[1,1]=1$, $R[1,1]=A[1,1]\neq 0$.
Nun sei eine Zerlegung $A[1:n,1:n] = L[1:n,1:n]R[1:n,1:n]$ berechnet. Der Ansatz

$$\begin{pmatrix} L[1:n,1:n] & 0_n \\ L[n+1,1:n] & 1 \end{pmatrix}\begin{pmatrix} R[1:n,1:n] & R[1:n,n+1] \\ 0_n^T & R[n+1,n+1] \end{pmatrix} = \begin{pmatrix} A[1:n,1:n] & A[1:n,n+1] \\ A[n+1,1:n] & A[n+1,n+1] \end{pmatrix}$$

ergibt die Gleichungen

$$\begin{aligned} L[1:n,1:n]R[1:n,n+1] &= A[1:n,n+1]\,, \\ L[n+1,1:n]R[1:n,1:n] &= A[n+1,1:n]\,, \\ L[n+1,1:n]R[1:n,n+1]+R[n+1,n+1] &= A[n+1,n+1]\,. \end{aligned}$$

Da $A[1:n,1:n]$ regulär ist, sind auch $L[1:n,1:n]$ und $R[1:n,1:n]$ regulär, also gilt

$$\begin{aligned} R[1:n,n+1] &= L[1:n,1:n]^{-1}A[1:n,n+1]\,, \\ L[n+1,1:n] &= A[n+1,1:n]R[1:n,1:n]^{-1}\,, \\ R[n+1,n+1] &= A[n+1,n+1]-L[n+1,1:n]R[1:n,n+1]\,. \end{aligned}$$

Damit ist eine LR-Zerlegung $A[1:n+1,1:n+1]=L[1:n+1,1:n+1]R[1:n+1,1:n+1]$ berechnet, und aus

$$\begin{aligned} 0\neq \det A[1:n+1,1:n+1] &= \det R[1:n+1,1:n+1] \\ &= \det\big(R[1:n,1:n]\big)\,R[n+1,n+1] \end{aligned}$$

folgt $R[n+1,n+1]\neq 0$. Also ist $R[1:n+1,1:n+1]$ regulär.
(ii) Eindeutigkeit:
Sei $A=\tilde{L}\tilde{R}$ eine weitere LR-Zerlegung von A. Dann ist $LR=\tilde{L}\tilde{R}$, also $\tilde{L}^{-1}L=\tilde{R}R^{-1}$. Aus Satz (2.2) folgt, dass $\tilde{L}^{-1}L$ und $\tilde{R}R^{-1}$ sowohl untere normierte Dreiecksmatrix ist wie auch obere Dreieckmatrix. Somit gilt $\tilde{L}^{-1}L=\tilde{R}R^{-1}=I_N$, also ist $\tilde{L}=L$, $\tilde{R}=R$.
(iii) Aufwand:
In jedem Schritt werden zwei Dreieckssysteme mit $O(n^2)$ Operationen gelöst, N Schritte werden benötigt. Damit ist die Anzahl der Operationen in der Größenordnung $O(N^3)$.
□

Bemerkung
Die LR-Zerlegung respektiert die Hüllenstruktur. Für $k<n$ gilt:

$$\begin{aligned} A[n,1:k]=0_k^T &\implies L[n,1:k]=0_k^T \\ A[1:k,n]=0_k &\implies R[1:k,n]=0_k \end{aligned}$$

Algorithmus 1 Berechnung einer LR-Zerlegung, Vorwärts- und Rückwärtssubstitution.

```
function x = lr_solve(A,b)
  N = size(A,1);
  for n=1:N-1
    A(n+1:N,n) = A(n+1:N,n)/A(n,n);
    A(n+1:N,n+1:N) = A(n+1:N,n+1:N) - A(n+1:N,n) * A(n,n+1:N);
  end
  x = b;
  for n=2:N
    x(n) = x(n) - A(n,1:n-1) * x(1:n-1);
  end
  for n=N:-1:1
    x(n) = (x(n) - A(n,n+1:N)*x(n+1:N))/A(n,n);
  end
return
```

(2.5) Definition
Eine Matrix $A \in \mathbb{R}^{N\times N}$ heißt diagonaldominant*, wenn*

$$|A[n,n]| \geq \sum_{\substack{k=1\\k\neq n}}^{N} |A[n,k]| \qquad (n = 1,\ldots,N)$$

gilt, und strikt diagonaldominant*, falls*

$$|A[n,n]| > \sum_{\substack{k=1\\k\neq n}}^{N} |A[n,k]| \qquad (n = 1,\ldots,N)$$

(2.6) Folgerung
Wenn $A \in \mathbb{R}^{N\times N}$ strikt diagonal dominant ist, dann existiert eine LR-Zerlegung.

Beweis. Wenn keine LR-Zerlegung existiert, dann existiert ein n, für das die Untermatrix $A[1:n,1:n]$ singulär ist, und es existiert $x \in \mathbb{R}^n$ mit $A[1:k,1:k]x = 0$ und $x \neq 0_n$. Wähle $k \in \{1,\ldots,n\}$ mit $|x[k]| \geq |x[j]|$ für alle $j = 1,...,n$. Dann folgt aus

$$\begin{aligned} |A[k,k]|\,|x[k]| = \Big|\sum_{\substack{j=1\\j\neq k}}^{n} A[k,j]x[j]\Big| &\leq \sum_{\substack{j=1\\j\neq k}}^{n} |A[k,j]|\,|x[j]| \\ &\leq \sum_{j\neq k} |A[k,j]|\,|x[k]| < |x[k]|\,|A[k,k]| \end{aligned}$$

und $x[k] \neq 0$ ein Widerspruch.

□

(2.7) Satz
Sei $A \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit. Dann existiert genau eine Cholesky-Zerlegung $A = LL^T$ mit regulärer unterer Dreiecksmatrix L.

Beweis. Da A positiv definit ist, ist $\det(A[1:n,1:n]) > 0$; es existiert also eine LR-Zerlegung $A = \tilde{L}\tilde{R}$.
Definiere $D = \operatorname{diag}(\tilde{R})$. Aus

$$\prod_{k=1}^{n} D[k,k] = \det(\tilde{R}[1:n,1:n]) = \det(A[1:n,1:n]) > 0$$

folgt induktiv $D[k,k] > 0$. Wir skalieren die Spalten von $\tilde{L}$ und Zeilen von $\tilde{R}$ und setzen $L = \tilde{L}D^{\frac{1}{2}}$, $R = D^{-\frac{1}{2}}\tilde{R}$. Damit folgt

$$LR = \tilde{L}\tilde{R} = A = A^T = \tilde{R}^T\tilde{L}^T = R^T L^T$$

Hieraus ergibt sich $R^{-T}L = L^T R^{-1} = I_N$, woraus $R = L^T$ folgt.

□

Die Berechnung der Cholesky-Zerlegung benötigt nur halb soviele Operationen wie die Berechnung einer LR-Zerlegung.

Algorithmus 2 Berechnung einer Cholesky-Zerlegung, Vorwärts- und Rückwärtssubstitution.

```
function x = cholesky_solve(A,b)
  N = size(A,1);
  for n=1:N
    A(n:N,n) = A(n:N,n) - A(n:N,1:n-1) * A(n,1:n-1)';
    A(n:N,n) = A(n:N,n) / sqrt(A(n,n));
  end
  x = b
  for n=1:N
    x(n) = (x(n) - A(n,1:n-1) * x(1:n-1))/ A(n,n);
  end
  for n=N:-1:1
    x(n) = (x(n) - A(n+1:N,n)' * x(n+1:N))/ A(n,n);
  end
return
```

Die LR-Zerlegung mit Pivotsuche

Durch Zeilenvertauschungen von A lässt sich immer garantieren, dass eine LR-Zerlegung existiert.

(2.8) Definition
Sei $\pi \in S_N$ eine Permutation. Dann heißt $P_\pi = \left(e_{\pi^{-1}(1)} \middle| \cdots \middle| e_{\pi^{-1}(N)}\right) \in \mathbb{R}^{N\times N}$ Permutationsmatrix zu π.

Die Multiplikation einer Matrix $A \in \mathbb{R}^{M\times N}$ mit einer Permutationsmatrix P_π von links vertauscht die Zeilen (bzw. mit P_σ von rechts vertauscht die Spalten).

$$(P_\pi A)[n,k] = A[\pi(n),k]$$
$$(AP_\sigma)[n,k] = A[n,\sigma^{-1}(k)]$$

Also ist die Zeile n von A die Zeile $\pi(n)$ von $P_\pi A$.

Als Beispiel betrachte $\pi \in S_3$ mit $\pi(1)=3$, $\pi(2)=1$ und $\pi(3)=2$ und $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$.

Dann gilt $P_\pi = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $P_\pi A = \begin{pmatrix} 7 & 8 & 9 \\ 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$ und $AP_\pi = \begin{pmatrix} 2 & 3 & 1 \\ 5 & 6 & 4 \\ 8 & 9 & 7 \end{pmatrix}$.

(2.9) Satz
Die Permutationsmatrizen in $\mathbb{R}^{N\times N}$ bilden eine Gruppe.
Es gilt $P_\pi P_\sigma = P_{\pi\circ\sigma}$ und $(P_\pi)^{-1} = P_\pi^T$.

(2.10) Satz
Sei $A \in \mathbb{R}^{N\times N}$ regulär. Dann existiert eine Permutationsmatrix P, so dass PA eine LR-Zerlegung mit $|L[n,k]| \le 1$ besitzt.

Beweis. Wir konstruieren die Zerlegung induktiv. Für $N=1$ ist nichts zu zeigen. Nun sei $N>1$. Suche in der ersten Spalte $n=1$ das *Pivotelement* $A[p,1]$ mit dem größten Betrag, d.h. $|A[p,1]| \ge |A[k,1]|$ für alle $k=1,...,N$. Da A regulär ist, ist $A[1:N,1] \neq 0_N$, also $A[p,1] \neq 0$. Falls $p \neq 1$, vertausche die Zeilen 1 und p und setze $\tau = [1\,p]$. Sonst setze $\tau = \mathrm{id}$. Nun setze $A_1 = P_\tau A$, d.h. $A_1[1,1] = A[p,1]$,

$$L_1 = \begin{pmatrix} 1 & 0_{N-1}^T \\ -A_1[2:N,1]/A_1[1,1] & I_{N-1} \end{pmatrix}$$

$$\begin{aligned} L_1A_1 &= \begin{pmatrix} 1 & 0_{N-1}^T \\ -A_1[2:N,1]/A_1[1,1] & I_{N-1} \end{pmatrix} \begin{pmatrix} A_1[1,1] & A_1[1,2:N] \\ A_1[2:N,1] & A_1[2:N,2:N] \end{pmatrix} \\ &= \begin{pmatrix} A_1[1,1] & A_1[1,2:N] \\ 0_{N-1} & A_2 \end{pmatrix} \end{aligned}$$

mit der Restmatrix $A_2 = A_1[2:N,2:N] + L_1[2:N,1]A_1[1,2:N]$. Dann gilt per Konstruktion $L_1A_1 = R_1$. Aus

$$A_1[1,1]\det A_2 = \det(L_1)\det(A_1) = \det(P_\tau A) = \underbrace{\det(P_\tau)}_{=\pm 1}\det A \neq 0$$

folgt $\det A_2 \neq 0$. Also existiert in $\mathbb{R}^{N-1\times N-1}$ per Induktionsvoraussetzung eine LR-Zerlegung $P_2A_2 = L_2R_2$. Wir setzen

$$\hat{P}_2 = \begin{pmatrix} 1 & 0_{N-1}^T \\ 0_{N-1} & P_2 \end{pmatrix}, \hat{L}_2 = \begin{pmatrix} 1 & 0_{N-1}^T \\ 0_{N-1} & L_2 \end{pmatrix}, R = \begin{pmatrix} A_1[1,1] & A_1[1,2:N] \\ 0_{N-1} & R_2 \end{pmatrix}.$$

Aus $P = \hat{P}_2 P_\tau$ folgt $\hat{P}_2 L_1 A_1 = \hat{L}_2 R$. Damit erhalten wir eine LR-Zerlegung mit Pivotsuche:

$$PA = \hat{P}_2 P_\tau A = \hat{P}_2 A_1 = \hat{P}_2 L_1^{-1} L_1 A_1 = L_1^{-1}\hat{P}_2 L_1 A_1 = L_1^{-1}\hat{L}_2 R = LR$$

□

Das System $Ax = b$ wird wie folgt gelöst: Es gilt $LRx = PAx = Pb$. Berechne erst $y = Rx$ durch Vorwärtssubstitution von $Ly = Pb$ und dann x durch Rückwärtssubstitution von $Rx = y$.

Die *LR*-Zerlegung mit *Spaltenpivotsuche* benötigt zusätzlich $O(N^2)$ Operationen. Die *Stabilität* (siehe unten) der LR-Zerlegung lässt sich durch sogenannte *totale Pivotsuche* erhöhen: Durch den Zeilen- und Spaltentausch ersetze in jedem Schritt $A_n[1,1]$ durch den maximalen Eintrag in der Restmatrix A_n. Hierfür werden allerdings $O(N^3)$ Operationen benötigt.

Algorithmus 3 Berechnung einer LR-Zerlegung mit Pivotsuche, Vorwärts- und Rückwärtssubstitution.

```
function x = lr_pivot_solve(A,b)
  N = size(A,1);
  p = (1:N)';
  for n = 1:N-1
    [r,m] = max(abs(A(n:N,n)));
    m = m+n-1;
    if abs(A(m,n))<eps
      error('*** ERROR *** Matrix fast singulär');
    end
    if (m ~= n)
      A([n m],:) = A([m n],:);   p([n m]) = p([m n]);
    end
    A(n+1:N,n) = A(n+1:N,n)/A(n,n);
    A(n+1:N,n+1:N) = A(n+1:N,n+1:N) - A(n+1:N,n)*A(n,n+1:N);
  end
  x = b(p);
  for n=2:N
    x(n) = x(n) - A(n,1:n-1) * x(1:n-1);
  end
  for n=N:-1:1
    x(n) = (x(n) - A(n,n+1:N)*x(n+1:N))/A(n,n);
  end
return
```

Störungsrechnung

(2.11) Satz
Sei $A \in \mathbb{R}^{N\times N}$ regulär, und sei $\Delta A \in \mathbb{R}^{N\times N}$ so klein, dass $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$ gilt. Dann ist die Matrix $\tilde{A} = A + \Delta A$ regulär. Sei $b \in \mathbb{R}^N$, $b \neq 0_N$, $\Delta b \in \mathbb{R}^N$ klein und $\tilde{b} = b + \Delta b$.
Dann gilt für die Lösung $x \in \mathbb{R}^N$ von $Ax = b$ und $\tilde{x} \in \mathbb{R}^N$ von $\tilde{A}\tilde{x} = \tilde{b}$

$$\frac{|\Delta x|}{|x|} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{\|\Delta A\|}{\|A\|}} \left(\frac{|\Delta b|}{|b|} + \frac{\|\Delta A\|}{\|A\|} \right).$$

Dabei ist $\Delta x = \tilde{x} - x$ der absolute Fehler, *$\frac{|\Delta x|}{|x|}$ der* relative Fehler, *und*

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

die Kondition *von A.*

Beweis. Aus der Voraussetzung $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$ folgt, dass die Reihe $\sum_{k=0}^{\infty} \left(-\Delta A A^{-1}\right)^k$ konvergiert. Also ist $\tilde{A} = A + \Delta A \in \mathbb{R}^{N\times N}$ regulär, es gilt für die Inverse die Darstellung

$$\left(A + \Delta A\right)^{-1} = A^{-1} \sum_{k=0}^{\infty} \left(-\Delta A A^{-1}\right)^k$$

und die Normabschätzung $\|(A+\Delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\|\Delta A\|}$.
Ferner folgt aus $(A+\Delta A)(x+\Delta x) = b + \Delta b$ und $Ax = b$ die Gleichung $(A+\Delta A)\Delta x = \Delta b - \Delta A x$ und somit

$$\begin{aligned}
|\Delta x| &\leq \|(A+\Delta A)^{-1}\|(|\Delta b| + \|\Delta A\||x|) \\
&\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|} \left(\frac{|\Delta b|}{|x|} + \|\Delta A\| \right) |x| \\
&\leq \frac{\|A\|\|A^{-1}\|}{1 - \|A\|\|A^{-1}\|\frac{\|\Delta A\|}{\|A\|}} \left(\frac{|\Delta b|}{\|A\||x|} + \frac{\|\Delta A\|}{\|A\|} \right) |x|.
\end{aligned}$$

Mit $\kappa(A) = \|A\|\,\|A^{-1}\|$ und $\frac{|\Delta b|}{\|A\||x|} \leq \frac{|\Delta b|}{|b|}$ folgt die Behauptung.

□

Vektor- und Matrixnormen

Wir verwenden für $x \in \mathbb{R}^N$ und $A \in \mathbb{R}^{K\times N}$

$\|x\|_1 = \sum_{n=1}^{N} \|x[n]\|$	1-Norm
$\|x\|_2 = \sqrt{x^T x}$	2-Norm / euklidische Norm
$\|x\|_\infty = \max_{n=1,\ldots,\infty} \|x[n]\|$	Maximumsnorm
$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$	zugeordnete Operatornorm $(p = 1, 2, \infty)$.

(2.12) Satz
Sei $A \in \mathbb{R}^{K\times N}$. $\|\cdot\|_p$ ist submultiplikativ und es gilt

$$\|A\|_1 = \max_{n=1,\dots,N} \sum_{k=1}^{K} |A[k,n]| \qquad \textit{Spaltensummennorm}$$

$$\|A\|_2 = \sqrt{\rho(A^T A)} \qquad \textit{Spektralnorm}$$

$$\|A\|_\infty = \max_{k=1,\dots,K} \sum_{n=1}^{N} |A[k,n]| \qquad \textit{Zeilensummennorm}.$$

Dabei ist

$$\rho(A) = \max\{|\lambda| \colon \lambda \in \sigma(A)\} \qquad \textit{Spektralradius von } A,$$
$$\sigma(A) = \{\lambda \in \mathbb{C} \colon \det(A - \lambda I_N) = 0\} \qquad \textit{Spektrum von } A.$$

Beweis. Zur Submultiplikativität der Norm:

$$\|AB\|_p = \sup_{x\neq 0} \frac{|ABx|_p}{|x|_p} \leq \sup_{x\neq 0} \frac{\|A\|_p |Bx|_p}{|x|_p} = \|A\|_p \|B\|_p$$

$p = 1$: Es gilt

$$\begin{aligned} |Ax|_1 &= \sum_{k=1}^{K} \left| \sum_{n=1}^{N} A[k,n]x[n] \right| \\ &leq \sum_{k=1}^{K} \sum_{n=1}^{N} |A[k,n]||x[n]| \\ &\leq \max_{n=1,\dots,N} \sum_{k=1}^{K} |A[k,n]| \sum_{n=1}^{N} |x[n]| \\ &= \|A\|_1 |x|_1 \end{aligned}$$

Für n_0 mit $\sum_{k=1}^{K} |A[k,n_0]| = \|A\|_1$ ist

$$|Ae_{n_0}|_1 = \sum_{k=1}^{K} |A[k,n_0]e_{n_0}[n_0]| = \|A\|_1$$

Also gilt

$$\|A\|_1 = \min\{C \geq 0 \mid |Ax|_1 \leq C|x|_1\} = \sup_{x\neq 0} \frac{|Ax|_1}{|x|_1}$$

$p = 2$: Zu $A^T A \in \mathbb{R}^{N,N}$ symmetrisch existiert eine orthogonale Matrix $Q \in \mathbb{R}^{N,N}$ (d.h $Q^{-1} = Q^T$) mit

$$Q^T A^T A Q = \operatorname{diag}(\lambda_1, \dots \lambda_N) =: \Lambda$$

und Eigenwerten $\lambda_n \geq 0$ von $A^T A$.
Es gilt

$$|Qy|_2^2 = (Qy)^T Qy = y^T Q^T Qy = y^T y = |y|_2^2$$
$$|AQy|_2^2 = (AQy)^T AQy = y^T Q^T A^T AQy = y^T \Lambda y \leq |\sqrt{\Lambda} y|_2^2.$$

Damit ergibt sich (mit der Substitution $x = Qy$)

$$\|A\| = \sup_{x\neq 0} \frac{|Ax|_2}{|x|_2} = \sup_{y\neq 0} \frac{|AQy|_2}{|Qy|_2} = \sup_{y\neq 0} \frac{|\sqrt{\Lambda}y|_2}{|y|_2} \leq \max \sqrt{|\lambda_n|}\,,$$

es gilt also $\|A\|_2 \leq \sqrt{\rho(A^TA)}$.
Wähle n_0 mit $\lambda_{n_0} = \rho(A^TA)$ und $q \in \mathbb{R}^N$ mit $Aq = \lambda_{n_0}q$, $q \neq 0_N$ mit $A^TAq = \lambda_{n_0}q$. Dann gilt

$$\|A\|_2 = \sup_{x\neq 0} \frac{|Ax|_2}{|x|_2} \geq \frac{|Aq|_2}{|q|_2} = \sqrt{\lambda_{n_0}} = \rho(A^TA)\,,$$

da $|Aq|_2^2 = q^TA^TAq = \lambda_{n_0}q^Tq = \lambda_{n_0}|q|_2^2$.
$p = \infty$: siehe Analysis II

□

Bemerkung
Wenn A symmetrisch ist mit Eigenwerten $\lambda_1, \ldots, \lambda_N$, dann gilt

$$\|A\|_2 = \sqrt{\rho(A^2)} = \rho(A) = \max_{n=1,\ldots,N} |\lambda_n|$$

und

$$\|A^{-1}\|_2 = \max_{n=1,\ldots,N} \frac{1}{|\lambda_n|} = \frac{1}{\min\limits_{n=1,\ldots,N} |\lambda_n|}.$$

Also ist

$$\kappa(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\max\limits_{n=1,\ldots,N} |\lambda_n|}{\min\limits_{n=1,\ldots,N} |\lambda_n|}\,.$$

Grundlagen der Arithmetik

In der Praxis muss das Schema Eingabe—Algorithmus—Resultat ergänzt werden:
Untersuche Eingabe mit Fehlern—Algorithmus mit Fehlern—Resultat mit Fehlern.

(2.13) Definition
a) *Die* Gleitkommazahlen *zur Basis $B \in \{2,3,4,\ldots\}$, Mantissenlänge $M \in \mathbb{N}$ und Exponent $E \in \mathbb{N}$ ist die Menge*

$$\mathrm{FL} = \left\{ \pm B^e \sum_{m=1}^{M} a_m B^{-m} \in \mathbb{Q} \mid e = e^- + \sum_{k=0}^{E-1} e_k B^{-k},\, a_k, e_k \in \{0,\ldots,B-1\} \right\}$$

b) *Eine Gleitkommaarithmetik wird durch eine Abbildung*

$$\mathrm{fl}\colon \mathbb{R} \longrightarrow \mathrm{FL}$$

mit $\mathrm{fl}(x) = x$ für $x \in \mathrm{FL}$ definiert: $x \oplus y = \mathrm{fl}(x+y)$, $x \odot y = \mathrm{fl}(xy)$.

Mit $\text{eps} := \sup\left\{\frac{|x-\text{fl}(x)|}{|x|} \mid x \in \mathbb{R}\backslash\text{FL}\right\}$ bezeichnen wir die Maschinengenauigkeit.
Als Ergebnis einer Rechenoperation werden zusätzlich folgende Fehlerzustände definiert:

NaN (not a number)	nicht definiertes Ergebnis
overflow	Ergebnis $>$ max FL oder $<$ min FL
underflow	$\text{fl}(x) = 0$, aber $x \neq 0$

Beispiele:

NaN (not a number)	Division durch 0, $\sqrt{-1}$
overflow	$\exp(100)$
underflow	$\exp(-100)$

Im IEEE-Standard wird double mit 64 Bit = 8 Byte dargestellt:

$$\begin{array}{lll} 1\ \textit{Bit} & \textit{Vorzeichen} & \\ 52\ \textit{Bits} & \textit{Mantisse} & m = \sum_{i=0}^{51} a_i 2^i \qquad a_i \in \{0,1\} \\ 11\ \textit{Bits} & \textit{Exponenten} & e = E^- + \sum_{j=0}^{10} b_j 2^j \qquad b_j \in \{0,1\} \end{array}$$

Es ist $M = 2^{52} - 1$, $E^- = -1022$, $E^+ = 1023$ und damit gilt für die Maschinengenauigkeit

$$\text{eps} \approx 10^{-16}.$$

Außerdem erhalten wir

$$\text{FL} \subset [-10^{308}, -10^{-308}] \cup \{0\} \cup [10^{-308}, 10^{308}].$$

Folgende Probleme treten wegen der begrenzten Zahlenmenge auf:

(1) Unvermeidliche *Rundungsfehler*:
Setze double $x = 4 \cdot \arctan(1)$. Dies sollte π sein. Wir erhalten:

$$x = 3.14159265358979316... \qquad \pi = 3.14159265358979323 84...$$

(2) *Auslöschung*:
Auslöschung tritt immer dann auf, wenn zwei fast gleich große Zahlen voneinander abgezogen werden.
Sei beispielsweise double $x = \exp(1)$. Berechne nun $y = x + 10^{-8}$ und $z = \text{fl}(x - y)$. Dann gilt

$$|z - 10^{-8}| \approx 1e^{-16},$$

und für den relativen Fehler

$$\frac{|z - 10^{-8}|}{|z|} \approx 1e^{-8}.$$

Also ist die Berechnung nur auf 8 Dezimalstellen genau.

Ein typisches Beispiel für vermeidbare Rundungsfehler ist die Auswertung von $\exp(-x)$ für $x > 0$: Die Berechnung der Exponentialfunktion ist gut konditioniert, aber die direkte numerische Auswertung der Exponentialreihe für negative Argumente ist (durch Auslöschung) sehr ungenau. Die Auswertung $\exp(-x) = 1/\exp(x)$ ist numerisch stabil.

Kondition und Stabilität

Bei vielen Anwendungen sind die Daten (z.B. durch ungenaue Messungen) unsicher. Dazu führen wir einige Begriffe ein.

Bezeichnung

a) *Ein Problem heißt* sachgemäß gestellt*, wenn es eindeutig lösbar ist und die Lösung stetig von den Daten abhängt.*

b) *Die* Kondition *eines Problems ist ein Maß dafür, wie stark die Abhängigkeit der Lösung von den Daten ist.*

c) *Die* Stabilität *eines numerischen Algorithmus ist ein Maß dafür, wie stark die Datenabhängigkeit der numerischen Lösung im Vergleich zu der tatsächlichen Lösung ist.*

Also: Ein Problem ist gut konditioniert, wenn kleine Änderungen der Daten die Lösung nur wenig ändern.
Ein numerischer Algorithmus ist stabil, wenn durch kleine Änderungen der Daten die Änderung der numerischen Lösung durch den Algorithmus nicht zusätzlich verstärkt wird.
Damit ergibt sich folgende Charakterisierung:

Kondition des Problems:	Unvermeidbare Fehlerverstärkung bei optimaler Problemlösung.
Stabilität des Algorithmus:	Der gewählte Algorithmus vergrößert den Fehler nicht stärker als die unvermeidliche Fehlerverstärkung.

Beispiel

a) *Wetterberechnung ist schlecht konditioniert: Kleine Veränderungen der Ausgangslage können langfristig große Auswirkungen haben.*

b) *Sei $V = L_2(0,1)$ ausgestattet mit der Norm $\|v\|_V^2 = \int_0^1 |v(t)|^2 dt$; $V_N = \mathbb{P}_N$ seien die Polynome mit Grad kleiner oder gleich N.*
Zu $v(t) = \frac{1}{1-t}$ bestimme die euklidische Bestapproximation in $\mathbb{P}_N$. Nach Kapitel 1 ist $P(t) = \sum_{k=0}^N x[k]t^k$ mit $Ax = b$ in $\mathbb{R}^{N+1}$

$$A = \left(\int_0^1 t^k t^j dt \right)_{k,j=0,\ldots,N} = \left(\frac{1}{1+k+j} \right)_{k,j=0,\ldots,N} \qquad (Hilbertmatrix)$$
$$b = \left(\int_0^1 v(t) t^k dt \right)_{k=0,\ldots,N}$$

A ist sehr schlecht konditioniert, aber das Problem selber ist gut konditioniert.

c) *Die Berechnung von $x^T y\,(x, y \in \mathbb{R}^N$ ist gut konditioniert, d.h. es existiert ein Algorithmus ohne Rundungsfehlerverstärkung.*

d) *Die Polynomauswertung mit dem Hornerschema*

$$\begin{aligned} P(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 \\ &= a_0 + t(a_1 + t(a_2 + t(a_3 + a_4 t))) \end{aligned}$$

ist stabil.

e) *Bei der Auswertung von Differenzenquotienten*

$$f\prime(t) \approx \frac{1}{h}\left(f(t+h) - f(t)\right)$$

ist Auslöschung unvermeidbar. Ein Kompromiss ist die Wahl $h = |x|\sqrt{\text{eps}}$. *Die Genauigkeit beträgt liegt dann im Bereich von* $\frac{|f(x)|}{|x}\sqrt{\text{eps}}$.

f) *Das Berechnen einer Orthonormalbasis mit dem Gram-Schmidt-Verfahren ist nicht stabil.*

Konditionszahlen

(2.14) Definition
Sei $f : \mathbb{R}^N \longrightarrow \mathbb{R}^K$ *differenzierbar,* $x, \delta x_k \in \mathbb{R}^N$.
Dann heißt $\kappa^{nk}_{abs}(x) = |\partial_n f_k(x)|$ absolute Konditionszahl, *und ist* $|f_k(x+\delta x) - f_k(x)|$ *der* absolute Fehler *von* f_k.
Falls $f_k(x) \neq 0$, *so heißt* $\kappa^{nk}_{rel} = |\partial_n f_k(x)|\frac{|x_n|}{|f_k(x)|}$ relative Konditionszahl, *und* $\frac{|f_k(x+\delta x)-f_k(x)|}{|f_k(x)|}$ *ist der* relative Fehler *von* f_k.

Eine Taylor-Entwicklung erster Ordnung liefert

$$f_k(x+\delta x) = f_k(x) + \sum_{n=1}^{N} \partial_n f_k(x)\delta x_n + o(\delta x)\,.$$

Daher gilt für den absoluten Fehler

$$|f_k(x+\delta x) - f_k(x)| \leq \sum_{n=1}^{N} \kappa^{nk}_{abs}(x)|\delta x_n| + o(\delta x)$$

und den *relativen Fehler*

$$\frac{|f_k(x+\delta x) - f_k(x)|}{|f_k(x)|} \leq \sum_{n=1}^{N} \kappa^{nk}_{rel}(x)\frac{|\delta x_n|}{|x_n|} + o(\delta x)\,.$$

Beispiel

a) $f(x_1,x_2) = x_1 + x_2$; *die Jacobi-Matrix ist* $J_f(x_1,x_2) = (1,1)$, *d.h. es gilt* $\kappa^{nk}_{abs} \equiv 1$ *und* $\kappa^{n1}_{rel}(x) = \frac{|x_n|}{|x_1+x_2|} \leq 1$ *für* $x_1, x_2 \geq 0$.

b) $f(x_1,x_2) = x_1 \cdot x_2$; *die Jacobi-Matrix ist* $J_f(x_1,x_2) = (x_2,x_1)$.
Es ist $\kappa^{11}_{rel}(x) = \frac{|x_1||x_2|}{|x_1x_2|} = 1$.
Addition und Multiplikation sind gut konditioniert.

c) $f(b) = A^{-1}b$; *es gilt* $\partial_n f_k(b) = e_k^T A^{-1} e_n$. *Die relative Konditionszahl ist durch die Kondition von A beschränkt:*

$$\kappa^{nk}_{rel}(b) = |e_k^T A^{-1} e_n|\frac{|b_n|}{|e_k^T A^{-1} b|} \leq \|A^{-1}\|\frac{|AA^{-1}b|}{|A^{-1}b|} \leq \|A^{-1}\|\|A\| = \kappa(A)$$

Stabilitätsanalyse

a) Ein Algorithmus ist *vorwärtsstabil*, wenn die einzelnen Schritte in

$$f = f_K \circ f_{K-1} \circ \ldots \circ f_2 \circ f_1$$

nicht wesentlich schlechter konditioniert sind als das Problem f.

b) Zu einem gestörten Ergebnis $\tilde{y} = \tilde{f}(x)$ suche $\tilde{x}$ mit $f(\tilde{x}) = \tilde{y}$ (exaktes Ergebnis mit gestörten Eingangsdaten).
Falls $\frac{|x-\tilde{x}|}{|x|}$ klein ist, heißt der Algorithmus *rückwärtsstabil*.

Satz von Wilkinson

Sei $PA \approx \tilde{L}\tilde{R}$ eine numerisch berechnete LR-Zerlegung. Dann gilt

$$\frac{\|PA - \tilde{L}\tilde{R}\|_\infty}{\|A\|_\infty} \leq 2N^3 r(A)\,\text{eps}$$

mit

$$r(A) = \frac{\max\{|a| \mid a \in \mathbb{R} \text{ tritt im Algorithmus auf}\}}{\max\{|A[n,k]|\}}$$

und

$r(A) \leq 2^{N-1}$	LR-Zerlegung mit Spaltenpivotsuche
$r(A) \leq 1$	Cholesky-Zerlegung
$r(A) \leq 2$	A tridiagonal oder strikt diagonaldominant

$PA \approx \tilde{L}\tilde{R}$ impliziert $A^{-1} \approx (P^T\tilde{L}\tilde{R})^{-1} = B$.