Aufgabe: Zu gegebener Matrix $A \in \mathbb{R}^{n \times n}$ und rechter Seite $b \in \mathbb{R}^n$ ist ein $x \in \mathbb{R}^n$ gesucht, so dass $Ax = b$ gilt.

Satz 0.0.1. Sei $n \in \mathbb{N}$, $A \in \mathbb{R}^{n \times n}$ mit $\det(A) \neq 0$. Dann existiert zu jedem $b \in \mathbb{R}^n$ genau ein $x \in \mathbb{R}^n$ so, dass $Ax = b$.

Beweis Lineare Abbildung zu Matrxi A ist bijektiv, wenn $\det(A) \neq 0$. Dann existiert auch die inverse Matrix A^{-1} und $x = A^{-1}b$ □

Berechnung der Lösung?

"Direkte Verfahren": berechne x (bis auf Rundungsfehler, Zahlendarstellung ...)

"Iterative verfahren": ausgehend von $x_0 \in \mathbb{R}^n$ berechne Folge $x_1, x_2, \ldots, \in \mathbb{R}^n$ und $x_i \to x(i \to \infty$ bzw. $x_1, \ldots, x_N \in \mathbb{R}^N$ mit $||x_N - x|| \leq TOL$

Jetzt:

0.1 Direkte Verfahren

Es gibt einige spezielle Matrizen, für die sich die Lösung einfach berechnen lässt:

Diagonalmatrizen:

$$A = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{NN} \end{pmatrix} \Rightarrow x_i = \frac{b_i}{a_{ii}} \quad a_{ii} \neq 0, \text{ wenn } \det(A) \neq 0, \text{ für alle } i = 1, \ldots, n$$

Dreiecksmatrizen

Definition 0.1.1. Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt rechte obere Dreiecksmatrix, wenn

$$\forall j < i : a_{ij} = 0, a_{ii} \neq 0, \ i = 1, \ldots, n$$

Bei einer oberen Dreiecksmatrix lässt sich das lineare Gleichungssystem einfach von unten nach oben auflösen:

$$x_{11} = \frac{b_n}{a_{nn}}, \ x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j>i} a_{ij} b_j \right), \quad i = n-1, \ldots, 1$$

Definition 0.1.2. Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt linke untere Dreiecksmatrix, wenn

$$\forall j > i : a_{ij} = 0, a_{ii} \neq 0, \ i = 1, \ldots, n$$

auflösen:

$$x_i = \frac{b_1}{a_{11}},\ x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j<i} a_{ij} b_j \right), \quad i = 2, \ldots, n$$

Produkte aus Dreiecksmatrizen z.B. $A = L \cdot R$

$$Ax = b \Leftrightarrow L \cdot \underbrace{Rx}_{y} = b \Leftrightarrow \begin{cases} Ly &= b \\ Rx &= y \end{cases}$$

Definition 0.1.3. Die Zerlegung einer Matrix $A \in \mathbb{R}^{n \times n}$ in ein Produkt $A = LR$ mit linker unterer Dreiecksmatrix L und oberer rechter Dreiecksmatrix R löst "LR-Zerlegung"

Bemerkung 0.1.4. Im englischen: "LU-decomposition", L: lower, U: upper

0.2 LR-Zerlegung einer Matrix

Bemerkung 0.2.1. Nicht jede reguläre Matrix A mit $\det(A) \neq 0$ besitzt eine LR- Zerlegung.

Beispiel 0.2.2. $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \neq LR$, da

$$\begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{pmatrix} \implies (LR)_{11} = l_{11} r_{11} = a_{11} = 0 \implies l_{11} = 0 \lor r_{11} = 0 \lightning$$

Bei Vertauschung der Zeilen funktioniert es aber

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Zu jeder regulären Matrix $A \in \mathbb{R}^{n \times n}$ mit $\det(A) \neq 0$ gibt es eine Permutationsmatrix P so, dass PA eine LR-Zerlegung besitzt.

Berechnung der LR-Zerlegung: Gauß Algorithmus
Wir nehmen zunächst an, dass alle Operationen durchgeführt werden können, dass alle auftretenden Diagonalelemente$\neq 0$ sind. Erster Schritt:

$$A^{(1)} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix} \to A^{(2)} \begin{pmatrix} a_{11} & \ldots & \ldots & a_{1n} \\ 0 & \tilde{a}_{22} & \ldots & \tilde{a}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & \ldots & \ldots & a_{nn} \end{pmatrix}$$

durch subtrahieren von Vielfachen der ersten Zeile von den anderen Zeilen.

Diese Operation lässt sich auch als Matrix-Produkt darstellen: $A^{(2)} = L_1 A^{(1)}$ mit linker unterer Dreiecksmatrix L_1

$$L_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ -\frac{a_{n1}}{a_{11}} & 0 & \cdots & 1 \end{pmatrix}$$

Weiter entsprechend

$$A^{(i)} = \begin{pmatrix} * & * & * & * \\ 0 & \ddots & * & * \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{pmatrix} \rightsquigarrow \begin{pmatrix} * & * & * \\ 0 & \ddots & * \\ 0 & 0 & * \end{pmatrix}$$

durch Matrix

$$L_i = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & i & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{-a_{i\tilde{+}1,i}}{a_{ii}} & \ddots & 0 & 0 & 0 \\ 0 & 0 & \vdots & 0 & 0 & \ddots & 0 \\ 0 & 0 & -\frac{\tilde{a}_{n,i}}{a_{ii}} & 0 & 0 & 0 & \ddots \end{pmatrix}$$

für $i = 2, \ldots, n-1$:

$$L_{n-1} \ldots L_i L_{i-1} \ldots L_1 A = R$$

Man zeigt leicht:

$$L_{n-1} \ldots L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ l_{2n} & \ldots & l_{n,n-1} & 1 \end{pmatrix}$$

und

$$(L_{n-1} \ldots L_1)^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -l_{21} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & 0 \\ -l_{n,1} & \ldots & l_{n,n-1} & 1 \end{pmatrix} =: L$$

Dementsprechend

$$(L_{n-1} \cdots L_1)^{-1} (L_{n-1} \cdots L_1) A = (L_{n-1} \cdots L_1)^{-1} R \implies A = LR$$

Tritt beim Algorithmus ein Diagonalelement $\tilde{a}_{jj} = 0$ auf, dann muss durch Zeilentausch von Zahlen j und k, $k > j$, ein Diagonalelement erzeugt werden, das $\neq 0$ ist. Falls $\det(A) \neq 0$, muss das immer möglich sein.
Numerisch ist es aus Konditions- und Stabilitätsgründen vorteilhaft, wenn alle $|l_{ij}| \leq 1$ sind für $j < 1$.
Tausch der Zeilen j und k kann auch durch multiplikation mit einer "Permutationsmatrix" $P = P_{ik}$ beschrieben werden

$$P_{jk} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & \cdots & 1 & & & \\ & & & \vdots & 1 & \vdots & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & \vdots & 1 & \vdots & & & \\ & & & 1 & \cdots & 0 & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}$$

Matrixbeispiel siehe Tafelbild " Tafel_20221212_4.jpg" unten rechts.

Der Gauß-Algorithmus inklusive Zeilentausch lässt sich durch ein Matrix- Produkt darstellen

$$L_{n-1}P_{n-1}\cdots L_2P_2L_1P_{1k_1}A = R$$

Man kann zeigen:

$$i < j : P_jL_i = \tilde{L}_iP_j$$

für unsere Matrizen L_i von oben, wobei in $\tilde{L}_i$ nur Einträge $l_{j,i}, l_{k,i}$ vertauscht sind gegenüber L_i, also

$$L_{n-1}P_{n-1}\cdots L_1P_1A = \underbrace{\left(\widetilde{L_{n-1}}\cdots\tilde{L_1}\right)}_{(\ )^{-1}=L}\underbrace{(P_{n-1}\cdots P_1)}_{P}A = R \implies PA = LR$$

Diese Matrizen L_i heißen "Frobenius-Matrizen"

Satz 0.2.3. Ist $A \in \mathbb{R}^{n\times n}$ mit $\det(A) \neq 0$, dann gibt es eine Permutationsmatrix P so, dass PA eine LR-Zerlegung besitzt.

$$PA = LR = \frac{1}{2}L \cdot 2R$$

ist die Diagonale von L $\begin{pmatrix} L_{11} & 0 & 0 \\ * & \ddots & 0 \\ * & * & L_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ * & \ddots & 0 \\ & * & 1 \end{pmatrix}$, dann die Zerlegung eindeutig. (für die Permutationsmatrix P)

Bemerkung 0.2.4. LR-Zerlegung kann auf dem Rechner sparsam gespeichert werden:

$$\begin{pmatrix} r_{11} & \dots & \dots & r_{1n} \\ P_{21} & \ddots & R & \vdots \\ \vdots & L & \ddots & \vdots \\ l_{n1} & & l_{n,n-1} & r_n \end{pmatrix}$$

Bemerkung 0.2.5. Matlab/Octave: lu(...)

Für manche Matrizen kann die LR-Zerlegung mit deutlich weniger Rechenoperationen berechent werden:

0.2.1 LR-Zerlegung von Bandmatrizen

Definition 0.2.6. Eine Matrix $A \in \mathbb{R}^{n\times n}$, $A = (a_{ij})_{i,j}$ heißt "Bandmatrix" vom Bandtyp (m_l, m_r), mit $0 \le m_l, m_r \le n-1$, wenn gilt

$$a_{jk} = 0 \text{ für } k < j - m_l \text{ oder } k > j + m_r$$

$1 + m_l + m_r$ heißt "Bandbreite" der Matrix.

Beispiel 0.2.7.

Typ $(\mathbf{0}, \mathbf{0})$: Diagonalmatrix,

Typ $(\mathbf{1}, \mathbf{1})$: Tridiagonalmatrix

Typ $(\mathbf{n-1}, \mathbf{0})$: linke untere Dreiecksmatrix,

Typ $(\mathbf{0}, \mathbf{n-1})$: rechte obere Dreiecksmatrix

!D Randwertproblem für gewöhnliche DGL 2. Ordnung

$$\begin{aligned} -u''(x) &= f(x) \qquad x \in (0,1) \\ u(0) &= u_0 \\ u(1) &= u_1 \end{aligned}$$

Bild

$$\begin{aligned} -u''(x) &\approx \frac{-u(x-h) + 2u(x) - u(x+h)}{h^2} \\ &\approx \frac{-u_{i-1} + 2u_i - u_{i+1}}{h^2} \end{aligned}$$

$\Rightarrow$Tridiagonalmatrix

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & 0 \\ -1 & \diagdown & \diagdown & \\ & \diagdown & \diagdown & -1 \\ 0 & & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} f_1 + u_0 \\ f_2 \\ \vdots \\ f_{n-1} \\ f_n + u_1 \end{pmatrix}$$

2D Randproblem für partielle DGL 2. Ordnung:

$$-\Delta u(x_i, y_i) \approx \frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} u$$

$$\begin{aligned} -\Delta u(x) &= f(x) \text{ in } \Omega = (0,1)^2 \\ u(x) &= g(x) \text{ auf (da ist ein komisches gespiegeltes C)}\Omega \end{aligned}$$

Satz 0.2.8. Sei $A \in \mathbb{R}^{n \times n}$ Bandmatrix vom Typ (m_l, m_r), die eine LR-Zerlegung (ohne Zeilentausch) erlaubt, dann sind die Faktoren L, R ebenfalls Bandmatrizen vom Typ $(m_l, 0)$ bzw $(0, m_r)$. Der Aufwand für die Berechnung der LR-Zerlegung ist dann

$$\frac{1}{3} n \cdot m_l \cdot m_r + \mathcal{O}\left(n \cdot (m_l + m_r)\right)$$

Beweis Nachrechnen 😂 □

Beispiel 0.2.9. Tridiagonalmatrix $(1,1)$: $\mathcal{O}(n)$ Operationen $\rightsquigarrow$ lösen eines LGS mit Tridiagonalmatrix

0.2.2 Cholesky-Zerlegung

Spezialfall für symmetrische positiv definite Matrizen.

Satz 0.2.10. Sei $A \in \mathbb{R}^{n \times n}$ eine Symmetrische und positiv definite Matrix. Dann besitzt A eine LR-Zerlegung (ohne Zeilentausch) mit <u>positiven</u> $\tilde{a}_{ii}$, $i = 1, \dots, n$.

Beweis In ersten Schritt der LR-Zerlegung: $a_{11} > 0$, da A positiv definit ist, dann $e_1^T A e_1 = a_{11} > 0,\ e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ Elimination der ersten Spalte:

$$\tilde{a}_{JK} = a_{JK} - \frac{a_{j1}}{a_{11}} \cdot a_{1K} = a_{KJ} - \frac{a_{k1}}{a_{11}} \cdot a_{1J} = \tilde{a}_{KJ} \Rightarrow \tilde{A}\text{ist Symmetrisch}$$

Ist $\tilde{A}$ positiv definit?
Sei

$$\tilde{x} = (\tilde{x}_2, \dots, \tilde{x}_n)^T \in \mathbb{R}^{n-1}$$

beliebig. Setze

$$x_1 := \frac{-1}{a_{11}} \cdot \sum_{K=2}^{n} a_{1K} x_K.$$

Dann ist $x_1, \dots, x_n)^T \in \mathbb{R}^n$.
Dann ist (A pos. def.).

$$0 < \underbrace{\sum_{J,K=1}^{n} a_{JK} x_J x_K}_{x^T A x}$$

$$= \sum_{J,K=1}^{n} a_{JK} x_J x_K + a_{11} x_1^2 + 2a_{11} \sum_{K=2}^{n} a_{1K} x_K + \underbrace{\frac{1}{a_{11}} \left(\sum_{K=2}^{n} a_{1K} x_K \right)^2 - \frac{1}{a_{11}} \left(\sum_{J,K=1}^{n} a_{1J} a_{1K} x_J x_K \right)}_{=0}$$

$$= \sum_{j,k=2}^{n} \left(a_{jk} - \frac{a_{k1}}{a_{11}} \cdot a_{ij} \right) x_j x_k + a_{11} \underbrace{\left(x_1 + \frac{1}{a_{11} \sum_{k=2}^{n} a_{1k} x_k} \right)^2}_{=0,\text{ Wahl von } x_1}$$

$$= \tilde{x}^T \tilde{A} \tilde{x}$$

Also ist $\tilde{A}$ positiv definit. □

Satz 0.2.11. Symmetrische positiv definite Matrizen $A \in \mathbb{R}^{n \times n}$ gestatten eine "Cholesky-Zerlegung"

$$A = LDL^T = \tilde{L}\tilde{L}^T$$

mit positiver Diagonalmatrix

$$D = \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & d_{nn} \end{pmatrix}, \quad d_{ii} > 0$$

bzw (unskalierter) linker unterer Dreiecksmatrix

$$\tilde{L} = LD^{\frac{1}{2}}, \quad D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{d_{11}} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & \sqrt{d_{nn}} \end{pmatrix}$$

Beweis $A = LR$ mit $L = \begin{pmatrix} 1 & 0 & 0 \\ * & \ddots & 0 \\ * & * & 1 \end{pmatrix}$ "skaliert", $R = \begin{pmatrix} r_{11} & * & * \\ 0 & \ddots & * \\ 0 & 0 & r_{nn} \end{pmatrix}$, $r_{ii} > 0$. Dann ist $R = D\tilde{R}$ mit

$$D = \begin{pmatrix} r_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & r_{nn} \end{pmatrix}, \quad \tilde{R} = \begin{pmatrix} 1 & & \frac{r_{ij}}{r_{ii}} \\ 0 & \ddots & \\ 0 & 0 & 1 \end{pmatrix}$$

Also ist

$$A = A^T = (LR)^T = (LD\tilde{R})^T = \tilde{R}DL^T$$

mit linker unterer Dreiecksmatrix $\tilde{R}^T$, sklaiert, und rechter oberer Dreiecksmatrix DL^T. Also

$$\tilde{R}^T = L, \quad DL^T = R$$

□

$\Rightarrow$ R muss gar nicht explizit berechnet werden, es genügt L und D zu kennen.
$\Rightarrow$ Rechenoperationen etwa nur halb so viele nötig, Speicherplatz ähnlich.

0.2.3 "Lösung" nicht regulärer Systeme

Jetzt muss die Matrix nicht quadratisch sein.

$$A \in \mathbb{R}^{m \times n}, \; b \in \mathbb{R}^m$$

gegeben $\Rightarrow$ lineares Gleichungssystem

$$Ax = b \text{ für } x \in \mathbb{R}^n$$

Lineare Algebra: Ist Rang $(A) =$ Rang $(A|b)$ dann ist das lineare Geichungssystem lösbar ($b \in$Span der Spalten von A), aber im Allgemeinen nicht eindeutig lösbar.
Ist Rang $(A) <$ Rang $(A|b)$, dann ist das lineare Gleichungssystem nicht klassisch lösbar. Man kann aber versuchen, den "Defekt" $d := Ax - b$ zu minimieren, z.B. bezüglich der euklidischen Norm mit der "Methode der kleinsten Fehlerquadrate"

$$||d||_2 := \sqrt{\sum_{i=1}^{n} d_i^2}$$

Satz 0.2.12 ("Least-squares-Lösung")**.** $A \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ gegeben. Dann existiert immer eine Lösung $\overline{x} \in \mathbb{R}^n$ mit kleinstem Fehlerquadrat

$$||A\overline{x} - b||_2 = \min_{x \in \mathbb{R}^n} ||Ax - b||_2$$

Dies ist äquivalent dazu, dass $\overline{x} \in \mathbb{R}^n$ eine Lösung der "Normalgleichung"

$$A^T A x = A^T b$$

ist. Ist Rang $(A) = n$, dann ist $\overline{x}$ eindeutig bestimmt, ansonsten ist mit $\overline{x}$ auch für jedes $y \in \ker(A)$ $\overline{x} + y$ eine Lösung, und dies beschreibt alle Lösungen.

Satz 0.2.13 (Least-Squares Lösungen)**.** Sei $A \in \mathbb{R}^{m \times n}, b \in R^m$. Dann existiert immer ein $\tilde{x} \in R^n$ mit

$$||A\tilde{x} - b||^2 = \min_{x \in \mathbb{R}^n} ||Ax - b||^2$$

und $\tilde{x}$ ist Lösung der Normalgleichung

$$A^T A \tilde{x} = A^T b$$

$\tilde{x}$ ist eindeutig, wenn Rang $(A) = n$

Beweis Sei $\tilde{x}$ Lösung von $A^T A \tilde{x} = A^T b$. Wir wissen aus

Analysis: Minimum einer stetigen Funktion,

$$\min_i |x_i| \to \infty \implies \underbrace{||Ax - b||_2^2}_{:=F(x)} \to \infty \quad \text{Außer } A = 0$$

also

$$\forall M > 0 \exists r_0 : F(x) \geq M \forall x : \min |x_i| \geq r_0$$

dementsprechend nimmt $F(x)$ auf der kompletten Menge $B_{r_0}(0)$ ihr Minimum an.

Lineare Algebra: Es gelten die orthogonalen Zerlegungen:

$$R^m = \mathrm{Im}(A) + \ker\left(A^T\right)$$

$$R^n = \mathrm{Im}(A^T) + \ker(A)$$

Damit gibt es zu $b \in \mathbb{R}^m$ genau ein $s \in \mathrm{Im}(A)$ und genau ein $r \in \ker\left(A^T\right)$, sodass $b = s + r$. Sei dann $\tilde{x} \in \mathbb{R}^n$ mit $A\tilde{x} = s$ Dann ist

$$A^T A \tilde{x} = A^T s = A^T (s + r) = A^T b$$

also ist $\tilde{x}$ eine Lösung der Normalgleichung

□

Ist $A^T A \in \mathbb{R}^{n \times m}$ regulär, dann auch positiv definit und symmetrisch. Dann können wir die Normalgleichungen mit Hilfe der Cholesky-Zerlegung von $A^T A$ lösen.

$$\min_x ||Ax - b||_2^2 \iff A^T A \tilde{x} = A^T b,\ A^T A = \tilde{L}\tilde{L}^T \rightsquigarrow \tilde{L}\tilde{L}^T \tilde{x} = A^T b$$

Eine besser konditionierte Lösungsmethode:

Bemerkung 0.2.14 (QR Zerlegung von A bzw A^TA)**.** Dabei ist Q eine orthogonale Matrix, also bilden Zeilen bzw. Spalten eine Orthonormalbasis, R rechte obere Dreiecksmatrix. Dementsprechend ist

$$||Qx||_2 = ||x||_2 \, \forall x \in R^n$$

Für $A \in \mathbb{R}^{m\times n}, \ m > n$ gilt

$$QA = \begin{pmatrix} R \\ \underbrace{0}_{m-n} \end{pmatrix}$$

Um nun

$$QAx = Qb \iff \begin{pmatrix} R \\ 0 \end{pmatrix} x = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

zu lösen, reicht es

$$Rx = b_1$$

zu lösen, mit $R \in \mathbb{R}^{n\times m}, \ x, b_1 \in \mathbb{R}^n$ Dann ist

$$||Ax - b||_2^2 = ||Q(Ax - b)||_2^2 = ||Rx_1 - b_1||_2^2 + ||b_2||_2^2$$

und es wird minimiert, wenn

$$Rx_1 = b_1$$

Ausgleichsrechnung Bei der Interpolation wollten wir eine genaue Beschreibung von gegebenen Daten durch Funktion (Polynom, Spline,...) von evtl hohen Polynomgrad. Jetzt suchen wir eine möglichst gute Approximation der Daten mit einfachen Funktionen (kleiner Grad z.B.). Zu gegebenen Funktionen $u_1, \ldots, u_n$ und Daten $(x_1, y_1), \ldots, (x_m, y_m)$ mit $m > n$ ist eine Linearkombination

$$u(x) = \sum_{j=1}^{m} c_j \cdot u_j(x)$$

gesucht, sodass

$$\sum_{j=1}^{m} (u(x_i) - y_i)^2$$

minimal ist. Das lässt sich als Gleichungssystem bzw. Normalgleichung schreiben

$$A = (a_{ij}), \ a_{ij} = u_j(x_i), \ \underbrace{\begin{pmatrix} u_1(x_1) & \ldots & u_n(x_1) \\ \vdots & \vdots & \vdots \\ u_1(x_m) & \ldots & u_n(x_m) \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{b}, \ \min_x ||Ax - b||_2^2$$

Die optimalen Koeffizienten c_i lösen gerade die Normalgleichungen $A^TA\tilde{x} = A^Tb$

Beispiel 0.2.15. Polynome $p \in \mathbb{P}_{n-1}$, Basisfunktionen sind die Monome $1, x, x^2, x^{n-1}$. Es ist

$$\underbrace{\begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{pmatrix}}_{A} \begin{pmatrix} c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots y_n \end{pmatrix}$$

löse

$$\underbrace{A^T A}_{\in \mathbb{R}^{n \times m}} \tilde{x} = A^T b$$

Für höhere Polynomgrade ist die Vandermonde Matrix immer schlechter konditioniert, deshalb besser QR-Zerlegung benutzen.

0.2.4 Kondition und Stabilität von linearen Gleichungssystemen

Wir betrachten die Mathematische Aufgabe $(A, b) \mapsto x$, die Lösung von $Ax = b$. Diese ist schlecht konditioniert, wenn kleine Störungen in den Daten relativ große Fehler versuchachen.

Frage. Wie misst man die Störungen in den Daten?

Wir betrachten die Vektor- und Matrixnormen

$$||A + \delta A|| \, , \; ||b + \delta b||$$

Vektornormen

- $||x||_1 := \sum_{i=1}^{n} |x_i|$
- $||x||_2 := \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}$
- $||x||_\infty := \max_{i=1,\dots,n} |x_i|$

<u>induzierte</u> **Matrixnormen**

- $||A||_1 := \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|$ max. Spaltensumme
- $||A||_2 := \sqrt{\zeta(A^T A)} = \max \left\{ \sqrt{\lambda} : \lambda \text{Eigenwert von } A^T A \right\}$. Ist A symmetrisch, dann ist $||A||_2 = \zeta(A)$
- $||A||_\infty := \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ji}|$ max. Zeilensumme

Definition 0.2.16. Die Konditionszahl einer regulären Matrix A in einer Matrixnorm ist

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

Ist A nicht invertierbar, dann $\kappa(A) := \infty$

Bemerkung 0.2.17. Bezüglich der 2-Norm

$$\kappa(A) = ||A||_2 \left|\left|A^{-1}\right|\right|_2 = \frac{\lambda_{\max}}{\lambda_{\min}}$$

wobei $\lambda_{\max}, \lambda_{\min}$ die betragsmäßig größten/kleinsten Eigenwerte von A sind, falls A symmetrisch ist.

Beispiel 0.2.18.

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 \\ -1 & \ddots & \cdots \\ \ddots & \ddots & \ddots \end{pmatrix}$$

aus Diskretisierung von $-n''(x) = f(x)$ hat Eigenwerte $h^2 \cdot k^2\pi^2,\ k = 1, \dots, n$ also

$$\lambda_{\min} = h^2\pi^2,\ \lambda_{\max} = h^2 n^2 \pi^2$$

Dementsprechend

$$\kappa(A) = \mathcal{O}\left(n^2\right) 0 \mathcal{O}\left(\frac{1}{h^2}\right)$$

$$Ax = b \implies (A + \delta A)(x + \delta x) = (b + \delta b)$$

zur Lösbarkeit des gestörten Systems

Lemma 0.2.19. Eine Matrix $B \in \mathbb{R}^{n\times m}$ habe Norm $||B|| < 1$. Dann ist $I + B$ regulär und es gilt

$$\left|\left|(I+B)^{-1}\right|\right| \leq \frac{1}{1 - ||B||}$$

Beweis

$$\forall x \in \mathbb{R}^n : x = Ix = (I+B)x - Bx$$

also

$$||x|| \leq ||(I+B)x|| + ||Bx||$$

und

$$\forall x \neq 0 : ||(I+B)x|| \geq ||x|| - ||Bx|| \geq ||x|| - ||B|| \cdot ||x|| = \underbrace{(1 - ||B||)}_{>0} ||x|| > 0$$

Also ist $\ker(I+B) = \{0\}$, $I+B$ regulär. Weiter ist

$$\begin{aligned} 1 = ||I|| &= \left|\left|(I+B)(I+B)^{-1}\right|\right| \\ &= \left|\left|(I+B)^{-1} + B(I+B)^{-1}\right|\right| \\ &\geq \left|\left|(I+B)^{-1}\right|\right| - ||B||\,||(I+B)||^{-1} \\ &= (1 - ||B||)\left|\left|(I+B)^{-1}\right|\right| \end{aligned}$$

□

Damit können wir zeigen

Satz 0.2.20 (Störungssatz). $A \in \mathbb{R}^{n\times n}$ regulär und $\delta A \in \mathbb{R}^{n\times n}$ mit

$$||\delta A|| \leq \frac{1}{||A^{-1}||}$$

Damit ist die Matrix $A+\delta A$ ebenfalls regulär und für den relativen Fehler der Lösung δx gilt:

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

Beweis

$$\left|\left|A^{-1}\cdot\delta A\right|\right| \leq \left|\left|A^{-1}\right|\right|\,||\delta A|| \leq 1$$

Damit ist

$$A + \delta A = A(I + A^{-1}\delta A)$$

regulär nach dem Lemma vorher. Für die Lösung $x+\delta x$ gilt

$$(A+\delta A)(x+\delta x) = b+\delta b \Leftrightarrow (A+\delta A)\delta x = b + \delta b - (A+\delta A)x = \delta b - \delta A x$$

Also

$$\begin{aligned} ||\delta x|| &\leq \left|\left|(A+\delta A)^{-1}\right|\right|(||\delta b|| + ||\delta A||\,||x||) \\ &= \left|\left|(A(I+A^{-1}\delta A))^{-1}\right|\right| \\ &= \left|\left|(I+A^{-1}\delta A)^{-1}A^{-1}\right|\right| \\ &\leq \left|\left|(I+A^{-1}\delta A)^{-1}\right|\right|\cdot\left|\left|A^{-1}\right|\right|(||\delta b|| + ||\delta A||\,||x||) \\ &\leq \frac{1}{1-||A^{-1}\delta A||}\left|\left|A^{-1}\right|\right|(||\delta b|| + ||\delta A||\,||x||) \\ &\leq \frac{1}{1-||A^{-1}||\,||\delta A||\frac{||A||}{||A||}}\cdot\frac{||A||\left|\left|A^{-1}\right|\right|}{||A||}\,||x||\left(\frac{||\delta b||}{||x||} + ||\delta A||\right) \\ &\leq \frac{1}{1-\kappa(A)\cdot\frac{||\delta A||}{A}}\cdot\kappa(A)\cdot\left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||}\right)\cdot||x|| \end{aligned}$$

□

Bemerkung 0.2.21. Ist $\kappa(A)\frac{||\delta A||}{||A||} \leq 1$, dann gilt im Wesentlichen

$$\frac{||\delta x||}{||x||} \leq \kappa(A) \cdot \left(\frac{||\delta b||}{||b||} + \frac{||\delta A||}{||A||}\right)$$

das heißt Fehler in den Daten werden im Wesentlichen durch die <u>Kondition der Matrix</u> verstärkt. Dies kann groß sein:

$$\text{Kond}\left[\begin{pmatrix} 2 & -1 & 0 \\ -1 & \ddots & \ddots \\ 0 & \ddots & \ddots \end{pmatrix}\right] = \mathcal{O}\left(\frac{1}{h^2}\right) = \mathcal{O}\left(n^2\right)$$

Was gilt zusätzlich für die Stabilität der Gauß-Elimination?

Satz 0.2.22 (Rundungsfehlereinfluss)**.** $A \in \mathbb{R}^{n\times n}$ regulär, das Gleichungssystem $Ax = b$ werde durch Gauß-Elimination mit Spaltenpriorisierung gelöst. Dann ist die unter Einfluss von Rundungsfehlernberechnete Lösung $x + \delta x$ die exakte Lösung eines gelösten Problems $(A - \delta A)(x + \delta x) = b$ mit

$$\frac{||\delta A||_\infty}{||A||_\infty} \leq 1.01 \cdot 2^{n-1} \cdot \left(n^3 + 2n^2\right) \cdot \text{eps}_{=\text{Maschinengenauigkeit}}$$

(Ohne Beweis)

Mit obigem Satz folgt:

$$\frac{||\delta x||_\infty}{||x||_\infty} \leq \frac{\kappa(A)}{1 - \kappa(A)\frac{||\delta A||}{||A||}}$$

Der Faktor 2^{n-1} ist eine obere Abschätzung für allgemeine reguläre Matrizen, hat etwas mit dem größten auftretenden Pivot-Element zu tun.
Es gibt allerdings Matrizen, die diesen Faktor realisieren:

$$\text{Bsp.: Wilkinson}\begin{pmatrix} 1 & & 0 & & 1 \\ -1 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ -1 & \cdots & \cdots & -1 & 1 \end{pmatrix}$$

(Wikipedia zeigt eine andere Matrix als Schmidt)
Für spezielle Matrizen ist der Faktor deutich kleiner:

n bei der oberen Hessenbergmatrix

2 Matrix strikt diagonaldominant

2 Matrix tridiagonal

1 Matrix Symmetrisch positiv definit

$n^{\frac{2}{3}}$ Erwartungswert für statistische Matrix

Rundungsfehler Einfluss dann mit der LR-Zerlegung also stark verstärkt werden.

$$A \to L_1 A \to L_2 L_1 A \to \cdots \to L_{n-1} \cdots L_1 A = R$$

und jedes L_i kann die Rundungsfehler verstärken. bessere Stabilität ist zu erwarten, wenn man beschränkte Transformationen nutzt, z.B. ortogonale.
$\rightsquigarrow$ QR-Zerlegung, Transformation mit orthogonalen Matrizen: $A = QR$.
Orthogonale Matrix, R rechte obere Dreiecksmatrix.
Orthogonal: $||Qx|| = ||x||$, $Q^{-1} = Q^T$ ebenfalls orthogonal: Verfahren generiert $Q^T A = R$

Q^T ist wieder Produkt einfacher orthogonaler Matrizen: $Q^T = Q_{n-1} \cdots Q_1$

Q_1 bringt erste Spalte von $Q_1 A$ auf $\begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ Gestalt, etc.

Beispiel 0.2.23 (2D)**.** $\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \to \begin{pmatrix} \alpha \\ 0 \end{pmatrix}$ durch orthogonale Transformation
(Bild)
geht zwar mit einfachen geometrischen Operationen:

- Drehung um (0,0) um Winkel-β
  Multiplikation mit Matrix $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$, $c^2+s^2 = 1,\ c = \cos((-\beta),\ s = \sin((-\beta)))$
- Spiegelung an Ebene/Gerade:
  Matrix $I - 2\frac{vv^T}{v^T v}$ Spiegelung an Ebene $\perp \checkmark$

Beide Ideen werden in numerischen Verfahren genutzt:

- QR-Zerlegung mittels "girens-Rotationen":
  nutze Rotationsmatritzen um nacheinander-

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & \cdots \\ \vdots & & & \\ \vdots & & & \\ a_{n-1,1} & \cdots & \cdots & \cdots \\ a_{n,1} & \cdots & \cdots & \cdots \end{pmatrix} \to \mathrm{Q}_{n,n} A = \begin{pmatrix} a_{11} & \cdots & \cdots & \cdots \\ \vdots & & & \\ a_{n-1,1} & & & \\ \alpha & * & \cdots & * \\ 0 & * & \cdots & * \end{pmatrix};$$

$$\mathrm{Q}_{n1} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & c & s \\ & & -s & c \end{pmatrix}; \mathrm{Q}_{n2}\mathrm{Q}nnA = \begin{pmatrix} * & & \\ \vdots & & \\ * & * & \\ 0 & & \\ 0 & & \end{pmatrix} \cdots$$

Eine 2×2-Rotation für jedes Matrixelement unterhalb der Diagonalen.

Aufwand: etwa 4-mal so groß wie bei der LR-Zerlegung aber stabiler und auch anwendbar bei überbestimmten Gleichungssystemen, (statt Normalgleichungen) $QA = \begin{pmatrix} R \\ 0 \end{pmatrix} = Qb$

- QR-Zerlegung mittels "Householder-Reflesionen" 1. Spiegelung die $\begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \end{pmatrix}$ auf $\begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ spiegelt mit $Q_1 = I - 2\frac{vv^T}{v^Tv}$, v relativer Fehler einfach Berechenbar. Spiegelungsmatrizen Q_i symmetrische orthogonale Matrizen.

Aufwand: insgesamt etwa doppelt so groß wie bei der LR-Zerlegung, aber stabil und auchReflesionen anwendbar bei überbestimmten Gleichungssystemen.