Kapitel 2

Lösung linearer Gleichungssysteme I: Direkte Verfahren

Prolog: Wiederholung aus der Linearen Algebra

Einige Fakten über lineare Gleichungssysteme und Matrizen werden kurz wiederholt. Es seien

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \in \mathbb{K}^{m\times n}, \qquad b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \in \mathbb{K}^m$$

gegeben. Man beachte, dass A eine rechteckige Matrix ist, die Anzahl der Zeilen und Spalten muss nicht übereinstimmen. Gesucht sind alle Vektoren

$$x = (x_k)_{k=1,\dots,n} \in \mathbb{K}^n \qquad \text{mit} \qquad Ax = b.$$

Der entscheidende Begriff ist der Rang von Matrizen. Für $r = \mathrm{Rang}(A)$ gibt es die folgenden gleichwertigen Definitionen:

- r ist die maximale Anzahl linear unabhängiger Spalten von A,
- r ist die maximale Anzahl linear unabhängiger Zeilen von A,
- r ist die Dimension des Bildes der linearen Abbildung $f_A : \mathbb{K}^n \to \mathbb{K}^m$, $f(x) = Ax$.

Zusätzlich verwenden wir den Rang der erweiterten Matrix

$$\text{Rang}\left([A,b]\right) = \text{Rang}\left[\begin{array}{ccc|c} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{m1} & \dots & a_{mn} & b_m \end{array}\right].$$

Satz:

Das lineare Gleichungssystem $Ax = b$ besitzt genau dann mindestens eine Lösung $x \in \mathbb{K}^n$, wenn $\text{Rang}(A) = \text{Rang}\left([A,b]\right)$ gilt.
Im Fall $m = n$ (also für quadratische Matrizen) sind die folgenden Aussagen äquivalent:

(i) $Ax = b$ ist für jedes $b \in \mathbb{K}^n$ eindeutig lösbar.

(ii) Das homogene Gleichungssystem $Ax = 0$ hat als eindeutige Lösung $x = 0$.

(iii) $\text{Rang}(A) = n$.

(iv) $\det(A) \neq 0$.

(v) A ist regulär; d.h. es existiert die eindeutige Inverse $A^{-1} \in \mathbb{K}^{n\times n}$ mit $A^{-1}A = I_n$, wobei I_n die $n \times n$-Einheitsmatrix ist.

(vi) Alle Eigenwerte von A sind ungleich Null.

Zum letzten Punkt (vi) wiederholen wir die Definitionen von Eigenwerten und Eigenvektoren sowie die Jordan-Normalform:

- Die Eigenwerte $\lambda_j \in \mathbb{C}$, $j = 1, \dots, n$, einer Matrix $A \in \mathbb{C}^{n\times n}$ sind die Nullstellen des charakteristischen Polynoms

  $$p(\lambda) = \det(A - \lambda I_n),$$

  die Vielfachheit r_j einer Nullstelle λ_j ist die *algebraische Vielfachheit* dieses Eigenwerts. Zu jedem Eigenwert λ_j wird der *Eigenraum* definiert als

  $$\text{Eig}(A, \lambda_j) = \{v \in \mathbb{C}^n : Av = \lambda_j v\};$$

  seine Dimension $1 \leq s_j \leq r_j$ ist die *geometrische Vielfachheit* von λ_j.

- Falls für alle Eigenwerte $s_j = r_j$ gilt, so ist A *diagonalisierbar*: Die Matrix $S \in \mathbb{C}^{n\times n}$, deren Spalten eine Basis von Eigenvektoren von A sind, ergibt

  $$S^{-1}AS = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix},$$

  auf der rechten Seite steht die Diagonalmatrix der Eigenwerte (in passender Reihenfolge zu den Spalten von S).

- Falls für einige Eigenwerte $s_j < r_j$ gilt, so berechnet man weitere sog. "Hauptvektoren" (oder "verallgemeinerte Eigenvektoren") von A und erhält eine reguläre Matrix $S \in \mathbb{C}^{n\times n}$ mit

$$S^{-1}AS = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} + N,$$

wobei $N \in \mathbb{C}^{n\times n}$ überall Nullen enthält mit Ausnahme der oberen Nebendiagonalen, in der Nullen oder Einsen stehen. Dies ist die *Jordan-Normalform* von A.

Beispiel: $J = \begin{bmatrix} 3 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$ ist die Jordan-Normalform einer Matrix A mit doppeltem Eigenwert $\lambda_1 = \lambda_2 = 3$ der geometrischen Vielfachheit 1 und einfachem Eigenwert $\lambda_3 = 5$.

Eine wichtige Klasse von Matrizen sind die reell-symmetrischen und die hernmiteschen Matrizen.

1. $A \in \mathbb{C}^{n\times n}$ ist *hermitesch*, wenn $A = A^*$ gilt, d.h. $a_{jk} = \overline{a_{kj}}$ für alle $j, k = 1, \ldots, n$. (Reelle Matrizen mit dieser Eigenschaft sind *symmetrisch*.) Die Notation A^* bezeichnet die transponierte und komplex konjugierte Matrix.

2. Hermitesche Matrizen haben nur **reelle** Eigenwerte. Sie sind diagonalisierbar mit einer **unitären Matrix** $S \in \mathbb{C}^{n\times n}$, $S^{-1} = S^*$, deren Spalten eine Orthonormalbasis des $\mathbb{C}^n$ bilden. (Für reell-symmetrisches A ist diese Matrix S eine reelle Orthogonalmatrix und $S^{-1} = S^T$.)

2.1 Störungsanalyse

Wir behandeln zunächst den Fall quadratischer linearer Gleichungssysteme ($n = m$). Das mathematische Problem lautet:

Zu gegebener regulärer Matrix $A \in \mathbb{K}^{n\times n}$ und $b \in \mathbb{K}^n$
finde die Lösung von $Ax = b$.

Für die Betrachtung der natürlichen Stabilität (=Kondition) dieses Problems benötigen wir spezielle Normen auf dem Vektorraum $\mathbb{K}^{n\times n}$ der $n \times n$-Matrizen. (Der Vektorraum $K^{n\times n}$ hat die Dimension n^2. Er ist endlich-dimensional, also sind alle Normen äquivalent.)

2.1.1 Definition (Matrixnorm):

a) Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$ heißt *Matrixnorm*, wenn sie *submultiplikativ* ist, d.h.

(MN1) $\|A \cdot B\| \le \|A\| \cdot \|B\|$ für alle $A, B \in \mathbb{K}^{n\times n}$.

b) Eine Norm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$ heißt *verträglich* mit der (Vektor-)Norm $\|\cdot\|'$ auf $\mathbb{K}^n$, wenn

(MN2) $\|Ax\|' \le \|A\| \cdot \|x\|'$ für alle $A \in \mathbb{K}^{n\times n}$, $x \in \mathbb{K}^n$ gilt.

2.1.2 Die vier wichtigsten Matrixnormen:

- *Frobeniusnorm* (auch Schur-Norm genannt)

$$\|A\|_F = \left(\sum_{j,k=1}^{n} |a_{j,k}|^2\right)^{1/2} = (\mathrm{Spur}(A^*A))^{1/2},$$

  wobei $\mathrm{Spur}(A^*A)$ die Summe der Diagonalelemente von A^*A bezeichnet; hier wird also eine euklidische Norm über alle Einträge der Matrix A gebildet.

- *Zeilensummennorm*

$$\|A\|_\infty = \max_{1\le j\le n} \sum_{k=1}^{n} |a_{j,k}|,$$

- *Spaltensummennorm*

$$\|A\|_1 = \max_{1\le k\le n} \sum_{j=1}^{n} |a_{j,k}|,$$

- *Spektralnorm*

$$\|A\|_2 := \max\{\sqrt{\lambda} : \lambda \text{ ist Eigenwert von } A^*A\}.$$

Warnung: Der Spektralradius

$$\varrho(A) = \max\{|\lambda| : \lambda \text{ ist Eigenwert von } A\}.$$

ist im Fall $n \geq 2$ keine Norm; insbesondere das Axiom (N1) in 1.2.1 ist verletzt: jede strikte obere Dreiecksmatrix hat den Spektralradius 0.

Woher kommen die gerade eingeführten Bezeichnungen mit dem Index $p = 1, 2, \infty$?

2.1.3 Definition (natürliche Matrixnorm):
Zu einer gegebenen (Vektor-)Norm $\|\cdot\|$ auf $\mathbb{K}^n$ definieren wir

$$\|A\|_{\text{op}} := \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x \in \mathbb{K}^n,\ \|x\|=1} \|Ax\|.$$

Die Norm $\|\cdot\|_{\text{op}}$ heißt die *natürliche Matrixnorm* (oder *Operatornorm*) zur gegebenen Vektor-Norm $\|\cdot\|$ auf $\mathbb{K}^n$.

Hierbei handelt es sich wirklich um eine Matrixnorm, denn:

2.1.4 Hilfssatz:
Die natürliche Matrixnorm $\|\cdot\|_{\text{op}}$ ist submultiplikativ und verträglich mit der gegebenen Vektor-Norm $\|\cdot\|$ auf $\mathbb{K}^n$.

Speziell gilt $\|I_n\|_{\text{op}} = 1$ für die Einheitsmatrix $I_n \in \mathbb{K}^{n \times n}$.

Nun können wir die Schreibweise mit den Indizes der Matrixnormen in 2.1.2 erklären.

2.1.5 Hilfssatz:
Die natürliche Matrixorm

a) zur Maximums-Norm $\|\cdot\|_\infty$ auf $\mathbb{K}^n$ ist die Zeilensummennorm,

b) zur Vektor-Norm $\|\cdot\|_1$ auf $\mathbb{K}^n$ ist die Spaltensummennorm,

c) zur euklidischen Norm $\|\cdot\|_2$ auf $\mathbb{K}^n$ ist die Spektralnorm.

2.1.6 Definition (Konditionszahl):
Gegeben sei eine Matrixnorm $\|\cdot\|$ auf $\mathbb{K}^{n\times n}$. Für eine reguläre Matrix $A \in \mathbb{K}^{n\times n}$ heißt

$$\mathrm{cond}(A) = \mathrm{cond}_{\|\cdot\|}(A) = \|A\| \cdot \|A^{-1}\|$$

die *Konditionszahl* von A. Für die p-Normen in 2.1.2, $p = 1, 2, \infty$, schreiben wir auch $\mathrm{cond}_p(A)$.

Die Konditionszahl von A hängt laut ihrer Definition von der verwendeten Matrixnorm ab. Die Äquivalenz der Normen besagt aber, dass die Größenordnung der Konditionszahlen von A bzgl. verschiedener Normen vergleichbar ist.

2.1.7 Beispiel: Die Matrix in der Einleitung $A = \begin{bmatrix} 2.0001 & 1.9999 \\ 1.9999 & 2.0001 \end{bmatrix}$ hat die Inverse

$$A^{-1} = \begin{bmatrix} 2500.125 & -2499.875 \\ -2499.875 & 2500.125 \end{bmatrix}.$$

Die Zeilen- und Spaltensummennormen stimmen überein, weil A und A^{-1} symmetrisch sind. Es ist

$$\mathrm{cond}_1(A) = \mathrm{cond}_\infty(A) = 4 \cdot 5000 = 20000.$$

Die Eigenwerte von A sind $\lambda_1 = 4$ und $\lambda_2 = 0.0002$; am Ende des Abschnitts begründen wir die einfache Formel

$$\mathrm{cond}_2(A) = \frac{\lambda_1}{\lambda_2} = 20000,$$

wiederum weil A symmetrisch ist.

Beispiele mit unterschiedlicher Konditionszahl je nach p-Norm lassen sich leicht mit Hilfe von Zufallsmatrizen finden: `A=randn(3)` liefert in Matlab eine reelle 3×3-Matrix mit höchstwahrscheinlich unterschiedlichen Konditionszahlen für $p = 1, 2, \infty$, wie man mit dem Befehl `cond(A,p)` testet.

2.1.8 Proposition: Eigenschaften der Konditionszahl
Es sei $\|\cdot\|$ eine natürliche Matrixnorm auf $\mathbb{K}^{n\times n}$. Dann erfüllt die zugehörige Konditionszahl

$$\mathrm{cond}(I_n) = 1, \qquad \mathrm{cond}(A) \geq 1, \qquad \mathrm{cond}(\alpha A) = \mathrm{cond}(A)$$

für beliebiges reguläres A und $\alpha \in \mathbb{K} \setminus \{0\}$. Speziell für die Spektralnorm gilt:

$\mathrm{cond}_2(A) = 1 \iff \alpha A$ ist unitär ($\mathbb{K} = \mathbb{C}$) bzw. orthogonal ($\mathbb{K} = \mathbb{R}$) für ein $\alpha > 0$

Wie bereits allgemein in Kapitel 1 betrachtet, beschreibt die Konditionszahl den Verstärkungsfaktor des relativen Fehlers bei gestörten Eingabedaten. Wir betrachten zuerst den einfachen Fall, dass nur die Daten der rechten Seite $b \in \mathbb{K}^n$ des Gleichungssystems gestört werden, aber die Matrixeinträge von A exakt sind. Im folgenden seien die verwendeten Vektor- und Matrixnormen verträglich im Sinne von Definition 2.1.1.

2.1.9 Störungssatz (einfache Form):
Die Matrix $A \in \mathbb{K}^{n\times n}$ sei regulär, es sei $b \in \mathbb{K}^n \setminus \{0\}$ sowie $\widetilde{b} = b + \Delta b$ eine gestörte rechte Seite. Die Vektoren $x, \widetilde{x} \in \mathbb{K}^n$ seien die Lösungen von

$$Ax = b \qquad \text{bzw.} \qquad A\widetilde{x} = \widetilde{b}.$$

Dann gilt für den relativen Fehler

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \operatorname{cond}(A) \frac{\|\widetilde{b} - b\|}{\|b\|}.$$

Der *Beweis* erfolgt mit einfacher Rechnung (ohne Differentialrechnung):

$$\frac{\|\widetilde{x} - x\|}{\|x\|} = \frac{\|A^{-1}(\widetilde{b} - b)\|}{\|A^{-1}b\|} = \frac{\|A\|\,\|A^{-1}(\widetilde{b} - b)\|}{\|A\|\,\|A^{-1}b\|} \leq \|A\|\,\|A^{-1}\|\,\frac{\|\widetilde{b} - b\|}{\|b\|}.$$

Im letzten Schritt wurde im Zähler und Nenner die Submultiplikativität verwendet, im Nenner in der Form $\|A\|\,\|A^{-1}b\| \geq \|AA^{-1}b\| = \|b\|$.

Im Beispiel der Einleitung hatten wir die 2×2-Matrix A mit $\operatorname{cond}_\infty(A) = 20000$, der relative Fehler der speziellen rechten Seite war 0.0001 und der relative Fehler der Lösungen war $2 = 200\,\%$. Hier beobachtet man also schon einen Fall, in dem die Fehlerverstärkung tatsächlich so groß ist wie die Konditionszahl. Generell beschreibt die Konditionszahl den "worst case" der Fehlerverstärkung, für spezielle Störungen kann die Verstärkung auch geringer ausfallen. Eine genauere Untersuchung findet in den Übungen statt.

Für den allgemeinen Fall, dass sowohl die Daten der rechten Seite b als auch die Matrixeinträge von A gestört werden, beweisen wir zunächst eine Aussage über die Regularität von $A + \Delta A$. Kurzgefasst besagt das nachfolgende Korollar, dass kleine Störungen wieder eine reguläre Matrix ergeben.

2.1.10 Hilfssatz:
Es sei $\|\cdot\|$ eine Matrixnorm, also submultiplikativ, und $B \in \mathbb{K}^{n\times n}$ erfülle $\|B\| < 1$. Dann ist die Matrix $I_n - B$ regulär und es gilt

$$\|(I_n - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

Im Hauptteil des Beweises verwenden wir ein Hilfsmittel der Funktionalanalysis, die sog. *Neumann-Reihe*. Sie stellt eine Verallgemeinerung der geometrischen Reihe in $\mathbb{C}$ (oder $\mathbb{R}$) auf den Vektorraum der $n \times n$-Matrizen dar und ist auch in viel allgemeinerem Zusammenhang wichtig.

1. Zuerst halten wir fest, dass die Potenzen B^k im normierten Vektorraum $\mathbb{K}^{n\times n}$ gegen die Nullmatrix konvergieren, denn die Submultiplikativität ergibt

$$\|B^k\| \leq \|B\|^k \to 0 \quad \text{für} \quad k \to \infty.$$

2. Mit Hilfe der Dreiecksungleichung folgt weiter, dass die Reihe

$$\sum_{k=0}^{\infty} B^k \qquad \text{(Neumann-Reihe zu } B\text{)}$$

konvergiert. Denn ihre Partialsummen bilden eine Cauchy-Folge, wie man mit

$$\left\| \sum_{k=m}^{s} B^k \right\| \overset{(N3)}{\leq} \sum_{k=m}^{s} \left\| B^k \right\| \leq \sum_{k=m}^{s} \|B\|^k \leq \frac{\|B\|^m}{1-\|B\|}$$

für $s \geq m$ sieht. Der normierte Vektorraum der $n \times n$-Matrizen ist vollständig (weil endlich-dimensional), also existiert der Grenzwert

$$C = \lim_{m\to\infty} \sum_{k=0}^{m} B^k.$$

3. Multiplikation der Partialsumme bis m mit $(I_n - B)$ ergibt (genau wie im Reellen)

$$(I_n - B) \sum_{k=0}^{m} B^k = I_n - B^{m+1} \to I_n \quad \text{für} \quad m \to \infty.$$

Die Konvergenz im letzten Schritt wurde schon oben gezeigt. Also gilt für den Grenzwert C der Neumann-Reihe

$$(I_n - B)C = I_n,$$

mit anderen Worten, C ist die Inverse von $I_n - B$. Nun ist alles gezeigt: $I_n - B$ ist regulär und

$$\|C\| \leq \sum_{k=0}^{\infty} \|B\|^k = \frac{1}{1-\|B\|}.$$

2.1.11 Korollar:
Es seien $A, B \in \mathbb{K}^{n\times n}$. Die Matrix A sei regulär und die Matrix B erfülle $\|B\| < \frac{1}{\|A^{-1}\|}$. Dann ist auch $A + B$ regulär und

$$\|(A+B)^{-1}\| \leq \frac{\|A^{-1}\|}{1-\|A^{-1}B\|} \leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|B\|}.$$

Beweis: Hilfssatz 2.1.10 für $H = -A^{-1}B$ mit $\|H\| \leq \|A^{-1}\|\,\|B\| < 1$ besagt, dass $I_n - H = A^{-1}A + A^{-1}B = A^{-1}(A+B)$ regulär ist, also ist auch $A + B$ regulär. Die erste Ungleichung der Normabschätzung folgt direkt aus dem Hilfssatz, die zweite folgt mit der Submultiplikativität.

Die Fehleranalyse bei gestörter Matrix $\widetilde{A} = A + \Delta_A$ und gestörter rechter Seite $\widetilde{b} = b + \Delta_b \in \mathbb{K}^n$ liefert der folgende Satz.

2.1.12 Störungssatz (vollständige Form)
Die Matrix $A \in \mathbb{K}^{n\times n}$ sei regulär, der absolute Fehler $\Delta_A = \widetilde{A} - A$ erfülle $\|\Delta_A\| < \frac{1}{\|A^{-1}\|}$. Die Vektoren $x, \widetilde{x} \in \mathbb{K}^n$ seien die Lösungen von

$$Ax = b \qquad \text{bzw.} \qquad \widetilde{A}\widetilde{x} = \widetilde{b}. \qquad (b \neq 0)$$

Weiter seien Vektor- und Matrixnorm verträglich. Dann gilt für den relativen Fehler

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \|A^{-1}\|\,\|\Delta_A\|}\left(\frac{\|\Delta_b\|}{\|b\|} + \frac{\|\Delta_A\|}{\|A\|}\right) = \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\cdot\frac{\|\Delta_A\|}{\|A\|}}\left(\frac{\|\Delta_b\|}{\|b\|} + \frac{\|\Delta_A\|}{\|A\|}\right).$$

Beweis: Wie üblich in der Analysis fügen wir Summanden ein und verwenden die Dreiecksungleichung. Es gilt

$$\|\widetilde{x} - x\| = \|\widetilde{A}^{-1}\widetilde{b} - A^{-1}b\| \leq \|\widetilde{A}^{-1}(\widetilde{b} - b)\| + \|(\widetilde{A}^{-1} - A^{-1})b\|.$$

Aus dem ersten Teil wird nach Division durch $\|x\|$

$$\frac{\|\widetilde{A}^{-1}(\widetilde{b} - b)\|}{\|A^{-1}b\|} \leq \|\widetilde{A}^{-1}\|\,\|A\|\,\frac{\|\widetilde{b} - b\|}{\|b\|}.$$

Für den zweiten Teil verwenden wir die Umformung

$$\widetilde{A}^{-1} - A^{-1} = \widetilde{A}^{-1}(A - \widetilde{A})A^{-1}$$

und erhalten nach Division durch $\|x\|$

$$\frac{\|(\widetilde{A}^{-1} - A^{-1})b\|}{\|A^{-1}b\|} \leq \|\widetilde{A}^{-1}(A - \widetilde{A})\| \leq \|\widetilde{A}^{-1}\|\,\|A\|\,\frac{\|\widetilde{A} - A\|}{\|A\|}.$$

Die Behauptung folgt durch Einsetzen der Normabschätzung aus Korollar 2.1.11. Hierbei wird $B = \Delta A$ gesetzt, also ist $\widetilde{A} = A + B$.

2.1.13 Faustregel
Wir erklären den Störungssatz anhand der Anzahl der genauen Dezimalstellen im Lösungsvektor. Dies ist eher eine heuristische Analyse, sie gibt aber den Kern des Störungssatzes wieder.

Es seien $k, s \in \mathbb{N}$, $k > s$ und

$$\operatorname{cond}(A) \approx 10^s, \quad \frac{\|\Delta_A\|}{\|A\|} \approx 10^{-k}, \quad \frac{\|\Delta_b\|}{\|b\|} \approx 10^{-k}.$$

Dann gilt $\frac{\|\Delta_x\|}{\|x\|} \approx 10^{s-k}$, d.h. man verliert circa s Dezimalstellen Genauigkeit (in jeder Komponente bei Verwendung der Maximumsnorm).

2.1.14 Bemerkung: Bei großer Konditionszahl von A ist das *Residuum*

$$r = b - A\widetilde{x} = A(x - \widetilde{x})$$

ein schlechtes Maß für die Güte von $\widetilde{x}$: Es gilt

$$\frac{\|r\|}{\|b\|} \leq \frac{\|A\| \cdot \|\widetilde{x} - x\|}{\|Ax\|} \leq \operatorname{cond}(A) \cdot \frac{\|\widetilde{x} - x\|}{\|x\|}$$

und

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\|A\| \cdot \|A^{-1}(A\widetilde{x} - Ax)\|}{\|b\|} \leq \operatorname{cond}(A) \cdot \frac{\|r\|}{\|b\|}.$$

Die untere und obere Schranke von $\frac{\|\widetilde{x}-x\|}{\|x\|}$ in der Ungleichungskette

$$\frac{1}{\operatorname{cond}(A)} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \operatorname{cond}(A) \cdot \frac{\|r\|}{\|b\|}$$

liegen für große Konditionszahlen sehr weit auseinander, deshalb ist das Residuum r ungeeignet zur Abschätzung des relativen Fehlers an x.

2.1.15 Positiv-definite Matrizen

Zum besseren Verständnis der Spektralnorm wollen wir den Begriff der *positiv-definiten Matrizen* behandeln. Dieser wird üblicherweise in der Vorlesung Lineare Algebra II eingeführt. Wir benötigen den Begriff auch weiterhin in Abschnitt 2.3, 2.4 und Kapitel 6.

Definition:
Die Matrix $A \in \mathbb{C}^{n\times n}$ sei hermitesch. A heißt *positiv definit*, wenn

$$x^* Ax = \sum_{j,k=1}^{n} a_{jk}\overline{x_j}x_k > 0 \qquad \text{für alle } x \in \mathbb{C}^n \setminus \{0\}.$$

Gilt nur die schwache Ungleichung "≥ 0", so heißt A *positiv semi-definit*.

Äquivalent zur positiven (Semi-)Definitheit ist die Eigenschaft, dass alle Eigenwerte von A positiv (bzw. nichtnegativ) sind.

Man rechnet leicht nach, dass zu beliebiger Matrix $A \in \mathbb{C}^{n\times n}$ (sogar im rechteckigen Fall $A \in \mathbb{C}^{m\times n}$) die Matrix $B = A^*A$ hermitesch und positiv semi-definit ist, denn

$$x^*A^*Ax = (Ax)^*(Ax) = \|Ax\|_2^2 \geq 0.$$

Also sind alle Eigenwerte von A^*A größer oder gleich 0 und daher ist die Spektralnorm

$$\|A\|_2 := \max\{\sqrt{\lambda} : \lambda \text{ Eigenwert von } A^*A\} = \sqrt{\varrho(A^*A)}$$

wohldefiniert.

Wir beweisen, dass die Spektralnorm die *natürliche Matrixnorm* zur euklidischen Norm $\|\cdot\|_2$ in $\mathbb{C}^n$ ist. Dazu nehmen wir eine Orthonormalbasis $(v_1, \ldots, v_n)$ von $\mathbb{C}^n$, die aus Eigenvektoren von A^*A besteht, und schreiben jedes $x \in \mathbb{C}^n$ als

$$x = \sum_{j=1}^{n} \alpha_j v_j.$$

Die Orthonormalität liefert

$$\|x\|_2^2 = x^*x = \sum_{j,k=1}^{n} \overline{\alpha_j}\alpha_k \cdot v_j^* v_k = \sum_{j=1}^{n} |\alpha_j|^2.$$

Außerdem erhalten wir aus $A^*Av_j = \lambda_j v_j$ mit ähnlicher Rechnung

$$\|Ax\|_2^2 = x^* \cdot (A^*Ax) = \sum_{j=1}^{n} \overline{\alpha_j} v_j^* \cdot \sum_{k=1}^{n} \lambda_k \alpha_k v_k = \sum_{j=1}^{n} |\alpha_j|^2 \lambda_j.$$

Die rechte Seite schätzen wir nach oben ab durch $\varrho(A^*A)\|x\|^2$, und damit ist

$$\max_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \sqrt{\varrho(A^*A)}.$$

Die Gleichheit wird erzielt für einen Eigenvektor x zum größten Eigenwert $\lambda = \varrho(A^*A)$.

Die Spektralnorm hermitescher Matrizen lässt sich einfacher berechnen. Wegen $A^* = A$ ist ja die relevante Matrix $B = A^*A = A^2$. Sie hat als Eigenwerte genau die Quadrate der Eigenwerte von A und die gleichen Eigenvektoren wie A. Deshalb gilt

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ Eigenwert von } A^*A\} = \max\{|\lambda| : \lambda \text{ Eigenwert von } A\}.$$

Hier stimmen also *Spektralnorm* und *Spektralradius* überein!

Als Anwendung dieser Untersuchung gehen wir auf die Berechnung der Konditionszahl einer hermiteschen regulären Matrix ein. Die Inverse A^{-1} ist ebenfalls hermitesch, ihre Eigenwerte sind die Kehrwerte der Eigenwerte von A. Daraus ergibt sich ohne große Mühe

$$\operatorname{cond}_2(A) = \frac{\max\{|\lambda| : \lambda \text{ Eigenwert von } A\}}{\min\{|\lambda| : \lambda \text{ Eigenwert von } A\}}, \quad \text{falls } A \text{ hermitesch.}$$

2.2 Eliminationsverfahren, LR-Zerlegung

Zur Lösung des linearen Gleichungssystems $Ax = b$ verwendet man einfache Äquivalenz-Umformungen des Gleichungssystems.

2.2.1 Satz: Elementare Umformungen des LGS
Die Menge der Lösungen eines linearen Gleichungssystems bleibt unverändert, wenn man

E1 die Reihenfolge der Gleichungen vertauscht,

E2 beide Seiten einer Gleichung mit einer Zahl $\alpha \neq 0$ multipliziert,

E3 eine Gleichung ersetzt durch die Summe *dieser* Gleichung und dem Vielfachen einer *anderen* Gleichung.

E4 Vertauscht man die Reihenfolge der Unbekannten $x_1, \ldots, x_n$, setzt also

$$(y_1, \ldots, y_n) := (x_{\sigma_1}, \ldots, x_{\sigma_n})$$

mit einer *Permutation* $(\sigma_1, \ldots, \sigma_n)$ der Zahlen $(1, \ldots, n)$, so erhält man die Lösungsmenge des neuen Systems (bzgl. $\vec{y}$) aus der Lösungsmenge des alten (bzgl. $\vec{x}$) durch entsprechende Vertauschung der Komponenten.

2.2.2 Die elementaren Umformungen in Matlab/Octave:

- MATLAB verwendet die Notation

  `A(j,:)` für die j-te Zeile von A

  `A(:,k)` für die k-te Spalte von A

  Damit lassen sich die elementaren Umformungen einer Matrix einfach schreiben:

`A([j1,j2],:)=A([j2,j1],:)`	Zeilentausch von Zeile $j1$ und $j2$
`A(:,[k1,k2])=A(:,[k2,k1])`	Spaltentausch von Spalte $k1$ und $k2$
`A(j,:)=t*A(j,:)`	Multiplikation der Zeile j mit der Zahl t
`A(j2,:)=A(j2,:)+t*A(j1,:)`	Addition des t-fachen der Zeile $j1$ zur Zeile $j2$

Bei der mathematische Analyse der Eliminationsverfahren verwendet man die folgende Darstellung dieser Operationen.

2.2.3 Die elementaren Umformungen als Matrix-Multiplikation:

E1 Zeilentausch $j_1 \leftrightarrow j_2$: $P_{j_1,j_2} \cdot A$ mit *einfacher Permutationsmatrix*

$$P_{j_1,j_2} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & & 1 & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & & 0 & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \leftarrow\ j_1 \\ \\ \\ \\ \leftarrow\ j_2 \\ \\ \\ \end{matrix}$$

Die Multiplikation von rechts, also $A \cdot P_{k_1,k_2}$, führt den Spaltentausch (E4) durch.

E3$^+$ Für $j = k+1, \ldots, n$: subtrahiere das q_j-fache der Zeile k von der Zeile j:
$(I_n - L_k(q)) \cdot A$ mit *strikter unterer Dreiecksmatrix*

$$L_k(q) = \begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & q_{k+1} & 0 & \\ & & \vdots & & \ddots \\ & & q_n & & 0 \end{pmatrix} \leftarrow\ k, \qquad q = \begin{pmatrix} q_{k+1} \\ \vdots \\ q_n \end{pmatrix} \in \mathbb{K}^{n-k}.$$

2.2.4 Hilfssatz: Eigenschaften der elementaren Umformungsmatrizen

a) Die einfachen Permutationsmatrizen sind symmetrisch und orthogonal, also $P_{j_1,j_2}^{-1} = P_{j_1,j_2}$.

b) Für $1 \le j_1 \le j_2 \le n$ und Vektoren p, q ist

$$(I_n - L_{j_1}(p)) \cdot (I_n - L_{j_2}(q)) = I_n - L_{j_1}(p) - L_{j_2}(q).$$

Insbesondere ist $(I_n - L_{j_1}(p))^{-1} = I_n + L_{j_1}(p)$.

c) Für $1 \le j_1 < j_2 < j_3 \le n$ ist

$$P_{j_2,j_3}(I_n - L_{j_1}(q)) = (I_n - L_{j_1}(\widetilde{q}))P_{j_2,j_3},$$

mit $\widetilde{q} = P_{j_2,j_3}(0, \ldots, 0, q_{j_1+1}, \ldots, q_n)^T$.

2.2.5 Definition: Dreiecksmatrizen
Eine Matrix der Form

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & r_{nn} \end{bmatrix}$$

heißt *obere Dreiecksmatrix*.

Eine Matrix der Form

$$L = \begin{bmatrix} \ell_{11} & 0 & \cdots & 0 \\ \ell_{21} & \ell_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \ell_{n,1} & \cdots & \ell_{n,n-1} & \ell_{nn} \end{bmatrix}$$

heißt *untere Dreiecksmatrix*.

Sind zusätzlich alle Diagonalelemente gleich Null, so spricht man von *strikter* oberer bzw. unterer Dreiecksmatrix.

Bemerkung:

- Das Produkt $R_1 R_2$ oberer Dreiecksmatrizen ist eine obere Dreiecksmatrix.
- Die obere Dreiecksmatrix R ist genau dann regulär, wenn alle Diagonalelemente ungleich Null sind. Es gilt $\det R = \prod_{j=1}^{n} r_{jj}$.

2.2.6 LGS mit Dreiecksmatrizen: Liegt ein Gleichungssystem vor, dessen Matrix bereits die obere oder untere Dreiecksform hat, erfolgt die Lösung durch Vorwärts- bzw. Rückwärtseinsetzen:

- Lösen von $Ly = \widetilde{b}$ mit unterer Dreiecksmatrix L durch Vorwärtseinsetzen

$$y_1 = \frac{\widetilde{b}_1}{\ell_{11}}, \quad y_2 = \frac{1}{\ell_{22}}(\widetilde{b}_2 - \ell_{21} y_1), \quad \ldots$$

- Lösen von $Rx = y$ mit oberer Dreiecksmatrix R durch Rückwärtseinsetzen

$$x_n = \frac{y_n}{r_{nn}}, \quad x_{n-1} = \frac{1}{r_{n-1,n-1}}(y_{n-1} - r_{n-1,n} x_n), \quad \ldots$$

Zur Lösung des linearen Gleichungssystems $Ax = b$ mit regulärer Matrix A führt man die elementaren Umformungen (E1=Zeilentausch) und (E3$^+$=Elimination der Spalte) nach einem fest vorgegebenen System durch, um auf eine Dreiecksform der Matrix zu kommen. Wir beschreiben nun den *Gauß-Algorithmus zur LR-Zerlegung* und berücksichtigen dabei eine sog. *Pivot-Strategie*, hier die Spaltenpivotsuche. Das Verfahren entspricht weitgehend der klassischen Gauß-Elimination aus der Linearen Algebra. Darüberhinaus erzielen wir sogar die *LR-Zerlegung*

$$PA = LR, \tag{2.1}$$

wobei P eine Permutationsmatrix ist, die alle durchgeführten Zeilenvertauschungen "bündelt", und L, R reguläre untere bzw. obere Dreiecksmatrizen sind. Diese LR-Zerlegung erfordert keine zusätzlichen Operationen im Vergleich zur Gauß-Elimination, sie ergänzt die Elimination nur durch die "Buchführung" der Zeilenvertauschungen und der Koeffizienten q_j im Schritt (E3$^+$).

Mit der LR-Zerlegung in (2.1) löst man anschließend das LGS durch Vorwärts- und Rückwärtseinsetzen

$$Ax = b \quad \Longleftrightarrow \quad PAx = Pb \quad \Longleftrightarrow \quad (Ly = Pb, \quad Rx = y). \tag{2.2}$$

Hierbei entsteht der Vektor Pb aus der gegebenen rechten Seite b durch Anwendung der Zeilenvertauschungen.

2.2.7 Gauß-Algorithmus zur LR-Zerlegung mit Spaltenpivotsuche

Wir bezeichnen die Matrixeinträge durchgehend mit `A(j,k)` wie in der Matlab-Notation; Einträge werden durch neue überschrieben und wieder als `A(j,k)` aufgerufen. Deshalb benötigen wir keinen zusätzlichen Index an der Matrix für die Schrittzahl k im Algorithmus.

Als Hilfsvektor führen wir den Array `pindex` mit, der die Zeilenvertauschungen enthält und durch `pindex=[1:n]` initialisiert wird.

k-ter Eliminations-Schritt für $1 \le k \le n-1$:

- Suche in der k-ten Spalte abwärts von der Diagonalen das "Pivotelement" `A(j,k)`, $k \le j \le n$, mit maximalem Absolutbetrag und führe evtl. eine Zeilenvertauschung $k \leftrightarrow j$ durch:

```
[v,s]=max(abs(A(k:n,k)));  j=s(1)+k-1;
if j>k,  A([k,j],:)=A([j,k],:);  pindex([k,j])=pindex([j,k]);  end
```

- Eliminiere die Einträge der k-ten Spalte unter der Diagonalen durch Subtraktion eines Vielfachen der k-ten Zeile und speichere die Faktoren q:

```
for j=k+1:n
    q=A(j,k)/A(k,k);
    A(j,k+1:n)=A(j,k+1:n)-q*A(k,k+1:n);   % Elimination in Zeile j
    A(j,k)=q;  % Speichern des Eintrags von L
end
```

Aufstellen der Matrizen P, L, R für $PA = LR$:
P ist die Permutationsmatrix, deren Zeile j den kanonischen Einheitsvektor $e_{\texttt{pindex(j)}}$ enthält.
R ist die obere Dreiecksmatrix in der oberen Hälfte von `A` nach Durchführung der Elimination.
$L = I_n + N$ ist die untere Dreiecksmatrix mit Einsen auf der Diagonalen und den

Faktoren q der Elimination, die unterhalb der Diagonalen in `A` eingetragen wurden.

```
P=eye(n);  P=P(pindex,:);
R=triu(A);  % obere Hälfte nach Elimination
L=eye(n)+tril(A,-1);  % Diagonale 1 und strikte untere Hälfte
```

Bemerkungen:

(i) Falls im k-ten Schritt der Rechnung das Pivotelement `A(k,k)` mit sehr kleinem Betrag entsteht, wird eine Warnung oder Fehlermeldung "A fast singulär" ausgegeben. Denn dann sind alle Einträge der k-ten Spalte von der Diagonalen nach unten, also in `A(k:n,k)`, fast Null, damit ist die Matrix aufgrund von Rundungsfehlern nicht mehr von einer singulären Matrix zu unterscheiden.

(ii) Anstatt die Permutationsmatrix P aufzustellen genügt es, den Vektor `pindex` zu verwenden. Die Multiplikation Pb der rechten Seite des LGS ist gleich der Zeilenvertauschung `b(pindex)`.

2.2.8 Beispiel: $PA = LR$ in kompakter Schreibweise

$A = \begin{pmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{pmatrix}$, `pindex=[1,2,3]`

1. Schritt:

- Pivot-Element $a_{11} = 3$ in 1. Spalte, also kein Zeilentausch erforderlich, `pindex` ist unverändert.
- Elimination in 1. Spalte: $A^{(1)} = (I - L_1((\frac{2}{3}, \frac{1}{3})) * A$

$$\begin{array}{ccc} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \\ \hline \end{array}$$

$$\begin{array}{c|cc} 3 & 1 & 6 \\ \overline{2/3} & 1/3 & -1 \\ 1/3 & 2/3 & -1 \\ \hline \end{array}$$

Dabei Buchführung über die wesentlichen Einträge von $L_1((\frac{2}{3}, \frac{1}{3})$ in der 1. Spalte

2. Schritt:

- Pivot-Element $a_{32} = 2/3$ in 2. Spalte, also Zeilentausch mit $P_{2,3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. Dieser wird kurz dargestellt in `pindex=[1,3,2]`.
- Elimination in 2. Spalte: $A^{(2)} = (I - L_2((\frac{1}{2})) * P_{2,3} * A^{(1)}$

$$\begin{array}{c|cc} 3 & 1 & 6 \\ \hline 1/3 & 2/3 & -1 \\ 2/3 & 1/3 & -1 \end{array}$$

$$\begin{array}{c|cc} 3 & 1 & 6 \\ \hline 1/3 & 2/3 & -1 \\ 2/3 & \boxed{1/2} & -1/2 \end{array}$$

Dabei Buchführung über die wesentlichen Einträge von $L_2((\frac{1}{2}))$ in der 2. Spalte

Ablesen des Ergebnisses:

$$PA = \begin{pmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{pmatrix} = LR.$$

In der Linearen Algebra wurde bereits bewiesen, dass für jede reguläre Matrix A der Gauß-Algorithmus mit Spaltenpivotsuche durchführbar ist: die lineare Unabhängigkeit der Spalten von A sichert, dass im k-ten Schritt mindestens ein Eintrag der Spalte `A(k:n,k)` ungleich Null ist, d.h. der Zeilentausch liefert ein Diagonalelement `A(k,k)`$\neq 0$.

Die Identität $PA = LR$ in (2.1) beweist man mit den berechneten Matrizen des Algorithmus unter Zuhilfenahme von Hilfssatz 2.2.4.

- In Matrixschreibweise liefert der Algorithmus die obere Dreiecksmatrix R durch die Operationen

  $$R = (I_n - L_{n-1}(q_{n-1}))P_{n-2,j_{n-2}} \cdots P_{3,j_3}(I_n - L_2(q_{n-1}))P_{2,j_2}(I_n - L_1(q_1))P_{1,j_1} * A.$$

  Dabei ist jeweils $j_k \geq k$, im Fall $j_k = k$ ist die Permutationsmatrix die Einheitsmatrix.

- Die erste Beobachtung fußt auf Teil (c) in 2.2.4 und schiebt alle Permutationsmatrizen nach rechts,

  $$R = (I_n - L_{n-1}(q_{n-1})) \cdots (I_n - L_2(\widetilde{q}_2))(I_n - L_1(\widetilde{q}_1))P_{n-2,j_{n-2}} \cdots P_{3,j_3}P_{2,j_2}P_{1,j_1} * A.$$

  Dabei erfahren die Vektoren q_k entsprechende Zeilenvertauschungen, die im Algorithmus im Befehl

  ```
  if j>k, A([k,j],:)=A([j,k],:);  pindex([k,j])=pindex([j,k]); end
  ```

  enthalten sind: der vordere Teil der k-ten und j-ten Zeile enthält ja bereits die gespeicherten Faktoren q aus den vorherigen Eliminationsschritten. Dies wird im Beispiel im 2. Schritt deutlich, weil die Faktoren 2/3 und 1/3 in der ersten Spalte die Plätze tauschen.

- Das Produkt der einfachen Permutationsmatrizen liefert die Matrix P, die durch den Vektor `pindex` angegeben ist.

- Nun wird mit den Inversen der unteren Dreiecksmatrizen multipliziert. Aus Hilfssatz 2.2.4 (a) und (b) ergibt sich

$$\begin{aligned} PA &= (I_n - L_1(\widetilde{q}_1))^{-1}(I_n - L_2(\widetilde{q}_2))^{-1} \cdots (I_n - L_{n-1}(q_{n-1}))^{-1} * R \\ &= (I_n + L_1(\widetilde{q}_1))(I_n + L_2(\widetilde{q}_2)) \cdots (I_n + L_{n-1}(q_{n-1})) * R \\ &= (I_n + L_1(\widetilde{q}_1) + L_2(\widetilde{q}_2) + \cdots + L_{n-1}(q_{n-1})) * R. \end{aligned}$$

  In der Klammer steht die berechnete Matrix L mit Diagonalelementen Eins und den Faktoren aus der Buchführung der Elimination.

Damit haben wir den folgenden Satz bewiesen.

2.2.9 Satz zur LR-Zerlegung:
Zu jeder regulären Matrix $A \in \mathbb{K}^{n\times n}$ existiert eine Permutationsmatrix P, eine untere Dreiecksmatrix L mit Diagonalelementen 1 sowie eine obere Dreiecksmatrix R mit

$$P \cdot A = L \cdot R.$$

Die Matrizen P, L und R ergeben sich aus der Gauß–Elimination mit Spaltenpivotierung.

2.2.10 Komplexität der LR-Zerlegung: Zählt man eine Multiplikation mit anschließender Addition als eine Rechenoperation, so erfordert Schritt k

- die Berechnung von $n-k$ Quotienten (=Einträge von L in Spalte k),
- die Berechnung der neuen Einträge `A(j,r)` im Bereich $k+1 \le j, r \le n$.

Der Gesamtaufwand der LR-Zerlegung ist also

$$\sum_{k=1}^{n-1}(n-k+(n-k)^2) = \sum_{k=1}^{n-1}(k+k^2) = \frac{(n-1)n}{2} + \frac{(n-1)n(2n-1)}{6} = \frac{n^3}{3} + \mathcal{O}(n^2).$$

Das anschließende Vorwärts- und Rückwärtseinsetzen erfordert zusammen $n^2+\mathcal{O}(n)$ Operationen.

Wenn man $Ax = b$ mit der LR-Zerlegung löst, stellt sich die Frage, wie sich $\mathrm{cond}(A)$ zu $\mathrm{cond}(L)$ und $\mathrm{cond}(R)$ verhält. Zuerst geben wir eine grobe Auskunft über L und R aus dem Buch *Demmel, Applied Numerical Linear Algebra, SIAM Publ. 1997, S. 49.*

2.2.11 Konditionierung der Gauß–Elimination mit Pivotierung:
Sei $A \in \mathbb{K}^{n\times n}$ regulär und $PA = LR$ die LR-Zerlegung nach dem Gauß-Algorithmus mit Spaltenpivotierung.

a) Die Einträge der Matrix L erfüllen $|\ell_{jk}| \le 1$.

b) Die Einträge der Matrix R erfüllen

$$\max_{1\le j,k\le n} |r_{jk}| \le 2^{n-1} \max_{1\le j,k\le n} |a_{jk}|.$$

Der Faktor 2^{n-1} in der oberen Schranke ist scharf, er wird angenommen für Matrizen des Typs

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix}.$$

Ein weiteres Resultat von Wilkinson befasst sich mit der "Rückwärts-Stabilität" des numerischen Verfahrens in der folgenden Form.

2.2.12 Satz: Numerische Stabilität der $LR-$Zerlegung:
Seien $A \in \mathbb{K}^{n\times n}$ regulär und $b \in \mathbb{K}^n$. Das Gleichungssystem $A \cdot x = b$ werde mit Gauß–Elimination mit Spaltenpivotierung gelöst. Dann ist die unter dem Einfluss von Rundungsfehlern tatsächlich berechnete Lösung

$$\widetilde{x} = x + \Delta_x$$

die exakte Lösung eines Gleichungssystems

$$(A + \Delta_A) \cdot (x + \Delta_x) = b$$

mit einer Störung ΔA der relativen Größe

$$\frac{\|\Delta_A\|_\infty}{\|A\|_\infty} \le 1.01 \cdot 2^{n-1} \cdot (n^3 + 2n^2) \cdot \text{eps}.$$

Diese Abschätzung ist jedoch für praktische Fälle oft zu pessimistisch. Insbesondere für *dünnbesetzte* Matrizen sind wesentlich bessere Resultate zu erwarten.

2.2.13 Ergänzungen zur LR-Zerlegung::

I. Totalpivotierung

II. Determinanten- und Rangbestimmung

III. Nachiteration

IV. Berechnung der Inversen: Austauschschritte

Wir bezeichnen die Matrix nach dem k-ten Eliminationsschritt mit $A^{(k)} = \left(a_{ij}^{(k)}\right)$.

I. Totalpivotierung:

Im k-ten Schritt wird anstatt des betragsgrößten Elements $a_{j,k}^{(k-1)}$, $j = k : n$, der k-ten Spalte in dem gesamten Bereich $A^{(k-1)}(k : n, k : n)$ nach dem betragsgrößten Element gesucht. Dieses wird mittels Zeilen- und **Spalten-**Tausch (E1 und E4) an die Stelle des Pivot-Elements $a_{k,k}$ gebracht. Verschiedene Webseiten stellen hierfür ein Matlab-Programm `gecp.m` zur Verfügung, ich empfehle die Version von N. Higham auf github unter `carandraug/testmatrix/`.

- Dadurch wird die maximal mögliche Vergrößerung der Elemente von A von 2^{n-1} (siehe 2.2.10) auf den Faktor n reduziert.
- Der gesamte Aufwand zur Suche der Pivot-Elemente steigt allerdings von $n^2/2$ (bei Spaltenpivotisierung) auf $n^3/3$.

II. Determinanten- und Rangbestimmung:

- Determinante:

  Wegen $\det(I_n + L_k(q)) = 1$ (mit strikter unterer Dreiecksmatrix $L_k(q)$) folgt mit dem Determinanten-Multiplikationssatz $\det L = 1$.

  Für einfache Permutationsmatrizen ist $\det P_{j,k} = -1$, $1 \le j < k \le n$.

  Also ist

  $$\pm \det A = \det(PA) = \det(LR) = \det R = \prod_{k=1}^{n} r_{kk}.$$

  Das Vorzeichen richtet sich danach, ob eine gerade oder ungerade Anzahl von Permutationen durchgeführt wurde.

- Rangbestimmung:

  Bei regulärem A ist $\operatorname{Rang} A = \operatorname{Rang} R = n$, weil die elementaren Umformungen den Rang nicht verändern.

  Im Fall $\operatorname{Rang} A < n$ kann der Fall auftreten, dass alle Spalten-Einträge $a_{jk}^{(k-1)}$, $j = k : n$, im k-ten Schritt Null sind. Dann ist Spalten-Vertauschung erforderlich (siehe I.). Damit wird die Stufenform von A erzeugt, die eine oder mehrere Nullzeilen enthält. Der Rang von A ist die Anzahl der von Null verschiedenen Zeilen in der Stufenform.

III. Nachiteration:

Problem: Löse $Ax = b$ mit der LR-Zerlegung $PA = LR$.

- Wir betrachten den Fall, dass die LR-Zerlegung stark verfälscht wurde und $\Delta_A = \widetilde{L}\widetilde{R} - PA$ einen großen relativen Fehler

$$\epsilon_1 := \frac{\|\Delta_A\|}{\|A\|} \gg \text{eps}$$

aufweist. Dies kann z.B. durch Abspeichern von L und R im "single" Zahlenformat und späteres Einlesen erfolgt sein. Die Lösung von $\widetilde{L}\widetilde{R}\widetilde{x} = Pb$ erfüllt nach Satz 2.1.12

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \kappa(A) \cdot \epsilon_1, \qquad \kappa(A) = \frac{\text{cond}(A)}{1 - \text{cond}(A) \cdot \frac{\|\Delta_A\|}{\|A\|}}.$$

- Das Residuum

$$r^0 = b - A\widetilde{x} \qquad \text{(berechnet mit dem exakten } A \text{ und Rundungsfehler eps)}$$

ergibt den Korrekturvektor k^1 als Lösung von

$$\widetilde{L}\widetilde{R}k^1 = Pr^0.$$

- Als verbesserte Lösung setzen wir $x^1 = \widetilde{x} + k^1$, denn

$$\begin{aligned} x^1 - x &= \widetilde{x} - x + (\widetilde{L}\widetilde{R})^{-1}(Pb - PA\widetilde{x}) \\ &= (I_n - (\widetilde{L}\widetilde{R})^{-1}PA)(\widetilde{x} - x) \\ &= (\widetilde{L}\widetilde{R})^{-1}(\underbrace{\widetilde{L}\widetilde{R} - PA}_{=\Delta_A})(\widetilde{x} - x), \end{aligned}$$

also

$$\frac{\|x^1 - x\|}{\|x\|} \leq \|(\widetilde{L}\widetilde{R})^{-1}\| \, \|\Delta_A\| \, \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \|(\widetilde{L}\widetilde{R})^{-1}\| \, \|A\| \, \frac{\|\Delta_A\|}{\|A\|} \, \frac{\|\widetilde{x} - x\|}{\|x\|}.$$

Mit $\widetilde{L}\widetilde{R} = PA + \Delta_A$ folgt wie im Beweis von Satz 2.1.12

$$\frac{\|x^1 - x\|}{\|x\|} \leq \kappa(A) \, \frac{\|\Delta_A\|}{\|A\|} \, \frac{\|\widetilde{x} - x\|}{\|x\|} \leq \kappa(A)^2 \, \epsilon_1^2,$$

also eine Verdopplung der exakten Dezimalstellen! Weitere Schritte liefern Verbesserung bis zur Rechengenauigkeit eps. In der Praxis genügen $1 - 3$ Schritte der Nachiteration.

IV. Berechnung der Inversen: Löse $AX = I_n$
1. Variante:

1. Aufstellen der LR-Zerlegung $PA = LR$.
2. Lösen von $LY = P$ mit Vorwärtseinsetzen (simultan für jede Spalte der rechten Seite).
3. Lösen von $RX = Y$ mit Rückwärtseinsetzen.

2. Variante:
Vorwärts- und Rückwärts-Elimination in A mit Skalierung der Zeilen

1. Starte mit der erweiterten Matrix $(A \mid I_n)$.
2. Führe Elimination mit Spaltenpivotierung und Skalierung (E2) durch, um auf die Form $(I_n \mid A^{-1})$ zu kommen.

In Matlab/Octave: das Kommando `rref([A,eye(n)])` liefert das obige Ergebnis, rref steht für "row-reduced-echolon-form"; siehe auch die schrittweise Ausführung in `rrefmovie([A,eye(n)])` in alten Matlab-Versionen

3. Variante: Gauß–Jordan-Algorithmus (Austauschverfahren)

Dieses Verfahren wird auch in der Linearen Optimierung eingesetzt (im Simplex-Verfahren). Betrachte das LGS $Ax = y$ mit **variablen** Vektoren $x = (x_1, \ldots, x_n)^T$, $y = (y_1, \ldots, y_n)^T$, und ersetze nacheinander die Komponenten von x_j durch die y_k's:

Wenn $a_{pq} \neq 0$ gilt, kann Gleichung p nach x_q aufgelöst werden. Man nennt dies einen Austauschschritt, weil die Variable x_q durch die Variable y_p ausgetauscht wird:

$$-\frac{a_{p1}}{a_{pq}}x_1 - \cdots - \frac{a_{p,q-1}}{a_{pq}}x_{q-1} + \frac{1}{a_{pq}}y_p - \frac{a_{p,q+1}}{a_{pq}}x_{q+1} - \cdots - \frac{a_{p,n}}{a_{pq}}x_n = x_q. \qquad \text{(Pivotzeile)}$$

Dies ist die neue p-te Zeile des äquivalenten Gleichungssystems

$$A^{(1)}(x_1, \ldots, x_{q-1}, y_p, x_{q+1}, \ldots, x_n)^T = (y_1, \ldots, y_{p-1}, x_q, y_{p+1}, \ldots, y_n)^T.$$

In den anderen Gleichungen ist x_q ebenfalls zu ersetzen; dies liefert die Einträge von $A^{(1)}$ zum 1. Austauschschritt:

$$\begin{array}{lll}
\text{Pivotelement} & : & a^{(1)}_{pq} = \dfrac{\mathbf{1}}{\mathbf{a_{pq}}}, \\
\text{Pivotzeile} & : & a^{(1)}_{pk} = -\dfrac{a_{pk}}{a_{pq}}, \qquad k \neq q, \\
\text{Pivotspalte} & : & a^{(1)}_{jq} = \dfrac{a_{jq}}{a_{pq}}, \qquad j \neq p, \\
\text{sonstige} & : & a^{(1)}_{jk} = a_{jk} - a_{jq}\frac{a_{pk}}{a_{pq}}, \qquad j \neq p,\ k \neq q
\end{array}$$

In den folgenden Austausch-Schritten wird nur außerhalb der p-ten Zeile und q-ten Spalte nach dem Pivotelement gesucht, um einen "Rücktausch" zu vermeiden.

Bemerkung: Aufwand $n^3 + \mathcal{O}(n^2)$.

2.3 Spezielle Gleichungssysteme, Cholesky-Zerlegung

Wir betrachten Gleichungssysteme mit spezieller Struktur, bei denen die LR-Zerlegung ohne Pivotierung Vorteile hat.

2.3.1 Satz: LR-Zerlegung ohne Pivotierung
Es sei $A \in \mathbb{K}^{n\times n}$ regulär. Die LR-Zerlegung $A = LR$ mit unterer Dreiecksmatrix L, deren Diagonalelemente alle Eins sind, und oberer Dreiecksmatrix R existiert genau dann, wenn jede Teilmatrix

$$A_k = (a_{ij})_{i,j=1,\dots,k}\,, \qquad k = 1,\dots,n,$$

regulär ist. Weiterhin ist dann die LR-Zerlegung von A eindeutig.

Beweis:
1. Falls die LR-Zerlegung $A = LR$ existiert, gilt für $1 \le k \le n$

$$A_k = \begin{pmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \vdots & & \ddots & \\ \ell_{k1} & \cdots & \ell_{k,k-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ & r_{22} & \cdots & r_{2k} \\ & & \ddots & \vdots \\ & & & r_{kk} \end{pmatrix},$$

also $\det A_k = \prod_{j=1}^{k} r_{jj} \neq 0$. Deshalb ist A_k regulär.

2. Für die umgekehrte Schlussrichtung führen wir Induktion nach n.
Für $n = 1$, also $A \in \mathbb{K}$ mit $A = \det A \neq 0$, ist $A = 1 \cdot A$ die LR-Zerlegung.

Sei nun $n \in \mathbb{N}$, $n \ge 2$, $A \in \mathbb{K}^{n\times n}$ regulär und $\det A_k \neq 0$ für $k = 1,\dots,n$.

Induktionsannahme: alle regulären Matrizen $B \in \mathbb{K}^{(n-1)\times(n-1)}$ mit der Eigenschaft $\det B_k \neq 0$ für $k = 1,\dots,n-1$ besitzen eine LR-Zerlegung.

Wegen $a_{11} = \det A_1 \neq 0$ kann der 1. Eliminationsschritt ohne Zeilenvertauschung ausgeführt werden. Dies liefert

$$A = (I_n + L_1(q)) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & B & \\ & & & \end{pmatrix}.$$

mit $q = \frac{1}{a_{11}}(a_{21},\dots,a_{n1})^T$ und $B \in \mathbb{K}^{(n-1)\times(n-1)}$. Die Teilmatrix A_k von A ergibt sich hieraus als

$$A_k = \begin{pmatrix} 1 & & & & \\ q_1 & 1 & & & \\ q_2 & 0 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ q_{k-1} & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ 0 & & B_{k-1} & \\ & & & \end{pmatrix}$$

für $2 \le k \le n$, mit der entsprechenden Teilmatrix B_{k-1} von B. Wegen $0 \neq \det A_k = a_{11} \det B_{k-1}$ ist $\det B_k \neq 0$ für $k = 1, \ldots, n-1$. Also besitzt die Matrix B nach der Induktionsannahme die LR-Zerlegung $B = \widetilde{L}\widetilde{R}$, und damit ist

$$A = (I_n + L_1(q)) \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & \widetilde{L}\widetilde{R} & \end{pmatrix} = \underbrace{(I_n + L_1(q)) \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & \widetilde{L} & \end{pmatrix}}_{=L} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & & \widetilde{R} & \end{pmatrix}.$$

3. Die Eindeutigkeit der LR-Zerlegung folgt so: Aus $A = L_1 R_1 = L_2 R_2$ mit unteren Dreiecksmatrizen L_1, L_2, deren Diagonalelemente 1 sind, und oberen Dreiecksmatrizen R_1, R_2 folgt

$$L_2^{-1} A R_1^{-1} = L_2^{-1} L_1 = R_2 R_1^{-1}.$$

Die Matrix $L_2^{-1} L_1$ ist untere Dreiecksmatrix mit Diagonalelementen 1, die Matrix $R_2 R_1^{-1}$ ist obere Dreiecksmatrix. Aus der Gleichheit folgt

$$L_2^{-1} L_1 = I_n = R_2 R_1^{-1}.$$

Die Eindeutigkeit der Inversen liefert $L_2 = L_1$ und $R_2 = R_1$.

Die Berechnung der LR-Zerlegung ohne Zeilenvertauschung erfolgt entweder mit dem Gauß-Algorithmus in 2.2.7, wobei die Pivotsuche einfach ausgelassen wird, oder durch den folgenden Ansatz.

2.3.2 Direkte LR-Zerlegung (Verfahren von Crout)
Falls die LR-Zerlegung ohne Zeilenvertauschung durchgeführt werden kann, liefert der *Ansatz*

$$A = LR = \begin{pmatrix} 1 & & & \\ \ell_{21} & 1 & & \\ \vdots & & \ddots & \\ \ell_{n1} & \cdots & \ell_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}$$

durch Betrachten der Einträge von A nacheinander die Gleichungen

1. Zeile	$k = 1, \ldots, n:$	$a_{1k} = r_{1k}$		
1. Spalte	$j = 2, \ldots, n:$	$a_{j1} = \ell_{j1} r_{11}$	$\Rightarrow$	$\ell_{j1} = \dfrac{a_{j1}}{r_{11}}$
	$\vdots$			
p-te Zeile	$k = p, \ldots, n:$	$a_{pk} = \sum_{\mu=1}^{p-1} \ell_{p\mu} r_{\mu k} + r_{pk}$	$\Rightarrow$	$r_{pk} = a_{pk} - \sum_{\mu=1}^{p-1} \ell_{p\mu} r_{\mu,k}$
p-te Spalte	$j = p+1, \ldots, n:$	$a_{jp} = \sum_{\mu=1}^{p} \ell_{j\mu} r_{\mu p}$	$\Rightarrow$	$\ell_{jp} = \dfrac{1}{r_{pp}} \left(a_{jp} - \sum_{\mu=1}^{p-1} \ell_{j\mu} r_{\mu p} \right)$

Die Bedingung in Satz 2.3.1 ist insbesondere für die positiv-definiten Matrizen erfüllt: in der Linearen Algebra wird die folgende Aussage bewiesen.

2.3.3 Satz: Positiv definite Matrizen
Eine symmetrische Matrix $A \in \mathbb{R}^{n\times n}$ ist positiv-definit genau dann, wenn für die Teilmatrizen $A_k = (a_{ij})_{i,j=1,\dots,k}$ gilt

$$\det A_k > 0, \qquad k = 1, \dots, n.$$

An Stelle der LR-Zerlegung nach Crout wird eine ähnliche Zerlegung bevorzugt.

2.3.4 Folgerung: Cholesky-Zerlegung
Zu einer symmetrischen positiv-definiten Matrix $A \in \mathbb{R}^{n\times n}$ existiert eine untere Dreiecksmatrix L mit

$$A = LL^T.$$

Achtung: L hat positive Diagonalelemente, die nicht alle 1 sein müssen (im Gegensatz zum L in der LR-Zerlegung).

Man berechnet den Cholesky-Faktor L ähnlich wie im Algorithmus von Crout:

2.3.5 Algorithmus zur Cholesky-Zerlegung:
Eingabe: $A \in \mathbb{R}^{n\times n}$ symmetrisch

Berechnung: Für $k = 1, \dots, n$

(1) Berechne $\quad \ell_{k,k} = \left(a_{k,k} - \sum_{\mu=1}^{k-1} \ell_{k,\mu}^2 \right)^{1/2}$

(2) Berechne für $j = k+1, \dots, n \quad \ell_{j,k} = \frac{1}{\ell_{k,k}} \left(a_{j,k} - \sum_{\mu=1}^{k-1} \ell_{j,\mu}\ell_{k,\mu} \right)$

Bemerkung:

(i) Ob A positiv-definit ist, stellt man im Algorithmus fest: der Ausdruck unter der Wurzel muss positiv sein! Eine Fehlernachricht *"A nicht positiv definit"* wird ausgegeben, falls diese Bedingung verletzt wird.

(ii) Der Rechenaufwand für die Berechnung der Cholesky–Zerlegung einer positiv definiten Matrix (Wurzel zählt als eine Operation wie Mult./Add.) ist

$$\sum_{k=1}^{n} (k + (n-k)k) = \sum_{k=1}^{n} k(n-k+1) = \frac{n^3}{6} + \mathcal{O}\left(n^2\right).$$

In Matlab/Octave (mit Tests der Eingangsdaten auf richtige Dimension, Symmetrie, positive Definitheit und Warnung bei sehr kleinem Diagonalelement von L):

```
function L = mychol(A)
% teste Dimension von A
[m,n]=size(A);
if m ~= n
 error('Matrix ist nicht quadratisch')
end
% teste auf Symmetrie
if norm(triu(A,1)-tril(A,-1)')>n*eps
 error('Matrix ist nicht symmetrisch')
end
L=zeros(n,n);
for k=1:n
 g=A(k,k)-L(k,1:(k-1))*L(k,1:(k-1))';
 % teste auf positive Definitheit
 if g<=0
  error('Matrix ist nicht positiv definit')
 end
 L(k,k)=sqrt(g);
 % teste auf numerische Invertierbarkeit
 if L(k,k)<eps
  fprintf('Warnung: Matrix ist eventuell nicht invertierbar')
 end
 L((k+1):n,k)=1/L(k,k)*(A((k+1):n,k)-L((k+1):n,1:(k-1))*L(k,1:(k-1))');
end
```

Der Test `L=mychol(hilb(3))` ergibt

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{\sqrt{12}} & 0 \\ \frac{1}{3} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{180}} \end{pmatrix}$$

Probe: `norm(hilb(3)-L*L')`

2.3.6 Definition: Bandmatrizen
Eine Matrix $A \in \mathbb{K}^{n\times n}$ heißt Bandmatrix vom Typ (p,q) mit $0 \le p,q \le n-1$, wenn

$$a_{j,k} = 0 \quad \text{für alle Paare } (j,k) \text{ mit} \quad k < j-p \quad \text{oder} \quad k > j+q,$$

d.h. A hat Nullen außer in der Diagonalen, den p unteren und q oberen Nebendiagonalen. Die Größe $p+q+1$ ist die *Bandbreite* von A.

Bemerkung:

- Eine "kompakte" Speicherung einer Bandmatrix vom Typ (p,q) erfordert nur $n(p+q+1)$ Speicherplätze, jeweils n für die Diagonale und Nebendiagonalen.
- Für eine symmetrische Bandmatrix vom Typ (p,p) speichert man nur die Diagonale und die unteren Nebendiagonalen, also nur $n(p+1)$ Einträge.

Beispiel: symmetrische Tridiagonalmatrix (Typ (1,1)):

$$A = \begin{pmatrix} a_1 & b_1 & & & \\ b_1 & a_2 & b_2 & & \\ & & \ddots & & \\ & & b_{n-2} & a_{n-1} & b_{n-1} \\ & & & b_{n-1} & a_n \end{pmatrix} \quad \leftrightarrow \quad \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ b_1 & \cdots & b_{n-1} & 0 \end{pmatrix}$$

2.3.7 Satz:
Es sei $A \in \mathbb{K}^{n\times n}$ eine Bandmatrix vom Typ (p,q). Falls die $LR-$Zerlegung $A = L \cdot R$ existiert, dann ist L eine Bandmatrix vom Typ $(p,0)$ und R eine Bandmatrix vom Typ $(0,q)$.

Bemerkung:

- Die LR-Zerlegung $A = LR$ kann also in kompakter Speicherung durchgeführt werden. Der Rechenaufwand in jedem Eliminationsschritt ist ungefähr gleich:

$$p \text{ Divisionen}, \quad \sum_{j=1}^{p} (q+j) \text{ Mult./Add.},$$

  also insgesamt $n\left(pq + \frac{p^2}{2}\right) + \mathcal{O}\left(n \cdot (p+q)\right)$ Operationen.

- Rechnung ohne Pivotierung erhält die Bandstruktur. Hingegen zerstört Spalten- oder Totalpivotierung die Bandstruktur und führt zum sogenannten "Fill-In" von Matrixeinträgen, die in A noch Null waren.

Für positiv-definite Bandmatrizen vom Typ (p, p) ist der Cholesky-Faktor L eine Bandmatrix vom Typ $(p, 0)$. Der folgende Algorithmus behandelt positiv-definite *Tridiagonalmatrizen.*

2.3.8 Algorithmus: Cholesky-Zerlegung positiv-definiter Tridiagonalmatrizen

Eingabe: Diagonale $(a_1, \ldots, a_n)$ und untere Nebendiagonale $(b_1, \ldots, b_{n-1})$

Berechnung:

(1) $\ell_{1,1} = \sqrt{a_1}, \qquad \ell_{2,1} = \frac{b_1}{\ell_{1,1}}.$

(2) Für $k = 2, \ldots, n-1$

$$\ell_{k,k} = \sqrt{a_k - \ell_{k,k-1}^2}, \qquad \ell_{k+1,k} = \frac{b_k}{\ell_{k,k}}.$$

(3) $\ell_{n,n} = \sqrt{a_n - \ell_{n,n-1}^2}.$

Bemerkung: Der Rechenaufwand für die Berechnung der Cholesky–Zerlegung einer positiv definiten Tridiagonalmatrix (Wurzel zählt als eine Operation wie Mult./Add.) ist $2n$.

2.3.9 Beispiel: Positiv definite Bandmatrizen in den Anwendungen

Die Poisson-Gleichung (auch Potentialgleichung)

$$-\Delta u = f, \qquad (\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2})$$

beschreibt das elektrostatische Potential u bei gegebener Ladungsdichte f in einem Gebiet $\Omega \subset \mathbb{R}^2$. Δ ist ein Differentialoperator 2. Ordnung, der *Laplace-Operator*. Mit Vorgabe von u auf dem Rand $\partial\Omega$ (Randbedingungen) stellt sich das *Dirichlet-Problem*:
Finde $u \in C^2(\Omega)$ mit

$$\begin{aligned} -\Delta u(x,y) &= f(x,y), \quad (x,y) \in \Omega, \\ u(x,y) &= g(x,y), \quad (x,y) \in \partial\Omega. \end{aligned}$$

Beispiel: $\Omega = [0,1]^2$, $f(x,y) = x^2(x-1)(2-6y) + y^2(y-1)(2-6x)$ in Ω, und homogene Randbedingungen $u(x,y) = 0$ für $(x,y) \in \partial\Omega$ ergibt

$$u(x,y) = x^2(1-x)y^2(1-y).$$

Der Funktionsgraph in Matlab/Octave:

```
[xx,yy]=meshgrid(0:.05:1);
mesh(xx,yy,xx.^2.*(1-xx).*yy.^2.*(1-yy))   für 3D-Plot
contour(xx,yy,xx.^2.*(1-xx).*yy.^2.*(1-yy))   für Kontur-Plot
```

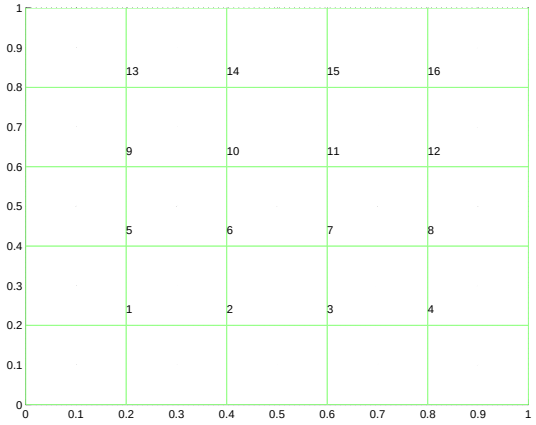
Differenzenverfahren: Eine numerische Näherungslösung u_h wird durch Diskretisierung des Laplace-Operators als "5-Punkte-Stern" erzielt:

$$-\Delta u(x,y) \approx \frac{4u(x,y) - u(x+h,y) - u(x-h,y) - u(x,y+h) - u(x,y-h)}{h^2}$$

an den Stellen $(x,y) = (jh, kh)$ mit $1 \le j,k \le N$ und $h = \frac{1}{N+1}$. Die Nummerierung dieser Stellen ergibt den Lösungsvektor

$$\vec{u} = (u_1, \ldots, u_N, \quad u_{N+1}, \ldots, u_{2N}, \ldots, u_{N^2})^T$$

und den entsprechenden Vektor der rechten Seite $\vec{f}$.



Verwendung der homogenen Randbedingungen führt dann zum linearen Gleichungssystem

$$\begin{pmatrix} B & -\mathrm{I} & & & \\ -\mathrm{I} & B & -\mathrm{I} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathrm{I} & B & -\mathrm{I} \\ & & & -\mathrm{I} & B \end{pmatrix} \vec{u} = h^2 \vec{f}, \qquad B = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{N \times N},$$

mit der $N \times N$ Einheitsmatrix I. Die gesamte Matrix $A \in \mathbb{R}^{N^2 \times N^2}$ ist

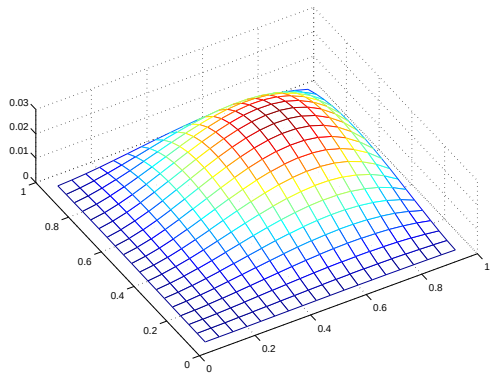
- symmetrische Bandmatrix vom Typ (N, N),
- *schwach diagonaldominant*, d.h. $\sum_{k \neq j} |a_{jk}| \leq |a_{j,j}|$ für alle $j = 1, \ldots, N^2$; weil die Diagonalelemente positiv sind, ist A zumindest positiv-semi-definit;
- in mehreren Zeilen sogar *stark* diagonaldominant.

Man zeigt, dass A sogar positiv-definit ist.

Der Aufwand zur Lösung mit Cholesky-Zerlegung ist $\frac{N^4}{2}$ (im Vergleich zu $\frac{N^6}{6}$ bei voll besetzter Matrix, denn $n = N^2$).

Beachte: Der Cholesky-Faktor L ist Bandmatrix vom Typ $(N, 0)$ mit von Null verschiedenen Elementen in **allen** unteren Nebendiagonalen (sog. "Fill-in").

Das Matlab/Octave-Skript `poisson.m` liefert die diskrete Lösung u_h als $N \times N$-Matrix.



2.4 Lineare Ausgleichsprobleme, QR-Zerlegung

Wir betrachten nun lineare Gleichungssysteme

$$Ax = b \quad \text{mit} \quad A \in \mathbb{R}^{m\times n}, \quad b \in \mathbb{R}^m,$$

von m Gleichungen mit n Unbekannten. Insbesondere müssen wir davon ausgehen, dass das System **nicht lösbar** (also **überbestimmt**) ist, d.h. die Beziehung

$$\text{Rang}(A) < \text{Rang}([A, b])$$

gilt.

An Stelle der Lösung des Gleichungssystems suchen wir alle Vektoren $x \in \mathbb{R}^n$, die das *Lineare Ausgleichsproblem*

$$\|Ax - b\|_2 = \min_{y\in\mathbb{R}^n} \|Ay - b\|_2;$$

lösen, d.h. der Vektor Ax soll den kleinsten euklidischen Abstand von der rechten Seite b unter allen Vektoren Ay annehmen: dieses Problem wird geometrisch dadurch gelöst, dass $Ax = s$ der Fusspunkt des Lotes vom Punkt b auf den Bildraum von A ist. Die Lösungsmenge des linearen Ausgleichsproblems umfasst dann alle $x \in \mathbb{R}^n$ mit $Ax = s$, und diese Menge ist nichtleer.

Bemerkung:

a) $Ax = b$ ist genau dann lösbar, wenn $b \in \text{Bild}(A)$.

b) $\dim \text{Bild}(A) = \text{Rang}(A) \leq \min(m, n)$.

Wir kommen schnell zu einer allgemeinen Beschreibung der Lösungen dieses Problems.

2.4.1 Satz: Lösung des linearen Ausgleichsproblems
Es seien $m, n \in \mathbb{N}$ sowie $A \in \mathbb{R}^{m\times n}$ und $b \in \mathbb{R}^m$.

a) Es existiert stets eine Lösung $x \in \mathbb{R}^n$ des linearen Ausgleichsproblems

$$\|Ax - b\|_2 = \min_{y\in\mathbb{R}^n} \|Ay - b\|_2; \tag{2.3}$$

b) Die Lösungen x von (2.3) sind genau die Lösungen der *Normalengleichung*

$$A^T Ax = A^T b. \tag{2.4}$$

c) Die Lösungsmenge für das lineare Ausgleichsproblem in a) und für die Normalengleichung in b) hat die Form $x + \text{Kern}(A)$, wobei x irgendeine Lösung ist.

Genau im Fall $\text{Rang}(A) = n$ hat das lineare Ausgleichsproblem eine eindeutige Lösung.

Bemerkung zur Linearen Algebra:

- Für eine Matrix $A \in \mathbb{R}^{m\times n}$ ist die Matrix $A^TA \in \mathbb{R}^{n\times n}$ symmetrisch und positiv semi-definit. Es gilt $\text{Rang}(A) = \text{Rang}(A^TA)$.
- Im Fall $\text{Rang}(A) = n$ ist A^TA also symmetrisch und positiv definit, insbesondere invertierbar. Zur Lösung der Normalengleichung kann man die Cholesky-Zerlegung von A^TA verwenden.
- Der Rechenaufwand ist $\frac{mn^2}{2} + \frac{n^3}{6}$ zum Aufstellen (der unteren Hälfte) der symmetrischen Matrix A^TA und der Cholesky-Zerlegung.

Beweis: a) (geometrisch!) Wir bestimmen zur rechten Seite b einen Punkt $s \in \text{Bild}(A)$ so, dass

$$\|b - s\|_2 = \min\{\|b - t\|_2 : t \in \text{Bild}(A)\}.$$

Sodann ist jedes $x \in \mathbb{R}^n$ mit $Ax = s$ (also das Urbild von s) eine Lösung der Ausgleichsaufgabe. Wir wissen aus der Linearen Algebra, dass die Lösungsmenge von $Ax = s$ ein affiner Teilraum der Form $x + \text{Kern}(A)$ ist, wobei x irgendeine der Lösungen ist. Dann ist auch der erste Teil in c) bewiesen.

1. Die Untervektorräume

   $$\text{Bild}(A) = \{Ax : x \in \mathbb{R}^n\} \preceq \mathbb{R}^m, \quad \text{Kern}(A^T) = \{y \in \mathbb{R}^m : A^Ty = 0\} \preceq \mathbb{R}^m$$

   sind orthogonal und erfüllen $\mathbb{R}^m = \text{Bild}(A) \oplus \text{Kern}(A^T)$: denn

   $$\begin{aligned} A^Ty = 0 &\Leftrightarrow (A^Ty)^Tx = 0 \text{ für alle } x \in \mathbb{R}^n \\ &\Leftrightarrow y^T(Ax) = 0 \text{ für alle } x \in \mathbb{R}^n \\ &\Leftrightarrow y \perp \text{Bild}(A). \end{aligned}$$

2. Die rechte Seite $b \in \mathbb{R}^m$ besitzt eine eindeutige Zerlegung

   $$b = s + r \quad \text{mit} \quad s \in \text{Bild}(A),\ r \in \text{Kern}(A^T).$$

   Nach dem Satz des Pythagoras ist für jeden Punkt $t \in \text{Bild}(A)$

   $$\|b-t\|^2 = \|b-s\|^2 + \|s-t\|^2, \quad \text{weil } b - s = r \text{ und } s - t \in \text{Bild}(A) \text{ orthogonal sind.}$$

   Damit haben wir den gewünschten Punkt s gefunden, er ist der Lotfußpunkt des Lots von b auf den Teilraum $\text{Bild}(A)$.

Die Lösungen von (2.3) sind also genau die Lösungen des linearen Gleichungssystems $Ax = s$. Dieses Gleichungssystem besitzt Lösungen wegen $s \in \text{Bild}(A)$, und damit ist Teil a) bewiesen.

b) Jede Lösung x von $Ax = s$ erfüllt die Normalengleichung, denn

$$A^TAx = A^Ts = A^T(b - r) = A^Tb, \quad \text{weil } r \in \text{Kern}(A^T).$$

Umgekehrt sei x eine Lösung von $A^TAx = A^Tb$. Wir wollen zeigen, dass $Ax = s$ gilt, denn dann ist Teil b) bewiesen. Wegen $s \in \text{Bild}(A)$ existiert $y \in \mathbb{R}^n$ mit $s = Ay$. Wie gerade bewiesen wurde, gilt $A^TAy = A^Tb$, und damit folgt

$$\|Ax - s\|_2^2 = \|A(x-y)\|_2^2 = (x-y)^TA^TA(x-y) = (x-y)^T \underbrace{(A^TAx - A^Tb)}_{=0} = 0,$$

also $Ax = s$. Da beide Lösungsmengen in a) und b) übereinstimmen, ist auch der zweite Teil von c) bewíesen.

Der Zusatz ist eine Aussage der Linearen Algebra: Im Fall $\text{Rang}(A) = n$ liefert die Dimensionsformel $\dim \text{Kern}(A) = 0$, also hat die Lösungsmenge genau ein Element.

2.4.2 Anwendungsbeispiele:

a) **Gaußsche Ausgleichsparabel, Methode der kleinsten Fehlerquadrate (siehe `polyfit`):**

Gegeben: $n \in \mathbb{N}$ (Polynomgrad ist $n-1$), $\quad m \geq n$,
Punkte (x_k, y_k) mit $k = 1, 2, \ldots, m$

Gesucht: Polynom $p(x) = c_1 + c_2x + \cdots + c_nx^{n-1}$ mit

$$\sum_{k=1}^{m} |y_k - p(x_k)|^2 \;\to\; \text{min!}$$

Lösung: lineares Ausgleichsproblem $Ax = b$ für den Vektor $x = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ und

$$A = \begin{pmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ & \vdots & & \\ 1 & x_m & \cdots & x_m^{n-1} \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Es gilt $\text{Rang}(A) = n$ genau dann, wenn mindestens n verschiedene Argumente x_k gegeben sind: dann ist die hierzu ausgewählte $n \times n$-Teilmatrix eine Vandermonde-Matrix, hat also vollen Rang.

b) Gaußsche Ausgleichsrechnung mit Basisfunktionen $u_1, \ldots, u_n : [a, b] \to \mathbb{R}$

Gesucht: Funktion $g(x) = c_1u_1(x) + c_2u_2(x) + \cdots + c_nu_n(x)$ mit

$$\sum_{k=1}^{m} |y_k - g(x_k)|^2 \;\to\; \text{min!}$$

Lösung: lineares Ausgleichsproblem $Ax = b$ für den Vektor $x = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ und

$$A = \begin{pmatrix} u_1(x_1) & u_2(x_1) & \cdots & u_n(x_1) \\ u_1(x_2) & u_2(x_2) & \cdots & u_n(x_2) \\ & \vdots & & \\ u_1(x_m) & u_2(x_m) & \cdots & u_n(x_m) \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

Ob $\text{Rang}(A) = n$ gilt, hängt von den Funktionen $u_1, \ldots, u_n$ und der Lage der Stützstellen $x_1, \ldots, x_m$ ab: es gilt $\text{Rang}(A) = n$ genau dann, wenn keine nichttriviale Linearkombination

$$g = \sum_{j=1}^{n} c_j u_j, \qquad (c_1, \ldots, c_n)^T \in \mathbb{R}^n \setminus \{0\},$$

in allen x_k, $1 \le k \le m$, den Wert Null annimmt. Ist $u_1, \ldots, u_n$ z.B. ein *Tschebyscheff-System* ($\rightarrow$ Numerik II), so gilt die gleiche Aussage wie für Polynome: es genügt, dass mindestens n verschiedene Stellen x_k vorliegen.

c) Nichtlineare Ausgleichsrechnung, Koordinatentransformation

Anstatt des in b) betrachteten Ansatzes werden oft *nichtlineare* Ansätze, z.B.

$$g(x) = \frac{c_1}{1 + c_2 x}, \qquad c_1, c_2 \in \mathbb{R},$$

zur Approximation der Daten (x_k, y_k), $1 \le k \le n$, verwendet. Dieser Ansatz ist nichtlinear in Bezug auf die unbekannten Koeffizienten c_1, c_2. Durch Koordinaten-Transformation (von y und/oder x) erzeugt man einen linearen Ansatz

$$\frac{1}{g(x)} = \frac{1}{c_1} + \frac{c_2}{c_1} x$$

für die neuen "Variablen"

$$\widetilde{c}_1 = \frac{1}{c_1}, \quad \widetilde{c}_2 = \frac{c_2}{c_1}$$

und die transformierten Daten $(\widetilde{x}_k, \widetilde{y}_k)$ mit

$$\widetilde{x}_k = x_k, \quad \widetilde{y}_k = \frac{1}{y_k}, \qquad 1 \le k \le n.$$

Man führt die Gaußsche Ausgleichsrechnung für die transformierten Größen durch und bestimmt daraus die Funktion g. (Übung!)

2.4.1 Lösung mit der QR-Zerlegung

Wir haben bereits die geometrische Deutung der Ausgleichsaufgabe

$$\|Ax - b\|_2 = \min_{y \in \mathbb{R}^n} \|Ay - b\|_2$$

mit Hilfe der Orthogonalprojektion im Beweis von Satz 2.4.1 kennengelernt. Wir wollen nun versuchen, die Lösung(en) ohne das Aufstellen der Matrix $A^T A$ zu berechnen. Dazu wird wesentlich sein, dass die Multiplikation des Abstandsvektors $Ay - b$ mit einer Orthogonalmatrix Q die Länge nicht verändert, also gilt

$$\|Ay - b\|_2 = \|Q^T A y - Q^T b\|_2 \quad \text{für alle} \quad Q \in O(m). \tag{2.5}$$

Falls es gelingt, ein $Q \in O(m)$ so zu bestimmen, dass

$$Q^T A = \begin{pmatrix} R \\ \hline 0_{(m-n)\times n} \end{pmatrix} \quad \text{mit} \quad R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n},$$

gilt, so wird mit der Bezeichnung $Q^T b = \binom{c}{d}$ mit $c \in \mathbb{R}^n$, $d \in \mathbb{R}^{m-n}$

$$\|Ay - b\|_2^2 = \left\| \begin{pmatrix} R \\ \hline 0_{(m-n)\times n} \end{pmatrix} y - \begin{pmatrix} c \\ \hline d \end{pmatrix} \right\|_2^2 = \|Ry - c\|_2^2 + \|d\|_2^2.$$

Wir beobachten, dass

- der Term $\|d\|_2^2$ bei der Minimierung über $y \in \mathbb{R}^n$ keine Rolle spielt: er trägt am Ende zum Residuum $\|Ax - b\|_2^2$ bei, kann aber nicht vermieden werden.
- die Bestimmung des Minimums x im linearen Ausgleichsproblem (2.3) sich vereinfacht zu
$$\|Rx - c\|_2 = \min_{y\in\mathbb{R}^n} \|Ry - c\|_2,$$
wobei R eine obere $n \times n$-Dreiecksmatrix ist.

Falls R invertierbar ist (also alle Diagonalelemente ungleich Null sind), erhalten wir sofort die eindeutige Lösung $x = R^{-1}c$ des linearen Ausgleichsproblems (2.3). Die Berechnung von x erfolgt durch Rückwärtseinsetzen in 2.2.6.

Wir formulieren die QR-Zerlegung einer $m \times n$-Matrix gleichzeitig für reelle und komplexe Matrizen.

2.4.3 Satz: QR-Zerlegung
Es sei $A \in \mathbb{K}^{m\times n}$ mit $\mathrm{Rang}(A) = n$ gegeben. Dann existiert eine Orthogonalmatrix $Q \in O(m)$ (falls $\mathbb{K} = \mathbb{R}$) bzw. unitäre Matrix $Q \in U(m)$ (falls $\mathbb{K} = \mathbb{C}$) sowie eine obere Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ mit reellen Diagonalelementen $r_{jj} > 0$, $1 \le j \le n$, so dass

$$A = Q \begin{pmatrix} R \\ \hline 0_{(m-n)\times n} \end{pmatrix}$$

gilt. Dabei gilt:

(i) Die Matrix R sowie die ersten n Spalten von Q sind eindeutig bestimmt.

(ii) Die ersten n Spalten von Q sind eine Orthonormalbasis von $\mathrm{Bild}(A)$.

(iii) Die letzten $m - n$ Spalten von Q sind eine Orthonormalbasis von $\mathrm{Kern}(A^*)$.

Beweis und erste Konstruktion von Q: Die Matrix Q soll eine $m \times m$-Matrix sein, deren Spalten eine komplette ONB des $\mathbb{K}^m$ bilden. Die ersten n Spalten $q_1, \ldots, q_n$ konstruieren wir als die Vektoren im Gram-Schmidt-Orthonormalisierungsverfahren, das ausgeht von den Spalten $a_1, \ldots, a_n \in \mathbb{K}^m$ von A:

- $q_1 = \frac{1}{\|a_1\|} \cdot a_1$,
- $q_2 = \frac{1}{\|a_2 - \langle a_2, q_1\rangle q_1\|}(a_2 - \langle a_2, q_1\rangle q_1)$,

usw. Weil die Spalten von A linear unabhängig sind (wegen $\mathrm{Rang}(A) = n$), erhält man eine ONB von $\mathrm{Bild}(A)$. Man erkennt sogar, dass für die ersten j Vektoren gilt

$$\mathrm{Spann}(a_1, \ldots, a_j) = \mathrm{Spann}(q_1, \ldots, q_j), \qquad 1 \le j \le n.$$

Dies liefert schon die gewünschte Matrix R. Denn das spaltenweise Aufschreiben des Matrixprodukts $A = (a_1, \ldots, a_n) = (q_1, \ldots, q_n)R$ mit oberer Dreiecksmatrix $R \in \mathbb{K}^{n\times n}$ ergibt

$$\begin{aligned} a_1 &= r_{11}q_1, \\ a_2 &= r_{12}q_1 + r_{22}q_2, \\ &\vdots \\ a_n &= r_{1n}q_1 + \cdots + r_{n-1,n}q_{n-1} + r_{nn}q_n. \end{aligned} \tag{2.4.2a}$$

Die Koeffizienten r_{jk} ergeben sich durch Umstellen der Gram-Schmidt Gleichungen als

$$\begin{aligned} &r_{11} = \|a_1\|, \\ &r_{12} = \langle a_2, q_1\rangle q_1, \quad r_{22} = \|a_2 - r_{12}q_1\| \\ &\vdots \\ &r_{1n} = \langle a_n, q_1\rangle, \ldots, r_{n-1,n} = \langle a_n, q_{n-1}\rangle, \quad r_{nn} = \|a_n - r_{1n}q_1 - \cdots - r_{n-1,n}q_{n-1}\|. \end{aligned}$$

Hierbei sind alle $r_{jj} > 0$ für $1 \le j \le n$. Damit ist der wesentliche Teil der QR-Zerlegung konstruiert.

Wir wählen zuätzlich noch eine beliebige ONB $(q_{n+1}, \ldots, q_m)$ im orthogonalen Komplement $\mathrm{Bild}(A)^\perp = \mathrm{Kern}(A^*)$ und bilden die orthogonale (bzw. unitäre) Matrix $Q = (q_1, \ldots, q_m) \in \mathbb{K}^{m\times m}$. Am Matrixprodukt ändert sich nichts, wenn wir Nullen an die Spalten von R anhängen,

$$A = (q_1, \ldots, q_n)R = Q \begin{pmatrix} R \\ \hline 0_{(m-n)\times n} \end{pmatrix}.$$

Die Eindeutigkeit von R mit $r_{jj} > 0$ und der ersten n Vektoren $q_1, \ldots, q_n$ folgt aus den Gleichungen (2.4.2a).

2.4.4 Bemerkung:

a) Mit $Q^*Q = I_m$ erhalten wir, dass die Matrix R in Satz 2.4.2

$$R^*R = R^*Q^*QR = A^*A$$

erfüllt. Also ist $L = R^*$ der Cholesky-Faktor von A^*A. Die QR-Zerlegung liefert demnach auch die Cholesky-Zerlegung von A^*A, OHNE diese Matrix aufzustellen.

b) Matlab liefert mit `[Q,R]=qr(A)` die volle $m \times m$-Matrix Q und die um Nullen ergänzte Matrix R, dagegen mit `[Q,R]=qr(A,0)` die "economy size" Version der ersten n Spalten von Q und die $n \times n$-Matrix R.

Die Gram-Schmidt-Orthonormalisierung ist numerisch nicht besonders stabil: Nach wenigen Schritten ($n \approx 5$) beobachtet man bereits den Verlust der paarweisen Orthogonalität der q_j. Daher wird im Folgenden eine **2. Konstruktion** der QR-Zerlegung entwickelt, die numerisch stabiler ist (mit Diagonalelementen $r_{jj} \neq 0$ nicht notwendig positiv).

2.4.5 Definition: Householder-Transformation
Für einen Einheitsvektor $v \in \mathbb{K}^m$, $\|v\|_2 = 1$, heißt die Matrix

$$H_v = I_m - 2vv^* = I - 2 \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} (\overline{v_1}, \ldots, \overline{v_m}) \in \mathbb{K}^{m \times m}$$

eine *Householder-Transformation.*

2.4.6 Hilfssatz: Eigenschaften von H_v:
(siehe Übungen)

(i) $H_v = H_v^* = H_v^{-1}$; also ist H_v hermitesch und orthogonal (für $\mathbb{K} = \mathbb{R}$) bzw. unitär (für $\mathbb{K} = \mathbb{C}$).

(ii) $H_v^2 = I_m$.

(iii) $H_v y = y$ für alle $y \in (\mathrm{Span}(v))^\perp$; deshalb ist $n-1$-facher Eigenwert.

(iv) $H_v y = -y$ für alle $y \in (\mathrm{Span}(v))$; deshalb ist 1 ein einfacher Eigenwert.

Für $\mathbb{K} = \mathbb{R}$ beschreibt H_v eine Spiegelung des $\mathbb{R}^m$ an der Hyperebene $(\mathrm{Span}(v))^\perp$.

Die Householder-Matrizen werden zur Elimination von Spalten unterhalb der Diagonalen einer Matrix $A \in \mathbb{K}^{m \times n}$ eingesetzt. Wir schreiben zuerst einen einzelnen Eliminationsschritt dazu auf.

2.4.7 Lemma: Householder-Transformation zur Elimination
Es sei $y \in \mathbb{R}^m \setminus \{0\}$. Mit e_1 bezeichnen wir den ersten kanonischen Einheitsvektor.
Wir setzen

$$\alpha = \mathrm{sign}(y_1)\ \|y\|_2, \quad \widetilde{v} = y + \alpha e_1, \quad v = \frac{1}{\|\widetilde{v}\|}\ \widetilde{v}.$$

Dann gilt

$$H_v \cdot y = -\alpha e_1.$$

Bemerkung zur Wahl von α:

(i) Alternativ kann man auch $\alpha = -\text{sign}(y_1)\ \|y\|_2$ wählen; dies ergibt jedoch im Spezialfall $y = ce_1$ den Vektor $\widetilde{v} = 0$, der zur Konstruktion von H_v nicht zugelassen ist.

Dagegen vermeidet unsere Wahl im Lemma mit $\alpha = \text{sign}(y_1)\ \|y\|_2$ die Stellenauslöschung, weil in der ersten Koordinate von

$$\widetilde{v} = (y_1 + \alpha, y_2, \ldots, y_m)^T$$

Zahlen mit gleichem Vorzeichen addiert werden.

(ii) Im Sonderfall $y_1 = 0$ setzen wir $\alpha = \|y\|_2$ oder $\alpha = -\|y\|_2$.

(iii) Im Komplexen wählt man $\alpha = e^{i\arg(y_1)}\ \|y\|_2$ und erhält numerisch stabile Werte für $\widetilde{v} = y + \alpha e_1$.

2.4.8 Bemerkung: Praktischer Umgang mit Householder-Transformationen

(i) Bei Berechnungen $H_v \cdot A$ stellt man die Matrix H_v nicht auf, sondern berechnet die einzelnen Spalten

$$H_v a_j = (I_m - 2vv^*)a_j = a_j - 2(v^* a_j)\ v = a_j - 2\langle a_j, v\rangle\ v.$$

Dies ist eine elementare *Spalten*umformung, fast wie beim Gauß-Algorithmus. Auch auf die Normierung des Vektors $\widetilde{v}$ kann verzichtet werden gemäß

$$H_v a_j = a_j - \frac{2\langle a_j, \widetilde{v}\rangle}{\|\widetilde{v}\|^2}\ \widetilde{v}.$$

(ii) Im Reellen: für $\alpha = \pm\text{sign}(y_1)\ \|y\|$ ist $\alpha y_1 = \pm|y_1|\ \|y\|$. Also gilt für $\widetilde{v} = y + \alpha e_1$

$$\|\widetilde{v}\|^2 = \|y\|^2 + 2\alpha y_1 + |\alpha|^2 = 2\|y\|(\|y\| \pm |y_1|).$$

Hiermit rechnet man für $x \in \mathbb{R}^m$ zuerst

$$\beta = \frac{2\langle x, \widetilde{v}\rangle}{\|\widetilde{v}\|^2} = \frac{\langle x, y\rangle + \alpha x_1}{\|y\|(\|y\| \pm |y_1|)}$$

und dann

$$H_v x = x - \beta\widetilde{v} = x - \beta y - \alpha\beta e_1.$$

Speziell ergibt sich für $x = y$ damit

$$\beta = 1, \qquad H_v y = -\alpha e_1.$$

2.4.9 Beispiel:

Elimination der 1. Spalte von $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$ durch Householder-Transformation:

$$y = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \alpha = \|y\| = \sqrt{2}, \quad \widetilde{v} = \begin{pmatrix} 1+\sqrt{2} \\ 0 \\ 1 \end{pmatrix}.$$

Berechne $A^{(1)} = H_v A$ spaltenweise:

$$\begin{array}{c|ccc} \beta & 1 & \sqrt{2} & 2-\sqrt{2} \\ \hline & -\sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ & 0 & 1 & 1 \\ & 0 & -\sqrt{2} & \sqrt{2} \end{array}$$

Zur Probe beachte man, dass die euklidische Norm der Spalten sich nicht ändert.

2.4.10 Algorithmus: QR-Zerlegung mit Householder-Matrizen

Es sei $A \in \mathbb{R}^{m\times n}$ mit $m \geq n = \mathrm{Rang}(A)$. Wir berechnen Matrizen

$$A^{(j)} = H_{v^{(j)}} \cdots H_{v^{(1)}} A = \left(\begin{array}{ccc|cc} r_{11} & \cdots & & \cdots & r_{1n} \\ & \ddots & & & \vdots \\ & & r_{jj} & \cdots & r_{jn} \\ \hline & & & & \\ & 0 & & \widetilde{A}^{(j)} & \end{array}\right), \qquad j = 1, \ldots, n,$$

mit einer im nächsten Schritt zu bearbeitenden Teilmatrix

$$\widetilde{A}^{(j)} = \begin{pmatrix} \widetilde{a}^{(j)}_{1,1} & \cdots & \widetilde{a}^{(j)}_{1,n-j} \\ \vdots & & \vdots \\ \widetilde{a}^{(j)}_{m-j,1} & \cdots & \widetilde{a}^{(j)}_{m-j,n-j} \end{pmatrix} \in \mathbb{R}^{(m-j)\times(n-j)}$$

($\widetilde{A}^{(n)}$ tritt nicht auf; falls $m = n$, so endet die Berechnung bei $j = n-1$).

Der Vektor $v^{(j)} \in \mathbb{R}^m$ der Householder-Matrix $H_{v^{(j)}}$ ist ein Einheitsvektor der Form

$$v^{(j)} = (0, \ldots, 0, v^{(j)}_j, \ldots, v^{(j)}_m)^T.$$

- Initialisierung: $A = A^{(0)} = \widetilde{A}^{(0)}$
- Für $j = 1, \ldots, n$ (bzw. $j = 1, \ldots, n-1$ falls $m = n$)

1. $y \in \mathbb{R}^{m-j+1}$ bezeichne die erste Spalte von $\widetilde{A}^{(j-1)}$, setze

$$\alpha_j = \operatorname{sign}(y_1)\ \|y\|_2, \quad \widetilde{v}^{(j)} = (y_1 + \alpha_j, y_2, \ldots, y_{m-j+1})^T \in \mathbb{R}^{m-j+1},$$
$$v^{(j)} = \frac{1}{\|\widetilde{v}^{(j)}\|}\ (0, \ldots, 0, \widetilde{v}_1^{(j)}, \ldots, \widetilde{v}_{m-j+1}^{(j)})^T \in \mathbb{R}^m.$$

2. berechne $A^{(j)} = H_{v^{(j)}} A^{(j-1)}$ wie folgt:

– Zeilen und Spalten 1 bis $j-1$ von $A^{(j-1)}$ bleiben unverändert,

– Ersetze Spalte 1 von $\widetilde{A}^{(j-1)}$ durch $(-\alpha_j, 0, \ldots, 0)^T$; insbesondere ist $r_{jj} = -\alpha_j$.

– Für $k = 2, \ldots, n-j+1$ (nur falls $j < n$) ersetze Spalte k von $\widetilde{A}^{(j-1)}$ durch

$$\widetilde{a}_k^{(j-1)} - \beta_{jk}\widetilde{v}^{(j)} \quad \text{mit} \quad \beta_{jk} = \frac{2\langle \widetilde{a}_k^{(j-1)}, \widetilde{v}^{(j)} \rangle}{\|\widetilde{v}^{(j)}\|^2}.$$

Ergebnis: $A = \underbrace{H_{v^{(1)}} \cdots H_{v^{(n)}}}_{=Q \in \mathbb{R}^{m\times m}} \cdot A^{(n)}$ und $A^{(n)} = \begin{pmatrix} R \\ 0_{m-n,n} \end{pmatrix}$.

2.4.11 Bemerkung:

a) Rechenaufwand für $A^{(j)}$:

– Norm der Spalte $\widetilde{a}_1^{(j-1)} \in \mathbb{R}^{m-j+1}$: $m-j+1$ Mult./Add. und eine Wurzel,

– $\beta_{j,k}$, $k = 2, \ldots, n-j+1$: jeweils $m-j+1$ Mult./Add. und 1 Division (gleicher Nenner),

– Update der Spalten $k = 2, \ldots, n-j+1$ von $\widetilde{A}^{(j-1)}$: jeweils $m-j+1$ Mult./Add.

Insgesamt für $A^{(j)}$ also

$$(m-j+1)(2(n-j)+1) \text{ Mult./Add.}, \quad n-j \text{ Div.}, \quad 1 \text{ Wurzel.}$$

Gesamter Rechenaufwand der QR-Zerlegung:

$$\frac{2}{3}n^3 + n^2(m-n) + \mathcal{O}(n^2 + n(m-n))$$

b) Die QR-Zerlegung wird auch zur Lösung linearer Gleichungssysteme $Ax = b$ eingesetzt (also für reguläres $A \in \mathbb{R}^{n\times n}$). Ihr Vorteil ist, dass die Konditionszahl (bzgl. der Spektral- und der Frobeniusnorm)

$$\operatorname{cond}(A) = \operatorname{cond}(A^{(1)}) = \cdots = \operatorname{cond}(A^{(n)}) = \operatorname{cond}(R)$$

erhalten bleibt. Die Lösung von $Rx = Q^T b$ mittels Rücksubstitution erfolgt also mit einer oberen Dreiecksmatrix R mit der gleichen Kondition wie A.

Der Nachteil ist der doppelt so hohe Rechenaufwand gegenüber der LR-Zerlegung.

2.4.12 Beispiel:

QR-Zerlegung der Matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$ durch Householder-Transformationen:

Der 1. Schritt wurde in Beispiel 2.4.9 mit

$$v^{(1)} = \frac{1}{\sqrt{4+2\sqrt{2}}} \begin{pmatrix} 1+\sqrt{2} \\ 0 \\ 1 \end{pmatrix}$$

berechnet. Für den 2. Schritt setze

$$y = \begin{pmatrix} 1 \\ -\sqrt{2} \end{pmatrix}, \quad \alpha_2 = \|y\| = \sqrt{3}, \quad \widetilde{v}^{(2)} = \begin{pmatrix} 1+\sqrt{3} \\ -\sqrt{2} \end{pmatrix}.$$

Berechne $A^{(2)} = H_{v^{(2)}} A^{(1)}$ durch Update des unteren rechten 2×2-Blocks:

β	$-$	1	$-1+\frac{2\sqrt{3}}{3}$
	$-\sqrt{2}$	$-\sqrt{2}$	$-\sqrt{2}$
	0	$-\sqrt{3}$	$\frac{\sqrt{3}}{3}$
	0	0	$\frac{2\sqrt{6}}{3}$

Es ist

$$v^{(2)} = \frac{1}{\sqrt{6+2\sqrt{3}}} \begin{pmatrix} 0 \\ 1+\sqrt{3} \\ -\sqrt{2} \end{pmatrix}$$

2.4.13 Bemerkung: Vergleich mit Satz 2.4.2
Die Diagonaleinträge von R in der (reellen) QR-Zerlegung mit Householder-Matrizen sind $r_{jj} = -\alpha_j$. Um positive Diagonaleinträge $r_{jj} > 0$ wie in Satz 2.4.2 zu erhalten, muss man im j-ten Schritt die Wahl

$$\alpha_j = -\|y\|$$

treffen, ohne Berücksichtigung des Vorzeichens der Komponente y_1. Dadurch kann man aber Stellenauslöschung in der 1. Komponente von $\widetilde{v}^{(j)}$ erhalten.

Achtung: Dies führt im Spezialfall $y = ce_1$ mit $c > 0$ zu $\widetilde{v}_j = 0$, ist also gar nicht zulässig. Für diesen Fall setzt man statt $H_{v^{(j)}}$ die Einheitsmatrix I ein (, die keine Householder-Matrix ist, warum?), führt also keinen Update der Spalten von $\widetilde{A}^{(j-1)}$ durch.

2.5 Nicht-reguläre Systeme, Singulärwertzerlegung

Das *Lineare Ausgleichsproblem* (2.3) mit $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $\mathrm{Rang}(A) < n$ besitzt unendlich viele Lösungen

$$\mathbb{L} = \{x \in \mathbb{R}^n : A^T Ax = A^T b\}.$$

Wir erweitern die Aufgabenstellung und bestimmen hiervon die *Minimallösung* $x_* \in \mathbb{L}$ mit

$$\|x_*\|_2 = \min\{\|x\|_2 : x \in \mathbb{L}\}. \tag{2.6}$$

(Geometrisch: $\|x^*\|_2$ ist der Abstand des affinen Raumes $\mathbb{L} = x+\mathrm{Kern}(A)$ vom Nullpunkt.)

Verfahren: Es sei $r = \mathrm{Rang}(A) \le \min\{m, n\}$. Falls es gelingt, orthogonale Matrizen $U \in O(m)$, $V \in O(n)$ so zu bestimmen, dass

$$U^T AV = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) =: \Sigma \in \mathbb{R}^{m\times n}$$

gilt, wobei die Diagonalelemente $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$ und sonst nur Nullen in Σ stehen, so ist durch Multiplikation mit Orthogonalmatrizen

$$\|Ax - b\|_2 = \|U^T Ax - U^T b\|_2 = \|U^T AVV^T x - U^T b\|_2 = \|\Sigma V^T x - U^T b\|_2.$$

Setzen wir $y = V^T x \in \mathbb{R}^n$ und $c = U^T b \in \mathbb{R}^m$, so erhalten wir

$$\|Ax - b\|_2^2 = \sum_{k=1}^{r} (\sigma_k y_k - c_k)^2 + \sum_{k=r+1}^{m} c_k^2. \tag{2.7}$$

Minimierung über alle $x \in \mathbb{R}^n$ ist dasselbe wie Minimierung über $y = V^T x \in \mathbb{R}^n$, weil V regulär ist. Alle Lösungen des linearen Ausgleichsproblems (2.3) sind daher gegeben durch

$$y_k = \frac{c_k}{\sigma_k} = \frac{\langle b, u_k\rangle}{\sigma_k}, \qquad 1 \le k \le r,$$
$$y_k \ \text{ beliebig}, \qquad r+1 \le k \le n.$$

Hierbei bezeichnen wir die Spalten von U mit $u_1, \ldots, u_n$. Die Umrechnung in die x-Koordinaten wird mit $x = Vy$ erzielt, denn V ist ja eine orthogonale Matrix. Wir bezeichnen die Spalten von V mit $v_1, \ldots, v_n$ und erhalten sämtliche Lösungen des linearen Ausgleichsproblems

$$x = \sum_{k=1}^{r} \frac{\langle b, u_k\rangle}{\sigma_k}\, v_k + \sum_{k=r+1}^{n} y_k v_k \tag{2.8}$$

mit beliebigen $y_{k+1}, \ldots, y_n \in \mathbb{R}$. Das Quadrat der euklidischen Norm der Lösungsvektoren in (2.8) ist

$$\|x\|_2^2 = \sum_{k=1}^{r} \left(\frac{\langle b, u_k \rangle}{\sigma_k} \right)^2 + \sum_{k=r+1}^{n} y_k^2,$$

weil die Vektoren v_k eine Orthonormalbasis des $\mathbb{R}^n$ sind. Die eindeutige Lösung des erweiterten Problems (2.6) ist deshalb

$$x_* = \sum_{k=1}^{r} \frac{\langle b, u_k \rangle}{\sigma_k} \, v_k.$$

2.5.1 Satz: Singulärwertzerlegung
Es seien $m, n \in \mathbb{N}$ sowie $A \in \mathbb{R}^{m \times n}$ gegeben, weiter sei $r = \mathrm{Rang}(A)$.

Dann existieren orthogonale Matrizen $U \in O(m)$ und $V \in O(n)$ und reelle Zahlen

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0,$$

so dass

$$A = U \Sigma V^T \quad \text{mit} \quad \Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array} \right) \in \mathbb{R}^{m \times n}.$$

Die Zahlen σ_k, $1 \leq k \leq r$, sind eindeutig bestimmt und heißen *Singulärwerte* von A.

Bemerkung: Im Fall $r = \mathrm{Rang}(A) = n < m$ hat die Matrix Σ die Gestalt

$$\Sigma = \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & 0_{(m-n) \times n} & \end{array} \right).$$

und im Fall $r = \mathrm{Rang}(A) = m < n$ ist

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0_{m \times (n-m)} \\ & & \sigma_m & \end{array} \right).$$

Beweismethode 1: Eigenwerte und -vektoren von $A^T A$ und $A A^T$:

Es sei $r = \mathrm{Rang}(A)$. Die Matrix $A^T A \in \mathbb{R}^{n \times n}$ besitzt den Rang r, ist symmetrisch und positiv semi-definit. Zu ihren Eigenwerten

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \qquad \lambda_{r+1} = \cdots = \lambda_n = 0$$

wählen wir eine ONB des $\mathbb{R}^n$ aus Eigenvektoren $(v_1, \ldots, v_n)$. Wir setzen $u_j = \frac{1}{\sqrt{\lambda_j}} A v_j$ für $j = 1, \ldots, r$. Es gilt

$$u_j^T u_k = \frac{1}{\sqrt{\lambda_j \lambda_k}} (Av_j)^T (Av_k) = \frac{1}{\sqrt{\lambda_j \lambda_k}} v_j^T (A^T A v_k) = \frac{\sqrt{\lambda_k}}{\sqrt{\lambda_j}} v_j^T v_k = \begin{cases} 1 & , \text{ falls } j = k, \\ 0 & , \text{ falls } j \neq k, \end{cases}$$

also bilden $u_1, \ldots, u_r$ ein Orthonormalsystem in $\mathbb{R}^m$. Wir ergänzen zu einer ONB $(u_1, \ldots, u_m)$ des $\mathbb{R}^m$, setzen $\sigma_j = \sqrt{\lambda_j}$, $j = 1, \ldots, r$, und erhalten die Singulärwertzerlegung

$$U^T A V = U^T \begin{pmatrix} \sigma_1 u_1 & \cdots & \sigma_r u_r & 0 & \cdots & 0 \end{pmatrix} = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right)$$

mit den Orthogonalmatrizen $U = (u_1, \ldots, u_m)$ und $V = (v_1, \ldots, v_n)$.

Beweismethode 2: vollständige Induktion nach $r = \mathrm{Rang}(A)$:

1. Im Fall $r = 0$ ist $A = 0$, also gilt $A = U \Sigma V$ für alle Orthogonalmatrizen $U \in O(m)$ und $V \in O(n)$ mit der Nullmatrix $\Sigma = 0 \in \mathbb{R}^{m \times n}$.

2. Nun sei $r \geq 1$. Dann ist $A \neq 0$ und die Spektralnorm $\sigma_1 = \|A\|_2 > 0$. Wähle einen Einheitsvektor $v_1 \in \mathbb{R}^n$ mit $\|Av_1\|_2 = \sigma_1$ und setze $u_1 = \frac{1}{\sigma_1} v_1$. Sodann ergänze zu Orthogonalmatrizen $V = (v_1, \ldots, v_n) \in O(n)$ und $U = (u_1, \ldots, u_m) \in O(m)$. Dann gilt

$$U^T A V = U^T (\sigma_1 u_1,\ Av_2, \ldots,\ Av_n) = \left(\begin{array}{c|c} \sigma_1 & w^T \\ \hline 0 & B \\ & \end{array}\right) \tag{2.9}$$

mit $w \in \mathbb{R}^{n-1}$ und $B \in \mathbb{R}^{(m-1) \times (n-1)}$. Aus $\mathrm{Rang}(A) = \mathrm{Rang}(U^T A v) = r$ folgt direkt $\mathrm{Rang}(B) = r - 1$. Wir zeigen $w = 0$: Transposition der Gleichung (2.9) und schreiben der 1. Spalte ergibt

$$V^T A^T U e_1 = \begin{pmatrix} \sigma_1 \\ w \end{pmatrix}$$

und damit

$$\sigma_1 = \|A\|_2 = \|A^T\|_2 = \|V^T A^T U\|_2 \geq \|V^T A^T U e_1\|_2 = \left\| \begin{pmatrix} \sigma_1 \\ w \end{pmatrix} \right\|_2 = \sqrt{\sigma_1^2 + \|w\|^2}.$$

Also folgt $w = 0$.

Die Induktionsannahme für Matrizen vom Rang $r - 1$ liefert Matrizen $\widetilde{U} \in O(m-1)$ und $\widetilde{V} \in O(n-1)$ so, dass

$$B = \widetilde{U} \widetilde{\Sigma} \widetilde{V}^T \quad \text{mit} \quad \widetilde{\Sigma} = \left(\begin{array}{ccc|c} \sigma_2 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right) \in \mathbb{R}^{(m-1) \times (n-1)}.$$

Hierbei wurde $\mathrm{Rang}(B) = r-1$ für die Angabe der Singulärwerte $\sigma_2 \geq \cdots \geq \sigma_r > 0$ verwendet. (Dass $\sigma_1 \geq \sigma_2$ gilt, erkennt man anhand von (2.9) und $\sigma_2 = \|B\|_2$.) Insgesamt ergibt sich mit (2.9) die Singulärwertzerlegung

$$A = U \left(\begin{array}{c|c} \sigma_1 & 0 \\ \hline 0 & B \end{array}\right) V^T = \underbrace{U \left(\begin{array}{c|c} 1 & \\ \hline & \widetilde{U} \end{array}\right)}_{\in O(m)} \left(\begin{array}{c|c} \sigma_1 & \\ \hline & \widetilde{\Sigma} \end{array}\right) \underbrace{\left(\begin{array}{c|c} 1 & \\ \hline & \widetilde{V}^T \end{array}\right) V^T}_{\in O(n)}.$$

2.5.2 Bemerkung: In den obigen Überlegungen treten einige einfache Beziehungen zu Tage für U, V und Σ. Es sei $r = \mathrm{Rang}(A)$ und $A = U\Sigma V^T$ die Singulärwertzerlegung von $A \in \mathbb{R}^{m\times n}$.

(i) Die Spektralnorm von A ist $\|A\|_2 = \sigma_1$, also der größte Singulärwert von A. Die Frobeniusnorm ist $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$.

(ii) Es gilt
$$\mathrm{Kern}(A) = \mathrm{Kern}(A^TA) = \mathrm{Span}(v_{r+1}, \ldots, v_n),$$
$$\mathrm{Kern}(A^T) = (\mathrm{Bild}(A))^\perp = \mathrm{Span}(u_{r+1}, \ldots, u_m).$$

(iii) Die Vektoren $v_1, \ldots, v_r$ sind Eigenvektoren von A^TA, und die Vektoren $u_1, \ldots, u_r$ sind Eigenvektoren von AA^T, zu den gleichen Eigenwerten $\lambda_j = \sigma_j^2 > 0$.

(iv) Die Singulärwertzerlegung von A kann geschrieben werden als
$$A = \sum_{j=1}^{r} \sigma_j \, u_j v_j^T;$$
hierbei ist $u_jv_j^T$ eine $m\times n$-Matrix vom Rang 1, die Summe hat dann den Rang r. Dies verallgemeinert die Spektraldarstellung symmetrischer Matrizen aus der Linearen Algebra.

Als Algorithmus zur Berechnung der Singulärwertzerlegung wird das iterative Verfahren von Golub und Reinsch empfohlen, siehe G. Golub, C. van Loan, *Matrix Computations*, Abschnitt 8.6, Johns Hopkins University Press; 3. Auflage, 1996. Der Matlab-Befehl lautet `[U,S,V]=svd(A)`.

Wir wollen noch drei Ergänzungen angeben, die auf der SVD aufbauen.

2.5.3 Definition
Gegeben sei $A \in \mathbb{R}^{m\times n}$ mit der Singulärwertzerlegung $A = U\Sigma V^T$. Die *Moore-Penrose-Pseudoinverse* $A^+ \in \mathbb{R}^{n\times m}$ von A (Matlab/Octave `pinv`) ist

$$A^+ = V\Sigma^+ U^T \quad \text{mit} \quad \Sigma^+ = \left(\begin{array}{ccc|c} \frac{1}{\sigma_1} & & & \\ & \ddots & & 0 \\ & & \frac{1}{\sigma_r} & \\ \hline & 0 & & 0 \end{array}\right) \in \mathbb{K}^{n\times m}.$$

Bemerkung:

(i) Man rechnet die charakterisierenden Eigenschaften

$$AA^+ \text{ und } A^+A \text{ symmetrisch}, \qquad AA^+A = A, \quad A^+AA^+ = A^+$$

leicht nach. (Übungen!)

(ii) A^+ kann geschrieben werden als

$$A^+ = \sum_{j=1}^{r} \frac{1}{\sigma_j} v_j u_j^T.$$

Also lässt sich die Minimallösung des linearen Ausgleichsproblems in (2.6) schreiben als

$$x_* = A^+ b.$$

(iii) Falls A quadratisch und regulär ist, gilt $A^+ = A^{-1}$.
Im Fall $\text{Rang}(A) = n$ ist $A^+ = (A^TA)^{-1}A^T$.
Im Fall $\text{Rang}(A) = m$ ist $A^+ = A^T(AA^T)^{-1}$.

(iv) Im Fall $\text{Rang}(A) = n$ erhält man A^+ auch aus der QR-Zerlegung $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$. Es gilt

$$A^+ = (R^{-1} \mid 0)Q^T.$$

Damit erhält man den größten und kleinsten Singulärwert auch aus der QR-Zerlegung gemäß

$$\sigma_1 = \|R\|_2, \qquad \sigma_n = \frac{1}{\|R^{-1}\|_2}.$$

Wir betrachten erneut das lineare Ausgleichsproblem mit $\text{Rang}(A) = r < p := \min\{m, n\}$. Um die Minimallösung in (2.6) mit der SVD zu bestimmen, wird durch die Singulärwerte von A geteilt. Falls durch Rundungsfehler sehr kleine Singulärwerte

$\sigma_{r+1}, \ldots, \sigma_p > 0$ berechnet werden, die ja eigentlich Null sein sollten, führt dies zu einer großen Verfälschung

$$\widetilde{x}_* = \sum_{k=1}^{r} \frac{\langle b, u_k \rangle}{\sigma_k}\, v_k + \sum_{k=r+1}^{p} \frac{\langle b, u_k \rangle}{\sigma_k}\, v_k.$$

Deshalb sollte die Summation nur bis zu einem sinnvollen berechneten Singulärwert von A laufen, der nicht zu klein ist.

2.5.4 Definition: Numerischer Rang:
Für $A \in \mathbb{R}^{m\times n}$ und $\epsilon > 0$ heißt

$$\mathrm{Rang}(A, \epsilon) = \min\{\mathrm{Rang}(B) : \|A - B\|_2 \le \epsilon\}$$

der *numerische Rang* von A. Die Matrix A heißt *numerisch Rang-defizient*, falls

$$\mathrm{Rang}(A, \epsilon) < p = \min\{m, n\} \quad \text{für} \quad \epsilon = \mathtt{eps}\|A\|_2.$$

2.5.5 Satz: Rang-k Approximation
Die Matrix $A \in \mathbb{R}^{m\times n}$ habe die Singulärwertzerlegung $A = U\Sigma V^T$. Wir definieren die "abgeschnittene Singulärwertzerlegung" für $1 \le k < r = \mathrm{Rang}(A)$ als

$$A_k = \sum_{j=1}^{k} \sigma_j\, u_j v_j^T.$$

Dann gilt

$$\|A - A_k\|_2 = \min\{\|A - B\|_2 : \mathrm{Rang}(B) \le k\} = \sigma_{k+1}.$$

Insbesondere gilt $\mathrm{Rang}(A, \epsilon) = k$, falls $\sigma_{k+1} \le \epsilon < \sigma_k$ gilt.

Beweis: Es gilt $\mathrm{Rang}(A_k) = k$ und $A_k = U\Sigma_k V^T$ mit $\Sigma_k = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_k & \\ \hline & 0 & & 0 \end{array}\right)$.

Also ist

$$\|A - A_k\|_2 = \|U^T(A - A_k)V\|_2 = \|\Sigma - \Sigma_k\|_2 = \sigma_{k+1},$$

weil σ_{k+1} der größte Singulärwert von $\Sigma - \Sigma_k$ ist.

Sei nun $B \in \mathbb{R}^{m\times n}$ mit $\mathrm{Rang}(B) \le k$. Dann ist $\dim(\mathrm{Kern}(B)) \ge n - k$, also

$$\dim\left(\mathrm{Span}(v_1, \ldots, v_{k+1}) \cap \mathrm{Kern}(B)\right) \ge 1.$$

Wähle $x = \sum_{j=1}^{k+1} \alpha_j v_j$ mit $\|x\|_2 = 1$ und $x \in \text{Kern}(B)$. Dann ist

$$\|(A-B)x\|_2 = \|Ax\|_2 = \left\| \sum_{j=1}^{k+1} \alpha_j \sigma_j u_j \right\|_2 = \left(\sum_{j=1}^{k+1} \alpha_j^2 \, \sigma_j^2 \right)^{1/2} \geq \sigma_{k+1} \underbrace{\left(\sum_{j=1}^{k+1} \alpha_j^2 \right)^{1/2}}_{=\|x\|_2=1}.$$

Also ist die Spektralnorm von $A - B$ größer oder gleich σ_{k+1}.

Zum Abschluss dieses Kapitels wollen wir noch den Begriff der Kondition von Matrizen auf rechteckige Matrizen erweitern.

2.5.6 Definition: Konditionszahl rechteckiger Matrizen
Für $A \in \mathbb{R}^{m \times n}$ mit $\text{Rang}(A) = p = \min\{m, n\}$ und Singulärwerten $\sigma_1 \geq \cdots \geq \sigma_p > 0$ heißt

$$\text{cond}_2(A) = \frac{\sigma_1}{\sigma_p}$$

die (*Spektral-*)*Kondition* von A.

Bemerkung:

(i) Für die Moore-Penrose-Pseudoinverse A^+ gilt $\|A^+\|_2 = \frac{1}{\sigma_r}$, also ist (fast wie für invertierbare Matrizen)

$$\text{cond}_2(A) = \|A\|_2 \, \|A^+\|_2.$$

Die Definitionen stimmen also im Fall quadratischer regulärer Matrizen überein.

(ii) Im Fall $\text{Rang}(A) = n$ gilt

$$\text{cond}_2(A) = \frac{\sigma_1}{\sigma_n} = \left(\frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)} \right)^{1/2} = \sqrt{\text{cond}_2(A^T A)}.$$

Andererseits ergibt sich für die QR-Zerlegung

$$\text{cond}_2(A) = \text{cond}_2(R).$$

Hieran sieht man deutlich, dass die Konditionszahl für das LGS der Normalengleichungen größer ist als für die Lösung des linearen Ausgleichsproblems mit der QR-Zerlegung.

Zusammenfassung: Methoden zur Lösung des linearen Ausgleichsproblems

$$\|Ax - b\|_2 \to \min!, \qquad A \in \mathbb{R}^{m \times n}, \;\; \text{Rang}(A) = n.$$

Methode	Lösung	Kondition	Aufwand
Normalengleichung	$A^T Ax = A^T b$	$\text{cond}(A^T A) = (\text{cond}(A))^2$	$\frac{mn^2}{2} + \frac{n^3}{6}$
QR-Zerlegung	$Rx = (q_1, \ldots, q_n)^T b$	$\text{cond}(R) = \text{cond}(A)$	$mn^2 - \frac{n^3}{3}$
Singulärwertzerlegung	$x = \sum_{k=1}^{n} \frac{\langle b, u_k \rangle}{\sigma_k} v_k$	$\text{cond}(\Sigma) = \text{cond}(A)$	$2mn^2 + 4n^3$