3.5 Die Kondition eines linearen Gleichungssystems (einer Matrix)

Neu benötigt: Matrixnorm und ihre Eigenschaften.

|| . || Matrixnorm: ||A|| > 0 für A ≠ 0

||a*A|| = |a|*||A|| für a∈R

||A+B|| <= ||A|| + ||B||

Besonders wichtig sind Matrixnormen, die zu einer Vektornorm „passen“ :

3.5.1. Eine Matrixnorm ist mit einer Vektornorm verträglich, wenn $\|Ax\| \leq \|A\| \cdot \|x\|$

Vektor Matrix Vektornorm

3.5.2. Eine Matrixnorm heißt submultiplikativ, wenn stets gilt

$$\|A \cdot B\| \leq \|A\| \cdot \|B\|$$

3.5.3. Eine submultiplikative Matrixnorm, verträglich mit einer vorgegeben Vektornorm, kann leicht definiert werden durch

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

die sog. Grenzen-Norm oder lub-Norm (least upper bound):

$$\frac{\|Ay\|}{\|y\|} \leq \sup_x \frac{\|Ax\|}{\|x\|} = \|A\| \Rightarrow \|Ay\| \leq \|A\| \cdot \|y\| \qquad \forall y$$

$$\|A \cdot B\| = \sup \frac{\|ABx\|}{\|x\|} = \sup_x \left(\frac{\|ABx\|}{\|Bx\|} \cdot \frac{\|Bx\|}{\|x\|} \right) \leq$$

$$\leq \sup_{y=Bx} \frac{\|Ay\|}{\|y\|} \cdot \sup_x \frac{\|Bx\|}{\|x\|} \leq \|A\| \cdot \|B\|$$

So erhält man zu den Vektornormen $\|x\|_1 = \sum_{i=1}^{n} |x_i|$

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

dazugehörige Matrix-Grenzen-Normen $\|\cdot\|_1$ und $\|\cdot\|_\infty$.

3.5.4. Das innere Produkt $(x, y) = x^T y = \sum x_i \cdot y_i$

führt auf die

3.5.5. Euklid'sche Norm: $\|x\|_2^2 = (x, x) = x^T \cdot x = \sum x_i^2$

und die damit verträgliche **Matrixnorm**

3.5.6. (2-Norm): $\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$ als maximale Längenänderung

mit der Eigenschaft

$$\|Ax\|_2^2 = (Ax, Ax) = (Ax)^T (Ax) = x^T A^T Ax = x^T (A^T A) x$$

$$\|A\|_2^2 = \sup_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x \neq 0} \frac{x^T (A^T A) x}{x^T x} = \lambda_{\max}(A^T A)$$

3.5.7. Anmerkung:

Speziell wichtige Klasse von Matrizen mit Norm 1:

Q ist orthogonale Matrix ←→ $Q^{-1} = Q^T$

oder $Q\,Q^T = I = Q^T Q$,

$$\|Qx\|_2^2 = x^T Q^T Q x = x^T x = \|x\|_2^2;$$

$$\|Qx\|_2 = \|x\|_2 \quad \rightarrow \quad \|Q\|_2 = \sup_{x \neq 0} \frac{\|Qx\|_2}{\|x\|_2} = 1$$

$$\|A\|_2 = \|U^T \Lambda U\|_2 = \|\Lambda\|_2 = \max_{i=1,\dots,n} |\lambda_i|$$

Eigenwertzerlegung von $A=A^T$. U ist orthogonale Matrix

Sei A nun eine allgemeine Matrix → Verallgemeinerung: SVD
Singuläre Werte

A^TA und AA^T sind symmetrisch → ONB Eigenvektoren

$A = U\Sigma V$ Wobei U^T die Eigenvektoren von AA^T sind,
V die Eigenvektoren von A^TA, und
Σ, die sog. Singulären Werte von A
sind die Wurzeln aus den Eigenwerten
von A^TA und auch von AA^T.

$$\|A\|_2 = \|U\Sigma V\|_2 = \|\Sigma\|_2 = \max_{i=1,\dots,n} |\lambda_i|^2 = \sigma_{\max}$$

Also $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$

wobei $\lambda_{\max}(A^T A)$ der größte Eigenwert von $A^T A$ ist, und $\sigma_{\max}(A)$ der größte Singulärwert von A.

Diese Matrixnorm ist i.A. nicht einfach berechenbar!

3.5.8. Def. der Frobeniusnorm: $\|A\|_F := \sqrt{\sum |a_{j,k}|^2}$

Fasse die Matrix A als Vektor auf und benutze für diesen langen Vektor die euklid'sche Vektornorm → $\|.\|_F$

Die Frobeniusnorm ist mit der euklid'scher Vektornorm verträglich und submultiplikativ.

3.6 Gestörte Eingabedaten

Für ein lineares Gleichungssystem **A x = b** mit Matrix **A,** Vektor der rechten Seite **b** und gesuchtem Lösungsvektor **x,** untersuchen wir wieder den Einfluss von Eingabefehlern bei sonst exakter Rechnung

→ Kondition der Matrix

Eingabedaten mit Fehler: $b \rightarrow b + \Delta b = \tilde{b}$

Damit ergibt sich an Stelle der exakten Lösung x die Näherung

$$x \rightarrow x + \Delta x = \tilde{x}$$

Mit **Ax = b** und $A\tilde{x} = A(x + \Delta x) = b + \Delta b = \tilde{b}$

gilt: $A\Delta x = \Delta b$ **oder** $\Delta x = A^{-1}\Delta b$

Die Matrix **A** wird hier als exakt angenommen!

Damit erhält man die Ungleichungen

$$\|\Delta x\|_2 = \|A^{-1}\Delta b\|_2 \le \|A^{-1}\|_2 \|\Delta b\|_2$$

$$\|b\|_2 = \|Ax\|_2 \le \|A\|_2 \|x\|_2$$

wegen der Verträglichkeit der euklid'schen Vektor- und Matrixnorm.

Also auch $\frac{1}{\|x\|_2} \le \frac{\|A\|_2}{\|b\|_2}$ und damit insgesamt:

3.6.1.

$$\frac{\|\Delta x\|_2}{\|x\|_2} \le \left(\|A^{-1}\|_2 \cdot \|A\|_2\right) \cdot \frac{\|\Delta b\|_2}{\|b\|_2}$$

rel.Fehler in x — **Kondition von A** — **rel. Fehler in b**

3.6.2 Definition: Die Kondition der Matrix A bzgl. der euklid'schen Norm ist gegeben durch

$$cond_2(A) = \left\| A^{-1} \right\|_2 \cdot \left\| A \right\|_2$$

(Genauso kann man Konditionszahlen bzgl. anderer verträglicher Normen definieren.)

Die Konditionszahl beschreibt also wieder, wie sich ein relativer Eingabefehler in Vektor b auf das Resultat, den Lösungsvektor x, auswirkt.

cond(A) groß → kleine Störungen in b bewirken große Fehler in x

$$cond_2(Q) = \| Q^{-1} \|_2 \| Q \|_2 = \| Q^T \|_2 \| Q \|_2 = 1$$

Außerdem gilt

$$cond_2(QA) = cond_2(A)$$

denn $$\|QA\|_2 = \sup_{x \neq 0} \frac{\|QAx\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \|A\|_2$$

Orthogonale Matrizen sind selbst gut konditioniert und lassen bei Multiplikation mit einer Matrix die Kondition der Ausgangsmatrix unverändert!

Man beachte: Der relative Fehler des Gesamtvektors wird abgeschätzt, nicht der Fehler einzelner Komponenten!

Als Vektor betrachtet ist $\begin{pmatrix} 10^{-5} \\ 1 \end{pmatrix} \approx \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ 

Aber komponentenweise ist natürlich $10^{-5} \not\approx 0$

Beispiel: $A = \begin{pmatrix} 10^{-9} & 1 \\ 0 & 1 \end{pmatrix}$ *und* $A^{-1} = \begin{pmatrix} 10^{9} & -10^{9} \\ 0 & 1 \end{pmatrix}$

Normen: $\|A\|_2 \approx \sqrt{2}\,,\ \|A^{-1}\|_2 \approx \sqrt{2}\cdot 10^9\,,\ cond_2(A) = 2\cdot 10^9$

Wähle speziell: $b = \begin{pmatrix} 1 \\ 1 \end{pmatrix},\quad x = A^{-1}b = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\quad \Delta b = \begin{pmatrix} \Delta b_1 \\ 0 \end{pmatrix},$

Dann folgt $|\Delta x_1| = 10^9 \cdot |\Delta b_1|, \Delta x_2 = 0 \quad \|x\| = 1, \quad \|b\| = \sqrt{2}$

Damit ergibt sich $\dfrac{\|\Delta x\|}{\|x\|} = \dfrac{10^9 \cdot |\Delta b_1|}{1} = \sqrt{2}\cdot 10^9 \cdot \dfrac{\|\Delta b\|}{\|b\|}$

Vgl.: $cond(A) = \|A\| \cdot \|A^{-1}\| = \sqrt{2}\cdot\sqrt{2}\cdot 10^9 = 2\cdot 10^9$

Ein kleiner Fehler in der ersten Komponente von b wirkt sich verstärkt um Faktor 10^9 in der ersten Komponente von x aus.

3.7 Kosten der Gauss-Elimination

Zunächst Dreieckssystem:

$$x_n = b_n / a_{nn};$$

$$für \quad i = n-1, n-2, \ldots, 1:$$

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{ij} x_j}{a_{ii}}$$

Programm:

```
FOR i = n,...,1 DO
    x(i) = b(i);
    FOR j=i+1,…,n DO
        x(i) = x(i) - a(i,j)x(j);
    ENDFOR
    x(i) = x(i)/a(i,i);
ENDFOR
```

In jedem Schritt i fallen eine Division und jeweils n-i Additionen und Multiplikationen an:

Also $$\sum_{i=1}^{n-1}(n-i)=\sum_{j=1}^{n-1} j=\frac{n(n-1)}{2}=\frac{n^2}{2}-\frac{n}{2}$$

Additionen und genauso viele Multiplikationen.

Dazu kommen n Divisionen.

3.7.2. Definition: Unter flop verstehen wir eine elementare ‚floating point operation' (Gleitpunktoperation + , - , * , /). Die Kosten eines Algorithmus werden üblicherweise in flop angegeben. Dazu gibt man in erster Näherung nur den Term höchster Ordnung an.

In unserem Fall: $$2\cdot\left(\frac{n^2}{2}-\frac{n}{2}\right)+n=n^2 \quad\rightarrow\quad n^2 \qquad flop$$

oder $$O(n^2) \qquad flop$$

3.7.3. Exkurs: Landau'sche Symbole:

1. Betrachte Funktion in n für n → ∞ :

$$f(n) = \mathrm{O}(g(n)), \quad falls \quad \left|\frac{f(n)}{g(n)}\right| \leq M < \infty \quad für \quad n \geq N$$

Beispiel: $n^2 + n = \mathrm{O}(n^2)$; denn n^2 ist der am stärksten wachsende Term → f(n) / g(n) = 1 + 1/n <= M

2. Betrachte Funktion in h für h → 0 :

$$f(h) = \mathrm{O}(g(h)), \quad falls \quad \left|\frac{f(h)}{g(h)}\right| \leq c < \infty \quad für \quad |h| \leq \delta$$

Beispiel: $h^2 + h = \mathrm{O}(h)$; denn h ist der am langsamsten schrumpfende Term → f(h) / g(h) = h + 1 <= c

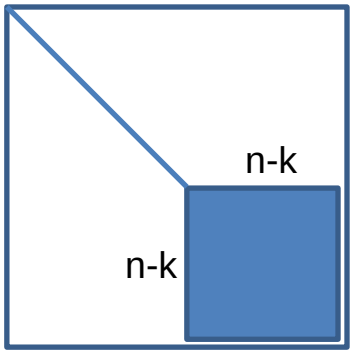
O(..) bezeichnet also jeweils den dominanten Term!

3.7.4. Kosten der Gauss-Elimination:

Im k-ten Teilschritt arbeitet man in einer n-k x n-k Untermatrix A_k

In dieser Matrix wird für i=k+1,...,n neu berechnet :

$$a_{ij} = a_{ij} - l_{ik} a_{kj} \; ; \qquad \begin{pmatrix} & * \\ 0 & \end{pmatrix}$$



Das sind ca. $(n-k)^2$ Additionen und genauso viele Multiplikationen.

Insgesamt also

$$2\sum_{k=1}^{n-1}(n-k)^2 = 2\sum_{j=1}^{n-1} j^2 = \frac{2(n-1)(2n-1)n}{6} = \frac{2}{3}n^3 + \mathrm{O}(n^2) \qquad flop$$

Dazu kommen $O(n^2)$ flop für die Spaltenpivotsuche und $O(n^2)$ flop für das Auflösen des Dreiecksgleichungssystems.

Diese Kosten fallen aber praktisch nicht ins Gewicht gegenüber den obigen ***2/3*n^3*** .

3.7.5 Beispiel zur Verdeutlichung der Kondition

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 \\ -1 & 1 & \ddots & \vdots & 1 \\ -1 & -1 & \ddots & 0 & 1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 & 1 \end{pmatrix} \begin{matrix} \downarrow \\ \\ \\ \\ \\ \end{matrix} \quad \begin{matrix} \\ \downarrow \\ \\ \vdots \\ \\ \end{matrix} \quad \begin{matrix} + \\ \\ \\ \\ \downarrow \end{matrix}$$

Eliminiere die erste Spalte durch Addieren der ersten Zeile:

$$\begin{pmatrix} 1 & & & & 1 \\ 0 & 1 & & & 2 \\ 0 & -1 & 1 & & 2 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & -1 & \cdots & -1 & 2 \end{pmatrix} \begin{matrix} \downarrow \\ \\ \\ \\ \\ \end{matrix} \quad \begin{matrix} \\ \\ \\ \vdots \\ \\ \end{matrix} \quad \begin{matrix} + \\ \\ \\ \\ \downarrow \end{matrix}$$

Im nächsten Schritt

$$\begin{pmatrix} 1 & & & & 1 \\ 0 & 1 & & & 2 \\ 0 & 0 & 1 & & 4 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 4 \end{pmatrix}$$

und schließlich

$$U = \begin{pmatrix} 1 & & & & & 1 \\ & 1 & & & & 2 \\ & & 1 & & & 4 \\ & & & \ddots & & \vdots \\ & & & & 1 & 2^{n-2} \\ & & & & & 2^{n-1} \end{pmatrix}$$

In jedem Schritt verdoppelt sich der größte Eintrag in der Matrix!

Kondition von A selbst ist **O(n),** Kondition von U ist **O(2^{n-1}) !**

Im Verlauf der Gauss-Elimination *kann* die Kondition der Matrizen sehr stark anwachsen!

Aber: In der Praxis kommt das *so gut wie* nie vor!

Gauss-Elimination mit Pivotsuche *gilt als* numerisch stabil.

3.8 Methode der kleinsten Quadrate

(Least Squares, Normalgleichung)

Ausgangspunkt: Überbestimmtes System.

Mehr Gleichungen als Unbekannte

$$\boxed{A}\, x = b$$

Sei A eine m x n – Matrix mit m>n und maximal vollem Rang:

rang(A) = n, d.h. A bildet den $\mathbf{R^m}$ in den ganzen $\mathbf{R^n}$ ab.

Das System $A x = b$ ist dann i.A. nicht lösbar!

Versuche, das Problem so gut wie möglich zu lösen!

Minimiere dazu die Abweichung $A x - b$ in passender Norm!

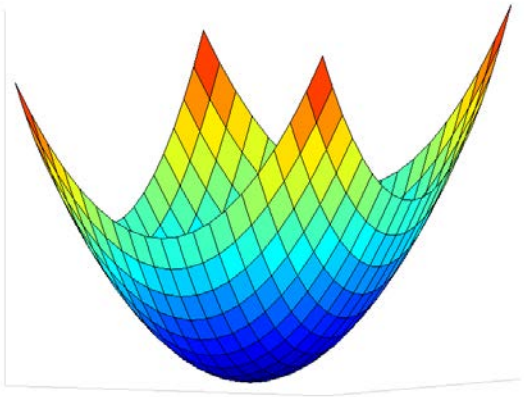
Am besten eignet sich dazu die euklid'sche Norm, da sie auf eine differenzierbare Funktion f führt:

3.8.1. : $\min_x \|Ax-b\|_2^2$

$$f(x_1,\ldots,x_n) := \|Ax-b\|_2^2 =$$

$$= (Ax-b)^T (Ax-b) = x^T A^T Ax - 2x^T A^T b + b^T b =$$

$$= \left\| \begin{pmatrix} (\sum_{j=1}^{n} a_{1,j} x_j) - b_1 \\ \vdots \\ (\sum_{j=1}^{n} a_{m,j} x_j) - b_m \end{pmatrix} \right\|_2^2 = \sum_{k=1}^{m} \left((\sum_{j=1}^{n} a_{k,j} x_j) - b_k \right)^2$$



Die Funktion f beschreibt einen Paraboloiden (n-dim. Parabel).

Das eindeutige Minimum dieser Funktion ist an der Stelle, an der die Ableitung gleich Null ist (waagrechte Tangente).

$$0 = \frac{df}{dx_i} = 2\sum_{k=1}^{m}\left(\sum_{j=1}^{n} a_{k,j}x_j - b_k\right)a_{k,i} \qquad \text{für i=1,...,n}$$

oder $$\sum_{k=1}^{m} a_{k,i}\sum_{j=1}^{n} a_{k,j}x_j = \sum_{k=1}^{m} a_{k,i}b_k$$

In Matrixschreibweise: $\left(A^T Ax\right)_i = \left(A^T b\right)_i$, i=1,...,n

3.8.2. Normalgleichung zu Ax=b: $A^T Ax = A^T b$

Die Matrix $\mathbf{A^TA}$ ist eine n x n – Matrix von Rang n
(da A Rang n hat) und beschreibt daher ein eindeutig lösbares, quadratisches lineares Gleichungssystem.

Allerdings ist die Kondition von A^TA oft sehr viel schlechter als die von A, denn:

$$cond_2(A^T A) = \left\| A^T A \right\|_2 \cdot \left\| (A^T A)^{-1} \right\|_2 =$$

$$= \sqrt{\lambda_{\max}(A^T A A^T A) * \lambda_{\max}((A^T A)^{-1}(A^T A)^{-1})} =$$

$$= \lambda_{\max}(A^T A) \cdot \lambda_{\max}(inv(A^T A)) =$$

$$= \lambda_{\max}(A^T A) / \lambda_{\min}(A^T A) = \sigma^2_{\max}(A) / \sigma^2_{\min}(A) =$$

$$= \left\| A \right\|_2^2 \cdot \left\| A^{-1} \right\|_2^2 = cond^2(A)$$

Im folgenden Abschnitt werden wir daher ein besseres Verfahren zur Lösung dieses Problems kennen lernen.

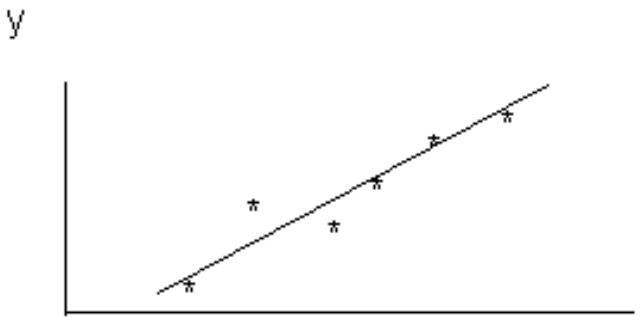
Dazu werden besser orthogonale Matrizen verwendet, um diese Konditionsverschlechterung zu vermeiden.

3.8.3. Lineares Ausgleichsproblem (Ausgleichsgerade)

Gegeben: Punktepaare in der Ebene , (x_i, y_i), i=1,...,n;

Gesucht: beste Gerade, die möglichst nahe an den Punkten liegt.

y = g(x) = ax + b .



Es soll also gelten:

$$\begin{pmatrix} a + bx_1 \\ \vdots \\ a + bx_n \end{pmatrix} \approx \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

oder in Matrixschreibweise

$$A\begin{pmatrix} a \\ b \end{pmatrix} \approx y \quad \text{mit} \quad A = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \text{und} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Die Normalgleichung lautet also

$$A^T A\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n & \sum_{j=1}^{n} x_j \\ \sum_{j=1}^{n} x_j & \sum_{j=1}^{n} x_j^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} x_j y_j \end{pmatrix}.$$

Die Lösung dieses 2 x 2 – Gleichungssystems liefert a und b, und damit die gesuchte Gerade y = ax + b .

Allgemeiner:

Ansatzfunktionen $g_1(x),\ldots,g_m(x)$ und

Punkte $(x_1,y_1),\ldots,(x_n,y_n)$, n>m

Gesucht : $f(x)=\sum_{k=1}^{m} a_k g_k(x)$ mit $f(x_j)\approx y_j,\ j=1,\ldots,n$

Mit $G=\begin{pmatrix} g_1(x_1) & \cdots & g_m(x_1) \\ \vdots & & \vdots \\ g_1(x_n) & \cdots & g_m(x_n) \end{pmatrix}$ ist dann $G^T G \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = G^T \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

zu lösen.

Ergebnis ist die ‚näheste' Funktion an den vorgegebenen Punkten, die aus den g_1,..., g_m linear zusammengesetzt ist.

3.9. Die QR-Zerlegung einer Matrix

Schon vorher haben wir bemerkt:

- cond(U) in der Gauß-Elimination ev. groß, auch bei kleinem cond(A);
- falls A schlecht konditioniert: was ist der Rang von A? Welche Pivotelemente werten wir als 0?
- cond($A^T A$) oft sehr groß,
- andererseits: cond(QA) = cond(A), falls Q orthogonal.

Also sind orthogonale Matrizen sehr gut für äquivalente Umformungen von A geeignet (vgl. LU-Zerlegung).

Außerdem gilt: $\mathbf{Q^{-1} = Q^T}$.

Also sind Gleichungssysteme in **Q** sehr leicht zu lösen.

Versuche daher, analog zur LU-Zerlegung A=LR ,
eine Zerlegung der Form

$$\mathbf{A = QR}$$

zu bestimmen mit

Q orthogonal

R obere Dreiecksmatrix

Vorteile:

- numerisch stabiler als LU
- ähnliche Kosten
- Systeme in Q und R leicht zu lösen

3.9.2 Elementare orthogonale Matrizen

Orthogonale 2 x 2 – Matrix :

$$G = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{pmatrix}$$

heißt **Givensreflexion.** Denn

$$G^T G = GG = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{pmatrix} \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{pmatrix} =$$

$$= \begin{pmatrix} \cos^2(\varphi) + \sin^2(\varphi) & \cos(\varphi)\sin(\varphi) - \sin(\varphi)\cos(\varphi) \\ \sin(\varphi)\cos(\varphi) - \cos(\varphi)\sin(\varphi) & \sin^2(\varphi) + \cos^2(\varphi) \end{pmatrix} =$$

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

G ist eindeutig bestimmt durch den ‘Winkel’ φ.
Bestimme nun φ so, dass

$$\tilde{A} = GA = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} \\ \tilde{a}_{21} & \tilde{a}_{22} \end{pmatrix}$$

obere Dreiecksmatrix wird.

Dazu muss gelten:

$$\tilde{a}_{21} = \begin{pmatrix} \sin(\varphi) & -\cos(\varphi) \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \sin(\varphi) a_{11} - \cos(\varphi) a_{21} \overset{!}{=} 0$$

Lösung:

$$\cot(\varphi) = \frac{a_{11}}{a_{21}}; \quad \varphi = arccctg\left(\frac{a_{11}}{a_{21}}\right) \quad \text{oder} \quad \varphi = arctg\left(\frac{a_{21}}{a_{11}}\right)$$

Ist $a_{21} = 0$, so ist keine weitere Transformation nötig!

Numerisch stabilere Art der Berechnung :

(a_{21} oder a_{11} könnten fast 0 sein):

$$\rho = sign(a_{11})\sqrt{a_{11}^2 + a_{21}^2}; \quad \cos(\varphi) = \frac{a_{11}}{\rho}; \quad \sin(\varphi) = \frac{a_{21}}{\rho};$$

3.9.3. Givens-Reflexion für den allgemeinen n x n – Fall:

n-dimensionale Givens-Reflexion ist im Wesentlichen wie die Einheitsmatrix, bis auf den gerade zu betrachtenden 2 x 2 – Block.

Dieser Block wird wieder - wie oben definiert - abhängig von φ bestimmt.

Man eliminiert wieder in der ersten Spalte $a_{2,1},\ldots,a_{m,1}$, und dann entsprechend in der zweiten Spalte die Unterdiagonale, usw. wie bei Gauss.

Gestalt der allgemeinen Givens-Matrix

$$
G_{ij} = \begin{pmatrix}
1 & & & & & & & & \\
& \ddots & & & & & & & \\
& & 1 & & & & & & \\
& & & \cos(\varphi) & & & \sin(\varphi) & & \\
& & & & 1 & & & & \\
& & & & & \ddots & & & \\
& & & & & & 1 & & \\
& & & \sin(\varphi) & & & -\cos(\varphi) & & \\
& & & & & & & 1 & \\
& & & & & & & & \ddots \\
& & & & & & & & 1
\end{pmatrix}
\begin{matrix}
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & j \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & i \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & &
\end{matrix}
\qquad
\begin{matrix}
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & j & & & i & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & & \\
 & & & & & & & &
\end{matrix}
$$

Zur Elimination eines Elementes a_{ij} der Matrix A multiplizieren wir $G_{ij} \cdot A$.

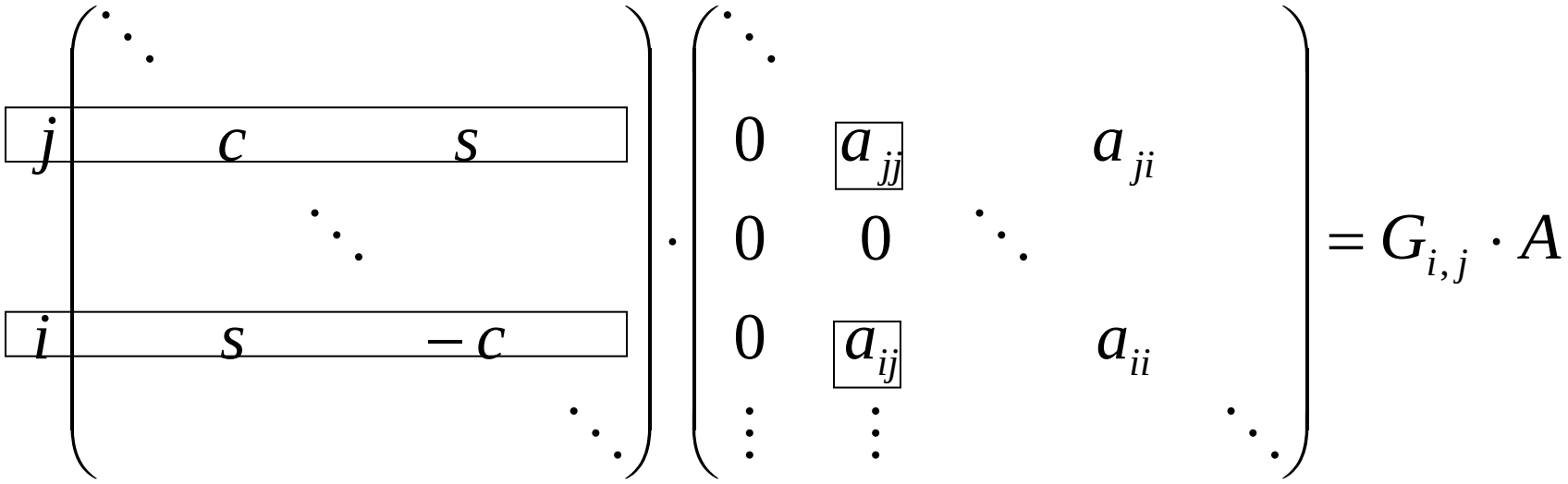
Dieses Produkt verändert nur die i-te und die j-te Zeile von A. Es genügt, vom Gesamt-System nur diesen 2 x 2 – Teil zu betrachten. Also muss wieder

$$\varphi = arcctg\left(\frac{a_{jj}}{a_{ij}}\right)$$ gesetzt sein wie oben. (1 → j und 2 → i)

Mit einer solchen Matrix G_{21} wird dann im ersten Schritt a_{21} zu Null gemacht.

$$G_{21} = \begin{pmatrix} G & 0 \\ 0 & I \end{pmatrix}, \quad G = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ \sin(\varphi) & -\cos(\varphi) \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

Genauere Analyse eines allgemeinen Eliminationsschritts:



verändert nur j-te und i-te Zeile;

i-te Zeile: $a_{ik} \to sa_{jk} - ca_{ik}$ für k=j,...,n

speziell: $a_{ij} \to sa_{jj} - ca_{ij} = 0$! soll Null werden

Legt daher φ fest.

j-te Zeile: $a_{jk} \to ca_{jk} + sa_{ik}$ für k=j,...,n

mit c und s zu obigen φ

Verwende der Reihe nach $G_{21}, G_{31}, \ldots, G_{n1}$

zur Bearbeitung der ersten Spalte,

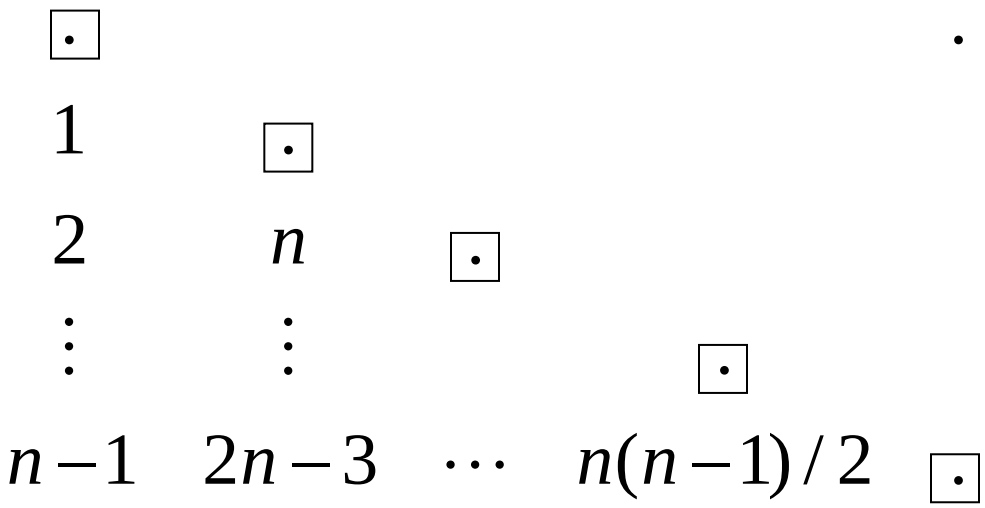
also um $a_{21}, a_{31}, \ldots, a_{n1}$ zu Null zu machen,

und danach $G_{32}, G_{42}, \ldots, G_{n2}$, ... , G_{n-1n-2}, G_{nn-2} , und G_{nn-1}

um $a_{32}, a_{42}, \ldots, a_{n2}$, ... , a_{n-1n-2}, a_{nn-2} , und a_{nn-1}

zu Null zu machen.

Die Reihenfolge, in der die $a_{j,i}$ zu Null gemacht werden, ist gegeben durch:



Jeweils nötig ist eine Multiplikation mit Givens-Reflexion

$G_{i,j}$, i=1,...,n-1 und **j=i+1,...,n.**

Also benötigt man insgesamt n(n-1)/2 Givensreflexionen um eine quadratische n x n –Matrix auf Dreiecksgestalt zu transformieren.

Man benutze also immer das Diagonalelement a_{jj} und eine Kombination von i-ter/j-ter Zeile, um a_{ij} zu Null zu machen.

$$Q^T := G_{n,n-1} G_{n,n-2} \cdots G_{n-1,n-2} \cdots G_{n1} \cdots G_{21}$$

$$Q^T A = G_{n,n-1} G_{n,n-2} \cdots G_{n-1,n-2} \cdots G_{n1} \cdots G_{21} A = R$$

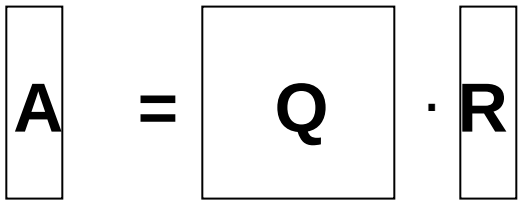
mit einer oberen Dreiecksmatrix R.

Daher ist

$Q = G_{21} \cdots G_{n1} \cdots G_{n,n-1}$, da $G_{ij}^T = G_{ij}$ und A=Q*R.

Q ist gegeben durch die einzelnen G_{ij};
jedes G_{ij} ist eindeutig gegeben durch das φ_{ij}, das nötig war, um genau ein a_{ij} zu eliminieren.

Genauso kann man für eine **m** x **n** Matrix **A** **(m>n)** mit **rank(A)=n** eine QR- Zerlegung berechnen



Wie bei der Gauss-Elimination eliminiert man also mit den Diagonalelementen der Reihe nach sämtliche Unterdiagonalelemente.

Der Vorteil der QR-Zerlegung:

cond(A) = cond(QR) = cond(R)
Gut für schlecht konditionierte Systeme
Anwendbar auf rechteckige Systeme

Andere Orthogonalisierungsverfahren:

- Gram-Schmidt (orthonormalisiere Vektoren)
- Householder (erzeuge in einem Schritt eine ganze Nullspalte).

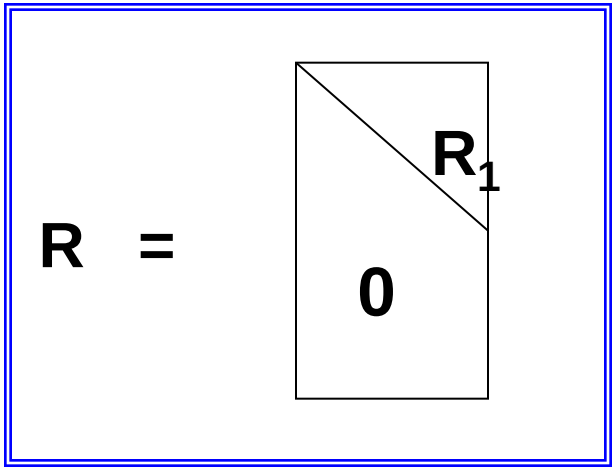
3.9.4. Anwendung bei Linearer Ausgleichsrechnung:

$$\min_x \|Ax-b\|_2 = \min_x \|QRx-b\|_2 =$$

$$= \min_x \|Q^T(QRx-b)\|_2 = \min_x \|Rx-Q^Tb\|_2$$

da Q orthogonal und euklid'sche Norm.

R ist obere Dreiecksmatrix der Dimension m x n und vollen Ranges n.



Das obige Minimum erhält man wegen

$$\min_x \left\| \begin{pmatrix} R_1 \\ 0 \end{pmatrix} x - Q^T b \right\|_2^2 = \min_x \left\| \begin{pmatrix} R_1 x \\ 0 \end{pmatrix} - \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} \right\|_2^2 = \min_x \left\| R_1 x - \tilde{b}_1 \right\|_2^2 + \left\| \tilde{b}_2 \right\|_2^2$$

aus der Lösung des Dreieckssystems

$$R_1 x = \tilde{b}_1$$

Der Wert des Minimums ist gegeben durch $\left\| \tilde{b}_2 \right\|_2$

Beispiel: QR-Zerlegung für Least Squares:

$$\min_x \left\| \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} x - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\|_2 = \min_x \|Ax - b\|_2$$

Erster Schritt: $a_{21} \rightarrow 0$:

$$\begin{pmatrix} c & s & 0 \\ s & -c & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} c+s & c+s \\ s-c & s-c \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$$

mit

$$c = s = 1/\sqrt{2}, \qquad \varphi = \pi/4$$

Zweiter Schritt: a_{32} → 0 :

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & s & -c \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & s \\ 0 & -c \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

mit $c = 0, \; s = 1, \qquad \varphi = \pi / 2$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$$\mathbf{Q^T} \cdot \mathbf{A} = \mathbf{R}$$

Also $$Q = \begin{pmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \end{pmatrix}$$

Anwendung auf Minimierungsproblem :

$$\min_x \left\| Ax - \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\|_2 = \min_x \left\| \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \\ 0 & 0 \end{pmatrix} x - \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\|_2 =$$

$$= \min_x \left\| \begin{pmatrix} \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \end{pmatrix} \cdot x - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ \hline 0 \end{pmatrix} \right\|_2 = \min_x \left\| \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \end{pmatrix} \cdot x - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\|_2$$

Lösung x als Lösung des Dreiecksgleichungssystems:

$$x = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

In diesem Fall liefert x sogar eine genaue Lösung von Ax=b, da der Fehlerterm $\|b_2\|$ gleich Null ist.

QR-Zerlegung ist in dieser Form anwendbar für beliebige rechteckige Matrix A, so lange A vollen Rang besitzt.

Kosten des QR-Verfahrens mit Givens für n x n – Matrix:

$2n^3 + O(n^2)$ (also teuerer als Gauss-Elimination mit $2n^3/3$)

Ein Eliminationsschritt bei Spalte k:

(2 mult +1 add)2k = 6k flop´s

Insgesamt: $\sum_{k=n-1}^{2}(k-1)*6k = 2n^3 + O(n^2)$

Bei m x n – Matrix mit m>n und Rang(A)=n: n^2 (3m-n)

Anwendung des QR-Verfahren bei

- schlecht konditioniertem Gleichungssystem

- überbestimmtem Gleichungssystem mit vollem Rang (an Stelle der Normalgleichung), wie oben beschrieben

- allgemeinem nichtquadratischen System in der Form QAP = R mit Permutation P zum Vertauschen von Spalten. (P ist nötig, um einen Block vollen Ranges nach vorne/oben zu transportieren)

Beispiel: $\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}$

- Entdeckung fast linear abhängiger (eigentlich überflüssiger) Gleichungen
  (numerische Bestimmung des Rangs von A)

- Reduktion der Matrix auf den wesentlichen Teil
  (Noise-reduction)

3.10 Regularisierung

In vielen praktischen Anwendungen hat man zwar ein überbestimmtes lineares Gleichungssystem vorliegen, aber so, dass die Normalmatrix $\mathbf{A^TA}$ auch noch (fast) singulär ist!

Dadurch erhält man bei der Lösung dieses Problems einen Vektor x, der extrem groß ist:

Ist in Bx=b die Matrix B (fast) singulär →

→ ||inv(B)|| sehr groß →

→ || x || = || inv(B)*b || sehr groß

Durch Mess/Rundungsfehler enthält aber die rechte Seite b viele kleine Störungen (noise, Rauschen), die in der berechneten Näherungslösung x dann sehr groß werden, so dass - selbst bei exakter Rechnung - x unbrauchbar ist.

$$\tilde{x} = B^{-1}(b + \Delta b) = B^{-1}b + B^{-1}\Delta b = x + \underbrace{B^{-1}\Delta b}$$

Störanteil
Viel größer als x

Ausweg:

Suche ‚vernünftige' Least Squares Lösung durch Minimierung mit Nebenbedingung:
‚x soll nicht zu groß werden'.

$$\min_x \left(\|Ax - b\|_2^2 + \gamma^2 \|x\|_2^2 \right)$$

Minimerung ←→ Nullstelle der Ableitung führt auf das sog. regularisierte Gleichungssystem

$$\left(A^T A + \gamma^2 I \right) x = A^T b$$

Idee: Verschiebe A^TA durch Aufaddieren von γ^2 I , so dass die neue Matrix besser konditioniert ist.

Dann ist $\left\| inv\left(A^T A + \gamma^2 I\right)\right\|_2 << \left\| inv(A^T A)\right\|_2$

Daher führen in dem neuen Gleichungssystem die Rausch-komponenten in b nicht mehr zu einem extremen Anwachsen der Lösung x.

Man weiß, dass die gesuchte Lösung x nicht zu groß sein kann, und dies wird durch die Regularisierung gewährleistet.

γ heißt Regularisierungsparameter und die hier beschriebene Methode heißt

Tikhonov-Regularisierung.

Regularisierung muss häufig angewendet werden bei Problemen der Bildverarbeitung

(z.B. bei verrauschten, unscharfen Bilder)

- Gauss-Elimination,
- Normalengleichung und QR-Zerlegung,
- Regularisierung

sind die wichtigsten Werkzeuge zur direkten Lösung von $Ax=b$.