Točnost i stabilnost numeričkih algoritama

Pašić, Matteo

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Matteo Pašić

TOČNOST I STABILNOST NUMERIČKIH ALGORITAMA

Diplomski rad

Voditelj rada:
prof. dr. sc. Nela Bosner

Zagreb, rujan, 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Svojoj obitelji za svu podršku i pomoć tijekom studiranja

Sadržaj

Uvod

U ovom radu se bavimo analizom greške QR faktorizacije te analizom greške problema najmanjih kvadrata. QR faktorizacija je svestran računski alat koji se može koristiti u npr. rješavanju sustava linearnih jednadžbi, svojstvenim problemima te služi i za rješavanje problema najmanjih kvadrata. Također, iz QR faktorizacije je nastao QR algoritam koji je proglašen jednim od 10 najvažnijih algoritama 20. stoljeća. Može se izračunati na 3 različita načina. Gram-Schmidtov proces, koji redom ortogonalizira stupce matrice, je najstarija metoda. Givensove transformacije su preferirane kad matrica ima specijalne strukture s puno 0, npr. tridijagonalnu, Hessenbergovu strukturu. Householderova transformacija daje najopćenitiji način za izračunati QR faktorizaciju. Problem najmanjih kvadrata se bavi aproksimativnim rješavanjem sustava u kojem je više jednadžbi nego nepoznanica. Najčešće služi da bi skup mjerenih podataka aproksimirali funkcijom. Pojavljuje se jako puno u statistici no može se pojaviti u svim znanostima koje mjere podatke. Također, pojavljuje se i pri npr. numeričkom rješavanju integralnih jednadžbi. Među najpoznatijim načinima za naći rješenje problema najmanjih kvadarata su normalna jednadžba, koja je ujedno i najstariji način za riješiti problem najmanjih kvadrata, SVD dekompozicija te QR faktorizacija za koju ćemo napraviti analizu greške.

Poglavlje 1

Osnovni pojmovi i metode

1.1 QR faktorizacija

QR faktorizacija matrice $A \in \mathbb{R}^{m\times n}$, $m \geq n$ je faktorizacija:

$$A = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$$

gdje je $Q \in \mathbb{R}^{m\times m}$ ortogonalna, a $R_1 \in \mathbb{R}^{n\times n}$ je gornje trokutasta. Matricu R zovemo gornje trapezastom, jer je matrica $R \in \mathbb{R}^{m\times n}$, a termin trokutasta se primjenjuje samo na kvadratne matrice. Ovisno o kontekstu, punu faktorizaciju $A = QR$ ili "ekonomičniju" verziju $A = Q_1 R_1$ se može zvati QR faktorizacijom. Egzistencija QR faktorizacije za matrice punog stupčanog ranga se može dokazati pomoću Cholesky faktorizacije (faktorizacija simetrične, pozitivno definitne matrice $A = R^T R$, gdje je R gornje trokutasta matrica). Kako A ima puni rang, tada tvrdimo da je matrica $A^T A$ simetrična i pozitivno definitna. Vrijedi da je $(A^T A)^T = A^T (A^T)^T = A^T A$, tj. $A^T A$ je simetrična. Za pozitivnu definitnost trebamo pokazati da je $x^T (A^T A)x > 0, \forall x \neq 0$. Kako je za $x \neq 0$:

$$x^T (A^T A)x = (Ax)^T (Ax) = \langle Ax, Ax \rangle = \|Ax\|_2 \geq 0$$

S obzirom da je matrica A punog ranga slijedi po teoremu o rangu i defektu ($d(A)+r(A) = n$, gdje je $d(A) = \dim(\mathrm{Ker}(A))$, a $r(A) = \dim(\mathrm{Im}(A))$) da je skup $\mathrm{Ker}(A) = \{0\}$. Kako je $x \neq 0$, onda $x \notin \mathrm{Ker}(A)$ pa je $Ax \neq 0$, odnosno:

$$x^T (A^T A)x = \|Ax\|_2 > 0.$$

Tada postoji gornje trokutasta matrica R s pozitivnim dijagonalnim elementima tako da je $A^T A = R^T R$. Zbog pozitivnih dijagonalnih elemenata je matrica R invertibilna pa $A = AR^{-1} \cdot R$. Vrijedi:

$$(AR^{-1})^T (AR^{-1}) = (R^T)^{-1} A^T A R^{-1} = (R^T)^{-1} R^T R R^{-1} = I$$

gdje smo u 2. jednakosti koristili Cholesky faktorizaciju $A^TA = R^TR$, pa imamo faktorizaciju matrice A na produkt ortogonalne matrice AR^{-1} i gornje trokutaste matrice R. Općenito, QR faktorizacija je jedinstvena ako A ima puni stupčani rang i zahtijevamo da R ima pozitivne dijagonalne elemente, jer inače ako je $A = QR$, onda je i $A = QD \cdot DR$ QR faktorizacija, za $D = \text{diag}(\pm 1)$. Jer je $D^2 = I$, onda je i $(QD)^TQD = I$, dok je DR gornje trokutasta matrica kojoj i-ti redak odgovara i-tom retku matrice R, ako je $d_{ii} = 1$, a ako je $d_{ii} = -1$ onda i-tom retku matrice $-R$.

Householderova transformacija

Householderova matrica (Householderova transformacija, Householderov reflektor) je matrica koja ima oblik:

$$P = I - \frac{2}{v^Tv}vv^T, \quad v \neq 0 \in \mathbb{R}^n.$$

Householderova matrica ima svojstvo simetričnosti (jer je $I^T = I$ i $(vv^T)^T = vv^T$), ortogonalnosti te involutivnosti. Vrijedi:

$$P^TP = P^2 = I - \frac{4}{v^Tv}vv^T + \frac{4}{v^Tv}vv^T = I.$$

Primjenom matrice P na vektor dobivamo:

$$Px = x - \left(\frac{2v^Tx}{v^Tv}\right)v$$

Iz ove formule možemo vidjeti zašto se matrica P zove Householderov reflektor: ona reflektira vektor x spram hiperravnine $\text{span}(v)^\perp$. Zanima nas, ako imamo dane vektore x i y, možemo li naći Householderovu matricu P tako da je $Px = y$. Kako je P ortogonalna matrica, vrijedi $\|Px\|_2 = \|x\|_2$, tj. $\|x\|_2 = \|y\|_2$, znači $\|x\|_2 = \|y\|_2$ je nužan uvjet. Vrijedi:

$$Px = y \iff x - \left(\frac{2v^Tx}{v^Tv}\right)v = y,$$

tj. imamo jednadžbu oblika $\alpha v = x - y$. Kako je Householderova matrica za vektor αv, $\alpha \neq 0$ jednaka Householderovoj matrici za vektor v ($I - \frac{2}{(\alpha v)^T(\alpha v)}(\alpha v)(\alpha v)^T = I - \frac{2}{v^Tv}vv^T = P$), možemo uzeti $\alpha = 1$. Dobili smo $v = x - y$. Želimo dobiti $Px = y$ pa računamo:

$$v^Tv = (x^T - y^T)(x - y) = x^Tx - x^Ty - y^Tx + y^Ty = x^Tx + y^Ty - 2x^Ty,$$

gdje smo u zadnjoj jednakosti koristili simetričnost skalarnog produkta u $\mathbb{R}$ ($\langle x, y\rangle = x^Ty$). Također nam treba:

$$v^Tx = (x^T - y^T)x = x^Tx - y^Tx = \frac{1}{2}x^Tx + \frac{1}{2}y^Ty - x^Ty = \frac{1}{2}v^Tv,$$

gdje smo u 3. jednakosti koristili činjenicu da je $\|x\|_2 = \|y\|_2$. Tada imamo:

$$Px = x - \left(\frac{2v^T x}{v^T v}\right) v = x - \frac{v^T v}{v^T v} v = x - v = y.$$

Zaključujemo da, uz uvjet $\|x\|_2 = \|y\|_2$ te $x \neq y$, možemo naći Householderovu matricu P tako da vrijedi $Px = y$.

Ideja kako Householderovim transformacijama izračunati QR faktorizaciju je sljedeća: primjenjivat ćemo Householderove transformacije na matricu A te matricu A tako pretvoriti u gornje trokutastu matricu R. Prva Householderova transformacija će u prvom stupcu postaviti sve elemente ispod dijagonalnog na 0, druga u drugom stupcu itd. Kako bi takvu ideju realizirali, trebamo znati konstruirati takav P. Za to nam koristi zadnja rasprava, jer imamo stupac matrice A i stupac koji želimo dobiti (0 ispod dijagonalnog elementa).

Neka je x proizvoljan vektor, $x \neq 0$ te neka je y vektor koji će imati sve 0, osim na 1. komponenti, tj. $y = \sigma e_1$. Da bismo našli Householderovu matricu P, zahtijevali smo da je $\|x\|_2 = \|y\|_2$. Tako je onda $\sigma = \pm\|x\|_2$. Tada je $v = x - y = x - \sigma e_1$. Da bi izbjegli mogućnost katastrofalnog kraćenja, u praksi se najčešće koristi formula:

$$\sigma = -\text{sign}(x_1)\|x\|_2, \quad v = x - \sigma e_1$$

jer je tada $v_1 = x_1 + \text{sign}(x_1)\|x\|_2$ pa imamo zbrajanje dva broja istog predznaka te tako izbjegavamo mogućnost katastrofalno kraćenje. Ovakav pristup je doveo do mišljenja da je drugi izbor predznaka neprikladan. Pri odabiru drugog predznaka, može doći do katastrofalnog kraćenja, ali formulu možemo zapisati malo drugačije da imamo samo zbrajanje brojeva istih predznaka pa možemo izbjeći mogućnost katastrofalnog kraćenja. Formula će glasiti:

$$\begin{aligned}
&\sigma = \text{sign}(x_1)\|x\|_2, \\
&v_1 = x_1 - \sigma = (x_1 - \sigma)\,\frac{x_1 + \sigma}{x_1 + \sigma} = \frac{x_1^2 - \|x\|_2^2}{x_1 + \sigma} = \frac{-(x_2^2 + \ldots + x_n^2)}{x_1 + \sigma}.
\end{aligned}$$

Tako i u brojniku i u nazivniku imamo zbrajanje brojeva istog predznaka. Iz ovih formula možemo izračunati da je $\beta = \frac{2}{v^T v} = -\frac{1}{\sigma v_1}$. Napravit ćemo to za slučaj $\sigma = -\text{sign}(x_1)\|x\|_2$, dok je račun za drugi predznak vrlo sličan. Imamo:

$$\begin{aligned}
v^T v &= v_1^2 + v_2^2 + \ldots + v_n^2 = (x_1 + \text{sign}(x_1)\|x\|_2)^2 + x_2^2 + \ldots + x_n^2 = \\
&= x_1^2 + 2x_1\text{sign}(x_1)\|x\|_2 + \|x\|_2^2 + x_2^2 + \ldots + x_n^2 = 2\|x\|_2^2 + 2x_1\text{sign}(x_1)\|x\|_2
\end{aligned}$$

tj. $v^T v = 2\|x\|_2\,(\|x\|_2 + x_1\text{sign}(x_1))$. Kako je $\text{sign}(x_1)^2 = 1$, vrijedi:

$$v^T v = 2\,\text{sign}(x_1)\|x\|_2(\text{sign}(x_1)\|x\|_2 + x_1) = -2\sigma(x_1 - \sigma) = -2\sigma v_1.$$

Tada je:

$$\beta = \frac{2}{v^T v} = \frac{2}{-2\,\sigma\, v_1} = -\frac{1}{\sigma\, v_1}.$$

Nakon što smo objasnili proces nalaženja Householderove matrice P za dane vektore, možemo sad točno opisati kako pomoću Householderovih matrica izračunati QR faktorizaciju. Proces koji smo ranije spominjali kako iz matrice A dobivamo gornje trokutastu matricu R ćemo ilustrirati na 5×3 matrici.

$$A = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \xrightarrow{P_1} \left[\begin{array}{c|cc} * & * & * \\ \hline 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{array}\right] \xrightarrow{P_2}$$

$$\left[\begin{array}{cc|c} * & * & * \\ 0 & * & * \\ \hline 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{array}\right] \xrightarrow{P_3} \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Opisat ćemo k-ti korak procesa redukcije matrice $A \in \mathbb{R}^{m\times n}$ na gornje trokutastu matricu. S $A_1 = A$, na početku k-tog koraka, imamo:

$$A_k = \left[\begin{array}{c|cc} R_{k-1} & z_k & B_k \\ \hline 0 & x_k & C_k \end{array}\right], \quad R_{k-1} \in \mathbb{R}^{(k-1)\times(k-1)}, \quad x_k \in \mathbb{R}^{m-k+1}$$

gdje je R_{k-1} gornje trokutasta matrica. Cilj nam je vektor x_k pretvoriti u vektor koji ima sve osim prve komponente 0. Pa nalazimo Householderovu matricu $\tilde{P}_k$ tako da je $\tilde{P}_k x_k = \sigma e_1$ i proširujemo tu matricu na matricu:

$$P_k = \begin{bmatrix} I_{k-1} & 0 \\ 0 & \tilde{P}_k \end{bmatrix}$$

Matrica $P_k \in \mathbb{R}^{m\times m}$ je također Householderova matrica, jer ako je $\tilde{v}_k$ vektor s kojim se konstruira matrica $\tilde{P}_k$, onda matricu P_k možemo konstruirati vektorom:

$$v_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tilde{v}_{k1} \\ \vdots \\ \tilde{v}_{k(m-k+1)} \end{bmatrix}$$

tj. $P_k = I - \frac{2}{v_k^T v_k} v_k v_k^T$. Tada definiramo $A_{k+1} = P_k A_k$. Izvršavanjem n koraka, dobivamo $R = P_n P_{n-1} \dots P_1 A$ (u slučaju $m = n$ je $P_n = I$) te $Q = P_1 P_2 \dots P_n$. Kako su Householderove matrice P_k ortogonalne, slijedi da je matrica Q ortogonalna. Dobili smo $A = QR$, gdje je Q ortogonalna, a R gornje trokutasta. U praksi se Householderove matrice P_k nikad ne formiraju, pohranjuju i koriste se samo Householderovi vektori v_k. Npr., za izračunati A_{k+1} trebamo izračunati produkt $\tilde{P}_k C_k$. To možemo raspisati:

$$\tilde{P}_k C_k = \left(I - \beta v v^T\right) C_k = C_k - \beta v \left(v^T C_k\right)$$

iz čega vidimo da nam je dovoljan vektor v za izračunati produkt. Ovaj pristup je također puno efikasniji nego formiranje matrice $\tilde{P}_k$ i množenje matrica, jer ovdje imamo množenje matrice i vektora te vanjski produkt vektora nakon toga. Sveukupno, Householderova redukcija na gornje trokutastu matricu zahtjeva $2n^2(m - \frac{n}{3})$ flopsa. Eksplicitna formacija matrice $Q = P_1 P_2 \dots P_n$ se vrši s desna na lijevo (efikasnije nego s lijeva na desno) zbog samih struktura matrica P_k (efektivna dimenzija problema raste s $m-n$ na m, dok s lijeva na desno, je ona cijelo vrijeme m). Računanje s desna na lijevo zahtjeva $4(m^2 n - mn^2 + n^3/3)$ flopsa, odnosno $2n^2(m-\frac{n}{3})$ flopsa, ako se računa samo prvih n stupaca matrice Q. Za većinu primjena (poput rješavanja problema najmanjih kvadrata) Q će se ostaviti u faktoriziranoj formi.

Givensova transformacija

Drugi način za izračunati QR faktorizaciju je Givensovim rotacijama. Givensova rotacija (ili rotacija ravnine) $G = G(i, j, \theta) \in \mathbb{R}^{n \times n}$ je jednaka identiteti osim dijela matrice:

$$G([i, j], [i, j]) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

uz oznake, $c = \cos\theta$ i $s = \sin\theta$. Koristimo Matlab notaciju te ona označava:

$$G([i, j], [i, j]) = \begin{bmatrix} G_{ii} & G_{ij} \\ G_{ji} & G_{jj} \end{bmatrix}$$

Produkt $y = G(i, j, \theta)x$ rotira x za θ radijana u smjeru kazaljke na sat u (i, j) ravnini. Algebarski imamo:

$$y_k = \begin{cases} x_k, & k \neq i, j, \\ c x_i + s x_j, & k = i, \\ -s x_i + c x_j, & k = j. \end{cases}$$

Slično kao i s Householderovim transformacijama, ideja je naći kut Givensove rotacije tako da je $y_j = 0$. Kako su Givensove rotacije ortogonalne matrice, problem nam se svodi na:

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x_i \\ x_j \end{bmatrix} = \begin{bmatrix} \left\| \begin{bmatrix} x_i \\ x_j \end{bmatrix} \right\|_2 \\ 0 \end{bmatrix} = \begin{bmatrix} \sqrt{x_i^2 + x_j^2} \\ 0 \end{bmatrix}$$

Rješavanjem ovog sustava se dobije:

$$s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}}, \quad c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} \tag{1.1}$$

Dobili smo način kako staviti $y_j = 0$ te i način kako izračunati tu Givensovu rotaciju bez nalaženja točnog kuta, računat ćemo $\sin\theta$ i $\cos\theta$ pomoću formule (1.1). U praksi, može doći do nepreciznog izračuna radijusa $r = \sqrt{x_i^2 + x_j^2}$ radi "overflowa" pa da bi to izbjegli, možemo skalirati problem.

Da bi izračunali QR faktorizaciju, Givensove rotacije koristimo da bi eliminirali elemente ispod dijagonale. Redoslijed eliminacije elemenata ćemo ilustrirati na matrici 5×3:

$$A = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \xrightarrow{G_{45}} \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ 0 & * & * \end{bmatrix} \xrightarrow{G_{34}} \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \xrightarrow{G_{23}} \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \xrightarrow{G_{12}}$$

$$\left[\begin{array}{c|cc} * & * & * \\ \hline 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{array}\right] \xrightarrow{G_{45}} \left[\begin{array}{c|cc} * & * & * \\ \hline 0 & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{array}\right] \xrightarrow{G_{34}} \left[\begin{array}{c|cc} * & * & * \\ \hline 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{array}\right] \xrightarrow{G_{23}} \left[\begin{array}{cc|c} * & * & * \\ 0 & * & * \\ \hline 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{array}\right] \xrightarrow{G_{45}}$$

$$\left[\begin{array}{cc|c} * & * & * \\ 0 & * & * \\ \hline 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{array}\right] \xrightarrow{G_{34}} \left[\begin{array}{ccc} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] = R$$

Iz ilustracije, intuitivno je da je potrebno više operacija za Givens QR faktorizaciju nego za Householder QR faktorizaciju. Za generalnu $m \times n$ matricu ($m \geq n$) potrebno je $3n^2(m - \frac{n}{3})$ flopsa što je za 50% više nego za Householder QR faktorizaciju. Također vidimo iz ove ilustracije da kad bi imali matricu punu 0, broj rotacija bi se smanjivao pa je tako

Givens QR faktorizacija najkorisnija za matrice s puno 0, poput tridijagonalne matrice ili Hessenbergove matrice.

Za analizu greške će nam trebati pojam disjunktnih Givensovih rotacija. Rotacije $G_{i_1,j_1}, \ldots, G_{i_r,j_r}$ su disjunktne ako vrijedi $\{i_s, j_s\} \cap \{i_t, j_t\} = \emptyset$ za $s \neq t$. Disjunktne rotacije komutiraju jer djeluju na različite dijelove vektora, matematički ni numerički nije bitno kojim se redom disjunktne rotacije primjenjuju. Naš pristup će biti uzeti dani niz rotacija i rasporediti ih u grupe disjunktnih rotacija. Raspoređeni algoritam će biti numerički ekvivalentan originalnom, ali će nam olakšat analizu greške.

Kao primjer niza rotacija koji je poredan u disjunktne grupe, dajemo ilustraciju matrice 6×3:

$$\begin{bmatrix} * & * & * \\ 1 & * & * \\ 2 & 3 & * \\ 3 & 4 & 5 \\ 4 & 5 & 6 \\ 5 & 6 & 7 \end{bmatrix}$$

Cijeli broj k na poziciji (i, j) označava da će element (i, j) biti eliminiran u k-tom koraku rotacijom u (j, i) ravnini (G_{ji} će poništiti (i, j)-ti element). Sve rotacije u k-tom koraku su disjunktne (u 5. koraku imamo rotacije G_{16}, G_{25}, G_{34}, te po definiciji vidimo da su one disjunktne). Za matricu $m \times n$ s $m > n$ imamo $r = m + n - 2$ koraka, i Givensova QR faktorizacija se može zapisati kao $W_r W_{r-1} \ldots W_1 A = R$, gdje su W_i produkti od najviše n disjunktnih rotacija.

Gram-Schmidtova ortogonalizacija

Treća i najstarija metoda za računanje QR faktorizacije je Gram-Schmidtova metoda ortogonalizacije. Može se direktno izvesti iz formule $A = QR$, gdje je $A \in \mathbb{R}^{m \times n}$, $Q \in \mathbb{R}^{m \times n}$ ortogonalna i $R \in \mathbb{R}^{n \times n}$ gornje trokutasta. Primijetit ćemo da se ovdje formula razlikuje od definicije te da se zapravo ne računa puna QR faktorizacija (jer se računa samo Q_1). Označimo s a_j, q_j, r_j j-ti stupac od matrica A, Q, R respektivno. Zbog toga što je R gornje trokutasta imamo:

$$r_j = \begin{bmatrix} r_{1j} \\ r_{2j} \\ \vdots \\ r_{jj} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Iz formule onda imamo:

$$a_j = Qr_j = \sum_{k=1}^{j} r_{kj} q_k \tag{1.2}$$

Kako je Q ortonormalna matrica, množenjem s q_i^T dobivamo:

$$q_i^T a_j = r_{ij}, \quad i = 1 : j-1,$$

Dok iz (1.2) imamo:

$$q_j = \frac{q_j'}{r_{jj}}, \quad r_{jj} = \|q_j'\|_2,$$

gdje je:

$$q_j' = a_j - \sum_{k=1}^{j-1} r_{kj} q_k.$$

Ovakvim postupkom možemo računati Q i R stupac po stupac. Da bi osigurali $r_{jj} > 0$, zahtijevamo da A ima puni rang.

Algoritam 1. (klasični Gram-Schmidt (CGS)) Dana nam je matrica $A \in \mathbb{R}^{m\times n}$ ranga n, ovaj algoritam računa QR faktorizaciju $A = QR$, gdje je $Q \in \mathbb{R}^{m\times n}$ i $R \in \mathbb{R}^{n\times n}$, s Gram-Schmidtovom metodom.

for j = 1 : n **do**
  for i = 1 : j-1 **do**
    $r_{ij} = q_i^T a_j$
  end for
  $q_j' = a_j - \sum_{k=1}^{j-1} r_{kj} q_k$
  $r_{jj} = \|q_j'\|_2$
  $q_j = \frac{q_j'}{r_{jj}}$
end for

CGS metoda zahtjeva $2mn^2$ flopsa ($\frac{2n^3}{3}$ flopsa više nego Householder QR faktorizacija u faktoriziranoj formi).

U metodi CGS se a_j pojavljuje samo u $j-$tom koraku. Metoda se može preurediti tako da čim izračunamo q_j, sve preostale vektore ćemo ortogonalizirati spram q_j. To nam daje modificiranu Gram-Schmidt metodu (MGS).

Algoritam 2. (Modificirana Gram-Schimdt metoda(MGS)) Dana nam je matrica $A \in \mathbb{R}^{m\times n}$ ranga n, ovaj algoritam računa QR faktorizaciju $A = QR$, gdje je $Q \in \mathbb{R}^{m\times n}$ i $R \in \mathbb{R}^{n\times n}$, s MGS metodom.

$a_k^{(1)} = a_k$, $k = 1 : n$

for k = 1 : n **do**
  $r_{kk} = \|a_k^{(k)}\|_2$
  $q_k = \frac{a_k^{(k)}}{r_{kk}}$
  for j = $k + 1 : n$ **do**
    $r_{kj} = q_k^T a_j^{(k)}$
    $a_j^{(k+1)} = a_j^{(k)} - r_{kj} q_k$
  end for
end for

MGS metoda također zahtjeva $2mn^2$ flopsa. U MGS metodi možemo primijetiti da se preostali vektori matrice A ažuriraju jednom po koraku da bi bili ortogonalni s novo izračunati vektorom matrice Q. Takvim postupkom su preostali vektori ortogonalni sa svim vektorima do tad izračunatim. Također, možemo primijetiti da u MGS metodi je $r_{kj} = q_k^T a_j^{(k)}$, gdje smo koristili parcijalno ortogonalan vektor $a_j^{(k)}$.

MGS metoda se može izraziti i matrično. Definirajmo $A_k = \left[q_1, \ldots, q_{k-1}, a_k^{(k)}, \ldots, a_n^{(k)}\right]$. MGS transformira $A_1 = A$ u $A_{n+1} = Q$ nizom transformacija $A_k = A_{k+1} R_k$, gdje je:

$$R_k = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ 0 & 0 & \ldots & r_{kk} & r_{kk+1} & \ldots & r_{kn} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}$$

odnosno R_k je jednaka identiteti osim u k-tom retku, gdje se poklapa s matricom R (iz MGS metode). Ako preciznije pogledamo $A_k = A_{k+1} R_k$, vidjet ćemo da u j-tom stupcu, gdje je $j < k$, imamo $q_j = q_j$. Za $j = k$ imamo $a_k^{(k)} = r_{kk} q_k$, tj. $q_k = a_k^{(k)} / r_{kk}$. Za $j > k$ imamo $a_j^{(k)} = a_j^{(k+1)} + r_{kj} q_k$, tj. $a_j^{(k+1)} = a_j^{(k)} - r_{kj} q_k$. Iz ovoga vidimo kako je matrični zapis ekvivalentan MGS metodi.

Kako bi pojednostavili analizu greške MGS metode, pokazat ćemo da postoji veza između MGS metode i Householder QR faktorizacije. Tako ćemo u analizi greške MGS metode moći koristiti rezultate koje ćemo dokazati za Householder QR faktorizaciju. Za matricu $A \in \mathbb{R}^{m \times n}$, definiramo proširenu matricu $\begin{bmatrix} 0_n \\ A \end{bmatrix} \in \mathbb{R}^{(m+n) \times n}$. Promotrimo njenu Householder QR faktorizaciju. Postoji matrica $P \in \mathbb{R}^{(m+n) \times (m+n)}$ i matrica $R \in \mathbb{R}^{n \times n}$, takve da vrijedi:

$$P^T \begin{bmatrix} 0_n \\ A \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad P^T = P_n \ldots P_2 P_1.$$

Pokazat ćemo da istovremeno možemo izračunati rezultat MGS metode za matricu A i Householder QR faktorizacije za proširenu matricu $\begin{bmatrix} 0_n \\ A \end{bmatrix}$. Kao u MGS metodi, za prvi stupac matrice A, definiramo $q_1 = a_1/\|a_1\|_2$ te definiramo:

$$P_1 = I - v_1 v_1^T, \quad v_1 = \begin{bmatrix} -e_1 \\ q_1 \end{bmatrix}, \quad v_1^T v_1 = 2,$$

iz množenja $A_2 = P_1 \begin{bmatrix} 0_n \\ A \end{bmatrix}$ dobijamo:

$$A_2 = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & 0 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 0 \\ 0 & a_2^{(2)} & \dots & a_n^{(2)} \end{bmatrix},$$

gdje smo množenjem dobili vektore $a_k^{(2)}$ koji su identični vektorima iz MGS metode te prvi red matrice R. Tada možemo definirati $q_2 = a_2^{(2)}/\|a_2^{(2)}\|_2$. Ponavljamo postupak za $k = 2, \dots, n$, te svaki put definiramo:

$$P_k = I - v_k v_k^T, \quad v_k = \begin{bmatrix} -e_k \\ q_k \end{bmatrix}, \quad , v_k^T v_k = 2,$$

te iz rezultata množenja $A_{k+1} = P_k A_k$ možemo definirati vektor $q_{k+1} = a_{k+1}^{(k+1)}/\|a_{k+1}^{(k+1)}\|_2$. Na kraju ćemo dobiti sve vektore MGS metode $q_1, \dots, q_n$, matricu R tako da je $A = QR$ te matricu P tako da je

$$P^T \begin{bmatrix} 0_n \\ A \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix}.$$

1.2 Problem najmanjih kvadrata

Neka je $m \geq n$. Promotrimo problem pronalaska vektora $x \in \mathbb{R}^n$ takvog da je $Ax = b$, gdje su nam dani $A \in \mathbb{R}^{m \times n}$ i $b \in \mathbb{R}^m$. Kad imamo više jednadžbi nego nepoznanica, sustav najčešće nema rješenje pa nam postaje cilj minimizirati izraz $\|Ax - b\|_p$. Problem najmanjih kvadrata je minimizacijski problem

$$\min_x \|Ax - b\|_p.$$

Mi ćemo se baviti problemom najmanjih kvadrata u kojem matrica A ima puni stupčani rang, $r(A) = n$. Također zanimat će nas 2- norma jer je unitarno invarijantna. U slučaju

$m = n$, matrica A je regularna pa je rješenje problema najmanjih kvadrata jednostavno i ono je rješenje sustava $Ax = b$, $x = A^{-1}b$. Neka je $x, z \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, $\alpha \in \mathbb{R}$ te matrica $A \in \mathbb{R}^{m\times n}$.

Neka je $r = r(A) \leq n$. Promotrimo sljedeću jednakost:

$$\|A(x + \alpha z) - b\|_2^2 = \|Ax - b\|_2^2 + 2\alpha z^T A^T(Ax - b) + \alpha^2\|Az\|_2^2. \tag{1.3}$$

Ako je x rješenje problema najmanjih kvadrata $\min_x \|Ax - b\|_2$ slijedit će da je $A^T(Ax - b) = 0$. Inače, ako bi uzeli $z = -A^T(Ax - b)$ te dovoljno mali α, dobili bi kontradikciju $\|A(x + \alpha z) - b\|_2 < \|Ax - b\|_2$. Jednadžbe $A^T(Ax - b) = 0$ možemo zapisati i na sljedeći način $A^TAx = A^Tb$. Takve jednadžbe zovemo normalne jednadžbe.

Pomoću normalnih jednadžbi možemo pokazati egzistenciju rješenja problema najmanjih kvadrata. Prvo ćemo pokazati da normalne jednadžbe uvijek imaju rješenje. Dovoljno nam je pokazati da je $A^Tb \in \mathrm{Im}(A^TA)$ jer onda slijedi da postoji $x \in \mathbb{R}^n$ takav da je $A^TAx = A^Tb$. Tvrdimo da je $\mathrm{Im}(A^TA) = \mathrm{Im}(A^T)$ pa kako je $A^Tb \in \mathrm{Im}(A^T)$ slijedi da je $A^Tb \in \mathrm{Im}(A^TA)$. Vrijedi da je $\mathrm{Im}(A^TA) \subseteq \mathrm{Im}(A^T)$ jer za $x \in \mathrm{Im}(A^TA)$ postoji $y \in \mathbb{R}^n$ tako da je $x = A^TAy = A^T(Ay) = A^Tz$ pa je $x \in \mathrm{Im}(A^T)$. Još ostaje pokazati $\mathrm{Im}(A^T) \subseteq \mathrm{Im}(A^TA)$. Neka je $x \in \mathrm{Im}(A^T)$, tj. postoji $y \in \mathbb{R}^m$ tako da je $x = A^Ty$. Kako je $\mathbb{R}^m = \mathrm{Ker}(A^T) \oplus \mathrm{Im}(A)$, slijedi da je $y = y_1 + y_2$, gdje je $y_2 \in \mathrm{Im}(A)$, tj. postoji $w \in \mathbb{R}^n$ tako da je $Aw = y_2$. Tada je

$$x = A^Ty = A^Ty_1 + A^Ty_2 = A^TAw,$$

gdje smo u 3. nejednakosti iskoristili činjenicu da je $y_1 \in \mathrm{Ker}(A^T)$ i $y_2 = Aw$. Tada je $x \in \mathrm{Im}(A^TA)$. Sad tvrdimo da ako je $x \in \mathbb{R}^n$ rješenje normalne jednadžbe da je $x \in \mathbb{R}^n$ rješenje problema najmanjih kvadrata. Neka je $y \in \mathbb{R}^n$ proizvoljan vektor. Vrijedi:

$$\begin{aligned}\|A(x + y) - b\|_2^2 &= \|Ax - b\|_2^2 + 2y^TA^T(Ax - b) + \|Ay\|_2^2\\ &= \|Ax - b\|_2^2 + \|Ay\|_2^2\\ &\geq \|Ax - b\|_2^2,\end{aligned}$$

gdje smo koristili u 2. jednakosti činjenicu da je x rješenje normalne jednadžbe. S ovime smo pokazali egzistenciju rješenja problema najmanjih kvadrata.

Također, iz jednakosti (1.3) možemo zaključiti da ako su x i $x + \alpha z$ rješenja problema najmanjih kvadrata da će vrijediti da je $z \in \mathrm{Ker}(A)$. U našim slučajevima kako je $r(A) = n$ po teoremu o rangu i defektu će vrijediti da je $d(A) = 0$, odnosno $\mathrm{Ker}(A) = 0$ pa slijedi da imamo jedinstveno rješenje problema najmanjih kvadrata.

Rješenje QR faktorizacijom

Neka je $A \in \mathbb{R}^{m\times n}$ i neka je $r(A) = n$. Neka je

$$A = Q\begin{bmatrix}R\\0\end{bmatrix}$$

QR faktorizacija matrice A. Tada je

$$\|Ax-b\|_2^2 = \|Q^TAx - Q^Tb\|_2^2 =: \left\| \begin{bmatrix} Rx-c \\ -d \end{bmatrix} \right\|_2$$
$$= \|Rx-c\|_2^2 + \|d\|_2^2. \tag{1.4}$$

Kako je $r(A) = n$, matrica R je regularna pa je jedinstveno rješenje problema najmanjih kvadrata $x = R^{-1}c$ te je $\|r\|_2 = \|d\|_2$. Stoga se problem najmanjih kvadrata može rješiti s relativno malo posla osim računanja QR faktorizacije. Matrica Q nije potrebna eksplicitno, trebamo samo znati primijeniti Q^T na vektor. Ako se koristi QR faktorizacija pomoću Householderovih reflektora potrebno je $2n^2(m - \frac{n}{3})$ flopsa.

Rješenje MGS metodom

MGS metoda se može koristiti za rješavanje problema najmanjih kvadrata. Naime, ako je $A = QR$ QR faktorizacija matrice A te imamo problem najmanjih kvadrata $\min_x \|b-Ax\|_2 = \min_x \|b - QRx\|$, ne smijemo x računati kao $x = R^{-1}(Q^Tb)$ zbog nedostatka ortogonalnosti izračunate matrice $\hat{Q}$. To bi utjecalo na stabilnost pa moramo x izračunati na drugi način. Primijenit ćemo MGS metodu na proširenu matricu $\begin{bmatrix} A & b \end{bmatrix}$. Dobijamo:

$$\begin{bmatrix} A & b \end{bmatrix} = \begin{bmatrix} Q_1 & q_{n+1} \end{bmatrix} \begin{bmatrix} R & z \\ 0 & \rho \end{bmatrix},$$

gdje je $z \in \mathbb{R}^n$, a $\rho \in \mathbb{R}$. Tada imamo:

$$Ax - b = \begin{bmatrix} A & b \end{bmatrix} \begin{bmatrix} x \\ -1 \end{bmatrix} = \begin{bmatrix} Q_1 & q_{n+1} \end{bmatrix} \begin{bmatrix} Rx - z \\ -\rho \end{bmatrix}$$
$$= Q_1(Rx - z) - \rho q_{n+1}.$$

Vrijedi da je $\|b - Ax\|_2^2 = \|Rx - z\|_2^2 + \rho^2$ jer je Q_1 ortogonalna matrica, $\|q_{n+1}\|_2 = 1$ te jer je q_{n+1} ortogonalan sa stupcima matrice Q_1. Tada je rješenje $x = R^{-1}z$.

1.3 Uvod u analizu greške

Računajući QR faktorizaciju u računalu, s obzirom na realnu aritmetiku računala, u svakoj operaciji dolazi do zaokruživanja, tj. dolazi do greške. Realna aritmetika računala nije egzaktna te radi toga trebamo analizirati grešku da vidimo koliko se izračunata faktorizacija razlikuje od teoretske. Također, u analizi greške algoritma je bitan pojam stabilnosti algoritma. Stabilni algoritmi prigušavaju grešku, dok je nestabilni pojačavaju. Prvo ćemo definirati potrebne pojmove te opisati realnu aritmetiku računala pa preći na analizu greške računanja QR faktorizacije.

Greške i mjere za greške

Označimo s $\hat{x}$ izračunatu ili približnu vrijednost od x. Zapravo $\hat{x}$ je aproksimacija od x. Najkorisnije mjere točnosti za $\hat{x}$ su apsolutna greška:

$$E_{\text{abs}}(\hat{x}) = |\hat{x} - x|,$$

te relativna greška za $x \neq 0$:

$$E_{\text{rel}}(\hat{x}) = \frac{|\hat{x} - x|}{|x|}$$

Često se koristi i oznaka $\Delta x = \hat{x} - x$ pa je $E_{\text{abs}}(\hat{x}) = |\Delta x|$. Relativna greška ima i alternativnu definiciju. Ako zapišemo $\hat{x} = x(1+\rho)$, onda je $E_{\text{rel}}(\hat{x}) = |\rho|$. U tom slučaju, relativna greška mjeri koliko se $1 + \rho$ razlikuje od 1. U praksi, zanimljivija nam je relativna greška, jer je invarijantna na skaliranje: ako $x \to \alpha x$ te $\hat{x} \to \alpha x$, $E_{\text{rel}}(\hat{x})$ ostaje nepromijenjen.

Model aritmetike

Da bi mogli provesti analizu greške metode, moramo uvesti pretpostavke o točnosti osnovnih aritmetičkih operacija. Definiramo skup $F \subset \mathbb{R}$ čiji elementi imaju sljedeću formu:

$$y = \pm m\beta^{e-t}, \tag{1.5}$$

gdje je β baza, t preciznost, e eksponent za koji vrijedi $e_{\min} \leq e \leq e_{\max}$ te cijeli broj m zovemo signifikand za kojeg vrijedi $0 \leq m \leq \beta^t - 1$. Da bi se osigurala jedinstvena reprezentacija $y \in F$ koji je različit od 0 pretpostavljamo da vrijedi da je $m \geq \beta^{t-1}$. Takav sustav je normaliziran. Broj 0 nema normaliziranu reprezentaciju. Raspon skupa F je $\beta^{e_{\min}-1} \leq |y| \leq \beta^{e_{\max}}(1 - \beta^{-t})$. Alternativni zapis $y \in F$ je sljedeći:

$$y = \pm\beta^e \left(\frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \ldots + \frac{d_t}{\beta_t} \right) = \pm\beta^e \times .d_1 d_2 \ldots d_t,$$

gdje svaka znamenka d_i zadovoljava $0 \leq d_i \leq \beta - 1$ i $d_1 \neq 0$ za normalizirane brojeve.

Neka $G \subset \mathbb{R}$ označava skup čiji elementi imaju formu (1.5), ali bez ograničenja na eksponent e. Ako je $x \in \mathbb{R}$ onda $fl(x)$ označava element iz G koji je najbliži x. Transformacija $x \to fl(x)$ se naaziva zaokruživanje. Sljedeći rezultat pokazuje da se svaki realan broj x koji leži u rasponu skupa F može aproksimirati elementom iz F s relativnom greškom koja nije veća od $u = \frac{1}{2}\beta^{1-t}$. Veličinu u zovemo jediničnom greškom zaokruživanja.

Teorem 1.3.1. *(Higham [4, p. 38]) Ako $x \in \mathbb{R}$ leži u rasponu skupa F tada*

$$fl(x) = x(1 + \delta), \quad |\delta| \leq u.$$

Tipično se koristi IEEE standardna aritmetika u kojoj je $\beta = 2$ i podupire dvije preciznosti. U jednostrukoj preciznosti je $t = 24$, $e_{\min} = -125$, $e_{\max} = 128$ i $u = 2^{-24} \approx 5.96 \times 10^{-8}$. U dvostrukoj preciznosti je $t = 53$, $e_{\min} = -1021$, $e_{\max} = 1024$ i $u = 2^{-53} \approx 1.11 \times 10^{-16}$. Najčešća pretpostavka o točnosti osnovnih aritmetičkih operacija je dana sljedećim modelom:

$$fl(x \text{ op } y) = (x \text{ op } y)(1 + \delta), \quad |\delta| \leq u,\ x, y \in F, \quad \text{op} = +, -, \cdot, /.$$

Također u našu pretpostavku ćemo uključiti i korijen, tj. vrijedit će

$$fl(\sqrt{x}) = \sqrt{x}(1 + \delta), \quad |\delta| \leq u, \quad x \in F.$$

Greška unaprijed i greška unazad

Neka je $\hat{y}$ izračunata vrijednost od $y = f(x)$ u danoj aritmetici, gdje je f realna skalarna funkcija realne skalarne varijable. Definiramo grešku unaprijed kao $E_{\text{abs}}(\hat{y})$ ili $E_{\text{rel}}(\hat{y})$. Za izračunatu vrijednost $\hat{y}$ se možemo zapitati, za koje x' je $\hat{y}$ egzaktno rješenje problema $y = f(x)$, tj. za koje Δx je $\hat{y} = f(x + \Delta x)$. Δx može biti puno, ali nas zanima najmanji. Definiramo grešku unatrag kao vrijednost $|\Delta x|$ (mislimo na $\min |\Delta x|$) ili kao $\frac{|\Delta x|}{|x|}$. Ako označimo s $x' = x + \Delta x$, onda je greška unatrag zapravo $E_{\text{abs}}(x')$ ili $E_{\text{rel}}(x')$.

Metodu za računanje $y = f(x)$ zovemo stabilnom unatrag, ako za proizvoljan x, daje $\hat{y}$ s malom greškom unatrag, tj. ako je $\hat{y} = f(x + \Delta x)$ za dovoljno mali Δx. Općenito, problem $y = f(x)$ može imati više metoda za računanje rješenja, od kojih će neke metode biti stabilne unatrag, neke ne. Neke metode ne moraju imati relativno mali Δx, ali mogu zadovoljavati slabiju relaciju. Definiramo mješanu grešku unaprijed-unazad rezultata s relacijom:

$$\hat{y} + \Delta y = f(x + \Delta x), \quad |\Delta y| \leq \epsilon |y|,\ |\Delta x| \leq \eta |x|. \tag{1.6}$$

Ono što ona zapravo kaže, je da je $\hat{y}$ skoro rješenje problema za skoro točne podatke.

Općenito, algoritam zovemo numerički stabilnim ako zadovoljava relaciju (1.6) za dovoljno male ϵ, η. Kako ϵ može biti 0, možemo zaključiti da su i algoritmi stabilni unatrag, numerički stabilni.

Vektorske i matrične norme

Vektorska norma je funkcija $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ koja zadovoljava sljedeće uvjete:

1. $\|x\| \geq 0$, gdje jednakost vrijedi ako i samo ako je $x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\|$ za svaki $\alpha \in \mathbb{C}$, $x \in \mathbb{C}^n$.
3. $\|x + y\| \leq \|x\| + \|y\|$ za svaki $x, y \in \mathbb{C}^n$.

Nama će najkorisnija biti Euklidska norma ili 2-norma, koja je dana formulom

$$||x||_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}},$$

a ona je specijalan slučaj Hölderove p-norme:

$$||x||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \quad p \geq 1.$$

Koristit ćemo da je 2-norma ortogonalno invarijantna, tj. za ortogonalnu matricu Q, za koju vrijedi $Q^T Q = I$, će vrijediti da je $||Qx||_2^2 = (Qx)^T Qx = xQ^T Qx = x^T x = ||x||_2^2$. Također koristit ćemo Cauchy-Schwarz nejednakost:

$$|x^T y| \leq ||x||_2 ||y||_2, \quad x, y \in \mathbb{R}^n.$$

Matrična norma je funkcija $|| \cdot || : \mathbb{C}^{m \times n} \to \mathbb{R}$ koja zadovoljava analogne uvjete onima za vektorsku normu. Nama će najkorisnije matrične norme biti Frobeniusova norma koja je dana formulom

$$||A||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}} = (tr(A^T A))^{\frac{1}{2}}$$

te matrična 2-norma koja je dana formulom

$$||A||_2 = \left(\rho(A^T A) \right)^{\frac{1}{2}} = \sigma_{\max}(A),$$

gdje je

$$\rho(A) = \max \{ |\lambda| : \det(A - \lambda I) = 0 \}$$

te $\sigma_{\max}(A)$ najveća singularna vrijednost matrice A. Matrična 2-norma je specijalan slučaj operatorske norme. Operatorske norme su matrične norme inducirane vektorskim normama te se za vektorsku normu $|| \cdot ||_p$ definira:

$$||A||_p = \max_{x \neq 0} \frac{||Ax||_p}{||x||_p}.$$

Iz definicije se lako vidi da je $||Ax||_p \leq ||A||_p ||x||_p$. Kažemo da je norma konzistentna ako vrijedi $||AB|| \leq ||A|| ||B||$ kad god je produkt AB definiran. Frobeniusova norma te sve operatorske norme su konzistentne. Frobeniusova norma i matrična 2-norma su ortogonalno invarijatne norme, tj. $||UAV|| = ||A||$, za sve U, V ortogonalne. Korisna će nam biti relacija ekvivalencije između Frobeniusove i matrične 2-norme. Vrijedi:

$$||A||_2 \leq ||A||_F \leq \sqrt{r(A)} ||A||_2,$$

gdje smo s $r(A)$ označili rang matrice A. Sad kad smo uveli matrične norme, možemo definirati pojam uvjetovanosti matrice $\kappa(A) = \|A\|\|A^{-1}\|$ koji ćemo dosta koristiti u analizi greške QR faktorizacije. Označimo j-ti stupac matrice A s a_j te definirajmo vektor $e = (1, 1, \ldots, 1)^T$. Navodimo još tehničku lemu koja će biti jako korisna u analizi.

Lema 1.3.2. *(Higham [4, p. 111-112]) Neka su $A, B \in \mathbb{R}^{m\times n}$.*

a) Ako je $\|a_j\|_2 \leq \|b_j\|_2,\ j = 1, \ldots, n$, tada je:

$$\|A\|_F \leq \|B\|_F, \quad \|A\|_2 \leq \sqrt{r(B)}\|B\|_2, \quad |A| \leq ee^T|B|.$$

b) Ako $|A| \leq B$ tada $\|A\|_2 \leq \|B\|_2$.

c) Ako $|A| \leq |B|$ tada $\|A\|_2 \leq \sqrt{r(B)}\|B\|_2$.

d) $\|A\|_2 \leq \||A|\|_2 \leq \sqrt{r(A)}\|A\|_2$.

Osnove analize greške

Nakon što smo definirali model aritmetike, navest ćemo neke od osnovnih rezultata koji su nam potrebni za provesti analizu greške QR faktorizacije. Sljedeća lema uvodi γ_n notaciju koja će biti fundamentalna za analizu greške.

Lema 1.3.3. *(Higham [4, p. 63]) Ako je $|\delta_i| \leq u$ i $\rho_i = \pm 1$ za $i = 1, \ldots, n$ i $nu < 1$, tada*

$$\prod_{i=1}^{n}(1 + \delta_i)^{\rho_i} = (1 + \theta_n),$$

gdje

$$|\theta_n| \leq \frac{nu}{1 - nu} =: \gamma_n.$$

Kao posljedica leme, za vektore $x, y \in \mathbb{R}^n$, vrijedi:

$$fl(x^T y) = (x + \Delta x)^T y = x^T(y + \Delta y), \quad |\Delta x| \leq \gamma_n|x|, \quad |\Delta y| \leq \gamma_n|y|, \tag{1.7}$$

gdje za vektore $a, b \in \mathbb{R}^n$ kažemo da je $|a| \leq |b|$, ako je $|a_i| \leq |b_i|$, $i = 1, \ldots, n$. Ovo će vrijediti za proizvoljan odabir redoslijeda evaluacije skalarnog produkta $x^T y$. Navodimo lemu koja nam daje osnove aritmetike s θ_n i γ_n notacijom.

Lema 1.3.4. *(Higham [4, p. 67]) Za bilo koji prirodan broj k, označimo s θ_k veličinu koja je ograničena prema $|\theta_k| \leq \gamma_k$. Sljedeće relacije vrijede:*

$$(1+\theta_k)(1+\theta_j) = (1+\theta_{k+j}),$$
$$\frac{1+\theta_k}{1+\theta_j} = \begin{cases} 1+\theta_{k+j}, & j \leq k, \\ 1+\theta_{k+2j}, & j > k, \end{cases}$$
$$\gamma_k\gamma_j \leq \gamma_{\min(k,j)}, \quad za\ \max(j,k)u \leq \frac{1}{2},$$
$$i\gamma_k \leq \gamma_{ik},$$
$$\gamma_k + u \leq \gamma_{k+1},$$
$$\gamma_k + \gamma_j + \gamma_k\gamma_j \leq \gamma_{k+j}.$$

Također, u nekim analizama neće biti bitno točno pratiti konstante uz članove γ_n pa je korisna sljedeća notacija:

$$\tilde{\gamma}_k := \frac{cku}{1-cku},$$

gdje je c neki mali prirodan broj. Ona sama po sebi stvara malu aritmetiku, npr. $3\gamma_n = \tilde{\gamma}_n$, $n\tilde{\gamma}_m = m\tilde{\gamma}_n = \tilde{\gamma}_{mn}$. Za matricu $A \in \mathbb{R}^{m\times n}$ te vektor $x \in \mathbb{R}^n$, neka je $y = Ax$. Tada se iz (1.7) lako može izvesti:

$$\hat{y} = (A + \Delta A)x, \quad |\Delta A| \leq \gamma_n |A|.$$

Sljedeće leme će biti korisna u analizi greške Householderove QR faktorizacije te MGS QR faktorizacije.

Lema 1.3.5. *(Higham [4, p. 73]) Ako $X_j + \Delta X_j \in \mathbb{R}^{n\times n}$ zadovoljava $\|\Delta X_j\| \leq \delta_j \|X_j\|$, za svaki $j = 0, \ldots, m$ i gdje je norma konzistentna, tada*

$$\left\| \prod_{j=0}^{m} (X_j + \Delta X_j) - \prod_{j=0}^{m} X_j \right\| \leq \left(\prod_{i=0}^{j} (1+\delta_j) - 1 \right) \prod_{j=0}^{m} \|X_j\|.$$

Lema 1.3.6. *(Higham [4, p. 74]) Neka su $a, b, x \in \mathbb{R}^n$ i neka je $y = (I - ab^T)x$. Za $\hat{y} = fl(x - a(b^T x))$ će vrijediti da je $\hat{y} = y + \Delta y$, gdje*

$$|\Delta y| \leq \gamma_{n+3}(I + |a||b^T|)|x|,$$

te iz toga

$$\|\Delta y\|_2 \leq \gamma_{n+3}(1 + \|a\|_2\|b\|_2)\|x\|_2.$$

Sljedeće navodimo rezultate koji će nam biti potrebni za analizu greške MGS QR faktorizacije.

Teorem 1.3.7. *(Björck i Paige, [1]) Za bilo koje matrice koje zadovoljavaju:*

$$\begin{bmatrix} \Delta A_1 \\ A + \Delta A_2 \end{bmatrix} = \begin{bmatrix} P_{11} \\ P_{21} \end{bmatrix} R, \quad P_{11}^T P_{11} + P_{21}^T P_{21} = I,$$

gdje obje matrice P_{11}, P_{21} imaju barem redaka koliko i stupaca, tada postoji ortogonalna matrica Q takva da $A + \Delta A = QR$, gdje:

$$\Delta A = F\Delta A_1 + \Delta A_2, \quad \|F\|_2 \leq 1.$$

Definiramo polarnu dekompoziciju matrice $A \in \mathbb{R}^{m\times n}$ kao faktorizaciju matrice oblika $A = UH$, gdje $U \in \mathbb{R}^{m\times n}$ ima ortonormalne stupce, a $H \in \mathbb{R}^{n\times n}$ je simetrična pozitivno semidefinitna matrica. Polarna dekompozicija uvijek postoji te je P uvijek jedinstvena.

Neka je $A \in \mathbb{R}^{m\times n}$ te neka je r rang matrice A. Definiramo kompaktnu dekompoziciju singularnih vrijednosti (SVD) matrice A kao faktorizaciju oblika $A = U\Sigma V^T$ gdje su matrica $\Sigma \in \mathbb{R}^{r\times r}$ dijagonalna matrica, čije su vrijednosti singularne vrijednosti matrice A koje su različite od 0, matrica $U \in \mathbb{R}^{m\times r}$ matrica koja ima ortonormalne stupce te matrica $V \in \mathbb{R}^{n\times r}$ koja ima ortonormalne stupce takve da vrijedi $U^TU = V^TV = I_r$. Postoji veza između polarne dekompozicije te SVD dekompozicije. Neka je $A = U_1\Sigma V^T$ SVD dekompozicija te $A = U_2H$ polarna dekompozicija. Tada je:

$$\begin{aligned} H &= V\Sigma V^T, \\ U_2 &= U_1V^T. \end{aligned} \tag{1.8}$$

Pomoću ovih dekompozicija dobijamo sljedeće rezultate.

Lema 1.3.8. *(Higham [3]) Neka matrica $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ ima polarnu dekompoziciju $A = UH$. Tada vrijedi:*

$$\frac{\|A^TA - I\|_2}{1 + \|A\|_2} \leq \|A - U\|_2 \leq \frac{\|A^TA - I\|_2}{1 + \sigma_{\min}(A)}$$

Teorem 1.3.9. *(Higham [2, p. 7]) Neka matrica $A \in \mathbb{R}^{m\times n}$, $m \geq n$, ima polarnu dekompoziciju $A = UH$. Tada ako $Q \in \mathbb{R}^{m\times n}$ ima ortonormalne stupce vrijedi:*

$$\|A - U\|_2 \leq \|A - Q\|_2.$$

Za matricu $A \in \mathbb{R}^{m\times n}$ definiramo pseudo-inverz $X \in \mathbb{R}^{n\times m}$ koji se definira kao jedinstvena matrica koja zadovoljava četiri Moore-Penrose uvjeta:

$$\begin{aligned} &1. \quad AXA = A, \quad 2. \quad XAX = X, \\ &3. \quad AX = (AX)^T, \quad 4. \quad XA = (XA)^T. \end{aligned}$$

Često ćemo pseudo-inverz matrice A označavati s A^+. U slučaju da je $m = n$ i da A ima puni rang, pseudo-inverz je jednak inverzu matrice A. U slučaju $m > n$ i da A ima puni stupčani rang, može se pokazati da je pseudo-inverz $A^+ = (A^TA)^{-1}A^T$ pa slijedi da je $A^+A = (A^TA)^{-1}A^TA = I_n$. U slučaju pravokutne matrice $A \in \mathbb{R}^{m\times n}$ definiramo uvjetovanost matrice kao $\kappa(A) = \|A\|\,\|A^+\|$. Sljedeća lema je rezultat koji se često koristi u analizi matrica.

Teorem 1.3.10. *Neka je $A \in \mathbb{R}^{m\times m}$. Ako je $\|A\| < 1$ te $\|\cdot\|$ konzistentna norma onda je $I_m - A$ regularna matrica te vrijedi:*

$$(I_m - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

Dokaz. Zbog konzistentnosti norme vrijedi $\|A^k\| \leq \|A\|^k$, za $k \in \mathbb{N}$ te vrijedi:

$$\sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|},$$

jer je $\|A\| < 1$, dobili smo da je $\sum_{k=0}^{\infty} A^k$ apsolutno konvergentan red. Kako je $\mathbb{R}^{m\times m}$ Banachov prostor slijedi da je $\sum_{k=0}^{\infty} A^k$ konvergentan red. Iz:

$$(I_m - A)(I_m + A + \ldots + A^n) = (I_m - A^{n+1})$$

Puštanjem limesa $n \to \infty$ dobijamo:

$$(I_m - A)\sum_{k=0}^{\infty} A^k = I_m,$$

jer je $\lim_{n\to\infty} A^{n+1} = 0$ radi $\|A\| < 1$. Analogno se dobije:

$$\left(\sum_{k=0}^{\infty} A^k\right)(I_m - A) = I_m,$$

pa zaključujemo da je:

$$(I_m - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

□

Glavni rezultat analize greške problema najmanjih kvadrata je Wedinov perturbacijski teorem. Wedinov perturbacijski teorem mjeri osjetljivost problema najmanjih kvadrata na perturbacije. Da bi mogli dokazati Wedinov perturbacijski teorem, potrebne su nam dvije leme. Prvo navodimo Weylov teorem koji će nam biti potreban za dokazati prvu od dvije leme.

Teorem 1.3.11. *(Weyl, [8], [6, p. 3]) Neka su $A, B \in \mathbb{R}^{m\times n}$, $m \geq n$. Tada vrijedi:*

$$|\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|_2, \quad k = 1, \ldots, n.$$

Lema 1.3.12. *Neka su $A, B \in \mathbb{R}^{m\times n}$. Ako je $r(A) = r(B)$ i $\eta = \|A^+\|_2\|A - B\|_2 < 1$, onda vrijedi:*

$$\|B^+\|_2 \leq \frac{1}{1-\eta}\|A^+\|_2.$$

Dokaz. Neka je $r = r(A)$. Po Teoremu 1.3.11 vrijedi $|\sigma_r(A) - \sigma_r(B)| \leq \|A - B\|_2$, odnosno:

$$\sigma_r(B) \geq \sigma_r(A) - \|A - B\|_2.$$

Kako su matrice A i B ranga r, znamo da je $\|A^+\|_2 = \frac{1}{\sigma_r(A)}$ i $\|B^+\|_2 = \frac{1}{\sigma_r(B)}$. Tada je:

$$\frac{1}{\|B^+\|_2} \geq \frac{1}{\|A^+\|_2} - \|A - B\|_2 = \frac{1-\eta}{\|A^+\|_2} > 0.$$

Iz toga slijedi:

$$\|B^+\|_2 \leq \frac{1}{1-\eta}\|A^+\|_2.$$

□

Radi potrebe sljedeće leme te samog dokaza Wedinovog teorema definiramo $P_A = AA^+$, ortogonalan projektor na $Im(A)$.

Lema 1.3.13. *Neka su $A, B \in \mathbb{R}^{m\times n}$. Ako je $r(A) = r(B)$ tada je*

$$\|P_A(I - P_B)\|_2 = \|P_B(I - P_A)\|_2 \leq \|A - B\|_2 \min\{\|A^+\|_2, \|B^+\|_2\}.$$

Dokaz. Vrijedi:

$$\begin{aligned}
\|P_B(I - P_A)\|_2 &= \left\|(P_B(I - P_A))^T\right\|_2 = \|(I - P_A)P_B\|_2 \\
&= \left\|(I - AA^+)BB^+\right\|_2 = \left\|(I - AA^+)(A - (B - A))B^+\right\|_2 \\
&= \left\|(A + (B - A))B^+ - (AA^+A + AA^+B - AA^+A)B^+\right\|_2 \\
&= \left\|((B - A) - AA^+B + A)B^+\right\|_2 = \left\|(I(B - A) - AA^+(B - A))B^+\right\|_2 \\
&= \left\|(I - AA^+)(B - A)B^+\right\|_2 \leq \|A - B\|_2\|B^+\|_2,
\end{aligned}$$

gdje smo 6. jednakosti iskoristili svojstvo pseudo-inverza $AA^+A = A$ te u zadnjoj nejednakosti činjenicu $\|I - AA^+\|_2 \leq 1$, što se i dokaže tijekom dokaza 2.4.1. Na isti način se dobije $\|P_A(I - P_B)\|_2 \leq \|A - B\|_2\|A^+\|_2$. Rezultat slijedi iz netrivijalne jednakosti $\|P_A(I - P_B)\|_2 = \|P_B(I - P_A)\|_2$ koja je dokazana u [5, Thm 2.3]. □

Sljedeći rezultat je rezultat iz analize greške trokutastih sustava koji će nam pomoći u analizi greške problema najmanjih kvadrata riješenih pomoću QR faktorizacije.

Teorem 1.3.14. *[4, p. 142] Neka je $Tx = b$ trokutasti sustav, gdje je $T \in \mathbb{R}^{n\times n}$ regularna matrica. Neka je sustav riješen substitucijom s bilo kojim redoslijedom. Tada izračunato rješenje $\hat{x}$ zadovoljava*

$$(T + \Delta T)\hat{x} = b, \quad |\Delta T| \leq \gamma_n |T|.$$

Poglavlje 2

Analiza greške

2.1 Analiza greške Householderove transformacije

U ovoj sekciji, bavit ćemo se stupčanom analizom greške Householderove transformacije. Stupčane granice daju dodatne informacije koje su nam bitne u određenim primjenama.

Lema 2.1.1. *Neka je $x \in \mathbb{R}^n$. Razmotrimo dvije konstrukcije $\beta \in \mathbb{R}$ i $v \in \mathbb{R}^n$ takve da je $Px = \sigma e_1$, gdje $P = I - \beta vv^T$ je Householderova transformacija s $\beta = 2/(v^T v)$:*

$$\begin{aligned} &sign(\sigma) = -sign(x_1) && sign(\sigma) = sign(x_1) \\ &v = x && v = x \\ &s = sign(x_1)\|x\|_2 && s = sign(x_1)\|x\|_2 \\ &v_1 = v_1 + s && v_1 = \frac{v_2^2 + \dots v_n^2}{v_1 + s} \\ &\beta = \frac{1}{sv_1} && \beta = \frac{1}{sv_1} \end{aligned}$$

U realnoj aritmetici računala, izračunati $\hat{\beta}$ i $\hat{v}$ iz obje konstrukcije zadovoljavaju $\hat{v}(2:n) = v(2:n)$ i:

$$\hat{\beta} = \beta\left(1 + \tilde{\theta}_n\right), \quad \hat{v}_1 = v_1\left(1 + \tilde{\theta}_n\right),$$

gdje je $|\tilde{\theta}_n| \leq \tilde{\gamma}_n$

Dokaz. Svako pojavljivanje δ označava različiti broj ograničen s $|\delta| \leq u$. Iz (1.7) možemo zapisati: $fl(x^T x) = (1+\theta_n)x^T x$. Računamo: $fl(\|x\|_2) = fl(\sqrt{x^T x}) = (1+\delta)(1+\theta_n)^{1/2}(x^T x)^{1/2}$. Po Lagrangeovom teoremu srednje vrijednosti vrijedi:

$$\sqrt{1+\theta_n} = 1 + c\theta_n = 1 + \tilde{\theta}_n, \quad c \lesssim \frac{1}{2}.$$

Vrijedi da je $|\tilde{\theta}_n| \leq |\theta_n| \leq \gamma_n$ pa je $fl(\|x\|_2) = (1 + \tilde{\theta}_{n+1})\|x\|_2$, gdje vrijedi:

$$|\tilde{\theta}_{n+1}| = |\tilde{\theta}_n + \delta + \tilde{\theta}_n\delta| \leq \gamma_n + u + \gamma_n u = \frac{nu + u - nu^2 + nu^2}{1 - nu} = \frac{(n+1)u}{1 - nu} \leq \gamma_{n+1}$$

Stoga je $\hat{s} = (1 + \tilde{\theta}_{n+1})s$.

Radi jednostavnosti zapisa, definirajmo $w = v_1 + s$. Tada je:

$$\hat{w} = (v_1 + \hat{s})(1 + \delta) = v_1 + s + (v_1 + s)\delta + \tilde{\theta}_{n+1}s(1 + \delta) = w(1 + \tilde{\theta}_{n+2}),$$

gdje smo definirali $\tilde{\theta}_{n+2} = \delta + \frac{s}{v_1+s}\tilde{\theta}_{n+1}(1 + \delta)$. Vrijedi:

$$|\tilde{\theta}_{n+2}| \leq u + \gamma_{n+1}(1 + u) = \frac{u - (n+1)u^2 + (n+1)u + (n+1)u^2}{1 - (n+1)u} = \frac{(n+2)u}{1 - (n+1)u} \leq \gamma_{n+2},$$

gdje smo koristili činjenicu da je $\left|\frac{s}{v_1+s}\right| \leq 1$ jer su v_1 i s istog predznaka. Stoga je:

$$\begin{aligned}\hat{\beta} = fl\left(\frac{1}{\hat{s}\hat{w}}\right) &= \frac{(1 + \delta_1)}{(1 + \delta_2)(1 + \tilde{\theta}_{n+1})s(1 + \tilde{\theta}_{n+2})w} \\ &= \frac{1 + \theta_2}{(1 + \tilde{\theta}_{2n+3})sw} = (1 + \tilde{\theta}_{4n+8})\beta\end{aligned}$$

Drugi algoritam ide sličnim računom. □

Radi jednostavnosti ćemo preformulirati Householderove matrice. Householderova matrica će biti matrica forme $I - vv^T$, što zahtjeva $\|v\|_2 = \sqrt{2}$. To se lako može postići ako za matricu $P = I - \beta vv^T$, definiramo $v_1 = \sqrt{\beta}v$ pa imamo da je $P = I - v_1v_1^T$. Tada rezultat leme možemo zapisati malo drugačije:

$$\hat{v} = v + \Delta v, \quad |\Delta v| \leq \tilde{\gamma}_m|v| \ \ (v \in \mathbb{R}^m, \ \|v\|_2 = \sqrt{2}). \tag{2.1}$$

Sljedeći rezultat se bavi analizom greške primjene Householderove matrice na vektor. Zanimat će nas P kakav je definiran u Lemi 2.1.1, ali ćemo dopustiti da je P proizvoljan. Tada je i v proizvoljan, ali ćemo pretpostavljati da izračunati $\hat{v}$ zadovoljava (2.1).

Lema 2.1.2. *Neka je $b \in \mathbb{R}^m$. Tada je $y = \hat{P}b = (I - \hat{v}\hat{v}^T)b = b - \hat{v}(\hat{v}^Tb)$, gdje $\hat{v} \in \mathbb{R}^m$ zadovoljava* (2.1)*. Izračunati $\hat{y}$ zadovoljava:*

$$\hat{y} = (P + \Delta P)b, \quad \|\Delta P\|_F \leq \tilde{\gamma}_m$$

Dokaz. Za $a \in \mathbb{R}$ i $x \in \mathbb{R}^m$ vrijedi:

$$fl(ax) = \begin{bmatrix} ax_1(1+\delta_1) \\ \vdots \\ ax_m(1+\delta_m) \end{bmatrix} = ax + a\Delta, \quad |\delta_i| \leq u,\ i = 1, \ldots, m, \tag{2.2}$$

gdje smo definirali

$$\Delta := \begin{bmatrix} \delta_1 x_1 \\ \vdots \\ \delta_m x_m \end{bmatrix}.$$

Tada je $|\Delta_i| \leq u|x_i|$, $i = 1, \ldots, m$, iz čega slijedi $|\Delta| \leq u|x|$. Definiramo $w := \hat{v}(\hat{v}^T b)$. Tada je

$$\hat{w} = fl(\hat{v}(\hat{v}^T b)) = (\hat{v} + \Delta\hat{v})(\hat{v}^T(b + \Delta b)), \quad |\Delta\hat{v}| \leq u|\hat{v}|, \quad |\Delta b| \leq \gamma_m |b|,$$

gdje smo koristili (2.2) i (1.7). Uvrštavanjem $\hat{v} = v + \Delta v$, dobivamo:

$$\hat{w} = (v + \Delta v + \Delta\hat{v})(v + \Delta v)^T(b + \Delta b).$$

Množenjem dobijamo $\hat{w} = v(v^T b) + \Delta w$, gdje je $\Delta w = vv^T\Delta b + v(\Delta v)^T(b + \Delta b) + (\Delta v + \Delta\hat{v})(v + \Delta v)^T(b + \Delta b)$. Kako je

$$|\Delta\hat{v}| \leq u|\hat{v}| = u|v + \Delta v| \leq u(|v| + |\Delta v|) \leq u|v| + u\tilde{\gamma}_m|v|,$$

vrijedi:

$$\begin{aligned} |\Delta w| &\leq |v||v^T||\Delta b| + |v||(\Delta v)^T|(|b| + |\Delta b|) + (|\Delta v| + |\Delta\hat{v}|)(|v^T| + (\Delta v)^T)(|b| + |\Delta b|) \\ &\leq \gamma_m|v||v^T||b| + \tilde{\gamma}_m|v||v^T|(|b| + \gamma_m|b|) + (\tilde{\gamma}_m|v| + u|v| + u\tilde{\gamma}_m|v|)(|v^T| + \tilde{\gamma}_m|v^T|)(|b| + \gamma_m|b|) \\ &\leq \tilde{\gamma}_m|v||v^T||b| \end{aligned}$$

gdje smo koristili aritmetiku notacije $\tilde{\gamma}_m$. Računamo:

$$\hat{y} = fl(b - \hat{w}) = \begin{bmatrix} (b - \hat{w})_1(1+\delta_1) \\ \vdots \\ (b - \hat{w})_m(1+\delta_m) \end{bmatrix} = b - \hat{w} + \Delta y_1 = b - v(v^T b) - \Delta w + \Delta y_1,$$

gdje smo definirali

$$\Delta y_1 = \begin{bmatrix} (b - \hat{w})_1\delta_1 \\ \vdots \\ (b - \hat{w})_m\delta_m \end{bmatrix}, \quad |\delta_i| \leq u,\ i = 1, \ldots, m.$$

Vrijedi $|\Delta y_1| \leq u|b - \hat{w}|$. Definiramo $\Delta y := \Delta y_1 - \Delta w$ te je tada $\hat{y} = Pb + \Delta y$. Vrijedi:

$$\begin{aligned}
|\Delta y| &\leq |\Delta w| + |\Delta y_1| \\
&\leq \tilde{\gamma}_m |v||v^T||b| + u|b| + u|\hat{w}| \\
&\leq \tilde{\gamma}_m |v||v^T||b| + u|b| + u(|v||v^T||b| + |\Delta w|) \\
&\leq \tilde{\gamma}_m |v||v^T||b| + u|b| + u(|v||v^T||b| + \tilde{\gamma}_m |v||v^T||b|) \\
&\leq (\tilde{\gamma}_m + u + u\tilde{\gamma}_m)|v||v^T||b| + u|b| \\
&\leq \tilde{\gamma}_m |v||v^T||b| + u|b|
\end{aligned}$$

Koristeći to dobivamo:

$$\begin{aligned}
||\Delta y||_2 &\leq \tilde{\gamma}_m |v^T||b| \sqrt{\sum_{i=1}^{m} |v_i|^2} + \tilde{\gamma}_m \sqrt{\sum_{i=1}^{m} |b_i|^2} \\
&\leq \tilde{\gamma}_m ||v^T||_2 ||b||_2 ||v||_2 + \tilde{\gamma}_m ||b||_2 \\
&\leq 2\tilde{\gamma}_m ||b||_2 + \tilde{\gamma}_m ||b||_2 \\
&= \tilde{\gamma}_m ||b||_2
\end{aligned}$$

gdje smo u prvoj nejednakosti koristili nejednakost trokuta, a u drugoj nejednakosti iskoristili Cauchy-Schwarz nejednakost. Ako defiramo $\Delta P := \frac{\Delta y b^T}{b^T b}$, tada je $\hat{y} = (P + \Delta P)b$, te vrijedi:

$$||\Delta P||_F = \frac{1}{||b||_2^2} \left(\sum_{i=1}^{m} \sum_{j=1}^{m} \left(\Delta y_i b_j \right)^2 \right)^{\frac{1}{2}} = \frac{1}{||b||_2^2} \left(\sum_{i=1}^{m} (\Delta y_i)^2 \sum_{j=1}^{m} (b_j)^2 \right)^{\frac{1}{2}} = \frac{1}{||b||_2^2} ||\Delta y||_2 ||b||_2 = \frac{||\Delta y||_2}{||b||_2},$$

što je onda po zaključku koji smo ranije dobili $||\Delta P|| \leq \tilde{\gamma}_m$. □

Kako smo objasnili, u metodi Householder QR faktorizacije primjenjujemo niz Householderovih matrica na početnu matricu A. Sljedeća lema je upravo analiza greške primjenjivanja niza Householderovih matrica na proizvoljnu matricu $A \in \mathbb{R}^{m \times n}$. Kako množimo Householderovim matricama s lijeva matricu A, primjenjujemo Householderove matrice na stupce matrice A. Tu ćemo iskoristiti rezultat Leme 2.1.2 te ćemo dobiti stupčane ocjene.

Pretpostavit ćemo da vrijedi:

$$r\tilde{\gamma}_m < \frac{1}{2},$$

gdje je r broj Householderovih matrica koje primjenjujemo na matricu A.

Lema 2.1.3. *Promotrimo niz transformacija:*

$$A_{k+1} = P_k A_k, \quad k = 1, \ldots, r,$$

gdje je $A_1 = A \in \mathbb{R}^{m\times n}$ *te* $P_k = I - v_k v_k^T \in \mathbb{R}^{m\times m}$ *Householderova matrica. Pretpostavimo da transformacije izvodimo koristeći izračunate Householderove vektore* $\hat{v}_k \approx v_k$ *koji zadovoljavaju* (2.1)*. Izračunata matrica* $\hat{A}_{r+1}$ *zadovoljava:*

$$\hat{A}_{r+1} = Q^T(A + \Delta A),$$

gdje je $Q^T = P_r P_{r-1} \dots P_1$ *i*

$$\|\Delta a_j\|_2 \leq r\tilde{\gamma}_m \|a_j\|_2, \quad j = 1, \dots, n,$$

gdje su a_j *stupci od* A *i* Δa_j *stupci od* ΔA*.*
U specijalnom slučaju $n = 1$*, gdje je* $A \equiv a$*, imamo*

$$\hat{a}^{(r+1)} = (Q + \Delta Q)^T a,$$

gdje vrijedi $\|\Delta Q\|_F \leq r\tilde{\gamma}_m$*.*

Dokaz. Iz niza transformacija koje radimo vidimo da će j-ti stupac od A poprimiti sljedeći oblik:

$$a_j^{(r+1)} = P_r \dots P_1 a_j.$$

Po Lemi 2.1.2 imamo da je:

$$\hat{a}_j^{(r+1)} = (P_r + \Delta P_r) \dots (P_1 + \Delta P_1) a_j,$$

gdje svaki ΔP_k ovisi o j i zadovoljava $\|\Delta P_k\|_F \leq \tilde{\gamma}_m$. To se može zapisati i na sljedeći način:

$$\hat{a}_j^{(r+1)} = Q^T(a_j + \Delta a_j)$$

gdje smo Δa_j definirali kao umnožak matrice Q i $\prod_{k=1}^{r}(P_k + \Delta P_k) - Q^T$ te vektora a_j. Vrijedi:

$$\begin{aligned}
\|\Delta a_j\|_2 = \|Q^T \Delta a_j\|_2 = \|\hat{a}_j^{(r+1)} - Q^T a_j\|_2 &= \left\| \left(\prod_{k=1}^{r}(P_k + \Delta P_k) - \prod_{k=1}^{r} P_k \right) a_j \right\|_2 \\
&\leq \left\| \prod_{k=1}^{r}(P_k + \Delta P_k) - \prod_{k=1}^{r} P_k \right\|_2 \|a_j\|_2 \leq ((1 + \tilde{\gamma}_m)^r - 1)\|a_j\|_2 \\
&\leq \frac{r\tilde{\gamma}_m}{1 - r\tilde{\gamma}_m} \|a_j\|_2 \leq r\tilde{\gamma}'_m \|a_j\|_2.
\end{aligned}$$

gdje smo u 2. nejednakosti koristili Lemu 1.3.5, a u 3. nejednakosti Lemu 1.3.3 te pretpostavku $r\tilde{\gamma}_m < \frac{1}{2}$. Lemu 1.3.5 smo mogli koristiti zbog činjenice da je matrična 2-norma

konzistentna te jer vrijedi $\|\Delta P_k\|_2 \leq \|\Delta P_k\|_F \leq \tilde{\gamma}_m \|P_k\|_2$ jer je P_k ortogonalna matrica, tj. $\|P_k\|_2 = 1$.

Na kraju, ako je $n = 1$, onda imamo da je matrica A zapravo vektor stupac te je $\hat{a}^{(r+1)} = Q^T(a + \Delta a)$. Kao u dokazu Leme 2.1.2, to možemo preoblikovati u $\hat{a}^{(r+1)} = (Q + \Delta Q)^T a$, gdje je $\Delta Q^T = (Q^T \Delta a)a^T/(a^T a)$. Vrijedi:

$$\|\Delta Q\|_F = \|\Delta Q^T\|_F = \frac{1}{\|a\|_2^2}\|Q^T(\Delta a a^T)\|_F = \frac{1}{\|a\|_2^2}\|\Delta a a^T\|_F = \frac{1}{\|a\|_2}\|\Delta a\|_2 \leq r\tilde{\gamma}_m,$$

gdje smo u zadnjoj nejednakosti koristili rezultat dobiven u 1. dijelu. □

Pomoću Leme 2.1.3 dobijamo glavni rezultat za Householder QR faktorizaciju.

Teorem 2.1.4. *Neka je $\hat{R} \in \mathbb{R}^{m\times n}$ izračunata gornje trapezasta matrica, koja je QR faktor od $A \in \mathbb{R}^{m\times n} (m \geq n)$, dobivena iz Householder QR algoritma (s bilo kojim izborom predznaka u algoritmu). Tada postoji ortogonalna matrica $Q \in \mathbb{R}^{m\times m}$ takva da vrijedi:*

$$A + \Delta A = Q\hat{R},$$

gdje vrijedi:

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2, \quad j = 1, \ldots, n.$$

Matrica Q je dana eksplicitno formulom $Q = (P_n P_{n-1} \ldots P_1)^T$, gdje su matrice P_k Householderove matrice koje odgovaraju egzaktnoj primijeni k-tog koraka algoritma na matricu $\hat{A}_k$.

Dokaz. Radimo direktnu primjenu Leme 2.1.3. Radi se $r = n$ transformacija te će $\hat{R} = \hat{A}_{n+1}$ i vrijedit će $\hat{R} = Q^T(A + \Delta A)$, tj. $A + \Delta A = Q\hat{R}$. Ono što preostaje je definirati P_k transformacije koje su u Lemi 2.1.3 bile proizvoljne. P_k će biti definirane kao Householderove matrice koje u produktu u k-tom koraku s matricom $\hat{A}_k$ daju 0 u k-tom stupcu ispod dijagonale. Jedna suptilnost je da se ne računa eksplicitno donji trokut od $\hat{R}$, nego ga postavljamo na 0. Važno je za primijetiti u Lemama 2.1.2 i 2.1.3 sve još uvijek vrijedi. Glavni razlog tome leži u Lemi 2.1.2. Ako gledamo $\hat{y} = (P + \Delta P)b$, mi u $\hat{y}$ postavljamo nakon nekog elementa, recimo ispod k-tog elementa vektora, sve 0 (govorimo o Lemi 2.1.2 u smislu procesa QR faktorizacije), teoretski Pb nakon tog k-tog elementa ima sve 0, pa onda prisilno nakon k-tog elementa vektora će i ΔPb morati imati sve 0. To znači da će u ΔP svi redci ispod k-tog morati biti 0. To naime neće utjecati na ocjenu $\|\Delta P\|_F \leq \tilde{\gamma}_m$ jer smo na ovaj način zapravo smanjili $\|\Delta P\|_F$ jer je Frobeniusova norma korijen zbroja kvadriranih elemenata matrice. □

Faktor $\tilde{\gamma}_{mn}$ u ocjeni se može za svaki stupac smanjiti na $\tilde{\gamma}_{mj}$ za j-ti stupac. Naime, kako je $\hat{a}_j^{(n+1)} = (P_n + \Delta P_n)(P_{n-1} + \Delta P_{n-1}) \ldots (P_1 + \Delta P_1)a_j$, vrijedit će da su $\Delta P_k = 0$ za

$k > j$. Nakon j-te transformacije, na j-ti stupac će se djelovati s jediničnom matricom (kao što je u algoritmu opisano) pa ne nastaju numeričke greške, tj. $\Delta P_k = 0$ za $k > j$. Tada računom istim kao u Lemi 2.1.3 dolazimo do faktora $\tilde{\gamma}_{mj}$. Druga stvar koja se iz teorema da primijetiti je da iz stupčanih ocjena $\|\Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2$ preko Leme 1.3.2 dolazimo do slabije ocjene $\|\Delta A\|_F \leq \tilde{\gamma}_{mn}\|A\|_F$. Teorem 2.1.4 se može izraziti i pomoću te ocjene. Mana te ocjene je to što je puno slabija za matrice čiji stupci jako variraju u normi.

Primijetimo da je u Teoremu 2.1.4 matrica Q bila teoretski egzaktna, tj. nismo je računali. Upravo ta činjenica da je Q bila egzaktno ortogonalna čini ovaj rezultat toliko korisnim. Matrica je zadana formulom $Q = (P_n P_{n-1} \dots P_1)^T$. Zanimat će nas može li se u Teoremu 2.1.4 teoretska matrica Q zamijeniti s izračunatom matricom $\hat{Q}$ te kako će to utjecati na ocjenu.

Pretpostavimo da je $Q = P_1 P_2 \dots P_n$. Računat ćemo je s desna na lijevo. Po Lemi 2.1.3 dobivamo (uz $A_1 = I_m$):

$$\hat{Q} = Q(I_m + \Delta I), \quad \|\Delta I(:, j)\| \leq \tilde{\gamma}_{mn}\|I_m(:, j)\|_2 = \tilde{\gamma}_{mn}, \quad j = 1, 2, \dots, m,$$

gdje $\Delta I(:, j)$ i $I_m(:, j)$ označavaju j-te stupce matrica ΔI, I_m (Matlab notacija). Tada je:

$$\|\hat{Q} - Q\|_F = \|Q\Delta I\|_F = \|\Delta I\|_F = \sqrt{\sum_{j=1}^{m} \|\Delta I(:, j)\|_2^2} \leq \tilde{\gamma}_{mn} \sqrt{\sum_{j=1}^{m} 1} = \sqrt{m}\tilde{\gamma}_{mn}.$$

Ovom ocjenom zapravo pokazujemo da je matrica $\hat{Q}$ vrlo blizu ortogonalnoj matrici. Treba još vidjeti kako će zamjena matrice Q s $\hat{Q}$ utjecati na stupce greške:

$$\|\Delta A_2(:, j)\|_2 = \|(A - \hat{Q}\hat{R})(:, j)\|_2,$$

gdje ćemo s ΔA_2 označiti matricu greške kad zamijenimo Q s $\hat{Q}$, odnosno $A + \Delta A_2 = \hat{Q}\hat{R}$. Vrijedi:

$$\begin{aligned} \|(A - \hat{Q}\hat{R})(:, j)\|_2 &= \|(A - Q\hat{R} + Q\hat{R} - \hat{Q}\hat{R})(:, j)\|_2 \\ &\leq \|(A - Q\hat{R})(:, j)\|_2 + \|((Q - \hat{Q})\hat{R})(:, j)\|_2 \\ &\leq \tilde{\gamma}_{mn}\|a_j\|_2 + \sqrt{m}\tilde{\gamma}_{mn}\|\hat{R}(:, j)\|_2 \end{aligned} \tag{2.3}$$

Ostaje nam ocijeniti $\|\hat{R}(:, j)\|_2$. Vrijedi:

$$\|\hat{R}(:, j)\|_2 = \|(Q^T(A + \Delta A))(:, j)\|_2 = \|(A + \Delta A)(:, j)\|_2 \leq \|a_j\|_2 + \|\Delta a_j\|_2 \leq (1 + \tilde{\gamma}_{mn})\|a_j\|_2,$$

gdje smo koristili unitarnu invarijatnost Euklidske norme te Teorem 2.1.4. Vrijedi da je $\tilde{\gamma}_{mn}^2 \leq \tilde{\gamma}_{mn}$ pa nastavljajući (2.3) dobijamo:

$$\begin{aligned} \|(A - \hat{Q}\hat{R})(:, j)\|_2 &\leq \tilde{\gamma}_{mn}\|a_j\|_2 + \sqrt{m}\tilde{\gamma}_{mn}(1 + \tilde{\gamma}_{mn})\|a_j\|_2 \\ &\leq \sqrt{m}\tilde{\gamma}_{mn}\|a_j\|_2, \end{aligned} \tag{2.4}$$

gdje smo u zadnjoj nejednakosti zamijenili konstantu c s $3c$ u $\tilde{\gamma}_{mn}$. Pokazali smo da $A + \Delta A_2 = \hat{Q}\hat{R}$, gdje je $\hat{Q}$ približno ortogonalna te da imamo ocjenu za grešku unatrag po stupcima koja se razlikuje za samo konstantni faktor od onih u Teoremu 2.1.4.

2.2 Analiza greške Givensovih rotacija

Analiza greške Givensovih rotacija je slična analizi Householderovih transformacija, ali ipak malo lakša. Kao i u slučaju analize greške Householderovih transformacija, prvo promatramo samo računanje Givensovih rotacija.

Lema 2.2.1. *Neka je $G(i, j, \theta)$ Givensova rotacija koju konstruiramo s* (1.1)*. Izračunati $\hat{c}$ i $\hat{s}$ zadovoljavaju:*

$$\hat{c} = c(1 + \theta_4), \quad \hat{s} = s(1 + \theta_4'), \tag{2.5}$$

gdje $|\theta_4|, |\theta_4'| \leq \gamma_4$.

Dokaz. Vrijedi:

$$\hat{c} = fl\left(\frac{x_i}{\sqrt{x_i^2 + x_j^2}}\right) = \frac{x_i(1 + \delta_1)}{(1 + \delta_5)\sqrt{(x_i^2(1 + \delta_2) + x_j^2(1 + \delta_3))(1 + \delta_4)}} \tag{2.6}$$

uz $|\delta_i| \leq u,\ i = 1, 2, 3, 4, 5$. Definiramo:

$$\delta := \frac{x_i^2\delta_2 + x_j^2\delta_3}{x_i^2 + x_j^2}.$$

te vrijedi:

$$|\delta| = \left|\frac{x_i^2\delta_2 + x_j^2\delta_3}{x_i^2 + x_j^2}\right| \leq \frac{x_i^2 u + x_j^2 u}{x_i^2 + x_j^2} = u.$$

Tada je $x_i^2\delta_2 + x_j^2\delta_3 = (x_i^2 + x_j^2)\delta$, tj. tada (2.6) postaje

$$\hat{c} = \frac{x_i(1 + \delta_1)}{(1 + \delta_5)\sqrt{(1 + \delta_4)}\sqrt{(x_i^2 + x_j^2)(1 + \delta)}}.$$

Želimo pokazati da je $\sqrt{1 + \delta} = 1 + \delta'$ i $\sqrt{1 + \delta_4} = 1 + \delta''$, gdje su $|\delta'|, |\delta''| \leq u$. Koristeći Taylorov teorem srednje vrijednosti za funkciju $f(x) = \sqrt{1 + x}$ u točki 0, dobijamo:

$$\sqrt{1 + x} = 1 + \frac{1}{2}x + O(x^2).$$

Tada je $\sqrt{1+\delta} = 1 + \frac{1}{2}\delta + O(\delta^2) = 1 + \delta'$, gdje je $\delta' = \frac{1}{2}\delta + O(\delta^2)$. Vrijedi:

$$|\delta'| \leq \frac{1}{2}u + O(u^2) \leq u,$$

jer je u dovoljno mali broj. Analogno se pokaže da je $\sqrt{1+\delta_4} = 1 + \delta''$, gdje je $\delta'' = \frac{1}{2}\delta_4 + O(\delta_4^2)$ te da vrijedi $|\delta''| \leq u$. Konačno:

$$\hat{c} = \frac{x_i(1+\delta_1)}{(1+\delta')(1+\delta'')(1+\delta_5)\sqrt{x_i^2+x_j^2}} = c(1+\theta_4)$$

po Lemi 1.3.3, gdje je $\theta_4 \leq \gamma_4$. Kako je $s = \frac{x_j}{\sqrt{x_i^2+x_j^2}}$, što je u smislu numeričkog računanja, isti izraz kao i c, vrijedit će isti račun, odnosno:

$$\hat{s} = s(1+\theta_4'),$$

gdje je $|\theta_4'| \leq \gamma_4$. □

Kao i u analizi greške Householderovih reflektora, slijedi nam primjena Givensovih rotacija na vektor.

Lema 2.2.2. *Neka je $x \in \mathbb{R}^m$. Promotrimo računanje $y = \hat{G}_{ij}x$, gdje je $\hat{G}_{ij}$ izračunata Givensova rotacija u (i, j) ravnini za koju $\hat{c}$ i $\hat{s}$ zadovoljavaju* (2.5)*. Izračunati $\hat{y}$ zadovoljava:*

$$\hat{y} = (G_{ij} + \Delta G_{ij})x, \quad \|\Delta G_{ij}\|_F \leq \sqrt{2}\gamma_6,$$

gdje je G_{ij} egzaktna Givensova rotacija bazirana na c i s iz (2.5)*. Svi redci od ΔG_{ij} osim i-tog i j-tog su* 0*.*

Dokaz. Kao što smo i u uvodu napisali, $\hat{y}$ se razlikuje od x samo na i-tom i j-tom elementu. Pa imamo:

$$\hat{y}_i = fl(\hat{c}x_i + \hat{s}x_j) = \Big(c(1+\theta_4)x_i(1+\delta_1) + sx_j(1+\theta_4')(1+\delta_2)\Big)(1+\delta_3) = cx_i(1+\theta_6) + sx_j(1+\theta_6'),$$

gdje $|\theta_6|, |\theta_6'| \leq \gamma_6$. Kako imamo istu situaciju u $\hat{y}_j$, analogno zaključujemo da je $\hat{y}_j = cx_j(1+\theta_6'') - sx_i(1+\theta_6'''), |\theta_6''|, |\theta_6'''| \leq \gamma_6$. Kako je $(G_{ij}x)_k = x_k$ za $k \neq i, j$ te $(G_{ij}x)_i = cx_i + sx_j$, $(G_{ij}x)_j = cx_j - sx_i$ onda možemo definirati matricu ΔG_{ij} kojoj su svi elementi 0 osim:

$$\Delta G_{ij}([i,j],[i,j]) = \begin{bmatrix} c\theta_6 & s\theta_6' \\ -s\theta_6''' & c\theta_6'' \end{bmatrix}.$$

Tada je:

$$\hat{y} - G_{ij}x = \Delta G_{ij}x,$$

tj. $\hat{y} = (G_{ij} + \Delta G_{ij})x$ te vrijedi:

$$\|\Delta G_{ij}\|_F = \sqrt{c^2\theta_6^2 + s^2\theta_6'^2 + s^2\theta_6''^2 + c^2\theta_6''^2} \leq \sqrt{2c^2\gamma_6^2 + 2s^2\gamma_6^2} = \gamma_6\sqrt{2(c^2+s^2)} = \sqrt{2}\gamma_6.$$

□

Nastavljamo sa rezultatom primjene niza Givensovih rotacija na matricu. Razlika u odnosu na analizu Householderovih transformacija je to što ćemo ovdje radi lakše analize, ipak primjenjivati grupe disjunktnih Givensovih rotacija. Kao što smo u uvodu naveli, algoritam sa disjunktnim Givensovim rotacijama je numerički ekvivalentan algoritmu sa standardnim izborom i rasporedom rotacija.

Lema 2.2.3. *Promatrajmo niz transformacija:*

$$A_{k+1} = W_k A_k, \quad k = 1, \ldots, r,$$

gdje je $A_1 = A \in \mathbb{R}^{m\times n}$ i svaki W_k je produkt disjunktnih Givensovih rotacija. Pretpostavimo da Givensove rotacije koje definiraju matrice W_k koriste izračunate vrijednosti sinusa i kosinusa kao u (2.5)*. Tada izračunata matrica $\hat{A}_{r+1}$ zadovoljava:*

$$\hat{A}_{r+1} = Q^T(A + \Delta A),$$

gdje je $Q^T = W_r W_{r-1} \ldots W_1$ i

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_r \|a_j\|_2, \quad j = 1, \ldots, n.$$

U specijalnom slučaju $n = 1$, td. je $A = a$, imamo $\hat{a}^{(r+1)} = (Q + \Delta Q)^T a$ s $\|\Delta Q\|_F \leq \tilde{\gamma}_r$.

Dokaz. Promatrajmo j-ti stupac od A, a_j, koji prolazi kroz r transformacija. Stupac nakon r transformacija izgleda $a_j^{(r+1)} = W_r \ldots W_1 a_j$. Kako je $W_1 = G_{i_1 j_1} \ldots G_{i_l j_l}$, uzastopnom primjenom Leme 2.2.2 dobijamo da je $\hat{y} = fl(W_1 x) = (G_{i_1 j_1} + \Delta G_{i_1 j_1}) \ldots (G_{i_l j_l} + \Delta G_{i_l j_l})x = (W_1 + \Delta W_1)x$. Analogno sada dobijamo da je $\hat{a}_j^{(r+1)} = (W_r + \Delta W_r) \ldots (W_1 + \Delta W_1)a_j$, gdje po Lemi 2.2.2 svaki ΔW_k ovisi o j. Tvrdimo da vrijedi $\|\Delta W_k\|_2 \leq \sqrt{2}\gamma_6$. Neka je $W_k = G_{i_1 j_1} \ldots G_{i_t j_t}$, gdje je $t \leq n$. Kako su te rotacije disjunktne, vrijedit će:

$$(W_k x)_l = \begin{cases} x_l, & l \neq i_1, \ldots, i_t, j_1, \ldots, j_t, \\ cx_{i_{l'}} + sx_{j_{l'}}, & l = i_{l'},\ l' \in \{1, \ldots, t\}, \\ cx_{j_{l'}} - sx_{i_{l'}}, & l = j_{l'},\ l' \in \{1, \ldots, t\}. \end{cases}$$

Također, zbog toga što su rotacije disjunktne, znamo kako će $\hat{y}$ izgledati jer će svaka matrica $G_{i_l j_l} + \Delta G_{i_l j_l}$ djelovati na različite x_{i_l} i x_{j_l}. Po Lemi 2.2.2 vrijedi:

$$\hat{y}_l = \begin{cases} x_l, & l \neq i_1, \ldots, i_t, j_1, \ldots, j_t, \\ cx_{i_{l'}}(1+\theta_6) + sx_{j_{l'}}(1+\theta_6'), & l = i_{l'},\ l' \in \{1, \ldots, t\}, \\ cx_{j_{l'}}(1+\theta_6) - sx_{i_{l'}}(1+\theta_6'), & l = j_{l'},\ l' \in \{1, \ldots, t\}. \end{cases}$$

Sad možemo izračunati:

$$\begin{aligned}\|\hat{y} - W_k x\|_2 &= \sqrt{\sum_{l'=1}^{t} (\hat{y} - W_k x)_{i_{l'}}^2 + \sum_{l'=1}^{t} (\hat{y} - W_k x)_{j_{l'}}^2} \\ &= \sqrt{\sum_{l'=1}^{t} \left((\hat{y} - W_k x)_{i_{l'}}^2 + (\hat{y} - W_k x)_{j_{l'}}^2 \right)} = \sqrt{\sum_{l'=1}^{t} \|\Delta G_{i_{l'} j_{l'}} x\|_2^2}, \end{aligned} \tag{2.7}$$

gdje je $\Delta G_{i_{l'} j_{l'}}$ definiran u Lemi 2.2.2 i gdje smo u zadnjoj jednakosti iskoristili činjenicu da je

$$\|\Delta G_{i_{l'} j_{l'}} x\|_2 = \sqrt{(c\theta_6 x_{i_{l'}} + s\theta'_6 x_{j_{l'}})^2 + (c\theta_6 x_{j_{l'}} - s\theta'_6 x_{i_{l'}})^2} = \sqrt{(\hat{y} - W_k x)_{i_{l'}}^2 + (\hat{y} - W_k x)_{j_{l'}}^2}.$$

Vrijedi:

$$\begin{aligned}|c\theta_6 x_{i_{l'}} + s\theta_6 x_{j_{l'}}| &= \left\| \begin{bmatrix} c\theta_6 \\ s\theta'_6 \end{bmatrix}^T \begin{bmatrix} x_{i_{l'}} \\ x_{j_{l'}} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} c\theta_6 \\ s\theta'_6 \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} x_{i_{l'}} \\ x_{j_{l'}} \end{bmatrix} \right\|_2 = \sqrt{c^2\theta_6^2 + s^2\theta_6'^2} \sqrt{x_{i_{l'}}^2 + x_{j_{l'}}^2} \\ &\leq \gamma_6 \sqrt{c^2 + s^2} \sqrt{x_{i_{l'}}^2 + x_{j_{l'}}^2} = \gamma_6 \sqrt{x_{i_{l'}}^2 + x_{j_{l'}}^2}, \end{aligned}$$

gdje smo u 1. nejednakosti koristili Cauchy-Schwarz nejednakost. Na isti način se dobije $|c\theta_6 x_{j_{l'}} - s\theta_6 x_{i_{l'}}| \leq \gamma_6 \sqrt{x_{i_{l'}}^2 + x_{j_{l'}}^2}$. Koristeći to dobijamo:

$$\|\Delta G_{i_{l'} j_{l'}} x\|_2 = \sqrt{(c\theta_6 x_{i_{l'}} + s\theta'_6 x_{j_{l'}})^2 + (c\theta_6 x_{j_{l'}} - s\theta'_6 x_{i_{l'}})^2} \leq \sqrt{2}\gamma_6 \sqrt{x_{i_{l'}}^2 + x_{j_{l'}}^2}.$$

Nastavljajući (2.7) dobijamo:

$$\|\hat{y} - W_k x\|_2 \leq \sqrt{2}\gamma_6 \sqrt{\sum_{l'=1}^{t} \left(x_{i_{l'}}^2 + x_{j_{l'}}^2 \right)} \leq \sqrt{2}\gamma_6 \|x\|_2.$$

Kako je:

$$\|\Delta W_k\|_2 = \max_{x \neq 0} \frac{\|\Delta W_k x\|_2}{\|x\|_2} = \max_{x \neq 0} \frac{\|\hat{y} - W_k x\|_2}{\|x\|_2} \leq \max_{x=0} \frac{\sqrt{2}\gamma_6 \|x\|_2}{\|x\|_2} = \sqrt{2}\gamma_6.$$

Možemo $\hat{a}_j^{(r+1)}$ zapisati kao $\hat{a}_j^{(r+1)} = Q^T(a_j + \Delta a_j)$, gdje je $Q^T = W_r \dots W_1$, a $\Delta a_j = Q(\prod_{i=1}^r (W_i + \Delta W_i) - \prod_{i=1}^r W_i)a_j$. Tada vrijedi:

$$\begin{aligned}
\|\Delta a_j\|_2 &= \left\| Q\left(\prod_{i=1}^r (W_i + \Delta W_i) - \prod_{i=1}^r W_i\right) a_j \right\|_2 \\
&= \left\| \left(\prod_{i=1}^r (W_i + \Delta W_i) - \prod_{i=1}^r W_i\right) a_j \right\|_2 \\
&\leq \left\| \prod_{i=1}^r (W_i + \Delta W_i) - \prod_{i=1}^r W_i \right\|_2 \|a_j\|_2. \qquad (2.8)
\end{aligned}$$

Po Lemi 1.3.5 kako je $\|\Delta W_k\|_2 \leq \sqrt{2}\gamma_6 \|W_k\|_2$ ($\|W_k\|_2 = 1$), nastavljanjem (2.8) vrijedi:

$$\|\Delta a_j\|_2 \leq \left(\prod_{i=1}^r (1 + \sqrt{2}\gamma_6) - 1\right) \|a_j\|_2 \leq ((1 + \sqrt{2}\gamma_6)^r - 1)\|a_j\|_2.$$

Kako vrijedi:

$$\begin{aligned}
(1 + \sqrt{2}\gamma_6)^r - 1 \leq e^{r\sqrt{2}\gamma_6} - 1 &= \sum_{n=0}^{\infty} \frac{(r\sqrt{2}\gamma_6)^n}{n!} - 1 = r\sqrt{2}\gamma_6 \sum_{n=0} \frac{(r\sqrt{2}\gamma_6)^n}{(n+1)!} \\
&\leq r\sqrt{2}\gamma_6 \sum_{n=0}^{\infty} \left(\frac{r\sqrt{2}\gamma_6}{2}\right)^n = r\sqrt{2}\gamma_6 \frac{1}{1 - \frac{r\sqrt{2}\gamma_6}{2}} \leq \tilde{\gamma}_r.
\end{aligned}$$

gdje smo u 2. nejednakosti koristili činjenicu da je $(n+1)! \geq 2^n$ te na kraju pretpostavku da je $r\sqrt{2}\gamma_6 < 2$ što je razumna pretpostavka s obzirom da je γ_6 jako mali broj. Tada dobivamo:

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_r \|a_j\|_2.$$

Dokaz za $n = 1$ je isti dokazu u Lemi 2.1.3 za $n = 1$. □

Nakon uvodnih lema, spremni smo dati rezultat za Givens QR faktorizaciju.

Teorem 2.2.4. *Neka je $\hat{R} \in \mathbb{R}^{m\times n}$ izračunat gornji trapezasti QR faktor od $A \in \mathbb{R}^{m\times n}(m \geq n)$ dobiven preko Givens QR algoritma, s bilo kojim standardnim izabirom i rasporedom rotacija. Tada postoji ortogonalna matrica $Q \in \mathbb{R}^{m\times n}$ takva da*

$$A + \Delta A = Q\hat{R}, \quad \|\Delta a_j\|_2 \leq \tilde{\gamma}_{m+n-2}\|a_j\|_2, \quad j = 1, \dots, n.$$

(Matrica Q je produkt Givensovih rotacija, gdje k-ta rotacija odgovara egzaktnoj rotaciji koja se primjenjuje na $\hat{A}_k$ u k-tom koraku algoritma.)

Dokaz. Kao što smo i u uvodu napravili, s bilo kojim odabirom i rasporedom rotacija, preuredit ćemo algoritam da bi imali grupe disjunktnih rotacija. Takav algoritam je numerički ekvivalentan algoritam pa ćemo tvrdnju dokazati za preuređen algoritam. Za matricu $m \times n$, gdje je $m > n$, Givens QR faktorizacija se može napisati kao $W_r \dots W_1 A = R$, gdje su W_i produkti najviše n rotacija, te je $r = m + n - 2$ (broj antidijagonala u donjem trokutu matrice). U slučaju $m = n$, $r = m + n - 3$, no $\gamma_{m+n-3} \leq \gamma_{m+n-2}$ pa je ocjena koju ćemo dobiti za $m \geq n$. Primjenom Leme 2.2.3 u kojoj je $r = m + n - 2$, dobijamo $A + \Delta A = Q\hat{R}$ i ocjene:

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_{m+n-2} \|a_j\|_2, \quad j = 1, \dots, n.$$

Kao što je to bio i slučaj u Teoremu 2.1.4 i ovdje ćemo umjesto eksplicitnog računanja donjeg trokuta $\hat{R}$ staviti 0. Sličnom argumentacijom kao i u Teoremu 2.1.4, to neće utjecati na ocjenu u Lemi 2.2.2 jer će postavljanje $\hat{y}_j = 0$ te teorijska vrijednost $(G_{ij}x)_j = 0$ natjerati da je i $(\Delta G_{ij}x)_j = 0$ pa to ne utječe na ocjenu $\|\Delta G_{ij}\|_F \leq \sqrt{2}\gamma_6$. □

2.3 Analiza greške Gram-Schmidtove ortogonalizacije

Gram-Schmidtove metode računaju Q eksplicitno, za razliku od Householderove i Givensove metode. Ovo je i prednost jer nema dodatnog posla za formiranje matrice Q, ali je i mana jer izračunati Q ne mora biti ortogonalan. Ortonormalnost matrice Q je posljedica ortogonalizacije vektora u metodama, no te metode ortogonalizacije vektora se mogu iskvariti greškama zaokruživanja. U analizi greške metoda ćemo vidjeti da izračunate matrice Q ne moraju nužno ispast ortogonalne.

Za uvod u analizu greške, pogledat ćemo analizu CGS metode za slučaj $n = 2$. Metode CGS i MGS su identične za $n = 2$. Već tu ćemo vidjeti kako će gubitak ortogonalnosti biti ograničen s uvjetovanošću matrice. Neka su nam dani $a_1, a_2 \in \mathbb{R}^m$. Pretpostavimo da se $q_1 = \frac{a_1}{\|a_1\|_2}$ računa egzaktno te s time formiramo nenormirani vektor $q_2 = a_2 - (q_1^T a_2)q_1$. Izračunati vektor će zadovoljavati:

$$\begin{aligned}\hat{q_2} = fl(a_2 - (q_1^T a_2)q_1) &= fl\left(\begin{bmatrix} a_{21} - (q_1^T a_2)q_{11} \\ \vdots \\ a_{2m} - (q_1^T a_2)q_{1m} \end{bmatrix}\right) \\ &= \begin{bmatrix} \Big(a_{21} - (q_1^T(a_2 + \Delta a_2))q_{11}(1 + \delta_{11})\Big)(1 + \delta_{21}) \\ \vdots \\ \Big(a_{2m} - (q_1^T(a_2 + \Delta a_2))q_{1m}(1 + \delta_{1m})\Big)(1 + \delta_{2m}) \end{bmatrix}\end{aligned}$$

gdje je $|\Delta a_2| \leq \gamma_m |a_2|$ te $|\delta_{1i}|, |\delta_{2i}| \leq u, i = 1, \ldots, m$. Definiramo:

$$\Delta\tilde{q}_2 = \begin{bmatrix} a_{21}\delta_{21} - q_1^T(a_2 + \Delta a_2)q_{11}(\delta_{11} + \delta_{21} + \delta_{11}\delta_{21}) \\ \vdots \\ a_{2m}\delta_{2m} - q_1^T(a_2 + \Delta a_2)q_{1m}(\delta_{1m} + \delta_{2m} + \delta_{1m}\delta_{2m}) \end{bmatrix}$$

Iz toga je:

$$\hat{q}_2 = a_2 - q_1^T(a_2 + \Delta a_2)q_1 + \Delta\tilde{q}_2$$

Ako gledamo Lemu 1.3.3, možemo zaključiti da iz $(1 + \theta_2) = (1 + \delta_1)(1 + \delta_2)$ vrijedi da je $\theta_2 = \delta_1 + \delta_2 + \delta_1\delta_2$. Tada je:

$$|\Delta\tilde{q}_2| \leq u|a_2| + \gamma_2 |q_1^T(a_2 + \Delta a_2)q_1|$$

Stoga:

$$\hat{q}_2 = q_2 + \Delta q_2, \quad \Delta q_2 = \Delta\tilde{q}_2 - (q_1^T \Delta a_2)q_1,$$

pa je:

$$\begin{aligned} |\Delta q_2| &\leq \gamma_m |q_1^T||a_2||q_1| + u|a_2| + \gamma_2 |q_1^T||a_2||q_1| + \gamma_2\gamma_m |q_1^T||a_2||q_1| \\ &= \gamma_m |q_1^T||a_2||q_1| + u|a_2| + \gamma_2(1 + \gamma_m)|q_1^T||a_2||q_1|. \end{aligned} \tag{2.9}$$

Sad kada smo našli granicu za grešku izračunatoga vektora, preostaje pogledati ortogonalnost među izračunatim vektorima. Vrijedi:

$$\left| q_1^T \frac{\hat{q}_2}{||\hat{q}_2||_2} \right| = \left| \frac{q_1^T q_2 + q_1^T \Delta q_2}{||\hat{q}_2||_2} \right| = \frac{1}{||\hat{q}_2||_2} |q_1^T \Delta q_2| \leq \frac{1}{||\hat{q}_2||_2} ||\Delta q_2||_2,$$

gdje smo u 2. jednakosti koristili da su q_1 i q_2 ortogonalni te Cauchy - Schwarzovu nejednakost u 1. nejednakosti te činjenicu da je q_1 normiran. Preostaje još ograničiti $||\Delta q_2||_2$:

$$\begin{aligned} ||\Delta q_2||_2 &= \sqrt{\sum_{i=1}^{m} |(\Delta q_2)_i|^2} \\ &\leq \gamma_m |q_1^T||a_2|\,||q_1||_2 + u||a_2||_2 + \gamma_2(1 + \gamma_m)|q_1^T||a_2|\,||q_1||_2 \\ &\leq \gamma_m ||a_2||_2 + u||a_2||_2 + \gamma_2(1 + \gamma_m)||a_2||_2 = (\gamma_m + \gamma_2 + \gamma_2\gamma_m)||a_j||_2 + u||a_j||_2 \\ &\leq (\gamma_{m+2} + u)||a_j||_2 \leq \gamma_{m+3}||a_j||_2, \end{aligned}$$

gdje smo u 1. nejednakosti koristili (2.9), u 2. nejednakosti Cauchy- Schwarzovu nejednakost te činjenicu da je q_1 normiran, a u 3. i 4. nejednakosti Lemu 1.3.4. Možemo približno reći da vrijedi:

$$\left| q_1^T \frac{\hat{q}_2}{||\hat{q}_2||_2} \right| \lesssim (m + 3)u \frac{||a_2||_2}{||q_2||_2}. \tag{2.10}$$

Da bi to dalje ocijenili, pomaže nam sljedeća jednakost:

$$\sin \angle(a_1, a_2) = \sqrt{1 - \cos^2 \angle(a_1, a_2)} = \sqrt{1 - \left(\frac{a_1^T a_2}{\|a_1\|_2 \|a_2\|_2}\right)^2}$$
$$= \sqrt{1 - \left(\frac{q_1^T a_2}{\|a_2\|_2}\right)^2} = \frac{1}{\|a_2\|_2} \sqrt{\|a_2\|_2 - (q_1^T a_2)^2},$$

gdje smo u 3. jednakosti iskoristili činjenicu da je $q_1 = \frac{a_1}{\|a_1\|_2}$. Kako je $q_2 = a_2 - (q_1^T a_2) q_1$, vrijedi:

$$\|q_2\|_2 = \sqrt{q_2^T q_2} = \sqrt{a_2^T a_2 - 2(q_1^T a_2)(a_2^T q_1) + (q_1^T a_2)^2 q_1^T q_1} = \sqrt{\|a_2\|_2^2 - (q_1^T a_2)^2},$$

gdje smo u zadnjoj jednakosti iskoristili da vrijedi $a_2^T q_1 = q_1^T a_2$ te da je $\|q_1\|_2 = 1$. Tada je:

$$\sin \angle(a_1, a_2) = \frac{\|q_2\|_2}{\|a_2\|_2}.$$

Nastavljajući 2.10 vrijedi:

$$\left| q_1^T \frac{\hat{q}_2}{\|\hat{q}_2\|_2} \right| \lesssim \frac{(m+3)u}{\sin \angle(a_1, a_2)} = \frac{(m+3)u \operatorname{ctg} \angle(a_1, a_2)}{\cos \angle(a_1, a_2)} \leq \frac{(m+3)u\kappa_2(A)}{c \cos \angle(a_1, a_2)},$$

gdje smo koristili činjenicu da je $\kappa_2(A) \geq c \operatorname{ctg} \angle(a_1, a_2)$, gdje je

$$A = [a_1\ a_2], \quad c = \frac{\max(\|a_1\|_2, \|a_2\|_2)}{\min(\|a_1\|_2, \|a_2\|_2}.[4, p.\ 379, 564]$$

Poznavajući vezu između MGS metode i Householder QR faktorizacije, možemo dokazati teorem koji nam daje granice za grešku MGS metode te granicu za ortogonalnost matrice $\hat{Q}$, ali u ovom slučaju za proizvoljan n.

Teorem 2.3.1. *Pretpostavimo da se MGS metoda primjenjuje na matricu $A \in \mathbb{R}^{m \times n}$ koja je ranga n te tako daje matrice $\hat{Q} \in \mathbb{R}^{m \times n}$ i $\hat{R} \in \mathbb{R}^{n \times n}$. Tada postoje konstante $c_i \equiv c_i(m, n)$ takve da vrijedi:*

$$A + \Delta A_1 = \hat{Q}\hat{R}, \quad \|\Delta A_1\|_2 \leq c_1 u \|A\|_2, \tag{2.11}$$
$$\|\hat{Q}^T \hat{Q} - I\|_2 \leq c_2 u \kappa_2(A) + O((u\kappa_2(A))^2) \tag{2.12}$$

i postoji ortogonalna matrica Q takva da vrijedi:

$$A + \Delta A_2 = Q\hat{R}, \quad \|\Delta A_2(:, j)\|_2 \leq c_3 u \|a_j\|_2, \quad j = 1, \ldots, n. \tag{2.13}$$

Dokaz. Da bi pokazali (2.11) koristimo matričnu formu MGS metode ($A_k = A_{k+1}R_k$). Vrijedi:

$$\hat{A}_{k+1}\hat{R}_k = \begin{bmatrix} \hat{q}_1 & \hat{q}_2 & \dots & \hat{q}_k & \hat{a}_{k+1}^{(k+1)} & \dots & \hat{a}_n^{(k+1)} \end{bmatrix} \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ 0 & 0 & \dots & \hat{r}_{kk} & \hat{r}_{kk+1} & \dots & \hat{r}_{kn} \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \hat{q}_1 & \hat{q}_2 & \dots & \hat{q}_{k-1} & \hat{r}_{kk}\hat{q}_k & \hat{a}_{k+1}^{(k+1)} + \hat{r}_{kk+1}\hat{q}_k & \dots & \hat{a}_n^{(k+1)} + \hat{r}_{kn}\hat{q}_k \end{bmatrix}.$$

Iz algoritma znamo da vrijedi, za $j > k$:

$$a_j^{(k+1)} = \hat{a}_j^{(k)} - \hat{r}_{kj}\hat{q}_k \tag{2.14}$$

To možemo zapisati u terminima matrične forme MGS metode na sljedeći način:

$$(A_{k+1})_j = (\hat{A}_k)_j - \hat{r}_{kj}\hat{q}_k.$$

Tada računamo:

$$(\hat{A}_{k+1})_j = fl((\hat{A}_k)_j - \hat{r}_{kj}\hat{q}_k) = \begin{bmatrix} \left((\hat{A}_k)_{1j} - \hat{r}_{kj}\hat{q}_{1k}(1+\delta_{11})\right)(1+\delta_{12}) \\ \vdots \\ \left((\hat{A}_k)_{mj} - \hat{r}_{kj}\hat{q}_{mk}(1+\delta_{m1})\right)(1+\delta_{m2}) \end{bmatrix}$$

Tada za svaki $i \in \{1, \dots, m\}$ vrijedi:

$$(\hat{A}_k)_{ij} = (1+\delta_{i2})^{-1}(\hat{A}_{k+1})_{ij} + \hat{r}_{kj}\hat{q}_{ik}(1+\delta_{i1}).$$

Prema Lemi 1.3.3 $(1+\delta_{i2})^{-1} = (1+\theta_{i2})$, gdje je $|\theta_{i2}| \leq \gamma_1 \approx u$, za svaki $i \in \{1, \dots, m\}$. Tada je:

$$(\hat{A}_k)_{ij} = (1+\theta_{i2})(\hat{A}_{k+1})_{ij} + \hat{r}_{kj}\hat{q}_{ik}(1+\delta_{i1}) = (\hat{A}_{k+1}\hat{R}_k)_{ij} + (\Delta_k)_{ij},$$

gdje smo definirali $(\Delta_k)_{ij} = \theta_{i2}(\hat{A}_{k+1})_{ij} + \hat{r}_{kj}\hat{q}_{ik}\delta_{i1}$, za $j > k$, $i \in \{1, \dots, m\}$. Vrijedi:

$$\begin{aligned} |\Delta_k|_{ij} &\leq |\theta_{12}|\hat{A}_{k+1}|_{ij} + |\hat{r}_{kj}||\hat{q}_{ij}||\delta_{i1}| \\ &\leq u(|\hat{A}_{k+1}|_{ij} + |\hat{r}_{kj}||\hat{q}_{ij}|) \\ &= u(|\hat{A}_{k+1}||\hat{R}_k|)_{ij}, \quad j > k,\ i \in \{1, \dots, m\}. \end{aligned}$$

Za $j < k$ se stupci matrica $\hat{A}_k$ i $\hat{A}_{k+1}\hat{R}_k$ poklapaju pa možemo definirati $(\Delta_k)_{ij} = 0$, za $i \in \{1, \ldots, m\}$. Iz algoritma znamo da je $q_k = \frac{\hat{a}_k^{(k)}}{\hat{r}_{kk}}$. Tada je:

$$\hat{q}_k = fl\left(\frac{\hat{a}_k^{(k)}}{\hat{r}_{kk}}\right) = fl\left(\frac{(\hat{A}_k)_k}{\hat{r}_{kk}}\right) = \begin{bmatrix} \left(\frac{(\hat{A}_k)_{1k}}{\hat{r}_{kk}}\right)(1+\delta_{1k}) \\ \vdots \\ \left(\frac{(\hat{A}_k)_{mk}}{\hat{r}_{kk}}\right)(1+\delta_{mk}) \end{bmatrix}$$

Tada je $(\hat{A}_k)_{ik} = (1+\delta_{i1})^{-1}\hat{r}_{kk}\hat{q}_{ik} = (1+\theta_{i1})\hat{r}_{kk}\hat{q}_{ik}$ za svaki $i \in \{1, \ldots, m\}$. U drugoj jednakosti smo iskoristili Lemu 1.3.3 te vrijedi $|\theta_{i1}| \leq \gamma_1 \approx u$ za svaki $i \in \{1, \ldots, m\}$. Tada je $(\hat{A}_k)_{ik} = (\hat{A}_{k+1}\hat{R}_k)_{ik} + (\Delta_k)_{ik}$, gdje smo definirali $(\Delta_k)_{ik} = \theta_{i1}\hat{r}_{kk}\hat{q}_{ik}$. Za svaki $i \in \{1, \ldots, m\}$ vrijedi:

$$|\Delta_k|_{ik} \leq u|\hat{r}_{kk}||\hat{q}_{ik}| = u(|\hat{A}_{k+1}||\hat{R}_k|)_{ik}.$$

Sve skupa smo pokazali da vrijedi:

$$\hat{A}_k = \hat{A}_{k+1}\hat{R}_k + \Delta_k, \quad |\Delta_k| \leq u|\hat{A}_{k+1}||\hat{R}_k|. \tag{2.15}$$

Koristeći (2.15) dobivamo:

$$A = \hat{Q}\hat{R} + \Delta_n\hat{R}_{n-1}\ldots\hat{R}_1 + \Delta_{n-1}\hat{R}_{n-2}\ldots\hat{R}_1 + \ldots + \Delta_2\hat{R}_1 + \Delta_1. \tag{2.16}$$

Iz (2.15) i (2.16) dobijamo:

$$|A - \hat{Q}\hat{R}| \leq u\left(|\hat{A}_{n+1}||\hat{R}_n||\hat{R}_{n-1}|\ldots|\hat{R}_1| + \ldots + |\hat{A}_3||\hat{R}_2||\hat{R}_1| + |\hat{A}_2||\hat{R}_1|\right). \tag{2.17}$$

Lako se provjeri da zbog strukture $|\hat{R}_k|$ proizvoljan član desne strane ima sljedeći oblik:

$$|\hat{A}_k||\hat{R}_{k-1}|\ldots|\hat{R}_1| = \left[|\hat{q}_1| \quad \ldots \quad |\hat{q}_{k-1}| \quad |\hat{a}_k^k| \quad \ldots|\hat{a}_n^k|\right] S_{k-1},$$

gdje je prvih $k-1$ redaka matrice S_{k-1} jednako prvih $k-1$ redaka matrice $|\hat{R}|$, a preostalih $n-k+1$ redaka odgovara identiteti. Radi jednostavnosti pretpostavimo da vrijedi $\|\hat{q}_i\|_2 \equiv 1$, $i = 1, \ldots, n$. Ta pretpostavka neće utjecati na krajnji rezultat. Iz algoritma znamo da vrijedi $a_j^{(k+1)} = \hat{a}_j^{(k)} - \hat{q}_k(\hat{q}_k^T\hat{a}_j^{(k)}) = (I - \hat{q}_k\hat{q}_k^T)\hat{a}_j^{(k)}$. Tada možemo primjeniti Lemu 1.3.6 pa iz nje dobijamo da je $\hat{a}_j^{(k+1)} = a_j^{(k+1)} + \Delta a_j^{(k+1)}$ i vrijedi:

$$\|\Delta a_j^{(k+1)}\|_2 \leq \gamma_{m+3}(1 + \|\hat{q}_k\|_2\|\hat{q}_k^T\|_2)\|\hat{a}_j^{(k)}\|_2 = 2\gamma_{m+3}\|\hat{a}_j^{(k)}\|_2. \tag{2.18}$$

Kako vrijedi da je $a_j^{(k+1)} = \hat{a}_j^{(k)} - \hat{q}_k(\hat{q}_k^T\hat{a}_j^{(k)})$, iz toga slijedi:

$$\begin{aligned} \|a_j^{(k+1)}\|_2^2 &= \|\hat{a}_j^{(k)} - \hat{q}_k(\hat{q}_k^T\hat{a}_j^{(k)})\|_2^2 \\ &= \langle \hat{a}_j^{(k)} - \langle\hat{q}_k, \hat{a}_j^{(k)}\rangle\hat{q}_k, \hat{a}_j^{(k)} - \langle\hat{q}_k, \hat{a}_j^{(k)}\rangle\hat{q}_k\rangle \\ &= \|\hat{a}_j^{(k)}\|_2^2 - 2\langle\hat{q}_k, \hat{a}_j^{(k)}\rangle^2 + \|\hat{q}_k\|_2^2\langle\hat{q}_k, \hat{a}_j^{(k)}\rangle^2 \\ &= \|\hat{a}_j^{(k)}\|_2^2 - \langle\hat{q}_k, \hat{a}_j^{(k)}\rangle^2 \leq \|\hat{a}_j^{(k)}\|_2^2 \end{aligned}$$

Koristeći (2.18) i $\|a_j^{(k+1)}\|_2 \leq \|\hat{a}_j^{(k)}\|_2$ dobijamo:

$$\|\hat{a}_j^{(k+1)}\|_2 \leq \|a_j^{(k+1)}\|_2 + \|\Delta a_j^{(k+1)}\|_2 \leq \|\hat{a}_j^{(k)}\|_2 + 2\gamma_{m+3}\|\hat{a}_j^{(k)}\|_2 = (1+2\gamma_{m+3})\|\hat{a}_j^{(k)}\|_2 \tag{2.19}$$

što implicira:

$$\|\hat{A}_{k+1}\|_F = \sqrt{\sum_{j=1}^{n} \|\hat{a}_j^{(k+1)}\|_2^2} \leq (1+2\gamma_{m+3}) \sqrt{\sum_{j=1}^{n} \|\hat{a}_j^{(k)}\|_2^2} = (1+2\gamma_{m+3})\|\hat{A}_k\|_F.$$

Rekurzivno se dobije:

$$\|\hat{A}_{k+1}\|_F \leq (1+2\gamma_{m+3})^k \|\hat{A}_1\|_F = (1+2\gamma_{m+3})^k \|A\|_F. \tag{2.20}$$

Iz $\hat{r}_{kj} = fl(\hat{q}_k^T \hat{a}_j^{(k)})$ vrijedi $\hat{r}_{kj} = (\hat{q}_k^T + \Delta\hat{q}_k^T)\hat{a}_j^{(k)}$, gdje je $|\Delta\hat{q}_k^T| \leq \gamma_m |\hat{q}_k^T|$. Tada je:

$$\begin{aligned}
\|\hat{R}\|_F &= \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} |\hat{r}_{kj}|^2} \leq (1+\gamma_m) \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} (|\hat{q}_k^T \|\hat{a}_j^{(k)}|)^2} \\
&\leq (1+\gamma_m) \sqrt{\sum_{k=1}^{n}\sum_{j=1}^{n} \|\hat{a}_j^{(k)}\|_2^2} = (1+\gamma_m) \sqrt{\sum_{k=1}^{n} \|\hat{A}_k\|_F^2} \\
&\leq \sqrt{n}(1+\gamma_m)(1+2\gamma_{m+3})^{n-1}\|A\|_F,
\end{aligned} \tag{2.21}$$

gdje smo u 2. nejednakosti koristili Cauchy-Schwarzovu nejednakost te činjenicu da je $\|\hat{q}_k^T\|_2 = 1$, a u 3. nejednakosti koristili (2.20), tj. da je $\|\hat{A}_k\|_F \leq (1+2\gamma_{m+3})^{n-1}\|A\|_F$, $k = 1, \ldots, n$. Želimo izračunati ogradu za $\|A - \hat{Q}\hat{R}\|_F$. Koristeći (2.17) dobijamo:

$$\begin{aligned}
\|A - \hat{Q}\hat{R}\|_F &\leq u \left\| |\hat{A}_{n+1}\|\hat{R}_n\|\hat{R}_{n-1}| \ldots |\hat{R}_1| + \ldots + |\hat{A}_3\|\hat{R}_2\|\hat{R}_1| + |\hat{A}_2\|\hat{R}_1| \right\|_F \\
&\leq u \left(\|\hat{Q}\|_F \|\hat{R}\|_F + \left\| |\hat{A}_n\|\hat{R}_{n-1}| \ldots |\hat{R}_1| \right\|_F + \ldots + \left\| |\hat{A}_2\|\hat{R}_1| \right\|_F \right).
\end{aligned} \tag{2.22}$$

Znamo da je $\|\hat{Q}\|_F = \sqrt{n}$ (zbog pretpostavke da je $\|\hat{q}_i\|_2 = 1$) te da je $\|\hat{R}\|_F \leq \sqrt{n}(1+\gamma_m)(1+2\gamma_{m+3})^{n-1}\|A\|_F$ pa je

$$\|\hat{Q}\|_F \|\hat{R}\|_F \leq n(1+\gamma_m)(1+2\gamma_{m+3})^{n-1}\|A\|_F. \tag{2.23}$$

Još ostaje ocijeniti članove oblika $\left\| |\hat{A}_k\|\hat{R}_{k-1}| \ldots |\hat{R}_1| \right\|_F = \left\| |\hat{A}_k\| S_{k-1}| \right\|_F$. Ako k-ti redak matrice $\hat{R}$ označimo s:

$$\hat{r}_k = \begin{bmatrix} 0 & 0 & \ldots & \hat{r}_{kk} & \hat{r}_{kk+1} & \ldots & \hat{r}_{kn} \end{bmatrix},$$

te s e_i vektore standardne ortonormirane baze za $\mathbb{R}^n$ imamo:

$$|\hat{A}_k||S_{k-1}| = \begin{bmatrix} |\hat{q}_1| & \dots & |\hat{q}_{k-1}| & |\hat{a}_k^{(k)}| & \dots & |\hat{a}_n^{(k)}| \end{bmatrix} \begin{bmatrix} |\hat{r}_1| \\ \vdots \\ |\hat{r}_{k-1}| \\ e_k^T \\ \vdots \\ e_n^T \end{bmatrix}$$
$$= \begin{bmatrix} |\hat{r}_{11}||\hat{q}_1| & \dots & \sum_{j=1}^{k-1} |\hat{r}_{jk-1}||\hat{q}_j| & |\hat{a}_k^{(k)}| + \sum_{j=1}^{k-1} |\hat{r}_{jk}||\hat{q}_j| & \dots & |\hat{a}_n^{(k)}| + \sum_{j=1}^{k-1} |\hat{r}_{jn}||\hat{q}_j| \end{bmatrix}. \tag{2.24}$$

Kako je $\|A\|_F = \sqrt{\sum_{i=1}^n \|a_i\|_2^2}$, gdje je a_i i-ti stupac matrice A, za $i \geq k$ računamo:

$$\left\| \left(|\hat{A}_k||S_{k-1}| \right)_i \right\|_2 = \left\| |\hat{a}_i^{(k)}| + \sum_{j=1}^{k-1} |\hat{r}_{ji}||\hat{q}_j| \right\|_2 \leq \|\hat{a}_i^{(k)}\|_2 + \sum_{j=1}^{k-1} |\hat{r}_{ji}| \left\| \hat{q}_j \right\|_2 = \|\hat{a}_i^{(k)}\|_2 + \sum_{j=1}^{k-1} |\hat{r}_{ji}|, \tag{2.25}$$

gdje smo u zadnjoj jednakosti koristili našu pretpostavku da je $\|\hat{q}_j\|_2 = 1$. Kako smo u (2.21) koristili da je $|\hat{r}_{ji}| \leq (1+\gamma_m)(1+2\gamma_{m+3})^{k-1}\|a_i\|_2$ onda je:

$$\sum_{j=1}^{k-1} |\hat{r}_{ji}| \leq \sum_{j=1}^{k-1} (1+\gamma_m)(1+2\gamma_{m+3})^{k-1}\|a_i\|_2 = (k-1)(1+\gamma_m)(1+2\gamma_{m+3})^{k-1}\|a_i\|_2$$
$$\leq (n-1)(1+\gamma_m)(1+2\gamma_{m+3})^{k-1}\|a_i\|_2.$$

Tada nastavljajući (2.25) dobijamo:

$$\left\| \left(|\hat{A}_k||S_{k-1}| \right)_i \right\|_2 \leq (1+2\gamma_{m+3})^{k-1}\|a_i\|_2 + (n-1)(1+\gamma_m)(1+2\gamma_{m+3})^{k-1}\|a_i\|_2$$
$$= (1+2\gamma_{m+3})^{k-1} \left(1 + (n-1)(1+\gamma_m)\right) \|a_i\|_2, \tag{2.26}$$

gdje smo u 1. nejednakosti rekurzivno primijenili (2.19). Da bi izračunali $\||\hat{A}_k||S_{k-1}|\|_F$ nedostaju nam ocjene stupaca $i < k$. Za $i < k$ iz (2.24) vidimo da vrijedi:

$$\left\| \left(|\hat{A}_k||S_{k-1}| \right)_i \right\|_2 \leq \sum_{j=1}^{i} |\hat{r}_{ji}| \left\| \hat{q}_j \right\|_2 \leq \sum_{j=1}^{k-1} |\hat{r}_{ji}| \leq (1+2\gamma_{m+3})^{k-1} \left(1 + (n-1)(1+\gamma_m)\right) \|a_i\|_2, \tag{2.27}$$

gdje smo u zadnjoj nejednakosti iskoristili račun napravljen za (2.26). Relacije (2.26) i (2.27) nam daju:

$$\begin{aligned}\left\||\hat{A}_k||S_{k-1}|\right\|_F &= \sqrt{\sum_{i=1}^{n}\left\|\left(|\hat{A}_k||S_{k-1}|\right)_i\right\|_2^2} \\ &\leq (1+2\gamma_{m+3})^{k-1}\,(1+(n-1)(1+\gamma_m))\,\sqrt{\sum_{i=1}^{n}\|a_i\|_2^2} \\ &\leq (1+2\gamma_{m+3})^{n-1}\,(1+(n-1)(1+\gamma_m))\,\|A\|_F. \end{aligned} \tag{2.28}$$

Pretpostavit ćemo da vrijedi $(1+\gamma_m)(1+2\gamma_{m+3})^{n-1} < 2$. Također onda vrijedi $(1+2\gamma_{m+3})^{n-1} \leq 2$ i $(1+\gamma_m) \leq 2$. Uz tu pretpostavku (2.23) postaje

$$\|\hat{Q}\|_F\|\hat{R}\|_F \leq 2n\|A\|_F \leq 4n\|A\|_F \tag{2.29}$$

te (2.28) postaje

$$\left\||\hat{A}_k||S_{k-1}|\right\|_F \leq 2\,(1+2(n-1))\,\|A\|_F \leq 4n\|A\|_F. \tag{2.30}$$

Konačno nastavljajući (2.22) iz (2.29) i (2.28) dobijamo:

$$\|A-\hat{Q}\hat{R}\|_F \leq u(4n+4n+\ldots+4n)\|A\|_F = 4n^2u\|A\|_F = c_1\|A\|_F,$$

gdje je $c_1 = 4n^2$. Da bi dokazali zadnja dvije tvrdnje teorema, iskoristit ćemo vezu MGS metode i QR faktorizacije pomoću Householderovih reflektora. Imamo matricu $\begin{bmatrix}0_n\\A\end{bmatrix} \in \mathbb{R}^{(m+n)\times n}$ te njenu Householder QR faktorizaciju:

$$P^T\begin{bmatrix}0_n\\A\end{bmatrix} = \begin{bmatrix}R\\0\end{bmatrix},$$

gdje je $P^T = P_n\ldots P_2P_1$. Po Teoremu 2.1.4 tada postoji ortogonalna matrica $\tilde{P} \in \mathbb{R}^{(m+n)\times(m+n)}$ takva da

$$\begin{bmatrix}\Delta A_3\\A+\Delta A_4\end{bmatrix} = \tilde{P}\begin{bmatrix}\hat{R}\\0\end{bmatrix} = \begin{bmatrix}\tilde{P}_{11}\\\tilde{P}_{21}\end{bmatrix}\hat{R} \tag{2.31}$$

gdje je $\tilde{P}_{11} \in \mathbb{R}^{n\times n}$ i $\tilde{P}_{21} \in \mathbb{R}^{m\times n}$ te vrijedi

$$\left\|\begin{bmatrix}\Delta A_3(:,j)\\\Delta A_4(:,j)\end{bmatrix}\right\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2, \quad j=1,\ldots,n. \tag{2.32}$$

Imamo $A+\Delta A_4 = \tilde{P}_{21}\hat{R}$, ali nam to nažalost ne daje (2.13) jer $\tilde{P}_{21}$ nije nužno ortogonalna matrica. No primjenom Leme 1.3.7 na (2.31) postoji ortogonalna matrica $Q \in \mathbb{R}^{m\times n}$ (koja je

ujedno i najbliža ortogonalna matrica matrici $\tilde{P}_{21}$ u 2-normi i Frobeniusovoj normi) takva da vrijedi $A + \Delta A_2 = Q\hat{R}$, gdje je $\Delta A_2 = F\Delta A_3 + \Delta A_4$, $\|F\|_2 \leq 1$. Tada vrijedi:

$$\begin{aligned}\|\Delta A_2(:, j)\| &\leq \|(F\Delta A_3)(:, j)\|_2 + \|\Delta A_4(:, j)\|_2 \leq \|F\|_2\|\Delta A_3(:, j)\|_2 + \|\Delta A_4(:, j)\|_2 \\ &\leq \|F\|_2\tilde{\gamma}_{mn}\|a_j\|_2 + \tilde{\gamma}_{mn}\|a_j\|_2 \leq 2\tilde{\gamma}_{mn}\|a_j\|_2 = c_3u\|a_j\|_2,\end{aligned}$$

gdje smo u 3. nejednakosti koristili (2.32) te na kraju definirali

$$c_3 := \frac{2cmn}{1 - cmnu}.$$

Ovime smo pokazali (2.13). Ocjene u (2.13) možemo pomoću Leme 1.3.2 zapisati na sljedeći način:

$$\|\Delta A_2\|_2 \leq \sqrt{n}c_3u\|A\|_2. \tag{2.33}$$

Iz $A + \Delta A_2 = Q\hat{R}$ slijedi da je $\hat{R} = Q^T(A + \Delta A_2) = Q^T(I_m + \Delta A_2A^+)A$, jer A ima puni stupčani rang pa je $A^+A = I_n$ pa onda uz pretpostavku regularnosti matrice $\hat{R}$ vrijedi $\hat{R}^{-1} = A^+(I_m + \Delta A_2A^+)^{-1}Q$. Tada je:

$$\|\hat{R}^{-1}\|_2 \leq \|A^+\|_2\|(I_m + \Delta A_2A^+)^{-1}\|_2\|Q\|_2 = \|A^+\|_2\|(I_m + \Delta A_2A^+)^{-1}\|_2. \tag{2.34}$$

Iz 2.33 slijedi:

$$\|\Delta A_2A^+\|_2 \leq \sqrt{n}c_3u\|A\|_2\|A^+\|_2 = \sqrt{n}c_3u\kappa_2(A). \tag{2.35}$$

Za sve osim matrica kojima je uvjetovanost jako velika vrijedit će $\sqrt{n}c_3u\kappa_2(A) < 1$, tj. $\|\Delta A_2A^+\|_2 < 1$. Onda možemo iskoristiti teorem 1.3.10 pa vrijedi:

$$(I_m + \Delta A_2A^+)^{-1} = \sum_{k=0}^{\infty}(\Delta A_2A^+)^k.$$

Tada iz (2.34) možemo dobiti sljedeću ocjenu:

$$\|\hat{R}^{-1}\|_2 \leq \|A^+\|_2 \sum_{k=0}^{\infty}\|\Delta A_2A^+\|_2^k \leq \|A^+\|_2 \sum_{k=0}^{\infty}\left(\sqrt{n}c_3u\kappa_2(A)\right)^k = \frac{\|A^+\|_2}{1 - \sqrt{n}c_3u\kappa_2(A)}, \tag{2.36}$$

gdje smo u 2. nejednakosti iskoristili (2.35). Ako uvedemo polarnu dekompoziciju matrice $\hat{Q} = UH$, tada po Lemi 1.3.8 :

$$\|\hat{Q}^T\hat{Q} - I\|_2 \leq (1 + \|\hat{Q}\|_2)\|\hat{Q} - U\|_2.$$

Jer je Q ortogonalna matrica primjenom Teorema 1.3.9 dobijamo:

$$\|\hat{Q} - U\|_2 \leq \|\hat{Q} - Q\|_2,$$

odnosno

$$\|\hat{Q}^T\hat{Q} - I\|_2 \leq (1 + \|\hat{Q}\|_2)\|\hat{Q} - Q\|_2. \tag{2.37}$$

Uz pretpostavku regularnosti matrice $\hat{R}$, iz $A + \Delta A_1 = \hat{Q}\hat{R}$ slijedi $\hat{Q} = (A + \Delta A_1)\hat{R}^{-1}$ te iz $A + \Delta A_2 = Q\hat{R}$ slijedi $Q = (A + \Delta A_2)\hat{R}^{-1}$ pa vrijedi:

$$\hat{Q} - Q = (\Delta A_1 - \Delta A_2)\hat{R}^{-1}. \tag{2.38}$$

Konačno:

$$\begin{aligned}\|\hat{Q}^T\hat{Q} - I\|_2 &\leq (1 + \|\hat{Q}\|_2)\|\hat{Q} - Q\|_2 \leq (1 + \|\hat{Q}\|_2)\|(\Delta A_1 - \Delta A_2)\|_2\|\hat{R}^{-1}\|_2 \\ &\leq (1 + \|\hat{Q}\|_2)(c_1 + \sqrt{n}c_3)u\|A\|_2\|\hat{R}^{-1}\|_2 \\ &\leq \frac{c_2 u\kappa_2(A)}{1 - \sqrt{n}c_3 u\kappa_2(A)} = c_2 u\kappa_2(A)(1 + O(u\kappa_2(A))),\end{aligned}$$

gdje smo definirali $c_2 = (1+\|\hat{Q}\|_2)(c_1+\sqrt{n}c_3)$ te redom koristili (2.37), (2.38), (2.11), (2.33) i (2.36) te na kraju Taylorov razvoj geometrijskog reda. Konstanta c_2 se može zapisati kao $c_2 = (2 + O(u))(c_1 + \sqrt{n}c_3)$ jer vrijedi:

$$\begin{aligned}\|\hat{Q}\|_2 &= \sqrt{\rho(\hat{Q}^T\hat{Q})} = \sqrt{\rho\left(\left((\hat{Q} - Q) + Q\right)^T\left((\hat{Q} - Q) + Q\right)\right)} \\ &= \sqrt{\rho\left(I + Q^T\left(\hat{Q} - Q\right) + \left(\hat{Q} - Q\right)^T Q + \left(\hat{Q} - Q\right)^T\left(\hat{Q} - Q\right)\right)} \\ &= \sqrt{1 + O(u)} = 1 + O(u).\end{aligned}$$

□

Teorem 2.3.1 nam govori 3 stvari. Prva je da greška unatrag mala pa metodu možemo smatrati stabilnom unatrag. Druga je da je udaljenost matrice $\hat{Q}$ od njoj najbliže ortogonalne matrice Q ograničena s višekratnikom od $\kappa_2(A)u$, što nam onda garantira da je matrica $\hat{Q}$ skoro ortogonalna kad je matrica A dobro uvjetovana. Treća je da je $\hat{R}$ dobar kao i R faktor izračunat Householder QR faktorizacijom (u smislu stupčane ocjene). MGS metoda je slabija od Householder QR faktorizacije jedino zbog toga što $\hat{Q}$ nije garantirano blizu ortogonalna matrica.

2.4 Teorija perturbacije problema najmanjih kvadrata

Glavni rezultat za problem najmanjih kvadrata kojim se bavimo je perturbacijski teorem po normi koji je dokazao Wedin.

Teorem 2.4.1. *(Wedin [7, Teorem 5.1]) Neka $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ i $A + \Delta A$ imaju puni stupčani rang $(n = r(A))$. Neka je:*

$$\begin{aligned} \|b - Ax\|_2 = \min, \qquad & r = b - Ax, \\ \|(b + \Delta b) - (A + \Delta A)y\|_2 = \min, \qquad & s = b + \Delta b - (A + \Delta A)y, \\ \|\Delta A\|_2 \leq \epsilon\|A\|_2 \qquad & \|\Delta b\|_2 \leq \epsilon\|b\|_2. \end{aligned} \tag{2.39}$$

Tada, uz pretpostavku da je $\kappa_2(A)\epsilon < 1$,

$$\frac{\|x - y\|_2}{\|x\|_2} \leq \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(2 + (\kappa_2(A) + 1)\frac{\|r\|_2}{\|A\|_2\|x\|_2}\right), \tag{2.40}$$

$$\frac{\|r - s\|_2}{\|b\|_2} \leq (1 + 2\kappa_2(A))\epsilon. \tag{2.41}$$

Ove granice su aproksimativno dostižne.

Dokaz. Neka je $B := A + \Delta A$. Kako je y rješenje $\|(b + \Delta b) - By\|_2 = \min$, slijedi da je y rješenje normalne jednadžbe $B^T By = B^T(b+\Delta b)$. Kako je po pretpostavci teorema B punog ranga slijedi da je $B^T B$ regularna matrica pa je $y = (B^T B)^{-1}B^T(b + \Delta b) = B^+(b + \Delta b)$. Na analogan način se može dobiti da je $x = A^+ b$ jer A ima puni rang. Imamo:

$$\begin{aligned} y - x &= B^+(b + \Delta b) - x = B^+(r + Ax + \Delta b) - x \\ &= B^+(r + Bx - \Delta Ax + \Delta b) - x \\ &= B^+(r - \Delta Ax + \Delta b) + B^+Bx - x \\ &= B^+(r - \Delta Ax + \Delta b), \end{aligned} \tag{2.42}$$

gdje smo u 2. jednakosti iskoristili $b = r + Ax$, a u zadnjoj da je $B^+B = I_n$. Vrijedi:

$$B^+r = B^+(BB^+)r = B^+P_Br = B^+P_B(I - P_A)r, \tag{2.43}$$

gdje smo u zadnjoj jednakosti koristili činjenicu da je

$$P_Ar = AA^+r = AA^+(b - Ax) = AA^+b - Ax = Ax - Ax = 0.$$

Sad možemo ocijeniti:

$$\begin{aligned} \|B^+r\|_2 &\leq \|B^+\|_2\,\|P_B(I - P_A)\|_2\,\|r\|_2 \\ &\leq \|B^+\|_2(\|B - A\|_2\|A^+\|_2)\|r\|_2 \\ &\leq \frac{\|A^+\|_2}{1 - \|A^+\|_2\|B - A\|_2}(\|B - A\|_2\|A^+\|_2)\|r\|_2 \\ &= \frac{\|A^+\|_2}{1 - \|A^+\|_2\|\Delta A\|_2}\|\Delta A\|_2\|A^+\|_2\|r\|_2, \end{aligned} \tag{2.44}$$

gdje smo u 2. nejednakosti iskoristili Lemu 1.3.13, a u 3. Lemu 1.3.12. Iz $||\Delta A||_2 \leq \epsilon||A||_2$ slijedi da je $||A^+||_2||\Delta A||_2 \leq \kappa_2(A)\epsilon$ iz čega slijedi da je

$$\frac{1}{1 - ||A^+||_2||\Delta A||_2} \leq \frac{1}{1 - \kappa_2(A)\epsilon} \tag{2.45}$$

zbog pretpostavke da je $\kappa_2(A)\epsilon < 1$. Tada nastavljajući (2.44) vrijedi:

$$\begin{aligned} ||B^+ r||_2 &\leq \frac{||A^+||_2}{1 - \kappa_2(A)\epsilon}\epsilon||A||_2||A^+||_2||r||_2 \\ &= \frac{\kappa_2(A)^2\epsilon}{1 - \kappa_2(A)\epsilon}\frac{||r||_2}{||A||_2}, \end{aligned} \tag{2.46}$$

gdje smo u 1. nejednakosti iskoristili $||\Delta A||_2 \leq \epsilon||A||_2$ te (2.45). Ostaje još ocijeniti ostatak u (2.42). Vrijedi:

$$\begin{aligned} ||B^+(-\Delta A x + \Delta b)||_2 &\leq ||B^+||_2(||\Delta A||_2||x||_2 + ||\Delta b||_2) \\ &\leq ||B^+||_2\,\epsilon\,(||A||_2||x||_2 + ||b||_2) \\ &\leq \frac{||A^+||_2}{1 - ||A^+||_2||\Delta A||_2}\epsilon\,(||A||_2||x||_2 + ||b||_2) \\ &\leq \frac{||A^+||_2||A||_2}{1 - \kappa_2(A)\epsilon}\epsilon\left(1 + \frac{||b||_2}{||A||_2||x||_2}\right)||x||_2 \\ &\leq \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(2 + \frac{||r||_2}{||A||_2||x||_2}\right)||x||_2, \end{aligned} \tag{2.47}$$

gdje smo u 2. nejednakosti iskoristili pretpostavku (2.39), u 3. nejednakosti Lemu 1.3.12, u 4. (2.45) te u 5. nejednakosti činjenicu da je $b = r + Ax$, tj. da je $||b||_2 \leq ||r||_2 + ||A||_2||x||_2$. Imamo:

$$\begin{aligned} \frac{||x - y||_2}{||x||_2} &\leq \frac{||B^+ r||_2 + ||B^+(-\Delta A x + \Delta b)||_2}{||x||_2} \\ &\leq \frac{\kappa_2(A)^2\epsilon}{1 - \kappa_2(A)\epsilon}\frac{||r||_2}{||A||_2||x||_2} + \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(2 + \frac{||r||_2}{||A||_2||x||_2}\right) \\ &= \frac{\kappa_2(A)\epsilon}{1 - \kappa_2(A)\epsilon}\left(2 + (\kappa_2(A) + 1)\frac{||r||_2}{||A||_2||x||_2}\right), \end{aligned}$$

gdje smo u 1. nejednakosti iskoristili (2.42), u 2. nejednakosti (2.46) i (2.47). Time smo dokazali (2.40). Da bi dokazali (2.41) prelazimo na reziduale. Vrijedi:

$$\begin{aligned} s - r &= (b + \Delta b) - By - b + Ax \\ &= \Delta b - \Delta A x + B(x - y) \\ &= \Delta b - \Delta A x - BB^+(r - \Delta A x + \Delta b) \\ &= (I_m - BB^+)(\Delta b - \Delta A x) - BB^+ r, \end{aligned} \tag{2.48}$$

gdje smo u 3. jednakosti iskoristili (2.42). Kako je BB^+ projektor ($(BB^+)^2 = BB^+BB^+ = BB^+$ jer B ima puni stupčani rang) onda je i $I_m - BB^+$ projektor. U slučaju $m = n$, zbog punog stupčanog ranga, vrijedi $B^+ = B^{-1}$ pa je $I_m - BB^+ = 0_m$. U slučaju $m > n$ je $I_m - BB^+$ netrivijalan projektor. Kako je spektar projektora sadržan u skupu $\{0, 1\}$ te imamo netrivijalan projektor slijedi da je $||I_m - BB^+||_2 = 1$. Zaključujemo da za $m \geq n$ vrijedi

$$||I_m - BB^+||_2 \leq 1. \tag{2.49}$$

Tada je iz (2.48):

$$\begin{aligned} ||r - s||_2 &\leq ||I_m - BB^+||_2 \left(||\Delta b||_2 + ||\Delta A x||_2\right) + ||BB^+ r||_2 \leq ||\Delta b||_2 + ||\Delta A x||_2 + ||BB^+ r||_2 \\ &\leq \epsilon \left(||b||_2 + ||A||_2||x||_2\right) + ||BB^+ r||_2, \end{aligned} \tag{2.50}$$

gdje smo u 3. nejednakosti iskoristili (2.39). Kako je BB^+ netrivijalan projektor vrijedi $||BB^+||_2 = 1$. Vrijedi:

$$\begin{aligned} ||BB^+ r||_2 &= ||BB^+ P_B(I - P_A)r||_2 \leq ||BB^+||_2||P_B(I - P_A)||_2||r||_2 \\ &\leq ||B - A||_2||A^+||_2||r||_2 = ||\Delta A||_2||A^+||_2||r||_2 \leq \kappa_2(A)\epsilon||r||_2, \end{aligned}$$

gdje smo u 1. jednakosti iskoristili (2.43), u 2. nejednakosti Lemu 1.3.13 i u 3. nejednakosti (2.39). Iz toga te iz (2.50) slijedi:

$$\frac{||r - s||_2}{||b||_2} \leq \epsilon \left(1 + \frac{||A||_2||x||_2}{||b||_2} + \kappa_2(A)\frac{||r||_2}{||b||_2}\right). \tag{2.51}$$

Iz $x = A^+b$ slijedi da je $r = b - Ax = b - AA^+b = (I_m - AA^+)b$. Kako matrica A ima puni stupčani rang, vrijedi da je AA^+ projektor $\left((AA^+)^2 = AA^+AA^+ = AA^+\right)$ pa je i $I_m - AA^+$ projektor pa kao što smo pokazali u (2.49) vrijedi da je $||I_m - AA^+||_2 \leq 1$. Tada je:

$$||r||_2 \leq ||I_m - AA^+||_2||b||_2 \leq ||b||_2. \tag{2.52}$$

Nastavljajući (2.51) dobijamo:

$$\frac{||r - s||_2}{||b||_2} \leq \epsilon(1 + 2\kappa_2(A))$$

gdje smo iskoristili činjenicu da je $x = A^+b$ pa je $||x||_2 \leq ||A^+||_2||b||_2$ te činjenicu (2.52). □

Izvest ćemo još perturbacijske ograde po komponentama. One će nam biti potrebne za naći ogradu za rezidual problema najmanjih kvadrata riješenim QR faktorizacijom. Kako je $r = b - Ax$ te rješenje problema najmanjih kvadrata zadovoljava normalnu jednadžbu $A^TAx = A^Tb$, tj. $A^Tr = 0$, uvodimo prošireni sustav:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}. \tag{2.53}$$

Matrica sustava je regularna jer

$$\det\left(\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}\right) = \det(I)\det(0 - A^T I^{-1} A) = \det(-A^T A) \neq 0,$$

jer A ima puni stupčani rang pa je matrica $A^T A$ pozitivno definitna. Također perturbirani problem možemo zapisati kao prošireni sustav:

$$\begin{bmatrix} I & A + \Delta A \\ (A + \Delta A)^T & 0 \end{bmatrix} \begin{bmatrix} s \\ y \end{bmatrix} = \begin{bmatrix} b + \Delta b \\ 0 \end{bmatrix}, \tag{2.54}$$

gdje smo pretpostavili

$$|\Delta A| \leq \epsilon E, \quad |\Delta b| \leq \epsilon f. \tag{2.55}$$

Oduzimanjem sustava (2.53) i (2.54) dobijamo

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} s - r \\ y - x \end{bmatrix} = \begin{bmatrix} \Delta b - \Delta A y \\ -\Delta A^T s \end{bmatrix}. \tag{2.56}$$

Koristeći činjenicu [4, Problem 13.8. p. 258] da je

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} I - A(A^T A)^{-1} A^T & A(A^T A)^{-1} \\ (A^T A)^{-1} A^T & -(A^T A)^{-1} \end{bmatrix} = \begin{bmatrix} I - AA^+ & (A^+)^T \\ A^+ & -(A^T A)^{-1} \end{bmatrix},$$

gdje smo iskoristili činjenicu da je $A^+ = (A^T A)^{-1} A^T$, množimo sustav (2.56) s inverzom te dobijamo

$$\begin{bmatrix} s - r \\ y - x \end{bmatrix} = \begin{bmatrix} I - AA^+ & (A^+)^T \\ A^+ & -(A^T A)^{-1} \end{bmatrix} \begin{bmatrix} \Delta b - \Delta A y \\ -\Delta A^T s \end{bmatrix}. \tag{2.57}$$

Koristeći sad (2.55) dobijamo perturbacijske ocjene:

$$|s - r| \leq \epsilon \left(\left| I - A^+ A \right| (f + E|y|) + \left| A^+ \right|^T E^T |s| \right),$$
$$|y - x| \leq \epsilon \left(\left| A^+ \right| (f + E|y|) + \left| \left(A^T A \right)^{-1} \right| E^T |s| \right).$$

2.5 Analiza greške problema najmanjih kvadrata riješenih pomoću QR faktorizacije

Sljedeći rezultat nam daje stupčane ograde za grešku unatrag problema najmanjih kvadrata riješenih QR faktorizacijom pomoću Householderovih reflektora (skoro isti rezultat će vrijediti i za QR faktorizaciju pomoću Givensovih rotacija). Kako će te ograde biti male, zbog Leme 1.3.2 zaključujemo da su QR faktorizacija pomoću Householderovih reflektora te QR faktorizacija pomoću Givensovih rotacija stabilne unatrag po normi metode za rješavanje problema najmanjih kvadrata.

Teorem 2.5.1. *Neka je $A \in \mathbb{R}^{m\times n}$, $m \geq n$. Neka A ima puni rang i pretpostavimo da problem najmanjih kvadrata $\min_x \|b - Ax\|_2$ rješavamo koristeći QR faktorizaciju pomoću Householderovih reflektora. Izračunato rješenje $\hat{x}$ je egzaktno rješenje problema najmanjih kvadrata:*

$$\min_x \|(b + \Delta b) - (A + \Delta A)\hat{x}\|_2,$$

gdje

$$\|\Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2, \quad j = 1, \ldots, n, \quad \|\Delta b\|_2 \leq \tilde{\gamma}_{mn}\|b\|_2.$$

Dokaz. Neka je $A = Q\begin{bmatrix} R \\ 0 \end{bmatrix}$ QR faktorizacija matrice A. Po Teoremu 2.1.4 izračunati gornje trokutasti QR faktor $\hat{R}$ zadovoljava $A + \Delta A = Q\begin{bmatrix} \hat{R} \\ 0 \end{bmatrix}$ s ocjenama $\|\Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2$, $j = 1, \ldots, n$. Iz (1.4) nas zanima vektor $Q^T b = \begin{bmatrix} c \\ d \end{bmatrix}$. Tada je po Lemi 2.1.3 izračunati vektor jednak $Q^T(b + \Delta b) = \begin{bmatrix} \hat{c} \\ \hat{d} \end{bmatrix}$ s ocjenom $\|\Delta b\|_2 \leq \tilde{\gamma}_{mn}\|b\|_2$. Iz (1.4) nas također zanima rješenje sustava $\hat{R}x = \hat{c}$. Prema Teoremu 1.3.14 izračunato rješenje $\hat{x}$ zadovoljava

$$(\hat{R} + \Delta\hat{R})\hat{x} = \hat{c}, \quad |\Delta\hat{R}| \leq \gamma_n|\hat{R}|,$$

tj. $\hat{x}$ je egzaktno rješenje sustava $(\hat{R} + \Delta\hat{R})x = \hat{c}$. Prema (1.4) tada je $\hat{x}$ egzaktno rješenje problema najmanjih kvadrata:

$$\min_x \left\| \begin{bmatrix} (\hat{R} + \Delta\hat{R})x - \hat{c} \\ -\hat{d} \end{bmatrix} \right\|_2 = \min_x \left\| Q \begin{bmatrix} \hat{R} + \Delta\hat{R} \\ 0 \end{bmatrix} x - (b + \Delta b) \right\|_2.$$

Ako definiramo

$$A + \overline{\Delta A} := A + \Delta A + Q\begin{bmatrix} \Delta\hat{R} \\ 0 \end{bmatrix},$$

tada je $\hat{x}$ egzaktno rješenje problema najmanjih kvadrata

$$\min_x \left\| (A + \overline{\Delta A})x - (b + \Delta b) \right\|_2.$$

Vrijedi:

$$\begin{aligned} \|\overline{\Delta a}_j\|_2 &\leq \|\Delta a_j\|_2 + \left\|(\Delta\hat{R})_j\right\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2 + \gamma_n\|\hat{r}_j\|_2 \\ &= \tilde{\gamma}_{mn}\|a_j\|_2 + \gamma_n\|a_j + \Delta a_j\|_2 \leq \tilde{\gamma}_{mn}\|a_j\|_2, \end{aligned}$$

gdje smo u 2. nejednakosti iskoristili činjenicu da je $\|Q\|_2 = 1$, u 1. jednakosti činjenicu da je $\left\|(Q^T(A + \Delta A))_j\right\|_2 = \left\|Q^T(a_j + \Delta a_j)\right\|_2 = \|a_j + \Delta a_j\|_2$ jer je Q ortogonalna matrica. □

Teorem 2.5.1 nam ipak ne daje ogradu za rezidual izračunatog rješenja, koji mi želimo minimizirati. Pitanje je koliko je $\|b - A\hat{x}\|_2$ blizu $\min_x \|b - Ax\|_2$. Definirajmo $\hat{r} = b + \Delta b - (A + \Delta A)\hat{x}$, $x := x_{\mathrm{LS}}$ (rješenje problema najmanjih kvadrata $\min_x \|b - Ax\|_2$) i $r = b - Ax$. Tada nam (2.57) daje

$$\hat{r} - r = \left(I - AA^+\right)(\Delta b - \Delta A\hat{x}) - (A^+)^T \Delta A^T \hat{r}$$

pa je

$$(b - A\hat{x}) - r = -AA^+ (\Delta b - \Delta A\hat{x}) - (A^+)^T \Delta A^T \hat{r}.$$

Ocjene iz Teorema 2.5.1 po Lemi 1.3.2 možemo zamijeniti s

$$|\Delta A| \leq \tilde{\gamma}_{mn} ee^T |A|, \quad |\Delta b| \leq \tilde{\gamma}_{mn} ee^T |b|. \tag{2.58}$$

Tada vrijedi:

$$\begin{aligned} \|(b - A\hat{x}) - r\|_2 &\leq \|AA^+\|_2 \|\Delta b - \Delta A\hat{x}\|_2 + \left\|(A^+)^T \Delta A^T \hat{r}\right\|_2 \\ &\leq \| \, |\Delta b| + |\Delta A||\hat{x}| \, \|_2 + \left\| \left|(A^+)^T\right| \left|\Delta A^T\right| |\hat{r}| \right\|_2 \\ &\leq \tilde{\gamma}_{mn} \left(\|ee^T(|b| + |A||\hat{x}|)\|_2 + \left\| \left|(A^+)^T\right| \left|A^T\right| ee^T |\hat{r}| \right\|_2 \right), \end{aligned} \tag{2.59}$$

gdje smo u 2. nejednakosti iskoristili činjenicu da je $\|AA^+\|_2 = 1$ jer je to projektor, činjenicu da je $\|a\|_2 = \| \, |a| \, \|_2$ za $a \in \mathbb{R}^n$ te nejednakost trokuta, a u 3. nejednakosti (2.58). Želimo zamijeniti $\hat{x}$ i $\hat{r}$ u ogradi pa ih ocjenjujemo:

$$\begin{aligned} |\hat{x}| &\leq |x| + |\hat{x} - x| \leq |x| + |A^+|(|\Delta b| + |\Delta A\hat{x}|) + \left|(A^TA)^{-1}\right| \left|\Delta A^T\right| |\hat{r}| \\ &\leq |x| + \tilde{\gamma}_{mn} |A^+| ee^T (|b| + |A||\hat{x}|) + \tilde{\gamma}_{mn} \left|(A^TA)^{-1}\right| ee^T |A||\hat{r}|, \end{aligned}$$

gdje smo u 2. nejednakosti iskoristili (2.57), a u 3. nejednakosti (2.58). Tada je

$$\begin{aligned} \tilde{\gamma}_{mn} \|ee^T(|b| + |A||\hat{x}|)\|_2 &\leq \tilde{\gamma}_{mn} \|ee^T(|b| + |A||x|)\|_2 \\ &\quad + \tilde{\gamma}_{mn}^2 \left\| |A| \left(|A^+| ee^T (b + |A||\hat{x}|) + \left|(A^TA)^{-1}\right| ee^T |A||\hat{r}| \right) \right\|_2 \\ &\leq \tilde{\gamma}_{mn} \|ee^T(|b| + |A||x|)\|_2 + O(u^2). \end{aligned} \tag{2.60}$$

Na analogan način ocjenujemo $|\hat{r}|$, samo što koristimo $\hat{r} - r$ iz (2.57) dok smo za ocjenu $\hat{x}$ koristili $\hat{x} - x$ iz (2.57). Dobijamo da je

$$|\hat{r}| \leq |r| + O(u)$$

pa je

$$\tilde{\gamma}_{mn} \left\| \left|(A^+)^T\right| \left|A^T\right| ee^T |\hat{r}| \right\|_2 \leq \tilde{\gamma}_{mn} \left\| \left|(A^+)^T\right| \left|A^T\right| ee^T |r| \right\|_2 + O(u^2). \tag{2.61}$$

Iz (2.60) i (2.61) nastavljući (2.59) dobijamo:

$$\begin{aligned}\|(b - A\hat{x}) - r\|_2 &\leq \tilde{\gamma}_{mn}\left(\|ee^T(|b| + |A||x|)\|_2 + \left\|\left|(A^+)^T\right|\left|A^T\right| ee^T|r|\right\|_2\right) + O(u^2)\\ &\leq \tilde{\gamma}_{mn}\|ee^T\|_F\left(\||b| + |A||x|\|_2 + \text{cond}_2(A^T)\|r\|_2\right) + O(u^2)\\ &= m\tilde{\gamma}_{mn}\left(\||b| + |A||x|\|_2 + \text{cond}_2\left(A^T\right)\|r\|_2\right) + O(u^2), \qquad (2.62)\end{aligned}$$

gdje smo u 2. nejednakosti iskoristili nejednakost trokuta, činjenicu da je $\|ee^T\|_2 \leq \|ee^T\|_F$ te definirali $\text{cond}_2(A) := \||A^+||A|\|_2$. Onda iz (2.62) vrijedi:

$$\begin{aligned}\|b - A\hat{x}\|_2 &\leq \|b - A\hat{x} - r\|_2 + \|r\|_2\\ &\leq m\tilde{\gamma}_{mn}\left(\||b| + |A||x|\|_2\right) + \left(1 + m\tilde{\gamma}_{mn}\text{cond}_2\left(A^T\right)\right)\|r\|_2 + O(u^2). \qquad (2.63)\end{aligned}$$

Ova ograda se sastoji od 2 dijela. Izraz $m\tilde{\gamma}_{mn}\left(\||b| + |A||x|\|_2\right)$ je višekratnik ograde za grešku u računanju $fl(b - Ax)$ te je ta greška jako mala. U 2. dijelu izraza je najzanimljiviji faktor $1 + m\tilde{\gamma}_{mn}\text{cond}_2\left(A^T\right)$. Taj faktor će biti manji na primjer od 1.1 sve dok $\text{cond}_2\left(A^T\right)$ nije prevelik. Vrijedi:

$$\text{cond}_2\left(A^T\right) \leq \||A^+|\|_2\|\,|A|\,\|_2 \leq n\|A^+\|_2\|A\|_2 = n\kappa_2(A),$$

gdje smo u 2. nejednakosti iskoristili Lemu 1.3.2. Zaključak je da, osim u slučaju u kojem je matrica A jako loše uvjetovana, rezidual $b - A\hat{x}$ neće preći zbroj reziduala $b - Ax$ i konstantnog višekratnika greške računanja $fl(r)$.

MGS metoda

Bjorck i Paige [1] su pokazali da MGS metoda daje stabilan unatrag po normi način za rješavanje problema najmanjih kvadrata. Dokazali su rezultat sličan Teoremu 2.5.1, razlika je u tome što je rezultat po normi, za razliku od stupčanih ocjena u Teoremu 2.5.1. No vrijedi i rezultat po stupcima [4, p. 566]. Rezultat koji su dokazali je bitan jer pokazuje da to što $\hat{Q}$ nije ortogonalna matrica ne utječe na stabilnost MGS metode kao način za rješavanje problema najmanjih kvadrata.

Poglavlje 3

Programski primjeri

Cilj je provjeriti greške metoda te ograde grešaka koje su dane u teoremima. Primjeri te same metode su isprogramirane u MATLAB R2021a.

3.1 QR faktorizacija

Za prvi primjer ćemo uzeti 3×3 matricu. Neka je

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}.$$

Matrica A ima puni rang te je $\kappa_2(A) = 13.9152$. Što se tiče samih grešaka, provjeravat će se $\|A - \hat{Q}\hat{R}\|_2$ te za QR faktorizaciju pomoću Householderovih reflektora i za QR faktorizaciju pomoću Givensovih rotacije će se provjeravati $\left\|(A - \hat{Q}\hat{R})(:, j)\right\|_2$, $j = 1, 2, 3$. Također, provjerit ćemo ortogonalnost izračunatih matrica. Greške su sljedeće:

$$\begin{aligned}
\text{Householder:} \quad & \|A - \hat{Q}\hat{R}\|_2 = 1.9 \cdot 10^{-14}, \quad && \|\hat{Q}^T\hat{Q} - I_3\|_2 = 6.8 \cdot 10^{-16}, \\
\text{Givens:} \quad & \|A - \hat{Q}\hat{R}\|_2 = 1.5 \cdot 10^{-14}, && \|\hat{Q}^T\hat{Q} - I_3\|_2 = 1.4 \cdot 10^{-16}, \\
\text{CGS:} \quad & \|A - \hat{Q}\hat{R}\|_2 = 7.1 \cdot 10^{-15}, && \|\hat{Q}^T\hat{Q} - I_3\|_2 = 4.0 \cdot 10^{-16}, \\
\text{MGS:} \quad & \|A - \hat{Q}\hat{R}\|_2 = 7.1 \cdot 10^{-15}, && \|\hat{Q}^T\hat{Q} - I_3\|_2 = 2.0 \cdot 10^{-16}.
\end{aligned}$$

Što se tiče samih ograda, ograda za QR faktorizaciju pomoću Householderovih reflektora je dana u (2.4), ograda za QR faktorizaciju pomoću MGS metode je dana u (2.11), a ograda za ortogonalnost izračunate matrice $\hat{Q}$ pomoću MGS metode je dana u (2.12). U ogradama ćemo zamijeniti $\tilde{\gamma}_{mn}$ s γ_{mn} jer vrijedi da je $\gamma_{mn} \leq \tilde{\gamma}_{mn}$.

Ograde su sljedeće:

$$\begin{aligned} \text{Householder:} \quad & \|A - \hat{Q}\hat{R}\|_2 \quad \leq 3.3 \cdot 10^{-13}, \\ \text{MGS:} \quad & \|A - \hat{Q}\hat{R}\|_2 \quad \leq 7.8 \cdot 10^{-13}, \\ \text{Ortogonalnost MGS:} \quad & \|\hat{Q}^T\hat{Q} - I_3\|_2 \quad \leq 2.9 \cdot 10^{-12}. \end{aligned}$$

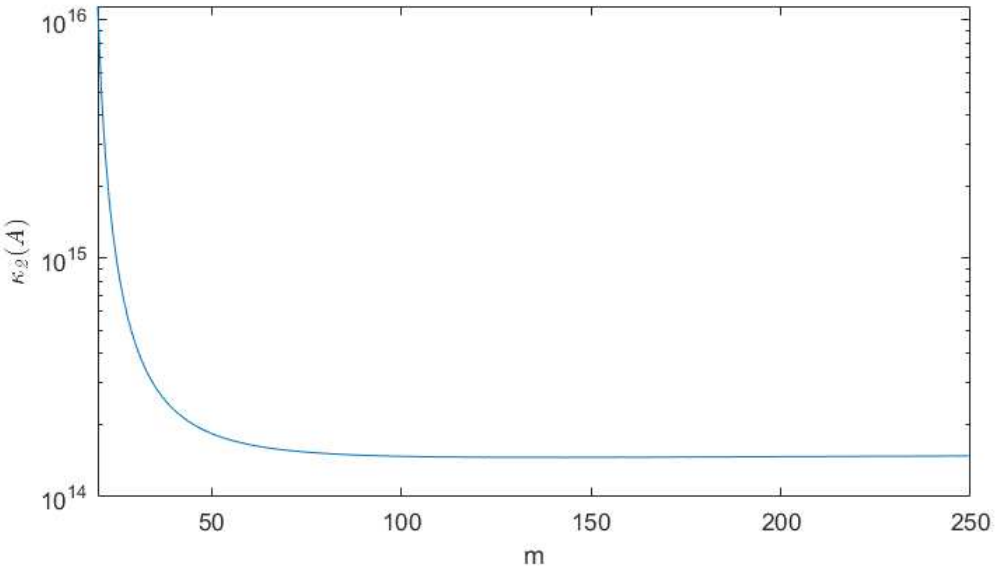
Ograde su zadovoljene te razlike između grešaka i ograda nisu toliko male. Što se tiče grešaka stupaca te ograda za greške stupaca, za QR faktorizaciju pomoću Householderovih reflektora su sljedeće:

$$\begin{aligned} j = 1 \quad & \left\|(A - \hat{Q}\hat{R})(:,1)\right\|_2 = 3.7 \cdot 10^{-15}, \quad & \left\|(A - \hat{Q}\hat{R})(:,1)\right\|_2 \leq 2.4 \cdot 10^{-14}, \\ j = 2 \quad & \left\|(A - \hat{Q}\hat{R})(:,2)\right\|_2 = 0, \quad & \left\|(A - \hat{Q}\hat{R})(:,2)\right\|_2 \leq 3.1 \cdot 10^{-13}, \\ j = 3 \quad & \left\|(A - \hat{Q}\hat{R})(:,3)\right\|_2 = 1.9 \cdot 10^{-14}, \quad & \left\|(A - \hat{Q}\hat{R})(:,3)\right\|_2 \leq 1.4 \cdot 10^{-13}. \end{aligned}$$

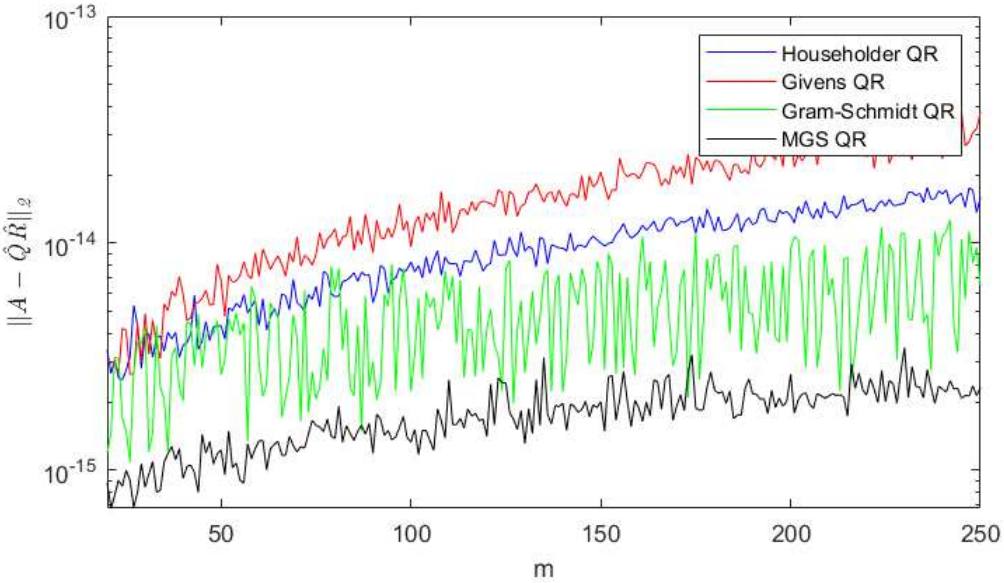
Za drugi primjer definirajmo vektor $x = (x_i)$, gdje su $x_i = \frac{i}{m-1}$, $i = 0, \ldots, m-1$. Za vektor x ćemo definirati Vandermondovu matricu

$$A = \begin{bmatrix} x_0^{n-1} & x_0^{n-2} & \ldots & x_0^1 & 1 \\ x_1^{n-1} & x_1^{n-2} & \ldots & x_1^1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m-1}^{n-1} & x_{m-1}^{n-2} & \ldots & x_{m-1}^1 & 1 \end{bmatrix}.$$

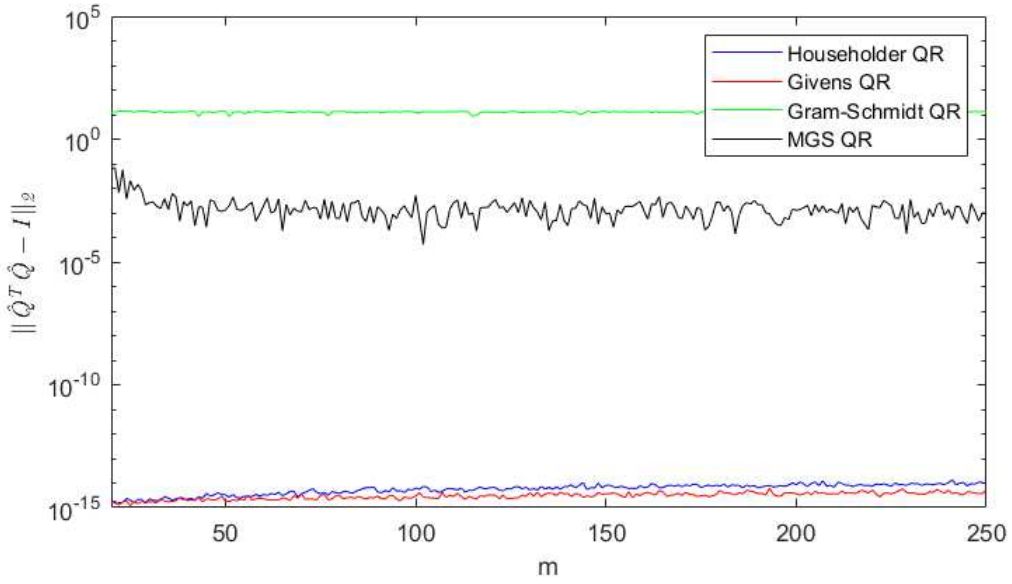
Kako je matrica $A \in \mathbb{R}^{m\times n}$, fiksirat ćemo $n = 20$ te ćemo varirati m. Za m ćemo krenuti od $m = 20$ te završiti s $m = 250$. Vandermondove matrice su loše uvjetovane matrice pa ćemo dobiti nešto drugačije rezultate nego u prvom primjeru. Prikazani grafovi će imati y-os u logaritamskoj skali. Uvjetovanost matrica dana je na Slici 3.1. Greška unatrag QR faktorizacije dobivene različitim metodama je dana na Slici 3.2. Ortogonalnosti izračunatih matrica dane su na Slici 3.3. Ono što možemo primijetiti je da je izračunata matrica $\hat{Q}$ iz QR faktorizacije pomoću Gram - Schmidtove metode nije niti blizu ortogonalna. Također, za izračunatu matricu $\hat{Q}$ znamo da će njena ortogonalnost ovisiti o $\kappa_2(A)$ pa se iz slike može primijetiti da je $\|\hat{Q}^T\hat{Q} - I_n\|_2$ nešto manji na dijelu gdje je $\kappa_2(A)$ nešto manji. Kao i očekivano, u QR faktorizaciji pomoću Householderovih reflektora te u QR faktorizaciji pomoću Givensovih rotacija izračunate matrice su blizu ortogonalne. Na Slici 3.4 vidimo usporedbu između $\|A - \hat{Q}\hat{R}\|_2$, $\max_j \left\|(A - \hat{Q}\hat{R})(:,j)\right\|_2$ te njihovih ograda. U oba slučaja je ograda dosta veća nego sama greška. Zbog loše uvjetovanosti matrice A, ogradu ortogonalnosti izračunate matrice $\hat{Q}$ iz QR faktorizacije pomoću MGS metode nema smisla provjeravati. Jedan od razloga je to što neće biti zadovoljen uvjet $\sqrt{n}c_3u\kappa_2(A) < 1$ pa dalje račun koji se napravio da se dođe do ograde neće biti zadovoljen.



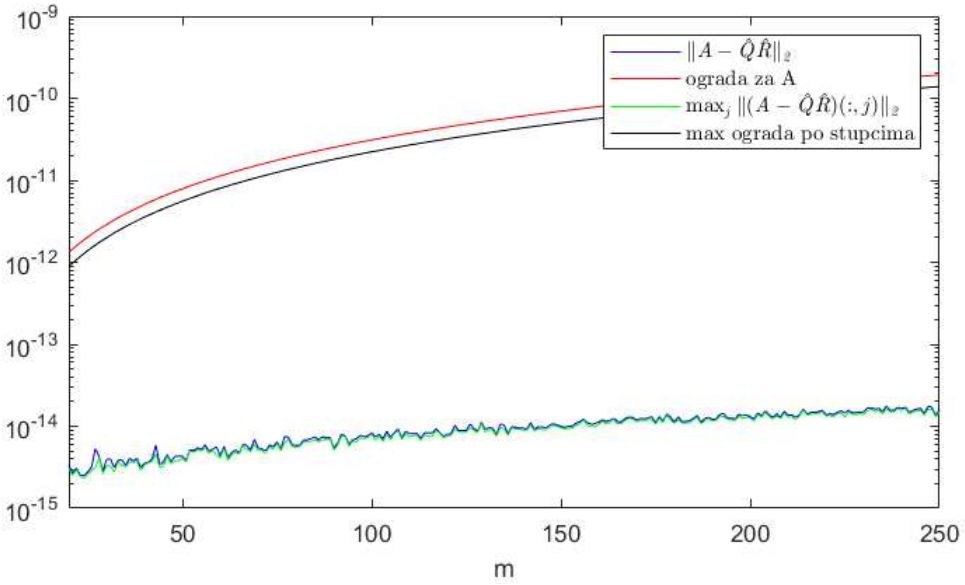
Slika 3.1: Uvjetovanost matrice A



Slika 3.2: Greška unatrag QR faktorizacije



Slika 3.3: Ortogonalnost izračunatih matrica



Slika 3.4: Usporedba greške QR faktorizacije pomoću Householderovih reflektora i ograde za grešku

3.2 Problem najmanjih kvadrata

Primjer će biti linearan sustav, odnosno problem najmanjih kvadrata u kojemu je

$$\min_x \|b - Ax\|_2 = 0.$$

Definiramo matricu i vektor

$$A = \begin{bmatrix} 1 & 3 & -2 \\ 3 & 5 & 6 \\ 2 & 4 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ 7 \\ 8 \end{bmatrix}.$$

Egzaktno rješenje sustava, odnosno problema najmanjih kvadrata je

$$x = \begin{bmatrix} -15 \\ 8 \\ 2 \end{bmatrix}.$$

S obzirom da imamo egzaktno rješenje, provjerit ćemo izraz $\|b - A\hat{x}\|_2$ za sve QR faktorizacije te njegovu ogradu koja je dana u (2.63) za QR faktorizaciju pomoću Householderovih reflektora. Također u ogradi ćemo opet zamijeniti $\tilde{\gamma}_{mn}$ s γ_{mn}. Ograda je sljedeća:

$$\|b - A\hat{x}\|_2 \leq 4.1 \cdot 10^{-13}.$$

Greške su sljedeće:

$$\begin{aligned} \text{Householder:} &\quad \|b - A\hat{x}\|_2 = 1.2 \cdot 10^{-14}, \\ \text{Givens:} &\quad \|b - A\hat{x}\|_2 = 6.2 \cdot 10^{-15}, \\ \text{CGS:} &\quad \|b - A\hat{x}\|_2 = 2.8 \cdot 10^{-14}, \\ \text{MGS:} &\quad \|b - A\hat{x}\|_2 = 2.0 \cdot 10^{-15}. \end{aligned}$$

Kod rješavanja problema najmanjih kvadrata MGS metodom, koristili smo postupak opisan u 1.2 na stranici 13. Greške su očekivano male, s obzirom da smo za sve osim QR faktorizacije pomoću Gram - Schmidtove ortogonalizacije pokazali da su stabilne unatrag. S obzirom da imamo egzaktno rješenje, možemo pogledati i grešku unaprijed:

$$\begin{aligned} \text{Householder:} &\quad \|\hat{x} - x\|_2 = 2.4 \cdot 10^{-14}, \\ \text{Givens:} &\quad \|\hat{x} - x\|_2 = 8.9 \cdot 10^{-16}, \\ \text{CGS:} &\quad \|\hat{x} - x\|_2 = 2.5 \cdot 10^{-13}, \\ \text{MGS:} &\quad \|\hat{x} - x\|_2 = 1.2 \cdot 10^{-14}. \end{aligned}$$

Bibliografija

[1] Å. Björck i C. C. Paige, *Loss and Recapture of Orthogonality in the Modified Gram–Schmidt Algorithm*, SIAM Journal on Matrix Analysis and Applications **13** (1992), br. 1, 176–190, https://doi.org/10.1137/0613015.

[2] Nicholas J. Higham, *Matrix Nearness Problems and Applications*, Applications of Matrix Theory (M. J. C. Gover i S. Barnett, ur.), Oxford University Press, 1989, str. 1–27.

[3] ———, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra and its Applications **212-213** (1994), 3–20, ISSN 0024-3795, https://www.sciencedirect.com/science/article/pii/002437959490393X.

[4] ———, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002, ISBN 0-89871-521-0.

[5] G. W. Stewart, *On the Perturbation of Pseudo-Inverses, Projections and Linear Least Squares Problems*, SIAM Review **19** (1977), br. 4, 634–662, https://doi.org/10.1137/1019104.

[6] ———, *Perturbation Theory for the Singular Value Decomposition*, Teh. izv., USA, 1990.

[7] Per Åke Wedin, *Perturbation theory for pseudo-inverses*, BIT Numerical Mathematics **13** (1973), 217–232.

[8] Hermann Von Weyl, *Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)*, Mathematische Annalen **71** (1912), 441–479.

Sažetak

U ovom radu smo prvo uveli osnovne pojmove vezane uz QR faktorizaciju, problem najmanjih kvadrata te analizu greške. Uveli smo načine za računanje QR faktorizacije te za rješavanje problema najmanjih kvadrata. Glavni cilj rada je bio napraviti analizu greške svakog od tih načina. Za QR faktorizaciju pomoću Householderovih reflektora te Givensovih rotacija smo prvo krenuli od stabilnosti računanja Householderovim reflektorima, odnosno Givensovim rotacija, te stabilnosti operacija s njima. Pomoću toga smo dokazali glavne rezultate, stabilnost unatrag QR faktorizacije pomoću Householderovih reflektora odnosno Givensovih rotacija. Za metode CGS i MGS smo pokazali da izračunati ortogonalni faktor nije nužno ortogonalan te smo dokazali stabilnost unatrag MGS metode. Za problem najmanjih kvadrata smo prvo dokazali Wedinov perturbacijski teorem te izveli perturbacijske ograde po komponentama. Dokazali smo da su QR faktorizacija pomoću Householderovih reflektora te QR faktorizacija pomoću Givensovih rotacija stabilne unatrag metode za rješavanje problema najmanjih kvadrata. Također smo naveli rezultat da je i MGS metoda stabilan unatrag način za rješavanje unatoč problemima s ortogonalnošću. Za kraj smo neke od ograda, iz rezultata koje smo dokazali, testirali programskim primjerima u MATLAB-u.

Summary

This thesis firstly introduces basic terms related to QR factorization, least square problem and error analysis. We introduced methods to compute QR factorization and to compute the solution of the least square problem. The main goal of this thesis was to do an error analysis for each of these methods. For Householder and Givens QR factorization, we first proved the stability of the construction and operations with Householder reflectors and Givens rotations. With that we proved our main results, backward stability of Householder and Givens QR factorization. For the CGS and MGS methods we showed that the computed orthogonal factor isn't necessarily orthogonal and then we proved backward stability for the MGS method. For the least square problem, we first proved Wedin's perturbation theorem and we derived componentwise perturbation bounds. We proved that Householder and Givens QR factorization provide a backward stable way to solve the least square problem. We also stated that the MGS method provides a backward stable way to solve the least square problem even though its computed orthogonal factor isn't orthogonal. In the end, we tested some of the bounds we obtained with programmed examples in MATLAB.

Životopis

Rođen sam 23. listopada 1997. u Zagrebu. Pohađao sam Osnovnu školu Većeslava Holjevca te V. Gimnaziju u Zagrebu. Upisao sam 2016. godine preddiplomski sveučilišni studij Matematika na Prirodoslovnom-matematičkom fakultetu u Zagrebu koji sam završio 2019. godine. Iste godine sam upisao diplomski sveučilišni studij Primijenjena matematika na istom fakultetu.